Applied Statistical Methods II

Larry Winner
University of Florida
Department of Statistics

August 16, 2021

Chapter 1

Simple Linear Regression

1.1 Introduction

Linear regression is used when there is a numeric response variable and numeric (and possibly categorical) predictor (explanatory) variable(s). The mean of the response variable is to be related to the predictor(s) with random error terms assumed to be independent and normally distributed with constant variance. The fitting of linear regression models is very flexible, allowing for fitting curvature and interactions between factors. In this chapter, the case of a single predictor variable is covered in detail. The methods described here generalize (for the most part) directly to the case when there are $p \geq 2$ predictor variables included in the model.

1.2 Basic Simple Linear Regression Model

When there is a single numeric predictor, the model is referred to as **Simple Regression**. The (random) response variable is denoted as Y and the predictor variable is denoted as X. The basic model is written as follows.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad\qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

Here β_0 is the intercept (mean of Y when X=0) and β_1 is the slope (the change in the mean of Y when X increases by 1 unit). Of primary concern is whether $\beta_1 = 0$, which implies the mean of Y is constant (β_0), and thus Y and X are not associated (at least in a linear pattern). In many applications, transformations of Y and/or X are needed to meet the assumptions of the model.

Example 1.1: Galvonometer Deflection in Experiments with Explosives

An experiment was conducted (McNab and Ristori (1899-1900), [23]), relating the deflection of a galvonome-

obsNum (i)	wireArea (X_i)	galvDef (Y_i)	obsNum (i)	wireArea (X_i)	galvDef (Y_i)
1	152	85	12	53	149
2	152	81.5	13	53	149
3	152	83.5	14	53	147
4	125	102	15	35	152.5
5	125	90.5	16	35	158.5
6	125	98.5	17	35	151
7	99	109	18	25	161
8	99	115.5	19	25	170
9	66	131	20	25	166
10	66	128.5	21	17	185.5
11	66	138.5	22	17	192

Table 1.1: Galvonometer Deflection (Y) and Wire Area (X) in an Explosion Experiment

ter (Y, in mm) as a linear function of the area of the wire (X in 1/100000 in) in explosion research, based on $n = 22$ experimental observations. The data are given in Table 1.1 and a plot of the data and the fitted simple regression line are given in Figure 1.1.

∇

1.3 Estimation of Model Parameters

We obtain a sample of n pairs $(X_i, Y_i) \quad i = 1, \ldots, n$. Our goal is to choose estimators of β_0 and β_1 that minimize the error sum of squares: $Q = \sum_{i=1}^n \epsilon_i^2$. The resulting estimators are (from calculus) given below after introducing some useful notation.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, \ldots, n \qquad\qquad Q = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n \left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2$$

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \qquad \overline{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad SS_{XX} = \sum_{i=1}^n (X_i - \overline{X})^2 \qquad SS_{XY} = \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \qquad SS_{YY} = \sum_{i=1}^n (Y_i - \overline{Y})^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{SS_{XY}}{SS_{XX}} = \sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right) Y_i$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \sum_{i=1}^n \left[\frac{1}{n} + \frac{\left(X_i - \overline{X}\right) \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] Y_i$$

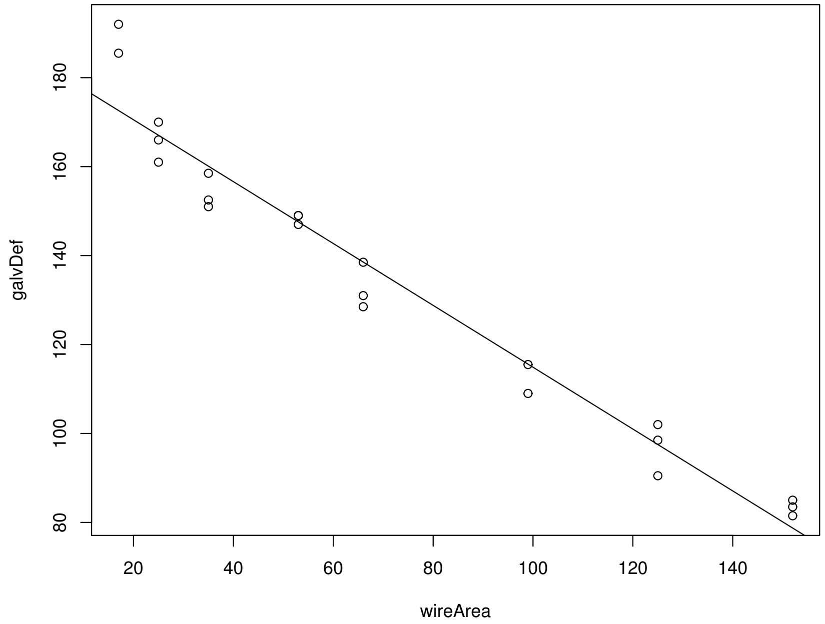

Figure 1.1: Plot of Galvonomer Deflection (Y) and Wire Area (X) and the fitted regression equation

Once estimates are computed, **fitted values** and **residuals** are obtained for each observation. The **error sum of squares (SSE)** is obtained as the sum of the squared residuals as follows.

$$\text{Fitted Values: } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad \text{Residuals: } e_i = Y_i - \hat{Y}_i \qquad i = 1, \ldots, n$$

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = SS_{YY} - \frac{(SS_{XY})^2}{SS_{XX}}$$

The (unbiased) estimate of the error variance σ^2 is $s^2 = MSE = \frac{SSE}{n-2}$, where MSE is the **Mean Square Error**. The subtraction of 2 can be thought of as the fact that two parameters have been estimated : β_0 and β_1.

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear functions of $Y_1, \ldots, Y_n$ and thus using basic rules of mathematical statistics, their sampling distributions are given below. Note first that for constants $a_1, \ldots, a_n$ and random variables $Y_1, \ldots, Y_n$, the following results are obtained.

$$E\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i E\left\{Y_i\right\} \qquad V\left\{\sum_{i=1}^{n} a_i Y_i\right\} = \sum_{i=1}^{n} a_i^2 V\left\{Y_i\right\} + 2\sum_{i=1}^{n-1}\sum_{i'=i+1}^{n} a_i a_{i'} \text{COV}\left\{Y_i, Y_{i'}\right\}$$

In the case of the simple linear regression model described above, the following results are used, where the X^s are assumed to be fixed constants.

$$E\left\{Y_i\right\} = \beta_0 + \beta_1 X_i \qquad V\left\{Y_i\right\} = \sigma^2 \qquad i \neq i' \Rightarrow \text{COV}\left\{Y_i, Y_{i'}\right\} = 0$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right) Y_i \qquad E\left\{\hat{\beta}_1\right\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right) (\beta_0 + \beta_1 X_i) = \beta_1$$

$$V\left\{\hat{\beta}_1\right\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \right)^2 \sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]\right)$$

The standard error is the square root of the variance, and the estimated standard error is the standard error with the unknown σ^2 replaced by MSE.

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

<u>**Example 1.2: Galvonometer Deflection in Experiments with Explosives**</u>

The computations necessary to obtain the parameter estimates and their estimated standard errors for the explosion experiment are contained in Table 1.2, they were originally computed in EXCEL.

$$\hat{\beta}_1 = \frac{-32715.80}{47048.36} = -0.6954 \qquad \hat{\beta}_0 = 133.864 - 72.727(-0.6954) = 184.438$$

$$SSE = 1076.63 \qquad s^2 = MSE = \frac{1076.63}{22-2} = 53.83$$

$$\hat{SE}\left\{\hat{\beta}_1\right\} = \sqrt{\frac{53.83}{47048.36}} = 0.0338 \qquad \hat{SE}\left\{\hat{\beta}_0\right\} = \sqrt{53.83\left[\frac{1}{22} + \frac{(72.727)^2}{47048.36}\right]} = 2.915$$

∇

i	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$\left(X_i - \overline{X}\right)^2$	$\left(Y_i - \overline{Y}\right)^2$	$\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$	$\hat{Y}_i$	e_i	e_i^2
1	152	85	79.273	-48.864	6284.165	2387.655	-3873.55	78.74	6.26	39.184
2	152	81.5	79.273	-52.364	6284.165	2741.95	-4151.01	78.74	2.76	7.616
3	152	83.5	79.273	-50.364	6284.165	2536.496	-3992.46	78.74	4.76	22.655
4	125	102	52.273	-31.864	2732.438	1015.291	-1665.6	97.52	4.48	20.114
5	125	90.5	52.273	-43.364	2732.438	1880.405	-2266.74	97.52	-7.02	49.212
6	125	98.5	52.273	-35.364	2732.438	1250.587	-1848.55	97.52	0.98	0.970
7	99	109	26.273	-24.864	690.256	618.2	-653.236	115.59	-6.59	43.490
8	99	115.5	26.273	-18.364	690.256	337.223	-482.463	115.59	-0.09	0.009
9	66	131	-6.727	-2.864	45.256	8.2	19.264	138.54	-7.54	56.878
10	66	128.5	-6.727	-5.364	45.256	28.769	36.083	138.54	-10.04	100.836
11	66	138.5	-6.727	4.636	45.256	21.496	-31.19	138.54	-0.04	0.002
12	53	149	-19.727	15.136	389.165	229.11	-298.599	147.58	1.42	2.012
13	53	149	-19.727	15.136	389.165	229.11	-298.599	147.58	1.42	2.012
14	53	147	-19.727	13.136	389.165	172.564	-259.145	147.58	-0.58	0.338
15	35	152.5	-37.727	18.636	1423.347	347.314	-703.099	160.10	-7.60	57.730
16	35	158.5	-37.727	24.636	1423.347	606.95	-929.463	160.10	-1.60	2.554
17	35	151	-37.727	17.136	1423.347	293.655	-646.508	160.10	-9.10	82.775
18	25	161	-47.727	27.136	2277.893	736.382	-1295.15	167.05	-6.05	36.623
19	25	170	-47.727	36.136	2277.893	1305.837	-1724.69	167.05	2.95	8.692
20	25	166	-47.727	32.136	2277.893	1032.746	-1533.78	167.05	-1.05	1.106
21	17	185.5	-55.727	51.636	3105.529	2666.314	-2877.55	172.61	12.89	166.033
22	17	192	-55.727	58.136	3105.529	3379.837	-3239.78	172.61	19.39	375.792
Mean	72.727	133.864						133.864	0.00	
Sum	1600	2945	0	0	47048.36	23826.09	-32715.8	2945.00	0.00	1076.63

Table 1.2: Galvonomer Deflection (Y) and Wire Area (X) calculations to obtain parameter estimates

1.4 Inferences Regarding β_1 and β_0

Primarily of interest are inferences regarding β_1. Note that if $\beta_1 = 0$, Y and X are not associated. Hypotheses can be tested and confidence intervals constructed based on the estimate β_1 and its estimated standard error. The t-test is conducted as follows. Note that the null value β_{10} is almost always 0, and that software packages that report these tests always are treating β_{10} as 0. Here, and in all other tests, TS represents Test Statistic, and RR represents Rejection Region.

$$H_0 : \beta_1 = \beta_{10} \qquad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{SE}\{\hat{\beta}_1\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P\text{-value} : 2P(t_{n-2} \geq |t_{obs}|)$$

One-sided tests use the same test statistic, but the Rejection Region and P-value are changed to reflect the alternative hypothesis:

$$H_A^+ : \beta_1 > \beta_{10} \qquad RR : t_{obs} \geq t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \geq t_{obs})$$

$$H_A^- : \beta_1 < \beta_{10} \qquad RR : t_{obs} \leq -t_{\alpha,n-2} \qquad P\text{-value} : P(t_{n-2} \leq t_{obs})$$

A $(1-\alpha)100\%$ Confidence Interval for β_1 is obtained as follows.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{SE}\{\hat{\beta}_1\}$$

Note that the confidence interval represents the values of β_{10} that the two-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ fails to reject the null hypothesis.

Inferences regarding β_0 are only of interest when it is meaningful to estimate the mean when $X = 0$ or when the data have been centered by subtracting off the mean ($x_i = X_i - \overline{X}$) where the intercept represents the mean when $X = \overline{X}$. Inference can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $\hat{SE}\{\hat{\beta}_0\}$.

Example 1.3: Galvonometer Deflection in Experiments with Explosives

A test of whether the mean galvonemeter deflection is (linearly) related to the wire area can be tested as $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. Based on the plot in Figure 1.1, it appears to be very certain that there is a negative association. The test, conducted as 2-sided with $\alpha = 0.05$ and 95% Confidence Intervals for β_1 and β_0 are given below.

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \quad TS : t_{obs} = \frac{-0.6954 - 0}{0.0338} = -20.57 \qquad RR : |t_{obs}| \geq t_{.025,22-2} = 2.086$$

$$P\text{-value} : 2P(t_{20} \geq |-20.57|) < .0001$$

$$95\% \text{ CI for } \beta_1 : \quad -0.6954 \pm 2.086(0.0338) \quad \equiv \quad -0.6954 \pm 0.0705 \quad \equiv \quad (-0.7659, -0.6249)$$

$$95\% \text{ CI for } \beta_0 : \quad 188.438 \pm 2.086(2.915) \quad \equiv \quad 188.438 \pm 6.081 \quad \equiv \quad (182.357, 194.519)$$

There is strong evidence that as the wire area increases, the mean galvonemeter deflection decreases. As it would be impossible to have a wire area of 0, there is no physical interpretation of the intercept β_0.

∇

1.5 Estimating the Mean and Predicting a New Observation @ $X = X^*$

In some cases, it is of interest to estimate the mean response at a specific level X^*. The parameter of interest is $\mu^* = \beta_0 + \beta_1 X^*$. The point estimate, standard error, and $(1-\alpha)100\%$ Confidence Interval are given below.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]} \qquad (1-\alpha)100\% \text{ CI } : \hat{Y}^* \pm t_{\alpha/2,n-2} \hat{SE}\left\{\hat{Y}^*\right\}$$

To obtain a simultaneous $(1-\alpha)100\%$ Confidence Interval for the entire regression line (not just a single point), the Working-Hotelling method can be used.

$$\hat{Y}^* \pm \sqrt{2F_{\alpha/2,2,n-2}} SE\left\{\hat{Y}^*\right\}$$

If interest is in predicting a new observation when $X = X^*$, there is uncertainty with respect to estimating the mean (as seen by the Confidence Interval above), and the random error for the new case (with standard deviation σ). The point prediction is the same as for the mean. The estimate, standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below.

$$\hat{Y}^*_{\text{New}} = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\% \text{ PI } : \hat{Y}^*_{\text{New}} \pm t_{\alpha/2,n-2} \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\}$$

Note that the Prediction Interval can be much wider than the Confidence Interval for the mean, particularly when MSE is large.

Example 1.4: Galvonometer Deflection in Experiments with Explosives

Suppose interest is in estimating the mean galvonometer deflection when the wire area is $X^* = 110$. Further, assume that an upcoming experiment will be run at the same level of wire area. Note that the first case is considering the long run average of all potential experiments at this level, while the second case is for a single outcome. The prediction is the same for both cases, the estimated standard errors differ.

$$\hat{Y}^* = \hat{Y}^*_{\text{New}} = 184.438 - 0.6954(110) = 184.438 - 76.494 = 107.944 \qquad \overline{X} = 72.727 \qquad MSE = 53.83$$

$$\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{1}{22} + \frac{(110 - 72.727)^2}{47048.36} = 0.0750$$

$$\hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{53.83(0.0750)} = 2.009 \qquad \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{53.83(1.0750)} = 7.607$$

$$\begin{aligned}(1-\alpha)100\% \text{ CI } &: 107.944 \pm 2.086(2.009) \quad \equiv \quad 107.944 \pm 4.191 \quad \equiv \quad (103.753, 112.135)\\ (1-\alpha)100\% \text{ PI } &: 107.944 \pm 2.086(7.607) \quad \equiv \quad 107.944 \pm 15.868 \quad \equiv \quad (92.076, 123.812)\end{aligned}$$

If the goal was to obtain simultaneous Confidence Intervals for the mean along the entire regression line, the multiplier of the standard error of the mean would replace $t_{.025,20} = 2.086$ with $\sqrt{2F_{.025,2,20}} = \sqrt{2(4.461)} = 2.987$. Thus, the individual confidence intervals would be almost 50% wider (2.987/2.086=1.43). The pointwise Confidence Interval for the mean and Prediction Interval for an individual observation are given in Figure 1.2.

∇

1.6 Analysis of Variance

When there is no linear association between Y and X ($\beta_1 = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = SS_{YY} = \sum_{i=1}^n (Y_i - \overline{Y})^2$, the **Total Sum of Squares**. Technically it is the Total (Corrected for the Mean) Sum of Squares.

When there is an association between Y and X ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (in terms of minimizing the sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on X (as opposed to having ignored X). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$$

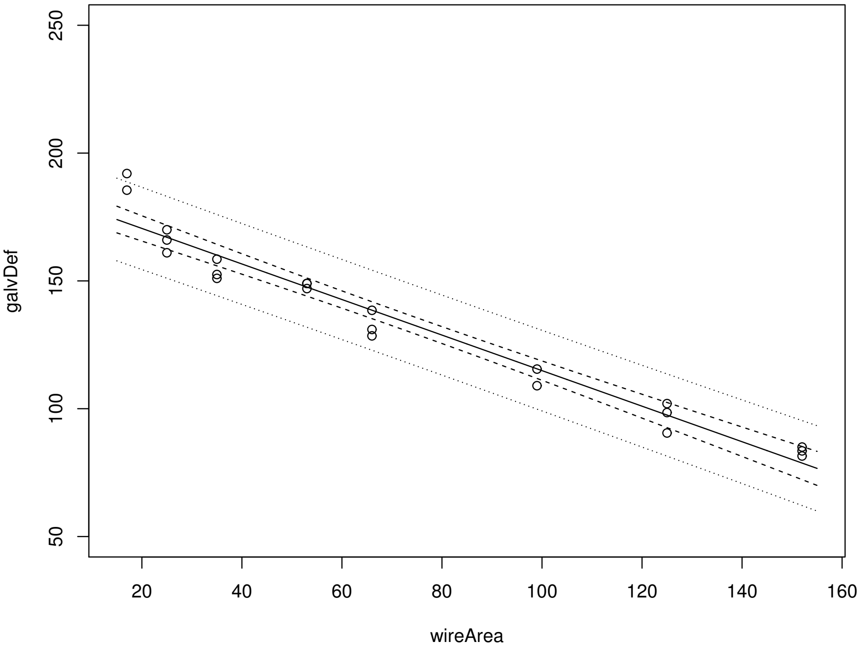

Figure 1.2: Plot of data, fitted equation (middle solid line), pointwise confidence intervals for the mean (inner dashed lines), and pointwise prediction intervals for individual observations - Explosive Experiment Data

Source	df	SS	MS	F_{obs}	P-value
Regression	1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{1,n-2} \geq F_{obs})$
Error (Residual)	$n-2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 1.3: Analysis of Variance Table for Simple Linear Regression

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** is $df_{\text{Regression}} = 1$ (for simple regression).

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - 2 + 1$$

The Error and Regression Sums of Squares each have a **Mean Square**, which is the Sum of Squares divided by its corresponding Degrees of Freedom: $MSE = SSE/(n-2)$ and $MSR = SSR/1$. It can be shown that these Mean Squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

Note that when $\beta_1 = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A second way of testing whether $\beta_1 = 0$ is by the F-test.

$$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0 \quad TS: F_{obs} = \frac{MSR}{MSE} \qquad RR: F_{obs} \geq F_{\alpha,1,n-2} \qquad P\text{-value}: P(F_{1,n-2} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table 1.3.

A measure often reported from a regression analysis is the **Coefficient of Determination** or r^2. This represents the variation in Y "explained" by X, divided by the total variation in Y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq r^2 \leq 1$$

The interpretation of r^2 is the proportion of variation in Y that is "explained" by X, and is often reported as a percentage ($100r^2$).

Example 1.5: Galvonometer Deflection in Experiments with Explosives

Source	*df*	SS	MS	F_{obs}	P-value
Regression	1	22749.46	22749.46	422.62	$< .0001$
Error (Residual)	20	1076.63	53.83		
Total (Corrected)	21	23826.09			

Table 1.4: Analysis of Variance Table for the Explosives Experiment

For the explosives data, the Total and Error sums of squares were previously computed, and the Regression sum of squares is obtained by subtraction.

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 23826.09 \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 1076.63 \qquad SSR = 23826.09 - 1076.63 = 22749.46$$

The F-statistic and test for $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$ is given below. The Analysis of Variance is given in Table 1.4.

$$MSR = \frac{22749.46}{1} = 22749.46 \qquad MSE = \frac{1076.63}{22-2} = 53.83 \qquad F_{obs} = \frac{22749.46}{53.83} = 422.62$$

The coefficient of determination is $r^2 = 22749.46/23826.09 = .9548$. Over 95% of the variation in galvonometer deflection is "explained" by the amount of wire area.

∇

1.7 Correlation

The regression coefficient β_1 depends on the units of Y and X. It also depends on which variable is the dependent variable and which is the independent variable. A second widely reported measure is the **Pearson Product Moment Coefficient of Correlation**. It is invariant to linear transformations of Y and X, and does not distinguish which is the dependent and which is the independent variable. This makes it a widely reported measure when researchers are interested in how two (or more) random variables vary together in a population. The population correlation coefficient is labeled ρ, and the sample correlation is labeled r, and is computed as follows.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} = \left(\frac{s_X}{s_Y}\right)\hat{\beta}_1$$

where s_X and s_Y are the standard deviations of X and Y, respectively. While $\hat{\beta}_1$ can take on any value, r lies between -1 and +1, taking on the extreme values if all of the points fall on a straight line. The test of whether $\rho = 0$ is mathematically equivalent to the t-test for testing whether $\beta_1 = 0$. The 2-sided test is given below.

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0 \quad TS : t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P - \text{value} : 2P(t_{n-2} \geq |t_{obs}|)$$

To construct a large-sample confidence interval, **Fisher's z transform** is used to transform r so that it has an approximately normal sampling distribution. A confidence interval is constructed for the transformed correlation, then "back transformed" for a Confidence Interval for ρ.

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (1-\alpha)100\%\text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) : z' \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Labeling the endpoints of the Confidence Interval as (a,b), the Confidence Interval for ρ is computed.

$$(1-\alpha)100\%\text{ Confidence Interval for } \rho : \left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right)$$

Example 1.6: Galvonometer Deflection in Experiments with Explosives

For the explosives experiment, the sample correlation, r, the t-test, and a 95% Confidence Interval for the population correlation, ρ, are computed below.

$$SS_{XY} = -32715.80 \quad SS_{XX} = 47048.36 \quad SS_{YY} = 23826.09 \quad r = \frac{-32715.80}{\sqrt{47048.36(23826.09)}} = \frac{-32715.80}{33481.02} = -.9771$$

$$H_0 : \rho = 0 \quad H_A : \rho \neq 0 \qquad TS : t_{obs} = \frac{-.9771}{\sqrt{\frac{1-(-.9771)^2}{22-2}}} = \frac{-.9771}{.0476} = 20.54$$

$$z' = \frac{1}{2}\ln\left(\frac{1+(-.9771)}{1-(-.9771)}\right) = \frac{1}{2}(-4.4582) = -2.2291 \qquad z_{.025}\sqrt{\frac{1}{22-3}} = 1.96(.2294) = .4497$$

$$a = -2.2291 - 0.4497 = -2.6788 \qquad b = -2.2291 + 0.4497 = -1.7794$$

$$95\%\text{CI for } \rho : \quad \left(\frac{e^{2(-2.6788)}-1}{e^{2(-2.6788)}+1}, \frac{e^{2(-1.7794)}-1}{e^{2(-1.7794)}+1}\right) \quad \equiv \quad \left(\frac{-.9953}{1.0047}, \frac{-.9715}{1.0285}\right) \quad \equiv \quad (-.9906, -.9446)$$

∇

1.8 Model Diagnostics

The inferences regarding the simple linear regression model (tests and confidence intervals) are based on the following assumptions.

- Relation between Y and X is linear
- Errors are normally distributed
- Errors have constant variance
- Errors are independent

These assumptions can be checked graphically, as well as by statistical tests. Further, models can be extended to allow for the assumptions not being met.

1.8.1 Assumption of Linearity

A plot of the residuals versus X (or $\hat{Y}$) should be a random cloud of points centered at 0 (they sum to 0). A "U-shaped", "inverted U-shaped", or "J-shaped" pattern is inconsistent with linearity.

A test for linearity can be conducted when there are repeat observations at certain X-levels (methods have also been developed to "group" X levels). Suppose there are c distinct X-levels, with n_j observations at the j^{th} level. It is useful for the data to be re-labeled as Y_{ij} where j represents the X group, and i represents the individual case within the group ($i = 1, \ldots, n_j$). The following quantities are computed.

$$\overline{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \qquad\qquad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$$

The Error Sum of Squares is decomposed into **Pure Error** and **Lack of Fit**.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j \left(\overline{Y}_j - \hat{Y}_j\right)^2 \qquad SSE = SSPE + SSLF$$

Partition the error degrees of freedom $(n-2)$ into Pure Error $(n-c)$ and Lack of Fit $(c-2)$. This leads to an F-test for testing H_0: Relation is Linear versus H_A: Relation is not Linear.

$$H_0 : E\left\{Y_{ij}\right\} = \mu_j = \beta_0 + \beta_1 X_j \qquad H_A : E\left\{Y_{ij}\right\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

$$TS : F_{LF} = \frac{[SSLF/(c-2)]}{[SSPE/(n-c)]} = \frac{MSLF}{MSPE} \qquad RR : F_{LF} \geq F_{\alpha,c-2,n-c} \qquad P\text{-Value } : P\left(F_{c-2,n-c} \geq F_{LF}\right)$$

If the relationship is not linear, polynomial terms can be added to allow for "bends" in the relationship between Y and X using multiple regression, or a nonlinear regression model (in the parameters) can be fit.

Example 1.7: Galvonometer Deflection in Experiments with Explosives

Group (j)	X_{ij}	Y_{ij}	$\overline{Y}_j$	$\hat{Y}_j$	e_{ij}	$Y_{ij} - \overline{Y}_j$	$\overline{Y}_j - \hat{Y}_j$
1	152	85.0	83.3333	78.7372	6.2628	1.6667	4.5961
1	152	81.5	83.3333	78.7372	2.7628	-1.8333	4.5961
1	152	83.5	83.3333	78.7372	4.7628	0.1667	4.5961
2	125	102.0	97.0000	97.5130	4.4870	5.0000	-0.5130
2	125	90.5	97.0000	97.5130	-7.0130	-6.5000	-0.5130
2	125	98.5	97.0000	97.5130	0.9870	1.5000	-0.5130
3	99	109.0	112.2500	115.5934	-6.5934	-3.2500	-3.3434
3	99	115.5	112.2500	115.5934	-0.0934	3.2500	-3.3434
4	66	131.0	132.6667	138.5416	-7.5416	-1.6667	-5.8749
4	66	128.5	132.6667	138.5416	-10.0416	-4.1667	-5.8749
4	66	138.5	132.6667	138.5416	-0.0416	5.8333	-5.8749
5	53	149.0	148.3333	147.5818	1.4182	0.6667	0.7515
5	53	149.0	148.3333	147.5818	1.4182	0.6667	0.7515
5	53	147.0	148.3333	147.5818	-0.5818	-1.3333	0.7515
6	35	152.5	154.0000	160.0990	-7.5990	-1.5000	-6.0990
6	35	158.5	154.0000	160.0990	-1.5990	4.5000	-6.0990
6	35	151.0	154.0000	160.0990	-9.0990	-3.0000	-6.0990
7	25	161.0	165.6667	167.0530	-6.0530	-4.6667	-1.3863
7	25	170.0	165.6667	167.0530	2.9470	4.3333	-1.3863
7	25	166.0	165.6667	167.0530	-1.0530	0.3333	-1.3863
8	17	185.5	188.7500	172.6162	12.8838	-3.2500	16.1338
8	17	192.0	188.7500	172.6162	19.3838	3.2500	16.1338
Sum of Squares					1076.63	246.92	829.72

Table 1.5: Galvonomer Deflection (Y) and Wire Area (X) calculations to obtain F-test for Lack of Fit

The explosives experiment was conducted at $c = 8$ distinct levels of wire area (X). The sample sizes were $n_j = 3$ for 6 of the levels and $n_j = 2$ for the other 2 levels. Table 1.5 contains the data, group means, group fitted values from the linear regression, and the error decomposition into Pure Error and Lack of Fit. The sums of squares are obtained in the bottom row, using the **SUMSQ** function in EXCEL.

$$n = 22 \quad c = 8 \quad SSPE = 246.92 \quad SSLF = 829.72 \qquad MSPE = \frac{246.92}{22-8} = 17.637 \quad MSLF = \frac{889.72}{8-2} = 138.287$$

$$TS : F_{LF} = \frac{138.287}{17.637} = 7.841 \qquad RR : F_{LF} \geq F_{\alpha,6,14} = 2.848 \qquad P\text{-Value } : P\left(F_{6,14} \geq 7.841\right) = .0008$$

There is strong evidence that the relation between mean galvonometer deflection and wire area is not linear. The plot of the residuals versus the fitted values does reveal a “U-shape”, as seen in Figure 1.3. A polynomial model will be considered later.

∇

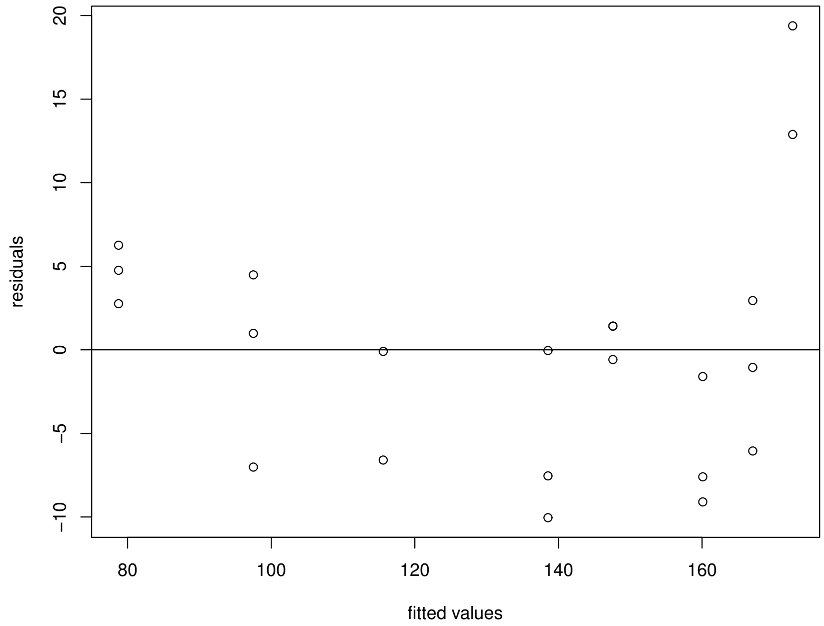

Figure 1.3: Plot of residuals versus fitted values for the Explosives Experiment

1.8.2 Assumption of Normality

A normal probability plot of the ordered residuals versus their predicted values based on a normal distribution should fall approximately on a straight line. A histogram should be mound-shaped. Neither of these methods work well with small samples (even data generated from a normal distribution will not necessarily look like it is normal).

Various tests are computed directly by statistical computing packages. The Shapiro-Wilk and Kolmogorov-Smirnov tests are commonly reported, with resulting P-values for testing H_0: Errors are normally distributed.

Example 1.8: Galvonometer Deflection in Experiments with Explosives

The Shapiro-Wilk statistic based on the residuals (obtained in R), yields a P-value of .0900. The normal probability plot is given in Figure 1.4. In the plot, the residuals are not particularly close to the straight line. The sample size is fairly small for checking normality.

∇

When data are not normally distributed, the **Box-Cox transformation** is often applied. This involves fitting regression models for various power transformations of Y on X, where:

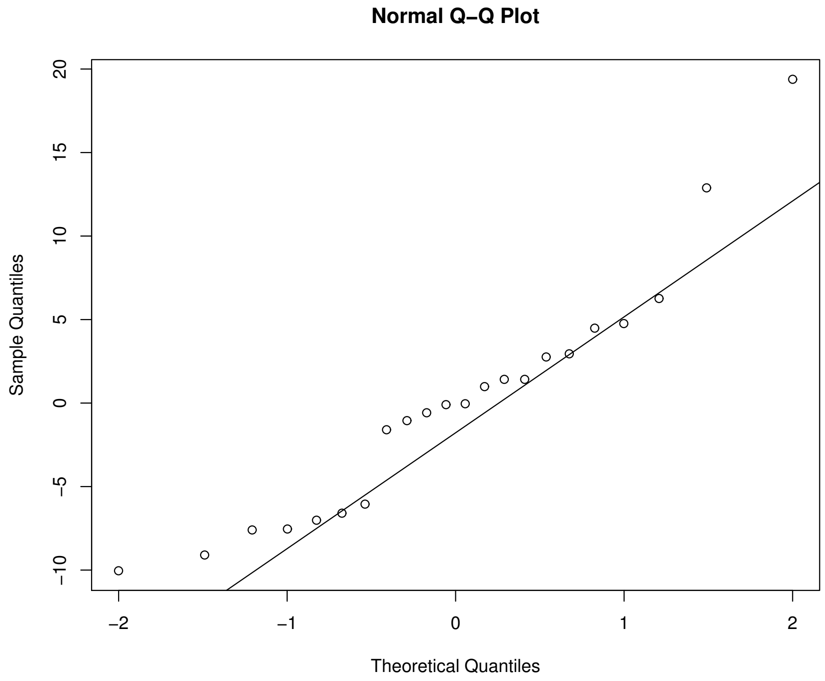

Figure 1.4: Normal probability plot for residuals from linear regression model for the Explosives Experiment

$$Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^{\lambda}-1}{\lambda(\dot{Y})^{(\lambda-1)}} & \lambda \neq 0 \\ \dot{Y}\ln(Y_i) & \lambda = 0 \end{cases}$$

Here $\dot{Y}$ is the geometric mean of $Y_1, \ldots, Y_n$, where all observations are strictly positive (a constant can be added to all observations to assure this).

$$\dot{Y} = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} = \exp\left\{\frac{\sum_{i=1}^{n}\ln(Y_i)}{n}\right\}$$

Values of λ between -2 and 2 by small increments are typically run, and the value of λ that has the smallest Error Sum of Squares (equivalently Maximum Likelihood) is identified. Statistical software packages will present a confidence interval for λ as well.

Example 1.9: Spread of Shotgun Pellets

A forensic experiment was conducted (Rowe and Hanson, 1985, [30]) to determine the relationship between the spread of shotgun pellets (Y, square root of area of the rectangle containing the pellet holes) and the distance of the shot (X, in feet). The study used two shotgun shell brands, this analysis is based on the

repNum	$X = 10$	$X = 20$	$X = 30$	$X = 40$	$X = 50$
1	2.60	6.84	6.51	10.28	11.80
2	3.35	6.32	6.72	11.47	13.74
3	3.33	6.96	8.24	14.10	15.18
4	3.06	5.85	7.38	12.54	20.13
5	3.38	5.95	9.84	16.13	16.94
6	3.85	6.29	9.42	11.03	14.09
Mean	3.26	6.37	8.02	12.59	15.31
SD	0.4126	0.4527	1.3929	2.1835	2.9045
Var	0.1702	0.2049	1.9401	4.7677	8.4359

Table 1.6: Shotgun spread data

second (Remington No. 4). The data are given in Table 1.6, there were 5 distinct distances (10 to 50 by 10), with 6 replicates per distance. A plot of the data and the fitted regression line are given in Figure 1.5. The result of the model fit is given below (all R programs are given at the end of the chapter).

```
> sg.mod1 <- lm(spread.Y2 ~ dist.X2)
> summary(sg.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01267    0.75413   0.017    0.987
dist.X2      0.30327    0.02274  13.337 1.18e-13 ***

Residual standard error: 1.761 on 28 degrees of freedom
Multiple R-squared:  0.864,     Adjusted R-squared:  0.8591
F-statistic: 177.9 on 1 and 28 DF,  p-value: 1.184e-13

> anova(sg.mod1)
Analysis of Variance Table
Response: spread.Y2
          Df Sum Sq Mean Sq F value    Pr(>F)
dist.X2    1 551.82  551.82  177.89 1.184e-13 ***
Residuals 28  86.86    3.10
```

The Shapiro-Wilk test for normality yields a P-value of .0579, and the normal probability plot shows several outlying residuals based on the assumption of normality in Figure 1.6. Clearly, based on Figure 1.5, there is also non-constant variance which will be considered below.

The Box-Cox transformation was run, with the point estimate of λ being 0.263, with a Confidence Interval of (-.020,0.586), suggesting a quarter root transformation for Y. The plot of the likelihood function versus λ and the Confidence Interval are given in Figure 1.7.

Based on the quarter root transformation on Y, the P-value for the Shapiro-Wilk test is .9958, confirming that the residuals from the transformed model are approximately normal, the normal probability plot (not shown) conforms to the lack of outliers.

∇

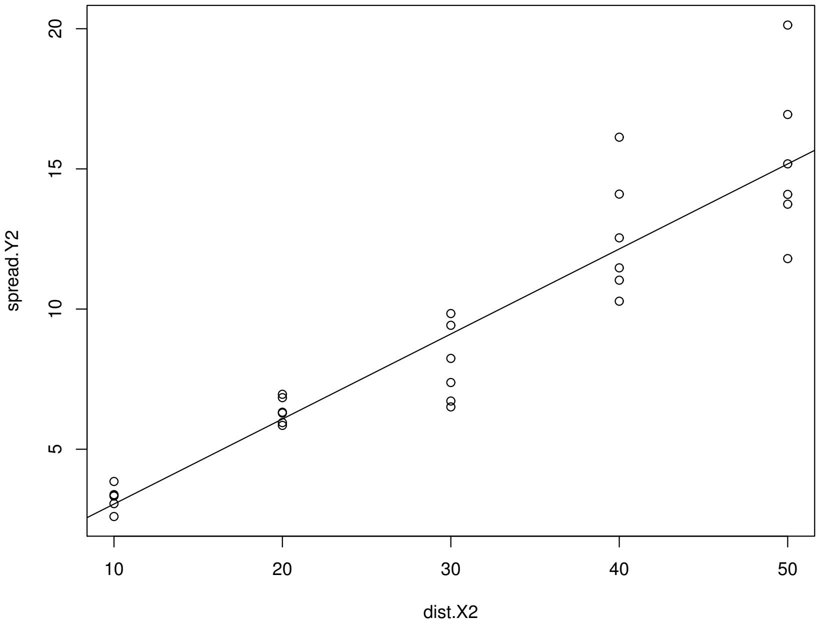

Figure 1.5: Plot of shotgun spread (Y) versus distance (X) and fitted regression equation

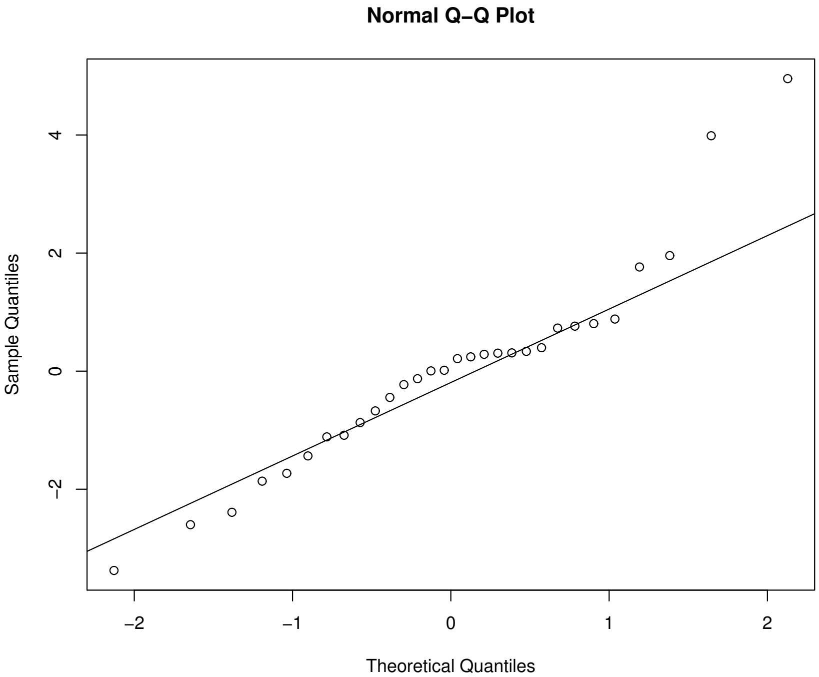

Figure 1.6: Normal probability plot of residuals - shotgun spread data

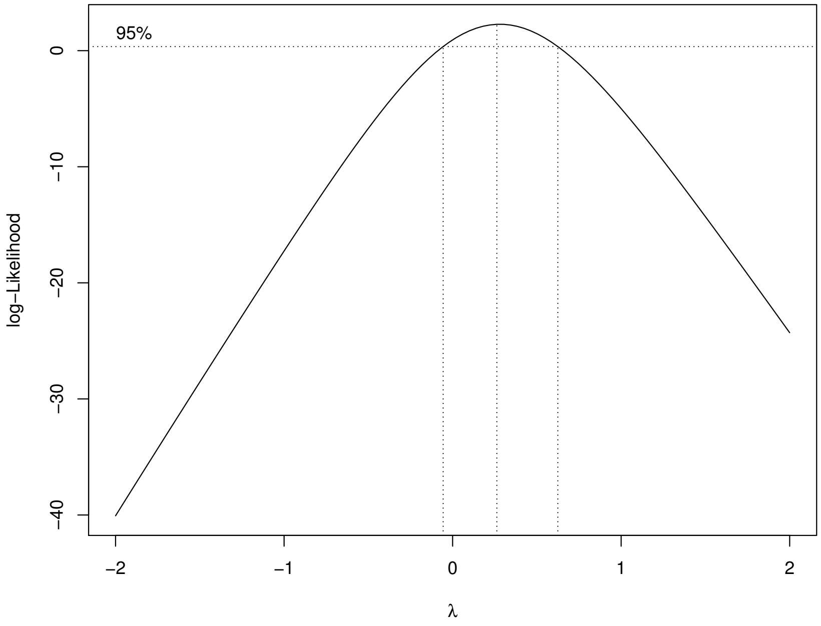

Figure 1.7: Plot of Box-Cox transformation - shotgun spread data

1.8.3 Assumption of Equal Variance

A plot of the residuals versus the fitted values $\hat{Y}$ (or X) should be a random cloud of points centered at 0. When the variances are unequal, the variance tends to increase with the mean, and a funnel-type shape is often observed.

Two tests for equal variance are the Brown-Forsyth test and the Breusch-Pagan test. When the variance is not constant, Y can possibly be transformed to obtain approximately constant variance.

Brown-Forsyth Test - Split data into two groups of approximately equal sample sizes based on their fitted values (any cases with the same fitted values should be in the same group). Then labeling the residuals $e_{11}, \ldots, e_{1n_1}$ and $e_{21}, \ldots, e_{2n_2}$, obtain the median residual for each group: $\tilde{e}_1$ and $\tilde{e}_2$, respectively. Then compute the following quantities.

$$d_{ij} = |e_{ij} - \tilde{e}_i| \quad i = 1, 2; j = 1, \ldots, n_i \qquad \overline{d}_i = \frac{\sum_{j=1}^{n_i} d_{ij}}{n_i} \qquad s_i^2 = \frac{\sum_{j=1}^{n_i} \left(d_{ij} - \overline{d}_i\right)^2}{n_i - 1} \qquad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Then, a 2-sample t-test is conducted to test H_0: Equal Variances in the 2 groups.

j	e_{1j}	$\hat{Y}_{1j}$	med(e_{1j})	d_{1j}	e_{2j}	$\hat{Y}_{2j}$	med(e_{2j})	d_{2j}
1	6.259899	78.7401	0.445271	5.814628	1.418695	147.5813	-0.81642	2.23512
2	2.759899	78.7401	0.445271	2.314628	1.418695	147.5813	-0.81642	2.23512
3	4.759899	78.7401	0.445271	4.314628	-0.58131	147.5813	-0.81642	0.23512
4	4.485025	97.51497	0.445271	4.039754	-7.59789	160.0979	-0.81642	6.781463
5	-7.01497	97.51497	0.445271	7.460246	-1.59789	160.0979	-0.81642	0.781463
6	0.985025	97.51497	0.445271	0.539754	-9.09789	160.0979	-0.81642	8.281463
7	-6.59448	115.5945	0.445271	7.039754	-6.05154	167.0515	-0.81642	5.23512
8	-0.09448	115.5945	0.445271	0.539754	2.948455	167.0515	-0.81642	3.76488
9	-7.54155	138.5416	0.445271	7.986822	-1.05154	167.0515	-0.81642	0.23512
10	-10.0416	138.5416	0.445271	10.48682	12.88553	172.6145	-0.81642	13.70195
11	-0.04155	138.5416	0.445271	0.486822	19.38553	172.6145	-0.81642	20.20195
Mean				4.63851				5.789889
SD				3.423818				6.276751

Table 1.7: Galvonomer Deflection (Y) and Wire Area (X) calculations to obtain the Brown-Forsythe test

$$TS: t_{BF} = \frac{\overline{d}_1 - \overline{d}_2}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \qquad RR: |t_{BF}| \geq t_{\alpha/2,n-2} \qquad P\text{-value} = 2P\left(t_{n-2} \geq |t_{BF}|\right)$$

Example 1.10: Galvonometer Deflection in Experiments with Explosives

The Brown-Forsyth test is conducted for the Explosives Experiment, with the data being split into 2 groups with $n_1 = n_2 = 11$ observations based on their fitted values. Calculations were obtained in EXCEL, and are given in Table 1.7.

$$\overline{d}_1 = 4.639 \quad \overline{d}_2 = 5.790 \qquad s_1 = 3.424 \quad s_2 = 6.277 \qquad s_p^2 = \frac{(11-1)3.424^2 + (11-1)6.277^2}{11+11-2} = 25.562$$

$$TS: t_{BF} = \frac{4.639 - 5.790}{\sqrt{25.562\left(\frac{1}{11}+\frac{1}{11}\right)}} = \frac{-1.151}{2.156} = -0.534 \qquad P\text{-value} = 2P\left(t_{22-2} \geq |-0.534|\right) = .5992$$

The test provides no evidence of non-constant error variance.

∇

Breusch-Pagan Test - Fits a regression of the squared residuals on X and tests whether the (natural) log of the variance is linearly related to X. When the regression of the squared residuals is fit, obtain SSR_{e^2}, the regression sum of squares. The test is conducted as follows, where SSE is the Error Sum of Squares for the original regression of Y on X.

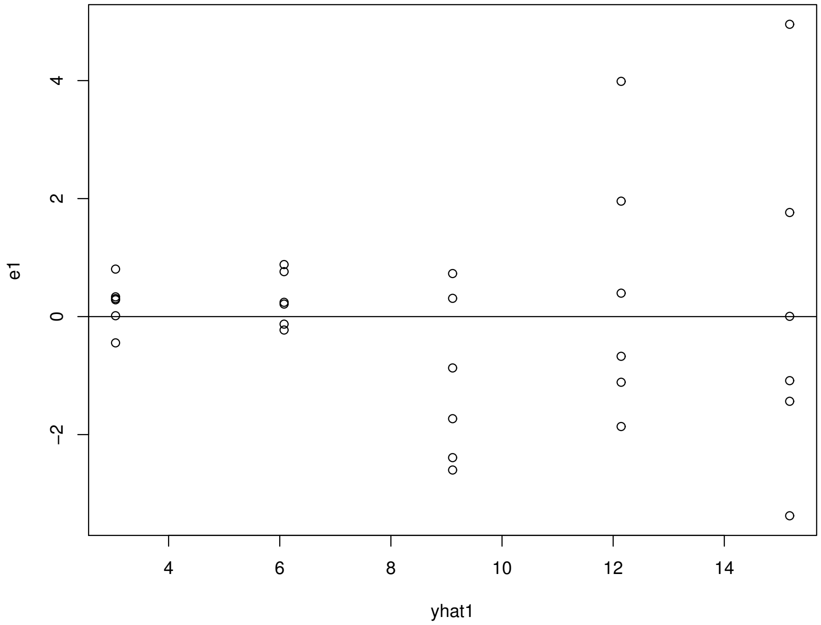

Figure 1.8: Plot of residuals versus predicted values for the shotgun spread data (original scale)

$$TS: X^2_{BP} = \frac{(SSR_{e^2}/2)}{(SSE/n)^2} \qquad RR: X^2_{BP} \geq \chi^2_{\alpha,1} \qquad P\text{-value: } P\left(\chi^2_1 \geq X^2_{BP}\right)$$

Example 1.11: Spread of Shotgun Pellets

A plot of the residuals versus the fitted values of the original regression model relating pellet spread to distance for the shotgun data is given in Figure 1.8. There is strong evidence of variance increasing with the mean. A second regression model is fit, relating the squared residuals to distance so the Breusch-Pagan test can be conducted. Recall that SSE from the original model is 86.86 and $n = 30$ from Example 1.9.

```
> sg.mod3 <- lm(I(e1^2) ~ dist.X1)
> anova(sg.mod3)
Analysis of Variance Table
Response: I(e1^2)
          Df Sum Sq Mean Sq F value   Pr(>F)
dist.X1    1 186.69 186.686  7.6855 0.009786 **
Residuals 28 680.14  24.291
```

$$SSR_{e^2} = 186.69 \qquad TS: X^2_{BP} = \frac{(186.686/2)}{(86.86/30)^2} = \frac{93.343}{8.383} = 11.135$$

$$RR: X^2_{BP} \geq \chi^2_{.05,1} = 3.841 \quad P = \left(\chi^2_1 \geq 11.135\right) = .0008$$

Making use of the **bptest** function in the **lmtest** package, the Breusch-Pagan test will be computed directly. The default is to use studentized residuals (see Section 1.8.5 for the definition). To use the current version of the test, use the studentize=FALSE option.

```
> bptest(sg.mod1, studentize=FALSE)

        Breusch-Pagan test
data:  sg.mod1
BP = 11.135, df = 1, p-value = 0.0008471
```

There is very strong evidence of non-constant variance. Note that as the Box-Cox transformation cured the non-normality of errors, it also reduces the non-constant variance (the P-value for the Breusch-Pagan test is .1713, not shown here). This "combined effect" of the Box-Cox transformation will not work for all data sets.

∇

Another possibility is to use **Estimated Weighted Least Squares** by relating the standard deviation (or a power of it) of the errors to the mean. This is an iterative process, where the weights are re-weighted at each iteration. The weights are the reciprocal of the estimated variance (or possibly a power of it)as a function of the mean. Iteration continues until the regression coefficient estimates stabilize.

When data have replicates at the various X levels, the reciprocal of the estimated variance (or possibly a power of it) of the observations at the various X levels can be used as weights. This gives a higher weight on the observations with the smaller variation. Most statistical software packages have options for weighting the data (this can be done directly using the matrix form of the regression model).

Example 1.12: Spread of Shotgun Pellets

From Table 1.6, the sample variances for the five distances range from 0.1702 when $X = 10$ to 8.4359 when $X = 50$. The estimated weighted least squares estimates are given below, where the weights of the observations are the reciprocal of the variance of the measurements at the corresponding X level. For observations at $X = 10$, the weight is 1/.1702=5.8754, at $X = 50$, the weight is 1/8.4359=0.1185.

```
> sg.mod4 <- lm(spread.Y2 ~ dist.X2, weight=reg.wt)
> summary(sg.mod4)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3821     0.2940   1.299    0.204
dist.X2       0.2923     0.0166  17.604   <2e-16
```

The slope $\left(\hat{\beta}_1\right)$ coefficients are very similar (0.3033 for ordinary least squares versus 0.2923 for weighted least squares), but the standard error is much smaller for weighted (0.0166) than for ordinary least squares (0.0227). This implies more precise estimates, with larger t-statistic and narrower Confidence Intervals when using weighted least squares.

∇

The variance of the individual measurements may vary as a function of a covariate or the mean. Note that for the shotgun pellet data, the variance clearly increases with the distance (as well as the mean). For instance, the variance may be related to the mean by a power function.

$$V\{\epsilon_i\} = \sigma_i^2 = \sigma^2\mu_i^{2\delta} \text{ or possibly in terms of a predictor variable } X\text{: } \sigma_i^2 = \sigma^2 X_i^{2\delta}$$

For this model, if $\delta = 0$, then the error variance is constant, otherwise the variance is said to be heteroskedastic. Estimated Generalized Least Squares (EGLS) can be used to estimate the parameters σ and δ (as well as estimated regression coefficients) with various statistical software packages.

Example 1.13: Spread of Shotgun Pellets

Using the **gls** function in the **nlme** R package, the variance power model is fit with the following estimates being obtained for the shotgun spread data. The computer output is given below.

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 0.3748 + 0.2889X_i \qquad \hat{\sigma}_i^2 = \hat{\sigma}^2\hat{\mu}_i^{2\hat{\delta}} = \left(0.0651\hat{\mu}_i^{1.3882}\right)^2$$

```
> sg.mod5 <- gls(spread.Y2 ~ dist.X2, weights = varPower(form = ~ fitted(.)), method="ML")
> summary(sg.mod5)
Generalized least squares fit by maximum likelihood
  Model: spread.Y2 ~ dist.X2
  Data: NULL
      AIC      BIC    logLik
  102.123 107.7277 -47.06148
Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
  power
1.38818
Coefficients:
                Value  Std.Error   t-value p-value
(Intercept) 0.3747855 0.25378056  1.476809  0.1509
dist.X2     0.2888985 0.01644691 17.565515  0.0000
Residual standard error: 0.06514301
Degrees of freedom: 30 total; 28 residual

> intervals(sg.mod5)
Approximate 95\% confidence intervals
 Coefficients:
                 lower      est.     upper
(Intercept) -0.1450604 0.3747855 0.8946314
dist.X2      0.2552085 0.2888985 0.3225884
 Variance function:
         lower    est.    upper
power 0.938246 1.38818 1.838114
 Residual standard error:
     lower       est.      upper
0.02475866 0.06514301 0.17139905
```

$$\nabla$$

When the distribution of Y is a from a known family of a certain type of probability distributions (e.g. Binomial, Poisson, Gamma), a **Generalized Linear Model** can be fit, which is covered in a subsequent chapter.

1.8.4 Assumption of Independence

When the data are a time (or spatial) series, the errors can be correlated over time (or space), referred to as being **autocorrelated**. A plot of residuals versus time should be random, not displaying a trending pattern (linear or cyclical). If it does show these patterns, autocorrelation may be present. If autocorrelation is present, the ordinary least squares estimators are still unbiased, however their estimated standard errors tend to be small, so that t-statistics tend to be too large, and Confidence Intervals tend to be too narrow.

The Durbin-Watson test is used to test for serial autocorrelation in the errors, where the null hypothesis is that the errors are uncorrelated. Unfortunately, the formal test can end in one of 3 possible outcomes: reject H_0, accept H_0, or inconclusive. Statistical software packages can report an approximate P-value. The test is obtained as follows.

$$TS: DW = \frac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2} \quad \text{Decision Rule: } DW < d_L \text{Reject } H_0 \quad DW > d_U \text{Accept } H_0 \quad \text{ow Inconclusive}$$

where tables of d_L and d_U are in standard regression texts and posted on the internet. These values are indexed by the number of predictor variables (1, in the case of simple regression) and the sample size (n).

A commonly used approach when autocorrelation is present is to use **Estimated Generalized Least Squares (EGLS)**. This uses the estimated covariance structure of the observations to obtain estimates of the regression coefficients and their estimated standard errors. There are many possible covariance structures for autocorrelated residuals. One particularly useful model is the autoregressive model of order 1 (AR(1)). For this model, the following results are obtained (assuming the errors are labeled $\epsilon_1, \ldots, \epsilon_n$, and that $u_2, \ldots, u_n$ are independent of one another).

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \quad t = 2, \ldots, n \quad |\rho| < 1 \quad E\{u_t\} = 0 \quad V\{u_t\} = \sigma^2$$

This leads to the following variance-covariance structure for the errors with σ^2 and ρ as parameters to be estimated, along with the EGLS estimates of the regression coefficients.

$$V\{\epsilon_t\} = \frac{\sigma^2}{1-\rho^2} \qquad \text{COV}\{\epsilon_t, \epsilon_{t-k}\} = \frac{\sigma^2\rho^{|k|}}{1-\rho^2} \quad k = \pm 1, \pm 2, \ldots$$

If $\rho = 0$, the errors are independent (assuming an AR(1) model).

Example 1.14: Annual Mean Temperatures in Minneapolis/St. Paul, Minnesota: 1900-2015

A plot of the annual mean temperature (Y, in degrees Fahrenheit) versus (X, Year-1900) for the years 1900-2015 is given in Figure 1.9, along with the ordinary least squares regression line, as well as a **loess** line which fits a smooth curve making use of "local" observations along the X-axis. The model fit is given below. There is evidence of a positive trend, but there is a large amount of variation around the mean. The fitted values, residuals, and formulas for obtaining the Durbin-Watson statistic are given below. The residuals plotted versus time order are displayed in Figure 1.10.

$$\hat{Y}_t = 44.3732 + 0.0160t \quad t = 0, \ldots., 115 = n-1 \qquad e_t = Y_t - \hat{Y}_t \qquad \frac{\sum_{t=1}^{n-1}\left(e_t - e_{t-1}\right)^2}{\sum_{t=0}^{n-1} e_t^2} = \frac{573.98}{364.40} = 1.575$$

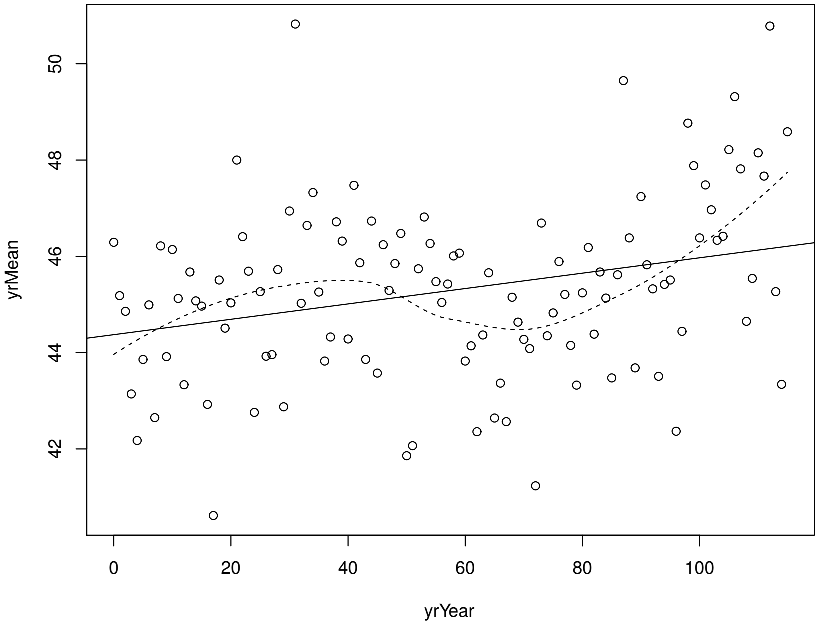

Figure 1.9: Annual mean temperature versus (Year-1900) for Minneapolis/St. Paul for years 1900-2015, fitted ordinary least squares regression equation (solid line) and smooth loess curve (dashed line)

For $\alpha = 0.05$, $n = 100$, $p = 1$ predictor variable, the critical values are $d_L = 1.65$ and $d_U = 1.69$, thus there is evidence of positive autocorrelation in the residuals. The P-value reported by the **durbinWatsonTest** function in the R package **car** is 0.016 (output given below). The autocorrelation parameter ρ is estimated as .1996. The loess shows that the observations tend to be below the fitted linear regression in early years, then above for about years 1910-1950, followed by below for about years 1950-1995, then above for 1995-2015. This is helpful in visualizing the autocorrelation in the residuals.

The generalized least squares fit for the model with an AR(1) error structure is given below. the following parameter estimates are obtained (note that it uses a different method of estimating ρ as the previous method used).

$$\hat{\rho} = 0.2247 \qquad \hat{\sigma} = 1.7972 \qquad \hat{\beta}_0 = 44.3805 \qquad \hat{\beta}_1 = 0.0160 \qquad \hat{V}\{\epsilon_t\} = \frac{1.7972^2}{1 - 0.2247^2} = \frac{3.2299}{0.9495} = 3.4016$$

∇

```
### Ordinary Least Squares Model Fit
> msw.mod1 <- lm(yrMean ~ yrYear)
> summary(msw.mod1)
Coefficients:
```

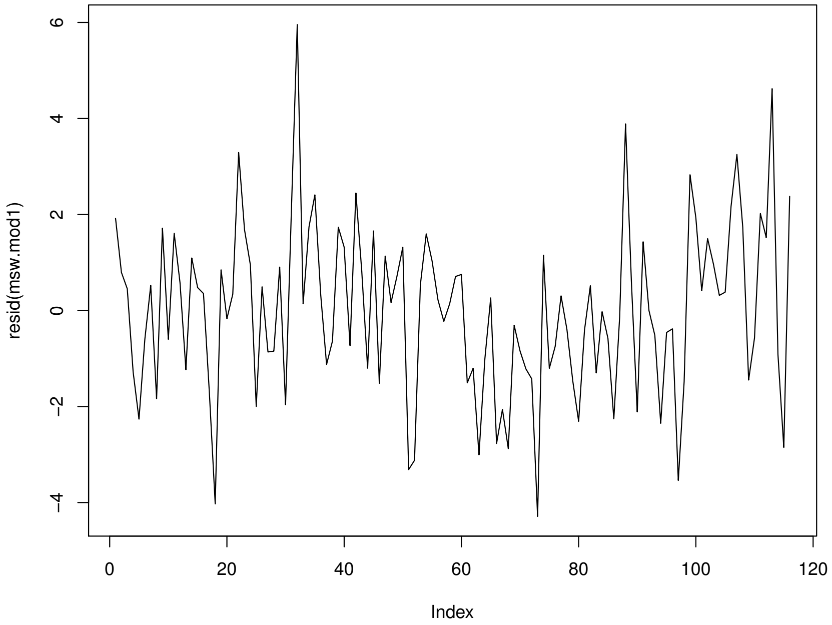

Figure 1.10: Line plot of residuals versus time order for Minneapolis/St. Paul annual mean temperature model

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 44.373176   0.329864 134.520  < 2e-16 ***
yrYear       0.015967   0.004957   3.221  0.00167 **

Residual standard error: 1.788 on 114 degrees of freedom
Multiple R-squared:  0.08341,   Adjusted R-squared:  0.07537
F-statistic: 10.37 on 1 and 114 DF,  p-value: 0.001666
> anova(msw.mod1)
Analysis of Variance Table
Response: yrMean
           Df Sum Sq Mean Sq F value   Pr(>F)
yrYear      1  33.16  33.159  10.374 0.001666 **
Residuals 114 364.40   3.196

### Direct computation of Durbin-Watson statistic
> cbind(DW1, DW2, DW)
       DW1      DW2       DW
2 573.9759 364.3984 1.575133

###  durbinWatsonTest function applied to OLS model fit
> library(car)
> durbinWatsonTest(msw.mod1)
 lag Autocorrelation D-W Statistic p-value
   1       0.1996232      1.575133   0.016
 Alternative hypothesis: rho != 0

### Generalized Least Squares Model Fit w/ AR(1) errors
> msw.mod3 <- gls(yrMean ~ yrYear, correlation=corAR1(), method="ML")
> summary(msw.mod3)
Generalized least squares fit by maximum likelihood
  Model: yrMean ~ yrYear
  Data: NULL
       AIC      BIC    logLik
  465.1704 476.1848 -228.5852
Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.2031676
Coefficients:
               Value Std.Error   t-value p-value
(Intercept) 44.37957 0.4032028 110.06763  0.0000
yrYear       0.01602 0.0060529   2.64654  0.0093

