

Heteroskedasticity

Pravat Uprety
Central Department of Statistics
Tribhuvan University

Heteroskedasticity

For the sample two variable model

$$Y_i = \alpha + \beta X_i + \epsilon_i \text{ -----(1)}$$

We assumed that the variances of disturbance term ϵ_i is constant for all observations

$$\text{i.e. } \text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma^2 \text{ (constant for all } i) \text{ -----(2)}$$

- this feature of disturbances term of the regression model is known as **homoskedasticity**.

However, it is quite common in regression analysis to have cases where the variance of disturbance term becomes variable rather than remaining constant.

- In this situation the disturbance is said to be **heteroskedasticity**.
- i.e. $\text{var}(\epsilon_i) = E(\epsilon_i^2) = \sigma_i^2 \text{ -----(3)}$
- which means that the variance of disturbance term can change for every different observation in the sample $i = 1, 2, \dots, n$.

Sources

- 1) When we are dealing with micro-economic or cross sectional data, we are very likely to have a heteroskedasticity problem.
- 2) Presence of outliers in data may cause heteroskedasticity.
- 3) Heteroskedasticity may arise if some relevant variables have been mistakenly omitted.
- 4) Inclusion of explanatory variables in the model whose distributions are skewed.
- 5) Heteroskedasticity may also arise due to incorrect data transformation and incorrect functional form.

Consequences

Unbiased ness

$$E(\hat{\beta}) = \beta$$

This shows that $\hat{\beta}$ remains unbiased when the disturbance term of the model ϵ_i is heteroskedasticity.

Bestness

$$\text{Thus } \text{Var}(\hat{\beta})|_{\text{heteroskedasticity}} > \text{Var}(\hat{\beta})|_{\text{homoskedasticity}}$$

So, there is no longer a minimum variance and hence not best estimator.

$\hat{\beta}$ is unbiased but not the best.

Consistence

: $\hat{\beta}$ is consistent when the disturbance term is heteroskedastic.

Consequences

- The OLS estimators continue to remain unbiased and consistent under heteroskedasticity.
- Heteroskedasticity increases the variances of the distributions of estimator of B thereby turning the OLS estimators inefficient (not best

Heteroskedasticity also affects the variance of OLS and their standard error. In fact, the presence of heteroskedasticity, in general causes the OLS method to underestimate the variances and hence standard error of the estimators. As a consequence, we have higher than expected values of t and F statistic. As the OLS estimators are unbiased under heteroskedasticity, the forecasts generated on the basis of the estimated model will also be unbiased.

Detection Techniques

Graphical method

We may graphically examine the presence of heteroskedasticity by plotting the squared residuals (ϵ_i^2) against the explanatory variable (X_i) to which it is suspected the disturbance variance is related. Since ϵ_i^2 is unknown, its proxy measure $\epsilon_i^2 = e_i^2$ is used.

Breusch-Pagan-Godfrey Test

Breusch-Pagan Godfrey (1978) developed a Lagrange Multiplier (LM) test to examine the presence of heteroskedasticity in data

Considering the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon_i \quad \text{-----}(1)$$

And suppose that

$$\text{Var}(\epsilon_i) = \sigma_i^2 = f(\gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \dots + \gamma_r Z_{ri})$$

This implies that $\text{Var}(\epsilon_i)$ is the function of non-stochastic Z s. Here Z s represent a set of variables that we think determine the variance of the disturbance term ϵ_i . Usually, the explanatory variables of (1) are used for Z s.

The steps involved in the Breusch-Pagan Godfrey test are the following.

Step 1: Estimate model (1) by OLS method and obtain the estimated residuals $\hat{e}_i = \hat{\varepsilon}_i$.

Step 2: Run the auxiliary regression

$$e_i^2 = \gamma_0 + \gamma_1 Z_{1i} + \gamma_2 Z_{2i} + \dots + \gamma_r Z_{ri} + v_i \quad (2)$$

Step 3: Construct the null and alternative hypothesis

$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_r = 0$ (Homoskedasticity)

H_1 : At least one of the γ_s is non zero (Heteroskedasticity)

Step 4: Compute $LM = nR^2$ where n = number of observations used to estimate auxiliary regression model (2) and R^2 is the coefficient of determination of this regression.

(Note that LM-statistic follows a χ^2 distribution with degrees of freedom r)

Step 5: If $LM = \chi^2 \leq \chi_r^2$ at $\alpha\%$

Then we do not reject H_0

If $LM = \chi^2 > \chi_r^2$ at $\alpha\%$

Then we reject H_0

And we conclude that there is significant evidence of heteroskedasticity in data.

