Chapter 4
Prediction, Goodness-of-fit, and Modeling Issues

Walter R. Paczkowski
Rutgers University

Chapter Contents

4.1
Least Squares Prediction

- The ability to predict is important to:
  - business economists and financial analysts who attempt to forecast the sales and revenues of specific firms
  - government policy makers who attempt to predict the rates of growth in national income, inflation, investment, saving, social insurance program expenditures, and tax revenues
  - local businesses who need to have predictions of growth in neighborhood populations and income so that they may expand or contract their provision of services
- Accurate predictions provide a basis for better decision making in every type of planning context

- In order to use regression analysis as a basis for prediction, we must assume that y_0 and x_0 are related to one another by the same regression model that describes our sample of data, so that, in particular, SR1 holds for these observations

Eq. 4.1

$$y_0 = \beta_1 + \beta_2 x_0 + e_0$$

where e_0 is a random error.

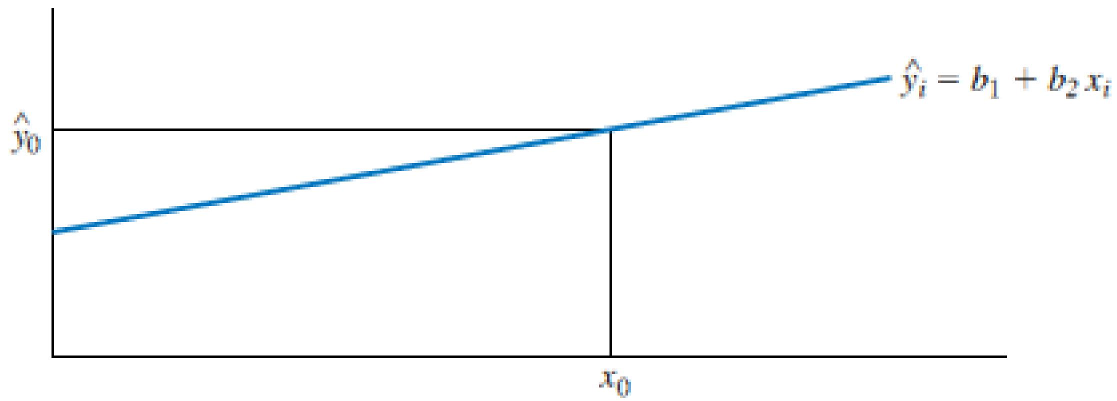
- The task of predicting y_0 is related to the problem of estimating $E(y_0) = \beta_1 + \beta_2 x_0$
  - Although $E(y_0) = \beta_1 + \beta_2 x_0$ is not random, the outcome y_0 is random
  - Consequently, there is a difference between the **interval estimate** (a confidence interval) of $\mathrm{E}(y_0) = \beta_1 + \beta_2 x_0$ and the **prediction interval** for y_0

- The least squares predictor of $E(y_0)$ comes from the fitted regression line

Eq. 4.2

$$\hat{y}_0 = b_1 + b_2 x_0$$

Figure 4.1 A point prediction



- To evaluate how well this predictor performs, we define the forecast error, which is analogous to the least squares residual:

Eq. 4.3

$$f = y_0 - \hat{y}_0 = \left(\beta_1 + \beta_2 x_0 + e_0\right) - \left(\mathrm{b}_1 + \mathrm{b}_2 x_0\right)$$

  – We would like the forecast error to be small, implying that our forecast is close to the value we are predicting

- Taking the expected value of f, we find that

$$E(f) = \beta_1 + \beta_2 x_0 + E(e_0) - \left[E(b_1) + E(b_2)x_0\right]$$
$$= \beta_1 + \beta_2 x_0 + 0 - [\beta_1 + \beta_2 x_0]$$
$$= 0$$

which means, on average, the forecast error is zero and $\hat{y}_0$ is an **unbiased predictor** of $E(y_0)$.

- However, unbiasedness does not necessarily imply that a particular forecast will be close to the actual value
  - $\hat{y}_0$ is the **best linear unbiased predictor** (*BLUP*) of $E(y_0)$ if assumptions SR1–SR5 hold

Eq. 4.4

- The variance of the forecast is

$$\operatorname{var}(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

- The variance of the forecast is smaller when:
  - the overall uncertainty in the model is smaller, as measured by the variance of the random errors σ^2
  - the sample size N is larger
  - the variation in the explanatory variable is larger
  - the value of $(x_0 - \bar{x})^2$ is small

- In practice we use

$$\widehat{\operatorname{var}(f)} = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

  for the variance

- The **standard error of the forecast** is:

Eq. 4.5

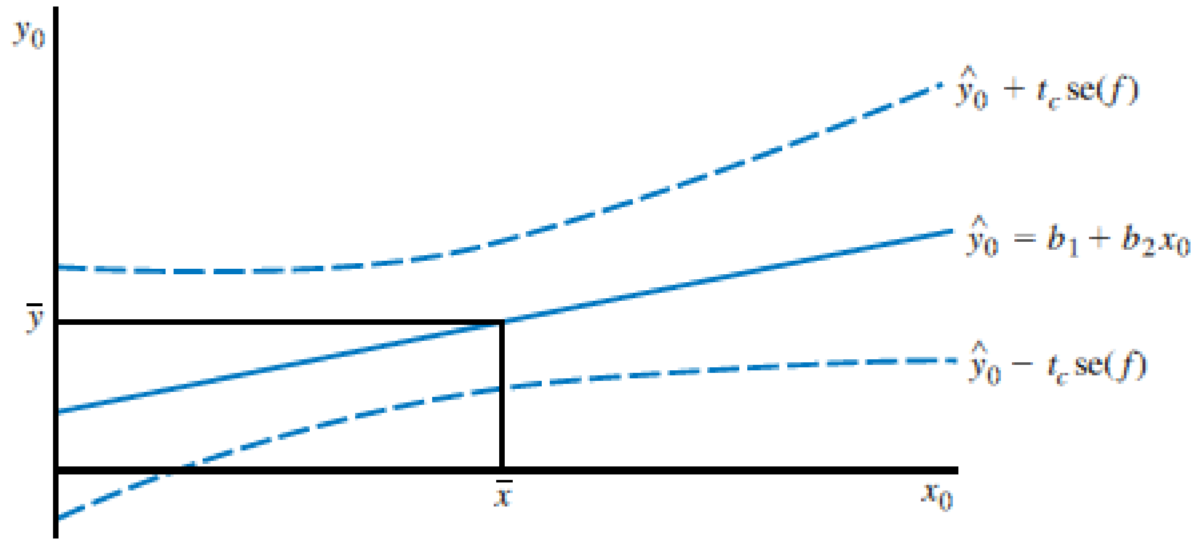
$$\operatorname{se}(f) = \sqrt{\widehat{\operatorname{var}(f)}}$$

- The $100(1 - \alpha)\%$ **prediction interval** of y_0 is:

Eq. 4.6

$$\hat{y}_0 \pm t_c \mathrm{se}\left(f\right)$$

Figure 4.2 Point and interval prediction



- For our food expenditure problem, we have:

$$\hat{y}_0 = b_1 + b_2 x_0 = 83.4160 + 10.2096(20) = 287.6089$$

- The estimated variance for the forecast error is:

$$\widehat{\operatorname{var}(f)} = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

$$= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2 \widehat{\operatorname{var}}(b_2)$$

- The 95% prediction interval for y_0 is:

$$\hat{y}_0 \pm t_c \text{se}(f) = 287.6089 \pm 2.0244(90.6328)$$
$$= [104.1323,\ 471.0854]$$

- The prediction interval is wide even for $x_0 = 20$, which is close to $\bar{x} = 19.6$. We observe that se(f) = 90.6 is very close to $\hat{\sigma} = 89.6$. Thus the uncertainty in the prediction of y_0 comes from the large uncertainty in the model. We can improve that by:
  - Changing the functional form of the model
  - Including additional independent variables

- There are two major reasons for analyzing the model

Eq. 4.7

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

1. to explain how the dependent variable (y_i) changes as the independent variable (x_i) changes
2. to predict y_0 given an x_0

- Closely allied with the prediction problem is the desire to use x_i to explain as much of the variation in the dependent variable y_i as possible.
  - In the regression model Eq. 4.7 we call x_i the "explanatory" variable because we hope that its variation will "explain" the variation in y_i

- To develop a measure of the variation in y_i that is explained by the model, we begin by separating y_i into its explainable and unexplainable components.

Eq. 4.8

$$y_i = E(y_i) + e_i$$

  – $E(y_i)$ is the explainable or systematic part
  – e_i is the random, unsystematic and unexplainable component

- Analogous to Eq. 4.8, we can write:

Eq. 4.9

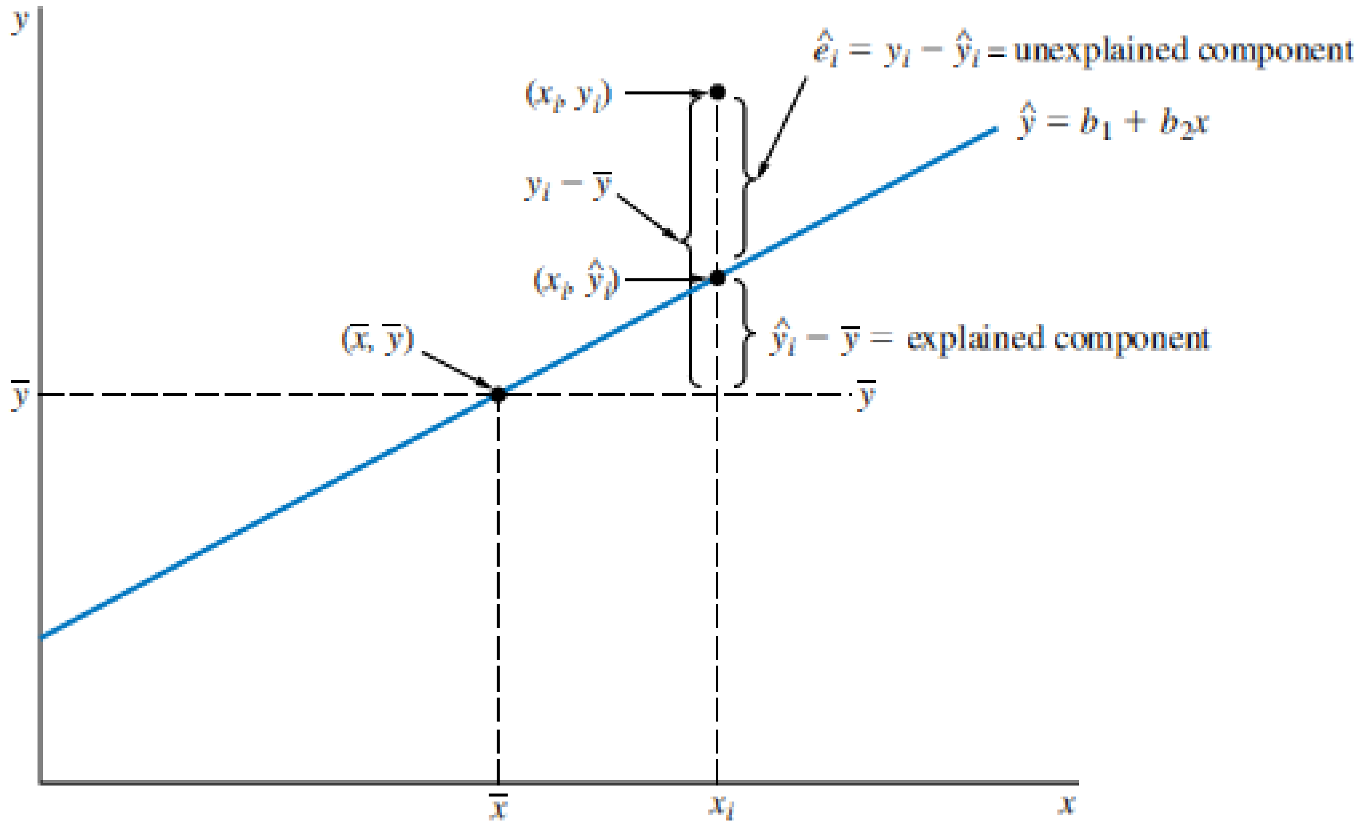
$$y_i = \hat{y}_i + \hat{e}_i$$

  – Subtracting the sample mean from both sides:

Eq. 4.10

$$y_i - \bar{y} = \left(\hat{y}_i - \bar{y}\right) + \hat{e}_i$$

Figure 4.3 Explained and unexplained components of y_i



- Recall that the sample variance of y_i is

$$s_y^2 = \frac{\sum \left(\hat{y}_i - \bar{y} \right)}{N-1}$$

- Squaring and summing both sides of Eq. 4.10, and using the fact that $\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$ we get:

Eq. 4.11

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2$$

- Eq. 4.11 decomposition of the "total sample variation" in y into explained and unexplained components
  - These are called "sums of squares"

- Specifically:

$$\sum\left(y_i-\bar{y}\right)^2=\text{total sum of squares} \;=\; \text{SST}$$

$$\sum\left(\hat{y}_i-\bar{y}\right)^2=\text{sum of squares due to regression} \;=\; \text{SSR}$$

$$\sum\hat{e}_i^2=\text{sum of squares due to error} \;=\; \text{SSE}$$

- We now rewrite Eq. 4.11 as:

$$SST = SSR + SSE$$

- Let's define the **coefficient of determination**, or R^2 , as the proportion of variation in y explained by x within the regression model:

Eq. 4.12

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- We can see that:
  - The closer R^2 is to 1, the closer the sample values y_i are to the fitted regression equation
  - If $R^2 = 1$, then all the sample data fall exactly on the fitted least squares line, so $SSE = 0$, and the model fits the data "perfectly"
  - If the sample data for y and x are uncorrelated and show no linear association, then the least squares fitted line is "horizontal," and identical to $\bar{y}$, so that $SSR = 0$ and $R^2 = 0$

- When $0 < R^2 < 1$ then R^2 is interpreted as "the proportion of the variation in y about its mean that is explained by the regression model"

- The correlation coefficient ρ_{xy} between x and y is defined as:

Eq. 4.13

$$\rho_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\sqrt{\operatorname{var}(y)}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Substituting sample values, as get the sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where:

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) / (N-1)$$

$$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (N-1)}$$

$$s_y = \sqrt{\sum (y_i - \bar{y})^2 / (N-1)}$$

– The sample correlation coefficient r_{xy} has a value between -1 and 1, and it measures the strength of the linear association between observed values of x and y

- Two relationships between R^2 and r_{xy}:
  1. $r^2_{xy} = R^2$
  2. R^2 can also be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = b_1 + b_2 x_i$
- The last property makes R^2 a measure of *goodness-of-fit* of the regression model, that can be extended to the case where we have more than one independent variable.

- For the food expenditure example, the sums of squares are:

$$SST = \sum \left(y_i - \bar{y} \right)^2 = 495132.160$$

$$SSE = \sum \left(y_i - \hat{y} \right)^2 = \sum \hat{e}_i^2 = 304505.176$$

- Therefore:

$$
\begin{aligned}
R^2 &= 1 - \frac{SSE}{SST} \\
&= 1 - \frac{304505.176}{495132.160} \\
&= 0.385
\end{aligned}
$$

  – We conclude that 38.5% of the variation in food expenditure (about its sample mean) is explained by our regression model, which uses only income as an explanatory variable

- The sample correlation between the y and x sample values is:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{478.75}{(6.848)(112.675)}$$

$$= 0.62$$

  – As expected:

$$r_{xy}^2 = 0.62^2 = 0.385 = R^2$$

- The key ingredients in a report are:
  1. the coefficient estimates
  2. the standard errors (or *t*-values)
  3. an indication of statistical significance
  4. R^2
- Avoid using symbols like *x* and *y*
  – Use abbreviations for the variables that are readily interpreted, defining the variables precisely in a separate section of the report.

- For our food expenditure example, we might have:

  $FOOD\_EXP$ = weekly food expenditure by a household of size 3, in dollars

  $INCOME$ = weekly household income, in \$100 units

- And:

$$\widehat{FOOD\_EXP} = 83.42 + 10.21 INCOME \qquad R^2 = 0.385$$

$$\text{(se)} \qquad (43.41)^{*} \; (2.09)^{***}$$

where

* indicates significant at the 10% level

** indicates significant at the 5% level

*** indicates significant at the 1% level

4.3
Modeling Issues

- There are a number of issues we must address when building an econometric model

- What are the effects of scaling the variables in a regression model?
  - Consider the food expenditure example
    - We report weekly expenditures in dollars
    - But we report income in $100 units, so a weekly income of $2,000 is reported as $x = 20$

- If we had estimated the regression using income in dollars, the results would have been:

$$\widehat{FOOD\_EXP} = 83.42 + 0.1021 INCOME(\$) \qquad R^2 = 0.385$$

$$(\text{se}) \qquad (43.41)^{*} \; (0.0209)^{***}$$

– Notice the changes

1. The estimated coefficient of income is now 0.1021
2. The standard error becomes smaller, by a factor of 100.
   – Since the estimated coefficient is smaller by a factor of 100 also, this leaves the *t*-statistic and all other results unchanged.

- Possible effects of scaling the data:
  1. Changing the scale of x: the coefficient of x must be multiplied by c, the scaling factor
     - When the scale of x is altered, the only other change occurs in the standard error of the regression coefficient, but it changes by the same multiplicative factor as the coefficient, so that their ratio, the t-statistic, is unaffected
     - All other regression statistics are unchanged

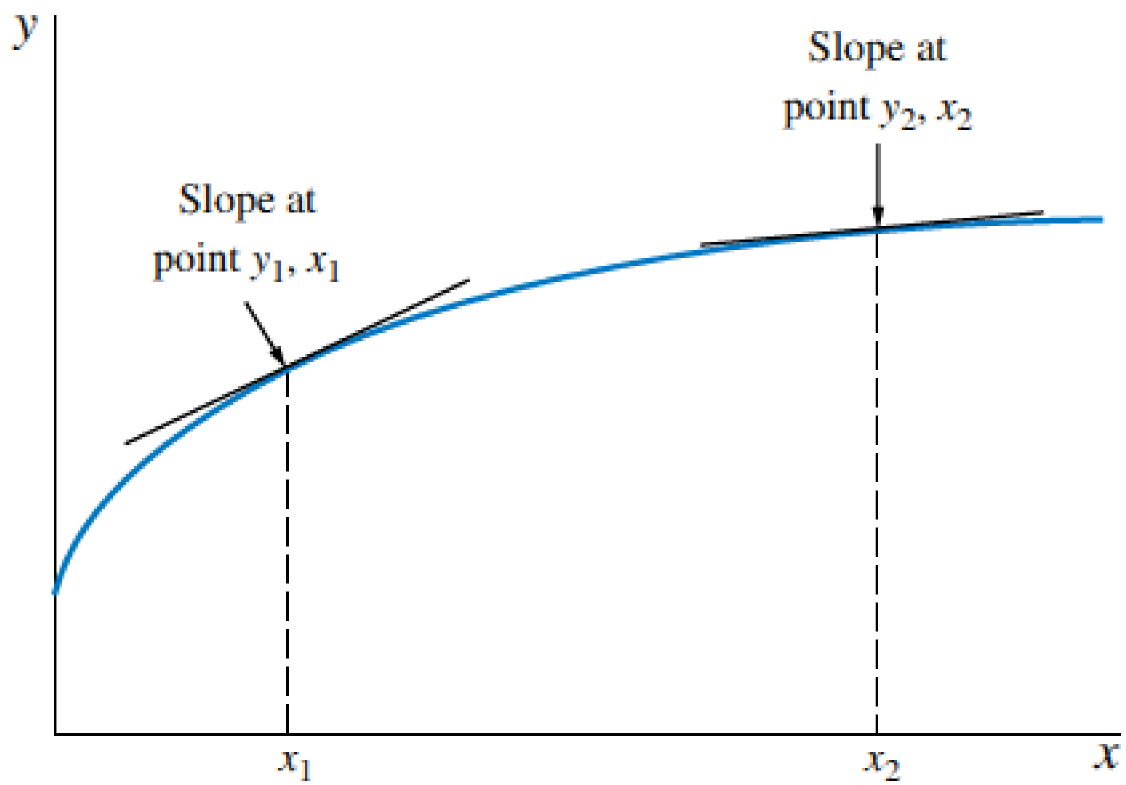
- Possible effects of scaling the data (Continued):
  2. Changing the scale of y: If we change the units of measurement of y, but not x, then all the coefficients must change in order for the equation to remain valid
     - Because the error term is scaled in this process the least squares residuals will also be scaled
     - This will affect the standard errors of the regression coefficients, but it will not affect t-statistics or R^2

- Possible effects of scaling the data (Continued):
  3. Changing the scale of y and x by the same factor: there will be no change in the reported regression results for b_2, but the estimated intercept and residuals will change
     - t-statistics and R^2 are unaffected.
     - The interpretation of the parameters is made relative to the new units of measurement.

