Chapter Review

Computer Exercises

C1. Use the data in GPA1 for this exercise.

i. Add the variables *mothcoll* and *fathcoll* to the equation estimated in (7.6) and report the results in the usual form. What happens to the estimated effect of PC ownership? Is *PC* still statistically significant?

ii. Test for joint significance of *mothcoll* and *fathcoll* in the equation from part (i) and be sure to report the *p*-value.

iii. Add $hsGPA^2$ to the model from part (i) and decide whether this generalization is needed.

C2. Use the data in WAGE2 for this exercise.

i. Estimate the model

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married \\ &\quad + \beta_5 black + \beta_6 south + \beta_7 urban + u\end{aligned}$$

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant?

ii. Add the variables $exper^2$ and $tenure^2$ to the equation and show that they are jointly insignificant at even the 20% level.

iii. Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.

iv. Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married blacks and married nonblacks?

C3. A model that allows major league baseball player salary to differ by position is

$$\begin{aligned}\log(salary) &= \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr \\ &\quad + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar \\ &\quad + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop \\ &\quad + \beta_{13} catcher + u,\end{aligned}$$

where outfield is the base group.

i. State the null hypothesis that, controlling for other factors, catchers and outfielders earn, on average, the same amount. Test this hypothesis using the data in MLB1 and comment on the size of the estimated salary differential.

ii. State and test the null hypothesis that there is no difference in average salary across positions, once other factors have been controlled for.

iii. Are the results from parts (i) and (ii) consistent? If not, explain what is happening.

C4. Use the data in GPA2 for this exercise.

i. Consider the equation

$$\begin{aligned}colgpa &= \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat \\ &\quad + \beta_5 female + \beta_6 athlete + u,\end{aligned}$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; *female* is a binary gender variable; and *athlete* is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

ii. Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

iii. Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

iv. In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

v. Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

C5. In Problem 2 in Chapter 4, we added the return on the firm's stock, *ros*, to a model explaining CEO salary; *ros* turned out to be insignificant. Now, define a dummy variable, *rosneg*, which is equal to one if $ros < 0$ and equal to zero if $ros \geq 0$. Use CEOSAL1 to estimate the model

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

C6. Discuss the interpretation and statistical significance of $\hat{\beta}_3$.

Use the data in SLEEP75 for this exercise. The equation of interest is

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

i. Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences in the two estimated equations?

ii. Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms *male*·*totwrk*, …, *male*·*yngkid* and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?

iii. Now, allow for a different intercept for males and females and determine whether the interaction terms involving *male* are jointly significant.

iv. Given the results from parts (ii) and (iii), what would be your final model?

C7. Use the data in WAGE1 for this exercise.

i. Use equation (7.18) to estimate the gender differential when $educ = 12.5$. Compare this with the estimated differential when $educ = 0$.

ii. Run the regression used to obtain (7.18), but with *female*·(*educ* − 12.5) replacing *female*·*educ*. How do you interpret the coefficient on *female* now?

iii. Is the coefficient on *female* in part (ii) statistically significant? Compare this with (7.18) and comment.

C8. Use the data in LOANAPP for this exercise. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal

to one if the applicant was white. The other applicants in the data set are black and Hispanic.

To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

i. If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?

ii. Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

iii. As controls, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?

iv. Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?

v. Using the model from part (iv), what is the effect of being white on the probability of approval when $obrat = 32$, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

C9. There has been much interest in whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS contains information on net financial assets (*nettfa*), family income (*inc*), a binary variable for eligibility in a 401(k) plan (*e401k*), and several other variables.

i. What fraction of the families in the sample are eligible for participation in a 401(k) plan?

ii. Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

iii. Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.

iv. Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?

v. Using the fitted values $\widehat{e401k}_i$ from part (iv), define $\widetilde{e401k}_i = 1$ if $\widehat{e401k}_i \geq .5$ and $\widetilde{e401k}_i = 0$ if $\widehat{e401k}_i < .5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?

vi. For the 5,638 families not eligible for a 401(k), what percentage of these are predicted not to have a 401(k), using the predictor $\widehat{e401k}_i$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one? (It is helpful if your econometrics package has a "tabulate" command.)

vii. The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answers in part (vi)?

viii. Add the variable *pira* as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for a 401(k) plan? Is it statistically different from zero at the 10% level?

C10. Use the data in NBASAL for this exercise.

i. Estimate a linear regression model relating points per game to experience in the league and position (guard, forward, or center). Include experience in quadratic form and use centers as the base group. Report the results in the usual form.

ii. Why do you not include all three position dummy variables in part (i)?

iii. Holding experience fixed, does a guard score more than a center? How much more? Is the difference statistically significant?

iv. Now, add marital status to the equation. Holding position and experience fixed, are married players more productive (based on points per game)?

v. Add interactions of marital status with both experience variables. In this expanded model, is there strong evidence that marital status affects points per game?

vi. Estimate the model from part (iv) but use assists per game as the dependent variable. Are there any notable differences from part (iv)? Discuss.

