

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.

To predict $Y =$ college GPA, for example, it is sensible to use several predictors in the same model. Possibilities include $X_1 =$ high school GPA, $X_2 =$ math college entrance exam score, $X_3 =$ verbal college entrance exam score, and $X_4 =$ rating by high school guidance counselor. This chapter presents models for the relationship between a response variable Y and a collection of explanatory variables.

A multivariable model provides better predictions of Y than does a model with a single explanatory variable. Such a 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. Thus, this model provides information not available with simple models that analyze only two variables at a time.

Sections 11.1 and 11.2 extend the regression model to a ***multiple regression model*** that can have multiple explanatory variables. Section 11.3 defines correlation and r -squared measures that describe association between Y and a set of explanatory variables. Section 11.4 presents inference procedures for multiple regression. Section 11.5 shows how to allow *statistical interaction* in the model, and Section 11.6 presents a test of whether a complex model provides a better fit than a simpler model. The final two sections introduce 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

Suppose there are two explanatory variables, denoted by X_1 and X_2 . As in earlier chapters, we use lower-case letters to denote observations or particular values of the variables. 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 . When there are additional explanatory variables, each has a βx term, for example $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ with four predictors.

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

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

((Fig. 11.1 in 3e))

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 $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$ traces a plane (a flat surface) cutting through three-dimensional space, as Figure 11.1 portrays.

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 variable.

Example 11.1 Do Higher Levels of Education Cause Higher Crime Rates?

Exercise 39 in Chapter 9 contains recent data on several variables for the 67 counties in the state of Florida. For each county, let 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.

The relation between crime rate and both predictors 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

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

((Fig. 11.2 in 3e))

($\bar{x}_1 = 70$) 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.

((Fig. 11.3 in 3e))

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 original one. 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 on 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).

((Fig. 11.4 in 3e))

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* predictors, the beta coefficient of a predictor describes the change in the mean of Y for a one-unit increase in that predictor, controlling for the other variables in the model. The parameter α represents the mean of Y when each explanatory variable equals 0.

The parameters β_1, β_2, \dots 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.

This multiple regression model assumes that the slope of the partial relationship between Y and each predictor is identical for *all* combinations of values of the other explanatory variables. This means that the model is appropriate when there is *no statistical interaction*, in the sense of Section 10.3. 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 11.5 will show this model.

A partial slope in a multiple regression model usually differs from the slope in the bivariate model for that predictor, but it need not. With two predictors, 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 .

Prediction Equation and Residuals

Corresponding to the multiple regression equation, software finds a prediction equation by estimating the model parameters using sample data. For simplicity of notation, so far we've used just two predictors. In general, let k denote the number of predictors.

Notation for Prediction Equation

The prediction equation that estimates the multiple regression equation $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$ is denoted by $\hat{y} = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$.

For multiple regression, it is almost imperative to use computer software to find the prediction equation. The calculation formulas are complex and are not shown in this text.

We get the predicted value of Y for a subject by substituting 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 predictor.

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 form $\hat{Y} = a + b_1x_1 + \cdots + b_kx_k$.

11.2 Example with Multiple Regression Computer Output

We illustrate the methods of this chapter with the data introduced in the following example:

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 measure¹ 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 on the three variables for a random sample of 40 adults in the county. [These data are based on a larger survey. The authors thank Dr. Charles Holzer for permission to use the study as the basis of this example.] Table 11.2 summarizes the sample means and standard deviations of the three variables.

Table 11.1: Scores on 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: Estimated 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

¹Developed by E. Paykel et al., *Arch. Gen. Psychiatry*, vol. 75, 1971, pp. 340–347.

□

Scatterplot Matrix for Bivariate Relationships

Plots of the data provide an informal check of whether the relationships are linear. Most software can construct scatterplots on a single diagram for each pair of the variables. 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 one 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 has a mild positive effect and SES has a mild negative effect on mental impairment.

Partial Plots for Partial Relationships

The multiple regression model states that each predictor has a linear effect with common slope, controlling for the other predictors. To check this, we could use software to 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. A more informative single picture is provided by the *partial regression 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.

For example, here's how to find the partial regression plot for the effect of x_1 when the multiple regression model also has explanatory variables x_2 and x_3 . Find the residuals from the model using x_2 and x_3 to predict Y . Also find the residuals from the model using x_2 and x_3 to predict x_1 . Then plot 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 (from SPSS) for $Y =$ mental impairment and $x_1 =$ life events, controlling for $x_2 =$ SES. It plots the residuals on the y -axis from the model $E(Y) = \alpha + \beta x_2$ against the residuals on the x -axis from the model $E(X_1) = \alpha + \beta x_2$. 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.

11.2. EXAMPLE WITH MULTIPLE REGRESSION COMPUTER OUTPUT 445

Figure 11.7 shows the partial regression plot for SES. It shows that its partial effect is also approximately linear but is negative. It is simple to obtain partial regression plots with standard software such as SPSS. (See the appendix.)

Sample Computer Printouts

Tables 11.3 and 11.4 are SPSS printouts of the coefficients table for the bivariate relationships between mental impairment and the separate explanatory variables. The estimated regression coefficients fall in the column labelled 'B'. The prediction equations are

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

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 greater the SES level, the lower the mental impairment tends to be. The correlations between mental impairment and the explanatory variables are modest, 0.372 for life events and -0.399 for SES (listed by SPSS as 'Standardized coefficients'; the 'beta' label is misleading and refers to the alternate term *beta weights* for standardized regression coefficients).

Table 11.3: Bivariate Regression Analysis for $Y = \text{Mental Impairment (IMPAIR)}$ and $x_1 = \text{Life Events (LIFE)}$

Model		Coefficients(a)				
		Unstandardized		Standardized		
		Coefficients		Coefficients		
	B	Std. Error	Beta	t	Sig.	
1	(Constant)	23.309	1.807	12.901	.000	
	LIFE	.090	.036	.372	2.472	.018

a. Dependent Variable: IMPAIR

Table 11.5 shows part of a SPSS printout for the multiple regression model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$. The prediction equation is

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

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 1-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,

Table 11.4: Bivariate Regression Analysis for $Y =$ Mental Impairment and $x_2 =$ Socioeconomic Status (SES)

Model		Coefficients(a)				
		Unstandardized		Standardized		
		Coefficients		Coefficients		
	B	Std. Error	Beta	t	Sig.	
1	(Constant)	32.172	1.988		16.186	.000
	SES	-.086	.032	-.399	-2.679	.011

a Dependent Variable: IMPAIR

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.

Table 11.5: Fit of Multiple Regression Model for $Y =$ Mental Impairment, $x_1 =$ Life Events (LIFE), and $x_2 =$ Socioeconomic Status (SES)

	Unstandardized		Standardized		
	Coefficients		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

Dependent Variable: IMPAIR

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 predictor does little to alter the effect of the other

one. This suggests that these predictors 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 measure the joint association of the explanatory variables with the response variable, and shows how to interpret the R^2 value listed for the multiple regression model.

Table 11.6: Summary of Regression Models for Mental Impairment

Effect	Predictors in Regression Model		
	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
(n)	(40)	(40)	(40)

11.3 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

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 **multiple correlation** for a regression model 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 example, above we saw that the first subject in the sample had $y = 17$ and $\hat{y} = 24.8$. 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 the y - and \hat{y} -values is the multiple correlation. It is denoted by R .

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 a predictor x , which falls between -1 and $+1$. The larger the multiple correlation R , the better the predictions of y by the set of explanatory variables.

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, \dots, x_k): The best predictor is then the sample mean, \bar{y} .

Rule 2 (Predict Y using x_1, \dots, x_k): The best predictor is the prediction equation $\hat{y} = a + b_1x_1 + b_2x_2 + \dots + b_kx_k$.

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 $\text{TSS} = \sum(y - \bar{y})^2$, called the *total sum of squares*. For rule 2, it is $\text{SSE} = \sum(y - \hat{y})^2$, the sum of squared errors using the prediction equation, called the *residual sum of squares*.

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

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 it is the square of the multiple correlation. 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, introduced in Example 11.2, the prediction equation is $\hat{y} = 28.23 + 0.103x_1 - 0.097x_2$. Table 11.5 showed some output for this model. Software also reports ANOVA tables with sums of squares and R and R^2 tables. Table 11.7 shows some SPSS output.

Table 11.7: ANOVA Table and Model Summary for Regression of Mental Impairment (IMPAIR) on Life Events (LIFE) and Socioeconomic Status (SES)

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,

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

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.

SPSS reports R and R^2 in a separate ‘Model Summary’ table, as Table 11.7 shows. Most software also reports an adjusted version of R^2 that is a less biased estimate of the population value. Exercise 63 defines this measure, and Table 11.7 reports its value of 0.303.

□

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, \dots, x_k) collectively predict y .
- $R^2 = 1$ only when all the residuals are 0, that is, when all $y = \hat{y}$, so that $\text{SSE} = 0$. In that case, the prediction equation passes through all the data points.
- $R^2 = 0$ when the predictions do not vary as any of the x -values vary. In that case, $b_1 = b_2 = \dots = b_k = 0$, and \hat{y} is identical to \bar{y} , since the explanatory variables do not add any predictive power. When this happens, the correlation between y and each explanatory variable equals 0.
- 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 $r_{YX_1}^2$ when Y as a linear function of x_1 , $r_{YX_2}^2$ when Y as a linear function of x_2 , and so forth.

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, the multiple correlation for the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ is at least as large as the multiple correlation for the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$.

