CHAPTER 8

SOLUTIONS TO PROBLEMS

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.

8.3 False. The unbiasedness of WLS and OLS hinges crucially on Assumption MLR.4, and, as we know from Chapter 4, this assumption is often violated when an important variable is omitted. When MLR.4 does not hold, both WLS and OLS are biased. Without specific information on how the omitted variable is correlated with the included explanatory variables, it is not possible to determine which estimator has a small bias. It is possible that WLS would have more bias than OLS or less bias. Because we cannot know, we should not claim to use WLS in order to solve "biases" associated with OLS.

8.5 (i) No. For each coefficient, the usual standard errors and the heteroskedasticity-robust ones are practically very similar.

(ii) The effect is –.029(4) = –.116, so the probability of smoking falls by about .116.

(iii) As usual, we compute the turning point in the quadratic: .020/[2(.00026)] $\approx$ 38.46, so about 38 and one-half years.

(iv) Holding other factors in the equation fixed, a person in a state with restaurant smoking restrictions has a .101 lower chance of smoking. This is similar to the effect of having four more years of education.

(v) We just plug the values of the independent variables into the OLS regression line:

$$\widehat{smokes} = .656 - .069 \cdot \log(67.44) + .012 \cdot \log(6,500) - .029(16) + .020(77) - .00026(77^2) \approx .0052.$$

Thus, the estimated probability of smoking for this person is close to zero. (In fact, this person is not a smoker, so the equation predicts well for this particular observation.)

8.7 (i) This follows from the simple fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances: $\text{Var}(f_i + v_{i,e}) = \text{Var}(f_i) + \text{Var}(v_{i,e}) = \sigma_f^2 + \sigma_v^2$.

(ii) We compute the covariance between any two of the composite errors as

$$\begin{aligned}\text{Cov}(u_{i,e}, u_{i,g}) &= \text{Cov}(f_i + v_{i,e},\, f_i + v_{i,g}) = \text{Cov}(f_i,\, f_i) + \text{Cov}(f_i, v_{i,g}) + \text{Cov}(v_{i,e},\, f_i) + \text{Cov}(v_{i,e}, v_{i,g}) \\ &= \text{Var}(f_i) + 0 + 0 + 0 = \sigma_f^2,\end{aligned}$$

where we use the fact that the covariance of a random variable with itself is its variance and the assumptions that $f_i, v_{i,e}$, and $v_{i,g}$ are pairwise uncorrelated.

(iii) This is most easily solved by writing

$$m_i^{-1}\sum_{e=1}^{m_i} u_{i,e} = m_i^{-1}\sum_{e=1}^{m_i}(f_i + u_{i,e}) = f_i + m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}.$$

Now, by assumption, f_i is uncorrelated with each term in the last sum; therefore, f_i is uncorrelated with $m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}$. It follows that

$$\begin{aligned}\text{Var}\left(f_i + m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}\right) &= \text{Var}(f_i) + \text{Var}\left(m_i^{-1}\sum_{e=1}^{m_i} v_{i,e}\right)\\ &= \sigma_f^2 + \sigma_v^2/m_i,\end{aligned}$$

where we use the fact that the variance of an average of m_i uncorrelated random variables with common variance (σ_v^2 in this case) is simply the common variance divided by m_i – the usual formula for a sample average from a random sample.

(iv) The standard weighting ignores the variance of the firm effect, σ_f^2. Thus, the (incorrect) weight function used is $1/h_i = m_i$. A valid weighting function is obtained by writing the variance from (iii) as $\text{Var}(\bar{u}_i) = \sigma_f^2[1 + (\sigma_v^2/\sigma_f^2)/m_i] = \sigma_f^2 h_i$. But obtaining the proper weights requires us to know (or be able to estimate) the ratio σ_v^2/σ_f^2. Estimation is possible, but we do not discuss that here. In any event, the usual weight is incorrect. When the m_i are large or the ratio σ_v^2/σ_f^2 is small – so that the firm effect is more important than the individual-specific effect – the correct weights are close to being constant. Thus, attaching large weights to large firms may be quite inappropriate.

