Chapter 13

Combining Regression and ANOVA: Quantitative and Categorical Predictors

Chapter 11 introduced multiple regression to analyze the relationship between a quantitative response variable and *quantitative* explanatory variables. Chapter 12 showed that multiple regression can also handle *categorical* explanatory variables, as in analysis of variance with dummy variables. Not surprisingly, multiple regression can also handle simultaneously quantitative and categorical explanatory variables. The model combines elements of ordinary regression analysis, for which the predictors are quantitative, and analysis of variance, for which the predictors are categorical.

Controlling for a Covariate

One-way ANOVA compares the mean of the response variable for several groups. Two-way ANOVA compares means while controlling for another categorical variable. In many applications, it's useful to compare means while controlling for a quantitative variable. For example, in comparing mean income for men and women, we might control for possibly differing levels of job experience between men and women. The quantitative control variable is called a ***covariate***. The use of regression for this type of comparison is often called ***analysis of covariance***. It is one of the many statistical contributions of the brilliant British statistician, R. A. Fisher.

Because effects may change after controlling for a variable, the results of analysis of covariance differ from the results of analysis of variance. For instance, job experience is usually positively correlated with income. If men tend to have higher levels of experience than women at a particular job, the results of a comparison of mean income for men and women will depend on whether we control for experience.

For simplicity, this chapter illustrates concepts using a single categorical predictor and a single quantitative predictor, but the basic ideas extend to multiple predictors. The first section shows graphic representations of the potential effects. Sections 13.2 and 13.3 show that regression models with dummy variables provide the basis for the analyses. Sections 13.4 and 13.5 present inferences for the models. Section 13.6 shows how to adjust the sample means of Y to reflect their predicted values after controlling for the covariate.

13.1 Comparing Means and Comparing Regression Lines

Denote the quantitative explanatory variable by x, and the categorical explanatory variable by z. When the categorical predictor has two categories, z is a dummy variable; when it has several categories, we use a set of dummy variables.

The analysis of the effect of the quantitative predictor x has a regression flavor. It refers to the regression of Y on x within each category of the categorical variable, treating z as a control variable. The analysis of the effect of the categorical predictor has an ANOVA flavor. It refers to comparing the means of Y for the groups defined by the categorical variable, treating x as the control variable.

Comparing Regression Lines

Table 9.4 in Section 9.5 introduced a data file on Y = selling price of homes. One quantitative predictor is x = size of home. One categorical predictor is z = whether a house is new (1 = yes, 0 = no). Studying the effect of x on Y while controlling for z is equivalent to analyzing the regression of Y on x separately for new and older homes. We could find the best-fitting straight line for each set of points, one line for new homes and a separate line for older homes. We could then compare characteristics of the lines, for instance, whether they climb with similar or different slopes.

In this context, *no interaction* means that the true slope of the line relating expected selling price to the size of home is the same for new and older homes. Equality of slopes implies that the regression lines are parallel (see Figure 13.1a). When the y-intercepts are also equal, the regression lines coincide (see Figure 13.1b). If the rate of increase in selling price as a function of size of home differed for new and existing homes, then the two regression lines would not be parallel. There is then interaction (see Figure 13.1c).

Figure 13.1: Regression Lines Between Two Quantitative Variables, Within Categories of a Categorical Variable with Two Categories

((Fig. 13.1 in 3e))

The effect of x while controlling for z may differ in substantial ways from the bivariate relationship. For instance, the effect could disappear when we control for z. Figure 13.2 displays a set of points having an overall positive relationship when z is ignored. Within each category of z, however, the regression line relating Y to x is horizontal. The overall positive trend is due to the tendency for the categories with high (low) scores on Y to have high (low) scores on x also. Example 10.1 in Chapter 10 presented an example of this type, with Y = math achievement and x = height. The categorical variable was grade of school, with students coming from grades 2, 5, and 8.

Figure 13.2: An Association Between Two Quantitative Variables that Disappears After Controlling for a Categorical Variable

(Fig. 13.2 in 3e))

Figure 13.3: Y = Income by x = Number of Years Experience and z = Gender

((Fig. 13.3 in 3e))

Comparing Means on Y, Controlling for X

Likewise, the effect of the categorical variable z may change substantially when we control for x. For example, consider the relationship between Y = annual income and z = gender for managerial employees of a chain of fast-food restaurants. From a two-sample comparison of men and women, mean annual income is higher for men than for women. In this company, annual income of managers tends to increase with x = number of years of experience. In addition, only recently have women received many managerial appointments, so on the average they have less experience than the men. In summary, men tend to have greater experience, and greater experience tends to correlate with higher income. Perhaps this is why the overall mean annual income is higher for men. A chain relationship may exist, with gender affecting experience, which itself affects income. The difference between the mean incomes of men and women could disappear when we control for experience.

To study whether the difference in mean incomes can be explained by differing experience levels of men and women, we compare mean incomes for men and women having equal levels of experience. If there is no interaction, then the regression line between income and experience for the male employees is parallel to the one for the female employees. In that case, the difference between mean incomes for men and women is identical for all fixed values of x = number of years of experience. Figure 13.3a illustrates this. If the same regression line applies to each gender, as in Figure 13.3b, the mean income for each gender is identical at each level of experience. In that case, no difference occurs between male and female incomes, controlling for experience.

The results of this analysis may differ considerably from the results of the one-way analysis of variance, which compares mean incomes while ignoring rather than controlling for experience. For example, Figure 13.3b depicts a situation in which the sample mean income for men is much greater than that for women. However, the reason for the difference is that men have more experience. In fact, the same regression line fits the relationship between income and experience for both genders. It appears that the mean incomes are equal, controlling for experience.

If interaction exists, then the regression lines are not parallel. In that case, the difference between the mean incomes varies by level of experience. In Figure 13.3c, for example, the mean income for men is higher than the mean income for women at all experience levels, and the difference increases as experience increases.

13.2 Regression with Quantitative and Categorical Predictors

This section presents the regression model without interaction. As in Section 12.3, we represent the categorical predictor using dummy variables, one fewer than the number of categories. We include x in the model in the usual linear manner.

Quantitative and Dummy Explanatory Variables

Suppose there are three categories, as in the following example. The regression model is

$$E(Y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2.$$

Here, β (without a subscript) describes the effect of x on the mean of Y for each group (category). Setting $(z_1 = 1, z_2 = 0)$ gives the line for category one, setting $(z_1 = 0, z_2 = 1)$ gives the line for category two, and setting $(z_1 = 0, z_2 = 0)$ gives the line for category three.

Example 13.1 Regression of Income on Education and Racial–Ethnic Group

For a sample of adult Americans aged over 25, Table 13.1 shows Y = annual income (thousands of dollars), x = number of years of education (where 12 = high school graduate, 16 = college graduate), and z = racial–ethnic group (black, Hispanic, white). The data exhibit patterns of a much larger sample taken by the U.S. Bureau of the Census. The sample contains $n_1 = 16$ blacks, $n_2 = 14$ Hispanics, and $n_3 = 50$ whites, for a total sample size of $N = 80$.

Table 13.2 reports the mean income and education for these subjects. Although the mean incomes differ among the three groups, these differences could result from the differing educational levels. For instance, although white subjects had higher mean incomes than blacks or Hispanics, they also had higher mean education. Perhaps the differences would disappear if we could control for education, making comparisons among the racial–ethnic groups at fixed levels of education.

