Answer Key: Problem Set 6

1. Consider a linear model to explain monthly beer consumption:

\[ beer = \beta_0 + \beta_1 inc + \beta_2 price + \beta_3 educ + \beta_4 female + u \]

\[ E(u | inc, price, educ, female) = 0 \]

\[ \text{var}(u | inc, price, educ, female) = \sigma^2 inc^2 . \]

Write the transformed equation that has a homoskedastic error term.

(Ans)

With \( \text{Var}(u | inc, price, educ, female) = \sigma^2 inc^2 \), \( h(x) = inc^2 \), where \( h(x) \) is the heteroskedasticity function defined in equation (8.21). Therefore, \( \sqrt{h(x)} = inc \), 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 \( \beta_1 \), 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.

2. Using the data in GPA3.DTA, the following equation was estimated for the fall and second semester students:

\[ \tilde{trmgpa} = -2.12 + 0.900 crsgpa + 0.193 cumpa + 0.0014 tothrs \]

\[ (0.55) (0.175) (0.064) (0.0012) \]

\[ [0.55] [0.166] [0.074] [0.0012] \]

\[ + 0.0018 sat - 0.0039 hsperrc + 0.351 female - 0.157 season \]

\[ (0.0002) (0.0018) (0.085) (0.098) \]

\[ [0.0002] [0.0019] [0.079] [0.080] \]

\[ n = 269, \quad R^2 = 0.465 \]

Here, \( \tilde{trmgpa} \) is term GPA, \( crsgpa \) is a weighted average of overall GPA in courses taken, \( cumpa \) is GPA prior to the current semester, \( tothrs \) is total credit hours prior to the semester, \( sat \) is SAT score, \( hsperrc \) is graduating percentile in high school class, \( female \) is a gender dummy, and \( season \) is a dummy variable equal to unity if the
student's sport is in season during the fall. The usual and heteroskedasticity-robust standard errors are reported in parentheses and brackets, respectively.

i. Do the variables crsgpa, cumgpa, and tothrs have the expected estimated effects? Which of these variables are statistically significant at the 5% level? Does it matter which standard errors are used?

(Ans)
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. Why does the hypothesis \( H_0: \beta_{crsgpa} = 1 \) make sense? Test this hypothesis against the two-sided alternative at the 5% level, using both standard errors. Describe your conclusion.

(Ans)
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 heteroskedasticity-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. Test whether there is an in-season effect on term GPA, using both standard errors. Does the significance level at which the null can be rejected depend on the standard error used?

(Ans)

The in-season effect is given by the coefficient on \textit{season}, which implies that, other things equal, an athlete’s GPA is about .16 points lower when his/her sport is competing. The \textit{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 \textit{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 \textit{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.

**Computer Exercise**

3. Use the data set GPA1.DTA for this exercise

i. Use OLS to estimate a model relating \textit{colGPA} to \textit{hsGPA}, \textit{ACT}, \textit{skipped}, and \textit{PC}. Obtain the OLS residuals

(Ans)

The equation estimated by OLS is

\[
\text{\textit{colGPA}} = 1.36 + .412 \text{ \textit{hsGPA}} + .013 \text{ \textit{ACT}} - .071 \text{ \textit{skipped}} + .124 \text{ \textit{PC}}
\]

\[
\begin{array}{cccc}
.33 & .092 & .010 & .026 & .057 \\
\end{array}
\]

\(n = 141, \; R^2 = .259, \; \tilde{R}^2 = .238\)

ii. Compute the special case of the White test for heteroskedasticity. In the regression of \(\hat{u}^2_i\) on \(\text{\textit{colGPA}_i}, \; \text{\textit{colGPA}_i}^2\), obtain the fitted values, say \(\hat{h}_i\).

(Ans)

The \(F\) statistic obtained for the White test is about 3.58. With 2 and 138 \(df\), this gives \(p\)-value \(\approx .031\). So, at the 5% level, we conclude there is evidence of heteroskedasticity in the errors of the \textit{colGPA} equation. (As an aside, note that the \textit{t} statistics for each of the terms is very small, and we could have simply dropped the
iii. Verify that the fitted values from part ii are all strictly positive. Then, obtain the weighted least squares estimates using weights $1/\hat{h}_i$. Compare the weighted least squares estimates for the effect of skipping lectures and the effect of PC ownership with the corresponding OLS estimates. What about their statistical significance?

(Ass)

In fact, the smallest fitted value from the regression in part (ii) is about .027, while the largest is about .165. Using these fitted values as the $\hat{h}_i$ in a weighted least squares regression gives the following:

$$
\text{colGPA} = 1.40 + .402 \text{hsGPA} + .013 \text{ACT} - .076 \text{skipped} + .126 \text{PC}
$$

(Ans)

There is very little difference in the estimated coefficient on PC, and the OLS $t$ statistic and WLS $t$ statistic are also very close. Note that we have used the usual OLS standard error, even though it would be more appropriate to use the heteroskedasticity-robust form (since we have evidence of heteroskedasticity). The $R^2$-squared in the weighted least squares estimation is larger than that from the OLS regression in part (i), but, remember, these are not comparable.

iv. With the WLS estimation from part iii, obtain heteroskedasticity-robust standard errors. In other words, allow for the fact that the variance function estimated in part ii might be misspecified. Do the standard error change much from part iii? [Hint: In Stata, use option 'robust' at the end of the 'regress' command to get the heteroskedasticity-robust standard errors.]

(Ass)

With robust standard errors – that is, with standard errors that are robust to misspecifying the function $h(x)$ – the equation is

$$
\text{colGPA} = 1.40 + .402 \text{hsGPA} + .013 \text{ACT} - .076 \text{skipped} + .126 \text{PC}
$$

(Ans)
The robust standard errors do not differ by much from those in part (iii); in most cases, they are slightly higher, but all explanatory variables that were statistically significant before are still statistically significant. But the confidence interval for $\beta_{pc}$ is a bit wider.