The numerator of R^2 , $\text{TSS} - \text{SSE}$, 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 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 there are 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 "house selling price" data set at the text website (introduced in Example 9.10), 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 predictors 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 predictors, shows a common occurrence. The size of the set for a predictor in this figure represents the size of its r^2 -value in predicting Y . The amount a set for a predictor overlaps with the set for another predictor represents its association with that predictor. The part of the set for a predictor that does not overlap with other sets represents the part of the variability in Y explained uniquely by that predictor. In Figure 11.8, all three predictors 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 .

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 if we want to include certain variables in the model for theoretical reasons.

In practice, 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 (for example, at least about 40 for 4 explanatory variables).

11.4 Inferences for Multiple Regression Coefficients

The multiple regression function

$$E(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

describes the relationship between the explanatory variables and the mean of the response variable. For particular values of the explanatory variables, $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$ represents the mean of Y for the population having those values.

To make inferences about the parameters, 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, \dots, x_k .
- The standard deviation, σ , of the conditional distribution of responses on Y is the same at each combination of values of x_1, \dots, x_k .
- 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 Section 14.2.

Two types of significance tests are used in multiple regression. 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

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0.$$

This states that the mean of Y does not depend on the values of x_1, \dots, x_k . Under the inference assumptions, this states that Y is statistically independent of all k explanatory variables.

The alternative hypothesis is

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

This states that *at least one* explanatory variable is related to Y , controlling for the others. The test judges whether using x_1, \dots, x_k together to predict y , with the prediction equation $\hat{y} = a + b_1 x_1 + \cdots + b_k x_k$, is better than using \bar{y} .

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

$$H_0 : \text{Population multiple correlation} = 0 \quad H_a : \text{Population multiple correlation} > 0.$$

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 k predictors, the test statistic equals

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

The sampling distribution of this statistic is called the ***F distribution***. We next study this distribution and its properties.

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.

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

$df_1 = k$, the number of explanatory variables in the model.

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

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

The mean of the F distribution is approximately equal to 1.² 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.

Table D at the end of the text lists the F scores having P -values of 0.05, 0.01, and 0.001, for various combinations of df_1 and df_2 . This table allows us to determine whether $P > 0.05$, $0.01 < P < 0.05$, $0.001 < P < 0.01$, or $P < 0.001$. Software for regression reports the actual P -value.

Example 11.4 *F* Test for Mental Health Impairment Data

In Example 11.2, we used multiple regression for $n = 40$ observations on $Y =$ mental impairment, with $k = 2$ explanatory variables, life events and SES. The null hypothesis that mental impairment is statistically independent of life events and SES is H_0 : $\beta_1 = \beta_2 = 0$.

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

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

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

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

From Table D, when $df_1 = 2$ and $df_2 = 37$, the F -value with right-tail probability of 0.001 falls between 8.77 and 8.25. Since the observed F test statistic of 9.5 falls above these two, it is farther out in the tail and has smaller tail probability than 0.001. Thus, the P -value is $P < 0.001$. Part of the SPSS printout 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		

in which we see the F statistic. The P -value, which rounded to three decimal places is $P = 0.000$, appears under the heading ‘Sig’ in the ANOVA table.

This extremely small P -value provides strong evidence against H_0 . It suggests that at least one of the explanatory variables is related to mental impairment. Equivalently, we can conclude that the population multiple correlation and R -squared are positive. So, we obtain significantly better predictions of y using the multiple regression equation than by using \bar{y} .

□

Normally, unless the sample size 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

Suppose the P -value is small for the F test that all the regression coefficients equal 0. 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, the most important being randomization and that the regression function describes well how the mean of Y depends on the explanatory variables.

Consider an arbitrary explanatory variable x_i , with coefficient β_i in the multiple regression 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 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 $df = n - (k + 1)$. The df value 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 $df = n - (k + 1)$. 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 Separate 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 consider 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

This tells us that the point estimate of β_1 is $b_1 = 0.103$ and has standard error $se = 0.032$. The test statistic equals

$$t = \frac{b_1}{se} = \frac{0.103}{0.032} = 3.2.$$

This appears under the heading 't' in the table in the row for the variable LIFE. The statistic has $df = n - (k + 1) = 40 - 3 = 37$. The P -value appears under 'Sig' in the row for LIFE. It is 0.003, the probability that the t statistic exceeds 3.2 in absolute value. There is strong evidence that mental impairment is related to 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 equals

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

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. 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 this interval contains only positive numbers, the relationship between mental impairment and life events is positive, controlling for SES. It may be simpler to interpret the interval (0.04, 0.17) by noting that an increase of 100 units in life events corresponds to anywhere from a $100(0.04) = 4$ to a $100(0.17) = 17$ unit increase in mean mental impairment. The interval is relatively wide, because of the small sample size.

□

How is the t test for a partial regression coefficient different from the t test of $H_0: \beta = 0$ for the bivariate model, $E(Y) = \alpha + \beta x$, studied 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 there is multicollinearity, that is, a lot of 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 β s. Any particular variable may explain uniquely little of the variation in Y , even though together the variables explain a lot of the 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

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

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 predictors. (When a model has only $k = 1$ predictor, df simplifies to $n - 2$, the term in the s formula of Section 9.3.)

Part of the SPSS printout 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		

containing the sums of squares for the multiple regression model with the mental impairment data. We see that $\text{SSE} = 768.2$. Since $n = 40$ for $k = 2$ predictors, we have $df = n - (k + 1) = 40 - 3 = 37$ 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 has the R and R^2 values (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 called the **mean square error**, often abbreviated by MSE. 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 the above table. Some software (such as SAS) better labels the conditional standard deviation estimate s as ‘Root MSE,’ because it is the square root of the mean square error.

The F Statistic Is a Ratio of Mean Squares*

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

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

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

The regression mean square equals the regression sum of squares divided by its degrees of freedom. The df equals k , the number of explanatory variables in the model, which is df_1 for the F test. On the printout shown above, 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 seen that the F distribution is used 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 $df_2 = n - (k + 1)$.

To illustrate, in Example 11.5, for $H_0: \beta_1 = 0$ and $H_a: \beta_1 \neq 0$, the test statistic was $t = 3.18$ 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.

11.5 Interaction between Predictors in their Effects

The multiple regression equation

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

assumes that the partial relationship between Y and each x_i is linear and that the slope β_i of that relationship 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 illustrated.

This model is sometimes too simple to be adequate. Often, there is *interaction*, with the relationship between two variables changing according to the value of a third variable. Section 10.3 introduced this concept.

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, suppose the relationship between x_1 and the mean of Y is $E(Y) = 2 + 5x_1$ when $x_2 = 0$, it is $E(Y) = 4 + 15x_1$ when $x_2 = 50$, and it is $E(Y) = 6 + 25x_1$ when $x_2 = 100$. The slope for the partial effect of x_1 changes markedly as the fixed value for x_2 changes. There is then interaction between x_1 and x_2 in their effects on Y .

Cross-product Terms

A common approach for allowing interaction 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

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

So, for fixed x_2 , the mean of Y changes linearly as a function of x_1 . The slope of the relationship is $\beta' = (\beta_1 + \beta_3 x_2)$. This depends on the value of x_2 . As x_2 changes, the

11.5. INTERACTION BETWEEN PREDICTORS IN THEIR EFFECTS 459

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 depends on the value of x_2 .

Note that now we can interpret β_1 as 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_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 Interaction Model for 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, $x_1 = 46$, $x_2 = 84$, so $x_3 = 46(84) = 3864$. Software makes it easy to create this variable without doing the calculations yourself. Table 11.8 shows part of the printout for the interaction model. The prediction equation is

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

Table 11.8: Interaction Model for $Y =$ Mental Impairment, $X_1 =$ Life Events, and $X_2 =$ SES

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

	R	R Square
	.589	.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

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

$$Y = 20.0 + 0.069x_1.$$

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. This suggests that subjects who possess greater resources, in the form of higher SES, are better able to withstand the mental stress of potentially traumatic life events.

□

Testing an Interaction Term

For two explanatory variables, the model allowing interaction is

$$E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2.$$

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, and this makes it difficult to estimate effects precisely. Studies based on larger sample sizes (e.g., Holzer 1977) have shown that interaction of the type seen in this example does exist for these variables.

In Table 11.8, neither the test of $H_0: \beta_1 = 0$ or of $H_0: \beta_2 = 0$ have small P -values. Yet, the tests of both $H_0: \beta_1 = 0$ and $H_0: \beta_2 = 0$ are highly significant for the ‘no interaction’ model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$; from Table 11.5, the P -values are 0.003 and 0.002. This loss of significance occurs because $x_3 = x_1x_2$ is quite strongly correlated with x_1 and x_2 , with $r_{X_1X_3} = 0.779$ and $r_{X_2X_3} = 0.646$. These substantial correlations are not surprising, since $x_3 = x_1x_2$ is completely determined by x_1 and x_2 .

Since considerable overlap occurs in the variation in Y that is explained by x_1 and by x_1x_2 , and also by x_2 and x_1x_2 , the *partial* variability explained by each is relatively small. For example, much of the predictive power contained in x_1 is also contained in x_2 and x_1x_2 . The *unique* contribution of x_1 (or x_2) to the model is relatively small, and nonsignificant, when x_2 (or x_1) and x_1x_2 are in the model.

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 health data, we've seen that x_1 and x_2 are highly significant in the model with only those predictors (see Table 11.5) but lose their significance after entering the interaction term, even though the interaction is not significant (see Table 11.8). We also saw 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.

Suppose we center the scores for each variable around 0, by subtracting the mean. Letting $x_1^C = x_1 - \mu_{X_1}$ and $x_2^C = x_2 - \mu_{X_2}$, we then 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 .

When we rerun the interaction model for the mental health data after centering the predictors about their sample means, that is, with LIFE_CEN = LIFE - 44.425 and SES_CEN = SES - 56.60, we get

	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

The estimate for the interaction term is the same as for the model with uncentered predictors, but now 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. Also, their statistical significance is similar as in that model.

