Heteroscedasticity:

Testing and Correcting in SPSS

1) Introduction

Recall that for estimation of coefficients and for regression inference to be correct we have to assume that:

1. Equation is correctly specified:
2. Error Term has zero mean
3. **Error Term has constant variance**
4. Error Term is not autocorrelated
5. **Explanatory variables are fixed**
6. No linear relationship between RHS variables

When assumption 3 holds, the errors u_i in the regression equation have common variance, and then we have what is called *homoscedasticity*, or a "scalar error covariance matrix" (assuming also that there is no autocorrelation), where "scalar" is another word for *constant*. When assumption 3 breaks down, we have the opposite of homoscedasticity: *heteroscedasticity*, or a "non-scalar error covariance matrix"

a) *Scalar Error Covariance Matrix*

Assumption 4 of OLS requirements states that the sampling distributions for each of the residuals are not correlated with any of the others. So, for any two observations, the residual terms are uncorrelated: $cov(u_1, u_2) = 0$; or more generally: $cov(u_i, u_j) = 0 \quad \forall i, j$.
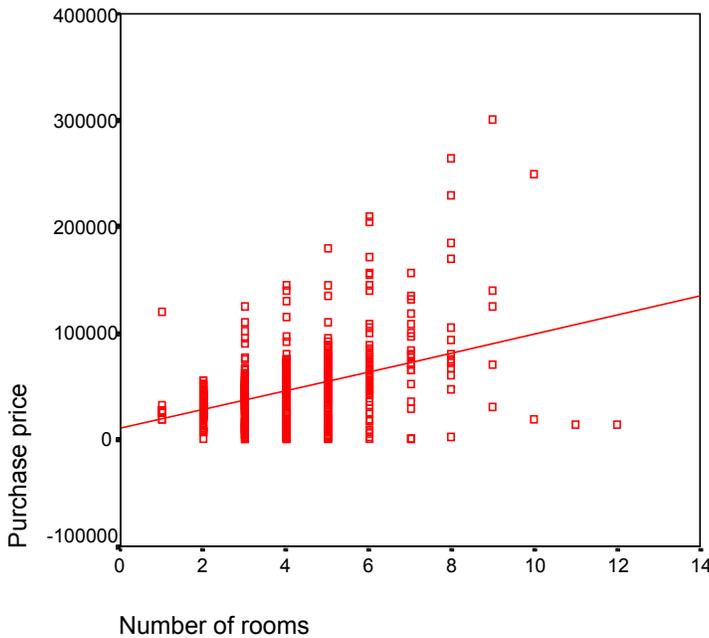
$$cov(u_1, u_2, ... u_n) = \begin{bmatrix} var(u_1) & cov(u_1, u_2) & \cdots & cov(u_1, u_n) \\ cov(u_2, u_1) & var(u_2) & & cov(u_2, u_n) \\ \vdots & & \ddots & \vdots \\ cov(u_n, u_1) & cov(u_n, u_2) & \cdots & var(u_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \quad \text{where } \sigma^2 \text{ is a scalar}$$

Assumption 3 (the one we are most concerned with here) states that the variance of each of the sampling distributions should be the same, so a covariance matrix of residuals from repeated samples should have a constant value ("scalar") down the diagonal and zero's off the diagonal.

b) Homoscedastic errors have a scalar error covariance matrix:

To understand what we mean by the variance of the residual, you have to first understand **assumption 5**, that the regressors (I.e. explanatory variables) are fixed. This means that, as in an experiment, the regressors (or control variables) can be repeated. For each value of the control variable, the scientist will observe a particular effect (i.e. a particular value of the dependent variable). In repeated experiments, she can keep the values of the control variables the same, and observe the effects on the dependent variable. There will thus be a range of values of y for each controlled and repeatable value of x. If we plot observed values of y for given values of x repeated samples, then the regression line will run through the mean of each of these conditional distributions of y.


Note, however, that each time a regression is run, it is run on a particular sample, for which there may only be one value of y for a given x (as assumed in the above diagram) or many values, depending on the experiment. As such, for each sample, there will be a slightly different line of best fit and estimates of a and b (the intercept and slope coefficients) will vary from sample to sample.

The variability of b across samples is measured by the *standard error* of b, which is an estimate of the variation of b across regressions run on repeated samples. Although we don't know SE(b) for sure (unless we run *all possible* repeated samples), we can estimate it from within the current sample because the variability of the slope parameter estimate will be linked to the variability of the y-values about the hypothesised line of best fit within the current sample. In particular, it is likely that the greater the variability of y for each given value of x, the greater the variability of estimates of a and b in repeated samples and so we can work backwards from the variability of y for a given value of x in our sample to provide an estimate of the sampling variability of b.

We can apply a similar logic to the variability of the residuals across samples. Recall that the value of the Residual for each observation i is the vertical distance between the *observed* value of the dependent variable and the *predicted* value of the dependent variable (i.e. the difference between the observed value of the dependent variable and the line of best fit value). Assume in the following figure that this is a plot from a single sample, this time with multiple observations of y for each given value of x).

Each one of the residuals has a sampling distribution, each of which should have the same variance -- "*homoscedasticity*". Clearly, this is not the case within in this sample, and so is unlikely to be true across samples. Although the sampling distribution of a residual cannot be estimated precisely from within one sample (by definition, one would need to run the same regression on repeated samples) as with SE(*b*), one can get an idea of how it might vary *between* samples by looking at how it varies *within* the current sample.
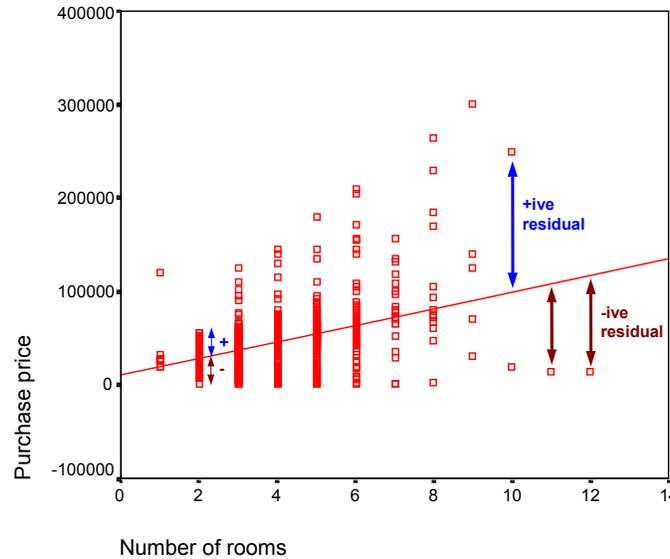
Another way to look at the residual is to plot it against one of the explanatory variables (it is particularly useful to use an explanatory variable we feel may be the cause of the heterowscedasticity). If we plot the residual against Rooms, we can see that its variance increases with the number rooms. Here we have superimposed imaginary sampling distributions of particular residuals for selected values of *x*.
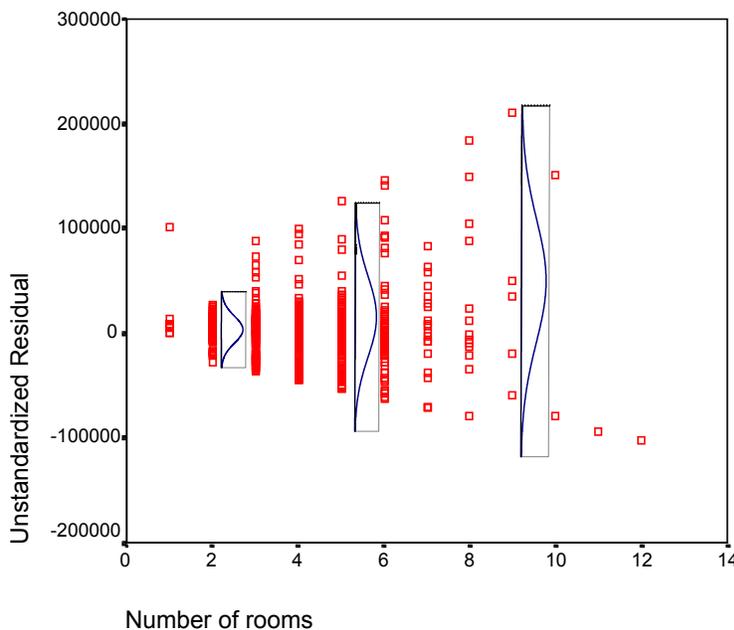
2) Causes

