Chapter 13

Regression Analysis: Inference on Two or More Numerical Variables

It is often necessary to examine the directional relationship between two variables. For example:

- Does the amount of exposure to radioactive materials affect the cancer mortality rate?
- Do SAT scores predict college success?
- Does the speed on highways affect noise level?

Or, it may be necessary to examine the effect of two or more independent variables on one dependent variable. For example:

- Are the number of grams of carbohydrates and of fiber related to the number of calories in breakfast cereal?
- Do SAT scores and high school GPA predict college success?
- Are age, cholesterol level, and amount of exercise a person gets related to his or her chance of having a heart attack?

This chapter describes how to use SPSS for Windows to perform linear regression analysis to estimate statistical relationships from a sample of data. Because a scatter plot and correlation coefficient are indispensable in interpreting regression results, procedures for obtaining these are reviewed as well.

13.1 THE SCATTER PLOT AND CORRELATION COEFFICIENT

Two important steps in regression analysis involve examining a scatter plot of two variables and calculating the correlation coefficient. Although both of these procedures are described in Chapter 5, we will illustrate them here using the "cancer.sav" data file. In this example, we wish to examine the relationship between the amount of exposure to radioactive materials and cancer mortality rate. To create a scatter plot of these variables, open the data file and:

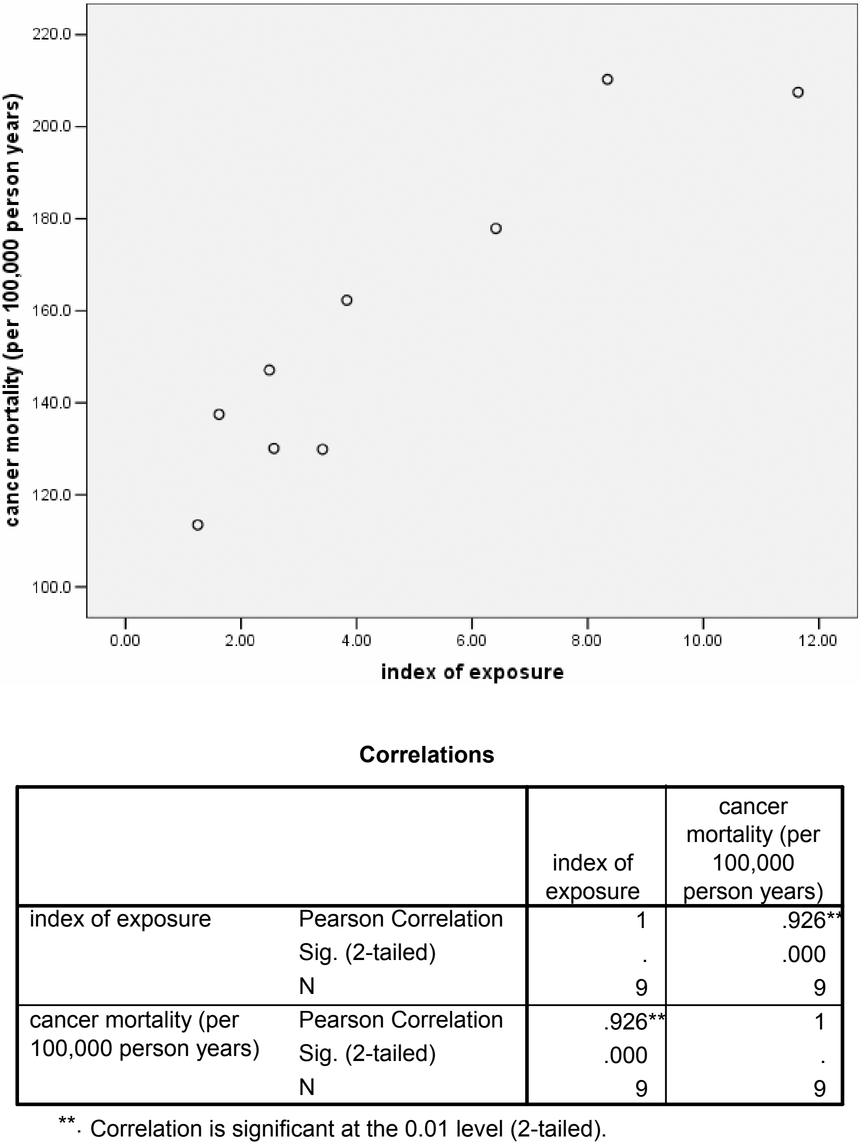
1. Click on **Graphs** from the menu bar.
2. Click on **Scatter/Dot** to open the Scatterplot dialog box.
3. Click on **Simple Scatter** and then on **Define** to open the Simple Scatterplot dialog box.
4. Click on the name of the independent variable (*x*) that you wish to examine ("expose") and move it to the X Axis box using the **second right arrow button**.
5. Click on the name of the dependent variable (*y*) that you wish to examine ("mortalit") and move it to the Y Axis box using the **top right arrow button**.
6. Click on **OK**.

The correlation coefficient can be calculated by using the following commands:

1. Click on **Analyze** from the menu bar.
2. Click on **Correlate** from the pull-down menu.
3. Click on **Bivariate** to open the Bivariate Correlations dialog box.
4. Click on the variable(s) that you wish to correlate, each followed by the **right arrow button** to move them into the Variables box.
5. Click on **OK**.

The output of both procedures appears in Figure 13.1.

The swarm of points in the scatter plot goes from lower left to upper right. We also see that there are no apparent outliers. In addition, the association appears linear, rather than (for instance) curvilinear.



Correlations

		index of exposure	cancer mortality (per 100,000 person years)
index of exposure	Pearson Correlation	1	.926**
	Sig. (2-tailed)	.	.000
	N	9	9
cancer mortality (per 100,000 person years)	Pearson Correlation	.926**	1
	Sig. (2-tailed)	.000	.
	N	9	9

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 13.1 Scatter Plot and Correlation Coefficient of Exposure and Mortality

The correlation between exposure and mortality is +0.926, indicating that it is both positive and strong. Thus, higher levels of exposure to radioactive materials are strongly associated with higher levels of cancer mortality. The *P* value results from a test of significance (*t*-test of the hypothesis that the correlation is zero). The *t*-statistic for this case is not printed, but it is:

$$t = (0.926)\sqrt{\frac{(9-2)}{(1-0.926^2)}} = 6.507$$

The *P* value is obtained by referring this statistic to the t-distribution with 7 degrees of freedom. Here, because *P* < .0005, we conclude that the variables are significantly (positively) related.

