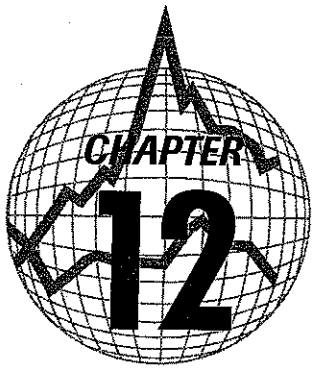


Simple Regression

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Use simple regression to test the statistical significance of a bivariate relationship involving one dependent and one independent variable
- Use Pearson's correlation coefficient as a measure of association between two continuous variables
- Interpret statistics associated with regression analysis
- Write up the model of simple regression
- Assess assumptions of simple regression

This chapter completes our discussion of statistical techniques for studying relationships between two variables by focusing on those that are *continuous.* Several approaches are examined: simple regression; the Pearson's correlation coefficient; and a nonparametric alterative, Spearman's rank correlation coefficient.

Although all three techniques can be used, we focus particularly on simple regression. Regression allows us to predict outcomes based on knowledge of an independent variable. It is also the foundation of time

series analysis, which is useful for budgeting and planning, and it is the essential foundation for studying relationships among three or more variables. Such relationships are examined in subsequent chapters and include control variables, which were introduced in Chapter 9. We begin with simple regression.

SIMPLE REGRESSION

Simple regression is used to analyze the relationship between two continuous variables. For example, we might study the relationship between productivity and job satisfaction when both variables are measured on a continuous scale. Continuous variables assume that the distances between ordered categories are determinable.[1] In simple regression, one variable is defined as the dependent variable and the other as the independent variable.

Scatterplot

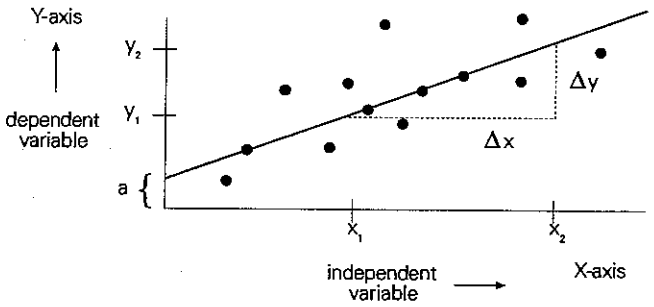
The relationship between two continuous variables can be portrayed in a ***scatterplot.*** A scatterplot is merely a plot of the data points for two continuous variables, as shown in Figure 12.1 (without the straight line). By convention, the dependent variable is shown on the vertical (or Y-) axis, and the independent variable on the horizontal (or X-) axis. The relationship between the two variables is estimated as a straight line relationship. The line is defined by the equation $y = a + bX$, where a is the intercept (or constant), and b is the slope. The slope, b, is defined as $\Delta y/\Delta x$, or $(y_2 - y_1)/(x_2 - x_1)$. The line is mathematically calculated such that the sum of distances from each observation to the line is minimized.[2] By definition, the slope indicates the change in y as a result of a unit change in x. The straight line is also called the ***regression line,*** and the slope (b) is called the ***regression coefficient.***

A positive regression coefficient indicates a positive relationship between the variables, shown by the upward slope in Figure 12.1. A negative regression coefficient indicates a negative relationship between the variables and is indicated by a downward-sloping line.

Test of Significance

The ***test of significance of the regression coefficient*** is a key test of hypothesis regression analysis that tells us whether the slope (b) is statistically different from zero. The slope is calculated from a sample, and we wish to know whether it is significant. When the regression line is horizontal ($b = 0$), no relationship exists between the two variables. Then, changes in the

Figure 12.1 — Scatterplot



independent variable have no effect on the dependent variable. The following hypotheses are thus stated:

H_0: $b = 0$, or the two variables are unrelated.
H_A: $b \neq 0$, or the two variables are (positively or negatively) related.

