Introductory Econometrics I

Heteroskedasticity

Yingjie Feng

School of Economics and Management

Tsinghua University

May 10, 2024

Review: Assumptions for OLS

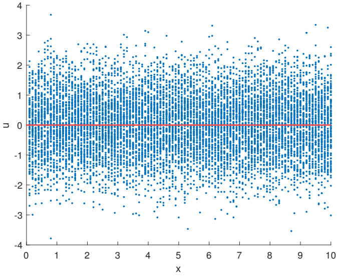
- Recall the Gauss-Markov Assumptions for OLS regression:
  - MLR.1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
  - MLR.2: random sampling from the population
  - MLR.3: no perfect collinearity in the sample
  - MLR.4: $\mathbb{E}[u|x_1, ..., x_k] = \mathbb{E}[u] = 0$ (exogenous explanatory variables)
  - MLR.5: $\mathbb{V}[u|x_1, ..., x_k] = \mathbb{V}[u] = \sigma^2$ (**homoskedasticity**)
- Under MLR.1-MLR.5, OLS is BLUE (and asymptotically efficient) in a broad class of estimators
- Add normality (MLR.6): tests and confidence intervals are **exact** given any sample size
- Without normality (MLR.6): the usual test statistics and CIs are **approximately** valid in large samples

Relaxing Assumptions for OLS

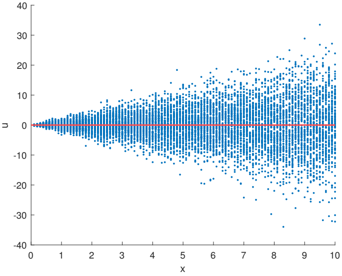
- Now we will relax some assumptions for OLS
  - Relax MLR.5: homoskedasticity fails (heteroskedasticity)
  - Relax MLR.4: $\mathbb{E}[u|x_1, \cdots, x_k] \neq 0$ (endogeneity)
  - Relax MLR.2: non-i.i.d. data (e.g., time series data)
  - Relax MLR.1: nonlinear models (e.g., limited dependent variable models)
- Today, we focus on relaxing Assumption MLR.5
  - $\mathbb{V}[u|\mathbf{x}]$ depends on $\mathbf{x} = (x_1, \cdots, x_k)$

Heteroskedasticity

- Homoskedasticity: $\mathbb{V}[u|\mathbf{x}] = \sigma^2$
- Heteroskedasticity: $\mathbb{V}[u|\mathbf{x}] = \sigma^2(\mathbf{x})$



Homoskedasticity



Heteroskedasticity

Outline

1. Consequences of Heteroskedasticity for OLS
2. Heteroskedasticity-Robust Inference after OLS Estimation
3. Testing for Heteroskedasticity*
4. Weighted Least Squares
5. The Linear Probability Model Revisited

Consequences of Heteroskedasticity for OLS

- We drop MLR.5 and act as if we know **nothing** about

$$\mathbb{V}[u|x_1, ..., x_k] = \mathbb{V}[u|\mathbf{x}]$$

- OLS is still **unbiased and consistent**
  - Recall unbiasedness and consistency only rely on MLR.1 to MLR.4.
  - **Important** conclusion: Heteroskedasticity does not cause bias or inconsistency in $\hat{\beta}_j$s
- R^2 and $\bar{R}^2$ are still valid as goodness-of-fit measures and remain consistent estimators of the population R-squared:

$$\rho^2 = 1 - \frac{\sigma_u^2}{\sigma_y^2}, \quad \sigma_u^2 = \mathbb{V}[u], \ \sigma_y^2 = \mathbb{V}[y]$$

  - SSR/n or $SSR/(n-k-1)$ are consistent for σ_u^2 regardless of $\mathbb{V}[u|\mathbf{x}] = \mathbb{V}[u]$. SST/n and $SST/(n-1)$ are consistent for σ_y^2.

Consequences of Heteroskedasticity for OLS

- But if $\mathbb{V}[u|\mathbf{x}]$ depends on $\mathbf{x}$ ("**heteroskedasticity**"), OLS is not BLUE
  - In principle, it is possible to find unbiased estimators that have smaller variances than OLS estimators.
- And more importantly, with heteroskedasticity,
  - The variance formula $\mathbb{V}[\hat{\beta}_j] = \frac{\sigma^2}{SST_j(1-R_j^2)}$ assuming homoskedasticity is **wrong**
  - The standard errors based on this formula are wrong
  - The t statistics and confidence intervals that use these standard errors cannot be trusted.
  - Joint hypotheses tests using the usual F statistic are no longer valid as well
  - This is true even in large samples

Consequences of Heteroskedasticity for OLS

- Without MLR.5, there are still good reasons to use OLS, but we need to modify the usual test statistics to make them valid in the presence of heteroskedasticity.

