Heteroskedasticity

Occurs when the Gauss Markov assumption that the residual variance is constant across all observations in the data set

Heteroskedasticity

Occurs when the Gauss Markov assumption that the residual variance is constant across all observations in the data set

so that $E(u_i^2/X_i) \neq \sigma^2 \quad \forall i$

Heteroskedasticity

Occurs when the Gauss Markov assumption that the residual variance is constant across all observations in the data set

so that $E(u_i^2/X_i) \neq \sigma^2 \quad \forall i$

In practice this means the spread of observations at any given value of X will not now be constant

Heteroskedasticity

Occurs when the Gauss Markov assumption that the residual variance is constant across all observations in the data set

so that $E(u_i^2/X_i) \neq \sigma^2 \quad \forall i$

In practice this means the spread of observations at any given value of X will not now be constant

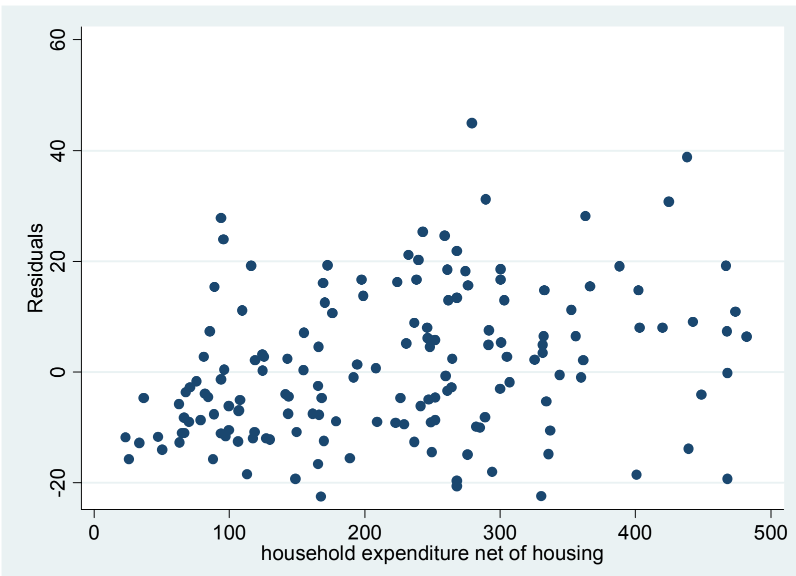
Eg. food expenditure is known to vary much more at higher levels of income than at lower levels of income, the level of profits tends to vary more across large firms than across small firms)

Example: the data set food.dta contains information on food expenditure and income. A graph of the *residuals* from a regression of food spending on total household expenditure clearly that the residuals tend to be more spread out at higher levels of income – this is typical pattern associated with heteroskedasticity.

```
. reg food expnethsum

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  107.19
       Model |  22490.0823     1  22490.0823           Prob > F      =  0.0000
    Residual |  41544.8096   198  209.822271           R-squared     =  0.3512
-------------+------------------------------           Adj R-squared =  0.3479
       Total |  64034.8918   199  321.783376           Root MSE      =  14.485

------------------------------------------------------------------------------
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  expnethsum |   .0355189   .0034308    10.35   0.000     .0287534    .0422844
       _cons |   28.55002    1.56964    18.19   0.000     25.45466    31.64537
------------------------------------------------------------------------------
. predict res, resid
. two (scatter res expnet if expnet<500)
```



Consequences of Heteroskedasticity

Can show:

1. OLS estimates of coefficients remains unbiased (as with autocorrelation)

Consequences of Heteroskedasticity

Can show:

1. OLS estimates of coefficients remains unbiased (as with autocorrelation)

- since given

$$Y_i = b_0 + b_1 X_i + u_I$$

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased
(as with autocorrelation)

- since given

$$Y_i = b_0 + b_1X_i + u_I$$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)}$$

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased
(as with autocorrelation)

- since given

$$Y_i = b_0 + b_1 X_i + u_I$$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)}$$ sub in $Y_i = b_0 + b_1 X_i + u_I$

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased
   (as with autocorrelation)

   - since given

$$Y_i = b_0 + b_1 X_i + u_I$$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)} \quad \text{sub in } Y_i = b_0 + b_1 X_i + u_I \Rightarrow b_1 + \frac{Cov(X,u)}{Var(X)}$$

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased
(as with autocorrelation)

- since given

$Y_i = b_0 + b_1X_i + u_I$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)} \quad \text{sub in } Y_i = b_0 + b_1X_i + u_I \quad \Rightarrow b_1 + \frac{Cov(X,u)}{Var(X)}$$

heteroskedasticity assumption does not affect Cov(X,u) = 0
needed to prove unbiasedness,

