20 Simple Regression: Statistical Inference

Calculation workbook: `regression_inference.xlsx`
Data workbook: `ch20_data.xlsx`

You get who you pay for.

In 2012, more than 2.3 million Americans were employed as customer service representatives, a figure that is expected to grow to 2.6 million by 2022.[1] Given this vast number of employees and the diversity of their employers, it is no surprise that the caliber of customer service varies tremendously from company to company. Among retailers, Costco is known for the quality of its customer service; Kmart is not. Among airlines, Virgin America is recognized for its strength in customer service; Spirit is not.[2]

A firm staffing its customer service department faces a tradeoff. By paying higher wages, the firm can attract better qualified applicants and increase its quality of service. But higher wages have a direct impact on profits. Likewise, keeping wages low holds labor expenses down, but at the cost of having many unhappy customers.

Suppose we wanted to understand the relationship between the wages of customer service workers and customer satisfaction among technology firms in Silicon Valley. If we observed the wages and customer satisfaction ratings of all such firms, we could use the regression line to summarize the relationship between these variables. But typically, we can only obtain observations about a random sample of firms. How can we use the results of a sample to draw inferences about the relation between wages and customer satisfaction in the population as a whole?

The previous chapter introduced the regression line as a descriptive statistics for bivariate data sets. We derived the regression line as the best linear predictor of *y* values from *x* values, explained how the correlation coefficient quantifies how well the regression line fits the data relative to the mean line, and compared the regression line to various alternative “lines of best fit.”

[1] The data and projection are from the Bureau of Labor Statistics: www.bls.gov/ooh/office-and-administrative-support/customer-service-representatives.htm.
[2] 2017 customer service rankings from temkingroup.com/temkin-ratings/.

In this chapter, we consider regression in settings in which the data we observe comes from a random sample and is to be used for statistical inference. To start, we introduce new probability models that describe the ex ante properties of the sample in terms of unknown parameters: the intercept α and slope β of the conditional mean line, and the conditional variance σ^2. As with the i.i.d. trials model, the probability models in this chapter may describe the behavior of an inherently random process or random sampling from a population. The two models we focus on—the classical regression model and the random sampling regression model—differ in whether one or both of the variables under study are generated by a random process, but the two models lead to inference procedures that are virtually identical. The models impose considerable structure on the population or random process under study, requiring linearity of conditional means and constant conditional variances. After studying these basic models, we introduce normality assumptions under which inferences may be drawn from small samples.

With the probability models in place, we introduce point estimators, interval estimators, and hypothesis tests, and other inference procedures for the context of regression. We also revisit analysis of residuals, and show how both sums of squared residuals and the correlation coefficient can be used as the basis for hypothesis tests about the slope parameter β. We conclude the chapter with a short discussion of regression and causation, complementing those from Chapters 18 and 19.

20.1 The Classical and Random Sampling Regression Models

We now present the two basic regression probability models: the **classical regression model** and the **random sampling regression model**. As in our previous probability models for statistical inference, these models specify the properties of the sample in terms of certain unknown parameters. Here there are three: α, β, and σ^2.

The classical regression model.

(C1) *Fixed x sampling*: $x_1, \dots, x_n$ *are fixed and not all identical;* $Y_1, \dots, Y_n$ *are independent random variables.*

(C2) *Linearity of conditional means*: $\mathrm{E}(Y_i) = \alpha + \beta x_i$.

(C3) *Constant conditional variances*: $\mathrm{Var}(Y_i) = \sigma^2$.

The random sampling regression model.

(R1) *Random sampling*: $(X_1, Y_1), \dots, (X_n, Y_n)$ *are independent as i varies;* $\mathrm{SD}(X_i) > 0$.

(R2) *Linearity of conditional means*: $\mathrm{E}(Y_i|X_i = x) = \alpha + \beta x$.

(R3) *Constant conditional variances*: $\mathrm{Var}(Y_i|X_i = x) = \sigma^2$.

20.1.1 Fixed *x* sampling vs. random sampling

The important difference between the two models comes in their initial assumptions, which describe how the x and y values in the sample are obtained. The classical regression model is based on assumption (C1), **fixed *x* sampling**. This assumption says that the x values are set in advance to at least two distinct values. The y values corresponding to each x value are random, and in particular are independent random variables.

The simplest interpretation of this assumption is in the context of an experiment. Suppose that an agricultural researcher would like to understand the relationship between fertilizer quantities and crop yields. According to assumption (C1), the researcher is able to choose the amounts of fertilizer x_i to use in each trial, and then can observe the crop yields Y_i that result. Since the researcher chooses the amount of fertilizer x_i, she views it as a fixed number. The crop yield Y_i is influenced by the choice of x_i, but is inherently random, being affected by various uncertain environmental conditions. To learn the relationship between fertilizer levels and crop yields, the researcher will want to try a variety of x_i values, so the assumption that these are not all the same is innocuous. The assumption that the Y_i are independent means, for instance, that learning that crop yield Y_1 was higher than its expected value provides no information about whether this is true of crop yield Y_2.

The random sampling regression model instead starts from assumption (R1), **random sampling**, under which each pair (X_i, Y_i) in the sample consists of two random variables. This is the case if each of these pairs is a random draw from a population described by a bivariate data set $\{(x_j, y_j)\}_{j=1}^{N}$. The values of X_i and Y_i are both obtained from a random draw of a single individual from the population. For instance, the data set might represent education and income levels of all adults in the United States. By looking at the education and income levels of randomly chosen U.S. adults, we can estimate how education and income are linked in the U.S. population as a whole. The assumption that the standard deviation of the x values is positive says that education levels in the U.S. population aren't all identical—again, an innocuous assumption.

It is usual to think of the classical regression model as describing an experiment, and the random regression model as describing sampling from a bivariate population, and our examples of these models will follow this pattern. However, these are not the only possibilities; in fact, all four combinations of model and application are possible.

For instance, suppose that a paper manufacturer experimenting with a new production process gets to observe two properties—the tensile strength and whiteness—of the paper the process produces. In this case, the experimenter does not choose anything, and his observations are randomly determined x and y values. This experiment is described by assumption (R1) of the random regression model.

Likewise, the classical regression model can be used to describe a structured approach to sampling from a population called **stratified sampling**. Instead of picking individuals at random from the population, the researcher instead prespecifies the x values to be used in his sample. As in Section 19.3, we can think of each of these x values as defining a **subpopulation** of size N_x of the full population of size N. By fixing the x values, the researcher is specifying in advance which subpopulations he will sample from. In the education/income example, the researcher might prespecify that the first member of his sample be someone whose education ended immediately after high school. Then the corresponding y observation would be obtained by choosing an adult at random from the subpopulation with $x = 12$ years of schooling. Since it fixes the x values in advance, stratified sampling agrees with assumption (C1) of the classical regression model.

Because the classical regression model has fewer moving parts, it is easier to analyze, and we will initially derive our procedures for statistical inference under its assumptions. Happily, the same procedures work equally well in the context of the random sampling model, for reasons we explain in Appendix 20.A.1. Because our procedures work equally well in both cases, we use examples based on both the classical and random sampling models without further ado.

20.1.2 Linearity of conditional means

The remaining assumptions of both probability models are about the conditional distributions of the y variables. Assumptions (C2) and (R2), called **linearity of conditional means**, say that the expected y values can be described by a linear function of the corresponding x values, specifically, the function $f(x) = \alpha + \beta x$. Like the parameter μ from the i.i.d. trials model, the parameters α and β appearing here are numbers that the experimenter does not know in advance. The point of obtaining the sample is to estimate and draw inferences about these unknown parameters.

Let's first interpret assumption (C2) of the classical regression model, in which the x values are fixed by the experimenter. This assumption says that if the experimenter sets the x value of the ith trial at x_i, then the expected value of the y variable is $\mathrm{E}(Y_i) = \alpha + \beta x_i$. In our earlier example, this means that within the relevant range of fertilizer levels, the expected crop yield is a linear function of the amount of fertilizer applied, with slope equal to β and intercept equal to α. It follows that increasing the amount of fertilizer applied by one unit always increases the expected yield by the same amount, namely, β units.

Under the random sampling model, the pair (X_i, Y_i) is determined by a single random draw from the population, so the value of X_i is not known until we take the sample. The equation $\mathrm{E}(Y_i|X_i = x) = \alpha + \beta x$ in assumption (R2) says that if we happen to sample an individual from subpopulation x, then this individual's expected y value is $\alpha + \beta x$.

Ultimately, assumption (R2) is a statement about the population $\{(x_j, y_j)\}_{j=1}^N$ from which the random sample is drawn. By conditioning on the event that $X_i = x$, we specify that Y_i is the y value of some member of subpopulation x. In Section 19.3, we called the average of such y values the **conditional mean** (or subpopulation mean) for subpopulation x. It is denoted $\mu_{y|x}$, and defined by

$$\mu_{y|x} = \frac{1}{N_x} \sum_{j\,:\,x_j=x} y_j.$$

In words, we find the N_x data pairs (x_j, y_j) from subpopulation x (i.e., for which $x_j = x$), sum up the corresponding y values, and divide by the number of data pairs in the subpopulation.

In Section 13.4, we argued that the traits of a random draw from a population are equal to the corresponding traits of the population from which the sample is drawn. Here, this connection implies that $\mathrm{E}(Y_i|X_i = x)$, the expected y value of a random draw from subpopulation x, is equal to the descriptive statistic $\mu_{y|x}$. We can therefore rewrite assumption (R2) as

$$\mu_{y|x} = \alpha + \beta x.$$

Thus in the random sampling regression model, the subpopulation means are assumed to be a linear function of x; the slope of the line is β and the intercept is α.

In Chapter 19, we used the notation $f(x) = \alpha + \beta x$ for something else. There it denoted the population regression line, the line that minimizes the sum of squared residuals

$$\sum_{j=1}^{N} \left(y_j - (a + bx_j)\right)^2 \tag{20.1}$$

over all choices of the intercept a and slope b. While it may seem that this disagrees with our notation here, it actually does not. We explained in Section 19.3.2 that when the conditional mean function is linear, as (C2) and (R2) posit, then it *is* the regression line.[3] So we began with the strong assumption that $f(x) = \alpha + \beta x$ is the conditional mean function, and this assumption implies that $f(x) = \alpha + \beta x$ is also the regression line, as the notation suggested in the first place. This agreement helps explain why we call the probability models above "regression models." It also justifies using regressions on the sample data to estimate the unknown parameters α and β, as we explain in Section 20.2.

20.1.3 Constant conditional variances

The remaining assumptions, (C3) and (R3), are called **constant conditional variances**. These assumptions say that the dispersion of the y values does not depend on the corresponding x value.[4] In the classical model, the experimenter chooses each value of x_i. While assumption (C2) says that the expected value of the observation Y_i is a linear function of this choice, assumption (C3) says that the dispersion of Y_i around its mean is equal to σ^2 regardless of this choice. In the agriculture example, the assumption says that within the relevant range of fertilizer levels, the amount of dispersion in crop yields is the same regardless of the fertilizer level chosen.

[3] If you don't remember why this is true, you should review Section 19.3.2 now. The explanation there is in the context of bivariate population data, but the claim also holds in the context of an experiment.

[4] Assumptions (C3) and (R3) also go by the name *homoskedasticity*. Econometricians enjoy not only Greek letters, but also Greek words!

For the random sampling model, assumption (R3) conditions on the event that the individual sampled is from subpopulation x. It requires that regardless of the subpopulation in question, the conditional variance in the y observation, $\mathrm{Var}(Y_i|X_i = x)$, be equal to σ^2.

This statement too is ultimately about the bivariate data set $\{(x_j, y_j)\}_{j=1}^N$ describing the population from which the sample is drawn. In Section 19.3, we defined the **conditional variance** (or subpopulation variance) for subpopulation x by

$$\sigma^2_{y|x} = \frac{1}{N_x} \sum_{j\,:\,x_j=x} \left(y_j - \mu_{y|x}\right)^2 .$$

Thus $\sigma^2_{y|x}$ is the average of the squared deviations $(y_j - \mu_{y|x})^2$ of the y values from the subpopulation x's mean.

Since the traits of a sample are the traits of the population from which the sample is drawn, the conditional variance $\mathrm{Var}(Y_i|X_i = x)$ from condition (R3) is equal to $\sigma^2_{y|x}$. Assumption (R3) says that all of these conditional variances have the same value, a requirement we can express equivalently as

$$\sigma^2_{y|x} = \sigma^2.$$

Like α and β, the parameter σ^2 generally isn't known in advance, but must be estimated using the results of the sample. Also, it is worth emphasizing that the common conditional variance $\sigma^2_{y|x} = \sigma^2$ is not equal to σ^2_y, the variance of the y data. The total variance in the y data must account not only for the variance within each subpopulation, but also for the fact that the subpopulation means vary around the overall mean μ_y. Thus $\sigma^2_y \geq \sigma^2$, with equality holding only if all of the conditional means are the same.[5]

20.1.4 How reasonable are the assumptions?

The assumptions of linear conditional means and constant conditional variances are rather strong, so we should not expect them to hold in every application. For the latter of these, let us return to the education (x) and income (y) example. In this context, assumption (R3) states that whether we consider the subpopulations with 8 years, 12 years, or 16 years of education, the dispersions of the income levels within the subpopulations are the same. There is reason to think that this will not be true. Instead, we might expect that subpopulations with higher levels of education will also exhibit more dispersion in their income levels: while limited education usually constrains one to a low-wage job, an advanced education gives one more flexibility in choosing a career, introducing a wider range of incomes. If the income dispersion varies across education levels, assumption (R3) is violated.[6]

[5] The exact relation between conditional and total variance is given by the decomposition of variance formula—see Exercise 4.M.3.

[6] The fancy name for this variation in subpopulation variances is *heteroskedasticity*.

Similarly, the assumption of linearity of conditional means may be a reasonable approximation in some applications, but not in others. For instance, suppose we studied the relationship between the number of cars using a certain segment of an interstate highway, and the speed at which these cars traveled. In this case, we expect that as long as the number of cars is relatively small, average speed will be close to the speed limit, but that as the road becomes congested, the average speed will drop sharply. In this case, the effect of a change in the number of cars on average speed will differ depending on whether the current number of cars on the road is small or large. This means that assumption (R2) is violated.

■ Example *The assumptions of the regression models in pictures.*

What does population data that satisfies the assumptions of the regression models look like? Suppose we collect a complete set of students' scores on a standardized math test in a small school district. Students from grades 3 through 12 take the test, and enrollments in each grade range from 20 to 36. The computerized adaptive test, which is scored on a 600-point scale, adjusts the difficulty in the questions according to the student's performance on previous questions, making it suitable for all grade levels.

Figures 20.1–20.3 present three possible scatterplots of grade levels (x) and test scores (y) for the school district. In each figure, the data points are slightly opaque, so that locations with many overlapping data points are darker

Figure 20.1: Population data that satisfies both conditional distribution assumptions.

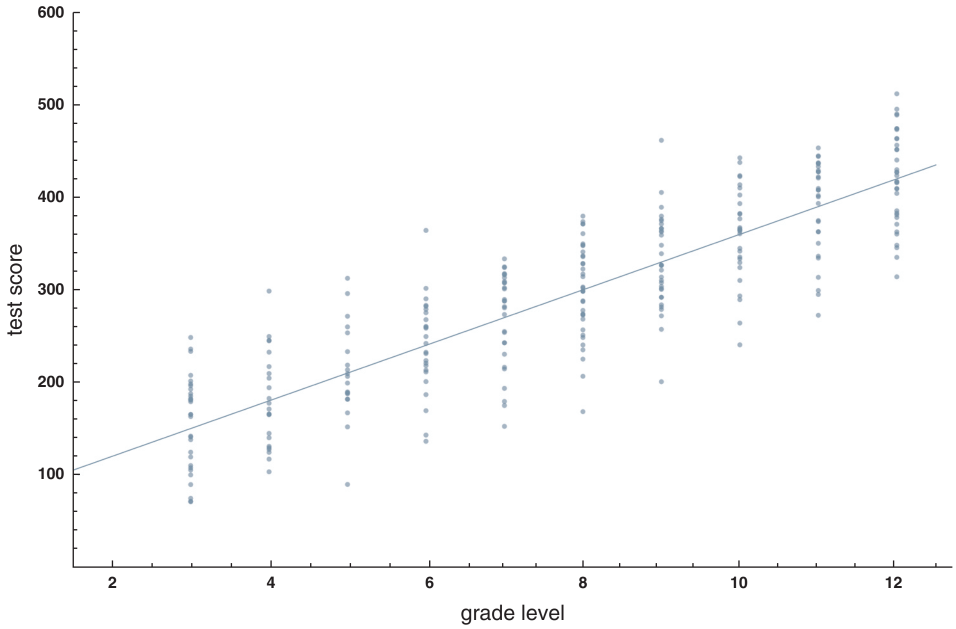


Figure 20.2: Population data that fails linearity of conditional means (R2).

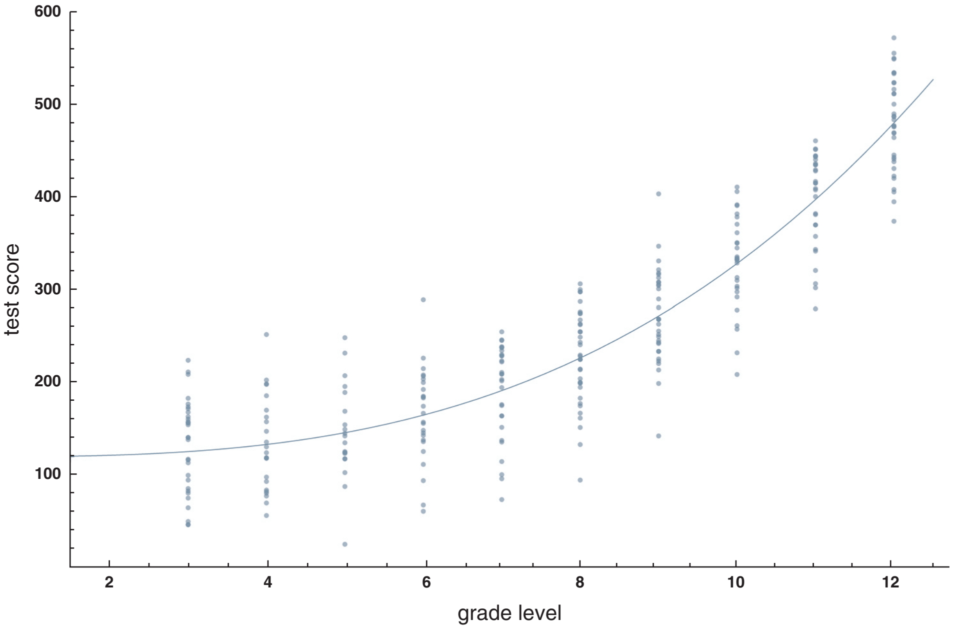
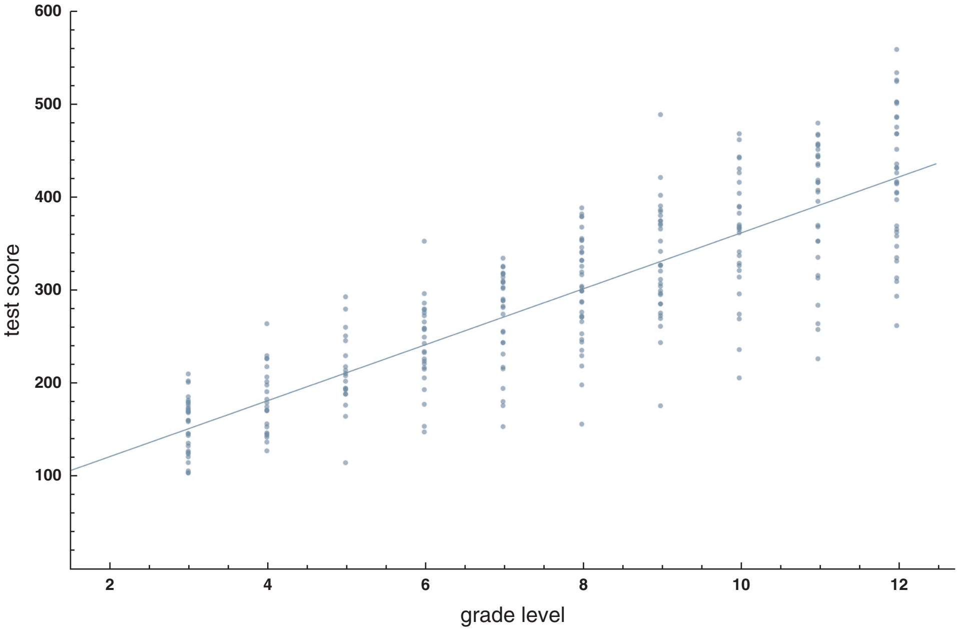


Figure 20.3: Population data that fails constant conditional variances (R3).



than isolated data points. Which of the scatterplots satisfy the assumptions on conditional distributions from our regression models? For any that do not, which assumptions are violated?

Figure 20.1 presents population data that satisfy both of the regression models' assumptions on conditional distributions. For each grade level, the mean test score lies on the linear conditional mean function $\mu_{y|x} = 40 + 30x$, which is drawn in the figure. Moreover, the subpopulation variances of scores are the same for each grade level, namely, $\sigma^2_{y|x} = 50^2 = 2500$.

In Figure 20.2, the subpopulation means do not change at a constant rate: the rate increase is faster at higher grade levels than at lower ones. Thus linearity of conditional means (R2) fails to hold for the data presented here. In fact, the conditional means are described by the function $\mu_{y|x} = 100 + .2083\,x^3$, which is drawn in the figure. The data in the figure do satisfy constant conditional variances (R3), again with $\sigma^2_{y|x} = 50^2 = 2500$.

In Figure 20.3, the subpopulation means again vary linearly according to the function $\mu_{y|x} = 40 + 30x$. But in this figure, test scores become more dispersed as the grade level increases. Indeed, the subpopulation variances for this data set are described by the function $\sigma^2_{y|x} = (15 + 5x)^2$. Thus constant conditional variances (R3) is violated. ■

In practice, how well the assumptions of the regression models are satisfied is a matter of degree: sometimes they are reasonably good approximations, and sometimes not. When they are not, one can turn to regression probability models that use weaker assumptions. In Appendix 20.A.2, we present a regression model that makes no assumptions at all about the conditional distributions of the y variables. We argue that versions of the main inference procedures described in the main text can still be used, although with differences in both their details and their interpretations.

To close this discussion, we should note one violation of the basic assumptions that is not so easily resolved. Suppose that our observations are **time series**, describing the evolution of certain quantities over time. For instance, macroeconomic analyses of gross national products, inflation rates, unemployment rates, and exchange rates make use of time series data. Time series observations generally do not satisfy the independence condition from assumptions (C1) and (R1), but instead exhibit **serial correlation**. For example, if Y_i represents the inflation rate in year i, it seems natural to expect correlation over time, with abnormally high inflation in year i often being followed by abnormally high inflation in year $i + 1$. Econometric analysis of time series data uses methods that are quite different from those that we present here and are covered in more advanced books on econometrics.

