ACE 564
Spring 2006

Lecture 9

Violations of Basic Assumptions II: Heteroskedasticity

by
Professor Scott H. Irwin

Readings:

Griffiths, Hill and Judge. "Heteroskedastic Errors," Chapter 15 in *Learning and Practicing Econometrics*

Kennedy. "Introduction," Section 8.1; "Consequences of Violation," Section 8.2; "Heteroskedasticity," Section 8.3 in *A Guide to Econometrics*

The Nature of Heteroskedasticity

In the first part of this course, we introduced the linear economic model to explain the relationship between food expenditure and income,

$$y = \beta_1 + \beta_2 x$$

This linear economic model predicts that food expenditure for a given level of income will be the same for all households

We specified a statistical model by recognizing that actual expenditure for a given level of income will not be the same for all households,

$$y_t = \beta_1 + \beta_2 x_t + e_t \qquad t = 1,...,T$$

where,

- y_t is the dependent variable
- x_t is the independent, or explanatory, variable
- e_t is the error, or disturbance, term
- T is the number of households in the sample

Before re-stating the estimation results for the sample of 40 households, consider the following questions:

- Will it be easier to predict food expenditure for low-income or high-income households?
- Who will have more choices regarding food expenditure?

Estimates for Food Expenditure Data

$$b_1 = 7.3832 \quad b_2 = 0.2323 \quad \hat{\sigma}^2 = 46.853$$

$$\hat{\text{var}}(b_1) = 16.0669 \qquad \hat{\text{var}}(b_2) = 0.0031$$

$$\hat{\text{cov}}(b_1, b_2) = -0.2134$$

Based on this information and the assumption that the statistical model is correctly specified, we can estimate the distributions of e_t and y_t as,

$$e_t \sim N(0, 46.853) \qquad y_t \sim N(7.382 + 0.2323x_t, 46.853)$$

We also can estimate the sampling distributions of b_1 and b_2 as,

$$b_1 \sim N(7.382, 16.0669) \qquad b_2 \sim N(0.2323, 0.0031)$$

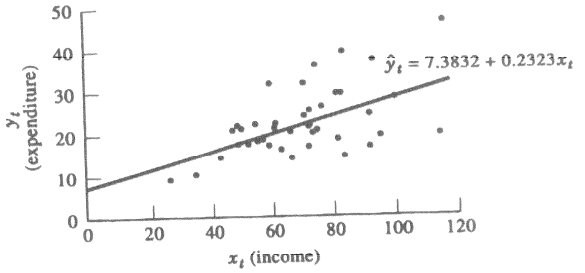

Figure 15.1 The least squares estimated relationship between expenditure on food and income.

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

The scatter plot suggests that data points have a tendency to deviate more and more from the estimated mean function (line) as income increases

Another way to say this is that the least squares residuals,

$$\hat{e}_t = y_t - \beta_1 - \beta_2 x_t$$

increase in absolute value as income grows

Since the observable least squares residuals $\hat{e}_t$ are proxies for the true, unobservable errors,

$$e_t = y_t - \beta_1 - \beta_2 x_t$$

the sample evidence suggests that the unobservable, true errors also increase in absolute value as income grows

Since the "spread" of errors is controlled by the variance of the error term, we are suggesting that the variance of the error term increases as income grows

One of the assumptions of the classical linear regression model is that the variance of the (population) error term is constant,

- $E(e_t^2) = \sigma^2 \quad \forall\, t$
- Termed homoskedasticity

When this assumption is violated, the variance of the (population) error term is not constant

- $E(e_t^2) = \sigma_t^2 \quad \forall\, t$
- Subscript on variance indicates that it is different for different levels of the independent variables
- Termed heteroskedasticity

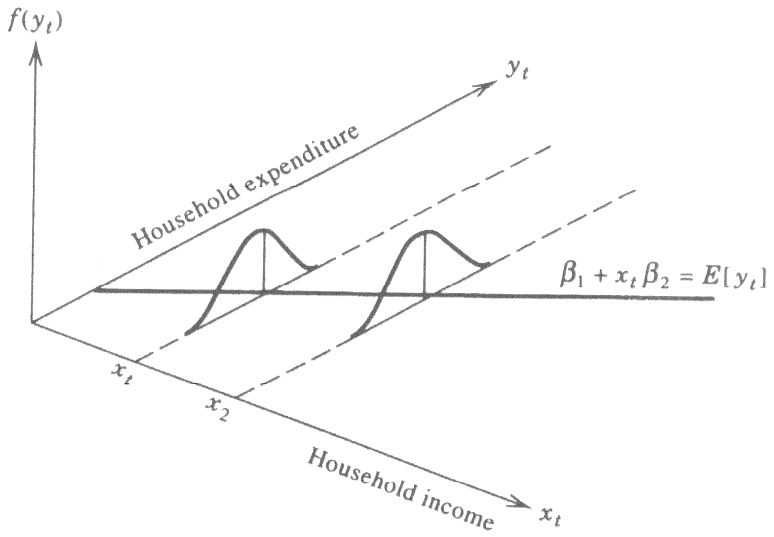

Figure 5.5 The probability density function for y_t at two levels of income.

Griffiths, W.E., R.C. Hill and G.C. Judge. *Learning and Practicing Econometrics*. John Wiley & Sons, Inc., New York, NY, 1993.