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased (as with autocorrelation)

- since given

$$Y_i = b_0 + b_1 X_i + u_I$$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)} \quad \text{sub in } Y_i = b_0 + b_1 X_i + u_I \quad \Rightarrow b_1 + \frac{Cov(X,u)}{Var(X)}$$

heteroskedasticity assumption does not affect Cov(X,u) = 0 needed to prove unbiasedness, so OLS estimate of **coefficients** remains unbiased in presence of heteroskedasticity

Consequences of Heteroskedasticity

Can show:

1) OLS estimates of coefficients remains unbiased
(as with autocorrelation)

- since given

$$Y_i = b_0 + b_1 X_i + u_I$$

and

$$\hat{b}_1^{ols} = \frac{COV(X,Y)}{Var(X)} \qquad = b_1 + \frac{Cov(X,u)}{Var(X)}$$

and heteroskedasticity assumption does not affect Cov(X,u) = 0 needed to prove unbiasedness, so OLS estimate of **coefficients** remains unbiased in presence of heteroskedasticity

but

2) can show that heteroskedasticity (like autocorrelation) means the OLS estimates of the standard errors (and hence t and F tests) are biased.

2) can show that heteroskedasticity (like autocorrelation) means the OLS estimates of the standard errors (and hence t and F tests) are biased.

(intuitively, if all observations are distributed unevenly about the regression line then OLS is unable to distinguish the “quality” of the observations - observations further away from the regression line should be given less weight in the calculation of the standard errors (since they are more unreliable) but OLS can’t do this, so the standard errors are biased).

Testing for Heteroskedasticity

1. Residual Plots

In absence of Heteroskedasticity there should be no obvious pattern to the **spread** of the residuals, so useful to plot the residuals against the X variable thought to be causing the problem,

Testing for Heteroskedasticity

1. Residual Plots

In absence of Heteroskedasticity there should be no obvious pattern to the **spread** of the residuals, so useful to plot the residuals against the X variable thought to be causing the problem,

- assuming you know which X variable it is (often difficult)

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

i) Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)
ii) Drop the middle “c” observations where c is approximately 30% of your sample

Fine if certain which variable causing the problem, less so if unsure.

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

i) Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)

ii) Drop the middle "c" observations where c is approximately 30% of your sample

iii) Run separate regressions for the high and low variance sub-samples

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

i) Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)

ii) Drop the middle "c" observations where c is approximately 30% of your sample

iii) Run separate regressions for the high and low variance sub-samples

iv) Compute

$$F = \frac{RSS_{high\,\mathrm{var}iance\,sub-sample}}{RSS_{low\,\mathrm{var}iance\,sub-sample}} \sim F\left[\frac{N-c-2k}{2}, \frac{N-c-2k}{2}\right]$$

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

iii) Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)

iv) Drop the middle "c" observations where c is approximately 30% of your sample

v) Run separate regressions for the high and low variance sub-samples

vi) Compute

$$F = \frac{RSS_{high\,variance\,sub-sample}}{RSS_{low\,variance\,sub-sample}} \sim F\left[\frac{N-c-2k}{2}, \frac{N-c-2k}{2}\right]$$

vii) If estimated F>Fcritical, **reject** null of no heteroskedasticity (intuitively the residuals from the high variance sub-sample are much larger than the residuals from the high variance sub-sample)

2. Goldfeld-Quandt

Again assuming know which variable is causing the problem then can test formally whether the residual spread varies with values of the suspect X variable.

viii) Order the data by the size of the X variable and split the data into 2 equal sub-groups (one high variance the other low variance)

ix) Drop the middle “c” observations where c is approximately 30% of your sample

x) Run separate regressions for the high and low variance sub-samples

xi) Compute

$$F = \frac{RSS_{high\,variance\,sub-sample}}{RSS_{low\,variance\,sub-sample}} \sim F\left[\frac{N-c-2k}{2}, \frac{N-c-2k}{2}\right]$$

xii) If estimated F>Fcritical, **reject** null of no heteroskedasticity (intuitively the residuals from the high variance sub-sample are much larger than the residuals from the high variance sub-sample)

Fine if certain which variable causing the problem, less so if unsure.

Breusch-Pagan Test

In most cases involving more than one right hand side variable it is unlikely that you will know which variable is causing the problem.

Breusch-Pagan Test

In most cases involving more than one right hand side variable it is unlikely that you will know which variable is causing the problem.

A more general test is therefore to regress an approximation of the (unknown) residual variance on **all** the right hand side variables and test for a significant causal effect (if there is then you suspect heteroskedasticity)

Breusch-Pagan Test

In most cases involving more than one right hand side variable it is unlikely that you will know which variable is causing the problem.

