5 Simple Regression

Estimating Models with a Single Continuous Predictor

We have used the simple single-parameter model to illustrate the use of models, the notion of error, and inference procedures to be used in comparing augmented and compact models. We have focused on this single-parameter model in so much detail because the estimation and inference procedures that we developed within this very simple context generalize to much more complicated models. That is, regardless of the complexity of a model, estimation from here on will be done by minimizing the sum of squared errors, just as we did in the single-parameter case, and inference will be done by comparing augmented and compact models using PRE and *F*. So the detail on single-parameter models has been necessitated by our desire to present in a simple context all of the statistical tools that we will use in much more complex situations.

However, as we noted, single-parameter models are only infrequently of substantive or theoretical interest. In many ways, the example from the last chapter, where we wanted to test the hypothesis that people were willing to pay the expected value of $50, is unusual in the behavioral sciences. More frequently such a priori values do not exist, and instead we may be asking whether the mean in one group of respondents (e.g., those who were trained in the meaning of expected values) differs from the mean in another group of respondents (e.g., those who received no such training). Or, returning to the data on internet access, while it is certainly possible that we would be interested in testing whether some a priori percentage (e.g., 75%) is a good estimate of mean internet access, it is much more likely that we would be interested in examining the determinants or correlates of internet access rates. In other words, our interest is more likely to center on attempts to explain the internet access data than on tests of alternative values for the mean access rate.

To examine these types of substantive issues, we need to consider models having more than a single parameter. Initially, we will consider only two-parameter models, taking the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The exact definition of the terms in this two-parameter model will be detailed below. For the present we simply note that we are making predictions of Y_i conditional upon some other variable X_i, since the model's predictions from this two-parameter model change as X_i changes, assuming that β_1 takes on some value other than zero.

Actually, there are two variations on this two-parameter model, each of which is illustrated by one of the two examples we have just discussed. In the first example, concerning whether training in the meaning of expected values influences how much one is willing to pay for a lottery ticket, we want to examine whether those who receive such training are willing to pay more or less than those who do not. This amounts to asking whether we need different predictions from the model for the two groups of respondents or whether a single prediction suffices regardless of whether training was received or not. In other words, we want to compare a model in which predictions are made conditional on knowing whether or not a given respondent received the training with one where the identical prediction is made for every respondent. For this comparison, the augmented model is a two-parameter model, defining X_i in such a way that it identifies whether or not the respondent received training. We might, for instance, define X_i as follows:

$X_i = -1$, if a respondent did not receive training

$X_i = +1$, if a respondent did receive training

If the estimated value of β_1 in the above two-parameter model is something other than zero, the model then gives different predicted values for participants in the two groups. For example, if b_0 (the estimated value of β_0) equals 46 and b_1 (the estimated value of β_1) equals 3, then the prediction for the respondents without training is:

$$\hat{Y} = b_0 + b_1X_i = 46 + 3(-1) = 46 - 3 = 43$$

and the prediction for the respondents who receive training is:

$$\hat{Y} = b_0 + b_1X_i = 46 + 3(1) = 46 + 3 = 49$$

Notice that there are only two possible values for X_i in this example, and hence only two predicted values. Respondents either receive training or not, and our decision about the numerical values used to represent this training is arbitrary. For instance, had we defined X_i differently, giving respondents with training a value of 2 on X_i and those without training a value of 4, the two-parameter model would still generate different predictions for the two groups of students, assuming the estimated value of β_1 does not equal zero.

Now consider the second example. Suppose we wanted to explain variation in the internet access data and we suspected that average educational level of residents in the US states, measured as the percentage of residents with a college degree, might be a reasonable candidate for an explanatory factor. In other words, we thought that internet access rates might be higher in states where more people had graduated from college. So we might use the two-parameter model to make predictions for states conditional on college graduation rates, defining this as X_i. This variable has many possible values, and it would be unlikely that any two states would have exactly the same graduation rate and, hence, the exact same values on X_i. Therefore, instead of making two different predictions as in the lottery example, our two-parameter model now is likely to make different predictions for each state, since each state is likely to have a unique value of X_i. Another difference between this two-parameter model and the lottery example is that here values on X_i are less arbitrary than they were in the lottery example. Each state has an actual college graduation rate that can be compared with other states and the information about such state-to-state differences needs to be represented in the values we assign to X_i for each state.

The difference between these two examples lies in the nature of the units of measurement of the predictor variable, the variable upon which the predictions are conditional. In the lottery example, no training versus training is a categorical variable, in the sense that all participants are in one group or the other. This means that while X_i needs to code the distinction between the two training conditions, it does not matter which group of respondents is given the higher value on X_i nor does it matter which two values are used. On the other hand, college graduation rate is what we call a continuous predictor variable, in the sense that different states have different values and the differences among these values are meaningful.

While all of the procedures for building models, estimating parameters, and comparing augmented and compact models can be used regardless of whether the predictor variable (or variables) is categorical or continuous, it is conceptually useful to treat models with categorical predictor variables separately from models whose predictors are assumed to be continuously measured. Initially, we will consider models that contain only predictors presumed to be continuously measured.

In the current chapter, we treat two-parameter models having a single, continuously measured predictor variable. Then, in Chapters 6 and 7, we consider models with multiple continuously measured predictors. In traditional statistical jargon these three chapters (Chapters 5, 6, and 7) deal with simple and multiple regression, including polynomial and moderated regression models. Then in Chapters 8 and 9 we turn our attention to models having categorical predictors. Again, in traditional statistical jargon, these chapters deal with analysis of variance. Finally, in Chapter 10 we consider models in which some predictors are categorical variables and some are continuous variables. Such models are traditionally referred to as analysis of covariance models. Our approach, however, to each type of model, regardless of the chapter, will be uniform. Rather than describing seemingly different statistical techniques for multiple regression, analysis of variance, and analysis of covariance, we will estimate parameters just as we have done in the simple single-parameter case, and we will test hypotheses by comparing augmented and compact models. So, while our treatment of categorial predictors is located in different chapters from our treatment of continuous predictors, the same procedures will be used throughout.

DEFINING A LINEAR TWO-PARAMETER MODEL

We now confine our attention to two-parameter models with a single continuous predictor variable. As an example, we will use the data contained in Figure 5.1 to ask whether differences between US states in their internet access rates are predictable from differences in their college graduation rates. As we speculated above, it seems reasonable that internet access may be higher in states where the population is relatively better educated.

Figure 5.2 is a scatterplot of the tabular data from Figure 5.1. The vertical axis represents internet access rates, and the horizontal axis represents college graduation rates. Each point in this plot represents one of the 50 US states. The question that we would like to ask is whether we can use graduation rates to predict internet access. Or, expressed differently, do our predictions of internet access improve by making those predictions conditional on knowledge of graduation rates?

We will use a simple linear model to generate such conditional predictions. As already discussed, this model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i is a state's internet access rate and X_i is its college graduation rate. Returning to our generic equation:

DATA = MODEL + ERROR

we see that the model in this two-parameter equation is represented by

$$\beta_0 + \beta_1 X_i$$

In terms of estimated parameter values, the predictions made by this model for each state's access rate are given by:

$$\hat{Y}_i = b_0 + b_1 X_i$$

Because ERROR equals DATA minus MODEL, the residuals in this two-parameter model can be expressed as follows:

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i \\ &= Y_i - (b_0 + b_1 X_i) \end{aligned}$$

Let us examine each of the parameter estimates in this model and see what each is telling us. First, consider b_0. In the single-parameter model, we saw that b_0 equaled the mean value of Y_i, assuming that we define error as the sum of squared errors. Another way of saying the same thing is that in the single-parameter model b_0 is our predicted value for each state. However, in this two-parameter model we wish to take further information into account in making each state's prediction. We are making each state's prediction conditional on its college graduation rate. Therefore, b_0 is not the predicted value for each state, because the predictions vary as a function of graduation rates:

$$\hat{Y}_i = b_0 + b_1 X_i$$

There is one case, however, when this model predicts an internet access rate equal to b_0. This is clearly when X_i equals zero, for then:

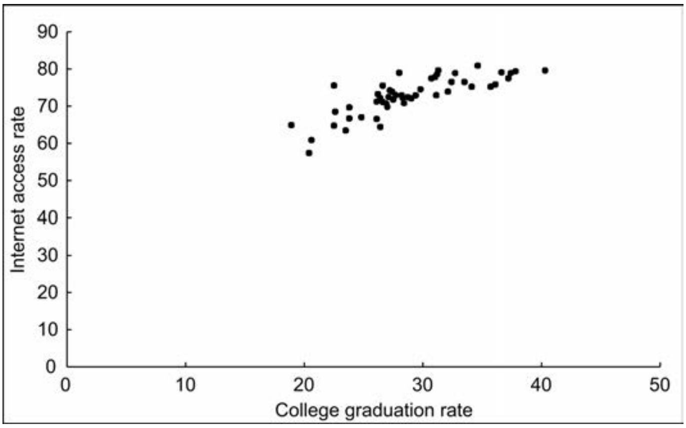
$$\hat{Y}_i = b_0 + b_1 (0) = b_0$$

This, then, provides the interpretation for the parameter estimate b_0 in this two-parameter model: b_0 is our prediction of Y_i when X_i equals zero. As we will see for our example, this prediction may not be very useful because the data from which we estimate this parameter may not include data points having values of X_i near zero.

