Introduction to Econometrics (4th Edition)

by

James H. Stock and Mark W. Watson

Answers to End-of-Chapter "Review the Concepts" Questions

(This version September 14, 2018)

Chapter 1

1.1 The experiment that you design should have one or more treatment groups and a control group; for example, one treatment could be studying for four hours, and the control would be not studying (no treatment). Students would be randomly assigned to the treatment and control groups, and the causal effect of hours of study on midterm performance would be estimated by comparing the average midterm grades for each of the treatment groups to that of the control group. The largest impediment is to ensure that the students in the different treatment groups spend the correct number of hours studying. How can you make sure that the students in the control group do not study at all, since that might jeopardize their grade? How can you make sure that all students in the treatment group actually study for four hours?

1.2 This experiment needs the same ingredients as the experiment in the previous question: treatment and control groups, random assignment, and a procedure for analyzing the resulting experimental data. Here there are two treatment levels: not wearing a seatbelt (the control group) and wearing a seatbelt (the treated group). These treatments should be applied over a specified period of time, such as the next year. The effect of seat belt use on traffic fatalities could be estimated as the difference between fatality rates in the control and treatment group. One impediment to this study is ensuring that participants follow the treatment (do or do not wear a seat belt). More importantly, this study raises serious ethical concerns because it instructs participants to engage in known unsafe behavior (not wearing a seatbelt).

1.3

a. You will need to specify the treatment(s) and randomization method, as in Questions 1.1 and 1.2.

b. One such cross-sectional data set would consist of a number of different firms with the observations collected at the same point in time. For example, the data set might contain data on training levels and average labor productivity for 100 different firms

during 2018. Chapter 4 introduces linear regression as a way to estimate causal effects using cross-sectional data.

c. The time series data would consist of observations for a single firm collected at different points in time. For example, the data set might contain data on training levels and average labor productivity for the firm for each year between 1980 and 2018. Chapter 15 discusses how linear regression can be used to estimate causal effects using time series data.

d. Panel data would consist of observations from different firms, each observed at different points in time. For example, the data might consist of training levels and average labor productivity for 100 different firms, with data on each firm in 1995, 2005, and 2015. Chapter 10 discusses how linear regression can be used to estimate causal effects using panel data.

Chapter 2

2.1 These outcomes are random because they are not known with certainty until they actually occur. You do not know with certainty the gender of the next person you will meet, the time that it will take to commute to school, and so forth.

2.2 If X and Y are independent, then $\Pr(Y \leq y \mid X = x) = \Pr(Y \leq y)$ for all values of y and x. That is, independence means that the conditional and marginal distributions of Y are identical so that learning the value of X does not change the probability distribution of Y: Knowing the value of X says nothing about the probability that Y will take on different values.

2.3 Although there is no apparent causal link between rainfall and the number of children born, rainfall could tell you something about the number of children born. Knowing the amount of monthly rainfall tells you something about the season, and births are seasonal. Thus, knowing rainfall tells you something about the month, which tells you something about the number of children born. Thus, rainfall and the number of children born are not independently distributed.

2.4 The average weight of four randomly selected students is unlikely to be exactly 145 lbs. Different groups of four students will have different sample average weights, sometimes greater than 145 lbs. and sometimes less. Because the four students were selected at random, their sample average weight is also random.

2.5 All of the distributions will have a normal shape and will be centered at 1, the mean of Y. However they will have different spreads because they have different variances. The variance of $\overline{Y}$ is $4/n$, so the variance shrinks as n gets larger. In your plots, the spread of the normal density when $n = 2$ should be wider than when $n = 10$, which should be wider than when $n = 100$. As n gets very large, the variance approaches zero, and the normal density collapses around the mean of Y. That is, the distribution of $\overline{Y}$ becomes

highly concentrated around μ_Y as n grows large (the probability that $\bar{Y}$ is close to μ_Y tends to 1), which is just what the law of large numbers says.

2.6 The normal approximation does not look good when $n = 5$, but looks good for $n = 25$ and $n = 100$. Thus $\Pr(\bar{Y} \leq 0.1)$ is approximately equal to the value computed by the normal approximation when n is 25 or 100, but is not well approximated by the normal distribution when $n = 5$.

2.7 The probability distribution looks liked Figure 2.3b, but with more mass concentrated in the tails. Because the distribution is symmetric around $\mu_Y = 0$, $\Pr(Y > c) = \Pr(Y < c)$ and, because this is substantial mass in the tails of the distribution, $\Pr(Y > c)$ remains significantly greater than zero even for large values of c.

Chapter 3

3.1 The population mean is the average in the population. The sample average $\bar{Y}$ is the average of a sample drawn from the population.

