STATISTICAL METHODS FOR THE SOCIAL SCIENCES

Fifth Edition

Alan Agresti

University of Florida

Pearson

Chapter 11

MULTIPLE REGRESSION AND CORRELATION

CHAPTER OUTLINE

Chapter 9 introduced regression modeling of the relationship between two quantitative variables. Multivariate relationships require more complex models, containing several explanatory variables. Some of these may be predictors of theoretical interest, and some may be control variables.

This chapter extends the regression model to a ***multiple regression model*** that can have multiple explanatory variables. Such a model provides better predictions of y than does a model with a single explanatory variable. The model also can analyze relationships between variables while controlling for other variables. This is important because Chapter 10 showed that after controlling for a variable, an association can appear quite different from when the variable is ignored.

After defining the multiple regression model and showing how to interpret its parameters, we present correlation and r-squared measures that describe association between y and a set of explanatory variables, and we present inference procedures for the model parameters. We then show how to allow *statistical interaction* in the model, whereby the effect of an explanatory variable changes according to the value of another explanatory variable. A significance test can analyze whether a complex model, such as one permitting interaction, provides a better fit than a simpler model. The final two sections introduce correlation-type measures that summarize the association between the response variable and an explanatory variable while controlling other variables.

11.1 The Multiple Regression Model

Chapter 9 modeled the relationship between the explanatory variable x and the mean of the response variable y by the straight-line (linear) equation $E(y) = \alpha + \beta x$. We refer to this model containing a *single* predictor as a ***bivariate model***, because it contains only two variables.

THE MULTIPLE REGRESSION FUNCTION

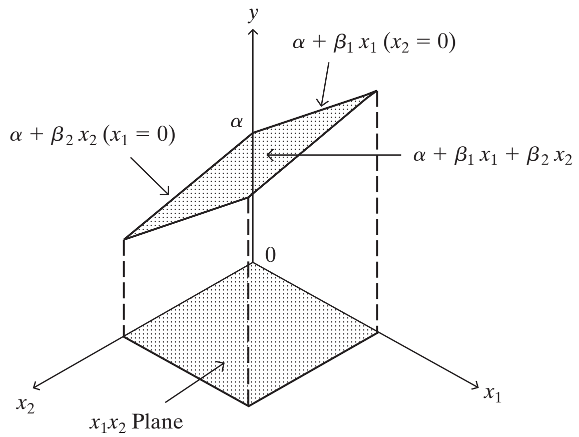
With two explanatory variables, denoted by x_1 and x_2, the bivariate regression function generalizes to the ***multiple regression function***

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

In this equation, α, β_1, and β_2 are parameters discussed below. For particular values of x_1 and x_2, the equation specifies the population mean of y for all subjects with those values of x_1 and x_2. With additional explanatory variables, each has a βx term, such as $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ with four explanatory variables.

The multiple regression function is more difficult to portray graphically than the bivariate regression function. With two explanatory variables, the x_1 and x_2 axes are perpendicular but lie in a horizontal plane and the y axis is vertical and perpendicular to both the x_1 and x_2 axes. The equation $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ traces a plane (a flat surface) cutting through three-dimensional space, as Figure 11.1 portrays.

FIGURE 11.1: Graphical Depiction of a Multiple Regression Function with Two Explanatory Variables


y
$\alpha + \beta_1 x_1 \ (x_2 = 0)$
α
$\alpha + \beta_2 x_2 \ (x_1 = 0)$
$\alpha + \beta_1 x_1 + \beta_2 x_2$
0
x_2
x_1
x_1x_2 Plane

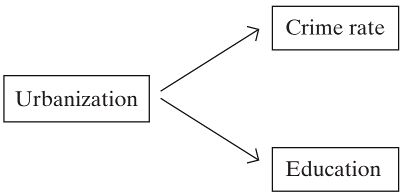
The simplest interpretation treats all but one explanatory variable as control variables and fixes them at particular levels. This leaves an equation relating the mean of y to the remaining explanatory variables.

Example 11.1

Do Higher Levels of Education Cause Higher Crime Rates? The `Florida` data file at the text website, shown partly in Table 9.15 on page 283, contains data for the 67 counties in the state of Florida on y = crime rate (annual number of crimes per 1000 population), x_1 = education (percentage of adult residents having at least a high school education), and x_2 = urbanization (percentage living in an urban environment). The bivariate relationship between crime rate and education is approximated by $E(y) = -51.3 + 1.5x_1$. Surprisingly, the association is moderately *positive*, the correlation being $r = 0.47$. As the percentage of county residents having at least a high school education increases, so does the crime rate.

A closer look at the data reveals strong positive associations between crime rate and urbanization ($r = 0.68$) and between education and urbanization ($r = 0.79$). This suggests that the association between crime rate and education may be spurious. Perhaps urbanization is a common causal factor. See Figure 11.2. As urbanization increases, both crime rate and education increase, resulting in a positive correlation between crime rate and education.

FIGURE 11.2: The Positive Association between Crime Rate and Education May Be Spurious, Explained by the Effects of Urbanization on Each



The relation between crime rate and both explanatory variables considered together is approximated by the multiple regression function

$$E(y) = 58.9 - 0.6x_1 + 0.7x_2.$$

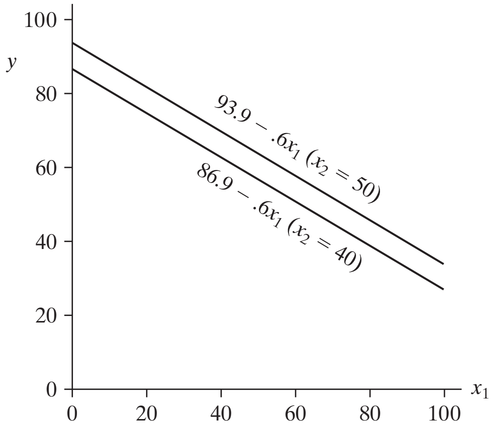
For instance, the expected crime rate for a county at the mean levels of education ($\bar{x}_1 = 70$) and urbanization ($\bar{x}_2 = 50$) is $E(y) = 58.9 - 0.6(70) + 0.7(50) = 52$ annual crimes per 1000 population.

Let's study the effect of x_1, controlling for x_2. We first set x_2 at its mean level of 50. Then, the relationship between crime rate and education is

$$E(y) = 58.9 - 0.6x_1 + 0.7(50) = 58.9 - 0.6x_1 + 35.0 = 93.9 - 0.6x_1.$$

Figure 11.3 plots this line. Controlling for x_2 by fixing it at 50, the relationship between crime rate and education is negative, rather than positive. The slope decreased and changed sign from 1.5 in the bivariate relationship to -0.6. At this fixed level of urbanization, a negative relationship exists between education and crime rate. We use the term *partial* regression equation to distinguish the equation $E(y) = 93.9 - 0.6x_1$ from the regression equation $E(y) = -51.3 + 1.5x_1$ for the *bivariate* relationship between y and x_1. The *partial* regression equation refers to *part* of the potential observations, in this case counties having $x_2 = 50$.

FIGURE 11.3: Partial Relationships between $E(y)$ and x_1 for the Multiple Regression Equation $E(y) = 58.9 - 0.6x_1 + 0.7x_2$. These partial regression equations fix x_2 to equal 50 or 40.



Next we fix x_2 at a different level, say $x_2 = 40$ instead of 50. Then, you can check that $E(y) = 86.9 - 0.6x_1$. Thus, decreasing x_2 by 10 units shifts the partial line relating y to x_1 downward by $10\beta_2 = 7.0$ units (see Figure 11.3). The slope of -0.6 for the partial relationship remains the same, so the line is parallel to the one with $x_2 = 40$. Setting x_2 at a variety of values yields a collection of parallel lines, each having slope $\beta_1 = -0.6$.

Similarly, setting x_1 at a variety of values yields a collection of parallel lines, each having slope 0.7, relating the mean of y to x_2. In other words, controlling for education, the slope of the partial relationship between crime rate and urbanization is $\beta_2 = 0.7$.

In summary, education has an overall positive effect on crime rate, but it has a negative effect when controlling for urbanization. The partial association has the opposite direction from the bivariate association. This is called ***Simpson's paradox***. Figure 11.4 illustrates how this happens. It shows the scatterplot relating crime rate to education, portraying the overall positive association between these variables. The diagram circles the 19 counties that are highest in urbanization. That subset of points for which urbanization is nearly constant has a negative trend between crime rate and education. The high positive association between education and urbanization is reflected by the fact that most of the highlighted observations that are highest on urbanization also have high values on education. ■

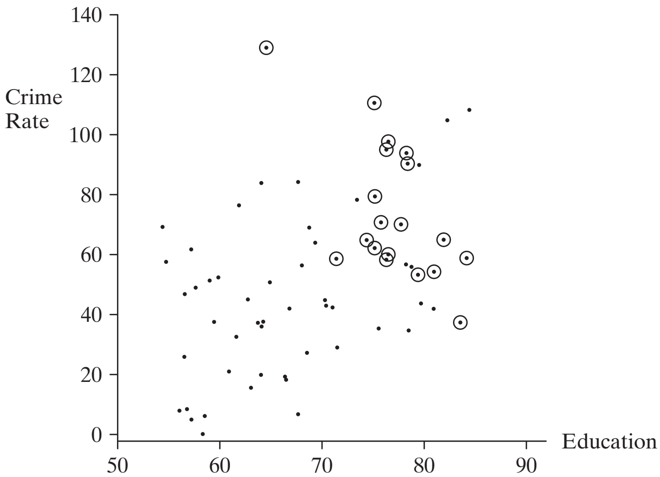


FIGURE 11.4: Scatterplot Relating Crime Rate and Education. The circled points are the counties highest in urbanization. A regression line fitting the circled points has negative slope, even though the regression line passing through *all* the points has positive slope (Simpson's paradox).

INTERPRETATION OF REGRESSION COEFFICIENTS

We have seen that for a fixed value of x_2, the equation $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$ simplifies to a straight-line equation in x_1 with slope β_1. The slope is the same for each fixed value of x_2. When we fix the value of x_2, we are holding it constant: We are *controlling* for x_2. That's the basis of the major difference between the interpretation of slopes in multiple regression and in bivariate regression:

- In *multiple regression*, a slope describes the effect of an explanatory variable while *controlling* effects of the other explanatory variables in the model.
- *Bivariate regression* has only a single explanatory variable. So, a slope in bivariate regression describes the effect of that variable while *ignoring* all other possible explanatory variables.

The parameter β_1 measures the *partial effect* of x_1 on y, that is, the effect of a one-unit increase in x_1, holding x_2 constant. The partial effect of x_2 on y, holding x_1 constant, has slope β_2. Similarly, for the multiple regression model with *several* explanatory variables, the beta coefficient of a particular explanatory variable describes the change in the mean of y for a one-unit increase in that variable, controlling for the other variables in the model. The parameter α represents the mean of y when each explanatory variable equals 0.

The parameters $\beta_1, \beta_2, \ldots$ are called ***partial regression coefficients***. The adjective *partial* distinguishes these parameters from the regression coefficient β in the *bivariate* model $E(y) = \alpha + \beta x$, which *ignores* rather than *controls* effects of other explanatory variables.

A partial slope in a multiple regression model usually differs from the slope in the bivariate model for that explanatory variable, but it need not. With two explanatory variables, the partial slopes and bivariate slopes are equal if the correlation between x_1 and x_2 equals 0. When x_1 and x_2 are independent causes of y, the effect of x_1 on y does not change when we control for x_2.

LIMITATIONS OF THIS MULTIPLE REGRESSION MODEL

In interpreting partial regression coefficients in observational studies, we need to be cautious not to regard the estimated effects as implying causal relations. For example, for a sample of college students, suppose y = math achievement test score (scale of

0 to 100) and the explanatory variables are x_1 = number of years of math education, x_2 = mother's number of years of math education, and x_3 = GPA, and we fit the multiple regression model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

Suppose that the estimate of β_1 is 5. In interpreting the effect of x_1, we might say, "A one-year increase in math education corresponds to an increase in the predicted math achievement test score of 5, controlling for the mother's math education and GPA." However, this does not imply that if a student attains another year of math education, her or his math achievement test score is predicted to change by 5. To validly make such a conclusion, we'd need to conduct an experiment that adds a year of math education for each student and then observes the results. Otherwise, a higher mean test score at a higher math education level could at least partly reflect the correlation of several other variables with both math test score and math education level, such as the student's IQ, father's number of years of math education, and number of years of science courses.

What the above interpretation actually means is this: "The difference between the estimated mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year of math education equals 5, when we control (keep constant) the mother's math education and GPA." However, we need to be cautious even with this interpretation. It is unnatural and even inconsistent with the data for some observational studies to envision increasing one explanatory variable while keeping all the others fixed. For example, x_1 and x_2 are likely to be positively correlated, so increases in x_1 naturally tend to occur with increases in x_2. In some data sets, one might not even observe a one-unit range in an explanatory variable when the other explanatory variables are all held constant. As an extreme example, suppose y = height, x_1 = length of left leg, and x_2 = length of right leg. The correlation between x_1 and x_2 is extremely close to 1. It does not make much sense to imagine how y changes as x_1 changes while x_2 is controlled.

Because of this limitation, some methodologists prefer to use more cautious wording than "controlling." In interpreting an estimate of 5 for β_1, they would say, "The difference between the estimated mean math achievement test score of a subpopulation of students having a certain number of years of math education and a subpopulation having one fewer year equals 5, when both subpopulations have the same estimated value for $\beta_2 x_{i2} + \beta_3 x_{i3}$." More concisely, "The effect of the number of years of math education on the estimated mean math achievement test score equals 5, *adjusting* for student's age and mother's math education." In the rest of the text, we will use the simpler "controlling" wording, but we should keep in mind its limitations.

Finally, this multiple regression model also has a structural limitation. It assumes that the slope of the partial relationship between y and each explanatory variable is identical for *all* combinations of values of the other explanatory variables. This means that the model is appropriate when there is a lack of *statistical interaction*, in the sense explained in Section 10.3 (page 294). If the true partial slope between y and x_1 is very different at $x_2 = 50$ than at $x_2 = 40$, for example, we need a more complex model. Section 11.4 will show this model and Section 11.5 will show how to analyze whether it fits significantly better.

PREDICTION EQUATION AND RESIDUALS

Corresponding to the multiple regression equation, software finds a prediction equation by estimating the model parameters using sample data. In general, we let p denote the number of explanatory variables.

Notation for Prediction Equation

The prediction equation that estimates the multiple regression equation $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ is denoted by $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$.

We use statistical software to find the prediction equation. The calculation formulas are complex and are not shown in this text.

To get the predicted value of y for a subject, we substitute the x-values for that subject into the prediction equation. Like the bivariate model, the multiple regression model has ***residuals*** that measure prediction errors. For a subject with predicted response $\hat{y}$ and observed response y, the residual is $y - \hat{y}$. The next section shows an example.

The ***sum of squared errors*** (SSE),

$$\text{SSE} = \sum (y - \hat{y})^2,$$

summarizes the closeness of fit of the prediction equation to the response data. Most software calls SSE the ***residual sum of squares***. The formula for SSE is the same as in Chapter 9. The only difference is that the predicted value $\hat{y}$ results from using *several* explanatory variables instead of just a single one.

The parameter estimates in the prediction equation satisfy the ***least squares*** criterion: The prediction equation has the *smallest* SSE value of all possible equations of the form $\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$.

Example 11.2

Multiple Regression for Mental Health Study A study in Alachua County, Florida, investigated the relationship between certain mental health indices and several explanatory variables. Primary interest focused on an index of mental impairment, which incorporates various dimensions of psychiatric symptoms, including aspects of anxiety and depression. This measure, which is the response variable y, ranged from 17 to 41 in the sample. Higher scores indicate greater psychiatric impairment.

The two explanatory variables used here are x_1 = life events score and x_2 = socioeconomic status (SES). The life events score is a composite measure of both the number and severity of major life events the subject experienced within the past three years. These events range from severe personal disruptions, such as a death in the family, a jail sentence, or an extramarital affair, to less severe events, such as getting a new job, the birth of a child, moving within the same city, or having a child marry. This measure ranged from 3 to 97 in the sample. A high score represents a greater number and/or greater severity of these life events. The SES score is a composite index based on occupation, income, and education. Measured on a standard scale, it ranged from 0 to 100. The higher the score, the higher the status.

Table 11.1 shows data[1] on the three variables for a random sample of 40 adults in the county. Table 11.2 summarizes the sample means and standard deviations of the three variables. ■

SCATTERPLOT MATRIX FOR BIVARIATE RELATIONSHIPS

Plots of the data provide an informal check of whether the relationships are linear. Software can construct scatterplots on a single diagram for each pair of the variables.

[1] These data are based on a much larger survey. Thanks to Dr. Charles Holzer for permission to use the study as the basis of this example.

TABLE 11.1: Mental Data File from the Text Website with y = Mental Impairment, x_1 = Life Events, and x_2 = Socioeconomic Status

y	x_1	x_2	y	x_1	x_2	y	x_1	x_2
17	46	84	26	50	40	30	44	53
19	39	97	26	48	52	31	35	38
20	27	24	26	45	61	31	95	29
20	3	85	27	21	45	31	63	53
20	10	15	27	55	88	31	42	7
21	44	55	27	45	56	32	38	32
21	37	78	27	60	70	33	45	55
22	35	91	28	97	89	34	70	58
22	78	60	28	37	50	34	57	16
23	32	74	28	30	90	34	40	29
24	33	67	28	13	56	41	49	3
24	18	39	28	40	56	41	89	75
25	81	87	29	5	40			
26	22	95	30	59	72			

TABLE 11.2: Sample Means and Standard Deviations of Mental Impairment, Life Events, and Socioeconomic Status (SES)

Variable	Mean	Standard Deviation
Mental impairment	27.30	5.46
Life events	44.42	22.62
SES	56.60	25.28

Figure 11.5 shows the plots for the variables from Table 11.1. This type of plot is called a ***scatterplot matrix***. Like a correlation matrix, it shows each pair of variables twice. In one plot, a variable is on the y-axis and in the other it is on the x-axis. Mental impairment (the response variable) is on the y-axis for the plots in the first row of Figure 11.5, so these are the plots of interest to us. The plots show no evidence of nonlinearity, and models with linear effects seem appropriate. The plots suggest that life events have a mild positive effect and SES has a mild negative effect on mental impairment.

