Chapter 8 Heteroskedasticity

Chapter Contents

- 8.1 The Nature of Heteroskedasticity
- 8.2 Heteroskedasticity in the Multiple Regression Model
- 8.3 Heteroskedasticity Robust Variance Estimator
- 8.4 Generalized Least Squares: Known Form of Variance
- 8.5 Generalized Least Squares: Unknown Form of Variance
- 8.6 Detecting Heteroskedasticity
- 8.7 Heteroskedasticity in the Linear Probability Model

8.1 The Nature of Heteroskedasticity 1 of 3

- Consider our basic linear function: $FOOD\_EXP_i = \beta_1 + \beta_2 INCOME_i + e_i$
- The random error e_i represents the collection of all the factors other than income that affect household expenditure on food
- The assumption of **strict exogeneity** says that when using information on household income our best prediction of the random error is zero
- If sample values are randomly selected, then the technical expression for this assumption is that given income the conditional expected value of the random error e_i is zero

8.1 The Nature of Heteroskedasticity 2 of 3

- If the assumption of strict exogeneity holds then the regression function is

$$E(FOOD\_EXP_i|INCOME_i) = \beta_1 + \beta_2 INCOME$$

- Holding income constant, and given our model, what is the source of the variation in household food expenditures? It must be from the random error
- Recall that the random error in the regression is the difference between any observation on the outcome variable and its conditional expectation, that is:

$$(8.2) \quad e_i = FOOD\_EXP_i - E(FOOD\_EXP_i|INCOME_i)$$

8.1 The Nature of Heteroskedasticity 3 of 3

- In such a case, when the error variances for all observations are not the same, we say that **heteroskedasticity** exists
  - Alternatively, we say the random error e_i is **heteroskedastic**
- Conversely, if all observations come from probability density functions with the same variance, we say that **homoskedasticity** exists, and e_i is **homoskedastic**

8.2 Heteroskedasticity in the Multiple Regression Model 1 of 2

- The existence of heteroskedasticity is a violation of one of our least squares assumptions
- For the multiple regression model $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{iK} + e_i, i = 1, \ldots, N$, assumption MR3 is $var(e_i|X) = var(y_i|X) = \sigma^2$
- The simplest statement of the conditional heteroskedasticity is

(8.3) $var(e_i|X) = var(y_i|X) = \sigma_i^2$

8.2 Heteroskedasticity in the Multiple Regression Model 2 of 2

- Heteroskedasticity often arises when using cross-sectional data
- The term cross-sectional data refers to having data on a number of economic units such as firms or households, at a *given point in time*
- Heteroskedasticity is not a property that is necessarily restricted to cross-sectional data
- With time series data it is possible that the conditional error variance will change

8.2.1 The Heteroskedastic Regression Model

- The multiple regression model is $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$
  - We assume we have a random sample
- The heteroskedasticity assumption in (8.3) becomes

(8.4) $\text{var}(y_i|x_i) = \text{var}(e_i|x_i) = \sigma^2 h(x_i) = \sigma_i^2$

- where $h(x_i)^0 > 0$ is a function of x_i that is sometimes called the skedastic function
- Where $\sigma^2 > 0$ is a constant

8.2.2 Heteroskedasticity Consequences for the OLS Estimator

- There are two implications of heteroskedacity

1. The least squares estimator is still a linear and unbiased estimator, but it is no longer best. There is another estimator with a smaller variance

2. The standard errors usually computed for the least squares estimator are incorrect. Confidence intervals and hypothesis tests that use these standard errors may be misleading

8.3 Heteroskedasticity Robust Variance Estimator 1 of 2

- Calculation of a correct estimate for the OLS variance
- (8.8) $\operatorname{var}(b_2|x) = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{-1} \sum_{i=1}^{N}\left[(x_i - \bar{x})^2 \sigma_i^2\right] \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{-1}$
- The **White heteroskedasticity-consistent estimator (HCE)** that is valid in large samples for the simple regression model is

