Problem Set 15
"Multiple Regression: Heteroskedasticity"

(15.1) [Consequences of heteroskedasticity] True, false, or uncertain? Explain your answer.

a. Heteroskedasticity causes the least squares estimators for the intercept and slope coefficients to be biased and inconsistent.
b. Heteroskedasticity causes the usual standard errors to be too large.

(15.2) [GQ test] The regression equation $y = \beta_1 + \beta_2 x + \varepsilon$ was estimated using 50 cross-sectional observations on states, by ordinary least squares. To check for heteroskedasticity related to population, separate regressions were run for the 17 states with the lowest populations and the 17 states with the highest populations. The sum of squared residuals for the low-population states was **270**. The sum of squared residuals for the high-population states was **90**.

a. Compute unbiased estimates of the variance of the error term in the two subsamples.
b. Given these results, which subsample appears to lie closer to the true regression line: the low-population-states or the high-population states? Explain your answer.
c. Test the null hypothesis of homoskedasticity, against the (one-sided) alternative hypothesis that high-population states have lower error variance, at 5% significance using a Goldfeld-Quandt test. Give
   - the *value* of the test statistic
   - the *degrees of freedom* for the test statistic
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program)
   - your conclusion: whether you reject the null hypothesis of homoskedasticity at 5% significance.
d. Regardless of your conclusion for part (c), suppose you believe that heteroskedasticity is indeed present and that the variance of the error term is inversely proportional to state population: $\text{Var}(\varepsilon_i) = \alpha/\text{pop}_i$, where α = an unknown constant and pop_i = population of state *i*, in millions Explain how you would transform the data to satisfy the classical assumptions.

(15.3) [BP test] Suppose we have estimated the following regression by ordinary least squares using data for 50 states:

$$\textit{average life expectancy} = \beta_1 + \beta_2\ \textit{fraction smokers} + \beta_2\ \textit{fraction obese} + \varepsilon$$

However, we are concerned that heteroskedasticity related to state population may invalidate the standard errors, so we wish to perform a Breusch-Pagan test, which requires an auxiliary regression, at 5% significance.

a. What is the dependent variable of the Breusch-Pagan auxiliary regression?
b. What is the regressor of the Breusch-Pagan auxiliary regression?
c. Would you expect to the estimated coefficient of this regressor to be positive or negative? [Hint: Note that the dependent variable in the original equation is an average.]

Suppose the R^2 value from the auxiliary regression is 0.082.

d. Give
   - the *value* of the Breusch-Pagan test statistic
   - the *degrees of freedom* for the test statistic
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program)
   - your conclusion: whether you can reject the null hypothesis of homoskedasticity at 5% significance.

(15.4) [BP test] Suppose we have estimated the following regression by ordinary least squares using data for 400 workers:

$$\textit{annual salary} = \beta_1 + \beta_2\ \textit{years of education} + \beta_2\ \textit{years of experience} + \textit{hour of work per week} + \varepsilon$$

However, we are concerned that heteroskedasticity related to both *education* and *experience* may invalidate the standard errors, so we wish to perform a Breusch-Pagan test, which requires an auxiliary regression, at 5% significance.

a. What is the dependent variable of the Breusch-Pagan auxiliary regression?
b. What are the regressors of the Breusch-Pagan auxiliary regression?

Suppose the R^2 value from the auxiliary regression is 0.011.

c. Give
   - the *value* of the Breusch-Pagan test statistic
   - the *degrees of freedom* for the test statistic
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program)
   - your conclusion: whether you can reject the null hypothesis of homoskedasticity at 5% significance.

(15.5) [White test] Suppose we wish to test for heteroskedasticity in the following regression equation explaining firms' costs, using White's test, at 5 percent significance.

$$cost = \beta_1 + \beta_2\ output + \beta_3\ oil\ price + \varepsilon$$

Recall that White's test involves estimating an auxiliary regression equation.

a. What is the dependent variable of the White test auxiliary regression?
b. What are the regressors of the White test auxiliary regression?
c. Assume there are 80 observations and the R^2 value from the auxiliary regression is 0.16. Give
   - the *value* of the White test statistic
   - the *degrees of freedom* for the test statistic
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program)
   - your conclusion: whether you can reject the null hypothesis of homoskedasticity at 5% significance.

