Eco 213: Basic Data Analysis and Econometrics
Lecture 7: Heteroskedasticity
March 16, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

Outline

- Heteroskedasticity
- OLS and Heteroskedasticity
- Tests for Heteroskedasticity
- Remedial measures

Homoskedasticity

- What happens if we relax the assumption that?

$$Var(\varepsilon_i) = \sigma^2$$

Heteroskedasticity

- The variance of ε_i is NOT a constant σ^2.
- The variance of ε_i is greater for some observations than for others.

$$Var(\varepsilon_i) = \sigma_i^2$$

Causes of Heteroskedasticity

- Learning: reduces errors; driving practice, driving errors and accidents

    typing practice and typing errors, defects in productions; improved machines
- Growth: saving and variance of saving increases with income
- Improved data collection: better formulas and goods software
- Outliers affect the value of estimates
- Specification Errors and omitted variables:- in a demand model if you regress demand of a product to only its own price, there is a danger variables such as the prices of complements and income may appear in the error term.

More heteroscedasticity exists in cross section than in time series data.

Heteroskedasticity (cont.)

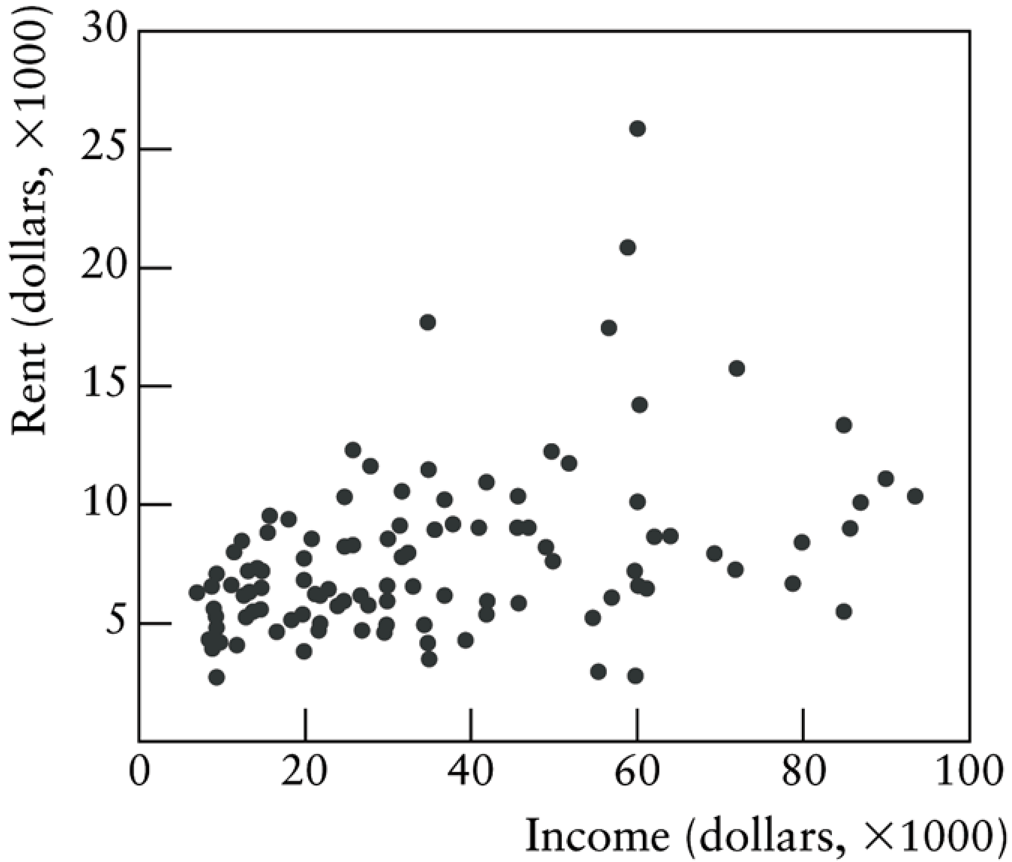
- For example, consider a regression of housing expenditures on income.

$$Rent_i = \beta_0 + \beta_1 Income_i + \varepsilon_i$$

- Consumers with low values of income have little scope for varying their rent expenditures. $Var(\varepsilon_i)$ is low.

