# Heteroskedasticity in Cross-Sectional Data

The classical linear regression model (CLRM) assumes that the error term $u_i$ in the regression model has homoscedasticity (equal variance) across observations, denoted by $\sigma^2$, that is, $var(e_i) = \sigma^2$.

However, if the assumption of homoscedasticity, or equal variance, is not satisfied, we have the problem of heteroskedasticity, or unequal variance, denoted by $\sigma_i^2$ (note the subscript *i*), that is, $var(e_i) = \sigma_i^2$.

**There are two types of heteroskedasticity:**

1- **Pure heteroskedasticity**: arises if the model is correctly specified, but the errors are heteroskedastic, e.g., the DGP is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i$$

where $var(e_i) = \sigma_i^2$

There are many ways to specify the heteroskedastic variance $\sigma_i^2$. A very simple specification is discrete heteroskedasticity, where the errors are drawn from one of two distributions, a "wide" distribution (with Large variance $\sigma_L^2$) or a "narrow" distribution (with Small variance $\sigma_S^2$). A common specification is to assume that the error variance is proportional to the square of variable *Z* (that may or may not be one of the independent variables). In this case, $var(e_i) = \sigma_i^2 = \sigma^2 Z_i^2$ and each observation's error is drawn from its own distribution with mean zero and variance $\sigma^2 Z_i^2$.

2- **Impure heteroskedasticity**: can arise if the model is mis-specified (e.g., due to an omitted variable) and the specification error induces heteroskedasticity. For example, suppose the DGP is

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

but we estimate $Y_i = \beta_0 + \beta_1 X_{1i} + e_i^*$

where $e_i^* = \beta_2 X_{2i}$

and where $e_i$ is a classical error term

Then, if $X_{2i}$ itself has a heteroskedastic component (e.g., the value of $X_{2i}$ comes from either a "wide" or "narrow" distribution), then omitting it from the model makes the mis-specified error term $e_i^*$ behave heteroskedastically. Of course, the solution here is simple: do not omit $X_2$ from the model!

**Heteroskedasticity has the following consequences:**

1- Heteroskedasticity does not alter the unbiasedness and consistency properties of OLS estimators.

Let $Y_i = \alpha + \beta X_i + e_i$

and $E[(X_i - \bar{X})e_i] = 0$

and let $e$ be heteroskedastic: $E[(e_i)^2] = \sigma_i^2$

$$\hat{\beta} = \beta + \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})e_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

$$E[\hat{\beta}] = \beta$$

2- But OLS estimators are no longer of minimum variance or efficient. That is, they are not best linear unbiased estimators (BLUE); they are simply linear unbiased estimators (LUE).

$$V[\beta] = E\left[\left(\beta - E(\beta)\right)^2\right] = E\left[\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})e_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)^2\right]$$

$$= \frac{1}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} E\left[\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 e_i\right)^2\right]$$

$$= \frac{1}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} E[(X_1 - \bar{X})(X_1 - \bar{X})e_1 e_1 + (X_1 - \bar{X})(X_2 - X)e_1 e_2 + \cdots$$
$$+ (X_{n-1} - \bar{X})(X_n - \bar{X})e_{n-1}e_n + (X_n - \bar{X})(X_n - \bar{X})e_n e_n]$$

$$= \frac{1}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 (e_i)^2\right]$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 E[(e_i)^2]}{(\sum_{i=1}^{n}(X_i - \bar{X}))^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sigma_i^2}{(\sum_{i=1}^{n}(X_i - \bar{X}))^2}$$

This the OLS variance under heteroskedasticity which is wrong compared to the following estimated variance under homoscedasticity

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n}\hat{e}_i^2/(n-k)}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- Another way to derive the variance of OLS under heteroskedasticity is:

$$V[\beta] = var\left(\sum_{i=1}^{n} w_i e_i\right)$$
$$= \sum_{i=1}^{n} w_i^2 var(e_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} w_i w_j cov(e_i e_j)$$
$$= \sum_{i=1}^{n} w_i^2 \sigma_i^2$$
$$= \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 \sigma_i^2]}{[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2}$$

- If $\sigma_i^2$ is constant, then we can take it out of the numerator summation and the numerator summation on deviations of *x* cancels one of the denominator summations, leaving the usual formula:

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

- If the variance is not constant, we cannot do this and the ordinary estimator is incorrect.

