Discussion Section

Week 5

Bret Stevens

September 8, 2019

University of California, Davis

Admin

- HW 4 is graded, so you should be able to decide whether or not you want to do HW 5
- Doing this pushed the midterm grading back a bit. It should be done at some point this weekend

Heteroskedasticity

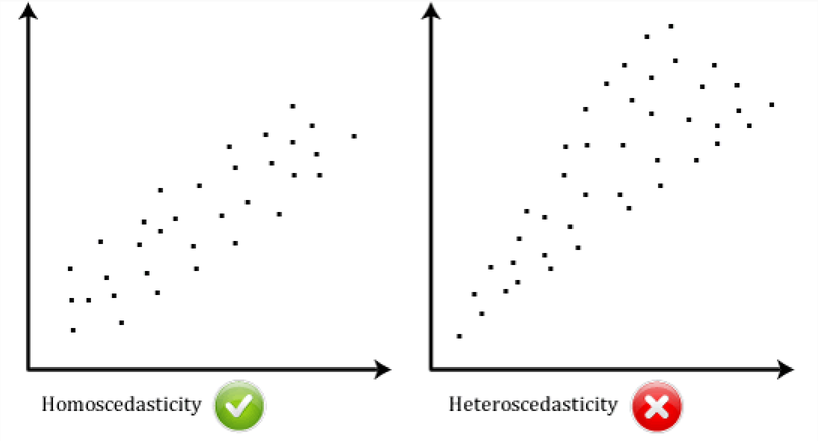
In words:

- Homoskedasticity - The errors all have the same variance
- Heteroskedasticity - The errors all have different variances

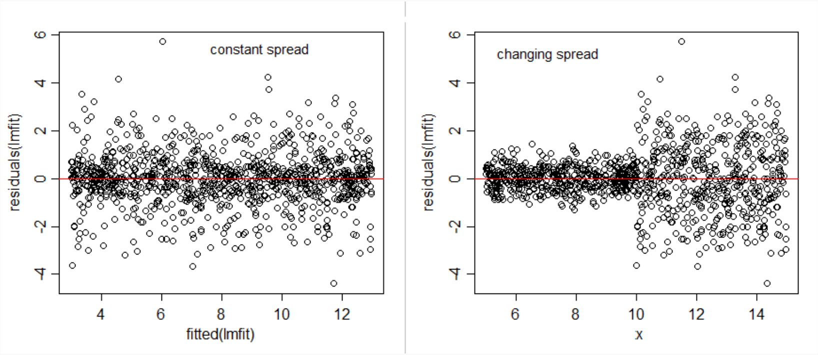
In math:

- Homoskedasticity - $Var(\epsilon_i) = \sigma^2$
- Heteroskedasticity - $Var(\epsilon_i) = \sigma_i^2$

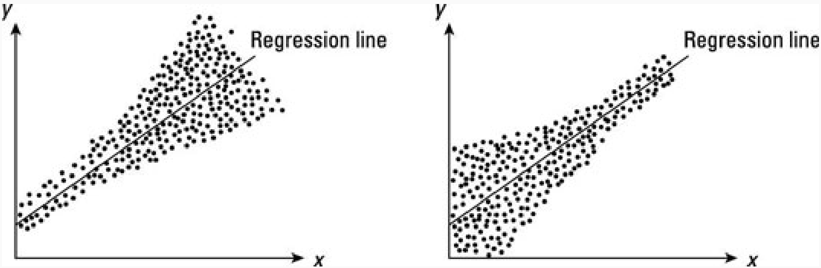
Heteroskedasticity



Heteroskedasticity



Heteroskedasticity



We’ve said its bad, what are the consequences of heteroskedasticity?

Heteroskedasticity

We’ve said its bad, what are the consequences of heteroskedasticity?

- Does it cause bias? (Show on board)

Heteroskedasticity

We've said its bad, what are the consequences of heteroskedasticity?

- Does it cause bias? (Show on board) Nope.

Heteroskedasticity

We’ve said its bad, what are the consequences of heteroskedasticity?

- Does it cause bias? (Show on board) Nope.
- This effects our *efficiency*
- The way we normally calculate standard errors is incorrect when we have heteroskedasticity

Heteroskedasticity

We've said its bad, what are the consequences of heteroskedasticity?

