

Key Terms

Best Linear Unbiased Estimator (BLUE)	Inclusion of an Irrelevant Variable	Population Model
Biased Toward Zero	Intercept	Residual
Ceteris Paribus	Micronumerosity	Residual Sum of Squares
Degrees of Freedom (df)	Misspecification Analysis	Sample Regression Function (SRF)
Disturbance	Multicollinearity	Slope Parameter
Downward Bias	Multiple Linear Regression (MLR) Model	Standard Deviation of $\hat{\beta}_j$
Endogenous Explanatory Variable	Multiple Regression Analysis	Standard Error of $\hat{\beta}_j$
Error Term	OLS Intercept Estimate	Standard Error of the Regression (SER)
Excluding a Relevant Variable	OLS Regression Line	Sum of Squared Residuals (SSR)
Exogenous Explanatory Variable	OLS Slope Estimate	Total Sum of Squares (SST)
Explained Sum of Squares (SSE)	Omitted Variable Bias	True Model
First Order Conditions	Ordinary Least Squares	Underspecifying the Model
Frisch-Waugh Theorem	Overspecifying the Model	Upward Bias
Gauss-Markov Assumptions	Partial Effect	Variance Inflation Factor (VIF)
Gauss-Markov Theorem	Perfect Collinearity	

Problems

- 1 Using the data in GPA2 on 4,137 college students, the following equation was estimated by OLS:

$$\widehat{colgpa} = 1.392 - .0135 hspc + .00148 sat$$

$$n = 4,137, R^2 = .273,$$

where $colgpa$ is measured on a four-point scale, $hsperc$ is the percentile in the high school graduating class (defined so that, for example, $hsperc = 5$ means the top 5% of the class), and sat is the combined math and verbal scores on the student achievement test.

- Why does it make sense for the coefficient on $hsperc$ to be negative?
 - What is the predicted college GPA when $hsperc = 20$ and $sat = 1,050$?
 - Suppose that two high school graduates, A and B, graduated in the same percentile from high school, but Student A's SAT score was 140 points higher (about one standard deviation in the sample). What is the predicted difference in college GPA for these two students? Is the difference large?
 - Holding $hsperc$ fixed, what difference in SAT scores leads to a predicted $colgpa$ difference of .50, or one-half of a grade point? Comment on your answer.
- 2 The data in WAGE2 on working men was used to estimate the following equation:

$$\widehat{educ} = 10.36 - .094 sibs + .131 meduc + .210 feduc$$

$$n = 722, R^2 = .214,$$

where $educ$ is years of schooling, $sibs$ is number of siblings, $meduc$ is mother's years of schooling, and $feduc$ is father's years of schooling.

- Does $sibs$ have the expected effect? Explain. Holding $meduc$ and $feduc$ fixed, by how much does $sibs$ have to increase to reduce predicted years of education by one year? (A noninteger answer is acceptable here.)
- Discuss the interpretation of the coefficient on $meduc$.
- Suppose that Man A has no siblings, and his mother and father each have 12 years of education. Man B has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between B and A?

- 3 The following model is a simplified version of the multiple regression model used by Biddle and Hamermesh (1990) to study the tradeoff between time spent sleeping and working and to look at other factors affecting sleep:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + u,$$

where *sleep* and *totwrk* (total work) are measured in minutes per week and *educ* and *age* are measured in years. (See also Computer Exercise C3 in Chapter 2.)

- If adults trade off sleep for work, what is the sign of β_1 ?
- What signs do you think β_2 and β_3 will have?
- Using the data in SLEEP75, the estimated equation is

$$\widehat{sleep} = 3,638.25 - .148 totwrk - 11.13 educ + 2.20 age$$

$$n = 706, R^2 = .113.$$

If someone works five more hours per week, by how many minutes is *sleep* predicted to fall? Is this a large tradeoff?

- Discuss the sign and magnitude of the estimated coefficient on *educ*.
 - Would you say *totwrk*, *educ*, and *age* explain much of the variation in *sleep*? What other factors might affect the time spent sleeping? Are these likely to be correlated with *totwrk*?
- 4 The median starting salary for new law school graduates is determined by

$$\log(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + \beta_5 rank + u,$$

where *LSAT* is the median LSAT score for the graduating class, *GPA* is the median college GPA for the class, *libvol* is the number of volumes in the law school library, *cost* is the annual cost of attending law school, and *rank* is a law school ranking (with *rank* = 1 being the best).

