- Nonparametric
- Semiparametric
- Statistical power
- Correlation
- Pearson's r
- Absolute value
- Collinearity
- Construct validity
- Sum of products
- Sum of squares

22 Regression

In this chapter we build on the concept of correlation with ordinary least squares (OLS) regression, even though the latter was invented first.[188] If we think about our scatter plot, we could draw one line through the data that best fits all of our data points. Correlation is a statement about how close the points are to the line. The objective of regression is to determine the best fitting line for the data. Using regression, we can determine the average effect of our independent variable and make predictions about cases outside our sample. We will begin with the simplest regression model, bivariate, where the dependent variable is a function of a single independent variable, before expanding to consider multivariate models with multiple independent variables.

Bivariate Regression

As social scientists we primarily want to explain why variables of interest vary and vary together. Regression allows us the ability to measure the effect of one variable on another. It tells us the effect of an independent variable on a dependent variable. Furthermore, it provides us with the degree of the effect, thereby providing more explanatory leverage than in any other technique we have discussed thus far. Not unlike correlation we can find the strength and direction of association between two variables. Here, however, we can also get at the specific nature of the relationship; i.e., how much variance in the dependent variable is "explained" by the independent variable.

In discussing correlation we implied without much specificity that one could draw a straight line that passed through the set of points in a manner that represented the overall pattern, positive or negative, steep or shallow. In addition to moving to thinking about causal relationships between independent and dependent variables (which was not required for correlation), with regression we also ask: Which linear relationship? In other words, of all the lines in Figure 22.1 that pass through the graph **centroid** – the point where $\overline{X}$ and $\overline{Y}$ intersect and marked in the figure by the intersection of the dotted lines – which fits the data the best?

Before delving into the math, it is useful to graphically illustrate the characteristics of the best fitting regression line, as in Figure 22.2. For any line, we can measure the vertical distance between the line and each observation, which is called a **residual**. Our goal is to minimize these residuals; or more specifically, the sum of squared residuals,

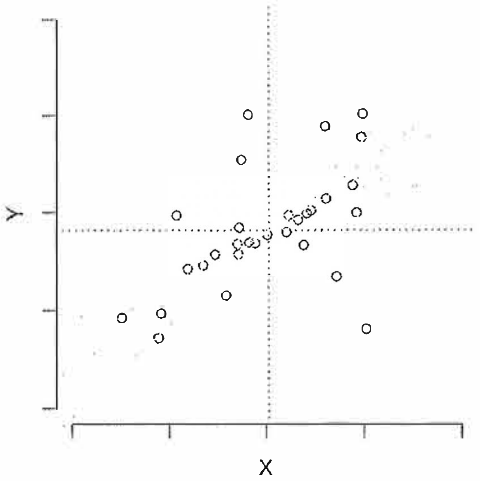

Figure 22.1 Fitting lines through a scatter plot

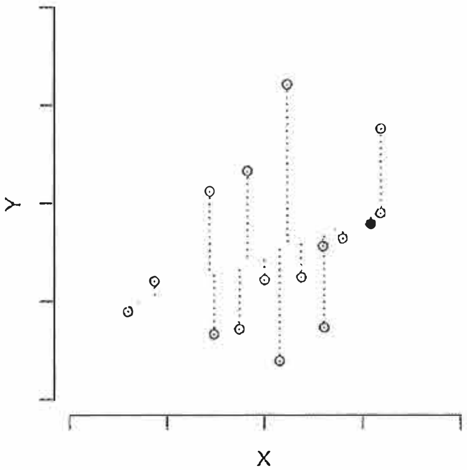

Figure 22.2 Residuals from best fitting line in a scatter plot

because if we did not square them our residuals would sum to zero – as you might recall from the previous chapter. The line which does so is the best fit.

Obviously, choosing possible lines by trial and error and calculating the sum of squared residuals for each line would not be very efficient. Fortunately, determining the correct line can be easily done using some of the values we calculated last time for the correlation coefficient r. We just need the sum of products and the sum of squares. In fact, the general process should be quite familiar. We will begin by summarizing the data into a single equation which states the relationship between X and Y. This equation produces two test statistics, a and b, that describe the relationship. Next we translate them into t-ratios to test the null hypothesis, at which point the rest of the procedure for checking significance is the same as the previous examples with t-ratios (e.g., correlation, difference of means test).

The regression line is based on a simple mathematical equation similar to what you used to draw a line in elementary geometry:

$$Y = a + bX + e. \qquad (22.1)$$

The regression line is a statement about the relationship between X and Y in the population. The equation simply states that the predicted value of Y is the sum of three components: a, which is a constant that applies to each case; bX, which is the product of an average effect b and the specific value of the independent variable X; and e, which is a random component that varies by observation.

Generally, a is referred to as "the constant" and b "the coefficient." In order to determine the best fitting regression line, we calculate a specific numerical value for these terms. The error term, e, however, is not determined. Thus we will not calculate e, but it is important nonetheless from a statistical inference perspective. In the equation, e is merely a symbol to represent the fact that our relationship is probabilistic, not deterministic. For any given case, we would not expect our raw value of Y to equal the predicted value of $\hat{Y}$ (spoken "y-hat") because of this factor. We place a hat over Y for the predicted value to show that it is calculated from the equation and not the same as the Y variable in the data. More colloquially, we can think of the error term as collecting all the junk in the equation. It represents the random error in the stochastic model that makes our predictions less than perfect. However, for a large number of cases and on average, we would still expect the predicted value of $\hat{Y}$. Since we do not calculate the random error term, the actual regression line we calculate looks like:

$$\hat{Y} = a + bX. \qquad (22.2)$$

But, again, this does not mean we can ignore the theoretical importance of e.

