DATA ANALYSIS FOR BUSINESS, ECONOMICS, AND POLICY

Gábor Békés
Central European University, Vienna and Budapest

Gábor Kézdi
University of Michigan, Ann Arbor

CAMBRIDGE UNIVERSITY PRESS

10 Multiple Linear Regression

Why and how to carry out multiple linear regression analysis, and how to interpret its results

Motivation

There is a substantial difference in the average earnings of women and men in all countries. You want to understand more about the potential origins of that difference, focusing on employees with a graduate degree in your country. You have data on a large sample of employees with a graduate degree, with their earnings and some of their characteristics, such as age and the kind of graduate degree they have. Women and men differ in those characteristics, which may affect their earnings. How should you use this data to uncover gender difference that are not due to differences in those other characteristics? And can you use regression analysis to uncover patterns of associations between earnings and those other characteristics that may help understand the origins of gender differences in earnings?

You have analyzed your data on hotel prices in a particular city to find hotels that are underpriced relative to how close they are to the city center. But you have also uncovered differences in terms of other features of the hotels that measure quality and are related to price. How would you use this data to find hotels that are underpriced relative to all of their features? And how can you visualize the distribution of hotel prices relative to what price you would expect for their features in a way that helps identify underpriced hotels?

After understanding simple linear regression, we can turn to multiple linear regression, which has more than one explanatory variable. Multiple linear regression is the most used method to uncover patterns of associations between variables. There are multiple reasons to include more explanatory variables in a regression. We may be interested in uncovering patterns of association between y and several explanatory variables, which may help uncover patterns of association that could be investigated in subsequent analysis. Or, we may be interested in the effect of an x variable, but we want to compare observations that are different in x but similar in other variables. Finally, we may want to predict y, and we want to use more x variables to arrive at better predictions.

We discuss why and when we should estimate multiple regression, how to interpret its coefficients, and how to construct and interpret confidence intervals and test the coefficients. We discuss the relationship between multiple regression and simple regression. We explain that piecewise linear splines and polynomial regressions are technically multiple linear regressions without the same interpretation of the coefficients. We discuss how to include categorical explanatory variables as well as interactions that help uncover different slopes for groups. We include an informal discussion on how to decide what explanatory variables to include and in what functional form. Finally, we discuss why a typical multiple regression with observational data can get us closer to causal interpretation without fully uncovering it.

The first case study in this chapter, **Understanding the gender difference in earnings**, uses the `cps-earnings` dataset to illustrate the use of multiple regression to understand potential sources of gender differences in earnings. We also go back to our question on finding underpriced hotels relative to their location and quality in the case study **Finding a good deal among hotels with multiple regression**, using the `hotels-vienna` dataset, to illustrate the use of multiple regression in prediction and residual analysis.

Learning outcomes

After working through this chapter, you should be able to

- identify questions that are best answered with the help of multiple regression from available data;
- estimate multiple linear regression coefficients and present and interpret them;
- estimate appropriate standard errors, create confidence intervals and tests of regression coefficients, and interpret those;
- select the variables to include in a multiple regression guided by the purpose of the analysis;
- understand the relationship between the results of a multiple regression and causal effects when using observational data.

10.1 Multiple Regression: Why and When?

There are three broad reasons to carry out multiple regression analysis instead of simple regression. The first is exploratory data analysis: we may want to uncover more patterns of association, typically to generate questions for subsequent analysis. The other two reasons are the two ultimate aims of data analysis: making a better prediction by explaining more of the variation, and getting closer to establishing cause and effect in observational data by comparing observations that are more comparable.

The first example of the introduction is about understanding the reasons for a difference. It's a causal question of sorts: we are interested in what causes women to earn less than men. The second example is one of prediction: we want to capture average price related to hotel features that customers value in order to identify hotels that are inexpensive compared to what their price "should be."

10.2 Multiple Linear Regression with Two Explanatory Variables

Multiple regression analysis uncovers average y as a function of more than one x variable: $y^E = f(x_1, x_2, ...)$. It can lead to better predictions of $\hat{y}$ by considering more explanatory variables. It may improve the interpretation of slope coefficients by comparing observations that are different in terms of one of the x variables but similar in terms of all other x variables.

Multiple linear regression specifies a linear function of the explanatory variables for the average y. Let's start with the simplest version with two explanatory variables:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{10.1}$$

This is the standard notation for a multiple regression: all coefficients are denoted by the Greek letter β, but they have subscripts. The intercept has subscript 0, the first explanatory variable and its coefficient have subscript 1, and the second explanatory variable and its coefficient have subscript 2.

Having another right-hand-side variable in the regression means that we further condition on that other variable when we compare observations. The slope coefficient on x_1 shows the difference in average y across observations with different values of x_1 but with the same value of x_2. Symmetrically, the slope coefficient on x_2 shows difference in average y across observations with different values of x_2 but with the same value of x_1. This way, multiple regression with two explanatory variables compares observations that are similar in one explanatory variable to see the differences related to the other explanatory variable.

The interpretation of the slope coefficients takes this into account. β_1 shows how much larger y is on average for observations in the data with one unit larger value of x_1 but the same value of x_2. β_2 shows how much larger y is on average for observations in the data with one unit larger value of x_2 but with the same value of x_1.

Review Box 10.1 Multiple linear regression

Multiple linear regression with two explanatory variables:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Interpretation of the coefficients:

- β_0 (intercept): average y for observations if both x_1 and x_2 are zero in the data.
- β_1 (slope of x_1): on average, y is β_1 units larger in the data for observations with one unit larger x_1 but with the same x_2.
- β_2 (slope of x_2): on average, y is β_2 units larger in the data for observations with one unit larger x_2 but with the same x_1.

10.3 Multiple Regression and Simple Regression: Omitted Variable Bias

It is instructive to examine the difference in the slope coefficient of an explanatory variable x_1 when it is the only variable on the right-hand side of a regression compared to when another explanatory variable, x_2, is included as well. In notation, the question is the difference between β in the simple linear regression

$$y^E = \alpha + \beta x_1 \tag{10.2}$$

and β_1 in the multiple linear regression

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{10.3}$$

For example, we may use time series data on sales and prices of the main product of our company, and regress month-to-month change in log quantity sold (y) on month-to-month change in log price (x_1). Then β shows the average percentage change in sales when our price increases by 1%.

For example, $\hat{\beta} = -0.5$ would show that sales tend to decrease by 0.5% when our price increases by 1%.

But we would like to know what happens to sales when we change our price but competitors don't. The second regression has change in the log of the average price charged by our competitors (x_2) next to x_1. Here β_1 would answer our question: average percentage change in our sales in months when we increase our price by 1%, but competitors don't change their price. Suppose that in that regression, we estimate $\hat{\beta_1} = -3$. That is, our sales tend to drop by 3% when we increase our price by 1% and our competitors don't change their prices. Also suppose that the coefficient on x_2 is positive, $\hat{\beta_2} = 3$. That means that our sales tend to increase by 3% when our competitors increase their prices by 1% and our price doesn't change.

As we'll see, whether the answer to the two questions is the same will depend on whether x_1 and x_2 are related. To understand that relationship, let us introduce the regression of x_2 on x_1, called the ***x*–*x* regression**, where δ is the slope parameter:

$$x_2^E = \gamma + \delta x_1 \tag{10.4}$$

In our own price–competitor price example, δ would tell us how much the two prices tend to move together. In particular, it tells us about the average percentage change in competitor price in months when our own price increases by 1%. Let's suppose that we estimate it to be $\hat{\delta} = 0.83$: competitors tend to increase their price by 0.83% when our company increases its price by 1%. The two prices tend to move together, in the same direction.

To link the two original regressions, plug this x–x regression back in the multiple regression (this step is fine even though that may not be obvious; see Under the Hood section 10.U1):

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2(\gamma + \delta x_1) = \beta_0 + \beta_2\gamma + (\beta_1 + \beta_2\delta)x_1 \tag{10.5}$$

Importantly, we find that with regards to x_1, the slope coefficients in the simple (β) and multiple regression (β_1) are different:

$$\beta - \beta_1 = \delta\beta_2 \tag{10.6}$$

The slope of x_1 in a simple regression is different from its slope in the multiple regression, the difference being the product of its slope in the regression of x_2 on x_1 and the slope of x_2 in the multiple regression. Or, put simply, the slope in simple regression is different from the slope in multiple regression by the slope in the x–x regression times the slope of the other x in the multiple regression.

This difference is called the **omitted variable bias**. If we are interested in the coefficient on x_1 with x_2 in the regression, too, it's the second regression that we need; the first regression is an incorrect regression as it omits x_2. Thus, the results from that first regression are biased, and the bias is caused by omitting x_2. We will discuss omitted variables bias in detail when discussing causal effects in Chapter 21, Section 21.3.

In our example, we had that $\hat{\beta} = -0.5$ and $\hat{\beta_1} = -3$, so that $\hat{\beta} - \hat{\beta_1} = -0.5 - (-3) = +2.5$. In this case our simple regression gives a biased estimate of the slope coefficient on x_1 compared to the multiple regression, and the bias is positive (the simple regression estimate is less negative). Recall that we had $\hat{\delta} = 0.83$ and $\hat{\beta_2} = 3$. Their product is approximately 2.5. This positive bias is the result of two things: a positive association between the two price changes (δ) and a positive association between competitor price and our own sales (β_2).

In general, the slope coefficient on x_1 in the two regressions is different unless x_1 and x_2 are uncorrelated ($\delta = 0$) or the coefficient on x_2 is zero in the multiple regression ($\beta_2 = 0$). The slope in the

simple regression is more positive or less negative if the correlation between x_2 and x_1 has the same sign as β_2 (both are positive or both are negative).

