EE 425: Econometrics

Relaxing the assumption in the classical model: Heteroscedastic Disturbances

- 1. Nature of heteroscedasticity : variances of the disturbance term are not constant.
- 2. Consequences
- 3. Detection of heteroscedasticity
- 4. Remedial measures

Reasons for heteroscedasticity

- Economic reason: income-consumption behaviour
- Error-learning model:with more experience, errors and inconsistency decrease
- Data collection improves overtime, and error is likely to decrease
- The presence of outlier
- Problem with specification
- etc.
- Note: heteroscedasticity is more common in cross-sectional data

Examples:

Figure 11.1
Homoscedastic disturbance

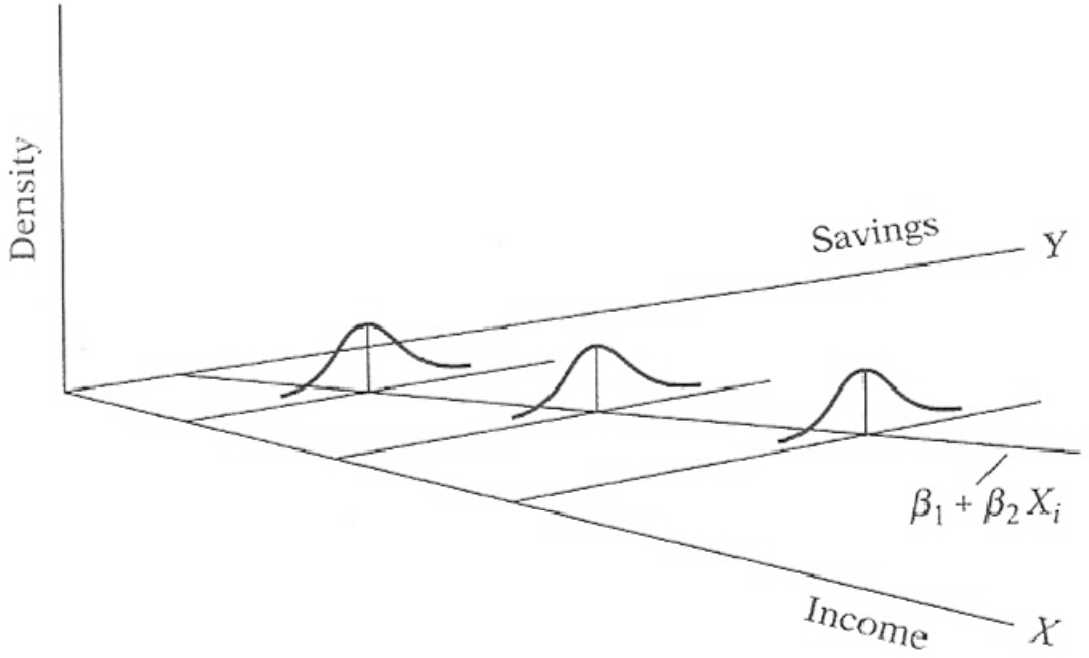


Figure 11.2
Heteroscedastic disturbance

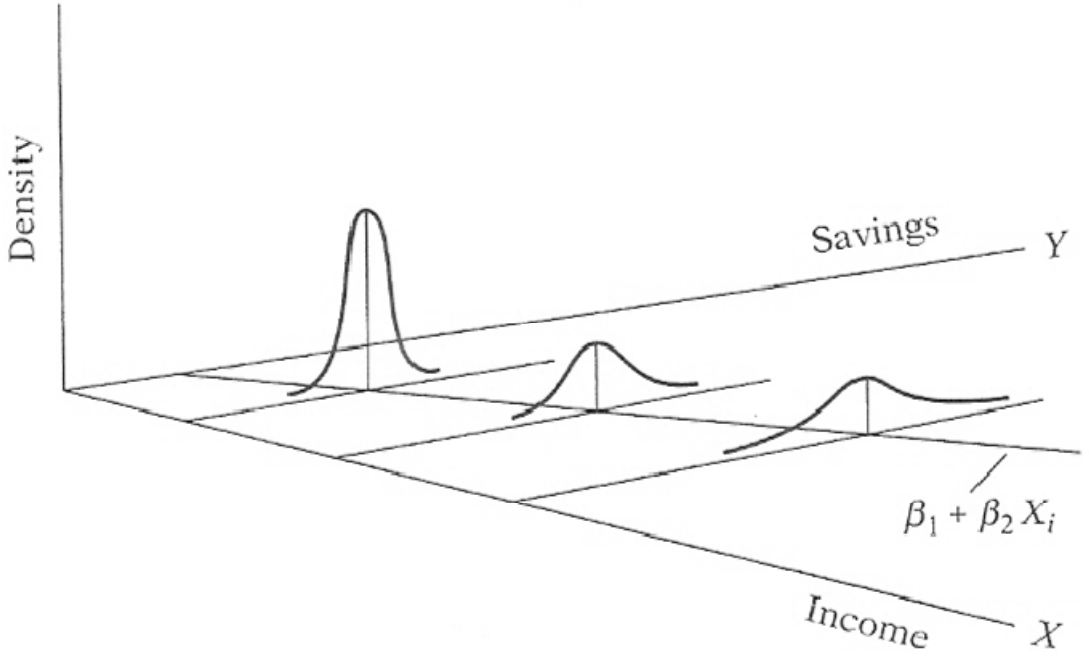
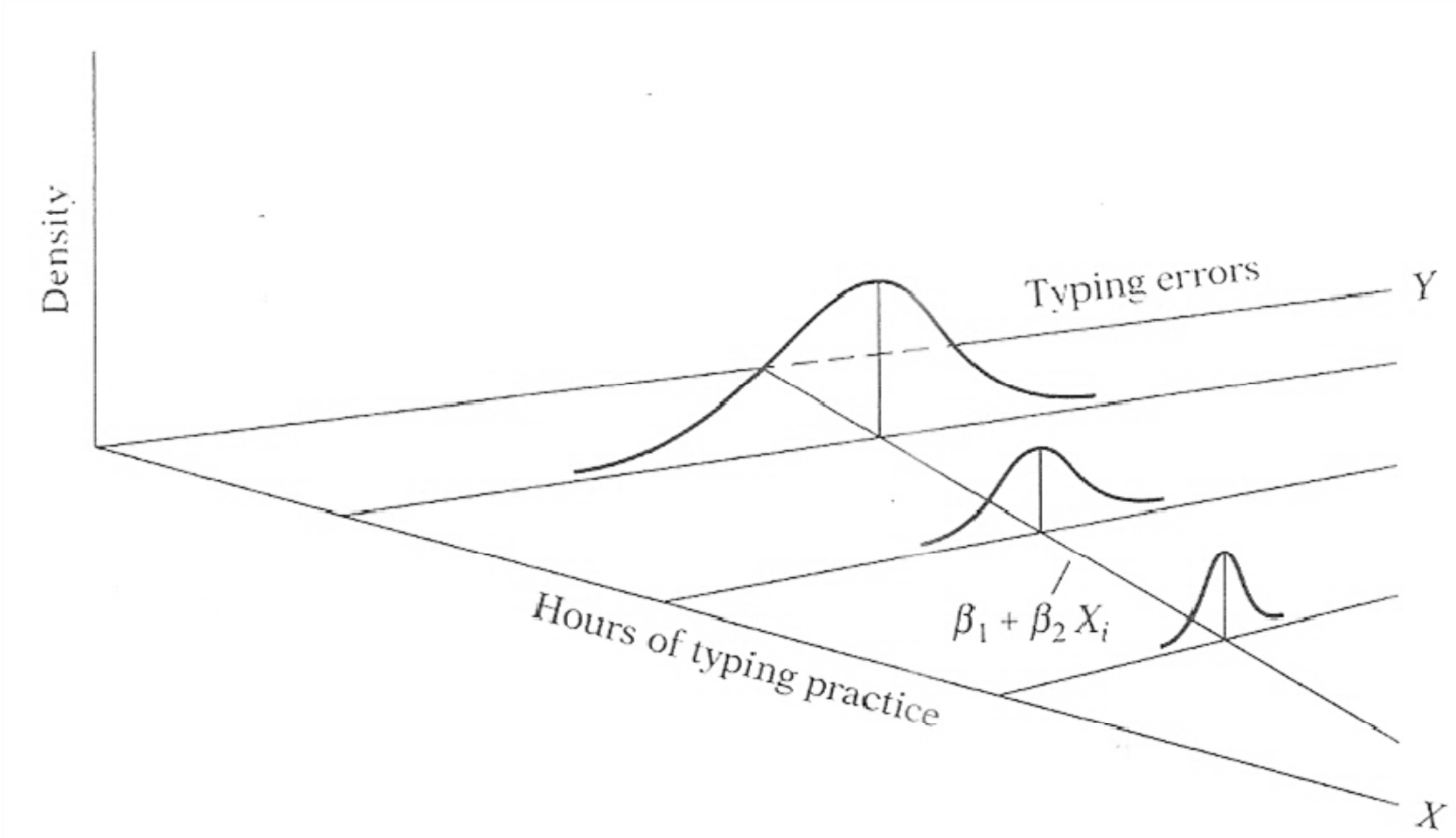
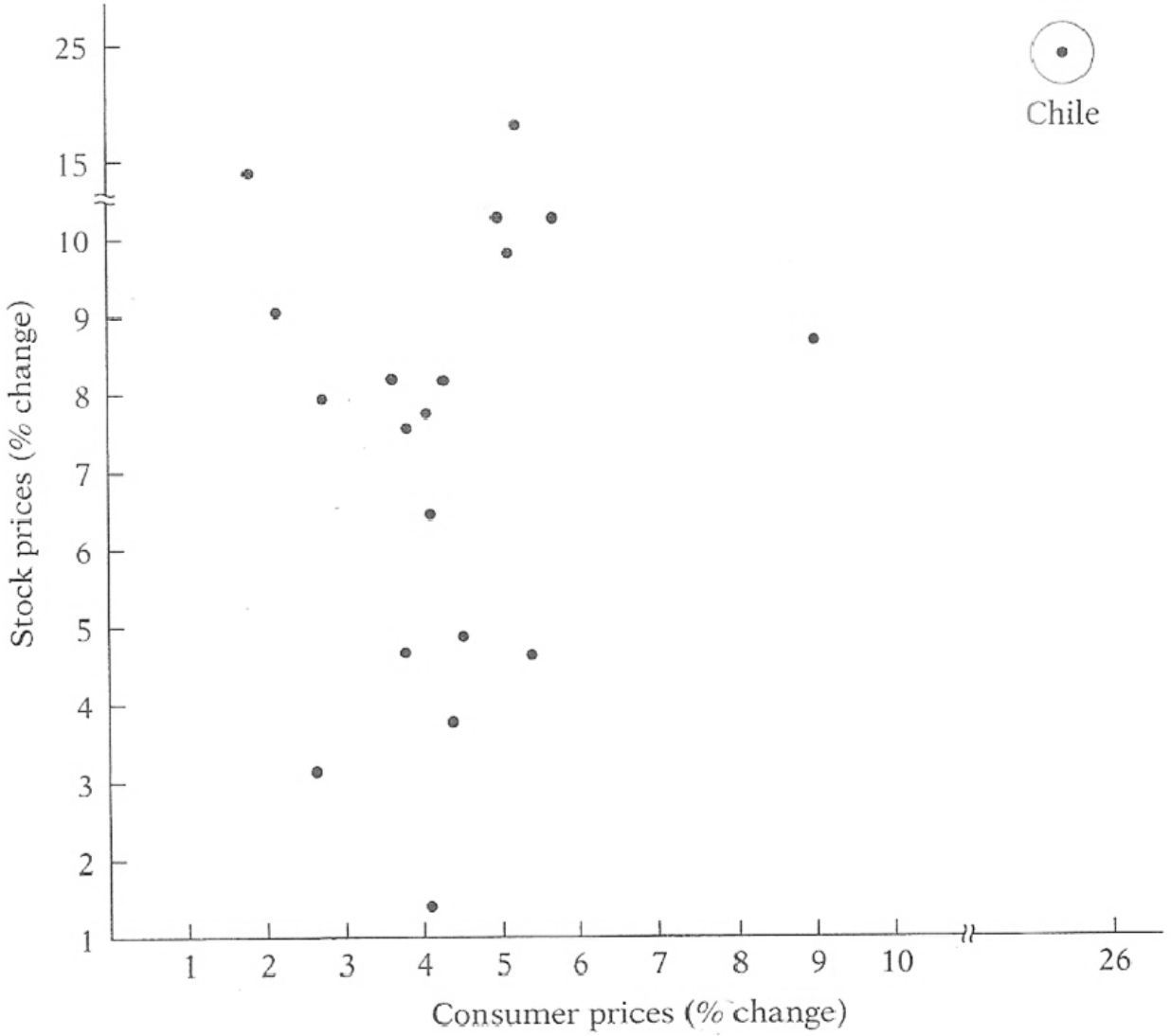


