Applied Econometric QEM
Theme 2
Regression Model

Chapters from 2 to 6 of PoE

Michał Rubaszek
Based on presentation by Walter R. Paczkowski

Economic and Econometric Model

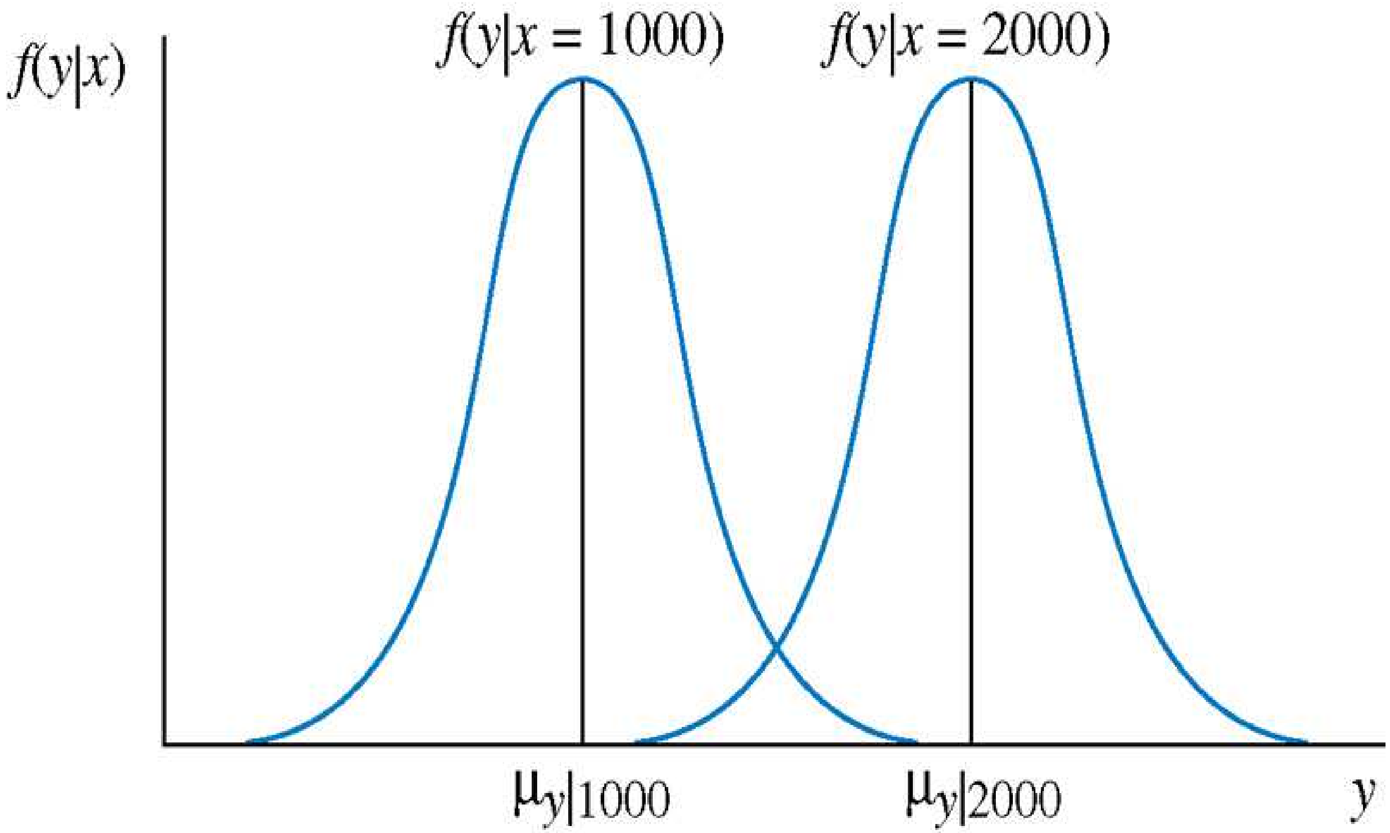
- Economists interested in relationships between variables

- Example: the theory tells us that expenditures *y* depends on income *x*

- We call *y* the “**dependent variable**” and *x* the “**independent**” or “**explanatory**” variable

- In econometrics *y* is a **random variable** and we need to use data to learn about the relationship

- The econometric model helps to calculate conditional mean $\boldsymbol{E(y|x)} = \boldsymbol{\mu_{y|x}}$ and the conditional variance $\boldsymbol{\sigma^2}$, which give us valuable information about the population we are considering



(*b*)

- To investigate the relationship we build an **economic model** and a corresponding **econometric model:**

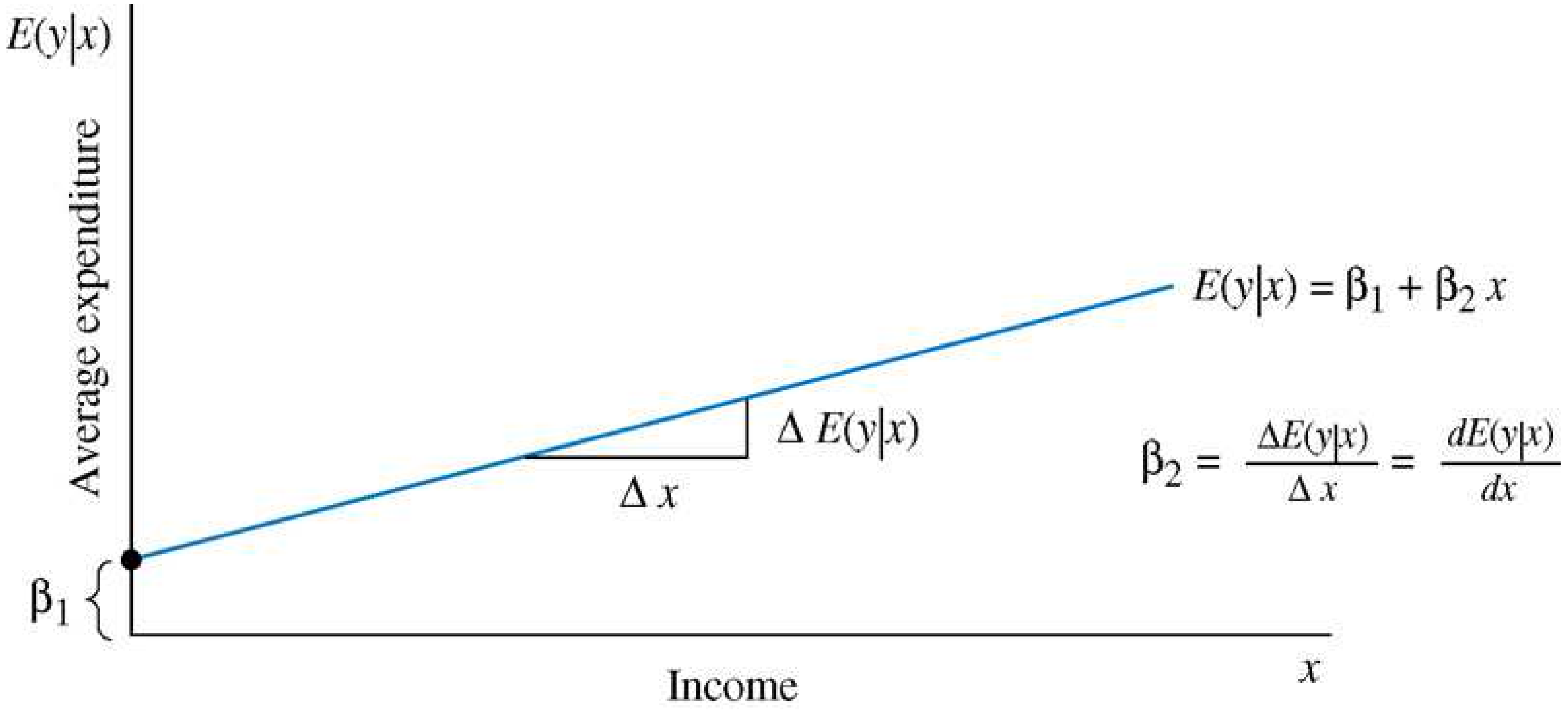
$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$$

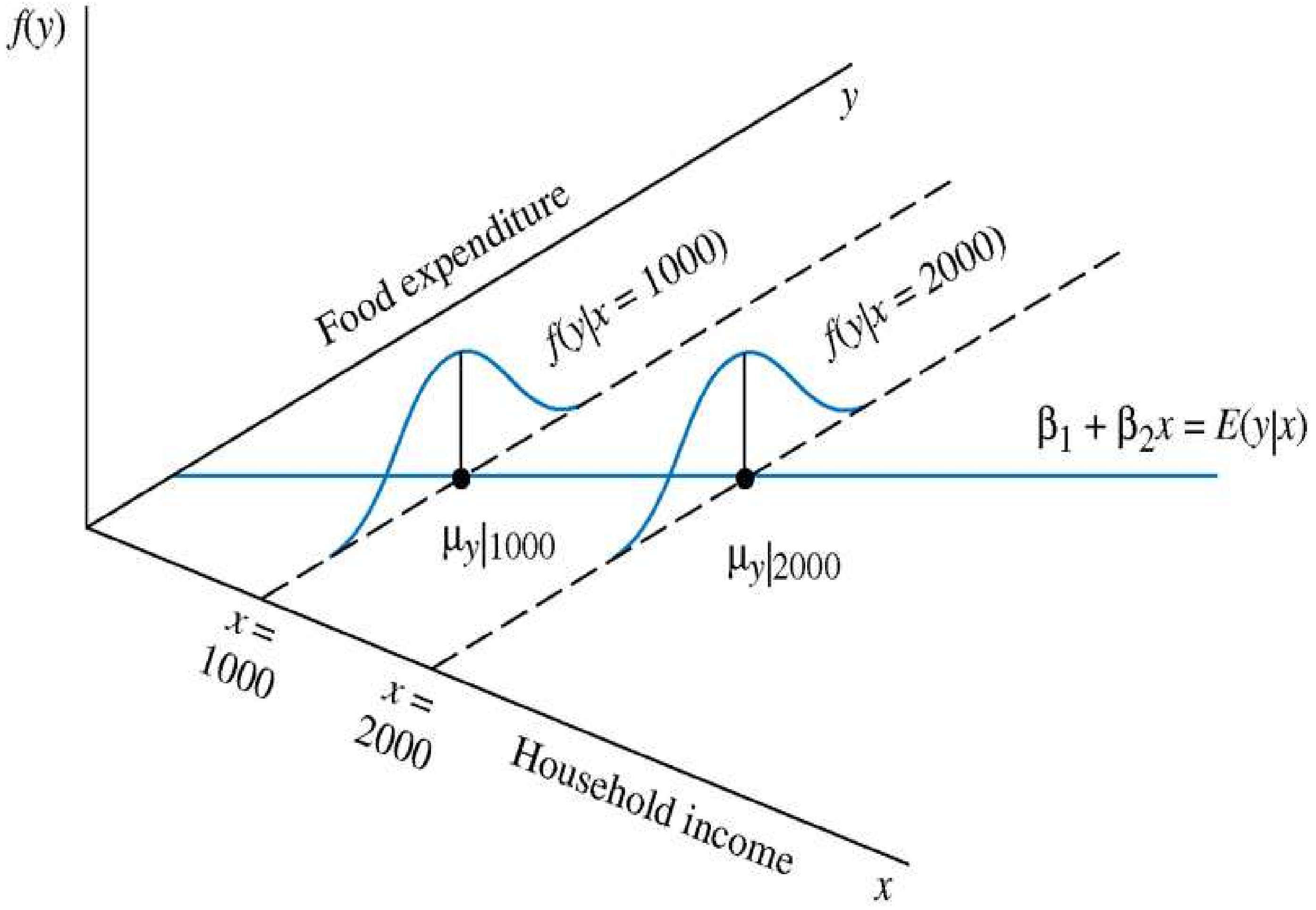
β_1 – intercept

β_2 – slope

- Interpretation of the slope – derivative of the expected value of *y* given an *x* value:

$$\beta_2 = \frac{\Delta E(y\,|\,x)}{\Delta x} = \frac{dE(y\,|\,x)}{dx}$$


$E(y|x)$
Average expenditure
$E(y|x) = \beta_1 + \beta_2 x$
$\Delta E(y|x)$
Δx
$\beta_2 = \dfrac{\Delta E(y|x)}{\Delta x} = \dfrac{dE(y|x)}{dx}$
β_1
Income
x


$f(y)$
y
Food expenditure
$f(y|x = 1000)$
$f(y|x = 2000)$
$\beta_1 + \beta_2 x = E(y|x)$
$\mu_{y|1000}$
$\mu_{y|2000}$
$x =$ 1000
$x =$ 2000
Household income
x

Multiple regression model – a general case:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_K x_K + e$$

β_k measures the effect of a change in x_k upon the expected value of y, all other variables held constant (ceteris paribus)

$$\beta_k = \left.\frac{\Delta E(y)}{\Delta x_k}\right|_{\text{other xs held constant}} = \frac{\partial E(y)}{\partial x_k}$$

Economic vs. econometric model

Economic model

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

Econometric model

$$y_i = \beta_1 + \beta_2 x_i + e_i, \qquad e_i \sim N(0, \sigma^2)$$

Assumptions of linear econometric model:

A1: The value of *y*, for each value of *x*, is:

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

A2: The expected value of the random error *e* is:

$$E(e) = 0 \leftrightarrow E(y|x) = \beta_1 + \beta_2 x$$

A3: The variance of the random error *e* is:

$$var(e) = var(y) = \sigma^2$$

A4: The covariance between e_i and e_j for $i \neq j$ is:

$$cov(e_i, e_j) = 0$$

A5: Variable *x* is not random and takes at least 2 different values

A6+: Random term *e* is *normally distributed:*

$$e \sim N(0, \sigma^2)$$

Assumptions for a multiple regression model:

A1. $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i,\ i = 1, \ldots, N$

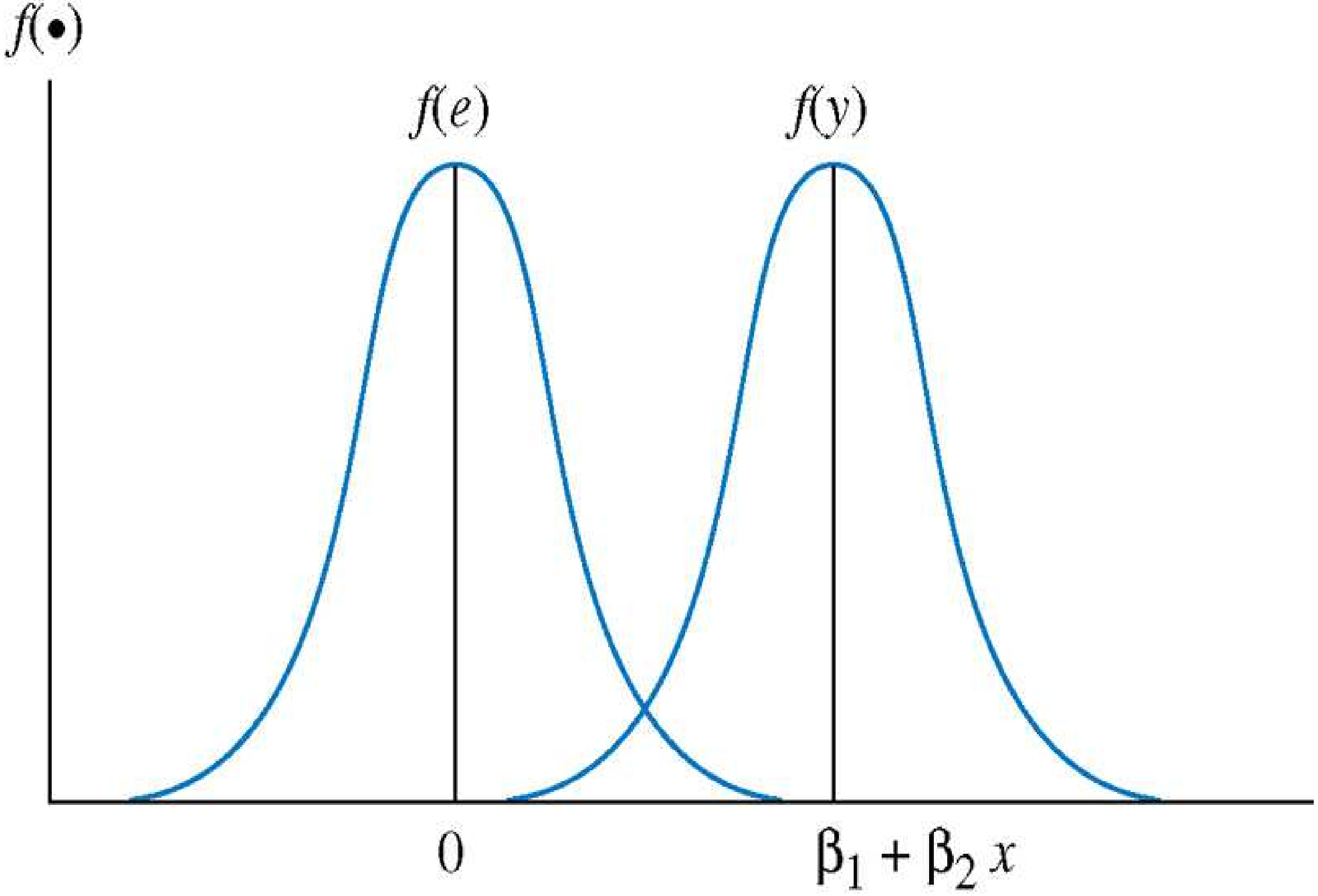
A2. $E(y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \Leftrightarrow E(e_i) = 0$

A3. $\operatorname{var}(y_i) = \operatorname{var}(e_i) = \sigma^2$

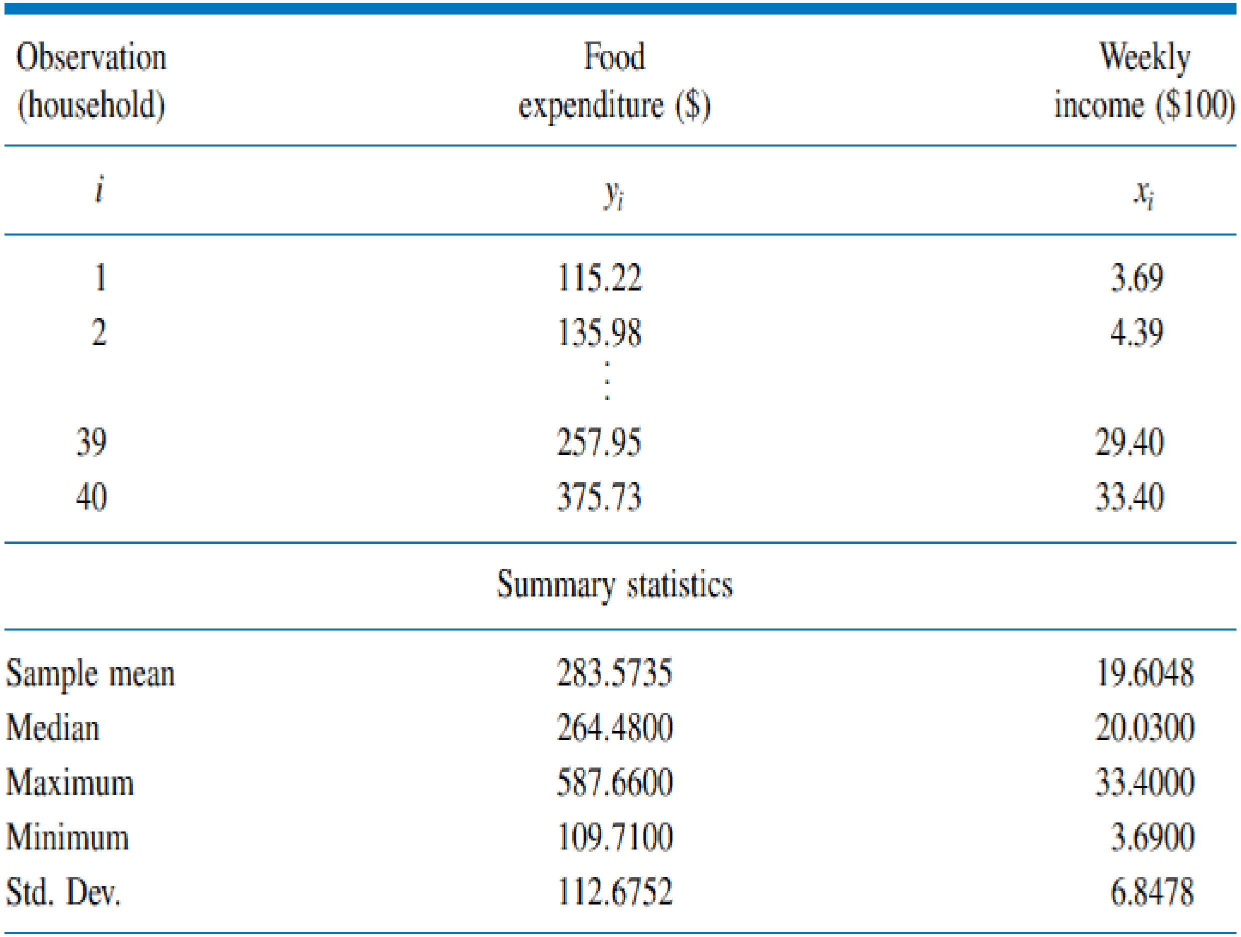
A4. $\operatorname{cov}(y_i, y_j) = \operatorname{cov}(e_i, e_j) = 0$

A5. The values of each x_{tk} are not random and are not exact linear functions of the other explanatory variables

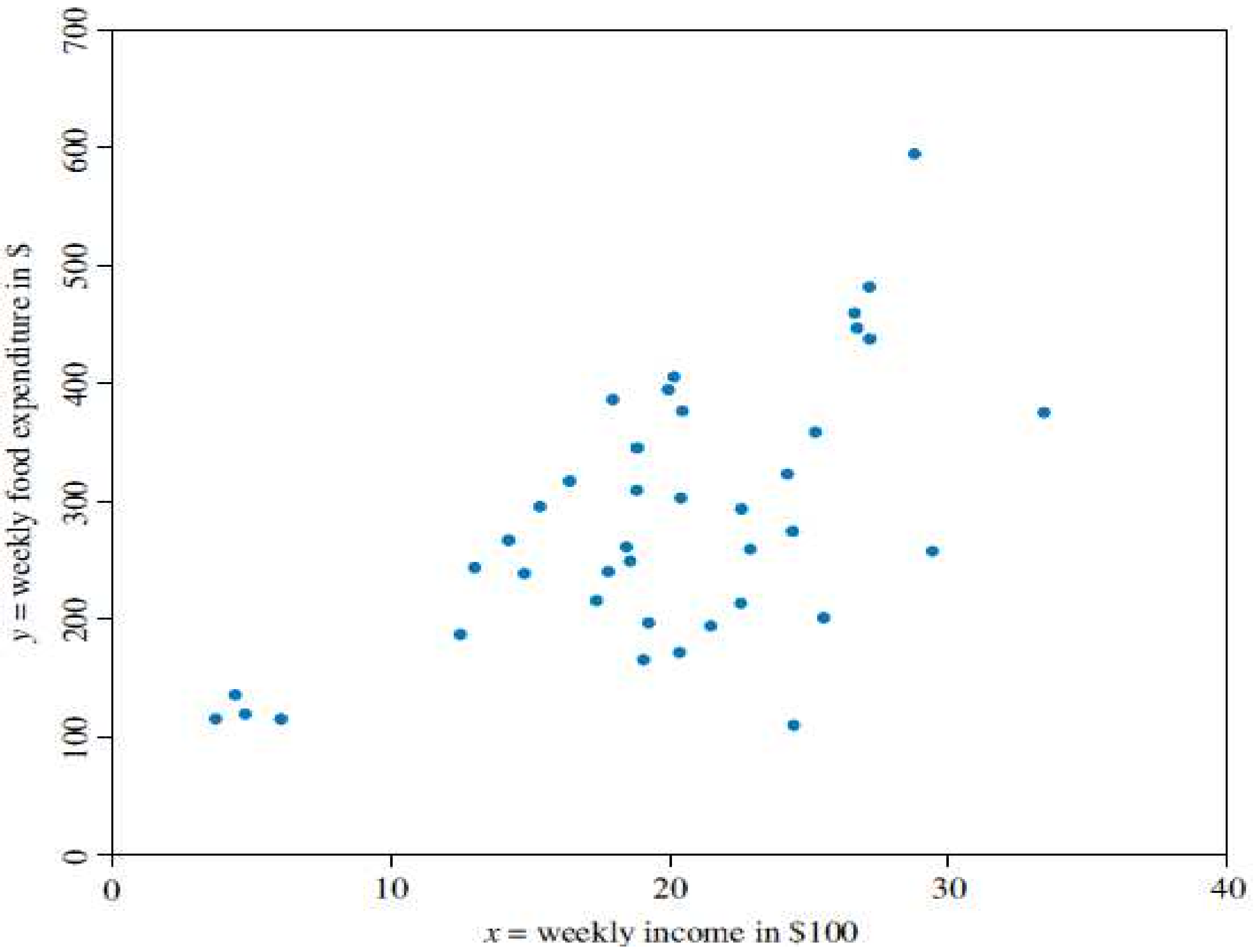
A6. $y_i \sim N\left[(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}), \sigma^2\right] \Leftrightarrow e_i \sim N(0, \sigma^2)$


$f(\bullet)$
$f(e)$
$f(y)$
0
$\beta_1 + \beta_2 x$

Estimating the Regression Parameters



Observation (household)	Food expenditure ($)	Weekly income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	$\vdots$	
39	257.95	29.40
40	375.73	33.40
	Summary statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. Dev.	112.6752	6.8478


700
600
500
400
300
200
100
0
y = weekly food expenditure in $
0
10
20
30
40
x = weekly income in $100

Fitted values, residuals and least squares

- For any values b_1 and b_2 we can calculate **fitted values**:

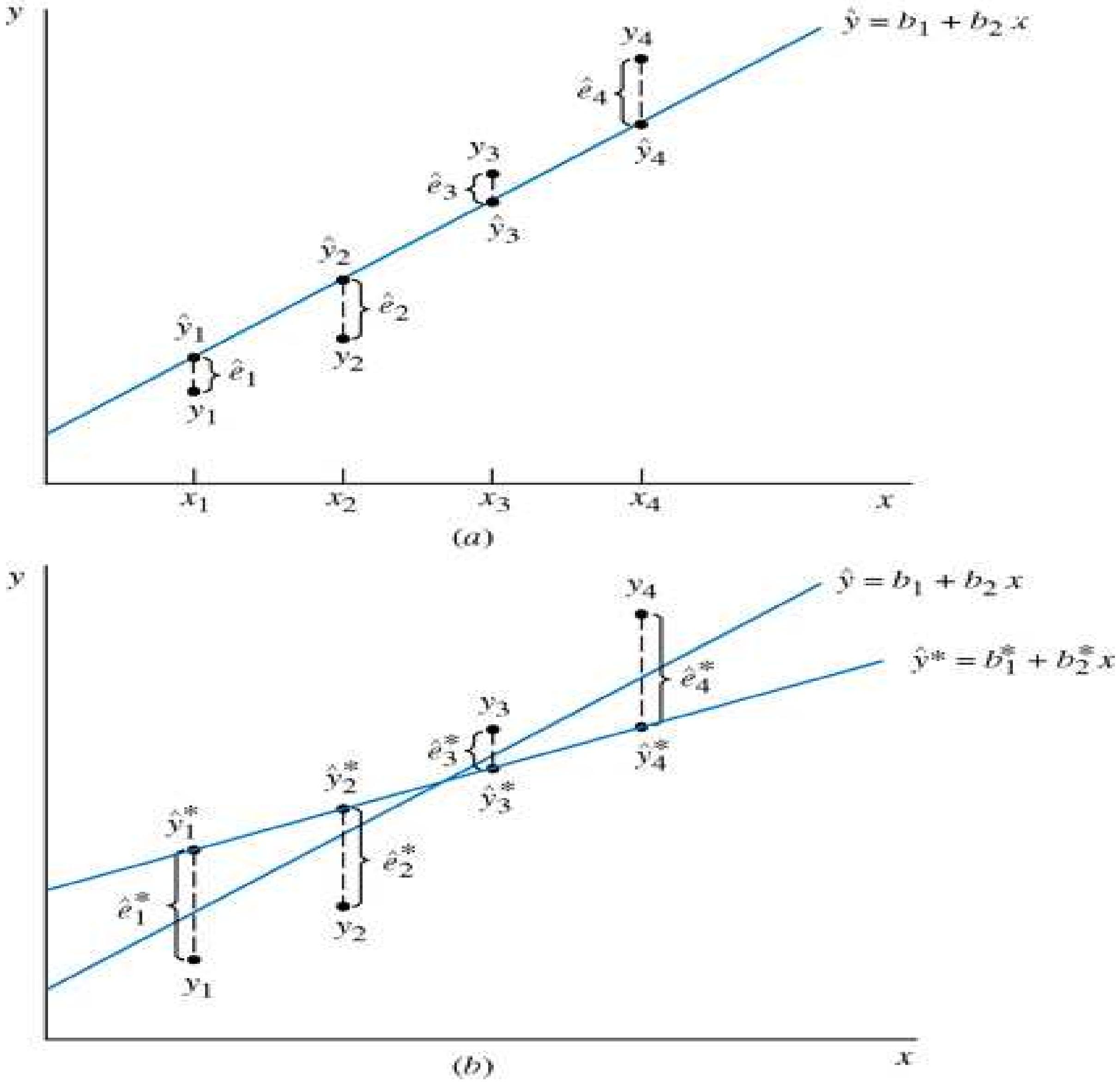
$$\hat{y}_i = b_1 + b_2 x_i$$

and **residuals:**

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

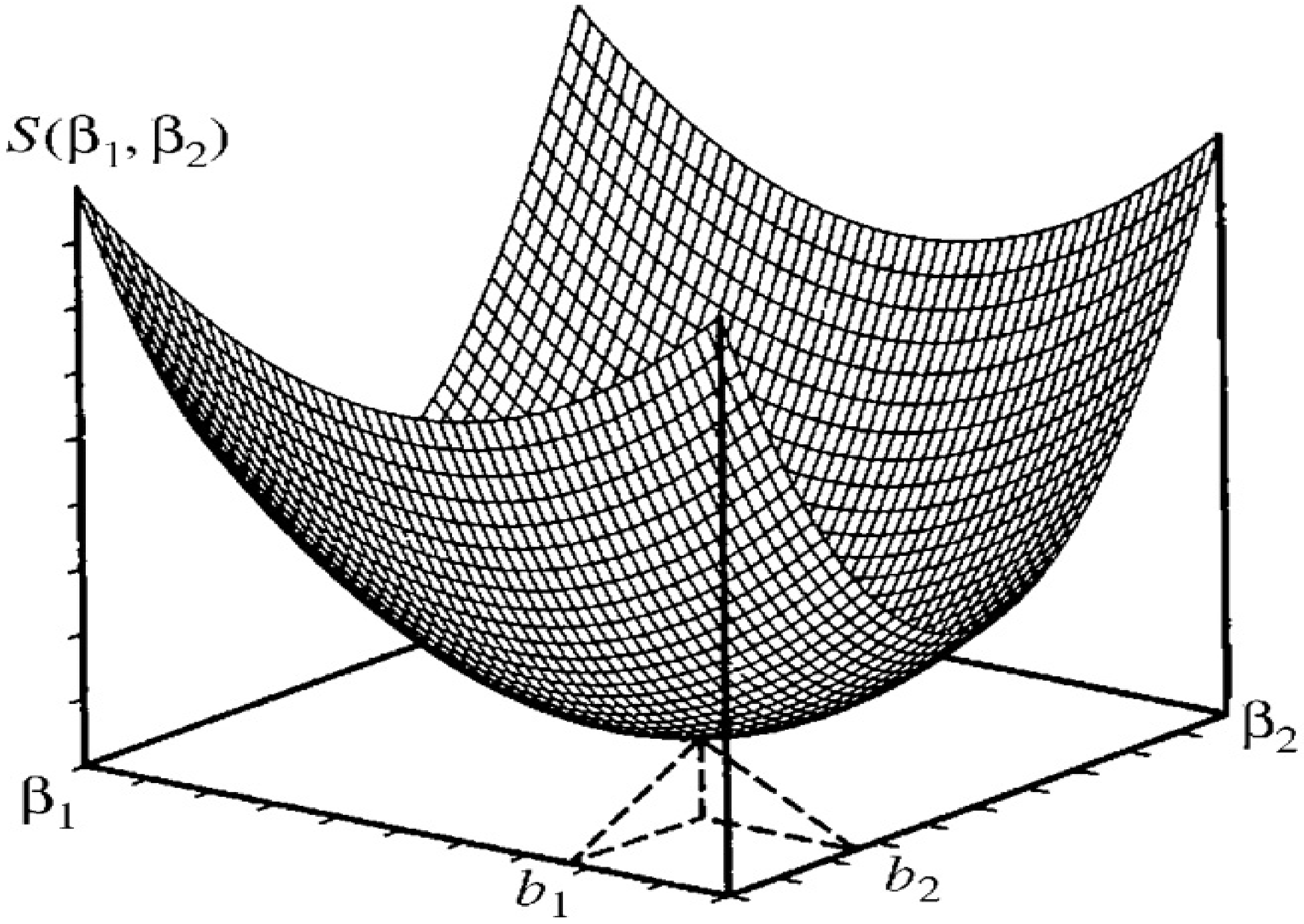
- The least squares values of b_1 and b_2 minimize the sum of squared residuals:

$$SSE = \sum_{i=1}^{N} \hat{\boldsymbol{e}}_i^2 = \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2 = S(b_1, b_2)$$


y
$\hat{y} = b_1 + b_2 x$
y_4
$\hat{e}_4$
$\hat{y}_4$
y_3
$\hat{e}_3$
$\hat{y}_3$
$\hat{y}_2$
$\hat{e}_2$
y_2
$\hat{y}_1$
$\hat{e}_1$
y_1
x_1
x_2
x_3
x_4
x
y
$\hat{y} = b_1 + b_2 x$
y_4
$\hat{y}^* = b_1^* + b_2^* x$
$\hat{e}_4^*$
y_3
$\hat{e}_3^*$
$\hat{y}_4^*$
$\hat{y}_2^*$
$\hat{y}_3^*$
$\hat{y}_1^*$
$\hat{e}_2^*$
$\hat{e}_1^*$
y_2
y_1
x

(a)

(b)


$S(\beta_1, \beta_2)$
β_1
b_1
b_2
β_2

Least squares estimator

- Least squares estimates for the unknown parameters β$_1$ and β$_2$ are obtained my minimizing the sum

$$SSE = \sum_{i=1}^{N} \hat{\boldsymbol{e}}_i^2 = \sum_{i=1}^{N} (y_i - b_1 - b_2 x_i)^2$$

- Solution for one explanatoty variable case:

$$b_2 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$ and $$b_1 = \overline{y} - b_2 \overline{x}$$

Least squares estimator – multiple regression

Multiple regression

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + e_i$$

in a vector form:

$$y_i \ = \boldsymbol{\beta}' \, \boldsymbol{x}_i \ + \ e_i$$

$\boldsymbol{x}_i \ = [1 \; x_{1i} \ \ldots x_{Ki}]'$ - the vector of explanatory variables
$\boldsymbol{\beta} \ \ = [\beta_1 \; \beta_2 \ \ldots \beta_K]'$ - the vector of parameters.

We observe y_i and $\boldsymbol{x}_i$, but don't know the values of β and need to estimate it

Let $\boldsymbol{b}$ be the estimate of $\boldsymbol{\beta}$ so that:

Fitted values: $\hat{y}_i = E(y_i) = \boldsymbol{b}'\boldsymbol{x}_i$

Residuals: $e_i = y_i - \hat{y}_i$

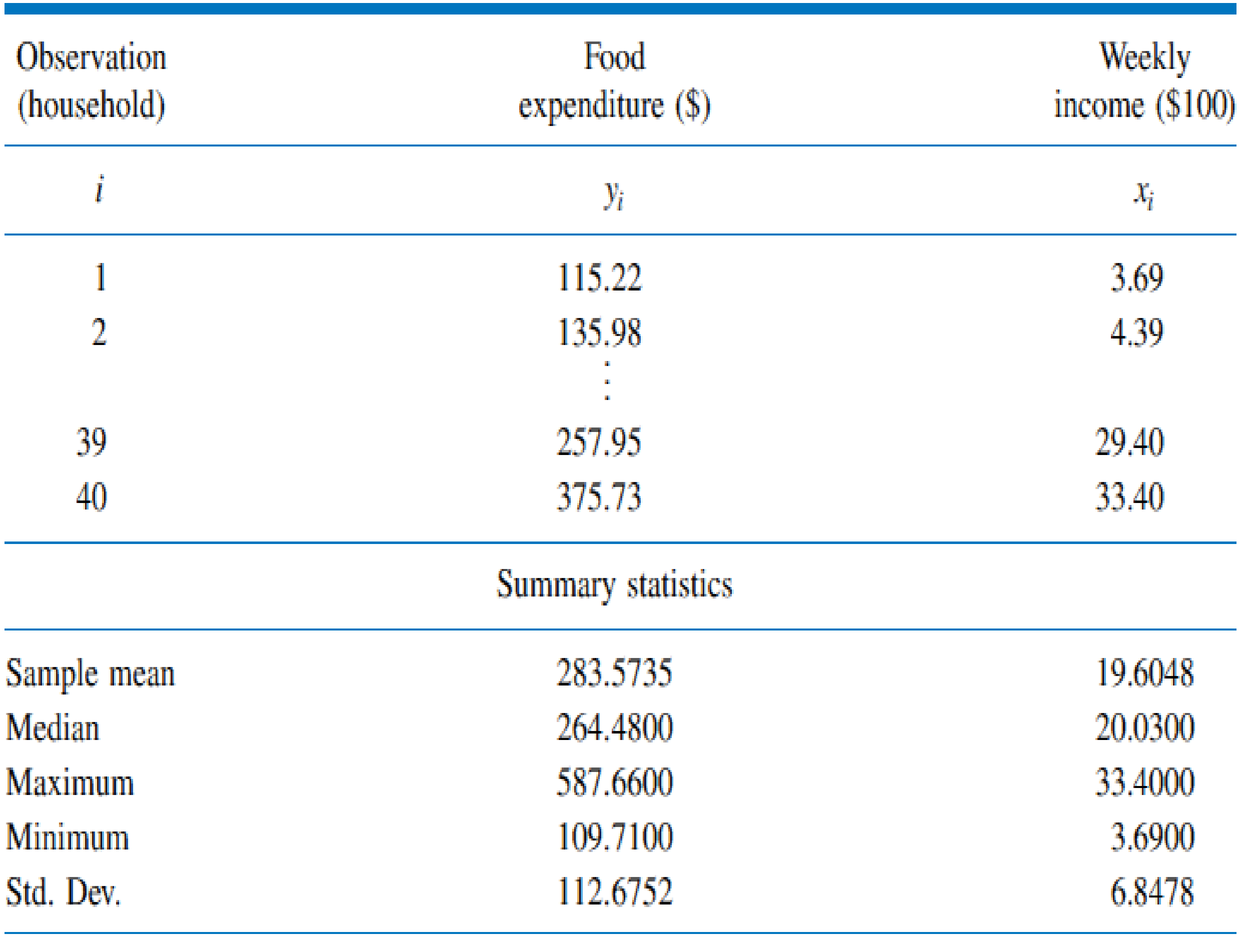
Sum of sq. residuals: $SSE(\boldsymbol{b}) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \boldsymbol{b}'\boldsymbol{x}_i)^2$

Since SSE depends on $\boldsymbol{b}$, we can find $\boldsymbol{b}$ such that the SEE is minimum. The solution is the formula for LS estimator:

$$\boldsymbol{b} = \left(\sum_{t=1}^{T} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{t=1}^{T} \boldsymbol{x}_i y_i\right)$$

- **The LS estimator is a general formula** and is a **random variable**, the properties of which depend on the structure of the model (described by assumptions).
- **LS estimates are numbers** that we obtain by applying the general formulas to the observed data.

