2.1 - 2.3: Regression

Financial Econometrics

Prof. Huei-Wen Teng

2.1 An economic model

Fig 2.6: Data for the food expenditure example.

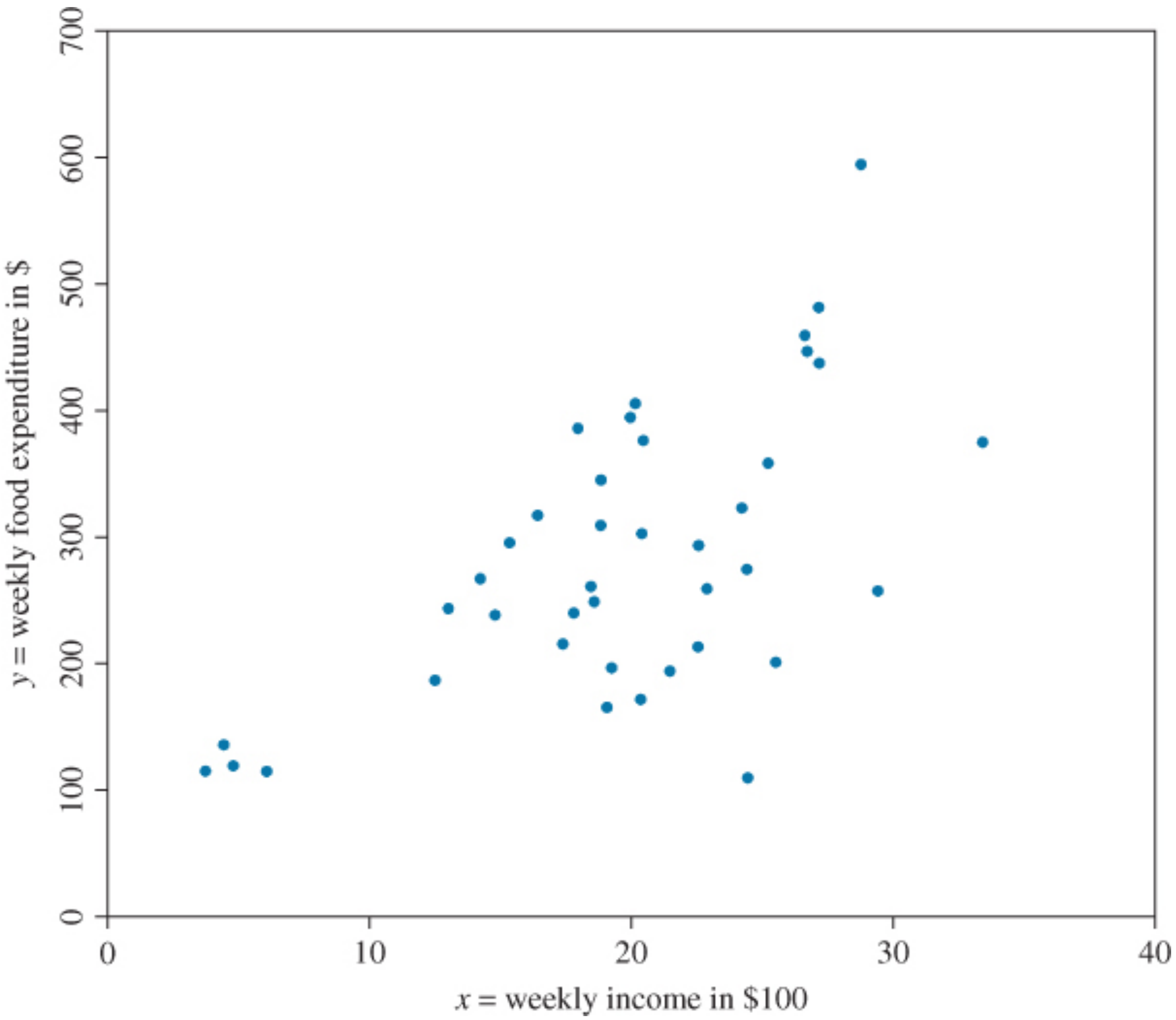
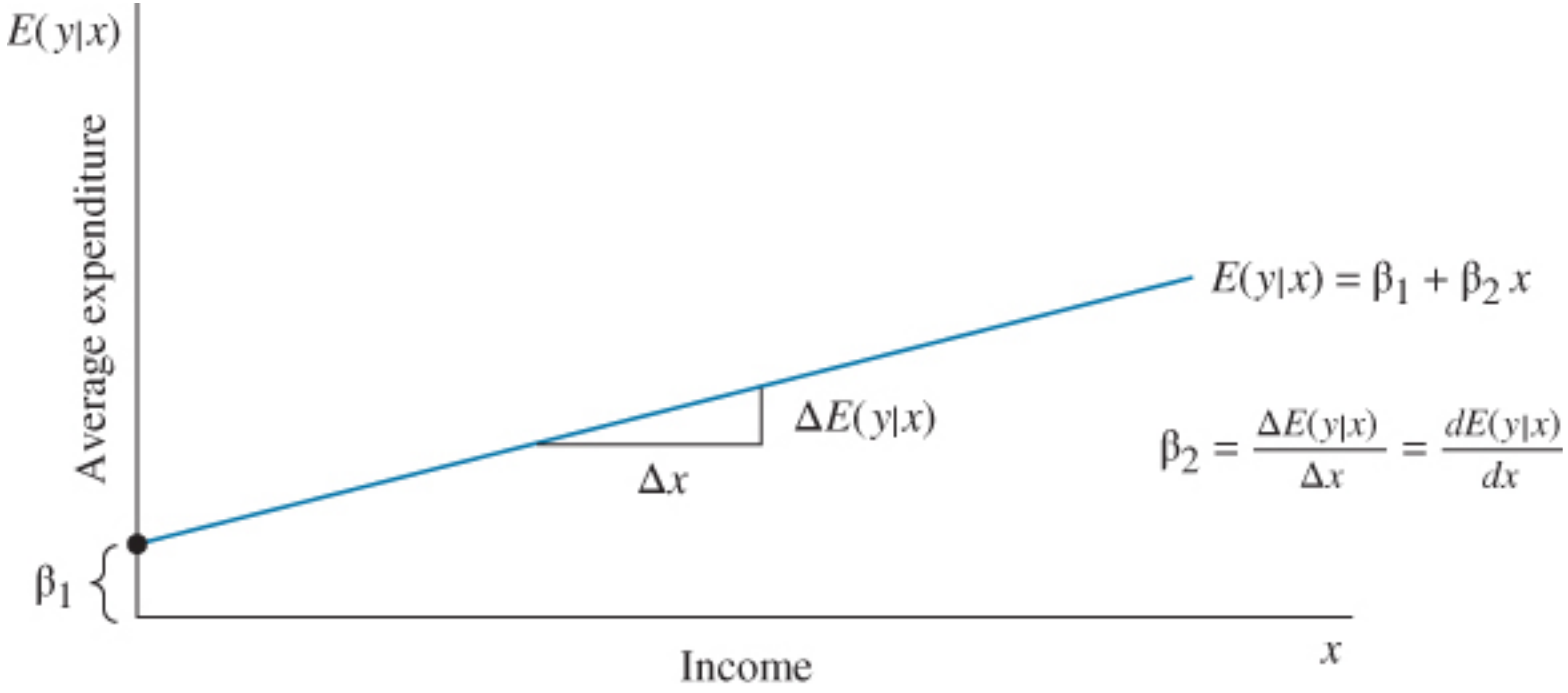


Fig 2.2. The economic model: a linear relationship between average per person food expenditure and income.