Figure 11.3 Illustration of Heteroscedasticity



Example of Outliers

The Relationship between Stock Prices and Consumer Prices



OLS Estimation in the Presence of Heteroscedasticity

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

Applying the usual formula, the OLS estimator of β_2 is

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \qquad (11.2.1)$$

but its variance is now given by the following expression (see Appendix 11A, Section 11A.1):

$$\text{var}\,(\hat{\beta}_2) = \frac{\sum x_i^2 \sigma_i^2}{\left(\sum x_i^2\right)^2} \qquad (11.2.2)$$

which is obviously different from the usual variance formula obtained under the assumption of homoscedasticity, namely,

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \qquad (11.2.3)$$

Consequences of heteroscedasticity to OLS estimators

- OLS estimators are still unbiased but they are not best (not efficient) because the assumption of homoscedasticity has been violated.
- Formula for variance of OLS estimators are biased.
- t and F tests are not valid.

Detection of heteroscedasticity

_Graphical method

_Formal methods

Detection of Heteroscedasticity: Graphical Method

Figure 11.8 Hypothetical Patterns of Estimatied Squared Residuals.

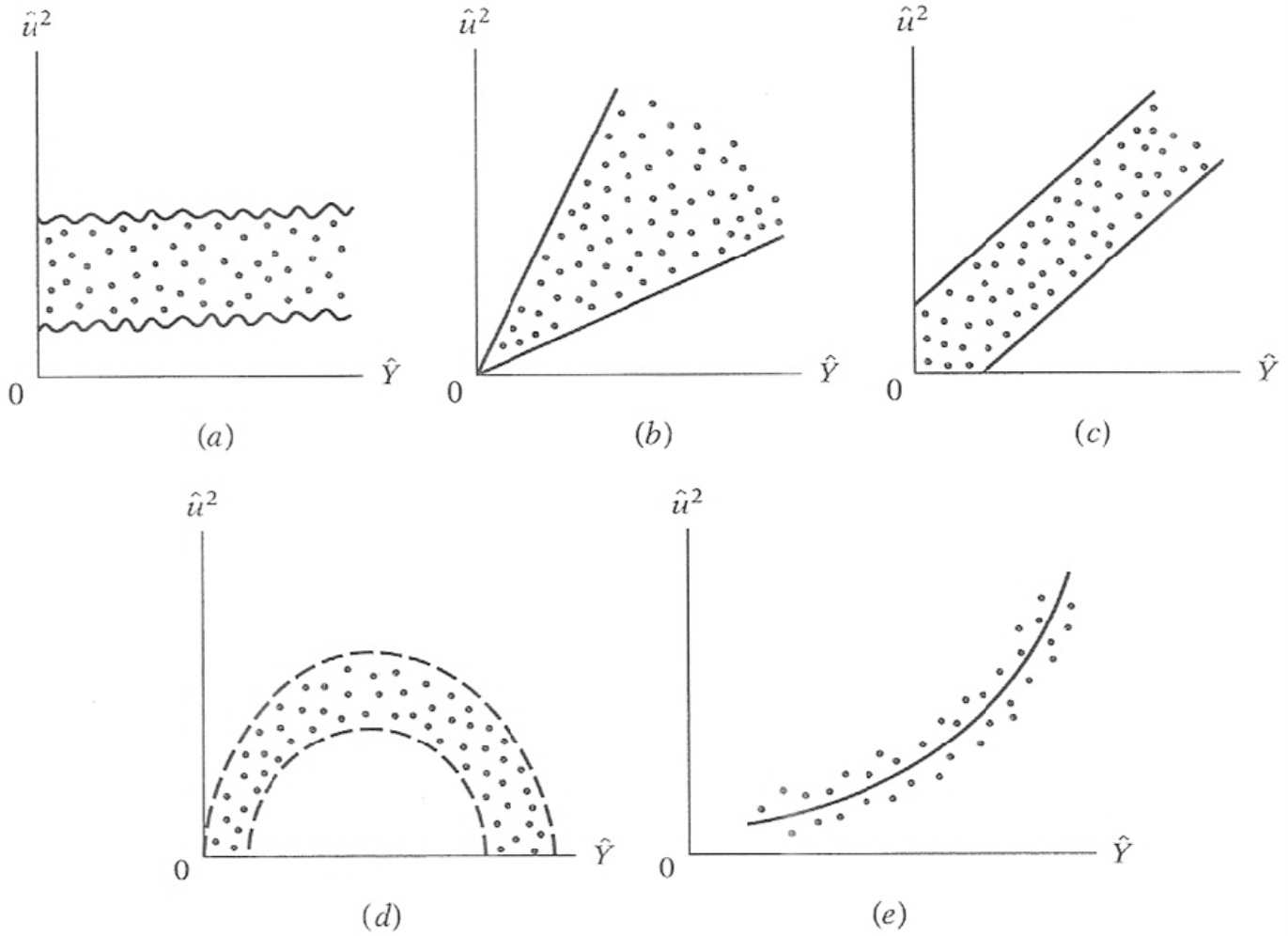
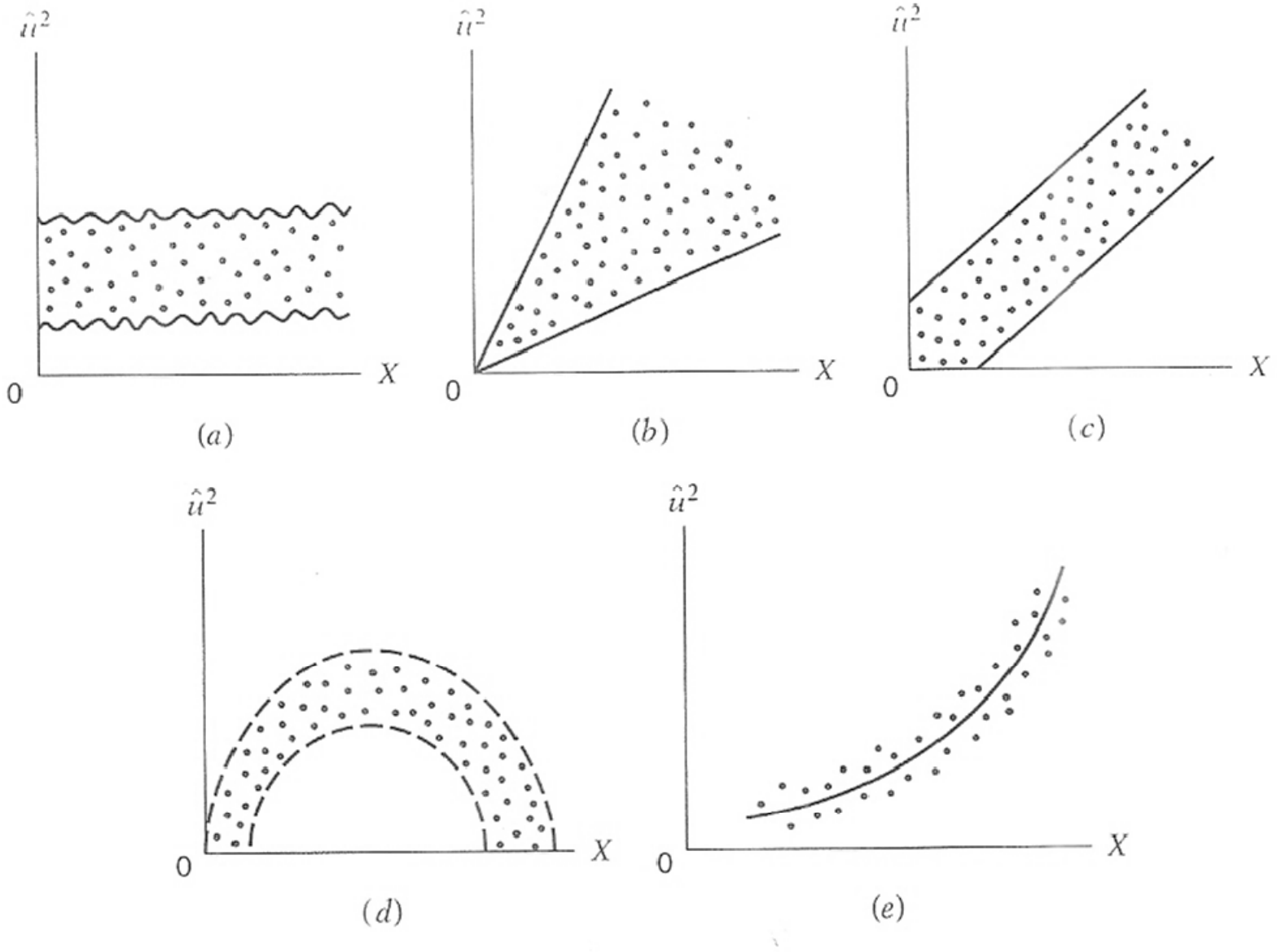