3.2 An estimator is a procedure for computing an educated guess of the value of a population parameter, such as the population mean. An estimate is the number that the estimator produces in a given sample. $\bar{Y}$ is an example of an estimator. It gives a procedure (add up all of the values of Y in the sample and divide by n) for computing an educated guess of the value of the population mean. If a sample of size $n = 4$ produces values of Y of 100, 104, 123, and 96, then the estimate computed using the estimator $\bar{Y}$ is 105.75.

3.3 In all cases the mean of $\bar{Y}$ is 10. The variance of $\bar{Y}$ is $var(Y)/n$, which yields $var(\bar{Y}) = 1.6$ when $n = 10$, $var(\bar{Y}) = 0.16$ when $n = 100$, and $var(\bar{Y}) = 0.016$ when $n = 1000$. Since $var(\bar{Y})$ converges to zero as n increases, then, with probability approaching 1, $\bar{Y}$ will be close to 10 as n increases. This is what the law of large numbers says.

3.4 The central limit theorem plays a key role when hypotheses are tested using the sample mean. Since the sample mean is approximately normally distributed when the sample size is large, critical values for hypothesis tests and p-values for test statistics can be computed using the normal distribution. Normal critical values are also used in the construction of confidence intervals.

3.5 These are described in Section 3.2.

3.6 A confidence interval contains all values of the parameter (for example, the mean) that cannot be rejected when used as a null hypothesis. Thus, it summarizes the results from a very large number of hypothesis tests.

3.7 The treatment (or causal) effect is the difference between the mean outcomes of treatment and control groups when individuals in the *population* are randomly assigned to the two groups. The differences-in-mean estimator is the difference between the mean outcomes of treatment and control groups for a randomly selected *sample* of individuals in the population, who are then randomly assigned to the two groups.

3.8 The plot for (a) is upward sloping, and the points lie exactly on a line. The plot for (b) is downward sloping, and the points lie exactly on a line. The plot for (c) should show a positive relation, and the points should be close to, but not exactly on an upward-sloping line. The plot for (d) shows a generally negative relation between the variables, and the points are scattered around a downward-sloping line. The plot for (e) has no apparent linear relation between the variables.

Chapter 4

4.1 β_1 is the value of the slope in the population regression. This value is unknown. $\hat{\beta}_1$ (an estimator) gives a formula for estimating the unknown value of β_1 from a sample. Similarly, u_i is the value of the regression error for the i^{th} observation; u_i is the difference between Y_i and the population regression line $\beta_0 + \beta_1 X_i$. Because the values of β_0 and β_1 are unknown, the value of u_i is unknown; that is, u_i cannot be constructed from Y_i and X_i because β_0 and β_1 are unknown. In contrast, $\hat{u}_i$ is the difference between Y_i and $\hat{\beta}_0 + \hat{\beta}_1 X_i$; thus, $\hat{u}_i$ is an estimator of u_i. Finally, $E(Y \mid X) = \beta_0 + \beta_1 X$ is unknown because the values of β_0 and β_1 are unknown; an estimator for this is the OLS predicted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X$.

4.2 There are many examples. Here is one for each assumption. If the value of X is assigned in a randomized controlled experiment, then (1) is satisfied. For the class size regression, if X = class size is correlated with other factors that affect test scores, then u and X are correlated and (1) is violated. If entities (for example, workers or schools) are randomly selected from the population, then (2) is satisfied. For the class size regression, if only rural schools are included in the sample while the population of interest is all schools, then (2) is violated, If u is normally distributed, then (3) is satisfied. For the class size regression, if some test scores are misreported as 100,000 (out of a possible 1000), then large outliers are possible and (3) is violated.

4.3 *SER* is an estimate of the standard deviation of the error term in the regression. The error term summarizes the effect of factors other than X for explaining Y. If the standard deviation of the error term is large, these omitted factors have a large effect on Y. The units of *SER* are the same as the units of Y. R^2 measures the *fraction* of the variability of Y explained by X, and $1–R^2$ measures the fraction of the variability of Y explained by the factors comprising the regression's error term. If R^2 is large, most of the variability in Y is explained by X. R^2 is "unit free" and takes on values between zero and one.

4.4 The value of the R^2 indicates how dispersed the points are around the estimated regression line. When $R^2 = 0.9$, the scatter of points should lie very close to the regression line. When $R^2 = 0.5$, the points should be more dispersed about the line. The R^2 does not indicate whether the line has a positive or a negative slope.

Chapter 5

5.1. The *p*-value for a two-sided test of H_0: $\mu_Y = 0$ using an i.i.d. set of observations Y_i, $i = 1, \ldots, n$ can be constructed in three steps: (1) compute the sample mean and the standard error $SE(\bar{Y})$; (2) compute the *t*-statistic for this sample $t^{act} = \bar{Y}/SE(\bar{Y})$; (3) using the standard normal table, compute the *p*-value = $\Pr(|Z| > |t^{act}|) = 2\times\Phi(-|t^{act}|)$. A similar three-step procedure is used to construct the *p*-value for a two-sided test of H_0: $\beta_1 = 0$: (1) compute the OLS estimate of the regression slope and the standard error $SE(\hat{\beta}_1)$; (2) compute the *t*-statistic for this sample $t^{act} = \hat{\beta}_1/SE(\hat{\beta}_1)$; (3) using the standard normal table, compute the *p*-value = $\Pr(|Z| > |t^{act}|) = 2\times\Phi(-|t^{act}|)$.

