Solution for the suggested questions

Chapter 7

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 ≈ 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 ≈ −9.06, which is very statistically significant. The coefficient implies that one more hour of work (60 minutes) is associated with .163(60) ≈ 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.

7.4 (i) The approximate difference is just the coefficient on *utility* times 100, or –28.3%. The *t* statistic is −.283/.099 ≈ −2.86, which is very statistically significant.

(ii) 100 · [exp(−.283) – 1] ≈ −24.7%, and so the estimate is somewhat smaller in magnitude.

(iii) The proportionate difference is .181 − .158 = .023, or about 2.3%. One equation that can be estimated to obtain the standard error of this difference is

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \delta_1 consprod + \delta_2 utility + \delta_3 trans + u,$$

where *trans* is a dummy variable for the transportation industry. Now, the base group is *finance*, and so the coefficient δ_1 directly measures the difference between the consumer products and finance industries, and we can use the *t* statistic on *consprod*.

7.10 (i) Yes, simple regression does produce an unbiased estimator of the effect of the voucher program. Because participation was randomized, we can write

$$score = \beta_0 + \beta_1 voucher + u,$$

where *voucher* is independent of *u*, that is, all other factors affecting *score*. Therefore, the key assumption for unbiasedness of simple regression, Assumption SLR.3, is satisfied.

(ii) No, we do not need to control for background variables. In the equation from part (i), these are factors in the error term, *u*. But *voucher* was assigned to be independent of all factors, including the listed background variables.

(iii) We should include the background variables to reduce the sampling error of the estimated voucher effect. By pulling background variables out of the error term, we reduce the error variance – perhaps substantially. Further, we can be sure that multicollinearity is not a problem because the key variable of interest, *voucher*, is uncorrelated with all of the added explanatory variables. (This zero correlation will only be approximate in any random sample, but in large samples it should be very small.) The one case where we would not add these variables – or, at least, when there is no benefit from doing so – is when the background variables themselves have no affect on the test score. Given the list of background variables, this seems unlikely in the current application.

Chapter 7 (computer problems)

C7.4 (i) The two signs that are pretty clear are $\beta_3 < 0$ (because *hsperc* is defined so that the smaller the number the better the student) and $\beta_4 > 0$. The effect of size of graduating class is not clear. It is also unclear whether males and females have systematically different GPAs. We may think that $\beta_6 < 0$, that is, athletes do worse than other students with comparable characteristics. But remember, we are controlling for ability to some degree with *hsperc* and *sat*.

(ii) The estimated equation is

$$\begin{aligned} \widehat{colgpa} = {} & 1.241 - .0569\, hsize + .00468\, hsize^2 - .0132\, hsperc \\ & (0.079)\ (.0164)\qquad (.00225)\qquad (.0006) \\ & + .00165\, sat + .155\, female + .169\, athlete \\ & \quad (.00007)\qquad (.018)\qquad (.042) \end{aligned}$$

$n = 4{,}137, \quad R^2 = .293.$

Holding other factors fixed, an athlete is predicted to have a GPA about .169 points *higher* than a nonathlete. The *t* statistic .169/.042 $\approx$ 4.02, which is very significant.

(iii) With *sat* dropped from the model, the coefficient on *athlete* becomes about .0054 (se $\approx$ .0448), which is practically and statistically not different from zero. This happens because we do not control for SAT scores, and athletes score lower on average than nonathletes. Part (ii) shows that, once we account for SAT differences, athletes do better than nonathletes. Even if we do not control for SAT score, there is no difference.

(iv) To facilitate testing the hypothesis that there is no difference between women athletes and women nonathletes, we should choose one of these as the base group. We choose female nonathletes. The estimated equation is

$$\underset{}{colgpa} = \underset{(0.076)}{1.396} - \underset{(.0164)}{.0568\, hsize} + \underset{(.00225)}{.00467\, hsize^2} - \underset{(.0006)}{.0132\, hsperc}$$

$$+ \underset{(.00007)}{.00165\, sat} + \underset{(.084)}{.175\, femath} + \underset{(.049)}{.013\, maleath} - \underset{(.018)}{.155\, malenonath}$$

$n = 4{,}137, \quad R^2 = .293.$

The coefficient on *femath* = *female*·*athlete* shows that *colgpa* is predicted to be about .175 points higher for a female athlete than a female nonathlete, other variables in the equation fixed. The hypothesis that there is no difference between female athletes and female nonathletes is tested by using the *t* statistic on *femath*. In this case, *t* = 2.08, which is statistically significant at the 5% level against a two-sided alternative.