- We are **not** talking about a new estimation method. It is still OLS estimation to obtain the $\hat{\beta}_j$.

- But we need to use **heteroskedasticity-robust inference** after OLS estimation.

- **Reminder**: We are talking about **conditional** heteroskedasticity of y or u given $\mathbf{x}$. The unconditional variance must be a constant.

Outline

Heteroskedasticity-Robust Inference after OLS

- Standard errors and all test statistics can be modified to be valid in the presence of **heteroskedasticity of unknown form**.
  - Homoskedasticity (MLR.5) can be covered as a special case
- Most regression packages include an option with OLS estimation that computes **heteroskedasticity-robust standard errors**, which then produces **heteroskedasticity-robust** t **statistics** and **heteroskedasticity-robust confidence intervals**.
- In Stata, the general command is:
  - `reg y x1 x2 ... xk, robust`

  where "robust" means "robust to heteroskedasticity of any form".

Heteroskedasticity-Robust Inference after OLS

- Consider the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad \mathbb{E}[u_i|x_i] = 0, \;\; \mathbb{V}[u_i|x_i] = \sigma_i^2$$

- OLS estimator

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- The “correct” variance formula would be

$$\mathbb{V}[\hat{\beta}_1|\mathbf{x}] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \mathbb{V}[u_i|\mathbf{x}]}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2}$$

- The “correct” variance estimator (White, 1980)

$$\hat{\mathbb{V}}[\hat{\beta}_1|\mathbf{x}] = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2}$$

- Intuition: $\mathbb{E}[u_i^2|\mathbf{x}] = \sigma_i^2$

Heteroskedasticity-Robust Inference after OLS

- More generally, for multiple regression,

$$\hat{\mathbb{V}}[\hat{\beta}_j|\mathbf{x}] = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{[\sum_{i=1}^{n} \hat{r}_{ij}^2]^2}$$

  - $\hat{r}_{ij}$ is the residual from regressing x_j on all other explanatory variables
  - $\hat{u}_i$ is the residual from regressing y on $x_1, \cdots, x_k$

- Heteroskedasticity-robust standard error:

$$(\hat{\mathbb{V}}[\hat{\beta}_j|\mathbf{x}])^{1/2}$$

  - An estimate of the standard deviation of $\hat{\beta}_j$ which is valid regardless of heteroskedasticity

- Robust t statistic

$$t = \frac{\text{estimate} - \text{hypothesized value}}{\text{robust standard error}}$$

Heteroskedasticity-Robust Inference after OLS

- **Question:** If we can compute standard errors that work with or without MLR.5, how come we bother with the usual standard errors at all?
- **Answers:**
  1. Tradition (not necessarily a good answer).
  2. A (slightly) better answer: The heteroskedasticity-robust test statistics and CIs only have asymptotic justification, even if MLR.1-MLR.6 hold.
- With small sample sizes, the heteroskedasticity-robust statistics need not be well behaved. Sometimes they can have more bias than the usual statistics.
- Some researchers, especially with large sample sizes, only report the heteroskedasticity-robust statistics.

Heteroskedasticity-Robust Inference after OLS

- Use WAGE1.DTA:

$$\begin{array}{rcccccccc}
\widehat{lwage} & = & 0.481 & - & .344 & female + & .009 & exper + & .091 & educ \\
 & & (.1050) & & (.0377) & & (.0014) & & (.0071) & \\
 & & [.1174] & & [.0374] & & [.0015] & & [.0081] & \\
n & = & \multicolumn{8}{l}{526,\ R^2 = .353,\ \bar{R}^2 = .349}
\end{array}$$

- In this example, the robust standard errors in brackets [] are slightly larger than the usual standard errors in parentheses () except for $female$, but this has little consequence. (CIs are slightly wider, t statistics slightly lower.)

Heteroskedasticity-Robust Inference after OLS

```
reg lwage educ female exper

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(3, 522)       =     94.75
       Model |  52.2939096         3  17.4313032   Prob > F        =    0.0000
    Residual |  96.0358418       522  .183976708   R-squared       =    0.3526
-------------+----------------------------------   Adj R-squared   =    0.3488
       Total |  148.329751       525   .28253286   Root MSE        =    .42893

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0912897   .0071232    12.82   0.000     .0772962    .1052833
      female |  -.3435967   .0376668    -9.12   0.000    -.4175939   -.2695996
       exper |   .0094139   .0014493     6.50   0.000     .0065667     .012261
       _cons |   .4808357   .1050163     4.58   0.000     .2745292    .6871421
------------------------------------------------------------------------------
```