- Explain why we expect $\beta_5 \leq 0$.
- What signs do you expect for the other slope parameters? Justify your answers.
- Using the data in LAWSCH85, the estimated equation is

$$\widehat{\log(salary)} = 8.34 + .0047 LAST + .248 GPA + .095 \log(libvol) + .038 \log(cost) - .0033 rank$$

$$n = 136, R^2 = .842.$$

What is the predicted ceteris paribus difference in salary for schools with a median GPA different by one point? (Report your answer as a percentage.)

- Interpret the coefficient on the variable $\log(libvol)$.
 - Would you say it is better to attend a higher ranked law school? How much is a difference in ranking of 20 worth in terms of predicted starting salary?
- 5 In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168.
- In the model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u,$$

does it make sense to hold *sleep*, *work*, and *leisure* fixed, while changing *study*?

- Explain why this model violates Assumption MLR.3.
- How could you reformulate the model so that its parameters have a useful interpretation and it satisfies Assumption MLR.3?

- 6 Consider the multiple regression model containing three independent variables, under Assumptions MLR.1 through MLR.4:

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + u.$$

You are interested in estimating the sum of the parameters on x_1 and x_2 ; call this $\theta_1 = \beta_0 + \beta_1$.

- (i) Show that $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ is an unbiased estimator of θ_1 .
 - (ii) Find $\text{Var}(\hat{\theta}_1)$ in terms of $\text{Var}(\hat{\beta}_1)$, $\text{Var}(\hat{\beta}_2)$, and $\text{Corr}(\hat{\beta}_1, \hat{\beta}_2)$.
- 7 Which of the following can cause OLS estimators to be biased?
- (i) Heteroskedasticity.
 - (ii) Omitting an important variable.
 - (iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

- 8 Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$\text{avgprod} = \beta_0 + \beta_1\text{avgtrain} + \beta_2\text{avgabil} + u.$$

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that *avgtrain* and *avgabil* are negatively correlated, what is the likely bias in $\tilde{\beta}_1$ obtained from the simple regression of *avgprod* on *avgtrain*?

- 9 The following equation describes the median housing price in a community in terms of amount of pollution (*nox* for nitrous oxide) and the average number of rooms in houses in the community (*rooms*):

$$\log(\text{price}) = \beta_0 + \beta_1\log(\text{nox}) + \beta_2\text{rooms} + u.$$

- (i) What are the probable signs of β_1 and β_2 ? What is the interpretation of β_1 ? Explain.
- (ii) Why might *nox* [or more precisely, $\log(\text{nox})$] and *rooms* be negatively correlated? If this is the case, does the simple regression of $\log(\text{price})$ on $\log(\text{nox})$ produce an upward or a downward biased estimator of β_1 ?
- (iii) Using the data in HPRICE2, the following equations were estimated:

$$\widehat{\log(\text{price})} = 11.71 - 1.043 \log(\text{nox}), n = 506, R^2 = .264.$$

$$\widehat{\log(\text{price})} = 9.23 - .718 \log(\text{nox}) + .306 \text{rooms}, n = 506, R^2 = .514.$$

Is the relationship between the simple and multiple regression estimates of the elasticity of *price* with respect to *nox* what you would have predicted, given your answer in part (ii)? Does this mean that $-.718$ is definitely closer to the true elasticity than -1.043 ?

- 10 Suppose that you are interested in estimating the ceteris paribus relationship between y and x_1 . For this purpose, you can collect data on two control variables, x_2 and x_3 . (For concreteness, you might think of y as final exam score, x_1 as class attendance, x_2 as GPA up through the previous semester, and x_3 as SAT or ACT score.) Let $\tilde{\beta}_1$ be the simple regression estimate from y on x_1 and let $\hat{\beta}_1$ be the multiple regression estimate from y on x_1, x_2, x_3 .
- (i) If x_1 is highly correlated with x_2 and x_3 in the sample, and x_2 and x_3 have large partial effects on y , would you expect $\tilde{\beta}_1$ and $\hat{\beta}_1$ to be similar or very different? Explain.
 - (ii) If x_1 is almost uncorrelated with x_2 and x_3 , but x_2 and x_3 are highly correlated, will $\tilde{\beta}_1$ and $\hat{\beta}_1$ tend to be similar or very different? Explain.
 - (iii) If x_1 is highly correlated with x_2 and x_3 , and x_2 and x_3 have small partial effects on y , would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.
 - (iv) If x_1 is almost uncorrelated with x_2 and x_3 , x_2 and x_3 have large partial effects on y , and x_2 and x_3 are highly correlated, would you expect $\text{se}(\tilde{\beta}_1)$ or $\text{se}(\hat{\beta}_1)$ to be smaller? Explain.

