Heteroscedasticity

1. NATURE OF HETEROSCEDASTICITY

Heteroscedasticity refers to unequal variances of the error ε_i for different observations. It may be visually revealed by a "funnel shape" in the plot of the residuals e_i against the estimates $\hat{}Y_i$ or against one of the independent variables X_k. Effects of heteroscedasticity are the following

- heteroscedasticity *does not* bias OLS coefficient estimates
- heteroscedasticity means that OLS standard errors of the estimates are incorrect (often underestimated); therefore statistical inference is invalid
- heteroscedasticity means that OLS is not the best (= most efficient, minimum variance) estimator of β

2. FORMAL DIAGNOSTIC TESTS FOR HETEROSCEDASTICITY

There are many diagnostic tests for heteroscedasticity. Tests vary with respect to the statistical assumptions required and their sensitivity to departure from these assumptions (robustness).

1. (Optional) Brown-Forsythe Test

Properties

This test is robust against even serious departures from normality of the errors.

Principle

Find out whether the error variance σ_i^2 increases or decreases with values of an independent variable X_k (or with values of the estimates ^Y) by the following procedure:

1. split the observations into 2 groups: one group with low values of X_k (or low values of ^Y) and another group with high values of X_k (or high values of ^Y)
2. calculate the median value of the residuals within each group, and the absolute deviations of the residuals from their group median
3. then do a t-test of the difference in the means of these absolute deviations between the two groups; the test statistic is distributed as t with (n-2) df where n is the total number of cases

An example is shown at the following link:

Exhibit: brown-Forsythe test with the Afifi & Clark depression data

2. Breusch-Pagan *aka* Cook-Weisberg Test

Properties

This is a large sample test; it assumes normality of errors; it assumes σ_i^2 is a specific function of one or several X_k.

Principle

Compare the SSR from regressing e_i^2 on the X_k to SSE from regressing of Y on the X_k, with each SS divided by its df; resulting ratio is distributed as χ^2 with p-1 df.
This is a large-sample test that assumes that the logarithm of the variance σ^2 of the error term e_i is a linear function of X.
The B-P test statistic is the quantity

$$\chi^2_{BP} = (SSR^*/(p-1) / (SSE/n)^2$$

where

> SSR* is the regression sum of squares of the regression of e^2 on the X_k
> SSE is the error sum of squares of the regression of Y on the X_k

When n is sufficiently large and σ^2 is constant, χ^2_{BP} is distributed as a chi-square distribution with 1 df. Large values of χ^2_{BP} lead to the conclusion that σ^2 is not constant.

B-P Test in STATA

STATA calls it the Cook-Weisberg test. The test is obtained with the option **hettest** used after **regress**. The STATA manual states

> **hettest** [*varlist*] performs 2 flavors of the Cook and Weisberg (1983) test for heteroscedasticity. This test amounts to testing **t=0** in Var(e) = σ^2exp(**zt**). If *varlist* is not specified, the fitted values are used for **z**. If varlist is specified, the variables specified are used for **z**.

- Exhibit: Cook-Weisberg heteroscedasticity test in STATA with the Afifi & Clark depression data

References

This test was developed independently by Breusch and Pagan (1979) and Cook and Weisberg (1983).

- Cook, R. D. and S. Weisberg. 1983. "Diagnostics for Heteroscedasticity in Regression." *Biometrika* 70:1-10.
- Breusch, T. S. and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47:1287-1294.

3. (Optional) Goldfeld-Quandt Test

Properties

Test does not assume a large sample.

Principle

Sort cases with respect to variable believed related to residual variance; omit about 20% middle cases; run separate regressions in the low group (obtain SSE_{low}) and high group (obtain SSE_{high}); test F-distributed ratio SSE_{high}/SSE_{low} with (N-d-2p)/2 df in both the numerator and the denominator (where N is the total number of cases, d is the number of omitted cases, and p is the total number of independent variables including the constant term).

Reference

- Wilkinson, Blank, and Gruber (1996:274-277).

3. REMEDIAL APPROACH I: TRANSFORMING Y

If heteroscedasticity is found the first strategy is to try finding a transformation of Y that stabilizes the error variance. One can try various transformations along the ladder of powers or estimate the optimal transformation using the Box-Cox procedure. One variant of the Box-Cox procedure automatically finds the optimal transformation of Y given a multiple regression model with p independent variables. (See STATA reference [R] **boxcox**. Note that transforming Y can change the regression relationship with the independent variables X_k.