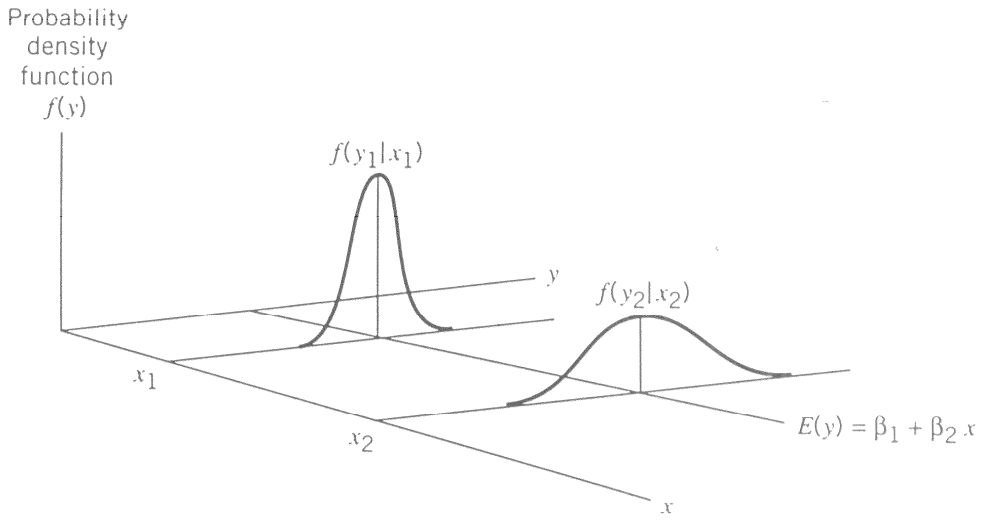

FIGURE 10.2 Heteroskedastic errors

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.

Why Does Heteroskedasticity Arise?

Cross-sectional data

Scale effects

- As income grows, people have more discretion over their consumption choices
- Larger firms have more flexibility in production and investment plans than small firms

Time-series data

Learning curves

- As people learn, errors (hopefully!) become smaller
- Forecast errors in new futures markets

Improvements in technology

- New seed variety that is more drought-resistant

Improvements in data collection techniques

Correlation in "shocks" to economic systems

- Large shocks tend to be followed by large shocks and vice versa

Consequences of Heteroskedasticity for the Least Squares Estimators

If heteroskedasticity is present, then,

- The least squares estimator is still a linear and unbiased estimator, but no longer the best linear unbiased estimator (no longer BLUE)

- The standard errors usually computed for the least squares estimator are incorrect, and hence, confidence intervals and hypothesis tests that use these standard errors may be misleading

To explore these issues, it is most straightforward to continue working with the simple linear regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all assumptions are the same as before except the variance of the regression is assumed to be heteroskedastic,

$$\text{var}(e_t) = E(e_t^2) = \sigma_t^2 \quad \forall\, t$$

When exploring the sampling properties of the least squares estimator of the slope parameter in the simple linear regression model, we noted that the estimator could be written as,

$$b_2 = \beta_2 + \sum_{t=1}^{T} w_t e_t$$

where

$$w_t = \frac{(x_t - \bar{x})}{\sum_{t=1}^{T} (x_t - \bar{x})^2}$$

The mean is derived by taking the expectation of b_2,

$$E(b_2) = E\left(\beta_2 + \sum_{t=1}^{T} w_t e_t \right)$$

$$E(b_2) = E(\beta_2) + \sum_{t=1}^{T} w_t E(e_t)$$

$$E(b_2) = \beta_2$$

This shows that the mean of the sampling distribution of b_2 is β_2, the population slope parameter, even when the variance of the error term is heteroskedastic

We can derive the variance of b_2 as follows,

$$\text{var}(b_2) = \text{var}\left(\beta_2 + \sum_{t=1}^{T} w_t e_t \right) = \text{var}\left(\sum_{t=1}^{T} w_t e_t \right)$$

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \, \text{var}(e_t) + \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \, \text{cov}(e_t, e_s) \qquad t \neq s$$

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \, \text{var}(e_t)$$

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \sigma_t^2$$

If we note that ,

$$w_t^2 = \frac{(x_t - \bar{x})^2}{\left[\sum_{t=1}^{T} (x_t - \bar{x})^2 \right]^2} \quad \text{and} \quad \sum_{t=1}^{T} w_t^2 = \frac{\sum_{t=1}^{T} (x_t - \bar{x})^2}{\left[\sum_{t=1}^{T} (x_t - \bar{x})^2 \right]^2}$$

Then,

$$\text{var}(b_2) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sigma_t^2}{\left[\sum_{t=1}^{T}(x_t - \bar{x})^2\right]^2}$$

This can be compared the sampling variance in the homoskedastic case,

$$\text{var}(b_2) = \sigma^2 \left[\frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\right]$$

Conclusions:

- Least squares estimator of sampling variance is no longer minimum variance, or best
- Least squares is no longer BLUE
- If we use the original least squares formula when heteroskedasticity is present, our estimate of the standard error will be biased

- In general, we cannot predict the direction of the bias in the least squares estimator of the sampling variance of b_2

- The nature of the bias depends on the relationship between the variance of the regression and the independent variable

- Using least squares in the presence of heteroskedasticity will make confidence interval estimates and hypothesis test results misleading

White's Approximate Estimator for the Sampling Variance of Least Squares Estimators

One approach to the problem of heteroskedasticity is to seek "correct" standard error estimates for least squares parameter estimates

- White has developed once such method, based on replacing σ_t^2 with $\hat{e}_t^2$ in the formula for standard error
- The argument is that large variances are likely to lead to large estimated squared residuals
- Because of this approximation, White standard error estimates are valid only "large" sample sizes
- Sometimes White standard errors are called "heteroskedastic-consistent variance-covariance estimates"
- Most econometric packages have commands or options to compute White standard errors

To derive the White estimator, recall that the formula for the sampling variance of the least squares slope estimator with non-constant variance is,

$$\text{var}(b_2) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sigma_t^2}{\left[\sum_{t=1}^{T}(x_t - \bar{x})^2\right]^2}$$

White's estimator is,

$$\hat{\text{var}}_w(b_2) = \frac{\sum_{t=1}^{T}(x_t - \bar{x})^2 \hat{e}_t^2}{\left[\sum_{t=1}^{T}(x_t - \bar{x})^2\right]^2}$$

