EE 325

Heteroscedasticity

The nature of Heteroscedasticity

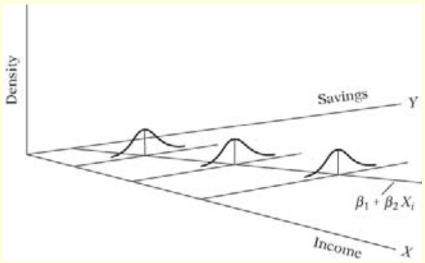
One of the important assumptions of CLRM is that the variance of each disturbance term μ_i, conditional on the chosen values of the explanatory variables, is some constant number equal to σ^2 (Homoscedasticity)

$$E(u_i^2) = \sigma^2 \qquad i = 1, \ldots, n$$

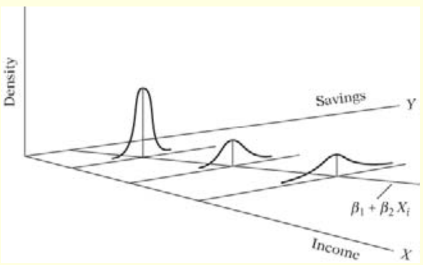
The conditional variance of Y_i increases as X increases. Here, the variances of Y_i are not the same. Here, there is heteroscedasticity.

$$E(u_i^2) = \sigma_i^2 \qquad i = 1, \ldots, n$$

Homoscedastic disturbances



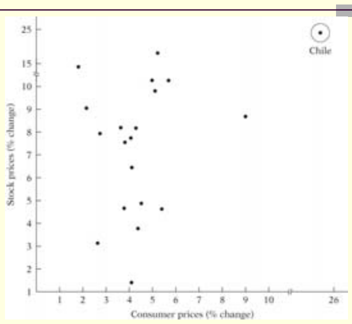
Heteroscedastic disturbances



Several reasons why the variances of u_i may be variable

- As incomes grow, people have more discretionary income and hence more scope for choice about the disposition of their income. Hence σ_i^2 is likely to increase with income.
- As data collecting techniques improve, σ_i^2 is likely to decrease

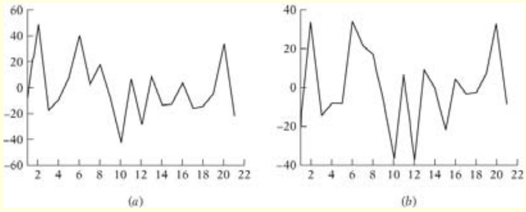
Heteroscedasticity can also arise as a result of the presence of **outliers**



EE325

Specification error

- Some important variables are omitted from the model



EE325

- Incorrect data transformation
- Incorrect functional form

EE325

OLS Estimation in the Presence of Heteroscedasticity

EE325

$\hat{\beta}_2$ *is best linear unbiased estimator (BLUE)*

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \quad \textit{Homoscedasticity}$$

EE325

Is $\hat{\beta}_2$ still BLUE when we drop only homoscedasticity assumption and replace it with the assumption of heteroscedasticity?

$\hat{\beta}_2$ is **no longer best and the minimum variance** given by

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{(\sum x_i^2)^2}$$

EE325

The Method of Generalized Least Squares (GLS)

EE325

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i \quad where \quad X_{0i} = 1 \quad for \quad each \quad i$$

Now assume that the heteroscedastic variance σ_i^2 are known

$$\frac{Y_i}{\sigma_i} = \beta_1\left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2\left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right)$$

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^*$$

EE325

$$\operatorname{var}(u_i^*) = E(u_i^*)^2 = E\left(\frac{u_i}{\sigma_i}\right)^2 \quad Since \ E(u_i^*) = 0$$

$$= \frac{1}{\sigma_i^2} E(u_i^2) \quad Since \ \sigma_i^2 \ is \ known$$

$$= \frac{1}{\sigma_i^2}(\sigma_i^2) Since \ E(u_i^2) = \sigma_i^2$$

$$= 1$$

EE325

- Since we are still retaining the other assumptions of the classical model, the finding that it is u_i^* that is homoscedastic suggests that if we apply OLS to the transformed

$$\frac{Y_i}{\sigma_i} = \beta_1\left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2\left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right)$$

it will produce estimators that are BLUE. In short, the estimated β_1^* *and* β_2^* are now BLUE and not the OLS estimators $\hat{\beta}_1$ *and* $\hat{\beta}_2$