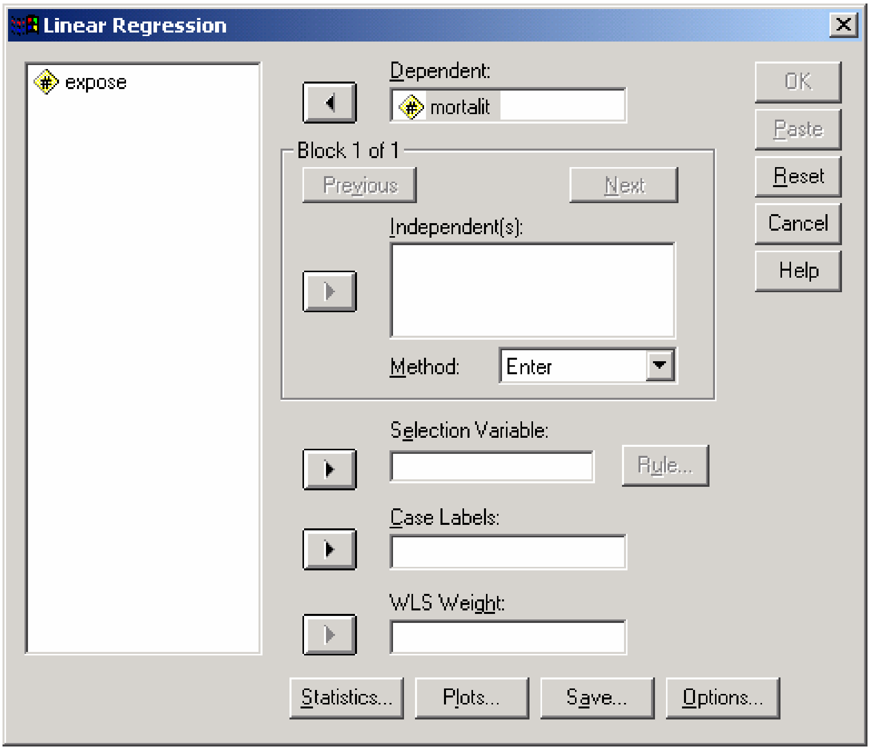
13.2 SIMPLE LINEAR REGRESSION ANALYSIS

In linear regression analysis, sample data are used to estimate the intercept and slope of the "line of best fit" — the regression line — in a scatter plot. Simple regression analysis refers to a situation in which there is one independent and one dependent variable. The equation for the regression line in simple regression is y = α + βx; where α and β are the *y*-intercept and slope, respectively. The slope β is usually of most interest because it tells the number of units increase (or decrease) in the dependent variable (*y*) associated with a one-unit increase in the independent variable (*x*).

We will illustrate this procedure using the "cancer.sav" data set to examine the association between mortality rates and exposure to radioactive materials.

After opening the "cancer.sav" data file, the steps for a regression analysis are:

1. Click on **Analyze** on the menu bar.
2. Click on **Regression** from the pull-down menu.
3. Click on **Linear** to open the Linear Regression dialog box (see Fig. 13.2).
4. Click on the variable that is your dependent variable ("mortalit"), and then click on the **top right arrow button** to move the variable name into the Dependent variable box.
5. Click on the variable that is your independent variable ("expose"), and then click on the **middle right arrow button** to move the variable name into the Independent(s) variable box.
6. Click on the **Statistics** button to open the Linear Regression: Statistics dialog box (Fig. 13.3).
7. By default, the **Estimates** and **Model fit** options are selected. Although we computed the correlation coefficient in a separate procedure, it is possible to calculate is as part of the regression procedure. To do so, select **Descriptives**. Another useful statistic is the confidence interval of the regression coefficients. To obtain this, click on **Confidence intervals**.



expose

Dependent: mortalit

Independent(s):

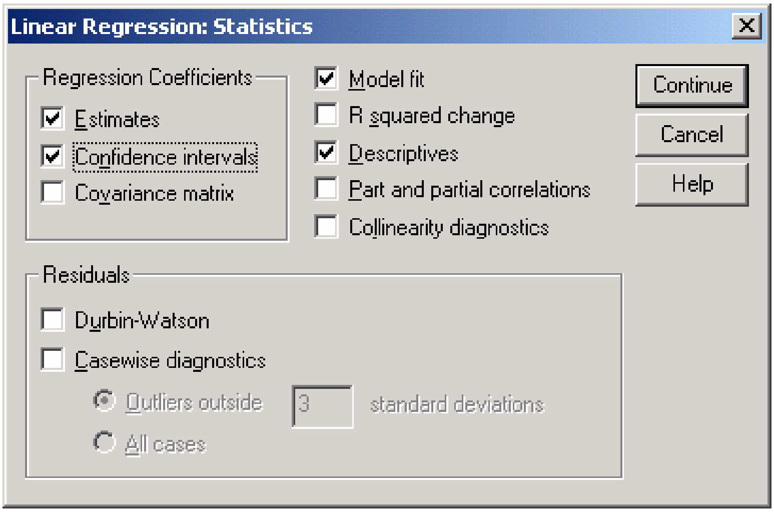
Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Figure 13.2 Linear Regression Dialog Box



Regression Coefficients

☑ Estimates

☑ Confidence intervals

□ Covariance matrix

☑ Model fit

□ R squared change

☑ Descriptives

□ Part and partial correlations

□ Collinearity diagnostics

Residuals

□ Durbin-Watson

□ Casewise diagnostics

● Outliers outside 3 standard deviations

○ All cases

Figure 13.3 Linear Regression: Statistics Dialog Box

8. Click on **Continue.**
9. Click on **OK**.

The complete output is shown in Figure 13.4.

The first two tables — Descriptive Statistics and Correlations — are the result of selecting the **Descriptives** option. We see that the correlation between index of exposure and mortality rate is .926, exactly the same coefficient displayed in Figure 13.1. The only difference between these two tables is that the *P* value in Figure 13.4 is given as a one-tailed *P* value.

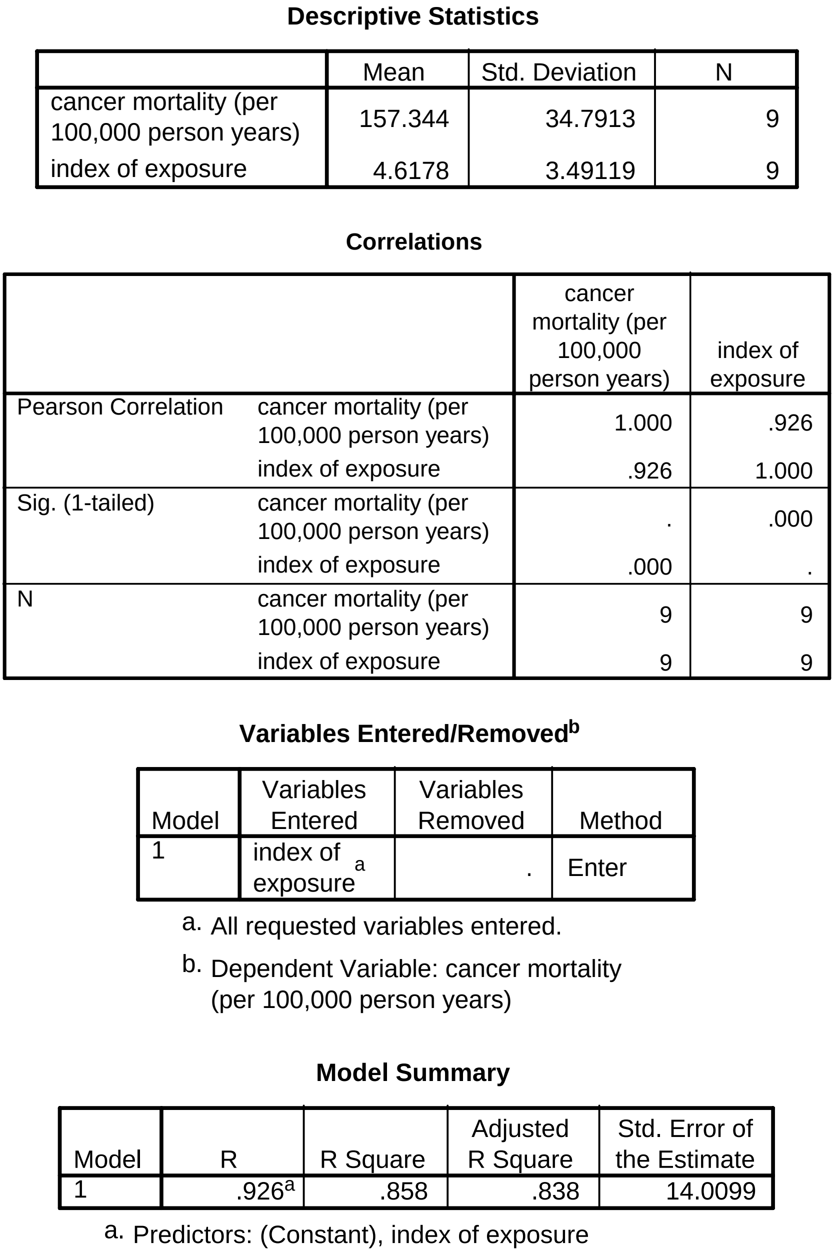
The square of the correlation ($0.926^2 = 0.858$) is the proportion of variation in *y* attributable to *x*; that is, 85.8% of the variation in cancer mortality is attributable to variation in radiation exposure. This is a very strong association.

