Jointly Insignificant
Jointly Statistically Significant
Minimum Variance Unbiased Estimators
Multiple Hypotheses Test
Multiple Restrictions
Normality Assumption
Null Hypothesis
Numerator Degrees of Freedom
One-Sided Alternative
One-Tailed Test
Overall Significance of the Regression
p-Value
Practical Significance
R-squared Form of the *F* Statistic
Rejection Rule
Restricted Model
Significance Level
Statistically Insignificant
Statistically Significant
t Ratio
t Statistic
Two-Sided Alternative
Two-Tailed Test
Unrestricted Model

PROBLEMS

4.1 Consider an equation to explain salaries of CEOs in terms of annual firm sales, return on equity (*roe*, in percentage form), and return on the firm's stock (*ros*, in percentage form):

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 ros + u.$$

(i) In terms of the model parameters, state the null hypothesis that, after controlling for *sales* and *roe*, *ros* has no effect on CEO salary. State the alternative that better stock market performance increases a CEO's salary.

(ii) Using the data in CEOSAL1.RAW, the following equation was obtained by OLS:

$$\widehat{\log(salary)} = \underset{(.32)}{4.32} + \underset{(.035)}{.280}\log(sales) + \underset{(.0041)}{.0174}\, roe + \underset{(.00054)}{.00024}\, ros$$
$$n = 209,\ R^2 = .283.$$

By what percentage is *salary* predicted to increase if *ros* increases by 50 points? Does *ros* have a practically large effect on *salary*?

(iii) Test the null hypothesis that *ros* has no effect on *salary* against the alternative that *ros* has a positive effect. Carry out the test at the 10% significance level.

(iv) Would you include *ros* in a final model explaining CEO compensation in terms of firm performance? Explain.

4.2 Which of the following can cause the usual OLS t statistics to be invalid (that is, not to have t distributions under H_0)?

(i) Heteroskedasticity.

(ii) A sample correlation coefficient of .95 between two independent variables that are in the model.

(iii) Omitting an important explanatory variable.

4.3 In Example 4.7, we used data on nonunionized manufacturing firms to estimate the relationship between the scrap rate and other firm characteristics. We now look at this example more closely and use all available firms.

(i) The population model estimated in Example 4.7 can be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales) + \beta_3 \log(employ) + u.$$

Using the 43 observations available for 1987, the estimated equation is

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\, hrsemp - \underset{(.370)}{.951}\log(sales) + \underset{(.360)}{.992}\log(employ)$$
$$n = 43,\ R^2 = .310.$$

Compare this equation to that estimated using only the 29 nonunionized firms in the sample.

(ii) Show that the population model can also be written as

$$\log(scrap) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(sales/employ) + \theta_3 \log(employ) + u,$$

where $\theta_3 = \beta_2 + \beta_3$. [*Hint:* Recall that $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.] Interpret the hypothesis H_0: $\theta_3 = 0$.

(iii) When the equation from part (ii) is estimated, we obtain

$$\widehat{\log(scrap)} = \underset{(4.57)}{11.74} - \underset{(.019)}{.042}\ hrsemp - \underset{(.370)}{.951}\ \log(sales/employ) + \underset{(.205)}{.041}\ \log(employ)$$

$$n = 43,\ R^2 = .310.$$

Controlling for worker training and for the sales-to-employee ratio, do bigger firms have larger statistically significant scrap rates?

(iv) Test the hypothesis that a 1% increase in *sales/employ* is associated with a 1% drop in the scrap rate.

4.4 The following table was created using the data in CEOSAL2.RAW:

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	—	.112 (.050)	.100 (.049)
profmarg	—	−.0023 (.0022)	−.0022 (.0021)
ceoten	—	—	.0171 (.0055)
comten	—	—	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 (0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353

The variable *mktval* is market value of the firm, *profmarg* is profit as a percentage of sales, *ceoten* is years as CEO with the current company, and *comten* is total years with the company.

(i) Comment on the effect of *profmarg* on CEO salary.

(ii) Does market value have a significant effect? Explain.

(iii) Interpret the coefficients on *ceoten* and *comten*. Are these explanatory variables statistically significant?

(iv) What do you make of the fact that longer tenure with the company, holding the other factors fixed, is associated with a lower salary?

4.5 In Section 4.5, we used as an example testing the rationality of assessments of housing prices. There, we used a log-log model in *price* and *assess* [see equation (4.47)]. Here, we use a level-level formulation.

(i) In the simple regression model

$$price = \beta_0 + \beta_1 assess + u,$$

the assessment is rational if $\beta_1 = 1$ and $\beta_0 = 0$. The estimated equation is

$$\widehat{price} = \underset{(16.27)}{-14.47} + \underset{(.049)}{.976}\ assess$$
$$n = 88,\ \text{SSR} = 165{,}644.51,\ R^2 = .820.$$

First, test the hypothesis that H_0: $\beta_0 = 0$ against the two-sided alternative. Then, test H_0: $\beta_1 = 1$ against the two-sided alternative. What do you conclude?