Least squares estimator - example



Observation (household)	Food expenditure ($)	Weekly income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. Dev.	112.6752	6.8478

Least squares estimator - example

- We can calculate:

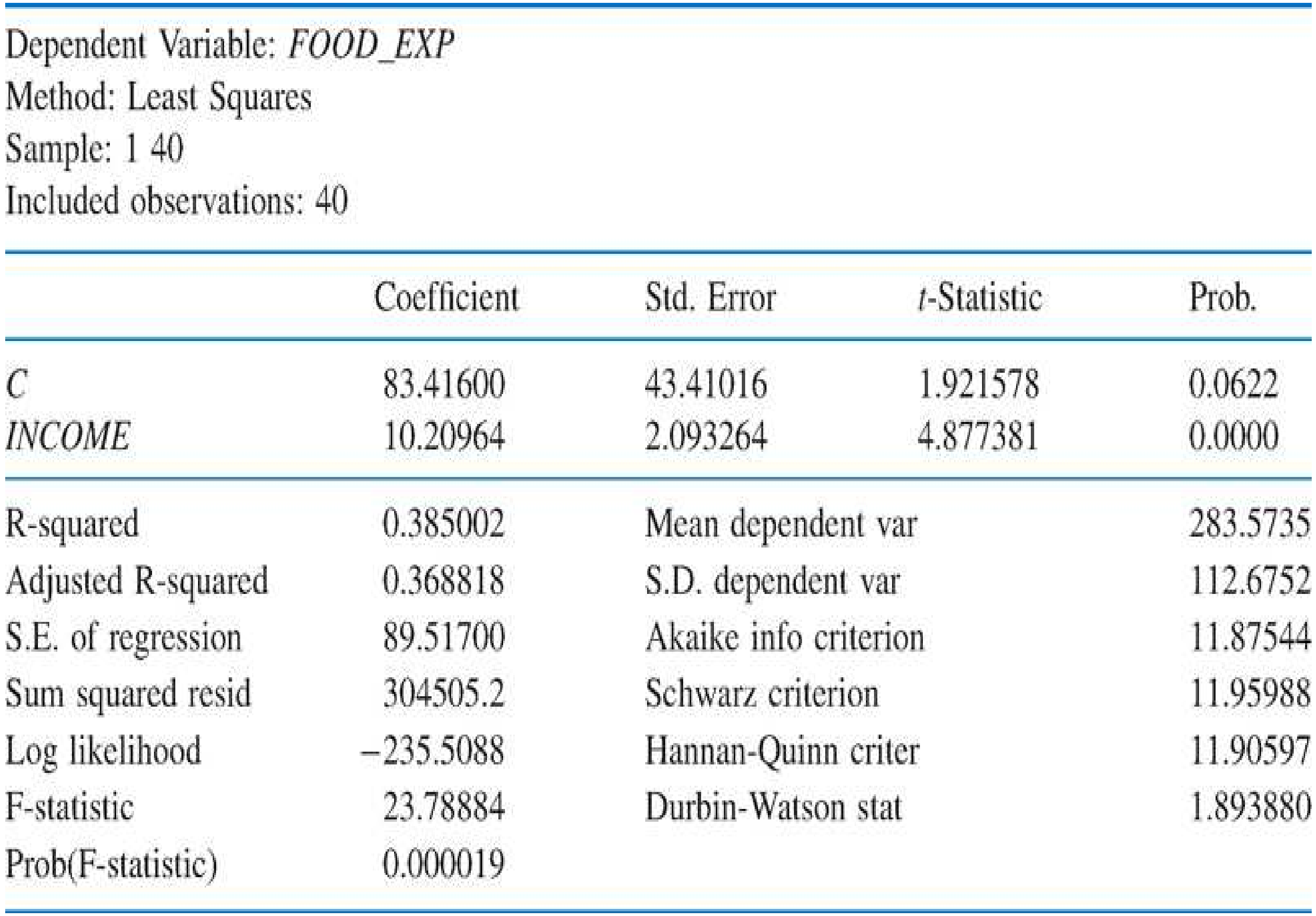
$$b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{18671.2684}{1828.7876} = 10.2096$$

$$b_1 = \bar{y} - b_2 \bar{x} = 283.5735 - (10.2096)(19.6048) = 83.4160$$

- And report that:

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

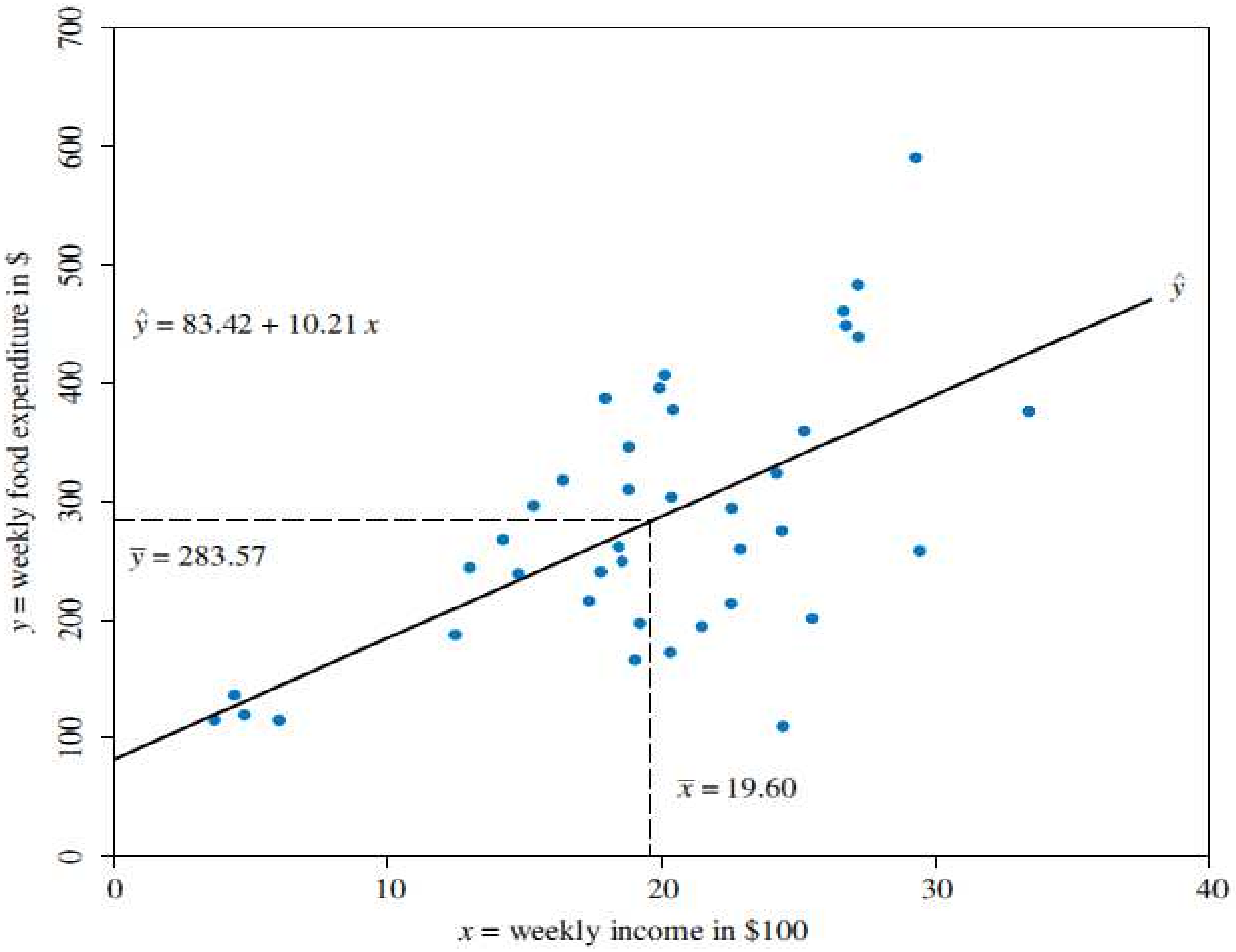
- What interpretation of b_1 and b_2?



Dependent Variable: *FOOD_EXP*
Method: Least Squares
Sample: 1 40
Included observations: 40

	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

R-squared	0.385002	Mean dependent var	283.5735
Adjusted R-squared	0.368818	S.D. dependent var	112.6752
S.E. of regression	89.51700	Akaike info criterion	11.87544
Sum squared resid	304505.2	Schwarz criterion	11.95988
Log likelihood	−235.5088	Hannan-Quinn criter	11.90597
F-statistic	23.78884	Durbin-Watson stat	1.893880
Prob(F-statistic)	0.000019		


$\hat{y} = 83.42 + 10.21\,x$
$\bar{y} = 283.57$
$\bar{x} = 19.60$
$\hat{y}$
y = weekly food expenditure in $
x = weekly income in $100
0
100
200
300
400
500
600
700
0
10
20
30
40

Point prediction

Suppose that we wanted to predict food expenditure for a household with income of $2000, so that x = 20. We obtain:

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

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

Assessing the Least Squares Fit

Notice that LS estimators (do not confuse with estimates) are random variables so we can calculate their expected values, variances, covariances or probability distributions

Given that:

$$b_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\beta_2(x_i - \bar{x}) + e_i)}{\sum(x_i - \bar{x})^2}$$

We can derive:

$$b_2 = \beta_2 + \frac{1}{\sum(x_i - \bar{x})^2}\sum(x_i - \bar{x})e_i = \beta_2 + \sum w_i e_i$$

A5 [x is not random] and A2 [$E(e) = 0$] imply that:

$$E(w_i e_i) = E\left(\frac{x_i - \bar{x}}{\sum(x_i - \bar{x})^2} e_i\right) = 0$$

This means that the estimator b_2 is **unbiased:**

$$E(b_2) = E(\beta_2 + \sum w_i e_i) = \beta_2 + \sum E(w_i e_i) = \beta_2$$

Important: unbiasedness does not say that an estimate from any one sample is close to the true parameter value (estimate ≠ estimator). For different samples the estimates of b_1 and b_2 are different – they are just single draws from the distribution of the estimator

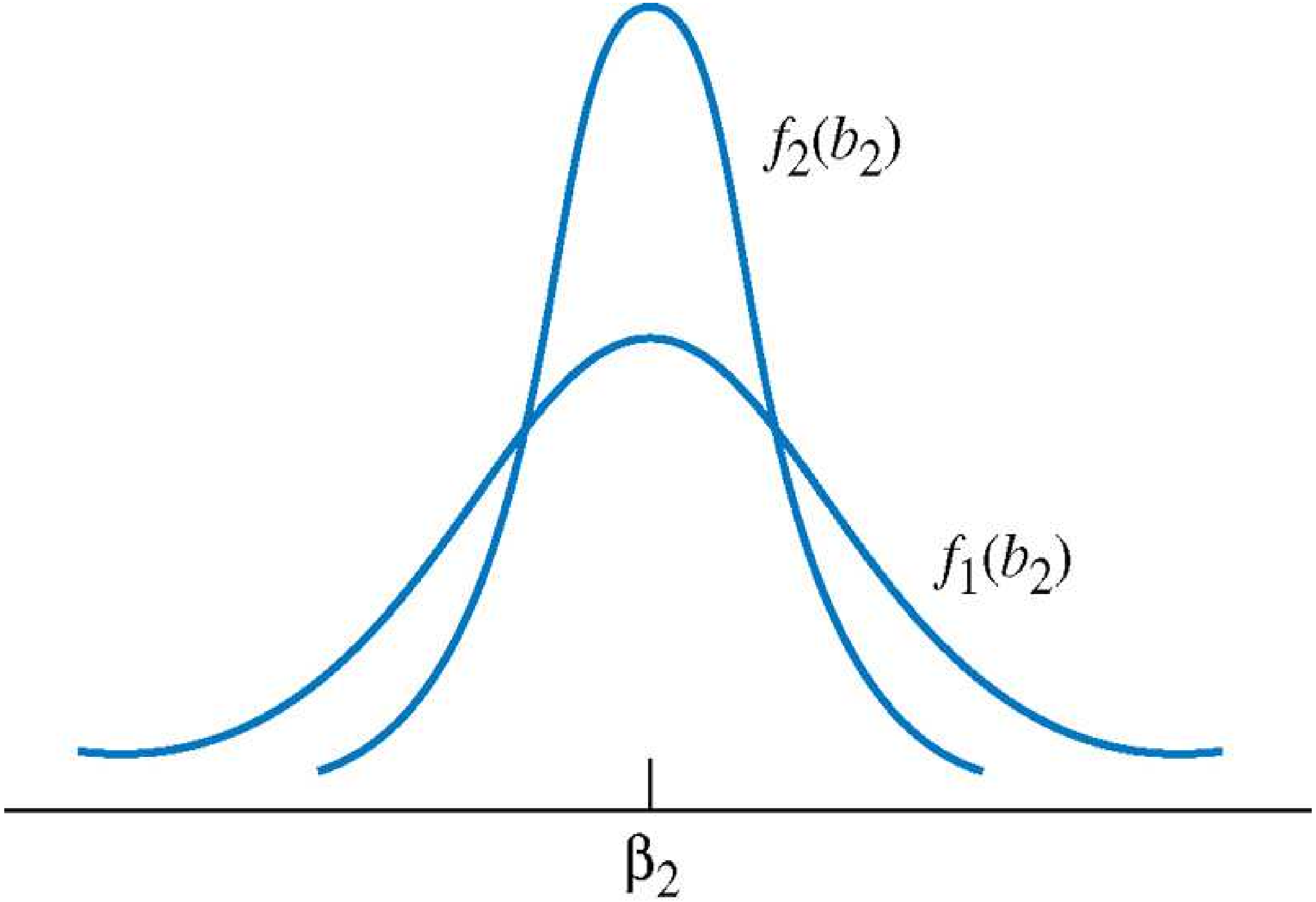
- Question: *what is the variance of the LS estimator?*

- If A1-A5 hold then the variances and covariance of b_1 and b_2 are

$$\operatorname{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - x)^2} \right] \qquad \operatorname{var}(b_2) = \frac{\sigma^2}{\sum (x_i - x)^2}$$

- Precision of estimates decreases with σ^2 and increases with N
- Consistent estimators: for $N \to \infty$ the variance converges to 0
- Effective estimators: estimators with the smallest variance

The variance of b_2 is defined as $\mathrm{var}(b_2) = E[b_2 - E(b_2)]^2$



Gauss-Markov theorem

Under A1-A5 of the linear regression model, the LS estimators have the smallest variance of all linear and unbiased estimators. They are the **Best Linear Unbiased Estimators (BLUE)**

- Notice that:
  1. The LS estimators are "best" when compared to other linear and unbiased estimators - the Theorem does *not* say about all *possible* estimators.
  2. The LS estimators are the best within their class because they have the minimum variance.
  3. In order for the Gauss-Markov Theorem to hold, assumptions A1-A5 must be true. If any of these assumptions are *not* true, then LS is *not* the best linear unbiased estimator.

