Prediction, Goodness-of-Fit, and Modeling Issues

Chapter 4

Prepared by Vera Tabakova, East Carolina University

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

- 4.1 Least Squares Prediction
- 4.2 Measuring Goodness-of-Fit
- 4.3 Modeling Issues
- 4.4 Log-Linear Models

4.1 Least Squares Prediction

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \tag{4.1}$$

where e_0 is a random error. We assume that $E(y_0) = \beta_1 + \beta_2 x_0$ and $E(e_0) = 0$. We also assume that $\text{var}(e_0) = \sigma^2$ and $\text{cov}(e_0, e_i) = 0 \quad i = 1, 2, \ldots, N$

$$\hat{y}_0 = b_1 + b_2 x_0 \tag{4.2}$$

4.1 Least Squares Prediction

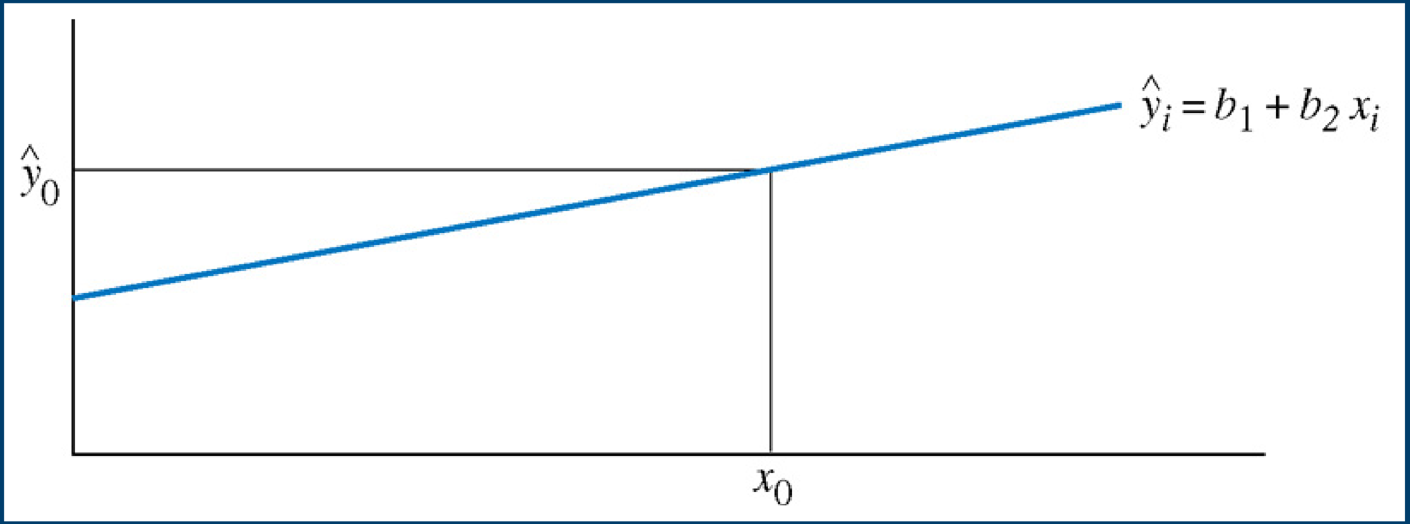

Figure **4.1** A point prediction

4.1 Least Squares Prediction

$$f = y_0 - \hat{y}_0 = \left(\beta_1 + \beta_2 x_0 + e_0\right) - \left(b_1 + b_2 x_0\right) \qquad (4.3)$$

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

$$= \beta_1 + \beta_2 x_0 + 0 - \left[\beta_1 + \beta_2 x_0\right] = 0$$

$$\operatorname{var}(f) = \sigma^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right] \qquad (4.4)$$

4.1 Least Squares Prediction

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 is small.