Residual standard error: 1.772726
Degrees of freedom: 116 total; 114 residual
> intervals(msw.mod3)
Approximate 95\% confidence intervals
Coefficients:
                   lower        est.       upper
(Intercept) 43.580833187 44.37957476 45.17831633
yrYear       0.004028463  0.01601912  0.02800978
 Correlation structure:
        lower      est.     upper
Phi 0.0187394 0.2031676 0.3742249
 Residual standard error:
   lower     est.    upper
1.550264 1.772726 2.027112
```

1.8.5 Detecting Outliers and Influential Observations

These measures are widely used in multiple regression, as well, when there are p predictors, and $p' = p + 1$ parameters (including intercept, β_0). Many of the "rules of thumb" are based on p', which is 1+1=2 for simple regression. Most of these methods involve matrix algebra, but are obtained from statistical software packages. Their matrix forms are not given here (see references).

Also, many of these methods make use of the estimated variance when the i^{th} case was removed (to remove its effect if it is an outlier):

$$MSE_{(i)} = \frac{SSE_{(i)}}{n - p' - 1} = \frac{SSE - e_i^2}{n - p' - 1} \qquad \text{for simple regression } p' = 2$$

Studentized Residuals - Residuals divided by their estimated standard error, with their contribution to SSE having been removed (see above). Since residuals have mean 0, the studentized residuals are like t-statistics. Since we are simultaneously checking whether n of these are outliers, we conclude any cases are outliers if the absolute value of their studentized residuals exceed $t_{\alpha/2n,n-p'-1}$, where p' is the number of independent variables plus one (for simple regression, p'=2).

Leverage Values (Hat Values) - These measure each case's potential to influence the regression due to its X levels. Cases with high leverage values (often denoted v_{ii} or h_{ii}) have X levels "away" from the center of the distribution. The leverage values sum to p' (2 for simple regression), and cases with leverage values greater than $2p'/n$ (twice the average) are considered to be potentially influential due to their X-levels. Note that R flags cases with leverage values greater than $3p'/n$ as being potentially influential.

DFFITS - These measure how much an individual case's fitted value shifts when it is included in the regression fit, and when it is excluded. The shift is divided by its standard error, so we are measuring how many standard errors a fitted value shifts, due to its being included in the regression model. Cases with the DFFITS values greater than $2\sqrt{p'/n}$ in absolute value are considered influential on their own fitted values. Note that R flags cases with DFFITS greater than $3\sqrt{p'/n}$ in absolute value.

DFBETAS - One of these is computed for each case, for each regression coefficient (including the intercept). DFBETAS measures how much the estimated regression coefficient shifts when that case is included and excluded from the model, in units of standard errors. Cases with DFBETAS values larger than $2/\sqrt{n}$ in absolute value are considered to be influential on the estimated regression coefficient. Note that R flags cases with DFBETAS greater than 1 in absolute value.

Cook's D - A measure that represents each case's aggregate influence on all regression coefficients, and all cases' fitted values. Cases with Cook's D larger than $F_{.50,p',n-p'}$ are considered influential.

COVRATIO - This measures each case's influence on the estimated standard errors of the regression coefficients (inflating or deflating them). Cases with COVRATIO outside of $1 \pm 3p'/n$ are considered influential.

Example 1.15: Spring Migration of trans-Saharan Bird Species

A biology study considered the spring migrations of $n = 38$ bird species in trans-Saharan Africa (Rubolini, Spina, and Saino, 2005, [31]). The dependent variable was median migration date (Y=Days from April 1) and the independent variable was mean wintering latitude (X = degrees latitude). Note that these are obtained across many birds within the species. The data are given in Table 1.8 and a plot of the data and fitted equation are given in Figure 1.11. Note that this example should probably use weighted least squares with weights being the number of birds per species, but will use Ordinary Least Squares for this example.

The fitted equation is $\hat{Y} = 29.9918 - 0.2467X$ with $r^2 = .3119$ (R output given below). The "rules of thumb" for influential cases are computed below, based on $n = 38$ and $p' = 1 + 1 = 2$.

Studentized Residuals: $t_{.05/2(38),38-2-1} = t_{.00066,35} = 3.4925$ Hat: $\frac{2(2)}{38} = 0.1053 \quad \frac{3(2)}{38} = 0.1579$

DFFITS: $2\sqrt{\frac{2}{38}} = 0.4588 \quad 3\sqrt{\frac{2}{38}} = 0.6882$ DFBETAS: $\frac{2}{\sqrt{38}} = 0.3244$

Cook's D: $F_{.50,2,38-2} = F_{.50,2,36} = 0.7067$ COVRATIO: $1 \pm \frac{3(2)}{38} \equiv 1 \pm 0.1579 \equiv (0.8421, 1.1579)$

The results of the studentized residuals and influence measures are given in Table 1.9. Based on the original criteria, species' 3 for $\hat{\beta}_0$, 10, 13, and 18 for $\hat{\beta}_1$, 14 and 8 for Hat have values above the "critical value," all being very close. No species exceed any of the R criteria (there would be asterisks in the influence column). Either way, there is no evidence of any highly influential cases.

∇

1.9 Repeated Sampling From a Population of (X, Y) Pairs when X is Random

So far, we have treated the X variable as fixed, which particularly holds for the Explosives and Shot Gun pellet experiments, not so much for the bird migrations example. As long as the X and ϵ are "generated" independently, there is no problem, and the estimated standard errors of $\hat{\beta}_1$ and $\hat{\beta}_0$ are said to be conditional on the observed X values in the sample. Coverage rates for Confidence Intervals should be as expected if the error terms are independent and normally distributed with constant variance.

Example 1.16: Predicted Over/Under and Total Points in NBA Games

Oddsmakers provide an Over/Under score for the combined total points in many sporting events. Gamblers can choose to bet that the combined score will be above (over) or below (under) the posted score. In this example, we consider the $N = 1154$ games from the 2014-2015 NBA regular season that finished in regulation (48 minutes), removing games that went into overtime. The oddsmakers' final Over/Under score (as reported by covers.com) is treated as X, as it is posted in advance (although there is randomness in its chosen level). The response is the total points in the game, and is Y. For this "population" of games, we have the following model and "parameters." Note that due to estimation issues, centered X values are used with $x = X - \mu_X$, and both X and Y are measured in 100s (a typical combined score is 200 points).

speciesID	species1	species2	migDate (Y)	latBreed	latWntr (X)
1	Acrocephalus	arundinaceus	33	46	-10.3
2	Acrocephalus	schoenobaenus	35	57.5	-7.5
3	Acrocephalus	scirpaceus	38	48	0
4	Anthus	campestris	32	43.5	6
5	Anthus	trivialis	27	55.3	-10
6	Calandrella	brachydactyla	27.5	39.5	15.5
7	Caprimulgus	europaeus	35	47.5	-7.5
8	Coturnix	coturnix	30	50.3	18.5
9	Cuculus	canorus	31	51	-15
10	Delichon	urbica	29	48.5	-15
11	Emberiza	hortulana	30.8	51.5	7.5
12	Ficedula	albicollis	30	48.8	-10
13	Ficedula	hypoleuca	28	59	7.5
14	Hippolais	icterina	39	56	-19
15	Hirundo	rustica	27	49	-10
16	Jynx	torquilla	23	49	9
17	Lanius	senator	30	39	9.5
18	Locustella	naevia	31.2	54.5	13
19	Luscinia	megarhynchos	25	42.5	7.5
20	Merops	apiaster	33	42.5	-2.5
21	Monticola	saxatilis	25.7	40.5	2.5
22	Motacilla	flava	28	49	-7.5
23	Muscicapa	striata	36	49	-13
24	Oenanthe	hispanica	23	37	13
25	Oenanthe	oenanthe	26	49	4.5
26	Oriolus	oriolus	35	45.5	-12
27	Otus	scops	25	44	6
28	Phoenicurus	phoenicurus	29	49	7.5
29	Phylloscopus	sibilatrix	31	53.8	1.5
30	Phylloscopus	trochilus	27	56.5	-9
31	Riparia	riparia	34	52.5	-2.5
32	Saxicola	rubetra	31	56	0
33	Streptopelia	turtur	31	45.5	12.5
34	Sylvia	atricapilla	22	48	11
35	Sylvia	borin	38	52.5	-10
36	Sylvia	cantillans	23	37	13
37	Sylvia	communis	34	48	-2.5
38	Upupa	epops	22	44	16

Table 1.8: Median migration date (Y) and mean wintering latitude (X) for n=38 trans-Saharan bird species

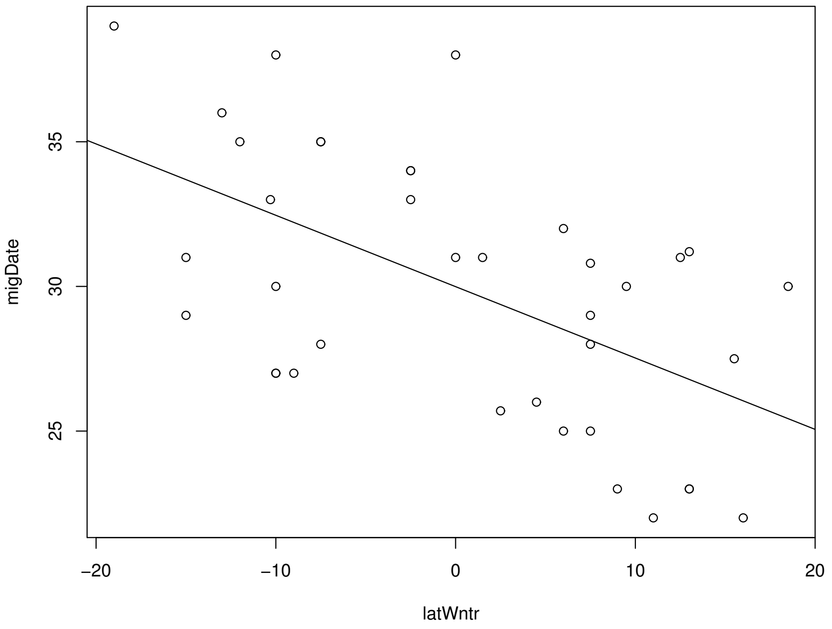

Figure 1.11: Plot of Median migration date (Y) and mean wintering latitude (X) for n=38 trans-Saharan bird species and fitted OLS regression equation

speciesID	Std. Resid	dfb.1	dfb.ltWn	dffit	cov.r	cook.d	hat	inf
1	0.1227	0.0215	-0.0215	0.0297	1.1190	0.0005	0.0553	
2	0.8310	0.1426	-0.1070	0.1744	1.0620	0.0153	0.0422	
3	2.2091	0.3636	-0.0169	0.3636	0.8370	0.0597	0.0264	
4	0.9157	0.1472	0.0812	0.1716	1.0440	0.0148	0.0339	
5	-1.4746	-0.2573	0.2509	-0.3513	0.9910	0.0597	0.0537	
6	0.3551	0.0560	0.0879	0.1065	1.1450	0.0058	0.0826	
7	0.8310	0.1426	-0.1070	0.1744	1.0620	0.0153	0.0422	
8	1.2615	0.1987	0.3800	0.4374	1.0840	0.0941	0.1073	
9	-0.7233	-0.1312	0.1850	-0.2220	1.1240	0.0250	0.0861	
10	-1.2804	-0.2323	0.3274	-0.3929	1.0560	0.0759	0.0861	
11	0.6961	0.1114	0.0787	0.1395	1.0710	0.0099	0.0386	
12	-0.6483	-0.1131	0.1103	-0.1544	1.0920	0.0121	0.0537	
13	-0.0369	-0.0059	-0.0042	-0.0074	1.1000	0.0000	0.0386	
14	1.1993	0.2256	-0.3935	0.4448	1.1100	0.0978	0.1209	
15	-1.4746	-0.2573	0.2509	-0.3513	0.9910	0.0597	0.0537	
16	-1.2733	-0.2029	-0.1753	-0.2745	1.0110	0.0370	0.0444	
17	0.6174	0.0983	0.0901	0.1365	1.0860	0.0095	0.0466	
18	1.1877	0.1878	0.2429	0.3142	1.0460	0.0488	0.0654	
19	-0.8250	-0.1320	-0.0933	-0.1653	1.0590	0.0138	0.0386	
20	0.6221	0.1037	-0.0297	0.1066	1.0650	0.0058	0.0285	
21	-0.9627	-0.1567	-0.0312	-0.1614	1.0320	0.0131	0.0273	
22	-1.0156	-0.1743	0.1308	-0.2132	1.0420	0.0227	0.0422	
23	0.7473	0.1334	-0.1651	0.2076	1.1040	0.0218	0.0716	
24	-1.0130	-0.1601	-0.2072	-0.2680	1.0680	0.0359	0.0654	
25	-0.7523	-0.1215	-0.0485	-0.1331	1.0570	0.0090	0.0303	
26	0.5425	0.0961	-0.1106	0.1432	1.1130	0.0105	0.0652	
27	-0.9221	-0.1482	-0.0818	-0.1728	1.0440	0.0150	0.0339	
28	0.2233	0.0357	0.0253	0.0448	1.0970	0.0010	0.0386	
29	0.3568	0.0583	0.0058	0.0590	1.0790	0.0018	0.0266	
30	-1.4002	-0.2426	0.2149	-0.3169	0.9970	0.0489	0.0487	
31	0.8872	0.1478	-0.0424	0.1520	1.0420	0.0116	0.0285	
32	0.2608	0.0429	-0.0020	0.0429	1.0820	0.0009	0.0264	
33	1.0957	0.1734	0.2148	0.2826	1.0550	0.0397	0.0624	
34	-1.4233	-0.2258	-0.2431	-0.3398	0.9990	0.0561	0.0539	
35	1.4985	0.2615	-0.2549	0.3570	0.9870	0.0616	0.0537	
36	-1.0130	-0.1601	-0.2072	-0.2680	1.0680	0.0359	0.0654	
37	0.8872	0.1478	-0.0424	0.1520	1.0420	0.0116	0.0285	
38	-1.0976	-0.1730	-0.2815	-0.3375	1.0820	0.0566	0.0864	
Rule 1	3.4925	0.3244	0.3244	0.4588	.842-1.158	0.7067	0.1053	
R Rule	3.4925	1.0000	1.0000	0.6882	.842-1.159	0.7067	0.1579	

Table 1.9: Studentized Residuals and influence measures for n=38 trans-Saharan bird species

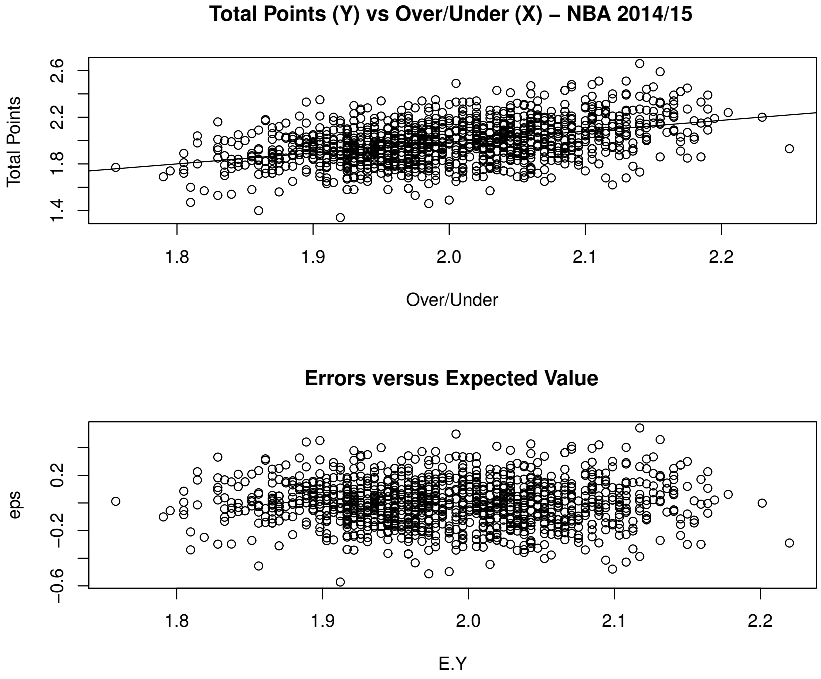

Figure 1.12: Plots of Total Points versus Over/Under and Errors versus Expected Values - NBA Oddsmaker population data

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad \epsilon \sim NID\left(0, \sigma^2\right) \qquad E\{X\} = \mu_X \qquad V\{X\} = \sigma_X^2 \qquad \text{COV}\{X_i, \epsilon_i\} = 0$$

$$\beta_0 = 1.9871 \qquad \beta_1 = 0.9328 \qquad \sigma^2 = 0.0263 \qquad \sigma = 0.1622 \qquad \mu_X = 2.0004 \qquad \sigma_X = 0.838$$

Total points (Y) plotted versus Over/Under (X) as well as a plot of errors versus expected values are given in Figure 1.12. Histograms of the un-centered X values and the error terms are given in Figure 1.13. The model assumptions seem reasonable.

100000 random samples, each with 25 games were obtained, and for each sample $\hat{\beta}_0$, $\hat{\beta}_1$, s^2, and $\hat{SE}\left\{\hat{\beta}_1\right\}$ were obtained. Further 95% Confidence Intervals were obtained to observe the overall coverage rate. Among the 100000 95% Confidence Intervals for β_1, 94.85% contained the true value. The empirical sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in Figure 1.14, both are mound shaped and centered at the true parameter values (lines at center of plots).

∇

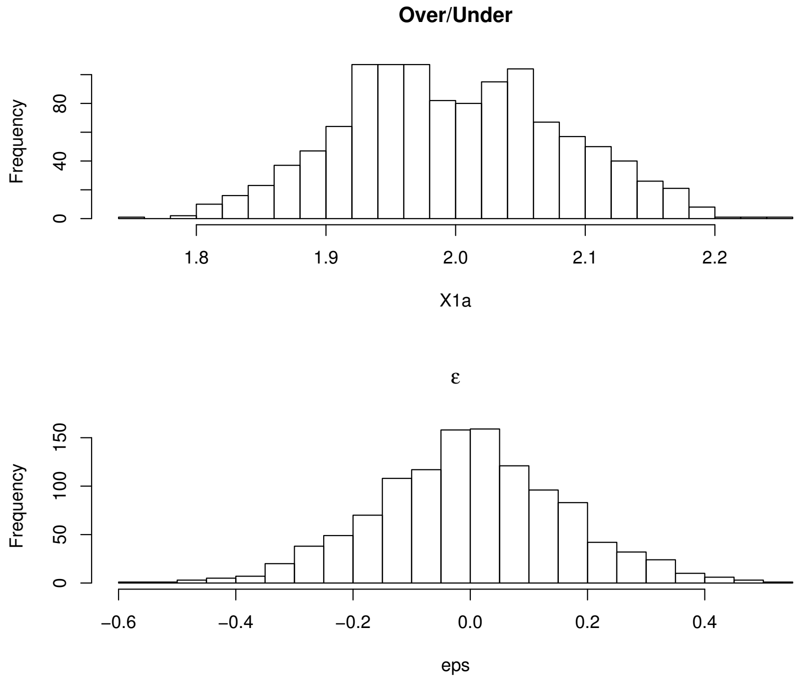

Figure 1.13: Histograms of Over/Under and Errors - NBA Oddsmaker population data

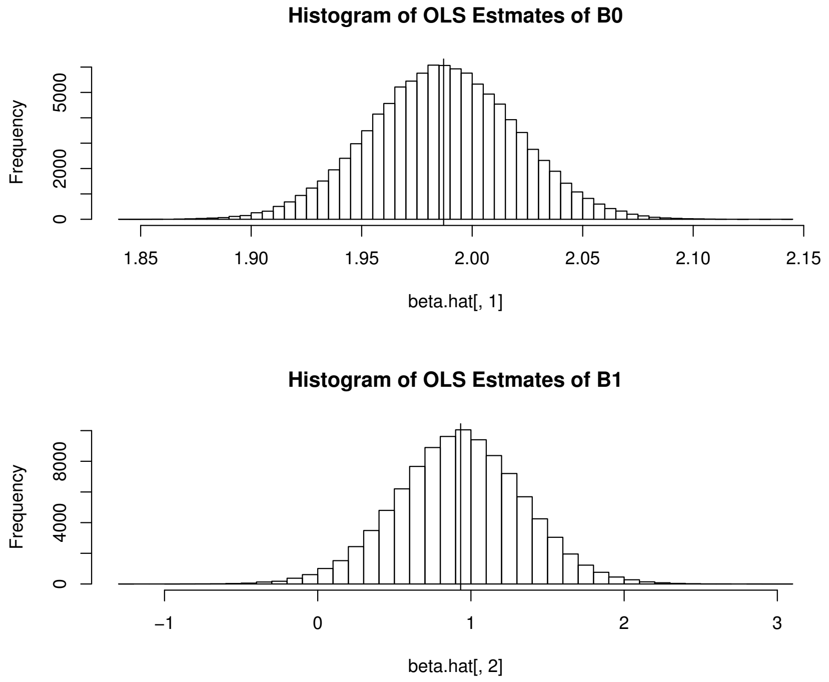

Figure 1.14: Empirical Sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$ - NBA Oddsmaker population data

1.10 R Programs for Chapter 1 Examples

1.10.1 Explosives Experiment

```
## Read in explosives data
explosives <- read.table("http://www.stat.ufl.edu/~winner/data/explosives1.dat",
       header=F, col.names=c("coupling", "wireArea", "galvDef"))
attach(explosives)

##### Example 1.1
explosives                         ## prints data frame
plot(galvDef ~ wireArea)           ## Plot Y ~ X
abline(lm(galvDef ~ wireArea))     ## Add Fitted Equation


##### Example 1.2
explo.mod1 <- lm(galvDef ~ wireArea)    ## Fit model using lm(Y ~ X) function
summary(explo.mod1)                     ## Print results of regression
Yhat.mod1 <- predict(explo.mod1)        ## Predicted values with predict(model) function
e.mod1 <- resid(explo.mod1)             ## Residuals with resid(model) function
(df_E.mod1 <- df.residual(explo.mod1))  ## Error df with df.residual(model) function
(SSE.mod1 <- deviance(explo.mod1))      ## Error SS with deviance(model) function
(MSE.mod1 <- sigma(explo.mod1)^2)       ## Obtain (s)^2 = MSE
(b0.mod1 <- coef(explo.mod1)[1])        ## Obtain b0 with coef(model) function
(b1.mod1 <- coef(explo.mod1)[2])        ##    "   b1   "    "    "       "

## Direct computations
(n <- length(galvDef))                  ## Obtain sample size with length(Y) function
(Xbar <- mean(wireArea))                ## Obtain X-bar
(Ybar <- mean(galvDef))                 ## Obtain Y-bar
(SS_XX <- sum((wireArea - mean(wireArea))^2))     ## Obtain SS_XX
(SS_YY <- sum((galvDef - mean(galvDef))^2))       ## Obtain SS_YY
(SS_XY <- sum((wireArea - mean(wireArea)) *
            (galvDef - mean(galvDef))))           ## Obtain SS_XY
(beta.hat1 <- SS_XY / SS_XX )                     ## Obtain beta.hat1
(beta.hat0 <- Ybar - beta.hat1 * Xbar)            ## Obtain beta.hat0
Y.hat <- beta.hat0 + beta.hat1 * wireArea         ## Obtain Y.hat
e <- galvDef - Y.hat                              ## Obtain e (residuals)
(df_E <- n-2)                                     ## Obtain Error df
(SSE <- sum(e^2))                                 ## Obtain SSE
(MSE <- SSE / df_E)                               ## Obtain s^2 = MSE
(se.b0 <- sqrt(MSE*(1/n + Xbar^2/SS_XX)))         ## Obtain SE{beta.hat0}
(se.b1 <- sqrt(MSE/SS_XX))                        ## Obtain SE{beta.hat1}


##### Example 1.3
summary(explo.mod1)                     ## Print results of regression (t-tests)
confint(explo.mod1)                     ## Confidence Intervals for beta0, beta1

## Direct Calculations
t_025 <- qt(.975,df_E)    ## Obtain t_(.025,df_E)
t.b0 <- beta.hat0 / se.b0;   p.t.b0 <- 2*(1-pt(abs(t.b0),df_E))          ## t-test for beta0
t.b1 <- beta.hat1 / se.b1;   p.t.b1 <- 2*(1-pt(abs(t.b1),df_E))          ## t-test for beta1
b0.lo <- beta.hat0 - t_025*se.b0; b0.hi <-  beta.hat0 + t_025*se.b0     ## 95%CI for beta0
b1.lo <- beta.hat1 - t_025*se.b1; b1.hi <-  beta.hat1 + t_025*se.b1     ## 95%CI for beta1

## Set-up for printing
## rbind "stacks" Variables on top of each other; cbind places them side-by-side
beta.hat <- rbind(beta.hat0, beta.hat1)
se.b <- rbind(se.b0, se.b1)
```

```
t.b <- rbind(t.b0, t.b1)
p.t.b <- rbind(p.t.b0, p.t.b1)
b.lo <- rbind(b0.lo, b1.lo)
b.hi <- rbind(b0.hi, b1.hi)
beta.est <- cbind(beta.hat, se.b, t.b, p.t.b, b.lo, b.hi)
colnames(beta.est) <- c("Estimate", "Std. Error", "t", "Pr(>|t|)","LL","UL")
rownames(beta.est) <- c("Intercept","wireArea")
round(beta.est,4)


##### Example 1.4
Xstar <- 110                                      ## Assign X* = 110
predict(explo.mod1, list(wireArea=Xstar))         ## Obtain predicted value
predict(explo.mod1, list(wireArea=Xstar),int="c") ## Obtain CI for mean
predict(explo.mod1, list(wireArea=Xstar),int="p") ## Obtain PI for individual

## Direct Calculations
Ystar <- beta.hat0 + beta.hat1 * Xstar                          ## Yhat.star
SE_Ystar <- sqrt(MSE * (1/n + (Xstar - Xbar)^2/SS_XX))          ## SE{Mean}
SE_Ystar_New <- sqrt(MSE * (1 + 1/n + (Xstar - Xbar)^2/SS_XX))  ## SE{Individual}
t_025 <- qt(.975,df_E)                                          ## t_{.025, df_E}
CI_LB <- Ystar - t_025 * SE_Ystar; CI_UB <- Ystar + t_025 * SE_Ystar           ## CI for mean
PI_LB <- Ystar - t_025 * SE_Ystar_New; PI_UB <- Ystar + t_025 * SE_Ystar_New   ## PI for individual

## Print Results"
ci.pi.out <- cbind(Xstar, Ystar, SE_Ystar, SE_Ystar_New, CI_LB, CI_UB, PI_LB, PI_UB)
colnames(ci.pi.out) <-
      cbind("X*", "Y*", "SE{Mean}", "SE{Indiv}", "CI LL", "CI UL", "PI LL", "PI UL")
round(ci.pi.out,4)

## Plot Data, fitted equation, Pointwise CI and PI

Xstar1 <- seq(15, 155, 0.10)                                ## Range of X* values
Ystar1 <- predict(explo.mod1, list(wireArea=Xstar1))        ## Obtain predicted values
Ystar_CI <- predict(explo.mod1, list(wireArea=Xstar1), int="c")   ## Obtain CI's
Ystar_PI <- predict(explo.mod1, list(wireArea=Xstar1), int="p")   ## Obtain PI's

plot(galvDef ~ wireArea, xlim=c(15,155), ylim=c(50,250))   ## Plot raw data and set plot ranges
lines(Xstar1, Ystar1)                            ## Plot fitted equation
lines(Xstar1, Ystar_CI[,2], lty=2)               ## Plot CI Lower Bound (Column 2 of Ystar_CI)
lines(Xstar1, Ystar_CI[,3], lty=2)               ## Plot CI Upper Bound (Column 3 of Ystar_CI)
lines(Xstar1, Ystar_PI[,2], lty=3)               ## Plot PI Lower Bound (Column 2 of Ystar_PI)
lines(Xstar1, Ystar_PI[,3], lty=3)               ## Plot PI Upper Bound (Column 3 of Ystar_PI)

##### Example 1.5

anova(explo.mod1)      ## Analysis of Variance and F-test based on anova(model)

## Direct Calculations

TSS <- SS_YY
SSR <- SS_YY - SSE;   df_R <- 1;   MSR <- SSR / df_R   ## Regression  SS, df, MS
F_obs <- MSR / MSE;  p.F <- 1 - pf(F_obs, df_R, df_E)  ## F-test
round(cbind(MSR, MSE, F_obs, p.F), 4)
(r2 <- SSR / TSS)                                      ## R-square

##### Example 1.6

cor.test(galvDef, wireArea)    ## Correlation t-test and CI based on cor.test(X, Y) function

##### Example 1.7

## Fit 1-Way ANOVA model with X as factor variable
explo.mod2 <- lm(galvDef ~ factor(wireArea))
```

```
anova(explo.mod2)              ## Error Sum of Squares is SSPE  Error df is df_PE
anova(explo.mod1, explo.mod2)  ## Compares Regression and 1-Way ANOVA - difference is F_LF

## Direct Calculations
SSPE <- deviance(explo.mod2);   df_PE <- df.residual(explo.mod2); MSPE <- SSPE / df_PE
SSLF <- SSE - SSPE;      df_LF <- df_E - df_PE;   MSLF <- SSLF / df_LF
F_LF <- MSLF / MSPE;   p.F.LF <- 1 - pf(F_LF, df_LF, df_PE)
round(cbind(MSLF, MSPE, F_LF, p.F.LF), 4)

## Plot of residuals versus predicted for linear regression model
plot(resid(explo.mod1) ~ predict(explo.mod1), xlab="fitted values",
   ylab="residuals")
abline(h=0)


##### Example 1.8

shapiroTest(resid(explo.mod1))    ## Shapiro-Wilk Test for residuals from linear model
qqnorm(resid(explo.mod1)); qqline(resid(explo.mod1))    ## Normal probability plot
```

1.10.2 Shotgun Pellet Spread Experiment

```
## Read in shotgun spread data
sg1 <- read.table("http://www.stat.ufl.edu/~winner/data/shotgun_spread.dat",
    header=F, col.names=c("crtrdg","dist.X","spread.Y","sd.sprd"))
attach(sg1)

## Select only cases where Cartridge is brand 2
spread.Y2 <- spread.Y[crtrdg == 2]    ## Y = spread (spread.Y2)
dist.X2 <- dist.X[crtrdg == 2]        ## X = distance (dist.X2)
sd.sprd2 <- sd.sprd[crtrdg == 2]      ## Std. Dev. of spreads at the X levels

##### Example 1.9
cbind(spread.Y2, dist.X2, sd.sprd2)   ## Prints the data
plot(spread.Y2 ~ dist.X2)             ## Scatterplot of Y vs X
abline(lm(spread.Y2 ~ dist.X2))       ## Adds fitted equation - OLS

sg.mod1 <- lm(spread.Y2 ~ dist.X2)    ## Fits OLS regression
summary(sg.mod1)                      ## Summary of model
anova(sg.mod1)                        ## ANOVA of model
e1 <- resid(sg.mod1)                  ## Save model residuals
yhat1 <- predict(sg.mod1)             ## Save model fitted values

plot(e1 ~ yhat1)                      ## Plot residuals versus fitted values
abline(h=0)                           ## Add horizontal line at e = 0

shapiro.test(e1)                      ## Shapiro-Wilk test for normaility of residuals
qqnorm(e1); qqline(e1)                ## Normal Probability Plot of residuals

library(MASS)

bc.mod1 <- boxcox(sg.mod1,plotit=T)   ## Run Box-Cox transformation on sg.mod1
print(cbind(bc.mod1$x,bc.mod1$y))     ## Print out results (lambda,log-like)
print(bc.mod1$x[which.max(bc.mod1$y)])  ## Print out "best" lambda
ci.bc <- max(bc.mod1$y)-0.5*qchisq(0.95,1)   ## Obtain cut-off for 95% CI (in log-like)
print(bc.mod1$x[bc.mod1$y>= ci.bc])   ## Print Values of lambda in 95% CI

## Fit model suggested by Box-Cox transformation and perform residual analysis
sg.mod2 <- lm(I(spread.Y2^(1/4)) ~ dist.X2)
summary(sg.mod2)
e2 <- resid(sg.mod2)
```

```
yhat2 <- predict(sg.mod2)

shapiro.test(e2)
qqnorm(e2); qqline(e2)


##### Example 1.11
plot(e1 ~ yhat1)                    ## Plot residuals vs fitted values for model 1

sg.mod3 <- lm(I(e1^2) ~ dist.X2)    ## Fit model relating e^2 vs X from model 1
anova(sg.mod3)                      ## Obtain ANOVA for SSReg_{e^2} for "manual test"

## Use lmtest package and bptest function for direct test (default is studentized residuals)
library(lmtest)
bptest(sg.mod1, studentize=FALSE)

## Conduct Breusch-Pagan test for Model 2 from Box-Cox transformation
library(lmtest)
bptest(sg.mod2, studentize=FALSE)


##### Example 1.12
reg.wt <- 1/(sd.sprd2^2)            ## Assign weights to individual cases = 1/(Group var)

sg.mod4 <- lm(spread.Y2 ~ dist.X2, weight=reg.wt)   ## Run weighted least squares
summary(sg.mod4)                                    ## Obtain model summary
anova(sg.mod4)                                      ## Obtain model ANOVA
e4 <- resid(sg.mod4)                                ## Save residual values
yhat4 <- predict(sg.mod4)                           ## Save fitted values
plot(e4 ~ yhat4)                                    ## Plot residuals vs fitted values


##### Example 1.13
### GLS - Power variance model - Variance is power of mean
##  sigma_i = sigma*(mu_i^delta)

library(nlme)
sg.mod5 <- gls(spread.Y2 ~ dist.X2, weights = varPower(form = ~ fitted(.)), method="ML")
summary(sg.mod5)
intervals(sg.mod5)

e5 <- resid(sg.mod5, type="p")       ## Save the Pearson (studentized) residuals
yhat5 <- predict(sg.mod5)            ## Save the fitted values
plot(e5 ~ yhat5)                     ## Plot studentized residuals vs fitted values

## Re-fit the original model using the gls function and ML estimation for comparison
sg.mod1a <- gls(spread.Y2 ~ dist.X2, method="ML")
summary(sg.mod1a)

anova(sg.mod1a, sg.mod5)    ## Compare Power Variance model with constant var model (delta=0)
```

1.10.3 Minneapolis/St. Paul Annual Temperature Data 1900-2015

```
## Read in Minneapolis/St. Paul Temperature Data
minnspw <- read.csv("http://www.stat.ufl.edu/~winner/data/minn_stp_weather.csv")
attach(minnspw); names(minnspw)

## Obtain year mean temperatures (data are given by month) = yrMean
## Convert years so that first year is 0 and last is n-1   = yrYear
(yrMean <- as.numeric(tapply(meanTemp,Year,mean)))
(yrMin <- min(Year))
```

```
(yrMax <- max(Year))
(yrYear <- seq((yrMin-yrMin),(yrMax-yrMin)))

##### Example 1.14
plot(yrYear, yrMean)                ## Plot(X=year, Y=mean temp)
abline(lm(yrMean ~ yrYear))         ## Add fitted line relating temp to year
mod1.lo <- loess(yrMean ~ yrYear)   ## Fit smooth loess model
xv <- 0:115                         ## Assign X-values for fitted
yv <- predict(mod1.lo, data.frame(yrYear=xv))  ## Obtain fitted values
lines(xv, yv, lty=2)                ## Add loess fit to original plot

msw.mod1 <- lm(yrMean ~ yrYear)     ## Fit OLS model
summary(msw.mod1)                   ## Summary of model
anova(msw.mod1)                     ## ANOVA of model
e1 <- resid(msw.mod1)               ## Save residuals
plot(e1, type="l")                  ## Line plot of residuals

## Direct Calculation of Durbin-Watson statistic
DW1 <- 0
for (i in 2:length(yrMean)) DW1 <- DW1 + (e1[i]-e1[i-1])^2
DW2 <- sum(e1^2)
DW <- DW1 / DW2
cbind(DW1, DW2, DW)

## durbinWatsonTest function in car package
library(car)
durbinWatsonTest(msw.mod1)

## gls function used to fit EGLS with AR(1) and ARMA(0,1)=MA(1) errors
library(nlme)
msw.mod2 <- gls(yrMean ~ yrYear, method="ML")
msw.mod3 <- gls(yrMean ~ yrYear, correlation=corAR1(), method="ML")
summary(msw.mod3)
intervals(msw.mod3)
msw.mod4 <- gls(yrMean ~ yrYear, correlation=corARMA(p=0,q=1), method="ML")
anova(msw.mod2, msw.mod3)     ## Compares OLS w/ AR(1)
anova(msw.mod3, msw.mod4)     ## Compares AR(1) w/ MA(1)

## Plot of autocorrelation function - leads to MA(1) as possible model
plot(ACF(msw.mod2, maxLag=15), alpha=0.05)
```

1.10.4 Bird Migration in trans-Sahara

```
## Read in bird migration data
bird.mig <- read.csv("http://www.stat.ufl.edu/~winner/data/bird_migration.csv")
attach(bird.mig); names(bird.mig)

##### Exercise 1.15
bird.mod1 <- lm(migDate ~ latWntr)    ## Linear regression model
summary(bird.mod1)
anova(bird.mod1)
e1 <- resid(bird.mod1)
yhat1 <- predict(bird.mod1)

plot(migDate ~ latWntr)
abline(bird.mod1)
plot(e1 ~ yhat1)
abline(h=0)

rstudent(bird.mod1)
influence.measures(bird.mod1)
```

```
qt(1-.05/76,35)      ## Critical value for Studentized Residuals
qf(.50,2,38-2)       ## Critical value for Cook's D
```

1.10.5 NBA Over/Under and Total Points

```
## Read in NBA data
nba1415 <- read.csv("http://www.stat.ufl.edu/~winner/data/nbaodds201415.csv",
                 header=T)
attach(nba1415); names(nba1415)

## Exercise 1.16
## Keep only non-Overtime games
(N.pop <- length(TotalPts[OT==0]))   ## Population Size
Y <- TotalPts[OT==0]/100             ## Y = Total Points / 100
X1a <- OvrUndr[OT==0]/100            ## X1a = OvrUndr / 100
X1 <- X1a - mean(X1a)                ## X1 = X1a - mean(X1a)

nba.mod1 <- lm(Y ~ X1)               ## Fit population model
(beta0 <- coef(nba.mod1)[1])         ## Obtain beta0
(beta1 <- coef(nba.mod1)[2])         ## Obtain beta1
(sigma2 <- sigma(nba.mod1)^2)        ## Obtain sigma^2
mean(X1a); var(X1a); sd(X1a)         ## Obtain mean, var, SD of X
eps <- resid(nba.mod1)               ## Obtain epsilons
E.Y <- predict(nba.mod1)             ## Obtain Expected Values

## Various plots

plot(X1a,eps)             ## X=X1a, Y = eps
lines(lowess(eps ~ X1a))

par(mfrow=c(2,1))
plot(X1a,Y,main="Total Points (Y) vs Over/Under (X) - NBA 2014/15",
xlab="Over/Under",ylab="Total Points")
abline(lm(Y ~ X1a))
plot(E.Y,eps,main="Errors versus Expected Value")

hist(X1a,breaks=30,main="Over/Under")
hist(eps,breaks=30,main=expression(paste(epsilon)))
mean(X1); sd(X1); mean(eps); sd(eps); cor(X1,eps)
par(mfrow=c(1,1))

## Take n.sim samples, each of size n.sample and save estimates, SE's
## Program uses matrix form for quicker computation
set.seed(135678)
n.sim <- 100000
n.sample <- 25
beta.hat <- matrix(rep(0,2*n.sim),ncol=2)
s2 <- numeric(n.sim)
s2.beta <- matrix(rep(0,2*n.sim),ncol=2)

for (i in 1:n.sim) {
nba.sample <- sample(1:N.pop,n.sample,replace=F)
X.s <- cbind(rep(1,n.sample), X1[nba.sample])
Y.s <- Y[nba.sample]
eps.s <- eps[nba.sample]
XPX.s <- t(X.s) %*% X.s
beta.hat.s <- solve(XPX.s) %*% t(X.s) %*% Y.s
beta.hat[i,] <- t(beta.hat.s)
e.s <- Y.s - X.s%*%beta.hat.s
s2[i]  <- (t(e.s) %*% (e.s))/(n.sample-2)
s2.b <- s2[i] * solve(XPX.s)
s2.beta[i,1] <- s2.b[1,1]; s2.beta[i,2] <- s2.b[2,2]
}

## Summarize Results
mean(s2)
mean(beta.hat[,1]); mean(beta.hat[,2])
beta1.CI.lo <- beta.hat[,2] + qt(.025,n.sample-2) * sqrt(s2.beta[,2])
beta1.CI.hi <- beta.hat[,2] + qt(.975,n.sample-2) * sqrt(s2.beta[,2])

## Empirical coverage rate of 95% CI's for Beta_1
sum(beta1.CI.lo <= beta1 & beta1.CI.hi >= beta1) / n.sim

par(mfrow=c(2,1))
hist(beta.hat[,1],breaks=50,main="Histogram of OLS Estmates of B0")
abline(v=beta[1])
hist(beta.hat[,2],breaks=50,main="Histogram of OLS Estmates of B1")
abline(v=beta[2])
par(mfrow=c(1,1))
plot(beta.hat[,1],beta.hat[,2],pch=16,cex=.4,
  main="Scatterplot of Estimates of B1 vs B0",xlab="B0",ylab="B1")
```

Chapter 2

Multiple Linear Regression

Typically, there is a group of more than one potential predictor variable, and the model generalizes to multiple linear regression. The calculations become more complex and make use of matrix algebra, but conceptually, the ideas remain the same. The notation used here will use p as the number of predictors, and $p' = p + 1$ as the number of parameters in the model (including the intercept). Note that the matrix form of the models are identical, they just differ in terms of the dimensions of the matrices used in computations. The model can be written as follows

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

Computer packages are used to obtain least squares (and maximum likelihood) estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize the error sum of squares. The fitted values, residuals, and error sum of squares are obtained as follow.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2$$

The degrees of freedom for error are now $n - p' = n - (p + 1)$, as there are $p' = p + 1$ estimated parameters (regression coefficients), and the degrees of freedom for regression is p. In the multiple linear regression model, β_j represents the change in $E\{Y\}$ when X_j increases by 1 unit, with all other predictor variables being held constant. It is referred to as the **partial regression coefficient**.

The model is flexible, allowing polynomial terms, indicator variables for categorical predictors, and cross-product terms to allow for interactions among predictors. These situations will be covered in this chapter.

Example 2.1: Recycling Program in Scotland

A study was conducted in Scotland to determine whether local authority differences in recycling policies were associated with amount of recycling (Baird, Curry, and Reid, 2013, [7]). There were three predictor variables (each at the household level) for $n = 31$ local authority districts. They were: recycling capacity (X_1, litres/week/hhold), residual waste capacity (X_2, litres/week/hhold), and number of extended recycling materials collected (X_3). The response was the yield of extended materials collected (Y, kg/week/hhold). The data are given in Table 2.1 and a scatterplot matrix of the data is given in Figure 2.1.

∇

recArea	recycCap	residCap	extMat	extYld
Aberdeen City	62.5	240	5	2.11
Angus	55	120	4	1.87
Argyll and Bute	120	120	5	2.8
Clackmannanshire	147.5	120	8	4.17
Dumfries and Galloway	40	240	1	0.88
Dundee City	87.5	240	6	2.35
East Ayrshire	147.5	120	4	3.68
East Dunbartonshire	110	240	5	2.65
East Lothian	50	240	5	2.81
East Renfrewshire	75	240	6	2.36
Edinburgh, City of	92.5	240	6	2.04
Eileen Siar	120	120	5	2.97
Falkirk	147.5	120	9	3.89
Fife	90	90	3	3.02
Glasgow City	60	240	3	1.79
Highland	27.5	240	2	1.63
Inverclyde	120	120	4	3.44
Midlothian	88	120	5	3.85
Moray	80	120	6	2.66
North Ayrshire	133.75	120	7	4.16
North Lanarkshire	155	120	8	4.37
Orkney	27.5	240	3	1.35
Perth and Kinross	120	120	4	2.75
Renfrewshire	145	120	6	3.1
Scottish Borders	70	180	7	2.88
Shetland Islands	74	240	4	1.24
South Ayrshire	87.5	120	6	3.58
South Lanarkshire	155	120	7	3.84
Stirling	55	120	8	3.27
West Dunbartonshire	82.5	240	6	2.22
West Lothian	120	120	5	3.03

Table 2.1: Recycling Capacity (X_1), Residual (non-recycling) capacity (X_2), number of extended materials recycled (X_3) and Yield of extended materials recycled (Y) for 31 Scottish localities

2.1 Testing and Estimation for Partial Regression Coefficients

Once the model is fit, the estimated regression coefficients and the standard errors for each coefficient are also computed. Actually, the estimated variance-covariance matrix for the coefficients is obtained which is used to obtain variances and standard errors of linear functions of regression coefficients.

To test whether Y is associated with X_j, after controlling for the remaining $p-1$ predictors, is to test whether $\beta_j = 0$. This is equivalent to the t-test from simple regression (in general, the test can be whether a regression coefficient is any specific number, although software packages are testing whether it is 0 by default).

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-p'} \qquad P\text{-value} : 2P(t_{n-p'} \geq |t_{obs}|)$$

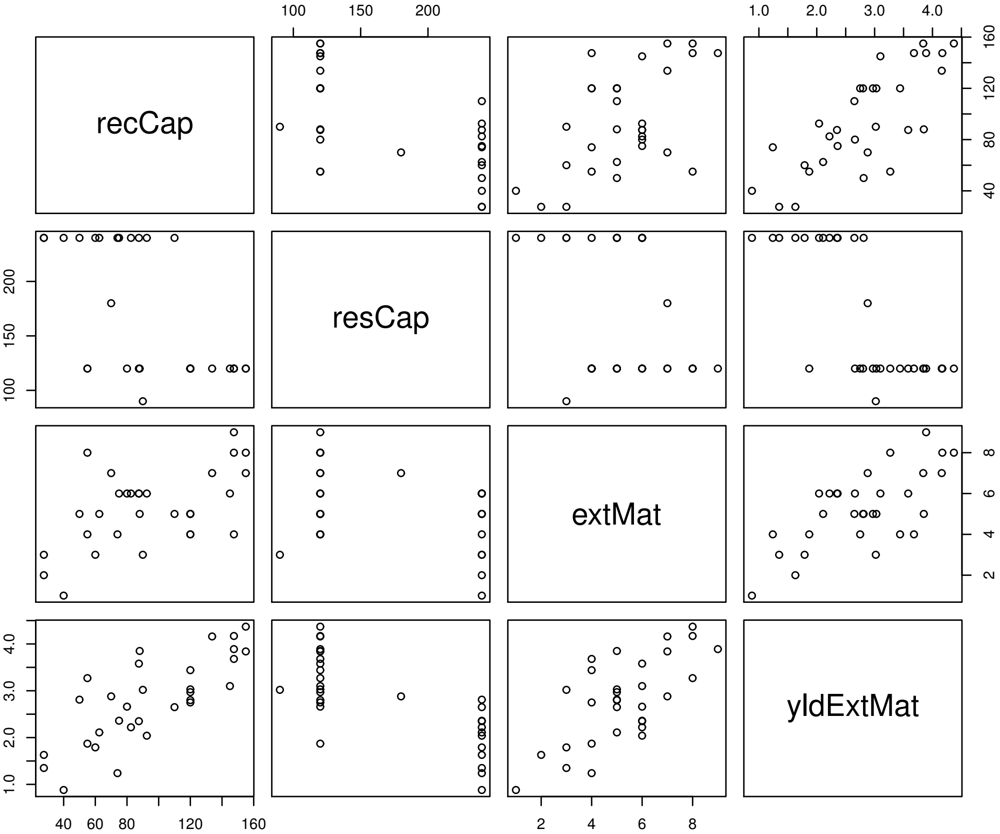

Figure 2.1: Scatterplot matrix for Scottish recycling study

One-sided tests make the same adjustments as in simple linear regression.

$$H_A^+ : \beta_j > \beta_{j0} \qquad RR : t_{obs} \geq t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \geq t_{obs})$$

$$H_A^- : \beta_j < \beta_{j0} \qquad RR : t_{obs} \leq -t_{\alpha,n-p'} \qquad P\text{-value} : P(t_{n-p'} \leq t_{obs})$$

A $(1-\alpha)100\%$ confidence interval for β_j is obtained as follows.

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'} \hat{SE}\{\hat{\beta}_j\}$$

Note that the Confidence Interval represents the values of β_{j0} for which the two-sided test: $H_0 : \beta_j = \beta_{j0} \quad H_A : \beta_j \neq \beta_{j0}$ fails to reject the null hypothesis.

Example 2.2: Recycling Program in Scotland

The tests for whether yield is associated with each predictor, controlling for all other predictors are all significant with t-values of $t_1 = 2.573$ for recycling capacity, $t_2 = -3.848$ for residual capacity and $t_3 = 3.775$ for number of extra materials. All else being equal, as recycling capacity and number of extra materials increase, yield of extra materials being recycled increases. As residual capacity increases, yield of extra materials being recycled decreases. A plot of residuals versus fitted values is given in Figure 2.2, it shows no evidence of non-constant error variance. The P-value for the Shapiro-Wilk test of normality of error terms is .6705, providing no concern about non-normality of errors.

∇

```
> recycle.mod1 <- lm(yldExtMat ~ recCap + resCap + extMat)
> summary(recycle.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.176991   0.496792   4.382 0.000160 ***
recCap       0.007360   0.002860   2.573 0.015887 *
resCap      -0.006347   0.001650  -3.848 0.000661 ***
extMat       0.187319   0.049617   3.775 0.000799 ***

Residual standard error: 0.4207 on 27 degrees of freedom
Multiple R-squared:  0.8097,    Adjusted R-squared:  0.7885
F-statistic: 38.29 on 3 and 27 DF,  p-value: 7.253e-10

> confint(recycle.mod1)
                   2.5 %       97.5 %
(Intercept)  1.157657813  3.196325143
recCap       0.001491434  0.013229452
resCap      -0.009731842 -0.002962482
extMat       0.085514199  0.289123976

> e1 <- resid(recycle.mod1)
> yhat1 <- predict(recycle.mod1)
> plot(e1 ~ yhat1)
> abline(h=0)
> shapiro.test(e1)
        Shapiro-Wilk normality test
data:  e1
W = 0.97519, p-value = 0.6705
```

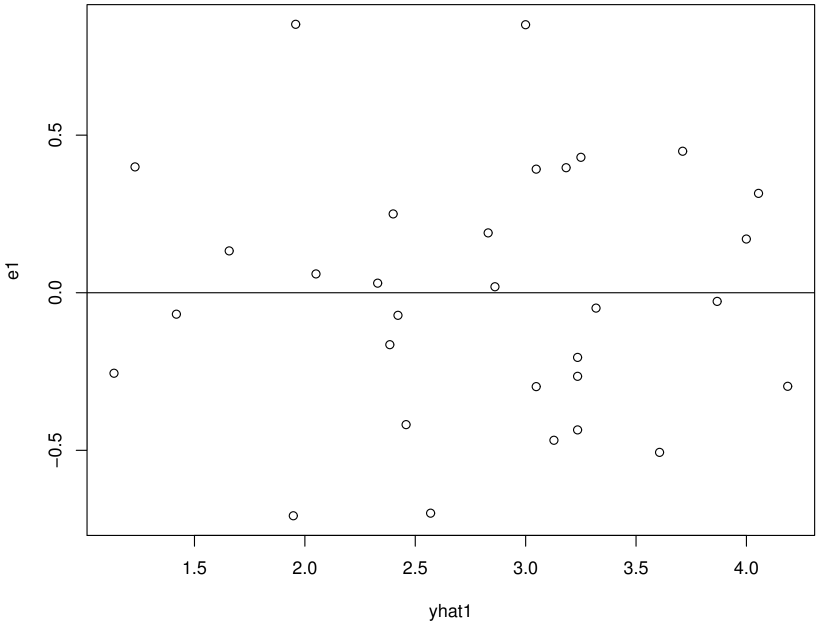

Figure 2.2: Plot of residuals versus fitted values for Scottish recycling study

2.2 Analysis of Variance

When there is no association between Y and the set of predictors $X_1, \ldots, X_p$ ($\beta_1 = \cdots = \beta_p = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **Total Sum of Squares**, just as with simple regression.

When there is an association between Y and at least one of $X_1, \ldots, X_p$ (not all $\beta_i = 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation ”explained” by the regression of Y on $X_1, \ldots, X_p$ (as opposed to having ignored $X_1, \ldots, X_p$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

Each sum of squares has a **Degrees of Freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - p'$. The **Regression Degrees of**

Source	df	SS	MS	F_{obs}	P-value
Regression (Model)	p	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{p}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{p,n-p'} \geq F_{obs})$
Error (Residual)	$n - p'$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-p'}$		
Total (Corrected)	$n - 1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 2.2: Analysis of Variance Table for Multiple Linear Regression

Freedom is $df_{\text{Regression}} = p$. Note that when there is $p = 1$ predictor, this generalizes to simple regression.

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - p' + p$$

Error and Regression Sums of Squares have **Mean Squares**, which are the Sums of Squares divided by their corresponding Degrees of Freedom: $MSE = SSE/(n - p')$ and $MSR = SSR/p$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \sum_{i=1}^{n}\left(\sum_{j=1}^{p} X_{ij}\beta_j\right)^2 \geq \sigma^2$$

Note that when $\beta_1 = \cdots = \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \cdots = \beta_p = 0$ is by the F-test.

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0$$

$$TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P\text{-value} : P(F_{p,n-p'} \geq F_{obs})$$

The Analysis of Variance is typically set up in a table as in Table 2.2.

A measure often reported from a regression analysis is the **Coefficient of Determination** or R^2. This represents the variation in Y "explained" by $X_1, \ldots, X_p$, divided by the total variation in Y.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq R^2 \leq 1$$

The interpretation of R^2 is the proportion of variation in Y that is "explained" by $X_1, \ldots, X_p$, and is often reported as a percentage ($100R^2$).

Example 2.3: Recycling Program in Scotland

Referring back to Example 2.2, the F-statistic for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is $F_{obs} = 38.29$, based on 3 and 27 degrees of freedom. There is strong evidence that $\beta_1 \neq 0$ and/or $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$, which is not surprising given that all of the partial t-tests were significant. The coefficient of determination is $R^2 = 0.8097$, thus the model explains a fairly large fraction of variation in yield. R doesn't explicitly give SSR, it gives sums of squares attributable to the various predictor variables, as described below.

∇

2.2.1 Sequential and Partial Sums of Squares

The **Sequential Sums of Squares** are the regression sums of squares for each independent variable as they are added to a model one-at-a-time. If there are p variables, and in the model fit are entered in order $X_1, X_2, \ldots, X_p$, then the sequential sums of squares are computed as follow.

- Fit the simple regression model: $\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and obtain $SSR\,(X_1)$, its regression sum of squares
- Fit the 2 variable model: $\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ and obtain $SSR\,(X_1, X_2)$, the regression sum of squares
- Continue until the p variable model: $\hat{Y}_p = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$ and obtain $SSR\,(X_1, X_2, \ldots, X_p)$

$$X_1 : \quad SSR\,(X_1) \qquad X_2 : \quad SSR\,(X_2|X_1) = SSR\,(X_1, X_2) - SSR\,(X_1) \qquad \ldots$$

$$X_p : \quad SSR\,(X_p|X_1, X_2, \ldots, X_{p-1}) = SSR\,(X_1, X_2, \ldots, X_p) - SSR\,(X_1, X_2, \ldots, X_{p-1})$$

The **Partial Sums of Squares** are the regression sum of squares for each independent variable as each is added to a model containing the remaining $p-1$ predictors. For X_1 and X_p for instance, there are the following definitions.

$$X_1 : \quad SSR\,(X_1, X_2, \ldots, X_p) - SSR\,(X_2, X_2, \ldots, X_p) \qquad X_p : \quad SSR\,(X_1, X_2, \ldots, X_p) - SSR\,(X_1, X_2, \ldots, X_{p-1})$$

The sequential sums of squares add up to the regression sum of squares for the model. Only if the X variables are uncorrelated in a controlled experiment will the partial sums of squares sum to the regression sum of squares.

Software packages will compute F-tests for each independent variable by taking each sum of squares divided by MSE for the model. The F-tests based on the partial sums of squares are identical to the 2-sided t-tests for the corresponding partial regression coefficients.

Example 2.4: Recycling Program in Scotland

The sequential and partial sums of squares for the recycling data are given below. The sequential sums of squares are obtained from the **anova** function and the partial sums of squares are given by the **drop1** function, the F-tests are not default and must be selected as an option for the **drop1** function. Note that the last variable in the model by definition has the same sequential and partial sums of squares. The F-statistics for the partial sums of squares are the square of the t-statistics for the corresponding partial regression coefficients: $6.6216 = (2.573)^2$, $14.8050 = (-3.848)^2$ and $14.2531 = (3.775)^2$.

∇

```
> anova(recycle.mod1)
Analysis of Variance Table
Response: yldExtMat
          Df  Sum Sq Mean Sq F value    Pr(>F)
recCap     1 15.1165 15.1165  85.396 7.508e-10 ***
resCap     1  2.6918  2.6918  15.206 0.0005770 ***
extMat     1  2.5230  2.5230  14.253 0.0007994 ***
Residuals 27  4.7794  0.1770

> drop1(recycle.mod1,test="F")
Single term deletions
Model:
yldExtMat ~ recCap + resCap + extMat
       Df Sum of Sq    RSS     AIC F value    Pr(>F)
<none>              4.7794 -49.960
recCap  1    1.1721 5.9516 -45.160  6.6216 0.0158873 *
resCap  1    2.6207 7.4001 -38.407 14.8050 0.0006612 ***
extMat  1    2.5230 7.3025 -38.819 14.2531 0.0007994 ***
```

The sequential sums of squares for the p variables sum to the regression sum of squares for the model. In many cases, it is useful to obtain the sum of squares for a group of $p - g$ predictors. For instance if there are $p = 5$ predictors and interest is in the contribution of the last $p - g = 3$ predictors, given the first $g = 2$ predictors were included in a model, obtain the following sums of squares. These can be computed directly in R by fitting separate models, see examples below. The second equality holds since if there are two models, based on the same dataset, then $TSS = SSR_1 + SSE_1 = SSR_2 + SSE_2$.

$$SSR\left(X_3, X_4, X_5 | X_1, X_2\right) = SSR\left(X_1, X_2, X_3, X_4, X_5\right) - SSR\left(X_1, X_2\right) = SSE\left(X_1, X_2\right) - SSE\left(X_1, X_2, X_3, X_4, X_5\right)$$

2.2.2 Coefficients of Partial Determination

Just as R^2 describes the predictive ability of a set of predictors, the amount of variation explained by subsequent predictor variables can be measured. Suppose the variables have been entered into the model in order $X_1, X_2, \ldots, X_p$. Then consider the following sequence of coefficients of partial determination with TSS representing the Total sum of squares.

$$R^2_{Y1} = \frac{SSR\left(X_1\right)}{TSS} \qquad R^2_{Y2|1} = \frac{SSR\left(X_1, X_2\right) - SSR\left(X_1\right)}{TSS - SSR\left(X_1\right)} \qquad \cdots$$

$$R^2_{Yp|1\ldots p-1} = \frac{SSR\left(X_1, \ldots, X_p\right) - SSR\left(X_1, \ldots, X_{p-1}\right)}{TSS - SSR\left(X_1, \ldots, X_{p-1}\right)}$$

Each coefficient measures the fraction of the variation that was not explained by the previous predictor(s) that is explained by the current predictor.

Example 2.5: Recycling Program in Scotland

For the Scottish recycling study, when the sequential and error sums of squares are added up, the Total sum of squares is $TSS = 25.1107$. The coefficients of partial determination (for the order of the independent variables used here) are given below.

$$SSR\left(X_1\right) = 15.1165 \qquad SSR\left(X_1, X_2\right) = 15.1165 + 2.6918 = 17.8083$$

$$SSR\left(X_1, X_2, X_3\right) = 17.8083 + 2.5230 = 20.3313$$

$$R^2_{Y1} = \frac{15.1165}{25.1107} = .6020 \qquad R^2_{Y2|1} = \frac{17.8083 - 15.1165}{25.1107 - 15.1165} = \frac{2.6918}{9.9942} = .2693$$

$$R^2_{Y3|12} = \frac{20.3313 - 17.8083}{25.1107 - 17.8083} = \frac{2.5230}{7.3024} = .3455$$

Recycling capacity explains about 60% of the variation in yield. Residual capacity explains about 27% of the variation in yield that is not explained by recycling capacity. Finally, the extra number of recyclable materials explains about 34.5% of the variation not explained by recycling and residual capacity.

∇

2.3 Testing a Subset of $\beta^s = 0$

The F-test from the Analysis of Variance and the t-tests represent extremes as far as model testing (all variables simultaneously versus one-at-a-time). Often interest is in testing whether a group of predictors do not improve prediction, after controlling for the remaining predictors.

Suppose that after controlling for g predictors, the goal is to test whether the remaining $p - g$ predictors are associated with Y. That is, to test between the following hypotheses.

$$H_0 : \beta_{g+1} = \cdots \beta_p = 0 \qquad H_A : \text{ Not all of } \beta_{g+1}, \ldots, \beta_p = 0$$

Note that, the t-tests control for all other predictors, while here, the purpose is to control for only $X_1, \ldots, X_g$. To do this, fit two models: the **Complete** or **Full Model** with all p predictors, and the **Reduced Model** with only the g "control" variables. For each model, compute the Regression and Error sums of squares, as well as R^2. This leads to the test statistic, rejection region, and P-value.

$$TS : F_{obs} = \frac{\left[\frac{SSE(R) - SSE(F)}{(n-g') - (n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F) - SSR(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2 - R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR: F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P\text{-value}: P(F_{p-g,n-p'} \geq F_{obs})$$

Example 2.6: LPGA Prize Winnings and Performance Measures - 2009

A multiple regression model is fit, relating prize money won (Y=log(winnings/tournaments)) to $p = 7$ performance variables for $n = 146$ golfers for the Ladies Professional Golf Association 2009 season. The predictors are: **drive** (X_1, average drive distance, yards), **frwy** (X_2, percent of fairways hit), **grnReg** (X_3, percent of greens in regulation), **putts**, (X_4, average putts per round), **sandSave** (X_5, percent of saves from sand), **pctile** (X_6, average percentile in tournaments, higher is better), and **strksRnd** (X_7, average strokes per round). Note that in particular that the expected signs for putts and strokes per round should be negative as fewer putts and strokes per round are better in golf. A scatterplot matrix of the data is given in Figure 2.3. The summary of the regression model is given below.

Note that log of winnings per round was used because winnings per round is highly skewed. A Box-Cox transformation (not shown here) leads to the choice. The model explains 86.6% of the variance in the (transformed) winnings. The variables drive distance, fairway percent, percentile, and strokes per round all have significant partial regression coefficients at the $\alpha = 0.05$ significance level. Keep in mind that the predictors are highly correlated among themselves. This issue will be discussed in a later section on multicollinearity.

Consider testing whether $\beta_3 = \beta_4 = \beta_5 = 0$. That is whether, controlling the four significant predictors given above are in the model that neither greens in regulation, putts per round, or sand save percent are associated with winnings. This test is different from the t-tests, in that the t-tests for the individual predictors control for all of the other predictors. The second model contains only the $g = 4$ control variables: drive, frwy, pctile, and strksRnd. The F-test is directly computed below. The R **anova** function can be used to compute the F-statistic and P-value by giving both model fits as inputs, see the R output below).

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0 \quad SSE(F) = 29.948 \quad p = 7 \quad n - p' = 138 \quad SSE(R) = 30.171 \quad g = 4 \quad n - g' = 141$$

$$TS: F_{obs} = \frac{\left[\frac{30.171-29.948}{141-138}\right]}{\left[\frac{29.948}{138}\right]} = \frac{0.0743}{0.2170} = 0.3424 \qquad P\left(F_{3,138} \geq 0.3424\right) = .7947$$

Fail to reject the null hypothesis. The reduced model is a "better" fit in terms of being more parsimonious, without losing predictive ability.

∇

```
### Full Model (p=7)
> lpga.mod1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
> summary(lpga.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.875511  14.615851   4.302 3.19e-05 ***
drive       -0.015586   0.007131  -2.186 0.030533 *
frwy        -0.032081   0.009554  -3.358 0.001015 **
grnReg      -0.011984   0.034657  -0.346 0.730025
putts        0.117793   0.152163   0.774 0.440181
```

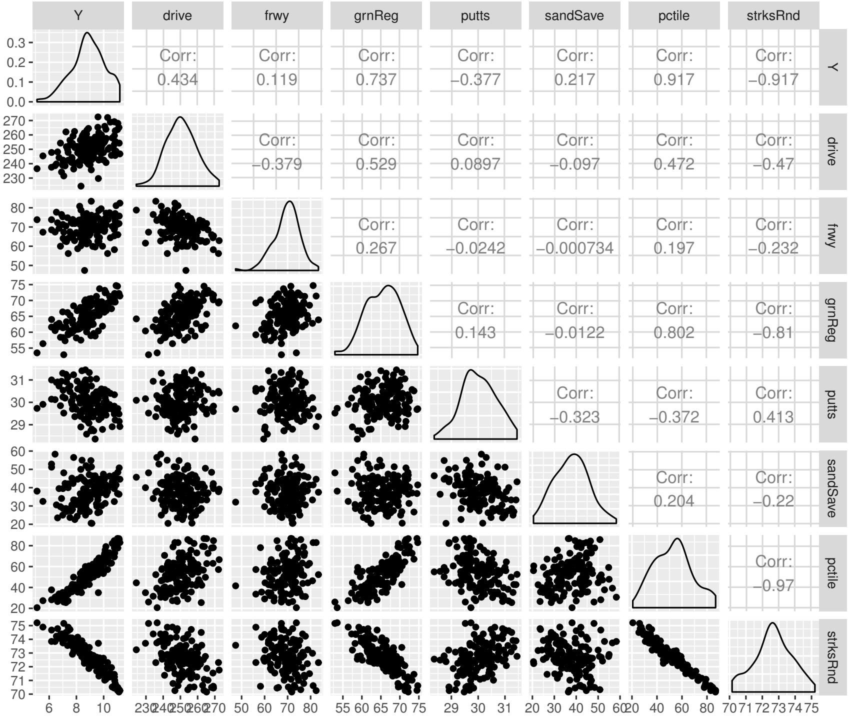

Figure 2.3: Scatterplot matrix for LPGA Performance data

```
sandSave    -0.000756   0.005560  -0.136 0.892038
pctile       0.033704   0.010689   3.153 0.001982 **
strksRnd    -0.720633   0.205876  -3.500 0.000626 ***

Residual standard error: 0.4658 on 138 degrees of freedom
Multiple R-squared:  0.8658,    Adjusted R-squared:  0.859
F-statistic: 127.2 on 7 and 138 DF,  p-value: < 2.2e-16
> anova(lpga.mod1)
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)
drive       1 42.049  42.049 193.7629 < 2.2e-16 ***
frwy        1 20.919  20.919  96.3948 < 2.2e-16 ***
grnReg      1 59.657  59.657 274.9007 < 2.2e-16 ***
putts       1 54.432  54.432 250.8273 < 2.2e-16 ***
sandSave    1  1.021   1.021   4.7059 0.0317719 *
pctile      1 12.462  12.462  57.4263 4.611e-12 ***
strksRnd    1  2.659   2.659  12.2523 0.0006262 ***
Residuals 138 29.948   0.217

### Reduced Model (g=4)
> lpga.mod2 <- lm(Y ~ drive+frwy+pctile+strksRnd)
> summary(lpga.mod2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 57.108691  11.430003   4.996 1.70e-06 ***
drive       -0.012210   0.006001  -2.035 0.043743 *
frwy        -0.028932   0.008589  -3.369 0.000974 ***
pctile       0.035226   0.010475   3.363 0.000993 ***
strksRnd    -0.619578   0.142332  -4.353 2.56e-05 ***
```

```
Residual standard error: 0.4626 on 141 degrees of freedom
Multiple R-squared:  0.8648,    Adjusted R-squared:  0.861
F-statistic: 225.5 on 4 and 141 DF,  p-value: < 2.2e-16

> anova(lpga.mod2)
Analysis of Variance Table
Response: Y
           Df  Sum Sq Mean Sq F value    Pr(>F)
drive       1  42.049  42.049 196.512 < 2.2e-16 ***
frwy        1  20.919  20.919  97.762 < 2.2e-16 ***
pctile      1 125.954 125.954 588.635 < 2.2e-16 ***
strksRnd    1   4.055   4.055  18.949 2.563e-05 ***
Residuals 141  30.171   0.214

> anova(lpga.mod2, lpga.mod1)
Analysis of Variance Table
Model 1: Y ~ drive + frwy + pctile + strksRnd
Model 2: Y ~ drive + frwy + grnReg + putts + sandSave + pctile + strksRnd
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    141 30.171
2    138 29.948  3   0.22303 0.3426 0.7946
```

2.4 Models with Categorical Predictors, Interaction, and Polynomial Terms

In this section, three generalizations of the multiple regression model are considered. These models are fit in the exact same manner, and tests are conducted as in the case where the model contains p distinct numeric predictors. The first case allows for categorical predictors and makes use of **dummy** or **indicator variables** to represent the levels of the variable(s). The second case includes interaction terms which allows for the slope with respect to predictor variables to depend on level(s) that the other predictor(s) take on. The third case allows for **polynomial terms** which represent levels of a numeric variable raised to powers. Note that many models make use of two or more of these generalizations.

2.4.1 Models With Categorical (Qualitative) Predictors

Often, one or more categorical variables are included in a model. If there is a categorical variable with m levels, $m-1$ **dummy** or **indicator variables** will need to be created. The variable will take on 1 if the i^{th} observation is in that level of the variable, 0 otherwise. Note that one level of the variable will have 0^s for all $m-1$ dummy variables, making it the reference group. The β^s for the other groups (levels of the qualitative variable) reflect the difference in the mean for that group with the reference group, controlling for all other predictors.

Note that if the qualitative variable has 2 levels, there will be a single dummy variable, and the test for differences in the effects of the 2 levels will be a t-test, controlling for all other predictors. If there are $m-1 \geq 2$ dummy variables associated with the predictor, the F-test is used to test whether all $m-1$ β^s are 0, controlling for all other predictors.

Example 2.7: Crop Subsidence and Water Table Level in 3 Crop Varieties

A study in the Florida Everglades related annual subsidence (Y, in cm) to Water Table (X_1, in cm) for 3 crop varieties: Pasture, Truck Crop, and Sugarcane (Stephens and Johnson, 1951, [35]; Shih and Shih, 1978, [33]). The data are given in Table 2.3 and displayed in Figure 2.4. The crop variable is categorical with $m = 3$ levels, so $m - 1 = 2$ dummy variables are generated: $X_2 = 1$ if Pasture, 0 otherwise; $X_3 = 1$ if Truck Crop, 0 otherwise. The choice of "reference group" is arbitrary (Sugarcane in this case), and all relevant estimates are the same, regardless of which level is chosen. The model is given below.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i \quad i = 1, \ldots, 24$$

The model fits with (wr.mod1) and without (wr.mod2) the two crop dummy variables are given below, with the full model displayed in Figure 2.5.

With Crop Dummys: $\hat{Y}_1 = -1.1818 + 0.0639X_1 + 1.4965X_2 + 1.3291X_3 \quad SSE_1 = 1.8534 \quad df_{E1} = 24 - 4 = 20$

Without Crop Dummys: $\hat{Y}_2 = 0.1930 + 0.0572X_1 \quad SSE_2 = 12.293 \quad df_{E2} = 24 - 2 = 22$

Model 1: Pasture: $\hat{Y} = 0.3147 + 0.0639X_1$ Truck Crop: $\hat{Y} = 0.1473 + 0.0639X_1$ Sugarcane: $\hat{Y} = -1.1818 + 0.0639X_1$

The F-test for crop effects, controlling for water table level is testing $H_0 : \beta_2 = \beta_3 = 0$, with $p = 3$ independent variables in the full model, and $g = 1$ predictor in the reduced model. The F-test is given below. There is strong evidence of crop variety differences.

$$TS : F_{obs} = \frac{\left[\frac{12.293 - 1.8534}{22 - 20}\right]}{\left[\frac{1.8534}{20}\right]} = \frac{5.2198}{0.0927} = 56.31 \qquad P\left(F_{2,20} \geq 56.31\right) < .0001$$

$$\nabla$$

```
### Model 1 (Full Model)
> wr.mod1 <- lm(subsidence ~ waterTbl + pasture + truckCrop)
> summary(wr.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.181172   0.274270  -4.307 0.000344 ***
waterTbl     0.063882   0.003655  17.476 1.39e-13 ***
pasture      1.496518   0.154130   9.709 5.18e-09 ***
truckCrop    1.329058   0.153716   8.646 3.44e-08 ***

Residual standard error: 0.3044 on 20 degrees of freedom
Multiple R-squared:  0.9481,    Adjusted R-squared:  0.9403
F-statistic: 121.7 on 3 and 20 DF,  p-value: 5.157e-13

> anova(wr.mod1)
Analysis of Variance Table
Response: subsidence
          Df  Sum Sq Mean Sq F value    Pr(>F)
waterTbl   1 23.3931 23.3931 252.436 8.292e-13 ***
pasture    1  3.5119  3.5119  37.898 5.150e-06 ***
```

crop	subsidence	waterTbl	crop	subsidence	waterTbl	crop	subsidence	waterTbl
1	1.90	30.5	2	1.94	32.1	3	1.48	35.7
1	2.88	43.0	2	2.76	46.3	3	2.35	51.5
1	4.08	56.7	2	4.00	58.2	3	3.12	62.5
1	4.16	57.6	2	4.12	60.4	3	2.97	67.4
1	5.23	71.9	2	5.03	70.4	3	3.50	78.3
1	5.15	77.7	2	4.98	78.9	3	4.49	83.8
1	5.56	80.8	2	5.64	78.9	3	4.00	86.9
1	5.44	80.8	2	4.98	79.9	3	3.91	86.0

Table 2.3: Subsidence and Water Table Levels for 3 Crop Varieties: 1=Pasture, 2=Truck Crop, 3=Sugarcane

```
truckCrop  1  6.9277  6.9277  74.757 3.440e-08 ***
Residuals 20  1.8534  0.0927

### Model 2 (Reduced Model)
> wr.mod2 <- lm(subsidence ~ waterTbl)
> summary(wr.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.193047   0.593322   0.325    0.748
waterTbl    0.057214   0.008843   6.470 1.65e-06 ***

Residual standard error: 0.7475 on 22 degrees of freedom
Multiple R-squared:  0.6555,    Adjusted R-squared:  0.6399
F-statistic: 41.87 on 1 and 22 DF,  p-value: 1.649e-06

> anova(wr.mod2)
Analysis of Variance Table
Response: subsidence
          Df Sum Sq Mean Sq F value    Pr(>F)
waterTbl   1 23.393 23.3931  41.865 1.649e-06 ***
Residuals 22 12.293  0.5588

### Comparison of Models 1 and 2
> anova(wr.mod2, wr.mod1)
Analysis of Variance Table
Model 1: subsidence ~ waterTbl
Model 2: subsidence ~ waterTbl + pasture + truckCrop
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     22 12.2930
2     20  1.8534  2     10.44 56.327 6.068e-09 ***
```

2.4.2 Models With Interaction Terms

When the effect of one predictor depends on the level of another predictor (and vice versa), the predictors are said to **interact**. The way to model interaction(s) is to create a new variable that is the product of the 2 predictors (higher order interactions can also be included). Suppose the model contains Y, and 2 numeric predictors: X_1 and X_2. Create a new predictor $X_3 = X_1X_2$, and consider the following model.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = \beta_0 + \beta_2 X_2 + (\beta_1 + \beta_3 X_2) X_1$$

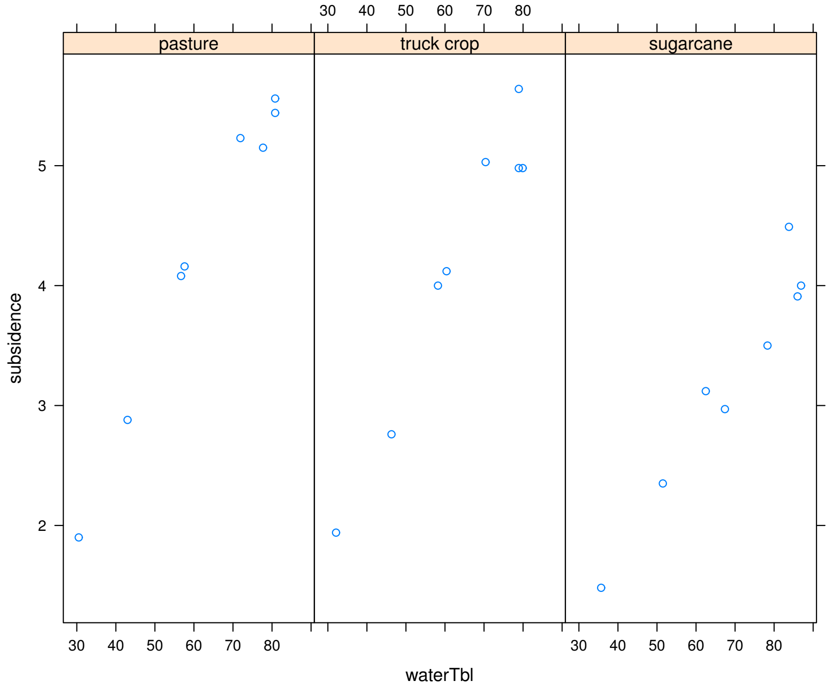

Figure 2.4: Plot of Subsidence by Water Table level by Crop Type

Thus, the slope with respect to X_1 depends on the level of X_2, unless $\beta_3 = 0$, which can be tested with a t-test. This logic extends to qualitative variables as well. Cross-product terms are obtained between numeric (or other categorical) predictors with the $m-1$ dummy variables representing the qualitative predictor. Then t-test $(m-1=1)$ or F-test $(m-1 \geq 2)$ can be conducted to test for interactions among predictors.

Example 2.8: Crop Subsidence and Water Table Level in 3 Crop Varieties

The interaction model allows the slopes relating subsidence to water table level to vary across the three crops. Two interaction terms are added to the additive model in Example 2.6: $X_4 = X_1X_2$ and $X_5 = X_1X_3$. The model is given below.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 \qquad \text{Sugarcane: } E\{Y\} = \beta_0 + \beta_1 X_1$$

$$\text{Pasture: } E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1 \qquad \text{Truck Crop: } E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$$

The test for interaction between water table level and crop type with respect to subsidence is of the form $H_0 : \beta_4 = \beta_5 = 0$. The (full) model containing the two interaction terms has $SSE(F) = 1.2607$ with $n - p' = 24 - 6 = 18$. It will be compared with the additive (reduced) model with $SSE(R) = 1.8534$ and $n - g' = 24 - 4 = 20$. There is evidence of differences in slopes among the crop types.

$$TS : F_{obs} = \frac{\left[\frac{1.8534-1.2607}{20-18}\right]}{\left[\frac{1.2607}{18}\right]} = \frac{0.2964}{0.0700} = 4.234 \qquad P\left(F_{2,18} \geq 4.234\right) = .0311$$

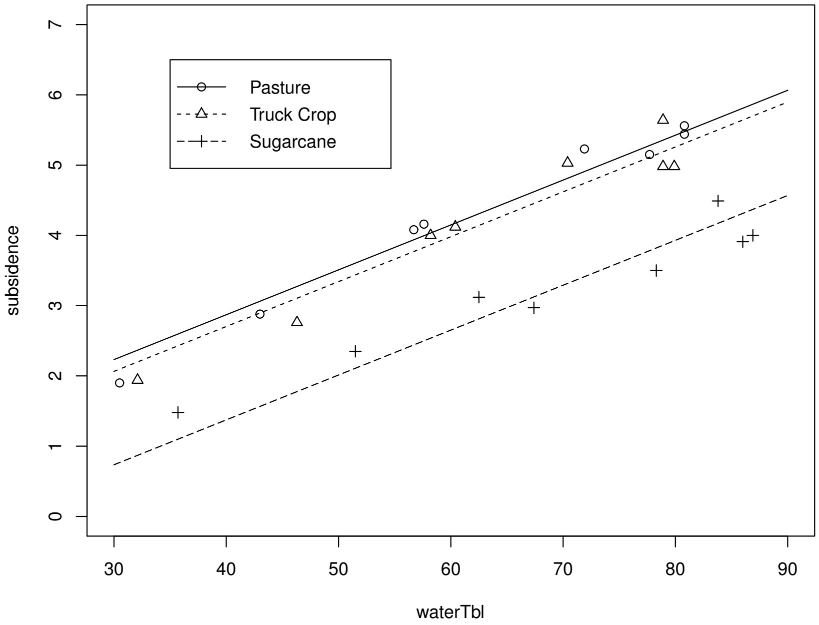

Figure 2.5: Fitted equations relating subsidence to water table level, separately by crop type - Additive model

The R output is given below and the fitted model is given in Figure 2.6. The plot reveals that the lines are very similar for the pasture and truck crop. Consider a model that forces the lines to be the same for these two crops and have a different intercept and slope than sugarcane. This can be done simply by creating a new dummy variable for the combined pasture and truck crop category, say $X_{23} = X_2 + X_3$, and fitting the following model (labeled Model 4 below).

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_{23} + \beta_3 X_1 X_{23}$$

The test comparing Models 3 and 4 is testing whether $H_0 : \beta_2 = \beta_3, \beta_4 = \beta_5$ in Model 3. Based on Model 4, $SSE(R) = 1.3804$ with $n - g' = 24 - 4 = 20$. The F-test and fitted models are given below. The simpler model is appropriate.

$$TS : F_{obs} = \frac{\left[\frac{1.3804-1.2607}{20-18}\right]}{\left[\frac{1.2607}{18}\right]} = \frac{0.0599}{0.0700} = 0.856 \qquad P\left(F_{2,18} \geq 0.856\right) = .4414$$

$$\text{Sugarcane: } \hat{Y} = -0.2976 + 0.0511X \qquad \text{Pasture/Truck Crop: } \hat{Y} = -0.1795 + 0.0705X$$

∇

```
## Model 3 (Interaction Model}
> wr.mod3 <- lm(subsidence ~ waterTbl + pasture + truckCrop +
+        I(waterTbl*pasture) + I(waterTbl*truckCrop))
> summary(wr.mod3)
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             -0.297627   0.386286  -0.770   0.4510
waterTbl                 0.051080   0.005431   9.406 2.27e-08 ***
pasture                  0.219489   0.518799   0.423   0.6773
truckCrop               -0.001777   0.538829  -0.003   0.9974
I(waterTbl * pasture)    0.019111   0.007620   2.508   0.0219 *
I(waterTbl * truckCrop)  0.019887   0.007918   2.512   0.0218 *

Residual standard error: 0.2647 on 18 degrees of freedom
Multiple R-squared:  0.9647,    Adjusted R-squared:  0.9549
F-statistic:  98.3 on 5 and 18 DF,  p-value: 2.014e-12

> anova(wr.mod3)
Analysis of Variance Table
Response: subsidence
                        Df  Sum Sq Mean Sq  F value    Pr(>F)
waterTbl                 1 23.3931 23.3931 333.9966 4.541e-13 ***
pasture                  1  3.5119  3.5119  50.1421 1.331e-06 ***
truckCrop                1  6.9277  6.9277  98.9103 9.716e-09 ***
I(waterTbl * pasture)    1  0.1508  0.1508   2.1537   0.15948
I(waterTbl * truckCrop)  1  0.4418  0.4418   6.3081   0.02178 *
Residuals               18  1.2607  0.0700

> anova(wr.mod1, wr.mod3)
Analysis of Variance Table
Model 1: subsidence ~ waterTbl + pasture + truckCrop
Model 2: subsidence ~ waterTbl + pasture + truckCrop + I(waterTbl * pasture) +
    I(waterTbl * truckCrop)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     20 1.8534
2     18 1.2607  2   0.59267 4.2309 0.03118 *

### Model 4 (Common Lines for Pasture and Truck Crop)
> p.tC <- pasture + truckCrop
> wr.mod4 <- lm(subsidence ~ waterTbl + p.tC + I(waterTbl*p.tC))
> summary(wr.mod4)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.297627   0.383467  -0.776  0.44674
waterTbl            0.051080   0.005391   9.475 7.77e-09 ***
p.tC                0.118066   0.459280   0.257  0.79975
I(waterTbl * p.tC)  0.019355   0.006648   2.912  0.00863 **

Residual standard error: 0.2627 on 20 degrees of freedom
Multiple R-squared:  0.9613,    Adjusted R-squared:  0.9555
F-statistic: 165.7 on 3 and 20 DF,  p-value: 2.726e-14

> anova(wr.mod4)
Analysis of Variance Table
Response: subsidence
                   Df  Sum Sq Mean Sq F value    Pr(>F)
waterTbl            1 23.3931 23.3931 338.926 5.218e-14 ***
p.tC                1 10.3275 10.3275 149.627 9.682e-11 ***
I(waterTbl * p.tC)  1  0.5851  0.5851   8.477  0.008629 **
Residuals          20  1.3804  0.0690

> anova(wr.mod4, wr.mod3)
Analysis of Variance Table

Model 1: subsidence ~ waterTbl + p.tC + I(waterTbl * p.tC)
Model 2: subsidence ~ waterTbl + pasture + truckCrop + I(waterTbl * pasture) +
    I(waterTbl * truckCrop)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 1.3804
2     18 1.2607  2   0.11971 0.8546  0.442
```

2.4.3 Models With Curvature

When a plot of Y versus one or more of the predictors displays curvature, polynomial terms can be included to "bend" the regression line. Often, to avoid multicollinearity, predictor(s) are centered, by subtracting off their mean(s). If the data show k bends, $k+1$ polynomial terms should be included. Suppose there is a single predictor variable, with 2 "bends" appearing in a scatterplot. Then, the model should include terms up to the a third order (cubic). Note that even if lower order terms are not significant, when a higher order term is significant, the lower order terms should be kept in the model (unless there is some physical reason not to). The following (cubic, with a single predictor variable, in this case) model could be fit.

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

If the goal is to test whether the fit is linear, as opposed to "not linear," the test $H_0 : \beta_2 = \beta_3 = 0$, using the F-test to compare the two model fits would be used.

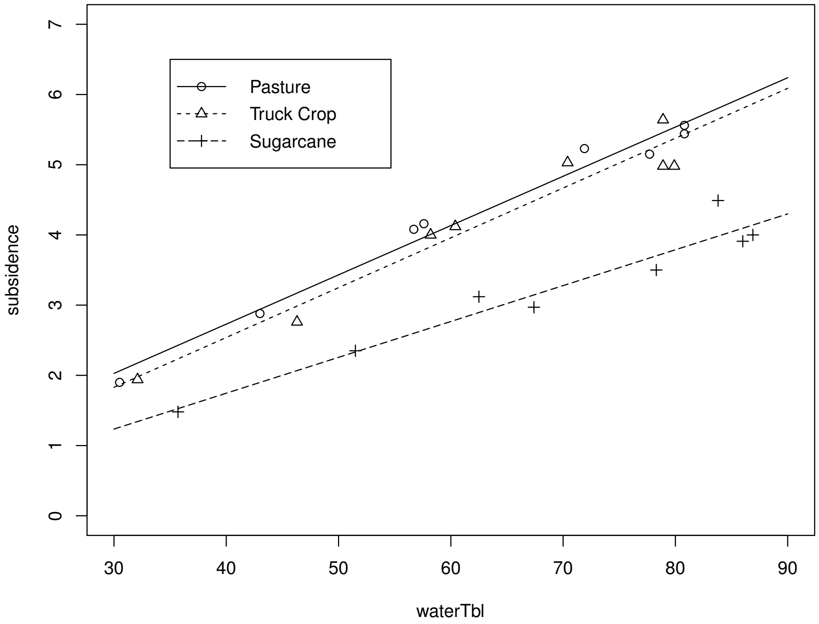

Figure 2.6: Subsidence versus Water Table level - Interaction Model

Example 2.9: Galvonometer Deflection in Experiments with Explosives

In Example 1.7, it was determined that there was evidence of a non-linearity in the relation between mean galvonometer deflection and wire area. There was evidence of a "single bend" so consider adding a quadratic term to the original linear model. Many researchers center the independent variable X when fitting polynomial models, however it is fit in the original units here which makes plotting results easier. The fitted linear and quadratic equations are given below, along with the t-test for the quadratic term. A plot of the data and the linear and quadratic models is given in Figure 2.7. There is strong evidence that $\beta_2 \neq 0$. A cubic model was considered, but the t-statistic was not significant (P=.1783, not shown here).

$$\text{Linear:}\hat{Y}_L = 184.4 - 0.695X \qquad \text{Quadratic: } \hat{Y}_Q = 196.6 - 1.119X + 0.00251X^2$$

$$H_0 : \beta_2 = 0 \qquad H_A : \beta_2 \neq 0 \qquad \hat{\beta}_2 = 0.002510 \qquad \hat{SE}\left\{\hat{\beta}_2\right\} = 0.0008162$$

$$t_{obs} = \frac{0.002510}{0.0008162} = 3.075 \qquad P = 2P\left(t_{19} \geq |3.075|\right) = .0062$$

∇

```
## Linear Model (n=22, p'=2)
```

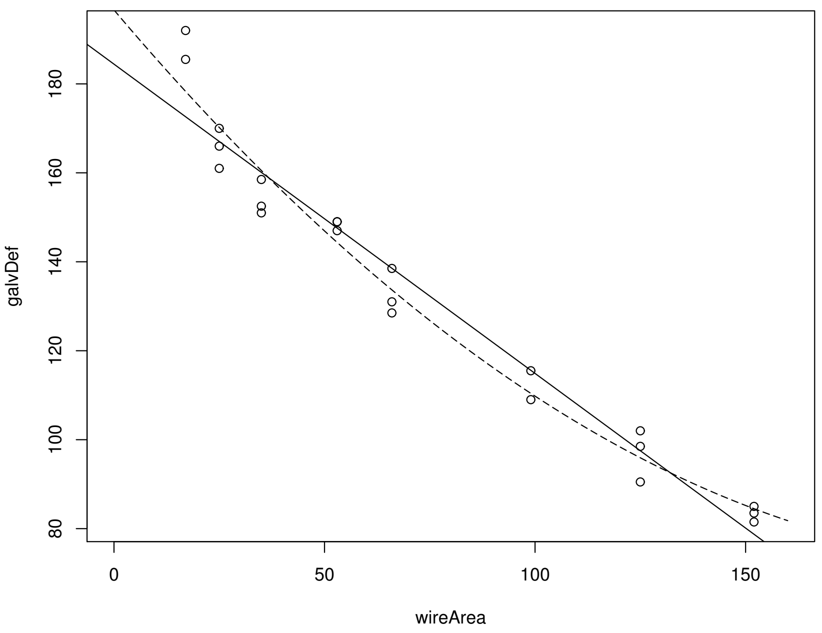

Figure 2.7: Plot of galvonometer deflection versus wire area with linear (solid) and quadratic (dashed) fitted equations.