The intuition is the following. In the simple regression $y^E = \alpha + \beta x_1$, we compare observations that are different in x_1 without considering whether they are different in x_2. If x_1 and x_2 are uncorrelated this does not matter. In this case observations that are different in x_1 are, on average, the same in x_2, and, symmetrically, observations that are different in x_2 are, on average, the same in x_1. Thus the extra step we take with the multiple regression to compare observations that are different in x_1 but similar in x_2 does not matter here.

If, however, x_1 and x_2 are correlated, comparing observations with or without the same x_2 value makes a difference. If they are positively correlated, observations with higher x_2 tend to have higher x_1. In the simple regression we ignore differences in x_2 and compare observations with different values of x_1. But higher x_1 values mean higher x_2 values, too. Corresponding differences in y may be due to differences in x_1 but also due to differences in x_2.

In our sales–own price–competitors' price example, the drop in sales when our own price increases (and competitors do what they tend to do) is smaller than the drop in sales when our own price increases but competitor prices don't change. That's because when our own price increases, competitors tend to increase their prices, too, which in itself would push up our sales. The two work against each other: the increase in our price makes sales decrease, but the increase in competitors' prices that tends to happen at the same time makes sales increase.

Review Box 10.2 Multiple linear regression and simple linear regression

- The difference of slope β in $y^E = \alpha + \beta x_1$ and the slope β_1 in $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is

$$\beta - \beta_1 = \delta\beta_2$$

  where δ is the slope in the regression of x_2 on x_1: $x_2^E = \gamma + \delta x_1$
- In words, briefly: slope in the simple regression differs from the slope in the multiple regression by the product of the slope in x–x regression and slope of other x in multiple regression.

10.A1 CASE STUDY – Understanding the Gender Difference in Earnings

Multiple linear regression

We continue investigating patterns in earnings, gender, and age. The data is the same `cps-earnings` dataset that we used earlier in Chapter 9, Section 9.A1: it is a representative sample of all people of age 15 to 85 in the USA in 2014.

Compared to Chapter 9, Section 9.A1, where we focused on a single occupation, we broaden the scope of our investigation here to all employees with a graduate degree – that is, a degree higher than a four-year college degree: these include professional, master's, and doctoral degrees.

We use data on people of age 24 to 65 (to reflect the typical working age of people with these kinds of graduate degrees). We excluded the self-employed (their earnings is difficult to measure) and included those who reported 20 hours or more as their usual weekly time worked. We have 18 241 observations.

The dependent variable is log hourly earnings (ln wage). Table 10.1 shows the results from three regressions: (1) is a simple regression of *ln wage* on a binary *female* variable; (2) is a multiple regression that includes age as well, in a linear fashion; and (3) is a simple regression with *age* as the dependent variable and the *female* binary variable.

Table 10.1 Gender differences in earnings – log earnings and gender

	(1)	(2)	(3)
Variables	**ln wage**	**ln wage**	**age**
female	−0.195**	−0.185**	−1.484**
	(0.008)	(0.008)	(0.159)
age		0.007**	
		(0.000)	
Constant	3.514**	3.198**	44.630**
	(0.006)	(0.018)	(0.116)
Observations	18 241	18 241	18 241
R-squared	0.028	0.046	0.005

Note: Robust standard error estimates in parentheses. ** $p<0.01$, * $p<0.05$.

Source: `cps-earnings` dataset. 2014, USA. Employees of age 24–65 with a graduate degree and 20 or more work hours per week.

According to column (1), women in this sample earn 19.5 log points (around 21%) less than men, on average. Column (2) suggests that when we compare employees of the same age, women in this sample earn 18.5 log points (around 20%) less than men, on average. This is a slightly smaller gender difference than in column (1). While the log approximation is not perfect at these magnitudes, from now on, we will ignore the difference between log units and percent. For example, we will interpret a 0.195 coefficient as a 19.5% difference.

The estimated coefficients differ, and we know where the difference should come from: average difference in age. Let's use the formula for the difference between the coefficient on *female* in the simple regression and in the multiple regression. Their difference is $-0.195-(-0.185)=-0.01$. This should be equal to the product of the coefficient of *female* in the regression of *age* on *female* (our column (3)) and the coefficient on *age* in column (2): $-1.48\times 0.007\approx -0.01$. These two are indeed equal.

Intuitively, we can see that women of the same age have a slightly smaller earnings disadvantage in this data because they are somewhat younger, on average, and employees who are younger tend to earn less. Part of the earnings disadvantage of women is thus due to the fact that they are younger. This is a small part, though: around one percentage point of the 19.5% difference, which is a 5% share of the entire difference.

But why are women employees younger, on average, in the data? It's because there are fewer female employees with graduate degrees over age 45 than below. Figure 10.1 shows two density plots overlaid: the age distributions of male and female employees with graduate degrees. There are relatively few below age 30. From age 30 and up, the age distribution is close to uniform for men. But not for women: the proportion of female employees with graduate degrees drops above age 45, and again above age 55.

In principle, this could be due to two things: either there are fewer women with graduate degrees in the 45+ generation than among the younger ones, or fewer of them are employed (i.e., employed for 20 hours or more for pay, which is the criterion to be in our subsample). Further investigation reveals that it is the former: women are less likely to have a graduate degree if they were born before 1970 (those 45+ in 2014) in the USA. The proportion of women working for pay for more than 20 hours is very similar among those below age 45 and above.

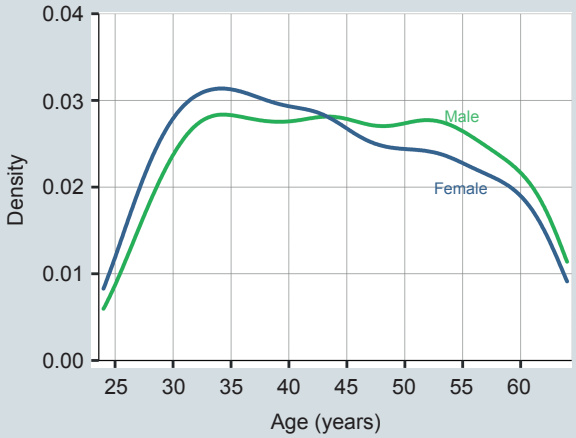


Figure 10.1 Age distribution of employees with graduate degree by gender

Note: Kernel density estimates of the age distribution of employees with a graduate degree; female and male employees separately.

Source: `cps-earnings` dataset. 2014, USA. Employees of age 24–65 with a graduate degree and 20 or more work hours per week. N=18 241.

10.4 Multiple Linear Regression Terminology

Multiple regression with two explanatory variables (x_1 and x_2) allows for assessing the differences in expected y across observations that differ in x_1 but are similar in terms of x_2. This difference is called **conditional** on that other explanatory variable x_2: difference in y by x_1, conditional on x_2. It is also called the controlled difference: difference in y by x_1, controlling for x_2. We often say that we condition on x_2, or control for x_2, when we include it in a multiple regression that focuses on average differences in y by x_1.

When we focus on x_1 in the multiple regression, the other right-hand-side variable, x_2, is called a **covariate**. In some cases, it is also called a **confounder**: if omitting x_2 makes the slope on x_1 different, it is said to confound the association of y and x_1 (we'll discuss confounders in Chapter 19, Section 19.14).

Review Box 10.3 Multiple linear regression terminology

In a regression $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ that focuses on β_1,

- if we estimate a multiple regression $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and we are interested in β_1, x_2 is a covariate, and we say that we condition on x_2 or control for x_2;
- if, instead, we estimate $y^E = \alpha + \beta x_1$, we say x_2 is an omitted variable.

10.5 Standard Errors and Confidence Intervals in Multiple Linear Regression

The concept of statistical inference and the interpretation of confidence intervals in multiple regressions is similar to that in simple regressions. For example, the 95% confidence interval of the slope coefficient of x_1 in a multiple linear regression shows where we can expect the coefficient in the population, or general pattern, represented by the data.

Similarly to the coefficients in the simple regression, the 95% CI of a slope in a multiple regression is the coefficient value estimated from the data plus-or-minus two standard errors. Again similarly to the simple regression case, we can get the standard error either by bootstrap or using an appropriate formula. And, as usual, the simple SE formula is not a good approximation in general: it assumes homoskedasticity (same fit of the regression over the range of the explanatory variables). There is a robust SE formula for multiple regression, too, that works in general, both under homoskedasticity and heteroskedasticity . Thus, just as with simple regressions we should make the software calculate **robust SE** as default.

While not correct in general, the simple formula is good to examine because it shows what makes the SE larger in a simpler more intuitive way than the robust formula. The simple SE formula for the slope $\hat{\beta}_1$ is

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1-R_1^2}} \tag{10.7}$$

Similarly to the simple SE formula for the simple linear regression in Chapter 9, Section 9.2, this formula has $\sqrt{n}$ in its denominator. But, similarly again to the simple linear regression, the correct number to divide with would be slightly different: the degrees of freedom instead of the number of the observations (see Under the Hood section 9.U4). Here that would be $\sqrt{n-k-1}$, where k is the number of right-hand-side variables in the regression. Similarly to the simple regression, this makes little practical difference in most cases. However, in contrast with the simple regression case, it may make a difference not only when we have too few observations, but also when we have many right-hand-side variables relative to the number of observations. We'll ignore that issue for most of this textbook, but it will come back in Chapter 21, Section 21.4.

This formula is very similar to what we have for simple regressions in other details, too, except for that new $\sqrt{1-R_1^2}$ term in the denominator. R_1^2 is the R-squared of the regression of x_1 on x_2. Recall that the R-squared of a simple regression is the square of the correlation between the two variables in the regression. Thus, R_1^2 is the square of the correlation between x_1 and x_2. The stronger this correlation, the larger R_1^2, the smaller $\sqrt{1-R_1^2}$, but then the larger $1/\sqrt{1-R_1^2}$ ($\sqrt{1-R_1^2}$ is in the denominator). So, the stronger the correlation between x_1 and x_2, the larger the SE of $\hat{\beta}_1$. Note the

symmetry: the same would apply to the SE of $\hat{\beta}_2$. As for the familiar terms in the formula: the SE is smaller, the smaller the standard deviation of the residuals (the better the fit of the regression), the larger the sample, and the larger the standard deviation of x_1.

