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

Chapter Contents

4.1 Least Squares Prediction 1 of 7

- The ability to predict is important to:
  - Business economists and financial analysts who attempt to forecast the sales and revenues of specific firms
  - Government policymakers 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 service
- Accurate predictions provide a basis for better decision making in every type of planning context

4.1 Least Squares Prediction 2 of 7

- 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
  - (4.1) $y_0 = \beta_1 + \beta_2 x_0 + e_0$
- where e_0 is a random error.

4.1 Least Squares Prediction 3 of 7

- The task of predicting y_0 is related to the problem of estimating $E(y_0| x_0)= \beta_1 + \beta_2 x_0$
- Although $E(y_0 \mid x_0) = \beta_1 + \beta_2 x_0$ is not random, the outcome y_0 is random
- Consequently, as we will see, there is a difference between the **interval estimate** of $E(y_0 \mid x_0) = \beta_1 + \beta_2 x_0$ and the **prediction interval** for y_0
- The **least squares predictor** of y_0 comes from the fitted regression line
  - (4.2) $$\hat{y}_0 = b_1 + b_2 x_0$$

4.1 Least Squares Prediction 4 of 7

- To evaluate how well this predictor performs, we define the forecast error, which is analogous to the least squares residual:
  - (4.3) $f = y_0 - \hat{y}_0 = \left(\beta_1 + \beta_2 x_0 + e_0\right) - \left(b_1 + b_2 x_0\right)$
- We would like the forecast error to be small, implying that our forecast is close to the value we are predicting

4.1 Least Squares Prediction 5 of 7

- Taking the expected value of *f*, we find that:
- $E(f|x) = \beta_1 + \beta_2 x_0 + E(e_0) - [E(b_1) + E(b_2)x_0] = \beta_1 + \beta_2 x_0 + 0 - [\beta_1 + \beta_2 x_0] = 0$
- which means, on average, the forecast error is zero and $\hat{y}_0$ is an **unbiased predictor** of 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 y_0 if assumptions SR1–SR5 hold

4.1 Least Squares Prediction 6 of 7

- The variance of the forecast is: (4.4) $var(f|x) = \sigma^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$
- 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

4.1 Least Squares Prediction 7 of 7

- In practice we use $\widehat{var}(f|x) = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]$ for the variance

- The **standard error of the forecast** is: (4.5) $\text{se}(f) = \sqrt{\widehat{var}(f|x)}$

- The 100(1 – α)% **prediction interval** is:

  - (4.6) $\hat{y}_0 \pm t_c \text{se}(f)$

Figure 4.2 Point and interval prediction.

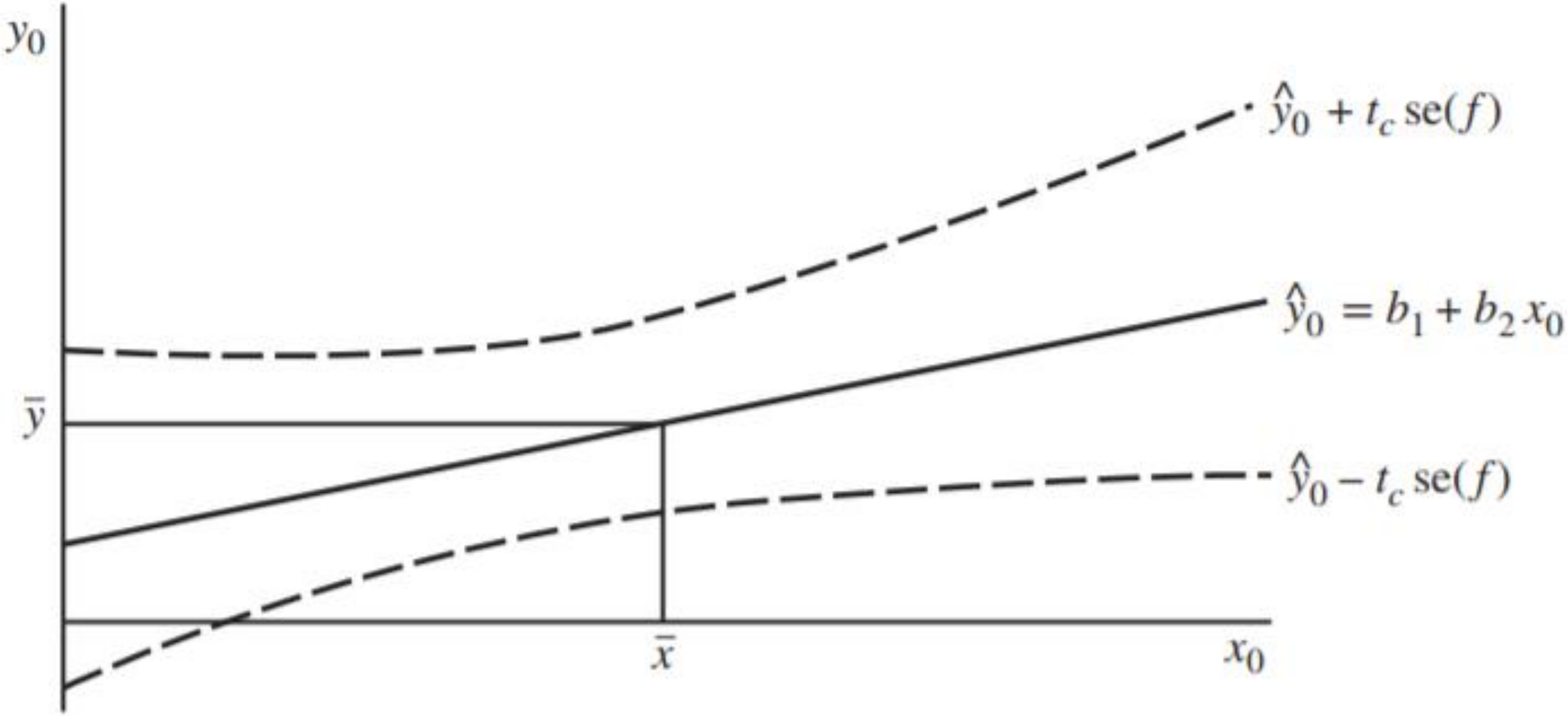


FIGURE 4.2 **Point and interval prediction.**

4.2 Measuring Goodness-of-Fit 1 of 6

- There are two major reasons for analyzing the model
  - (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

4.2 Measuring Goodness-of-Fit 2 of 6

- 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
  - (4.8) $y_i = E(y_i|x) + e_i$
  - $E(y_i|x)$ is the explainable or systematic part
  - e_i is the random, unsystematic and unexplainable component