- The starting point in all econometric analyses is economic theory
  - What does economics really say about the relation between food expenditure and income, holding all else constant?
  - We expect there to be a positive relationship between these variables because food is a normal good
  - But nothing says the relationship must be a straight line

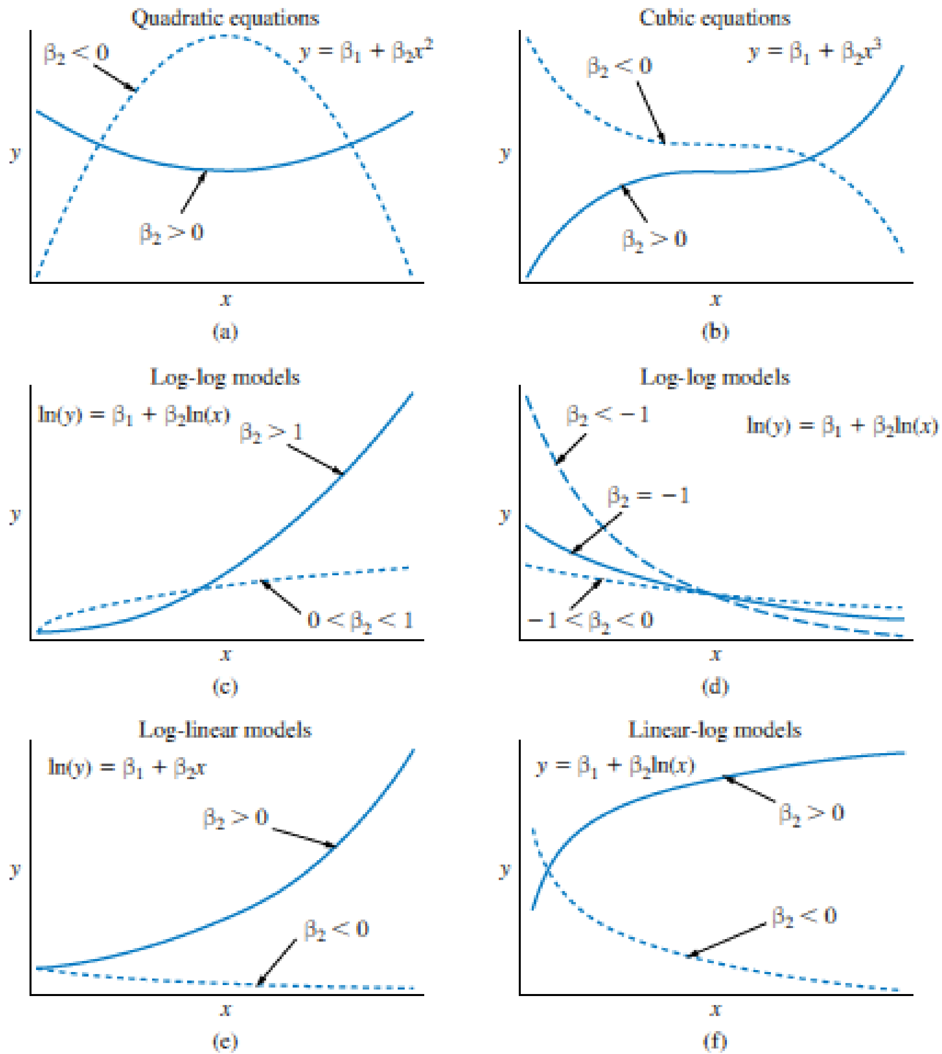
- In fact, we do *not* expect that as household income rises that food expenditure will continue to rise indefinitely at the same constant rate.
- Instead, as income rises we expect food expenditures to rise, but we expect such expenditures to rise in a *decreasing rate*.

Figure 4.4 A nonlinear relationship between food expenditure and income



- By transforming the variables *y* and *x* we can represent many curved, nonlinear relationships and still use the linear regression model
  - Choosing an algebraic form for the relationship means choosing transformations of the original variables
  - The most common are:
    - **Power**: If x is a variable, then x^p means raising the variable to the power p
      - Quadratic (x^2)
      - Cubic (x^3)
    - **Natural logarithm**: If x is a variable, then its natural logarithm is $\ln(x)$

Figure 4.5 Alternative functional forms



- Summary of three configurations:
  1. In the log-log model both the dependent and independent variables are transformed by the "natural" logarithm. Thus,

$$\beta_2 = \frac{d \ln y}{d \ln x} = \frac{dy / y}{dx / x} = \varepsilon$$

  - The parameter β_2 is the elasticity of y with respect to x. An 1% increase in x leads to a β_2% change in y.

  2. In the log-linear model only the dependent variable is transformed by the logarithm. Thus,

$$\beta_2 = \frac{d \ln y}{dx} = \frac{dy / y}{dx}$$

- A unit change in x leads to a 100 β_2 % change in y.

3. For the linear-log model, we can write that:

$$\beta_2 = \frac{dy}{d \ln x} = \frac{dy}{dx / x}$$

- Thus, in the linear-log model we can say that a 1% increase in x leads to a β_2 /100 unit change in y.

Table 4.1 Some Useful Functions, their Derivatives, Elasticities and Other Interpretation

Name	Function	Slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-Linear	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1 unit change in x leads to (approximately) a 100 β_2% change in y		
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

- We wish to choose a functional form for the food expenditure model that is consistent with Figure 4.4.
- One option is the linear-log model.
- The food expenditure model in logs is:

$$FOOD\_EXP = \beta_1 + \beta_2 \ln(INCOME)$$

- The estimated version is:

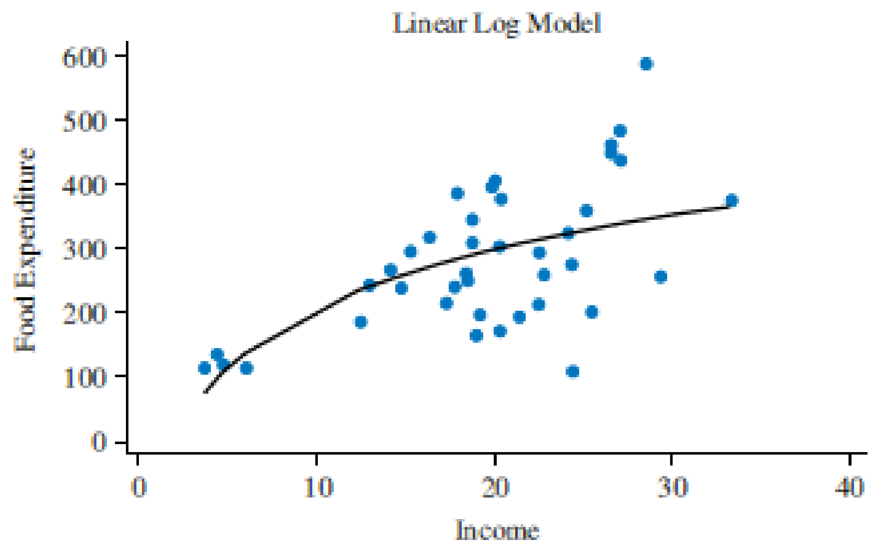
$$\widehat{FOOD\_EXP} = -97.19 + 132.17 \ln(INCOME) \qquad R^2 = 0.357$$

$$\text{(se)} \qquad (84.24) \quad (28.80)^{***}$$

- For a household with \$1,000 weekly income, we estimate that the household will spend an additional \$13.22 on food from an additional \$100 income
  - Whereas we estimate that a household with \$2,000 per week income will spend an additional \$6.61 from an additional \$100 income
  - The marginal effect of income on food expenditure is smaller at higher levels of income
    - This is a change from the linear, straight-line relationship we originally estimated, in which the marginal effect of a change in income of \$100 was \$10.21 for all levels of income

- Alternatively, we can say that a 1% increase in income will increase food expenditure by approximately \$1.32 per week, or that a 10% increase in income will increase food expenditure by approximately \$13.22

Figure 4.6 The fitted linear-log model



GUIDELINES FOR CHOOSING A FUNCTIONAL FORM

1. Choose a shape that is consistent with what economic theory tells us about the relationship.
2. Choose a shape that is sufficiently flexible to "fit" the data.
3. Choose a shape so that assumptions SR1–SR6 are satisfied, ensuring that the least squares estimators have the desirable properties described in Chapters 2 and 3

- Even when we have chosen an adequate functional form, one or more of the model assumptions may not hold.
  1. Examine the regression results vis-à-vis theory.
  2. Look at the least squares residuals in order to see if there are any evidence that assumptions SR3 (homoskedasticity), SR4 (no serial correlation) and SR6 (normality) are violated.
     - There are formal statistical tests to check for:
       - Homoskedasticity
       - Serial correlation
       - Normality
     - Use residual plots