- Wealthy consumers can choose to spend a lot of money on rent, or to spend less, depending on tastes. $Var(\varepsilon_i)$ is high.

Figure 10.1 Rents and Incomes for a Sample of New Yorkers



OLS and Heteroskedasticity

- What are the implications of heteroskedasticity for OLS?
- Under the Gauss–Markov assumptions (including homoskedasticity), OLS was the Best Linear Unbiased Estimator.
- Under heteroskedasticity, is OLS still Unbiased?
- Is OLS still Best?

OLS and Heteroskedasticity (cont.)

- Implications of Heteroskedasticity:
  - OLS is still unbiased.
  - OLS is no longer efficient; some other linear estimator will have a lower variance.
  - Estimated Standard Errors will be incorrect; *C.I.*'s and hypothesis tests (both t- and F- tests) will be incorrect.

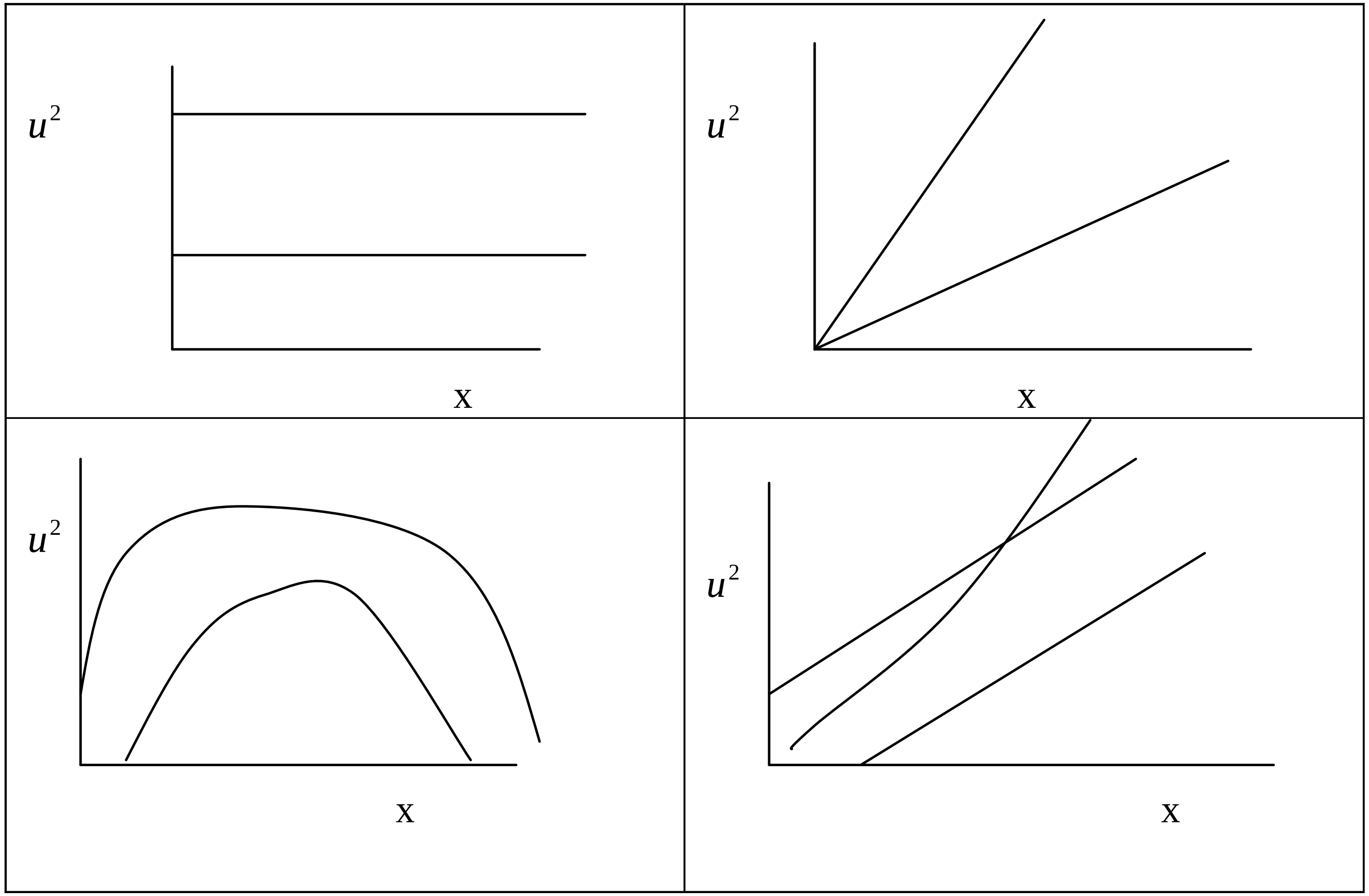
Nature and Causes

- LS assumption: variance of e_i is constant $\text{var}\left[e_i\right] = \sigma^2$ for every ith observation, $\text{var}\left(\hat{\beta}_2\right) = \hat{\sigma}^2 \left[\frac{1}{\sum_i \left(x_i - \bar{x}\right)^2} \right]$ but it is possible that

$$\sigma_i^2 = \sigma^2 x_i$$

Causes: Learning, growth, improved data collection, outliers, omitted variables;

Detection of Heteroscedasticity: Informal (Graphical) method



Consequence of Heteroscedasticity

OLS estimators give **unbiased** and **linear** estimates but not best because they have large variance with the heteroscedasticity.

Assume a simple model: $Y_i = \beta_1 + \beta_2 x_i + e_i$

- OLS estimators are still unbiased $E(\hat{\beta}_2) = \beta_2$

Proof:

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum w_i y_i = \sum w_i (\beta_1 + \beta_2 x_i + e_i)$$

where $w_i = \dfrac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$

$$E(\hat{\beta}_2) = E\left[\sum w_i y_i\right] = E\left[\sum w_i (\beta_1 + \beta_2 x_i + e_i)\right] = E\left[\sum w_i \beta_1 + \beta_2 \sum w_i x_i + \sum w_i e_i\right] = \beta_2$$

