Chapter 9
Correlation and Simple Regression

The methods we have discussed so far allow us to examine whether means of a numeric variable (interval/ratio) vary across two or more groups measured at the nominal level and whether two or more nominal, or possibly ordinal variables, are associated. However, we often want to determine whether two continuous variables are related. In these cases, none of the methods we have covered is appropriate. Instead, we may turn to two alternate methods: the Pearson correlation coefficient and the simple linear regression model. These methods form the basis for the more widely used multiple regression model, which we will discuss in the next chapter.

9.1 Measuring Linear Association

9.1.1 The Covariance

Consider the hypothetical variables x and y shown in the scatterplots in Fig. 9.1. In the upper left plot in the figure, x and y are clearly very strongly related to each other: larger values of x correspond to larger values of y (and vice versa). In the upper right plot, the two variables do not appear to be patterned. In the lower left plot, x and y are very strongly related, but in the opposite direction compared to the upper left plot. Here, larger values of x correspond with smaller values of y. Finally, in the lower right plot x and y appear to be positively related, but not as strongly as in the first plot.

How can we measure the extent of the linear association between two variables like those in each plot? The covariance is a first measure of the strength of linear association between two continuous variables. It is computed as:

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}. \tag{9.1}$$

S.M. Lynch, *Using Statistics in Social Research: A Concise Approach*,

DOI 10.1007/978-1-4614-8573-5_9, © Springer Science+Business Media New York 2013

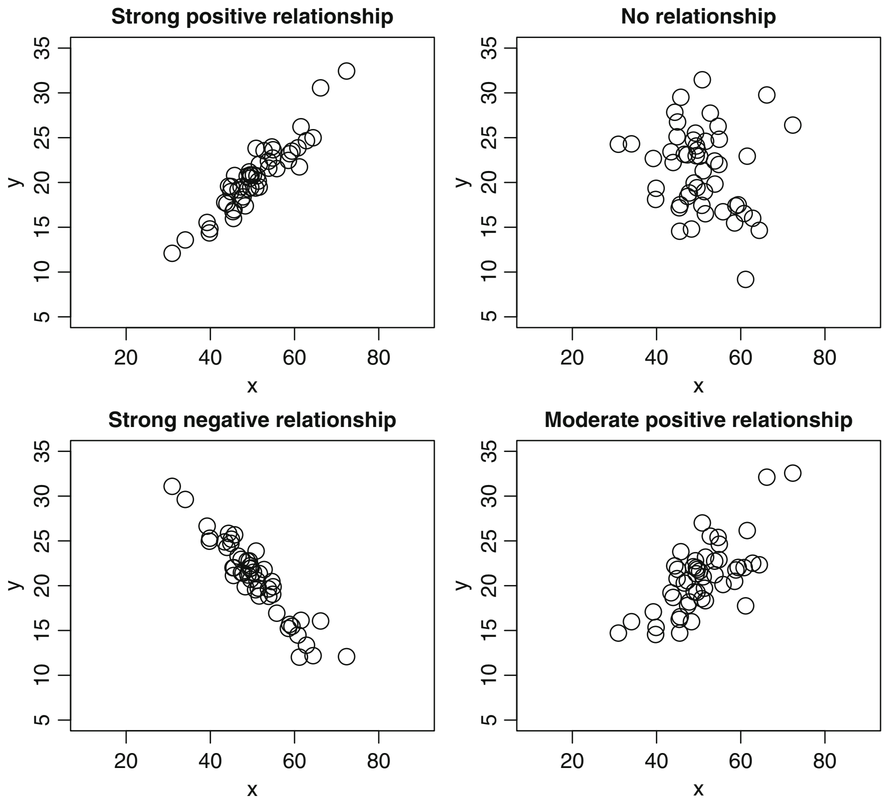

Fig. 9.1. Scatterplot of hypothetical variables x and y.

This equation looks like the variance formula from Chap. 4, but it contains information on two variables (x and y), not just one. Specifically, whereas the numerator of the variance formula involves summing the squared deviations of the sample values of a variable from its mean, the numerator of the covariance requires summing the product of x deviations from its mean with y deviations from its mean. In both the variance and covariance calculations, the numerator is then "averaged" by dividing by $n - 1$.