- 11 Suppose that the population model determining y is

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

and this model satisfies Assumptions MLR.1 through MLR.4. However, we estimate the model that omits x_3 . Let $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\beta}_2$ be the OLS estimators from the regression of y on x_1 and x_2 . Show that the expected value of $\tilde{\beta}_1$ (given the values of the independent variables in the sample) is

$$E(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{i1} x_{i3}}{\sum_{i=1}^n \hat{r}_{i1}^2},$$

where the \hat{r}_{i1} are the OLS residuals from the regression of x_1 on x_2 . [Hint: The formula for $\tilde{\beta}_1$ comes from equation (3.22). Plug $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ into this equation. After some algebra, take the expectation treating x_{i3} and \hat{r}_{i1} as nonrandom.]

- 12 The following equation represents the effects of tax revenue mix on subsequent employment growth for the population of counties in the United States:

$$growth = \beta_0 + \beta_1 share_p + \beta_2 share_i + \beta_3 share_s + other\ factors,$$

where *growth* is the percentage change in employment from 1980 to 1990, *share_p* is the share of property taxes in total tax revenue, *share_i* is the share of income tax revenues, and *share_s* is the share of sales tax revenues. All of these variables are measured in 1980. The omitted share, *share_F*, includes fees and miscellaneous taxes. By definition, the four shares add up to one. Other factors would include expenditures on education, infrastructure, and so on (all measured in 1980).

- (i) Why must we omit one of the tax share variables from the equation?
 - (ii) Give a careful interpretation of β_1 .
- 13 (i) Consider the simple regression model $y = \beta_0 + \beta_1 x + u$ under the first four Gauss-Markov assumptions. For some function $g(x)$, for example $g(x) = x^2$ or $g(x) = \log(1 + x^2)$, define $z_i = g(x_i)$. Define a slope estimator as

$$\tilde{\beta}_1 = \left(\sum_{i=1}^n (z_i - \bar{z}) y_i \right) / \left(\sum_{i=1}^n (z_i - \bar{z}) x_i \right).$$

Show that $\tilde{\beta}_1$ is linear and unbiased. Remember, because $E(ux) = 0$, you can treat both x_i and z_i as nonrandom in your derivation.

- (ii) Add the homoskedasticity assumption, MLR.5. Show that

$$\text{Var}(\tilde{\beta}_1) = \sigma^2 \left(\sum_{i=1}^n (z_i - \bar{z})^2 \right) / \left(\sum_{i=1}^n (z_i - \bar{z}) x_i \right)^2.$$

- (iii) Show directly that, under the Gauss-Markov assumptions, $\text{Var}(\hat{\beta}_1) \leq \text{Var}(\tilde{\beta}_1)$, where $\hat{\beta}_1$ is the OLS estimator. [Hint: The Cauchy-Schwartz inequality in Appendix B implies that

$$\left(n^{-1} \sum_{i=1}^n (z_i - \bar{z})(x_i - \bar{x}) \right)^2 \leq \left(n^{-1} \sum_{i=1}^n (z_i - \bar{z})^2 \right) \left(n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right);$$

notice that we can drop \bar{x} from the sample covariance.]

- 14 Suppose you have a sample of size n on three variables, y , x_1 , and x_2 , and you are primarily interested in the effect of x_1 on y . Let $\tilde{\beta}_1$ be the coefficient on x_1 from the simple regression and $\hat{\beta}_1$ the coefficient on x_1 from the regression y on x_1, x_2 . The standard errors reported by any regression package are

$$se(\tilde{\beta}_1) = \frac{\tilde{\sigma}}{\sqrt{SST_1}}$$

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_1}} \cdot \sqrt{VIF_1},$$

where $\hat{\sigma}$ is the SER from the simple regression, $\hat{\sigma}$ is the SER from the multiple regression, $VIF_1 = 1/(1 - R_1^2)$, and R_1^2 is the R -squared from the regression of x_1 on x_2 . Explain why $se(\hat{\beta}_1)$ can be smaller or larger than $se(\tilde{\beta}_1)$.

