1 Simple Linear Regression

Text: RPD, Chapter 1
Problems:

1.1 Statistical Model

In simple linear regression, the model contains a random **dependent** (or **response** or **outcome** or **end point**) variable Y, that is hypothesized to be associated with an **independent** (or **predictor** or **explanatory**) variable X. The simple linear regression model specifies that the mean, or expected value of Y is a linear function of the level of X. Further, X is presumed to be set by the experimenter (as in controlled experiments) or known in advance to the activity generating the response Y. The experiment consists of obtaining a sample of n pairs (X_i, Y_i) from a population of such pairs (or nature). The model with respect to the mean is:

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

where β_0 is the mean of when when $X = 0$ (assuming this is a reasonable level of X), or more generally the Y–intercept of the regression line; β_1 is the change in the mean of Y as X increases by a single unit, or the slope of the regression line. Note that in practice β_0 and β_1 are unknown parameters that will be estimated from sample data.

Individual measurements are assumed to be independent, and normally distributed around the mean at their corresponding X level, with standard deviation σ. This can be stated as below:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim NID(0, \sigma^2)$$

where ε_i is a random error term and $NID(0, \sigma^2)$ means normally and independently distributed with mean 0, and variance σ^2.

1.1.1 Examples

The following two examples are based on applications of regression in pharmacodynamics and microeconomics.

Example 1 – Pharmacodynamics of LSD

The following data were published by J.G. Wagner, et al, in the 1968 article: "Correlation of Performance Test Scores with 'Tissue Concentration' of Lysergic Acid Diethylamide in Human Subjects," (*Clinical Pharmacology & Therapeutics*, 9:635–638).

Y — Mean score on math test (relative to control) for a group of five male volunteers.

X — Mean tissue concentration of LSD among the volunteers.

A sample of $n = 7$ points were selected, with X_i and Y_i being measured at each point in time. These 7 observations are treated as a sample from all possible realizations from this experiment. The parameter β_1 represents the systematic change in mean score as tissue concentration increases by one unit, and β_0 represents the true mean score when the concentration is 0. The data are given in Table 1.1.1.

i	X_i	Y_i
1	1.17	78.93
2	2.97	58.20
3	3.26	67.47
4	4.69	37.47
5	5.83	45.65
6	6.00	32.92
7	6.41	29.97

Table 1: LSD concentrations and math scores – Wagner, et al (1968)

.

Example 2 – Estimating Cost Functions of a Hosiery Mill

The following (approximate) data were published by Joel Dean, in the 1941 article: "Statistical Cost Functions of a Hosiery Mill," (*Studies in Business Administration*, vol. 14, no. 3).

Y — Monthly total production cost (in \$1000s).

X — Monthly output (in thousands of dozens produced).

A sample of $n = 48$ months of data were used, with X_i and Y_i being measured for each month. The parameter β_1 represents the change in mean cost per unit increase in output (unit variable cost), and β_0 represents the true mean cost when the output is 0, without shutting plant (fixed cost). The data are given in Table 1.1.1 (the order is arbitrary as the data are printed in table form, and were obtained from visual inspection/approximation of plot).

1.1.2 Generating Data from the Model

To generate data from the model using a computer program, use the following steps:

1. Specify the model parameters: β_0, β_1, σ
2. Specify the levels of X_i, $i = 1, \ldots, n$. This can be done easily with do loops or by brute force.
3. Obtain n standard normal errors $Z_i \sim N(0, 1)$, $i = 1, \ldots, n$. Statistical routines have them built in, or transformations of uniform random variates can be obtained.
4. Obtain random response $Y_i = \beta_0 + (\beta_1 X_i) + (\sigma Z_i)$, $i = 1, \ldots, n$.
5. For the case of random X_i, these steps are first completed for X_i in 2), then continued for Y_i. The Z_i used for X_i must be independent of that used for Y_i.

1.2 Least Squares Estimation

The parameters β_0 and β_1 can take on any values in the range (∞, ∞), and σ can take on any values in the range $[0, \infty)$ (if it takes on 0, then the model is deterministic, and not probabilistic). The most common choice of estimated regression equation (line in the case of simple linear regression), is to choose the line that minimizes the sum of squared vertical distances between the observed

i	X_i	Y_i	i	X_i	Y_i	i	X_i	Y_i
1	46.75	92.64	17	36.54	91.56	33	32.26	66.71
2	42.18	88.81	18	37.03	84.12	34	30.97	64.37
3	41.86	86.44	19	36.60	81.22	35	28.20	56.09
4	43.29	88.80	20	37.58	83.35	36	24.58	50.25
5	42.12	86.38	21	36.48	82.29	37	20.25	43.65
6	41.78	89.87	22	38.25	80.92	38	17.09	38.01
7	41.47	88.53	23	37.26	76.92	39	14.35	31.40
8	42.21	91.11	24	38.59	78.35	40	13.11	29.45
9	41.03	81.22	25	40.89	74.57	41	9.50	29.02
10	39.84	83.72	26	37.66	71.60	42	9.74	19.05
11	39.15	84.54	27	38.79	65.64	43	9.34	20.36
12	39.20	85.66	28	38.78	62.09	44	7.51	17.68
13	39.52	85.87	29	36.70	61.66	45	8.35	19.23
14	38.05	85.23	30	35.10	77.14	46	6.25	14.92
15	39.16	87.75	31	33.75	75.47	47	5.45	11.44
16	38.59	92.62	32	34.29	70.37	48	3.79	12.69

Table 2: Production costs and Output – Dean (1941)

.

responses (Y_i) and the fitted regression line ($\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$), where $\hat{\beta}_0$ and $\hat{\beta}_1$ are sample based estimates of β_0 and β_1, respectively.

Mathematically, we can label the error sum of squares as the sum of squared distances between the observed data and their mean values based on the model:

$$Q = \sum_{i=1}^{n}(Y_i - E(Y_i))^2 = \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2$$

The least squares estimates of β_0 and β_1 that minimize Q, which are obtained by taking derivatives, setting them equal to 0, and solving for $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = \sum_{i=1}^{n} Y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} X_i = 0 \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)X_i = \sum_{i=1}^{n} X_i Y_i - \beta_0 \sum_{i=1}^{n} X_i - \beta_1 \sum_{i=1}^{n} X_i^2 = 0 \qquad (2)$$

From equations (1) and (2) we obtain the so–called "normal equations":

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad (3)$$

$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad = \quad \sum_{i=1}^{n} X_i Y_i \tag{4}$$

Multipliying equation (3) by $\sum_{i=1}^{n} X_i$ and equation (4) by n, we obtain the following two equations:

$$n\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 (\sum_{i=1}^{n} X_i)^2 \quad = \quad (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i) \tag{5}$$

$$n\hat{\beta}_0 \sum_{i=1}^{n} X_i + n\hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \quad = \quad n\sum_{i=1}^{n} X_i Y_i \tag{6}$$

Subtracting equation (5) from (6), we get:

$$\hat{\beta}_1 (n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2) \quad = \quad n\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)$$

$$\Rightarrow \quad \hat{\beta}_1 \quad = \quad \frac{\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} \quad = \quad \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{7}$$

Now, from equation (1), we get:

$$n\hat{\beta}_0 \quad = \quad \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i \quad \Rightarrow \quad \hat{\beta}_0 \quad = \quad \overline{Y} - \hat{\beta}_1 \overline{X} \tag{8}$$

and the estimated (or fitted or prediction) equation ($\hat{Y}_i$):

$$\hat{Y}_i \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad i = 1, \ldots, n \tag{9}$$

The residuals are defined as the difference between the observed responses (Y_i) and their predicted values ($\hat{Y}_i$), where the residuals are denoted as e_i (they are estimates of ε_i):

$$e_i \quad = \quad Y_i - \hat{Y}_i \quad = \quad Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \qquad i = 1, \ldots, n \tag{10}$$

The residuals sum to 0 for this model:

$$\begin{aligned} e_i \quad &= \quad (Y_i - \hat{Y}_i) \quad = \quad Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\ &= \quad Y_i - \{[\overline{Y} - \hat{\beta}_1 \overline{X}] + \hat{\beta}_1 X_i]\} \end{aligned}$$

$$\Rightarrow \quad \sum_{i=1}^{n} e_i \quad = \quad \sum_{i=1}^{n} Y_i - \{n\overline{Y} - n\hat{\beta}_1 \overline{X} + \hat{\beta}_1 \sum_{i=1}^{n} X_i\} \quad = \quad \sum_{i=1}^{n} Y_i - \{\sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i\} \quad = \quad 0$$

1.2.1 Examples

Numerical results for the two examples desribed before are given below.

Example 1 – Pharmacodynamics of LSD

This dataset has $n = 7$ observations with a mean LSD tissue content of $\overline{X} = 4.3329$, and a mean math score of $\overline{Y} = 50.0871$.

$$\sum_{i=1}^{n} X_i = 30.33 \quad \sum_{i=1}^{n} X_i^2 = 153.8905 \quad \sum_{i=1}^{n} Y_i = 350.61 \quad \sum_{i=1}^{n} Y_i^2 = 19639.2365 \sum_{i=1}^{n} X_iY_i = 1316.6558$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_iY_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{1316.6558 - \frac{(30.33)(350.61)}{7}}{153.8905 - \frac{(30.33)^2}{7}} =$$

$$= \frac{1316.6558 - 1519.1430}{153.8905 - 131.4146} = \frac{-202.4872}{22.4759} = -9.0091$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 50.0871 - (-9.0091)(4.3329) = 89.1226$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 89.1226 - 9.0091X_i \qquad i = 1, \ldots, 7$$

$$e_i = Y_i - \hat{Y}_i = Y_i - (89.1226 - 9.0091X_i) \qquad i = 1, \ldots, 7$$

Table 1.2.1 gives the raw data, their fitted values, and residuals.

i	X_i	Y_i	$\hat{Y}_i$	e_i
1	1.17	78.93	78.5820	0.3480
2	2.97	58.20	62.3656	-4.1656
3	3.26	67.47	59.7529	7.7171
4	4.69	37.47	46.8699	-9.3999
5	5.83	45.65	36.5995	9.0505
6	6.00	32.92	35.0680	-2.1480
7	6.41	29.97	31.3743	-1.4043

Table 3: LSD concentrations, math scores, fitted values and residuals – Wagner, et al (1968)

.

A plot of the data and regression line are given in Figure 1.

Example 2 – Estimating Cost Function of a Hosiery Mill

This dataset has $n = 48$ observations with a mean output (in 1000s of dozens) of $\overline{X} = 31.0673$, and a mean monthly cost (in \$1000s) of $\overline{Y} = 65.4329$.

$$\sum_{i=1}^{n} X_i = 1491.23 \quad \sum_{i=1}^{n} X_i^2 = 54067.42 \quad \sum_{i=1}^{n} Y_i = 3140.78 \quad \sum_{i=1}^{n} Y_i^2 = 238424.46 \sum_{i=1}^{n} X_iY_i = 113095.80$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_iY_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{113095.80 - \frac{(1491.23)(3140.78)}{48}}{54067.42 - \frac{(1491.23)^2}{48}} =$$

$$= \frac{113095.80 - 97575.53}{54067.42 - 46328.48} = \frac{15520.27}{7738.94} = 2.0055$$

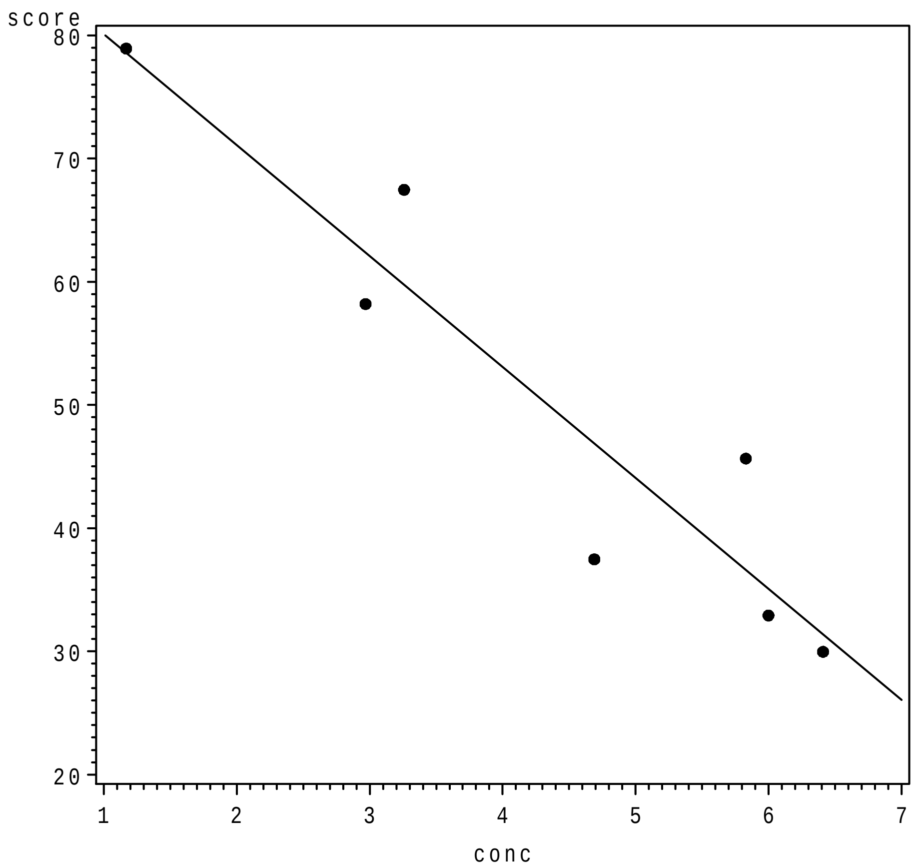

Figure 1: Regression of math score on LSD concentration (Wagner, et al, 1968)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} = 65.4329 - (2.0055)(31.0673) = 3.1274$$
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 3.1274 + 2.0055X_i \qquad i = 1, \ldots, 48$$
$$e_i = Y_i - \hat{Y}_i = Y_i - (3.1274 + 2.0055X_i) \qquad i = 1, \ldots, 48$$

Table 1.2.1 gives the raw data, their fitted values, and residuals.

A plot of the data and regression line are given in Figure 2.

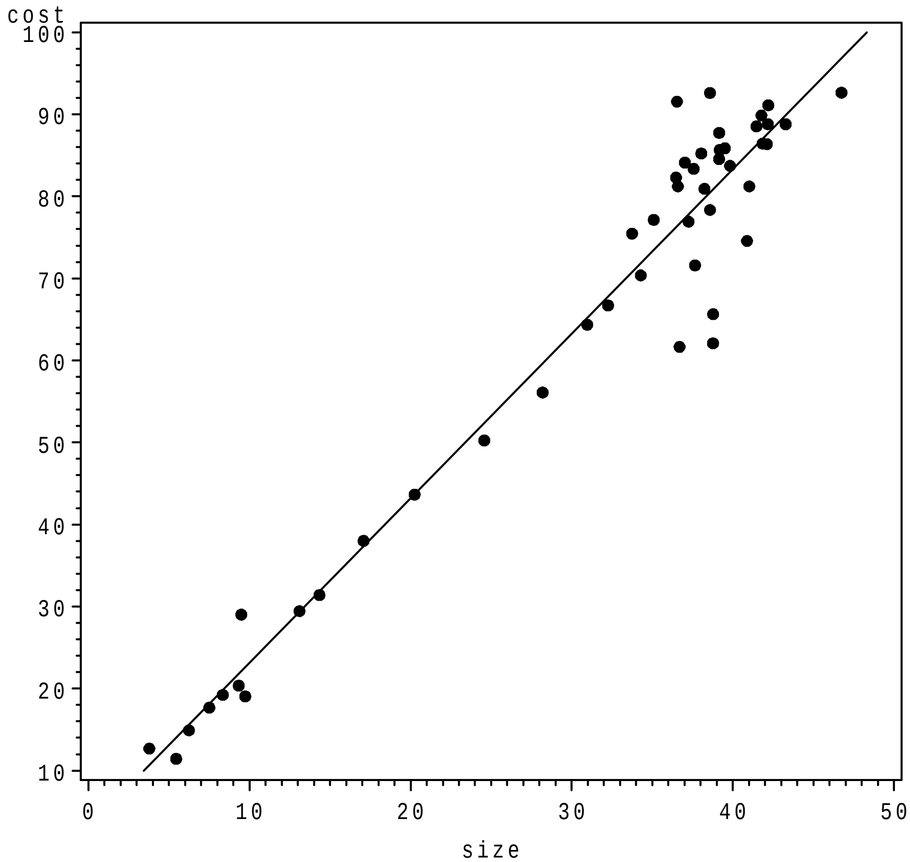

Figure 2: Estimated cost function for hosiery mill (Dean, 1941)

1.3 Analysis of Variance

The total variation in the response (Y) can be partitioned into parts that are attributable to various sources. The response Y_i can be written as follows:

$$Y_i = \hat{Y}_i + e_i \qquad i = 1, \ldots, n$$

We start with the total (uncorrected) sum of squares for Y:

$$SS(\text{TOTAL UNCORRECTED}) = \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (\hat{Y}_i + e_i)^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{Y}_i e_i$$

i	X_i	Y_i	$\hat{Y}_i$	e_i
1	46.75	92.64	96.88	-4.24
2	42.18	88.81	87.72	1.09
3	41.86	86.44	87.08	-0.64
4	43.29	88.80	89.95	-1.15
5	42.12	86.38	87.60	-1.22
6	41.78	89.87	86.92	2.95
7	41.47	88.53	86.30	2.23
8	42.21	91.11	87.78	3.33
9	41.03	81.22	85.41	-4.19
10	39.84	83.72	83.03	0.69
11	39.15	84.54	81.64	2.90
12	39.20	85.66	81.74	3.92
13	39.52	85.87	82.38	3.49
14	38.05	85.23	79.44	5.79
15	39.16	87.75	81.66	6.09
16	38.59	92.62	80.52	12.10
17	36.54	91.56	76.41	15.15
18	37.03	84.12	77.39	6.73
19	36.60	81.22	76.53	4.69
20	37.58	83.35	78.49	4.86
21	36.48	82.29	76.29	6.00
22	38.25	80.92	79.84	1.08
23	37.26	76.92	77.85	-0.93
24	38.59	78.35	80.52	-2.17
25	40.89	74.57	85.13	-10.56
26	37.66	71.60	78.65	-7.05
27	38.79	65.64	80.92	-15.28
28	38.78	62.09	80.90	-18.81
29	36.70	61.66	76.73	-15.07
30	35.10	77.14	73.52	3.62
31	33.75	75.47	70.81	4.66
32	34.29	70.37	71.90	-1.53
33	32.26	66.71	67.82	-1.11
34	30.97	64.37	65.24	-0.87
35	28.20	56.09	59.68	-3.59
36	24.58	50.25	52.42	-2.17
37	20.25	43.65	43.74	-0.09
38	17.09	38.01	37.40	0.61
39	14.35	31.40	31.91	-0.51
40	13.11	29.45	29.42	0.03
41	9.50	29.02	22.18	6.84
42	9.74	19.05	22.66	-3.61
43	9.34	20.36	21.86	-1.50
44	7.51	17.68	18.19	-0.51
45	8.35	19.23	19.87	-0.64
46	6.25	14.92	15.66	-0.74
47	5.45	11.44	14.06	-2.62
48	3.79	12.69	10.73	1.96

Table 4: Approximated Monthly Outputs, total costs, fitted values and residuals – Dean (1941)

.