EE325

GLS is OLS on the transformed variables that satisfy the standard least-squares assumptions

The estimators thus obtained are known as GLS estimators, and it is these estimators that are BLUE

EE325

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_1^*\left(\frac{X_{0i}}{\sigma_i}\right) + \hat{\beta}_2^*\left(\frac{X_i}{\sigma_i}\right) + \left(\frac{\hat{u}_i}{\sigma_i}\right)$$

$$Y_i^* = \hat{\beta}_1^* X_{0i}^* + \hat{\beta}_2^* X_i^* + \hat{u}_i^*$$

$$\sum \hat{u}_i^{2*} = \sum (Y_i^* - \hat{\beta}_1^* X_{0i}^* - \hat{\beta}_2^* X_i^*)^2$$

$$\sum\left(\frac{\hat{u}_i}{\sigma_i}\right)^2 = \sum\left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^*\left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_2^*\left(\frac{X_i}{\sigma_i}\right)\right]^2$$

EE325

the GLS estimator of β_2^*

$$\hat\beta_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2}$$

$$\mathrm{var}(\hat\beta_2^*) = \frac{\sum w_i}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2}$$

where $w_i = 1/\sigma^2$

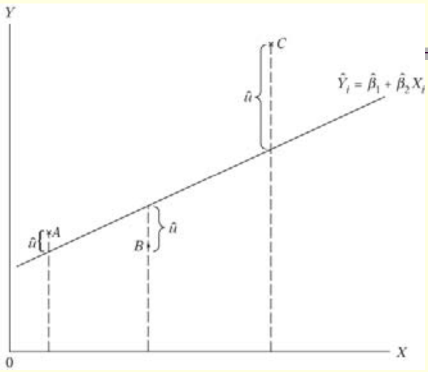
Difference between OLS and GLS

OLS we minimize

$$\sum \hat u_i^2 = \sum (Y_i - \hat\beta_1 - \hat\beta_2 X_i)^2$$

GLS we minimize

$$\sum w_i \hat u_i^2 = \sum w_i (Y_i - \hat\beta_1^* X_{0i} - \hat\beta_2^* X_i)^2, w_i = 1/\sigma_i^2$$



Example P 374-375

	Standard error of $\hat\beta_1$			Standard error of $\hat\beta_2$		
Value of α	OLS	OLS_{het}	GLS	OLS	OLS_{het}	GLS
0.5	0.164	0.134	0.110	0.285	0.277	0.243
1.0	0.142	0.101	0.048	0.246	0.247	0.173
2.0	0.116	0.074	0.0073	0.200	0.220	0.109
3.0	0.100	0.064	0.0013	0.173	0.206	0.056
4.0	0.089	0.059	0.0003	0.154	0.195	0.017

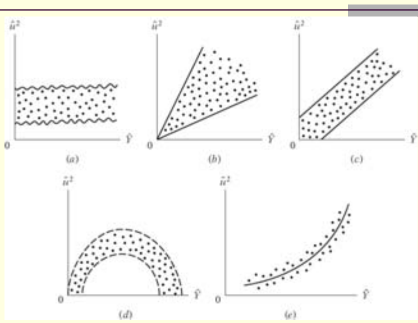
The most striking feature of these results is that OLS, with or without correction for heteroskedasticity, consistently overestimates the true standard error obtained by the (correct) GLS procedure, especially for large values of α , thus establishing the superiority of GLS

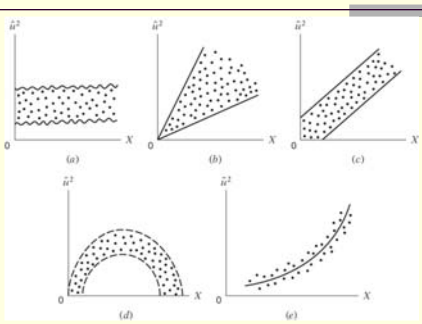
Detection of Heteroscedasticity

Detection of Heteroscedasticity

- Informal method
  - **Graphical method**
- Formal methods
  - Park Test
  - Glejser Test
  - Spearman's Rank Correlation Test
  - Goldfeld-Quandt Test
  - Breusch-Pagan-Godfrey Test
  - **White's General Heteroscedasticity Test**