4.1 Least Squares Prediction

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

$$\text{se}(f) = \sqrt{\widehat{\text{var}}(f)} \qquad (4.5)$$

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

4.1 Least Squares Prediction

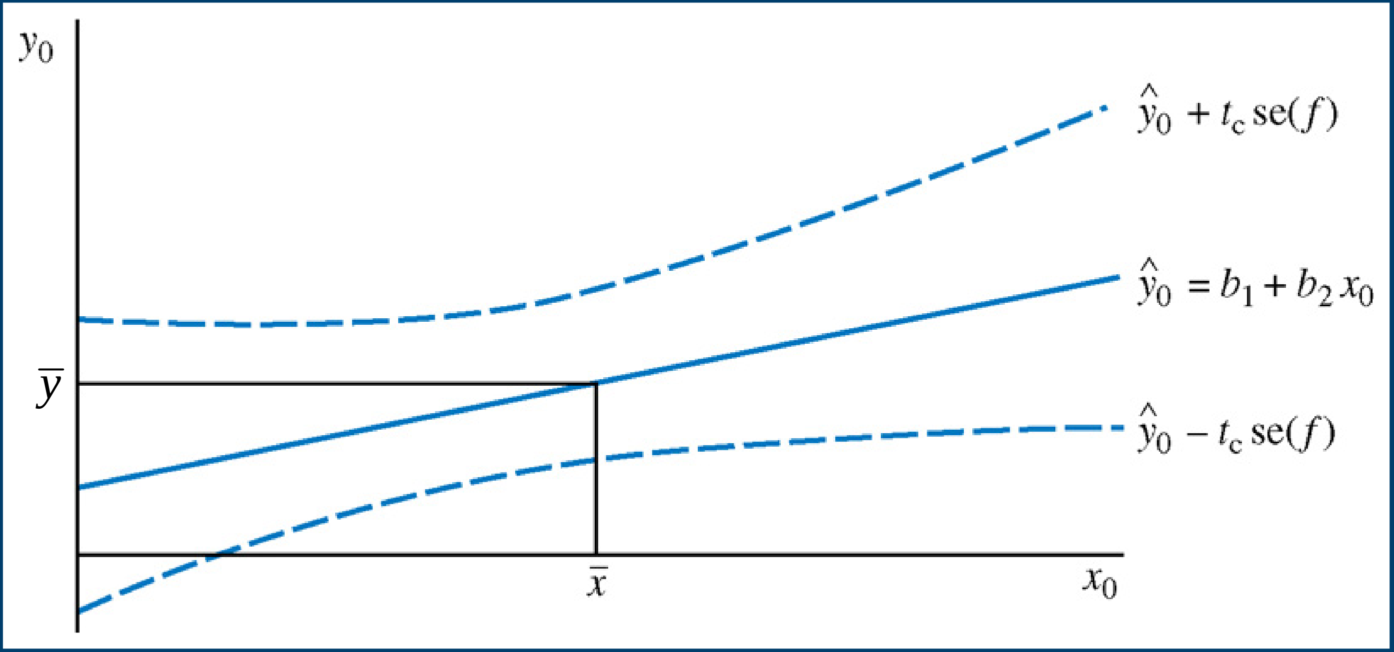

Figure **4.2** Point and interval prediction

4.1.1 Prediction in the Food Expenditure Model

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

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

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

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

$$\hat{y}_0 \pm t_c \text{se}(f) = 287.6069 \pm 2.0244(90.6328) = \left[104.1323, 471.0854\right]$$