(8.9) $\widehat{var}(b_2) = \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{-1} \left\{\sum \left[(x_i - \bar{x})^2 \left(\frac{N}{N-2}\right) \hat{e}_i^2\right]\right\} \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]^{-1}$

8.3 Heteroskedasticity Robust Variance Estimator 2 of 2

- Where $\hat{e}_i$ is the least squares residual from the regression model, $y_i = \beta_1 + \beta_2 x_i + e_i$
- This variance estimator is robust because it is valid whether heteroskedasticity is present or not
- If we are not sure whether the random errors are heteroskedastic or homoskedastic, then we can use a robust variance estimator and be confident that our standard errors, t-tests, and interval estimates are valid in large samples
- This does not address the implication that the least square estimator is no longer the best

8.4 Generalized Least Squares: Known Form of Variance

- To develop an estimator that is better than the least squares estimator, we need to make a further assumption about how the variances σ_i^2 change with each observation
- This means making an assumption about the skedastic function h(x_i)
- The further assumption is necessary because the best linear unbiased estimator in the presence of heteroskedasticity, an estimator known as the **generalized least squares (GLS) estimator**, depends on the unknown σ_i^2

8.4.1 Transforming the Model: Proportional Heteroskedasticity 1 of 5

- An estimator known as the **generalized least squares estimator**, depends on the unknown σ^2_i
- To make the generalized least squares estimator operational, some structure is imposed on σ^2_i
- One possibility:

$$(8.11)\ \text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 h(x_i) = \sigma^2 x_i, x_i > 0$$

8.4.1 Transforming the Model: Proportional Heteroskedasticity 2 of 5

- We change or transform the model into one with homoskedastic errors:
- Leaving the basic structure of the model intact, we turn the heteroskedastic error model into a homoskedastic error model
- After the transformation, applying OLS to the transformed model gives a best linear unbiased estimator

$$\frac{y_i}{\sqrt{x_i}} = \beta_1\left(\frac{1}{\sqrt{x_i}}\right) + \beta_2\left(\frac{x_i}{\sqrt{x_i}}\right) + \frac{e_i}{\sqrt{x_i}} \quad (8.12)$$

8.4.1 Transforming the Model: Proportional Heteroskedasticity 3 of 5

- Define the following transformed variables:
- (8.13) $y_i^* = \frac{y_i}{\sqrt{x_i}}, \quad x_{i1}^* = \frac{1}{\sqrt{x_i}}, \quad x_{i2}^* = \frac{x_i}{\sqrt{x_i}} = \sqrt{x_i}, \quad e_i^* = \frac{e_i}{\sqrt{x_i}}$
  - Our model is now

(8.14) $y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + e_i^*$

- The beauty of this transformed model is that the new transformed error term e∗ i is homoskedastic

8.4.1 Transforming the Model: Proportional Heteroskedasticity 4 of 5

- If X is a random variable and a is a constant, then var(aX) = a^2var(X). Applying that rule here we have

$$(8.15)\quad \text{var}(e_i^*) = \text{var}\left(\frac{e_i}{\sqrt{x_i}}\right) = \frac{1}{x_i}\text{var}(e_i) = \frac{1}{x_i}\sigma^2 x_i = \sigma^2$$

- The transformed error term will retain the properties of zero mean and zero correlation between different observations

8.4.1 Transforming the Model: Proportional Heteroskedasticity 5 of 5

- To obtain the best linear unbiased estimator for a model with heteroskedasticity of the type specified in (8.11):

  1. Calculate the transformed variables given in (8.13)
  2. Use OLS to estimate the transformed model given in (8.14), yielding estimates $\widehat{\beta_1}$ and $\widehat{\beta_2}$

- The estimator obtained in this way is called a generalized least squares estimator

8.4.2 Weighted Least Squares: Proportional Heteroskedasticity

- One way of viewing the generalized least squares estimator is as a **weighted least squares** estimator

- Minimizing the sum of squared transformed errors:

$$\sum_{i=1}^{n} \frac{(y_i - \beta_1 - \beta_2 x_{i2})^2}{x_i}$$

- The squared errors are weighted by $x_i^{-1/2}$

Example 8.3 Applying GLS/WLS to the Food Expenditure Data

- Applying the generalized (weighted) least squares procedure to our food expenditure problem:

8.17
$$\widehat{FOOD\_EXP}_i = 78.68 + 10.45 INCOME_i$$
$$(se) \qquad (23.79) \ (1.39)$$

- A 95% confidence interval for β_2 is given by:

$$\hat{\beta}_2 \pm t_c \text{se}(\hat{\beta}_2) = 10.451 \pm 2.024 \times 1.386 = [7.65, 13.26]$$

8.5 Generalized Least Squares: Unknown Form of Variance 1 of 4

- In order to deal with the more general specification we need a model that is flexible, parsimonious, and for which $\sigma_i^2 > 0$
- One specification that works well is:

$$\begin{aligned} \sigma_i^2 &= \exp(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \exp(\alpha_1)\exp(\alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ &= \sigma^2 h(z_{i2}, \cdots, z_{iS}) \end{aligned} \quad (8.18)$$

8.5 Generalized Least Squares: Unknown Form of Variance 2 of 4

- A Equation (8.18) is called the model of **multiplicative heteroskedasticity**

- It includes homoskedasticity as a special case; when α2 =···=αS = 0 the error variance is $\sigma_i^2 = \exp(\alpha_1) = \sigma^2$

- It is called a multiplicative model because

$$\exp(\alpha_1)\exp(\alpha_2 z_{i2} + \cdots + \alpha_S z_{iS}) \\ = \exp(\alpha_1)\exp(\alpha_2 z_{i2})\ldots\exp(\alpha_s z_{is})$$

8.5 Generalized Least Squares: Unknown Form of Variance 3 of 4

- **Multiplicative Heteroskedasticity, Special Case 1:** $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha 2}$
- There are three plausible variance functions, they are a special cases of
  - $\text{var}(e_i|x_i) = \sigma_i^2 = \sigma^2 x_i^{\alpha 2}$
  - Where $\propto_2$ is an unknown parameter

8.5 Generalized Least Squares: Unknown Form of Variance 4 of 4

- **Multiplicative Heteroskedasticity, Special Case 2: Grouped Heteroskedasticity**
- Suppose we are considering just two groups
- $D_i = 1$ if an observation is in one group and $D_i = 0$ for observations in the other group, Then the variance function is:

$$\text{var}(e_i|x_i) = \exp(\alpha_1 + \alpha_1 D_i) = \begin{cases} \exp(\alpha_1) = \sigma^2 & D_i = 0 \\ \exp(\alpha_1 + \alpha_2) = \sigma^2 \exp(\alpha 2) & D_i = 1 \end{cases}$$

8.5.1 Estimating the Multiplicative Model 1 of 2

- How do we proceed with estimation with an assumption like (8.18)
- With the model of multiplicative heteroskedasticity, we use several estimation steps
- **FEASIBLE GLS PROCEDURE**

1. Estimate the original model by OLS, saving the OLS residuals $\hat{e}_i$
2. Use the least squares residuals and the variables $z_{i2}, \cdots, z_{iS}$ to estimate

   $\propto_1, \propto_2 \cdots, \propto_S$

8.5.1 Estimating the Multiplicative Model 2 of 2

3. Calculate the estimated skedastic function $\hat{h}\,(z_{i2}, \cdots, z_{iS})$

4. . Divide each observation by $\sqrt{\hat{h}\,(z_{i2}, \cdots, z_{iS})}$ and apply OLS to the transformed data, or use WLS regression with weighting factor $1/\,\hat{h}\,(z_{i2}, \cdots, z_{iS})$

- The resulting estimates are called **feasible generalized least squares (FGLS) estimates** or **estimated generalized least squares (EGLS)** estimates

8.6 Detecting Heteroskedasticity

- In many applications, there is uncertainty about the presence, or absence, of heteroscedasticity
- There are two methods we can use to detect heteroskedasticity
  1. An informal way using residual charts
  2. A formal way using statistical tests