5.2. The wage gender gap for 2015 can be estimated using the regression in Equation (5.19) (page 148) and the data summarized in the 2015 row of Table 3.1 (page 80). The dependent variable is the hourly earnings of the i^{th} person in the sample. The independent variable is a binary variable that equals 1 if the person is a male and equals 0 if the person is a female. The wage gender gap in the population is the population coefficient β_1 in the regression, which can be estimated using $\hat{\beta}_1$. The wage gender gap for the other years can be estimated in a similar fashion.

5.3 Homoskedasticity means that the variance of *u* is unrelated to the value of *X*. Heteroskedasticity means that the variance of *u* is related to the value of *X*. If the value of *X* is chosen using a randomized controlled experiment, then *u* is homoskedastic. In a regression of a worker's earnings (*Y*) on years of education (*X*), *u* would heteroskedastic if the variance of earnings is higher for college graduates than for non-college graduates. Figure 5.3 (page 151) suggests that this is indeed the case.

5.4 In this regression β_0 denotes the average values of earnings for non-college graduates (X=0) and $\beta_0 + \beta_1$ denotes the average value of earnings for college graduates (X=1).

Thus β_1 denotes the difference in average earnings between college graduates and non-college graduates. If $\beta_1 = 8.1$, then on average, college graduates earn \$8.10 more per hour than non-college graduates.

Chapter 6

6.1 It is likely that $\hat{\beta}_1$ will be biased because of omitted variables. Schools in more affluent districts are likely to spend more on all educational inputs and thus would have smaller class sizes, more books in the library, and more computers. These other inputs may lead to higher average test scores. Thus, $\hat{\beta}_1$ will be biased upward because the number of computers per student is positively correlated with omitted variables that have a positive effect on average test scores.

6.2 If X_1 increases by 3 units and X_2 is unchanged, then Y is expected to change by $3\beta_1$ units. If X_2 decreases by 5 units and X_1 is unchanged, then Y is expected to change by $-5\beta_2$ units. If X_1 increases by 3 units and X_2 decreases by 5 units, then Y is expected to change by $3\beta_1 - 5\beta_2$ units.

6.3 Because "least squares" regression makes *SSR* as small as possible and $R^2 = 1 - SSR/TSS$, R^2 will increase (in general) when an additional regressor is added to a regression, even if the additional regressor is not important for explaining Y. $\overline{R}^2$ adjusts R^2 to eliminate this bias.

6.4 The regression cannot determine the effect of a change in one of the regressors assuming no change in the other regressors, because if the value of one of the perfectly multicollinear regressors is held constant, then so is the value of the other. That is, there is no independent variation in one multicollinear regressor. Two examples of perfectly multicollinear regressors are (1) a person's weight measured in pounds and the same person's weight measured in kilograms, and (2) the fraction of students who are male and the constant term, when the data come from all-male schools.

6.5 If X_1 and X_2 are highly correlated, most of the variation in X_1 coincides with the variation in X_2. Thus there is little variation in X_1, holding X_2 constant that can be used to estimate the partial effect of X_1 on Y.

Chapter 7

7.1 The null hypothesis that $\beta_1 = 0$ can be tested using the *t*-statistic for β_1 as described in Key Concept 7.1. Similarly, the null hypothesis that $\beta_2 = 0$ can be tested using the *t*-statistic for β_2. The null hypothesis that $\beta_1 = 0$ *and* $\beta_2 = 0$ can be tested using the *F*-statistic from Section 7.2. The *F*-statistic is necessary to test a joint hypothesis because the test will be based on both $\hat{\beta}_1$ and $\hat{\beta}_2$, and this means that the testing procedure must use properties of their joint distribution.

7.2 Here is one example. Using data from several years of her econometrics class, a professor regresses students' scores on the final exam (*Y*) on their score from the midterm exam (*X*). This regression will have a high R^2, because people who do well on the midterm tend to do well on the final. However, this regression produces a biased estimate of the causal effect of midterm scores on the final. Students who do well on the midterm tend to be students who attend class regularly, study hard, and have an aptitude for the subject. The variables are correlated with the midterm score but are determinants of the final exam score, so omitting them leads to omitted variable bias.

7.3 Control variables are regressors that capture the effects of omitted variables in a regression. These variables can eliminate or attenuate omitted variable bias for the coefficient on the variable of interest. Coefficients on the control will, in general, be biased estimates of causal effects because (by design) they capture the effect of omitted variables. In Table 7.1, student-teacher ratio is the variable of interest and the other variables are control variables.