Centering the predictor 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. The cross-product term with centered variables does not overlap with the other terms like it does in the ordinary 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.

The interaction terms in the above model are called *second-order*, to distinguish them from *higher-order* interaction terms with products of more than two variables at a time. Such terms are occasionally used in more complex models, not considered in this chapter.

11.6 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 predictors may be needed in the model. We next present a test of whether a model fits significantly better than a simpler model containing only some of the predictors.

Complete and Reduced Models

We refer to the full model with all the predictors as the *complete model*. The model containing only some of these predictors 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 predictors 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 is better than the reduced model.

For instance, a complete model with three explanatory variables and all the second-order 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 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, $SSE_r \geq SSE_c$, because the reduced model has fewer predictors and tends to make poorer predictions. Even if H_0 were true, we would not expect the estimates of the extra parameters and the difference ($SSE_r - SSE_c$) 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, $SSE_r - SSE_c$, that results from adding the extra variables. It has $df =$ the number of extra terms in the complete model. An equivalent statistic uses the R^2 values, R_c^2 for the complete model and R_r^2 for the reduced model. The test statistic equals

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_c/df_2} = \frac{(R_c^2 - R_r^2)/df_1}{(1 - R_c^2)/df_2},$$

where df_1 is the number of extra terms in the complete model and df_2 is the usual residual df for the complete model, which is $df_2 = n - (k + 1)$. A relatively large reduction in error (or relatively large increase in R^2) yields a large F test statistic and 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

$$E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2$$

to the reduced model

$$E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$$

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

The sum of squared errors for the complete model is $SSE_c = 758.8$ (Table 11.8), while for the reduced model it is $SSE_r = 768.2$ (Table 11.7). The difference

$$SSE_r - SSE_c = 768.2 - 758.8 = 9.4$$

has $df_1 = 1$ since the complete model has one more parameter. Since the sample size is $n = 40$, $df_2 = n - (k + 1) = 40 - (3 + 1) = 36$, the df for SSE in Table 11.8. The F test statistic equals

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_c/df_2} = \frac{9.4/1}{758.8/36} = 0.45.$$

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

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

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 equals

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

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 the final subsection in Section 11.4.)

□

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 $H_0: \beta_1 = \cdots = \beta_k = 0$ 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.

11.7 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 predictor, whereas the multiple correlation and R^2 describe the association between Y and the entire set of predictors in the model. The *partial correlation* is based on the ordinary correlations between each pair of variables. For a single control variable, it is defined as follows:

Partial Correlation

The sample **partial correlation** between Y and X_2 , controlling for X_1 , is

$$r_{YX_2 \cdot X_1} = \frac{r_{YX_2} - r_{YX_1}r_{X_1X_2}}{\sqrt{(1 - r_{YX_1}^2)(1 - r_{X_1X_2}^2)}}.$$

In the symbol $r_{YX_2 \cdot X_1}$, the variable to the right of the dot represents the controlled variable. The analogous formula for $r_{YX_1 \cdot X_2}$ (i.e., controlling X_2) is

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{(1 - r_{YX_2}^2)(1 - r_{X_1X_2}^2)}}.$$

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 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, equals

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{(1 - r_{YX_2}^2)(1 - r_{X_1X_2}^2)}} = \frac{0.468 - 0.678(0.791)}{\sqrt{(1 - 0.678^2)(1 - 0.791^2)}} = -0.152.$$

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, such as a range of -1 to $+1$, larger absolute values representing stronger associations, and value free of the units. 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 $\hat{y} = a + b_1x_1 + b_2x_2$. This happens 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 11.4), $r_{YX_1 \cdot X_2}$ 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 correlation computed for the points in the *partial regression plot* (Section 11.2).

Interpreting Squared Partial Correlations

Like r^2 and R^2 , the square of a partial correlation has a proportional reduction in error (PRE) interpretation. It states that $r_{YX_2 \cdot X_1}^2$ 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

$$r_{YX_2 \cdot X_1}^2 = \frac{R^2 - r_{YX_1}^2}{1 - r_{YX_1}^2} = \frac{\text{Partial proportion explained uniquely by } X_2}{\text{Proportion unexplained by } X_1}.$$

Recall from Section 9.4 that $r_{YX_1}^2$ represents the proportion of the variation in Y explained by X_1 . The remaining proportion $(1 - r_{YX_1}^2)$ 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, $R^2 - r_{YX_1}^2$ 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 $1 - r_{YX_1}^2$, the proportion of variation yet to be explained after accounting for the influence of X_1 . The additional explained variation $R^2 - r_{YX_1}^2$ 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. It shows the ratio of the partial contribution of X_2 beyond that of X_1 , namely, $R^2 - r_{YX_1}^2$, divided by the proportion $(1 - r_{YX_1}^2)$ left unexplained by X_1 . Similarly, the square of $r_{YX_1 \cdot X_2}$ equals

$$r_{YX_1 \cdot X_2}^2 = \frac{R^2 - r_{YX_2}^2}{1 - r_{YX_2}^2},$$

the proportion of variation in Y explained by X_1 , out of that part unexplained by X_2 .

Example 11.9 Partial Correlation of Life Events with Mental Impairment

We return to the mental health study, with Y = mental impairment, X_1 = life events, 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 $r_{X_1X_2} = 0.123$. By its definition, the partial correlation between mental impairment and life events, controlling for SES, is

$$r_{YX_1 \cdot X_2} = \frac{r_{YX_1} - r_{YX_2}r_{X_1X_2}}{\sqrt{(1 - r_{YX_2}^2)(1 - r_{X_1X_2}^2)}} = \frac{0.372 - (-0.399)(0.123)}{\sqrt{[1 - (-0.399)^2](1 - 0.123^2)}} = 0.463.$$

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

Since $r_{YX_1 \cdot X_2}^2 = (0.463)^2 = 0.21$, controlling for SES, 21% of the variation in mental impairment is explained by life events. Alternatively, since $R^2 = 0.339$ (Table 11.7),

$$r_{YX_1 \cdot X_2}^2 = \frac{R^2 - r_{YX_2}^2}{1 - r_{YX_2}^2} = \frac{0.339 - (-0.399)^2}{1 - (-0.399)^2} = 0.21.$$

□

Higher-Order Partial Correlations

One reason we showed the connection between squared partial correlation values and R -squared is that this approach also works when the number of control variables exceeds one. For example, with three predictors, let $R_{Y(X_1, X_2, X_3)}^2$ 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 predictors, which we denote by $R_{Y(X_1, X_2)}^2$. The squared partial correlation is

$$r_{YX_3 \cdot X_1, X_2}^2 = \frac{R_{Y(X_1, X_2, X_3)}^2 - R_{Y(X_1, X_2)}^2}{1 - R_{Y(X_1, X_2)}^2}.$$

In this expression, $R_{Y(X_1, X_2, X_3)}^2 - R_{Y(X_1, X_2)}^2$ is the increase in the proportion of explained variance from adding x_3 to the model. The denominator $1 - R_{Y(X_1, X_2)}^2$ is the proportion of the variation left unexplained when x_1 and x_2 are the only predictors 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 .

Inference for Partial Correlations

Controlling for a certain set of variables, the slope of the partial effect of a predictor is 0 in the same situations in which the partial correlation between Y and that predictor is 0. An alternative formula for the t test for a partial effect uses the partial correlation.

With k predictors in the model, the equivalent t test statistic is

$$t = \frac{\text{partial correlation}}{\sqrt{(1 - \text{squared partial correlation})/[n - (k + 1)]}}.$$

This statistic has the t distribution with $df = n - (k + 1)$. It equals the t statistic based on the partial slope estimate and, hence, has the same P -value.

We illustrate by testing that the population partial correlation between mental impairment and life events, controlling for SES, is 0. From Example 11.9, $r_{YX_1 \cdot X_2} = 0.463$. There are $k = 2$ explanatory variables and $n = 40$ observations. The test statistic equals

$$t = \frac{r_{YX_1 \cdot X_2}}{\sqrt{(1 - r_{YX_1 \cdot X_2}^2)/[n - (k + 1)]}} = \frac{0.463}{\sqrt{[1 - (0.463)^2]/37}} = 3.18.$$

This equals the test statistic for $H_0: \beta_1 = 0$ in Table 11.5. Thus, the P -value is also the same, $P = 0.003$.

When no variables are controlled (i.e., the number of explanatory variables is $k = 1$), the t statistic formula simplifies to

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}.$$

This is the statistic for testing that the population bivariate correlation equals 0 (Section 9.5). Confidence intervals for partial correlations are more complex. They require a log transformation such as shown for the correlation in Exercise 65 in Chapter 9.

11.8 Standardized Regression Coefficients*

As in bivariate regression (Recall Section 9.4), 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 β_1^* , β_2^* , \dots .

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. We standardize the partial regression coefficients by adjusting for the differing standard deviation of Y and each X_i . Let s_y denote the sample standard deviation of Y , and let $s_{x_1}, s_{x_2}, \dots, s_{x_k}$ denote the sample standard deviations of the explanatory variables.

The estimates of the standardized regression coefficients are

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

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 $s_{x_2} = 25.3$. Since the unstandardized coefficient of x_1 is $b_1 = 0.103$, 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 equals

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

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 11.9, which repeats Table 11.5, shows how SPSS reports the estimated standardized regression coefficients. It uses the heading BETA, reflecting the alternative name *beta weights* for these coefficients.

□

Table 11.9: SPSS Printout for Fit of Multiple Regression Model to Mental Impairment Data

	Unstandardized		Standardized		
	Coefficients		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 67), and it usually takes 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 predictor 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.