For racial–ethnic status we set up the dummy variables

$$\begin{aligned} z_1 = &\ 1 \text{ if subject is black, } z_1 = 0 \text{ otherwise;} \\ z_2 = &\ 1 \text{ if subject is Hispanic, } z_2 = 0 \text{ otherwise;} \\ z_1 = &\ z_2 = 0 \text{ if subject is white.} \end{aligned}$$

Table 13.3 shows part of a printout from using software to fit the regression model. The [race=b] and [race=h] parameters refer to the coefficients of the dummy variables z_1 for blacks and z_2 for Hispanics. The prediction equation is

$$\hat{y} = -15.7 + 4.4x - 10.9z_1 - 4.9z_2.$$

For blacks, $z_1 = 1$ and $z_2 = 0$, so the prediction equation is

$$\hat{y} = -15.7 + 4.4x - 10.9(1) - 4.9(0) = -26.6 + 4.4x.$$

Table 13.1: Observations on Y = Annual Income (in Thousands of Dollars) and X = Number of Years of Education, for Three Racial–Ethnic Groups

Black		Hispanic		White		White		White	
y	x	y	x	y	x	y	x	y	x
16	10	32	16	30	14	62	16	50	16
18	7	16	11	48	14	24	10	50	14
26	9	20	10	40	7	50	13	22	11
16	11	58	16	84	18	32	10	26	12
34	14	30	12	50	10	34	16	46	16
22	12	26	10	38	12	52	18	22	9
42	16	20	8	30	12	24	12	24	9
42	16	40	12	76	16	22	14	64	14
16	9	32	10	48	16	20	13	28	12
20	10	22	11	36	11	30	14	32	12
66	16	20	10	40	11	24	13	38	14
26	12	56	14	44	12	120	18	44	12
20	10	32	12	30	10	22	10	22	12
30	15	30	11	60	15	82	16	18	10
20	10			24	9	18	12	24	12
30	19			88	17	26	12	56	20
				46	16	104	14		

The prediction equations for the other two racial–ethnic groups are

$$\begin{aligned} \hat{y} &= -20.6 + 4.4x \quad \text{(Hispanics)} \\ \hat{y} &= -15.7 + 4.4x \quad \text{(whites)} \end{aligned}$$

Figure 13.4 is a scatterplot showing the prediction equations for the three groups. The lines are parallel, since they each have the same slope, 4.4. This is the coefficient of x in each prediction equation.

□

Table 13.2: Mean Income and Education, by Racial–Ethnic Group

	Black	Hispanic	White	Overall
Mean income	$\overline{y}_1 = 27.8$	$\overline{y}_2 = 31.0$	$\overline{y}_3 = 42.4$	$\overline{y} = 37.6$
Mean education	$\bar{x}_1 = 12.2$	$\bar{x}_2 = 11.6$	$\bar{x}_3 = 13.1$	$\bar{x} = 12.7$
Sample size	$n_1 = 16$	$n_2 = 14$	$n_3 = 50$	$N = 80$

Table 13.3: Printout for Fitting Model Assuming No Interaction to Table 13.1 on Y = Income and Explanatory Variables Education and Racial–Ethnic Status (with Dummy Variables for Black and Hispanic Categories)

Parameter	B	Std. Error	t	Sig	95% Conf. Int. Lower	Upper
Intercept	-15.663	8.412	-1.862	.066	-32.4	1.09
education	4.432	.619	7.158	.000	3.2	5.7
[race = b]	-10.874	4.473	-2.431	.017	-19.8	-2.0
[race = h]	-4.934	4.763	-1.036	.304	-14.4	4.6
[race = w]	0	.	.	.		

race=w parameter is set to zero because it is redundant

R-Squared = .462

Figure 13.4: Plot of Prediction Equation for Model Assuming No Interaction. Each line has the same slope, so the lines are parallel.

((Fig. 13.4 in 3e revised))

Interpretation of Parameters: No Interaction Model

Let's consider parameter interpretation. For category 1, $z_1 = 1$ and $z_2 = 0$. Then, the relationship between x and the mean of Y is

$$\begin{aligned} E(Y) &= \alpha + \beta x + \beta_1(1) + \beta_2(0) \\ &= (\alpha + \beta_1) + \beta x. \end{aligned}$$

The mean of Y is linearly related to x with slope β and Y-intercept $\alpha + \beta_1$. Similarly, for category 2, $E(Y) = (\alpha + \beta_2) + \beta x$. In the last category, all the dummy variables equal 0, and the regression equation reduces to $E(Y) = \alpha + \beta x$.

The coefficient β of x is the slope of the regression line for each category. The equation $E(Y) = \alpha + \beta x$ is the regression equation for the final category. The coefficients of the dummy variables tell us how the Y-intercept changes for the other categories. For instance, β_1 (the coefficient of z_1) is the difference between the Y-intercept for the first and the final categories. Since the regression lines are parallel, β_1 is the vertical distance between those two regression lines *at any fixed value of x.* That is, controlling for x, β_1 is the difference between the means of Y for the first and last categories. Table 13.4 displays the separate equations and the parameter interpretations. Figure 13.5 graphically displays the model.

Table 13.4: Summary of Regression Equations and Parameters for Model with No Interaction, When Categorical Predictor Has 3 Categories

Category	y-Intercept	Slope	Mean $E(Y)$ at Fixed x	Difference From Mean of Category 3, Controlling for x
1	$\alpha + \beta_1$	β	$(\alpha + \beta_1) + \beta x$	β_1
2	$\alpha + \beta_2$	β	$(\alpha + \beta_2) + \beta x$	β_2
3	α	β	$\alpha + \beta x$	0

Figure 13.5: Graphic Portrayal of a Model with No Interaction, When the Categorical Predictor Has Three Categories

((Fig. 13.5 in 3e, with all X changed to x))

In summary, the coefficients of the dummy variables estimate differences in means between each category and the final category, which does not have its own dummy variable. Analogous results hold for sample prediction equations. Example 13.1 had prediction equation

$$\hat{y} = -15.7 + 4.4x - 10.9z_1 - 4.9z_2,$$

where z_1 is a dummy variable for blacks and z_2 is one for Hispanics. The coefficient -10.9 of z_1 is the vertical distance between the lines for blacks and whites. This means that the estimated mean annual income is \$10,900 lower for blacks than for whites, at each fixed level of education. Similarly, the coefficient -4.9 of z_2 represents the difference ($-$\$4900) between the estimated mean income for Hispanics and whites, controlling for education.

13.3 Permitting Interaction between Quantitative and Categorical Predictors

The analysis of covariance model that allows no interaction has the same slope for the effect of x in each category of the categorical predictor. The model permitting interaction allows different slopes. To allow interaction, as usual we take cross products of the explanatory variables.

Example 13.2 Regression of Income on Education and Racial–Ethnic Group, Permitting Interaction

For Table 13.1, let's now allow interaction between education and racial–ethnic group in their effects on income. The model has dummy variables z_1 and z_2 for the first two racial-ethnic groups (blacks and Hispanics) and their cross-products $x \times z_1$ and $x \times z_2$ with the education predictor. Software provides the results shown in Table 13.5.

From Table 13.5, the overall prediction equation is

$$\hat{y} = -25.9 + 5.2x + 19.3z_1 + 9.3z_2 - 2.4(x \times z_1) - 1.1(x \times z_2).$$

The prediction equation with both dummy variables equal to zero ($z_1 = z_2 = 0$) refers to the final racial–ethnic category, namely, whites. For that group,

$$\hat{y} = -25.9 + 5.2x + 19.3(0) + 9.3(0) - 2.4x(0) - 1.1x(0) = -25.9 + 5.2x.$$

For the first category (blacks), $z_1 = 1$ and $z_2 = 0$, and

$$\hat{y} = -6.6 + 2.8x.$$

For the second category (Hispanics), $z_1 = 0, z_2 = 1$, and

$$\hat{y} = -16.6 + 4.1x.$$

The coefficient of z_1 (i.e., 19.3) again describes the difference between the y-intercepts for the first and the last category. However, this is the difference *only* at $x = 0$, since the equations have different slopes. Since the coefficient of x (i.e., 2.6) represents the slope for the last (third) category, the coefficient of $(x \times z_1)$ (i.e., -2.4) represents the *difference in slopes* between the first and last categories. The two lines are parallel only when that coefficient equals 0. Similarly, for the second category,

Table 13.5: Printout for Fitting Interaction Model to Table 13.1 on Income, Education, and Racial–Ethnic Status (with Dummy Variables for Black and Hispanic Categories).

```
Parameter                  B     Std. Error     t      Sig
Intercept             -25.869     10.498     -2.464   .016
education               5.210       .783      6.655   .000
[race=b]               19.333     18.293      1.057   .294
[race=h]                9.264     24.282       .382   .704
[race=w]                    0          .          .      .
[race=b]*education     -2.411      1.418     -1.700   .093
[race=h]*education     -1.121      2.006      -.559   .578
[race=w]*education          0          .          .      .

race=w parameters are set to zero because they are redundant

R-Squared   0.482
```

the coefficient of z_2 is the difference between the y-intercepts for the second and the last category, and the coefficient of $(x \times z_2)$ is the difference between their slopes. Table 13.6 summarizes the interpretations of the estimated parameters in the model.