Heteroskedasticity-Robust Inference after OLS

```
. reg lwage educ female exper, robust

Linear regression                               Number of obs     =        526
                                                F(3, 522)         =      80.28
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3526
                                                Root MSE          =     .42893
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0912897	.00809	11.28	0.000	.0753968	.1071827
female	-.3435967	.0374432	-9.18	0.000	-.4171546	-.2700389
exper	.0094139	.0014749	6.38	0.000	.0065164	.0123113
_cons	.4808357	.1174382	4.09	0.000	.2501262	.7115452

Heteroskedasticity-Robust Inference after OLS

- It is sometimes **incorrectly** claimed that the heteroskedasticity-robust standard errors for OLS are always larger than the usual standard errors.

- As we have seen in the previous example, it can go either way, even within the same regression.
  - The robust s.e. on $female$ is .0374, below the usual s.e. .0377.

- Remember, at this point we do not know whether the error in the equation is heteroskedastic.
  - We can compute the heteroskedasticity-robust standard errors in either case.
  - The large difference in some standard errors is suggestive, but it does not constitute a formal test.

Heteroskedasticity-Robust Inference after OLS

- The usual F statistic for testing multiple hypotheses can also be modified to allow for unknown heteroskedasticity.
- In Stata
  - `reg y x1 x2 x3 ... xk, robust`
  - `test x1 x2 x3`
- This will automatically compute a heteroskedasticity-robust joint test of x_1, x_2, and x_3.
- In the following example the robust test rejects at the 5% level while the nonrobust one is not even close.

Heteroskedasticity-Robust Inference after OLS

Use APPLE.DTA

```
. qui reg ecolbs ecoprc regprc lfaminc numlt5 num5_17 num18_64 numgt64 age

. test lfaminc numlt5 num5_17 num18_64 numgt64 age

 ( 1)  lfaminc = 0
 ( 2)  numlt5 = 0
 ( 3)  num5_17 = 0
 ( 4)  num18_64 = 0
 ( 5)  numgt64 = 0
 ( 6)  age = 0

       F(  6,   651) =    1.25
            Prob > F =    0.2764

. qui reg ecolbs ecoprc regprc lfaminc numlt5 num5_17 num18_64 numgt64 age, robust

. test lfaminc numlt5 num5_17 num18_64 numgt64 age

 ( 1)  lfaminc = 0
 ( 2)  numlt5 = 0
 ( 3)  num5_17 = 0
 ( 4)  num18_64 = 0
 ( 5)  numgt64 = 0
 ( 6)  age = 0