If two variables, x and y, are linearly related, then the covariance will be large, either positively or negatively. Why? Consider the relationship between x and y again with some of the components of the covariance calculation superimposed in the plot (see Fig. 9.2). If x and y are linearly associated, then whenever x is far from its mean (i.e., $x - \bar{x}$ is large), y should also be far from its mean (i.e., $y - \bar{y}$ is large). The product of two large numbers is large itself, and the sum of a set of large numbers will be large, so the covariance will be large. If x and y are not linearly related, then when $x - \bar{x}$ is large, $y - \bar{y}$ will, on average, be close to 0, and when $y - \bar{y}$ is large, $x - \bar{x}$ will, on average, be close to 0. The product of large numbers with numbers close to 0 is small, so is their sum, and so the covariance will be small.

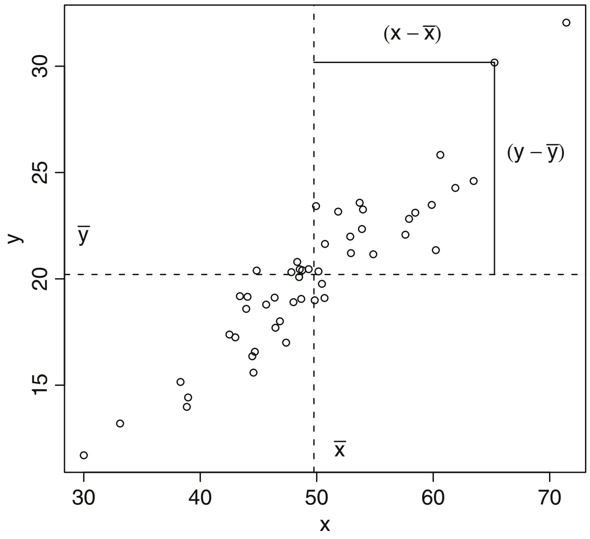

Fig. 9.2. Scatterplot of hypothetical variables x and y with deviations from means superimposed for a single observation.

9.1.2 The Pearson Correlation

A key limitation of using the covariance to assess the strength of the relationship between two variables is that its scale is a function of the scale of the variables used to compute it. If you change the scale, the size of the covariance will change. The correlation (denoted r) corrects for this scale problem by dividing the covariance by the standard deviation of each variable:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{sd(x)sd(y)}. \tag{9.2}$$

Performing this division constrains the correlation to be between -1 and $+1$. Thus, the correlation represents the expected standard unit change in one variable for a standard unit change in the other variable. In this metric, a perfect positive association is represented by a correlation of +1, a perfect negative association is represented by a correlation of −1, and a 0 represents no linear association at all.

Figure 9.3 is a scatterplot of education and family income from the 2004 GSS (unmarried persons and persons with 0 income have been excluded). Education and income appear to be related linearly, but the scale of the variables makes it difficult to discern the strength of the relationship.

Figure 9.4 shows the same information, but after the two variables have been standardized. The figure contains a 45-degree line; if the data were to fall along this line, the correlation between education and income would be 1. Instead, the

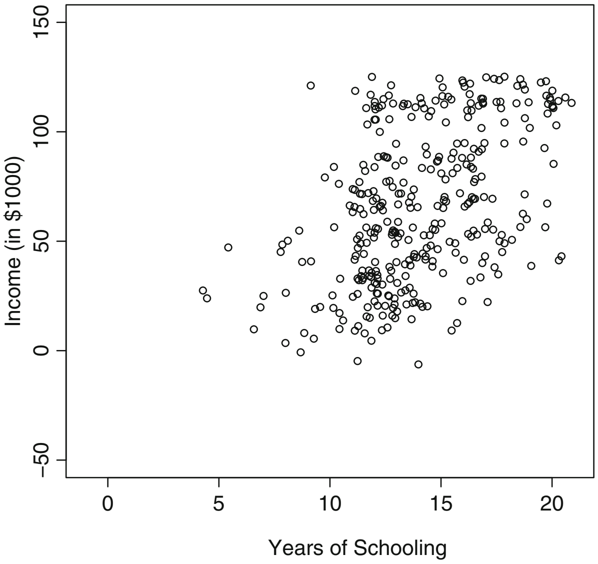

Fig. 9.3. Scatterplot of education and income (2004 GSS Data).