FIGURE 5.1 Internet access rates and college graduation rates

US state	*Internet access rate*	*College graduation rate*
AK	79.0	28.0
AL	63.5	23.5
AR	60.9	20.6
AZ	73.9	27.4
CA	77.9	31.0
CO	79.4	37.8
CT	77.5	37.2
DE	74.5	29.8
FL	74.3	27.2
GA	72.2	28.3
HI	78.6	31.2
IA	72.2	26.4
ID	73.2	26.2
IL	74.0	32.1
IN	69.7	23.8
KS	73.0	31.1
KY	68.5	22.6
LA	64.8	22.5
MA	79.6	40.3
MD	78.9	37.4
ME	72.9	28.2
MI	70.7	26.9
MN	76.5	33.5
MO	69.8	27.0
MS	57.4	20.4
MT	72.1	29.0
NC	70.8	28.4
ND	72.5	27.1
NE	72.9	29.4
NH	80.9	34.6
NJ	79.1	36.6
NM	64.4	26.4
NV	75.6	22.5
NY	75.3	34.1
OH	71.2	26.1
OK	66.7	23.8
OR	77.5	30.7
PA	72.4	28.7
RI	76.5	32.4
SC	66.6	26.1
SD	71.1	26.6
TN	67.0	24.8
TX	71.8	27.5
UT	79.6	31.3
VA	75.8	36.1
VT	75.3	35.7
WA	78.9	32.7
WI	73.0	27.7
WV	64.9	18.9
WY	75.5	26.6

FIGURE 5.2 Scatterplot of internet access rates and college graduation rates for each of the 50 US states



The second parameter estimate in the model, b_1, tells us how our predictions change as X_i changes. Suppose we had two observations differing in their values on X_i by one unit, with X_i for the first observation being one unit larger than X_i for the second. According to the model, our predictions for the two data points would differ by b_1 since:

$$\begin{aligned}\hat{Y}_1 - \hat{Y}_2 &= (b_0 + b_1X_1) - (b_0 + b_1X_2)\\ &= b_1X_1 - b_1X_2\\ &= b_1(X_1 - X_2)\\ &= b_1\end{aligned}$$

So, b_1 tells us by how much our predictions of Y_i change as X_i increases by one unit. Notice that in this derivation we did not specify what the actual values of X_1 and X_2 were. We only specified that they were one unit apart from each other. Hence, this implies that b_1 in this two-parameter model is constant, regardless of the level of X_i. This is what was meant by the definition of this sort of two-parameter model as a linear model. As X_i changes by some set amount, our predictions of Y_i change by a constant amount, regardless of the value of X_i.

To review, b_0 and b_1 tell us very different things: b_0 is a predicted value (a $\hat{Y}_i$) at a particular value of X_i, namely when X_i equals zero; b_1 is not a predicted value, rather it is the difference between two predicted values as we move from a smaller X_i to one that is one unit larger.

Let us look at this two-parameter model graphically for the example in which internet access is predicted from college graduation rates. Figure 5.3 presents the graph of the model set against the data. All of the predictions $\hat{Y}_i$ lie on the line defined by the model. Errors of prediction, e_i, as in the single-parameter model, are defined as vertical distances between the line and an actual observation. That is, an error or residual is the difference between Y_i and $\hat{Y}_i$. b_0 is the value of $\hat{Y}_i$ when X_i equals zero; it is frequently called the intercept because it is the value on the vertical axis of the graph where the prediction function crosses or "intercepts" it. b_1 is the difference in $\hat{Y}_i$ for each unit increase in X_i. We can think of it as the slope of the line, since algebraically it is the difference in predicted values between any two points on the line per their difference in X_i values: *rise over run*:

$$b_1 = \frac{(\hat{Y}_2 - \hat{Y}_1)}{(X_2 - X_1)}$$

where the subscripts designate any two points on the line. Notice that the slope can take on any positive or negative value. If the slope is positive, it means that the model predicts higher values of Y_i as X_i increases. If the slope is negative, the model predicts lower values of Y_i as X_i increases.

ESTIMATING A LINEAR TWO-PARAMETER MODEL

Given some sample of data, how do we estimate the parameters of this sort of model? We want to use our sample of data to generate values of b_0 and b_1 in the equation:

$$\hat{Y}_i = b_0 + b_1X_i$$

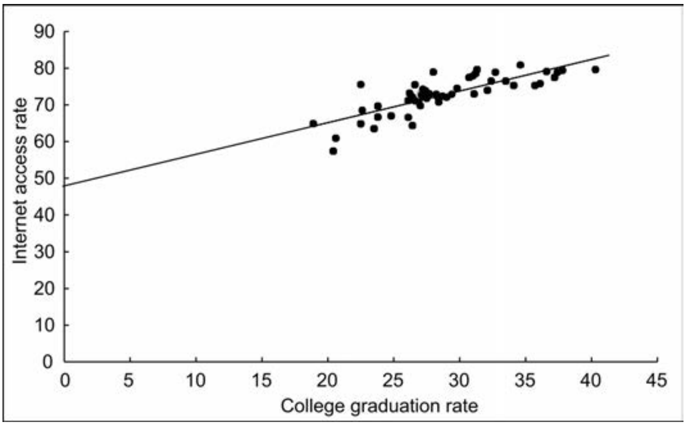
that are good estimates of the true (but unknown) parameters β_0 and β_1. To do this, we decided in Chapter 3 that for the single-parameter model we would derive estimates that minimize the sum of squared errors. This preference was due to the fact that if the errors are normally distributed, least-squares parameter estimates are unbiased, consistent, and relatively efficient. This continues to be the case in the context of the present two-parameter model and will continue to be the case in more complicated models with many parameters, which we consider in subsequent chapters. For now, we want to derive estimated values of β_0 and β_1 that minimize $\Sigma(Y_i - \hat{Y}_i)^2$. The resulting *least-squares* parameter estimates are given as:

$$b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

The derivation of these estimates is given in Box 5.1.

FIGURE 5.3 Scatterplot with two-parameter model predicting internet access from college graduation rates



Box 5.1 Algebraic Derivation of Least-Squares Estimates of β_0 and β_1

$$\text{SSE} = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - b_0 - b_1 X_i)^2$$

given that $\hat{Y}_i = b_0 + b_1 X_i$. We now add $(\bar{Y} - \bar{Y})$ and $b_1(\bar{X} - \bar{X})$ inside the parentheses to this expression for SSE. Since both of these expressions equal zero, we have not changed the equality. Thus:

$$\text{SSE} = \Sigma(Y_i - \bar{Y} + \bar{Y} - b_0 - b_1 X_i + b_1\bar{X} - b_1\bar{X})^2$$

Grouping terms yields the equivalent expression:

$$\text{SSE} = \Sigma[(Y_i - \bar{Y}) + (\bar{Y} - b_0 - b_1\bar{X}) - b_1(X_i - \bar{X})]^2$$

If we square the term in brackets and distribute the summation sign, this gives the equivalent expression:

$$\begin{aligned}\text{SSE} = {} & \Sigma(Y_i - \bar{Y})^2 + 2(\bar{Y} - b_0 - b_1\bar{X})\,\Sigma(Y_i - \bar{Y}) \\ & -2b_1\,\Sigma(Y_i - \bar{Y})(X_i - \bar{X}) + n\,(\bar{Y} - b_0 - b_1\bar{X})^2 \\ & -2b_1(\bar{Y} - b_0 - b_1\bar{X})\,\Sigma(X_i - \bar{X}) + b_1^2\,\Sigma(X_i - \bar{X})^2\end{aligned}$$

Since both $\Sigma(Y_i - \bar{Y})$ and $\Sigma(X_i - \bar{X})$ equal zero, this expression reduces to:

$$\begin{aligned}\text{SSE} = {} & \Sigma(Y_i - \bar{Y})^2 - 2b_1\,\Sigma(Y_i - \bar{Y})(X_i - \bar{X}) + n(\bar{Y} - b_0 - b_1\bar{X})^2 \\ & + b_1^2\,\Sigma(X_i - \bar{X})^2\end{aligned}$$

Since the third term in this expression, $n(\bar{Y} - b_0 - b_1\bar{X})^2$, is necessarily positive, to minimize SSE we would like to set it equal to zero. Therefore, we wish values of b_0 and b_1 such that:

$$n(\bar{Y} - b_0 - b_1\bar{X}) = 0$$

Dividing both sides of this equality by n gives us:

$$\bar{Y} - b_0 - b_1\bar{X} = 0$$

or, equivalently:

$$b_0 = \bar{Y} - b_1\bar{X}$$

We have now reduced our expression for SSE, assuming the desire to minimize it, to:

$$\begin{aligned}\text{SSE} &= \Sigma(Y_i - \bar{Y})^2 - 2b_1\,\Sigma(Y_i - \bar{Y})(X_i - \bar{X}) + b^2{}_1\,\Sigma(X_i - \bar{X})^2 \\ &= \Sigma(Y_i - \bar{Y})^2 + \Sigma(X_i - \bar{X})^2\left[b_1^2 - 2b_1\frac{\Sigma(Y_i - \bar{Y})(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}\right]\end{aligned}$$

Let us now add to and subtract from this expression the quantity:

$$\Sigma(X_i - \bar{X})^2\left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}\right)^2$$

Thus:

$$
\begin{aligned}
\text{SSE} &= \Sigma(Y_i - \bar{Y})^2 \\
&\quad + \Sigma(X_i - \bar{X})^2\left[b_1^2 - 2b_1\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} + \left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}\right)^2\right] \\
&\quad - \Sigma(X_i - \bar{X})^2\left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}\right)^2
\end{aligned}
$$

Rearranging terms and taking the square root of the term in brackets gives us:

$$
\begin{aligned}
\text{SSE} &= \Sigma(Y_i - \bar{Y})^2 - \Sigma(X_i - \bar{X})^2\left(\frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}\right)^2 \\
&\quad + \Sigma(X_i - \bar{X})^2\left[b_1 - \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}\right]^2
\end{aligned}
$$

The last term in this expression for SSE is necessarily positive. Therefore, to minimize SSE we want this last term to equal zero. This occurs if:

$$
b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}
$$

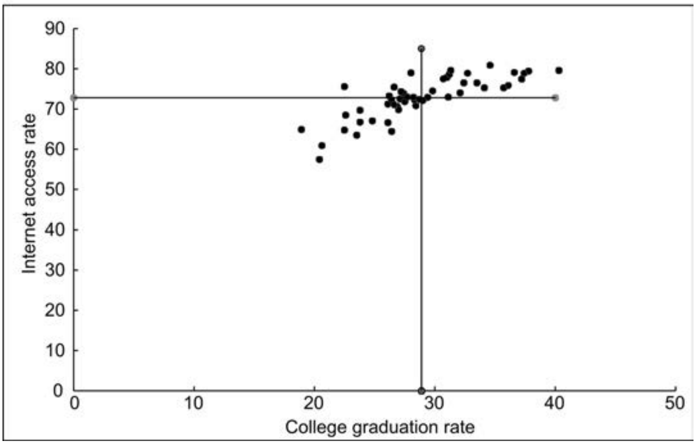
We can think about the formula for the estimated slope in a couple of different ways. One way is to divide both the numerator and denominator of the formula by $n - 1$:

$$
\begin{aligned}
b_1 &= \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/(n - 1)}{\Sigma(X_i - \bar{X})^2/(n - 1)} \\
&= \frac{s_{XY}}{s^2_X}
\end{aligned}
$$

In this last expression, the numerator of the slope, s_{XY}, is known as the covariance of X and Y, and the denominator is, of course, the variance of X.