To determine whether the slope equals zero, a t-test is performed. The test statistic is defined as the slope, b, divided by the standard error of the slope, se(b). The standard error of the slope is a measure of the distribution of the observations around the regression slope, which is based on the standard deviation of those observations to the regression line:

$$\frac{b}{se(b)}.$$

Thus, a regression line with a small slope is more likely to be statistically significant when observations lie closely around it (that is, the standard error of the observations around the line is also small, resulting in a larger test statistic). By contrast, the same regression line might be statistically insignificant when observations are scattered widely around it. Observations that lie farther from the regression line will have larger standard deviations, and hence larger standard errors. *The computer calculates the slope, intercept, standard error of the slope, and the level at which the slope is statistically significant.*

Consider the following example. A management analyst with the Department of Defense wishes to evaluate the impact of teamwork on the productivity of naval shipyard repair facilities. Although all shipyards are required to use teamwork management strategies, these strategies are assumed to vary in practice. Coincidentally, a recently implemented

Table 12.1 — Simple Regression Output

Model Fit

R	R square	SEE
0.272	0.074	0.825

Dependent variable: Productivity

Coefficients

Constant	4.026	0.213	18.894	0.000
Teamwork	0.223	0.044	5.053	0.000

Note: SEE = standard error of the estimate; SE = standard error; Sig. = significance.

employee survey asked about the perceived use and effectiveness of teamwork. These items have been aggregated into a single index variable that measures teamwork. Employees were also asked questions about perceived performance, as measured by productivity, customer orientation, planning and scheduling, and employee motivation. These items were combined into an index measure of work productivity. Both index measures are continuous variables. The analyst wants to know whether a relationship exists between perceived productivity and teamwork. Table 12.1 shows the computer output obtained from a simple regression. The slope, b, is 0.223; the slope coefficient of teamwork is positive; and the slope is significant at the 1 percent level. Thus, perceptions of teamwork are positively associated with productivity. The t-test statistic, 5.053, is calculated as 0.223/0.044 (rounding errors explain the difference from the printed value of t). Other statistics shown in Table 12.1 are discussed below. The appropriate notation for this relationship is shown below. Either the t-test statistic or the standard error should be shown in parentheses, directly below the regression coefficient; analysts should state which statistic is shown. Here, we show the t-test statistic:[3]

$$\text{PRODUCTIVITY} = 4.026 + 0.223^{**}\text{TEAMWORK}$$
$$(5.05)$$
$$^{**}\, p < .01;\ ^{*}\, p < .05$$

The level of significance of the regression coefficient is indicated with asterisks, which conforms to the p-value legend that should also be shown. Typically, two asterisks are used to indicate a 1 percent level of significance, one asterisk for a 5 percent level of significance, and no asterisk for coefficients that are insignificant.[4]

Table 12.1 also shows R-square (R^2), which is called the ***coefficient of determination.*** R-square is of great interest: its value is interpreted as *the percentage of variation in the dependent variable that is explained by the independent variable.* R-square varies from zero to one, and is called a goodness-of-fit measure.[5] In our example, teamwork explains only 7.4 percent of the variation in productivity. Although teamwork is significantly associated with productivity, it is quite likely that other factors also affect it. It is conceivable that other factors might be more strongly associated with productivity and that, when controlled for other factors, teamwork is no longer significant. Typically, values of R^2 below 0.20 are considered to indicate weak relationships, those between 0.20 and 0.40 indicate moderate relationships, and those above 0.40 indicate strong relationships. Values of R^2 above 0.65 are considered to indicate very strong relationships. R is called the *multiple correlation coefficient* and is always $0 \leq R \leq 1$.

To summarize up to this point, simple regression provides three critically important pieces of information about bivariate relationships involving two continuous variables: (1) the level of significance at which two variables are associated, if at all (*t-statistic*), (2) whether the relationship between the two variables is positive or negative (b), and (3) the strength of the relationship (R^2).

The primary purpose of regression analysis is hypothesis testing, not prediction. In our example, the regression model is used to test the hypothesis that teamwork is related to productivity. However, if the analyst wants to predict the variable "productivity," the regression output also shows the SEE, or the ***standard error of the estimate*** (see Table 12.1). This is a measure of the spread of y values around the regression line as calculated *for the mean value of the independent variable, only, and assuming a large sample.* The standard error of the estimate has an interpretation in terms of the normal curve, that is, 68 percent of y values lie within one standard error from the calculated value of y, as calculated for the *mean* value of x using the preceding regression model. Thus, if the *mean* index value of the variable "teamwork" is 5.0, then the calculated (or predicted) value of "productivity" is [4.026 + 0.223*5 =] 5.141. Because SEE = 0.825, it follows that 68 percent of productivity values will lie ±0.825 from 5.141 when "teamwork" = 5. Predictions of y for other values of x have larger standard errors.[6]

Assumptions and Notation

Simple regression assumes that the relationship between two variables is *linear.* The linearity of bivariate relationships is easily determined through visual inspection, as shown in Figure 12.2. In fact, all analysis of relationships involving continuous variables should begin with a scatterplot. When variable relationships are nonlinear (parabolic or otherwise heavily curved), it is not appropriate to use linear regression. Then, one or both variables must be transformed, as discussed in Chapter 11.

