4.1 Least Squares Prediction

\[ y_i = \beta_0 + \beta_2 x_i + e_i \]  \hspace{1cm} (4.1)

where \( e_i \) is a random error. We assume that \( E(y_i) = \beta_0 + \beta_2 x_i \) and \( E(e_i) = 0 \). We also assume that \( \text{var}(e_i) = \sigma^2 \) and \( \text{cov}(e_i, e_j) = 0 \) for \( i, j = 1, 2, \ldots, N \).

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_i \]  \hspace{1cm} (4.2)
4.1 Least Squares Prediction

\[ f = y - \hat{y} = (\hat{\beta}_1 + \hat{\beta}_2 x) - (\hat{\beta}_1 + \hat{\beta}_2 x) \]

\[ E(f) = \hat{\beta}_1 + \hat{\beta}_2 x + E(e) - \left[ E(\hat{\beta}_1) + E(\hat{\beta}_2) x \right] \]

\[ = \hat{\beta}_1 + \hat{\beta}_2 x + 0 - \left[ \hat{\beta}_1 + \hat{\beta}_2 x \right] = 0 \]

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

The variance of the forecast error is smaller when:

i. the overall uncertainty in the model is smaller, as measured by the variance of the random errors;
ii. the sample size \( N \) is larger;
iii. the variation in the explanatory variable is larger; and
iv. the value of \( \sigma \) is small.
4.1 Least Squares Prediction

\[
\hat{\sigma}^2 = \frac{1}{N} \sum (e_i - \overline{e})^2
\]

\[
\text{se}(f) = \sqrt{\text{var}(f)}
\]

\[
\hat{y}_i \pm t_{a/2, N-n} \text{se}(f)
\]

Figure 4.2 Point and interval prediction

4.1.1 Prediction in the Food Expenditure Model

\[
\hat{y}_i = b_0 + b_1 x_i = 83.4160 + 10.2096(20) = 287.6089
\]

\[
\text{var}(f) = \hat{\sigma}^2 \left[ 1 + \frac{1}{N} \left( \frac{\sum (x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} \right) \right]
\]

\[
= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} \left( \frac{\sum (x_i - \overline{x})^2}{\sum (x_i - \overline{x})^2} \right)
\]

\[
= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} \left( \frac{1}{N} \sum (x_i - \overline{x})^2 \right) \text{var}(b_0)
\]

\[
\hat{y}_i \pm t_{a/2, N-n} \text{se}(f) = 287.6069 \pm 2.0244(90.6328) = [104.1323, 471.0854]
\]
4.2 Measuring Goodness-of-Fit

\[ y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \epsilon_i \]  
\[ y_i = E(y_i) + \epsilon_i \]  
\[ y_i = \hat{y}_i + \hat{\epsilon}_i \]  
\[ y_i - \overline{y} = (\hat{y}_i - \overline{y}) + \hat{\epsilon}_i \]

Figure 4.3 Explained and unexplained components of \( y_i \)

\[ \hat{\sigma}^2 = \frac{\sum (y_i - \overline{y})^2}{N-1} \]
\[ \sum (y_i - \overline{y})^2 = \sum (\hat{y}_i - \overline{y})^2 + \sum \hat{\epsilon}_i^2 \]
4.2 Measuring Goodness-of-Fit

1. \( \sum (y_i - \bar{y})^2 = \text{total sum of squares} = \text{SST} \): a measure of total variation in \( y \) about the sample mean.
2. \( \sum (y_i - \bar{y})^2 = \text{sum of squares due to the regression} = \text{SSR} \): that part of total variation in \( y \) about the sample mean, that is explained by, or due to, the regression. Also known as the “explained sum of squares.”
3. \( \sum e_i^2 = \text{sum of squares due to error} = \text{SSE} \): that part of total variation in \( y \) about its mean that is not explained by the regression. Also known as the unexplained sum of squares, the residual sum of squares, or the sum of squared errors.
4. \( \text{SST} = \text{SSR} + \text{SSE} \)

\[ R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \] (4.12)

- The closer \( R^2 \) is to one, the closer the sample values \( y_i \) are to the fitted regression equation \( \hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i \). If \( R^2 = 1 \), then all the sample data fall exactly on the fitted least squares line, so \( \text{SSE} = 0 \), 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,” so that \( \text{SSR} = 0 \) and \( R^2 = 0 \).