At the polar case of a correlation of one (or negative one) that corresponds to $R_1^2 = 1$, the SE of the two coefficients does not exist. A correlation of one means that x_1 and x_2 are linear functions of each other. It is not only the SE formulae that cannot be computed in this case; the regression coefficients cannot be computed either. In this case the explanatory variables are said to be **perfectly collinear**.

Strong but imperfect correlation between explanatory variables is called **multicollinearity**. It allows for calculating the slope coefficients and their standard errors, but it makes the standard errors large. Intuitively, this is because we would like to compare observations that are different in one of the variables but similar in the other. But strong correlation between the two implies that there are not many observations that are the same in one variable but different in the other variable. Indeed, the problem of multicollinearity is very similar to the problem of having too few observations in general. We can see it in the formula as well: the role of $(1 - R^2)$ and n are the same.

Consider our example of using monthly data to estimate how sales of the main product of our company tend to change when our price changes but the prices of competitors do not. In that example our own price and the competitors' prices tended to move together. One consequence of this is that omitting the change in the competitors' price would lead to omitted variable bias; thus we need to include that in our regression. But here we see that it has another consequence. Including both price variables in the regression makes the SE of the coefficient of our own price larger, and its confidence interval wider, too. Intuitively, that's because there are fewer months when our price changes but the competitors' prices don't change, and it is changes in sales in those months that contain the valuable information for estimating the coefficient on our own price. Months when our own and competitors' prices change the same way don't help. So the reason why we want competitors' price in our regression (strong co-movement) is exactly the reason for having imprecise estimates with wide confidence intervals.

That's true in general, too. Unfortunately, there is not much we can do about multicollinearity in the data we have, just as there is not much we can do about having too few observations. More data helps both, of course, but that is not much help when we have to work with the data that's available. Alternatively, we may decide to change the specification of the regression and drop one of the strongly correlated explanatory variables. However, that results in a different regression. Whether we want a different regression or not needs to be evaluated keeping the substantive question of the analysis in mind.

Review Box 10.4 Inference in multiple regression

- In the linear regression $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, the 95% confidence interval (CI) of $\hat{\beta}_1$ tells us about the range in which we can expect, with 95% confidence, the difference in y to fall in the general pattern, or population, that our data represents, when comparing observations with the same x_2 but differing in x_1 by one unit.
- In the linear regression $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, the simple SE formula of $\hat{\beta}_1$ is

$$SE(\hat{\beta}_1) = \frac{Std[e]}{\sqrt{n}Std(x_1)\sqrt{1 - R_1^2}}$$

where e is the residual $e = y - \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ and R_1^2 is the R-squared in the simple linear regression of x_1 on x_2.

- The standard error of $\hat{\beta}_1$ is smaller:
  - the smaller the standard deviation of the residual (the better the fit of the regression);
  - the larger the sample;
  - the larger the variance of x_1;
  - the smaller the correlation between x_1 and x_2 (the smaller R_1^2).

10.6 Hypothesis Testing in Multiple Linear Regression

Testing hypotheses about coefficients in a multiple regression is also very similar to that in a simple regression. The standard errors are estimated in a different way but with the appropriate SE, all works just the same. For example, testing whether $H_0: \beta_1 = 0$ against $H_A: \beta_1 \neq 0$, we need the p-value or the t-statistic. Standard regression output produced by most statistical software shows those statistics. If our level of significance is 0.05, we reject H_0 if the p-value is less than 0.05, or – which is the same information in a different form – the t-statistic is less than -2 or greater than +2.

Besides testing a hypothesis that involves a single coefficient, we sometimes test a hypothesis that involves more coefficients. As we explained in Chapter 9, Section 9.5, these come in two forms: a single null hypothesis about two or more coefficients (e.g., if they are equal), or a list of null hypotheses (e.g., that several slope coefficients are zero). The latter is called testing joint hypotheses .

Testing joint hypotheses are based on a test statistic called the F-statistic, and the related test is called the **F-test**. The underlying logic of hypothesis testing is the same here: reject the null if the test statistic is larger than a critical value, which shows that the estimated coefficients are too far from what's in the null. The technical details are different. But the meaning of the p-value is the same as always. Thus, we advise getting the p-value when testing a joint hypothesis.

In fact, the test that asks whether all slope coefficients are zero in the regression has its own name: the **global F-test**, or simply "the" F-test. Its results are often shown by statistical software by default. More frequently, we use joint testing of joint hypotheses to decide whether a subset of the coefficients (such as the coeffients on all geographical variables) are all zero.

Similarly to testing hypotheses about single coefficients, the F-test needs appropriate standard error estimates. In cross-sectional data, those appropriate estimates are usually the robust SE estimates.

10.A2 CASE STUDY – Understanding the Gender Difference in Earnings

Statistical inference

Let's revisit the results in Table 10.1, taking statistical inference into account. The data represents employees with a graduate degree in the USA in 2014. According to the estimate in column (1), women in this sample earn 19.5 percent less than men, on average. The appropriately estimated (robust) standard error is 0.008, implying a 95% CI of approximately [−0.21, −0.18]. We can be

95% confident that women earned 18 to 21 percent less, on average, than men among employees with graduate degrees in the USA in 2014.

Column (2) suggests that when we compare employees of the same age, women in this sample earn approximately 18.5 percent less than men, on average. The 95% CI is approximately [−0.20,−0.17]. It turns out that the estimated −0.195 in column (1) is within this CI, and the two CIs overlap. Thus it is very possible that there is no difference between these two coefficients in the population. We uncovered a difference in the data between the unconditional gender wage gap and the gender gap conditional on age. However, that difference is small. Moreover, it may not exist in the population. These two facts tend to go together: small differences are harder to pin down in the population, or general pattern, represented by the data. Often, that's all right. Small differences are rarely very important. When they are, we need more precise estimates, which may come with larger sample size.

10.7 Multiple Linear Regression with Three or More Explanatory Variables

We spent a lot of time on multiple regression with two right-hand-side variables. That's because that regression shows all the important differences between simple regression and multiple regression in intuitive ways. In practice, however, we rarely estimate regressions with exactly two right-hand-side variables. The number of right-hand-side variables in a multiple regression varies from case to case, but it's typically more than two. In this section we describe multiple regressions with three or more right-hand-side variables. Their general form is

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \tag{10.8}$$

All of the results, language, and interpretations discussed so far carry forward to multiple linear regressions with three or more explanatory variables. Interpreting the slope of x_1: on average, y is β_1 units larger in the data for observations with one unit larger x_1 but with the same value for all other x variables. The interpretation of the other slope coefficients is analogous. The language of multiple regression is the same, including the concepts of **conditioning**, **controlling**, **omitted**, or **confounder variables**.

The standard error of coefficients may be estimated by bootstrap or a formula. As always, the appropriate formula is the robust SE formula. But the simple formula contains the things that make even the robust SE larger or smaller. For any slope coefficient $\hat{\beta}_k$ the simple SE formula is

$$SE(\hat{\beta}_k) = \frac{Std[e]}{\sqrt{n}Std[x_k]\sqrt{1-R_k^2}} \tag{10.9}$$

Almost all is the same as with two right-hand-side variables. In particular, the SE is smaller, the smaller the standard deviation of the residuals (the better the fit of the regression), the larger the sample, and the larger the standard deviation of x_k. The new-looking thing is R_k^2. But that's simply the generalization of R_1^2 in the previous formula. It is the R-squared of the regression of x_k on all other x variables. The smaller that R-squared, the smaller the SE.

Review Box 10.5 Multiple linear regression with three or more explanatory variables

- Equation: $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots$
- Interpretation of β_k (slope of x_k):
  - On average, y is β_k units larger in the data for observations with one unit larger x_k but with the same value for all other x variables.
  - $SE(\hat{\beta}_k) = \frac{Std[e]}{\sqrt{n}Std[x_k]\sqrt{1-R_k^2}}$ where e is the regression residual and R_k^2 is the R-squared of the regression of x_k on all other x variables.

10.8 Nonlinear Patterns and Multiple Linear Regression

In Chapter 8 we introduced piecewise linear splines, quadratics, and other polynomials to approximate a nonlinear $y^E = f(x)$ regression.

From a substantive point of view, piecewise linear splines and polynomials of a single explanatory variable are not multiple regressions. They do not uncover differences with respect to one right-hand-side variable conditional on one or more other right-hand-side variables. Their slope coefficients cannot be interpreted as the coefficients of multiple regressions: it does not make sense to compare observations that have the same x but a different x^2. But such regressions are multiple linear regressions from a technical point of view. This means that the way their coefficients are calculated is the exact same way the coefficients of multiple linear regressions are calculated. Their standard errors are calculated the same way, too and so are their confidence intervals, test statistics, and p-values.

Testing hypotheses can be especially useful here, as it can help choose the functional form. With a piecewise linear spline, we can test whether the slopes are the same in adjacent line segments. If we can't reject the null that they are the same, we may as well join them instead of having separate line segments. Testing hypotheses helps in choosing a polynomial, too. Here an additional complication is that the coefficients don't have an easy interpretation in themselves. At the same time, testing if all nonlinear coefficients are zero may help decide whether to include them at all.

However, testing hypotheses to decide whether to include a higher-order polynomial has its issues. Recall that a multiple linear regression requires that the right-hand-side variables are not perfectly collinear. In other words, they cannot be linear functions of each other. With a polynomial on the right-hand side, those variables are exact functions of each other: x^2 is the square of x. But they are not a linear function of each other, so, technically, they are not perfectly collinear. That's why we can include both x and x^2 and, if needed, its higher-order terms, in a linear regression. While they are not perfectly collinear, explanatory variables in a polynomial are often highly correlated. That multicollinearity results in high standard errors, wide confidence intervals, and high p-values . As with all kinds of multicollinearity, there isn't anything we can do about that once we have settled on a functional form.