8.6.1 Residual Plots 1 of 2

- If the errors are homoskedastic, there should be no patterns of any sort in the residuals
- If the errors are heteroskedastic, they may tend to exhibit greater variation in some systematic way
- We discovered that the absolute values of the residuals do indeed tend to increase as income increases

8.6.1 Residual Plots 2 of 2

- This method of investigating heteroskedasticity can be followed for any simple regression
- In a regression with more than one explanatory variable we can plot the least squares residuals against each explanatory variable, or against $\hat{y}_i$, to see if they vary in a systematic way relative to the specified variable

8.6.2 The Goldfeld–Quandt Test

- The **Goldfeld–Quandt** test uses the estimated error variances from separate sub-sample regressions as a basis for the test
- Let the first sub-sample contain N_1 observations
- Let the regression odel in this partition have K_1 parameters
- The test statistic is 8.22 $$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(N_M - K_1, N_2 - K_2)}$$

Example 8.7 The Goldfeld–Quandt Test in the Food Expenditure Model 1 of 2

- With the observations ordered according to income x_i, and the sample split into two equal groups of 20 observations each, yields:

$$\hat{\sigma}_1^2 = 3574.8 \qquad \hat{\sigma}_2^2 = 12921.9$$

- Calculate:

$$F = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_1^2} = \frac{12921.9}{3574.8} = 3.61$$

Example 8.7 The Goldfeld–Quandt Test in the Food Expenditure Model 2 of 2

- Believing that the variances could increase, but not decrease with income, we use a one-tail test with 5% critical value $F(0.95, 18, 18) = 2.22$
- Since 3.61 > 2.22, a null hypothesis of homoskedasticity is rejected in favor of the alternative that the variance increases with income

8.6.3 A General Test for Conditional Heteroskedasticity 1 of 5

- In this section we consider a test for conditional heteroskedasticity that is related to some "explanatory" variables
- Under assumptions MR1–MR5 the OLS estimator is the best linear unbiased estimator of the parameters $\beta_1, \beta_2, \ldots \beta_k,$
- When conditional heteroskedasticity is a possibility, we suppose that the variance of the random error, e_i , depends on a set of explanatory variables $z_{i2}, z_{i3}, \ldots z_{ik}$ that may include some or all of the explanatory variables $x_{i2}, \ldots x_{ik}$

8.6.3 A General Test for Conditional Heteroskedasticity 2 of 5

- Assume a general expression for the conditional variance

$$\text{var}(e_i|z_i) = \sigma_i^2 = E(e_i^2|z_i) = h(\alpha_1 + \alpha_2 z_{i2} + \cdots + \alpha_s z_{is})$$

- Where h(•) is some smooth function and $\alpha_2, \alpha_3, \ldots, \alpha_s$ are nuisance parameters
- We will test for any relationship between the variance of the error term and any function of the selected variables

8.6.3 A General Test for Conditional Heteroskedasticity 3 of 5

- The null and alternative hypotheses for a test for heteroskedasticity based on the variance function are
- homoskedasticity ↔ $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_S = 0$
- heteroskedasticity ↔ H_1 : not all the α_s in H_0 are zero
- if the random errors are homoskedastic, then the sample size multiplied by R^2, N× R^2 or simply NR^2, has a chi-square ($\chi2$) distribution with S − 1 degrees of freedom

8.6.3 A General Test for Conditional Heteroskedasticity 4 of 5

- The test statistic is:
  - 8.30 $NR^2 \overset{a}{\sim} \chi^2_{(S-1)}$ if the null hypothesis of homoskedasticity is true
- There are several important features of this test:

1. It is a large sample test. The result in (8.30) holds approximately in large samples
2. You will often see the test referred to as a Lagrange multiplier test (LM test) or a Breusch–Pagan test for heteroskedasticity

8.6.3 A General Test for Conditional Heteroskedasticity 5 of 5

3. One of the amazing features of the Breusch–Pagan/LM test is that the value of the statistic computed from the linear function is valid for testing an alternative hypothesis of heteroskedasticity where the variance function can be of any form given by (8.24)

4. The Breusch–Pagan test is for conditional heteroskedasticityUnconditional heteroskedasticity exists when the error term variance is completely random