SOLUTIONS TO COMPUTER EXERCISES

C8.1 (i) Given the equation

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + \beta_6 male + u,$$

the assumption that the variance of *u* given all explanatory variables depends only on gender is

$$Var(u \mid totwrk, educ, age, yngkid, male) = Var(u \mid male) = \delta_0 + \delta_1 male$$

Then the variance for women is simply δ_0 and that for men is $\delta_0 + \delta_1$; the difference in variances is δ_1.

(ii) After estimating the above equation by OLS, we regress $\hat{u}_i^2$ on $male_i$, $i = 1,2, \ldots,706$ (including, of course, an intercept). We can write the results as

$$\hat{u}^2 = \underset{(20,546.4)}{189,359.2} - \underset{(27,296.5)}{28,849.6}\, male + residual$$

$$n = 706, \quad R^2 = .0016.$$

Because the coefficient on *male* is negative, the estimated variance is higher for women.

(iii) No. The *t* statistic on *male* is only about –1.06, which is not significant at even the 20% level against a two-sided alternative.

C8.3 After estimating equation (8.18), we obtain the squared OLS residuals $\hat{u}^2$. The full-blown White test is based on the *R*-squared from the auxiliary regression (with an intercept),

$$\hat{u}^2 \text{ on } llotsize,\ lsqrft,\ bdrms,\ llotsize^2,\ lsqrft^2,\ bdrms^2,$$
$$llotsize \cdot lsqrft,\ llotsize \cdot bdrms, \text{ and } lsqrft \cdot bdrms,$$

where "*l* " in front of *lotsize* and *sqrft* denotes the natural log. [See equation (8.19).] With 88 observations the *n*-*R*-squared version of the White statistic is $88(.109) \approx 9.59$, and this is the outcome of an (approximately) χ^2_9 random variable. The *p*-value is about .385, which provides little evidence against the homoskedasticity assumption.

C8.5 (i) By regressing *sprdcvr* on an intercept only we obtain $\hat{\mu} \approx .515$ se $\approx .021$). The asymptotic *t* statistic for H_0: $\mu = .5$ is $(.515 - .5)/.021 \approx .71$, which is not significant at the 10% level, or even the 20% level.

(ii) 35 games were played on a neutral court.

(iii) The estimated LPM is

$$\widehat{sprdcvr} = \underset{(.045)}{.490} + \underset{(.050)}{.035}\, favhome + \underset{(.095)}{.118}\, neutral - \underset{(.050)}{.023}\, fav25 + \underset{(.092)}{.018}\, und25$$

$$n = 553, \quad R^2 = .0034.$$

The variable *neutral* has by far the largest effect – if the game is played on a neutral court, the probability that the spread is covered is estimated to be about .12 higher – and, except for the intercept, its *t* statistic is the only *t* statistic greater than one in absolute value (about 1.24).

(iv) Under H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, the response probability does not depend on any explanatory variables, which means neither the mean nor the variance depends on the explanatory variables. [See equation (8.38).]

(v) The *F* statistic for joint significance, with 4 and 548 *df*, is about .47 with *p*-value $\approx .76$. There is essentially no evidence against H_0.

(vi) Based on these variables, it is not possible to predict whether the spread will be covered. The explanatory power is very low, and the explanatory variables are jointly very insignificant. The coefficient on *neutral* may indicate something is going on with games played on a neutral court, but we would not want to bet money on it unless it could be confirmed with a separate, larger sample.

C8.7 (i) The heteroskedasticity-robust standard error for $\hat{\beta}_{white} \approx .129$ is about .026, which is notably higher than the nonrobust standard error (about .020). The heteroskedasticity-robust 95% confidence interval is about .078 to .179, while the nonrobust CI is, of course, narrower, about .090 to .168. The robust CI still excludes the value zero by some margin.

(ii) There are no fitted values less than zero, but there are 231 greater than one. Unless we do something to those fitted values, we cannot directly apply WLS, as $\hat{h}_i$ will be negative in 231 cases.