Chapter 8

8.1 The regression function will look like the quadratic regression in Figure 8.3 or the logarithmic function in Figure 8.4. The first of these is specified as the regression of Y onto X and X^2, and the second as the regression of Y onto $\ln(X)$. There are many economic relations with this shape. For example, this shape might represent the decreasing marginal productivity of labor in a production function.

8.2 Taking logarithms of both sides of the equation yields $\ln(Q) = \beta_0 + \beta_1\ln(K) + \beta_2\ln(L) + \beta_3\ln(M) + u$, where $\beta_0 = \ln(\lambda)$. The production function parameters can be estimated by regressing the logarithm of production on the logarithms of capital, labor, and raw materials.

8.3 $\overline{R}^2$ can be used to compare the fit of regressions with same dependent variable. Thus, $\overline{R}^2$ can be used to compare the fit of a log-log and log-linear regression because the dependent variable is the logarithm, say $\ln(Y)$, in both cases. It cannot be used to compare the fit of a log-log and linear-log regression because the dependent variable is the logarithm, say $\ln(Y)$, in the first, and it is the level, say Y, in the second.

8.4 Write $HiSTR = 1 - LoSTR$ and $HiEL = (1 - LoEL)$. The regression in (8.30) can then be written as

$$\begin{aligned}\widehat{TestScore} &= 664.1 - 1.9\left(1 - LoSTR\right) - 18.2\left(1 - LoEL\right) - 3.5\left(\left(1 - LoSTR\right)\times\left(1 - LoEL\right)\right)\\ &= \left(664.1 - 1.9 - 18.2 - 3.5\right) + \left(1.9 + 3.5\right)LoSTR + \left(18.2 + 3.5\right)LoEL - 3.5\left(LoSTR \times LoEL\right)\end{aligned}$$

8.5 Augmenting the equation in Question 8.2 with an interaction term yields yields $\ln(Q) = \beta_0 + \beta_1\ln(K) + \beta_2\ln(L) + \beta_3\ln(M) + \beta_4[\ln(K) \times \ln(L)] + u$. The partial effect of $\ln(L)$ on $\ln(Q)$ is now $\beta_2 + \beta_4\ln(K)$.

8.6 You want to compare the fit of your linear regression to the fit of a nonlinear regression. Your answer will depend on the nonlinear regression that you choose for the comparison. You might test your linear regression against a quadratic regression by adding X^2 to the linear regression. If the coefficient on X^2 is significantly different from zero, then you can reject the null hypothesis that the relationship is linear in favor of the alternative that it is quadratic.

Chapter 9

9.1 See Key Concept 9.1 (page 316) and the item (1) in the chapter summary.

9.2 Including an additional variable that belongs in the regression will eliminate or reduce omitted variable bias. However, including an additional variable that does not belong in the regression will, in general, reduce the precision (increase the variance) of the estimator of the other coefficients.

9.3 It is important to distinguish between measurement error in Y and measurement error in X. If Y is measured with error, then the measurement error becomes part of the regression error term, u. If the assumptions of Key Concept 6.4 (page 201) continue to hold, this will not affect the internal validity of OLS regression, although by making the variance of the regression error term larger, it increases the variance of the OLS estimator. If X is measured with error, however, this can result in correlation between the regressor and regression error, leading to inconsistency of the OLS estimator. As suggested by Equation (9.2), this inconsistency becomes more severe the larger is the measurement error [that is, the larger is σ_w^2 in Equation (9.2)].

9.4 Schools with higher-achieving students could be more likely to volunteer to take the test, so that the schools volunteering to take the test are not representative of the population of schools, and sample selection bias will result. For example, if all schools with a low student–teacher ratio take the test, but only the best-performing schools with a high student–teacher ratio do, the estimated class size effect will be biased.

9.5 Cities with high crime rates may decide that they need more police protection and spend more on police, but if police do their job then more police spending reduces crime. Thus, there are causal links from crime rates to police spending and from police spending to crime rates, leading to simultaneous causality bias.

9.6 If the regression has homoskedastic errors, then the homoskedastic and heteroskedastic standard errors generally are similar, because both are consistent. However, if the errors are heteroskedastic, then the homoskedastic standard errors are inconsistent, while the heteroskedastic standard errors are consistent. Thus, different values for the two standard errors constitutes evidence of heteroskedasticity, and this suggests that the heteroskedastic standard errors should be used.

Chapter 10

10.1 Panel data (also called longitudinal data) refers to data for n different entities observed at T different time periods. One of the subscripts, i, identifies the entity, and the other subscript, t, identifies the time period.

10.2 A person's ability or motivation might affect both education and earnings. More able individuals tend to complete more years of schooling, and, for a given level of education, they tend to have higher earnings. The same is true for highly motivated people. The state of the macroeconomy is a time-specific variable that affects both earnings and education. During recessions, unemployment is high, earnings are low, and enrollment in colleges increases. Person-specific and time-specific fixed effects can be included in the regression to control for person-specific and time-specific variables. In this case, the effect of education on earnings is estimated using the variation in earnings for individuals whose education changed during the 10-year sample.