Figure 13.6 plots the three prediction equations. The sample slopes are all positive. Over nearly the entire range of education values observed, whites have the highest estimated mean income, and blacks have the lowest.

When interaction exists, the difference between means of Y for two groups varies as a function of x. For example, the difference between the estimated mean of Y for

Table 13.6: Summary of Prediction Equation Allowing Interaction for y = Income, x = Education, and Racial–Ethnic Status (Dummy Variables z_1 for Blacks and z_2 for Hispanics)

Category	y-Intercept	Slope	$\hat{y} = -25.9 + 5.2x$ $+19.3z_1 + 9.3z_2$ $-2.4(xz_1) - 1.1(xz_2)$	Difference From Category 3 of y-Intercept	Difference From Category 3 of Slope
1 (Black)	$-25.9 + 19.3$	$5.2 - 2.4$	$(-25.9 + 19.3) + (5.2 - 2.4)x$	19.3	-2.4
2 (Hispanic)	$-25.9 + 9.3$	$5.2 - 1.1$	$(-25.9 + 9.3) + (5.2 - 1.1)x$	9.3	-1.1
3 (White)	-25.9	5.2	$-25.9 + 5.2x$	0	0

Figure 13.6: Plot of Prediction Equations for Model with Interaction Terms

((Fig. 13.6 in 3e, amended))

whites and Hispanics at a particular x-value is

$$(-25.9 + 5.2x) - (-16.6 + 4.1x) = -9.3 + 1.1x.$$

This depends on the value of x. As x changes, the difference in the estimated means changes. In this case, the difference between the estimated means is larger at higher levels of education. Figure 13.6 shows that the difference between the mean incomes of whites and blacks also gets larger at higher education levels.

□

Comparing R or R^2 for Different Models

To summarize how much better the model permitting interaction fits, we can check the increase in R^2 or in the multiple correlation R. Recall that R^2 represents the proportional reduction in error from using the prediction equation to predict the response instead of using the overall mean $\overline{y}$.

From the printout for the no-interaction model (Table 13.3), $R^2 = 0.462$. From the printout for the interaction model (Table 13.5), $R^2 = 0.482$. The corresponding multiple correlation values are $\sqrt{0.462} = 0.680$ and $\sqrt{0.482} = 0.695$. There is little gained by fitting the more complex model, as R^2 and R do not increase much.

Regression with Multiple Categorical and Quantitative Predictors

It is straightforward to generalize the regression models to include additional predictors of either type. To introduce additional quantitative variables, add a βx term for each one. To introduce another categorical variable, add a set of dummy variables for its categories. To permit interaction, introduce cross-product terms: Cross-products of x terms for interaction between quantitative predictors, cross-products between an x and a set of dummy variables for interaction between a quantitative and categorical predictor, and cross-products between dummy variables for interaction between two categorical predictors.

When there are several predictors, the number of potential models is quite large when we consider the possible main effect and interaction terms. When a variable occurs in an interaction term, recall that it is inappropriate to use the main effect term to summarize its effect, because that variable's effect changes as the value changes of a variable with which it interacts. Also, some variables may overlap considerably in the variation they explain in the response variable, so it may be possible to simplify the model by dropping some terms. We can use inference, as described in the next section, to help us select a model.

13.4 Inference for Regression with Quantitative and Categorical Predictors

This section presents significance tests and the following section presents estimation methods for analysis of covariance models. As in other multivariable models, we first test the hypothesis of no interaction. If a lack of interaction is plausible, then further analyses of the main effects treat the regression lines as parallel.

We test hypotheses about model parameters using the F test comparing complete and reduced regression models, introduced in Section 11.6. For instance, the test of no interaction compares the complete model containing the interaction terms to the reduced model deleting them. The test statistic for comparing models is

$$F = \frac{(\text{SSE}_r - \text{SSE}_c)/df_1}{\text{SSE}_c/df_2} = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2},$$

where SSE_r and SSE_c are the residual sums of squares (i.e., the sum of squared errors) for the reduced and complete models, df_1 equals the difference between the number of terms in the two models, and df_2 is the df value for the mean square for the complete model. Equivalently, we can use the second expression for the R^2 values of the complete and reduced models. Software presents such tests in an ANOVA table.

Test of No Interaction

For testing H_0: no interaction, the model under H_0 is the reduced model without the cross-product terms, which has a common slope for all lines. Figure 13.7 depicts the hypotheses for this test. This test has a small P-value if the addition of the cross-product interaction terms provides a significant improvement in the fit.

Figure 13.7: Graphical Representation of Null and Alternative Hypotheses in a Test of No Interaction, When the Categorical Predictor Has Three Categories

(Fig. 13.7 in 3e))

Example 13.3 Testing Interaction of Education and Racial–Ethnic Group in Their Effects on Income

For Table 13.1, we now test H_0: No interaction between education and racial–ethnic group, in their effects on income. The complete model

$$E(Y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x \times z_1 + \beta_4 x \times z_2,$$

contains two interaction terms. The null hypothesis is H_0: $\beta_3 = \beta_4 = 0$.

Table 13.7 shows how software summarizes sums of squares explained by various sets of terms in the model with interaction terms. The variability explained by the interaction terms, 691.8, equals the difference between the SSE values without and with those terms in the model. These sums of squares are ***partial sums of squares***, also called ***Type III sums of squares*** by some software (see Section 12.5 and Exercise 60 in Chapter 11). They represent the variability explained after the other terms are already in the model.

Table 13.7: Computer Printout of Partial Sums of Squares Explained by Education, Racial–Ethnic Group, and Their Interaction, in the Analysis of Covariance Model Permitting Interaction

Source	Type III Sum of Squares	df	Mean Square	F	Sig
Race	267.319	2	133.659	.566	.570
Education	6373.507	1	6373.507	26.993	.000
Race*Education	691.837	2	345.918	1.465	.238
Error	17472.412	74	236.114		
Total	33761.950	79			

For H_0: No interaction, the F test statistic is the ratio of the interaction mean square to the MSE. Table 13.7 shows that the test statistic is $F = 345.9/236.1 = 1.46$, with a P-value of 0.24. There is not much evidence of interaction. We are justified in using the simpler model without cross-product terms.

□

Test for Effect of Categorical Variable, Controlling for X

Possibly the model can be simplified further, if either of the main effects are not significant. Here, we test the main effect for the categorical predictor, racial–ethnic group. The null hypothesis states that each racial–ethnic group has the same regression line between x and Y. Equivalently, each group has the same mean on Y, controlling for x. This H_0 resembles H_0 in one-way ANOVA, except that this test compares the means while controlling for x.

For a categorical predictor with three categories, this test compares the complete model

$$E(Y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2$$

to the reduced model

$$E(Y) = \alpha + \beta x$$

lacking effects of the categorical predictor. The null hypothesis is

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{(coefficients of dummy variables = 0).}$$

The complete model represents three different but parallel regression lines between income and education, one for each racial–ethnic group. The reduced model states that the same regression line applies for all three groups. Figure 13.8 depicts this test. The P-value is small if the complete model with separate parallel lines provides a significantly better fit to the data than the reduced model of a common line.

Figure 13.8: Graphical Representation of Null and Alternative Hypotheses in a Test of Equivalence of Regression Lines, When the Categorical Variable Has Three Categories (Test Assumes No Interaction)

((Fig. 13.8 in 3e))

Example 13.4 Testing Effect of Racial–Ethnic Group on Income, Controlling for Education

Table 13.8 shows how software reports the results of tests for the no interaction model. The F statistic for the test of no effect of racial–ethnic group equals 3.06. Its P-value equals 0.053. There is some evidence, but not strong, that the regressions of Y on x are different for at least two of the racial–ethnic groups. The sample sizes for two of the three groups are very small, so this test does not have much power.