4. (Optional) REMEDIAL APPROACH II: WEIGHTED LEAST SQUARES (WLS)

Weighted least squares is an alternative to finding a transformation that stabilizes Y. However WLS has drawbacks (explained at the end of this section). Because of this the robust standard errors approach explaine in Section 5 below has become more popular.

1. Principle of WLS

Unequal error variance implies that the variance-covariance matrix of the errors ε_i, $\boldsymbol{\sigma^2\{\varepsilon\}}$ =

σ_1^2	0	...	0
0	σ_2^2	...	0
...	...	...	...
0	0	...	σ_n^2

is such that the variance σ_i^2 of ε_i may be different for each observation. Errors are still assumed uncorrelated across observations. Hence the off-diagonal entries of $\boldsymbol{\sigma^2\{\varepsilon\}}$ are zeroes and the matrix is diagonal.

Assume (for sake of argument) that the σ_i^2 are known.

Then the weighted least squares (WLS) criterion is to minimize

$$Q_w = \Sigma_{i=1 \text{ to } n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_{p-1} X_{i,p-1})^2$$

where the weights $w_i=1/\sigma_i^2$ are inversely proportional to the σ_i^2; thus WLS gives *less weight* to observations with *large error variance*, and vice-versa.

2. WLS in Practice

1. Estimating the σ_i^2

In practice the σ_i^2 (and the weights w_i) are not known and must be estimated. The general strategy for estimating the σ_i^2 (and w_i) is

- estimate the regression of Y on the X_k with OLS and obtain the residuals e_i; then
  - e_i^2 is an estimator of σ_i^2
  - $|e_i|$ (the absolute value of e_i) is an estimator of σ_i
- on the basis of visual evidence (residual plots), regress either e_i^2 (to estimate the *variance function*) or $|e_i|$ (to estimate the *standard deviation function*) on
  - one X_k, or
  - several X_k, or
  - ^Y (from the OLS regression), or
  - a polynomial function of any of the above
- the fitted value (estimate) from the regression is an estimate
  - $\hat{v}_i$ of the variance σ_i^2 (if dependent variable is e_i^2), or
  - $\hat{s}_i$ of the standard deviation σ_i (if dependent variable is $|e_i|$)
- calculate the weights w_i as either
  - $w_i = 1/(\hat{s}_i)^2$ (if $\hat{s}_i$ was estimated), or
  - $w_i = 1/\hat{v}_i$ (if $\hat{v}_i$ was estimated)

2. Estimating the WLS Regression

Having estimated the w_i, the WLS regression can be done either

- using a WLS-capable program, by simply providing the program with a variable containing the weights, say w; the program automatically minimizes Q_w; for example, in SYSTAT enter the command **weight=w** prior to the regression
- using OLS, by multiplying each variable (both dependent and independent, including the constant) by *the square root of the w_i* corresponding to a given observation and running an OLS regression without a constant with the transformed data

These steps can be iterated more than once until the estimates converge (= Iteratively Reweighted Least Squares - IRLS).

3. Examples of WLS Estimation

Example 1

The following exhibits replicate the analysis of blood pressure as function of age in ALSM5e pp. <>; ALSM4e pp. <406-407>.

- Exhibit: Scatter plot of blood pressure by age (cf NKNW Figure 10.1 (a) p. 406)
- Exhibit: Scatter plot of residual by estimate (equivalent here to plot of residual by age cf NKNW Figure 10.1 (b) p. 406)
- Exhibit: Scatter plot of absolute residual by age (cf NKNW Figure 10.1(c) p. 406)
- Exhibit: Scatter plot of squared residual by age
- Exhibit: SYSTAT program replicating weighted least squares estimation of blood presure

example (cf NKNW pp. 406-407)

Example 2

The following exhibit carries out a WLS analysis of the depression model with the Afifi & Clark data.

- Exhibit: WLS estimation of the depression model

3. Weighted Least Squares (WLS) as Generalized Least Squares (GLS)

In this section we show that WLS is a special case of a more general approach called Generalized Least Squares (GLS).