4.2 Measuring Goodness-of-Fit

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{4.7}$$

$$y_i = E(y_i) + e_i \tag{4.8}$$

$$y_i = \hat{y}_i + e_i \tag{4.9}$$

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + e_i \tag{4.10}$$

4.2 Measuring Goodness-of-Fit

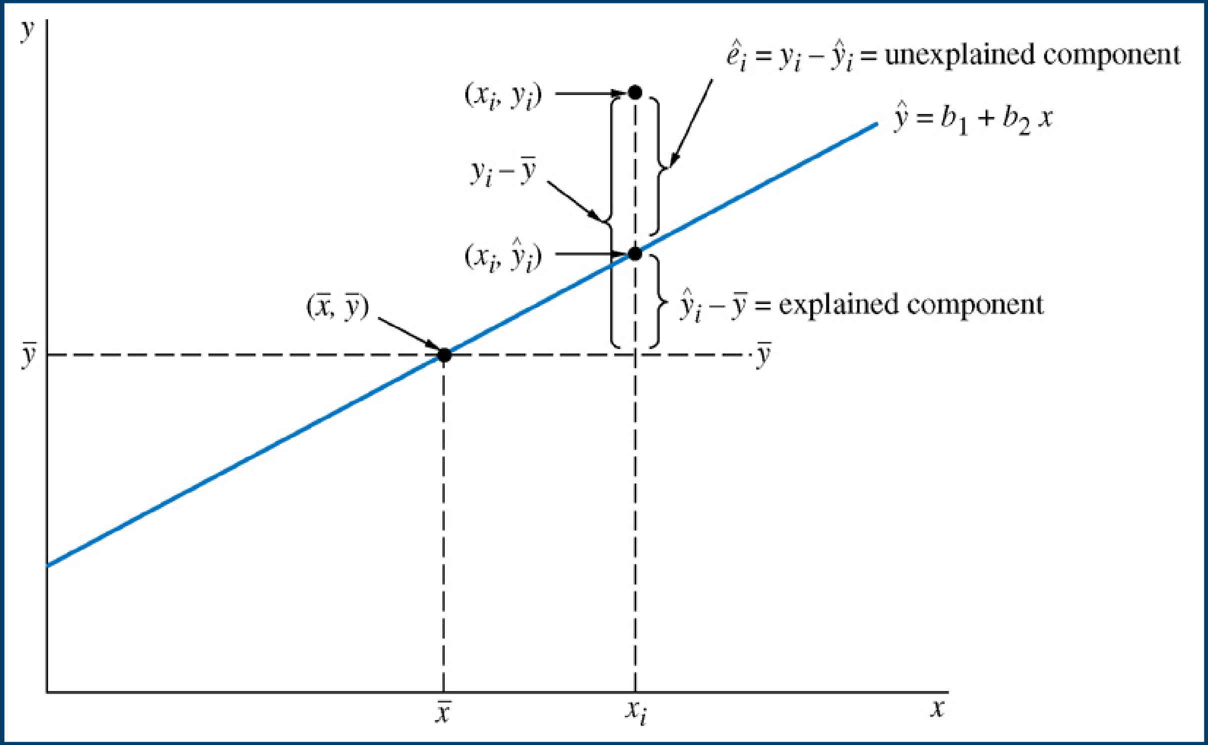

Figure **4.3** Explained and unexplained components of y_i

4.2 Measuring Goodness-of-Fit

$$\hat{\sigma}_y^2 = \frac{\sum \left(y_i - \bar{y}\right)^2}{N - 1}$$

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 \qquad (4.11)$$

4.2 Measuring Goodness-of-Fit

- $\sum(y_i - \bar{y})^2$ = total sum of squares = *SST*: a measure of *total variation* in *y* about the sample mean.
- $\sum(\hat{y}_i - \bar{y})^2$ = sum of squares due to the regression = *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."
- $\sum \hat{e}_i^2$ = sum of squares due to error = *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.
- $SST = SSR + SSE$

4.2 Measuring Goodness-of-Fit

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (4.12)$$

- The closer R^2 is to one, the closer the sample values y_i are to the fitted regression equation $\hat{y}_i = b_1 + b_2 x_i$. 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," so that $SSR = 0$ and $R^2 = 0$.

4.2.1 Correlation Analysis

$$\rho_{xy} = \frac{\operatorname{cov}(x,y)}{\sqrt{\operatorname{var}(x)}\sqrt{\operatorname{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.13}$$

$$r_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \tag{4.14}$$

$$\begin{aligned} \hat{\sigma}_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \big/ (N-1) \\ \hat{\sigma}_x &= \sqrt{\sum (x_i - \bar{x})^2 \big/ (N-1)} \\ \hat{\sigma}_y &= \sqrt{\sum (y_i - \bar{y})^2 \big/ (N-1)} \end{aligned} \tag{4.15}$$

4.2.2 Correlation Analysis and R^2

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

$$R^2 = r_{y\hat{y}}^2$$

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

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

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

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{304505.176}{495132.160} = .385$$

$$r_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{478.75}{\left(6.848\right)\left(112.675\right)} = .62$$

4.2.4 Reporting the Results

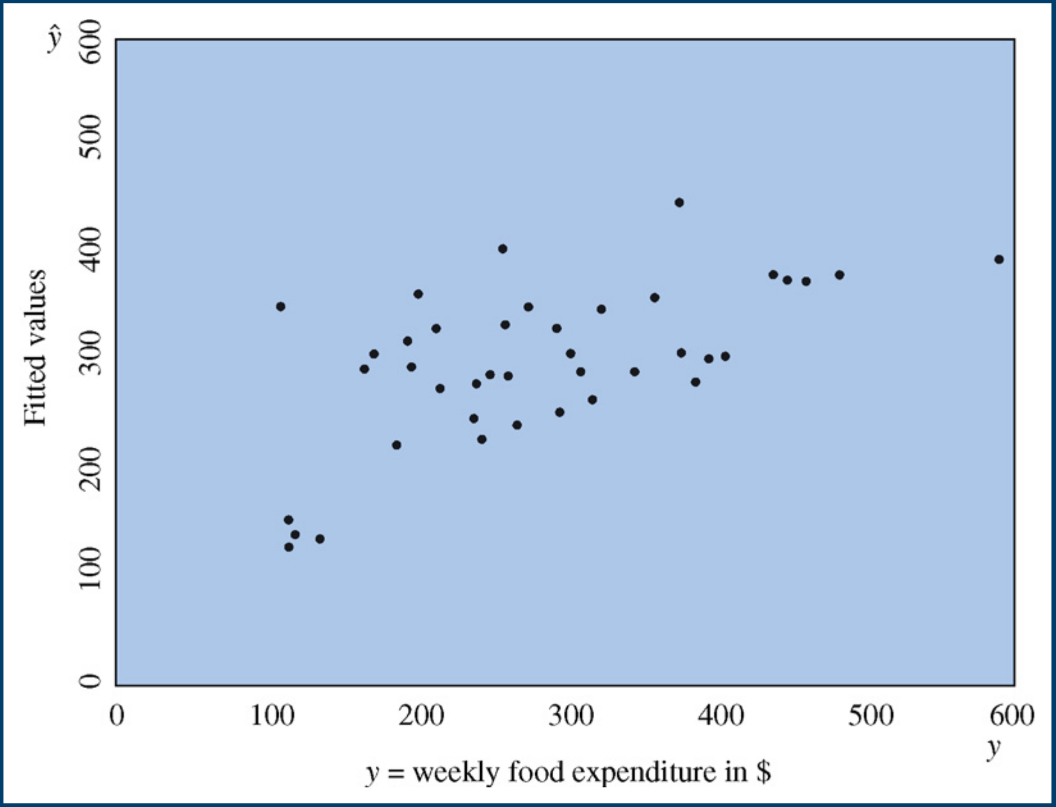

Figure **4.4** Plot of predicted y, $\hat{y}$ against y

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

$$FOOD\_EXP = 83.42 + 10.21\, INCOME \qquad R^2 = .385$$
$$\text{(se)} \qquad (43.41)^{*} \;\; (2.09)^{***}$$

