LECTURE #10

Econometrics I

HETEROSKEDASTICITY

Jiri Kukacka, Ph.D.
Institute of Economic Studies, Faculty of Social Sciences, Charles University

Summer semester 2024, April 30

In the previous lecture #9

- We extended the analysis with dummy variables:

multiple dummy categories	$y = \beta_0 + \beta_1 x + \boxed{\beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3} + u$
ordinal information	$Q = 0, 1, 2, 3$
interactions among dummies	$y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \boxed{\beta_4 D_1 D_2} + u$
slope dummy variables	$y = \beta_0 + \beta_1 x + \beta_2 D + \boxed{\beta_3 D x} + u$

- We derived the **Chow test** for differences between groups:

$$F = \frac{SSR_P - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \frac{n - 2(k+1)}{k+1} \sim F_{k+1, n-2(k+1)}.$$

- We introduced the **linear probability model (LPM)** for a binary dependent variable and discussed its shortcomings:

$$p(X) \equiv \mathrm{P}(y = 1|X) = \mathbb{E}(y|X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

- Readings for lecture #10:
  - Chapter 8 (**8.4, sections 'What If...' & 'Prediction and...'**, and **8.5 mandatory after lecture**)

Outline

Heteroskedasticity-robust inference
- Consequences for OLS
- White's standard errors
- Heteroskedasticity-robust LM test

Testing for heteroskedasticity

Weighted least squares estimation
- With a known heteroskedasticity form: WLS
- With an unknown heteroskedasticity form: FGLS

Outline

Outline

Heteroskedasticity: Consequences for OLS

- **MLR.5 Homoskedasticity:** The error u has the same variance given any values of the independent variables, i.e.,

$$\text{Var}(u|x_1, \ldots, x_k) = \sigma^2 \mathbb{I}.$$

- Violation of homoskedasticity is called **heteroskedasticity**.

Consequences:

1. OLS remains unbiased and consistent (under MLR.1–4).
   - estimated coefficients remain unaffected.
   - coefficients of determination R^2 and $\bar{R}^2$ remain unaffected.
2. True variance of the $\hat{\beta}^{OLS}$ distribution increases.
   - because the heteroskedastic error term explains a larger proportion of fluctuations of the dependent variable.

   $\Rightarrow$ OLS is no longer ~~B~~LUE, even not asymptotically efficient.

   $\Rightarrow$ OLS more likely to misestimate the true β.

Heteroskedasticity: Consequences for OLS

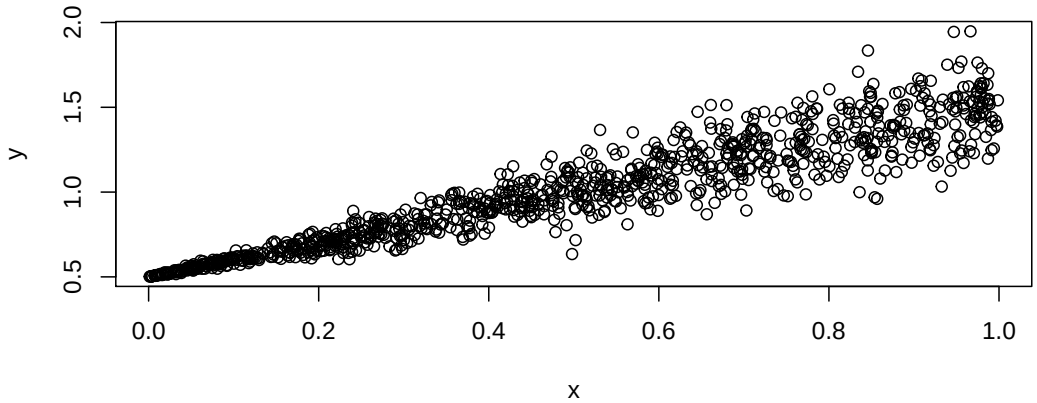
3. But (!) estimators of $\text{Var}(\hat{\beta}_j)$ are biased, usually down (as long as positive correlation between σ_i^2 and $(x_{ij} - \overline{x}_j)^2$).
   - increase of the (true) variance is, however, 'masked' by OLS because it assumes a homoskedastic error.
   - OLS thus attributes the impact of the heteroskedastic error to the independent variables.

   $\Rightarrow$ standard errors tend to be smaller under heteroskedasticity, and statistical inference becomes unreliable and incorrect:
   - $\Rightarrow$ t statistics and CIs invalid even for large samples!
   - $\Rightarrow$ also F statistics and LM statistics invalid, i.e., no longer asymptotically F and χ^2 distributed!