4.2 Measuring Goodness-of-Fit 3 of 6

- Recall that the sample variance of y_i is $s_y^2 = \dfrac{\sum(\hat{y}_i - \bar{y})}{N-1}$
- Squaring and summing both sides of (4.10), and using the fact that $\sum(\hat{y}_i - \bar{y})e_i = 0$

  we get: (4.11) $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum e_i^2$
- Eq. 4.11 decomposition of the “total sample variation” in y into explained and unexplained components
  - These are called “sums of squares”

4.2 Measuring Goodness-of-Fit 4 of 6

- Specifically:

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

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

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

- Using these abbreviations, (4.11) becomes ***SST = SSR + SSE***

4.2 Measuring Goodness-of-Fit 5 of 6

- Let's define the **coefficient of determination**, or R^2 , as the proportion of variation in *y* explained by *x* within the regression model:
  - (4.12) $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- The closer R2 is to 1, the closer the sample values y_i are to the fitted regression equation

4.2 Measuring Goodness-of-Fit 6 of 6

- 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 $R^2 = 0$

Figure 4.3 Explained and unexplained components of y_i

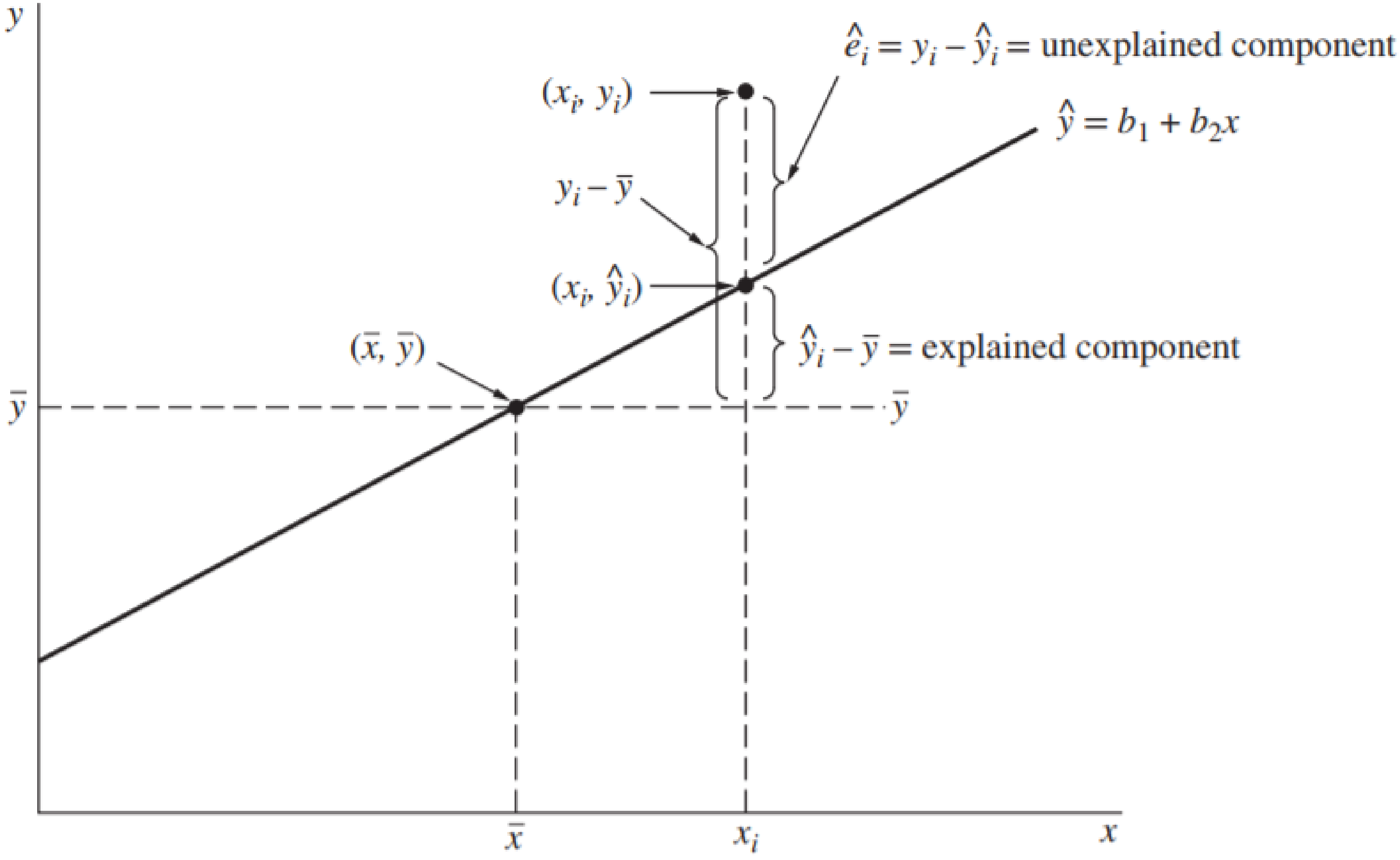


FIGURE 4.3 **Explained and unexplained components of y_i.**

4.2.1 Correlation Analysis 1 of 2

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

- (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:

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

4.2.1 Correlation Analysis 2 of 2

- 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

4.2.2 Correlation Analysis and R^2

- 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 $\hat{y}_i = b_1 + b_2 x_i$

4.3.1 The Effects of Scaling the Data 1 of 4

- 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 $2,000 is reported as x = 20
- If we had estimated the regression using income in dollars, the results would have been:
- FOOD_EXP = 83.42 + 0.1021 INCOME($) R^2 = 0.385 (se) (43.41) $*$(0.0209) $***$

4.3.1 The Effects of Scaling the Data 2 of 4

- 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

4.3.1 The Effects of Scaling the Data 3 of 4

- Possible effects of scaling the data:

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

4.3.1 The Effects of Scaling the Data 4 of 4

- Possible effects of scaling the data:

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.

4.3.2 Choosing a Functional Form 1 of 3

- 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

4.3.2 Choosing a Functional Form 2 of 3

- 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.