* 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*:

$$y = \beta_1 + \beta_2 x + e = \beta_1 + (c\beta_2)(x/c) + e = \beta_1 + \beta_2^* x^* + e$$

where $\beta_2^* = c\beta_2$ and $x^* = x/c$

- Changing the scale of *y*:

$$y/c = (\beta_1/c) + (\beta_2/c)x + (e/c) \text{ or } y^* = \beta_1^* + \beta_2^* x + e^*$$

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

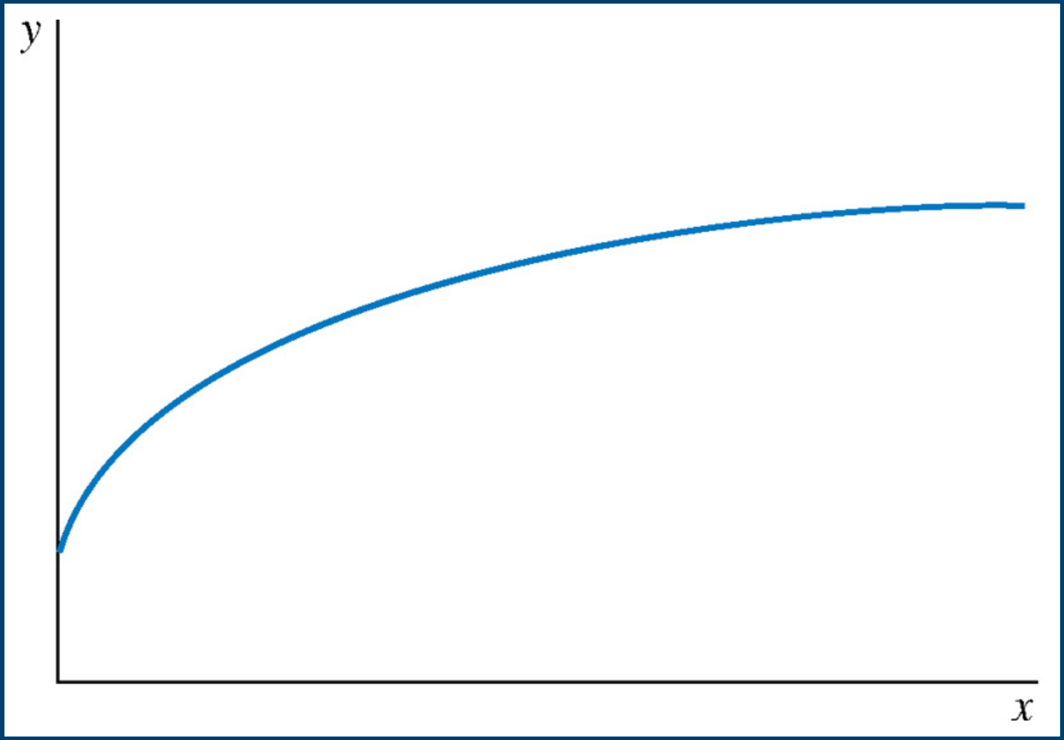

Figure **4.5** A nonlinear relationship between food expenditure and income

4.3.2 Choosing a Functional Form

- The **log-log model**

$$\ln(y) = \beta_1 + \beta_2 \ln(x)$$

The parameter β is the elasticity of *y* with respect to *x*.

- The **log-linear model**

$$\ln(y_i) = \beta_1 + \beta_2 x_i$$

A one-unit increase in *x* leads to (approximately) a $100 \times \beta_2$ percent change in *y*.

- The **linear-log** model

$$y = \beta_1 + \beta_2 \ln(x) \text{ or } \frac{\Delta y}{100(\Delta x / x)} = \frac{\beta_2}{100}$$

A 1% increase in *x* leads to a $\beta_2/100$ *unit* change in *y*.

4.3.3 The Food Expenditure Model

- The reciprocal model is

$$FOOD\_EXP = \beta_1 + \beta_2 \frac{1}{INCOME} + e$$

- The linear-log model is

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

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?