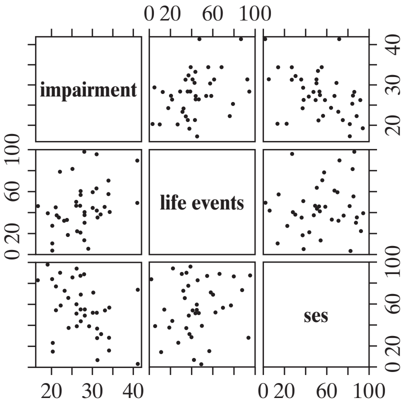


FIGURE 11.5: Scatterplot Matrix: Scatterplots for Pairs of Variables from Table 11.1

PARTIAL PLOTS FOR PARTIAL RELATIONSHIPS

The multiple regression model states that each explanatory variable has a linear effect with common slope, controlling for the other predictors. Although the regression formula is relatively simple, this itself is quite a strong assumption. To check it, we can compare the fit of the model to the fit of a more complex model, as we'll explain in Section 11.4. We can also plot y versus each predictor, for subsets of points that are nearly constant on the other predictors. With a single control variable, for example, we could sort the observations into four groups using the quartiles as boundaries, and then either construct four separate scatterplots or mark the observations on a single scatterplot according to their group.

With several control variables, however, keeping them all nearly constant can reduce the sample to relatively few observations and is impractical. A more informative single picture is provided by the ***partial regression plot*** (also called *added-variable plot*). It displays the relationship between the response variable and an explanatory variable after removing the effects of the other predictors in the multiple regression model. It does this by plotting the residuals from models using these two variables as responses and the other explanatory variables as predictors.

Here is how software constructs the partial regression plot for the effect of x_1 when the multiple regression model also has explanatory variables x_2 and x_3. It finds the residuals from the models (i) using x_2 and x_3 to predict y and (ii) using x_2 and x_3 to predict x_1. Then it plots the residuals from the first analysis (on the y-axis) against the residuals from the second analysis. For these residuals, the effects of x_2 and x_3 are removed. The least squares slope for the points in this plot is necessarily the same as the estimated partial slope b_1 for the multiple regression model.

Figure 11.6 shows a partial regression plot for y = mental impairment and x_1 = life events, controlling for x_2 = SES. It plots the residuals on the y-axis from $\hat{y} = 32.2 - 0.086x_2$ against the residuals on the x-axis from $\hat{x}_1 = 38.2 + 0.110x_2$. Both axes have negative and positive values, because they refer to residuals. Recall that residuals (prediction errors) can be positive or negative, and have a mean of 0. Figure 11.6 suggests that the partial effect of life events is approximately linear and is positive.

FIGURE 11.6: Partial Regression Plot for Mental Impairment and Life Events, Controlling for SES. This plots the residuals from regressing mental impairment on SES against the residuals from regressing life events on SES.

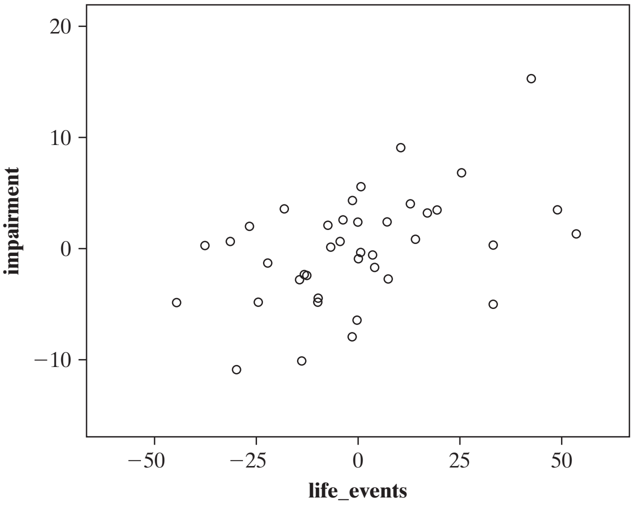


Figure 11.7 shows the partial regression plot for SES. It shows that its partial effect is also approximately linear but is negative.

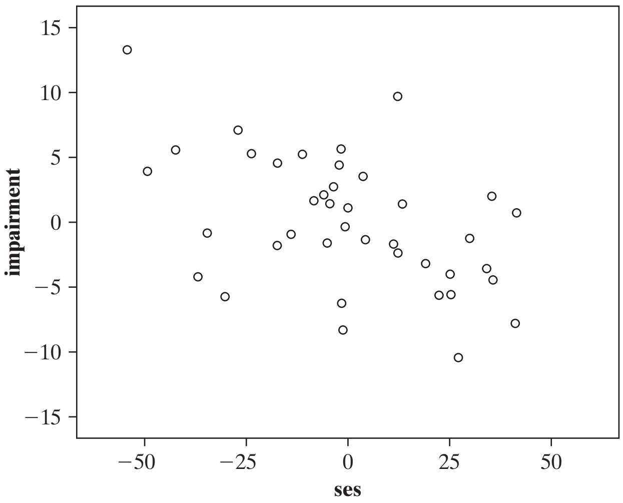


FIGURE 11.7: Partial Regression Plot for Mental Impairment and SES, Controlling for Life Events. This plots the residuals from regressing mental impairment on life events against the residuals from regressing SES on life events.

SOFTWARE OUTPUT FOR MENTAL IMPAIRMENT EXAMPLE

Tables 11.3 and 11.4 are Stata outputs of the coefficients table for the bivariate relationships between mental impairment and the separate explanatory variables. The estimated regression coefficients fall in the column labeled *Coef.* The prediction equations are

$$\hat{y} = 23.31 + 0.090x_1 \text{ and } \hat{y} = 32.17 - 0.086x_2.$$

In the sample, mental impairment is positively related to life events, since the coefficient of x_1 (0.090) is positive. The greater the number and severity of life events in the previous three years, the higher the mental impairment (i.e., the poorer the mental health) tends to be. Mental impairment is negatively related to socioeconomic status. The correlations between mental impairment and the explanatory variables are modest, 0.372 for life events and −0.399 for SES (the appropriate square roots of the r^2-values reported).

TABLE 11.3: Bivariate Regression Analysis for y = Mental Impairment and x_1 = Life Events from `Mental` Data File

```
                                           R-squared = 0.1385
impair |     Coef.    Std. Err.      t     P>|t|    [95% Conf. Interval]
  life |  .0898257    .0363349     2.47    0.018      .0163     .1634
 _cons |  23.30949    1.806751    12.90    0.000      19.65     26.97
```

TABLE 11.4: Bivariate Regression Analysis for y = Mental Impairment and x_2 = Socioeconomic Status (SES) from `Mental` Data File

```
                                           R-squared = 0.1589
impair |     Coef.    Std. Err.      t     P>|t|    [95% Conf. Interval]
  ses  |  -.086078    .0321317    -2.68    0.011     -.1511    -.0210
 _cons |  32.17201    1.987649    16.19    0.000     28.148    36.196
```

Table 11.5 shows output for the multiple regression model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$. The prediction equation is

$$\hat{y} = a + b_1x_1 + b_2x_2 = 28.230 + 0.103x_1 - 0.097x_2.$$

TABLE 11.5: Fit of Multiple Regression Model for y = Mental Impairment, x_1 = Life Events, and x_2 = Socioeconomic Status from `Mental` Data File

```
                                            R-squared = 0.3392
impair |      Coef.    Std. Err.       t     P>|t|   [95% Conf. Interval]
  life |   .1032595    .0324995     3.18    0.003      .0374     .1691
   ses |  -.0974755    .0290848    -3.35    0.002     -.1564    -.0385
 _cons |   28.22981    2.174222    12.98    0.000      23.82     32.64
```

Controlling for SES, the sample relationship between mental impairment and life events is positive, since the coefficient of life events ($b_1 = 0.103$) is positive. The estimated mean of mental impairment increases by about 0.1 for every one-unit increase in the life events score, controlling for SES. Since $b_2 = -0.097$, a negative association exists between mental impairment and SES, controlling for life events. For example, over the 100-unit range of potential SES values (from a minimum of 0 to a maximum of 100), the estimated mean mental impairment changes by $100(-0.097) = -9.7$. Since mental impairment ranges only from 17 to 41 with a standard deviation of 5.5, a decrease of 9.7 points in the mean is noteworthy.

From Table 11.1, the first subject in the sample had $y = 17$, $x_1 = 46$, and $x_2 = 84$. This subject's predicted mental impairment is

$$\hat{y} = 28.230 + 0.103(46) - 0.097(84) = 24.8.$$

The prediction error (residual) is $y - \hat{y} = 17 - 24.8 = -7.8$.

Table 11.6 summarizes some results of the regression analyses. It shows standard errors in parentheses below the parameter estimates. The partial slopes for the multiple regression model are similar to the slopes for the bivariate models. In each case, the introduction of the second explanatory variable does little to alter the effect of the other one. This suggests that these explanatory variables may have nearly independent sample effects on y. In fact, the sample correlation between x_1 and x_2 is only 0.123. The next section shows how to interpret the R^2-value listed for the multiple regression model.

TABLE 11.6: Summary of Regression Models for Mental Impairment

	Explanatory Variables in Regression Model		
Effect	Multiple	Life Events	SES
Intercept	28.230	23.309	32.172
Life events	0.103 (0.032)	0.090 (0.036)	—
SES	−0.097 (0.029)	—	−0.086 (0.032)
R^2	0.339	0.138	0.159

11.2 Multiple Correlation and R^2

The correlation r and its square describe strength of linear association for bivariate relationships. This section presents analogous measures for the multiple regression model. They describe the strength of association between y and the set of explanatory variables acting together as predictors in the model.

THE MULTIPLE CORRELATION R

The explanatory variables collectively are strongly associated with y if the observed y-values correlate highly with the $\hat{y}$-values from the prediction equation. The correlation between the observed and predicted values summarizes this association.

Multiple Correlation

> The sample ***multiple correlation*** for a regression model, denoted by R, is the correlation between the observed y-values and the predicted $\hat{y}$-values.

For each subject, the prediction equation provides a predicted value $\hat{y}$. So, each subject has a y-value and a $\hat{y}$-value. For the first three subjects in Table 11.1, the observed and predicted y-values are

y	$\hat{y}$
17	24.8
19	22.8
20	28.7

The sample correlation computed between all 40 of the y- and $\hat{y}$-values is R, the multiple correlation. The larger the value of R, the better the predictions of y by the set of explanatory variables.

The predicted values cannot correlate negatively with the observed values. The predictions must be at least as good as the sample mean $\bar{y}$, which is the prediction when all partial slopes = 0, and $\bar{y}$ has zero correlation with y. So, R always falls between 0 and 1. In this respect, the correlation between y and $\hat{y}$ differs from the correlation between y and an explanatory variable x, which falls between -1 and $+1$.

R^2: THE COEFFICIENT OF MULTIPLE DETERMINATION

Another measure uses the *proportional reduction in error* concept, generalizing r^2 for bivariate models. This measure summarizes the relative improvement in predictions using the prediction equation instead of $\bar{y}$. It has the following elements:

Rule 1 (Predict y without using $x_1, \ldots, x_p$): The best predictor is then the sample mean, $\bar{y}$.

Rule 2 (Predict y using $x_1, \ldots, x_p$): The best predictor is the prediction equation $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_px_p$.

Prediction Errors: The prediction error for a subject is the difference between the observed and predicted values of y. With rule 1, the error is $y - \bar{y}$. With rule 2, it is the residual $y - \hat{y}$. In either case, we summarize the error by the sum of the squared prediction errors. For rule 1, this is TSS $= \sum(y - \bar{y})^2$, the *total sum of squares*. For rule 2, it is SSE $= \sum(y-\hat{y})^2$, the sum of squared errors using the prediction equation, which is the *residual sum of squares*.

Definition of Measure: The proportional reduction in error from using the prediction equation $\hat{y} = a + b_1x_1 + \cdots + b_px_p$ instead of $\bar{y}$ to predict y is ***R-squared***, also called the ***coefficient of multiple determination***.

***R*-Squared: The Coefficient of Multiple Determination**

$$R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

R^2 measures the proportion of the total variation in y that is explained by the predictive power of all the explanatory variables, through the multiple regression model. The symbol reflects that R^2 *is the square of the multiple correlation* R. The uppercase notation R^2 distinguishes this measure from r^2 for the bivariate model. Their formulas are identical, and r^2 is the special case of R^2 applied to a regression model with one explanatory variable. For the multiple regression model to be useful for prediction, it should provide improved predictions relative not only to $\bar{y}$ but also to the separate bivariate models for y and each explanatory variable.

Example 11.3

Multiple Correlation and R^2 for Mental Impairment For the data on y = mental impairment, x_1 = life events, and x_2 = socioeconomic status in Table 11.1, Table 11.5 showed some output. Software (SPSS) also reports ANOVA tables with sums of squares and shows R and R^2. See Table 11.7.

TABLE 11.7: ANOVA Table and Model Summary for Regression of Mental Impairment on Life Events and Socioeconomic Status from `Mental` Data File

```
 ANOVA
             Sum of Squares   df    Mean Square     F        Sig.
 Regression    394.238         2      197.119      9.495     .000
 Residual      768.162        37       20.761
 Total        1162.400        39

               Model Summary
  R    R Square   Adjusted R Square   Std. Error of the Estimate
.582     .339           .303                     4.556

Predictors: (Constant), SES, LIFE
Dependent Variable: IMPAIR
```

From the *Sum of Squares* column, the total sum of squares is TSS $= \sum(y-\bar{y})^2 =$ 1162.4, and the residual sum of squares from using the prediction equation to predict y is SSE $= \sum(y-\hat{y})^2 = 768.2$. Thus,

$$R^2 = \frac{\text{TSS} - \text{SSE}}{\text{TSS}} = \frac{1162.4 - 768.2}{1162.4} = 0.339.$$

Using life events and SES together to predict mental impairment provides a 33.9% reduction in the prediction error relative to using only $\bar{y}$. The multiple regression model provides a substantially larger reduction in error than either bivariate model (Table 11.6 reported r^2-values of 0.138 and 0.159 for them). It is more useful than those models for predictive purposes.

The multiple correlation between mental impairment and the two explanatory variables is $R = +\sqrt{0.339} = 0.582$. This equals the correlation between the observed y- and predicted $\hat{y}$-values for the model. ■

PROPERTIES OF R AND R^2

The properties of R^2 are similar to those of r^2 for bivariate models.

- R^2 falls between 0 and 1.
- The larger the value of R^2, the better the set of explanatory variables $(x_1, \ldots, x_p)$ collectively predicts y.
- $R^2 = 1$ only when all the residuals are 0, that is, when all $y = \hat{y}$, so that predictions are perfect and SSE = 0.

- $R^2 = 0$ when the predictions do not vary as any of the x-values vary. In that case, $b_1 = b_2 = \cdots = b_p = 0$, and $\hat{y}$ is identical to $\bar{y}$, since the explanatory variables do not add any predictive power. The correlation is then 0 between y and each explanatory variable.

- R^2 cannot decrease when we add an explanatory variable to the model. It is impossible to explain *less* variation in y by adding explanatory variables to a regression model.

- R^2 for the multiple regression model is at least as large as the r^2-values for the separate bivariate models. That is, R^2 for the multiple regression model is at least as large as $r^2_{yx_1}$ for y as a linear function of x_1, $r^2_{yx_2}$ for y as a linear function of x_2, and so forth.

- R^2 tends to overestimate the population value, because the sample data fall closer to the sample prediction equation than to the true population regression equation. Most software also reports a less biased estimate, called ***adjusted*** R^2. Exercise 11.61 shows its formula. For the mental impairment example, Table 11.7 reports its value of 0.303, compared to ordinary $R^2 = 0.339$.

Properties of the multiple correlation R follow directly from the ones for R^2, since R is the positive square root of R^2. For instance, R for the model $E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ is at least as large as R for the model $E(y) = \alpha + \beta_1x_1 + \beta_2x_2$.

The numerator of R^2, TSS − SSE, summarizes the variation in y explained by the multiple regression model. This difference, which equals $\sum(\hat{y} - \bar{y})^2$, is called the ***regression sum of squares***. The ANOVA table in Table 11.7 lists the regression sum of squares as 394.2. (Some software, such as Stata and SAS, labels this the *Model* sum of squares.) The total sum of squares TSS of the y-values about $\bar{y}$ partitions into the variation explained by the regression model (regression sum of squares) plus the variation not explained by the model (the residual sum of squares, SSE).

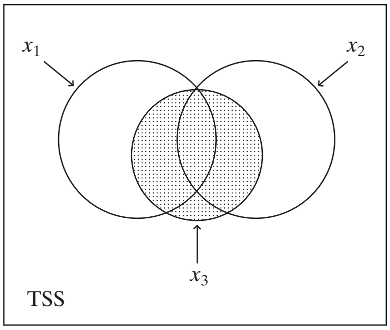
MULTICOLLINEARITY WITH MANY EXPLANATORY VARIABLES

When a study has many explanatory variables but the correlations among them are strong, once you have included a few of them in the model, R^2 usually doesn't increase much more when you add additional ones. For example, for the `Houses` data file at the text website (introduced in Example 9.10 on page 268), r^2 is 0.71 with the house's tax assessment as a predictor of selling price. Then, R^2 increases to 0.77 when we add house size as a second predictor. But then it increases only to 0.79 when we add number of bathrooms, number of bedrooms, and whether the house is new as additional predictors.

When R^2 does not increase much, this does not mean that the additional variables are uncorrelated with y. It means merely that they don't add much new power for predicting y, given the values of the explanatory variables already in the model. These other variables may have small associations with y, given the variables already in the model. This often happens in social science research when the explanatory variables are highly correlated, no one having much unique explanatory power. Section 14.3 discusses this condition, called ***multicollinearity***.

Figure 11.8, which portrays the portion of the total variability in y explained by each of three explanatory variables, shows a common occurrence. The size of the set for an explanatory variable in this figure represents the size of its r^2-value in predicting y. The amount a set for an explanatory variable overlaps with the set for another explanatory variable represents its association with that predictor. The part of the set for an explanatory variable that does not overlap with other sets represents the part of the variability in y explained uniquely by that explanatory variable. In Figure 11.8, all three explanatory variables have moderate associations with y, and together they explain considerable variation. Once x_1 and x_2 are in the model, however, x_3 explains little additional variation in y, because of its strong correlations with x_1 and x_2. Because of this overlap, R^2 increases only slightly when x_3 is added to a model already containing x_1 and x_2.

