exercise shows that interpretation also applies under conditional mean independence. Consider the hypothetical experiment in Exercise 7.11.

**a.** Suppose that you estimate the regression $Y_i = \gamma_0 + \gamma_1 X_{1i} + u_i$ using only the data on returning students. Show that $\gamma_1$ is the class size effect for returning students, that is, that $\gamma_1 = E(Y_i|X_{1i} = 1, X_{2i} = 0) - E(Y_i|X_{1i} = 0, X_{2i} = 0)$. Explain why $\hat{\gamma}_1$ is an unbiased estimator of $\gamma_1$.

**b.** Suppose that you estimate the regression $Y_i = \delta_0 + \delta_1 X_{1i} + u_i$ using only the data on new students. Show that $\delta_1$ is the class size effect for new students, that is, that $\delta_1 = E(Y_i|X_{1i} = 1, X_{2i} = 1) - E(Y_i|X_{1i} = 0, X_{2i} = 1)$. Explain why $\hat{\delta}_1$ is an unbiased estimator of $\delta_1$.

**c.** Consider the regression for both returning and new students, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$. Use the conditional mean independence assumption $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$ to show that $\beta_1 = \gamma_1$, $\beta_1 + \beta_3 = \delta_1$, and $\beta_3 = \delta_1 - \gamma_1$ (the difference in the class size effects).

**d.** Suppose that you estimate the interaction regression in (c) using the combined data and that $E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$. Show that $\hat{\beta}_1$ and $\hat{\beta}_3$ are unbiased but that $\hat{\beta}_2$ is in general biased.

## Empirical Exercises

**E8.1** Use the data set **CPS08** described in Empirical Exercise 4.1 to answer the following questions.

**a.** Run a regression of average hourly earnings (*AHE*) on age (*Age*), gender (*Female*), and education (*Bachelor*). If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**b.** Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**c.** Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on ln(*Age*), *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**d.** Run a regression of the logarithm of average hourly earnings, ln(*AHE*), on *Age*, $Age^2$, *Female*, and *Bachelor*. If *Age* increases from

25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

**e.** Do you prefer the regression in (c) to the regression in (b)? Explain.

**f.** Do you prefer the regression in (d) to the regression in (b)? Explain.

**g.** Do you prefer the regression in (d) to the regression in (c)? Explain.

**h.** Plot the regression relation between *Age* and ln(*AHE*) from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degrees?

**i.** Run a regression of ln(*AHE*), on *Age*, $Age^2$, *Female*, *Bachelor*, and the interaction term *Female* × *Bachelor*. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of ln(*AHE*)? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of ln(*AHE*)? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of ln(*AHE*)? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of ln(*AHE*)? What is the predicted difference between Bob's and Jim's earnings?

**j.** Is the effect of *Age* on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question.

**k.** Is the effect of *Age* on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

**l.** After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

**E8.2** Using the data set **TeachingRatings** described in Empirical Exercise 4.2, carry out the following exercises.

**a.** Estimate a regression of *Course_Eval* on *Beauty*, *Intro*, *OneCredit*, *Female*, *Minority*, and *NNEnglish*.

**b.** Add *Age* and $Age^2$ to the regression. Is there evidence that *Age* has a nonlinear effect on *Course_Eval*? Is there evidence that *Age* has any effect on *Course_Eval*?

**c.** Modify the regression in (a) so that the effect of *Beauty* on *Course_Eval* is different for men and women. Is the male–female difference in the effect of *Beauty* statistically significant?

**d.** Professor Smith is a man. He has cosmetic surgery that increases his beauty index from one standard deviation below the average to one standard deviation above the average. What is his value of *Beauty* before the surgery? After the surgery? Using the regression in (c), construct a 95% confidence for the increase in his course evaluation.

**e.** Repeat (d) for Professor Jones, who is a woman.

**E8.3** Use the data set **CollegeDistance** described in Empirical Exercise 4.3 to answer the following questions.

**a.** Run a regression of *ED* on *Dist*, *Female*, *Bytest*, *Tuition*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *MomColl*, *Cue80*, and *Stwmfg80*. If *Dist* increases from 2 to 3 (that is, from 20 to 30 miles), how are years of education expected to change? If *Dist* increases from 6 to 7 (that is, from 60 to 70 miles), how are years of education expected to change?

**b.** Run a regression of ln(*ED*) on *Dist*, *Female*, *Bytest*, *Tuition*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *MomColl*, *Cue80*, and *Stwmfg80*. If *Dist* increases from 2 to 3 (from 20 to 30 miles), how are years of education expected to change? If *Dist* increases from 6 to 7 (from 60 to 70 miles), how are years of education expected to change?

**c.** Run a regression of *ED* on *Dist*, $Dist^2$, *Female*, *Bytest*, *Tuition*, *Black*, *Hispanic*, *Incomehi*, *Ownhome*, *DadColl*, *MomColl*, *Cue80*, and *Stwmfg80*. If *Dist* increases from 2 to 3 (from 20 to 30 miles), how are years of education expected to change? If *Dist* increases from 6 to 7 (from 60 to 70 miles), how are years of education expected to change?

