

## CHAPTER 7

### SOLUTIONS TO PROBLEMS

**7.1** (i) The coefficient on *male* is 87.75, so a man is estimated to sleep almost one and one-half hours more per week than a comparable woman. Further,  $t_{male} = 87.75/34.33 \approx 2.56$ , which is close to the 1% critical value against a two-sided alternative (about 2.58). Thus, the evidence for a gender differential is fairly strong.

(ii) The  $t$  statistic on *totwrk* is  $-.163/.018 \approx -9.06$ , which is very statistically significant. The coefficient implies that one more hour of work (60 minutes) is associated with  $.163(60) \approx 9.8$  minutes less sleep.

(iii) To obtain  $R_r^2$ , the  $R$ -squared from the restricted regression, we need to estimate the model without *age* and  $age^2$ . When *age* and  $age^2$  are both in the model, *age* has no effect only if the parameters on both terms are zero.

**7.3** (i) The  $t$  statistic on  $hsize^2$  is over four in absolute value, so there is very strong evidence that it belongs in the equation. We obtain this by finding the turnaround point; this is the value of *hsize* that maximizes  $\hat{s}at$  (other things fixed):  $19.3/(2 \cdot 2.19) \approx 4.41$ . Because *hsize* is measured in hundreds, the optimal size of graduating class is about 441.

(ii) This is given by the coefficient on *female* (since *black* = 0): nonblack females have SAT scores about 45 points lower than nonblack males. The  $t$  statistic is about  $-10.51$ , so the difference is very statistically significant. (The very large sample size certainly contributes to the statistical significance.)

(iii) Because *female* = 0, the coefficient on *black* implies that a black male has an estimated SAT score almost 170 points less than a comparable nonblack male. The  $t$  statistic is over 13 in absolute value, so we easily reject the hypothesis that there is no ceteris paribus difference.

(iv) We plug in *black* = 1, *female* = 1 for black females and *black* = 0 and *female* = 1 for nonblack females. The difference is therefore  $-169.81 + 62.31 = -107.50$ . Because the estimate depends on two coefficients, we cannot construct a  $t$  statistic from the information given. The easiest approach is to define dummy variables for three of the four race/gender categories and choose nonblack females as the base group. We can then obtain the  $t$  statistic we want as the coefficient on the black female dummy variable.

**7.5** (i) Following the hint,  $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0(1 - noPC) + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT = (\hat{\beta}_0 + \hat{\delta}_0) - \hat{\delta}_0 noPC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$ . For the specific estimates in equation (7.6),  $\hat{\beta}_0 = 1.26$  and  $\hat{\delta}_0 = .157$ , so the new intercept is  $1.26 + .157 = 1.417$ . The coefficient on *noPC* is  $-.157$ .

(ii) Nothing happens to the  $R$ -squared. Using *noPC* in place of *PC* is simply a different way of including the same information on *PC* ownership.

(iii) It makes no sense to include both dummy variables in the regression: we cannot hold *noPC* fixed while changing *PC*. We have only two groups based on *PC* ownership so, in addition to the overall intercept, we need only to include one dummy variable. If we try to include both along with an intercept we have perfect multicollinearity (the dummy variable trap).

**7.7** (i) Write the population model underlying (7.29) as

$$\begin{aligned} \text{inlf} = & \beta_0 + \beta_1 \text{nwifeinc} + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{exper}^2 + \beta_5 \text{age} \\ & + \beta_6 \text{kidslt6} + \beta_7 \text{kidsage6} + u, \end{aligned}$$

plug in  $\text{inlf} = 1 - \text{outlf}$ , and rearrange:

$$\begin{aligned} 1 - \text{outlf} = & \beta_0 + \beta_1 \text{nwifeinc} + \beta_2 \text{educ} + \beta_3 \text{exper} + \beta_4 \text{exper}^2 + \beta_5 \text{age} \\ & + \beta_6 \text{kidslt6} + \beta_7 \text{kidsage6} + u, \end{aligned}$$

or

$$\begin{aligned} \text{outlf} = & (1 - \beta_0) - \beta_1 \text{nwifeinc} - \beta_2 \text{educ} - \beta_3 \text{exper} - \beta_4 \text{exper}^2 - \beta_5 \text{age} \\ & - \beta_6 \text{kidslt6} - \beta_7 \text{kidsage6} - u, \end{aligned}$$

The new error term,  $-u$ , has the same properties as  $u$ . From this we see that if we regress *outlf* on all of the independent variables in (7.29), the new intercept is  $1 - .586 = .414$  and each slope coefficient takes on the opposite sign from when *inlf* is the dependent variable. For example, the new coefficient on *educ* is  $-.038$  while the new coefficient on *kidslt6* is  $.262$ .