Importantly, when thinking about functional form, we should always keep in mind the substantive focus of our analysis. As we emphasized in Chapter 8, Section 8.1, we should go back to that original focus when deciding whether we want to include a piecewise linear spline or a polynomial to approximate a nonlinear pattern. There we said that we want our regression to have a good

approximation to a nonlinear pattern in x if our goal is prediction or analyzing residuals. We may not want that if all we care about is the average association between x and y, except if that nonlinearity messes up the average association. This last point is a bit subtle, but usually means that we may want to transform variables to relative changes or take logs if the distribution of x or y is very skewed.

Here we have multiple x variables. Should we care about whether each is related to average y in a nonlinear fashion? The answer is the same as earlier: yes, if we want to do prediction or analyze residuals; no, if we care about average associations (except we may want to have transformed variables here, too). In addition, when we focus on a single average association (with, say, x_1) and all the other variables (x_2, x_3, ...) are covariates to condition on, the only thing that matters is the coefficient on x_1. Even if nonlinearities matter for x_2 and x_3 themselves, they only matter for us if they make a difference in the estimated coefficient on x_1. Sometimes they do; very often they don't.

10.A3 CASE STUDY – Understanding the Gender Difference in Earnings

Nonlinear patterns and multiple linear regression

This step in our case study illustrates the point we made in the previous section. The regressions in Table 10.2 enter age in linear ways. Using part of the same data, in Chapter 9, Section 9.A2 we found that log earnings and age follow a nonlinear pattern. In particular, there we found that average log earnings are a positive and steep function of age for younger people, but the pattern becomes gradually flatter for the middle-aged and may become completely flat, or even negative, among older employees.

Should we worry about the nonlinear age–earnings pattern when our question is the average earnings difference between men and women? We investigated the gender gap conditional on age. Table 10.2 shows the results for multiple ways of doing it. Column (1) shows the regression with the unconditional difference that we showed in Table 10.1, for reference. Column (2) enters age in linear form. Column (3) enters it as quadratic. Column (4) enters it as a fourth-order polynomial.

The unconditional difference is −19.5%; the conditional difference is −18.5% according to column (2), and −18.3% according to columns (3) and (4). The various estimates of the conditional difference are very close to each other, and all of them are within each others' confidence intervals. Thus, apparently, the functional form for age does not really matter if we are interested in the average gender gap.

At the same time, all coefficient estimates of the high order polynomials are statistically significant, meaning that the nonlinear pattern is very likely true in the population and not just a chance event in the particular dataset. The R-squared of the more complicated regressions are larger. These indicate that the complicated polynomial specifications are better at capturing the patterns. That would certainly matter if our goal was to predict earnings. But it does not matter for uncovering the average gender difference in earnings.

Table 10.2 Gender differences in earnings – log earnings and age, various functional forms

Variables	(1) ln wage	(2) ln wage	(3) ln wage	(4) ln wage
female	−0.195**	−0.185**	−0.183**	−0.183**
	(0.008)	(0.008)	(0.008)	(0.008)
age		0.007**	0.063**	0.572**
		(0.000)	(0.003)	(0.116)
age^2			−0.001**	−0.017**
			(0.000)	(0.004)
age^3				0.000**
				(0.000)
age^4				−0.000**
				(0.000)
Constant	3.514**	3.198**	2.027**	−3.606**
	(0.006)	(0.018)	(0.073)	(1.178)
Observations	18 241	18 241	18 241	18 241
R-squared	0.028	0.046	0.060	0.062

Note: Robust standard error estimates in parentheses. ** $p<0.01$, * $p<0.05$.

Source: `cps-earnings` dataset. 2014 USA. Employees of age 24–65 with a graduate degree and 20 or more work hours per week.

10.9 Qualitative Right-Hand-Side Variables

A great advantage of multiple linear regression is that it can deal with binary and other qualitative explanatory variables (also called categories, factor variables), together with quantitative variables, on the right-hand side.

To include such variables in the regression, we need to have them as binary, zero–one variables – also called **dummy variables** in the regression context. That's straightforward for variables that are binary to begin with: assign values zero and one (as we did with *female* in the case study). We need to transform other kinds of qualitative variables into binary ones, too, each denoting whether the observation belongs to that category (one) or not (zero). Then we need to include all those binary variables in the regression. Well, all except one.

We should select one binary variable denoting one category as a **reference category**, or **reference group** – also known as the "left-out category." Then we have to include the binary variables for all other categories but not the reference category. That way the slope coefficient of a binary variable created from a qualitative variable shows the difference between observations in the category captured by the binary variable and the reference category. If we condition on other explanatory variables, too, the interpretation changes in the usual way: we compare observations that are similar in those other explanatory variables.

As an example, suppose that x is a categorical variable measuring the level of education with three values $x = low, medium, high$. We need to create binary variables and include two of the three in the

regression. Let the binary variable x_{med} denote if $x = \textit{medium}$, and let the binary x_{high} variable denote if $x = \textit{high}$. Include x_{med} and x_{high} in the regression. The third potential variable for $x = \textit{low}$ is not included. It is the reference category.

$$y^E = \beta_0 + \beta_1 x_{med} + \beta_2 x_{high} \tag{10.10}$$

Let us start with the constant, β_0; this shows average y in the reference category. Here, β_0 is average y when both $x_{med} = 0$ and $x_{high} = 0$: this is when $x = \textit{low}$. β_1 is the difference in average y between observations that are different in x_{med} but the same in x_{high}. Thus β_1 shows the difference of average y between observations with $x = \textit{medium}$ and $x = \textit{low}$, the reference category. Similarly, β_2 shows the difference of average y between observations with $x = \textit{high}$ and $x = \textit{low}$, the reference category.

Which category to choose for the reference? In principle that should not matter: choose a category and all others are compared to that, but we can easily compute other comparisons from those. For example, the difference in y^E between observations with $x = \textit{high}$ and $x = \textit{medium}$ in the example above is simply $\beta_2 - \beta_1$ (both coefficients compare to $x = \textit{low}$, and that drops out of their difference). But the choice may matter for practical purposes. Two guiding principles may help this choice, one substantive, one statistical. The substantive guide is simple: we should choose the category to which we want to compare the rest. Examples include the home country, the capital city, the lowest or highest value group. The statistical guide is to choose a category with a large number of observations. That is relevant when we want to infer differences from the data for the population, or general pattern, it represents. If the reference category has very few observations, the coefficients that compare to it will have large standard errors, wide confidence intervals, and large p-values.

Review Box 10.6 Qualitative right-hand-side variables in multiple linear regression

- We should include qualitative right-hand-side variables with more categories as a series of binary ("dummy") variables.
- For a qualitative right-hand-side variable with k categories, we should enter $k-1$ binary variables; the category not represented by those binary variables is the reference category.
- Coefficients on each of the $k - 1$ binary variables show average differences in y compared to the reference category.

10.A4 CASE STUDY – Understanding the Gender Difference in Earnings

Qualitative variables

Let's use our case study to illustrate qualitative variables as we examine earnings differences by categories of educational degree. Recall that our data contains employees with graduate degrees. The dataset differentiates three such degrees: master's (including graduate teaching degrees, MAs, MScs, MBAs), professional (including MDs), and PhDs.

Table 10.3 shows the results from three regressions. As a starting point, column (1) repeats the results of the simple regression with *female* on the right-hand side; column (2) includes two education categories *ed_Profess* and *ed_PhD*; and column (3) includes another set of education categories, *ed_Profess* and *ed_MA*. The reference category is MA degree in column (2) and PhD in column (3).

Table 10.3 Gender differences in earnings – log earnings, gender and education

	(1)	(2)	(3)
Variables	**ln wage**	**ln wage**	**ln wage**
female	−0.195**	−0.182**	−0.182**
	(0.008)	(0.009)	(0.009)
ed_Profess		0.134**	−0.002
		(0.015)	(0.018)
ed_PhD		0.136**	
		(0.013)	
ed_MA			−0.136**
			(0.013)
Constant	3.514**	3.473**	3.609**
	(0.006)	(0.007)	(0.013)
Observations	18 241	18 241	18 241
R-squared	0.028	0.038	0.038

Note: MA, Professional, and PhD are three categories of graduate degree. Column (2): MA is the reference category. Column (3): the reference category is Professional or PhD. Robust standard error estimates in parentheses. ** $p<0.01$, * $p<0.05$.

Source: `cps-earnings` dataset. USA, 2014. Employees of age 24–65 with a graduate degree and 20 or more work hours per week.

The coefficients in column (2) of Table 10.3 show that comparing employees of the same gender, those with a professional degree earn, on average, 13.4% more than employees with an MA degree, and those with a PhD degree earn, on average, 13.6% more than employees with an MA degree. The coefficients in column (3) show that, among employees of the same gender, those with an MA degree earn, on average, 13.6% less than those with a PhD degree, and those with a professional degree earn about the same on average as those with a PhD degree. These differences are consistent with each other.

This is a large dataset so confidence intervals are rather narrow whichever group we choose as a reference category. Note that the coefficient on *female* is smaller, −0.182, when education is included in the regression. This suggests that part of the gender difference is due to the fact that women are somewhat more likely to be in the lower-earner MA group than in the higher-earner professional or PhD groups. But only a small part. We shall return to this finding later when we try to understand the causes of gender differences in earnings.

10.10 Interactions: Uncovering Different Slopes across Groups

Including binary variables for various categories of a qualitative variable uncovers average differences in y. But sometimes we want to know something more: whether and how much the slope with respect to a third variable differs by those categories. Multiple linear regression can uncover that too, with appropriate definition of the variables.