FIGURE 11.8: R^2 Does Not Increase Much when x_3 Is Added to the Model Already Containing x_1 and x_2



For predictive purposes, we gain little by adding explanatory variables to a model that are strongly correlated with ones already in the model, since R^2 will not increase much. Ideally, we should use explanatory variables having weak correlations with each other but strong correlations with y. In practice, this is not always possible, especially when we include certain variables in the model for theoretical reasons.

The sample size you need to do a multiple regression well gets larger when you want to use more explanatory variables. Technical difficulties caused by multicollinearity are less severe for larger sample sizes. Ideally, the sample size should be at least about 10 times the number of explanatory variables (e.g., at least about 40 for 4 explanatory variables).

11.3 Inferences for Multiple Regression Coefficients

To make inferences about the parameters in the multiple regression function

$$E(y) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p,$$

we formulate the entire *multiple regression model*. This consists of this equation together with a set of assumptions:

- The population distribution of y is normal, for each combination of values of $x_1, \ldots, x_p$.
- The standard deviation, σ, of the conditional distribution of responses on y is the same at each combination of values of $x_1, \ldots, x_p$.
- The sample is randomly selected.

Under these assumptions, the true sampling distributions exactly equal those quoted in this section. In practice, the assumptions are never satisfied perfectly. Two-sided inferences are robust to the normality and common σ assumptions. More important are the assumptions of randomization and that the regression function describes well how the mean of y depends on the explanatory variables. We'll see ways to check the latter assumption in Sections 11.4 and 14.2.

Multiple regression analyses use two types of significance tests. The first is a global test of independence. It checks whether *any* of the explanatory variables are statistically related to y. The second studies the partial regression coefficients individually, to assess which explanatory variables have significant partial effects on y.

TESTING THE COLLECTIVE INFLUENCE OF THE EXPLANATORY VARIABLES

Do the explanatory variables collectively have a statistically significant effect on the response variable? We check this by testing

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

that the mean of y does not depend on the values of $x_1, \ldots, x_p$. Under the inference assumptions, this states that y is statistically independent of all p explanatory variables.

The alternative hypothesis is

$$H_a\colon \text{At least one } \beta_i \neq 0.$$

This states that *at least one* explanatory variable is associated with y, controlling for the others. The test judges whether using $x_1, \ldots, x_p$ together to predict y, with the prediction equation $\hat{y} = a + b_1x_1 + \cdots + b_px_p$, is significantly better than using $\bar{y}$.

These hypotheses about $\{\beta_i\}$ are equivalent to

H_0: Population multiple correlation $= 0$ and H_a: Population multiple correlation > 0.

The equivalence occurs because the multiple correlation equals 0 only in those situations in which all the partial regression coefficients equal 0. Also, H_0 is equivalent to H_0: population R-squared $= 0$.

For these hypotheses about the p predictors, the test statistic is

$$F = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}.$$

The sampling distribution of this statistic is called the ***F* distribution**.

THE *F* DISTRIBUTION

The symbol for the F test statistic and its distribution honors the most eminent statistician in history, R. A. Fisher, who discovered the F distribution in 1922. Like the chi-squared distribution, the F distribution can take only nonnegative values and it is somewhat skewed to the right. Figure 11.9 illustrates this.

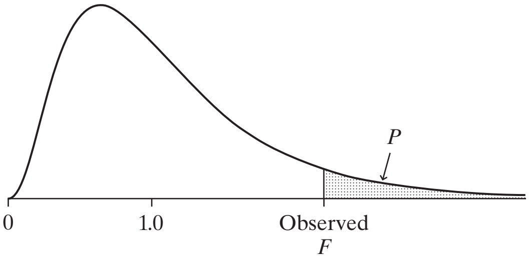


FIGURE 11.9: The F Distribution and the P-Value for F Tests. Larger F-values give stronger evidence against H_0.

The shape of the F distribution is determined by two degrees of freedom terms, denoted by df_1 and df_2:

$$df_1 = p, \text{ the number of explanatory variables in the model.}$$

$$df_2 = n - (p + 1) = n - \text{ number of parameters in regression equation.}$$

The first of these, $df_1 = p$, is the divisor of the numerator term (R^2) in the F test statistic. The second, $df_2 = n - (p + 1)$, is the divisor of the denominator term $(1 - R^2)$. The number of parameters in the multiple regression model is $p + 1$, representing the p beta terms and the y-intercept (α) term.

The mean of the F distribution is approximately[2] equal to 1. The larger the R^2-value, the larger the ratio $R^2/(1 - R^2)$, and the larger the F test statistic becomes. Thus, larger values of the F test statistic provide stronger evidence against H_0. Under the presumption that H_0 is true, the P-value is the probability the F test statistic is larger than the observed F-value. This is the right-tail probability under the F distribution beyond the observed F-value, as Figure 11.9 shows. Software for regression and Internet applets[3] report the P-value.

Example 11.4

***F* Test for Mental Impairment Data** For Table 11.1 (page 313), we used multiple regression for $n = 40$ observations on y = mental impairment, with $p = 2$ explanatory variables, life events and SES. The null hypothesis that mental impairment is statistically independent of life events and SES is H_0: $\beta_1 = \beta_2 = 0$.

In Example 11.3 (page 318), we found that this model has $R^2 = 0.339$. The F test statistic value is

$$F = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]} = \frac{0.339/2}{0.661/[40 - (2 + 1)]} = 9.5.$$

The two degrees of freedom terms for the F distribution are $df_1 = p = 2$ and $df_2 = n - (p + 1) = 40 - 3 = 37$, the two divisors in this statistic.

Part of the SPSS software output in Table 11.7 showed the ANOVA table

```
           Sum of Squares   df    Mean Square     F       Sig.
Regression      394.238      2      197.119     9.495     .000
Residual        768.162     37       20.761
```

which contains the F statistic. The P-value, which rounded to three decimal places is $P = 0.000$, appears under the heading *Sig* in the table. (R reports it as *p-value*, Stata reports it as *Prob > F*, and SAS reports it as *Pr > F*.) This extremely small P-value provides strong evidence against H_0. We infer that at least one of the explanatory variables is associated with mental impairment. Equivalently, we can conclude that the population multiple correlation and R-squared are positive. ■

Normally, unless n is small and the associations are weak, this F test has a small P-value. If we choose variables wisely for a study, at least one of them should have *some* explanatory power.

INFERENCES FOR INDIVIDUAL REGRESSION COEFFICIENTS

When the P-value is small for the F test, this does not imply that *every* explanatory variable has an effect on y (controlling for the other explanatory variables in the model), but merely that *at least one* of them has an effect. More narrowly focused analyses judge *which* partial effects are nonzero and estimate the sizes of those effects. These inferences make the same assumptions as the F test.

For a particular explanatory variable x_i in the model, the test for its partial effect on y has H_0: $\beta_i = 0$. If $\beta_i = 0$, the mean of y is identical for all values of x_i, controlling

[2] The mean equals $df_2/(df_2 - 2)$, which is usually close to 1 unless n is quite small.

[3] For example, the *F distribution* applet at www.artofstat.com/webapps.html.

for the other explanatory variables in the model. The alternative can be two-sided, H_a: $\beta_i \neq 0$, or one-sided, H_a: $\beta_i > 0$ or H_a: $\beta_i < 0$, to predict the direction of the partial effect.

The test statistic for H_0: $\beta_i = 0$, using sample estimate b_i of β_i, is

$$t = \frac{b_i}{se},$$

where se is the standard error of b_i. As usual, the t test statistic takes the best estimate (b_i) of the parameter (β_i), subtracts the H_0 value of the parameter (0), and divides by the standard error. The formula for se is complex, but software provides its value. If H_0 is true and the model assumptions hold, the t statistic has the t distribution with $df = n - (p + 1)$, which is the same as df_2 in the F test.

It is more informative to estimate the size of a partial effect than to test whether it is zero. Recall that β_i represents the change in the mean of y for a one-unit increase in x_i, controlling for the other variables. A confidence interval for β_i is

$$b_i \pm t(se).$$

The t-score comes from the t table, with $df = n - (p + 1)$. For example, a 95% confidence interval for the partial effect of x_1 is $b_1 \pm t_{.025}(se)$.

Example 11.5

Inferences for Individual Predictors of Mental Impairment For the multiple regression model for y = mental impairment, x_1 = life events, and x_2 = SES,

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2,$$

let's analyze the effect of life events. The hypothesis that mental impairment is statistically independent of life events, controlling for SES, is H_0: $\beta_1 = 0$. If H_0 is true, the multiple regression equation reduces to $E(y) = \alpha + \beta_2 x_2$. If H_0 is false, then $\beta_1 \neq 0$ and the full model provides a better fit than the bivariate model.

Table 11.5 contained the results,

```
                  B     Std. Error        t      Sig.
(Constant)      28.230    2.174        12.984    .000
LIFE              .103     .032         3.177    .003
SES              -.097     .029        -3.351    .002
```

The point estimate of β_1 is $b_1 = 0.103$, which has standard error $se = 0.032$. The test statistic is

$$t = \frac{b_1}{se} = \frac{0.103}{0.032} = 3.177.$$

This appears under the heading t in the table in the row for the variable LIFE. The statistic has $df = n - (p + 1) = 40 - 3 = 37$. The P-value is 0.003, the probability that the t statistic exceeds 3.177 in absolute value. The evidence is strong that mental impairment is associated with life events, controlling for SES.

A 95% confidence interval for β_1 uses $t_{0.025} = 2.026$, the t-value for $df = 37$ having a probability of 0.05/2 = 0.025 in each tail. This interval is

$$b_1 \pm t_{0.025}(se) = 0.103 \pm 2.026(0.032), \quad \text{which is} \quad (0.04, 0.17).$$

Controlling for SES, we are 95% confident that the change in mean mental impairment per one-unit increase in life events falls between 0.04 and 0.17. An increase of 100 units in life events corresponds to anywhere from a 100(0.04) = 4-unit to a 100(0.17) = 17-unit increase in mean mental impairment. The interval is relatively wide because of the relatively small sample size. The interval does not contain 0. This is in agreement with rejecting H_0: $\beta_1 = 0$ in favor of H_a: $\beta_1 \neq 0$ at the $\alpha = 0.05$

level. Since the confidence interval contains only positive numbers, the association between mental impairment and life events is positive, controlling for SES. ■

How is the t test for a partial regression coefficient different from the t test of H_0: $\beta = 0$ for the bivariate model, $E(y) = \alpha + \beta x$, presented in Section 9.5? That t test evaluates whether y and x are associated, *ignoring* other variables, because it applies to the bivariate model. By contrast, the test just presented evaluates whether variables are associated, *controlling* for other variables.

A note of caution: Suppose multicollinearity occurs, that is, much overlap among the explanatory variables in the sense that any one is well predicted by the others. Then, possibly none of the individual partial effects has a small P-value, even if R^2 is large and a large F statistic occurs in the global test for the βs. Any particular variable may uniquely explain little variation in y, even though together the variables explain much variation.

VARIABILITY AND MEAN SQUARES IN THE ANOVA TABLE*

The precision of the least squares estimates relates to the size of the conditional standard deviation σ that measures variability of y at fixed values of the predictors. The smaller the variability of y-values about the regression equation, the smaller the standard errors become. The estimate of σ is

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n - (p + 1)}} = \sqrt{\frac{\text{SSE}}{df}}.$$

The degrees of freedom value is also df for t inferences for regression coefficients, and it is df_2 for the F test about the collective effect of the explanatory variables. (When a model has only $p = 1$ predictor, df simplifies to $n - 2$, the term in the s formula in Section 9.3 on page 256.)

From the ANOVA table in Table 11.7 (page 318) that contains the sums of squares for the multiple regression model with the mental impairment data, SSE = 768.2. Since $n = 40$ for $p = 2$ predictors, we have $df = n - (p + 1) = 40 - 3 = 37$ and

$$s = \sqrt{\frac{\text{SSE}}{df}} = \sqrt{\frac{768.2}{37}} = \sqrt{20.76} = 4.56.$$

If the conditional distributions are approximately bell shaped, nearly all mental impairment scores fall within about 14 units (3 standard deviations) of the mean specified by the regression function.

SPSS reports the conditional standard deviation under the heading *Std. Error of the Estimate* in the Model Summary table that also shows R and R^2 (see Table 11.7). This is a poor choice of label by SPSS, because s refers to the variability in y-values, not the variability of a sampling distribution of an estimator.

The square of s, which estimates the conditional variance, is often called the ***error mean square***, often abbreviated by MSE, or the ***residual mean square***. Software shows it in the ANOVA table in the *Mean Square* column, in the row labeled *Residual* (or *Error* in some software). For example, MSE = 20.76 in Table 11.7. Some software (such as Stata and SAS) better labels the conditional standard deviation estimate s as *Root MSE*, because it is the square root of the error mean square. R reports it as *Residual standard error*.

THE F STATISTIC IS A RATIO OF MEAN SQUARES*

An alternative formula for the F test statistic for testing H_0: $\beta_1 = \cdots = \beta_p = 0$ uses the two mean squares in the ANOVA table. Specifically, for our example,

$$F = \frac{\text{Regression mean square}}{\text{Residual mean square (MSE)}} = \frac{197.1}{20.8} = 9.5.$$

This gives the same value as the F test statistic formula (page 321) based on R^2.

The regression mean square equals the regression sum of squares divided by its degrees of freedom. The df equals p, the number of explanatory variables in the model, which is df_1 for the F test. In the ANOVA table in Table 11.7, the regression mean square equals

$$\frac{\text{Regression SS}}{df_1} = \frac{394.2}{2} = 197.1.$$

RELATIONSHIP BETWEEN F AND t STATISTICS*

We've used the F distribution to test that all partial regression coefficients equal 0. Some regression software also lists F test statistics instead of t test statistics for the tests about the individual regression coefficients. The two statistics are related and have the same P-values. The square of the t statistic for testing that a partial regression coefficient equals 0 is an F test statistic having the F distribution with $df_1 = 1$ and $df_2 = n - (p + 1)$.

To illustrate, in Example 11.5 for H_0: $\beta_1 = 0$ and H_a: $\beta_1 \neq 0$, the test statistic $t = 3.18$ with $df = 37$. Alternatively, we could use $F = t^2 = 3.18^2 = 10.1$, which has the F distribution with $df_1 = 1$ and $df_2 = 37$. The P-value for this F-value is 0.002, the same as Table 11.5 reports for the two-sided t test.

In general, if a statistic has the t distribution with d degrees of freedom, then the square of that statistic has the F distribution with $df_1 = 1$ and $df_2 = d$. A disadvantage of the F approach is that it lacks information about the direction of the association. It cannot be used for one-sided alternative hypotheses.

11.4 Modeling Interaction Effects

The multiple regression equation

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

assumes that the slope β_i of the partial relationship between y and each x_i is identical for all values of the other explanatory variables. This implies a parallelism of lines relating the two variables, at various values of the other variables, as Figure 11.3 (page 309) illustrated.

This model is sometimes too simple to be adequate. Often, the relationship between two variables changes according to the value of a third variable. There is ***interaction***, a concept introduced in Section 10.3 (page 294).

Interaction

For quantitative variables, ***interaction*** exists between two explanatory variables in their effects on y when the effect of one variable changes as the level of the other variable changes.

For example, for y = annual income (thousands of dollars), x_1 = years of working experience, and x_2 = number of years of education, suppose $E(y) = 18 + 0.25x_1$

when $x_2 = 10$, $E(y) = 25 + 0.50x_1$ when $x_2 = 12$, and $E(y) = 39 + 1.00x_1$ when $x_2 = 16$. The slope for the partial effect of x_1 changes markedly as the value for x_2 changes. Interaction occurs between x_1 and x_2 in their effects on y.

CROSS-PRODUCT TERMS

The most common approach for modeling interaction effects introduces ***cross-product terms*** of the explanatory variables into the multiple regression model. With two explanatory variables, the model is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

This is a special case of the multiple regression model with three explanatory variables, in which x_3 is an artificial variable created as the cross product $x_3 = x_1 x_2$ of the two primary explanatory variables.

Let's see why this model permits interaction. Consider how y is related to x_1, controlling for x_2. We rewrite the equation in terms of x_1 as

$$E(y) = (\alpha + \beta_2 x_2) + (\beta_1 + \beta_3 x_2)x_1 = \alpha' + \beta' x_1,$$

where

$$\alpha' = \alpha + \beta_2 x_2 \qquad \text{and} \qquad \beta' = \beta_1 + \beta_3 x_2.$$

So, for fixed x_2, the mean of y changes linearly as a function of x_1. But the slope of the relationship, $\beta' = (\beta_1 + \beta_3 x_2)$, depends on the value of x_2. As x_2 changes, the slope for the effect of x_1 changes. In summary, the mean of y is a linear function of x_1, but the slope of the line changes as the value of x_2 changes.

For the model containing the cross-product term, β_1 is the effect of x_1 only when $x_2 = 0$. Unless $x_2 = 0$ is a particular value of interest for x_2, it is not particularly useful to form confidence intervals or perform significance tests about β_1 (or β_2) in this model.

Similarly, the mean of y is a linear function of x_2, but the slope varies according to the value of x_1. The coefficient β_2 of x_2 refers to the effect of x_2 only at $x_1 = 0$.

