Heteroscedasticity

Q1) $y_i = \beta x_i + \varepsilon_i$

$E(\varepsilon_i) = 0$, $E(\varepsilon_i \varepsilon_j)$ for $i \neq j$

$E(\varepsilon_i^2) = a \cdot x_i^2$; $\sum x_i^2 = n$

a. $\hat{\beta}_{OLS} = \frac{\sum x_i y_i}{\sum x_i^2}$ $\qquad \hat{\beta} = (X'X)^{-1} \cdot X'y$

$$E(\hat{\beta}_{OLS}) = E\left(\frac{\sum x_i (\beta x_i + \varepsilon_i)}{\sum x_i^2}\right)$$

$$= \beta + \frac{\sum x_i \overbrace{E(\varepsilon_i)}^{=0}}{\sum x_i^2} = \beta$$

$$Var(\hat{\beta}_{OLS}) = E(\hat{\beta} - E(\hat{\beta}))^2 =$$

$$= E(\hat{\beta} - \beta)^2 = E\left(\beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2} - \beta\right)^2$$

$$= E\left(\frac{\sum x_i \varepsilon_i}{\sum x_i^2}\right)^2 = \left(\frac{1}{\sum x_i^2}\right)^2 \cdot$$

$$E\left(\sum x_i^2 \varepsilon_i^2 + \sum_j \sum_i x_i x_j \varepsilon_i \varepsilon_j\right) =$$

$$\left(\frac{1}{\sum x_i^2}\right)^2 \left(\sum x_i^2 \underbrace{E(\varepsilon_i^2)}_{a \cdot x_i^2} + \sum\sum x_i x_j \underbrace{E(\varepsilon_i \varepsilon_j)}_{0} \right)$$

$$= \frac{a \sum x_i^4}{\left(\sum x_i^2\right)^2} = \frac{a \sum x_i^4}{h^2}$$

$x_i = 1$: $\varepsilon_i \sim N(0, a)$ $\quad$ $x_i = 2$: $\varepsilon_i \sim N(0, 4a)$

b. $y_i = \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, a \cdot x_i^2)$

$E(\varepsilon_i) = 0$, $E(\varepsilon_i \varepsilon_j)$ for $i \neq j$

$E(\varepsilon_i^2) = a \cdot x_i^2$; $\sum x_i^2 = n$

WLS :

$$\frac{y_i}{x_i} = \beta + \frac{\varepsilon_i}{x_i} \qquad u_i \sim N\left(0, \frac{a x_i^2}{x_i^2}\right)$$

u_i $\qquad v_i \sim N(0, a)$

$$\hat{\beta}_{WLS} = \frac{\sum v_i / x_i}{h}$$

$$E(\hat{\beta}_{WLS}) = \frac{1}{h} E\left(\sum \frac{\beta \cdot x_i + \varepsilon_i}{x_i}\right) =$$

$$= \frac{1}{n} \cdot E\left(\sum \beta + \frac{\varepsilon_i}{x_i}\right) =$$

$$= \frac{n}{n} \cdot \beta = \beta$$

$$Var(\hat{\beta}_{WLS}) = E(\hat{\beta} - \beta)^2$$

$$E\left(\frac{1}{n} \cdot \sum \frac{y_i}{x_i} - \beta\right) =$$

$$E\left(\beta + \frac{1}{n}\sum \frac{\varepsilon_i}{x_i} - \beta\right)^2 =$$

$$= \frac{1}{n^2} E\left(\sum \frac{\varepsilon_i}{x_i}\right)^2 =$$

$$= \frac{1}{n^2} E\left(\sum \left(\frac{\varepsilon_i}{x_i}\right)^2 + \sum_j \sum_i \frac{\varepsilon_i}{x_i} \frac{\varepsilon_i}{x_j}\right) =$$

$\left.\begin{matrix} E(\varepsilon_i \varepsilon_j) = 0 \\ E(\varepsilon_i^2) = a x_i^2 \end{matrix}\right\} = \frac{1}{n^2} \cdot \sum \frac{a \cdot \cancel{x_i^2}}{\cancel{x_i^2}} = \frac{na}{n^2} = \frac{a}{n}$