The strength of association of the independent variable(s) with the dependent variable is also available in the Model Summary Table. This table represents the multiple correlation of the set of all independent variables (predictors) with the dependent variable. Because in simple regression there is only one predictor variable, the simple and multiple correlation coefficients are identical in number. However, the multiple correlation is always positive. The data analyst must remember that the multiple correlation does not indicate the direction of association! The *R* Square in this table represents the 85.8% of variation accounted for in mortality by index of exposure, as discussed previously.

Test of Significance for the Model

The output also includes a table labeled ANOVA (Analysis of Variance) located below the information about multiple correlation. This is a test of the significance of the model (that is, of the set of independent variables in predicting the dependent variable). The null hypothesis states that the set of all independent variables is not significantly related to the dependent variable. The Sig. column represents the *P* value for the test of significance of the model. In this case, $P < .0005$, so we conclude that the independent variable is significantly related to the dependent variable.

The other columns provide the detail and the building blocks from which the *P* value is determined. The sum of squares for Regression (8309.56) divided by the number of degrees of freedom (1) is the Mean Square for Regression (8309.56), which is the numerator of the *F*-ratio. The sum of squares labeled Residual (1373.95) is the sum of squared differences between the predicted values and the actual values of y, that is, the sum of squared deviations of the data around the regression line. These are combined to yield the proportion-of-variation statistic, r^2. The residual sum of squares divided by the number of degrees of freedom (7) is the variance of the residuals, 196.28 in the column labeled Mean Square. The square root of this value, 14.01, is the standard error of estimate $s_{y.x}$. The *F*-ratio is the ratio of these two mean squares,



Descriptive Statistics

	Mean	Std. Deviation	N
cancer mortality (per 100,000 person years)	157.344	34.7913	9
index of exposure	4.6178	3.49119	9

Correlations

		cancer mortality (per 100,000 person years)	index of exposure
Pearson Correlation	cancer mortality (per 100,000 person years)	1.000	.926
	index of exposure	.926	1.000
Sig. (1-tailed)	cancer mortality (per 100,000 person years)	.	.000
	index of exposure	.000	.
N	cancer mortality (per 100,000 person years)	9	9
	index of exposure	9	9

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	index of exposure[a]	.	Enter

a. All requested variables entered.

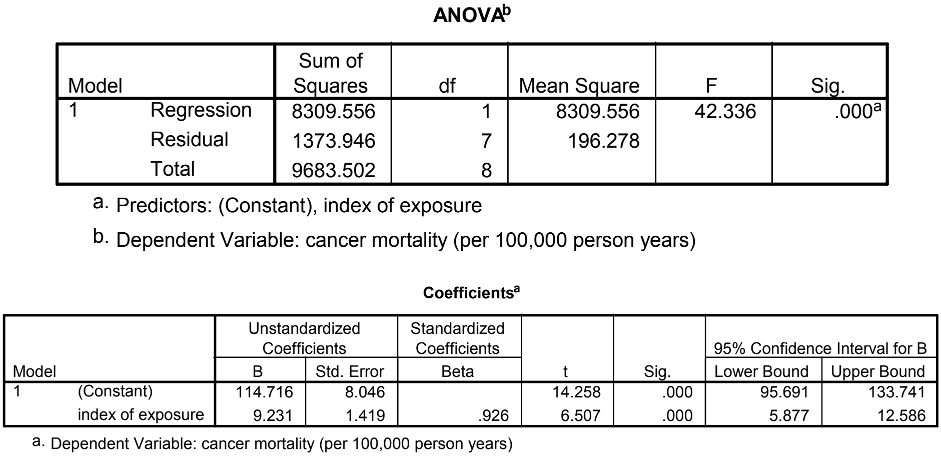
b. Dependent Variable: cancer mortality (per 100,000 person years)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.926[a]	.858	.838	14.0099

a. Predictors: (Constant), index of exposure

Figure 13.4 Regression Analysis Output with Descriptives and Confidence Intervals



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8309.556	1	8309.556	42.336	.000[a]
	Residual	1373.946	7	196.278		
	Total	9683.502	8			

a. Predictors: (Constant), index of exposure

b. Dependent Variable: cancer mortality (per 100,000 person years)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	114.716	8.046		14.258	.000	95.691	133.741
	index of exposure	9.231	1.419	.926	6.507	.000	5.877	12.586

a. Dependent Variable: cancer mortality (per 100,000 person years)

Figure 13.4 Regression Analysis Output with Descriptives and Confidence Intervals*, continued*

$$F = \frac{8309.556}{196.278} = 42.336 \cdot$$

The sum of squares are combined to obtain the proportion of variance in y explained by x, that is, the squared multiple correlation. In this example, this is

$$R^2 = \frac{8309.556}{9683.502} = .858 \cdot$$

Test of Significance for β

A test of the hypotheses H_0: $\beta = 0$ and H_1: $\beta \neq 0$ is given in the Coefficients table of the regression output. The t-statistic is $t = \frac{9.23}{1.418} = 6.507$. The P value, listed under Sig. in the output, is .000. Since this is smaller than most potential values of α (e.g., .05 or .01 or even .001), H_0 is rejected. We conclude that there is a nonzero (positive) association between exposure and cancer mortality in the population of counties represented by this sample.

At the outset of this study, researchers had reason to believe that a positive association might be found. Thus, a one-tailed test would have been appropriate with H_0: $\beta \leq 0$ and H_1: $\beta > 0$. The P value printed by SPSS is for a two-tailed test. To reject H_0 in favor of a one-sided alternative, $P/2$ must be less than α and the sign of the regression weight must be consistent with H_1. Both conditions are met in this example and H_0 is rejected in a one-tailed test as well.

The careful reader may notice that this t value and P are the same in the test of significance for β as they are for the test of the correlation coefficient (Section 13.1). When a study has just one numerical independent variable and one numerical dependent variable, the regression coefficient and the correlation coefficient have the same sign (+ or –) and the tests of significance are identical.

Estimating the Regression Equation

The least-squares estimates of the intercept and slope of the regression line are displayed in the Coefficients table of the output (Fig. 13.4) under the title Unstandardized Coefficients. Two values are listed in the column headed B; these are the intercept (a, 114.716) and the regression weight (b, 9.231), respectively. The equation of the least-squares line is thus $y = 114.716 + 9.231x$. (Instructions for having SPSS add the regression line to the scatter plot are given in a later section.)

SPSS also prints a form of β called the Standardized Coefficient, labeled Beta in the output. The standardized weight is the *number of standard deviations change in y associated with a one-standard deviation change in x*. Thus, in this example, a one-standard deviation increment in exposure is associated with a 0.93-standard deviation increment in cancer mortality — a large effect. When the units of x and y are familiar (e.g., income, time, body weight) the unstandardized ("raw") coefficient is easily interpreted. When scales are in less familiar units (e.g., psychological test scores) the standardized weight is a convenient way to express the relationship of x and y.

The Confidence Interval option produced two 95% intervals in the output, one for the slope and one for the intercept. The interval for the slope indicates that we are 95% confident that a one-unit increment in exposure is associated with an increase in mortality of at least 5.88 deaths per 100,000 person years and perhaps as much as 12.59 additional deaths. These values were obtained by adding to, and subtracting from, 9.23 a multiple of the standard error required to give the preselected confidence interval. In this example, the standard error is 1.42 (see Std. Error in Fig. 13.3), and the multiplier from the t-distribution with 7 degrees of freedom is 2.37.