1. Matrix Representation of WLS

Assume the variance-covariance matrix of ε, $\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}$ as above, with diagonal elements σ_i^2 and zeros elsewhere.
The matrix W of weights $w_i = 1/\sigma_i^2$ is defined as W =

w_1	0	...	0
0	w_2	...	0
...	...	...	...
0	0	...	w_n

Then the WLS estimator of **β**, $\mathbf{b}_W$ is given by

$$(\mathbf{X'WX})\mathbf{b}_W = \mathbf{X'WY} \quad \text{(normal equations)}$$
$$\mathbf{b}_W = (\mathbf{X'WX})^{-1}\mathbf{X'WY}$$

Likewise one can show that

$$\boldsymbol{\sigma}^2\{\mathbf{b}_W\} = \sigma^2(\mathbf{X'WX})^{-1}$$
$$\mathbf{s}^2\{\mathbf{b}_W\} = MSE_W(\mathbf{X'WX})^{-1}$$
$$MSE_W = \Sigma w_i(Y_i - \hat{}Y_i)^2/(n - p)$$

The WLS estimates can also be obtained by applying OLS to the data transformed by the "square root" $\mathbf{W}^{1/2}$ of $\mathbf{W}$, where $\mathbf{W}^{1/2}$ contains the square roots of the w_i on the diagonal, and zeros elsewhere.
Since $\mathbf{W}^{1/2}$ is symmetric and $\mathbf{W}^{1/2}\mathbf{W}^{1/2} = \mathbf{W}$, it follows that

$$((\mathbf{W}^{1/2}\mathbf{X})'(\mathbf{W}^{1/2}\mathbf{X}))^{-1}(\mathbf{W}^{1/2}\mathbf{X})'(\mathbf{W}^{1/2}\mathbf{Y})$$
$$= (\mathbf{X'W}^{1/2}\mathbf{W}^{1/2}\mathbf{X})^{-1}(\mathbf{X'W}^{1/2}\mathbf{W}^{1/2}\mathbf{Y})$$
$$= (\mathbf{X'WX})^{-1}(\mathbf{X'WY}) = \mathbf{b}_W$$

Thus one can obtain $\mathbf{b}_W$ by multiplying **Y** and **X** by the square root of the weight and applying OLS to the transformed data.

2. WLS is a Special Case of Generalized Least Squares (GLS)

The standard regression model $\mathbf{Y} = \mathbf{Xb} + \varepsilon$ assumes that the variance-covariance matrix of the ε_i is scalar, that is $E\{\varepsilon\varepsilon'\} = \sigma^2\mathbf{I}$. Then the OLS estimator

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

has variance matrix

$$\sigma^2\{\mathbf{b}\} = E\{\mathbf{bb'}\} = E\{(\mathbf{X'X})^{-1}\mathbf{X'YY'X}(\mathbf{X'X})^{-1}\}$$
$$\sigma^2\{\mathbf{b}\} = (\mathbf{X'X})^{-1}\mathbf{X'}E\{\mathbf{YY'}\}\mathbf{X}(\mathbf{X'X})^{-1}$$
$$\sigma^2\{\mathbf{b}\} = (\mathbf{X'X})^{-1}\mathbf{X'}E\{\varepsilon\varepsilon'\}\mathbf{X}(\mathbf{X'X})^{-1}$$

When the error variance is the same for all observations (homoscedasticity) then the well-known result for OLS follows:

$$\sigma^2\{\mathbf{b}\} = (\mathbf{X'X})^{-1}\mathbf{X'}\sigma^2\mathbf{IX}(\mathbf{X'X})^{-1} \quad \text{(because } E\{\varepsilon\varepsilon'\} = \sigma^2\mathbf{I}\text{)}$$
$$\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}$$
$$\sigma^2\{\mathbf{b}\} = \sigma^2(\mathbf{X'X})^{-1} \quad \text{(after cancellation)}$$

And the covariance matrix of errors is estimated as before as

$$\mathbf{s}^2\{\mathbf{b}\} = MSE(\mathbf{X'X})^{-1} \quad \text{(estimating } \sigma^2 \text{ as MSE)}$$

and the OLS estimator **b** is the BLUE of $\boldsymbol{\beta}$ by the Gauss-Markov theorem.
When $E\{\varepsilon\varepsilon'\}$ is *not* scalar, it must be represented as $E\{\varepsilon\varepsilon'\} = \Omega$ where Ω is a (positive definite) symmetric matrix. Then OLS is no longer the BLUE of $\boldsymbol{\beta}$. Instead, Aitken's (or Generalized Least Squares) theorem states that the BLUE of $\boldsymbol{\beta}$ is

$$\mathbf{b}_{GLS} = (\mathbf{X'}\Omega^{-1}\mathbf{X})^{-1}\mathbf{X'}\Omega^{-1}\mathbf{Y}$$

where $\mathbf{b}_{GLS}$ is termed the *generalized least squares* (GLS) estimator.
The matrix Ω is usually unknown. When it is possible to estimate Ω from the data, the resulting estimator is