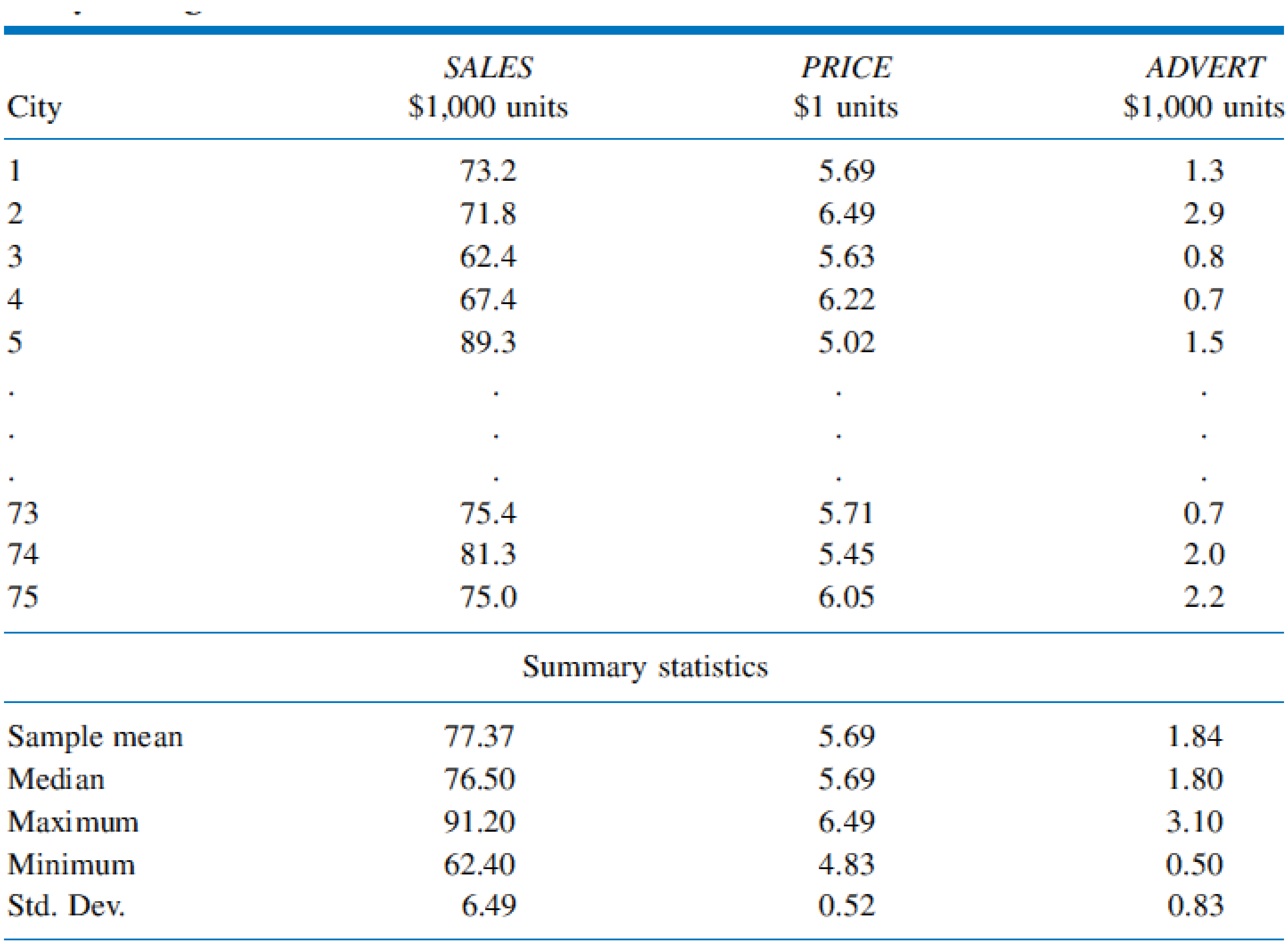
Interval estimation

Let us focus on a multiple regression model in which sales revenue depends on price and advertising expenditure:

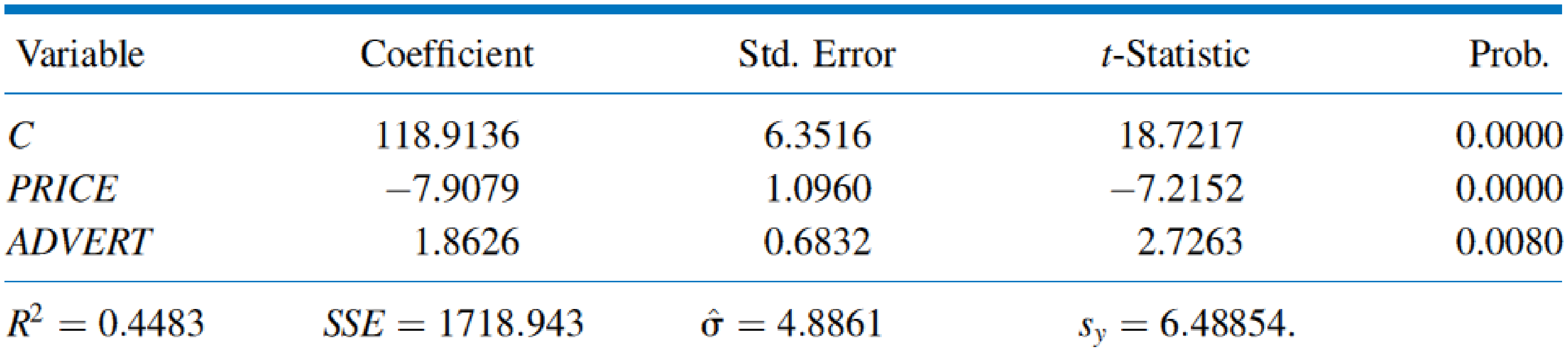
$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT$$

The econometric model is:

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$



	SALES	*PRICE*	*ADVERT*
City	$1,000 units	$1 units	$1,000 units
1	73.2	5.69	1.3
2	71.8	6.49	2.9
3	62.4	5.63	0.8
4	67.4	6.22	0.7
5	89.3	5.02	1.5
.	.	.	.
.	.	.	.
.	.	.	.
73	75.4	5.71	0.7
74	81.3	5.45	2.0
75	75.0	6.05	2.2
	Summary statistics		
Sample mean	77.37	5.69	1.84
Median	76.50	5.69	1.80
Maximum	91.20	6.49	3.10
Minimum	62.40	4.83	0.50
Std. Dev.	6.49	0.52	0.83



Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	118.9136	6.3516	18.7217	0.0000
PRICE	−7.9079	1.0960	−7.2152	0.0000
ADVERT	1.8626	0.6832	2.7263	0.0080

$R^2 = 0.4483$ $SSE = 1718.943$ $\hat{\sigma} = 4.8861$ $s_y = 6.48854.$

- Interpretations of the results:
  1. The coefficient on *PRICE:*
     with advertising held constant, an increase in price of $1 will lead to a fall in monthly revenue of $7,908
  2. The coefficient on *ADVERT*:
     with price held constant, an increase in advertising expenditure of $1,000 will lead to an increase in sales revenue of $1,863

How to assess the precision of our estimates?

If A1–A5 hold, and the errors are normally distributed (A6), then the LS estimators are normally distributed

$$\boldsymbol{b} \sim N(\beta, \Sigma)$$

The variance of LS estimator is:

$$\Sigma = var(\boldsymbol{b}) = \begin{bmatrix} var(b_1) & \cdots & cov(b_1, b_K) \\ \vdots & \ddots & \vdots \\ cov(b_K, b_1) & \cdots & var(b_K) \end{bmatrix} = \sigma^2 (\textstyle\sum_{t=1}^{T} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}$$

However, we don't know the variance $\boldsymbol{\sigma^2}$… so we need to substitute it with the unbiased estimator:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N-K}$$

where $\mathrm{N} - K$ is the number of degrees of freedom

For sales model $SSE = 1718.943$ so that:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{75} \hat{e}_i^2}{N-K} = \frac{1718.943}{75-3} = 23.874$$

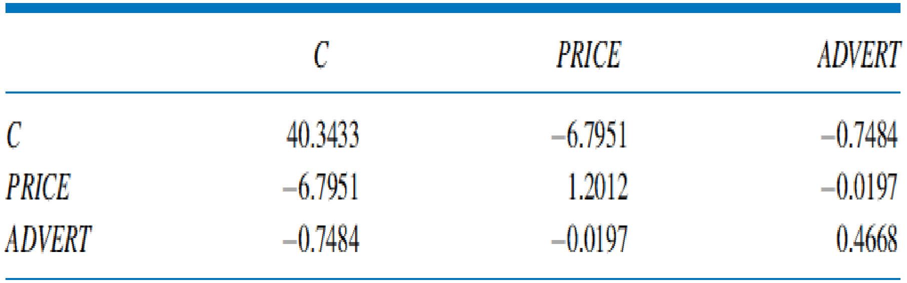
$$\hat{\sigma} = \sqrt{23.874} = 4.8861$$

Now we are ready to calculate the precision of estimates with the feasible formula:

$$\hat{\Sigma} = \widehat{var}(\boldsymbol{b}) = \hat{\sigma}^2 \left(\sum_{t=1}^{T} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}$$

For the sales model we have:

$$\hat{\Sigma} = \begin{bmatrix} 40.3 & -6.80 & -0.75 \\ & 1.20 & -0.02 \\ & & -.47 \end{bmatrix}$$

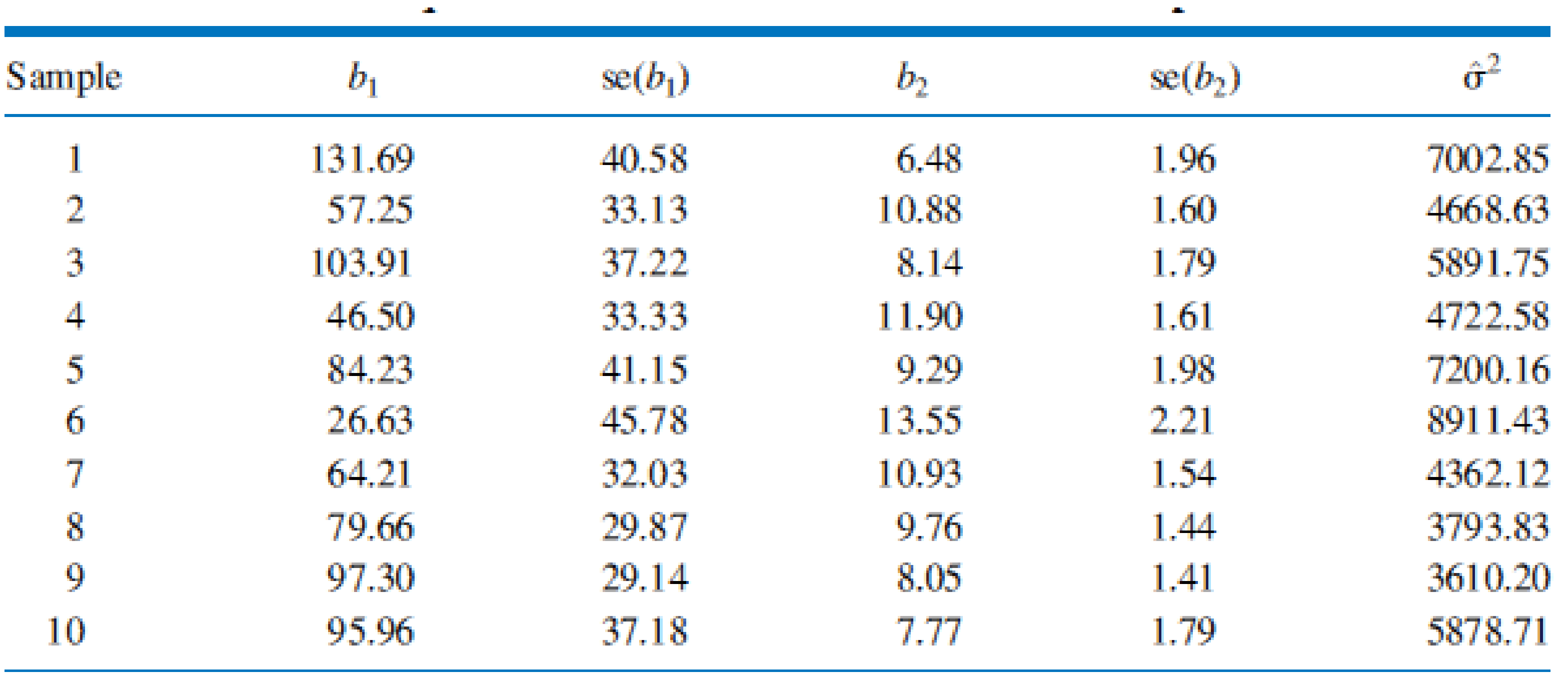


	C	*PRICE*	*ADVERT*
C	40.3433	−6.7951	−0.7484
PRICE	−6.7951	1.2012	−0.0197
ADVERT	−0.7484	−0.0197	0.4668

The standard errors are:

$$se(b_1) = \sqrt{40.3} = 6.35$$
$$se(b_2) = \sqrt{1.20} = 1.096$$
$$se(b_3) = \sqrt{0.47} = 0.68$$

Monte Carlo experiment:



Sample	b_1	$se(b_1)$	b_2	$se(b_2)$	$\hat{\sigma}^2$
1	131.69	40.58	6.48	1.96	7002.85
2	57.25	33.13	10.88	1.60	4668.63
3	103.91	37.22	8.14	1.79	5891.75
4	46.50	33.33	11.90	1.61	4722.58
5	84.23	41.15	9.29	1.98	7200.16
6	26.63	45.78	13.55	2.21	8911.43
7	64.21	32.03	10.93	1.54	4362.12
8	79.66	29.87	9.76	1.44	3793.83
9	97.30	29.14	8.05	1.41	3610.20
10	95.96	37.18	7.77	1.79	5878.71

IMPORTANT!!!