More generally, we can use the method of linear regression analysis to uncover how association between y and x varies by values of a third variable z. Such variation is called an **interaction**, as it shows how x and z interact in shaping average y. In medicine, when estimating the effect of x on y, if that effect varies by a third variable z, that z is called a **moderator variable**. Examples include whether malnutrition, immune deficiency, or smoking can decrease the effect of a drug to treat an illness. Non-medical examples of interactions include whether and how the effect of monetary policy differs by the openness of a country, or whether and how the way customer ratings are related to hotel prices differs by hotel stars.

Multiple regression offers the possibility to uncover such differences in patterns. For the simplest case, consider a regression with two explanatory variables: x_1 is quantitative; x_2 is binary. We wonder if the relationship between average y and x_1 is different for observations with $x_2 = 1$ than for $x_2 = 0$. How shall we uncover that difference?

A multiple regression that includes x_1 and x_2 estimates two parallel lines for the y–x_1 pattern: one for those with $x_2 = 0$ and one for those with $x_2 = 1$.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{10.11}$$

The slope of x_1 is β_1 and is the same in this regression for observations in the $x_2 = 0$ group and observations in the $x_2 = 1$ group. β_2 shows the average difference in y between observations that are different in x_2 but have the same x_1. Since the slope of x_1 is the same for the two x_2 groups, this β_2 difference is the same across the range of x_1. This regression does not allow for the slope in x_1 to be different for the two groups. Thus, this regression cannot uncover whether the y–x_1 pattern differs in the two groups.

Denote the expected y conditional on x_1 in the $x_2 = 0$ group as y_0^E, and denote the expected y conditional on x_1 in the $x_2 = 1$ group as y_1^E. Then, the regression above implies that the intercept is different (higher by β_2 in the $x_2 = 1$ group) but the slopes are the same:
First group, $x_2 = 0$

$$y_0^E = \beta_0 + \beta_1 x_1 \tag{10.12}$$

Second group, $x_2 = 1$

$$y_1^E = \beta_0 + \beta_2 \times 1 + \beta_1 x_1 \tag{10.13}$$

If we want to allow for different slopes in the two x_2 groups, we have to do something different. That difference is including the interaction term. An **interaction term** is a new variable that is created from two other variables, by multiplying one by the other. In our case:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \tag{10.14}$$

Not only are the intercepts different; the slopes are different, too:

$$y_0^E = \beta_0 + \beta_1 x_1 \tag{10.15}$$

$$y_1^E = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1 \tag{10.16}$$

It turns out that the coefficients of this regression can be related to the coefficients of two simple regressions of y on x_1, estimated separately in the two x_2 groups:

$$y_0^E = \gamma_0 + \gamma_1 x_1 \tag{10.17}$$

$$y_1^E = \gamma_2 + \gamma_3 x_1 \tag{10.18}$$

What we have is $\gamma_0 = \beta_0$; $\gamma_1 = \beta_1$; $\gamma_2 = \beta_0 + \beta_2$; and $\gamma_3 = \beta_1 + \beta_3$.

In other words, the separate regressions in the two groups and the regression that pools observations but includes an interaction term yield exactly the same coefficient estimates. The coefficients of the separate regressions are easier to interpret. But the pooled regression with interaction allows for a direct test of whether the slopes are the same. $H_0: \beta_3 = 0$ is the null hypothesis for that test; thus the simple t-test answers this question.

We can mix these tools to build ever more complicated multiple regressions. Binary variables can be interacted with other binary variables . Binary variables created from qualitative explanatory variables with multiple categories can all be interacted with other variables. Piecewise linear splines or polynomials may be interacted with binary variables. More than two variables may be interacted as well. Furthermore, quantitative variables can also be interacted with each other, although the interpretation of such interactions is more complicated.

Review Box 10.7 Interactions of right-hand-side variables in multiple linear regression

- Interactions between right-hand-side variables in a linear regression allow for the slope coefficient of a variable to differ by values of another variable.
- Interactions between two right-hand-side variables are modeled in a linear regression as $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
- β_1 shows average differences in y corresponding to a one-unit difference in x_1 when $x_2 = 0$.
- β_2 shows average differences in y corresponding to a one-unit difference in x_2 when $x_1 = 0$.
- β_3 is the coefficient on the interaction term. It shows the additional average differences in y corresponding to a one-unit difference in x_1 when x_2 is one unit larger, too. (It's symmetrical in x_1 and x_2 so it also shows the additional average differences in y corresponding to a one-unit difference in x_2 when x_1 is one unit larger, too.)
- When one of the two right-hand-side variables is binary, a simpler interpretation is also true. Say, $x_2 = 0$ *or* 1. Then,
  - β_1 shows the average difference in y corresponding to a one-unit difference in x_1 when $x_2 = 0$;
  - $\beta_1 + \beta_3$ shows the average difference in y corresponding to a one-unit difference in x_1 when $x_2 = 1$;
  - the coefficients of the regression are the same as the coefficients of two separate regressions on two parts of the data, one with $x_2 = 0$ and one with $x_2 = 1$:
    if $x_2 = 0$: $y^E = \beta_0 + \beta_1 x_1$
    if $x_2 = 1$: $y^E = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$

10.A5 CASE STUDY – Understanding the Gender Difference in Earnings

Interactions

We turn to illustrating the use of interactions as we consider the question of whether the patterns with age are similar or different for men versus women. As we discussed, we can investigate this in two ways that should lead to the same result: estimating regressions separately for men and women and estimating a regression that includes age interacted with gender. This regression model with an interaction is

$$(\ln w)^E = \beta_0 + \beta_1 \times age + \beta_2 \times female + \beta_3 \times age * female \quad (10.19)$$

Table 10.4 shows the results with age entered in a linear fashion. Column (1) shows the results for women, column (2) for men, column (3) for women and men pooled, with interactions. To have a better sense of the differences, which are often small, the table shows coefficients up to three digits.

Table 10.4 Gender differences in earnings – log earnings, gender, age, and their interaction

	(1)	(2)	(3)
Variables	**Women**	**Men**	**All**
female			−0.036
			(0.035)
age	0.006**	0.009**	0.009**
	(0.001)	(0.001)	(0.001)
female × age			−0.003**
			(0.001)
Constant	3.081**	3.117**	3.117**
	(0.023)	(0.026)	(0.026)
Observations	9 685	8 556	18 241
R-squared	0.011	0.028	0.047

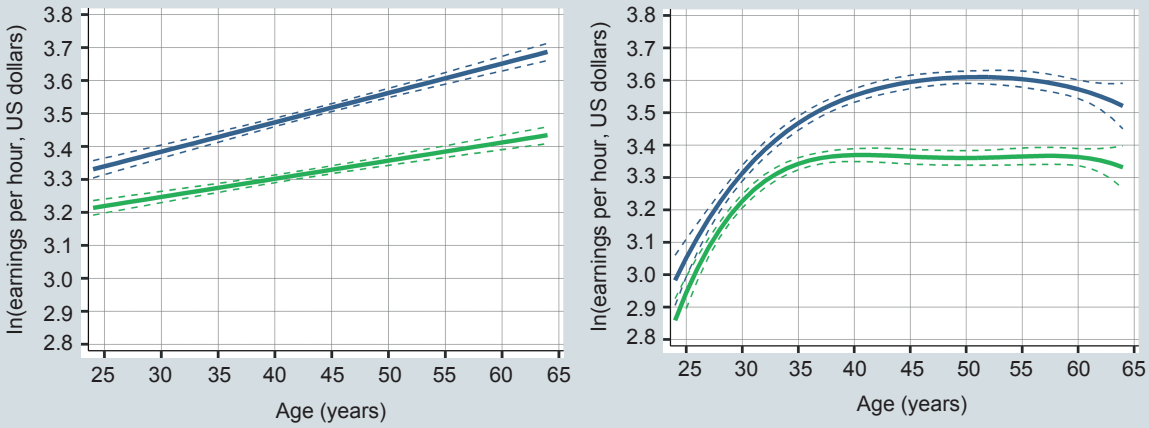
Note: Column (1) is women only; column (2) is men only; column (3) includes all employees. Robust standard error estimates in parentheses. ** $p<0.01$, * $p<0.05$.

Source: `cps-earnings` dataset. USA, 2014. Employees of age 24–65 with a graduate degree and 20 or more work hours per week.

According to column (1) of Table 10.4, women who are one year older earn 0.6% more, on average. According to column (2), men who are one year older earn 0.9% more, on average. Column (3) repeats the age coefficient for men. Then it shows that the slope of the log earnings–age pattern is, on average, 0.003 less positive for women, meaning that the earnings advantage of women who are one year older is 0.3 percentage points smaller than the earnings advantage of men who are one year older.

The advantage of the pooled regression, is its ability to allow for direct inference about gender differences. The 95% CI of the gender difference in the average pattern of ln wages and age is [−0.005, −0.001]. Among employees with a post-graduate degree in the USA in 2014, the wage difference corresponding to a one year difference in age was 0.1 to 0.5 percentage points less for women than for men. This confidence interval does not include zero. Accordingly, the t-test of whether the difference is zero rejects its null at the 5% level, suggesting that we can safely consider the difference as real in the population (as opposed to a chance event in the particular dataset); we are less than 5% likely to make a mistake by doing so.

The coefficient on the *female* variable in the pooled regression is −0.036. This is equal to the difference of the two regression constants: 3.081 for women and 3.117 for men. Those regression constants do not have a clear interpretation here (average log earnings when age is zero are practically meaningless). Their difference, which is actually the coefficient on *female* in the pooled regression, shows the average gender difference between employees with zero age. Similarly to the constants in the separate regressions, the coefficient is meaningless for any substantive purpose. Nevertheless, the regression needs it to have an intercept with the ln *w* axis.