To elaborate with the help of Figure 22.3, a geometrically represents the y-intercept, the point where the line crosses the y-axis. Substantively, this question

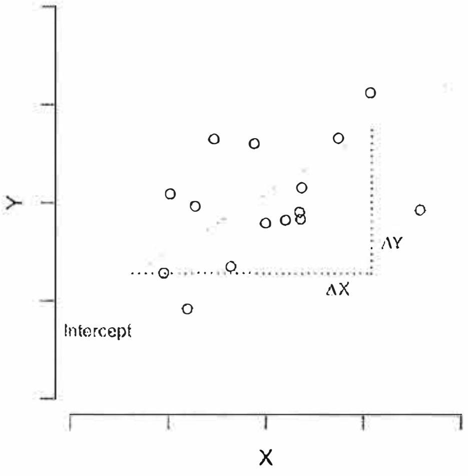

Figure 22.3 Regression line

asks about our expected value for Y when the value of the independent variable is zero ($x = 0$):

$$\begin{aligned}\hat{Y} &= a + b \times 0\\ &= a.\end{aligned} \tag{22.3}$$

The key part of the regression equation is the bX term. The b term tells us the average effect of a unit change of X on Y. More simply stated, it tells us how much Y changes, ΔY, as X changes, ΔX. Thus, if we take our observed value of X and multiply it by our average effect b, and then add this product to the constant a, we would get our predicted value of Y, which is the regression line. Of course, the value of b also has a geometric interpretation: the slope of the regression line. Unlike our correlation coefficient, the slope is not limited in range. The value of the slope can be zero, or any positive or negative value. The one exception, of course, is that the slope cannot be infinite, as this would imply a vertical line.

In order to solve the regression equation we begin by calculating b, which is simply the sum of the products divided by the sum of squares:

$$b = \frac{SP}{SS_x}. \tag{22.4}$$

We can then solve for a:

$$a = \overline{Y} - b\overline{X}. \tag{22.5}$$

With numeric values for both a and b we can rewrite the regression equation in terms of the predicted value of Y, $\hat{Y}$. Again, we use $\hat{Y}$ instead of Y to denote that it is expected or predicted, as opposed to actual or observed. In other words, if you compute the disturbances over several trials (the differences between $\hat{Y}$ and Y) and sum them, you should get 0. At this point we can interpret the relationship between X and Y.

When interpreting a regression equation, the two most important substantive findings relate to the values of a and b. We want to relate these values to our real-world question. Foremost we note the interpretation of b. For every one unit increase in the value of X, we expect a b unit increase in the value of Y, on average. Notice this sounds very much like the mathematical interpretation of a slope, for obvious reasons.

Although we do not determine a specific value for e, we can make a statement about how much of our relationship is systematic (the bX term) and how much is random (e). This too is related to the concepts we discussed with correlation. Correlation told us about the strength of a relationship. The stronger the relationship, the less important the random component is in determining individual values of Y. The weaker the relationship, the more important the random component is in determining the values of Y. Here, when we say random we mean that the explanation is due to something outside of the equation. The e term is picking up any variance we cannot explain with our independent variable. Remember that our regression line is the one that minimizes the squared residuals. However, how small we can actually make that sum depends on the correlation. If the relationship is weak, we can only minimize that sum to a small extent. In fact, we can think about the residuals as the random component itself, for each individual observation.

Regression is flexible. Although hypothesis testing with it requires that we have an interval and normally distributed dependent variable, the independent variable can be of any level of measurement: nominal, ordinal, or interval. The regression line allows us to make a number of statements about the predicted value of Y given values of X. That is, rather than interpreting a single statistic, in regression we modify the value of X in line with our research questions to calculate substantively meaningful predictions of Y. The researcher must ask herself what value of X makes substantive sense, which does, of course, relate to the variable's level of measurement. We might be interested in what the average of an interval level independent variable explains and thus set X to $\overline{X}$. Alternatively, we might be interested in what the lowest value of X tells us about Y by setting X to its minimum value. We might even be interested in knowing what the highest levels of X predict in terms of Y, or anything in between. In sum, we choose a value or a series of values for X that make sense given the question we are asking and solve for $\hat{Y}$.

For one value of X we can make a statement that leverages the information given to us by the constant. When X is at zero, $\hat{Y}$ is a. It is important to note, however, that the interpretation of a depends on zero being a substantively meaningful value for X. For example, we would not think to ask about how many parties we would expect in a country with zero issue dimensions because all countries deal with at least some issues. Thus, though mathematically this is a fair interpretation, it does not always make substantive sense. The constant merely tells us what value of $\hat{Y}$ we should expect given that $X = 0$.

Relatedly, the regression line also allows us to make out-of-sample predictions, or extrapolations, about Y. **Extrapolation** is the process of making predictions about cases outside the range of the X variable in the sample. First, note that we do not take e into consideration when making predictions. The random component is exactly that, so we do not make predictions about it. Besides, this prediction is just an expected value. Just as we do not make statements about the constant that would not make substantive sense, we also exercise caution in making predictions beyond the sample. For example, we would not want to make predictions about countries with negative issue dimensions, which makes no sense though we could extend the regression line into the negative values of X. Nor would we want to predict Y for a hypothetical country with 250 issue dimensions, which is also substantively goofy. In sum, when making predictions, we always want to consider whether we are making extreme or nonsensical counterfactuals.

Interpolation is the process of making predictions about cases within the range of the X variable for the sample but for which no values in the sample exist. For example, if X ranged from 0 to 100 but no units in the sample had values between

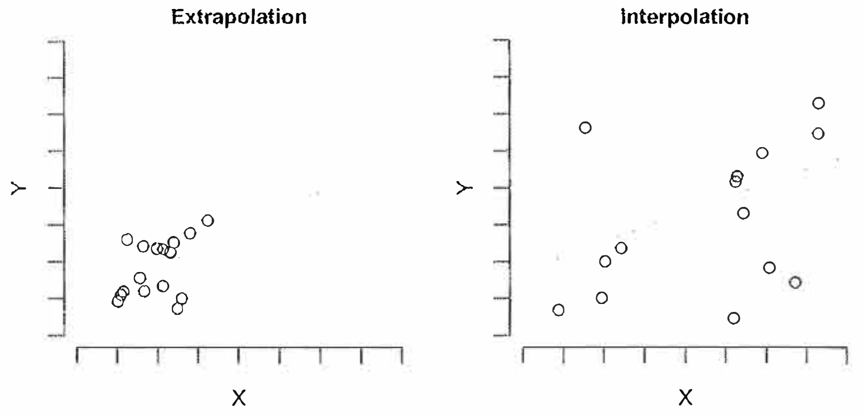

Figure 22.4 Extrapolation and interpolation

30 and 70, interpolation involves making predictions of the value of Y for Xs between 30 and 70. In general, and as should be obvious from Figure 22.4, extrapolation requires more caution than interpolation.