Consequence of Heteroscedasticity

- Variance of estimated parameters and the dependent variable

- $\operatorname{var}(\hat{\beta}_2) = \hat{\sigma}^2 \left[\frac{1}{\sum_i (x_i - \bar{x})^2} \right]$

It was proved above that

$$E(\hat{\beta}_2) = E\left[\sum w_i y_i\right] = E\left[\sum w_i (\beta_1 + \beta_2 x_i + e_i)\right] = E\left[\sum w_i \beta_1 + \beta_2 \sum w_i x_i + \sum w_i e_i\right] = \beta_2$$

$$\operatorname{var}(\hat{\beta}_2) = E\left[E(\hat{\beta}_2) - \beta_2\right]^2 =$$

$$E\left[\sum w_i e_i\right]^2 = \left[\sum w_i^2 \operatorname{var}(e_i)^2 + \sum_{i \neq j} \sum w_i w_j \operatorname{cov}(e_i e_j)\right] = \sum w_i^2 \sigma_i^2 = \frac{\sum_i (x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$$

$\operatorname{var}(\hat{\beta}_2) = \frac{\sum_i (x_i - \bar{x})^2 \sigma_i^2}{\left[\sum_i (x_i - \bar{x})^2\right]^2}$, thus variance of parameter is no longer constant. It rises with observations. When variances are larger, the standard errors are large and calculated t becomes smaller and coefficients become insignificant, though we may have correct variables in the model.

Tests for Heteroscedasticity

There are a series of formal methods developed in the econometrics literature to detect the existence of Heteroscedasticity in a given regression model.

Park test

Model $Y_i = \beta_1 + \beta_2 x_i + e_i$ (1)

Error square: $\sigma_i^2 = \sigma^2 x_i^{\beta} e^{v_i}$ (2)

Or taking log

$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_i + v_i$ (2')

steps : run the OLS regression for (1) and get the estimates of error terms e_i. Square e_i, and then run a regression of $\ln e_i^2$ with x variable. Do t-test H_0: $\beta = 0$ against H_A: $\beta \neq 0$. If β is significant then that is the evidence of heteroscedasticity.