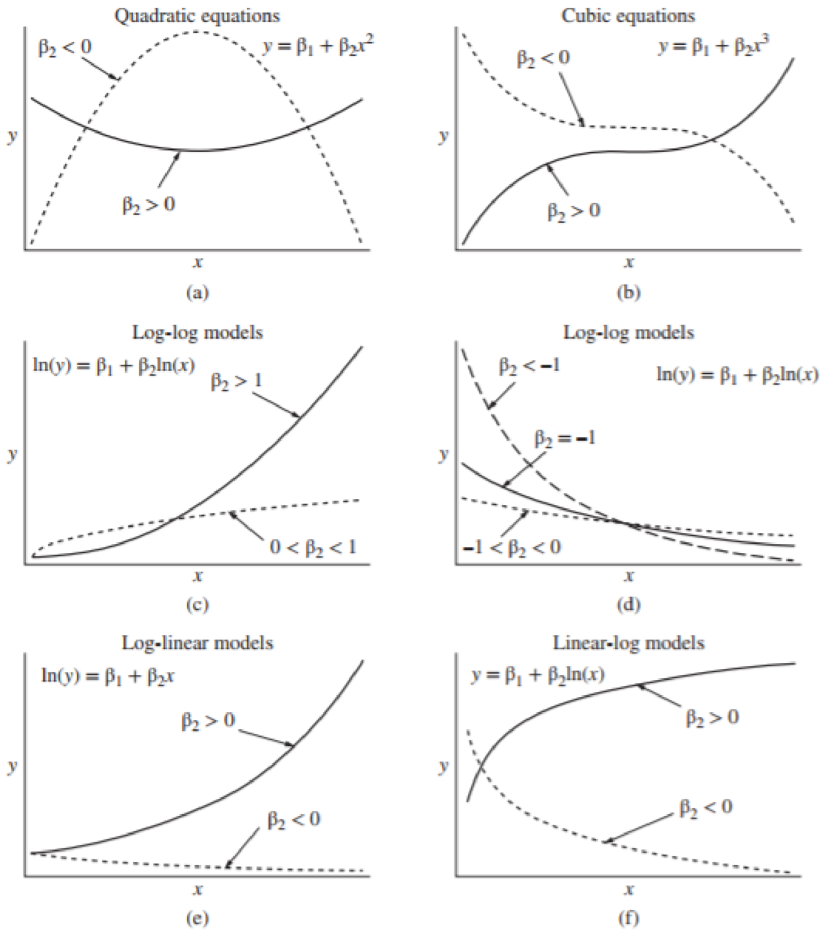


FIGURE 4.5 Alternative functional forms.

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

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 $100\beta_2\%$ 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		

4.3.2 Choosing a Functional Form 3 of 3

- Summary of three configurations:

1. In the log-log model both the dependent and independent variables are transformed by the "natural" logarithm
   - The parameter β_2 is the elasticity of y with respect to *x*
2. In the log-linear model only the dependent variable is transformed by the logarithm
3. In the linear-log model the variable *x* is transformed by the natural logarithm

4.3.3 A Linear-Log Food Expenditure Model 1 of 2

- A linear-log equation has a linear, untransformed term on the left-hand side and a logarithmic term on the right-hand side: $y = \beta_1 + \beta_2 \ln(x)$
  - The elasticity of y with respect to x is: $\varepsilon = \text{slope} \times x/y = \beta_2/y$
  - A convenient interpretation is:
    - The change in y, represented in its units of measure, is approximately β_2 =100 times the percentage change in x

$$\Delta y = y_1 - y_0 = \beta_2 \left[\ln(x_1) - \ln(x_0) \right]$$

$$= \frac{\beta_2}{100} \times 100 \left[\ln(x_1) - \ln(x_0) \right]$$

$$\approx \frac{\beta_2}{100} (\% \Delta x)$$

4.3.3 A Linear-Log Food Expenditure Model 2 of 2

- Given alternative models that involve different transformations of the dependent and independent variables, and some of which have similar shapes, what are some guidelines for choosing a functional form?

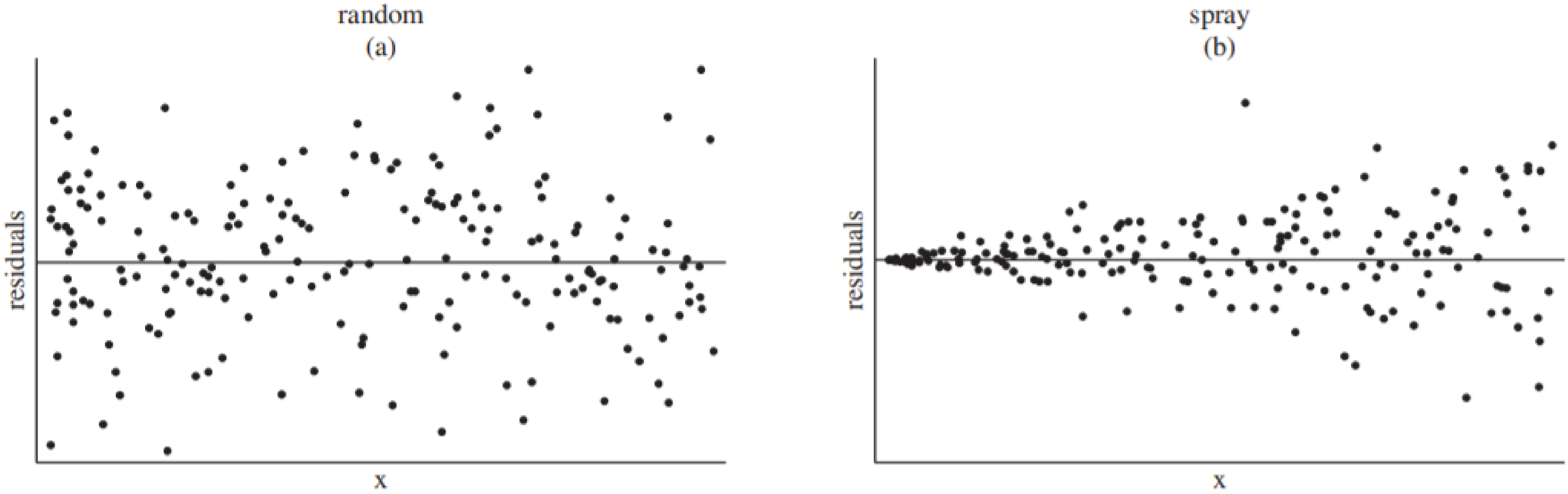
1. Choose a shape that is consistent with what economic theory tells us about the relationship
2. Choose a shape that is sufficiently flexible to “fit” the data
3. Choose a shape so that assumptions SR1–SR6 are satisfied, ensuring that the least squares estimators have the desirable properties described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

