14

CORRELATION ANALYSIS AND LINEAR REGRESSION

Assessing the Covariability of Two Quantitative Properties

14.0 LEARNING OBJECTIVES

In this chapter, we discuss two related techniques for assessing a possible association between two quantitatively assessed properties of a set of phenomena. These related techniques—*correlation analysis* and *regression analysis*—are based on the measure of association said to be the *covariance* as developed from probability theory (Chapter 8). In this chapter, we discuss the following:

- constructing what is said to be a *scatter plot* to provide a preliminary assessment of a possible association relationship between two properties based on *curve fitting*;
- assessing the *covariance* of the occurrences of the two properties in terms of the *correlation coefficient* and then using the *Central Limit Theorem* to assess the statistical significance of that *correlation coefficient*; and
- using the *correlation coefficient* to construct a *linear* mathematical model to *estimate* the covariability of the two properties as a mathematical function (this linear mathematical model is said to be a *linear regression*).

14.1 MOTIVATION

Recalling the discussion of an *explanatory* (or *association*) study presented in the introduction to Part IV, each such study begins with the following question:

> "Why do these different phenomena (people, places, or things) have different values for this property of interest?"

Following is the answer:

> "These phenomena have different values for the property of interest (the behavioral property, **Y**) because they were found to also have different values for another property (the explanatory property, **X**)."

In Chapter 11, we described the scenario in which *both* the behavioral property (**Y**) and the explanatory property (**X**) were *qualitative*. In this case, the phenomena can be sorted into groups according to both properties, and a suspected association between the properties may be assessed using the method of *Chi-Square Analysis* based on the premise of stochastic independence. Then, in Chapters 12 and 13, we described scenarios in which the behavioral property (**Y**) was *quantitative* rather than *qualitative*. As in the case in which the two properties were qualitative, the phenomena may be sorted and assessed using the method of *Chi-Square Analysis.* However, given that the behavioral property is quantitative, we can expand the informational content of the analysis by noting that each group of phenomena can be characterized by a typical (mean) value of the behavioral property, and we can determine whether the different groups have different typical (mean) values for the behavioral property. In the scenario in which there are only two groups to compare, the method of assessing the significance of the groups' differences is the *t-test* method of analysis (Chapter 12), and in the scenario in which there are several groups to compare, we use the method of *Analysis of Variance* (*ANOVA*) (Chapter 13).

Now, we will describe the investigative scenario in which *both* the behavioral property and the explanatory property are *quantitative*. To begin, we could start our analysis by sorting the phenomena into groups according to both properties in order to assess a possible association using either the method of *Chi-Square Analysis* or the method of *Analysis of Variance*. Typically, however, neither of these is the method of choice because neither takes full advantage of some useful information. That is, because both properties of interest are quantitative, not only can we assess whether specific values of the behavioral property tend to co-occur with specific values of the explanatory property, but we also can determine whether specific quantitative differences in the behavioral property tend to co-occur with specific quantitative differences in the explanatory property. For example,

- a political scientist might assess the extent to which individuals who spend more time on the Internet (daily hours) might have greater, or lesser, knowledge of American history (assessed as a quiz score);
- an economist might assess the extent to which job creation (new jobs per 1000 persons) might be related to the rate of net imports of goods and services; or
- a marketing manager might assess the extent to which those consumers who spend more time on social media also spend more money with online shopping.

This type of association is said to be *covariance* (Chapter 8), and two useful—and related—statistical methods of assessing the covariance of two quantitative properties of a set of phenomena are said to be *correlation analysis* and *regression analysis*. They are related in the following way:

1) The *covariance* of a set of coexisting properties is a measure of the extent to which

    a) *higher* (lower) values of one property tend to co-occur with *higher* (lower) values of the other property;

    b) *higher* (lower) values of one property tend to co-occur with *lower* (higher) values of the other property; or

    c) *higher* values of one property tend to co-occur with *higher and lower* values of the other property.

    Under scenario (a), the covariance will be a positive number, and this is said to represent a "direct" association. Under scenario (b), the covariance will be a negative number, and this is said to represent an "inverse" association. Finally, under scenario (c), the covariance will be a number near zero, and this is said to represent a non-association, or *stochastic independence*.

2) If the coexisting properties are converted to their *standardized* form (Chapter 5), the *covariance* is then said to be the *correlation coefficient*. Similar in interpretation to the covariance, the values of the *correlation coefficient* will range from a maximum of +1, representing scenario (a), to a minimum of –1, representing scenario (b), with a value near zero representing scenario (c).

3) The *correlation coefficient* can be used to construct a *linear mathematical model* describing the association between the two coexisting properties. This mathematical model is said to represent a *regression model*. If the two properties are **X** (explanatory) and **Y** (behavioral), the model will be of the form

$$\mathbf{Y} = B\,\mathbf{X} + C,$$

where B and C are constants.

We describe this analytical technique in Section 14.2 through an example.

Before we proceed, however, we should inject an important *caveat*. The *covariance* of two properties is only one of many types of possible "functional" relationships between two properties. For example, two properties might be related in the following way:

> Starting at $x = 0$, values of property **Y** *increase* with *increasing* values of property **X**; then, at value $x = a$, values of property **Y** *decrease* with *increasing* values of property **X**.

This type of "real" relationship would fail the conditions of a *covariance*. The *covariance*—or a *linear relationship*—is simply the most basic form of a functional relationship. Moreover, linear relationships are well represented in the natural world, so covariance analysis remains a useful start to any association study.

14.2 AN EXAMPLE

Dr. M, an educational psychologist at a high school district covering a large student body, has read several national studies suggesting that students who sleep more perform better academically. While the results seem reasonable, Dr. M knows that many intervening demographic factors might be involved, and the national results might not be as profound for the students in her district. Moreover, Dr. M was a light sleeper, she used her waking hours to study, and she was the class *valedictorian*. Consequently, Dr. M has decided to test this "more sleep, better performance" proposition using data from her district. Because her district is relatively homogeneous in its demographic profile, the potential effects of demographic differences are less likely to be present. To this end, Dr. M initiates the following study. Selecting 100 students at random, she asks them to monitor their daily school night sleeping habits (minutes of sleep) for 2 weeks. From these records, each student's average sleeping time can be determined. Adding to this, she has each student's academic record (grade point average, or GPA). The summary statistics for these two properties for these students are found in Table 14.1.

Now, while Dr. M could use the *Chi-Square Analysis* or *ANOVA* method to assess the extent to which students with specific sleeping habits have specific grade point averages, she is more interested in the following proposition: Do students

TABLE 14.1 ■ Sleeping Habits and GPA of 100 Randomly Selected Students in District A

Property	Mean	Standard Deviation
Sleep (minutes)	470.2	76.554
GPA (0–4)	2.8790	0.6525

who tend to sleep more have (a) higher grade point averages, (b) lower grade point averages, or (c) both high and low grade point averages? Following our discussion of Chapter 8, these two properties are said to be *coexisting,* with scenario (a) representing a positive association, scenario (b) representing a negative, or inverse, association, and scenario (c) representing a non-association, or *stochastic independence*.

As a first step, Dr. M chooses to use a technique of visual analysis to assess the extent to which either scenario (a), scenario (b), or scenario (c) seems to describe the pattern in the co-occurrences of these two properties. This visual technique is said to be the construction of a *scatter plot.* A *scatter plot* is a pictorial representation of a *contingency table* that is useful for displaying observations of phenomena described by two coexisting quantitative properties, and it is constructed using *Cartesian coordinates* (named for the philosopher and mathematician René Descartes, 1596–1652). As a first step of analysis, the *scatter plot* provides an indication of the appropriateness of constructing a linear model to describe a potential association. That is, in some cases, a different mathematical model might better fit the observed phenomena, and where a linear model might suggest that no relationship exists, a different mathematical model might reveal a "nonlinear" relationship between the two properties. As a technical note, while this analytical step is useful, it is also optional. See Section 14.3 for further details.

14.3 VISUAL INTERPRETATION WITH A SCATTER PLOT (OPTIONAL)

A *scatter plot* is a pictorial representation of a *contingency table*, and it is constructed using *Cartesian coordinates*:

- Two orthogonal axes—one vertical and the other horizontal—are drawn to represent the two coexisting properties as *dimensions*. The horizontal axis is typically identified as the "X-axis," and it is typically used to represent the *independent* (*explanatory*) property. The vertical axis is typically identified as the "Y-axis," and it is typically used to represent the *dependent* (*behavioral*) property.
- The zero point of each scale is placed at the intersection of the two axes. This intersection is said to be the *origin* of the coordinate system.
  a) Positive values of the scale used to assess property **X** are placed on the horizontal axis to the right—or "east"—of the origin, and negative scale values of **X** (if relevant) are placed to the left—or "west"—of the origin.
  b) Positive values of the scale used to assess **Y** are placed on the vertical axis above—or to the "north" of—the origin, and negative scale values of **X** (if relevant) are placed below—or to the "south" of—the origin.

- Each observation has a value for property **X**, and each observation also has a value for property **Y**; thus, each observation is represented as a dot according to its **X**-value and its **Y**-value. The **X** and **Y** values of an observation are said to be its "coordinates."

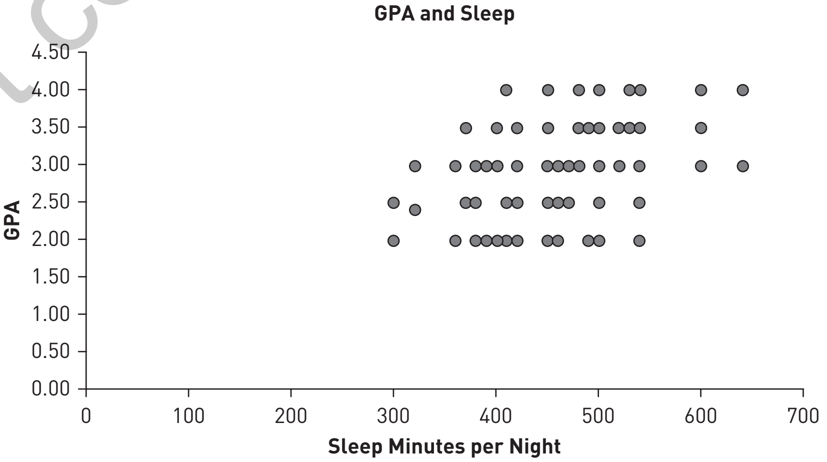
Depicted in this format, a set of observations may be compared with any number of different mathematical models in an exercise said to be *curve fitting.* Some commonly used mathematical models include the parabola, the hyperbola, the "S-curve," the exponential curve, and an oscillating sine wave. Each of these patterns has a mathematical form, and choosing the mathematical form that best fits a set of observations is the "curve-fitting" exercise. Of all the mathematical models, however, the linear model is the simplest; thus, it is the model of choice unless compelling evidence suggests otherwise. As a technical note, the mathematical models that are not the "linear" model are said to be "nonlinear."

Now, from the students' responses, Dr. M constructs the scatter plot shown in Figure 14.1. From this depiction, Dr. M sees a potential *linear* association between these two properties. That is, while the observations do not follow a line,

- they do indicate a tendency for students with more sleeping hours to have higher grade point averages; and
- the increase in grade point averages does not seem to accelerate or decelerate as typical sleep time increases.

Thus, Dr. M will proceed on the basis that the most logical model to describe these observations is a *linear model*.