Graphical method





Park Test

Park formalizes the graphical method by suggesting that σ_i^2 is some function of the explanatory variable X_i . The functional form he suggests is

$$\sigma_i^2 = \sigma^2 X_i^{\beta} e^{vi}$$

or

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$

Where v_i is the stochastic disturbance term

Since σ_i^2 is generally not known. Park suggests using $\hat{u}_i^2$ as a proxy and running the following regression:

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$
$$= \alpha + \beta \ln X_i + v_i$$

If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data

Example

Table 11.1 Relationship between compensation and productivity

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Y = average compensation in thousands of dollars
X = average productivity in thousands of dollars
i = ith employment size of the establishment

Step 1

Run the OLS regression disregarding the heteroscedasticity question

$$\hat{Y}_i = 1992.3452 + 0.2329X_i$$
$$se = (936.4791) \quad (0.0998)$$
$$t = (2.1275) \quad (2.333)$$
$$R^2 = 0.4375$$

Step 2

We obtain $\hat{u}_i$ from this regression, and then in the second stage we run the regression

$$\ln \hat{u}_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i$$
$$= \alpha + \beta \ln X_i + v_i$$

$$\widehat{\ln \hat{u}_i^2} = 35.817 - 2.8099 \ln X_i$$
$$se = (38.319) \quad (4.216)$$
$$t = (0.934) \quad (-0.667)$$
$$R^2 = 0.0595$$

Glejser Test

The Glejser test is similar to Park test. After obtaining the residuals $\hat{u}_i$ from OLS regression, Glejser suggests repressing the absolute values of $\hat{u}_i$ on the X variable that is thought to be closely associated with σ_i^2

$$|\hat{u}_i| = \beta_1 + \beta_2 X_i + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i$$
$$|\hat{u}_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$
$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$
$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

Where v_i is the stochastic disturbance term

Glejser technique may be used for large samples and may be used in the small samples strictly as a qualitative device to learn something about heteroscedasticity

Example

Table 11.1 Relationship between compensation and productivity

$$\widehat{|\hat{u}_i|} = 407.2783 - 0.0203X_i$$
$$se = (633.1621) \quad (0.0675)$$
$$t = (0.6432) \quad (-0.3012)$$
$$r^2 = 0.0127$$

$$|\hat{u}_i| = \beta_1 + \beta_2 X_i + v_i$$

$$|\hat{u}_i| = \beta_1 + \beta_2 \sqrt{X_i} + v_i$$

$$|\hat{u}_i| = \beta_1 + \beta_2 \frac{1}{X_i} + v_i$$

$$|\hat{u}_i| = \beta_1 + \beta_2 \frac{1}{\sqrt{X_i}} + v_i$$

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i} + v_i$$

$$|\hat{u}_i| = \sqrt{\beta_1 + \beta_2 X_i^2} + v_i$$

Where V_i is the stochastic disturbance term

Glejser technique may be used for large samples and may be used in the small samples strictly as a qualitative device to learn something about heteroscedasticity

Example

Table 11.1 Relationship between compensation and productivity

$$|\widehat{u_i}| = 407.2783 - 0.0203 X_i$$

$$se = (633.1621) \quad (0.0675)$$

$$t = (0.6432) \quad (-0.3012)$$

$$r^2 = 0.0127$$

Spearman's Rank Correlation Test

$$r_s = 1 - 6\left[\frac{\sum d_i^2}{n(n^2-1)}\right]$$

d_i = difference in the ranks assigned to two different characteristics of the ith individual

n = number of individuals ranked

Assume $Y_i = \beta_0 + \beta_1 X_i + u_i$

Step 1 Fit the regression to the data on Y and X and obtain the residual $\hat{u}_i$

Step 2 Ignoring the sign of $\hat{u}_i$, taking the absolute value $|\hat{u}_i|$, rank both $|\hat{u}_i|$ and X_i (or Y_i) according to an ascending or decending order and compute the Spearman's rank correlation coefficient

Step 3. Assuming that the population rank correlation coefficient ρ_s is zero and n>8, the significance of the sample r_s can be tested by the t-test as follows

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}, \text{with } df = n-2$$

If the computed t value exceeds the critical t value, we may accept the hypothesis of heteroscedasticity ;otherwise we may reject it.

