Sample Examinations

Each of the following sample exams follows the same format. Question 1 consists of four identifications from the chapters and one short application of the material. Question 2 asks the student "did you read the book?" about one of the most important topics covered. Question 3 asks the student to apply the materials of the chapters in a manner that parallels work already done, and Question 4 attempts to see if the student can apply skills learned in the chapters to a regression result that he or she has not seen before. Questions from different sample exams can thus be combined to form tests that cover more chapters or that are harder or easier than these are. Answers to the questions (typically just textbook page references) follow each sample exam.

By the way, if your students seem especially nervous before a particular exam, you might relax them a bit with the following actual quote from a mayonnaise jar: "Keep cool but don't freeze."

Sample Exam for Chapters 1–3

1. Briefly identify the following in words or equations as appropriate:
   (a) Degrees of freedom
   (b) Estimated regression equation
   (c) The Six Steps in Applied Regression Analysis
   (d) Ordinary Least Squares
   (e) The meaning of β_1 in:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

2. Two of the most important econometric concepts to date have been the stochastic error term and the residual. Carefully distinguish between these two concepts, being sure to:
   (a) Define both terms.
   (b) State how they are similar.
   (c) State how they are different.
   (d) Give an example of an equation that contains a stochastic error term.
   (e) Give an example of an equation that contains a residual.

3. A source of constant discussion among applied econometricians is the degree to which measures of the overall fit of an estimated equation also measure the quality of that regression. To date, we have introduced something like four different measures of overall fit, but the two most used are R^2 and $\bar{R}^2$.
   (a) Carefully distinguish between R^2 and $\bar{R}^2$.
   (b) Of the two, which do you recommend typically using? Why?
   (c) What drawbacks are there to the use of the measure you chose (as your answer to part (b) above) as the primary determinant of the overall quality of a regression?

4. Consider the following estimated equation:

$$\hat{C}_t = 18.5 - 0.07P_t + 0.93YD_t - 0.74D_{1t} - 1.3D_{2t} - 1.3D_{3t}$$

where: C_t = per-capita pounds of pork consumed in the United States in quarter t
P_t = the price of a hundred pounds of pork (in dollars) in quarter t
YD_t = per capita disposable income (in dollars) in quarter t
D_{1t} = dummy equal to 1 in the first quarter (Jan.–Mar.) of the year and 0 otherwise
D_{2t} = dummy equal to 1 in the second quarter of the year and 0 otherwise
D_{3t} = dummy equal to 1 in the third quarter of the year and 0 otherwise

(a) What is the meaning of the estimated coefficient of YD?
(b) Specify expected signs for each of the coefficients. Explain your reasoning.
(c) Suppose we changed the definition of D_{3t} so that it was equal to 1 in the fourth quarter and 0 otherwise and re-estimated the equation with all the other variables unchanged. Which of the estimated coefficients would change? Would your answer to part (b) above change? Explain your answer.

Answers:

1. (a) See page 54.
(b) See Section 1.3.
(c) See Section 3.1.
(d) See Section 2.1.
(e) See page 40.

2. See Sections 1.2 and 1.3.

3. See Section 2.4.

4. We'd expect pork consumption to be the highest in the fourth quarter due to holidays, so the expected signs of the dummy coefficients are negative. For part (c), the coefficients of P and YD shouldn't change, but the others should, because the omitted condition is now the third quarter and not the fourth quarter. Thus the expected signs of the coefficients of D_1 and D_2 are no longer negative. (Some of the best students will note that the estimate of the coefficient of D_2 will be almost exactly zero.)

Sample Exam for Chapters 4–5

Note: To do this test, students will need to consult a *t*-table like Statistical Table B-1 in the text. Since the table on page 585 is accompanied by a description of how to use it, we have repeated Table B-1 without the description on the inside front cover of this text. This will allow students to use a *t*-table (as long as no notes are written on it) without seeing the additional material if you so desire.

1. Briefly identify the following in words or equations as appropriate:
   (a) *p*-value
   (b) Sampling distribution of $\hat{\beta}$
   (c) Type I Error
   (d) Level of significance
   (e) Given $\hat{\beta} = 4.0$ and SE ($\hat{\beta}$) = 2.0, test the null hypothesis that $\beta = 0$ (versus the alternative hypothesis that it does not equal zero) with 13 degrees of freedom at the 1% level of significance.