8.6.4 The White Test 1 of 3

- The variance tests used so far presupposes we have knowledge of what variables will appear in the variance function if the alternative hypothesis of heteroskedasticity is true
- We may wish to test for heteroskedasticity without precise knowledge of the relevant variables

8.6.4 The White Test 2 of 3

- Suppose:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

- The White test without cross-product terms (interactions) specifies:

$$z_2 = x_2 \quad z_3 = x_3 \quad z_4 = x_2^2 \quad z_5 = x_3^2$$

- Including interactions adds one further variable

$$z_5 = x_2 x_3$$

8.6.4 The White Test 3 of 3

- The White test is performed using the NR^2 test defined in (8.29)
- or an F-test
- One difficulty with the White test is that it can detect problems other than heteroskedasticity
- Thus, while it is a useful diagnostic, be careful about interpreting the result of a significant White test

8.6.5 Model Specification and Heteroskedasticity 1 of 4

- As hinted at the end of the previous section, heteroskedasticity can be present because of a model specification error
- If data partitions are not recognized, or important variables omitted, or an incorrect functional form selected, then heteroskedasticity can appear to be present
- Don’t necessarily believe that a significant heteroskedasticity test means that heteroskedasticity is the problem and that using robust standard errors will be an adequate fix

8.6.5 Model Specification and Heteroskedasticity 2 of 4

- Critically examine the model from the point of view of economic reasoning and look for any specification problems
- One very common specification issue with economic data is the choice of functional form
- Using a logarithmic transformation of the dependent variable has another feature, **variance stabilization,** that is useful in the context of heteroskedastic data

8.6.5 Model Specification and Heteroskedasticity 3 of 4

- Economic variables like wages, incomes, house prices, and expenditures are right-skewed, with a long tail to the right
- The log-normal probability distribution is useful when modeling such variables
- If the random variable y has a log-normal probability density function, then ln(y) has a normal distribution, which is symmetrical and bell-shaped, and not skewed

8.6.5 Model Specification and Heteroskedasticity 4 of 4

- The feature of the log-normal random variable that we are now interested in is that its variance increases when its mean and median increase
- By choosing a log-linear or log-log model we are implicitly assuming a curvilinear and heteroskedastic relationship between the variables y and x
- However, there is a linear and homoskedastic relation between ln(y) and x