A more general test is therefore to regress an approximation of the (unknown) residual variance on **all** the right hand side variables and test for a significant causal effect (if there is then you suspect heteroskedasticity)

A test that is valid asymptotically (ie in large samples) that does not rely on knowing which variable is causing the problem is the Breusch-Pagan test

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals $\hat{u}$

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals $\hat{u}$

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals $\hat{u}$

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

$$\hat{u}^2_t = g + g_1X_1 + g_2X_2 + u_i \qquad (2)$$

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

$$\hat{u}^2_t = g + g_1X_1 + g_2X_2 + u_i \qquad (2)$$

Using (2) either compute

$$F = \frac{R^2{}_{auxillary} / k-1}{(1-R^2{}_{auxillary}) / N-k} \sim F[k-1, N-k]$$

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

$$\hat{u}^2_t = g + g_1X_1 + g_2X_2 + u_i \qquad (2)$$

Using (2) either compute

$$F = \frac{R^2{}_{auxillary} / k-1}{(1 - R^2{}_{auxillary}) / N - k} \sim F[k-1, N-k]$$

ie test of goodness of fit for the model in this auxillary regression

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

$$\hat{u}^2_t = g + g_1X_1 + g_2X_2 + u_i \qquad (2)$$

Using (2) either compute

$$F = \frac{R^2{}_{auxillary} / k-1}{(1-R^2{}_{auxillary}) / N-k} \sim F[k-1, N-k]$$

ie test of goodness of fit for the model in this auxillary regression

Given

$$Y_i = a + b_1X_1 + b_2X_2 + u_i \qquad (1)$$

i) Estimate (1) by OLS and save residuals

ii) Square residuals and regress these on all the original X variables in (1) - these squared OLS residuals proxy the unknown true residual variance

$$\hat{u}^2_t = g + g_1X_1 + g_2X_2 + u_i \qquad (2)$$

Using (2) either compute

$$F = \frac{R^2{}_{auxillary} / k-1}{(1-R^2{}_{auxillary}) / N-k} \sim F[k-1, N-k]$$

ie test of goodness of fit for the model in this auxillary regression
or
compute $N^*R^2_{auxillary} \sim \chi^2(k-1)$

If F or $N^*R^2_{auxillary}$ > respective critical values reject null of no heterosked.

Example: Breusch-Pagan Test of Heteroskedastcity

The data set smoke.dta contains information on the smoking habits, wages age and gender of a cross-section of individuals

```
. u smoke.dta                      /* read in data */

.  reg lhw age age2 female smoke

      Source |       SS       df       MS              Number of obs =    7970
-------------+------------------------------           F(  4,  7965) =  284.04
       Model |  304.964893     4  76.2412233           Prob > F      =  0.0000
    Residual |  2137.94187  7965  .268417059           R-squared     =  0.1248
-------------+------------------------------           Adj R-squared =  0.1244
       Total |  2442.90677  7969  .306551232           Root MSE      =  .51809
------------------------------------------------------------------------------
         lhw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0728466   .0031712    22.97   0.000     .0666301    .0790631
        age2 |   -.000847   .0000382   -22.17   0.000    -.0009219   -.0007721
      female |  -.2583456   .0116394   -22.20   0.000    -.2811618   -.2355294
      smokes |  -.1501679   .0128866   -11.65   0.000    -.1754291   -.1249068
       _cons |   .8732505    .062907    13.88   0.000     .7499363    .9965646
------------------------------------------------------------------------------

/* save residuals */
. predict reshat, resid

. g reshat2=reshat^2                                    /* square them */

/* regress square of residuals on all original rhs variables */

. reg reshat2 age age2 female smoke

      Source |       SS       df       MS              Number of obs =    7970
-------------+------------------------------           F(  4,  7965) =    6.59
       Model |  13.2179958     4  3.30449895           Prob > F      =  0.0000
    Residual |  3996.90523  7965  .501808566           R-squared     =  0.0033
-------------+------------------------------           Adj R-squared =  0.0028
       Total |  4010.12323  7969  .503215363           Root MSE      =  .70838
```

```
------------------------------------------------------------------------------
     reshat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0012546    .004336     0.29   0.772    -.0072452    .0097544
        age2 |   .0000252   .0000522     0.48   0.630    -.0000772    .0001276
      female |   .0022702   .0159145     0.14   0.887    -.0289264    .0334668
      smokes |  -.0174587   .0176199    -0.99   0.322    -.0519983    .0170808
       _cons |   .1766929   .0860128     2.05   0.040     .0080854    .3453004
------------------------------------------------------------------------------
```