- 15 The following estimated equations use the data in MLB1, which contains information on major league baseball salaries. The dependent variable, *lsalary*, is the log of salary. The two explanatory variables are years in the major leagues (*years*) and runs batted in per year (*rbisyr*):

$$\widehat{lsalary} = 12.373 + .1770 \text{ years}$$

$$(.098) \quad (.0132)$$

$$n = 353, \text{ SSR} = 326.196, \text{ SER} = .964, R^2 = .337$$

$$\widehat{lsalary} = 11.861 + .0904 \text{ years} + .0302 \text{ rbisyr}$$

$$(.084) \quad (.0118) \quad (.0020)$$

$$n = 353, \text{ SSR} = 198.475, \text{ SER} = .753, R^2 = .597$$

- How many degrees of freedom are in each regression? How come the SER is smaller in the second regression than the first?
- The sample correlation coefficient between *years* and *rbisyr* is about 0.487. Does this make sense? What is the variance inflation factor (there is only one) for the slope coefficients in the multiple regression? Would you say there is little, moderate, or strong collinearity between *years* and *rbisyr*?
- How come the standard error for the coefficient on *years* in the multiple regression is lower than its counterpart in the simple regression?

- 16 The following equations were estimated using the data in LAWSCH85:

$$\widehat{lsalary} = 9.90 - .0041 \text{ rank} + .294 \text{ GPA}$$

$$(.24) \quad (.0003) \quad (.069)$$

$$n = 142, R^2 = .8238$$

$$\widehat{lsalary} = 9.86 - .0038 \text{ rank} + .295 \text{ GPA} + .00017 \text{ age}$$

$$(.29) \quad (.0004) \quad (.083) \quad (.00036)$$

$$n = 99, R^2 = .8036$$

How can it be that the R -squared is smaller when the variable *age* is added to the equation?

Computer Exercises

- C1 A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

$$bwght = \beta_0 + \beta_1 \text{cigs} + \beta_2 \text{faminc} + u.$$

- What is the most likely sign for β_2 ?
- Do you think *cigs* and *faminc* are likely to be correlated? Explain why the correlation might be positive or negative.

- (iii) Now, estimate the equation with and without *faminc*, using the data in BWGHT. Report the results in equation form, including the sample size and *R*-squared. Discuss your results, focusing on whether adding *faminc* substantially changes the estimated effect of *cigs* on *bwght*.

C2 Use the data in HPRICE1 to estimate the model

$$price = \beta_0 + \beta_1sqrft + \beta_2bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

- (i) Write out the results in equation form.
 - (ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?
 - (iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).
 - (iv) What percentage of the variation in price is explained by square footage and number of bedrooms?
 - (v) The first house in the sample has *sqrft* = 2,438 and *bdrms* = 4. Find the predicted selling price for this house from the OLS regression line.
 - (vi) The actual selling price of the first house in the sample was \$300,000 (so *price* = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?
- C3** The file CEOSAL2 contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.
- (i) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.
 - (ii) Add *profits* to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?
 - (iii) Add the variable *ceoten* to the model in part (ii). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?
 - (iv) Find the sample correlation coefficient between the variables $\log(mktval)$ and *profits*. Are these variables highly correlated? What does this say about the OLS estimators?
- C4** Use the data in ATTEND for this exercise.
- (i) Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.
 - (ii) Estimate the model

$$atndrte = \beta_0 + \beta_1priGPA + \beta_2ACT + u,$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

- (iii) Discuss the estimated slope coefficients. Are there any surprises?
 - (iv) What is the predicted *atndrte* if *priGPA* = 3.65 and *ACT* = 20? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?
 - (v) If Student A has *priGPA* = 3.1 and *ACT* = 21 and Student B has *priGPA* = 2.1 and *ACT* = 26, what is the predicted difference in their attendance rates?
- C5** Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out for Example 3.2. This first requires regressing *educ* on *exper* and *tenure* and saving the residuals, \hat{r}_1 . Then, regress $\log(wage)$ on \hat{r}_1 . Compare the coefficient on \hat{r}_1 with the coefficient on *educ* in the regression of $\log(wage)$ on *educ*, *exper*, and *tenure*.
- C6** Use the data set in WAGE2 for this problem. As usual, be sure all of the following regressions contain an intercept.
- (i) Run a simple regression of *IQ* on *educ* to obtain the slope coefficient, say, $\tilde{\delta}_1$.
 - (ii) Run the simple regression of $\log(wage)$ on *educ*, and obtain the slope coefficient, $\tilde{\beta}_1$.