20.2 The OLS Estimators

Our two regression models are defined in terms of three unknown parameters: α and β, which describe the conditional expectation function, and the conditional variance σ^2. In this section and the next, we consider how to estimate these parameters using the results of a sample. We begin by considering point estimation of α and β.

20.2.1 Defining the OLS estimators

According to the assumptions of our regression models, the function $f(x) = \alpha + \beta x$ describes the conditional means of y values given x values. But as we explained in Section 20.1.2, this implies that $f(x) = \alpha + \beta x$ is also the population regression line. The latter fact suggests an approach to estimating α and β.

In earlier chapters, we used the sample mean $\bar{X}_n$ to estimate the population mean μ_x. Here, we will use the sample analogues of α and β—that is, the intercept and slope of the *sample regression line*—as our estimators for α and β, the intercept and slope of the population regression line.[7] In other words, to estimate the line that minimizes the sum of squared residuals in the population, we use the line that minimizes the sum of squared residuals in the sample.

In Chapter 19, we expressed the parameters α and β defining the population regression line in terms of other descriptive statistics for the population:

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2} \quad \text{and} \quad \alpha = \mu_y - \beta\mu_x.$$

We define the sample regression line using the sample analogues of these equations.

We focus on the simpler case of the classical regression model, in which the x values are picked by the researcher.[8] To avoid conflict with our notation for population descriptive statistics, we introduce the notations

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

for the mean and variance of the x values the researcher has chosen. As usual, we let

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

denote the sample mean of the y variables, and we let

$$S_{xY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$$

[7] This approach to coming up with estimators is sometimes called the *sample-analogue principle*.

[8] For the random sampling regression model, see Appendix 20.A.1.

denote the covariance between the x values and y variables in the sample. We use capital $\mathcal{S}$ here to emphasize that $\mathcal{S}_{xY}$ is a random variable: it depends on the values of the random variables Y_i.[9]

With these preliminaries addressed, we can define our estimators for the classical regression model, which are known as the **ordinary least squares estimators**, or OLS estimators for short. The name refers to the fact that these estimators minimize the sum of squared residuals in the sample data.[10]

Definition.

In the classical regression model, the OLS estimators for α and β are

$$B = \frac{\mathcal{S}_{xY}}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad A = \bar{Y} - B\bar{x}. \tag{20.2}$$

■ Example *Demand more beer.*

A microbrewer has decided to expand its distribution network to take advantage of the growing popularity of its flagship beer, Trial by Hops. Before setting the price for Trial by Hops in the new territory, the CEO asks the marketing director to estimate the expected-demand curve for their product. To do so, the director runs an experiment, varying the wholesale keg price of Trial by Hops at 33 of their existing distributors, all of which cover territories of comparable size and demographics, and observing the number of kegs each distributor sells during the next week. The data can be found in the workbook `ch20_data.xlsx/trial_by_hops`.

The firm believes that demand for Trial by Hops satisfies the assumptions of the classical regression model. Thus, for each fixed dollar price x, the expected number of kegs sold is $\alpha + \beta x$, and the variance in the number of kegs sold is σ^2, where the parameters α, β, and σ^2 are unknown.[11]

After the running the experiment, the microbrewer compiles the following basic statistics summarizing its results:

$$\bar{x} = 80.727,\ s_x^2 = 41.205,\ \bar{Y} = 28.394,\ \text{and}\ \mathcal{S}_{xY} = -33.295.$$

[9] Defining s_x^2 and $\mathcal{S}_{xY}$ by dividing by $n-1$ makes these definitions agree with those for the random sampling regression model (see Appendix 20.A.1). We could instead have divided by n, so as to match the descriptive statistic formulas from Chapter 19. This wouldn't affect the formulas for the OLS estimators, since the divisors cancel in the first formula for B in equation (20.2).

[10] The traditional notations for the OLS estimators are $\hat{\beta}$ and $\hat{\alpha}$ (pronounced "beta hat" and "alpha hat"). We instead use capital letters for these estimators to emphasize that they are random variables, and so have distributions, expected values, variances, and the like.

[11] In this experiment, the x variable is the price, and the y variable is the quantity demanded. If you've taken any microeconomics, you may have noticed that this choice of variables is not the one typically used in graphs of demand curves, which put quantity on the horizontal axis and price on the vertical axis. But you may also remember that such graphs are called "inverse demand curves." This is because the price is chosen first and then the demand at that price is realized, not the other way around.

The OLS estimates for β and α are therefore

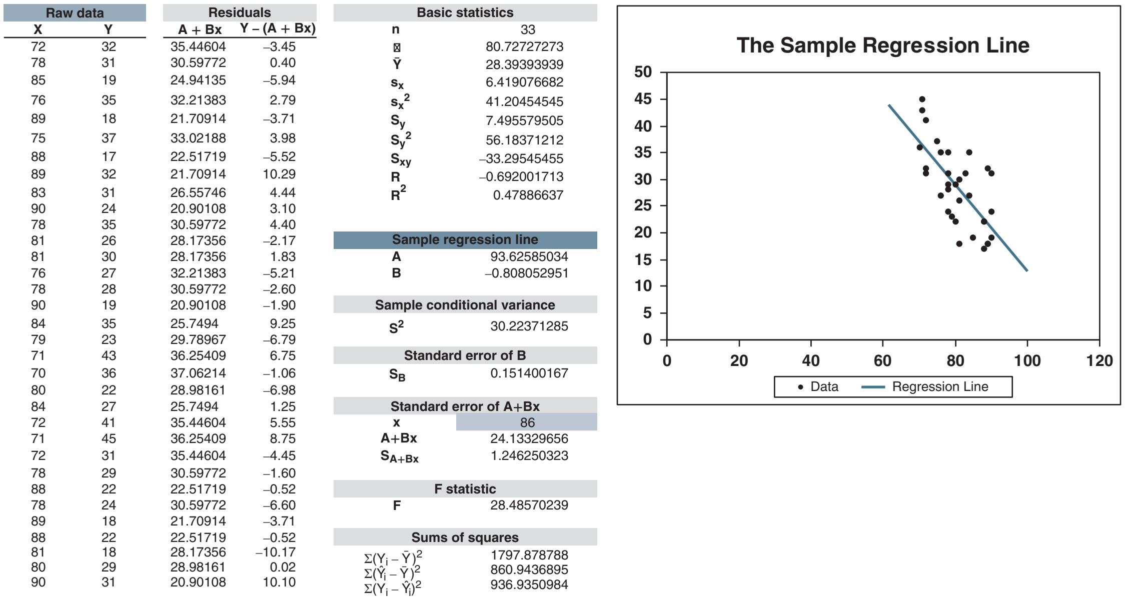
$$B = \frac{\mathcal{S}_{xY}}{s_x^2} = -.8080 \text{ and } A = \bar{Y} - B\mu_x = 93.63.$$

The equation for the sample regression line, which here is the estimate of the expected-demand curve, is $y = 93.63 - .8080x$. Based on the results of this experiment, the microbrewer estimates that increasing the price it charges by one dollar reduces the expected number of kegs sold by about .8. ■

Excel calculation: *Computing OLS estimates in Excel*

Computing the sample regression line directly from raw data involves a lot of calculation, which is best done using a computer. The workbook `regression_inference.xlsx` provides a template for doing these calculations. Figure 20.4 presents the workbook's output for the Trial by Hops data.

Figure 20.4: `regression_inference.xlsx`



Raw data		Residuals	
X	Y	A + Bx	Y – (A + Bx)
72	32	35.44604	–3.45
78	31	30.59772	0.40
85	19	24.94135	–5.94
76	35	32.21383	2.79
89	18	21.70914	–3.71
75	37	33.02188	3.98
88	17	22.51719	–5.52
89	32	21.70914	10.29
83	31	26.55746	4.44
90	24	20.90108	3.10
78	35	30.59772	4.40
81	26	28.17356	–2.17
81	30	28.17356	1.83
76	27	32.21383	–5.21
78	28	30.59772	–2.60
90	19	20.90108	–1.90
84	35	25.7494	9.25
79	23	29.78967	–6.79
71	43	36.25409	6.75
70	36	37.06214	–1.06
80	22	28.98161	–6.98
84	27	25.7494	1.25
72	41	35.44604	5.55
71	45	36.25409	8.75
72	31	35.44604	–4.45
78	29	30.59772	–1.60
88	22	22.51719	–0.52
78	24	30.59772	–6.60
89	18	21.70914	–3.71
88	22	22.51719	–0.52
81	18	28.17356	–10.17
80	29	28.98161	0.02
90	31	20.90108	10.10

Basic statistics	
n	33
⊠	80.72727273
$\bar{Y}$	28.39393939
s_x	6.419076682
$s_x{}^2$	41.20454545
S_y	7.495579505
$S_y{}^2$	56.18371212
S_{xy}	–33.29545455
R	–0.692001713
R^2	0.47886637
Sample regression line	
A	93.62585034
B	–0.808052951
Sample conditional variance	
S^2	30.22371285
Standard error of B	
S_B	0.151400167
Standard error of A+Bx	
x	86
A+Bx	24.13329656
S_{A+Bx}	1.246250323
F statistic	
F	28.48570239
Sums of squares	
$\Sigma(Y_i - \bar{Y})^2$	1797.878788
$\Sigma(\hat{Y}_i - \bar{Y})^2$	860.9436895
$\Sigma(Y_i - \hat{Y}_i)^2$	936.9350984

To use the workbook, enter up to 1000 pairs of sample data points in the first pair of columns. The top part of the third pair of columns reports various basic statistics describing the sample data. Immediately below these are the OLS estimators—the intercept A and slope B of the sample regression line. The diagram on the right plots this line on top of a scatterplot of the sample data.

The workbook computes a number of additional quantities that are used to draw inferences about the unknown parameters α, β, and σ^2, and that we discuss below.

20.2.2 Basic properties of the OLS estimators

The OLS estimators A and B, which describe the sample regression line, seem like natural candidates to estimate the parameters α and β, which define the true regression line. But of course, a more convincing justification for these estimators requires our usual criteria of unbiasedness, consistency, and efficiency from Chapter 14.

In order to evaluate these criteria, we need to be able to determine the traits—the means, variances, and covariance—of the OLS estimators. To do so, we first express these estimators in a very convenient form.

Linearity of the OLS estimators.

Under the classical regression model, the OLS estimators are linear functions of the random variables $\{Y_i\}_{i=1}^n$: they can be written as

$$B = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{(n-1)s_x^2} \right) Y_i \quad \text{and} \quad A = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}\,(x_i - \bar{x})}{(n-1)s_x^2} \right) Y_i. \tag{20.3}$$

Exercise 20.M.1 works through the derivations of these formulas from the definitions in (20.2) above.

Knowing that the OLS estimators are linear functions of the random variables Y_i is very useful. For one thing, it allows us to compute their traits using the formulas for linear functions of random variables from Chapters 3 and 4.

Traits of the OLS estimators.

Under the classical regression model, the OLS estimators are unbiased:

$$\mathrm{E}(B) = \beta \;\; \textit{and} \;\; \mathrm{E}(A) = \alpha. \tag{20.4}$$

The variances and covariance of the OLS estimators are

$$\mathrm{Var}(B) = \frac{\sigma^2}{(n-1)s_x^2}, \;\; \mathrm{Var}(A) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right),$$

$$\text{and} \quad \mathrm{Cov}(A,B) = -\frac{\sigma^2 \bar{x}}{(n-1)s_x^2}. \tag{20.5}$$

The calculations of the mean and variance of B are presented in Appendix 20.A.3, and the remaining calculations are described in Exercises 20.M.2 and 20.M.3.

Display (20.4) says that A and B are unbiased estimators of α and β: the estimators are correct on average, where the average is taken ex ante over all possible results of the sample.

Display (20.5) reports the variances and covariance of A and B. These measures of the estimators' dispersion and comovements are needed to construct interval estimators and hypothesis tests. Notice that increasing the conditional variance σ^2 increases $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$. This makes sense: spreading out the y values in each subpopulation makes it harder to obtain an accurate estimate of the true regression line. On the other hand, increasing the variance s_x^2 of the x values reduces the dispersion of the estimators. Intuitively, the sample regression line is more likely to be close to the true regression line if it is based on a wide range of x values than if all of the x values are packed together.

Another desirable property for estimators introduced in Chapter 14 is consistency, the requirement that as the size of the sample grows large, the probability that the estimator takes a value close to the parameter it estimates approaches 1. Is this true of the OLS estimators? Since the OLS estimators are unbiased, it is enough to show that their variances, $\mathrm{Var}(B)$ and $\mathrm{Var}(A)$, approach zero as the sample size grows large.[12] Looking at the formulas for $\mathrm{Var}(B)$ and $\mathrm{Var}(A)$ above, we see that this depends on the values of the x_i, which in the classical regression model are chosen by the experimenter. As long as the experimenter includes some variation in the x values—in particular, as long as s_x^2 is kept some fixed distance away from 0—then $\mathrm{Var}(B)$ and $\mathrm{Var}(A)$ will go to zero, implying consistency.

Consistency of the OLS estimators.

Under the classical regression model, so long as there is nonnegligible variation in the x values, the OLS estimators B and A are consistent.

20.2.3 Estimating conditional means

In some applications, we are particularly interested in estimating the mean y value corresponding to a particular x value. For instance, our brewer might want to estimate the expected number of kegs sold at a particular price point.

According to the classical regression model, the mean y value in subpopulation x is $\alpha + \beta x$. Since we use the OLS estimators A and B to estimate the parameters α and β, the natural point predictor of the conditional mean is $A + Bx$. It is easy to check that this estimator is unbiased:

$$\mathrm{E}(A + Bx) = \mathrm{E}(A) + \mathrm{E}(B)x = \alpha + \beta x.$$

A somewhat longer calculation determines its variance:

$$\mathrm{Var}(A + Bx) = \mathrm{Var}(A) + \mathrm{Var}(Bx) + 2\mathrm{Cov}(A, Bx)$$

[12] This follows directly from Chebyshev's inequality, which we introduced in Chapter 7 to prove the law of large numbers.

$$
\begin{aligned}
&= \mathrm{Var}(A) + x^2\mathrm{Var}(B) + 2x\mathrm{Cov}(A, B)\\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) + x^2\frac{\sigma^2}{(n-1)s_x^2} - 2x\frac{\sigma^2\bar{x}}{(n-1)s_x^2}\\
&= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2 + x^2 - 2x\bar{x}}{(n-1)s_x^2}\right)\\
&= \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right). \qquad (20.6)
\end{aligned}
$$

Finally, since A and B are consistent estimators of α and β, $A + Bx$ is a consistent estimator of $A + Bx$. Let's summarize these conclusions.

Estimating conditional means.

Under the classical regression model, $A + Bx$ is an unbiased estimator of $\alpha + \beta x$, with variance given by (20.6). *If there is non-negligible variation in the x values, $A + Bx$ is also a consistent estimator of $\alpha + \beta x$.*

Looking more closely at expression (20.6), we see that increasing the distance between x and $\bar{x}$ makes $\mathrm{Var}(A + Bx)$ larger. In other words, we have better information about y values corresponding to x values that are "in the middle" than to ones that are extreme. Why is this so? Using the definitions of α and A, we can express the conditional mean and its estimator in terms of $x - \bar{x}$:

$$
\begin{aligned}
\alpha + \beta x &= (\mu_y - \beta\bar{x}) + \beta x = \mu_y + \beta(x - \bar{x});\\
A + Bx &= (\bar{Y} - B\bar{x}) + Bx = \bar{Y} + B(x - \bar{x}).
\end{aligned}
$$

Comparing the final expressions, we see that when x is close to $\bar{x}$, so that $x - \bar{x}$ is close to zero, errors in estimating β will have little effect on our estimates of $\alpha + \beta x$. But when x is far from $\bar{x}$, so that $x - \bar{x}$ is not close to zero, the impact of these errors will be large as well.

20.2.4 Approximate normality of the OLS estimators

In order to construct interval estimators and hypothesis tests using the OLS estimators, we need information about their distributions. In the case of the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ of independent and identically distributed random variables, we used the central limit theorem to conclude that the $\bar{X}_n$ is approximately normally distributed. We can reach the same conclusion about the OLS estimators, but for somewhat more complicated reasons.

Approximate normality of the OLS estimators.

In the classical regression model, if n is not small and there is non-negligible variation in the x values, then the OLS estimators A and B are approximately normally distributed. Under the same conditions, for any choice of x, $A + Bx$ is approximately normally distributed.

To see why this statement is true, let's focus on the slope estimator B. We know from equation (20.3) that we can write B as a weighted sum of the Y_i. The latter random variables are independent, but they are not identically distributed—since $\mathrm{E}(Y_i) = \alpha + \beta x_i$, the means of the Y_i differ, and their distributions may also differ in other respects. For these reasons, the central limit theorem introduced in Chapter 7 cannot be applied to B.

Happily, it turns out that the conclusions of the central limit theorem do not require all of the structure we imposed in Chapter 7. If we have a collection of independent random variables, and we take a weighted sum that does not put too much weight on any one term, there is a generalization of the central limit theorem that allows us to conclude that this weighted sum has an approximately normal distribution.[13] This is precisely the result we need to conclude that the OLS estimators are approximately normally distributed.

20.2.5 Efficiency of the OLS estimators: The Gauss-Markov theorem*

When we introduced point estimation in Chapter 14, we proposed the notion of efficiency as one of our criteria for evaluating the quality of estimators. By choosing efficient estimators—that is, by choosing estimators whose variance is as small as possible—we ensure that we extract the maximum amount of information from the sample.

In the context of regression inference, we have the following basic efficiency result.[14]

The Gauss-Markov theorem.

Of all unbiased linear estimators of α and β, the OLS estimators A and B have the smallest variance. Likewise, for any choice of x, of all unbiased linear estimators of $\alpha + \beta x$, the estimator $A + Bx$ has the smallest variance.

We present a proof of this result in Appendix 20.A.4.

The first part of the Gauss-Markov theorem considers the class of unbiased linear estimators of the regression parameters α and β. As we argued in Chapter 14, it makes sense to restrict attention to unbiased estimators because they are correct "on average" from the ex ante point of view—that is, from the perspective of the time before the sample is taken. The restriction to linear estimators—those that can be expressed as a linear function of the random variables Y_i—can be defended on the grounds of simplicity and convenience. The theorem tells us that among all estimators satisfying these conditions, the OLS estimators have the lowest variance.

[13] This general version of the central limit theorem is known as the *Lindeberg-Feller central limit theorem*.

[14] We encountered Carl Friedrich Gauss, the originator of the regression line, in Section 19.1. Russian mathematician Andrey Andreyevich Markov (1856–1922) made basic contributions to probability theory, with fundamental work on a class of random processes, now called *Markov processes*, that are used in an endless variety of applications.

Thus within this class, one cannot make better use of the information provided by the sample in estimating the regression parameters.

The second part of the Gauss-Markov theorem extends this conclusion to estimators of subpopulation means $\alpha + \beta x$. Among all unbiased linear estimators of $\alpha + \beta x$, the OLS estimator $A + Bx$ has the lowest variance.

This latter result can help us appreciate the strength of the assumptions of the classical regression model. You might expect that to estimate the subpopulation mean $\mu_{y|x}$, the best choice of estimator would be the **sample subpopulation mean**

$$\bar{Y}_{|x} = \frac{1}{n_x} \sum_{i\,:\,x_i = x} Y_i; \tag{20.7}$$

here n_x is the number of times x appears in the list $\{x_i\}_{i=1}^n$ of values used in the sample. $\bar{Y}_{|x}$ is a linear and unbiased estimator of $\mu_{y|x}$, and it seems particularly appealing when n_x is large. But the Gauss-Markov theorem tells us that the estimator $A + Bx$ is preferable to $\bar{Y}_{|x}$; the former is also linear and unbiased, but it has a lower variance than $\bar{Y}_{|x}$.

This fact is a consequence of the structure imposed by the classical regression model. Clearly, observation Y_i from subpopulation x provides information about the subpopulation mean $\mu_{y|x} = \mathrm{E}(Y_i)$. But since all y observations, whether from subpopulation x or not, are used to compute the OLS estimators A and B, all of these observations provide information about the parameters α and β. And since $\mu_{y|x} = \alpha + \beta x$ (by assumption (C2)), it follows that every y observation is also informative about this conditional mean. Indeed, as the estimators A and B use the information from all y observations to estimate α and β as efficiently as possible, it makes sense that $A + Bx$ is the most efficient estimator of $\mu_{y|x} = \alpha + \beta x$.

20.3 The Sample Conditional Variance

The conditional variance σ^2 describes the dispersion of the y values within each subpopulation. The smaller is this conditional variance, the less likely it is that our sample observations will be far from the population regression line $y = \alpha + \beta x$, and so the better the estimate the sample regression line $y = A + Bx$ is likely to provide.

To create interval estimators and hypothesis tests for the parameters α and β and the conditional means $\alpha + \beta x$, we need estimators of the variances $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$ and the covariance $\mathrm{Cov}(A, B)$ of the OLS estimators A and B. The formulas for $\mathrm{Var}(A)$, $\mathrm{Var}(B)$, and $\mathrm{Cov}(A, B)$ were presented in display (20.5). The terms in these formulas involving the x values and the sample size n are known in advance; but the conditional variance σ^2 is unknown. Thus by constructing an estimator of σ^2, we also obtain the sought-after estimators of $\mathrm{Var}(A)$, $\mathrm{Var}(B)$, and $\mathrm{Cov}(A, B)$, allowing us to draw inferences about α, β, and $\alpha + \beta x$.

By definition, the conditional variance σ^2 is the variance of each observation Y_i. In other words, σ^2 is the expected value of the squared deviation of each Y_i from its mean $E(Y_i) = \alpha + \beta x_i$. If we knew the conditional means $E(Y_i) = \alpha + \beta x_i$, then the natural way to estimate σ^2 would be to take the average of the squared deviations $(Y_i - (\alpha + \beta x_i))^2$. But since we don't know the parameters α or β, we don't know the conditional means either.

We faced a similar problem in Section 14.4 when seeking an estimator of the variance $\sigma_X^2 = \mathrm{E}(X_i - \mu_X)^2$ of i.i.d. trials $\{X_i\}_{i=1}^n$. If we knew the mean μ_X of the trials, we could have estimated σ_X^2 using the average of the squared deviations $(X_i - \mu_X)^2$. But not knowing μ_X, we replaced it with the sample mean $\bar{X}_n$, and defined the sample variance $\mathcal{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. In computing the "average" that defines the sample variance, we divided by $n-1$ rather than n in order to an obtain an unbiased estimator of σ_X^2.

The solution to our current problem follows these same lines. Since the conditional means $\mathrm{E}(Y_i) = \alpha + \beta x_i$ are unknown, we replace them with the estimators $A + Bx_i$. Taking the "average" of the squared deviations then leads to the **sample conditional variance**,

$$\mathcal{S}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - (A + Bx_i))^2. \tag{20.8}$$

This time, the "average" divides by $n-2$ rather than n.