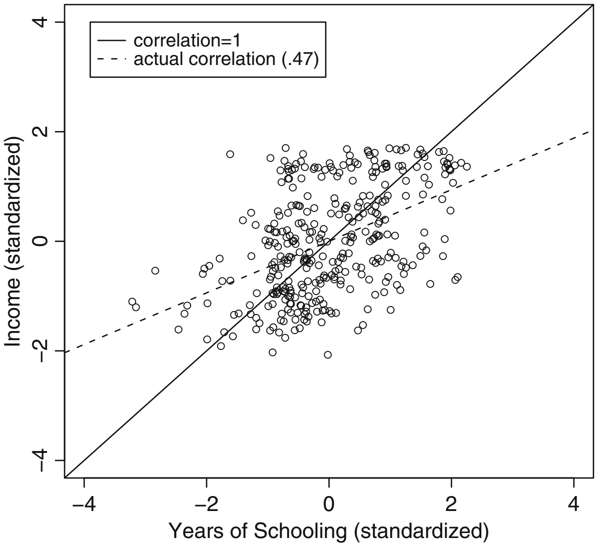

Fig. 9.4. Scatterplot of standardized education and income with reference lines

correlation between education and income is .47, which is considered to be a moderately strong linear relationship between the two variables. In general, in social science research a correlation between .1 and .3 is considered a weak association, a correlation between .3 and .6 is considered to be a moderate association, and a correlation above .6 is considered to be a strong association.

9.1.3 Confidence Intervals for the Correlation

In previous chapters, we have seen that sampling variability produces sample means ($\bar{x}$) that differ from the true population mean (μ). Sampling variability similarly produces estimates of the correlation (r) that differ from the true value in the population (ρ). In previous chapters, we constructed confidence intervals around our sample estimate in order to make inference about a reasonable value for μ. We can construct a similar interval estimate for the correlation.

However, the correlation is bounded by -1 and 1, and even though its sampling distribution becomes more normal in appearance as sample sizes increase, its distribution will remain bounded and slightly skewed so long as the correlation is not exactly 0. Ronald Fisher, a prolific and influential statistician from the early twentieth century, proposed a transformation of the correlation that has an approximately normal distribution and allows for construction of reasonable confidence intervals. The process of interval construction involves the following steps:

1. Transform r into z_f using: $z_f = .5\,[\ln(1+r) - \ln(1-r)]$ (z_f is "Fisher's z")
2. Compute the standard error of z_f: $\hat{\sigma}_{z_f} = \frac{1}{\sqrt{n-3}}$
3. Compute the interval in the z_f metric as:

$$z_f \pm z_{\alpha/2}\hat{\sigma}_{z_f}$$

4. Convert the lower and upper bounds (b) back to the correlation scale using:

$$b = \frac{e^{2z}-1}{e^{2z}+1} \tag{9.3}$$

How does this transformation work to create a more normally distributed sampling distribution for the correlation? Consider the calculation. When r is 0, the transformation is:

$$z_f = .5\,[\ln(1+r) - \ln(1-r)] \tag{9.4}$$

$$= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{9.5}$$

$$(1/2)\ln(1) \tag{9.6}$$

$$= 0 \tag{9.7}$$

When r is 1, the transformation is:

$$z_f = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{9.8}$$

$$= (1/2)\ln(2/0) \tag{9.9}$$

$$\approx \infty \tag{9.10}$$

When r is -1, the transformation is:

$$z_f = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{9.11}$$

$$= (1/2)\ln(0/2) \tag{9.12}$$

$$\approx -\infty \tag{9.13}$$

These latter two sets of equations are approximate, because both division by 0 and the log of 0 are undefined. However, as r *approaches* these values, z_f approaches positive or negative infinity. In other words, Fisher's z transformation stretches the correlation's bounds from ± 1 over the real line.