FIGURE 14.1 ■ Scatter Plot of Observed Values of Sleeping Habits and GPA for 100 Randomly Selected Students



14.4 ASSESSING AN ASSOCIATION AS A COVARIANCE

As described in Chapter 8, the *covariance* is a measure of the *covariability* expected in the co-occurrences of two quantitative properties coexisting in the elements of a *sample space*. When describing the covariability in a set of *actual observations*, the *covariance* is interpreted as the "typical" covariability of the values of two such quantitative properties coexisting in a set of phenomena. As a *model* of covariability, the *covariance* is based on the following:

- Values of a quantitative property can be classified as "high" or "low."
- If "high" values of one property co-occur with "high" values of the coexisting property and "low" values of one property co-occur with "low" values of the coexisting property, this is identified as a "direct" association.
- If "high" values of one property co-occur with "low" values of the coexisting property, this is identified as an "inverse" association.

As a *measure* of covariability, the covariance is based on the following:

- Adopting the *method of moments*, "high" values and "low" values of a property are assessed in comparison with the mean for a set of observations. Expressed as an arithmetic difference (i.e., subtraction), a "high" value will yield a positive difference, and a "low" value will yield a negative difference.
- Adopting a model of interaction from the physical sciences, coexisting properties of a phenomenon are represented as the product of their respective values.
  a) A "high" value of one property co-occurring with a "high" value of the coexisting property will result in a positive number. Similarly, a "low" value of one property co-occurring with a "low" value of the coexisting property will also result in a positive number. Thus, a positive product represents a direct association.
  b) A "high" value of one property co-occurring with a "low" value of the coexisting property will result in a negative number. Thus, a negative product represents an inverse association.
- Combining the *method of moments* and physical science models, we can represent the "typical" co-occurrence of the two properties as the mean of the combined co-occurrences. This is said to be the *covariance.*
  a) If the mean is positive, the typical co-occurrence is consistent with a direct association.

b) If the mean is negative, the typical co-occurrence is consistent with an inverse association.

c) If the mean is zero, the mix of "direct" associations and "inverse" associations is equal, thereby indicating a non-pattern of association. This was identified in Chapter 8 as a "non-association," and the two properties would be identified as *stochastically independent*.

- In this way, the *covariance* technique provides an objective and interpretable assessment of an association between two properties.

In statistical practice, however, the *standardized* version of the *covariance*—defined in Chapter 8 as the *correlation coefficient*—is typically found to be more useful. To find the *correlation coefficient* for a set of phenomena with coexisting values of properties **X** and **Y**, we take the following steps:

1) For the set of phenomena, we find the mean value $\bar{\mathbf{x}}$ for property **X** and the mean value $\bar{\mathbf{y}}$ for property **Y**.

2) For the set of phenomena, we then find the standard deviation $\mathbf{s}_x$ for property **X** and the standard deviation $\mathbf{s}_y$ for property **Y**, using the Bessel Correction (Chapter 10) in both cases.

3) For each phenomenon, we standardize its value x for property **X** by the z-transformation,

$$x^* = (x - \bar{\mathbf{x}}) / \mathbf{s}_x,$$

and we standardize its value y for property **Y** by the z-transformation,

$$y^* = (y - \bar{\mathbf{y}}) / \mathbf{s}_y.$$

4) Then, for each phenomenon, we take the following steps:

a) We compare its standardized value for property **X** with the standardized mean for property **X**, which is 0. This gives us $x^* - 0$.

b) We compare its standardized value for property **Y** with the standardized mean for property **Y**, which is 0. This gives us $y^* - 0$.

c) We multiply the two comparisons,

$$(x^* - 0) \bullet (y^* - 0) = x^* \bullet y^*.$$

5) Finally, we add all these products together and find the mean by dividing the total by the number n of phenomena minus 1 (the Bessel Correction):

$$(\sum (x^* \bullet y^*)) / (n - 1).$$

The result is the "standardized" covariance, or *correlation coefficient*, and it is denoted as *r*. In this way, the *correlation coefficient* provides an *index of covariation* as a percentage of the combined variability of the two coexisting properties. As an index, the *correlation coefficient* will always be within the range of –1 to +1, with –1 indicating a strong inverse association, +1 indicating a strong direct association, and 0 indicating a non-association or stochastic independence. Furthermore, in statistical practice, the *correlation coefficient* is typically interpreted in the following way:

- A value of *r* between –1.00 and –0.60 is interpreted as a strong inverse association between the two properties. This means that phenomena having high values for one property are very likely to have low values for the other property, and phenomena having low values for one property are very likely to have high values for the other property.
- A value of *r* between –0.59 and -0.01 is interpreted as a weak inverse association between the two properties. This means that phenomena having high values for one property are somewhat more likely to have low values for the other property, and phenomena having low values for one property are somewhat more likely to have high values for the other property.
- A value of *r* of 0 is interpreted as a non-association, indicating the stochastic independence of the two properties. This means that phenomena having high values for one property are just as likely to have high values for the other property as they are to have low values for the other property, and phenomena having low values for one property are just as likely to have high values for the other property as they are to have low values for the other property.
- A value of *r* between 0.01 and 0.59 is interpreted as a weak direct association between the two properties. This means that phenomena having high values for one property are somewhat more likely to have high values for the other property, and phenomena having low values for one property are somewhat more likely to have low values for the other property.
- A value of *r* between 0.60 and 1.00 is interpreted as a strong direct association between the two properties. This means that phenomena having high values for one property are very likely to also have high values for the other property, and phenomena having low values for one property are very likely to also have low values for the other property.

Finally, as a technical note, the *correlation coefficient* may also be found by first finding the "unstandardized" *covariance* of a set of observations and then dividing the "unstandardized" *covariance* by the standard deviations of the two coexisting properties. While these two methods are mathematically equivalent, the method of first standardizing the observations and then finding the *covariance* better represents the *meaningfulness* of the *correlation coefficient*, and this fact will be usefully employed in our later discussion of *regression*.

Returning, then, to Dr. M's investigation, she has assessed the *correlation coefficient* describing the association between the sleeping habits of the surveyed students and their grade point averages to be 0.427. Thus, Dr. M concludes that she has found *tentative* evidence of a weak positive association between these two properties. That is, those students who sleep more than an average amount are somewhat more likely to have better than average grade point averages, while students who sleep less than an average amount are somewhat more likely to have lower than average grade point averages. Dr. M considers this conclusion to be tentative, however, because she understands the need to rule out the likelihood that the result represents only the normal variability of sampling. That is, it could be that there is no "real" relationship between the two properties, and her results might have been different with a different sample of students. In other words, Dr. M needs to assess the *statistical significance* of her results, and here she can use the *t-test for statistical significance*.

As described in Chapter 12, the *t-test* uses the *t-Probability Model* to assess the probability that two samples with *different means* were drawn from the *same population*. If that probability is very low—less than 0.05—we can conclude that the two samples *probably were not* drawn from the same population. Now, in the case of the *correlation coefficient*, it *is* a sample mean. Moreover, we can identify what the "ideal" *correlation coefficient* would be if the two properties were not related (i.e., *stochastically independent*). This "ideal" *correlation coefficient* is zero. Thus, it would be useful if we could use the *t-Probability Model* to assess the probability of drawing our actual sample—with a *correlation coefficient* of 0.427—and an "ideal" sample—with a *correlation coefficient* of zero—from the same "ideal" population in which the two properties *are not related*. If that probability is very low—less than 0.05—we would conclude that the actual sample *probably was not drawn* from the "ideal" population, and we would conclude that the observed correlation coefficient *probably was not* the result of normal sampling variability. Now, the question is, can we use the *t-Probability Model* for this assessment? The answer is "yes," and Box 14.1 provides a brief mathematical explanation.

Now, to use the *t-test* in this case, Dr. M compares the *correlation coefficient* of her sample (r) with the "ideal" *correlation coefficient* of a hypothetical population in which the two properties are not related. This "ideal" *correlation coefficient* is 0 based on the premise of stochastic independence. Recalling that the actual correlation coefficient and the ideal correlation coefficient both are sample "means," this gives a difference of the two sample means as

$$(r - 0) = r.$$

BOX 14.1

As with the *Central Limit Theorem*, the *t-test* and *t-Probability Model* are based on the mathematical fact regarding samples drawn from a population described by a quantitative property **X** with a mean for the population of μ:

- most samples will have means identical or nearly identical to the population mean;
- a slightly lesser number of samples will have means that differ moderately from the population mean; and
- very few samples will have means that differ greatly from the population mean.

This pattern fits a *Normal Probability Model* (Chapter 9).

Now, suppose a population of phenomena are described by the quantitative properties **X** and **Y**, with population means of μ_x and μ_y, respectively. Any sample from this population will have a mean $\bar{x}$ for property **X** and a mean $\bar{y}$ for property **Y**. Separately, for the property **X**,

- most samples will have a mean for property **X** identical or nearly identical to the population mean;
- a slightly lesser number of samples will have a mean for property **X** that differs moderately from the population mean; and
- very few samples will have a mean for property **X** that differs greatly from the population mean.

Similarly, for the property **Y**,

- most samples will have a mean for property **Y** identical or nearly identical to the population mean;
- a slightly lesser number of samples will have a mean for property **Y** that differs moderately from the population mean; and
- very few samples will have a mean for property **Y** that differs greatly from the population mean.

In mathematical terms, the sample means for property **X** are normally distributed, and the sample means for property **Y** are also normally distributed.

Now, from probability theory (see Feller, William, 1968, *An Introduction to Probability Theory and Its Applications*, Vol. 1, New York: John Wiley), we know that two normally distributed, and stochastically independent, "random variables" drawn from the same "sample space" are also "bivariate normal." That is, suppose **X** is a property of a set of phenomena, **Y** is a property of those phenomena, and the two properties are *stochastically independent*. Furthermore, suppose the values of **X** are normally distributed among the phenomena and the values of **Y** are also normally distributed among the phenomena. We can then describe each phenomenon as *xy*, where *x* is the **X**-value of the phenomenon and *y* is the **Y**-value. The set of all possible values of *xy* is then denoted as **XY**. Now, as a fact from probability theory (see Feller's *Introduction* cited above), we know that the expected value of **XY** will be $\mu_x \bullet \mu_y$, where μ_x and μ_y are the respective means for properties **X** and **Y**. We also know the following regarding the distribution of the values of *xy* among the phenomena:

- most phenomena will have an **XY** value identical or nearly identical to $\mu_x \bullet \mu_y$;
- a slightly lesser number of phenomena will have an **XY** value that differs moderately from $\mu_x \bullet \mu_y$; and
- very few phenomena will have an **XY** value that differs greatly from $\mu_x \bullet \mu_y$.

In other words, the **XY**—or bivariate—values of the phenomena are "normally" distributed.

Returning to the *t-test*, for the purpose of *significance testing*, we can presume that we have a population of hypothetical phenomena represented by the coexisting properties x^* and y^*, both of which are standardized. Moreover, we can presume that the two properties are *stochastically independent*. Thus, we have the expected value of $x^* y^*$ as $0 \bullet 0 = 0$. Furthermore, regarding samples drawn from this population,

(Continued)

(Continued)

- most samples will have a mean $x^* y^*$ value identical or nearly identical to 0;
- a slightly lesser number of samples will have a mean $x^* y^*$ value that differs moderately from 0; and
- very few phenomena will have an $x^* y^*$ value that differs greatly from 0.

That is, the occurrence of sample mean $x^* y^*$ values will follow a Normal Probability Model.

As to the *expected difference* between two sample means drawn from this population, she uses the formula

$$\sqrt{\ }((1 - r^2)/(n - 2)),$$

where n is the size of the sample and the Bessel Correction (–2) has been applied to reflect the appropriate degrees of freedom. (See Box 12.1 for a derivation of this formula.) In combination, this gives the *t-statistic* as

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

$$0.427/\sqrt{\ }(1 - 0.182)/98) =$$

$$0.427/\sqrt{\ }(0.818/98) =$$

$$0.427/\sqrt{0.00834} =$$

$$0.427/0.09136 = 4.673.$$

Then, using the *t-Probability Model* appropriate for 98 degrees of freedom, Dr. M finds the *p-value* of such a *t-value* to be less than 0.001. She interprets this in the following way:

> The probability of having selected her sample—with a *correlation coefficient* of 0.427—from a hypothetical population in which the two properties are not related—thus, with a *correlation coefficient* of zero—is less than 0.001. Therefore, given this "improbability," it is reasonable to believe that the sample *probably did not* come from a hypothetical population in which the two properties *were not related* and, thus, that the two properties *probably are related*. In technical terms, she would say that the "null hypothesis"—that the two properties *are not related*—has been *rejected*, and the relationship may be judged to be *statistically significant*.

