ECO 372: Introduction to Econometrics

Spring 2025

Important Concepts-II

Sakib Bin Amin, Ph.D.

Associate Professor in Economics

Director, Accreditation Project Team (APT)

Outline

Our objectives for this lecture will be to learn:

- Small and Large Sample Properties
- OLS is BLUE
- Homoskedasticiy and Heteroskedasticity
- Detecting Heteroskedasticity
- Dealing with Heteroskedasticity
- Slope Homogeneity

Small and Large Sample Properties

- Two types of properties are mainly used for checking the estimator’s effectiveness.
- These are : small sample properties and Large sample properties.

<u>Small Sample Properties :</u> Applicable to small samples

Unbiasedness:

1. An estimator is said to be unbiased if $E(\hat{\beta}) = \beta$. This means estimated parameters will be same as population parameters.
2. In laymen perspective, this means if repeated samples of a fixed sized are drawn, then the average value of all different $\hat{\beta}$s obtained will be equal to true β.
3. OLS is an unbiased estimator.

Minimum Variance:

1. An estimator is said to be minimum variance or best estimator of β if its variance is less than the variance of any other estimator taken into consideration.
2. It shows the degree of reliability of the estimator.
3. OLS is a minimum variance estimator.

Small and Large Sample Properties

Efficiency:

1. An estimator is said to be efficient if $\hat{\beta}$ is (i) unbiased and(ii) $var(\hat{\beta}) \leq var(\hat{\vartheta})$
2. OLS is an efficient estimator if all conditions are met.

<u>**Large Sample Properties:**</u> Applicable when sample is large and reaches to infinity

Asymptotic Unbiasedness:

1. An estimator is said to be asymptotically unbiased if $\lim_{n\to\infty} E(\hat{\beta}) = \beta$
2. This condition may hold even if the estimator is biased in small samples.
3. OLS is asymptotically unbiased.

Asymptotic Efficiency:

1. Also known as consistency.
2. An estimator is said to be consistent if it has a smaller variance than others.
3. That is, $\lim_{n\to\infty} var(\hat{\beta}) = 0$ and $\lim_{n\to\infty} E[(\hat{\beta}) - \beta] = 0$
4. OLS is a consistent estimator, given conditions are met.

Small and Large Sample Properties

Large Sample Properties: Applicable when sample is large and reaches to infinity

Asymptotic Unbiasedness:

1. An estimator is said to be asymptotically unbiased if $\lim_{n\to\infty} E(\hat{\beta}) = \beta$
2. This condition may hold even if the estimator is biased in small samples.
3. OLS is asymptotically unbiased.

Asymptotic Efficiency:

1. Also known as consistency.
2. An estimator is said to be consistent if it has a smaller variance than others.
3. That is, $\lim_{n\to\infty} var(\hat{\beta}) = 0$ and $\lim_{n\to\infty} E[(\hat{\beta}) - \beta] = 0$. The condition will be achieved quick with the consistent estimator.
4. OLS is a consistent estimator, given conditions are met.

OLS is BLUE

- Gauss–Markov theorem states that if the Gauss–Markov conditions hold, then the OLS estimator is the best (most efficient) conditionally linear unbiased estimator (is BLUE)
- Gauss–Markov conditions are:
  - $E(u_i|X_i) = 0$ [No correlation between regressor and error term]
  - $Var(u_i|X_i) = E(u_i^2) = \sigma^2$ [Constant variance]
  - $cov(u_i, u_j) = E(u_i, u_j) = 0$ [There is no co-movement among the errors]
- Yes, they are the first 3 assumptions of CLRM!

OLS is Linear:

The best way to show this is to express $\hat{\beta}_1$ as a linear combinations of Y_i

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})Y_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \sum_{i=1}^{N} \omega_i Y_i$$

So, $\hat{\beta}_1 = \sum_{i=1}^{N} \omega_i Y_i = \omega_1 Y_1 + \omega_2 Y_2 + \cdots + \omega_N Y_N$

OLS is BLUE

OLS is an Unbiased Estimator:

The idea is to show estimated parameter is same as population parameter

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^{N}(X_i - \bar{X})u_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

$$E(\hat{\beta}_1) = \beta_1 + E\left[\frac{\sum_{i=1}^{N}(X_i - \bar{X})u_i}{\sum_{i=1}^{N}(X_i - \bar{X})^2}\right] = \beta_1 \qquad [E(u_i) = 0]$$

OLS has Minimum Variance:

The idea is to compare two estimators and show OLS has minimum variance

$$var(\hat{\beta}) = E(\textstyle\sum_{i=1}^{N} \omega_i\, u_i)^2$$

$$= E(\textstyle\sum_{i=1}^{N} \omega_i\, u_i)^2$$

$$= \sigma^2 \textstyle\sum_{i=1}^{N} \omega_i^2$$

Now consider another estimator's variance $var(\hat{\vartheta}) = E(\sum_{i=1}^{N} \delta_i\, G_i)^2$. Consider estimated error variance $\varphi^2 > \sigma^2$

Therefore, $var(\hat{\beta}) < var(\hat{\vartheta})$, indicating OLS is the best estimator.