```
> summary(explo.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 184.43569    2.91526   63.27  < 2e-16 ***
wireArea     -0.69537    0.03383  -20.56 6.39e-15 ***

Residual standard error: 7.337 on 20 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9526
F-statistic: 422.6 on 1 and 20 DF,  p-value: 6.386e-15

## Quadratic Model (n=22, p'=3)
> explo.mod2 <- lm(galvDef ~ wireArea + I(wireArea^2))
> summary(explo.mod2)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.966e+02  4.655e+00  42.239  < 2e-16 ***
wireArea      -1.119e+00  1.407e-01  -7.954 1.83e-07 ***
I(wireArea^2)  2.510e-03  8.162e-04   3.075  0.00623 **

Residual standard error: 6.151 on 19 degrees of freedom
Multiple R-squared:  0.9698,    Adjusted R-squared:  0.9667
F-statistic: 305.4 on 2 and 19 DF,  p-value: 3.596e-15
```

2.4.4 Response Surface Models

Response surfaces are often fit when there are 2 or more numeric predictors, and include “linear effects,” “quadratic effects,” and “interaction effects.” The goal is to choose levels of the predictors that optimize (maximize or minimize) the response. In the case of 3 predictors, a full (second order) model would be of the following form.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3$$

Tests are used to remove unnecessary terms, to make the model more parsimonious when possible. Generally multiple runs are made at the “center points” so that a goodness-of-fit test can be conducted (a direct extension of the Lack-of-Fit test in Chapter 1).

Example 2.10: Optimization of 3 Factors Producing Cordyceps Rice Wine

An experiment was conducted (Yang, Gu, and Gu, 2016, [38]) using a response surface design with 3 factors: Liquid-to-solid-ratio (X_1, mL/g), Koji addition (X_2, percent), and Temperature (X_3, degrees C) in cordyceps rice wine formulations. There were two response variables, Cordycepin production (mg/L) and Total acids (g/L). The response considered here is the Cordycepin measurement. There were $n = 17$ runs in a **Box-Behnken design**. The data are given in Table 2.4. The table includes both coded and actual levels for the factors. Begin by fitting the following full second-order response surface, and then consider a reduced model, removing non-significant second order terms (and any main effects not included in the remaining second order terms).

$$\text{Model 1: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2$$

$$\text{Model 2: } E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2$$

RunID	liqSol.C	koji.C	tempC.C	liqSol	koji	tempC	cordycepin	totAcid
1	-1	-1	0	1.5	6	26	44.2	7.8
2	1	-1	0	2.5	6	26	36.31	7.2
3	-1	1	0	1.5	10	26	32.34	8.1
4	-1	1	0	2.5	10	26	32.62	7.5
5	-1	0	-1	1.5	8	24	38.27	7.7
6	1	0	-1	2.5	8	24	34.84	7.1
7	-1	0	1	1.5	8	28	40.99	7.7
8	1	0	1	2.5	8	28	34.22	7.3
9	0	-1	-1	2	6	24	38.42	6.6
10	0	1	-1	2	10	24	32.56	7.4
11	0	-1	1	2	6	28	37.61	7.1
12	0	1	1	2	10	28	31.68	7.4
13	0	0	0	2	8	26	41.55	7.1
14	0	0	0	2	8	26	40.12	7.2
15	0	0	0	2	8	26	42.15	7.3
16	0	0	0	2	8	26	42	7.1
17	0	0	0	2	8	26	39.87	7.2

Table 2.4: Liquid-to-Solid Ratio, Koji percent, Temperature and Output variables - Cordyceps Rice Wine Experiment

$$SSE_1 = 8.9823 \quad df_{E1} = 17 - 10 = 7 \qquad SSE_2 = 11.7724 \quad df_{E2} = 17 - 8 = 9$$

$$H_0 : \beta_{13} = \beta_{23} = 0 \qquad TS : F_{obs} = \frac{\left[\frac{11.7724 - 8.9823}{9-7}\right]}{\left[\frac{8.9823}{7}\right]} = \frac{1.3951}{1.2832} = 1.087 \quad P\left(F_{2,7} \geq 1.087\right) = .3880$$

The R output is given below, as well as results from the **rsm** function in the **rsm** package of R. It gives the values of the input factors that optimize (maximize in this case) the response surface. For this data, the estimated optimal inputs are $X_1^* = 1.188, X_2^* = 6.013, X_3^* = 26.532$. Also included are contour plots of predicted cordyceps output for each pair of predictor variables at the optimal level of the other predictor in Figure 2.8.

The goodness-of-fit test, based on model 1, has an F-statistic of $F_{LOF} = 1.2810$ based on 3 numerator and 4 denominator degrees of freedom has P-value of .3946. There is no evidence that the second order model is inappropriate. Output from "brute force" and **rsm** fits are given below.

∇

```
### Model 1
> crw.mod1 <- lm(cordycepin ~ liqSol + koji + tempC + I(liqSol*koji) +
+     I(liqSol*tempC) + I(koji*tempC) + I(liqSol^2) + I(koji^2) +
+     I(tempC^2))
> summary(crw.mod1)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.777e+02  1.024e+02  -4.667 0.002296 **
liqSol            2.298e+01  1.778e+01   1.293 0.237141
koji              7.886e+00  4.444e+00   1.774 0.119276
tempC             3.656e+01  7.356e+00   4.970 0.001620 **
```

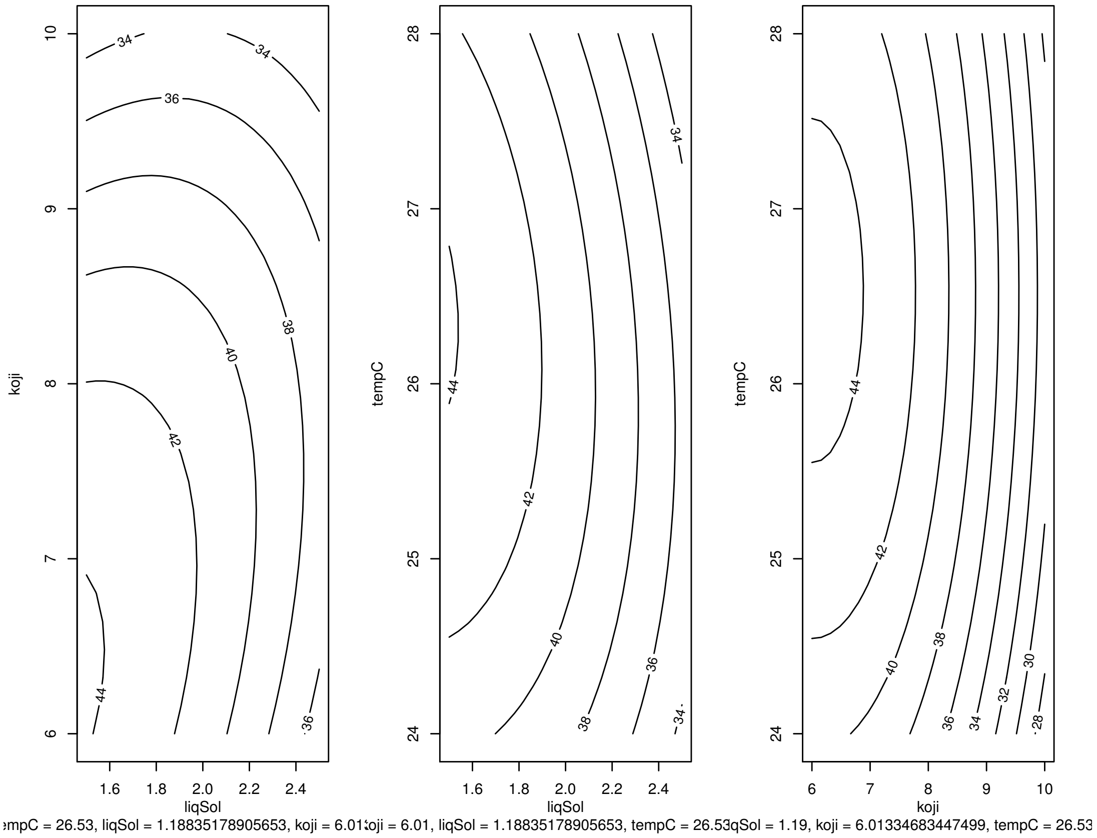

Figure 2.8: Contour plots of predicted Cordycepin production across pairs of predictor variables at the optimum level of the third variable

```
I(liqSol * koji)   2.042e+00  5.664e-01   3.606 0.008668 **
I(liqSol * tempC) -8.350e-01  5.664e-01  -1.474 0.183910
I(koji * tempC)   -4.375e-03  1.416e-01  -0.031 0.976214
I(liqSol^2)       -5.516e+00  2.208e+00  -2.498 0.041114 *
I(koji^2)         -8.479e-01  1.380e-01  -6.143 0.000471 ***
I(tempC^2)        -6.697e-01  1.380e-01  -4.853 0.001850 **

Residual standard error: 1.133 on 7 degrees of freedom
Multiple R-squared:  0.965,     Adjusted R-squared:  0.9201
F-statistic: 21.47 on 9 and 7 DF,  p-value: 0.0002702

### Model 2
> crw.mod2 <- lm(cordycepin ~ liqSol + koji + tempC + I(liqSol*koji) +
+     I(liqSol^2) + I(koji^2) + I(tempC^2))
> summary(crw.mod2)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -433.3523    94.3975  -4.591 0.001308 **
liqSol              1.2715    10.0555   0.126 0.902157
koji                7.7722     2.5139   3.092 0.012893 *
tempC              34.8526     7.2486   4.808 0.000963 ***
I(liqSol * koji)    2.0425     0.5718   3.572 0.006008 **
I(liqSol^2)        -5.5160     2.2295  -2.474 0.035331 *
I(koji^2)          -0.8479     0.1393  -6.085 0.000183 ***
I(tempC^2)         -0.6698     0.1393  -4.807 0.000965 ***

Residual standard error: 1.144 on 9 degrees of freedom
Multiple R-squared:  0.9542,    Adjusted R-squared:  0.9185
F-statistic: 26.78 on 7 and 9 DF,  p-value: 2.484e-05

> anova(crw.mod2, crw.mod1)
Analysis of Variance Table
Model 1: cordycepin ~ liqSol + koji + tempC + I(liqSol * koji) + I(liqSol^2) +
    I(koji^2) + I(tempC^2)
Model 2: cordycepin ~ liqSol + koji + tempC + I(liqSol * koji) + I(liqSol *
    tempC) + I(koji * tempC) + I(liqSol^2) + I(koji^2) + I(tempC^2)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1      9 11.7724
2      7  8.9823  2    2.7901 1.0872  0.388

### rsm model fit
> crw.rsm1 <- rsm(cordycepin ~ SO(liqSol,koji,tempC))
> summary(crw.rsm1)
                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  -477.682250  102.354283 -4.6669 0.0022957 **
liqSol         22.981500   17.777762  1.2927 0.2371413
koji            7.886000    4.444440  1.7744 0.1192756
tempC          36.557625    7.355975  4.9698 0.0016196 **
liqSol:koji     2.042500    0.566389  3.6062 0.0086682 **
liqSol:tempC   -0.835000    0.566389 -1.4743 0.1839101
koji:tempC     -0.004375    0.141597 -0.0309 0.9762138
liqSol^2       -5.516000    2.208191 -2.4980 0.0411140 *
koji^2         -0.847875    0.138012 -6.1435 0.0004705 ***
tempC^2        -0.669750    0.138012 -4.8528 0.0018503 **

Multiple R-squared:  0.965,     Adjusted R-squared:  0.9201
F-statistic: 21.47 on 9 and 7 DF,  p-value: 0.0002702

Analysis of Variance Table
Response: cordycepin
                         Df  Sum Sq Mean Sq F value    Pr(>F)
FO(liqSol, koji, tempC)   3 133.105  44.368 34.5767 0.0001442
TWI(liqSol, koji, tempC)  3  19.477   6.492  5.0596 0.0356962
PQ(liqSol, koji, tempC)   3  95.381  31.794 24.7772 0.0004206
```

```
Residuals               7   8.982   1.283
Lack of fit             3   4.401   1.467  1.2810 0.3946206
Pure error              4   4.581   1.145

Stationary point of response surface:
   liqSol      koji     tempC
 1.188352  6.013347 26.531574
```

2.5 Model Building

When there are many predictors, algorithms can be used to determine which variables to include in the model. These variables can be main effects, interactions, and polynomial terms. Note that there are two common approaches. One method involves testing variables based on t-tests, or equivalently F-tests for partial regression coefficients. An alternative method involves comparing models based on model based measures, such as Akaike Information Criterion (AIC), or Schwartz Bayesian Information criterion (BIC or SBC). These measures can be written as follows (note that different software packages print different versions, as some parts are constant for all potential models). The goal is to minimize the measures.

$$AIC(\text{Model}) = n\ln(SSE(\text{Model})) + 2p' - n\ln(n) \qquad BIC(\text{Model}) = n\ln(SSE(\text{Model})) + [\ln(n)]p' - n\ln(n)$$

Note that $SSE(\text{Model})$ depends on the variables included in the current model. The measures put a penalty on excess predictor variables, with BIC placing a higher penalty when $\ln(n) > 2$. Note that p' is the number of parameters in the model (including the intercept), and n is the sample size.

2.5.1 Backward Elimination

This is a "top-down" method, which begins with a "Complete" Model, with all potential predictors. The analyst then chooses a significance level to stay in the model (SLS). The model is fit, and the predictor with the lowest t-statistic in absolute value (largest P-value) is identified. If the P-value is larger than SLS, the variable is dropped from the model. Then the model is re-fit with all other predictors (this will change all regression coefficients, standard errors, and P-values). The process continues until all variables have P-values below SLS.

The model based approach fits the full model, with all predictors and computes AIC (or BIC). Then, each variable is dropped one-at-a-time, and AIC (or BIC) is obtained for each model. If none of the models with one dropped variable has AIC (or BIC) below that for the full model, the full model is kept, otherwise the model with the lowest AIC (or BIC) is kept as the new full model. The process continues until no variables should be dropped (none of the "drop one variable models" has a lower AIC (or BIC) than the "full model").

2.5.2 Forward Selection

This is a "bottom-up" approach, which begins with all "Simple" Models, each with one predictor. The analyst then chooses a significance level to enter into the model (SLE). Each model is fit, and the predictor with the highest t-statistic in absolute value (smallest P-value) is identified. If the P-value is smaller than SLE, the variable is entered into the model. Then all two variable models including the best predictor in the first round are fit. The best second variable is identified, and its P-value is compared with SLE. If its P-value is below SLE, the variable is added to the model. The process continues until no potential added variables have P-values below SLE.

The model based approach fits each simple model, with one predictor and computes AIC (or BIC). The best variable is identified (assuming its AIC (or BIC) is smaller than that for the null model, with no predictors). Then, each potential variable is added one-at-a-time, and AIC (or BIC) is obtained for each model. If none of the models with one added variable has AIC (or BIC) below that for the best simple model, the simple model is kept, otherwise the model with the lowest AIC (or BIC) is kept as the new full model. The process continues until no variables should be added (none of the "add one variable models" has a lower AIC (or BIC) than the "reduced model").

2.5.3 Stepwise Regression

This approach is a hybrid of forward selection and backward elimination. It begins like forward selection, but then applies backward elimination at each step. In forward selection, once a variable is entered, it stays in the model. In stepwise regression, once a new variable is entered, all previously entered variables are tested, to confirm they should stay in the model, after controlling for the new entrant, as well as the other previous entrants.

Example 2.11: LPGA Predictors of Prize Winnings

The LPGA 2009 data with Y as the log of prize winnings per tournament is used as an example of model building. The **stepAIC** function in the **MASS** library is used. It will fit Backward Elimination, Forward Selection and Stepwise Regression. The output for the three methods is given below. Note that fit1 is the full model with all predictors and fit2 is the intercept only model. In some confirmatory studies, the reduced model may contain a set of predictors that are "forced" to be in a model. The output for Stepwise Regression is not included, as it is identical to Forward Selection (no variables are ever removed after entering the model). All methods fit the same model with predictors: drive distance, fairway percent, percentile, and strokes per round. These methods will not always be in agreement.

∇

```
> library(MASS)
> fit1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
> fit2 <- lm(Y ~ 1)
### Backward Elimination
> stepAIC(fit1,direction="backward")
Start:  AIC=-215.29
Y ~ drive + frwy + grnReg + putts + sandSave + pctile + strksRnd
           Df Sum of Sq    RSS     AIC
```

```
- sandSave  1   0.00401 29.952 -217.27
- grnReg    1   0.02595 29.974 -217.16
- putts     1   0.13005 30.078 -216.65
<none>                  29.948 -215.29
- drive     1   1.03661 30.984 -212.32
- pctile    1   2.15756 32.105 -207.13
- frwy      1   2.44707 32.395 -205.82
- strksRnd  1   2.65889 32.606 -204.87
Step:  AIC=-217.27
Y ~ drive + frwy + grnReg + putts + pctile + strksRnd
           Df Sum of Sq    RSS     AIC
- grnReg    1   0.02274 29.974 -219.16
- putts     1   0.12628 30.078 -218.65
<none>                  29.952 -217.27
- drive     1   1.06354 31.015 -214.17
- pctile    1   2.16280 32.114 -209.09
- frwy      1   2.47367 32.425 -207.68
- strksRnd  1   2.75837 32.710 -206.41
Step:  AIC=-219.16
Y ~ drive + frwy + putts + pctile + strksRnd
           Df Sum of Sq    RSS     AIC
- putts     1    0.1963 30.171 -220.20
<none>                  29.974 -219.16
- drive     1    1.0765 31.051 -216.00
- pctile    1    2.1513 32.126 -211.04
- frwy      1    2.5910 32.565 -209.05
- strksRnd  1    4.1502 34.124 -202.22
Step:  AIC=-220.2
Y ~ drive + frwy + pctile + strksRnd
           Df Sum of Sq    RSS     AIC
<none>                  30.171 -220.20
- drive     1    0.8860 31.057 -217.98
- pctile    1    2.4198 32.590 -210.94
- frwy      1    2.4281 32.599 -210.90
- strksRnd  1    4.0546 34.225 -203.79

Call:
lm(formula = Y ~ drive + frwy + pctile + strksRnd)
Coefficients:
(Intercept)        drive         frwy       pctile     strksRnd
   57.10869     -0.01221     -0.02893      0.03523     -0.61958

### Forward Selection
> stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
Start:  AIC=63.94
Y ~ 1
           Df Sum of Sq     RSS      AIC
+ pctile    1   187.706  35.441 -202.699
+ strksRnd  1   187.628  35.518 -202.380
+ grnReg    1   121.166 101.980  -48.389
+ drive     1    42.049 181.098   35.453
+ putts     1    31.774 191.373   43.510
+ sandSave  1    10.485 212.661   58.910
+ frwy      1     3.147 219.999   63.862
<none>                  223.146   63.936
Step:  AIC=-202.7
Y ~ pctile
           Df Sum of Sq    RSS     AIC
+ strksRnd  1   2.83652 32.604 -212.88
+ frwy      1   0.87845 34.562 -204.36
<none>                  35.441 -202.70
+ putts     1   0.33926 35.101 -202.10
+ sandSave  1   0.20866 35.232 -201.56
+ grnReg    1   0.00080 35.440 -200.70
+ drive     1   0.00063 35.440 -200.70
Step:  AIC=-212.88
Y ~ pctile + strksRnd
           Df Sum of Sq    RSS     AIC
+ frwy      1   1.54752 31.057 -217.98
<none>                  32.604 -212.88
+ grnReg    1   0.11883 32.485 -211.41
+ sandSave  1   0.08992 32.514 -211.28
+ putts     1   0.03870 32.565 -211.05
+ drive     1   0.00543 32.599 -210.90
Step:  AIC=-217.98
Y ~ pctile + strksRnd + frwy
           Df Sum of Sq    RSS     AIC
+ drive     1   0.88596 30.171 -220.20
<none>                  31.057 -217.98
+ sandSave  1   0.05108 31.006 -216.22
+ grnReg    1   0.02351 31.033 -216.09
+ putts     1   0.00573 31.051 -216.00
Step:  AIC=-220.2
Y ~ pctile + strksRnd + frwy + drive
           Df Sum of Sq    RSS     AIC
<none>                  30.171 -220.20
+ putts     1  0.196276 29.974 -219.16
+ grnReg    1  0.092734 30.078 -218.65
+ sandSave  1  0.006446 30.164 -218.24
Call:
lm(formula = Y ~ pctile + strksRnd + frwy + drive)
Coefficients:
(Intercept)       pctile     strksRnd         frwy        drive
   57.10869      0.03523     -0.61958     -0.02893     -0.01221
```

2.5.4 All Possible Regressions

All possible regression models can be fit, and model based measures used to choose the "best" model. Commonly used measures are: Adjusted-R^2 (equivalently MSE), Mallow's C_p statistic, AIC, and BIC. The formulas, and decision criteria are given below (where p' is the number of parameters in the "current" model being fit.

R^2-Adjusted - $1-\left(\frac{n-1}{n-p'}\right)\frac{SSE}{TSS}$ - Goal is to maximize

C_p - $C_p = \frac{SSE(\text{Model})}{MSE(\text{Complete})} + 2p' - n$ - Goal is to have $C_p \leq p'$

AIC - $AIC(\text{Model}) = n\ln(SSE(\text{Model})) + 2p' - n\ln(n)$ - Goal is to minimize

BIC - $BIC(\text{Model}) = n\ln(SSE(\text{Model})) + [\ln(n)]p' - n\ln(n)$ - Goal is to minimize

Example 2.12: LPGA Predictors of Prize Winnings

All possible regressions are run on the LPGA 2009 data using the **regsubsets** function in the **leaps** package. Note that a data frame must be given that contains Y and all possible predictors. AIC is not computed in the function, but can be computed from BIC. The following output prints the quantities for the best 4 models with each number of predictor variables. Based on BIC, which places a higher penalty on extra predictor variables, the 3 variable model with fairway percent, percentile, and strokes per round is selected. Based on C_p and AIC, the 4 variable model which also includes drive distance is selected. Note that BIC is very close for the two models.

∇

```
> with(aprout,round(cbind(which,rsq,adjr2,cp,bic,aic),3))     ## Prints "readable" results
  (Intercept) drive frwy grnReg putts sandSave pctile strksRnd   rsq adjr2      cp      bic      aic
1           1     0    0      0     0        0      1        0 0.841 0.840  21.312 -258.668 -264.636
1           1     0    0      0     0        0      0        1 0.841 0.840  21.669 -258.349 -264.317
1           1     0    0      1     0        0      0        0 0.543 0.540 327.930 -104.358 -110.325
1           1     1    0      0     0        0      0        0 0.188 0.183 692.507  -20.516  -26.484
2           1     0    0      0     0        0      1        1 0.854 0.852  10.241 -265.864 -274.815
2           1     0    1      0     0        0      0        1 0.850 0.848  14.155 -262.109 -271.060
2           1     0    1      0     0        0      1        0 0.845 0.843  19.264 -257.349 -266.300
2           1     0    0      0     1        0      1        0 0.843 0.840  21.749 -255.089 -264.040
3           1     0    1      0     0        0      1        1 0.861 0.858   5.110 -267.980 -279.915
3           1     0    0      1     0        0      1        1 0.854 0.851  11.694 -261.414 -273.348
3           1     0    0      0     0        1      1        1 0.854 0.851  11.827 -261.284 -273.218
3           1     0    0      0     1        0      1        1 0.854 0.851  12.063 -261.054 -272.988
4           1     1    1      0     0        0      1        1 0.865 0.861   3.028 -267.222 -282.140
4           1     0    1      0     0        1      1        1 0.861 0.857   6.875 -263.237 -278.155
4           1     0    1      1     0        0      1        1 0.861 0.857   7.002 -263.107 -278.025
4           1     0    1      0     1        0      1        1 0.861 0.857   7.084 -263.023 -277.942
5           1     1    1      0     1        0      1        1 0.866 0.861   4.123 -263.191 -281.093
5           1     1    1      1     0        0      1        1 0.865 0.860   4.600 -262.688 -280.590
5           1     1    1      0     0        1      1        1 0.865 0.860   4.998 -262.270 -280.171
5           1     0    1      1     0        1      1        1 0.861 0.856   8.842 -258.287 -276.188
6           1     1    1      1     1        0      1        1 0.866 0.860   6.018 -258.319 -279.204
6           1     1    1      0     1        1      1        1 0.866 0.860   6.120 -258.212 -279.097
6           1     1    1      1     0        1      1        1 0.865 0.859   6.599 -257.706 -278.591
6           1     0    1      1     1        1      1        1 0.861 0.855  10.777 -253.370 -274.255
7           1     1    1      1     1        1      1        1 0.866 0.859   8.000 -253.355 -277.223
```

2.5.5 Cross-Validation

Regression models tend to "over-fit" to the current dataset and may not apply as well to data not used to fit the model. **Cross-Validation** is used to fit the model based on one set of observations, referred to as the "training sample," and is then used to predict outcomes for an external or "validation sample." In k-fold cross-validation, the full set of data is randomly split into k subsamples. The model is fit k times, with each subsample being "left out" and the prediction error is obtained for each possible model. Models can be compared by their mean square prediction error for the "left out" observations. Note that different splits of the data into the k folds will give different results.

Example 2.13: LPGA Predictors of Prize Winnings

The all possible regressions suggested two models: one based on BIC with 3 predictors and one based on AIC and C_p which added a 4th predictor. Making use of **CVlm** function in the R package **DAAG**, a 10-fold cross-validation is run. The output is long, containing the results for observations within all 10 folds for each model. A subset of the output is given below. The 3 variable model has a slightly smaller mean square prediction error (0.224) than the 4 variable model (0.228), suggesting the more parsimonious 3 variable model is better.

∇

```
>
> lpgacv.mod1 <-  lm(Y ~ drive+frwy+pctile+strksRnd,
+      data=lpga2)
> CVlm(lpga2, lpgacv.mod1, m=10)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
drive       1   42.0    42.0   196.5 < 2e-16 ***
frwy        1   20.9    20.9    97.8 < 2e-16 ***
pctile      1  126.0   126.0   588.6 < 2e-16 ***
strksRnd    1    4.1     4.1    18.9 2.6e-05 ***
Residuals 141   30.2     0.2

fold 1
Observations in test set: 14
               49     56     59    64     65    102     103    120     123    125    136    138   140    145
Predicted   7.905  9.358  8.480 8.813  8.927  7.756 10.8063  9.088 11.1040  8.944 10.181 11.165  7.58  7.046
cvpred      7.986  9.399  8.529 8.863  9.053  7.763 10.8207  9.102 11.1158  8.935 10.237 11.232  7.67  7.057
Y           8.315  9.008  8.476 9.270  8.583  7.621 10.8451  8.605 11.1885  8.775  9.700 10.786  6.43  6.532
CV residual 0.328 -0.391 -0.053 0.407 -0.469 -0.142  0.0244 -0.497  0.0726 -0.159 -0.537 -0.446 -1.24 -0.525
Sum of squares = 3.24    Mean square = 0.23    n = 14
...
fold 10
Observations in test set: 14
               16      20     23    27     30    33     61     82     86    89     91   108   111     116
Predicted   8.667 10.8948 8.5314 8.302  7.419 6.948  7.288  9.350 8.2780 9.019  9.267 9.520 9.212 10.2031
cvpred      8.679 10.8937 8.5469 8.279  7.421 6.945  7.299  9.366 8.3134 9.006  9.250 9.505 9.208 10.2197
Y           9.007 10.9124 8.5991 8.745  6.725 7.392  7.023  8.747 8.3356 9.263  9.082 9.834 9.375 10.1759
CV residual 0.328  0.0187 0.0522 0.466 -0.696 0.447 -0.276 -0.619 0.0222 0.258 -0.169 0.329 0.168 -0.0438
Sum of squares = 1.71    Mean square = 0.12    n = 14

Overall (Sum over all 14 folds)
   ms
0.228

> lpgacv.mod2 <-  lm(Y ~ frwy+pctile+strksRnd,
+      data=lpga2)
> CVlm(lpga2, lpgacv.mod2, m=10)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
frwy        1    3.1     3.1    14.4 0.00022 ***
pctile      1  185.4   185.4   847.9 < 2e-16 ***
strksRnd    1    3.5     3.5    16.0 0.00010 ***
Residuals 142   31.1     0.2

fold 1
Observations in test set: 14
              49    56      59    64     65    102   103   120    123    125    136    138   140    145
Predicted   7.91  9.18  8.4768 8.732  8.910  7.729 10.71  9.14 11.049  9.089 10.091 11.162  7.47  7.080
cvpred      7.98  9.19  8.5203 8.764  9.020  7.732 10.70  9.16 11.043  9.102 10.123 11.216  7.54  7.096
Y           8.31  9.01  8.4762 9.270  8.583  7.621 10.85  8.61 11.188  8.775  9.700 10.786  6.43  6.532
CV residual 0.33 -0.18 -0.0441 0.506 -0.436 -0.111  0.14 -0.55  0.145 -0.326 -0.423 -0.429 -1.11 -0.565
Sum of squares = 2.97    Mean square = 0.21    n = 14
...
fold 10
Observations in test set: 14
               16      20    23    27     30    33     61     82    86    89     91   108   111     116
Predicted   8.596 10.8684 8.392 8.399  7.415 6.967  7.213  9.228 8.080 9.069  9.342 9.592 9.202 10.0855
```

```
cvpred      8.591 10.8675 8.371 8.386  7.410 6.959  7.204  9.215 8.081 9.062  9.337 9.585 9.188 10.0836
Y           9.007 10.9124 8.599 8.745  6.725 7.392  7.023  8.747 8.336 9.263  9.082 9.834 9.375 10.1759
CV residual 0.415  0.0449 0.228 0.358 -0.686 0.433 -0.182 -0.469 0.254 0.202 -0.255 0.249 0.187  0.0924
Sum of squares = 1.54    Mean square = 0.11    n = 14

Overall (Sum over all 14 folds)
   ms
0.224
```

2.6 Issues of Collinearity

When the predictor variables are highly correlated among themselves, the regression coefficients become unstable, with increased standard errors. This leads to smaller t-statistics for tests regarding the partial regression coefficients and wider confidence intervals. At its most extreme case, the sign of a regression coefficient can change when a new predictor variable is included. One widely reported measure of collinearity is the **Variance Inflation Factor (VIF)**. This is computed for each predictor variable, by regressing it on the remaining $p-1$ predictors. Then $VIF_J = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficent of determination of the regression of X_j on the remaining predictors. Values of VIF_j greater than 10 are considered problematic. Collinearity is not problematic when the primary goal of the model is for prediction.

Various remedies exist. One is determining which variable(s) make the most sense theoretically for the model, and removing other variables, which are correlated with the other more meaningful predictors. A second method involves generating uncorrelated predictor variables from the original set of predictors. While this method based on **principal components** removes the collinearity problem, the new variables may lose their meaning, thus making it harder to describe the process. A third method, **ridge regression**, introduces a bias factor into the regression that reduces the inflated variance due to collinearity, and through that reduces the Mean Square Error of the regression coefficients. Unfortunately, there is no simple rule on choosing the bias factor.

Example 2.14: LPGA Predictors of Prize Winnings

For the full LPGA 2009 data (with all 7 predictors), the **vif** function in the **DAAG** package is used to obtain the variance inflation factors among the 7 predictors. It is seen that there is high collinearity among the predictors with highest VIF's for strksRnd (37.0), pctile (17.5), and greenReg (15.4).

∇

```
> Y <- log(prize/tourneys)
> lpga.mod1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
> lpga.vif <- vif(lpga.mod1)
> lpga.vif
   drive     frwy   grnReg    putts sandSave   pctile strksRnd
  2.6402   1.9206  15.3700   6.4040   1.2559  17.5170  37.0050
>
```

2.7 R Programs for Chapter 2 Examples

2.7.1 Scottish Recycling Study

```
### Read in Scottish Recycling Data
sr1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/scottish_recycle.dat",
      header=F, width=c(25,8,8,8,8), col.names=c("locAuth","recCap",
      "resCap","extMat","yldExtMat"))
attach(sr1)

##### Example 2.1
plot(sr1[,2:5])

##### Example 2.2-2.4
recycle.mod1 <- lm(yldExtMat ~ recCap + resCap + extMat)
summary(recycle.mod1)
confint(recycle.mod1)
anova(recycle.mod1)
drop1(recycle.mod1,test="F")

e1 <- resid(recycle.mod1)
yhat1 <- predict(recycle.mod1)
plot(e1 ~ yhat1)
abline(h=0)
shapiro.test(e1)
```

2.7.2 LPGA 2009 Data

```
lpga1 <- read.table("http://www.stat.ufl.edu/~winner/data/lpga2009.dat",
  header=F, col.names=c("glfrID","drive","frwy","grnReg",
  "putts","sandSave","prize","logPrize","tourneys","grpph",
  "compTrn","pctile","rounds","strksRnd"))
attach(lpga1)

##### Example 2.5
Y <- log(prize/tourneys)
lpga2 <- data.frame(Y,drive,frwy,grnReg,putts,sandSave,pctile,strksRnd)

# install.packages("GGally")
GGally::ggpairs(lpga2)

lpga.mod1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
summary(lpga.mod1)
anova(lpga.mod1)

lpga.mod2 <- lm(Y ~ drive+frwy+pctile+strksRnd)
summary(lpga.mod2)
anova(lpga.mod2)

anova(lpga.mod2, lpga.mod1)     ## Complete vs Reduced Model Comparison

##### Example 2.11
######### Perform Backward Elimination, Forward Selection, and Stepwise Regression
######### Based on Model AIC (not individual regression coefficients)
######### fit1 and fit2 represent "extreme" models

library(MASS)
```

```
fit1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
fit2 <- lm(Y ~ 1)
stepAIC(fit1,direction="backward")
stepAIC(fit2,direction="forward",scope=list(upper=fit1,lower=fit2))
stepAIC(fit2,direction="both",scope=list(upper=fit1,lower=fit2))

##### Example 2.12
########## Perform all possible regressions (aka all subset regressions)
########## Prints out best 4 models of each # of predictors
install.packages("leaps")
library(leaps)

all_lpga <- regsubsets(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd,
 nbest=4,data=lpga2)
aprout <- summary(all_lpga)
n <- length(lpga2$Y)
p <- apply(aprout$which, 1, sum)   ### p includes intercept
aprout$aic <- aprout$bic - log(n) * p + 2 * p     ### Compute AIC from BIC
with(aprout,round(cbind(which,rsq,adjr2,cp,bic,aic),3))     ## Prints "readable" results

##### Example 2.13
#install.packages("DAAG")
library(DAAG)
set.seed(12345)

lpgacv.mod1 <-  lm(Y ~ drive+frwy+pctile+strksRnd,
     data=lpga2)
CVlm(lpga2, lpgacv.mod1, m=10)

lpgacv.mod2 <-  lm(Y ~ frwy+pctile+strksRnd,
     data=lpga2)
CVlm(lpga2, lpgacv.mod2, m=10)

##### Example 2.14
library(DAAG)
lpga.mod1 <- lm(Y ~ drive+frwy+grnReg+putts+sandSave+pctile+strksRnd)
lpga.vif <- vif(lpga.mod1)
lpga.vif
```

2.7.3 Subsidence and Water Table Study

```
## Read in water resource data
wr1 <- read.csv("http://www.stat.ufl.edu/~winner/data/water_resource.csv")
attach(wr1); names(wr1)

crop.f <- factor(crop, labels=c("pasture","truck crop","sugarcane"))

##### Example 2.6
library(lattice)
xyplot(subsidence ~ waterTbl | crop.f, layout=c(3,1))

wr.mod1 <- lm(subsidence ~ waterTbl + pasture + truckCrop)
summary(wr.mod1)
anova(wr.mod1)

wr.mod2 <- lm(subsidence ~ waterTbl)
summary(wr.mod2)
anova(wr.mod2)

anova(wr.mod2, wr.mod1)
```

```
wt.seq <- seq(30,90,.01)
yhat_p <- coef(wr.mod1)[1] + coef(wr.mod1)[2]*wt.seq + coef(wr.mod1)[3]
yhat_t <- coef(wr.mod1)[1] + coef(wr.mod1)[2]*wt.seq + coef(wr.mod1)[4]
yhat_s <- coef(wr.mod1)[1] + coef(wr.mod1)[2]*wt.seq

plot(subsidence ~ waterTbl, pch=crop, xlim=c(30,90),ylim=c(0,7))
lines(wt.seq,yhat_p, lty=1)
lines(wt.seq,yhat_t, lty=2)
lines(wt.seq,yhat_s, lty=5)
legend(35,6.5,c("Pasture","Truck Crop","Sugarcane"),pch=c(1,2,3),lty=c(1,2,5))

##### Example 2.7
wr.mod3 <- lm(subsidence ~ waterTbl + pasture + truckCrop +
       I(waterTbl*pasture) + I(waterTbl*truckCrop))
summary(wr.mod3)
anova(wr.mod3)

anova(wr.mod1, wr.mod3)

wt.seq <- seq(30,90,.01)
yhat_p <- coef(wr.mod3)[1] + coef(wr.mod3)[2]*wt.seq + coef(wr.mod3)[3] +
       coef(wr.mod3)[5]*wt.seq
yhat_t <- coef(wr.mod3)[1] + coef(wr.mod3)[2]*wt.seq + coef(wr.mod3)[4] +
       coef(wr.mod3)[6]*wt.seq
yhat_s <- coef(wr.mod3)[1] + coef(wr.mod3)[2]*wt.seq

plot(subsidence ~ waterTbl, pch=crop, xlim=c(30,90),ylim=c(0,7))
lines(wt.seq,yhat_p, lty=1)
lines(wt.seq,yhat_t, lty=2)
lines(wt.seq,yhat_s, lty=5)
legend(35,6.5,c("Pasture","Truck Crop","Sugarcane"),pch=c(1,2,3),lty=c(1,2,5))

p.tC <- pasture + truckCrop

wr.mod4 <- lm(subsidence ~ waterTbl + p.tC + I(waterTbl*p.tC))
summary(wr.mod4)
anova(wr.mod4)

anova(wr.mod4, wr.mod3)
```

2.7.4 Explosives Experiment

```
## Read in explosives data
explosives <- read.table("http://www.stat.ufl.edu/~winner/data/explosives1.dat",
       header=F, col.names=c("coupling", "wireArea", "galvDef"))
attach(explosives)

##### Example 2.7
## Fit linear model
explo.mod1 <- lm(galvDef ~ wireArea)
summary(explo.mod1)
## Fit quadratic model
explo.mod2 <- lm(galvDef ~ wireArea + I(wireArea^2))
summary(explo.mod2)

x.seq <- seq(0,160,0.1)
yhat2 <- coef(explo.mod2)[1] + coef(explo.mod2)[2]*x.seq +
         coef(explo.mod2)[3]*x.seq^2
plot(galvDef ~ wireArea, xlim=c(0,160))
abline(explo.mod1)
```

```
lines(x.seq,yhat2, lty=3)

## Fit cubic model
explo.mod3 <- lm(galvDef ~ wireArea + I(wireArea^2) + I(wireArea^3))
summary(explo.mod3)
```

2.7.5 Cordyceps Rice Wine Experiment

```
crw1 <- read.csv("http://www.stat.ufl.edu/~winner/data/cordyceps_ricewine.csv")
attach(crw1); names(crw1)

##### Example 2.10
crw.mod1 <- lm(cordycepin ~ liqSol + koji + tempC + I(liqSol*koji) +
    I(liqSol*tempC) + I(koji*tempC) + I(liqSol^2) + I(koji^2) +
    I(tempC^2))
summary(crw.mod1)

crw.mod2 <- lm(cordycepin ~ liqSol + koji + tempC + I(liqSol*koji) +
    I(liqSol^2) + I(koji^2) + I(tempC^2))
summary(crw.mod2)
anova(crw.mod2, crw.mod1)

# install.packages("rsm")
library(rsm)
crw.rsm1 <- rsm(cordycepin ~ SO(liqSol,koji,tempC))
summary(crw.rsm1)
par(mfrow=c(1,3))
contour(crw.rsm1, ~ liqSol + koji + tempC,
    at=summary(crw.rsm1)$canonical$xs)
```

Chapter 3

Factorial Designs

In STA 6166, methods to compare a set of treatments were covered for the Completely Randomized Design (independent samples) and the Randomized Block Design (paired/matched samples). Many times there are more than one set of treatments that are to be compared simultaneously. For instance, drug trials are generally run at different medical centers. In this case, the drug a subject receives would be factor A (active or placebo), while the center he/she is located at would be factor B. Then tests for drug effects and center effects can be conducted.

An interaction would exist if the drug effects differ among centers. That is an undesirable situation, but one that should be tested for. If interest is to measure the interaction there will have to be more than one measurement (replicate) corresponding to each combination of levels of the 2 factors. In this situation, that would mean having multiple subjects receiving each treatment at each center.

Models can have two or more factors. Factors can be **fixed** or **random**. Fixed factors have all levels of interest for the factor included in the experiment or observational study. The effects of the fixed factor levels are unknown parameters to be estimated. Random factors have a sample of levels from a larger population included, their effects are unobservable random variables. For random factors, inference is typically in the variance of the effects. In the drug example given above, the drug would be considered a fixed factor, while the medical center would be a random factor.

Factors can be **crossed** or **nested**. Crossed factors have treatments that are the combinations of the levels of the individual factors. Nested factors have a hierarchy, with levels of one factor being different than those within different levels of another factor. In the drug example, if within each medical center, both the drug and placebo were given in a random manner, with individual patients receiving only drug or placebo, the experiment would be crossed. If, on the other hand, centers were randomly assigned to either drug or placebo, and patients within a center received whichever the center had been assigned to, the experiment would be nested. This second scenario would be a poor design, it is only given as an example of a nested design.

3.1 Multiple-Factor Analysis of Variance - Crossed Factors

In this section, models with two or more factors are considered. The description begins with 2-factor models. Denoting the k^{th} measurement (replicate) observed under the i^{th} level of factor A and the j^{th} level of factor B, the model is written as follows.

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, n \quad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Here μ is the overall mean, α_i is the effect of the i^{th} level of factor A, β_j is the effect of the j^{th} level of factor B, $(\alpha\beta)_{ij}$ is the effect of the interaction of the i^{th} level of factor A and the j^{th} level of factor B, and ϵ_{ijk} is the random error term representing the fact that units within each treatment combination will vary. Here,the model where both factors A and B are **fixed**, with all levels of interest present in the study is considered. As before, the assumption is that ϵ_{ijk} is normally distributed with mean 0 and variance σ^2.

When factors A and B are fixed, the effects are unknown parameters to be estimated. One common way of parameterizing the model is as follows.

$$E\{Y_{ijk}\} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \qquad V\{Y_{ijk}\} = \sigma^2 \qquad \sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \forall j, i$$

Some interesting hypotheses to test are as follow.

1. $H_0 : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$ (No interaction effect).
2. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No effects among the levels of factor A)
3. $H_0 : \beta_1 = \cdots = \beta_b = 0$ (No effects among the levels of factor B)

The total variation in the set of observed measurements can be decomposed into four parts: variation in the means of the levels of factor A, variation in the means of the levels of factor B, variation due to the interaction of factors A and B, and error variation. The formulas for the means and sums of squares are given here.

$$\begin{aligned}
\overline{y}_{ij.} &= \frac{\sum_{k=1}^{n} y_{ijk}}{n} \\
\overline{y}_{i..} &= \frac{\sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}}{bn} \\
\overline{y}_{.j.} &= \frac{\sum_{i=1}^{a} \sum_{k=1}^{n} y_{ijk}}{an} \\
\overline{y}_{...} &= \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}}{abn} \\
n_{..} &= abn \\
s_{ij}^2 &= \frac{\sum_{k=1}^{n} \left(y_{ijk} - \overline{y}_{ij.}\right)^2}{n-1} \\
TSS &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} (y_{ijk} - \overline{y}_{...})^2
\end{aligned}$$

$$\begin{aligned}
SSA &= bn\sum_{i=1}^{a}(\overline{y}_{i..}-\overline{y}_{...})^2 \\
SSB &= an\sum_{j=1}^{b}(\overline{y}_{.j.}-\overline{y}_{...})^2 \\
SSAB &= n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...})^2 \\
SSE &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\overline{y}_{ij.})^2
\end{aligned}$$

The error sum of squares can also be computed from the within cell standard deviations, which is helpful as many research articles provide the treatment means and standard deviations.

$$SSE=\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\overline{y}_{ij.})^2=(n-1)\sum_{i=1}^{a}\sum_{j=1}^{b}s_{ij}^2$$

Note that this type of analysis is almost always done on a computer (either a statistical package or spreadsheet). The analysis of variance can be set up as shown in Table 3.1, assuming that n measurements are made at each combination of levels of the two factors.

ANOVA

Source of	Degrees of	Sum of	Mean		
Variation	Freedom	Squares	Square	F	$Pr(>F)$
A	$a-1$	SSA	$MSA=\frac{SSA}{a-1}$	$F_A=\frac{MSA}{MSE}$	$P\left(F_{a-1,ab(n-1)}\geq F_A\right)$
B	$b-1$	SSB	$MSB=\frac{SSB}{b-1}$	$F_B=\frac{MSB}{MSE}$	$P\left(F_{b-1,ab(n-1)}\geq F_B\right)$
AB	$(a-1)(b-1)$	$SSAB$	$MSAB=\frac{SSAB}{(a-1)(b-1)}$	$F_{AB}=\frac{MSAB}{MSE}$	$P\left(F_{(a-1)(b-1),ab(n-1)}\geq F_{AB}\right)$
ERROR	$ab(n-1)$	SSE	$MSE=\frac{SSE}{ab(n-1)}$		
TOTAL	$abn-1$	TSS			

Table 3.1: The Analysis of Variance Table for a Balanced 2-Factor Factorial Design with Fixed Effects

The expectations of the mean squares for the fixed effects model are given below.

$$E\{MSA\}=\sigma^2+\frac{bn\sum_{i=1}^{a}\alpha_i^2}{a-1} \qquad E\{MSB\}=\sigma^2+\frac{an\sum_{j=1}^{b}\beta_j^2}{b-1}$$

$$E\{MSAB\}=\sigma^2+\frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)} \qquad E\{MSE\}=\sigma^2$$

The tests for interactions and the main effects of factors A and B involve the three F–statistics, and can be conducted as follow. Note that under each of the three null hypotheses, the corresponding expected mean square in the numerator simplifies to $\sigma^2 = E\{MSE\}$.

1. $H_0^{AB}: (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{ab} = 0$ (No interaction effect).
2. H_A^{AB} : Not all $(\alpha\beta)_{ij} = 0$ (Interaction effects exist)
3. T.S. $F_{AB} = \frac{MSAB}{MSE}$
4. R.R.: $F_{AB} \geq F_{\alpha,(a-1)(b-1),ab(n-1)}$
5. P-value: $P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$

Assuming no interaction effects exist, the test for differences among the effects of the levels of factor A as follows.

1. $H_0^{A}: \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A^{A} : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSE}$
4. R.R.: $F_A \geq F_{\alpha,(a-1),ab(n-1)}$
5. P-value: $P\left(F_{a-1,ab(n-1)} \geq F_A\right)$

Assuming no interaction effects exist, the test for differences among the effects of the levels of factor B as follows.

1. $H_0^{B}: \beta_1 = \cdots = \beta_b = 0$ (No factor B effect).
2. H_A^{B} : Not all $\beta_j = 0$ (Factor B effects exist)
3. T.S. $F_B = \frac{MSB}{MSE}$
4. R.R.: $F_B \geq F_{\alpha,(b-1),ab(n-1)}$
5. P-value: $P\left(F_{(b-1),ab(n-1)} \geq F_{obs}\right)$

Note that if interaction effects exist, comparisons are made among the ab individual combinations of factors A and B separately (as in the Completely Randomized Design), and don't interpret the tests for main effects among levels of factors A and B.

Example 3.1: Halo Effect - Essay Evaluation

A study was conducted to observe evidence of the "halo effect," the fact that people tend to judge items in one dimension, given they have evidence of quality in other (irrelevant) dimensions (Landy and Sigall, 1974, [17]). There were two factors, A: Essay Quality ($a = 2$ levels: good and poor), and B: Appearance of the student who had written the essay ($b = 3$ levels: attractive photo, no photo (control), and unattractive photo). There were a total of $abn = 60$ subjects who rated the essays, with $n = 10$ subjects per treatment. The model can be written as follows, and data that were generated to reproduce the cell means and standard deviations are given in Table 3.2. An interaction plot of the treatment means is given in Figure 3.1.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where μ is the overall mean, α_i is the effect of the i^{th} level of factor A (essay quality 1=Good, 2=Poor), β_j is the effect of the j^{th} level of factor B (photo of essay writer: 1=attractive, 2=none, 3=unattractive), $(\alpha\beta)_{ij}$ is the effect of the interaction of the i^{th} level of essay quality and the j^{th} level of photo of essay writer. The sums of squares are computed below.

$$SSA = 3(10)\left[(17.10 - 14.72)^2 + (12.33 - 14.72)^2\right] = 340.77 \quad df_A = 2 - 1 = 1$$

$$SSB = 2(10)\left[(16.40 - 14.72)^2 + (15.65 - 14.72)^2 + (12.10 - 14.72)^2\right] = 211.00 \quad df_B = 3 - 1 = 2$$

$$SSAB = 10\left[(17.90 - 17.10 - 16.40 + 14.72)^2 + \cdots + (8.70 - 12.33 - 12.10 + 14.72)^2\right] = 36.58 \quad df_{AB} = 1(2) = 2$$

$$SSE = (10 - 1)\left[4.82^2 + 3.60^2 + 4.70^2 + 3.31^2 + 5.99^2 + 3.68^2\right] = 1067.92 \quad df_E = 3(2)(10 - 1) = 54$$

$$MSA = \frac{340.77}{1} = 340.77 \quad MSB = \frac{211.00}{2} = 105.50 \quad MSAB = \frac{36.58}{2} = 18.29 \quad MSE = \frac{1067.92}{54} = 19.78$$

The F-tests for the interaction between essay quality and student's photograph condition and for the main effects are given below. Note that $F_{.05,1,54} = 4.020$ and $F_{.05,2,54} = 3.168$.

$$H_0^{AB} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{23} = 0 \qquad TS : F_{AB} = \frac{18.29}{19.78} = 0.9247 \qquad P\left(F_{2,54} \geq .9247\right) = .4028$$

$$H_0^{A} : \alpha_1 = \alpha_2 = 0 \qquad TS : F_A = \frac{340.77}{19.78} = 17.2312 \qquad P\left(F_{1,54} \geq 17.2312\right) = .0001$$

$$H_0^{B} : \beta_1 = \beta_2 = \beta_3 = 0 \qquad TS : F_B = \frac{105.50}{19.78} = 5.3347 \qquad P\left(F_{2,54} \geq 5.3347\right) = .0077$$

The interaction effect is not significant, a more parsimonious model would combine the interaction and error sums of squares and degrees of freedom and form an additive model. Some practitioners suggest against doing this when the error degrees of freedom are small, but given the rather large number here (54), the test for interaction certainly has reasonable power. The additive Analysis of Variance Table is given in Table 3.3. The R output for the additive and interaction model, as well as a Complete versus Reduced F-test are given below.

∇

```
> options(contrasts=c("contr.sum","contr.poly"))
> halo.mod3 <- aov(grade ~ essayqual + picture)
> anova(halo.mod3)
Analysis of Variance Table

Response: grade
```

Essay Quality	Good ($i = 1$)	Good ($i = 1$)	Good ($i = 1$)	Poor ($i = 2$)	Poor ($i = 2$)	Poor ($i = 2$)
Photograph	Att ($j = 1$)	None ($j = 2$)	Unatt ($j = 3$)	Att ($j = 1$)	None ($j = 2$)	Unatt ($j = 3$)
k=1	8.76	18.51	18.54	14.05	23.13	4.58
k=2	24.26	26.09	12.49	18.47	14.76	2.29
k=3	13.38	14.41	23.81	15.47	20.09	14.49
k=4	25.30	19.90	15.36	10.26	4.63	4.94
k=5	16.62	18.28	21.79	19.45	17.15	8.73
k=6	16.19	16.83	12.82	10.75	6.97	10.17
k=7	17.25	19.79	16.47	19.42	16.87	11.07
k=8	19.67	13.65	9.10	13.11	8.50	11.02
k=9	19.08	15.06	11.24	13.71	10.13	10.25
k=10	18.49	16.48	13.37	14.30	11.77	9.47
Mean	17.90	17.90	15.50	14.90	13.40	8.70
Std. Dev.	4.82	3.60	4.70	3.31	5.99	3.68
Factor A Means	$\overline{y}_{1..} = 17.10$			$\overline{y}_{2..} = 12.33$		
Factor B (and Overall)	$\overline{y}_{.1.} = 16.40$	$\overline{y}_{.2.} = 15.65$	$\overline{y}_{.3.} = 12.10$			$\overline{y}_{...} = 14.72$

Table 3.2: Essay evaluations in Beauty is Talent Study

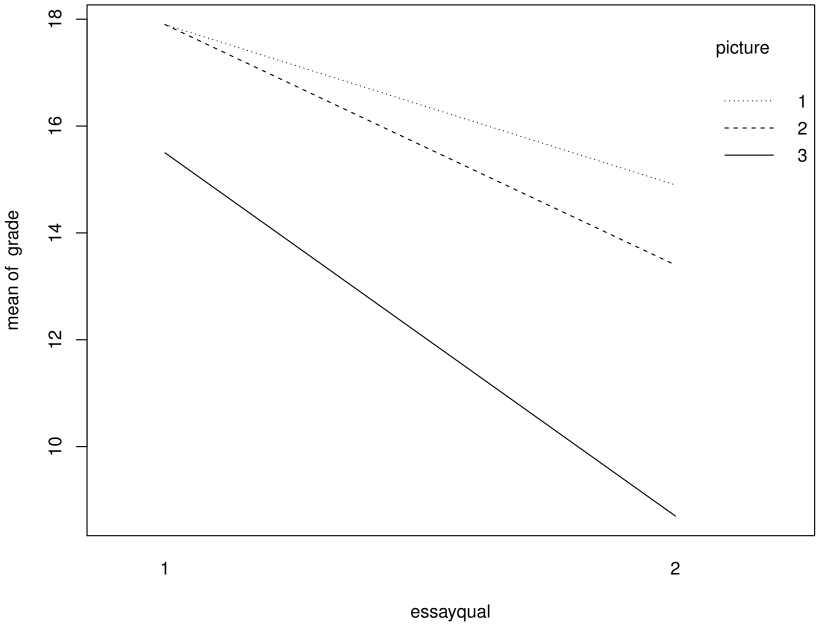

Figure 3.1: Essay evaluation means for the Beauty is Talent study

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(>F)$
Essay Quality (A)	1	340.77	$\frac{340.77}{1} = 340.77$	$\frac{340.77}{19.72} = 17.278$	.0001
Photo Condition (B)	2	211.00	$\frac{211.00}{2} = 105.50$	$\frac{105.50}{19.72} = 5.349$	.0075
ERROR	56	1104.49	$\frac{1104.49}{56} = 19.72$		
TOTAL	59	1656.264			

Table 3.3: Additive Analysis of Variance Table for Beauty and Talent study

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
essayqual  1  340.77  340.77  17.278 0.0001117 ***
picture    2  211.00  105.50   5.349 0.0074832 **
Residuals 56 1104.49   19.72

> summary.lm(halo.mod3)

Call:
aov(formula = grade ˜ essayqual + picture)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.7165     0.5733  25.668  < 2e-16 ***
essayqual1    2.3832     0.5733   4.157 0.000112 ***
picture1      1.6830     0.8108   2.076 0.042529 *
picture2      0.9335     0.8108   1.151 0.254502

Residual standard error: 4.441 on 56 degrees of freedom
Multiple R-squared:  0.3331,    Adjusted R-squared:  0.2974
F-statistic: 9.325 on 3 and 56 DF,  p-value: 4.269e-05

>
> halo.mod4 <- aov(grade ˜ essayqual*picture)
> anova(halo.mod4)
Analysis of Variance Table

Response: grade
                  Df  Sum Sq Mean Sq F value    Pr(>F)
essayqual          1  340.77  340.77 17.2312 0.0001182 ***
picture            2  211.00  105.50  5.3347 0.0076873 **
essayqual:picture  2   36.58   18.29  0.9247 0.4028316
Residuals         54 1067.92   19.78

> summary.lm(halo.mod4)
Call:
aov(formula = grade ˜ essayqual * picture)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          14.7165     0.5741  25.633  < 2e-16 ***
essayqual1            2.3832     0.5741   4.151 0.000118 ***
picture1              1.6830     0.8119   2.073 0.042971 *
picture2              0.9335     0.8119   1.150 0.255314
essayqual1:picture1  -0.8827     0.8119  -1.087 0.281804
essayqual1:picture2  -0.1332     0.8119  -0.164 0.870332

Residual standard error: 4.447 on 54 degrees of freedom
Multiple R-squared:  0.3552,    Adjusted R-squared:  0.2955
F-statistic:  5.95 on 5 and 54 DF,  p-value: 0.000188
```

```
>
> anova(halo.mod3, halo.mod4)
Analysis of Variance Table
Model 1: grade ~ essayqual + picture
Model 2: grade ~ essayqual * picture
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     56 1104.5
2     54 1067.9  2    36.575 0.9247 0.4028
```

Example 3.2: Penetration of Arrowheads by Clothing Fit and Type

A forensic science experiment was conducted to determine penetration of 4 arrowhead types on two clothing fits and types (MacPhee, et al, 2018, [21]). This analysis considers only the first arrowhead type (bullet style) with two factors: Clothing Fit (Tight ($i = 1$) and Loose ($i = 2$)) and Clothing Type (T-Shirt/95% Cotton ($j = 1$), Jeans/65% Cotton ($j = 2$), and Jeans/95% Cotton ($j = 3$)). The response was penetration (centimeters), with $n = 4$ replicates per combination of Clothing Fit and Type. The data are given in Table 3.4. The Analysis of Variance is given in Table 3.5 and an interaction plot is given in Figure 3.2. Due to the small amount of variation within treatments, both main effects and the interaction are highly significant. The R output is given below.

∇

Clothing Fit	$i = 1$	$i = 1$	$i = 1$	$i = 2$	$i = 2$	$i = 2$
Clothing Type	$j = 1$	$j = 2$	$j = 3$	$j = 1$	$j = 2$	$j = 3$
k=1	17.5	14.5	15.9	16.7	11.2	13.4
k=2	17.6	14.4	15.4	16.2	11	13.5
k=3	17.1	14.6	15.9	16.6	11.1	12.9
k=4	17.2	14.7	15.7	16.4	11.2	13.4
Mean	17.35	14.55	15.725	16.475	11.125	13.3
SD	0.238	0.129	0.236	0.222	0.096	0.271
Factor A Means	$\overline{y}_{1..} = 15.875$			$\overline{y}_{2..} = 13.633$		
Factor B (and Overall)	$\overline{y}_{.1.} = 16.913$	$\overline{y}_{.2.} = 12.838$	$\overline{y}_{.3.} = 14.513$			$\overline{y}_{...} = 14.754$

Table 3.4: Arrowhead penetration by Clothing Fit and Type (Bullet style arrow)

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(> F)$
Clothing Fit (A)	1	30.150	$\frac{30.150}{1} = 30.150$	$\frac{30.150}{0.043} = 693.556$	$< .0001$
Clothing Type (B)	2	67.123	$\frac{67.123}{2} = 33.562$	$\frac{33.562}{0.043} = 772.026$	$< .0001$
Interaction (AB)	2	6.603	$\frac{6.603}{2} = 3.302$	$\frac{3.302}{0.043} = 75.949$	$< .0001$
ERROR	18	0.783	$\frac{0.783}{18} = 0.043$		
TOTAL	23	104.660			

Table 3.5: Analysis of Variance Table for Arrowhead penetration experiment

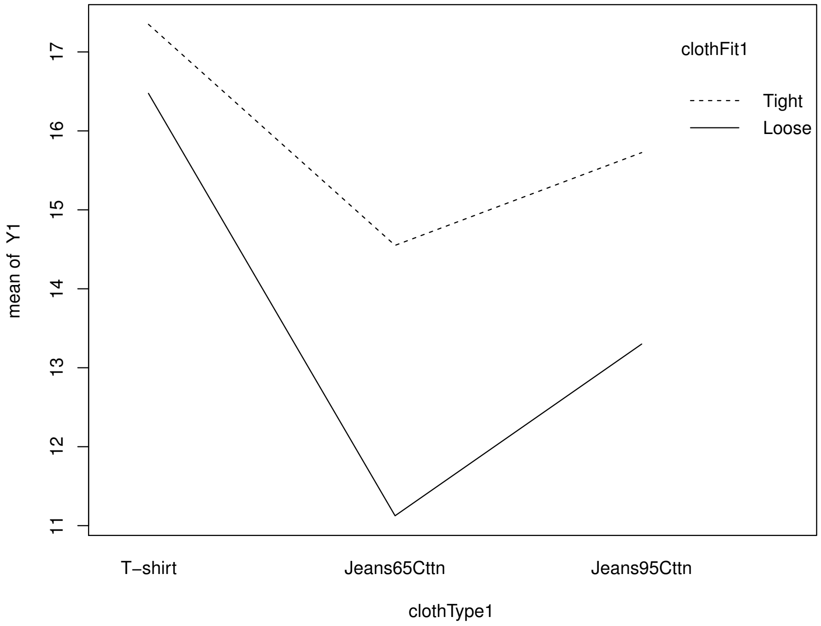

Figure 3.2: Arrowhead penetration by Clothing Fit and Type (Bullet style)

```
> arrow.mod1 <- aov(Y1 ~ clothFit1 * clothType1)
> anova(arrow.mod1)
Analysis of Variance Table
Response: Y1
                     Df Sum Sq Mean Sq F value    Pr(>F)
clothFit1             1 30.150  30.150 693.556 7.959e-16 ***
clothType1            2 67.123  33.562 772.026 < 2.2e-16 ***
clothFit1:clothType1  2  6.603   3.302  75.949 1.682e-09 ***
Residuals            18  0.783   0.043
> summary.lm(arrow.mod1)
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            14.75417    0.04256  346.67  < 2e-16 ***
clothFit11              1.12083    0.04256   26.34 7.96e-16 ***
clothType11             2.15833    0.06019   35.86  < 2e-16 ***
clothType12            -1.91667    0.06019  -31.84  < 2e-16 ***
clothFit11:clothType11 -0.68333    0.06019  -11.35 1.22e-09 ***
clothFit11:clothType12  0.59167    0.06019    9.83 1.16e-08 ***
Residual standard error: 0.2085 on 18 degrees of freedom
```

3.1.1 Contrasts and Post-Hoc Comparisons

When there is no interaction, contrasts and pairwise comparisons can be made among levels of Factor A and Factor B, respectively. In the case of general contrasts among the levels of Factors A and B, the following results are the same as those that are used with the 1-Way ANOVA. Recall that the coefficients for contrasts $\{a_i\}$ and $\{b_j\}$ must sum to zero.

$$C_A = \sum_{i=1}^{a} a_i \mu_{i.} = \sum_{i=1}^{a} a_i \left(\mu + \alpha_i + \overline{\beta}_{.}\right) = \sum_{i=1}^{a} a_i \alpha_i \qquad \text{s.t.} \sum_{i=1}^{a} a_i = 0$$

$$C_B = \sum_{j=1}^{b} b_j \mu_{.j} = \sum_{j=1}^{b} b_j \left(\mu + \overline{\alpha}_{.} + \beta_j\right) = \sum_{j=1}^{b} b_j \beta_j \qquad \text{s.t.} \sum_{j=1}^{b} b_j = 0$$

$$\hat{C}_A = \sum_{i=1}^{a} a_i \overline{Y}_{i..} \qquad V\left\{\hat{C}_A\right\} = \frac{\sigma^2}{bn} \sum_{i=1}^{a} a_i^2 \qquad \hat{SE}\left\{\hat{C}_A\right\} = \sqrt{\frac{MSE}{bn} \sum_{i=1}^{a} a_i^2}$$

$$\hat{C}_B = \sum_{j=1}^{b} b_j \overline{Y}_{.j.} \qquad V\left\{\hat{C}_B\right\} = \frac{\sigma^2}{an} \sum_{j=1}^{b} b_j^2 \qquad \hat{SE}\left\{\hat{C}_B\right\} = \sqrt{\frac{MSE}{an} \sum_{j=1}^{b} b_j^2}$$

Sums of Squares for the contrasts are also computed as in the 1-Way ANOVA.

$$SS_{C_A} = \frac{\left(\hat{C}_A\right)^2}{\frac{1}{bn}\sum_{i=1}^{a} a_i^2} = \frac{bn\left(\hat{C}_A\right)^2}{\sum_{i=1}^{a} a_i^2} \qquad SS_{C_B} = \frac{\left(\hat{C}_B\right)^2}{\frac{1}{an}\sum_{j=1}^{b} b_j^2} = \frac{an\left(\hat{C}_B\right)^2}{\sum_{j=1}^{b} b_j^2}$$

Tests of whether $C_A = 0$ or $C_B = 0$ can be conducted via t-tests or F-tests. Further, Confidence Intervals of C_A and C_B can be obtained from the estimates and their estimated standard errors. The degrees of freedom will be based on whether or not the interaction term is included or excluded from the Analysis of Variance. If the interaction is included, $df_E = ab(n-1)$, if it is removed, $df_E = abn - a - b + 1$. The rules for the F-test, t-test and Confidence Intervals are given below for C_A, with obvious changes for C_B.

$$H_0^{C_A} : C_A = 0 \qquad TS : F_{C_A} = \frac{SS_{C_A}}{MSE} \qquad RR : F_{C_A} \geq F_{\alpha,1,df_E} \qquad P = P\left(F_{1,df_E} \geq F_{C_A}\right)$$

$$H_0^{C_A} : C_A = 0 \qquad TS : t_{C_A} = \frac{\hat{C}_A}{\hat{SE}\left\{\hat{C}_A\right\}} \qquad RR : |t_{C_A}| \geq t_{\alpha/2,df_E} \qquad P = 2P\left(t_{df_E} \geq |t_{C_A}|\right)$$

$$(1-\alpha)100\% \text{ CI for } C_A\text{:} \quad \hat{C}_A \pm t_{\alpha/2,df_E} \hat{SE}\left\{\hat{C}_A\right\}$$

Example 3.3: Halo Effect - Essay Evaluation

Suppose for the Beauty is Talent study, the researchers were interested in contrasting the Attractive and Control Photograph conditions with the Unattractive condition. For this contrast, $b_1 = b_2 = 1$ and $b_3 = -2$, and will make use of the additive model.

$$a = 2 \quad b = 3 \quad n = 10 \quad MSE = 19.72 \quad df_E = 56 \quad \overline{y}_{.1.} = 16.40 \quad \overline{y}_{.2.} = 15.65 \quad \overline{y}_{.3.} = 12.10$$

$$\hat{C}_B = 16.40 + 15.65 - 2(12.10) = 7.85 \qquad \frac{1}{2(10)}\left(1^2 + 1^2 + (-2)^2\right) = 0.30 \qquad SS_{C_B} = \frac{7.85^2}{0.30} = 205.408$$

$$\hat{SE}\left\{\hat{C}_B\right\} = \sqrt{19.72(0.30)} = 2.432$$

$$H_0^{C_B}: C_B = 0 \quad TS: F_{C_B} = \frac{205.408}{19.72} = 10.416 \quad RR: F_{C_B} \geq F_{.05,1,56} = 4.013 \quad P = P\left(F_{1,56} \geq 10.416\right) = .0021$$

$$H_0^{C_B}: C_B = 0 \quad TS: t_{C_B} = \frac{7.85}{2.432} = 3.228 \quad RR: |t_{C_B}| \geq t_{\alpha/2,56} = 2.003 \quad P = 2P\left(t_{56} \geq |3.228|\right) = .0021$$

$$(1-\alpha)100\% \text{ CI for } C_B\text{: } 7.85 \pm 2.003(2.432) \quad \equiv \quad 7.85 \pm 4.87 \quad \equiv \quad (3.98, 12.72)$$

$$\nabla$$

When interactions are present, contrasts are often made among the cell means. The results based on Factor level means generalize to cell means.

$$C_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij}\mu_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij}\left(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}\right) \quad \text{s.t.} \sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij} = 0$$

$$\hat{C}_{AB} = \sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij}\overline{Y}_{ij.} \qquad SS_{C_{AB}} = \frac{\left(\hat{C}_{AB}\right)^2}{\frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij}^2} \qquad \hat{SE}\left\{\hat{C}_{AB}\right\} = \sqrt{\frac{MSE}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} ab_{ij}^2}$$

The F-test and t-test as well as Confidence Intervals are obtained as for main effects described above.