(ii) The standard errors will not change. In the case of the slopes, changing the signs of the estimators does not change their variances, and therefore the standard errors are unchanged (but the  $t$  statistics change sign). Also,  $\text{Var}(1 - \hat{\beta}_0) = \text{Var}(\hat{\beta}_0)$ , so the standard error of the intercept is the same as before.

(iii) We know that changing the units of measurement of independent variables, or entering qualitative information using different sets of dummy variables, does not change the  $R$ -squared. But here we are changing the *dependent* variable. Nevertheless, the  $R$ -squareds from the regressions are still the same. To see this, part (i) suggests that the squared residuals will be identical in the two regressions. For each  $i$  the error in the equation for  $\text{outlf}_i$  is just the negative of the error in the other equation for  $\text{inlf}_i$ , and the same is true of the residuals. Therefore, the SSRs are the same. Further, in this case, the total sum of squares are the same. For *outlf* we have

$$\text{SST} = \sum_{i=1}^n (\text{outlf}_i - \overline{\text{outlf}})^2 = \sum_{i=1}^n [(1 - \text{inlf}_i) - (1 - \overline{\text{inlf}})]^2 = \sum_{i=1}^n (-\text{inlf}_i + \overline{\text{inlf}})^2 = \sum_{i=1}^n (\text{inlf}_i - \overline{\text{inlf}})^2,$$

which is the SST for *lnlf*. Because  $R^2 = 1 - SSR/SST$ , the  $R$ -squared is the same in the two regressions.

**7.9** (i) Plugging in  $u = 0$  and  $d = 1$  gives  $f_1(z) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z$ .

(ii) Setting  $f_0(z^*) = f_1(z^*)$  gives  $\beta_0 + \beta_1 z^* = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)z^*$  or  $0 = \delta_0 + \delta_1 z^*$ .

Therefore, provided  $\delta_1 \neq 0$ , we have  $z^* = -\delta_0 / \delta_1$ . Clearly,  $z^*$  is positive if and only if  $\delta_0 / \delta_1$  is negative, which means  $\delta_0$  and  $\delta_1$  must have opposite signs.

(iii) Using part (ii) we have  $totcoll^* = .357 / .030 = 11.9$  years.

(iv) The estimated years of college where women catch up to men is much too high to be practically relevant. While the estimated coefficient on *female · totcoll* shows that the gap is reduced at higher levels of college, it is never closed – not even close. In fact, at four years of college, the difference in predicted log wage is still  $-.357 + .030(4) = -.237$ , or about 21.1% less for women.

## SOLUTIONS TO COMPUTER EXERCISES

**C7.1** (i) The estimated equation is

$$\begin{aligned} \widehat{colGPA} = & 1.26 + .152 PC + .450 hsGPA + .0077 ACT - .0038 mothcoll \\ & (0.34) \quad (.059) \quad (.094) \quad (.0107) \quad (.0603) \\ & + .0418 fathcoll \\ & \quad (.0613) \\ n = & 141, \quad R^2 = .222. \end{aligned}$$

The estimated effect of *PC* is hardly changed from equation (7.6), and it is still very significant, with  $t_{pc} \approx 2.58$ .

(ii) The  $F$  test for joint significance of *mothcoll* and *fathcoll*, with 2 and 135 *df*, is about .24 with  $p$ -value  $\approx .78$ ; these variables are jointly very insignificant. It is not surprising the estimates on the other coefficients do not change much when *mothcoll* and *fathcoll* are added to the regression.

(iii) When  $hsGPA^2$  is added to the regression, its coefficient is about .337 and its  $t$  statistic is about 1.56. (The coefficient on *hsGPA* is about  $-1.803$ .) This is a borderline case. The quadratic in *hsGPA* has a U-shape, and it only turns up at about  $hsGPA^* = 2.68$ , which is hard to interpret. The coefficient of main interest, on *PC*, falls to about .140 but is still significant. Adding  $hsGPA^2$  is a simple robustness check of the main finding.