4.2.1 Correlation Analysis

\[ \rho = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x) \text{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \] (4.13)

\[ \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \] (4.14)

\[ \hat{\sigma}_x = \hat{\sigma}_y = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}} \] (4.15)
4.2.2 Correlation Analysis and $R^2$

\[ r_{xy}^2 = R^2 \]
\[ R^2 = \frac{s_{\hat{y}}^2}{s^2} \]

$R^2$ measures the linear association, or goodness-of-fit, between the sample data and their predicted values. Consequently $R^2$ is sometimes called a measure of "goodness-of-fit."

4.2.3 The Food Expenditure Example

\[ \text{SST} = \sum (y_i - \bar{y})^2 = 495132.160 \]
\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum e^2 = 304505.176 \]
\[ R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{304505.176}{495132.160} = .385 \]
\[ r_{xy} = \frac{s_{\hat{y}}}{s_{y}} = \frac{478.75}{6.848} = \frac{112.675}{6.848} = .62 \]

4.2.4 Reporting the Results

Figure 4.4 Plot of predicted $\hat{y}$ against $y$
4.2.4 Reporting the Results

- FOOD_EXP = weekly food expenditure by a household of size 3, in dollars
- INCOME = weekly household income, in $100 units

\[
\text{FOOD}\_\text{EXP} = 83.42 + 10.21\text{INCOME} \quad R^2 = 0.385
\]

* indicates significant at the 10% level
** indicates significant at the 5% level
*** indicates significant at the 1% level

4.3 Modeling Issues

- 4.3.1 The Effects of Scaling the Data
  - Changing the scale of \( x \):
    \[
    y = \beta_0 + \beta_1 x + e = \beta_0 + (\beta_1)(x/c) + e
    \]
    where \( \beta_0' = \beta_0 \) and \( x' = x/c \)
  - Changing the scale of \( y \):
    \[
    y/c = (\beta_0'/c) + (\beta_1'/c)x + e/c \quad \text{or} \quad y' = \beta_0' + \beta_1'x + e'
    \]

4.3.2 Choosing a Functional Form

Variable transformations:
- Power: if \( x \) is a variable then \( x^p \) means raising the variable to the power \( p \); examples are quadratic (\( x^2 \)) and cubic (\( x^3 \)) transformations.
- The natural logarithm: if \( x \) is a variable then its natural logarithm is \( \ln(x) \).
- The reciprocal: if \( x \) is a variable then its reciprocal is \( 1/x \).
4.3.2 Choosing a Functional Form

- The log-log model
  \[ \ln(y) = \beta_0 + \beta_1 \ln(x) \]
  The parameter \( \beta \) is the elasticity of \( y \) with respect to \( x \).

- The log-linear model
  \[ \ln(y) = \beta_0 + \beta_1 x \]
  A one-unit increase in \( x \) leads to (approximately) a \( 100 \beta \) percent change in \( y \).

- The linear-log model
  \[ y = \beta_0 + \beta_1 \ln(x) \]
  A 1% increase in \( x \) leads to a \( \frac{\beta_1}{100} \) unit change in \( y \).

4.3.3 The Food Expenditure Model

- The reciprocal model is
  \[ \text{FOOD\_EXP} = \beta_0 + \beta_1 \frac{1}{\text{INCOME}} + \epsilon \]

- The linear-log model is
  \[ \text{FOOD\_EXP} = \beta_0 + \beta_1 \ln(\text{INCOME}) + \epsilon \]
### 4.3.3 The Food Expenditure Model

**Remark:** Given this array of 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 Are the Regression Errors Normally Distributed?

- The Jarque-Bera statistic is given by
  \[
  JB = \frac{N}{6} \left( S^2 + \frac{(K-3)^2}{4} \right)
  \]

  where \( N \) is the sample size, \( S \) is skewness, and \( K \) is kurtosis.