Here is a proof that the final term on the right-hand side is 0 (which is very easy in matrix algebra):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = (\overline{Y} - \left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\overline{X}) + X_i \left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \quad (11)$$

$$e_i = Y_i - \hat{Y}_i = Y_i - (\overline{Y} - \left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\overline{X}) + X_i \left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \quad (12)$$

Combining equations (11) and (12), we get:

$$\begin{aligned}
e_i\hat{Y}_i &= Y_i \left[\overline{Y} - \overline{X}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right] + X_i Y_i \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] - \\
&\left[\overline{Y} - \overline{X}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right]^2 - X_i^2 \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 - \\
&2X_i \left[\overline{Y} - \overline{X}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right] \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \\
&= Y_i \left[\overline{Y} - \overline{X}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)\right] + X_i Y_i \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] - \overline{Y}^2 - \\
&\overline{X}^2 \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 + 2\overline{Y}\,\overline{X} \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] - \\
&X_i^2 \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 - 2X_i\overline{X} \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 \\
&= \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 (2X_i\overline{X} - X_i^2 - \overline{X}^2) + \\
&\left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] (-Y_i\overline{X} + X_i Y_i + 2\overline{Y}\,\overline{X} - 2X_i\overline{Y}) + Y_i\overline{Y} - \overline{Y}^2
\end{aligned}$$

Now summing $e_i\hat{Y}_i$ over all observations:

$$\begin{aligned}
\sum_{i=1}^n e_i\hat{Y}_i &= \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 (2\overline{X}\sum_{i=1}^n X_i - \sum_{i=1}^n X_i^2 - n\overline{X}^2) + \\
&\left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] (\sum_{i=1}^n X_i Y_i - \overline{X}\sum_{i=1}^n Y_i + 2n\overline{XY} - 2\overline{Y}\sum_{i=1}^n X_i) + \overline{Y}\sum_{i=1}^n Y_i - n\overline{Y}^2 \\
&= \left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 (2n\overline{X}^2 - \sum_{i=1}^n X_i^2 - n\overline{X}^2) + \\
&\left[\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] (\sum_{i=1}^n X_i Y_i - n\overline{XY} + 2n\overline{XY} - 2n\overline{XY}) + n\overline{Y}^2 - n\overline{Y}^2
\end{aligned}$$

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

Now making use of the following two facts:

$$\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 + n\overline{X}^2 - 2\overline{X}\sum_{i=1}^n X_i = \sum_{i=1}^n X_i^2 - n\overline{X}^2$$

$$\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n X_i Y_i + n\overline{XY} - \overline{X}\sum_{i=1}^n Y_i - \overline{Y}\sum_{i=1}^n X_i = \sum_{i=1}^n X_i Y_i - n\overline{XY}$$

We obtain the desired result:

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

$$= -\left[\frac{(\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}))^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] + \left[\frac{(\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}))^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 0$$

Thus we can partition the total (uncorrected) sum of squares into the sum of squares of the predicted values (the model sum of squares) and the sum of squares of the errors (the residual sum of squares).

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n \hat{Y}_i^2 + \sum_{i=1}^n e_i^2$$

$$SS(\text{TOTAL UNCORRECTED}) = SS(\text{MODEL}) + SS(\text{RESIDUAL})$$

The computational formulas are obtained as follows:

$$SS(\text{Model}) = \sum_{i=1}^n \hat{Y}_i^2 = \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 X_i)^2$$

$$= n\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_1^2 \sum_{i=1}^n X_i^2 = n(\overline{Y} - \hat{\beta}_1\overline{X})^2 + 2(\overline{Y} - \hat{\beta}_1\overline{X})\hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_1^2 \sum_{i=1}^n X_i^2$$

$$= n\overline{Y}^2 + n\hat{\beta}_1^2\overline{X}^2 - 2n\hat{\beta}_1\overline{YX} + 2n\hat{\beta}_1\overline{YX} - 2\hat{\beta}_1^2 n\overline{X}^2 + \hat{\beta}_1^2 \sum_{i=1}^n X_i^2$$

$$= n\overline{Y}^2 - n\hat{\beta}_1^2\overline{X}^2 + \hat{\beta}_1^2 \sum_{i=1}^n X_i^2 = n\overline{Y}^2 + \hat{\beta}_1^2(\sum_{i=1}^n X_i^2 - n\overline{X}^2)$$

$$= n\overline{Y}^2 + \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2$$

$$SS(\text{RESIDUAL}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = SS(\text{TOTAL UNCORRECTED}) - SS(\text{MODEL})$$

The total (uncorrected) sum of squares is of little interest by itself, since it depends on the level of the data, but not the variability (around the mean). The total (corrected) sum of squares measures the sum of the squared deviations around the mean.

$$\begin{aligned}
SS(\text{TOTAL CORRECTED}) &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \\
= (SS(\text{MODEL}) - n\overline{Y}^2) + SS(\text{RESIDUAL}) &= \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n} e_i^2 \\
&= SS(\text{REGRESSION}) + SS(\text{RESIDUAL})
\end{aligned}$$

Note that the model sum of squares considers both β_0 and β_1, while the regression sum of squares considers only the slope β_1.

For a general regression model, with p independent variables, we have the following model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots \beta_p X_{ip} \quad i = 1, \ldots, n$$

which contains $p' = p + 1$ model parameters. The Analysis of Variance is given in Table 1.3, which contains sources of variation, their degrees of freedom, and sums of squares.

Source of Variation	Degrees of Freedom	Sum of Squares
Total (Uncorrected)	n	$\sum_{i=1}^{n} Y_i^2$
Correction Factor	1	$n\overline{Y}^2$
Total (Corrected)	$n-1$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$
Model	$p' = p + 1$	$\sum_{i=1}^{n} \hat{Y}_i^2$
Correction Factor	1	$n\overline{Y}^2$
Regression	p	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} \hat{Y}_i^2 - n\overline{Y}^2$
Residual	n-p’	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Table 5: The Analysis of Variance

.

The mean squares for regression and residuals are the corresponding sums of squares divided by their respective freedoms:

$$MS(\text{REGRESSION}) = \frac{SS(\text{REGRESSION})}{p}$$

$$MS(\text{RESIDUAL}) = \frac{SS(\text{RESIDUAL})}{n - p'}$$

The expected value of the mean squares are given below (for the case where there is a single independent variable), the proof is given later. These are based on the assumption that the model is fit is the correct model.

$$E[MS(\text{REGRESSION})] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$E[MS(\text{RESIDUAL})] \quad = \quad \sigma^2$$

The **Coefficient of Determination** (R^2) is the ratio of the regression sum of squares to the total (corrected) sum of squares, and represents the fraction of the variation in Y that is "explained" by the set of independent variables $X_1, \ldots, X_p$.

$$R^2 \quad = \quad \frac{SS(\text{REGRESSION})}{SS(\text{TOTAL CORRECTED})} \quad = \quad 1 - \frac{SS(\text{RESIDUAL})}{SS(\text{TOTAL CORRECTED})}$$

1.3.1 Examples

Numerical results for the two examples desribed before are given below.

Example 1 – Pharmacodynamics of LSD

The Analysis of Variance for the LSD/math score data are given in Table 1.3.1. Here, $n = 7$, $p = 1$, and $p' = 2$. All relevant sums are obtained from previous examples.

Source of Variation	Degrees of Freedom	Sum of Squares
Total (Uncorrected)	$n = 7$	$\sum_{i=1}^n Y_i^2 = 19639.24$
Correction Factor	1	$n\overline{Y}^2 = 17561.02$
Total (Corrected)	$n - 1 = 6$	$\sum_{i=1}^n (Y_i - \overline{Y})^2 = \sum_{i=1}^n Y_i^2 - n\overline{Y}^2 = 2078.22$
Model	$p' = p + 1 = 2$	$\sum_{i=1}^n \hat{Y}_i^2 = 19385.32$
Correction Factor	1	$n\overline{Y}^2 = 17561.02$
Regression	p=1	$\sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^n \hat{Y}_i^2 - n\overline{Y}^2 = 1824.30$
Residual	n-p'=5	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 253.92$

Table 6: The Analysis of Variance for the LSD/math score data

.

The mean squares for regression and residuals are as follow:

$$MS(\text{REGRESSION}) \quad = \quad \frac{SS(\text{REGRESSION})}{p} \quad = \quad \frac{1824.30}{1} \quad = \quad 1824.30$$

$$MS(\text{RESIDUAL}) \quad = \quad \frac{SS(\text{RESIDUAL})}{n - p'} \quad = \quad \frac{253.92}{5} \quad = \quad 50.78$$

The coefficient of determination for this data is:

$$R^2 \quad = \quad \frac{SS(\text{REGRESSION})}{SS(\text{TOTAL CORRECTED})} \quad = \quad \frac{1824.30}{2078.22} \quad = \quad 0.8778$$

Approximately 88% of the variation in math scores is "explained" by the linear relation between math scores and LSD concentration.

Example 2 – Estimating Cost Function of a Hosiery Mill

The Analysis of Variance for the output/cost data are given in Table 1.3.1. Here, $n = 48$, $p = 1$, and $p' = 2$. All relevant sums are obtained from previous examples and computer output.

Source of Variation	Degrees of Freedom	Sum of Squares
Total (Uncorrected)	$n = 48$	$\sum_{i=1}^{n} Y_i^2 = 238424.46$
Correction Factor	1	$n\overline{Y}^2 = 205510.29$
Total (Corrected)	$n - 1 = 47$	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = 32914.17$
Model	$p' = p + 1 = 2$	$\sum_{i=1}^{n} \hat{Y}_i^2 = 236636.27$
Correction Factor	1	$n\overline{Y}^2 = 205510.29$
Regression	p=1	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} \hat{Y}_i^2 - n\overline{Y}^2 = 31125.98$
Residual	n-p'=46	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 1788.19$

Table 7: The Analysis of Variance for the hosiery mill cost function data

.

The mean squares for regression and residuals are as follow:

$$MS(\text{REGRESSION}) = \frac{SS(\text{REGRESSION})}{p} = \frac{31125.98}{1} = 31125.98$$

$$MS(\text{RESIDUAL}) = \frac{SS(\text{RESIDUAL})}{n - p'} = \frac{1788.19}{46} = 38.87$$

The coefficient of determination for this data is:

$$R^2 = \frac{SS(\text{REGRESSION})}{SS(\text{TOTAL CORRECTED})} = \frac{31125.98}{32914.17} = 0.9457$$

Approximately 95% of the variation in math scores is "explained" by the linear relation between math scores and LSD concentration.

1.4 Precision and Distribution of Estimates

Important results from mathematical statistics regarding linear functions of random variables. Let $U = \sum_{i=1}^{n} a_i Y_i$, where $a_1, \ldots, a_n$ are fixed constants and Y_i are random variables with $E(Y_i) = \mu_i$, $Var(Y_i) = \sigma^2$, and $Cov(Y_i, Y_j) = 0$, $i \neq j$:

$$E[U] = E[\sum_{i=1}^{n} a_i Y_i] = \sum_{i=1}^{n} a_i E[Y_i] = \sum_{i=1}^{n} a_i \mu_i \quad (13)$$

$$Var[U] = Var[\sum_{i=1}^{n} a_i Y_i] = \sum_{i=1}^{n} a_i^2 Var[Y_i] = \sum_{i=1}^{n} a_i \sigma_i^2 \quad (14)$$

$$E[e^{tU}] = E[e^{t\sum_{i=1}^{n} a_i Y_i}] = E\left[[e^{ta_1 Y_1} \cdots e^{ta_n Y_n}\right] = \prod_{i=1}^{n} E\left[e^{t_i^* Y_i}\right] \qquad t_i^* = a_i t \quad (15)$$

1.4.1 Distribution of $\hat{\beta}_1$

Consider $\hat{\beta}_1$ as a linear function of $Y_1, \ldots, Y_n$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i - \overline{Y}\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} Y_i = \sum_{i=1}^n a_i Y_i$$

Under the simple linear regression model:

$$E[Y_i] = \mu_i = \beta_0 + \beta_1 X_i \qquad Var[Y_i] = \sigma_i^2 = \sigma^2$$

From equation (13):

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

$$= \frac{\beta_0}{\sum_{i=1}^n (X_i - \overline{X})^2}\sum_{i=1}^n (X_i - \overline{X}) + \frac{\beta_1}{\sum_{i=1}^n (X_i - \overline{X})^2}\sum_{i=1}^n (X_i - \overline{X})X_i = \frac{\beta_1}{\sum_{i=1}^n (X_i - \overline{X})^2}\sum_{i=1}^n (X_i - \overline{X})X_i$$

$$= \frac{\beta_1}{\sum_{i=1}^n (X_i - \overline{X})^2}\left[\sum_{i=1}^n X_i^2 - \overline{X}\sum_{i=1}^n X_i\right] = \frac{\beta_1}{\sum_{i=1}^n (X_i - \overline{X})^2}\left[\sum_{i=1}^n X_i^2 - n\overline{X}^2\right]$$

$$= \frac{\beta_1}{\sum_{i=1}^n (X_i - \overline{X})^2}\sum_{i=1}^n (X_i - \overline{X})^2 = \beta_1$$

From Equation (14):

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

$$= \left[\frac{1}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^2 \sigma^2 \sum_{i=1}^n (X_i - \overline{X})^2 = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

With the further assumption that Y_i (and, more specifically, ε_i) being normally distributed, we can obtain the specific distribution of $\hat{\beta}_1$.

$$E\left[e^{t\hat{\beta}_1}\right] = E\left[e^{t\sum_{i=1}^n \sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} Y_i}\right] = E\left[e^{\sum_{i=1}^n t_i^* Y_i}\right]$$

where $t_i^* = t\frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}$. If $Y \sim N(\mu, \sigma^2)$, then the moment generating function for Y is:

$$m_Y(t) = E\left[e^{tY}\right] = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

$$\Rightarrow E\left[e^{t_i^* Y_i}\right] = exp\left\{(\beta_0 + \beta_1 X_i)\left(\frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)t + \left(\frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2}\right)^2 \frac{\sigma^2 t^2}{2}\right\}$$

By independence of the Y_i, we get:

$$E\left[e^{t\hat{\beta}_1}\right] = E\left[e^{t\sum_{i=1}^n \frac{X_i - \overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2} Y_i}\right] = \prod_{i=1}^n E\left[e^{t_i^* Y_i}\right]$$

$$\prod_{i=1}^{n} exp\{(\beta_0+\beta_1 X_i)\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)t+\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^2\frac{\sigma^2 t^2}{2}\}$$

$$= exp\left\{\sum_{i=1}^{n}(\beta_0+\beta_1 X_i)\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)t+\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^2\frac{\sigma^2 t^2}{2}\right\} \tag{16}$$

Expanding the first term in the exponent in equation (16), we get:

$$\begin{aligned}\sum_{i=1}^{n}(\beta_0+\beta_1 X_i)\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)t &= \frac{t}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\{\beta_0\sum_{i=1}^{n}(X_i-\overline{X})+\beta_1\sum_{i=1}^{n}X_i(X_i-\overline{X})\}\\ &= \frac{t}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\{0+\beta_1\sum_{i=1}^{n}(X_i-\overline{X})^2\} = \beta_1 t \end{aligned}\tag{17}$$

Expanding the second term in the exponent in equation (16), we get:

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}\right)^2\frac{\sigma^2 t^2}{2} = \frac{\sigma^2 t^2}{2(\sum_{i=1}^{n}(X_i-\overline{X})^2)^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{\sigma^2 t^2}{2\sum_{i=1}^{n}(X_i-\overline{X})^2} \tag{18}$$

Putting equations (17) and (18) back into equation 1(16), we get:

$$m_{\hat{\beta}_1}(t) = E\left[e^{t\hat{\beta}_1}\right] = exp\{\beta_1 t+\frac{\sigma^2 t^2}{2\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

which is the moment generating function of a normally distributed random variable with mean β_1 and variance $\sigma^2/\sum_{i=1}^{n}(X_i-\overline{X})^2$. Thus, we have the complete sampling distribution of $\hat{\beta}_1$ under the model's assumptions.

1.4.2 Distribution of $\hat{\beta}_0$

Consider the following results from mathematical statistics:

$$U = \sum_{i=1}^{n}a_i Y_i \qquad W = \sum_{i=1}^{n}d_i Y_i$$

where $\{a_i\}$ and $\{d_i\}$ are constants and $\{Y_i\}$ are random variables. Then:

$$Cov[U,W] = \sum_{i=1}^{n}a_i d_i V(Y_i)+\sum_{i=1}^{n}\sum_{j\neq i}a_i d_j Cov[Y_i,Y_j]$$

Then, we can write $\hat{\beta}_0$ as two linear functions of the $\{Y_i\}$:

$$\hat{\beta}_0 = \overline{Y}-\hat{\beta}_1\overline{X} = \sum_{i=1}^{n}\frac{1}{n}Y_i-\sum_{i=1}^{n}\frac{\overline{X}(X_i-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}Y_i = U-W \tag{19}$$

The expected values of the the two linear functions of the Y_i in equation (19) are as follow:

$$E[U] = \sum_{i=1}^{n}\frac{1}{n}(\beta_0+\beta_1 X_i) = \frac{1}{n}(n\beta_0)+\frac{1}{n}\beta_1\sum_{i=1}^{n}X_i = \beta_0+\beta_1\overline{X}$$

$$
\begin{aligned}
E[W] &= \sum_{i=1}^{n} \frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}(\beta_0 + \beta_1 X_i) \\
&= \frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}[\beta_0 X_i + \beta_1 X_i^2 - \beta_0 \overline{X} - \beta_1 \overline{X} X_i] \\
&= \frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}[\beta_0 \sum_{i=1}^{n}(X_i - \overline{X}) + \beta_1 (\sum_{i=1}^{n} X_i^2 - n\overline{X}^2)] \\
& \frac{\overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}[0 + \beta_1 \sum_{i=1}^{n}(X_i - \overline{X})^2] = \beta_1 \overline{X}
\end{aligned}
$$

Putting these together in equation (19):

$$
E[\hat{\beta}_0] = E[U - W] = E[U] - E[W] = \beta_0 + \beta_1 \overline{X} - \beta_1 \overline{X} = \beta_0
$$

Now to get the variance of $\hat{\beta}_0$ (again assuming that $Cov[Y_i, Y_j] = 0 for i \neq j$) :

$$
Var[U - W] = Var[U] + Var[W] - 2Cov[U, W]
$$

$$
Var[U] = Var[\sum_{i=1}^{n} \frac{1}{n} Y_i] = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 Var[Y_i] = n\left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}
$$

$$
Var[W] = Var[\hat{\beta}_1 \overline{X}] = \overline{X}^2 Var[\hat{\beta}_1] = \overline{X}^2 \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}
$$

$$
Cov[U, W] = \sum_{i=1}^{n} \frac{1}{n}\left(\frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right) Var[Y_i] = \frac{\sigma^2 \overline{X}}{n \sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X}) = 0
$$

$$
\Rightarrow \quad Var[\hat{\beta}_0] = Var[U] + Var[W] - 2Cov[U, W] = \frac{\sigma^2}{n} + \frac{\overline{X}^2 \sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]
$$

Note that $Cov[U, W] = Cov[\overline{Y}, \hat{\beta}_1 \overline{X}] = 0$, then $\overline{Y}$ and $\hat{\beta}_1$ are independent. We can also write $\hat{\beta}_0$ as a single linear function of Y_i, allowing use of the moment generating function method to determine it's normal sampling distribution:

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

1.4.3 Distribution of $\hat{Y}_i$

The distibution of $\hat{Y}_i$, which is an estimate of the population mean of Y_i at the level X_i of the independent variable is obtained as follows:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = (\overline{Y} - \hat{\beta}_1\overline{X}) + \hat{\beta}_1 X_i = \overline{Y} + \hat{\beta}_1(X_i - \overline{X})$$

$$E[\hat{Y}_i] = E[\overline{Y}] + (X_i - \overline{X})E[\hat{\beta}_1] = \beta_0 + \beta_1\overline{X} + \beta_1(X_i - \overline{X}) = \beta_0 + \beta_1 X_i$$

$$\begin{aligned} Var[\hat{Y}_i] &= Var[\overline{Y}] + (X_i - \overline{X})^2 Var[\hat{\beta}_1] + 2(X_i - \overline{X})Cov[\overline{Y}, \hat{\beta}_1] = \frac{\sigma^2}{n} + \frac{(X_i - \overline{X})^2\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} + 0 \\ &= \sigma^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \end{aligned}$$

Further, since $\hat{Y}_i$ is a linear function of the Y_i, then $\hat{Y}_i$ has a normal sampling distribution.

1.4.4 Prediction of future observation Y_0 when $X = X_0$

The predicted value of a future observation Y_0 is $\hat{Y}_{pred0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$. The prediction error is $Y_0 - \hat{Y}_0$, and the quantity $E[(\hat{Y}_0 - Y_0)^2]$ is referred to as the mean square error of prediction. Assuming the model is correct:

$$E[\hat{Y}_0 - Y_0] = 0$$

$$Var[\hat{Y}_{pred0}] = Var[\hat{Y}_0 - Y_0] = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] + \sigma^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$$

1.4.5 Estimated Variances

For all of the sampling distributions derived above, the unknown observation variance σ^2 appears in the estimators' variances. To obtain the estimated variance for each of the estimators, we replace σ^2 with $s^2 = MS$(RESIDUAL). It's important to keep in mind that this estimator is unbiased for σ^2 only if the model is correctly specified. The estimated variances for each estimator and predictor are given below:

- $s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$ Estimated variance of $\hat{\beta}_1$
- $s^2(\hat{\beta}_0) = s^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$ Estimated variance of $\hat{\beta}_0$
- $s^2(\hat{Y}_i) = s^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$ Estimated variance of estimated mean at X_i
- $s^2(\hat{Y}_{pred0}) = s^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]$ Estimated variance of prediction at X_0

1.4.6 Examples

Estimated variances are computed for both of the previous examples.