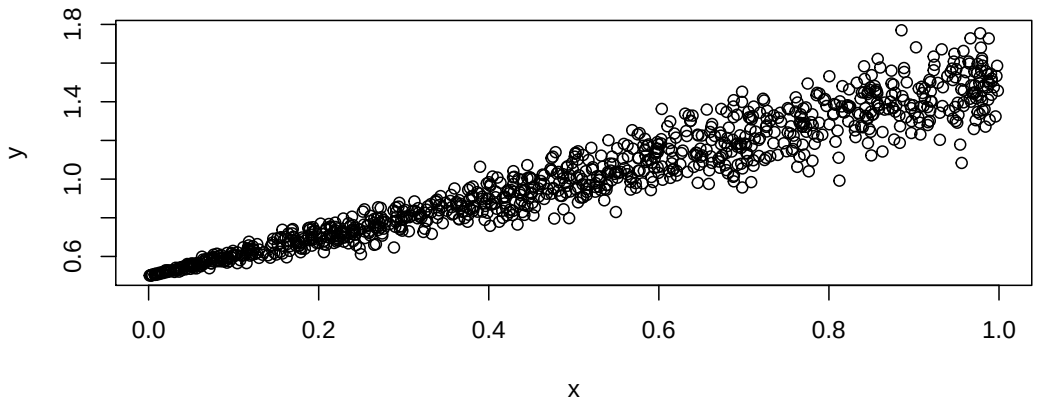
- Fortunately, the OLS standard errors can be modified to be asymptotically valid under MLR.1–4, i.e., without MLR.5.

Heteroskedasticity: Illustration in R

(a) $y = 0.5 + 1x + u, \quad u \sim N(0, sd = 0.2x)$



(b) $y = 0.5 + 1x + u, \quad u \sim N(0, sd = 0.2 \log |1 + x|)$



Outline

Heteroskedasticity-robust inference

- Consider a simple regression population model

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

- Assume MLR.1 through MLR.4 hold but

$$\text{Var}(u_i|x_i) = \sigma_i^2,$$

  i.e., variance of u_i depends on x_i.

- Using the rewritten OLS estimator

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

  and under MLR.1 through MLR.4, we have

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n \sigma_i^2 (x_i - \bar{x})^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sum_{i=1}^n \sigma_i^2 (x_i - \bar{x})^2}{SST_x^2}.$$

- Under MLR.5, this reduces to

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}.$$

White's standard errors

- In the 1980s, Halbert White (following the ideas of Friedhelm Eicker and Peter Huber from the 1960s) proposed a **valid estimator** of the OLS estimator variance, which works without the MLR.5 assumption.
- This leads to the **heteroskedasticity-consistent/robust standard error** for $\hat{\beta}_j$ (or, giving credit to the main contributors, either White's or White-Huber-Eicker standard errors) or sometimes just **robust standard errors**.
- In empirical practice, robust standard errors are **often larger** than the usual OLS standard errors.
- Provided automatically by most econometric packages.

▸ Formula

White's standard errors

- Intuition behind the White's standard errors is well understandable from their simple regression form:

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} \hat{u}_i^2 (x_i - \bar{x})^2}{SST_x^2}$$

- As we know, little x sample variation (or multicollinearity) can cause large standard errors.
- For **large samples**, the White's standard errors are asymptotically valid for any form of heteroskedasticity.
- Once we have the robust standard errors, we can construct the t **statistics and confidence intervals** in the same way as for the usual OLS standard errors and justify them **even without MLR.5–6**.

Why bother with the usual OLS std. errors, then?

- White's standard errors work **only for large samples**, i.e., due to the LLN and CLT.
- It is thus suggested to use them only with at least 100 degrees of freedom, i.e., $n - k - 1 \geq 100$.
- Even for large samples, they are only asymptotically valid; no statements are made about bias, consistency, or efficiency.
- OTOH, if the assumptions MLR.1–6 are met, OLS t statistics have exact t-distribution regardless of the sample size.