Example 11.6

Allowing Interaction in Modeling Mental Impairment For the data set on y = mental impairment, x_1 = life events, and x_2 = SES, we create a third explanatory variable x_3 that gives the cross product of x_1 and x_2 for the 40 individuals. For the first subject, for example, $x_1 = 46$ and $x_2 = 84$, so $x_3 = 46(84) = 3864$. Software makes it easy to create this variable without doing the calculations yourself. Table 11.8 shows some software output for the model that permits interaction. The prediction equation is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060x_2 - 0.00087x_1x_2.$$

Figure 11.10 portrays the relationship between predicted mental impairment and life events for a few distinct SES values. For an SES score of $x_2 = 0$, the relationship between $\hat{y}$ and x_1 is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060(0) - 0.00087x_1(0) = 26.0 + 0.156x_1.$$

When $x_2 = 50$, the prediction equation is

$$\hat{y} = 26.0 + 0.156x_1 - 0.060(50) - 0.00087(50)x_1 = 23.0 + 0.113x_1.$$

When $x_2 = 100$, the prediction equation is

$$y = 20.0 + 0.069x_1.$$

TABLE 11.8: Output for Model Allowing Interaction, for y = Mental Impairment, x_1 = Life Events, and x_2 = SES from `Mental` Data File

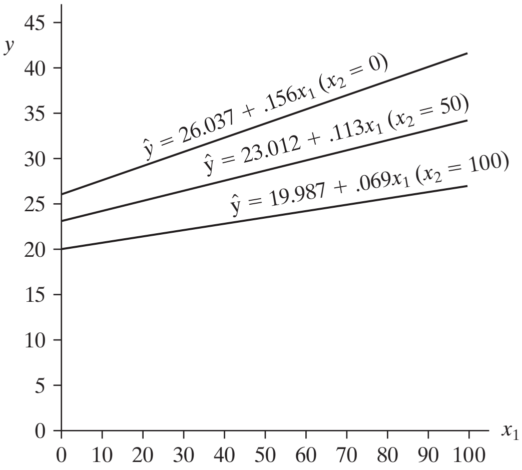
```
              Sum of            Mean
              Squares   DF     Square       F          Sig
Regression    403.631    3    134.544     6.383      0.0014
Residual      758.769   36     21.077
Total        1162.400   39

              R: .589       R Square:  .347

                 B         Std. Error       t              Sig
(Constant)    26.036649     3.948826      6.594          0.0001
LIFE           0.155865     0.085338      1.826          0.0761
SES           -0.060493     0.062675     -0.965          0.3409
LIFE*SES      -0.000866     0.001297     -0.668          0.5087
```

The higher the value of SES, the smaller the slope between predicted mental impairment and life events, and so the weaker is the effect of life events. Perhaps subjects who possess greater resources, in the form of higher SES, are better able to withstand the mental stress of potentially traumatic life events. ■

FIGURE 11.10: Portrayal of Interaction between x_1 and x_2 in Their Effects on y



TESTING SIGNIFICANCE OF AN INTERACTION TERM

For two explanatory variables, the model allowing interaction is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

The simpler model assuming no interaction is the special case $\beta_3 = 0$. The hypothesis of no interaction is H_0: $\beta_3 = 0$. As usual, the t test statistic divides the estimate of the parameter (β_3) by its standard error.

From Table 11.8, $t = -0.00087/0.0013 = -0.67$. The P-value for H_a: $\beta_3 \neq 0$ is $P = 0.51$. Little evidence exists of interaction. The variation in the slope of the relationship between mental impairment and life events for various SES levels could be due to sampling variability. The sample size here is small, however, which makes it difficult to estimate effects precisely.

When the evidence of interaction is weak, as it is here with a P-value of 0.51, it is best to drop the interaction term from the model before testing hypotheses about partial effects such as H_0: $\beta_1 = 0$ or H_0: $\beta_2 = 0$. On the other hand, *if the evidence of interaction is strong, it no longer makes sense to test these other hypotheses*. If there is interaction, then the effect of each variable exists and differs according to the level of the other variable.

CENTERING THE EXPLANATORY VARIABLES*

For the mental impairment data, x_1 and x_2 are highly significant in the model with only those predictors (Table 11.5) but lose their significance after entering the interaction term, even though the interaction is not significant (Table 11.8). But we've noted that the coefficients of x_1 and x_2 in an interaction model are not usually meaningful, because they refer to the effect of a predictor only when the other predictor equals 0.

An alternative way to parameterize the interaction model gives estimates and significance for the effect of x_1 and x_2 similar to those for the no-interaction model. The method *centers* the scores for each explanatory variable around 0, by subtracting the mean. Let $x_1^C = x_1 - \mu_{x_1}$ and $x_2^C = x_2 - \mu_{x_2}$, so that each new explanatory variable has a mean of 0. Then, we express the interaction model as

$$\begin{aligned} E(y) &= \alpha + \beta_1 x_1^C + \beta_2 x_2^C + \beta_3 x_1^C x_2^C \\ &= \alpha + \beta_1(x_1 - \mu_{x_1}) + \beta_2(x_2 - \mu_{x_2}) + \beta_3(x_1 - \mu_{x_1})(x_2 - \mu_{x_2}). \end{aligned}$$

Now, β_1 refers to the effect of x_1 at the mean of x_2, and β_2 refers to the effect of x_2 at the mean of x_1. Their estimates are usually similar to the estimated effects for the no-interaction model.

When we rerun the interaction model for the mental health data after centering the predictors about their sample means, that is, with

$$\text{LIFE\_CEN} = \text{LIFE} - 44.425 \text{ and } \text{SES\_CEN} = \text{SES} - 56.60,$$

we get software output shown in Table 11.9. The estimate for the interaction term is the same as for the model with uncentered predictors. Now, though, the estimates (and standard errors) for the effects of x_1 and x_2 alone are similar to the values for the no-interaction model. This happens because the coefficient for a variable represents its effect at the mean of the other variable, which is typically similar to the effect for the no-interaction model. Also, the statistical significance of x_1 and x_2 is similar as in the no-interaction model.

TABLE 11.9: Output for Model Allowing Interaction, Using Centered Explanatory Variables

```
                          B        Std. Error       t        Sig
(Constant)          27.359555     0.731366      37.409    0.0001
LIFE_CEN             0.106850     0.033185       3.220    0.0027
SES_CEN             -0.098965     0.029390      -3.367    0.0018
LIFE_CEN*SES_CEN    -0.000866     0.001297      -0.668    0.5087
```

In summary, centering the explanatory variables before using them in a model allowing interaction has two benefits. First, the estimates of the effects of x_1 and x_2 are more meaningful, being effects at the mean rather than at 0. Second, the estimates and their standard errors are similar as in the no-interaction model.

GENERALIZATIONS AND LIMITATIONS*

When the number of explanatory variables exceeds two, a model allowing interaction can have cross products for each pair of explanatory variables. For example, with three explanatory variables, an interaction model is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3.$$

This is a special case of multiple regression with six explanatory variables, identifying $x_4 = x_1 x_2$, $x_5 = x_1 x_3$, and $x_6 = x_2 x_3$. Significance tests can judge which, if any, of the cross-product terms are needed in the model.

When interaction exists and the model contains cross-product terms, it is more difficult to summarize simply the relationships. One approach is to sketch a collection of lines such as those in Figure 11.10 to describe graphically how the relationship between two variables changes according to the values of other variables. Another possibility is to divide the data into groups according to the value on a control variable (e.g., high on x_2, medium on x_2, low on x_2) and report the slope between y and x_1 within each subset as a means of describing the interaction.

11.5 Comparing Regression Models

When the number of explanatory variables increases, the multiple regression model becomes more difficult to interpret and some variables may become redundant. This is especially true when some explanatory variables are cross products of others, to allow for interaction. Not all the variables may be needed in the model. We next present a significance test of whether a model fits significantly better than a simpler model containing only some of the explanatory variables.

COMPLETE AND REDUCED MODELS

We refer to the full model with all the explanatory variables as the ***complete model***. The model containing only some of these variables is called the ***reduced model***. The reduced model is said to be *nested* within the complete model, being a special case of it.

The complete and reduced models are identical if the partial regression coefficients for the extra variables in the complete model all equal 0. In that case, none of the extra explanatory variables increases the explained variability in y, in the population of interest. Testing whether the complete model is identical to the reduced model is equivalent to testing whether the extra parameters in the complete model equal 0. The alternative hypothesis is that at least one of these extra parameters is not 0, in which case the complete model fits better than the reduced model.

For instance, a complete model with three explanatory variables and all two-variable interaction terms is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3.$$

The reduced model without the interaction terms is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

The test comparing the complete model to the reduced model has H_0: $\beta_4 = \beta_5 = \beta_6 = 0$.

COMPARING MODELS BY COMPARING SSE VALUES OR R^2-VALUES

The test statistic for comparing two regression models compares the residual sums of squares for the two models. Denote SSE $= \sum(y-\hat{y})^2$ for the reduced model by SSE_r and for the complete model by SSE_c. Now, $\text{SSE}_r \geq \text{SSE}_c$, because the reduced model has fewer explanatory variables and makes poorer overall predictions. Even if H_0 were true, we would not expect the estimates of the extra parameters and the difference $(\text{SSE}_r - \text{SSE}_c)$ to equal 0. Some reduction in error occurs from fitting the extra terms because of sampling variability.

The test statistic uses the reduction in error, $\text{SSE}_r - \text{SSE}_c$, that results from adding the extra variables. An equivalent statistic uses the R^2-values, R_c^2 for the complete model and R_r^2 for the reduced model. The two expressions for the test statistic are

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

Here, df_1 is the number of extra terms in the complete model (e.g., 3 when we add three interaction terms to get the complete model) and df_2 is the residual df for the complete model. A relatively large reduction in error (or relatively large increase in R^2) yields a large F test statistic and a small P-value. As usual for F statistics, the P-value is the right-tail probability.

Example 11.7 **Comparing Models for Mental Impairment** For the mental impairment data, a comparison of the complete model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

to the reduced model

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

analyzes whether interaction exists. The complete model has only one additional term, and the null hypothesis is H_0: $\beta_3 = 0$.

The sum of squared errors for the complete model is $\text{SSE}_c = 758.8$ (Table 11.8 on page 327), while for the reduced model it is $\text{SSE}_r = 768.2$ (Table 11.7 on page 318). The difference

$$\text{SSE}_r - \text{SSE}_c = 768.2 - 758.8 = 9.4$$

has $df_1 = 1$ since the complete model has one more parameter. From Table 11.8, $df_2 = n-(p+1) = 40-(3+1) = 36$, the residual df in that table. The F test statistic equals

$$F = \frac{(\text{SSE}_r - \text{SSE}_c)/df_1}{\text{SSE}_c/df_2} = \frac{9.4/1}{758.8/36} = 0.45.$$

Equivalently, the R^2-values for the two models are $R_r^2 = 0.339$ and $R_c^2 = 0.347$, so

$$F = \frac{\left(R_c^2 - R_r^2\right)/df_1}{\left(1-R_c^2\right)/df_2} = \frac{(0.347-0.339)/1}{(1-0.347)/36} = 0.45.$$

From software, the P-value from the F distribution with $df_1 = 1$ and $df_2 = 36$ is $P = 0.51$. There is little evidence that the complete model is better. The null hypothesis seems plausible, so the reduced model is adequate.

When H_0 contains a single parameter, the t test is available. In fact, from the previous section (and Table 11.8), the t statistic is

$$t = \frac{b_3}{se} = \frac{-0.00087}{0.0013} = -0.67.$$

It also has a P-value of 0.51 for H_a: $\beta_3 \neq 0$. We get the same result with the t test as with the F test for complete and reduced models. In fact, the F test statistic equals the square of the t statistic. (Refer to page 325.) ■

The t test method is limited to testing one parameter at a time. The F test can test *several* regression parameters together to analyze whether at least one of them is nonzero, such as in the global F test of H_0: $\beta_1 = \cdots = \beta_p = 0$ or the test comparing a complete model to a reduced model. F tests are equivalent to t tests only when H_0 contains a single parameter.

11.6 Partial Correlation*

Multiple regression models describe the effect of an explanatory variable on the response variable while controlling for other variables of interest. Related measures describe the strength of the association. For example, to describe the association between mental impairment and life events, controlling for SES, we could ask, "Controlling for SES, what proportion of the variation in mental impairment does life events explain?"

These measures describe the partial association between y and a particular explanatory variable, whereas the multiple correlation and R^2 describe the association between y and the entire set of explanatory variables in the model. The *partial correlation* is based on the ordinary correlations between each pair of variables. For a single control variable, it is calculated as follows:

Partial Correlation

The sample ***partial correlation*** between y and x_1, controlling for x_2, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}}.$$

In the symbol $r_{yx_1 \cdot x_2}$, the variable to the right of the dot represents the controlled variable. The analogous formula for $r_{yx_2 \cdot x_1}$ (i.e., controlling x_1) is

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}}.$$

Since one variable is controlled, the partial correlations $r_{yx_1 \cdot x_2}$ and $r_{yx_2 \cdot x_1}$ are called ***first-order partial correlations***.

Example 11.8

Partial Correlation between Education and Crime Rate Example 11.1 (page 308) discussed a data set for counties in Florida, with y = crime rate, x_1 = education, and x_2 = urbanization. The pairwise correlations are $r_{yx_1} = 0.468$, $r_{yx_2} = 0.678$, and $r_{x_1x_2} = 0.791$. It was surprising to observe a positive correlation between crime rate and education. Can it be explained by their joint dependence on urbanization? This is plausible if the association disappears when we control for urbanization.

The partial correlation between crime rate and education, controlling for urbanization, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}} = \frac{0.468 - 0.678(0.791)}{\sqrt{(1 - 0.678^2)(1 - 0.791^2)}} = -0.152.$$

Not surprisingly, $r_{yx_1 \cdot x_2}$ is much smaller than r_{yx_1}. It even has a different direction, illustrating Simpson's paradox. The relationship between crime rate and education may well be spurious, reflecting their joint dependence on urbanization. ■

INTERPRETING PARTIAL CORRELATIONS

The partial correlation has properties similar to those for the ordinary correlation between two variables. We list the properties below for $r_{yx_1 \cdot x_2}$, but analogous properties apply to $r_{yx_2 \cdot x_1}$.

- $r_{yx_1 \cdot x_2}$ falls between -1 and $+1$.
- The larger the absolute value of $r_{yx_1 \cdot x_2}$, the stronger the association between y and x_1, controlling for x_2.
- The value of a partial correlation does not depend on the units of measurement of the variables.
- $r_{yx_1 \cdot x_2}$ has the same sign as the partial slope (b_1) for the effect of x_1 in the prediction equation $\hat{y} = a + b_1x_1 + b_2x_2$, because the same variable (x_2) is controlled in the model as in the correlation.
- Under the assumptions for conducting inference for multiple regression (see the beginning of Section 11.3), $r_{yx_1 \cdot x_2}$ estimates the correlation between y and x_1 at every *fixed* value of x_2. If we could control x_2 by considering a subpopulation of subjects all having the same value on x_2, then $r_{yx_1 \cdot x_2}$ estimates the correlation between y and x_1 for that subpopulation.
- The sample partial correlation is identical to the ordinary bivariate correlation computed for the points in the *partial regression plot* (page 314).

INTERPRETING SQUARED PARTIAL CORRELATIONS

Like r^2 and R^2, the square of a partial correlation has a proportional reduction in error (PRE) interpretation. For example, $r^2_{yx_2 \cdot x_1}$ is the proportion of variation in y explained by x_2, controlling for x_1. This squared measure describes the effect of removing from consideration the portion of the total sum of squares (TSS) in y that is explained by x_1, and then finding the proportion of the remaining unexplained variation in y that is explained by x_2.

Squared Partial Correlation

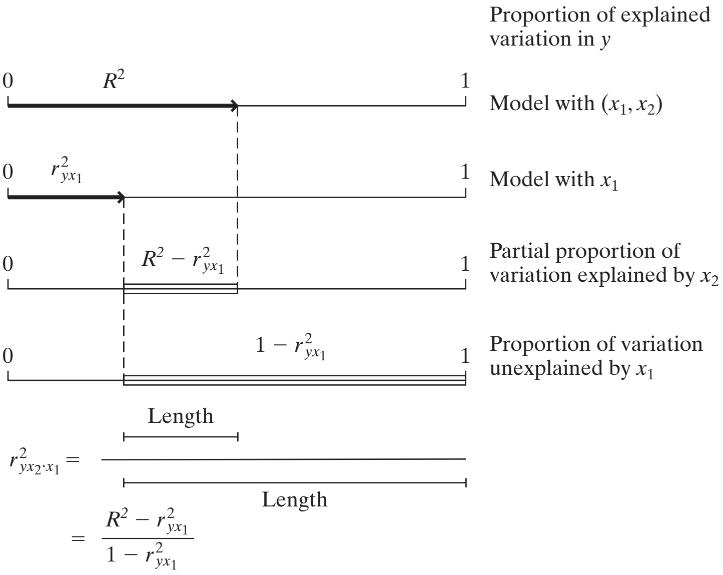
The square of the partial correlation $r_{yx_2 \cdot x_1}$ represents the proportion of the variation in y that is explained by x_2, out of that left unexplained by x_1. It equals

$$r^2_{yx_2 \cdot x_1} = \frac{R^2 - r^2_{yx_1}}{1 - r^2_{yx_1}} = \frac{\text{Partial proportion explained uniquely by } x_2}{\text{Proportion unexplained by } x_1}.$$

From Section 9.4 (page 259), $r^2_{yx_1}$ represents the proportion of the variation in y explained by x_1. The remaining proportion $(1 - r^2_{yx_1})$ represents the variation left unexplained. When x_2 is added to the model, it accounts for some additional variation. The total proportion of the variation in y accounted for by x_1 and x_2 jointly is R^2 for the model with both x_1 and x_2 as explanatory variables. So, $R^2 - r^2_{yx_1}$ is the additional proportion of the variability in y explained by x_2, after the effects of x_1

have been removed or controlled. The maximum this difference could be is $1 - r^2_{yx_1}$, the proportion of variation yet to be explained after accounting for the influence of x_1. The additional explained variation $R^2 - r^2_{yx_1}$ divided by this maximum possible difference is a measure that has a maximum possible value of 1. In fact, as the above formula suggests, this ratio equals the squared partial correlation between y and x_2, controlling for x_1. Figure 11.11 illustrates this property of the squared partial correlation.

FIGURE 11.11: Representation of $r^2_{yx_2 \cdot x_1}$ as the Proportion of Variability That Can Be Explained by x_2, of That Left Unexplained by x_1



Proportion of explained variation in y
0
R^2
1
Model with (x_1, x_2)
0
$r^2_{yx_1}$
1
Model with x_1
0
$R^2 - r^2_{yx_1}$
1
Partial proportion of variation explained by x_2
0
$1 - r^2_{yx_1}$
1
Proportion of variation unexplained by x_1
Length
$r^2_{yx_2 \cdot x_1} =$
Length
$= \frac{R^2 - r^2_{yx_1}}{1 - r^2_{yx_1}}$

Example 11.9

Partial Correlation of Life Events with Mental Impairment We return to the example with y = mental impairment, x_1 = life events, and x_2 = SES. Software reports the correlation matrix

```
            IMPAIR        LIFE          SES
IMPAIR      1.000         .372         -.399
LIFE         .372        1.000          .123
SES         -.399         .123         1.000
```

So, $r_{yx_1} = 0.372$, $r_{yx_2} = -0.399$, and $r_{x_1x_2} = 0.123$. The partial correlation between mental impairment and life events, controlling for SES, is

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right)\left(1 - r^2_{x_1x_2}\right)}} = \frac{0.372 - (-0.399)(0.123)}{\sqrt{[1 - (-0.399)^2](1 - 0.123^2)}} = 0.463.$$

The partial correlation, like the correlation of 0.37 between mental impairment and life events, is moderately positive.