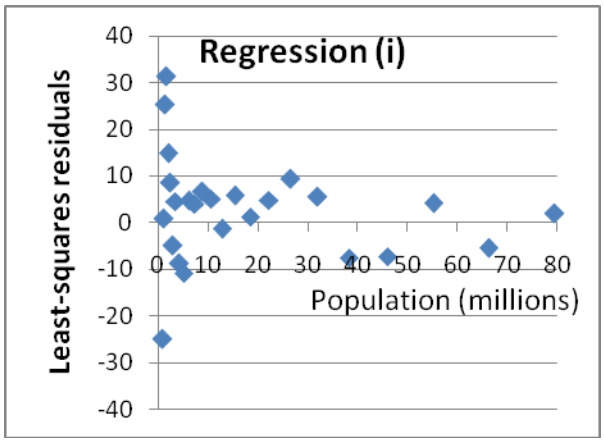
(15.6) [White test] Suppose we wish to test for heteroskedasticity in the regression equation

$$y = \beta_1 + \beta_2\ x + \beta_3\ x^2 + \varepsilon$$

using White's test, at 5 percent significance. Note that one regressor is the square of another here. Recall that White's test involves estimating an auxiliary regression equation.

a. What variable should be on the left side of this auxiliary regression? That is, what should be the dependent variable?
b. What variables should be on the right side of this auxiliary regression? That is, what should be the regressors? [Hint: Eliminate duplicate regressors.]
c. Assume there are 150 observations and the R^2 value from the auxiliary regression is 0.040. Give
   - the *value* of the White test statistic
   - the *degrees of freedom* for the test statistic
   - the *critical point* from the appropriate table at the back of your textbook (or compute the *p-value* using a spreadsheet program)
   - your conclusion: whether you can reject the null hypothesis of homoskedasticity at 5% significance.

(15.7) [Weighted least-squares] Consider the plots of least-squares residuals for the two ordinary least-squares regressions shown below. Let pop_i denote the population of each observation.



a. In regression (i), does the variance of the error term appear to be *positively* or *negatively* related to population?

b. To correct for heteroskedasticity in regression (i), should the data be *multiplied* by $\sqrt{pop_i}$ or *divided* by $\sqrt{pop_i}$?

c. In regression (ii), does the variance of the error term appear to be *positively* or *negatively* related to population?

d. To correct for heteroskedasticity in regression (ii), should the data be *multiplied* by $\sqrt{pop_i}$ or *divided* by $\sqrt{pop_i}$?

(15.8) [Weighted least-squares] Suppose we wish to estimate the following equation

$$average\ income = \beta_1 + \beta_2\ average\ education + \varepsilon$$

using state-level observations. However, we want to eliminate any heteroskedasticity in the error term so that our estimates of the coefficients are more precise.

a. Note the dependent variable is an *average*. Is the variance of the error term more likely to be *proportional* to population (pop_i) or *inversely proportional* to population? Why?

b. Given your answer to part (a), should the data be *multiplied* by $\sqrt{pop_i}$ or *divided* by $\sqrt{pop_i}$ to correct for heteroskedasticity?

c. The first two observations, and state population, are given below.

Observation #	Average income (thousands of $)	Average education (years)	State population (millions)
1	38	11.2	6.250
2	41.5	14.5	10.240

Given your answer to part (b), transform the data to eliminate heteroskedasticity by completing the table below.

Observation #	Transformed average income	Transformed replacement for intercept	Transformed average education
1			
2			

(15.9) [Weighted least-squares] Suppose we wish to estimate the following equation

$$total\ consumption = \beta_1 + \beta_2\ total\ income + \varepsilon$$

using state-level observations. However, we want to eliminate any heteroskedasticity in the error term so that our estimates of the coefficients are more precise.

a. Note that the dependent variable is a *total*. Is the variance of the error term more likely to be *proportional* to population (pop_i) or *inversely proportional* to population? Why?

b. Given your answer to part (a), should the data be *multiplied* by $\sqrt{pop_i}$ or *divided* by $\sqrt{pop_i}$ to correct for heteroskedasticity?

c. The first two observations, and state population, are given below.

Observation #	Total consumption (billions of $)	Total income (billions of $)	State population (millions)
1	525	560	12.250
2	324	378	7.290

Given your answer to part (b), transform the data to eliminate heteroskedasticity by completing the table below.

Observation #	Transformed total consumption	Transformed replacement for intercept	Transformed total income
1			
2			

[end of problem set]