Example 3.4: Penetration of Arrowheads by Clothing Fit and Type

Suppose interest is in comparing the Tight ($i = 1$) versus Loose ($i = 2$) fit for the two types of jeans ($j = 2, 3$). That is, $C_{AB} = (\mu_{12} - \mu_{22}) - (\mu_{13} - \mu_{23})$ so that $ab_{11} = ab_{21} = 0, ab_{12} = ab_{23} = 1, ab_{22} = ab_{13} = -1$.

$$a = 2 \quad b = 3 \quad n = 4 \quad MSE = 0.043 \quad df_E = 18 \quad \overline{y}_{12.} = 14.550 \quad \overline{y}_{13.} = 15.725 \qquad \overline{y}_{22.} = 11.125 \quad \overline{y}_{23.} = 13.300$$

$$\hat{C}_{AB} = (14.550 - 11.125) - (15.725 - 13.300) = 3.425 - 2.425 = 1.000 \qquad \frac{1}{4}\left(0^2 + 0^2 + 1^2 + (-1)^2 + (-1)^2 + 1^2\right) = 1.000$$

$$SS_{C_{AB}} = \frac{1^2}{1} = 1 \qquad \hat{SE}\left\{\hat{C}_{AB}\right\} = \sqrt{\frac{0.043}{4}4} = 0.207$$

$$H_0^{C_{AB}} C_{AB} = 0 \quad TS: F_{C_{AB}} = \frac{1}{0.043} = 23.26 \quad RR: F_{C_{AB}} \geq F_{.05,1,18} = 4.414 \quad P = P\left(F_{1,18} \geq 23.26\right) < .0001$$

$$H_0^{C_{AB}} C_{AB} = 0 \quad TS: t_{C_{AB}} = \frac{1}{0.207} = 4.831 \quad RR: |t_{C_{AB}}| \geq t_{.025,18} = 2.101 \quad P = 2P\left(t_{18} \geq 4.831\right) < .0001$$

$$(1-\alpha)100\% \text{ CI for } C_{AB}\text{: } 1.000 \pm 2.101(0.207) \quad \equiv \quad 1.000 \pm 0.435 \quad \equiv \quad (0.565, 1.435)$$

Due to the very small within treatment variance, conclude that there is evidence of an interaction between jeans type and the fit of the jean in terms of arrow penetration.

$$\nabla$$

Special cases of contrasts are **pairwise comparisons** among means. For Factor A, Bonferroni's or Tukey's method (among many others) can be used to obtain simultaneous confidence intervals of the following forms. If the additive form of the Analysis of Variance is used, $df_E = abn - a - b + 1$, if the interaction is included (even if not significant) $df_E = ab(n-1)$.

Bonferroni (with $c_A^* = a(a-1)/2$ comparisons):

$$\left(\overline{y}_{i..} - \overline{y}_{i'..}\right) \pm t_{\alpha/2c_A^*, df_E} \sqrt{MSE\left(\frac{2}{bn}\right)}$$

Tukey

$$\left(\overline{y}_{i..} - \overline{y}_{i'..}\right) \pm q_{\alpha,a,df_E} \sqrt{MSE\left(\frac{1}{bn}\right)} \quad \equiv \quad \left(\overline{y}_{i..} - \overline{y}_{i'..}\right) \pm \frac{q_{\alpha,a,df_E}}{\sqrt{2}} \sqrt{MSE\left(\frac{2}{bn}\right)}$$

For Factor B, the following formulas are used.

Bonferroni (with $c_B^* = b(b-1)/2$ comparisons)

$$\left(\overline{y}_{.j.} - \overline{y}_{.j'.}\right) \pm t_{\alpha/2c_B^*, df_E} \sqrt{MSE\left(\frac{2}{an}\right)}$$

Tukey

$$\left(\overline{y}_{.j.} - \overline{y}_{.j'.}\right) \pm q_{\alpha,b,df_E} \sqrt{MSE\left(\frac{1}{an}\right)} \quad \equiv \quad \left(\overline{y}_{.j.} - \overline{y}_{j'..}\right) \pm \frac{q_{\alpha,b,df_E}}{\sqrt{2}} \sqrt{MSE\left(\frac{2}{an}\right)}$$

Example 3.5: Halo Effect - Essay Evaluation

In this study, the interaction between essay quality and photograph was not significant. The following results were obtained for the additive model.

$$MSE = 19.72 \quad df_E = 56 \quad a = 2 \quad \overline{y}_{1..} = 17.10, \overline{y}_{2..} = 12.33 \quad b = 3 \quad \overline{y}_{.1.} = 16.40, \overline{y}_{.2.} = 15.65, \overline{y}_{.3.} = 12.10 \quad n = 10$$

There are two essay qualities, so there is only one comparison. Bonferroni's and Tukey's methods will give the same interval. The following critical values were obtained in R with the **qt** and **qtukey** functions.

$$\overline{y}_{1..} - \overline{y}_{2..} = 17.10 - 12.33 = 4.44 \quad t_{.05/2,56} = 2.003 \quad \sqrt{19.72\left(\frac{2}{3(10)}\right)} = 1.147$$

$$4.44 \pm 2.003(1.147) \equiv 4.44 \pm 2.30 \equiv (2.14, 6.74)$$

$$q_{0.05,2,56} = 2.833 \qquad \sqrt{19.72\left(\frac{1}{3(10)}\right)} = 0.811 \qquad 2.833(0.811) = 2.30$$

The intervals are entirely positive, providing evidence that the true mean score is higher for the "Good" essay than the "Poor" essay. In terms of photo, there were 3 conditions. First, compute the minimum significant difference for each method, and then form simultaneous confidence intervals.

$$c_B^* = \frac{3(3-1)}{2} = 3 \qquad t_{.05/(2(3)),56} = 2.468 \qquad \sqrt{19.72\left(\frac{2}{2(10)}\right)} = 1.404 \qquad 2.468(1.404) = 3.47$$

$$q_{.05,3,56} = 3.405 \qquad \sqrt{19.72\left(\frac{1}{2(10)}\right)} = 0.993 \qquad 3.405(0.993) = 3.38$$

$$\overline{y}_{.1.} - \overline{y}_{.2.} = 16.40 - 15.65 = 0.75 \quad \overline{y}_{.1.} - \overline{y}_{.3.} = 16.40 - 12.10 = 4.30 \quad \overline{y}_{.2.} - \overline{y}_{.3.} = 15.65 - 12.10 = 3.55$$

The intervals for comparing Attractive with Unattractive and Control with Unattractive are entirely positive based on both Bonferroni's and Tukey's methods. Thus, conclude there is evidence of the halo effect. The interval comparing Attractive with control does contain zero, and those conditions do not differ significantly. The results are given in Table 3.6. The R output for Tukey's method is given below. Note that it subtracts the lower labeled group from the higher labeled group, so the intervals are the negative versions of those in Table 3.6.

j, j'	$\overline{y}_{.j.} - \overline{y}_{.j'.}$	Bonferroni CI	Tukey CI
1,2	0.75	$(-2.72, 4.22)$	$(-2.63, 4.13)$
1,3	4.30	$(0.83, 7.77)$	$(0.92, 7.68)$
2,3	3.55	$(0.08, 7.02)$	$(0.17, 6.93)$

Table 3.6: Simultaneous Confidence Intervals for Photo Conditions in Beauty is Talent Study

```
> anova(halo.mod1)
Analysis of Variance Table
Response: grade
          Df  Sum Sq Mean Sq F value    Pr(>F)
essayqual  1  340.77  340.77  17.278 0.0001117 ***
```

```
picture    2  211.00  105.50   5.349 0.0074832 **
Residuals 56 1104.49   19.72

> TukeyHSD(halo.mod1,"essayqual")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = grade ~ essayqual + picture)
$essayqual
         diff       lwr       upr     p adj
2-1 -4.766333 -7.063409 -2.469258 0.0001117

> TukeyHSD(halo.mod1,"picture")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = grade ~ essayqual + picture)
$picture
       diff       lwr        upr     p adj
2-1 -0.7495 -4.130658  2.6316579 0.8551372
3-1 -4.2995 -7.680658 -0.9183421 0.0093295
3-2 -3.5500 -6.931158 -0.1688421 0.0375422
```

When there is an interaction present, that means the effects of the levels of Factor A depend on the level of Factor B, and vice versa. Then it only makes sense to make comparisons among levels of one factor within the levels of the other factor (or simply compare all ab means). These comparisons are often called "slices" as they make comparisons of levels of one factor within the levels of the other factor. Note that these intervals will be wider than those for the main effects, as they are based on fewer observations per sample mean.

To compare all levels of Factor A, when Factor B is at level j, simultaneous confidence intervals can be up in the forms given below.

Bonferroni (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{ij.} - \overline{y}_{i'j.}) \pm t_{\alpha/2c_A^*,ab(n-1)}\sqrt{MSE\left(\frac{2}{n}\right)},$$

Tukey

$$(\overline{y}_{ij.} - \overline{y}_{i'j.}) \pm q_{\alpha,a,ab(n-1)}\sqrt{MSE\left(\frac{1}{n}\right)}.$$

For comparing levels of Factor B, when Factor A is at level i, the following formulas can be used.

Bonferroni (with $c_B^* = b(b-1)/2$)

$$(\overline{y}_{ij.} - \overline{y}_{ij'.}) \pm t_{\alpha/2c_B^*,ab(n-1)}\sqrt{MSE\left(\frac{2}{n}\right)},$$

Tukey

$$(\overline{y}_{ij.} - \overline{y}_{ij'.}) \pm q_{\alpha,b,ab(n-1)}\sqrt{MSE\left(\frac{1}{n}\right)}.$$

Example 3.6: Penetration of Arrowheads by Clothing Fit and Type

In this study, there was a significant interaction effect. Comparisons could be made among Clothing Fits within Types, with one Fit comparison within each of the 3 Types. Similarly, the 3 Types could be compared within each Fit. Further, all 6 treatments could be compared simultaneously. First, obtain minimum significant differences for comparing the Fits within Types.

$$t_{.05/2,18} = 2.101 \qquad \sqrt{0.043\left(\frac{2}{4}\right)} = 0.147 \qquad 2.101(0.147) = 0.308$$

$$q_{.05,2,18} = 2.971 \qquad \sqrt{0.043\left(\frac{1}{4}\right)} = 0.104 \qquad 2.971(0.104) = 0.308$$

Next, obtain minimum significant differences for comparing the Types within Fits.

$$t_{.05/(2(3)),18} = 2.639 \qquad \sqrt{0.043\left(\frac{2}{4}\right)} = 0.147 \qquad 2.639(0.147) = 0.388$$

$$q_{.05,3,18} = 3.609 \qquad \sqrt{0.043\left(\frac{1}{4}\right)} = 0.104 \qquad 3.609(0.104) = 0.375$$

If the goal is to compare all pairs among the $ab = 2(3) = 6$ treatments, then there are $c^*_{AB} = 6(5)/2 = 15$ comparisons.

$$t_{.05/(2(15)),18} = 3.380 \qquad \sqrt{0.043\left(\frac{2}{4}\right)} = 0.147 \qquad 3.380(0.147) = 0.497$$

$$q_{.05,6,18} = 4.494 \qquad \sqrt{0.043\left(\frac{1}{4}\right)} = 0.104 \qquad 4.494(0.104) = 0.467$$

The simultaneous comparisons among the 6 treatments are given in Table 3.7. All intervals are either entirely positive or negative. The highest penetration occurs when $(i, j) = (1, 1)$, which corresponds to the Tight fit with a t-shirt. The lowest penetration occurs when $(i, j) = (2, 2)$, which corresponds to a Loose fit with Jeans made from 65% cotton.

∇

$(i, i'), (j, j')$	$\overline{y}_{ij.} - \overline{y}_{i'j'.}$	Bonferroni CI	Tukey CI
(1,1),(1,2)	2.800	$(2.303, 3.297)$	$(2.333, 3.267)$
(1,1),(1,3)	1.625	$(1.128, 2.122)$	$(1.158, 3.237)$
(1,1),(2,1)	0.875	$(0.378, 1.372)$	$(0.408, 1.342)$
(1,1),(2,2)	6.225	$(5.728, 6.722)$	$(5.758, 6.692)$
(1,1),(2,3)	4.050	$(3.553, 4.547)$	$(3.583, 517)$
(1,2),(1,3)	-1.175	$(-1.672, -0.678)$	$(-1.642, -0.708)$
(1,2),(2,1)	-1.925	$(-2.422, -1.428)$	$(-2.392, -1.458)$
(1,2),(2,2)	3.425	$(2.928, 3.922)$	$(2.958, 3.892)$
(1,2),(2,3)	1.250	$(0.753, 1.747)$	$(0.783, 1.717)$
(1,3),(2,1)	-0.750	$(-1.247, -0.253)$	$(-1.217, -0.283)$
(1,3),(2,2)	4.600	$(4.103, 5.097)$	$(4.133, 5.067)$
(1,3),(2,3)	2.425	$(1.928, 2.922)$	$(1.958, 2.892)$
(2,1),(2,2)	5.350	$(4.853, 5.847)$	$(4.883, 5.817)$
(2,1),(2,3)	3.175	$(2.678, 3.672)$	$(2.708, 3.642)$
(2,2),(2,3)	-2.175	$(-2.672, -1.678)$	$(-2.642, -1.708)$

Table 3.7: Simultaneous Confidence Intervals for the Arrowhead Penetration Study

Unbalanced Data

When the numbers of replicates per treatment vary, the sums of squares computations cannot be used and a regression model is fit. Note that software packages do it this way regardless of whether the data are balanced or unbalanced. The model makes use of dummy type variables for the various levels of Factors A and B, while interaction effects make use of cross-products of the dummy variables. Let $X_1^A, \ldots, X_{a-1}^A$ and $X_1^B, \ldots, X_{b-1}^B$ be defined as follow (this is a parameterization with the effects summing to 0).

$$X_i^A = \begin{cases} 1 & : \text{ if Factor } A \text{ is at level } i \;\; i = 1, \ldots, a-1 \\ -1 & : \text{ if Factor } A \text{ is at level } a \\ 0 & : \text{ otherwise} \end{cases}$$

$$X_j^B = \begin{cases} 1 & : \text{ if Factor } B \text{ is at level } j \;\; j = 1, \ldots, b-1 \\ -1 & : \text{ if Factor } B \text{ is at level } b \\ 0 & : \text{ otherwise} \end{cases}$$

Then the regression model can be fit with the following equation.

$$Y = \beta_0 + \beta_1^A X_1^A + \cdots + \beta_{a-1}^A X_{a-1}^A + \beta_1^B X_1^B + \cdots + \beta_{b-1}^B X_{b-1}^B + \beta_{11}^{AB} X_1^A X_1^B + \cdots + \beta_{a-1,b-1}^{AB} X_{a-1}^A X_{b-1}^B + \epsilon$$

To test for interactions between levels of Factors A and B, is to test $H_0^{AB} : \beta_{11}^{AB} = \ldots = \beta_{a-1,b-1}^{AB} = 0$, and can be conducted with a Complete versus Reduced F-test. Tests for main effects of Factors A and B (controlling for the interaction) can be done in a similar manner (see Example 3.5 below).

When each factor has 2 levels, the model simplifies as follows, and the tests for interaction and main effects are t-tests that can be obtained without having to fit Complete and Reduced models.

$$Y = \beta_0 + \beta_1^A X_1^A + \beta_1^B X_1^B + \beta_{11}^{AB} X_1^A X_1^B + \epsilon$$

Example 3.7: Lead Content in Lip Products

A European regulatory study measured lead content (mg/kg) in 223 manufactured lip products (Piccinini, Piecha, and Torrent, 2013, [28]). The products were classified by the factors: Color (Factor A, $a = 4$: Red ($i = 1$), Purple ($i = 2$), Pink ($i = 3$), and Brown ($i = 4$)) and Product Type (Factor B, $b = 2$: Lip Stick ($j = 1$) and Lip Gloss ($j = 2$)). The number of brands, n_{ij} varied among the "treatments." The products included here are treated as a random sample from a population of all possible lip product formulations that could be manufactured for these colors and product types. The summary statistics are given in Table 3.8 and the interaction plot is given in Figure 3.3. The analysis is based on the raw data.

Four models are fit. Model 1 contains all main effects and interactions, model 2 contains the main effects only, model 3 contains Factor B main effects and interactions, model 4 contains Factor A main effects and interactions. Two more models will be considered. Model 5 contains Factor A main effects only, while Model 6 contains Factor B main effects only.

Model1 1: $\hat{Y}_1 = 0.564 - 0.163X_1^A + 0.086X_2^A + 0.032X_3^A + 0.188X_1^B - 0.033X_1^A X_1^B + 0.092X_2^A X_1^B + 0.026X_3^A X_1^B$

$$SSE_1 = 66.898 \quad df_{E1} = 223 - 8 = 215$$

Model1 2: $\hat{Y}_2 = 0.563 - 0.173X_1^A + 0.120X_2^A + 0.040X_3^A + 0.185X_1^B \quad SSE_2 = 66.898 \quad df_{E2} = 223 - 5 = 218$

Model1 3: $\hat{Y}_3 = 0.570 + 0.185X_1^B - 0.093X_1^A X_1^B + 0.124X_2^A X_1^B + 0.036X_3^A X_1^B \quad SSE_3 = 67.952 \quad df_{E3} = 223 - 5 = 218$

Model1 4: $\hat{Y}_4 = 0.631 - 0.148X_1^A + 0.110X_2^A - 0.007X_3^A - 0.062X_1^A X_1^B + 0.022X_2^A X_1^B + 0.103X_3^A X_1^B$

$$SSE_4 = 72.481 \quad df_{E4} = 223 - 7 = 216$$

Model1 5: $\hat{Y}_5 = 0.570 + 0.185X_1^B \quad SSE_5 = 68.918 \quad df_{E5} = 223 - 2 = 221$

Model1 6: $\hat{Y}_6 = 0.628 - 0.167X_1^A + 0.121X_2^A + 0.024X_3^A \quad SSE_6 = 73.657 \quad df_{E4} = 223 - 4 = 219$

To test whether there is a significant interaction between Color and Product Type on Lead content, Models 1 (Complete) and 2 (Reduced) are compared.

$$H_0^{AB} : \beta_{11}^{AB} = \beta_{21}^{AB} = \beta_{31}^{AB} = 0 \quad TS : F_{AB} = \frac{\left[\frac{66.898-66.160}{218-215}\right]}{\left[\frac{66.160}{215}\right]} = \frac{0.246}{0.308} = 0.799 \quad P\left(F_{3,215} \geq 0.799\right) = .4956$$

In practice, given the interaction is not significant (and the overall sample size is quite large), to test for main effects for Factors A and B, Model 2 (Complete) and Models 5 (Factor A) and 6 (Factor B) are compared. First, Models 3 and 4 will be compared with Model 1 to mimic the results of a single computer pass of the interaction model (this is the method used by the SAS Systems Type III sums of squares).

$$H_0^A : \beta_1^A = \beta_2^A = \beta_3^A = 0 \quad TS : F_A = \frac{\left[\frac{67.592-66.160}{218-215}\right]}{\left[\frac{66.160}{215}\right]} = \frac{0.477}{0.308} = 1.551 \quad P\left(F_{3,215} \geq 1.551\right) = .2023$$

$$H_0^B : \beta_1^B = 0 \quad TS : F_B = \frac{\left[\frac{72.481-66.160}{216-215}\right]}{\left[\frac{66.160}{215}\right]} = \frac{6.321}{0.308} = 20.54 \quad P\left(F_{1,215} \geq 20.54\right) < .0001$$

Had Model 2 (additive) been used as the Complete Model for the main effects tests (with Models 5 and 6 as the Reduced Models), the F-tests show that the main effect for Color is much closer to being significant when the interaction is no longer being controlled for.

$$H_0^A : \beta_1^A = \beta_2^A = \beta_3^A = 0 \quad TS : F_A = \frac{\left[\frac{68.918-66.898}{221-218}\right]}{\left[\frac{66.898}{218}\right]} = \frac{0.673}{0.307} = 2.194 \quad P\left(F_{3,218} \geq 2.194\right) = .0897$$

$$H_0^B : \beta_1^B = 0 \quad TS : F_B = \frac{\left[\frac{73.657-66.898}{219-218}\right]}{\left[\frac{66.898}{218}\right]} = \frac{6.759}{0.307} = 22.02 \quad P\left(F_{1,218} \geq 22.02\right) < .0001$$

The primary conclusion is that there are significant differences in Product Type (Lip Stick has higher mean Lead levels than Lip Gloss). There is not a significant Color effect or Color/Product Type interaction. The output for the R program is given below.

∇

Color (i)	Product Type (j)	n_{ij}	$\overline{y}_{ij}$	s_{ij}
Red (1)	Lip Stick (1)	31	0.557	0.758
Red (1)	Lip Gloss (2)	14	0.246	0.217
Purple (2)	Lip Stick (1)	25	0.931	0.657
Purple (2)	Lip Gloss (2)	12	0.369	0.212
Pink (3)	Lip Stick (1)	51	0.811	0.580
Pink (3)	Lip Gloss (2)	30	0.381	0.302
Brown (4)	Lip Stick (1)	42	0.712	0.591
Brown (4)	Lip Gloss (2)	18	0.507	0.462

Table 3.8: Summary statistics for lip product lead content Study

```
### Regression Model

> anova(ll.mod2, ll.mod1)
Analysis of Variance Table
Model 1: Pb ~ X1.A + X2.A + X3.A + X1.B
Model 2: Pb ~ X1.A + X2.A + X3.A + X1.B + I(X1.A * X1.B) + I(X2.A * X1.B) +
    I(X3.A * X1.B)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    218 66.898
2    215 66.160  3   0.73838 0.7998 0.4952
```

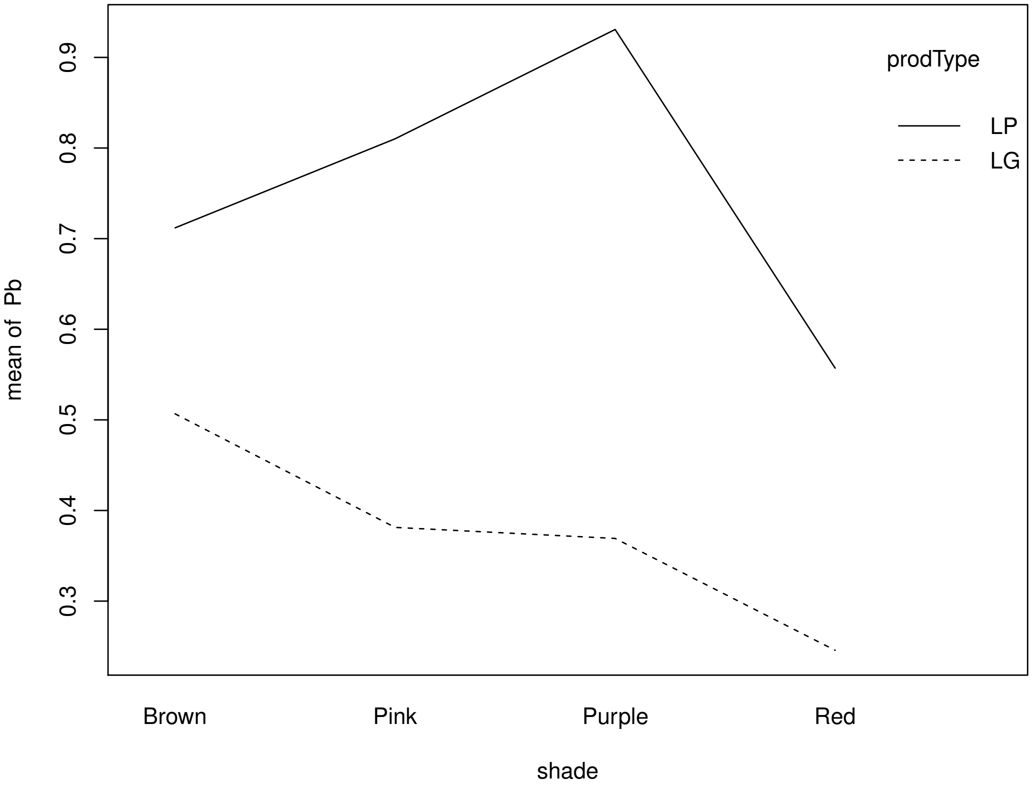

Figure 3.3: Interaction for Lip Product Lead content study

```
> anova(ll.mod3, ll.mod1)
Analysis of Variance Table
Model 1: Pb ~ X1.B + I(X1.A * X1.B) + I(X2.A * X1.B) + I(X3.A * X1.B)
Model 2: Pb ~ X1.A + X2.A + X3.A + X1.B + I(X1.A * X1.B) + I(X2.A * X1.B) +
    I(X3.A * X1.B)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    218 67.592
2    215 66.160  3     1.432 1.5511 0.2023

> anova(ll.mod4, ll.mod1)
Analysis of Variance Table
Model 1: Pb ~ X1.A + X2.A + X3.A + I(X1.A * X1.B) + I(X2.A * X1.B) + I(X3.A *
    X1.B)
Model 2: Pb ~ X1.A + X2.A + X3.A + X1.B + I(X1.A * X1.B) + I(X2.A * X1.B) +
    I(X3.A * X1.B)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    216 72.481
2    215 66.160  1    6.3213 20.542 9.678e-06 ***

> anova(ll.mod5, ll.mod2)
Analysis of Variance Table
Model 1: Pb ~ X1.B
Model 2: Pb ~ X1.A + X2.A + X3.A + X1.B
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    221 68.918
2    218 66.898  3    2.0196 2.1937 0.08972 .

> anova(ll.mod6, ll.mod2)
Analysis of Variance Table

Model 1: Pb ~ X1.A + X2.A + X3.A
Model 2: Pb ~ X1.A + X2.A + X3.A + X1.B
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    219 73.657
2    218 66.898  1    6.7591 22.026 4.751e-06 ***

### Using aov Function - Interaction Model
## Sequential Sums of Squares
> anova(ll.aov1)
Analysis of Variance Table
Response: Pb
                Df Sum Sq Mean Sq F value    Pr(>F)
shade            3  1.879  0.6264  2.0357    0.1099
prodType         1  6.759  6.7591 21.9652 4.926e-06 ***
shade:prodType   3  0.738  0.2461  0.7998    0.4952
Residuals      215 66.160  0.3077

### Main Effects given other main effects, Interaction given all main effects
> Anova(ll.aov1, Type="III")
Anova Table (Type II tests)
Response: Pb
               Sum Sq  Df F value    Pr(>F)
shade           2.020   3  2.1877   0.09046 .
prodType        6.759   1 21.9652 4.926e-06 ***
shade:prodType  0.738   3  0.7998   0.49517
Residuals      66.160 215

### Using aov Function - Additive Model
## Sequential Sums of Squares
> anova(ll.aov2)
Analysis of Variance Table
Response: Pb
           Df Sum Sq Mean Sq F value    Pr(>F)
shade       3  1.879  0.6264  2.0414     0.109
```

```
prodType    1  6.759  6.7591 22.0259 4.751e-06 ***
Residuals 218 66.898  0.3069

### Main Effects given other main effects
> Anova(ll.aov2, Type="III")
Anova Table (Type II tests)
Response: Pb
          Sum Sq  Df F value    Pr(>F)
shade      2.020   3  2.1937   0.08972 .
prodType   6.759   1 22.0259 4.751e-06 ***
Residuals 66.898 218
```

For contrasts among main effects, which are relevant when the model is additive, the following sums of squares and variances are obtained. Inferences are then made as in the balanced case.

$$SS_{C_A} = \frac{\left(\hat{C}_A\right)^2}{\sum_{i=1}^{a} \frac{a_i^2}{n_{i.}}} \qquad V\left\{\hat{C}_A\right\} = \sigma^2 \sum_{i=1}^{a} \frac{a_i^2}{n_{i.}} \qquad SS_{C_B} = \frac{\left(\hat{C}_B\right)^2}{\sum_{j=1}^{b} \frac{b_j^2}{n_{.j}}} \qquad V\left\{\hat{C}_B\right\} = \sigma^2 \sum_{j=1}^{b} \frac{b_j^2}{n_{.j}}$$

In the case of contrasts among cell means, the following results apply.

$$SS_{C_{AB}} = \frac{\left(\hat{C}_{AB}\right)^2}{\sum_{i=1}^{a} \sum_{j=1}^{b} \frac{ab_{ij}^2}{n_{ij}}} \qquad V\left\{\hat{C}_{AB}\right\} = \sigma^2 \sum_{i=1}^{a} \sum_{j=1}^{b} \frac{ab_{ij}^2}{n_{ij}}$$

The following adjustments are made to the Bonferroni and Tukey methods for the unbalanced case.

Bonferroni (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*, df_E} \sqrt{MSE\left(\frac{1}{n_{i.}} + \frac{1}{n_{i'.}}\right)}$$

Tukey-Kramer

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha, a, df_E} \sqrt{\frac{MSE}{2}\left(\frac{1}{n_{i.}} + \frac{1}{n_{i'.}}\right)}$$

Example 3.8: Lead Content in Lip Products

Although the Color was not a significant main effect at the $\alpha = 0.05$ significance level, suppose interest was in comparing the colors. First, consider a contrast of Red ($i = 1$) versus Purple($i = 2$)/Pink($i = 3$)/Brown($i = 4$), with $a_1 = 3, a_2 = a_3 = a_4 = -1$. Then obtain all pairwise comparisons among colors. The additive model (Model 2) is used.

$$MSE = \frac{66.898}{218} = 0.307 \quad n_{1.} = 45 \quad n_{2.} = 37 \quad n_{3.} = 81 \quad n_{4.} = 60 \quad \overline{y}_{1..} = 0.460 \quad \overline{y}_{2..} = 0.749 \quad \overline{y}_{3..} = 0.651 \quad \overline{y}_{4..} = 0.650$$

$$\hat{C}_A = 3(0.460) - 0.749 - 0.651 - 0.650 = -0.670 \qquad \frac{3^2}{45} + \frac{(-1)^2}{37} + \frac{(-1)^2}{81} + \frac{(-1)^2}{60} = 0.256$$

$$SS_{C_A} = \frac{(-0.670)^2}{0.256} = 1.753 \qquad \hat{SE}\left\{\hat{C}_A\right\} = \sqrt{0.307(0.256)} = 0.280$$

$$H_0^{C_A}: C_A = 0 \quad TS: F_{C_A} = \frac{1.753}{0.307} = 5.710 \quad RR: F_{C_A} \geq F_{.05,1,218} = 3.884 \quad P = P\left(F_{1,218} \geq 5.710\right) = 0.0177$$

$$H_0^{C_A}: C_A = 0 \quad TS: t_{C_A} = \frac{-0.670}{0.280} = -2.393 \quad RR: |t_{C_B}| \geq t_{\alpha/2,218} = 1.971 \quad P = 2P\left(t_{218} \geq |-2.393|\right) = .0176$$

$$(1-\alpha)100\% \text{ CI for } C_A: \; -0.670 \pm 1.971(0.280) \quad \equiv \quad -0.670 \pm 0.552 \quad \equiv \quad (-1.222, -0.118)$$

The contrast is significantly different from 0. Note that this contrast consisted of comparing the lowest mean with the average of the 3 higher means.

There are $a = 4$ colors, which implies $c_A^* = 6$ pairwise comparisons. Consider the the pairwise comparison between Red ($i = 1$) and Purple ($i' = 2$).

$$\overline{y}_{1..} - \overline{y}_{2..} = 0.460 - 0.749 = -0.289 \quad \frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{45} + \frac{1}{37} = 0.049249 \quad t_{.05/(2(6)),218} = 2.663 \quad q_{.05,4,218} = 3.661$$

$$\text{Bonferroni CI: } -0.289 \pm 2.663\sqrt{0.307(0.049249)} \quad \equiv \quad -0.289 \pm 0.327 \quad \equiv \quad (-0.616, 0.038)$$

$$\text{Tukey CI: } -0.289 \pm 3.661\sqrt{\frac{0.307}{2}0.049249} \quad \equiv \quad -0.289 \pm 0.318 \quad \equiv \quad (-0.607, 0.029)$$

The results for all pairs are given in Table 3.9. No pairs of colors are significantly different.

i, i'	$\overline{y}_{i..} - \overline{y}_{i'..}$	$\hat{SE}\{\overline{y}_{i..} - \overline{y}_{i'..}\}$	Bonferroni CI	Tukey CI
1,2	-0.289	0.123	$(-0.616, 0.038)$	$(-0.607, 0.029)$
1,2	-0.191	0.103	$(-0.465, 0.083)$	$(-0.458, 0.076)$
1,2	-0.190	0.109	$(-0.481, 0.101)$	$(-0.473, 0.093)$
1,2	0.098	0.110	$(-0.195, 0.391)$	$(-0.187, 0.383)$
1,2	0.099	0.116	$(-0.209, 0.407)$	$(-0.201, 0.399)$
1,2	0.001	0.094	$(-0.250, 0.252)$	$(-0.243, 0.245)$

Table 3.9: Simultaneous Confidence Intervals for Lip Product Colors in Lead study

Higher Order Models

Models can be extended to any number of factors. The case of a balanced experiment with three factors A, B, and C with a, b, and c levels respectively, and n replicates per treatment is considered here. Extensions to more factors and unbalanced data can be generalized from this and previous sections. The model is given below, along with representative sums of squares.

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}$$

$$i = 1, \ldots, a \quad j = 1, \ldots, b \quad k = 1, \ldots, c \quad l = 1, \ldots, n \qquad \epsilon_{ijkl} \sim N\left(0, \sigma^2\right)$$

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{k=1}^{c} \gamma_i = \sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = \sum_{i=1}^{a} (\alpha\beta\gamma)_{ijk} = \sum_{j=1}^{b} (\alpha\beta\gamma)_{ijk} = \sum_{k=1}^{c} (\alpha\beta\gamma)_{ijk} = 0 \quad \forall i, j, k$$

$$SSA = bcn \sum_{i=1}^{a} \left(\overline{Y}_{i\ldots} - \overline{Y}_{\ldots.}\right)^2 \qquad SSAB = cn \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij..} - \overline{Y}_{i\ldots} - \overline{Y}_{.j..} + \overline{Y}_{\ldots.}\right)^2$$

$$SSABC = n \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(\overline{Y}_{ijk.} - \overline{Y}_{ij..} - \overline{Y}_{i.k.} - \overline{Y}_{.jk.} + \overline{Y}_{i\ldots} + \overline{Y}_{.j..} + \overline{Y}_{..k.} - \overline{Y}_{\ldots.}\right)^2$$

$$SSE = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \sum_{l=1}^{n} \left(Y_{ijkl} - \overline{Y}_{ijk.}\right)^2$$

$$df_A = a - 1 \qquad df_{AB} = (a-1)(b-1) \qquad df_{ABC} = (a-1)(b-1)(c-1) \qquad df_E = abc(n-1)$$

The F-tests for interactions and main effects are conducted as in the case of the 2-factor model.

Example 3.9: Oil Holding Capacity of Banana Cultivars

A food chemistry experiment was conducted to measure oil holding capacity of bananas (Anyasi, Jideani, and Mchau, 2015, [4]). The response was oil holding capacity (g/g dry weight) and there were three factors: Banana Cultivar (Luvhele ($i = 1$), Mabonde ($i = 2$), M-red ($i = 3$)), pre-treatment acid (ascorbic ($j = 1$), citric ($j = 2$), lactic ($j = 3$)), and amount of acid (10 g/L ($k = 1$), 15 ($k = 2$), 20 ($k = 3$)). There were $n = 3$ replicates per treatment. Although the amount of acid is numeric, it is treated as a nominal factor in this analysis.

The data means are given in Table 3.10. The Analysis of Variance is given in Table 3.11, note that none of the interactions or main effects are significant. Apparently none of these factors are associated with oil holding capacity of bananas. The R output is given below.

∇

```
> ban.mod1 <- aov(OHC ~ cultivar * acidType * acidDose)
> anova(ban.mod1)
Analysis of Variance Table
Response: OHC
                           Df  Sum Sq Mean Sq F value Pr(>F)
cultivar                    2  0.9923 0.49614  1.9031 0.1590
acidType                    2  0.2811 0.14053  0.5391 0.5864
acidDose                    2  0.3556 0.17778  0.6819 0.5099
cultivar:acidType           4  0.1939 0.04847  0.1859 0.9447
cultivar:acidDose           4  0.4634 0.11585  0.4444 0.7760
acidType:acidDose           4  0.1558 0.03894  0.1494 0.9625
cultivar:acidType:acidDose  8  1.2886 0.16107  0.6179 0.7590
Residuals                  54 14.0777 0.26070
```

ijk	Mean	ijk	Mean	ijk	Mean	ijk	Mean
111	1.7300	211	1.6000	311	1.9300	.11	1.7533
112	1.5300	212	1.4700	312	1.8700	.12	1.6233
113	1.6000	213	1.4700	313	1.6000	.13	1.5567
121	1.7300	221	1.8000	321	1.6000	.21	1.7100
122	1.8000	222	1.6700	322	1.9300	.22	1.8000
123	1.7300	223	1.2700	323	1.6700	.23	1.5567
131	1.6000	231	1.0700	331	2.0000	.31	1.5567
132	1.4000	232	1.4700	332	1.9300	.32	1.6000
133	1.5300	233	1.6000	333	1.3300	.33	1.4867
11.	1.6200	21.	1.5133	31.	1.8000	.1.	1.6444
12.	1.7533	22.	1.5800	32.	1.7333	.2.	1.6889
13.	1.5100	23.	1.3800	33.	1.7533	.3.	1.5478
1.1	1.6867	2.1	1.4900	3.1	1.8433	..1	1.6733
1.2	1.5767	2.2	1.5367	3.2	1.9100	..2	1.6744
1.3	1.6200	2.3	1.4467	3.3	1.5333	..3	1.5333
1..	1.6278	2..	1.4911	3..	1.7622	...	1.6270

Table 3.10: Oil Holding Capacity for Bananas

Source	df	SS	MS	F	$F_{.05}$	$Pr > F$
A	2	0.9923	0.4961	1.9031	3.1682	0.1590
B	2	0.2811	0.1405	0.5391	3.1682	0.5864
C	2	0.3556	0.1778	0.6819	3.1682	0.5099
AB	4	0.1939	0.0485	0.1859	2.5429	0.9447
AC	4	0.4634	0.1159	0.4444	2.5429	0.7760
BC	4	0.1558	0.0389	0.1494	2.5429	0.9625
ABC	8	1.2886	0.1611	0.6179	2.1152	0.7590
Error	54	14.0777	0.2607			
Total	80	17.8083				

Table 3.11: Analysis of Variance of Oil Holding Capacity for Bananas

Example 3.10: Lead Content in Lip Products

This study included a third factor, Price with 3 levels containing ranges of prices (≤ 5 euros ($k = 1$), 5-15 euros ($k = 2$), ≥ 15 euros ($k = 3$)). The model would now include two new dummy variables, as well as extended numbers of interactions. The **aov** function will be used directly for the analysis. The R output is given below with no changes in interpretation. Price and none of the interactions are significant, the primary important factor is product.

∇

```
## Interaction Model
> ll.aov3 <- aov(Pb ~ shade * prodType * priceCatgry)
> anova(ll.aov3)
Analysis of Variance Table
Response: Pb
                            Df Sum Sq Mean Sq F value    Pr(>F)
shade                        3  1.879  0.6264  1.9836    0.1175
prodType                     1  6.759  6.7591 21.4031 6.556e-06 ***
priceCatgry                  1  0.004  0.0043  0.0135    0.9076
shade:prodType               3  0.741  0.2469  0.7817    0.5054
shade:priceCatgry            3  0.516  0.1721  0.5450    0.6520
prodType:priceCatgry         1  0.044  0.0442  0.1401    0.7086
shade:prodType:priceCatgry   3  0.222  0.0739  0.2341    0.8725
Residuals                  207 65.371  0.3158

> Anova(ll.aov3, Type="II")
Anova Table (Type II tests)
Response: Pb
                           Sum Sq  Df F value    Pr(>F)
shade                       1.960   3  2.0685    0.1055
prodType                    6.729   1 21.3081 6.857e-06 ***
priceCatgry                 0.006   1  0.0205    0.8862
shade:prodType              0.749   3  0.7907    0.5003
shade:priceCatgry           0.493   3  0.5205    0.6686
prodType:priceCatgry        0.044   1  0.1401    0.7086
shade:prodType:priceCatgry  0.222   3  0.2341    0.8725
Residuals                  65.371 207

## Additive Model
> ll.aov4 <- aov(Pb ~ shade + prodType + priceCatgry)
> anova(ll.aov4)
Analysis of Variance Table
Response: Pb
             Df Sum Sq Mean Sq F value    Pr(>F)
shade         3  1.879  0.6264  2.0321    0.1103
prodType      1  6.759  6.7591 21.9262 4.992e-06 ***
priceCatgry   1  0.004  0.0043  0.0138    0.9065
Residuals   217 66.894  0.3083

> Anova(ll.aov4, Type="II")
Anova Table (Type II tests)
Response: Pb
            Sum Sq  Df F value    Pr(>F)
shade        1.995   3  2.1572   0.09404 .
prodType     6.763   1 21.9381 4.964e-06 ***
priceCatgry  0.004   1  0.0138   0.90646
Residuals   66.894 217
```

3.1.2 Mixed Effects Models

When Factor A is a **Fixed Factor** (all levels of interest are included in the study), and Factor B is a **Random Factor** (only a sample of its levels are included, such as medical centers in a clinical trial), it is called a **Mixed Effects Model**. The computation of the sums of squares in the Analysis of Variance is the same, but the tests for treatment effects change. The model is given below.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where μ is the overall mean, α_i is the (fixed) effect of the i^{th} level of factor A, β_j is the (random) effect of the j^{th} level of factor B, $(\alpha\beta)_{ij}$ is the (random) interaction of the factor A at level i and factor B at level j. One (of several) ways this model is parameterized is to assume the following model structure.

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \beta_j \sim N\left(0, \sigma_b^2\right) \qquad (\alpha\beta)_{ij} \sim N\left(0, \sigma_{ab}^2\right) \qquad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

All random effects and error terms are assumed mutually independent in this (particular) formulation. The Analysis of Variance for the mixed effects model is given in Table 3.12. Data are no longer assumed independent when they are measured on the same level of a random factor.

$$E\left\{Y_{ijk}\right\} = \mu + \alpha_i \qquad V\left\{Y_{ijk}\right\} = \sigma_b^2 + \sigma_{ab}^2 + \sigma^2$$

$$\text{COV}\left\{Y_{ijk}, Y_{i'j'k'}\right\} = \begin{cases} \sigma_b^2 + \sigma_{ab}^2 + \sigma^2 & : \quad i = i', j = j', k = k' \\ \sigma_b^2 + \sigma_{ab}^2 & : \quad i = i', j = j', k \neq k' \\ \sigma_b^2 & : \quad i \neq i', j = j', \forall k, k' \\ 0 & : \quad \text{otherwise} \end{cases}$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(> F)$
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSAB}$	$P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$
B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSAB}$	$P\left(F_{b-1,(a-1)(b-1)} \geq F_B\right)$
AB	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$	$P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$
ERROR	$ab(n-1)$	SSE	$MSE = \frac{SSE}{ab(n-1)}$		
TOTAL	$abn-1$	TSS			

Table 3.12: The Analysis of Variance Table for a 2-Factor Factorial Design - Factor A fixed, Factor B random

The expectations of the mean squares for the mixed effects model (with A fixed, and B random) are given below.

$$E\{MSA\} = \sigma^2 + n\sigma_{ab}^2 + \frac{bn\sum_{i=1}^{a}\alpha_i^2}{a-1} \qquad E\{MSB\} = \sigma^2 + n\sigma_{ab}^2 + an\sigma_b^2$$

$$E\{MSAB\} = \sigma^2 + n\sigma_{ab}^2 \qquad E\{MSE\} = \sigma^2$$

Tests concerning interactions and main effects for the mixed model are carried out as follow. Note that expected mean squares for Factors A and B contain the interaction variance component. Thus, under their null hypotheses their expected mean squares simplify to $E\{MSAB\}$, which is why their F-tests use $MSAB$ as the error term.

1. $H_0^{AB} : \sigma_{ab}^2 = 0$ (No interaction effect).
2. $H_A^{AB} : \sigma_{ab}^2 > 0$ (Interaction effects exist)
3. T.S. $F_{AB} = \frac{MSAB}{MSE}$
4. R.R.: $F_{AB} \geq F_{\alpha,(a-1)(b-1),ab(n-1)}$
5. P-value: $P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$

Assuming no interaction effects exist, the test for differences among the effects of the levels of factor A as follows.

1. $H_0^{A} : \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A^{A} : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSAB}$
4. R.R.: $F_A \geq F_{\alpha,a-1,(a-1)(b-1)}$
5. P-value: $P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$

Assuming no interaction effects exist, we can test for differences among the effects of the levels of factor B as follows.

1. $H_0^{B} : \sigma_b^2 = 0$ (No factor B effect).
2. $H_A^{B} : \sigma_b^2 > 0$ (Factor B effects exist)
3. T.S. $F_B = \frac{MSB}{MSAB}$
4. R.R.: $F_B \geq F_{\alpha,b-1,(a-1)(b-1)}$
5. P-value: $P\left(F_{b-1,(a-1)(b-1)} \geq F_B\right)$

Unbiased (ANOVA) estimates of the variance components are obtained from the mean squares (see their expectations above). Note that these can be negative (except the estimate of the error variance).

$$s^2 = MSE \qquad s_{ab}^2 = \frac{MSAB - MSE}{n} \qquad s_b^2 = \frac{MSB - MSAB}{an}$$

Assuming the interaction is not significant, pairwise comparisons among levels of Factor A can be made based on simultaneous confidence intervals.

Bonferroni (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*,(a-1)(b-1)} \sqrt{MSAB\left(\frac{2}{bn}\right)},$$

Tukey's

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha,a,(a-1)(b-1)} \sqrt{MSAB\left(\frac{1}{bn}\right)}.$$

Example 3.11: Women's Professional Bowling Association Scores - 2009

This data comes from the qualifying rounds of the Women's Professional Bowling Association tournament held at Alan Park, Michigan in 2009. There were two factors: oil pattern (Factor A, fixed) and bowler (Factor B, random). There were $a = 4$ oil patterns used on the lane (1=Viper, 2=Chameleon, 3=Scorpion, 4=Shark). There were $b = 15$ bowlers (1=Diandra Abaty, 2=Shalin Zulkiffi, 3=Liz Johnson, 4=Kelly Kulick, 5=Clara Guerrero, 6=Jennifer Petrick, 7=Wendy MacPherson, 8=Shannon Pluhowski, 9=Stephanie Nation, 10=Tammy Boomershine, 11=Amanda Fagan, 12=Aumi Guerra, 13=Michelle Feldman, 14=Shannon O'Keefe, 15=Jodie Woessner). Each bowler bowled 2 sets of 7 games on each of the oil patterns. The response Y, is the score for an individual game for a bowler, making $n = 2(7) = 14$ replicates per oil pattern/bowler combination. The means (SDs) are given in Table 3.13 and the interaction plot is given in Figure 3.4. The Analysis of Variance and F-tests are given in Table 3.14.

The ANOVA results in significant main effects for oil pattern ($F_A = 3.573, p = .0217$) and for bowlers ($F_B = 2.611, p = .0082$), but not a significant interaction ($F_{AB} = 1.262, p = .1271$). Estimates of the variance components are given here.

$$s^2 = 649.6 \qquad s_{ab}^2 = \frac{819.6 - 649.6}{14} = 12.14 \qquad s_b^2 = \frac{2140.3 - 819.6}{4(14)} = 23.58$$

In particular, the estimate of the standard deviation of scores for the same bowler on the same oil pattern is $\sqrt{649.6} = 25.5$ and the standard deviation of the effects of the different bowlers is $\sqrt{23.58} = 4.86$.

Since there is no evidence of an oil pattern/bowler interaction, Bonferroni's and Tukey's methods can be used to obtain simultaneous Confidence Intervals to compare the $a = 4$ oil patterns (thus there are 4(3)/2=6 comparisons). The results are given in Table 3.15. The only significant difference is between oil patterns 1 (Viper) and 3 (Scorpion), all other simultaneous Confidence Intervals contain 0.

$$\text{Bonferroni:} \qquad t_{.05/(2(6)),42} = 2.769 \qquad \sqrt{819.6\left(\frac{2}{15(14)}\right)} = 2.794 \qquad 2.769(2.794) = 7.74$$

$$\text{Tukey:} \qquad q_{.05,4,42} = 3.783 \qquad \sqrt{819.6\left(\frac{1}{15(14)}\right)} = 1.976 \qquad 3.783(1.976) = 7.47$$

The output from R is given below, including likelihood based tests that give the correct F-statistics based on Estimated Generalized Least Squares. Note that the F-tests based on the **aov** function are incorrect, as they use MSE for the tests for the main effects of Factors A and B.

∇

Bowler	Viper	Chameleon	Scorpion	Shark	Mean
1	223.29 (42.44)	208.79 (22.98)	195.57 (27.27)	211.00 (28.85)	209.66
2	211.79 (28.77)	195.71 (27.06)	196.43 (24.31)	191.57 (19.77)	198.88
3	218.14 (24.51)	208.50 (26.23)	209.64 (27.78)	215.50 (14.09)	212.95
4	219.43 (18.72)	216.64 (21.74)	212.43 (27.52)	216.21 (35.62)	216.18
5	210.57 (20.77)	198.43 (18.8)	204.71 (28.7)	219.07 (24.82)	208.20
6	211.00 (30.91)	203.29 (16.9)	193.07 (23.72)	187.14 (30.82)	198.63
7	223.36 (34.26)	199.29 (30.29)	194.43 (20.24)	221.07 (22.84)	209.54
8	209.57 (25.17)	214.21 (31.64)	208.64 (20.76)	201.29 (25.94)	208.43
9	199.57 (27.98)	198.57 (21.67)	193.29 (18.57)	204.43 (26.78)	198.96
10	205.86 (33.02)	213.71 (20.73)	198.36 (31.23)	219.29 (27.84)	209.30
11	202.50 (26.88)	205.29 (14.49)	194.36 (21.4)	207.57 (17.67)	202.43
12	206.21 (31.98)	182.64 (25.8)	196.14 (16.73)	194.00 (26.59)	194.75
13	198.50 (15.51)	207.86 (22.39)	210.71 (25.88)	193.86 (33.64)	202.73
14	212.00 (30.03)	205.86 (25.49)	208.29 (22.43)	220.21 (13.01)	211.59
15	199.57 (27.98)	209.79 (23.93)	204.86 (20.27)	208.64 (13.62)	205.71
Mean	210.09	204.57	201.40	207.39	205.86

Table 3.13: Means (SDs) by Oil Pattern/Bowler combination - Women's Professional Bowlers Association 2009 Data

```
> wpba.mod1 <- aov(score ~ pattern + bowler + bowler:pattern)
> summary(wpba.mod1)
## This gives incorrect F-tests for Pattern and Bowler
               Df Sum Sq Mean Sq F value   Pr(>F)
pattern         3   8785  2928.4   4.508  0.00382 **
bowler         14  29965  2140.3   3.295 3.84e-05 ***
pattern:bowler  42  34423   819.6   1.262  0.12711
Residuals      780 506679   649.6

> wpba.mod2 <- aov(score ~ pattern + bowler + Error(bowler:pattern))
> summary(wpba.mod2)
## This gives correct F-tests for Pattern and Bowler
Error: bowler:pattern
          Df Sum Sq Mean Sq F value  Pr(>F)
```

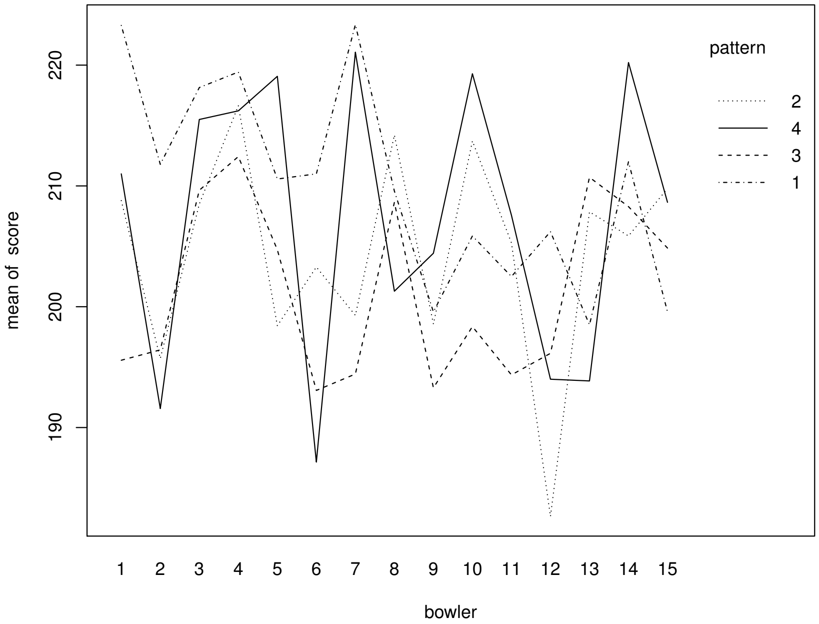

Figure 3.4: Interaction Plot of scores by Oil Pattern/Bowler - WPBA 2009 data

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(>F)$
Oil Pattern	$4-1=3$	8785	$\frac{8785}{3}=2928.4$	$\frac{2928.4}{819.6}=3.573$	.0217
Bowler	$15-1=14$	29965	$\frac{29965}{14}=2140.3$	$\frac{2140.3}{819.6}=2.611$	.0082
INTERACTION	$3(14)=42$	344232	$\frac{344232}{42}=819.6$	$\frac{819.6}{649.6}=1.262$	.1271
ERROR	$4(15)(14-1)=780$	506679	$\frac{506679}{780}$=649.6		
TOTAL	$4(15)(14)-1839$	579852			

Table 3.14: The Analysis of Variance Table for the WPBA 2009 data

i,i'	$\overline{y}_{i..}-\overline{y}_{i'..}$	Bonferroni CI	Tukey CI
1,2	$210.09-204.57=5.52$	$(-2.22,13.26)$	$(-1.95,12.99)$
1,3	$210.09-201.40=8.69$	$(0.95,16.43)$	$(1.22,16.16)$
1,4	$210.09-207.39=2.70$	$(-5.04,10.44)$	$(-4.77,10.17)$
2,3	$204.57-201.40=3.17$	$(-4.57,10.91)$	$(-4.30,10.64)$
2,4	$204.57-207.39=-2.82$	$(-10.56,4.92)$	$(-10.29,4.65)$
3,4	$201.40-207.39=-5.99$	$(-13.73,1.75)$	$(-13.46,1.48)$

Table 3.15: Simultaneous Confidence Intervals for Oil Patterns for the WPBA 2009 data

```
pattern    3   8785  2928.4   3.573 0.02172 *
bowler    14  29965  2140.3   2.611 0.00824 **
Residuals 42  34423   819.6

Error: Within
           Df Sum Sq Mean Sq F value Pr(>F)
Residuals 780 506679   649.6

> wpba.mod3 <- lme(fixed = score ~ pattern, random = ~1|bowler/pattern)
> summary(wpba.mod3)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  7851.303 7884.403 -3918.652

Random effects:
 Formula: ~1 | bowler
        (Intercept)
StdDev:    4.856495

 Formula: ~1 | pattern \%in\% bowler
        (Intercept) Residual
StdDev:    3.484903 25.48701

Fixed effects: score ~ pattern
                Value Std.Error  DF   t-value p-value
(Intercept) 205.86190  1.596276 780 128.96385  0.0000
pattern1      4.22857  1.710901  42   2.47155  0.0176
pattern2     -1.29048  1.710901  42  -0.75427  0.4549
pattern3     -4.46667  1.710901  42  -2.61071  0.0125

> intervals(wpba.mod3)
Approximate 95\% confidence intervals
 Fixed effects:
                  lower       est.      upper
(Intercept) 202.7283989 205.861905 208.995411
pattern1      0.7758329   4.228571   7.681310
pattern2     -4.7432148  -1.290476   2.162262
pattern3     -7.9194052  -4.466667  -1.013928

 Random Effects:
  Level: bowler
                   lower     est.    upper
sd((Intercept)) 2.627534 4.856495 8.976303
  Level: pattern
                   lower     est.   upper
sd((Intercept)) 1.221161 3.484903 9.94508
 Within-group standard error:
   lower     est.    upper
24.25312 25.48701 26.78367
> anova(wpba.mod3)
            numDF denDF   F-value p-value
(Intercept)     1   780 16631.675  <.0001
pattern         3    42     3.573  0.0217
>
> library(lmerTest)
> wpba.mod4 <- lmer(score~pattern+(1|bowler)+(1|pattern:bowler))
> summary(wpba.mod4)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: score ~ pattern + (1 | bowler) + (1 | pattern:bowler)
Random effects:
 Groups         Name        Variance Std.Dev.
 pattern:bowler (Intercept)  12.14    3.485
 bowler        (Intercept)  23.58    4.856
 Residual                  649.59   25.487
Number of obs: 840, groups:  pattern:bowler, 60; bowler, 15

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  205.862      1.596  14.000 128.966   <2e-16 ***
pattern1       4.229      1.711  42.000   2.472   0.0176 *
pattern2      -1.290      1.711  42.000  -0.754   0.4549
pattern3      -4.467      1.711  42.000  -2.611   0.0125 *
> anova(wpba.mod4)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
        Sum Sq Mean Sq NumDF DenDF F.value  Pr(>F)
pattern 6962.8  2320.9     3    42  3.5729 0.02172 *

> difflsmeans(wpba.mod4)
## No adjustment for simultaneous CIs
Differences of LSMEANS:
              Estimate Standard Error   DF t-value Lower CI Upper CI p-value
pattern 1 - 2      5.5           2.79 42.0    1.98   -0.119   11.157   0.055 .
pattern 1 - 3      8.7           2.79 42.0    3.11    3.057   14.334   0.003 **
pattern 1 - 4      2.7           2.79 42.0    0.97   -2.938    8.338   0.339
pattern 2 - 3      3.2           2.79 42.0    1.14   -2.462    8.814   0.262
pattern 2 - 4     -2.8           2.79 42.0   -1.01   -8.457    2.819   0.319
pattern 3 - 4     -6.0           2.79 42.0   -2.15  -11.633   -0.357   0.038 *
```

When there are 3 or more factors in a balanced design, the expected mean squares depend on which terms are fixed and which are random. The primary results are given in Table 3.16 for two cases, Case 1: A and B are fixed, and C is random, Case 2: A is fixed, and B and C are random. Note that when there is only one replicate per treatment ($n = 1$), MSE has no degrees of freedom and σ^2 cannot be independently estimated from $\sigma^2 + \sigma^2_{ABC}$ which is $E\{MSABC\}$. To simplify the table, the following notation is introduced.

$$A \text{ Fixed: } \theta_A^2 = \frac{\sum_{i=1}^a \alpha_i^2}{a-1} \qquad B \text{ Fixed: } \theta_B^2 = \frac{\sum_{j=1}^b \beta_j^2}{b-1} \qquad A,B \text{ Fixed: } \theta_{AB}^2 = \frac{\sum_{i=1}^a \sum_{j=1}^b (\alpha_i\beta_j)^2}{(a-1)(b-1)}$$

Source of Variation	Degrees of Freedom	A,B Fixed C Random	A Fixed B,C Random
A	$a-1$	$\sigma^2 + n\sigma^2_{ABC} + bn\sigma^2_{AC} + bcn\theta^2_A$	$\sigma^2 + n\sigma^2_{ABC} + bn\sigma^2_{AC} + cn\sigma^2_{AB} + bcn\theta^2_A$
B	$b-1$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC} + acn\theta^2_B$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC} + cn\sigma^2_{AB} + acn\sigma^2_B$
C	$c-1$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC} + bn\sigma^2_{AC} + abn\sigma^2_C$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC} + bn\sigma^2_{AC} + abn\sigma^2_C$
AB	$(a-1)(b-1)$	$\sigma^2 + n\sigma^2_{ABC} + cn\theta^2_{AB}$	$\sigma^2 + n\sigma^2_{ABC} + cn\sigma^2_{AB}$
AC	$(a-1)(c-1)$	$\sigma^2 + n\sigma^2_{ABC} + bn\sigma^2_{AC}$	$\sigma^2 + n\sigma^2_{ABC} + bn\sigma^2_{AC}$
BC	$(b-1)(c-1)$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC}$	$\sigma^2 + n\sigma^2_{ABC} + an\sigma^2_{BC}$
ABC	$(a-1)(b-1)(c-1)$	$\sigma^2 + n\sigma^2_{ABC}$	$\sigma^2 + n\sigma^2_{ABC}$
Error	$abc(n-1)$	σ^2	σ^2

Table 3.16: Expected Mean Squares for Balanced 3-Way Mixed Model

The goal is to isolate the variance components $\sigma^2_\bullet$ and the fixed effects components $\theta^2_\bullet$ for estimators and tests. Beginning with the bottom of the table, the following ANOVA estimators are obtained.

$$s^2 = MSE \qquad s^2_{ABC} = \frac{MSABC - MSE}{n} \qquad s^2_{BC} = \frac{MSBC - MSABC}{an} \qquad s^2_{AC} = \frac{MSAC - MSABC}{bn}$$

$$s_C^2 = \frac{MSC - MSAC - MSBC + MSABC}{abn}$$

$$\text{Case2: } s_{AB}^2 = \frac{MSAB - MSABC}{cn} \qquad s_B^2 = \frac{MSB - MSAB - MSBC + MSABC}{acn}$$

When tests have single numerator and denominator mean squares, the F-tests are obtained directly as in the 2-factor Mixed model. In other cases, combinations of mean squares are needed, and degrees of freedom must be estimated, one such method is **Satterthwaite's approximation**. Consider Factor A for each Case.

$$\text{Case1: } E\{MSA\} = \sigma^2 + n\sigma_{ABC}^2 + bn\sigma_{AC}^2 + bcn\theta_A^2 \quad E\{MSAC\} = \sigma^2 + n\sigma_{ABC}^2 + bn\sigma_{AC}^2$$

$$\text{Case2: } E\{MSA\} = \sigma^2 + n\sigma_{ABC}^2 + bn\sigma_{AC}^2 + cn\sigma_{AB}^2 + bcn\theta_A^2$$

$$E\{MSAB\} + E\{MSAC\} - E\{MSA\} = \sigma^2 + n\sigma_{ABC}^2 + bn\sigma_{AC}^2 + cn\sigma_{AB}^2$$

Thus to isolate θ_A^2 for Case 1, only $MSAC$ is needed. However for Case 2, $MSAB + MSAC - MSABC$ is needed. A direct F-test can be used for Case 1, while a "synthetic" F-test is needed for Case 2. The approximate degrees of freedom based on Satterthwaite's approximation can be obtained as follows for a linear function of Mean Squares.

$$MS_\bullet = g_1 MS_1 + \cdots + g_m MS_m \quad \Rightarrow \quad df_\bullet \approx \frac{(MS_\bullet)^2}{\frac{(g_1 MS_1)^2}{df_1} + \cdots + \frac{(g_m MS_m)^2}{df_m}}$$

These calculations will be performed based on two examples given. The first example considered spatula performance by cooks, which has spatula length and angle (both fixed factors) and subject utilizing the spatula (random). The second study involves measurement reliability of foot inversion with a Phillips biometer by testers (treated as fixed) for 12 subjects on 2 days, with 2 replicates per day. The first study has one replicate per treatment (which is common in higher order studies), so that σ^2 and σ_{ABC}^2 cannot be estimated independently (without assuming the 3-way interaction variance component is 0). However, the main effects and 2-way interactions can still be analyzed.

Example 3.12: Shoveling Times for Spatulas

An ergonomic experiment was conducted to compare spatula length and angles on food shoveling times (Wu and Hsieh, 2002, [37]). There were 4 spatula lengths (20, 25, 30, 40cm), and 4 angles (15, 25, 35, 45 degrees). There were 8 chefs, who each used each of the 16 combinations once. The spatula lengths and angles are fixed factors, and the chefs (subjects) are treated as random. The response was shoveling time for 2000 grams of green beans from pan to plate. The analysis of variance is given in Table 3.17.

Estimated variance components are given below.

$$s^2 + s_{LAC}^2 = 13.08 \qquad s_{AC}^2 = \frac{13.15 - 13.08}{4(1)} = 0.0175 \qquad s_{LC}^2 = \frac{15.63 - 13.08}{4(1)} = 0.6375$$

$$s_C^2 = \frac{861.46 - 15.63 - 13.15 + 13.08}{4(4)(1)} = 52.86$$

Tests for fixed effects are given below.

$$H_0^{LA} : (\alpha\beta)_{11} = \cdots = (\alpha\beta)_{44} = \theta_{LA}^2 = 0 \qquad TS : F_{LA} = \frac{MSLA}{MSLAC} = \frac{12.59}{13.08} = 0.963$$

$$RR : F_{LA} \geq F_{.05,9,63} = 2.032 \qquad P = P\left(F_{9,63} \geq 0.963\right) = .4788$$

$$H_0^{L} : \alpha_1 = \cdots = \alpha_4 = \theta_L^2 = 0 \qquad TS : F_L = \frac{MSL}{MSLC} = \frac{325.31}{15.63} = 20.81 \qquad RR : F_L \geq F_{.05,3,21} = 3.072$$

$$P = P\left(F_{3,21} \geq 20.81\right) < .0001$$

$$H_0^{A} : \beta_1 = \cdots = \beta_4 = \theta_A^2 = 0 \quad TS : F_A = \frac{MSA}{MSAC} = \frac{289.15}{13.15} = 21.99 \quad RR : F_A \geq F_{.05,3,21} = 3.072$$

$$P = P\left(F_{3,21} \geq 21.99\right) < .0001$$

The evidence is for important main effects for Length, Angle, and Chef with no evidence of any important interactions. Table 3.18 gives the comparisons among pairs of Lengths and Angles (with added Bonferroni adjusted P-values). The Likelihood Ratio tests for the variance components (which are different from the F-test from the ANOVA) are given below. Each of these is a chi-square statistic with 1 degree of freedom $\left(\chi^2_{.05,1} = 3.841\right)$. A portion of the extensive R output is given below.

$$H_0^C : \sigma_C^2 = 0 \quad TS : X_C^2 = 46.80 \qquad H_0^{LC} : \sigma_{LC}^2 = 0 \quad TS : X_{LC}^2 = 0.273 \qquad H_0^{AC} : \sigma_{AC}^2 = 0 \quad TS : X_{AC}^2 = 0.00024$$

∇

Source	Df	Sum Sq	Mean Sq
Length	3	975.9	325.31
Angle	3	867.4	289.15
Chef	7	6030.2	861.46
Length:Angle	9	113.3	12.59
Length:Chef	21	328.3	15.63
Angle:Chef	21	276.1	13.15
Length:Angle:Chef	63	823.8	13.08
Residuals	0	0.0	

Table 3.17: Analysis of Variance for Spatula experiment

```
> spat.mod2 <- lmer(shovtime ~ length*angle + (1|subject) + (1|subject:length) +
+              (1|subject:angle))
> summary(spat.mod2)
Random effects:
 Groups         Name        Variance Std.Dev.
```

Differences of LSMEANS								
Comparison	Estimate	Std Err	DF	t-value	Lo Bnd	Hi Bnd	P	Adj P
length 1 - 2	0.5	0.988	21.0	0.55	-1.511	2.6006	0.587	1.0000
length 1 - 3	-1.4	0.988	21.0	-1.44	-3.481	0.6306	0.164	0.987
length 1 - 4	-6.5	0.988	21.0	-6.53	-8.506	-4.3944	$< 2e-16$	$< .0001$
length 2 - 3	-2.0	0.988	21.0	-1.99	-4.026	0.0856	0.059	.3586
length 2 - 4	-7.0	0.988	21.0	-7.08	-9.051	-4.9394	$< 2e-16$	$< .0001$
length 3 - 4	-5.0	0.988	21.0	-5.08	-7.081	-2.9694	$< 2e-16$	.0003
angle 1 - 2	3.7	0.906	21.0	4.12	1.847	5.6177	$5e-04$	.0029
angle 1 - 3	2.0	0.906	21.0	2.24	0.142	3.9127	0.036	.2162
angle 1 - 4	-3.3	0.906	21.0	-3.60	-5.145	-1.3748	0.002	.0101
angle 2 - 3	-1.7	0.906	21.0	-1.88	-3.590	0.1802	0.074	.4443
angle 2 - 4	-7.0	0.906	21.0	-7.71	-8.878	-5.1073	$< 2e-16$	$< .0001$
angle 3 - 4	-5.3	0.906	21.0	-5.83	-7.173	-3.4023	$< 2e-16$	$< .0001$

Table 3.18: Pairwise comparisons among Lengths and Angles for Spatula experiment

```
 subject:angle  (Intercept)  0.01819 0.1349
 subject:length (Intercept)  0.63906 0.7994
 subject        (Intercept) 52.85968 7.2705
 Residual                   13.07619 3.6161

> anova(spat.mod2)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
             Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
length       816.35 272.116     3    21 20.8101 1.701e-06 ***
angle        862.64 287.546     3    21 21.9900 1.106e-06 ***
length:angle 113.28  12.587     9    63  0.9626    0.4791

> difflsmeans(spat.mod2)
Differences of LSMEANS:
                         Estimate Standard Error    DF t-value Lower CI Upper CI p-value
length 1 - 2                  0.5          0.988  21.0    0.55   -1.511   2.6006   0.587
length 1 - 3                 -1.4          0.988  21.0   -1.44   -3.481   0.6306   0.164
length 1 - 4                 -6.5          0.988  21.0   -6.53   -8.506  -4.3944  <2e-16
length 2 - 3                 -2.0          0.988  21.0   -1.99   -4.026   0.0856   0.059
length 2 - 4                 -7.0          0.988  21.0   -7.08   -9.051  -4.9394  <2e-16
length 3 - 4                 -5.0          0.988  21.0   -5.08   -7.081  -2.9694  <2e-16
angle 1 - 2                   3.7          0.906  21.0    4.12    1.847   5.6177   5e-04
angle 1 - 3                   2.0          0.906  21.0    2.24    0.142   3.9127   0.036
angle 1 - 4                  -3.3          0.906  21.0   -3.60   -5.145  -1.3748   0.002
angle 2 - 3                  -1.7          0.906  21.0   -1.88   -3.590   0.1802   0.074
angle 2 - 4                  -7.0          0.906  21.0   -7.71   -8.878  -5.1073  <2e-16
angle 3 - 4                  -5.3          0.906  21.0   -5.83   -7.173  -3.4023  <2e-16
```

Example 3.13: Reliability of Foot Joint Inversion Measurements

A study reported results of a reliability experiment measuring foot joint inversion and eversion using a Phillips biometer (Freeman, el al, 2007, [14]). This analysis is based on the inversion measurement (angle). There were 2 testers, for the course of this analysis they are treated as fixed (for instance they may be the only two people trained at a medical unit to operate the machine). There were 12 subjects who were each measured twice ($n = 2$) on each of 2 days by the 2 testers. Subjects and days are treated as random. The Analysis of Variance is given in Table 3.19. The "Inv" label on the variables refers to the fact that these are only the inversion (not the eversion) measurements.

