Econometrics-2022-2023
Home assignment 4. Specification. Heteroscedasticity.
To be submitted by December 13, 23:55

1. [30 marks] A student of ICEF considers for her graduation thesis the hedonic price function for houses. The hedonic price refers to the implicit price of a house given certain attributes (e.g., the number of bedrooms). The data contains the sale price of 546 houses sold in the summer of 2017 in the Moscow region (in thousands of rubles) along with their important features. The following characteristics are available: the lot size of the property in square meters (*lotsize*), the numbers of bedrooms (*bedrooms*), the number of full bathrooms (*bathrooms*), and a dummy indicating the presence of air-conditioning (*airco*).
Consider the following ordinary least squares results

$$\widehat{\log(price)}_i = \underset{\substack{(0.232)\\[0.233]}}{9.894} + \underset{\substack{(0.028)\\[0.028]}}{0.400}\log(lotsize)_i + \underset{\substack{(0.015)\\[0.017]}}{0.078}\,bedrooms_i + \underset{\substack{(0.023)\\[0.024]}}{0.096}\,bathrooms_i + \underset{\substack{(0.024)\\[0.023]}}{0.212}\,airco_i \quad n = 546 \quad \textbf{(1)}$$

The usual standard errors are in parentheses, the heteroskedasticity robust standard errors are in square brackets, and *RSS* measures the residual sum of squares.

1.1. [10 marks] □ Interpret the parameter estimates on $\log(lotsize)$, *bedrooms* and *airco*. Discuss also the intercept. Comment the differences in standard errors of two types.
□ The variable *airco* is a dummy variable so it can take only two values so interpretation based on the properties of infinitely small values above is not quite correct. Calculate the effect of having airconditioning on the price of a flat and compare it with the result based on the analysis of differentials obtained above.

1.2. [10 marks] □ The student suspects the presence of heteroscedasticity in the data. She noticed that the variance of $\log(price)$ is increasing with $\log(lotsize)$. To test for heteroscedasticity she orders the 546 observations by the magnitude. Then she runs the same regression (1) using 212 observations with the smallest values of $\log(lotsize)$, obtaining $RSS_1 = 7.781$, and then runs the same regression for 90 observations with the largest values of $\log(lotsize)$, obtaining $RSS_2 = 6.214$ (an unequal number of observations was taken due to the fact that number of lots with high prices turned out to be significantly less). Help the student to run the appropriate test and to make the conclusion.

1.3. [10 marks] □ Student's friend advised her to use White test instead. She runs test for regression (1) obtaining for the auxiliary equation of the White test (with cross terms) the value of $R^2 = 0.0317$. Help the student to complete the test briefly explaining her the logic of this test, indicating distribution, degrees of freedom, critical values and final conclusion.

2. [20 marks] A researcher obtains data on household annual expenditure on books, *B*, and annual household income, *Y*, for 100 UK households in 2015. He hypothesizes that *B* is related to *Y* and the average cognitive ability of adults in the household, *IQ*, by the relationship (where *u* satisfies the Gauss–Markov conditions)

$$\log B = \beta_1 + \beta_2 \log Y + \beta_3 \log IQ + u \qquad \textbf{(A)}$$

He also considers the possibility that $\log B$ may be determined by $\log Y$ alone:

$$\log B = \beta_1 + \beta_2 \log Y + u \qquad \textbf{(B)}$$

He does not have data on *IQ* and decides to use average years of schooling of the adults in the household, *S*, as a proxy in specification **(A)**. It may be assumed that *Y* and *S* are both nonstochastic. In the sample the correlation between $\log Y$ and $\log S$ is 0.86. He performs the following regressions: **(1)** $\log B$ on both $\log Y$ and $\log S$, and **(2)** $\log B$ on $\log Y$ only, with the results shown in the table (standard errors in parentheses):

$$\widehat{\log B} = \underset{(2.28)}{-6.89} + \underset{(0.69)}{1.10}\log Y + \underset{(0.35)}{0.59}\log S \quad R^2 = 0.29 \qquad \textbf{(1)}$$

$$\widehat{\log B} = \underset{(0.89)}{-3.37} + \underset{(0.35)}{2.10}\log Y \quad R^2 = 0.27 \qquad \textbf{(2)}$$

2.1. [10 marks] □ Assuming that **(A)** is the correct specification, explain, with a mathematical proof, whether you would expect the coefficient of $\log Y$ to be greater in regression **(2)**.
□ Assuming that (A) is the correct specification, describe the various benefits from using $\log S$ as a proxy for $\log IQ$, as in regression (1), if log *S* is a good proxy.

2.2. [10 marks] □ Explain whether the low value of R^2 in regression **(1)** implies that $\log S$ is not a good proxy.
□ Assuming that **(A)** is the correct specification, provide an explanation of why the coefficients of $\log Y$ and $\log S$ in regression **(1)** are not significantly different from zero.
□ Assuming that **(B)** is the correct specification, explain whether you would expect the coefficient of $\log Y$ to be lower in regression **(1)**.
□ Assuming that **(B)** is the correct specification, explain whether the standard errors in regression **(1)** are valid estimates.

3. [**50 marks]** For your data set **ha04_data__** regress ***EARNINGS*** on ***ASVABC*** and ***S*** **(Regression 1)**, ***EARNINGS*** on ***ASVABC*** only **(Regression 2)**, and ***EARNINGS*** on ***S*** only **(Regression 3).**
3.1. [10 marks] □ Compare the results of all three regressions. Assuming Regression (1) corresponds to the correct specification explain the difference between regressions, paying attention to comparing regression coefficients and various characteristics of regression quality. The mathematical justification for your conclusions is not expected here.
□ Compare again the results of all three regressions but now assuming that correct specification corresponds to the regression (2) and (3) (in turn).
□ Which of the assumptions about the correct specification looks most convincing, taking into account the results of your analysis?

3.2. [10 marks] □ Suppose that data on the variable ***S*** is not available, but there is data on variables ***SM*** and ***SF***. How can these variables be used as proxies for building regressions, what are the advantages of regressions using proxies, what are their disadvantages?

3.3. [10 marks] □ Why heteroscedasticity in regression (1)-(3) can be expected? Provide a meaningful explanation with an example of regression (2) or (3).
□ Is there any sign of heteroscedastisity in (1-3) judging by some graphical tools?
□ How could heteroscedasticity affect the estimated equations (1-3)?

3.4. [10 marks] □ For equations (1-3), run Breusch-Pagan test and White's tests (with and without cross terms) and conclude whether heteroscedasticity is present. Discuss breafly the comparative advantages and disadvantages of these tests.
□ For equation (3) (regression EARNINGS ON S), perform the Goldfeld-Quandt test and conclude whether heteroscedasticity is present.

3.5. [10 marks] □ Use weighted least squares and corrected standard errors to estimate regression (3) (regression of EARNINGS on S) using as weighting series expanatory variable. Test obtained equation for heteroscedasticity and compare it with the unweighted one.
□ Use weighted least squares and corrected standard errors to estimate regression (1) using as weighting series expanatory variables in turn. Test obtained equations for heteroscedasticity.
□ Use logarithm in dependent variable to reduce the risk of heteroscedasticity. Explain.