Breusch-Pagan test is $N*R^2$

```
. di 7970*.0033
26.301
```

which is chi-squared k-1 degrees of freedom (4 in this case) and the critical value is 9.48. So estimated value exceeds critical value

Similarly the F test for goodness of fit in stata output in the top right corner is test for joint significance of all the rhs variables in this model (excluding the constant)

From F tables, $F_{critical}^{5\%\ level}(4,7970) = 2.37$

So estimated F = 6.59 > Fcritical, so **reject** null of **no** heteroskedasticity
Or could use Stata's version of the Breusch-Pagan test

What to do if heteroskedasticity present?

1. Try different functional form
Sometimes taking logs of dependent or explanatory variable can reduce the problem

```
. reg food expnethsum if exp<1000
      Source |       SS       df       MS              Number of obs =     192
-------------+------------------------------           F(  1,   190) =  110.00
       Model |  21179.4196     1  21179.4196           Prob > F      =  0.0000
    Residual |  36583.9436   190  192.547072           R-squared     =  0.3667
-------------+------------------------------           Adj R-squared =  0.3633
       Total |  57763.3632   191  302.425986           Root MSE      =  13.876
------------------------------------------------------------------------------
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  expnethsum |   .0504532   .0048106    10.49   0.000     .0409641    .0599423
       _cons |   24.60655   1.770093    13.90   0.000     21.11499     28.0981
------------------------------------------------------------------------------
. bpagan expn

Breusch-Pagan LM statistic:   7.54351  Chi-sq( 1)  P-value =   .006
```

The Breusch-Pagan test indicates the presence of heteroskedasticity (estimated chi-squared value > critical value). This means the standard errors, t statistics etc are biased

If use the log of the dependent variable rather than in levels

```
. g lfood=log(food)
. reg lfood expnethsum

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =   93.00
       Model |  14.6377436     1  14.6377436           Prob > F      =  0.0000
    Residual |  31.1642937   198  .157395423           R-squared     =  0.3196
-------------+------------------------------           Adj R-squared =  0.3162
       Total |  45.8020374   199  .230160992           Root MSE      =  .39673
------------------------------------------------------------------------------
       lfood |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  expnethsum |   .0009062    .000094     9.64   0.000     .0007209    .0010915
       _cons |   3.290222   .0429903    76.53   0.000     3.205444    3.374999
------------------------------------------------------------------------------
. bpagan expn

Breusch-Pagan LM statistic:  .4280017  Chi-sq( 1)  P-value =   .513
```

2. Drop "Outliers"

Sometimes heteroskedasticity can be influenced by 1 or 2 observations in the data set which stand a long way from the main concentration of data - "outliers".

2. Drop "Outliers"

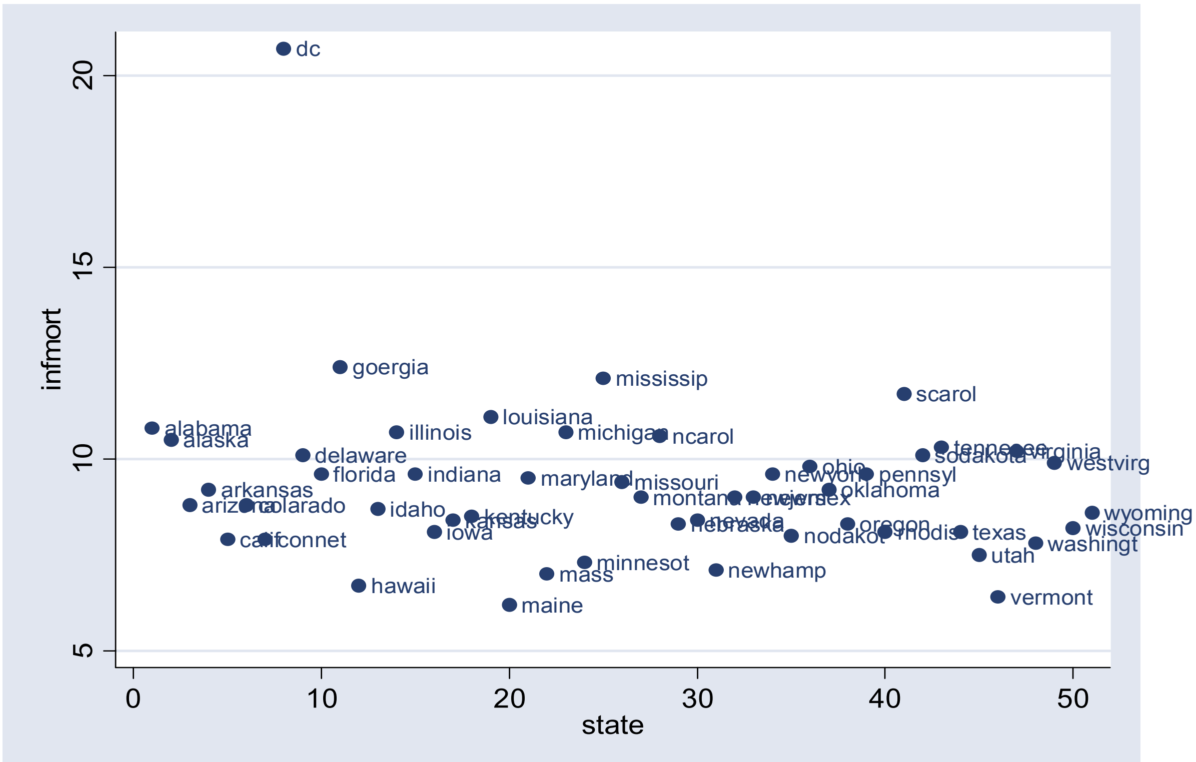
Sometimes heteroskedasticity can be influenced by 1 or 2 observations in the data set which stand a long way from the main concentration of data - "outliers".