- Least squares with White standard error estimates is still not BLUE
- Instead, White standard error estimator is consistent
- In very large samples, White standard error estimator converges towards the true estimator

X=Weekly Income	Y=Food Expenditure	*Residuals*	*ehat2*	x-bar	(x-xbar)2	numterm
25.83	9.46	0.93	0.86	69.80	1933.36	1664.51
34.31	10.56	-0.77	0.60		1259.54	753.07
42.50	14.81	0.77	0.60		745.29	443.56
46.75	21.71	6.27	39.28		531.30	20871.07
48.29	22.79	6.84	46.77		462.68	21639.88
48.77	18.19	2.08	4.33		442.26	1914.06
49.65	22.00	5.60	31.36		406.02	12731.40
51.94	18.12	0.96	0.93		318.98	295.96
54.33	23.13	5.18	26.87		239.32	6430.95
54.87	19.00	0.88	0.77		222.90	170.82
56.46	19.46	0.81	0.66		177.96	116.82
58.83	17.83	-1.60	2.57		120.34	309.09
59.13	32.81	13.28	176.31		113.85	20072.94
60.73	22.13	2.07	4.28		82.26	352.41
61.12	23.46	3.27	10.70		75.34	806.08
63.10	16.81	-4.03	16.27		44.89	730.18
65.96	21.35	-0.44	0.19		14.75	2.83
66.40	14.87	-7.06	49.89		11.56	576.71
70.42	33.00	9.74	94.85		0.38	36.46
70.48	25.19	1.91	3.64		0.46	1.69
71.98	17.77	-6.01	36.08		4.75	171.45
72.00	22.44	-1.34	1.80		4.84	8.73
72.23	22.87	-0.99	0.98		5.90	5.77
72.23	26.52	2.66	7.08		5.90	41.81
73.44	21.00	-3.26	10.62		13.25	140.69
74.25	37.52	12.99	168.84		19.80	3343.45
74.77	21.69	-3.01	9.05		24.70	223.48
76.33	27.40	2.19	4.78		42.64	203.91
81.02	30.69	3.93	15.43		125.89	1941.94
81.85	19.56	-7.48	55.90		145.20	8116.72
82.56	30.58	3.31	10.95		162.82	1782.65
83.33	41.12	13.59	184.81		183.06	33831.79
83.40	15.38	-12.17	148.07		184.96	27387.81
91.81	17.87	-12.46	155.17		484.44	75168.53
91.81	25.54	-4.79	22.91		484.44	11099.07
92.96	39.00	8.29	68.78		536.39	36894.44
95.17	20.44	-11.00	120.92		643.64	77829.48
101.40	30.10	-3.39	11.52		998.56	11504.76
114.13	20.90	-16.80	282.22		1965.15	554595.34
115.46	48.71	10.57	111.75		2084.84	232990.08
					15324.63	1167202.41

White's var(b2)	0.00497011
White's se(b2)	0.07049902

Applying White's sampling variance estimator, we obtain the following results for the food expenditure data,

$$\hat{y}_t = 7.3832 + 0.2323x_t \qquad R^2 = 0.317$$
$$(4.2920)\ \ (0.0705) \qquad \text{(White s.e.)}$$

These can be compared to the "incorrect" least squares sampling variance estimates,

$$\hat{y}_t = 7.3832 + 0.2323x_t \qquad R^2 = 0.317$$
$$(4.0080)\ \ (0.0553) \qquad \text{(LS s.e.)}$$

In this case, ignoring heteroskedasticity and using incorrect standard errors tends to overstate the precision of estimation; we tend to get confidence intervals that are narrower than they should be.

We can construct two corresponding 95% confidence intervals for β_2

White: $b_2 \pm t_c \hat{\text{se}}(b_2) = 0.2323 \pm 2.024(0.0705) =$ [0.0896, 0.3750]

Incorrect: $b_2 \pm t_c \hat{\text{se}}(b_2) = 0.2323 \pm 2.024(0.0553) =$ [0.1204, 0.3442]

White's estimator helps overcome the problem of drawing incorrect inferences with least squares in the presence of heteroskedasticity

However, we can go further and ask if it is possible to obtain an estimator for the regression parameters that is superior to least squares or least squares with White standard errors

In other words, is there a BLUE estimator in the presence of heteroskedasticity?

Generalized Least Squares When σ_t^2 is Known

In the case of heteroskedastic variances of the error term, we would like to have an estimator that:

- Places more weight on observations drawn from error distributions with lower variances, and
- Places less weight on observations drawn from error distributions with higher variances

Least squares estimators do not meet this criterion

- Places same weight, or "influence," on each observation, regardless of the disturbance population from which it was drawn

An estimator known as generalized least squares does meet the above criterion

- Also known as weighted least squares

Start with the following regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all "classical" assumptions hold except the variance of the regression is assumed to be heteroskedastic,

$$\text{var}(e_t) = E(e_t^2) = \sigma_t^2 \quad \forall\, t$$

Assume that all the heteroskedastic variances σ_t^2 are known

Next, divide through the original model by σ_t as follows,

$$\frac{y_t}{\sigma_t} = \beta_1 \frac{1}{\sigma_t} + \beta_2 \frac{x_t}{\sigma_t} + \frac{e_t}{\sigma_t}$$

which can be re-written as,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + e_t^*$$

where

$$y_t^* = \frac{y_t}{\sigma_t}, \quad x_{1,t}^* = \frac{1}{\sigma_t}, \quad x_{2,t}^* = \frac{x_t}{\sigma_t}, \quad e_t^* = \frac{e_t}{\sigma_t}$$

The effect of transforming the model on the variance of the error term is

$$\text{var}(e_t^*) = E(e_t^*)^2 = E\left(\frac{e_t}{\sigma_t}\right)^2$$

$$= \frac{1}{\sigma_t^2} E(e_t^2)$$

$$= \frac{1}{\sigma_t^2} \sigma_t^2$$

$$= 1$$

The variance of the error in the transformed model is a constant, and therefore, homoskedastic