- Does it cause bias? (Show on board) Nope.
- This effects our *efficiency*
- The way we normally calculate standard errors is incorrect when we have heteroskedasticity
    - Why does this matter?

Heteroskedasticity

Lets review hypothesis testing

Heteroskedasticity

Lets review hypothesis testing

- What is the formula for a t-test?

Heteroskedasticity

Lets review hypothesis testing

- What is the formula for a t-test?

$$t = \frac{b - \beta_H}{s.e.(b)}$$

- If the standard error of b is calculated incorrectly, how will this effect our t-test?

Heteroskedasticity

Lets review hypothesis testing

- What is the formula for a t-test?

$$t = \frac{b - \beta_H}{s.e.(b)}$$

- If the standard error of b is calculated incorrectly, how will this effect our t-test?
    - If it is too small, our t-stat will be too large
    - If it is too large, or t-stat will be too small
- Remember, we typically use t-stats to tell if our b is "statistically significant"
- If out t-stats are too large, we will think we have an effect more often than we actually do
    - This is typically the problem with heteroskedasticity

Heteroskedasticity

Heteroskedasticity can be seen easiest when working with averages

- We want to measure the effect of income on educational attainment at the person level, but can only get county-wide averages
- So we have a counties average years of schooling and average income for all counties in the united states
- Our original regression would look like:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$$

- Lets say at the individual level, the errors are homoskedastic

$$Var(\epsilon_{ij}) = \sigma^2$$

- At the county level we have:

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \bar{\epsilon}_i$$

Heteroskedasticity

- What would our error look like then?

$$
\begin{aligned}
Var(\bar{\epsilon}_i) &= Var\left(\frac{1}{n_i}\sum_j \epsilon_{ij}\right) \\
&= \frac{1}{n_i^2}\sum_j Var\left(\epsilon_{ij}\right) \\
&= \frac{1}{n_i^2}\sum_j \sigma^2 \\
&= \frac{1}{n_i^2} n_i \sigma^2 \\
&= \frac{\sigma^2}{n_i} \\
\Rightarrow \frac{\partial Var(\bar{\epsilon}_i)}{n_i} &= -\frac{1}{n_i^2}
\end{aligned}
$$

- Thus, as the size of the county gets larger, the standard error will decrease, causing heteroskedasticity

Heteroskedasticity

How do we know we have heteroskedasticity?

Heteroskedasticity

How do we know we have heteroskedasticity?

- We can test for it using a Breusch-Pegan Test or White's Test
  - We can then fix it using White's Correction
- If we know the source of the heteroskedasticity, we can calculate the exact problem and fix it (like with the averages)
- Logs also help with heteroskedasticity sometimes

Heteroskedasticity

Breusch-Pegan Test

1. Estimate your regression
$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$
2. Save the residuals
3. Square those residuals
4. Regress those squared residuals on the Xs in your original regression
$$e_i^2 = b_0 + b_1 X_{1i} + b_2 X_{2i} + u_i$$
5. From this auxiliary regression we will get an R^2, multiply it by N, this is now our test statistic (like t)
6. Compare NR_a^2 to a χ^2 distribution with degrees of freedom equal to the number of X variables
7. The null hypothesis is that there is homoskedasticity, so if the test statistic is larger than the critical value there is heteroskedasticity

Heteroskedasticity

White's Test

1. Estimate your regression

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

2. Save the residuals
3. Square those residuals (up to now, same as BP)
4. Regress squared residuals on original regression, the squares of those Xs and their interaction

$$e_i^2 = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 + b_3 X_{2i} + b_4 x_{2i}^2 + b_5 X_{1i} X_{2i} + u_i$$

5. From this auxiliary regression we will get an R^2, multiply it by N, this is now our test statistic (like t)
6. Compare NR_a^2 to a χ^2 distribution with degrees of freedom equal to the number of X variables **in your auxiliary regression**
7. The null hypothesis is that there is homoskedasticity, so if the test statistic is larger than the critical value there is heteroskedasticity

Heteroskedasticity

White's Correction

- Normally we calculate the variance as:

$$\begin{aligned} s^2_{b_1} &= \frac{\sum_i x_i^2 s^2}{\left(\sum_i x_i^2\right)^2} \\ &= \frac{s^2 \sum_i x_i^2}{\left(\sum_i x_i^2\right)^2} \\ &= \frac{s^2}{\sum_i x_i^2} \end{aligned}$$

  Where $s^2 = var(e_i)$

- With White's standard errors:

$$Vb_1 = \frac{\sum_i x_i^2 e_i^2}{\left(\sum_i x_i^2\right)}$$

Sample Selection

What is sample selection?