Often these observations may be genuine – in which case you should not drop them – but sometimes they may be the result of measurement error or miscoding in which case you may have a case for dropping them.

Example

The data infmort.dta gives infant mortality for 51 U.S. states along with the number of doctors per capita ine ach state. A graph of infant mortality against number of doctors clearly shows that Washington D.C. is something of an outlier (it has lots of doctors but also a very high infant mortality rate)

```
. twoway (scatter infmort state, mlabel(state)), ytitle(infmort) ylabel(, labels) xtitle(state)
```



A regression of infant mortality on (the log of) doctor numbers for all 51 observations suffers from heteroskedasticity

```
. reg  infmort ldocs

      Source |       SS       df       MS              Number of obs =      51
-------------+------------------------------           F(  1,    49) =    4.08
       Model |  17.7855153     1  17.7855153           Prob > F      =  0.0488
```

```
    Residual |  213.461954    49   4.3563664           R-squared     =  0.0769
-------------+------------------------------           Adj R-squared =  0.0581
       Total |  231.247469    50  4.62494938           Root MSE      =  2.0872
------------------------------------------------------------------------------
     infmort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldocs |   2.130049   1.054189     2.02   0.049     .0115765    4.248522
       _cons |  -1.959674   5.572467    -0.35   0.727    -13.15797    9.238617
------------------------------------------------------------------------------

. bpagan ldocs

Breusch-Pagan LM statistic:  67.14974  Chi-sq( 1)  P-value =  2.5e-16
```

However if the outlier is excluded then

```
. reg  infmort ldocs if dc==0

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    5.13
       Model |  9.49879378     1  9.49879378           Prob > F      =  0.0280
    Residual |  88.8244081    48   1.8505085           R-squared     =  0.0966
-------------+------------------------------           Adj R-squared =  0.0778
       Total |  98.3232019    49  2.00659596           Root MSE      =  1.3603
------------------------------------------------------------------------------
     infmort |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldocs |  -1.915912   .8456428    -2.27   0.028    -3.616191   -.2156336
       _cons |   19.12582   4.448765     4.30   0.000     10.18098    28.07066
------------------------------------------------------------------------------

. bpagan ldocs

Breusch-Pagan LM statistic:  .0825086  Chi-sq( 1)  P-value =  .7739
```

Can see that the problem of heteroskedasticty disappears – though the D.C. observation is genuine so you need to think carefully about the benefits of dropping it against the costs.

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^{\,2}$ (and not say $\sigma^2 X_2^{\,3}$)

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X)$

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^2 Var(u_i)$

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^{\,2}$ (and not say $\sigma^2 X_2^{\,3}$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^2 Var(u_i) = 1/X_i^{2} * \sigma^2 X_i^2$

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^{\ 2}$ (and not say $\sigma^2 X_2^{\ 3}$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^{\ 2} Var(u_i) = 1/X_i^{\ 2} * \sigma^2 X_i^{\ 2} = \sigma^2$

So the variance of this **is** constant for all observations in the data set

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^2 Var(u_i) = 1/X_i^2 * \sigma^2 X_i^2 = \sigma^2$

So the variance of this **is** constant for all observations in the data set

This means if we divide all the observations by $1/X_i$

$$Y_i = b_0 + b_1 X_i + u_i \qquad (1)$$

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^2 Var(u_i) = 1/X_i^2 * \sigma^2 X_i^2 = \sigma^2$