- In the food expenditure example
  \[
  JB = \frac{40}{6} \left( -0.95^2 + \frac{(2.99 - 3)^2}{4} \right) = 0.63
  \]
### 4.3.5 Another Empirical Example

**Figure 4.7** Scatter plot of wheat yield over time

\[
YIELD_t = \beta_0 + \beta_1 \text{TIME}_t + \epsilon_t
\]

\[
YIELD_t = .638 + .0210 \text{ TIME}_t, \quad R^2 = .649
\]

(\text{se}) \quad (.064) (.0022)

**Figure 4.8** Predicted, actual and residual values from straight line

*Principles of Econometrics, 3rd Edition* Slide 4-28
4.3.5 Another Empirical Example

Figure 4.9 Bar chart of residuals from straight line

Slide 4-31 Principles of Econometrics, 3rd Edition

\[
YIELD_t = \beta_1 + \beta_2 \text{TIME}_t + \epsilon_t
\]

\[
\text{TIMECUBE} = \text{TIME}_t^3 / 1000000
\]

\[
\text{YIELD} = 0.874 + 9.68 \text{TIMECUBE}, \quad R^2 = 0.751
\]

(see) (.036) (.082)

Slide 4-32 Principles of Econometrics, 3rd Edition

4.3.5 Another Empirical Example

Figure 4.10 Fitted, actual and residual values from equation with cubic term

Slide 4-33 Principles of Econometrics, 3rd Edition
4.4 Log-Linear Models

4.4.1 The Growth Model

\[ \ln(YIELD_t) = \ln(YIELD_0) + \ln(1 + g)t \]
\[ = \beta_1 + \beta_2 t \]

\[ \ln(YIELD_t) = -0.3434 + 0.0178t \]

(\text{se}) \quad (0.0584) \quad (0.0021)

4.4.2 A Wage Equation

\[ \ln(WAGE) = \ln(WAGE_0) + \ln(1 + r) \times EDUC \]
\[ = \beta_1 + \beta_2 \times EDUC \]

\[ \ln(WAGE) = 0.7884 + 0.1038 \times EDUC \]

(\text{se}) \quad (0.0849) \quad (0.0063)

4.4.3 Prediction in the Log-Linear Model

\[ \hat{y}_t = \exp(\ln(y)) \exp(\hat{\beta}_1 + \hat{\beta}_2 x) \]
\[ \hat{y}_t = \hat{E}(y) = \exp(\hat{\alpha} + \hat{\beta}_2 x + \hat{\sigma}^2/2) = \bar{y}_t \times e^{\hat{\sigma}^2/2} \]

\[ \ln(WAGE) = 0.7884 + 0.1038 \times EDUC = 0.7884 + 0.1038 \times 12 = 2.0335 \]
\[ \hat{y}_t = \hat{E}(y) = \bar{y}_t \times e^{\hat{\sigma}^2/2} = 7.6408 \times 1.1276 = 8.6161 \]
4.4 Log-Linear Models

4.4.4 A Generalized R² Measure

\[ R^2 = [\text{corr}(y, \hat{y})]^2 = r_{y, \hat{y}}^2 \]

\[ R^2 = [\text{corr}(y, \hat{y})]^2 = 0.4739^2 = 0.2246 \]

R² values tend to be small with microeconomic, cross-sectional data, because the variations in individual behavior are difficult to fully explain.

4.4 Log-Linear Models

4.4.5 Prediction Intervals in the Log-Linear Model

\[ \exp\left[\ln(y) + t \cdot \text{se}(f)\right], \exp\left[\ln(y) + t \cdot \text{se}(f)\right] \]

\[ \left[\exp(2.0335 - 1.96 \cdot .4905), \exp(2.0335 + 1.96 \cdot .4905)\right] = [2.9184, 20.0046] \]

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
- R²
- residual
- skewness
Chapter 4 Appendices

- Appendix 4A Development of a Prediction Interval
- Appendix 4B The Sum of Squares Decomposition
- Appendix 4C The Log-Normal Distribution

Appendix 4A
Development of a Prediction Interval

\[ f = y^* - \hat{y} = (\beta_1 + \beta_2 x_i + \epsilon_i) - (\hat{\beta}_1 + \hat{\beta}_2 x_i) \]

\[
\text{var}(\hat{y}_i) = \text{var}(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \text{var}(\hat{\beta}_1) + x_i^2 \text{var}(\hat{\beta}_2) + 2x_i \text{cov}(\hat{\beta}_1, \hat{\beta}_2)
\]

\[
= \frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2} + x_i^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} + 2x_i \sigma^2 \frac{-\bar{x}}{\sum (x_i - \bar{x})^2}
\]