Figure 4.7 Randomly scattered residuals

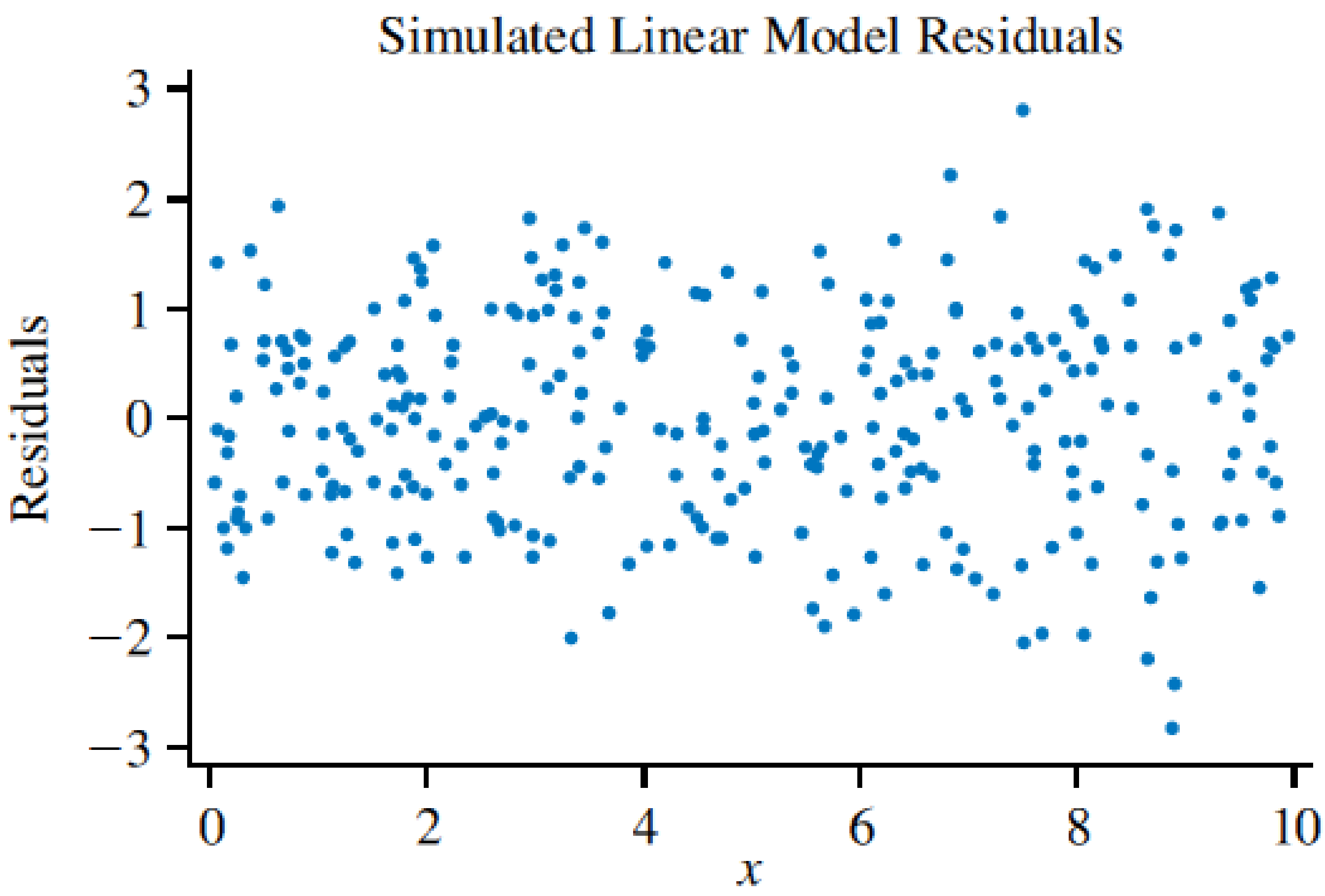
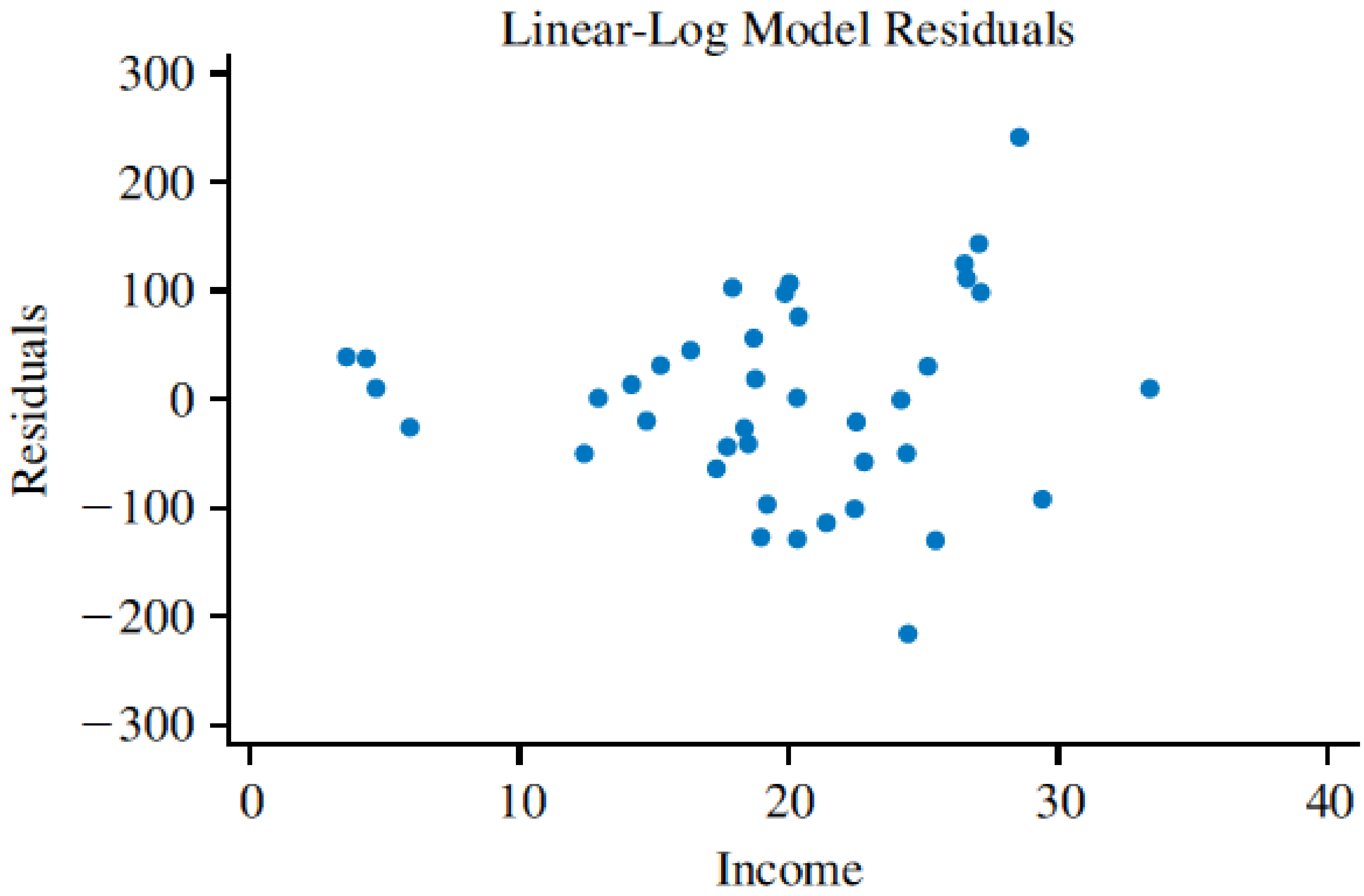
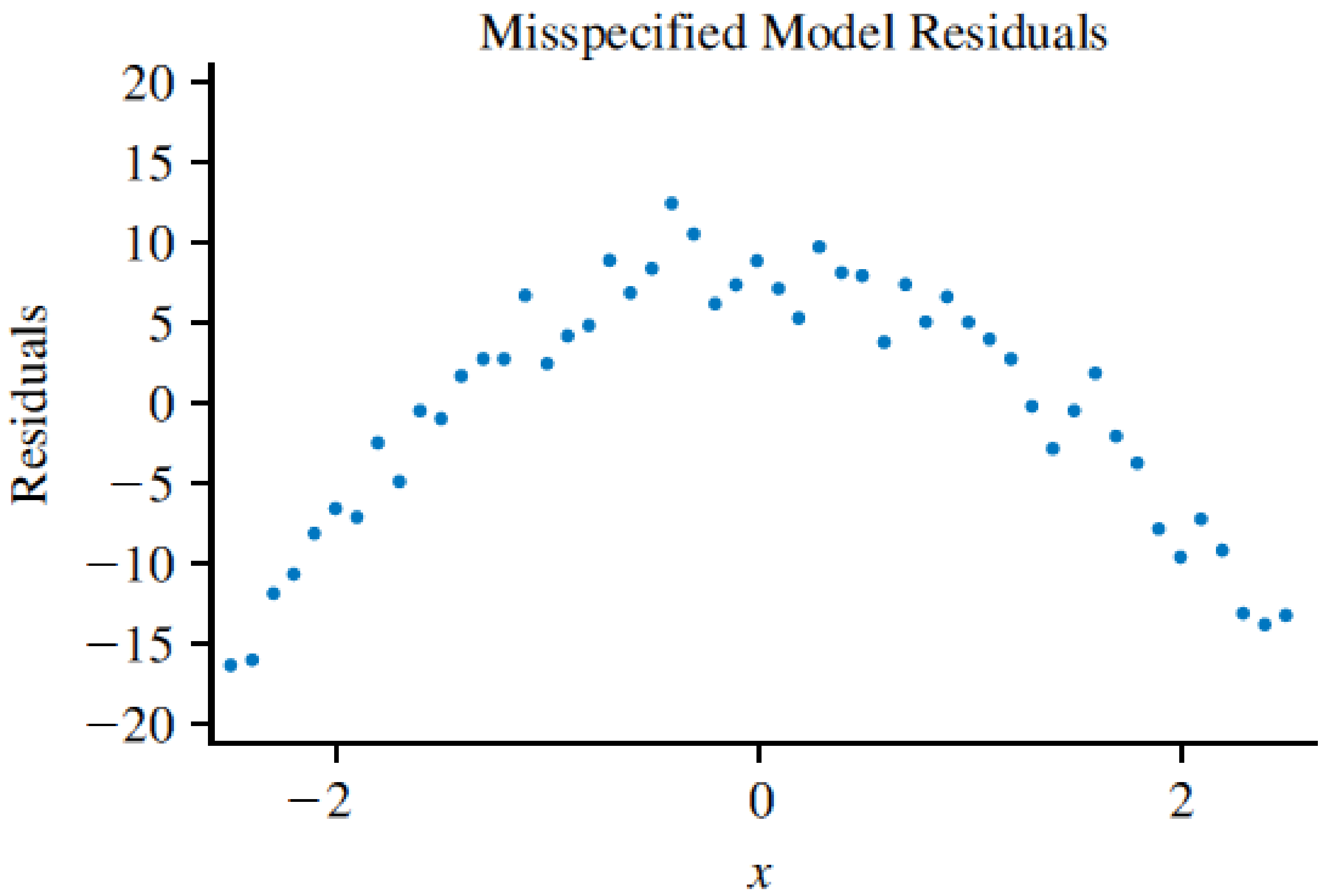