At this point we should be curious about the kind of leverage that regression actually provides. In the context of regression r^2 is referred to as the **coefficient of determination**. In the context of correlation, this does not give us a lot of additional information since r^2 is completely determined by the value of r. In regression, this statistic takes on a more important interpretation, particularly when we have a multivariate regression with more than one independent variable, which we discuss in the next section.

With a strong relationship, the independent variable explains more of the variance in the dependent variable. This means that the fit of the line is better; the sum of residuals is smaller, and the correlation is higher. This is where the r^2 statistic comes in. Squaring r gives us the amount of variation in Y explained by X. Thus the remainder, $1 - r^2$ is random variation, i.e., not explained by the variables in our equation.

The notion of explaining variance can be difficult to grasp. Fortunately, there is a more intuitive way of looking at this information. The r^2 statistic also belongs to a class of statistics called **proportionate reduction in error** statistics (PRE). This interpretation is mathematically equivalent to the variance explained description, but the logic is somewhat different. The power of our regression comes from its ability to make accurate predictions. Our regression allows us to make a specific prediction about the value of Y for a given observation. However, these predictions are not perfect even for our sample; there are residuals.

But, if we did not have information about the X variable, what predictions would we make about Y? The most logical answer is that we would guess the mean of Y. Recall that a central tendency gives us a good summary description of Y. So the question becomes: How much more accurate are our predictions using our regression instead of just guessing the mean of Y every time? The answer is the r^2

value. By knowing X and the resulting regression equation, we are able to reduce the amount of prediction error relative to the mean. Thus the calculated r^2 tells us how much variance in Y is accounted for by our predicted relationship $(a + bX)$.

Earlier we noted that Pearsons r has the following equation:

$$\frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma\left(X-\overline{X}\right)^2\Sigma\left(Y-\overline{Y}\right)^2},} \tag{22.6}$$

which can be rewritten in simpler terms:

$$\frac{SP}{\sqrt{SS_X \times SS_Y}}. \tag{22.7}$$

While we could square r to get the coefficient of determination, r^2, in a regression equation, we can also derive r^2 based on our knowledge of the standard error. R^2 is a ratio of the expected or explained sum of squares to the total sum of squares. The **explained sum of squares** (ESS_Y) is the difference in the predicted and mean values of Y:

$$ESS_Y = \Sigma(\hat{Y} - \overline{Y})^2. \tag{22.8}$$

Recall that the total sum of squares for Y (SS_Y) is

$$SS_Y = \Sigma\left(Y - \overline{Y}\right)^2. \tag{22.9}$$

This ratio gives us the r^2 as well:

$$r^2 = \frac{ESS_Y}{SS_Y}. \tag{22.10}$$

Similarly, we can just as easily calculate r^2 based on the residuals instead of the expected values. We take 1 less the **residual sum of squared errors** (RSS_e) over the total sum of squares. The RSS_e is the sum of the squared differences between the actual and predicted values of Y:

$$RSS_e = \Sigma(Y - \hat{Y})^2. \tag{22.11}$$

Thus r^2 can be thought of in terms of unexplained variance, since the remainder is a simple ratio of unexplained variance in the model's errors to the total variance in the data:

$$r^2 = 1 - \frac{RSS_e}{SS_Y}. \tag{22.12}$$

It should be clear now that the formulae for r^2 are equivalent:

$$r^2 = 1 - \frac{RSS_e}{SS_Y} = \frac{ESS_Y}{SS_Y}. \tag{22.13}$$

Having arrived at an understanding of how to calculate and interpret our regression equation, we next move to testing our hypothesis about the population given the sample data. That is, in practice we also need to check the statistical significance of our results, as we have done in the past (e.g., difference of means and correlation). In particular, we need to make sure that b is relevant in our relationship. But why b?

In regression our null hypotheses refer to the coefficients. The null hypothesis for the constant is that $a = 0$. If we can reject the null, we can be confident that the value of Y is greater than zero, when $X = 0$. Of course, this is likely to be substantively uninteresting.

The null hypothesis for the coefficient, b, however, is the one we typically care about. Thus in regression our null hypothesis predominantly refers to b: $b = 0$. In other words, the null holds that the independent variable has no effect on the dependent variable. Thus we focus on b because it modifies the independent variable, X. Graphically, the null hypothesis predicts a horizontal regression line, i.e., a slope of 0. If we can reject the null hypothesis, we are confident that the relationship we find in the sample between the independent variable and dependent variable exists in the population. Again, this is our major concern in hypothesis testing. The effect could be substantively large or small, but we need to test whether it is likely to exist beyond the sample.