What might cause the variance of the residuals to change over the course of the sample? The error term may be correlated with either the dependent variable and/or the explanatory variables in the model, or some combination (linear or non-linear) of all variables in the model or those that should be in the model. But why?

a) Non-constant coefficient

Suppose that the slope coefficient varies across observations *i*:

$$y_i = a + b_i x_i + u_i$$

and suppose that it varies randomly around some fixed value β:

$$b_i = \beta + \varepsilon_i$$

then the regression actually estimated by SPSS will be:

$$y_i = a + (\beta + \varepsilon_i)\,x_i + u_i$$
$$= a + \beta\,x_i + (\varepsilon_i\,x_i + u_i)$$

where $(\varepsilon_i x + u_i)$ is the error term in the SPSS regression. The error term will thus vary with x.

b) Omitted variables

Suppose the "true" model of y is:
$$y_i = a + b\,x_i + c\,z_i + u_i$$
but the model we estimate fails to include z:
$$y_i = a + b\,x_i + v_i$$
then the error term in the model estimated by SPSS (v_i) will be capturing the effect of the omitted variable, and so it will be correlated with z:
$$v_i = c\,z_i + u_i$$
and so the variance of v_i will be non-scalar.

c) Non-linearities

If the true relationship is non-linear:
$$y_i = a + b\,x_i^2 + u_i$$
but the regression we attempt to estimate is linear:
$$y_i = a + b\,x_i + v_i$$
then the residual in this estimated regression will capture the non-linearity and its variance will be affected accordingly:
$$\boldsymbol{v_i = f(x_i^2, u_i)}$$

d) Aggregation

Sometimes we aggregate our data across groups. For example, we might use quarterly time series data on income which is calculated as the average income of a group of households in a given quarter. If this is so, and the size of groups used to calculate the averages varies, then the variation of the mean not be constant (larger groups will have a smaller standard error of the mean). This means that the measurement errors of each value of our variable will be correlated with the sample size of the groups used.

Since measurement errors will be captured by the regression residual, the implication is that the regression residual will vary the sample size of the underlying groups on which the data is based.

3) Consequences

Heteroscedasticity by itself does not cause OLS estimators to be biased or inconsistent (for the difference between these two concepts see the graphs below) since neither bias nor consistency are determined by the covariance matrix of the error term. However, if heteroscedasticity is a symptom of omitted variables, measurement errors, or non-constant parameters, then OLS estimators will be biased and inconsistent. Note that in such cases, heteroscedasticity does not causes the bias: it is merely one of the side effects of a failure of one of the other assumptions that also causes bias and inconsistency.

Asymptotic Distribution of OLS Estimate β^{hat}
The Estimate is Unbiased and Consistent since as the sample size increases, the mean of the
distribution tends towards the population value of the slope coefficient β
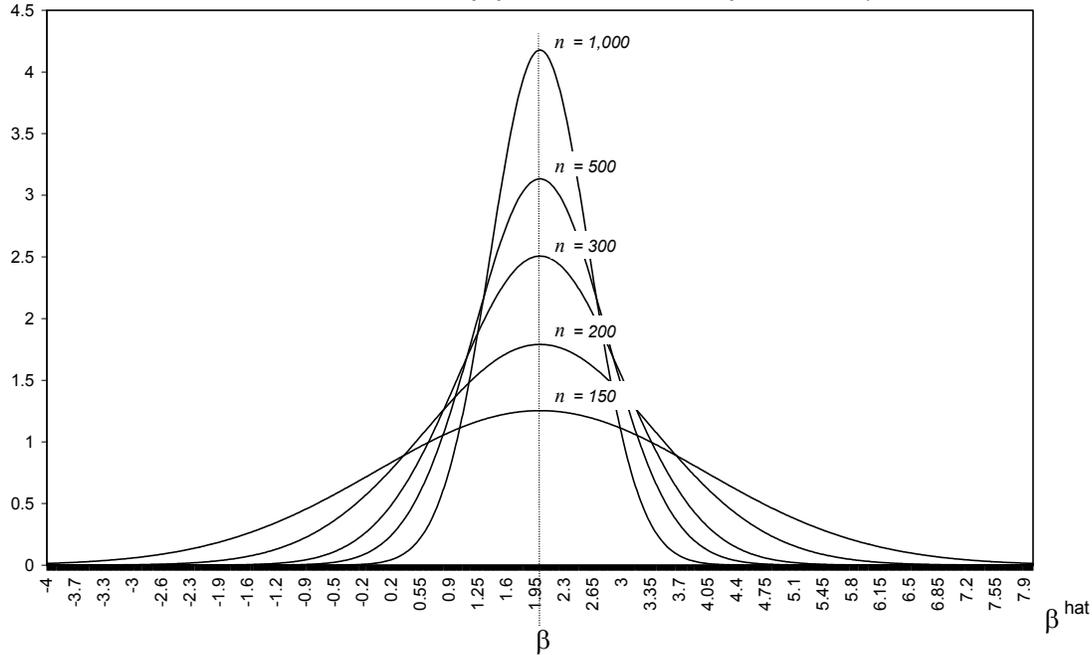
Asymptotic Distribution of OLS Estimate β^{hat}
The Estimate is Biased but Consistent since as the sample size increases, the mean of the
distribution tends towards the population value of the slope coefficient β
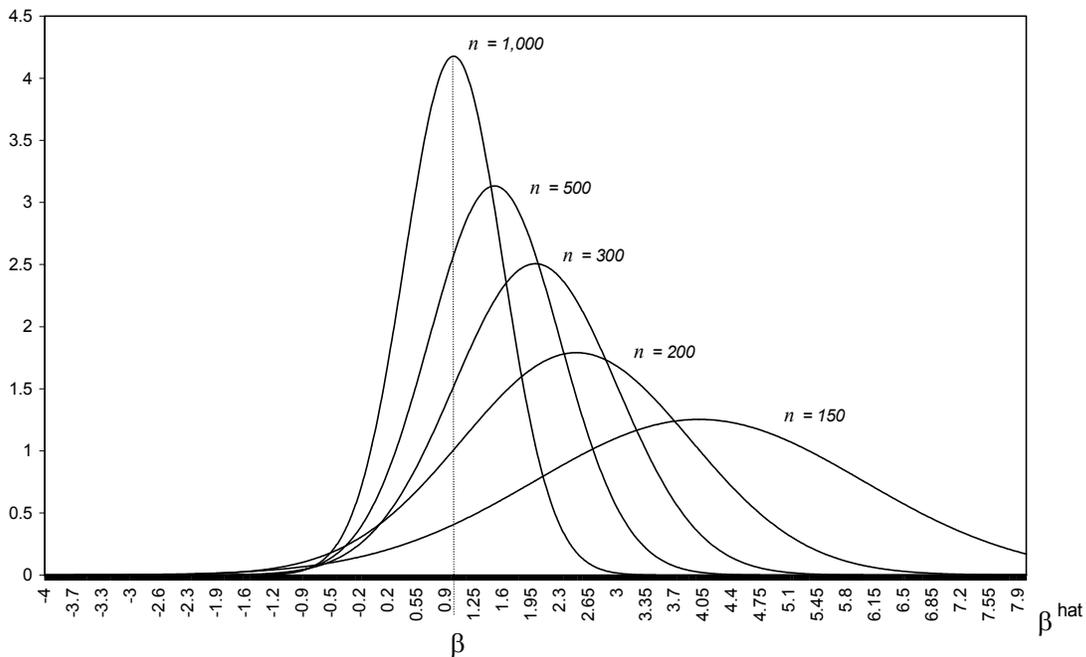
So testing for heteroscedasticity is closely related to tests for misspecification generally and many of the tests for heteroscedasticity end up being general mispecification tests. Unfortunately, there is no straightforward way to identify the cause of heteroscedasticity.

Whilst not biasing the slope estimates, heteroscedasticity **does**, however, bias the OLS estimated standard errors of those slope estimates, SE(b^{hat}), which means that the *t* tests will not be reliable (since $t = b^{hat} / \text{SE}(b^{hat})$). F-tests are also no longer reliable. In particular, it has been found that Chow's first Test no longer reliable (Thursby).