Drawing the Regression Line

SPSS will draw the least-squares line (regression line) on a scatter plot. This is an option you may request as you create the plot or after doing so. For the latter instance, follow these steps:

1. Double click on the scatter plot chart to open the SPSS Chart Editor.

2. Select (click on) all the data points; they will become highlighted.
3. Click on **Elements** in the menu bar.
4. Select **Fit Line at Total** from the pull-down menu.
5. The Properties dialog box will open, with the option **Linear** fit line selected. Accept this default by clicking **Close**.
6. Click on the **X** in the upper right corner to close the Chart Edit Window.

The output is shown in Figure 13.5.

13.3 ANOTHER EXAMPLE: INVERSE ASSOCIATION OF X AND Y

As another example, open the data file "noise.sav," which has data on the relationship between the acceleration noise of an automobile and the speed for a section of highway. Make a scatter plot (with the regression line superimposed) with acceleration noise as the dependent variable (y) and speed (mph) as the independent variable (x), and perform a simple regression analysis (including the descriptive statistics). Your output from both procedures should look like Figure 13.6.

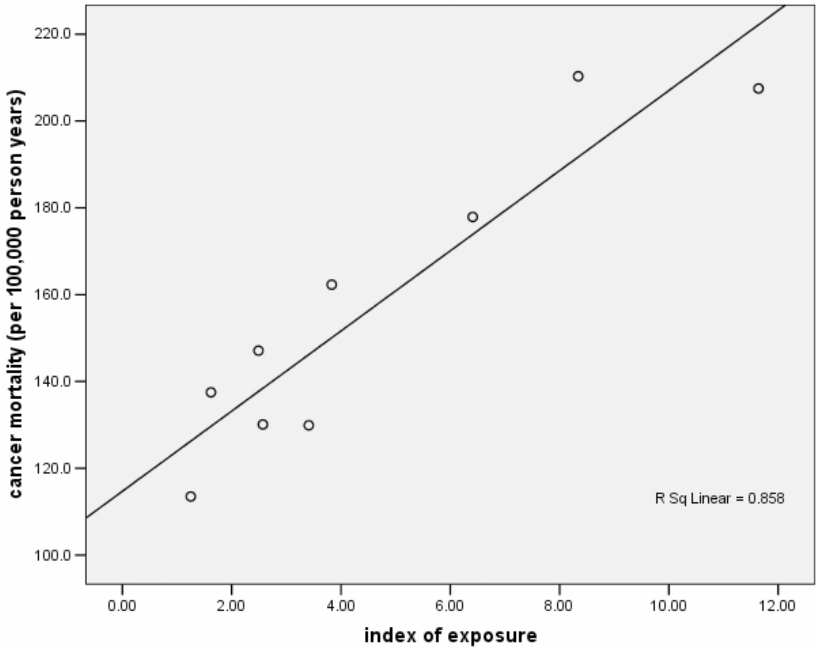
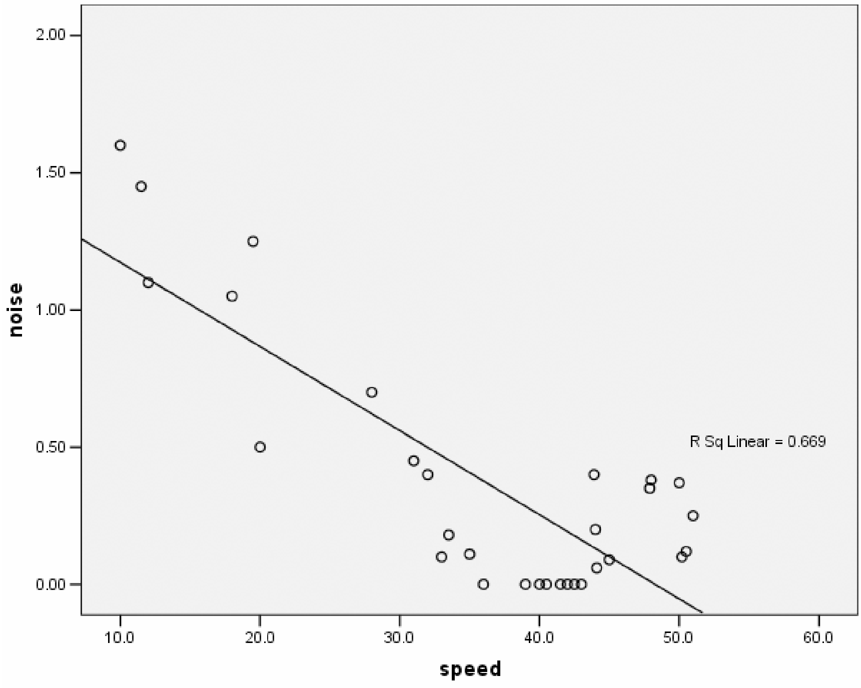


Figure 13.5 Scatter Plot with Regression Line



Descriptive Statistics

	Mean	Std. Deviation	N
NOISE	.3737	.46268	30
SPEED	36.087	12.3453	30

Correlations

		NOISE	SPEED
Pearson Correlation	NOISE	1.000	-.818
	SPEED	-.818	1.000
Sig. (1-tailed)	NOISE	.	.000
	SPEED	.000	.
N	NOISE	30	30
	SPEED	30	30

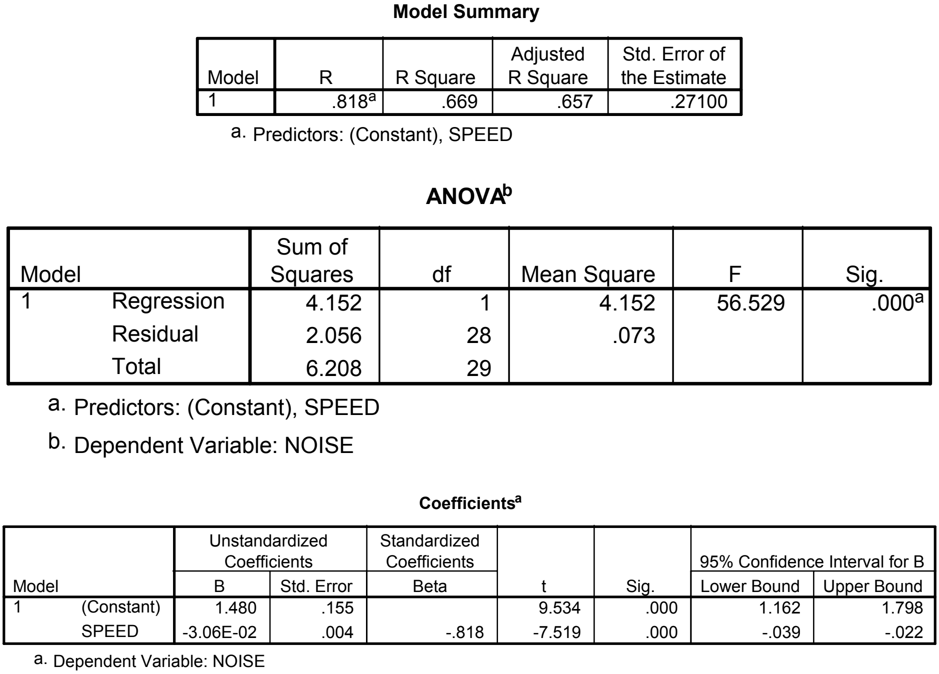
Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	SPEED[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: NOISE

Figure 13.6 Scatter Plot and Regression Analysis of Noise and Speed



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.818[a]	.669	.657	.27100

a. Predictors: (Constant), SPEED

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.152	1	4.152	56.529	.000[a]
	Residual	2.056	28	.073		
	Total	6.208	29			

a. Predictors: (Constant), SPEED

b. Dependent Variable: NOISE

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1.480	.155		9.534	.000	1.162	1.798
	SPEED	-3.06E-02	.004	-.818	-7.519	.000	-.039	-.022

a. Dependent Variable: NOISE

Figure 13.6 Scatter Plot and Regression Analysis of Noise and Speed, *continued*

Notice that there is an inverse relationship between speed and acceleration noise. We first see this with the negative correlation (–0.818), indicating a strong, negative linear association. The inverse relationship is also indicated by the negative slope of the line in the scatter plot. The scatter plot also shows that there are no apparent outliers.