Since $r^2_{yx_1 \cdot x_2} = (0.463)^2 = 0.21$, controlling for SES, 21% of the variation in mental impairment is explained by life events. Alternatively, since $R^2 = 0.339$ (Table 11.7),

$$r^2_{yx_1 \cdot x_2} = \frac{R^2 - r^2_{yx_2}}{1 - r^2_{yx_2}} = \frac{0.339 - (-0.399)^2}{1 - (-0.399)^2} = 0.21.$$

■

HIGHER-ORDER PARTIAL CORRELATIONS

The connection between squared partial correlation values and R-squared also applies when the number of control variables exceeds one. For example, with three explanatory variables, let $R^2_{y(x_1,x_2,x_3)}$ denote the value of R^2. The square of the partial correlation between y and x_3, controlling for x_1 and x_2, relates to how much larger this is than the R^2-value for the model with only x_1 and x_2 as explanatory variables, which we denote by $R^2_{y(x_1,x_2)}$. The squared partial correlation is

$$r^2_{yx_3 \cdot x_1,x_2} = \frac{R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}}{1 - R^2_{y(x_1,x_2)}}.$$

In this expression, $R^2_{y(x_1,x_2,x_3)} - R^2_{y(x_1,x_2)}$ is the increase in the proportion of explained variance from adding x_3 to the model. The denominator $1 - R^2_{y(x_1,x_2)}$ is the proportion of the variation left unexplained when x_1 and x_2 are the only explanatory variables in the model.

The partial correlation $r_{yx_3 \cdot x_1,x_2}$ is called a ***second-order partial correlation***, since it controls two variables. It has the same sign as b_3 in the prediction equation $\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$, which also controls x_1 and x_2 in describing the effect of x_1.

11.7 Standardized Regression Coefficients*

As in bivariate regression, the sizes of regression coefficients in multiple regression models depend on the units of measurement for the variables. To compare the relative effects of two explanatory variables, it is appropriate to compare their coefficients only if the variables have the same units. Otherwise, *standardized* versions of the regression coefficients provide more meaningful comparisons.

Standardized Regression Coefficient

The ***standardized regression coefficient*** for an explanatory variable represents the change in the mean of y, in y standard deviations, for a one standard deviation increase in that variable, controlling for the other explanatory variables in the model. We denote them by $\beta_1^*, \beta_2^*, \ldots.$

If $|\beta_2^*| > |\beta_1^*|$, for example, then a standard deviation increase in x_2 has a greater partial effect on y than does a standard deviation increase in x_1.

THE STANDARDIZATION MECHANISM

The standardized regression coefficients represent the values the regression coefficients take when the units are such that y and the explanatory variables all have equal standard deviations, such as when we use standardized variables. We can obtain the standardized regression coefficients from the unstandardized coefficients. Let s_y denote the sample standard deviation of y, and let $s_{x_1}, s_{x_2}, \ldots, s_{x_p}$ denote the sample standard deviations of the explanatory variables.

The estimates of the standardized regression coefficients relate to the estimates of the unstandardized coefficients by

$$b_1^* = b_1\left(\frac{s_{x_1}}{s_y}\right), \quad b_2^* = b_2\left(\frac{s_{x_2}}{s_y}\right), \ldots.$$

Example 11.10 **Standardized Coefficients for Mental Impairment** The prediction equation relating mental impairment to life events and SES is

$$\hat{y} = 28.23 + 0.103x_1 - 0.097x_2.$$

Table 11.2 reported the sample standard deviations $s_y = 5.5$, $s_{x_1} = 22.6$, and $s_{x_2} = 25.3$. The unstandardized coefficient of x_1 is $b_1 = 0.103$, so the estimated standardized coefficient is

$$b_1^* = b_1 \left(\frac{s_{x_1}}{s_y}\right) = 0.103 \left(\frac{22.6}{5.5}\right) = 0.43.$$

Since $b_2 = -0.097$, the standardized value is

$$b_2^* = b_2 \left(\frac{s_{x_2}}{s_y}\right) = -0.097 \left(\frac{25.3}{5.5}\right) = -0.45.$$

The estimated change in the mean of y for a standard deviation increase in x_1, controlling for x_2, has similar magnitude as the estimated change for a standard deviation increase in x_2, controlling for x_1. However the partial effect of x_1 is positive, whereas the partial effect of x_2 is negative.

Table 11.10, which repeats Table 11.5, shows how SPSS reports the estimated standardized regression coefficients. It uses the heading BETA (as does Stata), reflecting the alternative name ***beta weights*** for these coefficients. ■

TABLE 11.10: SPSS Output for Fit of Multiple Regression Model to Mental Impairment Data from `Mental` Data File, with Standardized Coefficients

```
          Unstandardized coefficients   Standardized coefficients
                   B      Std. Error       Beta          t          Sig.
(Constant)     28.230     2.174                        12.984      .000
LIFE             .103      .032            .428         3.177      .003
SES             -.097      .029           -.451        -3.351      .002
```

PROPERTIES OF STANDARDIZED REGRESSION COEFFICIENTS

For bivariate regression, standardizing the regression coefficient yields the correlation. For the multiple regression model, the standardized partial regression coefficient relates to the partial correlation (Exercise 11.65), and it usually takes a similar value.

Unlike the partial correlation, however, b_i^* need not fall between -1 and $+1$. A value $|b_i^*| > 1$ occasionally occurs when x_i is highly correlated with the set of other explanatory variables in the model. In such cases, the standard errors are usually large and the estimates are unreliable.

Since a standardized regression coefficient is a multiple of the unstandardized coefficient, one equals 0 when the other does. The test of H_0: $\beta_i^* = 0$ is equivalent to the t test of H_0: $\beta_i = 0$. It is unnecessary to have separate tests for these coefficients. In the sample, the magnitudes of the $\{b_i^*\}$ have the same relative sizes as the t statistics from those tests. For example, the explanatory variable with the greatest standardized partial effect is the one that has the largest t statistic, in absolute value.

STANDARDIZED FORM OF PREDICTION EQUATION

Regression equations have an expression using the standardized regression coefficients. In this equation, the variables appear in standardized form.

Let $z_y, z_{x_1}, \ldots, z_{x_p}$ denote the standardized versions of the variables $y, x_1, \ldots, x_p$. For instance, $z_y = (y - \bar{y})/s_y$ represents the number of standard deviations that an observation on y falls from its mean. Each subject's scores on $y, x_1, \ldots, x_p$ have corresponding z-scores for $z_y, z_{x_1}, \ldots, z_{x_p}$. If a subject's score on x_1 is such that $z_{x_1} = (x_1 - \bar{x}_1)/s_{x_1} = 2.0$, for instance, then that subject falls two standard deviations above the mean $\bar{x}_1$ on x_1.

Let $\hat{z}_y = (\hat{y} - \bar{y})/s_y$ denote the predicted z-score for the response variable. For the standardized variables and the estimated standardized regression coefficients, the prediction equation is

$$\hat{z}_y = b_1^* z_{x_1} + b_2^* z_{x_2} + \cdots + b_p^* z_{x_p}.$$

This equation predicts how far an observation on y falls from its mean, in standard deviation units, based on how far the explanatory variables fall from their means, in standard deviation units. The standardized coefficients are the weights attached to the standardized explanatory variables in contributing to the predicted standardized response variable.

Example 11.11 **Standardized Prediction Equation for Mental Impairment** Example 11.10 found that the estimated standardized regression coefficients for the life events and SES predictors of mental impairment are $b_1^* = 0.43$ and $b_2^* = -0.45$. The prediction equation relating the standardized variables is therefore

$$\hat{z}_y = 0.43 z_{x_1} - 0.45 z_{x_2}.$$

A subject who is two standard deviations above the mean on life events but two standard deviations below the mean on SES has a predicted standardized mental impairment of

$$\hat{z}_y = 0.43(2) - 0.45(-2) = 1.8.$$

The predicted mental impairment for that subject is 1.8 standard deviations above the mean. If the distribution of mental impairment is approximately normal, this subject might well have mental health problems, since only about 4% of the scores in a normal distribution fall at least 1.8 standard deviations above their mean. ■

In the prediction equation with standardized variables, no intercept term appears. Why is this? When the standardized explanatory variables all equal 0, those variables all fall at their means. Then, $\hat{y} = \bar{y}$, so

$$\hat{z}_y = \frac{\hat{y} - \bar{y}}{s_y} = 0.$$

So, this merely tells us that a subject who is at the mean on each explanatory variable is predicted to be at the mean on the response variable.

CAUTIONS IN COMPARING STANDARDIZED REGRESSION COEFFICIENTS

To assess which explanatory variable in a multiple regression model has the greatest impact on the response variable, it is tempting to compare their standardized regression coefficients. Make such comparisons with caution. In some cases, the observed differences in the b_i^* may simply reflect sampling error. In particular, when multicollinearity exists, the standard errors are high and the estimated standardized coefficients may be unstable.

For a standardized regression coefficient to make sense, the variation in the explanatory variable must be representative of the variation in the population of interest. It is inappropriate to compare the standardized effect of an explanatory variable to others if the study purposely sampled values of that variable in a narrow range. This comment relates to a warning in Section 9.6 (page 272) about the correlation: Its value depends strongly on the range of explanatory variable values sampled.

Keep in mind also that the effects are partial ones, depending on which other variables are in the model. An explanatory variable that seems important in one system of variables may seem unimportant when other variables are controlled. For example, it is possible that $|b_2^*| > |b_1^*|$ in a model with two explanatory variables, yet when a third explanatory variable is added to the model, $|b_2^*| < |b_1^*|$.

It is unnecessary to standardize to compare the effect of the same variable for two groups, such as in comparing the results of separate regressions for females and males, since the units of measurement are the same in each group. In fact, it is usually unwise to standardize in this case, because the standardized coefficients are more susceptible than the unstandardized coefficients to differences in the standard deviations of the explanatory variables. Two groups that have the same value for an estimated regression coefficient have different standardized coefficients if the standard deviation of the explanatory variable differs for the two groups.

11.8 Chapter Summary

This chapter generalized the bivariate regression model to include additional explanatory variables. The ***multiple regression equation*** relating a response variable y to a set of p explanatory variables is

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

- The $\{\beta_i\}$ are ***partial regression coefficients***. The value β_i is the change in the mean of y for a one-unit change in x_i, controlling for the other variables in the model.
- The ***multiple correlation*** R describes the association between y and the collective set of explanatory variables. It equals the correlation between the observed and predicted y-values. It falls between 0 and 1.
- $R^2 = (\text{TSS} - \text{SSE})/\text{TSS}$ represents the *proportional reduction in error* from predicting y using the prediction equation $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$ instead of $\bar{y}$. It equals the square of the multiple correlation.
- A ***partial correlation***, such as $r_{yx_1 \cdot x_2}$, describes the association between two variables, controlling for others. It falls between -1 and $+1$. The squared partial correlation between y and x_i represents the proportion of the variation in y that can be explained by x_i, out of that part left unexplained by a set of control variables.
- An F ***statistic*** tests H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, that the response variable is independent of all the explanatory variables. F-values are nonnegative and have two df values. A large F test statistic and small P-value suggest that the response variable is correlated with at least one of the explanatory variables.
- Individual t tests and confidence intervals for $\{\beta_i\}$ analyze partial effects of each explanatory variable, controlling for the other variables in the model.

- ***Interaction*** between x_1 and x_2 in their effects on y means that the effect of either explanatory variable changes as the value of the other one changes. We can allow this by adding cross products of explanatory variables to the model, such as the term $\beta_3(x_1x_2)$.
- To ***compare regression models***, a *complete* model and a simpler *reduced* model, an F test compares the SSE values or the R^2-values.
- ***Standardized regression coefficients*** do not depend on the units of measurement. The estimated standardized coefficient b_i^* describes the change in y, in y standard deviation units, for a one standard deviation increase in x_i, controlling for the other explanatory variables.

To illustrate, with $p = 3$ explanatory variables, the prediction equation is

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3.$$

Fixing x_2 and x_3, a straight line describes the relation between y and x_1. Its slope b_1 is the change in $\hat{y}$ for a one-unit increase in x_1, controlling for x_2 and x_3. The multiple correlation R is at least as large as the absolute values of the correlations r_{yx_1}, r_{yx_2}, and r_{yx_3}. The squared partial correlation $r^2_{yx_3 \cdot x_1, x_2}$ is the proportion of the variation of y that is explained by x_3, out of that part of the variation left unexplained by x_1 and x_2. The estimated standardized regression coefficient $b_1^* = b_1(s_{x_1}/s_y)$ describes the effect of a standard deviation change in x_1, controlling for x_2 and x_3.

Table 11.11 summarizes the basic properties and inference methods for these measures and those introduced in Chapter 9 for bivariate regression.

TABLE 11.11: Summary of Bivariate and Multiple Regression

	Bivariate Regression	Multiple Regression	
Model	$E(y) = \alpha + \beta x$	$E(y) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$	
Prediction equation	$\hat{y} = a + bx$	$\hat{y} = a + b_1 x_1 + \cdots + b_p x_p$	
	Overall effect of x	Simultaneous effect of $x_1, \ldots, x_p$	Partial effect of one x_i
Measures	b = Slope r = Correlation, standardized slope, $-1 \leq r \leq 1$, r has the same sign as b r^2 = PRE measure, $0 \leq r^2 \leq 1$	R = Multiple correlation, $0 \leq R \leq 1$ R^2 = PRE measure, $0 \leq R^2 \leq 1$	b_i = Partial slope b_i^* = Standardized regression coefficient Partial correlation, $-1 \leq r_{yx_1 \cdot x_2} \leq 1$, same sign as b_i and b_i^*, $r^2_{yx_1 \cdot x_2}$ is PRE measure
Tests of no association	H_0: $\beta = 0$ or H_0: $\rho = 0$, y not associated with x	H_0: $\beta_1 = \cdots = \beta_p = 0$, y not associated with $x_1, \ldots, x_p$	H_0: $\beta_i = 0$, y not associated with x_i, controlling for other x variables
Test statistic	$t = \frac{b}{se} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $df = n - 2$	$F = \frac{\text{Regression MS}}{\text{Residual MS}} = \frac{R^2/p}{(1-R^2)/[n-(p+1)]}$, $df_1 = p$, $df_2 = n - (p+1)$	$t = \frac{b_i}{se}$ $df = n - (p + 1)$

Regression with Categorical Predictors: Analysis of Variance Methods

Chapter 12

CHAPTER OUTLINE

The regression models presented so far have quantitative explanatory variables. This chapter shows how a regression model can contain categorical explanatory variables.

Chapter 7 presented methods for comparing the means of two groups. Those methods extend for comparing means of *several* groups. The methods relate to the association between a *quantitative* response variable and a *categorical* explanatory variable. The mean of the quantitative response variable is compared among groups that are categories of the explanatory variable. For example, for a comparison of mean annual income among blacks, whites, and Hispanics, the quantitative response variable is annual income and the categorical explanatory variable is racial–ethnic status. We can use the regression methods of this chapter to do this.

Artificial variables called *dummy variables* can represent the categories of a categorical explanatory variable in a regression model. The inferential method for testing equality of several means is often called the ***analysis of variance***, abbreviated as ***ANOVA***. We'll see that the name refers to the way the significance test focuses on two types of variability in the data.

Categorical explanatory variables in ANOVA are called ***factors***. ANOVA methods extend to incorporate multiple factors, for example, to compare mean income across categories of both racial–ethnic status and gender. We first present analyses for *independent samples*. When each sample has the same subjects or the samples are matched, the samples are *dependent* and different methods apply. We also present such methods, referred to as ***repeated-measures ANOVA***.

12.1 Regression Modeling with Dummy Variables for Categories

We can use a regression model for the relationship between a quantitative response variable and a *categorical* explanatory variable. We shall use the following example to illustrate methods.

Example 12.1

Political Ideology by Political Party ID Table 12.1 summarizes observations on political ideology for three groups, based on data from subjects of ages 18–27 in the 2014 General Social Survey. The three groups are the (Democrat, Independent, Republican) categories of the explanatory variable, political party identification (ID). Political ideology, the response variable, is measured on a seven-point scale, ranging from extremely liberal (1) to extremely conservative (7). For each party ID, Table 12.1 shows the number of subjects who made each response. For instance, of 83 Democrats, 5 responded extremely liberal, 18 responded liberal, . . . , 2 responded extremely conservative.

TABLE 12.1: Political Ideology by Political Party Identification (ID), for Respondents of Age 18–27

	Political Ideology							Sample		Standard
Party ID	1	2	3	4	5	6	7	Size	Mean	Deviation
Democrat	5	18	19	25	7	7	2	83	3.48	1.43
Independent	4	19	27	79	13	9	6	157	3.82	1.23
Republican	1	3	1	11	10	11	1	38	4.66	1.36

Note: For political ideology, 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative.

Since Table 12.1 displays the data as counts in a contingency table, we could use methods for categorical data (Chapter 8). The chi-squared test treats both variables as nominal, however, whereas political ideology is ordinal. That test is not directed toward detecting whether responses have a higher or lower mean in some groups than others. The ordinal measure of association gamma is inappropriate, because it requires both variables to be ordinal. Here, the groups, which are the categories of political party ID, are nominal.

When an ordinal response has many categories, one approach assigns scores to its levels and treats it as a quantitative variable. This is a reasonable strategy when we want to focus on a measure of center (such as the mean) rather than on the proportions in particular categories, and when the observations do not mainly fall at one of the boundary categories. For Table 12.1, for instance, interest might focus on how liberal or conservative the responses tend to be for each group, in some average sense. We analyze these data by assigning the scores (1, 2, 3, 4, 5, 6, 7) to the levels of political ideology and then comparing means. The higher the mean score, the more conservative the group's responses tended to be. For these scores, Table 12.1 shows the mean and standard deviation for each group. The overall sample mean is $\bar{y} = 3.83$, not far from the score of 4.0 corresponding to moderate ideology. ■

REGRESSION WITH DUMMY (INDICATOR) VARIABLES

How can we enter a categorical explanatory variable such as party ID in a regression model? We set up an indicator variable to equal 1 if an observation comes from a particular category and 0 otherwise. With three categories (e.g., the party IDs), we use two indicator variables. The first, denoted by z_1, equals 1 for observations from the first category and equals 0 otherwise. The second, denoted by z_2, equals 1 for observations from the second category and equals 0 otherwise. That is,

$z_1 = 1$ and $z_2 = 0$: observations from category 1.
$z_1 = 0$ and $z_2 = 1$: observations from category 2.
$z_1 = 0$ and $z_2 = 0$: observations from category 3.