4) Detection: Specific Tests/Methods

a) *Visual Examination of Residuals*
A number of residual plots are worth examining and are easily accessible in SPSS. These are:
- **histogram of residuals** – you would like normal a normal distribution (note that a non-normal distribution is not necessarily problematic since only inference is effected, but non-normality can be a symptom of misspecification).
- **normal probability plot of residuals** – another way of visually testing for normality (normally distributed errors will lie in a straight line along the diagonal – non-linearities not captured by the model and other misppecifications may cause the residuals to deviate from this line).

- **Scatter plot of the standardised residuals** on the standardised predicted values (ZRESID as the Y variable, and ZPRED as the X variable – this plot will allow you to detect outliers and non-linearities since "well behaved" residuals will be spherical i.e. scattered randomly in an approximate circular pattern). If the plot fans out in (or fan in) a funnel shape, this is a sign of heteroscedasticity. If the residuals follow a curved pattern, then this is a sign that non-linearities have not been accounted for in the model.
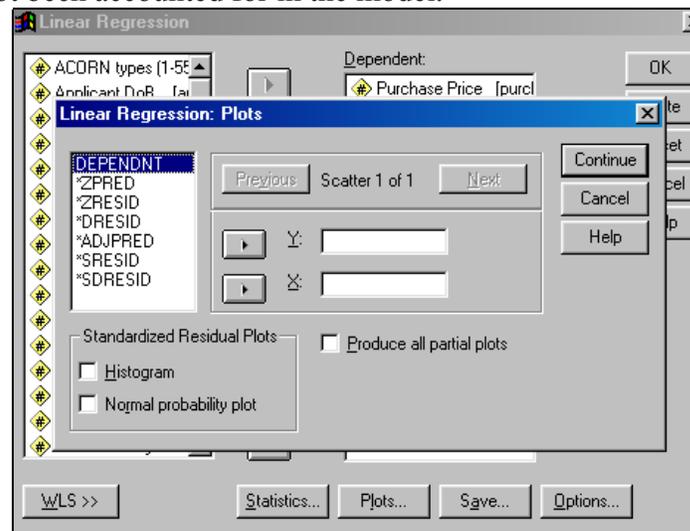
These can all be included as part of the regression output by clicking on "Plots" in the Linear Regression Window, check the "Histogram" and "Normal Probability Plot" boxes, and select the ZRESID on ZPRED scatter plot. Alternatively, you can add

```
/SCATTERPLOT=(*ZRESID ,*ZPRED )
 /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

to the end of your regression syntax before the full stop.

Example of Visual Plots:

A regression of house price on floor area produces the following plots:

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT purchase
  /METHOD=ENTER floorare
  /SCATTERPLOT=(*ZRESID ,*ZPRED )
  /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

Histogram

Dependent Variable: Purchase Price
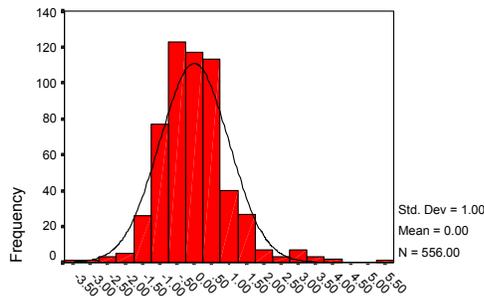
Std. Dev = 1.00
Mean = 0.00
N = 556.00

Regression Standardized Residual

Normal P-P Plot of Regression Stand

Dependent Variable: Purchase Price

Observed Cum Prob

Scatterplot

Dependent Variable: Purchase Price

Regression Standardized Predicted Value

The residuals are pretty much normally distributed but there is evidence of heteroscedasticity since the residual plot "fans out". If we re-run the regression using the log of purchase price as the dependent variable, we find that the residuals become spherical again (one should check whether taking logs has a detrimental effect on other diagnostics such as the Adjusted R^2 and t-values – in this case the impact is negligible):

```
COMPUTE price_l = ln(purchase).
EXECUTE.

REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT price_l
 /METHOD=ENTER floorare
 /SCATTERPLOT=(*ZRESID ,*ZPRED )
 /RESIDUALS HIST(ZRESID) NORM(ZRESID) .
```

Scatterplot
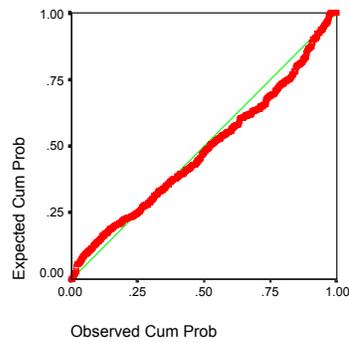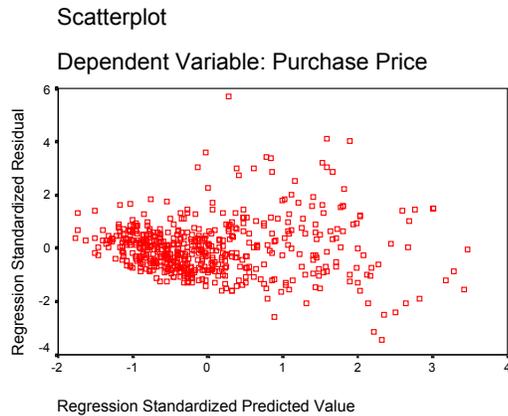
Dependent Variable: PRICE_L

Regression Standardized Predicted Value
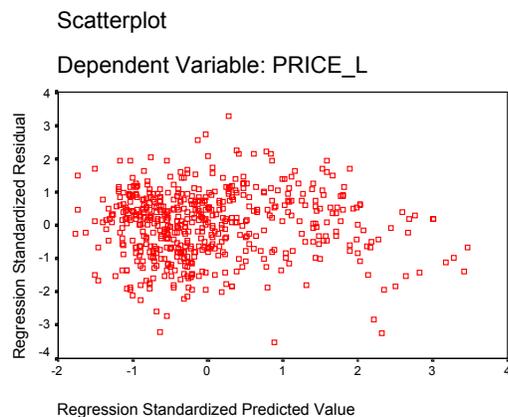
b) Levene's Test

We came across the Levene's test in Module I when we tested for the equality of means between two populations. You may recall out that there are two t-test statistics, one for the case of homogenous variances and one for the case of heterogeneous variances. In order to decide which t-test statistic to use, we used the Levene's test for equality of variances. We can apply this here:

Step 1:   save the residuals from your regression
Step 2:   Decide on which variable might be the cause of the heteroscedasticity.
Step 3:   run a Levene's test across two segments of your sample, using the variable you believe to be the cause of the heteroscedasticity as the grouping variable.

> To do the Levene's test:
> - go to Analyse, Compare Means, Independent Samples T-Test, select the residual you have created as the Test Variable.
> - Then select the variable you believe to be the cause of heteroscedasticity as the grouping variable (e.g. age of dwelling) – note that you may want to miss out observations in the middle range of your grouping variable (e.g. those in the middle two quartiles) in order to capture variation in the residual across the extremes of your grouping variable.
> - Click on Define Groups and select a cut off point for your grouping variable (this might be the mean value for example)
> - Click Paste and run the syntax (ignore the t-test portion on the right hand side of the output table – just focus on the Levene's test results).

Example of using the Levene's Test:

Use the Levene's test to test for heteroscedasticity caused by age of dwelling in a regression of floor area on age of dwelling, rooms, bedrooms. Also test for heteroscedasticity caused by floor area (e.g. variance of the residuals increases with floor area).

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT floorare
 /METHOD=ENTER age_dwel  bedrooms bathroom
 /save resid(res_1).

T-TEST
 GROUPS=age_dwel(62.5)
 /MISSING=ANALYSIS
```

```
 /VARIABLES=res_1
/CRITERIA=CIN(.95) .
```

Group Statistics

	Age of the dwelling in years	N	Mean	Std. Deviation	Std. Error Mean
Unstandardized Residual	>= 62.500000000	252	-.8801041	23.65955	1.4904118
	< 62.500000000	304	.7295600	24.46173	1.4029765

Indepe

		Levene's Test for Equality of Variances	
		F	Sig.
Unstandardized Residual	Equal variances assumed	.110	.740
	Equal variances not assumed		

H_0: equal variances Age dwelling < 62.5 and age dwelling > 62.5. Since the significance level is so high, we cannot reject the null of equal variances. In other words, the Levene's test is telling us that the variance of the residual term does not vary by age of dwelling. This seems surprising given the residual plots we did earlier, but the standard deviations of the residual across the two groups reported in the Group Statistics table seems to confirm this (i.e. the standard deviations are very similar).