If we apply least squares to the transformed model, the least squares estimators will once again be BLUE

Basic idea of generalized least squares is to transform the variables so that the classical linear regression assumptions hold

We can gain further insight about the GLS (generalized least squares) estimator by noting that the objective of GLS is to minimize the sum of transformed squared errors,

$$\min WSSE = \sum_{t=1}^{T} e_t^{*2} = \sum_{t=1}^{T} \frac{e_t^2}{\sigma_t^2}$$

This clearly shows that the squared errors are weighted by the reciprocal of σ_t^2

- When σ_t^2 is small, the observation contains more information, and the observation is weighted more heavily
- When σ_t^2 is large, the observation contains less information, and the observation receives less weight

This shows how GLS takes advantage of heteroskedasticity to improve parameter estimation

Important to emphasize that GLS is BLUE when σ_t^2 is assumed to be known

Generalized Least Squares When σ_t^2 is Unknown: Proportional Heteroskedasticity

Normal situation is that σ_t^2 is unknown

With GLS, an estimation problem is created because we only have T sample observations and T different error variances plus the intercept and slope to estimate

$\Rightarrow$ More parameters than observations!

But, econometricians are a clever bunch (or they like to think so!) and have developed ways of getting around this problem

- The solution is to make further assumptions regarding the process generating error variances
- Typically, assume error variance is a proportional to x_t, the square of x_t, or some other simple functional form, such as quadratic or absolute value

Generalized Least Squares: Heteroskedasticity is Proportional to x_t

Start with the following regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all "classical" assumptions hold except the variance of the regression is assumed to be heteroskedastic,

$$\text{var}(e_t) = E(e_t^2) = \sigma_t^2 = \sigma^2 x_t \quad \forall\, t$$

Implies that the variance of the t^{th} error term is given by a positive unknown constant (σ^2) multiplied by the positive variable x_t

- At high levels of x_t, error variance will be high
- At low levels of x_t, error variance will be low

Our earlier observation of the least squares residuals for the food expenditure problem is consistent with this model; error variance increases as income increases

GLS transformation begins by dividing both sides of the regression model by the square root of x_t (we will see momentarily why we use the square root),

$$\frac{y_t}{\sqrt{x_t}} = \beta_1 \frac{1}{\sqrt{x_t}} + \beta_2 \frac{x_t}{\sqrt{x_t}} + \frac{e_t}{\sqrt{x_t}}$$

which can be re-written as,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + e_t^*$$

where

$$y_t^* = \frac{y_t}{\sqrt{x_t}}, \quad x_{1,t}^* = \frac{1}{\sqrt{x_t}}, \quad x_{2,t}^* = \frac{x_t}{\sqrt{x_t}} = \sqrt{x_t}, \quad e_t^* = \frac{e_t}{\sqrt{x_t}}$$

We can show that the error term for the transformed model is homoskedastic,

The effect of transforming the model on the variance of the error term is,

$$\text{var}(e_t^*) = E(e_t^*)^2 = E\left(\frac{e_t}{\sqrt{x_t}}\right)^2$$

$$= \frac{1}{x_t} E(e_t^2)$$

$$= \frac{1}{x_t} \sigma^2 x_t$$

$$= \sigma^2$$

The variance of the error in the transformed model is a constant, and therefore, homoskedastic

The estimated model will be of the form,

$$\hat{y}_t^* = b_1^* x_{1,t}^* + b_2 x_{2,t}^*$$

which is a multiple regression model without an intercept

- b_1^* is the estimate of β_1
- b_2 is the estimate of β_2
- Usual problem in interpreting R^2 in a model without an intercept

Some important points to note:

- By assuming that error variance is a multiplicative function of x_t, we have solved the problem of having $T+2$ parameters to estimate and only T observations; now only have two parameters to estimate

- The transformed model is linear in the unknown parameters β_1 and β_2, the original parameters that we are interested in estimating

- Transformation of variables with GLS should be viewed as a device for converting a heteroskedastic error model into a homoskedastic error model, not as something that changes the meaning of the coefficients

- The transformed error term will retain the properties $E(e_t^*) = 0$ and zero correlation between different observations, $\text{cov}(e_t^*, e_s^*) = 0$ for $t \neq s$

- As a consequence, we can apply least squares to the transformed variables, y_t^*, $x_{1,t}^*$ and $x_{2,t}^*$ to obtain the best linear unbiased estimators for β_1 and β_2

- The transformed model satisfies the conditions of the Gauss-Markov Theorem, and the least squares estimators defined in terms of the transformed variables are BLUE

- The estimator obtained in this way is also called the weighted least squares estimator (WLS)