Notation for Standardized Variables

Let $z_Y, z_{X_1}, \dots, z_{X_k}$ denote the standardized versions of the variables Y, X_1, \dots, X_k . 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, \dots, x_k have corresponding z -scores for $z_Y, z_{X_1}, \dots, z_{X_k}$. 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 that variable.

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

$$\hat{z}_Y = b_1^* z_{X_1} + b_2^* z_{X_2} + \dots + b_k^* z_{X_k}.$$

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

$$\hat{z}_Y = 0.43z_{X_1} - 0.45z_{X_2}.$$

Consider a subject who is two standard deviations above the mean on life events but two standard deviations below the mean on SES. This subject 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 that

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

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

Cautions in Comparing Standardized Regression Coefficients

To assess which predictor 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.

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 predictors. For instance, Section 9.6 showed that the correlation depends strongly on the range of x -values sampled. Two groups that have the same value for an estimated regression coefficient have different standardized coefficients if the standard deviation of the predictor differs for the two groups.

Finally, if an explanatory variable is highly correlated with the set of other explanatory variables, it is artificial to conceive of that variable changing while the others remain fixed in value. 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.

11.9 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 k explanatory variables is

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

- 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 $\hat{y} = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$ instead of \bar{y} . It equals the square of the multiple correlation.
- A **partial correlation**, such as $r_{Y X_1 \cdot X_2}$, 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 $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$, that the response variable is independent of all the predictors. A small *P*-value suggests that at least one predictor affects the response.
- Individual *t* tests and confidence intervals for $\{\beta_i\}$ analyze partial effects of each predictor, 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 predictor changes as the value of the other predictor changes. We can allow this by introducing 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, the *F* test compares the SSE values or 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 $k = 2$ explanatory variables, the prediction equation is

$$\hat{Y} = a + b_1x_1 + b_2x_2.$$

Fixing x_2 , 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 . The multiple correlation R is at least as large as the correlations between Y and each predictor. The squared partial correlation $r_{YX_2 \cdot X_1}^2$ is the proportion of the variation of Y that is explained by x_2 , out of that part of the variation left unexplained by x_1 . 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 .

Table 11.10 summarizes the basic properties and inference methods for these measures and those introduced in Chapter 9 for bivariate regression.

The model studied in this chapter is still somewhat restrictive in the sense that all the predictors are quantitative. The next chapter shows how to include categorical predictors in the model.

PROBLEMS

Practicing the Basics

1. For students at Walden University, the relationship between $Y =$ college GPA (with range 0–4.0) and $X_1 =$ high school GPA (range 0–4.0) and $X_2 =$ college board score (range 200–800) satisfies $E(Y) = 0.20 + 0.50x_1 + 0.002x_2$.
 - a) Find the mean college GPA for students having (i) high school GPA = 4.0 and college board score = 800, (ii) $x_1 = 3.0$ and $x_2 = 300$.
 - b) Show that the relationship between Y and x_1 for those students with $x_2 =$

Table 11.10: Summary of Bivariate and Multiple Regression

	BIVARIATE REGRESSION	MULTIPLE REGRESSION	
Model	$E(Y) = \alpha + \beta x$	$E(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$	
Prediction equation	$\hat{y} = a + bx$	$\hat{y} = a + b_1 x_1 + \cdots + b_k x_k$	
		Simultaneous effect of x_1, \dots, x_k	Partial effect of one x_i
Properties of 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, $-1 \leq r_{Y X_1 \cdot X_2} \leq 1$, same sign as b_i and b_i^* , $r_{Y X_1 \cdot X_2}^2$ is PRE measure
Tests of no association	$H_0: \beta = 0$ or $H_0: \rho = 0$, Y not associated with x	$H_0: \beta_1 = \cdots = \beta_k = 0$ (Y not associated with x_1, \dots, x_k)	$H_0: \beta_i = 0$, or H_0 : popul. partial corr. = 0, Y not associated with x_i , controlling for other x variables
Test statistic	$t = \frac{b}{se} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ $df = n - 2$	$F = \frac{\text{Regression MS}}{\text{Residual MS}}$ $= \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$, $df_1 = k, df_2 = n - (k + 1)$	$t = \frac{b_i}{se}$ $df = n - (k + 1)$

500 is $E(Y) = 1.2 + 0.5x_1$.

- c)** Show that when $x_2 = 600$, $E(Y) = 1.4 + 0.5x_1$. Thus, increasing x_2 by 100 shifts the line relating Y to x_1 upward by $100\beta_2 = 0.2$ units.
- d)** Show that setting x_1 at a variety of values yields a collection of parallel lines, each having slope 0.002, relating the mean of Y to x_2 .
2. For recent data in Florida on Y = selling price of home (in dollars), X_1 = size of home (in square feet), X_2 = lot size (in square feet), the prediction equation is $\hat{y} = -10,536 + 53.8x_1 + 2.84x_2$.
- a)** A particular home of 1240 square feet on a lot of 18,000 square feet sold for \$145,000. Find the predicted selling price and the residual, and interpret.
- b)** For fixed lot size, how much is the house selling price predicted to increase for each square foot increase in home size? Why?
3. Refer to the previous exercise:
- a)** For fixed home size, how much would lot size need to increase to have the same impact as a one square foot increase in home size?
- b)** Suppose house selling prices are changed from dollars to thousands of dollars. Explain why the prediction equation changes to $\hat{y} = -10.536 + 0.0538x_1 + 0.00284x_2$.
4. Use software with the “2005 statewide crime” data file at the text website, with murder rate (number of murders per 100,000 people) as the response variable and with percent of high school graduates and the poverty rate (percentage of the population with income below the poverty level) as explanatory variables.
- a)** Construct the partial regression plots. Interpret.
- b)** Report the prediction equation. Explain how to interpret the estimated coefficients.
- c)** Re-do the analyses after deleting the D.C. observation. Does this observation have much influence on the results?
5. A regression analysis with recent U.N. data from several nations on Y = percentage of people who use the Internet, X_1 = per capita gross domestic product (in thousands of dollars), and X_2 = percentage of people using cell phones has results shown in Table 11.11.
- a)** Write the prediction equation.
- b)** Find the predicted Internet use for a country with per capita GDP of 10 thousand dollars and 50% using cell phones.
- c)** Find the prediction equations when cell-phone use is (i) 0 %, (ii) 100%, and use them to interpret the effect of GDP.
- d)** Use the equations in (c) to explain the ‘no interaction’ property of the model.
6. Refer to the previous exercise.
- a)** Show how to obtain R -squared from the sums of squares in the ANOVA table. Interpret it.
- b)** $r^2 = 0.78$ when GDP is the sole predictor. Why do you think R^2 does not increase much when cell-phone use is added to the model, even though it is

Table 11.11:

	B	Std. Error	t	Sig
(Constant)	-3.601	2.506	-1.44	0.159
GDP	1.2799	0.2703	4.74	0.000
CELLULAR	0.1021	0.0900	1.13	0.264

R Square .796

ANOVA

	Sum of Squares	DF
Regression	10316.8	2
Residual Error	2642.5	36
Total	12959.3	38

itself highly associated with Y (with $r = 0.67$)? (Hint: Would you expect X_1 and X_2 to be highly correlated? If so, what's the effect?)

7. Table 9.17 showed data from Florida counties on $Y =$ crime rate (number per 1000 residents), $X_1 =$ median income (thousands of dollars), and $X_2 =$ percent in urban environment.
 - a) Figure 11.12 shows a scatterplot relating Y to X_1 . Predict the sign that the estimated effect of X_1 has in the prediction equation $\hat{Y} = a + bx_1$. Explain.
 - b) Figure 11.13 shows a partial regression plot relating Y to X_1 , controlling for X_2 . Predict the sign that the estimated effect of X_1 has in the prediction equation $\hat{Y} = a + b_1x_1 + b_2x_2$. Explain.
 - c) Table 11.12 shows part of a printout for the bivariate and multiple regression models. Report the prediction equation relating Y to x_1 , and interpret the slope.
 - d) Report the prediction equation relating Y to both x_1 and x_2 . Interpret the coefficient of x_1 , and compare to (c).
 - e) The correlations are $r_{YX_1} = 0.43$, $r_{YX_2} = 0.68$, $r_{X_1X_2} = 0.73$. Use these to explain why the x_1 effect seems so different in (c) and (d).
 - f) Report the prediction equations relating crime rate to income at urbanization levels of (i) 0, (ii) 50, (iii) 100. Interpret.

8. Refer to the previous exercise. Using software with the "Florida crime" data file at the text website:
 - a) Construct box plots for each variable and scatterplots and partial regression plots between Y and each of x_1 and x_2 . Interpret these plots.
 - b) Find the prediction equations for the bivariate effects of x_1 and of x_2 . Interpret.
 - c) Find the prediction equation for the multiple regression model. Interpret.