So the variance of this **is** constant for all observations in the data set

This means if we divide all the observations by $1/X_i$

$$Y_i = b_0 + b_1 X_i + u_i \qquad (1)$$

becomes

$$Y_i/ X_i = b_0/ X_i + b_1 X_i/ X_i + u_i/ X_i \qquad (2)$$

2. Feasible GLS

If (and this is a big if) you think you know the exact functional form of the heteroskedasticity

eg you know that $var(u_i)=\sigma^2 X_1^2$ (and not say $\sigma^2 X_2^3$)

so that there is a common component to the variance, σ^2, and a part that rises with the square of the level of the variable X_1

Consider the term $Var(u_i/X) = 1/X_i^2 Var(u_i) = 1/X_i^{2} * \sigma^2 X_i^2 = \sigma^2$

So the variance of this **is** constant for all observations in the data set

This means if we divide all the observations by $1/X_i$

$$Y_i = b_0 + b_1 X_i + u_i \qquad (1)$$

becomes

$$Y_i / X_i = b_0 / X_i + b_1 X_i / X_i + u_i / X_i \qquad (2)$$

and the estimates of b_0 and b_1 in (2) will not be affected by heterosked.

This is called a Feasible Generalised Least Squares Estimator (FGLS) and will be more efficient than OLS

IF

This is called a Feasible Generalised Least Squares Estimator (FGLS) and will be more efficient than OLS

IF

The assumption about the form of heteroskedasticity is correct

This is called a Feasible Generalised Least Squares Estimator (FGLS) and will be more efficient than OLS

IF

The assumption about the form of heteroskedasticity is correct

If not the “solution” may be much worse than OLS

Example

```
. reg hourpay age

      Source |       SS       df       MS              Number of obs =   12098
-------------+------------------------------           F(  1, 12096) =  133.08
       Model |  5207.03058     1  5207.03058           Prob > F      =  0.0000
    Residual |  473292.608 12096  39.1280264           R-squared     =  0.0109
-------------+------------------------------           Adj R-squared =  0.0108
       Total |  478499.638 12097  39.5552317           Root MSE      =  6.2552
------------------------------------------------------------------------------
     hourpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0586134    .005081    11.54   0.000     .0486539    .0685729
       _cons |   6.168383   .2066433    29.85   0.000     5.763329    6.573437
------------------------------------------------------------------------------

. bpagan age

Breusch-Pagan LM statistic:  17.27396  Chi-sq( 1)  P-value =  3.2e-05
```

Test suggests heteroskedasticity present

Suppose you decide that heteroskedasticity is given by $var(u_i)=\sigma^2 Age_i$

So transform variables by dividing by SQUARE ROOT of Age
(including the constant)

```
. g ha=hourpay/sqrt(age)
. g aa=age/sqrt(age)
. g ac=1/sqrt(age)                  /* this is new constant term */

. reg ha aa ac, nocon

      Source |       SS       df       MS              Number of obs =   12098
-------------+------------------------------           F(  2, 12096) =10990.27
       Model |   22854.251     2  11427.1255           Prob > F      =  0.0000
    Residual |  12576.8073 12096  1.03974928           R-squared     =  0.6450
-------------+------------------------------           Adj R-squared =  0.6450
```

```
       Total |  35431.0584 12098  2.92867072           Root MSE      =  1.0197
------------------------------------------------------------------------------
          ha |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          aa |   .0932672   .0049446    18.86   0.000     .0835749    .1029594
          ac |   4.813435    .184437    26.10   0.000     4.451908    5.174961
------------------------------------------------------------------------------
```

If heteroskedastic assumption is correct these are the GLS estimates and should be preferred to OLS. If assumption is not correct they will be misleading.

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don’t know precise form of heteroskedasticity

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2{}_u}{N * Var(X)}$$

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2{}_u}{N * Var(X)}$$

Can show that true OLS variance in presence of heteroskedasticity is given by

3. **White adjustment (OLS robust standard errors)**

As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2{}_u}{N * Var(X)}$$

Can show that true OLS variance in presence of heteroskedasticity is given by

$$Var(\hat{\beta}_{ols}) = \frac{w_i s^2{}_u}{N * Var(X)}$$

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2{}_u}{N * Var(X)}$$

Can show that true OLS variance in presence of heteroskedasticity is given by

$$Var(\hat{\beta}_{ols}) = \frac{w_i s^2{}_u}{N * Var(X)}$$

where w_i depends on the distance of the X_i observation from the mean of X (distant observations have larger weight). This is the basis for the White correction.