$$\mathbf{b}_{EGLS} = (\mathbf{X'}\hat{}\Omega^{-1}\mathbf{X})^{-1}\mathbf{X'}\hat{}\Omega^{-1}\mathbf{Y}$$

where $\hat{}\Omega$ denotes the estimated matrix Ω. $\mathbf{b}_{EGLS}$ is termed the *estimated generalized least squares* (EGLS) or *feasible generalized least squares* (FGLS) estimator.
It may be possible to derive a "square root" of $\hat{}\Omega^{-1}$, i.e. a symmetric matrix $\hat{}\Omega^{-1/2}$ such that $(\hat{}\Omega^{-1/2})(\hat{}\Omega^{-1/2}) = \hat{}\Omega^{-1}$. Then an alternative procedure for EGLS estimation is to premultiply **X** and **Y** by $\hat{}\Omega^{-1/2}$ and use OLS with the transformed data.
In practice, GLS (or EGLS/FGLS) is used when one has an *a priori* hypothesis concerning the structure of Ω. For example

- in the heteroscedasticity case one assumes that Ω is a diagonal matrix with elements σ_i^2 repressenting the error variance for observation i. Then one only has to estimate the n error variances σ_i^2 to estimate Ω. One can see that WLS is a special case of EGLS, with $\hat{}\Omega^{-1} = \mathbf{W}$.
- in regression models for time series data with a first order autoregressive error structure the

entries of the Ω matrix decrease exponentially away from the diagonal (see Module 14). On the basis of this systematic pattern one can estimate the matrix Ω and estimate **β** by EGLS.

- in regression models for panel data in which one has t observations over time on n individual units, one assumes that the error terms contains components that are specific to each unit and/or each time period. Then Ω has a distinctive block-diagonal structure that can be reconstructed by estimating a small number of parameters. Again one can estimate Ω and estimate **β** by EGLS.

4. Recommendations on WLS

The WLS approach to heteroscedasticity has at least two drawbacks.

1. WLS usually necessitates strong assumptions about the nature of the error variance, e.g. that it is a function of particular X variable or of ^Y. Sometimes the assumption appears reasonable (e.g., error variance is proportional to population size, when the units are areal units); other times it is not.
2. WLS produces an alternative unbiased estimate of **β**; but the OLS estimate is also unbiased. When $\mathbf{b}_{OLS}$ and $\mathbf{b}_{WLS}$ differ, which one should one choose?

Today researchers tend to prefer the robust standard errors approach to heteroscedasticity explained next.

5. REMEDIAL APPROACH III: ROBUST STANDARD ERRORS

The following discussion relies heavily on Long and Ervin (2000).

1. Principle of Robust Standard Errors

When heteroscedasticity is present transforming the variables or the use of WLS may be undesirable when

- a transformation of the variables that stabilizes the variances cannot be found
- a suitable transformation is found, but the resulting non-linear model is difficult to interpret substantively
- the weights to use in WLS cannot be found, as when the functional form of the heteroscedasticity is not known

The alternative strategy can be used even when the form of the heteroscedasticity is unknown. It consists of

1. estimating **b** using OLS as usual
2. use a *heteroscedasticity consistent covariance matrix* (HCCM) to estimate the standard errors of the estimates; these standard errors are then called *robust standard errors*

There are 3 variants of the strategy, labelled HC1, HC2, and HC3. To explain the principle of HCCM start with the usual multiple regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where E{ε} = **0** and E{εε'} = Ω is a positive definite matrix.
Then the covariance matrix of the OLS estimate $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ is

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = (\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1}$$

When the errors are homoscedastic, $\Omega = \sigma^2\mathbf{I}$ and the expression for $\boldsymbol{\sigma}^2\{\mathbf{b}\}$ reduces to the usual

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \sigma^2(\mathbf{X'X})^{-1}$$
$$\text{OLSCM} = \mathbf{s}^2\{\mathbf{b}\} = \text{MSE}(\mathbf{X'X})^{-1} \quad (\text{where MSE} = \Sigma e_i^2/(n-p))$$

OLSCM denotes the usual OLS covariance matrix of estimates.