As in the past, in order to test the null hypothesis for statistical significance we need formulae for the t-ratio and the standard error of the coefficient. The standard error of b depends on the ratio between the **mean squared errors** (MSe), which captures the average variance in Y that is unexplained by X, and a product of sample variance and size. In the bivariate case, we calculate it as such:

$$MSe = \frac{\Sigma(Y - \hat{Y})^2}{N - 2}. \tag{22.14}$$

Notice that the numerator of this equation is what we referred to above as the sum of squared errors. Thus we can abbreviate the equation for the mean squared errors,

$$MSe = \frac{RSSe}{N - 2}, \tag{22.15}$$

and calculate the standard error of b accordingly:

$$s_b = \sqrt{\frac{MSe}{s_X^2 \times N - 1}}. \tag{22.16}$$

Again, the test statistic is simply a ratio between the coefficient and its standard error:

$$t = \frac{b}{s_b}. \tag{22.17}$$

After calculating the t-ratio and standard error we proceed as usual by selecting a level of statistical significance (conventionally, $\alpha = 0.05$), noting the degrees of freedom, and checking the t-distribution table (Appendix Table A.2) for the corresponding p-value in order to decide whether or not to reject the null.[189]

Example: Education and Income

We can demonstrate regression with a simple research question: Does education lead to greater income? Perhaps part of the value of an education is that it provides skills that translate into more lucrative job opportunities. We thus hypothesize a positive relationship wherein years of post-secondary education predict annual income. The null hypothesis is that education is not related to income. We are going to use regression to solve for the predicted relationship:

$$Income = a + b \times Education + e. \tag{22.18}$$

Assume that Table 22.1 contains data from a simple random sample survey of ten adults' levels of income and education. The second and third columns contain our collected data and we use the subsequent columns to calculate our statistics. We begin by calculating the necessary statistics for use in the formulae. In the data, the mean of income, $\overline{Y} = 38.6$ and the mean of education is $\overline{X} = 2.8$. The variance for income is 142.93 and for education it is 3.29. The process for calculating the sum of squares for income, 1286.4, should be familiar by now.

Table 22.1 Calculating regression for income and education

	Income in	P – S			$(X - \overline{X})\times$		
Respondent	Thousands	Education	$Y - \overline{Y}$	$X - \overline{X}$	$(Y - \overline{Y})$	$(Y - \overline{Y})^2$	$(X - \overline{X})^2$
1	44	4	5.4	1.2	6.48	29.16	1.44
2	30	3	–8.6	0.2	–1.72	73.96	0.04
3	51	2	12.4	–0.8	–9.92	153.76	0.64
4	40	4	1.4	1.2	1.68	1.96	1.44
5	14	0	–24.6	–2.8	68.88	605.16	7.84
6	44	5	5.4	2.2	11.88	29.16	4.84
7	34	2	–4.6	–0.8	3.68	21.16	0.64
8	56	5	17.4	2.2	38.28	302.76	4.84
9	42	3	3.4	0.2	0.68	11.56	0.04
10	31	0	–7.6	–2.8	21.28	57.76	7.84
Sum	$\Sigma Y = 386$	$\Sigma X = 28$					
Sum of products					$SP = 141.2$		
Sum of squares						$SSy = 1286.4$	$SSx = 29.6$

Recall that our covariance can be abbreviated as $\frac{SP}{N}$, where

$$\begin{aligned} SP &= \Sigma(X - \overline{X}) \times (Y - \overline{Y}) \\ &= 141.2. \end{aligned} \tag{22.19}$$

With these statistics, calculating the slope of the line is trivial:

$$\begin{aligned} b &= \frac{SP}{SS_X} \\ &= \frac{141.2}{29.6} \\ &= 4.77. \end{aligned} \tag{22.20}$$

To situate the line on the x-axis we calculate the constant:

$$\begin{aligned} a &= \overline{Y} - b\overline{X} \\ &= 38.6 - 4.77 \times 2.8 \\ &= 25.24. \end{aligned} \tag{22.21}$$

At this point we can interpret our regression results substantively. For instance, we might be interested in how often those of average income participate. To that end, we calculate our predicted value of Y at the mean of X:

$$\begin{aligned} \hat{Y} &= a + b\overline{X} \\ &= 25.24 + 4.77 \times 2.8 \\ &= 38.6. \end{aligned} \tag{22.22}$$

Thus we expect an individual with the average amount of education to earn about $38,600. What would we expect for an individual with the highest amount of education in our sample to make?

With the regression parameters in hand we move to testing statistical significance. We will continue with the tabular format to illustrate the calculation of the sum of squared errors in Table 22.2. Our first step is to calculate the $\hat{Y}$ for each of the values of X. This allows us to draw our regression line, as in Figure 22.5.

We begin by finding the mean sum of squares:

$$\begin{aligned} MSe &= \frac{RSSe}{N - 2} \\ &= \frac{\Sigma(Y - \hat{Y})^2}{(N - 2)} \\ &= \frac{612.84}{8} \\ &= 76.61. \end{aligned} \tag{22.23}$$

Table 22.2 Calculating the sum of squared errors

Income in thousands	P – S Education	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
44	4	44.32	−0.32	0.10
30	3	39.55	−9.55	91.20
51	2	34.78	16.22	263.09
40	4	44.32	−4.32	18.66
14	0	25.24	−11.24	126.34
44	5	49.09	−5.09	25.91
34	2	34.78	−0.78	0.61
56	5	49.09	6.91	47.75
42	3	39.55	2.45	6.00
31	0	25.24	5.76	33.18
Sum of squares				*RSSe* = 612.84

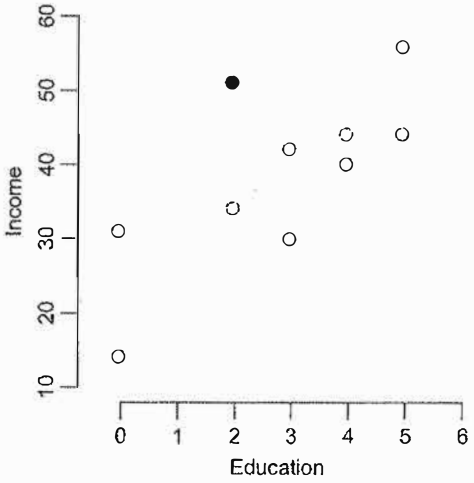

Figure 22.5 Plot regression line for education and income

With the mean sum of squares we can calculate the standard error of b:

$$\begin{aligned} s_b &= \sqrt{\frac{MSe}{s_X^2 \times N - 1}} \\ &= \sqrt{\frac{76.61}{3.29 \times 9}} \\ &= 1.61. \end{aligned} \tag{22.24}$$

Finally, we can calcuate the t-ratio,

$$\begin{aligned} t &= \frac{b}{s_b} \\ &= \frac{4.77}{1.61} \\ &= 2.96. \end{aligned} \tag{22.25}$$

We now check the t-ratio against the t critical value in Appendix Table A.2 for the appropriate degrees of freedom and confidence level. We find that we can reject the null hypothesis of no relationship between income and participation.