GLS Transformation of the Food Expenditure Data

yt	xt	sqrt(xt)	yt*	x1t*	x2t*
9.46	25.83	5.08	1.86	0.20	5.08
10.56	34.31	5.86	1.80	0.17	5.86
14.81	42.50	6.52	2.27	0.15	6.52
21.71	46.75	6.84	3.18	0.15	6.84
22.79	48.29	6.95	3.28	0.14	6.95
18.19	48.77	6.98	2.60	0.14	6.98
22.00	49.65	7.05	3.12	0.14	7.05
18.12	51.94	7.21	2.51	0.14	7.21
23.13	54.33	7.37	3.14	0.14	7.37
19.00	54.87	7.41	2.56	0.13	7.41
19.46	56.46	7.51	2.59	0.13	7.51
17.83	58.83	7.67	2.32	0.13	7.67
32.81	59.13	7.69	4.27	0.13	7.69
22.13	60.73	7.79	2.84	0.13	7.79
23.46	61.12	7.82	3.00	0.13	7.82
16.81	63.10	7.94	2.12	0.13	7.94
21.35	65.96	8.12	2.63	0.12	8.12
14.87	66.40	8.15	1.82	0.12	8.15
33.00	70.42	8.39	3.93	0.12	8.39
25.19	70.48	8.40	3.00	0.12	8.40
17.77	71.98	8.48	2.09	0.12	8.48
22.44	72.00	8.49	2.64	0.12	8.49
22.87	72.23	8.50	2.69	0.12	8.50
26.52	72.23	8.50	3.12	0.12	8.50
21.00	73.44	8.57	2.45	0.12	8.57
37.52	74.25	8.62	4.35	0.12	8.62
21.69	74.77	8.65	2.51	0.12	8.65
27.40	76.33	8.74	3.14	0.11	8.74
30.69	81.02	9.00	3.41	0.11	9.00
19.56	81.85	9.05	2.16	0.11	9.05
30.58	82.56	9.09	3.37	0.11	9.09
41.12	83.33	9.13	4.50	0.11	9.13
15.38	83.40	9.13	1.68	0.11	9.13
17.87	91.81	9.58	1.87	0.10	9.58
25.54	91.81	9.58	2.67	0.10	9.58
39.00	92.96	9.64	4.04	0.10	9.64
20.44	95.17	9.76	2.10	0.10	9.76
30.10	101.40	10.07	2.99	0.10	10.07
20.90	114.13	10.68	1.96	0.09	10.68
48.71	115.46	10.75	4.53	0.09	10.75

GLS Regression Results in Excel for the Food Expenditure Data

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.1970
R Square	0.0388
Adjusted R Square	-0.0128
Standard Error	0.7699
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	0.9094	0.45472	0.76719	0.4716
Residual	38	22.5228	0.5927		
Total	40	23.4322			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	5.7821	3.2566	1.7755	0.0838	-0.8105	12.3747
X Variable 2	0.2552	0.0489	5.2210	0.0000	0.1562	0.3541

The GLS estimation results are,

$$\hat{y}_t = 5.782 + 0.2552 x_t$$
$$(3.257)\ \ (0.0489) \qquad \text{(GLS s.e.)}$$

These can be compared to the White and "incorrect" least squares estimates,

$$\hat{y}_t = 7.3832 + 0.2323 x_t \qquad R^2 = 0.317$$
$$(4.2920)\ \ (0.0691) \qquad \text{(White s.e.)}$$

$$\hat{y}_t = 7.3832 + 0.2323 x_t \qquad R^2 = 0.317$$
$$(4.0080)\ \ (0.0553) \qquad \text{(LS s.e.)}$$

Since GLS is a "better" estimation procedure than least squares or least squares with White standard errors, we expect the GLS standard errors to be lower

The smaller standard errors have the advantage of producing narrower more informative confidence intervals

95% confidence interval estimates for β_2,

GLS:	[0.1562, 0.3542]
White:	[0.0896, 0.3750]
LS:	[0.1204, 0.3442]

GLS with proportional heteroskedasticity is BLUE, assuming the form of the proportional relationship is known

Generalized Least Squares: Heteroskedasticity is Proportional x_t^2

Start with the following regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all "classical" assumptions hold except the variance of the regression is assumed to be heteroskedastic,

$$\text{var}(e_t) = E(e_t^2) = \sigma_t^2 = \sigma^2 x_t^2 \quad \forall\, t$$

Implies that the variance of the t^{th} error term is given by a positive unknown constant (σ^2) multiplied by the positive variable x_t^2

- At high levels of x_t^2, error variance will be high
- At low levels of x_t^2, error variance will be low

GLS transformation begins by dividing both sides of the regression model by x_t (we will see momentarily why we do not use the square root in this case),

$$\frac{y_t}{x_t} = \beta_1 \frac{1}{x_t} + \beta_2 \frac{x_t}{x_t} + \frac{e_t}{x_t}$$

which can be re-written as,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + e_t^*$$

where

$$y_t^* = \frac{y_t}{x_t}, \quad x_{1,t}^* = \frac{1}{x_t}, \quad x_{2,t}^* = \frac{x_t}{x_t} = 1, \quad e_t^* = \frac{e_t}{x_t}$$

We can show that the error term for the transformed model is homoskedastic,

The effect of transforming the model on the variance of the error term is,

$$\text{var}(e_t^*) = E(e_t^*)^2 = E\left(\frac{e_t}{x_t}\right)^2$$

$$= \frac{1}{x_t^2} E(e_t^2)$$

$$= \frac{1}{x_t^2}\sigma^2 x_t^2 = \sigma^2$$

The variance of the error in the transformed model is a constant, and therefore, homoskedastic

If we apply least squares to the transformed model, the least squares estimators will once again be BLUE

The estimated model is,

$$\hat{y}_t^* = b_1^* x_{1,t}^* + b_2 x_{2,t}^*$$

Multiple regression model without an intercept

- b_1^* is the estimate of β_1
- b_2 is the estimate of β_2

In practice, simply regress $\frac{y_t}{x_t}$ on $\frac{1}{x_t}$ and a constant,

but reverse the interpretation of the parameter estimates!