2. Huber-White Robust Standard Errors HC1

The basic idea of robust standard errors is that when the errors are heteroscedastic one can estimate the observation-specific variance σ_i^2 with the single observation on the residual as

$$\hat{}\Omega_{ii} = (e_i - 0)^2/1 = e_i^2$$
$$\hat{}\Omega = \text{diag}\{e_i^2\}$$

This leads to the HCCM

$$\text{HC1} = (n/(n-p))\ (\mathbf{X'X})^{-1}\mathbf{X'}\text{diag}\{e_i^2\}\mathbf{X}(\mathbf{X'X})^{-1}$$

where n/(n-p) is a degree of freedom correction factor that becomes negligible for large samples.
HC1 is called the Huber-White estimator (after Huber 1967; White 1980) or the "sandwich" estimator because of the appearance of the formula. (See it?)
HC1 is obtained in STATA using the **robust** option (e.g., **regress y x1 x2, robust**).

3. HC2

An alternative to HC1 proposed by MacKinnon and White (1985) is to use a better estimate of the variance of ε_i based on $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$ where h_{ii} represent the leverage of observation i (diagonal element of the hat matrix H); the alternative formula divides the squared residual by $(1 - h_{ii})$

$$\text{HC2} = (\mathbf{X'X})^{-1}\mathbf{X'}\text{diag}\{e_i^2/(1 - h_{ii})\}\mathbf{X}(\mathbf{X'X})^{-1}$$

HC2 is obtained in STATA using the **hc2** option (e.g., **regress y x1 x2, hc2**).

4. HC3

A third possibility has a less straightforward theoretical motivation (Long and Ervin 2000; although compare the formula for HC3 with that for the deleted residual d_i in Module 10). The idea is to "overcorrect" for high variance residuals by dividing the squared residual by $(1 - h_{ii})^2$. This yields

$$\text{HC3} = (\mathbf{X'X})^{-1}\mathbf{X'}\text{diag}\{e_i^2/(1 - h_{ii})^2\}\mathbf{X}(\mathbf{X'X})^{-1}$$

HC3 is obtained in STATA using the **hc3** option (e.g., **regress y x1 x2, hc3**).

5. Relative Performance of HC1, HC2 and HC3 Robust Variance Estimators

Long and Erwin (2000) conclude from an extensive series of computer simulations that the HC3 gives the best results overall in small samples in the presence of heteroscedasticity of various forms. They state

> "1. If there is an a priori reason to suspect that there is heteroscedasticity, HCCM-based tests should be used."
> "2. For samples less than 250, HC3 should be used; when samples are 500 or larger, other versions of the HCCM can also be used. The superiority of HC3 over HC2 lies in its better properties when testing coefficients that are most strongly affected by heteroscedasticity."
> "3. The decision to correct for heteroscedasticity should not be based on the results of a screening test for heteroscedasticity."

"Given the relative costs of correcting for heteroscedasticity using HC3 when there is homoscedasticity and using OLSCM tests when there is heteroscedasticity, we recommend that HC3-based tests should be used routinely for testing individual coefficients in the linear regression model."

6. Example of Robust Standard Errors Estimation

The following exhibit shows the use of the HC1 (**robust**), HC2 (**hc2**) and HC3 (**hc3**) robust standard errors with STATA

- Exhibit (REPEAT): Robust standard error estimation in STATA for the depression model - Afifi & Clark data
- Exhibit: Summary comparison of OLS, HC1, HC2, and HC3 estimation for the depression model - Afifi & Clark data

6. CONCLUSION: DEALING WITH HETEROSCEDASTICITY

Provisional guidelines for dealing with the possibility of heteroscedasticity are

1. look at the plot of OLS residuals against estimates; if there is a suggestion of a funnel shape use a test of heteroscedasticity; use the Breusch-Pagan a.k.a. Cook-Weisberg test as it is easy to do in STATA; use one of the other tests (modified Levene or Goldfeld-Quandt) if you have a reason to, such as a small sample or doubts about normality of errors
2. if there is heteroscedasticity look first for a reasonable transformation that might stabilize the variances of the errors, but without introducing problems of interpretation or upsetting the functional relationship of Y with the independent variables; if such a transformation is found it is a desirable solution
3. if a suitable transformation cannot be found, investigate the possibility of WLS; try estimating the variance function or the standard deviation function; if a convincing function is found (one that has substantial R^2 and/or one that makes substantive sense, such as when the error variance is proportional to some measure of the size of the unit) then try WLS; otherwise, use the robust standard error approach instead (next)
4. if the transformation approach and the WLS approach do not seem promising, then use the robust standard errors approach; follow the recommendations of Long and Ervin (2000) to choose between HC1, HC2 and HC3, at least until someone comes up with evidence to the contrary; alternatively, adopt this approach right away after failing to find a good variance-stabilizing transformation, bypassing WLS

Last modified 17 Apr 2006