Dr. M can then be relatively confident in the findings of her study, which suggest the following:

- students who tend to sleep more than average are somewhat more likely to have higher than average grade point averages; and

- students who tend to sleep less than average are somewhat more likely to have lower than average grade point averages.

14.5 REGRESSION ANALYSIS: REPRESENTING A CORRELATION AS A LINEAR MATHEMATICAL MODEL

Having found evidence of a *direct* association (or *positive correlation*) between sleeping habits and academic performance, Dr. M is curious to see whether the relationship might be used to *predict* exactly how much students' grade point averages tend to change with each additional minute of sleep. Mathematically, a predictive model based on the *correlation* (whether positive or negative) is a linear model of the form

$$\mathbf{Y} = (b \bullet \mathbf{X}) + c.$$

In this model, *b* and *c* are constants, and the model describes the expected grade point average (**Y**) associated with each possible value of "daily minutes sleeping" (**X**). More important, this model offers the following predictions:

- for each additional unit (minute) of sleep, a student's grade point average can be expected to change by *b* units (points); and
- a student who sleeps zero hours (i.e., $\mathbf{X} = 0$) can be expected to have a grade point average of *c* (i.e., $\mathbf{Y} = c$).

In technical terms,

- *b* is said to be the coefficient of the "variable" **X**; and
- *c* is said to be the "constant" coefficient, referring to the fact that the property **X** is "held constant."

Moreover, in describing the model in terms of *Cartesian coordinates*, *b* is said to represent the "slope" of the line represented by the model. It represents the *expected* change in **Y** (the behavioral property) due to a *given* change in **X** (the explanatory property).

- It is correspondingly written as $\Delta y / \Delta x$.
- *c* is said to be the *y* "intercept," corresponding to the value of **Y** when $\mathbf{X} = 0$.

Now, how is a linear model constructed from a set of observations? There are two methods, both of which yield the same result.

A) The method of estimation said to be *ordinary least squares regression* was developed by the parallel—and contentious—efforts of Pierre Simone Laplace, Adrien Marie Legendre, and Carl Friedrich Gauss at the end of the 18th century and beginning of the 19th century (see Stigler, Stephen M., 1986, *The History of Statistics*, Cambridge, MA: Belknap Press). Using the calculus, a hypothetical line is proposed to represent "the best fit" among the observations. The difference between the hypothetical line and each observation is measured as a distance, and the calculus is used to find the hypothetical line that minimizes all of the differences *squared*. The resulting minimization process is said to result in the "least squares"; thus, the resulting hypothetical line is said to be the "least squares" model. The equation of this line will be

$$\mathbf{Y} = ((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \mathbf{X} - (((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \overline{\mathbf{x}}) + \overline{\mathbf{y}},$$

where r is the correlation coefficient, $\overline{\mathbf{x}}$ is the mean observed value of property **X**, $\mathbf{s}_x$ is the observed standard deviation in the values of property **X**, $\overline{\mathbf{y}}$ is the mean observed value of property **Y**, and $\mathbf{s}_y$ is the observed standard deviation in the values of property **Y**. In constructing this line, it should be remembered that it represents the *average* of the co-occurrences of the two properties **X** and **Y** and that some of the actual co-occurrences will be consistent with the line and others will not. Put somewhat differently, the *regression line* in this form represents an *estimate* of what we might *expect* to find for the co-occurrences of the two properties **X** and **Y**.

B) A linear model can also be constructed using the correlation coefficient:

$$y^* = r \bullet x^*,$$

where y^* and $x^*$are the standardized versions of the properties **Y** and **X**. This is said to be the *standardized* version of the linear regression model. This "standardized" version can then be "unstandardized" by "reversing" the z-transformation. That is, the z-transformation is as follows:

$$y^* = (\mathbf{Y} - \overline{\mathbf{y}})/\mathbf{s}_y \text{ and } x^* = (\mathbf{X} - \overline{\mathbf{x}})/\mathbf{s}_x,$$

where $\overline{\mathbf{x}}$ is the mean observed value of property **X**, $\mathbf{s}_x$ is the observed standard deviation in the values of property **X**, $\overline{\mathbf{y}}$ is the mean observed value of property **Y**, and $\mathbf{s}_y$ is the observed standard deviation in the values of property **Y**. Substituting these values into the standardized model, we have

$$(\mathbf{Y} - \overline{\mathbf{y}})/\mathbf{s}_y = r \bullet ((\mathbf{X} - \overline{\mathbf{x}})/\mathbf{s}_x).$$

This gives the "unstandardized" model as

$$\mathbf{Y} = (\mathbf{s}_y \bullet r) \bullet ((\mathbf{X} - \overline{\mathbf{x}})/\mathbf{s}_x) + \overline{\mathbf{y}} =$$
$$\mathbf{Y} = ((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \mathbf{X} - (((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \overline{\mathbf{x}}) + \overline{\mathbf{y}},$$

which is equivalent to the result found using calculus.

Returning to our example, Dr. M has constructed the following linear model from her observations:

$$\mathbf{Y} = 0.004\,\mathbf{X} + 1.167.$$

From this model, she then makes the following *tentative* predictions:

- for every additional minute of sleep, an increase of 0.004 of a point can be expected in grade point average; and
- students sleeping zero hours can be expected to have a grade point average of 1.167.

Why are the preceding predictions *tentative*? Because the linear model is based on sample statistics, and sample statistics carry the normal variability of sampling. Consequently, Dr. M understands that she needs to assess the statistical significance of the model. In testing the statistical significance of the model, there are two parts to the model and two separate tests of significance.

First, Dr. M can ask whether the "slope" of the model is "real"—that is, *statistically significant*—or is simply a result of normal sampling variability. If the "slope" is not "real," the predictive value of the model is questionable. Now, because the "slope" of the "unstandardized" model is based on the slope of the "standardized" model, the two models are tested simultaneously. That is, from the equation of the linear model, we have

$$\mathbf{Y} = ((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \mathbf{X} - (((\mathbf{s}_y \bullet r)/\mathbf{s}_x) \bullet \overline{\mathbf{x}}) + \overline{\mathbf{y}}.$$

This means that $b = ((\mathbf{s}_y \bullet r)/\mathbf{s}_x)$, and if r is *not* statistically significant—that is, not "really" different from zero—neither is b. Thus, the significance testing results of the *correlation coefficient*—identified previously as the *t-test*—are extended to the significance testing of the slope of both the "standardized" and "unstandardized" models. In our example, Dr. M conducted the *t-test* to assess the statistical significance of the correlation coefficient and found the correlation coefficient to be statistically significant.

Second, Dr. M can ask whether the y "intercept" of the model is "real"—that is, *statistically significant*—or is simply a result of normal sampling variability. Again, if the intercept is not "real," the predictive value of the model is questionable. Having tested the slope of the regression line for statistical significance, it remains to test the y-intercept as well. That is, we want to know if the value c is a proper representation of the relationship between the two attributes or simply reflective of the normal variability in sampling. In this case, the *null hypothesis* is that the "real" *y*-intercept of the regression line is actually zero. Here, we might pose a three-part question:

> "Why is the significance of the *y*-intercept not simply presumed given that the *correlation coefficient* was found to be statistically significant, how would it be tested, and why do we care?"

To answer the *first* question:

> While the *correlation coefficient* and the regression model slope are based on relative differences in the two properties, the *y*-intercept is an absolute value, and two models might have similar slopes, but different *y*-intercepts. Thus, the *y*-intercept of the model needs to be tested using the actual observations rather than their standardized versions.

Then, to answer the *third* question:

> If the *y*-intercept is not "correct," it cannot be used for making predictions with the model.

Returning to the *second* question:

> The *y*-intercept of the model is actually a mean of the *possible y*-intercepts based on the actual observations of the sample. That is, a possible *y*-intercept can be constructed for each of the observations of the sample by using the slope of the model to project the value of **Y** where the value of **X** is zero. The mathematical description of these projections is presented in Box 14.2.

Now, how might we represent the "normal variability of sampling" with regard to the *y*-intercept of the sample? The answer: We can use the *variability* of the projected *y*-intercepts. That is, we can

a) compare each of the projected *y*-intercepts with the intercept of the model *c*;

b) square all of the differences; and

c) find the average (mean) of the squared differences.

The result is said to be the *standard error* of the *y*-intercept. With this, we can use the *t-test* for the comparison of two means, where the *t-statistic* (Chapter 12) is

$$t = (c - 0) / \text{"standard error of the differences."}$$

Fortunately, we can use a formula based on r, n, $\mathbf{s}_y$, $\bar{\mathbf{x}}$, and $\mathbf{s}_x$ to describe the standard error of the *y*-intercept:

$$\sqrt{}\ (((1 - r^2) \bullet n \bullet \mathbf{s}_y) / (n - 2)) \bullet ((1/n) + (\bar{\mathbf{x}}^{\,2} / (n \bullet \mathbf{s}_x))).$$

BOX 14.2

In mathematical terms, hypothetical projections of the y-intercept are made in the following way. Let x_1 be an observed value of **X**, and let y_1 be its observed co-occurring value of **Y**. Then, let x_2 be a *given* hypothetical value of **X**, and let y_2 be its *unknown*, hypothetical co-occurring value of **Y**. From the regression model, we have

$$b = (y_1 - y_2) / (x_1 - x_2), \text{ so}$$

$$b \bullet (x_1 - x_2) = (y_1 - y_2), \text{ and}$$

$$(b \bullet (x_1 - x_2)) - y_1 = -\, y_2, \text{ or}$$

$$y_1 - (b \bullet (x_1 - x_2)) = y_2.$$

Now, suppose the given hypothetical value x_2 is zero. Thus, its corresponding unknown hypothetical y-value (y_2) is a projected *hypothetical* y-intercept. In a formula, this projection would be

$$y_2 = y_1 - (b \bullet x_1).$$

So, for every observed value of **X** and its observed co-occurring value of **Y**, a hypothetical y-intercept can be projected.

This gives us the *t-statistic*:

$$t = c / \surd\ (((1 - r^2) \bullet n \bullet \mathbf{s}_y) / (n - 2)) \bullet ((1 / n) + (\bar{\mathbf{x}}\ ^2 / (n \bullet \mathbf{s}_x))).$$

With this, we can use the *t-probability model* to assess the probability (*p-value*) of having drawn two samples from the same population with the different means of *c* and 0. If that probability is very low—less than 0.05—we can conclude that the two samples probably were not drawn from the same sample, and the differences between these means is statistically significant. As to which version of the *t-Probability Model* to use in testing the *t-statistic*, it would be the version appropriate for (n – 2) degrees of freedom (i.e., applying the Bessel Correction for property **X** and property **Y**). In the case of Dr. M, her *t-statistic* was

$$t = 1.167 / 0.371 = 3.147.$$

Then, consulting the *t-Probability Model* for (100 – 2) = 98 degrees of freedom, she finds the *p-value* to be 0.02. Because this *p-value* is less than 0.05, Dr. M is confident in concluding that the "real" *y*-intercept of the model *probably is not zero*, and the *y*-intercept of her model *is statistically significant*.