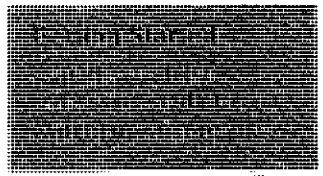
Simple regression also assumes that the *linear relationship is constant* over the range of observations. This assumption is violated when the relationship is "broken," for example, by having an upward slope for the first half of independent variable values and a downward slope over the remaining values. Then, analysts should consider using two regression models each for these different, linear relationships. The linearity assumption is also violated when no relationship is present in part of the independent variable values. This is particularly problematic because regression analysis will calculate a regression slope based on all observations. In this case, analysts may be misled into believing that the linear pattern holds for all observations. Hence, regression results always should be verified through visual inspection.

Linear regression also assumes that the variables are continuous. In Chapter 13, we will see that regression can also be used for nominal and dichotomous independent variables. The dependent variable, however, must be continuous. When the dependent variable is dichotomous, logistic regression should be used (Chapter 14).

Finally, the following notations are commouly used in regression analysis. The predicted value of y (defined, based on the regression model, as $y = a + bX$) is typically different from the ***observed value of*** y. The ***predicted value of the dependent variable*** y is sometimes indicated as $\hat{y}$ (pronounced "y-hat"). Only when $R^2 = 1$ are the observed and predicted values identical for each observation. The difference between y and $\hat{y}$ is called the regression *error* or ***error term*** (e). Hence the expressions

$$\hat{y} = a + b^*x \text{ and}$$
$$y = a + b^*x + e$$

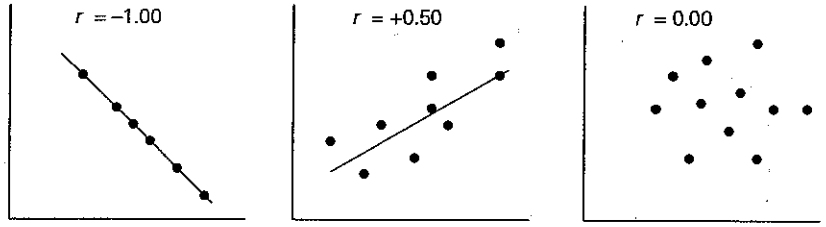
are equivalent, as is $y = \hat{y} + e$. Certain assumptions about e are important, such as that it is normally distributed. When error term assumptions are violated, incorrect conclusions may be made about the statistical significance of relationships. This important issue is discussed in greater detail in Chapter 13 and, for time series data, in Chapter 15. Hence, the above is a pertinent, but incomplete list of assumptions.



PEARSON'S CORRELATION COEFFICIENT

Pearson's correlation coefficient, r, measures the association (significance, direction, and strength) between two continuous variables; it is a measure of

Figure 12.2 ——〰—— Three Examples of *r*



association for two continuous variables. Also called the Pearson's product-moment correlation coefficient, it does not assume a causal relationship, as does simple regression. The correlation coefficient, *r*, indicates the extent to which the observations lie closely or loosely clustered around the regression line. The coefficient *r* ranges from –1 to +1. The sign indicates the direction of the relationship, which, in simple regression, is always the same as the slope coefficient. A "–1" indicates a perfect negative relationship, that is, that all observations lie exactly on a downward-sloping regression line; a "+1" indicates a perfect positive relationship, whereby all observations lie exactly on an upward-sloping regression line. Of course, such values are rarely obtained in practice because observations seldom lie exactly on a line. An *r* value of zero indicates that observations are so widely scattered that it is impossible to draw any well-fitting line. Figure 12.2 illustrates some values of *r*.

It is important to avoid confusion between Pearson's correlation coefficient and the coefficient of determination. For the two-variable, simple regression model, $r^2 = R^2$, but whereas $0 \leq R \leq 1$, *r* ranges from –1 to +1. Hence, the sign of *r* tells us whether a relationship is positive or negative, but the sign of *R*, in regression output tables such as Table 12.1, is always positive and cannot inform us about the direction of the relationship. In simple regression, only the regression slope, *b*, informs us about the direction of the relationship. Statistical software programs usually show *r* rather than r^2. Note also that the Pearson's correlation coefficient can be used only to assess the association between two continuous variables, whereas regression can be extended to deal with more than two variables, as discussed in Chapter 13. Pearson's correlation coefficient assumes that both variables are normally distributed.