Important points:

- The GLS transformation "works" (generates BLUE estimators) because it is a function of known x_t values
- GLS transformation generalizes to the multiple variable regression case
- Simply transform all variables as in the two-variable case
- Estimation and interpretation is the same as in two-variable case
- Heteroskedasticity can be a multiplicative function of more than one independent variable

A final caution regarding GLS

If assumed form of heteroskedasticity is correct,

- GLS will produce BLUE parameter estimators
- White's correction is less efficient than GLS

If assumed form of heteroskedasticity is incorrect,

- GLS will produce biased parameter estimates and biased standard error estimates
- A form of specification error

If there is "substantial" uncertainty about form of heteroskedasticity, probably better to use White standard error correction, as parameter estimators are unbiased and "consistent"

An Introduction to Estimated Generalized Least Squares: A Sample with a Heteroskedastic Partition

In each of the preceding examples of generalized least squares, the transformation eliminated the need to directly estimate the changing variance parameter

There may be situations where such transformations cannot be applied

Such situations requires the use of estimated generalized least squares (EGLS)

The Problem

Need to model the supply of wheat in an area of Australia

Information on supply response to price is important for government policy purposes

Government pays a guaranteed price for wheat and needs to know how much wheat will be produced at the guaranteed price

The Economic Model

Production economics teaches us that the quantity of wheat supplied will depend on the production technology of the firm, price expectations of the firm, and weather conditions

We can depict this economic model as,

$$\text{Quantity} = f(\text{Price, Technology, Weather})$$

The Statistical Model

Choice of statistical model will depend on the type of data available

Cross-section

- Adjust prices for differences in transportation and input costs
- Technology is likely to be similar across farmers
- Weather will be similar for farmers in a the same region

Time-series

- Prices will vary across time
- Production technology will presumably improve through time
- Weather conditions will vary from year-to-year

Time-series is available for 26 years of aggregate quantity supplied and price

- Use a simple linear time-trend variable to proxy changes in production technology
- Effect of weather is argued to be random and included in the error term

Table 15.1 Data on Quantity, Price, and Trend for an Australian Wheat-Growing District

q	p	t
197.6	1.47	1
140.1	1.30	2
162.3	1.59	3
166.5	1.44	4
159.5	1.89	5
195.6	1.49	6
207.0	1.94	7
218.4	1.52	8
239.0	2.15	9
208.2	2.09	10
253.4	1.74	11
278.7	2.51	12
221.1	2.14	13
240.0	2.42	14
236.1	2.45	15
234.5	2.44	16
239.0	2.26	17
258.4	2.50	18
247.9	2.41	19
272.2	2.83	20
266.2	2.79	21
284.1	3.17	22
283.4	2.83	23
277.4	2.69	24
301.0	3.65	25
281.4	3.36	26

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics John Wiley & Sons, Inc. New York. 1993.

The statistical model is specified as,

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t \qquad t = 1,\ldots,26$$

where

- y_t is the quantity of wheat produced in year t
- $x_{2,t}$ is the price of wheat guaranteed in year t
- $x_{3,t}$ is a trend variable to capture changes in production technology and it takes on values 1,..,26
- e_t is the random error term that accounts for random influences on wheat production, including weather

To complete the model, as before, we need to specify the statistical assumptions for the error term

- We could assume the error term is homoskedastic
- But, we have additional information that suggests this is not correct

We know that after year 13 new wheat varieties were introduced that were less susceptible to variations in weather conditions

- Mean yields did not change
- Yield variability did change

This is modeled as follows,

$$E(e_t) = 0 \quad t = 1,..,26$$

$$\text{var}(e_t) = E(e_t^2) = \sigma_1^2 \quad t = 1,..,13$$

$$\text{var}(e_t) = E(e_t^2) = \sigma_2^2 \quad t = 14,..,26$$

$$\text{cov}(e_t, e_s) = 0 \quad \forall t, s \text{ where } t \neq s$$

We further assume that,

$$\sigma_1^2 > \sigma_2^2$$

With the previous assumptions we can write our model in the following "partitioned" format,

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t \quad \text{var}(e_t) = \sigma_1^2 \quad t = 1,...,13$$

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t \quad \text{var}(e_t) = \sigma_2^2 \quad t = 14,...,26$$

We can transform this partitioned model so that the two partitions have the same error variance as follows,

$$\frac{y_t}{\sigma_1} = \beta_1 \frac{1}{\sigma_1} + \beta_2 \frac{x_{2,t}}{\sigma_1} + \beta_3 \frac{x_{3,t}}{\sigma_1} + \frac{e_t}{\sigma_1} \qquad t = 1,...,13$$

$$\frac{y_t}{\sigma_2} = \beta_1 \frac{1}{\sigma_2} + \beta_2 \frac{x_{2,t}}{\sigma_2} + \beta_3 \frac{x_{3,t}}{\sigma_2} + \frac{e_t}{\sigma_2} \qquad t = 14,...,26$$

- Using the same logic as we have followed before, it can be proven that the error variance in both partitions is the same (equals one)
- Transformed model is homoskedastic

We can write the transformed model using one equation,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + \beta_3 x_{3,t}^* + e_t^*$$

where

$$y_t^* = \frac{y_t}{\sigma_i}, \quad x_{1,t}^* = \frac{1}{\sigma_i}, \quad x_{2,t}^* = \frac{x_{2,t}}{\sigma_i}, \quad x_{3,t}^* = \frac{x_{3,t}}{\sigma_i}, \quad e_t^* = \frac{e_t}{\sigma_i}$$

$i=1$ when $t=1,..13$ and $i=2$ when $t=14,\ldots,26$

Providing σ_1 and σ_2 are known, the transformed model provides a set of new transformed variables to which we can apply LS to obtain the best linear unbiased estimators for (β_1, β_2 and β_3).

Like before, the complete process of transforming variables, then applying least squares to the transformed variables, is called generalized least squares or weighted least squares.

However, we cannot simply apply least squares as before because the transformed model depends on the unknown parameters σ_1^2 and σ_2^2

We now move into a new estimation method: estimated generalized least squares (EGLS)

In this case, EGLS is a three-step process

Step 1:

- Apply LS to the first half of the sample to obtain an estimate of σ_1^2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.814796
R Square	0.663893
Adjusted R Square	0.596671
Standard Error	25.33063
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	12673.94	6336.972	9.8762	0.004289
Residual	10	6416.407	641.6407		
Total	12	19090.35			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	121.1745	44.20106	2.741439	0.020781	22.68839	219.6606
X Variable 1	19.14782	32.42234	0.590575	0.567912	-53.0937	91.38931
X Variable 2	6.885293	3.004995	2.291283	0.044916	0.189746	13.58084

- Apply LS to the second half of the sample to obtain an estimate of σ_2^2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.949223
R Square	0.901025
Adjusted R Square	0.88123
Standard Error	7.599905
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	5258.095	2629.047	45.51789	9.5E-06
Residual	10	577.5855	57.75855		
Total	12	5835.68			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	137.2942	14.67813	9.353663	2.92E-06	104.5893	169.9991
X Variable 1	22.06256	9.280747	2.377239	0.038795	1.383761	42.74136
X Variable 2	3.257443	0.994419	3.275726	0.008349	1.04174	5.473146

- Results: $\hat{\sigma}_1^2 = 641.64$ and $\hat{\sigma}_2^2 = 57.76$

Step 2:

Transform the original observations using the square root of the variance estimates obtained in step 1

q	p	t	q*	int*	p*	t*
197.6	1.47	1	7.800833	0.039478	0.058033	0.039478
140.1	1.3	2	5.530854	0.039478	0.051321	0.078956
162.3	1.59	3	6.407263	0.039478	0.06277	0.118434
166.5	1.44	4	6.57307	0.039478	0.056848	0.157912
159.5	1.89	5	6.296725	0.039478	0.074613	0.197389
195.6	1.49	6	7.721877	0.039478	0.058822	0.236867
207	1.94	7	8.171925	0.039478	0.076587	0.276345
218.4	1.52	8	8.621973	0.039478	0.060006	0.315823
239	2.15	9	9.435218	0.039478	0.084877	0.355301
208.2	2.09	10	8.219299	0.039478	0.082509	0.394779
253.4	1.74	11	10.0037	0.039478	0.068692	0.434257
278.7	2.51	12	11.00249	0.039478	0.09909	0.473735
221.1	2.14	13	8.728563	0.039478	0.084483	0.513213
240	2.42	14	31.57934	0.131581	0.318425	1.842128
236.1	2.45	15	31.06618	0.131581	0.322372	1.973709
234.5	2.44	16	30.85565	0.131581	0.321057	2.10529
239	2.26	17	31.44776	0.131581	0.297372	2.23687
258.4	2.5	18	34.00043	0.131581	0.328951	2.368451
247.9	2.41	19	32.61883	0.131581	0.317109	2.500031
272.2	2.83	20	35.81624	0.131581	0.372373	2.631612
266.2	2.79	21	35.02676	0.131581	0.36711	2.763193
284.1	3.17	22	37.38205	0.131581	0.41711	2.894773
283.4	2.83	23	37.28994	0.131581	0.372373	3.026354
277.4	2.69	24	36.50046	0.131581	0.353952	3.157934
301	3.65	25	39.60576	0.131581	0.480269	3.289515
281.4	3.36	26	37.02678	0.131581	0.442111	3.421096

Step 3:

Apply LS to the entire set of 26 transformed observations

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.9975**
R Square	**0.9950**
Adjusted R Square	**0.9511**
Standard Error	**1.0135**
Observations	**26**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**3**	**4703.7078**	**1567.9026**	**1526.4681**	**0.0000**
Residual	**23**	**23.6243**	**1.0271**		
Total	**26**	**4727.3321**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**0.0000**	**#N/A**	**#N/A**	**#N/A**	**#N/A**	**#N/A**
X Variable 1	**138.0541**	**12.8210**	**10.7678**	**0.0000**	**111.5319**	**164.5763**
X Variable 2	**21.7198**	**8.9239**	**2.4339**	**0.0231**	**3.2592**	**40.1803**
X Variable 3	**3.2834**	**0.8226**	**3.9914**	**0.0006**	**1.5817**	**4.9852**

EGLS results:

$$\hat{y}_t = 138.05 + 21.72x_{2,t} + 3.28x_{3,t}$$
$$(12.82) \quad (8.92) \qquad (0.82) \qquad (s.e.)$$

For comparison, LS results:

$$\hat{y}_t = 139.90 + 19.54x_{2,t} + 3.64x_{3,t} \qquad R^2 = 0.809$$
$$(23.22) \quad (17.41) \quad (1.42) \qquad (s.e.)$$

- Parameter estimates are little changed between LS and EGLS
- Standard error estimates are substantially lower with EGLS

Important points:

- EGLS does not produce BLUE estimators
- EGLS only produces consistent estimators

Detecting Heteroskedasticity

Economic theory may be of limited guidance as to whether we should expect heteroskedasticity in a particular applied research problem

Nevertheless, always try to use economic theory, or other non-sample information, as the first means of detecting this problem

Without strong *a priori* information, all one can do is estimate the regression model using LS assuming there is no heteroskedasticity and then conduct a "post-mortem" analysis on the estimated residuals

If evidence of heteroskedasticity is found, then GLS or EGLS can be applied as a correction

> $\Rightarrow$Monte Carlo studies suggest this "two-step" estimator is reasonably efficient, if the overall model specification is correct

Graphical Methods

Plot LS residuals and examine whether they have any pattern

- Plots may use raw residuals or squared residuals
- Generate plots against each independent variable
- If the errors are homoskedastic, there should be no pattern of any kind
- If a pattern is detected, then a more formal investigation is warranted

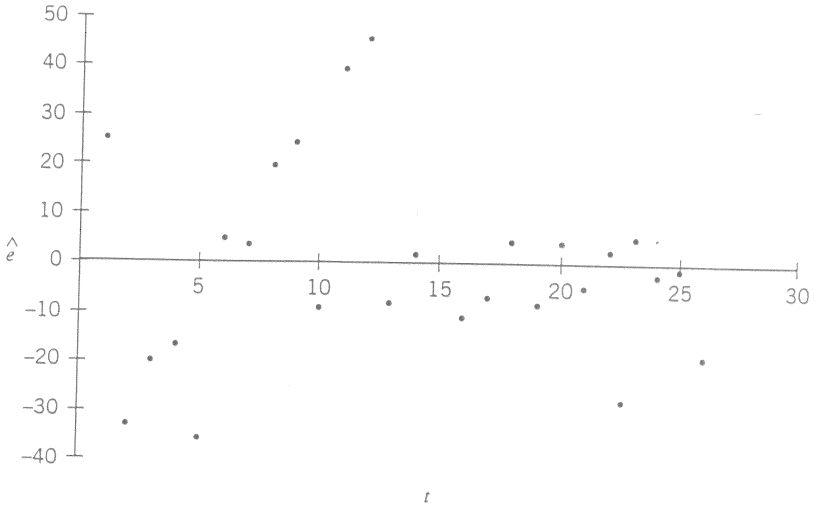

FIGURE 10.3 Least squares residuals plotted against time

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.