Examining the crossproduct of $(X_i - \bar{X})(Y_i - \bar{Y})$ for any given observation helps to conceptualize the meaning of the covariance between two variables. To aid with this conceptualization, in Figure 5.4 we have added a horizontal line at the mean of Y and a vertical line at the mean of X to the scatterplot of the data. Any given observation will have a positive value for its crossproduct if it lies either in the upper right quadrant of the scatterplot or the lower left quadrant of the scatterplot, where the quadrants are defined by the intersecting lines at the two means. Positive values of the crossproduct thus come from observations that are either above the mean on both variables or below the mean on both. On the other hand, observations with negative values for their crossproducts will lie in the other two quadrants, having values that are below the mean on one variable but above the mean on the other. The covariance is (roughly) the average of all these

FIGURE 5.4 Scatterplot with horizontal line at $\bar{Y}$ and a vertical line at $\bar{X}$



individual crossproducts. Given that the denominator of the slope is always positive in value, the sign of the slope is determined by the sign of the covariance. This means that the slope will be positive if most of the observations in a scatterplot are either above the means on both variables or below the means on both variables. A negative slope happens when most of the observations fall into the other two quadrants: above the mean on one variable but below it on the other. A slope near zero would occur when the observations are randomly distributed throughout all four quadrants.

The other way to think conceptually about the meaning of the slope involves applying a bit of algebra to the original formula for the slope given above, yielding:

$$b_1 = \Sigma\, w_i\left[\frac{Y_i - \bar{Y}}{X_i - \bar{X}}\right], \text{ where } w_i = \frac{(X_i - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}$$

The term in the brackets in this equation can be thought of as the slope suggested by each individual observation—it is the rise over run of a line that goes between the joint mean of the two variables and a particular observation. In Figure 5.5 we have added a few of these individual "slopes" for a few observations. The w_i can be thought of as a weight assigned to each observation, where the weight represents the proportion of the total sum of squares of X that is attributable to the particular observation. In essence, we can think of each observation as having a slope that it "prefers" (between the joint mean and itself), that gets a certain weight or vote in determining the value of the slope for all the observations. The more extreme the observation is on X, the greater the vote.

Using the above formulas to calculate both parameter estimates, b_0 and b_1, or more efficiently using a regression routine in one of the various statistical software packages, we can calculate the estimated intercept and slope for the model where we regress internet access rates on college graduation rates. (Notice here that in regression terminology one regresses Y on X, not the other way around.) The resulting estimated two-parameter model for these data is:

$$\hat{Y}_i = 47.912 + 0.862X_i$$

This prediction function is graphed in Figure 5.6 as a straight line on the scatterplot we saw before.

FIGURE 5.5 Scatterplot with individual slopes

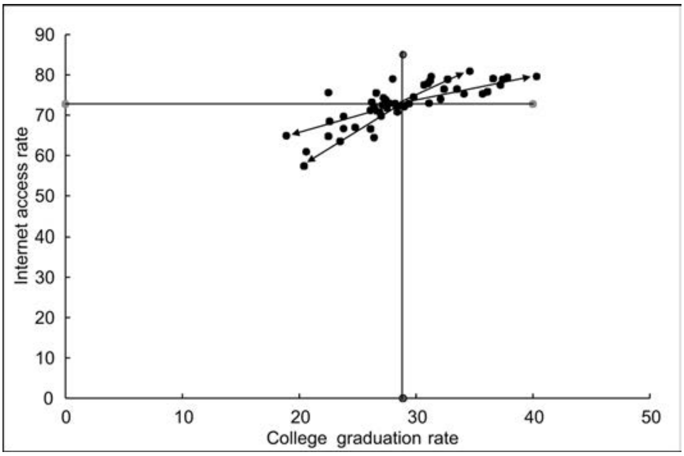
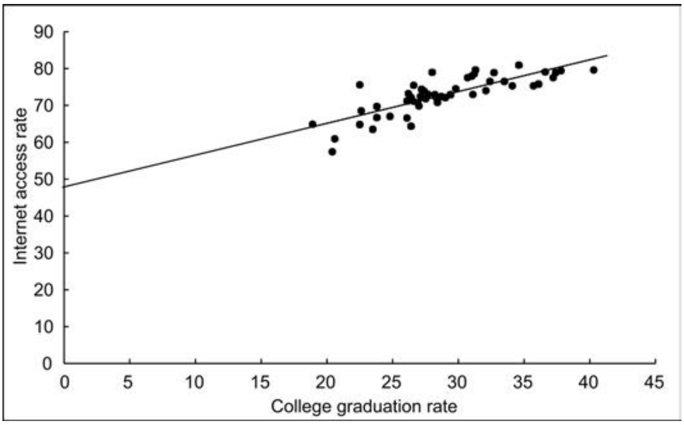


FIGURE 5.6 Scatterplot with two-parameter model: $\hat{Y}_i = 47.912 + 0.862X_i$



(This figure is identical to Figure 5.3 where we plotted the slope and intercept prior to indicating how they were estimated.)

Let us interpret each of the regression coefficients, b_0 and b_1, in this equation. The first equals 47.912. This is the value predicted by the model for a state's internet access rate if none of the population in the state had graduated from college. While this intercept is the best unbiased estimate of this prediction based on a linear model of these data, it is clearly a relatively meaningless value, because no state in the data had a college graduation rate anywhere near zero.

The value of the slope, 0.862, tells us that if we found two states differing in college graduation rates by 1%, our model predicts that the internet access rate would be .862% higher in the better educated state.

In Figure 5.7 we present for each state its Y_i, X_i, $\hat{Y}_i$, residual, and squared residual. The sum of these squared residuals, $\Sigma(Y_i - \hat{Y}_i)^2$, across all 50 states is also given. Having used the least-squares criterion guarantees that no other values of b_0 and b_1 would give us a smaller sum of squared residuals for these data.

FIGURE 5.7 Predicted values and residuals for internet access data by state

US state	Y_i	X_i	$\hat{Y}_i$	e_i	e_i^2
AL	63.5	23.5	68.169	–4.669	21.780
AK	79.0	28.0	72.048	6.952	48.330
AZ	73.9	27.4	71.531	2.369	5.612
AR	60.9	20.6	65.669	–4.769	22.743
CA	77.9	31.0	74.634	3.266	10.667
CO	79.4	37.8	80.496	–1.096	1.201
CT	77.5	37.2	79.978	–2.478	6.140
DE	74.5	29.8	73.600	0.900	0.810
FL	74.3	27.2	71.358	2.942	8.655
GA	72.2	28.3	72.307	–0.107	0.011
HI	78.6	31.2	74.806	3.794	14.394
ID	73.2	26.2	70.496	2.704	7.312
IL	74.0	32.1	75.582	–1.582	2.503
IN	69.7	23.8	68.428	1.272	1.618
IA	72.2	26.4	70.669	1.531	2.344
KS	73.0	31.1	74.720	–1.720	2.958
KY	68.5	22.6	67.393	1.107	1.225
LA	64.8	22.5	67.307	–2.507	6.285
ME	72.9	28.2	72.220	0.680	0.462
MD	78.9	37.4	80.151	–1.251	1.565
MA	79.6	40.3	82.651	–3.051	9.309
MI	70.7	26.9	71.100	–0.400	0.160
MN	76.5	33.5	76.789	–0.289	0.084
MS	57.4	20.4	65.497	–8.097	65.561
MO	69.8	27.0	71.186	–1.386	1.921
MT	72.1	29.0	72.910	–0.810	0.656
NE	72.9	29.4	73.255	–0.355	0.126
NV	75.6	22.5	67.307	8.293	68.774
NH	80.9	34.6	77.737	3.163	10.005
NJ	79.1	36.6	79.461	–0.361	0.130
NM	64.4	26.4	70.669	–6.269	39.300
NY	75.3	34.1	77.306	–2.006	4.024
NC	70.8	28.4	72.393	–1.593	2.538
ND	72.5	27.1	71.272	1.228	1.508
OH	71.2	26.1	70.410	0.790	0.624
OK	66.7	23.8	68.428	–1.728	2.986
OR	77.5	30.7	74.375	3.125	9.766
PA	72.4	28.7	72.651	–0.251	0.063
RI	76.5	32.4	75.841	0.659	0.434
SC	71.1	26.1	70.410	–3.810	14.516
SD	67.0	26.6	70.841	0.259	0.063
TN	71.8	24.8	69.290	–2.290	5.244
TX	79.6	27.5	71.617	0.183	0.033
UT	75.3	31.3	74.893	4.707	22.156
VT	75.8	35.7	78.685	–3.385	11.458
VA	78.9	36.1	79.030	–3.230	10.433
WA	64.9	32.7	76.099	2.801	7.846
WV	73.0	18.9	64.204	0.696	0.484
WI	75.5	27.7	71.789	1.211	1.467
WY	75.5	26.6	70.841	4.659	21.706
					SSE = 480.003

We can divide the sum of squared errors by the remaining degrees of freedom for error, $n - p$ (which in this case equals $n - 2$), to calculate the mean square error:

$$\text{MSE} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{480.003}{48} = 10.00$$

Just as b_0 and b_1 are unbiased estimates of β_0 and β_1 under the least-squares criterion, so also the mean square error is an unbiased estimate of the variance of ε_i. It estimates how variable the errors of prediction are at each level of X_i. As we will discuss, it is assumed that the variance of these errors is constant across all values of X_i. The square root of this mean square error is known as the *standard error of prediction*.

We will do one more example using two other variables from the states dataset. For this example, we are going to examine whether a state's population density (measured in 2010 as hundreds of people per square mile) can be used to predict the automobile fatality rate in the state (measured in 2010 as the number of fatalities per 100 million vehicle miles traveled). One certainly might expect more automobile accidents in states that are more densely populated, but it is less clear what one might expect in terms of fatalities from such accidents. On the one hand, if the accident rate is higher in more densely populated states, one might also predict a higher fatality rate. On the other hand, in more densely populated states, perhaps accidents are less likely to result in fatalities since more of the accidents are likely to be simply fender-benders rather than more serious high-speed collisions.