Figure 11.9 Scatter of Estimated Squared Residuals Against X.



Formal Methods: 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

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln X_i + v_i \qquad (11.5.1)$$

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:

$$\begin{aligned} \ln \hat{u}_i^2 &= \ln \sigma^2 + \beta \ln X_i + v_i \\ &= \alpha + \beta \ln X_i + v_i \end{aligned} \qquad (11.5.2)$$

If β turns out to be statistically significant, it would suggest that heteroscedasticity is present in the data. If it turns out to be insignificant, we may accept the assumption of homoscedasticity. The Park test is thus a two-stage procedure. In the first stage we run the OLS regression disregarding the heteroscedasticity question. We obtain $\hat{u}_i$ from this regression, and then in the second stage we run the regression (11.5.2).

Example 11.1: Relationship between Compensation and Productivity

To illustrate the Park approach, we use the data given in Table 11.1 to run the following regression:

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

where Y = average compensation in thousands of dollars, X = average productivity in thousands of dollars, and $i = i$th employment size of the establishment. The results of the regression are as follows:

$$\begin{aligned} \hat{Y}_i &= 1992.3452 + 0.2329X_i \\ \text{se} &= (936.4791) \quad (0.0998) \\ t &= \quad (2.1275) \quad (2.333) \qquad R^2 = 0.4375 \end{aligned} \tag{11.5.3}$$

The results reveal that the estimated slope coefficient is significant at the 5 percent level on the basis of a one-tail t test. The equation shows that as labor productivity increases by, say, a dollar, labor compensation on the average increases by about 23 cents.

The residuals obtained from regression (11.5.3) are then regressed on X_i as suggested in Eq. (11.5.2), giving the following results:

$$\begin{aligned} \widehat{\ln \hat{u}_i^2} &= 35.817 - 2.8099 \ln X_i \\ \text{se} &= (38.319) \quad (4.216) \\ t &= \quad (0.934) \quad (-0.667) \qquad R^2 = 0.0595 \end{aligned} \tag{11.5.4}$$

Obviously, there is no statistically significant relationship between the two variables. Following the Park test, one may conclude that there is no heteroscedasticity in the error variance.[13]

Glejser Test

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

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

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

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

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

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

where v_i is the error term.

Again as an empirical or practical matter, one may use the Glejser approach. But Goldfeld and Quandt point out that the error term v_i has some problems in that its expected value is nonzero, it is serially correlated (see Chapter 12), and, ironically, it is heteroscedastic.[15] An additional difficulty with the Glejser method is that models such as

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

and

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

are nonlinear in the parameters and therefore cannot be estimated with the usual OLS procedure.

Example 11.2: Relationship between Compensation and Productivity

Continuing with Example 11.1, the absolute value of the residuals obtained from regression (11.5.3) were regressed on average productivity (X), giving the following results:

$$\widehat{|\hat{u}_i|} = 407.2783 - 0.0203X_i$$
$$\text{se} = (633.1621) \quad (0.0675) \qquad r^2 = 0.0127 \qquad \textbf{(11.5.5)}$$
$$t = \quad (0.6432) \quad (-0.3012)$$

As you can see from this regression, there is no relationship between the absolute value of the residuals and the regressor, average productivity. This reinforces the conclusion based on the Park test.

Goldfeld-Quandt Test

This popular method is applicable if one assumes that the heteroscedastic variance, σ_i^2, is positively 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

$$\sigma_i^2 = \sigma^2 X_i^2 \qquad (11.5.10)$$

where σ^2 is a constant.[18]

Assumption (11.5.10) postulates that σ_i^2 is proportional to the square of the X variable. Such an assumption has been found quite useful by Prais and Houthakker in their study of family budgets. (See Section 11.5, informal methods.)

If Eq. (11.5.10) is appropriate, it would mean σ_i^2 would be larger, the larger the values of X_i. If that turns out to be the case, heteroscedasticity is most likely to be present in the model. To test this explicitly, Goldfeld and Quandt suggest the following steps:

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 $(n - c)/2$ observations and the last $(n - c)/2$ observations, and obtain the respective residual sums of squares RSS_1 and

RSS_2, RSS_1 representing the RSS from the regression corresponding to the smaller X_i values (the small variance group) and RSS_2 that from the larger X_i values (the large variance group). These RSS each have

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

where k is the number of parameters to be estimated, including the intercept. (Why?) For the two-variable case k is of course 2.