4.3.4 Using Diagnostic Residual Plots 1 of 5

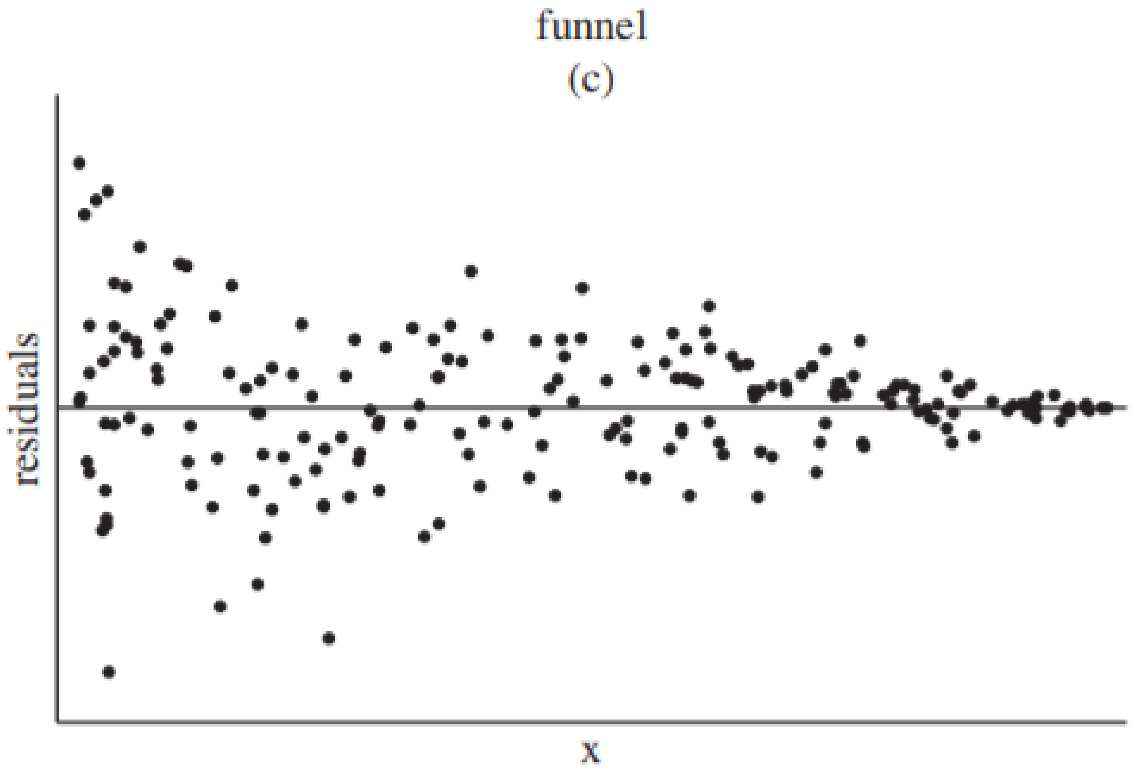
- When specifying a regression model, we may inadvertently choose an inadequate or incorrect functional form

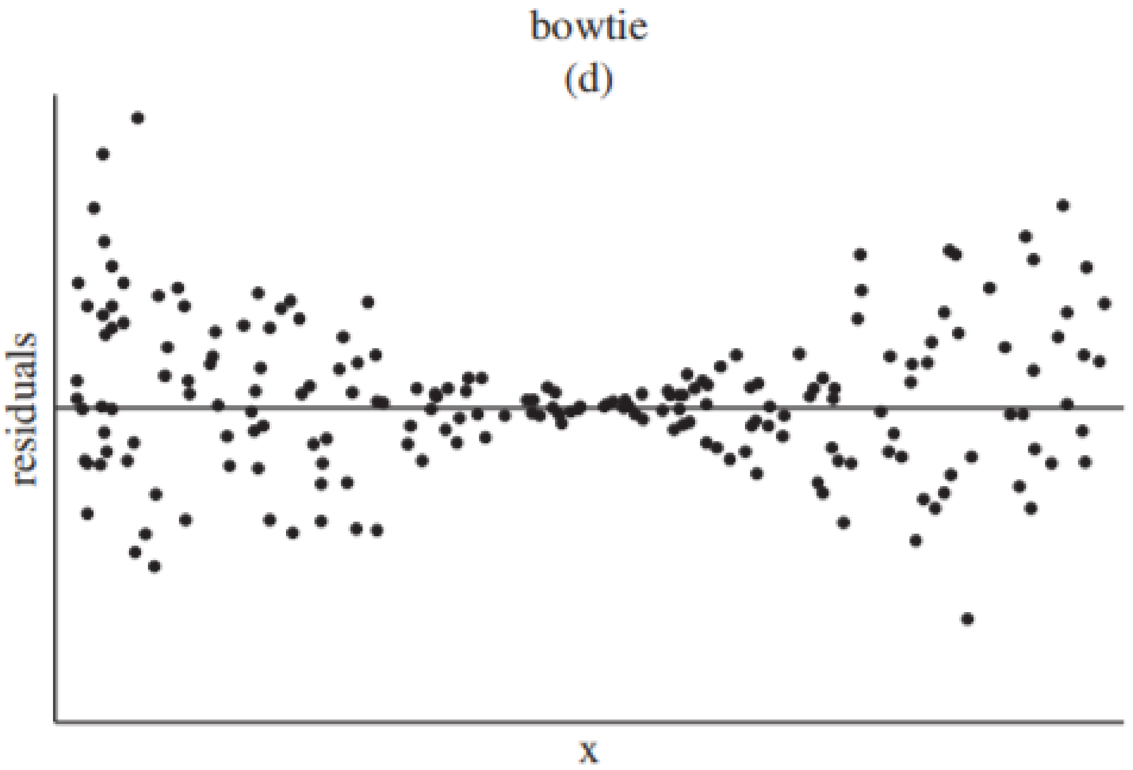
1. Examine the regression results
   - There are formal statistical tests to check for:
     - Homoskedasticity
     - Serial correlation
2. Use residual plots