ANOVA estimates of the variance components are given below, with tester having 2 levels, subject having 12, and day having 2, with $n = 2$ replicates per treatment.

$$s^2 = 0.323 \qquad s^2_{TSD} = \frac{0.783 - 0.323}{2} = 0.230 \qquad s^2_{TS} = \frac{10.366 - 0.783}{2(2)} = 2.396$$

$$s^2_{TD} = \frac{1.76 - 0.783}{12(2)} = 0.041 \qquad s^2_{SD} = \frac{3.094 - 0.783}{2(2)} = 0.578$$

$$s^2_S = \frac{173.048 - 10.366 - 3.094 + 0.783}{2(2)(2)} = 20.046 \qquad s^2_D = \frac{11.344 - 1.76 - 3.094 + 0.783}{2(12)(2)} = 0.152$$

The F-test for testing for Tester effects is conducted as follows.

$$H_0^T : \alpha_1 = \alpha_2 = \theta_T^2 = 0 \quad TS : F_T = \frac{MST}{MSTS + MSTD - MSTSD} = \frac{0.844}{10.366 + 1.76 - 0.783} = \frac{0.844}{11.343} = 0.074$$

$$df_1 = df_T = 1 \quad df_2 = \frac{(11.343)^2}{\left[\frac{(10.366)^2}{11} + \frac{(1.76)^2}{1} + \frac{(-0.783)^2}{11}\right]} = 9.957$$

$$RR : F_T \geq F_{.05,1,9.957} = 4.970 \qquad P = P\left(F_{1,9.957} \geq 0.074\right) = .7912$$

There is no evidence of differences in the Testers' means (which is good from a clinical standpoint). The output from the R program is given below. Note that if any of the ANOVA estimates of the variance components had been negative, the REML estimates of the variance components and the F-test would not have been the same.

∇

Source	Df	Sum Sq	Mean Sq
testerInv	1	0.84	0.844
subjInv	11	1903.53	173.048
dayInv	1	11.34	11.344
testerInv:subjInv	11	114.03	10.366
testerInv:dayInv	1	1.76	1.76
subjInv:dayInv	11	34.03	3.094
testerInv:subjInv:dayInv	11	8.61	0.783
Residuals	48	15.5	0.323

Table 3.19: Analysis of Variance for foot joint inversion measurement experiment

```
> foot.mod2 <- lmer(angleInv ~ testerInv + (1|subjInv) + (1|dayInv) +
+  (1|testerInv:subjInv) + (1|testerInv:dayInv) + (1|subjInv:dayInv) +
+  (1|testerInv:subjInv:dayInv))
> summary(foot.mod2)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
```

```
Random effects:
 Groups                   Name        Variance Std.Dev.
 testerInv:subjInv:dayInv (Intercept)  0.23011 0.4797
 subjInv:dayInv           (Intercept)  0.57765 0.7600
 testerInv:subjInv        (Intercept)  2.39583 1.5478
 subjInv                  (Intercept) 20.04639 4.4773
 testerInv:dayInv         (Intercept)  0.04072 0.2018
 dayInv                   (Intercept)  0.15152 0.3892
 Residual                              0.32292 0.5683

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept) 19.34375    1.37424 11.52600  14.076 1.29e-08 ***
testerInv1   0.09375    0.34375  9.95600   0.273    0.791

> anova(foot.mod2)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
            Sum Sq  Mean Sq NumDF  DenDF F.value Pr(>F)
testerInv 0.024019 0.024019     1 9.9566 0.07438 0.7906

> difflsmeans(foot.mod2)
Differences of LSMEANS:
                Estimate Standard Error   DF t-value Lower CI Upper CI p-value
testerInv 1 - 2      0.2          0.688 10.0    0.27    -1.35     1.72     0.8
> rand(foot.mod2)
Analysis of Random effects Table:
                         Chi.sq Chi.DF p.value
subjInv                  28.986      1   7e-08 ***
dayInv                    0.461      1    0.50
testerInv:subjInv        16.302      1   5e-05 ***
testerInv:dayInv          0.377      1    0.54
subjInv:dayInv            4.831      1    0.03 *
testerInv:subjInv:dayInv  4.158      1    0.04 *
```

3.1.3 Random Effects Models

When both Factors A and B are random, it is referred to as a **Random Effects Model**. The computation of the sums of squares in the Analysis of Variance is the same, but the tests for treatment effects change. The model is written as follows.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

where μ is the overall mean, α_i is the (random) effect of the i^{th} level of factor A, β_j is the (random) effect of the j^{th} level of factor B, $(\alpha\beta)_{ij}$ is the (random) interaction of factor A at level i and factor B at level j. This model is parameterized assuming the following model structure.

$$\alpha_i \sim N\left(0, \sigma_A^2\right) \qquad \beta_j \sim N\left(0, \sigma_B^2\right) \qquad (\alpha\beta)_{ij} \sim N\left(0, \sigma_{AB}^2\right) \qquad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

All random effects and error terms are assumed mutually independent in this formulation. The Analysis of Variance for the random effects model is given in Table 3.20. Data are no longer assumed independent when they are measured on the same level of a random factor.

$$E\left\{Y_{ijk}\right\} = \mu \qquad V\left\{Y_{ijk}\right\} = \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma^2$$

$$\text{COV}\left\{Y_{ijk}, Y_{i'j'k'}\right\} = \begin{cases} \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma^2 & : \quad i = i', j = j', k = k' \\ \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 & : \quad i = i', j = j', k \neq k' \\ \sigma_A^2 & : \quad i = i', j \neq j', \forall k, k' \\ \sigma_B^2 & : \quad i \neq i', j = j', \forall k, k' \\ 0 & : \quad \text{otherwise} \end{cases}$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$Pr(> F)$
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSAB}$	$P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$
B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$	$F_B = \frac{MSB}{MSAB}$	$P\left(F_{b-1,(a-1)(b-1)} \geq F_B\right)$
AB	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	$F_{AB} = \frac{MSAB}{MSE}$	$P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$
ERROR	$ab(n-1)$	SSE	$MSE = \frac{SSE}{ab(n-1)}$		
TOTAL	$abn-1$	TSS			

Table 3.20: The Analysis of Variance Table for a 2-Factor Factorial Design - Factors A and B random

The expectations of the mean squares for the random effects model (with A and B random) are given below.

$$E\{MSA\} = \sigma^2 + n\sigma_{AB}^2 + bn\sigma_A^2 \qquad E\{MSB\} = \sigma^2 + n\sigma_{AB}^2 + an\sigma_B^2$$

$$E\{MSAB\} = \sigma^2 + n\sigma_{AB}^2 \qquad E\{MSE\} = \sigma^2$$

Tests concerning interactions and main effects for the random model are carried out as follow. Note that expected mean squares for Factors A and B contain the interaction variance component. Thus, under their null hypotheses their expected mean squares simplify to $E\{MSAB\}$, which is why their F-tests use $MSAB$ as the error term.

1. $H_0^{AB} : \sigma_{AB}^2 = 0$ (No interaction effect).
2. $H_A^{AB} : \sigma_{AB}^2 > 0$ (Interaction effects exist)
3. T.S. $F_{AB} = \frac{MSAB}{MSE}$
4. R.R.: $F_{AB} \geq F_{\alpha,(a-1)(b-1),ab(n-1)}$
5. P-value: $P\left(F_{(a-1)(b-1),ab(n-1)} \geq F_{AB}\right)$

The test for differences among the effects of the levels of factor A as follows.

1. $H_0^A : \sigma_A^2 = 0$ (No factor A effect).

2. $H_A^A : \sigma_A^2 > 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSAB}$
4. R.R.: $F_A \geq F_{\alpha,a-1,(a-1)(b-1)}$
5. P-value: $P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$

The test for differences among the effects of the levels of factor B as follows.

1. $H_0^B : \sigma_B^2 = 0$ (No factor B effect).
2. $H_A^B : \sigma_B^2 > 0$ (Factor B effects exist)
3. T.S. $F_B = \frac{MSB}{MSAB}$
4. R.R.: $F_B \geq F_{\alpha,b-1,(a-1)(b-1)}$
5. P-value: $P\left(F_{b-1,(a-1)(b-1)} \geq F_B\right)$

Unbiased (ANOVA) estimates of the variance components are obtained from the mean squares (see their expectations above). Note that these can be negative (except the estimate of the error variance).

$$s^2 = MSE \qquad s_{AB}^2 = \frac{MSAB - MSE}{n} \qquad s_A^2 = \frac{MSA - MSAB}{bn} \qquad s_B^2 = \frac{MSB - MSAB}{an}$$

Example 3.14: Repeatability and Reproducibility of Measurements

In engineering, focus is often on accuracy of measurements and variability of products being manufactured. One factor is the Product, and the other is the Operator (or possibly Machine) that measures the product. The studies are referred to as Gage Repeatability and Reproducibility (GR&R)experiments. The same operator/machine will measure the same product multiple times (in random order). In these studies, σ^2 is referred to as the repeatability variance, and the sum of the operator and product/operator interaction variance is referred to as the reproducibility variance. The sum of the repeatability and reproducibility variances is called the gage variance.

$$\sigma_{\text{Total}}^2 = \sigma_P^2 + \sigma_O^2 + \sigma_{OP}^2 + \sigma^2 \qquad \sigma_{\text{Repeatability}}^2 = \sigma^2 \qquad \sigma_{\text{Reproducibility}}^2 = \sigma_O^2 + \sigma_{OP}^2$$

$$\sigma_{\text{Gage}}^2 = \sigma_{\text{Reproducibility}}^2 + \sigma_{\text{Repeatability}}^2$$

An experiment was conducted (Li and Al-Refaie, 2008, [19]) for measuring diameter of $p = 10$ drilled holes in wood (treated as "parts") by $o = 3$ operators, there were $n = 3$ measurements for each part by each operator. The means for all combinations of part and operator are given in Table 3.21 and the Analysis of Variance is given in Table 3.22. Measurements have been multiplied by 100 for ease of viewing calculations. Based on the F-tests, the interaction has a P-value of .0703 and for operator, $P = .0539$; while the part has a large F-statistic, with a very small P-value. The estimated variances are given below. The variance in parts accounts for about 77% of the total variation in measurements (30.8188/39.9114). R output that estimates the variance components is given below.

$$s^2 = 6.6778 \quad s^2_{OP} = \frac{11.1827 - 6.6778}{3} = 1.5016 \quad s^2_P = \frac{288.579 - 11.1827}{3(3)} = 30.8188$$

$$s^2_O = \frac{38.5778 - 11.1827}{10(3)} = 0.9132 \qquad s^2_{\text{Repeatability}} = 6.6778$$

$$s^2_{\text{Reproducibility}} = 0.9132 + 1.5016 = 2.4148 \qquad s^2_{\text{Total}} = 6.6778 + 2.4148 + 30.8188 = 39.9114$$

∇

	operator1	operator2	operator3	Mean
part1	2559.33	2559.00	2559.33	2559.22
part2	2561.00	2564.00	2565.67	2563.56
part3	2560.00	2560.67	2564.33	2561.67
part4	2560.67	2559.67	2561.33	2560.56
part5	2557.33	2564.33	2565.00	2562.22
part6	2569.00	2566.33	2565.00	2566.78
part7	2562.67	2563.33	2567.00	2564.33
part8	2569.00	2569.33	2569.33	2569.22
part9	2547.67	2546.67	2550.00	2548.11
part10	2557.67	2561.67	2560.00	2559.78
Mean	2560.43	2561.50	2562.70	2561.54

Table 3.21: Means by Part (Drill Hole)/Operator combination - Gage R&R experiment

Source	df	SS	MS	F	F(.05)	$P(>F)$
Part	9	2597.211	288.5790	25.8058	2.4563	$<.0001$
Operator	2	77.1556	38.5778	3.4498	3.5546	0.0539
P*O	18	201.2889	11.1827	1.6746	1.7784	0.0703
Error	60	400.6667	6.6778			
Total	89	3276.322				

Table 3.22: The Analysis of Variance Table for the Gage R&R experiment

```
> wd.mod3 <- lmer(Ymeas ~ 1 + (1|Part) + (1|Operator) + (1|Part:Operator))
> summary(wd.mod3)
summary from lme4 is returned
some computational error has occurred in lmerTest
Linear mixed model fit by REML ['lmerMod']
Formula: Ymeas ~ 1 + (1 | Part) + (1 | Operator) + (1 | Part:Operator)

Random effects:
 Groups        Name        Variance Std.Dev.
 Part:Operator (Intercept)  1.5016  1.2254
 Part          (Intercept) 30.8218  5.5517
 Operator      (Intercept)  0.9132  0.9556
 Residual                  6.6778  2.5841
Number of obs: 90, groups:  Part:Operator, 30; Part, 10; Operator, 3

Fixed effects:
            Estimate Std. Error t value
(Intercept) 2561.544      1.874    1367
```

The case of the 3-Way Random Effects model is very similar to the 3-way Mixed Effects model with two random factors with σ_A^2 replacing θ_A^2 in Table 3.16. Then all variance component estimates are obtained in a similar manner as was done in Example 3.13. Note that the ANOVA estimate for the (random) Tester effect would have been negative. This is consistent with the very small F-statistic ($F_T < 1$) for Testers in Example 3.13.

$$s_T^2 = \frac{MST - MSTS - MSTD + MSTSD}{bcn} = \frac{0.844 - 10.366 - 1.76 + 0.783}{12(2)(2)} = -0.219$$

3.2 Nested Designs

In some designs, one factor's levels are nested within levels of another factor. Thus, the levels of Factor B that are exposed to one level of Factor A are different from those that receive a different level of Factor A. There will be n replications under each "combination" of factor levels in balanced designs. The statistical model is written as follows.

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk} \quad i = 1, \ldots, a \quad j(i) = 1, \ldots, b_i \quad k = 1, \ldots, n \qquad \epsilon \sim N\left(0, \sigma^2\right)$$

Here μ is the overall mean, α_i is the effect of the i^{th} level of Factor A, $\beta_{j(i)}$ is the effect of the j^{th} level of Factor B nested under the i^{th} level of Factor A, and ϵ_{ijk} is the random error term. In general, there will be a levels for Factor A, b_i levels of Factor B within the i^{th} level of Factor A, and n replicates per cell. In practice, Factor A will be fixed or random, and Factor B will be either fixed or random. In any event, the Analysis of Variance is the same, and is obtained as follows, once the data have been observed.

$$\begin{aligned}
\overline{y}_{ij.} &= \frac{\sum_{k=1}^{n} y_{ijk}}{n} \\
\overline{y}_{i..} &= \frac{\sum_{j=1}^{b_i} \sum_{k=1}^{n} y_{ijk}}{b_i n} \\
N &= n \sum_{i=1}^{a} b_i \\
\overline{y}_{...} &= \frac{\sum_{i=1}^{a} \sum_{j=1}^{b_i} \sum_{k=1}^{n} y_{ijk}}{N} \\
TSS &= \sum_{i=1}^{a} \sum_{j=1}^{b_i} \sum_{k=1}^{n} \left(y_{ijk} - \overline{y}_{...}\right)^2
\end{aligned}$$

$$\begin{aligned}
SSA &= n\sum_{i=1}^{a} b_i\left(\overline{y}_{i..}-\overline{y}_{...}\right)^2 \\
SSB(A) &= n\sum_{i=1}^{a}\sum_{j=1}^{b_i}\left(\overline{y}_{ij.}-\overline{y}_{i..}\right)^2 \\
SSE &= \sum_{i=1}^{a}\sum_{j=1}^{b_i}\sum_{k=1}^{n}\left(y_{ijk}-\overline{y}_{ij.}\right)^2
\end{aligned}$$

3.2.1 Factors A and B Fixed

In the case where both A and B are fixed factors, the effects are fixed (unknown) constants, and the following assumptions are made.

$$\sum_{i=1}^{a}\alpha_i = 0 \qquad \sum_{j=1}^{b_i}\beta_{j(i)} = 0 \quad \forall i \qquad \epsilon_{ijk}\sim N\left(0,\sigma_e^2\right)$$

The Analysis of Variance when both factors A and $B(A)$ are fixed is given in Table 3.23, where $b_. = \sum_{i=1}^{a} b_i$.

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-Value
A	$a-1$	SSA	$MSA=\frac{SSA}{a-1}$	$F_A=\frac{MSA}{MSE}$	$P\left(F_{a-1,b_.(n-1)}\geq F_A\right)$
$B(A)$	$b_.-a$	$SSB(A)$	$MSB(A)=\frac{SSB(A)}{b_.-a}$	$F_{B(A)}=\frac{MSB(A)}{MSE}$	$P\left(F_{b_.-a,b_.(n-1)}\geq F_{B(A)}\right)$
ERROR	$b_.(n-1)$	SSE	$MSE=\frac{SSE}{b_.(r-1)}$		
TOTAL	$nb_.-1$	TSS			

Table 3.23: The Analysis of Variance Table for a 2-Factor Nested Design – A and B Fixed Factors

The tests for effects of factors A and $B(A)$ involve the two F–statistics, and are conducted as follow. The test for differences among the effects of the levels of factor A is as follows.

1. $H_0^A:\alpha_1=\cdots=\alpha_a=0$ (No factor A effect).
2. H_A^A : Not all $\alpha_i=0$ (Factor A effects exist)
3. T.S. $F_A=\frac{MSA}{MSE}$
4. R.R.: $F_A\geq F_{\alpha,a-1,b_.(n-1)}$
5. P-value: $P\left(F_{a-1,b_.(n-1)}\geq F_A\right)$

The test for differences among the effects of the levels of factor $B(A)$ is as follows.

1. $H_0^{B(A)} : \beta_{1(1)} = \ldots = \beta_{b_a(a)} = 0$ (No factor B effect).

2. $H_A^{B(A)}$: Not all $\beta_{j(i)} = 0$ (Factor B effects exist)

3. T.S. $F_{B(A)} = \frac{MSB(A)}{MSE}$

4. R.R.: $F_{B(A)} \geq F_{\alpha,(b-1),b_.(r-1)}$

5. P-value: $P\left(F_{b_.-a,b_.(n-1)} \geq F_{B(A)}\right)$

Pairwise comparisons among levels of Factor A are based on constructing simultaneous confidence intervals as follow.

Bonferroni (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*,b_.(n-1)}\sqrt{MSE\left(\frac{1}{nb_i} + \frac{1}{nb_{i'}}\right)},$$

Tukey

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha,a,b_.(n-1)}\sqrt{\frac{MSE}{2}\left(\frac{1}{nb_i} + \frac{1}{nb_{i'}}\right)}$$

To compare levels of Factor B under a particular level of Factor A, simultaneous confidence intervals are constructed as follow.

Bonferroni (with $c_{B_i}^* = b_i\left(b_i - 1\right)/2$):

$$(\overline{y}_{ij.} - \overline{y}_{ij'.}) \pm t_{\alpha/2c_{B_i}^*,b_.(n-1)}\sqrt{MSE\left(\frac{2}{n}\right)},$$

Tukey

$$(\overline{y}_{ij.} - \overline{y}_{ij'.}) \pm q_{\alpha,b_i,b_.(n-1)}\sqrt{MSE\left(\frac{1}{n}\right)}$$

When entering the data for the Factor B levels, it is helpful to make the levels distinct such as $1, \ldots, b_1, b_1 + 1, \ldots, b_1 + b_2, \ldots, b_.$, not as $1, \ldots, b_1, 1, \ldots, b_2, \ldots, 1, \ldots, b_a$. The latter way is more useful as a generic notation.

Example 3.15: Measurement of Alcohol Content in Distilled Beverages

An experiment was conducted to measure alcohol content in distilled beverages by thermal infrared enthalpimetry (Oliveira, et al, 2017, [27]). The response was the difference between the measured alcohol content and the amount stated on the label. There were $a = 3$ types of distilled spirits (Vodka ($i = 1$), Whiskey ($i = 2$), Cachaca ($i = 3$)), and within each type of distilled spirits there were $b = 3$ brands (the brands within Vodka are different than those within Whiskey and Cachaca, and so on). For each brand there were $n = 24$ measurements. For the purposes of this analysis, the assumption is that these are the only 3 types and the only 3(3)=9 brands of interest to the researchers. Data have been simulated to match the means and standard deviations of the alcohol contents in the paper, the units are percent alcohol. The summary statistics are given in Table 3.24. The sums of squares can be obtained as follow. Notice that for $SSB(A)$, deviations for the brand means from the spirit type means are obtained, not from the overall mean. The Analysis of Variance is given in Table 3.25.

$$SSA = 24(3)\left[((-0.304)-(-0.219))^2 + ((-0.306)-(-0.219))^2 + ((-0.047)-(-0.219))^2\right] = 72(0.0444) = 3.1968$$

$$SSB(A) = 24\left[((-0.431)-(-0.304))^2 + \cdots + ((-0.130)-(-0.047))^2\right] = 24(.1703) = 4.0872$$

$$SSE = (24-1)\left[0.330^2 + \cdots + 0.410^2\right] = 23(1.0843) = 24.9389$$

$$df_A = 3-1 = 2 \quad MSA = \frac{3.1968}{2} = 1.5984 \qquad df_{B(A)} = 3(3-1) = 6 \quad MSB(A) = \frac{4.0872}{6} = 0.6812$$

$$df_E = 3(3)(24-1) = 207 \quad MSE = \frac{24.9389}{207} = 0.1205$$

From the ANOVA table, it is clear that there are significant Spirit Type ($F_A = 13.27, P < .0001$) and Brand within Spirit Type effects $\left(F_{B(A)} = 5.65, P < .0001\right)$. Pairwise comparisons among Types and Brands within Types are obtained as follow.

Bonferroni for Type: $C_A^* = 3(3-1)/2 = 3$

$$t_{.05/(2(3)),207} = 2.414 \qquad \sqrt{0.1205\left(\frac{1}{3(24)} + \frac{1}{3(24)}\right)} = 0.0579 \qquad 2.414(.0579) = 0.140$$

Tukey for Type: $a = 3$

$$q_{.05,3,207} = 3.339 \qquad \frac{3.339}{\sqrt{2}}0.0579 = 0.137$$

Since the Brands within Types have the same number of levels and use the same error term, the Bonferroni and Tukey critical values are the same for making pairwise comparisons among them as for the Spirit Types. Vodka and Whiskey are both significantly lower than Cachaca. Within Vodka, Brand 1 is significantly lower than Brands 2 and 3; within Whiskey, Brand 1 is significantly lower than Brands 2 and 3; and within Cachaca, Brands 1 and 3 are significantly lower than Brand 2. The R output is given below.

∇

```
> anova(dist.mod1)
Analysis of Variance Table

Response: Y
```

Type	Vodka ($i = 1$)	Whiskey ($i = 2$)	Cachaca ($i = 3$)	
Brand1	-0.431 (0.330)	-0.589 (0.340)	-0.090 (0.339)	
Brand2	-0.240 (0.279)	-0.139 (0.340)	0.079 (0.420)	
Brand3	-0.240 (0.280)	-0.191 (0.359)	-0.130 (0.410)	
Mean	-0.304	-0.306	-0.047	-0.219

Table 3.24: Means (SDs) by Spirit Type and Brand within Type

ANOVA

Source of	Degrees of	Sum of	Mean		
Variation	Freedom	Squares	Square	F	P-Value
Spirit Type	2	3.1968	$\frac{3.1968}{2} = 1.5984$	$1.59840.1205 = 13.2747$	$< .0001$
Brand(Type)	6	4.0872	$\frac{4.0872}{6} = 0.6812$	$F_{B(A)} = \frac{0.6812}{0.1205} = 5.6531$	$< .0001$
ERROR	207	24.9389	$\frac{24.9389}{207} = 0.1205$		
TOTAL	215	32.2229			

Table 3.25: The Analysis of Variance Table for the Alcohol Content in Distilled Spirits experiment – A and B Fixed Factors

```
                     Df  Sum Sq Mean Sq F value    Pr(>F)
spiritType            2  3.1967 1.59834 13.2669 3.791e-06 ***
spiritType:brandSprt  6  4.0872 0.68120  5.6543 1.850e-05 ***
Residuals           207 24.9385 0.12048


> TukeyHSD(dist.mod1,"spiritType")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ spiritType + spiritType/brandSprt)
$spiritType
         diff        lwr       upr     p adj
2-1 -0.0025000 -0.1390648 0.1340648 0.9989709
3-1  0.2568056  0.1202407 0.3933704 0.0000435
3-2  0.2593056  0.1227407 0.3958704 0.0000362
```

3.2.2 Factor A Fixed and B Random

In the case where A is fixed and B is random, the effects for levels of Factor A are fixed (unknown) constants, the effects of levels of Factor B are random variables, and the following assumptions are made.

$$\sum_{i=1}^{a} \alpha_i = 0 \qquad \beta_{j(i)} \sim N\left(0, \sigma^2_{B(A)}\right) \qquad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

Further, assume all random effects of levels of Factor $B(A)$ and all random error terms are mutually independent. The sums of squares are the same as in the previous subsection, but the error term for Factor A changes. The Analysis of Variance when Factor A is fixed and $B(A)$ is random is given in Table 3.26, where $b_. = \sum_{i=1}^{a} b_i$. The covariance structure and Expected Mean Squares for the factors are given below.

$$E\{Y_{ijk}\} = \mu + \alpha_i \qquad V\{Y_{ijk}\} = \sigma^2_{B(A)} + \sigma^2$$

$$\text{COV}\{Y_{ijk}, Y_{i'j'k'}\} = \begin{cases} \sigma^2_{B(A)} + \sigma^2 & : \; i=i', j=j', k=k' \\ \sigma^2_{B(A)} & : \; i=i', j=j', k \neq k' \\ 0 & : \; \text{otherwise} \end{cases}$$

$$E\{MSE\} = \sigma^2 \qquad E\{MSB(A)\} = \sigma^2 + n\sigma^2_{B(A)} \qquad E\{MSA\} = \sigma^2 + n\sigma^2_{B(A)} + \frac{n\sum_{i=1}^{a} b_i\alpha_i^2}{a-1}$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-Value
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSB(A)}$	$P\left(F_{a-1,b_.-a} \geq F_A\right)$
$B(A)$	$b_. - a$	$SSB(A)$	$MSB(A) = \frac{SSB(A)}{b_.-a}$	$F_{B(A)} = \frac{MSB(A)}{MSE}$	$P\left(F_{b_.-a,b_.(n-1)} \geq F_{B(A)}\right)$
ERROR	$b_.(n-1)$	SSE	$MSE = \frac{SSE}{b_.(n-1)}$		
TOTAL	$nb_. - 1$	TSS			

Table 3.26: The Analysis of Variance Table for a 2-Factor Nested Design – A Fixed and B Random

The tests for effects of factors A and $B(A)$ involve the two F–statistics, and can be conducted as follow. The test for differences among the effects of the levels of factor A is as follows.

1. $H_0^A : \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A^A : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSB(A)}$
4. R.R.: $F_A \geq F_{\alpha,(a-1),b_.-a}$
5. P-value: $P\left(F_{a-1,b_.-a} \geq F_A\right)$

The test for differences among the effects of the levels of factor $B(A)$ is as follows.

1. $H_0^{B(A)} : \sigma^2_{B(A)} = 0$ (No factor B effect).
2. $H_A^{B(A)} : \sigma^2_{B(A)} > 0$ (Factor B effects exist)
3. T.S. $F_{B(A)} = \frac{MSB(A)}{MSE}$

4. R.R.: $F_{B(A)} \geq F_{\alpha,(b_{.}-a),b_{.}(n-1)}$
5. P-value: $P\left(F_{b_{.}-a,b_{.}(n-1)} \geq F_{B(A)}\right)$

Pairwise comparisons among levels of Factor A are based on constructing simultaneous confidence intervals as follow.

Bonferroni (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*,b_{.}-a}\sqrt{MSB(A)\left(\frac{1}{nb_i} + \frac{1}{nb_{i'}}\right)}$$

Tukey

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha,a,b_{.}-a}\sqrt{\frac{MSB(A)}{2}\left(\frac{1}{nb_i} + \frac{1}{nb_{i'}}\right)}$$

ANOVA estimators for the variance components are given below, see the Expected Mean Squares given above.

$$s^2 = MSE \qquad s^2_{B(A)} = \frac{MSB(A) - MSE}{n}$$

Example 3.16: Momentum Measurements for Animal Trap Models

An experiment compared $a = 8$ models of animal traps (Cook and Proulx, 1989, [9]). The response was trap momentum at HDISP (when both jaws are displaced halfway). There were $b = 3$ traps per model, and each trap was measured $n = 10$ times. Data were generated to match reported summary statistics. For this analysis, the models are treated as fixed and the individual traps within each model are treated as random. The summary data are given in Table 3.27 and the Analysis of Variance is given in Table 3.28.

There is strong evidence of differences among models ($F_A = 48.65, P < .0001$) and among traps nested within models ($F_{B(A)} = 10.14, P < .0001$). To make comparisons among the models, there are $a = 8$ models and $C_A^* = 8(7)/2 - 28$ pairs.

Bonferroni (with $c_A^* = a(a-1)/2 = 8(8-1)/2 = 28$):

$$t_{.05/2(28),16} = 3.7398 \qquad \sqrt{0.010525\left(\frac{1}{10(3)} + \frac{1}{10(3)}\right)} = 0.02649 \qquad 3.7398(0.02649) = 0.0991$$

Tukey

$$q_{.05,8,16} = 4.8962 \qquad \sqrt{\frac{0.010525}{2}\left(\frac{1}{10(3)} + \frac{1}{10(3)}\right)} = 0.01873 \qquad 4.8962(0.01873) = 0.0917$$

Based on Tukey's HSD, Models 1-3 have significantly lower momentums than the others, Model 4 is significantly lower than Models 6-8, and Model 5 is significantly lower than Model 8.

ANOVA estimators for the variance components and standard deviations are given below.

$$s^2 = MSE = 0.001038 \quad (s = 0.03222) \qquad s^2_{B(A)} = \frac{0.010525 - 0.001038}{10} = 0.00095 \quad \left(s_{B(A)} = 0.03080\right)$$

Partial output from the R program is given below.

$$\nabla$$

Model	Trap1	Trap2	Trap3	Mean
1	0.5055 (0.0192)	0.5479 (0.0202)	0.5618 (0.0202)	0.5384
2	0.5503 (0.0362)	0.5609 (0.0350)	0.5652 (0.0371)	0.5588
3	0.5668 (0.0364)	0.5891 (0.0371)	0.6179 (0.0355)	0.5913
4	0.6567 (0.0361)	0.6705 (0.0336)	0.7377 (0.0355)	0.6883
5	0.7494 (0.0333)	0.7754 (0.0344)	0.7911 (0.0361)	0.7720
6	0.8076 (0.0317)	0.8091 (0.0311)	0.8176 (0.0300)	0.8114
7	0.7659 (0.0201)	0.8257 (0.0194)	0.8848 (0.0195)	0.8255
8	0.8333 (0.0392)	0.8830 (0.0374)	0.8902 (0.0401)	0.8688
Mean				0.7068

Table 3.27: Means (SDs) by Animal Trap Model and Trap within Model ($n = 10$) measurements per trap

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-Value
Trap Model	7	3.5845	$\frac{3.5848}{7} = 0.5121$	$\frac{0.5121}{0.010525} = 48.65$	$< .0001$
Trap (Model)	16	0.1684	$\frac{0.1684}{16} = 0.010525$	$F_{B(A)} = \frac{0.010525}{0.001038} = 10.141$	$< .0001$
ERROR	216	0.2242	$\frac{0.2242}{216} = 0.001038$		
TOTAL	239	3.9771			

Table 3.28: The Analysis of Variance Table for the Animal Trap Momentum Experiment – A Fixed and B Random Factors

```
> trap.mod1 <- aov(momentum ~ model + model/trapModel)
> anova(trap.mod1)
Analysis of Variance Table
Response: momentum
                 Df Sum Sq Mean Sq F value    Pr(>F)
model             7 3.5845 0.51207 493.374 < 2.2e-16 ***
model:trapModel  16 0.1684 0.01053  10.141 < 2.2e-16 ***
Residuals       216 0.2242 0.00104
```

```
> trap.mod2 <- aov(momentum ~ model + Error(trapModel))
> summary(trap.mod2)
Error: trapModel
          Df Sum Sq Mean Sq F value   Pr(>F)
model      7  3.585  0.5121   48.65 1.32e-09 ***
Residuals 16  0.168  0.0105
Error: Within
           Df Sum Sq  Mean Sq F value Pr(>F)
Residuals 216 0.2242 0.001038

> library(lmerTest)
> trap.mod4 <- lmer(momentum ~ model + (1|model:trapModel))
> summary(trap.mod4)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: momentum ~ model + (1 | model:trapModel)

Random effects:
 Groups          Name        Variance  Std.Dev.
 model:trapModel (Intercept) 0.0009487 0.03080
 Residual                    0.0010379 0.03222
Number of obs: 240, groups:  model:trapModel, 24

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)  0.706808   0.006622 16.000000 106.731  < 2e-16 ***
model1      -0.168408   0.017521 16.000000  -9.612 4.75e-08 ***
model2      -0.148008   0.017521 16.000000  -8.447 2.72e-07 ***
model3      -0.115542   0.017521 16.000000  -6.594 6.17e-06 ***
model4      -0.018509   0.017521 16.000000  -1.056  0.30648
model5       0.065158   0.017521 16.000000   3.719  0.00187 **
model6       0.104625   0.017521 16.000000   5.971 1.96e-05 ***
model7       0.118658   0.017521 16.000000   6.772 4.48e-06 ***
```

3.2.3 Factors A and B Random

In the case where A and B are both random, the effects for levels of Factor A Factor B are random variables, and the following assumptions are made.

$$\alpha_i \; N\left(0, \sigma_a^2\right) \qquad \beta_{j(i)} \; N\left(0, \sigma_{b(a)}^2\right) \qquad \epsilon_{ijk} \; N\left(0, \sigma_e^2\right)$$

Further, it is assumed that all random effects of levels of Factors A and B and all random error terms are mutually independent. The sums of squares are the same as in the previous subsections, and the error term for Factor A is the same as in the mixed case. The Analysis of Variance when Factors A and $B(A)$ are random is given in Table 3.29, where $b_{.} = \sum_{i=1}^{a} b_i$. The covariance structure and Expected Mean Squares for the factors are given below.

$$E\left\{Y_{ijk}\right\} = \mu \qquad V\left\{Y_{ijk}\right\} = \sigma_A^2 + \sigma_{B(A)}^2 + \sigma^2$$

$$\text{COV}\left\{Y_{ijk}, Y_{i'j'k'}\right\} = \begin{cases} \sigma_A^2 + \sigma_{B(A)}^2 + \sigma^2 & : \quad i = i', j = j', k = k' \\ \sigma_A^2 + \sigma_{B(A)}^2 & : \quad i = i', j = j', k \neq k' \\ \sigma_A^2 & : \quad i = i', j \neq j', k \neq k' \\ 0 & : \quad \text{otherwise} \end{cases}$$

$$E\left\{MSE\right\} = \sigma^2 \qquad E\left\{MSB(A)\right\} = \sigma^2 + n\sigma_{B(A)}^2 \qquad E\left\{MSA\right\} = \sigma^2 + n\sigma_{B(A)}^2 + n\frac{b_\cdot}{a}\sigma_A^2$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-value
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSB(A)}$	$P\left(F_{a-1,b_\cdot-a} \geq F_A\right)$
$B(A)$	$b_\cdot - a$	$SSB(A)$	$MSB(A) = \frac{SSB(A)}{b_\cdot-a}$	$F_{B(A)} = \frac{MSB(A)}{MSE}$	$P\left(F_{b_\cdot-a,b_\cdot(n-1)} \geq F_{B(A)}\right)$
ERROR	$b_\cdot(n-1)$	SSE	$MSE = \frac{SSE}{b_\cdot(n-1)}$		
TOTAL	$nb_\cdot - 1$	TSS			

Table 3.29: The Analysis of Variance Table for a 2-Factor Nested Design – A and B Random

The tests for interactions and for effects of factors A and B involve the two F–statistics, and can be conducted as follow. The test for differences among the effects of the levels of factor A is as follows.

1. $H_0^A : \sigma_A^2 = 0$ (No factor A effect).
2. $H_A^A : \sigma_A^2 > 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSB(A)}$
4. R.R.: $F_A \geq F_{\alpha,(a-1),b_\cdot-a}$
5. P-value: $P\left(F_{a-1,b_\cdot-a} \geq F_A\right)$

The test for differences among the effects of the levels of factor B as follows.

1. $H_0^{B(A)} : \sigma_{B(A)}^2 = 0$ (No factor B effect).
2. $H_A^{B(A)} : \sigma_{B(A)}^2 = 0$ (Factor B effects exist)
3. T.S. $F_{B(A)} = \frac{MSB(A)}{MSE}$
4. R.R.: $F_{B(A)} \geq F_{\alpha,(b_\cdot-a),b_\cdot(n-1)}$
5. P-value: $P\left(F_{b_\cdot-a,b_\cdot(n-1)} \geq F_{B(A)}\right)$

ANOVA estimators for the variance components are given below, see the Expected Mean Squares given above.

$$s^2 = MSE \qquad s^2_{B(A)} = \frac{MSB(A) - MSE}{n} \qquad s^2_A = \frac{MSA - MSB(A)}{n\frac{b}{a}}$$

Example 3.17: Variation in Semiconductor Wafers

A study reported variation in a measurement made on silicon wafers (Jensen, 2002, [16]). The measurement was not specified due to proprietary reasons. A random sample of $a = 20$ batches of wafers were selected, with $b = 2$ wafers being randomly selected within each batch, and $n = 9$ measurements were made at random locations on each wafer. The Analysis of Variance is given in Table 3.30.

There is strong evidence of batch-to-batch $(F_A = 7.5475, P < .0001)$ and wafer within lot $\left(F_{B(A)} = 5.1488, P < .0001\right)$ variation. ANOVA estimates of the variance components and standard deviations are given below.

$$s^2 = 19.03 \quad s = 4.36 \qquad s^2_{B(A)} = \frac{97.98 - 19.03}{9} = 8.77 \quad s_{B(A)} = 2.96 \qquad s^2_A = \frac{739.52 - 97.98}{2(9)} = 35.64 \quad s_A = 5.97$$

Partial output from the R program is given below.

∇

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	P-Value
Batch	19	14050.87	$\frac{14050.87}{19} = 739.52$	$\frac{739.52}{97.98} = 7.5475$	$< .0001$
Wafer(Batch)	20	1959.63	$\frac{1959.63}{20} = 97.98$	$F_{B(A)} = \frac{97.98}{19.03} = 5.1488$	$< .0001$
ERROR	320	6089.58	$\frac{6089.58}{320} = 19.03$		
TOTAL	359	22100.09			

Table 3.30: The Analysis of Variance Table for the Semiconductor Variation Study – A and B Random Factors

```
> semi.mod1 <- aov(Y ~ batch + batch/wafer)
> anova(semi.mod1)
Analysis of Variance Table
## The F-test for batch uses wrong error term
Response: Y
             Df  Sum Sq Mean Sq F value    Pr(>F)
batch        19 14050.9  739.52 38.8608 < 2.2e-16 ***
batch:wafer  20  1959.6   97.98  5.1488 3.708e-11 ***
Residuals   320  6089.6   19.03
```

```
> semi.mod2 <- lmer(Y ~ 1 + (1|batch/wafer))
> summary(semi.mod2)
summary from lme4 is returned
some computational error has occurred in lmerTest
Linear mixed model fit by REML ['lmerMod']
Formula: Y ~ 1 + (1 | batch/wafer)

Random effects:
 Groups      Name        Variance Std.Dev.
 wafer:batch (Intercept)  8.772   2.962
 batch       (Intercept) 35.641   5.970
 Residual                19.030   4.362
Number of obs: 360, groups:  wafer:batch, 40; batch, 20

Fixed effects:
            Estimate Std. Error t value
(Intercept)  174.342      1.433   121.6
```

3.3 Split-Plot Designs

In some experiments, with two or more factors, there is a restriction on randomization when assigning units to combinations of treatments. This may be due to measurements being made at multiple time points, or in the logistics of conducting the experiment. In this setting, there are larger experimental units (**whole plots**), which are made up of smaller subunits (**subplots**). Factors that are assigned to the whole plots are called the **Whole Plot Factor**. Not surprisingly, the factor applied to the sub units is called the **Sub Plot Factor**. The experiment can be set up as a Completely Randomized Design, with whole plot units being randomly assigned to whole-plot treatments and sub-plot units within whole plots receiving the sub-plot treatments. Often the experiment will be replicated in various blocks (maybe locations in a field trial or days in a laboratory experiment) as a Randomized Block Design for the Whole Plot units.

An experiment to compare 4 heating temperatures and 6 additive ingredients to bread flour may be conducted as follow. First, it is described as a Completely Randomized Design (CRD), then as a Randomized Block Design (RBD).

- Select 12 large sections of (homogeneous) bread flour (whole plots)
- Randomly assign each section to a temperature setting, such that three sections receive each temperature (whole plot factor levels)
- Break each piece into 6 subsections (sub plots)
- Randomly Assign an additive to each subsection, such that each full section of flour receives each additive (sub-plot factor levels)

- Select 4 large sections of (homogeneous) bread flour (whole plots)
- Randomly assign each piece to a temperature setting (whole plot factor levels)
- Break each piece into 6 subparts (sub plots)
- Randomly Assign an additive to each subsection, such that each full piece of flour receives each additive (sub-plot factor levels)

- Repeat the experiment on 3 days (blocks) with separate randomizations

Note that with extended cooking times, it would be unrealistic to individually prepare 24 combinations of temperature and additive in a single day. Thus, there is a restriction on randomization and cannot use a fully crossed Completely Randomized Design. In this study, temperature is the whole plot factor, additive is the sub-plot factor, and days serve as blocks.

For the CRD with respect to the whole plots, with whole plot units nested within the whole-plot factor levels, the model can be written as follows.

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ijk}$$

Here μ is the overall mean, α_i is the effect of the i^{th} level of (Whole Plot) Factor A, $\beta_{j(i)}$ is the effect of the j^{th} whole-plot nested within the i^{th} whole-plot factor level, γ_k is the effect of the k^{th} level of (Sub-Plot) Factor C, $(\alpha\gamma)_{ik}$ is the interaction between the i^{th} level of Factor A and the k^{th} level of Factor C, and ϵ_{ijk} is the random error term. When the whole plot and sub plot factors are fixed, the usual assumptions are as follow.

$$\sum_{i=1}^{a}\alpha_i = \sum_{k=1}^{c}\gamma_k = \sum_{i=1}^{a}(\alpha\gamma)_{ik} = \sum_{k=1}^{c}(\alpha\gamma)_{ik} = 0 \quad \forall k,i \qquad \beta_{j(i)} \sim N\left(0,\sigma^2_{B(A)}\right) \qquad \epsilon_{ijk} \sim N\left(0,\sigma^2\right)$$

All random effects are assumed to be independent of one another.

The general form of the model for a Split-Plot experiment when conducted as an RBD with respect to the Whole-Plot units is as follows.

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\alpha\gamma)_{ik} + \epsilon_{ijk}$$

Here μ is the overall mean, α_i is the effect of the i^{th} level of (Whole Plot) Factor A, β_j is the effect of the j^{th} block, $(\alpha\beta)_{ij}$ is the interaction between the i^{th} level of Factor A and Block j, γ_k is the effect of the k^{th} level of (Sub-Plot) Factor C, $(\alpha\gamma)_{ik}$ is the interaction between the i^{th} level of Factor A and the k^{th} level of Factor C, and ϵ_{ijk} is the random error term.

In general, there will be a levels for Factor A, b whole plot units within each level of Factor A (CRD) or blocks (RBD), and c levels of Factor C. In practice, Factor A will be fixed or random, and Factor C will be either fixed or random, and Whole Plot Units (CRD) or Blocks (RBD) will be random. In any event, the Analysis of Variance is the same, and is obtained as follows, based on the observed data.

$$\begin{aligned}
\overline{y}_{ij.} &= \frac{\sum_{k=1}^{c} y_{ijk}}{c} \\
\overline{y}_{i.k} &= \frac{\sum_{j=1}^{b} y_{ijk}}{b} \\
\overline{y}_{.jk} &= \frac{\sum_{i=1}^{a} y_{ijk}}{a}
\end{aligned}$$

$$\begin{aligned}
\overline{y}_{i..} &= \frac{\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}}{bc} \\
\overline{y}_{.j.} &= \frac{\sum_{i=1}^{a}\sum_{k=1}^{c} y_{ijk}}{ac} \\
\overline{y}_{..k} &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ijk}}{ab} \\
\overline{y}_{...} &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} y_{ijk}}{abc} \\
TSS &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(y_{ijk}-\overline{y}_{...}\right)^2 \\
SSA &= bc\sum_{i=1}^{a}\left(\overline{y}_{i..}-\overline{y}_{...}\right)^2 \\
\text{CRD: } SSB(A) &= c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{y}_{ij.}-\overline{y}_{i..}\right)^2 \\
\text{RBD: } SSB &= ac\sum_{j=1}^{b}\left(\overline{y}_{.j.}-\overline{y}_{...}\right)^2 \\
\text{RBD: } SSAB &= c\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{y}_{ij.}-\overline{y}_{i..}-\overline{y}_{.j.}+\overline{y}_{...}\right)^2 \\
SSC &= ab\sum_{k=1}^{c}\left(\overline{y}_{..k}-\overline{y}_{...}\right)^2 \\
SSAC &= b\sum_{i=1}^{a}\sum_{k=1}^{c}\left(\overline{y}_{i.k}-\overline{y}_{i..}-\overline{y}_{..k}+\overline{y}_{...}\right)^2 \\
SSE &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\left(y_{ijk}-\overline{y}_{ij.}-\overline{y}_{i.k}+\overline{y}_{i..}\right)^2
\end{aligned}$$

In the case of the CRD, the final error term is the interaction between the sub-plot factor and the whole-plot units nested within the whole-plot factor levels. For the RBD, the final error term represents the sum of the BC interaction and three-way ABC interaction, and thus assumes there is no sub-plot by block interaction. These two error terms are identical for a given set of abc measurements. The cases of fixed whole-plot and sub-plot factors are considered here. The Analysis of Variance for the CRD is given in Table 3.31 and for the RBD is given in Table 3.32. The Expected Mean Squares are given here that lead to the appropriate F-tests, the Sub-Plot portion of the table is the same for both designs.

Completely Randomized Design for Whole Plot Units

$$E\{MSE\}=\sigma^2 \qquad E\{MSAC\}=\sigma^2+\frac{b\sum_{i=1}^{a}\sum_{k=1}^{c}(\alpha\gamma)_{ik}^2}{(a-1)(c-1)} \qquad E\{MSC\}=\sigma^2+\frac{ab\sum_{k=1}^{c}\gamma_i^2}{c-1}$$

$$E\{MSB(A)\}=\sigma^2+c\sigma_{B(A)}^2 \qquad E\{MSA\}=\sigma^2+c\sigma_{B(A)}^2+\frac{bc\sum_{i=1}^{a}\alpha_i^2}{a-1}$$

Randomized Block Design for Whole Plot Units

$$E\{MSAB\} = \sigma^2 + c\sigma^2_{AB} \quad E\{MSB\} = \sigma^2 + c\sigma^2_{AB} + ac\sigma^2_B \quad E\{MSA\} = \sigma^2 + c\sigma^2_{AB} + \frac{bc\sum_{i=1}^{a}\alpha_i^2}{a-1}$$

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$P(>F)$
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSB(A)}$	$P\left(F_{a-1,a(b-1)} \geq F_A\right)$
$B(A)$	$a(b-1)$	$SSB(A)$	$MSB(A) = \frac{SSB(A)}{a(b-1)}$		
C	$c-1$	SSC	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$	$P\left(F_{c-1,a(b-1)(c-1)} \geq F_C\right)$
AC	$(a-1)(c-1)$	$SSAC$	$MSAC = \frac{SSAC}{(a-1)(c-1)}$	$F_{AC} = \frac{MSAC}{MSE}$	$P\left(F_{(a-1)(c-1),a(b-1)(c-1)} \geq F_{AC}\right)$
ERROR	$a(b-1)(c-1)$	SSE	$MSE = \frac{SSE}{a(b-1)(c-1)}$		
TOTAL	$abc-1$	TSS			

Table 3.31: The Analysis of Variance Table for a Split-Plot Design in a CRD – A and B Fixed Factors

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$P(>F)$
A	$a-1$	SSA	$MSA = \frac{SSA}{a-1}$	$F_A = \frac{MSA}{MSAB}$	$P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$
B	$b-1$	SSB	$MSB = \frac{SSB}{b-1}$		
AB	$(a-1)(b-1)$	$SSAB$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$		
C	$c-1$	SSC	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$	$P\left(F_{c-1,a(b-1)(c-1)} \geq F_C\right)$
AC	$(a-1)(c-1)$	$SSAC$	$MSAC = \frac{SSAC}{(a-1)(c-1)}$	$F_{AC} = \frac{MSAC}{MSE}$	$P\left(F_{(a-1)(c-1),a(b-1)(c-1)} \geq F_{AC}\right)$
ERROR	$a(b-1)(c-1)$	SSE	$MSE = \frac{SSE}{a(b-1)(c-1)}$		
TOTAL	$abc-1$	TSS			

Table 3.32: The Analysis of Variance Table for a Split-Plot Design in a RBD – A and B Fixed Factors

The tests for interaction and main effects of factors A, C involve the three F–statistics, and can be conducted as follow. First, conduct the test for an interaction between the whole plot factor (A) and the sub-plot factor (C).

1. $H_0 : (\alpha\gamma)_{11} = \cdots = (\alpha\gamma)_{ac} = 0$ (No factor AC interaction).
2. H_A : Not all $(\alpha\gamma)_{ik} = 0$ (AC interaction exists)
3. T.S. $F_{AC} = \frac{MSAC}{MSE}$

4. R.R.: $F_{AC} \geq F_{\alpha,(a-1)(c-1),a(b-1)(c-1)}$
5. P-value: $P\left(F_{(a-1)(c-1),a(b-1)(c-1)} \geq F_{AC}\right)$

Assuming no interaction, the test for differences among the effects of the levels of factor C is conducted as follows.

1. $H_0 : \gamma_1 = \ldots = \gamma_c = 0$ (No factor C effect).
2. H_A : Not all $\gamma_k = 0$ (Factor C effects exist)
3. T.S. $F_C = \frac{MSC}{MSE}$
4. R.R.: $F_C \geq F_{\alpha,c-1,a(b-1)(c-1)}$
5. P-value: $P\left(F_{c-1,a(b-1)(c-1)} \geq F_C\right)$

Assuming no interaction exists, the test for differences among the effects of the levels of factor A is conducted as follows for the CRD.

1. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_A = \frac{MSA}{MSB(A)}$
4. R.R.: $F_A \geq F_{\alpha,(a-1),a(b-1)}$
5. P-value: $P\left(F_{a-1,a(b-1)} \geq F_A\right)$

Assuming no interaction exists, the test for differences among the effects of the levels of factor A is conducted as follows for the RBD.

1. $H_0 : \alpha_1 = \cdots = \alpha_a = 0$ (No factor A effect).
2. H_A : Not all $\alpha_i = 0$ (Factor A effects exist)
3. T.S. $F_{obs} = \frac{MSA}{MSAB}$
4. R.R.: $F_{obs} \geq F_{\alpha,(a-1),(a-1)(b-1)}$
5. P-value: $P\left(F_{a-1,(a-1)(b-1)} \geq F_A\right)$

When there is no interaction, pairwise comparisons can be made among levels of Factor A based on constructing simultaneous confidence intervals as follow.

Bonferroni (CRD) (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*,a(b-1)} \sqrt{MSB(A)\left(\frac{2}{bc}\right)}$$

Tukey (CRD)

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha,a,a(b-1)}\sqrt{MSB(A)\left(\frac{1}{bc}\right)}$$

Bonferroni (RBD) (with $c_A^* = a(a-1)/2$):

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm t_{\alpha/2c_A^*,(a-1)(b-1)}\sqrt{MSAB\left(\frac{2}{bc}\right)}$$

Tukey (RBD)

$$(\overline{y}_{i..} - \overline{y}_{i'..}) \pm q_{\alpha,a,(a-1)(b-1)}\sqrt{MSAB\left(\frac{1}{bc}\right)}$$

To compare levels of Factor C, the simultaneous confidence intervals are constructed as follow. These are the same, whether the design was constructed as a CRD or RBD with respect to the Whole-Plot units.

Bonferroni (with $c_C^* = c(c-1)/2$):

$$(\overline{y}_{..k} - \overline{y}_{..k'}) \pm t_{\alpha/2c_C^*,a(b-1)(c-1)}\sqrt{MSE\left(\frac{2}{ab}\right)},$$

Tukey's

$$(\overline{y}_{..k} - \overline{y}_{..k'}) \pm q_{\alpha,c,a(b-1)(c-1)}\sqrt{MSE\left(\frac{1}{ab}\right)}$$

When interaction is present, sub-plot factor levels can be compared within the same level of the whole-plot factor or the whole-plot levels could be compared within the same sub-plot. The estimated standard errors of the differences are given below, along with their degrees of freedom based on Satterthwaite's Approximation.

$$\text{SP within WP: } \hat{SE}\left\{\overline{Y}_{i.k} - \overline{Y}_{i.k'}\right\} = \sqrt{\frac{2MSE}{b}} \quad df = a(b-1)(c-1)$$

$$\text{WP within/across SP (CRD): } \hat{SE}\left\{\overline{Y}_{i.k} - \overline{Y}_{i'.k'}\right\} = \sqrt{\frac{2\left[MSB(A) + (c-1)MSE\right]}{bc}}$$

$$df = \frac{\left(MSB(A) + (c-1)MSE\right)^2}{\left(\frac{(MSB(A))^2}{a(b-1)} + \frac{((c-1)MSE)^2}{a(b-1)(c-1)}\right)}$$

$$\text{WP within/across SP (RBD): } \hat{SE}\left\{\overline{Y}_{i.k} - \overline{Y}_{i'.k'}\right\} = \sqrt{\frac{2\left[MSAB + (c-1)MSE\right]}{bc}}$$

$$df = \frac{\left(MSAB + (c-1)MSE\right)^2}{\left(\frac{(MSAB)^2}{(a-1)(b-1)} + \frac{((c-1)MSE)^2}{a(b-1)(c-1)}\right)}$$

In the case of the Completely Randomized Design, the ANOVA estimator of variation in the Whole-Plot unit effects can be obtained as follows.

$$s^2 = MSE \qquad s^2_{B(A)} = \frac{MSB(A) - MSE}{c}$$

In the case of the Randomized Block Design, the ANOVA estimators of the variance components for Blocks and Whole-Plot/Block interaction can be obtained as follow.

$$s^2 = MSE \qquad s^2_{AB} = \frac{MSAB - MSE}{c} \qquad s^2_B = \frac{MSB - MSAB}{ac}$$

Two examples are considered. The first is an observational study that was conducted as a Completely Randomized Design (except that subjects were sampled from and not assigned to the two whole-plot factor groups). The second involves a Split-Plot experiment conducted within a Randomized Block Design.

Example 3.18: Axion Densities in Eyes of Normal and Alzheimers Patients

An observational study reported axion densities in right/left eyes of Normal/Alzheimers patients (Armstrong, 2013, [5]). The whole-plot factor was patient status (Normal ($i = 1$), Alzheimers($i = 2$), $a = 2$), with $b = 12$ subjects sampled from each patient group. Within each patient, the axion density was measured in each eye (sub-plot factor, Right ($k = 1$), Left ($k = 2$), $c = 2$). Data are given in Table 3.33.

Due to the simplicity of the dataset, the sums of squares are set-up directly (albeit using a spreadsheet for actual calculations).

$$\text{Whole-Plot (A): } SSA = 12(2)\left[(886.125 - 734.938)^2 + (583.750 - 734.938)^2\right] = 1097168$$

$$\text{Subj(WP) (B(A)): } SSB(A) = 2\left[(719.5 - 886.125)^2 + \cdots + (627 - 583.750)^2\right] = 1392056$$

$$\text{Sub-Plot (C): } SSC = 2(12)\left[(748.208 - 734.938)^2 + (721.667 - 734.938)^2\right] = 8453$$

$$\text{WPxSP (AC): } SSAC = 12\left[(829.167 - 886.125 - 748.208 + 734.938)^2 + \cdots + (563.250 - 538.750 - 721.667 + 734.938)^2\right] = 2$$

$$\text{Error: } SSE = (673 - 719.5 - 892.167 + 886.125)^2 + \cdots + (374 - 376 - 563.250 + 721.667)^2 = 167889$$

$$df_A = 2-1 = 1 \qquad df_{B(A)} = 2(12-1) = 22 \qquad df_C = 2-1 = 1 \qquad df_{AC} = (2-1)(2-1) = 1 \qquad df_E = 2(12-1)(2-1) = 22$$

The Analysis of Variance is given in Table 3.34. It is clear that their is a significant Patient Group effect with little evidence of either an Eye main effect or an interaction between Patient Group and Eye. As there are only two levels of Patient Group, there is only one comparison: Normal versus Alzheimer's; the results of Bonferroni's and Tukey's methods are identical. The ANOVA estimates of the variance components are given below,

$$\overline{y}_{1..} - \overline{y}_{2..} = 886.125 - 583.750 = 302.375 \quad t_{.025,22} = 2.074 \quad \sqrt{MSB(A)\left(\frac{2}{2(12)}\right)} = \sqrt{\frac{63275(2)}{12(2)}} = 72.615$$

$$302.375 \pm 2.074(72.615) \equiv 302.375 \pm 150.603 \equiv (151.8, 453.0)$$

$$s^2 = 7631 \qquad s^2_{B(A)} = \frac{63275 - 7631}{2} = 27822$$

The R partial output is given below.

∇

Normal Patients				Alzheimers Patients			
Subject	Right	Left	SubjMean	Subject	Right	Left	SubjMean
1	673	766	719.5	13	538	377	457.5
2	899	956	927.5	14	583	555	569
3	616	605	610.5	15	696	298	497
4	749	858	803.5	16	568	583	575.5
5	1078	1017	1047.5	17	649	700	674.5
6	978	861	919.5	18	284	458	371
7	706	569	637.5	19	862	746	804
8	1005	991	998	20	848	774	811
9	1420	1258	1339	21	716	698	707
10	1003	997	1000	22	508	563	535.5
11	818	982	900	23	378	374	376
12	761	701	731	24	621	633	627
Mean	$\overline{y}_{1.1} = 892.167$	$\overline{y}_{1.2} = 880.083$	$\overline{y}_{1..} = 886.125$	Mean	$\overline{y}_{2.1} = 604.250$	$\overline{y}_{2.2} = 563.250$	$\overline{y}_{2..} = 583.750$
Mean	$\overline{y}_{..1} = 748.208$	$\overline{y}_{..2} = 721.667$	$\overline{y}_{...} = 734.938$				

Table 3.33: Individual measurements for Axion densities is Normals/Alzheimers patients

```
## ANOVA w/ incorrect WP error term
> eyes.mod1 <- aov(axondens ~ alz_grp + alz_grp/subject + eye + alz_grp:eye)
> anova(eyes.mod1)
Analysis of Variance Table
Response: axondens
                Df  Sum Sq Mean Sq  F value    Pr(>F)
alz_grp          1 1097168 1097168 143.7713 4.037e-11 ***
eye              1    8454    8454   1.1077    0.3040
alz_grp:subject 22 1392056   63275   8.2915 2.812e-06 ***
alz_grp:eye      1    2509    2509   0.3287    0.5722
Residuals       22  167889    7631

## ANOVA w/ correct WP error term
> eyes.mod2 <- aov(axondens ~ alz_grp * eye + Error(subject))
> summary(eyes.mod2)
Error: subject
          Df  Sum Sq Mean Sq F value   Pr(>F)
alz_grp    1 1097168 1097168   17.34 0.000404 ***
```

ANOVA

Source of	Degrees of	Sum of	Mean		
Variation	Freedom	Squares	Square	F	$P(>F)$
Patient	1	1097168	$\frac{1097168}{1} = 1097168$	$\frac{1097168}{63275} = 17.3396$	0.0004
Subj(Patient Grp)	22	1392056	$\frac{1392056}{22} = 63275$		
Eye	1	8454	$\frac{8454}{1} = 8454$	$\frac{8454}{7631} = 1.1077$	0.3040
Patient Grp x Eye	1	2509	$\frac{2509}{1} = 2509$	$\frac{2509}{7631} = 0.3287$	0.5722
Error	22	167889	$\frac{16788}{22} = 7631$		
Total	47	2668075			

Table 3.34: The Analysis of Variance Table for Axion densities in Normal/Alzheimers Patients' Right/Left Eyes

```
Residuals 22 1392056   63275
Error: Within
            Df Sum Sq Mean Sq F value Pr(>F)
eye          1   8454    8454   1.108  0.304
alz_grp:eye  1   2509    2509   0.329  0.572
Residuals   22 167889    7631

## Mixed Effects Model
> eyes.mod3 <- lmer(axondens ~ alz_grp * eye + (1|alz_grp:subject))
> summary(eyes.mod3)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: axondens ~ alz_grp * eye + (1 | alz_grp:subject)

Random effects:
 Groups          Name        Variance Std.Dev.
 alz_grp:subject (Intercept) 27822    166.80
 Residual                     7631     87.36
Number of obs: 48, groups:  alz_grp:subject, 24

Fixed effects:
              Estimate Std. Error      df t value Pr(>|t|)
(Intercept)    734.938     36.307  22.000  20.242 8.88e-16 ***
alz_grp1       151.187     36.307  22.000   4.164 0.000404 ***
eye1            13.271     12.609  22.052   1.052 0.303971
alz_grp1:eye1   -7.229     12.609  22.052  -0.573 0.572219

> anova(eyes.mod3)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
            Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
alz_grp     132324  132324     1 22.00 17.3396 0.0004041 ***
eye           8454    8454     1 22.03  1.1077 0.3039824
alz_grp:eye   2509    2509     1 22.03  0.3287 0.5722247

> difflsmeans(eyes.mod3)
Differences of LSMEANS:
                        Estimate Standard Error     DF t-value Lower CI Upper CI p-value
alz_grp 1 - 2              302.4           72.6   22.0    4.16    151.8    453.0   4e-04 ***
eye 1 - 2                   26.5           25.2   22.0    1.05    -25.8     78.8   0.304
```

Example 3.19: Chymosin Treatment and Ripening Time Effects on Mozzarella Cheese

An experiment was conducted to measure various responses due to different Chymosin treatments and Ripening times on mozzarella cheese (Moynihan, et al, 2014, [24]). The Whole-Plot factor was Chymosin Treatment (High Bovine Calf (HBCC, $i = 1$), Low Bovine Calf (LBCC, $i = 2$), High Camel (HCC, $i = 3$), and Low Camel (LCC, $i = 4$)). The Sub-Plot factor was Ripening Time (14 days ($k = 1$), 28 ($k = 2$), 56 ($k = 3$), and 84 ($k = 4$)). The experiment was replicated on $b = 3$ cheesemaking days (blocks). On a given cheesemaking day, there would be a random ordering of the 4 types of cheese to be made. Then once one of the Chymosin types was prepared, it would be broken into 4 subsections, which would be randomly assigned to the 4 ripening times. The 4 ripening times were randomly assigned to the subsections, then the samples were refrigerated and stored for the appropriate ripening time, then observed.

There were 3 response variables: hardness of melted cheese (Y_1), adhesiveness of mass (Y_2), and blister quantity (Y_3), this analysis will be based on blister quantity. The cell means for Chymosin Type and Ripening Time are given in Table 3.35, each mean is an average of $b = 3$ measurements (one per cheesemaking day). The Analysis of Variance is given in Table 3.36, there are significant main effects for Chymosin Treatment ($F_A = 14.87, P = .0035$) and Ripening Time ($F_C = 13.52, P < .0001$). There is no evidence of a Chymosin Treatment/Ripening Time interaction ($F_{AC} = 1.14, P = .3754$).

Pairwise comparisons among Chymosin Treatments and Ripening Times, as well as estimated variance components are given below. There are $a = c = 4$ Whole-Plot and Sub-Plot treatments with $c_A^* = c_C^* = 4(3)/2 = 6$ comparisons for each.

$$\text{Bonferroni (Chymosin): } t_{.05/(2(6)),6} = 3.863 \qquad \sqrt{\frac{2(1.22)}{3(4)}} = 0.451 \qquad 3.863(0.451) = 1.742$$

$$\text{Tukey (Chymosin): } q_{.05,4,6} = 4.896 \qquad \sqrt{\frac{1.22}{3(4)}} = 0.319 \qquad 4.896(0.319) = 1.561$$

$$\text{Bonferroni (Ripening): } t_{.05/(2(6)),24} = 2.875 \qquad \sqrt{\frac{2(1.05)}{4(3)}} = 0.418 \qquad 2.875(0.418) = 1.203$$

$$\text{Tukey (Ripening): } q_{.05,4,24} = 3.901 \qquad \sqrt{\frac{1.05}{3(4)}} = 0.296 \qquad 3.901(0.296) = 1.154$$

$$s^2 = 1.05 \qquad s_{AB}^2 = \frac{1.22 - 1.05}{4} = 0.43 \qquad s_B^2 = \frac{41.30 - 1.22}{4(4)} = 2.51$$

Based on Tukey's method, in terms of the Chymosin Treatments, HBCC and LBCC have significantly higher means than HCC and LCC; no other pairs of means are significantly different. In terms of Ripening Times, 84 days is significantly higher than all other times, 56 days is significantly higher than 14 days; no other pairs of times are significantly different. The partial R output is given below.

∇

	Ripe1 (14)	Ripe2 (28)	Ripe3 (56)	Ripe4 (84)	Mean
Chym1 (HBCC)	8.770	10.280	9.300	11.750	10.025
Chym2 (LBCC)	8.730	9.550	10.340	11.590	10.052
Chym3 (HCC)	7.550	8.080	8.690	9.160	8.370
Chym4 (LCC)	6.590	6.100	8.410	9.270	7.592
Mean	7.910	8.502	9.185	10.442	9.010

Table 3.35: Mean blister quantities by Chymosin Treatment and Ripening Times - Split-Plot experiment in RBD

		ANOVA			
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$P(> F)$
Chymosin (WP)	3	54.43	18.14	$\frac{18.14}{1.22} = 14.8719$	0.0035
Day (Block)	2	82.60	41.30	$\frac{41.30}{1.22} = 33.8525$	0.0005
ChyxDay (WPxBlk)	6	7.32	1.22	$\frac{1.22}{1.05} = 1.1619$	0.3588
RipeTime (SP)	3	42.60	14.20	$\frac{14.20}{1.05} = 13.5247$	0.0000
ChyxRipe (WPxSP)	9	10.76	1.20	$\frac{1.20}{1.05} = 1.1385$	0.3754
Error	24	25.20	1.05		
Total	47	222.91			

Table 3.36: The Analysis of Variance Table for Blister quantities for Chymosin Treatment/Ripening Times Split-Plot experiment

```
## ANOVA w/ incorrect Whole-Plot F-test
> chym.mod1 <- aov(blister ~ c.trt*c.blk + c.time + c.trt:c.time)
> summary(chym.mod1)
             Df Sum Sq Mean Sq F value   Pr(>F)
c.trt         3  54.43   18.14  17.280 3.42e-06 ***
c.blk         2  82.60   41.30  39.333 2.66e-08 ***
c.time        3  42.60   14.20  13.525 2.27e-05 ***
c.trt:c.blk   6   7.32    1.22   1.162    0.359
c.trt:c.time  9  10.76    1.20   1.139    0.375
Residuals    24  25.20    1.05

## Mixed Effects Model Outbut
> chy.mod2 <- lmer(blister ~ c.trt*c.time + (1|c.blk) + (1|c.trt:c.blk))
> summary(chy.mod2)
Linear mixed model fit by REML t-tests use Satterthwaite approximations to
  degrees of freedom [lmerMod]
Formula: blister ~ c.trt * c.time + (1 | c.blk) + (1 | c.trt:c.blk)
Random effects:
 Groups      Name        Variance Std.Dev.
 c.trt:c.blk (Intercept) 0.0425   0.2062
 c.blk       (Intercept) 2.5050   1.5827
 Residual                1.0500   1.0247
Number of obs: 48, groups:  c.trt:c.blk, 12; c.blk, 3

Fixed effects:
               Estimate Std. Error      df t value Pr(>|t|)
(Intercept)      9.0100     0.9276  2.0000   9.713  0.01043 *
c.trt1           1.0150     0.2761  6.0000   3.676  0.01038 *
c.trt2           1.0425     0.2761  6.0000   3.775  0.00923 **
c.trt3          -0.6400     0.2761  6.0000  -2.318  0.05964 .
c.time1         -1.1000     0.2562 24.0000  -4.294  0.00025 ***
c.time2         -0.5075     0.2562 24.0000  -1.981  0.05915 .
c.time3          0.1750     0.2562 24.0000   0.683  0.50107
c.trt1:c.time1  -0.1550     0.4437 24.0000  -0.349  0.72989
c.trt2:c.time1  -0.2225     0.4437 24.0000  -0.501  0.62062
c.trt3:c.time1   0.2800     0.4437 24.0000   0.631  0.53397
c.trt1:c.time2   0.7625     0.4437 24.0000   1.718  0.09858 .
c.trt2:c.time2   0.0050     0.4437 24.0000   0.011  0.99110
c.trt3:c.time2   0.2175     0.4437 24.0000   0.490  0.62845
c.trt1:c.time3  -0.9000     0.4437 24.0000  -2.028  0.05376 .
c.trt2:c.time3   0.1125     0.4437 24.0000   0.254  0.80201
c.trt3:c.time3   0.1450     0.4437 24.0000   0.327  0.74666

> anova(chy.mod2)
Analysis of Variance Table of type III  with  Satterthwaite
approximation for degrees of freedom
             Sum Sq Mean Sq NumDF DenDF F.value    Pr(>F)
c.trt        46.847 15.6155     3     6 14.8719  0.003478 **
c.time       42.603 14.2010     3    24 13.5247 2.269e-05 ***
c.trt:c.time 10.759  1.1954     9    24  1.1385  0.375426

> difflsmeans(chy.mod2)
Differences of LSMEANS:
                         Estimate Standard Error   DF t-value Lower CI Upper CI p-value
c.trt 1 - 2                   0.0         0.4509  6.0   -0.06  -1.1309   1.0759   0.953
c.trt 1 - 3                   1.7         0.4509  6.0    3.67   0.5516   2.7584   0.010 *
c.trt 1 - 4                   2.4         0.4509  6.0    5.39   1.3291   3.5359   0.002 **
c.trt 2 - 3                   1.7         0.4509  6.0    3.73   0.5791   2.7859   0.010 **
c.trt 2 - 4                   2.5         0.4509  6.0    5.46   1.3566   3.5634   0.002 **
c.trt 3 - 4                   0.8         0.4509  6.0    1.72  -0.3259   1.8809   0.135
c.time 1 - 2                 -0.6         0.4183 24.0   -1.42  -1.4559   0.2709   0.170
c.time 1 - 3                 -1.3         0.4183 24.0   -3.05  -2.1384  -0.4116   0.005 **
c.time 1 - 4                 -2.5         0.4183 24.0   -6.05  -3.3959  -1.6691  <2e-16 ***
c.time 2 - 3                 -0.7         0.4183 24.0   -1.63  -1.5459   0.1809   0.116
```

```
c.time 2 - 4                 -1.9        0.4183 24.0   -4.64  -2.8034  -1.0766   1e-04 ***
c.time 3 - 4                 -1.3        0.4183 24.0   -3.01  -2.1209  -0.3941   0.006 **
```

Split-plot experiments can also contain random effects, although most published examples tend to have fixed effects among whole-plot and sub-plot factors. Below is a description of an example with the sub-plot factor being a random factor.

Example 3.20: Rating of Whey Syrup in a Sensory Study

An experiment is described of sensory ratings of whey syrup used in Norwegian ice cream (Steinsholt, 1998, [32]). There were 4 varieties of whey syrup, and 8 raters, and the plan was to have 3 replicates per variety/rater. Had this been done as a fully crossed design, there would have had to been 4(8)(3)=96 batches produced, which would have been very time consuming. Instead, there were 4(3) = 12 batches made among the 4 syrup varieties in a CRD, and the batches were stored frozen for 14 days. Then the raters rated the 12 batches in random order (separate for each rater). In this example, the Varieties (whole-plot factor) are Fixed and the Raters (sub-plot factor) are random (although there is heated debate in the food preference literature of whether to treat trained raters as fixed or random). Note that the author used a different model formulation than that used here and in most mixed effects software packages. For this model, Factor A is variety with $a = 4$ levels, Factor $B(A)$ is batch nested within variety with $b = 3$ replicates per variety, and Factor C is rater, with $c = 8$ levels.

The model and Expected Mean Squares are given below for fixed variety (Whole-Plot factor), random rater (Sub-Plot factor) and and random variety/rater interaction effects, with all random effects assumed independent.

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + (\alpha\beta)_{ik} + \epsilon_{ijk} \qquad i = 1, \ldots, a = 4 \quad j = 1, \ldots, b = 3 \quad k = 1, \ldots, 8$$

$$\sum_{i=1}^{a} \alpha_i = 0 \quad \beta_{j(i)} \sim N\left(0, \sigma^2_{B(A)}\right) \quad \gamma_K \sim N\left(0, \sigma^2_{B(A)}\right) \quad (\alpha\gamma)_{ik} \sim N\left(0, \sigma^2_{AC}\right) \quad \epsilon_{ijk} \sim N\left(0, \sigma^2\right)$$

$$E\left\{MSE\right\} = \sigma^2 \quad E\left\{MSAC\right\} = \sigma^2 + b\sigma^2_{AC} \qquad E\left\{MSC\right\} = \sigma^2 + b\sigma^2_{AC} + ab\sigma^2_C$$

$$E\left\{MSB(A)\right\} = \sigma^2 + c\sigma^2_{B(A)} \qquad E\left\{MSA\right\} = \sigma^2 + b\sigma^2_{AC} + c\sigma^2_{B(A)} + bc\sum_{i=1}^{a}\alpha_i^2$$

Note that the F-test for variety effects is not simply the ratio of the Variety mean square to the Batch(Variety) mean square, and thus Satterthwaite's approximation must be used. There are two ways the F-test can be conducted, but note that software packages such as the **lmerTest** package in R will use the second version.

$$H_0^A : \alpha_1 = \cdots = \alpha_a = 0 \qquad TS_1 : F_{A1} = \frac{MSA + MSE}{MSB(A) + MSAC} \qquad TS_2 : F_{A2} = \frac{MSA}{MSB(A) + MSAC - MSE}$$

The first F-statistic will always be positive, but both the numerator and denominator degrees of freedom must be estimated. The second F-statistic can be negative, but only the denominator degrees of freedom must be estimated. The ANOVA table reported in paper is given in Table 3.37.