Figure 4.8 Residuals from linear-log food expenditure model



- The well-defined quadratic pattern in the least squares residuals indicates that something is wrong with the linear model specification
  - The linear model has "missed" a curvilinear aspect of the relationship
  - *Or* the error term is not uncorrelated (assumption SR4)

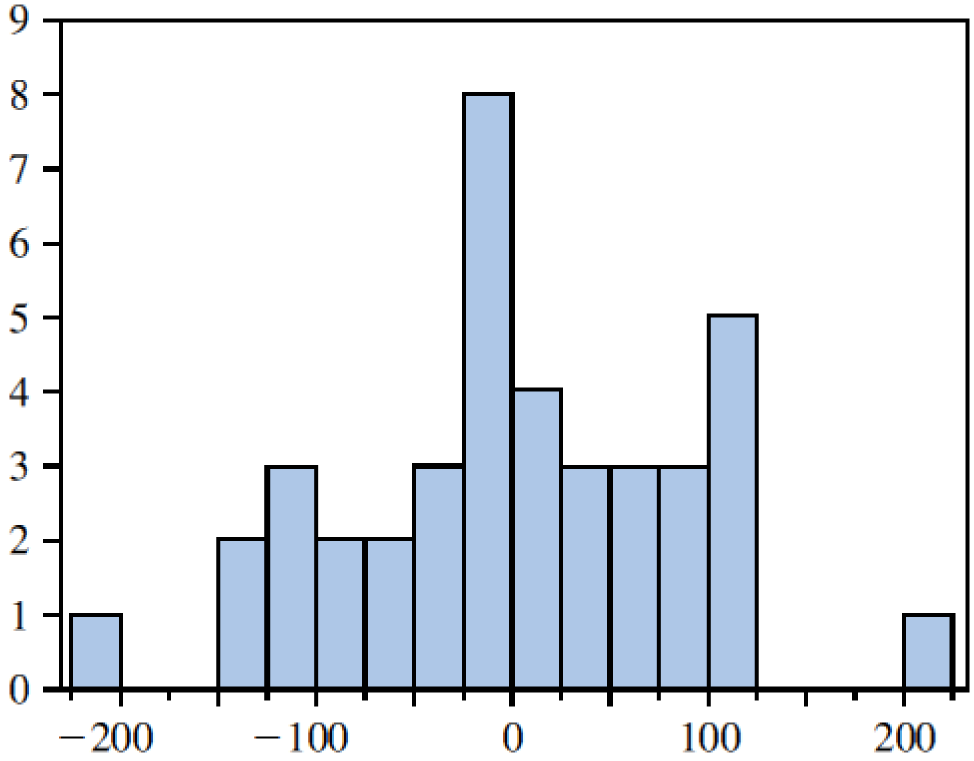
Figure 4.9 Least squares residuals from a linear equation fit to quadratic data



- Hypothesis tests and interval estimates for the coefficients rely on the assumption that the errors, and hence the dependent variable *y*, are normally distributed
  - Are they normally distributed?

- We can check the distribution of the residuals using:
  - A histogram
  - Formal statistical test
    - Merely checking a histogram is not a formal test
    - Many formal tests are available
      - A good one is the **Jarque–Bera test** for normality

Figure 4.10 EViews output: residuals histogram and summary statistics for food expenditure



Series: Residuals
Sample 1 40
Observations 40

Mean	6.93e-15
Median	−6.324473
Maximum	212.0440
Minimum	−223.0255
Std. Dev.	88.36190
Skewness	−0.097319
Kurtosis	2.989034
Jarque-Bera	0.063340
Probability	0.968826

- The Jarque–Bera test for normality is based on two measures, skewness and kurtosis
  - Skewness refers to how symmetric the residuals are around zero
    - Perfectly symmetric residuals will have a skewness of zero
    - The skewness value for the food expenditure residuals is -0.097
  - Kurtosis refers to the "peakedness" of the distribution.
    - For a normal distribution the kurtosis value is 3

- The Jarque–Bera statistic is given by:

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

where

$N =$ sample size

$S =$ skewness

$K =$ kurtosis

- When the residuals are normally distributed, the Jarque–Bera statistic has a chi-squared distribution with two degrees of freedom
  - We reject the hypothesis of normally distributed errors if a calculated value of the statistic exceeds a critical value selected from the chi-squared distribution with two degrees of freedom
    - The 5% critical value from a χ^2-distribution with two degrees of freedom is 5.99, and the 1% critical value is 9.21

- For the food expenditure example, the Jarque–Bera statistic is:

$$JB = \frac{40}{6}\left(-0.097^2 + \frac{(2.99-3)^2}{4}\right) = 0.063$$

  – Because 0.063 < 5.99 there is insufficient evidence from the residuals to conclude that the normal distribution assumption is unreasonable at the 5% level of significance

- We could reach the same conclusion by examining the *p*-value
  - The *p*-value appears in Figure 4.10 described as "Probability"
  - Thus, we also fail to reject the null hypothesis on the grounds that $0.9688 > 0.05$

4.4
Polynomial Models

- In addition to estimating linear equations, we can also estimate quadratic and cubic equations

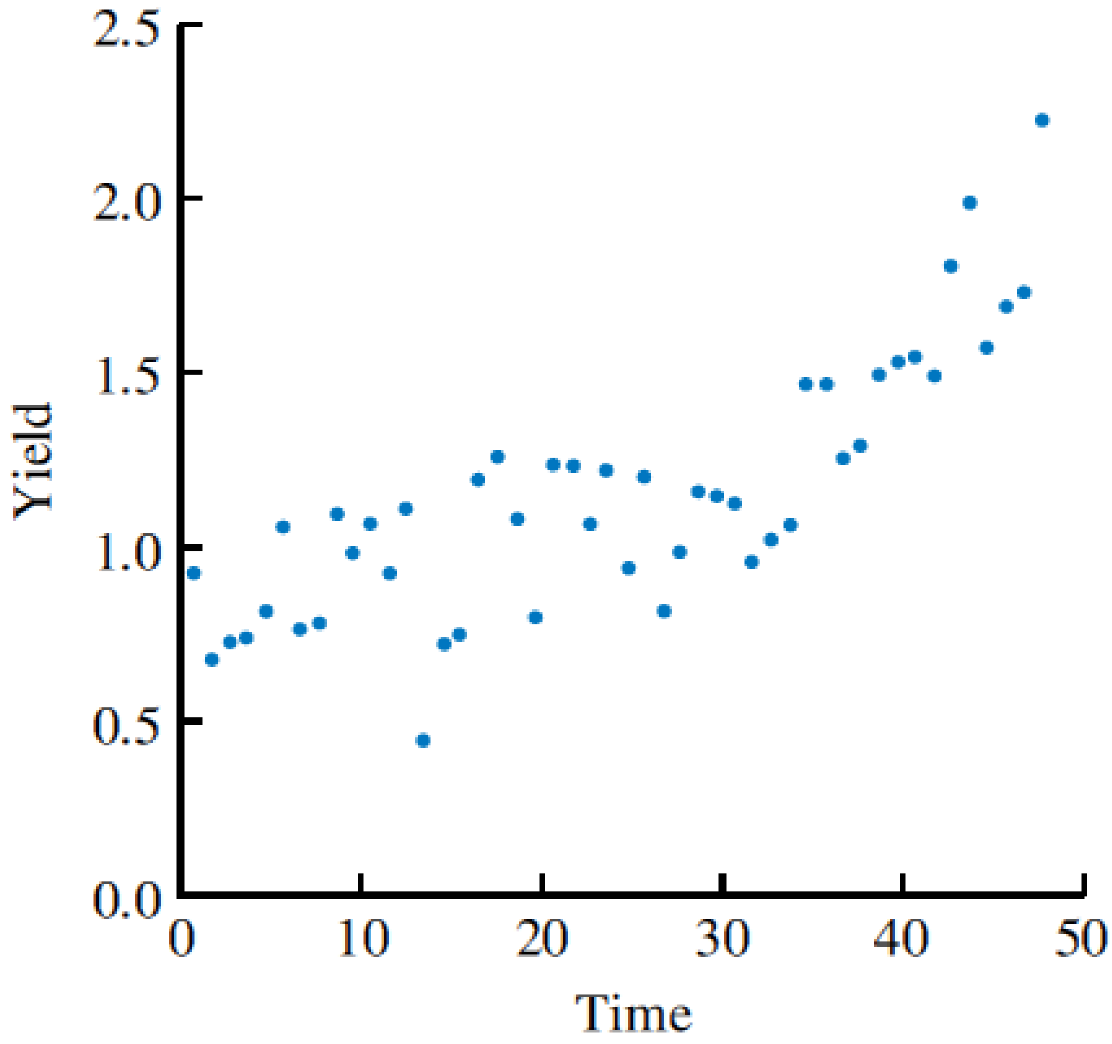
- The general form of a quadratic equation with one independent variable is:

$$y = \beta_1 + \beta_2 x^2$$

- The general form of a cubic equation is:

$$y = \beta_1 + \beta_2 x^3$$

Figure 4.11 Scatter plot of wheat yield over time



- One problem with the linear equation

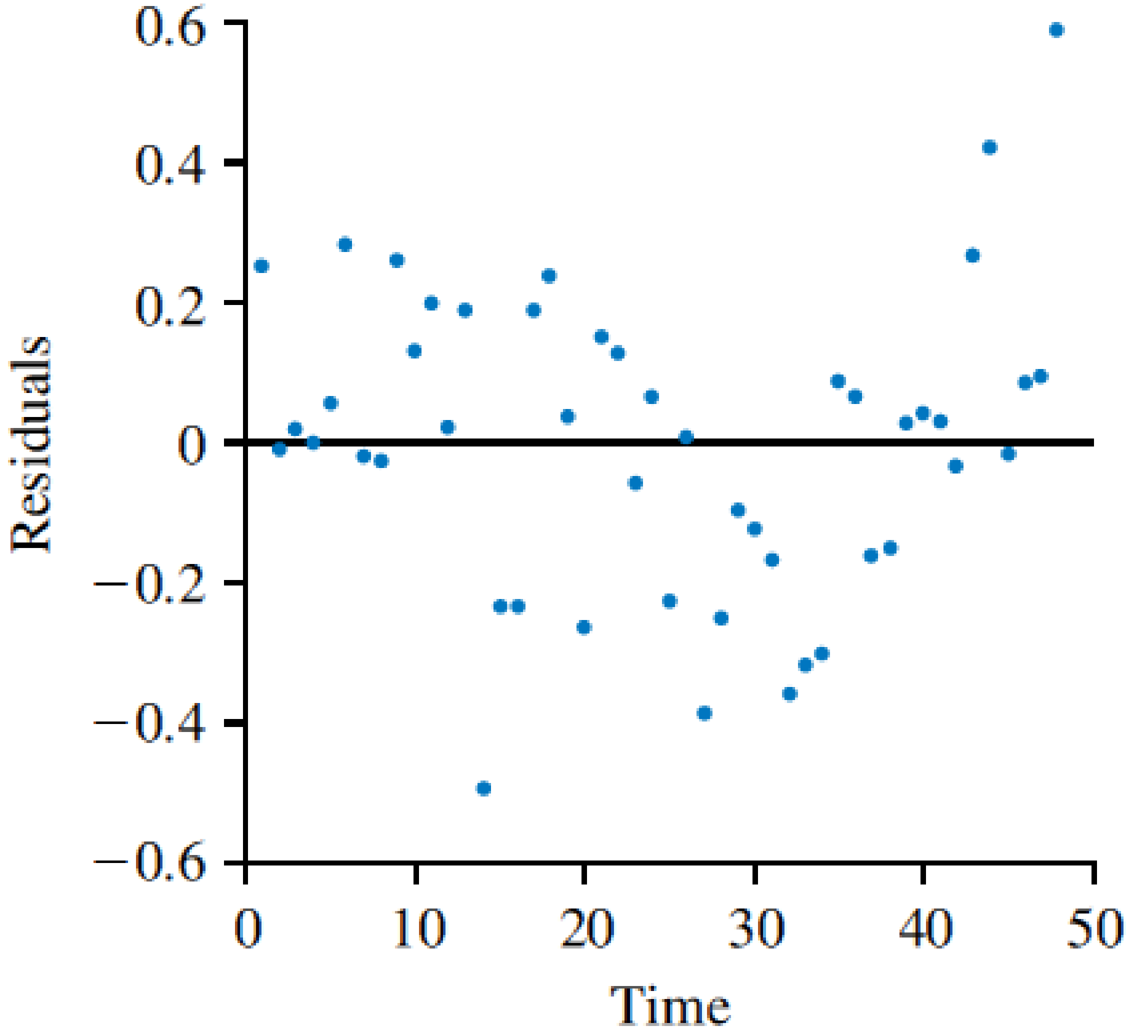
$$YIELD_t = \beta_2 + \beta_2 TIME + e_t$$

is that it implies that yield increases at the same constant rate β_2, when, from Figure 4.11, we expect this rate to be increasing

- The least squares fitted line is:

$$\widehat{YIELD_t} = 0.638 + 0.0210 TIME_t \qquad R^2 = 0.649$$
$$(\text{se}) \qquad (0.064)^{***} \quad (0.0022)^{***}$$

Figure 4.12 Residuals from a linear yield equation



- Perhaps a better model would be:

$$YIELD_t = \beta_1 + \beta_2 TIME_t^3 + e_t$$

But note that the values of $TIME_t^3$ can get very large

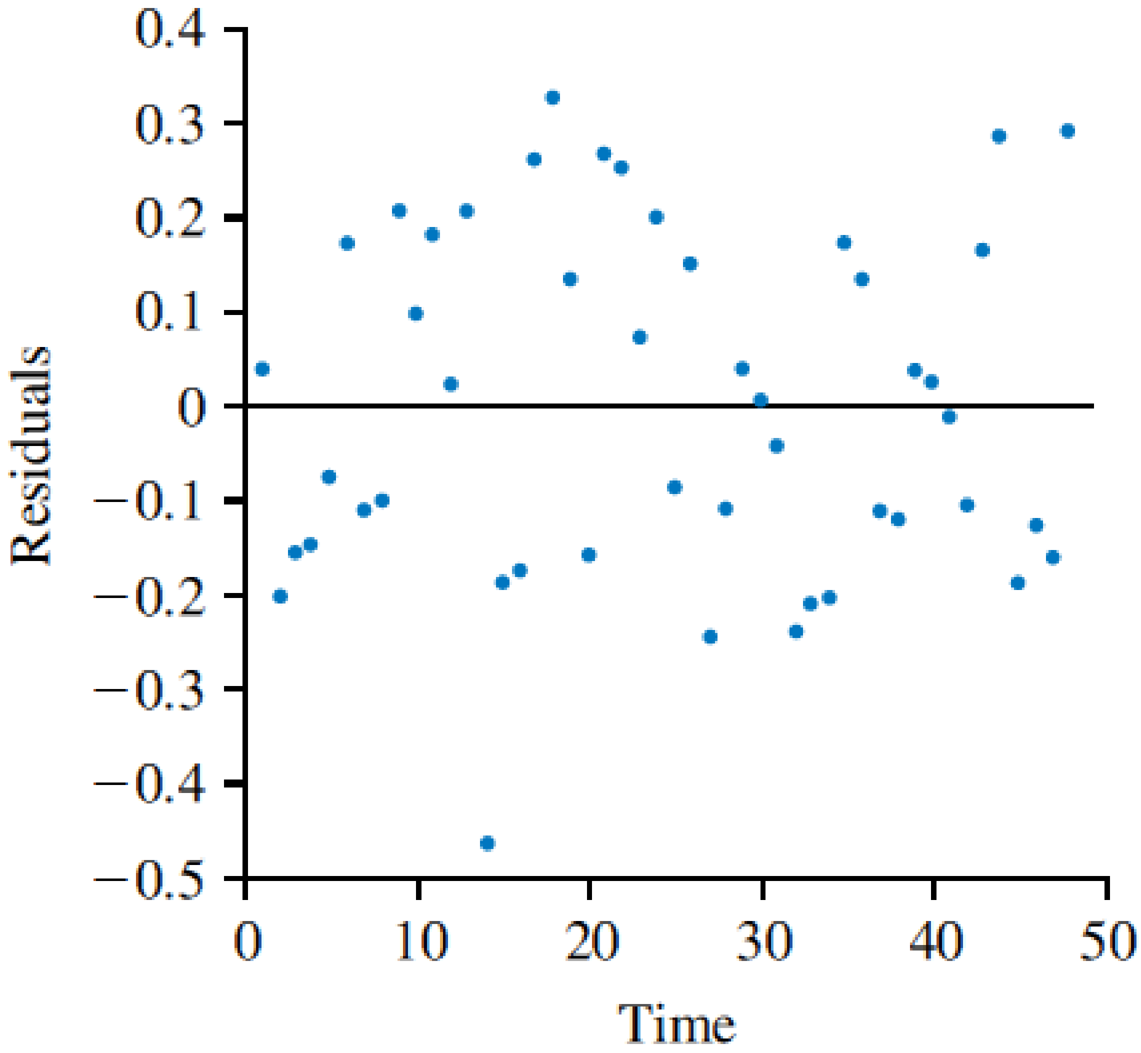
– This variable is a good candidate for scaling. Define $TIMECUBE_t = TIME_t^3/1000000$

- The least squares fitted line is:

$$\widehat{YIELD}_t = 0.874 + 9.68 TIMECUBE_t \qquad R^2 = 0.751$$

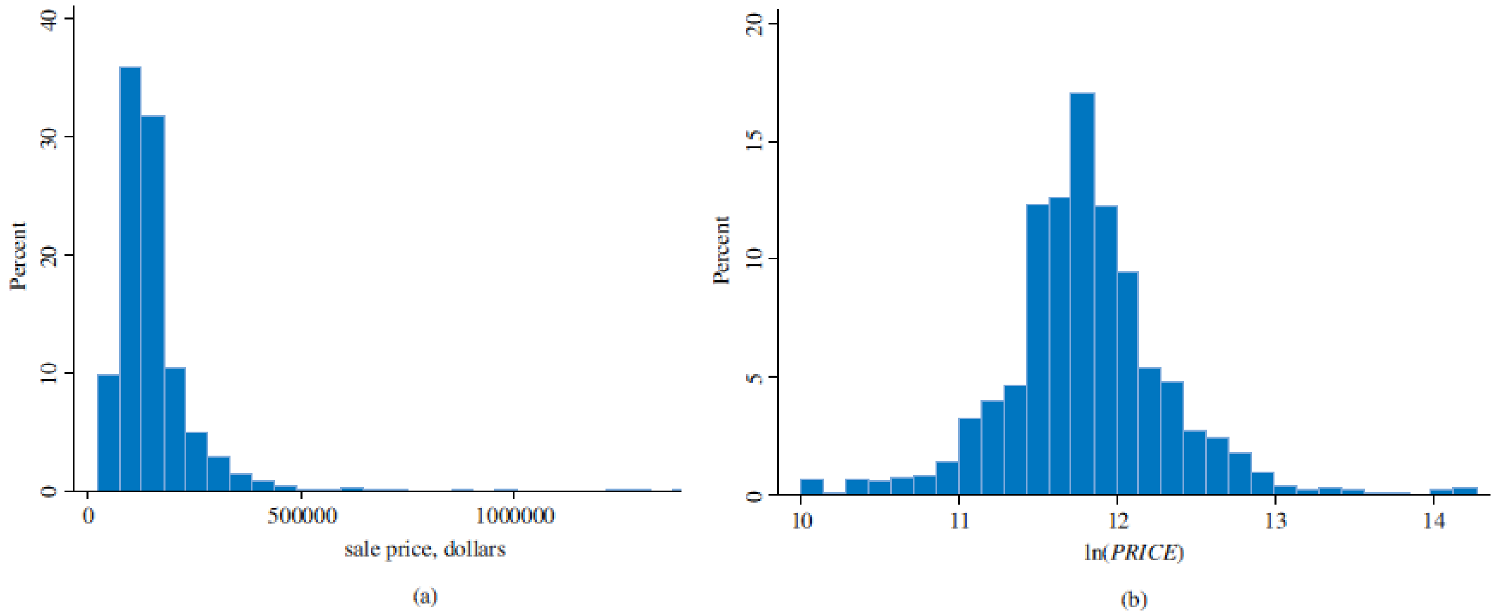
$$(\text{se}) \qquad (0.036)^{***} \quad (0.822)^{***}$$