The parameter estimates from the regression model make clear how these variables are related:

$$\hat{Y}_i = 1.28 - 0.05X_i$$

Let us interpret both parameter estimates in this model. Doing so will make clear that their interpretation depends on the metric in which the two variables are measured. First, the intercept, 1.28, represents the predicted number of fatalities (per 100 million vehicle miles driven) if a state's population density were zero. Of course this number is not very informative, since no state has a population density that is zero. Yet, it is the best linear prediction from these data, albeit well outside of the range of actual values of density found in the data. The slope, –0.05, is negative, meaning that in more densely populated states the fatality rates are lower. The exact interpretation is that for every increase in population density of 100 people per square mile (the measurement metric of X_i) we predict a decrease in the fatality rate of .05 per 100 million vehicle miles driven (the measurement metric of Y_i).

AN ALTERNATIVE SPECIFICATION

It will prove useful at later points to be able to specify regression models in which the predictor variables have been put into "mean-deviation" form or "centered." What this means is that for every observation we have taken the value of the variable and subtracted from it the variable's mean value. Thus, if the predictor variable is X_i, the mean-deviated or centered variable is $(X_i - \bar{X})$. This centered variable will necessarily have a mean of zero, i.e., $(\bar{X} - \bar{X}) = 0$. We will then regress Y_i on $(X_i - \bar{X})$ rather than on X_i:

$$Y_i = b_0' + b_1' (X_i - \bar{X})$$

The question is how these new parameter estimates, when the predictor is centered, differ from the parameter estimates that result from the estimation we have considered to this point, with an uncentered predictor:

$$\hat{Y}_i = b_0 + b_1 X_i$$

To answer this question, we can examine the formulas for the parameter estimates that we gave earlier, but this time with a centered X_i:

$$b_1' = \frac{\Sigma(((X_i - \bar{X}) - (\bar{X} - \bar{X}))(Y_i - \bar{Y}))}{\Sigma((X_i - \bar{X}) - (\bar{X} - \bar{X}))^2}$$

$$b_0' = \bar{Y} - b_1(\bar{X} - \bar{X})$$

Since $(\bar{X} - \bar{X}) = 0$, it follows that

$$b_1' = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} = b_1$$

$$b_0' = \bar{Y}$$

In other words, centering the predictor has no effect upon the slope, i.e., $b_1' = b_1$, but it does change the intercept. The intercept with the predictor in mean-deviated or centered form will be the mean of Y_i.

Conceptually, if we think graphically about the scatterplot of data and the superimposed regression line, by centering all we are really doing is changing the scale of the X_i axis in the scatterplot, redefining the value of X_i so that its mean equals zero. Such a scatterplot with the X_i axis centered is given in Figure 5.8. As this makes clear we have not changed the observations in the scatterplot at all; we have simply shifted the origin point in the plot. In a fundamental sense, our prediction function has not changed at all; the same line, having the same slope, minimizes the squared errors of prediction. The only change derives from the change in the location of the zero point on the horizontal axis. That zero point is now at the mean of X_i and, accordingly, the value of the intercept (i.e., the value of Y_i where the prediction function crosses the vertical axis) changes. It is now $\bar{Y}$. Obviously, this means that the regression line inevitably goes through the point defined by the joint means of the two variables.

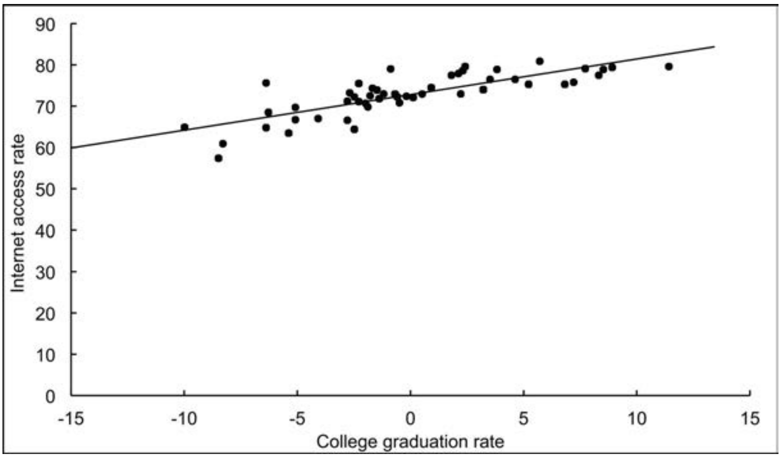
In the case of the estimated model predicting internet access rates from college graduation rates, when the latter variable is centered, the resulting least-squares parameter estimates are:

$$\hat{Y}_i = 72.806 + 0.862(X_i - \bar{X})$$

The slope is unchanged by the centering of X_i but the intercept has changed. It no longer equals 47.912, rather it equals $\bar{Y}$, which is 72.806. One can still use the conventional interpretation for the intercept: It remains the predicted value when the predictor equals zero, i.e., when $(X_i - \bar{X})$ equals zero. And of course $(X_i - \bar{X})$ equals zero when $X_i = \bar{X}$.

Because all that has changed with centering the predictor is the zero point on the horizontal access of the scatterplot, we are still dealing fundamentally with the same

FIGURE 5.8 Scatterplot and prediction function with centered *X*



data and the same regression line, making the same predictions for each observation. Unsurprisingly, then, centering the predictor leaves the mean squared error of the model unchanged. In a deep sense the regression results are unchanged by this transformation.

STATISTICAL INFERENCE IN TWO-PARAMETER MODELS

Now that the basics of estimation and interpretation in simple regression models are clear, we turn to the issue of statistical inference, asking questions about parameter values in such models. Our approach to statistical inference in the case of two-parameter models will be identical to the approach we adopted in the single-parameter case. That is, we will compare an augmented model (in which both parameters are estimated from the data) to a compact one (in which one or more parameters are fixed at a priori values). We will calculate the sum of squared errors associated with both the augmented and compact models; from these we will then compute PRE, the proportional reduction in error as we go from the compact to the augmented model. This PRE, and its associated *F* statistic, can then be compared to their critical values, making the assumptions of normality, constant variance, and independence of residuals. Such a comparison permits a test of the null hypothesis that the values of the parameters fixed in the compact model are in fact the true unknown parameter values. Put the other way around, we are testing whether the estimated parameter values in the augmented model depart significantly from their a priori values specified in the compact model.

Given that the augmented model for such comparisons is now the two-parameter simple regression model, there are alternative compact models with which it may be compared. On the one hand, one may be interested in asking questions about the slope, comparing the augmented model to a compact one in which the slope parameter has been set to some a priori value. On the other hand, there may arise occasions when one is interested in testing a null hypothesis about the intercept in this two-parameter model, comparing it to a compact one in which the intercept has been fixed at some a priori value. We consider each one in turn.

Inferences about β_1

To ask statistical inference questions about the slope is to ask about associations: as the predictor variable (X_i) increases, what is our conclusion about whether the dependent variable (Y_i) increases or decreases, and at what rate. Our augmented two-parameter model for such questions is the one we have been using throughout this chapter:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

The compact one, in its generic form, with which we will be making comparisons, is:

$$\text{MODEL C: } Y_i = \beta_0 + \mathrm{B}_1 X_i + \varepsilon_i$$

where B_1 is some a priori value to which the slope parameter has been set. As before, the null hypothesis that is to be tested by the comparison between these two models is:

$$\mathrm{H}_0 : \beta_1 = \mathrm{B}_1$$

In words, the inferential test is whether the slope differs significantly from the a priori value.

By far the most common form this comparison takes is when the a priori value of the slope in the compact model equals zero. In this case the compact model is:

$$\begin{aligned}\text{MODEL C: } & Y_i = \beta_0 + 0X_i + \varepsilon_i \\ & Y_i = \beta_0 + \varepsilon_i\end{aligned}$$

and the null hypothesis is:

$$\mathrm{H}_0 : \beta_1 = 0$$

The question being asked is whether X_i is a useful predictor of Y_i. If it is a useful predictor, then the predicted values of Y_i should change, either increasing or decreasing, as X_i changes. If we do no better with the augmented model (in which the slope is estimated from the data) than with the compact one (where the slope is constrained to equal zero) then it implies that the two variables may be unrelated. We do just as well making a constant prediction of all Y_i values regardless of an observation's X_i value as we do making conditional predictions.

For this comparison, the compact model is what was considered as the augmented model in the previous chapter. This will frequently be the case throughout the remainder of the book as we consider more complex models and additional inferential questions: What is for one question the augmented model becomes the compact model for a different question. So, in Chapter 4, we tested null hypotheses about a constant predicted value for every observation. Now we are testing whether we need to make conditional predictions and the compact model is one in which we estimate from the data a constant predicted value for each observation. That constant predicted value, when we estimate b_0 in Model C from the data, will not be the same as the intercept in Model A, when we estimate that model in our data. The best least-squares estimate of β_0 in Model C will be the mean of Y_i, just as it was (when we treated it as the augmented model) in Chapter 4. In the estimated augmented model (with two parameters), however, the estimate of β_0 will in general not be the mean of Y_i (unless of course the predictor has been centered). This makes clear another important point that will remain true as we consider more complex models in later chapters: the best estimate for a given parameter in general depends on what other parameters are estimated in the model. In the compact single-parameter model that we are considering, the best estimate of β_0 will not in general be the same as the best estimate of β_0 in the two-parameter augmented model.

To ask whether the slope equals zero, and thus whether X_i is a useful predictor of Y_i, is the most common inferential question that one might ask about the slope. But it is certainly not the only question one might ask. Other Model Cs, with other values of B_1, and thus other null hypotheses, might occasionally also be of interest. For instance, there are occasions when we are interested in testing the null hypothesis that:

$$H_0 : \beta_1 = 1$$

In this case, the augmented model remains the same two-parameter simple regression model, but Model C becomes:

$$\text{MODEL C: } Y_i = \beta_0 + 1X_i + \varepsilon_i$$

By casting all statistical inference questions in the form of Model A/Model C comparisons, testing such a null hypothesis becomes entirely feasible, even if standard statistical texts and software programs do not routinely provide such tests.