Example

The data pertain to the average annual return (E,%) and the standard deviation of annual return (σ_i ,%) of 10 mutual funds

TABLE 11.2 Rank Correlation Test of Heteroscedasticity

Name of Mutual Fund	E_i, Average Annual Return, %	σ_i, Standard Deviation of Annual Return, %	$\hat{E}_i^{*\dagger}$	$\lvert\hat{u}_i\rvert^{\ddagger}$ Residuals, $\lvert(E_i - \hat{E}_i)\rvert$	Rank of $\lvert\hat{u}_i\rvert$	Rank of σ_i	d, Difference between Two Rankings	d^2
Boston Fund	12.4	12.1	11.37	1.03	9	4	5	25
Delaware Fund	14.4	21.4	15.64	1.24	10	9	1	1
Equity Fund	14.6	18.7	14.40	0.20	4	7	−3	9
Fundamental Investors	16.0	21.7	15.78	0.22	5	10	−5	25
Investors Mutual	11.3	12.5	11.56	0.26	6	5	1	1
Loomis-Sales Mutual Fund	10.0	10.4	10.59	0.59	7	2	5	25
Massachusetts Investors Trust	16.2	20.8	15.37	0.83	8	8	0	0
New England Fund	10.4	10.2	10.50	0.10	3	1	2	4
Putnam Fund of Boston	13.1	16.0	13.16	0.06	2	6	−4	16
Wellington Fund	11.3	12.0	11.33	0.03	1	3	−2	4
Total							0	110

†Obtained from the regression: $\hat{E}_i = 5.8194 + 0.4590\,\sigma_i$.
‡Absolute value of the residuals.
Note: The ranking is in ascending order of values.

The capital market line of portfolio theory postulates a linear relationship between expected return (E_i) and risk (as measured by the standard deviation, σ) of a portfolio as follows:

$$E_i = \beta_1 + \beta_2 \sigma_i$$

$$r_s = 1 - 6\left[\frac{110}{10(100-1)}\right] = 0.3333$$

$$t = \frac{0.3333\sqrt{10-2}}{\sqrt{1-0.1110}} = 0.9998,\ with\ \ df = 8$$

t value is not significance. There is no evidence of a systematic relationship between the explanatory variable and the absolute values of the residuals, which might suggest that there is no heteroscedasticity

Goldfeld-Quandt Test

Assume that the heteroscedastic variance is positive related to one of the explanatory variables in the regression model. For simplicity, consider the usual two variable model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Suppose σ_i^2 is positively related to X_i as

$$\sigma_i^2 = \sigma^2 X_i^2$$

Where σ^2 is a constant

Step 1 Order or rank the observations according to the values of X_i ,beginning with the lowest X value

Step 2 Omit c central observations, where c is specified a priori, and divide the remaining (n-c) observations into two groups each of (n-c)/2 observations

Step 3 Fit separate OLS regressions to the first (n-c)/2 observation and the last (n-c)/2 observations, and obtain the respective residual sums of squares RSS_1 and RSS_2

$RSS_1 = RSS$ *from the regression corresponding to the smaller* X_i *values*
$RSS_2 = RSS$ *from the larger* X_i *values*

These RSS each have

$$\frac{(n-c)}{2} - k \; or \; \left(\frac{n-c-2k}{2}\right) df$$

Step 4 Compute the ratio

$$\lambda = \frac{RSS_2 / df}{RSS_1 / df}$$

If we assume u_i are normally distributed, and if the assumption of homoscedasticity is valid, then it can be shown that λ follows the F-distribution with numerator and denominator df each of (n-c-2k)/2

If in an application the computed $\lambda (= F)$ is greater than the critical F at the chosen level of significance, we can reject the hypothesis of homoscedasticity, we can say that heteroscedasticity is very likely

Example

TABLE 11.3 Hypothetical Data on Consumption Expenditure Y($) and Income X($) to Illustrate the Goldfeld–Quandt Test

		Data Ranked by X Values		
Y	X	Y	X	
55	80	55	80	
65	100	70	85	
70	85	75	90	
80	110	65	100	
79	120	74	105	
84	115	80	110	
98	130	84	115	
95	140	79	120	
90	125	90	125	
75	90	98	130	
74	105	95	140	
110	160	108	145	
113	150	113	150	
125	165	110	160	Middle 4 observations
108	145	125	165	
115	180	115	180	
140	225	130	185	
120	200	135	190	
145	240	120	200	
130	185	140	205	
152	220	144	210	
144	210	152	220	
175	245	140	225	
180	260	137	230	
135	190	145	240	
140	205	175	245	
178	265	189	250	
191	270	180	260	
137	230	178	265	
189	250	191	270	