C11. Use the data in 401KSUBS for this exercise.

i. Compute the average, standard deviation, minimum, and maximum values of *nettfa* in the sample.

ii. Test the hypothesis that average *nettfa* does not differ by 401(k) eligibility status; use a two-sided alternative. What is the dollar amount of the estimated difference?

iii. From part (ii) of Computer Exercise C9, it is clear that *e401k* is not exogenous in a simple regression model; at a minimum, it changes by income and age. Estimate a multiple linear regression model for *nettfa* that includes income, age, and *e401k* as explanatory variables. The income and age variables should appear as quadratics. Now, what is the estimated dollar effect of 401(k) eligibility?

iv. To the model estimated in part (iii), add the interactions *e401k*·(*age* − 41) and $e401k \cdot (age - 41)^2$. Note that the average age in the sample is about 41, so that in the new model, the coefficient on *e401k* is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?

v. Comparing the estimates from parts (iii) and (iv), do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.

vi. Now, drop the interaction terms from the model, but define five family size dummy variables: *fsize1, fsize2, fsize3, fsize4,* and *fsize5*. The variable *fsize5* is unity for families with five or more members. Include the family size dummies in the model estimated from part (iii); be sure to choose a base group. Are the family dummies significant at the 1% level?

vii. Now, do a Chow test for the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r, is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is

$\mathbf{SSR}_{ur} = \mathbf{SSR}_1 + \mathbf{SSR}_2 + \ldots + \mathbf{SSR}_5$, where SSR_f is the sum of squared residuals for the equation estimated using only family size *f*. You should convince yourself that there are 30 parameters in the unrestricted model (5 intercepts plus 25 slopes) and 10 parameters in the restricted model (5 intercepts plus 5 slopes). Therefore, the number of restrictions being tested is $q = 20$, and the *df* for the unrestricted model is $9{,}275 - 30 = 9{,}245$.

C12. Use the data set in BEAUTY, which contains a subset of the variables (but more usable observations than in the regressions) reported by Hamermesh and Biddle (1994).

i. Find the separate fractions of men and women that are classified as having above average looks. Are more people rated as having above average or below average looks?

ii. Test the null hypothesis that the population fractions of above-average-looking women and men are the same. Report the one-sided *p*-value that the fraction is higher for women. (*Hint:* Estimating a simple linear probability model is easiest.)

iii. Now estimate the model

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

separately for men and women, and report the results in the usual form. In both cases, interpret the coefficient on *belavg*. Explain in words what the hypothesis $\mathbf{H_0}: \beta_1 = 0$ against $\mathbf{H_1}: \beta_1 < 0$ means, and find the *p*-values for men and women.

iv. Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.

v. For both men and women, add the explanatory variables *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity*, and *service*. Do the effects of the "looks" variables change in important ways?

vi. Use the SSR form of the Chow *F* statistic to test whether the slopes of the regression functions in part (v) differ across men and women. Be sure to allow for an intercept shift under the null.

C13. Use the data in APPLE to answer this question.

i. Define a binary variable as $ecobuy = 1$ if $ecolbs > 0$ and $ecobuy = 0$ if $ecolbs = 0$. In other words, *ecobuy* indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?

ii. Estimate the linear probability model

$$\begin{aligned} ecobuy &= \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc \\ &\quad + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u, \end{aligned}$$

and report the results in the usual form. Carefully interpret the coefficients on the price variables.

iii. Are the nonprice variables jointly significant in the LPM? (Use the usual *F* statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?

iv. In the model from part (ii), replace *faminc* with $\log(faminc)$. Which model fits the data better, using *faminc* or $\log(faminc)$? Interpret the coefficient on $\log(faminc)$.

v. In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

vi. For the estimation in part (iv), compute the percent correctly predicted for each outcome, $ecobuy = 0$ and $ecobuy = 1$. Which outcome is best predicted by the model?

C14. Use the data in CHARITY to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organization. The variable *resplast* is a dummy variable equal to one if the person responded to the previous mailing, *avggift* is the average of past gifts (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

i. Estimate a linear probability model relating *respond* to *resplast* and *avggift*. Report the results in the usual form, and interpret the coefficient on *resplast*.

ii. Does the average value of past gifts seem to affect the probability of responding?

iii. Add the variable *propresp* to the model, and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)

iv. What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?

v. Add *mailsyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?

C15. Use the data in FERTIL2 to answer this question.

i. Find the smallest and largest values of *children* in the sample. What is the average of *children*? Does any woman have exactly the average number of children?

ii. What percentage of women have electricity in the home?

iii. Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find. Test whether the population means are the same using a simple regression.

iv. From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.

v. Estimate a multiple regression model of the kind reported in equation (7.37), but add age^2, *urban*, and the three religious affiliation dummies. How does the estimated effect of having electricity compare with that in part (iii)? Is it still statistically significant?

vi. To the equation in part (v), add an interaction between *electric* and *educ*. Is its coefficient statistically significant? What happens to the coefficient on *electric*?

vii. The median and mode value for *educ* is 7. In the equation from part (vi), use the centered interaction term $electric \cdot (educ - 7)$ in place of $electric \cdot educ$. What happens to the coefficient on *electric* compared with part (vi)? Why? How does the coefficient on *electric* compare with that in part (v)?

C16. Use the data in CATHOLIC to answer this question.

i. In the entire sample, what percentage of the students attend a Catholic high school? What is the average of *math*12 in the entire sample?

ii. Run a simple regression of *math*12 on *cathhs* and report the results in the usual way. Interpret what you have found.

iii. Now add the variables *lfaminc*, *motheduc*, and *fatheduc* to the regression from part (ii). How many observations are used in the regression? What happens to the coefficient on *cathhs*, along with its statistical significance?

iv. Return to the simple regression of *math*12 on *cathhs*, but restrict the regression to observations used in the multiple regression from part (iii). Do any important conclusions change?

v. To the multiple regression in part (iii), add interactions between *cathhs* and each of the other explanatory variables. Are the interaction terms individually or jointly significant?

vi. What happens to the coefficient on *cathhs* in the regression from part (v). Explain why this coefficient is not very interesting.

vii. Compute the average partial effect of *cathhs* in the model estimated in part (v). How does it compare with the coefficients on *cathhs* in parts (iii) and (v)?