The F-test for Variety effects is conducted here using the second version given above. ANOVA estimates of the variance components are also obtained.

$$TS: F_{A2} = \frac{5.566}{1.845 + 1.447 - 1.164} = \frac{5.566}{2.128} = 2.616 \qquad df_2 = \frac{(2.128)^2}{\left[\frac{(1.845)^2}{8} + \frac{(1.447)^2}{21} + \frac{(-1.164)^2}{56}\right]} = \frac{4.528}{0.549} = 8.25$$

$$RR_2 : F_{A2} \geq F_{.05,3,8.25} = 4.009 \quad P = .1210$$

$$s^2 = 1.164 \quad s^2_{AC} = \frac{1.447 - 1.164}{3} = 0.094 \quad s^2_C = \frac{33.118 - 1.447}{4(3)} = 2.639 \quad s^2_{B(A)} = \frac{1.845 - 1.164}{8} = 0.085$$

∇

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
A	$4-1=3$	16.698	$\frac{16.698}{3} = 5.566$
$B(A)$	$4(3-1)=8$	14.833	$\frac{14.833}{8} = 1.845$
C	$8-1=7$	231.823	$\frac{231.823}{7} = 33.118$
AC	$(4-1)(8-1)=21$	30.385	$\frac{30.385}{21} = 1.447$
ERROR	$4(3-1)(8-1)=56$	65.167	$\frac{65.167}{56}1.164$
TOTAL	$4(3)(8)-1=95$	358.906	

Table 3.37: The Analysis of Variance Table for a Split-Plot Design in a CRD – Ice Cream Rating Example

3.4 Repeated Measures Designs

In some experimental situations, subjects are assigned to treatments, and measurements are made at repeated points over some fixed period of time. This can be thought of as a CRD, where more than one measurement is being made on each experimental unit. The goal is still to detect differences among the treatment means (effects), but must account for the fact that measurements are being made over time. Previously, the error was differences among the units within the treatments. Now various measurements are observed on each unit nested within each treatment, and have a new error term. The measurement Y_{ijk}, representing the outcome for the i^{th} treatment on the j^{th} unit (that receives the treatment) at the k^{th} time point, can be written as follows (this is a special case of the Split-Plot experiment where the whole-plot units (often subjects) are assigned to treatments in a CRD). Note that many (particularly behavioral) studies will refer to a Randomized Block Design as a Repeated Measures Design where the time points are replaced by the various treatments and units (subjects) are the blocks (where each unit receives each treatment).

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \alpha\gamma_{ik} + \epsilon_{ijk} \qquad i = 1, \ldots, c \quad j(i) = 1 \ldots, b_i \quad k = 1, \ldots, c$$

where:

- μ is the overall mean
- α_i is the fixed effect of the i^{th} treatment $(i = 1, \ldots, a)$
- $\beta_{j(i)}$ is the random effect of the j^{th} unit receiving the i^{th} treatment $(j = 1, \ldots, b_i)$
- γ_k is the fixed effect of the k^{th} time point $(k = 1, \ldots, c)$
- $\alpha\gamma_{ik}$ is the interaction of the i^{th} treatment and the k^{th} time point
- ϵ_{ijk} is the random error component that is assumed to be $N\left(0, \sigma^2\right)$.

The means sums of squares can be obtained as follows for the case where $b_1 = \cdots = b_a = b$ and the data have been observed.

$$\begin{aligned}
\overline{y}_{ij.} &= \frac{\sum_{k=1}^{c} y_{ijk}}{c} \\
\overline{y}_{i.k} &= \frac{\sum_{j=1}^{b} y_{ijk}}{b} \\
\overline{y}_{i..} &= \frac{\sum_{j=1}^{b} \sum_{k=1}^{c} y_{ijk}}{bc} \\
\overline{y}_{..k} &= \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} y_{ijk}}{ab} \\
\overline{y}_{...} &= \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} y_{ijk}}{abc} \\
TSS &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(y_{ijk} - \overline{y}\ldots\right)^2 \\
SSA &= bc \sum_{i=1}^{a} \left(\overline{y}_{i..} - \overline{y}\ldots\right)^2 \\
SSB(A) &= c \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{y}_{ij.} - \overline{y}_{i..}\right)^2 \\
SSC &= ab \sum_{k=1}^{c} \sum_{k=1}^{c} \left(\overline{y}_{..k} - \overline{y}\ldots\right)^2 \\
SSAC &= b \sum_{i=1}^{a} \sum_{k=1}^{c} \left(\overline{y}_{i.k} - \overline{y}_{i..} - \overline{y}_{..k} + \overline{y}\ldots\right)^2 \\
SSE &= \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} \left(y_{ijk} - \overline{y}_{ij.} - \overline{y}_{i.k} + \overline{y}_{i..}\right)^2
\end{aligned}$$

In practice, treatments and time points will always be treated as fixed effects and units nested within treatments are random effects. The Analysis of Variance is given in Table 3.38 (this is always done on a computer). The degrees of freedom are based on the experiment consisting of a treatments, b subjects

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	$P(> F)$
Trts	SSA	$a-1$	$MSA = \frac{SSA}{a-1}$	$FA = \frac{MSA}{MSB(A)}$	$P\left(F_{a-1,a(b-1)} \geq F_A\right)$
Units(Trts)	$SSB(A)$	$a(b-1)$	$MSB(A) = \frac{SSB(A)}{a(b-1)}$		
Time	SSC	$c-1$	$MSC = \frac{SSC}{c-1}$	$F_C = \frac{MSC}{MSE}$	$P\left(F_{c-1,a(b-1)(c-1)} \geq F_C\right)$
TrtxTime	$SSAC$	$(a-1)(c-1)$	$MSAC = \frac{SSAC}{(a-1)(c-1)}$	$F_{AC} = \frac{MSAC}{MSE}$	$P\left(F_{((a-1)(c-1),a(b-1)(c-1)} \geq F_{AC}\right)$
Error	SSE	$a(b-1)(c-1)$	$MSE = \frac{SSE}{a(b-1)(c-1)}$		
TOTAL	TSS	$abc-1$			

Table 3.38: Univariate Analysis of Variance Table for a Repeated Measures Design

receiving each treatment, and measurements being made at c points in time. Note that if the number of subjects per treatment differ (b_i subjects receiving treatment i), replace $a(b-1)$ with $\sum_{i=1}^{a}(b_i - 1)$.

One primary hypothesis to test is for a treatment effect. This test is of the form:

1. $H_0^A : \alpha_1 = \cdots = \alpha_a = 0$ (No treatment effect)
2. H_A^A : Not all $\alpha_i = 0$ (Treatment effects)
3. T.S.: $F_A = \frac{MSA}{MSB(A)}$
4. R.R.: $F_A \geq F_{\alpha,a-1,a(b-1)}$
5. P-value: $P\left(F_{a-l,a(b-1)} \geq F_A\right)$

To test for time effects and time by treatment interaction, tests are of the following form.

1. $H_0^C : \gamma_1 = \cdots = \gamma_c = 0$ (No time effect)
2. H_A^C : Not all $\gamma_k = 0$ (Time effects)
3. T.S.: $F_C = \frac{MSC}{MSE}$
4. R.R.: $F_C \geq F_{\alpha,c-1,a(b-1)(c-1)}$
5. P-value: $P\left(F_{(c-1),a(b-1)(c-1)} \geq F_C\right)$

1. $H_0^{AC} : (\alpha\gamma)_{11} = \cdots = (\alpha\gamma)_{ac} = 0$ (No trt by time interaction)
2. H_A^{AC} : Not all $(\alpha\gamma)_{ik} = 0$ (Trt by Time interaction)
3. T.S.: $F_{AC} = \frac{MSAC}{MSE}$

4. R.R.: $F_{AC} \geq F_{\alpha,(a-1)(c-1),a(b-1)(c-1)}$
5. P-value: $P\left(F_{(a-1)(c-1),a(b-1)(c-1)} \geq F_{AC}\right)$

As this model is the same as the Split-Plot with the Whole-Plot units in a CRD, comparisons among Treatments and Times are conducted in the same manner.

Example 3.21: Heart Rates among Skydivers

A study compared heart rates at $c = 5$ time points of a skydiving flight among novice and experienced skydivers (Singley, Hale, and Russell, 2012, [34]). There were $b = 11$ novice (tandem) and 11 experienced (solo) jumpers ($a = 2$). The time points represented: baseline, take-off, 1524 meters, 3028 meters, and landing. Data have been generated that match reported means, F-statistic for time effect, and total sum of squares. Summary data with respect to treatment and time are given in Table 3.39, and the Analysis of Variance is given in Table 3.40. There is strong evidence of a time effect on heart rate ($F_C = 134.10, P < .0001$), but no evidence of a Group main effect of a Group/Time interaction ($F_A = 2.897, P = .1042; F_{AC} = 2.045, P = .0959$). The critical difference for comparing pairs of time means based on the Bonferroni method with $c_c^* = 5(5-1)/2 = 10$ comparisons is given below. Times 5 (88.59) and 2 (93.89) are not significantly different, all other pairs are. Although the groups (Expert vs Novice) are not significantly different based on the F-test, a 95% Confidence Interval for the difference is given below. Further, the ANOVA estimates of the variance components are given below.

$$\text{Times: } t_{.05/(2(10)),80} = 2.887 \quad \sqrt{\frac{2(52.24)}{2(11)}} = 2.179 \quad 2.887(2.179) = 6.291$$

$$\text{Groups: } 95.13 - 97.81 = -2.68 \quad t_{.025,20} = 2.086 \quad \sqrt{\frac{2(68.48)}{11(5)}} = 1.578 \quad -2.68 \pm 2.086(1.578) \equiv -2.68 \pm 3.29 \equiv (-5.97, 0.61)$$

$$s^2 = 53.24 \qquad s = 7.30 \qquad s_{B(A)}^2 = \frac{68.48 - 53.24}{5} = 3.90 \qquad s_{B(A)} = 1.98$$

Partial R Output is given below.

∇

	Expert	Novice	Mean
Time1	76.91	72.36	74.64
Time2	92.27	95.51	93.89
Time3	98.18	105.55	101.87
Time4	122.09	124.64	123.37
Time5	86.18	91.00	88.59
Mean	95.13	97.81	96.47

Table 3.39: Treatment/Time Heart Rate means among skydivers

		ANOVA			
Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F	$P(> F)$
Treatments (Group)	2-1=1	198.40	$\frac{198.40}{1} = 198.40$	$\frac{198.40}{68.48} = 2.897$	$P\left(F_{1,20} \geq 2.897\right) = .1042$
Subjects(Trts)	2(11-1)=20	1369.64	$\frac{1369.64}{20} = 68.48$		
Time	5-1=4	28555.23	$\frac{28555.23}{4} = 7138.81$	$\frac{7138.81}{53.24} = 134.10$	$P\left(F_{4,80} \geq 134.10\right) < .0001$
TrtxTime	1(4)=4	435.48	$\frac{435.48}{4} = 108.87$	$\frac{108.87}{53.24} = 2.045$	$P\left(F_{4,80} \geq 2.045\right) = .0959$
Error	2(11-1)(5-1)=80	5430.3	$\frac{4258.80}{80} = 53.24$		
TOTAL	2(11)(5)-1=109	34817.55			

Table 3.40: The Analysis of Variance Table for Skydiving Heart Rate Example

```
> ## AOV with incorrect error term for expGrp
> sd.mod1 <- aov(heartRt ~ expGrp*jumpTime + subjGrp:expGrp)
> summary(sd.mod1)
                Df Sum Sq Mean Sq F value Pr(>F)
expGrp           1    198     198   3.728  0.057 .
jumpTime         4  28555    7139 134.092 <2e-16 ***
expGrp:jumpTime  4    435     109   2.044  0.096 .
expGrp:subjGrp  20   1370      68   1.287  0.213
Residuals       80   4259      53
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> library(nlme)
> options(contrasts=c("contr.sum","contr.poly"))
>
> sd2 <- groupedData(heartRt ~ jumpTime | expGrp/subjGrp)
>
> ## Generalized Least Squares with Compound Symmetry
> sd.mod2a <- gls(heartRt ~ expGrp * jumpTime,
+    corr =  corCompSymm(form = ~ 1 | subjGrp))
> summary(sd.mod2a)
Generalized least squares fit by REML
  Model: heartRt ~ expGrp * jumpTime
  Data: NULL
       AIC      BIC    logLik
  747.6508 778.9128 -361.8254

Correlation Structure: Compound symmetry
 Formula: ~1 | subjGrp
 Parameter estimate(s):
       Rho
0.05420198

Coefficients:
                      Value Std.Error   t-value p-value
(Intercept)        96.46927 0.7890838 122.25479  0.0000
expGrp1            -1.34327 0.7890838  -1.70232  0.0918
jumpTime1         -21.83336 1.3913657 -15.69204  0.0000
jumpTime2          -2.57973 1.3913657  -1.85410  0.0667
jumpTime3           5.39573 1.3913657   3.87801  0.0002
jumpTime4          26.89618 1.3913657  19.33078  0.0000
expGrp1:jumpTime1   3.61736 1.3913657   2.59987  0.0107
expGrp1:jumpTime2  -0.27718 1.3913657  -0.19922  0.8425
expGrp1:jumpTime3  -2.34173 1.3913657  -1.68304  0.0955
```

```
expGrp1:jumpTime4   0.06873 1.3913657   0.04940  0.9607

 Correlation: 
                  (Intr) expGr1 jmpTm1 jmpTm2 jmpTm3 jmpTm4 eG1:T1 eG1:T2
expGrp1            0.00                                                 
jumpTime1          0.00   0.00                                          
jumpTime2          0.00   0.00  -0.25                                   
jumpTime3          0.00   0.00  -0.25  -0.25                            
jumpTime4          0.00   0.00  -0.25  -0.25  -0.25                     
expGrp1:jumpTime1  0.00   0.00   0.00   0.00   0.00   0.00              
expGrp1:jumpTime2  0.00   0.00   0.00   0.00   0.00   0.00  -0.25       
expGrp1:jumpTime3  0.00   0.00   0.00   0.00   0.00   0.00  -0.25  -0.25
expGrp1:jumpTime4  0.00   0.00   0.00   0.00   0.00   0.00  -0.25  -0.25
                  eG1:T3
expGrp1                 
jumpTime1               
jumpTime2               
jumpTime3               
jumpTime4               
expGrp1:jumpTime1       
expGrp1:jumpTime2       
expGrp1:jumpTime3       
expGrp1:jumpTime4 -0.25 

Standardized residuals:
          Min            Q1           Med            Q3           Max 
-2.035431e+00 -7.970632e-01  4.735347e-15  6.970365e-01  1.948673e+00 

Residual standard error: 7.502542 
Degrees of freedom: 110 total; 100 residual
> anova(sd.mod2a)
Denom. DF: 100 
                numDF   F-value p-value
(Intercept)         1 14946.234  <.0001
expGrp              1     2.898  0.0918
jumpTime            4   134.092  <.0001
expGrp:jumpTime     4     2.044  0.0939
> AIC(sd.mod2a)
[1] 747.6508
> getVarCov(sd.mod2a)
Marginal variance covariance matrix
        [,1]    [,2]    [,3]    [,4]    [,5]
[1,] 56.2880  3.0509  3.0509  3.0509  3.0509
[2,]  3.0509 56.2880  3.0509  3.0509  3.0509
[3,]  3.0509  3.0509 56.2880  3.0509  3.0509
[4,]  3.0509  3.0509  3.0509 56.2880  3.0509
[5,]  3.0509  3.0509  3.0509  3.0509 56.2880
  Standard Deviations: 7.5025 7.5025 7.5025 7.5025 7.5025 
> intervals(sd.mod2a)
Approximate 95% confidence intervals

 Coefficients:
                        lower         est.       upper
(Intercept)        94.9037530  96.46927273  98.0347924
expGrp1            -2.9087924  -1.34327273   0.2222470
jumpTime1         -24.5937935 -21.83336364 -19.0729338
jumpTime2          -5.3401572  -2.57972727   0.1807026
jumpTime3           2.6352974   5.39572727   8.1561572
jumpTime4          24.1357519  26.89618182  29.6566117
expGrp1:jumpTime1   0.8569338   3.61736364   6.3777935
expGrp1:jumpTime2  -3.0376117  -0.27718182   2.4832481
expGrp1:jumpTime3  -5.1021572  -2.34172727   0.4187026
expGrp1:jumpTime4  -2.6917026   0.06872727   2.8291572
attr(,"label")
```

```
[1] "Coefficients:"

 Correlation structure:
         lower       est.     upper
Rho -0.07678504 0.05420198 0.2392377
attr(,"label")
[1] "Correlation structure:"

 Residual standard error:
   lower     est.    upper
6.526298 7.502542 8.624818
```

The univariate model described above assumes that the measurements at the various time points have equal variances and correlations between pairs of measurements within subjects are all equal (Compound Symmetry). If that assumption holds, or the less stringent Huynh-Feldt assumption that the variance of the differences among all pairs of measurements within subjects are equal, then this analysis is appropriate. When these assumptions do not hold, researchers often conduct a multivariate Analysis of Variance and make adjustments to the degrees of freedom for the within subjects factors (Time and Treatment/Time interaction in this case). There are two widely used adjustments: Greenhouse-Geisser and Huynh-Feldt. These methods use methods beyond the scope of this course and are not discussed here.

With the advent of more flexible mixed model statistical programs (first developed as **Proc Mixed** in the SAS System), it is possible to allow for more complex correlation structure within subjects. For instance, if measurements are made at equally space time points over an extended period of time, the correlations between measurements further apart in time may tend to decrease multiplicatively. This is often modeled as an Autoregressive process of order 1 (AR(1)) for the errors. For example based tutorials of this based on Proc Mixed, see the following two papers (Littell, Pendergast, and Natarajan, 2000, [20]; Bagiella, Sloan, and Heitjan, 2000, [6]). There are many possibilities for the correlation/covariance structures that are covered in the papers. The most general structure allows all c variances and all $c(c-1)/2$ covariances to be distinct. That is considered to be the unstructured (symmetric) case. Examples of many variance/covariance structures in S, and applicable in R are given in Pinheiro and Bates, 2000, [29].

Example 3.22: Heart Rates among Skydivers

The model fit in Example 3.21 for the Skydivers' heart rate measurements assumed a Compound Symmetry pattern for the within subjects errors. This assumes that the variances in measurements are constant over the $c = 5$ time points, and that the correlation/covariance among measurements at all pairs of times are the same. Now, consider the model fit allowing for an unstructured covariance structure within subjects. The R output is given below (there is no closed form way of obtaining this). The F-tests for Group and Group/Time interaction are not quite significant at the $\alpha = 0.05$ level, but are closer to significance ($F_A = 3.871, P = .0631; F_{AC} = 2.256, P = .0703$). There is little evidence of the variances at the various time points being different: the variance function estimates (standard deviation multipliers, with time 1 as the reference) range from 0.936 to 1.093. The individual correlations range from -0.214 to 0.326. Note that this model has $c + c(c-1)/2 = 5 + 5(5-1)/2 = 15$ variance parameters, while the model in Example 3.21 had 2. The Likelihood Ratio test to test between the two models yields a Chi-square statistic of 5.729 with $15 - 2 = 13$ degrees of freedom and a P-value of .9555. There is no reason to prefer the more complex model over the simpler model for this data.

∇

```
> ## Generalized Least Squares with Unstructured Covariance Structure
> sd.mod3a <- gls(heartRt ~ expGrp * jumpTime,
+    corr =  corCompSymm(form = ~ 1 | subjGrp),
+    weight = varIdent(form = ~ 1 | jumpTime))
> summary(sd.mod3a)
Generalized least squares fit by REML
  Model: heartRt ~ expGrp * jumpTime
  Data: NULL
       AIC      BIC    logLik
  755.1105 796.7932 -361.5552

Correlation Structure: Compound symmetry
 Formula: ~1 | subjGrp
 Parameter estimate(s):
       Rho
0.05968999
Variance function:
 Structure: Different standard deviations per stratum
 Formula: ~1 | jumpTime
 Parameter estimates:
        1         2         3         4         5
1.0000000 1.0269539 1.1082193 0.9470501 0.9917593

Coefficients:
                      Value Std.Error   t-value p-value
(Intercept)        96.46927 0.7963595 121.13785  0.0000
expGrp1            -1.34327 0.7963595  -1.68677  0.0948
jumpTime1         -21.83336 1.3715664 -15.91856  0.0000
jumpTime2          -2.57973 1.3991575  -1.84377  0.0682
jumpTime3           5.39573 1.4838254   3.63636  0.0004
jumpTime4          26.89618 1.3181668  20.40423  0.0000
expGrp1:jumpTime1   3.61736 1.3715664   2.63740  0.0097
expGrp1:jumpTime2  -0.27718 1.3991575  -0.19811  0.8434
expGrp1:jumpTime3  -2.34173 1.4838254  -1.57817  0.1177
expGrp1:jumpTime4   0.06873 1.3181668   0.05214  0.9585

 Correlation:
                  (Intr) expGr1 jmpTm1 jmpTm2 jmpTm3 jmpTm4 eG1:T1 eG1:T2
expGrp1            0.000
jumpTime1         -0.016  0.000
jumpTime2          0.011  0.000 -0.248
jumpTime3          0.089  0.000 -0.274 -0.281
jumpTime4         -0.070  0.000 -0.223 -0.232 -0.261
expGrp1:jumpTime1  0.000 -0.016  0.000  0.000  0.000  0.000
expGrp1:jumpTime2  0.000  0.011  0.000  0.000  0.000  0.000 -0.248
expGrp1:jumpTime3  0.000  0.089  0.000  0.000  0.000  0.000 -0.274 -0.281
expGrp1:jumpTime4  0.000 -0.070  0.000  0.000  0.000  0.000 -0.223 -0.232
                  eG1:T3
expGrp1
jumpTime1
jumpTime2
jumpTime3
jumpTime4
expGrp1:jumpTime1
expGrp1:jumpTime2
expGrp1:jumpTime3
expGrp1:jumpTime4 -0.261

Standardized residuals:
          Min            Q1           Med            Q3           Max
-2.084396e+00 -8.046225e-01 -1.094309e-14  7.051100e-01  2.089758e+00

Residual standard error: 7.387175
Degrees of freedom: 110 total; 100 residual
> anova(sd.mod3a)
```

```
Denom. DF: 100
                numDF   F-value p-value
(Intercept)         1 14989.005  <.0001
expGrp              1     2.450  0.1206
jumpTime            4   147.423  <.0001
expGrp:jumpTime     4     2.009  0.0989
> AIC(sd.mod3a)
[1] 755.1105
> getVarCov(sd.mod3a)
Marginal variance covariance matrix
        [,1]    [,2]    [,3]    [,4]    [,5]
[1,] 54.5700  3.3451  3.6098  3.0848  3.2305
[2,]  3.3451 57.5520  3.7071  3.1680  3.3175
[3,]  3.6098  3.7071 67.0210  3.4187  3.5801
[4,]  3.0848  3.1680  3.4187 48.9440  3.0594
[5,]  3.2305  3.3175  3.5801  3.0594 53.6750
  Standard Deviations: 7.3872 7.5863 8.1866 6.996 7.3263
> intervals(sd.mod3a)
Approximate 95% confidence intervals

 Coefficients:
                       lower         est.       upper
(Intercept)        94.889318  96.46927273  98.0492272
expGrp1            -2.923227  -1.34327273   0.2366818
jumpTime1         -24.554512 -21.83336364 -19.1122150
jumpTime2          -5.355616  -2.57972727   0.1961613
jumpTime3           2.451860   5.39572727   8.3395947
jumpTime4          24.280976  26.89618182  29.5113873
expGrp1:jumpTime1   0.896215   3.61736364   6.3385122
expGrp1:jumpTime2  -3.053070  -0.27718182   2.4987068
expGrp1:jumpTime3  -5.285595  -2.34172727   0.6021401
expGrp1:jumpTime4  -2.546478   0.06872727   2.6839328
attr(,"label")
[1] "Coefficients:"

 Correlation structure:
          lower       est.     upper
Rho -0.07419729 0.05968999 0.2482478
attr(,"label")
[1] "Correlation structure:"

 Variance function:
      lower      est.    upper
2 0.6396169 1.0269539 1.648853
3 0.6995196 1.1082193 1.755705
4 0.6006061 0.9470501 1.493331
5 0.6100048 0.9917593 1.612424
attr(,"label")
[1] "Variance function:"

 Residual standard error:
    lower      est.     upper
 5.282376  7.387175 10.330645
>
>
> anova(sd.mod2a, sd.mod3a)
         Model df      AIC      BIC    logLik   Test   L.Ratio p-value
sd.mod2a     1 12 747.6508 778.9128 -361.8254
sd.mod3a     2 16 755.1105 796.7932 -361.5552 1 vs 2 0.5403307  0.9695
```

3.5 R Programs for Chapter 3 Examples

3.5.1 Halo Effect - Essay Evaluation

```
halo <- read.table("http://www.stat.ufl.edu/~winner/data/halo1.dat",
     header=F, col.names=c("essayqual","picture","grade"))
attach(halo)

essayqual <- factor(essayqual)
picture <- factor(picture)

interaction.plot(essayqual,picture,grade)

## Additive Model - Trt Level 1 is Reference
halo.mod1 <- aov(grade ~ essayqual + picture)
anova(halo.mod1)
summary.lm(halo.mod1)
TukeyHSD(halo.mod1,"essayqual")
TukeyHSD(halo.mod1,"picture")

## Interaction Model
halo.mod2 <- aov(grade ~ essayqual*picture)
anova(halo.mod2)
summary.lm(halo.mod2)

## Compare Additive and Interaction Models
anova(halo.mod1, halo.mod2)

## Switch to Trt Effects sum to 0
options(contrasts=c("contr.sum","contr.poly"))
halo.mod3 <- aov(grade ~ essayqual + picture)
anova(halo.mod3)
summary.lm(halo.mod3)

halo.mod4 <- aov(grade ~ essayqual*picture)
anova(halo.mod4)
summary.lm(halo.mod4)

anova(halo.mod3, halo.mod4)
```

3.5.2 Penetration of Arrowheads by Clothing Fit and Type

```
arrow1 <- read.csv("http://www.stat.ufl.edu/~winner/data/arrowhead_clothing.csv")
attach(arrow1); names(arrow1)

## Select only the first arrowhead type data (Bullet)
Y1 <- pntrt[arrowhead == 1]
clothFit1 <- clothFit[arrowhead == 1]
clothType1 <- clothType[arrowhead == 1]

clothFit1 <- factor(clothFit1, levels=1:2, labels=c("Tight","Loose"))
clothType1 <- factor(clothType1, levels=1:3,
     labels=c("T-shirt", "Jeans65Cttn", "Jeans95Cttn"))

options(contrasts=c("contr.sum","contr.poly"))
arrow.mod1 <- aov(Y1 ~ clothFit1 * clothType1)
anova(arrow.mod1)
```

```
summary.lm(arrow.mod1)

interaction.plot(clothType1, clothFit1, Y1)
sum((Y1 - mean(Y1))^2)   ## Total SS
```

3.5.3 Lead Content in Lip Products

```
lead_lip <- read.csv("http://www.stat.ufl.edu/~winner/data/lead_lipstick.csv")
attach(lead_lip); names(lead_lip)

tapply(Pb, list(shade, prodType), length)
tapply(Pb, list(shade, prodType), mean)
tapply(Pb, list(shade, prodType), sd)

interaction.plot(shade, prodType, Pb)

## Generate X's for Regression Model
n.tot <- length(Pb)
X1.A <- rep(0,n.tot)
X2.A <- rep(0,n.tot)
X3.A <- rep(0,n.tot)
X1.B <- rep(0,n.tot)

X1.A <- ifelse(shade == "Red", 1, ifelse(shade == "Brown", -1, 0))
X2.A <- ifelse(shade == "Purple", 1, ifelse(shade == "Brown", -1, 0))
X3.A <- ifelse(shade == "Pink", 1, ifelse(shade == "Brown", -1, 0))
X1.B <- ifelse(prodType == "LP", 1, -1)

## Full Model
ll.mod1 <- lm(Pb ~ X1.A + X2.A + X3.A + X1.B + I(X1.A * X1.B) +
     I(X2.A * X1.B) + I(X3.A * X1.B))
summary(ll.mod1)
anova(ll.mod1)

## Drop Interactions
ll.mod2 <- lm(Pb ~ X1.A + X2.A + X3.A + X1.B)
summary(ll.mod2)
anova(ll.mod2)

## Drop Factor A
ll.mod3 <- lm(Pb ~ X1.B + I(X1.A * X1.B) +
     I(X2.A * X1.B) + I(X3.A * X1.B))
summary(ll.mod3)
anova(ll.mod3)

## Drop Factor B
ll.mod4 <- lm(Pb ~ X1.A + X2.A + X3.A + I(X1.A * X1.B) +
     I(X2.A * X1.B) + I(X3.A * X1.B))
summary(ll.mod4)
anova(ll.mod4)

## Drop Factor A and Interactions
ll.mod5 <- lm(Pb ~ X1.B)
summary(ll.mod5)
anova(ll.mod5)

## Drop Factor B and Ineractions
ll.mod6 <- lm(Pb ~ X1.A + X2.A + X3.A)
summary(ll.mod6)
anova(ll.mod6)
```

```
## F-tests for Model Comparisons
anova(ll.mod2, ll.mod1)
anova(ll.mod3, ll.mod1)
anova(ll.mod4, ll.mod1)
anova(ll.mod5, ll.mod2)
anova(ll.mod6, ll.mod2)
anova(ll.mod5, ll.mod1)
anova(ll.mod6, ll.mod1)

options(contrasts=c("contr.sum","contr.poly"))
ll.aov1 <- aov(Pb ~ shade * prodType)
anova(ll.aov1)
ll.aov2 <- aov(Pb ~ shade + prodType)
anova(ll.aov2)

library(car)
Anova(ll.aov1, Type="II")
Anova(ll.aov1, Type="III")
Anova(ll.aov2, Type="II")
Anova(ll.aov2, Type="III")


## 3-Way ANOVA
ll.aov3 <- aov(Pb ~ shade * prodType * priceCatgry)
anova(ll.aov3)
ll.aov4 <- aov(Pb ~ shade + prodType + priceCatgry)
anova(ll.aov4)

library(car)
Anova(ll.aov3, Type="II")
Anova(ll.aov3, Type="III")
Anova(ll.aov4, Type="II")
Anova(ll.aov4, Type="III")
```

3.5.4 Oil Holding Capacity of Banana Cultivars

```
ban1 <- read.table("http://www.stat.ufl.edu/~winner/data/banana_pretreat.dat",
    header=F,col.names=c("cultivar","acidType","acidDose","OHC"))
attach(ban1)

cultivar <- factor(cultivar)
acidType <- factor(acidType)
acidDose <- factor(acidDose)

ban.mod1 <- aov(OHC ~ cultivar * acidType * acidDose)
anova(ban.mod1)
```

3.5.5 Women's Professional Bowling Scores - 2009

```
wpba2009 <- read.table("http://www.stat.ufl.edu/~winner/data/wpba2009.dat",
  header=F, col.names=c("bowler","pattern","set","game","score"))

attach(wpba2009)

bowler <- factor(bowler)
pattern <- factor(pattern)
```

```
tapply(score,bowler,mean); tapply(score,bowler,sd)
tapply(score,pattern,mean); tapply(score,pattern,sd)
tapply(score,list(bowler,pattern),mean); tapply(score,list(bowler,pattern),sd)

options(contrasts=c("contr.sum","contr.poly"))

wpba.mod1 <- aov(score ~ pattern + bowler + bowler:pattern)
summary(wpba.mod1)

interaction.plot(bowler, pattern, score)

wpba.mod2 <- aov(score ~ pattern + bowler + Error(bowler:pattern))
summary(wpba.mod2)

library(nlme)
wpba.mod3 <- lme(fixed = score ~ pattern, random = ~1|bowler/pattern)
summary(wpba.mod3)
intervals(wpba.mod3)
anova(wpba.mod3)

library(lmerTest)
wpba.mod4 <- lmer(score~pattern+(1|bowler)+(1|pattern:bowler))
summary(wpba.mod4)
anova(wpba.mod4)
lsmeans(wpba.mod4)
difflsmeans(wpba.mod4)
confint(wpba.mod4)
```

3.5.6 Shoveling Times for Spatulas

```
spatula <- read.table("http://www.stat.ufl.edu/~winner/data/chopstick3.dat",
                     header=F,col.names=c("length","angle","subject","shovtime"))
attach(spatula)
names(spatula)
# spatula
length <- factor(length)
angle <- factor(angle)
subject <- factor(subject)

options(contrasts=c("contr.sum","contr.poly"))
spat.mod1 <- aov(shovtime ~ length * angle * subject)
anova(spat.mod1)

# install.packages("lmerTest")
library(lmerTest)

spat.mod2 <- lmer(shovtime ~ length*angle + (1|subject) + (1|subject:length) +
            (1|subject:angle))
summary(spat.mod2)
anova(spat.mod2)
difflsmeans(spat.mod2)
rand(spat.mod2)
```

3.5.7 Reliability of Foot Joint Inversion Measurements

```
foot <- read.table("http://www.stat.ufl.edu/~winner/data/biometer_foot.dat",
        header=F,col.names=c("subj","inv_env","tester","day","trial","angle"))
```

```
attach(foot)
subj <- factor(subj)
inv_env <- factor(inv_env)
tester <- factor(tester)
day <- factor(day)

## Select only Inversion measurements
subjInv <- subj[inv_env == 1]
testerInv <- tester[inv_env == 1]
dayInv <- day[inv_env == 1]
angleInv <- angle[inv_env == 1]

foot.mod1 <- aov(angleInv ~ testerInv * subjInv * dayInv)
anova(foot.mod1)

# install.packages("lmerTest")
library(lmerTest)

options(contrasts=c("contr.sum","contr.poly"))
foot.mod2 <- lmer(angleInv ~ testerInv + (1|subjInv) + (1|dayInv) +
 (1|testerInv:subjInv) + (1|testerInv:dayInv) + (1|subjInv:dayInv) +
 (1|testerInv:subjInv:dayInv))
summary(foot.mod2)
anova(foot.mod2)
difflsmeans(foot.mod2)
rand(foot.mod2)
```

3.5.8 Repeatability and Reproducibility of Measurements

```
wd <- read.csv("http://www.stat.ufl.edu/~winner/data/wood_drill_gage.csv")
attach(wd); names(wd)

Part <- factor(Part)
Operator <- factor(Operator)

Ymeas <- 100*Ymeas

options(contrasts=c("contr.sum","contr.poly"))
wd1.mod1 <- aov(Ymeas ~ Part*Operator)
anova(wd1.mod1)

library(nlme)
wd2 <- groupedData(Ymeas ~ 1 | Part/Operator)  ## Set up grouped data

wd.mod2 <- lme(Ymeas ~ 1, data=wd2, random = ~ 1 | Part/Operator)
summary(wd.mod2)

library(lmerTest)
wd.mod3 <- lmer(Ymeas ~ 1 + (1|Part) + (1|Operator) + (1|Part:Operator))
summary(wd.mod3)
```

3.5.9 Measurement of Alcohol Content In Distilled Spirits

```
wac <- read.csv("whisky_alccont.csv")
attach(wac); names(wac)

spiritType <- factor(spiritType)
```

```
brandSprt <- factor(brandSprt)
Y <- alcCntnt - labelAC

dist.mod1 <- aov(Y ~ spiritType + spiritType/brandSprt)
anova(dist.mod1)
TukeyHSD(dist.mod1,"spiritType")
```

3.5.10 Momentum Measurements for Animal Traps

```
trap <- read.csv("http://www.stat.ufl.edu/~winner/data/animal_trap.csv")
attach(trap); names(trap)

model <- factor(model)
trapModel <- factor(trapModel)

trap.mod1 <- aov(momentum ~ model + model/trapModel)
anova(trap.mod1)

trap.mod2 <- aov(momentum ~ model + Error(trapModel))
summary(trap.mod2)

options(contrasts=c("contr.sum","contr.poly"))

library(nlme)

trap2 <- groupedData(momentum ~ model | model/trapModel)

trap.mod3 <- lme(momentum ~ model, data=trap2, random= ~1|model/trapModel)
summary(trap.mod3)

library(lmerTest)
trap.mod4 <- lmer(momentum ~ model + (1|model:trapModel))
summary(trap.mod4)
```

3.5.11 Variation in Semiconductor Wafers

```
semicon <- read.table("http://www.stat.ufl.edu/~winner/data/semicon_qual.dat",
   header=F,col.names=c("batch","waferBtch","wafer","location","Y"))
attach(semicon)

batch <- factor(batch)
wafer <- factor(wafer)

semi.mod1 <- aov(Y ~ batch + batch/wafer)
anova(semi.mod1)

library(lmerTest)

semi.mod2 <- lmer(Y ~ 1 + (1|batch/wafer))
summary(semi.mod2)
```

3.5.12 Axion Densities in Eyes of Normal and Alzheimers Patients

```
axioneyes <- read.csv("http://www.stat.ufl.edu/~winner/data/alzheimers_eyes.csv")
attach(axioneyes); names(axioneyes)

subject <- factor(subject)
alz_grp <- factor(alz_grp)
eye <- factor(eye)

options(contrasts=c("contr.sum","contr.poly"))
eyes.mod1 <- aov(axondens ~ alz_grp + alz_grp/subject + eye + alz_grp:eye)
summary(eyes.mod1)
eyes.mod2 <- aov(axondens ~ alz_grp * eye + Error(subject))
summary(eyes.mod2)

library(lmerTest)

eyes.mod3 <- lmer(axondens ~ alz_grp * eye + (1|alz_grp:subject))
summary(eyes.mod3)
anova(eyes.mod3)
difflsmeans(eyes.mod3)
```

3.5.13 Chymosin Treatment and Ripening Time Effects on Mozzarella Cheese

```
chym <- read.csv("http://www.stat.ufl.edu/~winner/data/camel_cheese.csv")
attach(chym); names(chym)

c.trt <- factor(c.trt)
c.blk <- factor(c.blk)
c.time <- factor(c.time)

options(contrasts=c("contr.sum","contr.poly"))
chym.mod1 <- aov(blister ~ c.trt*c.blk + c.time + c.trt:c.time)
summary(chym.mod1)

library(lmerTest)

chy.mod2 <- lmer(blister ~ c.trt*c.time + (1|c.blk) + (1|c.trt:c.blk))
summary(chy.mod2)
anova(chy.mod2)
difflsmeans(chy.mod2)
```

3.5.14 Heart Rates Among Skydivers

```
sd1 <- read.csv("http://www.stat.ufl.edu/~winner/data/skydive.csv")
attach(sd1); names(sd1)

expGrp <- factor(expGrp)
subjGrp <- factor(subjGrp)
jumpTime <- factor(jumpTime)

## AOV with incorrect error term for expGrp
sd.mod1 <- aov(heartRt ~ expGrp*jumpTime + subjGrp:expGrp)
summary(sd.mod1)
```

```
library(nlme)
options(contrasts=c("contr.sum","contr.poly"))

sd2 <- groupedData(heartRt ~ jumpTime | expGrp/subjGrp)

## Generalized Least Squares with Compound Symmetry
sd.mod2a <- gls(heartRt ~ expGrp * jumpTime,
   corr =  corCompSymm(form = ~ 1 | subjGrp))
summary(sd.mod2a)
anova(sd.mod2a)
AIC(sd.mod2a)
getVarCov(sd.mod2a)
intervals(sd.mod2a)

## Generalized Least Squares with Unstructured Covariance Structure
sd.mod3a <- gls(heartRt ~ expGrp * jumpTime,
   corr =  corCompSymm(form = ~ 1 | subjGrp),
   weight = varIdent(form = ~ 1 | jumpTime))
summary(sd.mod3a)
anova(sd.mod3a)
AIC(sd.mod3a)
getVarCov(sd.mod3a)
intervals(sd.mod3a)

anova(sd.mod2a, sd.mod3a)
```

Chapter 4

Analysis of Covariance

The Analysis of Covariance is generally applied when the goal is to compare treatments or groups in terms of a numeric response variable after controlling for one or more numeric predictors (covariates) that are believed to be related to the response. In many situations the covariate is a baseline or pre-treatment score for the unit, and the response is the post-treatment score. There can be one or more treatment factors and one or more covariates. Mechanically, the analysis is fit as a multiple regression model with dummy variables for the treatments and numeric variable(s) for the covariate(s).

4.1 Model with 1 Treatment Factor and 1 Covariate

Consider the model with a single factor with a levels, a single covariate X, and response variable Y. Define $a-1$ dummy variables $W_1, \ldots, W_{a-1}$ as follow.

$$W_i = \begin{cases} 1 & : \text{ if Factor } A \text{ is at level } i \quad i = 1, \ldots, a-1 \\ 0 & : \text{ otherwise} \end{cases}$$

Observations for treatment a receive $W_1 = \cdots = W_{a-1} = 0$ (as in the multiple regression models with categorical predictors). The use of W_i instead of X is to simplify notation. Many practitioners center the covariate(s) around their mean(s) when fitting the model, and that will be done here as well, as it makes interpreting parameters easier. First, the additive model is described, followed by a model with interaction.

4.1.1 Additive Model - Common Slopes

Let Y_{ij} represent the the response for the j^{th} unit within the i^{th} treatment or group, X_{ij} be its covariate value, and $W_{1ij}, \ldots, W_{a-1,ij}$ be its dummy values for the factors. The additive model can be written as follows.

$$Y_{ij} = \beta_0 + \beta_1 \left(X_{ij} - \overline{X}_{..}\right) + \gamma_1 W_{1ij} + \cdots + \gamma_{a-1} W_{a-1,ij} + \epsilon_{ij} \quad i = 1, \ldots, a; j = 1, \ldots, n_i \quad \epsilon_{ij} \sim N\left(0, \sigma^2\right)$$

In this model, the slope relating the covariate to the response is assumed to be the same for all treatments/groups. An interaction model allows the slopes to differ, and is considered below. In this model, β_0 is interpreted as the mean for treatment a when the covariate is at the overall mean level $\overline{X}_{..}$, β_1 is the slope relating the response to the covariate, and γ_i represents the difference in the mean response between treatments i and a, **controlling for the covariate**. If all treatments had the same mean for their covariates, there would be no need for including the covariate, and it would be more efficient to conduct a 1-Way ANOVA.

The test for treatment effects, after controlling for the covariate can be done using the general linear test using a comparison of a Complete model (with the covariate and treatment dummy variables) and a Reduced model with only the covariate, as was done in Chapter 2.

Example 4.1: Skin Smoothing Study

A study compared $a = 3$ treatments in terms of smoothing skin (Ma'Or, Yehuda, and Voss, 1997, [22]). The three treatments were: Formulated gel ($i = 1$), Formulated gel plus 1% Dead Sea concentrate ($i = 2$), and Placebo Control ($i = 3$). The response was a roughness value measured with laser on the skin surface after 4 weeks of treatment, the covariate was the baseline (pre-treatment) value of the same roughness value. There were $n_1 = n_2 = n_3 = 20$ subjects per treatment. Data have been generated to match mean, SD, min, max, and correlations between pre- and post-treatment scores within each treatment. Summary statistics are given in Table 4.1. A plot of the data is given in Figure 4.1, along with a vertical line at $\overline{X}_{..} = 186.89$.

The first (Reduced) model contains only the centered covariate as a predictor. The model summary is given below.

$$\text{Model 1: } \hat{Y} = 137.537 + 0.697\left(X - \overline{X}_{..}\right) \quad SSE_1 = 46773 \quad df_{E1} = 60 - 2 = 58 \quad R_1^2 = .7195$$

The second (Full) model contains the centered covariate as well as dummy variables for Gel formulation (W_1) and Gel plus Dead Sea concentrate (W_2). The model summary is given below.

$$\text{Model 2: } \hat{Y} = 167.118 + 0.707\left(X - \overline{X}_{..}\right) - 32.245W_1 - 56.599W_2 \quad SSE_2 = 14666 \quad df_{E2} = 60 - 4 = 56 \quad R_2^2 = .9121$$

A test of whether there are differences among the treatment effects, controlling for pre-treatment score is given below, making use of the F-test for a subset of regression coefficients in Section 2.3.

$$H_0 : \gamma_1 = \gamma_2 = 0 \quad TS : F_{obs} = \frac{\left[\frac{46773-14666}{58-56}\right]}{\left[\frac{14666}{56}\right]} = \frac{16053.5}{261.9} = 61.3 \quad P\left(F_{2,56} \geq 61.3\right) < .0001$$

There is strong evidence of differences among the treatment effects. The t-tests, given with the R output below show that both gel only and gel plus Dead Sea concentrate have significantly lower roughness scores

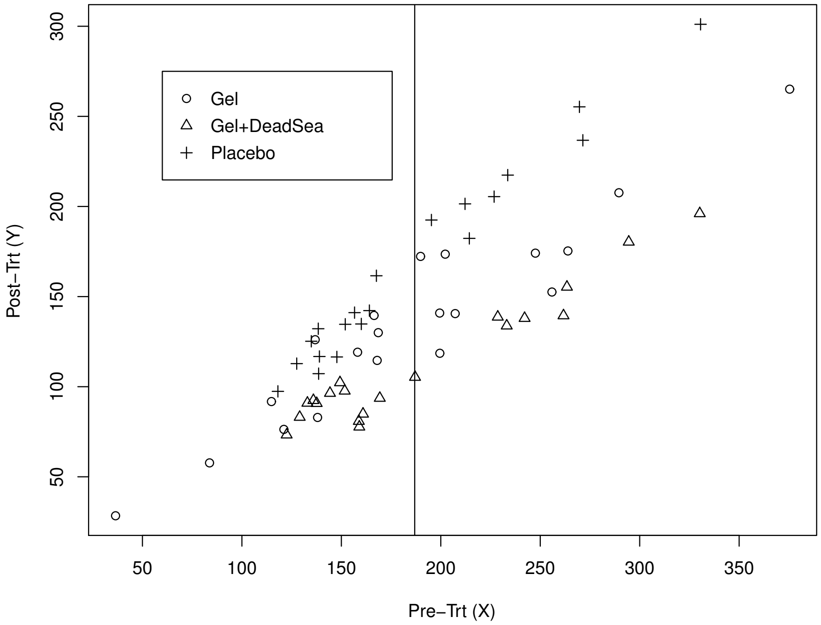

Figure 4.1: Plot of Post-Treatment Roughness Scores versus Pre-Treatment by Treatment Group

than the control group, controlling for baseline score. For gel only $t_1 = -32.245/5.118 = -6.301$, while for gel plus Dead Sea concentrate $t_2 = -56.599/5.120 = -11.035$; both P-values are less than .0001. A plot of the additive model (using the original X values) is given in Figure 4.2.

∇

	Gel ($i = 1$)	Gel+DeadSea ($i = 2$)	Placebo ($i = 3$)
Pre-Trt (X)	186.14 (76.44)	189.61 (61.9)	184.93 (57.6)
Post-Trt (Y)	134.34 (54.06)	112.54 (35.53)	165.73 (55.83)

Table 4.1: Roughness Means (SDs) by Treatment Pre-Treatment and Post-Treatment (20 subjects per treatment)

```
> ## Covariate Only Model
> dsm.mod1 <- lm(post_y ~ xc)
> summary(dsm.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.53667    3.66612   37.52   <2e-16 ***
xc            0.69686    0.05713   12.20   <2e-16 ***

Residual standard error: 28.4 on 58 degrees of freedom
Multiple R-squared:  0.7195,    Adjusted R-squared:  0.7147
```

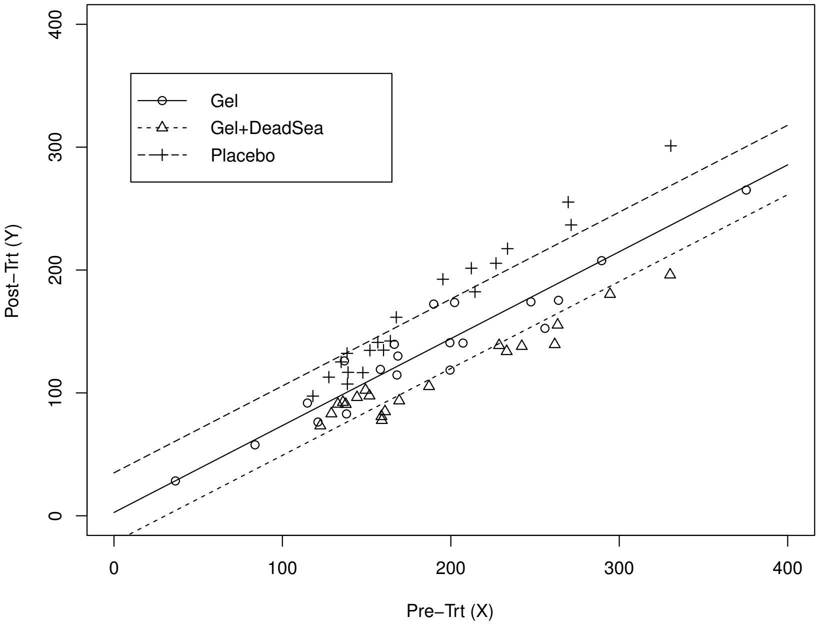

Figure 4.2: Plot of Additive Model - Skin Softening Study

```
F-statistic: 148.8 on 1 and 58 DF,  p-value: < 2.2e-16

> anova(dsm.mod1)
Analysis of Variance Table
Response: post_y
          Df Sum Sq Mean Sq F value    Pr(>F)
xc         1 119991  119991  148.79 < 2.2e-16 ***
Residuals 58  46773     806

> ## Additive Model
> dsm.mod2 <- lm(post_y ~ xc + gel + gelDS)
> summary(dsm.mod2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 167.11784    3.61923  46.175  < 2e-16 ***
xc            0.70732    0.03257  21.716  < 2e-16 ***
gel         -32.24500    5.11771  -6.301 4.91e-08 ***
gelDS       -56.49853    5.11983 -11.035 1.14e-15 ***

Residual standard error: 16.18 on 56 degrees of freedom
Multiple R-squared:  0.9121,    Adjusted R-squared:  0.9073
F-statistic: 193.6 on 3 and 56 DF,  p-value: < 2.2e-16

> anova(dsm.mod2)
Analysis of Variance Table

Response: post_y
          Df Sum Sq Mean Sq  F value    Pr(>F)
xc         1 119991  119991 458.1660 < 2.2e-16 ***
gel        1    214     214   0.8177    0.3697
gelDS      1  31893   31893 121.7764 1.141e-15 ***
```

```
Residuals 56  14666     262

> anova(dsm.mod1,dsm.mod2)
Analysis of Variance Table
Model 1: post_y ~ xc
Model 2: post_y ~ xc + gel + gelDS
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     58 46773
2     56 14666  2     32107 61.297 7.888e-15 ***
```

The **Adjusted Means** are the estimated responses for the treatment groups when the covariate(s) are at their mean levels. For the model with the X values centered, that removes the term involving $\hat{\beta}_1$ from the fitted equation. Note that the intercept, $\hat{\beta}_0$ is the adjusted mean for the reference group.

$$\overline{Y}_1^{\text{Adj}} = \hat{\beta}_0 + \hat{\gamma}_1 \quad \ldots \quad \overline{Y}_{a-1}^{\text{Adj}} = \hat{\beta}_0 + \hat{\gamma}_{a-1} \quad \overline{Y}_a^{\text{Adj}} = \hat{\beta}_0$$

While it is easy to obtain the standard errors of the adjusted mean for the reference group, and the differences of the adjusted means for the $a-1$ other groups with the reference group from summary output, other means and differences must be obtained from the variance-covariance matrix of the estimated regression coefficients. This can be obtained with the **vcov** function applied to an **lm** object (and many others) in R. For instance, to obtain the estimated standard error for the adjusted mean for Treatment 1 or the difference in adjusted means for Treatments 1 and 2, the following results can be used.

$$\hat{SE}\left\{\overline{Y}_1^{\text{Adj}}\right\} = \hat{SE}\left\{\hat{\beta}_0 + \hat{\gamma}_1\right\} = \sqrt{\hat{V}\left\{\hat{\beta}_0\right\} + \hat{V}\left\{\hat{\gamma}_1\right\} + 2\hat{\text{COV}}\left\{\hat{\beta}_0, \hat{\gamma}_1\right\}}$$

$$\hat{SE}\left\{\overline{Y}_1^{\text{Adj}} - \overline{Y}_2^{\text{Adj}}\right\} = \hat{SE}\left\{\hat{\gamma}_1 - \hat{\gamma}_2\right\} = \sqrt{\hat{V}\left\{\hat{\gamma}_1\right\} + \hat{V}\left\{\hat{\gamma}_2\right\} - 2\hat{\text{COV}}\left\{\hat{\gamma}_1, \hat{\gamma}_2\right\}}$$

Example 4.2: Skin Smoothing Study

The estimated adjusted means and their differences are given below for the skin smoothing study.

$$\overline{Y}_1^{\text{Adj}} = 167.118 - 32.245 = 134.873 \qquad \overline{Y}_2^{\text{Adj}} = 167.118 - 56.499 = 110.619 \qquad \overline{Y}_3^{\text{Adj}} = 167.118$$

$$\overline{Y}_1^{\text{Adj}} - \overline{Y}_2^{\text{Adj}} = -32.245 - (-56.499) = 24.254 \quad \overline{Y}_1^{\text{Adj}} - \overline{Y}_3^{\text{Adj}} = -32.245 \quad \overline{Y}_2^{\text{Adj}} - \overline{Y}_3^{\text{Adj}} = -56.499$$

The estimated variance-covariance matrix is given below. Estimated standard errors and t-tests and/or Confidence Intervals are obtained from it.

```
> round(vcov(dsm.mod2),4)
            (Intercept)      xc      gel    gelDS
```

```
(Intercept)     13.0988  0.0021 -13.0972 -13.1045
xc               0.0021  0.0011  -0.0013  -0.0050
gel            -13.0972 -0.0013  26.1910  13.1007
gelDS          -13.1045 -0.0050  13.1007  26.2127
```

$$\hat{SE}\left\{\overline{Y}_1^{\text{Adj}}\right\} = \sqrt{13.0988 + 26.1910 + 2(-13.0972)} = \sqrt{13.0954} = 3.619$$

$$\hat{SE}\left\{\overline{Y}_2^{\text{Adj}}\right\} = \sqrt{13.0988 + 26.2127 + 2(-13.1045)} = \sqrt{13.1025} = 3.620$$

$$\hat{SE}\left\{\overline{Y}_3^{\text{Adj}}\right\} = \sqrt{13.0988} = 3.619$$

$$\hat{SE}\left\{\overline{Y}_1^{\text{Adj}} - \overline{Y}_2^{\text{Adj}}\right\} = \sqrt{26.1919 + 26.2127 - 2(13.1007)} = \sqrt{26.2032} = 5.119$$

$$\hat{SE}\left\{\overline{Y}_1^{\text{Adj}} - \overline{Y}_3^{\text{Adj}}\right\} = \sqrt{26.1910} = 5.118$$

$$\hat{SE}\left\{\overline{Y}_2^{\text{Adj}} - \overline{Y}_3^{\text{Adj}}\right\} = \sqrt{26.2127} = 5.120$$

Note that for 56 degrees of freedom, $t_{.025} = 2.003$. 95% Confidence Intervals for the three treatment population means and differences among pairs of them are given below (without making adjustments for simultaneous intervals).

$$\text{Trt 1:}\ 134.873 \pm 2.003(3.619) \equiv 134.873 \pm 7.249 \equiv (127.624, 142.122)$$

$$\text{Trt 2:}\ 110.619 \pm 2.003(3.620) \equiv 110.619 \pm 7.251 \equiv (103.368, 1417.870)$$

$$\text{Trt 1:}\ 167.118 \pm 2.003(3.619) \equiv 167.118 \pm 7.249 \equiv (159.869, 174.367)$$

$$\text{Trt 1-2:}\ 24.254 \pm 2.003(5.119) \equiv 24.254 \pm 10.253 \equiv (14.001, 34.507)$$

$$\text{Trt 1-3:}\ -32.245 \pm 2.003(5.118) \equiv -32.245 \pm 10.251 \equiv (-42.496, -21.994)$$

$$\text{Trt 2-3:}\ -56.499 \pm 2.003(5.120) \equiv -56.499 \pm 10.255 \equiv (-66.754, -46.244)$$

All pairs of treatments are significantly different. The Gel formulation plus Dead Sea concentrate gives the best (lowest) roughness values, followed by Gel formulation, and then Placebo. Output from R making use of the **lsmeans** package, computing the adjusted means and simultaneous 95% Confidence Intervals making a Tukey adjustment is given below. These intervals are wider since they have simultaneous coverage with 95% confidence. To obtain them, replace $t_{.025,56}$ above with $q_{.05,3,56}/\sqrt{2} = 2.408$. Note that the uncentered X values were used and the least squares means are obtained at $\overline{X}_{..}$.

∇

```
> dsm.mod2b <- lm(post_y ~ factor(Trt) + pre_x)
> marginal = lsmeans(dsm.mod2b,
+                    ~ Trt:pre_x)
>
> cld(marginal,
+     alpha   = 0.05,
+     Letters = letters,     ### Use lower-case letters for .group
+     adjust  = "tukey")     ###  Tukey-adjusted comparisons
 Trt    pre_x   lsmean       SE df lower.CL upper.CL .group
   2 186.8928 110.6193 3.619743 56 101.7102 119.5284  a
   1 186.8928 134.8728 3.618745 56 125.9662 143.7795   b
   3 186.8928 167.1178 3.619227 56 158.2100 176.0257    c

Confidence level used: 0.95
Conf-level adjustment: sidak method for 3 estimates
P value adjustment: tukey method for comparing a family of 3 estimates
significance level used: alpha = 0.05
```

4.1.2 Interaction Model - Different Slopes

The possibility that the slope relating the covariate to the response can differ by treatment can be tested in the regression framework. As before, interaction terms are obtained by taking cross-product terms of the numeric covariate(s) with the dummy variables representing the treatments. In the case of a single centered covariate X and a single treatment factor with a levels, the model is given as follows.

$$Y_{ij} = \beta_0 + \beta_1 \left(X_{ij} - \overline{X}_{..}\right) + \gamma_1 W_{1ij} + \cdots + \gamma_{a-1} W_{a-1,ij} + \delta_1 \left(X_{ij} - \overline{X}_{..}\right) W_{1ij} + \cdots + \delta_{a-1} \left(X_{ij} - \overline{X}_{..}\right) W_{a-1,ij} + \epsilon_{ij}$$

The δ coefficients represent the difference in slope for the various treatments and that for the reference category.

$$i = 1: \quad E\left\{Y_{1j}\right\} = \beta_0 + \beta_1 \left(X_{ij} - \overline{X}_{..}\right) + \gamma_1(1) + \delta_1 \left(X_{ij} - \overline{X}_{..}\right)(1) = \left(\beta_0 + \gamma_1\right) + \left(\beta_1 + \delta_1\right)\left(X_{ij} - \overline{X}_{..}\right)$$

$$i = a: \quad E\left\{Y_{aj}\right\} = \beta_0 + \beta_1 \left(X_{ij} - \overline{X}_{..}\right)$$

The test for interaction is of the form $H_0 : \gamma_1 = \cdots \gamma_{a-1} = 0$, and involves comparing the full model containing the centered X values, $a-1$ dummy variables for the treatments, and $a-1$ cross-product terms to the reduced model with the cross-product terms removed.

Example 4.3: Skin Smoothing Study

The interaction model was fit for the skin smoothing study with the following results.

$$\text{Model 3: } \hat{Y} = 167.611 + 0.958 \left(X - \overline{X}_{..}\right) - 32.767 W_1 - 56.564 W_2 - 0.290 \left(X - \overline{X}_{..}\right) W_1 - 0.408 \left(X - \overline{X}_{..}\right) W_2$$

$$SSE_3 = 8719 \quad df_{E3} = 60 - 6 = 54 \quad R_3^2 = .9477$$

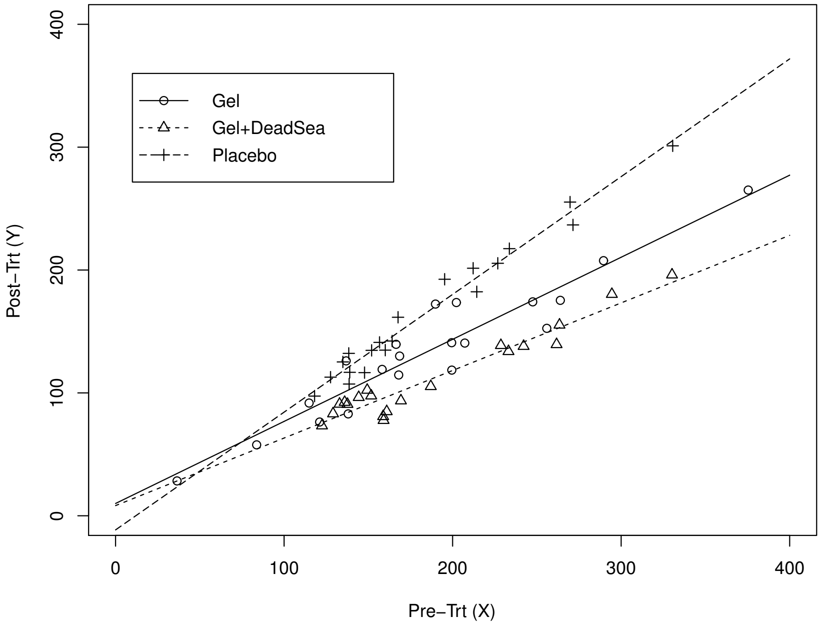

Figure 4.3: Plot of Interaction Model - Skin Softening Study

The test for Interaction effects ($H_0 : \delta_1 = \delta_2 = 0$) compares the full model (Model 3) with the reduced model (Model 2). The F-test is given below. The test is highly significant indicating that the slopes differ by treatment. A plot of the fitted equations based on the original X values is given in Figure 4.3.

$$H_0 : \delta_1 = \delta_2 = 0 \quad TS : F_{obs} = \frac{\left[\frac{14666-8719}{56-54}\right]}{\left[\frac{8719}{54}\right]} = \frac{2973.5}{161.5} = 18.4 \quad P\left(F_{2,54} \geq 18.4\right) < .0001$$

Since there is a significant interaction, the treatment effects differ depending on the pre-treatment score. The adjusted means could be compared at various levels of X. Based on Figure 4.3, it appears the higher the pre-treatment score (higher baseline roughness) the larger the differences in the treatments. All pairs of treatments are significantly different at $X = 140$, $X = 170$, and $X = 230$, which represent approximately the lower quartile, median, and upper quartile, respectively, of the baseline scores. Calculations making use of matrix algebra lead to the results in Table 4.2.

∇

X=140	1-2	1-3	2-3
Mean	40.32342	-73.4192	-113.743
SE	9.238539	9.811427	10.48674
LB	21.80126	-93.0899	-134.767
UB	58.84558	-53.7484	-92.7179
X=170	1-2	1-3	2-3
Mean	43.86492	-82.1303	-125.995
SE	10.88745	11.5685	12.43188
LB	22.03689	-105.324	-150.92
UB	65.69295	-58.9368	-101.071
X=230	1-2	1-3	2-3
Mean	50.94792	-99.5525	-150.5
SE	14.29506	15.18427	16.42047
LB	22.28806	-129.995	-183.421
UB	79.60778	-69.1098	-117.579

Table 4.2: Comparison of Adjusted means among treatments for Interaction model - Skin Softening Study

4.2 Extended Models

Making use of the regression framework, Analysis of Covariance (aka ANCOVA) models can be generalized to more than one covariate and more than one treatment factor, although it is much more difficult to visualize graphically. Suppose a model has 2 Factors: A with a levels and B with b levels and p covariates. An additive model with respect to the covariates could be written as follows with subscript i representing factor A level, j factor B level and k representing replicate number within treatment.

$$Y_{ijk} = \beta_0 + \sum_{m=1}^{p} \beta_m \left(X_{mijk} - \overline{X}_{m...} \right) + \sum_{i=1}^{a-1} W_{ijk}^{A} + \sum_{j=1}^{b-1} W_{ijk}^{B} + \sum_{i=1}^{a-1} \sum_{j=1}^{b-1} W_{ijk}^{A} W_{ijk}^{B} + \epsilon_{ijk}$$

Here, X_{mijk} represents the level of the m^{th} covariate for the experimental unit, with W_{ijk}^{A} and W_{ijk}^{B} being dummy variables for the levels of factors A and B.

Example 4.4: Factors Associated with Project Quality

An experiment was conducted to measure the effects of factors that effect project quality (Eubanks, Murphy, and Mumford, 2010, [11]). The researchers measured the quality of students' plans for setting up a college psychology club. There were three factors. Factor A was intuition, which was classified based on the student's scores on a series of decisions, with $a = 2$ levels (High/Low). Factor B was affect, which was assigned at random to students with $b = 2$ levels: positive and neutral; these were two different musical passages the students were exposed to. Factor C was the student's training method with $c = 4$ levels: associational model, mental model, fit appraisal, and control (no training). The model included $p = 3$ covariates: intelligence, openness, and class year (1=Freshman, etc). Intelligence and openness were measured using accepted scales from the literature. There were a total of $n_{..} = 320$ participants and the analysis included all main effects, 2-way, and the 3-way interactions among the 3 factors, along with the three covariates.

∇

4.3 R Programs for Chapter 4 Examples

4.4 Skin Smoothing Study

```
dsm <- read.csv("deadseaminerals.csv")
attach(dsm); names(dsm)

plot(post_y ~ pre_x, pch=Trt, xlab="Pre-Trt (X)", ylab="Post-Trt (Y)")
abline(v=mean(pre_x))
legend(60,275,c("Gel","Gel+DeadSea","Placebo"),pch=1:3)

xc <- pre_x - mean(pre_x)

## Covariate Only Model
dsm.mod1 <- lm(post_y ~ xc)
summary(dsm.mod1)
anova(dsm.mod1)

## Additive Model
dsm.mod2 <- lm(post_y ~ xc + gel + gelDS)
summary(dsm.mod2)
anova(dsm.mod2)
round(vcov(dsm.mod2),4)

anova(dsm.mod1,dsm.mod2)

dsm.mod2a <- lm(post_y ~ pre_x + gel + gelDS)

x.range <- seq(0,400,0.1)
yh.plac <- coef(dsm.mod2a)[1] + x.range*coef(dsm.mod2a)[2]
yh.gel <- coef(dsm.mod2a)[1] + x.range*coef(dsm.mod2a)[2] + coef(dsm.mod2a)[3]
yh.gelDS <- coef(dsm.mod2a)[1] + x.range*coef(dsm.mod2a)[2] + coef(dsm.mod2a)[4]

plot(post_y ~ pre_x, pch=Trt, xlab="Pre-Trt (X)", ylab="Post-Trt (Y)",
    xlim=c(0,400), ylim=c(0,400))
lines(x.range,yh.gel,lty=1)
lines(x.range,yh.gelDS,lty=2)
lines(x.range,yh.plac,lty=5)
legend(10,360,c("Gel","Gel+DeadSea","Placebo"),pch=c(1,2,3),lty=c(1,2,5))

dsm.mod2b <- lm(post_y ~ factor(Trt) + pre_x)

# install.packages("lsmeans")
library(lsmeans)

marginal = lsmeans(dsm.mod2b,
                   ~ Trt:pre_x)

cld(marginal,
    alpha   = 0.05,
    Letters = letters,     ### Use lower-case letters for .group
    adjust  = "tukey")     ###  Tukey-adjusted comparisons
```

```
## Interaction Model
dsm.mod3 <- lm(post_y ~ xc + gel + gelDS + I(xc*gel) + I(xc*gelDS))
summary(dsm.mod3)
anova(dsm.mod3)
round(vcov(dsm.mod3),4)
anova(dsm.mod2,dsm.mod3)

dsm.mod3a <- lm(post_y ~ pre_x + gel + gelDS + I(pre_x*gel) + I(pre_x*gelDS))

x.range <- seq(0,400,0.1)
yh.plac <- coef(dsm.mod3a)[1] + x.range*coef(dsm.mod3a)[2]
yh.gel <- coef(dsm.mod3a)[1] + x.range*coef(dsm.mod3a)[2] +
    coef(dsm.mod3a)[3] + x.range*coef(dsm.mod3a)[5]
yh.gelDS <- coef(dsm.mod3a)[1] + x.range*coef(dsm.mod3a)[2] +
    coef(dsm.mod3a)[4] + x.range*coef(dsm.mod3a)[6]

plot(post_y ~ pre_x, pch=Trt, xlab="Pre-Trt (X)", ylab="Post-Trt (Y)",
    xlim=c(0,400), ylim=c(0,400))
lines(x.range,yh.gel,lty=1)
lines(x.range,yh.gelDS,lty=2)
lines(x.range,yh.plac,lty=5)
legend(10,360,c("Gel","Gel+DeadSea","Placebo"),pch=c(1,2,3),lty=c(1,2,5))
```

Chapter 5

Generalized Linear Models

Previous chapters have been based on the data (or some transformation of the data) being normally distributed. In this chapter models are fit for data that follow other distributions. When the distributions are in the exponential family, these are referred to as **generalized linear models**. In this chapter, models for Binomial, Poisson, Negative Binomial, Gamma, and Beta random variables will be covered. The chapter begins with examples of data from each of the families without predictor variables. Then the models will be fit allowing for predictor variables.

5.1 Examples of Random Variables from Non-Normal Distributions

In this section, examples of random variables that are modeled by the Binomial, Poisson, Negative Binomial, Gamma, and Beta distributions are given. The first three distributions are discrete, with the random variables taking on only a finite or countably infinite set of distinct outcomes. The final two distributions are continuous and the random variables can take on any value along a continuum. In this section, parameter estimation and testing examples are given.

5.1.1 Binomial Distribution

A binomial experiment consists of a set of n independent Bernoulli trials, each of which can end in one of two outcomes: Success or Failure. The probability of success, labeled π, is assumed to be constant across trials. The random variable Y, is the number of observed Successes in the n trials and can only take on the values $y = 0, 1, \ldots, n$. In practice, the goal is to estimate and make inference regarding the unknown π based on sample data. The probability distribution of Y based on a sample of n trials is given below, along with its mean and variance.

$$P(Y = y) = p(y) = \left(\frac{n!}{y!(n-y)!}\right)\pi^y\,(1-\pi)^{n-y} \quad y = 0, 1, \ldots, n \quad 0 < \pi < 1 \qquad E\{Y\} = n\pi \quad V\{Y\} = n\pi\,(1-\pi)$$

The first term in the probability gives the number of arrangements of the n trials that can end in exactly y successes, the second term gives the probability of each of those distinct outcomes assuming independence.

An unbiased estimator for π is given below along with its mean, variance and standard error. In large samples, the estimator's sampling distribution is approximately normal.

$$\hat{\pi} = \frac{Y}{n} \qquad E\left\{\hat{\pi}\right\} = \pi \qquad V\left\{\hat{\pi}\right\} = \frac{\pi(1-\pi)}{n} \qquad SE\left\{\hat{\pi}\right\} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

Once the experiment has been conducted or a sample is taken, the observed number of successes, y, is used to make inference regarding π.

A large-sample $(1-\alpha)100\%$ Confidence Interval is obtained below.

$$\hat{\pi} \pm z_{\alpha/2}\hat{SE}\left\{\hat{\pi}\right\} \quad \equiv \quad \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}\left(1-\hat{\pi}\right)}{n}}$$

An exact Confidence Interval can be obtained as the set of π values which are "consistent" with having observed y successes in n trials and can be obtained with a computer package or spreadsheet.

A large-sample 2-sided test of whether $\pi = \pi_0$ can be conducted as follows, again an exact test can be obtained with a computer package or spreadsheet.

$$H_0 : \pi = \pi_0 \quad H_A : \pi \neq \pi_0 \quad TS : z_{obs} = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \qquad RR : |z_{obs}| \geq z_{\alpha/2} \quad P = 2P\left(Z \geq |z_{obs}|\right)$$

Suppose there are m independent binomial experiments, each with common success probability π. For the i^{th} experiment there is n_i trials and y_i observed successes. The joint probability mass function is the product of the individual probability mass functions.

$$p\left(y_1,\ldots,y_m\right) = \prod_{i=1}^{m}\left(\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\right)\pi_i^{y}\left(1-\pi\right)^{n_i-y_i} = \left(\prod_{i=1}^{m}\left(\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\right)\right)\pi^{y_.}\left(1-\pi\right)^{n_.-y_.}$$

$$n_. = \sum_{i=1}^{m} n_i \quad y_. = \sum_{i=1}^{m} y_i$$

Once the data $\left(y_1,n_1\right),\ldots,\left(y_m,n_m\right)$ have been observed, then $p\left(y_1,\ldots,y_m\right)$ is a function of the unknown parameter π, and is called the **likelihood function**, and often denoted by $L(\pi)$. The **Maximum Likelihood Estimator (MLE)** of π is the value $\hat{\pi}$ that maximizes the likelihood function. Calculus is used to obtain the estimator, however it is often easier to work with the log of the likelihood function. The logarithm is a monotonic function, so it is maximized at the same value of the variable to be maximized as the likelihood function.