Testing the null hypothesis that $\beta_1 = 0$

In the context of our model that predicted states' internet access rates from their college graduation rates, let us test the first of the above null hypotheses, asking whether predictions of internet access are improved by making them conditional on college graduation rates, compared to a compact model that sets the slope at zero and thus makes a constant prediction for internet access for all states.

In terms of parameters, the models to be compared are:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\text{MODEL C: } Y_i = \beta_0 + \varepsilon_i$$

These are estimated as:

$$\text{MODEL A: } \hat{Y}_i = 47.0912 + 0.862X_i$$
$$\text{MODEL C: } \hat{Y}_i = 72.806$$

In Figures 5.9 and 5.10, we present the results of these two models for each state. In the first two columns of values in Figure 5.9 both Y_i (internet access rate) and X_i (college graduation rate) values for each state are given. Then the next three columns provide the results of Model C. The first of these provides the predicted values, $\hat{Y}_{iC}$. These are necessarily the same value for every state, since Model C predicts simply the mean internet access rate for each. Next, for each state we give its error, e_i, and its squared error, e_i^2. Across all 50 states, the sum of squared errors for Model C equals 1355.028. This model has a single estimated parameter, hence n – PC equals 49 (with 50 states). Thus, the mean square error for Model C is:

$$1355.028/49 = 27.654$$

Given that this is the simplest single-parameter model, as defined in the previous chapter, this mean square error is also called the variance of Y_i.

FIGURE 5.9 Model comparison for internet access data

US state	*Internet access rate*	*College graduation rate*	$\hat{Y}_{iC}$	e_{iC}	e^2_{iC}	$\hat{Y}_{iA}$	e_{iA}	e^2_{iA}	$(\hat{Y}_{iA}-\hat{Y}_{iC})^2$
AL	63.5	23.5	72.806	–9.306	86.602	68.169	–4.669	21.780	21.502
AK	79.0	28.0	72.806	6.194	38.366	72.048	6.952	48.330	0.575
AZ	73.9	27.4	72.806	1.094	1.197	71.531	2.369	5.612	1.626
AR	60.9	20.6	72.806	–11.906	141.753	65.669	–4.769	22.743	50.937
CA	77.9	31.0	72.806	5.094	25.949	74.634	3.266	10.667	3.342
CO	79.4	37.8	72.806	6.594	43.481	80.496	–1.096	1.201	59.136
CT	77.5	37.2	72.806	4.694	22.034	79.978	–2.478	6.140	51.438
DE	74.5	29.8	72.806	1.694	2.870	73.600	0.900	0.810	0.630
FL	74.3	27.2	72.806	1.494	2.232	71.358	2.942	8.655	2.097
GA	72.2	28.3	72.806	–0.606	0.367	72.307	–0.107	0.011	0.249
HI	78.6	31.2	72.806	5.794	33.570	74.806	3.794	14.394	4.000
ID	73.2	26.2	72.806	0.394	0.155	70.496	2.704	7.312	5.336
IL	74.0	32.1	72.806	1.194	1.426	75.582	–1.582	2.503	7.706
IN	69.7	23.8	72.806	–3.106	9.647	68.428	1.272	1.618	19.167
IA	72.2	26.4	72.806	–0.606	0.367	70.669	1.531	2.344	4.567
KS	73.0	31.1	72.806	0.194	0.038	74.720	–1.720	2.958	3.663
KY	68.5	22.6	72.806	–4.306	18.542	67.393	1.107	1.225	29.301
LA	64.8	22.5	72.806	–8.006	64.096	67.307	–2.507	6.285	30.239
ME	72.9	28.2	72.806	0.094	0.009	72.220	0.680	0.462	0.343
MD	78.9	37.4	72.806	6.094	37.137	80.151	–1.251	1.565	53.949
MA	79.6	40.3	72.806	6.794	46.158	82.651	–3.051	9.309	96.924
MI	70.7	26.9	72.806	–2.106	4.435	71.100	–0.400	0.160	2.910
MN	76.5	33.5	72.806	3.694	13.646	76.789	–0.289	0.084	15.864
MS	57.4	20.4	72.806	–15.406	237.345	65.497	–8.097	65.561	53.422
MO	69.8	27.0	72.806	–3.006	9.036	71.186	–1.386	1.921	2.624
MT	72.1	29.0	72.806	–0.706	0.498	72.910	–0.810	0.656	0.011
NE	72.9	29.4	72.806	0.094	0.009	73.255	–0.355	0.126	0.202
NV	75.6	22.5	72.806	2.794	7.806	67.307	8.293	68.774	30.239
NH	80.9	34.6	72.806	8.094	65.513	77.737	3.163	10.005	24.315
NJ	79.1	36.6	72.806	6.294	39.614	79.461	–0.361	0.130	44.289
NM	64.4	26.4	72.806	–8.406	70.661	70.669	–6.269	39.300	4.567
NY	75.3	34.1	72.806	2.494	6.220	77.306	–2.006	4.024	20.250
NC	70.8	28.4	72.806	–2.006	4.024	72.393	–1.593	2.538	0.171
ND	72.5	27.1	72.806	–0.306	0.094	71.272	1.228	1.508	2.353
OH	71.2	26.1	72.806	–1.606	2.579	70.410	0.790	0.624	5.741
OK	66.7	23.8	72.806	–6.106	37.283	68.428	–1.728	2.986	19.167
OR	77.5	30.7	72.806	4.694	22.034	74.375	3.125	9.766	2.462
PA	72.4	28.7	72.806	–0.406	0.165	72.651	–0.251	0.063	0.024
RI	76.5	32.4	72.806	3.694	13.646	75.841	0.659	0.434	9.211
SC	66.6	26.1	72.806	–6.206	38.514	70.410	–3.810	14.516	5.741
SD	66.6	26.6	72.806	–1.706	2.910	70.841	0.259	0.063	3.861
TN	71.1	24.8	72.806	–5.806	33.710	69.290	–2.290	5.244	12.362
TX	67.0	27.5	72.806	–1.006	1.012	71.617	0.183	0.033	1.414
UT	71.8	31.3	72.806	6.794	46.158	74.893	4.707	22.156	4.356
VT	79.6	35.7	72.806	2.494	6.220	78.685	–3.385	11.458	34.563
VA	75.3	36.1	72.806	2.994	8.964	79.030	–3.230	10.433	38.738
WA	75.8	32.7	72.806	6.094	37.137	76.099	2.801	7.846	10.844
WV	78.9	18.9	72.806	–7.906	62.505	64.204	0.696	0.484	73.994
WI	64.9	27.7	72.806	0.194	0.038	71.789	1.211	1.467	1.034
WY	75.5	26.6	72.806	2.694	7.258	70.841	4.659	21.706	3.861

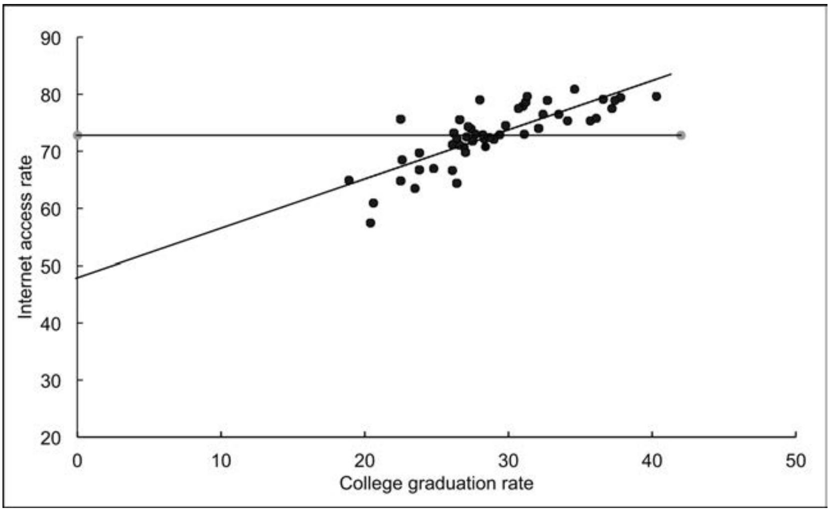
The next three columns in Figure 5.9 give the parallel results for Model A, first the predicted values ($\hat{Y}_{iA}$) and then errors of prediction and squared errors. Notice in this case that the predicted values now are not constant; rather, they are a linear function of the value of X_i for each state. And, again, the sum of the squared errors across the 50 states is SSE for Model A, equaling 480.00. In Model A we have estimated two parameters, accordingly PA equals 2, n – PA equals 48, and the mean square error for this model is 10.00.

Model A, since it uses an additional parameter to model the data, necessarily does better than Model C in terms of sums of squared errors. That is not to say, however, that in every state the prediction made by Model A is better than that made by Model C. Examine Kansas (KS) for instance. Its internet access rate is 73.0%. Model C predicts it to be the mean of 72.806, which is a difference of 0.194. Model A, on the other hand, making a prediction conditional on Kansas' college graduate rate, predicts it to be 74.720, missing by 1.720. For this particular state, the Model A prediction is not as accurate as the Model C prediction. Yet, on average across states the squared errors of prediction must be at least as small from Model A as they are from Model C, simply because an additional parameter has been estimated.

Figure 5.10 presents the results graphically. The horizontal line superimposed on the data points is the prediction function of Model C. The other line is the prediction function of Model A. The inferential question that we now want to ask is whether the reduction in the SSEs when we replace Model C with Model A has been worth the reduction in the error degrees of freedom due to estimating an additional parameter. Graphically, the question is: Do we do sufficiently better with the sloped line to merit the added complexity over and above the horizontal prediction function? To answer this, we compute PRE, the proportional reduction in error as we move from Model C to Model A:

$$\text{PRE} = \frac{\text{SSE(C)} - \text{SSE(A)}}{\text{SSE(C)}} = \frac{1355.028 - 480.003}{1355.028} = .6458$$

FIGURE 5.10 Model C and Model A



Thus, when we make conditional predictions in this example, the total errors of prediction are reduced by more than 64% compared to simply predicting the mean level of internet access for each state.

The numerator of PRE is the sum of squares reduced (SSR) and can be calculated directly by taking the squared difference between the predicted value for Model A and that for Model C, and summing these across the 50 states:

$$\text{SSR} = \text{SSE(C)} - \text{SSE(A)} = \Sigma(\hat{Y}_{iA} - \hat{Y}_{iC})^2$$

In the final column of Figure 5.9, we present these squared differences in predicted values for every state. The sum of the numbers in this final column necessarily equals the difference in the sum of squared errors between the two models, i.e., 1355.028 – 480.003 or 875.025.