2.2 An econometric model

Fig 2.3 Conditional probability densities for *e* and *y*

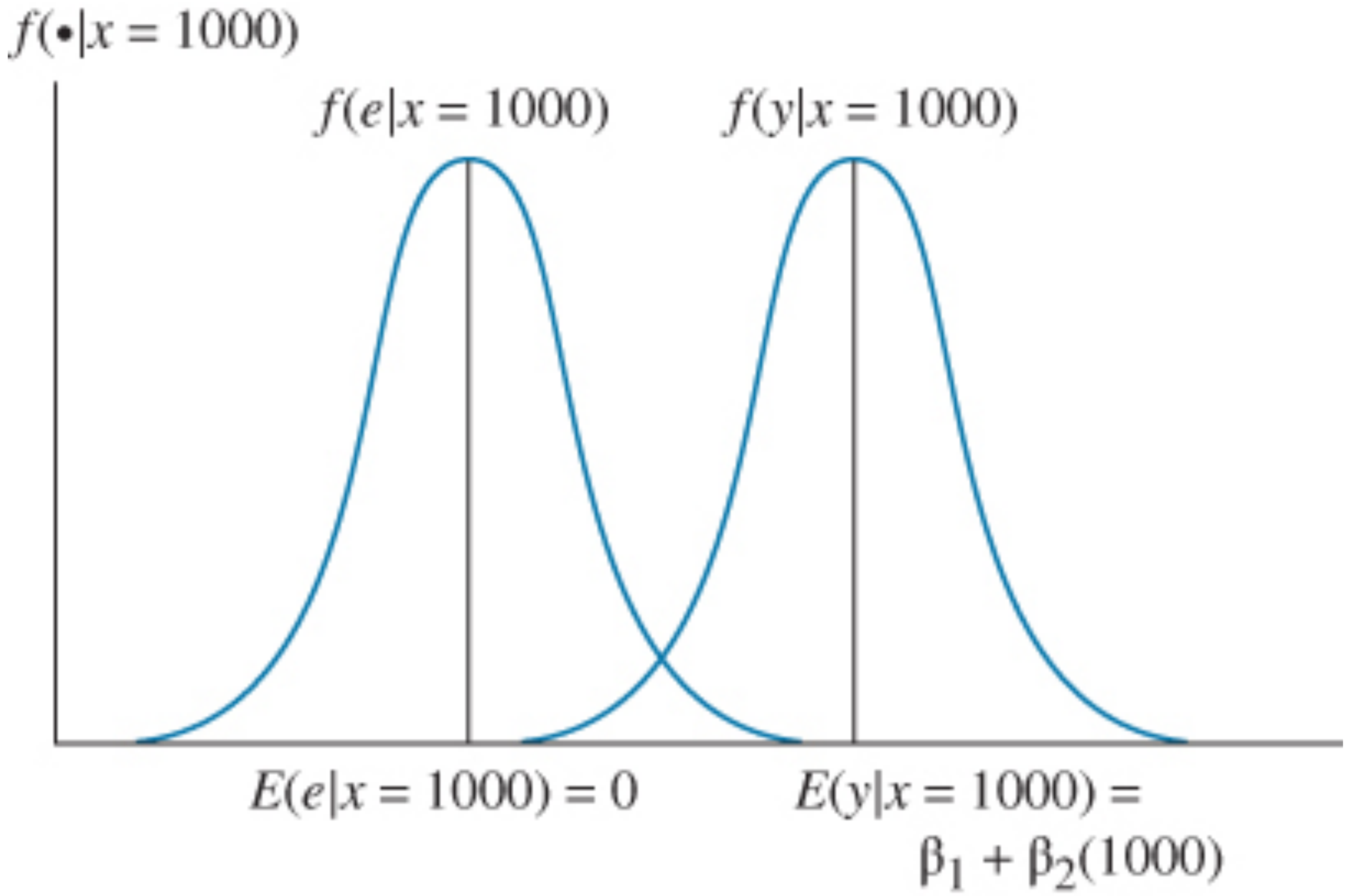


Fig 2.4. The random error

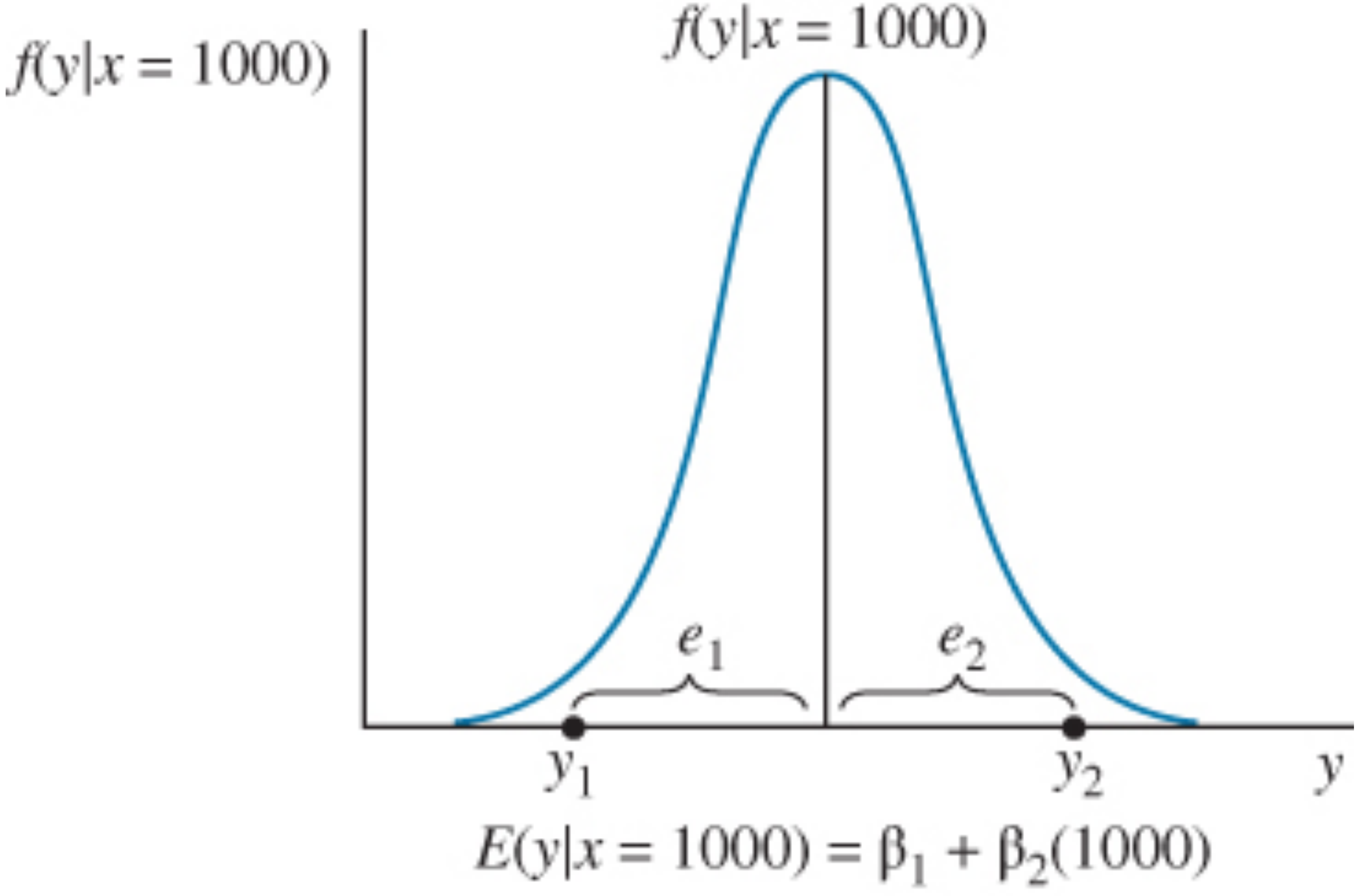
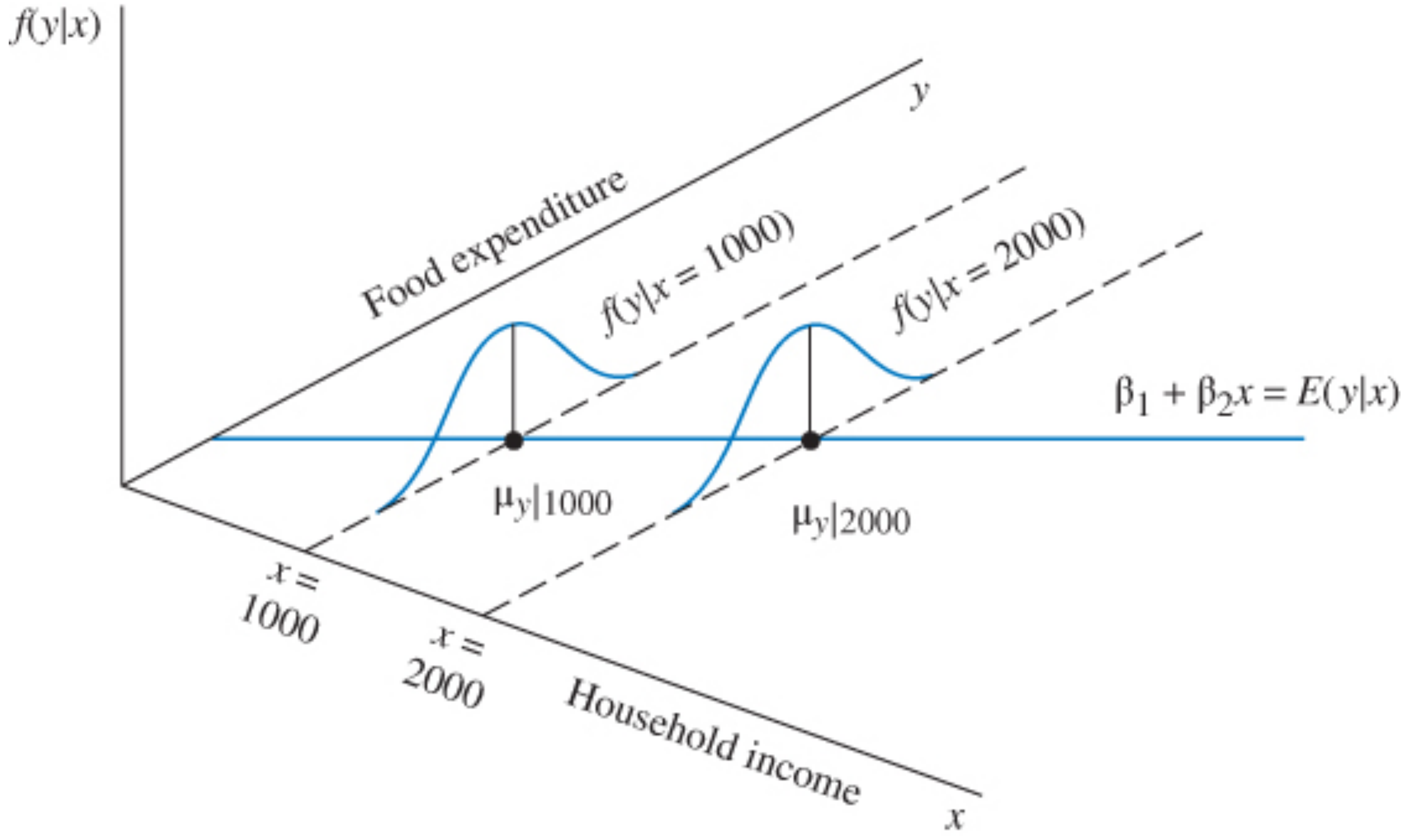


Fig 2.5 The conditional probability density function for *y*, food expenditure, at two levels of income.



2.3 Estimating the Regression Parameters

Review i

Economists are interested in studying relationship between variables

1. y: dependent variable (also known as target or response variable)
2. x: independent, explanatory variable. In a multiple regression,$x = (x_1, \ldots, x_p)$ indicate a vector of p dimensions.

Recipe in fitting a model i

1. An economic model is of the following form,

$$y = f(x, \beta) \, .$$

Suppose a functional form of $f(x,\beta)$ is decided based on domain knowledge or explanatory data analysis. β is called the parameter.

2. Econometric model links the data in practice to the theoretic economic model by an error term:

$$y = f(x, \beta) + e,$$

where e is the random error.

Recipe in fitting a model ii

3. Collect the data.
4. Estimate the parameter using one method from the following:

   least squares estimates (LSE), maximum likelihood estimate (MLE), Moment methods, others.

5. Perform model diagnostic to check model misspecification.

   5.1 Visualization tools: scatter plot (check weird pattern), density plot (check normality), qqplot (check normality).
   5.2 Goodness-of-fit test: Komogorove-Smirnov typed test, Others.

6. If Step (e) is passed, interpret the model parameters.

   Otherwise, it is meaningless to do so.