       F(  6,   651) =    2.43
            Prob > F =    0.0250
```

Outline

Testing for Heteroskedasticity

- Before the discovery of heteroskedasticity-robust inference, a common approach is to abandon OLS and use a new estimator
  - First test for heterskedasticity
  - If it was found (at a sufficiently small significance level), use "weighted least squares".
- But with simple adjustments to the usual OLS test statistics, there is less of a case for even testing for heteroskedasticity.
  - Still use the OLS estimators but fix the standard errors
- If you do have direct interest in heteroskedasticity, many tests are available (e.g., Breusch-Pagan test and White test). Read Section 8-3 of Textbook.

Outline

Weighted Least Squares

- If heteroskedasticity is present and we think $\mathbb{E}[y|\mathbf{x}]$ has been properly modeled, we might want to improve on OLS, as OLS is no longer BLUE (because Assumption MLR.5 fails).
- To ensure we get a better estimator than OLS we need to know the form of the heteroskedasticity.
- Even if we do not correctly specify the form of heteroskedasticity, sometimes we can do better than OLS by using an incorrect variance function. See Section 8-4c in Wooldridge for a discussion of weighted least squares.

Weighted Least Squares

- Consider the linear regression model satisfying MLR.1-MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- Assume the heteroskedasticity takes the following form

$$\mathbb{V}[u|\mathbf{x}] = \sigma^2 h(\mathbf{x})$$

- For each observation i,

$$\sigma_i^2 = \mathbb{V}[u_i|\mathbf{x}_i] = \sigma^2 h(\mathbf{x}_i) = \sigma^2 h_i$$

- Get a model with homoskedastic errors

$$\frac{y_i}{\sqrt{h_i}} = \beta_0 \frac{1}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + ... + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}.$$

- Note that

$$\mathbb{E}[u_i/\sqrt{h_i}|\mathbf{x}] = 0, \quad \mathbb{V}[u_i/\sqrt{h_i}|\mathbf{x}] = \sigma^2$$

Weighted Least Squares

- Therefore, we can run the following regression

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^*$$
$$y_i^* = \frac{y_i}{\sqrt{h_i}}, \quad x_{i0}^* = \frac{1}{\sqrt{h_i}}, \quad \cdots, \quad x_{ik}^* = \frac{x_{ik}}{\sqrt{h_i}}, \quad u_i^* = \frac{u_i}{\sqrt{h_i}}$$

- The estimators from the above regression is denoted by $\hat{\beta}_j^*$, which is different from the original OLS
- $\hat{\beta}_j^*$ are called **weighted least squares** estimator since they are solutions to

$$\min_{b_0, \cdots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2 / h_i$$

  - h_i^{-1} plays the role of weights: h_i is greater, then observation i is more noisy, and thus contributes less to the regression
- This transformed model satisfies MLR.1-MLR.5, so $\hat{\beta}_j^*$ is BLUE
- $\hat{\beta}_j^*$ are examples of **generalized least squares estimators (GLS)**

Generalized Least Squares

- Sometimes we have good reasons to use GLS and h_i is known
- Examples: relation between amount a worker contributes to a pension plan and other factors such as the earning of the plan
  - Regression model for "individual-level" data

$$contrib_{i,e} = \beta_0 + \beta_1 earns_{i,e} + u_{i,e},$$

    i stands for firm i and e stands for employee e

  - But you only have "firm-level" data that are averaged across worker in each firm i (size is m_i):

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{earns}_i + \bar{u}_i, \quad \bar{u}_i = \frac{1}{m_i} \sum_{e=1}^{m_i} u_{i,e}$$

  - Suppose that $u_{i,e}$ is i.i.d. over i and e and independent of $earns_{i,e}$:

$$\mathbb{V}[u_{i,e}] = \sigma^2, \quad \mathbb{V}[\bar{u}_i] = \frac{\sigma^2}{m_i}, \quad h_i = \frac{1}{m_i}$$

Feasible Generalized Least Squares

- In practice, h_i is unknown. The previous regression is infeasible
- We have to estimate h_i. For example, assume

$$\mathbb{V}(u|\mathbf{x}) = \mathbb{E}[u^2|\mathbf{x}] = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta x_k)$$

  - Taking exponential guarantees $\mathbb{V}[u|\mathbf{x}]$ is nonnegative
- Then, we can write

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta x_k) v$$

  v is independent of $\mathbf{x}$ and has mean equal to 1
- Taking logs, we can equivalently express the model as

$$\log u^2 = \alpha_0 + \delta_0 + \delta_1 x_1 + ... + \delta x_k + e$$

  $\alpha_0 = \log \sigma^2$, $e = \log v$.

Feasible Generalized Least Squares

- Given the model for u^2

$$\begin{aligned}\log u^2 &= \alpha_0 + \delta_0 + \delta_1 x_1 + ... + \delta x_k + e \\ &= \delta_0' + \delta_1 x_1 + ... + \delta_k x_k + e, \quad \delta_0' = \alpha_0 + \delta_0\end{aligned}$$

- We can obtain the feasible GLS estimator by
  1. Regress y_i on $1, x_1, \cdots, x_k$ to get the residual $\hat{u}_i$
  2. Regress $\log \hat{u}_i^2$ on 1, x_1, $\cdots$, x_k to get the fitted value $\hat{g}_i = \hat{\delta}_0' + \hat{\delta}_1 x_{i1} + \cdots + \hat{\delta}_k x_{ik}$
  3. $\hat{\sigma}_i^2 = \exp(\hat{g}_i)$
  4. Use $1/\hat{\sigma}_i^2$ as weights in weighted least squares (i.e., $y_i^* = y_i/\hat{\sigma}_i$ and $x_{ij}^*/\hat{\sigma}_i$)
  5. Estimate β_j by weighted least squares (WLS)

- Note: here the new error term $u_i^* = u_i/\sigma_i$ has conditional variance equal to 1

Feasible Generalized Least Squares

- If the heteroskedasticity function is correctly specified
  - Feasible generalized least squares (FGLS) is consistent and more efficient than OLS (recall OLS is also consistent even when MLR.5 fails)
- If the heteroskedasticity function is incorrectly specified ($\mathbb{V}[u|\mathbf{x}] \neq \sigma^2 h(\mathbf{x})$)
  - FGLS may not be more efficient
  - We still have to use heteroskedasticity-robust standard errors for inference after FGLS
- If $\mathbb{E}[u|\mathbf{x}] = 0$ is true, OLS and FGLS should be close in large samples; if not, it might suggest $\mathbb{E}[u|\mathbf{x}] \neq 0$

Outline

Linear Probability Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

- We noted earlier that if y is binary, there must be heteroskedasticity except in the special case that no x_j affects y

$$\mathbb{V}[u|\mathbf{x}] = \mathbb{V}[y|\mathbf{x}] = p(\mathbf{x})(1 - p(\mathbf{x})), \quad p(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$$

- The simplest solution is to use heteroskedasticity-robust inference after OLS.
- Even for binary y, the Stata command
  - `reg y x1 x2 ... xk, robust`

  produces heteroskedasticity-robust inference.