We can also compute the *F* statistic associated with the comparison between these two models. Below we do this with the equivalent expressions for *F*, either in terms of PRE or in terms of the sums of squares of the two models:

$$F = \frac{\text{PRE/(PA–PC)}}{(1 - \text{PRE})/(n - \text{PA})} = \frac{.6458/1}{(1 - .6458)/48} = 87.50$$

$$= \frac{\text{SSR/(PA – PC)}}{\text{SSE(A)}/(n - \text{PA})} = \frac{875.025/1}{480.003/48} = 87.50$$

Using the appropriate tables in the Appendix, we compare the computed values of either PRE or *F* with their critical values, given that the assumptions of normality, constant variance, and independence of errors are met. The critical values at $\alpha = .05$ (with 1 and 48 degrees of freedom) are approximately .08 for PRE and 4.04 for *F*. Clearly, our computed values exceed these, and hence we can reject Model C in favor of the two-parameter Model A that makes conditional predictions. Our conclusion is that college graduation rates are a useful predictor of internet access rates in these data. Further, once we have rejected the null hypothesis of no relationship between the two variables, we can make a conclusion about the direction of the relationship between them, based on the sign of the estimated slope: In states where the percentage of people who graduated from college is higher, there are higher rates of internet access.

Figure 5.11 summarizes the results of the statistical analysis in an ANOVA table of the same type as we developed in Chapter 4. The first row provides information about the reduction in error achieved by including X_i as a predictor in the model (i.e., by using Model A instead of Model C). The entry in the SS column for that row is the SSR computed earlier. Associated with this SSR is a single degree of freedom, PA – PC = 1, for this comparison. The next row provides information about the error remaining in Model A, and its associated degrees of freedom (n – PA). The final row provides similar information for Model C. Calculating MS = SS/df provides the basis for calculating *F* according to the sum of squares formula presented above. In the *p* column is indicated the fact that the computed PRE and *F* exceed the critical value with $\alpha = .05$. Finally, the value of PRE is given, computed as the SSR divided by SSE for Model C.

One final comment is appropriate about the statistics we have just computed. Throughout this book, we will use PRE to refer to the proportional reduction in error when replacing any compact model with an augmented one. In this sense, PRE is a very general statistic that is informative regardless of the specifics of the models that are being compared. However, in the history of inferential statistics, a variety of specific names for PRE have been developed for particular model comparisons. For the case at hand, testing whether the slope in a single predictor model is different from zero, the square root of PRE is known as the Pearson correlation coefficient, commonly referred to simply as the correlation or *r*. Thus, the inferential test we have just conducted is equivalent to a test of whether the true correlation between two variables differs from zero. If PRE and *F* do not surpass their critical values, then it means that the true value of PRE (η^2) may equal zero. In fact, since the *F* statistic we have computed has only a single degree of freedom in its numerator, its square root is a *t* statistic with n – PA ($n - 2$ in this case) degrees of freedom. It is simple algebra to show that the square root of the *F* formula we have given is the same as the *t* that is typically given in traditional statistics texts for testing whether a correlation between two variables is significantly different from zero:

$$t_{n-2} = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}}$$

As a result, our conclusion in this case that Model A does significantly better than Model C is equivalent to concluding that the two variables are related, that β_1 differs from zero, that the true correlation differs from zero, and that the true PRE (η^2) differs from zero.

Testing null hypotheses for other values of β_1

There may be occasions when theory makes strong predictions about particular values of slopes other than zero. For instance, if two variables are measured in the same metric, then we might be interested in whether a one-unit difference on one variable is associated with a one-unit difference on the other. To illustrate, we continue to use the internet access/college graduation example. Since they are both measured in the same metric (percentage of the state who either has internet access or has attended college), we will ask the question of whether a 1% increase in college graduation is associated with a 1% increase in internet usage on average across the states. Obviously, we do not have any strong theory that would make this prediction; we use it for illustrative purposes only. The model comparison for this question is:

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MODEL C: $Y_i = \beta_0 + 1X_i + \varepsilon_i$

And the null hypothesis is that $\beta_1 = 1$.

FIGURE 5.11 ANOVA source table test of H_0: $\beta_1 = 0$ in simple regression context

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Reduction (using $b_1 = 0.862$)	875.025	1	875.025	87.50	< .001	.646
Error (using $b_1 = 0.862$)	480.003	48	10.000			
Total error (using $b_1 = 0$)	1355.028	49	27.654			

Unfortunately few statistical programs readily permit the estimation of parameters in a model in which some parameters are fixed at values other than zero. In this case, however, we can easily estimate β_0 in Model C by noting that if we subtract X_i from both sides of Model C we obtain:

$$Y_i - X_i = \beta_0 + \varepsilon_i$$

so the mean of the constructed variable $Y_i - X_i = 43.922$ estimates β_0 in Model C. The estimated models are then:

$$\text{MODEL A: } Y_i = 47.912 + 0.862X_i + e_i$$

$$\text{MODEL C: } Y_i = 43.922 + 1X_i + e_i$$

The SSE for Model C equals 502.486, again with n – PC equal to 49. Model A makes better predictions, since as we saw above its sum of squared errors equals 480.003. SSR equals 22.483, with PRE and F computed as:

$$\text{PRE} = \frac{22.483}{502.486} = .045$$

$$F_{1,48} = \frac{22.483/1}{502.486/48} = 2.15$$

These values do not beat their respective critical values. Thus we cannot reject the null hypothesis—there is no evidence to conclude that the true slope is different from 1.00.

Confidence intervals of β_1

Recall from Chapter 4 that the confidence interval defines the range of values for the parameter for which we would fail to reject the null hypothesis. In other words, if we tested a null hypothesis that the parameter equals a value that lies within the confidence interval, we would fail to reject that null hypothesis. If the null hypothesis to be tested specifies that the parameter equals a value that lies outside of the confidence interval, we would reject that null hypothesis. In this sense, the confidence interval is entirely redundant with inferential tests about the value of a parameter.

In the case of β_1 in the two-parameter model we are considering, its confidence interval is given as:

$$b_1 \pm \sqrt{\frac{F_{\text{crit; }1,\, n-2;\alpha}\text{MSE}}{(n-1)s_X^2}}$$

where $F_{\text{crit }1,n-2;\alpha}$ is the critical value of F at level α, with degrees of freedom of PA – PC = 1 and n – PA = n – 2. MSE is the mean square error from the augmented model, again based on n – 2 degrees of freedom. And s^2 is the variance of the predictor variable.

For the internet access/college graduation example, the critical F with 1 and 48 degrees of freedom, using $\alpha = .05$, equals 4.04, MSE from Model A equals 10.00, and the variance of the predictor variable (college graduation rates) equals 24.04. Accordingly, the confidence interval for β_1 is:

$$0.862 \pm \sqrt{\frac{4.04(10.00)}{49(24.04)}} = 0.862 \pm 0.185$$

or

$$0.677 \leq \beta_1 \leq 1.047$$

Based on the present data, we can thus say we are confident that the true value for the slope in this two-parameter regression model, predicting internet access rates from college graduation rates, lies somewhere between 0.677 and 1.047. Notice that zero lies outside of this interval and that 1.00 lies within it. Both of these are consistent with the results already reported for our two null hypotheses: the null hypothesis that the parameter equals zero was rejected; the null hypothesis that it equaled 1.00 was not.

Although we might lament the fact that many statistical packages do not permit the estimation of models in which parameter values are fixed at values other than zero, the confidence interval permits a general approach for testing any null hypothesis about the slope. If the value that is specified by the null hypothesis lies within the interval, it would not be rejected. All other null hypotheses would be.

The formula for the confidence interval provides some insights into the factors that influence statistical power—the probability of rejecting the null hypothesis when it is false—and how to improve it. In general, the narrower the confidence interval, the more precision we have in estimating a parameter and statistical power means that we have greater precision, i.e., narrower confidence intervals. According to the formula, what are the factors that affect the width of the interval, and therefore power?

First, the critical value of F affects its width. If we use a smaller α level, thus reducing Type I errors, the critical value of F increases, thereby widening the confidence interval and resulting in less power and greater probability of Type II errors. This is the inherent tradeoff between Type I and Type II statistical errors.

Second, the width of the confidence interval is affected by the mean square error from the augmented model. As the variability of errors of prediction is reduced, the confidence interval becomes narrower. Thus, whatever we can do to reduce error, such as improving the quality of measurement of Y_i, will increase power.

Third, as n increases, all else being equal, power increases. This is reflected by the fact that $n - 1$ appears in the denominator of the confidence interval.

And, finally, the variance of the predictor variable X_i appears in the denominator. As X_i becomes more variable, the interval narrows and power improves. Given some predictor variable, we will examine its effects as a predictor with more precision (assuming a linear model) if we make sure that we sample widely across its values.

Power analysis in tests of simple regression

In Chapter 4 we performed "what if" power analyses for the simple model making a constant prediction for each observation. We can use exactly the same process to ask "what if" power analysis questions for simple regression models using one predictor variable. We perform "what if" power analyses for particular values of the true proportional reduction in error, η^2, which may be of interest, in exactly the same way as before. For example, for the internet access data, we might want to know the power of detecting a relationship between it and some variable—detecting that the slope for a predictor is different from zero—when in fact we think that $\eta^2 = .20$. To do this, we can use a software program, such as R or SAS, to calculate power as we mentioned in the last chapter. For this Model A/Model C comparison, PA – PC = 1 and n – PA = 48.

If $\eta^2 = .20$ then the power calculator in SAS informs us that the power of our test is roughly .92. That is, if we were to do a study with an n of 50 and expect to find a relationship between a predictor and Y_i with a true PRE of .20, we would have roughly a 92% chance of correctly rejecting the null hypothesis.