Consumption expenditure in relation to income for a cross section of 30 families. Suppose we postulate that consumption expenditure is linearly related to income but that heteroscedasticity is present in the data

Regression based on the first 13 observations

$$\hat{Y}_i = 3.4094 + 0.6968 X_i$$

$$RSS_1 = 377.17 \quad df = 11$$

Regression based on the last 13 observations

$$\hat{Y}_i = -28.0272 + 0.7941 X_i$$

$$RSS_1 = 1536.8 \quad df = 11$$

$$\lambda = \frac{1536.8/11}{377.17/11} = 4.07$$

$$4.07 > critical \; F_{5\%,11,11}$$

Breusch-Pagan-Godfrey (BPG)Test

Consider the k-variables linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

Assume that the error variance σ_i^2 is described as

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi})$$

That is σ_i^2 is some function of the nonstochastic Z variables; some or all of the X's can serve as Z's. Specifically, assume that

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi}$$

EE325

That is, σ_i^2 is a linear function of the Z's. If

$$\alpha_2 = \alpha_3 = ... = \alpha_m = 0, \quad \sigma_i^2 = \alpha_1$$

, which is a constant. Therefore, to test whether σ_i^2 is homoscedastic, one can test the hypothesis that $\alpha_2 = \alpha_3 = ... = \alpha_m = 0$. That is the basic idea behind the Breusch-Pagan-Godfrey test.

EE325

☺ Test procedure ☺

Step 1 Estimate Equation

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

by OLS and obtain the residuals $\hat{u}_1, \hat{u}_2, ..., \hat{u}_n$

Step 2 Obtain $\tilde{\sigma}^2 = \sum \hat{u}_i^2 / n$ (This is the maximum likelihood estimator of σ^2

EE325

Step 3 Construct variables p_i defined as

$$p_i = \hat{u}_i^2 / \tilde{\sigma}^2$$

which is simply each residual squared divided by $\tilde{\sigma}^2$

Step 4 Regress p_i thus constructed on the Z's as

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$

v_i is the residual term of this regression

EE325

Step 5 Obtain the ESS (Explained sum of squares) from

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + ... + \alpha_m Z_{mi} + v_i$$

and define $\Theta = \frac{1}{2}(ESS)$

Assuming u_i are normally distributed, one can show that if there is homoscedasticity and if the sample size n increases indefinitely, then

$$\Theta \underset{asy}{\sim} \chi^2_{m-1}$$

EE325

Therefore, if an application the computed Θ exceeds the critical χ^2 value at the chosen level of significance, one can reject the hypothesis of homoscedasticity; otherwise one does not reject it

EE325

Example P. 386

- Data on consumption expenditure in relation to income for a cross section of 30 families. Suppose we postulate that consumption expenditure is linearly related to income but that heteroscedasticity is present in the data

TABLE 11.3 Hypothetical Data on Consumption Expenditure Y($) and Income X($) to Illustrate the Goldfeld–Quandt Test

		Data Ranked by X Values		
Y	X	Y	X	
55	80	55	80	
65	100	70	85	
70	85	75	90	
80	110	65	100	
79	120	74	105	
84	115	80	110	
98	130	84	115	
95	140	79	120	
90	125	90	125	
75	90	98	130	
74	105	95	140	
110	160	108	145	
113	150	113	150	
125	165	110	160	Middle 4 observations
108	145	125	165	
115	180	115	180	
140	225	130	185	
120	200	135	190	
145	240	120	200	
130	185	140	205	
152	220	144	210	
144	210	152	220	
175	245	140	225	
180	260	137	230	
135	190	145	240	
140	205	175	245	
178	265	189	250	
191	270	180	260	
137	230	178	265	
189	250	191	270	


```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =  496.72
       Model |  41886.7134     1  41886.7134           Prob > F      =  0.0000
    Residual |  2361.15325    28  84.3269018           R-squared     =  0.9466
-------------+------------------------------           Adj R-squared =  0.9447
       Total |  44247.8667    29  1525.78851           Root MSE      =   9.183

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .6377846   .0286167    22.29   0.000      .579166    .6964031
       _cons |   9.290307   5.231386     1.78   0.087      -1.4257    20.00632
------------------------------------------------------------------------------
```


Step 1

$$\hat{Y}_i = 9.2903 + 0.6378 X_i$$

$$se = (5.2314)(0.0286)$$

$$RSS = 2361.153 \qquad R^2 = 0.9466$$

Step 2

$$\tilde{\sigma}^2 = \sum \hat{u}_i^2 / 30 = 2361.153 / 30 = 78.7051$$

Step 3 Divide the squared residuals $\hat{\mu}_i$ obtained from regression in step 1 by 78.7051 to construct the variable p_i

Step 4 Assuming that p_i are linearly related to $X_i (= Z_i)$, we obtained the regression

$$\hat{p}_i = -0.7426 + 0.0101 X_i$$

$$se = (0.7529) \ (0.0041)$$

$$ESS = 10.4280 \qquad R^2 = 0.18$$


```
      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  1,    28) =    5.97
       Model |  10.4280234     1  10.4280234           Prob > F      =  0.0211
    Residual |  48.9099163    28  1.74678273           R-squared     =  0.1757
-------------+------------------------------           Adj R-squared =  0.1463
       Total |  59.3379398    29  2.04613585           Root MSE      =  1.3217

------------------------------------------------------------------------------
        phat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0100632   .0041187     2.44   0.021      .0016265    .0184999
       _cons |  -.7426137   .7529274    -0.99   0.332     -2.284916    .7996882
------------------------------------------------------------------------------
```

- Step 5

$$\Theta = \frac{1}{2}(ESS) = 5.2140$$

Under the assumptions of the BPG Test asymptotically follows the chi-square distribution with 1 df. Now from the chi-square table we find that for 1 df the 5 percent critical χ^2 is 3.8414 and the 1 percent critical χ^2 is 6.6349. Thus, the observed chi-square value of 5.2140 is significant at the 5 percent but not the 1 percent level of significance.

The BPG test is **an asymptotic, or large sample, test** and in the present example 30 observations may not constitute a large sample. It should also be pointed out that in small samples the test is sensitive to the assumption that the disturbances u_i are normally distributed.

White's General Heteroscedasticity Test

Consider the following three-variable regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

The White test proceeds as follows:

Step 1 Given the data, we estimate

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

And obtain the residuals $\hat{u}_i$

Step 2 We then run the following (auxiliary) regression

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$

Obtain the R-Squared from this (auxiliary) regression

Step 3 Under the null hypothesis that there is no heteroscedasticity, it can be shown that sample size (n) times the R-squared obtained from the auxiliary regression asymptotically follows the chi-square distribution with df equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is,

$$n \cdot R^2 \underset{asy}{\sim} \chi^2_{df}$$

Step 4 If the chi-square value obtained in

$$n \cdot R^2 \underset{asy}{\sim} \chi^2_{df}$$

Exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. It is does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i$$

$$, \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$$

Example P.387-388

From Cross-sectional data on 41 countries

$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$

$Y = ratio\ of\ Trade\ taxes\ to\ total\ Government\ revenue$

$X_2 = ratio\ of\ the\ sum\ of\ Exports\ plus\ imports\ to\ GNP$

$X_3 = GNP\ per\ capita$

- By applying White's heteroscedasticity test to the residuals obtained from regression, the following results were obtained.

$$\hat{u}_i^2 = -5.8417 + 2.5629 \ln Trade_i + 0.6918 \ln GNP_i$$
$$-0.4081(\ln Trade_i)^2 - 0.0491(\ln GNP_i)^2$$
$$+0.0015(\ln Trade_i)(\ln GNP_i)$$
$$R^2 = 0.1148$$

$$n \cdot R^2 = 41(0.1148) = 4.7068$$

The 5 percent critical chi-square value for 5 df is 11.0705.

4.7068<11.0705 On the basis of the White test, that there is no heteroscedasticity

Remedial Measures

Remedial Measures

- When σ_i^2 is known: The Method of Weighted Least Squares
- When σ_i^2 is not known

When σ_i^2 is known

If σ_i^2 is known, the most straightforward method of correcting heteroscedasticity is by means of weighted least squares, for the estimators thus obtained are BLUE.

Example

TABLE 11.4 Illustration of Weighted Least-Squares Regression

Source: Data on Y and σ_i (standard deviation of compensation) are from Table 11.1. Employment size: 1 = 1–4 employees, 2 = 5–9 employees, etc. The latter data are also from Table 11.1.

