In Section 6-3, we discussed the somewhat subtle problem of relying too much on R^2 or $\overline{R}^2$ in arriving at a final model: it is possible to control for too many factors in a regression model. For this reason, it is important to think ahead about model specification, particularly the ceteris paribus nature of the multiple regression equation. Explanatory variables that affect y and are uncorrelated with all the other explanatory variables can be used to reduce the error variance without inducing multicollinearity.

In Section 6-4, we demonstrated how to obtain a confidence interval for a prediction made from an OLS regression line. We also showed how a confidence interval can be constructed for a future, unknown value of y.

Occasionally, we want to predict y when $\log(y)$ is used as the dependent variable in a regression model. Section 6-4 explains this simple method. Finally, we are sometimes interested in knowing about the sign and magnitude of the residuals for particular observations. Residual analysis can be used to determine whether particular members of the sample have predicted values that are well above or well below the actual outcomes.

Key Terms

Adjusted *R*-Squared
Average Partial Effect (APE)
Beta Coefficients
Bootstrap
Bootstrap Standard Error
Interaction Effect
Nonnested Models
Over Controlling
Population *R*-Squared
Prediction Error
Prediction Interval
Predictions
Quadratic Functions
Resampling Method
Residual Analysis
Smearing Estimate
Standardized Coefficients
Variance of the Prediction Error

Problems

1 The following equation was estimated using the data in CEOSAL1:

$$\widehat{\log(salary)} = \underset{}{4.322} + \underset{(.324)}{.276}\log(sales) + \underset{(.033)}{.0215}\,roe - \underset{(.0129)}{.00008}\,roe^2 \qquad \underset{(.00026)}{}$$
$$n = 209,\ R^2 = .282.$$

This equation allows *roe* to have a diminishing effect on log(*salary*). Is this generality necessary? Explain why or why not.

2 Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$, $i = 1, 2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$, are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \hat{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Hint*: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.]

3 Using the data in RDCHEM, the following equation was obtained by OLS:

$$\widehat{rdintens} = \underset{}{2.613} + \underset{(.429)}{.00030}\,sales - \underset{(.00014)}{.0000000070}\,sales^2 \qquad \underset{(.0000000037)}{}$$
$$n = 32,\ R^2 = .1484.$$

(i) At what point does the marginal effect of *sales* on *rdintens* become negative?
(ii) Would you keep the quadratic term in the model? Explain.

(iii) Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1{,}000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that $salesbil^2 = sales^2/(1{,}000)^2$.]

(iv) For the purpose of reporting the results, which equation do you prefer?

4 The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ \cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

(i) Show that, in decimal form, the return to another year of education in this model is

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

What sign do you expect for β_2? Why?

(ii) Using the data in WAGE2, the estimated equation is

$$\begin{aligned}\widehat{\log(wage)} &= 5.65 + .047\ educ + .00078\ educ \cdot pareduc + \\ &\quad (.13)\ \ (.010)\qquad (.00021) \\ &\quad .019\ exper + .010\ tenure \\ &\quad (.004)\qquad (.003) \\ n &= 722,\ R^2 = .169.\end{aligned}$$

(Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, $pareduc = 32$ if both parents have a college education, or $pareduc = 24$ if both parents have a high school education—and to compare the estimated return to *educ*.

(iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\begin{aligned}\widehat{\log(wage)} &= 4.94 + .097\ educ + .033\ pareduc - .0016\ educ \cdot pareduc \\ &\quad (.38)\ \ (.027)\qquad (.017)\qquad\quad (.0012) \\ &\quad + .020\ exper + .010\ tenure \\ &\quad (.004)\qquad (.003) \\ n &= 722,\ R^2 = .174.\end{aligned}$$

Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

5 In Example 4.2, where the percentage of students receiving a passing score on a tenth-grade math exam (*math*10) is the dependent variable, does it make sense to include *sci11*—the percentage of eleventh graders passing a science exam—as an additional explanatory variable?

6 When $atndrte^2$ and $ACT \cdot atndrte$ are added to the equation estimated in (6.19), the *R*-squared becomes .232. Are these additional terms jointly significant at the 10% level? Would you include them in the model?