(v) Whether we add the interaction *female*·*sat* to the equation in part (ii) or part (iv), the outcome is practically the same. For example, when *female*·*sat* is added to the equation in part (ii), its coefficient is about .000051 and its *t* statistic is about .40. There is very little evidence that the effect of *sat* differs by gender.

C7.12 (i) For women, the fraction rated as having above average looks is about .33; for men, it is .29. The proportion of women rated as having below average looks is only .135; for men, it is even lower at about .117.

(ii) The difference is about .04, that is, the percent rated as having above average looks is about four percentage points higher for women than men. A simple way to test whether the difference is statistically significant is to run a simple regression of *abvavg* on *female* and do a *t* test (which is asymptotically valid). The *t* statistic is about 1.48 with two-sided *p*-value = .14. Therefore, there is not strong evidence against the null that the population fractions are the same, but there is some evidence.

(iii) The regression for men is

$$\underset{}{\log(wage)} = \underset{(0.024)}{1.884} - \underset{(.060)}{.199\, belavg} - \underset{(.042)}{.044\, abvavg}$$

$n = 824 \quad R^2 = .013.$

and the regression for women is

$$\underset{}{\log(wage)} = \underset{(0.034)}{1.309} - \underset{(.076)}{.138\, belavg} + \underset{(.055)}{.034\, abvavg}$$

$n = 436 \quad R^2 = .011.$

Using the standard approximation, a man with below average looks earns almost 20% less than a man of average looks, and a woman with below average looks earns about 13.8% less than a woman with average looks. (The more accurate estimates are about 18% and 12.9%, respectively.) The null

hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ means that the null is that people with below average looks earn the same, on average, as people with average looks; the alternative is that people with below average looks earn less than people with average looks (in the population). The one-sided *p*-value for men is .0005, and for women, it is .036. We reject H_0 more strongly for men because the estimate is larger in magnitude and the estimate has less sampling variation (as measured by the standard error).

(iv) Women with above average looks are estimated to earn about 3.4% more, on average, than women with average looks. But the one-sided *p*-value is .272, and this provides very little evidence against H_0: $\beta_2 = 0$.

(v) Given the number of added controls, with many of them very statistically significant, the coefficients on the looks variables do not change by much. For men, the coefficient on *belavg* becomes −.143 (*t* = −2.80) and the coefficient on *abvavg* becomes −.001 (*t* = −.03). For women, the changes in magnitude are similar: the coefficient on *belavg* becomes −.115 (*t* = −1.75) and the coefficient on *abvavg* becomes .058 (*t* = 1.18). In both cases, the estimates on *belavg* move closer to zero but are still reasonably large.

(vi) The SSR for women is 83.6933. For men, it is 166.0841, and so the unrestricted SSR is 249.7774 (rounded to four decimal places). The SSR obtained by adding the dummy variable *female* to the regression in part (v) is 261.1771. Therefore, the *F* statistic (Chow statistic) is

$$F = \frac{(261.1771 - 249.7774)}{249.7774} \cdot \frac{(1260 - 2 \cdot 14)}{13} \approx 4.33.$$

With 13 and 1,232 *df,* the *p*-value is essentially zero. So we strongly reject the common slopes formulation even if we allow for a different intercept. A look at the estimated regression functions for men and women shows that some of the slopes are quite different.

Chapter 8

8.1 Parts (ii) and (iii). The homoskedasticity assumption played no role in Chapter 5 in showing that OLS is consistent. But we know that heteroskedasticity causes statistical inference based on the usual *t* and *F* statistics to be invalid, even in large samples. As heteroskedasticity is a violation of the Gauss-Markov assumptions, OLS is no longer BLUE.

8.2 Var(*u*|*inc,price,educ,female*) = $\sigma^2 inc^2$, $h(\mathbf{x}) = inc^2$, where $h(\mathbf{x})$ is the heteroskedasticity function defined in equation (8.21). Therefore, $\sqrt{h(\mathbf{x})} = inc$, and so the transformed equation is obtained by dividing the original equation by *inc*:

$$\frac{beer}{inc} = \beta_0(1/inc) + \beta_1 + \beta_2(price/inc) + \beta_3(educ/inc) + \beta_4(female/inc) + (u/inc).$$

Notice that β_1, which is the slope on *inc* in the original model, is now a constant in the transformed equation. This is simply a consequence of the form of the heteroskedasticity and the functional forms of the explanatory variables in the original equation.

8.4 (i) These coefficients have the anticipated signs. If a student takes courses where grades are, on average, higher – as reflected by higher *crsgpa* – then his/her grades will be higher. The better the student has been in the past – as measured by *cumgpa* – the better the student does (on average) in the current semester. Finally, *tothrs* is a measure of experience, and its coefficient indicates an increasing return to experience.