- (iii) Run the multiple regression of $\log(\text{wage})$ on educ and IQ , and obtain the slope coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively.
- (iv) Verify that $\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2\tilde{\delta}_1$.

C7 Use the data in MEAP93 to answer this question.

- (i) Estimate the model

$$\text{math10} = \beta_0 + \beta_1\log(\text{expend}) + \beta_2\text{lnchprg} + u,$$

and report the results in the usual form, including the sample size and R -squared. Are the signs of the slope coefficients what you expected? Explain.

- (ii) What do you make of the intercept you estimated in part (i)? In particular, does it make sense to set the two explanatory variables to zero? [Hint: Recall that $\log(1) = 0$.]
- (iii) Now run the simple regression of math10 on $\log(\text{expend})$, and compare the slope coefficient with the estimate obtained in part (i). Is the estimated spending effect now larger or smaller than in part (i)?
- (iv) Find the correlation between $\text{lexpend} = \log(\text{expend})$ and lnchprg . Does its sign make sense to you?
- (v) Use part (iv) to explain your findings in part (iii).

C8 Use the data in DISCRIM to answer this question. These are ZIP code–level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

- (i) Find the average values of prpbck and income in the sample, along with their standard deviations. What are the units of measurement of prpbck and income ?
- (ii) Consider a model to explain the price of soda, psoda , in terms of the proportion of the population that is black and median income:

$$\text{psoda} = \beta_0 + \beta_1\text{prpbck} + \beta_2\text{income} + u.$$

Estimate this model by OLS and report the results in equation form, including the sample size and R -squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpbck . Do you think it is economically large?

- (iii) Compare the estimate from part (ii) with the simple regression estimate from psoda on prpbck . Is the discrimination effect larger or smaller when you control for income?
- (iv) A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

$$\log(\text{psoda}) = \beta_0 + \beta_1\text{prpbck} + \beta_2\log(\text{income}) + u.$$

If prpbck increases by .20 (20 percentage points), what is the estimated percentage change in psoda ? (Hint: The answer is 2.xx, where you fill in the “xx.”)

- (v) Now add the variable prppov to the regression in part (iv). What happens to $\hat{\beta}_{\text{prpbck}}$?
- (vi) Find the correlation between $\log(\text{income})$ and prppov . Is it roughly what you expected?
- (vii) Evaluate the following statement: “Because $\log(\text{income})$ and prppov are so highly correlated, they have no business being in the same regression.”

C9 Use the data in CHARITY to answer the following questions:

- (i) Estimate the equation

$$\text{gift} = \beta_0 + \beta_1\text{mailsyear} + \beta_2\text{giftlast} + \beta_3\text{propresp} + u$$

by OLS and report the results in the usual way, including the sample size and R -squared. How does the R -squared compare with that from the simple regression that omits giftlast and propresp ?

- (ii) Interpret the coefficient on *mailsyear*. Is it bigger or smaller than the corresponding simple regression coefficient?
- (iii) Interpret the coefficient on *propresp*. Be careful to notice the units of measurement of *propresp*.
- (iv) Now add the variable *avggift* to the equation. What happens to the estimated effect of *mailsyear*?
- (v) In the equation from part (iv), what has happened to the coefficient on *giftlast*? What do you think is happening?

C10 Use the data in HTV to answer this question. The data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

- (i) What is the range of the *educ* variable in the sample? What percentage of men completed twelfth grade but no higher grade? Do the men or their parents have, on average, higher levels of education?
- (ii) Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

by OLS and report the results in the usual form. How much sample variation in *educ* is explained by parents' education? Interpret the coefficient on *motheduc*.

- (iii) Add the variable *abil* (a measure of cognitive ability) to the regression from part (ii), and report the results in equation form. Does "ability" help to explain variations in education, even after controlling for parents' education? Explain.
- (iv) (Requires calculus) Now estimate an equation where *abil* appears in quadratic form:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u.$$

Using the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$, use calculus to find the value of *abil*, call it *abil**, where *educ* is minimized. (The other coefficients and values of parents' education variables have no effect; we are holding parents' education fixed.) Notice that *abil* is measured so that negative values are permissible. You might also verify that the second derivative is positive so that you do indeed have a minimum.