The test of whether the regression weight is significantly different from zero appears in the Coefficients table in the regression output. The sample regression weight is –0.0306. The *t*-statistic is $t = \frac{-0.0306}{0.004} = -7.519$ where 0.004 is the standard error of the regression coefficient. The *P* value given under the label Sig. is .000, which implies that $P < .0005$; noise is significantly inversely related to speed on the highway.

Given that we have a significant negative association of noise with speed, we ask about the strength of the relationship. Because speed is measured in familiar units (miles per hour, or mph), we may prefer to interpret the unstandardized regression weight (labeled B in the output). This tells us that every 1 mph increase in average speed on sections of the highway is associated with a .031-unit decrease in acceleration noise.

We have already seen that the correlation between speed and noise is strong and negative. In addition, the output produced by the regression analysis shows that the square of the correlation is 0.669. (Recall that the Multiple R is the absolute value of the correlation.) Thus, 66.9% of the variability in noise level is accounted for by speed and, by subtraction from 100%, 33.1% of variation in noise level is explained by other factors that were not included in this study.

No Relationship

Does it ever happen that a predictor variable is not related to the dependent variable? Yes, it does. This can be illustrated with the data in file "weather.sav" by examining the relationship between the amount of precipitation in an area and the temperature. Again, construct a scatter plot of the variables amount of precipitation ("precip") and temperature ("temp") (with the regression line) and compute the correlation coefficient as described in Section 13.1. The output is displayed in Figure 13.7.

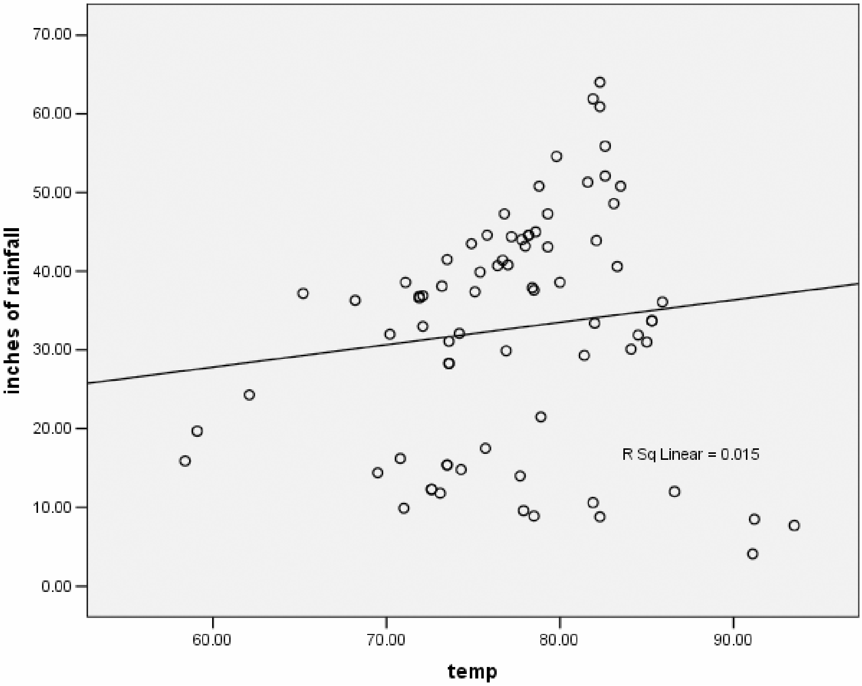
The regression line appears to be rather flat and many of the points are far away from it. The correlation itself is small ($r = 0.124$) and nonsignificant ($P = .284$); less than 2% of the variability in "precipitation" is attributed to "temperature" ($0.124^2 = 0.015$). There is little to be gained by attempting to predict precipitation from temperature.

13.4 MULTIPLE REGRESSION ANALYSIS

The goal of multiple regression analysis is to explore how much variation in the dependent variable can be explained by variability in two or more independent variables. The equation of the regression line for two independent variables is: $y = \alpha + \beta_1 x_1 + \beta_2 x_2$. SPSS provides estimates and tests of the two most important parameters, β_1 and β_2, which determine the slope of the line; they reflect the "partial" contribution each independent variable has toward explaining the outcome (y).

Selecting the Order of Entry of the Independent Variables

When a regression analysis includes two or more independent variables, the order in which they are entered into the analysis is an important consideration. There are several approaches to entering the variables, depending on both the purpose of the research and philosophy of the researcher. In one approach, the researcher decides the order of entry according to the conceptual importance of the independent variables. This “hierarchical” procedure produces tests of the first independent variable, the *additional* contribution of the second independent variable, and so on. Other approaches use the strength of the partial correlations of the variables in the sample to determine which independent measures are included at each step. These procedures minimize the number of predictors in the final model. The examples in this chapter use the former (hierarchical) approach.



Correlations

		temp	inches of rainfall
temp	Pearson Correlation	1	.124
	Sig. (2-tailed)		.284
	N	77	77
inches of rainfall	Pearson Correlation	.124	1
	Sig. (2-tailed)	.284	
	N	77	77

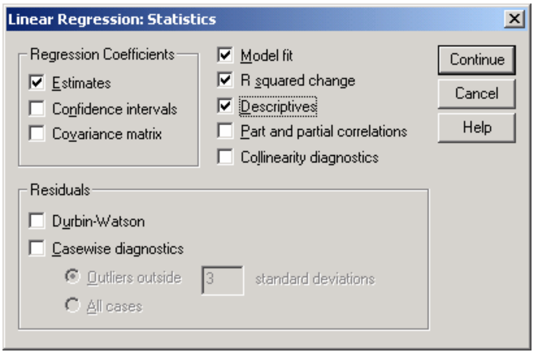
Figure 13.7 Scatter Plot and Correlation Coefficient of Precipitation and Weather

We will illustrate the multiple regression procedure using the “cereal.sav” data set. Our interest is in examining the relationship between amount of fiber and carbohydrates in cereals and the number of calories per serving. Specifically, we wish to examine the additional effect of variations in fiber, over and above that of carbohydrate differences.

After opening the “cereal.sav” data file, do the following:

1. Click on **Analyze** on the menu bar.
2. Click on **Regression** from the pull-down menu.
3. Click on **Linear** to open the Linear Regression dialog box (see Fig. 13.2).
4. Click on the variable that is your dependent variable (“calories”), and then click on the **top right arrow button** to move the variable name into the Dependent variable box.
5. Click on the independent variable that you want to enter first into the model (“carbo”), and then click on the **second right arrow button** to move the variable name into the Independent(s) variable box.
6. Click on the **Next** button located above the Independent(s) box. The block should change to 2 of 2.
7. Click on the second independent variable (“fiber”), and then click on the **second right arrow button** to move the variable name into the Independent(s) variable box.
8. Click on Statistics to open the Linear Regression: Statistics dialog box (see Fig. 13.8).
9. Click on the **R squared change** and **Descriptives** (in addition to the default **Estimates** and **Model fit**).
10. Click on **Continue** to close the dialog box.
11. Click on **OK** to run the regression.

The complete output is shown in Figure 13.9.