Step 4. Compute the ratio

$$\lambda = \frac{RSS_2/\text{df}}{RSS_1/\text{df}} \tag{11.5.11}$$

If we assume u_i are normally distributed (which we usually do), and *if the assumption of homoscedasticity is valid*, then it can be shown that λ of Eq. (11.5.10) 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, that is, we can say that heteroscedasticity is very likely.

Example 11.4:
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	

Example 11.4: The Goldfeld-Quandt Test (Cont')

Regression based on the first 13 observations:

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

(8.7049) (0.0744) $r^2 = 0.8887$ $RSS_1 = 377.17$ df = 11

Regression based on the last 13 observations:

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

(30.6421) (0.1319) $r^2 = 0.7681$ $RSS_2 = 1536.8$ df = 11

From these results we obtain

$$\lambda = \frac{RSS_2/df}{RSS_1/df} = \frac{1536.8/11}{377.17/11}$$

$$\lambda = 4.07$$

The critical F value for 11 numerator and 11 denominator df at the 5 percent level is 2.82. Since the estimated $F\ (= \lambda)$ value exceeds the critical value, we may conclude that there is heteroscedasticity in the error variance. However, if the level of significance is fixed at 1 percent, we may not reject the assumption of homoscedasticity. (Why?) Note that the p value of the observed λ is 0.014.

Breusch-Pagan-Godfrey Test

The success of the Goldfeld–Quandt test depends not only on the value of c (the number of central observations to be omitted) but also on identifying the correct X variable with which to order the observations. This limitation of this test can be avoided if we consider the Breusch–Pagan–Godfrey (BPG) test.

To illustrate this test, consider the k-variable linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \qquad (11.5.12)$$

Assume that the error variance σ_i^2 is described as

$$\sigma_i^2 = f(\alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}) \qquad (11.5.13)$$

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

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} \qquad (11.5.14)$$

that is, σ_i^2 is a linear function of the Z's. If $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$, $\sigma_i^2 = \alpha_1$, which is a constant. Therefore, to test whether σ_i^2 is homoscedastic, one can test the hypothesis that $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$. This is the basic idea behind the Breusch–Pagan–Godfrey test.

The actual test procedure is as follows.

Step 1. Estimate Eq. (11.5.12) by OLS and obtain the residuals $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n$.

Step 2. Obtain $\tilde{\sigma}^2 = \sum \hat{u}_i^2/n$. Recall from Chapter 4 that this is the maximum likelihood (ML) estimator of σ^2. (*Note:* The OLS estimator is $\sum \hat{u}_i^2/[n-k]$.)

Breusch-Pagan-Godfrey Test

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

$$p_i = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi} + v_i \qquad (11.5.15)$$

where v_i is the residual term of this regression.

Step 5. Obtain the ESS (explained sum of squares) from Eq. (11.5.15) and define

$$\Theta = \frac{1}{2}(\text{ESS}) \qquad (11.5.16)$$

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

$$\Theta \underset{\text{asy}}{\sim} \chi^2_{m-1} \qquad (11.5.17)$$

that is, Θ follows the chi-square distribution with $(m - 1)$ degrees of freedom. (*Note:* asy means asymptotically.)

Therefore, if in an application the computed Θ $(= \chi^2)$ 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.

Example 11.5: The Breusch-Pagan-Godfrey Test

As an example, let us revisit the data (Table 11.3) that were used to illustrate the Goldfeld–Quandt heteroscedasticity test. Regressing Y on X, we obtain the following:

Step 1.

$$\hat{Y}_i = 9.2903 + 0.6378X_i$$
$$\text{se} = (5.2314) \quad (0.0286) \qquad \text{RSS} = 2361.153 \qquad R^2 = 0.9466 \qquad \textbf{(11.5.18)}$$

Step 2.

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

Step 3. Divide the squared residuals $\hat{u}_i$ obtained from regression (11.5.18) by 78.7051 to construct the variable p_i.

Step 4. Assuming that p_i are linearly related to $X_i (= Z_i)$ as per Eq. (11.5.14), we obtain the regression

$$\hat{p}_i = -0.7426 + 0.0101X_i$$
$$\text{se} = (0.7529) \quad (0.0041) \qquad \text{ESS} = 10.4280 \qquad R^2 = 0.18 \qquad \textbf{(11.5.19)}$$

Step 5.

$$\Theta = \frac{1}{2}(\text{ESS}) = 5.2140 \qquad \textbf{(11.5.20)}$$

Under the assumptions of the BPG test Θ in Eq. (11.5.20) asymptotically follows the chi-square distribution with 1 df. (*Note:* There is only one regressor in Eq. [11.5.19].) Now from the chi-square table we find that for 1 df the 5 percent critical chi-square value is 3.8414 and the 1 percent critical χ^2 value 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. Therefore, we reach the same conclusion as the Goldfeld–Quandt test. But keep in mind that, strictly speaking, 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. Of course, we can test the normality assumption by the tests discussed in Chapter 5.[23]

White's General Heterosced-asticity Test

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad (11.5.21)$$

The White test proceeds as follows:

Step 1. Given the data, we estimate Eq. (11.5.21) and obtain the residuals, $\hat{u}_i$.