It is unnecessary and redundant to create a variable for the last (third) category, because values of 0 for z_1 and z_2 identify observations from it.

The indicator variables z_1 and z_2 are called ***dummy variables***. They indicate the category for an observation. That is, they give a classification, not a magnitude, for the factor. Table 12.2 summarizes the dummy variables for three categories.

For three groups, denote the population means on y by μ_1, μ_2, and μ_3. For the dummy variables just defined, consider the multiple regression equation

$$E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2.$$

TABLE 12.2: The Two Dummy Variables for a Categorical Explanatory Variable with Three Categories

Category	z_1	z_2
1	1	0
2	0	1
3	0	0

For observations from category 3, $z_1 = z_2 = 0$. The equation then simplifies to

$$E(y) = \alpha + \beta_1(0) + \beta_2(0) = \alpha.$$

So, α represents the population mean μ_3 of y for the last category. For observations from category 1, $z_1 = 1$ and $z_2 = 0$, so

$$E(y) = \alpha + \beta_1(1) + \beta_2(0) = \alpha + \beta_1$$

equals the population mean μ_1 for that category. Similarly, $\alpha + \beta_2$ equals the population mean μ_2 for category 2 (let $z_1 = 0$ and $z_2 = 1$).

Since $\alpha + \beta_1 = \mu_1$ and $\alpha = \mu_3$, the parameter $\beta_1 = \mu_1 - \mu_3$. Similarly, $\beta_2 = \mu_2 - \mu_3$. Table 12.3 summarizes the parameters of the regression model and their correspondence with the population means. The β coefficient of a dummy variable represents the difference between the mean for the category that dummy variable represents and the mean of the category not having its own dummy variable.

TABLE 12.3: Interpretation of Coefficients of Dummy Variables in Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ Having Explanatory Variable with Three Categories

Category	z_1	z_2	Mean of y	Interpretation of β
1	1	0	$\mu_1 = \alpha + \beta_1$	$\beta_1 = \mu_1 - \mu_3$
2	0	1	$\mu_2 = \alpha + \beta_2$	$\beta_2 = \mu_2 - \mu_3$
3	0	0	$\mu_3 = \alpha$	

Dummy variable coding works because it allows the population means to take arbitrary values, with no assumed distances between categories. Using a single variable with coding such as $z = 1$ for category 1, $z = 2$ for category 2, and $z = 3$ for category 3 would not work. The model $E(y) = \alpha + \beta z$ would then assume an ordering as well as equal distances between categories. It treats the factor as if it were quantitative, which is improper. Whereas we need only one term in a regression model to represent the linear effect of a quantitative explanatory variable, for a categorical explanatory variable we need one fewer term than the number of categories.

Example 12.2

Regression Model for Political Ideology and Party ID For Table 12.1, the categorical explanatory variable (political party ID) has three categories. The regression model for $y =$ political ideology is

$$E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2,$$

with $z_1 = 1$ only for Democrats, $z_2 = 1$ only for Independents, and $z_1 = z_2 = 0$ for Republicans. Table 12.4 shows some software output for fitting this regression model. No dummy variable estimate appears in the table for party 3 (Republicans), because it is redundant to include a dummy variable for the last category.

TABLE 12.4: Software Output for Fitting Regression Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ to Data on $y =$ Political Ideology with Dummy Variables z_1 and z_2 for Political Party ID

```
IDEOLOGY         Coef.   Std. Err.      t        P>|t|
(Constant)       4.658    0.2126      21.91    <0.0001
PARTY      1    -1.176    0.2567      -4.58    <0.0001
           2    -0.836    0.2369      -3.53     0.0005
           3     0.000
```

The prediction equation is $\hat{y} = 4.66 - 1.18z_1 - 0.84z_2$. The coefficients in the prediction equation relate to the sample means in the same manner that the regression parameters relate to the population means. Just as $\alpha = \mu_3$, so does its estimate $4.66 = \bar{y}_3$, the sample mean for Republicans. Similarly, the coefficient of z_1 is $-1.18 = \bar{y}_1 - \bar{y}_3$ and the coefficient of z_2 is $-0.84 = \bar{y}_2 - \bar{y}_3$.

Some software codes factors so that the first category is the one lacking its own dummy variable. The reported model parameter estimates then differ, but they yield the same estimates for differences between category means. For example, R software sets up dummy variables for categories 2 and 3 and yields estimates

```
                Estimate
(Intercept)       3.48
party2            0.34
party3            1.18
```

The estimate for party1 (Democrats) is 0, so the estimated difference between the means for Democrats and Republicans is still $0 - 1.18 = -1.18$. ■

USING REGRESSION FOR A SIGNIFICANCE TEST COMPARING MEANS

For the three groups that are categories of a categorical explanatory variable with three categories, consider H_0: $\mu_1 = \mu_2 = \mu_3$. If H_0 is true, then $\mu_1 - \mu_3 = 0$ and $\mu_2 - \mu_3 = 0$. Recall that $\mu_1 - \mu_3 = \beta_1$ and $\mu_2 - \mu_3 = \beta_2$ in the regression model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ with dummy variables for categories 1 and 2. So, the hypothesis is equivalent to H_0: $\beta_1 = \beta_2 = 0$ in that model. If all β-values in the model equal 0, then the mean of the response variable equals α for each category.

As usual, we assume randomization. This could be either a single random sample, with subjects then classified by group, or *independent random samples* from the groups. The assumption for inferences in regression modeling that the conditional distributions of y about the regression equation are normal with constant standard deviation corresponds here to the population distributions for the groups being normal, with identical standard deviations.

We can perform the test using the F test of H_0: $\beta_1 = \beta_2 = 0$ for the regression model. As shown in Section 11.3 (page 320), the P-value is the right-tail probability that the F test statistic exceeds the observed F-value. The larger the F test statistic, the smaller the P-value. Table 12.5 shows the ANOVA table for the regression model on political ideology and party ID. The F test statistic equals 10.51, with $df_1 = 2$ and $df_2 = 275$, for testing H_0: $\beta_1 = \beta_2 = 0$, which is equivalently H_0: $\mu_1 = \mu_2 = \mu_3$ for the three party IDs. The P-value is <0.0001, strong evidence against H_0. We conclude that a difference exists among the population mean political ideology values for the three political party IDs.

TABLE 12.5: Software Output of ANOVA Table for Regression Model $E(y) = \alpha + \beta_1 z_1 + \beta_2 z_2$ for $y =$ Political Ideology and Political Party ID. The "regression sum of squares" is called the "model sum of squares" by Stata and SAS.

```
               Sum of            Mean
               Squares    df    Square    F Value    Prob>F
Regression      36.11      2    18.05       10.51   <0.0001
Residual       472.28    275     1.72
Total          508.39    277
```

ROBUSTNESS AND EFFECTS OF VIOLATIONS OF ASSUMPTIONS

In addition to randomization, each method presented in this chapter assumes that the groups have population distributions that are normal with identical standard deviations. These are stringent assumptions that are never exactly satisfied in practice.

Moderate departures from normality of the population distributions can be tolerated. The F distribution still provides a good approximation to the actual sampling distribution of the F test statistic. This is particularly true for larger sample sizes, since the sampling distributions then have weaker dependence on the shape of the population distribution. Moderate departures from equal standard deviations can also be tolerated. When the sample sizes are identical for the groups, the F test is very robust to violations of this assumption.

Constructing histograms for each sample data distribution helps to check for extreme deviations from these assumptions. Misleading results may occur in the F tests if the population distributions are highly skewed and the sample size is small, or if there are relatively large differences among the population standard deviations (say, the largest sample standard deviation is several times as large as the smallest one) and the sample sizes are unequal. When the distributions are very highly skewed, the mean may not even be an appropriate summary measure.

As in other inferences, the quality of the sample is most crucial. Conclusions may be invalid if the observations in the separate groups compared are not independent random samples.

12.2 Multiple Comparisons of Means

When the P-value is small for comparing several means for groups corresponding to categories of the explanatory variable, this does not indicate which means are different or how different they are. In practice, it is more informative to estimate differences between the population means than merely to test whether they are all equal. Confidence intervals do this. Even if the P-value is not small, it still is informative to determine the plausible sizes of the differences among the population means.

CONFIDENCE INTERVALS COMPARING PAIRS OF MEANS

We can construct a confidence interval for each mean or for each difference between a pair of means. For a categorical variable with g categories corresponding to g groups, denote the sample means by $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g$ and the corresponding populations by $\mu_1, \mu_2, \ldots, \mu_g$. Let $N = n_1 + n_2 + \cdots + n_g$ denote the total sample size.

Confidence Intervals for Pairwise Comparisons of Means

A confidence interval for $\mu_i - \mu_j$ is

$$(\bar{y}_i - \bar{y}_j) \pm ts\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

In this formula, s^2 is the residual mean square in the regression model for g groups. The t-value for the chosen confidence level has $df = N - g$.

The t-value is based on df for the variance estimate s^2, which is $df = N - g$ since the model has g parameters. Evidence exists of a difference between μ_i and μ_j when the interval[1] does not contain 0.

Confidence intervals, like tests, do not depend strongly on the normality assumption. When the standard deviations are quite different, with the ratio of the largest to smallest exceeding about 2, it is preferable to use intervals based on separate standard deviations for the groups rather than a single pooled value. For instance, the confidence interval method presented in Section 7.3 for two groups does not assume equal standard deviations.

Example 12.3

Comparing Mean Ideology of Democrats and Republicans For Table 12.1, let's compare population mean ideology of Democrats (group 1) and Republicans (group 3). From Table 12.1 (page 352), $\bar{y}_1 = 3.48$ for $n_1 = 83$ Democrats and $\bar{y}_3 = 4.66$ for $n_3 = 38$ Republicans. From the regression results in Table 12.5 (page 355), the estimate of the population standard deviation is $s = \sqrt{1.72} = 1.31$, with $df = 275$. For a 95% confidence interval with $df = 275$, the t-score is $t_{.025} = 1.97$. The confidence interval for $\mu_3 - \mu_1$ is

$$(\bar{y}_3 - \bar{y}_1) \pm t_{.025}s\sqrt{\frac{1}{n_1} + \frac{1}{n_3}} = (4.66 - 3.48) \pm 1.97(1.31)\sqrt{\frac{1}{83} + \frac{1}{38}}$$
$$= 1.18 \pm 0.51, \quad \text{or} \quad (0.67, 1.68).$$

We infer that population mean ideology was between 0.67 and 1.68 units higher for Republicans than for Democrats. Since the interval contains only positive numbers, we conclude that $\mu_3 - \mu_1 > 0$; that is, μ_3 exceeds μ_1. On the average, Republicans were more conservative than Democrats, with difference between the means 0.67 to 1.68 categories on the seven-category scale. ■

ERROR RATES WITH LARGE NUMBERS OF CONFIDENCE INTERVALS

With g groups, we can compare $g(g-1)/2$ pairs of groups. When g is relatively large, the number of comparisons can be very large. Confidence intervals for some pairs of means may suggest they are different *even if all of the population means are equal*.

When $g = 10$, for example, there are $g(g - 1)/2 = 45$ pairs of means. Suppose we form a 95% confidence interval for the difference between each pair. The error probability of 0.05 applies for each comparison. For the 45 comparisons, the expected number of intervals that would not contain the true differences of means is $45(0.05) = 2.25$.

For 95% confidence intervals, the error probability of 0.05 is the probability that any particular confidence interval will not contain the true difference in population

[1] For $g = 2$ groups, $df = N - g = n_1 + n_2 - 2$; this interval then simplifies to the one in Section 7.5 (page 193) introduced for $\mu_2 - \mu_1$ assuming a common standard deviation.

means. When we form a large number of confidence intervals, the probability that *at least* one confidence interval will be in error is much larger than the error probability for any particular interval. The larger the number of groups to compare, the greater is the chance of at least one incorrect inference.

BONFERRONI MULTIPLE COMPARISONS OF MEANS

When we plan many comparisons, methods are available that control the probability that *all* intervals will contain the true differences. Such methods are called ***multiple comparison*** methods. They fix the probability that *all* intervals contain the true differences of population means *simultaneously*, rather than individually.

For example, with a multiple comparison method applied with $g = 10$ means and 95% confidence, the probability equals 0.95 that *all* 45 of the intervals will contain the pairwise differences $\mu_i - \mu_j$. Equivalently, the probability that *at least one* interval is in error equals 0.05. This probability is called the ***multiple comparison error rate***.

The ***Bonferroni multiple comparison*** method is simple and applies to a wide variety of situations. This method uses the same formula for a confidence interval introduced at the beginning of this section. However, it uses a more stringent confidence level for each interval, to ensure that the overall confidence level is sufficiently high.

To illustrate, suppose we would like a multiple comparison error rate of 0.10, that is, a probability of 0.90 that all confidence intervals are simultaneously correct. If we plan four comparisons of means, then the Bonferroni method uses error probability $0.10/4 = 0.025$ for each one. That is, it uses a 97.5% confidence level for each interval. This approach is somewhat conservative: It ensures that the actual overall error rate is *at most* 0.10 and that the overall confidence level is *at least* 0.90. The method is based on a probability inequality employed by the Italian probabilist Carlo Bonferroni in 1935. It states that the probability that at least one of a set of events occurs can be no greater than the sum of the separate probabilities of the events. For instance, if the probability of an error equals 0.025 for each of four confidence intervals, then the probability that at least one of the four intervals will be in error is no greater than $(0.025 + 0.025 + 0.025 + 0.025) = 0.10$.

Example 12.4

Bonferroni Intervals for Political Ideology Comparisons For the $g = 3$ political party IDs in Table 12.1, let's compare the mean political ideologies: μ_1 with μ_2, μ_1 with μ_3, and μ_2 with μ_3. We construct confidence intervals having overall confidence level at least 0.95. For a multiple comparison error rate of 0.05 with three comparisons, the Bonferroni method uses error probability $0.05/3 = 0.0167$ for each interval. These use the t-score with two-tail probability 0.0167, or single-tail probability 0.0083. For $df = 275$, $t_{0.0083} = 2.41$.

The interval for $\mu_3 - \mu_1$, the difference between the population mean ideology of Republicans and Democrats, is

$$(\bar{y}_2 - \bar{y}_1) \pm ts\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (4.66 - 3.48) \pm 2.41(1.31)\sqrt{\frac{1}{83} + \frac{1}{38}}$$

$$= 1.18 \pm 0.62, \quad \text{or} \quad (0.56, 1.79).$$

We construct the intervals for the other two pairs of means in a similar way. Table 12.6 displays them. The interval comparing Democrats and Independents contains 0. They are not significantly different. The intervals comparing Republicans to Democrats and to Independents do not contain 0. They show significant evidence of a difference between the population means for Republicans and the other two groups. ■

Multiple Regression with Quantitative and Categorical Predictors

Chapter 13

CHAPTER OUTLINE

Chapter 11 introduced the multiple regression model to analyze the relationship between a quantitative response variable and *quantitative* explanatory variables. Chapter 12 showed that multiple regression models can also handle *categorical* explanatory variables, by constructing dummy variables. In this chapter, we see that multiple regression can simultaneously handle quantitative and categorical explanatory variables.

In the last chapter, we learned that models with a single categorical explanatory variable focus on comparing the mean of y for several groups. The analysis of variance (ANOVA) F test relates to that model. In many applications, it is useful to compare means while controlling for other variables, some of which may be quantitative. For example, in comparing mean income for men and women in some profession, we might control for possibly differing levels of job experience between men and women. The quantitative control variable measuring job experience 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 R. A. Fisher, the brilliant British statistician.

Because effects may change after controlling for a variable, the results of analysis of covariance may 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 in the profession studied, the results of a comparison of mean income for men and women will depend on whether we control for experience.

In this chapter, we first show graphic representations of using both categorical and quantitative explanatory variables. In regression models, we again use dummy variables for qualitative explanatory variables. The models enable us to analyze effects of variables while controlling for both quantitative and categorical explanatory variables. For example, we can adjust sample means of y for different groups to reflect their predicted values after controlling for covariates.

The final section of the chapter introduces a more general model, called the *linear mixed model*, which can have both quantitative and categorical explanatory variables but also includes *random effects*. Whereas the ordinary regression model assumes that all observations are independent, the linear mixed model handles situations in which some observations are correlated. This type of model is useful for repeated-measures experiments, longitudinal studies, and for applications with clusters of observations such as families, since the observations are not all independent in such studies.

13.1 Models with Quantitative and Categorical Explanatory Variables

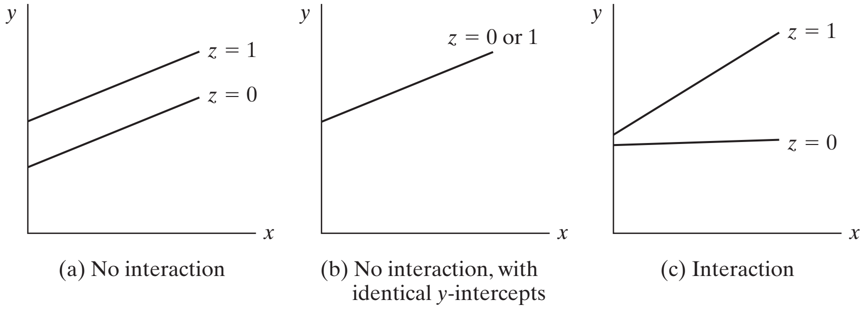
We introduce concepts using a single quantitative explanatory variable, denoted by x, and a single categorical factor, denoted by z. When the categorical variable has two categories, z is a dummy variable; when it has several categories, we use a set of dummy variables. The analysis of the effect of x 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 variable z refers to comparing the means of y for the groups defined by z, treating x as the control variable.

COMPARING REGRESSION LINES

Table 9.5 (page 268) introduced a data file on $y =$ selling price of homes. One quantitative explanatory variable is $x =$ size of home. One categorical variable 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.