C8.9 (i) I now get $R^2 = .0527$, but the other estimates seem okay.

(ii) One way to ensure that the unweighted residuals are being provided is to compare them with the OLS residuals. They will not be the same, of course, but they should not be wildly different.

(iii) The *R*-squared from the regression $\breve{u}_i^2$ on $\breve{y}_i$, $\breve{y}_i^2$, $i = 1,...,807$ is about .027. We use this as $R^2_{\hat{u}^2}$ in equation (8.15) but with $k = 2$. This gives $F = 11.15$, and so the *p*-value is essentially zero.

(iv) The substantial heteroskedasticity found in part (iii) shows that the feasible GLS procedure described on page 279 does not, in fact, eliminate the heteroskedasticity. Therefore, the usual standard errors, *t* statistics, and *F* statistics reported with weighted least squares are not valid, even asymptotically.

(v) Weighted least squares estimation with robust standard errors gives

$$\widehat{cigs} = \underset{(37.31)}{5.64} + \underset{(.54)}{1.30}\log(income) - \underset{(8.97)}{2.94}\log(cigpric) - \underset{(.149)}{.463}\,educ$$

$$+ \underset{(.115)}{.482}\,age - \underset{(.0012)}{.0056}\,age^2 - \underset{(.72)}{3.46}\,restaurn$$

$n = 807$, $R^2 = .1134$

The substantial differences in standard errors compared with equation (8.36) further indicate that our proposed correction for heteroskedasticity did not fully solve the heteroskedasticity problem. With the exception of *restaurn*, all standard errors got notably bigger; for example, the standard error for log(*cigpric*) doubled. All variables that were statistically significant with the nonrobust standard errors remain significant, but the confidence intervals are much wider in several cases.

C8.11 (i) The usual OLS standard errors are in (·), the heteroskedasticity-robust standard errors are in [·]:

$$\widehat{nettfa} = -17.20 + .628\,inc + .0251\,(age - 25)^2 + 2.54\,male$$
$$(2.82) \quad (.080) \quad (.0026) \quad (2.04)$$
$$[3.23] \quad [.098] \quad [.0044] \quad [2.06]$$

$$- 3.83\,e401k + .343\,e401k \cdot inc$$
$$(4.40) \quad (.124)$$
$$[6.25] \quad [.220]$$

$n = 2{,}017$, $R^2 = .131$

Although the usual OLS t statistic on the interaction term is about 2.8, the heteroskedasticity-robust t statistic is just under 1.6. Therefore, using OLS, we must conclude the interaction term is only marginally significant. But the coefficient is nontrivial: it implies a much more sensitive relationship between financial wealth and income for those eligible for a 401(k) plan.

(ii) The WLS estimates, with usual WLS standard errors in (·) and the robust ones in [·], are

$$\widehat{nettfa} = -14.09 + .619\,inc + .0175\,(age - 25)^2 + 1.78\,male$$
$$(2.27) \quad (.084) \quad (.0019) \quad (1.56)$$
$$[2.53] \quad [.091] \quad [.0026] \quad [1.31]$$

$$- 2.17\,e401k + .295\,e401k \cdot inc$$
$$(3.66) \quad (.130)$$
$$[3.51] \quad [.160]$$

$n = 2{,}017$, $R^2 = .114$

The robust t statistic is about 1.84, and so the interaction term is marginally significant (two-sided p-value is about .066).

(iii) The coefficient on $e401k$ literally gives the estimated difference in financial wealth at $inc = 0$, which obviously is not interesting. It is not surprising that it is not statistically different from zero; we obviously cannot hope to estimate the difference at $inc = 0$, nor do we care to.

(iv) When we replace $e401k \cdot inc$ with $e401k \cdot (inc - 30)$, the coefficient on $e401k$ becomes 6.68 (robust $t = 3.20$). Now, this coefficient is the estimated difference in *nettfa* between those with and without 401(k) eligibility at roughly the average income, \$30,000. Naturally, we can estimate this much more precisely, and its magnitude (\$6,680) makes sense.