Step 2. We then run the following (*auxiliary*) regression:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + v_i \qquad (11.5.22)^{25}$$

That is, the squared residuals from the original regression are regressed on the original X variables or regressors, their squared values, and the cross product(s) of the regressors. Higher powers of regressors can also be introduced. Note that there is a constant term in this equation even though the original regression may or may not contain it. Obtain the R^2 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^2 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,

$$n \cdot R^2 \underset{\text{asy}}{\sim} \chi^2_{\text{df}} \qquad (11.5.23)$$

where df is as defined previously. In our example, there are 5 df since there are 5 regressors in the auxiliary regression.

Step 4. If the chi-square value obtained in Eq. (11.5.23) exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity. If it does not exceed the critical chi-square value, there is no heteroscedasticity, which is to say that in the auxiliary regression (11.5.22), $\alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = 0$ (see footnote 25).

Example 11.6: White's General Heteroscedasticity Test

From cross-sectional data on 41 countries, Stephen Lewis estimated the following regression model:[26]

$$\ln Y_i = \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i \qquad \textbf{(11.5.24)}$$

where Y = ratio of trade taxes (import and export taxes) to total government revenue, X_2 = ratio of the sum of exports plus imports to GNP, and X_3 = GNP per capita; and ln stands for natural log. His hypotheses were that Y and X_2 would be positively related (the higher the trade volume, the higher the trade tax revenue) and that Y and X_3 would be negatively related (as income increases, government finds it is easier to collect direct taxes—e.g., income tax—than it is to rely on trade taxes).

The empirical results supported the hypotheses. For our purpose, the important point is whether there is heteroscedasticity in the data. Since the data are cross-sectional involving a heterogeneity of countries, a priori one would expect heteroscedasticity in the error variance. By applying White's heteroscedasticity test to the residuals obtained from regression (11.5.24), the following results were obtained:[27]

$$\begin{aligned}\widehat{u_i^2} = {} & -5.8417 + 2.5629 \ln \text{Trade}_i + 0.6918 \ln \text{GNP}_i \\ & -0.4081(\ln \text{Trade}_i)^2 - 0.0491(\ln \text{GNP}_i)^2 \\ & +0.0015(\ln \text{Trade}_i)(\ln \text{GNP}_i) \qquad R^2 = 0.1148\end{aligned} \qquad \textbf{(11.5.25)}$$

Note: The standard errors are not given, as they are not pertinent for our purpose here.

Now $n \cdot R^2 = 41(0.1148) = 4.7068$, which has, asymptotically, a chi-square distribution with 5 df (why?). The 5 percent critical chi-square value for 5 df is 11.0705, the 10 percent critical value is 9.2363, and the 25 percent critical value is 6.62568. For all practical purposes, one can conclude, on the basis of the White test, that there is no heteroscedasticity.

The Model of Generalized Least Squares(GLS)

now-familiar two-variable model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad (11.3.1)$$

which for ease of algebraic manipulation we write as

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + u_i \qquad (11.3.2)$$

where $X_{0i} = 1$ for each i. The reader can see that these two formulations are identical.

Now assume that the heteroscedastic variances σ_i^2 are *known*. Divide Eq. (11.3.2) through by σ_i to obtain

$$\frac{Y_i}{\sigma_i} = \beta_1 \left(\frac{X_{0i}}{\sigma_i}\right) + \beta_2 \left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \tag{11.3.3}$$

which for ease of exposition we write as

$$Y_i^* = \beta_1^* X_{0i}^* + \beta_2^* X_i^* + u_i^* \tag{11.3.4}$$

where the starred, or transformed, variables are the original variables divided by (the known) σ_i. We use the notation β_1^* and β_2^*, the parameters of the transformed model, to distinguish them from the usual OLS parameters β_1 and β_2.

What is the purpose of transforming the original model? To see this, notice the following feature of the transformed error term u_i^*:

$$\begin{aligned}
\operatorname{var}(u_i^*) = E(u_i^*)^2 &= E\left(\frac{u_i}{\sigma_i}\right)^2 && \text{since } E(u_i^*) = 0 \\
&= \frac{1}{\sigma_i^2} E\left(u_i^2\right) && \text{since } \sigma_i^2 \text{ is known} \qquad (11.3.5) \\
&= \frac{1}{\sigma_i^2}\left(\sigma_i^2\right) && \text{since } E\left(u_i^2\right) = \sigma_i^2 \\
&= 1
\end{aligned}$$

GLS Estimators: The Actual Mechanics of estimating β_1 and β_2

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

or

$$Y_i^* = \hat{\beta}_1^* X_{0i}^* + \hat{\beta}_2^* X_i^* + \hat{u}_i^* \tag{11.3.6}$$

Now, to obtain the GLS estimators, we minimize

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

that is,

$$\sum \left(\frac{\hat{u}_i}{\sigma_i}\right)^2 = \sum \left[\left(\frac{Y_i}{\sigma_i}\right) - \hat{\beta}_1^* \left(\frac{X_{0i}}{\sigma_i}\right) - \hat{\beta}_2^* \left(\frac{X_i}{\sigma_i}\right)\right]^2 \tag{11.3.7}$$

The actual mechanics of minimizing Eq. (11.3.7) follow the standard calculus techniques and are given in Appendix 11A, Section 11A.2. As shown there, the GLS estimator of β_2^* is

$$\hat{\beta}_2^* = \frac{(\sum w_i)(\sum w_i X_i Y_i) - (\sum w_i X_i)(\sum w_i Y_i)}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{11.3.8}$$

and its variance is given by

$$\text{var}\,(\hat{\beta}_2^*) = \frac{\sum w_i}{(\sum w_i)(\sum w_i X_i^2) - (\sum w_i X_i)^2} \tag{11.3.9}$$

where $w_i = 1/\sigma_i^2$.