A simple regression model i

A simple regression model is proposed by

$$y = \beta_1 + \beta_2 x + e,$$

where β_1 is the intercept and β_2 is the slope. Further assumptions:

1. The variable x is not random and must take at least two different values.
2. e are i.i.d. $N(0, \sigma_2)$.

These equal the following five assumptions in the textbook:

Assumption 1: $E(y|x) = \beta_1 + \beta_2 x$

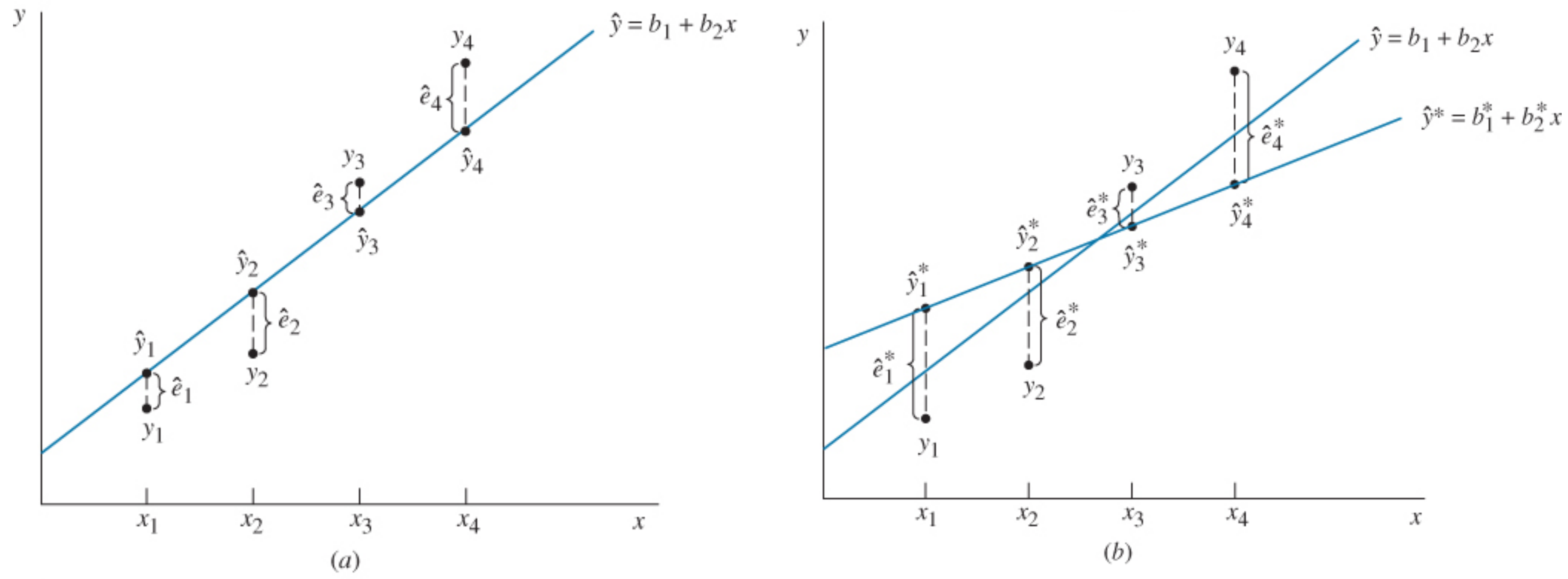
Assumption 2: $var(y|x) = \sigma^2$

Assumption 3: $cov(y_i, y_j) = 0$

Assumption 4: The variable x is not random and must take at least two different values.

Assumption 5: $y \sim N(\beta_1 + \beta_2 x, \sigma^2)$.

Fig 2.7 (a) The relationship among y, $\hat{e}$, and the fitted regression line, (b) The residuals from another fitted line.



Estimate parameters using LSE.

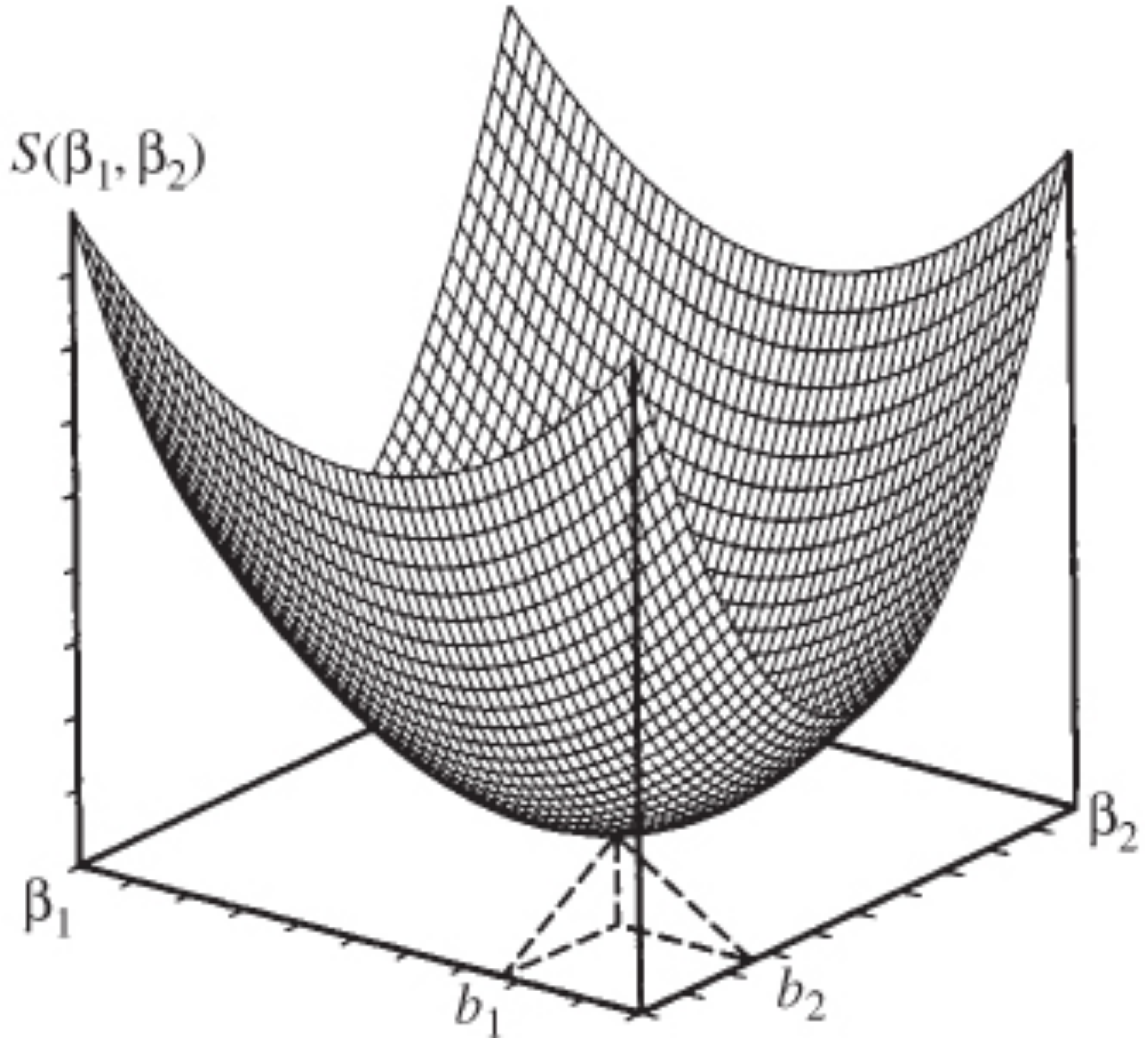
Define the sum of squares:

$$S(\beta_1, \beta_2) = \sum_{i=1}^{N} (y_i - \beta_1 - \beta_2 x_i)^2 \,.$$

Least squares estimates b_1, b_2 satisfy

$$b_1, b_2 = \arg\min_{\beta_1, \beta_2} S(\beta_1, \beta_2) \,.$$

Fig. 2A.1: The sum of squares function and the minimizing values b_1 and b_2



The resulting estimators are:

$$b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2},$$

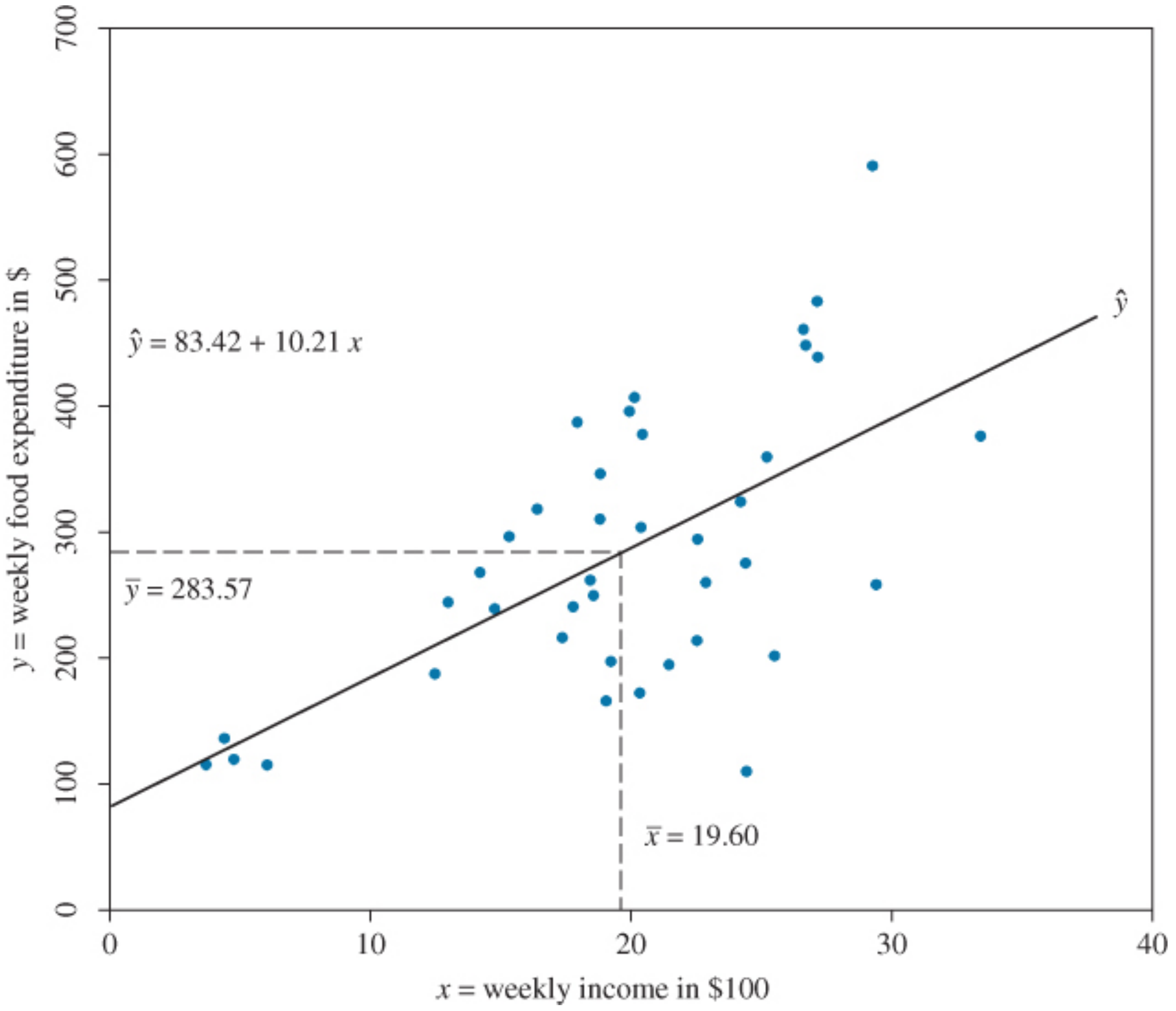
$$b_1 = \bar{y} - b_2 \bar{x}\,.$$

Derivations to obtain b_1 and b_2 are deferred to the Appendix.