14.6 ASSESSING THE EXPLANATORY VALUE OF THE MODEL

Finally, for an *explanatory* study, we can address the "ultimate" question: "To what extent can the differences (variability) observed in the values of the explanatory property **X** be used to *explain* the differences (variability) observed in the values of the behavioral property **Y**?" This is said to be an assessment of the "fit" of the model, with a reference to the extent to which observations of the sample—or the "dots" on the *scatter plot*—*are* consistent with the regression line:

- a "perfect" fit is one in which *every* observation—or dot—is consistent with the regression model;
- a "less perfect" fit is one in which the differences between the observations and the regression model are relatively small; and
- a "poor" fit is one in which the differences between the observations and the regression model are large.

In standard statistical practice, a useful way of assessing the "fit" of a regression model is to employ the *method of moments* to find the "typical" variation of the actual observations to the projections of the model. As a first step, we need to define what we mean by an "error" in a prediction of the model. Suppose we have phenomena with the values (x, y) for properties **X** and **Y**. In terms of the predictions of the linear model, we have $y(x)$ is the expected value of **Y** for the given value x. We then have the following:

$$y = y(x) + e, \text{ or}$$

$$e = y - y(x),$$

where e is the *error* in the prediction of the model. This error is also said to be a *residual*, and it represents the "misfit" of the model for that particular phenomenon.

Now, returning to our question of variability, we can employ the method of moments to describe the variability in each observation of **Y** as $(y - \bar{\mathbf{y}})$. Furthermore, we can express this difference in terms of the projected value of **Y** and the consequent error:

$$(y - \bar{\mathbf{y}}) = (y - y(x)) + (y(x) - \bar{\mathbf{y}}).$$

Moreover, with some algebraic manipulation (see Box 14.3 below for an explanation), we have the mathematical fact that

$$\sum(y - \bar{\mathbf{y}})^2 = \sum(y - y(x))^2 + \sum(y(x) - \bar{\mathbf{y}})^2.$$

Here, we have the following interpretations:

- $\sum(y - \bar{\mathbf{y}})^2$ is said to be the "total sum of squares," and it represents the "total variability in **Y**."
- $\sum(y(x) - \bar{\mathbf{y}})^2$ is said to be the "regression sum of squares," and it represents the "variability in **Y** explicitly related to the variability in **X**." Moreover, it is said to be the variability in the behavioral property "explained" by the explanatory property.

- $\sum(y - y(x))^2$ is said to be the "sum of the squares of the residuals," and it represents the "total of the errors in the model predictions." It is also said to be the "unexplained" variability in the behavioral property.

This gives us the following:

total variability in **Y** =
variability in **Y** directly related to the variability in **X** ("fit") +
total model errors ("misfit").

Now, with some additional algebraic manipulation, we have

("total sum of squares"/"total sum of squares") =
("regression sum of squares"/"total sum of squares") +
("sum of the squares of the residuals"/"total sum of squares").

In this way, we have

1 = percentage of total variability in **Y** explained by the variability in **X** +
percentage of total variability in **Y** unexplained.

Finally, we have the following definition *and* mathematical fact:

- The term ("regression sum of squares"/"total sum of squares") is said to be the *coefficient of determination*, and it is denoted as $\mathbf{R}^2$. It represents the percentage of total variability in the behavioral property "explained" by the variability in the explanatory property.
- $\mathbf{R}^2$ is mathematically equal to the square of the correlation coefficient *r* (see Box 14.3).

This relationship is often written as

1 = percentage explained variability + percentage unexplained variability, or
1 = $\mathbf{R}^2$ + percentage unexplained variability.

Interpreted as a percentage, the $\mathbf{R}^2$ value of a regression model is typically interpreted in the following way:

- If the value of $\mathbf{R}^2$ is 0.6 or greater, the variability in the explanatory property is seen to explain 60% of the variability in the behavioral property. In such cases, the model is judged to offer a "good" explanation of the behavioral property.

- If the value of **R²** is less than 0.6, the variability in the explanatory is seen to explain less than 60% of the variability in the behavioral property. We would typically interpret this to mean that other properties may be involved in "determining" the values of the behavioral property, and the model does not offer a strong explanation of the behavioral property.

Returning to the example of Dr. M's study of sleeping times and grade point averages, the **R²** value of her regression model is $(0.427)^2 = 0.182$, thereby suggesting that the variability in sleeping times explains only 18.2% of the variability in grade point averages. However, this conclusion remains tentative. Why? Because in representing the "explained" variability in a behavioral property, the *coefficient of determination* is placed in contrast to the "unexplained" variability in that property, and it may be the case that the difference between these two "variabilities" is simply the result of the normal variability of sampling. Consequently, the difference between these two variances needs to be tested for its statistical significance, and an appropriate method for doing so is with an *Analysis of Variance*.

Of course, one might ask why this is necessary. That is, because the *coefficient of determination* is the square of the *correlation coefficient*, one might reasonably suppose the statistical significance of the *correlation coefficient* would signal the statistical significance of the *coefficient of determination*. While this logic has some merit, it misses the differences in the two scenarios:

- the significance testing of the correlation coefficient compares the regression model with a non-association; and
- the significance testing of the *coefficient of determination* compares the "explained" variability in **Y** with the "unexplained" variability in **Y**.

Now, it may be recalled that the *Analysis of Variance* is based on the premise that two samples (*S1* and *S2*) have been drawn from the *same hypothetical population*, but they have *different* variances (Var1 and Var2). The difference between the two sample variances is represented as their ratio, and this ratio is said to be the *F-statistic*:

$$\text{Var1}/\text{Var2} = F.$$

From this, the *F-Probability Model* can then be used to find the probability of having selected such samples yielding such a ratio. It may also be recalled that the *F-Probability Model* has different versions based on the *degrees of freedom* of the two variances; if Var1 has *df1* degree of freedom and Var2 has *df2* degrees of freedom, the appropriate *F-Probability Model* is designated as (*df1*, *df2*). Furthermore, the *F-statistic* will have the expected value of

$$df2/(df2-2).$$

In applying the *F-test* to the *coefficient of determination*, we have the following:

- Var1 is the "regression sum of squares," which has 1 degree of freedom. Why 1 degree of freedom? Because all of the values constituting the "regression sum of squares" are determined by the slope of the model—a constant—and the choice of the *y*-intercept.
- Var2 is the mean of the "residual sum of squares," which has 2 degrees of freedom. Why the mean? Because the "residual sum of squares" represents a set of *n different* variances (*n*), and the mean represents the "typical" variance. Applying the Bessel Correction, we get *df2* = *n* – 2.

Applying the *F-test* to her regression model, Dr. M finds the following *F-value*:

1) Var1 = 7.691, with *df1* = 1 degree of freedom.
2) The mean "residual sum of squares" is 34.455/(*df2*), with *df2* = (100 – 2) = 98. This gives Var2 as 34.455/98 = 0.352.
3) The *F-value* is Var1/Var2 = 7.691/0.352 = 21.849.

Consulting the *F-Probability Model* for (1, 98), she finds the following probability, or *p-value*:

$$p < 0.001.$$

Thus, the probability that these two sets of variances represent samples drawn from the same hypothetical population is very low. Consequently, Dr. M is relatively confident the two sets of variances probably *do not* represent samples drawn from the *same* hypothetical population and that the sample represented by the "regression sum of squares" and the sample represented by the "mean residual sum of squares" were not drawn from the same hypothetical population; thus, she is relatively confident that the difference between these two variances is statistically significant. With this result, Dr. M has the following interpretation of her study:

- Students who sleep more than the group average of 470.2 minutes per night are more likely to have better grade point averages. Moreover, grade point averages tend to improve by 0.004 point for every additional minute of sleep. Furthermore, these results are not likely due to normal sampling variability.
- Although sleeping time seems to have a direct relationship with academic performance, sleeping time explains only 18.2% of the variability in the grade point averages in the group of students studied. Thus, other factors may also be involved.

BOX 14.3

An explanation of the "fit" of a model can be developed from the standardized versions of the properties of interest. Suppose we are interested in comparing the typical variation in property **Y** with

- the variability directly related to the variability in property **X**; and
- the variability unrelated to the variability in the property **X**.

For simplicity, we can consider **X** and **Y** in their standardized forms *x** and *y**. Then, for every observed value of *y**, we have

$$y^* = rx^* + e,$$

where

- $\dot{x}$ is the co-occurring value of *x**;
- $r\dot{x}$ is the value of *y** projected by the model; and
- *e* is the error, or misfit, with the model.

Now, using the method of moments, we can find the typical variability of the values of *y** as the *variance*:

$$(\sum y^{*2}) / n = 1.$$

Moreover, this is equal to $(\sum (rx^* + e)^2)/n$. Now, if we expand $(rx^* + e)^2$ for each value of *y**, we have

$$r^2 x^{*2} + 2rx \bullet e + e^2.$$

In summary, for all the observed values of *y**, we have

$$r^2 \sum x^{*2} + 2r \sum x^* e + \sum e^2 =$$

$$r^2 \sum x^{*2} + 2r(\sum e\ (\sum x^*)) + \sum e^2.$$

But, we also have the following:

- because $(\sum x^{*2}) / n = 1$, $\sum x^{*2} = n$; and
- because $(\sum x^*) / n = 0$, $\sum x^* = 0$.

Thus, we have

$$r^2 \sum x^{*2} + 2r(\sum e\ (\sum x^*)) + \sum e^2 = nr^2 + \sum e^2.$$

This gives us

$$(\sum y^{*2}) / n = (nr^2 + \sum e^2) / n.$$

Finally, because $(\sum y^{*2}) / n = 1$, we have

$$1 = r^2 + (\sum e^2) / n.$$

In this formulation,

- 1 = the total variability in *y**;
- r^2 = the variability in *y** related to the variability in *x** as a percentage; and
- $(\sum e^2) / n$ represents the variability in *y** *not* related to the variability in *x**, also as a percentage.

Now, to test the statistical significance of r^2, we have the following:

- total sum of squares = $(\sum y^{*2})$;
- regression sum of squares = nr^2;
- mean residual sum of squares = $(\sum e^2) / n = 1 - r^2$; and
- *F* = regression sum of squares / mean residual sum of squares = $nr^2 / (1 - r^2)$.

For the "sleep" and "grade point average" example, $r = 0.427$ and $n = 98$ (with the Bessel Correction). This gives us the following:

a) $r^2 = 0.182$;

b) $nr^2 = 98 \bullet 0.182 = 17.836$;

c) $1 - r^2 = 1 - 0.182 = 0.818$; so

d) $F = 17.836 / 0.818 = 21.804$, which is within rounding error of the *F-value* of 21.849 obtained using the unstandardized values of the two properties.

14.7 Summary

1) The objective of an *explanatory study* is to assess the extent to which the different values observed in one property (**Y**, the behavioral property) of a set of *n* phenomena tend to co-occur with different values of another property (**X**, the explanatory property) of those phenomena. Where both properties are quantitative, we have the following:

- In a *direct* relationship, high values of property **X** tend to co-occur with high values of property **Y**, and low values of property **X** tend to co-occur with low values of property **Y**.
- In an *inverse* relationship, high values of property **X** tend to co-occur with low values of property **Y**, and low values of property **X** tend to co-occur with high values of property **Y**.
- In a mixed relationship, or "non-relationship," high values of property **X** tend to co-occur with both high and low values of property **Y**, and low values of property **X** tend to co-occur with both low and high values of property **Y**. Such a "non-relationship" is said to be *stochastic independence*.

2) The covariability of two properties may be visually assessed using a *contingency table*. However, when both of the properties are quantitative, a more practical visual display is the *scatter plot* using Cartesian coordinates.