Tests for Heteroscedasticity

Glejser test

$Y_i = \beta_1 + \beta_2 x_i + e_i$

There are several tests

$$|e_i| = \beta_1 + \beta_2 X_i + v_i$$

$$|e_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i$$

$$|e_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$

$$|e_i| = \sqrt{\beta_1 + \beta_2 X_{i_i}^2} + v_i$$

In each case do t-test H_0: $\beta = 0$ against H_A: $\beta \neq 0$. If β is significant then that is the evidence of heteroscedasticity.

Goldfeld-Quandt test

Model $Y_i = \beta_1 + \beta_2 x_i + e_i$ (1)

Steps:

1. Rank observations in ascending order of one of the x variable
2. Omit c numbers of central observations leaving two groups with $\frac{n-c}{2}$ number of osbervations
3. Fit OLS to the first $\frac{n-c}{2}$ and the last $\frac{n-c}{2}$ observations and find sum of the squared errors from both of them.
4. Set hypothesis

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ against } H_A : \sigma_1^2 \neq \sigma_2^2.$$

5. compute $\lambda = \frac{RSS_2 / df2}{RSS_1 / df1}$ it follows F distribution.

Tests for Heteroscedasticity

Breusch-Pagan,Godfrey test	White Test
$Y_i = \beta_1 + \beta_2 x_{2,i} + .. + \beta_k x_{k,i} + e_i$ 1.run OLS and obtain error squares 2. Obtain average error square $\tilde{\sigma}^2 = \sum \frac{\hat{e}_i^2}{n}$ and $p_i = \frac{\hat{e}_i^2}{\tilde{\sigma}^2}$ 3. regress p_i on a set of explanatory variables $p_i = \alpha_1 + \alpha_2 x_{2,i} + .. + \alpha_k x_{k,i} + e_i$ 4. obtain squares of explained sum (ESS) 5. $\theta = \frac{1}{2}(ESS)$ 6. $\theta = \frac{1}{m-1}(ESS) \approx \chi^2_{m-1}$ $H_0 : \alpha_2 = \alpha_2 = .. = \alpha_k = 0$ No heteroscedasticity and $\sigma_i^2 = \alpha_1$ a constant. If calculated χ^2_{m-1} is greater than table value there is an evidence of heteroscedasticity.	This is a more general test Model $Y_i = \beta_1 + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i$ Run OLS to this and get $\hat{e}_i$ $\hat{e}_i^2 = \alpha_1 + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{2,i}^2 + \alpha_5 x_{3,i}^2$ $\alpha_6 x_{2,i} x_{3,i} + v_i$ Compute the test statistics $n.R^2 \sim \chi^2_{df}$ Again if the calculated χ^2_{df} is greater than table value there is an evidence of heteroscedasticity.

Remedial Measures

Weighted Least Square and GLS when σ_i^2 known,

divide the whole equation by σ_i

Apply OLS to transformed variables.

$$\frac{Y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \beta_1 \frac{X_1}{\sigma_i} + \beta_2 \frac{X_2}{\sigma_i} + \beta_3 \frac{X_3}{\sigma_i} + \beta_4 \frac{X_4}{\sigma_i} + \ldots\ldots + \beta_k \frac{X_k}{\sigma_i} + \frac{e_i}{\sigma_i}$$

Variance of this transformed model equals 1.

Generalized Least Squares

- Note: we derive the same BLUE Estimator (Generalized Least Squares) whether we:

    1. Find the optimal weights for heteroskedastic data, or
    2. Transform the data to be homoskedastic, then use OLS weights

GLS: An Example (cont.)

- We want to estimate the relationship

$$rent_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

- We are concerned that higher income individuals are less constrained in how much income they spend in rent. Lower income individuals cram into what housing they can afford; higher income individuals find housing to suit their needs/tastes.
- That is, $Var(\varepsilon_i)$ may vary with income.