In the case of the Binomial Distribution, the log-likelihood is given as follows.

$$l(\pi) = \log\left(L(\pi) = \log\left[\left(\prod_{i=1}^{m}\left(\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\right)\right)\pi^{y_.}\left(1-\pi\right)^{n_.-y_.}\right] =$$

$$\log\left(\prod_{i=1}^{m}\left(\frac{n_i!}{y_i!\left(n_i-y_i\right)!}\right)\right) + y_.\log(\pi) + \left(n_. - y_.\right)\log\left(1-\pi\right)$$

Take the derivative of $l(\pi)$ with respect to π, set the derivative to 0, and solve for $\hat{\pi}$. The derivative is given below along with the MLE, which is obtained by algebra once the derivative is set to 0.

$$\frac{dl(\pi)}{d\pi} = \frac{y_.}{\pi} + \frac{n_. - y_.}{1-\pi} \qquad \Rightarrow \qquad \hat{\pi} = \frac{y_.}{n_.}$$

Example 5.1: WNBA Free Throw Shooting - Maya Moore

Maya Moore is a Women's professional basketball star player. Interest in her in game free throw shooting skill could involve estimating her true underlying success probability or testing whether or not it is equal to some null value, say $\pi_0 = .80$. Note that basketball free throws are unguarded from a constant distance to the basket. During the 2014 season, Maya attempted $n = 181$ free throws, having successfully made $y = 160$ of them. For the purposes of this example, it is assumed that her 181 observed free throw attempts are a random sample of all free throws she could have taken at that point in her career.

A 95% Confidence Interval for her true underlying success rate π and a test of whether $\pi = 0.80$ are given below.

$$\hat{\pi} = \frac{160}{181} = 0.884 \qquad \hat{SE}\left\{\hat{\pi}\right\} = \sqrt{\frac{0.884\,(1-0.884)}{181}} = 0.0238$$

$$95\% \text{ CI for } \pi\text{: } 0.884 \pm 1.96(0.0238) \equiv 0.884 \pm 0.047 \equiv (0.837, 0.931)$$

$$H_0: \pi = 0.80 \qquad TS: z_{obs} = \frac{0.884 - 0.800}{\sqrt{\frac{0.80(1-0.80)}{181}}} = \frac{0.084}{0.0297} = 2.828 \qquad 2P(Z \geq 2.828) = .0047$$

Her true success rate appears to have been between .837 and .931, with strong evidence it exceeds .80.

R output is given below. It reports the estimated logit (log(odds)). The estimate of π is obtained by back-transforming. Let $\hat{\gamma}$ be the estimated logit, then $\hat{\pi}$ can be obtained as follows. Note that $\hat{\gamma}$ can take on any value on the real line, while $\hat{\pi}$ is constrained to $(0,1)$.

$$\hat{\gamma} = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) \qquad \Rightarrow \qquad \hat{\pi} = \frac{e^{\hat{\gamma}}}{1+e^{\hat{\gamma}}} = \frac{1}{1+e^{-\hat{\gamma}}}$$

$$\hat{\gamma} = 2.0307 \qquad \Rightarrow \qquad \hat{\pi} = \frac{e^{2.0307}}{1+e^{2.0307}} = \frac{7.6194}{1+7.6194} = 0.884$$

$$\nabla$$

```
> mm_ft <- c(rep(1,160),rep(0,21))
>
> table(mm_ft)
mm_ft
  0   1
 21 160
```

```
> mm.mod1 <- glm(mm_ft ~ 1,binomial("logit"))
> summary(mm.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0307     0.2321   8.749   <2e-16 ***
AIC: 131.93
```

5.1.2 Poisson Distribution

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable Y is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is is problematic in many applications where the variance can be larger than the mean. The probability distribution, mean and variance of Y are given below.

$$P(Y=y) = p(y) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0,1,\ldots \quad \mu > 0 \qquad E\{Y\} = \mu \qquad V\{Y\} = \mu$$

When a random sample of n Poisson random variables are observed with common μ, the joint probability mass function is obtained as the product of the individual probability mass functions. This is also the likelihood function for μ once $y_1, \ldots, y_n$ are observed. The following are the joint probability mass/likelihood function, log likelihood function, its derivative and the MLE for μ.

$$L(\mu) = p(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{e^{-\mu}(\mu)^{y_i}}{y_i!} = \frac{e^{-n\mu}(\mu)^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y_i!}$$

$$l(\mu) = \log(L) = -n\mu + \sum_{i=1}^{n} y_i \log(\mu) - \log\left(\prod_{i=1}^{n} y_i!\right) \qquad \frac{dl(\mu)}{d\mu} = -n + \frac{\sum_{i=1}^n y_i}{\mu} \quad \Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^n y_i}{n}$$

An unbiased estimator of μ is given below along with its mean, variance and standard error. The sampling distribution of the estimator is approximately normal in large samples.

$$\hat{\mu} = \frac{Y}{n} \qquad E\{\hat{\mu}\} = \mu \qquad V\{\hat{\mu}\} = \frac{\mu}{n} \qquad SE\{\hat{\mu}\} = \sqrt{\frac{\mu}{n}}$$

A large-sample $(1-\alpha)100\%$ Confidence Interval for μ is given below. A test of whether $\mu = \mu_0$ is also described.

$$\hat{\mu} \pm z_{\alpha/2} S\hat{E}\{\hat{\mu}\} \qquad \equiv \qquad \hat{\mu} \pm z_{\alpha/2}\sqrt{\frac{\hat{\mu}}{n}}$$

$$H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0 \quad TS: z_{obs} = \frac{\hat{\mu} - \mu_0}{\sqrt{\frac{\mu_0}{n}}} \qquad RR: |Z_{obs}| \geq z_{\alpha/2} \quad P = 2P(Z \geq |z_{obs}|)$$

Example 5.2: English Premier League Football Total Goals per Game - 2013/14 Season

Suppose interest is in estimating the population mean combined goals per game among the 2013/14 English Premier League (EPL) teams, based on the sample of games played in the season (380 total games). There are 20 teams, and each team plays each other team twice, once at Home, once Away. Assuming a Poisson model (which may not be reasonable, as different teams play in different games), estimate the underlying population mean μ. There were 380 games, with a total of 1063 goals, and sample mean and variance of 2.768 and 3.002, respectively. The number of goals and frequencies are given in Table 5.1.

Goals	0	1	2	3	4	5	6	7	8	9
Games	27	75	82	70	63	39	17	4	1	2

Table 5.1: Frequency Tabulation for EPL 2013/2014 Total Goals per Game

$$\hat{\mu} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{1063}{380} = 2.768 \qquad \hat{V}\left\{\hat{\mu}\right\} = \frac{\hat{\mu}}{n} = \frac{2.768}{380} = 0.007285$$

A 95% Confidence Interval for μ is obtained below.

$$\hat{\mu} \pm z_{.025}\sqrt{\hat{V}\left(\hat{\mu}\right)} \quad \equiv \quad 2.768 \pm 1.96\sqrt{0.007285} \quad \equiv \quad 2.768 \pm 0.167 \quad \equiv \quad (2.601, 2.935)$$

Table 5.2 gives the categories (goals), observed and expected counts, for the Poisson and Negative Binomial (next subsection) and the Chi-Square Goodness-of-fit tests for the two distributions. The goodness-of-fit test statistics are computed as follows, where O_i is the **Observed** count for the i^{th} category, E_i is the **Expected** count for the i^{th} category, and N is the total number of observations.

$$\hat{\pi}_i = P\left(Y = i\right) = \frac{e^{\hat{\mu}}\hat{\mu}^i}{i!} \quad i = 0 \ldots, 6 \qquad E_i = N \cdot \hat{\pi}_i \quad i = 0 \ldots, 6 \qquad O_7 = N - \sum_{i=0}^{6} O_i \quad E_7 = N - \sum_{i=0}^{6} E_i$$

$$X^2_{\text{GOF}} = \sum_{i=0}^{7} \frac{\left(O_i - E_i\right)^2}{E_i}$$

Goals	Observed	Expected(Poisson)	Expected(Neg Bin)
0	27	23.85	26.71
1	75	66.02	68.05
2	82	91.39	89.41
3	70	84.34	80.69
4	63	58.37	56.22
5	39	32.32	32.23
6	17	14.91	15.82
≥ 7	7	8.80	10.87

Table 5.2: Frequency Tabulation and Expected Counts for EPL 2013/2014 Total Goals per Game

The degrees of freedom for the Chi-Square Goodness-of-Fit test is one less than the number of categories minus the number of estimated parameters. In the case of the EPL Total goals per game with a Poisson distribution, there are 8 categories $(0, 1, \ldots, 7^+)$ and one estimated parameter (μ), for 8-1-1=6 degrees of freedom.

$$X^2_{\text{GOF-Poi}} = \frac{\left(27 - 23.85\right)^2}{23.85} + \cdots + \frac{\left(7 - 8.80\right)^2}{8.80} = 9.695 \qquad \chi^2\left(0.05, 6\right) = 12.592 \qquad P\left(\chi^2_6 \geq 9.695\right) = .1381$$

Fail to reject the hypothesis that total goals per game follows a Poisson distribution. A comparison of the Poisson model and the Negative Binomial model is given below. Output from an R program estimating the log of μ is given along with the Negative Binomial case below.

∇

5.1.3 Negative Binomial Distribution

The negative binomial distribution is used in two quite different contexts. The first is where a binomial type experiment is being conducted, except instead of having a fixed number of trials, the experiment is completed when the r^{th} success occurs. The random variable Y is the number of trials needed until the r^{th} success, and can take on any integer value greater than or equal to r. The probability distribution, its mean and variance areas follow.

$$P(Y=y)=p(y)=\frac{(y-1)!}{(r-1)!(y-r)!}\pi^r\left(1-\pi\right)^y-r \quad y=r,r+1,\ldots \qquad E\left\{Y\right\}=\frac{r}{\pi} \qquad V\left\{Y\right\}=\frac{r\left(1-\pi\right)}{\pi^2}.$$

A second use of the negative binomial distribution is as a model for count data. It arises from a mixture of Poisson models. In this setting it has 2 parameters and is more flexible than the Poisson (which has the variance equal to the mean), and can take on any non-negative integer value. In this form, the negative binomial distribution and its mean and variance can be written as follows (see e.g. Cameron and Trivedi, 2005, [8], and Agresti, 2002, [1]).

$$P(Y=y)=p(y)=\frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad \Gamma(w)=\int_0^\infty x^{w-1}e^{-x}dx=(w-1)\,\Gamma\left(w-1\right).$$

The mean and variance of this form of the Negative Binomial distribution are as follow.

$$E\left\{Y\right\}=\mu \qquad V\left\{Y\right\}=\mu\left(1+\alpha\mu\right)$$

If a random sample of n observations from a Negative Binomial distribution with parameters μ and α^{-1} have been observed, the likelihood function is of the following form based on $y_1,\ldots,y_n$.

$$L\left(\mu,\alpha^{-1}\right)=\prod_{i=1}^{n}\frac{\Gamma\left(\alpha^{-1}+y_i\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y_i+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y_i}$$

There are not closed-form estimators for μ and α^{-1}, and they are estimated using iterative methods in standard statistical software packages or in matrix form using a computer matrix language.

Example 5.3: English Premier League Football Total Goals per Game - 2013/14 Season

The Poisson model does appear to fit the data well, and the variance (3.002) is not much larger than the mean (2.768). As an example, a Negative Binomial model is fit to the data with Maximum Likelihood estimates for μ and α^{-1} being 2.768 and 32.00, respectively. Based on these values, the fitted probabilities lead to the expected counts for the 8 goal categories given in the last column of Table 5.2. As with the

Poisson model, a goodness-of-fit test can be conducted, the degrees of freedom for the chi-square statistic are now 8-1-2=5. The statistic is set up below.

$$X^2_{\text{GOF-NB}} = \frac{(27-26.71)^2}{26.71} + \cdots + \frac{(7-10.87)^2}{10.87} = 9.414 \qquad \chi^2(0.05,5) = 11.071 \qquad P\left(\chi^2_5 \geq 9.414\right) = .0937$$

Fail to reject the hypothesis that total goals per game follows a Negative Binomial distribution. A comparison of the Poisson model and the Negative Binomial model leads to a Likelihood Ratio test statistic of 1.256 on 1 degree of freedom (the difference in the numbers of parameters). The P-value is .262, leading to choose the simpler Poisson model over the more complex Negative Binomial. R output is given below for the Poisson and Negative Binomial models. Note that the log of μ is being estimated. The output labels α^{-1} as "Theta." Making use of the estimates $\log(\hat{\mu})$ and $\hat{\alpha^{-1}}$, the mean and variance are estimated as follow.

$$\hat{\mu} = e^{1.01828} = 2.768 \qquad \hat{V}\{Y\} = \hat{\mu}\left(1 - \frac{\hat{\mu}}{\hat{\alpha^{-1}}}\right) = 2.768\left(1 + \frac{2.768}{32.0}\right) = 3.007$$

∇

```
> library(MASS)
> mod1 <- glm(totEng~1,family="poisson")
> summary(mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.01828    0.03083   33.03   <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)
AIC: 1466.2

> mod2 <- glm.nb(totEng~1)
> summary(mod2)
Call:
glm.nb(formula = totEng ~ 1, init.theta = 32.00072139, link = log)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.01828    0.03214   31.68   <2e-16 ***

(Dispersion parameter for Negative Binomial(32.0007) family taken to be 1)
AIC: 1467

              Theta:  32.0
          Std. Err.:  30.4
 2 x log-likelihood:  -1462.983
>
> X2.poi.nb <- -2*(logLik(mod1)-logLik(mod2))
> X2.05 <- qchisq(.95,1)
> X2.pval <- 1-pchisq(X2.poi.nb,1)
>
> print(round(cbind(X2.poi.nb,X2.05,X2.pval),3))
     X2.poi.nb X2.05 X2.pval
[1,]     1.256 3.841   0.262
```

5.1.4 Gamma Distribution

The gamma family of distributions are used to model non-negative random variables that are often right-skewed. There are two widely used parameterizations. The first given here is in terms of *shape* and *scale* parameters:

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \alpha\beta^2$$

Here, $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}dy$ and is built-in to virtually all statistical packages and spreadsheets. It also has two simple properties.

$$\alpha > 1: \quad \Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1) \qquad\qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}.$$

Thus, if α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. The second version given here is in terms of *shape* and *rate* parameters.

$$f(y) = \frac{\theta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-y\theta} \quad y \geq 0, \alpha > 0, \theta > 0 \qquad E\{Y\} = \mu_Y = \frac{\alpha}{\theta} \qquad V\{Y\} = \frac{\alpha}{\theta^2}$$

Once a sample of size n has been obtained, the likelihood function for the second parameterization is the following.

$$L(\alpha, \theta) = \prod_{i=1}^{n} \frac{\theta^{\alpha}}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-y_i\theta} = \left[\frac{\theta^{\alpha}}{\Gamma(\alpha)}\right]^n \left(\prod_{i=1}^{n} y_i^{\alpha-1}\right) e^{-\sum_{i=1}^{n} y_i\theta}$$

Note that different software packages use different parameterizations in generating samples and giving tail-areas and critical values. For instance, EXCEL uses the first parameterization and *R* uses the second. As in the case of the Negative Binomial distribution, the parameters must be estimated iteratively with no closed-form solutions.

Example 5.4: Running Speeds Among Females at a Marathon

The running speeds (miles per hour) among $n = 1045$ females who completed the Rock and Roll Marathon in Washington are all positive and are seen to be skewed right. The histogram and corresponding gamma distribution are shown in Figure 5.1. Here times are treated as a random sample of times from a larger conceptual population. Maximum Likelihood estimates of α and θ for the second parameterization are 49.381 and 8.456, respectively. The gamma density with these parameters (multiplied by 1045) is included in Figure 5.1. The mean and variance of are given speeds are given below.

$$E\{Y\} = \frac{49.381}{8.456} = 5.840 \qquad\qquad V\{Y\} = \frac{49.381}{(8.456)^2} = 0.691$$

R Output is given below. Note that it is estimating the log of the mean with the Intercept parameter, and the reciprocal of α with the dispersion parameter.

$$\hat{\alpha} = \frac{1}{0.02025082} = 49.381 \qquad \frac{\hat{\alpha}}{\hat{\theta}} = e^{1.764703} = 5.840 \quad \Rightarrow \quad \hat{\theta} = [0.02025082(5.840)]^{-1} = 8.456$$

$$\nabla$$

```
> rrf.mod1 <- glm(f.mph ~ 1, family=Gamma(link="log"))
> summary(rrf.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.764703   0.004402   400.9   <2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.02025082)
AIC: 2530
```

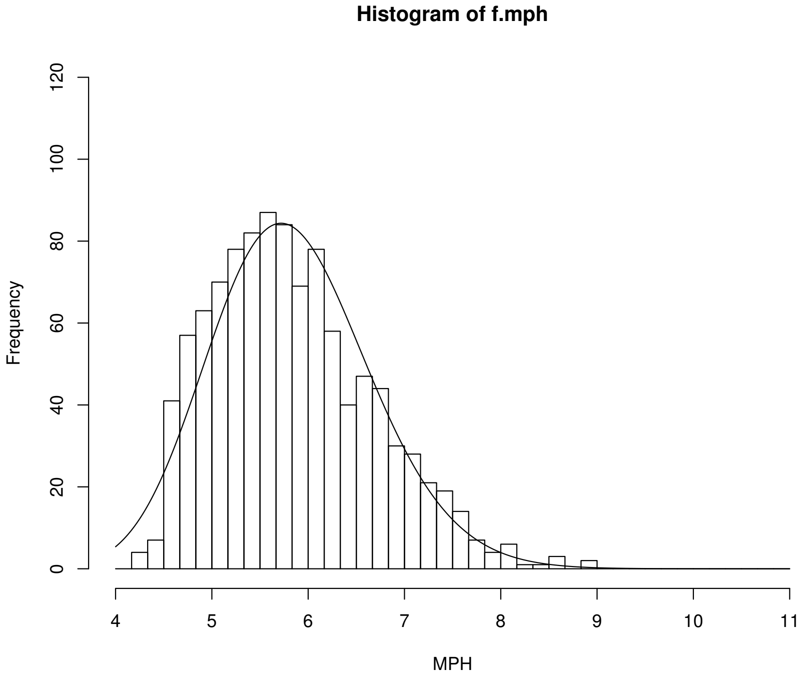

Figure 5.1: Marathon speeds for female runners at 2015 Rock & Roll Marathon in Washington, DC

5.1.5 Beta Distribution

The Beta distribution can be used to model data that are proportions (or percentages divided by 100). The traditional model for the Beta distribution is as follows.

$$f\left(y;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}y^{\alpha-1}\left(1-y\right)^{\beta-1}\qquad 0<y<1;\quad \alpha>0,\beta>0$$

$$E\left\{Y\right\}=\frac{\alpha}{\alpha+\beta}\qquad V\left\{Y\right\}=\frac{\alpha\beta}{\left(\alpha+\beta\right)^{2}\left(\alpha+\beta+1\right)}$$

An alternative formulation of the distribution involves setting a re-parameterizing as follows. It is the formulation used in the **betareg** package in R.

$$\mu=\frac{\alpha}{\alpha+\beta}\qquad \phi=\alpha+\beta\quad\Rightarrow\quad \alpha=\mu\phi\qquad \beta=\left(1-\mu\right)\phi$$

$$f\left(y;\alpha,\beta\right)=\frac{\Gamma\left(\phi\right)}{\Gamma\left(\mu\phi\right)\Gamma\left(\left(1-\mu\right)\phi\right)}y^{\mu\phi-1}\left(1-y\right)^{\left(1-\mu\right)\phi-1}$$

Once a sample of size n has been obtained, the likelihood function for the second parameterization is the following.

$$L\left(\alpha,\beta\right)=\prod_{i=1}^{n}\frac{\Gamma\left(\phi\right)}{\Gamma\left(\mu\phi\right)\Gamma\left(\left(1-\mu\right)\phi\right)}y_i^{\mu\phi-1}\left(1-y_i\right)^{\left(1-\mu\right)\phi-1}=\left[\frac{\Gamma\left(\phi\right)}{\Gamma\left(\mu\phi\right)\Gamma\left(\left(1-\mu\right)\phi\right)}\right]^{n}\prod_{i=1}^{n}y_i^{\mu\phi-1}\left(1-y_i\right)^{\left(1-\mu\right)\phi-1}$$

In this formulation, the logit (log odds) of the mean (of the proportions) is estimated as the Intercept and ϕ is estimated as the precision parameter.

$$\mu = \frac{\alpha}{\alpha+\beta} \quad \gamma = \log\left(\frac{\mu}{1-\mu}\right) \quad \mu = \frac{e^{\gamma}}{1+e^{\gamma}} \quad \phi = \alpha+\beta$$

Example 5.5: NBA Team/Game Free Throw Proportions

For the National Basketball Association 2016/17 regular season, data is obtained for each team's free throw proportion made by game. There are 30 teams, each playing 82 games, thus there are 2460 team/games. Algorithms can have problems when some proportions are 0 or 1, so in this example, "Wilson-Agresti-Coull" method is used for the game proportions. This involves the addition of 2 successes and 2 failures to the observed number of successes and failures. Thus, if a team made 20 free throws out of 24 attempts, the proportion for that game would be $(20+2)/(24+4) = 22/28 = .7857$. Maximum Likelihood estimates for γ and ϕ are 0.983 and 27.439, respectively. This leads to the following estimates.

$$\hat{\mu} = \frac{e^{0.983}}{1+e^{0.983}} = 0.728 \qquad \hat{\alpha} = 27.439(0.728) = 19.967 \qquad \hat{\beta} = 27.439(1-0.728) = 7.471$$

These lead to the following fitted means and variances directly from $\hat{\alpha}$ and $\hat{\beta}$.

$$\hat{\mu} = \frac{19.967}{19.967+7.471} = \frac{19.967}{27.438} = 0.728 \qquad \hat{V}\{Y\} = \frac{19.967(7.471)}{(27.438)^2(27.438+1)} = \frac{149.173}{21409.373} = 0.00697$$

A histogram of the team/game proportions and the beta density (multiplied by 2460) are given in Figure 5.2. R output is given below.

$$\nabla$$

```
> library(betareg)
> FT.mod1 <- betareg(Ftprop.r ~ 1)
> summary(FT.mod1)
Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.98300    0.00849   115.8   <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  27.4386     0.7714   35.57   <2e-16 ***

Type of estimator: ML (maximum likelihood)
Log-likelihood:  2654 on 2 Df
```

5.2 Regression Models

In the previous section, measurements were considered to be from the same population, and parameters of the particular distributions were estimated by maximum likelihood. In this section models are fit that allow for units to have independent variables (covariates) associated with them, and allow these to be related to

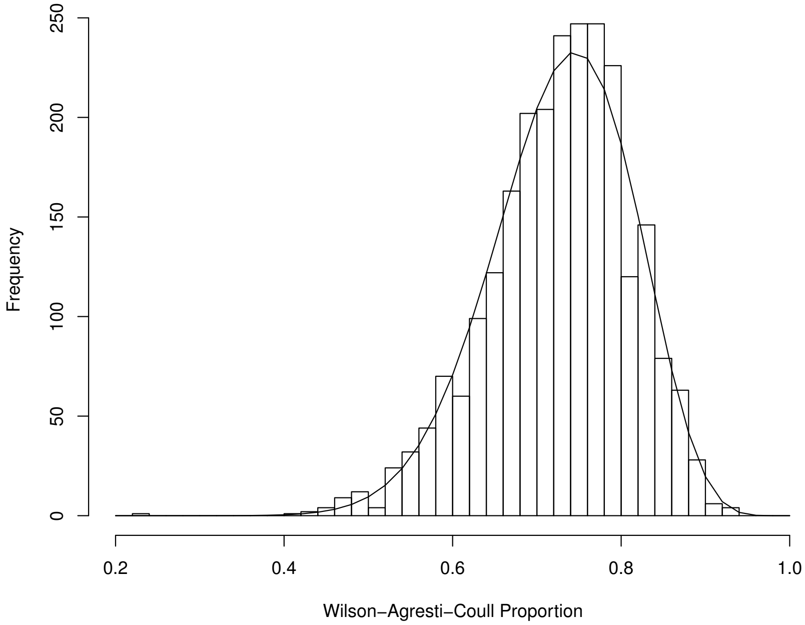

Figure 5.2: Team/Game free throw proportions and scaled Beta Density

the response variable. For instance, in the Maya Moore binomial free throw example, her probability of success may depend on whether the game was Home or Away, and/or the number of days between games. The soccer goal example could be extended to multiple premier leagues and tests could be conducted of whether the various leagues have the same true mean.

The same types of tests can be conducted as was done for the Normal linear regression model, however the F-tests and t-tests are replaced with the χ^2-tests and z-tests. These are referred to as **Likelihood Ratio** and **Wald** tests. Likelihood Ratio tests compare Complete and Reduced models in terms of the difference in their evaluated log likelihoods under the two (or more) models. These are similar to F-tests for the Normal models. Wald tests can be used for individual partial regression coefficients by comparing the estimate with the null parameter value (typically 0) in standard error units, using z-tests of χ^2-tests on the squared z value. Wald tests can be extended to multiple parameter tests with matrix form which is not covered here.

In this section, regression and ANOVA type models will be covered for the five probability distributions from Section 5.1. These will make use of the **glm**, **glm.nb**, and **betareg** functions in R that were used in Section 5.1 to estimate the parameters under the models with no predictor variables.

Each model type has a **link** function, a function of the mean response that is linear in the predictor variables. Note that categorical predictors, polynomials, and interactions can be included just as in the Normal model.

The following notation is used in testing hypotheses and comparing models. The **Null Model** is the model with no predictor variables (the models fit in Section 5.1, for instance). The **Saturated Model** is the model with each unit's predicted value being equal to its observed value (examples will be given for the individual distributions below). The maximized log-likelihood is obtained for the null model (l_0) and for the saturated model (l_s), where $l_s \geq l_0$ by definition. The quantity $D_0 = -2\left(l_0 - l_s\right)$ is referred to as the **Null Deviance**. For a particular model M, with say p predictors, the maximized log-likelihood is l_M, with $l_0 \leq l_M \leq l_s$, the quantity $D_M = -2\left(l_M - l_s\right)$ is the **Residual Deviance** for model M. These are used to test between different nested models, as the F-tests were used in Normal regression.

To test whether all of the regression coefficients of the predictors in model M are 0, the likelihood ratio test is conducted as follows.

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \qquad TS : X_{LR}^2 = -2\left(l_0 - l_M\right) = D_0 - D_M \qquad RR : X_{LR}^2 \geq \chi_{\alpha,p}^2 \qquad P = P\left(\chi_p^2 \geq X_{LR}^2\right)$$

Similarly, to test whether a group of $p - g$ are note associated with the response after controlling for the g remaining predictors, let F refer to the Full model with all p predictors and R be the Reduced model with the subset of $g < p$ predictors. Then labeling l_F and l_R as the log-likelihoods for the models the test is conducted as foolows.

$$H_0 : \beta_{g+1} = \cdots = \beta_p = 0 \quad TS : X_{LR}^2 = -2\left(l_R - l_F\right) = D_R - D_F \quad RR : X_{LR}^2 \geq \chi_{\alpha,p-g}^2 \quad P = P\left(\chi_{p-g}^2 \geq X_{LR}^2\right)$$

A test for a single partial regression coefficient, say β_j can be conducted as a special case of the previous Complete/Reduced test or as a Wald test, which is given below, the first is a z-test, the second is an equivalent Chi-square test. R and Stata print the z version, while SAS and SPSS print the Chi-Square version.

$$H_0 : \beta_j = 0 \qquad TS : z_W = \frac{\hat{\beta}_j}{\hat{SE}\left\{\hat{\beta}_j\right\}} \qquad RR : |z_W| \geq z_{\alpha/2} \qquad P = 2P\left(Z \geq |z_W|\right)$$

$$H_0 : \beta_j = 0 \qquad TS : X_W^2 = \left(\frac{\hat{\beta}_j}{\hat{SE}\left\{\hat{\beta}_j\right\}}\right) \qquad RR : X_W^2 \geq \chi_{\alpha,1}^2 \qquad P = P\left(\chi_1^2 \geq X_W^2\right)$$

There are two commonly used types of residuals: **Pearson** and **Deviance**. These will be defined in the special cases in the following subsections, as they depend on the probability distribution being modeled. These are used to test Goodness-of-Fit for the various models being fit and compared.

There are various Pseudo-R^2 statistics used to describe model fit. These include McFadden's, Cox & Snell's, Nagelkerke's, and Efron's described below, where l_M is the log-likelihood and L_M is the likelihood for model M.

$$\text{McFadden: } R_{McF}^2 = 1 - \frac{l_M}{l_0} \qquad AdjustedVersion : 1 - \frac{l_M - \#\text{parameters}}{l_0}$$

$$\text{Cox \& Snell: } R_{CS}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n} \qquad L_* = e^{l_*}$$

$$\text{Nagelkerke: } R_N^2 = \frac{1 - \left(\frac{L_0}{L_M}\right)^{2/n}}{1 - \left(L_0\right)^{2/n}} \qquad \text{Can be as large as 1}$$

$$\text{Efron: } R_E^2 = 1 - \frac{\sum_{i=1}^n \left(y_i - \hat{y_i}\right)^2}{\sum_{i=1}^n \left(y_i - \overline{y}\right)^2} \qquad \text{Similar to definition for OLS Regression}$$

5.2.1 Logistic Regression for Binomial/Bernoulli Outcomes

In the case where responses are outcomes of a Binomial experiment, or simply a set of individual Bernoulli trials, one commonly fit model is the **Logistic Regression** model. In this model, the link function of the mean that is linear in the predictors is the logit (see Section 5.1.1), where $\mu = \pi$.

$$g(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad \Rightarrow \qquad \pi = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

Note that $g(\pi)$ can take on any values along the real line, π is bounded to be between 0 and 1. Further, if $\beta_1 = \cdots = \beta_p = 0$, then π is constant (not associated with any of the predictors). Once ML estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are obtained, they replace the unknown parameters for the fitted probabilities.

In the linear regression model, the slopes β_j represent the change in the mean of Y when X_j is increased by 1 unit, with all other predictors held constant. In the logistic regression model, this interpretation applies to the logit (log odds), which is not particularly intuitive. Consider the odds of Success at X for the case with $p = 1$ predictor, then at $X + 1$.

$$odds(X) = \frac{\pi(X)}{1-\pi(X)} = \frac{\left[\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right]}{\left[1-\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right]} = \frac{\left[\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}\right]}{\left[\frac{1}{1+e^{\beta_0+\beta_1 X}}\right]} = e^{\beta_0+\beta_1 X} = e^{\beta_0}e^{\beta_1 X}$$

$$odds(X+1) = e^{\beta_0+\beta_1(X+1)} = e^{\beta_0}e^{\beta_1(X+1)} = e^{\beta_0}e^{\beta_1 X}e^{\beta 1}$$

Consider the Odds Ratio: $OR = odds(X+1)/odds(X) = e^{\beta_1}$. Thus, regardless of the level of X, e^{β_1} represents the multiplicative change in the odds of a Success as X increases 1 unit. When there are $p > 1$ predictors, e^{β_j} represents that change when X_j increases 1 unit, while all other predictors are held constant. To obtain a Confidence Interval for the Odds Ratio for X_j, first obtain a Confidence Interval for β_j, then exponentiate the end points.

$$(1-\alpha)100\% \text{ CI for } \beta_j\text{: } \hat{\beta}_j \pm z_{\alpha/2}\hat{SE}\left\{\hat{\beta}_j\right\} \equiv \left(\beta_j^L, \beta_j^U\right) \qquad \Rightarrow \qquad (1-\alpha)100\% \text{ CI for } OR_j = e^{\beta_j}\text{: } \left(e^{\beta_j^L}, e^{\beta_j^U}\right)$$

If the regression coefficient is significantly different from 0, the Odds Ratio will be significantly different from 1.

For grouped Binomial data, with m groups and observed numbers of successes and trials $(y_1, n_1), \ldots, (y_1, n_1)$, the fitted values and residuals are obtained as follow, where $X_{i1}, \ldots, X_{ip}$ are the levels of the p predictor variables for the i^{th} group.

$$\text{Fitted: } \hat{y}_i = n_i\hat{\pi}_i = n_i\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\cdots+\hat{\beta}_p X_{ip}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\cdots+\hat{\beta}_p X_{ip}}} \qquad \text{Residual: } e_i = y_i - \hat{y}_i$$

The Pearson and Deviance residuals and Chi-square statistics for the Logistic Regression model are given below, in the case of Deviance residuals, the ln represents the natural log and $0ln(0) \equiv 0$ and sign$(*)$ is $+$ if $*$ is positive and $-$ if negative.

$$\text{Pearson Residual: } e_i^P = \frac{y_i - \hat{y}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

$$\text{Pearson Chi-square: } X_P^2 = \sum_{i=1}^{m} \left(e_i^P\right)^2 = \sum_{i=1}^{m} \frac{\left(y_i - \hat{y}_i\right)^2}{n_i \hat{\pi}_i \left(1 - \hat{\pi}_i\right)} \qquad df = m - (p+1)$$

$$\text{Deviance Residual: } e_i^D = \operatorname{sign}\left(y_i - \hat{y}_i\right) \left[2\left(y_i \ln\left(\frac{y_i}{\hat{y}_i}\right) + \left(n_i - y_i\right) \ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right)\right]$$

$$\text{Deviance Chi-square: } X_D^2 = \sum_{i=1}^{m} \left(e_i^D\right)^2 \qquad df = m - (p+1)$$

When the data are based on individual (ungrouped) observations, Chi-square approximations for the Pearson and Deviance Chi-square statistics for Goodness-of-Fit is questionable (they are based on large numbers of cases within groups). One commonly used test in this situation is the **Hosmer-Lemeshow Test**. The approximate test is conducted as follows.

1. Group individual cases into g groups based on their predicted probabilities (typically, $g = 10$)
2. For each group, obtain n_i (total number of cases), o_i (total number of successes), $\bar{\hat{\pi}}_i$ (average predicted probability)

The approximate Chi-square statistic has $g - 2$ degrees of freedom.

$$TS: X_{HL}^2 = \sum_{i=1}^{g} \frac{\left(o_i - n_i \bar{\hat{\pi}}_i\right)^2}{n_i \bar{\hat{\pi}}_i \left(1 - \bar{\hat{\pi}}_i\right)^2} \qquad RR: X_{HL}^2 \geq \chi_{\alpha, g-2}^2 \qquad P = P\left(\chi_{g-2}^2 \geq X_{HL}^2\right)$$

Example 5.6: Motorcycles and Erectile Dysfunction

A Japanese study investigated the association between motorcycle riding, age, and presence/absence of erectile dysfunction (ED) in men (Ochiai, et al, 2006, [26]). Their study involved 234 motorcycle riders and 752 healthy controls. Subjects were classified by age (20-29, 30-39, 40-49, and 50-59 years). The ED was classified in 5 categories (No ED, mild, mild-to-moderate, moderate, and severe). For the purposes of this example, all subjects are given the age of the center of their range (25 for 20-29, and so on). Also, all ED classifications (mild through severe) are combined into "ED present." The model being fit will use motorcycle riding and age to estimate the probability of incidence of ED. Frequency counts are given in Table 5.3. Note that this is purely looking for association and not causation. Men might choose to ride motorcycles after developing ED.

Note that for this study, $m = 4(2) = 8$, representing the combinations of age and motorcycle riding. Consider the following 4 models, where π is the probability of having from ED, with A=age, and M=1 if motorcycle rider, 0 if control. Model 1 is the null model with no predictors. Model 2 contains Age as a predictor. Model 3 is an additive model with Age and Motorcycle Riding. Model 4 includes an interaction term.The fitted equations and log-likelihoods are given below.

$$\text{Null Model (1): } \log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = -0.2694 \qquad l_1 = -69.909 \quad df_1 = 8 - 1 = 7$$

$$\text{Age Model (2): } \log\left(\frac{\hat{\pi}}{1 - \hat{\pi}}\right) = -0.9155 + 0.0165A \qquad l_2 = -66.943 \quad df_2 = 8 - 2 = 6$$

Age/MR Additive Model (3): $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.6027{+}0.0250A{+}1.4688M \qquad l_3 = -23.001 \quad df_3 = 8{-}3 = 5$

$$\text{Age/MR Interaction Model (4): } \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.2575 + 0.0164A + 0.0131M + 0.0395AM$$

$$l_4 = -20.500 \qquad df_4 = 8 - 4 = 4$$

Consider testing whether any of the main effects of age and motorcycle riding or their interaction are significant. This involves comparing models 1 and 4.

$$H_0 : \beta_A = \beta_M = \beta_{AM} = 0 \quad TS : X^2_{LR} = -2((-69.909){-}(-20.500)) = 98.82 \quad RR : X^2_{LR} \geq 7.815 \quad P < .0001$$

Clearly, at least one of these predictors is associated with the probability of ED. Now consider whether motorcycle riding and/or the interaction are significant after controlling for Age. This test compares models 2 and 4.

$$H_0 : \beta_M = \beta_{AM} = 0 \qquad TS : X^2_{LR} = -2((-66.943){-}(-20.500)) = 92.89 \qquad RR : X^2_{LR} \geq 5.991 \qquad P < .0001$$

Again, at least one of the regression coefficients is significantly different from 0. Finally, consider a test of whether there is a significant interaction, after controlling for main effects. This compares models 3 and 4.

$$H_0 : \beta_{AM} = 0 \qquad TS : X^2_{LR} = -2((-23.001) - (-20.500)) = 5.002 \qquad RR : X^2_{LR} \geq 3.841 \qquad P = .0253$$

The interaction between Age and Motorcycle riding is significant. The Wald statistic (see R output below) is $z_W = 2.189, X^2_W = (2.189)^2 = 4.792, P = .0286$. The model 3 fitted probabilities, predicted counts, Pearson and Deviance residuals and Goodness-of-Fit tests are given in Table 5.4. The model appears to fit the data very well (both P-values are above 0.5). A plot of the fitted curves is given in Figure 5.3.

Several Pseudo-R^2 measures for model 3 are given below with $l_0 = -69.909$, $l_3 = -20.500$.

$$R^2_{McF} = 1{-}\frac{-20.500}{-69.909} = 0.7068 \qquad R^2_{CS} = 1{-}\left(\frac{e^{-69.909}}{e^{-20.500}}\right)^{2/8} = 1.0000 \qquad R^2_N = \frac{1-\left(\frac{e^{-69.909}}{e^{-20.500}}\right)^{2/8}}{1-\left(e^{-69.909}\right)^{2/8}} = 1.0000$$

R output is given below.

∇

	Controls			Motorcycle Riders		
Age	ED	No ED	Total	ED	No ED	Total
20-29	35	77	112	35	25	60
30-39	101	205	306	55	32	87
40-49	77	131	208	44	14	58
50-59	53	73	126	27	2	29
Total	266	486	752	161	73	234

Table 5.3: Frequency Tabulation for Motorcycle Riding/Erectile Dysfunction Study

Age	MR	n_i	y_i	$\hat{\pi}_i$	$\hat{y}_i$	e_i^P	e_i^D
25	0	112	35	0.2963	33.19	0.3756	0.1396
35	0	306	101	0.3316	101.47	-0.0567	-0.0032
45	0	208	77	0.3689	76.73	0.0391	0.0015
55	0	126	53	0.4078	51.38	0.2928	0.0855
25	1	60	35	0.5338	32.03	0.7684	0.5941
35	1	87	55	0.6670	58.03	-0.6888	-0.4668
45	1	58	44	0.7779	45.12	-0.3537	-0.1226
55	1	29	27	0.8597	24.93	1.1064	1.4646
					GOF Chi-Sq	2.6456	2.7578
					df	4	4
					$X^2_{.05}$	9.4877	9.4877
					P-value	0.6188	0.5991

Table 5.4: Fitted Values, Pearson and Deviance residuals and Goodness-of-Fit statistics - Motorcycle Riding/Erectile Dysfunction Study

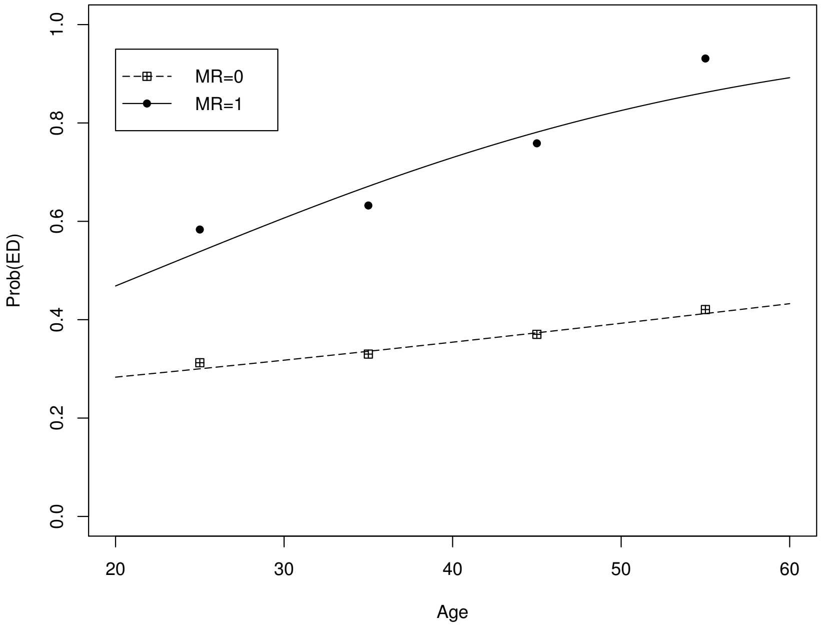

Figure 5.3: Observed Proportions and fitted probabilities - Motorcycle Riding/Erectile Dysfunction Study

```
> print(cbind(age,mr,y))
     age mr ed.1 ed.0
[1,]  25  0   35   77
[2,]  35  0  101  205
[3,]  45  0   77  131
[4,]  55  0   53   73
[5,]  25  1   35   25
[6,]  35  1   55   32
[7,]  45  1   44   14
[8,]  55  1   27    2
>
> mod0 <- glm(y~1, family=binomial("logit"))
> summary(mod0)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.26937    0.06427  -4.191 2.78e-05 ***

    Null deviance: 101.62  on 7  degrees of freedom
Residual deviance: 101.62  on 7  degrees of freedom
AIC: 141.82
> logLik(mod0)
’log Lik.’ -69.90915 (df=1)

> mod1 <- glm(y~age, family=binomial("logit"))
> summary(mod1)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.915497   0.274247  -3.338 0.000843 ***
age          0.016486   0.006785   2.430 0.015114 *

    Null deviance: 101.618  on 7  degrees of freedom
Residual deviance:  95.684  on 6  degrees of freedom
AIC: 137.89
> logLik(mod1)
’log Lik.’ -66.94255 (df=2)
>
> anova(mod0,mod1,test="Chisq")
Analysis of Deviance Table
Model 1: y ~ 1
Model 2: y ~ age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         7    101.618
2         6     95.684  1   5.9332  0.01486 *


> mod2 <- glm(y~age+mr, family=binomial("logit"))
> summary(mod2)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.602726   0.299229  -5.356  8.5e-08 ***
age          0.025036   0.007186   3.484 0.000494 ***
mr           1.468761   0.163629   8.976  < 2e-16 ***

    Null deviance: 101.6176  on 7  degrees of freedom
Residual deviance:   7.8012  on 5  degrees of freedom
AIC: 52.002
> logLik(mod2)
’log Lik.’ -23.00093 (df=3)

> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table
Model 1: y ~ age
Model 2: y ~ age + mr
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
1         6     95.684
2         5      7.801  1   87.883 < 2.2e-16 ***

> mod3 <- glm(y~age+mr+age:mr, family=binomial("logit"))
> summary(mod3)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.257466   0.334854  -3.755 0.000173 ***
age          0.016436   0.008141   2.019 0.043509 *
mr           0.013092   0.676970   0.019 0.984570
age:mr       0.039497   0.018047   2.189 0.028628 *

    Null deviance: 101.6176  on 7  degrees of freedom
Residual deviance:   2.7998  on 4  degrees of freedom
AIC: 49.001
> logLik(mod3)
'log Lik.' -20.50026 (df=4)

> anova(mod2,mod3,test="Chisq")
Analysis of Deviance Table
Model 1: y ~ age + mr
Model 2: y ~ age + mr + age:mr
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         5     7.8012
2         4     2.7998  1   5.0014  0.02533 *
```

Example 5.7: NFL Field Goal Attempts - 2008 Regular Season

Data were obtained on all $n = 1039$ field goal attempts by kickers during the 2008 National Football League (NFL) regular season. Two variables were used as predictors: the distance of the kick (yards) and a home game indicator (1 if Home game, 0 if Away). While kickers surely have different skill levels, they are all professionals, so kicker effects are ignored here. Unlike the Mortorcycle/ED data, these data are treated as being ungrouped. Three models were considered: Null Model (1) with no predictors, a model with only distance (2), and a model with Distance and Home (3).

$$\hat{\pi}_1 = \frac{e^{1.8679}}{1 + e^{1.8679}} \quad l_1 = -408.8614 \quad df_1 = 1039 - 1 = 1038$$

$$\hat{\pi}_2 = \frac{e^{6.7627 - 0.1208D}}{6.7627 - 0.1208D} \quad l_2 = -343.4635 \quad df_2 = 1039 - 2 = 1037$$

$$\hat{\pi}_3 = \frac{e^{6.8987 - 0.1208D - 0.2596H}}{6.8987 - 0.1208D - 0.2596H} \quad l_3 = -342.5926 \quad df_3 = 1039 - 3 = 1036 \qquad \hat{SE}\left\{\hat{\beta}_H\right\} = 0.1972$$

The likelihood ratio tests for Distance (alone) and Home given Distance are given below, as well as a 95% Confidence Interval for the Home Field Odds Ratio (controlling for Distance).

$$H_0^D : \beta_D = 0 \qquad TS : X_{LR}^2 = -2((-408.8614) - (-343.4635)) = 130.80 \qquad RR : X_{LR}^2 \geq 3.841 \quad P < .0001$$

$$H_0^H : \beta_H = 0 \qquad TS : X_{LR}^2 = -2((-343.4635) - (-342.5926)) = 1.7418 \qquad RR : X_{LR}^2 \geq 3.841 \quad P = .1869$$

$$H_0^H : \beta_H = 0 \qquad TS : X_W^2 = \left(\frac{-0.2569}{0.1972}\right)^2 = 1.6971 \qquad RR : X_W^2 \geq 3.841 \quad P = .1927$$

$$(1-\alpha)100\% \text{ CI for } \beta_H\text{: } -0.2569 \pm 1.96(0.1972) \equiv -0.2569 \pm 0.3865 \equiv (-0.6434, .1296)$$

$$(1-\alpha)100\% \text{ CI for } OR_H\text{: } \left(e^{-0.6434}, e^{0.1296}\right) \equiv (0.526, 1, 138)$$

There is no evidence of a Home Field effect, controlling for Distance (the point estimate was actually negative).

The Hosmer-Lemeshow test is given below, with $g = 10$ groups. The groups are set up so that all kicks of the same yardage are in the same group. A plot of the fitted equation and the sample proportions of Successes at each of the yardage distances is given in Figure 5.4. McFadden's Pseudo-R^2 based on the model with Distance as the only predictor is $R^2_{McF} = 1 - ((-343.46)/(-408.86)) = 0.1600$ The R output is given below.

$$\nabla$$

Group (i)	DistRange	n_i	$\bar{\hat{\pi}}_i$	o_i	$n_i\bar{\hat{\pi}}_i$	HL X^2
1	18-23	117	0.9847	116	115.2043	0.358088
2	24-26	84	0.9762	82	81.99916	3.6E-07
3	27-30	117	0.9649	112	112.8906	0.200031
4	31-33	102	0.9470	96	96.59051	0.068071
5	34-36	89	0.9272	83	82.5218	0.03807
6	37-39	108	0.8984	99	97.02217	0.396656
7	40-42	93	0.8580	79	79.79007	0.055076
8	43-45	98	0.8114	86	79.51427	2.804541
9	46-48	96	0.7434	59	71.36537	8.349341
10	49-76	135	0.6156	88	83.10177	0.751015
					H-L GOF	13.02089
					df	8
					$X^2_{.05}$	15.50731
					P-value	0.111133

Table 5.5: Hosmer-Lemeshow test for NFL Field Goal Attempts - 2008 Regular Season

```
> fga.mod1 <- glm(GOOD ~ 1, binomial("logit"))
> summary(fga.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.86792    0.09113    20.5   <2e-16 ***

    Null deviance: 817.72  on 1038  degrees of freedom
Residual deviance: 817.72  on 1038  degrees of freedom
AIC: 819.72
> logLik(fga.mod1)
’log Lik.’ -408.8614 (df=1)
>
> fga.mod2 <- glm(GOOD ~ distance, binomial("logit"))
> summary(fga.mod2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.76271    0.54443  12.422   <2e-16 ***
distance    -0.12084    0.01229  -9.836   <2e-16 ***

    Null deviance: 817.72  on 1038  degrees of freedom
Residual deviance: 686.93  on 1037  degrees of freedom
AIC: 690.93
> logLik(fga.mod2)
’log Lik.’ -343.4635 (df=2)
>
> fga.mod3 <- glm(GOOD ~ distance + homekick, binomial("logit"))
> summary(fga.mod3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.89871    0.55851  12.352   <2e-16 ***
```

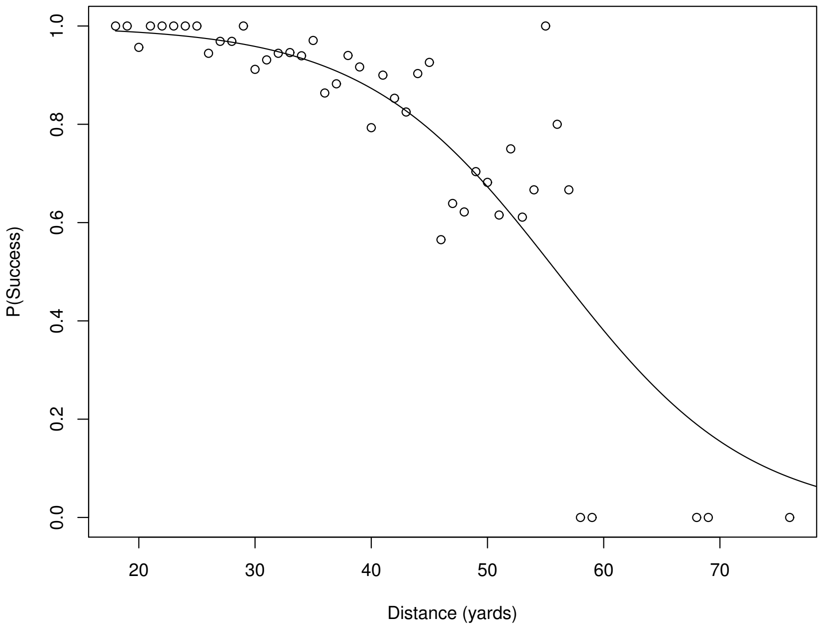

Figure 5.4: Fitted equation relating P(Success to Distance) and observed proportions - NFL Field Goal Attempts

```
distance    -0.12083    0.01233  -9.795   <2e-16 ***
homekick    -0.25959    0.19716  -1.317    0.188

    Null deviance: 817.72  on 1038  degrees of freedom
Residual deviance: 685.19  on 1036  degrees of freedom
AIC: 691.19
> logLik(fga.mod3)
'log Lik.' -342.5926 (df=3)
```

Overdispersion

When the model is based on grouped data and the Pearson Chi-square statistic based on the sum of squared Pearson residuals is much larger than $m-(p+1)$, its degrees of freedom, there is evidence of overdispersion. This is when the data are more variable than expected based on the binomial model. This can be evidence of a mis-specified model, with missing important predictors and/or polynomial and/or interaction terms.

One remedy is to fit a **quasibinomial** model and adjust the standard errors of the regression coefficients used in the Wald tests for the individual coefficients. An approximate F-test for multiple regression coefficients can be conducted as well (see e.g. Faraway, 2006, Section 2.11, [12]).

$$\hat{\sigma}^2 = \frac{X_P^2}{m-(p+1)} \qquad \hat{SE}^*\left\{\hat{\beta}_j\right\} = \hat{\sigma}\hat{SE}\left\{\hat{\beta}_j\right\} Z_W^* = \frac{\hat{\beta}_j}{\hat{SE}^*\left\{\hat{\beta}_j\right\}} \qquad F_{LR}^* = \frac{\left[\frac{D_R-D_F}{df_R-df_F}\right]}{\hat{\sigma}^2}$$

Here D_R and D_F are the deviances for the Reduced and Full models and df_R and df_F are their residual degrees of freedom $(m - \frac{\#parms}{)}$.

Example 5.8: Toxicity of Chemicals on Beetles

A study considered the effects of two chemicals on beetles (Hewlett, 1969, [15]). There were $m = 13$ combinations of levels of pyrethrin (X_1) and piperonyl butoxide (X_2), with approximately 150 beetles exposed in each combination (and 200 in the control condition). The data, fitted values, Pearson residuals and χ_P^2 statistic are given in Table 5.6 for the following second order model in the logit form.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2$$

The Pearson Goodness-of-Fit test rejects the null, and there is evidence of overdispersion with $\hat{\sigma}^2 = 40.7113/7 = 5.8159$. For the quasibinomial model, the estimated standard errors of the regression coefficients are multiplied by $\hat{\sigma} = \sqrt{5.8159} = 2.4116$. From the R output below and the original binomial fit, the estimate, standard error, and Wald test statistic for β_{12} are given below, followed by the adjustment for the quasibinomial model.

$$\text{Binomial: } \hat{\beta}_{12} = 0.3897 \qquad \hat{SE}\left\{\hat{\beta}_{12}\right\} = 0.1722 \qquad Z_W = 2.263 \quad 2P\left(Z \geq 2.263\right) = .0237$$

$$\text{Quasibinomial: } \hat{\beta}_{12} = 0.3897 \quad \hat{SE}^*\left\{\hat{\beta}_{12}\right\} = 2.4116(0.1722) = 0.4153 \quad Z_W^* = 0.938 \quad 2P\left(Z \geq 0.938\right) = .3793$$

The scaling of the standard error by 2.4 reduces the z_W statistic, and it is no longer significant for the interaction term. The residual deviance for the Full model (with interaction and quadratic terms) is

$D_F = 36.157$ with degrees of freedom $df_F = 13 - 6 = 7$. For the Reduced model with only main effects ($\beta_{12} = \beta_{11} = \beta_{22} = 0$), the deviance is $D_R = 304.11$ with degrees of freedom $df_R = 13 - 3 = 10$. The approximate F-test is given below.

$$H_0 : \beta_{12} = \beta_{11} = \beta_{22} = 0 \qquad TS : F_{LR}^* = \frac{\left[\frac{304.11-36.157}{10-7}\right]}{5.8159} = 15.357 \qquad P\left(F_{3,7} \geq 15.357\right) = .0018$$

There is evidence at least one of the higher order terms is significant. R output is given below.

∇

i	pyreth (X_1)	pipBut (X_2)	numExps (n_i)	Mortality (y_i)	y_i/n_i	$\hat{\pi}_i$	$n_i\hat{\pi}_i$	e_i^P
1	1.5	0	150	138	0.9200	0.9096	136.4459	0.4426
2	1.06	0	149	75	0.5034	0.6475	96.4833	-3.6840
3	0.75	0	150	32	0.2133	0.2365	35.4755	-0.6678
4	1.1	0.25	151	129	0.8543	0.7772	117.3589	2.2766
5	0.78	0.25	151	65	0.4305	0.3628	54.7821	1.7294
6	0.55	0.25	150	19	0.1267	0.1013	15.1975	1.0289
7	0.8	2.5	149	143	0.9597	0.9365	139.5334	1.1643
8	0.57	2.5	150	112	0.7467	0.7094	106.4097	1.0053
9	0.4	2.5	140	37	0.2643	0.3528	49.3938	-2.1921
10	0.65	10	150	141	0.9400	0.9654	144.8115	-1.7030
11	0.46	10	150	117	0.7800	0.7622	114.3282	0.5124
12	0.32	10	149	56	0.3758	0.3669	54.6652	0.2269
13	0	0	200	1	0.0050	0.0006	0.1141	2.6228
							X_P^2	40.7113
							df	7
							$\chi^2(.05)$	14.0671
							P-value	0.0000

Table 5.6: Beetle Mortality in relation pyrethrine and piperonyl butoxide concentration

```
> trib.mod1 <- glm(y.trib ~ 1, binomial("logit"))
> summary(trib.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14202    0.04496   3.159  0.00158 **

    Null deviance: 1038.1  on 12  degrees of freedom
Residual deviance: 1038.1  on 12  degrees of freedom
AIC: 1100.1
> logLik(trib.mod1)
'log Lik.' -549.0736 (df=1)

> trib.mod2 <- glm(y.trib ~ pyreth + pipBut,
+                     binomial("logit"))
> summary(trib.mod2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.77521    0.21048  -17.94   <2e-16 ***
pyreth       4.53425    0.24630   18.41   <2e-16 ***
pipBut       0.28567    0.01706   16.74   <2e-16 ***

    Null deviance: 1038.05  on 12  degrees of freedom
```

```
Residual deviance:  304.11  on 10  degrees of freedom
AIC: 370.21
> logLik(trib.mod2)
'log Lik.' -182.1043 (df=3)

> trib.mod4 <- glm(y.trib ~ pyreth + pipBut,
+                      quasibinomial("logit"))
> summary(trib.mod4)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77521    1.12229  -3.364   0.0072 **
pyreth       4.53425    1.31330   3.453   0.0062 **
pipBut       0.28567    0.09098   3.140   0.0105 *

    Null deviance: 1038.05  on 12  degrees of freedom
Residual deviance:  304.11  on 10  degrees of freedom
AIC: NA
> logLik(trib.mod4)
'log Lik.' NA (df=3)
>
> trib.mod5 <- glm(y.trib ~ pyreth + pipBut + I(pyreth*pipBut) +
+                       I(pyreth^2) + I(pipBut^2),
+                      binomial("logit"))
> summary(trib.mod5)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.46808    0.85850  -8.699  < 2e-16 ***
pyreth             10.27147    1.74038   5.902 3.59e-09 ***
pipBut              1.38568    0.19737   7.021 2.21e-12 ***
I(pyreth * pipBut)  0.38969    0.17221   2.263  0.02365 *
I(pyreth^2)        -2.50217    0.86490  -2.893  0.00382 **
I(pipBut^2)        -0.11212    0.01187  -9.443  < 2e-16 ***

    Null deviance: 1038.053  on 12  degrees of freedom
Residual deviance:   36.157  on  7  degrees of freedom
AIC: 108.25
> logLik(trib.mod5)
'log Lik.' -48.12553 (df=6)
> e.p5 <- resid(trib.mod5,type="pearson")
> e.d5 <- resid(trib.mod5,type="deviance")
> sum(e.p5^2)
[1] 40.71129
> sum(e.d5^2)
[1] 36.15705
>
> trib.mod6 <- glm(y.trib ~ pyreth + pipBut + I(pyreth*pipBut) +
+                       I(pyreth^2) + I(pipBut^2),
+                      quasibinomial("logit"))
> summary(trib.mod6)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -7.46808    2.07037  -3.607  0.00866 **
pyreth             10.27147    4.19712   2.447  0.04428 *
pipBut              1.38568    0.47599   2.911  0.02262 *
I(pyreth * pipBut)  0.38969    0.41531   0.938  0.37933
I(pyreth^2)        -2.50217    2.08582  -1.200  0.26931
I(pipBut^2)        -0.11212    0.02863  -3.916  0.00578 **

    Null deviance: 1038.053  on 12  degrees of freedom
Residual deviance:   36.157  on  7  degrees of freedom
AIC: NA
> # logLik(trib.mod5)
> anova(trib.mod4, trib.mod6, test="F")
Analysis of Deviance Table
```

```
Model 1: y.trib ~ pyreth + pipBut
Model 2: y.trib ~ pyreth + pipBut + I(pyreth * pipBut) + I(pyreth^2) +
    I(pipBut^2)
  Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
1        10    304.115
2         7     36.157  3   267.96 15.358 0.001838
```

5.2.2 Poisson Regression for Counts

When data are counts, a Poisson Regression model is often fit. The mean of the distribution must be positive, but the log of the mean can take on positive or negative values, so the log of the mean is typically modeled as a linear function of the predictors.

$$\log(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad \Rightarrow \qquad \mu = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$$

The interpretation of the regression coefficients is that the mean of Y changes multiplicatively by e^{β_j} when X_j increases by 1 unit, holding all other predictors constant. The Poisson distribution is restricted to have its mean and variance to be the same. In many applications, data display overdispersion where the variance exceeds the mean. Adjustments can be made as in the case of the Logistic Regression model described in the previous subsection, or a 2-parameter Negative Binomial model can be fit, which allows for the variance to be larger than the mean. The Negative Binomial will be described in the next subsection.

When there is a fixed number of n distinct $\mathbf{X}$ levels, the Pearson and Likelihood-Ratio (Deviance) Goodness-of-Fit statistics given below have approximate chi-square distributions with $n - p'$ degrees of freedom (see e.g. Agresti, 1996, pp.89-90, [3]).

$$\text{Pearson: } X_P^2 \quad = \quad \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad = \quad \sum_{i=1}^{n} e_{iP}^2 \qquad e_{iP} = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}$$

$$\text{Deviance: } X_D^2 = G^2 = 2\sum_{i=1}^{n}\left[y_i \ln\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right] \quad = \quad \sum_{i=1}^{n} e_{iD}^2 \qquad e_{iD} = \text{sign}\{y_i - \hat{\mu}_i\}\left[2y_i \ln\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right]^{1/2}$$

When the independent variable(s) have many distinct level(s) or combinations, observations can be grouped based on their X levels in cases where there is $p = 1$ independent variable, or grouped based on their predicted means in general. The sums of events (y) and their corresponding predicted values ($\hat{\mu}_i$) are obtained for each group. Pearson residuals are obtained for each group based on the sums of their observed and predicted values. If we have g groups and p predictors, the approximate Pearson chi-square statistic will have $g - p'$ degrees of freedom (see e.g. Agresti, 1996, p. 90, [3]).

Example 5.9: NASCAR Crashes - 1972-1979 Seasons - Poisson Model

NASCAR is a professional stock car racing organization in the United States. The top division is currently called the Monster Energy Cup. We consider all races during the 1972-1979 season (Winner, 2006, [36]).

The response is the number of Caution Flags (a proxy for crashes) for each race (Y), and the predictors considered are: Number of **D**rivers, **T**rack Length, and Number of **L**aps. During this period, there were $n = 151$ races. Table 5.7 contains summaries (quantiles) and correlations for D, T,L, and Y. The pairs of independent variables with high correlations are Track Length is highly correlated with Laps ($-.901$) and Drivers (.731). The following models are fit, using R's **glm** function.

$$\text{Model 1:} \log(\hat{\mu}_i) = 1.5657 \quad \hat{\mu}_i = e^{1.5657} = 4.7861 \quad l_1 = -353.6927 \quad D_1 = 215.49 \quad df_1 = 151 - 1 = 150$$

$$\text{Model 2:} \log(\hat{\mu}_i) = -0.7963 + 0.0365D + 0.1145T + 0.0026L \qquad \hat{\mu}_i = e^{-0.7963+0.0365D+0.1145T+0.0026L}$$

$$l_2 = -331.5551 \qquad D_2 = 171.22 \qquad df_2 = 151 - 4 = 147$$

$$\hat{SE}\left\{\hat{\beta}_D\right\} = 0.00843 \quad Z_{WD} = \frac{0.0365}{0.00843} = 5.081 \qquad \hat{SE}\left\{\hat{\beta}_L\right\} = 0.00034 \quad Z_{WD} = \frac{0.0021}{0.00034} = 6.153$$

A test of whether any of the three predictors are related to the number of crashes is given below, followed by the partial tests for the individual predictors.

$$H_0 : \beta_D = \beta_T = \beta_L = 0 \qquad TS : X^2_{LR} = -2\left(l_1 - l_2\right) = D_1 - D_2 = 44.27 \qquad P\left(X^2_3 \geq 44.27\right) < .0001$$

$$H_{0j} : \beta_j = 0 \qquad Z_{WD} = \frac{0.0365}{0.0125} = 2.924 \qquad Z_{WT} = \frac{0.1145}{0.1684} = 0.680 \qquad Z_{WL} = \frac{0.0026}{0.0008} = 3.289$$

There is evidence that as the number of drivers D and number of laps L increase, so do the number of crashes. Note that track length and number of laps have a large negative correlation, when laps is already in the model, adding track length does not improve the fit. A third model is fit, removing track length.

$$\text{Model 3:} \log(\hat{\mu}_i) = -0.6876 + 0.0428D + 0.0021L \qquad \hat{\mu}_i = e^{-0.6876+0.0428D+0.0021L}$$

$$l_3 = -331.7863 \qquad D_3 = 171.68 \qquad df_3 = 151 - 3 = 148$$

The Pearson residuals and Goodness-of-Fit test based on $g = 12$ groups and $12 - 3 = 9$ degrees of freedom and the group means and variances are given in Table 5.8. There is evidence that the Poisson model is not a good fit. The Pearson Chi-square statistic based on the raw (not grouped) data is $\chi^2_P = 159.17$ with degrees of freedom $df = 151 - 3 = 148$.

$$\hat{\sigma}^2 = \frac{159.17}{148} = 1.0755 \qquad \hat{\sigma} = \sqrt{1.0755} = 1.0371$$

The quasipoisson model will only make minor adjustments to the Poisson model.

$$\hat{SE}^*\left\{\hat{\beta}_D\right\} = 1.0371(0.00843) = 0.000874 \quad Z^*_{WD} = \frac{0.0365}{0.00874} = 4.899$$

$$\hat{SE}^*\left\{\hat{\beta}_L\right\} = 1.0371(0.00034) = 0.00036 \quad Z^*_{WD} = \frac{0.0021}{0.00036} = 5.933$$

R output is given below.

∇

```
> race.mod1 <- glm(cautions ~ 1, poisson("log"))
> summary(race.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.56751    0.03716   42.18   <2e-16 ***
    Null deviance: 215.49  on 150  degrees of freedom
```

	Quantiles						Correlations			
Variable	Min	25%	Med	Mean	75%	Max	TrkLen	Drivers	Laps	Cautions
TrkLen	0.526	0.625	1.366	1.446	2.500	2.660	1	0.731	-0.901	-0.317
Drivers	22	30	36	35.2	40	50	0.731	1	-0.508	0.160
Laps	95	200	367	339.7	420	500	-0.901	-0.508	1	0.297
Cautions	0	3	5	4.795	6	12	-0.317	0.160	0.297	1

Table 5.7: NASCAR Caution Flag, Track Length, Drivers, Laps: Summaries and Correlations

Group	# Races	Total Obs	Total Expected	e_i^P	Mean	Variance
1	15	37	46.155	-1.348	2.467	3.552
2	11	50	39.374	1.693	4.545	2.273
3	11	39	42.136	-0.483	3.545	3.473
4	14	58	57.210	0.104	4.143	4.901
5	16	53	67.806	-1.798	3.312	4.763
6	4	20	17.308	0.647	5.000	0.667
7	17	80	76.852	0.359	4.706	5.346
8	21	129	108.131	2.007	6.143	5.529
9	2	10	10.815	-0.248	5.000	2.000
10	8	40	45.360	-0.796	5.000	2.857
11	24	167	154.125	1.037	6.958	6.216
12	8	41	58.729	-2.313	5.125	9.839
			X_P^2	19.856		
			df	9		
			$X_{.05}^2$	16.919		
			P-value	0.0188		

Table 5.8: NASCAR Caution Flag Goodness-of-Fit Test and Group Means and Variances

```
Residual deviance: 215.49  on 150  degrees of freedom
AIC: 709.39
> logLik(race.mod1)
’log Lik.’ -353.6927 (df=1)
>
> race.mod2 <- glm(cautions ~ drivers + trklen + laps, poisson("log"))
> summary(race.mod2)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7962699  0.4116942  -1.934  0.05310 .
drivers      0.0365253  0.0124932   2.924  0.00346 **
trklen       0.1144986  0.1684236   0.680  0.49662
laps         0.0025963  0.0007893   3.289  0.00100 **

    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.22  on 147  degrees of freedom
AIC: 671.11
> logLik(race.mod2)
’log Lik.’ -331.5551 (df=4)

> race.mod3 <- glm(cautions ~ drivers + laps, poisson("log"))
> summary(race.mod3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.6876335  0.3776880  -1.821   0.0687 .
drivers      0.0428077  0.0084250   5.081 3.75e-07 ***
laps         0.0021136  0.0003435   6.153 7.59e-10 ***

    Null deviance: 215.49  on 150  degrees of freedom
Residual deviance: 171.68  on 148  degrees of freedom
AIC: 669.57
> logLik(race.mod3)
’log Lik.’ -331.7863 (df=3)
>
> round(gof.grp,3)
      # Races Total Obs Total Exp Pearson r  Mean Variance
 [1,]      15        37    46.155    -1.348 2.467    3.552
 [2,]      11        50    39.374     1.693 4.545    2.273
 [3,]      11        39    42.136    -0.483 3.545    3.473
 [4,]      14        58    57.210     0.104 4.143    4.901
 [5,]      16        53    67.806    -1.798 3.312    4.763
 [6,]       4        20    17.308     0.647 5.000    0.667
 [7,]      17        80    76.852     0.359 4.706    5.346
 [8,]      21       129   108.131     2.007 6.143    5.529
 [9,]       2        10    10.815    -0.248 5.000    2.000
[10,]       8        40    45.360    -0.796 5.000    2.857
[11,]      24       167   154.125     1.037 6.958    6.216
[12,]       8        41    58.729    -2.313 5.125    9.839

> race.mod3a <- glm(formula = cautions ~ drivers + laps,
+  family=quasipoisson("log"))
> summary(race.mod3a)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6876335  0.3916893  -1.756   0.0812 .
drivers      0.0428077  0.0087373   4.899 2.49e-06 ***
laps         0.0021136  0.0003562   5.933 2.02e-08 ***
```

Models with Varying Exposures

In many studies, interest is in comparing rates of events in groups or observations with different amounts of exposure to the outcome of interest. In these cases, the response is the number of of observations per unit of exposure. A log linear model is assumed for the expectation of the ratio. The fixed exposure (in the log model) is referred to as an **offset**. The model is as follows.

$$\text{Sample Rate: } \frac{Y_i}{t_i} \qquad E\left\{\frac{Y_i}{t_i}\right\} = \frac{\mu_i}{t_i} \qquad \log\left(\frac{\mu_i}{t_i}\right) = \log\left(\mu_i\right) - \log\left(t_i\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

Note that we will place $\log\left(t_i\right)$ on the right-hand side of the equal sign, but do not want to put a regression coefficient on it. In statistical software packages, an offset option is typically available. The predicted values for the observations are given below.

$$\hat{Y}_i = t_i\hat{\mu}_i = \exp\{\log\left(t_i\right) + \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}\} = t_i \exp\{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}\}$$

Inferences are conducted as in the previously described Poisson model.

Example 5.10: Friday the 13th, Gender, and Traffic Deaths

A study reported incidences of traffic deaths by gender on Friday the 13th and other Fridays in Finland over the years 1971-1997 (Nayha (2002)). The response was the number of traffic deaths and the exposure was the number of person-days (100000s). The groups were the 4 combinations of Friday type ($X_{i1} = 1$ if Friday the 13th, 0 otherwise) and Gender ($X_{i2} = 1$ if Female, 0 if Male). The model contains an interaction term, $X_{i3} = X_{i1}X_{i2}$, which allows the Friday the 13th effect to differ by Gender (and vice versa). Table 5.9 gives the data, exposure, the independent variables, the predicted mean, and Total Death Rate per 100000 exposures for the four classifications/groups.