3. **White adjustment (OLS robust standard errors)**
As with autocorrelation, best fix may be to make OLS standard errors unbiased (if inefficient) if don't know precise form of heteroskedasticity

In absence of heteroskedasticity we know OLS estimate of variance on any coefficient is

$$Var(\hat{\beta}_{ols}) = \frac{s^2{}_u}{N * Var(X)}$$

Can show that true OLS variance in presence of heteroskedasticity is given by

$$Var(\hat{\beta}_{ols}) = \frac{w_i s^2{}_u}{N * Var(X)}$$

where w_i depends on the distance of the X_i observation from the mean of X (distant observations have larger weight). This is the basis for the white correction.

Again should only really do this in **large** samples

Example

```
. reg food expnethsum

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  107.19
       Model |  22490.0823     1  22490.0823           Prob > F      =  0.0000
    Residual |  41544.8096   198  209.822271           R-squared     =  0.3512
-------------+------------------------------           Adj R-squared =  0.3479
       Total |  64034.8918   199  321.783376           Root MSE      =  14.485

------------------------------------------------------------------------------
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  expnethsum |   .0355189   .0034308    10.35   0.000     .0287534    .0422844
       _cons |   28.55002    1.56964    18.19   0.000     25.45466    31.64537
------------------------------------------------------------------------------

. reg food expnethsum, robust

Linear regression                                      Number of obs =     200
                                                       F(  1,   198) =   80.37
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.3512
                                                       Root MSE      =  14.485

------------------------------------------------------------------------------
             |               Robust
        food |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  expnethsum |   .0355189   .0039619     8.97   0.000     .0277059    .0433319
       _cons |   28.55002   1.549629    18.42   0.000     25.49412    31.60591
------------------------------------------------------------------------------
```

The Breusch-Pagan test indicates the presence of heteroskedasticity (estimated chi-squared value > critical value). This means the standard errors, t statistics etc are biased, so decide to fix up the standard errors using the white correction

Note that the OLS coefficients are unchanged, only the standard errors and t values change

Testing for Heteroskedasticity in Time Series Data (ARCH)

Sometimes what appears to be autocorrelation in time series data can be cause by heteroskedasticity

Testing for Heteroskedasticity in Time Series Data (ARCH)

Sometimes what appears to be autocorrelation in time series data can be cause by heteroskedasticity

The difference is that it is that heteroskedasicity means that *the squared values* of the residuals rather than the levels of the residuals are correlated over time as with autocorrelation

Testing for Heteroskedasticity in Time Series Data (ARCH)

Sometimes what appears to be autocorrelation in time series data can be cause by heteroskedasticity

The difference is that it is that heteroskedasicity means that *the squared values* of the residuals rather than the levels of the residuals are correlated over time as with autocorrelation

ie

$$u_t^2 = \beta u_{t-1}^2 + e_t \qquad (1)$$

Testing for Heteroskedasticity in Time Series Data (ARCH)

Sometimes what appears to be autocorrelation in time series data can be cause by heteroskedasticity

The difference is that it is that heteroskedasicity means that *the squared values* of the residuals rather than the levels of the residuals are correlated over time as with autocorrelation

ie

$$u_t^2 = \beta u_{t-1}^2 + e_t \qquad (1)$$

and not

$$u_t = \rho u_{t-1} + e_t \qquad (2)$$

Testing for Heteroskedasticity in Time Series Data (ARCH)

Sometimes what appears to be autocorrelation in time series data can be cause by heteroskedasticity

The difference is that it is that heteroskedasicity means that *the squared values* of the residuals rather than the levels of the residuals are correlated over time as with autocorrelation

ie

$$u_t^2 = \beta u_{t-1}^2 + e_t \qquad (1)$$

and not

$$u_t = \rho u_{t-1} + e_t \qquad (2)$$

Model (1) is called an Autoregressive Conditional Heteroskedasticty (ARCH) model

Testing for the presence of heteroskedasticity in time series data is very easy, just

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

ii) Save the estimated residuals, $\hat{u}_t$

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

ii) Save the estimated residuals, $\hat{u}_t$

iii) Square them, $\hat{u}^2_t$

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

ii) Save the estimated residuals, $\hat{u}_t$

iii) Square them, $\hat{u}^2_t$

iv) Regress the squared OLS residuals on their value lagged by 1 period and a constant

$$\hat{u}^2_t = g_0 + g_1\hat{u}^2_{t-1} + v_t$$

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

ii) Save the estimated residuals, $\hat{u}_t$

iii) Square them, $\hat{u}^2_t$

iv) Regress the squared OLS residuals on their value lagged by 1 period and a constant

$$\hat{u}^2_t = g_0 + g_1\hat{u}^2_{t-1} + v_t$$

v) If the t value on the lag is significant conclude that residual variances are correlated over time – there is heteroskedasticity