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime
    - Only look at poor neighborhoods

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime
    - Only look at poor neighborhoods
- Want to find effect of drinking on GPA

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime
  - Only look at poor neighborhoods
- Want to find effect of drinking on GPA
  - Only sample frats and sororities

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime
    - Only look at poor neighborhoods
- Want to find effect of drinking on GPA
    - Only sample frats and sororities
- Want to find effect of driving on pollution

Sample Selection

What is sample selection?

- Want to find effect of police presence on crime
  - Only look at poor neighborhoods
- Want to find effect of drinking on GPA
  - Only sample frats and sororities
- Want to find effect of driving on pollution
  - ?

Sample Selection

Most sample selection happens on accident

Sample Selection

Most sample selection happens on accident

- Lets say we are a manager at a grocery store. want to find the effect of a price decrease on the sales of a certain product. So we send out a coupon to some of our store's club members. We then collect the data on how much each person bought. We run the following regression:

$$Q_i = b_0 + b_1 Coupon + e_i$$

Sample Selection

Most sample selection happens on accident

- Lets say we are a manager at a grocery store. want to find the effect of a price decrease on the sales of a certain product. So we send out a coupon to some of our store's club members. We then collect the data on how much each person bought. We run the following regression:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- Do you see a problem with that?

Sample Selection

Most sample selection happens on accident

- Lets say we are a manager at a grocery store. want to find the effect of a price decrease on the sales of a certain product. So we send out a coupon to some of our store's club members. We then collect the data on how much each person bought. We run the following regression:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- Do you see a problem with that?
    - If we only give coupons to people who are already in the store's club, they may react differently to a coupon
    - If they join the club, they are the type of person who may be looking for coupons, thus they may react differently to one than some random person.

Sample Selection

How do we say this in math?

- We have:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- What may be in e_i?

Sample Selection

How do we say this in math?

- We have:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- What may be in e_i?
  - Maybe how much they care about coupons?

Sample Selection

How do we say this in math?

- We have:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- What may be in e_i?
    - Maybe how much they care about coupons?
- In our case, whether or not you cared about sales effected the probability of getting a coupon:

$$Cov(Coupon, e_i) \neq 0$$

Sample Selection

How do we say this in math?

- We have:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- What may be in e_i?
  - Maybe how much they care about coupons?
- In our case, whether or not you cared about sales effected the probability of getting a coupon:

$$Cov(Coupon, e_i) \neq 0$$

- We've broken CR 5!
- What if we gave out the coupons randomly at the door?

Sample Selection

How do we say this in math?

- We have:

$$Q_i = b_0 + b_1 Coupon + e_i$$

- What may be in e_i?
  - Maybe how much they care about coupons?
- In our case, whether or not you cared about sales effected the probability of getting a coupon:

$$Cov(Coupon, e_i) \neq 0$$

- We've broken CR 5!
- What if we gave out the coupons randomly at the door?

$$Cov(Coupon, e_i) = 0$$

Sample Selection

This works if we can run our own experiment, however often we can't. What if we have selection in observational data?

- We can use the Heckman selection model!
  - This model tries to predict what kind of person someone is based on the data we have about them
- So in the coupon case, we would estimate how likely someone is to use a coupon based on what we know about them
- We then use this prediction to account for the fact that they are different from the other people in the sample
  - In the coupon example, it would account for the fact that they are the type of person who likes coupons

Sample Selection

Lets say we have some data where a company just posted a coupon online. They want to see how lowering the price to the sale price will effect sales. They know that most of the people who use the online coupon will be people who take advantage of deals. You have some demographic information on the people and know whether or not they are in the rewards club.

1. Run a regression predicting whether or not a customer is in the rewards club

$$Club_i = a_0 + a_1 Z_1 + \ldots + a_k Z_k + u_i$$

2. Find the predicted values from this regression
3. Include these predicted values in the second stage regression

$$Q_i = b_0 + b_1 X_1 + \ldots + b_k X_k + b_{k+1} Coupon_i + b_{k+2} Cl\hat{u}b_i + e_i$$

This will then control for the fact that many of the people who took advantage of the coupon were people who are really into coupons.