7 The following three equations were estimated using the 1,534 observations in 401K:

$$\begin{aligned}\widehat{prate} &= 80.29 + 5.44\ mrate + .269\ age - .00013\ totemp \\ &\quad (.78)\ \ (.52)\qquad (.045)\qquad (.00004) \\ R^2 &= .100,\ \overline{R}^2 = .098.\end{aligned}$$

$$\widehat{prate} = 97.32 + 5.02\, mrate + .314\, age - 2.66 \log(totemp)$$
$$(1.95)\quad (0.51)\qquad (.044)\qquad (.28)$$
$$R^2 = .144, \overline{R}^2 = .142.$$

$$\widehat{prate} = 80.62 + 5.34\, mrate + .290\, age - .00043\, totemp$$
$$(.78)\quad (.52)\qquad (.045)\qquad (.00009)$$
$$+ .0000000039\, totemp^2$$
$$(.00000000010)$$
$$R^2 = .108, \overline{R}^2 = .106.$$

Which of these three models do you prefer? Why?

8 Suppose we want to estimate the effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized test score (say, *SAT*) and high school GPA (*hsGPA*) are also available.

(i) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret β_{alcohol}.)

(ii) Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

9 If we start with (6.38) under the CLM assumptions, assume large n, and ignore the estimation error in the $\hat{\beta}_j$, a 95% prediction interval for y^0 is $[\exp(-1.96\hat{\sigma})\exp(\widehat{logy^0}), \exp(1.96\hat{\sigma})\exp(\widehat{logy^0})]$. The point prediction for y^0 is $\hat{y}^0 = \exp(\hat{\sigma}_2)\exp(\widehat{logy^0})$.

(i) For what values of $\hat{\sigma}$ will the point prediction be in the 95% prediction interval? Does this condition seem likely to hold in most applications?

(ii) Verify that the condition from part (i) is satisfied in the CEO salary example.

10 The following two equations were estimated using the data in MEAPSINGLE. The key explanatory variable is *lexppp*, the log of expenditures per student at the school level.

$$\widehat{math4} = 24.49 + 9.01\, lexppp - .422\, free - .752\, lmedinc - .274\, pctsgle$$
$$(59.24)\quad (4.04)\qquad (.071)\qquad (5.358)\qquad (.161)$$
$$n = 229, R^2 = .472, \overline{R}^2 = .462.$$

$$\widehat{math4} = 149.38 + 1.93\, lexppp - .060\, free - 10.78\, lmedinc - .397\, pctsgle + .667\, read4$$
$$(41.70)\quad (2.82)\qquad (.054)\qquad (3.76)\qquad (.111)\qquad (.042)$$
$$n = 229, R^2 = .749, \overline{R}^2 = .743.$$

(i) If you are a policy maker trying to estimate the causal effect of per-student spending on math test performance, explain why the first equation is more relevant than the second. What is the estimated effect of a 10% increase in expenditures per student?

(ii) Does adding *read4* to the regression have strange effects on coefficients and statistical significance other than β_{lexppp}?

(iii) How would you explain to someone with only basic knowledge of regression why, in this case, you prefer the equation with the smaller adjusted R-squared?

Computer Exercises

C1 Use the data in KIELMC, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.

(i) To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.

(ii) To the simple regression model in part (i), add the variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths*, and *age*, where *intst* is distance from the home to the interstate, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why (i) and (ii) give conflicting results.

(iii) Add $[\log(intst)]^2$ to the model from part (ii). Now what happens? What do you conclude about the importance of functional form?

(iv) Is the square of log(*dist*) significant when you add it to the model from part (iii)?

C2 Use the data in WAGE1 for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\widehat{\Delta wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

C3 Consider a model where the return to education depends upon the amount of work experience (and vice versa):

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

(i) Show that the return to another year of education (in decimal form), holding *exper* fixed, is $\beta_1 + \beta_3 exper$.

(ii) State the null hypothesis that the return to education does not depend on the level of *exper*. What do you think is the appropriate alternative?

(iii) Use the data in WAGE2 to test the null hypothesis in (ii) against your stated alternative.

(iv) Let θ_1 denote the return to education (in decimal form), when $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_1$ and a 95% confidence interval for θ_1. (*Hint:* Write $\beta_1 = \theta_1 - 10\beta_3$ and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for θ_1.)

C4 Use the data in GPA2 for this exercise.

(i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

where *hsize* is the size of the graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

(iii) Is this analysis representative of the academic performance of *all* high school seniors? Explain.

(iv) Find the estimated optimal high school size, using log(*sat*) as the dependent variable. Is it much different from what you obtained in part (ii)?

C5 Use the housing price data in HPRICE1 for this exercise.

(i) Estimate the model

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

and report the results in the usual OLS format.

(ii) Find the predicted value of log(*price*), when $lotsize = 20{,}000$, $sqrft = 2{,}500$, and $bdrms = 4$. Using the methods in Section 6-4, find the predicted value of *price* at the same values of the explanatory variables.

(iii) For explaining variation in *price*, decide whether you prefer the model from part (i) or the model

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

C6 Use the data in VOTE1 for this exercise.