4.3.4 Using Diagnostic Residual Plots 2 of 5

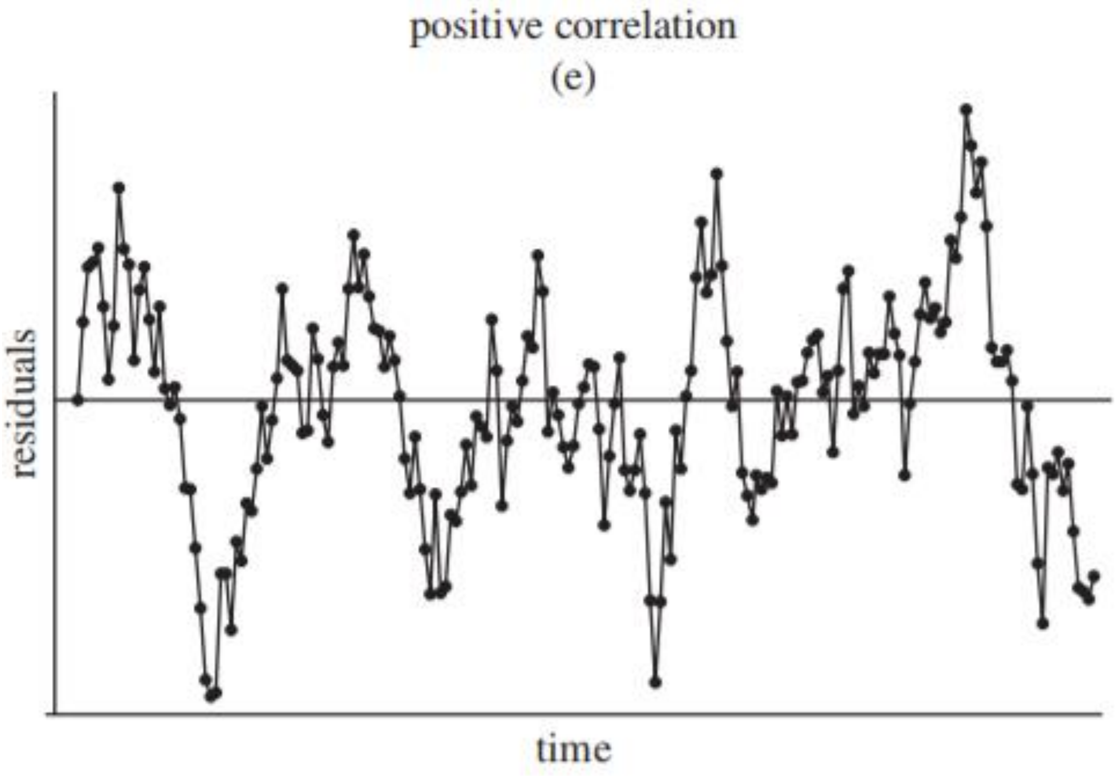


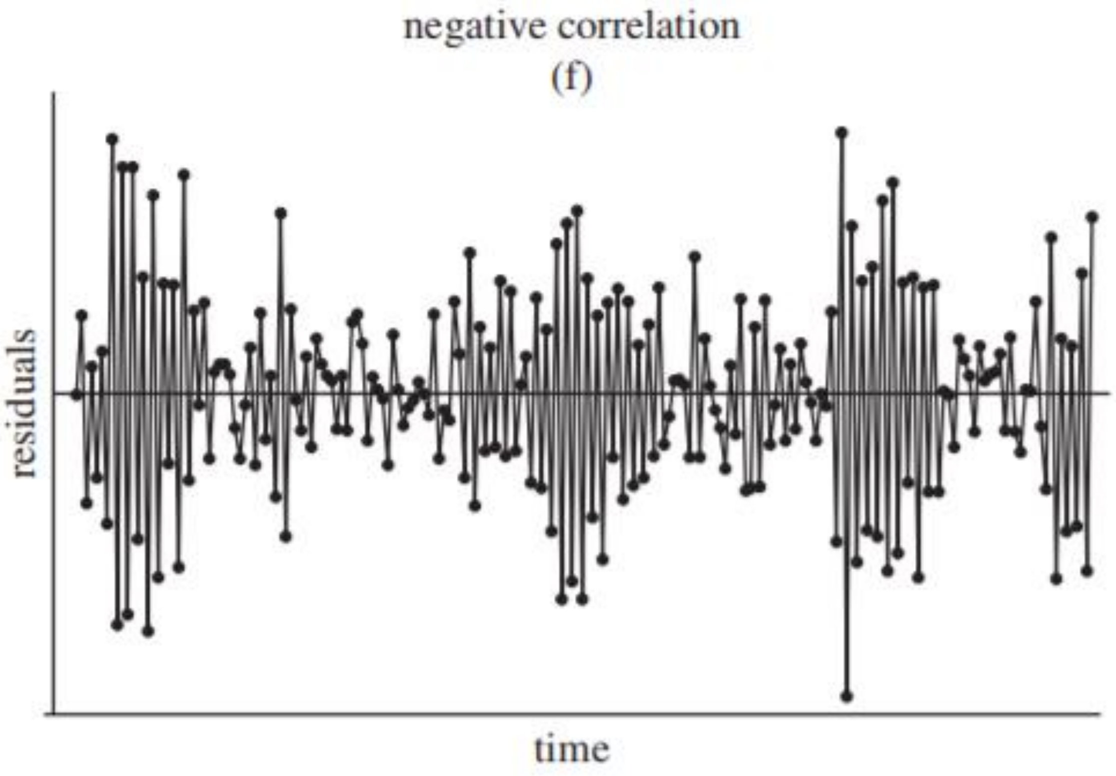
4.3.4 Using Diagnostic Residual Plots 3 of 5