Given that we now know the procedure for asking questions to determine the power with which we can assess whether one variable is significantly related to another in a simple regression model, we need to know what values of η^2 are appropriate and expected for such comparisons. As before, prior experience in a research domain may provide values of PRE that can be used directly in the power table. For the simple regression case, we might have estimates of PRE available from previous research, typically reported as the correlation between two variables, rather than as PRE itself. In this case, we need to square the correlation coefficient to obtain the estimate of PRE, since PRE = r^2 for this question. As before, we would want to convert past empirical values of PRE to unbiased estimates of η^2, using the same adjustment formula as before:

$$\text{Unbiased estimate of } \eta^2 = 1 - \frac{(1 - \text{PRE})(n - \text{PC})}{n - \text{PA}}$$

To illustrate, suppose prior research has reported a correlation between two variables of .33 based on a sample size of 30. We intend to do a study, examining the relationship between the same two variables, but we intend to use an n of 50. What we would like to know is the power of our planned study, if in fact the two variables are related as strongly as they were reported to be in the prior research. To do this, we first convert the previously reported correlation to PRE, by squaring it: $.33^2 = .11$. We then convert this PRE to an unbiased estimate of η^2 using the above formula and the sample size from the prior study that reported the .33 correlation:

$$\text{Unbiased estimate of } \eta^2 = 1 - \frac{(1 - .11)(29)}{28} \approx .078$$

We then use SAS to estimate our power in the new study we plan to conduct. With an n of 50, n – PA for our study will be 48, and with an anticipated η^2 of .078, our power is approximately .51. Given this result, we may want to think further about our anticipated study. It might be worthwhile, for instance, to recruit a larger sample to increase power.

If we do not have relevant prior experience for estimating η^2, we can use the values suggested in Chapter 4 for "small" ($\eta^2 = .03$), "medium" ($\eta^2 = .10$), and "large" ($\eta^2 = .30$) effects. From these and the anticipated n for a new study, we can get the approximate power.

A third and final approach for finding an appropriate value of η^2 for "what if" power analyses for the simple regression model involves guesses about the parameter values and variances, just as in Chapter 4. Again, to have reasonable expectations about the parameter values and variances generally requires as much or more experience in a research domain as is necessary to know typical values of PRE. We present it, however, for completeness.

The formula that relates true values of the parameters to η^2, the true value of PRE, is:

$$\eta^2 = \beta_1^2 \frac{\sigma_X^2}{\sigma_Y^2}$$

where β_1 is the true parameter value for the slope, σ_x^2 is the true variance of the predictor variable, and σ_y^2 is the true variance of the dependent variable. In other words, given some alternative hypothesis that specifies what we think is the correct value for the slope, β_1, and given that we want to determine the power of our test of the null hypothesis that β_1 equals zero, we can calculate the corresponding η^2 using the above expression, assuming we have estimates of the variances of both X_i and Y_i. We can then take that value, and the anticipated $n - 2$ for the study in the planning, and estimate power.

Inferences about β_0

Our discussion so far has concentrated exclusively on inferences about the slope in the two-parameter simple regression model. But one may certainly also compare this augmented two-parameter model with a compact one that fixes the intercept, rather than the slope, at some a priori value:

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MODEL C: $Y_i = \mathrm{B}_0 + \beta_1 X_i + \varepsilon_i$

where B_0 represents an a priori value for the intercept. The null hypothesis tested by such a comparison would then be:

$\mathrm{H}_0 : \beta_0 = \mathrm{B}_0$

Although perhaps not of frequent theoretical interest in the behavioral sciences, one particular form of this compact model involves what is known as "regression through the origin" in which the intercept is fixed at zero:

MODEL C: $Y_i = 0 + \beta_1 X_i + \varepsilon_i$

Recall that the intercept is defined as the value predicted by the model when X_i equals zero. Thus, regression through the origin means that the line goes through the (0, 0) point on the scatterplot.

Importantly, null hypothesis tests about the intercept are of a fundamentally different nature than those about the slope. When we are making inferences about the slope, we are asking about the rate of change in predicted values. And when we test that the slope is zero, we are asking whether there is any change in the predicted values when the predictor varies, i.e., whether X_i and Y_i are related. On the other hand, inferences about the intercept are inferences about predicted values, not inferences about changes in predicted values. Specifically, we are asking whether the predicted value of Y_i *when X_i equals zero* differs significantly from some a priori value, be it zero (in regression through the origin) or some other value.

In many cases, there is no intrinsic interest in the predicted value of Y_i when X_i equals zero. This may be because the value of zero lies outside of the range of the X_i values represented in the dataset, as in the data that we have been using where no state has a college graduation rate that is near zero. Even if zero is a value within the range of X_i in the dataset, it may not be a point that has much theoretical interest. However, with simple transformations of X_i, such as centering it or deviating it from its mean (discussed earlier), the zero value becomes of considerably greater potential interest. Earlier in this chapter we saw that when X_i is centered the intercept in the estimated model will equal the mean of Y_i (i.e., $\bar{Y}$). Accordingly, with a centered predictor, the following Model A/Model C comparison is equivalent to asking questions about the mean of Y_i:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

$$\text{MODEL C: } Y_i = \text{B}_0 + \beta_1 (X_i - \bar{X}) + \varepsilon_i$$

with the following null hypothesis:

$$\text{H}_0\text{: } \beta_0 = \text{B}_0 \text{ or } \mu_Y = \text{B}_0$$

where μ_Y is the true mean of Y_i.

Let us illustrate such a test with the internet access data that we have used throughout this chapter, asking the same question that we did at the end of the last chapter—whether the projection of a mean rate of 75% was overly optimistic for the year 2013. But this time, we will ask the question in the context of a simple regression model that makes conditional predictions of internet access based on states' college graduation rates. We will then compare our test in this model with our results from Chapter 4 to examine how the present test differs from that used there.

Estimating a model in which internet access rates are regressed on college graduation rates, with the latter variable in its centered or mean-deviated form, gives the following estimates:

$$\text{MODEL A: } \hat{Y}_i = 72.806 + 0.862(X_i - \bar{X})$$

with a sum of squared errors of 480.003. As explained earlier, the intercept in this model is now the mean value for the internet access variable, while the slope has not changed compared to the model in which X_i was not centered. Additionally, in a deep sense, this model is identical to the one with X_i uncentered, in that it makes the same predictions and therefore necessarily has the same sum of squared errors.

To test the null hypothesis that the true mean of Y_i equals 75, we want to compare this model to a compact one in which the intercept has been fixed at 75:

$$\text{MODEL C: } \hat{Y}_i = 75 + b_1(X_i - \bar{X})$$

This comparison permits a test of the null hypothesis:

$$\text{H}_0\text{: } \beta_0 = 75 \text{ or } \mu_Y = 75$$

As we said previously, many computer packages for data analysis do not readily permit the estimation of models in which parameters have been fixed at values other than zero. In this case, with a centered predictor, it can be shown that the best least-squares estimate for β_1 does not change even if we fix the intercept at some a priori value.[1] Accordingly, Model C is estimated as:

$$\text{MODEL C: } \hat{Y}_i = 75 + 0.862(X_i - \bar{X})$$

The sum of squared errors from this model can be directly computed across the observations. More simply, it can be computed by first calculating the sum of squares reduced (SSR) as we move from Model C to Model A.

Recall that SSR = $\Sigma(Y_{iA} - Y_{iC})^2$, accordingly:

$$\text{SSR} = \Sigma((72.806 + 0.862(X_i - \bar{X})) - (75 + 0.862(X_i - \bar{X})))^2$$

Because both predicted values have the same slope for the centered predictor, this reduces to:

$$\text{SSR} = \Sigma((72.806) - (75))^2 = 50(-2.194)^2 = 240.682$$

Accordingly, the sum of squared errors for the compact model is:

$$\text{SSE(C)} = \text{SSE(A)} + \text{SSR} = 480.003 + 240.682 = 720.685$$

Now that we have the sums of squared errors, we can calculate PRE and *F* for the comparison and the test of the null hypothesis:

$$\text{PRE} = \frac{\text{SSR}}{\text{SSE(C)}} = \frac{240.682}{720.685} = .334$$

$$F = \frac{.334/1}{(1 - .334)/48} = \frac{240.682/1}{480.003/48} = 24.07$$

All of this is summarized in the ANOVA source table of Figure 5.12.

Let us now compare these results with the test of the same null hypothesis reported at the end of Chapter 4, in the context of the simplest single-parameter model. There, estimated Models A and C were:

MODEL A: $\hat{Y}_i = 72.806$

MODEL C: $\hat{Y}_i = 75$

The sum of squared errors associated with Model A was 1355.046, that is, the total sum of squares of Y_i around its mean. While the SSR was found to be:

$$\text{SSR} = \Sigma((72.806) - (75))^2 = 50(-2.194)^2 = 240.682$$

Thus, the results of this test of the same null hypothesis yielded the ANOVA source table in Figure 5.13.

Although both tests resulted in the rejection of the null hypothesis, clearly the two approaches differ substantially in terms of the obtained PRE and *F*. Those values, in the context of the two-parameter simple regression model, are more than twice what they were for the same test in the context of the single-parameter model of Chapter 4. And this difference is attributable entirely to the difference in the sum of squared errors for Model A (and the concomitant difference in n – PA). The sum of squared errors for Model A in the context of the simple regression model equals 480.003, with n – PA equal to 48. The same values in the context of the single-parameter model of Chapter 4 are SSE = 1355.046 and n – PA = 49. As a result, the MS error values (denominators of *F*) are markedly different: 10.00 (for the simple regression model) versus 27.654 (for the Chapter 4 single-parameter model). Importantly, the numerator of *F* is the same in both tests, with SSR equal to 240.682 and PA – PC = 1.

The difference in the sums of squared errors for Model A in the two source tables is, not surprisingly, the sum of squares that is attributable to the predictor variable

FIGURE 5.12 ANOVA source table test of H_0: $\beta_0 = \mu_Y = 75$ in simple regression context

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Reduction (using $b_0 = 72.806$)	240.682	1	240.682	24.07	< .001	.334
Error (using $b_0 = 72.806$)	480.003	48	10.000			
Total error (using $b_0 = 75$)	720.685	49	14.708			

FIGURE 5.13 ANOVA source table test of H_0: $\beta_0 = \mu_Y = 75$ in single-parameter model (from Chapter 4)

Source	*SS*	*df*	*MS*	*F*	*p*	*PRE*
Reduction (using $b_0 = 72.806$)	240.682	1	240.682	8.71	< .01	.151
Error (using $b_0 = 72.806$)	1355.046	49	27.654			
Total error (using $b_0 = 75$)	1595.728	50	31.915			

(college graduation rates), on the basis of which conditional predictions are made in the simple regression model. That is, earlier, we reported that the SSR attributable to a Model A that made conditional predictions of internet access, using college graduation rates as a predictor, compared to a Model C that predicted the mean value of Y_i, was 875.025. This is exactly the difference between each Model A used in these two tests. Model A for the Chapter 4 single-parameter version of the test makes unconditional predictions of Y_i; Model A for the simple regression version of the test makes predictions of Y_i conditional on (centered) X_i. Making these conditional predictions means that Model A in the simple regression context has one fewer degrees of freedom for error (n – PA = 48) than Model A in the Chapter 4 version of the test (where n – PA = 49). But the loss in degrees of freedom has been more than compensated for by the substantial difference between the sums of squared errors of the two Model As. As a result, the test in the context of the conditional simple regression model has substantially more statistical power than the same test conducted in the context of the single-parameter model of Chapter 4.