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the SSR in the restricted model. This amounts to computing $\sum_{i=1}^{n}(price_i - assess_i)^2$, where $n = 88$, since the residuals in the restricted model are just $price_i - assess_i$. (No estimation is needed for the restricted model because both parameters are specified under H_0.) This turns out to yield SSR = 209,448.99. Carry out the F test for the joint hypothesis.

(iii) Now, test H_0: $\beta_2 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$ in the model

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u.$$

The R-squared from estimating this model using the same 88 houses is .829.

(iv) If the variance of *price* changes with *assess*, *lotsize*, *sqrft*, or *bdrms*, what can you say about the F test from part (iii)?

4.6 Regression analysis can be used to test whether the market efficiently uses information in valuing stocks. For concreteness, let *return* be the total return from holding a firm's stock over the four-year period from the end of 1990 to the end of 1994. The *efficient markets hypothesis* says that these returns should not be systematically related to information known in 1990. If firm characteristics known at the beginning of the period help to predict stock returns, then we could use this information in choosing stocks.

For 1990, let *dkr* be a firm's debt to capital ratio, let *eps* denote the earnings per share, let *netinc* denote net income, and let *salary* denote total compensation for the CEO.

(i) Using the data in RETURN.RAW, the following equation was estimated:

$$\widehat{return} = \underset{(6.89)}{-14.37} + \underset{(.201)}{.321}\ dkr + \underset{(.078)}{.043}\ eps - \underset{(.0047)}{.0051}\ netinc + \underset{(.0022)}{.0035}\ salary$$
$$n = 142,\ R^2 = .0395.$$

Test whether the explanatory variables are jointly significant at the 5% level. Is any explanatory variable individually significant?

(ii) Now, reestimate the model using the log form for *netinc* and *salary*:

$$\widehat{return} = \underset{(39.37)}{-36.30} + \underset{(.203)}{.327}\ dkr + \underset{(.080)}{.069}\ eps - \underset{(3.39)}{4.74}\ \log(netinc) + \underset{(6.31)}{7.24}\ \log(salary)$$
$$n = 142,\ R^2 = .0330.$$

Do any of your conclusions from part (i) change?

(iii) In this sample, some firms have zero debt and others have negative earnings. Should we try to use log(*dkr*) or log(*eps*) in the model to see if these improve the fit? Explain.

(iv) Overall, is the evidence for predictability of stock returns strong or weak?

4.7 Consider the estimated equation from Example 4.3, which can be used to study the effects of skipping class on college GPA:

$$\widehat{colGPA} = \underset{(.33)}{1.39} + \underset{(.094)}{.412}\ hsGPA + \underset{(.011)}{.015}\ ACT - \underset{(.026)}{.083}\ skipped$$
$$n = 141,\ R^2 = .234.$$

(i) Using the standard normal approximation, find the 95% confidence interval for β_{hsGPA}.

(ii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = .4$ against the two-sided alternative at the 5% level?

(iii) Can you reject the hypothesis H_0: $\beta_{hsGPA} = 1$ against the two-sided alternative at the 5% level?

4.8 In Problem 3.4, we estimated the equation

$$\widehat{sleep} = \underset{(112.28)}{3{,}638.25} - \underset{(.017)}{.148}\ totwrk - \underset{(5.88)}{11.13}\ educ + \underset{(1.45)}{2.20}\ age$$
$$n = 706,\ R^2 = .113,$$

where we now report standard errors along with the estimates.

(i) Is either *educ* or *age* individually significant at the 5% level against a two-sided alternative? Show your work.

(ii) Dropping *educ* and *age* from the equation gives

$$\widehat{sleep} = \underset{(38.91)}{3{,}586.38} - \underset{(.017)}{.151}\ totwrk$$
$$n = 706,\ R^2 = .103.$$

Are *educ* and *age* jointly significant in the original equation at the 5% level? Justify your answer.

(iii) Does including *educ* and *age* in the model greatly affect the estimated tradeoff between sleeping and working?

(iv) Suppose that the sleep equation contains heteroskedasticity. What does this mean about the tests computed in parts (i) and (ii)?

4.9 Are rent rates influenced by the student population in a college town? Let *rent* be the average monthly rent paid on rental units in a college town in the United States. Let *pop* denote the total city population, *avginc* the average city income, and *pctstu* the student population as a percentage of the total population. One model to test for a relationship is

$$\log(rent) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(avginc) + \beta_3 pctstu + u.$$

(i) State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.

(ii) What signs do you expect for β_1 and β_2?

