Statistics
Tutorial 08

Philipp Scherer & Jens Wiederspohn

24.06.2020

Disclaimer!

- **The content of the slides partly relies on material of Philipp Prinz, a former Statistics tutor. Like us, he's just a student. Therefore we provide no guarantee for the content of the slides or other data/information of the tutorial.**
- **Please note that the slides will not cover the entire lecture content. To pass the exam, it is still absolutely necessary to deal with the Wooldridge in detail!**

Homoskedasticity

- $Var(u|x) = \sigma^2 \rightarrow$ Variance of unobserved error is constant
  - conditional on explanatory variables
- Homoskedasticity is a requirement for using t tests, F tests, and confidence intervals (depends on t value)
  - Like bias, homoskedasticity remains even if we increase n
- Homoskedasticity fails when variance of unobserved error u changes with explanatory variables x_j
  - Different error variance across different segments of population

$\rightarrow$ How can we detect and counter violation of homoskedasticity?

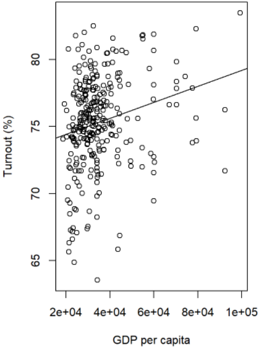
Heteroskedasticity

- Heteroskedasticity = dependency of $Var(u)$ on x_j
  - Violation of Gauss-Markov assumption GM.5

    → OLS is no longer BLUE
    - Possible to find more efficient estimators than OLS
- No impact on unbiasedness or consistency of OLS estimates
  - R^2 also remains a consistent estimator of population R^2
- But: $Var(\hat{\beta}_j)$ is biased under heteroskedasticity
  - If $Var(\hat{\beta}_j)$ is biased, so is $se(\hat{\beta}_j)$
  - Confidence intervals and t statistics are no longer valid

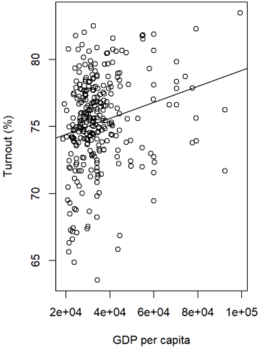
→ T & F statistics are not valid under heteroskedasticity

Heteroskedasticity



- Left plot shows signs of heteroskedasticity
  → $Var(\hat{u}_i)$ larger in districts with lower GDP
- Robust standard error larger than regular standard error
  - $se(\hat{\beta}_{GDP} = 0.000015 < 0.000017 = \widehat{se}(\hat{\beta}_{GDP})$

Heteroskedasticity



- Residuals in right plot are more evenly distributed

  → $Var(\hat{u}_i)$ does not depend too much on % catholic
- Similar robust and regular standard errors
  - $se(\hat{\beta}_{\% C} = 0.01178 < 0.01113 = \widehat{se}(\hat{\beta}_{\% C})$

Heteroskedasticity-robust inference

- Luckily, even with heteroskedasticity, we can still use OLS
  - Solution only holds asymptotically → requires large n!
    (Justification based on law of large numbers and CLT)
- Adjust $se(\hat{\beta}_j)$ to account for heteroskedasticity
  - White/Huber/(Eicker) or just "robust" standard errors
- $\widehat{se}(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)} = \sqrt{\frac{\sum_{i=1}^{n} \hat{r}_{ij} \hat{u}_i^2}{SSR_j^2}}$
  - $\hat{r}_{ij} = i^{th}$ residual from regressing x_j on all other x
  - $SSR_j =$ sum of squared residuals from above regression
    - → We can also use $SSR_j = SST_j(1 - R_j^2)$
    - → *Under what conditions is $\widehat{se}(\hat{\beta}_j)$ large?* (Hint: 8.2)
  - With large n, df-correction $(\frac{n}{n-k-1})$ is equivalent

Heteroskedasticity-robust inference

- Computation of robust t statistic looks familiar:

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\widehat{se}(\hat{\beta}_j)}$$

- Why do we not always use robust standard errors?
  - Robust t statistics can have distributions that are not very close to t distribution if n is small

    → Negative impact on inference (even under LMR.1-LMR.6)
  - For large n, it is argued that one should always use robust se
    - Applies for cross-sectional data!
    - Good practice to also report normal standard errors

    → Standard phrase in almost all empirical papers:
    "robust standard errors are reported in parentheses"

Testing for heteroskedasticity

- H_0: Homoskedasticity $\rightarrow E(u^2|x) = \sigma^2$
- Test H_0 with regression of $\hat{u}^2$ on independent variables
  - Use same independent variables as in regression of Y on x
  - Test overall significance of regression with F statistic
    - $\rightarrow$ If F is low, there is little evidence that $\hat{u}^2$ depends on x
    - $\rightarrow$ In this case, it is good if we cannot reject H_0
  - Caution: if functional form of $E(y|x)$ is misspecified, the test can reject H_0, even if the error variance is constant!
- Alternatively, we can use White test to check $Cov(x, u) = 0$
  - Regress $\hat{u}^2$ on $\hat{y}$ and $\hat{y}^2$
    - $\rightarrow$ White test allows independent variables to have a non-linear (and interactive, see 8.19) effect on error variances
  - Check F statistic to determine if we can reject H_0
    - $\rightarrow$ Is there a joint effect of explanatory variables on $\hat{u}^2$?