3) When both co-occurring properties of a set of phenomena are quantitative, their co-occurrence in each phenomenon can be modeled as an interaction, and their interaction can be modeled mathematically by multiplying their respective values to yield their *product*. Using this representation, the covariability of the two properties within a set of phenomena can be quantified as the *covariance:*

$$\mathbf{Cov(XY)} = (\textstyle\sum (x - \bar{\mathbf{x}})(y - \bar{\mathbf{y}})) / (n - 1),$$

where x is the **X**-value for a phenomenon, y is the co-occurring **Y**-value for that phenomenon, $\bar{\mathbf{x}}$ is the mean value of **X** for the set of phenomena, $\bar{\mathbf{y}}$ is the mean value of **Y** for the set of phenomena, and $(n - 1)$ reflects the application of the Bessel Correction:

a) a positive value indicates a direct association;

b) a negative value indicates an inverse association; and

c) a zero value indicates a non-association, or the *stochastic independence*, of the two properties.

4) If both properties are first standardized, the "standardized" covariance is said to be the *correlation coefficient*. It is calculated as

$$(\textstyle\sum (x^*_i \bullet y^*_i)) / (n - 1),$$

where x^*_i and y^*_i are the standardized values of each pair of values (x_i, y_i). Alternatively, the correlation coefficient can also be calculated by dividing the "unstandardized" covariance by the standard deviation ($\mathbf{s}_x$) of property **X** and the standard deviation ($\mathbf{s}_y$) of property **Y**. The correlation coefficient is denoted as *r*, and its value will range between −1 and +1:

- a value between −1 and −0.6 is interpreted as a strong inverse association;
- a value between −0.59 and −0.01 is interpreted as a weak inverse association;
- a value of 0 is *defined* as a non-association, or *stochastic independence*;
- a value between 0.01 and 0.59 is interpreted as a weak direct association; and
- a value between 0.6 and 1.0 is interpreted as a strong direct association.

5) Given the normal variability of sampling, it is always possible that a non-zero correlation coefficient found for a sample set of observations might reflect such sampling variability rather than a "real" relationship between the two properties. Because the correlation coefficient is a mean, the *t-test* may be used to compare the sample correlation coefficient to a hypothetical sample with a correlation coefficient of zero representing the premise that the two properties of interest (**X** and **Y**) are

(Continued)

(Continued)

stochastically independent. The *t-statistic* is derived as

$$t = r / \surd\ ((1 - r^2) / (n - 2)),$$

and the *p-value* is found from a *t-Probability Model* for (*n* – 2) degrees of freedom. If the *p-value* is less than 0.05, we conclude that the sample probably *was not* drawn from a hypothetical population in which the two properties are stochastically independent, and the correlation coefficient of the sample probably is *not reflective of normal sampling variability*. Otherwise, we cannot rule out the role of normal sampling variability as explaining the value of the correlation coefficient.

6) In some cases, an investigator might attempt to describe a relationship observed between two quantitative properties as a mathematical function, or model. This is said to be *curve fitting,* and it is facilitated by an examination of the scatter plot. In some cases, a *linear* mathematical model might seem appropriate, and such a model would have the form

$$\mathbf{Y} = b\mathbf{X} + c.$$

In this model,

- *b* is said to be the *slope* of the line, and it represents the expected difference in **Y**-values given a difference in **X**-values. It is often represented symbolically as $\Delta y / \Delta x$.
- *c* is said to be "y-intercept," and it represents the expected value of **Y** associated with a zero value of **X**.

With such a model, the *expected* value of **Y** for a phenomenon can be projected from its co-occurring value *x* of **X**. This projected value is denoted as *y*(*x*), and it is said to be the value of the behavioral property *explained* by the explanatory property.

7) A linear model of a set of observations may be constructed analytically in two different ways, and while the approaches are different, the results are equivalent:

- Using the "correlation coefficient method," a model is constructed directly as

$$y^* = r \bullet x^*,$$

where *y** and *x** are the standardized versions of the properties **Y** and **X**. This is said to be the *standardized* version of the linear model.

- Using calculus, a hypothetical line that minimizes the square of the differences between each observed value *y* and its hypothetical projection *y*(*x*) can be constructed. The resulting model is of the form

$$\mathbf{Y} = ((\mathbf{s_y} \bullet r) / \mathbf{s_x}) \bullet \mathbf{X} - (((\mathbf{s_y} \bullet r) / \mathbf{s_x}) \bullet \bar{\mathbf{x}}) + \bar{\mathbf{y}}.$$

This is said to be the *unstandardized* version of the linear model.

Because both versions of the model are described using the correlation coefficient, and because the correlation coefficient is based on the average difference of the observed values from the mean, the model is said to be a *regression* model. In addition, because the "calculus method" is based on minimizing the squared differences between the model and the observations of the sample, the resulting model (regardless of the version) is said to be the "ordinary" least squares (or *OLS*) model.

8) Because the "relationships" suggested by an OLS model (slope and intercept) are based on sample statistics, it is possible that they simply represent the normal variability of sampling rather than a "real" association. Thus, the parts of the OLS model should be tested for their statistical significance:

- Because the slope of the model will be based on the correlation coefficient, the statistical significance of the slope is based on the statistical significance of the correlation coefficient.
- Because the *y*-intercept of the model is based on the average of a set of projected hypothetical *y*-intercepts, the *y*-intercept can be tested using the *t-test*. The premise of the *t-test* is that the *y*-intercept *c* of the model is "really" zero, and the *t-statistic* is

$$t = c / \surd\ (((1 - r^2) \bullet n \bullet s_y) / (n - 2))\ .\ ((1 / n) + (\bar{\mathbf{x}}^2 / (n \bullet s_x))).$$

- Consulting the *t-Probability Model* for $(n - 2)$ degrees of freedom, if the *p-value* is less than 0.05, we can conclude that the *y*-intercept probably is not "really" zero and that *c* is the appropriate *y*-intercept for this model. Otherwise, we cannot be confident that the proper *y*-intercept of the model is not zero.

9) Having constructed a linear model describing a set of sample observations, it is reasonable to ask the following question:

> To what extent do the observed values *y* of the behavioral property exactly conform to their "explained" values $y(x)$ projected by the model?

This is said to be the "fit" of the model, and the differences between the observed values of **Y** and the "explained values" of **Y** are said to be "errors" or "residuals." That is,

$$y = y(x) + e.$$

Now, using the method of moments, we can describe the "fit" of a model as a percentage of the total variability of the behavioral property, starting with the following mathematical relationship:

$$\sum(y - \bar{\mathbf{y}})^2 = \sum(y(x) - \bar{\mathbf{y}})^2 + \sum(y - y(x))^2.$$

In this formula, we have the following interpretation:

- $\sum(y - \bar{\mathbf{y}})^2$ represents the "total variability in **Y**" and is said to be the "total sum of squares";
- $\sum(y(x) - \bar{\mathbf{y}})^2$ represents the "variability of **Y** *explained* as its covariability with **X**," and it is said to be the "regression sum of squares"; and
- $\sum(y - y(x))^2$ represents the total "*unexplained* variability in **Y**," and it is said to be the "residual sum of squares."

From this, we have

1 = "regression sum of squares" / "total sum of squares" + "residual sum of squares" / "total sum of squares."

That is,

1 = percentage of total variability explained by the model + percentage of total variability unexplained by the model.

Furthermore, we have

"regression sum of squares" / "total sum of squares" = r^2.

In this context, r^2 is denoted as $\mathbf{R}^2$, it is said to be the *coefficient of determination*, and it is interpreted as the "percentage of total variability explained by the model":

- If $\mathbf{R}^2$ is 0.60 or above, we say that the regression model is a good fit and offers a good explanation of the variability in the behavioral property.
- If $\mathbf{R}^2$ is below 0.60, we say that the regression model is a poor fit and offers only a partial explanation of the variability in the behavioral property.

10) Understanding that $\mathbf{R}^2$ is a sample statistic subject to the normal variability of sampling, a proper interpretation of an $\mathbf{R}^2$ value will include a test of statistical significance. In this context, the test compares the difference between the "explained variability" and the "unexplained variability," and the procedure most typically used is ANOVA under the premise that the "explained variability" is not "really" different from the "*typical* unexplained variability." We would construct the following *F-statistic*:

F = "explained variability" / "*typical* unexplained variability" = ("regression sum of squares") / ("residual sum of squares" / $(n - 2)$).

Consulting the *F-Probability Model* for $(1, n - 2)$ degrees of freedom, we would interpret the *p-value* in the following way:

- If this probability is *below* 0.05, we would say that we are relatively confident the "explained variability" and the "typical

(Continued)

(Continued)

unexplained variability" probably are *not* the same and, thus, that the "fit" of the model assessed as **R**^{2} probably is *not* the result of normal sampling variability and probably does represent the "real" fit of the model.

- Otherwise, we would say that *we cannot be confident* that the "explained variability" and the "typical unexplained variability" are *not* the same and, thus, that the "fit" of the model—assessed as **R**^{2}—cannot be judged to be statistically significant.

14.8 SPSS TUTORIAL

An educational psychologist is interested in the effect that Internet use might have on students' academic performance, and her "suspicion" reflects two possible behavioral scenarios:

- students might use the Internet primarily for knowledge acquisition; or
- students might use the Internet primarily as entertainment.

Moreover, Internet use can be meaningfully assessed quantitatively as "hours per day," and academic performance can be meaningfully assessed quantitatively as "grade point average," so both properties can be assessed in comparative terms of "high" and "low" according to their respective scales of measurement. Of course, there are any number of possible patterns of co-occurrence that may be imagined—and modeled—for these two properties. However, two models are of particular interest here:

A) It may be the case that specific amounts of Internet use co-occur with specific grade point averages. Following are some examples:

  - "1 hour" of Internet use may co-occur among the students with a grade point average of 3.0, suggesting that this amount of Internet use is useful;
  - "2 hours" of Internet use may co-occur among the students with a grade point average of 2.5, suggesting that this amount of Internet use is a distraction;
  - "3 hours" of Internet use may co-occur with a grade point average of 3.5, suggesting that this amount of Internet use is useful; and
  - "4 hours" of Internet use may co-occur among the students with a grade point average of 2.0, suggesting that this amount of Internet use is a distraction.

B) It may be the case that comparative values of Internet use co-occur with similarly comparative values for grade point averages:

  - If increasingly higher amounts of Internet use co-occur with increasingly higher grade point averages, we would say that this represents a *direct* relationship. This would suggest that the Internet is a useful learning tool.

- If increasingly higher amounts of Internet use co-occur with increasingly lower grade point averages, we would say that this represents an *inverse* relationship. This would suggest that the Internet is a distraction.

With these analytical models in mind, the psychologist collects a random sample of 50 students, asking each student his or her daily Internet use and grade point average. The results are found in Table 14.2.

