Heteroskedasticity

Manu Navjeevan

February 23, 2020

1 Theory Overview

In the past few sections, we have gone over extensions of our workhouse linear model

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{1}$$

where we place the following general assumptions on the model

$$\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2) \tag{2}$$
$$0 = E[\epsilon_i | X] \tag{3}$$

In general, we have focused on relaxing the assumptions in equation (1) about the form of the relationship between our outcome variable Y and our explanatory variable X. We've allowed for multiple explanatory variables and nonlinear transformations on both sides of the eqaution. We've also discussed how our normality of the errors assumption (2) can be relaxed when we have a large sample size. Today we are going to focus on relaxing another assumption, namely that each of the error terms is distributed identically *with the same variance.*

1.1 Heteroskedasticity

The assumption that all the error terms are distributed with the same variance is called **Homoskedasticity**. When we relax this assumption, we allow the error terms to have different variances and for these variances to depend on the value X. This is called allowing for **Heteroskedasticity**. For example, suppose that we are trying to predict income with age. We hypothesize the linear relationship

$$Income = \alpha + \beta \cdot Age + \epsilon_i \tag{4}$$

Then we note that our error term is equal to

$$\epsilon_i = Income - E[Income | Age] \tag{5}$$

The variance of our error term, given age is then

$$Var(\epsilon_i|Age) = E[(Income - E[Income|Age])^2|Age] \tag{6}$$

It is somewhat natural to think that this variance may be increasing in age. Visually this can be illustrated below in Figure 1. As we can see, the regression points become further from the regression line as age increases.

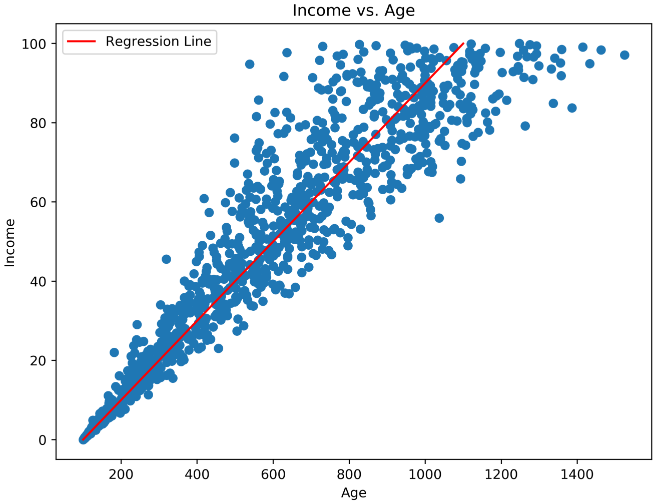

Figure 1

Heteroskedasticity does not affect the way we estimate the base parameters of the model, α and β, but it does affect how we do inference on these parameters. The confidence intervals and hypohesis tests we did in previous chapters were all done using standard errors calculated under the assumption of homoskedasticity. We will have to adjust these to allow for Heteroskedasticity. For example, in the linear case with one explanatory variable, we can calculate the heteroskedasticity consistent variance of our slope parameter

$$V\hat{ar}(\beta) = \frac{N}{N-2} \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})^2 \hat{\sigma}_i^2\right)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \tag{7}$$

In general, we can get this from Stata output by indicating **robust** as an option when running a regression. Compared to the homoskedasticity, this generally makes our confidence intervals wider and makes it harder to find significance. [1]

[1] This is not a bad thing! It's better to not know something about the relationship between X and Y then to think you know something false. Insignificant results are just as important as significant ones.

2 Practice Problems

1. Why might homoskedasticity be a troublesome assumption when using indicator variables or categorical data?

2. Show that the heteroskedasticity variance estimator simplifies to the homoskedasticity variance under homoskedasticity. That is, show that when $\sigma_i^2 = \sigma^2, \forall i$:

$$\frac{\hat{\sigma^2}}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}\left((x_i - \bar{x})^2 \hat{\sigma}_i^2\right)}{\left(\sum\limits_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \tag{8}$$

3. How may we go about estimating σ_i^2 in the case of indicator variables or categorical data?