Solutions to the Practice questions -Chapter 8 (Heteroskedasticity)

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: β_{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 β_{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 heteroskedasticity-robust standard error gives $t \approx -.60$. In either case we fail to reject H_0: β_{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.

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

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

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

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,

$$\widehat{\log(price)} = \underset{\substack{(0.65)\\ [0.76]}}{-1.30} + \underset{\substack{(.038)\\ [.041]}}{.168}\log(lotsize) + \underset{\substack{(.093)\\ [.101]}}{.700}\log(sqrft) + \underset{\substack{(.028)\\ [.030]}}{.037}\, bdrms$$

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

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*.)