In addition, consider what this relationship means for education more generally. To what extent does education predict income? To answer this question we can calculate the coefficient of determination:

$$\begin{aligned} r^2 &= 1 - \frac{RSS_e}{SS_y} \\ &= 1 - \frac{\Sigma(Y - \hat{Y})^2}{\Sigma(Y - \overline{Y})^2} \\ &= 1 - \frac{612.84}{1286.4} \\ &= 1 - 0.48 \\ &= 0.52. \end{aligned} \tag{22.26}$$

In this example, the postulated relationship explains 52% of variation in Y. In other words, 48% ($1 - r^2$) of the variation in income is unexplained by education.

Note that we can do some rearranging of our formulae to arrive at the same answers. We begin by calculating the sample **standard error of the estimate**, SEe (also written as s_e). This can be interpreted much like a typical standard deviation, but this time in terms of the regression line. Given normally distributed errors about 68% of the observations will fall within one standard error of the line, 95% within two and and over 99% within three.

$$\begin{aligned} SEe &= \sqrt{\frac{RSSe}{N - K - 1}} \\ &= \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{(N - 2)}} \\ &= \sqrt{\frac{612.84}{8}} \\ &= 8.75. \end{aligned} \tag{22.27}$$

Then the standard error of b can be calculated with either the SEe or the MSE:

$$\begin{aligned} s_b &= \frac{SEe}{\sqrt{SSx}} \\ &= \frac{8.75}{\sqrt{29.6}} \\ &= 1.61 \\ s_b &= \sqrt{\frac{MSe}{SSx}} \\ &= \sqrt{\frac{76.61}{29.6}} \\ &= 1.61. \end{aligned} \tag{22.28}$$

Similarly, here is the way to calculate r^2 from Pearson's r:

$$\begin{aligned} r &= \frac{\frac{SP}{N - 1}}{s_x s_y} \\ &= \frac{\frac{141.2}{9}}{1.81 \times 11.96} \\ &= 0.72 \\ r^2 &= r \times r \\ &= 0.72^2 \\ &= 0.52. \end{aligned} \tag{22.29}$$

In order to present our regression results we collect these statistics into an easy-to-read table. Typically regression tables list the names of the independent variable(s) in rows of the first column with their corresponding coefficients and standard errors in one or more subsequent columns headed by the name of the dependent variable. Standard errors are presented in parentheses to distinguish them from the coefficients. Oftentimes the coefficients are followed by stars that serve as a visual heuristic denoting that the standard error is small enough to reject the corresponding null hypothesis. In Table 22.3 we have done exactly this. The star, as indicated by the note at the table, tells us what we discovered above: that the ratio of the coefficient on the education variable, 4.77, to its standard error is larger than the critical value for t at the 95% confidence level. Thus it is clear that we reject the null hypothesis of no relationship between education and income.

Multivariate Regression

In this section we provide an introduction to the multivariate version of the ordinary least squares regression model introduced above. We focus on describing

Table 22.3 Explaining income with education

	Dependent variable:
	Income
Education	4.77* (2.13)
Constant	25.24* (5.29)
N	10
R^2	0.52

Note: *p<0.05

the major intuition behind multivariate models and the interpretation of the statistics in this context, while avoiding the matrix algebra necessary to calculate the statistics in this section by hand. In addition, we note that the regression model depends on several assumptions, which can be easily violated – and often are. Therefore we also provide a discussion of the most frequent assumption violations and their potential effects on model-based inferences.

The purpose of multivariate regression is the same as in the bivariate case. We would like to estimate the effect of an independent variable on a continuous dependent variable. However, in the multivariate case we also want to **control** for and estimate the effects of other independent variables. Control in multivariate statistics is a process of parsing out the specific effect of each of two or more independent variables on a dependent variable. That is, with multivariate analysis we can "hold constant" the effects of one or more independent variables in order to get a more precise estimate of the effect of another independent variable.

Controlling for multiple independent variables is important when the researcher believes there to be a **confounding** relationship. Recall our discussion of confounders in Chapter 6. We think of a confounding variable as one that is correlated with both the independent and dependent variable. In terms of causality – a topic we return to in Chapter 23 – the concern is that change in the confounder leads to change in both the independent and dependent variables, which is not perceived by the researcher when the confounder is excluded from the model. Indeed the researcher may incorrectly infer that the independent and dependent variable are related, when in truth it is the omitted confounder that causes the variables in the model to appear to correlate. In contrast, when we include the confounder in the model with the other independent variable, we can estimate the specific effect of each independent variable on the dependent variable, thereby ensuring that our estimates are not the result of the previously omitted variable. The bias introduced from omitting a relevant variable in the linear regression model is eponymously called **omitted variable bias** and is a violation of one of the major assumptions of the regression model, which we further discuss below.