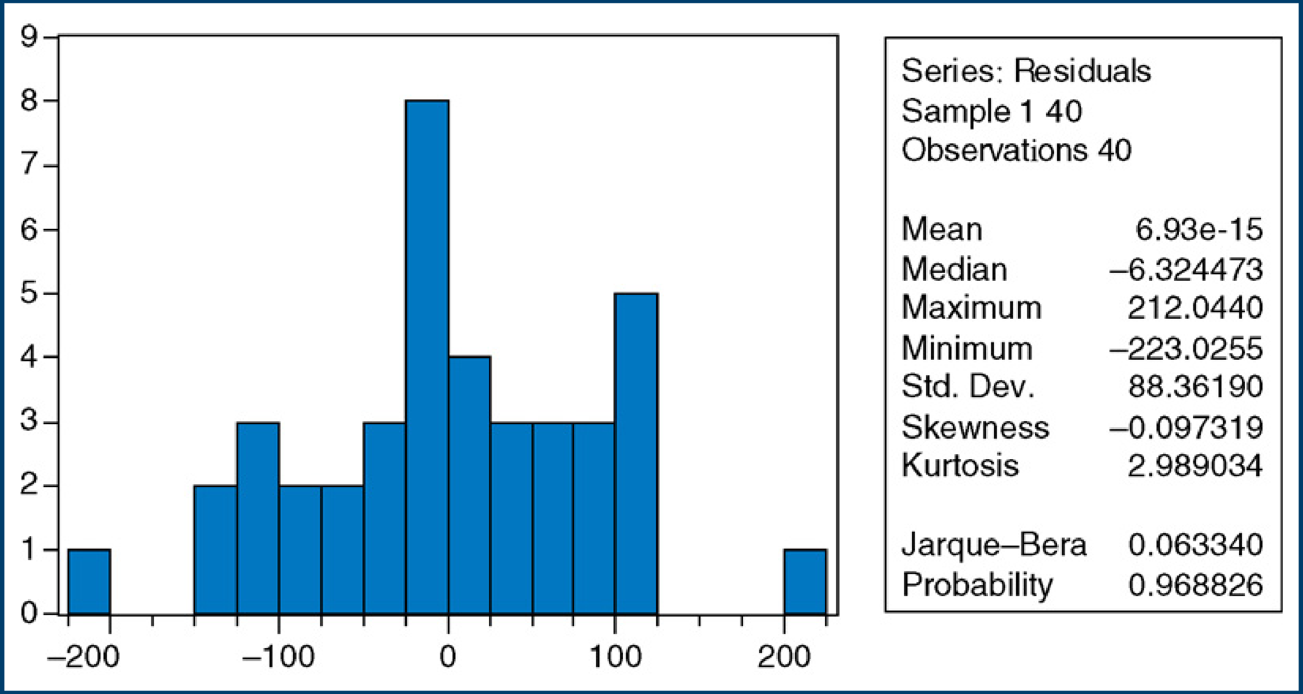

Figure **4.6** EViews output: residuals histogram and summary statistics for food expenditure example

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

$$JB = \frac{40}{6}\left(-.097^2 + \frac{(2.99-3)^2}{4} \right) = .063$$

4.3.5 Another Empirical Example

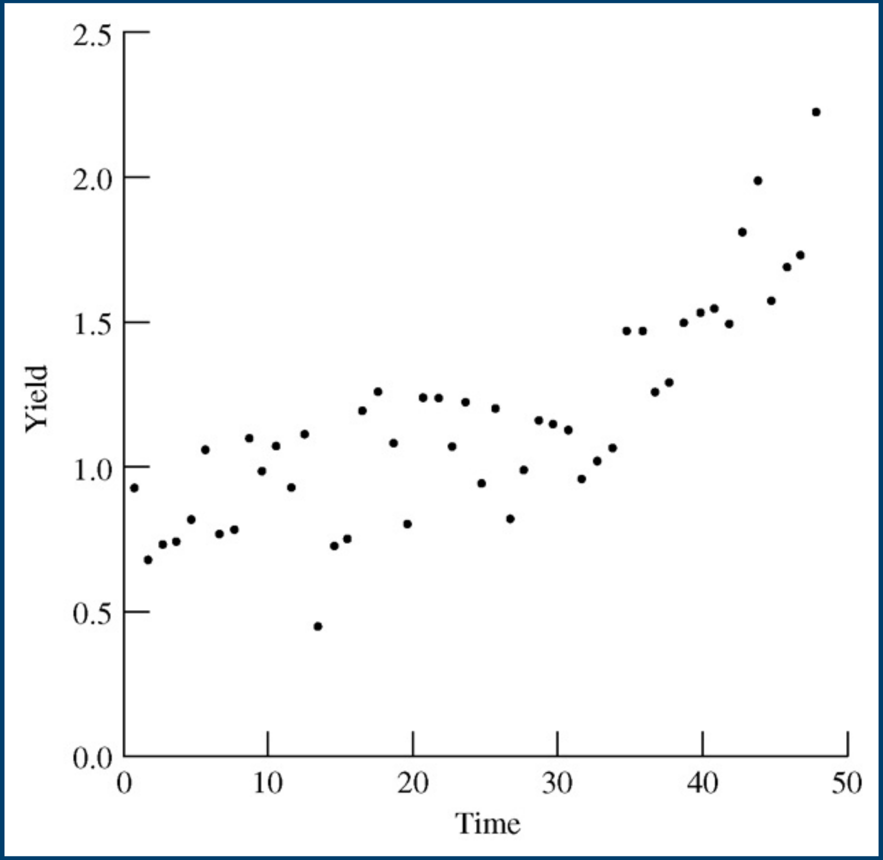

Figure **4.7** Scatter plot of wheat yield over time

4.3.5 Another Empirical Example

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

$$\widehat{YIELD}_t = .638 + .0210\, TIME_t \qquad R^2 = .649$$
$$\text{(se)} \qquad (.064)\ (.0022)$$