10.3 When person-specific fixed effects are included in a regression, they capture all features of the individual that do not vary over the sample period. Since sex does not vary over the sample period, its effect on earnings cannot be determined separately from the person-specific fixed effect. Similarly, time fixed effects capture all features of the time period that do not vary across individuals. The national unemployment rate is the same for all individuals in the sample at a given point in time, and thus its effect on earnings cannot be determined separately from the time-specific fixed effect.

10.4 There are several factors that will lead to serial correlation. For example, the economic conditions in a particular individual's city or industry might be different from the economy-wide average that is captured by the regression's time fixed effect. If these conditions vary slowly over time, they will lead to serial correlation in the error term. As another example, suppose that in 2009 the individual is lucky and finds a particularly

high-paying job that she keeps through 2014. Other things equal, this will lead to negative values of u_{it} before 2009 (when the individual's earning are lower than her average earnings over 2008-2017), and positive values in 2014 and later (when the individual's earning are higher than her average earnings over 2008-2017).

Chapter 11

11.1 Because Y is binary, its predicted value is the probability that $Y = 1$. A probability must be between 0 and 1, so the value of 1.3 is nonsensical.

11.2 The results in column (1) are for the linear probability model. The coefficients in a linear probability model show the effect of a unit change in X on the probability that $Y = 1$. The results in columns (2) and (3) are for the logit and probit models. These coefficients are difficult to interpret. To compute the effect of a change in X on the probability that $Y = 1$ for logit and probit models, use the procedures outlined in Key Concept 11.2.

11.3 She should use a logit or probit model. These models are preferred to the linear probability model because they constrain the regression's predicted values to be between 0 and 1. Usually, probit and logit regressions give similar results, and she should use the method that is easier to implement with her software.

11.4 OLS cannot be used because the regression function is not a linear function of the regression coefficients (the coefficients appear inside the nonlinear functions Φ or F). The maximum likelihood estimator is efficient and can handle regression functions that are nonlinear in the parameters.

Chapter 12

12.1 An increase in the regression error, u, shifts out the demand curve, leading to an increase in both price and quantity. Thus $\ln(P^{butter})$ is positively correlated with the regression error. Because of this positive correlation, the OLS estimator of β_1 is inconsistent and is likely to be larger than the true value of β_1.

12.2 The number of trees per capita in the state is exogenous because it is plausibly uncorrelated with the error in the demand function. However, it probably is also uncorrelated with $\ln(P^{cigarettes})$, so it is not relevant. A valid instrument must be exogenous and relevant, so the number of trees per capita in the state is not a valid instrument.

12.3 The number of lawyers is arguably correlated with the incarceration rate, so it is relevant (although this should be checked using the methods in Section 12.3). However, states with high crime rates (with positive regression errors) are likely to have more lawyers (criminals must be defended and prosecuted), so the number of lawyers will be positively correlated with the regression error. This means that the number of lawyers is not exogenous. A valid instrument must be exogenous and relevant, so the number of lawyers is not a valid instrument.

12.4 If the difference in distance is a valid instrument, then it must be correlated with X, which in this case is a binary variable indicating whether the patient received cardiac catheterization. Instrument relevance can be checked using the procedure outlined in Section 12.3 – that is from regressing X (the binary treatment variable) on Z (the distance variable) and any included exogenous variables W. Checking instrument exogeneity is more difficult. If there are more instruments than endogenous regressors, then joint exogeneity of the instruments can be tested using the J-test outlined in Key Concept 12.6. However, if the number of instruments is equal to the number of

endogenous regressors, then it is impossible to test for exogeneity statistically. In the McClellan, McNeil, and Newhouse study (1994) there is one endogenous regressor (treatment) and one instrument (difference in distance), so the *J*-test cannot be used. Expert judgment is required to assess the exogeneity.

Chapter 13

13.1 It would be better to assign the treatment level randomly to each parcel. The research plan outlined in the problem may be flawed because the different groups of parcels might differ systematically. For example, the first 25 parcels of land might have poorer drainage than the other parcels and this would lead to lower crop yields. The treatment assignment outlined in the problem would place these 25 parcels in the control group, thereby overestimating the effect of the fertilizer on crop yields. This problem is avoided with random assignment of treatments.

13.2 The treatment effect could be estimated as the difference in average cholesterol levels for the treated group and the untreated (control) group. Data on the weight, age, and gender of each patient could be used to improve the estimate using the differences estimator with additional regressors shown in Equation (13.2). This regression may produce a more accurate estimate because it controls for these additional factors that may affect cholesterol. If you had data on the cholesterol levels of each patient before he or she entered the experiment, then the differences-in-differences estimator could be used. This estimator controls for individual-specific determinants of cholesterol levels that are constant over the sample period, such as the person's genetic predisposition to high cholesterol.