(iii) The equation estimated using 1990 data from RENTAL.RAW for 64 college towns is

$$\widehat{\log(rent)} = \underset{(.844)}{.043} + \underset{(.039)}{.066}\ \log(pop) + \underset{(.081)}{.507}\ \log(avginc) + \underset{(.0017)}{.0056}\ pctstu$$
$$n = 64,\ R^2 = .458.$$

What is wrong with the statement: "A 10% increase in population is associated with about a 6.6% increase in rent"?

(iv) Test the hypothesis stated in part (i) at the 1% level.

4.10 Consider the multiple regression model with three independent variables, under the classical linear model assumptions MLR.1 through MLR.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

You would like to test the null hypothesis H_0: $\beta_1 - 3\beta_2 = 1$.

(i) Let $\hat{\beta}_1$ and $\hat{\beta}_2$ denote the OLS estimators of β_1 and β_2. Find $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ in terms of the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ and the covariance between them. What is the standard error of $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Write the *t* statistic for testing H_0: $\beta_1 - 3\beta_2 = 1$.

(iii) Define $\theta_1 = \beta_1 - 3\beta_2$ and $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Write a regression equation involving β_0, θ_1, β_2, and β_3 that allows you to directly obtain $\hat{\theta}_1$ and its standard error.

4.11 The variable *rdintens* is expenditures on research and development (R&D) as a percentage of sales. Sales are measured in millions of dollars. The variable *profmarg* is profits as a percentage of sales.

Using the data in RDCHEM.RAW for 32 firms in the chemical industry, the following equation is estimated:

$$\widehat{rdintens} = \underset{(1.369)}{.472} + \underset{(.216)}{.321} \log(sales) + \underset{(.046)}{.050}\, profmarg$$
$$n = 32,\ R^2 = .099.$$

(i) Interpret the coefficient on log(*sales*). In particular, if *sales* increases by 10%, what is the estimated percentage point change in *rdintens*? Is this an economically large effect?

(ii) Test the hypothesis that R&D intensity does not change with *sales* against the alternative that it does increase with sales. Do the test at the 5% and 10% levels.

(iii) Interpret the coefficient on *profmarg*. Is it economically large?

(iv) Does *profmarg* have a statistically significant effect on *rdintens*?

COMPUTER EXERCISES

C4.1 The following model can be used to study whether campaign expenditures affect election outcomes:

$$voteA = \beta_0 + \beta_1 \log(expendA) + \beta_2 \log(expendB) + \beta_3 prtystrA + u,$$

where *voteA* is the percentage of the vote received by Candidate A, *expendA* and *expendB* are campaign expenditures by Candidates A and B, and *prtystrA* is a measure of party strength for Candidate A (the percentage of the most recent presidential vote that went to A's party).

(i) What is the interpretation of β_1?

(ii) In terms of the parameters, state the null hypothesis that a 1% increase in A's expenditures is offset by a 1% increase in B's expenditures.

(iii) Estimate the given model using the data in VOTE1.RAW and report the results in usual form. Do A's expenditures affect the outcome? What about B's expenditures? Can you use these results to test the hypothesis in part (ii)?

(iv) Estimate a model that directly gives the t statistic for testing the hypothesis in part (ii). What do you conclude? (Use a two-sided alternative.)

C4.2 Use the data in LAWSCH85.RAW for this exercise.

(i) Using the same model as Problem 3.4, state and test the null hypothesis that the rank of law schools has no ceteris paribus effect on median starting salary.

(ii) Are features of the incoming class of students—namely, *LSAT* and *GPA*—individually or jointly significant for explaining *salary*? (Be sure to account for missing data on *LSAT* and *GPA*.)

(iii) Test whether the size of the entering class (*clsize*) or the size of the faculty (*faculty*) needs to be added to this equation; carry out a single test. (Be careful to account for missing data on *clsize* and *faculty*.)

(iv) What factors might influence the rank of the law school that are not included in the salary regression?

C4.3 Refer to Problem 3.14. Now, use the log of the housing price as the dependent variable:

$$\log(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u.$$

(i) You are interested in estimating and obtaining a confidence interval for the percentage change in *price* when a 150-square-foot bedroom is added to a house. In decimal form, this is $\theta_1 = 150\beta_1 + \beta_2$. Use the data in HPRICE1.RAW to estimate θ_1.

(ii) Write β_2 in terms of θ_1 and β_1 and plug this into the log(*price*) equation.

(iii) Use part (ii) to obtain a standard error for $\hat{\theta}_1$ and use this standard error to construct a 95% confidence interval.

C4.4 In Example 4.9, the restricted version of the model can be estimated using all 1,388 observations in the sample. Compute the R-squared from the regression of *bwght* on *cigs*, *parity*, and *faminc* using all observations. Compare this to the R-squared reported for the restricted model in Example 4.9.