Example 1 – Pharmacodynamics of LSD

Here we obtain estimated variances for $\hat{\beta}_1$, $\hat{\beta}_0$, the true mean, and a future score when the tissue concentration is 5.0:

$$s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{50.78}{22.48} = 2.26$$

$$s^2(\hat{\beta}_0) = s^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 50.78\left[\frac{1}{7} + \frac{4.3329^2}{22.48}\right] = 49.66$$

$$s^2(\hat{Y}_5) = s^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 50.78\left[\frac{1}{7} + \frac{(5 - 4.3329)^2}{22.48}\right] = 8.26$$

$$s^2(\hat{Y}_{pred0}) = s^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 50.78\left[1 + \frac{1}{7} + \frac{(5 - 4.3329)^2}{22.48}\right] = 59.04$$

Example 2 – Estimating Cost Function of a Hosiery Mill

Here we obtain estimated variances for $\hat{\beta}_1$, $\hat{\beta}_0$, the true mean, and a future cost when the production output is 30.0:

$$s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{38.87}{7738.94} = 0.0050$$

$$s^2(\hat{\beta}_0) = s^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 38.87\left[\frac{1}{48} + \frac{31.0673^2}{7738.94}\right] = 5.66$$

$$s^2(\hat{Y}_{30}) = s^2\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 38.87\left[\frac{1}{48} + \frac{(30 - 31.0673)^2}{7738.94}\right] = 0.82$$

$$s^2(\hat{Y}_{pred0}) = s^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] = 38.87\left[1 + \frac{1}{48} + \frac{(30 - 31.0673)^2}{7738.94}\right] = 39.69$$

1.5 Tests of Significance and Confidence Intervals

Under the model assumptions of independence, normality and constant error variance; we can make inferences concerning model parameters. We can conduct t–tests, F–tests, and obtain confidence intervals regarding the unknown parameters.

1.5.1 Tests of Significance

The t-test can be used to test hypotheses regarding β_0 or β_1, and can be used for 1–sided or 2–sided alternative hypotheses. The form of the test is as follows, and can be conducted regarding any of the regression coefficients:

- $H_0 : \beta_i = m$ (m specified, usually 0 when testing β_1)
- (1) $H_a : \beta_i \neq m$
  (2) $H_a : \beta_1 > m$
  (3) $H_a : \beta_1 < m$
- $TS : t_0 = \frac{\hat{\beta}_i - m}{s(\hat{\beta}_i)}$
- (1) $RR : |t_0| \geq t_{(\alpha/2, n-p')}$ ($p' = 2$ for simple regression)
  (2) $RR : t_0 \geq t_{(\alpha, n-p')}$ ($p' = 2$ for simple regression)
  (3) $RR : t_0 \leq -t_{(\alpha, n-p')}$ ($p' = 2$ for simple regression)
- (1) P–value: $2 \cdot P(t \geq |t_0|)$
  (2) P–value: $P(t \geq t_0)$
  (3) P–value: $P(t \leq t_0)$
  Using tables, we can only place bounds on these p–values, but statistical computing packages will print them directly.

A second test is available to test whether the slope parameter is 0 (no linear association exists between Y and X). This is based on the Analysis of Variance and the F-distribution:

1. $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0$ (This will always be a 2–sided test)
2. T.S.: $F_0 = \frac{MS(\text{REGRESSION})}{MS(\text{RESIDUAL})}$
3. R.R.: $F_0 > F_{(\alpha, 1, n-p'}$ ($p' = 2$ for simple regression)
4. p-value: $P(F > F_0)$ (You can only get bounds on this from tables, but computer outputs report them exactly)

Under the null hypothesis, the test statistic should be near 1, as β_1 moves away from 0, the test statistic should increase.

1.5.2 Confidence Intervals

Confidence intervals for model parameters can be obtained under all the previously stated assumptions. The $100(1-\alpha)100\%$ confidence intervals can be obtained as follows:

$$\beta_0 : \qquad \hat{\beta}_0 \pm t_{(\alpha/2, n-p')} s(\hat{\beta}_0)$$

$$\beta_1 : \qquad \hat{\beta}_1 \pm t_{(\alpha/2, n-p')} s(\hat{\beta}_1)$$

$$\beta_0 + \beta_1 X_i : \qquad \hat{Y}_i \pm t_{(\alpha/2, n-p')} s(\hat{Y}_i)$$

Prediction intervals for future observations at $X = X_0$ can be obtained as well in an obvious manner.

1.5.3 Examples

The previously described examples are continued here.

Example 1 – Pharmacodynamics of LSD

To determine whether there is a negative association between math scores and LSD concentration, we conduct the following test at $\alpha = 0.05$ significance level. Note that $s(\hat{\beta}_1) = \sqrt{s^2(\hat{\beta}_1)} = \sqrt{2.26} = 1.50$.

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 < 0$$

$$TS: t_0 = \frac{\hat{\beta}_1 - 0}{s(\hat{\beta}_1)} = \frac{-9.01}{1.50} = -6.01 \quad RR: t_0 \leq -t_{.05,5} = -2.015 \quad P-val = P(t \leq -6.01)$$

Next, we obtain a confidence interval for the true mean score when the tissue concentration is $X = 5.0$. The estimated standard error of $\hat{Y}_5$ is $s(\hat{Y}_5) = \sqrt{s^2(\hat{Y}_5)} = \sqrt{8.26} = 2.87$, and $t_{(0.025,5)} = 2.571$. The 95% confidence interval for $\beta_0 + \beta_1(5)$ is:

$$\hat{Y}_5 = 89.12 - 9.01(5) = 44.07 \qquad 44.07 \pm 2.571(2.87) \quad \equiv \quad 44.07 \pm 7.38 \quad \equiv \quad (36.69, 51.45)$$

Example 2 – Estimating Cost Function of a Hosiery Mill

Here, we use the F-test to determine whether there is an association between product costs and the production output at $\alpha = 0.05$ significance level.

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

$$TS: F_0 = \frac{MS(\text{REGRESSION})}{MS(\text{RESIDUAL})} = \frac{31125.98}{38.87} = 800.77 \quad RR: F_0 \geq F_{(.05,1,46)} \approx 1.680 \quad P-val = P(F \geq 800.77)$$

Unit variable cost is the average increment in total production cost per unit increase in production output (β_1). We obtain a 95% confidence interval for this parameter:

$$\hat{\beta}_1 = 2.0055 \quad s^2(\hat{\beta}_1) = .0050 \quad s(\hat{\beta}_1) = \sqrt{.0050} = .0707 \quad t_{(.025,46)} \approx 2.015$$

$$\hat{\beta}_1 \pm t_{(.025,46)} s(\hat{\beta}_1) \quad 2.0055 \pm 2.015(.0707) \quad 2.0055 \pm 0.1425 \quad (1.8630, 2.1480)$$

As the production output increases by 1000 dozen pairs, we are very confident that mean costs increase by between 1.86 and 2.15 \$1000. The large sample size $n = 48$ makes our estimate very precise.

1.6 Regression Through the Origin

In some practical situations, the regression line is expected (theoretically) to pass trough the origin. It is important that $X = 0$ is a reasonable level of X in practice for this to be the case. For instance, in the hosiery mill example, if firm knows in advance that production will be 0 they close plant and have no costs if they are able to work in "short–run," however most firms still have "long–run" costs if they know they will produce in future. If a theory does imply that the mean response (Y) is 0 when $X = 0$, we have a new model:

$$Y_i = \beta_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

The least squares estimates are obtained by minimizing (over β_1):

$$Q = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

This is obtained by taking the derivative of Q with respect to β_1, and setting it equal to 0. The value $\hat{\beta}_1$ that solves that equality is the least squares estimate of β_1.

$$\frac{\partial Q}{\partial \beta_1} = 2\sum_{i=1}^{n}(Y_i - \beta_1 X_i)(-X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

The estimated regression equation and residuals are:

$$\hat{Y}_i = \hat{\beta}_1 X_i \qquad e_i = Y - \hat{Y}_i = Y - \hat{\beta}_1 X_i$$

Note that for this model, the residuals do not necessarily sum to 0:

$$e_i = Y - \hat{Y}_i = Y - \hat{\beta}_1 X_i = Y_i - \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) X_i$$

$$\Rightarrow \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right)\sum_{i=1}^{n} X_i$$

This last term is not necessarily (and will probably rarely, if ever, in practice be) 0.
The uncorrected sum of squares is:

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{Y}_i e_i$$

The last term (the cross–product term) is still 0 under the no–interecept model:

$$\sum_{i=1}^{n} e_i \hat{Y}_i = \sum_{i=1}^{n} (Y_i - \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) X_i)(\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) X_i)$$

$$= \sum_{i=1}^{n} Y_i(\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) X_i) - \sum_{i=1}^{n}(\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) X_i)^2 = \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right)\sum_{i=1}^{n} X_i Y_i - \left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right)^2 \sum_{i=1}^{n} X_i^2 = 0$$

So, we obtain the same partitioning of the total sum of squares as before:

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2$$

$$SS(\text{TOTAL UNCORRECTED}) = SS(\text{MODEL}) + SS(\text{RESIDUAL})$$

The model sum of squares is based on only one parameter, so it is not broken into the components of mean and regression as it was before. Similarly, the residual sum of squares has $n-1$ degrees of freedom. Assuming the model is correct:

$$E[MS(\text{MODEL})] = \sigma^2 + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$E[MS(\text{RESIDUAL})] = \sigma^2$$

The variance of the estimator $\hat{\beta}_1$ is:

$$Var[\hat{\beta}_1] = Var\left[\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right] = \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} \sum_{i=1}^{n} X_i^2 Var[Y_i]$$

$$= \frac{1}{(\sum_{i=1}^{n} X_i^2)^2} (\sum_{i=1}^{n} X_i^2)\sigma^2 = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

Similarly, the variance of $\hat{Y}_0 = X_0 \hat{\beta}_1$ is:

$$Var[\hat{Y}_0] = Var[X_0 \hat{\beta}_1] = X_0^2 Var[\hat{\beta}_1] = \frac{\sigma^2 X_0^2}{\sum_{i=1}^{n} X_i^2}$$

Estimates are obtained by replacing σ^2 with $s^2 = MS(\text{RESIDUAL})$.

1.6.1 Example – Galton's Height Measurements

In what is considered by many to be the first application of regression analysis, Sir Frances Galton (1889, *Natural Inheritance*) obtained heights of $n = 928$ adult children (Y) and the"midheight" of their parents (X). Since the mean heights of adult children and their parents were approximately the same (68.1" for adult children and 68.3" for their parents). Once both datasets have been centered around their means, Galton found that adult chidrens heights were less extreme than their parents. This phenomenon has been observed in many areas of science, and is referred to as *regression to the mean.*

Here we fit a regression model through the origin, which for this centered data is the point (68.1,68.3). We have the following quantities based on the centered data given in Galton's table:

$$n = 928 \quad \sum_{i=1}^{n} X_i^2 = 3044.92 \quad \sum_{i=1}^{n} Y_i^2 = 5992.48 \quad \sum_{i=1}^{n} X_i Y_i = 1965.46$$

From this data, we obtain the following quantities:

$$\hat{\beta}_1 \quad = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} = \frac{1965.46}{3044.92} = 0.6455$$

$$SS(\text{MODEL}) = \sum_{i=1}^{n} \hat{Y}_i^2 = \sum_{i=1}^{n} (\hat{\beta}_1 X_i)^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} X_i^2 = (0.6455)^2(3044.92) = 1268.73$$

$$SS(\text{RESIDUAL}) = \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} \hat{Y}_i^2 = 5992.48 - 1268.73 = 4723.75$$

$$s^2 = MS(\text{RESIDUAL}) = \frac{SS(\text{RESIDUAL})}{n-1} = \frac{4723.75}{927} = 5.10$$

$$s^2(\hat{\beta}_1) = \frac{s^2}{\sum_{i=1}^{n} X_i^2} = \frac{5.10}{3044.92} = .0017 \qquad s(\hat{\beta}_1) = .0409$$

From this, we get a 95% confidence interval for β_1:

$$bh_1 \pm z_{(.025)} s(\hat{\beta}_1) \equiv 0.6455 \pm 1.96(.0409) \equiv 0.6455 \pm 0.0802 \equiv (0.5653, 0.7257)$$

Note that there is a positive association between adult children's heights and their parent's heights. However, as the parent's height increases by 1", the adult child's height increases by between 0.5633" and 0.7257" on average. This is an example of regression to the mean.

1.7 Models with Several Independent Variables

As was discussed in the section on the Analysis of Variance, models can be generalized to contain $p < n$ independent variables. However, the math to obtain estimates and their estimated variances and standard errors is quite messier. This can be avoided by making use of matrix algebra, which is introduced shortly. The general form of the multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad \varepsilon \sim NID(0, \sigma^2)$$

The least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the values that minimize the residual sum of squares:

$$SS(\text{RESIDUAL}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip})^2$$

An unbiased estimate of σ^2 is:

$$s^2 = SS(\text{RESIDUAL}) = \frac{SS(\text{RESIDUAL})}{n - (p+1)}$$

We will obtain these estimates after we write the model in matrix notation.

1.8 SAS Programs and Output

In this section, SAS code and its corresponding output are given for the two examples in Rawlings, Pantula, and Dickey (RPD).

2 Introduction to Matrices

Text: RPD, Sections 2.1-2.6

Problems:

In this section, important definitions and results from matrix algebra that are useful in regression analysis are introduced. While all statements below regarding the columns of matrices can also be said of rows, in regression applications we will typically be focusing on the columns.

A **matrix** is a rectangular array of numbers. The **order** or **dimension** of the matrix is the number of rows and columns that make up the matrix. The **rank** of a matrix is the number of linearly independent columns (or rows) in the matrix.

A subset of columns is said to be **linearly independent** if no column in the subset can be written as a linear combination of the other columns in the subset. A matrix is **full rank (nonsingular)** if there are no linear dependencies among its columns. The matrix is **singular** if lineardependencies exist.

The **column space** of a matrix is the collection of all linear combinations of the columns of a matrix.

The following are important types of matrices in regression:

Vector – Matrix with one row or column

Square Matrix – Matrix where number of rows equals number of columns

Diagonal Matrix – Square matrix where all elements off main diagonal are 0

Identity Matrix – Diagonal matrix with 1's everywhere on main diagonal

Symmetric Matrix – Matrix where element $a_{ij} = a_{ji}\ \forall i, j$

Scalar – Matrix with one row and one column (single element)

The **transpose** of a matrix is the matrix generated by interchanging the rows and columns of the matrix. If the original matrix is $\boldsymbol{A}$, then its transpose is labelled $\boldsymbol{A'}$. For example:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 7 \\ 1 & 7 & 2 \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}' = \begin{bmatrix} 2 & 1 \\ 4 & 7 \\ 7 & 2 \end{bmatrix}$$

Matrix addition (subtraction) can be performed on two matrices as long as they are of equal order (dimension). The new matrix is obtained by elementwise addition (subtraction) of the two matrices. For example:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 7 \\ 1 & 7 & 2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 3 & 0 \\ 2 & 4 & 8 \end{bmatrix} \quad \Rightarrow \quad \mathbf{A}+\mathbf{B} = \begin{bmatrix} 3 & 7 & 7 \\ 3 & 11 & 10 \end{bmatrix}$$

Matrix multiplication can be performed on two matrices as long as the number of columns of the first matrix equals the number of rows of the second matrix. The resulting has the same number of rows as the first matrix and the same number of columns as the second matrix. If $\mathbf{C} = \mathbf{AB}$ and $\mathbf{A}$ has s columns and $\mathbf{B}$ has s rows, the element in the i^{th} row and j^{th} column of $\mathbf{C}$, which we denote c_{ij} is obtained as follows (with similar definitions for a_{ij} and b_{ij}):

$$c_{ij} \quad = \quad a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots a_{is}b_{sj} \quad = \quad \sum_{k=1}^{s} a_{ik}b_{kj}$$

For example:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 7 \\ 1 & 7 & 2 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 5 & 6 \\ 2 & 0 & 1 \\ 3 & 3 & 3 \end{bmatrix} \quad \Rightarrow$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 2(1)+4(2)+7(3) & 2(5)+4(0)+7(3) & 2(6)+4(1)+7(3) \\ 1(1)+7(2)+2(3) & 1(5)+7(0)+2(3) & 1(6)+7(1)+2(3) \end{bmatrix} = \begin{bmatrix} 31 & 31 & 37 \\ 21 & 11 & 19 \end{bmatrix}$$

Note that $\mathbf{C}$ has the same number of rows as $\mathbf{A}$ and the same number of columns as $\mathbf{C}$. Note that in general $\mathbf{AB} \neq \mathbf{BA}$; in fact, the second matrix may not exist due to dimensions of matrices. However, the following equality does hold: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.

Scalar Multiplication can be performed between any scalar and any matrix. Each element of the matrix is multiplied by the scalar. For example:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 7 \\ 1 & 7 & 2 \end{bmatrix} \quad \Rightarrow \quad 2\mathbf{A} = \begin{bmatrix} 4 & 8 & 14 \\ 2 & 14 & 4 \end{bmatrix}$$

The **determinant** is scalar computed from the elements of a matrix via well–defined (although rather painful) rules. Determinants only exist for square matrices. The determinant of a matrix $\mathbf{A}$ is denoted as $|\mathbf{A}|$.

For a scalar (a 1×1 matrix): $|\mathbf{A}| = \mathbf{A}$.
For a 2×2 matrix: $|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$.
For $n \times n$ matrices ($n > 2$):

1. $\mathbf{A_{rs}} \equiv (n-1) \times (n-1)$ matrix with row r and column s removed from $\mathbf{A}$
2. $|\mathbf{A_{rs}}| \equiv$ the **minor** of element a_{rs}
3. $\theta_{rs} = (-1)^{r+s}|\mathbf{A_{rs}}| \equiv$ the **cofactor** of element a_{rs}
4. The determinant is obtained by summing the product of the elements and cofactors for any row or column of $\mathbf{A}$. By using row i of $\mathbf{A}$, we get $|\mathbf{A}| = \sum_{j=1}^{n} a_{ij}\theta_{ij}$

Example – Determinant of a 3×3 matrix
We compute the determinant of a 3×3 matrix, making use of its first row.

$$\mathbf{A} \quad = \quad \begin{bmatrix} 10 & 5 & 2 \\ 6 & 8 & 0 \\ 2 & 5 & 1 \end{bmatrix}$$

$$a_{11} = 10 \quad \mathbf{A_{11}} = \begin{bmatrix} 8 & 0 \\ 5 & 1 \end{bmatrix} \quad |\mathbf{A_{11}}| = 8(1) - 0(5) = 8 \quad \theta_{11} = (-1)^{1+1}(8) = 8$$

$$a_{12} = 5 \quad \mathbf{A_{12}} = \begin{bmatrix} 6 & 0 \\ 2 & 1 \end{bmatrix} \quad |\mathbf{A_{11}}| = 6(1) - 0(2) = 6 \quad \theta_{12} = (-1)^{1+2}(6) = -6$$

$$a_{13} = 2 \quad \mathbf{A_{13}} = \begin{bmatrix} 6 & 8 \\ 2 & 5 \end{bmatrix} \quad |\mathbf{A_{13}}| = 6(5) - 8(2) = 14 \quad \theta_{13} = (-1)^{1+3}(14) = 14$$

Then the determinant of $\mathbf{A}$ is:

$$|\mathbf{A}| \quad = \quad \sum_{j=1}^{n} a_{1j}\theta_{1j} \quad = \quad 10(8) + 5(-6) + 2(14) \quad = \quad 78$$

Note that we would have computed 78 regardless of which row and column we used.

An important result in linear algebra states that if $|\mathbf{A}| = 0$, then $\mathbf{A}$ is singular, otherwise $\mathbf{A}$ is nonsingular (full rank).