However, it may be that it is only at the extremes of age that the heteroscedasticity occurs. We should try running the Levene's test on the first and last quartile (i.e. group age of dwelling as below the 25 percentile and above the 75 percentile). You can find out percentiles by going to Analyse, Custom Tables, Basic Tables, enter Age of dwelling into the Summary, click statistics and select the relevant percentiles from the list available. This gives you the following syntax and output:

```
* Basic Tables.
TABLES
  /FORMAT BLANK MISSING('.')
  /OBSERVATION age_dwel
  /TABLES age_dwel
BY (STATISTICS)
  /STATISTICS
  mean( )
  ptile 25(  'Percentile 25')
  ptile 75(  'Percentile 75')
  median( ).
```

	Mean	Percentile 25	Percentile 75	Median
Age of the dwelling in years	62.4586331	21.000000	99.000000	49.0000000

Now run the Levene's test again, but this time screen out the middle two quartiles from the sample using the "TEMPORARY. SELECT IF age_dwel **le 21 or** age_dwel **ge 99**" syntax before the T-TEST syntax.

"le" means less than or equal to, and "ge" means greater than or equal to. Note that you must run the "TEMPORARY. SELECT IF…" and the "T-TEST…." syntax all in one go (i.e. block off all seven lines and run):

```
TEMPORARY.
SELECT IF age_dwel  le 21 or age_dwel ge 99.
T-TEST
 GROUPS=age_dwel(62.5)
 /MISSING=ANALYSIS
 /VARIABLES=res_1
             /CRITERIA=CIN(.95) .
```

Now there is more evidence of heteroscedasticity (compare the standard deviations) but the difference is still not statistically significant difference according to the Levene's test (sig. = 0.375 so if we reject the null of homoscedasticity there is nearly a 40% chance that we will have done so incorrectly):

Group Statistics

	Age of the dwelling in years	N	Mean	Std. Deviation	Std. Error Mean
Unstandardized Residual	>= 62.500000000	168	.1709786	25.22774	1.9463624
	< 62.500000000	141	1.4224648	28.13163	2.3691107

Independ

		Levene's Test for Equality of Variances	
		F	Sig.
Unstandardized Residual	Equal variances assumed	.789	.375
	Equal variances not assumed		

c) Goldfeld-Quandt Test:

Goldfeld and Quandt (1965) suggested the following test procedure for null and alternative hypotheses of the form:

H$_0$: σ_i^2 is not correlated with a variable z
H$_1$: σ_i^2 is correlated with a variable z

(i) order the observations in ascending order of x.

(ii) omit p central observations (as a rough guide take $p \approx n/3$ where n is the total sample size). This enables us to easily identify the differences in variances.

(iii) Fit the separate regression to both sets of observations. The number of observations in each sample would be $(n - p)/2$, so we need $(n - p)/2 > k$ where k is the number of explanatory variables.

(iv) Calculate the test statistic G where:
$$G = \frac{RSS_2/\left(\tfrac{1}{2}(n - p) - k\right)}{RSS_1/\left(\tfrac{1}{2}(n - p) - k\right)}$$
Where G has an F distribution:
$$G \sim F\left[\tfrac{1}{2}(n - p) - k, \ \tfrac{1}{2}(n - p) - k\right]$$

NB G must be > 1, if not, invert it.

Problems with the G-Q test:

In practice we don't usually know what z is. If there are various possible z's then it may not matter which one you choose if they are all highly correlated which each other.

Given that the G-Q test is very similar to the Levene's test considered above, we shall not spend any time on it here.

5) Detection: General Tests

a) Breusch-Pagan Test :

Assumes that:
$$\sigma_I^2 = a_1 + a_2 z_1 + a_3 z_3 + a4 z_4 \ldots a_m z_m \quad [1]$$
where z's are all independent variables. Z's can be some or all of the original regressors or some other variables or some transformation of the original regressors which you think cause the heteroscedasticity:
$$\text{e.g. } \sigma_I^2 = a_1 + a_2 \exp(x_1) + a_3 x_3^2 + a_4 x_4$$

Procedure for B-P test:

Step 0: Test for non-normality in the errors. If they are normal, procede. If not, see Koenker (1981) version below.

Step 1: Obtain OLS residuals u_I^{hat} from the original regression equation and construct a new variable g:

$$g_I = u^{hat\,2} / \sigma_I^{hat\,2}$$
$$\text{where } \sigma_I^{hat\,2} = RSS / n$$

Step 2:   Regress g_I on the z's (include a constant in the regression)

Step3:   Calculate B where,
$B = \frac{1}{2}(REGSS)$ from the regression of g_I on the z's,
and where B has a Chi-square distribution with m-1 degrees of freedom
where m is the number of z's.

Problems with B-P test:
B-P test is not reliable if the errors are not normally distributed and if the sample size is small
Koenker (1981) offers an alternative calculation of the statistic which is less sensitive to non-
normality in small samples:
$$B^{Koenker} = nR^2 \sim \chi^2_{m-1}$$
where n and R^2 are from the regression of $u^{hat\,2}$ on the z's, where $B^{Koenker}$ has a Chi-square
distribution with m-1 degrees of freedom.

Example of applying the B-P test:
     Use the B-P test to test for heteroscedasticity in a regression of floor area on age of dwelling,
     rooms, bedrooms.

Step 0:   Test for non-normality in the errors. If they are normal, proceed. If not, see
     Koenker (1981) version below.

          We can test for normality by looking at the histogram and normal
          probability plots of the residuals, but we can also use the **skew** and
          kurtosis measures available in descriptive statistics.
          - Go to Analysis, Descriptive Statistics, Descriptives, and select the
            appropriate standardised residual variable you are interested in.
          - Then click on options and tick kurtosis and skewness.
          - Alternatively you can add KURTOSIS SKEWNESS to your
            Descriptives syntax – see example below.

          Kurtosis is a measure of the extent to which observations cluster
          around a central point. For a *normal distribution*, the value of the
          kurtosis statistic is *zero*. Positive kurtosis indicates that the
          observations cluster more and have longer tails than those in the
          normal distribution. Negative kurtosis indicates the observations
          cluster less and have shorter tails.

          Skewness is a measure of the asymmetry of a distribution. The
          normal distribution is symmetric, and has a skewness *value of zero*. A
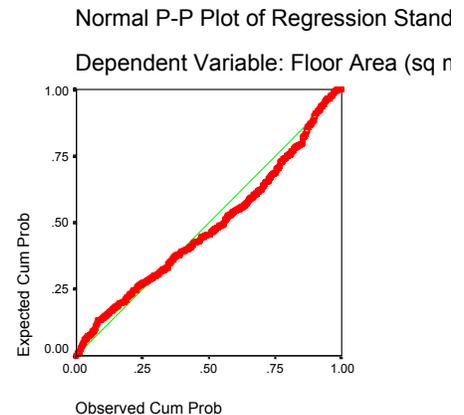
distribution with a significiant postive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a rough guide, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT floorare
  /METHOD=ENTER age_dwel  bedrooms bathroom
  /RESIDUALS HIST(ZRESID) NORM(ZRESID)
  /save resid(res_4).