13.3 If the students who were transferred to small classes differed systematically from the other students, then internal validity is compromised. For example, if the transferred students tended to have higher incomes and more learning opportunities outside of school, then they would tend to perform better on standardized tests. The experiment would incorrectly attribute this performance to the smaller class size. Information on original random assignment could be used as an instrument in a regression like Equation (13.3) to restore internal validity. The original random assignment is a valid instrument because it is exogenous (uncorrelated with the regression error) and is relevant (correlated with the actual assignment).

13.4 The Hawthorne effect is unlikely to be a problem in the fertilizer example, unless (for example) workers cultivated the different parcels of land more or less intensively depending on the treatment. Patients in the cholesterol study might be more diligent taking their medication than patients not in an experiment. Making the cholesterol experiment double-blind, so that neither the doctor nor the patient knows whether the patient is receiving the treatment or the placebo, would reduce experimental effects for the differences estimator. Experimental effects may be important in experiments like STAR, if the teachers feel that the experiment provides them with an opportunity to prove that small class sizes are best.

13.5 Military service may affect civilian earnings for some workers more than others. For example, a worker may learn a trade such as construction or electronics in military. This education may increase civilian earnings for high school graduates more than for college graduates. Thus, β_{1i} might be higher for non-college-graduates than college graduates. The lottery affects the probability of military service for those who have not already enlisted in the military. Low-wage workers might be more likely to enlist, so that π_{1i} may be lower for these workers than for others. These workers may also have higher average values of β_{1i}. TSLS will therefore estimate the effect on earnings for the subset of the population who are less likely to have enlisted (have a large value of π_{1i}) and may benefit less from military experience (have a small value of β_{1i}).

Chapter 14

14.1 The regression is not useful for determining the causal effects of reduced-price lunches on test scores. The reduced-price lunch variable is correlated with other factors (income and related learning activities outside of school, English language proficiency, etc.) that influence test scores. This means that the "causal" regression suffers from omitted variable bias. On the other hand, because the R^2 is high, the regression is useful for predicting test scores.

14.2 Cross-validation simulates out-of-sample observations by randomly dividing the in-sample observations into a subset used for estimation and another independent ('psuedo-out-of-sample) subset used to estimated the MSPE.

14.3 Mean squared error (MSE) and mean squared prediction error (MSPE) are the sum of two components: squared bias plus variance. Shrinkage estimators increase the first component (squared bias), but reduce the second component (variance). A reduction in variance that is larger than the increase in squared bias reduces MSE (or MSPE).

14.4 Both Lasso and Ridge penalize the SSR using a term that is increasing in the size of the regression coefficients. Lasso uses the sum of the absolute values of the coefficients, $\sum_{i=1}^{k}|b_j|$, to measure size, while Ridge uses the sum of the squares of the coefficients, $\sum_{i=1}^{k}b_j^2$.

14.5 A flat scree plot says that each principal component explains the same fraction of the sample variability in the *X*s. This will occur when the *Xs* are mutually uncorrelated. In this case, principal components will not simplify the predictive regression, because they will be the same as the *X*s.

Chapter 15

15.1 It does not appear stationary. The most striking characteristic of the series is that it has an upward trend. That is, observations at the end of the sample are systematically larger than observations at the beginning. This suggests that the mean of the series is not constant, which would imply that it is not stationary. The first difference of the series may look stationary, because first differencing eliminates the large trend. However, the level of the first difference series is the slope of the plot in Figure 15.2c. Looking at the figure, the slope is steeper in 1960–1975 than in 1976–1999, which in turn is steeper than in 2000-2017. Thus, it appears that there was a change in the mean of the first difference series. If there was a change in the population mean of the first difference series, then it too is nonstationary.

15.2 One way to do this is to construct pseudo out-of-sample forecasts for the random walk model and the financial analyst's model. If the analyst's model is better, then it should have a lower RMSFE in the pseudo out-of-sample period. Even if the analyst's model outperformed the random walk in the pseudo out-of-sample period, you might still be wary of his claim. If he had access to the pseudo out-of-sample data, then his model may have been constructed to fit these data very well, so it still could produce poor true out-of-sample forecasts. Thus, a better test of the analyst's claim is to use his model and the random walk to forecast future stock returns and compare true out-of-sample performance.

15.3 Yes. The usual 95% confidence interval is $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$, which in this case produces the interval 0.91–0.99. This interval does not contain 1.0. However, this method for constructing a confidence interval is based on the central limit theorem and the large-sample normal distribution of $\hat{\beta}_1$. When $\beta_1 = 1.0$, the normal approximation is not appropriate and this method for computing the confidence interval is not valid. Instead, we need to use the general method for constructing a confidence interval

outlined in Sections 3.3 and 5.2. To find out whether 1.0 is in the 95% confidence interval, we need to test the null hypothesis $\beta_1 = 1.0$ at the 5% level. If we do not reject this null, then 1.0 is in the confidence interval. The value of the *t*-statistic for this null is −2.50. From Table 15.4, the 5% critical value is −2.86, so the null hypothesis is not rejected. Thus $\beta_1 = 1.0$ is in the 95% confidence interval.