C4.5 Use the data in MLB1.RAW for this exercise.

(i) Use the model estimated in equation (4.31) and drop the variable *rbisyr*. What happens to the statistical significance of *hrunsyr*? What about the size of the coefficient on *hrunsyr*?

(ii) Add the variables *runsyr* (runs per year), *fldperc* (fielding percentage), and *sbasesyr* (stolen bases per year) to the model from part (i). Which of these factors are individually significant?

(iii) In the model from part (ii), test the joint significance of *bavg*, *fldperc*, and *sbasesyr*.

C4.6 Use the data in WAGE2.RAW for this exercise.

(i) Consider the standard wage equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u.$$

State the null hypothesis that another year of general workforce experience has the same effect on log(*wage*) as another year of tenure with the current employer.

(ii) Test the null hypothesis in part (i) against a two-sided alternative, at the 5% significance level, by constructing a 95% confidence interval. What do you conclude?

C4.7 Refer to the example used in Section 4.4. You will use the data set TWOYEAR.RAW.

(i) The variable *phsrank* is the person's high school percentile. (A higher number is better. For example, 90 means you are ranked better than 90 percent of your graduating class.) Find the smallest, largest, and average *phsrank* in the sample.

(ii) Add *phsrank* to equation (4.26) and report the OLS estimates in the usual form. Is *phsrank* statistically significant? How much is 10 percentage points of high school rank worth in terms of wage?

(iii) Does adding *phsrank* to (4.26) substantively change the conclusions on the returns to two- and four-year colleges? Explain.

(iv) The data set contains a variable called *id*. Explain why if you add *id* to equation (4.17) or (4.26) you expect it to be statistically insignificant. What is the two-sided *p*-value?

C4.8 The data set 401KSUBS.RAW contains information on net financial wealth (*nettfa*), age of the survey respondent (*age*), annual family income (*inc*), family size (*fsize*), and participation in certain pension plans for people in the United States. The wealth and income variables are both recorded in thousands of dollars. For this question, use only the data for single-person households (so $fsize = 1$).

(i) How many single-person households are there in the data set?

(ii) Use OLS to estimate the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u,$$

and report the results using the usual format. Be sure to use only the single-person households in the sample. Interpret the slope coefficients. Are there any surprises in the slope estimates?

(iii) Does the intercept from the regression in part (ii) have an interesting meaning? Explain.

(iv) Find the *p*-value for the test H_0: $\beta_2 = 1$ against H_0: $\beta_2 < 1$. Do you reject H_0 at the 1% significance level?

(v) If you do a simple regression of *nettfa* on *inc*, is the estimated coefficient on *inc* much different from the estimate in part (ii)? Why or why not?

C4.9 Use the data in DISCRIM.RAW to answer this question. (See also Computer Exercise C3.8 in Chapter 3.)

(i) Use OLS to estimate the model

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + \beta_3 prppov + u,$$

and report the results in the usual form. Is $\hat{\beta}_1$ statistically different from zero at the 5% level against a two-sided alternative? What about at the 1% level?

(ii) What is the correlation between log(*income*) and *prppov*? Is each variable statistically significant in any case? Report the two-sided *p*-values.

(iii) To the regression in part (i), add the variable, log(*hseval*). Interpret its coefficient and report the two-sided *p*-value for H_0: $\beta_{\log(hseval)} = 0$.

(iv) In the regression in part (iii), what happens to the individual statistical significance of log(*income*) and *prppov*? Are these variables jointly significant? (Compute a *p*-value.) What do you make of your answers?

(v) Given the results of the previous regressions, which one would you report as most reliable in determining whether the racial makeup of a zip code influences local fast-food prices?

C4.10 Use the data in ELEM94_95 to answer this question. The findings can be compared with those in Table 4.1. The dependent variable *lavgsal* is the log of average teacher salary and *bs* is the ratio of average benefits to average salary (by school).

(i) Run the simple regression of *lavgsal* on *bs*. Is the estimated slope statistically different from zero? Is it statistically different from -1?

(ii) Add the variables *lenrol* and *lstaff* to the regression from part (i). What happens to the coefficient on *bs*? How does the situation compare with that in Table 4.1?

(iii) How come the standard error on the *bs* coefficient is smaller in part (ii) than in part (i)? (*Hint:* What happens to the error variance versus multicollinearity when *lenrol* and *lstaff* are added?)

(iv) How come the coefficient on *lstaff* is negative? Is it large in magnitude?

(v) Now add the variable *lunch* to the regression. Holding other factors fixed, are teachers being compensated for teaching students from disadvantaged backgrounds? Explain.

(vi) Overall, is the pattern of results that you find with ELEM94_95.RAW consistent with the pattern in Table 4.1?