- (v) Argue that only a small fraction of men in the sample have "ability" less than the value calculated in part (iv). Why is this important?
- (vi) If you have access to a statistical program that includes graphing capabilities, use the estimates in part (iv) to graph the relationship between the predicted education and *abil*. Set *motheduc* and *fatheduc* at their average values in the sample, 12.18 and 12.45, respectively.

C11 Use the data in MEAPSINGLE to study the effects of single-parent households on student math performance. These data are for a subset of schools in southeast Michigan for the year 2000. The socioeconomic variables are obtained at the ZIP code level (where ZIP code is assigned to schools based on their mailing addresses).

- (i) Run the simple regression of *math4* on *pctsgle* and report the results in the usual format. Interpret the slope coefficient. Does the effect of single parenthood seem large or small?
- (ii) Add the variables *lmedinc* and *free* to the equation. What happens to the coefficient on *pctsgle*? Explain what is happening.
- (iii) Find the sample correlation between *lmedinc* and *free*. Does it have the sign you expect?
- (vi) Does the substantial correlation between *lmedinc* and *free* mean that you should drop one from the regression to better estimate the causal effect of single parenthood on student performance? Explain.
- (v) Find the variance inflation factors (VIFs) for each of the explanatory variables appearing in the regression in part (ii). Which variable has the largest VIF? Does this knowledge affect the model you would use to study the causal effect of single parenthood on math performance?

- C12** The data in ECONMATH contain grade point averages and standardized test scores, along with performance in an introductory economics course, for students at a large public university. The variable to be explained is *score*, the final score in the course measured as a percentage.
- How many students received a perfect score for the course? What was the average score? Find the means and standard deviations of *actmth* and *acteng*, and discuss how they compare.
 - Estimate a linear equation relating *score* to *colgpa*, *actmth*, and *acteng*, where *colgpa* is measured at the beginning of the term. Report the results in the usual form.
 - Would you say the math or English ACT score is a better predictor of performance in the economics course? Explain.
 - Discuss the size of the *R*-squared in the regression.

APPENDIX 3A

3A.1 Derivation of the First Order Conditions in Equation (3.13)

The analysis is very similar to the simple regression case. We must characterize the solutions to the problem

$$\min_{b_0, b_1, \dots, b_k} \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2.$$

Taking the partial derivatives with respect to each of the b_j (see Appendix A), evaluating them at the solutions, and setting them equal to zero gives

$$\begin{aligned} -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0 \\ -2 \sum_{i=1}^n x_{ij} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0, \quad \text{for all } j = 1, \dots, k. \end{aligned}$$

Canceling the -2 gives the first order conditions in (3.13).

3A.2 Derivation of Equation (3.22)

To derive (3.22), write x_{i1} in terms of its fitted value and its residual from the regression of x_1 on x_2, \dots, x_k : $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, for all $i = 1, \dots, n$. Now, plug this into the second equation in (3.13):

$$\sum_{i=1}^n (\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0. \tag{3.63}$$

By the definition of the OLS residual \hat{u}_i , since \hat{x}_{i1} is just a linear function of the explanatory variables x_{i2}, \dots, x_{ik} , it follows that $\sum_{i=1}^n \hat{x}_{i1} \hat{u}_i = 0$. Therefore, equation (3.63) can be expressed as

$$\sum_{i=1}^n \hat{r}_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0. \tag{3.64}$$

Since the \hat{r}_{i1} are the residuals from regressing x_1 on x_2, \dots, x_k , $\sum_{i=1}^n x_{ij} \hat{r}_{i1} = 0$, for all $j = 2, \dots, k$. Therefore, (3.64) is equivalent to $\sum_{i=1}^n \hat{r}_{i1} (y_i - \hat{\beta}_1 x_{i1}) = 0$. Finally, we use the fact that $\sum_{i=1}^n \hat{x}_{i1} \hat{r}_{i1} = 0$, which means that $\hat{\beta}_1$ solves

$$\sum_{i=1}^n \hat{r}_{i1} (y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0.$$

Now, straightforward algebra gives (3.22), provided, of course, that $\sum_{i=1}^n \hat{r}_{i1}^2 > 0$; this is ensured by Assumption MLR.3.