DESCRIPTIVES
  VARIABLES=res_4
 /STATISTICS=MEAN KURTOSIS SKEWNESS .
```

Histogram

Dependent Variable: Floor Area (sq meters)

Std. Dev = 1.00
Mean = 0.00
N = 556.00

Regression Standardized Residual

Normal P-P Plot of Regression Stand

Dependent Variable: Floor Area (sq m

Observed Cum Prob
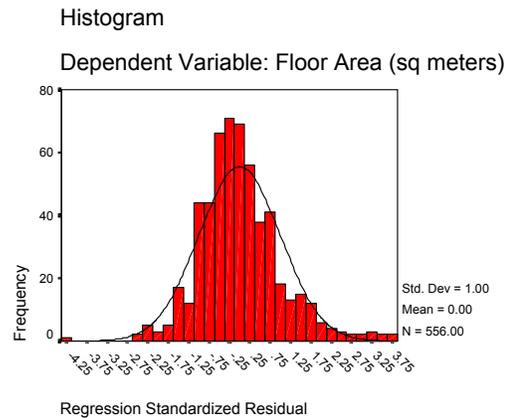
Descriptive Statistics

	N	Mean	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
RES_4	556	1.61E-15	.621	.104	1.823	.207
Valid N (listwise)	556					

The histogram and normal probability plot suggest that the errors are fairly normal. The positive value of the skewness statistic suggests that it is skewed to the left (long right tail) and since this is more than twice its standard error this suggests a degree of non-normality. The positive Kurtosis suggests that the distribution is more clustered than the normal distribution. I would say this was a borderline case so I shall present both the B-P statistic and the Koenker version. It is worth noting that the Koenker version is probably more reliable anyway so there is a case for dropping the B-P version entirely (the only reason to continue with it is because more people are familiar with it).

Step 1:   Square the residuals, and calculate RSS/n. Then calculate:
$g = (res\_4sq)/(RSS/n)$:

```
COMPUTE res_4sq = res_4 * res_4.
VARIABLE LABELS res_4sq "Square of saved residuals res_4".
EXECUTE.

DESCRIPTIVES
 VARIABLES=res_4sq
 /STATISTICS= sum  .
```

Descriptive Statistics

	N	Sum
RES_4SQ	556	322168.4
Valid N (listwise)	556	

Note that the sum of squared residuals = RSS = the figure reported in the ANOVA table, so you might want to check it against your ANOVA table to make sure you've calculated the squared residuals correctly.

```
COMPUTE g = (res_4sq)/(322168.419920 / 556).
EXECUTE.
```

Step 2:   Regress g_i on the z's (include a constant in the regression):

First you need to decide on what the *"z's"* are going to be. Lets say we used the original variables raised to the power of 1, 2, 3, and 4:

```
COMPUTE agedw_sq = age_dw * age_dw.
EXECUTE.
COMPUTE  agedw_cu = age_dw * age_dw * age_dw.
EXECUTE.
COMPUTE agedw_4 = agedw_cu * age_dw.
EXECUTE.

COMPUTE bedrm_sq = bedrooms * bedrooms.
EXECUTE.
COMPUTE  bedrm_cu = bedrooms * bedrooms * bedrooms.
EXECUTE.
COMPUTE bedrm_4 = bedrm_cu * bedrooms.
EXECUTE.

COMPUTE bath_sq = bathroom * bathroom.
EXECUTE.
COMPUTE  bath_cu = bathroom * bathroom * bathroom.
EXECUTE.
COMPUTE bath_4 = bath_cu * bathroom.
EXECUTE.


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT g
  /METHOD=ENTER age_dwel  bedrooms bathroom
agedw_sq agedw_cu agedw_4
bedrm_sq bedrm_cu  bedrm_4
bath_sq bath_cu bath_4.
```

The ANOVA table from this regression will give you the explained (or "regression") of squares REGSS = 218.293:

ANOVA^b

Model		Sum of Squares	df	M
1	Regression	218.293	9	
	Residual	1892.280	546	
	Total	2110.573	555	

a. Predictors: (Constant), BATH_4, AGEDW_SQ
   BATHROOM, AGE_DWEL, BEDRM_SQ, AGE

b. Dependent Variable: G

Step3:   Calculate $B = \frac{1}{2}(\text{REGSS}) \sim \chi^2_{m-1}$ from the regression of g_I on the z's,

$$B = \frac{1}{2}(\text{REGSS}) = 0.5(218.293) = 109.1465 \sim \chi^2_{m-1}$$

Since 3 of the z's were automatically dropped out of the regression because they were perfectly correlated, the actual number entered was $9 = m$ (see first row of *df* in the ANOVA table from the regression on the z's). So the degrees of freedom for the Chi square test $= m - 1 = 8$.

You could use Chi-square tables which will give you the Chi square value for a particular significance level and *df*. In this case, for *df* = 8, and a sig. level of 0.05, $\chi^2 = 2.73$. Since our test statistic value of 109.1465 for χ^2 is way beyond this we can confidently reject the null of homoscedasticity (i.e. we have a problem with heteroscedasticity).

Alternatively you could calculate the significance level using SPSS syntax: CDF.CHISQ(quant, df) which returns the probability that Chi-square < quant:

    COMPUTE B_PChisq = 1 - CDF.CHISQ(109.1465, 8) .
    EXECUTE .

So our test statistic $= \chi^2_8 = 109.1465$ (sig. = 0.0000)

Calculate $B^{\text{Koenker}} = nR^2 \sim \chi^2_{m-1}$

Turning now to the Koenker version, we simply multiply the sample size by the R^2 (NB **not** the adjusted R^2) from the regression of g on z:

$$B^{\text{Koenker}} = nR^2 = 0.103 * 556 = 57.268.$$

    COMPUTE BPKChisq = 1 - CDF.CHISQ(57.268, 8) .
    EXECUTE .

This has a sig. value of 1.6E-9 $\approx$ 0. So both tests reject the null
hypothesis of homoscedasticity.

---

b) White Test

The most general test of heteroscedasticity
no specification of the form of hetero required
Procedure for White's test:

Step 1:   run an OLS regression – use the OLS regression to calculate $u^{hat\,2}$ (i.e.
square of residual).

Step 2:   use $u^{hat\,2}$ as the dependent variable in another regression, in which the
regressors are: (a) all "k" original independent variables, and (b) the square
of each independent variable, (excluding dummy variables), and all 2-way
interactions (or crossproducts) between the independent variables.

The square of a dummy variable is excluded because it will be perfectly
correlated with the dummy variable.
Call the total number of regressors (not including the constant term) in this
second equation, P.

Step 3:   From results of equation 2, calculate the test statistic:
$$nR^2 \quad \sim \quad \chi^2_P$$
where n = sample size, and R^2 = unadjusted coefficient of determination.

The statistic is asymptotically (I.e. in large samples) distributed as chi-squared with P degrees of
freedom, where P is the number of regressors in the regression, not including the constant.

Notes on White's test:
- The White test does not make any assumptions about the particular form of heteroskedasticity,
  and so is quite general in application.
- It does not require that the error terms be normally distributed.
- However, rejecting the null may be an indication of model specification error, as well as or
  instead of heteroskedasticity.
- Generality is both a virtue and a shortcoming. It might reveal heteroscedasticity, but it might
  also simply be rejected as a result of missing variables.
- It is "nonconstructive" in the sense that its rejection does not provide any clear indication of
  how to proceed.
- However, if you use White's standard errors, eradicating the heteroscedasticity is less
  important.

Problems:

- Note that although *t*-tests become reliable when you use White's standard errors, *F*-tests are still not reliable (in particular, Chow's first test is still not reliable).
- White's SEs have been found to be unreliable in small samples but revised methods for small samples have been developed to allow robust SEs to be calculated for small *n*.

Example:
Run a regression of the log of floor area on terrace semidet garage1 age_dwel bathroom bedrooms and use the White Test to investigate the existence of heteroscedasticity.

One could calculate this test manually.The only problem is that it can be quite time consuming constructing all the cross products.

* 1st step: Open up your data file.
* 2nd step: Run you OLS regression and save UNSTANDARDISED residuals as RES_1:.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT flarea_l
  /METHOD=ENTER terrace semidet garage1 age_dwel  bathroom
bedrooms
  /SAVE RESID(RES_1) .
```

* 3rd step: create a variable called ESQ = square of those residuals:.

```
COMPUTE ESQ = RES_1 * RES_1.
EXECUTE.
```

* 4th step: create cross products.
* First use the "KEEP" command to save a file with only the relevant variables in it.
```
SAVE OUTFILE= 'C:\TEMP\WHI_TEST.SAV'
/KEEP= ESQ  terrace semidet garage1 age_dwel  bathroom bedrooms .
GET FILE = 'C:\TEMP\WHI_TEST.SAV'.
```

```
* given n variables, there are (n-1)*n/2 crossproducts.
* When n=6, there are (6-1)*6/2 = 15 cross products, hence we
need cp1 to cp15 to hold
     the cross products.
* The only things to alter below are the cp(?F8.0) figure in
the first line (? = total number of cross products),
and the numbers following "TO" in lines three (=? -1) and
four (=?):.
```

```
VECTOR v=terrace TO bedrooms /cp(15F8.0).
COMPUTE #idx=1.
LOOP #cnt1=1 TO 14.
LOOP #cnt2=#cnt1 +1 TO 15.
COMPUTE cp(#idx)=v(#cnt1)*v(#cnt2).
COMPUTE #idx=#idx+1.
END LOOP.
END LOOP.
EXECUTE.
```
*This step is based on part of a routine written by Raynald Levesque (2002) to calculate all combinations of crossproducts, http://pages.infinit.net/rlevesqu/.