In Chapter 4 we mentioned that the test we reported is known as the single-sample *t*-test. The advantage of our model comparison approach is that we have been able to generalize this test to cases where conditional predictions are made by the models that are compared. In the jargon traditionally used in the statistical inference literature, we have just conducted a single-sample *t*-test while controlling for a "covariate."

A further advantage of our approach is that it permits us to conduct inferential tests about predicted values other than the mean. Suppose, for instance, that we had some reason to want to ask about internet usage rates in states where the college graduation rate was 70%. Rather than centering the predictor around its mean value, one could deviate the predictor from the value of 70. Then one could estimate a Model A in which the predictor variable was this deviated variable:

$$\text{MODEL A: } Y_i = \beta_0 + \beta_1(X_i - 70) + \varepsilon_i$$

In a deep sense we would be dealing with the same conditional model; we have simply moved the zero point on the x-axis to what was the value of 70. Accordingly, the slope remains the same, while the estimated intercept in the model would be the predicted value of Y_i when X_i equals 70. This model might then be compared to a Model C that uses the same deviated predictor but fixes the intercept at some a priori value of interest.[2]

Confidence interval for the intercept

The confidence interval for the intercept represents the range of values for the intercept that would not be rejected by an inferential statistical test. In simple regression models, with a single predictor variable, the confidence interval for the intercept is:

$$b_0 \pm \sqrt{\frac{F_{\text{crit; 1, } n-2; \alpha} \text{MSE}(\Sigma X_i^2)}{n\Sigma(X_i - \bar{X})^2}}$$

This is the confidence interval for the intercept regardless of whether the predictor has been deviated from some value or not. If it has been deviated, then of course the X_i terms in the confidence intervals are the new values following deviation.

When using a centered predictor, deviated from its mean value, then $\bar{X}$ for the centered predictor equals zero, and the above formula for the confidence interval reduces to:

$$b_0 \pm \sqrt{\frac{F_{\text{crit; 1, } n-2; \alpha} \text{MSE}}{n}}$$

which is the formula that we gave in Chapter 4 for the confidence interval for β_0 in the single-parameter model (except there the critical F value had 1 and $n - 1$ degrees of freedom). Of course, with a predictor variable in the model that is a useful predictor of Y_i, the MSE in the numerator of this confidence interval should be considerably smaller than the MSE in the numerator of the interval used in Chapter 4, without a predictor. This difference reflects the increase in power when conducting inferential tests about the mean in the context of a useful predictor compared to the same test in the single-parameter context of Chapter 4.

To illustrate, with the centered predictor from the simple regression model used in this chapter, the confidence interval for β_0 (equivalently for μ_Y) equals:

$$72.806 \pm \sqrt{\frac{4.04(10.000\,)}{50}}$$

$$71.907 \leq \beta_0 \leq 73.705$$

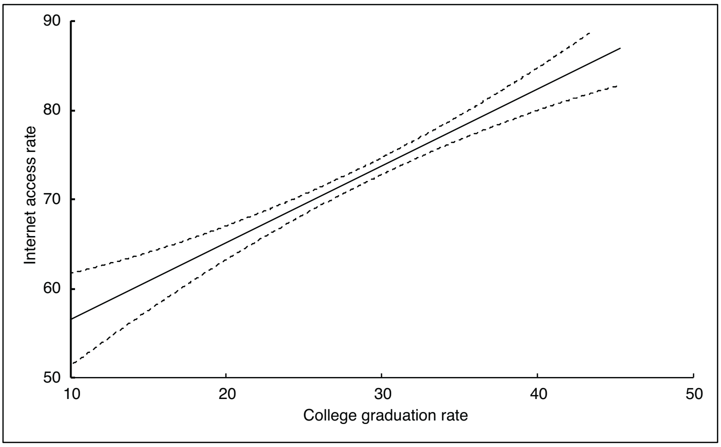
On the other hand, in the context of the single-parameter model of Chapter 4, the same confidence interval equals:

$$72.806 \pm \sqrt{\frac{4.03(27.654)}{50}}$$

$$71.313 \leq \beta_0 \leq 74.299$$

Clearly, the confidence interval for the mean of Y_i is smaller in the context of the simple regression model, reflecting the substantial increase in power resulting from the inclusion of the predictor variable (with its associated reduction in errors of prediction).

FIGURE 5.14 Confidence limits of predicted values



Using the general formula for the confidence interval for β_0 in this two-parameter model given above:

$$b_0 \pm \sqrt{\frac{F_{\text{crit};\,1,\,n-2;\alpha}\text{MSE}\,(\Sigma X_i^2)}{n\Sigma(X_i - \bar{X})^2}}$$

We can generate confidence intervals for predicted values at all possible levels of the predictor variable (by deviating the predictor from each of those levels and then calculating the resulting confidence intervals for the varying intercepts). In Figure 5.14 we have graphed the 95% confidence limits for the predicted values of internet access at values of the predictor (college graduation rate) ranging from 10% to about 45%. The middle straight line in the graph represents the predicted values and the curved lines above and below it represent the upper and lower confidence limits. What this figure makes clear is that the confidence interval is narrowest near the mean value of the predictor, where the predicted value is the mean of the dependent variable, and it becomes wider as we depart in either direction from that mean. Thus, we have greater precision in inferring predicted values near the joint mean of the bivariate distribution than when we move further away along the horizontal axis.

Two-Parameter Model Comparisons

To conclude this chapter on two-parameter simple regression models, we should note that MODEL A/MODEL C comparisons are now also possible involving PA – PC > 1. For instance, suppose we wanted to simultaneously ask about fixed a priori values both for the intercept and for the slope:

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MODEL C: $Y_i = \text{B}_0 + \text{B}_1 X_i + \varepsilon_i$

where both B_0 and B_1 are fixed a priori values (zero or any other value). The resulting null hypothesis from this comparison has two different components to it:

$$H_0 : \beta_0 = B_0; \beta_1 = B_1$$

To illustrate a situation where this sort of model comparison might be of interest, suppose we had data on the heights of each member of father–son pairs and we wanted to examine how these two height measures were related. We might estimate a Model A predicting each son's height from his father's height, estimating both the slope and intercept as in Model A above. This model might then be meaningfully compared with a Model C in which the intercept was fixed at 0 and the slope at 1, yielding the following model comparison and null hypothesis:

MODEL A: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

MODEL C: $Y_i = X_i + \varepsilon_i$

$H_0 : \beta_0 = 0; \beta_1 = 1$

Such a comparison asks whether sons and fathers are perfect resemblances of each other in terms of height. Except for error, Model C assumes that sons' heights equal their fathers' heights.

There is nothing statistically wrong with such a two-parameter model comparison and such a compound null hypothesis. The resulting PRE and *F* would be computed in the same ways as always, albeit with PA – PC = 2 degrees of freedom for the numerator. Since there is nothing to estimate in Model C, the predicted values and sum of squared errors from this model could be easily obtained.

The problem comes in interpreting the results. If Model C is rejected in favor of Model A (i.e., if the null hypothesis is rejected) we will not be able to be confident about *why* it was rejected. Maybe it is the case that the a priori value in Model C for the intercept is wrong. Maybe it is the case that the a priori value in Model C for the slope is wrong. Maybe both a priori values are wrong. All we can say is that Model A is preferred over Model C, but we will not know why.

It is for this reason that we prefer model comparisons where PA – PC = 1, where the numerator of the *F* statistic has only a single degree of freedom. There is nothing wrong with statistical tests involving more than one degree of freedom in the numerator of *F*, and we will occasionally discuss them and even recommend them. In general, however, we will refrain from such unfocused model comparisons. We simply note that in the context of the present two-parameter simple regression model they become possible for the first time.

SUMMARY

In this chapter we considered models with a single predictor, measured more or less continuously. We started by considering the definitions of both the intercept and slope in such models, with the former being a particular predicted value (when the predictor equals zero) and the latter being the unit difference between predicted values. We then provided formulas for estimating these parameters and used these to illustrate alternative ways of thinking about what a slope estimate means. Finally, we considered models in which the predictor is centered or put into mean-deviation form. In such cases, the slope of the predictor is unchanged while the intercept equals the mean of the data variable, *Y*.

The second half of the chapter was devoted to model comparisons, treating the two-parameter, single-predictor model as Model A and comparing it to an alternative Model C, testing inferences about the slope and the intercept (and both simultaneously). Inferential tests of the slope most frequently make comparisons with a Model C that fixes the slope at zero, thus testing the null hypothesis that the predictor is not a useful one, or equivalently that the predictor and *Y* are unrelated to each other. There are occasions, however, when other null hypotheses about the slope are of interest and we illustrated these. Inferences about the intercept are most frequently of interest when the predictor has been centered, thus permitting inferences about the mean of *Y*. Testing null hypotheses about the mean of *Y* will be more powerful in the presence of a predictor when in fact that predictor is a useful predictor; that is, it explains a significant amount of variation in *Y*. This was illustrated by making comparisons with the simplest model comparisons of Chapter 4.

Notes

[1] It is only in the case of a centered predictor in simple regression models that the estimate of the slope will remain constant regardless of whether the intercept is estimated or fixed at various a priori values. Unless predictors are centered around their mean, this will generally not be the case.

[2] Importantly, the estimated slope in such a Model C will differ from the estimated slope in Model A. As we mentioned previously, only when the predictor variable is centered around its mean will the estimated slope remain the same regardless of whether the intercept is estimated or fixed. For all other cases, the slope estimate will vary.