4.3.5 Another Empirical Example

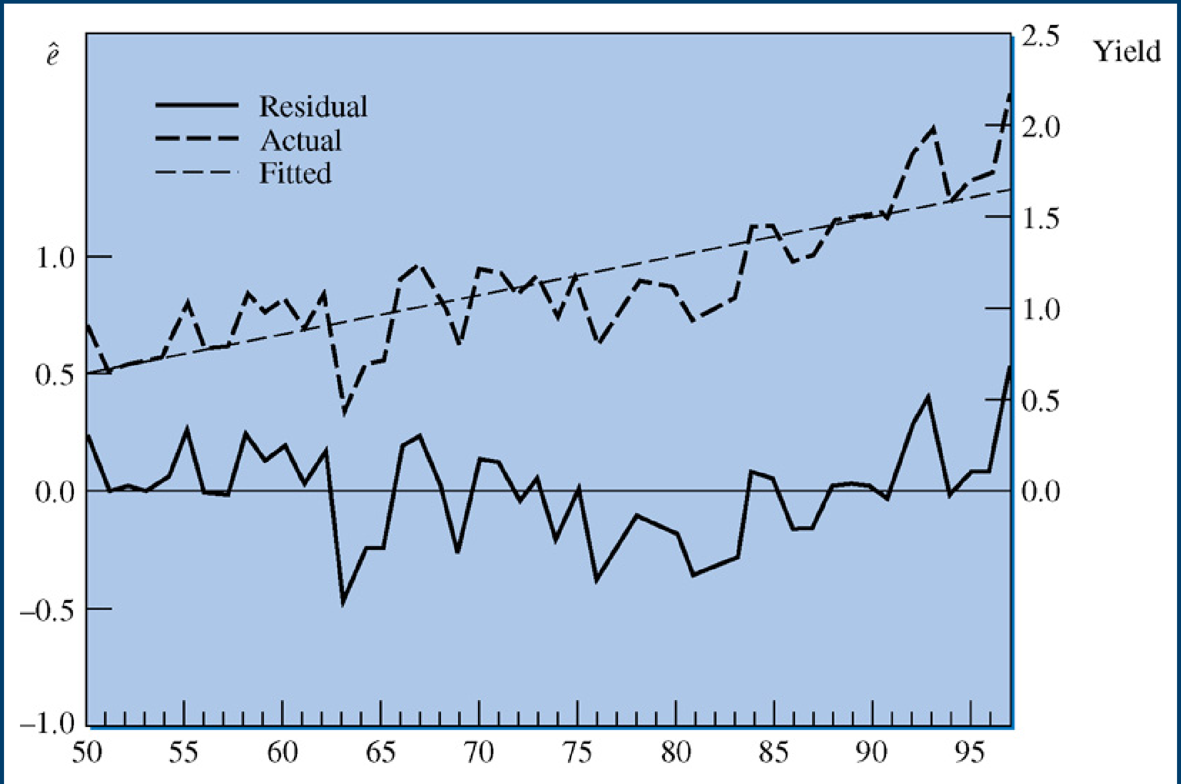

Figure **4.8** Predicted, actual and residual values from straight line

4.3.5 Another Empirical Example

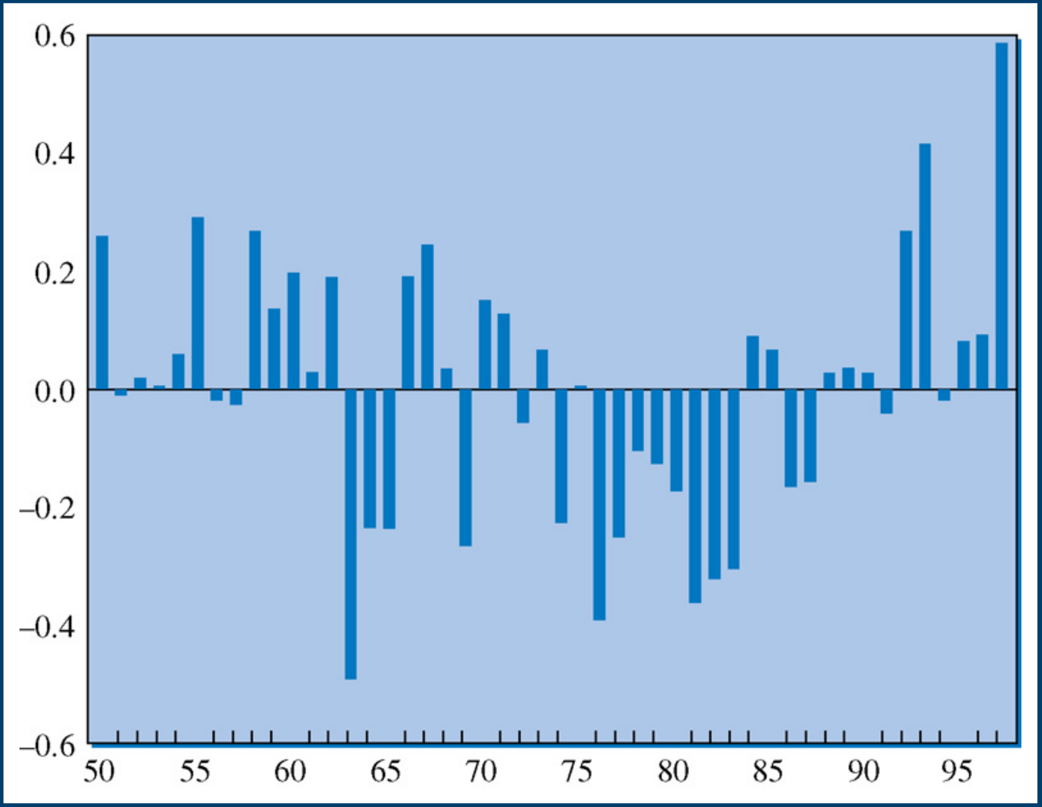

Figure **4.9** Bar chart of residuals from straight line

4.3.5 Another Empirical Example

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

$$TIMECUBE = TIME^3 / 1000000$$

$$\widehat{YIELD}_t = 0.874 + 9.68\, TIMECUBE_t \qquad R^2 = 0.751$$
$$\text{(se)} \qquad (.036)\ (.082)$$