Outline

Heteroskedasticity-robust LM test

As the heteroskedasticity-robust F test is rather complicated and not necessarily provided by the econometric packages, it is more convenient to introduce the robust version of the LM test for the exclusion restrictions testing. It follows these steps:

1. Obtain residuals $\tilde{u}$ from the restricted model.
2. Regress each of the excluded independent variables on all included independent variables. For q excluded variables, this gives us q sets of residuals $(\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_q)$.
3. Introduce q new variables constructed as products of residuals $\tilde{r}_j$ and residuals $\tilde{u}$.
4. Run the regression of 1 on $\tilde{r}_1\tilde{u}, \ldots, \tilde{r}_q\tilde{u}$ without an intercept. The heteroskedasticity-robust $\boxed{LM = n - SSR_1}$, where SSR_1 is the SSR from this final regression. Under the null hypothesis, the LM statistics is approximately χ^2_q distributed.

Outline

Testing for heteroskedasticity

- Before applying White's standard errors (or switching to the generalized least squares estimation), we should check whether the homoskedasticity assumption is, in fact, violated.
- Among the mostly utilized are the following two tests:
  - Breusch-Pagan test
  - White test
- Note that we assume MLR.1–4 to hold for both tests.
- Otherwise, especially without MLR.4, the tests may be seen as a general test for functional misspecification; we will discuss this in the next lecture #11.

Breusch-Pagan test

- **Breusch-Pagan test** is built on a simple assumption of linear dependence of the error variance on the independent variables.
- It can be performed in the following steps:
  1. estimate the OLS regression and keep the residuals $\hat{u}$,
  2. run the regression of $\hat{u}^2$ on all independent variables:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_x + v, \tag{1}$$

  3. apply the F test or the LM test with the **null hypothesis of homoskedasticity**, i.e., $H_0 : \delta_1 = 0, \ldots, \delta_k = 0$
     (both statistics would have asymptotic justification, even though $\hat{u}^2$ cannot be normally distributed).

White test

- **White test** assumes a very general form of heteroskedasticity: the error variance may be dependent on the **independent variables**, their **squares**, and **pairwise products**.
- Such a combination flexibly covers various forms of heteroskedasticity.
- For the case of two independent variables, we extend (1) to

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_1^2 + \delta_4 x_2^2 + \delta_5 x_1 x_2 + v.$$

- Number of parameters thus grows rapidly with k, which leads to an adjusted version of the test based on

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + v.$$

- In the same way as for the B-P test, we apply the F test or the LM test with the **null hypothesis of homoskedasticity**, i.e., $H_0 : \delta_1 = 0, \ldots, \delta_5 = 0$ or $\delta_1 = 0, \delta_2 = 0$, respectively.
- B-P test can be thus seen as a special case of the White test.

Outline

WLS: Why another method?

- **Weighted least squares (WLS)** estimation is a historically older method of treating heteroskedasticity compared to White's standard errors.
- If we have a correctly specified form of heteroskedasticity, WLS is **unbiased** and **more efficient** than OLS, and it leads to t-distributed t statistics and F-distributed F statistics only under MLR.1 through MLR.4 (it is, in fact, **BLUE**).
- Compare to White's standard errors being just 'valid.'