We can express $\mathcal{S}^2$ more succinctly by introducing sample versions of some definitions from the previous chapter. Let

$$\hat{Y}_i = A + Bx_i \quad \text{and} \quad U_i = Y_i - \hat{Y}_i \tag{20.9}$$

denote the ith **sample regression prediction** and the ith **sample regression residual**. Then the sample conditional variance is the "average" squared regression residual:

$$\mathcal{S}^2 = \frac{1}{n-2}\sum_{i=1}^n U_i^2.$$

Here are the key properties of $\mathcal{S}^2$.

Unbiasedness and consistency of the sample conditional variance.

In the classical regression model, the sample conditional variance $\mathcal{S}^2$ is an unbiased estimator of the conditional variance σ^2. If there is non-negligible variation in the x values, this estimator is also consistent.

We verify that $\mathcal{S}^2$ is unbiased, that is, that $\mathrm{E}(\mathcal{S}^2) = \sigma^2$, in Appendix 20.A.5.

In defining the sample conditional variance, we divide by $n-2$ rather than n in order to obtain an unbiased estimator of σ^2: dividing by n would yield an estimator that is biased low. The intuition here is similar to the one for dividing by $n-1$ in defining the sample variance (Section 14.4.2). By definition, the sample regression line minimizes the sum of squared sample residuals $Y_i - (A + Bx_i)$ for every realization of the sample.[15] These sample residuals thus tend to be smaller than

[15] Compare the discussion of the i.i.d. trials model in Section 14.4.2.

residuals $Y_i - (\alpha + \beta x_i)$ defined using the true regression line. By using a smaller divisor in (20.8)—by dividing by $n - 2$ rather than n—we compensate for this bias.

As in Section 14.4.2, we can remember to divide by $n - 2$ using the rule of thumb called **degrees of freedom**. To estimate σ^2, we first need to estimate α and β. There is a sense in which estimating these two parameters uses up the information from two of our n trials, leaving only $n - 2$ pieces of information—or "degrees of freedom"—for estimating σ^2. The rule of thumb says that dividing by the number of degrees of freedom rather than the number of trials leads to an unbiased estimator. Of course this does not prove that the rule of thumb is correct, but it gives us a guess that can be verified by calculation.

Excel calculation: *The sample conditional variance*

The `regression_inference.xlsx` workbook, introduced earlier to compute the OLS estimates, also computes the sample conditional variance. Figure 20.4 shows that the sample conditional variance for the Trial by Hops data is $\mathcal{S}^2 = 30.2237$. This is the brewer's estimate of the conditional variance σ^2, which here represents the dispersion in the number of kegs that would be sold at any fixed price x.

20.4 Interval Estimators and Hypothesis Tests

At this point, we have defined the OLS estimators A and B and have argued that they are unbiased and that they are approximately normally distributed if the sample size is not too small. We've also specified their variances and covariance in terms of the conditional variance σ^2, and have introduced the sample conditional variance $\mathcal{S}^2$ as an unbiased and consistent estimator of σ^2. With these tools in hand, we are in a position to construct interval estimators and hypothesis tests for the slope parameter β and the conditional means $\alpha + \beta x$. Choosing $x = 0$ in the latter case gives us inference procedures for α.

The procedures we introduce here work in essentially the same way as our procedures for inference about the unknown mean of i.i.d. trials, and for essentially the same reasons. We therefore start by reviewing the procedures for inference about an unknown mean, introducing new notation to make the transition to regression inference easier.

20.4.1 Review: Inference about an unknown mean

Suppose we want to estimate the unknown mean $\mathrm{E}(X_i) = \mu_X$ of a sequence $\{X_i\}_{i=1}^n$ of i.i.d. trials with variance $\mathrm{Var}(X_i) = \sigma_X^2$. Our point estimator for μ_X is the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. This estimator is unbiased with variance $\mathrm{Var}(\bar{X}_n) = \frac{\sigma_X^2}{n}$, and the central limit theorem implies that it is approximately normally distributed. To sum up, $\bar{X}_n \approx N(\mu_X, \frac{\sigma_X^2}{n})$.

Using this knowledge about the distribution of $\bar{X}_n$, we can show that

$$\text{(20.10)} \qquad \mathrm{P}\Big(\mu_X \in \Big[\bar{X}_n - z_{a/2}\tfrac{\sigma_X}{\sqrt{n}}, \bar{X}_n + z_{a/2}\tfrac{\sigma_X}{\sqrt{n}}\Big]\Big) \approx 1 - a.$$

In words, the mean μ_X is within $z_{a/2}\frac{\sigma_X}{\sqrt{n}}$ of the sample mean $\bar{X}_n$ with probability close to $1 - a$. The term in the brackets is therefore our $(1 - a)$ interval estimator for μ_X.[16] Likewise, we can show that

$$\text{(20.11)} \qquad \mathrm{P}\Big(\bar{X}_n > \mu_0 + z_a\tfrac{\sigma_X}{\sqrt{n}} \mid \mu_X = \mu_0\Big) \approx a.$$

In words, if the mean were μ_0, then the probability that the sample mean would exceed $\mu_0 + z_a\frac{\sigma_X}{\sqrt{n}}$ is approximately a. Thus the critical values for the one-tailed hypothesis test are $\mu_0 \pm z_a\frac{\sigma_X}{\sqrt{n}}$. In all of these expressions, the left-hand term consists of (i) a z-statistic and (ii) $\frac{\sigma_X}{\sqrt{n}}$, the standard deviation of the estimator $\bar{X}_n$.

Typically, the variance σ_X^2 is unknown, and so must be estimated using the sample variance $\mathcal{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. If the number of trials is large enough, it is a reasonable approximation to replace the actual standard deviation σ_X with the sample standard deviation $\mathcal{S}_X = \sqrt{\mathcal{S}_X^2}$ in the formulas above. Stating this another way, we replace $\mathrm{SD}(\bar{X}_n) = \frac{\sigma_X}{\sqrt{n}}$ in the formulas above with its estimator

$$\mathcal{S}_{\bar{X}} = \frac{\mathcal{S}_X}{\sqrt{n}},$$

known as the **standard error** of $\bar{X}_{n}$. With this substitution, formulas (20.10) and (20.11) become

$$\mathrm{P}\big(\mu_x \in \big[\bar{X}_n - z_{a/2}\mathcal{S}_{\bar{X}}, \bar{X}_n - z_{a/2}\mathcal{S}_{\bar{X}}\big]\big) \approx 1 - a \quad \text{and}$$
$$\mathrm{P}\big(\bar{X}_n > \mu_0 + z_a\mathcal{S}_{\bar{X}} \mid \mu_x = \mu_0\big) \approx a.$$

These formulas justify the procedures for inference about μ used in practice, which we summarize next:

Procedures for inference about μ.

Interval estimator endpoints, confidence level $1 - a$: $\bar{X}_n \pm z_{a/2}\mathcal{S}_{\bar{X}}$.

Critical value for one-tailed hypothesis test of $\mu = \mu_0$, significance level a: $\mu_0 \pm z_a s_{\bar{X}}$.

[16] We use a instead of α here since α is one of the parameters from the regression model.

Critical values for two-tailed hypothesis test of $\mu = \mu_0$, significance level a: $\mu_0 \pm z_{a/2} s_{\bar{X}}$.

In the formulas for the critical values, $s_{\bar{X}}$ is the realization of the standard error $\mathcal{S}_{\bar{X}}$.

20.4.2 Interval estimators and hypothesis tests for β

With this background, the inference procedures for the regression parameters are not hard to describe. We first consider the slope parameter β, which describes how increasing the x value affects the expected y value. For instance, if we are studying the relationship between education and income in a population, we likely care most about the effect of additional education on income prospects. In this case, the parameter β describes the effect of an additional year of education on average income.

Our point estimator for the slope parameter β is the random variable B. It is unbiased, has variance

$$\mathrm{Var}(B) = \frac{\sigma^2}{(n-1)s_x^2},$$

and is approximately normally distributed. Thus the same logic leading to equations (20.10) and (20.11) here leads to

$$\text{(20.12)} \qquad \mathrm{P}\left(\beta \in \left[B - z_{a/2}\frac{\sigma}{\sqrt{n-1}\,s_x}, B + z_{a/2}\frac{\sigma}{\sqrt{n-1}\,s_x}\right]\right) \approx 1 - a \quad \text{and}$$

$$\text{(20.13)} \qquad \mathrm{P}\left(B > \beta_0 + z_a\frac{\sigma}{\sqrt{n-1}\,s_x} \;\middle|\; \beta = \beta_0\right) \approx a.$$

If the conditional variance σ^2 were known, these formulas would define interval estimators and hypothesis tests for β. Of course, the conditional variance σ^2 is typically not known, so we estimate it using the sample conditional variance $\mathcal{S}^2$.

To simplify the formulas to come, we let

$$\text{(20.14)} \qquad \mathcal{S}_B = \frac{\mathcal{S}}{\sqrt{n-1}\,s_x}$$

denote the **standard error** of β. It is our estimator for the standard deviation of B, obtained by substituting $\mathcal{S}^2$ for σ^2 in the formula for Var(B), and then taking the square root. Substituting into equations (20.12) and (20.13), we obtain

$$\mathrm{P}\left(\beta \in \left[B - z_{a/2}\mathcal{S}_B, B + z_{a/2}\mathcal{S}_B\right]\right) \approx 1 - a \quad \text{and}$$
$$\mathrm{P}\left(B > \beta_0 + z_a\mathcal{S}_B \mid \beta = \beta_0\right) \approx a.$$

These equations yield the following procedures for inference about β, which are valid when the sample size n is large enough.

Procedures for inference about β.

Interval estimator endpoints, confidence level $1 - a$: $B \pm z_{a/2}\mathcal{S}_B$.

Critical value for one-tailed hypothesis test of $\beta = \beta_0$, *significance level* a: $\beta_0 \pm z_a s_B$.

Critical values for two-tailed hypothesis test of $\beta = \beta_0$, *significance level* a: $\beta_0 \pm z_{a/2} s_B$.

In the formulas for the critical values, s_B is the realization of the standard error $\mathcal{S}_B$.

■ Example *Pricing kegs.*

Our microbrewer estimated the relationship between the price of a keg of Trial by Hops (x) and the number of kegs sold per week by each distributor (y) as $y = A + Bx = 93.63 - .8080x$. Thus the brewer estimates that increasing the price of a keg by one dollar reduces each distributor's expected weekly sales by .8080 kegs.

To get a sense of the precision of this point estimate, we can construct an interval estimate of B with confidence level .95. To do so, we need to know the number of trials, $n = 33$, the variance of the prices chosen by the brewer, $s_x^2 = 41.205$, and the sample conditional variance, $\mathcal{S}^2 = 30.2237$. From these we can compute the standard error of β:

$$\mathcal{S}_B = \frac{\mathcal{S}}{\sqrt{n-1}\, s_x} = \frac{\sqrt{30.2237}}{\sqrt{32}\sqrt{41.205}} = .1514.$$

Using this standard error, we find that the .95 interval estimate for β has endpoints

$$B \pm z_{.025} s_B = -.8080 \pm 1.96 \times .1514 = -.8080 \pm .2967.$$

The interval itself is $[-1.1047, -.5113]$. This interval has the usual interpretation: it was generated by a procedure that captures the parameter β in 95% of possible samples.

The brewer is planning on raising keg prices by one dollar unless he is convinced that if he does so, each distributor's expected weekly keg sales per will fall by more than .75. He therefore considers the following null and alternative hypotheses:

$$H_0 : \beta = -.75,$$
$$H_1 : \beta < -.75.$$

With these hypotheses, rejection of the null is strong evidence that β is less than $-.75$.

To test the null hypothesis against the alternative at a 5% significance level, the brewer computes the critical value

$$\beta_0 - z_a s_B = -.75 - 1.645 \times .1514 = -.9991.$$

Since $B = -.8080 > -.9991$, he does not reject the null hypothesis. With this evidence in hand, the brewer raises the price. ■

■ Example *Greed and wrath.*

Experimental economists study decision making in a laboratory environment with the aim of confirming, refuting, and refining theoretical models of economic behavior. Many experiments consider versions of the *ultimatum game*. There are two subjects, a proposer and a responder, who have the opportunity to split a fixed amount of money. The proposer offers the responder a fraction of the cash—for instance, he may offer her 30%, which would mean keeping 70% for himself. The responder then chooses between accepting this offer or turning it down; in the latter case, both subjects go home empty-handed.

If the responder only cared about money, it would make sense for her to accept any positive offer. In reality, responders who feel they are being treated unfairly will turn down positive offers, preferring to take revenge instead of cash. One might expect the size of the stake to matter here. It is easy to turn down 30% of one dollar; it is not as easy to turn down 30% of a thousand dollars.[17]

We run ultimatum game experiments on randomly selected U.S. undergraduates, varying the stakes from \$1 to \$1000. After the stake s is announced, we ask the responder to tell us (but not the proposer) the minimum fraction of the stake she is willing to accept. Then the proposer makes his offer, and we divide the stake or pay nothing according to whether or not the offer meets the responder's threshold.

Suppose we believe that this experiment satisfies the assumptions of the classical regression model, with y representing the responder's threshold, and $x = \log s$ the (base 10) logarithm of the stake. Under this logarithmic transformation of the stake, any given value of x corresponds to a stake of $s = 10^x$ dollars, so that $x = 0$ corresponds to a stake of \$1, $x = 1$ to \$10, $x = 2$ to \$100, and $x = 3$ to \$1000.[18]

[17] The first experiment on the ultimatum game is Werner Güth, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3 (1982), 367–388. Studies of the effects of stakes on behavior include Lisa A. Cameron, "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Economic Inquiry* 37 (1999), 47–59, and Steffen Andersen et al., "Stakes Matter in Ultimatum Games," *American Economic Review* 101 (2011), 3427–3439.

[18] If this feels like a good moment to refresh your memories of log transformations, see Section 11.3.2.

Condition (C2) requires that $\mathrm{E}(Y_i) = \alpha + \beta x_i$ for some unknown parameters α and β. Because $x = \log s$, the slope parameter β represents the change in the expected threshold when the stake is increased by a factor of 10 from any initial level.

We run 40 trials, choosing a range of stakes such that $\bar{x} = 1.35$ and $s_x^2 = .9513$. Evaluating the results, we obtain the OLS estimates $A = .4532$ and $B = -.08295$ and the sample conditional variance $\mathcal{S}^2 = .01116$.

What is the .95 interval estimate for β? The standard error of β can be computed as

$$\mathcal{S}_B = \frac{\mathcal{S}}{\sqrt{n-1}\, s_x} = \frac{\sqrt{.01116}}{\sqrt{39}\sqrt{.9513}} = .01735.$$

Thus the .95 interval estimate for β has endpoints

$$B \pm z_{.025}\mathcal{S}_B = -.08295 \pm 1.96 \times .01735 = -.08295 \pm .03401.$$

and so the interval is $[-.11696, -.04894]$.

Suppose we would like to test the null hypothesis that doubling the stake lowers the expected threshold by .03, doing so against a two-sided alternative hypothesis at a 5% significance level. To accomplish this, observe that since $\log 2s = \log 2 + \log s$, doubling the stake s corresponds to increasing $x = \log s$ by $\log 2 \approx .3010$. Thus for doubling the stake to lower the expected threshold by .03, β must satisfy $\beta \log 2 = -.03$, implying that $\beta = -.03/\log 2 = -.0997$. Our null and alternative hypotheses are thus

$$H_0 : \beta = -.0997,$$
$$H_1 : \beta \neq -.0997.$$

The critical values for the hypothesis test are

$$c_{\pm} = \beta_0 \pm z_{.025} s_B = -.0997 \pm 1.96 \times .01735 = -.0997 \pm .03401,$$

so that $c_- = -.1337$ and $c_+ = -.0657$. Since $B = -.08295$ lies between these values, we do not reject the null hypothesis. (Actually, we could have anticipated this conclusion, since $-.0997$ was inside the .95 interval estimate for β—see Section 16.4.) ■

20.4.3 Interval estimators and hypothesis tests for conditional means

The other quantities we commonly want to estimate are the conditional means $\alpha + \beta x$ for given values of x. In an experiment to evaluate demand, $\alpha + \beta x$ is the expected quantity sold when the price is set at x. In a study of the relation between education and income, it is the mean income level in the subpopulation with education level x.

We saw in Section 20.2.3 that the estimator for the subpopulation mean, $A + Bx$, is unbiased ($\mathrm{E}(A + Bx) = \alpha + \beta x$), that its variance is

$$\mathrm{Var}(A + Bx) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right),$$

and that it is approximately normally distributed. Knowing the sample conditional variance $\mathcal{S}^2$, we can estimate the standard deviation of $A + Bx$ using the **standard error** of $A + Bx$, defined by

$$\mathcal{S}_{A+Bx} = \mathcal{S} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}. \tag{20.15}$$

Following the logic from the previous section, we find that

$$\mathrm{P}\big(\alpha + \beta x \in \big[A + Bx - z_{a/2}\mathcal{S}_{A+Bx}, A + Bx + z_{a/2}\mathcal{S}_{A+Bx}\big]\big) \approx 1 - a, \quad \text{and} \tag{20.16}$$

$$\mathrm{P}\big(A + Bx > m_0 + z_a \mathcal{S}_{A+Bx} \mid \alpha + \beta x = m_0\big) \approx a. \tag{20.17}$$

These facts are summarized in the following inference procedures for conditional means, which again are valid when the sample size n is large enough.

Procedures for inference about $\alpha + \beta x$.

Interval estimator endpoints, confidence level $1 - a$*:* $A + Bx \pm z_{a/2}\mathcal{S}_{A+Bx}$.

Critical value for one-tailed hypothesis test of $\alpha + \beta x = m_0$*, significance level* a*:* $m_0 \pm z_a s_{A+Bx}$.

Critical values for two-tailed hypothesis test of $\alpha + \beta x = m_0$*, significance level* a*:* $m_0 \pm z_{a/2} s_{A+Bx}$.

In the formulas for the critical values, s_{A+Bx} is the realization of the standard error $\mathcal{S}_{A+Bx}$.

Example *How many kegs?*

If the brewer of Trial by Hops decides to charge \$86 per keg, what can we say about the expected number of kegs sold weekly by each distributor? The expected number of kegs sold at a price of \$86 is $\alpha + 86\beta$. Using the OLS estimates, we obtain a point estimate of the expected number of kegs sold:

$$A + 86B = 93.62585 + 86 \times (-.80805) = 24.1333.$$

To compute a 95% confidence interval for $\alpha + 86\beta$, we first compute the standard error of this conditional mean. Using the facts that $n = 33$, $\bar{x} = 80.727$, $s_x^2 = 41.205$, and $\mathcal{S}^2 = 30.224$, we obtain

$$\mathcal{S}_{A+86B} = \mathcal{S} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}} = \sqrt{30.224} \times \sqrt{\frac{1}{33} + \frac{(86 - 80.727)^2}{32 \times 41.205}} = 1.246.$$

Thus the 95% confidence interval for the conditional mean has endpoints

$$A + 86B \pm z_{.025}\mathcal{S}_{A+86B} = 24.1333 \pm 1.96 \times 1.246 = 24.1333 \pm 2.4427,$$

so the interval itself is [21.6906, 26.5760].

What if we considered a price of \$95 a keg? The point estimate of the expected number of kegs sold becomes

$$A + 95B = 93.62585 + 95 \times (-.80805) = 16.8611,$$

with standard error

$$\mathcal{S}_{A+95B} = \mathcal{S}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}} = \sqrt{30.224} \times \sqrt{\frac{1}{33} + \frac{(95-80.727)^2}{32 \times 41.205}} = 2.363.$$

The endpoints of the 95% confidence interval for $\alpha + 95\beta$ are thus

$$A + 95B \pm z_{.025}\mathcal{S}_{A+95B} = 16.8611 \pm 1.96 \times 2.363 = 16.8611 \pm 4.6315,$$

so the interval itself is [12.2296, 21.4926].

Notice that the confidence interval when $x = 95$ is considerably wider than the confidence interval when $x = 86$. Because 95 is further than 86 from the mean $\bar{x} = 80.727$, the estimator of the conditional mean has a greater variance in the former case than the latter, resulting in the wider confidence interval. ■

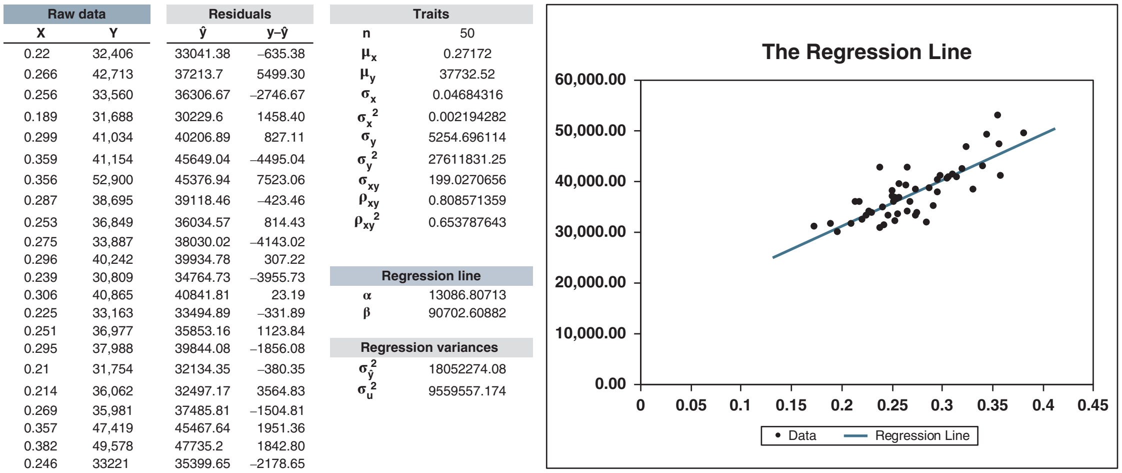
Excel calculation: *Standard errors*

The `regression_inference.xlsx` workbook computes standard errors for the slope β and for conditional means $\alpha + \beta x$ automatically. Figure 20.4 shows that for the Trial by Hops data, the standard error of β is $\mathcal{S}_B = .1514$. To obtain the standard error of $\alpha + \beta x$, you must enter the value of x of interest in the light blue cell. In the workbook shown in Figure 20.4, we entered $x = 86$; the workbook reports a standard error of $\mathcal{S}_{A+Bx} = 1.246$.

20.4.4 Population regressions vs. sample regressions

The previous and current chapters have used regression in two different contexts for two different purposes. In Chapter 19, we computed the population regression line $y = \alpha + \beta x$ for data describing a population. There, the slope β and intercept α were descriptive statistics; they conveniently summarized linear associations in the population data. In this chapter, we've computed the sample regression line $y = A + Bx$ for data obtained from a sample. In the case of random sampling from a population, the sample regression line provides an estimate of the population

Figure 20.5: An appropriate analysis of population data using `regression_descriptive.xlsx`.