The estimated or fitted regression line is:

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

Fig 2.8. The fitted regression



Example: Food expenditure model

We have

$$\hat{y}_i = 83.42 + 10.21x_i$$

Interpretations on the parameters:

The intercept estimate $b_1 = 83.42$ is an estimate of the weekly

food expenditure on food for a household with zero income.

2. The value $b_2 = 10.21$ is an estimate of β_2. We estimate that if the income goes up by \$100, expected weekly expenditure on food will increase approximately by \$10.21.

Elasticity

The elasticity of mean expenditure with respect to income is:

$$\varepsilon = \frac{\text{Percentage change in } y}{\text{percentage change in } x} = \frac{\triangle E(y)/E(y)}{\triangle x/x} = \frac{\triangle E(y)x}{\triangle xE(y)}$$

$$= \frac{\triangle E(y)}{\triangle x} \times \frac{x}{y} = \beta_2 \frac{x}{\beta_1 + \beta_2 x}$$

We estimate the elasticity by

$$\hat{\varepsilon} = b_2 \frac{\bar{x}}{\bar{y}} = 10.21 \times \frac{19.60}{283.57} = 0.71$$

Prediction

To predict weekly food expenditure for a household with a weekly income of \$2000, we plugging $x = 20$ into our estimated equation to obtained

$$\hat{y} = 83.42 + 10.21x_i = 83.42 + 10.21(20) = 287.61.$$

We *predict* that a household with a weekly income of \$2000 will spend \$287.61 per week on food.

R

```
# For details, see https://bookdown.org/ccolonescu/RPoE4/#
# plot the data
food = data(food);

plot(food$income, food$food_exp,
     ylim=c(0, max(food$food_exp)),
     xlim=c(0, max(food$income)),
     xlab="weekly income in $100",
     ylab="weekly food expenditure in $", type = "p")
```

```
# fit the model to the data: EXP = beta 1+ beta2 INCOME + e
mod1 <- lm(food_exp ~ income, data = food)
b1 <- coef(mod1)[[1]]
b2 <- coef(mod1)[[2]]
smod1 <- summary(mod1)
smod1
abline(b1,b2) # add the estimated (fitted) regression line


names(mod1)
names(smod1)
mod1$coefficients
smod1$coefficients
coef(mod1)
```

```
# retrieve the residuals and do a simple diagnostic test
r= resid(mod1)
plot(r) # scattor plot
hist(r) # histogram
plot(density(r)) # density plot
qqnorm(r) # qqplot
```

```
# prediction
newx <- data.frame(income = c(20, 25, 27))
yhat <- predict(mod1, newx)
names(yhat) <- c("income=$2000", "$2500", "$2700")
yhat  # prints the result
```

Appendix. Derivations of the LSE for b_1 and b_2 i

First-order-Condition requires the partial derivatives of $S(\beta_1, \beta_2)$ to be zeros:

$$\frac{\partial S(\beta_1, \beta_2)}{\partial \beta_1} = -2 \sum (y_i - \beta_1 - \beta_2 x_i) = 0,$$

$$\frac{\partial S(\beta_1, \beta_2)}{\partial \beta_2} = -2 \sum x_i (y_i - \beta_1 - \beta_2 x_i) = 0.$$

Simple math gives

$$\sum y_i = N\beta_1 + (\sum x_i)\beta_2, \qquad (1)$$

$$\sum x_i y_i = (\sum x_i)\beta_1 + (\sum x_i^2)\beta_2 \, . \qquad (2)$$

Appendix. Derivations of the LSE for b_1 andb_2 ii

Multiply (1) by $(\sum x_i)$ and multiply (2) by N, we have

$$(\sum x_i)(\sum y_i) = N(\sum x_i)\beta_1 + (\sum x_i)^2\beta_2, \tag{3}$$

$$N(\sum x_i y_i) = N(\sum x_i)\beta_1 + N(\sum x_i^2)\beta_2\,. \tag{4}$$

Let $\bar{x} = (\sum x_i)/N$ and $\bar{y} = (\sum y_i)/N$. We have

$$b_2 = \frac{N(\sum x_i y_i) - (\sum x_i \sum y_i)}{N(\sum x_i)^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - N\bar{x}\bar{y}}{\sum x_i^2 - N\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\,. \qquad 5()$$

Plug b_2 in (5) into (1), we have

$$b_1 = \bar{y} - \bar{x}b_2\,.$$

Appendix. Derivations of the LSE for b_1 and b_2 iii

For hand calculations, we obtain the following identities,

$$
\begin{aligned}
\sum (x_i - \bar{x})^2 &= \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \\
&= \sum x_i^2 - 2\bar{x}(N\bar{x}) + N\bar{x}^2 \\
&= \sum x_i^2 - N\bar{x}^2, \\
\sum (x_i - \bar{x})(y_i - \bar{y}) &= \sum (x_i y_i - \bar{x}y_i - \bar{y}x_i + \bar{x}\bar{y}) \\
&= \sum x_i y_i - \bar{x}(N\bar{y}) - \bar{y}N\bar{x} + N\bar{x}\bar{y} \\
&= \sum x_i y_i - N\bar{x}\bar{y}\,.
\end{aligned}
$$

2.4 - 2.9: Regression

Financial Econometrics

Prof. Huei-Wen Teng

2.4 Assessing the Least Squares Estimators

Expected values of b_1 and b_2

$$E(b_1 \mid \mathbf{x}) = \beta_1, E(b_2 \mid \mathbf{x}) = \beta_2 .$$

TABLE 2.2 Estimates from 10 Hypothetical Samples

Sample	b_1	b_2
1	93.64	8.24
2	91.62	8.90
3	126.76	6.59
4	55.98	11.23
5	87.26	9.14
6	122.55	6.80
7	91.95	9.84
8	72.48	10.50
9	90.34	8.75
10	128.55	6.99

Figure 1: Table 2.2. Estimates from 10 Hypothetical Samples

Variance and covariances

$$Var(b_1) = \sigma^2 \frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2}, \quad Var(b_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2},$$