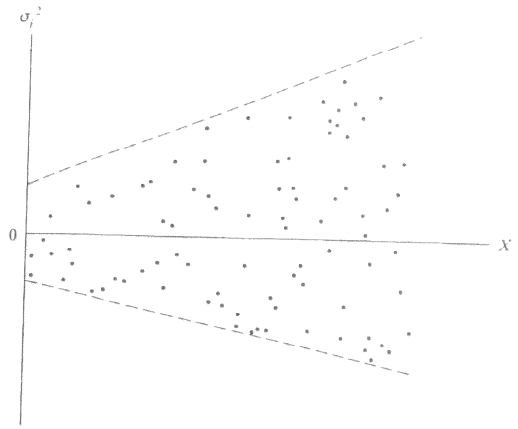

FIGURE 11.10
Error variance proportional to X.

Gujarati, Damodar N. Basic Econometrics. McGraw-Hill, Inc. 1995.

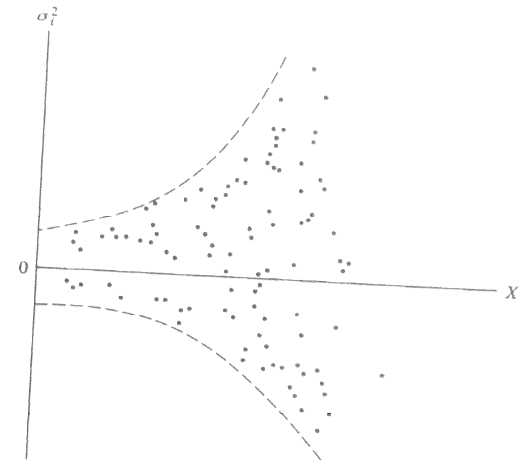

FIGURE 11.9
Error variance proportional to X^2.

Gujarati, Damodar N. Basic Econometrics. McGraw-Hill, Inc. 1995.

Formal Tests for Heteroskedasticity

There are a large number of tests for heteroskedasticity

Kennedy discusses three widely used tests in Ch.8

- Goldfield-Quandt test
- Breusch-Pagan test
- White test

Goldfield-Quandt Test

Basic idea is to compare estimated error variance from the "high variance" part of the observations to the "low variance" part of the observations

- Operationally, it takes the form of an *F*-test
- Assumes that heteroskedasticity is positively related to one of the *x*'s

1. Null hypothesis

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

2. Test statistic

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F(T_1 - K_1, T_2 - K_2)$$

where

T_1 is the number of observations in the first partition of the sample

K_1 is the number of parameters in the model estimated for the first partition

T_2 is the number of observations in the second partition of the sample

K_2 is the number of parameters in the model estimated for the second partition

- Given this setup it is not necessary to have the same number of observations in the two data partitions

- Always put the variance expected to be the largest in the numerator

3. Rejection region

Reject the null hypothesis if,

$$GQ > F_{\alpha}(T_1 - K_1, T_2 - K_2)$$

GQ test applied to wheat supply data

We first make sure that data are ordered by time

- Apply LS to the first half of the sample to obtain $\hat{\sigma}_1^2$
- Apply LS to the second half of the sample to obtain $\hat{\sigma}_2^2$

1. Null hypothesis

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

2. Test statistic

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{641.64}{57.76} = 11.11$$

3. Rejection Region

Reject the null hypothesis if,

$$GQ > F_{\alpha}(T_1 - K_1, T_2 - K_2)$$

If $\alpha = 0.05$, then,

$$F_{\alpha}(T_1 - K_1, T_2 - K_2) =$$
$$F_{0.05}(13-3, 13-3) =$$
$$F_{0.05}(10,10) = 2.98$$

Reject if $GQ \geq 2.98$

4. Decision

Since 11.11>2.98 we reject the null hypothesis and conclude that the error variance is larger in the first half of the sample than in the second half of the sample

New varieties of wheat did reduce the variance of yields

GQ test applied to food expenditure data

We first make sure that data are ordered by the magnitude of x_t, household income

- Apply LS to the second half of the sample (high income) to obtain $\hat{\sigma}_1^2$
- Apply LS to the first half of the sample (low income) to obtain $\hat{\sigma}_2^2$

1. Null hypothesis

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2$$

2. Test statistic

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{74.933}{22.377} = 3.35$$

3. Rejection Region

Reject the null hypothesis if,

$$GQ > F_{\alpha}(T_1 - K_1, T_2 - K_2)$$

If $\alpha = 0.05$, then,

$$F_{\alpha}(T_1 - K_1, T_2 - K_2) =$$
$$F_{0.05}(20 - 2, 20 - 2) =$$
$$F_{0.05}(18,18) = 2.22$$

Reject if $GQ \geq 2.22$

4. Decision

Since 3.35>2.22 we reject the null hypothesis and conclude that the error variance is larger in the higher income half of the sample than in the lower income half of the sample

The power of the GQ test can be improved by leaving out some of the central observations in the sample

- The ability of the test to detect heteroskedasticity is "sharpened" if central observations are omitted
- Monte Carlo studies suggests the middle 10-15% of observations should be deleted to maximize the effectiveness of the GQ test