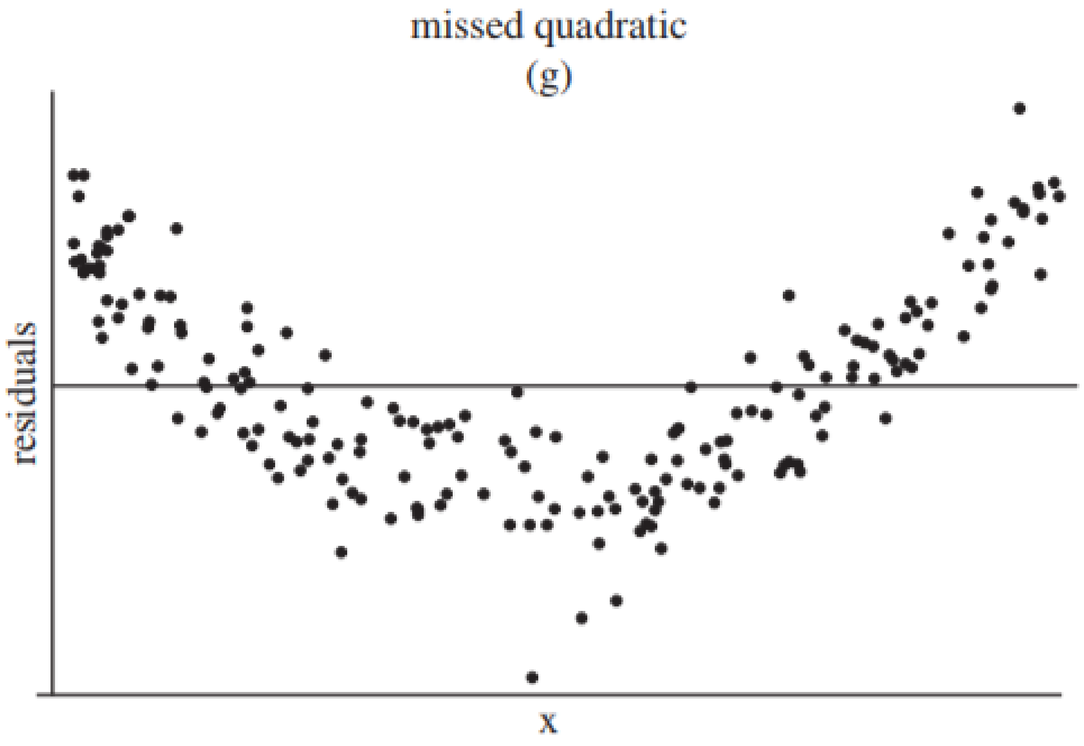


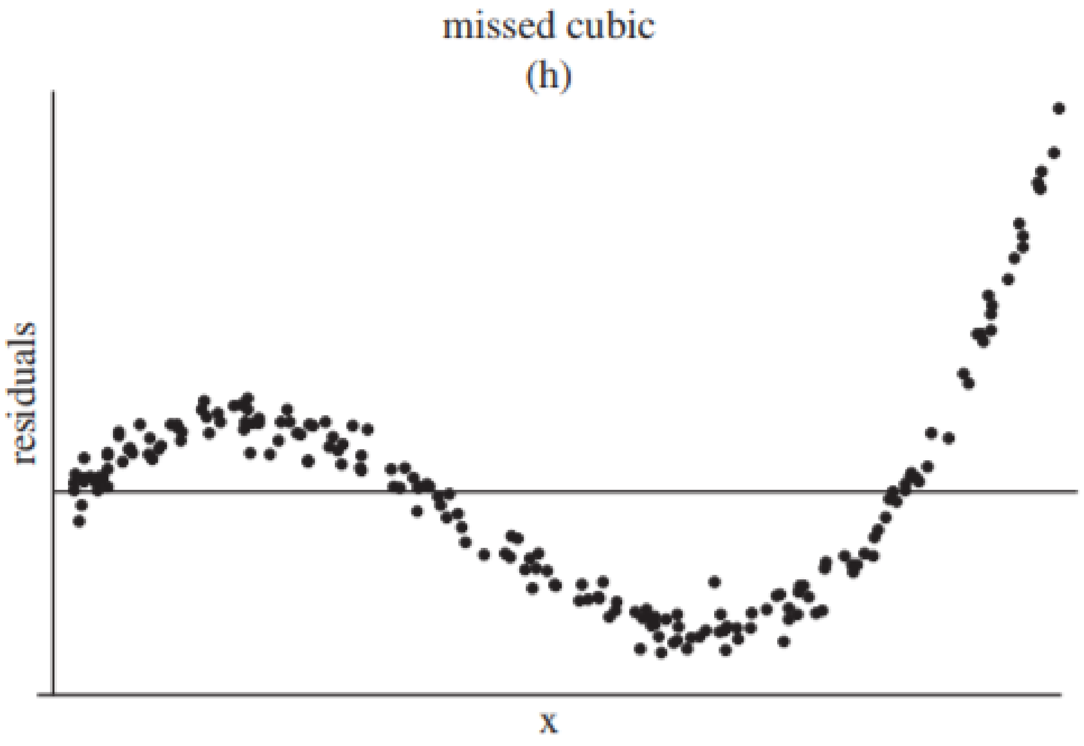
4.3.4 Using Diagnostic Residual Plots 4 of 5





4.3.4 Using Diagnostic Residual Plots 5 of 5





4.3.5 Are the Regression Errors Normally Distributed?

- Hypothesis tests and interval estimates for the coefficients rely on the assumption that the errors, and hence the dependent variable *y*, are normally distributed
- A histogram of the least squares residuals gives us a graphical representation of the empirical distribution
- There are many tests for normality
  - The Jarque–Bera test for normality is valid in large samples
  - It is based on two measures, **skewness** and **kurtosis**

4.3.6 Identifying Influential Observations 1 of 2

- One worry in data analysis is that we may have some unusual and/or **influential observations** Sometimes, these are termed “outliers”
  - If an unusual observation is the result of a data error, then we should correct it
  - Understanding how it came about, the story behind it, can be informative
- One way to detect whether an observation is influential is to delete it and re-estimate the model

4.3.6 Identifying Influential Observations 2 of 2

- The **studentized residual** is the standardized residual based on the delete-one sample
- If the studentized residual falls outside the 95% interval estimate interval, then the observation is worth examining because it is "unusually" large
- Another measure of the influence of a single observation on the least squares estimates is called DFBETAS

4.4 Polynomial Models

- In addition to estimating linear equations, we can also estimate quadratic and cubic equations
- Economics students will have seen many average and marginal cost curves (U-shaped) and average and marginal product curves (inverted-U shaped) in their studies

4.4.1 Quadratic and Cubic Equations

- The general form of a quadratic equation is:

$$y = a_0 + a_1 x + a_2 x^2$$

- The general form of a cubic equation is:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

- A problem with the linear equation is that it implies an increase at the same constant rate, when one might expect a rate to be increasing
- Polynomial models may provide a better fit

4.5 Log-Linear Models 1 of 2

- 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

4.5 Log-Linear Models 2 of 2

- 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 β_2 % change in y

$$100\left[\ln\left(y_1\right)-\ln\left(y_0\right)\right] \approx \%\Delta\, y = 100\beta_2\left(x_1 - x_0\right) = \left(100\beta_2\right)\times\Delta x$$

4.5.1 Prediction in the Log-Linear Model 1 of

- 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

4.5.1 Prediction in the Log-Linear Model 2 of

- 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:
    - $\hat{y}_c = \widehat{E(y)} = \exp(b_1 + b_2 x + \hat{\sigma}^2/2) = \hat{y}_n{}^{e^\wedge\left(\hat{\sigma}^2/2\right)}$
    - The subscript "*c*" is for "corrected"
    - This uses the properties of the **log-normal distribution**

4.5.1 Prediction in the Log-Linear Model 3 of

- 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 under predict the value of y in a log-linear model, and the correction offsets the downward bias in large samples

Example 4 .11 Prediction in a Log-linear Model

- The wage equation:
  - $\ln(\widehat{WAGE})$= 1.5968 + 0.0988 × EDUC = 1.5968 + 0.0988 × 12 = 2.7819
- The natural predictor is: $\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(2.7819) = 16.1493$
- The corrected predictor is:

$$\hat{y}_c = \widehat{E(y)} = \exp(b_1 + b_2 x + \hat{\sigma}^2/2) = \hat{y}_n{}^{e^\wedge(\hat{\sigma}^2/2)} = 16.1493 \times 1.1246 = 18.1622$$