□

Table 13.8: Printout of Partial Sums of Squares Explained by Education and Racial–Ethnic Group, in the Analysis of Covariance Model Without Interaction

Source	Type III Sum of Squares	df	Mean Square	F	Sig
Race	1460.584	2	730.292	3.056	.053
Education	12245.232	1	12245.232	51.235	.000
Error	18164.248	76	239.003		
Total	33761.950	79			

Comparison with Results of ANOVA

The test just described compares the mean of Y among categories of a categorical variable, controlling for x. This is a different test than one-way ANOVA, because of the control for x. Table 13.9 shows the table for one-way ANOVA with these data. The P-value equals 0.018. Somewhat stronger evidence exists of a difference among the means when we ignore, rather than control, education.

Table 13.9: ANOVA Printout for Fitting Model $E(Y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ Corresponding to One-Way ANOVA of Income by Racial–Ethnic Group

Source	Type III Sum of Squares	df	Mean Square	F	Sig
Race	3352.470	2	1676.235	4.244	.018
Error	30409.480	77	394.928		
Total	33761.950	79			

Chapter 12 showed that ANOVA methods partition the total sum of squares $\sum(y-\overline{y})^2$ about the overall sample mean into component parts. The analysis of covariance, by contrast, partitions the term $\sum(x-\bar{x})(y-\overline{y})$ into component parts (Snedecor and Cochran, 1967). This term measures the *covariation* between X and Y (Exercise 68 in Chapter 9). That is, it measures the way in which X and Y vary jointly around their means $\bar{x}$ and $\bar{y}$. The cross-product $(x-\bar{x})(y-\overline{y})$ for an observation summarizes the amount by which those observations are both above or both below their means (in which case it is positive), or one above and the other below their means (in which case the cross-product is negative). The name *analysis of covariance* originates from this representation of the method.

Test for Effect of X, Controlling for Categorical Variable

Assuming no interaction, we have tested for the effect of the categorical variable. Alternatively, we could test for the effect of the quantitative variable. This tests $H_0 : \beta = 0$ in the model

$$E(Y) = \alpha + \beta x + \beta_1 z_1 + \beta_2 z_2.$$

The hypothesis states that the straight line relating x to the mean of Y has slope 0 for each category of the categorical variable. The null model in which $\beta = 0$ is the model for one-way ANOVA. Thus, testing the effect of x corresponds to comparing the models for analysis of variance and analysis of covariance.

Since H_0 specifies a value for a single parameter, we can perform the test using the t test. From Table 13.3, the estimated slope of 4.432 has a standard error of 0.619. The test statistic equals $t = 4.432/0.619 = 7.2$, which has a P-value of 0.000. The evidence is very strong that the true slope is positive. Equivalently, the square of this

t statistic equals the F statistic of 51.2 reported for the effect of education in Table 13.8.

Table 13.10 summarizes the hypotheses, residual sums of squares, and R^2 values for the models of this section. In bivariate models, education is a good predictor of income ($R^2 = 0.42$), considerably better than racial–ethnic group ($R^2 = 0.10$). There is some further reduction in error from using three parallel lines, rather than one line, to predict income ($R^2 = 0.46$). This is summarized by the second test in the table, referring to the marginal evidence of the effect of the categorical predictor, controlling for the quantitative predictor. There is a highly significant effect of the quantitative predictor, controlling for the categorical one, as summarized by the third test. However, a small and insignificant reduction in error occurs by allowing different slopes for the three lines ($R^2 = 0.48$), the first test listed in the table.

Table 13.10: Summary of Comparisons of Four Models

	Anal. Covariance Interaction	Anal. Covariance No Interaction	Bivariate Regression	One-Way ANOVA
	$E(Y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$ $+\beta_3(xz_1) + \beta_4(xz_2)$	$E(Y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$	$E(Y) = \alpha + \beta x$	$E(Y) = \alpha$ $+\beta_1 z_1 + \beta_2 z_2$
SSE	17472.4	18164.2	9812.4	30409.5
R^2	0.48	0.46	0.42	0.10
H_0: No interaction $F = 1.5, P = 0.24$	Complete model	Reduced model	—	—
H_0: $\beta_1 = \beta_2 = 0$ (Equal means, cont. for x) $F = 3.1, P = 0.053$	—	Complete model	Reduced model	—
H_0: $\beta = 0$ (Zero slopes) $F = 51.2, P = 0.000$	—	Complete model	—	Reduced model

13.5 Adjusted Means*

As usual, we learn more by estimating parameters than by testing hypotheses. This section shows how to summarize and compare means on Y for the categories of the categorical predictor, while controlling for the quantitative covariate. These analyses are meaningful when a model assuming no interaction is adequate.

Adjusting Response Means, Controlling for the Covariate

To summarize the means of Y for the groups while taking into account the groups' differing means on the covariate, we can report the values expected for the means if the groups all had the same mean on x. These values, which adjust for the groups' differing distributions on x, are called *adjusted means*.

Adjusted Mean

The **adjusted mean** of Y for a particular group is the regression function for that group evaluated at the overall mean of the x-values for all the groups. It represents the expected value for Y at the mean of x for the combined population.

Figure 13.9 illustrates the adjusted means. Since adjusted means are relevant when the no interaction model applies, this figure shows parallel lines. Let μ_x denote the mean of x for the combined population. The adjusted mean of Y for a particular group equals that group's regression function evaluated at μ_x.

Figure 13.9: Population Adjusted Means, When a Categorical Explanatory Variable Has Three Categories

((Fig. 13.9 in 3e))

The ***sample adjusted mean*** of Y for a group is the prediction equation for that group evaluated at $\bar{x}$, the overall sample mean of x. This estimates the value expected for the group's mean of Y if the mean of x for the group had equaled the overall mean.

Notation for Adjusted Means

Denote the sample adjusted mean for group i by $\overline{y}'_i$. This is the value of the prediction equation for group i evaluated at $\bar{x}$.

Example 13.5 Adjusted Mean Incomes, Controlling for Education

From Table 13.3, the prediction equation for the model assuming no interaction is

$$\hat{y} = -15.7 + 4.4x - 10.9z_1 - 4.9z_2.$$

Table 13.11 lists the prediction equations for the three racial–ethnic groups. The table also shows the unadjusted mean incomes and the adjusted mean incomes, controlling for education.

For blacks, for instance, the prediction equation is

$$\hat{y} = -26.54 + 4.43x.$$

Table 13.11: Sample Unadjusted Mean Incomes, Adjusted Means (Controlling for Education), and Prediction Equations for Model Assuming No Interaction

Group	Prediction Equation	Mean of x	Mean of Y	Adjusted Mean of Y
Blacks	$\hat{y} = -26.54 + 4.43x$	12.2	27.8	29.7
Hispanics	$\hat{y} = -20.60 + 4.43x$	11.6	31.0	35.6
Whites	$\hat{y} = -15.66 + 4.43x$	13.1	42.5	40.6

From Table 13.2, the mean education for the combined sample of 80 observations is $\bar{x} = 12.7$. So, the sample adjusted mean income for blacks, controlling for education, is

$$\overline{y}_1' = -26.54 + 4.43\bar{x} = -26.54 + 4.43(12.7) = 29.7.$$

Similarly, the sample adjusted means for Hispanics and whites are

$$\begin{aligned} \overline{y}_2' &= -20.60 + 4.43(12.7) = 35.6, \\ \overline{y}_3' &= -15.66 + 4.43(12.7) = 40.6. \end{aligned}$$

The adjustment process adjusts an ordinary sample mean upward or downward according to whether mean education for the group is below or above average. For whites, for instance, the adjusted mean income of 40.6 is smaller than the unadjusted mean of 42.5. The reason is that the mean education for whites ($\bar{x}_3 = 13.1$) is larger than the mean education for the combined sample ($\bar{x} = 12.7$). Since a positive relationship exists between income and education, the model predicts that whites would have a lower mean income if their mean education were lower (equal to $\bar{x} = 12.7$).