2. In Chapter 4, we discussed seven explicit assumptions about the properties of a regression equation. These assumptions (or, more accurately, the first six of them) are usually referred to as the Classical Assumptions.
   (a) Carefully and accurately state the seven assumptions.
   (b) State what each assumption actually *means* in real-world terms (be extremely brief).

3. One of the shortest but most important sections in the book is that on the Gauss-Markov theorem.
   (a) What is the Gauss-Markov theorem?
   (b) Carefully explain what the precise properties specified by the Gauss-Markov theorem mean and why they are desireable for an equation.
   (c) What could cause the Gauss-Markov theorem to no longer hold? What should we do in such a situation?

4. Consider the following regression equation for the United States (standard errors in parentheses):

$$\hat{P}_t = 4.00 - 0.010PRP_t + 0.030PRB_t + 0.20YD_t$$
$$(0.005) \qquad (0.020) \qquad (0.04)$$
$$\bar{R}^2 = 0.98 \qquad n = 29$$

where: P_t = per capita pounds of pork consumed in time period t
PRP_t = the price of pork in time period t
PRB_t = the price of beef in time period t
YD_t = per capita disposable income in time period t

   (a) Hypothesize signs and specify the appropriate null and alternative hypotheses for the coefficients of each of these variables.
   (b) State your decision rules and then test your hypotheses on the above results using the *t*-test at a 5% level of significance.
   (c) If you could add one variable to the regression, what variable would you add? Why?

Answers:

1. (a) See page 135.
   (b) See Section 4.2.
   (c) See Section 5.1.
   (d) See Section 5.2.
   (e) t = 2.0, and the critical two-sided, 1%, 13 degree of freedom critical *t*-value is 3.012, so we can not reject the null hypothesis.

2. See Section 4.1.

3. See Section 4.3.

4.

Coefficient	β_{PRP}	β_{PRB}	β_{YD}
Hypothesized sign	−	+	+
Calculated *t*-score	−2.0	+1.5	+5.0
the critical *t*-value for a 5% one-sided test with 25 d.f. is 1.708, so	reject	do not reject	reject

For part (c), the two most important criteria are whether or not the suggested variable is a time series variable for the United States and whether or not that variable can be measured. “Tastes,” for example, are important but hard to measure.

Sample Exam for Chapter 6–7

1. Briefly identify the following in words or equations as appropriate:
   (a) The major consequence of including an irrelevant variable in a regression equation.
   (b) Four valid criteria for determining whether a given variable belongs in an equation.
   (c) The problem with sequential specification searches.
   (d) The elasticity of Y with respect to X in:

$$\ln Y = \beta_0 + \beta_1 \ln X + \varepsilon$$

   (e) The sign of the bias on the coefficient of age caused by omitting experience in an equation explaining the salaries of various workers.

2. Most of Chapter 6 is concerned with specification bias.
   (a) What exactly is specification bias?
   (b) What is known to cause specification bias?
   (c) Are unbiased estimates always better than biased estimates? Why or why not?
   (d) What’s the best way to avoid specification bias?

3. There are at least two different possible approaches to the problem of building a model of the costs of production of electric power.
   I: Model I hypothesizes that per-unit costs (C) as a function of the number of kilowatt-hours produced (Q) continually and smoothly falls as production is increased, but it falls at a decreasing rate.
   II: Model II hypothesizes that per-unit costs (C) decrease fairly steadily as production (Q) increases, but costs decrease at a much faster rate for hydroelectric plants than for other kinds of facilities.

Given this information,

(a) What functional form would you recommend for estimating Model I? Be sure to write out a specific equation.

(b) What functional form would you recommend for estimating Model II? Be sure to write out a specific equation.

(c) Would $\bar{R}^2$ be a reasonable way to compare the overall fits of the two equations? Why or why not?

Note: In the following question, you may want to change the absolute size of the coefficients, depending on the size of your school, but remember to change the size of the standard errors proportionally.