Table 11.12:

	B	Std. Error	t	Sig
(Constant)	-11.526	16.834	-0.685	0.4960
INCOME	2.609	0.675	3.866	0.0003

	B	Std. Error	t	Sig
(Constant)	40.261	16.365	2.460	0.0166
INCOME	-0.809	0.805	-1.005	0.3189
URBAN	0.646	0.111	5.811	0.0001

- d) Find R^2 for the multiple regression model, and show that it is not much larger than r^2 for the model using urbanization alone as the predictor. Interpret.
9. Recent UN data from several nations on Y = crude birth rate (number of births per 1000 population size), X_1 = women's economic activity (female labor force as percentage of male), and X_2 = GNP (per capita, in thousands of dollars) has prediction equation $\hat{y} = 34.53 - 0.13x_1 - 0.64x_2$.
- Interpret the coefficient of x_1 .
 - Sketch on a single graph the relationship between Y and x_1 when $x_2 = 0$, $x_2 = 10$, and $x_2 = 20$. Interpret the results.
 - The bivariate prediction equation with x_1 is $\hat{y} = 37.65 - 0.31x_1$. The correlations are $r_{YX_1} = -0.58$, $r_{YX_2} = -0.72$, and $r_{X_1X_2} = 0.58$. Explain why the coefficient of x_1 in the bivariate equation is quite different from in the multiple predictor equation.
10. For recent UN data for several nations, a regression of carbon dioxide use (CO₂, a measure of air pollution) on gross domestic product (GDP) has a correlation of 0.786. With life expectancy as a second explanatory variable, the multiple correlation is 0.787.
- Explain how to interpret the multiple correlation.
 - For predicting CO₂, did it help much to add life expectancy to the model? Does this mean that life expectancy is very weakly correlated with CO₂? Explain.
11. Table 11.13 shows a printout from fitting the multiple regression model to recent statewide data, excluding D.C., on Y = violent crime rate (per 100,000 people), X_1 = poverty rate (percentage with income below the poverty level), and X_2 = percent living in metropolitan areas.
- Report the prediction equation.
 - Massachusetts had $y = 805$, $x_1 = 10.7$, and $x_2 = 96.2$. Find its predicted

Table 11.13:

	Sum of Squares	DF	Mean Square	F	Sig
Regression	2448368.07	2	1224184.04	31.249	0.0001
Residual	1841257.15	47	39175.68		
Total	4289625.22	49			

R	R Square	Std Error of the Estimate
.7555	.5708	197.928

	B	Std. Error	t	Sig
(Constant)	-498.683	140.988	-3.537	0.0009
POVERTY	32.622	6.677	4.885	0.0001
METRO	9.112	1.321	6.900	0.0001

	Correlations		
	VIOLENT	POVERTY	METRO
VIOLENT	1.0000	.3688	.5940
POVERTY	.3688	1.0000	-.1556
METRO	.5940	-.1556	1.0000

- violent crime rate. Find the residual, and interpret.
- c) Interpret the fit by showing the prediction equation relating \hat{y} and x_1 for states with (i) $x_2 = 0$, (ii) $x_2 = 50$, (iii) $x_2 = 100$. Interpret.
- d) Interpret the correlation matrix.
- e) Report R^2 and the multiple correlation, and interpret.
12. Refer to the previous exercise.
- a) Report the F statistic for testing $H_0: \beta_1 = \beta_2 = 0$, report its df values and P -value, and interpret.
- b) Show how to construct the t statistic for testing $H_0: \beta_1 = 0$, report its df and P -value for $H_a: \beta_1 \neq 0$, and interpret.
- c) Construct a 95% confidence interval for β_1 , and interpret.
- d) Since these analyses use data for all the states, what relevance, if any, do the inferences have in (a)–(c)?
13. Refer to the previous two exercises. When we add $x_3 =$ percentage of single-parent families to the model, we get the results in Table 11.14.
- a) Report the prediction equation and interpret the coefficient of poverty rate.
- b) Why do you think the effect of poverty rate is much lower after x_3 is added to the model?

Table 11.14:

Variable	Coefficient	Std. Error
Intercept	-1197.538	
Poverty	18.283	(6.136)
Metropolitan	7.712	(1.109)
Single-parent	89.401	(17.836)
R^2	0.722	
n	50	

14. Table 11.15 comes from a regression analysis of $Y =$ number of children in family, $X_1 =$ mother's educational level in years (MEDUC), and $X_2 =$ father's socioeconomic status (FSES), for a random sample of 49 college students at Texas A&M University.
- a) Write the prediction equation. Interpret parameter estimates.
- b) For the first subject in the sample, $x_1 = 12$, $x_2 = 61$, and $y = 5$. Find the predicted value of y and the residual, and interpret.
- c) Report SSE. Use it to explain the least squares property of this prediction equation.
- d) Explain why it is not possible that $r_{YX_1 \cdot X_2} = 0.40$.
- b) Can you tell from the table whether r_{YX_1} is positive or negative? Explain.
15. The General Social Survey has asked subjects to rate various groups using the "feeling thermometer." The rating is between 0 and 100, more favorable as the

Table 11.15:

	Sum of Squares
Regression	31.8
Residual	199.3
	B
(Constant)	5.25
MEDUC	-0.24
FSES	0.02

score gets closer to 100 and less favorable as the score gets closer to 0. For a small data set from the GSS, Table 11.16 shows results of fitting the multiple regression model with feelings toward liberals as the response, using explanatory variables political ideology (scores 1 = extremely liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = extremely conservative) and religious attendance, using scores (1 = never, 2 = less than once a year, 3 = once or twice a year, 4 = several times a year, 5 = about once a month, 6 = 2-3 times a month, 7 = nearly every week, 8 = every week, 9 = several times a week). Standard errors are shown in parentheses.

- a) Report the prediction equation and interpret the ideology partial effect.
 - b) Report the predicted value and residual for the first observation, for which ideology = 7, religion = 9, and feelings = 10.
 - c) Report, and explain how to interpret, R^2 .
 - d) Tables of this form often put * by an effect having $P < 0.05$, ** by an effect having $P < 0.01$, and *** by an effect having $P < 0.001$. Show how this was determined for the ideology effect, and discuss the disadvantage of summarizing in this manner.
 - e) Explain how the F value can be obtained from the R^2 value reported. Report its df values, and explain how to interpret its result.
 - f) The estimated standardized regression coefficients are -0.79 for ideology and -0.23 for religion. Interpret.
16. Refer to Table 11.5. Test $H_0: \beta_2 = 0$ that mental impairment is independent of SES, controlling for life events. Report the test statistic, and report and interpret the P -value for (a) $H_a: \beta_2 \neq 0$, (b) $H_a: \beta_2 < 0$.
 17. For a random sample of 66 state precincts, data are available on

- Y = Percentage of adult residents who are registered to vote
- X_1 = Percentage of adult residents owning homes
- X_2 = Percentage of adult residents who are nonwhite
- X_3 = Median family income (thousands of dollars)
- X_4 = Median age of residents

Table 11.16:

Variable	Coefficient
Intercept	135.31
Ideology	-14.07 (3.16)**
Religion	-2.95 (2.26)
F	13.93**
R ²	0.799
Adj. R ²	0.742
(n)	(10)

X_5 = Percentage of residents who have lived in the
precinct at least ten years

Table 11.16 shows a portion of the printout used to analyze the data.

- a) Fill in all the missing values in the printout, indicating in each ‘Sig’ space whether $P > 0.05$, $0.01 < P < 0.05$, $0.001 < P < 0.01$, or $P < 0.001$.
 - b) Do you think it is necessary to include all five explanatory variables in the model? Explain.
 - c) To what test does the “F Value” refer? Interpret the result of that test.
 - d) To what test does the t -value opposite x_1 refer? Interpret the result of that test.
18. Refer to the previous exercise.
- a) Find a 95% confidence interval for the change in the mean of Y for a 1-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.
 - b) Find a 95% confidence interval for the change in the mean of Y for a 50-unit increase in the percentage of adults owning homes, controlling for the other variables. Interpret.
19. Use software with the “house selling price” data file at the text website to conduct a multiple regression analysis of Y = selling price of home (dollars), X_1 = size of home (square feet), X_2 = number of bedrooms, X_3 = number of bathrooms.
- a) Use graphics to display the effects of the predictors. Interpret, and explain how the highly discrete nature of x_2 and x_3 affects the plots.
 - b) Report the prediction equation and interpret the estimates.
 - c) Inspect the correlation matrix, and report the variables having the (i) strongest association, (ii) weakest association.
 - d) Report R^2 , and interpret.

Table 11.17:

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	----	---	----	----	----	----
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0500	----	----

- e) Find the F statistic for testing the overall effect of the three predictors, report its df values and its P -value, and interpret.
- f) Find the t test statistic for $H_0: \beta_3 = 0$, report its P -value for $H_a: \beta_3 > 0$, and interpret.
20. Refer to the previous exercise. Now use only number of bathrooms and number of bedrooms as predictors.
- Again test the partial effect of number of bathrooms, and interpret.
 - Construct a 95% confidence interval for the coefficient of number of bathrooms, and interpret.
 - Find the partial correlation between selling price and number of bathrooms, controlling for number of bedrooms. Compare it to the correlation, and interpret.
 - Find the estimated standardized regression coefficients for the model, and interpret.
 - Write the prediction equation using standardized variables. Interpret.
21. Exercise 11 showed a regression analysis for statewide data on $Y =$ violent crime rate, $X_1 =$ poverty rate, and $X_2 =$ percent living in metropolitan areas. When we add an interaction term, we get the prediction equation $\hat{y} = 158.9 - 14.72x_1 - 1.29x_2 + 0.76x_1x_2$.
- As the percentage living in metropolitan areas increases, does the effect of poverty rate tend to increase or decrease? Explain.
 - Show how to interpret the prediction equation, by finding how it simplifies

when $x_2 = 0, 50,$ and 100 .