□

Comparing Adjusted Means

The coefficients of the dummy variables in the 'no interaction' model refer to differences between adjusted means. To illustrate, the estimated difference between adjusted mean incomes of blacks and whites is $\overline{y}_1' - \overline{y}_3' = 29.7 - 40.6 = -10.9$ (i.e., $-\$10,900$). This is precisely the coefficient of the dummy variable z_1 for blacks in the above prediction equation. Similarly, the estimated difference between the adjusted means of Hispanics and whites equals $\overline{y}_2' - \overline{y}_3' = -4.9$, which is the coefficient of z_2. Figure 13.10 depicts the sample adjusted means. The vertical distances between the lines represent the differences between these adjusted means.

Since the groups had different means on the covariate (education), the adjusted means and their differences are not the same as the unadjusted means and differences. For example, from Table 13.11, $\overline{y}_2 - \overline{y}_3 = 31.0 - 42.5 = -11.5$ is the unadjusted

Figure 13.10: Sample Adjusted Means on Income, Controlling for Education, for Three Racial–Ethnic Groups

((Fig. 13.10 in 3e, amended))

difference in mean income between Hispanics and whites. This is somewhat larger than the adjusted difference of -5.0. If whites and Hispanics had the same average educational level, the model predicts that the difference in their mean incomes would be less than half as large as it actually is in this sample.

Graphical Interpretation

Figure 13.11 depicts the relationship between the adjusted and unadjusted means. The figure depicts the line for the first group, but the same reasoning applies to any group. The prediction equation predicts a value of $\overline{y}_i$ at the x-value of $x = \bar{x}_i$ for group i. In particular, the prediction line for the first category passes through the point with coordinates $(\bar{x}_1, \overline{y}_1)$. In other words, the *unadjusted mean* $\overline{y}_1$ is the value of the prediction equation for that group evaluated at the x-value of $\bar{x}_1$, the mean of the x-values *for that group alone* [see the point $(\bar{x}_1, \overline{y}_1)$ in Figure 13.11].

Figure 13.11: Graphical Depiction of Sample Adjusted and Unadjusted Means for the First Group. The ordinary mean is the predicted value at the mean of x for that group alone, whereas the adjusted mean is the predicted value at the mean $\overline{x}$ for all the data.

((Fig. 13.11 in 3e))

The *adjusted mean* $\overline{y}_1'$ for the first group is the value of that prediction equation evaluated at the *overall mean* $\bar{x}$ for the combined sample. Hence, the prediction line for that category also passes through the point $(\bar{x}, \overline{y}_1')$, as Figure 13.11 shows. The difference between the x-coordinates for the points $(\bar{x}, \overline{y}_1')$ and $(\bar{x}_1, \overline{y}_1)$ is the horizontal distance $\bar{x} - \bar{x}_1$, indicated in the figure. Denoting the slope of the sample prediction equation by b, the vertical change in the line over the x distance of $\bar{x} - \bar{x}_1$ is $b(\bar{x} - \bar{x}_1)$. That is, the predicted change in the y direction for a $(\bar{x} - \bar{x}_1)$ change in the x-direction is $b(\bar{x} - \bar{x}_1)$. This vertical change is precisely the difference $\overline{y}_1' - \overline{y}_1$ between the y coordinates of the points $(\bar{x}, \overline{y}_1')$ and $(\bar{x}_1, \overline{y}_1)$.

In summary, the difference between the adjusted and unadjusted means for the first group equals

$$\overline{y}_1' - \overline{y}_1 = b(\bar{x} - \bar{x}_1).$$

Equivalently,

$$\overline{y}_1' = \overline{y}_1 + b(\bar{x} - \bar{x}_1).$$

The process of controlling for x takes the sample mean for y and adjusts it by multiplying the difference $(\bar{x} - \bar{x}_1)$ in the means of x by the slope b of the prediction equation.

If the slope b is positive, then the sample mean $\overline{y}_1$ is adjusted upward if $\bar{x} > \bar{x}_1$ (i.e., if $\bar{x} - \bar{x}_1 > 0$), as in Figure 13.11. This predicts that the mean of y would have been larger had the distribution of x-values for that group had as large a mean as the combined samples. If group 1 is above average on x (i.e., $\bar{x}_1 > \bar{x}$), then $b(\bar{x} - \bar{x}_1)$ is negative when $b > 0$, so the mean is adjusted downward.

Similar reasoning applies to other groups. Figure 13.12 depicts the adjustment process on income for the three racial–ethnic groups. The mean is adjusted down for whites and up for blacks and Hispanics.

Figure 13.12: Adjustment Process for Income by Racial–Ethnic Group, Controlling for Education

((Fig. 13.12 in 3e))

The difference between a group's adjusted and unadjusted means depends directly on the difference between $\bar{x}$ for the combined sample and $\bar{x}_i$ for that group. The adjusted means are similar to the unadjusted means if the $\bar{x}_i$-values are close to the overall $\bar{x}$, or if the slope b of the prediction equations is small.

Multiple Comparisons of Adjusted Means

Following an analysis of variance, the Bonferroni method compares all pairs of means simultaneously with a fixed overall confidence level. This method extends directly to multiple comparison of *adjusted means*. Software for multiple regression reports standard errors for these estimates. We can form t confidence intervals using these estimates and their standard errors.

Example 13.6 Confidence Intervals for Comparing Adjusted Mean Incomes

Let's construct 95% confidence intervals for differences between the three pairs of adjusted mean incomes, using the Bonferroni multiple comparison approach. The error probability for each interval is $0.05/3 = 0.0167$. The t-score with single-tail probability $0.0167/2 = 0.0083$ and $df = 76$ (which is df for SSE for the no-interaction model) is 2.45.

The estimated difference between adjusted mean incomes of Hispanics and whites is the coefficient of the dummy variable z_2 for Hispanics in the prediction equation. Table 13.3 showed the racial–ethnic effects

Parameter	B	Std. Error	t	Sig
[race = b]	-10.874	4.473	-2.431	.017
[race = h]	-4.934	4.763	-1.036	.304

so this coefficient is -4.934 and has a standard error of 4.763. The Bonferroni confidence interval equals

$$-4.934 \pm 2.45(4.763), \quad \text{or } (-16.6, 6.7).$$

Controlling for education, the difference in mean incomes for Hispanics and whites is estimated to fall between $-\$16,600$ and \$6700. Since the interval contains 0, it is plausible that the true adjusted mean incomes are equal. The sample contained only 14 Hispanics, so the interval is wide.

Similarly, from Table 13.3, the confidence interval comparing blacks and whites is $-10.874 \pm 2.45(4.473)$, or $(-21.8, 0.1)$. To get the standard error for the estimate $b_1 - b_2 = (-10.87 - (-4.93)) = -5.94$ comparing blacks and Hispanics, we could fit the model with one of these categories as the baseline category lacking a dummy variable. Or, we could use the general expression to get se from the values se_1 for b_1 and se_2 for b_2 as

$$se = \sqrt{(se_1)^2 + (se_2)^2 - 2\text{Cov}(b_1, b_2)},$$

where $\text{Cov}(b_1, b_2)$ is taken from the *covariance matrix* of the parameter estimates, which software can provide. For these data, the standard error for $b_1 - b_2$ equals 5.67, and the confidence interval is $(-19.8,\ 8.0)$.

Table 13.12 summarizes the comparisons. We can be 95% confident that all three of these intervals contain the differences in population adjusted means. None of the intervals show a significant difference, which is not surprising because the F test of the group effect in the previous section had a P-value of 0.053. Nonetheless, the intervals show that the adjusted means could be quite a bit smaller for blacks or Hispanics than for whites. More precise estimation requires a larger sample.

□

For the model with no interaction, testing for identical adjusted means is equivalent to testing whether the regression lines are identical. That is, two parallel lines are identical if their corresponding adjusted means are equal. Therefore, just as confidence intervals reveal which pairs of adjusted means are significantly different, so do they reveal which pairs of lines are significantly different. For example, if the interval for the difference between a pair of adjusted means does not contain 0, then the regression lines between Y and x for those two categories are judged to have different intercepts.