To elaborate on how multivariate regression can help us control for confounders, consider two independent variables, X_1 and X_2, and a dependent variable, Y. If X_2 is related to both X_1 and Y, in a simple bivariate model of

$$Y = a + bX_1 + e, \tag{22.30}$$

some portion of the explained variance in Y attributed to X_1 may be due to the potential confounder X_2. In regression we control for X_2 by including it in the model, which now has an additional coefficient as well:

$$Y = a + b_1X_1 + b_2X_2 + e. \tag{22.31}$$

Regression will estimate partial slopes for each independent variable. Thus the estimate of b_1 is the average change in Y for each unit change in X_1, controlling for X_2. And, similarly, the estimate of b_2 is the average change in Y for each unit change in X_2, controlling for X_1. We can expand the regression equation to include k independent variables:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e. \tag{22.32}$$

Accordingly, the substantive interpretation of multivariate regression coefficients is akin to the bivariate model with the slight addition of the control, or "holding constant," terminology. That is, we would again describe the relationship between each independent variable and the dependent variable in terms of the respective slope coefficient by calculating $\hat{Y}$ for a particular value of, say, X_1, while holding X_2 constant, or at a set value. In a multivariate context it is particularly useful to look at the change in $\hat{Y}$ as a result of changing X_1 from one substantively meaningful value to another, perhaps a full unit or a standard deviation increase or decrease. Importantly, when we do so we hold X_2 constant, as well as any other independent variables in the model, to convey the specific effect of a change in one variable, X_1 in this case, on an expected change in the dependent variable, while holding constant the other independent variables in the model.

In multivariate regression the interpretation of the coefficient of determination, r^2, takes on a new meaning as well. Specifically, the proportionate reduction in error interpretation expands such that r^2 should still be thought of in terms of explained variance, but now for all independent variables in the model. The remainder, the variance in Y which is not explained by either X_1 or X_2, is the ratio of unexplained variance in the errors to the total variance; i.e., the variance in Y unexplained by all the independent variables in the model.

Assumptions

Like our other models, the linear regression model depends on a series of assumptions about the data-generating process that are required in order for us to arrive

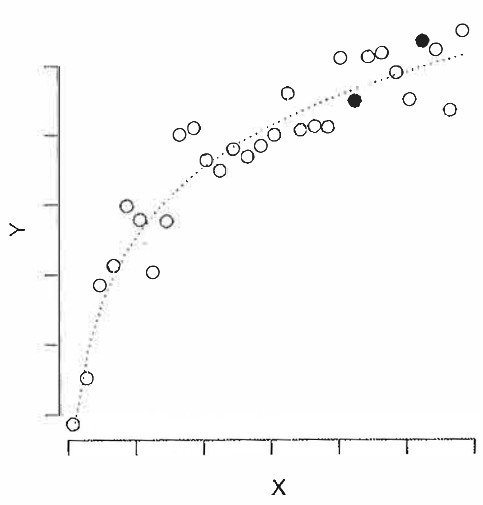

Figure 22.6 Nonlinear relationship

at good estimates. These are generally called the **Gauss-Markov assumptions**. While a full description of these is beyond the basic treatment we offer in this book, we focus on the general intuition behind these assumptions and note the most common violations that can lead to problems with inference.

As should be clear by now, regression models expect that the dependent variable is a linear function of a specific set of independent variables, plus the error term. We violate the **linearity and additivity specification assumption** when we misspecify the relationships in the model. We typically do so when we try to model a relationship that is nonlinear or include the wrong set of variables in our model. We discuss each of these in turn.

In the first case, modeling a relationship that is nonlinear (Figure 22.6), recall that the regression equation solves for a line. If the expected functional form of the relationship is not linear, the regression estimates will not properly capture the relationship. We assume that the change in Y associated with a unit increase in X_1, holding all other variables constant, is the same across all values of X_1. That is, the effect of a unit increase in X_1 does not depend on the value of X_1. Violations of this nature are typically dealt with in the regression context by transforming the data. One or more variables of a nonlinear function can be mathematically transformed (e.g., taking the log or a quadratic function) so as to create a linear relationship. One can either transform the independent variables or the entire equation via the dependent variable. To make these decisions one often relies on previous work and strong theory about the expected relationships.

Furthermore, because it is an additive model we are also assuming that the change in Y from a unit X_1 is constant regardless of the values of the other independent variables in the model. We can therefore state the relationship between X_1 and Y in terms of the average expected change while holding the other variables constant.

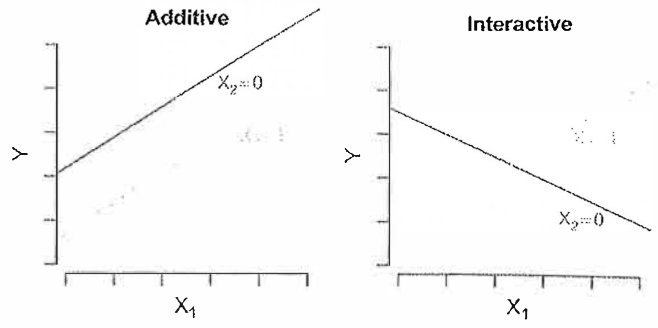

Figure 22.7 Additive and interactive relationships

Frequently in social science we would like to consider cases where the relationship between X_1 and Y is not constant, but instead depends on a third variable, say, X_2. These conditional relationships are called **interactions** and can be modeled in regression. Consider first what it would mean to have a relationship between X_1 and Y that differs depending on X_2. Most basically, we would expect different slopes for the effect of X_1 on Y depending on the value of X_2. If so, the standard posited linear and additive model is an incorrect depiction of the relationship.

For an interaction we adjust the linear regression model by making the relationship between X_1 conditional on X_2 via a multiplicative term: $X_1 \times X_2$. When we rewrite the regression line we maintain the direct, or lower order, effects and solve for the coefficient of the conditional relationship between X_1 and X_2:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_1X_2 + e. \qquad (22.33)$$

Figure 22.7 shows two hypothetical graphs of relationships. The left graph shows what we should expect from the standard regression model with two lines, distinguished by their different values of X_2, of similar slopes suggesting an additive relationship. The right graph shows a clearly non-additive relationship of the effect of X_1 on Y. Conditional on X_2 the slope of the relationship between X_1 and Y differs. In this example, those with a value of 1 for X_2 have a positive relationship and those with a value of 0 have a negative relationship.

In the second case, omission of a relevant variable, the remaining coefficients will be biased. Omitting a relevant variable can either raise or lower an estimator's mean squared error, depending on the relative size of the variance reduction and the bias. Of secondary concern, the estimate of the variance will be biased upward. To understand why, recall from above our discussion of omitted variable bias and the problem of attributing explained variance when there are potential confounders.