Raw data		Residuals	
X	Y	ŷ	y–ŷ
0.22	32,406	33041.38	–635.38
0.266	42,713	37213.7	5499.30
0.256	33,560	36306.67	–2746.67
0.189	31,688	30229.6	1458.40
0.299	41,034	40206.89	827.11
0.359	41,154	45649.04	–4495.04
0.356	52,900	45376.94	7523.06
0.287	38,695	39118.46	–423.46
0.253	36,849	36034.57	814.43
0.275	33,887	38030.02	–4143.02
0.296	40,242	39934.78	307.22
0.239	30,809	34764.73	–3955.73
0.306	40,865	40841.81	23.19
0.225	33,163	33494.89	–331.89
0.251	36,977	35853.16	1123.84
0.295	37,988	39844.08	–1856.08
0.21	31,754	32134.35	–380.35
0.214	36,062	32497.17	3564.83
0.269	35,981	37485.81	–1504.81
0.357	47,419	45467.64	1951.36
0.382	49,578	47735.2	1842.80
0.246	33221	35399.65	–2178.65

Traits	
n	50
μ_x	0.27172
μ_y	37732.52
σ_x	0.04684316
σ_x^2	0.002194282
σ_y	5254.696114
σ_y^2	27611831.25
σ_{xy}	199.0270656
ρ_{xy}	0.808571359
ρ_{xy}^2	0.653787643

Regression line	
α	13086.80713
β	90702.60882

Regression variances	
$\sigma_{\hat{y}}^2$	18052274.08
σ_u^2	9559557.174

regression line $y = \alpha + \beta x$ that we would have obtained if we had access to the complete population data.

Population regressions and sample regressions are different animals. Nevertheless, it is a common error to treat population data as though it were data from a sample. What are the consequences of doing so?

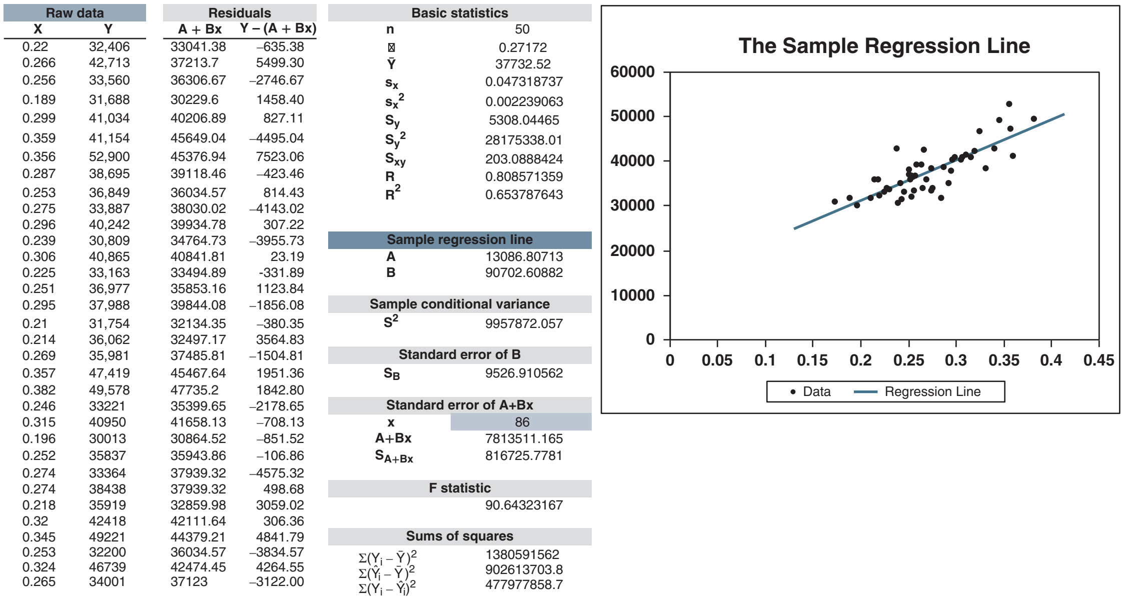
The worksheet `ch19_data.xlsx/states` presents the percentage of residents who are college graduates and per capita income in each of the 50 U.S. states in 2009.[19] Since this data set completely describes its population, the appropriate use of regression here is to obtain descriptive statistics. Figure 20.5, which presents the output of `regression_descriptive.xlsx` for this data set, shows that the regression line for the data set is $y = \alpha + \beta x = 13{,}087 + 90{,}703x$. Thus, to summarize the relationship between educational attainment and income in the 50 states, we can say that an increase in the percentage of residents who are college graduates of one point is associated with an additional \$907 in per capita income.

What would happen if we treated this population data as though it were sample data? Figure 20.6 shows the output of `regression_inference.xlsx` for the data set. Notice first that this workbook computes the "sample regression line" $y = A + Bx = 13{,}087 + 90{,}703x$. Evidently, the slope and intercept of the regression line are the same whether we treat the data as population data or sample data. This is because in either case, the regression line is the one that minimizes the sum of squared residuals.

Not all of the numbers reported in `regression_inference.xlsx` are so innocuous. For instance, the workbook reports that the standard error of B is

[19] The percentage of college graduates is among residents at least 25 years old. The data come from the U.S. Census Bureau and the Bureau of Economic Analysis of the U.S. Department of Commerce.

Figure 20.6: An inappropriate analysis of population data using `regression_inference.xlsx`.



Raw data		Residuals	
X	Y	A + Bx	Y − (A + Bx)
0.22	32,406	33041.38	–635.38
0.266	42,713	37213.7	5499.30
0.256	33,560	36306.67	–2746.67
0.189	31,688	30229.6	1458.40
0.299	41,034	40206.89	827.11
0.359	41,154	45649.04	–4495.04
0.356	52,900	45376.94	7523.06
0.287	38,695	39118.46	–423.46
0.253	36,849	36034.57	814.43
0.275	33,887	38030.02	–4143.02
0.296	40,242	39934.78	307.22
0.239	30,809	34764.73	–3955.73
0.306	40,865	40841.81	23.19
0.225	33,163	33494.89	-331.89
0.251	36,977	35853.16	1123.84
0.295	37,988	39844.08	–1856.08
0.21	31,754	32134.35	–380.35
0.214	36,062	32497.17	3564.83
0.269	35,981	37485.81	–1504.81
0.357	47,419	45467.64	1951.36
0.382	49,578	47735.2	1842.80
0.246	33221	35399.65	–2178.65
0.315	40950	41658.13	–708.13
0.196	30013	30864.52	–851.52
0.252	35837	35943.86	–106.86
0.274	33364	37939.32	–4575.32
0.274	38438	37939.32	498.68
0.218	35919	32859.98	3059.02
0.32	42418	42111.64	306.36
0.345	49221	44379.21	4841.79
0.253	32200	36034.57	–3834.57
0.324	46739	42474.45	4264.55
0.265	34001	37123	–3122.00

Basic statistics	
n	50
⊠	0.27172
$\bar{Y}$	37732.52
s_x	0.047318737
s_x^2	0.002239063
S_y	5308.04465
S_y^2	28175338.01
S_{xy}	203.0888424
R	0.808571359
R^2	0.653787643
Sample regression line	
A	13086.80713
B	90702.60882
Sample conditional variance	
S^2	9957872.057
Standard error of B	
S_B	9526.910562
Standard error of A+Bx	
x	86
A+Bx	7813511.165
S_{A+Bx}	816725.7781
F statistic	
	90.64323167
Sums of squares	
$\Sigma(Y_i - \bar{Y})^2$	1380591562
$\Sigma(\hat{Y}_i - \bar{Y})^2$	902613703.8
$\Sigma(Y_i - \hat{Y}_i)^2$	477977858.7

$\mathcal{S}_B = 9527$. If the data actually came from a sample, and the assumptions of the classical regression model were satisfied, then this number would be an estimate of the standard deviation of the OLS estimator B. It would indicate the likely size of the error in our estimate of the unknown parameter β, and so could be used to construct interval estimates and run hypothesis tests for β.

But our data is not from a sample—it is a complete account of the variables in question in all 50 states. The slope of the regression line, 90,703, is not an estimate of β; it *is* β. There is no random sample, and nothing is being estimated, so there is no role for a standard error. The output $\mathcal{S}_B = 9527$ computed by the workbook is meaningless.

Most commercial regression packages are designed with sample data in mind. There is nothing preventing a user from running these packages on population data, and then using the results to construct "confidence intervals," or to conduct "hypothesis tests." To a novice, an analysis of population data that "rejects a zero-slope null hypothesis at significance level .05" may sound impressively scientific. But when you have complete population data, applying the tools of statistical inference is not even wrong—it makes no sense at all.

20.5 Small Samples and the Classical Normal Regression Model

In Chapter 17, we introduced techniques for drawing inferences about an unknown mean based on small samples. The procedures required the additional assumption that each trial follow a normal distribution, and led to inference procedures based on the t distribution. Similar ideas apply in the context of regression inference, provided that one begins with a regression model that invokes a suitable normality assumption.

20.5.1 The classical normal regression model

In order to perform small sample inference in regression, we need to assume that conditional on the choice of the corresponding x variable, each y observation is drawn from a normal distribution. We therefore add this assumption to the classical regression model.

The classical normal regression model.

(N1)	*Fixed x sampling*:	$x_1, \dots, x_n$ *are fixed and not all identical;* $Y_1, \dots, Y_n$ *are independent random variables.*
(N2)	*Linearity of conditional means*:	$\mathrm{E}(Y_i) = \alpha + \beta x_i$.
(N3)	*Constant conditional variances*:	$\mathrm{Var}(Y_i) = \sigma^2$.
(N4)	*Conditional normality*:	Y_i *is normally distributed.*

One can likewise define the *random sampling normal regression model* by adding a conditional normality assumption to our earlier random sampling model, with similar consequences for small-sample inference—see Section 20.A.1.

The classical normal regression model completely specifies the distribution of observation Y_i in terms of the choice of subpopulation x_i and the model's parameters: succinctly, it requires that $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. Of course, the assumptions about the mean and variance of Y_i are the same as before; only the normality assumption is new.

■ Example *The classical normal regression model in pictures.*

To illustrate the assumptions of the classical normal regression model in the context of an experiment, we consider a firm whose choice of advertising expenditures (x, in millions of dollars) influences but does not completely determine its sales volume (y, in millions of units).

Figures 20.7 and 20.8 present possible conditional distribution of sales volumes y for each choice of advertising expenditures x. The height of the surface at point (x, y) represents the density at y of the conditional distribution of Y_i given x. Thus for any choice of x, the conditional distribution of Y_i given x is captured by the cross section of the surface at x. (For instance, the conditional distribution of sales volume when advertising expenditures are 0 is described by the edge of the surface following the y axis.)

Figure 20.7 describes an environment that satisfies the assumptions of the classical normal regression model, with conditional mean function $y = 2.5 + .2x$ and conditional variance $\sigma^2 = .49$. For each level of expenditures x, the conditional distribution of sales volume is described by the bell-shaped curve of the normal

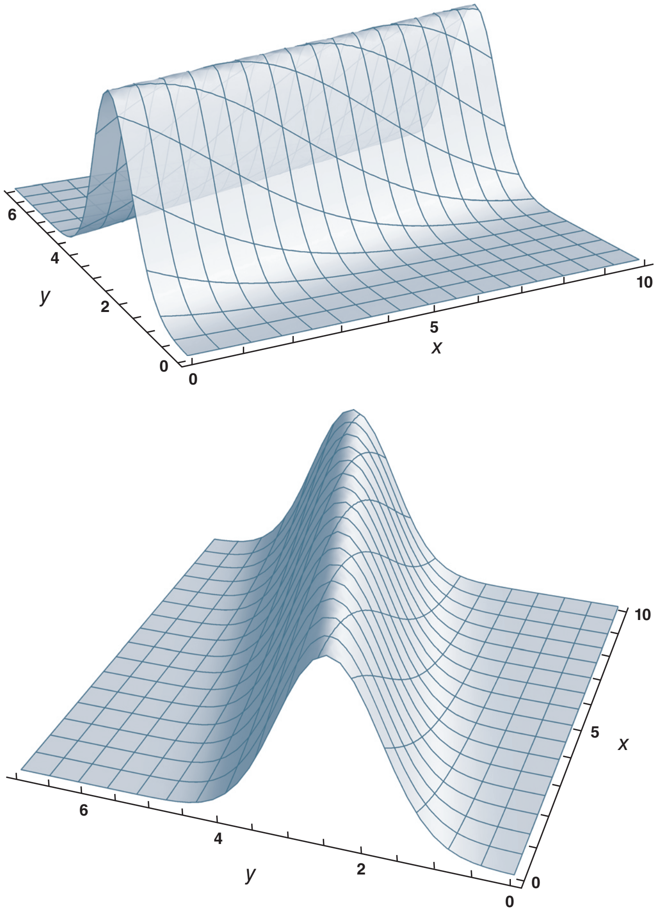


Figure 20.7: A collection of conditional distributions satisfying the classical normal regression model (two viewpoints).

distribution, as specified in (N4). Since the mean of a normal distribution corresponds to the highest point on its density function, the fact that the highest points as *x* varies lie on a straight line tells us that the conditional expectation function is linear, as required by (N2). Finally, the fact that the bell-shaped curve for each *x* value has the same "width" reflects that all conditional variances are equal, as required by (N3).

For a counterpoint, Figure 20.8 illustrates an environment in which all conditional distributions are normal but the other two assumptions about conditional distributions fail. Looking at the highest points on the bell-shaped curves, we see that increasing advertising expenditures increases expected sales volumes quickly when advertising expenditures are low, but more slowly when expenditures are high. This violates (N2). Also, the bell-shaped curves along each *x* cross section are tall and thin when *x* is small, but shorter and fatter when *x* is large, reflecting that the variance in sales volumes increases as advertising expenditures grow. This violates (N3).

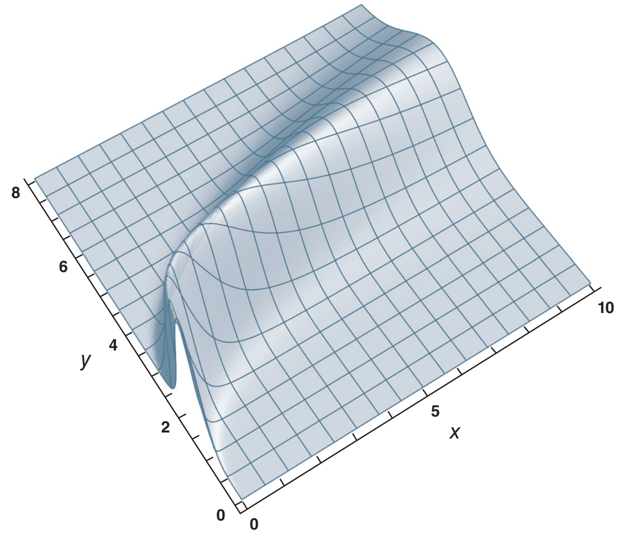


Figure 20.8: A collection of normal conditional distributions that fail (N2) and (N3).

20.5.2 Interval estimators and hypothesis tests for β

Chapter 17 introduced our small-sample inference procedures for i.i.d. normal trials $\{X_i\}_{i=1}^n$. Because the trials are normal and independent, their sample mean $\bar{X}_n = \sum_{i=1}^n X_i$ is normally distributed as well. In Section 17.A.4, we used this and other facts to derive the key result behind our inference procedures: again letting $\mathcal{S}_{\bar{X}} = \frac{\mathcal{S}_X}{\sqrt{n}}$ denote the standard error of $\bar{X}_n$, we have that

$$\frac{\bar{X}_n - \mu}{\mathcal{S}_{\bar{X}}} \sim t(n-1). \tag{20.18}$$

That is, the left-hand side of (20.18), called the t-statistic for μ for i.i.d. normal trials, follows a t distribution with $n-1$ degrees of freedom.

Analogous results hold for the classical normal regression model. First, since B is a linear function of the observations Y_i, and since the Y_i are independent normal random variables, B is normally distributed.[20] Second, using a more involved version of the arguments from Section 17.A.4, we can establish the following analogue of fact (20.18):

$$\frac{B - \beta}{\mathcal{S}_B} \sim t(n-2). \tag{20.19}$$

Here $\mathcal{S}_B$ is the standard error for β defined in (20.14). The left-hand side of (20.19) is called the ***t*-statistic** for β in the classical normal regression model.

[20] For a reminder of why this is so, see Sections 6.4.1 and 6.6.1.

Fact (20.19) says that this statistic follows a t distribution, this time with $n - 2$ degrees of freedom.

Fact (20.19) is the basis for our small sample inference procedures for the classical normal regression model. Concerning the procedures themselves, the main difference from Section 20.4.2 is the replacement of z-values with t-values from the $t(n-2)$ distribution.

Procedures for estimating β (small samples).

In the classical normal regression model, for any sample size n:

Interval estimator endpoints, confidence level $1 - a$*:* $B \pm t^{n-2}_{a/2} \mathcal{S}_B$.

Critical value for one-tailed hypothesis test of $\beta = \beta_0$*, significance level a:* $\beta_0 \pm t^{n-2}_{a} s_B$.

Critical values for two-tailed hypothesis test of $\beta = \beta_0$*, significance level a:* $\beta_0 \pm t^{n-2}_{a/2} s_B$.

■ Example *Customer service.*

A technology industry analyst is studying the relationship between wages paid to customer service representatives and customer satisfaction for Silicon Valley technology firms. For a random sample of 16 firms, she obtains the average wages (x, in dollars) and the average score on a customer satisfaction survey (y, on a 100-point scale). The analyst believes that the data for the population as a whole is well approximated by the classical normal regression model for some unknown parameters α, β, and σ^2.

The results of the sample and the output of the `regression_inference.xlsx` workbook are presented in Figure 20.9. The OLS estimate of the regression line is $y = A + Bx = -1.920 + 3.424x$. Thus, the analyst estimates that in the population as a whole, a one-dollar increase in average wage is associated with a 3.4-point increase in customer satisfaction ratings. The variance of the wages is $s_x^2 = 6.197$, and the sample conditional variance is $\mathcal{S}^2 = 35.356$.

To obtain a .95 confidence interval for β, the analyst computes the standard error of β:

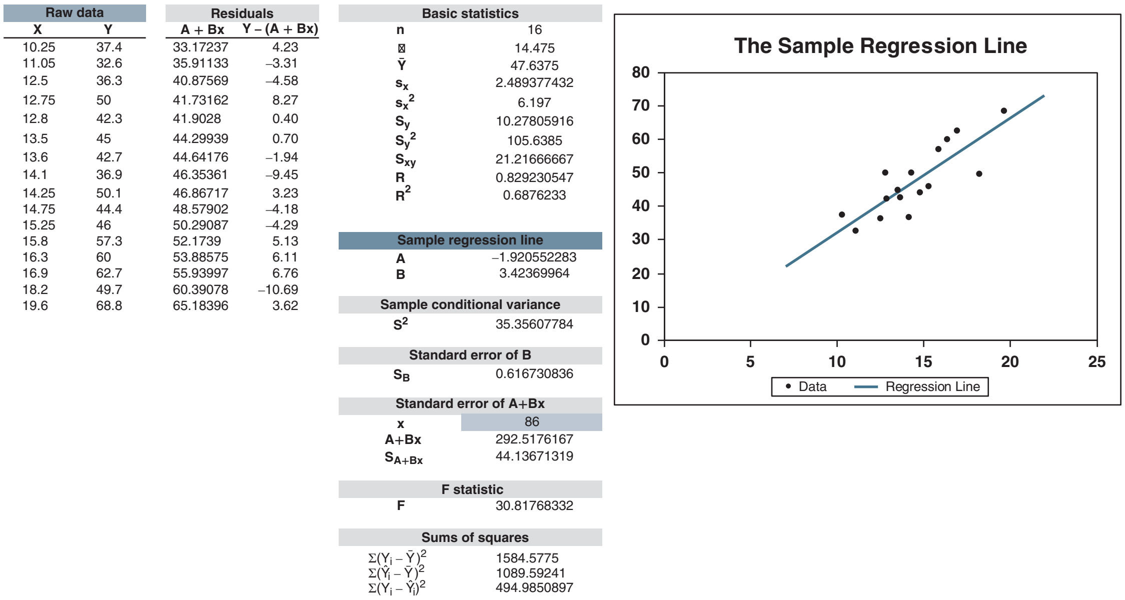
$$\mathcal{S}_B = \frac{\mathcal{S}}{\sqrt{n-1}\, s_x} = \frac{\sqrt{35.356}}{\sqrt{15}\sqrt{6.197}} = .6167.$$

(She also could have read this directly from the workbook.) Using the `distributions.xls` workbook, or inputting “`=T.INV(.975,14)`” into a blank workbook cell, the analyst obtains the t-value $t^{14}_{.025} = 2.145$. Thus the .95 interval estimate for β has endpoints

$$B \pm t^{14}_{.025} \mathcal{S}_B = 3.424 \pm 2.145 \times .6167 = 3.424 \pm 1.323,$$

and so the interval is $[2.101, 4.747]$.

Figure 20.9: Regressing customer satisfaction on wages of customer service representatives using `regression_inference.xlsx`.



Raw data		Residuals	
X	Y	A + Bx	Y – (A + Bx)
10.25	37.4	33.17237	4.23
11.05	32.6	35.91133	–3.31
12.5	36.3	40.87569	–4.58
12.75	50	41.73162	8.27
12.8	42.3	41.9028	0.40
13.5	45	44.29939	0.70
13.6	42.7	44.64176	–1.94
14.1	36.9	46.35361	–9.45
14.25	50.1	46.86717	3.23
14.75	44.4	48.57902	–4.18
15.25	46	50.29087	–4.29
15.8	57.3	52.1739	5.13
16.3	60	53.88575	6.11
16.9	62.7	55.93997	6.76
18.2	49.7	60.39078	–10.69
19.6	68.8	65.18396	3.62

Basic statistics	
n	16
⊠	14.475
$\bar{Y}$	47.6375
s_x	2.489377432
s_x^2	6.197
S_y	10.27805916
S_y^2	105.6385
S_{xy}	21.21666667
R	0.829230547
R^2	0.6876233

Sample regression line	
A	–1.920552283
B	3.42369964

Sample conditional variance	
S^2	35.35607784

Standard error of B	
S_B	0.616730836

Standard error of A+Bx	
x	86
A+Bx	292.5176167
S_{A+Bx}	44.13671319

F statistic	
F	30.81768332

Sums of squares	
$\Sigma(Y_i - \bar{Y})^2$	1584.5775
$\Sigma(\hat{Y}_i - \bar{Y})^2$	1089.59241
$\Sigma(Y_i - \hat{Y}_i)^2$	494.9850897

The analyst would like to test the null hypothesis that β is 4.5 against the alternative that it is smaller:

$$H_0 : \beta = 4.5$$
$$H_1 : \beta < 4.5.$$

The critical value for a hypothesis test with significance level .05 is

$$\beta_0 - t^{14}_{.05} s_B = 4.5 - 1.761 \times .6167 = 3.414.$$

Since $B = 3.424 > 3.414$, she does not reject the null hypothesis at this significance level.