(i) Consider a model with an interaction between expenditures:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

What is the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed? What is the partial effect of *expendA* on *voteA*? Is the expected sign for β_4 obvious?

(ii) Estimate the equation in part (i) and report the results in the usual form. Is the interaction term statistically significant?

(iii) Find the average of *expendA* in the sample. Fix *expendA* at 300 (for \$300,000). What is the estimated effect of another \$100,000 spent by Candidate B on *voteA*? Is this a large effect?

(iv) Now fix *expendB* at 100. What is the estimated effect of $\Delta expendA = 100$ on *voteA*? Does this make sense?

(v) Now, estimate a model that replaces the interaction with *shareA*, Candidate A's percentage share of total campaign expenditures. Does it make sense to hold both *expendA* and *expendB* fixed, while changing *shareA*?

(vi) (Requires calculus) In the model from part (v), find the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed. Evaluate this at $expendA = 300$ and $expendB = 0$ and comment on the results.

C7 Use the data in ATTEND for this exercise.

(i) In the model of Example 6.3, argue that

$$\Delta stndfnl / \Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Use equation (6.19) to estimate the partial effect when $priGPA = 2.59$ and $atndrte = 82$. Interpret your estimate.

(ii) Show that the equation can be written as

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u,$$

where $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. (Note that the intercept has changed, but this is unimportant.) Use this to obtain the standard error of $\hat{\theta}_2$ from part (i).

(iii) Suppose that, in place of $priGPA(atndrte - 82)$, you put $(priGPA - 2.59) \cdot (atndrte - 82)$. Now how do you interpret the coefficients on *atndrte* and *priGPA*?

C8 Use the data in HPRICE1 for this exercise.

(i) Estimate the model

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

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10{,}000$, $sqrft = 2{,}300$, and $bdrms = 4$; round this price to the nearest dollar.

(ii) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.

(iii) Let $price^0$ be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for $price^0$ and comment on the width of this confidence interval.

C9 The data set NBASAL contains salary information and career statistics for 269 players in the National Basketball Association (NBA).

(i) Estimate a model relating points-per-game (*points*) to years in the league (*exper*), *age*, and years played in college (*coll*). Include a quadratic in *exper*; the other variables should appear in level form. Report the results in the usual way.

(ii) Holding college years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?

(iii) Why do you think *coll* has a negative and statistically significant coefficient? (*Hint:* NBA players can be drafted before finishing their college careers and even directly out of high school.)

(iv) Add a quadratic in *age* to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and education are controlled for?

(v) Now regress log(*wage*) on *points*, *exper*, $exper^2$, *age*, and *coll*. Report the results in the usual format.

(vi) Test whether *age* and *coll* are jointly significant in the regression from part (v). What does this imply about whether age and education have separate effects on wage, once productivity and seniority are accounted for?

C10 Use the data in BWGHT2 for this exercise.

(i) Estimate the equation

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

by OLS, and report the results in the usual way. Is the quadratic term significant?

(ii) Show that, based on the equation from part (i), the number of prenatal visits that maximizes log(*bwght*) is estimated to be about 22. How many women had at least 22 prenatal visits in the sample?

(iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits? Explain.

(iv) Add mother's age to the equation, using a quadratic functional form. Holding *npvis* fixed, at what mother's age is the birth weight of the child maximized? What fraction of women in the sample are older than the "optimal" age?

(v) Would you say that mother's age and number of prenatal visits explain a lot of the variation in log(*bwght*)?

(vi) Using quadratics for both *npvis* and *age*, decide whether using the natural log or the level of *bwght* is better for predicting *bwght*.

C11 Use APPLE to verify some of the claims made in Section 6-3.

(i) Run the regression *ecolbs* on *ecoprc*, *regprc* and report the results in the usual form, including the *R*-squared and adjusted *R*-squared. Interpret the coefficients on the price variables and comment on their signs and magnitudes.

(ii) Are the price variables statistically significant? Report the *p*-values for the individual *t* tests.

(iii) What is the range of fitted values for *ecolbs*? What fraction of the sample reports *ecolbs* $= 0$? Comment.

(iv) Do you think the price variables together do a good job of explaining variation in *ecolbs*? Explain.

(v) Add the variables *faminc*, *hhsize* (household size), *educ*, and *age* to the regression from part (i). Find the *p*-value for their joint significance. What do you conclude?

(vi) Run separate simple regressions of *ecolbs* on *ecoprc* and then *ecolbs* on *regprc*. How do the simple regression coefficients compare with the multiple regression from part (i)? Find the correlation coefficient between *ecoprc* and *regprc* to help explain your findings.