(a) Log earnings on age interacted with gender

(b) Log earnings on fourth-order polynomial of age interacted with gender

Figure 10.2 Earning differences by gender as function of age

Note: Regression lines (curves) and 95% confidence intervals from regression of log earnings on age interacted with gender.
Source: `cps-earnings` dataset. USA, 2014. Employees of age 24–65 with a graduate degree and 20 or more work hours per week. N=18 241.

Taking the coefficients on *female*, *age*, and *female* × *age* together, the regression allows us to calculate the average gender difference by age. This exercise takes the linear functional form seriously, an assumption we know is false. We shall repeat this exercise with a better approximation of the nonlinear patterns. For now, let's stick to the linear specification, for educational purposes. The youngest people in our sample are 25 years old. Starting with the separate regressions, the predicted log wage for women of age 25 is $3.081 + 25 \times 0.006 \approx 3.231$. For men, $3.117 + 25 \times 0.009 \approx 3.342$. The difference is −0.11. We get the same number from the pooled regression: the gender difference at age 25 should be the gender difference at age zero implied by the coefficient on *female* plus 25 times the difference in the slope by age, the coefficient on the interaction term

female × *age*: $-0.036 + 25 \times -0.003 \approx -0.11$. Carrying out the same calculations for age 45 yields a difference of -0.17. These results imply that the gender difference in average earnings is wider for older ages.

Figure 10.2a shows the relationship graphically. It includes two lines with a growing gap: earnings difference is higher for older age. Remember, our regression can capture this growing gap because it includes the interaction. Without the interaction, we would not be able to see this, as that specification would force two parallel lines at constant distance.

However, we know that the pattern on age and (log) earnings is not linear. Our earlier results indicate that a fourth-order polynomial is a better approximation to that pattern. To explore whether the shapes of the age–earnings profiles are different between women and men, we re-estimated the regression with age in a fourth-order polynomial interacted with gender:

$$(\ln w)^E = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 age^4 + \beta_5 female + \beta_6 female \times age + \beta_7 female \times age^2 + \beta_8 female \times age^3 + \beta_9 female \times age^4 \quad (10.20)$$

This is a complicated regression with coefficients that are practically impossible to interpret. We don't show the coefficient estimates here. Instead we visualize the results. The graph in Figure 10.2b shows the predicted pattern (the regression curves) for women and men, together with the confidence intervals of the regression lines (curves here), as introduced in Chapter 9, Section 9.3.

Figure 10.2a suggests that the average earnings difference is a little less than 10% between ages 25 and 30, increases to around 15% by age 40, and reaches 22% by age 50, from where it decreases slightly to age 60 and more by age 65. These differences are likely similar in the population represented by the data as the confidence intervals around the regression curves are rather narrow, except at the two ends.

These results are very informative. Many factors may cause women with a graduate degree to earn less than men. Some of those factors are present at young age, but either they are more important in middle age, or additional factors start playing a role by then.

What can students infer from these results about the gender differences they may experience through their careers? Statistical inference established that the patterns are very likely present in the population represented by the data: employees with graduate degrees in the USA in 2014. The first question of external validity is whether similar patterns are likely to be true in the future as well and, if we are interested in another country, whether the patterns are similar there. The second question is the extent to which differences by age in the cross-section are informative about what we can expect to happen through time as people age. As we discussed earlier, those questions are impossible to answer with this data. Analyzing more data may help some but will never give a definitive answer to all questions. Nevertheless, the information produced by our analysis is a good starting point.

10.11 Multiple Regression and Causal Analysis

When interpreting regression coefficients, we advise being careful with the language, talking about differences and associations not effects and causation. But, can we say anything regarding the extent to which our results may indicate a causal link?

This question is all the more relevant because one main reason to estimate multiple regressions is to get closer to a causal interpretation. By conditioning on other observable variables, we can get closer to comparing similar objects – "apples to apples" – even in observational data. But getting closer is not the same as getting there.

For example, estimating the effect of a training program at a firm on the performance of employees would require comparing participants to non-participants who would perform similarly to how participants would without the program. A randomized experiment ensures such comparability. By randomly deciding who participates and who does not participate, we get two groups that are very similar in everything that is relevant, including what their future performance would be without the program. If, instead of a random rule, employees decided for themselves whether they participate in the program, a simple comparison of participants to non-participants would not measure the effect of the program because participants may have achieved different performance without the training.

The difference is between data from controlled experiments and observational data . Simple comparisons don't uncover causal relations in observational data. In principle, we may improve this by conditioning on every potential confounder: variables that would affect y and the causal variable x_1 at the same time. (In the training example, these are variables that would make participants and non-participants achieve different performance without the training, such as skills and motivation.) Such a comparison is called **ceteris paribus**.

But, importantly, conditioning on everything is impossible in general. Ceteris paribus prescribes what we want to condition on; a multiple regression can condition on what's in the data the way it is measured.

One more caveat. Not all variables should be included as covariates even if correlated both with the causal variable and the dependent variable. Such variables are called **bad conditioning variables**, or bad control variables. Examples include variables that are actually part of the **causal mechanism**, for example, how much participants in the training program actually learn.

What variables to include in a multiple regression and what variables not to include when aiming to estimate the effect of x on y is a difficult question. Chapter 19 will discuss this question along with the more general question of whether and when conditioning on other variables can lead to a good estimate of the effect of x on y, and what we mean by such an effect in the first place.

10.A6 CASE STUDY – Understanding the Gender Difference in Earnings

Thinking about cause and effect and getting closer to estimating it via multiple regression

Figure 10.2a showed a large and relatively stable average gender difference in earnings between ages 40 and 60 in the data and the population it represents (employees with a graduate degree in the USA in 2014). What might cause that difference?

One potential explanation is labor market discrimination. Labor market discrimination means that members of a group (women, minorities) earn systematically less per hour than members of another group (men, the majority) even if they have the same marginal product. Marginal product simply means their contribution to the sales of their employer by working one additional hour.

If one hour of work by women brings as much for the employer as one hour of work by men, they should earn the same, at least on average. There may be individual deviations for various reasons due to mistakes and special circumstances, but there should not be systematic differences in earnings per hour.

Note that this concept of labor market discrimination is quite narrow. For example, women may earn less on average because they are less frequently promoted to positions in which their work could have a higher effect on company sales. That would not count as labor market discrimination according to this narrow definition. A broader notion of discrimination would want to take that into account. An even broader concept of social inequality may recognize that women may choose occupations with flexible or shorter hours of work due to social norms about division of labor in the family. That may result in the over-representation of women in jobs that offer lower wages in return for more flexible hours.

Let's use our data to shed some light on these issues. Starting with the narrow definition of labor market discrimination, we have a clear steer as to what ceteris paribus analysis would be: condition on marginal product, or everything that matters for marginal product (and may possibly differ by gender). These may include cognitive skills, motivation, the ability to work efficiently in teams, and so on. Real-life data does not include all those variables. Indeed, our data has very little on skills: three broad categories of graduate degree and age. We may add race, ethnicity, and whether a person was born in the USA that may be related to the quality of education as well as other potential sources of discrimination.

To shed light on broader concepts of discrimination, we may want to enrich our regression by including more covariates. One example is occupation. Women may choose occupations that offer shorter and more flexible hours in exchange for lower wages. For the narrow concept of discrimination, we would like to condition on occupation, because we would want to compare women and men with the same work tasks. For the broad concept, we would not want to condition on it, because choice of occupation is affected by social norms about gender. Similar variables are industry, union status, hours worked, or whether the employer is private, nonprofit, or government.

Table 10.5 shows the results of those regressions. Some regressions have many explanatory variables. Instead of showing the coefficients of all, we show the coefficient and standard error of the variable of focus: *female*. The subsequent rows of the table indicate which variables are included as covariates. This is in fact a standard way of presenting results of large multiple regressions that focus on a single coefficient.

The data used for these regressions in Table 10.5 is a subset of the data used previously: it contains employees of age 40 to 60 with a graduate degree who work 20 hours per week or more. We focus on this age group because, as we have seen, this group has the largest average gender difference in earnings. We have 9816 such employees in our data.

Column (1) shows that women earn 22.4% less than men, on average, in the data (employees of age 40 to 60 with a graduate degree who work 20 hours or more). When we condition on age and the two binary variables of education, the difference is only slightly less, 21.2% (column (2)). This small difference appears to suggest that differences in age and education do not contribute to gender differences in earnings. However, our measures of education are only two binary variables of degree level, and more detailed data may imply a larger role of educational differences.

Table 10.5 Gender differences in earnings – regression with many covariates on a narrower sample

Variables	(1) ln wage	(2) ln wage	(3) ln wage	(4) ln wage
female	−0.224**	−0.212**	−0.151**	−0.141**
	(0.012)	(0.012)	(0.012)	(0.012)
Age and education		YES	YES	YES
Family circumstances			YES	YES
Demographic background			YES	YES
Job characteristics			YES	YES
Age in polynomial				YES
Hours in polynomial				YES
Observations	9816	9816	9816	9816
R-squared	0.036	0.043	0.182	0.195

Note: Education: professional, PhD. Family circumstances: marital status and number of children. Demographic background: race, ethnicity, whether US-born. Job characteristics: hours worked, whether employer is federal, state, local government, or nonprofit; union membership, two-digit industry, and two-digit occupation codes. Age and hours polynomials are fourth-order. Robust standard error estimates in parentheses. ** $p<0.01$, * $p<0.05$.

Source: `cps-earnings` dataset. USA, 2014. All employees of age 40–60 with a graduate degree and 20 or more work hours per week.

Column (3) includes all other covariates. The gender difference is 15.1%. When we compare people with the same personal and family characteristics and job features as measured in the data, women earn 15.1% less than men. Some of these variables are meant to measure job flexibility, but they are imperfect. Omitted variables include flexibility of hours and commuting time. Column (4) includes the same variables but pays attention to the potentially nonlinear relations with the two continuous variables, age and hours worked. The gender difference is very similar, 14.1%. The confidence intervals are reasonably narrow around these coefficients (±2%). They suggest that the average gender difference in the data, unconditional or conditional on the covariates, is of similar magnitude in the population represented by our data to what's in the data.