Figure 8.6 A log-linear relationship

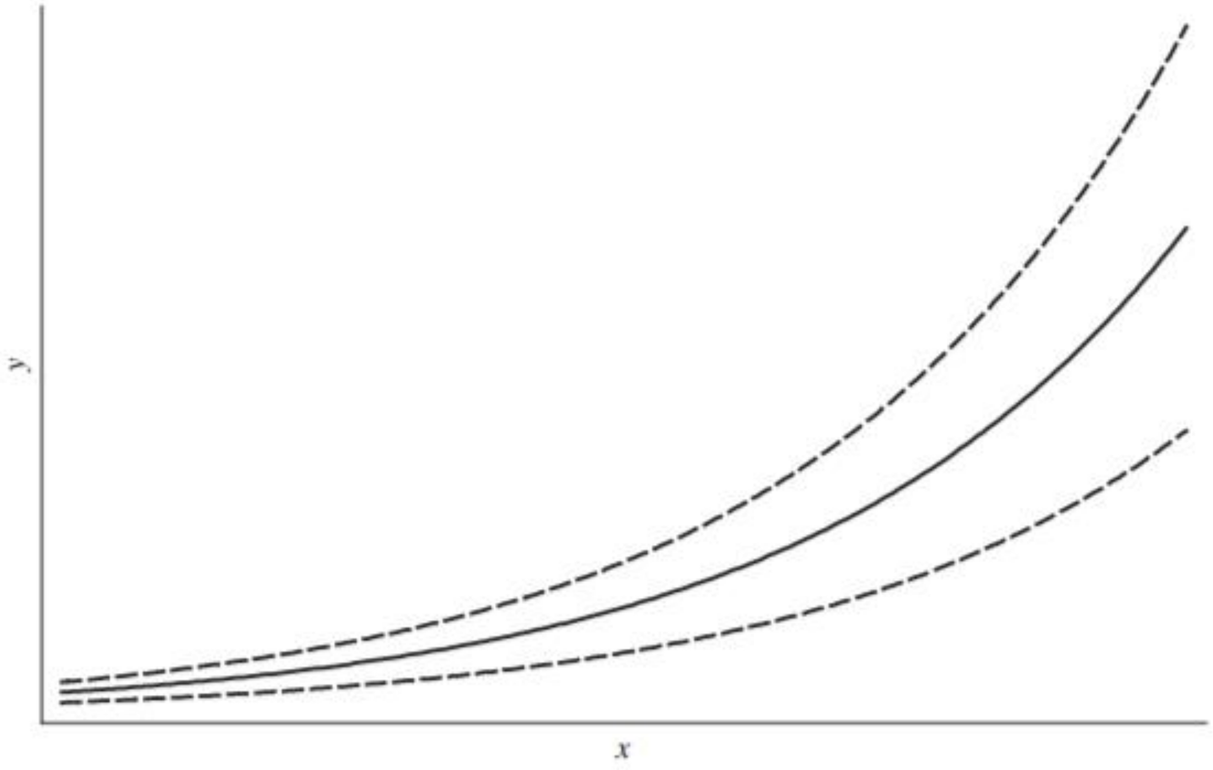

FIGURE 8.6 A log-linear relationship.

Example 8.8 Variance Stabilizing Log-transformation 1 of 3

- Consider the data file cex5_small
- Figure 8.7(a) shows a histogram of household expenditures on entertainment per person, ENTERT, for those households who have positive spending, and Figure 8.7(b) is the histogram for ln(ENTERT)
- The extremely skewed distribution of entertainment expenditures shows the effect of the log-transformation

Example 8.8 Variance Stabilizing Log-transformation 2 of 3

- The variation in ENTERT about the fitted line increases as INCOME increases
- Estimating the model ENTERT = β_1 + β_2INCOME + β_3COLLEGE + β_4ADVANCED + e
- Obtain the least squares residuals and then estimate by OLS the model

$$\widehat{e_i^2} = \alpha_1 + \alpha_2 INCOME_i + v_i$$

Example 8.8 Variance Stabilizing Log-transformation 3 of 3

- From this regression, $NR^2 = 31.34$. The critical value for a 1% level of significance
- heteroskedasticity test is 6.635, thus we conclude that heteroskedasticity is present
- There is little if any visual evidence of heteroskedasticity and the value of the heteroskedasticity test statistic is $NR^2 = 0.36$
- so we do not reject the null hypothesis of homoskedasticity. The log-transformation has “cured” the heteroskedasticity problem