Table 13.12: Bonferroni Multiple Comparisons of Differences in Adjusted Mean Income by Racial–Ethnic Group, Controlling for Education

Racial–Ethnic Groups	Estimated Difference in Adjusted Means	95% Bonferroni Confidence Intervals
Blacks, Whites	$\overline{y}'_1 - \overline{y}'_3 = -10.9$	$(-21.8, 0.1)$
Hispanics, Whites	$\overline{y}'_2 - \overline{y}'_3 = -4.9$	$(-16.6, 6.7)$
Blacks, Hispanics	$\overline{y}'_1 - \overline{y}'_2 = -5.9$	$(-19.8, 8.0)$

A Caution

Adjusted means are useful for comparing several groups by adjusting for differences in the means of a covariate x. Use them with caution, however, when the means on x are greatly different. The control process is a hypothetical one that infers what would happen *if* all groups had the same mean for x. If large differences exist among the groups in their means on x, the results of this control may be purely speculative. We must assume (1) that it makes sense to conceive of adjusting the groups on this covariate and (2) that the relationship between Y and x would continue to have the same linear form within each category as the x mean shifts for each category.

To illustrate, recall the relationship between annual income and experience and gender shown in Figure 13.3b. The same line fits the relationship between income and experience for each gender, so it is plausible that the adjusted means are equal. However, nearly all the women have less experience than the men. The conclusion that the mean incomes are equal, controlling for experience, assumes that the regression line shown also applies to women with more experience than those in the sample and to men with less experience. If it does not, then the conclusion is incorrect.

Figure 13.13 portrays a situation in which the conclusion would be misleading. The dotted lines show the relationship for each group over the x region not observed. At each fixed x value, a difference persists between the means of Y.

Figure 13.13: A Situation in Which Adjusted Means Are Misleading, Comparing Mean Incomes for Men and Women While Controlling for Experience

((Fig. 13.13 in 3e))

Whenever we use adjusted means, we should check the degree to which the distributions differ on the mean of x. Excessively large differences may mean that the conclusions need qualification. On the other hand, if relatively small differences exist among the $\bar{x}_i$, then controlling for x has little effect. The results of the comparisons of adjusted means are then similar to the results of the comparisons of unadjusted means in an analysis of variance.

13.6 Chapter Summary

This chapter showed that multiple regression can describe the relationship between a quantitative response variable and both quantitative and categorical explanatory variables. When the method is used to compare the mean of Y for different categories while controlling for a quantitative *covariate*, it is called the ***analysis of covariance***.

- The multiple regression model has linear terms (such as βx) for quantitative predictors and dummy variable terms for categorical predictors.
- There is ***no interaction*** if the slope of the line relating x to the mean of Y is the same for each group. The model then provides a set of parallel lines. We allow interaction and different slopes by entering cross products of quantitative predictors with dummy variables in the model.
- ***Adjusted means*** summarize the means on Y for the groups while controlling for x. They represent the model's prediction for the means of Y at the overall mean of the x values.

Adjusted means are meaningful only when there is no interaction. One can test the hypothesis of no interaction, as well as the hypothesis of equal adjusted means (or, equivalently, identical regression lines), using an F test for the relevant parameters in the model.

PROBLEMS

Practicing the Basics

1. The regression equation relating Y = education (number of years completed) to race ($z = 1$ for whites, $z = 0$ for nonwhites) in a certain country is $E(Y) = 11 + 2z$. The regression equation relating education to race and to father's education (x) is $E(Y) = 3 + 0.8x - 0.6z$.
   a) Ignoring father's education, find the mean education for whites, the mean education for nonwhites, and the difference between them.
   b) Plot the relationship between x and the mean of Y for whites and for nonwhites.
   c) Controlling for father's education, find the difference between the mean education of whites and nonwhites.

d) Find the mean education for whites and for nonwhites, when father's education equals 12 years.

2. Table 27 in Chapter 9 showed data for several nations on various indices. Let Y = percentage who use the Internet, x = per capita GDP (in thousands of dollars), and z = whether the nation is a Western nation (1 = yes, 0 = no).
a) The prediction equation for the effect of z is $\hat{y} = 10.5 + 30.4z$. Interpret the coefficients.
b) The prediction equation for the effects of x and z is $\hat{y} = 8.3 + 0.73x + 10.9z$. Interpret the coefficients.
c) Explain how the coefficient of z could be so different in (a) and (b).

3. A regression analysis for the 100th Congress predicted the proportion of each representative's votes on abortion issues that took the 'pro-choice' position (R. Tatalovich and D. Schier, *American Politics Quarterly*, Vol. 21, 1993, p. 125). The prediction equation was

$$\hat{y} = 0.350 + 0.011id + 0.094r + 0.005nw + 0.005inc + 0.063s - 0.167p$$

where r = religion (1 for non-Catholics), s = sex (1 for women), p = political party (1 for Democrats), id = ideology is the member's ADA score (ranging from 0 at most conservative to 100 at most liberal), nw = nonwhite is the percentage nonwhite of the member's district, and inc = income is the median family income of the member's district.
a) Interpret the coefficient for party.
b) Using standardized variables, the prediction equation is

$$\hat{z}_Y = 0.83z_{id} + 0.21z_r + 0.18z_{nw} + 0.05z_{inc} + 0.03z_s - 0.18z_p$$

Comment on the relative sizes of the partial effects. Interpret the coefficient of ideology.

4. For data[1] from 27 automotive plants on Y = number of assembly defects per 100 cars and x = time to assemble each vehicle (in hours, falling between 12 and 54), $\hat{y} = 61.3 + 0.35x$. Adding z = whether facility is Japanese (1 = yes, 0 = no) to the model, $\hat{y} = 105.0 - 0.78x - 36.2z$.
a) Interpret the estimated partial effects.
b) Explain why these equations satisfy Simpson's paradox, which states that a partial effect can have the reverse sign from the bivariate effect.
c) For these data, x values tended to be lower for the Japanese plants. Explain how this could be responsible for Simpson's paradox occurring.

5. Based on a national survey, Table 13.13 shows results of a prediction equation reported for Y = alcohol consumption, measured as the number of alcoholic

[1] *Source*: S. Chatterjee et al. *A Casebook for a First Course in Statistics and Data Analysis*, Wiley, 1995.

drinks the subject drank during the past month (D. Umberson and M. Chen, *American Sociological Review*, Vol. 59, 1994, p. 152).
a) Setting up dummy variables f for whether father died in the past three years, s for sex, and (m_1, m_2, m_3) for the four categories of marital status, and letting x = alcohol consumption three years ago, report the prediction equation.
b) Find the predicted alcohol consumption for a divorced male whose father died in the previous three years and whose consumption three years previously was ten drinks per month.
c) Interpret the coefficient of sex.

Table 13.13:

Explanatory Variable	Estimate	Std. Error
Intercept	8.3	
Death of father (0 = no)	9.8	2.9
Sex (0 = male)	-5.3	1.6
Marital status (0 = married)		
Divorced, separated	7.0	2.0
Widowed	2.0	3.6
Never married	1.2	2.4
Alcohol consumption, three years ago	0.501	0.023

6. For the survey in the previous exercise, the sample size was 1417.
a) Test the null hypothesis that sex has no effect on the response, controlling for the other predictors. Interpret.
b) Construct a 95% confidence interval for the sex effect, controlling for the other predictors. Interpret.
c) Marital status has three estimates. Dividing the coefficient of the divorced dummy variable by its standard error yields a t statistic. What hypothesis does it test?
d) What would you need to do to test the effect of marital status (all categories at once), controlling for the other variables?

7. For the "house selling price" data file at the text website, Table 13.14 shows results of modeling Y = selling price (in dollars) in terms of size of home (in square feet) and whether the home is new (1 = yes; 0 = no).
a) Report and interpret the prediction equation, and form separate equations relating selling price to size for new and for not new homes.
b) Find the predicted selling price for a home of 3 thousand square feet that is (i) new, (ii) not new.