**d.** Do you prefer the regression in (c) to the regression in (a)? Explain.

**e.** Consider a Hispanic female with *Tuition* = \$950, *Bytest* = 58, *Incomehi* = 0, *Ownhome* = 0, *DadColl* = 1, *MomColl* = 1, *Cue80* = 7.1, and *Stwmfg* = \$10.06.

i. Plot the regression relation between *Dist* and *ED* from (a) and (c) for *Dist* in the range of 0 to 10 (from 0 to 100 miles). Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for a white male with the same characteristics?

ii. How does the regression function (c) behave for *Dist* > 10? How many observation are there with *Dist* > 10?

**f.** Add the interaction term $DadColl \times MomColl$ to the regression in (c). What does the coefficient on the interaction term measure?

**g.** Mary, Jane, Alexis, and Bonnie have the same values of *Dist*, *Bytest*, *Tuition*, *Female*, *Black*, *Hispanic*, *Fincome*, *Ownhome*, *Cue80* and *Stwmfg80*. Neither of Mary's parents attended college. Jane's father attended college, but her mother did not. Alexis's mother attended college, but her father did not. Both of Bonnie's parents attended college. Using the regressions from (f):

i. What does the regression predict for the difference between Jane's and Mary's years of education?

ii. What does the regression predict for the difference between Alexis's and Mary's years of education?

iii. What does the regression predict for the difference between Bonnie's and Mary's years of education?

**h.** Is there any evidence that the effect of *Dist* on *ED* depends on the family's income?

**i.** After running all these regressions (and any others that you want to run), summarize the effect of *Dist* on years of education.

**E8.4** Using the data set **Growth** described in Empirical Exercise 4.4, excluding the data for Malta, run the following five regressions: *Growth* on (1) *TradeShare* and *YearsSchool*; (2) *TradeShare* and ln(*YearsSchool*); (3) *TradeShare*, ln(*YearsSchool*), *Rev_Coups*, *Assassinations* and ln(*RGDP60*); (4) *TradeShare*, ln(*YearsSchool*), *Rev_Coups*, *Assassinations*, ln(*RGDP60*), and *TradeShare* × ln(*YearsSchool*); and (5) *TradeShare*, $TradeShare^2$, $TradeShare^3$, ln(*YearsSchool*), *Rev_Coups*, *Assassinations*, and ln(*RGDP60*).

**a.** Construct a scatterplot of *Growth* on *YearsSchool*. Does the relationship look linear or nonlinear? Explain. Use the plot to explain why regression (2) fits better than regression (1).

**b.** In 1960, a country contemplated an education policy that would increase average years of schooling from 4 years to 6 years. Use regression (1) to predict the increase in *Growth*. Use regression (2) to predict the increase in *Growth*.

**c.** Test whether the coefficients on *Assassinations* and *Rev_Coups* are equal to zero using regression (3).

**d.** Using regression (4), is there evidence that the effect of *TradeShare* on *Growth* depends on the level of education in the country?

**e.** Using regression (5), is there evidence of a nonlinear relationship between *TradeShare* and *Growth*?

**f.** In 1960, a country contemplated a trade policy that would increase the average value of *TradeShare* from 0.5 to 1. Use regression (3) to predict the increase in *Growth*. Use regression (5) to predict the increase in *Growth*.

APPENDIX

# 8.1 Regression Functions That Are Nonlinear in the Parameters

The nonlinear regression functions considered in Sections 8.2 and 8.3 are nonlinear functions of the $X$'s but are linear functions of the unknown parameters. Because they are linear in the unknown parameters, those parameters can be estimated by OLS after defining new regressors that are nonlinear transformations of the original $X$'s. This family of nonlinear regression functions is both rich and convenient to use. In some applications, however, economic reasoning leads to regression functions that are not linear in the parameters. Although such regression functions cannot be estimated by OLS, they can be estimated using an extension of OLS called nonlinear least squares.

## Functions That Are Nonlinear in the Parameters

We begin with two examples of functions that are nonlinear in the parameters. We then provide a general formulation.

*Logistic curve.* Suppose that you are studying the market penetration of a technology, such as the adoption of database management software in different industries. The dependent variable is the fraction of firms in the industry that have adopted the software, a single independent variable $X$ describes an industry characteristic, and you have data on $n$ industries. The dependent variable is between 0 (no adopters) and 1 (100% adoption). Because a linear regression model could produce predicted values less than 0 or greater than 1, it makes sense to use instead a function that produces predicted values between 0 and 1.

The logistic function smoothly increases from a minimum of 0 to a maximum of 1. The logistic regression model with a single $X$ is

$$Y_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_i)}} + u_i. \tag{8.38}$$