$$\sigma_{b_1} = \sqrt{Var(b_1)}, \quad \sigma_{b_2} = \sqrt{Var(b_2)},$$

$$Cov(b_1, b_2) = \sigma^2 \frac{-\bar{x}}{\sum (x_i - \bar{x})^2}.$$

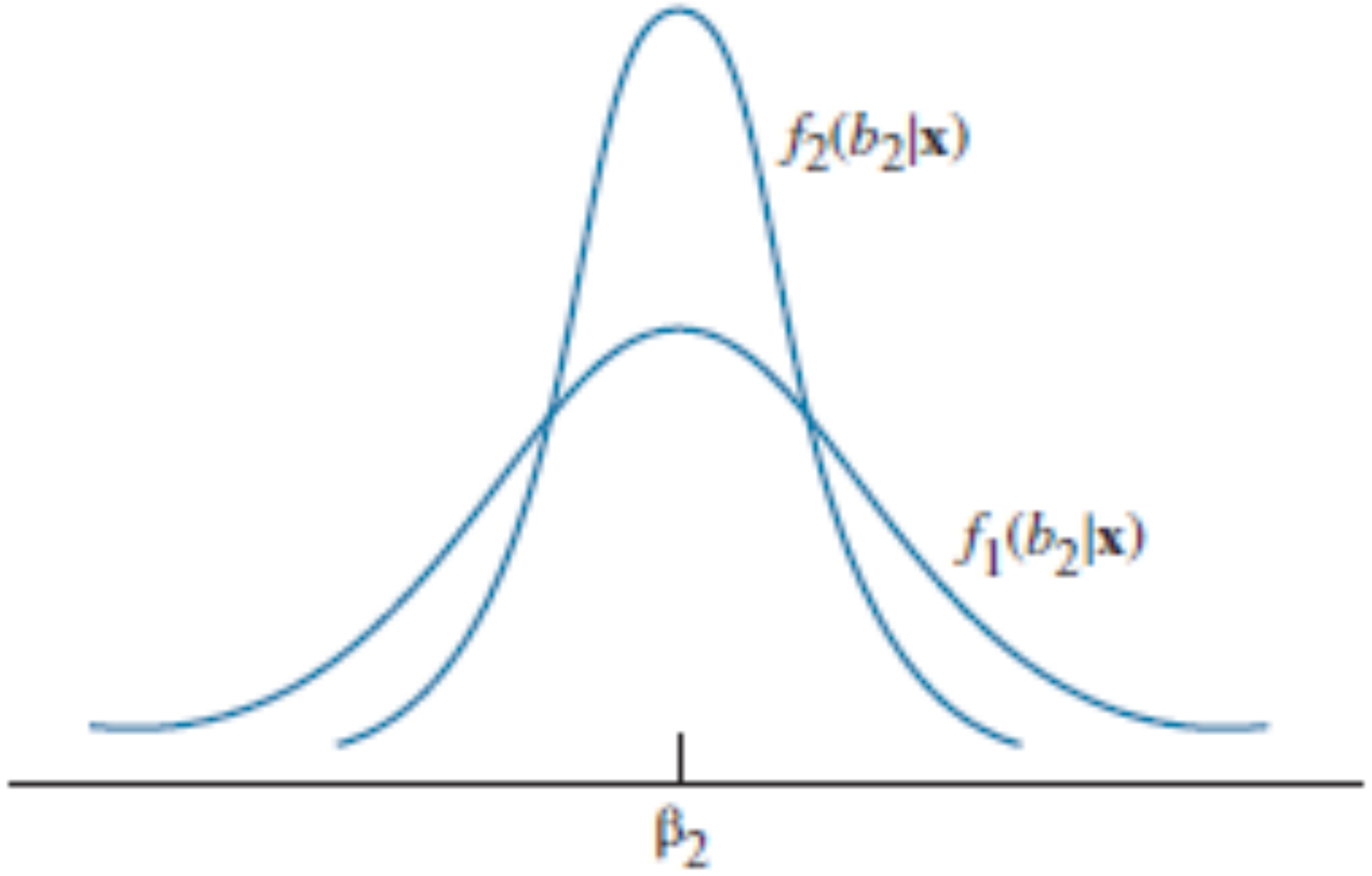


Figure 2: $Fig2.10$ Two possible probability density function for b_2

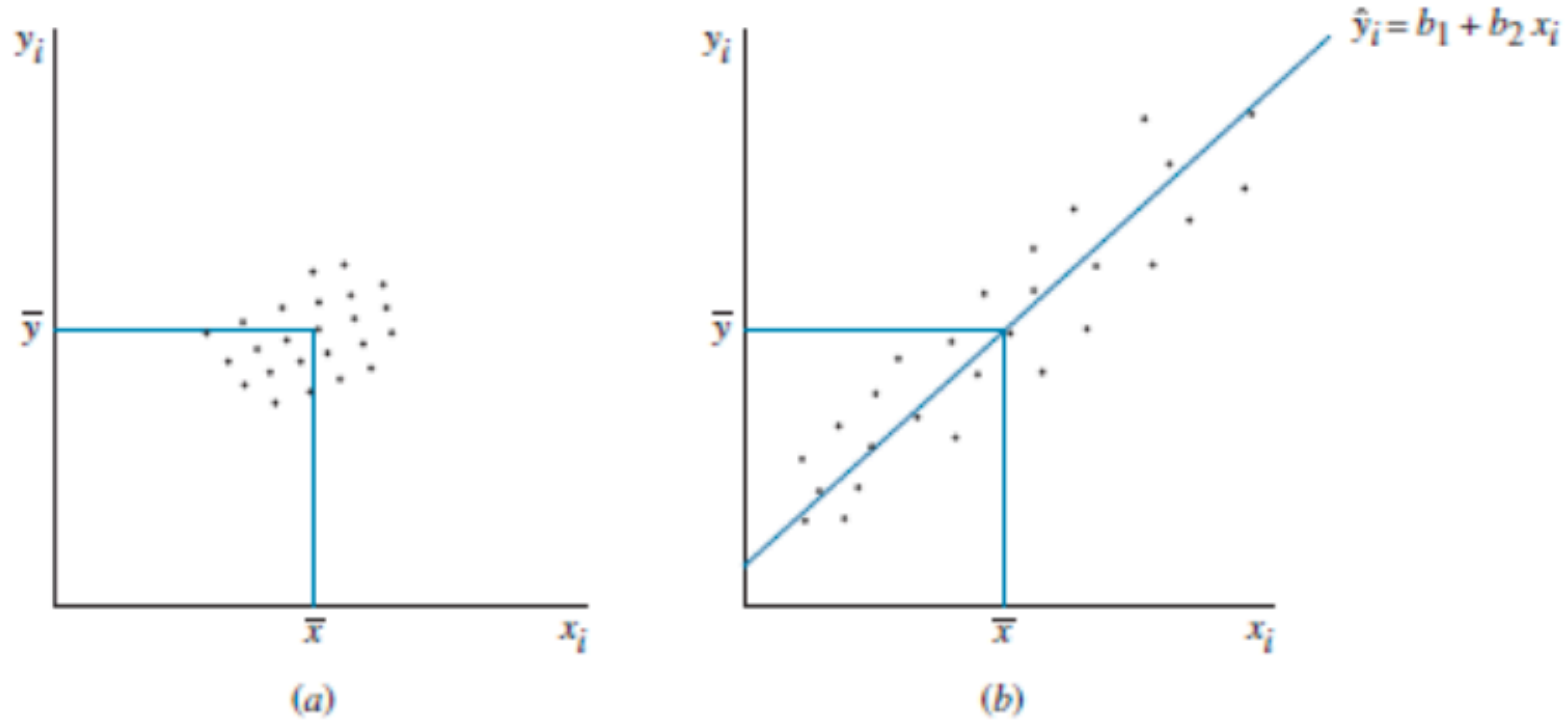


Figure 3: Fig 2.11 The influence of variation in the explanatory variable *x* on precision of estimation: (a) lower *x* variation, low precision: (b) high *x* variation, high precision.

2.5 Gauss-Markov Theorem

Assumptions in p. 58

Assumptions of the Simple Linear Regression Model

SR1: Econometric Model All data pairs (y_i, x_i) collected from a population satisfy the relationship

$$y_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

SR2: Strict Exogeneity The conditional expected value of the random error e_i is zero. If $\mathbf{x} = (x_1, x_2, \ldots, x_N)$, then

$$E(e_i|\mathbf{x}) = 0$$

If strict exogeneity holds, then the population regression function is

$$E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i, \quad i = 1, \ldots, N$$

and

$$y_i = E(y_i|\mathbf{x}) + e_i, \quad i = 1, \ldots, N$$

SR3: Conditional Homoskedasticity The conditional variance of the random error is constant.

$$\text{var}(e_i|\mathbf{x}) = \sigma^2$$

SR4: Conditionally Uncorrelated Errors The conditional covariance of random errors e_i and e_j is zero.

$$\text{cov}(e_i, e_j|\mathbf{x}) = 0 \quad \text{for} \quad i \neq j$$

SR5: Explanatory Variable Must Vary In a sample of data, x_i must take at least two different values.

SR6: Error Normality (*optional*) The conditional distribution of the random errors is normal.

$$e_i|\mathbf{x} \sim N(0, \sigma^2)$$

Gauss-Markov Theorem

Given x under the assumptions SR1-SR5 of the linear regression model, the estimators b_1 and b_2 have the smallest variance of all linear and unbiased estimators of β_1 and β_2. They are the best linear unbiased estimators (BLUE) of β_1 and β_2.

2.6 The Probability Distributions of Least Squares Estimators

Sampling distribution

If SR6 holds, we have:

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2}\right), \tag{1}$$

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right). \tag{2}$$

A Central Limit Theorem: If assumptions SR1-SR5 hold, and if the sample size N is sufficiently large, then the least square estimators have a distribution that approximate the normal distribution shown in (1) and (2).

Why we need (1) and (2)?

1. Confidence interval
2. Hypothesis test
    1.

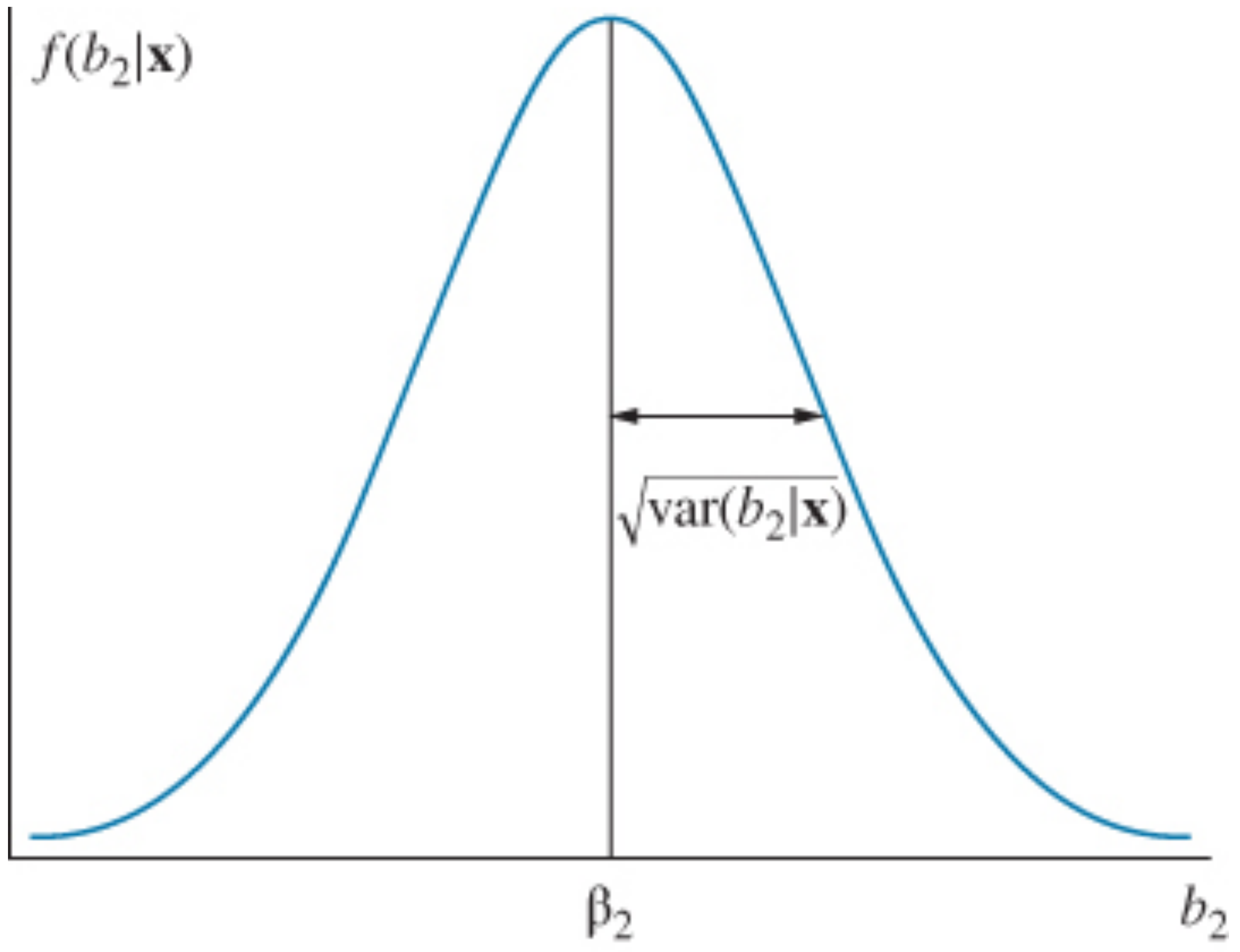
2.7 Estimating the Variance of the Error Term

How do we estimate $Var(b_2 | x)$ in (2)? Need to estimate σ^2!

Note $e_i \sim N(0,\sigma^2)$. Thus, we have

$$E(e_i^2) = \sigma^2 .$$

We use "a sample moment estimator" to estimate σ^2 .



The *residual* approximates the error:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i \,.$$

An unbiased estimator for σ^2 is

$$\hat{\sigma}^2 = \frac{\sum (y_i - b_1 - b_2 x_i)^2}{N - 2},$$

We have $E(\hat{\sigma}^2) = \sigma^2$.

We obtain estimates:

$$\hat{Var}(b_1) = \hat{\sigma}^2 \frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2}, \quad \hat{Var}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2},$$

$$\hat{Cov}(b_1, b_2) = \hat{\sigma}^2 \frac{-\bar{x}}{\sum (x_i - \bar{x})^2},$$

$$\hat{\sigma}_{b_1} = \sqrt{\hat{Var}(b_1)}, \quad \hat{\sigma}_{b_2} = \sqrt{\hat{Var}(b_2)}\,.$$

Summarize the estimated variances and covariance as:

$$\begin{pmatrix} \hat{Var}(b_1) & \hat{Cov}(b_1, b_2) \\ \hat{Cov}(b_1, b_2) & \hat{Var}(b_2) \end{pmatrix}.$$

R

```
# Many applications require estimates of the
# variances and covariances of the
# regression coefficients.
# R stores them in the a matrix vcov():
varb1 <- vcov(mod1)[1, 1]
varb1
varb2 <- vcov(mod1)[2, 2]
varb2
covb1b2 <- vcov(mod1)[1,2]
covb1b2
vcov(mod1)
```

2.8 Estimating Nonlinear Relationships

Overview

Recall the linear model of house price:

$$PRICE = \beta_1 + \beta_2 SQFT + e\,.$$

Many economic relationships are represented by curved lines, and are said to display *curvlinear* forms.