22. A study analyzes relationships among $Y =$ percentage vote for Democratic candidate, $X_1 =$ percentage of registered voters who are Democrats, and $X_2 =$ percentage of registered voters who vote in the election, for several congressional elections in 2006. The researchers expect interaction, since they expect a higher slope between Y and x_1 at larger values of x_2 than at smaller values. They obtain the prediction equation $\hat{Y} = 20 + 0.30x_1 + 0.05x_2 + 0.005x_1x_2$. Does this equation support the direction of their prediction? Explain.
23. Use software with the “house selling price” data file to allow interaction between number of bedrooms and number of bathrooms in their effects on selling price.
- Report the prediction equation.
 - Interpret the fit by showing the equation relating \hat{y} and number of bedrooms for homes with (i) two bathrooms, (ii) three bathrooms.
 - Use a test to analyze the significance of the interaction term. Interpret.
24. A multiple regression analysis investigates the relationship between $Y =$ college GPA and several explanatory variables, using a random sample of 195 students at Slippery Rock University. First, high school GPA and total SAT score are entered into the model. The sum of squared errors is $SSE = 20$. Next, parents’ education and parents’ income are added, to determine if they have an effect, controlling for high school GPA and SAT. For this expanded model $SSE = 19$. Test whether this complete model is significantly better than the one containing only high school GPA and SAT. Report and interpret the P -value.
25. Table 11.18 shows results of regressing $Y =$ birth rate (BIRTHS, number of births per 1000 population) on $x_1 =$ women’s economic activity (ECON) and $x_2 =$ literacy rate (LITERACY), using UN data for 23 nations.
- Report the value of each of the following:
 - r_{YX_1}
 - r_{YX_2}
 - R^2
 - TSS
 - SSE
 - mean square error
 - s
 - s_y
 - se for b_1
 - t for $H_0: \beta_1 = 0$
 - P for $H_0: \beta_1 = 0$ against $H_a: \beta_1 \neq 0$
 - P for $H_0: \beta_1 = 0$ against $H_a: \beta_1 < 0$
 - F for $H_0: \beta_1 = \beta_2 = 0$
 - P for $H_0: \beta_1 = \beta_2 = 0$
 - Report the prediction equation, and carefully interpret the three estimated regression coefficients.
 - Interpret the correlations r_{YX_1} and r_{YX_2} .
 - Report R^2 , and interpret its value.
 - Report the multiple correlation, and interpret.
 - Though inference may not be relevant for these data, report the F statistic for $H_0: \beta_1 = \beta_2 = 0$, report its P -value, and interpret.

g) Show how to construct the t statistic for $H_0: \beta_1 = 0$, report its df and P -value for $H_a: \beta_1 \neq 0$, and interpret.

Table 11.18:

	Mean	Std Deviation	N
BIRTHS	22.117	10.469	23
ECON	47.826	19.872	23
LITERACY	77.696	17.665	23

Correlations				
Correlation		BIRTHS	ECON	LITER
	BIRTHS	1.00000	-0.61181	-0.81872
	ECON	-0.61181	1.00000	0.42056
	LITERACY	-0.81872	0.42056	1.00000

Sig. (2-tailed)		BIRTHS	ECON	LITER
	BIRTHS	.	0.0019	0.0001
	ECON	0.0019	.	0.0457
	LITERACY	0.0001	0.0457	.

	Sum of Squares	DF	Mean Square	F	Sig
Regression	1825.969	2	912.985	31.191	0.0001
Residual	585.424	20	29.271		
Total	2411.393	22			

Root MSE (Std. Error of the Estimate) 5.410 R Square 0.7572

	Unstandardized Coeff.	Standardized		
	B	Std. Error	Coeff. (Beta)	t
(Constant)	61.713	5.2453		11.765
ECON	-0.171	0.0640	-0.325	-2.676
LITERACY	-0.404	0.0720	-0.682	-5.616

26. Refer to the previous exercise.
- Find the partial correlation between Y and X_1 , controlling for X_2 . Interpret both the partial correlation and its square.
 - Find the estimate of the conditional standard deviation, and interpret its value.
 - Show how to find the estimated standardized regression coefficient for x_1 using the unstandardized estimate and the standard deviations, and interpret its value.

- d) Write the prediction equation using standardized variables. Interpret.
 e) Find the predicted z -score for a country that is one standard deviation above the mean on both predictors. Interpret.
27. Refer to Examples 11.1 and 11.8. Explain why the partial correlation between crime rate and high school graduation rate is so different from the bivariate correlation. (This is an example of *Simpson's paradox*, which states that a bivariate association can have a different direction than a partial association.)
28. For a group of 100 children of ages varying from 3 to 15, the correlation between vocabulary score on an achievement test and height of child is 0.65. The correlation between vocabulary score and age for this sample is 0.85, and the correlation between height and age is 0.75.
- a) Show that the partial correlation between vocabulary and height, controlling for age, is 0.036. Interpret.
 b) Test whether this partial correlation is significantly nonzero. Interpret.
 c) Is it plausible that the relationship between height and vocabulary is spurious, in the sense that it is due to their joint dependence on age? Explain.
29. A multiple regression model describes the relationship among a collection of cities between $Y =$ murder rate (number of murders per 100,000 residents) and
- $X_1 =$ Number of police officers (per 100,000 residents)
 $X_2 =$ Median length of prison sentence given to convicted murderers
 (in years)
 $X_3 =$ Median income of residents of city (in thousands of dollars)
 $X_4 =$ Unemployment rate in city

These variables are observed for a random sample of thirty cities with population size exceeding 35,000. For these cities, the prediction equation is $\hat{y} = 30 - 0.02x_1 - 0.1x_2 - 1.2x_3 + 0.8x_4$, and $\bar{y} = 15$, $\bar{x}_1 = 100$, $\bar{x}_2 = 15$, $\bar{x}_3 = 13$, $\bar{x}_4 = 7.8$, $s_y = 8$, $s_{x_1} = 30$, $s_{x_2} = 10$, $s_{x_3} = 2$, $s_{x_4} = 2$.

- a) Can you tell from the coefficients of the prediction equation which explanatory variable has the greatest partial effect on Y ? Explain.
 b) Find the standardized regression coefficients and interpret their values.
 c) Write the prediction equation using standardized variables. Find the predicted z -score on murder rate for a city that is one standard deviation above the mean on x_1 , x_2 , and x_3 , and one standard deviation below the mean on x_4 . Interpret.
30. Exercise 11 showed a regression of violent crime rate on poverty rate and percent living in metropolitan areas. The estimated standardized regression coefficients are 0.473 for poverty rate and 0.668 for percent in metropolitan areas.
- a) Interpret the estimated standardized regression coefficients.
 b) Express the prediction equation using standardized variables, and explain how it is used.

Concepts and Applications

31. Refer to the student survey data set (Exercise 1.11). Using software, conduct a regression analysis using $Y =$ political ideology with predictors number of times per week of newspaper reading and religiosity. Prepare a report, summarizing your graphical analyses, bivariate models and interpretations, multiple regression models and interpretations, inferences, checks of effects of outliers, and overall summary of the relationships.
32. Repeat the previous exercise using $Y =$ college GPA with predictors high school GPA and number of weekly hours of physical exercise
33. Refer to the student data file you created in Exercise 1.12. For variables chosen by your instructor, fit a multiple regression model and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings.
34. Using software with the “2005 statewide crime” data file at the text website, conduct a regression analysis of murder rate with predictors poverty rate, the percent living in urban areas, and percent of high school graduates. Conduct descriptive and inferential analyses. Provide interpretations, and provide a paragraph summary of your conclusions at the end of your report.
35. Repeat the previous exercise using violent crime rate as the response variable.
36. Refer to Exercise 34. Repeat this problem, excluding the observation for D.C. Describe the effect on the various analyses of this observation.
37. Table 27 in Chapter 9 is the “UN data” data file at the text website. Construct a multiple regression model containing two explanatory variables that provide good predictions for the fertility rate. How did you select this model? (*Hint:* One way is based on entries in the correlation matrix.)
38. In about 200 words, explain to someone who has never studied statistics what multiple regression does and how it can be useful.
39. Analyze the “house selling price” data file at the text website (which were introduced in Example 9.10), using selling price of home, size of home, number of bedrooms, and taxes. Prepare a short report summarizing your analyses and conclusions.
40. For Example 11.2 on mental impairment, Table 11.19 shows the result of adding religious attendance as a predictor, measured as the approximate number of times the subject attends a religious service over the course of a year. Write a short report, interpreting the information from this table.
41. a study³ of mortality rates found in the U.S. that states with higher income inequality tended to have higher mortality rates. The effect of income inequality

³A. Muller, *BMJ*, vol. 324, 2002

Table 11.19:

Variable	Coefficient
Intercept	27.422
Life events	0.0935 (0.0313)**
SES	-0.0958 (0.0256)***
Religious attendance	-0.0370 (0.0219)
R^2	0.358
(n)	(40)

disappeared after controlling for the percentage of a state’s residents that had at least a high school education. Explain how these results relate to analyses conducted using bivariate regression and multiple regression.

42. A 2002 study⁴ relating the percentage of a child’s life spent in poverty to number of years of education completed by the mother and the percentage of a child’s life spent in a single parent home reported the results shown in Table 11.20. Prepare a one-page report explaining how to interpret the results in this table.