* 5th step: run a regression on the original explanatory variables plus all cross products.
*Note that SPSS will automatically drop out variables if that are perfectly correlated with variables already in the regression.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /NOORIGIN
  /DEPENDENT esq
  /METHOD=ENTER age_dwel bathroom bedrooms cp1 cp2 cp3
/ SAVE RESID(RES_2) .
```

* 6th Step: calculate the test statistic as nRsquare ~ Chi-square with degrees of freedom equal to P = the total number of regressors actually run in this last regression (i.e. not screened out because of perfect colinearity), not including the constant term. You can do this by hand or run the following syntax which will also calculate the significance
level of Chi-square test statistic (the only thing you will need to do is enter the value for P in the first line of MATRIX syntax).


```
MATRIX.
COMPUTE P = 6.
GET ESQ / VARIABLES = ESQ.
GET RES_2 / VARIABLES = RES_2.
COMPUTE RES2_SQ = RES_2 &**2.
COMPUTE N = NROW(ESQ).
COMPUTE RSS = MSUM(RES2_SQ).
COMPUTE ii_1 = MAKE(N, N, 1).
COMPUTE I = IDENT(N).
COMPUTE M0 = I - ((1/N) * ii_1).
COMPUTE TSS = TRANSPOS(ESQ)*M0*ESQ .
```

```
PRINT RSS
  / FORMAT = "E13".
PRINT TSS
  / FORMAT = "E13".
COMPUTE R_SQ = 1-(RSS / TSS).
PRINT R_SQ
  / FORMAT = "E13".
PRINT N
  / FORMAT = "E13".
PRINT P
  / FORMAT = "E13".
COMPUTE WH_TEST = N * (1-(RSS / TSS)).
PRINT WH_TEST
  / FORMAT = "E13"
  / TITLE = "White's General Test for Heterosced (CHI-SQUARE df = P)".
COMPUTE SIG = 1 - CHICDF(WH_TEST,P).
PRINT SIG
  / FORMAT = "E13"
  / TITLE = "SIGNIFICANCE LEVEL OF CHI-SQUARE df = P (H0 =
homoscedasticity)".
END MATRIX.
```

The output from this syntax is as follows:

```
RSS  2.385128E+00
TSS  2.487222E+00
R_SQ  4.104736E-02
N  5.560000E+02
White's General Test for Heterosced (CHI-SQUARE df = P)
  2.282233E+01
SIGNIFICANCE LEVEL OF CHI-SQUARE df = P (H0 = homoscedasticity)
  8.582205E-04
```
So we reject the null (i.e. we have a problem with heteroscedasticity)

___

3. Solutions
a) Weighted Least Squares
If the differences in variability of the error term can be predicted from another variable within the model, the Weight Estimation procedure (available in SPSS) can be used. The procedure computes the coefficients of a linear regression model using weighted least squares (WLS), such that the more precise observations (that is, those with less variability) are given greater weight in determining the regression coefficients. The Weight Estimation procedure tests a range of weight transformations and indicates which will give the best fit to the data.

Problems:
- Wrong choice of weights can produce biased estimates of the standard errors.
- We can never know for sure whether we have chosen the correct weights, this is a real problem.
- If the weights are correlated with the disturbance term, then the WLS slope estimates will be inconsistent.
- Other problems have been highlighted with WLS (e.g. Dickens (1990) found that errors in grouped data may be correlated within groups so that weighting by the square root of the group size may be inappropriate. See Binkley (1992) for an assessment of tests of grouped heteroscedasticity).
- In small sample sizes, tests for heteroscedasticity can fail to detect its presence (i.e. the tests tent to increase in power as sample size increases – see Long and Ervin 1999) and so it has been argued that in small samples corrected standard errors (see below) should be used.

b) ML Estimation (not covered)
The heteroscedasticity can actually be incorporated into the framework of the model if we use a more general estimation technique. However, this is an advanced topic and beyond the scope of the course. Those interested can consult Greene (1990) and the further references cited there.

c) Whites Standard Errors
White (op cit) developed an algorithm for correcting the standard errors in OLS when heteroscedasticity is present. The correction procedure does not assume any particular form of heteroscedasticity and so in some ways White has "solved" the heteroscedasticity problem. The argument is summarised by Long and Ervin (1999):

> "When the form and magnitude of heteroscedasticity are known, using weights to correct for heteroscedasticity is very simply using generalized least squares. If the form of heteroscedasticity involves a small number of unknown parameters, the variance of each residual can be estimated first and these estimates can be used as weights in a second step.In many cases, however, the form of heteroscedasticity is unknown, which makes the weighting approach impractical. When heteroscedasticity is caused by an incorrect functional form, it can be corrected by making variance-stabilizing transformations of the dependent variable (see, for example, Weisberg 1980:123-124) or by transforming both sides (Carroll and Ruppert 1988:115-173). While this approach can provide an efficient and elegant solution to the problems caused by heteroscedasticity, when the results need to be interpreted in the original scale of the variables, nonparametric methods may be necessary (Duan 1983; Carroll and Ruppert 1988:136-139). As noted by Emerson and Stoto (1983: 124), "...re-expression moves us into a scale that is often less familiar." Further, if there are theoretical reasons to believe that errors are heteroscedastistic around the correct functional form, transforming the dependent variable is inappropriate. An alternative approach, which is the focus of our paper, is to use tests based on a heteroscedasticity consistent covariance matrix, hereafter HCCM. The HCCM provides a consistent estimator of the covariance matrix of the regression coe cients in the presence of heteroscedasticity of an unknown form. This is particularly useful when the interpretation of nonlinear models that reduce heteroscedasticity is di cult, a suitable variance-stabilizing transformation cannot be found, or weights cannot be estimated for use in GLS. Theoretically, the use of HCCM allows a researcher to easily avoid the adverse e ects of heteroscedasticity even when nothing is known about the form of heteroscedasticity." (Long and Ervin 1999 p. 1)

i)   *HC0: Matrix Procedure for White's Standard Errors in SPSS when the sample is > 500:*

* SPSS PROCEDURE FOR CALCULATING White's Standard Errors: Full OLS and White's SE output.

* 1st step: Open up your data file and save it under a new name since the following procedure will alter it.
* 2nd step: Run you OLS regression and save UNSTANDARDISED residuals as RES_1:.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mp_pc
  /METHOD=ENTER xp_pc   gdp_pc
  /SAVE RESID(RES_1) .
```

* 3rd step: create a variable called ESQ = square of those residuals:.

```
COMPUTE ESQ = RES_1 * RES_1.
EXECUTE.
```

* 4th step: create a variable called CONSTANT = constant of value 1 for all observations in the sample.

```
FILTER OFF.
USE ALL.
EXECUTE .
COMPUTE CONSTANT = 1.
EXECUTE.
```

* 5th step: Filter out missing values and Enter Matrix syntax mode .

```
FILTER OFF.
USE ALL.
SELECT IF(MISSING(ESQ) = 0).
EXECUTE .
```