[If we have corresponding p-value then we can use the p-value approach]

Park Test

Considering the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_k X_{ki} + \epsilon_i \quad \text{-----(1)}$$

|

Step 1: Estimate model (1) by OLS method and obtain the estimated residuals $\hat{e}_i = \hat{\epsilon}_i$.

Step 2: Run the auxiliary regression

$$\ln(e_i^2) = \alpha_0 + \alpha_1 \ln Z_{1i} + \alpha_2 \ln Z_{2i} + \dots + \alpha_r \ln Z_{ri} + v_i \quad \text{-----(2)}$$


Step 3: Construct the null and alternative hypothesis

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0$ (Homoskedasticity)

H_1 : At least one of the α_s is non zero (Heteroskedasticity)

Step 4: Compute $LM = nR^2$ where n = number of observations used to estimate auxiliary regression model (2) and R^2 is the coefficient of determination of this regression.

(Note that LM-statistic follows a χ^2 distribution with degrees of freedom r)

 Step 5: If $LM = \chi^2 \leq \chi_r^2$ at $\alpha\%$

Then we do not reject H_0

If $LM = \chi^2 > \chi_r^2$ at $\alpha\%$

Then we reject H_0

And we conclude that there is significant evidence of heteroskedasticity in data.

[If we have corresponding p-value then we can use the p-value approach]

Goldfeld -Quandt Test

- Goldfeld and Quandt (1965) proposed a test of heteroskedasticity that may be applied when **one of the explanatory variables is suspected to be the heteroskedasticity culprit**.
- The basic idea behind this test is that if the variances of the disturbances are the same across all observations (i.e., homoskedasticity), then the variance of one part of the sample should be the same as the variance of another part of the sample. Under this test, we start by assuming that σ_i^2 is proportional to the size of X_i . It is also assumed that the disturbance term of the model (ϵ_i) is normally distributed and satisfies other regression assumptions.
- The hypothesis are
 - Null hypothesis (H_0): $\text{Var}(\epsilon_i | X_i) = \sigma^2$, a constant (Homoskedasticity)
 - Alternative hypothesis (H_1): $\text{Var}(\epsilon_i | X_i) = \sigma_i^2$, a variable (Heteroskedasticity)

Steps

- Step 1: Identify the variable to which the variance of disturbance term is suspected to be related.
- Step 2: Sort the raw data in ascending order (starting with lowest and going to be highest) of the values of X_i .
- Step 3: Cut out some central observations (c), breaking data set in two distinct sets-first one with low values of X_i . Note that there is no clear rule about how many observations to be cut out. Goldfeld and Quandt suggested the value of c as 8 if the sample size is about 30.
- Step 4: Fit separate regression by OLS to the first and last $(n-c)/2$ observations.
- Step 5: Compute residual sum squares for the two regressions, which are denoted by RSS_1 and RSS_2 , respectively.

Step 6: Compute the F-ratio of two RSSs as

$$F = \frac{RSS_2 / df_2}{RSS_1 / df_1}$$

Where $df_1 = df_2 = (n-c)/2 - k = (n-c-2k)/2$
(here k is no of parameters to be estimated)

Step 7: Compute critical value of F with df_1 , df_2 and level of significance (α).

If cal F > Critical value of F

We reject H_0

Then we conclude that there is heteroskedasticity in data

Limitations

- 1) The determination of an appropriate value of c sometimes becomes difficult.
- 2) In a model that involves a large number of explanatory variables, there may be difficulty in identifying the X-variable with which to sort the data.
- 3) It does not consider cases where heteroskedasticity is caused by more than one explanatory variable.

Remedial Techniques

Log Transformation

Weighted least squares method

Generalized Least Squares (GLS)

- Consider the model
- $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$ -----(1)
- And suppose that there is heteroskedasticity so that $\text{Var}(\epsilon_i) = \sigma_i^2$.
- If we know specifically the form of heteroskedasticity we may utilize such information to suitably transform our model to overcome the problem of heteroskedasticity and obtain efficient estimates for unknown parameters of our model.