C. $Var(\hat{\beta}_{OLS}) = \dfrac{a \sum x_i^4}{n^2}$

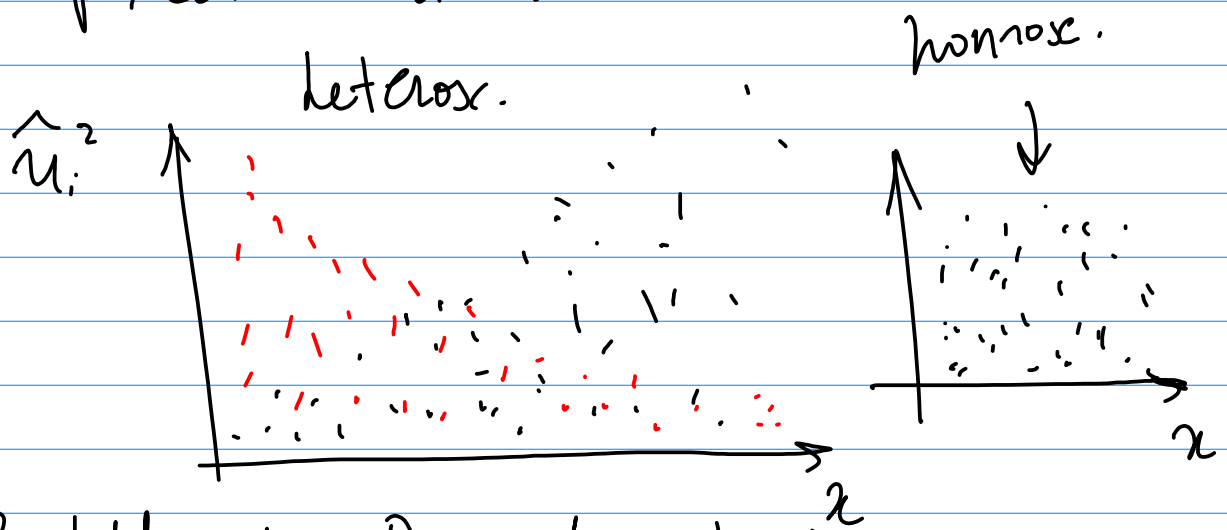
$Var(\hat{\beta}_{WLS}) = a/n$

$$\frac{Var(\hat{\beta}_{OLS})}{Var(\hat{\beta}_{WLS})} = \frac{\sum x_i^4}{n}$$

$$\sum x_i^2 = n \qquad \frac{\sum x_i^4}{n} \geq 1$$

Testing for Heteroscedasticity

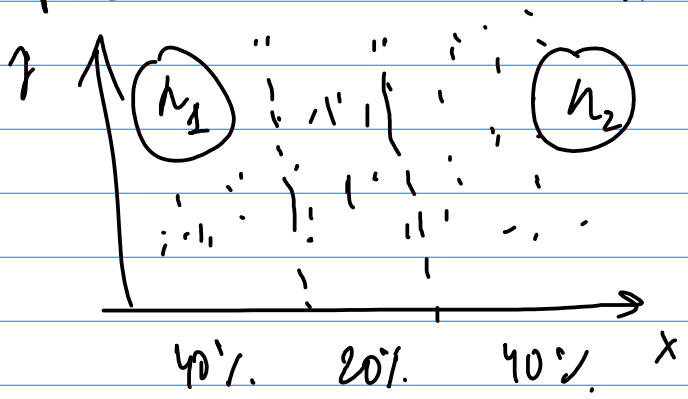
- Graphical Method



- Goldfeld - Quandt test

assumpt. σ_u proport. to x_i

1. Sample is div. in 3 parts



2. $$F = \frac{RRS_2 / n_2 - k}{RSS_1 / n_1 - k} \sim F_{n_2-k, n_1-k}$$

(★) $RSS_2 < RSS_1 \Rightarrow F^{-1}$

White test

$H_0: \sigma_1^2 = \dots = \sigma_n^2$

$H_a:$ heteroscedasticity

1) $y, | X \Rightarrow \hat{\varepsilon}_i$

2) $\hat{\varepsilon}_i^2 \left(\beta_0, x_1 \dots x_k, x_1^2, \dots, x_k^2, x_1 x_2, \dots, x_{k-1} \cdot x_k \right)$

$\hookrightarrow R^2$

$R^2 \cdot n \sim \chi^2_{k_{aux}}$

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

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

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.

a) ~~WLS~~ GLS

Robust to heterosc. s.e.

White form

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum (X_i - \bar{X})^2}$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sum x_i^2 \sigma_{u_i}^2}{\left(\sum x_i^2\right)^2}$$

□ 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.

$$GQ = \frac{RSS_2 / n_2 - k}{RSS_1 / n_1 - k} = \frac{5{,}87 \cdot 10^9 / 30 - 2}{4.81 \cdot 10^8 / 20 - 2} =$$

$= 7,84$

$F_{1\%;\,28,18} = 2,58$

(b) Then the researcher runs multiple regression

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

(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?

$n \cdot R^2 \sim \chi^2_{k_{aux}}$

$k = 4 + 2 + 6 = 12$ ← cross-terms

$N = 82$

$n \cdot R^2 = 82 \cdot 0,57 = 46,7$

$\chi^2_{1\%,12} = 23,2$

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

WLS:

$$(\hat{V}/I_i) = 2.99 - 1.19\cdot10^{-5} I_i - 16117.4\cdot(1/I_i) - 292.6(U_i/I_i) - 6249.5(R_i/I_i) \qquad R^2 = 0.22$$
$$(0.58)\ (8.71\cdot10^{-6})\ \ (9728.1)\qquad\qquad (215.7)\qquad\qquad (2062.8)$$

(3)

$$\log\hat{V}_i = 9.2 + 8.42\cdot10^{-5} I_i - 6.95\cdot10^{-10} I_i^2 - 0.02U_i - 0.11R_i \qquad R^2 = 0.85$$
$$(0.17)\ (8.82\cdot10^{-6})\ \ (1.09\cdot10^{-10})\ \ (0.006)\ \ (0.04)$$

(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?