1. Test of $E(u^2|x) = \sigma^2 \rightarrow$ Reject H_0

```
## Call:
## lm(formula = resid.sq ~ gdp + young + abi)
##
## Residual standard error: 15.98 on 295 degrees of freedom
## Multiple R-squared:  0.03955,    Adjusted R-squared:  0.02978
## F-statistic: 4.049 on 3 and 295 DF,  p-value: 0.007652
```

2. Test of $Cov(x, u) = 0 \rightarrow$ Keep H_0

```
## Call:
## lm(formula = resid.sq ~ predict + predict.sq)
##
## Residual standard error: 16.27 on 296 degrees of freedom
## Multiple R-squared:  0.001097
## Adjusted R-squared:  -0.005653
## F-statistic: 0.1625 on 2 and 296 DF,  p-value: 0.8501
```

→ *Why is GM.4’ not rejected when GM.4 is?*

Weighted Least Squares estimation

- Before robust standard errors, weighted least squares (WLS) was used to account for heteroskedasticity
  - If we have the correct function $E(u^2|x)$, WLS is more efficient than OLS (under heteroskedasticity; else OLS is BLUE)
- If squared residuals are correlated with some x_j, we assume that $Var(u)$ is proportional to $x_j \rightarrow Var(u|x_j) = \sigma^2 \cdot x_j$
  - In this case, form of heteroskedasticity is known
  - Divide all variables by $\sqrt{x_j}$ and estimate coefficients
  - Minimize weighted sum of squared residuals
    - → Give less weight to observations with higher error variance
    - → Estimates and se are different from OLS, but interpretation stays the same

LPM

- By construction, Linear Probability Model has heteroskedasticity (unless all coefficients are 0)
  - Variance depends on x
  - $Var(y|x) = p(x)[1 - p(x)]$, where $p(x) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$

1. Use OLS with robust standard errors
   - Ignores fact that we know form of heteroskedasticity
   - But: It is easy and the results are okay
2. Use WLS
   - $\hat{h}$ must be positive, but it can be negative if $\hat{y}_i < 0$ or $\hat{y}_i > 1$
   - If we have negative $\hat{h}$, bring all fitted values into unit interval
     Or: abandon WLS and use robust se

Exercise 1

Answer the following questions concerning the following regression model: $y_i = \beta_0 + \beta_1 x_i + u_i$**.**

Exercise 1a and b

Which equation(s) below do(es) represent the homoskedasticity assumption?

B $E(u_i^2|x_i) = \sigma^2$

C $Var(u_i|x_i) = \sigma^2$

→ Both equations represent a constant unobserved error variance

What is/are the consequence(s) of violating the homoskedasticity assumption?

B OLS estimates are not the best estimator.

"The Gauss-Markov Theorem, which says that OLS is best linear unbiased, relies crucially on the homoskedasticity assumption. If $Var(u_i|x_i)$ is not constant, OLS is no longer BLUE" (Wooldridge p.269)

Exercise 1c and d

What is the name of the square root of the following quantity?:

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

- Heteroskedasticity-robust standard error

Suppose that you have a null-hypothesis $\beta_1 = 0$. What is the name of the following quantity?: $\frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$

- Heteroskedasticity-robust t statistic.

Excercise 1e, f and g

What distribution does the quantity above have if certain conditions are fulfilled?

- t-distribution

What situation(s) suffice(s) the conditions above?

D Sample size is large enough.

To test against the null hypothesis that the homoskedasticity assumption is true, what do you have to do?

D Check the overall significance of the regression model, which regresses $\hat{u}^2$ on x.

Excercise 2a, b and c

What is the name of the estimator based on Model 1?

- Ordinary least squares

What is the name of the estimator based on Model 2?

- Weighted least squares

What is the name of the estimator based on Model 3?

- Feasible generalized least squares

Excercise 2d

Is the estimator based on Model 3 BLUE? Answer with yes or no.

- No
- If we could use h_i rather than $\hat{h}_i$ in the WLS procedure, we know that our estimators would be unbiased; in fact, they would be the best linear unbiased estimators, assuming that we have properly modeled the heteroskedasticity.
- Using $\hat{h}_i$ instead of h_i in the GLS transformation yields an estimator called the feasible GLS (FGLS) estimator.
- Having to estimate h_i using the same data means that the FGLS estimator is no longer unbiased (so it cannot be BLUE, either).
- Nevertheless, under heteroskedasticity the FGLS estimator is consistent and asymptotically more efficient than OLS.

Excercise 2e and f

What is/are true among the following statements?

A Model 1's estimator is a special case of Model 2's.

B Model 1's estimator is a special case of Model 3's.

C Model 2's estimator is a special case of Model 3's.

What Model's estimator is the most efficient?

- Model 3

Excercise 2g

Difference in estimates of the first two models is because Model 2 takes into account different level of error variance of observations. What observations are assumed to have a larger error variances in the above estimate?

C Districts with more citizens with university entrance diploma

- Here, $h(x) = h(diploma) = diploma$: the variance of the error is assumed to be proportional to the share of citizens with university entrance diploma. This means that, as the share of citizens with university entrance diploma increases, the variability in the electoral turnout at the electoral district level should also increase.