Case a: Suppose heteroskedasticity is of the form $\sigma_i^2 = \sigma^2 Z_i^2$

Now dividing the model (1) by Z_i , so that the transformed model is

$$\frac{Y_i}{Z_i} = \beta_1 \frac{1}{Z_i} + \beta_2 \frac{X_{2i}}{Z_i} + \beta_3 \frac{X_{3i}}{Z_i} + \frac{\varepsilon_i}{Z_i}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(2)}$$

Then there is no heteroskedasticity in the transformed model (2) because

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{Z_i}\right)^2 = \frac{E(\varepsilon_i^2)}{Z_i^2} = \frac{\sigma_i^2}{Z_i^2} = \frac{\sigma^2 Z_i^2}{Z_i^2} = \sigma^2 \text{ (Constant)}$$

Case b: Heterokedasticity is of the form $\sigma_i^2 = \sigma^2 X_{1i}$

In this case we transform model (1) by dividing through $\sqrt{X_{1i}}$

$$\frac{Y_i}{\sqrt{X_{1i}}} = \beta_1 \frac{1}{\sqrt{X_{1i}}} + \beta_2 \frac{X_{2i}}{\sqrt{X_{1i}}} + \beta_3 \frac{X_{3i}}{\sqrt{X_{1i}}} + \frac{\varepsilon_i}{\sqrt{X_{1i}}}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(3)}$$

|

For the transformed model, the variance of the disturbance term is again constant

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{\sqrt{X_{1i}}}\right)^2 = \frac{E(\varepsilon_i^2)}{X_{1i}} = \frac{\sigma_i^2}{X_{1i}} = \frac{\sigma^2 X_{1i}}{X_{1i}} = \sigma^2 \text{ (Constant)}$$

Thus the transformed model (3) is free from the problem of heteroskedasticity.

Case c Heteroskedasticity is of the form $\sigma^2 X_{1i}^2$

Here the transformed model is

$$\frac{Y_i}{X_{1i}} = \beta_1 \frac{1}{X_{1i}} + \beta_2 \frac{X_{2i}}{X_{1i}} + \beta_3 \frac{X_{3i}}{X_{1i}} + \frac{\varepsilon_i}{X_{1i}}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + v_i \text{ -----(3)}$$

Then there is no heteroskedasticity in the transformed model (3) because

$$\text{Var}(v_i) = E(v_i^2) = E\left(\frac{\varepsilon_i}{X_{1i}}\right)^2 = \frac{E(\varepsilon_i)^2}{X_{1i}^2} = \frac{\sigma^2 X_{1i}^2}{X_{1i}^2} = \sigma^2 \text{ (Constant)}$$

The above procedures of estimating the transformed model to overcome the heteroskedasticity problem are known as generalized least squares.

Weighted Least Squares (WLS)

The above cases are also referred to as cases of weighted least squares (WLS) method. This is because we may view $\frac{1}{Z_i}$ in (1), $\frac{1}{\sqrt{X_{1i}}}$ in (2) and $\frac{1}{X_{1i}}$ in (3) as weight (w_i)

Then above models can be expressed as

$$(w_i Y_i) = \beta_1 w_i + \beta_2 (w_i X_{2i}) + \beta_3 (w_i X_{3i}) + (\epsilon_i w_i) \text{ -----(4)}$$

Detection Techniques (in R)

Normal Q-Q plot checks

```
qqPlot(reg_hprice)
```

Breuch Pagan Test

- library(lmtest)
- bptest(reg)

```
bptest(reg_hprice)
```

White test

```
bptest(reg_hprice, ~fitted(reg_hprice)+I(fitted(reg_hprice)^2))
```

Log transformation

- ###Taking log

```
reg_lnhprice<-lm(log(price)~log(lotsize)+log(sqrft)+log(bdrms),  
data=hprice1)
```

BP test

```
bptest(reg_lnhprice)
```

White test

```
bptest(reg_lnhprice, ~fitted(reg_lnhprice)+I(fitted(reg_lnhprice)^2))
```

Another Example: SMOKE.dta (Wooldridge)

- `install.packages("wooldridge")`
- `library(wooldridge)`
- `data(smoke, package="wooldridge")`
`head(smoke)`
- `data(smoke, package = 'Wooldridge')`
- Or open SMOKE.dta file
- `dim(SMOKE)`
- `head(SMOKE)`

```
reg_smoke<-lm(cigs~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn, data=SMOKE)
summary(reg_smoke)
bptest(reg_smoke)
```

- `logu2<-log(resid(reg_smoke)^2)`
- `varreg<-lm(logu2~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn,
data=SMOKE)`
- `##Weight`
- `w<-1/exp(fitted(varreg))`
- `wls<-
lm(cigs~log(income)+log(cigpric)+educ+age+l(age^2)+restaurn,weight=w,
data=SMOKE)`
- `summary(wls)`
- `bptest(wls)`