Difference between OLS and GLS

Recall from Chapter 3 that in OLS we minimize

$$\sum \hat{u}_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \tag{11.3.10}$$

but in GLS we minimize the expression (11.3.7), which can also be written as

$$\sum w_i \hat{u}_i^2 = \sum w_i (Y_i - \hat{\beta}_1^* X_{0i} - \hat{\beta}_2^* X_i)^2 \tag{11.3.11}$$

where $w_i = 1/\sigma_i^2$ (verify that Eq. [11.3.11] and Eq. [11.3.7] are identical).

Example 11.7: Illustration of the Method of Weighted Least Squares

To illustrate the method, suppose we want to study the relationship between compensation and employment size for the data presented in Table 11.1. For simplicity, we measure employment size by 1 (1–4 employees), 2 (5–9 employees), . . . , 9 (1,000–2499 employees), although we could also measure it by the midpoint of the various employment classes given in the table.

Now letting Y represent average compensation per employee ($) and X the employment size, we run the following regression (see Eq. [11.3.6]):

$$Y_i/\sigma_i = \hat{\beta}_1^*(1/\sigma_i) + \hat{\beta}_2^*(X_i/\sigma_i) + (\hat{u}_i/\sigma_i) \qquad \textbf{(11.6.1)}$$

where σ_i are the standard deviations of wages as reported in Table 11.1. The necessary raw data to run this regression are given in Table 11.4.

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).

Example 11.7: Illustration of the Method of Weighted Least Squares (Cont')

Before going on to the regression results, note that Eq. (11.6.1) has no intercept term. (Why?) Therefore, one will have to use the regression-through-the-origin model to estimate β_1^* and β_2^*, a topic discussed in Chapter 6. But most computer packages these days have an option to suppress the intercept term (see *Minitab* or *EViews*, for example). Also note another interesting feature of Eq. (11.6.1): It has two explanatory variables, $(1/\sigma_i)$ and (X_i/σ_i), whereas if we were to use OLS, regressing compensation on employment size, that regression would have a single explanatory variable, X_i. (Why?)

The regression results of WLS are as follows:

$$\widehat{(Y_i/\sigma_i)} = 3406.639(1/\sigma_i) + 154.153(X_i/\sigma_i) \tag{11.6.2}$$
$$\text{se} = (80.983) \quad (16.959)$$
$$t = (42.066) \quad (9.090)$$
$$R^2 = 0.9993^{33}$$

For comparison, we give the usual or unweighted OLS regression results:

$$\hat{Y}_i = 3417.833 + 148.767 X_i \tag{11.6.3}$$
$$\text{se} = (81.136) \quad (14.418)$$
$$t = (42.125) \quad (10.318) \qquad R^2 = 0.9383$$

In Exercise 11.7 you are asked to compare these two regressions.

The heteroscedasticity-robust standard error of $\beta_j\hat{}$

- Since ESS/(n-2) is not an unbiased estimator of σ^2 , then how to obtain se($\beta_j\hat{}$)?
- White (1980) came up with consistent estimate of se ($\beta_j\hat{}$) which allow performing statistical inference about the parameters when sample size is large.
- White's heteroscedaticity-corrected standard errors are also called robust standard errors or heteroscedaticity-consistent covariance matrix

The heteroscedasticity-robust standard error of β_j^

$$\widehat{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ji}^2 \hat{u}_i^2}{RSS_j^2}$$

- $r\hat{}_{ji}$ is the residuals of regressing X_j on all other independent variables and RSS_j is residual sum of squares of this regression.
- Square root of the estimated variance of beta is se(β^)
- With large sample size, t-test is still valid

$$t = \frac{\hat{\beta}_j - \beta_{j,under H_0}}{robust - se(\hat{\beta}_j)}$$

When σ^2 is Not Known: White's Heteroscedasticity-Consistent Variances and Standard Errors

Example 11.8: Illustration of White's Procedure

Yi is per capita expenditure on public school, and income is per capita income by state

$$\hat{Y}_i = 832.91 - 1834.2\ \text{income} + 1587.04\ \text{income}^2$$

OLS se = (327.3) (829.0) (519.1)

White's se = (460.9) (1243.0) (830.0)

- Using OLS se, both income and income^2 are statistical significant, but base on White's se, they are not significant.
- Conclusion: If the White's se is available in the computer package used, should always report these value along with OLS se

When σ^2 is Not Known: Plausible Assumptions about Heteroscedasticity Pattern

Assumption 1

The error variance is proportional to X_i^2:

$$E(u_i^2) = \sigma^2 X_i^2 \qquad (11.6.5)^{38}$$

Assumption 2

The error variance is proportional to X_i. The **square root transformation:**

$$E(u_i^2) = \sigma^2 X_i \qquad (11.6.7)$$

Assumption 3

The error variance is proportional to the square of the mean valu (11.6.9)

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

Assumption 4

A log transformation such as

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \qquad (11.6.12)$$

very often reduces heteroscedasticity when compared with the regression $Y_i = \beta_1 + \beta_2 X_i + u_i$.