The **inverse** of a square matrix $\mathbf{A}$, denoted $\mathbf{A^{-1}}$, is a matrix such that $\mathbf{A^{-1}A} = \mathbf{I} = \mathbf{AA^{-1}}$ where $\mathbf{I}$ is the identity matrix of the same dimension as $\mathbf{A}$. A unique inverse exists if $\mathbf{A}$ is square and full rank.

The identity matrix, when multiplied by any matrix (such that matrix multiplication exists) returns the same matrix. That is: $\mathbf{AI} = \mathbf{A}$ and $\mathbf{IA} = \mathbf{A}$, as long as the dimensions of the matrices conform to matrix multiplication.

For a scalar (a 1×1 matrix): $\mathbf{A^{-1}}=1/\mathbf{A}$.

For a 2×2 matrix: $\mathbf{A^{-1}} = \frac{1}{|\mathbf{A}|} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$.

For $n \times n$ matrices ($n > 2$):

1. Replace each element with its cofactor (θ_{rs})
2. Transpose the resulting matrix
3. Divide each element by the determinant of the original matrix

Example – Inverse of a 3×3 matrix

We compute the inverse of a 3×3 matrix (the same matrix as before).

$$\mathbf{A} \quad = \quad \begin{bmatrix} 10 & 5 & 2 \\ 6 & 8 & 0 \\ 2 & 5 & 1 \end{bmatrix} \qquad |\mathbf{A}| = 78$$

$$|\mathbf{A_{11}}| = 8 \qquad |\mathbf{A_{12}}| = 6 \qquad |\mathbf{A_{13}}| = 14$$

$$|\mathbf{A_{21}}| = -5 \qquad |\mathbf{A_{22}}| = 6 \qquad |\mathbf{A_{23}}| = 40$$
$$|\mathbf{A_{31}}| = -16 \qquad |\mathbf{A_{32}}| = -12 \qquad |\mathbf{A_{33}}| = 50$$

$$\theta_{11} = 8 \quad \theta_{12} = -6 \quad \theta_{13} = 14 \quad \theta_{21} = 5 \quad \theta_{22} = 6 \quad \theta_{23} = -40 \quad \theta_{31} = -16 \quad \theta_{32} = 12 \quad \theta_{33} = 50$$

$$\mathbf{A^{-1}} = \frac{1}{78}\begin{bmatrix} 8 & 5 & -16 \\ -6 & 6 & 12 \\ 14 & -40 & 50 \end{bmatrix}$$

As a check:

$$\mathbf{A^{-1}A} = \frac{1}{78}\begin{bmatrix} 8 & 5 & -16 \\ -6 & 6 & 12 \\ 14 & -40 & 50 \end{bmatrix}\begin{bmatrix} 10 & 5 & 2 \\ 6 & 8 & 0 \\ 2 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{I_3}$$

To obtain the inverse of a diagonal matrix, simply compute the recipocal of each diagonal element.

The following results are very useful for matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and scalar λ, as long as the matrices' dimensions are conformable to the operations in use:

1. $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
2. $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
3. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
4. $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$
5. $\lambda(\mathbf{A} + \mathbf{B}) = \lambda\mathbf{A} + \lambda\mathbf{B}$
6. $(\mathbf{A}')' = \mathbf{A}$
7. $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
8. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
9. $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$
10. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
11. $(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1}$
12. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
13. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$

The length of a column vector $\mathbf{x}$ and the distance between two column vectors $\mathbf{u}$ and $\mathbf{v}$ are:

$$l(\mathbf{x}) = \sqrt{\mathbf{x}'\mathbf{x}} \qquad l((\mathbf{u} - \mathbf{v})) = \sqrt{(\mathbf{u} - \mathbf{v})'(\mathbf{u} - \mathbf{v})}$$

Vectors $\mathbf{x}$ and $\mathbf{w}$ are **orthogonal** if $\mathbf{x}'\mathbf{w} = 0$.

2.1 Linear Equations and Solutions

Suppose we have a system of r linear equations in s unknown variables. We can write this in matrix notation as:

$$\mathbf{Ax} = \mathbf{y}$$

where $\mathbf{x}$ is a $s \times 1$ vector of s unknowns; $\mathbf{A}$ is a $r \times s$ matrix of known coefficients of the s unknowns; and $\mathbf{y}$ is a $r \times 1$ vector of known constants on the right hand sides of the equations. This set of equations may have:

- No solution
- A unique solution
- An infinite number of solutions

A set of linear equations is **consistent** if any linear dependencies among rows of $\mathbf{A}$ also appear in the rows of $\mathbf{y}$. For example, the following system is inconsistent:

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 9 \end{bmatrix}$$

This is inconsistent because the coefficients in the second row of $\mathbf{A}$ are twice those in the first row, but the element in the second row of $\mathbf{y}$ is not twice the element in the first row. There will be no solution to this system of equations.

A set of equations is consistent if $r(\mathbf{A}) = r([\mathbf{Ay}])$ where $[\mathbf{Ay}]$ is the augmented matrix $[\mathbf{A}|\mathbf{y}]$. When $r(\mathbf{A})$ equals the number of unknowns, and A is square:

$$\mathbf{x} = \mathbf{A^{-1}y}$$

2.2 Projection Matrices

The goal of regression is to transform a n-dimensional column vector $\mathbf{Y}$ onto a vector $\hat{\mathbf{Y}}$ in a subspace (such as a straight line in 2-dimensional space) such that $\hat{\mathbf{Y}}$ is as close to $\mathbf{Y}$ as possible. Linear transformation of $\mathbf{Y}$ to $\hat{\mathbf{Y}}$, $\hat{\mathbf{Y}} = \mathbf{PY}$ is said to be a **projection** iff $\mathbf{P}$ is idempotent and symmetric, in which case $\mathbf{P}$ is said to be a **projection matrix**.

A square matrix $\mathbf{A}$ is **idempotent** if $\mathbf{AA} = \mathbf{A}$. If $\mathbf{A}$ is idempotent, then:

$$r(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = tr(\mathbf{A})$$

where $tr(\mathbf{A})$ is the **trace** of $\mathbf{A}$. The subspace of a projection is defined, or spanned, by the columns or rows of the projection matrix $\mathbf{P}$.

$\hat{\mathbf{Y}} = \mathbf{PY}$ is the vector in the subspace spanned by $\mathbf{P}$ that is closest to $\mathbf{Y}$ in distance. That is:

$$SS(\text{RESIDUAL}) = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}})$$

is at a minimum. Further:

$$\mathbf{e} \quad = \quad (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

is a projection onto a subspace orthogonal to the subspace defined by $\mathbf{P}$.

$$\hat{\mathbf{Y}}'\mathbf{e} = (\mathbf{PY})'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{P}'(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'\mathbf{P}(\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}'(\mathbf{P} - \mathbf{P})\mathbf{Y} = 0$$

$$\hat{\mathbf{Y}} + \mathbf{e} = \mathbf{PY} + (\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}$$

2.3 Vector Differentiation

Let f be a function of $\mathbf{x} = [x_1, \ldots, x_p]'$. We define:

$$\frac{df}{d\mathbf{x}} \quad = \quad \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix}$$

From this, we get for $p \times 1$ vector $\mathbf{a}$ and $p \times p$ symmetric matrix $\mathbf{A}$:

$$\frac{d(\mathbf{a}'\mathbf{x})}{d\mathbf{x}} = \mathbf{a} \qquad\qquad \frac{d(\mathbf{x}'\mathbf{A}\mathbf{x})}{d\mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

"Proof" – Consider $p = 3$:

$$\mathbf{a}'\mathbf{x} = a_1x_1 + a_2x_2 + a_3x_3 \quad \frac{d(\mathbf{a}'\mathbf{x})}{dx_i} = a_i \quad \Rightarrow \frac{d(\mathbf{a}'\mathbf{x})}{d\mathbf{x}} = \mathbf{a}$$

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{bmatrix} x_1a_{11} + x_2a_{21} + x_3a_{31} & x_1a_{12} + x_2a_{22} + x_3a_{32} & x_1a_{13} + x_2a_{23} + x_3a_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} =$$

$$= x_1^2a_{11} + x_1x_2a_{21} + x_1x_3a_{31} + x_1x_2a_{12} + x_2^2a_{22} + x_2x_3a_{32} + x_1x_3a_{13} + x_2x_3a_{23} + x_3^2a_{33}$$

$$\Rightarrow \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_i} = 2a_{ii}x_i + 2\sum_{j \neq i} a_{ij}x_j \qquad (a_{ij} = a_{ji})$$

$$\Rightarrow \quad \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_1} \\ \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_2} \\ \frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 + 2a_{13}x_3 \\ 2a_{21}x_1 + 2a_{22}x_2 + 2a_{23}x_3 \\ 2a_{31}x_1 + 2a_{32}x_2 + 2a_{33}x_3 \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

2.4 SAS Programs and Output

In this section, SAS code and its corresponding output are given for the examples in Rawlings, Pantula, and Dickey (RPD).

3 Multiple Regression in Matrix Notation

Multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots \beta_p X_{ip} + \varepsilon_i \qquad i = 1, \ldots, n$$

where i represents the observational unit, the second subscript on X represents the independent variable number, p is the number of independent variables, and $p' = p + 1$ is the number of model parameters (including the intercept term). For the model to have a unique set of regression coefficients, $n > p'$.

We can re–formulate the model in matrix notation:

$\mathbf{Y}$ — $n \times 1$ column vector of observations on the dependent variable Y

$\mathbf{X}$ — $n \times p'$ model matrix containing a column of 1's and p columns of levels of the independent variables $X_1, \ldots, X_p$

$\boldsymbol{\beta}$ — $p' \times 1$ column vector of regression coefficients (parameters)

$\boldsymbol{\varepsilon}$ — $n \times 1$ column vector of random errors

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

For our models, $\mathbf{X}$ will be of full column rank, meaning $r(\mathbf{X}) = p'$.

The elements of $\boldsymbol{\beta}$, are referred to as **partial regression coefficients**, β_j represents the change in $E(Y)$ as the j^{th} independent variable is increased by 1 unit, while all other variables are held constant. The terms "controlling for all other variables" and "ceteris parabis" are also used to describe the effect.

We will be working with many different models (that is, many different sets of independent variables). Often we will need to be more specific of which independent variables are in our model. We denote the partial regression coefficient for X_2 in a model containing X_1, X_2, and X_3 as $\beta_{2\cdot13}$.

3.1 Distributional Properties

We still have the same assumption on the error terms as before:

$$\varepsilon_i \sim NID(0, \sigma^2) \qquad i = 1, \ldots, n$$

This implies that the joint probability density function for the random errors is:

$$f(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) = \prod_{i=1}^{n} f_i(\varepsilon_i) = \prod_{i=1}^{n} \left[(2\pi)^{-1/2} \sigma^{-1} exp \left\{ \frac{-\varepsilon_i^2}{2\sigma^2} \right\} \right] = (2\pi)^{-n/2} \sigma^{-n} exp \left\{ \frac{-\sum_{i=1}^{n} \varepsilon_i^2}{2\sigma^2} \right\}$$

In terms of the observed responses $Y_1, \ldots, Y_n$, we have:

$$Y_i \sim NID(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \sigma^2) \quad \Rightarrow \quad Cov(Y_i, Y_j) = 0 \quad \forall i \neq j$$

From this, the joint probability density function for $Y_1, \ldots, Y_n$ is:

$$f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \left[(2\pi)^{-1/2} \sigma^{-1} exp \left\{ \frac{-(y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}))^2}{2\sigma^2} \right\} \right] =$$

$$= (2\pi)^{-n/2} \sigma^{-n} exp \left\{ \frac{-\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}))^2}{2\sigma^2} \right\}$$

The least squares estimates are Best Linear Unbiased Estimates (B.L.U.E.). Under normality assumption, maximum likelihood estimates are Minimum Variance Unbiased Estimates (M.V.U.E.). In either event, the estimate of β is:

Example 1 – Pharmacodynamics of LSD

For the LSD concentration/math score example, we have the following model for $\mathbf{Y} = \mathbf{X}\, \boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$\begin{bmatrix} 78.93 \\ 58.20 \\ 67.47 \\ 37.47 \\ 45.65 \\ 32.92 \\ 29.97 \end{bmatrix} = \begin{bmatrix} 1 & 1.17 \\ 1 & 2.97 \\ 1 & 3.26 \\ 1 & 4.69 \\ 1 & 5.83 \\ 1 & 6.00 \\ 1 & 6.41 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

Note that $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are unobservable and must be estimated.

3.2 Normal Equations and Least Squares Estimates

Consider the matrices $\mathbf{X'X}$ and $\mathbf{X'Y}$:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1p} & X_{2p} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \cdots & \sum_{i=1}^{n} X_{ip} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \cdots & \sum_{i=1}^{n} X_{i1} X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} X_{ip} & \sum_{i=1}^{n} X_{ip} X_{i1} & \cdots & \sum_{i=1}^{n} X_{ip}^2 \end{bmatrix}$$

For least squares estimation, we minimize $Q(\boldsymbol{\beta})$, the error sum of squares with respect to β:

$$Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}) = \mathbf{Y'Y} - \mathbf{Y'X}\,\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\,\boldsymbol{\beta} = \mathbf{Y'Y} - 2\mathbf{Y'X}\,\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\,\boldsymbol{\beta}$$

By taking the derivative of Q with respect to $\boldsymbol{\beta}$, and setting this to $\mathbf{0}$, we get:

$$\frac{dQ(\boldsymbol{\beta})}{d\,\boldsymbol{\beta}} = \mathbf{0} - 2\mathbf{X'Y} + 2\mathbf{X'X}\,\hat{\boldsymbol{\beta}} = \mathbf{0} \quad \Rightarrow \mathbf{X'X}\,\hat{\boldsymbol{\beta}} = \mathbf{X'Y}$$

This leads to the normal equations and the least squares estimates (when the $\mathbf{X}$ matrix is of full column rank).

$$\mathbf{X'X}\,\hat{\boldsymbol{\beta}} \quad = \quad \mathbf{X'Y} \qquad \Rightarrow \qquad \hat{\boldsymbol{\beta}} \quad = \quad (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Example 1 – Pharmacodynamics of LSD

For the LSD concentration/math score example, we have the following normal equations and least squares estimates:

$$\mathbf{X'X}\,\hat{\boldsymbol{\beta}} \quad = \quad \mathbf{X'Y} \quad \Rightarrow \begin{bmatrix} 7 & 30.33 \\ 30.33 & 153.8905 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 350.61 \\ 1316.6558 \end{bmatrix}$$

$$(\mathbf{X'X})^{-1} \quad = \quad \frac{1}{7(153.8905)-(30.33)^2} \begin{bmatrix} 153.8905 & -30.33 \\ -30.33 & 7 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \frac{1}{7(157.3246} \begin{bmatrix} 153.8905 & -30.33 \\ -30.33 & 7 \end{bmatrix} \begin{bmatrix} 350.61 \\ 1316.6558 \end{bmatrix} = \frac{1}{7(157.3246} \begin{bmatrix} 14021.3778 \\ -1417.4607 \end{bmatrix} = \begin{bmatrix} 89.129 \\ -9.0095 \end{bmatrix}$$

3.3 Fitted and Predicted Vectors

The vector of fitted (or predicted) values $\hat{\mathbf{Y}}$ is obtained as follows:

$$\hat{\mathbf{Y}}_{\mathbf{i}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 X_{11} + \cdots \hat{\beta}_p X_{1p} \\ \hat{\beta}_0 + \hat{\beta}_1 X_{21} + \cdots \hat{\beta}_p X_{2p} \\ \vdots \hat{\beta}_0 + \hat{\beta}_1 X_{n1} + \cdots \hat{\beta}_p X_{np} \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

$$= \mathbf{X}\,\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} = \mathbf{PY}$$

Here, $\mathbf{P}$ is the **projection** of **hat matrix**, and is of dimension $n \times n$. The hat matrix is symmetric and idempotent:

$$\mathbf{P} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} = (\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'})' = \mathbf{P'} \quad \Rightarrow \quad \text{Symmetric}$$

$$\mathbf{PP} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{P} \quad \Rightarrow \quad \text{Idempotent}$$

Example 1 – Pharmacodynamics of LSD

For the LSD concentration/math score example, we have the following hat matrix (generated in a computer matrix language):

$$\mathbf{P} = \begin{bmatrix} 1 & 1.17 \\ 1 & 2.97 \\ 1 & 3.26 \\ 1 & 4.69 \\ 1 & 5.83 \\ 1 & 6.00 \\ 1 & 6.41 \end{bmatrix} \frac{1}{7(157.3246} \begin{bmatrix} 153.8905 & -30.33 \\ -30.33 & 7 \end{bmatrix} \begin{bmatrix} 350.61 \\ 1316.6558 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1.17 & 2.97 & 3.26 & 4.69 & 5.83 & 6.00 & 6.41 \end{bmatrix} =$$

$$= \begin{bmatrix} 0.58796 & 0.33465 & 0.29384 & 0.09260 & -0.06783 & -0.09176 & -0.14946 \\ 0.33465 & 0.22550 & 0.20791 & 0.12120 & 0.05207 & 0.04176 & 0.01690 \\ 0.29384 & 0.20791 & 0.19407 & 0.12581 & 0.07139 & 0.06327 & 0.04370 \\ 0.09260 & 0.12120 & 0.12581 & 0.14853 & 0.16665 & 0.16935 & 0.17586 \\ -0.06783 & 0.05207 & 0.07139 & 0.16665 & 0.24259 & 0.25391 & 0.28122 \\ -0.09176 & 0.04176 & 0.06327 & 0.16935 & 0.25391 & 0.26652 & 0.29694 \\ -0.14946 & 0.01690 & 0.04370 & 0.17586 & 0.28122 & 0.29694 & 0.33483 \end{bmatrix}$$

The vector of residuals, e is the vector generated by elementwise subtraction between the data vector Y and the fitted vector $\hat{Y}$. It can be written as follows:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{PY} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$$

Also, note:

$$\hat{\mathbf{Y}} + \mathbf{e} = \mathbf{PY} + (\mathbf{I} - \mathbf{P})\mathbf{Y} = (\mathbf{P} + \mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{Y}$$

Example 1 – Pharmacodynamics of LSD

For the LSD concentration/math score example, we have the following fitted and residual vectors:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\boldsymbol{\beta}} = \begin{bmatrix} 89.12 - 9.01(1.17) = 78.58 \\ 89.12 - 9.01(2.97) = 62.36 \\ 89.12 - 9.01(3.26) = 59.75 \\ 89.12 - 9.01(4.69) = 46.86 \\ 89.12 - 9.01(5.83) = 36.59 \\ 89.12 - 9.01(6.00) = 35.06 \\ 89.12 - 9.01(6.41) = 31.37 \end{bmatrix} \qquad \mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} 78.93 - 78.58 = 0.35 \\ 58.20 - 62.36 = -4.16 \\ 67.47 - 59.75 = 7.72 \\ 37.47 - 46.86 = -9.39 \\ 45.65 - 36.59 = 9.06 \\ 32.92 - 35.06 = -2.14 \\ 29.97 - 31.37 = -1.40 \end{bmatrix}$$

3.4 Properties of Linear Functions of Random Vectors

Note that $\hat{\boldsymbol{\beta}}$, $\hat{Y}$, and e are all linear functions of the data vector $\mathbf{Y}$, and can be written as $\mathbf{AY}$:

- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \Rightarrow \quad \mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

- $\hat{\mathbf{Y}} = \mathbf{PY} \quad \Rightarrow \quad \mathbf{A} = \mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- $\mathbf{e} = (\mathbf{I} - \mathbf{P})\mathbf{Y} \quad \Rightarrow \quad \mathbf{A} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Consider a general vector $\mathbf{Z}$ that is of dimension 3×1. This can be easily expanded to $n \times 1$, but all useful results can be seen in the simpler case.

$$\mathbf{Z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

The **expectation vector** is the vector made up of the elementwise expected values of the elements of the random vector.

$$\mathbf{E}[\mathbf{Z}] = \begin{bmatrix} E(z_1) \\ E(z_2) \\ E(z_3) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \boldsymbol{\mu_z}$$

Note that the matrix $(\mathbf{Z} - \boldsymbol{\mu_z})(\mathbf{Z} - \boldsymbol{\mu_z})'$ is 3×3, and can be written as:

$$\begin{bmatrix} (z_1-\mu_1)^2 & (z_1-\mu_1)(z_2-\mu_2) & (z_1-\mu_1)(z_3-\mu_3) \\ (z_2-\mu_2)(z_1-\mu_1) & (z_2-\mu_2)^2 & (z_2-\mu_2)(z_3-\mu_3) \\ (z_3-\mu_3)(z_1-\mu_1) & (z_3-\mu_3)(z_2-\mu_2) & (z_3-\mu_3)^2 \end{bmatrix}$$

The **variance-covariance matrix** is the 3×3 matrix made up of variances (on the main diagonal) and the covariances (off diagonal) of the elements of $\mathbf{Z}$.

$$\mathbf{Var}[\mathbf{Z}] = \begin{bmatrix} Var(z_1) & Cov(z_1,z_2) & Cov(z_1,z_3) \\ Cov(z_2,z_1) & Var(z_2) & Cov(z_2,z_3) \\ Cov(z_3,z_1) & Cov(z_3,z_2) & Var(z_3) \end{bmatrix} = \mathbf{V_z} =$$

$$\begin{bmatrix} E[(z_1-\mu_1)^2] & E[(z_1-\mu_1)(z_2-\mu_2)] & E[(z_1-\mu_1)(z_3-\mu_3)] \\ E[(z_2-\mu_2)(z_1-\mu_1)] & E[(z_2-\mu_2)^2] & E[(z_2-\mu_2)(z_3-\mu_3)] \\ E[(z_3-\mu_3)(z_1-\mu_1)] & E[(z_3-\mu_3)(z_2-\mu_2)] & E[(z_3-\mu_3)^2] \end{bmatrix} =$$
$$= \mathbf{E}[(\mathbf{Z} - \boldsymbol{\mu_z})(\mathbf{Z} - \boldsymbol{\mu_z})'] = \mathbf{V_z}$$

Now let $\mathbf{A}$ be a $k \times n$ matrix of constants and $\mathbf{z}$ be a $n \times 1$ random vector with mean vector $\boldsymbol{\mu_z}$, and variance-covariance matrix $\mathbf{V_z}$. Suppose further that we can write $\mathbf{A}$ and $\mathbf{U} = \mathbf{Az}$ as follow:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a_1'} \\ \mathbf{a_2'} \\ \vdots \\ \mathbf{a_k'} \end{bmatrix}$$

$$\mathbf{U} = \mathbf{Az} = \begin{bmatrix} \mathbf{a_1'z} \\ \mathbf{a_2'z} \\ \vdots \\ \mathbf{a_k'z} \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_k \end{bmatrix}$$

where $\mathbf{a_i'}$ is a $1 \times n$ row vector of constants.