The idea of explained variance in the multivariate context is often aided by Venn diagrams. These diagrams convey shared relationships between variables as represented by overlapping circles. Figure 22.8, for example, shows three variables, each explaining some variance of the other. In regression, for example, we would like to estimate the partial slope for the relationship between X_1 and Y. In doing so we would like to hold constant the effect of X_2 on Y. What the diagram

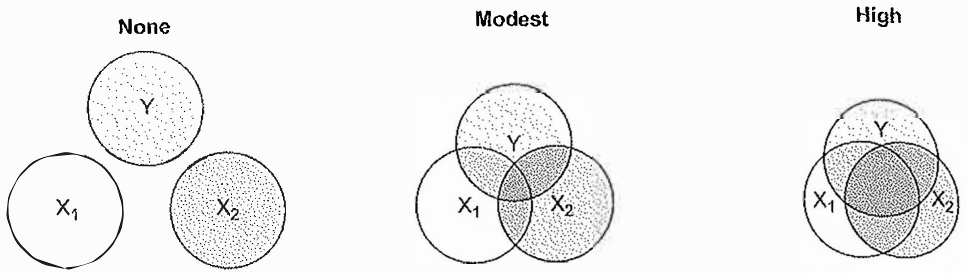

Figure 22.8 Venn diagrams by degree of collinearity

rightly suggests about multivariate regression is that in order to control for the effect of X_2 the slope coefficient for X_1 must not include the variance in Y that is explained by both X_1 and X_2, the darkest shading area in the diagram. That is, the partial slope coefficient for X_1 is the effect we get after removing the variance that is explained by both X_1 and X_2. Likewise, the partial slope coefficient for X_2 is the effect we get after removing the variance that is explained by both X_1 and X_2. Thus by including the confounder in the model we get an estimate of the effect of each variable on the dependent variable that is not tainted by the effects of the other.

We also note that the inclusion of an irrelevant variable is not helpful either. While the estimate of the explained variance remains unbiased in this case, the estimate will not be efficient. That is, we unnecessarily lose degrees of freedom as the means squared error is raised.

The second assumption we make about regression is that the **errors have an expected value of zero**. This means that on average and with enough data we get it right and the errors balance out. Related to discussion of sampling from the population, our model assumes that our estimates of the relationship will not be perfect, but that our errors will be randomly distributed around a mean value, which is correct, and thus zero. That is, to the extent we are off it is due to random error.

Our errors should also have the same variance and be uncorrelated with each other. The former assumption is referred to as **homoscedasticity**, which basically holds that the variance of each error term for each unit of observation is the same for each independent variable. Furthermore, it assumes that knowing something about the disturbance term from one observation tells us nothing about the disturbance term for another observation. Violations here are referred to as **heteroscedastic errors**, where the disturbances do not all have the same variance. The latter assumption of uncorrelated errors is referred to as **independence**, which is especially common when we have time-series or longitudinal data. Here, repeat observations create correlation between consecutive errors, which are often called **autocorrelated errors**, since the disturbances are correlated with one another.

We further assume that the independent variables are non-random and have finite variances. That is, we require that our measures of our independent variables are fixed, or reliable, such that if we were to repeat the data-gathering process from the same sample we would arrive at the same values on the independent variables.

Violations of this assumption can occur when: measuring the independent variable; autoregressing, or using a lagged value of the dependent variable as an independent variable; and in simultaneous equation estimation, or situations in which the dependent variables are determined by the simultaneous interaction of several relationships.[19]

Even more basic then the Gauss-Markov assumptions, it is also important to note that the mathematics behind the regression equation depends on a couple of key properties of the data. Foremost, the data must be **full rank matrix** for the regression estimator to work. This requires that we have at least as many if not more observations than we have independent variables. In solving equations we must have more knowns than unknowns or it is mechanically impossible to compute the estimates. In addition, there can be no exact linear relationships between the variables. These violations lead to **multicollinearity** problems, where the variables are so strongly correlated that it becomes difficult to parse out the partial effects of each.

In the presence of multicollinearity our estimates are still good; however, the variances of the estimates are quite large and unstable leading to potentially invalid predictions from particular independent variables, even though the full extent of prediction across all the variables will be correct. This is because there is not enough independent variation in a variable to precisely estimate its impact. Consider Figure 22.8. As the shared explained variance increases, it becomes more difficult to parse out the independent effects of each variable. Keep in mind that this is because the model only calculates the partial slopes. Typically, the best remedy for multicollinearity is to collect more data.

In sum, if the Gauss-Markov assumptions are met then the estimates we retrieve from our ordinary least square models are good. But what do we mean by *good*? If we change these assumptions (or they do not hold) then the regression estimator may no longer be optimal; indeed in almost all cases it will not be optimal if one of the assumptions is violated and we would want to choose a different kind of model. If it holds, then the estimator can be shown to be the **best estimator** among all the **unbiased** estimators, or all those that tell us how close our estimate is to the true parameter value (if it can be known). Usually there are several unbiased estimators and in choosing between them we like the estimator that has a sampling distribution with the smallest variance, i.e., the one that is the most **efficient**. Thus meeting the assumptions leads to the **best linear unbiased estimator** (BLUE), or an estimator that is linear and unbiased and has the minimum variance among all the linear unbiased estimators.

Example: Turnout by Region and Income

Consider the relationship between voter turnout rates in different states and the region of the country. It is not unreasonable to test this relationship as different regions of the country have had different historical experiences with the electoral system. We might suspect that those in the North, for example, with their longer

Table 22.4 Voter turnout by South and non-South regions

	Dependent variable:
	Turnout
Region: South	−12:89* (1.63)
Constant	57.23* (0.92)
N	50
R^2	0.57

Note: *p<0.05

history of enfranchisement, to turn out to vote at greater rates. Thus our hypothesis is that state turnout is a function of region of the country (South vs. non-South). We can measure turnout simply as the percentage of those who showed up at the polls out of all those who are eligible to vote.[191] South is a dummy variable where *southern* = 1 and *non-southern* = 0. The null hypothesis is that there is no relationship between region and turnout.