4.3.5 Another Empirical Example

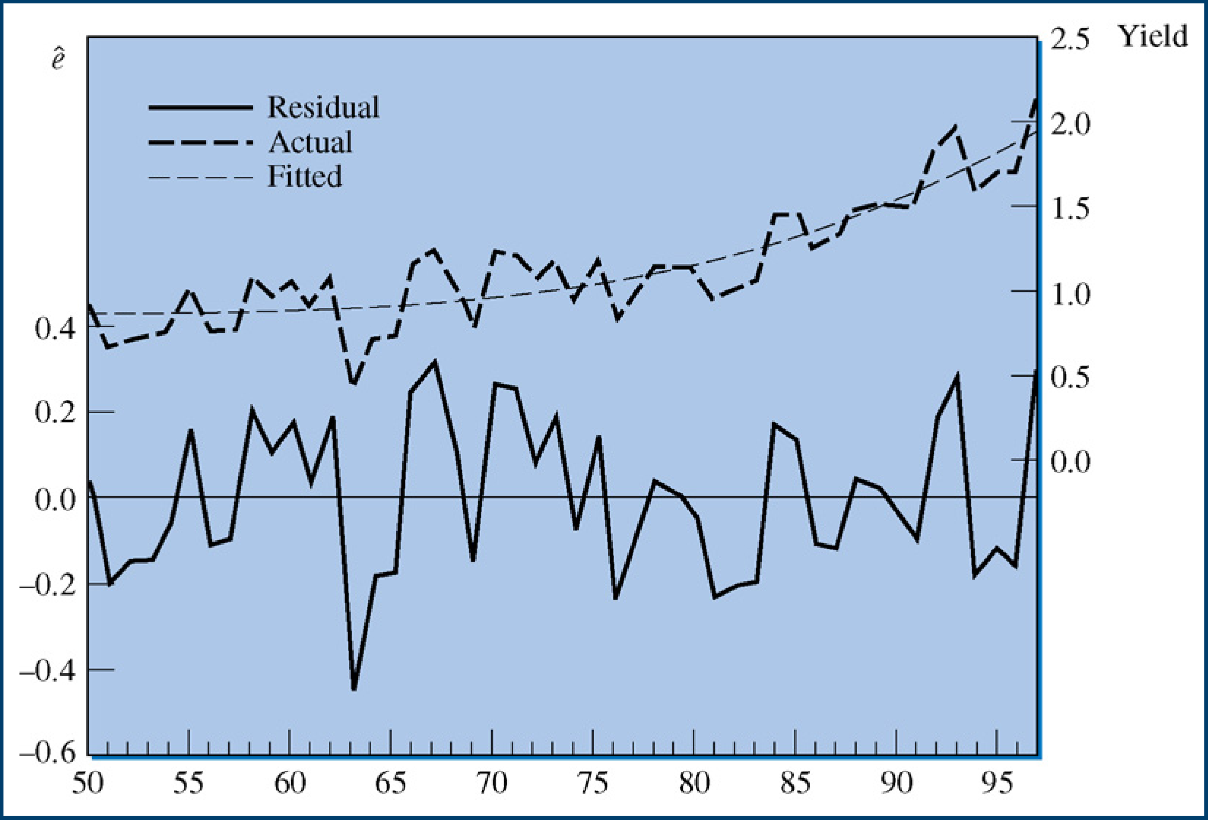

Figure **4.10** Fitted, actual and residual values from equation with cubic term

4.4 Log-Linear Models

- 4.4.1 The Growth Model

$$\ln\left(YIELD_t\right) = \ln\left(YIELD_0\right) + \ln\left(1+g\right)t$$

$$= \beta_1 + \beta_2 t$$

$$\widehat{\ln\left(YIELD_t\right)} = -.3434 \; + \; .0178t$$

$$(\text{se}) \qquad\qquad (.0584) \qquad (.0021)$$

4.4 Log-Linear Models

- 4.4.2 A Wage Equation

$$\ln(WAGE) = \ln(WAGE_0) + \ln(1+r)EDUC$$

$$= \beta_1 + \beta_2 EDUC$$

$$\widehat{\ln(WAGE)} = .7884 \;+\; .1038 \times EDUC$$

$$\text{(se)} \qquad (.0849) \qquad (.0063)$$

4.4 Log-Linear Models

- 4.4.3 Prediction in the Log-Linear Model

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

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

$$\widehat{\ln}(WAGE) = .7884 \;+\; .1038 \times EDUC = .7884 \;+\; .1038 \times 12 = 2.0335$$

$$\hat{y}_c = \widehat{E}(y) = y_n e^{\hat{\sigma}^2/2} = 7.6408 \times 1.1276 = 8.6161$$