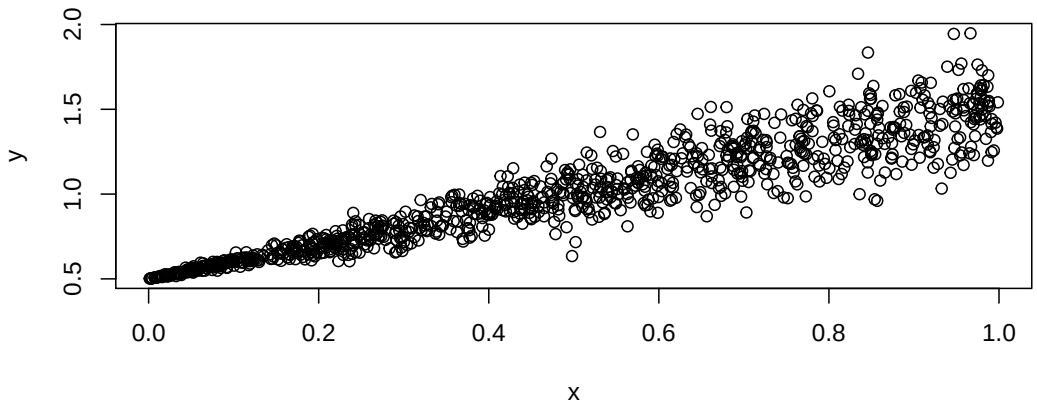
Outline

Heteroskedasticity form is known

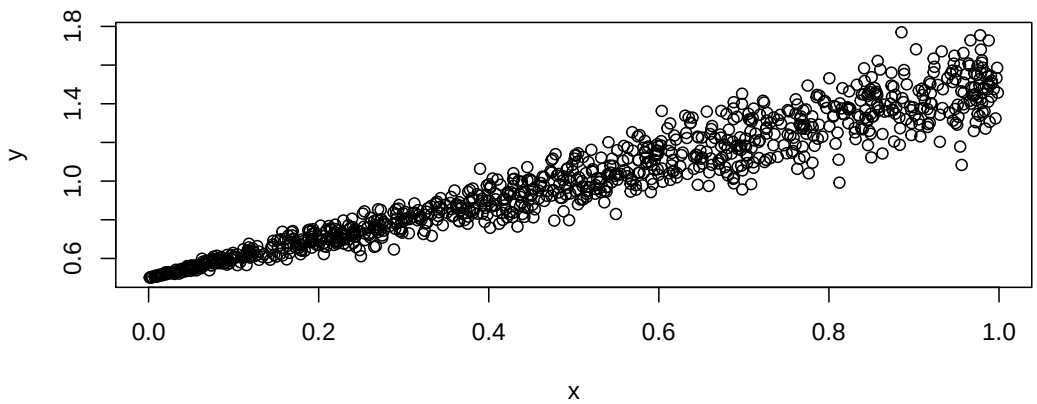
- Let us assume
$$\mathrm{Var}(u|X) = \sigma^2 h(X),$$
where $h(X) \equiv h$ is a function of the independent variables, determining the heteroskedasticity form.
- Let us define a new model as
$$\frac{y_i}{\sqrt{h_i}} = \frac{\beta_0}{\sqrt{h_i}} + \beta_1 \frac{x_{i1}}{\sqrt{h_i}} + \ldots + \beta_k \frac{x_{ik}}{\sqrt{h_i}} + \frac{u_i}{\sqrt{h_i}}$$
or, alternatively, written as
$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + u_i^*,$$
where $x_{i0}^* = \frac{1}{\sqrt{h_i}}$, i.e., there is no intercept in the new model.

Heteroskedasticity: Illustration in R

(c) $y = 0.5 + 1x + u, \quad u \sim N(0, sd = 0.2x)$



(d) $y = 0.5 + 1x + u, \quad u \sim N(0, sd = 0.2 \log |1 + x|)$



Properties of the new model

- Under MLR.1–4 for the original model, the new model:
  - is linear in parameters (MLR.1),
  - is randomly sampled (MLR.2),
  - does not suffer from perfect collinearity (MLR.3),
  - has zero conditional mean of the error (MLR.4),
  - has a constant variance of the error (MLR.5)

$$\text{Var}\left(\left(\frac{u_i}{\sqrt{h_i}}\right)^2\right) = \mathbb{E}\left(\left(\frac{u_i}{\sqrt{h_i}}\right)^2\right) = \frac{\mathbb{E}(u_i^2)}{h_i} = \frac{\sigma^2 h_i}{h_i} = \sigma^2,$$

  - if, in addition, MLR.6 holds for the original model, then it also holds for the new model (MLR.6).
- Therefore, the standard OLS approach can be used for the new model, and all 'tools' work as usually under the CLM assumptions (mainly the t and F statistics and CIs).

Weighted least squares

- In fact, we run the OLS estimation using variables weighted by $1/\sqrt{h_i}$, i.e., less weight is given to observations with a higher error variance.
- Hence the name **weighted least squares (WLS)**.
- WLS is a member of a broader family of the **generalized least squares (GLS)** methods.
- Estimates obtained by WLS are then **interpreted with respect to the original regression**

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i.$$

LPM (shortly) revisited (**readings: 8.5**)

- Example of a model with a 'known' form of heteroskedasticity: the linear probability model (LPM) from the last lecture #9:

$$\boxed{\text{Var}(u|X)} = \text{Var}(y|X) = \boxed{p(X)\big(1 - p(X)\big)},$$

  where
  $p(X) \equiv \text{P}(y = 1|X) = \mathbb{E}(y|X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.
- To apply WLS, we just need to estimate for each i:

$$\boxed{\hat{h}_i = \hat{y}_i(1 - \hat{y}_i)}.$$

- What is the catch here?
- And what are potential solutions?

Outline

Feasible GLS

- Most of the time, we do not know the form of heteroskedasticity, and it thus needs to be **estimated**.
- We thus need to **model** h_i first to obtain $\hat{h}_i$.
- Such procedure leads to the **feasible GLS (FGLS) estimator**, sometimes referred to as the estimated GLS (EGLS).
- Heteroskedasticity form can be estimated in various ways, and here we stick to one of the flexible approaches.