Compensation, Y	Employment Size, X	σ_i	Y_i/σ_i	X_i/σ_i
3,396	1	742.2	4.5664	0.0013
3,787	2	851.4	4.4480	0.0023
4,013	3	727.8	5.5139	0.0041
4,104	4	805.06	5.0978	0.0050
4,146	5	929.9	4.4585	0.0054
4,241	6	1,080.6	3.9247	0.0055
4,387	7	1,241.2	3.5288	0.0056
4,538	8	1,307.7	3.4702	0.0061
4,843	9	1,110.7	4.3532	0.0081

Note: In regression (11.6.2), the dependent variable is (Y_i/σ_i) and the independent variables are $(1/\sigma_i)$ and (X_i/σ_i).

$\hat{Y}_i = 3417.833 + 148.767 X_i$

```
      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =  106.46
       Model |  1327891.27     1  1327891.27           Prob > F      =  0.0000
    Residual |  87312.7333     7  12473.2476           R-squared     =  0.9383
-------------+------------------------------           Adj R-squared =  0.9295
       Total |     1415204     8    176900.5           Root MSE      =  111.68

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           X |   148.7667   14.4183    10.32   0.000     114.6728    182.8605
       _cons |   3417.833   81.13632   42.12   0.000     3225.976     3609.69
------------------------------------------------------------------------------
```


$\widehat{(Y_i/\sigma_i)} = 3406.639(1/\sigma_i) + 154.153(X_i/\sigma_i)$

```
      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  2,     7) = 4803.02
       Model |  175.811214     2  87.905607            Prob > F      =  0.0000
    Residual |  .128115078     7  .018302154           R-squared     =  0.9993
-------------+------------------------------           Adj R-squared =  0.9991
       Total |  175.939329     9  19.5488143           Root MSE      =  .13529

------------------------------------------------------------------------------
      Ysigma |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Xsigma |   154.2118   16.95407     9.10   0.000     114.1218    194.3018
    consigma |   3406.277   80.96623    42.07   0.000     3214.822    3597.731
------------------------------------------------------------------------------
```


When σ_i^2 is not known

Several assumptions about the pattern of heteroscedasticity

Assumption 1 The error variance is proportional to X_i^2

$$E(u_i^2) = \sigma^2 X_i^2$$

$$\frac{Y_i}{X_i} = \frac{\beta_1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$

$$= \beta_1 \frac{1}{X_i} + \beta_2 + v_i$$

$$E(v_i^2) = E\left(\frac{u_i}{X_i}\right)^2 = \frac{1}{X_i^2} E(u_i^2)$$

$$= \sigma^2$$

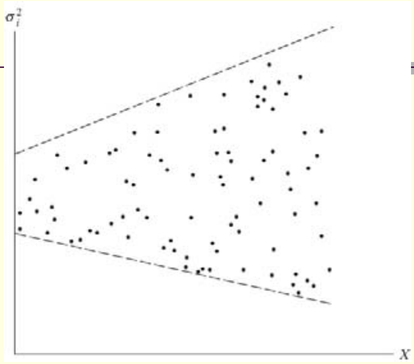
Assumption 2 The error variance is proportional to The **square root transformation** X_i

$$E(u_i^2) = \sigma^2 X_i$$

EE325

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$
$$= \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \nu_i$$

EE325



EE325

Assumption 3 The error variance is proportional to the square of the mean value of Y

$$E(u_i^2) = \sigma^2 \left[E(Y_i)\right]^2$$

EE325

$$E(Y_i) = \beta_1 + \beta_2 X_i$$

$$\frac{Y_i}{E(Y_i)} = \frac{\beta_1}{E(Y_i)} + \beta_2 \frac{X_i}{E(Y_i)} + \frac{u_i}{E(Y_i)}$$
$$= \beta_1 \left(\frac{1}{E(Y_i)}\right) + \beta_2 \frac{X_i}{E(Y_i)} + \nu_i$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
$$\frac{Y_i}{\hat{Y}_i} = \beta_1 \left(\frac{1}{\hat{Y}_i}\right) + \beta_2 \left(\frac{X_i}{\hat{Y}_i}\right) + \nu_i$$

EE325

Assumption 4 A log transformation such as

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$$

very often reduces heteroscedasticity when compared with the regression

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

EE325