8. Refer to the previous exercise. Table 13.15 shows results of fitting the model allowing interaction, where new*size refers to the cross-product term.

Table 13.14:

Parameter	B	Std. Error	t	Sig
Intercept	-40230.867	14696.140	-2.738	.007
size	116.132	8.795	13.204	.000
new	57736.283	18653.041	3.095	.003

a) Report the lines relating predicted selling price to size for homes that are (i) new, (ii) not new.
b) Find the predicted selling price for a home of 3 thousand square feet that is (i) new, (ii) not new.
c) Find the predicted selling price for a home of 1.5 thousand square feet that is (i) new, (ii) not new. Comparing to (b), explain how the difference in predicted selling prices changes as size of home increases.

Table 13.15:

Parameter	B	Std. Error	t	Sig
Intercept	-22227.808	15521.110	-1.432	.155
size	104.438	9.424	11.082	.000
new	-78527.502	51007.642	-1.540	.127
new*size	61.916	21.686	2.855	.005

9. The printouts in Table 13.16 show results of fitting two models to data from a study of the relationship between $Y =$ percentage of adults voting, percentage of adults registered to vote, and racial–ethnic representation, for a random sample of 40 precincts in the state of Texas for a gubernatorial election. Racial–ethnic representation of a precinct is the group (Anglo, black, or Mexican American) having the strongest representation in a precinct.
a) Report the prediction equation for the model assuming no interaction. Interpret the parameter estimates.
b) Report the prediction equation for the model allowing interaction. Interpret the parameter estimates and describe the nature of the estimated interaction.
c) Test whether the regression lines for the three categories have the same slope. Report the test statistic and P-value, and interpret.
d) For the model assuming no interaction, test whether the mean voting percentages are equal for the three categories of racial-ethnic representation, controlling for percentage registered. Report the test statistic and P-value, and interpret.

e) Report the test statistic and P-value for testing the null hypothesis that percentage voting and percentage registered are independent, controlling for racial–ethnic representation. Interpret.

Table 13.16:

NO INTERACTION MODEL

Source	Type III Sum of Squares	df	Mean Square	F	Sig
race	40.08	2	20.04	1.07	.354
register	2317.43	1	2317.43	123.93	.000

Source	Sum of Squares	df	Mean Square	Parameter		Estimate
Model	7936.734	3	2645.578	INTERCEPT		-2.7786
Error	673.166	36	18.699	REGISTER		0.7400
Total	8609.900	39		RACE	a	-1.3106
					b	-2.8522
					ma	0.0000

INTERACTION MODEL

Source	Type III Sum of Squares	df	Mean Square	F	Sig
race*register	53.79	2	27.89	1.47	.243

Source	Sum of Squares	df	Mean Square	Parameter		Estimate
Model	7990.523	5	1598.105	INTERCEPT		-8.245
Error	619.377	34	18.217	REGISTER		0.878
Total	8609.900	39		RACE	a	6.974
					b	9.804
					ma	0.000
				REGISTER*RACE	a	-0.175
					b	-0.283
					ma	0.000

10. Refer to the previous exercise. The means of percentage registered for the three categories are $\bar{x}_1 = 76.2$, $\bar{x}_2 = 49.5$, $\bar{x}_3 = 39.7$. The overall mean $\bar{x} = 60.4$.
a) Find the adjusted mean on percentage voting for Anglos. Compare it to the unadjusted mean of 52.3, and interpret.
b) Sketch a plot of the no interaction model for these data, and identify on it the unadjusted and adjusted means for Anglos.
c) Describe how to expand the no interaction model to include the additional

categorical variable, location of precinct (with categories urban, suburban, small town, and rural) and the additional quantitative explanatory variable, percentage of residents in precinct who are homeowners.

11. Table 13.1 did not report the observations for ten Asian Americans. Their (x, y) values were:

Subject	1	2	3	4	5	6	7	8	9	10
Education	16	14	12	18	13	12	16	16	14	10
Income	70	42	24	56	32	38	58	82	36	20

a) Fit the model using all four groups, assuming no interaction. (The data are available at the text website.) Interpret the parameter estimates.
b) Using all four groups, fit the interaction model. Interpret the parameter estimates.
c) For the model in (b), explain how the estimated difference in mean incomes between blacks and whites changes as education changes.
d) Report R^2 for the two models. Does the model allowing interaction seem to provide much stronger predictive power?

12. Refer to the previous exercise.
a) Test the hypothesis of no interaction. Report the test statistic, P-value, and interpret.
b) Assuming no interaction, test the hypothesis that income is independent of ethnic–racial group, controlling for education. Report the test statistic and P-value, and interpret.
c) Assuming no interaction, test the hypothesis that income is independent of education, controlling for ethnic–racial group. Report the test statistic and P-value, and interpret.
d) Construct a confidence interval for the difference between population mean incomes of blacks and whites, controlling for education. Use the appropriate confidence level for forming Bonferroni 94% confidence intervals for all six pairs of differences. Interpret.

13. Exercise 13.1 reported the regression equation relating Y = education to race ($z = 1$ for whites) and to father's education (x) of $E(Y) = 3 + 0.8x - 0.6z$. The mean education is 11 for nonwhites and 13 for whites. The overall mean of father's education is 12 years.
a) Find the adjusted mean educational levels for whites and nonwhites, controlling for father's education.
b) Explain why the adjusted means differ as they do from the unadjusted means.

Concepts and Applications

14. Refer to the student survey data set (Exercise 1.11). Using software, prepare a report presenting graphical, descriptive, and inferential analyses with
a) Y = political ideology and the predictors religiosity and whether a vegetarian.
b) Y = college GPA with predictors high school GPA, gender, and religiosity.

15. Refer to the data file your class created in Exercise 1.12. For variables chosen by your instructor, use regression analysis as the basis of descriptive and inferential statistical analyses. Summarize your findings in a report in which you describe and interpret the fitted models and the related analyses.

16. An article[2] on predicting attitudes toward homosexuality used GSS data to model a response variable with a 4-point scale in which homosexual relations were scaled from 1 = always wrong to 4 = never wrong, with x_1 = education (in years), x_2 = age, x_3 = political conservative (1 = yes, 0 = no), x_4 = religious fundamentalist (1 = yes, 0 = no), and x_5 = whether live in same city as when age 16 (1 = yes, 0 = no). The prediction equation allowing interaction between x_1 and x_3 is

$$\hat{y} = 0.94 + 0.13x_1 - 0.01x_2 + 1.10x_3 - 0.38x_4 - 0.15x_5 - 0.12(x_1 \times x_3).$$

a) Report the prediction equations for political conservatives and nonconservatives. Explain how these suggest that greater education corresponds to less negative views about homosexuality for nonconservatives but may have no effect for conservatives.
b) The model without interaction had fit

$$\hat{y} = 1.53 + 0.09x_1 - 0.01x_2 - 0.49x_3 - 0.39x_4 - 0.15x_5.$$

Summarize the effect of each predictor.

17. Table 13.17 shows results of fitting a regression model to data on salaries (in dollars) of about 35,000 college professors. Four predictors are categorical (binary), with dummy variable defined in parentheses. Write a short report, interpreting the effects of the predictors.

18. Table 13.18 is a printout based on GSS data. The response variable is an index of attitudes toward premarital, extramarital, and homosexual sex. Higher scores represent more permissive attitudes. The categorical explanatory variables are race (0 for whites, 1 for blacks), gender (0 for males, 1 for females), region (0 for South, 1 for non-South), and religion ($r_1 = 1$ for liberal Protestant sect, $r_2 = 1$ for conservative Protestant, $r_3 = 1$ for fundamentalist Protestant sect, $r_4 = 1$ for Catholic, $r_5 = 1$ for Jewish; no religious affiliation when $r_1 = \cdots = r_5 = 0$). The quantitative explanatory variables are age, education (number of years), attendance at church (higher values represent more frequent attendance), and a variable for which higher values represent greater intolerance of freedom of speech for atheists and communists.
a) State the prediction equation for this analysis.
b) Based on the parameter estimates, give a profile of a person you would expect to be (i) least permissive, (ii) most permissive, with respect to sexual attitudes.
c) Summarize the main conclusions that you make from studying this printout.