Table 22.4 provides the results (from hypothetical data) of our bivariate regression model. Given what we have discussed in the sections above, we should be able to interpret every statistic in the model. Most importantly, we would look to see if we can reject the null hypothesis for our question about the relationship between region and turnout. Looking at the row for the South variable, we can interpret the average difference between two kinds of states on turnout. Notice that we do not include both the South variable as well as a non-South variable in the model. This is because the values of the singular dummy variable already represent both. We need a baseline for comparison when each dummy variable is set at 1 in order to substantively interpret the line. Thus when dummying out a single variable into its composite categories we always include $k - 1$ categories, and we interpret the coefficient as moving from baseline category to particular dummy category in the model, holding all else constant. Comparing a state in the non-South to the South (a one unit change from 0 to 1) we see that the average turnout drops by nearly 13% ($b = -12.89$) for the South. This result is statistically significant at $p < 0.05$ with a standard error of 1.63.

Of course, careful observers of modern political behavior might raise a question about the regional diversity of the United States. Perhaps the differences across it are greater than just the historical geographic divide. That is, the South non-South dichotomy may capture more than just the South's history of disenfranchising blacks. The reason for today's lower turnout in the South may be due to any number of characteristics prevalent in the South other than its history of disenfranchisement. For example, we know there is greater poverty in the South. If poverty is likely to be associated with both the independent variable, South, and the dependent variable, turnout, how can we be sure then that poverty is not driving the low turnout rates instead of the history of disenfranchisement? To use a term we introduced above: is it possible that poverty is confounding the relationship we found in the bivariate model above?

Table 22.5 Voter turnout by region and income

	Dependent variable:
	Turnout
Region: South	−10.12* (1.68)
Income	0.004* (0.001)
Constant	36.66* (6.02)
N	50
R^2	0.65

Note: *p<0.05

Multivariate regression allows us to test the effect of both our original regional variable as well as the new income variable (per capita income) in the same model. By including them both as additive terms we can look at the effect of each on turnout while holding the other constant:

$$Y = a + b(South) + b(Income) + e. \tag{22.34}$$

In Table 22.5 we present the results of the multivariate regression model. Interpreting the constant we note that states in the North with a zero per capita income have an average turnout rate of only 37%. Of course no state has a zero average income so such an interpretation tells us substantively little. As before, we see a negative sign on the South coefficient. We interpret the slope of the line here to mean that in comparison to the North the southern states have about 10% lower turnout on average, holding income constant. On the contrary, the income coefficient tells us that there is a positive relationship between a state's per capita income and turnout. For a one unit increase in income we should expect a 0.004% increase in turnout. Of course, that is a very small increase given that the variable ranges several thousand points over the 50 states. We can make use of the standard deviation to provide a more substantively informative interpretation. Given a standard deviation of 614.47, we can say that a one standard deviation increase in income leads to a 2.5% increase in turnout. For both variables the relationships are significant, given the relatively small size of the standard errors. Finally, the r^2 indicates that 65% of the variance in turnout is explained by the independent variables. Note that including the income variable moves the explained variance up 15% from the bivariate model.

CONCLUSIONS

Regression is a frequently utilized tool for understanding the relationship between one or more independent variables of any level of measurement on a continuous dependent variable. It provides more explanatory leverage than any other technique we have discussed thus far by estimating the average effect of a unit change in the independent variable on the dependent variable. However, like all parametric models, regression makes various assumptions that are easily violated in practice. As such, a careful employer of regression will look to diagnose how well their model meets the assumptions before trusting their results.

KEY TERMS

- Centroid
- Residual
- Extrapolation
- Interpolation
- Coefficient of determination
- Proportionate reduction in error
- Explained sum of squares
- Residual sum of squared errors
- Mean squared errors
- Standard error of the estimate
- Control
- Confounding
- Omitted variable bias
- Gauss-Markov assumptions
- Linearity and additivity specification
- Interactions
- Errors have an expected value of zero
- Homoscedasticity
- Heteroscedastic errors
- Independence
- Autocorrelated errors
- Full rank matrix
- Multicollinearity
- Best estimator
- Unbiased estimators
- Efficient
- BLUE

23 Causal Inference

Much of social science research is concerned with causal relationships. In this chapter we explore the general framework for making causal statements using different methodological approaches. With experiments as our point of reference, we revisit regression in the context of a causal treatment variable, and then introduce a technique to evaluate the causal effect of a treatment with observational data by matching treated and control units. Before doing so we lay out the specific assumptions necessary for causality and the different motivating factors for each causal model.

Assumptions and Assignment

Why causal inference? Our statistical objective thus far has been to infer associations among variables. From these associations we can estimate probabilities of events with the statistical methods introduced above, provided that the external conditions remain the same. This allows us to answer important questions, like: What is the mean number of parties in democracies? Are turnout and geographic regions related? Are greater issue dimensions associated with greater numbers of parties? On the contrary, we need causal inference when we would like to infer probabilities under different conditions. That is, when we would like to know what would happen if something else happened. When we expect probabilities to change in response to external factors we must rely on causal analysis. Thus, causal questions ask somewhat different questions: What are the effects of worker-training programs? Does smoking cause cancer? Does viewing a campaign advertisement change vote preferences? In each case we are asking a question that posits different probabilities under different conditions: attending versus not attending a worker-training program; smoking versus not smoking; viewing versus not viewing a campaign ad.

Throughout the statistics section of this book we have identified hypothesis tests that are based on covariance between suspected cause and effect. However, the tests themselves are only covariational; that is, they are not explicitly causal. As we discussed in Chapters 21 and 22, in order to make claims of causality we need more than the evidence of covariation between cause and effect that we can attain from these statistical methods. Minimally, we also need the cause to precede the effect and to be able to eliminate plausible alternative causes. Thus, causal