To obtain $\mathbf{E}[\mathbf{U}] = \boldsymbol{\mu_u}$, consider each element of $\mathbf{U}$, namely u_i and it's expectation $E(u_i)$.

$$E[u_i] = E[\mathbf{a_i'z}] = E[a_{i1}z_1 + a_{i2}z_2 + \cdots + a_{in}z_n] = a_{i1}E[z_1] + a_{i2}E[z_2] + \cdots + a_{in}E[z_n] = \mathbf{a_i'}\boldsymbol{\mu_z} \quad i = 1, \ldots, k$$

Piecing these together, we get:

$$\mathbf{E}[\mathbf{U}] = \begin{bmatrix} E[u_1] \\ E[u_2] \\ \vdots \\ E[u_k] \end{bmatrix} = \begin{bmatrix} \mathbf{a_1'}\boldsymbol{\mu_z} \\ \mathbf{a_2'}\boldsymbol{\mu_z} \\ \vdots \\ \mathbf{a_k'}\boldsymbol{\mu_z} \end{bmatrix} = \begin{bmatrix} \mathbf{a_1'} \\ \mathbf{a_2'} \\ \vdots \\ \mathbf{a_k'} \end{bmatrix} \boldsymbol{\mu_z} = \mathbf{A}\boldsymbol{\mu_z} = \boldsymbol{\mu_u}$$

To obtain the variance covariance matrix of $\mathbf{U} = \mathbf{Az}$, first consider the definition of $\mathbf{V}[\mathbf{U}]$, then write in terms of $\mathbf{U} = \mathbf{Az}$:

$$\mathbf{Var}[\mathbf{U}] = \mathbf{V_u} = \mathbf{E}[(\mathbf{U} - \boldsymbol{\mu_u})(\mathbf{U} - \boldsymbol{\mu_u})'] =$$

$$= \mathbf{E}[(\mathbf{Az} - \mathbf{A}\boldsymbol{\mu_z})(\mathbf{Az} - \mathbf{A}\boldsymbol{\mu_z})'] = \mathbf{E}\{[\mathbf{A}(\mathbf{z} - \boldsymbol{\mu_z})][\mathbf{A}(\mathbf{z} - \boldsymbol{\mu_z})]'\} =$$

$$= \mathbf{E}[\mathbf{A}(\mathbf{z} - \boldsymbol{\mu_z})(\mathbf{z} - \boldsymbol{\mu_z})'\mathbf{A}'] = \mathbf{AE}[(\mathbf{z} - \boldsymbol{\mu_z})(\mathbf{z} - \boldsymbol{\mu_z})]\mathbf{A}' = \mathbf{AV_zA}'$$

Note that if $\mathbf{V_z} = \sigma^2\mathbf{I}$, then $\mathbf{V_u} = \sigma^2\mathbf{AA}'$.

3.5 Applications of Linear Functions of Random Variables

In this section, we consider two applications, each assuming independent observations ($Cov[Y_i, Y_j] = 0 \quad i \neq j$).

Case 1 – Sampling from a single population

$$E[Y_i] = \mu \quad i = 1, \ldots, n \qquad Var[Y_i] = \sigma^2 \quad i = 1, \ldots, n$$

Let $\mathbf{Y}$ be the $n \times 1$ vector made up of elements $Y_1, \ldots, Y_n$. Then:

$$\mathbf{E}[\mathbf{Y}] = \begin{bmatrix} E[Y_1] \\ E[Y_2] \\ \vdots \\ E[Y_n] \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} = \mu\mathbf{1}$$

where 1 is a $n \times 1$ column vector of $1's$.

$$\mathbf{Var}(\mathbf{Y}) = \begin{bmatrix} Var[Y_1] & Cov[Y_1, Y_2] & \cdots & Cov[Y_1, Y_n] \\ Cov[Y_2, Y_1] & Var[Y_2] & \cdots & Cov[Y_2, Y_n] \\ \vdots & \vdots & \ddots & \vdots \\ Cov[Y_n, Y_1] & Cov[Y_n, Y_2] & \cdots & Var[Y_n] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Now consider the estimator $\overline{Y}$:

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \mathbf{Y} = \mathbf{a}'\mathbf{Y}$$

Now we can obtain the mean and variance of $\overline{Y}$ from these rules, with $\mathbf{a}' = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}$:

$$E[\overline{Y}] = \mathbf{a}'\mathbf{E}[\mathbf{Y}] = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix} = \mathbf{a}`\mu\mathbf{1} = \sum_{i=1}^{n} \left(\frac{1}{n}\right) \mu = n \left(\frac{1}{n}\right) \mu = \mu$$

$$Var[\overline{Y}] = \mathbf{a}'\mathbf{Var}[\mathbf{Y}]\mathbf{a} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \sigma^2 \mathbf{I} \begin{bmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \cdots \\ \frac{1}{n} \end{bmatrix} = \sigma^2 \mathbf{a}'\mathbf{a} = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 = \sigma^2 n \left(\frac{1}{n}\right)^2 = \frac{\sigma^2}{n}$$

Case 2 – Multiple Linear Regression Model

$$E[Y_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \quad i = 1, \ldots, n \qquad Var[Y_i] = \sigma^2 \quad i = 1, \ldots, n$$

Let $\mathbf{Y}$ be the $n \times 1$ vector made up of elements $Y_1, \ldots, Y_n$. Then:

$$\mathbf{E}[\mathbf{Y}] = \begin{bmatrix} E[Y_1] \\ E[Y_2] \\ \vdots \\ E[Y_n] \end{bmatrix} = \mathbf{X}\ \boldsymbol{\beta}$$

$$\mathbf{Var}[\mathbf{Y}] \quad = \quad \sigma^2 \mathbf{I}$$

Now consider the least squares estimator $\hat{\boldsymbol{\beta}}$ of the regression coefficient parameter vector $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}'\mathbf{Y} \quad \Rightarrow \quad \mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The mean and variance of $\hat{\boldsymbol{\beta}}$ are:

$$\mathbf{E}[\ \hat{\boldsymbol{\beta}}] = \mathbf{E}[\mathbf{A}'\mathbf{Y}] = \mathbf{A}'\mathbf{E}[\mathbf{Y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\ \boldsymbol{\beta} = \ \boldsymbol{\beta}$$

$$\mathbf{Var}[\ \hat{\boldsymbol{\beta}}] = \mathbf{Var}[\mathbf{A}'\mathbf{Y}] = \mathbf{A}'\mathbf{Var}[\mathbf{Y}]\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^{\mathbf{2}}\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^{\mathbf{2}}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1} = \sigma^{\mathbf{2}}(\mathbf{X}'\mathbf{X})^{-1}$$

3.6 Multivariate Normal Distribution

Suppose a $n \times 1$ random vector $\mathbf{Z}$ has a multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. This would occur if we generated n independent standard normal random variables and put them together in vector form. The density function for $\mathbf{Z}$, evaluated any fixed point $\mathbf{z}$ is:

$$\mathbf{Z} \sim NID(\mathbf{0}, \sigma^2\mathbf{I}) \quad \Rightarrow \quad f_{\mathbf{Z}}(\mathbf{z}) = (2\pi)^{-n/2}|\sigma^2 I|^{-1/2} exp\{-\frac{1}{2}\mathbf{z}'(\boldsymbol{\sigma^2}\mathbf{I})^{-\mathbf{1}}\mathbf{z}\}$$

More generally, let $\mathbf{U} = \mathbf{AZ} + \mathbf{b}$ with $\mathbf{A}$ being a $k \times n$ matrix of constants, and $\mathbf{b}$ being a $n \times 1$ vector of constants. Then:

$$\mathbf{E}[\mathbf{U}] = \mathbf{AE}[\mathbf{Z}] + \mathbf{b} = \mathbf{b} = \boldsymbol{\mu_U} \qquad \mathbf{V}[\mathbf{U}] = \mathbf{AV}[\mathbf{Z}]\mathbf{A}' = \boldsymbol{\sigma^2}\mathbf{AA}' = \mathbf{V_U}$$

The density function for $\mathbf{U}$, evaluated any fixed point $\mathbf{u}$ is:

$$\mathbf{U} \sim NID(\boldsymbol{\mu_U}, \sigma^2\mathbf{V_U}) \quad \Rightarrow \quad f_{\mathbf{U}}(\mathbf{u}) = (2\pi)^{-k/2}|V_U|^{-1/2} exp\{-\frac{1}{2}(\mathbf{u} - \boldsymbol{\mu_U})'(\mathbf{V_U})^{-\mathbf{1}}(\mathbf{u} - \boldsymbol{\mu_U})\}$$

That is, any linear function of a normal random vector is normal.

3.6.1 Properties of Regression Estimates

Under the traditional normal theory linear regression model, we have:

$$\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \ \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\sigma^2}\mathbf{I}) \quad \Rightarrow \quad \mathbf{Y} \sim \mathbf{N}(\mathbf{X}\ \boldsymbol{\beta}, \boldsymbol{\sigma^2}\mathbf{I})$$

Then the density function of $\mathbf{Y}$ evaluated is:

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-n/2}\sigma^{-n} exp\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\ \boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\ \boldsymbol{\beta})\}$$

Assuming the model is correct, we've already obtained the mean and variance of $\hat{\boldsymbol{\beta}}$. We further know that its distibution is multivariate normal: $\hat{\boldsymbol{\beta}} \sim N(\beta, \boldsymbol{\sigma^2}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}})$.

The vector of fitted values $\hat{\mathbf{Y}} = \mathbf{X}\ \hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} = \mathbf{PY}$ is also a linear function of $\mathbf{Y}$ and thus also normally distributed with the following mean vector and variance-covariance matrix:

$$\mathbf{E}[\hat{\mathbf{Y}}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{E}[\mathbf{Y}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}\ \boldsymbol{\beta} = \mathbf{X}\ \boldsymbol{\beta}$$

$$\mathbf{Var}[\hat{\mathbf{Y}}] = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\boldsymbol{\sigma^2}\mathbf{IX}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \boldsymbol{\sigma^2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \boldsymbol{\sigma^2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' = \boldsymbol{\sigma^2}\mathbf{P}$$

That is $\hat{\mathbf{Y}} \sim N(\mathbf{X}\ \boldsymbol{\beta}, \boldsymbol{\sigma^2}\mathbf{P})$.

The vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ is also normal with mean vector and variance-covariance matrix:

$$\mathbf{E}[\mathbf{e}] = (\mathbf{I} - \mathbf{P})\mathbf{X}\ \boldsymbol{\beta} = (\mathbf{X} - \mathbf{P}\mathbf{X})\ \boldsymbol{\beta} = (\mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X})\ \boldsymbol{\beta} = (\mathbf{X} - \mathbf{X})\ \boldsymbol{\beta} = \mathbf{0}$$

$$\mathbf{Var}[\mathbf{e}] = (\mathbf{I} - \mathbf{P})\sigma^{\mathbf{2}}\mathbf{I}(\mathbf{I} - \mathbf{P})' = \sigma^{\mathbf{2}}(\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P})' = \sigma^{\mathbf{2}}(\mathbf{I} - \mathbf{P})(\mathbf{I} - \mathbf{P}) = \sigma^{\mathbf{2}}(\mathbf{I} - \mathbf{P})$$

Note the differences between the distributions of $\boldsymbol{\varepsilon}$ and $\mathbf{e}$:

$$\boldsymbol{\varepsilon} \sim N(\ \mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I}) \quad \Rightarrow \quad \mathbf{e} \sim N(\ \mathbf{0}, \sigma^{\mathbf{2}}(\mathbf{I} - \mathbf{P}))$$

Often the goal is to predict a future outcome when the set of independent levels are at a given setting, $\mathbf{x_0'} = \begin{bmatrix} 1 & x_{01} & \cdots & x_{0p} \end{bmatrix}$. The future observation Y_0 and its predicted value based on the estimated regression equation are:

$$Y_0 = \mathbf{x_0'}\ \boldsymbol{\beta} + \varepsilon_0 \qquad \varepsilon_0 \sim NID(0, \sigma^2)$$

$$\hat{Y}_0 = \mathbf{x_0'}\ \hat{\boldsymbol{\beta}} \qquad \hat{Y}_0 \sim N(\mathbf{x_0'}\ \boldsymbol{\beta}, \sigma^2\mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x_0})$$

It is assumed $\varepsilon_0 \sim N(0, \sigma^2)$ and is independent from the errors in the observations used to fit the regression model $(\varepsilon_1, \ldots, \varepsilon_n)$.

The prediction error is:

$$Y_0 - \hat{Y}_0 = \mathbf{x_0'}\ \boldsymbol{\beta} + \varepsilon_0 - \mathbf{x_0'}\ \hat{\boldsymbol{\beta}} = \mathbf{x_0'}(\ \boldsymbol{\beta} -\ \hat{\boldsymbol{\beta}}) + \varepsilon_0$$

which is normal with mean and variance:

$$E[Y_0 - \hat{Y}_0] = \mathbf{x_0'}(\ \boldsymbol{\beta} - \mathbf{E}[\ \hat{\boldsymbol{\beta}}]) + E[\varepsilon_0] = \mathbf{x_0'}(\ \boldsymbol{\beta} -\ \boldsymbol{\beta}) + 0 = 0$$

$$V[Y_0 - \hat{Y}_0] = V[Y_0] + V[\hat{Y}_0] = \sigma^2 + \sigma^2\mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x_0} = \sigma^2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x_0}]$$

and we have that:

$$Y_0 - \hat{Y}_0 \sim N(0, \sigma^2[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{x_0}])$$

4 Analysis of Variance and Quadratic Forms

The sums of square in the Analysis of Variance can be written as **quadratic forms** in $\mathbf{Y}$. The form we use is $\boldsymbol{Y'AY}$ where $\mathbf{A}$ is a matrix of coefficients, referred to as the **defining matrix**.

The following facts are important and particularly useful in regression models (for a very detailed discussion, see *Linear Models* (1971), by S.R. Searle).

1. Any sum of squares can be written as $\boldsymbol{Y'AY}$ where $\mathbf{A}$ is a square, symmetric nonnegative definite matrix
2. The degrees of freedom associated with any quadratic form is equal to the rank of the defining matrix, which is equal to its trace when the defining matrix is idempotent.
3. Two quadratic forms are orthogonal if the product of their defining matices is $\mathbf{0}$

4.1 The Analysis of Variance

Now consider the Analysis of Variance.

$$\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{e} \quad \mathbf{Y'Y} = \sum_{i=1}^{n} Y_i^2 = SS(\text{TOTAL UNCORRECTED})$$

Note that $\mathbf{Y'Y} = \mathbf{Y'IY}$, so that $\mathbf{I}$ is the defining matrix, which is symmetric and idempotent. The degrees of freedom for SS(TOTAL UNCORRECTED) is then the rank of $\mathbf{I}$, which is its trace, or n.

Now, we decompose the Total uncorrected sum of squares into it's model and error components.

$$\mathbf{Y'Y} = (\hat{\mathbf{Y}} + \mathbf{e})'(\hat{\mathbf{Y}} + \mathbf{e}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{Y}}'\mathbf{e} + \mathbf{e}'\hat{\mathbf{Y}} + \mathbf{e'e} =$$

$$= (\mathbf{PY})'(\mathbf{PY}) + (\mathbf{PY})'(\mathbf{I}-\mathbf{P})\mathbf{Y} + [(\mathbf{I}-\mathbf{P})\mathbf{Y}]'\mathbf{PY} + [(\mathbf{I}-\mathbf{P})\mathbf{Y}]'[(\mathbf{I}-\mathbf{P})\mathbf{Y}] =$$

$$= \mathbf{Y'P'PY} + \mathbf{Y'P'}(\mathbf{I}-\mathbf{P})\mathbf{Y} + \mathbf{Y'}(\mathbf{I}-\mathbf{P})'\mathbf{PY} + \mathbf{Y'}(\mathbf{I}-\mathbf{P})'(\mathbf{I}-\mathbf{P})\mathbf{Y} =$$

$$= \mathbf{Y'PPY} + (\mathbf{Y'PY} - \mathbf{Y'PPY}) + (\mathbf{Y'PY} - \mathbf{Y'PPY}) + (\mathbf{Y'IY} - \mathbf{Y'IPY} - \mathbf{Y'PIY} + \mathbf{Y'PPY}) =$$

$$= \mathbf{Y'PY} + (\mathbf{Y'PY} - \mathbf{Y'PY}) + (\mathbf{Y'PY} - \mathbf{Y'PY}) + (\mathbf{Y'Y} - \mathbf{Y'PY} - \mathbf{Y'PY} + \mathbf{Y'PY}) =$$

$$= \mathbf{Y'PY} + \mathbf{0} + \mathbf{0} + (\mathbf{Y'Y} - \mathbf{Y'PY}) = \mathbf{Y'PY} + \mathbf{Y'}(\mathbf{I}-\mathbf{P})\mathbf{Y} =$$

$$= \mathbf{Y'P'PY} + \mathbf{Y'}(\mathbf{I}-\mathbf{P})'(\mathbf{I}-\mathbf{P})\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \mathbf{e'e}$$

We obtain the degrees of freedom as follow, making use of the following identities regarding the trace of matrices:

$$tr(\mathbf{AB}) = tr(\mathbf{BA}) \qquad tr(\mathbf{A}+\mathbf{B}) = tr(\mathbf{A}) + tr(\mathbf{B})$$

$$SS(\text{MODEL}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y'PY}$$

$$df(\text{MODEL}) = tr(\mathbf{P}) = tr(\mathbf{X}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}) = tr((\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'X}) = tr(\mathbf{I_{p'}}) = p' = p+1$$

$$SS(\text{RESIDUAL}) = \mathbf{e'e} = \mathbf{Y'}(\mathbf{I}-\mathbf{P})\mathbf{Y}$$

$$df(\text{RESIDUAL}) = tr(\mathbf{I}-\mathbf{P}) = tr(\mathbf{I_n}) - tr(\mathbf{P}) = n - p' = n - p - 1$$

Table 8 gives the Analysis of Variance including degrees of freedom, and sums of squares (both definitional and computational forms).