- ***See the Handout for detail derivations***

Homoskedasticiy and Heteroskedasticity

- Let us recall the constant variance assumption from CLRM:

$$Var(\, u_i|X_i) = E\left(u_i^2\right) = \sigma^2$$

- If this assumption holds, the error term observations are all being drawn from the same distribution
- This is Homoskedasticiy.
- If this assumption is not satisfied (i.e., variance changes over observation) we have heteroskedasticity:

$$Var(\, u_i|X_i) = E\left(u_i^2\right) = \sigma_i^2, \qquad i = 1,2,3, \dots, N$$

- In other words, Heteroskedasticity means that the variance of the errors is not constant across observations.
- In particular the variance of the errors may be a function of explanatory or any exogenous variable (such as Z).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$Var(u_i) = E\left(u_i^2\right) = Z_i^2 \sigma^2$$

Homoskedasticiy and Heteroskedasticity

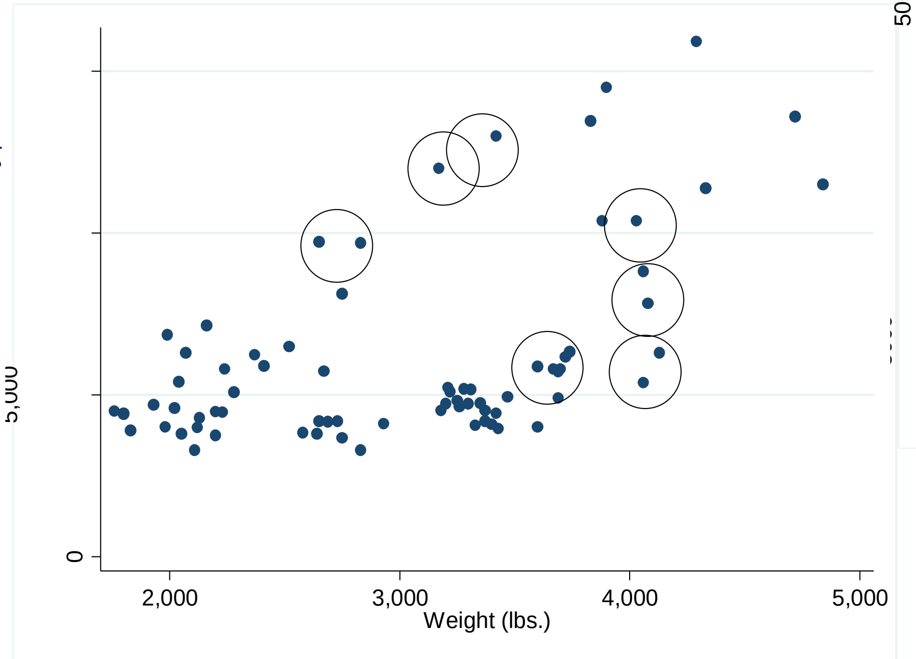
- Measurement error can cause heteroskedasticity. Some respondents might provide more accurate responses than others.
  - This is mostly observed in survey data (either cross-sectional or panel) from households and firms.
- Heteroskedasticity can also occur if there are subpopulation differences or other interaction effects.
  - For example, the effect of income on expenditures differs for rural and urban households.
- Heteroskedasticity does not result in biased OLS parameter estimates.
- However, OLS estimates are no longer BLUE. That is, among all the unbiased estimators, OLS does not provide the estimate with the smallest variance.
- Also, heteroskedasticity causes the estimated standard errors of the regression coefficients to be biased, leading to unreliable hypothesis testing.
  - In general, the t-statistics will actually appear to be more significant than they really are!

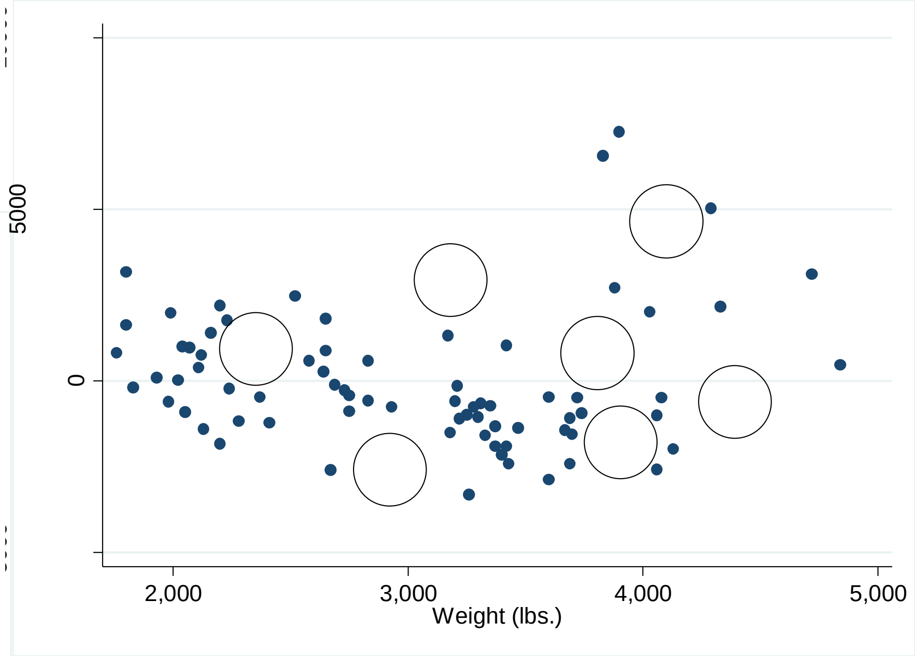
Detecting Heteroskedasticity: With Plotting Data