4. On your way to the cashier's office (to pay yet another dorm damage bill), you overhear U. R. Accepted (the Dean of Admissions) having a violent argument with I. M. Smart (the director of the Computer Center) about an equation that Smart built to understand the number of applications that the school received from high school seniors. In need of an outside opinion, they turn to you to help them evaluate the following regression results (standard errors in parentheses):

$$\hat{N}_t = 150 + 180A_t + 1.50 \ln T_t + 30.0P_t$$
$$(90) \qquad (1.50) \qquad (60.0)$$
$$\bar{R}^2 = 0.50 \qquad N = 22 \qquad \text{(annual)}$$

where: N_t = the number of high school seniors who apply for admission in year t

A_t = the number of people on the admission staff who visit high schools full time spreading information about the school in year t

T_t = dollars of tuition in year t

P_t = the percent of the faculty in year t that had PhDs in year t

How would you respond if they asked you to:

(a) Discuss the expected signs of the coefficients.

(b) Compare these expectations with the estimated coefficients by using the appropriate tests.

(c) Evaluate the possible econometric problems that could have caused any observed differences between the estimated coefficients and what you expected.

(d) Determine whether the semilog function for T makes theoretical sense.

(e) Make any suggestions you feel are appropriate for another run of the equation.

Answers:

1. (a) See Section 6.2.

(b) See Section 6.2.

(c) See Section 6.4.

(d) Constant = β_1.

(e) Positive, since the expected sign of the coefficient of experience is positive and since the simple correlation coefficient between age and experience is positive. See Section 6.1.3.

2. See Section 6.1. Note in particular that biased estimates can at times be closer to the true β than unbiased ones and that the solution to omitted variable bias *is not* to simply include every justifiable variable.

3. (a) A number of forms are possible, but a reciprocal form would be perhaps the most appropriate:

$$C_t = \beta_0 + \beta_1/Q_t + \varepsilon_t$$

(b) Such a hypothesis calls for the use of a slope dummy defined (for instance) as D = 1 if the plant is hydroelectric and 0 otherwise. The resulting equation would be:

$$C_t = \beta_0 + \beta_1 Q_t + \beta_2 D_t + \beta_3 D_t Q_t + \varepsilon_t$$

(c) $\bar{R}^2$ is perfectly appropriate for comparing the overall fits of the two equations because the number of independent variables changes but the functional form of the dependent variable does not.

4. (a)/(b)

Coefficient	β_1	β_2	β_3
Expected sign	+	−	+
t-score	+2.0	+1.0	+0.5
Decision	reject	do not reject	do not reject

(given the 5% level and a resulting critical *t*-score of 1.734)

(c) $\hat{\beta}_2$ has an insignificant unexpected sign, and $\hat{\beta}_3$ is not significant, so an irrelevant variable and an omitted variable (s) are both possible.

(d) Many would expect that the impact of an extra dollar of tuition (on applications) decreases as tuition increases. If this is so, the semilog functional form makes sense; if not, then we should use the linear form as a default.

(e) Percent PhDs appears to be intended as a proxy for quality of the school. If we could find a better proxy for quality, then it would make sense to substitute it for X_3, and that would be a possible answer. If the unexpected sign for lnT is being caused by an omitted variable, that omitted variable must be causing positive bias. Thus for a suggested omitted variable to be a good suggestion, it would have to be likely to have caused positive bias. Therefore it must be either positively correlated with tuition and have a positive expected coefficient (like the number of high school seniors in the country or in the region) or else have a negative correlation with tuition and have a negative expected coefficient (like the number of inadequate facilities, etc.).

Sample Exam for Chapters 8–10

Note: This exam uses heteroskedasticiy as the topic of the "major" question, Question 2, so the "IDs" in Question 1 avoid heteroskedasticity as a topic. Question 2 can easily be re-worded to cover serial correlation or multicollinearity, but then Question 1 should also be changed to cover heteroskedasticity.

1. Briefly identify the following in words or equations as appropriate:

(a) Impure serial correlation
(b) Dominant variable
(c) Variance inflation factor
(d) Generalized least squares
(e) Given a calculated Durbin-Watson d statistic of 2.58, a d_L of 1.21, and a d_U of 1.55, what would you conclude?