We may consider using $SQFT^2$ or ln(PRICE) as an alternative model.

For a general class of models, see chapter 4.1.

The quadratic model

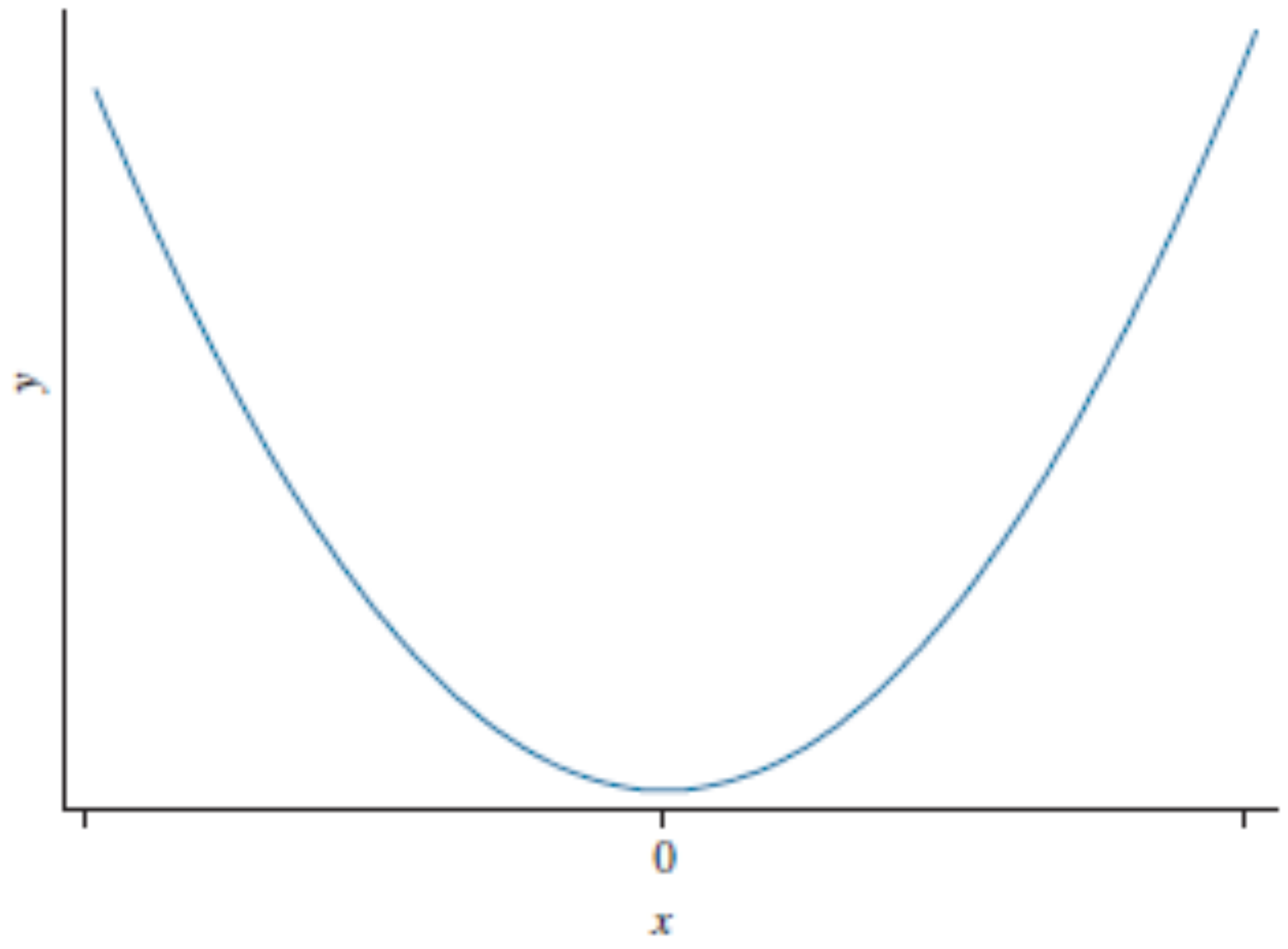


FIGURE 2.13 A quadratic function, $y = a + bx^2$.

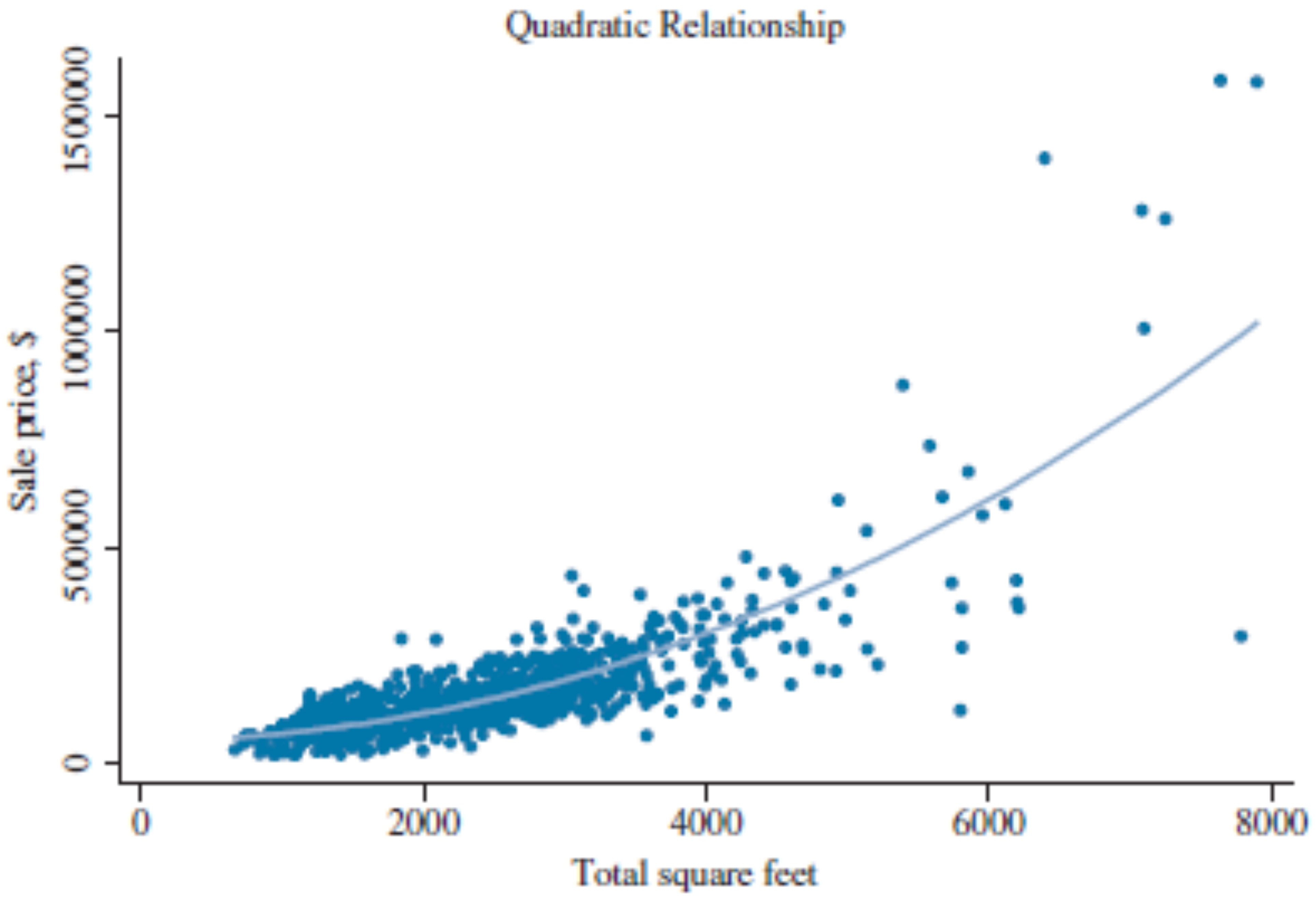


FIGURE 2.14 A fitted quadratic relationship.

The slope

Consider $SQFT^2$ as the explanatory variable:

$$\text{PRICE} = \beta_1 + \beta_2 SQFT^2 + e\,.$$

The slope is

$$m = \frac{dPRICE}{dSQFT} = 2\beta_2 SQFT\,.$$

We estimate the slope by

$$\hat{m} = 2b_2 SQFT\,.$$

If $b_2 > 0$, a larger house have large slop, and larger estimated price per additional square foot.

The elasticity

The elasticity is

$$\varepsilon = \frac{\triangle y/y}{\triangle x/x} = \frac{\triangle y}{\triangle x}\frac{x}{y} = m\frac{SQFT}{PRICE}$$

$$= (2\beta_2 SQFT)\frac{SQFT}{PRICE} = 2\beta_2\frac{SQFT^2}{PRICE} = 2\beta_2\frac{SQFT^2}{\beta_1 + \beta_2 SQFT^2}.$$

Hence, we estimate the elasticity by

$$\hat{\varepsilon} = 2b_2\frac{SQFT^2}{PRICE} = 2b_2\frac{SQFT^2}{b_1 + b_2}$$

R

```
# PRICE = beta1+ beta2*SQFT^2 + e

mod3 <- lm(price~I(sqft^2), data=br)

summary(mod3)

b1 <- coef(mod3)[[1]]

b2 <- coef(mod3)[[2]]

sqftx=c(2000, 4000, 6000)      #given values for pricex=b1+b2*sqftx^2 #prices corresponding to given sqft

DpriceDsqft <- 2*b2*sqftx.      # marginal effect of sqft on price

elasticity = DpriceDsqft*sqftx/pricex

curve(b1+b2*x^2, col="red", add=TRUE)    # add the quadratic curve to the scatter plot
```