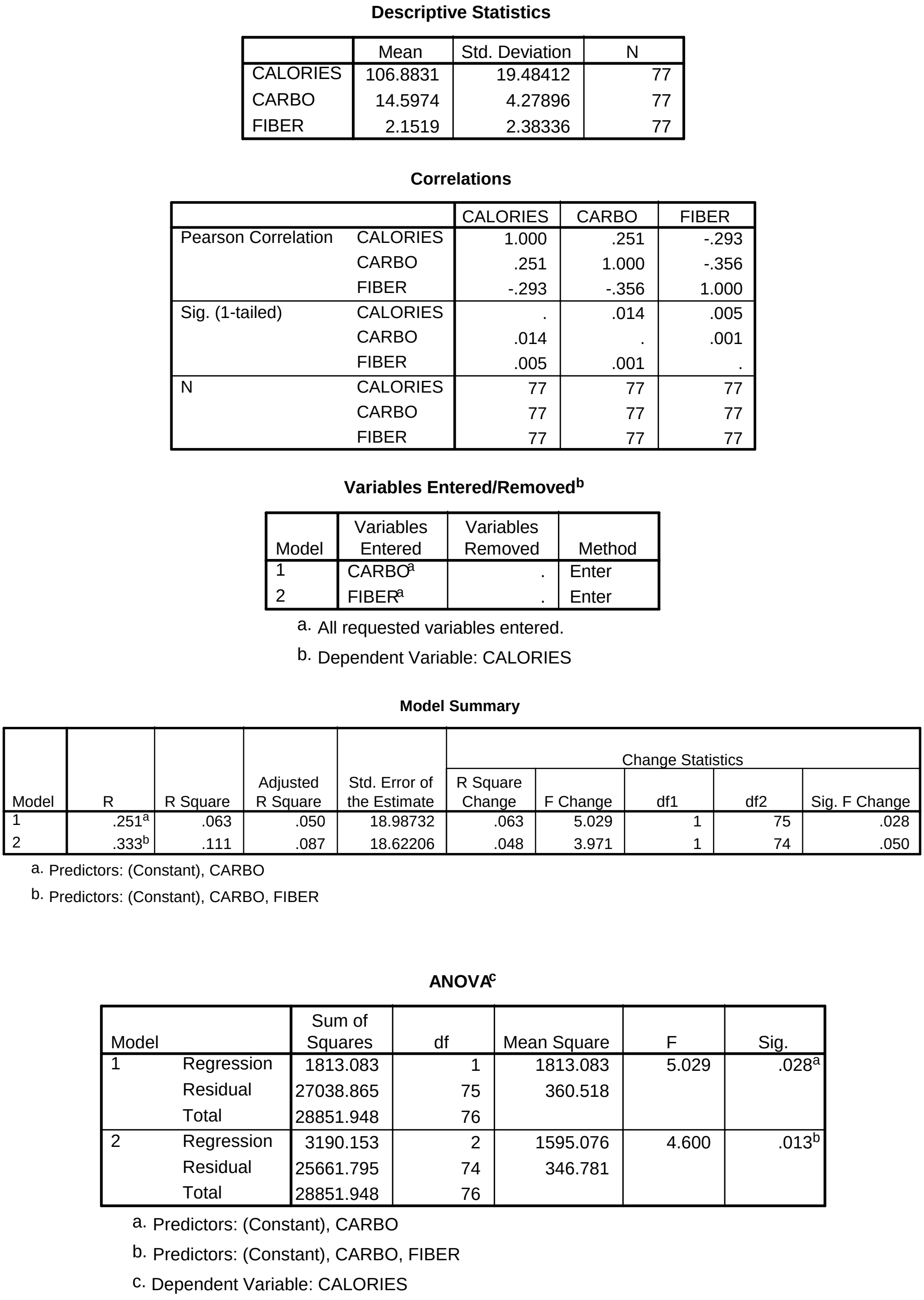
Linear Regression: Statistics

Regression Coefficients
☑ Estimates
□ Confidence intervals
□ Covariance matrix

☑ Model fit
☑ R squared change
☑ Descriptives
□ Part and partial correlations
□ Collinearity diagnostics

Residuals
□ Durbin-Watson
□ Casewise diagnostics
● Outliers outside 3 standard deviations
○ All cases

Figure 13.8 Linear Regression: Statistics Dialog Box



Descriptive Statistics

	Mean	Std. Deviation	N
CALORIES	106.8831	19.48412	77
CARBO	14.5974	4.27896	77
FIBER	2.1519	2.38336	77

Correlations

		CALORIES	CARBO	FIBER
Pearson Correlation	CALORIES	1.000	.251	-.293
	CARBO	.251	1.000	-.356
	FIBER	-.293	-.356	1.000
Sig. (1-tailed)	CALORIES	.	.014	.005
	CARBO	.014	.	.001
	FIBER	.005	.001	.
N	CALORIES	77	77	77
	CARBO	77	77	77
	FIBER	77	77	77

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	CARBO[a]	.	Enter
2	FIBER[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: CALORIES

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.251[a]	.063	.050	18.98732	.063	5.029	1	75	.028
2	.333[b]	.111	.087	18.62206	.048	3.971	1	74	.050

a. Predictors: (Constant), CARBO

b. Predictors: (Constant), CARBO, FIBER

ANOVA[c]

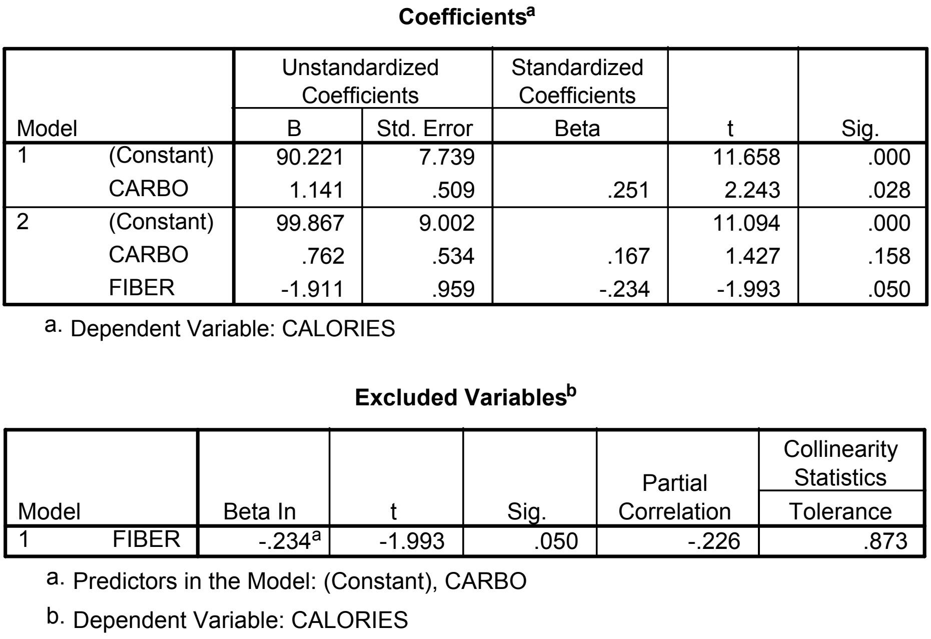
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1813.083	1	1813.083	5.029	.028[a]
	Residual	27038.865	75	360.518		
	Total	28851.948	76			
2	Regression	3190.153	2	1595.076	4.600	.013[b]
	Residual	25661.795	74	346.781		
	Total	28851.948	76			

a. Predictors: (Constant), CARBO

b. Predictors: (Constant), CARBO, FIBER

c. Dependent Variable: CALORIES

Figure 13.9 Multiple Regression Analysis Output



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	90.221	7.739		11.658	.000
	CARBO	1.141	.509	.251	2.243	.028
2	(Constant)	99.867	9.002		11.094	.000
	CARBO	.762	.534	.167	1.427	.158
	FIBER	-1.911	.959	-.234	-1.993	.050

a. Dependent Variable: CALORIES

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	FIBER	-.234[a]	-1.993	.050	-.226	.873

a. Predictors in the Model: (Constant), CARBO

b. Dependent Variable: CALORIES

Figure 13.9 *Continued*

Simple Correlations

The simple correlations between the independent and dependent variables are an important part of the results. These are shown in the Correlations table of the output. We see that there is a positive correlation between carbohydrates and calories (.251) and a negative correlation between fiber and calories (–.293). So, as the amount of carbohydrates in cereals increase, so does the number of calories. Conversely, cereals higher in fiber tend to have fewer calories. The output also gives *P* values for testing significance of each correlation. These are, however, simple correlations, and they do not include any consideration of the multiplicity of the independent variables.

The Full Model

The tests of significance for the full model are contained in the ANOVA table. The table lists two separate models because the variables were entered into the equation in separate steps. Model 1 corresponds to the simple model with calories regressed on carbohydrates. Model 2 represents the multiple regression with both carbohydrates and fiber as the independent variables. The *P* value is contained in the Sig. column. We see that *as a set*, the two predictors (carbohydrates and fiber) are significantly related to the number of calories in a cereal ($P < .013$).

The Model Summary table lists the multiple correlation for the two models. The squared correlation (R^2) indicated that carbohydrates alone account for 6.3% of variation in calories. For the full model, differences in fiber *and* carbohydrate content account for 11.1% of the variation in number of calories in cereals.

The Coefficients table lists the partial regression coefficients in raw and standardized form. From Model 2 we see that the regression equation is: CALORIES = 99.867 + .762 (CARBOHYDRATES) – 1.911 (FIBER). Each regression coefficient reflects the effect of the particular variable "above and beyond" (holding constant) the effect of the other independent variable(s). The coefficient for fiber is negative, indicating that independent of carbohydrates, higher fiber cereals have lower calories. That is, holding constant the amount of carbohydrates in a cereal, a one-gram increase in fiber content is associated with a decrease of 1.9 calories, on average.

The Sig. column contains the *P* values. Using a .05 significance level, the amount of fiber in a cereal is significantly related to calories, holding constant the amount of carbohydrates. Carbohydrates, on the other hand, are not related to calories, after controlling for fiber.

Incremental Models

On many occasions, the researcher is interested in examining the contribution of the independent variables to variation in the dependent variable in a particular order, that is, the first independent variable, the second above and beyond the first, and so on. An easy place to find these results is in the Model Summary table. It lists the proportion of variance accounted for (R Square Change), *F*-statistic (*F* Change), and *P* value (Sig. F Change) for each variable as it is entered. For instance, the squared multiple correlation for the first model (the model with carbohydrates as the only independent variable) is .063. (This is equal to the square of the simple correlation shown in the Correlations table.) The R^2 indicates that 6.3% of the variation in calories in cereals is explained by differences in the grams of carbohydrates. The *F* statistic (5.029) is significant at the .05 level ($P < .028$).

Model 2 shows that differences in fiber explain an *additional* 4.8% of the variation in calories in breakfast cereals (R Square Change). The additional explanatory power is significant at $P < .050$. Together, the two independent variables account for 11.1% of variation in calories (6.3% + 4.8%), or 100 times the squared multiple correlation for the full model.

Equivalently, the incremental effects may be found in the coefficients table. The regression weight for Model 1 (1.141) yields a *t*-statistic of 2.243 and *P* value of .028. The square of *t* (2.243^2) is 5.029, the *F*-statistic in the Model summary table; the *P* value is identical. The regression weight for FIBER in model 2 (–1.1911) yields a *t*-statistic of –1.993 and *P* value of .050. The square of *t* (1.993^2) is 3.971, the *F*-statistic in the Model Summary table; the *P* value is identical. The Model Summary table is the convenient place to find the changes in R^2 and tests for a hierarchical analysis. The Coefficients table is useful because it contains raw and standardized weights that also indicate the direction of the effects (positive or negative).

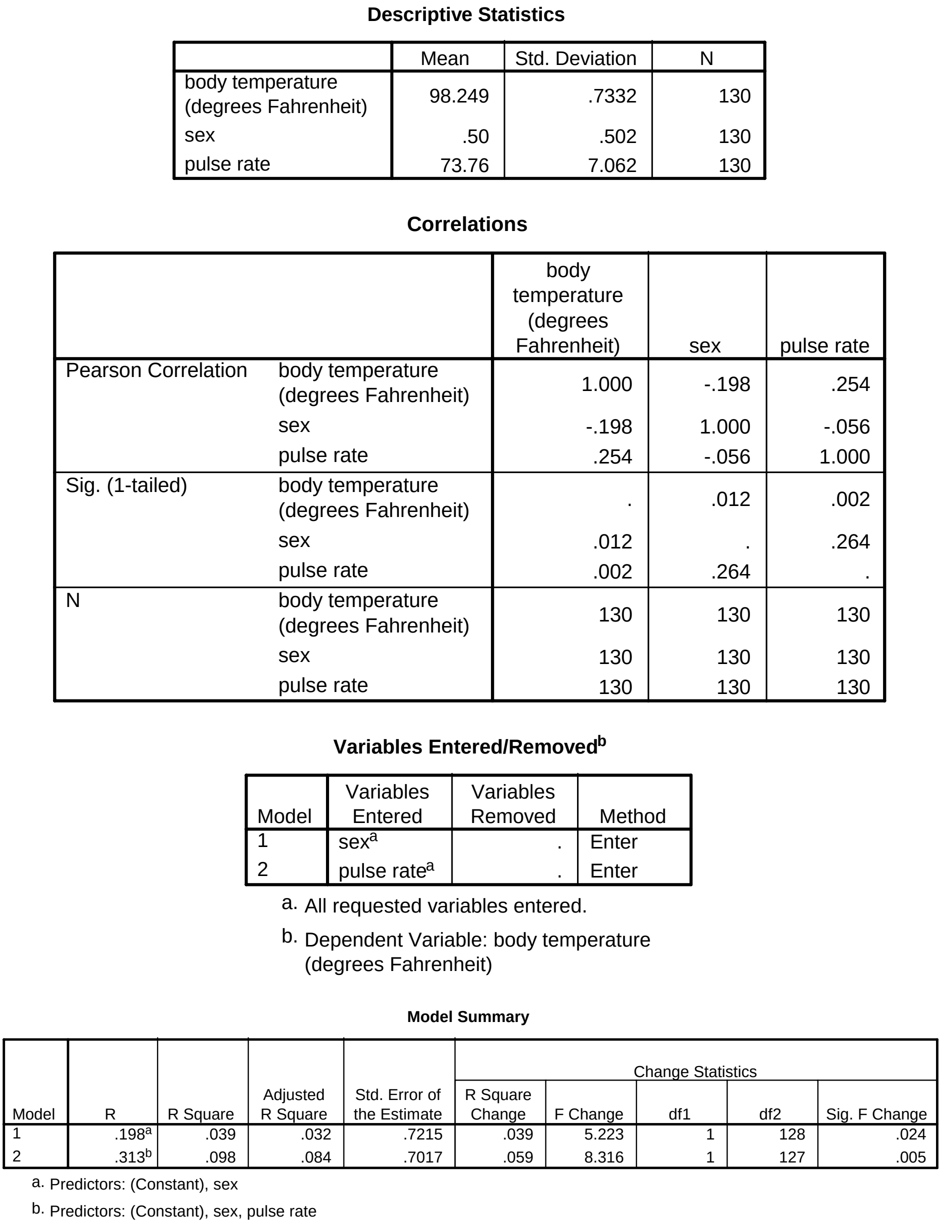
13.5 AN EXAMPLE WITH DUMMY CODING

Although regression requires numerical variables, it is possible to use categorical variables, especially dichotomous variables, as the independent variables. This is accomplished through a process called “dummy coding.” Dummy coding for dichotomous variables involves recoding the variables into the values 0 and 1. If gender were the variable, for instance, females could be coded 0 and males 1 (or vice versa).