When Pearson's correlation coefficients are calculated, a standard error of *r* can be determined, which then allows us to test the statistical significance of the bivariate correlation. For bivariate relationships, this is the same level of significance as shown for the slope of the regression coefficient. For the variables given earlier in this chapter, the value of *r* is .272 and the statistical significance of *r* is $p \leq .01$. Use of the Pearson's correlation coefficient assumes that the variables are normally distributed and that there are no significant departures from linearity.[7]

Comparing the measures *r* and *b* (the slope) sometimes causes confusion. The key point is that *r* does not indicate the regression slope but rather the extent to which observations lie close to it. A steep regression line (large *b*) can have observations scattered loosely or closely around it, as can a shallow (more horizontal) regression line. The purposes of these two statistics are very different.[8]

SPEARMAN'S RANK CORRELATION COEFFICIENT

The nonparametric alternative, ***Spearman's rank correlation coefficient*** (ρ, or "rho"), looks at correlation among the ranks of the data rather than among the values. The ranks of data are determined as shown in Table 12.2 (adapted from Table 10.8):

Table 12.2 ——〰—— Ranks of Two Variables

Observation	Variable 1 Value	Variable 1 Rank	Variable 2 Value	Variable 2 Rank
1	2.5	2	3.4	3
2	2.9	3	3.3	2
3	4.0	5	4.0	5
4	3.2	4	3.9	4
5	1.2	1	2.1	1

Because Spearman's rank correlation coefficient examines correlation among the ranks of variables, it can also be used with ordinal-level data.[9] For the data in Table 12.2, Spearman's rank correlation coefficient is .900 ($p = .035$).[10] Spearman's p-squared coefficient has a "percent variation explained" interpretation, similar to the measures described earlier. Hence, 90 percent of the variation in one variable can be explained by the other. For the variables given earlier, the Spearman's rank correlation coefficient is .274 ($p < .01$), which is comparable to *r* reported in preceding sections.

Box 12.1 illustrates another use of the statistics described in this chapter, in a study of the relationship between crime and poverty.

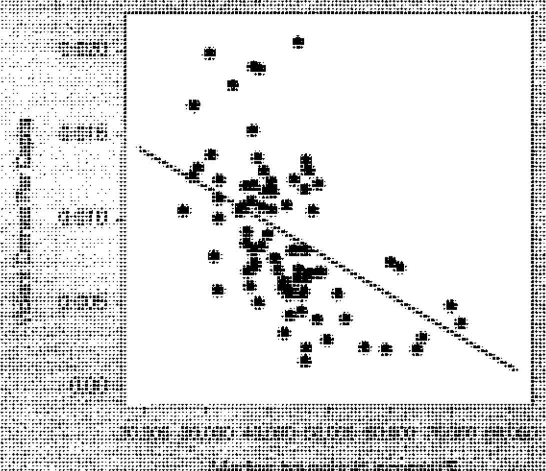
SUMMARY

When analysts examine relationships between two continuous variables, they can use simple regression or the Pearson's correlation coefficient. Both

In Greater Depth...

Box 12.1 Crime and Poverty

[illegible]



(continued)

Box 12.1 (continued)

[illegible]

measures show (1) the statistical significance of the relationship, (2) the direction of the relationship (that is, whether it is positive or negative), and (3) the strength of the relationship.

Simple regression assumes a causal and linear relationship between the continuous variables. The statistical significance and direction of the slope coefficient is used to assess the statistical significance and direction of the relationship. The coefficient of determination, R^2, is used to assess the strength of relationships; R^2 is interpreted as the percent variation explained. Regression is a foundation for studying relationships involving

three or more variables, such as control variables. The Pearson's correlation coefficient does not assume causality between two continuous variables.

A nonparametric alternative to testing the relationship between two continuous variables is Spearman's rank correlation coefficient, which examines correlation among the ranks of the data rather than among the values themselves. As such, this measure can also be used to study relationships in which one or both variables are ordinal.