Log transformation

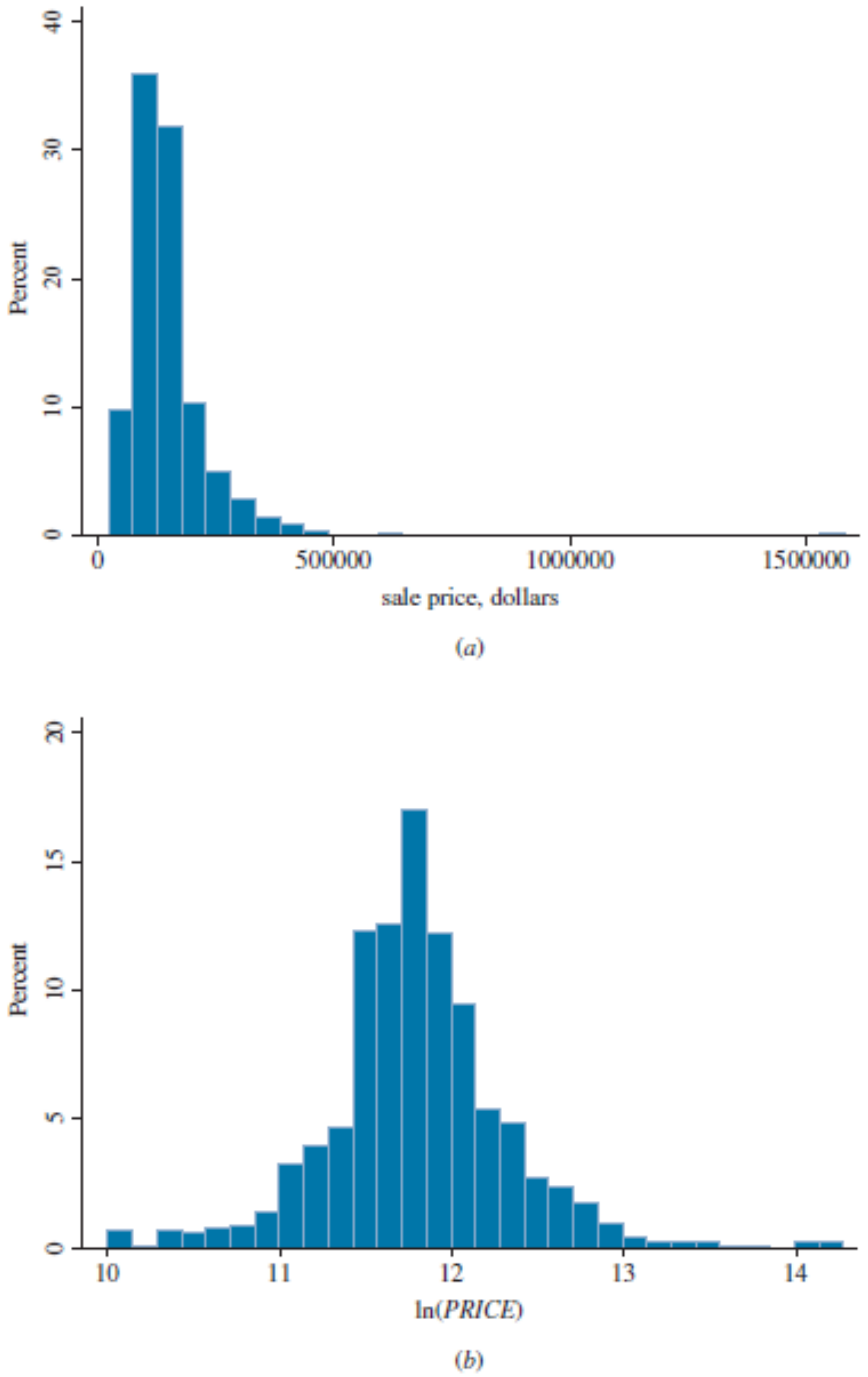
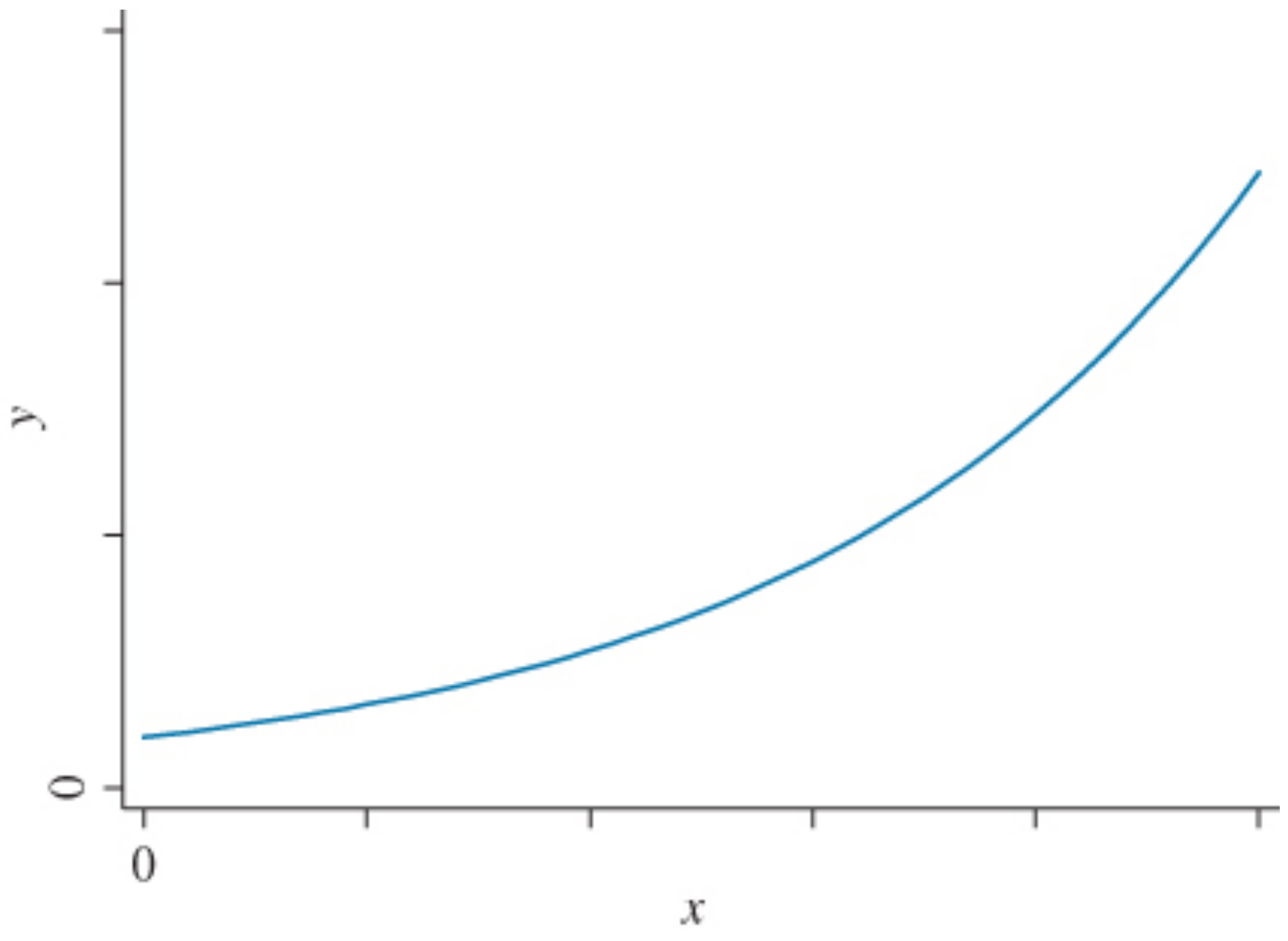
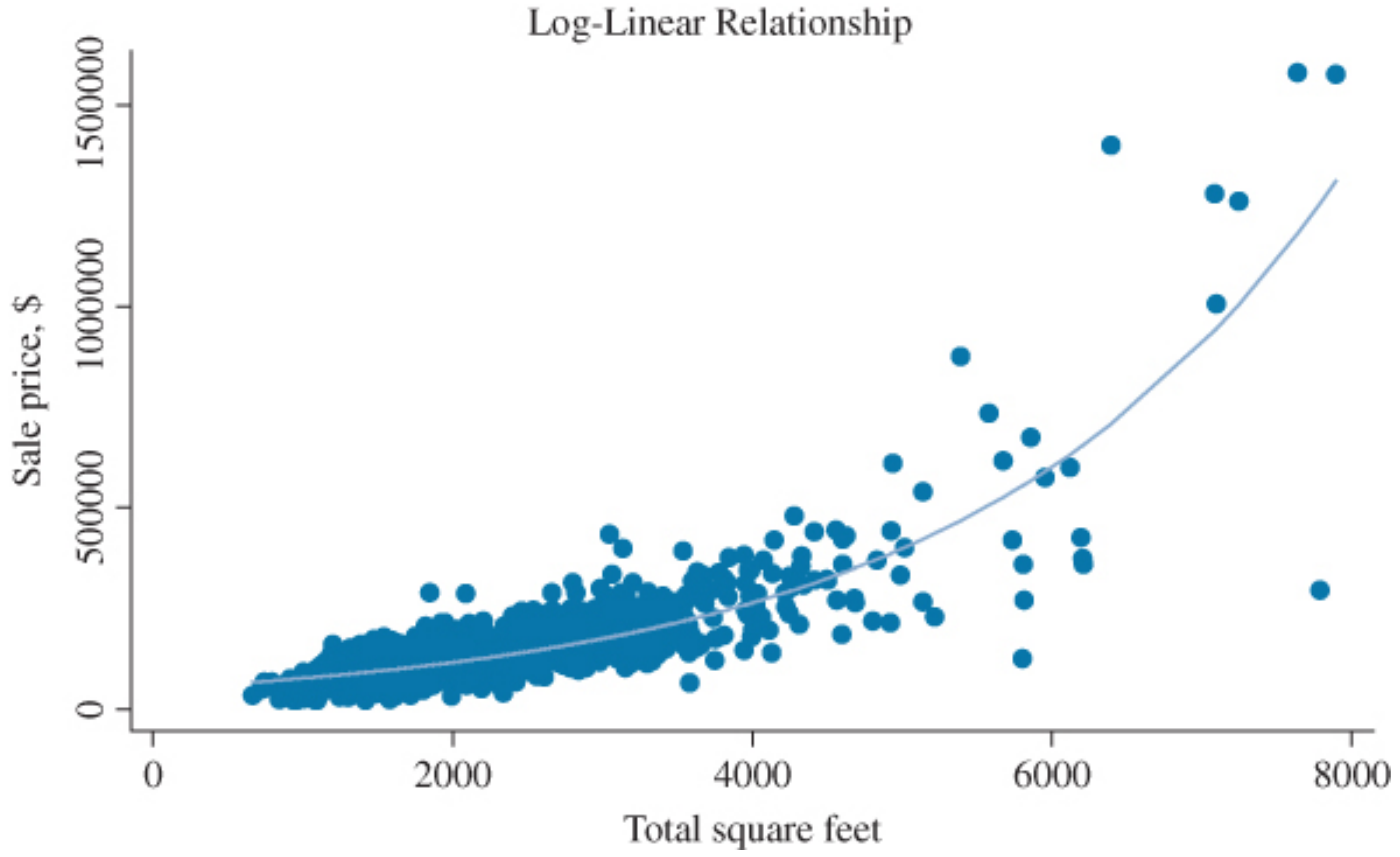


FIGURE 2.16 (*a*) **Histogram of** ***PRICE***. (*b*) **Histogram of ln(*****PRICE*****).**

The log-linear model





The slope

The log-linear equation:

$$log(PRICE) = \beta 1 + \beta 2SQFT\,.$$

It is easy to see

$$PRICE = \exp(\beta_1 + \beta_2 SQFT)\,.$$

The slope is

$$m = \frac{dy}{dx} = \beta_2 \exp(\beta_1 + \beta_2 SQFT),$$

$$\hat{m} = b_2 \exp(b_1 + b_2 SQFT)\,.$$

Interpretations: When the size of the house is SQFT , the expected PRICE increases about $\beta_2 exp(\beta_1 + \beta_2 SQFT)$ unit with an additional square foot.