Figure 8.7 Histograms of entertainment expenditures

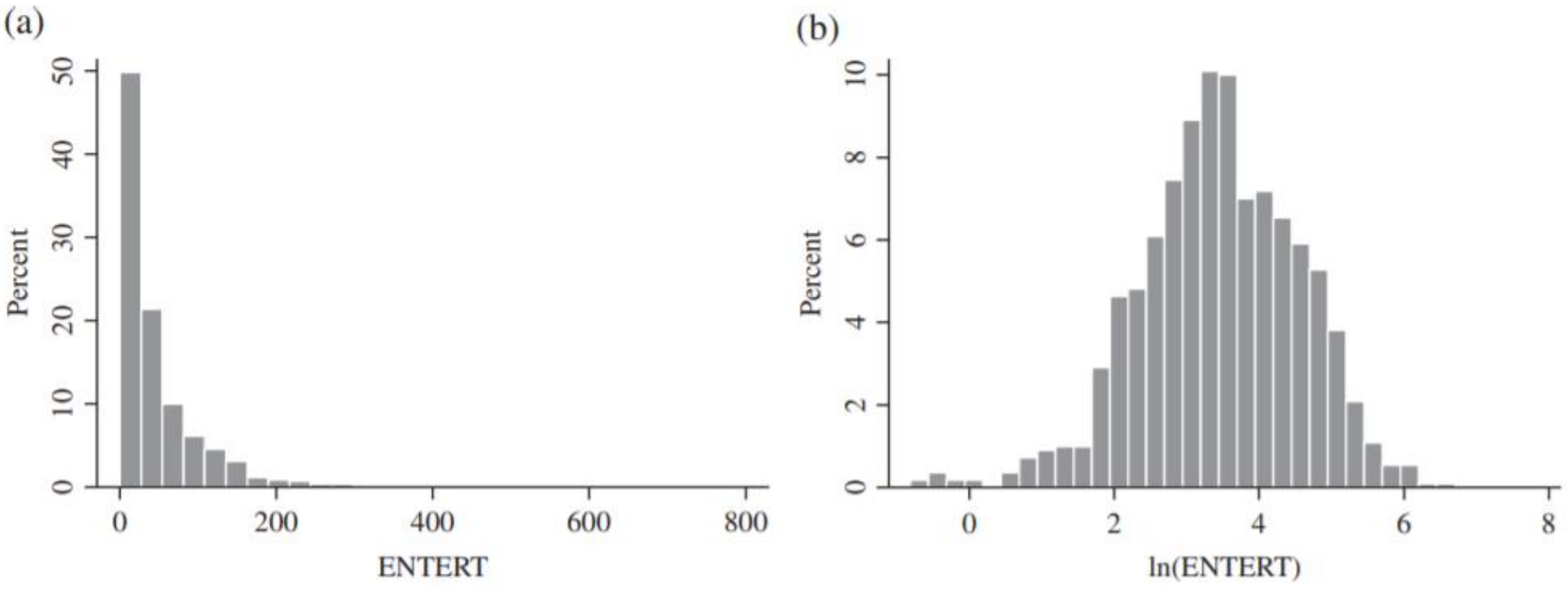

FIGURE 8.7 Histograms of entertainment expenditures.

Figure 8.8 Linear and log-linear models for entertainment expenditures.