TABLE 14.2 ■ Internet Use (hours daily) and GPA for a Sample of 50 Students

6 hours; 2.2	3 hours; 3.2	4 hours; 2.5	3 hours; 3.0	5 hours; 2.25
2 hours; 3.25	2 hours; 3.0	1 hour; 3.6	0 hours; 3.6	0 hours; 4.0
2 hours; 3.4	3 hours; 3.0	6 hours; 1.5	4 hours; 2.0	3 hours; 2.5
3 hours; 2.0	4 hours; 2.0	4 hours; 2.0	5 hours; 3.0	2 hours; 3.5
5 hours; 2.0	6 hours; 1.8	7 hours; 1.8	3 hours; 3.0	7 hours; 1.9
7 hours; 1.75	3 hours; 3.2	2 hours; 3.0	4 hours; 3.0	5 hours; 2.75
7 hours; 2.5	2 hours; 3.1	2 hours; 3.2	1 hour; 4.0	1 hour; 3.55
1 hour; 3.75	1 hour; 3.8	3 hours; 3.0	2 hours; 3.5	3 hours; 2.0
4 hours; 3.5	6 hours; 1.9	8 hours; 1.9	4 hours; 3.2	2 hours; 3.0
3 hours; 2.5	2 hours; 3.0	1 hour; 3.6	0 hours; 3.9	0 hours; 4.0

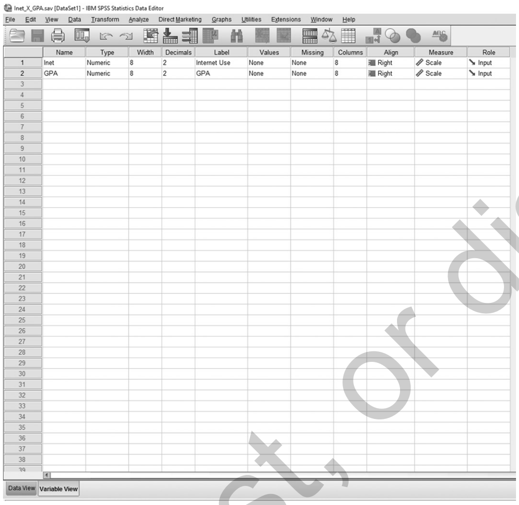
Now, with these observations, the psychologist first needs to determine which model she is going to adopt in order to analyze the observations for a potential relationship between the two properties:

- If Model "A" is adopted, the potential relationship between these two properties may be assessed using the *Analysis of Variance* techniques described in Chapter 13. That is, the students could be grouped together according to their hours of Internet use, and the average GPA could then be assessed for each group.
- If Model "B" is adopted, the potential relationship between the two properties is expressed in terms of the *relative differences* in one property *compared with* the *relative differences* in the other property, and the measure of association relevant for such models is the *covariance* as discussed in the current chapter.

With these modeling options in mind, the psychologist enters the observations into an SPSS data file for further analysis.

1) Entering the "Variable View," the psychologist defines her two variables as "Inet" and "GPA." Both are identified as "numeric" in type, both are set at a "width" of 8 characters, and both are set at "2 Decimals" in precision. Furthermore, each is given a label for the output, and each is identified as a "scale" measure (Screenshot 14.1).

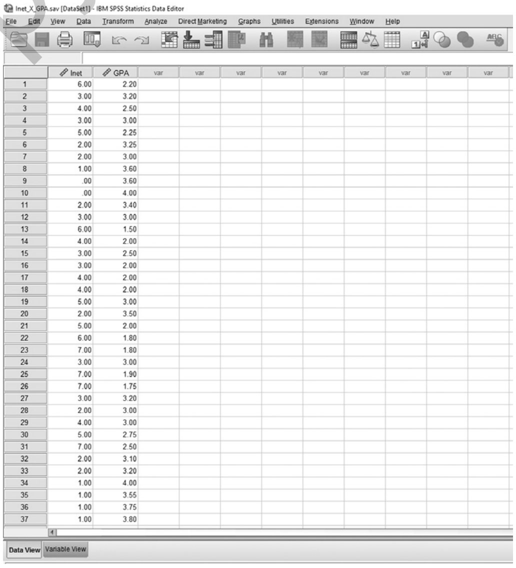
SCREENSHOT 14.1


Inet_X_GPA.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help
Data View
Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Inet	Numeric	8	2	Internet Use	None	None	8	Right	Scale	Input
2	GPA	Numeric	8	2	GPA	None	None	8	Right	Scale	Input

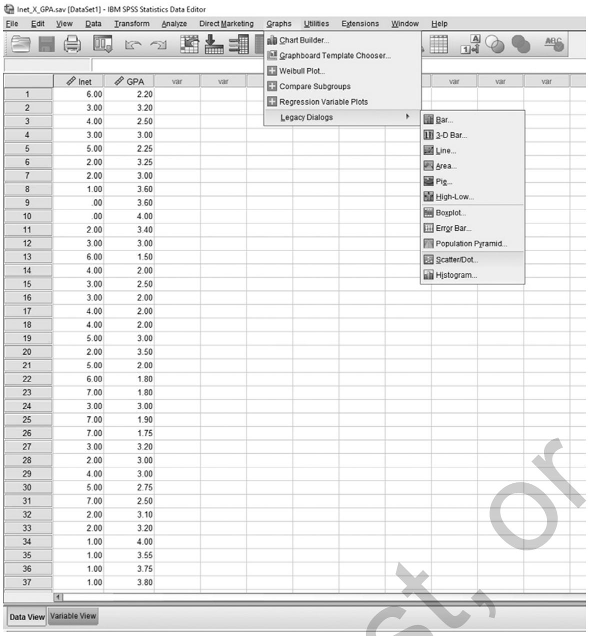
2) In the "Data View," the psychologist then enters the observations, with each student representing a case/row/record. The first student's Internet use is entered as 6, and his or her GPA is entered as 2.2 (Screenshot 14.2).

SCREENSHOT 14.2


Inet_X_GPA.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help
Data View
Variable View

	Inet	GPA	var	var	var	var	var	var	var	var	var
1	6.00	2.20									
2	3.00	3.20									
3	4.00	2.50									
4	3.00	3.00									
5	5.00	2.25									
6	2.00	3.25									
7	2.00	3.00									
8	1.00	3.60									
9	.00	3.60									
10	.00	4.00									
11	2.00	3.40									
12	3.00	3.00									
13	6.00	1.50									
14	4.00	2.00									
15	3.00	2.50									
16	3.00	2.00									
17	4.00	2.00									
18	4.00	2.00									
19	5.00	3.00									
20	2.00	3.50									
21	5.00	2.00									
22	6.00	1.80									
23	7.00	1.80									
24	3.00	3.00									
25	7.00	1.90									
26	7.00	1.75									
27	3.00	3.20									
28	2.00	3.00									
29	4.00	3.00									
30	5.00	2.75									
31	7.00	2.50									
32	2.00	3.10									
33	2.00	3.20									
34	1.00	4.00									
35	1.00	3.55									
36	1.00	3.75									
37	1.00	3.80									

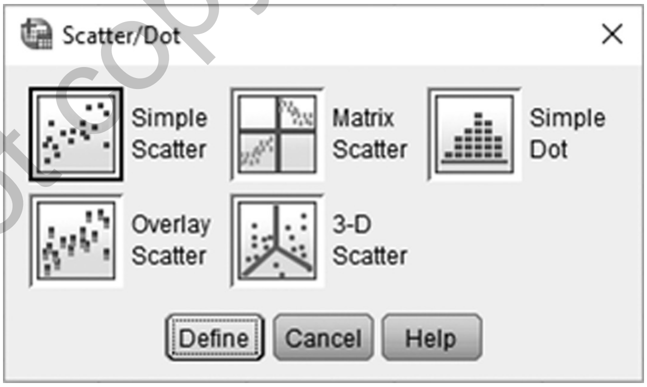
3) As a first step of analysis, the psychologist constructs a scatter plot. This is done by choosing "Graphs" from the task bar. This is followed by "Legacy Dialogs" and "Scatter/Dot" (Screenshot 14.3).


Inet_X_GPA.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Extensions Window Help
Chart Builder...
Graphboard Template Chooser...
Weibull Plot...
Compare Subgroups
Regression Variable Plots
Legacy Dialogs
Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...
Data View
Variable View

	Inet	GPA	var	var	var	var	var
1	6.00	2.20					
2	3.00	3.20					
3	4.00	2.50					
4	3.00	3.00					
5	5.00	2.25					
6	2.00	3.25					
7	2.00	3.00					
8	1.00	3.60					
9	.00	3.60					
10	.00	4.00					
11	2.00	3.40					
12	3.00	3.00					
13	6.00	1.50					
14	4.00	2.00					
15	3.00	2.50					
16	3.00	2.00					
17	4.00	2.00					
18	4.00	2.00					
19	5.00	3.00					
20	2.00	3.50					
21	5.00	2.00					
22	6.00	1.80					
23	7.00	1.80					
24	3.00	3.00					
25	7.00	1.90					
26	7.00	1.75					
27	3.00	3.20					
28	2.00	3.00					
29	4.00	3.00					
30	5.00	2.75					
31	7.00	2.50					
32	2.00	3.10					
33	2.00	3.20					
34	1.00	4.00					
35	1.00	3.55					
36	1.00	3.75					
37	1.00	3.80					

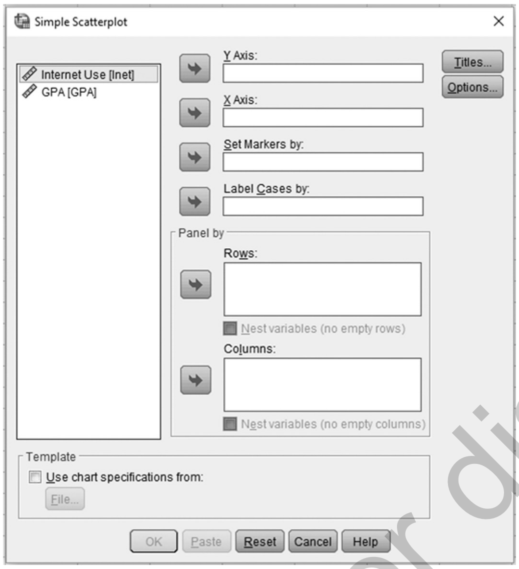
This leads to the Scatter/Dot menu. Here the psychologist chooses "Simple Scatter" (Screenshot 14.4).

SCREENSHOT 14.4



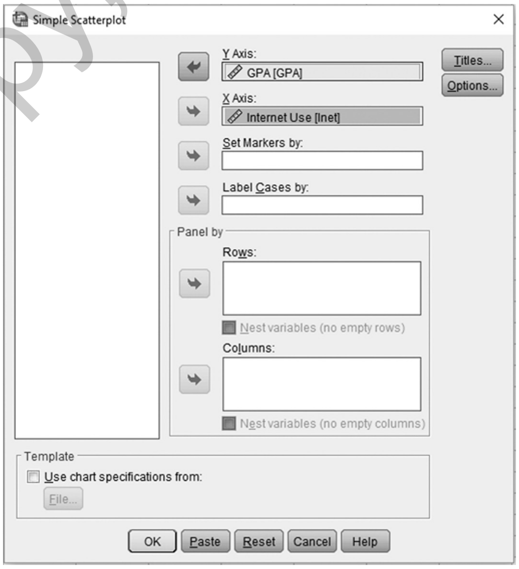
In the Scatter/Dot menu, the psychologist selects the "Define" option, which leads to the Simple Scatterplot menu (Screenshot 14.5). Here, the two variables of the dataset are identified.

SCREENSHOT 14.5



The psychologist then identifies the variable "GPA" as the Y-axis and identifies "Internet use" as the X-axis. This reflects the presumption that "Internet use" is the *explanatory* property and GPA is the *behavioral* property in this investigatory model. Having completed the definition, the "OK" option is chosen (Screenshot 14.6).

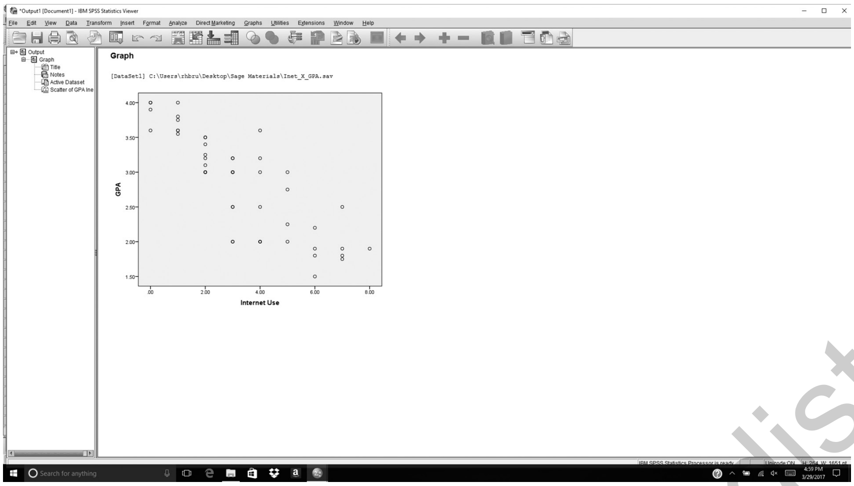
SCREENSHOT 14.6


Y Axis: GPA [GPA]

X Axis: Internet Use [Inet]

This results in the desired scatter plot (Screenshot 14.7).

