Linear models

Gabriela Ciołek

HETEROSKEDASTICITY

- Consider linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

- The homoskedasticity assumption for multiple regression states that the variance of the unobservable error, u, conditional on the explanatory variables, is constant. Homoskedasticity fails whenever the variance of the unobservables changes across different segments of the population, where the segments are determined by the different values of the explanatory variables.
- Homoskedasticity is needed to justify the usual t tests, F tests, and confidence intervals for OLS estimation of the linear regression model, even with large sample sizes.

HETEROSKEDASTICITY

In presence of heteroskedasticity it is known that

- the usual OLS t statistics do not have t distributions (the problem is not resolved by using large sample sizes)
- F statistics are no longer F distributed
- In summary, the statistics we used to test hypotheses under the Gauss-Markov assumptions are not valid in the presence of heteroskedasticity

- Many tests for heteroskedasticity have been suggested over the years. Some of them, while having the ability to detect heteroskedasticity, do not directly test the assumption that the variance of the error does not depend upon the independent variables.
- We assume that

  $$\mathbb{E}(u|x_1, \cdots, x_k) = 0$$

  so OLS is unbiased and consistent.
- Null hypothesis

  $$H_0 : Var(u|x_1, \cdots, x_k) = \sigma^2$$

  and

  $$Var(u|x_1, \cdots, x_k) = \mathbb{E}(u^2|x_1, \cdots, x_k) = \mathbb{E}(u^2).$$
- That is, we assume that the ideal assumption of homoskedasticity holds, and we require the data to tell us otherwise.
- If we cannot reject H_0 at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem. However, remember that we never accept H_0, we simply fail to reject it.

- In order to test for violation of the homoskedasticity assumption, we want to test whether u^2 is related (in expected value) to one or more of the explanatory variables.
- If H_0 is false, the expected value of u^2, given the independent variables, can be any function of the x_j.
- We consider

$$u^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v$$

  where v is an error term with mean zero given the x_j. Pay close attention to the dependent variable in this equation: it is the square of the error in the original regression equation.
-

$$H_0 : \delta_1 = \cdots = \delta_k = 0.$$

- Under the null hypothesis, it is often reasonable to assume that the error v, is independent of $x_1, \ x_2, \ \cdots, \ x_k$.
- It is known (see for instance chapter 5 of J. Wooldridge) that the F statistic for the overall significance of the independent variables in explaining u^2 can be used to test H_0.

- As we have emphasized before, we never know the actual errors in the population model, but we do have estimates of them: the OLS residual, $\hat{u}_i$, is an estimate of the error u_i for observation i. Thus, we can estimate the equation

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + error$$

- It turns out that using the OLS residuals in place of the errors does not affect the large sample distribution of the F statistic.

Correcting heteroskedasticity

The heteroskedasticity is known up to a multiplicative constant

$$Var(u|x) = \sigma^2 h(x),$$

where $h(x)$ is some function of the explanatory variables that determines the heteroskedasticity. Since variances must be positive, $h(x) > 0$ for all possible values of the independent variables.

Assumption: function h(x) is known. The population parameter σ^2 is unknown, but we will be able to estimate it from a data sample.

For a random drawing from the population, we can write

$$\sigma_i^2 = Var(u_i|x_i) = \sigma^2 h(x_i) = \sigma^2 h_i,$$

where we again use the notation x_i to denote all independent variables for observation i, and h_i changes with each observation because the independent variables change across observations.

We have assumed:

$$Var(u|x) = \sigma^2 h(x).$$

How can we use this information to estimate the β_j?

- take the original equation:

  $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + u_i,$$

  which contains heteroskedastic errors, and transform it into an equation that has homoskedastic errors
- since h_i is just a function of x_i, $u_i/\sqrt{h_i}$ has a zero expected value conditional on x_i
- Further, since

  $$Var(u_i|x_i) = E(u_i^2|x_i) = \sigma^2 h_i,$$

  the variance of $u_i/\sqrt{h_i}$ (conditional on x_i) is σ^2 :

  $$E\left((u_i/\sqrt{h_i})^2\right) = E(u_i^2)/h_i = \sigma^2 h_i/h_i = \sigma^2,$$

  where we have suppressed the conditioning on x_i for simplicity.

We consider now

$$y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \cdots \beta_k(x_{ik}/\sqrt{h_i}) + u_i/\sqrt{h_i},$$

or

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots \beta_k x_{ik}^* + u_i^*,$$

where $x_{i0}^* = 1/\sqrt{h_i}$ and the other starred variables denote the corresponding original variables divided by $\sqrt{h_i}$.

- The transformed equation is linear in its parameters.
- Further, u^* has a zero mean and a constant variance σ^2, conditional on x^*.
- If u_i has a normal distribution, then u_i^* has a normal distribution with variance σ^2.

Therefore, the transformed equation satisfies the classical linear model assumptions (MLR.1 through MLR.6), if the original model does so, except for the homoskedasticity assumption.