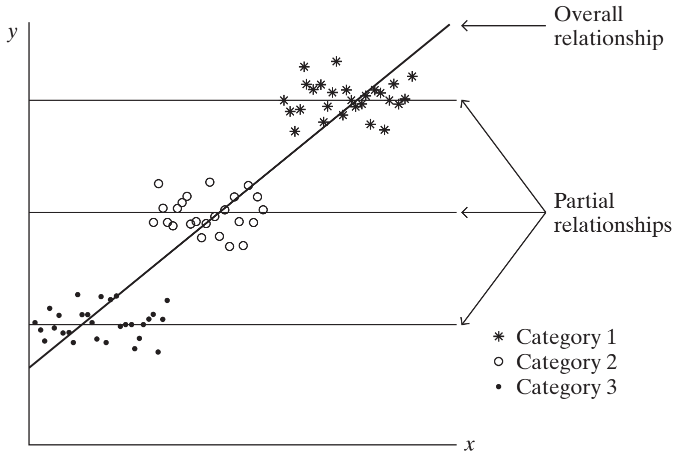
In this context, *no interaction* means that the true slope of the line relating expected selling price to the size of home is the same for new and older homes. Equality of slopes implies that the regression lines are parallel. See Figure 13.1a. When the y-intercepts are also equal, the regression lines 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 Quantitative Response and Quantitative Explanatory Variable, within Categories of a Categorical Variable with Two Categories



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

FIGURE 13.2: An Association between Two Quantitative Variables that Disappears after Controlling for a Categorical Variable



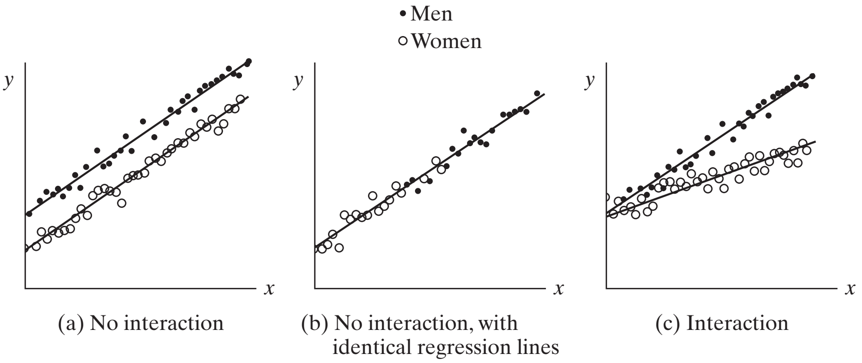
ignored. Within each category of z, however, the regression line relating y to x is horizontal. The overall positive trend is due to the tendency for the categories with high (low) scores on y to have high (low) scores on x also. Example 10.1 in Chapter 10 (page 291) presented an example of this type, with $y =$ math achievement test score and $x =$ height. The categorical variable was grade of school, with students coming from grades 2, 5, and 8.

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.

FIGURE 13.3: Three Scenarios for the Regression of $y =$ Income on $x =$ Number of Years of Experience and $z =$ Gender



(a) No interaction

(b) No interaction, with identical regression lines

(c) Interaction

The results of this analysis may differ considerably from a comparison of 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. Example 13.6 in this chapter shows an example of this type.

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 13.1: 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		

Note: The data are in the Income data file at the text website.

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.

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

	Black	Hispanic	White	Overall
Mean income	$\bar{y}_1 = 27.8$	$\bar{y}_2 = 31.0$	$\bar{y}_3 = 42.5$	$\bar{y} = 37.6$
Mean education	$\bar{x}_1 = 12.2$	$\bar{x}_2 = 11.6$	$\bar{x}_3 = 13.1$	$\bar{x} = 12.7$
Sample size	$n_1 = 16$	$n_2 = 14$	$n_3 = 50$	$N = 80$

As in Section 12.1, we represent a categorical factor in a regression model using dummy variables, one fewer than the number of categories. With three categories, 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 = education on the mean of y for each racial–ethnic group. For racial–ethnic status, one way to set up the dummy variables is

$z_1 =$ 1 if subject is black, $z_1 = 0$ otherwise;
$z_2 =$ 1 if subject is Hispanic, $z_2 = 0$ otherwise;
$z_1 = z_2 = 0$ if subject is white.

Table 13.3 shows some output 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

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

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

$$\hat{y} = -20.6 + 4.4x \quad \text{(Hispanics)};$$

$$\hat{y} = -15.7 + 4.4x \quad \text{(whites)}.$$

TABLE 13.3: Output for Fitting Model to Table 13.1 from the `Income` Data File on y = Income and Explanatory Variables Education and Racial–Ethnic Status, with Dummy Variables for Black and Hispanic Categories

```
                                                        95% Conf. Int.
Parameter     Coef.    Std. Error      t       Sig      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 is a scatterplot showing the prediction equations for the three groups. The lines are parallel, since they each have the same slope, 4.43. In each prediction equation, 4.43 is the coefficient of x, reflecting the increase for each group in the mean of y per one-year increase in education. The parallelism reflects the lack of interaction terms for this model. Since z_1 is a dummy variable for blacks, the coefficient -10.9 of z_1 represents the difference ($-\$10,900$) between the estimated annual mean income for blacks and for whites, controlling for education. The estimated mean income is \$10,900 lower for blacks than for whites, at each fixed level of education. Since z_2 is a dummy variable for Hispanics, the coefficient -4.9 of z_2 represents the difference ($-\$4900$) between the estimated mean income for Hispanics and whites, controlling for education. ■

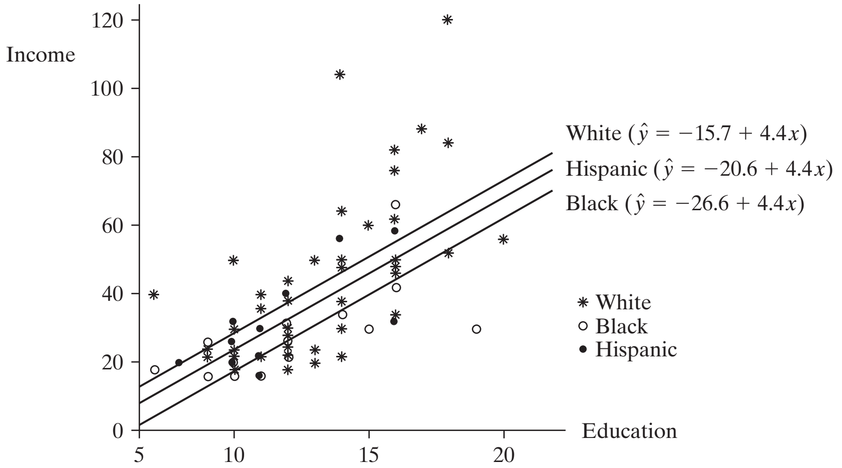


FIGURE 13.4: Plot of Prediction Equation for Model, Assuming No Interaction, with Quantitative and Categorical Explanatory Variables. Each line has the same slope, so the lines are parallel.

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. Some software (such as R and Stata) uses the first category instead of the final category as the baseline that does not have its own dummy variable. The coefficients of the dummy variables then estimate differences in means between each category and the first category, controlling for the other variables in the model.

Example 13.2

Regression of Income on Education and Racial–Ethnic Group, Permitting Interaction
A model that allows interaction between a quantitative explanatory variable x and a categorical factor z allows a different slope for the effect of x in each category of z. To allow interaction, as usual we take cross products of the explanatory variables. For Table 13.1, we take cross products $x \times z_1$ and $x \times z_2$ of the dummy variables z_1 and z_2 for blacks and Hispanics with the education explanatory variable.

Software provides the results shown in Table 13.4. The overall prediction equation is

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

TABLE 13.4: Output for Fitting Interaction Model to Table 13.1 from the `Income` Data File on Income, Education, and Racial–Ethnic Status

```
Parameter                 Coef.    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 prediction equation with both dummy variables equal to zero ($z_1 = z_2 = 0$) refers to the third 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), $z_1 = 1, z_2 = 0$, and

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

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

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

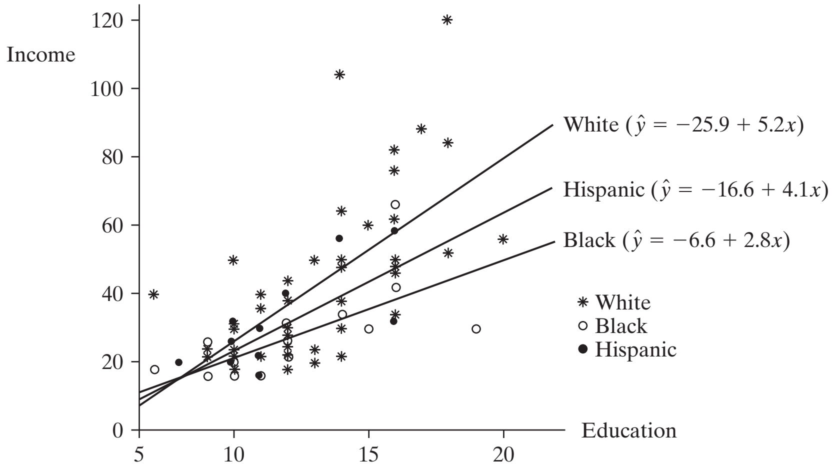
The coefficient 19.3 of z_1 describes the difference between the y-intercepts for blacks and whites. However, this is the difference *only* at $x = 0$, since the equations have different slopes. Since the 5.2 coefficient of x represents the slope for whites, the coefficient of ($x \times z_1$) (i.e., -2.4) represents the *difference in slopes* between blacks and whites. The two lines are parallel only when that coefficient equals 0. Similarly, for the second category, the coefficient of z_2 is the difference between the y-intercepts for Hispanics and whites, and the coefficient of ($x \times z_2$) is the difference between their slopes. Table 13.5 summarizes the interpretations of the estimated parameters in the model.

TABLE 13.5: Summary of Prediction Equation $\hat{y} = -25.9 + 5.2x + 19.3z_1 + 9.3z_2 - 2.4(x \times z_1) - 1.1(x \times z_2)$ Allowing Interaction, with $z_1 = 1$ for Blacks and $z_2 = 1$ for Hispanics

				Difference from White of	
Group	y-Intercept	Slope	Prediction Equation	y-Intercept	Slope
Black	$-25.9 + 19.3$	$5.2 - 2.4$	$(-25.9 + 19.3) + (5.2 - 2.4)x$	19.3	-2.4
Hispanic	$-25.9 + 9.3$	$5.2 - 1.1$	$(-25.9 + 9.3) + (5.2 - 1.1)x$	9.3	-1.1
White	-25.9	5.2	$-25.9 + 5.2x$	0	0

Figure 13.5 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.

FIGURE 13.5: Plot of Prediction Equations for Model with Interaction Terms



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 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 the education level x increases, the difference between the estimated mean incomes is larger. Figure 13.5 shows that the difference between the mean incomes of whites and blacks also gets larger at higher education levels. When a variable occurs in an interaction term, 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.

To summarize how much better the model permitting interaction fits, we can check the increase in R^2 or in the multiple correlation R. From the output for the no-interaction model (Table 13.3 on page 391), $R^2 = 0.462$. From the output for the interaction model (Table 13.4), $R^2 = 0.482$. The corresponding multiple correlation values are $\sqrt{0.462} = 0.680$ and $\sqrt{0.482} = 0.695$. Little is gained by fitting the more complex model, as R^2 and R do not increase much. ■

REGRESSION WITH MULTIPLE CATEGORICAL AND QUANTITATIVE PREDICTORS

The models generalize to add explanatory variables 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.

With several explanatory variables, the number of potential models is quite large when we consider the possible main effect and interaction terms. 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. Using inference, as described in the next section, helps us select a model.

13.2 Inference for Regression with Quantitative and Categorical Predictors

This section presents inference methods for models that contain both quantitative and categorical explanatory variables. As in other multivariable models, we first test whether the model needs interaction terms. We test hypotheses about model parameters using the F test comparing complete and reduced regression models, introduced in Section 11.5. For instance, the test of H_0: no interaction between two explanatory variables compares the complete model containing their cross-product interaction terms to the reduced model deleting them. This test has a small P-value if the addition of the interaction terms provides a significant improvement in the fit.

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: $\beta_3 = \beta_4 = 0$. The model under H_0 has a common slope β for all three lines relating $E(y)$ to x. Figure 13.6 depicts the hypotheses for this test.

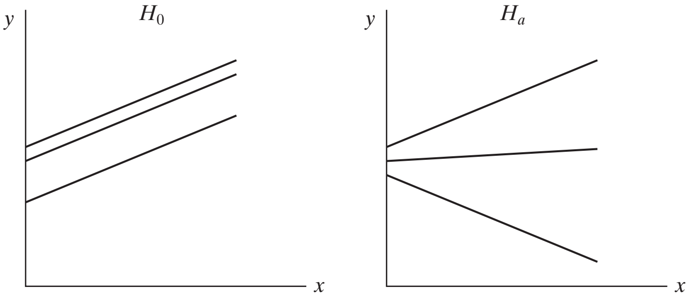


FIGURE 13.6: Graphical Representation of Null and Alternative Hypotheses in a Test of No Interaction, for a Categorical Factor with Three Categories and a Quantitative Explanatory Variable x

Table 13.6 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* (see page 368). They represent the variability explained after the other terms are already in the model.

TABLE 13.6: Software Output of Partial Sums of Squares Explained by Education, Racial–Ethnic Group, and Their Interaction, in the Model Permitting Interaction

```
                     Partial Sum          Mean
Source               of Squares   df     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
Residual (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 residual mean square. Table 13.6 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. ■

TESTS FOR INDIVIDUAL PARTIAL EFFECTS

Possibly the model can be simplified further, if either of the main effects is not significant. For the test of the main effect for the categorical factor, 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 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 racial–ethnic group. 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 13.7 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.

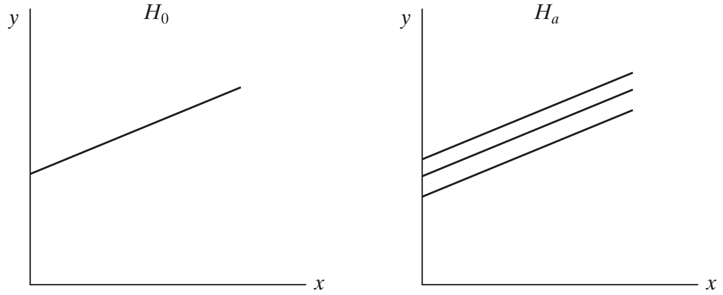


FIGURE 13.7: Graphical Representation of Null and Alternative Hypotheses in a Test of Equivalence of Regression Lines, when the Categorical Factor Represents Three Groups (Test Assumes No Interaction)

We can also test for the effect of the quantitative variable (x = education), by testing H_0: $\beta = 0$ in the model. The hypothesis states that the straight line relating x to the mean of y has slope 0 for each racial–ethnic group. Since H_0 specifies a value for a single parameter, we can perform the test using the t test.

Example 13.4

Testing Partial Effects of Racial–Ethnic Group and Education Table 13.7 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 is 730.29/239.00 = 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 13.7: Software Output of Inferences about Education and Racial–Ethnic Group, in the Model without Interaction for the `Income` Data File