TABLE 10.1
Rent and Income in New York

Dependent Variable: RENT
Method: Least Squares
Sample: 1 108
Included Observations: 108

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5455.483	602.7776	9.050573	0.0000
INCOME	0.063568	0.014390	4.417505	0.0000

R-squared	0.155475	Mean dependent var	7718.111
Adjusted R-squared	0.147508	S.D. dependent var	3577.000
S.E. of regression	3302.662	Akaike info criterion	19.06119
Sum squared resid	1.16*E* + 09	Schwarz criterion	19.11086
Log likelihood	−1027.304	F-statistic	19.51435
Durbin–Watson stat	2.012384	Prob(F-statistic)	0.000024

TABLE 10.5 Estimating a Transformed Rent–Income Relationship, $\mathrm{var}(\varepsilon_i) = \sigma^2 X_i^2$

Dependent Variable: RENT/INCOME
Method: Least Squares
Sample: 1 108
Included observations: 108

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	**0.085679**	**0.016701**	**5.130303**	**0.0000**
1/INCOME	**4811.862**	**322.2745**	**14.93094**	**0.0000**

R-squared	0.677746	Mean dependent var	0.291701
Adjusted R-squared	0.674706	S.D. dependent var	0.171429
S.E. of regression	0.097774	Akaike info criterion	−1.793980
Sum squared resid	1.013325	Schwarz criterion	−1.744311
Log likelihood	98.87494	F-statistic	222.9331
Durbin-Watson stat	1.900821	Prob(F-statistic)	0.000000

Checking Understanding

- If we have the correct model of heteroskedasticity, then OLS with the transformed data should be homoskedastic.

$$\frac{rent}{income_i} = \beta_0 \frac{1}{income_i} + \beta_1 + v_i$$

- We can apply either a White test or a Breusch–Pagan test for heteroskedasticity to the model with the transformed data.

Checking Understanding (cont.)

- To run the White test, we regress

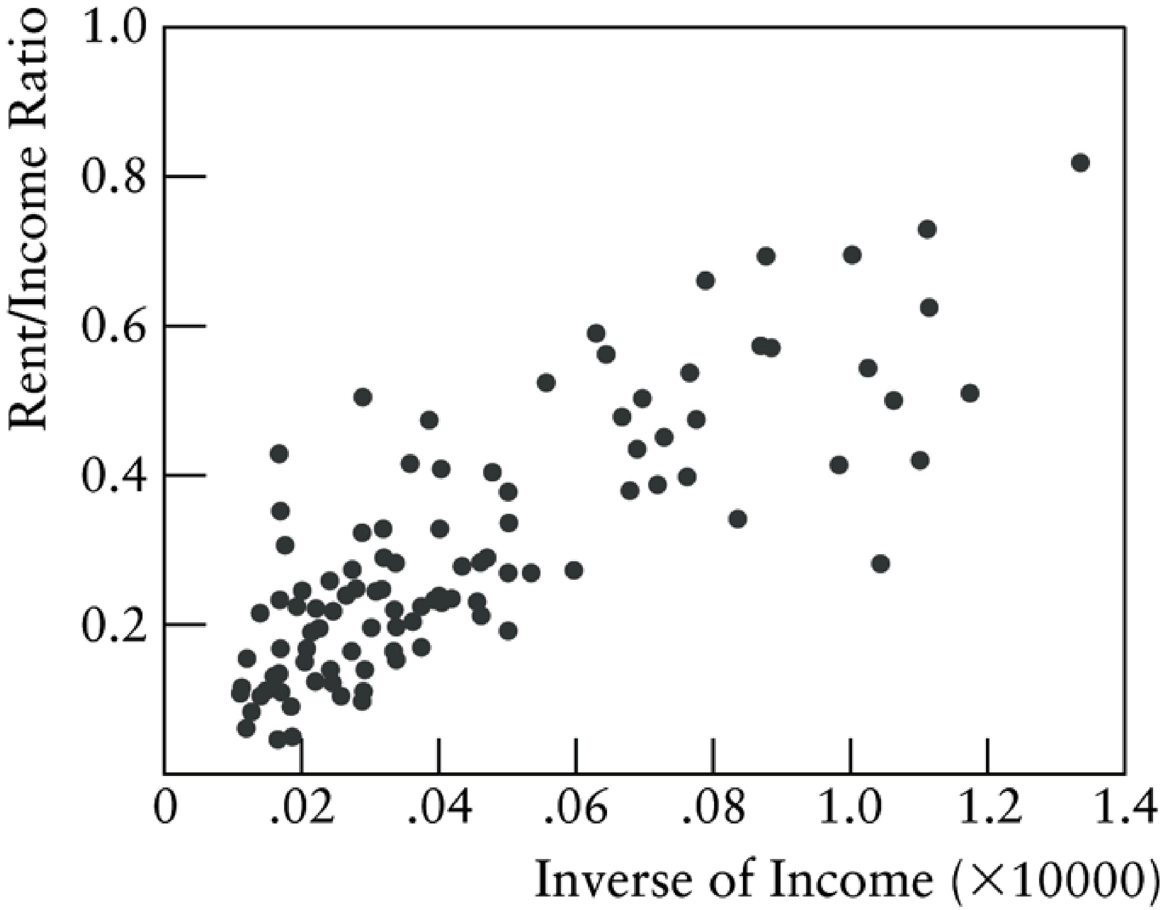
$$e_i = \alpha_0 + \alpha_1 \frac{1}{income_i} + \alpha_2 \frac{1}{income_i^{\,2}} + \eta_i$$