If instead the analyst uses a significance level of .10, she obtains a critical value of

$$\beta_0 - t^{14}_{.10} s_B = 4.5 - 1.345 \times .6167 = 3.671.$$

Since $B = 3.424 < 3.671$, she rejects the null hypothesis at this less demanding significance level. ■

■ Example *Sunny days and solar power.*

Insolation, a measure of the average daily amount of solar radiation at a given location over the course of a year, describes how much solar energy is available for conversion into electrical energy by solar panels. Of course, the amount of electrical energy obtained also depends on the size and quality of the panel.

Insolation in Honolulu is 6.02 kWh/(m$^2\cdot$ day) (kilowatt hours per square meter per day). But since the City and County of Honolulu encompasses the entire island of Oahu, this number masks a great deal of variation.

An Oahu-based property manager is considering adding SolarPro solar panels to the rooftops of all of the properties she manages. To test the quality of the panels, she installs them on 11 of her properties. She looks up the insolation at each property's address and measures the panels' average daily electrical output per square meter over the course of a year. Her data is presented in the table below.

insolation ($\frac{\text{kWh}}{\text{m}^2\cdot\text{day}}$)	electrical output ($\frac{\text{kWh}}{\text{m}^2\cdot\text{day}}$)
5.91	1.339
5.91	1.414
5.92	1.372
6.10	1.411
6.05	1.384
6.45	1.508
5.62	1.294
6.27	1.457
6.49	1.457
5.60	1.250
6.35	1.461

The manager believes that the performance of the solar panels satisfies the assumptions of the classical normal regression model. The parameter β is the expected increase in electrical output resulting from a unit increase in insolation, and so is a measure of the panels' efficiency. Since insolation and electrical output are measured in the same units, β is unit free.

The property manager only wants to buy the panels if she is convinced that the efficiency is at least .20. She thus considers the following null and alternative hypotheses:

$$H_0 : \beta = .20$$
$$H_1 : \beta > .20$$

With these hypotheses, rejecting the null is strong evidence that the efficiency is as high as the manager would like.

After some number crunching, the manager obtains the following statistics:

$$A = -.05130,\ B = .2387,\ \mathcal{S}^2 = .0007878,\ \bar{x} = 6.0609,\ s_x^2 = .09367.$$

Thus the standard error for β is

$$\mathcal{S}_B = \frac{\mathcal{S}}{\sqrt{n-1}\,s_x} = \frac{\sqrt{.0007878}}{\sqrt{10}\sqrt{.09367}} = .02900.$$

The critical value for a hypothesis test with significance level .05 is

$$\beta_0 + t^9_{.05} s_B = .20 + 1.833 \times .02900 = .253.$$

Since $B = .2387 < .253$, the manager does not reject the null hypothesis and so does not buy the panels.

The manager's point estimate of the panels' efficiency, $B = .2387$, is closer to .25 than to .20. However, given her small sample, this evidence was not enough to reject the null hypothesis. If the same values of B, $\mathcal{S}^2$, and s_x^2 were obtained with a larger sample, then the standard error $\mathcal{S}_B$ would have been smaller and might have led her to reject the null hypothesis in favor of the alternative—see Exercise 20.5.8. ■

20.5.3 Interval estimators and hypothesis tests for conditional means

The normality assumption (N4) also allows us to perform small-sample inference about conditional means $\alpha + \beta x$. The reasoning is similar to that above. Since the estimator $A + Bx$ is a linear function of the observations Y_i, and since the Y_i are independent, $A + Bx$ is normally distributed. In this case, the ***t*-statistic** and its distribution are

$$\frac{(A + Bx) - (\alpha + \beta x)}{\mathcal{S}_{A+Bx}} \sim t(n-2),$$

where $\mathcal{S}_{A+Bx}$ is the standard error for $\alpha + \beta x$ defined in (20.15). This fact can be used to derive the small-sample procedures below, which differ from those in Section 20.4.3 in the replacement of z-values with t-values from the $t(n-2)$ distribution.

Procedures for inference about $\alpha + \beta x$ (small samples).

In the classical normal regression model, for any sample size n:

Interval estimator endpoints, confidence level $1 - a$: $\quad A + Bx \pm t^{n-2}_{a/2} \mathcal{S}_{A+Bx}$.

Critical value for one-tailed hypothesis test of $\alpha + \beta x = m_0$, *significance level a:* $\quad m_0 \pm t^{n-2}_{a} s_{A+Bx}$.

Critical values for two-tailed hypothesis test of $\alpha + \beta x = m_0$, *significance level a:* $\quad m_0 \pm t^{n-2}_{a/2} s_{A+Bx}$.

■ Example *Sunny days revisited.*

The property manager on Oahu is considering installing solar panels at a location with insolation 6.10. What is the 95% confidence interval for the expected electrical output?

The expected electrical output is $\alpha + 6.10\beta$, which we estimate as

$$A + 6.10B = -.05130 + 6.10 \times .2387 = 1.4045.$$

The standard error of $\alpha + 6.10\beta$ is

$$\begin{aligned} \mathcal{S}_{A+6.10B} &= \mathcal{S}\sqrt{\frac{1}{n} + \frac{(6.10 - \bar{x})^2}{(n-1)s_x^2}} \\ &= \sqrt{.0007878} \times \sqrt{\frac{1}{11} + \frac{(6.10 - 6.0609)^2}{10 \times .09367}} \\ &= .008538. \end{aligned}$$

The 95% confidence interval has endpoints

$$A + 6.10B \pm t^9_{.025}\mathcal{S}_{A+6.10B} = 1.4045 \pm 2.262 \times .008538 = 1.4045 \pm .0193.$$

Thus the 95% confidence interval is [1.3852, 1.4238]. ■

20.5.4 Prediction intervals*

The classical normal regression model provides a setting for answering a new sort of question, one of prediction. As motivation, suppose that a monopolist is running an experiment to estimate how demand in its market varies with price. She believes that this relationship is as described in the classical normal regression model: given a price x_i, the quantity demanded Y_i is normally distributed with mean $\alpha + \beta x_i$ and variance σ^2. The parameters α, β, and σ^2 are unknown, but the monopolist can run an experiment with n trials, $\{(x_i, Y_i)\}_{i=1}^n$, in order to estimate them.

Suppose that after the experiment, the monopolist sets a price of x. This will lead to a quantity sold of Y, a normally distributed random variable with mean $\alpha + \beta x$ and variance σ^2. From our previous analyses, we know that the monopolist can estimate the *expected* quantity sold, $\alpha + \beta x$, using the estimator $A + Bx$. This estimator is unbiased and, under assumption (N4), normally distributed.

Now we want to consider a trickier question: Can the monopolist specify a random interval that provides a probabilistic statement about the *actual quantity* she will sell?

Random intervals that provide probabilistic statements about future observations are known as **prediction intervals**. Prediction intervals must account for two distinct sources of randomness. First, the estimate of the conditional mean

$\alpha + \beta x$ using the estimator $A + Bx$ comes with some dispersion. This dispersion is described by its variance, which we saw in equation (20.6) is

$$\text{Var}(A + Bx) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right). \tag{20.20}$$

In addition, the normally distributed observation Y itself will exhibit some dispersion around its mean of $\alpha + \beta x$. By assumption (N3), the variance of this observation is $\text{Var}(Y) = \sigma^2$.

Combining these facts with some basic properties of normal random variables will enable us to define the prediction interval. It is natural to center the interval at $A + Bx$, since this is our estimate of the expected value of Y. To determine how wide to make the interval, we need to know how far Y is likely to be from $A + Bx$. More precisely, we need to know the distribution of the difference $Y - (A + Bx)$.

To determine this distribution, we use the facts above, and one further fact that we have not mentioned: that $A + Bx$, which depends only on the initial n observations, and Y, the new observation, are independent. The mean and variance of $Y - (A + Bx)$ are thus

$$\text{E}(Y - (A + Bx)) = \text{E}(Y) - \text{E}(A + Bx) = (\alpha + \beta x) - (\alpha + \beta x) = 0;$$

$$\text{Var}(Y - (A + Bx)) = \text{Var}(Y) + \text{Var}(-(A + Bx)) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right).$$

The calculation of the variance uses the independence of $A + Bx$ and Y. Finally, since these two random variables are independent and normally distributed, their difference is normally distributed as well. Summing up, we have

$$Y - (A + Bx) \sim N\left(0, \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} + 1 \right)\right). \tag{20.21}$$

Proceeding with a typical normal distribution calculation, we can determine the endpoints of the prediction interval when σ^2 is known (Exercise 20.M.7). In the usual case in which σ^2 is unknown, we can derive the prediction interval from the fact that a suitably chosen t-statistic has a $t(n - 2)$ distribution. Instead of providing a detailed derivation, we simply state the end result.

Prediction intervals.

In the classical normal regression model, consider predicting the y *value Y of a randomly chosen individual with* x *value x. The random interval with endpoints*

$$(A + Bx) \pm t_{a/2}^{n-2} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} + 1}, \tag{20.22}$$

called the $\mathbf{(1-a)}$ ***prediction interval*** *for Y, will contain Y with probability $1 - a$.*

A straightforward calculation shows that the prediction interval (20.22) can also be expressed as

$$(A + Bx) \pm t^{n-2}_{a/2}\sqrt{\mathcal{S}^2_{A+Bx} + \mathcal{S}^2}. \tag{20.23}$$

This alternate formula is convenient when we already know the standard error $\mathcal{S}_{A+Bx}$—for instance, from the output of the `regression_inference.xlsx` workbook.

As with all of our inference procedures, the probability statements concerning prediction intervals should be understood in the ex ante sense. The random interval (20.22) contains the new observation Y with probability $1 - a$ at a time before either the random sample or Y itself is realized. The fact that the randomness is resolved in two stages—the sample is observed first, and only later is Y observed—makes prediction intervals even trickier to interpret than confidence intervals. We explore this point in Exercise 20.5.9.

■ Example *Still sunny.*

The Oahu property manager plans to install new solar panel at a location with insolation 6.10. What is the .95 prediction interval for the new panel's output?

The endpoints of the prediction interval are

$$(A + 6.10B) \pm t^9_{.025}\mathcal{S}\sqrt{\frac{1}{n} + \frac{(6.10 - \bar{x})^2}{(n-1)s_x^2} + 1}$$

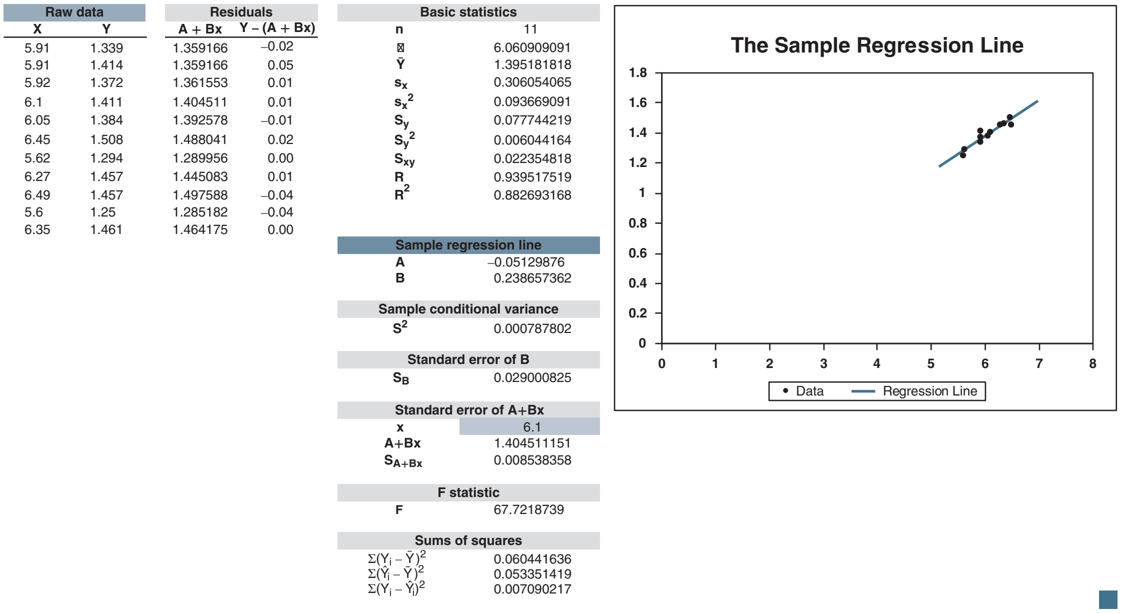
$$= 1.4045 \pm 2.262 \times \sqrt{.0007878} \times \sqrt{\frac{1}{11} + \frac{(6.10 - 6.0609)^2}{10 \times .09367} + 1}$$

$$= 1.4045 \pm .0664.$$

The .95 prediction interval is thus [1.3381, 1.4709]. Because it accounts not only for the randomness in the sample but also for the randomness in new panel's output, the prediction interval is wider—more than three times wider—than the .95 confidence interval for $\alpha + 6.10\beta$ computed earlier, whose endpoints were $1.4045 \pm .0193$.

Figure 20.10 presents output of the `regression_inference.xlsx` for the solar panel data. The values of A, B, $\mathcal{S}^2$, $\bar{x}$, s_x^2, and $\mathcal{S}_B$ reported above are all shown. To draw inferences about conditional means and to construct prediction intervals, we enter insolation level $x = 6.1$ in the blue cell in the third column. The workbook returns both the estimated conditional mean $A + 6.10B = 1.4045$ and the standard error $\mathcal{S}_{A+6.10B} = .008538$. The width of the prediction interval for a new solar panel's output can be obtained from $\mathcal{S}_{A+6.10B}$ and $\mathcal{S}^2$ using formula (20.23).

Figure 20.10: Analysis of the solar panel data using `regression_inference.xlsx`.



Raw data		Residuals	
X	Y	A + Bx	Y – (A + Bx)
5.91	1.339	1.359166	–0.02
5.91	1.414	1.359166	0.05
5.92	1.372	1.361553	0.01
6.1	1.411	1.404511	0.01
6.05	1.384	1.392578	–0.01
6.45	1.508	1.488041	0.02
5.62	1.294	1.289956	0.00
6.27	1.457	1.445083	0.01
6.49	1.457	1.497588	–0.04
5.6	1.25	1.285182	–0.04
6.35	1.461	1.464175	0.00

Basic statistics	
n	11
⊠	6.060909091
$\bar{Y}$	1.395181818
s_x	0.306054065
s_x^2	0.093669091
S_y	0.077744219
S_y^2	0.006044164
S_{xy}	0.022354818
R	0.939517519
R^2	0.882693168
Sample regression line	
A	–0.05129876
B	0.238657362
Sample conditional variance	
S^2	0.000787802
Standard error of B	
S_B	0.029000825
Standard error of A+Bx	
x	6.1
A+Bx	1.404511151
S_{A+Bx}	0.008538358
F statistic	
F	67.7218739
Sums of squares	
$\Sigma(Y_i - \bar{Y})^2$	0.060441636
$\Sigma(\hat{Y}_i - \bar{Y})^2$	0.053351419
$\Sigma(Y_i - \hat{Y}_i)^2$	0.007090217

20.6 Analysis of Residuals, R^2, and F Tests

20.6.1 Sums of squares and R^2

Section 19.4 introduced analysis of residuals of the population regression line $y = \alpha + \beta x$, which minimizes the sum of squared residuals over the entire population $\{(x_j, y_j)\}_{j=1}^N$. As we have seen, the sample regression line $y = A + Bx$ is defined in the same way, but using the results of a random sample. Because both lines are constructed in the same way, the analysis of residuals of the population regression line from Section 19.4 carries over to the sample regression line, but with one key change. The earlier equations about population data become equations that are true for all realizations of the sample. These equations are expressed in terms of the random variables that describe the sample from the ex ante point of view.

For example, one of our basic facts about the population regression residuals is the sum-of-squares equation,

$$\sum_{j=1}^{N} (y_j - \mu_y)^2 = \sum_{j=1}^{N} (\hat{y}_j - \mu_y)^2 + \sum_{j=1}^{N} (y_j - \hat{y}_j)^2. \tag{19.16}$$

This equation provided an alternate way of writing the decomposition of variance $\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_u^2$ (see (19.14)).

In the present context of the sample regression line, the sample analogue of (19.16) is true for every realization of the sample. It is therefore stated in ex ante terms, using the random variables that describe the sample.

The sum-of-squares equation (ex ante version).

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2. \tag{20.24}$$

Here $\hat{Y}_i = A + Bx_i$ is the ith sample prediction; see equation (20.9).

Our other main formula from Section 19.4 concerned the relative quality of the predictions of the regression line and the mean line, expressed in terms of the squared correlation coefficient $\rho_{x,y}^2$. We can write this formula using variances (19.10) or sums of squares (19.15):

$$\frac{\sigma_u^2}{\sigma_y^2} = \frac{\sum_{j=1}^{N}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{N}(y_j - \mu_y)^2} = 1 - \rho_{x,y}^2. \tag{19.15}$$

These equations say that the variance of prediction errors from the regression line is only $1 - \rho_{x,y}^2$ as large as the variance of prediction errors from the mean line. Thus if $\rho_{x,y}^2$ is near 1, the regression line tends to generate much more accurate predictions than the mean line (see Section 19.4.2).

To state the sample analogue of this relation, we introduce the **sample correlation coefficient**,

$$R = \frac{\mathcal{S}_{xY}}{s_x \mathcal{S}_Y}.$$

The fact that the sample analogue of (19.15) holds for every realization of the sample is expressed as

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - R^2. \tag{20.25}$$

Equation (20.25) shows that the random variable R^2, named the **squared sample correlation coefficient** (but usually called “R squared”), provides a measure of *relative quality of fit*. Specifically, it indicates how much better the sample data is fit by the sample regression line $y = A + Bx$ than by the sample mean line, $y = \bar{Y}$. An R^2 near 1 means that the regression line fits the sample data much better than the mean line, and an R^2 near 0 means that the fit of the latter and the former are nearly the same. These are the same interpretations we gave to ρ^2 earlier, but now in the context of sample data rather than population data.

Excel calculation: *Sums of squares and R^2*

The `regression_inference.xlsx` workbook computes sums of squares and R^2 automatically. Looking back at Figure 20.9, we see that for the wage/customer satisfaction data, these quantities are

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 1584.58,\ \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 1089.59,\ \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 494.99,$$

$$\text{and} \quad R^2 = .6876.$$

Notice that

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 1584.57 = 1089.59 + 494.99 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

and

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{494.99}{1584.58} = .3124 = 1 - .6876 = 1 - R^2,$$

as required by equations (20.24) and (20.25).

20.6.2 The F test for $\beta = 0$

We now describe how in the classical normal regression model, R^2 can be used to test the null hypothesis that the population regression slope β is zero against the two-sided alternative hypothesis. Specifically, large enough values of R^2 will lead us to reject the null hypothesis.

Why should such a test work? When β is equal to zero, then $\alpha = \mu_y - \beta\mu_x = \mu_y$, so the population regression line $y = \alpha + \beta x$ and the population mean line $y = \mu_y$ are identical. In this case, the sample regression line is likely to be close to the sample mean line. If this happens, both lines will fit the sample data about equally well, and so, by equation (20.25), R^2 will be close to zero. Therefore, drawing a sample whose R^2 is *not* close to zero would be an *unlikely* event if β were equal to zero. This suggests that an R^2 far from 0 should allow one to reject the null hypothesis that $\beta = 0$. In this section, we derive such a test under the assumptions of the classical normal regression model. Section 20.6.3 considers the robustness of the test to non-normal trials.

The hypothesis test for $\beta = 0$ based on R^2 makes use of the random variable

$$(n-2)\frac{R^2}{1-R^2},$$

known as the ***F*-statistic**. The test is based on the following fact, which relates the F-statistic to an F distribution (Section 17.A.5).

The *F*-statistic has an *F* distribution.

Suppose the assumptions of the classical normal regression model hold with $\beta = 0$. *Then*

$$(n-2)\frac{R^2}{1-R^2} \sim F(1, n-2). \tag{20.26}$$

For a point of comparison, recall that in Chapter 17, we defined a random variable we suggestively called the t-statistic, and based our inference procedures on the fact (that required a proof!) that this random variable has a t distribution. Here we are following similar steps, introducing a random variable suggestively called the F-statistic, and basing our inference procedures on the fact (that again requires a proof—see below) that this random variable has an F distribution.

To use (20.26) to define a hypothesis test, we need to use ***F*-values**, which are obtained from F distributions in the same way that t-values are obtained from t distributions. The main novelty here is that F distributions are defined by two parameters, k and d. Getting into specifics, recall from Appendix 17.A (online) that the right-tail F-value $\bar{F}_a^{k,d}$ is defined by $\mathrm{P}(\mathcal{F} > \bar{F}_a^{k,d}) = a$, where $\mathcal{F} \sim F(k, d)$. In words, a random variable drawn from an $F(k, d)$ distribution exceeds the F-value $\bar{F}_a^{k,d}$ with probability a.

We can now present the F test for simple regression.

The *F* test for $\beta = 0$ against the two-sided alternative.

In the classical normal regression model, we reject the null hypothesis $H_0:\ \beta = 0$ *in favor of the alternative* $H_1:\ \beta \neq 0$ *at significance level a if*

$$(n-2)\frac{R^2}{1-R^2} > \bar{F}_a^{1,n-2}.$$

■ Example In the wage/customer satisfaction example, can we reject the null hypothesis that $\beta = 0$ in favor of a two-sided alternative at significance level .001? Since the sample size is $n = 16$ and $R^2 = .6876$, the F-statistic is

$$(n-2)\frac{R^2}{1-R^2} = 14 \times \frac{.6876}{1-.6876} = 30.8177.$$

The F-statistic is also reported by the `regression_inference.xlsx` workbook—see Figure 20.9.

We need to compare the F-statistic to the F-value $\bar{F}_{.001}^{1,14}$. Using the `distributions.xlsx` workbook, or typing “`=F.INV.RT(.001,1,14)`” into a blank cell in Excel, we find that $\bar{F}_{.001}^{1,14} = 17.143$. Since $30.8177 > 17.143$, we reject the null hypothesis that $\beta = 0$ in favor of the two-sided alternative.

In this example, the F-statistic was far larger than the F-value specified by the hypothesis test, even with a demanding choice of significance level. This is a common occurrence in simple regression. While we may have only a rough sense of the value of β before we draw our sample, we often have little doubt that β is not near zero, and the results of the F test will reflect this. ■

In Appendix 20.A.6, we provide two different derivations of the fact (20.26) that underlies the F test. The first derivation shows that statement (20.26) is equivalent to statement (20.19) about the distribution of the t-statistic $(B - \beta)/\mathcal{S}_B$ in the case where $\beta = 0$. Put differently, the F test above amounts to a different way of writing the two-tailed t test for $\beta = 0$ from Section 20.5.2. In the second derivation, we obtain (20.26) from properties of the sum-of-squares equation (20.24).