The *t* statistic for *crsgpa* is very large, over five using the usual standard error (which is the largest of the two). Using the robust standard error for *cumgpa*, its *t* statistic is about 2.61, which is also significant at the 5% level. The *t* statistic for *tothrs* is only about 1.17 using either standard error, so it is not significant at the 5% level.

(ii) This is easiest to see without other explanatory variables in the model. If *crsgpa* were the only explanatory variable, H_0: $\beta_{crsgpa} = 1$ means that, without any information about the student, the best predictor of term GPA is the average GPA in the students' courses; this holds essentially by definition. (The intercept would be zero in this case.) With additional explanatory variables, it is not necessarily true that $\beta_{crsgpa} = 1$ because *crsgpa* could be correlated with characteristics of the student. (For example, perhaps the courses students take are influenced by ability – as measured by test scores – and past college performance.) But it is still interesting to test this hypothesis.

The *t* statistic using the usual standard error is $t = (.900 - 1)/.175 \approx -.57$; using the hetero-skedasticity-robust standard error gives $t \approx -.60$. In either case we fail to reject H_0: $\beta_{crsgpa} = 1$ at any reasonable significance level, certainly including 5%.

(iii) The in-season effect is given by the coefficient on *season*, which implies that, other things equal, an athlete's GPA is about .16 points lower when his/her sport is competing. The *t* statistic using the usual standard error is about –1.60, while that using the robust standard error is about –1.96. Against a two-sided alternative, the *t* statistic using the robust standard error is just significant at the 5% level (the standard normal critical value is 1.96), while using the usual standard error, the *t* statistic is not quite significant at the 10% level ($cv \approx 1.65$). So the standard error used makes a difference in this case. This example is somewhat unusual, as the robust standard error is more often the larger of the two.

Chapter 8 (computer problems)

C8.2 (i) The estimated equation with both sets of standard errors (heteroskedasticity-robust standard errors in parentheses) is

$$\begin{aligned} \widehat{price} = {} & -21.77 + .00207\, lotsize + .123\, sqrft + 13.85\, bdrms \\ & \quad (29.48)(.00064) \quad (.013) \quad (9.01) \\ & \quad [36.28][.00122] \quad [.017] \quad [8.28] \end{aligned}$$

$n = 88, \; R^2 = .672.$

The robust standard error on *lotsize* is almost twice as large as the usual standard error, making *lotsize* much less significant (the *t* statistic falls from about 3.23 to 1.70). The *t* statistic on *sqrft* also falls, but it

is still very significant. The variable *bdrms* actually becomes somewhat more significant, but it is still barely significant. The most important change is in the significance of *lotsize*.

(ii) For the log-log model,

$$\begin{array}{lcccc} \log(price) & = & -1.30 + .168\log(lotsize) & + .700\log(sqrft) & + .037\,bdrms \\ & & (0.65)\ \ (.038) & (.093) & (.028) \\ & & [0.76]\ \ [.041] & [.101] & [.030] \end{array}$$

$n = 88$, $R^2 = .643$.

Here, the heteroskedasticity-robust standard error is always slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, log(*lotsize*) and log(*sqrft*) still have very large *t* statistics, and the *t* statistic on *bdrms* is not significant at the 5% level against a one-sided alternative using either standard error.

(iii) As we discussed in Section 6.2, using the logarithmic transformation of the dependent variable often mitigates, if not entirely eliminates, heteroskedasticity. This is certainly the case here, as no important conclusions in the model for log(*price*) depend on the choice of standard error. (We have also transformed two of the independent variables to make the model of the constant elasticity variety in *lotsize* and *sqrft*.)

Chapter 12

12.3 (i) Because U.S. presidential elections occur only every four years, it seems reasonable to think that the unobserved shocks – that is, elements in u_t – in one election have pretty much dissipated four years later. This would imply that $\{u_t\}$ is roughly serially uncorrelated.

(ii) The *t* statistic for H_0: $\rho = 0$ is $-.068/.240 \approx -.28$, which is very small. Further, the estimate $\hat{\rho} = -.068$ is small in a practical sense, too. There is no reason to worry about serial correlation in this example.

(iii) Because the test based on $t_{\hat{\rho}}$ is only justified asymptotically, we would generally be concerned about using the usual critical values with $n = 20$ in the original regression. But any kind of adjustment, either to obtain valid standard errors for OLS as in Section 12.5 or a feasible GLS procedure as in Section 12.3, relies on large sample sizes, too. (Remember, FGLS is not even unbiased, whereas OLS is under TS.1 through TS.3.) Most importantly, the estimate of ρ is *practically* small, too. With $\hat{\rho}$ so close to zero, FGLS or adjusting the standard errors would yield similar results to OLS with the usual standard errors.