What did we learn from this exercise? We certainly could not safely pin down the role of labor market discrimination versus other reasons in driving the gender inequality in pay. Even their relative role is hard to assess from these results as the productivity measures are few, and the other covariates may be related to discrimination as well as preferences or other aspects of productivity. Thus, we cannot be sure that the 14.1% in column (4) is due to discrimination, and we can't even be sure if the role of discrimination is larger or smaller than that.

Nevertheless, our analysis provided some useful facts. The most important of them is that the gender difference is quite small below age 30, and it's the largest among employees between ages 40 and 60. Thus, gender differences, whether due to discrimination or other reasons, tend to be small among younger employees. In contrast, the disadvantages of women are large among

middle-aged employees who also tend to be the highest earning employees. This is consistent with many potential explanations, such as the difficulty of women to advance their careers relative to men due to "glass ceiling effects" (discrimination at promotion to high job ranks), or differences in preferences for job flexibility versus career advancement, which, in turn, may be due to differences in preferences or differences in the constraints the division of labor in families put on women versus men.

On the methods side, this case study illustrated how to estimate multiple linear regressions, and how to interpret and generalize their results. It showed how we can estimate and visualize different patterns of association, including nonlinear patterns, between different groups. It highlighted the difficulty of drawing causal conclusions from regression estimates using cross-sectional data. Nevertheless, it also illustrated that, even in the absence of clear causal conclusions, multiple regression analysis can advance our understanding of the sources of a difference uncovered by a simple regression.

10.12 Multiple Regression and Prediction

One frequent reason to estimate a multiple regression is to make a **prediction**: find the best guess for the dependent variable, or target variable y_j for a particular target observation j, for which we know the right-hand-side variables x but not y. Multiple regression offers a better prediction than a simple regression because it includes more x variables.

The predicted value of the dependent variable in a multiple regression for an observation j with known values for the explanatory variables $x_{1j}, x_{2j}, \ldots$ is simply

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \cdots \tag{10.21}$$

When the goal is prediction, we want the regression to produce as good a fit as possible. More precisely, we want as good a fit as possible to the general pattern that is representative of the target observation j. Good fit in a dataset is a good starting point – that is, of course, if our data is representative of that general pattern. But it's not necessarily the same. A regression with a very good fit in our data may not produce a similarly good fit in the general pattern. A common danger is **overfitting** the data: finding patterns in the data that are not true in the general pattern. Thus, when using multiple regression for prediction, we want a regression that provides good fit without overfitting the data. Finding a multiple regression means selecting right-hand-side variables and functional forms for those variables. We'll discuss this issue in more detail when we introduce the framework for prediction in Chapter 13.

But how can we assess the fit of multiple regressions? Just like with simple regressions, the most commonly used measure is the R-squared. The R-squared in a multiple regression is conceptually the same as in a simple regression that we introduced in Chapter 7:

$$R^2 = \frac{Var[\hat{y}]}{Var[y]} = 1 - \frac{Var[e]}{Var[y]} \tag{10.22}$$

where $Var[y] = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, $Var[\hat{y}] = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, and $Var[e] = \sum_{i=1}^{n}(e_i)^2$. Note that $\bar{\hat{y}} = \bar{y}$, and $\bar{e} = 0$.

The R-squared is a useful statistic to describe the fit of regressions. For that reason, it is common practice to report the R-squared in standard tables of regression results.

Unfortunately, the R-squared is an imperfect measure for selecting the best multiple regression for prediction purposes. The reason is that regressions with the highest R-squared tend to overfit the data.

When we compare two regressions, and one of them includes all the right-hand-side variables in the other one plus some more, the regression with more variables always produces a higher R-squared. Thus, regressions with more right-hand-side variables tend to produce higher R-squared. But that's not always good: regressions with more variables have a larger risk of overfitting the data. To see this, consider an extreme example. A regression with a binary indicator variable for each of the observations in the data (minus one for the reference category) produces a perfect fit with an R-squared of one. But such a regression would be completely useless to predict values outside the data. Thus, for variable selection, alternative measures are used, as we shall discuss it in Chapter 14.

Until we learn about more systematic methods to select the right-hand-side variables in the regression for prediction, all we can do is to use our intuition. The goal is to have a regression that captures patterns that are likely to be true for the general pattern for our target observations. Often, that means including variables that capture substantial differences in y, and not including variables whose coefficients imply tiny differences. That implies leaving out variables that capture detailed categories of qualitative variables or complicated interactions. To do really well, we will need the systematic tools we'll cover in Chapters 13 and 14.

The last topic in prediction is how we can visualize the fit of our regression. The purpose of such a graph is to compare values of y to the regression line. We visualized the fit of a simple regression with a scatterplot and the regression line in the x–y coordinate system. We did something similar with the age–gender interaction, too. However, with a multiple regression with more variables, we can't produce such a visualization because we have too many right-hand-side variables.

Instead, we can visualize the fit of a multiple regression by the **$\hat{y}$–y plot**. This plot has $\hat{y}$ on the horizontal axis and y on the vertical axis. The plot features the 45 degree line and the scatterplot around it. The 45 degree line is also the regression line of y regressed on $\hat{y}$. To see this consider that the regression of y on $\hat{y}$ shows the expected value of y for values of $\hat{y}$. But $\hat{y}$ is already the expected value of y conditional on the right-hand-side variables, so the expected value of y conditional on $\hat{y}$ is the same as $\hat{y}$. Therefore this line connects points where $\hat{y} = y$, so it is the 45 degree line.

The scatterplot around this line shows how actual values of y differ from their predicted value $\hat{y}$. The better the fit of the regression, the closer this scatterplot is to the 45 degree line (and the closer R-squared is to one). This visualization is more informative than the R-squared. For example, we can use the $\hat{y}$–y plot to identify observations with especially large positive or negative residuals. In this sense, it generalizes the scatterplot with a regression line when we only had a single x.

Review Box 10.8 Prediction with multiple linear regression

- The predicted value of the y variable from a multiple regression is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x1 + \hat{\beta}_2 x_2 + \cdots$$

- The $\hat{y}$–y plot is a good way to visualize the fit of a prediction. It's a scatterplot with $\hat{y}$ on the horizontal axis and y on the vertical axis, together with the 45 degree line, which is the regression line of y on $\hat{y}$.
  - observations to the right of the 45 degree line show overpredictions ($\hat{y} > y$).
  - observations to the left of the 45 degree line show underpredictions ($\hat{y} < y$).

10.B1 CASE STUDY – Finding a Good Deal among Hotels with Multiple Regression

Prediction with multiple regression

Let's return once more to our example of hotel prices and distance to the city center. Recall that the goal of the analysis is to find a good deal from among the hotels for the date contained in the data. A good deal is a hotel that is inexpensive relative to its characteristics. Of those characteristics two are especially important: the distance of the hotel to the city center and the quality of the hotel. In the earlier chapters we considered simple regressions with the distance to the city center as the only explanatory variable. Here we add measures of quality and consider a multiple regression. Those measures of quality are stars (3, 3.5, or 4) and rating (average customer rating, ranging from 2 to 5).

With prediction, capturing the functional form is often important. Based on earlier explorations of the price–distance relationship and similar explorations of the price–stars and price–ratings relationships, we arrived at the following specification. The regression has log price as the dependent variable, a piecewise linear spline in distance (knots at 1 and 4 miles), a piecewise linear spline in rating (one knot at 3.5), and binary indicators for stars (one for 3.5 stars, one for 4 stars; 3 stars is the reference category).

From a statistical point of view, this is prediction analysis. The goal is to find the best predicted (log) price that corresponds to distance, stars, and ratings of hotels. Then we focus on the difference of actual (log) price from its predicted value.

Good deals are hotels with large negative residuals from this regression. They have a (log) price that is below what's expected given their distance, stars, and rating. The more negative the residual, the lower their log price, and thus their price, compared to what's expected for them. Of course, our measures of quality are imperfect. The regression does not consider information on room size, view, details of location, or features that only photos can show. Therefore the result of this analysis should be a shortlist of hotels that the decision maker should look into in more detail.

Table 10.6 shows the five best deals: these are the hotels with the five most negative residuals. We may compare this list with the list in Chapter 7, Section 4.U1, that was based on the residuals of a simple linear regression of hotel price on distance. Only two hotels "21912" and "22080" featured on both lists; hotel "21912" is the best deal now, there it was the second best deal. The rest of the hotels from Chapter 7 did not make it to the list here. When considering stars and rating, they do not appear to be such good deals anymore because their ratings and stars are low. Instead, we have three other hotels that have good measures of quality and are not very far yet they have relatively low price. This list is a good shortlist to find the best deal after looking into specific details and photos on the price comparison website.

How good is the fit of this regression? Its R-squared is 0.55: 55 percent in the variation in log price is explained by the regression. In comparison, a regression with log price and piecewise linear spline in distance would produce an R-squared of 0.37. Including stars and ratings improved the fit by 18 percentage points.

The $\hat{y}$–y plot in Figure 10.3 visualizes the fit of this regression. The plot features the 45 degree line. Dots above the line correspond to observations with a positive residual: hotels that have

Table 10.6 Good deals for hotels: the five hotels with the most negative residuals

Hotel name	Price	Residual in ln(price)	Distance	Stars	Rating
21912	60	−0.565	1.1	4	4.1
21975	115	−0.405	0.1	4	4.3
22344	50	−0.385	3.9	3	3.9
22080	54	−0.338	1.1	3	3.2
22184	75	−0.335	0.7	3	4.1

Note: List of the five observations with the smallest (most negative) residuals from the multiple regression with log *price* on the left-hand side; right-hand-side variables are distance to the city center (piecewise linear spline with knots at 1 and 4 miles), average customer rating (piecewise linear spline with knot at 3.5), binary variables for 3.5 stars and 4 stars (reference category is 3 stars).