Since the F test is equivalent to the t test in the present context, introducing the F test here wasn't strictly necessary. We introduced it anyway because versions of the F test are basic tools for inference in more complicated econometric and statistical settings. For instance, F tests are a basic inference tool for *multiple regression*, where one uses multiple explanatory variables as the basis for predictions about the y variable (see Section 20.7.3). They are also key tools in *analysis of variance* (*ANOVA*), where they are used to test for differences in means across subpopulations defined in terms of one or more characteristics. F tests in these settings are not equivalent to t tests, but they can be derived from properties of suitable sum-of-squares equations.

20.6.3 What happens without normality? The robustness of the *F*-statistic*

Equation (20.26) states that under the assumptions of the classical normal regression model with $\beta = 0$, the F-statistic $(n-2)\frac{R^2}{1-R^2}$ follows an F distribution. This fact allows us to use the F-statistic to test the zero-slope null hypothesis against a two-sided alternative.

For equation (20.26) to be true, the observations Y_i must be normally distributed. What can be said if this is not the case? We describe two approaches to this question, each of which parallels an approach to inference about an unknown mean μ_X of i.i.d. trials $\{X_i\}_{i=1}^n$ whose variance σ_X^2 is unknown.[21]

To justify our small-sample inference procedures about μ_X for normally distributed trials, we used the fact that

$$\text{(20.27)} \qquad \frac{\bar{X}_n - \mu_X}{\frac{1}{\sqrt{n}}\mathcal{S}_n} \sim t(n-1) \text{ when each trial } X_i \text{ is normally distributed.}$$

If the trials are not normally distributed, then the t-statistic does not follow a t distribution. Nevertheless, we argued in Section 17.4 that the t-statistic is *robust*

[21] Again, since the F test is equivalent to a t test in the present context, this discussion is not strictly necessary, but the ideas here also apply in the more complicated settings mentioned above.

in the following sense: If the distribution of each trial is not too far from normal—if it is continuous, fairly symmetric, and not too skewed—then the the t-statistic will still have an approximate t distribution even if the sample size is fairly small.

As it turns out, the F-statistic is also robust, though to a lesser degree than the t-statistic. If the distribution of each trial is close to normal, then the F-statistic will have an approximate F distribution for fairly small sample sizes.[22] In such cases, using the F test introduced above is justified.

Alternatively, we can consider large samples. In the i.i.d. trials setting, having a large number of trials ensures that the sample mean $\bar{X}_n$ is approximately normally distributed (by the central limit theorem), and that the sample standard deviation $\mathcal{S}_n$ provides a good approximation of the unknown variance σ_X^2. Using these approximations, we argued in Chapter 15 that

$$\frac{\bar{X}_n - \mu_X}{\frac{1}{\sqrt{n}}\mathcal{S}_n} \approx N(0,1) \text{ when the sample size } n \text{ is large enough.} \tag{20.28}$$

This fact is the basis for our inference procedures from Chapters 15 and 16.

A related large-sample approximation holds for the F-statistic in the context of regression. In Appendix 20.A.6, we describe the logic behind the following fact:

The robustness of the *F*-statistic.

Suppose the assumptions of the classical regression model hold with $\beta = 0$. If the sample size n is large enough and there is non-negligible variation in the x values, then

$$(n-2)\frac{R^2}{1-R^2} \approx \chi^2(1). \tag{20.29}$$

Thus, if the number of trials is large enough, the statistic $(n-2)\frac{R^2}{1-R^2}$ has an approximate $\chi^2(1)$ distribution even if the Y_i are not normally distributed. This fact can be used to construct large-sample tests of the zero-slope null hypothesis based on the $\chi^2(1)$ distribution (see Exercise 20.6.8).

If we have a large i.i.d. sample from a distribution that is not too far from normal, inference procedures based on either (20.27) or (20.28) can be applied. The former will use t-values, the latter z-values. But as we saw in Chapter 17, the $t(n-1)$ distribution converges to the $N(0,1)$ distribution as the sample size n grows large. Thus in cases where either approach could be followed, procedures based on t-values and z-values do not differ in an important way.

Likewise, if the sample size is large and the distribution of each Y_i close to normal in our regression model, we have two options for inference procedures: starting from (20.26) and using F-values, or starting from (20.29) and using χ^2-values.

[22] For analysis and references, see Mukhtar M. Ali and Subhash C. Sharma, "Robustness to Nonnormality of Regression F-tests," *Journal of Econometrics* 71 (1996), 175–205.

But we saw in Chapter 17 that the $F(1, n-2)$ distribution converges to the $\chi^2(1)$ distribution as the sample size n grows large. Thus when either approach could be followed, procedures based on F-values and χ^2-values do not differ in an important way.

While we have emphasized the similarities in how failures of normality affect procedures based on t- and F-statistics, there are differences as well. Procedures based on t-statistics are robust to larger departures from normality than ones based on F-statistics. Also, as we discussed in Section 17.A.5, the convergence of t distributions to the standard normal distribution is faster than that of F distributions to averaged χ^2 distributions. Thus compared to those based on t-statistics, procedures using F-statistics should be employed with a bit more caution when trials are not normal.

20.7 Regression and Causation

We conclude the chapter with a discussion of regression inference and causation, complementing our discussions in Sections 18.4 and 19.5.3.

20.7.1 An alternate description of the classical regression model

We defined the classical regression model using conditions (C1)–(C3).

The classical regression model.

(C1) *Fixed x sampling*: $x_1, \ldots, x_n$ *are fixed and not all identical;* $Y_1, \ldots, Y_n$ *are independent random variables.*

(C2) *Linearity of conditional means*: $\mathrm{E}(Y_i) = \alpha + \beta x_i$.

(C3) *Constant conditional variances*: $\mathrm{Var}(Y_i) = \sigma^2$.

To begin our discussion, we observe that the classical regression model is also commonly expressed in the following mathematically equivalent way:

$$Y_i = \alpha + \beta x_i + \mathcal{E}_i, \quad \{\mathcal{E}_i\}_{i=1}^n \text{ independent}, \mathrm{E}(\mathcal{E}_i) = 0, \mathrm{Var}(\mathcal{E}_i) = \sigma^2. \tag{20.30}$$

In this description, the y observations are divided into two parts: their expected values $\alpha + \beta x_i$, and mean-zero **error terms** (or *disturbance terms*), $\mathcal{E}_i$.[23] The random variable $\mathcal{E}_i = Y_i - (\alpha + \beta x_i)$ represents the difference between the (random) observation Y_i and its (nonrandom) expected value. This description of the model has the advantage of describing the variation in the samples using independent random variables with zero mean and a common variance, which can be convenient for calculations (see Appendix 20.A.5).

[23] We use a capital $\mathcal{E}$ here to emphasize that $\mathcal{E}_i$ is a random variable.

20.7.2 Causal regression models

Conditions (C1)–(C3) are mathematically equivalent to the formulation (20.30) of the classical regression model. However, formulation (20.30) commonly serves a dual role, not only describing the probability model for the sampling process, but also providing a **causal model**, as introduced in Section 18.4. In other words, formulation (20.30) is often used to indicate a **causal assumption**: namely, that changes in the x variable *cause* changes in the y variable, regardless of how x itself is determined, and that there are no variables besides x that systematically influence y, apart from ones explicitly held fixed in the sample. If this causal assumption is correct, then the parameter β represents the rate at which changes in the x variable cause changes in the y variable.

If a regression model is used to model a controlled experiment, then a causal interpretation of the model is usually correct. For instance, if an agricultural researcher is varying the amount of fertilizer on distinct but identical plots of land, being sure to keep all other key influences on crop yields equal, then it is reasonable to conclude that the changes in the amount of fertilizer are causing the observed differences in crop yields. In such cases, the value of β is naturally interpreted as the causal effect of the x variable on the y variable.

If a regression model is used to model an observational study—for instance, if (20.30) is used to describe stratified sampling from a population (see Section 20.1.1), giving the model a causal interpretation is a bolder move. In this case, the causal interpretation of (20.30) says that if we fix an individual's x value, but have no control over the individual's other characteristics, then the individual's y value would be determined as in (20.30). This assumption often fails because of **confounding variables** that have a causal relationship with the x variable, the y variable, or both.[24]

■ Example

In the United States, refrigerators are labeled with a yellow EnergyGuide label that describes monthly electricity costs.[25] One would expect that refrigerators that use less energy, and so are less expensive to run, would be sold at higher prices. In fact, the opposite is true: refrigerators with low energy costs tend to have lower prices than ones with high energy costs. Do lower energy costs reduce the prices that consumers are willing to pay?

No. Both energy costs and prices depend on a third factor: size. Larger fridges use more energy, but they also keep more food cold, and it is the latter quality that consumers are willing to pay more for. If we look at refrigerators of a given size, the ones that use less energy tend to cost more, not less. ■

[24] We discussed confounding variables previously in Sections 10.5 and 18.4, where further examples can be found.

[25] See www.consumer.ftc.gov/articles/0072-shopping-home-appliances-use-energyguide-label.

■ Example We are analyzing the performances of sales staff in medical equipment companies. Regressing sales volumes on the sellers' levels of education, we find that an additional year of education is associated with an additional $7000 in annual sales. Should we conclude that the skills obtained during one more year in school are the cause of the $7000 increase in expected productivity?

Probably not. It seems likely that both sales volume and educational attainment are commonly influenced by other factors. For instance, intelligence and drive are likely to have direct effects on both educational attainment and career success. If we do not explicitly account for these factors in our analysis, then their effects on sales volume will show up in our regression in the only way they can: as part of the association between educational attainment and sales volume. Because of this **omitted variable bias**, only a portion of the association between schooling and sales volume appearing in our regression warrants a causal interpretation. ■

For better or worse, published research often presents models in the form (20.30) without being explicit about whether a causal interpretation is intended. As we discussed in Section 18.4, the only way to justify a causal interpretation of a statistical analysis is by way of causal assumptions. If researchers give the analysis of an observational study a causal interpretation, then they must be viewing (20.30) as a causal model. In this case, you need to ask yourself whether the causal model is convincing—in other words, whether it captures the main determinants of the y values other than those explicitly held fixed in the sample. If this is not the case, as in the examples above, then giving β a causal interpretation is not warranted.

20.7.3 Multiple regression

In observational studies, it is uncommon for a model with just one explanatory variable to have a causal interpretation. Instead, a collection of distinct explanatory variables (known as **independent variables**) are usually needed to capture the key causal influences on the y variable (known as the **dependent variable**).

As an example, a regression probability model with k independent variables can be written in the following form:

$$\text{(20.31)} \qquad \begin{gathered} Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \mathcal{E}_i, \\ \{\mathcal{E}_i\}_{i=1}^n \text{ independent}, \mathrm{E}(\mathcal{E}_i) = 0, \mathrm{Var}(\mathcal{E}_i) = \sigma^2. \end{gathered}$$

In performing statistical inference in this context, the central aim is to determine the values of the parameters β_1 through β_k, which describe how changes in the independent variables are associated with changes in the dependent variable. If the model can be given a causal interpretation—if all important influences on the dependent variable are included in the collection of independent variables—then each parameter β_j can be interpreted as the causal effect of increasing the jth independent variable on the value of the dependent variable when the other independent variables are held fixed.[26]

As you might expect, having to deal with more than two variables at a time makes analyses of multiple regression models distinctly more complicated than analyses of bivariate models, and large portions of econometrics textbooks are devoted to the novelties that arise. Nevertheless, many of the main concepts arising in multiple regression appear in a simple form in the bivariate models we have studied here.

20.A Appendix

20.A.1 Analysis of the random sampling regression model

Section 20.1 introduced the classical and random sampling regression models. The derivations of inference procedures in subsequent sections were for the classical model, but we claimed that these procedures work equally well under the assumptions of the random sampling model. In this section, we explain why.

First, we present the random regression model once again.[27]

The random sampling regression model.

(R1) *Random sampling*: $(X_1, Y_1), \dots, (X_n, Y_n)$ *are independent as i varies;* $\mathrm{SD}(X_i) > 0$.
(R2) *Linearity of conditional means*: $\mathrm{E}(Y_i|X_i = x) = \alpha + \beta x$.
(R3) *Constant conditional variances*: $\mathrm{Var}(Y_i|X_i = x) = \sigma^2$.

The OLS estimators for the random regression model are the same as those for the classical model, except from the fact that the x variable is now random. Thus the sample mean and sample variance,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \mathcal{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

are random variables, as is the sample covariance

$$\mathcal{S}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

[26] In econometrics, the term *structural equation* is often used to describe probability models like (20.31) when they are intended to have a causal interpretation.

[27] Here is a technicality we ignored the first time around. In writing down this model, we implicitly assume that the X_i are discrete random variables. This ensures that the event $\{X_i = x_i\}$ has positive probability for the realizations x_i that can actually occur, so that conditional probabilities that condition on this event make sense. Fortunately, nothing besides some technicalities changes if the X_i are continuous.

The formulas for the OLS estimators are the same as before, except that each x becomes an X:

$$B = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{and} \quad A = \bar{Y} - B\bar{X}.$$

We now explain why the random sampling regression model is more difficult to analyze than the classical model, and how the difficulty is resolved. We focus on the properties of the OLS slope estimator B.

Equation (20.3) showed that in the classical regression model, B is a linear function of the random variables $\{Y_i\}_{i=1}^n$:

$$B = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} \right) Y_i. \tag{20.3}$$

This fact made it easy to establish that B has mean $\mathrm{E}(B) = \beta$ and variance $\mathrm{Var}(B) = \sigma^2/((n-1)s_x^2)$ using our knowledge of the traits of Y_i and the fact that different Y_i are independent random variables (see Appendix 20.A.3 for the derivations). Likewise, under the additional assumption that the Y_i are normally distributed, this linearity property ensured that B is normally distributed as well.

With random sampling, the x values are random variables, so the corresponding expression for B is

$$B = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2} \right) Y_i.$$

While this is still a linear function of the Y_i, it is a *nonlinear function* of the X_i and Y_i in combination. Since the X_i and Y_i are unlikely to be independent—after all, the point of regression analysis is to look for relationships between the x and y variables—even computing the mean of B seems a daunting task.

The trick that gets us around this difficulty is to condition on the realization of the x values in the sample. If we condition on the event that $X_1 = x_1, \dots, X_n = x_n$, then B is again described by (20.3), and so is an unbiased estimator of β.

$$\mathrm{E}(B | X_1 = x_1, \dots, X_n = x_n) = \beta.$$

But the unconditional expected value of B is just an average of these conditional expected values.[28] Since the latter all equal β, the former must equal β as well: $\mathrm{E}(B) = \beta$.

This conditioning argument shows that even in the random sampling regression model, B is an unbiased estimator of β. Versions of this reasoning can be used to show that all of the inference procedures for the classical model apply equally well to the random sampling model. For the case of the normal random sampling regression model, see Exercise 20.M.9.

[28] This follows from the law of iterated expectation: see Exercise 20.M.8.

20.A.2 The unstructured regression model

Both the classical and random sampling regression models make strong assumptions about the relationship between x and y values in the population: subpopulation means depend linearly on the subpopulation x, and the subpopulation variances are all equal. While these assumptions are reasonable approximations in some applications, they may be violated in others. Fortunately, as we now explain, it is possible to use regression for inference without making any assumptions about the population whatsoever.

Consider the following regression probability model for inference about a population.

The unstructured regression model.

Independent random sampling: $(X_1, Y_1), \ldots, (X_n, Y_n)$ *are independent as i varies.*

The aim is to estimate the population regression line $y = \alpha + \beta x$, *where* $\beta = \frac{\sigma_{xy}}{\sigma_x^2}$ *and* $\alpha = \mu_y - \beta\mu_x$.

In the unstructured regression model, we make no assumptions at all about the joint distribution of x and y values in the population. Our goal in performing statistical inference is to estimate the regression line for the population. The regression line may not agree with the conditional expectation function, but as we argued in Chapter 19, it is still an important descriptive statistic in its own right.

The OLS estimators for the unstructured model are defined exactly as in the random sampling model from the previous section:

$$B = \frac{\mathcal{S}_{XY}}{\mathcal{S}_X^2} \quad \text{and} \quad A = \bar{Y} - B\bar{X}.$$

Are B and A still good estimators of β and α? Since the unstructured model imposes no assumptions on the population, it is too much to ask that B and A be unbiased. However, they can be shown to be *consistent* estimators of β and α. That is, when the sample size is large, B and A are very likely to be very close to β and α.[29]

What about hypothesis tests and confidence intervals for β? Advanced techniques can be used to show that the estimator B is approximately normally distributed when the sample size is large. To use this fact to define inference procedures, we need to be able to estimate the variance of B. This is a more complicated matter than in our earlier models: there is nothing like the common conditional variance σ^2 in the unstructured regression model, so the expression

[29] Here's why: We know (from Section 14.4) that the sample variance $\mathcal{S}_X^2$ is a consistent estimator of the population covariance σ_x^2, and it should not be hard to believe that the same connection holds between $\mathcal{S}_{XY}$ and σ_{xy}. Then Slutsky's theorem (Appendix 14.B) implies that the ratio $B = \mathcal{S}_{XY}/\mathcal{S}_X^2$ is a consistent estimator of the ratio $\beta = \sigma_{xy}/\sigma_x^2$.

for Var(B) is not as simple as before. Nevertheless, consistent estimators of Var(B) are available, and they are very widely used in economic research.[30]

To sum up, the use of regression for inference about a population really doesn't depend on imposing strong assumptions about the population under study, although the weaker assumptions change both the interpretations and details of the inference procedures.

20.A.3 Computation of the mean and variance of *B*

To compute the mean and variance of B, we will use the formula

$$B = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \right) Y_i, \tag{20.3}$$

that expresses B as a linear function of the random variables Y_i, along with the formulas for linear functions of random variables from Chapters 3 and 4 and assumptions (C1)–(C3) of the classical regression model. Throughout this calculation, we use the fact that the x_i are not random variables, but rather fixed numbers specified in advance by the researcher.

We start with the computation of E(B). Our basic facts about expected values show that

$$\begin{aligned}
\mathrm{E}(B) &= \mathrm{E}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} Y_i \right) \\
&= \sum_{i=1}^{n} \mathrm{E}\left(\frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} Y_i \right) \\
&= \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \mathrm{E}(Y_i).
\end{aligned}$$

Next, using the fact that $\mathrm{E}(Y_i) = \alpha + \beta x_i$ (assumption (C2)) shows that

$$\mathrm{E}(B) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} (\alpha + \beta x_i). \tag{20.32}$$

To simplify this expression, we use the fact that $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ twice, once to eliminate α, and once to introduce $\beta\bar{x}$. We obtain

$$\mathrm{E}(B) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n} (x_j - \bar{x})^2} (\beta x_i - \beta\bar{x}) \tag{20.33}$$

[30] Consistent estimators of Var(B) for this unstructured setting are known as *heteroskedasticity-robust standard errors* or *White standard errors*

.

$$
\begin{aligned}
&= \beta \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \\
&= \beta.
\end{aligned}
$$

To compute $\mathrm{Var}(B)$, use our formulas for variances of linear functions of random variables, along with the assumptions that the Y_i are independent (assumption (C1)) and have common variance σ^2 (assumption (C3)):

$$
\begin{aligned}
\mathrm{Var}(B) &= \mathrm{Var}\left(\sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} Y_i \right) \\
&= \sum_{i=1}^{n} \mathrm{Var}\left(\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} Y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right)^2 \mathrm{Var}(Y_i) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{j=1}^{n}(x_j - \bar{x})^2\right)^2} \sigma^2 \\
&= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sigma^2 \\
&= \frac{\sigma^2}{(n-1)s_x^2}.
\end{aligned}
$$

20.A.4 Proof of the Gauss-Markov theorem

In this appendix, we prove that under the assumptions of the classical regression model, the OLS estimator B has the lowest variance of all unbiased linear estimators of β. The proofs of the other claims of the theorem are similar—see Exercise 20.M.4.

This proof also uses the formula

$$
B = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \right) Y_i, \tag{20.3}
$$

expressing B as a linear function of the random variables Y_i. Rearranging this slightly shows that

$$
B = \sum_{i=1}^{n}(p + qx_i)Y_i, \quad \text{where } p = \frac{-\bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \text{ and } q = \frac{1}{\sum_{j=1}^{n}(x_j - \bar{x})^2}. \tag{20.34}
$$

In words, the weight on Y_i is the sum of two terms: a constant p (that does not depend on i), and the value x_i times a constant q (a constant that also does not depend on i). Put differently, the list of weights is a linear combination of a list of 1s and the list of x values. This is the key fact about the OLS estimator B used in the proof, specifically in display (20.37).

Now suppose that $\tilde{B}$ is a different linear unbiased estimator of β. Our goal is to show is that $\text{Var}(\tilde{B}) > \text{Var}(B)$. To start, we define $D = \tilde{B} - B$, so that $\tilde{B} = B + D$. Since $\tilde{B}$ and B are both linear, D is linear as well, meaning that we can write

$$D = \sum_{i=1}^{n} d_i Y_i \tag{20.35}$$

for some choice of the constants d_i. Since $\tilde{B}$ is different from B, at least one d_i is not zero.

Since $\tilde{B}$ and B are both unbiased, they both have mean β regardless of the values of α and β. It follows that

$$\text{E}(D) = \text{E}(\tilde{B} - B) = \text{E}(\tilde{B}) - \text{E}(B) = 0 - 0 = 0,$$

so D has mean zero regardless of the values of α and β. Combining the two previous equations, and using the fact that $\text{E}(Y_i) = \alpha + \beta x_i$ (assumption (C2)), we see that

$$0 = \text{E}(D) = \text{E}\left(\sum_{i=1}^{n} d_i Y_i\right) = \sum_{i=1}^{n} d_i \text{E}(Y_i) = \sum_{i=1}^{n} d_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} d_i + \beta \sum_{i=1}^{n} d_i x_i.$$

The only way that this statement can be true regardless of the values of α and β is if the following orthogonality conditions hold:

$$\sum_{i=1}^{n} d_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} d_i x_i = 0. \tag{20.36}$$

(Recall from Section 19.2.3 that lists $\{a_i\}_{i=1}^n$ and $\{b_i\}_{i=1}^n$ are orthogonal if $\sum_{i=1}^n a_i b_i = 0$.)

We are now ready to compare the variances of B and $\tilde{B}$. To begin, note that

$$\text{Var}(\tilde{B}) = \text{Var}(B + D) = \text{Var}(B) + 2\,\text{Cov}(B, D) + \text{Var}(D).$$

To evaluate the covariance term, we use our formulas from Chapter 4, the facts that the Y_i are independent (assumption (C1)) with common variance σ^2 (assumption (C3)), along with (20.34), (20.35), and (20.36):

$$\begin{aligned} \text{Cov}(B, D) &= \text{Cov}\left(\sum_{i=1}^{n} (p + qx_i) Y_i, \sum_{j=1}^{n} d_j Y_j\right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}\big((p + qx_i) Y_i, d_j Y_j\big) \end{aligned} \tag{20.37}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} (p + qx_i) d_i \mathrm{Var}(Y_i) \\
&= \sigma^2 \left(p \sum_{i=1}^{n} d_i + q \sum_{i=1}^{n} d_i x_i \right) \\
&= 0.
\end{aligned}
$$

In words, since the list of weights defining B is a linear combination of the list of 1s and the list of x values, and since both of the latter are orthogonal to the list of weights defining D, it follows that the lists of weights defining B and D are orthogonal to each other. This implies in turn that B and D are uncorrelated.

Furthermore, the independence of the Y_j and the fact that at least one d_j is not zero implies that

$$\mathrm{Var}(D) = \mathrm{Var}\left(\sum_{j=1}^{n} d_j Y_j \right) = \sum_{j=1}^{n} \mathrm{Var}\left(d_j Y_j \right) = \sum_{j=1}^{n} d_j^2 \mathrm{Var}(Y_j) = \sigma^2 \sum_{j=1}^{n} d_j^2 > 0.$$

Combining the last three displays allows us to conclude that

$$\mathrm{Var}(\tilde{B}) = \mathrm{Var}(B) + 0 + \mathrm{Var}(D) > \mathrm{Var}(B),$$

which is what we wanted to show.

20.A.5 Proof that the sample conditional variance is unbiased

As in (20.30), define the error term $\mathcal{E}_i = Y_i - (\alpha + \beta x_i)$ to be the difference between the ith y observation and its conditional mean. Then under the assumptions of the classical regression model, $\{\mathcal{E}_i\}_{i=1}^n$ is a sequence of independent random variables with mean $\mathrm{E}(\mathcal{E}_i) = 0$ and variance $\mathrm{Var}(\mathcal{E}_i) = \sigma^2$. Reviewing the argument from Appendix 14.B shows that the sample variance $\mathcal{S}_{\mathcal{E}}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})^2$ has mean $\mathrm{E}(\mathcal{S}_{\mathcal{E}}^2) = \sigma^2$. The novelty here is that this conclusion depends only on the independence and common mean and variance of the $\mathcal{E}_i$, and does not require them to be identically distributed.

The sample regression residuals can be expressed as

$$
\begin{aligned}
U_i &= Y_i - (A + Bx_i) \\
&= \mathcal{E}_i + (\alpha + \beta x_i) - (A + Bx_i) \\
&= \mathcal{E}_i - (A - \alpha) - (B - \beta) x_i. \qquad (20.38)
\end{aligned}
$$

Now, we know from our characterizations of the regression line from Chapter 19 that the sample regression residuals sum to 0. Therefore, summing (20.38) over i and dividing by n yields

$$0 = \bar{\mathcal{E}} - (A - \alpha) - (B - \beta)\bar{x}, \qquad (20.39)$$

where $\bar{\mathcal{E}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}_i$ denotes the sample mean of the sequence $\{\mathcal{E}_i\}_{i=1}^{n}$. Thus, subtracting (20.39) from (20.38) yields

$$U_i = (\mathcal{E}_i - \bar{\mathcal{E}}) - (B - \beta)(x_i - \bar{x}).$$

Taking the sum of squares of this equation yields

(20.40)

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})^2 + (B - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2(B - \beta) \sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})(x_i - \bar{x}).$$

In order to evaluate the expected value of this sum of squares, we repeatedly use the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ along with the definitions of $\mathcal{E}_i$ and B to rewrite the final term in (20.40):

$$\begin{aligned}
(B - \beta) \sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})(x_i - \bar{x}) &= (B - \beta) \sum_{i=1}^{n} \mathcal{E}_i(x_i - \bar{x}) \\
&= (B - \beta) \sum_{i=1}^{n} (Y_i - \beta x_i)(x_i - \bar{x}) \\
&= (B - \beta) \sum_{i=1}^{n} (Y_i - \bar{Y} - \beta(x_i - \bar{x}))(x_i - \bar{x}) \\
&= (B - \beta)\left(\sum_{i=1}^{n} (Y_i - \bar{Y})(x_i - \bar{x}) - \beta \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \\
&= (B - \beta)\left(B \sum_{i=1}^{n} (x_i - \bar{x})^2 - \beta \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \\
&= (B - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2.
\end{aligned}$$

By substituting this into (20.40), and using the facts that $\mathrm{E}(\mathcal{S}_{\mathcal{E}}^2) = \sigma^2$ and $\mathrm{Var}(B) = \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$, we find that

$$\begin{aligned}
\mathrm{E}\left(\sum_{i=1}^{n} U_i^2 \right) &= \mathrm{E}\left(\sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})^2 \right) + \mathrm{E}\left((B - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \\
&\quad - 2\mathrm{E}\left((B - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right) \\
&= (n - 1)\mathrm{E}(\mathcal{S}_{\mathcal{E}}^2) + \mathrm{Var}(B) \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\mathrm{Var}(B) \sum_{i=1}^{n} (x_i - \bar{x})^2
\end{aligned}$$

$$
\begin{aligned}
&= (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 \\
&= (n-2)\sigma^2.
\end{aligned}
$$

We therefore conclude that

$$\mathrm{E}(\mathcal{S}) = \mathrm{E}\left(\frac{1}{n-2}\sum_{i=1}^{n} U_i^2\right) = \frac{1}{n-2}\mathrm{E}\left(\sum_{i=1}^{n} U_i^2\right) = \sigma^2.$$

20.A.6 Deriving the distribution of the *F*-statistic

To start, we give two derivations of the distribution of the F-statistic in the classical normal regression model when $\beta = 0$, namely

$$(n-2)\frac{R^2}{1-R^2} \sim F(1, n-2). \tag{20.26}$$

We first derive it from the distribution of the t-statistic for inference about β in this model, namely

$$\frac{B-\beta}{\mathcal{S}_B} \sim t(n-2). \tag{20.19}$$

In Appendix 17.A (online), we saw that the square of a random variable with a $t(d)$ distribution has an $F(1, d)$ distribution. Thus, assuming that $\beta = 0$ and squaring the previous equation yields

$$\frac{B^2}{\mathcal{S}_B^2} \sim F(1, n-2).$$

Thus to establish (20.26) it is enough to show that

$$\frac{B^2}{\mathcal{S}_B^2} = (n-2)\frac{R^2}{1-R^2}. \tag{20.41}$$

We now derive (20.41). The definition of the ith predicted value is $\hat{Y}_i = A + Bx_i$, and the definition of A implies that $\bar{Y} = A + B\bar{x}$. Therefore

$$\hat{Y}_i - \bar{Y} = (A + Bx_i) - (A + B\bar{x}) = B(x_i - \bar{x}),$$

so taking sums of squares yields

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = B^2\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Using this equation and the definitions of $\mathcal{S}_B^2$ and R^2, we obtain

$$\begin{aligned}
\frac{B^2}{\mathcal{S}_B^2} &= \frac{B^2}{\dfrac{\mathcal{S}^2}{(n-1)s_x^2}} \\
&= \frac{\dfrac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}{\dfrac{\frac{1}{n-2}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \\
&= (n-2)\,\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2} \\
&= (n-2)\,\frac{\dfrac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}}{\dfrac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \\
&= (n-2)\,\frac{R^2}{1-R^2}.
\end{aligned}$$

One can also derive the distribution of the F-statistic from the following facts about the sum-of-squares equation. Unlike the derivation above, this approach generalizes to multiple regression.

Distributions of sums of squares in the classical normal regression model if $\beta = 0$.

Consider the sum-of-squares equation

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{20.24}$$

Under the assumptions of the classical normal regression model, if $\beta = 0$, then

(i) The two sums on the right-hand side of (20.24) *are independent random variables.*
(ii) If we divide each of the three sums in (20.24) *by σ^2, the resulting expressions have $\chi^2(n-1)$, $\chi^2(1)$, and $\chi^2(n-2)$ distributions.*

These facts about distributions of sums of squares are closely related to the derivation of the distribution of the t-statistic, which we discussed in Section 17.A.4.

It follows from these facts and from the definition (17.A.3) of the F distribution that when $\beta = 0$,

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \sim F(1, n-2). \tag{20.42}$$

Dividing the top and bottom of this fraction by $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and then applying the definition of R^2 yields

$$(n-2)\frac{R^2}{1-R^2} \sim F(1, n-2).$$

Next, we argue that even without the normality assumption, if the sample size n is large enough and there is nonnegligible variation in the x values, then when $\beta = 0$, the F-statistic has an approximate $\chi^2(1)$ distribution. We start from the fact that the F-statistic can be expressed as the left-hand side of (20.42). Since the denominator of this expression is the sample conditional variance $\mathcal{S}^2$, multiplying the numerator and denominator of this expression by $1/\sigma^2$ lets us express the F-statistic as

$$\frac{\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{\sigma^2}\mathcal{S}^2}. \tag{20.43}$$

We saw in Section 20.3 that $\mathcal{S}^2$ is a consistent estimator of σ^2: as n grows large, the denominator of (20.43) converges in probability to 1. This suggests (and in fact implies, by a version of Slutsky's theorem) that if the numerator of (20.43) converges in distribution to a $\chi^2(1)$ random variable, then so does (20.43) as a whole.

To evaluate the numerator of (20.43), we rewrite it as

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(A + Bx_i - \bar{Y})^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\bar{Y} - B\bar{x} + Bx_i - \bar{Y})^2 \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}B^2. \qquad (20.44)
\end{aligned}$$

We saw in Section 20.2 that the OLS estimator B is approximately normally distributed, with mean β and variance $\sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2$. Since $\beta = 0$ by assumption, we have

$$B \approx N\left(0, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Thus by the basic properties of normal random variables,

$$\frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\sigma}B \approx N(0,1). \tag{20.45}$$

By definition, the square of a standard normal random variable is a $\chi^2(1)$ random variable. This suggests (and in fact implies, by a version of the continuous mapping theorem) that if we square the left-hand side of (20.45), which has an approximately standard normal distribution, the result will have an approximate $\chi^2(1)$ distribution. But this square is (20.44). Thus the numerator of (20.43) has an approximate $\chi^2(1)$ distribution, which is what we set out to show.

KEY TERMS AND CONCEPTS

classical regression model (p. 2)
random sampling regression model (p. 2)
fixed *x* sampling (p. 3)
random sampling (p. 3)
stratified sampling (p. 3)
subpopulation (p. 4)
linearity of conditional means (p. 4)
conditional mean (p. 4)
constant conditional variances (p. 5)
conditional variance (p. 6)
time series (p. 9)
serial correlation (p. 9)
ordinary least squares (OLS) estimators (p. 11)
Gauss-Markov theorem (p. 16)
sample subpopulation mean (p. 17)
sample conditional variance (p. 18)
sample regression prediction (p. 18)
sample regression residual (p. 18)
degrees of freedom (p. 19)
standard error (p. 20, 21, 25)
population regressions vs. sample regressions (p. 26)
classical normal regression model (p. 29)
t-statistic (p. 31, 35)
prediction interval (p. 36, 37)
sum-of-squares equation (p. 40)
sample correlation coefficient (p. 40)
squared sample correlation coefficient (R^2) (p. 40)
F test (p. 41)
F-statistic (p. 42)
F-value (p. 42)
error term (p. 45)
causal model (p. 46)
causal assumption (p. 46)
confounding variable (p. 46)
omitted variable bias (p. 47)
independent variable (p. 47)
dependent variable (p. 47)
random sampling regression model (p. 48)
unstructured regression model (p. 50)

20.E Exercises

Data sets used in the exercises can be found in the `ch20_data.xlsx` workbook. Unless otherwise indicated, data should be analyzed using the `regression_inference.xlsx` workbook.

Section 20.1 exercises

Exercise 20.1.1. Bill's scores (y) during each attempt at a popular online game tend to be proportional to the amount of time the attempt lasts (x). While his scores on short attempts are quite predictable, those on longer attempts fall in much wider ranges. You will observe the durations and scores of Bill's next 100 attempts.

a. Which sampling assumption is more apt here, (C1) or (R1)? Why?
b. Does the sample satisfy the distributional assumptions of the regression model you indicated in part (a)? Explain.

Exercise 20.1.2. A firm's weekly expenses include a fixed component and a component that is proportional to the number of units it sells. The variability of weekly expenses is consistent across the sales figures that actually arise. We plan to take a stratified sample of this firm's weekly unit sales (x) and weekly expenses (y).

a. Which sampling assumption is more apt here, (C1) or (R1)? Why?
b. Does the sample satisfy the distributional assumptions of the regression model you indicated in part (a)? Explain.

Exercise 20.1.3. A global consulting firm hires thousands of new employees each year. The human resources department strictly limits the salary range of new hires in order to avoid setting costly precedents. Employees who remain with the firm for many years eventually get larger raises each year.

a. Would you expect population data on years of experience (x) and compensation (y) for this firm to satisfy the assumptions of the classical regression model? Explain.
b. Sketch a scatterplot similar to Figures 20.1–20.3 that captures the main features you expect the population data to have.

Exercise 20.1.4. Attendance at San Francisco Giants baseball games varies with weather conditions. The spring and fall include a mix of frigid and warm days, while the summer brings both warm and unpleasantly hot days. On pleasant days, the Giants nearly always fill every seat.

a. Would you expect data on temperature (x) and attendance (y) at Giants games to satisfy the assumptions of the classical regression model? Explain.
b. Draw a scatterplot similar to Figures 20.1–20.3 that captures the main features you expect the data to have.

Section 20.2 exercises

Exercise 20.2.1. Let $\{(x_i, Y_i)\}_{i=1}^n$ represent an experiment that satisfies the assumptions of the classical regression model with $n = 70$, $\bar{x} = 10.5$, $s_x^2 = 2.2$, $\alpha = .5$, $\beta = .8$, and $\sigma^2 = 1.6$.

a. Compute the means, variances, and covariance of A and B.
b. Compute the mean and variance of $A + Bx$ for an arbitrary value of x.
c. What is the approximate probability that $B > 1$?
d. What is the approximate probability that $A + 10.5B > 9$?

Exercise 20.2.2. Let $\{(x_i, Y_i)\}_{i=1}^n$ be a stratified sample that satisfies the assumptions of the classical regression model with $n = 100$, $\bar{x} = 2$, $s_x^2 = 0.75$, $\alpha = 2$, $\beta = .20$, and $\sigma^2 = 1.5$.

a. Compute the means, variances, and covariance of A and B.
b. Compute the mean and variance of $A + Bx$ for an arbitrary value of x.
c. Evaluate $\mathrm{P}(A < 1.5)$ and $\mathrm{P}(B < 0)$.

Exercise 20.2.3. Continue with the specification of the classical regression model from Exercise 20.2.2.

a. Suppose that $x = 2$ occurs $n_2 = 18$ times in the list $\{x_i\}_{i=1}^n$ of x values used in the sample. Compute the variance of the sample subpopulation mean

$$\bar{Y}_{|x=2} = \frac{1}{n_2} \sum_{i\,:\,x_i=2} Y_i.$$

b. Compute the variance of $A + 2B$.
c. Which is larger, $\mathrm{Var}(\bar{Y}_{|x=2})$ or $\mathrm{Var}(A + 2B)$?
d. Could you have answered part (c) without doing any computations? If so, how?

Section 20.3 exercises

Exercise 20.3.1. Drivers for a taxi company have flexible schedules. Their income is the difference between the percentage of the fares they get to keep, minus the cost of gasoline, which they must pay themselves. Some of the cost of gasoline is due to the journeys to and from the driver's home, which must be paid regardless of the shift length. The `drivers` worksheet presents the shift lengths (x) and incomes (y) from a random sample of 70 shifts. Assume that the assumptions of the random sampling regression model hold.

a. Provide interpretations of α, β, and σ^2 in this example.
b. Report the OLS estimators A and B and the sample conditional variance $\mathcal{S}^2$. Interpret each in the context of this example.
c. Estimate the expected income from a 10-hour shift.

Exercise 20.3.2. Crab fishermen in Alaska endure very dangerous working conditions to catch delicious Alaskan king crabs. The `crabs.xlsx` worksheet contains sample data on crew size (x) and catch (y, in tons) for 55 crabbing trips during the previous season. Assume that the assumptions of the random sampling regression model hold.

a. Report the sample regression line and the sample conditional variance.
b. Estimate the mean catch size for a ship with 8 crew members.
c. Suppose we are able to obtain data from a large sample of crabbing trips. What can we say about the OLS estimators and the sample conditional variance in this case?

Exercise 20.3.3. A pediatrician is running a study on sleep deprivation in children and its impact on cognitive skills. Seventy-two families with 13-year-olds agree to participate in the study. For each child, the pediatrician learns the average number of hours the child slept during the past week (x), as well as the child's score on a test of concentration and short-term memory (y, on a 100-point scale). Her data can be found in the `sleep` worksheet. Assume that the assumptions of the random sampling regression model hold.

a. Report the sample regression line and the sample conditional variance.
b. Estimate the expected test score of a student who sleeps 8.5 hours per night.
c. Suppose that the pediatrician is interested in the relationship between sleep and cognitive skills among all U.S. 13-year-olds. Will her experiment provide unbiased estimates of the slope and intercept of the true regression line for this population? Why or why not?

Exercise 20.3.4. Podcasts can make money from advertising if they can attract a large enough listener base. The `podcast` worksheet contains data reports the downloads per week (x, in thousands) and advertising revenues (y, in thousands of dollars per week) for 125 randomly selected podcasts. Assume that the assumptions of the random sampling regression model hold.

a. Report the sample regression line and the sample conditional variance.
b. Estimate the expected revenue of a podcast with 40,000 weekly downloads.

Exercise 20.3.5. A company is evaluating the effect of their employees' Internet use on productivity. It gathers data, reported in the `productivity` worksheet, on weekly personal Internet use (x, in minutes) and sales volume (y, in thousands of dollars) for 57 employees in the sales department. Assume that the assumptions of the random sampling regression model hold.

a. Report the sample regression line and the sample conditional variance.
b. Estimate the expected sales volume of an employee who uses the Internet for 2.5 hours per week.

Exercise 20.3.6. A consumer research group wants to estimate the relationship between miles driven (x) and sales price (y, in dollars) for Toyota Corollas that are 4–6 years old. Data on mileage and sales price from a random sample of 80 transactions can be found in the `used_cars` worksheet. Assume that the assumptions of the random sampling regression model hold.

a. Report the sample regression line and the sample conditional variance.
b. Estimate the expected sales price of a Corolla that has been driven 80,000 miles.
c. Estimate the decline in expected sales price from driving the car an additional 300 miles.

Exercise 20.3.7. Suppose that the assumptions of the classical regression model hold, and that α and β are known. Show that

$$\mathcal{V}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\alpha + \beta x_i))^2$$

is an unbiased estimator of the conditional variance σ^2.

Section 20.4 exercises

Exercise 20.4.1. A transportation engineer is investigating the relationship between driving speeds and traffic fatalities. Looking at a stratified sample of 221 of stretches of U.S. highway, he obtains data on the average driving speed (x, in miles per hour) and the number of fatalities per billion miles driven (y). He obtains the following descriptive statistics and estimates:

$$\bar{x} = 57.32 \qquad A = 277.55$$

$$s_x^2 = 77.83 \qquad B = .4880 \qquad \mathcal{S}^2 = 7128.$$

Assume that the sample satisfies the assumptions of the classical regression model.

a. What are the interpretations of β and σ^2 in this scenario?
b. Give a .95 interval estimate for β. Then interpret both B and the interval estimate.
c. Compute the standard error of $\mathcal{S}_{A+Bx}$ when $x = 55$ and when $x = 75$. Explain why these standard errors differ in the way that they do.
d. Can you reject the null hypothesis that $\alpha + 75\beta$ equals 300 in favor of the alternative that it is larger at a 5% significance level? What about the null hypothesis that $\alpha + 75\beta$ equals 290?

Exercise 20.4.2. An environmental economist is studying the relationship between temperature and home heating costs in Madison. Each member of his sample is a pair consisting of a nonsummer month and a house of between 2000 and 2200 square feet that is gas heated. She obtains a stratified sample of size 85, obtaining the average temperature (x, in degrees Fahrenheit) and heating costs (y, in dollars) for each member of the sample. Her descriptive statistics and estimates are as follows:

$$\bar{x} = 38.02 \qquad A = 178.32$$

$$s_x^2 = 96.33 \qquad B = -2.393 \qquad \mathcal{S}^2 = 27.77.$$

Assume that the sample satisfies the assumptions of the classical regression model.

a. What are the interpretations of β and σ^2 in this scenario?
b. Give a .95 interval estimate for β.
c. Using your answer to part (b), determine whether you can reject the null hypothesis that $\beta = -2.50$ in favor of the two-sided alternative at a 5% significance level.
d. Provide a 90% interval estimate for the mean home heating cost during a month in which the average temperature is 8°F.

Exercise 20.4.3. Revisit Exercise 20.3.1, which considered shift lengths (x) and incomes (y) of taxi drivers.

a. Provide an interpretation of the slope parameter β for this example.
b. Construct a .95 confidence interval for β.
c. Can you reject the null hypothesis that the effect of an additional hour of work on expected income is \$10.00 in favor of the alternative that it is more than \$10.00 at a 5% significance level?
d. Provide a .95 confidence interval for the expected income from driving a 10-hour shift.

Exercise 20.4.4. Revisit Exercise 20.3.2, which analyzed crew size (x) and catch (y, in tons) of crab-fishing boats.

a. Provide an interpretation of the slope parameter β for this example.
b. Using a 90% significance level, test the null hypothesis that the expected effect of an extra crew member on the catch is at least 3 tons against the alternative that the expected effect is smaller than 3 tons.
c. Provide a .99 confidence interval for the expected catch of a boat with 8 crewmen.
d. The threshold for profitability for a boat with an eight-man crew is a catch of 27 tons. Can you establish that the expected catch of such a crew is more than 27 tons at a 5% significance level?

Exercise 20.4.5. Revisit Exercise 20.3.3, which studied average number of sleep hours (x) and test scores (y) of 13-year-olds.

a. Provide an interpretation of β for this example.
b. Using a 99% confidence level, can you reject the null hypothesis that β is at most 2.5 in favor of the alternative that it is larger than 2.5?
c. Construct a 95% confidence interval for the expected test score of a student who sleeps 8.5 hours per night.

Exercise 20.4.6. Revisit Exercise 20.3.4, which considered downloads per week (x, in thousands) and advertising revenues (y in thousands of dollars per week) of podcasts.

a. Provide a .95 confidence interval for the average revenue of a podcast with 120,000 weekly downloads.

b. Provide a .95 confidence interval for the average revenue of a podcast with 240,000 weekly downloads.
c. Explain in words why the widths of these intervals differ in the way that they do.

Exercise 20.4.7. Revisit Exercise 20.3.5, which considered weekly personal Internet use (x, in minutes) and sales volume (y, in thousands of dollars) of sales employees.

a. Test the null hypothesis that $\beta = 0$ against the alternative that $\beta < 0$ at a 5% significance level.
b. Provide a .95 confidence interval for the expected sales of an employee who spends 4 hours per week on personal Internet use.
c. Provide a .95 confidence interval for the expected sales of an employee who spends 6 hours per week on personal Internet use.