SCREENSHOT 14.7


*Output1 [Document1] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Extensions Window Help
Output
Graph
Title
Notes
Active Dataset
Scatter of GPA Ine
GPA
4.00
3.50
3.00
2.50
2.00
1.50
.00
2.00
4.00
6.00
8.00
Internet Use
Search for anything

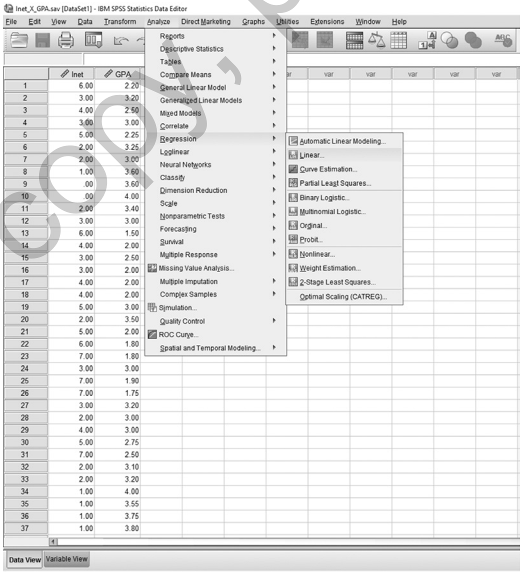
Graph

```
[DataSet1] C:\Users\rhbru\Desktop\Sage Materials\Inet_X_GPA.sav
```

On review, the pattern of dots appears to support a potential linear association, so the psychologist proceeds on this basis.

4) Returning to the "Data View," the psychologist proceeds to analyze her observations as a regression model. From the task bar, she chooses "Analyze," "Regression," and "Linear" (Screenshot 14.8).

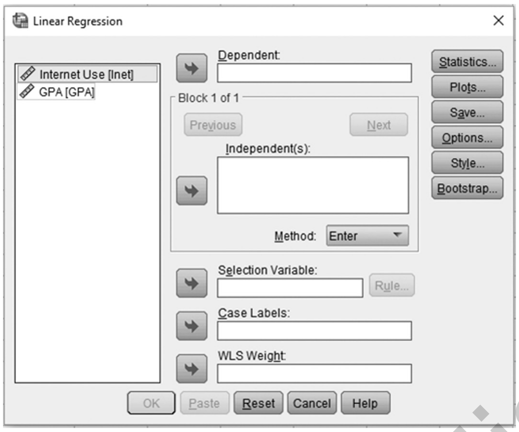
SCREENSHOT 14.8



	Inet	GPA
1	6.00	2.20
2	3.00	3.20
3	4.00	2.50
4	3.00	3.00
5	5.00	2.25
6	2.00	3.25
7	2.00	3.00
8	1.00	3.60
9	.00	3.60
10	.00	4.00
11	2.00	3.40
12	3.00	3.00
13	6.00	1.50
14	4.00	2.00
15	3.00	2.50
16	3.00	2.00
17	4.00	2.00
18	4.00	2.00
19	5.00	3.00
20	2.00	3.50
21	5.00	2.00
22	6.00	1.80
23	7.00	1.80
24	3.00	3.00
25	7.00	1.90
26	7.00	1.75
27	3.00	3.20
28	2.00	3.00
29	4.00	3.00
30	5.00	2.75
31	7.00	2.50
32	2.00	3.10
33	2.00	3.20
34	1.00	4.00
35	1.00	3.55
36	1.00	3.75
37	1.00	3.80

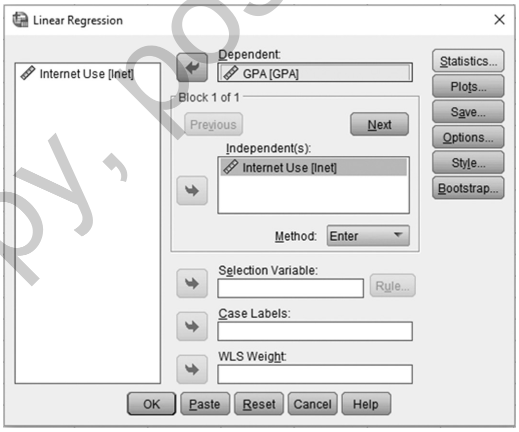
This leads to the "Linear Regression" dialog box. Here the two variables of the dataset are listed (Screenshot 14.9).

SCREENSHOT 14.9



The psychologist then identifies "Internet use" as the independent variable and GPA as the dependent variable (Screenshot 14.10). She also clicks the "Statistics" option.

SCREENSHOT 14.10



Dependent: GPA [GPA]

Independent(s): Internet Use [Inet]

Method: Enter

In the "Linear Regression Statistics" menu, she identifies the following (Screenshot 14.11):

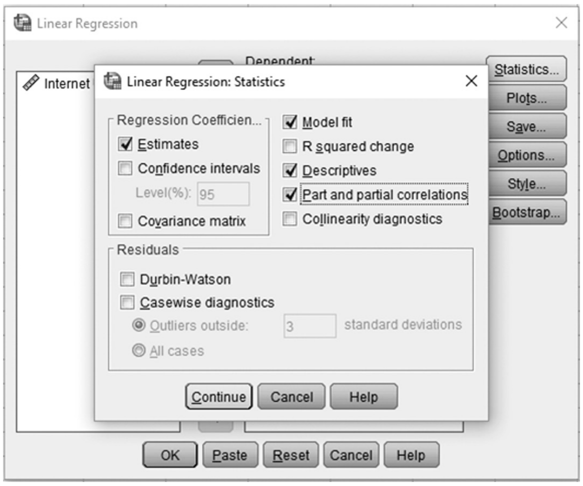
a) for "Regression Coefficients," she chooses "Estimates";

b) she chooses "Model Fit";

c) she chooses "Descriptives"; and

d) she chooses "Part and Partial Correlations."

SCREENSHOT 14.11



Linear Regression

Internet

Dependent:

Linear Regression: Statistics

Regression Coefficients

☑ Estimates

□ Confidence intervals

Level(%): 95

□ Covariance matrix

☑ Model fit

□ R squared change

☑ Descriptives

☑ Part and partial correlations

□ Collinearity diagnostics

Residuals

□ Durbin-Watson

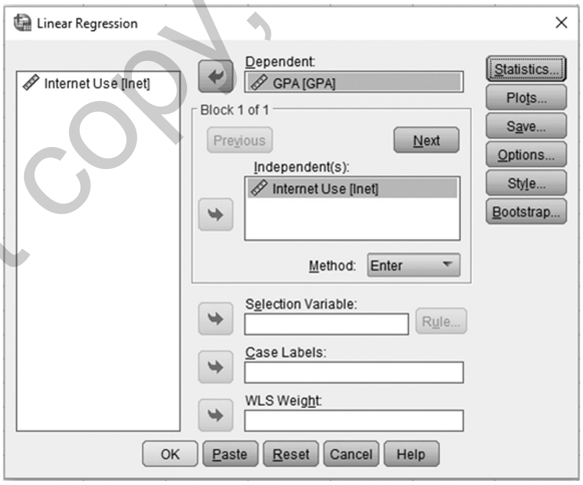
□ Casewise diagnostics

● Outliers outside: 3 standard deviations

○ All cases

She then completes the “Linear Regression Statistics” dialog by choosing “Continue.” This returns her to the “Linear Regression” dialog box (Screenshot 14.12). When she chooses “OK” in this dialog box, the program is started.

SCREENSHOT 14.12



Linear Regression

Internet Use [Inet]

Dependent: GPA [GPA]

Block 1 of 1

Independent(s): Internet Use [Inet]

Method: Enter

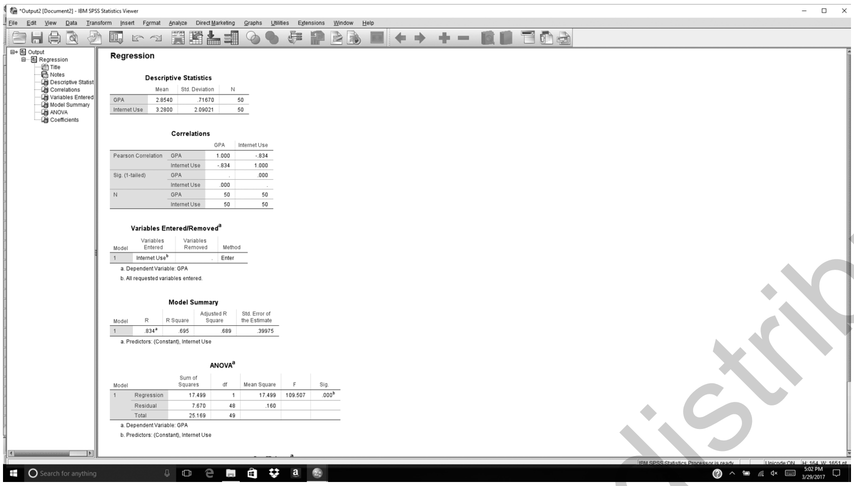
Selection Variable:

Case Labels:

WLS Weight:

When the program completes its calculations, the requested statistical reports are generated (Screenshots 14.13 and 14.14).

SCREENSHOT 14.13


*Output2 [Document2] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Extensions Window Help
Output
Regression
Title
Notes
Descriptive Statist
Correlations
Variables Entered
Model Summary
ANOVA
Coefficients
Search for anything

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
GPA	2.8540	.71670	50
Internet Use	3.2800	2.09021	50

Correlations

		GPA	Internet Use
Pearson Correlation	GPA	1.000	-.834
	Internet Use	-.834	1.000
Sig. (1-tailed)	GPA	.	.000
	Internet Use	.000	.
N	GPA	50	50
	Internet Use	50	50

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Internet Use[b]	.	Enter

a. Dependent Variable: GPA

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.834[a]	.695	.689	.39975

a. Predictors: (Constant), Internet Use

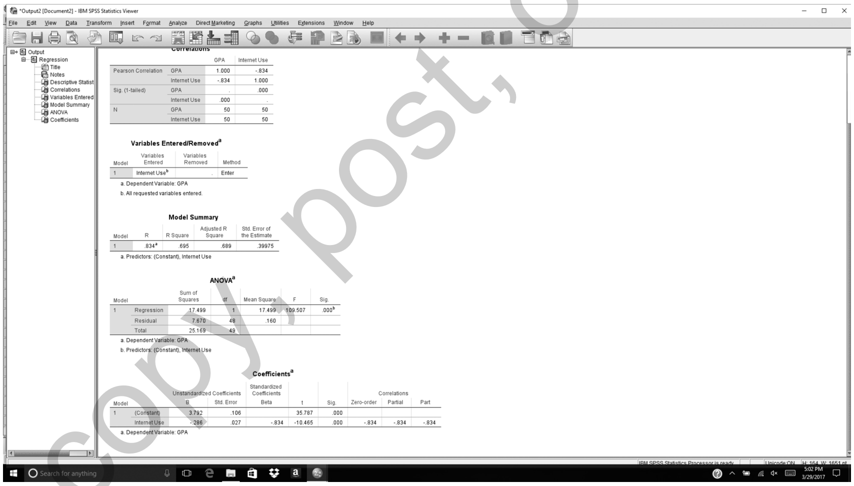
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17.499	1	17.499	109.507	.000[b]
	Residual	7.670	48	.160		
	Total	25.169	49			

a. Dependent Variable: GPA

b. Predictors: (Constant), Internet Use

SCREENSHOT 14.14


*Output2 [Document2] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Extensions Window Help
Output
Regression
Title
Notes
Descriptive Statist
Correlations
Variables Entered
Model Summary
ANOVA
Coefficients
Search for anything

