A note on Heteroskedasticity

Contents

The Source of the Heteroskasticity

Scale Effect

- Consumption function:

$$c = \beta * y + \epsilon$$

  y is income. Intuitively, rich people tend to fluctuate more in consumption on that poor people do. Therefore ϵ is likely to be a function of y , Leading to heteroskedasticity.

- Investment function:

$$I = \beta * v + \epsilon$$

  v is firm value. Intuitively, firm with large scale (large v), its fluctuation in investment may measured by billions, while firm with small scale, its fluctuation in investment may measure by thousand dollars. Putting large and small scale company in one regression leads to heteroskedasticity.

Across-Group Difference

- When individuals belong to different groups, and those groups are quite different, then heteroskedasticity also happens. For example, suppose the individuals contain both employers (entrepreneurs) and employees. The fluctuation of income(consumption) of employers are quite different from those of employees. This leads to the heteroskedasticity

Average Data

- If the outcome variable is an 'average' number, sometimes heteroskedasticity also happens. For example, we study the factors that determines the GDP per capital of a province. Since for provinces with large population, the variance of GDP per capital is low, there exists heteroskedasticity.

Omitted Variable

- Yes, Omitted Variable is also one source of heteroskedasticity. Consider a true model as:

$$y_i = x_i\beta + z_i\gamma + \epsilon_i \tag{1}$$

in which

$$E(\epsilon_i|x_i, z_i) = 0, var(\epsilon_i|x_i, z_i) = \sigma^2$$

Suppose now, however, that we estimate the following model:

$$y_i = x_i\beta + u_i \tag{2}$$

this implies that $u_i = z_i\gamma + \epsilon_i$. Therefore we have

$$cov(x_i, u_i) = cov(x_i, z_i\gamma + \epsilon_i) = cov(x_i, z_i\gamma)$$

$$var(u_i|x_i) = var(z_i\gamma + \epsilon_i|x_i) = var(z_i\gamma|x_i) + var(\epsilon_i|x_i) = \gamma^2 var(z_i|x_i) + var(\epsilon_i|x_i)$$

we know that

$$var(\epsilon_i|x_i) = E_z(var(\epsilon_i|x_i, z)) + var_z(E(\epsilon_i|x_i, z)) = \sigma^2 + 0$$

What is $var(z_i|x_i)$? According to the definition of conditional variance, we have

$$var(z|x) = \int [z - E(z|x)]^2 f(z|x)dz$$

Therefore we have the following results

- If z is independent of x, then $E(z|x) = E(z)$,$f(z|x) = f(z)$. Therefore $var(z|x)$ is not a function of x: There is no problem of heteroskedasticity in the (2). On the other hand, we also have $cov(x_i, u_i) = 0$: there is no problem of endogeneity.
- If z is not independent of x, then $var(z|x)$ is in general a function of x: There is a problem of heteroskedasticity in the (2). On the other hand, we also have $cov(x_i, u_i) \neq 0$: there is problem of endogeneity.

This example shows that, when we specify the model incorrectly, heteroskedasticity is a signal of possible misspecification of the model and omission of endogenous variables. This happens typical when we only include the linear function of x, while in the real model there is non-linear part of x. Therefore, when heteroskedasticity happens, we cannot exclude the possibility that there is uncaptured non-linear part of x.

Of course, if we are quite sure that our model specification is correct, then, heteroskedasticity is not a signal of omitting endogenous variables

Check for Heterosaksticity

Residual- X plot

A simple and quick way to check for heteroskedasticity is to examine scatter plots of the residuals against each of the predictor variables. Check the 'Heteroskedasticity.png' in the same directory.

- If there is a obvious trend of residual when predictor x increases, but the degree of 'variance' does not change much with x (a,b,c): there is omitted variable problem but no heteroskedasticity related to x. This implies that the residual, which is y net of the linear function of x (and of course also the linear function of other variables), is still correlated with x. Therefore, there is some non-linear function of x that also contributes to the y. Therefore we need to add some non-linear functions of x.
- If there is no obvious trend of residual when predictor x increases, but the degree of 'variance' increases (or decreases) obvious with predictor x (d): there is problem of heteroskedasticity, and we cannot exclude the problem of omitted variables especially if we specify a wrong model.

For omitted variable problem, we need to change the model specification. For heteroskedascity, we can (1) use wls (2) take log of y (3) do nothing but only report the robust standard error. But notice, as we mentioned, the heteroskedasticity is a possible sign of omitted variable!