* 6th step: Tell the matrix routine to get your variables.
   * you need to enter the names of the Y and X variables from your regression here.
            and Use matrix syntax to calculate White's standard errors for large samples:.
            *******Note that the only thing you need to do here is alter the variable names in lines 2 and 3 below so that they match those of your regression.

```
MATRIX.
GET Y / VARIABLES = mp_pc.
```

```
GET X / VARIABLES = CONSTANT, xp_pc, gdp_pc
/ NAMES = XTITLES.
GET RESIDUAL / VARIABLES = RES_1.
GET ESQ / VARIABLES = ESQ.
COMPUTE XRTITLES = TRANSPOS(XTITLES).
COMPUTE N = NROW(ESQ).
COMPUTE K = NCOL(X).
COMPUTE O = MDIAG(ESQ).
COMPUTE WHITEV = (INV(TRANSPOS(X) * X)) *TRANSPOS(X)* O *
X*INV(TRANSPOS(X) * X).
COMPUTE WDIAG = DIAG(WHITEV).
COMPUTE WHITE_SE = SQRT(WDIAG).
PRINT WHITE_SE
 / FORMAT = "E13"
 / TITLE = "White's (Large Sample) Corrected Standard Errors"
 / RNAMES = XRTITLES.
COMPUTE B = (INV(TRANSPOS(X) * X)) * (TRANSPOS(X) * Y).
PRINT B
/ FORMAT = "E13"
/TITLE = "OLS Coefficients"
/  RNAMES = XRTITLES.
COMPUTE WT_VAL = B / WHITE_SE.
PRINT WT_VAL
/ FORMAT = "E13"
/ TITLE = "t-values based on Whites (large sample) corrected SEs"
/  RNAMES = XRTITLES.
COMPUTE SIG_WT = 2*(1- TCDF(ABS(WT_VAL), N)) .
PRINT SIG_WT
/ FORMAT = "E13"
/ TITLE = "Prob(t < tc) based on Whites (large n) SEs"
/  RNAMES = XRTITLES.
COMPUTE SIGMASQ = (TRANSPOS(RESIDUAL)*RESIDUAL)/(N-K).
COMPUTE SE_SQ = SIGMASQ*INV(TRANSPOS(X)*X).
COMPUTE SESQ_ABS = ABS(SE_SQ).
COMPUTE SE = SQRT(DIAG(SESQ_ABS)).
PRINT SE
 / FORMAT = "E13"
 / TITLE = "OLS Standard Errors"
 / RNAMES = XRTITLES.
COMPUTE OLST_VAL = B / SE.
PRINT OLST_VAL
/ FORMAT = "E13"
/ TITLE = "OLS t-values"
/  RNAMES = XRTITLES.
COMPUTE SIG_OLST = 2*(1- TCDF(ABS(OLST_VAL), N)) .
PRINT SIG_OLST
```

```
/ FORMAT = "E13"
/ TITLE = "Prob(t < tc) based on OLS SEs"
/  RNAMES = XRTITLES.
COMPUTE WESTIM = {B, SE, WHITE_SE, WT_VAL, SIG_WT}.
PRINT WESTIM
/ FORMAT = "E13"
/ RNAMES = XRTITLES
/ CLABELS = B, SE, WHITE_SE, WT_VAL, SIG_WT.
END MATRIX.
```

Notes:

❑ Don't save your data file under the same name since the above procedure has removed from the data all observations with missing values.

❑ If you already have a variable called res_1, you will need to delete or rename it before you run the syntax.   This means that if you run the procedure on several regressions, you will need to delete the newly created res_1 and ESQ variables after each run.

❑ Note that the output will use scientific notation, so 20.7 will be written as **2.07E+01**, and 0.00043 will be written as **4.3E-04**.

❑ Note that the last table just collects together the results of five of the other tables.

❑ WT_VAL" is an abbreviation for "White's t-values" and "SIG_WT" is the significance level of these t values.

Example of White's Standard Errors:
If we run the matrix syntax on our earlier regression of floor area on age of dwelling, bedrooms and bathrooms, we get:

```
Run MATRIX procedure:

White's (Large Sample) Corrected Standard Errors
CONSTANT  4.043030E-02
AGE_DWEL  1.715285E-04
BATHROOM  2.735781E-02
BEDROOMS  1.284207E-02

OLS Coefficients
CONSTANT  3.536550E+00
AGE_DWEL  1.584464E-03
BATHROOM  2.258710E-01
BEDROOMS  2.721069E-01

t-values based on Whites (large sample) corrected SEs
CONSTANT  8.747276E+01
AGE_DWEL  9.237322E+00
BATHROOM  8.256180E+00
BEDROOMS  2.118870E+01

Prob(t < tc) based on Whites (large n) SEs
CONSTANT  0.000000E+00
AGE_DWEL  0.000000E+00
BATHROOM  2.220446E-16
BEDROOMS  0.000000E+00
```

```
OLS Standard Errors
CONSTANT  3.514394E-02
AGE_DWEL  1.640008E-04
BATHROOM  2.500197E-02
BEDROOMS  1.155493E-02

OLS t-values
CONSTANT  1.006304E+02
AGE_DWEL  9.661319E+00
BATHROOM  9.034130E+00
BEDROOMS  2.354899E+01

Prob(t < tc) based on OLS SEs
CONSTANT  0.000000E+00
AGE_DWEL  0.000000E+00
BATHROOM  0.000000E+00
BEDROOMS  0.000000E+00

WESTIM
                    B           SE      WHITE_SE        WT_VAL        SIG_WT
CONSTANT  3.536550E+00  3.514394E-02  4.043030E-02  8.747276E+01  0.000000E+00
AGE_DWEL  1.584464E-03  1.640008E-04  1.715285E-04  9.237322E+00  0.000000E+00
BATHROOM  2.258710E-01  2.500197E-02  2.735781E-02  8.256180E+00  2.220446E-16
BEDROOMS  2.721069E-01  1.155493E-02  1.284207E-02  2.118870E+01  0.000000E+00
```

If we compare the adjusted t-values with those from OLS, then we will see that they are marginally lower but all still highly significant in this case. The greater the heteroscedasticity, the larger the difference between the OLS t values and WT_VAL.

ii)   HC2 and HC3: Matrix Procedure for Corrected SEs when the sample is < 500:

When the sample size is small, it has been found that White's stand ard errors are not reliable MacKinnon and White (1985) proposed three tests to be used when the sample size is small. Long and Ervin (1999) found that the third of these tests, what they call HC3, is the most reliable, but unless one has a great deal of RAM on your computer, you may run into difficulties if your sample size is greater than 250. As a result, I would recommend the following:

n < 250   use HC3 irrespective of whether your tests for heteroscedasticity prove positive (Long and Ervin found that the tests are not very powerful in small samples).

250 < n < 500   use HC2 since this is more reliable than HC0 (HC0 = White's original SE as computed above).

n > 500   use either HC2 or HC0.

Syntax for computing HC2 is presented below. Follow the first 5 steps as before, and then run the following:

```
*HC2.
MATRIX.
GET Y / VARIABLES = flarea_l.
GET X / VARIABLES = CONSTANT, age_dwel, bathroom, bedrooms
/ NAMES = XTITLES.
GET RESIDUAL / VARIABLES = RES_1.
GET ESQ / VARIABLES = ESQ.
COMPUTE XRTITLES = TRANSPOS(XTITLES).
COMPUTE N = NROW(ESQ).
COMPUTE K = NCOL(X).
COMPUTE O = MDIAG(ESQ).
/*Computing HC2*/.
COMPUTE XX = TRANSPOS(X) * X.
COMPUTE XX_1 = INV(XX).
COMPUTE X_1 = TRANSPOS(X).
COMPUTE H = X*XX_1*X_1.
COMPUTE H_MONE = h * -1.
COMPUTE ONE_H = H_MONE + 1.
COMPUTE O_HC2 = O &/ ONE_H.
COMPUTE HC2_a = XX_1 * X_1 *O_HC2.
COMPUTE HC2 = HC2_a * X*XX_1.
COMPUTE HC2DIAG = DIAG(HC2).
COMPUTE HC2_SE = SQRT(HC2DIAG).
PRINT HC2_SE
 / FORMAT = "E13"
```

```
 / TITLE = "HC2 Small Sample Corrected Standard Errors"
 / RNAMES = XRTITLES.
COMPUTE B = XX_1 * X_1 * Y.
PRINT B
/ FORMAT = "E13"
/TITLE = "OLS Coefficients"
/  RNAMES = XRTITLES.
COMPUTE HC2_TVAL = B / HC2_SE.
PRINT HC2_TVAL
/ FORMAT = "E13"
/ TITLE = "t-values based on HC2 corrected SEs"
/  RNAMES = XRTITLES.
COMPUTE SIG_HC2T = 2*(1- TCDF(ABS(HC2_TVAL), N)) .
PRINT SIG_HC2T
/ FORMAT = "E13"
/ TITLE = "Prob(t < tc) based on HC2 SEs"
/  RNAMES = XRTITLES.
END MATRIX.
```