(vii) Consider a model that adds family income and the quantity demanded for regular apples:

$$ecolbs = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 reglbs + u.$$

From basic economic theory, which explanatory variable does not belong to the equation? When you drop the variables one at a time time, do the sizes of the adjusted *R*-squareds affect your answer?

C12 Use the subset of 401KSUBS with *fsize* $= 1$; this restricts the analysis to single-person households; see also Computer Exercise C8 in Chapter 4.

(i) What is the youngest age of people in this sample? How many people are at that age?

(ii) In the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

what is the literal interpretation of β_2? By itself, is it of much interest?

(iii) Estimate the model from part (ii) and report the results in standard form. Are you concerned that the coefficient on *age* is negative? Explain.

(iv) Because the youngest people in the sample are 25, it makes sense to think that, for a given level of income, the lowest average amount of net total financial assets is at age 25. Recall that the partial effect of *age* on *nettfa* is $\beta_2 + 2\beta_3 age$, so the partial effect at age 25 is $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; call this θ_2. Find $\hat{\theta}_2$ and obtain the two-sided *p*-value for testing H_0: $\theta_2 = 0$. You should conclude that $\hat{\theta}_2$ is small and very statistically insignificant. [*Hint:* One way to do this is to estimate the model $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, where the intercept, α_0 is different from β_0. There are other ways, too.]

(v) Because the evidence against H_0: $\theta_2 = 0$ is very weak, set it to zero and estimate the model

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

In terms of goodness-of-fit, does this model fit better than that in part (ii)?

(vi) For the estimated equation in part (v), set *inc* $= 30$ (roughly, the average value) and graph the relationship between *nettfa* and *age*, but only for *age* ≥ 25. Describe what you see.

(vii) Check to see whether including a quadratic in *inc* is necessary.

C13 Use the data in MEAP00 to answer this question.

(i) Estimate the model

$$math4 = \beta_0 + \beta_2 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

by OLS, and report the results in the usual form. Is each explanatory variable statistically significant at the 5% level?

(ii) Obtain the fitted values from the regression in part (i). What is the range of fitted values? How does it compare with the range of the actual data on *math*4?

(iii) Obtain the residuals from the regression in part (i). What is the building code of the school that has the largest (positive) residual? Provide an interpretation of this residual.

(iv) Add quadratics of all explanatory variables to the equation, and test them for joint significance. Would you leave them in the model?

(v) Returning to the model in part (i), divide the dependent variable and each explanatory variable by its sample standard deviation, and rerun the regression. (Include an intercept unless you also first subtract the mean from each variable.) In terms of standard deviation units, which explanatory variable has the largest effect on the math pass rate?

C14 Use the data in BENEFITS to answer this question. It is a school-level data set at the K–5 level on average teacher salary and benefits. See Example 4.10 for background.

(i) Regress *lavgsal* on *bs* and report the results in the usual form. Can you reject H_0: $\beta_{bs} = 0$ against a two-sided alternative? Can you reject H_0: $\beta_{bs} = -1$ against H_1: $\beta_{bs} > -1$? Report the *p*-values for both tests.

(ii) Define $lbs = \log(bs)$. Find the range of values for *lbs* and find its standard deviation. How do these compare to the range and standard deviation for *bs*?

(iii) Regress *lavgsal* on *lbs*. Does this fit better than the regression from part (i)?

(iv) Estimate the equation

$$lavgsal = \beta_0 + \beta_1 bs + \beta_2 lenroll + \beta_3 lstaff + \beta_4 lunch + u$$

and report the results in the usual form. What happens to the coefficient on *bs*? Is it now statistically different from zero?

(v) Interpret the coefficient on *lstaff*. Why do you think it is negative?

(vi) Add $lunch^2$ to the equation from part (iv). Is it statistically significant? Compute the turning point (minimum value) in the quadratic, and show that it is within the range of the observed data on *lunch*. How many values of *lunch* are higher than the calculated turning point?

(vii) Based on the findings from part (vi), describe how teacher salaries relate to school poverty rates. In terms of teacher salary, and holding other factors fixed, is it better to teach at a school with $lunch = 0$ (no poverty), $lunch = 50$, or $lunch = 100$ (all kids eligible for the free lunch program)?

APPENDIX 6A

6A. A Brief Introduction to Bootstrapping

In many cases where formulas for standard errors are hard to obtain mathematically, or where they are thought not to be very good approximations to the true sampling variation of an estimator, we can rely on a **resampling method**. The general idea is to treat the observed data as a population that we can draw samples from. The most common resampling method is the **bootstrap**. (There are actually several versions of the bootstrap, but the most general, and most easily applied, is called the *nonparametric bootstrap*, and that is what we describe here.)