Testing for the presence of heteroskedasticity in time series data is very easy, just

i) Estimate the model $y_t = b_0 + b_1X_t + u_t$ by OLS

ii) Save the estimated residuals, $\hat{u}_t$

iii) Square them, $\hat{u}^2_t$

iv) Regress the squared OLS residuals on their value lagged by 1 period and a constant

$$\hat{u}^2_t = g_0 + g_1\hat{u}^2_{t-1} + v_t$$

v) If the t value on the lag is significant conclude that residual variances are correlated over time – there is heteroskedasticity

vi) If so adjust the standard errors using the white robust correction

Example: This example uses Wooldridge's (2000) data, (*stocks.dta*) on New York Stock Exchange price movements to test the efficient markets hypothesis.

EMH suggests that information on returns (ie the percentage change in the share price) in the week before should not predict the percentage change in this week's share price. A simple way to test this is to regress current returns on lagged returns.

```
. u shares

. reg return return1

      Source |       SS       df       MS              Number of obs =     689
-------------+------------------------------           F(  1,   687) =    2.40
       Model |  10.6866237     1  10.6866237           Prob > F      =  0.1218
    Residual |  3059.73813   687   4.4537673           R-squared     =  0.0035
-------------+------------------------------           Adj R-squared =  0.0020
       Total |  3070.42476   688  4.46282668           Root MSE      =  2.1104
------------------------------------------------------------------------------
      return |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     return1 |   .0588984   .0380231     1.55   0.122    -.0157569    .1335538
       _cons |    .179634   .0807419     2.22   0.026     .0211034    .3381646
```

In this example it would appear that lagged returns have little power in predicting current price changes.

However it may be that the t value is influenced by heteroskedasticity in the variance of the residuals.

```
. predict reshat, resid
. g reshat2=reshat^2                         /* ols residuals squared */
. g reshat21=reshat[_n-1]                    /* lagged one period */
```

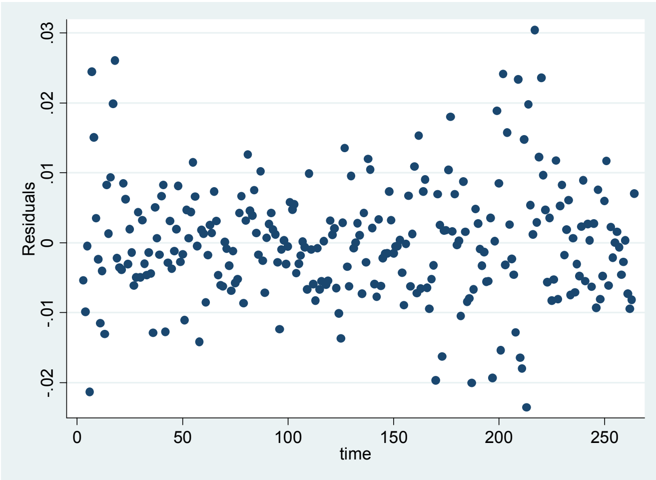
The graph of residuals over time, suggests heteroskedasticity may exist

```
Two (scatter reshat time  , yline(0))
```

As does the Breusch-Pagan test

```
. bpagan return1

 Breusch-Pagan LM statistic:  95.21722  Chi-sq( 1)  P-value =  1.7e-22
```



The ARCH test is found by a regression of the squared ols residuals on lagged values (in this case 1)

```
. reg return return1
. predict reshat, resid                    /* save residuals */

. g reshat2=reshat^2                       /* square residuals */
. g reshat21=reshat2[_n-1]                 /* lag by 1 period */

. reg reshat2 reshat21
      Source |       SS       df       MS              Number of obs =     688
-------------+------------------------------           F(  1,   686) =   87.92
       Model |  10177.7088     1  10177.7088           Prob > F      =  0.0000
    Residual |  79409.7826   686  115.757701           R-squared     =  0.1136
-------------+------------------------------           Adj R-squared =  0.1123
       Total |  89587.4914   687  130.403918           Root MSE      =  10.759
------------------------------------------------------------------------------
     reshat2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    reshat21 |   .3370622   .0359468     9.38   0.000     .2664833    .4076411
       _cons |   2.947434   .4402343     6.70   0.000     2.083065    3.811802
```

Since estimated t value on lagged dependent variable is highly significant reject null of homoskedasticity.

Need to fix up the standard errors in the original regression.