FI G U RE 4.13 Residuals from a cubic yield equation



4.5
Log-linear Models

- Econometric models that employ natural logarithms are very common
  - Logarithmic transformations are often used for variables that are monetary values
    - Wages, salaries, income, prices, sales, and expenditures
    - In general, for variables that measure the "size" of something
    - These variables have the characteristic that they are positive and often have distributions that are positively skewed, with a long tail to the right


Percent
0
10
20
30
40
0
500000
1000000
sale price, dollars
(a)
Percent
0
5
10
15
20
10
11
12
13
14
ln(PRICE)
(b)

- The log-linear model, $\ln(y) = \beta_1 + \beta_2 x$, has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side
  - Both its slope and elasticity change at each point and are the same sign as β_2
  - In the log-linear model, a one-unit increase in x leads, approximately, to a $100\,\beta_2$ % change in y

- Suppose that the yield in year t is $YIELD_t = (1+g)YIELD_{t-1}$, with g being the fixed growth rate in 1 year
  - By substituting repeatedly we obtain $YIELD_t = YIELD_0(1+g)^t$
  - Here $YIELD_0$ is the yield in year "0," the year before the sample begins, so it is probably unknown

- Taking logarithms, we obtain:

$$\ln\left(YIELD_t\right) = \ln\left(YIELD_0\right) + \left[\ln\left(1+g\right)\right] \times t$$
$$= \beta_1 + \beta_2 t$$

- The fitted model is:

$$\ln\left(\widehat{YIELD_t}\right) = -0.3434 + 0.0178t$$
$$(\text{se}) \qquad (0.0584)^{***} \; (0.0021)^{***}$$

- Using the property that $\ln(1+x) \approx x$ if x is small, we estimate that the growth rate in wheat yield is approximately $\hat{g} = 0.0178$, or about 1.78% per year, over the period of the data.

- Suppose that the rate of return to an extra year of education is a constant r
  - A model for wages might be:

$$\begin{aligned}\ln(WAGE) &= \ln(WAGE_0) + [\ln(1+r)] \times EDUC \\ &= \beta_1 + \beta_2 EDUC\end{aligned}$$

- A fitted model would be:

$$\ln\left(\widehat{WAGE}\right) = 1.6094 + 0.0904 \times EDUC$$
$$(\text{se}) \qquad (0.0864)^{***} \quad (0.0061)^{***}$$

  – An additional year of education increases the wage rate by approximately 9%

    • A 95% interval estimate for the value of an additional year of education is 7.8% to 10.2%

- In a log-linear regression the R^2 value automatically reported by statistical software is the percent of the variation in $\ln(y)$ explained by the model
  - However, our objective is to explain the variations in y, not $\ln(y)$
  - Furthermore, the fitted regression line predicts

$$\widehat{\ln(y)} = b_1 + b_2 x$$

whereas we want to predict y

- A natural choice for prediction is:

$$\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(b_1 + b_2 x)$$

  - The subscript "n" is for "natural"
  - But a better alternative is:

$$\hat{y}_c = \widehat{E(y)} = \exp\left(b_1 + b_2 x + \hat{\sigma}^2/2\right) = \hat{y}_n e^{\hat{\sigma}^2/2}$$

  - The subscript "c" is for "corrected"
  - This uses the properties of the **log-normal distribution**

- Recall that $\hat{\sigma}^2$ must be greater than zero and $e^0 = 1$
  - Thus, the effect of the correction is always to increase the value of the prediction, because $e^{\hat{\sigma}^2/2}$ is always greater than one
  - The natural predictor tends to systematically underpredict the value of *y* in a log-linear model, and the correction offsets the downward bias in large samples

- For the wage equation:

$$\widehat{\ln(WAGE)} = 1.6094 + 0.0904 \times EDUC = 1.6094 + 0.0904 \times 12 = 2.6943$$

- The natural predictor is:

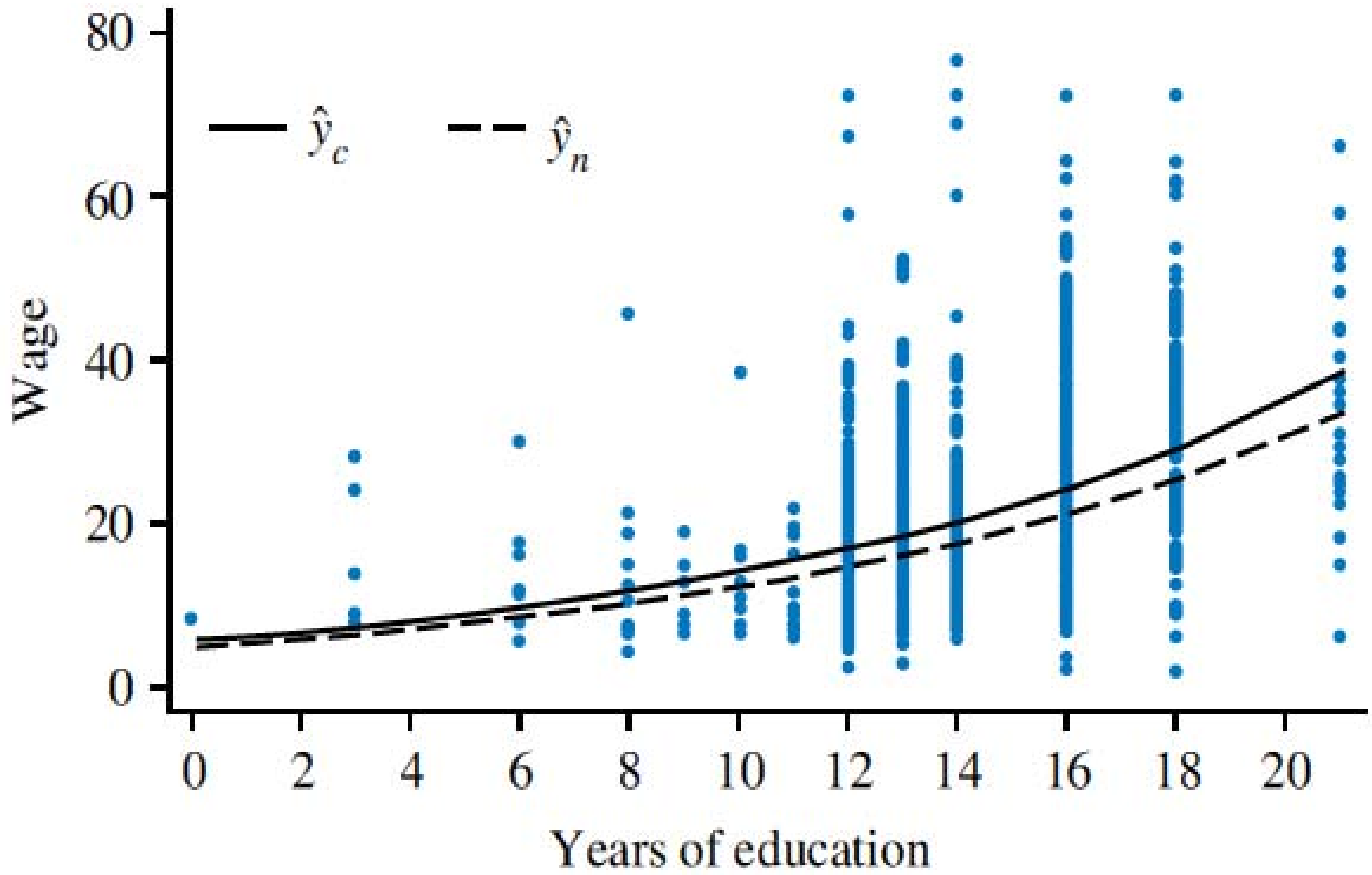
$$\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(2.6943) = 14.7958$$

- The corrected predictor is:

$$\hat{y}_c = \widehat{E(y)} = \hat{y}_n e^{\hat{\sigma}/2} = 14.7958 \times 1.1487 = 16.9964$$

  – We predict that the wage for a worker with 12 years of education will be \$14.80 per hour if we use the natural predictor, and \$17.00 if we use the corrected predictor

FIGURE 4.14 The natural and corrected predictors of wage



- A general goodness-of-fit measure, or general R^2, is:

$$R_g^2 = \left[\mathrm{corr}\left(y, \hat{y}\right)\right]^2 = r_{y\hat{y}}^2$$

- For the wage equation, the general R^2 is:

$$R_g^2 = \left[\operatorname{corr}\left(y, \hat{y}_c\right)\right]^2 = 0.4312^2 = 0.1859$$

– Compare this to the reported $R^2 = 0.1782$

- A 100(1 – α)% prediction interval for y is:

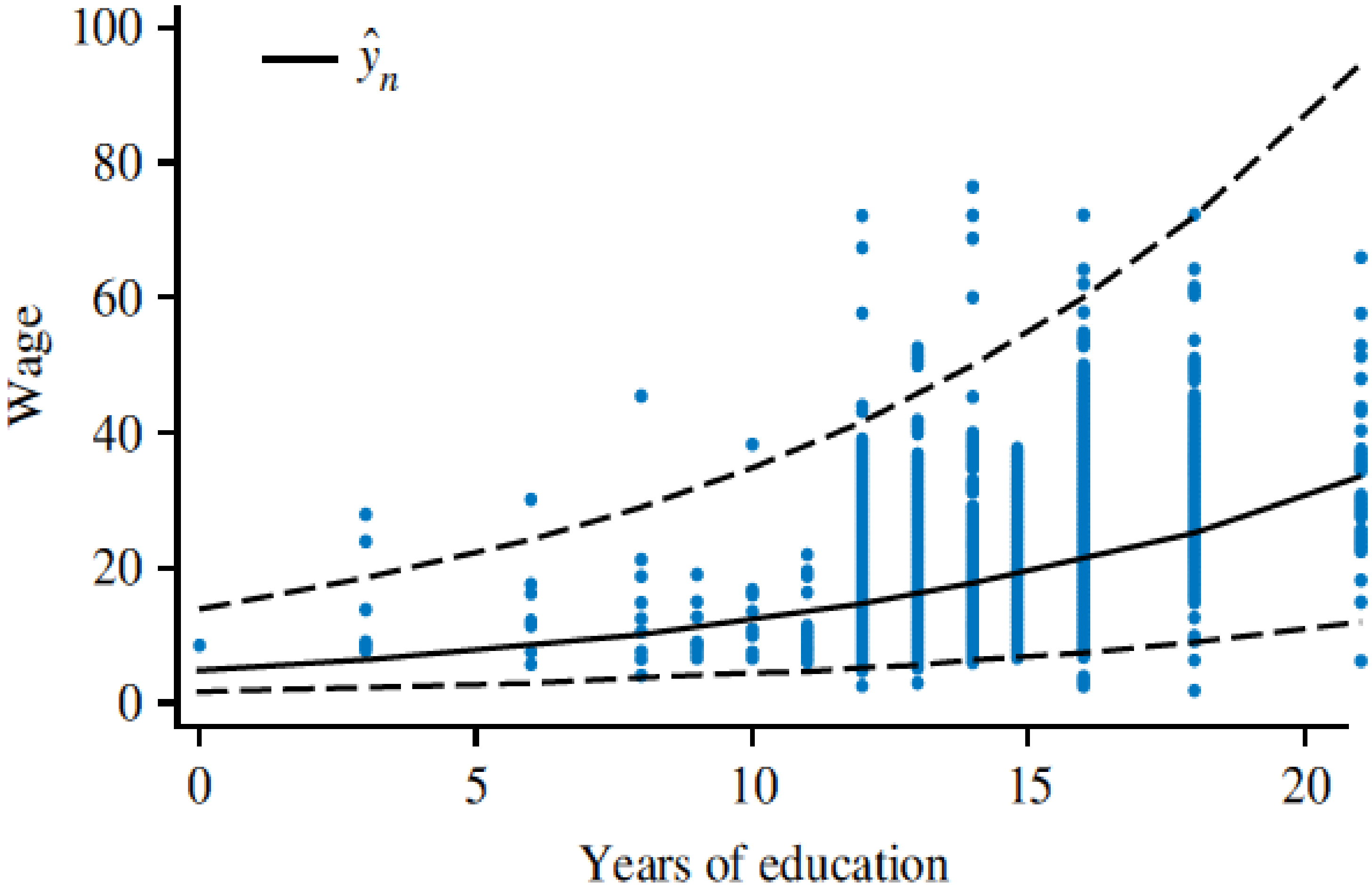
$$\left[\exp\left(\widehat{\ln(y)}-t_c se(f)\right),\exp\left(\widehat{\ln(y)}+t_c se(f)\right)\right]$$

- For the wage equation, a 95% prediction interval for the wage of a worker with 12 years of education is:

$$\left[\exp\left(2.6943 - 1.96 \times 0.5270\right), \exp\left(2.6943 + 1.96 \times 0.5270\right)\right]$$

$$= \left[52604,\ 41.6158\right]$$

FIGURE 4.15 The 95% prediction interval for wage

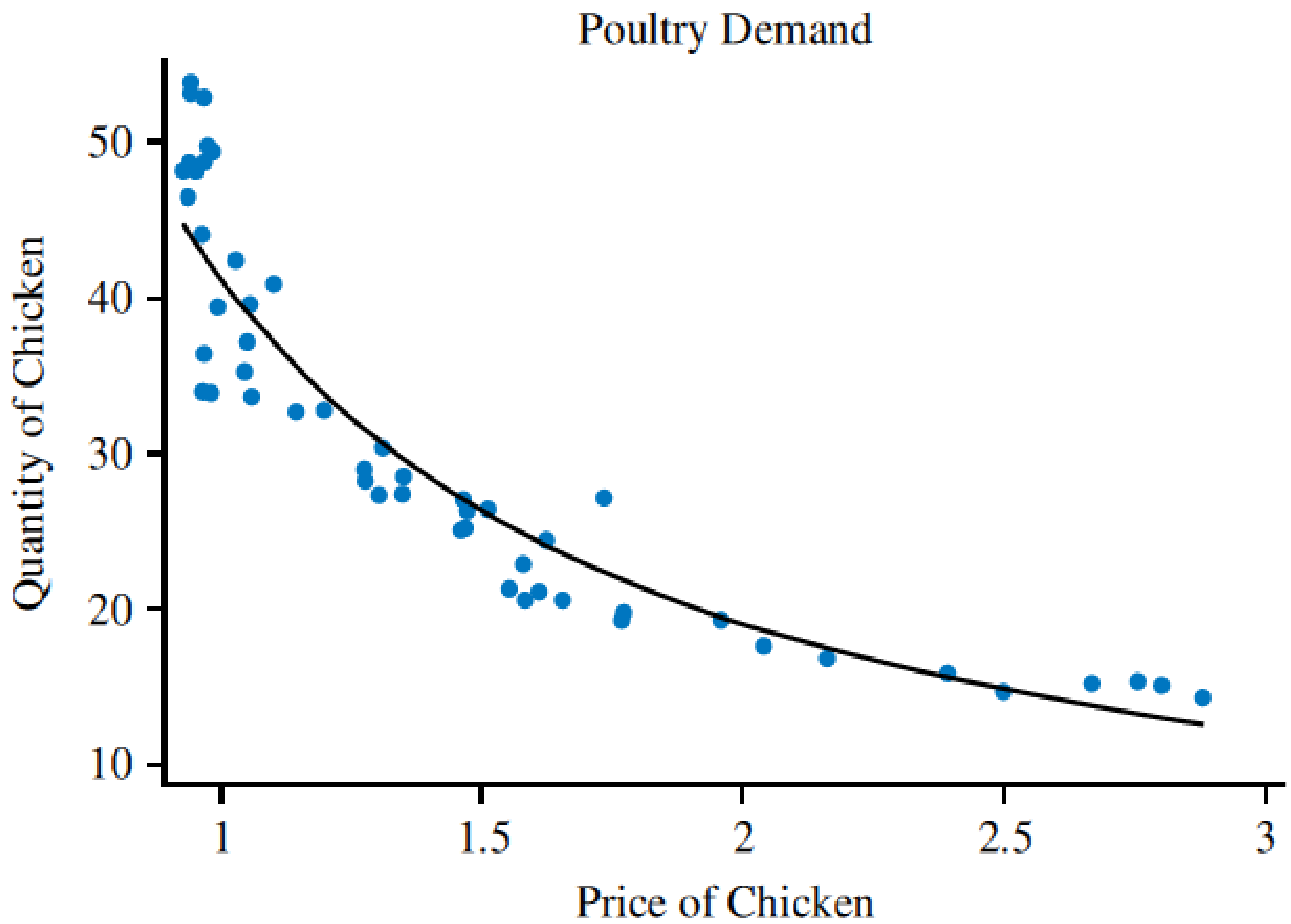


4.6
Log-log Models

- The log-log function, $\ln(y) = \beta_1 + \beta_2 \ln(x)$, is widely used to describe demand equations and production functions
  - In order to use this model, all values of y and x must be positive
  - The slopes of these curves change at every point, but the elasticity is constant and equal to β_2

- If $\beta_2 > 0$, then y is an increasing function of x
  - If $\beta_2 > 1$, then the function increases at an increasing rate
  - If $0 < \beta_2 < 1$, then the function is increasing, but at a decreasing rate
- If $\beta_2 < 0$, then there is an inverse relationship between y and x

FIGURE 4.16 Quantity and Price of Chicken



- The estimated model is:

Eq. 4.15

$$\widehat{\ln(Q)} = 3.717 - 1.121 \times \ln(P) \qquad R_g^2 = 0.8817$$
$$(\text{se}) \quad (0.022)^{***} \quad (0.049)^{***}$$

– We estimate that the price elasticity of demand is 1.121: a 1% increase in real price is estimated to reduce quantity consumed by 1.121%

- Using the estimated error variance $\hat{\sigma}^2 = 0.0139$, the corrected predictor is:

$$\begin{aligned}\hat{\mathrm{Q}}_{\mathrm{c}} &= \hat{Q}_n e^{\hat{\sigma}^2/2} \\ &= \exp\left(\widehat{\ln(Q)}\right) e^{\hat{\sigma}^2/2} \\ &= \exp\left(3.717 - 2.121 \times \ln(P)\right) e^{0.0139/2}\end{aligned}$$

- The generalized goodness-of-fit is:

$$R_g^2 = \left[\operatorname{corr}\left(Q, \hat{Q}_c\right)\right]^2 = 0.939^2 = 0.8817$$

Key Words

Keywords

- coefficient of determination
- correlation
- data scale
- forecast error
- forecast standard error
- functional form
- goodness-of-fit
- growth model
- Jarque-Bera test
- Kurtosis
- least squares predictor
- linear model
- linear relationship
- linear-log model
- log-linear model
- log-log model
- log-normal distribution
- Prediction
- prediction interval
- R^2
- Residual
- skewness