The elasticity

The elasticity is

$$\begin{aligned}\varepsilon &= \frac{\triangle y/y}{\triangle x/x} \\ &= m\frac{x}{y} \\ &= (\beta_2 \exp(\beta_1 + \beta_2 SQFT))\frac{SQFT}{PRICE} \\ &= \beta_2 SQFT,\end{aligned}$$

We estimate the elasticity by

$$\hat{\varepsilon} = b_2 SQFT\,.$$

Interpretations: While the size of the house is $SQFT$ and it increases one percent, the expected RRICE increases about $(b_2 SQFT) \times 100\,\%$.

R

```
# log(SQFT) = beta1 + beta2 SQFT + e_i
data(br)
hist(br$price, col='grey')
hist(log(br$price), col='grey')
mod4 <- lm(log(price)~sqft, data=br)
mod4
ordat <- br[order(br$sqft), ] #order the dataset
mod4 <- lm(log(price)~sqft, data=ordat)
mod4
plot(br$sqft, br$price, col="grey")
lines(exp(fitted(mod4))~ordat$sqft,
      col="blue", main="Log-linear Model")
```

Choosing a functional form

A challenging question:

- Sum squared of residuals (SSE), essentially $\hat{\sigma}^2$
- Informal way using visualization tools

2.9 Regression with Indicator Variables

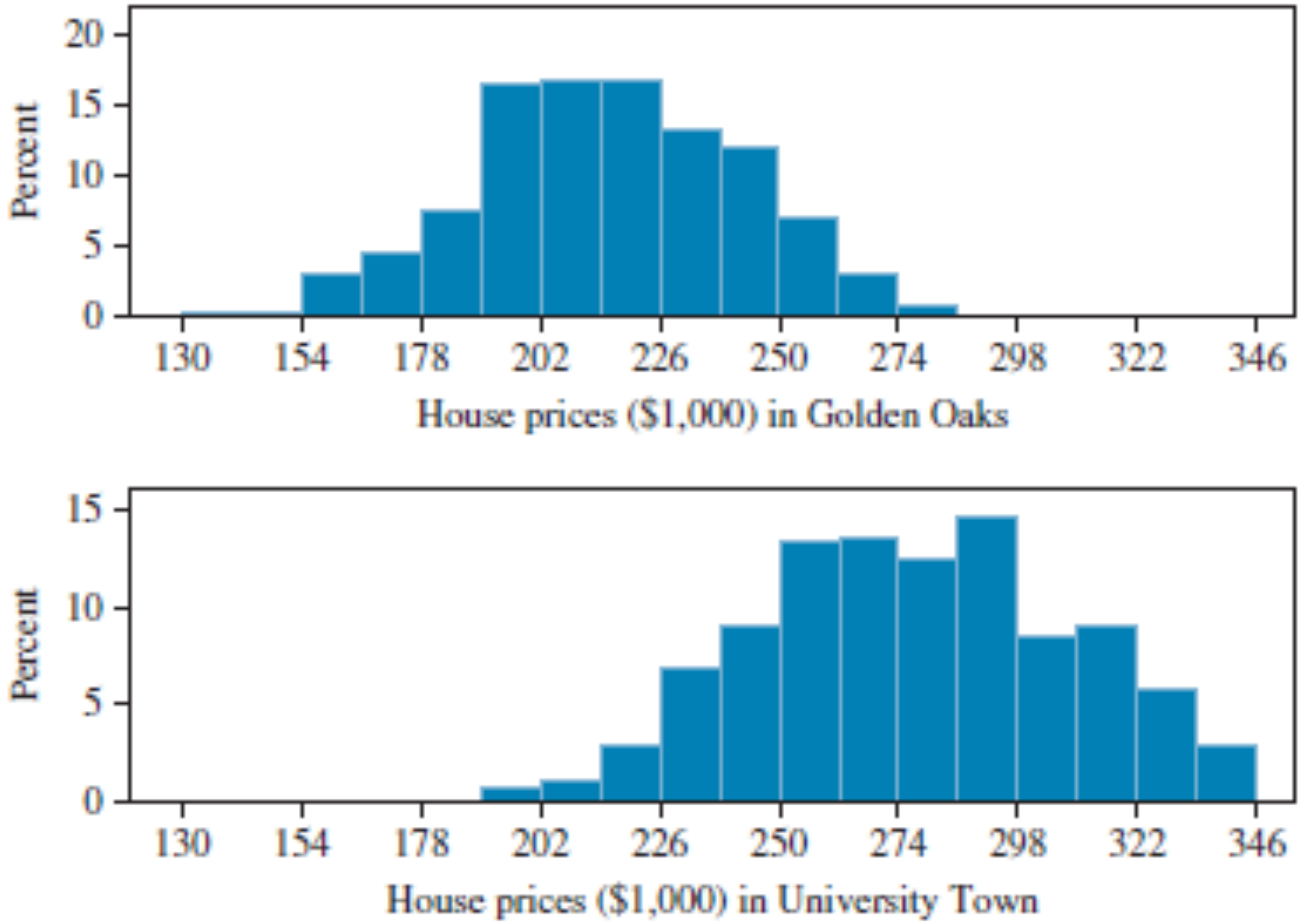


FIGURE 2.18 **Distributions of house prices.**

Dummy variable

Let

$$UTOWN = \begin{cases} 1, & \text{if a house is in University Town,} \\ 0, & \text{if a house is in G.} \end{cases}$$

Then, the model is

$$PRICE = \beta_1 + \beta_2 UTOWN + e\,.$$

- β_2 is the difference between the population means for house prices in the two neighborhoods.
- The expected price in University Town is $\beta_1 + \beta_2$.
- The expected price in Golden Oaks is β_1.

The estimated regression is

$$\widehat{PRICE} = b_1 + b_2 UTOWN = 215.7325 + 61.5091 UTOWN\,.$$

R

```
data(utown)
?utown
price0bar <- mean(utown$price[which(utown$utown==0)])
price1bar <- mean(utown$price[which(utown$utown==1)])
# See the difference
mod5 <- lm(price~utown, data=utown)
b1 <- coef(mod5)[[1]]
b2 <- coef(mod5)[[2]]
```

Appendix: Sampling distributions of b_1 and b_2

The sampling distributions of LSE

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N \sum (x_i - \bar{x})^2}\right),$$
$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum (x_i - \bar{x})^2}\right)$$

Theorem A: Linear combinations of independent normal distributions remain a
Linear combinations of independent normal distributions remain a normal distribution. Specifically, if $X_i \sim N(\mu_i, \sigma^2)$ and X_i are independent, then

$$\sum a_i X_i \sim N\left(\sum a_i \mu_i, \sigma^2 (\sum a_i^2)\right).$$

Derivations for $b_2 \sim N(\beta_2, \frac{\sigma^2}{\sum (x_i - \bar{x})^2})$.

Because $\sum (x_i - \bar{x}) = 0$, we rewrite b_2 by

$$\begin{aligned} b_2 = & \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ = & \sum \frac{(x_i - \bar{x})y_i}{\sum (x_i - \bar{x})^2} \\ = & \sum \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right) y_i \\ = & \sum w_i y_i, \end{aligned}$$

where $w_i = \dfrac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$.

b_2 is a linear combination of normal distribution, and hence a normal distribution.

To find the expectation and variance of b_2, note the following identities:

1. $\sum w_i = 0$
2. $\sum w_i x_i = \sum \frac{(x_i - \bar{x})x_i}{\sum (x_i - \bar{x})^2} = \sum \frac{(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}}{\sum (x_i - \bar{x})^2} = \sum \frac{T(x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 1$
3. $\sum w_i^2 = \sum \left(\frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right)^2 = \frac{1}{\sum (x_i - \bar{x})^2}$

Rewrite

$$
\begin{aligned}
b_2 &= \sum w_i(\beta_1 + \beta_2 x_i + e_i) \\
&= \beta_1 \sum w_i + \beta_2 \sum w_i x_i + \sum w_i e_i \\
&= \beta_2 + \sum w_i e_i \,.
\end{aligned}
$$

Therefore,

$$E(b_2) = \beta_2$$

$$var(b_2) = \sum w_i^2 \sigma^2 = \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2} \,.$$

Derivations for $b_1 \sim N(\beta_1, \sigma^2 \frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2})$.

Now, we write

$$
\begin{aligned}
b_1 &= \bar{y} - b_2\bar{x} \\
&= \sum (\frac{1}{N} - \bar{x}w_i)y_i \\
&= \sum (\frac{1}{N} - \bar{x}w_i)(\beta_1 + \beta_2 x_i + e_i) \\
&= (\beta_1 - \beta_1\bar{x} \sum w_i) + \left(\beta_2\bar{x} - \bar{x}\beta_2 \sum x_i w_i\right) + \sum (\frac{1}{N} - \bar{x}w_i)e_i \\
&= \beta_1 + \sum (\frac{1}{N} - \bar{x}w_i)e_i \,.
\end{aligned}
$$

Hence, b_1 is a normal distribution.

To find the expectation and variance of b_1, it is easy that

$$E(b_1) = \beta_1 + \sum [(\frac{1}{N} - \bar{x}w_i)0] = \beta_1$$

$$\begin{aligned}
var(b_1) &= \sum (\frac{1}{N} - \bar{x}w_i)^2\sigma^2 \\
&= \sigma^2 \left(\sum \frac{1}{N^2} - 2\sum \frac{\bar{x}}{N}w_i + \bar{x}^2 \sum w_i^2 \right) \\
&= \sigma^2 \left(\frac{1}{N} + \bar{x}^2 \frac{1}{\sum (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \frac{\sum (x_i - \bar{x})^2 + N\bar{x}^2}{N\sum (x_i - \bar{x})^2} \\
&= \sigma^2 \frac{(\sum x_i^2 - 2N\bar{x}^2 + N\bar{x}^2) + N\bar{x}^2}{N\sum (x_i - \bar{x})^2} = \sigma^2 \frac{\sum x_i^2}{N\sum (x_i - \bar{x})^2}.
\end{aligned}$$