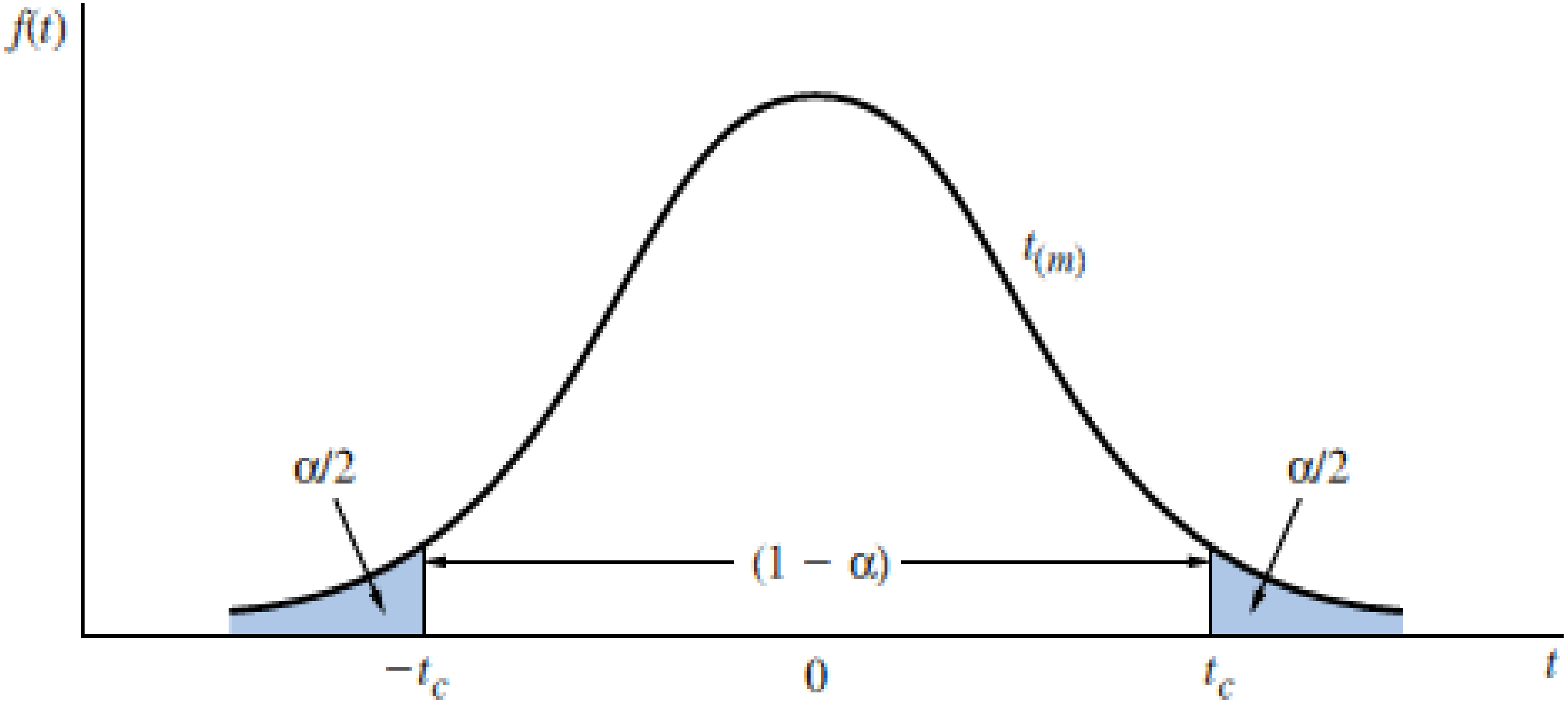
Replacing the variance of b_k with its estimate changes the distribution from normal to *t*-Student, so that:

$$\frac{b_k - \beta_k}{se(b_k)} \sim t_{(N-K)}$$

In general, if A1-A6 hold then:

$$t = \frac{b_k - \beta_k}{se(b_k)} \sim t_{(N-2)} \quad \text{for } k = 1,2$$

i



Interval estimation:

$$P\left(-t_{\alpha,N-K} \le \frac{b_k-\beta_k}{se(b_k)} \le t_{\alpha,N-K}\right) = 1-\alpha$$

$$P\big(b_k - t_{\alpha,N-K}se(b_k) \le \beta_k \le b_k - t_{\alpha,N-K}se(b_k)\big) = 1-\alpha$$

For *SALES* model we have [(*N*-*K*)=72]:

$$P\left[b_2 - 1.993\times \mathrm{se}(b_2) \ \le\ \beta_2 \ \le\ b_2 + 1.993\times \mathrm{se}(b_2)\right] = .95$$

$$\left(-7.9079-1.993\times 1.096, -7.9079+1.993\times 1.096\right)=\left(10.093,-5.723\right)$$

Interpretation: decreasing price by $1 will lead to an increase in revenue somewhere between $5,723 and $10,093.

Distribution for the linear combination of parameters

We may wish to obtain the distribution for a linear combination of parameters:

$$\lambda = c_1\beta_1 \ + c_2\beta_2$$

where c1 and c2 are constants that we specify

Then $\hat{\lambda} = c_1 b_1 \ + c_2 b_2$ we have:

$$E(\hat{\lambda}) = \lambda$$

$$\widehat{var}(\hat{\lambda}) = c_1^2\widehat{var}(b_1) + \mathrm{c}_2^2\widehat{var}(b_2) + 2\mathrm{c}_1\mathrm{c}_2\widehat{cov}(b_1, b_2)$$

$$se(\hat{\lambda}) = \sqrt{\widehat{var}}(\hat{\lambda})$$

$$t = (\hat{\lambda} - \lambda)/se(\hat{\lambda}) \sim t_{(N-2)}$$

Example:

Suppose we want to increase advertising by $800 and drop the price by 40 cents. The expected change in sales is:

$$\lambda = E(SALES_1) - E(SALES_2) = -0.4\beta_2 + 0.8\beta_3$$

The estimator is:

$$\hat{\lambda} = -0.4b_2 + 0.8b_3 = -0.4 \times (-7.91) + 0.8 \times 1.86 = 4.6532$$

$$se(\hat{\lambda}) = \sqrt{0.16 \times 1.2 + 0.64 \times 0.47 - 0.64 \times (-0.02)} = 0.7096$$

The 90% interval:

$$\left(4.6532 - 1.666 \times 0.7096, 4.6532 + 1.666 \times 0.7096\right) = \left(3.471, 5.835\right)$$

Indicates that the expected increase in sales will lie between $3,471 and $5,835 with 90% probability

Hypothesis Tests

- Hypothesis testing = comparison of a conjecture we have about a population to the information contained in a sample of data
- In econometric models hypotheses are represented as statements about model parameters
- Hypothesis tests use the information about a parameter from the sample: its LS estimate and standard error
- The procedurę consists of 4 steps:
  1. Setting H0 and H1
  2. Calculate a test statistic
  3. Calculate a rejection region
  4. A conclusion

- A null hypothesis is the belief we will maintain until we are convinced by the sample evidence that it is not true (the preasumption of innocence)
- The null hypothesis is stated as

$$H_0 : \beta_k = c$$

where c is a constant (usually 0)

- The alternative hypothesis depends to some extent on economic theory:

$$H_1 : \beta_k > c$$
$$H_1 : \beta_k < c$$
$$H_1 : \beta_k \neq c$$

- To choose between H0 and H1 we need a test statistic, for which the probability distribution is known when H0 is true (it has some other distribution if H1 is true)
- If A1-A5 holds then:

$$\frac{b_k - \beta_k}{se(b_k)} \sim t_{(N-K)}$$

- Hence, if $H_0 : \beta_k = c$ is true we can substitute and:

$$\frac{b_k - c}{se(b_k)} \sim t_{(N-K)}$$

- We can reject H0 or not - avoid saying that you ‘‘accept’’ the null - we only don’t have a proof to reject the null (which does not mean that is is true)

- The rejection region consists of values that have low probability of occurring when the null is true

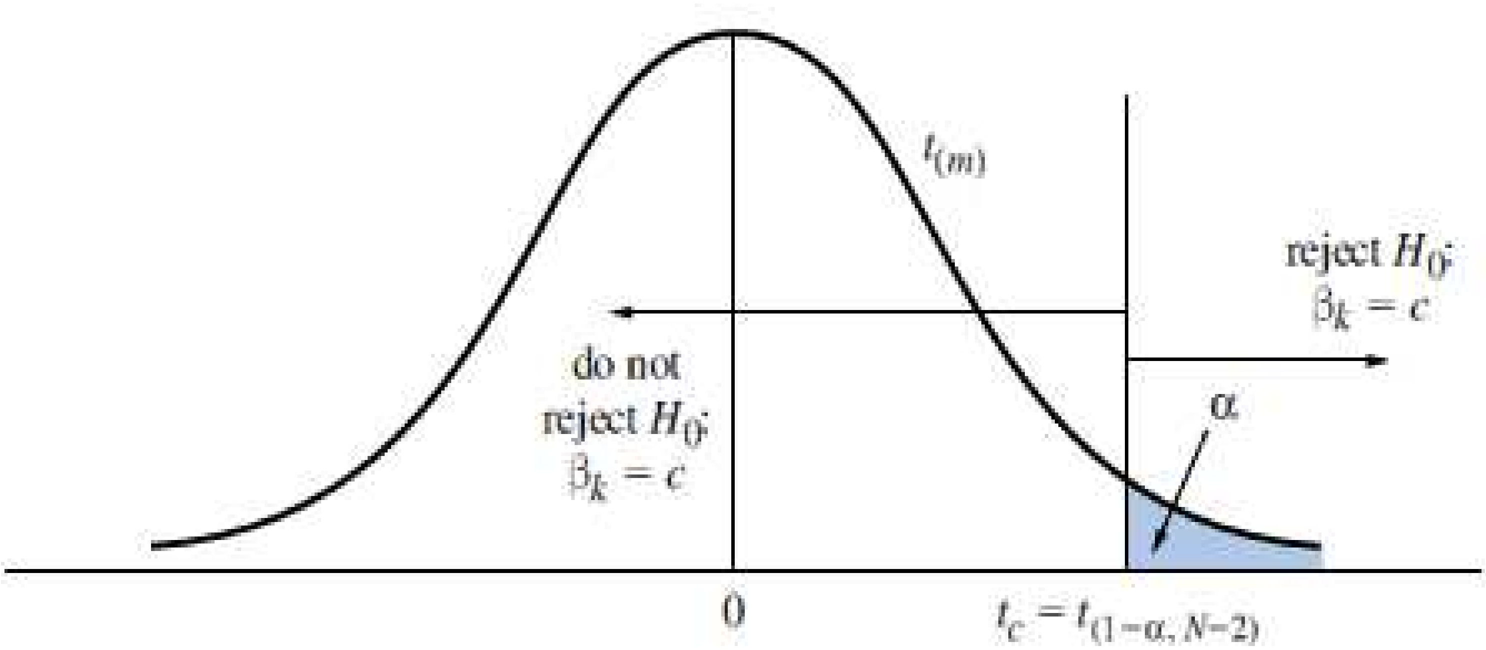
- The chain of logic is: *“If a value of the test statistic is obtained that falls in a region of low probability, then it is unlikely that the test statistic has the assumed distribution, and thus it is unlikely that the null hypothesis is true”*

- The probability α is called **the level of significance** and is interpreter as the probability of rejecting the null when it is true.