**C7.3** (i)  $H_0: \beta_{13} = 0$ . Using the data in `MLB1.RAW` gives  $\hat{\beta}_{13} \approx .254$ ,  $se(\hat{\beta}_{13}) \approx .131$ . The  $t$  statistic is about 1.94, which gives a  $p$ -value against a two-sided alternative of just over .05. Therefore, we would reject  $H_0$  at just about the 5% significance level. Controlling for the performance and experience variables, the estimated salary differential between catchers and outfielders is huge, on the order of  $100 \cdot [\exp(.254) - 1] \approx 28.9\%$  [using equation (7.10)].

(ii) This is a joint null,  $H_0: \beta_9 = 0, \beta_{10} = 0, \dots, \beta_{13} = 0$ . The  $F$  statistic, with 5 and 339  $df$ , is about 1.78, and its  $p$ -value is about .117. Thus, we cannot reject  $H_0$  at the 10% level.

(iii) Parts (i) and (ii) are roughly consistent. The evidence against the joint null in part (ii) is weaker because we are testing, along with the marginally significant *catcher*, several other insignificant variables (especially *thrdbase* and *shrtstop*, which has absolute  $t$  statistics well below one).

**C7.5** The estimated equation is

$$\widehat{\log(\text{salary})} = 4.30 + .288 \log(\text{sales}) + .0167 \text{roe} - .226 \text{rosneg}$$

$$(0.29) \quad (.034) \quad (.0040) \quad (.109)$$

$$n = 209, \quad R^2 = .297, \quad \bar{R}^2 = .286.$$

The coefficient on *rosneg* implies that if the CEO's firm had a negative return on its stock over the 1988 to 1990 period, the CEO salary was predicted to be about 22.6% lower, for given levels of *sales* and *roe*. The  $t$  statistic is about  $-2.07$ , which is significant at the 5% level against a two-sided alternative.

**C7.7** (i) When  $\text{educ} = 12.5$ , the approximate proportionate difference in estimated *wage* between women and men is  $-.227 - .0056(12.5) = -.297$ . When  $\text{educ} = 0$ , the difference is  $-.227$ . So the differential at 12.5 years of education is about 7 percentage points greater.

(ii) We can write the model underlying (7.18) as

$$\begin{aligned} \log(\text{wage}) &= \beta_0 + \delta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot \text{educ} + \text{other factors} \\ &= \beta_0 + (\delta_0 + 12.5 \delta_1) \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot (\text{educ} - 12.5) \\ &\quad + \text{other factors} \\ &\equiv \beta_0 + \theta_0 \text{female} + \beta_1 \text{educ} + \delta_1 \text{female} \cdot (\text{educ} - 12.5) + \text{other factors}, \end{aligned}$$

where  $\theta_0 \equiv \delta_0 + 12.5 \delta_1$  is the gender differential at 12.5 years of education. When we run this regression we obtain about  $-.294$  as the coefficient on *female* (which differs from  $-.297$  due to rounding error). Its standard error is about  $.036$ .

(iii) The  $t$  statistic on *female* from part (ii) is about  $-8.17$ , which is very significant. This is because we are estimating the gender differential at a reasonable number of years of education, 12.5, which is close to the average. In equation (7.18), the coefficient on *female* is the gender differential when  $educ = 0$ . There are no people of either gender with close to zero years of education, and so we cannot hope – nor do we want to – to estimate the gender differential at  $educ = 0$ .

**C7.9** (i) About  $.392$ , or  $39.2\%$ .

(ii) The estimated equation is

$$\widehat{e401k} = -.506 + .0124 inc - .000062 inc^2 + .0265 age - .00031 age^2 - .0035 male$$

$$(.081) \quad (.0006) \quad (.000005) \quad (.0039) \quad (.00005) \quad (.0121)$$

$$n = 9,275, \quad R^2 = .094.$$

(iii) 401(k) eligibility clearly depends on income and age in part (ii). Each of the four terms involving *inc* and *age* have very significant  $t$  statistics. On the other hand, once income and age are controlled for, there seems to be no difference in eligibility by gender. The coefficient on *male* is very small – at given income and age, males are estimated to have a  $.0035$  lower probability of being 401(k) eligible – and it has a very small  $t$  statistic.

(iv) Somewhat surprisingly, out of 9,275 fitted values, none is outside the interval  $[0,1]$ . The smallest fitted value is about  $.030$  and the largest is about  $.697$ . This means one theoretical problem with the LPM – the possibility of generating silly probability estimates – does not materialize in this application.

(v) Using the given rule, 2,460 families are predicted to be eligible for a 401(k) plan.

(vi) Of the 5,638 families actually ineligible for a 401(k) plan, about 81.7 are correctly predicted not to be eligible. Of the 3,637 families actually eligible, only 39.3 percent are correctly predicted to be eligible.

