13.4: Prediction in Regression

Prediction in regression is the idea that if two things are related, one can be used to estimate or approximate the other but this does not mean one causes the other or that we are mathematically seeing into the future. The idea is much simpler than that. When we use prediction in regression we are starting by establishing a pattern in data and then saying, essentially, "If a pattern is generally true based on known data points, it can be useful in estimating new or previously unknown data points." For example, suppose that in data collected from 100 people, there is a very strong, positive relationship between hours spent studying and exam scores. This would mean that, among those participants, exam scores tended to increase as hours of studying increased. This would also mean that as time spent studying decreased from person to person, exam scores also tended to decrease. Based on this pattern, if someone new reported they did not study at all we would predict that they would have a relatively low score on the exam. In keeping, if someone new reported they studied for many hours we would predict that they would have a relatively high score on the exam.

The Meaning of "Prediction"

It is important to clarify what prediction does and does not mean when this term is used with regression. When regression is used, **prediction** refers to the estimated, expected value of one variable (Y) based on the known value of another variable (X). Note that in this definition the prediction is "estimated" and "expected" rather than "precise" and "known." The terms prediction and estimation may be used interchangeably for regression.

Regression Lines

Correlations and regressions summarize relationships between two quantitative variables using a line with a consistent slope. When this is done with correlation, the summary line is often called a fit line but when regression is used the summary line is called a regression line. A **regression line** represents the best approximation of the linear relationship between two variables using a fit method, the most common of which is the least-squares regression method. A *fit method* refers to the way the line was created to best fit the data. There are different methods that can be used for creating the regression line. One is the least squares model. This fit method creates the regression line by angling it such that it minimizes residuals based on the data provided. Therefore, regressions using this are often called least-squares regression models. However, this method is so common and useful that it is often assumed when the form of regression is not specified. This is the version of regression which will be used throughout this chapter.

Least square regression refers to when the best-fitting line for the data is calculated such that the sum of squared deviations from Y (i.e. the vertical deviations) is minimized. To state this another way, this procedure is used to find the exact angle for the line that gets as close to the data on the graph (i.e. the dots on a scatterplot) as possible, on average. This should result in as many data points falling above the line as below the line. The line will only be able to actually pass through every data point when the correlation is perfect, which is quite rare. Instead, therefore, the line summarizes the pattern among the data such that the distance between the dots and the line is minimized when a correlation exists but is not perfect.

Consider what we already know about the fit lines for correlation in Chapter 12 (which are the same as the regression lines in this chapter). The stronger a relationship is between two variables, the closer the data points will tend to be to the line. Conversely, the weaker the relationship is between the two variables, the farther the dots will tend to be from the line. When the relationship is stronger, the line is a better fit to the data and when it is weaker, the line is a poorer fit to the data.

The Basics of β_1 (slope of a Regression Line)

Predictions in regression use the slope of the regression line plus something known as the y-intercept. For this reason, understanding the slope is key to understanding and using regression. The slope of the regression line summarizes the pattern between the X-variable and Y-variable. The symbol for the slope of a regression line is β_1. This symbol is the Greek letter beta. Thus, this symbol with the subscript of 1 is known as "beta one." We are only testing one regression and one slope at a time in this chapter so we will only be considering beta one. However, more complex techniques beyond the scope of this book may test multiple slopes at once. Additional slopes (i.e. betas), when applicable, are numbered consecutively starting from 2.

The slope of a line (β_1) is computed as change in the Y-variable divided by corresponding change in the X-variable. The symbol for change is Δ. Slope is often summarizes as "rise over run" meaning vertical change (i.e. y-axis change) divided by horizontal change (i.e. x-axis change). In mathematics the slope is, thus, often written as a division problem. Statisticians use the quotient of

that division problem for the slope known as β_1. Thus, slope can be presented as a fraction or as its quotient but is usually reported in quotient form in statistics.

$$\beta_1 = \frac{\Delta Y}{\Delta X}$$

When β_1 is used (in quotient form), it indicates the change in the Y-variable predicted for each one unit increase in the X-variable. Thus, if β_1 = 0.30 it translates to saying that for every one unit increase in the X-variable, there is a .30 unit increase predicted in the Y-variable. However, if β_1 = -0.30 it translates to saying that for every one unit increase in the X-variable, there is a 0.30 unit decrease predicted in the Y-variable.

When a hypothesis is tested, the generic terms of X-variable and Y-variable are replaced with the actual variable names of variables being tested when interpreting a slope. Suppose that a hypothesis is tested which states that "Hours spent studying will predict exam scores" and that when the data are tested, it is found that β_1 = 7.00. This slope would translate to saying that, "For every one hour increase in studying, there is a 7.00 unit increase predicted for exam scores." Another way to word this is to say that, "For every additional hour spent studying, a 7.00 point increase in exam scores is predicted."