Source of Variation	Degrees of Freedom	Sum of Squares Definitional	Sum of Squares Computational
TOTAL(UNCORRECTED)	n	$\mathbf{Y}'\mathbf{Y}$	$\mathbf{Y}'\mathbf{Y}$
MODEL	$p' = p+1$	$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{P}\mathbf{Y}$	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$
ERROR	$n - p'$	$\mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$	$\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$

Table 8: Analysis of Variance in Matrix form

Example 1 – Pharmacodynamics of LSD

We obtain the Analysis of Variance in matrix form:

$$SS(\text{TOTAL UNCORRECTED}) = \mathbf{Y}'\mathbf{Y} = \begin{bmatrix} 78.93 & 58.20 & 67.47 & 37.47 & 45.65 & 32.92 & 29.97 \end{bmatrix} \begin{bmatrix} 78.93 \\ 58.20 \\ 67.47 \\ 37.47 \\ 45.65 \\ 32.92 \\ 29.97 \end{bmatrix} =$$

$$= \sum_{i=1}^{n} Y_i^2 = 19639.2365 \quad df(\text{TOTAL UNCORRECTED}) = n = 7$$

$$SS(\text{MODEL}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 89.1239 & -9.0095 \end{bmatrix} \begin{bmatrix} 350.61 \\ 1316.6558 \end{bmatrix} =$$

$$89.1239(350.61) + (-9.0095)(1316.6558) = 19385.3201 \quad df(\text{MODEL}) = p' = 2$$

$$SS(\text{RESIDUAL}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = 19639.2365 - 19385.3201 = 253.9164 \quad df(\text{RESIDUAL}) = n - p' = 7 - 2 = 5$$

The total uncorrected sum of squares represents variation (in Y) around 0. We are usually interested in variation around the sample mean $\overline{Y}$. We will partition the model sum of squares into two components: $SS(\text{REGRESSION})$ and $SS(\mu)$. The first sum of squares is associated with β_1 and the second one is associated with β_0.

Model with only the mean $\mu = \beta_0 \quad (\beta_1 = 0)$

Consider the following model, we obtain the least squares estimates and model sum of squares.

$$Y_i = \beta_0 + \varepsilon_i = \mu + \varepsilon_i$$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbf{1} \quad \boldsymbol{\beta} = [\mu] = [\beta_0]$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = (\mathbf{1'1})^{-1}\mathbf{1'Y} \qquad \mathbf{1'1} = n \quad \mathbf{1'Y} = \sum_{i=1}^{n} Y_i$$

$$\Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{1'1})^{-1}\mathbf{1'Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \overline{Y}$$

$$SS(\mu) = \ \hat{\boldsymbol{\beta}}'\mathbf{X'Y} = \overline{Y}(\sum_{i=1}^{n} Y_i) = n\overline{Y}^2$$

$$= \mathbf{Y'1}(\mathbf{1'1})^{-1}\mathbf{1'Y} = \mathbf{Y}'(\frac{1}{n}\mathbf{11'})\mathbf{Y} =$$

$$\mathbf{Y}' \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \mathbf{Y} = \mathbf{Y}'(\frac{1}{n}\mathbf{J})\mathbf{Y}$$

where $\mathbf{J}$ is a $n \times n$ matrix of 1^s. Note that $(1/n)\mathbf{J}$ is an idempotent matrix:

$$(\frac{1}{n}\mathbf{J})(\frac{1}{n}\mathbf{J}) = \left(\frac{1}{n}\right)^2 \mathbf{JJ} = \left(\frac{1}{n}\right)^2 \begin{bmatrix} n & n & \cdots & n \\ n & n & \cdots & n \\ \vdots & \vdots & \ddots & \vdots \\ n & n & \cdots & n \end{bmatrix} = \frac{1}{n}\mathbf{J}$$

By subtraction, we get $SS(\text{REGRESSION}) = SS(\text{MODEL}) - SS(\mu)$:

$$SS(\text{REGRESSION}) = SS(\text{MODEL}) - SS(\mu) = \mathbf{Y'PY} - \mathbf{Y}'(\frac{1}{n}\mathbf{J})\mathbf{Y} = \mathbf{Y}'(\mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{Y}$$

To demonstrate that the defining matrix for $SS(\text{REGRESSION})$ is idempotent and that the three sum of squares are orthogonal, consider the following algebra where $\mathbf{X}^*$ is the matrix made up of the columns of $\mathbf{X}$ associated with the p independent variables and not the column for the intercept.

$$\mathbf{X} = [\mathbf{1}|\mathbf{X}^*] \qquad \mathbf{PX} = \mathbf{P}[\mathbf{1}|\mathbf{X}^*] = \mathbf{X} = [\mathbf{1}|\mathbf{X}^*]$$

$$\Rightarrow \quad \mathbf{P1} = \mathbf{1} \quad \Rightarrow \quad \mathbf{PJ} = \mathbf{J}$$

$$\mathbf{X}' = [\mathbf{1}|\mathbf{X}^*]' \qquad \mathbf{X'P} = [\mathbf{1}|\mathbf{X}^*]'\mathbf{P} = \mathbf{X}' = [\mathbf{1}|\mathbf{X}^*]'$$

$$\Rightarrow \quad \mathbf{1'P} = \mathbf{1}' \quad \Rightarrow \quad \mathbf{JP} = \mathbf{J}$$

$$(\mathbf{P}-\frac{1}{n}\mathbf{J})(\mathbf{P}-\frac{1}{n}\mathbf{J}) = \mathbf{PP} - \mathbf{P}(\frac{1}{n}\mathbf{J}) - \frac{1}{n}\mathbf{JP} + (\frac{1}{n}\mathbf{J})(\frac{1}{n}\mathbf{J}) = \mathbf{P} - (\frac{1}{n}\mathbf{J}) - (\frac{1}{n}\mathbf{J}) + (\frac{1}{n}\mathbf{J}) = \mathbf{P}-\frac{1}{n}\mathbf{J}$$

Summarizing what we have obtained so far (where all defining matrices are idempotent):

$$SS(\text{TOTAL UNCORRECTED}) = \mathbf{Y'IY} = \mathbf{Y'Y} \quad df(\text{TOTAL UNCORRECTED}) = tr(\mathbf{I_n}) = n$$

$$SS(\mu) = \mathbf{Y}'(\frac{1}{n}\mathbf{J})\mathbf{Y} \quad df(\mu) = tr(\frac{1}{n}\mathbf{J}) = \frac{1}{n}(n) = 1$$

$$SS(\text{REGRESSION}) = \mathbf{Y}'(\mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{Y} \quad df(\text{REGRESSION}) = tr((\mathbf{P}-\frac{1}{n}\mathbf{J}) = tr(\mathbf{P})-tr(\frac{1}{n}\mathbf{J}) = p'-1 = p+1-1 = p$$

$$SS(\text{RESIDUAL}) = \mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y} \quad df(\text{RESIDUAL}) = tr(\mathbf{I}-\mathbf{P}) = tr(\mathbf{I}) - tr(\mathbf{P}) = n - p'$$

To show that the sums of squares for the mean, regression, and residual are pairwise orthogonal, consider the products of their defining matrices: First for $SS(\mu)$ and $SS(\text{REGRESSION})$:

$$(\frac{1}{n}\mathbf{J})(\mathbf{P}-\frac{1}{n}\mathbf{J}) = \frac{1}{n}\mathbf{JP} - (\frac{1}{n}\mathbf{J})(\frac{1}{n}\mathbf{J}) = \frac{1}{n}\mathbf{J} - \frac{1}{n}\mathbf{J} = \mathbf{0}$$

Next for $SS(\mu)$ and $SS(\text{RESIDUAL})$:

$$(\frac{1}{n}\mathbf{J})(\mathbf{I}-\mathbf{P}) = \frac{1}{n}\mathbf{JI} - \frac{1}{n}\mathbf{JP} = \frac{1}{n}\mathbf{J} - \frac{1}{n}\mathbf{J} = \mathbf{0}$$

Finally for $SS(\text{REGRESSION})$ and $SS(\text{RESIDUAL})$:

$$(\mathbf{P}-\frac{1}{n}\mathbf{J})(\mathbf{I}-\mathbf{P}) = \mathbf{PI} - \mathbf{PP} - \frac{1}{n}\mathbf{JI} + \frac{1}{n}\mathbf{JP} = \mathbf{P} - \mathbf{P} - \frac{1}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} = \mathbf{0}$$

Example 1 – Pharmacodynamics of LSD

For the LSD concentration/math score example, we have the ANOVA in Table 9.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
TOTAL(UNCORRECTED)	7	19639.24	—
MEAN	1	17561.02	—
TOTAL (CORRECTED)	6	2078.22	—
REGRESSION	1	1824.30	1824.30
RESIDUAL	5	253.92	50.78

Table 9: Analysis of Variance for LSD data

A summary of key points regarding quadratic forms:

- The rank, $r(\mathbf{X})$ is the number of linearly independent columns in $\mathbf{X}$
- The model is **full rank** if $r(\mathbf{X}) = p'$ assuming $n > p'$
- A unique least squares solution exists iff the model is full rank.

- All defining matrices in the Analysis of Variance are idempotent.
- The defining matrices for the mean, regression, and residual are pairwise orthogonal and sum to $\mathbf{I}$. Thus they partition the total uncorrected sum of squares into orthogonal sums of squares.
- Degrees of freedom for quadratic forms are the ranks of their defining matrices; when idempotent, the trace of a matrix is its rank.

4.2 Expectations of Quadratic Forms

In this section we obtain the expectations of the sums of squares in the Analysis of Variance, making use of general results in quadratic forms. The proofs are given in Searle (1971). Suppose we have a random vector $\mathbf{Y}$ with the following mean vector and variance-covariance matrix:

$$\mathbf{E}[\mathbf{Y}] = \boldsymbol{\mu} \qquad \mathbf{Var}[\mathbf{Y}] = \mathbf{V_Y} = \mathbf{V}\sigma^2$$

Then, the expectation of a quadratic form $\mathbf{Y'AY}$ is:

$$\mathbf{E}[\mathbf{Y'AY}] = tr(\mathbf{AV_Y}) + \boldsymbol{\mu}'\mathbf{A}\ \boldsymbol{\mu} = \sigma^2 tr(\mathbf{AV}) + \boldsymbol{\mu}'\mathbf{A}\ \boldsymbol{\mu}$$

Under the ordinary least squares assumptions, we have:

$$\mathbf{E}[\mathbf{Y}] = \mathbf{X}\ \boldsymbol{\beta} \qquad \mathbf{Var}[\mathbf{Y}] = \sigma^2\mathbf{I_n}$$

Source of Variation	"$\mathbf{A}$" Matrix
TOTAL UNCORRECTED	$\mathbf{I}$
MODEL	$\mathbf{P} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$
REGRESSION	$\mathbf{P} - \frac{1}{n}\mathbf{J}$
RESIDUAL	$\mathbf{I} - \mathbf{P}$

Now applying the rules on expectations of quadratic forms:

$$E[SS(\text{MODEL})] = \mathbf{E}[\mathbf{Y'PY}] = \sigma^2 tr(\mathbf{PI}) + \boldsymbol{\beta}'\mathbf{X'PX}\ \boldsymbol{\beta} =$$

$$= \sigma^2 tr(\mathbf{P}) + \boldsymbol{\beta}'\mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'X}\ \boldsymbol{\beta} = \sigma^2 p' + \boldsymbol{\beta}\mathbf{X'X}\ \boldsymbol{\beta}$$

$$E[SS(\text{REGRESSION})] = \mathbf{E}[\mathbf{Y'}(\mathbf{P} - \frac{1}{n}\mathbf{J})\mathbf{Y}] = \sigma^2 tr(\mathbf{P} - \frac{1}{n}\mathbf{J}) + \boldsymbol{\beta}'\mathbf{X'}(\mathbf{P} - \frac{1}{n}\mathbf{J})\mathbf{X}\ \boldsymbol{\beta} =$$

$$\sigma^2(p'-1) + \boldsymbol{\beta}'\mathbf{X'X}\ \boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X'}\frac{1}{n}\mathbf{JX}\ \boldsymbol{\beta} = p\sigma^2 + \boldsymbol{\beta}'\mathbf{X'}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{X}\ \boldsymbol{\beta}$$

This last matrix can be seen to involve the regression coefficients: $\beta_1, \ldots, \beta_p$, and not β_0 as follows:

$$\mathbf{X'}(\mathbf{I} - \frac{1}{n}\mathbf{J})\mathbf{X} = \mathbf{X'X} - \mathbf{X'}\frac{1}{n}\mathbf{JX}$$

$$\mathbf{X'X} = \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \cdots & \sum_{i=1}^n X_{i1}X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \sum_{i=1}^n X_{ip}X_{i1} & \cdots & \sum_{i=1}^n X_{ip}^2 \end{bmatrix}$$

$$\frac{1}{n}\mathbf{X}'\mathbf{J}\mathbf{X} = \frac{1}{n}\begin{bmatrix} n & n & \cdots & n \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{i1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \sum_{i=1}^n X_{ip} & \cdots & \sum_{i=1}^n X_{ip} \end{bmatrix}\mathbf{X} =$$

$$= \frac{1}{n}\begin{bmatrix} n^2 & n\sum_{i=1}^n X_{i1} & \cdots & n\sum_{i=1}^n X_{ip} \\ n\sum_{i=1}^n X_{i1} & \left(\sum_{i=1}^n X_{i1}\right)^2 & \cdots & \left(\sum_{i=1}^n X_{i1}\right)\left(\sum_{i=1}^n X_{ip}\right) \\ \vdots & \vdots & \ddots & \vdots \\ n\sum_{i=1}^n X_{ip} & \left(\sum_{i=1}^n X_{ip}\right)\left(\sum_{i=1}^n X_{i1}\right) & \cdots & \left(\sum_{i=1}^n X_{ip}\right)^2 \end{bmatrix} =$$

$$= \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \frac{\left(\sum_{i=1}^n X_{i1}\right)^2}{n} & \cdots & \frac{\left(\sum_{i=1}^n X_{i1}\right)\left(\sum_{i=1}^n X_{ip}\right)}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \frac{\left(\sum_{i=1}^n X_{ip}\right)\left(\sum_{i=1}^n X_{i1}\right)}{n} & \cdots & \frac{\left(\sum_{i=1}^n X_{ip}\right)^2}{n} \end{bmatrix}$$

$$\Rightarrow \quad \mathbf{X}'(\mathbf{I}-\frac{1}{n}\mathbf{J})\mathbf{X} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X} =$$

$$= \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \sum_{i=1}^n X_{i1}^2 & \cdots & \sum_{i=1}^n X_{i1}X_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \sum_{i=1}^n X_{ip}X_{i1} & \cdots & \sum_{i=1}^n X_{ip}^2 \end{bmatrix} - \begin{bmatrix} n & \sum_{i=1}^n X_{i1} & \cdots & \sum_{i=1}^n X_{ip} \\ \sum_{i=1}^n X_{i1} & \frac{\left(\sum_{i=1}^n X_{i1}\right)^2}{n} & \cdots & \frac{\left(\sum_{i=1}^n X_{i1}\right)\left(\sum_{i=1}^n X_{ip}\right)}{n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{ip} & \frac{\left(\sum_{i=1}^n X_{ip}\right)\left(\sum_{i=1}^n X_{i1}\right)}{n} & \cdots & \frac{\left(\sum_{i=1}^n X_{ip}\right)^2}{n} \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n X_{i1}^2 - \frac{\left(\sum_{i=1}^n X_{i1}\right)^2}{n} & \cdots & \sum_{i=1}^n X_{i1}X_{ip} - \frac{\left(\sum_{i=1}^n X_{i1}\right)\left(\sum_{i=1}^n X_{ip}\right)}{n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \sum_{i=1}^n X_{ip}X_{i1} - \frac{\left(\sum_{i=1}^n X_{ip}\right)\left(\sum_{i=1}^n X_{i1}\right)}{n} & \cdots & \sum_{i=1}^n X_{ip}^2 - \frac{\left(\sum_{i=1}^n X_{ip}\right)^2}{n} \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \sum_{i=1}^n (X_{i1}-\overline{X}_1)^2 & \cdots & \sum_{i=1}^n (X_{i1}-\overline{X}_1)(X_{ip}-\overline{X}_p) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \sum_{i=1}^n (X_{ip}-\overline{X}_p)(X_{i1}-\overline{X}_1) & \cdots & \sum_{i=1}^n (X_{ip}-\overline{X}_p)^2 \end{bmatrix} =$$

Thus, $E[SS(\text{REGRESSION})]$ involves a quadratic form in $\beta_1, \ldots, \beta_p$, and not in β_0 since the first row and column of the previous matrix is made up of 0^s. Now, we return to $E[SS(\text{RESIDUAL})]$:

$$E[SS(\text{RESIDUAL})] = \mathbf{E}[\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}] = \sigma^2 tr(\mathbf{I}-\mathbf{P}) + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\ \boldsymbol{\beta} =$$

$$\sigma^2(n-p') + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\ \boldsymbol{\beta} - \boldsymbol{\beta}\mathbf{X}'\mathbf{P}\mathbf{X}\ \boldsymbol{\beta} = \sigma^2(n-p') + \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\ \boldsymbol{\beta} - \boldsymbol{\beta}\mathbf{X}'\mathbf{X}\ \boldsymbol{\beta} = \sigma^2(n-p')$$

Now we can obtain the expected values of the mean squares from the Analysis of Variance:

$$MS(\text{REGRESSION}) = \frac{SS(\text{REGRESSION})}{p} \quad \Rightarrow E[MS(\text{REGRESSION})] = \sigma^2 + \frac{1}{p}\ \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\frac{1}{n}\mathbf{J})\mathbf{X}\ \boldsymbol{\beta}$$

$$MS(\text{RESIDUAL} = \frac{SS(\text{RESIDUAL})}{n-p'} \quad \Rightarrow E[MS(\text{RESIDUAL})] = \sigma^2$$

Note that the second term in $E[MS(\text{REGRESSION})]$ is a quadratic form in $\boldsymbol{\beta}$, if any $\beta_i \neq 0$ $(i = 1, \ldots, p)$, then $E[MS(\text{REGRESSION})] > E[MS(\text{RESIDUAL})]$, otherwise they are equal.

4.2.1 The Case of Misspecified Models

The above statements presume that the model is correctly specified. Suppose:

$$\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \mathbf{Z}\ \boldsymbol{\gamma} + \ \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$$

but we fit the model:

$$\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \ \boldsymbol{\varepsilon}$$

Then $E[MS(\text{RESIDUAL})]$ can be written as:

$$E[SS(\text{RESIDUAL})] = \sigma^2 tr(\mathbf{I}-\mathbf{P}) + (\mathbf{X}\ \boldsymbol{\beta} + \mathbf{Z}\ \boldsymbol{\gamma})'(\mathbf{I}-\mathbf{P})(\mathbf{X}\ \boldsymbol{\beta} + \mathbf{Z}\ \boldsymbol{\gamma}) =$$

$$= \sigma^2(n-p') + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\ \boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{Z}\ \boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{X}\ \boldsymbol{\beta} + \boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}\ \boldsymbol{\gamma} =$$

$$= \sigma^2(n-p') + 0 + 0 + 0 + \boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}\ \boldsymbol{\gamma} \qquad \text{since } \mathbf{X}'(\mathbf{I}-\mathbf{P}) = (\mathbf{I}-\mathbf{P})\mathbf{X} = \mathbf{0}$$

$$E[MS(\text{RESIDUAL})] = \frac{E[SS(\text{RESIDUAL})]}{n-p'} = \sigma^2 + \frac{1}{n-p'}\ \boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}-\mathbf{P})\mathbf{Z}\ \boldsymbol{\gamma}$$

which is larger than σ^2 if the elements of $\boldsymbol{\gamma}$ are not all equal to 0 (which would make our fitted model correct).