2. Carefully *outline* (be brief!) a description of the problem typically referred to as pure heteroskedasticity.
   (a) What is it?
   (b) What are its consequences?
   (c) How do you diagnose it?
   (d) What do you do to get rid of it?

3. A model of the number of cars sold in the United States from 1980 through 2004 produced the following results (standard errors in parentheses):

$$\hat{C}_t = 3738 - 48.0P_t + 10.0Y_t + 6.0A_t - 360.0R_t$$
$$(12.0) \quad (2.0) \quad (2.0) \quad (120.0)$$
$$\bar{R}^2 = 0.85 \qquad DW = 1.86 \qquad N = 25 \text{ (annual)}$$

where: C_t = thousands of cars sold in year t
P_t = price index for domestic cars in year t
Y_t = disposable income (billions of dollars) in year t
A_t = billions of dollars of auto industry advertising expenditures in year t
R_t = the interest rate in year t

   (a) Hypothesize the expected signs of the coefficients and test the appropriate null hypotheses at the 1% level.
   (b) What econometric problems appear to be present in this equation? Why?
   (c) Suppose you were now told that the simple correlation coefficients between P, A, and Y were all between 0.88 and 0.94 and that a Park test with Y as Z produced a *t*-score of 0.50. Would your answer to part (b) above change? Why or why not? How would it change?
   (d) What suggestions would you have for another run of this regression?

4. In a study of the long-run and short-run demands for money, Chow estimated the following demand equation (standard errors in parentheses) for the United States from 1947:1 through 1965:4:

$$\hat{M}_t = 0.14 + 1.05\,Y_t^* - 0.01Y_t - 0.75R_t$$
$$(0.15) \quad (0.10) \quad (0.05)$$
$$\bar{R}^2 = 0.996 \qquad DW = 0.88 \qquad N = 76 \text{ (quarterly)}$$

where: M_t = the log of the money stock in quarter t
Y_t^* = the log of permanent income (a moving average of previous quarters' current income) in quarter t
Y_t = the log of current income in quarter t
r_T = the log of the rate of interest in quarter t

   (a) Hypothesize signs and test the appropriate null hypotheses at the 5% level of significance.
   (b) What econometric problems seem likely to be in this equation?
   (c) In particular, are there are any problems related to the coefficient of Y? If so, are these problems more likely to have been caused by multicollinearity, serial correlation, or heteroskedasticity?
   (d) What suggestions would you have for another estimation of this equation? Why?

Answers:

1. (a) See Section 9.1.
   (b) See Section 8.1.
   (c) See Section 8.3.
   (d) See Section 9.4.
   (e) The answer depends on whether you encourage your students to use one-sided or two-sided tests. For a one-sided test, the correct answer is that we can reject the null hypothesis of positive serial correlation. For a two-sided test, the correct answer is that the test is inconclusive.

2. The best answer would be only slightly more detailed than the summary on pages 361–362.

3. (a)

Coefficient	β_P	β_Y	β_A	β_R
Expected sign	−	+	+	−
Calculated *t*-score	–4.0	5.0	3.0	–3.0
Decision	reject	reject	reject	reject

   (given a 1% critical *t*-score of 2.528)

   In addition, the DW = 1.86 is higher than the d_U of 1.77, so there is no evidence of positive serial correlation.

   (b) There may be theoretically sound variables that have been omitted, for instance some measure of competition from foreign-made cars, but the results themselves give no indication of any problems.
   (c) These results are clear indications of multicollinearity but not of heteroskedasticity.
   (d) Unless a theoretically sound variable can be added measuring competition, the regression need not be changed at all.

4. (a)

Coefficient	β_{Y*}	β_Y	β_R
Expected sign	+	+	−
Calculated *t*-score	7.0	–0.1	−15.0
Decision	reject	do not reject	reject

   (given a 5% critical *t*-score of about 1.67)

   In addition, the DW = 0.88 is less than the d_L of 1.54, so we can reject the null hypothesis of no positive serial correlation.