```
                   Partial Sum          Mean
Source             of Squares     df    Square        F       Sig
Race                 1460.58       2      730.29     3.06     .053
Education           12245.23       1    12245.23    51.23     .000
Residual (Error)    18164.25      76      239.00
Total               33761.95      79
```

From Table 13.3 (page 391), the estimated slope for the effect of education on income of 4.432 has a standard error of 0.619. The test statistic is $t = 4.432/0.619 = 7.2$, 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 13.7. ■

Table 13.8 summarizes the hypotheses and R^2-values for the models. In bivariate models, education is a good predictor of income ($R^2 = 0.42$), considerably better than racial–ethnic group ($R^2 = 0.10$). Some further reduction in error results from using both explanatory variables, assuming no interaction, to predict income ($R^2 = 0.46$). A small and insignificant reduction in error occurs by allowing interaction ($R^2 = 0.48$).

TABLE 13.8: Summary of Comparisons of Four Models for Predicting Income (y) Using Education (x) and Racial–Ethnic Status (z)

Model:	$E(y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$ $+\beta_3(xz_1) + \beta_4(xz_2)$	$E(y) = \alpha + \beta x$ $+\beta_1 z_1 + \beta_2 z_2$	$E(y) = \alpha + \beta x$	$E(y) = \alpha$ $+\beta_1 z_1 + \beta_2 z_2$
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, control 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

13.3 Case Studies: Using Multiple Regression in Research

Multiple regression analysis is a common statistical tool in social research. Many studies start with a simple model containing an explanatory variable of primary focus, with the goal of studying its effect on the response and how that effect changes as other explanatory variables enter the model. Each new model adds potential confounding variables to try to help account for the bivariate effect of the primary explanatory variable on the response. The study often also adds potential mediating variables that could be responsible for the original association. Social scientists typically attempt to evaluate some causal dynamics by using a sequence of models, with primary interest in mediation processes and elimination of the possibility of spuriousness due to confounding variables.

The explanatory variables entered in the model often include both categorical variables and quantitative variables. Now that you've learned how regression can use both these types of explanatory variables, you have sufficient background to understand most regression analyses in social research. This section summarizes three such research studies[1] for which the conclusions are based on results of regression analyses.

Example 13.5

Regression for Modeling Adolescent Sexual Behavior A research study[2] about adolescent sexual behavior by Brian Soller and Dana Haynie used multiple regression with a response variable that is a composite measure of sexual risk-taking. This index incorporates information about inconsistent condom use, sexual intercourse without first discussing contraception or sexually transmitted infections, and sexual intercourse with more than one partner. The higher the composite measure, the greater

[1] Thanks to Prof. Alfred DeMaris for suggesting this section and two of these articles.

[2] Published in *Sociological Inquiry*, vol. 83 (2013), pp. 537–569; you may be able to access a pdf file of this article through your university's library at `http://onlinelibrary.wiley.com/doi/10.1111/soin.12019/abstract`.

the sexual risk-taking. The sample of 6255 adolescents, taken from a random sample of high schools in the United States, ranged from 7th to 12th graders. The explanatory variable of main interest was peer anticipation of college completion. If most of one's peers believed they would attend and complete college, does this tend to reduce a respondent's sexual risk-taking, controlling for relevant confounding variables?

Because of the very large sample size, multiple regression models can use many explanatory variables. The study measured each respondent's anticipation of college completion as a binary variable (1 = pretty likely or more, 0 = less than pretty likely). Peer anticipation of college completion was measured as the mean of the same binary variable for up to five male and five female friends. Other explanatory variables measured the consequences of pregnancy or romantic relationships. The study included control variables thought to be potential confounding variables, such as measures intended to capture individual and peer investment in scholastic achievement and measures of peer delinquency, impulsivity, family attachment, and religiosity. Other quantitative control variables were age, parental SES, and prior sexual risk-taking. Qualitative control variables included race, family structure, and whether the respondent had taken an abstinence pledge (i.e., to remain a virgin until married).

The authors fitted four regression models in which peer anticipation of college completion was an explanatory variable for y = sexual risk-taking. Model 1 analyzed its effect, adjusting for the control variables and the respondent's own anticipation of college completion. The estimated effect of peer anticipation was $\hat{\beta} = -0.13$ ($SE =$ 0.05). So, the estimated mean of sexual risk-taking was 0.13 lower for those whose peers all felt pretty likely or more to attend college (and so had a variable value of 1) than for those whose peers all felt less than pretty likely to attend college (and so had a variable value of 0), controlling for other variables.

The authors then investigated whether other variables mediated that association. Model 2 added the consequences of pregnancy variables. The estimated effect of peer anticipation then weakened a bit ($\hat{\beta} = -0.11, SE = 0.05$), suggesting a slight mediating effect. Model 3 removed the pregnancy variables and added relationship measures. The estimated effect of peer anticipation was then similar ($\hat{\beta} = -0.12$, $SE = 0.05$). Model 4 (the full model) added both the pregnancy variables and the relationship variables. The effect again weakened only slightly ($\hat{\beta} = -0.10, SE = 0.05$).

Table 13.9 shows some of the explanatory variables and their means and standard deviations and estimated effects in Model 4. Here are some results worth noting:

- The respondents' own anticipation of college completion was not significantly associated with sexual risk-taking, controlling for the other variables in the model.

- The authors concluded, "Results from our study underscore the importance of peers in shaping adolescent sexual behavior." For all the models, however, the effect of peer anticipation of college completion on sexual risk-taking is only about −0.1. This effect seems quite weak, because its magnitude is a small fraction of the standard deviation of y = sexual risk-taking, which was reported to equal 0.81. For example, Table 13.9 reports peer anticipation of college completion to have a standard deviation of 0.25, so the estimated standardized regression coefficient for this variable for the full model is only $(-0.10)(0.25)/0.81 = -0.03$. Although the effect was statistically significant at the 0.05 level for all these models and confirmed the authors' theoretical prediction that it would be negative, could this be a case of statistical significance but not practical significance, reflecting the very large n? In practice, often social scientists investigate whether a theoretical effect is truly there, even if it is quite small, to confirm a research hypothesis. With human behavior and imperfect measurement of constructs, large observed effect sizes are not common.

- The article does not mention anything about checking for interactions. Could the effect of peer anticipation of college completion depend on the respondent's level of anticipation of college completion, or on sex or some other variable?

■

TABLE 13.9: Explanatory Variables and Effects in Multiple Regression Model Predicting Adolescent Sexual Risk-Taking

Variable	Mean (Std. Dev.)	$\hat{\beta}$ (*SE*)
Control variables		
Age	16.63 (0.99)	0.05 (0.01)
Sex (male=1, female=0)	0.54	−0.10 (0.03)
SES	0.00 (0.79)	0.01 (0.02)
Religiosity	−0.02 (0.82)	−0.04 (0.02)
Abstinence pledge	0.12	−0.06 (0.03)
GPA	2.75 (0.79)	−0.04 (0.02)
Anticipation of college completion	0.76 (0.43)	0.02 (0.04)
Peer variables		
Peer anticipation of college completion	0.76 (0.25)	−0.10 (0.05)
Peer delinquency	0.93 (0.13)	0.25 (0.08)
Peer GPA	2.77 (0.50)	−0.05 (0.03)

Note: Standard deviations were not reported for sex and abstinence pledge.

Example 13.6

Regression for Modeling the Earnings Gender Gap For many professions, men and women have similar mean salaries when first employed after college graduation, but over time men's salaries tend to grow more quickly than women's salaries. The difference between the mean salaries increases with time. What is responsible for this? A research study[3] by Marianne Bertrand, Claudia Goldin, and Lawrence Katz addressed this using multiple regression models for a sample of 1856 men and 629 women MBA graduates from the University of Chicago.

Their use of multiple regression started with a dummy variable for gender, to enable comparing means, and then successively added various explanatory variables that could potentially explain differences in mean incomes. Some, such as undergraduate GPA and verbal and quantitative GMAT scores, were quantitative. Some, such as race, reasons for choosing the job, and whether the undergraduate institution was a "top 10" institution, were categorical. Some were potentially quantitative but were measured with ordered categories and represented by dummy variables, such as weekly hours worked (<20, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, ≥ 100) and number of years since receiving the MBA (0, 1, 3, 6, 9, ≥ 10). Fitting such models enabled the authors to study how the difference in mean income between men and women changed over time, controlling for relevant variables.

Upon starting a job after receiving the MBA, the mean salary was \$130,000 for men and \$115,000 for women. After nine years on the job, the mean was about \$400,000 for men and about \$250,000 for women. For the overall pooled sample, the mean was 36% higher for men. The initial regression model had a dummy variable for gender, dummy variables for five of the six categories for number of years since receiving the MBA, and five interaction terms to allow the difference between men and women to vary by time. This regression model had $R^2 = 0.15$.

[3] Published in *American Economic Journal: Applied Economics*, vol. 2 (2010), pp. 228–255; you may be able to access a pdf file of this article through your university's library at `https://www.aeaweb.org/articles.php?doi=10.1257/app.2.3.228`.

When the model adds the number of weekly hours worked as an explanatory variable, R^2 increases to 0.26. Controlling for this explanatory variable, the mean salary for men is now 19% higher than the mean salary for women. When the model next includes pre-MBA characteristics such as race and undergraduate GPA and GMAT scores, MBA GPA, and the fraction of MBA classes that were in finance, $R^2 = 0.40$, and controlling for all the explanatory variables, the mean salary for men is 10% higher than the mean salary for women. Finally, when the model adds a dummy variable for the presence of any post-MBA career interruption (such as caring for a baby) and variables dealing with reasons for choosing the job, the job function, and the employer type, $R^2 = 0.54$. At this stage, controlling for all explanatory variables in the model, the mean salary for men is only 4% higher than the mean salary for women, and the difference is not statistically significant.

- The authors concluded that three factors account for most of the gender gap in earnings: a modest male advantage in training prior to the MBA; greater weekly hours working for men, the difference increasing with years since MBA; greater career interruptions for women combined with large earnings losses associated with any career interruption. They noted that the greater career interruptions and shorter work hours for women than men were largely associated with motherhood.
- The authors used the logarithm of salary as the response variable but failed to clearly explain how to interpret the estimated regression coefficients. Data analysts sometimes use the log transform for variables such as income that have distributions very highly skewed to the right, as it "pulls in" values that are far out in the right tail and makes the distribution less skewed. Section 14.4 in the next chapter (page 435) presents an alternative model for such skewed response data, and Section 14.6 shows another setting in which the logarithm transform is effective.

■

Example 13.7 **Modeling the Consequences of Stigma for Self-Esteem of the Mentally Ill** A research study[4] by Bruce Link, Elmer Struening, Sheree Neese-Todd, Sara Asmussen, and Jo Phelan analyzed whether stigma affects the self-esteem of people who have serious mental illnesses, using a sample of 70 members of a clubhouse program for people with mental illness. To measure self-esteem, the study asked participants whether they strongly agreed, agreed, disagreed, or strongly disagreed with 10 statements such as "At times, you think you are no good at all." Each item had scores (1, 2, 3, 4) with a high score reflecting high self-esteem. The overall self-esteem measure was the mean of these 10 scores. The study measured self-esteem initially, after six months of an intervention designed to facilitate coping with stigma, and after 24 months. The initial measure had a mean of 2.7 and standard deviation of 0.5.

The primary explanatory variables were two quantitative stigma measures: One (perceived devaluation discrimination) measured the extent to which a person believes that other people devalue someone who has a mental illness. The other (stigma withdrawal) quantified the extent to which participants endorse withdrawal as a way to avoid rejection. Each of these was also scaled from 1 to 4, and had means of 2.76 and 2.82 and standard deviations of 0.50 and 0.42. Control variables included sex (male = 1, female = 0), diagnosis (schizophrenia and other nonaffective psychotic disorders = 1, other diagnoses = 0), and a quantitative assessment of depression (which ranged from 0 to 42).

[4] Published in *Psychiatric Service*, vol. 52 (2001), pp. 1621–1626; you can access a pdf file of this article at `http://ps.psychiatryonline.org/doi/pdf/10.1176/appi.ps.52.12.1621`.

Table 13.10 shows results of four regression models fitted to the self-esteem response after six months. The first model uses the control variables as explanatory variables and the initial self-esteem as a covariate. Models 2 and 3 add each stigma variable separately, and each shows a significant effect. Model 4 adds them together to determine their combined effect, which increased R^2 from 0.43 to 0.55. The stigma effects were negative in Models 2–4, although the effect of stigma withdrawal was weaker and not statistically significant in Model 4.

TABLE 13.10: Regression Analyses for Self-Esteem at Six Months

	Model 1	Model 2	Model 3	Model 4
Variable	Coeff. (*se*)	Coeff. (*se*)	Coeff. (*se*)	Coeff. (*se*)
Sex	−0.133 (0.088)	−0.165 (0.081)	−0.121 (0.084)	−0.152 (0.080)
Diagnosis	−0.098 (0.084)	−0.144 (0.077)	−0.081 (0.080)	−0.101 (0.076)
Self-esteem initial	0.352 (0.113)	0.343 (0.103)	0.338 (0.108)	0.337 (0.102)
Depression	−0.018 (0.007)	−0.010 (0.007)	−0.013 (0.007)	−0.008 (0.006)
Stigma devaluation		−0.321 (0.085)		−0.261 (0.091)
Stigma withdrawal			−0.302 (0.104)	−0.182 (0.107)
R^2	0.43	0.53	0.49	0.55

- The authors stated, "We also tested for interactions between the stigma variables and age, sex, diagnosis, and depressive symptoms. Only one interaction was significant, and, given that we tested 16, this one may have occurred by chance." So, they did not report results of models with interaction terms.
- The authors noted that an unmeasured confounding variable could potentially account for the association between stigma and self-esteem. However, they argued that the stigma measures strongly predicted self-esteem, and thus any unmeasured confounder would need to have very strong associations with both the stigma measures and self-esteem in order to eliminate the associations.
- How would you assess whether the stigma measures truly did strongly predict self-esteem, controlling for the other explanatory variables?

■

13.4 Adjusted Means*

We have seen that categorical explanatory variables often refer to groups to be compared. This section shows how to estimate means on y for the groups, while controlling for the other variables in the model.

ADJUSTING RESPONSE MEANS, CONTROLLING FOR OTHER VARIABLES

To estimate the means of y for the groups while taking into account the groups' differing means on the other explanatory variables, we can report the values expected for the means if the groups all had the same means on those other variables. These values, which adjust for the groups' differing distributions on the other variables, are *adjusted means* (also called *least squares means*).

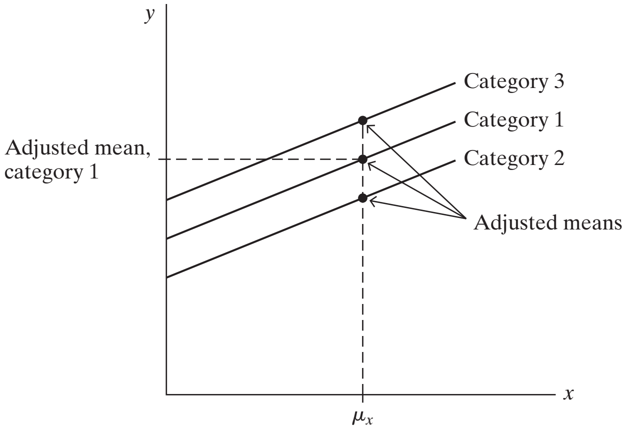
Adjusted Mean

> The ***adjusted mean*** of y for a particular group is the regression function for that group evaluated at the overall means of the explanatory variable values for all the groups. It represents the expected value for y at the means of the explanatory variables for the combined population.

Adjusted means are mainly relevant for models without interaction terms among the explanatory variables, so differences among them are the same at all values of those variables.

Figure 13.8 illustrates the population adjusted means for three groups when the model has a single covariate x. 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. 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 for the group's mean of y if the mean of x for the group had equaled the overall mean of x. We denote the sample adjusted mean for group i by $\bar{y}'_i$.

FIGURE 13.8: Population Adjusted Means with a Covariate x, when a Categorical Explanatory Factor Has Three Categories



Example 13.8

Adjusted Mean Incomes, Controlling for Education From Table 13.3 (page 391) for the example regressing income on education and racial–ethnic status, the prediction equation for the model is

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

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

TABLE 13.11: Prediction Equations, Sample Unadjusted Mean Incomes, and Adjusted Means (Controlling for x = Education)

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 (page 390), the mean education for the combined sample of 80 observations is $\bar{x} = 12.7$. Using the three prediction equations, the sample adjusted means for blacks, Hispanics, and whites are

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

$$\bar{y}'_2 = -20.60 + 4.43(12.7) = 35.6,$$

$$\bar{y}'_3 = -15.66 + 4.43(12.7) = 40.6.$$

■

COMPARING ADJUSTED MEANS, AND GRAPHICAL INTERPRETATION

The coefficients of the dummy variables in the model refer to differences between adjusted means. To illustrate, the estimated difference between adjusted mean incomes of blacks and whites is $\bar{y}'_1 - \bar{y}'_3 = 29.7 - 40.6 = -10.9$ (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 is $\bar{y}'_2 - \bar{y}'_3 = -4.9$, which is the coefficient of z_2. Figure 13.9 depicts the sample adjusted means. The vertical distances between the lines represent the differences between these adjusted means.

FIGURE 13.9: Sample Adjusted Means on Income, Controlling for Education, for Three Racial–Ethnic Groups

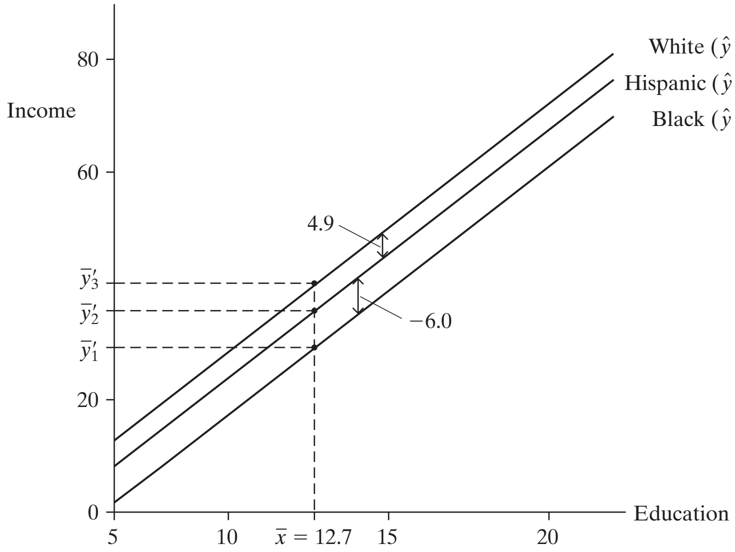
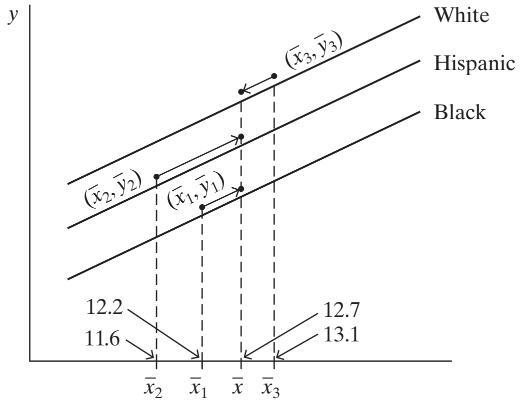


Figure 13.10 depicts the relationship between the adjusted and unadjusted means. The prediction equation predicts a value of $\bar{y}_i$ at the x-value of $x = \bar{x}_i$ for group i. In particular, the prediction line for the group i passes through the point with coordinates $(\bar{x}_i, \bar{y}_i)$. In other words, the *unadjusted mean* $\bar{y}_i$ is the value of the prediction equation for that group evaluated at the x-value of $\bar{x}_i$, the mean of the x-values *for that group alone* [see points such as $(\bar{x}_1, \bar{y}_1)$ in this figure].

FIGURE 13.10: Adjustment Process for Income by Racial–Ethnic Group, Controlling for Education. The ordinary mean for a group is the predicted value at the mean of x for that group alone, whereas the adjusted mean is the predicted value at the mean $\bar{x}$ for all the data.



The *adjusted mean* $\bar{y}'_i$ for group i 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 $(\bar{x}, \bar{y}_i')$, as Figure 13.10 shows for each of the three groups.

The adjustment process moves 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 $\bar{x} = 12.7$).

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 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*. We can form t confidence intervals using these estimates and their standard errors.

Example 13.9 **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 the residual df for the model) is 2.45.

Table 13.3 (page 391) showed the racial–ethnic effects from the model fit,

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

The estimated difference between adjusted mean incomes of Hispanics and whites is the coefficient -4.934 of the dummy variable z_2 for Hispanics in the prediction equation. This coefficient has a standard error of 4.763, so the Bonferroni confidence interval equals

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

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. 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 has 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. ■

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	$\bar{y}_1' - \bar{y}_3' = -10.9$	(−21.8, 0.1)
Hispanics, whites	$\bar{y}_2' - \bar{y}_3' = -4.9$	(−16.6, 6.7)
Blacks, Hispanics	$\bar{y}_1' - \bar{y}_2' = -5.9$	(−19.8, 8.0)

A CAUTION ABOUT HYPOTHETICAL ADJUSTMENT

Adjusted means can be 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 (page 389). 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 13.11 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.

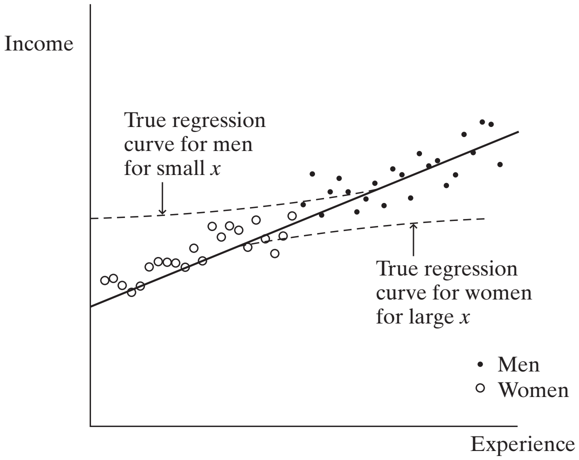


FIGURE 13.11: A Situation in Which Adjusted Means Are Misleading, Comparing Mean Incomes for Men and Women while Controlling for Experience