3- As a result, the t and F tests based under the standard assumptions of CLRM may not be reliable, resulting in erroneous conclusions regarding the statistical significance of the estimated regression coefficients.

**Detection of heteroskedasticity:**

- The eye-ball test is a simple but casual way to look for heteroskedasticity
    - Plot the residuals (or the squared residuals) against the explanatory variables or the predicted values of the dependent variable
    - If there is an apparent pattern, then there is heteroskedasticity of the type that the variance is related to x or **xβ**.
- The **Breusch-Pagan test** is a formal way to test whether the error variance depends on anything observable.
    - Suppose that $var(e_i) = \sigma_i^2 = E(e_i^2) = h(\alpha_1 + \alpha_2 z_{i,2} + \cdots + \alpha_S z_{i,S})$, where the *z* variables are may be the same or different from the *x* variables in the regression, and *h* may be any kind of function.

- To test this, we regression the squared residuals on z variables and test the hypothesis that $\alpha_1 \cdots \alpha_S$ are all zero:
  - $\hat{e}_i^2 = \alpha_1 + \alpha_2 z_{i,2} + \cdots + \alpha_S z_{i,S} + v_i$
  - Several possible tests of $H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_S = 0$:
    - Lagrange multiplier test is $NR^2 \sim \chi^2_{(S-1)}$
      - Reject homoskedasticity if test statistic >critical value
      - This is asymptotic test

- The **White test** is a test that is similar to the Breusch-Pagan test, using as the *z* variables
  - All of the *x* variables in the original equation
  - The squares of all the *x* variables
  - Optionally, the cross-products of all *x* variables
  - This leads to lots of variables if *K* is large
  - Dummies cannot be included as squares

- The **Goldfeld-Quandt** test is suitable for samples in which the data can be divided into two groups and with variance differing only between groups.
  - Suppose that the groups *A* and *B* with variances $\sigma_A^2$ and $\sigma_B^2$.
  - Run separate regressions for the two sub-samples *A* and *B* and calculate the estimated error variances from the residuals
  - $F = \frac{\hat{\sigma}_A^2/\sigma_A^2}{\hat{\sigma}_B^2/\sigma_B^2} \sim F_{(N_A-K, N_B-K)}$
  - If the null hypothesis is that $\sigma_A^2 = \sigma_B^2$, then the ratio of the estimated variances is the F statistic and we can do one-tailed or two-tailed test.
    - Must be careful with the two-tailed F test, though, because F tables only report the right-hand tail area and critical values.
    - Make *A* the sample with larger variance so that all the critical area is on the right.
      - The one-tailed test with alternative hypothesis $\sigma_A^2 > \sigma_B^2$ is just the ordinary *F* test with the usual critical value.
      - For the two-tailed test, 5% critical value become 10% critical value because of the possibility that the variance of *A* is smaller than the variance of *B*.

**Dealing with heteroskedasticity:**

In the presence of heteroskedasticity, the BLUE estimators are provided by the method of weighted least squares (WLS). However, this method assumes that the correct pattern of heteroskecdasticity is known. Alternatively, the White heteroskedasticity-consistent standard errors can be used which are given by the following robust estimator of the variance

$$V[\hat{\beta}] = \frac{\sum_{i=1}^{n} \hat{e}_i^2/(n-k)}{[\sum_{i=1}^{n}(X_i - \bar{X})^2]^2/n}$$



**References:**


- Gujarati, D., ***Econometrics by Example***, 2012. McGraw Hill.
- Section 8, Heteroskedasticity, http://www.reed.edu/economics/parker/s12/312/notes/Notes8.pdf
- The Classical Model, Heteroskedasticity and Correlations Across errors, http://www.sfu.ca/~pendakur/teaching/buec333/Heteroskedasticity%20and%20Correlations%20Across%20Errors.pdf