		GPA	Internet Use
Pearson Correlation	GPA	1.000	-.834
	Internet Use	-.834	1.000
Sig. (1-tailed)	GPA	.	.000
	Internet Use	.000	.
N	GPA	50	50
	Internet Use	50	50

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Internet Use[b]	.	Enter

a. Dependent Variable: GPA

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.834[a]	.695	.689	.39975

a. Predictors: (Constant), Internet Use

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	17.499	1	17.499	109.507	.000[b]
	Residual	7.670	48	.160		
	Total	25.169	49			

a. Dependent Variable: GPA

b. Predictors: (Constant), Internet Use

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients			Correlations		
Model		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	3.792	.106		35.787	.000			
	Internet Use	-.286	.027	-.834	-10.465	.000	-.834	-.834	-.834

a. Dependent Variable: GPA

5) The psychologist can then interpret the analytical results:

a) From the “Descriptive Statistics” report, the psychologist finds that the typical (mean) amount of daily Internet use among this set of students is 2.854 hours, and the typical (mean) grade point average among the students is 3.28.

b) From the “Correlations” report, the psychologist finds that the correlation coefficient (“Pearson Correlation”) between the properties “Internet Use” and “GPA” is assessed as −0.834. This result is found at the intersection of the row labeled “Pearson Correlation, GPA” and the column labeled “Internet Use.” Using the “standard” interpretation, this correlation is said to be a “strong inverse” association, meaning that those individuals using “greater” amounts of time on the Internet are likely to have “lower” grade point averages. This result supports the behavioral model suggesting that Internet use is more likely to be a distraction than a learning tool.

c) Also from the “Correlations” report, the psychologist finds the result of the *t-test* comparing the sample correlation coefficient of −0.834 with a hypothetical correlation coefficient of zero representing a hypothetical sample drawn from a population in which two properties are *stochastically independent*. The result of this *t-test* is found in the intersection of the row labeled “Sig, GPA” and the column labeled “Internet Use.” The value of 0.000 indicates that the probability of the actual sample and the hypothetical sample being drawn from the same population is less than 0.0005. This probability is sufficiently low—less than 0.05—for the psychologist to be relatively confident that the two samples were not drawn from such a population and that the sample correlation coefficient *is not due to normal sampling variability* and is, indeed, statistically significant. In formal terms, the *null hypothesis*—that the sample correlation coefficient is simply a product of normal sampling variability and not indicative of an association between the two properties—*may be confidently rejected*.

d) Given the strength and statistical significance of the association between these two properties, the psychologist feels reasonably confident in the explanatory value of this linear model of this association. In terms of the properties Internet use and grade point average, the model is

$$\text{GPA} = (b \bullet \text{Internet use}) + c.$$

Using this model, the likely difference between two students’ grade point averages can be predicted by their different Internet use habits. This linear model—based on the correlation coefficient—can be found in the “Coefficients” report:

- First, the *slope* of the estimated regression line is reported in the intersection of the row labeled “Internet Use” and the column labeled “Unstandardized Coefficients, B.” The assessed value of this coefficient is −0.286, and it may be interpreted as suggesting that if one student’s rate of Internet use is 1 hour greater than another student’s rate of Internet use, the student with the greater rate of Internet use is likely to have a GPA that is 0.286 points *lower* than the student with the lesser rate of Internet use. Conversely, the model suggests that if one student’s rate of Internet use is 1 hour less than another student’s rate of Internet use, the student with the lesser rate of Internet use is likely to have a GPA that is 0.286 points *higher* than the student with the higher rate of Internet use.

- Furthermore, in the intersection of the row labeled “Internet Use” and the column labeled “Sig” is a replication of the result of the *t-test* for statistical significance previously reported for the correlation coefficient, and its interpretation is the same. Given that the linear model was constructed from the correlation coefficient, the interpretation of result of the *t-test* is the same for the coefficient of the linear model as for the correlation coefficient. That is, the slope of the linear model is not simply a reflection of the normal variability of sampling and is statistically significant.

- As for the “constant” term of the linear model—or *y*-intercept—the same question of statistical significance can be addressed. In the “Coefficients” report, in the intersection of the row labeled “Constant” and the column labeled “Unstandardized Coefficients, B” is the value of the constant term in the linear model. It can be interpreted as the expected grade point average of the hypothetical student who typically uses the Internet for zero hours, and it is assessed as 3.792. In the intersection of the row labeled “Constant” and the column labeled “Sig” is the result of the *t-test* assessing the probability that the “real” *y*-intercept of this model is zero and that the assessed value of 3.792 is simply a reflection of the normal variability of sampling. In this case, the result of the *t-test* is a probability of less than 0.0005, and the psychologist interprets this to suggest that the probability that the “real” y-intercept is zero is sufficiently low for her to conclude that the “real” *y*-intercept is not zero and the assessed value of the *y*-intercept is statistically significant.

- Having affirmed the statistical significance of the constant term of the model, the model can be used to predict the likely grade point average of any student based on his or her Internet use. That is, if a student uses the Internet for *x* hours, his or her expected grade point average will be

$$3.792 + (-0.286 \bullet x) = 3.792 - (0.286 \bullet x),$$

in comparison with the hypothetical student who uses the Internet for zero hours. Thus, for a student who typically uses the Internet for 1 hour a day, his or her expected grade point average is

$$\text{GPA of hypothetical student at zero Internet hours} -$$

$$0.286 \bullet 1 \text{ hour} =$$

$$3.792 - (0.286 \bullet 1) = 3.506.$$

- As a note of interest, the “Coefficients” report also contains a column identified as “Standardized Coefficients.” In this column are the coefficients of the “standardized” version of the regression line standardized GPA = $r \bullet$ standardized Internet use,

where *r* is the correlation coefficient. In statistical practice, the coefficient of a linear model expressed in its standardized version is identified as "Beta," and the coefficient of a linear model in its unstandardized version is identified as "*b*." Moreover, in the standardized version of the linear model, the "constant" is defined to be zero and, thus, is excluded from the column of "Standardized Coefficients."

6) Finally, the psychologist can assess the extent to which the association model can be used to explain the differences observed among the students' grade point averages:

   a) In the report identified as the "Model Summary," the coefficient of determination is denoted in the column labeled "R Square." Its assessed value is 0.695, and this is interpreted as the percentage of the variation among the students' grade point averages "explained" as the covariability with their Internet use. Following standard statistical practice, the psychologist interprets this percentage to indicate that the model provides a "good" explanation of that variability.

   b) Turning to the report identified as "ANOVA," the psychologist can finally assess the extent to which the coefficient of determination is an accurate portrayal of the covariability of the two properties and not simply the result of the normal variability of sampling. To this end, the *F-statistic* representing the ratio of

   explained variance/unexplained variance =
   regression mean square/residual mean square =
   109.507.

   This is found in the intersection of the row labeled "Regression" and the column labeled "F." When this *F-statistic* value is compared with the *F-Distribution* Probability Model representing the scenario in which a sample has been drawn from a population with an "F-ratio" of 1, the probability of observing such an *F-statistic* value is found to be less than 0.0005. This assessment is found in the intersection of the row labeled "Regression" and the column labeled "Sig." On this basis, the psychologist concludes that it is *improbable* that this sample of students' Internet use habits and Internet use could have been drawn from such a population; thus, the resulting coefficient of determination is likely *not the result of normal sampling variability*. Instead, the psychologist concludes that it is *more likely* that the coefficient of determination *is a valid reflection of the covariability* of the two properties Internet use and grade point average for this sample of students.

14.9 Exercises

1) A real estate analyst is studying the relationship between home sale prices and living space (square footage) in a local housing market. The reason for this study is strategic; when offering a home for sale, it is important to choose a starting sales price that is close to the "market" price for homes with similar characteristics. Why? If a home is first offered at a selling price that is too far below the "market" price, the seller will not be able to realize the full value of selling his or her home. If a home is first offered at a selling price that is too far above the "market" price, the home will not be likely to attract any buyers, and the seller will be forced to make subsequent price reductions. In turn, these price reductions will be interpreted by buyers as desperation on the part of the seller, and buyers will act on this desperation by waiting for further price reductions. Again, this will leave the seller less likely to realize the full value of selling his or her home. It should be noted that this same strategic question is the basis for the financial analysis of "initial public offerings" (or IPOs) of corporate stocks.

The model underlying this study is based on practical real estate experience; the two most important characteristics of a home are its location and its living space, and because all the homes of interest are in the same geographic area, the remaining characteristic of importance is size. Presumably, a larger living space will command a higher price, so the "research question" is not whether or not these two quantitative properties of a home sale are related but more specifically the exact quantitative nature of the relationship. That is, given a home's size, what is a reasonable selling price to expect? To address this question, the analyst has conceptually constructed the linear model

$$\text{sales price} = (b \bullet \text{size}) + c,$$

where *b* represents the "differential" pricing applied to the square footage of the home being offered for sale.

Based on this conceptual model, and motivated by the desire to have a strategic advantage in constructing home selling offers for her clients, the analyst has collected the sales history of 25 homes sold in the town over a 3-month period. The sales prices, in thousands of dollars, and living spaces, in square footages, are provided in Table 14.3.

From these observations, the analyst can construct the appropriate statistical model for this particular market.

TABLE 14.3 ■ Home Sale Prices ($000) and Square Footage in a Nearby Town

$442; 2,975 sq. ft.	$170; 1,169 sq. ft.	$370; 1,680 sq. ft.	$263; 1,500 sq. ft.	$265; 1,262 sq. ft.
$275; 1,778 sq. ft.	$162; 1,600 sq. ft.	$134; 1,342 sq. ft.	$260; 1,300 sq. ft.	$746; 3,032 sq. ft.
$349; 1,826 sq. ft.	$118; 800 sq. ft.	$83; 1,000 sq. ft.	$427; 2,185 sq. ft.	$425; 2,450 sq. ft.
$285; 2,022 sq. ft.	$250; 1,050 sq. ft.	$443; 1,973 sq. ft.	$175; 1,110 sq. ft.	$400; 2,438 sq. ft.
$397; 1,493 sq. ft.	$306; 1,344 sq. ft.	$286; 1,211 sq. ft.	$270; 1,800 sq. ft.	$108; 386 sq. ft.

To assist the analyst, use SPSS to construct a linear model of the association between the sales price of a home and its size. Because the square footage of a home is not dependent on its sales price, you may assume that the "square footage" of a home is the independent property and the "sales price" is the dependent property.

a) Construct the data set with each home sale represented by the properties "Size" and "Price."

b) Construct a scatter plot of the observations to assess the applicability of using the covariance model.

c) Assuming that the covariance model is applicable, assess the linear regression model implied by the observations:

- What is the "slope" of the linear model? What does it imply with regard to the average difference in selling price based on the difference in home size?
- Is the "slope" statistically significant? What does this mean with regard to the preceding question?
- What is the "constant" of the model? What does it mean as a selling price?
- Is the "constant" statistically significant? What does this mean with regard to the preceding question?
- Use the model to construct a selling price for a home with 1,800 square feet of space.
- To what extent does the square footage of a home determine its selling price? That is, what proportion of the variation in selling prices is explained by the differences in square footage? In other words, what is the "fit" of the model? Is the "fitness" found for this model statistically significant?

2) Construct your own research investigation regarding a potential association between two quantitative properties of some set of phenomena of interest.

a) Describe the phenomena, the properties, and the motivation for the study.

b) Describe your speculative model of the association.

c) Obtain at least 25 observations of relevant phenomena.

d) Use SPSS to analyze your observations.

e) Report your analytical findings with regard to the speculative model you constructed.