In the education and income data shown in the figures, the correlation was .47, and the sample size was 342. Fisher's z, then, is:

$$z_f = .5\,[\ln(1+r) - \ln(1-r)] \tag{9.14}$$

$$= .5\,[\ln(1.47) - \ln(.53)] \tag{9.15}$$

$$= .5(.385 - (-.635)) \tag{9.16}$$

$$= .51 \tag{9.17}$$

The standard error is $(1/\sqrt{342-3}) = .0543$. A 95 % confidence interval for z_f can be computed by adding and subtracting 1.96 standard errors from the estimate z_f:

$$\begin{gathered}[.51 - (1.96)(.0543)\ ,\ .51 + (1.96)(.0543)] \\ [.404\ ,\ .616]\end{gathered} \tag{9.18}$$

We can now transform these bounds back into the correlation scale to obtain:

$$b_L = \frac{e^{2z}-1}{e^{2z}+1} \tag{9.19}$$

$$= \frac{e^{2(.404)}-1}{e^{2(.404)}+1} \tag{9.20}$$

$$= .383 \tag{9.21}$$

and

$$b_U = \frac{e^{2z} - 1}{e^{2z} + 1} \tag{9.22}$$

$$= \frac{e^{2(.616)} - 1}{e^{2(.616)} + 1} \tag{9.23}$$

$$= .548 \tag{9.24}$$

So, our 95 % confidence interval is $[.383, .548]$. Our interval estimate does not contain 0, and therefore we may conclude that a linear relationship between education and income exists in the population (the population correlation is not 0).

9.1.4 Hypothesis Testing on the Correlation

We can develop a hypothesis test for the correlation in the Fisher's z metric as an extension of the original z test formula presented in Chap. 6:

$$z = \frac{z_f - H_0}{\hat{\sigma}_{z_f}}. \tag{9.25}$$

The usual null hypothesis is that the correlation $\rho = 0$ in the population; that is, there is no linear association between the two variables. In the example from the previous section, Fisher's $z_f = .51$ and $\hat{\sigma}_{z_f} = .0543$. Furthermore, when $\rho = 0$, $z_f = 0$. Thus:

$$z = \frac{.51 - 0}{.0543} \tag{9.26}$$

$$= 9.39. \tag{9.27}$$

$$\tag{9.28}$$

This z score is more than large enough to reject the null hypothesis that $\rho = 0$ in the population. Thus, we conclude that there is a linear association between education and income in the population.

9.1.5 Limitations of the Correlation

The correlation is useful as a first step in examining the relationships between variables in an analysis. However, it is limited in several respects. First, without modifications, the correlation does not tell us the extent to which a relationship between two variables may be contingent on a third variable. That is, what part of the correlation—if not all of it—is spurious (i.e., false)? For example, there is a strong correlation between ice cream consumption rates and violent crime rates, but this relationship is entirely explained by temperature.

Second, the correlation only measures linear association, and so it does not tell us anything about the relationship between variables that are not linearly related.

Third, the correlation is agnostic with respect to the direction of the causal relationship between variables. Of course, it is well-known and often repeated that "correlation does not prove causation," and we will discuss this in the next chapter in some depth. Yet it is often the case that we may wish to single out one variable as a "cause" and another as an "effect" for theoretical reasons. In such cases, the correlation is not theoretically pleasing.

To be sure, no method can directly address all of these limitations, especially the last. However, linear regression analysis comes closer to addressing these issues—at least theoretically—and answering the types of questions we are interested in social science.

9.2 Simple Linear Regression

Linear regression is probably the most important method of quantitative data analysis you will learn. It is one of the most widely used methods in scientific research and is the foundation of more complex methods. The logic of regression analysis is simple; the mathematics can range from relatively simple to relatively complex. For the purposes of this book, I am primarily interested in introducing the purpose of regression modeling and providing a basis for understanding how regression modeling works and what kind of information it yields. Thus, I will keep the mathematics as simple as possible.