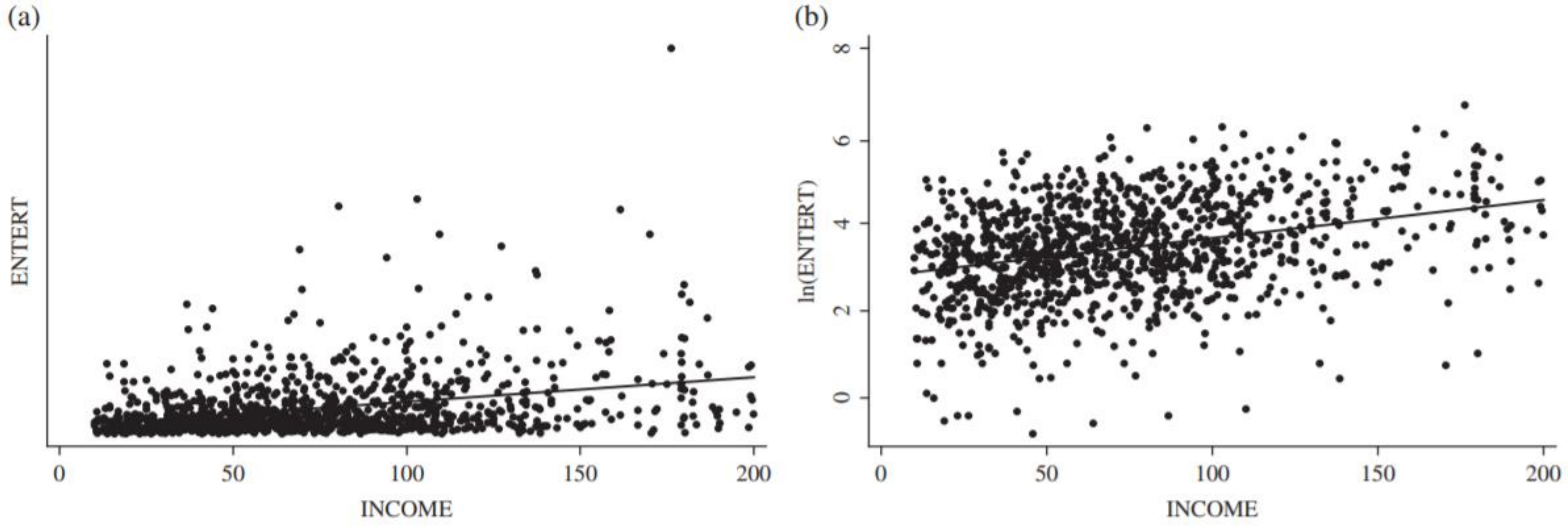

FIGURE 8.8 Linear and log-linear models for entertainment expenditures.

8.7 Heteroskedasticity in the Linear Probability Model 1 of 4

- We previously examined the linear probability model:

$$E(y_i|x_i) = p = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$$

- Defining the error ei as the difference $y_i - E(y_i|x_i)$ for the ith observation, we have the model

- (8.32) $y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i$

8.7 Heteroskedasticity in the Linear Probability Model 2 of 4

- This model can be estimated with least squares
- We know this model suffers from heteroskedasticity:
- (8.32) $$\operatorname{var}(y_i|x_i) = \operatorname{var}(e_i) = p_i(1 - p_i) = (\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK})(1 - \beta_1 - \beta_2 x_{i2} - \cdots - \beta_K x_{iK})$$
- The error variance depends on the values of the explanatory variables
- Instead of using least squares standard errors, we can use heteroskedasticity-robust standard errors

8.7 Heteroskedasticity in the Linear Probability Model 3 of 4

- The first step toward obtaining GLS estimates is to estimate the variance in (8.32)

- We can estimate p_i with the least squares predictions:

(8.33) $$\hat{p}_i = b_1 + b_2 x_{i2} + \cdots + b_K x_{iK}$$

- Giving an estimated variance of:

(8.34) $$\widehat{\text{var}(e_i|x)} = \hat{p}_i(1 - \hat{p}_i)$$

8.7 Heteroskedasticity in the Linear Probability Model 4 of 4

- A word of caution is required at this point. It is possible that some of the $\hat{p}_i$ obtained from (8.33) will not lie within the interval $0 < \hat{p}_i < 1$

- Generalized least squares estimates can be obtained by applying least squares to the transformed equation:

$$\frac{y_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \beta_1 \frac{1}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \beta_2 \frac{x_{i2}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \cdots + \beta_K \frac{x_{iK}}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

Example 8.9 The Marketing Example Revisited 1 of 2

TABLE 8.2 **Linear Probability Model Estimates**

	LS	LS-robust	GLS-trunc	GLS-omit
C	0.8902	0.8902	0.6505	0.8795
	(0.0655)	(0.0652)	(0.0568)	(0.0594)
PRATIO	−0.4009	−0.4009	−0.1652	−0.3859
	(0.0613)	(0.0603)	(0.0444)	(0.0527)
DISP_COKE	0.0772	0.0772	0.0940	0.0760
	(0.0344)	(0.0339)	(0.0399)	(0.0353)
DISP_PEPSI	−0.1657	−0.1657	−0.1314	−0.1587
	(0.0356)	(0.0343)	(0.0354)	(0.0360)

Example 8.9 The Marketing Example Revisited 2 of 2

- A suitable test for heteroskedasticity is the White test:

$$\chi^2 = N \times R^2 = 25.17 \qquad p\text{-value} = 0.0005$$

- This leads us to reject a null hypothesis of homoskedasticity at a 1% level of significance.

- Examining the estimates in Table 8.2, we see there is little difference in the four sets of standard errors

Key Words

- Breusch–Pagan test
- generalized least squares
- Goldfeld–Quandt test
- grouped heteroskadasticity
- heteroskedasticity
- Heteroskedasticity-consistent standard errors
- homoskedasticity
- Lagrange multiplier test
- linear probability model
- mean function
- residual plot
- robust standard errors
- skedastic function
- transformed model
- variance function
- weighted least squares
- White test

Copyright