Table 11.20:

	Unstandardized		Standardized		
	Coefficients		Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	56.401	2.121		12.662	.000
% single parent	0.323	.014	.295	11.362	.000
mother school	-3.330	.152	-.290	-11.294	.000

F 611.6 (df = 2, 4731) Sig .000
R 0.453 R Square 0.205

43. *The Economist* magazine⁵ developed a quality-of-life index for nations as the predicted value obtained by regressing an average of life-satisfaction scores from several surveys on gross domestic product (GDP, per capita, in dollars), life expectancy (in years), an index of political freedom (from 1 = completely free

⁴<http://www.heritage.org/Research/Family/cda02-05.cfm>

⁵<http://www.economist.com/media/pdf/QUALITYOFLIFE.pdf>

to 7 = unfree), the percentage unemployed, the divorce rate (on a scale of 1 for lowest rates to 5 for highest), latitude (to distinguish between warmer and cold climates), a political stability measure, gender equality defined as the ratio of average male and female earnings, and community life (1 if country has high rate of church attendance or trade-union membership, 0 otherwise). Table 11.21 shows results of the model fit for 74 countries, for which the multiple correlation is 0.92. The study used the prediction equation to predict the quality of life in 2005 for 111 nations. The top 10 ranks were for Ireland, Switzerland, Norway, Luxembourg, Sweden, Australia, Iceland, Italy, Denmark, and Spain. Other ranks included 13 for the U.S., 14 for Canada, 15 for New Zealand, 16 for Netherlands, and 29 for the U.K.

a) Which variables would you expect to have negative effects on quality of life? Is this supported by the results?

b) The study states that “GDP explains more than 50% of the variation in life satisfaction.” How does this relate to a summary measure of association?

c) The study reported that “Using so-called Beta coefficients from the regression to derive the weights of the various factors, life expectancy and GDP were the most important.” Explain what was meant by this.

d) Although GDP seems to be an important predictor, in a bivariate sense and a partial sense, Table 11.21 reports a very small coefficient, 0.00003. Why do you think this is?

e) The study mentioned other predictors that were not included because they provided no further predictive power. For example, the study stated that education seemed to have an effect mainly through its effects on other variables in the model, such as GDP, life expectancy, and political freedom. Does this mean there is no association between education and quality of life? Explain.

Table 11.21:

	Coefficients	Standard error	<i>t</i> statistic
Constant	2.796	0.789	3.54
GDP per person	0.00003	0.00001	3.52
Life expectancy	0.045	0.011	4.23
Political freedom	-0.105	0.056	-1.87
Unemployment	-0.022	0.010	-2.21
Divorce rate	-0.188	0.064	-2.93
Latitude	-1.353	0.469	-2.89
Political stability	0.152	0.052	2.92
Gender equality	0.742	0.543	1.37
Community life	0.386	0.124	3.13

44. A recent article⁶ used multiple regression to predict attitudes toward homosex-

⁶T. Shackelford and A. Besser, *Individual Differences Research*, 2007

uality. The researchers found that the effect of number of years of education on a measure of tolerance toward homosexuality varied from essentially no effect for political conservatives to a considerably positive effect for political liberals. Explain how this is an example of statistical interaction, and explain how it would be handled by a multiple regression model.

45. In the study mentioned in the previous exercise, a separate model did not contain interaction terms. The best predictor of attitudes toward homosexuality was educational level, with an estimated standardized regression coefficient of 0.21. The authors also reported, “Controlling for other variables, an additional year of education completed was associated with a .09 rating unit increase in attitudes toward homosexuality.” In comparing the effect of education with the effects of other predictors in the model, such as the age of the subject, explain the purpose of estimating standardized coefficients. Explain how to interpret the one reported for education.
46. For a linear model with two explanatory variables X_1 and X_2 , which of the following must be incorrect? Why?
 - a) $r_{YX_1} = 0.01$, $r_{YX_2} = -0.2$, $R = .75$
 - b) $r_{YX_1} = 0.01$, $r_{YX_2} = -0.75$, $R = 0.2$
 - c) $r_{YX_1} = 0.4$, $r_{YX_2} = 0.4$, $R = 0.4$
47. In Exercise 1 on $Y =$ college GPA, $X_1 =$ high school GPA, and $X_2 =$ college board score, $E(Y) = 0.20 + 0.50x_1 + 0.002x_2$. True or false: Since $\beta_1 = 0.50$ is larger than $\beta_2 = 0.002$, this implies that X_1 has the greater partial effect on Y . Explain.
48. Table 11.20 shows results of fitting various regression models to data on $Y =$ college GPA, $X_1 =$ high school GPA, $X_2 =$ mathematics entrance exam score, and $X_3 =$ verbal entrance exam score. Indicate which of the following statements are false. Give a reason for your answer.

Table 11.22:

Estimates	Model		
	$E(Y) = \alpha + \beta x_1$	$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
Coefficient of x_1	0.450	0.400	0.340
Coefficient of x_2		0.003	0.002
Coefficient of x_3			0.002
R^2	0.25	0.34	0.38

- a) The correlation between Y and X_1 is positive.
- b) A one-unit increase in x_1 corresponds to a change of 0.45 in the estimated mean of Y , controlling for x_2 and x_3 .
- c) The value of SSE increases as we add additional variables to the model.

- d) It follows from the sizes of the estimates for the third model that X_1 has the strongest partial effect on Y .
- e) The value of $r_{YX_3}^2$ is 0.40.
- f) The partial correlation $r_{YX_1 \cdot X_2}$ is positive.
- g) The partial correlation $r_{YX_1 \cdot X_3}$ could be negative.
- h) Controlling for X_1 , a 100-unit increase in X_2 corresponds to a predicted increase of 0.3 in college GPA.
- i) For the first model, the estimated standardized regression coefficient equals 0.50.
49. In regression analysis, which of the following statements must be false? Why?
- a) For the model $E(Y) = \alpha + \beta_1 x_1$, Y is significantly related to x_1 at the 0.05 level, but when x_2 is added to the model, Y is not significantly related to x_1 at the 0.05 level.
- b) The estimated coefficient of x_1 is positive in the bivariate model, but negative in the multiple regression model.
- c) When the model is refitted after Y is multiplied by 10, R^2 , r_{YX_1} , $r_{YX_1 \cdot X_2}$, b_1^* , the F statistics and t statistics do not change.
- d) $r_{YX_2 \cdot X_1}$ cannot exceed r_{YX_2} .
- e) The F statistic for testing that all the regression coefficients equal 0 has $P < 0.05$, but none of the individual t tests have $P < 0.05$.
- f) If you compute the standardized regression coefficient for a bivariate model, you always get the correlation.
- g) $r_{YX_1}^2 = r_{YX_2}^2 = 0.6$ and $R^2 = 0.6$.
- h) $r_{YX_1}^2 = r_{YX_2}^2 = 0.6$ and $R^2 = 1.2$.
- i) The correlation between Y and \hat{Y} equals -0.10 .
- j) If x_3 is added to a model already containing x_1 and x_2 , then if the prediction equation has $b_3 = 0$, R^2 stays the same.
- k) For every F test, there is an equivalent test using the t distribution.

For Exercises 50–54, select the correct answer(s) and indicate why the other responses are inappropriate. (More than one response may be correct.)

50. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$, then controlling for x_2 and x_3 , the predicted mean change in Y when x_1 is increased from 10 to 20 equals
 a) 3 b) 30 c) 0.3 d) Cannot be given—depends on specific values of x_2 and x_3 .
51. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$,
 a) The strongest correlation is between Y and X_3 .
 b) The variable with the strongest partial influence on Y is X_2 .
 c) The variable with the strongest partial influence on Y is X_3 , but one cannot tell from this equation which pair has the strongest correlation.
 d) None of the above.
52. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$,
 a) $r_{YX_3} < 0$

- b) $r_{YX_3 \cdot X_1} < 0$
 c) $r_{YX_3 \cdot X_1, X_2} < 0$
 d) Insufficient information to answer.
 e) Answers (a), (b), and (c) are all correct.
53. If $\hat{Y} = 2 + 3x_1 + 5x_2 - 8x_3$, and $H_0: \beta_3 = 0$ is rejected at the 0.05 level, then
 a) $H_0: \rho_{YX_3 \cdot X_1, X_2} = 0$ is rejected at the 0.05 level.
 b) $H_0: \rho_{YX_3} = 0$ is rejected at the .05 level.
 c) $r_{YX_3 \cdot X_1, X_2} > 0$
54. The F test for comparing a complete model to a reduced model
 a) Can be used to test the significance of a single regression parameter in a multiple regression model.
 b) Can be used to test $H_0: \beta_1 = \dots = \beta_k = 0$ in a multiple regression equation.
 c) Can be used to test H_0 : No interaction, in the model
- $$E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_1x_2 + \beta_5x_1x_3 + \beta_6x_2x_3.$$
- d) Can be used to test whether the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$ gives a significantly better fit than the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_3$.
55. Explain the difference in the purposes of the correlation, the multiple correlation, and the partial correlation.
56. Let Y = height, X_1 = length of right leg, X_2 = length of left leg. Describe what you expect for the relative sizes of the three pairwise correlations, R , and $r_{YX_2 \cdot X_1}$.
57. Give an example of three variables for which you expect $\beta \neq 0$ in the model $E(Y) = \alpha + \beta x_1$ but $\beta_1 = 0$ in the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$.
58. For the models $E(Y) = \alpha + \beta x$ and $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$, express null hypotheses in terms of correlations that are equivalent to the following:
 a) $H_0: \beta = 0$
 b) $H_0: \beta_1 = \beta_2 = 0$
 c) $H_0: \beta_2 = 0$
59. * Whenever X_1 and X_2 are uncorrelated, then R^2 for the model $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2$ satisfies $R^2 = r_{YX_1}^2 + r_{YX_2}^2$. In this case, draw a figure that portrays the variability in Y , the part of that variability explained by each of X_1 and X_2 , and the total variability explained by both of them together.
60. * Which of the following sets of correlations would you expect to yield the highest R^2 value? Why?
 a) $r_{YX_1} = 0.4$, $r_{YX_2} = 0.4$, $r_{X_1X_2} = 0.0$
 b) $r_{YX_1} = 0.4$, $r_{YX_2} = 0.4$, $r_{X_1X_2} = 0.5$
 c) $r_{YX_1} = 0.4$, $r_{YX_2} = 0.4$, $r_{X_1X_2} = 1.0$

61. * Suppose the correlation between Y and X_1 equals the multiple correlation between Y and X_1 and X_2 . What does this imply about the partial correlation $r_{YX_2 \cdot X_1}$? Interpret.
62. * Software reports four types of sums of squares in multiple regression models. The **Type I** (sometimes called *sequential*) sum of squares represents the variability explained by a variable, controlling for variables previously entered into the model. The **Type III** (sometimes called *partial*) sum of squares represents the variability explained by that variable, controlling for all other variables in the model.
- a) For any multiple regression model, explain why the Type I sum of squares for x_1 is the regression sum of squares for the bivariate model with x_1 as the predictor, whereas the Type I sum of squares for x_2 equals the amount by which SSE decreases when x_2 is added to the model.
- b) Explain why the Type I sum of squares for the last variable entered into a model is the same as the Type III sum of squares for that variable.
63. * The sample value of 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. This bias is greater if n is small or the number of predictors k is large. A somewhat better estimate is **adjusted R^2** ,

$$R_{\text{adj}}^2 = 1 - \frac{s^2}{s_Y^2} = R^2 - \left[\frac{k}{n - (k + 1)} \right] (1 - R^2),$$

where s^2 is the estimated conditional variance (i.e., the mean square error) and s_Y^2 is the sample variance of Y .