We illustrate this with the “bodytemp.sav” data file. This file contains information on body temperature, pulse rate, and sex (0 = female, 1 = male) for 130 adults. We perform a multiple regression analysis following the steps in Section 13.4 to determine whether pulse rate and gender are related to body temperature. We enter sex first, and pulse rate second. The output is given in Figure 13.10.

The simple correlations indicate that sex is negatively correlated with body temperature ($r = -.198$). Because sex is coded 0 = female and 1 = male, the negative correlation indicates that females have higher body temperature than do males (the correlation is statistically significant at the .05 level, with $P < .012$). The correlation matrix also indicates that pulse rate is positively correlated with body temperature ($r = .254$, $P < .002$).

The Model 2 of the ANOVA table indicates that together, the variables are significantly related to body temperature ($P < .001$). The change statistics (in the Model Summary table) show that sex is related to temperature ($P < .024$), and that it explains 3.9% of the variation in temperature. Pulse rate is also related to body temperature, after controlling for sex differences ($P < .005$). Pulse rate accounts for an additional 5.9% of variation in body temperature, over and above that explained by sex.



Descriptive Statistics

	Mean	Std. Deviation	N
body temperature (degrees Fahrenheit)	98.249	.7332	130
sex	.50	.502	130
pulse rate	73.76	7.062	130

Correlations

		body temperature (degrees Fahrenheit)	sex	pulse rate
Pearson Correlation	body temperature (degrees Fahrenheit)	1.000	-.198	.254
	sex	-.198	1.000	-.056
	pulse rate	.254	-.056	1.000
Sig. (1-tailed)	body temperature (degrees Fahrenheit)	.	.012	.002
	sex	.012	.	.264
	pulse rate	.002	.264	.
N	body temperature (degrees Fahrenheit)	130	130	130
	sex	130	130	130
	pulse rate	130	130	130

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	sex[a]	.	Enter
2	pulse rate[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: body temperature (degrees Fahrenheit)

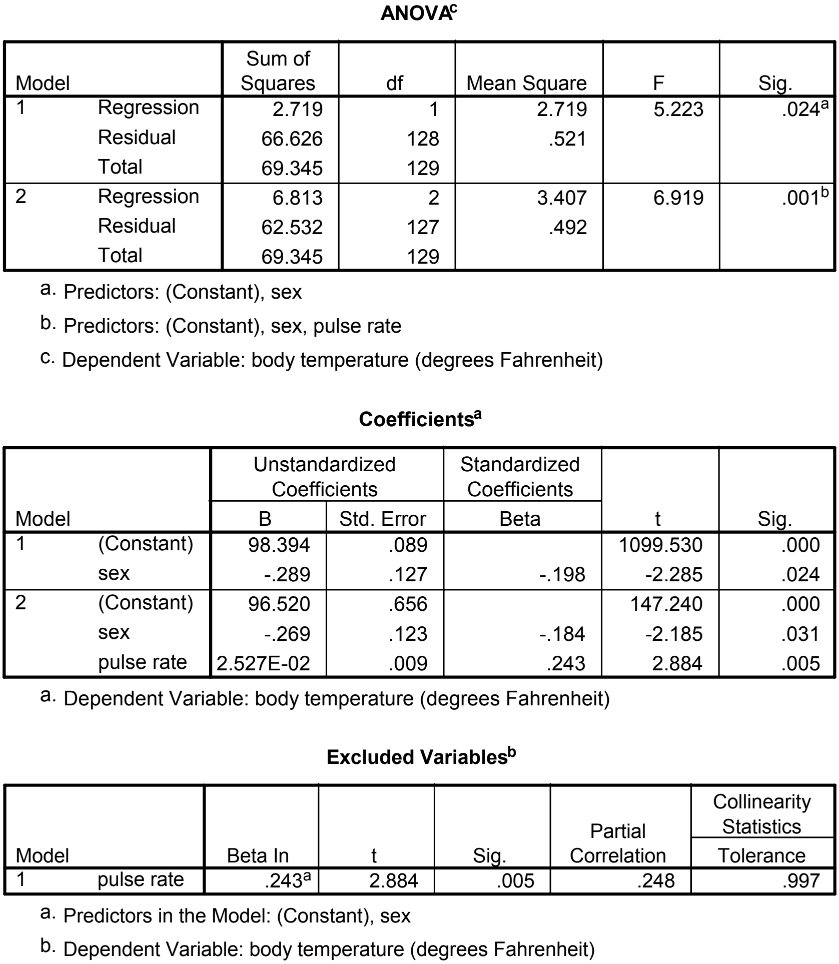
Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.198[a]	.039	.032	.7215	.039	5.223	1	128	.024
2	.313[b]	.098	.084	.7017	.059	8.316	1	127	.005

a. Predictors: (Constant), sex

b. Predictors: (Constant), sex, pulse rate

Figure 13.10 SPSS Regression Output



ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.719	1	2.719	5.223	.024[a]
	Residual	66.626	128	.521		
	Total	69.345	129			
2	Regression	6.813	2	3.407	6.919	.001[b]
	Residual	62.532	127	.492		
	Total	69.345	129			

[a] Predictors: (Constant), sex

[b] Predictors: (Constant), sex, pulse rate

[c] Dependent Variable: body temperature (degrees Fahrenheit)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	98.394	.089		1099.530	.000
	sex	-.289	.127	-.198	-2.285	.024
2	(Constant)	96.520	.656		147.240	.000
	sex	-.269	.123	-.184	-2.185	.031
	pulse rate	2.527E-02	.009	.243	2.884	.005

[a] Dependent Variable: body temperature (degrees Fahrenheit)

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	pulse rate	.243[a]	2.884	.005	.248	.997

[a] Predictors in the Model: (Constant), sex

[b] Dependent Variable: body temperature (degrees Fahrenheit)

Figure 13.10 *Continued*

Chapter Exercises

13.1 Using the “library.sav” data file:

- **a.** Use SPSS to make a scatter plot of the variables “volumes” and “staff,” and draw the least-squares line through the scatter plot.
- **b.** Is the relationship between libraries’ collection size and staff size positive or negative?

c. Judging from the scatter plot, would you estimate that the correlation is weak, moderate, or strong?

13.2 Using the data in "cereal.sav" use SPSS to:

a. Perform a regression analysis to examine whether the sugar content in cereal is related to rating of taste.

b. Compute the slope, the standard error of the slope, and a 95% confidence interval for the slope of the regression line.

c. Is there a significant relationship between sugar content and rating? State the test statistic and *P* value and explain the nature of the relationship, if one exists.

d. What is the value of the correlation between sugar content and rating?

e. What is the square of the correlation and what is its interpretation?

13.3 Open the "fire.sav" data file and use the variables "sex" and "agility" to perform the following:

a. Use SPSS to recode the sex variable so that males have a value of 1 and females have a value of 0. (Note: this is called "dummy coding" a variable.)

b. Perform a regression analysis with the agility score as the dependent variable and sex, after recoding, as the independent variable.

c. Is there a statistically significant relationship of agility with sex? What is the direction of the relationship? Be clear about which "sex" code is associated with better (lower) agility scores.

d. Calculate the mean agility scores for males and for females and the difference between the two means. Compare this difference with the regression weight you obtained.

e. Given the choice between the raw and standardized regression coefficient, which would you choose to emphasize in a written report of this study? Why?

13.4 Use the data in "sleep.sav" to perform a multiple regression analysis to examine whether the number of hours of dream sleep ("dream") is related to animals' likelihood of being preyed upon ("prey") and the exposure of the den during sleep ("sleepexp").

a. Is the overall relationship of predation and den exposure associated with the amount of time spent dreaming?

b. Are both predictors significant?

c. What is the strength of the overall relationship?

d. Redo this analysis by first including the body weight variable. How does the overall relationship change with three predictors? After controlling for body weight, are the predation and den exposure indices related to dream time?