15.4 You would add a binary variable, say D_t, that equals 0 for dates up to 1992:Q1 and equals 1 for dates after 1992:Q1. If the coefficient on D_t is significantly different from zero in the regression (as judged by its *t*-statistic), then this would be evidence of an intercept break in 1992:Q1. If the date of the break is unknown, then you would need to carry out this test for many possible break dates using the QLR procedure summarized in Key Concept 15.8.

Chapter 16

16.1 As discussed in Key Concept 16.2, causal effects can be estimated by a distributed lag model when the regressors are exogenous. In this context, exogeneity means that current and lagged values of the money supply are uncorrelated with the regression error. This assumption is unlikely to be satisfied. For example, aggregate supply disturbances (oil price shocks, changes in productivity) have important effects on GDP. The Federal Reserve and the banking system also respond to these factors, thus changing the money supply. This implies that the money supply is endogenous and is correlated with the regression error (which includes these omitted variables). Because the money supply is not exogenous, the distributed lag regression model cannot be used to estimate the dynamic causal effect of money on GDP.

16.2 In a correctly specified ADL model, the error term in not serially correlated. Thus, the ADL(1,1) model is likely to me misspecified. Adding more lags will eliminate the serial correlation in the error term and produce a consistent estimator of the dynamic causal effect when X is strictly exogenous.

16.3 Cumulating the dynamic multipliers for ΔY_t yields the dynamic multipliers for Y_t. To see this note that $Y_t = (Y_t - Y_{t-1}) + (Y_{t-1} - Y_{t-2}) + \ldots + (Y_{t-k} - Y_{t-k-1}) + Y_{t-k-1} = \Delta Y_t + \Delta Y_{t-1} + \ldots + \Delta Y_{t-k} + Y_{t-k-1}$.

16.4 The regression function that includes FDD_{t+1} can be written as $E(\%ChgP_t \mid FDD_{t+1}, FDD_t, FDD_{t-1}, \ldots) = \beta_0 + \beta_1 FDD_t + \beta_2 FDD_{t-1} + \beta_3 FDD_{t-2} + \ldots + \beta_7 FDD_{t-6} + E(u_t \mid FDD_{t+1}, FDD_t, FDD_{t-1}, \ldots)$. When FDD is strictly exogenous, then $E(u_t \mid FDD_{t+1}, FDD_t, FDD_{t-1}, \ldots) = 0$, so that FDD_{t+1} does not enter the regression. When FDD_t is exogenous, but not strictly exogenous, then it may be the case that $E(u_t \mid FDD_{t+1}, FDD_t, FDD_{t-1}, \ldots) \neq 0$, so that FDD_{t+1} will enter the regression.

Chapter 17

17.1 The macroeconomist wants to construct forecasts for nine variables. If four lags of each variable are used in a VAR, then each VAR equation will include 37 regression coefficients (the constant term and four coefficients for each of the nine variables). The sample period includes 4×48 = 192 quarterly observations. When 37 coefficients are estimated using 192 observations, the estimated coefficients are likely to be imprecise, leading to inaccurate forecasts. One alternative is to use a univariate autoregression for each variable. The advantage of this approach is that relatively few parameters need to be estimated, so that the coefficients will be precisely estimated by OLS. The disadvantage is that the forecasts are constructed using only lags of the variable being forecast, and lags of the other variables might contain additional useful forecasting information. A compromise is to use a set of time series regressions with additional predictors. For example, a GDP forecasting regression might be specified using lags of GDP, consumption, and long-term interest rates, but excluding the other variables. The short-term interest rate forecasting regression might be specified using lags of short-term rates, long-term rates, GDP, and inflation. The idea is to include the most important predictors in each of the regression equations, but leave out the variables that are not very important.

17.2 The forecast of Y_{t+2} is $Y_{t+2|t} = 0.7^2 \times 5 = 2.45$. The forecast of Y_{t+30} is $Y_{t+30|t} = 0.7^{30} \times 5 = 0.0001$. The result is reasonable. Because the process is moderately serially correlated ($\beta_1 = 0.7$), Y_{t+30} is only weakly related to Y_t. This means that the forecast of Y_{t+30} should be very close to μ_Y, the mean of Y. Since the process is stationary and $\beta_0 = 0$, $\mu_Y = 0$. Thus, as expected, $Y_{t+30|t}$ is very close to zero.