- a) Suppose $R^2 = 0.339$ for a model with $k = 2$ explanatory variables (such as in Table 11.5). Find R_{adj}^2 for the following sample sizes: 10, 40 (as in the text example), 100, and 1000. Show that R_{adj}^2 approaches R^2 in value as n increases.
- b) Show that R_{adj}^2 is negative when $R^2 < k/(n - 1)$. This is undesirable, and R_{adj}^2 is equated to 0 in such cases. (Also, unlike R^2 , R_{adj}^2 could decrease when we add an explanatory variable to a model.)
64. * Let $R_{Y(X_1, \dots, X_k)}^2$ denote R^2 for the multiple regression model with k explanatory variables. Explain why

$$r_{YX_k \cdot X_1, \dots, X_{k-1}}^2 = \frac{R_{Y(X_1, \dots, X_k)}^2 - R_{Y(X_1, \dots, X_{k-1})}^2}{1 - R_{Y(X_1, \dots, X_{k-1})}^2}.$$

65. * The numerator $R^2 - r_{YX_1}^2$ of the squared partial correlation $r_{YX_2 \cdot X_1}^2$ gives the increase in the proportion of explained variation from adding X_2 to the model. This increment, denoted by $r_{Y(X_2 \cdot X_1)}^2$, is called the squared **semipartial correlation**. One can use squared semipartial correlations to partition the

variation in the response variable. For instance, for three explanatory variables,

$$\begin{aligned} R_{Y(X_1, X_2, X_3)}^2 &= r_{YX_1}^2 + (R_{Y(X_1, X_2)}^2 - r_{YX_1}^2) + (R_{Y(X_1, X_2, X_3)}^2 - R_{Y(X_1, X_2)}^2) \\ &= r_{YX_1}^2 + r_{Y(X_2 \cdot X_1)}^2 + r_{Y(X_3 \cdot X_1, X_2)}^2. \end{aligned}$$

The total variation in Y explained by X_1 , X_2 , and X_3 together partitions into: (i) the proportion explained by X_1 (i.e., $r_{YX_1}^2$), (ii) the proportion explained by X_2 beyond that explained by X_1 (i.e., $r_{Y(X_2 \cdot X_1)}^2$), and (iii) the proportion explained by X_3 beyond that explained by X_1 and X_2 (i.e., $r_{Y(X_3 \cdot X_1, X_2)}^2$). These correlations have the same ordering as the t statistics for testing partial effects, and some researchers use them as indices of importance of the predictors.

a) In Example 11.2 on mental impairment, show that $r_{Y(X_2 \cdot X_1)}^2 = 0.20$ and $r_{Y(X_1 \cdot X_2)}^2 = 0.18$. Interpret.

b) Explain why the squared semipartial correlation $r_{Y(X_2 \cdot X_1)}^2$ cannot be larger than the squared partial correlation $r_{YX_2 \cdot X_1}^2$.

66. * The least squares prediction equation provides predicted values \hat{Y} with the strongest possible correlation with Y , out of all possible prediction equations of that form. That is, the least squares equation yields the best prediction of Y in the sense that it represents the linear reduction of X_1, \dots, X_k to the single variable that is most strongly correlated with Y . Based on this property, explain why the multiple correlation cannot decrease when one adds a variable to a multiple regression model. (*Hint:* The prediction equation for the simpler model is a special case of a prediction equation for the full model that has coefficient 0 for the added variable.)

67. * Let \bar{b}_i^* denote the estimated standardized regression coefficient when X_i is treated as the *response* variable and Y as an *explanatory* variable, controlling for the same set of other variables. Then, \bar{b}_i^* need not equal b_i^* . The partial correlation between Y and X_i , which is symmetric in the order of the two variables, satisfies

$$r_{YX_i \cdot \text{---}}^2 = b_i^* \bar{b}_i^*.$$

a) From this formula, explain why the partial correlation must fall between b_i^* and \bar{b}_i^* . (Note: When $a = \sqrt{bc}$, a is said to be the *geometric average* of b and c .)

b) Even though b_i^* does not necessarily fall between -1 and $+1$, explain why $b_i^* \bar{b}_i^*$ cannot exceed 1.

68. * Chapters 12 and 13 show how to incorporate categorical predictors in regression models, and this exercise provides a preview. Table 11.21 shows part of a printout for a model for the “house selling price 2” data set at the text website, with Y = selling price of home, X_1 = size of home, and X_2 = whether the house is new (1 = yes, 0 = no).

a) Report the prediction equation. By setting $x_2 = 0$ and then 1, construct the

two separate lines for older and for new homes. Note that the model implies that the slope effect of size on selling price is the same for each.

b) Since x_2 takes only the values 0 and 1, explain why the coefficient of x_2 estimates the difference of mean selling prices between new and older homes, controlling for house size.

Table 11.23:

	B	Std. Error	t	Sig
(Constant)	-26.089	5.977	-4.365	0.0001
SIZE	72.575	3.508	20.690	0.0001
NEW	19.587	3.995	4.903	0.0001

69. * Refer to the previous exercise. When we add an interaction term, we get $\hat{y} = -16.6 + 66.6x_1 - 31.8x_2 + 29.4(x_1x_2)$.

a) Interpret the fit by reporting the prediction equation between selling price and size of house separately for new homes ($x_2 = 1$) and for old homes ($x_2 = 0$). Interpret. (This fit is equivalent to fitting lines separately to the data for new homes and for old homes.)

b) Interpret the fit by reporting the difference between the predicted selling prices for new and old homes for houses with x_1 equal to (i) 1.5, (ii) 2.0, (iii) 2.5.

c) A plot of the data shows an outlier, a new home with a very high selling price. When that observation is removed from the data set and the model is re-fitted, $\hat{y} = -16.6 + 66.6x_1 + 9.0x_2 + 5.0(x_1x_2)$. Re-do (a), and explain how an outlier can have a large impact on a regression analysis.

11.9.1 Bibliography

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

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed., Wiley.

Holzer, C. E., III (1977). *The Impact of Life Events on Psychiatric Symptomatology*. Ph.D. dissertation, University of Florida, Gainesville.

Kutner, M. H., Nachtsheim, C. J., and Neter, J. (2004). *Applied Linear Regression Models*, 4th ed. McGraw-Hill/Irwin.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd ed., Wiley.

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

((Fig. 11.5 from 3e))

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.

((Include new Fig. 11.6))

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.

((Include new Fig. 11.7))

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

((Fig. 11.8 in 3e))

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

((Fig. 11.9 in 3e))

Figure 11.10: Portrayal of Interaction between x_1 and x_2 in their Effects on Y .

((Fig. 11.10 in 3e; if possible, change to lower-case letters))

Figure 11.11: Representation of $r_{Y|X_2, X_1}^2$ as the Proportion of Variability That Can Be Explained by X_2 , of that Left Unexplained by X_1

((Fig 11.11 in 3e))

Figure 11.12:

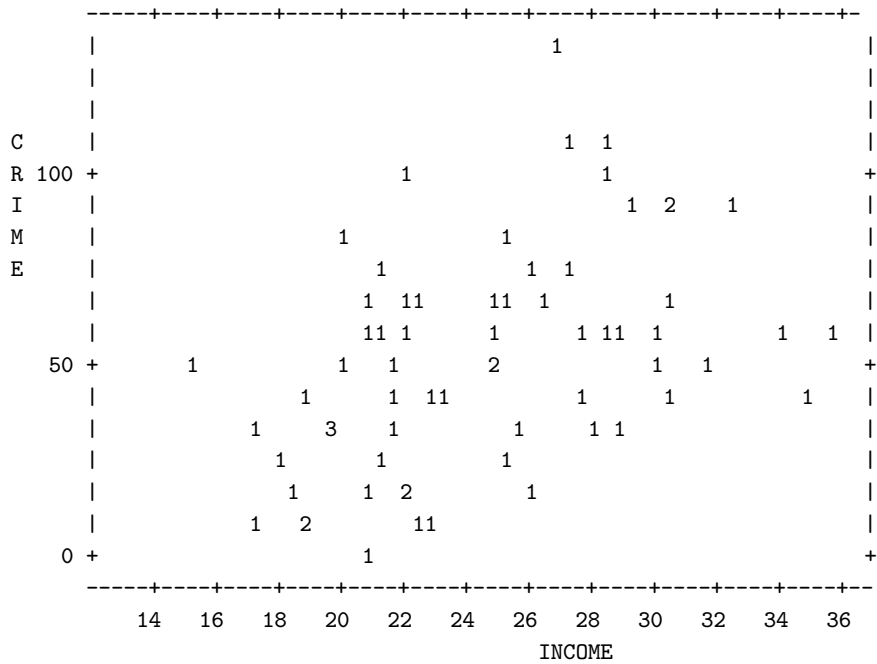


Figure 11.13:

