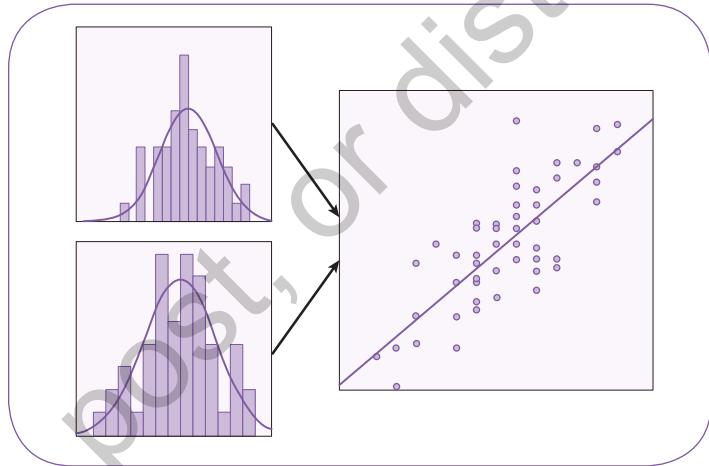


CHAPTER 8

Correlation and Regression— Pearson and Spearman

Correlation and regression show the relationship between continuous variables.



He who laughs most, learns best.

—John Cleese

LEARNING OBJECTIVES

Upon completing this chapter, you will be able to:

- Determine when it is appropriate to run Pearson regression and Spearman correlation analyses
- Interpret the direction and strength of a correlation
- Verify that the data meet the criteria for running regression and correlation analyses: normality, linearity, and homoscedasticity
- Order a regression analysis: correlation and scatterplot with regression line
- Interpret the test results
- Resolve the hypotheses
- Document the results in plain English
- Understand the criteria for causation: association/correlation, temporality, and nonspurious
- Differentiate between correlation and causation



VIDEOS

The videos for this chapter are **Ch 08 - Correlation and Regression - Pearson.mp4** and **Ch 08 - Correlation and Regression - Spearman.mp4**. These videos provide overviews of these tests, instructions for carrying out the pretest checklist, running the tests, and interpreting the results using the data sets **Ch 08 - Example 01 - Correlation and Regression - Pearson.sav** and **Ch 08 - Example 02 - Correlation and Regression - Spearman.sav**.



OVERVIEW—PEARSON CORRELATION

Regression involves assessing the correlation between two variables. Before proceeding, let us deconstruct the word *correlation*: The prefix *co* means *two*—hence, correlation is about *the relationship between two things*. *Regression* is about statistically assessing the correlation between two continuous variables.

Correlation involving two variables, sometimes referred to as *bivariate correlation*, is notated using a lowercase *r* and has a value between -1 and $+1$. Correlations have two primary attributes: direction and strength.

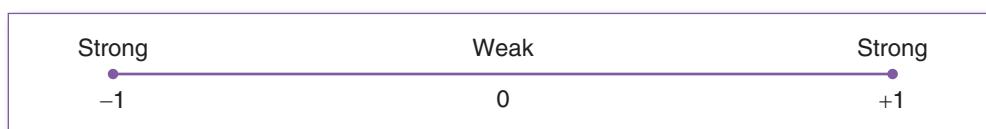
Direction is indicated by the sign of the *r* value: $-$ or $+$. *Positive correlations* ($r = 0$ to $+1$) emerge when the two variables move in the same direction. For example, we would expect that low homework hours would correlate with low grades, and high homework hours would correlate with high grades. *Negative correlations* ($r = -1$ to 0) emerge when the two variables move in different directions. For example, we would expect that high alcohol consumption would correlate with low grades, just as we would expect that low alcohol consumption would correlate with high grades (see Table 8.1).

Strength is indicated by the numeric value. A correlation wherein the *r* is close to 0 is considered weaker than those nearer to -1 or $+1$ (see Figure 8.1). Continuing with the

Table 8.1 Correlation Direction Summary.

Correlation	<i>r</i>	Variable Directions
Positive	0 to $+1$	$X \uparrow Y \uparrow$ or $X \downarrow Y \downarrow$
Negative	-1 to 0	$X \uparrow Y \downarrow$ or $X \downarrow Y \uparrow$

Figure 8.1 Correlation strength.



prior example, we would expect to find a strong positive correlation between homework hours and grade (e.g., $r = +.80$); conversely, we would expect to find a strong negative correlation between alcohol consumption and grade (e.g., $r = -.80$). However, we would not expect that a variable such as height would have much to do with academic performance, and hence we would expect to find a relatively weak correlation between height and grade (e.g., $r = +.02$ or $r = -.02$).

The concepts of correlation direction and strength will become clearer as we examine the test results, specifically upon inspecting the graph of the scatterplot with the regression line in the **Results** section.

In cases where the three pretest criteria are not satisfied for the Pearson test, the Spearman test, which is conceptually similar to the Pearson test, is the better option. Additionally, the Spearman test has some other uses, which are explained near the end of this chapter.



EXAMPLE 1—PEARSON REGRESSION

An instructor wants to determine if there is a relationship between how long a student spends taking a final exam (2 hours are allotted) and his or her grade on the exam (students are free to depart upon completion).

Research Question

Is there a correlation between how long it takes for a student to complete an exam and the grade on that exam?

Groups

Bivariate regression/correlation involves only one group, but two different continuous variables are gathered from each participant: In this case, the variables are (a) *time* taking the exam and (b) the *grade* on the exam.

Notice that in correlation analysis, you can mix apples and oranges; *time* is a measure of *minutes*, whereas *grade* is a measure of *academic performance*. The only constraints in this respect are that the two metrics must both be continuous variables, and of course, the comparison needs to inherently make sense. Whereas it is reasonable to consider the correlation between the amount of *time* a student spent taking an exam and the *grade* on that exam, it is implausible to assess the correlation between *shoe size* and exam *grade*, even though *shoe size* is a continuous variable.



Procedure

The instructor briefs the students that they are welcome to quietly leave the room upon completing the exam. At the start of the exam, the instructor will start a stopwatch. When each student hands in his or her exam, the instructor refers to the stopwatch and records the time (in minutes) on the back of each exam.

Hypotheses

H_0 : There is no correlation between the length of time spent taking the exam and the grade on the exam.

H_1 : There is a correlation between the length of time spent taking the exam and the grade on the exam.



Data Set

Use the following data set: **Ch 08 - Example 01 - Correlation and Regression - Pearson.sav**.

Codebook

Variable: name

Definition: Student's last name

Type: Alphanumeric

Variable: time

Definition: Number of minutes the student spent taking the exam

Type: Continuous (0 to 120) [2 hours = 120 minutes]

Variable: grade

Definition: Grade on exam

Type: Continuous (0 to 100)



Pretest Checklist

Correlation and Regression Pretest Checklist

1. Normality^a
2. Linearity^b
3. Homoscedasticity^b

a. Run prior to correlation and regression test.

b. Results produced upon correlation and regression test run.

The pretest criteria for running a correlation/regression involve checking the data for (a) **normality**, (b) **linearity**, and (c) **homoscedasticity** (pronounced *hoe-moe-skub-daz-tis-city*).



Pretest Checklist Criterion 1—Normality

The two variables involved in the correlation/regression each need to be inspected for normality. To do this, generate separate histograms with normal curves for *time* and *grade* (this is similar to the steps used to check for normality when using the *t* test and ANOVA [analysis of variance]).

For more details on this procedure, refer to Chapter 4 (“Descriptive Statistics”); see the star (★) icon on page 72 and follow the procedure in the section “SPSS—Descriptive Statistics: Continuous Variables (*Age*)”; instead of processing *age*, load the two variables *time* and *grade*. Alternatively, the following steps will produce histograms with a normal curve for *time* and *grade*:

1. From the main screen, select *Analyze, Descriptive Statistics, Frequencies*; this will take you to the *Frequencies* window.
2. On the *Frequencies* window, move *time* and *grade* from the left panel to the right (*Variables*) panel. This will order histograms for both variables at the same time.
3. Click on the *Charts* button; this will take you to the *Charts* window.
4. Click on the *Histograms* button, and check the *Show normal curve on histogram* checkbox.
5. Click on the *Continue* button; this will return you to the *Frequencies* window.
6. Click on the *OK* button, and the system will produce (two) histograms with normal curves for *time* and *grade* (Figures 8.2 and 8.3).

Figure 8.2 Histogram with normal curve for *time*.

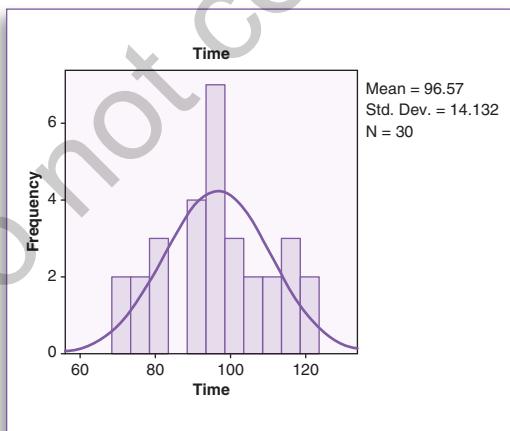
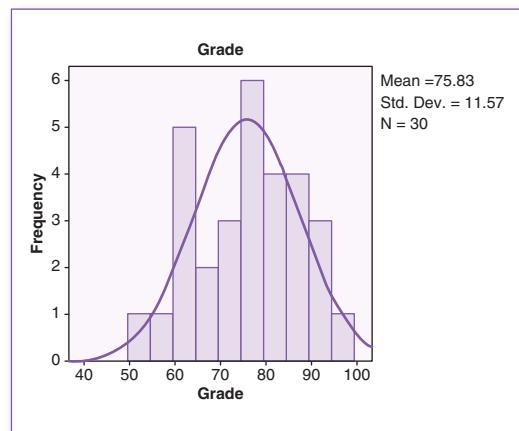


Figure 8.3 Histogram with normal curve for *grade*.



The curves on each histogram are reasonably symmetrically bell shaped; there is no notable skewing, and hence these criteria are satisfied.

The remaining two pretest criteria, **linearity** and **homoscedasticity**, are graphical in nature; they will be processed during the **Test Run** and finalized in the **Results** section.



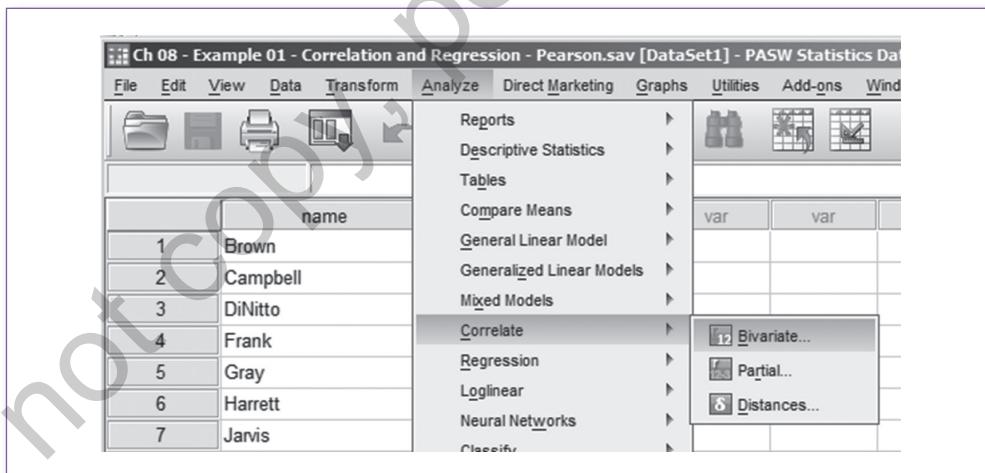
Test Run

The test run for correlation and regression involves two steps: First we will process the correlation table, which will render the correlation value (r) and the corresponding p value. Next, we will order a scatterplot, which will provide a clear graph showing the paired points from both variables on a chart along with the regression line, sometimes referred to as a *trend line*, which can be thought of as the average pathway through the points.

Correlation

1. To run a correlation, starting from the main screen, click on *Analyze, Correlate, Bivariate* (Figure 8.4).

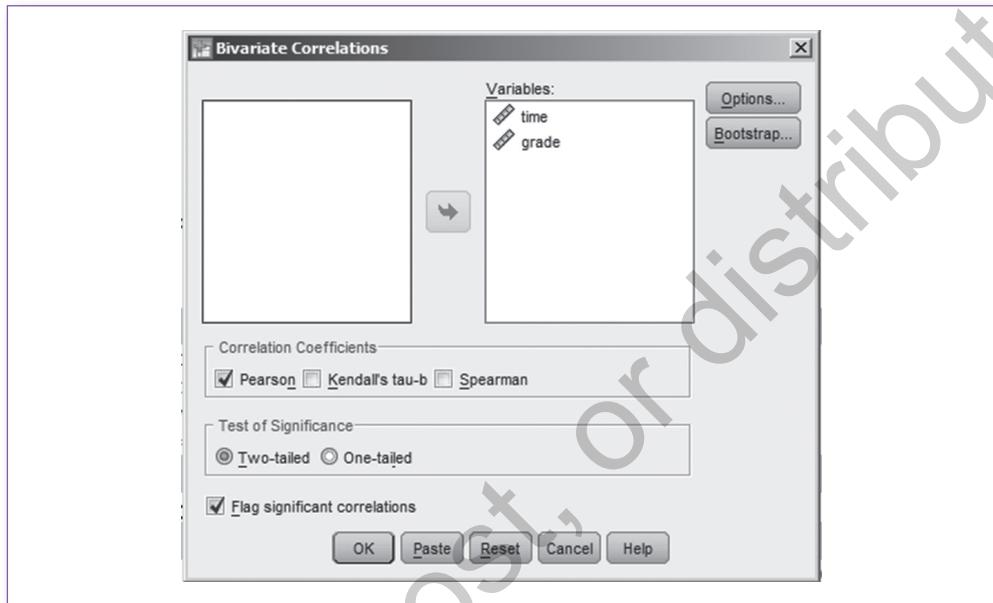
Figure 8.4 Accessing the *Bivariate Correlation* window: *Analyze, Correlate, Bivariate*.



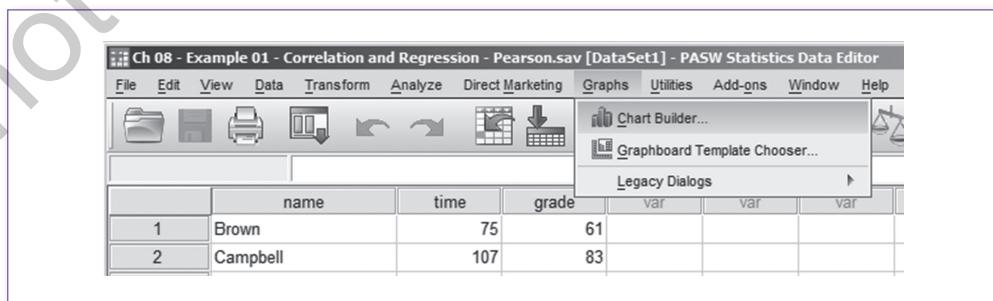
2. On the *Bivariate Correlations* window (Figure 8.5), move the *time* and *grade* variables from the left panel to the right (*Variables*) panel. Make sure that the *Pearson* checkbox is checked.

Figure 8.5

On the *Bivariate Correlation* window, move *time* and *grade* from the left panel to the right (*Variables*) panel.



3. Click the *OK* button, and the correlation will process. For now, set aside the correlations table that is produced; we will interpret it in the **Results** section.
4. To order a scatterplot with a regression line, from the main menu, click on *Graph, Chart Builder* (Figure 8.6).

Figure 8.6 Accessing the *Chart Builder* window: *Graphics, Chart Builder*.

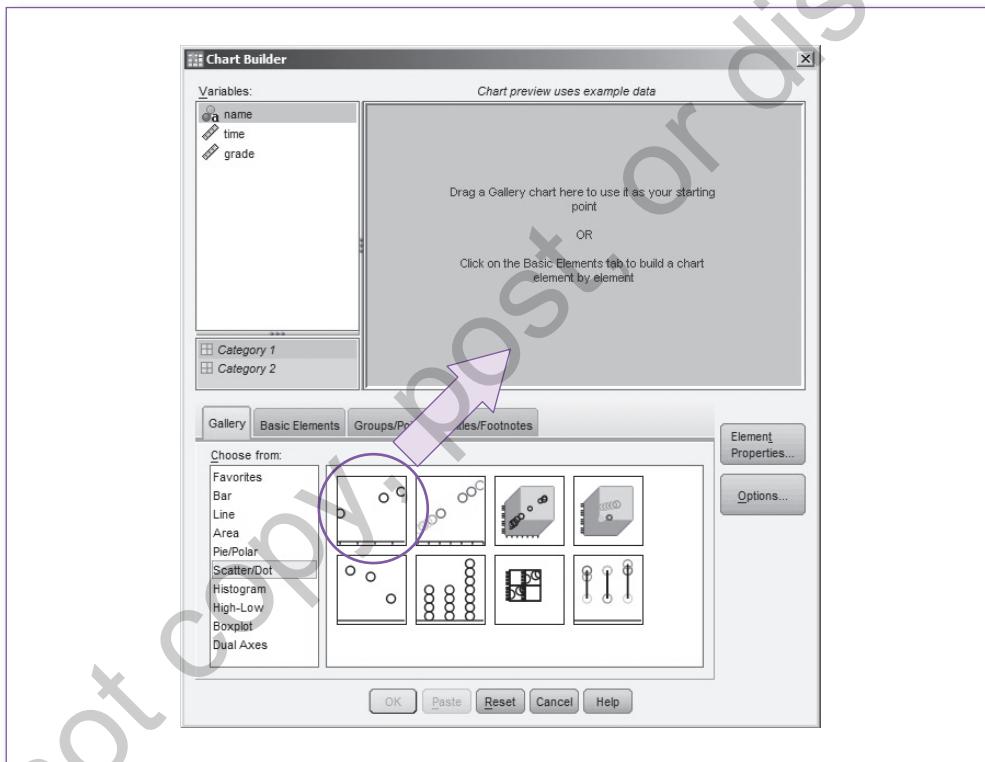


Regression (Scatterplot With Regression Line)

NOTE: SPSS graphics processing menus tend to differ across versions. If these instructions do not fit your version of the software, use the *Help* menu to guide you to order a scatterplot with a regression line. Indicate that you want the *time* variable on the X-axis and the *grade* variable on the Y-axis.

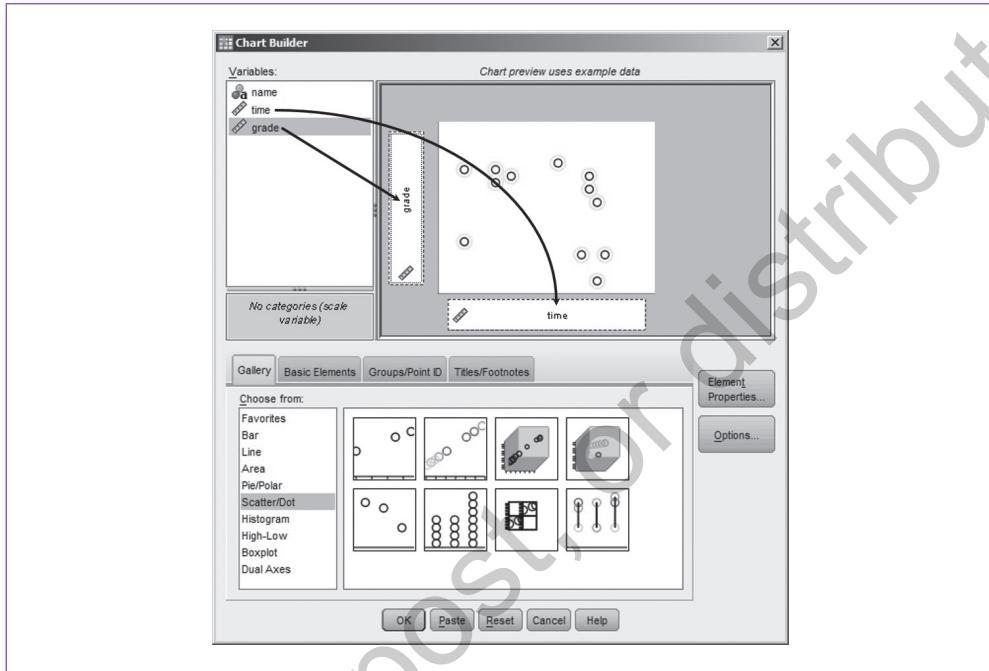
5. In the *Choose from* list, click on *Scatter/Dot* (Figure 8.7).

Figure 8.7 Chart Builder window.



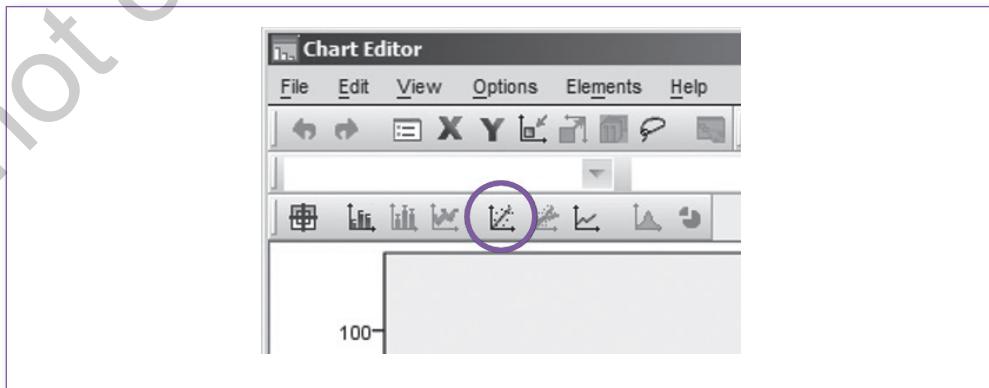
6. Double-click on the (circled) first choice, or click and drag this icon to the *Chart preview uses example data* window.
7. Click and drag *time* from the *Variables* panel to the *X-Axis* box (Figure 8.8).
8. Click and drag *grade* from the *Variables* panel to the *Y-Axis* box.
9. Click on the *OK* button, and the system will produce the scatterplot.

Figure 8.8 *Chart Builder* window—assign *time* to X-axis and *grade* to Y-axis.



When the scatterplot emerges, you will need to order the regression line: In the *Output* panel, double-click on the scatterplot. This will bring you to the *Chart Editor* (Figure 8.9).

Figure 8.9 *Chart Editor* window—click on *Add Fit Line* to include the regression line on the scatterplot.



10. Click on the *Add Fit Line at Total* icon to include the regression line on the scatterplot.
11. When you see the regression line emerge on the scatterplot, close the *Chart Editor*, and you will see that the regression line is now included on the scatterplot in the *Output* window.

Results

In this section, we will begin by explaining the two elements on the scatterplot: (a) the points and (b) the regression line. Next, we will finalize the two remaining pretest criteria (linearity and homoscedasticity), and finally, we will discuss the overall meaning of the scatterplot and correlation findings.

Scatterplot Points

The coordinates of each point on the scatterplot are derived from the two variables: *time* and *grade* for each record (individual).

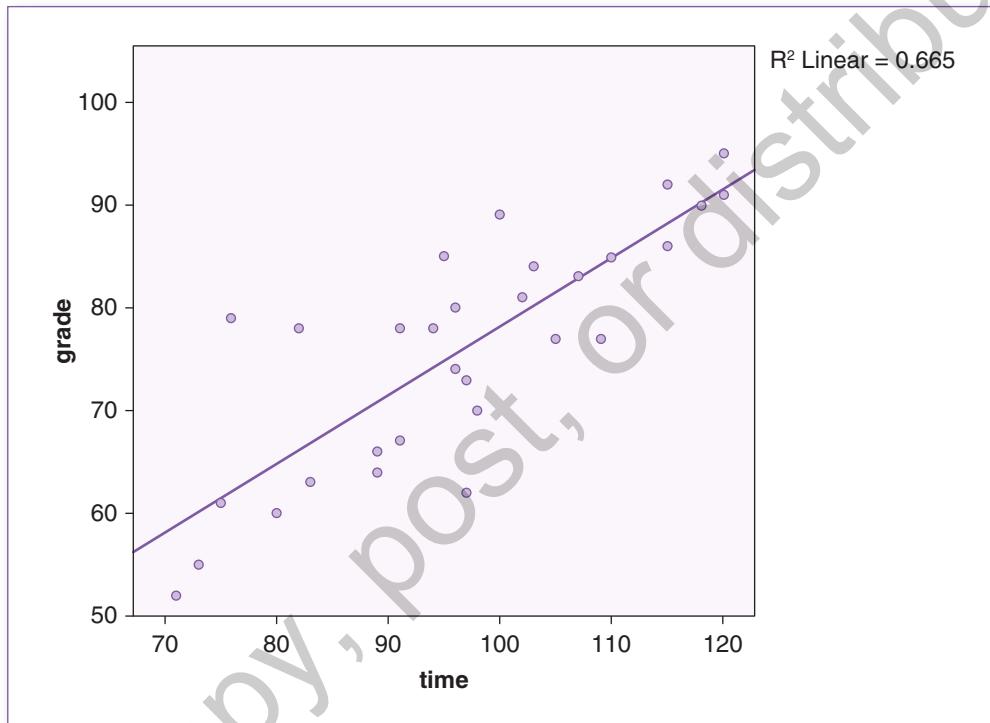
Figure 8.10 Source data for scatterplot: *time* and *grade*.

	name	time	grade
1	Brown	75	61
2	Campbell	107	83
3	DiNitto	80	60
4	Frank	91	67
5	Gray	91	78

The first record of the data set shows that Brown spent 75 minutes taking the exam and earned a *grade* of 61 on that exam (Figure 8.10). When we ordered the scatterplot, we placed *time* on the X-axis and *grade* on the Y-axis—hence, Brown’s dot on the (X,Y)

scatterplot is at coordinates (75, 61), Campbell's dot on the scatterplot is at (107, 83), and so on (Figure 8.11).