Method GLS

- The GLS estimators for correcting heteroskedasticity are called weighted least squares (WLS) estimators.
- the β_j^* minimize the weighted sum of squared residuals, where each squared residual is weighted by $\frac{1}{h_i}$.
- The idea is that less weight is given to observations with a higher error variance; OLS gives each observation the same weight because it is best when the error variance is identical for all partitions of the population.
- Mathematically, the WLS estimators are the values of the b_j that make

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2 / h_i$$

as small as possible.

Exercise 1

Test the homoskedasticity

- load HPRICE1.RAW
- description of the dataset: HPRICE1.DES
- Test the homoskedasticity

We start by performing regression

$$price = \beta_0 + \beta_1 bdrms + \beta_2 lotsize + \beta_3 sqrft + u.$$

We want to test

$$H_0 : Var(u|x_1, x_2, \cdots, x_k) = \sigma^2$$

Since we assume that u are zero mean, thus $Var(u|x) = \mathbb{E}(u^2|x)$ and

$$H_0 : \mathbb{E}(u^2|x_1, x_2, \cdots, x_k) = \mathbb{E}(u^2) = \sigma^2.$$

- load hprice1.raw
- Do regression :

$$price|bdrms,\ lotsize,\ sqrft$$

- $y = hprice1(:, 1);$
- $[n, k] = size(hprice1);$
- $X = [ones(n, 1), hprice1(:, [3, 4, 5])];$
- $[n, k] = size(X)$
- $beta = inv(X' * X) * X' * y$
- $u = y - X * beta;$

Thus, if H_0 is false, the expectation of u^2 may possibly be any function of x_j. We will test the model $\hat{u}^2 = \delta_0 + \delta_1 bdrms + \delta_2 lotsize + \delta_3 sqrft + \nu$, with the null hypothesis of homoskedasticity :

$$H0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0.$$

Test the homoskedasticity:

- $$u2 = u.^{2};$$
  $$y = u2;$$
- unrestricted model
  $$\beta = inv(X' * X) * X' * y$$
  $$u = y - X * beta;$$
  $$SSR0 = u' * u$$
- restricted model (all the coefficients except intercept are zeros)
  $$X = [ones(n, 1)];$$
  $$\beta = inv(X' * X) * X' * y$$
  $$u = y - X * beta;$$
  $$SSR1 = u' * u$$
- Compute F-statistic and p-value. Do we reject H_0 of homoskedasticity?

Exercise 2

Test the homoskedasticity when considering the logarithmic form

- Consider model:

$$log(price)|bdrms,\ lotsize,\ sqrft$$

- Perform regression
- Compute $\hat{u}^2$ and

$$\hat{u}^2 = \delta_0 + \delta_1 bdrms + \delta_2 lotsize + \delta_3 sqrtf + v$$

- Test $H_0 : \delta_1 = \delta_2 = \delta_3 = 0$
- compute F-statistic and p-value
- Do we reject null hypothesis of homoskedasticity?

Exercise 3

Assume that we know the form of the heteroskedasticity $Var(u|x) = \sigma^2 h(x)$ with $h(x) = lotsize$ that is, the variance of the error is proportional to the size of the terrain. Consider the model:

$$\frac{price}{\sqrt{lotsize}} = \frac{\beta_0}{\sqrt{lotsize}} + \frac{\beta_1 bdrms}{\sqrt{lotsize}} + \frac{\beta_2 lotsize}{\sqrt{lotsize}} + \frac{\beta_3 sqrft}{\sqrt{lotsize}} + \frac{u}{\sqrt{lotsize}}$$

Verify that the above model satisfies the hypothesis of homoskedasticity.

$$y = hprice1(:, 1);$$

$$X = [ones(n, 1), hprice1(:, [3, 4, 5])];$$

$$lotsize = hprice1(:, 4)$$

$$y = y./sqrt(lotsize)$$

$$fori = 1 : k$$

$$X(:, i) = X(:, i)./sqrt(lotsize)$$

$$end$$

$$[n, k] = size(X)$$

$$beta = inv(X' * X) * X' * y$$

$$u = y - X * beta;$$

Exercise 4

- Consider a linear model $sav = \beta_0 + \beta_1 inc + u$.

$$load \;\; saving.raw$$

- We know *a priori*

$$\sigma_i^2 = Var(u_i|inc_i) = \sigma^2 inc_i$$

- Conduct a correction by the method WLS
- Test the homoskedasticity for modified model.

$$y = saving(:,1);$$

$$inc = saving(:,2);$$

$$[n,k] = size(saving);$$

$$X = [ones(n,1), inc];$$

$$[n,k] = size(X)$$

$$beta = inv(X^{'} * X) * X^{'} * y$$

$$u = y - X * beta;$$

Correction by the method WLS :

$$ys = y./sqrt(inc);$$

$$Xs = [ones(n,1)./sqrt(inc)inc./sqrt(inc)];$$

$$beta = inv(Xs' * Xs) * Xs' * ys$$

$$u = ys - Xs * beta;$$