Appendix 4A
Development of a Prediction Interval

\[
\text{var}(\hat{y}_i) = \left[ \frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2} - \frac{\sigma^2 \bar{x}}{N \sum (x_i - \bar{x})^2} \right] \left[ \frac{\sigma^2}{\sum (x_i - \bar{x})^2} - \frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} - \frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} \right]
\]

\[
= \sigma^2 \left[ \frac{\sum (x_i - \bar{x})^2}{N \sum (x_i - \bar{x})^2} \right] \left[ \frac{\Sigma I - \lambda x_i \bar{x}}{\Sigma (x_i - \bar{x})^2} \right]
\]

\[
= \sigma^2 \left[ \frac{\bar{x} \Sigma (x_i - \bar{x})^2}{N \sum (x_i - \bar{x})^2} \right]
\]

\[
= \sigma^2 \left[ \frac{1}{\sum (x_i - \bar{x})^2} \right]
\]
Appendix 4A
Development of a Prediction Interval

\[
\frac{f}{\sqrt{\text{var}(f)}} - N(0,1) \\
\text{var}(f) = \delta_i^2 \left[ 1 + \frac{1}{N} \sum (x_i - \bar{x})^2 \right] \\
\frac{f}{\sqrt{\text{var}(f)}} = \frac{\hat{y}_i - \bar{y}}{\hat{\text{se}}(f)} \tag{4A.1} \\
P(-t_{\alpha/2} \leq \frac{f}{\sqrt{\text{var}(f)}} \leq t_{\alpha/2}) = 1 - \alpha \tag{4A.2}
\]

Appendix 4A
Development of a Prediction Interval

\[
P[-t_{\alpha/2} \leq \frac{\hat{y}_i - \bar{y}}{\hat{\text{se}}(f)} \leq t_{\alpha/2}] = 1 - \alpha \\
P[\hat{y}_i - t_{\alpha/2} \hat{\text{se}}(f) \leq \hat{y}_i \leq \hat{y}_i + t_{\alpha/2} \hat{\text{se}}(f)] = 1 - \alpha
\]

Appendix 4B
The Sum of Squares Decomposition

\[
(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + \hat{e}_i^2 + 2(\hat{y}_i - \bar{y})\hat{e}_i \\
\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2 + 2\sum (\hat{y}_i - \bar{y})\hat{e}_i \\
\sum (\hat{y}_i - \bar{y})\hat{e}_i = \sum \hat{e}_i \hat{y}_i - \bar{y} \sum \hat{e}_i = \sum (h_i + b_i x_i) \hat{e}_i - \bar{y} \sum \hat{e}_i \\
= h \sum \hat{e}_i + b \sum x \hat{e}_i - \bar{y} \sum \hat{e}_i
\]
**Appendix 4B**

**The Sum of Squares Decomposition**

\[
\sum \hat{e}_i = \sum (y_i - \hat{b}_1 x_i) = \sum y_i - \hat{b}_1 \sum x_i = 0
\]

\[
\sum x_i \hat{e}_i = \sum x_i (y_i - \hat{b}_1 x_i) = \sum x_i y_i - \hat{b}_1 \sum x_i^2 = 0
\]

\[
\sum (\hat{e}_i - \bar{y})\bar{e}_i = 0
\]

If the model contains an intercept it is guaranteed that \( SST = SSR + SSE \).

If, however, the model does not contain an intercept, then \( \sum \hat{e}_i \neq 0 \) and \( SST \neq SSR + SSE \).

---

**Appendix 4C**

**The Log-Normal Distribution**

Suppose that the variable \( y \) has a normal distribution, with mean \( \mu \) and variance \( \sigma^2 \).

If we consider \( w = \ln(y) \), then \( y \sim N(\mu, \sigma^2) \) is said to have a log-normal distribution.

\[
E(w) = e^{\mu + \frac{\sigma^2}{2}}
\]

\[
\text{var}(w) = e^{2\mu + \sigma^2} \left( e^{\sigma^2} - 1 \right)
\]
Appendix 4C
The Log-Normal Distribution

The growth and wage equations:

\[ \beta_i = \ln(1 + r) \quad \text{and} \quad r = e^{\beta_i} - 1 \]

\[ \beta_i \sim N(\beta, \text{var}(\beta_i) = \sigma^2 / \sum(x_i - \mu)^2) \]

\[ E[\exp(\beta_i)] = e^{\beta_i + \text{var}(\beta_i) / 2} \]

\[ \hat{r} = e^{\beta_i + \text{var}(\beta_i) / 2} - 1 \]

\[ \text{var}(\beta_i) = \hat{\sigma}^2 / \sum(x_i - \mu)^2 \]