Source: `hotels-vienna` dataset. Vienna, November 2017, weekday. Hotels with 3 to 4 stars within 8 miles of the city center, N=217.

higher price than expected based on the right-hand-side variables. Dots below the line correspond to observations with a negative residual: hotels that have lower price than expected. The dots that are furthest down from the line are the candidates for a good deal.

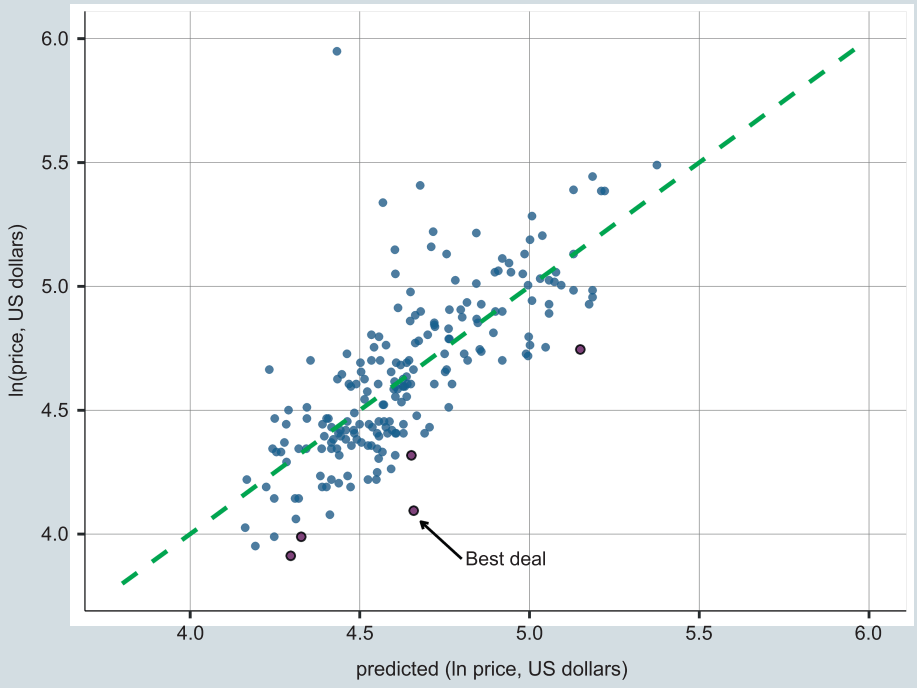


Figure 10.3 $\hat{y}$–y plot for log hotel price

Note: Results from a regression of ln price on distance to the city center (piecewise linear spline with knots at 1 and 4 miles), average customer rating (piecewise linear spline with knot at 3.5), binary variables for 3.5 stars and 4 stars (reference category is 3 stars). y is ln ***price***; $\hat{y}$ is predicted ln ***price*** from the regression. Five best deals denoted with purple.

Source: `hotels-vienna` dataset. Vienna, November 2017, weekday. Hotels with 3 to 4 stars within 8 miles of the city center; N=207.

This concludes the series of case studies using the `hotels-vienna` dataset to identify the hotels that are the best deals. We produced a shortlist of hotels that are the least expensive relative to their distance to the city center and their quality, measured by average customer ratings and stars.

10.13 Main Takeaways

Multiple regression allows for comparing mean y for different values of x for observations with the same values for the other variables.

- Doing so leads to better predictions and estimates of the slope on x that are usually closer to its true effect.
- Qualitative variables should be entered as binary variables on the right-hand side of multiple regressions.
- Interactions can uncover different slopes of one variable by values of another variable (e.g., denoting different groups).

PRACTICE QUESTIONS

1. The 95% CI of a slope coefficient in a multiple regression is narrower, the larger the R-squared of the regression. Why?
2. The 95% CI of a slope coefficient in a multiple regression is wider, the more the variable is correlated with the other right-hand-side variables. Why?
3. What's the difference between β in $y^E = \alpha + \beta x_1$ and β_1 in $y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$? Why is this difference called omitted variable bias?
4. You want to estimate differences in spending on cigarettes by family income, and you are also interested how this difference in spending varies by the level of education of the adults in the family. Write down a regression that can uncover those differences, and interpret the coefficients of that regression. (Hint: you may define education as a binary variable, high versus low.)
5. Give an example of a multiple regression with two binary right-hand-side variables and their interaction. Write down the regression and interpret its coefficients.
6. What's a $\hat{y}-y$ plot, and what is it good for? Give an example.
7. You want to predict y with the help of ten x variables using multiple linear regression. A regression that includes the ten variables produces an R-squared of 0.4. A regression that includes those ten variables together with many of their interactions has 100 variables altogether, and it produces an R-squared of 0.43. List pros and cons for choosing each of the two regressions for your prediction.
8. You want to estimate the effect of x on y with the help of a multiple regression. You have 10 z variables that you want to condition on to get closer to causality. A regression that includes the ten variables produces a coefficient on x that is 0.40, $SE = 0.05$. A regression that includes those ten variables together with many of their interactions gives a coefficient estimate of 0.35 and $SE = 0.10$. List pros and cons for choosing each of the two regressions for your causal analysis.

9. True or false? Why? Multiple regression analysis of observational data shows expected differences in the dependent variable corresponding to differences of an explanatory variable, ceteris paribus.
10. In a dataset of a cross-section of cars advertised on an online platform, log price is regressed on a binary variable that is one if the ad comes from a dealer and zero if it comes from the owner of the car. The slope coefficient is 0.1. When we add the age of the car to the regression, the slope coefficient on the same binary variable is 0.05. Interpret both numbers.
11. In the previous example the coefficient on the age of the car is −0.1. Are dealer-advertised cars older or younger on average? By how much?
12. Log hotel price is regressed on distance to the city center, average customer rating, and number of stars in a linear regression, using data on a cross-section of hotels in a city. The slope coefficient on stars is 0.2. Interpret this number.
13. The standard error on the slope coefficient in the previous question is 0.05. What's its 95% CI and what does that mean?
14. We are interested in how airline prices are related to the distance traveled and the size of the market (number of passengers). The data consists of average prices on the most popular markets (routes) of the USA (e.g., Boston–Chicago) for the year 2018. OLS estimates of our regression are the following:

$$(\ln price)^E = 4.4 + 0.3 distance - 0.06 passengers \tag{10.23}$$

where ln *price* is log price, *distance* is measured in thousand miles, and *passengers* is the number of passengers per day (thousands). Interpret the two slope coefficient estimates.
15. We include the interaction of distance and passengers in the same regression and get a coefficient estimate of −0.002. Interpret this number. What can we conclude from its inclusion about the interaction of distance and passengers if the SE is 0.002?

DATA EXERCISES

Easier and/or shorter exercises are denoted by [*]; harder and/or longer exercises are denoted by [**].

1. Re-do the case study on gender difference in earnings by age – using a different group of employees or a different year in the USA, a different educational group, or a different country. Compare your results to those in the text and try to explain what you see. [*]
2. Use the `hotels-europe` dataset and pick a different city or a different date than used in the case study. Estimate a regression to predict hotel prices (or their log) using stars of the hotel, average customer ratings, and distance to the city center. Pay attention to functional forms. Argue for the best regression specification, and use its results to create a shortlist of five hotels that are underpriced. [**]
3. Use the same data as in the previous exercise and consider adding other variables in the data in an appropriate functional form. Argue for the best regression specification. Use the results to create a shortlist of five hotels that are underpriced, and compare this list to the list you produced in the previous exercise. [**]
4. Use the `worldbank-lifeexpectancy` dataset on a cross-section of countries. Pick a year. Regress life expectancy on log GDP per capita separately for different groups of countries (e.g., by

continent). Then estimate a regression with the group dummies and their interactions with log GDP per capita. Interpret the coefficients and their 95% CI, and visualize the regression lines. What do you conclude from this exercise? [**]

5. Football is a global sport; FIFA, its governing body has 211 countries. There is plenty of evidence that countries' success in football is correlated with many socio-economic variables. Collect data on the results of all international games in a recent year, and pick a few socio-economic variables such as GDP/capita (pick three or more variables). Build a linear regression model to see which variables are correlated with the goal difference between teams. Create a $\hat{y}$–y graph to see which countries perform better or worse than expected. [**]

REFERENCES AND FURTHER READING

A great reading on the role of regressions in arguing social change is Golbeck (2017).

One example of multivariate regression used in real life is the widespread use of the hedonic price index. For more on price indexes and hedonic regression, a great resource is Eurostat's *Handbook on Residential Property Prices Indices (RPPIs)* (Eurostat, 2013).

10.U1 UNDER THE HOOD: A TWO-STEP PROCEDURE TO GET THE MULTIPLE REGRESSION COEFFICIENT

We can get β_1 in the multiple regression from a two-step procedure that involves two simple regressions. There is no practical use for this two-step procedure: we get the exact coefficient we need by running multiple regression in software. In fact, this procedure is inferior because it produces standard errors that are wrong. Nevertheless, this procedure may highlight the intuition of how multiple regression works and how we should interpret its results. Moreover, the procedure, or its underlying logic, may become useful in substantially more complicated models.

We can get coefficient $\hat{\beta}_1$ in

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{10.24}$$

by (1) regressing x_1 on x_2:

$$x_1^E = \kappa + \lambda x_2 \tag{10.25}$$

and saving the residual $e = x_1 - \hat{\kappa} - \hat{\lambda} x_2$ and then (2) regressing y on this residual:

$$y^E = \pi + \rho e \tag{10.26}$$

The estimated slope coefficients are exactly the same:

$$\hat{\beta}_1 = \hat{\rho} \tag{10.27}$$

The procedure is analogous with more right-hand-side variables, only we have to regress x_1 on all other right-hand-side variables in step (1).