[2] T. Shackelford and A. Besser, *Individual Differences Research*, 2007

Table 13.17:

Variable	Estimate (Std. Error)
Years of Experience	162.7
	(2.0)
Terminal degree	264.0
(Ph.D. = 1)	(40.1)
Degree quality	616.8
	(34.5)
Research productivity	616.5
	(7.7)
Marital status	850.3
(married = 1)	(51.2)
Race	-310.3
(nonwhite = 1)	(107.0)
Gender	-1863.7
(female = 1)	(62.1)
Intercept	10,033.2

Source: N. Langton and J. Pfeffer, *American Sociological Review*, Vol. 59 (1994), p. 236.

19. A researcher studies factors associated with fertility (a woman's number of children) in a Latin American city. Of particular interest is whether migrants from other cities or migrants from rural areas differ from natives of the city in their family sizes. The groups to be compared are urban natives, urban migrants, and rural migrants. Since fertility is negatively related to educational level, and since education might differ among the three groups, it is decided to control that variable. Table 13.19 shows data for a random sample of married women above age 45. Analyze these data. In your report, provide graphical presentations as well as interpretations for all your analyses, and present a summary of the main results.

20. Analyze the "house selling price 2" data file at the text website, by modeling selling price in terms of size of house and whether it is new.
a) Fit the model allowing interaction, and test whether the interaction term is needed in the model.
b) Figure 13.14 shows a scatterplot, labeling the points by a 1 when the home is new and a 0 when it is not. The observation with the highest selling price is a new home that is somewhat removed from the general trend of points. Table 13.20 shows a printout for the interaction model after removing this single observation. Again, test whether the interaction term is needed in the model. Note what a large impact one observation can have on the conclusions.

Table 13.18:

Analysis of Variance

	Sum of Squares	df	Mean Square	F
Regression	2583.326	12	215.277	54.098
Residual	4345.534	1092	3.979	

R Square 0.373

Variable	B	Std. Error	Beta	t	Sig
(Constant)	9.373				
RACE	0.993	0.2040	0.125	4.869	.000
AGE	-0.029	0.0042	-0.189	-6.957	.000
SEX	-0.289	0.1230	-0.058	-2.353	.019
EDUC	0.073	0.0223	0.092	3.281	.001
REGION	0.617	0.1401	0.115	4.403	.000
ATTEND	-0.286	0.0255	-0.304	-11.217	.000
R1	-0.296	0.2826	-0.049	-1.048	.295
R2	-0.605	0.2782	-0.113	-2.174	.030
R3	-1.187	0.3438	-0.128	-3.454	.001
R4	-0.127	0.2856	0.023	0.446	.656
R5	0.521	0.4417	0.034	1.179	.238
FREESPCH	-0.465	0.0581	-0.227	-8.011	.000

Table 13.19:

Urban Natives		Urban Migrants		Rural Migrants	
Education	Fertility	Education	Fertility	Education	Fertility
0	7	0	7	0	4
0	5	0	6	0	6
1	5	0	7	0	10
1	4	1	5	0	8
2	7	2	2	1	7
4	4	2	6	2	8
5	4	3	3	3	5
7	3	4	6	3	6
8	5	7	4	4	7
8	2	7	4	5	7
8	3	8	4	6	8
9	3	11	3	6	6
10	4	11	3	7	4
11	3	11	4	7	5
12	3	12	4	8	4
		12	2	8	6
		12	3	8	5
				9	3
				9	7
				10	4

Table 13.20:

```
          Parameter   Standard
Variable   Estimate    Error         t          Sig
---------------------------------------------------
INTERCEP   -16.600     5.794     -2.865      .0052
NEW          8.999    17.553      0.513      .6095
SIZE        66.604     3.446     19.328      .0001
NEW*SIZE     4.973    10.015      0.497      .6208
```

```
      ----+------+------+------+------+------+------+------+------+-----
  300 +                                                 1              +
P     |                                                         0      |
R     |                                                                |
I     |                                                                |
C     |                                                                |
E 200 +                                       1                        +
      |                                   1                            |
      |                                0                               |
      |                               1100  0        0                 |
      |                         11111 00 0  0                          |
  100 +                      11111 10001                               +
      |                   10000 100    0                               |
      |                 0  00000  0                                    |
      |              0  00                                             |
      |         0   0 0     0                                          |
    0 +                                                                +
      |                                                                |
      ----+------+------+------+------+------+------+------+------+-----
         0.0    0.5    1.0    1.5    2.0    2.5    3.0    3.5    4.0
                                    SIZE
```

Figure 13.14:

21. For the "2005 statewide crime" data file at the text website, let z be a dummy variable for whether a state is in the South, with $z = 1$ for AL, AR, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV.
a) Not including the observation for D.C., analyze the relationship between Y = violent crime rate and z, both ignoring and controlling for x = poverty rate. Summarize your analyses and results.
b) Repeat the analysis with D.C. in the data set, setting $z = 1$ for it. Is this observation influential?

22. You have two groups, and you want to compare their regressions of Y on x, in order to test the hypothesis that the true slopes are identical for the two groups. Explain how you can do this using regression modeling.

23. In analyzing GSS data relating Y = frequency of having sex in the past year to frequency of going to bars, DeMaris (2004, p. 62) noted that the slope for unmarried subjects is more than double the slope for married subjects. Introducing notation, state a model that you think would be appropriate.

24. Let Y = death rate and x = mean age of residents, measured for each county in Louisiana and in Florida. Sketch a hypothetical scatterplot, identifying points for each state, when the mean death rate is higher in Florida than in Louisiana when mean age is ignored but lower when it is controlled. (Hint: When you fit a line for each state, the line should be higher for Louisiana but the y-values for Florida should have an overall higher mean.)

25. Draw a scatterplot with sets of points representing two groups such that H_0: equal means would be rejected in a one-way ANOVA but not in an analysis of covariance.

26. A regression model is fitted to annual income (thousands of dollars), using predictors age and marital status. Table 13.21 shows the sample mean incomes and the adjusted means for the model. How could the adjusted means be so different from the unadjusted means? Draw a sketch to help explain.

Table 13.21:

Group	Mean Age	Mean Income	Adjusted Mean Income
Married	44	40	30
Divorced	35	30	30
Single	26	20	30

In Exercises 27 – 28, select the correct response(s).

27. In the model $E(Y) = \alpha + \beta_1 x + \beta_2 z$, where z is a dummy variable,
a) The categorical predictor has two categories.

b) One line has slope β_1 and the other has slope β_2.
c) β_2 is the difference between the mean of Y for the second and first categories of the categorical variable.
d) β_2 is the difference between the mean of Y for the second and first categories of the categorical variable, controlling for x.

28. In the United States, the mean annual income for blacks (μ_1) is smaller than for whites (μ_2), the mean number of years of education is smaller for blacks than for whites, and annual income is positively related to number of years of education. Assuming that there is no interaction, the difference in the mean annual income between whites and blacks, controlling for education, is
a) Less than $\mu_2 - \mu_1$
b) Greater than $\mu_2 - \mu_1$
c) Possibly equal to $\mu_2 - \mu_1$

29. Summarize the differences in purpose of a one-way analysis of variance and an analysis of covariance.

30. * Suppose we use a centered variable for the covariate and express the interaction model when the categorical predictor has two categories as

$$E(Y) = \alpha + \beta_1(x - \mu_X) + \beta_2 z + \beta_3(x - \mu_X) \times z.$$

Explain how to interpret β_2, and explain how this differs from the interpretation for the model without a centered covariate.

13.6.1 Bibliography

Blalock, H. M. (1979). *Social Statistics*, rev. 2nd ed. New York: McGraw-Hill.

DeMaris, A. (2004). *Regression with Social Data: Modeling Continuous and Limited Response Variables*. Wiley.

Kutner, M.H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*, 5th ed. McGraw-Hill/Irwin.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th ed. Ames: Iowa State University Press.