The output from this syntax is as follows:

```
HC2 Small Sample Corrected Standard Errors
CONSTANT  4.077517E-02
AGE_DWEL  1.726199E-04
BATHROOM  2.761153E-02
BEDROOMS  1.293651E-02


OLS Coefficients
CONSTANT  3.536550E+00
AGE_DWEL  1.584464E-03
BATHROOM  2.258710E-01
BEDROOMS  2.721069E-01


t-values based on HC2 corrected SEs
CONSTANT  8.673291E+01
AGE_DWEL  9.178915E+00
BATHROOM  8.180314E+00
BEDROOMS  2.103402E+01


Prob(t < tc) based on HC2 SEs
CONSTANT  0.000000E+00
AGE_DWEL  0.000000E+00
BATHROOM  1.998401E-15
BEDROOMS  0.000000E+00
```

For **HC3**, you need to make sure that your sample is not too large otherwise the computer may crash. You can temporarily draw a random sub-sample by using the TEMPORARY. SAMPLE

p. where p is the proportion of the sample (e.g. if p = 0.5, you have selected 40% of your sample for the following operations).

```
*HC3.
/*when Computing HC3 make sure n is < 250 (e.g. use TEMPORARY.
SAMPLE 0.4.) */.
TEMPORARY.
SAMPLE 0.4.
MATRIX.
GET Y / VARIABLES = flarea_l.
GET X / VARIABLES = CONSTANT, age_dwel, bathroom, bedrooms
/ NAMES = XTITLES.
GET RESIDUAL / VARIABLES = RES_1.
GET ESQ / VARIABLES = ESQ.
COMPUTE XRTITLES = TRANSPOS(XTITLES).
COMPUTE N = NROW(ESQ).
COMPUTE K = NCOL(X).
COMPUTE O = MDIAG(ESQ).
COMPUTE XX = TRANSPOS(X) * X.
COMPUTE XX_1 = INV(XX).
COMPUTE X_1 = TRANSPOS(X).
COMPUTE H = X*XX_1*X_1.
COMPUTE H_MONE =  h * -1.
COMPUTE ONE_H = H_MONE + 1.
/*Computing HC3*/.
COMPUTE  ONE_H_SQ = ONE_H &** 2.
COMPUTE O_HC3 = O &/ ONE_H_SQ.
COMPUTE HC3_a = XX_1 * X_1 *O_HC3.
COMPUTE HC3 = HC3_a * X*XX_1.
COMPUTE HC3DIAG = DIAG(HC3).
COMPUTE HC3_SE = SQRT(HC3DIAG).
COMPUTE B = XX_1 * X_1 * Y.
PRINT B
/ FORMAT = "E13"
/TITLE = "OLS Coefficients".
PRINT HC3_SE
  / FORMAT = "E13"
  / TITLE = "HC3 Small Sample Corrected Standard Errors"
  / RNAMES = XRTITLES.
COMPUTE HC3_TVAL = B / HC3_SE.
PRINT HC3_TVAL
/ FORMAT = "E13"
/ TITLE = "t-values based on HC3 corrected SEs"
/  RNAMES = XRTITLES.
COMPUTE SIG_HC3T = 2*(1- TCDF(ABS(HC3_TVAL), N)) .
PRINT SIG_HC3T
```

```
/ FORMAT = "E13"
/ TITLE = "Prob(t < tc) based on HC3 SEs"
/  RNAMES = XRTITLES.
END MATRIX.
```

The output from the above syntax is as follows:

```
OLS Coefficients
   3.530325E+00
   1.546620E-03
   2.213146E-01
   2.745376E-01

HC3 Small Sample Corrected Standard Errors
CONSTANT  4.518059E-02
AGE_DWEL  1.884062E-04
BATHROOM  3.106637E-02
BEDROOMS  1.489705E-02

t-values based on HC3 corrected SEs
CONSTANT  7.813809E+01
AGE_DWEL  8.208966E+00
BATHROOM  7.123928E+00
BEDROOMS  1.842899E+01

Prob(t < tc) based on HC3 SEs
CONSTANT  0.000000E+00
AGE_DWEL  2.220446E-15
BATHROOM  4.005019E-12
BEDROOMS  0.000000E+00
```

4. Conclusions

In conclusion, it is worth quoting Greene (1990),

> "It is rarely possible to be certain about the nature of the heteroscedasticity in a egression model. In one respect, this is only a minor problem. The weighted least squares estimator, …, is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances… But using the wrong set of weights has two other consequences which may be less benign. First, the improperly weighted least squares estimator is inefficient. This might be a moot point if the correct weights are unknown, but the GLS standard errors will also be incorrect. The asymptotic covariance matrix of the estimator … may not resemble the usual estimator. This underscores the usefulness of the White estimator… Finally, if the form of the heteroscedasticity is known but involves unknown parameters, it remains uncertain whether FGLS corrections are better than OLS. Asymptotically, the comparison is clear, but in small or moderate-sized samples, the additional variation incorporated by the estimated variance parameters may offset the gains to GLS." (W. H. Green, 1990, p. 407)

The corollary is that one should remove any heteroscedasticity caused by misspecification by removing (where possible) the source of that misspecificaiton (e.g. correct omitted variables by including the appropriate variable). Any heteroscedasticity that remains is unlikely to be particularly harmful and one should try solutions that do not distort the regression or confuse the interpretation of coefficients (taking logs of the dependent and/or independent variables is often quite effective at reducing heteroscedasticity and usually does not have adverse affects on interpretation or specification, though you should check this). Finally, one should report White's corrected standard errors (or t-values based on them). Even if your tests for heteroscedasticity suggest that it is not present, it is probably worth presenting White's standard errors anyway, rather than the usual OLS standard errors, since the tests for heteroscedasticity are not infallible (particularly in small samples) and they may have missed an important source of of systematic variation in the error term. In small samples ($n < 250$) White's standard errors are not reliable so you should use Mackinnon and White's HC3 (this should be used even if the tests for heteroscedasticity are clear because of the reduced power of these tests in small samples).

Reading

Kennedy (1998) "A Guide to Econometrics", Chapters 5,6,7 and 9
Maddala, G.S. (1992) "Introduction to Econometrics" chapter 12
Field, A. (2000) chapter 4, particularly pages 141-162.
Greene, W. H. (1990) Econometric Analysis, 2nd edition.

Further References:

i)   Original Papers for test statistics:

S.M. Goldfeld and R.E. Quandt, "Some Tests for Homoscedasticity," *Journal of the American Statistical Society*, Vol.60, 1965.

T.S. Breusch and A.R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation," *Econometrica*, Vol. 47, 1979.

H. White. 1980. "A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity." *Econometrica*, 48, 817-838.

MacKinnon, J.G. and H. White. (1985), 'Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties'. *Journal of Econometrics*, 29, 53-57.

ii)   Grouped Heteroscedasticity:

Binkley, J.K. (1992) "Finite Sample Behaviour of Tests for Grouped Heteroskedasticity", *Review of Economics and Statistics*, 74, 563-8.

Dickens, W.T. (1990) "Error components in grouped data: is it ever worth weighting?", *Review of Economics and Statistics*, 72, 328-33.

iii) Bresch Pagan critique:

Koenker, R. (1981) "A Note on Studentizing a Test for Heteroskedascity", *Journal of Applied Econometrics*, 3, 139-43.

iii) Critique of White's Standard Errors in small samples:

Long, J. S. and Laurie H. Ervin (1999) "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model", Mimeo, Indiana University (http://www.indiana.edu/~jsl650/files/hccm/99TAS.pdf)

Useful Link for SPSS syntax and macros:

http://pages.infinit.net/rlevesqu/

© Gwilym Pryce,
University of Glasgow
14th March 2002