Figure 8.11 Scatterplot with regression line for the *time* : *grade* correlation.



Scatterplot Regression Line

The simplest way to conceive the regression line, without presenting the formula, is to think of it as the average straight-line pathway through the cloud of points, based on their positions. Just as the descriptive statistics provide a summary of a single variable, the regression line provides a sort of graphical summary of the relationship between pairs of continuous variables—in this case, *time* and *grade*.

Pretest Checklist Criterion 2—Linearity

The points on the scatterplot should form a relatively straight line (Figure 8.12); the regression line should take a middle-of-the-road path through the cloud of points. If the



overall shape of the points departs into some other shape(s) that is not conducive to drawing a straight (regression) line through it (Figure 8.13), this would constitute a violation of the linearity assumption.

Figure 8.12 Linearity satisfied.

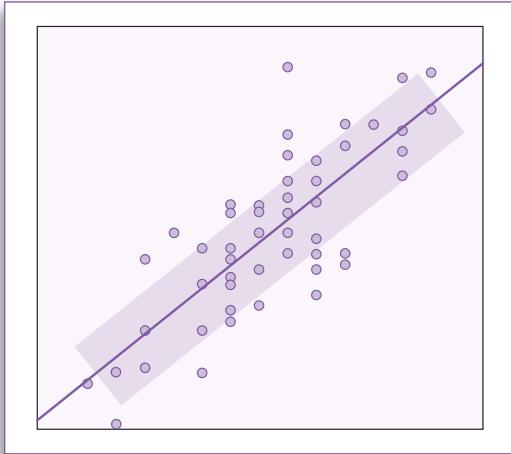
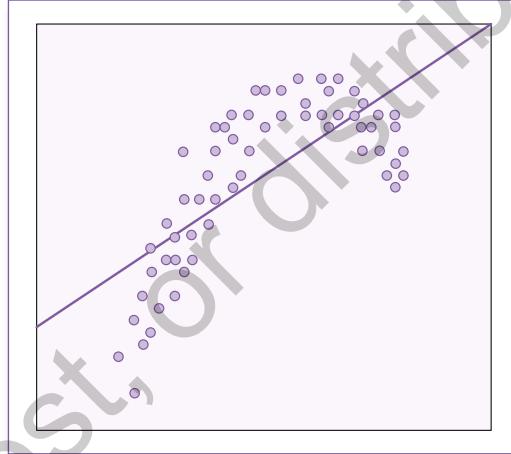


Figure 8.13 Linearity violated.



Pretest Checklist Criterion 3—Homoscedasticity

Homoscedasticity pertains to the density of the points along the regression line. The criterion of homoscedasticity is satisfied when the cloud of points is densest in the middle and tapers off at the ends (Figure 8.14), as opposed to the points being concentrated in some other way (Figure 8.15). The rationale for this distribution of points on the scatterplot follows the same notion as the shape of the normal curve of the histogram—the majority of the values are gathered around the mean, which accounts for the height of the normal bell-shaped curve on the histogram, whereas the tapered tails signify that there are considerably fewer very low and very high values. The positions of the points on the scatterplot are derived from the same data that rendered the normally distributed histograms for the two variables (Figures 8.2 and 8.3), so it follows that the middle of the cloud should contain considerably more points (and be denser) than the ends.

Correlation

Table 8.2 shows a positive correlation ($r = .815$) between time and grade, with a (Sig.) p value of .000. Despite the “.000” that is presented in the output, the p value never really reaches zero; in this case, $p = .0000000409310999$. When the p value is less than .001, it is typically notated as “ $p < .001$.” Since the p value is less than the α level of .05, and

Figure 8.14 Homoscedasticity satisfied.

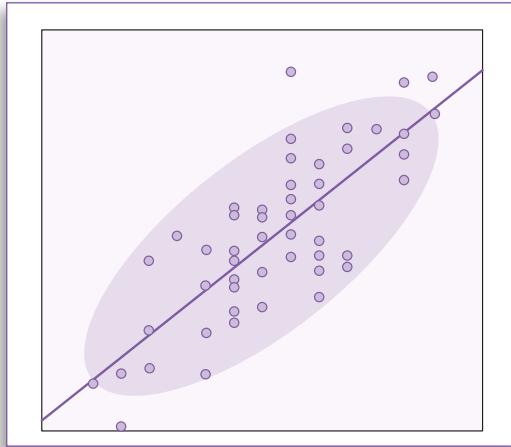
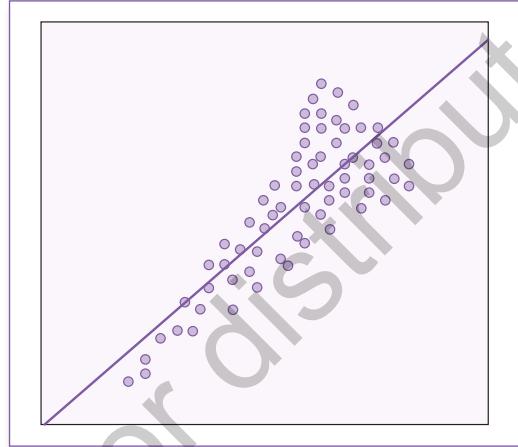


Figure 8.15 Homoscedasticity violated.



the r is greater than zero, we would say that there is a statistically significant positive correlation ($p < .001$, $\alpha = .05$) between *time* and *grade*. The positive correlation ($r = .815$) pertains to the positive slope of the regression line.

Notice that the *Correlations* table (Table 8.2) is double redundant; there are two $.815$'s and two $.000$'s in the table. This is because the correlation between *time* and *grade* is the same as the correlation between *grade* and *time*.

Table 8.2 Correlations Between *time* and *grade*.

		time	grade
time	Pearson Correlation	1	.815**
	Sig. (2-tailed)		.000
	N	30	30
grade	Pearson Correlation	.815**	1
	Sig. (2-tailed)	.000	
	N	30	30

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

H₀**Hypothesis Resolution**

REJECT H₀: There is no correlation between the length of time spent taking the exam and the grade on the exam.

ACCEPT H₁: There is a correlation between the length of time spent taking the exam and the grade on the exam.

Since the correlation calculation produced a p ($p < .001$) that is less than the specified .05 α level, we would say that there is a statistically significant (positive) correlation between the length of time that students spent taking the exam and the associated grade. As such, we would reject H₀ and not reject H₁.

**Documenting Results**

Prior to documenting the results, it can be helpful to run descriptive statistics for the two variables (*time* and *grade*) involved in the correlation (the procedure for running descriptive statistics can be found at the ★ icon on page 72; you can load *time* and *grade* into the *Variables* window together) (Table 8.3).

Table 8.3 Descriptive Statistics for *time* and *grade*.

Statistics			
		time	grade
N	Valid	30	30
	Missing	0	0
Mean		96.57	75.83
Median		96.50	78.00
Mode		89 ^a	78
Std. Deviation		14.132	11.570
Variance		199.702	133.868
Range		49	43
Minimum		71	52
Maximum		120	95

a. Multiple modes exist. The smallest value is shown

Discussing the n , means, and standard deviations of each variable along with the regression results can add to the substance of the abstract:

We were interested in discovering if there was a correlation between how long students spent taking an exam and the associated grade. The 30 students were allotted 2 hours to take their final exam. Students spent a mean of 96.6 ($SD = 14.1$) minutes taking the exam, and earned a mean grade of 75.8 ($SD = 11.6$) on the exam. Correlation analysis revealed a strong positive correlation between these two variables ($r = .815$), which was statistically significant ($p < .001$, $\alpha = .05$), suggesting that the more time students spend on their exams, the higher the grade, and vice versa.

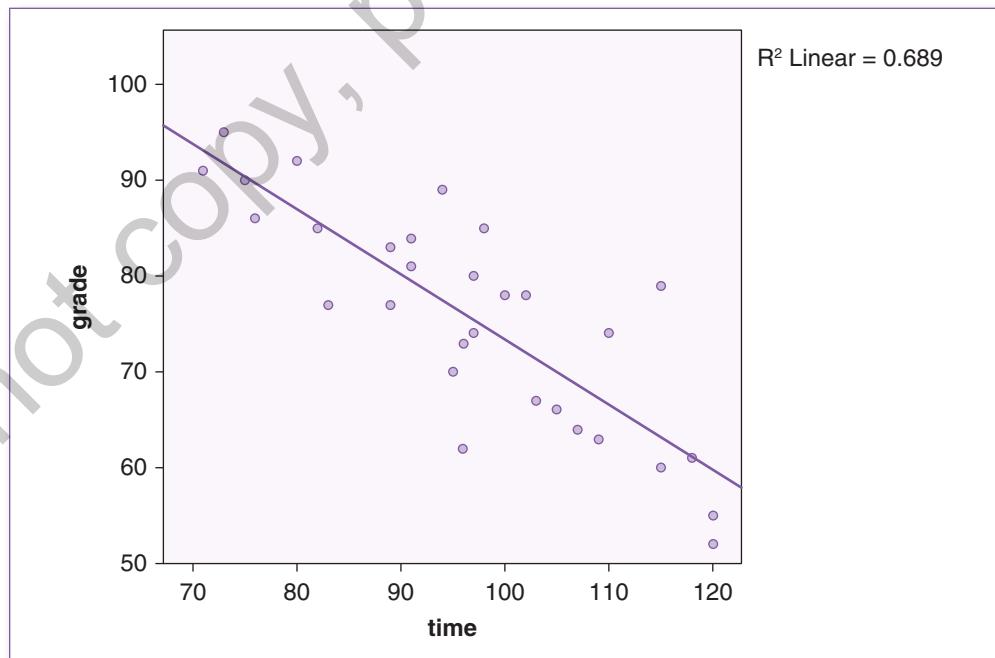
Before concluding this discussion of the Pearson statistic, wherein the example resulted in a statistically significant positive correlation between exam *time* and *grade*, let us take a brief look at two more examples: One that produces a *negative* correlation between *time* and *grade*, and another wherein there is *no* (significant) correlation.

Negative Correlation

Consider this result, where the data produced a statistically significant negative correlation ($r = -.803$, $p < .001$). The scatterplot would resemble Figure 8.16.

These findings reflect a negative (inverse) relationship between exam *time* and exam *grade*, suggesting that students who spent less time completing the exam scored higher, and vice versa.

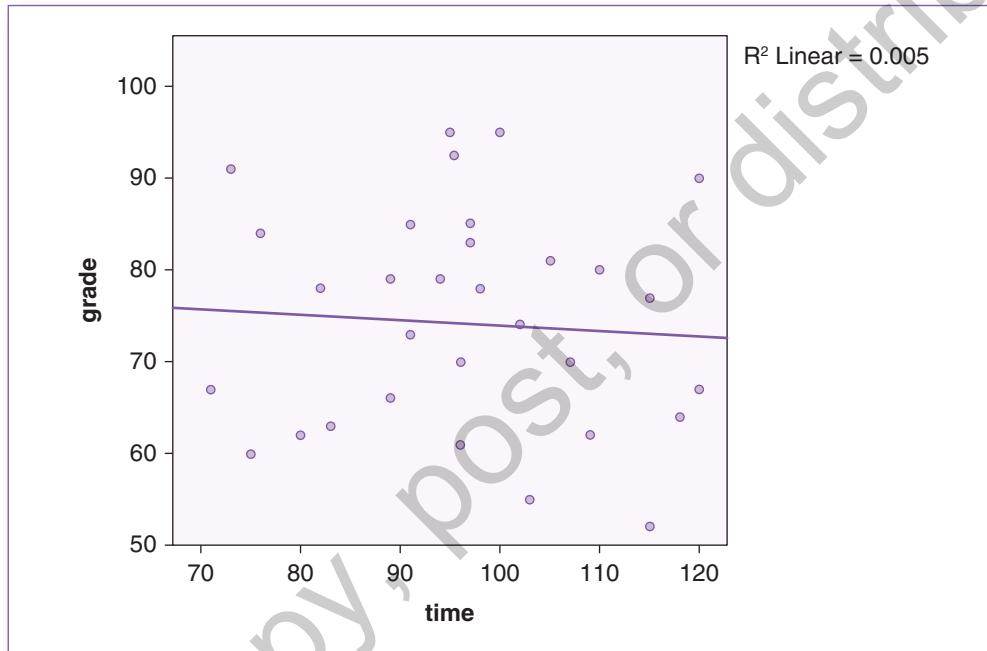
Figure 8.16 Scatterplot reflecting a statistically significant negative correlation between *time* and *grade* ($r = -.803$, $p < .001$).



No Correlation

Finally, it is possible that there is no statistically significant correlation between the two variables ($r = -.072, p = .704$), as shown in the scatterplot in Figure 8.17.

Figure 8.17 Scatterplot reflecting a statistically insignificant negative correlation between time and grade ($r = -.072, p < .704$).



Notice that as the r gets closer to 0, the relationship between the variables starts to break down. This is typically reflected in the scatterplot, wherein the points are scattered further from the regression line, and the regression line becomes more horizontal (less of a negative or positive slope).

Basically, this graph shows that there are about as many high and low grades among the students who spent a little time on the exam as those who spent a lot of time on the exam. In other words, the length of time a student spent taking the exam has nothing to do with how well he or she did on it.



OVERVIEW—SPEARMAN CORRELATION

The Spearman correlation, formally referred to as *Spearman's rho* (pronounced *row*), symbolized by the Greek letter ρ , can be thought of as a close cousin of the more commonly

used Pearson regression; however, whereas the Pearson statistic assesses the relationship between two continuous variables gathered from a data sample (e.g., height and weight), the Spearman correlation assesses the relationship between two rankings (ordered lists) using the same value range, -1 to $+1$, as the Pearson regression. The most common use of Spearman's rho is to determine how similarly two lists are sequenced.

For example, suppose you want to determine how similar Alice's color preferences are to Bill's. You could write the names of the colors on the front of cards, with the corresponding code number on the back (in this demonstration, the code numbers are shown on the front of each card for easy reference). Next, ask Alice and Bill to independently arrange their cards in their order of preference, with their favorite color at the top (Figure 8.18).

Figure 8.18 Two lists ranked identically produces a Spearman's rho of $+1$.

Alice	Bill
Red (1)	Red (1)
Green (2)	Green (2)
Blue (3)	Blue (3)

Since Alice and Bill arranged their cards in exactly the *same* order, this would produce a Spearman's rho of $+1$, signifying a perfectly positive correlation between the two prioritized lists (Figure 8.18).

If instead Alice sorted the colors the same way—*Red, Green, Blue* (1, 2, 3)—but Bill sequenced them *Blue, Green, Red* (3, 2, 1), these two rankings would be exactly *opposite* of each other, which would produce a Spearman's rho of -1 , signifying a (perfectly) negative correlation between the two lists (Figure 8.19).

Figure 8.19 Two lists ranked oppositely produces a Spearman's rho of -1 .

Alice	Bill
Red (1)	Blue (3)
Green (2)	Green (2)
Blue (3)	Red (1)

In this concise example, three items (colors) were used; however, there is no limit to the number of items that can constitute these lists. As you might expect, a variety of rankings of the items are possible, producing Spearman's rho values anywhere between -1 and $+1$. As with the Pearson correlation, the corresponding p value indicates if there is (or is not) a statistically significant difference between the (two) rankings.

EXAMPLE 2—SPEARMAN CORRELATION

A patient is referred to a dietician to collaboratively build a healthful eating plan. Part of this process involves ascertaining the patient's food preferences.

Research Question

Is there a statistically significant correlation between the dietician's recommended food ranking and the patient's current food preferences?

Groups

Unlike the Pearson regression, which gathers two continuous variables from each sample, the Spearman correlation gathers a sequence of ranked data from each of the two participants—in this case, food rankings from the dietician and the patient.

Procedure

At the initial consultation meeting, the dietician issues the patient five cards and asks the patient to arrange them in order of preference, with the favorite food at the top. The dietician will then use another set of cards to demonstrate the recommended diet in terms of which foods should be considered best (to worst) nutritional choices. The dietician will record the two card sequences and compare them using Spearman's rho.

Hypotheses

H_0 : There is no correlation between the dietician's recommended food ranking and the patient's food preferences.

H_1 : There is a correlation between the dietician's recommended food ranking and the patient's food preferences.



Data Set

Use the following data set: **Ch 08 - Example 02 - Correlation and Regression - Spearman.sav**.

Codebook

Variable: dietician

Definition: Dietician's recommended food ranking

Type: Categorical (1 = Vegetables, 2 = Fish, 3 = Poultry, 4 = Beef, 5 = Pork)

Variable: patient

Definition: Patient's food preferences

Type: Categorical (1 = Vegetables, 2 = Fish, 3 = Poultry, 4 = Beef, 5 = Pork)

The dietician and the patient independently arrange their five cards with the most preferable food at the top (Figure 8.20). On the *Data View* screen, you can click on the *Value Labels* icon to toggle between the numeric values and the corresponding assigned value labels.

Figure 8.20 Food rankings for dietician and patient.

Dietician	Patient
Vegetables (1)	Fish (2)
Fish (2)	Vegetables (1)
Poultry (3)	Poultry (3)
Beef (4)	Beef (4)
Pork (5)	Pork (5)

**Pretest Checklist**

Spearman's rho is a nonparametric (pronounced *non-pair-uh-metric*) test, meaning that the data are not expected to be normally distributed, and hence the pretest criteria for the Pearson regression (normality, linearity, and homoscedasticity) are not pertinent when it comes to running the Spearman correlation. Since each item is only present once

per variable, a bar chart, or histogram with a normal curve, would render all the bars at the same height, signifying one entry per value, which would be unrevealing.

The only real pretest criterion for Spearman's rho is to be certain that both lists consist of the same items—in this case, both the dietician and the patient ranked the same five food items, each in their own way.

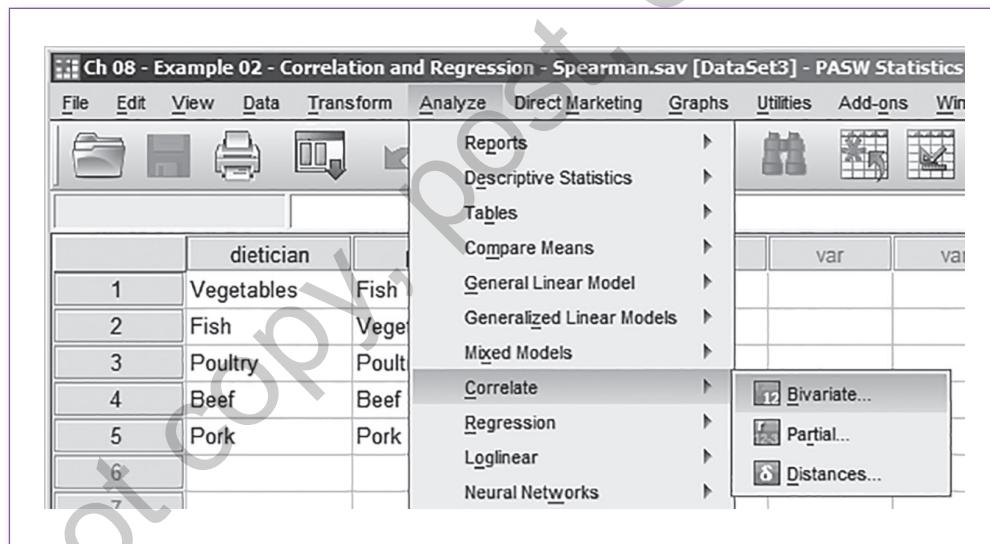


Test Run

The test run for the Spearman test involves the same order menu and results table as the Pearson correlation.

1. Click on *Analyze, Correlate, Bivariate* (Figure 8.21); this will take you to the *Bivariate Correlations* window (Figure 8.22).

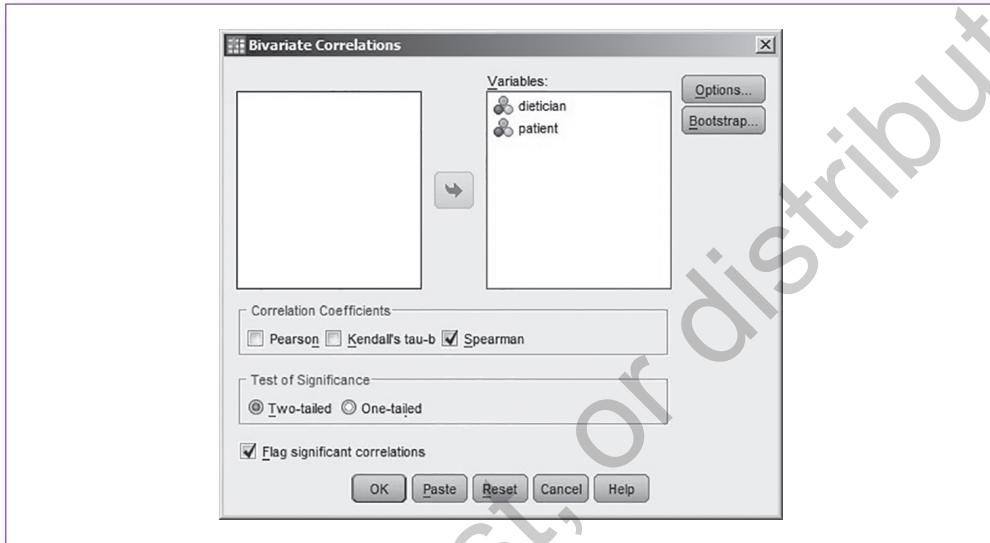
Figure 8.21 Accessing the *Bivariate Correlation* window: *Analyze, Correlate, Bivariate*.



2. On the *Bivariate Correlations* window (Figure 8.22), move both variables (*dietician* and *patient*) from the left panel to the right *Variables* panel.
3. Among the *Correlation Coefficients* options, uncheck *Pearson* and check *Spearman*.
4. Click *OK*, and the correlation will run.

Figure 8.22

On the *Bivariate Correlations* window, move the two variables into the right *Variables* panel, then uncheck *Pearson* and check *Spearman*.



Results

The results are presented in a single *Correlations* table (Table 8.4) indicating a Spearman’s rho of .900 with a corresponding *p* (Sig.) value of .037. This indicates a statistically significant positive correlation in the ranking of the two food lists. In other words, there is a strong similarity in the order of the foods on these two lists.

Table 8.4

Correlations Table Showing a Spearman’s Rho of .900 and Corresponding Statistically Significant *p* (Sig.) Value of .037

Correlations			dietician	patient
Spearman's rho	dietician	Correlation Coefficient	1.000	.900*
		Sig. (2-tailed)	.	.037
		N	5	5
patient		Correlation Coefficient	.900*	1.000
		Sig. (2-tailed)	.037	.
		N	5	5

*. Correlation is significant at the 0.05 level (2-tailed).

H₀**Hypothesis Resolution**

REJECT H₀: There is no correlation between the dietician's recommended food ranking and the patient's food preferences.

ACCEPT H₁: There is a correlation between the dietician's recommended food ranking and the patient's food preferences.

Spearman's rho is .900, indicating a strong positive correlation between the two lists; since the p value of .037 is less than the specified .05 α level, we would say that there is a statistically significant (positive) correlation between the food rankings of the dietician and the patient. As such, we would reject H₀ and not reject H₁.

**Documenting Results**

To work collaboratively with the patient in building a palatable healthy eating plan, as part of the initial encounter, the dietician asked the patient to sequence five food cards from favorite to least favorite without prompting. The dietician compared the patient's food preference (Fish, Vegetables, Poultry, Beef, Pork) with the recommended optimal nutrition for this patient (Vegetables, Fish, Poultry, Beef, Pork); Spearman's rho produced a statistically significant positive correlation of .900 ($p = .037$, $\alpha = .05$), indicating a strong concurrence between the two lists, suggesting that it should be fairly plausible to assemble a healthy dietary plan that is suitable to this patient's tastes.

**ALTERNATIVE USE FOR SPEARMAN CORRELATION**

The Spearman statistic is a viable alternative to the Pearson statistic when there is one or more substantial violation of the (Pearson) pretest criteria (normality, linearity, homoscedasticity).

Correlation Versus Causation

Correlation only means that two variables appear to move in a predictable direction with respect to each other (when one goes up, the other goes up; when one goes down, the other goes down; or when one goes up, the other goes down), but keep in mind, this is not necessarily due to *causation*, which would involve the change in one variable *causing* the change in the other. To make the leap from *correlation* to *causation*, three criteria must be met: (a) association/correlation, (b) temporality (timing), and (c) nonspurious (authentic) (Table 8.5).

Admittedly, the criteria to claim causation are strict, but without this rigor, numerous spurious (bogus) correlations could be wrongly attributed to causality, leading to inappropriate conclusions and potentially misguided interventions.

Table 8.5 Three Criteria for Satisfying Causality: Association/Correlation, Temporality, Nonspurious.

Causality Criteria		
Criteria	Rule	Example
1. Association/correlation	Variable A and Variable B must be empirically related; there must be a (scientific) logical relationship between A and B.	Taking a dose of aspirin lowers fever.
2. Temporality	A (cause [independent variable]) precedes B (effect [dependent variable]).	The person took aspirin, and <i>then</i> the fever went down, not the other way around.
3. Nonspurious	The relationship between A and B is not caused by other variable(s).	The drop in fever is not due to the room getting colder, submerging the person in an ice bath, or other factors.

For example, one might find a positive correlation between chocolate milk consumption and automobile theft—as chocolate milk sales go up, so do car thefts. Instead of concluding that chocolate milk causes people to steal cars or that car theft causes one to crave chocolate milk, anyone reasonable would continue his or her investigation and probably discover that *population* may be a variable worth consideration: In a town with a population of 2,000, we would find low chocolate milk sales and few car thefts, whereas in a city with a population of 2,000,000, chocolate milk sales and car thefts would both be considerably higher. In this case, we would be free to notice the positive correlation between chocolate milk consumption and car theft, but the causal criteria between these two variables clearly breaks down at all three levels.



OVERVIEW—OTHER TYPES OF STATISTICAL REGRESSION: MULTIPLE REGRESSION AND LOGISTIC REGRESSION

As we have seen, the most fundamental forms of statistical regression are the Pearson regression (r), which characterizes the relationship between two continuous variables, and the Spearman correlation (ρ), which compares the ranked order of two variables. The following sections provide an overview of two more advanced forms of regression: *multiple regression* and *logistic regression*, which are each capable of processing multiple variables.



Multiple Regression (R^2)

Multiple regression is best explained by example: Consider a smoking cessation intervention wherein 200 smokers are recruited to participate. In order to better understand the factors that are correlated with the treatment outcome, each participant is asked to complete a Smoking Cessation Survey at the conclusion of the treatment (Figure 8.23).

Figure 8.23 Smoking Cessation Survey for multiple regression analysis (Question 6 involves a continuous outcome variable).

Smoking Cessation Survey

1. What is your age? _____
2. What is your annual (gross) income? _____
3. Prior to this intervention, how many cigarettes did you smoke in an average day? _____
4. What is your gender?
 - Female Male
5. What is your race?
 - African American Asian Caucasian Latino Other
6. How many cigarettes do you now smoke in an average day? _____

Thank you for your participation.

Continuous Outcome

In Figure 8.23, Questions 1 through 5 are referred to as *predictor* (X) variables; they characterize the attributes of the individual that are thought to be correlated to Question 6 (“How many cigarettes do you now smoke in an average day?”)—the *outcome* (Y) variable. Notice that multiple regression accommodates continuous predictors (Questions 1, 2, and 3) and categorical predictors (Questions 4 and 5). Finally, notice that multiple regression involves one continuous outcome (Question 6).

Multiple regression produces a parsimonious (pronounced *par-sub-moe-nee-us*) model, meaning that it is the smallest set of predictors that provides the most explanation of the outcome. The multiple regression processor assesses the correlation between each predictor and the outcome variable. If a predictor variable is found to have a statistically significant correlation to the outcome variable, then it is included in the model, along with a percentage value; all other predictors are eliminated from the model, since they do not (significantly) predict the outcome.

The following is a sample of how such results would be documented:

We recruited 200 smokers to participate in a smoking cessation intervention. We gathered data detailing each participant's gender, race, age, income, baseline daily smoking rate, and daily smoking rate at the end of the intervention (0 = quit smoking).

Multiple regression analysis rendered an overall R^2 of .40, wherein gender accounts for 30% of the variability observed in the outcome variable (post-treatment smoking rate), race accounts for an additional 7%, and baseline smoking rate accounts for 3%.

Notice that *age* and *income* are not mentioned in the second paragraph of the above results documentation; this is because the regression processor determined that those two predictors are not statistically significantly correlated to the outcome variable (post-treatment smoking rate). Also notice that the overall R^2 accounts for only 40% of the variability observed in the outcome variable, so the question stands: *What about the other 60% ($100 - 40 = 60$)?* The answer to that is *error*. In this context, “error” does not imply that a mistake was made; rather, this model is saying that three of the predictor variables (*gender*, *race*, and *baseline smoking rate*) account for 40% of the variability observed in the outcome variable, leaving 60% unaccounted for. Basically, this is saying that there are other predictors that pertain to the outcome variable, which are not included in this model. If this study were to be repeated, we might consider retaining the three statistically significant predictors (*gender*, *race*, and *baseline smoking rate*), dropping the statistically insignificant predictors (*age* and *income*), and include some other, hopefully more relevant, predictor variables in order to increase the overall R^2 .



Logistic Regression

For clarity, we will use the same smoking cessation example with a slight modification to the outcome (dependent) variable.

Logistic regression is similar to multiple regression with regard to the predictor (independent) variables; both can process multiple continuous and categorical predictors. The major difference is that in multiple regression, the outcome variable is *continuous* (“How many cigarettes do you now smoke in an average day?”), whereas in logistic regression, the outcome variable is *dichotomous* (pronounced *die-cot-uh-muss*), meaning that it is a *categorical variable with two categories* (“What is your current smoking status? Still smoking Quit smoking”). Notice that the surveys for multiple regression (Figure 8.23) and logistic regression (Figure 8.24) are identical except for the outcome variable (Question 6).

The following is a sample of how such results would be documented:

We recruited 200 smokers to participate in a smoking cessation intervention. We gathered data detailing each participant's gender, race, age, income, baseline daily smoking, and smoking status at the end of the intervention (still smoking or quit smoking).

Logistic regression analysis revealed that men had 22 times the odds of quitting smoking compared with women. Those who indicated that their race designation was “Other” had 8 times the odds of quitting smoking compared with African Americans. Baseline smoking was also an influential factor; for every additional cigarette smoked per day, the odds of quitting smoking decrease by 5.4%.

Figure 8.24

Smoking Cessation Survey for logistic regression analysis (Question 6 involves a dichotomous outcome variable).

Smoking Cessation Survey

1. What is your age? _____
2. What is your annual (gross) income? _____
3. Prior to this intervention, how many cigarettes did you smoke in an average day? _____
4. What is your gender?
 - Female Male
5. What is your race?
 - African American Asian Caucasian Latino Other
6. What is your current smoking status?
 - Still smoking Quit smoking

Thank you for your participation.

Notice that multiple regression results are expressed as the percentage that each predictor accounts for the amount of variability observed in the outcome variable, whereas in logistic regression, categorical variables are reported in terms of odds ratios, and continuous variables are reported in terms of percentage increase or decrease with respect to the outcome variable.

GOOD COMMON SENSE

Reflecting back on the Pearson regression example involving homework hours and grade, even if a statistically significant, strong positive correlation were found between these two variables, it would be presumptuous to simply claim that homework hours (and nothing else) caused the grade. One or more underlying unaccounted-for factors that would not be revealed by correlation analysis may affect the homework hours,

which, in turn, may have affected the grade. For example, an adverse factor (e.g., household stress, health problem, adverse social issue) may be cutting into the student's ability to engage in longer homework hours; conversely, a pleasurable factor may be detracting from homework time (e.g., great social circle, multiple extracurricular activities, compelling video games). Alternatively, an overarching factor may affect both homework and grade, such as an undiagnosed learning disability, depression, exhaustion, effective tutor, and so forth.

The point is that correlation, no matter what the r or the p , is just that—an overall *correlation*; try to avoid jumping to conclusions regarding *causation*.

Key Concepts

- Pearson regression (r)
 - Correlation
 - Strength
- Direction
- Normality
- Linearity
 - Homoscedasticity
 - Bivariate correlation
 - Scatterplot
 - Regression
 - Spearman's rho correlation (ρ)
 - Correlation versus causation
 - Other statistical regression
 - Multiple regression (R^2)
 - Logistic regression

Practice Exercises

NOTE: Exercises 8.1 to 8.8 involve continuous data; use the Pearson statistic for these. Exercises 8.9 and 8.10 involve ordinal (categorical) data; as such, use the Spearman's rho statistic.

Exercise 8.1

An exercise advocate wants to determine the effect that walking rigorously has on weight loss. The researcher recruits participants to engage in a weeklong study. The researcher instructs participants to take a brisk walk as many days of the week as possible for as

long as they can. Participants will record the following data: weight prior to engaging in the walking regimen, the amount of time walked each day, and their weight at the end of the week. Participants will submit their data to the researcher at the end of the week. The researcher will preprocess the data to derive the total number of hours walked (*walkhrs*) and the change in weight for each participant (*wtloss* = weight at the end of the week – weight at the beginning of the week).

Data set: **Ch 08 - Exercise 01A.sav**

Codebook

Variable:	walkhrs
Definition:	Total hours walked in a week
Type:	Continuous
Variable:	wtloss
Definition:	Total weight loss in a week
Type:	Continuous

NOTE: In Data Set A, record 3, notice that the weight loss (*wtloss*) is -1.00 ; this indicates that the participant gained 1 pound. Data Set B, record 16, also signifies a half-pound weight gain ($wtloss = -0.50$) for that participant.

- Write the hypotheses.
- Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 01B.sav**.

Exercise 8.2

A social scientist has noticed that people seem to be spending a lot of nonwork hours on computers and wants to determine if this may, in some way, be associated with social relationship satisfaction (satisfaction derived from interacting with others). To determine if there is a correlation between nonwork computer hours and social satisfaction, the scientist recruited a group of participants and asked them to indicate (about) how many nonwork hours they spend on the computer each week. Next, each participant was given the Acme Social Satisfaction Inventory (ASSI); this self-administered instrument renders a score between 0 and 80 (0 = very low social satisfaction, 80 = very high social satisfaction).

Data set: **Ch 08 - Exercise 02A.sav**

Codebook

Variable: comphrs

Definition: Number of nonwork hours spent on the computer per week

Type: Continuous

Variable: aasi

Definition: Acme Social Satisfaction Inventory score

Type: Continuous (0 = very low social satisfaction, 80 = very high social satisfaction)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 02B.sav**.

Exercise 8.3

A social scientist and an economist working together want to discover if there is a correlation between income and happiness. The researchers recruit a group of participants and ask them to complete a confidential survey. This self-administered survey asks for the participant's annual income; it also includes the Acme Life Happiness Scale (ALHS), which renders a score between 0 and 100 (0 = very unhappy, 100 = very happy).

Data set: **Ch 08 - Exercise 03A.sav**

Codebook

Variable: income

Definition: Annual income in dollars rounded to the nearest thousand

Type: Continuous

Variable: alhs

Definition: Score on the Acme Life Happiness Scale

Type: Continuous (0 = very unhappy, 100 = very happy)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 03B.sav**.

Exercise 8.4

A political scientist wants to find out if there is a correlation between listening to a newscast and an individual's mood. This researcher recruits a group of participants and has them listen to a newscast that was recorded earlier that morning. Participants are instructed to listen for as long as they want; when they are done listening, the researcher writes down the listening duration and then asks each participant to complete the Acme Mood Report (AMR), a self-administered instrument that renders a score between 0 and 100 (0 = very bad mood, 100 = very good mood).

Data set: **Ch 08 - Exercise 04A.sav**

Codebook

Variable: minnews

Definition: Number of minutes of news listened to

Type: Continuous

Variable: amr

Definition: Acme Mood Report score

Type: Continuous (0 = very bad mood, 100 = very good mood)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.

- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 04B.sav**.

Exercise 8.5

An educational scientist wants to examine the correlation between years of education and job satisfaction. To address this question, the scientist recruits a group of participants and has each complete a self-administered survey; the first question asks how many years of education the participant has (e.g., 12 = high school diploma, 14 = associate's degree, 16 = bachelor's degree, 18 = master's degree). The remaining questions consist of the Acme Job Satisfaction Index (AJSI), which produces a score between 0 and 60 (0 = very unsatisfied with job, 60 = very satisfied with job).

Data set: **Ch 08 - Exercise 05A.sav**

Codebook

Variable: yearsed

Definition: Number of years of education

Type: Continuous

Variable: ajsi

Definition: Acme Job Satisfaction Index score

Type: Continuous (0 = very unsatisfied with job, 60 = very satisfied with job)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 05B.sav**.

Exercise 8.6

A dietician wants to discover if there is a correlation between age and number of meals eaten outside the home. The dietician recruits participants and administers a two-question survey: (a) “How old are you?” and (b) “How many times do you eat out (meals not eaten at home) in an average month?”

Data set: **Ch 08 - Exercise 06A.sav**

Codebook

Variable: age

Definition: Age of participant

Type: Continuous

Variable: mealsout

Definition: Number of means out participant eats per month

Type: Continuous

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 06B.sav**.

Exercise 8.7

A social scientist wants to determine if a person’s height might be correlated with his or her sense of self-confidence. To explore this, the scientist recruits a group of participants and gathers two metrics: First the researcher administers the Acme Self-Confidence Instrument (ASCI), a self-administered survey that produces a score between 0 and 50 (0 = very low self-confidence, 50 = very high self-confidence). Second, the scientist measures the height (in inches) of each participant.

Data set: **Ch 08 - Exercise 07A.sav**

Codebook

Variable: height

Definition: Height of participant (in inches)

Type: Continuous

Variable: asci

Definition: Acme Self-Confidence Instrument score

Type: Continuous (0 = very low self-confidence, 50 = very high self-confidence)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 07B.sav**.

Exercise 8.8

A sociologist has learned from a prior study that there is a strong positive correlation between time spent playing a video game and the score the player earns on that game (practice makes perfect). Since achieving such proficiency is time-consuming, this sociologist expects that there may be a (negative) correlation between game score and overall academic performance (grade: 0 to 100). To determine if there is such an inverse correlation, the sociologist recruits a group of participants to play a popular video game for 15 minutes, at which time the researcher records the score. Participants will also be asked to provide a copy of their most recent transcript.

Data set: **Ch 08 - Exercise 08A.sav**

Codebook

Variable: score

Definition: Score on video game

Type: Continuous

Variable: grade
 Definition: Overall academic grade
 Type: Continuous (0 to 100)

- a. Write the hypotheses.
- b. Run the criteria of the pretest checklist (normality [for both variables], linearity, homoscedasticity) and discuss your findings.
- c. Run the bivariate correlation, scatterplot with regression line, and descriptive statistics for both variables and document your findings (r and Sig. [p value], ns , means, standard deviations) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set: **Ch 08 - Exercise 08B.sav**.



NOTE: Exercises 8.9 and 8.10 involve categorical data; as such, use the Spearman's rho statistic.

Exercise 8.9

In order to better control inventory, Acme Motors wants to assess how similar customer car color preference is comparing the Pico Boulevard dealership with the Sepulveda Boulevard dealership (in the data sets, the most popular car color choice is at the top of the list for each dealership).

Data set: **Ch 08 - Exercise 09A.sav**

Variable: pico
 Definition: Customer car color preference at the Pico Boulevard dealership
 Type: Categorical (1 = Black, 2 = Blue, 3 = Red, 4 = Silver, 5 = White, 6 = Yellow)

Variable: sepulveda
 Definition: Customer car color preference at the Sepulveda Boulevard dealership
 Type: Categorical (1 = Black, 2 = Blue, 3 = Red, 4 = Silver, 5 = White, 6 = Yellow)

- a. Write the hypotheses.
- b. Verify the pretest checklist (both independently ranking the same set of items).

- c. Run the bivariate correlation for Spearman's rho, and document your findings (Spearman's rho and Sig. [p value]) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 09B.sav**.

Exercise 8.10

Ariel and Dusty want to determine how similar their movie preferences are. They independently rank the 13 movie categories with their favorite at the top.

Data set: **Ch 08 - Exercise 10A.sav**

Variable: ariel

Definition: Ariel's movie type preference

Type: Categorical (1 = Action/Adventure, 2 = Animation, 3 = Comedy, 4 = Cult Movie, 5 = Documentary, 6 = Fantasy, 7 = Film Noir, 8 = Horror, 9 = Romantic, 10 = Sci-Fi, 11 = Spy, 12 = Western, 13 = Zombies)

Variable: dusty

Definition: Dusty's movie type preference

Type: Categorical (1 = Action/Adventure, 2 = Animation, 3 = Comedy, 4 = Cult Movie, 5 = Documentary, 6 = Fantasy, 7 = Film Noir, 8 = Horror, 9 = Romantic, 10 = Sci-Fi, 11 = Spy, 12 = Western, 13 = Zombies)

- a. Write the hypotheses.
- b. Verify the pretest checklist (both independently ranking the same set of items).
- c. Run the bivariate correlation for Spearman's rho, and document your findings (Spearman's rho and Sig. [p value]) and hypothesis resolution.
- d. Write an abstract up to 200 words detailing a summary of the study, the bivariate correlation, hypothesis resolution, and implications of your findings.

Repeat this exercise using data set **Ch 08 - Exercise 10B.sav**.