Estimator	Theoretical Variance	Estimated Variance
$\hat{\boldsymbol{\beta}}$	$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	$s^2(\mathbf{X}'\mathbf{X})^{-1}$
$\hat{\mathbf{Y}}$	$\sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2\mathbf{P}$	$s^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = s^2\mathbf{P}$
$\mathbf{e}$	$\sigma^2(\mathbf{I}-\mathbf{P})$	$s^2(\mathbf{I}-\mathbf{P})$

Table 10: Theoretical and estimated variances of regression estimators in Matrix form

4.2.2 Estimated Variances

Recall the variance-covariance matrices of $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{Y}}$, and $\mathbf{e}$. Each of these depended on σ^2, which is in practice unknown. Unbiased estimators of these variances can be obtained by replacing σ^2 with an unbiased estimate:

$$s^2 = MS(\text{RESIDUAL}) = \frac{1}{n-p'}\mathbf{Y}'(\mathbf{I}-\mathbf{P})\mathbf{Y}$$

assuming the model is correct. Table 10 gives the true and estimated variances for these estimators.

4.3 Distribution of Quadratic Forms

We have obtained means of quadratic forms, but need their distributions to make statistical inferences. The assumptions for the traditional inferences to be made is that $\boldsymbol{\varepsilon}$ and $\mathbf{Y}$ are normally distributed, otherwise tests are approximate.

The following results are referred to as Cochran's Theorem, see Searle (1971) for proofs. Suppose $\mathbf{Y}$ is distributed as follows with nonsingular matrix $\mathbf{V}$:

$$\mathbf{Y} \sim N(\mu, \mathbf{V}\sigma^2) \qquad r(\mathrm{V}) = n$$

then:

1. $\mathbf{Y}'\left(\frac{1}{\sigma^2}\mathbf{A}\right)\mathbf{Y}$ is distributed noncentral χ^2 with:
   (a) Degrees of freedom $= r(\mathbf{A})$
   (b) Noncentrality parameter $= \Omega = \frac{1}{\sigma^2}\,\boldsymbol{\mu}'\mathbf{A}\,\boldsymbol{\mu}$ if $\mathbf{AV}$ is idempotent
2. $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independent if $\mathbf{AVB} = \mathbf{0}$
3. $\mathbf{Y}'\mathbf{AY}$ and linear function $\mathbf{BY}$ are independent if $\mathbf{BVA} = \mathbf{0}$

4.3.1 Applications to Normal Multiple Regression Model

The sums of squares for the Analysis of Variance are all based on idempotent defining matrices:

For the **Model** sum of squares:

$$\frac{SS(\text{MODEL})}{\sigma^2} = \mathbf{Y}'\left(\frac{1}{\sigma^2}\mathbf{P}\right)\mathbf{Y} \qquad \mathbf{AV} = \mathbf{PI} = \mathbf{P} \quad \mathbf{AVAV} = \mathbf{PP} = \mathbf{P}$$

$$df(\text{MODEL}) = r(\mathbf{A}) = r(\mathbf{P}) = p'$$

$$\Omega(\text{MODEL}) = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'\mathbf{PX}\,\boldsymbol{\beta} = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\,\boldsymbol{\beta}$$

For the **Mean** sum of squares:

$$\frac{SS(\mu)}{\sigma^2} = \mathbf{Y}'\left(\frac{1}{\sigma^2}\frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad \mathbf{AV} = \frac{1}{n}\mathbf{JI} = \frac{1}{n}\mathbf{J} \quad \mathbf{AVAV} = \frac{1}{n}\mathbf{J}\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{J}$$

$$\mathit{df}(\text{MEAN}) = r(\frac{1}{n}\mathbf{J}) = \frac{1}{n}\sum_{i=1}^{n} 1 = \frac{1}{n}n = 1$$

$$\Omega(\text{MEAN}) = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{JX}\,\boldsymbol{\beta} = \frac{1}{2\sigma^2}\frac{(\mathbf{1}'\mathbf{X}\,\boldsymbol{\beta})^2}{n}$$

The last equality is obtained by recalling that $\mathbf{J} = \mathbf{11}'$, and:

$$\boldsymbol{\beta}'\mathbf{X}'\mathbf{JX}\,\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{11}'\mathbf{X}\,\boldsymbol{\beta} = (\boldsymbol{\beta}'\mathbf{X}'\mathbf{1})(\mathbf{1}'\mathbf{X}\,\boldsymbol{\beta}) = (\mathbf{1}'\mathbf{X}\,\boldsymbol{\beta})^{\mathbf{2}}$$

For the **Regression** sum of squares:

$$\frac{SS(\text{REGRESSION})}{\sigma^2} = \mathbf{Y}'\left(\frac{1}{\sigma^2}(\mathbf{P}-\frac{1}{n}\mathbf{J})\right)\mathbf{Y} \qquad \mathbf{AV} = \mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{I}$$

$$\mathbf{AVAV} = \mathbf{PP} - \mathbf{P}\frac{1}{n}\mathbf{J} - \frac{1}{n}\mathbf{JP} + \frac{1}{n}\mathbf{J}\frac{1}{n}\mathbf{J} = \mathbf{P}-\frac{1}{n}\mathbf{J}$$

$$\mathit{df}(\text{REGRESSION}) = r(\mathbf{P}-\frac{1}{n}\mathbf{J})) = r(\mathbf{P}) - r(\frac{1}{n}\mathbf{J}) = p' - 1$$

$$\Omega(\text{REGRESSION}) = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J})\mathbf{X}\,\boldsymbol{\beta} = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J})\mathbf{X}\,\boldsymbol{\beta} = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\frac{\mathbf{1}}{\mathbf{n}}\mathbf{J})\mathbf{X}\,\boldsymbol{\beta}$$

For the **Residual** sum of squares:

$$\frac{SS(\text{RESIDUAL})}{\sigma^2} = \mathbf{Y}'\left(\frac{1}{\sigma^2}(\mathbf{I}-\mathbf{P})\right)\mathbf{Y} \qquad \mathbf{AV} = (\mathbf{I}-\mathbf{P})\mathbf{I} = (\mathbf{I}-\mathbf{P}) \quad \mathbf{AVAV} = (\mathbf{I}-\mathbf{P})(\mathbf{I}-\mathbf{P}) = (\mathbf{I}-\mathbf{P})$$

$$\mathit{df}(\text{RESIDUAL}) = r(\mathbf{A}) = r((\mathbf{I}-\mathbf{P})) = \mathbf{r}(\mathbf{I}) - \mathbf{r}(\mathbf{P}) = n - p'$$

$$\Omega(\text{RESIDUAL}) = \frac{1}{2\sigma^2}\,\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{P})\mathbf{X}\,\boldsymbol{\beta} = \frac{1}{2\sigma^2}(\,\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\,\boldsymbol{\beta} - \,\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\,\boldsymbol{\beta}) = 0$$

Since we have already shown that the quadratic forms for $SS(\mu)$, $SS(\text{REGRESSION})$, and $SS(\text{RESIDUAL})$ are all pairwise orthogonal, and in our current model $\mathbf{V} = \mathbf{I}$, then these sums of squares are all **independent** due to the second part of Cochran's Theorem.

Consider any linear function $\mathbf{K}'\hat{\beta} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y} = \mathbf{BY}$. Then, by the last part of Cochran's Theorem, $\mathbf{K}'\hat{\beta}$ is independent of $SS(\textbf{RESIDUAL})$:

$$\mathbf{B} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' \qquad \mathbf{V} = \mathbf{I} \qquad \mathbf{A} = (\mathbf{I}-\mathbf{P})$$

$$\Rightarrow \mathbf{BVA} = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P} =$$
$$\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0}$$

Consider the following random variable F:

$$F = \frac{X_1^2/\nu_1}{X_2^2/\nu_2}$$

where X_1^2 is distibuted noncentral χ^2 with ν_1 degrees of freedom and noncentrality parameter Ω_1, and X_2^2 is distibuted central χ^2 with ν_2 degrees of freedom. Further, assume that X_1^2 and X_2^2 are independent. Then, F is distrbuted noncental F with ν_1 numerator, ν_2 denominator degrees of freedom, and noncentrality parameter Ω_1.

This applies as follows for the F-test in the Analysis of Variance.

- $\frac{SS(\text{REGRESSION})}{\sigma^2} \sim$ noncentral-χ^2 with $df = p$ and $\Omega = \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{X}\,\boldsymbol{\beta}}{2\sigma^2}$
- $\frac{SS(\text{RESIDUAL})}{\sigma^2} \sim$ central-χ^2 with $df = n - p'$
- SS(REGRESSION) and SS(RESIDUAL) are independent
- $\frac{\left(\frac{SS(\text{REGRESSION})}{\sigma^2}\right)/p}{\left(\frac{SS(\text{RESIDUAL})}{\sigma^2}\right)/(n-p')} = \frac{MS(\text{REGRESSION})}{MS(\text{RESIDUAL})} \sim$ noncentral-F with p numerator and $n-p'$ denominator degrees' of freedom, and noncentrality parameter $\Omega = \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{X}\,\boldsymbol{\beta}}{2\sigma^2}$
- The noncentrality parameter for SS(REGRESSION) does not involve β_0, and for full rank $\mathbf{X}$, $\Omega = 0 \iff \beta_1 = \beta_2 = \cdots = \beta_p = 0$, otherwise $\Omega > 0$

This theory leads to the F-test to determine whether the set of p regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are all equal to 0:

- $H_0: \ \boldsymbol{\beta}^* = \mathbf{0}$ where $\boldsymbol{\beta}^* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$
- $H_A: \ \boldsymbol{\beta}^* \neq \mathbf{0}$
- $TS: F_0 = \frac{SS(\text{REGRESSION})/p}{MS(\text{RESIDUAL})/(n-p')} = \frac{MS(\text{REGRESSION})}{MS(\text{RESIDUAL})}$
- $RR: F_0 :\geq F_{(\alpha,p,n-p')}$
- P-value: $Pr\{F \geq F_0\}$ where $F \sim F_{p,n-p'}$
- The power of the test under a specific alternative can be found by finding the area under the relevant noncentral-F distribution to the right of the critical value defining the rejection region.

Example 1 – LSD Pharmacodynamics

Suppose that the true parameter values are: $\beta_0 = 90$, $\beta_1 = -10$, and $\sigma^2 = 50$ (these are consistent with the least squares estimates). Recall that the fact $\beta_0 \neq 0$ has no bearing on the F-test, only that $\beta_1 \neq 0$. Then:

$$\Omega = \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{P}-\frac{1}{n}\mathbf{J})\mathbf{X}\ \boldsymbol{\beta}}{2\sigma^2} = 22.475$$

Figure 3 gives the central-F distribution (the distribution of the test statistic under the null hypothesis) and the noncentral-F distribution (the distribution of the test statistic under this specific alternative hypothesis). Further, the power of the test under these specific parameter levels is the area under the noncentral-F distribution to the right of $F_{\alpha,1,5}$. Table 4.3.1 gives the power (the probability we reject H_0 under H_0 and several sets of values in the alternative hypothesis) for three levels of α, where $F_{(.100,1,5)} = 4.06$, $F_{(.050,1,5)} = 6.61$, and $F_{(.010,1,5)} = 16.26$. **The reason that the column for the noncentrality parameter is 2Ω is that SAS' function for returning a tail area from a noncentral-F distribution is twice the noncentrality parameter we use in this section's notation.)**

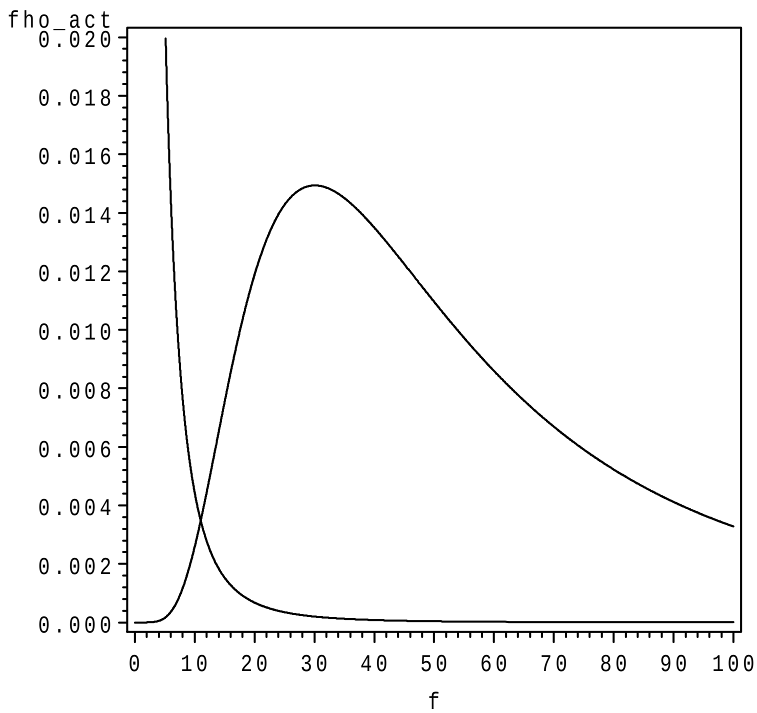

Figure 3: Central and noncentral-F distributions for LSD example, $\beta_1 = -10$, $\sigma^2 = 50$

Note that as the true slope parameter moves further away from 0 for a fixed α level, the power increases. Also, as α (the **size** of the rejection region) decreases, so does the power of the test. Under the null hypothesis ($\beta_1 = 0$), the size of the rejection region is the power of the test (by definition).

		α		
β_1	2Ω	0.100	0.050	0.010
0	0	0.100	0.050	0.010
-2	1.80	0.317	0.195	0.053
-4	7.19	0.744	0.579	0.239
-6	16.18	0.962	0.890	0.557
-8	28.77	0.998	0.987	0.831
-10	44.95	1.000	0.999	0.959

Table 11: Power = Pr(Reject H_0) under several configurations of Type I error rate (α) and slope parameter (β_1) for $\sigma^2 = 50$

4.4 General Tests of Hypotheses

Tests regarding linear functions of regression parameters are conducted as follow.

- Simple Hypothesis $\Rightarrow$ One linear function
- Composite Hypothesis $\Rightarrow$ Several linear functions

$$H_0 : \mathbf{K}' \boldsymbol{\beta} = \mathbf{m} \qquad H_A : \mathbf{K}' \boldsymbol{\beta} \neq \mathbf{m}$$

where $\mathbf{K}'$ is a $k \times p'$ matrix of coefficients defining k linear functions of the β_j^s to be tested ($k \leq p'$), and $\mathbf{m}$ is a $k \times 1$ column vector of constants (often, but not necessarily 0^s). The k linear functions must be linearly independent, but need not be orthogonal. This insures that $\mathbf{K}'$ will be full (row) rank (that is $r(\mathbf{K}') = k$) and that H_0 will be consistent $\forall \mathbf{m}$.

Estimator and its Variance

Parameter – $\mathbf{K}' \boldsymbol{\beta} - \mathbf{m}$

Estimator – $\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m}$ $\qquad \mathbf{E}[\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m}] = \mathbf{K}' \boldsymbol{\beta} - \mathbf{m}$

Variance of Estimator – $\mathbf{Var}[\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m}] = \mathbf{Var}[\mathbf{K}' \hat{\boldsymbol{\beta}}] = \mathbf{K}'\mathbf{Var}[\hat{\boldsymbol{\beta}}]\mathbf{K} = \mathbf{K}'\sigma^{\mathbf{2}}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K} = \mathbf{V}\sigma^{\mathbf{2}}$

Sum of Squares for testing $H_0 : K' \boldsymbol{\beta} = m$

A quadratic form is created from the estimator $\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m}$ by using a defining matrix that is the inverse of $\mathbf{V}$. This is can be thought of as a matrix version of "squaring a t-statistic.

$$Q = (\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K}]^{-\mathbf{1}}(\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m})$$

That is, Q is a quadratic form in $\mathbf{K}' \hat{\boldsymbol{\beta}} - \mathbf{m}$ with $\mathbf{A} = [\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{K}]^{-\mathbf{1}} = \mathbf{V}^{-\mathbf{1}}$. Making use of the earlier result, regarding expectations of quadratic forms, namely:

$$\mathbf{E}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = tr(\mathbf{A}\mathbf{V}_{\mathbf{Y}}) + \boldsymbol{\mu}'\mathbf{A}\,\boldsymbol{\mu} = \sigma^2 tr(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}'\mathbf{A}\,\boldsymbol{\mu}$$

we get:

$$\mathbf{E}[Q] = \sigma^2 tr[(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})] + (\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m}) =$$

$$= \sigma^2 tr[\mathbf{I_k}] + \mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m}) = k\sigma^2 + (\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})$$

Now, $\mathbf{AV} = \mathbf{I_k}$ is idempotent and $r(\mathbf{A}) = r(\mathbf{K}) = k$ (with the restrictions on $\mathbf{K'}$ stated above). So as long as $\boldsymbol{\varepsilon}$ holds our usual assumptions (normality, constant variance, independent elements), then Q/σ^2 is distributed noncentral-χ^2 with k degrees of freedom and noncentrality parameter:

$$\Omega_Q = \frac{\mu'\mathbf{A}\mu}{2\sigma^2} = \frac{(\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m})}{2\sigma^2}$$

$$\text{where } \Omega_Q = 0 \quad \Longleftrightarrow \quad \mathbf{K}'\ \boldsymbol{\beta} = \mathbf{m}$$

So, as before, for the test of $\boldsymbol{\beta}^* = \mathbf{0}$, we have a sum of squares for a hypothesis that is noncentral-χ^2, in this case having k degrees of freedom. Now, we show that Q is independent of SS(RESIDUAL), for the case $\mathbf{m} = \mathbf{0}$ (it holds regardless, but the math is messier otherwise).

$$Q = (\mathbf{K}'\ \hat{\beta})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \hat{\beta}) \quad SS(\text{RESIDUAL}) = \mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}$$

$$Q = \ \hat{\beta}'\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \hat{\beta}) = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Recall that $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independent if $\mathbf{BVA} = \mathbf{0}$. Here, $\mathbf{V} = \mathbf{I}$.

$$\mathbf{BA} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{P}) = \mathbf{0}$$

since $\mathbf{X}'\mathbf{P} = \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}'$. Thus Q is independent of SS(RESIDUAL). This leads to the F-test for the test.

- $H_0 : \mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m} = \mathbf{0}$
- $H_A : \mathbf{K}'\ \boldsymbol{\beta} - \mathbf{m} \neq \mathbf{0}$
- $TS : F_0 = \frac{Q/k}{s^2} = \frac{(\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m})/k}{MS(\text{RESIDUAL})}$
- $RR : F_0 \geq F_{(\alpha,k,n-p')}$
- P-value: $Pr(F \geq F_0)$ where $F \sim F_{(k,n-p')}$

4.4.1 Special Cases of the General Test

Case 1 - Testing a Simple Hypothesis ($k = 1$)

In this case, $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ is a scalar, with an inverse that is its recipocal. Also, $\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m}$ is a scalar.

$$Q = (\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m}) = \frac{(\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m})^2}{\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}}$$

$$\Rightarrow \quad F_0 = \frac{(\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m})^2}{s^2[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]} = \left(\frac{\mathbf{K}'\ \hat{\boldsymbol{\beta}} - \mathbf{m}}{sqrts^2[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]}\right)^2 = t_0^2$$

Case 2 - Testing k Specific $\beta_j^{'s} = 0$

In this case, $\mathbf{K}'\,\boldsymbol{\beta} - \mathbf{m}$ is simply a "subvector" of the vector $\boldsymbol{\beta}$, and $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ is a "submatrix" of $(\mathbf{X}'\mathbf{X})^{-1}$. Be careful of row and column labels because of β_0.