Feasible GLS

- Assume that the conditional variance has the following form

$$\mathrm{Var}(u|X) = \sigma^2 h(X) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k).$$

- Exponential fcn form is utilized mainly due to the need for $h(X) > 0$.
- We can thus write

$$u^2 = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k) v,$$

and transform it to

$$\log(u^2) = a_0 + \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e.$$

- Estimating u^2 with $\hat{u}^2$, we run a regression

$$\log(\hat{u}^2) = a_0 + \delta_0 + \delta_1 x_1 + \ldots + \delta_k x_k + e,$$

we save the fitted values, and transform them to $\boxed{\hat{h}_i = \exp\left(\widehat{\log(\hat{u}^2)}\right)}$.

- Finally, we apply WLS with weights $\frac{1}{\sqrt{\hat{h}_i}}$.
- Functional form inside the exponential can be freely adjusted.

WLS vs. FGLS

- If h_i is known, the WLS estimator is **BLUE**.
- What if the assumed heteroskedasticity function is wrong?
  - WLS still at least **consistent** under MLR.1–4 but **standard errors no longer** (even asymptotically) **valid**,
  - solution: WLS combined with robust standard errors.
  - **readings: 8.4**.

- If h_i needs to be estimated, the FGLS estimator is **biased** but **consistent** and **asymptotically more efficient** than OLS.

$\Rightarrow$ FGLS procedure based on $\hat{h}_i$ is thus an 'attractive' alternative to OLS **only for large samples**.

Predictions under heteroskedasticity (**readings: 8.4**)

- Point prediction based on OLS is still unbiased and consistent.
- Point prediction based on WLS is unbiased and consistent (for the known heteroskedasticity form) or consistent for FGLS (for an unknown form).
- For the standard errors of the predictions, we need to consider heteroskedasticity:
  - for the **mean prediction** $\mathbb{E}(y|x_{n+1})$, the process (substitution) is the same, but WLS/FGLS instead of OLS is used,
  - for the **individual prediction for a specific unit**, we use

$$se(\hat{e}^0) = \sqrt{\hat{\sigma}^2 \boxed{h(X^0)} + \text{Var}(\hat{y}^0)},$$

  again under WLS/FGLS instead of OLS.

A short guide

- Run OLS and test for homoskedasticity (B-P, White tests).
- **Homoskedastic?**
  - great (assuming MLR.1–4 to hold)!
  - if also MLR.6 holds, even better!
- **Heteroskedastic?**
  - Do we know the form?
  - **Yes:** WLS is BLUE, i.e., it works even for small samples.
    - Is the assumed heteroskedasticity form the correct one? Use the Hausman's test (not covered this semester, but it is good to know it exists).
  - **No:** apply FGLS (estimate $h(X)$ and run WLS using $\hat{h}(X)$) which is consistent and asymptotically efficient, i.e., you need a large sample.
    - You can directly support FGLS with **robust standard errors** as you need a large sample anyway.

Seminars and the next lecture

- Seminars:
  - consequences of heteroskedasticity
  - heteroskedasticity-robust standard errors and inference
  - testing for heteroskedasticity: B-P test, White tests
  - LPM estimation using WLS
- Next lecture #11:
  - functional form misspecification
  - proxy variables
  - OLS under measurement error
  - missing data, nonrandom samples, and outliers
- Readings for lecture #11:
  - Chapter 9: 9.1–9.5

Appendix: White's standard errors (not mandatory)

Valid estimator of the OLS estimator variance, which works without the MLR.5 assumption, is defined as

$$\boxed{\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2}} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{\left(SST_j(1 - R_j^2)\right)^2},$$

where $\hat{r}_{ij}$ denotes the ith residual from regressing x_j on all other independent variables, SSR_j and R_j^2 are the SSR and the R^2 from this regression, and $SST_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ is the total sample variation in x_j. The square root of the above quantity is referred to as White's standard errors.

Back

Appendix: CODE FOR LECTURE #10

R code to generate illustrative graphics:

```
n=1000
b0=0.5
b1=1
sigma2=0.04
x<-runif(n)
y<-b0+b1*x+rnorm(n,mean=0,sd=sqrt(sigma2)*x)
#y<-b0+b1*x+rnorm(n,mean=0,sd=sqrt(sigma2)*log(abs(1+x)))
plot(x,y)
```