Figure 4.13 The natural and corrected predictors of wage

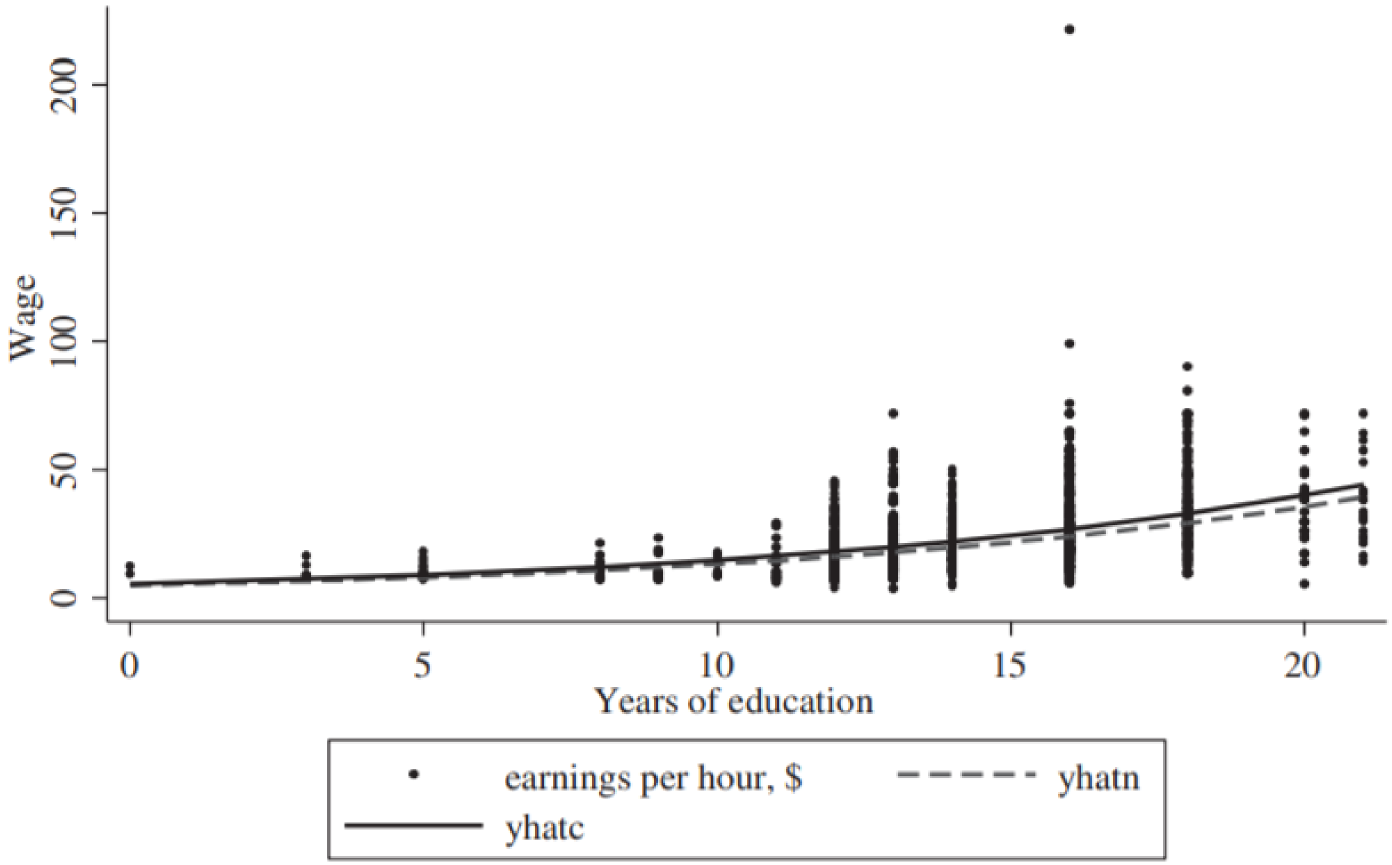


FIGURE 4.13 **The natural and corrected predictors of wage.**

4.5.2 A Generalized R^2 Measure

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

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

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

$$\mathrm{R}_{\mathrm{g}}^2 = \left[\text{corr}(\mathrm{y}, \hat{\mathrm{y}})\right]^2 = 0.4647^2 = 0.2159$$

- Compare this to the reported $R^2 = 0.2577$

4.5.3 Prediction Intervals in the Log-Linear Model

- If we prefer a prediction or forecast interval over a "point" predictor for y, then we must rely on the natural predictor y^n

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

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

Example 4.12 Prediction Intervals for a Log-linear Model

- For the wage equation, a 95% prediction interval for the wage of a worker with 12 years of education is:
- $[\exp(2.7819 - 1.96\ x\ 0.4850), \exp(2.7819 + 1.96 \times 0.4850)] = [6.2358, 41.8233]$
- The interval prediction is \$6.24–\$41.82, which is so wide that it is basically useless
- Our model is not an accurate predictor of individual behavior in this case