- $nR^2 = 7.17$
- The critical value at the 0.05 significance level for a Chi-square statistic with 2 degrees of freedom is 5.99
- We reject the null hypothesis.

GLS: An Example

- Our initial guess: $Var(\varepsilon_i) = \sigma^2 \cdot income_i^2$
- This guess didn't do very well. Can we do better?
- Instead of blindly guessing, let's try looking at the data first.

Figure 10.4 The Rent–Income Ratio Plotted Against the Inverse of Income



GLS: An Example

- We seem to have overcorrected for heteroskedasticity.
- Let's try

$$Var(\varepsilon_i) = \sigma^2 \cdot income_i$$

$$\frac{rent}{\sqrt{income}_i} = \beta_0 \frac{1}{\sqrt{income}_i} + \beta_1 \sqrt{income_i} + v_i$$

TABLE 10.6 Estimating a Second Transformed Rent–Income Relationship, $\operatorname{var}(\varepsilon_i) = \sigma^2 X_i$

Dependent Variable: RENT/(INCOME$^{0.5}$)
Method: Least Squares
Sample: 1 108
Included observations: 108

Variable	Coefficient	Std. Error	t-Statistic	Prob.
1/(INCOME$^{0.5}$)	5085.513	411.4241	12.36076	0.0000
INCOME$^{0.5}$	0.073963	0.014269	5.183325	0.0000

R-squared	0.121359	Mean dependent var	44.92013
Adjusted R-squared	0.113070	S.D. dependent var	17.41882
S.E. of regression	16.40451	Akaike info criterion	8.451335
Sum squared resid	28525.45	Schwarz criterion	8.501004
Log likelihood	−454.3721	Durbin–Watson stat	1.964951

- For example, consider the relationship

$$rent_i = \beta_0 + \beta_1 income_i + \varepsilon_i$$

- We are concerned that $Var(\varepsilon_i)$ may vary with income.
- We need to make an assumption about how $Var(\varepsilon_i)$ varies with income.

- If we have the correct model of heteroskedasticity, then OLS with the transformed data should be homoskedastic.

$$\frac{rent}{income_i} = \beta_0 \frac{1}{income_i} + \beta_1 + v_i$$

- Using a White test, we reject the null hypothesis of homoskedasticity of the model with transformed data.

Feasible GLS (cont.)

- To begin, we need to assume some model for the heteroskedasticity.

- Then we estimate the parameter/s of the model.

Feasible GLS (cont.)