For Males and Females, the Friday the 13th effects are given below.

$$\text{Males: } \frac{\exp\{\beta_0 + \beta_1\}}{\exp\{\beta_0\}} = \exp\{\beta_1\} \qquad \text{Females: } \frac{\exp\{\beta_0 + \beta_1 + \beta_2 + \beta_3\}}{\exp\{\beta_0 + \beta_2\}} = \exp\{\beta_1 + \beta_3\}$$

Thus, a significant interaction ($\beta_3 \neq 0$) implies the Friday the 13th effect is not the same among Males and Females. To obtain 95% Confidence Intervals for the Male and Female Friday the 13th effects, first obtain 95% CIs for β_1 and $\beta_1 + \beta_3$, then exponentiate the endpoints.

$$\beta_1: \quad \hat{\beta}_1 \pm 1.96\hat{SE}\left\{\hat{\beta}_1\right\} \qquad \beta_1 + \beta_3: \quad \hat{\beta}_1 + \hat{\beta}_3 \pm 1.96\sqrt{\hat{V}\left\{\hat{\beta}_1\right\} + \hat{V}\left\{\hat{\beta}_3\right\} + 2\hat{\text{COV}}\left\{\hat{\beta}_1, \hat{\beta}_3\right\}}$$

The R Program and Output are given below. The Friday the 13th effect is not significant for Males, as the 95% CI for their Risk Ratio (0.8442,1.3110) contains 1. For Females, there is evidence of a Friday the 13th effect, as the 95% CI for their Risk Ratio (1.1793,2.2096) is entirely above 1.

```
### Program
Y.13 <- c(41,82,789,2423)
t.13 <- c(86.5,79.9,2687.1,2483.7)
X0.13 <- c(1,1,1,1)
X1.13 <- c(1,1,0,0)
X2.13 <- c(1,0,1,0)
X3.13 <- c(1,0,0,0)

f13.mod1 <- glm(Y.13 ~ X1.13 + X2.13 + X3.13, offset=log(t.13),
    family=poisson("log"))
summary(f13.mod1)
vcov(f13.mod1)
beta1.hat <- coef(f13.mod1)[2]
se.beta1.hat <- sqrt(vcov(f13.mod1)[2,2])
beta13.hat <- beta1.hat + coef(f13.mod1)[4]
se.beta13.hat <- sqrt(vcov(f13.mod1)[2,2]+vcov(f13.mod1)[4,4]+
   2*vcov(f13.mod1)[2,4])
ll.beta1 <- beta1.hat - 1.96*se.beta1.hat
ul.beta1 <- beta1.hat + 1.96*se.beta1.hat
ll.beta13 <- beta13.hat - 1.96*se.beta13.hat
ul.beta13 <- beta13.hat + 1.96*se.beta13.hat

f13.eff.m <- cbind(beta1.hat,se.beta1.hat,ll.beta1,ul.beta1,
      exp(beta1.hat),exp(ll.beta1),exp(ul.beta1))
f13.eff.f <- cbind(beta13.hat,se.beta13.hat,ll.beta13,ul.beta13,
      exp(beta13.hat),exp(ll.beta13),exp(ul.beta13))
f13.eff <- rbind(f13.eff.m,f13.eff.f)
rownames(f13.eff) <- c("Males","Females")
colnames(f13.eff) <- c("Estimate","Std Err","LL","UL",
  "Risk Ratio", "LL RR", "UL RR")
round(f13.eff,4)

### Output
> summary(f13.mod1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.02474    0.02032  -1.218   0.2232
X1.13        0.05069    0.11228   0.451   0.6517
X2.13       -1.20071    0.04099 -29.293   <2e-16 ***
X3.13        0.42819    0.19562   2.189   0.0286 *

(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1.0258e+03  on 3  degrees of freedom
Residual deviance: 1.8230e-13  on 0  degrees of freedom
AIC: 37.942
Number of Fisher Scoring iterations: 2

> vcov(f13.mod1)
              (Intercept)         X1.13         X2.13         X3.13
(Intercept)  0.0004127115 -0.0004127115 -0.0004127115  0.0004127115
X1.13       -0.0004127115  0.0126078244  0.0004127115 -0.0126078244
X2.13       -0.0004127115  0.0004127115  0.0016801386 -0.0016801386
X3.13        0.0004127115 -0.0126078244 -0.0016801386  0.0382654231

> round(f13.eff,4)
        Estimate Std Err      LL     UL Risk Ratio  LL RR  UL RR
Males     0.0507  0.1123 -0.1694 0.2708     1.0520 0.8442 1.3110
Females   0.4789  0.1602  0.1649 0.7928     1.6143 1.1793 2.2096
```

∇

Group	i	Y_i	t_i	X_{i1}	X_{i2}	X_{i3}	$\log\left(\hat{\lambda}_i\right)$	$TDR_i = \hat{\lambda}_i$
Friday 13th/Female	1	41	86.5	1	1	1	-0.7466	0.4740
Friday 13th/Male	2	82	79.9	1	0	0	0.0259	1.0263
Other Friday/Female	3	789	2687.1	0	1	0	-1.2255	0.2936
Other Friday/Male	4	2423	2483.7	0	0	0	-0.0247	0.9756

Table 5.9: Friday the 13th and Gender for Finland Traffic Deaths

5.2.3 Negative Binomial Model

In this subsection, the two-parameter Negative Binomial model is fit to the NASCAR lead change data. The log link is used, as was used for the Poisson model. The predictors are **D**rivers, **T**rack Length, and **L**aps, for the $n = 151$ races.

Example 5.11: NASCAR Lead Changes - 1972-1979 Seasons - Negative Binomial Model

For the Lead Change outcome from the NASCAR 1972-1979 seasons there is much more overdispesion than in the case of the Crash outcome. For the Poisson model, all three predictors are highly significant with Z_W values greater than 5 (R Output given below). The Pearson Chi-Square statistic for the ungrouped data is $X_P^2 = 721.89$ with $df = 151 - 4 = 147$. This leads to the following overdispersion parameter and adjustments.

$$\hat{\sigma}^2 = \frac{721.89}{147} = 4.911 \qquad \hat{\sigma} = \sqrt{4.911} = 2.2160 \qquad \hat{SE}^* \left\{\hat{\beta}_j\right\} = 2.2160\hat{SE}\left\{\hat{\beta}_j\right\}$$

After the adjustment to the standard errors of the regression coefficients, all Z_W^* statistics remain at least 2.440.

When the data are grouped into $g = 10$ groups of races based on their predicted counts, the approximate (grouped) Pearson Chi-Square statistic is $X_P^2 = 107.4$ based on $g - p' = 10 - 4 = 6$. The P-value is virtually 0.

A Negative Binomial model is fit, leading to the following fitted equation and parameter estimates.

$$\hat{\mu}_i = e^{-0.5038+0.0597D_i+0.5153T_i+0.0017L_i} \qquad \hat{\alpha^{-1}} = 5.248 \qquad \hat{V}\left\{Y_i\right\} = \hat{\mu}_i\left(1 + \frac{\hat{\mu}_i}{5.248}\right)$$

The residual deviance for this model is $D = 162.8$ on $151 - 4 = 147$ degrees of freedom. This appears to fit the data well (as the ratio $162.8/147 \approx 1$). The grouped Goodness-of-Fit Pearson Chi-Square statistic with $g = 10$ groups leads to $X_P^2 = 1.79$, which is very small. The ratio of the variances to the mean and their group expected values are consistent (given in R output below).

∇

```
## Poisson Model
> race.mod2 <- glm(leadchng ~ drivers + trklen + laps, poisson("log"))
> summary(race.mod2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -0.490253   0.217796  -2.251   0.0244 *
drivers      0.051612   0.005679   9.089  < 2e-16 ***
trklen       0.610414   0.082949   7.359 1.85e-13 ***
laps         0.002138   0.000415   5.152 2.58e-07 ***

    Null deviance: 1388.59  on 150  degrees of freedom
Residual deviance:  687.61  on 147  degrees of freedom
AIC: 1395.2
> logLik(race.mod2)
’log Lik.’ -693.597 (df=4)

> (pearson.x2 <- sum((leadchng - muhat)^2/muhat))
[1] 655.6059
> (pearson.x2a <- sum(resid(race.mod2,type="pearson")^2))
[1] 655.6059
> (deviance.x2 <- sum(resid(race.mod2)^2))
[1] 687.6124

## Grouped Poisson Model
      # Races Total Obs Total Exp Pearson r   Mean Variance
 [1,]      15       113   130.541    -0.302  7.533   23.410
 [2,]      15       138   151.055    -0.195  9.200   34.457
 [3,]      14       178   155.015     0.334 12.714   41.297
 [4,]      17       321   270.637     0.422 18.882   56.360
 [5,]      19       485   390.052     0.554 25.526   89.930
 [6,]      15       191   326.424    -0.943 12.733   48.210
 [7,]      16       353   407.890    -0.306 22.062   74.329
 [8,]      16       491   453.298     0.189 30.688  183.696
 [9,]      11       349   373.342    -0.148 31.727  201.818
[10,]      13       574   541.275     0.138 44.154  229.474

> (pearson.X2.mg <- sum(pearson.r^2))
[1] 109.6503
> qchisq(.95,10-3-1)
[1] 12.59159
> (pval.mg <- 1-pchisq(pearson.X2.mg,10-3-1))
[1] 0

> ### quasipoisson takes SE(beta)*sqrt(phi)
> ### phi = pearson.x2/df
> race.mod2a <- glm(formula = leadchng ~ drivers + trklen + laps,
+  family=quasipoisson("log"))
> summary(race.mod2a)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4902529  0.4599522  -1.066  0.28823
drivers      0.0516117  0.0119924   4.304 3.05e-05 ***
trklen       0.6104140  0.1751757   3.485  0.00065 ***
laps         0.0021381  0.0008764   2.440  0.01590 *

    Null deviance: 1388.59  on 150  degrees of freedom
Residual deviance:  687.61  on 147  degrees of freedom
AIC: NA

## Negative Binomial Model
race.mod4 <- glm.nb(leadchng ~ drivers + trklen + laps, link=log)
> summary(race.mod4)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.5037750  0.4486264  -1.123  0.26147
drivers      0.0596900  0.0136021   4.388 1.14e-05 ***
trklen       0.5152983  0.1789336   2.880  0.00398 **
laps         0.0017422  0.0008671   2.009  0.04452 *
```

```
    Null deviance: 308.06  on 150  degrees of freedom
Residual deviance: 162.80  on 147  degrees of freedom
AIC: 1098.1
              Theta:  5.248
          Std. Err.:  0.809

 2 x log-likelihood:  -1088.054
> logLik(race.mod4)
'log Lik.' -544.0272 (df=5)

## Grouped data - Negative Binomial model
> round(gof.grp.nb,3)
      # Races Obs     Exp Pearson r   Mean     Var Var/Mean Exp V/M
 [1,]      15 113 130.541    -0.302  7.533  23.410    3.107   2.435
 [2,]      15 138 151.055    -0.195  9.200  34.457    3.745   2.753
 [3,]      14 178 155.015     0.334 12.714  41.297    3.248   3.422
 [4,]      17 321 270.637     0.422 18.882  56.360    2.985   4.597
 [5,]      19 485 390.052     0.554 25.526  89.930    3.523   5.863
 [6,]      15 191 326.424    -0.943 12.733  48.210    3.786   3.426
 [7,]      16 353 407.890    -0.306 22.062  74.329    3.369   5.203
 [8,]      16 491 453.298     0.189 30.688 183.696    5.986   6.846
 [9,]      11 349 373.342    -0.148 31.727 201.818    6.361   7.044
[10,]      13 574 541.275     0.138 44.154 229.474    5.197   9.411
>
> pearson.r <- (sum.mg - sum.muhat.mg) / sqrt(sum.muhat.mg + sum.muhat.mg^2*0.1905)
> (pearson.X2.mg <- sum(pearson.r^2))
[1] 1.785865
> qchisq(.95,10-3-1)
[1] 12.59159
> (pval.mg <- 1-pchisq(pearson.X2.mg,10-3-1))
[1] 0.9383018
```

5.2.4 Gamma Model

In the case of fitting a Gamma model to continuous, strictly positive data, three common link functions are used: the identity, the inverse, and the log. These are given below (the log link has been given above).

$$\text{Identity: } \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \qquad \textbf{Inverse} : \mu = \frac{1}{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

For the identity link, the predicted values for the individual cases make use of the regression equation directly, while for the reciprocal link, the inverse of regression equation is used.

Example 5.12: Napa Marathon Velocities - Gamma Model

The Napa Valley marathon in 2015 had 977 Males and 905 Females complete the 26.2 mile race. Consider a model relating runners' speeds in miles per hour (Y=mph) to Gender ($M = 1$ if Male, 0 if Female), Age (A, in Years), and an interaction term (AM, allowing for different slopes with respect to age for Males and Females. Figure 5.5 plots the reciprocal of mph and the log of mph separately for Males and Females. Both the inverse link and log link models are fit below using the **glm** function in R. Note that the "default" link for the gamma regression model in the glm function is the inverse link.

First, models with **A**ge, a dummy variable for **M**ale, and the **AM** cross-product term for the interaction were fit. The P-values for the cross-product terms were not significant based on the Wald tests (P=.1493 for identity, P=.0741 for log). The fitted equations, log-likelihoods, residual deviances, and degrees of freedom for the two link functions and the interaction models are given below.

$$\text{Identity Link/Interaction Model: } \hat{\mu} = (.1571 + .0003025A - .02281M + .0001751AM)^{-1}$$

$$l_{II} = -2732.766 \quad D_{II} = 53.699 \quad df = 1878$$

$$\text{Log Link/Interaction Model: } \hat{\mu} = e^{1.8494 - .0018A + .1522M - .0013AM}$$

*

$$l_{LI} = -2732.232 \quad D_{LI} = 53.668 \quad df = 1878$$

$$\text{Identity Link/Additive Model: } \hat{\mu} = (.1531 + .0004048A - .01574M)^{-1}$$

$$l_{IA} = -2733.829 \quad D_{IA} = 53.759 \quad df = 1879$$

$$\text{Log Link/Additive Model: } \hat{\mu} = e^{1.8776 - .0025A + .0972M} \qquad l_{LA} = -2733.85 \quad D_{LA} = 53.760 \quad df = 1879$$

The fits for the two link functions are very similar. The Additive models are appropriate for this data. This implies that the slopes for males and females are the same in the linear form of the models. The mean and variance for the Log Link/Additive model is given below. Note that the dispersion parameter reported by R is the reciprocal of the dispersion parameter.

$$\hat{\alpha} = \frac{1}{.02879762} = 34.725 \qquad \hat{\mu}_i = e^{1.8776 - .0025A + .0972M} \qquad \hat{V}\{Y_i\} = \frac{(\hat{\mu}_i)^2}{\hat{\alpha}}$$

For instance, the predicted means and variances for 25 year old females and 40 year old males are computed here.

$$\text{25 year old males: } \hat{\mu} = e^{1.8776 - .0025(25) + .0972(1)} = 6.769 \qquad \hat{V}\{Y\} = \frac{(6.769)^2}{34.725} = 1.319$$

$$\text{40 year old females: } \hat{\mu} = e^{1.8776 - .0025(40) + .0972(0)} = 5.916 \qquad \hat{V}\{Y\} = \frac{(5.916)^2}{34.725} = 1.008$$

The R output is given below.

∇

```
## Identity Link/Interaction Model
> napa.mod7 <- glm(mph~Age*gender,family=Gamma)
> summary(napa.mod7)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.571e-01  3.711e-03  42.331  < 2e-16 ***
Age          3.025e-04  9.275e-05   3.261  0.00113 **
genderM     -2.281e-02  5.068e-03  -4.501 7.19e-06 ***
Age:genderM  1.751e-04  1.214e-04   1.443  0.14930

(Dispersion parameter for Gamma family taken to be 0.02878207)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.699  on 1878  degrees of freedom
AIC: 5475.6
> logLik(napa.mod7)
'log Lik.' -2732.776 (df=5)

## Log Link/Interaction Model
```

```
> napa.mod8 <- glm(mph ~ Age*gender, family=Gamma(link="log"))
> summary(napa.mod8)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.8494178  0.0220581  83.843  < 2e-16 ***
Age         -0.0018116  0.0005461  -3.317 0.000927 ***
genderM      0.1521938  0.0315083   4.830 1.47e-06 ***
Age:genderM -0.0013388  0.0007422  -1.804 0.071404 .


(Dispersion parameter for Gamma family taken to be 0.02876127)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.668  on 1878  degrees of freedom
AIC: 5474.5
> logLik(napa.mod8)
’log Lik.’ -2732.232 (df=5)

## Identity Link/Additive Model
> napa.mod5 <- glm(mph~Age + gender,family=Gamma)
> summary(napa.mod5)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.531e-01  2.496e-03   61.34  < 2e-16 ***
Age          4.048e-04  5.988e-05    6.76 1.83e-11 ***
genderM     -1.574e-02  1.296e-03  -12.15  < 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.02879748)
    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.759  on 1879  degrees of freedom
AIC: 5475.7
> logLik(napa.mod5)
’log Lik.’ -2733.829 (df=4)

## Log Link/Additive Model
> napa.mod6 <- glm(mph ~ Age + gender, family=Gamma(link="log"))
> summary(napa.mod6)

Call:
glm(formula = mph ~ Age + gender, family = Gamma(link = "log"))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.877592   0.015512 121.044  < 2e-16 ***
Age         -0.002532   0.000370  -6.844 1.04e-11 ***
genderM      0.097190   0.007997  12.154  < 2e-16 ***

(Dispersion parameter for Gamma family taken to be 0.02879762)

    Null deviance: 58.586  on 1881  degrees of freedom
Residual deviance: 53.760  on 1879  degrees of freedom
AIC: 5475.7
> logLik(napa.mod6)
’log Lik.’ -2733.85 (df=4)
```

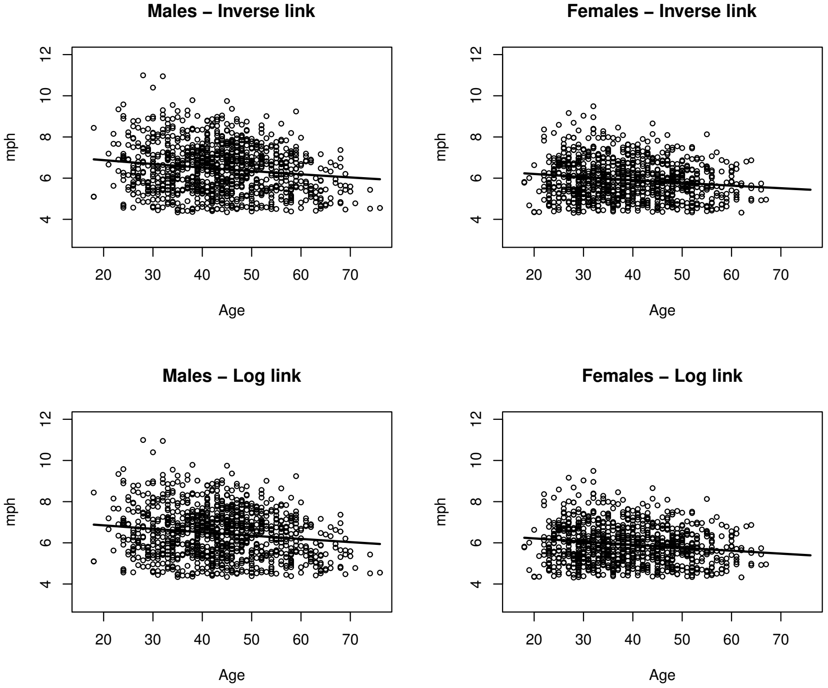

Figure 5.5: Data and Fitted Models by Gender and Link Function - 2015 Napa Marathon

5.3 Beta Regression

When data are rates or proportions, a regression model based on the Beta Distribution can be fit (Ferrari and Cribari-Neto, 2004, [13]). The mean and variance for a beta random variable are given below.

$$E\left\{Y\right\} = \mu = \frac{\alpha}{\alpha+\beta} \qquad V\left\{Y\right\} = \frac{\alpha\beta}{\left(\alpha+\beta+1\right)\left(\alpha+\beta\right)^2}$$

A re-parameterization of the model is used in estimating the regression model. Below is the re-parameterized likelihood function.

$$\phi = \alpha+\beta \qquad \Rightarrow \qquad \alpha = \mu\phi \qquad \beta = \left(1-\mu\right)\phi \qquad \Rightarrow \qquad E\left\{Y\right\} = \mu \qquad V\left\{Y\right\} = \frac{\mu\left(1-\mu\right)}{\phi+1}$$

Several link functions can be used, the logit link is used here as was done with the logistic regression function.

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 X_1 + \cdots \beta_P X_P \quad \rightarrow \quad \mu = \frac{e^{\beta_0+\beta_1 X_1+\cdots\beta_P X_P}}{1+e^{\beta_0+\beta_1 X_1+\cdots\beta_P X_P}}$$

In R, the "Precision Coefficient" is $\hat{\phi}$.

Example 5.13: Ford Prize Winnings in NASCAR Races: 1992-2000

The NASCAR Winston Cup series had $n = 267$ races during the years 1992-2000 (Winner, 2006, [36]). For each race, we obtain Y, the proportion of the prize money won by Ford cars. Variables used as predictors include: X_1, the proportion of cars in the race that are Fords, X_2, the track length (miles), X_3, the bank of the turns of the track (degrees), X_4, the number of laps, and dummy variables for the years 1993-2000. R output for the model is given below. Almost all predictors are significant with the exceptions of Bank and Year1994. Model 1 is the null model (intercept only) and Model 2 is the full model with the full set of predictors.

For the null model, the following parameter estimates are obtained, as well as mean and variance.

$$\hat{\mu} = \frac{e^{-0.10959}}{1+e^{-0.10959}} = .4726 \qquad \hat{\phi} = 62.548 \qquad \hat{\alpha} = .4726(62.548) = 29.562 \qquad \hat{\beta} = (1-.4726)(62.548) = 32.988$$

$$\hat{V}\{Y\} = \frac{\hat{\alpha}\hat{\beta}}{\left(\hat{\alpha}+\hat{\beta}\right)^2\left(\hat{\alpha}+\hat{\beta}+1\right)} = \frac{\hat{\mu}\,(1-\hat{\mu})}{\hat{\phi}+1} = \frac{.4726(1-.4726)}{62.548+1} = 0.003922$$

A Pseudo-R^2 measure can be obtained from the correlation between the logit transformed race proportions and the (linear) predicted values for them.

$$Y_i^* = \log\left(\frac{Y_i}{1-Y_i}\right) \qquad \hat{Y}_i* = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} \qquad \text{Pseudo-}R^2 = r^2_{Y_i^*,\hat{Y}_i^*} = .3906$$

∇

```
## Null model
> beta.mod1 <- betareg(FPrzp ~ 1)
> summary(beta.mod1)
Coefficients (mean model with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10959    0.01538  -7.128 1.02e-12 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   62.548      5.371   11.65   <2e-16 ***

Type of estimator: ML (maximum likelihood)
Log-likelihood: 360.9 on 2 Df
Number of iterations: 13 (BFGS) + 2 (Fisher scoring)
> (X2.P1 <- sum(resid(beta.mod1,type="pearson")^2))
[1] 265.6367
> (X2.D1 <- sum(resid(beta.mod1,type="deviance")^2))
[1] 264.766

## Full model (Year is a factor variable, with 1992 as reference)
> beta.mod2 <- betareg(FPrzp ~ FDrvp + TrkLng + Bank + Laps + Year)
> summary(beta.mod2)

Call:
betareg(formula = FPrzp ~ FDrvp + TrkLng + Bank + Laps + Year)

Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-3.7996 -0.6602  0.0031  0.6532  5.2351

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.7312457  0.2074226  -3.525 0.000423 ***
```

```
FDrvp         2.5440987  0.4128540   6.162 7.17e-10 ***
TrkLng       -0.1106140  0.0405332  -2.729 0.006353 **
Bank         -0.0019614  0.0015161  -1.294 0.195767
Laps         -0.0007601  0.0002544  -2.988 0.002808 **
Year1993     -0.2225534  0.0817810  -2.721 0.006502 **
Year1994     -0.0441460  0.0852535  -0.518 0.604584
Year1995     -0.1924006  0.0844577  -2.278 0.022722 *
Year1996     -0.2031800  0.0785715  -2.586 0.009712 **
Year1997     -0.1441680  0.0571996  -2.520 0.011721 *
Year1998     -0.1585144  0.0550342  -2.880 0.003973 **
Year1999     -0.1892330  0.0602789  -3.139 0.001694 **
Year2000     -0.1904757  0.0571511  -3.333 0.000860 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  102.877      8.861   11.61   <2e-16 ***

Type of estimator: ML (maximum likelihood)
Log-likelihood: 427.4 on 14 Df
Pseudo R-squared: 0.3906
Number of iterations: 23 (BFGS) + 2 (Fisher scoring)
> (X2.P2 <- sum(resid(beta.mod2,type="pearson")ˆ2))
[1] 264.143
> (X2.D2 <- sum(resid(beta.mod2,type="deviance")ˆ2))
[1] 265.6407
```

5.4 R Programs for Chapter 5 Examples

5.4.1 Maya Moore Free Throw Shooting

```
mm_ft <- c(rep(1,160),rep(0,21))

table(mm_ft)

mm.mod1 <- glm(mm_ft ˜ 1,binomial("logit"))
summary(mm.mod1)
```

5.4.2 English Premier League Football Total Goals per Game - 2013/14 Season

```
goals <- c(rep(0,27),rep(1,75),rep(2,82),rep(3,70),rep(4,63),
rep(5,39),rep(6,17),rep(7,4),rep(8,1),rep(9,2))

library(MASS)
mod1 <- glm(goals˜1,family="poisson")
summary(mod1)

mod2 <- glm.nb(goals˜1)
summary(mod2)

X2.poi.nb <- -2*(logLik(mod1)-logLik(mod2))
X2.05 <- qchisq(.95,1)
X2.pval <- 1-pchisq(X2.poi.nb,1)

print(round(cbind(X2.poi.nb,X2.05,X2.pval),3))
```

5.4.3 Example 5.4: Running Speeds Among Females at a Marathon

```
goals <- c(rep(0,27),rep(1,75),rep(2,82),rep(3,70),rep(4,63),
rep(5,39),rep(6,17),rep(7,4),rep(8,1),rep(9,2))

library(MASS)
mod1 <- glm(goals~1,family="poisson")
summary(mod1)

mod2 <- glm.nb(goals~1)
summary(mod2)

X2.poi.nb <- -2*(logLik(mod1)-logLik(mod2))
X2.05 <- qchisq(.95,1)
X2.pval <- 1-pchisq(X2.poi.nb,1)

print(round(cbind(X2.poi.nb,X2.05,X2.pval),3))
```

5.4.4 NBA Team/Game Free Throw Proportions

```
nbateam <- read.csv("http://www.stat.ufl.edu/~winner/data/nba_teamgame_20167.csv")
attach(nbateam); names(nbateam)
table(GameType)
Ftm.r <- Ftm[GameType == 1]
Fta.r <- Fta[GameType == 1]
Ftprop.r <- (Ftm.r + 2) / (Fta.r + 4)
N <- length(Ftprop.r)
mean(Ftprop.r); var(Ftprop.r)
hist(Ftprop.r, breaks=30)
summary(Ftprop.r)

library(betareg)
FT.mod1 <- betareg(Ftprop.r ~ 1)
summary(FT.mod1)
(gamma <- coef(FT.mod1)[1])
(phi <- coef(FT.mod1)[2])
(mu <- exp(gamma)/(1+exp(gamma)))
(alpha <- mu * phi)
(beta <- (1-mu) * phi)
(sigma2 <- (alpha*beta) / ((phi^2)*(phi+1)))

mean(Ftprop.r); var(Ftprop.r)
f.xft <- seq(0.2,1.0,0.02)
fb.yft <- dbeta(f.xft, alpha, beta)

hist(Ftprop.r, xlim=c(0.2,1.0), breaks=seq(0.2,1.0,0.02),
xlab="Wilson-Agresti-Coull Proportion",main="")
lines(f.xft,fb.yft*N*0.02)
```

5.4.5 Motorcycles and Erectile Dysfunction

```
age <- c(25,35,45,55,25,35,45,55)
mr  <- c(0,0,0,0,1,1,1,1)
ed.0 <- c(77,205,131,73,25,32,14,2)
ed.1 <- c(35,101,77,53,35,55,44,27)
```

```
# Note that y is made up of (number of Ss, # of Fs)
y <- cbind(ed.1,ed.0)
print(cbind(age,mr,y))

mod0 <- glm(y~1, family=binomial("logit"))
summary(mod0)
logLik(mod0)
mod1 <- glm(y~age, family=binomial("logit"))
summary(mod1)
logLik(mod1)
anova(mod0,mod1,test="Chisq")
mod2 <- glm(y~age+mr, family=binomial("logit"))
summary(mod2)
logLik(mod2)
anova(mod1,mod2,test="Chisq")
mod3 <- glm(y~age+mr+age:mr, family=binomial("logit"))
summary(mod3)
logLik(mod3)
anova(mod2,mod3,test="Chisq")

age.f <- factor(age)
mr.f <- factor(mr)

## Saturated model
mod4 <- glm(y ~ age.f*mr.f, family=binomial("logit"))
summary(mod4)

l0 <- logLik(mod0)
l3 <- logLik(mod3)

R2.CS <- 1 - (exp(l0)/exp(l3))^(2/8)

pi_hat <- ed.1/(ed.0+ed.1)

plot(age,pi_hat,type="n",xlim=c(20,60), ylim=c(0,1),
 xlab="Age", ylab="Prob(ED)")
points(age[mr==0],pi_hat[mr==0],pch=12)
points(age[mr==1],pi_hat[mr==1],pch=16)
ageseq=seq(20,60,0.1)
logodds.mr1 <- -1.257466+0.016436*ageseq+0.013092+0.039497*ageseq
logodds.mr0 <- -1.257466+0.016436*ageseq
lines(ageseq,(exp(logodds.mr1)/(1+exp(logodds.mr1))))
lines(ageseq,(exp(logodds.mr0)/(1+exp(logodds.mr0))),lty=5)
legend(20,.95,c("MR=0","MR=1"),pch=c(12,16),lty=c(5,1))
```

5.4.6 NFL Field Goal Attempts - 2008 Regular Season

```
fga <- read.csv("http://www.stat.ufl.edu/~winner/data/nfl2008_fga.csv")
attach(fga); names(fga)

n <- length(GOOD)

fga.mod1 <- glm(GOOD ~ 1, binomial("logit"))
summary(fga.mod1)
logLik(fga.mod1)

fga.mod2 <- glm(GOOD ~ distance, binomial("logit"))
summary(fga.mod2)
logLik(fga.mod2)

fga.mod3 <- glm(GOOD ~ distance + homekick, binomial("logit"))
```

```
summary(fga.mod3)
logLik(fga.mod3)

pihat <- predict(fga.mod2,type="response")
pi_good <- data.frame(pihat,GOOD,distance)
write.csv(pi_good, file="fga_fit.csv")

########
mindist <- min(distance); maxdist <- max(distance)
n.dist <- numeric(maxdist-mindist+1)
y.dist <- numeric(maxdist-mindist+1)
pi_hat.dist <- numeric(maxdist-mindist+1)
fga.dist <- numeric(maxdist-mindist+1)
cnt.dist <- 0

for (i in mindist:maxdist) {
cnt.dist <- cnt.dist+1
n.dist[cnt.dist] <- length(distance[distance==i])
y.dist[cnt.dist] <- sum(GOOD[distance==i])
fga.dist[cnt.dist] <- i
if (n.dist[cnt.dist] == 0) pi_hat.dist[cnt.dist] <- NA
else pi_hat.dist[cnt.dist] <- y.dist[cnt.dist] / n.dist[cnt.dist]
}

pi_hat.dist
fga.dist

x.seq <- seq(18,80,0.1)
pi_hat.seq <- predict(fga.mod2, list(distance=x.seq), type="response")
plot(pi_hat.dist ~ fga.dist, xlab="Distance (yards)", ylab="P(Success)")
lines(x.seq,pi_hat.seq)
```

5.4.7 Toxicity of Chemicals on Beetles

```
tribol <- read.table("http://www.stat.ufl.edu/~winner/data/tribol_tox.dat",
       header=F,col.names=c("pyreth","pipBut","numExps","numMor"))
attach(tribol)

y.trib <- cbind(numMor,numExps-numMor)

trib.mod1 <- glm(y.trib ~ 1, binomial("logit"))
summary(trib.mod1)
logLik(trib.mod1)
e.p1 <- resid(trib.mod1,type="pearson")
e.d1 <- resid(trib.mod1,type="deviance")
sum(e.p1^2)
sum(e.d1^2)

trib.mod2 <- glm(y.trib ~ pyreth + pipBut,
                    binomial("logit"))
summary(trib.mod2)
logLik(trib.mod2)
e.p2 <- resid(trib.mod2,type="pearson")
e.d2 <- resid(trib.mod2,type="deviance")
sum(e.p2^2)
sum(e.d2^2)

trib.mod3 <- glm(y.trib ~ pyreth + pipBut + I(pyreth*pipBut),
                    binomial("logit"))
summary(trib.mod3)
logLik(trib.mod3)
```

```
e.p3 <- resid(trib.mod3,type="pearson")
e.d3 <- resid(trib.mod3,type="deviance")
sum(e.p3^2)
sum(e.d3^2)

anova(trib.mod1,trib.mod2)
anova(trib.mod2,trib.mod3)

trib.mod4 <- glm(y.trib ~ pyreth + pipBut,
                    quasibinomial("logit"))
summary(trib.mod4)
logLik(trib.mod4)
e.p4 <- resid(trib.mod4,type="pearson")
e.d4 <- resid(trib.mod4,type="deviance")
sum(e.p4^2)
sum(e.d4^2)

trib.mod5 <- glm(y.trib ~ pyreth + pipBut + I(pyreth*pipBut) +
                     I(pyreth^2) + I(pipBut^2),
                    binomial("logit"))
summary(trib.mod5)
logLik(trib.mod5)
e.p5 <- resid(trib.mod5,type="pearson")
e.d5 <- resid(trib.mod5,type="deviance")
sum(e.p5^2)
sum(e.d5^2)

trib.mod6 <- glm(y.trib ~ pyreth + pipBut + I(pyreth*pipBut) +
                     I(pyreth^2) + I(pipBut^2),
                    quasibinomial("logit"))
summary(trib.mod6)
# logLik(trib.mod5)
e.p6 <- resid(trib.mod6,type="pearson")
e.d6 <- resid(trib.mod6,type="deviance")
sum(e.p6^2)
sum(e.d6^2)
```

5.4.8 NASCAR Crashes - 1972-1979 Seasons - Poisson Model

```
race1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/race7579.dat",
  width=c(8,8,8,8,8,8,8,8,12,40),
col.names=c('srace', 'yr', 'yrace', 'drivers', 'trklen', 'laps', 'roadtrk',
 'cautions', 'leadchng', 'trkid', 'track'))

race <- data.frame(drivers=race1$drivers, trklen=race1$trklen, laps=race1$laps,
 cautions=race1$cautions)
attach(race)

race.mod1 <- glm(cautions ~ 1, poisson("log"))
summary(race.mod1)
logLik(race.mod1)

race.mod2 <- glm(cautions ~ drivers + trklen + laps, poisson("log"))
summary(race.mod2)
logLik(race.mod2)
anova(race.mod2, test="Chisq")
drop1(race.mod2, test="Chisq")

race.mod3 <- glm(cautions ~ drivers + laps, poisson("log"))
summary(race.mod3)
logLik(race.mod3)
```

```
anova(race.mod3, test="Chisq")
drop1(race.mod3, test="Chisq")

muhat <- predict(race.mod3, type="response")

#print(cbind(cautions, muhat))
(pearson.x2 <- sum((cautions - muhat)^2/muhat))
(pearson.x2a <- sum(resid(race.mod3,type="pearson")^2))
(deviance.x2 <- sum(resid(race.mod3)^2))

mean.grp <- rep(0,length(cautions))
for (i in 1:length(cautions)) {
if (muhat[i] < 3.50) mean.grp[i] <- 1
else  if (muhat[i] < 3.70) mean.grp[i] <- 2
else  if (muhat[i] < 4.00) mean.grp[i] <- 3
else  if (muhat[i] < 4.15) mean.grp[i] <- 4
else  if (muhat[i] < 4.30) mean.grp[i] <- 5
else  if (muhat[i] < 4.40) mean.grp[i] <- 6
else  if (muhat[i] < 4.70) mean.grp[i] <- 7
else  if (muhat[i] < 5.25) mean.grp[i] <- 8
else  if (muhat[i] < 5.50) mean.grp[i] <- 9
else  if (muhat[i] < 6.00) mean.grp[i] <- 10
else  if (muhat[i] < 6.80) mean.grp[i] <- 11
else mean.grp[i] <- 12
}

count.mg <- rep(0,max(mean.grp))
sum.mg <- rep(0,max(mean.grp))
sum.muhat.mg <- rep(0,max(mean.grp))
mean.mg <- rep(0,max(mean.grp))
var.mg <- rep(0,max(mean.grp))

for (i in 1:max(mean.grp)) {
count.mg[i] <- length(cautions[mean.grp == i])
sum.mg[i] <- sum(cautions[mean.grp == i])
sum.muhat.mg[i] <- sum(muhat[mean.grp == i])
mean.mg[i] <- mean(cautions[mean.grp == i])
var.mg[i] <- var(cautions[mean.grp == i])
}

pearson.r <- (sum.mg - sum.muhat.mg) / sqrt(sum.muhat.mg)
(pearson.X2.mg <- sum(pearson.r^2))
qchisq(.95,12-2-1)
(pval.mg <- 1-pchisq(pearson.X2.mg,12-2-1))

gof.grp <- cbind(count.mg,sum.mg,sum.muhat.mg,pearson.r,mean.mg,var.mg)
colnames(gof.grp) <- c("# Races","Total Obs", "Total Exp", "Pearson r",
  "Mean","Variance")
round(gof.grp,3)

### quasipoisson takes SE(beta)*sqrt(phi)
### phi = pearson.x2/df
race.mod3a <- glm(formula = cautions ~ drivers + laps,
 family=quasipoisson("log"))
summary(race.mod3a)
anova(race.mod3a, test="Chisq")
```

5.4.9 NASCAR Lead Changes - 1972-1979 Seasons - Negative Binomial Model

```
race1 <- read.fwf("http://www.stat.ufl.edu/~winner/data/race7579.dat",
  width=c(8,8,8,8,8,8,8,8,8,12,40),
col.names=c('srace', 'yr', 'yrace', 'drivers', 'trklen', 'laps', 'roadtrk',
 'cautions', 'leadchng', 'trkid', 'track'))

race <- data.frame(drivers=race1$drivers, trklen=race1$trklen, laps=race1$laps,
 leadchng=race1$leadchng)
attach(race)

race.mod1 <- glm(leadchng ~ 1, poisson("log"))
summary(race.mod1)
logLik(race.mod1)

race.mod2 <- glm(leadchng ~ drivers + trklen + laps, poisson("log"))
summary(race.mod2)
logLik(race.mod2)
anova(race.mod2, test="Chisq")
drop1(race.mod2, test="Chisq")

race.mod3 <- glm(leadchng ~ drivers + laps, poisson("log"))
summary(race.mod3)
logLik(race.mod3)
anova(race.mod3, test="Chisq")
drop1(race.mod3, test="Chisq")

muhat <- predict(race.mod2, type="response")

#print(cbind(leadchng, muhat))
(pearson.x2 <- sum((leadchng - muhat)^2/muhat))
(pearson.x2a <- sum(resid(race.mod2,type="pearson")^2))
(deviance.x2 <- sum(resid(race.mod2)^2))

mean.grp <- rep(0,length(leadchng))
for (i in 1:length(leadchng)) {
if (muhat[i] < 9.4) mean.grp[i] <- 1
else  if (muhat[i] < 10.5) mean.grp[i] <- 2
else  if (muhat[i] < 11.6) mean.grp[i] <- 3
else  if (muhat[i] < 20) mean.grp[i] <- 4
else  if (muhat[i] < 21) mean.grp[i] <- 5
else  if (muhat[i] < 23) mean.grp[i] <- 6
else  if (muhat[i] < 26) mean.grp[i] <- 7
else  if (muhat[i] < 32) mean.grp[i] <- 8
else  if (muhat[i] < 36) mean.grp[i] <- 9
else mean.grp[i] <- 10
}

count.mg <- rep(0,max(mean.grp))
sum.mg <- rep(0,max(mean.grp))
sum.muhat.mg <- rep(0,max(mean.grp))
mean.mg <- rep(0,max(mean.grp))
var.mg <- rep(0,max(mean.grp))

for (i in 1:max(mean.grp)) {
count.mg[i] <- length(leadchng[mean.grp == i])
sum.mg[i] <- sum(leadchng[mean.grp == i])
sum.muhat.mg[i] <- sum(muhat[mean.grp == i])
mean.mg[i] <- mean(leadchng[mean.grp == i])
var.mg[i] <- var(leadchng[mean.grp == i])
}

gof.grp <- cbind(count.mg,sum.mg,sum.muhat.mg,pearson.r,mean.mg,var.mg)
colnames(gof.grp) <- c("# Races","Total Obs", "Total Exp", "Pearson r",
  "Mean","Variance")
round(gof.grp,3)

pearson.r <- (sum.mg - sum.muhat.mg) / sqrt(sum.muhat.mg)
```

```
(pearson.X2.mg <- sum(pearson.r^2))
qchisq(.95,10-3-1)
(pval.mg <- 1-pchisq(pearson.X2.mg,10-3-1))

### quasipoisson takes SE(beta)*sqrt(phi)
### phi = pearson.x2/df
race.mod2a <- glm(formula = leadchng ~ drivers + trklen + laps,
 family=quasipoisson("log"))
summary(race.mod2a)
anova(race.mod2a, test="Chisq")
drop1(race.mod2a, test="Chisq")

### Negative Binomial Model

library(MASS)
race.mod4 <- glm.nb(leadchng ~ drivers + trklen + laps, link=log)
summary(race.mod4)
logLik(race.mod4)
anova(race.mod4, test="Chisq")
drop1(race.mod4, test="Chisq")

race.mod5 <- glm.nb(leadchng ~ drivers +  laps, link=log)
summary(race.mod5)
logLik(race.mod5)
anova(race.mod5, test="Chisq")
drop1(race.mod5, test="Chisq")

anova(race.mod3, race.mod5, test="Chisq")

mean.grp <- rep(0,length(leadchng))
for (i in 1:length(leadchng)) {
if (muhat[i] < 9.4) mean.grp[i] <- 1
else  if (muhat[i] < 10.5) mean.grp[i] <- 2
else  if (muhat[i] < 11.6) mean.grp[i] <- 3
else  if (muhat[i] < 20) mean.grp[i] <- 4
else  if (muhat[i] < 21) mean.grp[i] <- 5
else  if (muhat[i] < 23) mean.grp[i] <- 6
else  if (muhat[i] < 26) mean.grp[i] <- 7
else  if (muhat[i] < 32) mean.grp[i] <- 8
else  if (muhat[i] < 36) mean.grp[i] <- 9
else mean.grp[i] <- 10
}

muhat.nb <- predict(race.mod4, type="response")
count.mg <- rep(0,max(mean.grp))
sum.mg <- rep(0,max(mean.grp))
sum.muhat.mg <- rep(0,max(mean.grp))
mean.mg <- rep(0,max(mean.grp))
var.mg <- rep(0,max(mean.grp))
var.mean.mg <- rep(0,max(mean.grp))
exp.v.m.mg <- rep(0,max(mean.grp))

for (i in 1:max(mean.grp)) {
count.mg[i] <- length(leadchng[mean.grp == i])
sum.mg[i] <- sum(leadchng[mean.grp == i])
sum.muhat.mg[i] <- sum(muhat.nb[mean.grp == i])
mean.mg[i] <- mean(leadchng[mean.grp == i])
var.mg[i] <- var(leadchng[mean.grp == i])
var.mean.mg[i] <- var.mg[i]/mean.mg[i]
exp.v.m.mg[i] <- 1 + mean.mg[i]*0.1905
}

gof.grp.nb <- cbind(count.mg,sum.mg,sum.muhat.mg,pearson.r,mean.mg,var.mg,
   var.mean.mg, exp.v.m.mg)
```

```
colnames(gof.grp.nb) <- c("# Races","Obs", "Exp", "Pearson r",
  "Mean","Var","Var/Mean","Exp V/M")
round(gof.grp.nb,3)

pearson.r <- (sum.mg - sum.muhat.mg) /
  sqrt(sum.muhat.mg + sum.muhat.mg^2*0.1905)
(pearson.X2.mg <- sum(pearson.r^2))
qchisq(.95,10-3-1)
(pval.mg <- 1-pchisq(pearson.X2.mg,10-3-1))
```

5.4.10 Napa Marathon Velocities - Gamma Model

```
napaf2015 <- read.csv("http://www.stat.ufl.edu/~winner/data/napa_marathon_fm2015.csv",
header=T)
attach(napaf2015); names(napaf2015)

gender <- factor(Gender)
summary(mph[gender=="F"])
summary(mph[gender=="M"])
length(mph[gender=="F"])
length(mph[gender=="M"])

napa.mod1 <- glm(mph~1,family=Gamma)
summary(napa.mod1)
deviance(napa.mod1)

napa.mod2 <- glm(mph~Age,family=Gamma)
summary(napa.mod2)

napa.mod3 <- glm(mph ~ Age, family=Gamma(link="log"))
summary(napa.mod3)

napa.mod4 <- glm(mph~gender,family=Gamma)
summary(napa.mod4)
deviance(napa.mod4)

napa.mod5 <- glm(mph~Age + gender,family=Gamma)
summary(napa.mod5)
logLik(napa.mod5)

napa.mod6 <- glm(mph ~ Age + gender, family=Gamma(link="log"))
summary(napa.mod6)
logLik(napa.mod6)

napa.mod7 <- glm(mph~Age*gender,family=Gamma)
summary(napa.mod7)
logLik(napa.mod7)

napa.mod8 <- glm(mph ~ Age*gender, family=Gamma(link="log"))
summary(napa.mod8)
logLik(napa.mod8)

age1 <- min(Age):max(Age)
genderf1m2 <- rep(1,length(mph))
for (i in 1:length(mph)) {
if (Gender[i] == "M") male[i] <- 2
}

par(mfrow=c(2,2))
yhat.F.inv <- 1/(coef(napa.mod5)[1] + coef(napa.mod5)[2]*age1)
yhat.M.inv <- 1/((coef(napa.mod5)[1]+coef(napa.mod5)[3]) +
 coef(napa.mod5)[2]*age1)

plot(Age[gender=="M"],mph[gender=="M"],xlab="Age",ylab="mph",
  main="Males - Inverse link",
  pch=1,cex=0.7,xlim=c(16,76),ylim=c(3,12))
lines(age1,yhat.M.inv,lty=1,lwd=2)

plot(Age[gender=="F"],mph[gender=="F"],xlab="Age",ylab="mph",
  main="Females - Inverse link",
  pch=1,cex=0.7,xlim=c(16,76),ylim=c(3,12))
lines(age1,yhat.F.inv,lty=1,lwd=2)

yhat.F.log <- exp(coef(napa.mod6)[1] + coef(napa.mod6)[2]*age1)
yhat.M.log <- exp((coef(napa.mod6)[1]+coef(napa.mod6)[3]) +
 coef(napa.mod6)[2]*age1)

plot(Age[gender=="M"],mph[gender=="M"],xlab="Age",ylab="mph",
  main="Males - Log link",
  pch=1,cex=0.7,xlim=c(16,76),ylim=c(3,12))
lines(age1,yhat.M.log,lty=1,lwd=2)

plot(Age[gender=="F"],mph[gender=="F"],xlab="Age",ylab="mph",
  main="Females - Log link",
  pch=1,cex=0.7,xlim=c(16,76),ylim=c(3,12))
lines(age1,yhat.F.log,lty=1,lwd=2)

anova(napa.mod5,napa.mod7,test="Chisq")

anova(napa.mod6,napa.mod8,test="Chisq")
mu8 <- predict(napa.mod8,type="response")
(dev.8 <- sum((log(mph/mu8))-((mph-mu8)/mu8)))
deviance(napa.mod8)
mu6 <- predict(napa.mod6,type="response")
(dev.6 <- sum((log(mph/mu6))-((mph-mu6)/mu6)))
```

5.4.11 Ford Prize Winnings in NASCAR Races: 1992-2000

```
ford <- read.csv("http://www.stat.ufl.edu/~winner/data/nas_ford_1992_2000a.csv",
header=T)
attach(ford); names(ford)

library(betareg)

Year <- factor(Year)
Track_id <- factor(Track_id)

beta.mod1 <- betareg(FPrzp ~ 1)
summary(beta.mod1)
(X2.P1 <- sum(resid(beta.mod1,type="pearson")^2))
(X2.D1 <- sum(resid(beta.mod1,type="deviance")^2))

beta.mod2 <- betareg(FPrzp ~ FDrvp + TrkLng + Bank + Laps + Year)
summary(beta.mod2)
(X2.P2 <- sum(resid(beta.mod2,type="pearson")^2))
(X2.D2 <- sum(resid(beta.mod2,type="deviance")^2))
```

Chapter 6

Nonlinear Regression

In many applications, theory leads to a nonlinear relationship between the response variable and a predictor variable or a set of predictors. These relationships can be based on growth models, differential equations, or simply observation. Note that these models go beyond polynomial models considered in Chapter 2, which are still linear with respect to the regression coefficients. Normal based models can be written in a general form as follows.

$$Y = g\left(X_1, \ldots, X_p; \beta_1, \ldots, \beta_q\right) + \epsilon \qquad eps \sim N\left(0, \sigma^2\right)$$

Nonlinear least squares can be used to estimate the parameters $\beta_1, \ldots, \beta_q$. This is done by calculus using matrix form of the models. Several examples are given here to illustrate nonlinear regression models. The usual (approximate) F-tests, t-tests, and Confidence Intervals can be used to make inference regarding parameters.

An important difference from previous models is that starting values must be given for parameters. Software packages are very good at solving for nonlinear least squares estimates as long as the signs and magnitudes are reasonable. It is especially important with models with exponential terms.

6.1 Examples of Nonlinear Regression Models

Example 6.1: Kentucky Derby Winning Times 1896-2016

It has been argued in the academic literature that there are limits to performance in animals (e.g. Denny, 2008, [10]). Denny studied historical results involving speed among horses, dogs, and humans with a wide variety of theoretically based nonlinear models relating performance to year. One model considered for Velocity was an "S-shaped" logistic function, of the following form, where Y_t is the winning velocity (meters per second) in year t.

$$Y_t = \beta_1 + \left(\beta_2 - \beta_1\right)\left[\frac{\exp\left\{\beta_3\left(t - \beta_4\right)\right\}}{1 + \exp\left\{\beta_3\left(t - \beta_4\right)\right\}}\right] + \epsilon_t$$

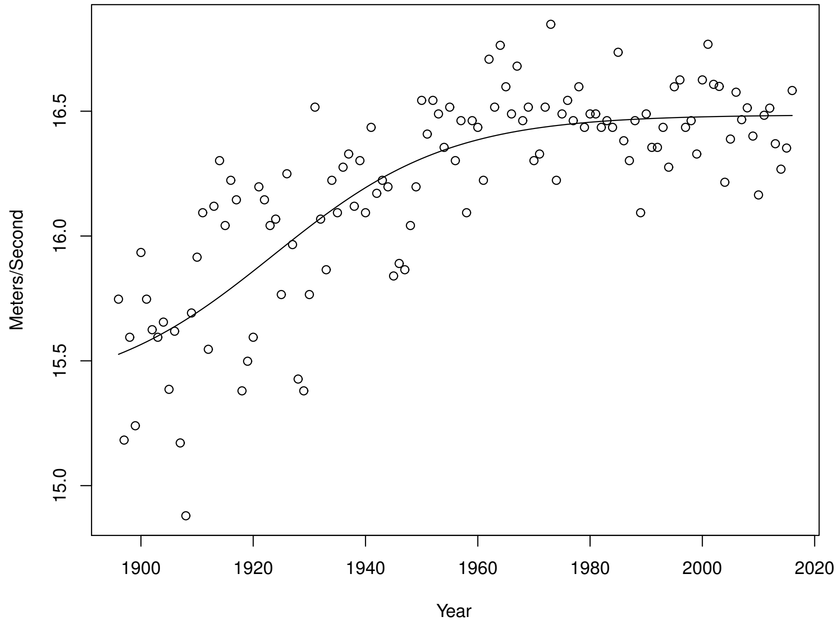

Figure 6.1: Kentucky Derby winning velocities and fitted logistic function - Years 1896-2016

In this model, β_1 is the lower asymptote (minimum mean speed), β_2 is the upper asymptote (maximum mean speed), β_3 is a shape parameter determining the steepness of the curve between lower and upper asymptotes. Finally, β_4 is the year when the curve is steepest, as well as half way between the lower and upper asymptotes. Here we consider the winning speeds of the Kentucky Derby for years 1896-2016, all years that the horse race was run at a distance of 1.25 miles. The variable t represents Year. The fitted equation below and is plotted with velocity data in Figure 6.1.

$$\hat{Y} = 15.36 + (16.49 - 15.36)\left[\frac{e^{.0638(t-1924)}}{1+e^{.0638(t-1924)}}\right] \qquad \sqrt{MSE} = 0.2376$$

In terms of the mean velocities have a lower asymptote of 15.36, and upper asymptote of 16.49, and the year where the mean is halfway between the two asymptotes being 1924. The R output is given below.

$$\nabla$$

```
> Speed125 <- 1609.34*Length125/Time125
> summary(Speed125)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.88   16.04   16.30   16.20   16.49   16.85
>
> kd.mod1 <- nls(Speed125 ~ b1 + (b2-b1)*exp(b3*(Year125-b4))/
+      (1+exp(b3*(Year125-b4))), start=c(b1=1,b2=20,b3=1,b4=1940))
> summary(kd.mod1)
Parameters:
    Estimate Std. Error t value Pr(>|t|)
b1 1.536e+01  2.823e-01  54.405  < 2e-16 ***
b2 1.649e+01  4.639e-02 355.370  < 2e-16 ***
b3 6.382e-02  2.297e-02   2.778  0.00637 **
b4 1.924e+03  9.271e+00 207.486  < 2e-16 ***

Residual standard error: 0.2376 on 117 degrees of freedom
> AIC(kd.mod1)
[1] 1.564131
```

Example 6.2: Beer Foam Heights Over Time for 3 Beer Brands

A study (Leike, 2002, [18]) reported results of an experiment measuring beer foam height over a 6 minute period for 3 brands of beer (Erdinger Weissbier, Augustinerbrau Munchen, and Budweiser). The data are given in Table 6.1. There are a total of $n = 3(15) = 45$ observations when "stacking" the data for 3 brands. An exponential decay model with additive errors is fit, allowing for different curves for the 3 brands, with t_i representing time, and dummy variables: $X_{i1} = 1$ if Erdinger, 0 otherwise; $X_{i2} = 1$ if Augustinerbrau, 0 otherwise; and $X_{i3} = 1$ if Budweiser, 0 otherwise. The fitted equations for the 3 brands are given below the model. Since the exponents have negative signs before $\beta_{11}, \beta_{12}, \beta_{13}$ these will have to be given positive and small starting values.

$$Y_i = \beta_{01}X_{i1}\exp\{-\beta_{11}X_{i1}t_i\} + \beta_{02}X_{i2}\exp\{-\beta_{12}X_{i2}t_i\} + \beta_{03}X_{i3}\exp\{-\beta_{13}X_{i3}t_i\} + \epsilon_i \qquad i = 1, \ldots, 45$$

$$\text{Erdinger: } \hat{Y} = 16.50e^{-.00396t} \qquad \text{August: } \hat{Y} = 13.23e^{-.006758t} \qquad \text{Bud: } \hat{Y} = 13.37e^{-.005625t}$$

The R program and output are given below. Note that the algorithm fails when $t_i = 0$, so replace it with $t_i = 0.0001$. A plot of the data and the fitted curves are given in Figure 6.2. Based on the plot and the coefficients given above, it appears that Augustinerbrau and Budweiser have similar curves. A simpler model would combine the dummy variables for brands 2 and 3 into a single dummy variable by adding them and fitting the following model with $X_{i23} = X_{i2} + X_{i3}$. The fitted equation is given below.

$$Y_i = \beta_{01}X_{i1}\exp\{-\beta_{11}X_{i1}t_i\} + \beta_{023}X_{i23}\exp\{-\beta_{123}X_{i22}t_i\} + \epsilon_i$$

$$\text{Erdinger: } \hat{Y} = 16.50e^{-.00396t} \qquad \text{August/Bud: } \hat{Y} = 13.29e^{-.006148t} \qquad SSE_F = 6.0277 \quad SSE_R = 10.3060$$

The plot in Figure 6.3 gives the data and fitted curves for the Reduced model. It appears that the curve tends to "miss low" typically for one of the two brands and "miss high" for the other.

A Complete versus Reduced model is used to test between the two models. For the Full model, the error degrees of freedom is $df_F = 45 - 6 = 39$ and for the Reduced model, it is $df_R = 45 - 4 = 41$. The (approximate) test is given below. The hypothesis of common curves for Augustinerbrau and Budweiser is rejected.

$$H_0 : \beta_{02} = \beta_{03}, \beta_{12} = \beta_{13} \qquad TS : F_{obs} = \frac{\left[\frac{10.3060-6.0277}{41-39}\right]}{\frac{6.0277}{39}} = \frac{2.1392}{0.1546} = 13.84$$

$$RR : F_{obs} \geq F_{.05,2,39} = 3.238 \qquad P < .0001$$

∇

Time (sec)	Erdinger	Augustinerbrau	Budweiser
0	17.0	14.0	14.0
15	16.1	11.8	12.1
30	14.9	10.5	10.9
45	14.0	9.3	10.0
60	13.2	8.5	9.3
75	12.5	7.7	8.6
90	11.9	7.1	8.0
105	11.2	6.5	7.5
120	10.7	6.0	7.0
150	9.7	5.3	6.2
180	8.9	4.4	5.5
210	8.3	3.5	4.5
240	7.5	2.9	3.5
300	6.3	1.3	2.0
360	5.2	0.7	0.9

Table 6.1: Beer Foam Heights for 3 Brands of Beer over Time

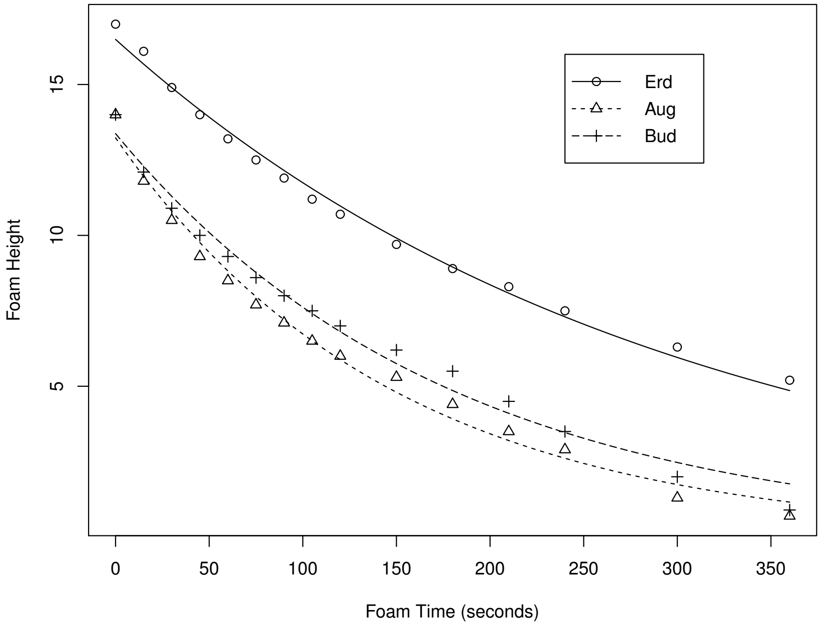

Figure 6.2: Beer Foam Height versus Time for 3 beer brands - Exponential Decay Model.

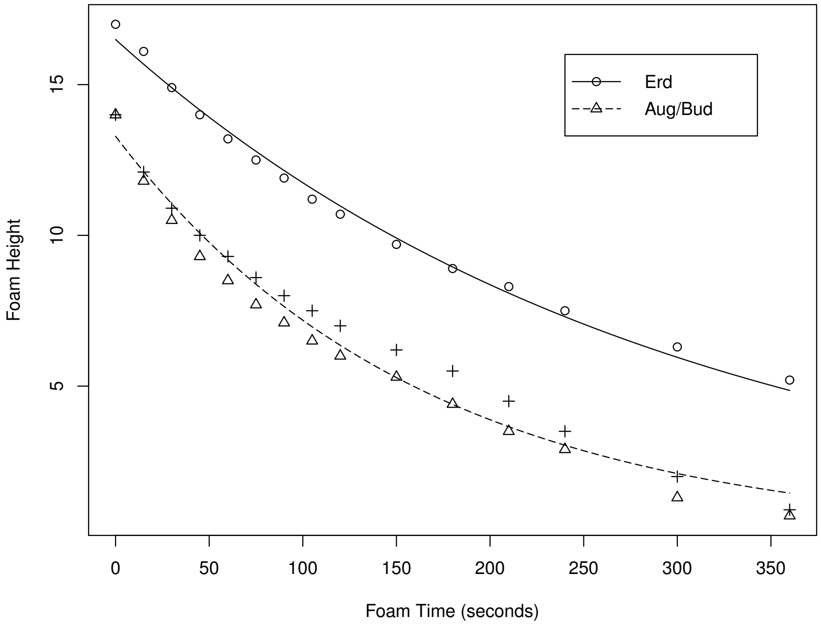

Figure 6.3: Beer Foam Height versus Time, with Brands 2 and 3 combined into a single brand

```
## Full Model
> foam.mod1 <- nls(foamHt ~ b01*brand1*exp(-b11*brand1*foamTime) +
+    b02*brand2*exp(-b12*brand2*foamTime) +
+    b03*brand3*exp(-b13*brand3*foamTime),
+    start=c(b01=10,b02=10,b03=10,b11=0.01,b12=0.01,b13=0.01))
> summary(foam.mod1)
Formula: foamHt ~ b01 * brand1 * exp(-b11 * brand1 * foamTime) + b02 *
    brand2 * exp(-b12 * brand2 * foamTime) + b03 * brand3 * exp(-b13 *
    brand3 * foamTime)
Parameters:
     Estimate Std. Error t value Pr(>|t|)
b01 1.650e+01  2.080e-01   79.32   <2e-16 ***
b02 1.323e+01  2.469e-01   53.61   <2e-16 ***
b03 1.337e+01  2.346e-01   57.02   <2e-16 ***
b11 3.396e-03  1.172e-04   28.98   <2e-16 ***
b12 6.758e-03  2.534e-04   26.67   <2e-16 ***
b13 5.625e-03  2.117e-04   26.57   <2e-16 ***

Residual standard error: 0.3931 on 39 degrees of freedom
> deviance(foam.mod1)
[1] 6.02771

## Reduced Model
> foam.mod2 <- nls(foamHt ~ b01*brand1*exp(-b11*brand1*foamTime) +
+    b023*brand23*exp(-b123*brand23*foamTime),
+    start=c(b01=10,b023=10,b11=0.01,b123=0.01))
> summary(foam.mod2)
Formula: foamHt ~ b01 * brand1 * exp(-b11 * brand1 * foamTime) + b023 *
    brand23 * exp(-b123 * brand23 * foamTime)
Parameters:
      Estimate Std. Error t value Pr(>|t|)
```

```
b01  1.650e+01  2.652e-01   62.20   <2e-16 ***
b023 1.329e+01  2.167e-01   61.30   <2e-16 ***
b11  3.396e-03  1.494e-04   22.73   <2e-16 ***
b123 6.148e-03  2.082e-04   29.53   <2e-16 ***

Residual standard error: 0.5014 on 41 degrees of freedom
> deviance(foam.mod2)
[1] 10.30604

> anova(foam.mod2,foam.mod1)
Analysis of Variance Table
Model 1: foamHt ~ b01 * brand1 * exp(-b11 * brand1 * foamTime) + b023 * brand23 * exp(-b123 * brand23 * foamTime)
Model 2: foamHt ~ b01 * brand1 * exp(-b11 * brand1 * foamTime) + b02 * brand2 * exp(-b12 * brand2 * foamTime) +
                  b03 * brand3 * exp(-b13 * brand3 * foamTime)
  Res.Df Res.Sum Sq Df Sum Sq F value    Pr(>F)
1     41    10.3060
2     39     6.0277  2 4.2783  13.841 2.869e-05 ***
```

6.2 R Programs for Chapter 6 Examples

6.2.1 Kentucky Derby Winning Times 1896-2016

```
kd <- read.csv("http://www.stat.ufl.edu/~winner/data/kentuckyderby.csv",
              header=TRUE)
attach(kd); names(kd)

Year125 <- Year[Length==1.25]
Time125 <- Time[Length==1.25]
Length125 <- Length[Length==1.25]
Year125.0 <- Year125-min(Year125)

Speed125 <- 1609.34*Length125/Time125
summary(Speed125)

kd.mod1 <- nls(Speed125 ~ b1 + (b2-b1)*exp(b3*(Year125-b4))/
    (1+exp(b3*(Year125-b4))), start=c(b1=1,b2=20,b3=1,b4=1940))
summary(kd.mod1)
AIC(kd.mod1)

plot(Year125,Speed125, xlab="Year", ylab="Meters/Second")
lines(Year125,predict(kd.mod1,Year125))

plot(Year125,resid(kd.mod1))
qqnorm(resid(kd.mod1)); qqline(resid(kd.mod1))
```

6.2.2 Beer Foam Heights Over Time for 3 Beer Brands

```
beerfoam <- read.csv("http://www.stat.ufl.edu/~winner/data/beerfoam2a.csv")
attach(beerfoam); names(beerfoam)

for (i in 1:length(foamTime)) {
if (foamTime[i] == 0) foamTime[i] <- 0.0001
}
```

```
foam.mod1 <- nls(foamHt ~ b01*brand1*exp(-b11*brand1*foamTime) +
   b02*brand2*exp(-b12*brand2*foamTime) +
   b03*brand3*exp(-b13*brand3*foamTime),
   start=c(b01=10,b02=10,b03=10,b11=0.01,b12=0.01,b13=0.01))
summary(foam.mod1)
deviance(foam.mod1)

time.x <- 0:360
yhat.b1 <- coef(foam.mod1)[1] * exp(-coef(foam.mod1)[4]*time.x)
yhat.b2 <- coef(foam.mod1)[2] * exp(-coef(foam.mod1)[5]*time.x)
yhat.b3 <- coef(foam.mod1)[3] * exp(-coef(foam.mod1)[6]*time.x)

plot(foamTime,foamHt,pch=beerBrnd, xlab="Foam Time (seconds)",
    ylab="Foam Height")
lines(time.x,yhat.b1,lty=1)
lines(time.x,yhat.b2,lty=2)
lines(time.x,yhat.b3,lty=5)
legend(240,16,c("Erd","Aug","Bud"),pch=c(1,2,3),lty=c(1,2,5))

### Reduced Model
brand23=brand2 + brand3

foam.mod2 <- nls(foamHt ~ b01*brand1*exp(-b11*brand1*foamTime) +
   b023*brand23*exp(-b123*brand23*foamTime),
   start=c(b01=10,b023=10,b11=0.01,b123=0.01))
summary(foam.mod2)
deviance(foam.mod2)

time.x <- 0:360
yhat.b1 <- coef(foam.mod2)[1] * exp(-coef(foam.mod2)[3]*time.x)
yhat.b23 <- coef(foam.mod2)[2] * exp(-coef(foam.mod2)[4]*time.x)

plot(foamTime,foamHt,pch=beerBrnd, xlab="Foam Time (seconds)",
    ylab="Foam Height")
lines(time.x,yhat.b1,lty=1)
lines(time.x,yhat.b23,lty=5)
legend(240,16,c("Erd","Aug/Bud"),pch=c(1,2),lty=c(1,5))

anova(foam.mod2,foam.mod1)
```

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis. 2nd Ed.* Wiley, New York.

[2] Agresti, A.(2007). *An Introduction to Categorical Analysis, 2nd Ed.* Wiley, New York.

[3] Agresti, A.(1996). *An Introduction to Categorical Analysis* Wiley, New York.

[4] Anyasi, T.A., A.I.O. Jideani, and G.R.A. Mchau (2015). "Effect of Organic Acid Pretreatment on Some Physical, Functional, and Antioxidant Properties of Flour Obtained from Three Unripe Banana Cultivars," *Food Chemistry*, Vol. 172, pp. 515-522.

[5] Armstrong, R.A. (2013). "Statistical Guidelines for the Analysis of Data Obtained from One or Both Eyes," *Opthalmic & Physiological Optics*, Vol. 33, pp. 7-14.

[6] Bagiella, E., R.P. Sloan, and D.F. Heitjan (2000). "Mixed Effects Models in Psychophysiology," *Psychophysiology*, Vol. 37, pp. 13-20.

[7] Baird, J., R. Curry, and T. Reid (2013). "Development and Application of a Multiple Linear Regression Model to Consider the Impact of Weekly Waste Container Capacity on the Yield from Kerbside Recycling Programmes in Scotland," *Waste Management & Research*, Vol. 31, #3, pp. 306-314.

[8] Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications.* Cambridge, Cambridge.

[9] Cook, S.R. and G. Proulx (1989). "Mechanical Evaluation and Performance Improvement of the Rotating Jaw Conibear 120 Trap," *Journal of Testing and Evaluation*, Vol. 17, # 3, pp. 190-195.

[10] Denny, M.W. (2008). Limits to Running Speeds in Dogs, Horses, and Humans, *The Journal of Experimental Biology,* Vol. 211, pp. 3836-3849.

[11] Eubanks, D.L., S.T. Murphy, and M.D. Mumford (2010). "Intuition as an Influence on Creative Problem-Solving: The Effects of Intuition, Positive Affect and Training," *Creative Research Journal*, Vol. 22, #2, pp. 170-184.

[12] Faraway, J.J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman & Hall/CRC, Boca Raton, FL.

[13] Ferrari, S.L.P. and F. Cribari-Neto (2004). "Beta Regression for Modeling Rates and Proportions," *Journal of Applied Statistics,* Vol. 31, #7, pp. 799-815.

[14] Freeman, D., A. Jaeger, R. Johnson, S. Geletta, K. Cooper, and P. Toney (2007). "Reliability Study of the Phillips Biometer for the Measurement of Subtalar Joint Range of Motion," *The Foot*, Vol. 17, pp. 102-110.

[15] Hewlett, P.S. (1969). "The Toxicity to Tribolium Castaneum (Herbst) (Coleoptora, Tenebrionidae) of Mixtures of Pyrethrins and Piperonyl Butoxide: Fitting a Mathematical Model," Journal of Stored Products Research, Vol. 5, Issue 1, pp 1-9.

[16] Jensen, C.R. (2002). "Variance Component Calculations: Common Methods and Misapplications in the Semiconductor Industry," *Quality Engineering*, Vol. 14, #4, pp. 647-657.

[17] Landy, D. and H. Sigall (1974). "Beauty is Talent: Task Evaluation as a Function of the Performer's Physical Attraction," *Journal of Personality and Social Psychology*, Vol. 29, pp. 299-304.

[18] Leike, A. (2002). Demonstration of the Exponential Decay Law Using Beer Froth, *European Journal of Physics*, Vol. 23, Number 1, pp. 21-26.

[19] Li, M-H. C. and A. Al-Refaie (2008). "Improving Wooden Parts' Quality by Adopting DMAIC Procedure," *Quality and Reliability Engineering International*, Vol. 24, pp. 351-360.

[20] Littell, R.C., J. Pendergast, and R. Natarajan (2000). "Modelling Covariance Structure in the Analysis of Repeated Measures Data," *Statistics in Medicine*, Vol. 19, pp. 1793-1819.

[21] MacPhee, N., A. Savage, N. Noton, E. Beattie, L. Milne, and J. Fraser (2018). "A Comparison of Penetration and Damage Caused by Different Types of Arrowheads on Loose and Tight Fit Clothing," *Science & Justice*, Vol. 58, pp. 109-120.

[22] Ma'Or, Z., S. Yehuda, and W. Voss (1997). "Skin Smoothing Effects of Dead Sea Minerals: Comparative Profilometric Evaluation of Skin Surface," *International Journal of Cosmetic Science*, Vol. 19, pp. 105-110.

[23] McNab, W. and E. E. Ristori (1899-1900). "Researches on Modern Explosives, Second Communication," *Proceedings of the Royal Society of London*, Vol. 66, pp. 221-232.

[24] Moynihan, A.C., S. Govindasamy-Lucey, J.J. Jaeggi, M.E. Johnson, J.A. Lucey, and P.L.H. McSweeney (2014). "Effect of Camel Chynosin on the Texture, Functionality, and Sensory Properties of Low-Moisture, Part-Skim Mozzarella Cheese," *Journal of Dairy Science*, Vol. 97, #1, pp. 85-96.

[25] Nyh, S. (2002). Traffic Deaths and Superstition on Friday the 13th, *American Journal of Psychiatry,* Vol. 159, #12, pp. 2110-2111.

[26] Ochiai, A., Y. Naya, J. Soh, Y. Ishida, S. Ushijima, Y. Mizutani, A. Kawauchi, and T. Miki (2006). "Do Motorcyclists Have Erectile Dysfunction? A Preliminary Study," *International Journal of Impotence Research*, Vol. 18, #4, pp. 396-399.

[27] Oliveira, A.S., F.M. Dalla Nora, R.O. Mello, P.A. Mello, B. Tischer, A.B. Costa, and J.S. Barin (2017). "One-Shot, Reagent-Free Determination of the Alcohol Content of Distilled Beverages by Thermal Infrared Enthalpimetry," *Talanta*, Vol. 171, pp. 335-340.

[28] Piccinini, P. , M. Piecha, and S.F. Torrent (2013). "European Survey of the Content in Lead in Lip Products," *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 76, pp. 225-233.

[29] Pinheiro, J.C. and D.M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer Verlag, New York.

[30] Rowe, W.F. and S.R. Hanson (1985). "Range-of-Fire Estimates from Regression Analysis Applied to the Spreads of Shotgun Pellets Patterns: Results of a Blind Study," *Forensic Science International*, Vol. 28, pp. 239-250.

[31] Rubolini, D., F. Spina, and N. Saino (2005). "Correlates of Timing of Spring Migration in Birds: a Comparative Study of Trans-Saharan Migrants," *Biological Journal of the Linnean Society*, Vol. 85, pp. 199-210.

[32] Steinsholt, K. (1998). "Are Assessors Levels of a Split-Plot Factor in The Analysis of Variance of Sensory Profile Experiments?," *Food Quality and Preference*, Vol. 9, #3, pp. 153-156.

[33] Shih, S.F. and W.F.P. Shih (1978). "Use of Dummy Variables in Water Resources Studies," *Journal of Hydrology*, Vol. 38, pp. 289-298.

[34] Singley, K.I., B.D. Hale, and D.M. Russell (2012). "Heart Rate, Anxiety and Hardiness in Novice (Tandem) and Experienced (Solo) Skydivers," *Journal of Sport Behavior*, Vol. 35, #4, pp. 453-469.

[35] Stephens, J.C. and L. Johnson (1951). "Subsidence of Organic Soils in the Upper Everglades Region of Florida," *Soil Science Society of Florida Proceedings*, Vol. 11, pp. 191-237.

[36] Winner, L. (2006). NASCAR Winston Cup Race Results for 1975-2003, *Journal of Statistics Education,* Vol.14,#3, www.amstat.org/publications/jse/v14n3/datasets.winner.html

[37] Wu, S-P. and C-S. Hsieh (2002). "Ergonomics Study on the Handle Length and Lift Angle for the Culinary Spatula," *Applied Ergonomics*, Vol. 33, pp. 493-501.

[38] Yang, R., D. Gu, and Z. Gu (2016). "Cordyceps Rice Wine: A Novel Brewing Process," *Journal of Food Process Engineering*, Vol. 39, pp. 581-590.