9.2.1 Logic of Simple Regression

The basic logic of regression analysis is as follows. Suppose I have some variable y that I think is affected by another variable x, and I am interested in (1) determining whether there is a relationship between x and y and (2) quantifying its strength (note that I have explicitly stated that x causes y and not the other way around). If I believe the relationship between x and y to be linear, I could specify the relationship as:

$$y = \alpha + \beta x, \tag{9.29}$$

where α is the "intercept," and β is the "slope." Notice that this equation looks very similar to the slope-intercept form of the equation for a line from algebra: $y = mx + b$. The only differences are that b has been replaced by α, m has been replaced by β, and the two terms on the righthand side have traded places. These are only cosmetic differences: the idea behind both equations is the same. There is some value of y for the case in which $x = 0$ (α; the intercept), and there is a relationship between x and y such that a one-unit increase in x translates into—produces—a β unit increase in y. To make this idea concrete, consider the relationship between education and income. We may think that obtaining more education increases income. Thus, x is education, and y is income, and we expect that β is positive. Since income is generally positive, even at minimal levels of schooling, we might expect that α is also positive.

There are a few key problems with the model specified above. First, this specification is deterministic, meaning that a change in x produces a guaranteed β unit change in y. This is almost never the case in the real world. Second, and similarly, this model suggests that x is the *only* factor that is important in affecting y. This is also almost never the case in the real world. For example, education may be *a* cause of income, but so is winning the lottery and receiving an inheritance from a dead relative. Smoking may be a cause of lung cancer, but so is radon and exposure to asbestos. Lack of exercise may be a cause of obesity, but so is overeating and genetics. We certainly do not expect to have a single cause for any outcome. Otherwise, the scatterplot in Fig. 9.3 would have shown exactly one value of income for each value of education.

These limitations suggest that, to make the model more realistic, we should make the model a little more flexible, and so, we can modify it a little:

$$y_i = \alpha + \beta x_i + e_i. \tag{9.30}$$

In this specification, we have added the subscript i and an "error" term e_i. This model now says that the value of y for any individual (i) is a function of an intercept (which is constant for everyone), an effect of x (β; also constant for everyone), and some individual-specific error. In other words, under this specification we do *not* expect that x is the only cause of y nor do we expect that a one-unit change in x is necessarily associated with a β unit increase in y for every individual—we *do* expect a β unit increase in y for a one unit increase in x, but this is only an *average* effect and may not hold true for everyone.

For example, consider again the relationship between education and income. If we expect that income is affected by education, such that persons with more education make more income than persons with less education, we may wish to estimate the model:

$$\text{Income}_i = \alpha + \beta \times \text{Education}_i + e_i. \tag{9.31}$$

If we estimate this model using a subset of the GSS data, we obtain:

$$\text{Income}_i = -11.8 + 5.6 \times \text{Education}_i + e_i. \tag{9.32}$$

This result implies that each person's income can be recovered from his education and his unique error term along with the estimated intercept and slope. For persons not in the sample, we may estimate (predict) their income from their education using the predicted value:

$$\widehat{\text{Income}} = -11.8 + 5.6 \times \text{Education}.$$

Notice in this equation, the error term has dropped out; the reason is that, for the linear regression model, while individuals are expected to have some error, the average error is expected to be 0. We are generally not terribly interested in the particular sample members we have, but instead we are interested in the general pattern. Here, the results say that each additional year of schooling is worth (on average) \$5,600 more in income (recall that income is in \$1,000 units).

Technically, the results show that mean income at each level of education is about \$5,600 higher than the level of education below it. Given that we have observed *different* people at each level of schooling, we cannot truly conclude that increasing a person's education would necessarily lead to a higher income for him/her.

9.2.2 Estimation of the Model Parameters

How are the values of α and β estimated? In general, we would like estimates of the parameters that yield a prediction line that closely represents the pattern observed in a scatterplot of the data. Reconsider, for example, the two lines shown in Fig. 9.4. The 45-degree line (solid line) in the figure does not seem to do as good of a job at "fitting" the observed pattern as the dashed line. But, what makes the dashed line better?

There are many criteria that could be used to find estimates of α and β, but the most commonly used criterion is the "least squares" criterion. The least squares criterion is that we want values for the intercept and slope that minimize the sum of the squared error terms (the e_i) around the line implied by the parameter estimates:

$$G = min\left(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right), \tag{9.33}$$

where y_i is the observed value of y for person i, and $\hat{y}_i$ is the model-predicted value of y for person i. Given that $\hat{y}_i = \alpha + \beta x_i$, we can substitute this linear combination in for $\hat{y}_i$:

$$G = min\left(\sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2\right), \tag{9.34}$$

Source	Sum of Squares	DF	Mean Squares	F
Model	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$MSR = \frac{SSR}{df(R)}$	$\frac{MSR}{MSE}$
Error	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$	$MSE = \frac{SSE}{df(E)}$	
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	(Sample Variance)	

Table 9.1. Generic ANOVA table for a simple regression model.

We can find the best values of α and β by using a little calculus. In particular, the minimum (or maximum) of a function is the point where the slope of the curve implied by the function is 0. Thus, the function G is minimized by first taking its partial derivative with respect to α and then with respect to β. Next, we set these partial derivatives equal to 0 and solve to find the value of the parameters that minimize the function. If we do this, we find:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{9.35}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{9.36}$$

Notice that we must find β before computing α. Also notice that the solution for β can be rewritten as:

$$\hat{\beta} = \frac{cov(x, y)}{var(x)}, \tag{9.37}$$

because the numerator is simply $n - 1$ times the covariance of x with y, and the denominator is $n - 1$ times the variance of x. Finally, notice that the parameters have a caret (hat) above them, indicating that they are *estimates* of the population parameters α and β.

9.2.3 Model Evaluation and Hypothesis Tests

Once estimates of the intercept and slope are obtained, we are usually interested in evaluating how well the model fits the data and testing hypotheses regarding the relationship between x and y in the population. The parameter estimates can be used for these purposes. First, we can construct an ANOVA table, much as we did in the previous chapter, for the regression model results. The regression ANOVA table follows the format shown in Table 9.1.

Source	Sum of Squares	DF	Mean Squares	F
Model	89240.19	1	89240.19	95.02
Error	319315.48	340	939.16	
Total	408555.7	341	(1198.11)	

Table 9.2. ANOVA table for model of income regressed on education (data from continued example)

This table differs only slightly from the ANOVA table shown in the previous chapter. First, the names for the sources of the sums of squares differ: for regression, we have "model" (or "regression") and "error" rather than "between" and "within." Second, the calculations of the sums of squares are in terms of the data (y), the model predicted values ($\hat{y}$), and the sample mean ($\bar{y}$). It is still true that $SST = SSR + SSE$ (total sum of squares equals regression sum of squares plus error sum of squares). In the regression context, the total sum of squares is partitioned into the part explained by the regression line ($\hat{y} - \bar{y}$) and the part not explained by the regression line ($y - \hat{y}$). The remaining calculations are carried out much as before, but note the change in degrees of freedom: the model degrees of freedom for the simple regression model are 1 (two parameters minus 1); the error degrees of freedom are $n - 2$ (sample size minus number of estimated regression parameters).

Table 9.2 shows the ANOVA table from the education-income example. From the model sum of squares and the total sum of squares, we can compute R^2 as we did in ANOVA; the result is .22, meaning that 22 % of the variance in income is explained by education (conversely, 78 % remains to be explained—is error). This is a pretty strong result: in social science research, we are usually happy to have a double digit R^2.

The F statistic from this table is 95.02, which has a corresponding p value of approximately 0. This result tells us that we can reject the null hypothesis that there is no linear relationship between education and income in the population. In other words, education and income are linearly related in the population.

In the simple regression model, with one predictor (x) variable, the F statistic suffices as a test of the relationship between x and y. In multiple regression, however, the F test is an "omnibus" test that simply tells us whether at least one of the x variables is related to y, just as the F in ANOVA did not tell us which groups' means differed; it only told us that at least one mean differed from the others.