4.4 Log-Linear Models

- 4.4.4 A Generalized R^2 Measure

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

$$R_g^2 = \left[\text{corr}\left(y, \hat{y}_c\right)\right]^2 = .4739^2 = .2246$$

R^2 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

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

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

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

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

$$f = y_0 - \hat{y}_0 = \left(\beta_1 + \beta_2 x_0 + e_0\right) - \left(b_1 + b_2 x_0\right)$$

$$\operatorname{var}\left(\hat{y}_0\right) = \operatorname{var}\left(b_1 + b_2 x_0\right) = \operatorname{var}\left(b_1\right) + x_0^2 \operatorname{var}\left(b_2\right) + 2x_0 \operatorname{cov}\left(b_1, b_2\right)$$

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

Appendix 4A
Development of a Prediction Interval

$$\operatorname{var}(\hat{y}_0) = \left[\frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2} - \left\{ \frac{\sigma^2 N \bar{x}^2}{N \sum (x_i - \bar{x})^2} \right\} \right] + \left[\frac{\sigma^2 x_0^2}{\sum (x_i - \bar{x})^2} + \frac{\sigma^2 (-2 x_0 \bar{x})}{\sum (x_i - \bar{x})^2} + \left\{ \frac{\sigma^2 N \bar{x}^2}{N \sum (x_i - \bar{x})^2} \right\} \right]$$

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

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

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

Appendix 4A
Development of a Prediction Interval

$$\frac{f}{\sqrt{\operatorname{var}(f)}} \sim N(0,1)$$

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

$$\frac{f}{\sqrt{\widehat{\operatorname{var}}(f)}} = \frac{y_0 - \hat{y}_0}{\operatorname{se}(f)} \sim t_{(N-2)} \qquad (4A.1)$$

$$P(-t_c \le t \le t_c) = 1 - \alpha \qquad (4A.2)$$

Appendix 4A
Development of a Prediction Interval

$$P[-t_c \le \frac{y_0 - \hat{y}_0}{\mathrm{se}(f)} \le t_c] = 1 - \alpha$$

$$P\left[\hat{y}_0 - t_c \mathrm{se}(f) \le y_0 \le y_0 + t_c \mathrm{se}(f)\right] = 1 - \alpha$$

Appendix 4B
The Sum of Squares Decomposition

$$\left(y_i - \bar{y}\right)^2 = \left[(\hat{y}_i - \bar{y}) + e_i\right]^2 = (\hat{y}_i - \bar{y})^2 + e_i^2 + 2(\hat{y}_i - \bar{y})e_i$$

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

$$\sum\left(\hat{y}_i - \bar{y}\right)e_i = \sum \hat{y}_i e_i - \bar{y}\sum e_i = \sum\left(b_1 + b_2 x_i\right)e_i - \bar{y}\sum e_i$$

$$= b_1\sum e_i + b_2\sum x_i e_i - \bar{y}\sum e_i$$

Appendix 4B
The Sum of Squares Decomposition

$$\sum \hat{e}_i = \sum \left(y_i - b_1 - b_2 x_i \right) = \sum y_i - Nb_1 - b_2 \sum x_i = 0$$

$$\sum x_i \hat{e}_i = \sum x_i \left(y_i - b_1 - b_2 x_i \right) = \sum x_i y_i - b_1 \sum x_i - b_2 \sum x_i^2 = 0$$

$$\sum \left(\hat{y}_i - \bar{y} \right) e_i = 0$$

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

$$E(w) = e^{\mu + \sigma^2/2}$$

$$\operatorname{var}(w) = e^{2\mu + \sigma^2}\left(e^{\sigma^2} - 1\right)$$

Appendix 4C
The Log-Normal Distribution

Given the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$

If we assume that $e \sim N(0, \sigma^2)$

$$E(y_i) = E\left(e^{\beta_1+\beta_2 x_i + e_i}\right) = E\left(e^{\beta_1+\beta_2 x_i} e^{e_i}\right) =$$

$$e^{\beta_1+\beta_2 x_i} E\left(e^{e_i}\right) = e^{\beta_1+\beta_2 x_i} e^{\sigma^2/2} = e^{\beta_1+\beta_2 x_i + \sigma^2/2}$$

$$\hat{E}(y_i) = e^{b_1+b_2 x_i + \hat{\sigma}^2/2}$$

Appendix 4C
The Log-Normal Distribution

The growth and wage equations:

$$\beta_2 = \ln(1 + r) \qquad \text{and} \qquad r = e^{\beta_2} - 1$$

$$b_2 \sim N\left(\beta_2, \operatorname{var}(b_2) = \sigma^2 \Big/ \sum (x_i - \bar{x})^2\right)$$

$$E\left[e^{b_2}\right] = e^{\beta_2 + \operatorname{var}(b_2)/2} \qquad \hat{r} = e^{b_2 + \widehat{\operatorname{var}}(b_2)/2} - 1$$

$$\widehat{\operatorname{var}}(b_2) = \hat{\sigma}^2 \Big/ \sum (x_i - \bar{x})^2$$