Elements of Econometrics. 2022-2023.
Class 10. Heteroscedasticity

Question 1 (UoL Exam)
Consider a regression model

$$y_i = \alpha x_i + u_i;\ i,\ j = 1,\ 2,\ \ldots,\ n$$

where $E(u_i) = 0$, $E(u_i^2) = \sigma^2 x_i^2$ and $E(u_i u_j) = 0$ if $i \neq j$ for all $i,\ j = 1,\ 2,\ \ldots,\ n$. x's are fixed.

(a) Derive the weighted least squares (WLS) estimator $\hat{\alpha}$, of α.

(b) Prove that WLS estimator of α is unbiased.

(c) Derive the variance of $\hat{\alpha}$.

(d) Prove that WLS estimator of α is consistent.

Question 2 (UoL/ICEF Exam) A researcher has data on school enrolment, *N*, and annual fixed expenditure, *EXP*, measured in thousands of Korean won, for the sample of 100 schools in South Korea and estimated cost function of the quadratic form

$$\hat{EXP} = \underset{(12,908)}{17,999} + \underset{(133)}{1,060}N - \underset{(0.29)}{1.27}N^2 \quad R^2 = 0.74 \quad (1)$$

Suspecting that the regression was subject to heteroscedasticity, the researcher runs the regression twice more, first with the 43 schools with lowest enrolments, then with the 31 schools with the highest enrolments The residual sum of squares (*RSS*) in the two regressions are 8,416,000 and 65,525,000 respectively.
The researcher defines a new variable, *EXPSTUD*, expenditure per student, as $EXPSTUD = EXP/N$, and runs the regression

$$\hat{EXPSTUD} = \underset{(33)}{1,316} - \underset{(0.15)}{1.92}N \quad R^2 = 0.63 \qquad (2)$$

He then runs regressions with the 43 smallest schools and the 31 largest schools and the residual sums of squares are 967,000 and 698,000 respectively.
Her colleague suggests that it might be a good idea to include the reciprocal of *N* in the regression. The researcher does this and obtains the following result:

$$\hat{EXPSTUD} = \underset{(91)}{1,078} - \underset{(0.26)}{1.31}N + \underset{(5,863)}{16,374}NRECIP \quad R^2 = 0.66 \qquad (3)$$

where $NRECIP = 1/N$.
He again runs regressions with the 43 smallest schools and the 31 largest schools and the residual sums of squares are 948,000 and 677,000.

(a) □ Explain why the researcher should be prepared for the presence of heteroscedasticity in the regression under consideration.
□ Explain what is meant by heteroscedasticity and describe the consequences of its presence in a regression model.

(b) □ Describe the Goldfeld-Quandt test for heteroscedasticity and explain why under certain conditions it may detect heteroscedasticity.
□ What assumptions must be made to correctly apply this test?
□ Why the researcher may believe that these assumptions are fulfilled?
□ Run a Goldfeld-Quandt test for heteroscedasticity on each of the regressions.

(c) □ Explain why the researcher ran the second regression.
□ Give an economic interpretation of regression (3) and explain why it may be preferable to regression (2).
□ What significance, if any, should be attached to the fact that the coefficient of *N* is smaller in regression (3) and its standard error is larger?

Question 3 (ICEF Exam)

The researcher studies the factors that affect the volume of paid services per capita V_i in 82 regions of Russia (in rubles). He suggests that this indicator may depend primarily on average per capita monthly income in rubles I_i (from 14000 to 70000 rubles depending on the region), as well as on the level of unemployment in percent U_i for each region. In addition, the researcher suggests that the situation with paid services in the central (near Moscow) and northwestern regions of Russia (near the city of St. Petersburg) may differ from the rest of the country, so he introduces a dummy variable R_i equal to 1 for central and northwestern regions, and equal to 0 for other regions of Russia.

(a) To assess the impact of income on paid services, the researcher first runs a simple linear regression

$$\hat{V}_i = -2448.2 + 2.05 I_i \qquad R^2 = 0.78$$
$$(3546.8)\ \ (0.12) \qquad\qquad \textbf{(1)}$$

The researcher is afraid that the equation may not be of sufficient quality due to possible heteroscedasticity.

□ What is heteroscedasticity? Explain how heteroscedasticity can arise here. What characteristics of the equation can heteroscedasticity influence and how? How to correct them? .

□ The researcher then rank all regions in order of increasing per capita income, and then regresses first for the 20 regions with the lowest income (getting RSS value equal $4.81 \cdot 10^8$), and then for the 30 regions with the highest income (getting RSS value equal $5.87 \cdot 10^9$). How can this information be used to check the data for heteroscedasticity? Carry out the necessary calculations, explaining your actions, and make the conclusion.

(b) Then the researcher runs multiple regression

$$\hat{V}_i = -27725.7 + 3.75 I_i - 2.26 \cdot 10^{-5} I_i^2 - 348.9 U_i - 6323.1 R_i \qquad R^2 = 0.82$$
$$(10858.6)\ \ (0.56)\ \ (6.89 \cdot 10^{-6})\ \ (354.8)\ \ (2389.7) \qquad\qquad \textbf{(2)}$$

After conducting White's test with all the cross-terms for equation (2), the researcher obtained the value of the determination coefficient $R^2 = 0.57$ for the auxiliary equation and concluded that heteroscedasticity is present.

□ Explain the mathematics of the White's test: how is the auxiliary equation constructed, how many regressors it includes.

□ Run White test and make the conclusion.

□ What are relative advantages and disadvantages of Goldfeld-Quandt and White tests?

(c) In an effort to get rid of heteroscedasticity, the researcher runs the following equations (both equations demonstrate an absence of heteroscedasticity)

$$\widehat{(V_i/I_i)} = 2.99 - 1.19 \cdot 10^{-5} I_i - 16117.4 \cdot (1/I_i) - 292.6(U_i/I_i) - 6249.5(R_i/I_i) \qquad R^2 = 0.22$$
$$(0.58)\ \ (8.71 \cdot 10^{-6})\ \ (9728.1) \qquad (215.7) \qquad (2062.8) \qquad \textbf{(3)}$$
$$\widehat{log\, V_i} = 9.2 + 8.42 \cdot 10^{-5} I_i - 6.95 \cdot 10^{-10} I_i^2 - 0.02 U_i - 0.11 R_i \qquad R^2 = 0.85$$
$$(0.17)\ (8.82 \cdot 10^{-6})\ \ (1.09 \cdot 10^{-10})\ \ (0.006)\ \ (0.04) \qquad\qquad \textbf{(4)}$$

□ Both equations reveal no heteroscedasticity. Explain why each of the specifications for equations (3) and (4) was able to eliminate heteroscedasticity.

□ In equation (3) the value R^2 is less than in equation (4), and the income factor became negative and insignificant? Is this an indication that the resulting equation is of poor statistical quality?