Suppose we wish to test that the last $q < p$ elements of $\boldsymbol{\beta}$ are 0, controlling for the remaining $p - q$ independent variables:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0 \qquad H_A : \text{Not all } \beta_i = 0 \quad (i = p-q+1, \ldots, p)$$

Here, $\mathbf{K}'$ is a $q \times p'$ matrix that can be written as $\mathbf{K}' = [\mathbf{0}|\mathbf{I}]$, where $\mathbf{0}$ is a $q \times p' - q$ matrix of 0^s and $\mathbf{I}$ is the $q \times q$ identity matrix. Then:

$$\mathbf{K}'\,\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_{p-q+1} \\ \hat{\beta}_{p-q+2} \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \qquad \mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K} =$$

$$= \begin{bmatrix} c_{p-q+1,p-q+1} & c_{p-q+1,p-q+2} & \cdots & c_{p-q+1,p} \\ c_{p-q+2,p-q+1} & c_{p-q+2,p-q+2} & \cdots & c_{p-q+2,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p,p-q+1} & c_{p,p-q+2} & \cdots & c_{p,p} \end{bmatrix}$$

where $c_{i,j}$ is the element in the $(i+1)^{st}$ row and $(i+1)^{st}$ column of $(\mathbf{X}'\mathbf{X})^{-1}$. Then Q is:

$$Q = (\mathbf{K}'\,\hat{\boldsymbol{\beta}})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}]^{-1}(\mathbf{K}'\,\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \hat{\beta}_{p-q+1} & \hat{\beta}_{p-q+2} & \vdots & \hat{\beta}_p \end{bmatrix} \begin{bmatrix} c_{p-q+1,p-q+1} & c_{p-q+1,p-q+2} & \cdots & c_{p-q+1,p} \\ c_{p-q+2,p-q+1} & c_{p-q+2,p-q+2} & \cdots & c_{p-q+2,p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p,p-q+1} & c_{p,p-q+2} & \cdots & c_{p,p} \end{bmatrix}^{-1} \begin{bmatrix} \hat{\beta} \\ \hat{\beta} \\ \\ \\ \end{bmatrix}$$

$$\Rightarrow F_0 = \frac{Q/q}{s^2}$$

Case 3 – Testing a single $\beta_j = 0$

This is a simplification of case 2, with $\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K}$ being the $(j+1)^{st}$ element of $(\mathbf{X}'\mathbf{X})^{-1}$, and $\mathbf{K}'\,\hat{\boldsymbol{\beta}} = \hat{\beta}_j$.

$$Q = \frac{(\hat{\beta}_j)^2}{c_{jj}} \quad \Rightarrow \quad F_0 = \frac{(\hat{\beta}_j)^2}{s^2 c_{jj}} = \left[\frac{\hat{\beta}_j}{\sqrt{s^2 c_{jj}}}\right]^2 = t_0^2$$

4.4.2 Computing Q from Differences in Sums of Squares

The sums of squares for a general test can be obtained by fitting various models, and taking differences in Residual suns of squares.

$$H_0 : \mathbf{K}'\,\boldsymbol{\beta} = \mathbf{m} \qquad H_A : \mathbf{K}'\,\boldsymbol{\beta} \neq \mathbf{m}$$

First, the **Full Model** is fit, that lets all parameters to be free (H_A), and the least squares estimate is obtained. The residual sum of squares is obtained and labelled $SS(\text{RESIDUAL}_{\text{FULL}})$. Under the full model, with no restriction on the parameters, p' parameters are estimated and this sum of squares has $n - p'$.

Next, the **Reduced Model** is fit, that places $k \leq p'$ constraints on the parameters (H_0). Any remaining parameters are estimated by least squares. The residual sum of squares is obtained and labelled $SS(\text{RESIDUAL}_{\text{REDUCED}})$. Note that since we are forcing certain parameters to take on specific values $SS(\text{RESIDUAL}_{\text{REDUCED}}) \geq SS(\text{RESIDUAL}_{\text{FULL}})$, with the equality only taking place if the estimates from the full model exactly equal the constrained values under H_0. With the k constraints, only $p' - k$ parameters are being estimated and the residual sum of squares has $n - (p' - k)$ degrees of freedom.

We obtain the sum of squares and degrees of frredom for the test by taking the difference in the residual sums of squares and in their corresponding degrees' of freedom:

$$Q = SS(\text{RESIDUAL}_{\text{REDUCED}}) - SS(\text{RESIDUAL}_{\text{FULL}}) \qquad df(Q) = (n - (p' - k)) - (n - p') = k$$

As before:

$$F = \frac{Q/k}{s^2} = \frac{\frac{(SS(\text{RESIDUAL}_{\text{REDUCED}}) - SS(\text{RESIDUAL}_{\text{FULL}})}{(n-(p'-k))-(n-p')}}{\frac{SS(\text{RESIDUAL}_{\text{FULL}})}{n-p'}}$$

Examples of Constraints and the Appropriate Reduced Models

Suppose that $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$.

$H_0 : \beta_1 = \beta_2 \qquad (k = 1)$

$$Y_i = \beta_0+\beta_1X_{i1}+\beta_1X_{i2}+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon_i == \beta_0+\beta_1(X_{i1}+X_{i2})+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon_i = \beta_0+\beta_1X^*_{i1}+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon_i \quad X^*_{i1} =$$

$H_0 : \beta_0 = 100 \quad \beta_1 = 5 \qquad (k = 2)$

$$Y_i = 100+5X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon_i \quad \Rightarrow \quad Y_i-100-5X_i1 = \beta_2X_{i2}+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon_i \quad Y_i^* = \beta_2X_{i2}+\beta_3X_{i3}+\beta_4X_{i4}+\varepsilon$$

Some notes regarding computation of Q:

- Q can always be computed from differences in residual sums of squares.
- When β_0 is in the model, and not involved in $H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}$ then we can use $Q = SS(\text{MODEL}_{\text{FULL}}) - SS(\text{MODEL}_{\text{REDUCED}})$.
- When $\beta_0 \neq 0$, is in the reduced model, you cannot use the difference in Regression sums of squares, since $SS(\text{TOTAL UNCORRECTED})$differs between the two models.

Best practice is always to use the error sums of squares.

4.4.3 R-Notation to Label Sums of Squares

Many times in practice, we wish to test that a subset of the partial regression coefficients are all equal to 0. We can write the model sum of squares for a model containing $\beta_0, \beta_1, \ldots, \beta_p$ as:

$$R(\beta_0, \beta_1, \ldots, \beta_p) = SS(\text{MODEL})$$

The logic is to include all β_i in $R(\cdot)$ that are in the model being fit. Returning to the case of testing the last $q < p$ regression coefficients are 0:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0 \qquad H_A : \text{Not all } \beta_i = 0 \quad (i = p-q+1, \ldots, p)$$

$$H_0 : R(\beta_0, \beta_1, \ldots, \beta_{p-q}) = SS(\text{MODEL}_{\text{REDUCED}})$$

$$H_0 : R(\beta_0, \beta_1, \ldots, \beta_p) = SS(\text{MODEL}_{\text{FULL}})$$

$$Q = SS(\text{MODEL}_{\text{FULL}} - SS(\text{MODEL}_{\text{REDUCED}} = R(\beta_0, \beta_1, \ldots, \beta_p) - R(\beta_0, \beta_1, \ldots, \beta_{p-q}) =$$

$$= R(\beta_{p-q+1}, \ldots, \beta_p | \beta_0, \beta_1, \ldots, \beta_{p-q})$$

Special cases include:

$$SS(\text{REGRESSION}) = SS(\text{MODEL}) - SS(\mu) = R(\beta_0, \beta_1, \ldots, \beta_p) - R(\beta_0) = R(\beta_1, \ldots, \beta_p | \beta_0)$$

Partial (TYPE III) Sums of Squares: $R(\beta_0, \ldots, \beta_{i-1}, \beta_i, \beta_{i+1}, \ldots, \beta_p) - R(\beta_0, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p) =$

$$= R(\beta_i | \beta_0, \ldots, \beta_{i-1}, \beta_{i+1}, \ldots, \beta_p)$$

Sequential (TYPE I) Sums of Squares: $R(\beta_0, \ldots, \beta_{i-1}, \beta_i) - R(\beta_0, \ldots, \beta_{i-1}) = R(\beta_i | \beta_0, \ldots, \beta_{i-1})$

$$SS(\text{REGRESSION}) = R(\beta_1, \ldots, \beta_p | \beta_0) = R(\beta_1 | \beta_0) + R(\beta_2 | \beta_1, \beta_0) + \cdots + R(\beta_p | \beta_1, \ldots, \beta_{p-1})$$

The last statement shows that sequential sums of squares (corrected for the mean) sum to the regression sum of squares. The partial sums of squares do not sum to the regression sum of squares unless the last p columns of $\mathbf{X}$ are mutually pairwise orthogonal, in which case the partial and sequential sums of squares are identical.

4.5 Univariate and Joint Confidence Regions

In this section confidence regions for the regression parameters are given. See RPD for cool pictures.

Confidence Intervals for Partial Regression Coefficients and Intercept

Under the standard normality, constant variance, and independence assumptions; as well as the independence of $\mathbf{K}'\,\hat{\boldsymbol{\beta}}$ and SSE(RESIDUAL), we have:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 c_{jj}) \quad \text{where } c_{jj} \text{ is the } (j+1)^{st} \text{ diagonal element of } (\mathbf{X}'\mathbf{X})^{-1}$$

$$\Rightarrow \quad \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 c_{jj}}} \sim t_{(n-p')} \quad \Rightarrow \quad Pr\{\hat{\beta}_j - t_{(\alpha/2,n-p')}\sqrt{s^2 c_{jj}} \le \beta_j \le \hat{\beta}_j - t_{(\alpha/2,n-p')}\sqrt{s^2 c_{jj}}\} = 1-\alpha$$

$$\Rightarrow \quad (1-\alpha)100\% \text{ Confidence Interval for } \beta_j\text{:} \quad \hat{\beta}_j \pm t_{(\alpha/2,n-p')}\sqrt{s^2 c_{jj}}$$

Confidence Interval for $\beta_0 + \beta_1 X_{10} + \cdots + \beta_p X_{p0} = \mathbf{x_0'}\,\boldsymbol{\beta}$

By a similar argument, we have a $(1-\alpha)100\%$ confidence interval for the mean at a given combination of levels of the independent variables, where $\hat{Y}_0 = \mathbf{x_0'}\,\hat{\boldsymbol{\beta}}$:

$$\hat{Y}_0 \pm t_{(\alpha/2,n-p')}\sqrt{s^2 \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}}$$

Prediction Interval for Future Observation Y_0 when $X_1 = X_{10}, \ldots, X_p = X_{p0}$ ($\mathbf{x_0}$)

For a $(1-\alpha)100\%$ prediction interval for a single outcome (future observation) at a given combination of levels of the independent variables, where $\hat{Y}_0 = \mathbf{x_0'}\,\hat{\boldsymbol{\beta}}$:

$$\hat{Y}_0 \pm t_{(\alpha/2,n-p')}\sqrt{s^2\,[1 + \mathbf{x_0'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_0}]}$$

Bonferroni's Method for Simultaneous Confidence Statements

If we want to construct c confidence statements, with simultaneous conficence coefficient $1-\alpha$, then we can generate the c confidence intervals, each at level $(1-\frac{\alpha}{c})$. That is, each confidence interval is more conservative (wider) than if they had been constructed one-at-a-time.

Joint Confidence Regions for $\boldsymbol{\beta}$

From the section on the general linear tests, if we set $\mathbf{K}' = \mathbf{I_{p'}}$, we have the following distributional property:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'[(X'X)^{-1}]^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{p's^2} \sim F_{(p',n-p')}$$

$$\Rightarrow \quad Pr\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le p's^2 F_{(1-\alpha,p',n-p')}\} = 1-\alpha$$

Values of $\boldsymbol{\beta}$ in this set constitute a joint $(1-\alpha)100\%$ confidence region for $\boldsymbol{\beta}$.

4.6 A Test for Model Fit

A key assumption for the model is that the relation between $\mathbf{Y}$ and $\mathbf{X}$ is linear (that is $\mathbf{E}[\mathbf{Y}] = \mathbf{X}\,\boldsymbol{\beta}$). However, the relationship may be nonlinear. S-shaped functions are often seen in biological and business applications, as well as the general notion of "diminishing marginal returns," for instance.

A test can be conducted, when replicates are obtained at various combinations of levels of the independent variables. Suppose we have c unique levels of $\mathbf{x}$ in our sample. It's easiest to consider the test when there is a single independent variable (but it generalizes straightforwardly). We obtain the sample size (n_j), mean ($\overline{Y}_j$) and variance (s_j^2) at each unique level of X ($\overline{Y}$ is the overall sample mean for Y). We obtain the a partition of SS(TOTAL CORRECTED) to test:

$$H_0 : E[Y_i] = \beta_0 + \beta_1 X_i \qquad H_A : E[Y_i] = \mu_i \neq \beta_0 + \beta_1 X_i$$

where μ_i is the mean of all observations at the level X_i and is **not linear** in X_i. The alternative can be interpreted as a 1–way ANOVA where the means are not necessarily equal. The partition is given in Table 12, with the following identities with respect to sums of squares:

$$SS(\text{TOTAL CORR}) = SS(\text{REG}) + SS(\text{LF}) + SS(\text{PE})$$

where $SS(\text{RESIDUAL}) = SS(\text{LF}) + SS(\text{PE})$. Intuitively, these sums of squares and their degrees of freedom can be written as:

$$SS(\text{LF}) = \sum_{i=1}^{n} (\overline{Y}_{(i)} - \hat{Y}_i)^2 = \sum_{j=1}^{c} n_j (\overline{Y}_j - \hat{Y}_{(j)})^2 \qquad df_{LF} = c - 2$$

$$SS(\text{PE}) = \sum_{i=1}^{n} (Y_i - \overline{Y}_{(i)})^2 = \sum_{j=1}^{c} (n_j - 1) s_j^2 \qquad df_{LF} = n - c$$

where $\overline{Y}_{(i)}$ is the mean for the group of observations at the same level of X as observation i and $\hat{Y}_{(j)}$ is the fitted value. The test for goodness-of-fit is conducted as follows:

Source	df	SS
Regression (REG)	1	$\sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$
Lack of Fit (LF)	$c - 2$	$\sum_{j=1}^{c} n_j (\overline{Y}_j - \hat{Y}_{(j)})^2$
Pure Error (PE)	$n - c$	$\sum_{j=1}^{c} (n_j - 1) s_j^2$

Table 12: ANOVA for Lack of Fit F-test

- $H_0 : E[Y_i] = \beta_0 + \beta_1 X_i$ (Relation is linear)
- $H_A : E[Y_i] = \mu_i \neq \beta_0 + \beta_1 X_i$ (Relationship is nonlinear)
- $TS : F_0 = \frac{MS(LF)}{MS(PE)} = \frac{SS(LF)/(c-2)}{SS(PE)/(n-c)}$
- $RR : F_0 \geq F_{(\alpha, c-2, n-c)}$

Example – Building Costs

A home builder has 5 floor plans: $1000 ft^2$, 1500, 2000, 2500, and 3000. She knows that the price to build individual houses varies, but believes that the mean price may be linearly related to size in this size range. She samples from her files the records of $n = 10$ houses, and tabulates the total building cost for each of the houses. She samples $n_i = 2$ homes at each of the $c = 5$ levels of X. Consider each of the following models:

$$\text{Model 1:} \quad E[Y] = 5000 + 10X \quad \sigma = 500$$

$$\text{Model 2:} \quad E[Y] = -12500 + 30X - 0.05X^2 \quad \sigma = 500$$

These are shown in Figure 12.

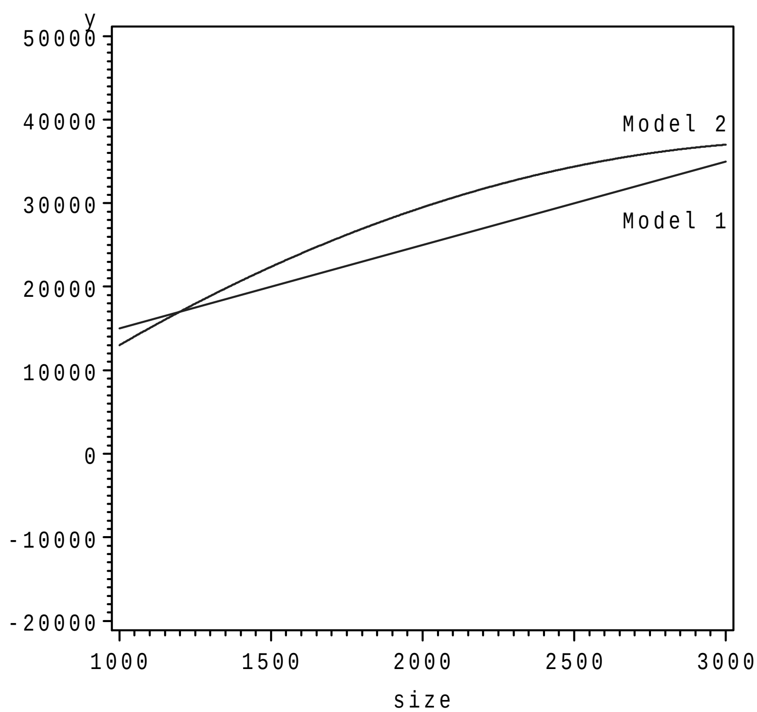

Figure 4: Models 1 and 2 for the building cost example

Data were generated from each of these models, and the **simple linear regression model** was fit. The least squares estimates for model 1 (correctly specified) and for model 2 (incorrectly fit) were obtained:

$$\text{Model 1:} \quad \hat{Y} = 5007.22 + 10.0523X$$

$$\text{Model 2:} \quad \hat{Y} = 4638.99 + 10.1252X$$

The observed values, group means, and fitted values from the simple linear regression model are obtained for both the correct model (1) and the incorrect model (2) in Table 4.6.

The sums of squares for lack of fit are obtained by taking deviations between the group means (which are estimates of $E[Y_i]$ under H_A) and the fitted values (which are estimates of $E[Y_i]$ under H_0).

$$\text{Model 1} \quad SS(\text{LF}) = \sum_{i=1}^{n} n_j(\overline{Y}_{(i)} - \hat{Y}_i)^2 = 2(15303.96 - 15059.40)^2 + \cdots + 2(35496.89 - 35164.04)^2 = 631872.71$$

		Correct Model (1)			Incorrect Model (2)		
i	X_i	Y_i	$\overline{Y}_{(i)}$	$\hat{Y}_i$	Y_i	$\overline{Y}_{(i)}$	$\hat{Y}_i$
1	1000	14836.70	15303.96	15059.49	12557.72	12429.03	14764.23
2	1000	15771.22	15303.96	15059.49	12300.33	12429.03	14764.23
3	1500	20129.51	19925.80	20085.63	21770.87	21045.46	19826.86
4	1500	19722.08	19925.80	20085.63	20320.05	21045.46	19826.86
5	2000	25389.36	25030.88	25111.77	27181.07	27242.41	24889.48
6	2000	24672.40	25030.88	25111.77	27303.75	27242.41	24889.48
7	2500	29988.95	29801.31	30137.90	31139.83	30931.27	29952.10
8	2500	29613.68	29801.31	30137.90	30722.71	30931.27	29952.10
9	3000	35362.12	35496.89	35164.04	33335.17	32799.24	35014.73
10	3000	35631.66	35496.89	35164.04	32263.31	32799.24	35014.73

Table 13: Observed, fitted, and group mean values for the lack of fit test

$$\text{Model 2} \quad SS(\text{LF}) = \sum_{i=1}^{n} n_j(\overline{Y}_{(i)} - \hat{Y}_i)^2 = 2(12429.03 - 14764.23)^2 + \cdots + 2(32799.24 - 35014.73)^2 = 36683188.83$$

The sum of squares for **pure error** are obtained by taking deviations between the observed outcomes and their group means (this is used as an unbiased estimate of σ^2 under H_A, after dividing through by its degrees of freedom).

$$\text{Model 1} \quad SS(\text{PE}) = \sum_{i=1}^{n}(Y_i - \overline{Y}_{(i)})^2 = (14836.70 - 15303.96)^2 + \cdots + (35631.66 - 35496.89)^2 = 883418.93$$

$$\text{Model 2} \quad SS(\text{PE}) = \sum_{i=1}^{n}(Y_i - \overline{Y}_{(i)})^2 = (12557.72 - 12429.03)^2 + \cdots + (32263.31 - 32799.24)^2 = 1754525.81$$

The F-statistics for testing between H_0 (that the linear model is the true model) and H_A (that the true model is not linear) are:

$$\text{Model 1} \quad F_0 = \frac{MS(LF)}{MS(PE)} = \frac{SS(LF)/(c-2)}{SS(PE)/(n-c)} = \frac{631872.71/(5-2)}{883418.93/(10-5)} = 1.19$$

$$\text{Model 2} \quad F_0 = \frac{MS(LF)}{MS(PE)} = \frac{SS(LF)/(c-2)}{SS(PE)/(n-c)} = \frac{36683188.83/(5-2)}{1754525.81/(10-5)} = 34.85$$

The rejection region for these tests, based on $\alpha = 0.05$ significance level is:

$$RR : F_0 \geq F_{(.05,3,5)} = 5.41$$

Thus, we fail to reject the null hypothesis that the model is correct when the data were generated from the correct model. Further, we do reject the null hypothesis when data were generated from the incorrect model.