- Plotting the residuals is always a good first step.
- One can also check the outcome variable's relationship with respect to any predictors.
- The thing one should look in the plots is the spread of data points.
- look at the scatter plots of price of a car and weight of a car. Easily we can see data is scattered a lot!

SATA Commands:

sysuse auto
scatter price weight
regress price weight foreign##c.mpg
predict residual, resid
scatter residual weight





Detecting Heteroskedasticity: With Tests

- Most useful test is the Breusch-Pagan test/ Cook-Weisberg test for heteroskedasticity.
- Its a Chi-squared test. this test runs null hypothesis of constant variance (homoskedastic) against alternative hypothesis of variable variance (heteroskedastic).
- We can reject the null if p-value of the computed Chi-squared within 5-10%.
- There are other tests as well which can be easily performed with STATA.
- Like another famous test known as the White's test. Use the last command for the White's test.

<u>SATA Commands:</u>

sysuse auto
scatter price weight
reg price weight foreign##c.mpg
estat hettest
estat imtest, white

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of price

         chi2(1)      =     6.50
         Prob > chi2  =   0.0108
```

Dealing with Heteroskedasticity

- Re-specify the model/transform the variables. For example, Switching from a linear model to a double-log model might do it.
- Use robust standard errors. This will adjust the standard errors of the coefficients. **<u>This is the most used method</u>**.
- One can also use calculated weights. Weighted Least Squares:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + Z_i u_i$$

Where, $Var(u_i) = E(u_i^2) = \sigma^2$

If we transform the equation by dividing both sides by Z_i we obtain a new regression equation that is homoskedastic.

$$\frac{Y_i}{Z_i} = \frac{\beta_0}{Z_i} + \frac{\beta_1 X_{1i}}{Z_i} + \frac{\beta_2 X_{2i}}{Z_i} + u_i$$

This OLS is BLUE!

Dealing with Heteroskedasticity

Linear regression

Number of obs	=	74
F(4, 69)	=	24.59
Prob > F	=	0.0000
R-squared	=	0.5516
Root MSE	=	2031.4

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	4.613589	.9973318	4.63	0.000	2.623966	6.603211
foreign						
Foreign	11240.33	3351.065	3.35	0.001	4555.138	17925.52
mpg	263.1875	163.5432	1.61	0.112	-63.07226	589.4472
foreign#c.mpg						
Foreign	-307.2166	131.3249	-2.34	0.022	-569.2025	-45.23065
_cons	-14449.58	6351.996	-2.27	0.026	-27121.47	-1777.695

<u>SATA Commands:</u>

reg price weight foreign##c.mpg

reg price weight foreign##c.mpg, vce(r)

**vce(r) option asks STATA to report robust standard errors

Source	SS	df	MS
Model	350319665	4	87579916.3
Residual	284745731	69	4126749.72
Total	635065396	73	8699525.97

Number of obs	=	74
F(4, 69)	=	21.22
Prob > F	=	0.0000
R-squared	=	0.5516
Adj R-squared	=	0.5256
Root MSE	=	2031.4

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	4.613589	.7254961	6.36	0.000	3.166263	6.060914
foreign						
Foreign	11240.33	2751.681	4.08	0.000	5750.878	16729.78
mpg	263.1875	110.7961	2.38	0.020	42.15527	484.2197
foreign#c.mpg						
Foreign	-307.2166	108.5307	-2.83	0.006	-523.7294	-90.70368
_cons	-14449.58	4425.72	-3.26	0.002	-23278.65	-5620.51

Slope Homogeneity

- The concept of slope homogeneity is relatively new concept in econometrics.
- As we discussed the β_is are the slopes (except constant).
- It is a general assumption that the slopes remain same across the cross-section, which is commonly referred as slope homogeneity.
- However, some recent studies have found that there can be differences in slopes across the cross-sections, which might lead to more intuitive explanations of different economic theories or issues.
- However, not always we can go for slope heterogeneity. Among others, some reasons are issue of degrees of freedom, cross-sectional strong unobserved homogeneity, couple with assumptions and predicting power of the estimators.
- Slope homogeneous model and slope heterogeneous model in general looks like as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$Y_i = \beta_0 + \beta_{1i} X_{1i} + \beta_{2i} X_{2i} + u_i$$