17.3 If Y and C are cointegrated, then the error correction term $Y - C$ is stationary. A plot of the series $Y - C$ should appear stationary. Cointegration can be tested by carrying out a

Dickey-Fuller or DF-GLS unit root test for the series $Y - C$. This is an example of a test for cointegration with a known cointegrating coefficient.

17.4 When u_{t-1}^2 is unusually large, then σ_t^2 will be large. Since σ_t^2 is the conditional variance of u_t, then u_t^2 is likely to be large. This will lead a large value of σ_{t+1}^2 and so forth.

17.5 OLS estimates of 110 regression coefficients with a sample size of $T = 150$ are likely to be very imprecise. (There are less than 2 observations per estimated coefficient). In the the dynamic model, the 110 predictors are replaced by a handful (say $r = 5$) principal components. OLS estimates with $r = 5$ regressors and $T = 150$ are much more precise, and therefore yield more accurate forecasts.

Chapter 18

18.1 If Assumption 4 in Key Concept 18.1 is true, in large samples a 95% confidence interval constructed using the heteroskedastic-robust standard error will contain the true value of β_1 with a probability of 95%. If assumption 4 in Key Concept 18.1 is false, the homoskedasticity-only variance estimator is inconsistent. Thus, in general, in large samples a 95% confidence interval constructed using the homoskedasticity-only standard error will not contain the true value of β_1 with a probability of 95% if the errors are heteroskedastic, so the confidence interval will not be valid asymptotically.

18.2 From Slutsky's theorem, A_nB_n has an asymptotic $N(0,9)$ distribution. Thus, $\Pr(A_nB_n < 2)$ is approximately equal to $\Pr(Z < (2/3))$, where Z is a standard normal random variable. Evaluating this probability yields $\Pr(Z < (2/3)) = 0.75$.

18.3 For values of $X_i \leq 10$, the points should lie very close to the regression line because the variance of u_i is small. When $X_i > 10$, the points should be much farther from the regression line because the variance of u_i is large. Since the points with $X_i \leq 10$ are much closer to the regression line, WLS gives them more weight.

18.4 The Gauss-Markov theorem implies that the averaged estimator cannot be better than WLS. To see this, note that the averaged estimator is a linear function of $Y_1, \ldots, Y_n$ (the OLS estimators are linear functions, as is their average) and is unbiased (the OLS estimators are unbiased, as is their average). The Gauss Markov theorem implies the WLS is the best linear conditionally unbiased estimator. Thus, the averaged estimator cannot be better than WLS.

Chapter 19

19.1 Each entry of the first column of X is 1. The entries in the second and third columns are zeros and ones. The first column of the matrix X is the sum of the second and third columns; thus the columns are linearly dependent, and X does not have full column rank. The regression can be respecified by eliminating either X_{1i} or X_{2i} .

19.2

a. Estimate the regression coefficients by OLS and compute heteroskedasticity-robust standard errors. Construct the confidence interval as $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$.

b. Estimate the regression coefficients by OLS and compute heteroskedasticity-robust standard errors. Construct the confidence interval as $\hat{\beta}_1 \pm 1.96\text{SE}(\hat{\beta}_1)$.

Alternatively, compute the homoskedasticity-only standard error $\widetilde{\text{SE}}(\hat{\beta}_1)$ and form the confidence interval as $\hat{\beta}_1 \pm 1.96\widetilde{\text{SE}}(\hat{\beta}_1)$.

c. The confidence intervals could be constructed as in (b). These use the large-sample normal approximation. Under assumptions 1–6, the exact distribution can be used to form the confidence interval: $\hat{\beta}_1 \pm t_{n-k-1,0.975}\widetilde{\text{SE}}(\hat{\beta}_1)$, where $t_{n-k-1,0.975}$ is the 97.5^{th} percentile of the t distribution with $n-k-1$ degrees of freedom. Here $n = 500$ and $k = 1$. An extended version of Appendix Table 2 shows $t_{498,0.975} = 1.9648$.

19.3 No, this result requires normally distributed errors.

19.4 The BLUE estimator is the GLS estimator. You must know $\boldsymbol{\Omega}$ to compute the exact GLS estimator. However, if $\boldsymbol{\Omega}$ is a known function of some parameters that in turn can be consistently estimated, then estimators for these parameters can be used to construct an estimator of the covariance matrix $\boldsymbol{\Omega}$. This estimator can then be used to construct a

feasible version of the GLS estimator. This estimator is approximately equal to the BLUE estimator when the sample size is large.

19.5 There are many examples. Here is one. Suppose that $X_i = Y_{i-1}$ and u_i is i.i.d. with mean 0 and variance σ^2. [That is, the regression model is an AR(1) model from Chapter 15.] In this case X_i depends on u_j for $j < i$ but does not depend on u_j for $j \geq i$. This implies $E(u_i \mid X_i) = 0$. However, $E(u_{i-1} \mid X_i) \neq 0$, and this implies $E(\boldsymbol{U} \mid \boldsymbol{X}) \neq \mathbf{0}_n$.