White Test

Main Idea: if there is no heteroskedasticity, then the standard error and robust standard error must be close to each other. (Recall that under homoskedasticity, the standard error and robust standard error are the same). ### BP Test Main Idea: Directly regress OLS residual on the explanatory variables (all or a part of).If the coefficients are jointly not zero, then there exists heteroskedasticity. (One limitation of this test is that it presume the functional form of heteroskedasticity)

The Consequence of Neglecting Heteroskasticity

Now we assume we have specified the model correctly, i.e, there is no problem of omission of endogenous variables. ### For models to which we can apply moment estimation(OLS): Notice that for models that we can apply moment estimation (mostly linear model, where we can use OLS), the moment condition does not assume homoskedasticity. Therefore, if we neglect the heteroskedasticity, the estimator under OLS is still consistent but no longer efficient (Since BLUE is derived under homoskedasticity). Intuitively, observations with larger variance contains less information. OLS, by giving the same weight on every observations, leads to estimators with large variance.

Of course, if we neglect the heteroskedasticity and estimate using MLE, the estimator would be no longer consistent: we are simply optimizing a wrong likelihood function if we neglect the heteroskedasticity. For example,

$$y = \beta x + \epsilon, var(\epsilon|x) = \sigma^2 x^2$$

then the correct log likelihood function contains the following part:

$$\sum_{i=1}^{n} \frac{(y_i - \beta x_i)^2}{2\sigma^2 x_i^2}$$

if we neglect the heteroskedasticity and simply assume $var(\epsilon|x) = \sigma^2$, then the counterpart of the above term is now

$$\sum_{i=1}^{n} \frac{(y_i - \beta x_i)^2}{2\sigma^2}$$

i.e., we are optimizing a wrong likelihood function. ### For models to which we can not apply moment estimation: For most non-linear models (binary choices, poisson regression and so on), we can only use MLE to get the estimation. Again, if we neglect the heteroskedasticity, then we are simply optimizing a wrong likelihood function and hence get a wrong estimation. consider a binary choice model, where we have

$$y_i^* = x_i\beta + \epsilon_i, y_i = 1(y_i^* > 0), \epsilon_i|x\tilde{\mathrm{N}}(0, exp(\sigma x_i))$$

there is heteroskedasticity, and the right log likelihood function is

$$\sum_{1}^{n} \left[y_i \Phi(\frac{-x_i\beta}{\sqrt{exp(\sigma x_i)}}) + (1 - y_i) \left(1 - \Phi(\frac{-x_i\beta}{\sqrt{exp(\sigma x_i)}}) \right) \right]$$

Again, if we neglect such heteroskedasticity and simply assume

$$\epsilon_i|x\tilde{\mathrm{N}}(0, exp(\sigma))$$

, then the likelihood function becomes

$$\sum_{1}^{n} \left[y_i \Phi(\frac{-x_i\beta}{\sqrt{exp(\sigma)}}) + (1 - y_i) \left(1 - \Phi(\frac{-x_i\beta}{\sqrt{exp(\sigma)}}) \right) \right]$$

Still we are optimization a wrong likelihood function.

In summary,suppose we neglect heteroskedasticity, for models where we can apply moment condition estimation, we can get consistent estimation of the parameter as long as we use moment condition estimation; for models where we can only estimate using MLE, we cannot get consistent estimation.

Dealing with Heteroskasticity (For linear model)

OLS + Robust Standard Error

Details can be found in most textbooks.

FGLS (WLS)

Details can be found in most textbooks.

Take logarithm of y

Sometimes, taking logarithm of y may be a strategy. Still consider:

$$y_i = x\beta + \epsilon_i, var(\epsilon_i|x_i) = \sigma^2 x_i^2$$

now, we define $z_i \equiv \log y_i$ and specify a new model

$$z_i = x_i\gamma + u_i$$

we therefore have

$$var(u_i|x_i) = var(log(x_i\beta + \epsilon_i) - x_i\gamma|x_i) = var\left(log(1 + \frac{\epsilon_i}{x_i\beta}) + log(x_i\beta) - x_i\gamma|x_i\right)$$

$$= var\left(log(1 + \frac{\epsilon_i}{x_i\beta})|x_i\right) \approx var(\frac{\epsilon_i}{x_i\beta}|x_i) = \frac{1}{x_i^2\beta^2} var(\epsilon_i|x_i) = \frac{\sigma^2}{\beta^2}$$

Therefore in the new model there is less heteroskadasticity.

By the way, taking log of y can help dealing with the data that is not normally distributed. To be specific, after the regression, we may want to take the residuals and see the distribution. If the distribution of residuals is highly right skewed (which is often the case),then the model we estimated violates normal error distribution (This is not a big issue in the large sample theory). We can take log of y to resolve this problem.