   (b) The poor significance of $\hat{\beta}_Y$ could have been caused by multicollinearity, an omitted variable, or by an irrelevant variable (Y) and in addition, as mentioned above, we have serial correlation.
   (c) Of the three, only multicollinearity between Y^* and Y could have caused the problem, since serial correlation and heteroskedasticty would cause the *t*-scores to appear higher than they actually should be.
   (d) This is not an easy question, but we would not suggest dropping Y, nor would we consider GLS right now. It turns out that making the equation a distributed lag by adding M_{t-1} to the right-hand side switches the insignificance to Y* (and raises the DW, naturally), giving another indication of multicollinearity. The best action right now would seem to be to review the theory in search of an omitted variable and failing to find one, do nothing. This problem was adapted from Edward Tower, *Econometrics Exams, Puzzles and Problems* (Durham, NC: Eno River Press, 1985) p. 167.

Sample Exam for Chapters 12–16

1. Briefly identify the following in words or equations as appropriate:
   (a) Problems with ad hoc distributed lags
   (b) Unconditional forecasting
   (c) Moving-average process
   (d) How to test for serial correlation in a dynamic model
   (e) Difference-in-differences estimator

2. Virtually all of Chapter 14 is spent discussing the violation of the assumption that the error term is independent of the explanatory variables.
   (a) Under what circumstances might that assumption be violated?
   (b) What would the violation of that assumption be likely to cause?
   (c) What general technique is used to rid the equation of this problem? Specifically, how does it work?

3. Suppose you've been hired by your school's admissions department to help them decide whether to change admissions procedures. You are given the files of all of the students in the last graduating class (including those students who didn't graduate) and told to build a model to explain why some admitted students graduate and others don't.
   (a) Specify the functional form you would use in building such a model and carefully explain why that form is appropriate.
   (b) Specify the independent variables you would include in the equation and briefly explain how they apply to the dependent variable in question.
   (c) Carefully explain the meaning of the coefficient of your first independent variable.

4. You have been hired to forecast GDP (Y) for the Caribbean island of Tabasco. Tabasco has domestic food (F) and shelter (S) industries, a tourist (T) industry, and an export (X) industry. All tourists come from the United States, while the exports are split between Mexico and the United States. Investment is virtually zero, and government expenditures (G) can be considered to be exogenously determined. Imports (I) are a function of GDP. Thus the structural equations for a simultaneous model of the Tabascan economy would look something like:

$$Y = F + S + T + X + G - I$$
$$F = f_F(Y, ?)$$
$$S = f_S(Y, ?)$$
$$T = f_T(USGNP, ?)$$
$$X = f_X(USGNP, MEXICOGNP, ?)$$
$$I = f_I(Y, ?)$$
$$G = \text{exogenous}$$

   (a) Develop a theory for Tabasco's economy. Then choose which predetermined variables you would like to add to the simultaneous system and specify to which of the five stochastic structural equations (see question marks) you would like to add them. Explain your reasoning.
   (b) Comment on the identification properties of each of the five stochastic equations in the system you outlined in your answer to part (a) above.
   (c) How should the coefficients of the system be estimated?
   (d) What technique would you use to forecast Tabasco's GNP? Why?

Answers:

1. (a) See Section 12.1.
   (b) See Section 15.2.
   (c) See Section 15.3.
   (d) See Section 12.2.
   (e) See Section 16.2.

2. See Sections 14.2 and 14.3. This answer will also depend on whether or not Chapter 12 and/or Section 14.6 have been assigned.

3. The appropriate functional form is the logit because of the problems with the linear probability model outlined in Section 13.1. The key to choosing independent variables is the type of variable suggested; some students will misunderstand the disaggregate nature of the variables required by such a study and will suggest variables that are constant for all observations in the dataset. Each coefficient tells the impact of a one-unit change in the independent variable in question (holding constant all the other independent variables in the equation) on the log of the odds that the person graduated. his or her

4. OK, OK, we know this will be hard to grade, since each answer will be different depending on the exact variables and equations added, but this question tends to work well. The student will be forced to apply the identification, estimation, and forecasting techniques to a system of his or her own choosing in much the same way he or she will have to in his or her work later on.

   The key to the questions on estimation and forecasting have to do with the size and importance of the feedback loops (as compared to exogenous factors in determining GDP). In this case, there is a good chance that the most accurate forecast of Tabasco's GDP would be based on a "simplified reduced-form" equation that included only USGNP and MEXICOGNP as explanatory variables.