- One reasonable model for the error terms could be that the variance is proportional to some power of the explanator.

$$Var(\varepsilon_i) = \sigma^2 X_i^h$$

- For example, in the rent-income example, we tried both

$$Var(\varepsilon_i) = \sigma^2 income_i^2 (h = 2)$$

and $Var(\varepsilon_i) = \sigma^2 income_i (h = 1)$

TABLE 10.7 ln(*Squared Residual*) vs. ln(*Income*) Following *RENT* vs. *INCOME* by OLS

Dependent Variable: LOG (e^2)
Method: Least Squares
Sample: 1 108
Included observations: 108

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**2.083771**	**2.994793**	**0.695798**	**0.4881**
LOG(INCOME)	**1.216408**	**0.290830**	**4.182539**	**0.0001**

R-squared	0.141656	Mean dependent var	14.58387
Adjusted R-squared	0.133559	S.D. dependent var	2.142308
S.E. of regression	1.994121	Akaike info criterion	4.236629
Sum squared resid	421.5110	Schwarz criterion	4.286298
Log likelihood	−226.7780	F-statistic	17.49363
Durbin–Watson stat	1.843326	Prob(F-statistic)	0.000060

White Robust Standard Errors

- Heteroskedasticity is a common problem.
- We may not always be happy making the FGLS assumptions, especially if we don't really need that extra efficiency.
- OLS is unbiased. OLS may yield a sufficiently small standard error to allow reasonably precise estimates.

White Robust Standard Errors (cont.)

- The main problem in applying OLS under heteroskedasticity is that our *e.s.e.* formula is incorrect
- White's brilliant idea: use OLS and fix the estimated standard errors

White Robust Standard Errors (cont.)

- For OLS with an intercept and a single explanator, $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, we have derived the formula for the *e.s.e*:

$$e.s.e.(\hat{\beta}_1) = \sqrt{\frac{\Sigma e_i^2}{(n-2)\Sigma x_i^2}}$$

- However, we really used the homoskedasticity assumption only to simplify this formula.

White Robust Standard Errors (cont.)

- White Heteroskedastic Consistent standard errors (commonly called “robust” standard errors) correct for possible heteroskedasticity
- Software packages often provide White *e.s.e.*’s as an option
- If errors are homoskedastic, White *e.s.e.*’s are less efficient than OLS *e.s.e.*’s

TABLE 10.8 OLS Estimates of the Rent–Income Relationship with Robust Standard Errors

Dependent Variable: RENT
Method: Least Squares
Sample: 1 108
Included observations: 108
White Heteroskedasticity-consistent Standard Errors and Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5455.483	403.2469	13.52889	0.0000
INCOME	0.063568	0.014759	4.307218	0.0000

R-squared	0.155475	Mean dependent var	7718.111
Adjusted R-squared	0.147508	S.D. dependent var	3577.000
S.E. of regression	3302.662	Akaike info criterion	19.06119
Sum squared resid	1.16E + 09	Schwarz criterion	19.11086
Log likelihood	−1027.304	F-statistic	19.51435
Durbin-Watson stat	2.012384	Prob(F-statistic)	0.000024

White Robust Standard Errors

- Applying White estimated standard errors is a very easy fix for possible heteroskedasticity.
- Some economists simply use White *e.s.e.*'s routinely.
- This fix comes with a cost in efficiency:
  - OLS is not BLUE under heteroskedasticity.
  - White *e.s.e.*'s are inefficient under homoskedasticity.

White Robust Standard Errors (cont.)

- Note: It is CRUCIAL, when you present your own results, that you clarify which *e.s.e.* you have used. If you do use White standard errors, you MUST say so.

- For example, many tables of results include the footnote "White standard errors in parentheses" or "Robust standard errors in parentheses."

Heteroskedasticity

- Heteroskedasticity is not, in practice, a burdensome complication.
- Econometricians have easy-to-apply tests to detect heteroskedasticity (White tests, Breusch–Pagan tests, or Goldfeld–Quandt tests).
- If there is heteroskedasticity, econometricians have a number of options available.

Heteroskedasticity (cont.)

- If econometricians know the exact nature of the heteroskedasticity (i.e. if they know the d_i), then they can simply divide all variables by d_i and apply GLS.
- If the d_i are unknown, but econometricians are willing to make some assumptions about their functional form, then the d_i can be estimated by FGLS.

Heteroskedasticity (cont.)

- If econometricians are unwilling to make assumptions about the nature of the heteroskedasticity, they can implement OLS to get unbiased, but inefficient, estimates.

- Then they must correct the estimated standard errors using White Robust Standard Errors.