- Two types of error:
  - Type I error: we reject the null when it is true (with probability α)
  - Type II error: do not reject a null that is false

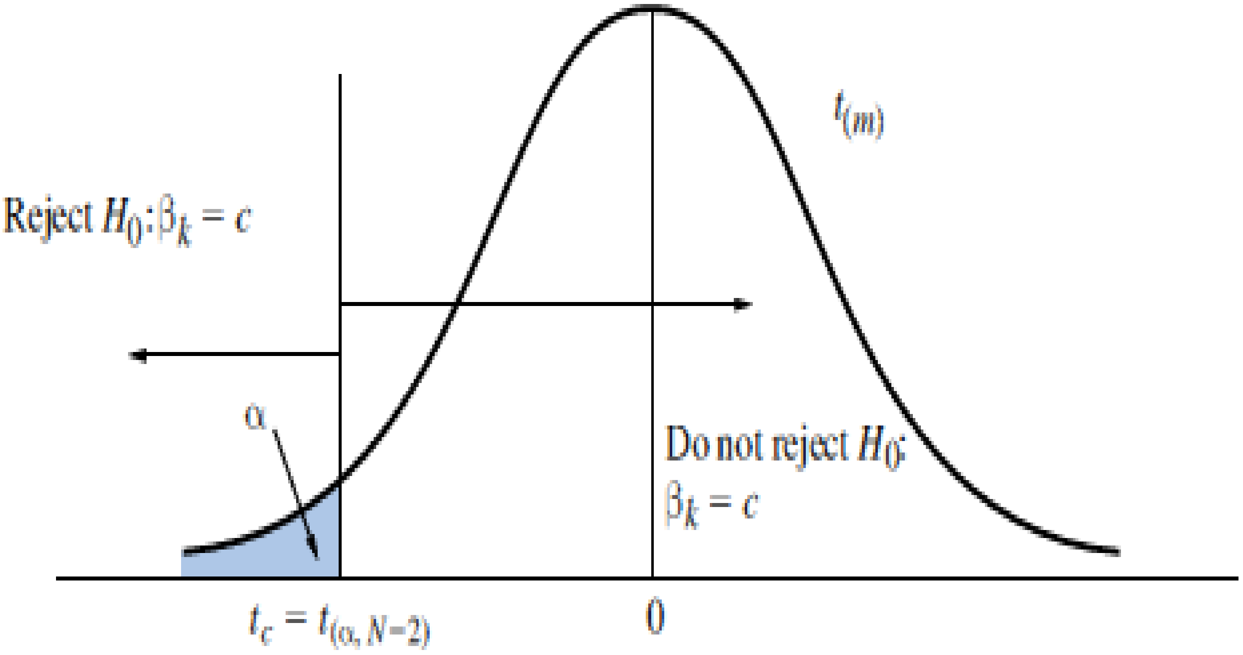
Inference for:

$$H0\colon \beta_k \;=\; c$$
$$H1\colon \beta_k \;>\; c$$


$t_{(m)}$
reject H_0: $\beta_k = c$
do not reject H_0: $\beta_k = c$
α
0
$t_c = t_{(1-\alpha, N-2)}$

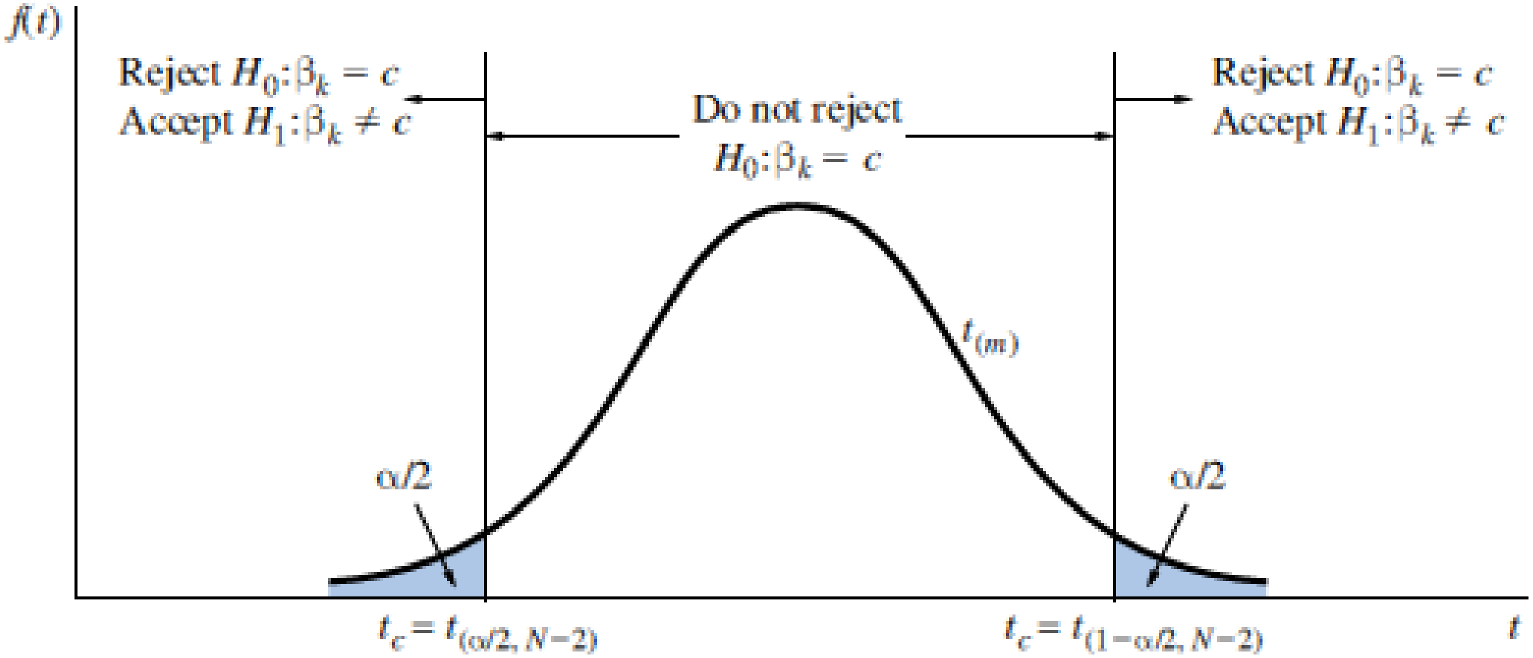
Inference for:

$$H0\colon \beta_k \;=\; c$$
$$H1\colon \beta_k < \; c$$

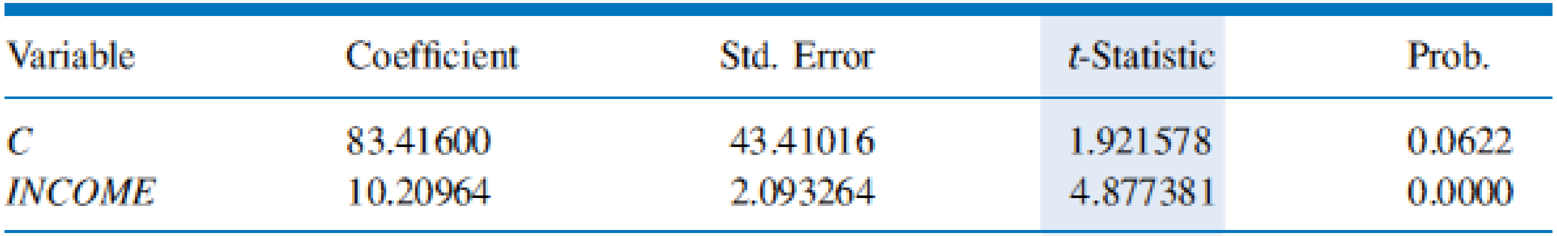


Inference for:

$$H0: \beta_k = \; c$$
$$H1: \; \beta_k \neq \; c$$


$f(t)$
Reject $H_0: \beta_k = c$
Accept $H_1: \beta_k \neq c$
Do not reject
$H_0: \beta_k = c$
Reject $H_0: \beta_k = c$
Accept $H_1: \beta_k \neq c$
$t_{(m)}$
$\alpha/2$
$\alpha/2$
$t_c = t_{(\alpha/2, N-2)}$
$t_c = t_{(1-\alpha/2, N-2)}$
t

Typical Eviews output



Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	83.41600	43.41016	1.921578	0.0622
INCOME	10.20964	2.093264	4.877381	0.0000

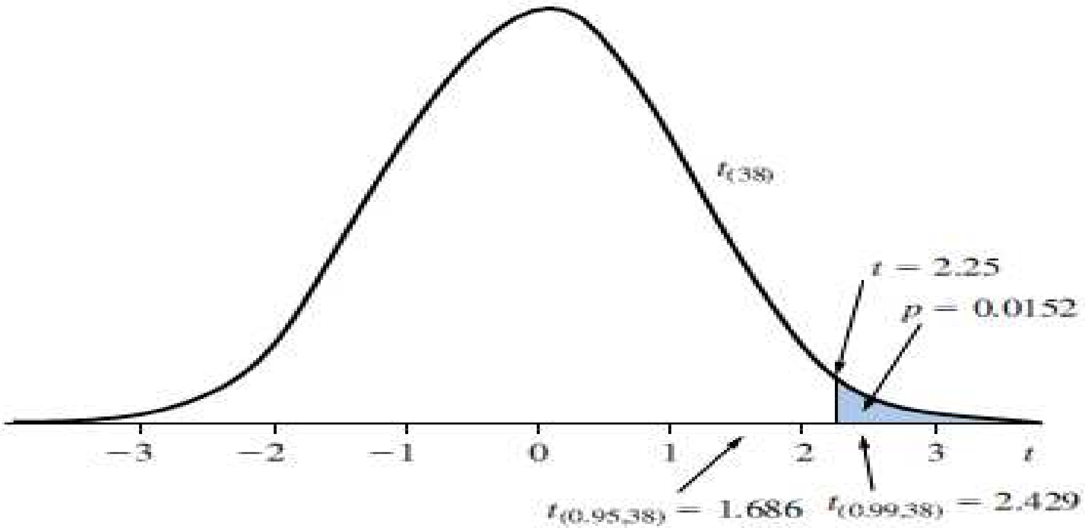
- Standard practice: report the ***p*-value** (an abbreviation for probability value) of the test.
- We compare the *p*-value to the significance level α

$$p \leq \alpha \text{ reject } H_0$$

$$p > \alpha \text{ do not reject } H_0$$

- For H0: $\beta_2 \leq 5.5$ against $H1$: $\beta_2 > 5.5$:

$$t\text{=2.25 and } P\left(t_{(38)} \geq 2.25\right) = 0.0152$$



The fit of the model

How to measure the fit of the model?

We can separate y_i into : $y_i = E(y_i) + e_i$

- $E(y_\mathrm{i})$ is the explainable or systematic part
- e_i is the random, unsystematic component

In terms of estimated model we have:

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

Or as deviations from the mean:

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

Use the fact that $\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$ to decompose the "total sample variation"

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

Specifically:

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

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

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

$$\text{SST} = \text{SSR} + \text{SSE}$$

The **coefficient of determination**, or R^2 , is defined as the proportion of variation in *y* explained by *x*:

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

Interpretation of *R*²: the proportion of the variation in *y* about its mean that is explained by the regression model

Example for the food expenditure model:

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

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

Hence:

$$R^2 = 1 - \frac{304505.176}{495132.160} = 0.385$$

Conclusion: 38.5% of the variation in food expenditure is explained by the regression model, which uses only income as an explanatory variable

Least Squares Prediction

Prediction = inference about *out-of-sample* observations

The ability to predict is important to:

- – business (e.g. forecasts of sales)
- – policy makers who (e.g. forecast of output, inflation)

Accurate predictions → better decisions

The LS predictor of y_0 comes from the fitted regression line (we assume that predition is for *t*=0):

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

Let us define the forecast error:

$$f_0 = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$$

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

The expected value of f_0 (unbiased forecast):

$$E(f_0) = \beta_1 + \beta_2 x_0 + E(e_0) \; - (E(b_1) + E(b_2) x_0) =$$
$$= \beta_1 + \beta_2 x_0 + 0 \; - (\beta_1 + \beta_2 x_0) = 0$$

The variance of the forecast is

$$var(f_0) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Two sources of forecast variance:

- random error
- estimation error

Multivariate case

True value:

$$y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + e_0$$

Prediction:

$$y_0^P = (\boldsymbol{x}_{\boldsymbol{0}}^P)'\boldsymbol{b}$$

Forecast error:

$$\begin{aligned} f_0 &= y_0 - y_0^P = \\ &= e_0 && \text{(stochastic error)} \\ &+(\boldsymbol{x}_{\boldsymbol{0}}^P)'(\boldsymbol{\beta} - \boldsymbol{b}) && \text{(estimation error)} \\ &+\left(\boldsymbol{x}_0 \ - \boldsymbol{x}_{\boldsymbol{0}}^P\right)'\boldsymbol{\beta} && \text{(exogenous vars. error)} \end{aligned}$$

The variance of theforecast:

$$\begin{aligned} var(y_0^P) &= var(e_0) + (\boldsymbol{x}_{\boldsymbol{0}}^P)'var(\boldsymbol{b})\ (\boldsymbol{x}_{\boldsymbol{0}}^P)[+\boldsymbol{\beta}'var(\boldsymbol{x}_{\boldsymbol{0}}^P)\boldsymbol{\beta}] \\ &= \sigma^2\left[1 + \boldsymbol{x}_{\boldsymbol{0}}'\ \widehat{\boldsymbol{\Sigma}}\ \boldsymbol{x}_{\boldsymbol{0}}\right] \end{aligned}$$