KEY TERMS

Coefficient of determination, R^2 (p. 202)
Error term (p. 203)
Observed value of y (p. 203)
Pearson's correlation coefficient (p. 203)
Predicted value of the dependent variable y, $\hat{y}$ (p. 203)
Regression coefficient (p. 199)
Regression line (p. 199)
Scatterplot (p. 199)
Spearman's rank correlation coefficient (p. 205)
Standard error of the estimate (p. 202)
Test of significance of the regression coefficient (p. 199)

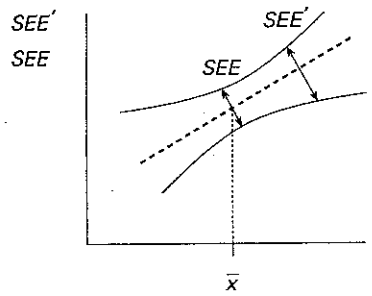
Notes

1. See Chapter 3 for a definition of continuous variables. Although the distinction between ordinal and continuous is theoretical (namely, whether or not the distance between categories can be measured), in practice ordinal-level variables with seven or more categories (including Likert variables) are sometimes analyzed using statistics appropriate for interval-level variables. This practice has many critics because it violates an assumption of regression (interval data), but it is often done because it doesn't (much) affect the robustness of results.
2. The method of calculating the regression coefficient (the slope) is called *ordinary least squares*, or OLS. This method estimates the slope by minimizing the sum of squared differences between each predicted value of $a + bX$ and the actual value of y. One reason for squaring these distances is to ensure that all distances are positive.
3. No consistent preference exists about what is shown in parentheses. The current practice in many political science journals is to report the standard error, but many public administrations report the t-test.
4. Some authors also identify other levels of significance, such as $p < .001$ or $p < .10$, but this does not affect study conclusions, of course.

5. The formula for R^2 is presented in Chapter 13, in our discussion of the F-test.
6. For predictions not based on the mean of x, the standard error of y is larger than the SEE, according to the following formula:

$$SEE' = SEE\sqrt{1+\frac{1}{N}+\frac{(x_i-\bar{x})^2}{(N-1)s_x^2}},$$

where s_x^2 is the variance of x, that is, $\Sigma(x-\bar{x})^2/(N-1)$. As can be seen, $SEE' = SEE$ only when N is large and the predicted values of y are calculated for the mean value of x (that is, $x_i = \bar{x}$). Graphically, the relationship between SEE′ and x is as follows:



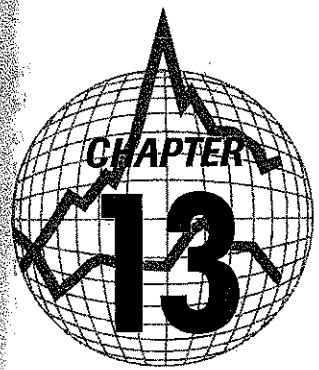
7. Based on visual inspection, these two variables are normally distributed. In addition, the Kolmogorov-Smirnov test (see Chapter 11) for the variable "teamwork" shows p = .084.
8. Pearson's correlation coefficient is also the basis for calculating Cronbach alpha, the measure of internal reliability discussed in Chapter 3. The formula for alpha is $\alpha = N^*\bar{r}/[1 + (N-1)^*\bar{r})]$, where N = the number of variables and $\bar{r}$ is the mean of the correlations among all of the different pairs of variables that make up the measure. This formula clearly shows that alpha is bounded by zero and one: when $\bar{r} = 1$, then $\alpha = 1$, and when $\bar{r} = 0$, then $\alpha = 0$.
9. Spearman's rank correlation coefficient would also be used when assumptions of normality are violated, or when variables are related in nonlinear ways.
10. The formula for Spearman's rank correlation coefficient is as follows:

$$r_s = 1-\frac{6\sum d^2}{n(n^2-1)}$$

where d is the difference between ranks in each observation (x,y). For the data shown in Table 12.2, consider the following calculation:

Observation	Variable 1 Rank	Variable 2 Rank	d	d^2
1	2	3	−1	1
2	3	2	1	1
3	5	5	0	0
4	4	4	0	0
5	1	1	0	0

Hence, the value of $r_s = 1 - [(6*2)/5(25-1)] = 0.9$.



Multiple Regression

CHAPTER OBJECTIVES

After reading this chapter, you should be able to

- Understand multiple regression as a full model specification technique
- Interpret standardized and unstandardized regression coefficients of multiple regression
- Know how to use nominal variables in regression as dummy variables
- Explain the importance of the error term plot
- Identify assumptions of regression, and know how to test and correct assumption violations

Multiple regression is one of the most widely used multivariate statistical techniques for analyzing three or more variables. This chapter uses multiple regression to examine such relationships, and thereby extends the discussion in Chapter 12. The popularity of multiple regression is due largely to the ease with which it takes *control variables* (or rival hypotheses) into account. In Chapter 9, we discussed briefly how contingency tables can be used for this purpose, but doing so is often a cumbersome and sometimes inconclusive effort. By contrast, multiple regression easily incorporates