Figure 4.14 The 95% prediction for wage

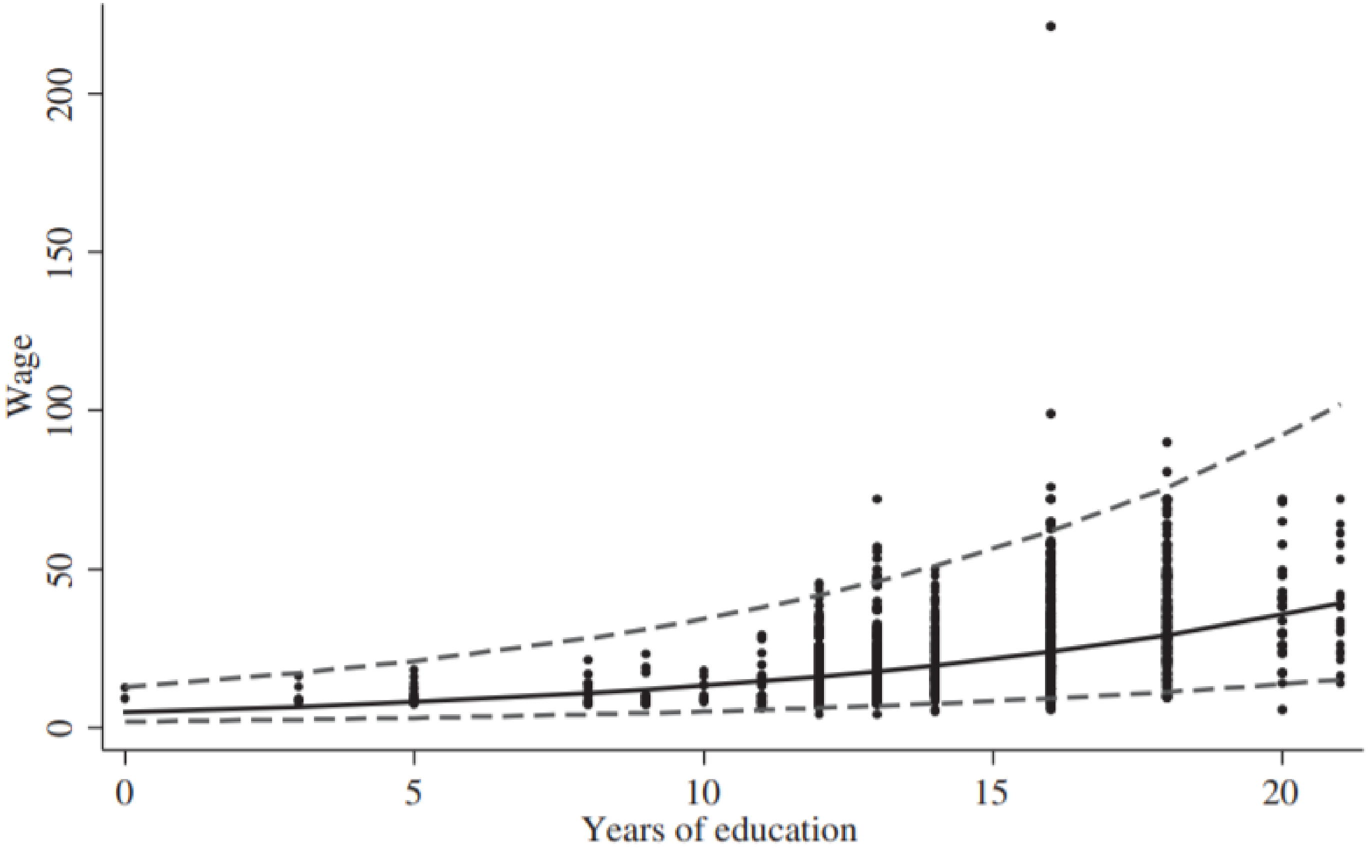


FIGURE 4.14 **The 95% prediction interval for wage.**

4.6 Log-Log Models 1 of 2

- 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

4.6 Log-Log Models 2 of 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

Example 4.13 A Log-log Poultry Demand Equation 1 of 2

- The estimated model is:
  - (4.15) $\widehat{\ln(Q)} = 3.717 - 1.121 \times \ln(\text{P})$ $R_g^2 = 0.8817$

    (se) (0.022) (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%

Example 4.13 A Log-log Poultry Demand Equation 2 of 2

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

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

- The generalized goodness-of-fit is: $R_g^2 = \left[\operatorname{corr}\left(Q, \hat{Q}_c\right)\right]^2 = 0.939^2 = 0.8817$

Figure 4.15 Quantity and price of Chicken.

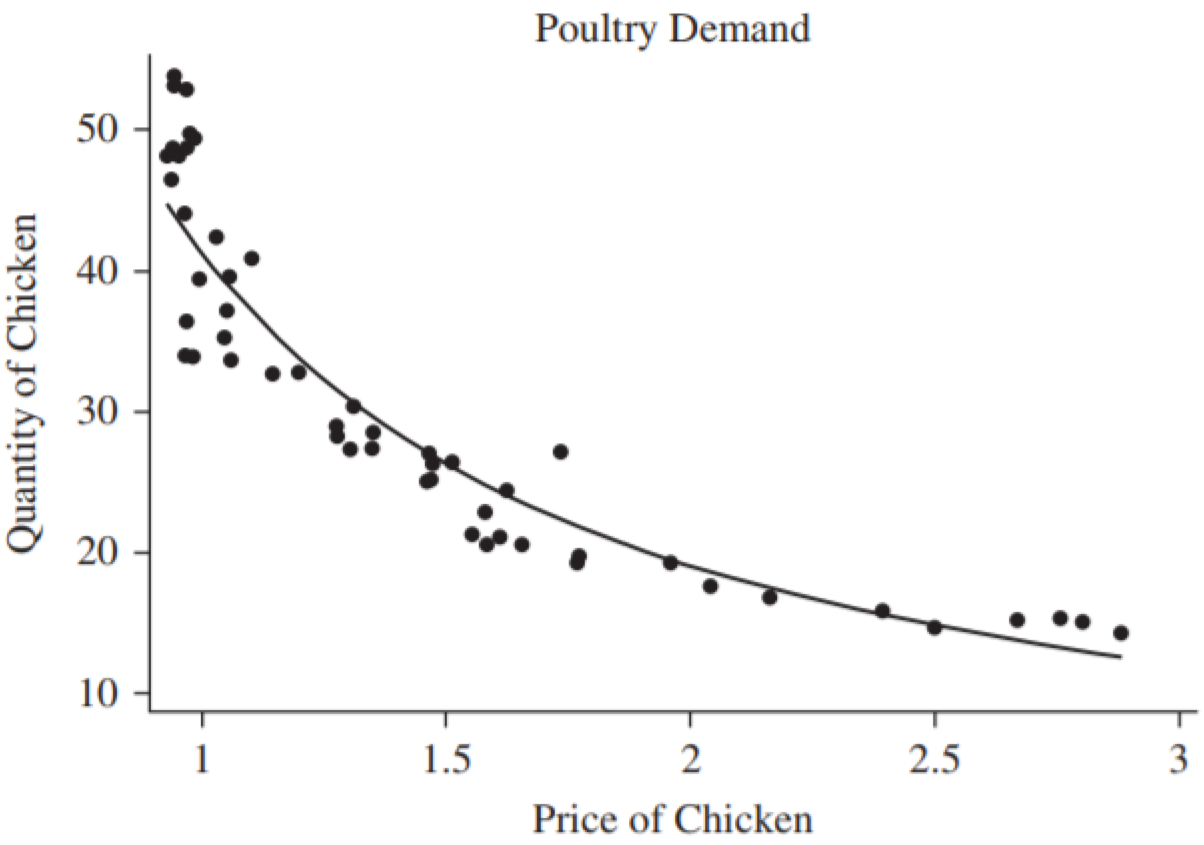


FIGURE 4.15 **Quantity and price of chicken.**

Key Words

- coefficient of determination
- correlation
- forecast error
- functional form
- goodness-of-fit
- growth model
- influential observations
- 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 diagnostics
- scaling data
- skewness
- standard error of the forecast

Copyright