(vii) The overall percent correctly predicted is a weighted average of the two percentages obtained in part (vi). As we saw there, the model does a good job of predicting when a family is ineligible. Unfortunately, it does less well – predicting correctly less than 40% of the time – in predicting that a family is eligible for a 401(k).

(viii) The estimated equation is

$$\widehat{e401k} = -.502 + .0123 inc - .000061 inc^2 + .0265 age - .00031 age^2$$

$$\begin{array}{cccccc}
 (.081) & (.0006) & (.000005) & (.0039) & (.00005) & \\
 - .0038 \textit{ male} & + .0198 \textit{ pira} & & & & \\
 (.0121) & (.0122) & & & & 
 \end{array}$$

$$n = 9,275, R^2 = .095.$$

The coefficient on *pira* means that, other things equal, IRA ownership is associated with about a .02 higher probability of being eligible for a 401(k) plan. However, the *t* statistic is only about 1.62, which gives a two-sided *p*-value = .105. So *pira* is not significant at the 10% level against a two-sided alternative.

**C7.11** (i) The average is 19.072, the standard deviation is 63.964, the smallest value is -502.302, and the largest value is 1,536.798. Remember, these are in thousands of dollars.

(ii) This can be easily done by regressing *nettfa* on *e401k* and doing a *t* test on  $\hat{\beta}_{e401k}$ ; the estimate is the average difference in *nettfa* for those eligible for a 401(k) and those not eligible. Using the 9,275 observations gives  $\hat{\beta}_{e401k} = 18.858$  and  $t_{e401k} = 14.01$ . Therefore, we strongly reject the null hypothesis that there is no difference in the averages. The coefficient implies that, on average, a family eligible for a 401(k) plan has \$18,858 more on net total financial assets.

(iii) The equation estimated by OLS is

$$\widehat{\textit{nettfa}} = 23.09 + 9.705 \textit{ e401k} - .278 \textit{ inc} + .0103 \textit{ inc}^2 - 1.972 \textit{ age} + .0348 \textit{ age}^2$$

$$\begin{array}{cccccc}
 (9.96) & (1.277) & (.075) & (.0006) & (.483) & (.0055)
 \end{array}$$

$$n = 9,275, R^2 = .202$$

Now, holding income and age fixed, a 401(k)-eligible family is estimated to have \$9,705 more in wealth than a non-eligible family. This is just more than half of what is obtained by simply comparing averages.

(iv) Only the interaction *e401k*·(*age* - 41) is significant. Its coefficient is .654 (*t* = 4.98). It shows that the effect of 401(k) eligibility on financial wealth increases with age. Another way to think about it is that *age* has a stronger positive effect on *nettfa* for those with 401(k) eligibility. The coefficient on *e401k*·(*age* - 41)<sup>2</sup> is -.0038 (*t* statistic = -.33), so we could drop this term.

(v) The effect of *e401k* in part (iii) is the same for all ages, 9.705. For the regression in part (iv), the coefficient on *e401k* from part (iv) is about 9.960, which is the effect at the average age, *age* = 41. Including the interactions increases the estimated effect of *e401k*, but only by \$255. If we evaluate the effect in part (iv) at a wide range of ages, we would see more dramatic differences.

(vi) I chose *fsize1* as the base group. The estimated equation is



(iv) The model with  $\log(\text{faminc})$  fits the data slightly better: the  $R$ -squared increases to about .112. (We would not expect a large increase in  $R$ -squared from a simple change in the functional form.) The coefficient on  $\log(\text{faminc})$  is about .045 ( $t = 1.55$ ). If  $\log(\text{faminc})$  increases by .10, which means roughly a 10% increase in  $\text{faminc}$ , then  $P(\text{ecobuy} = 1)$  is estimated to increase by about .0045, a pretty small effect.

(v) The fitted probabilities range from about .185 to 1.051, so none are negative. There are two fitted probabilities above 1, which is not a source of concern with 660 observations.

(vi) Using the standard prediction rule – predict one when  $\widehat{\text{ecobuy}}_i \geq .5$  and zero otherwise – gives the fraction correctly predicted for  $\text{ecobuy} = 0$  as  $102/248 \approx .411$ , so about 41.1%. For  $\text{ecobuy} = 1$ , the fraction correctly predicted is  $340/412 \approx .825$ , or 82.5%. With the usual prediction rule, the model does a much better job predicting the decision to buy eco-labeled apples. (The overall percent correctly predicted is about 67%.)