In order to conduct hypothesis tests on particular parameters so that we may determine whether particular x variables are related to y, we need standard errors of the parameter estimates. Just as we expect sample means to vary from the population mean and from sample to sample, we may expect that estimated regression parameters will also vary from sample to sample. Indeed, the Central Limit Theorem applies to regression coefficients much the same way as it applies to means: the distribution of sample regression coefficients is asymptotically normal with a mean equal to the true population regression coefficient (slope), and a standard deviation equal to a function of the mean squared error of the regression (MSE; see the ANOVA table) and the variance of x. Specifically:

$$s.e.(\hat{\alpha}) = \sqrt{\frac{\sigma_e^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}} \tag{9.38}$$

$$s.e.(\hat{\beta}) = \sqrt{\frac{\sigma_e^2}{\sum (x_i - \bar{x})^2}}, \tag{9.39}$$

where σ_e^2 is estimated using the MSE.

In the education-income example, the standard error for α ($\alpha = -11.8$) was 8.15 and the standard error for β ($\beta = 5.6$) was .57. In previous z and t tests, we established a hypothesized value for the population parameter (mean; μ) and evaluated how probable it would be for our sample mean to occur if the hypothesized value were true. We follow the same strategy in regression modeling, and we usually test the hypothesis that $\beta = 0$. If $\beta = 0$, then there is no linear relationship between x and y in the population; therefore, if we can reject this null hypothesis, the implication is that there is a linear relationship between the two variables in the population. We can test a similar hypothesis for the intercept, but because the intercept corresponds to the expected value of y when $x = 0$, and $x = 0$ may be an unreasonable value, we generally aren't terribly interested in testing whether the intercept is 0. Given that the MSE is an estimate of σ_e^2, the resulting test is a t test. In the education-income example, the t-test on the slope (β) is:

$$t = \frac{5.6 - 0}{.57} = 9.75. \tag{9.40}$$

The sample is large enough that we can use the z table to find the p value; that value is approximately 0, so we can reject the null hypothesis that $\beta = 0$ in the population. Education is linearly related to income.

9.3 Conclusions

In this chapter, we have developed statistical methods for assessing the relationship between two continuous variables, including the covariance, the correlation, and the simple regression model. As we discovered, the covariance is limited because of its measurement scale dependence. The correlation overcomes this limitation, but it suffers from its own limitations, including that it is agnostic with respect to the direction of causality between variables, and it cannot directly rule out spurious explanations for the association between variables. The simple regression model resolves, at least at a theoretical level, the first problem: it specifies a causal direction. The multiple regression model, an extension of the simple regression model, helps overcome the latter problem, as we will discuss next.

9.4 Items for Review

- Covariance
- Correlation coefficient
- Fisher's z transformation
- Confidence bounds on the correlation
- Hypothesis testing on the correlation
- Intercept and slope
- Error term
- Least squares criterion
- Estimators for intercept and slope
- Regression ANOVA table
- Standard errors of intercept and slope
- Regression hypothesis tests

9.5 Homework

Below is a small data set of 10 persons measured on 4 variables: life satisfaction, health, happiness, and education. Assume all variables are continuous.

Person	Satisfaction	Health	Happiness	Education
1	21	3	2	17
2	19	2	1	14
3	25	2	1	13
4	26	3	1	16
5	26	2	1	12
6	24	3	1	16
7	20	1	3	17
8	10	1	0	8
9	23	2	1	13
10	25	2	2	12

1. Some use life satisfaction and happiness measures as exchangeable, arguing that satisfaction and happiness are the same thing. Compute the correlation between

satisfaction and happiness and construct a confidence interval for it. How strong is the relationship? Use the confidence interval and hypothesis testing approach to answer this question.

2. Some argue that education influences health. Regress health on education and test the hypothesis that education and health are not related. Construct the ANOVA table and interpret all results.
3. Compute the correlation between health and happiness and construct a confidence interval for it.
4. Some argue that health limits one's ability to obtain more years of schooling. Regress education on health and test this hypothesis. Construct the ANOVA table and interpret. Given these results versus those in the previous question, can you say anything about the direction of causality?