Exercise 20.4.8. Revisit Exercise 20.3.6, which looked at miles driven (x) and sales price (y, in dollars) of late-model Toyota Corollas.

a. Test the null hypothesis that $\beta \geq -.07$ against the alternative hypothesis that $\beta < -.07$ at a 5% significance level.
b. Provide a .95 confidence interval for the expected price of a Corolla with 60,000 miles on it.
c. Provide a .95 confidence interval for the expected price of a Corolla with 120,000 miles on it.
d. Explain in words why the widths of these intervals differ in the way that they do.

Exercise 20.4.9. Following the logic used in Chapter 15, derive the probability statement (20.12) that describes the defining property of the interval estimator for B.

Exercise 20.4.10. Following the logic used in Chapter 16, derive the probability statement (20.17) that provides the critical value for a one-tailed hypothesis test about the conditional mean $A + Bx$.

Exercise 20.4.11. A researcher painstakingly collects data in each of the 55 African countries on the average daily protein intake of pregnant women (x) and the average birthweight of children (y). He reports that the positive relationship he finds between these two variables is statistically significant at a 1% level. Knowing nothing else about his data, can you identify any problems with his claim? Discuss.

Section 20.5 exercises

Exercise 20.5.1. Suppose that $\{(x_i, Y_i)\}_{i=1}^{n}$ represents an experiment that satisfies the assumptions of the classical normal regression model with $\alpha = 10$, $\beta = 7$, and $\sigma^2 = 9$. Also, suppose that $x_1 = 8$ and $x_2 = 11$.

a. What is $P(Y_1 \geq 70)$?
b. What is $P(Y_2 \leq 80)$?
c. What is $P(Y_1 \geq 70$ and $Y_2 \leq 80)$?

Exercise 20.5.2. Suppose that $\{(x_i, Y_i)\}_{i=1}^n$ represents a stratified sample that satisfies the assumptions of the classical normal regression model with $\alpha = 110$, $\beta = -6$, and $\sigma^2 = 16$. Also, suppose that $x_1 = 12$ and $x_2 = 16$.
a. What is $P(Y_1 \geq 40)$?
b. What is $P(Y_2 \geq 20)$?
c. What is $P(\frac{1}{2}(Y_1 + Y_2) \geq 30)$?

Exercise 20.5.3. Let $\{(x_i, Y_i)\}_{i=1}^n$ represent an experiment that satisfies the assumptions of the classical normal regression model with $n = 10$, $\bar{x} = 45$, $s_x^2 = 60$, $\alpha = 60$, $\beta = -.35$, and $\sigma^2 = 40$.
a. Compute $P(B < .25)$.
b. Compute $P(A + 90B < 25)$.

Exercise 20.5.4. Suppose that stratified sample of size 5 satisfies the assumptions of the classical normal regression model. Knowing only this, is it possible to compute $P((B - \beta)/\mathcal{S}_B > 2)$? If so, compute it.

Exercise 20.5.5. A maker of exclusive collectible dolls is evaluating the relationship between the amount of time she spends creating a doll (x, in days) and the sales price of the doll at auction (y, in dollars), given the following data for her previous 12 dolls:

$$\bar{x} = 9.8 \qquad A = 890$$
$$s_x^2 = 6.3 \qquad B = 283 \qquad \mathcal{S}^2 = 7{,}562{,}500.$$

Assume that this experiment satisfies the assumptions of the classical normal regression model.
a. Compute a 95% confidence interval for β.
b. Compute a 95% confidence interval for the expected sales price of a doll that takes 100 hours to make.
c. Compute a 95% prediction interval for the sales price of the next doll she spends 100 hours making.

Exercise 20.5.6. A researcher is trying to determine whether a particular chemical's presence in tap water increases cancer rates. She measures the chemical's concentration in tap water (x, in ppm (parts per million)) and the cancer rate (y, in new cases per 1000 people during the past year) in 12 communities, all of which draw tap water from different sources. This data is reported in the `carcinogen` workbook. Assume that the sample satisfies the assumptions of the random sampling normal regression model.

a. Report the sample regression line and the sample conditional variance.
b. Test the null hypothesis that $\beta = 0$ against the alternative that $\beta > 0$ at a 5% significance level level.
c. Provide a 90% prediction interval for the cancer rate in a community where the concentration of the chemical in the tap water is 2.5 ppm.

Exercise 20.5.7. A consultant is estimating consumer water demand in Southern California. He gathers data on price (x, in dollars per gallon) and per capita water usage (y, in gallons per person per day) from 15 different towns, each of which sets via legislative fiat the price that the water utility charges. This data is reported in the `water` worksheet. Assume that the sample satisfies the assumptions of the random sampling normal regression model.

a. Report the sample regression line and the sample conditional variance.
b. Provide a 95% confidence interval for the expected demand for water in a town where the price of water is 2 cents per gallon.
c. Provide a 95% prediction interval for the demand for water in a California town where the price of water is 2 cents per gallon.
d. Test the null hypothesis that $\beta = 0$ against the alternative that $\beta < 0$ at a 5% significance level.

Exercise 20.5.8. Consider the solar panel example from Section 20.5.2.

a. Suppose that the same values of B, $\mathcal{S}^2$, and s_x^2 were obtained with a sample size of $n = 30$. Should the manager reject the null hypothesis?
b. What about if the sample size were $n = 60$?
c. What is the smallest sample size that would lead her to reject the null hypothesis?

Exercise 20.5.9. In the solar panel example from Section 20.5.2, the manager conducted an experiment by installing solar panels at 11 locations, and then used the results to construct a .95 prediction interval for the electrical output at a new location with insolation 6.10. The interval she found is $[1.3381, 1.4709]$. Let Y be the electrical output of the new panel. At the time before she installs the new panel, is the probability that Y lies in the interval $[1.3381, 1.4709]$ equal to .95? If so, explain why; if not, say what you can about this probability, and explain what the probability .95 refers to.

Exercise 20.5.10. You plan to construct a .95 prediction interval using data from a very large sample.

a. Give a simple approximate formula for the endpoints of this random interval in terms of the estimators A, B, and $\mathcal{S}$. (Hint: If n is very large, then $\frac{1}{n}$ is approximately zero.)
b. The prediction interval is very likely to be close to a fixed interval that can be described in terms of the parameters of the model. What is this fixed interval? Explain why this interval takes the form that it does.

Section 20.6 exercises

Exercise 20.6.1. You plan to run a regression on a stratified sample of size 15 that satisfies the assumptions of the classical normal regression model. You then intend to test the null hypothesis that $\beta = 0$ against the two-sided alternative. After taking the sample, you obtain an R^2 of .4582.

a. Can you reject the null hypothesis at a 5% significance level?
b. What is the lowest significance level at which the null hypothesis could be rejected?

Exercise 20.6.2. You plan to run a regression on a stratified sample of size 10 that satisfies the assumptions of the classical normal regression model, and to test the null hypothesis that $\beta = 0$ against the two-sided alternative.

a. What values of R^2 would allow you to reject the null hypothesis at a 5% significance level?
b. What values of R^2 would allow you to reject the null hypothesis at a 1% significance level?

Exercise 20.6.3. In Exercise 20.5.6, which considered the relation between the concentration of a chemical in tap water (x, in parts per million) and the cancer rate (y, in new cases per 1000 people during the past year), the sample size was 12, and it can be shown that $R^2 = .2069$.

a. Can you reject the null hypothesis that $\beta = 0$ in favor of the two-sided alternative at a 5% significance level?
b. What is the lowest significance level at which the null hypothesis can be rejected?

Exercise 20.6.4. Exercise 20.5.7 considered the relation between price (x, in dollars per gallon) and per capita water usage (y, in gallons per person per day) in California towns. The sample size was 15, and it can be shown that the sample correlation is $R = -0.5869$. Consider testing the null hypothesis that $\beta = 0$ against the two-sided alternative. What is the lowest significance level at which the null hypothesis can be rejected?

Exercise 20.6.5. Revisit Exercises 20.3.3 and 20.4.5, in which a pediatrician studied average number of sleep hours (x) and test scores (y) of 13-year-olds. The researcher used a sample of size 72, and her regression resulted in an R^2 of .3821. Suppose that the conditional distributions of test scores are approximately normal.

a. Can she reject the null hypothesis that there is no relation between hours of sleep and expected test scores in favor of the two-sided alternative at a significance level of .001?
b. What is the lowest significance level at which the null hypothesis could be rejected?

Exercise 20.6.6. Carefully explain the differences between the sum of squares equation (19.16) for population regressions Chapter 19 and its counterpart (20.24) for regressions based on random samples.

Exercise 20.6.7. Express the F-statistic in terms of two of the sums of squares from equation (20.24).

Exercise 20.6.8. Use fact (20.29) to show that in the classical regression model, if the sample size n is large enough, we should reject the null hypothesis $H_0 : \beta = 0$ in favor of the alternative $H_1 : \beta \neq 0$ at significance level a if

$$(n-2)\frac{R^2}{1-R^2} > c^1_{1-a}.$$

(Recall from Appendix 17.A (online) that the right-tail χ^2-value c^d_{1-a} is defined by $\mathrm{P}(C > c^d_{1-a}) = a$, where $C \sim \chi^2(d)$.)

Section 20.7 exercises

Exercise 20.7.1. In a study of a sample of hospitals from mid-Atlantic states, a regression of survival rates (the percentage of admitted patients who survived and were discharged) on the number of physicians per patient returned a negative OLS estimate of β. What confounding variable might account for this result?

Exercise 20.7.2. Recall Exercises 20.3.5 and 20.4.7, which considered weekly personal Internet use (x, in minutes) and sales volume (y, in thousands of dollars) of sales employees. The sample was assumed to satisfy the conditions of the random sampling regression model. Is it reasonable to interpret β as the expected causal effect of Internet use on sales volume? Discuss.

Exercise 20.7.3. Recall Exercises 20.3.1 and 20.4.3, which considered shift lengths (x) and incomes (y) for taxi drivers. The sample was assumed to satisfy the conditions of the random sampling regression model.

a. Suppose that passengers hail taxis from the sidewalks, and that after a driver drops off a passenger it is easy to find another fare. In this case, is it reasonable to interpret β as the causal effect of shift length on income? Explain.
b. Now suppose that it is not always so easy to find passengers, and experienced drivers are more skilled at locating them. Also, assume that drivers who work long shifts tend to be long-time employees. In this case, is it reasonable to interpret β as the causal effect of shift length on income? Explain.

Exercise 20.7.4. Consider Exercise 20.5.7, which considered the relationship between the price of water (x, in dollars per gallon) and per capita water usage (y, in gallons per person per day) in Southern California towns. The sample was assumed to satisfy the conditions of the random sampling regression model.

a. In the original story, water was provided by a public utility with prices set by legislative fiat. Is it reasonable to interpret β as the causal effect of the price of water on water usage? Explain.
b. Now suppose that the towns were instead served by several privately owned water suppliers. Explain why in this case it may not be appropriate to give β a causal interpretation.

Chapter exercises

Exercise 20.C.1. An education researcher is studying the effect of study time on eighth-graders' test scores. He obtains a stratified sample of 120 children, recording the amount of studying the week before the exam (x, in hours) and test scores (y). His descriptive statistics and estimates are as follows:

$$\begin{array}{lll} \bar{X} = 2.5 & A = 56.0 & \\ s_X^2 = .81 & B = 7.5 & \mathcal{S}^2 = 8.8. \end{array}$$

Assume that the sample satisfies the assumptions of the classical regression model.

a. Interpret β.
b. Construct a .95 confidence interval for β.
c. Do you think the regression model has a causal interpretation? Explain.

Exercise 20.C.2. A fiction writer is trying to finish the much-anticipated next book in his epic series and wants to know whether his coffee consumption is helping or hurting his output. He varies his coffee intake (x, in cups) over the course of 20 working days according to a predetermined schedule, and records how many pages he wrote each day (y). His descriptive statistics and estimates are as follows:

$$\begin{array}{lll} \bar{x} = 3.5 & A = 7.22 & \\ s_x^2 = 1.21 & B = .25 & \mathcal{S}^2 = 2.7. \end{array}$$

Assume that this experiment satisfies the assumptions of the classical normal regression model.

a. Test the null hypothesis that $\beta = 0$ against the alternative that $\beta > 0$ at a significance level of 5%.
b. Construct a .90 confidence interval for the writer's expected output when he drinks 5 cups of coffee.
c. Construct a .90 prediction interval for the writer's output the next time he drinks 5 cups of coffee.
d. Explain why your answers to parts (b) and (c) differ in the way that they do.

Exercise 20.C.3. A doctor is investigating the effect of exercise on cholesterol levels among those whose cholesterol levels are high. She recruits 65 subjects with high LDL ("bad") cholesterol, who agree to monitor amount the time (x, in hours) spent exercising over a 90-day period and to have their blood drawn to find the change in cholesterol levels since the beginning of the period (y, in mg/dL (milligrams per deciliter)). His descriptive statistics and estimates are as follows:

$$\bar{X} = 30 \qquad A = -1.20$$

$$\mathcal{S}_X^2 = 9.77 \qquad B = -.075 \qquad \mathcal{S}^2 = 15.58.$$

Assume that the sample satisfies the assumptions of the random sampling normal regression model.

a. Consider testing the null hypothesis that β is greater than or equal to 0 against the alternative that it is less than 0. What is the P-value of the sample?
b. Do you think β has a causal interpretation? Explain.

Exercise 20.C.4. You work for a major soft drink maker and would like to better understand the impact of advertising on your firm's sales. To do so you perform an experiment in 15 cities of similar sizes and with similar sales figures prior to the experiment. You assign each city an advertising budget (x, in dollars) for one month, and then record the sales figures (y, in dollars) for that city during the second half of that month and the first half of the next month. The results of the experiment can be found in the `advertising` worksheet. Assume that the experiment satisfies the assumptions of the classical normal regression model.

a. Report the sample regression line and the sample conditional variance.
b. Consider testing the null hypothesis that $\beta = 0$ against the alternative hypothesis that $\beta > 0$. What is the P-value of your sample?
c. Construct a 95% confidence interval for expected sales in a market in which the brewer spends \$20,000 on advertising.
d. Construct a 95% confidence interval for expected sales in a market in which the brewer spends \$15,000 on advertising.
e. Explain why the widths of the intervals from parts (c) and (d) differ in the way that they do.

Exercise 20.C.5. The Department of Education wants to estimate the effect of school district funding on the preparedness of high school graduates for college, as measured by their average SAT scores. The `schools.xls` worksheet contains data on per pupil funding (x, in dollars) and average SAT score (y) for 120 randomly chosen school districts in parts of the country where the SAT is the standard college entrance exam. Assume that the sample satisfies the assumptions of the random sampling normal regression model.

a. Report the sample regression line and the sample conditional variance.
b. Consider testing the null hypothesis that $\beta = .02$ against the alternative hypothesis that $\beta > .02$. What is the P-value of your sample? Can you reject the null hypothesis at a 1% significance level?

c. Construct a 95% confidence interval for the mean average SAT score in a districts in which spending per pupil is $10,000.
d. Do you think that β has a causal interpretation? Explain.

Exercise 20.C.6. The admissions office at a state university's flagship campus is determining how much weight to put on an applicant's high school GPA in making admissions decisions. It has collected the high school GPAs (x) and the college first-year GPAs (y) of a random sample of 75 students who just completed their first year at the university. The data is compiled in the `GPAs` worksheet. Assume that the sample satisfies the assumptions of the random sampling regression model.

a. Report the sample regression line.
b. What is the mean first-year GPA of students in the sample whose high school GPA was 4.0?
c. What is the best linear prediction of the mean first-year GPA of students whose high school GPA was 4.0?
d. Explain in detail why your answers to parts (b) and (c) differ.
e. Construct a 95% confidence interval for the first-year GPA of a student with a 4.0 high school GPA.
f. Construct a 95% confidence interval for the first-year GPA of a student with a 3.0 high school GPA.
g. Explain why the widths of the intervals from parts (e) and (f) differ in the way that they do.
h. Consider testing the null hypothesis that $\beta = 0$ against the alternative hypothesis that $\beta > 0$. What is the P-value of the sample?

Exercise 20.C.7. A pesticide company is conducting research on the effectiveness of its product at eliminating potato beetles. It has conducted an experiment by stocking 10 greenhouses with 200 potato plants and 1000 potato beetles each, varying the amount of pesticide solution (x, in liters) sprayed in each, and then counting each greenhouse's beetle population (y). The results of this experiment are given in the `potatoes` worksheet. Assume that the experiment satisfies the assumptions of the classical normal regression model.

a. Report A, B, and $\mathcal{S}^2$.
b. For what amount of pesticide solution would your point estimate of the expected change in the beetle population equal zero?
c. Construct a 95% confidence interval for β.
d. Construct a 95% confidence interval for α. Give an interpretation of both α and the confidence interval in the context of this example.

Exercise 20.C.8. A private toll bridge operator is considering raising the toll. To estimate the demand curve, the operator performs an experiment, setting a different toll each week over the course of 15 weeks. The results of this experiment are presented in the `tolls` worksheet, first in their original units (prices in dollars, and numbers of vehicles), and then after a logarithmic transformation. Assume that the assumptions of the classical normal regression model hold for the transformed variables, logarithm of price and logarithm of number of vehicles.

a. Explain what the assumptions of the classical normal regression model require in terms of the original (untransformed) variables.
b. Report the sample regression line and the sample conditional variance.

When a demand function is represented using logarithmically transformed variables, the absolute value of this function's slope is the *price elasticity of demand*, a measure of the sensitivity of quantity demanded to price charged.[31] Demand is said to be *elastic* when elasticity is greater than 1, and *inelastic* when elasticity is less than 1.

Under the assumptions of our regression models, the line $y = \alpha + \beta x$ is the expected demand curve. Thus $|\beta|$ is the price elasticity of *expected* demand and is assumed to be independent of the price charged.

c. Test the null hypothesis that expected demand has elasticity less than 1 against the alternative hypothesis that expected demand is elastic at a 5% significance level.
d. What is the P-value of the sample for the hypothesis test in part (c)?

Mathematical exercises

Exercise 20.M.1. Consider the classical regression model.

a. Starting from the formula

$$B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

show that B can be written in the linear form

$$B = \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{(n-1)s_x^2}\right)Y_i.$$

(Hint: Use the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ to eliminate $\bar{Y}$.)

[31] If p represents price, q quantity, and $q = Q(p)$ the demand function, the *elasticity of demand* at price p is defined as $E(p) = |\frac{\mathrm{d}}{\mathrm{d}p}Q(p) \cdot \frac{p}{Q(p)}|$. To establish the claim in the text for natural logarithms, write $\tilde{p} \equiv \ln p$ and $\tilde{q} \equiv \ln q$, and let $\tilde{q} = \tilde{Q}(\tilde{p})$ be the demand function expressed in terms of the transformed variables. Then $\tilde{Q}(\tilde{p}) = \ln Q(\exp(\tilde{p}))$ by definition, so the claim follows from differentiation (via the chain rule) and substitution:

$$\left|\frac{\mathrm{d}}{\mathrm{d}\tilde{p}}\tilde{Q}(\tilde{p})\right| = \left|\frac{1}{Q(\exp(\tilde{p}))} \cdot \frac{\mathrm{d}}{\mathrm{d}p}Q(\exp(\tilde{p})) \cdot \exp(\tilde{p})\right| = \left|\frac{1}{Q(p)} \cdot \frac{\mathrm{d}}{\mathrm{d}p}Q(p) \cdot p\right| = E(p).$$

If the base 10 logarithm is used instead, a very similar calculation yields the same result.

b. Starting from the formula $A = \bar{Y} - B\bar{x}$, and again using the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, show that A can be written in the linear form

$$A = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{x}\,(x_i - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right)Y_i.$$

Exercise 20.M.2. Using the assumptions of the classical regression model and the basic facts about means of linear functions of random variables, show that

a. $\mathrm{E}(\bar{Y}) = \alpha + \beta\bar{x}$.

b. $\mathrm{E}(A) = \alpha$.

Exercise 20.M.3. Consider the classical regression model

a. Show that

$$\mathrm{Var}(A) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right),$$

and hence that $\mathrm{Var}(A) = \frac{\sigma^2}{n} + \bar{x}^2\sigma_B^2$.

b. Show that

$$\mathrm{Cov}(A, B) = -\frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

and hence that $\mathrm{Cov}(A, B) = -\bar{x}\sigma_B^2$.

(Hints: Start from the expressions for A and B as a linear functions of the Y_i, and use the formulas for variances and covariances of linear functions of random variables from Chapters 3 and 4. During the calculation, use the fact that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ as necessary to simplify the expressions you obtain.)

Exercise 20.M.4.

a. Prove that the OLS estimator A has a lower variance than any other unbiased linear estimator of α. (Hint: Starting from the linear expression for A in Section 20.2.2, show that $A = \sum_{i=1}^{n}(r + sx_i)Y_i$ for some choices of r and s. Then follow the line of argument from Appendix 20.A.4.)

b. Prove that the OLS estimator $A + Bx$ has a lower variance than any other unbiased linear estimator of $\mathrm{E}(Y|x) = \alpha + \beta x$.

Exercise 20.M.5.

a. Suppose that in the classical regression model, an experimenter chooses $x_1 = 1$, followed by $x_i = 0$ for all $i > 1$. Show that in this case, the variance of B does not approach 0 as the sample size grows large, implying that B is not a consistent estimator of β for this choice of x values.

b. Find a specification of the x_i for which A is not a consistent estimator of α.

Exercise 20.M.6. Show that in the random sampling regression model, the sample covariance $\mathcal{S}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is an unbiased estimator of the covariance $\sigma_{xy} = \text{Cov}(X_i, Y_i)$. (Hint: Mimic the proof that the sample variance is an unbiased estimator of the variance (Appendix 20.A.5.))

Exercise 20.M.7. Use the information about the distribution of $Y - (A + Bx)$ from equation (20.21) to show that

$$\text{P}(Y \in [(A + Bx) - d, (A + Bx) + d]) = 1 - c,$$

$$\text{where } d = z_{c/2}\sigma\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2}}.$$

This formula tells us the endpoints of the $1 - c$ prediction interval for Y when σ is known.

Exercise 20.M.8. Show that in the random sampling regression model, since $\text{E}[B|X_1 = x_1, \ldots, X_n = x_n] = \beta$ for all realizations $x_1, \ldots, x_n$, it must be that $\text{E}(B) = \beta$. (Hint: Use the law of iterated expectation (Exercise 4.M.2).)

Exercise 20.M.9. Consider the *random sampling normal regression model*, which is obtained from the random sampling regression model by adding the assumption that Y_i is normally distributed conditional on $X_i = x_i$ for any value of x_i.

a. Show that conditional on the event $\{X_1 = x_1, \ldots, X_n = x_n\}$, B is normally distributed with mean β and variance $\sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2$. Thus B is not normally distributed itself, but instead has what is known as a *mixture distribution* composed from normal distributions.

b. Use part (a) to show that

$$Z_B = \frac{B - \beta}{\dfrac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

has a standard normal distribution. This result provides a basis for our inference procedures for β under random sampling. (Hint: Apply the law of iterated expectation to an indicator random variable that equals 1 when $Z_B \leq z$, where z is an arbitrary number.)