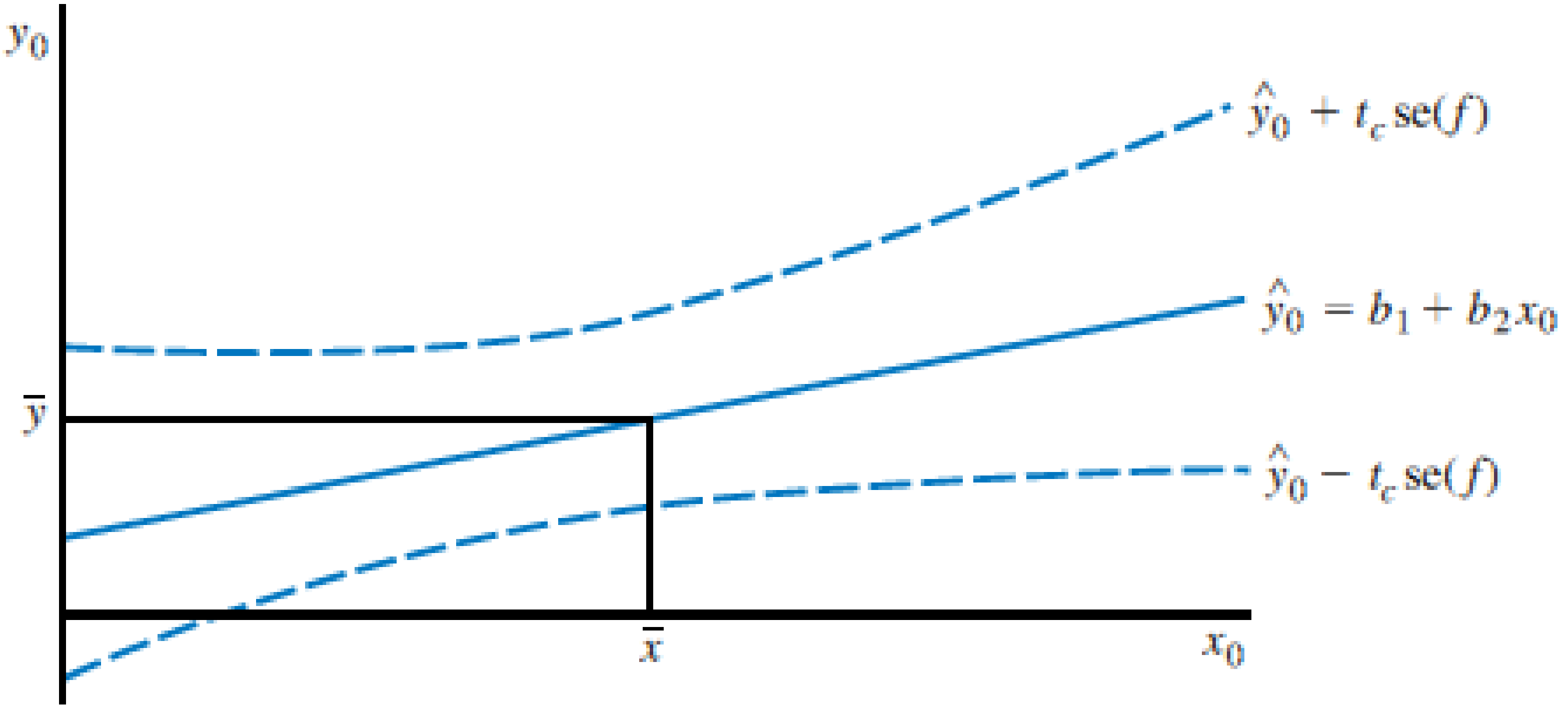
In practice we need to use:

$$se(f_0) = \sqrt{\widehat{var}(f_0)}$$

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

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

Important: prediction most accurate for $x_0 = \bar{x}$



Normal distribution of the error term

Hypothesis tests and interval estimates often rely on the assumption that the errors are normally distributed

We can check this using:

– a histogram

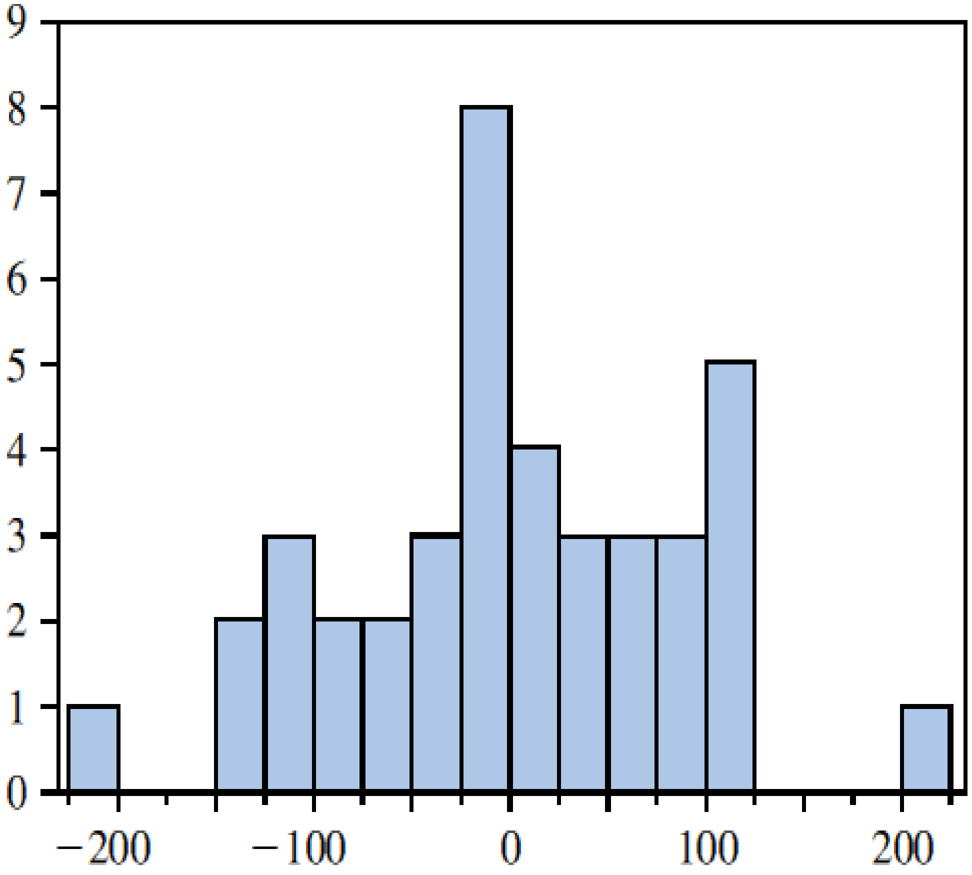
– formal statistical test, e.g. **Jarque–Bera test**

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

N - sample size, S – skewness, K – kurtosis

Under the null, $JB \sim \chi^2(2)$

Food expenditure example



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

Food expenditure example

The Jarque–Bera statistic is:

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

- – Because 0.063 < 5.99 (critical value for 5% significance level) there is insufficient evidence from the residuals to conclude that the normal distribution assumption is unreasonable
- – The same conclusion on the basis of p-value, as 0.9688 > 0.05

Joint Hypothesis Testing

A null hypothesis with multiple conjectures is called a **joint hypothesis.** For example, for the model

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$$

a possible joint hypothesis could be:

$$H_0 : \beta_3 = 0, \beta_4 = 0$$

$$H_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both are nonzero}$$

Unrestricted model: the restrictions in the null have not been imposed on the model

Restricted model: assumes the parameter restrictions in H_0 are true, i.e.:

$$SALES = \beta_1 + \beta_2 PRICE + e$$

F-test for the joint hypothesis: a comparison of the sums of squared errors from the unrestricted model SSE_U and the restricted one SSE_R (*J*-the number of restrictions)

$$F = \frac{\left(SSE_R - SSE_U\right)\big/ J}{SSE_U \big/ \left(N - K\right)}$$

If the null hypothesis is true, then the statistic *F* has the *F*-distribution with *J* numerator degrees of freedom and *N* - *K* denominator degrees of freedom

$$F \sim F(J, N - K)$$

Example, continuation:

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(1896.391-1532.084\right)/2}{1532.084/\left(75-4\right)} = 8.44$$

Since $F = 8.44 > F_{c,2,71} = 3.126$ we reject the null

The p-value is $p = P(F_{(2,\,71)} > 8.44) = 0.0005$

Conclusion: advertising does have a significant effect upon sales revenue

Overall significance test of the regression model

For the model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + e$$

we examine:

$$H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$$

$$H_1 : \textit{At least one} \text{ of the } \beta_k \text{ is nonzero for } k = 2, 3, \ldots K$$

The restricted model is:

$$y_i = \beta_1 + e_i$$

Comparison of F and LM tests

The *F*-statistic of the Wald test:

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)} \sim F(J, N-K)$$

Lagrange Multiplier test:

$$LM = \frac{SSE_R - SSE_U}{\hat{\sigma}^2} \sim \chi^2(J)$$

Given the LS estimator $\hat{\sigma}^2 = \frac{SSE_U}{N-K}$:

$$F = \frac{LM}{J}$$

When testing

$$H_0 : \beta_3 = \beta_4 = 0$$

in the equation

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e_i$$

we get

$$F = 8.44 \qquad p\text{-value} = .0005$$

$$\chi^2 = 16.88 \qquad p\text{-value} = .0002$$

Model Specification

Model specification: the most important issue in any econometric investigation

Model specification = the set of explanatory variables + functional form

A model could be misspecified if:

- – we have omitted important variables
- – included irrelevant ones
- – chosen a wrong functional form
- – have a model that violates the LS assumptions

Steps of choosing a specification of a model

1. Choose variables and a functional form on the basis of your theoretical considerations (economic theory)
2. If an estimated equation has coefficients with unexpected signs or unrealistic values – a sign of model misspecification (e.g. omitted variables)
3. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests
4. Consider various model selection criteria
5. The adequacy of a model can be tested using a general specification test known as RESET

Ommited variable bias

Let the true model be

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

But we estimate

$$y = \beta_1 + \beta_2 x_2 + e$$

Omitting x_3 is equivalent to imposing incorrect restriction $\beta_3 = 0$. This leads to the endogeneity bias (subject of future meeting):

$$bias(b_2) = \beta_3 \frac{cov(x_2, x_3)}{var(x_2)}$$

Inflated variance due to irrelevant variables

A strategy to avoid omitted variables bias - to include as many variables as possible in your model

However, this might complicate the model unnecessarily and inflate the variances of the estimator due to the presence of **irrelevant variables**

As a result – this is not a good strategy…

Model selection criteria

The common feature of information criteria of model selection:

- the best fit to the data (minimum *SSE*)
- the most parsimonious specification (minimum *K*)

Akaike information criterion (*AIC*):

$$AIC = \ln\left(\frac{SSE}{N}\right) + \frac{2K}{N}$$

Schwarz information criterion (*SC*) = Bayesian information criterion (*BIC*) :

$$SC = \ln\left(\frac{SSE}{N}\right) + \frac{K\ln(N)}{N}$$

RESET (REgression Specification Error Test)

RESET test - designed to incorrect functional form

Let $\hat{y}$ be the predicted values of

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Consider the artificial model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_2 \hat{y}^3 + e$$

A test for misspecification

$$H_0{:}\gamma_1 = \gamma_2 = 0 \text{ against } H_1{:}\ \gamma_1 \neq 0 \text{ or } \gamma_2 \neq 0$$

Collinearity

Exact collinearity: there is a linear relationship among the explanatory variables. In this case the LS estimator is not defined and we cannot obtain estimates of β

Close colinearity: high correlation ammong explanatory variables → imprecise LS estimates

How to detect the problem? If R^2 of auxilary regression

$$x_2 = a_1 x_1 + a_3 x_3 + \cdots + a_K x_K + error$$

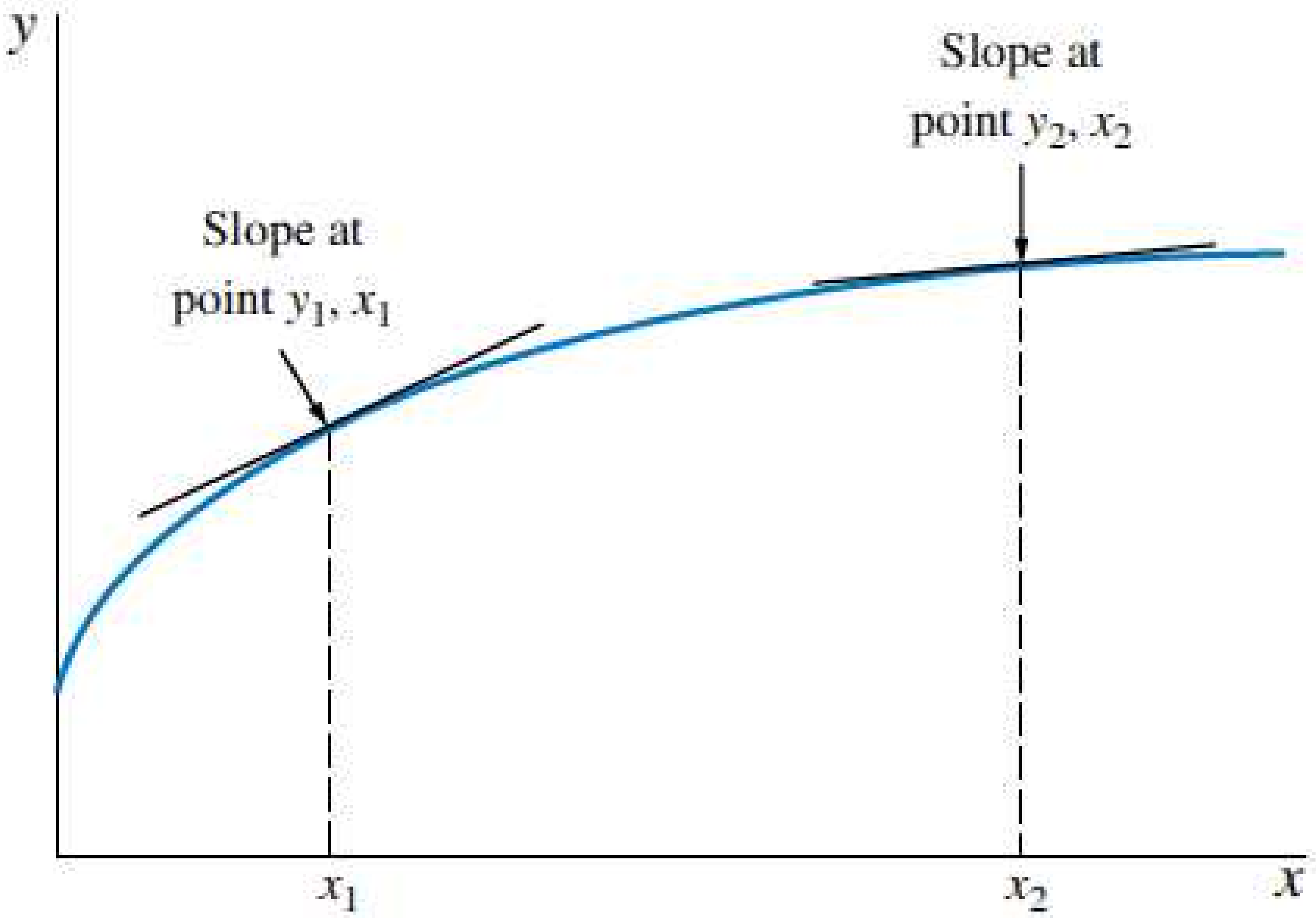
is above 80%

What to do: add nonsample information in the form of restrictions on the parameters

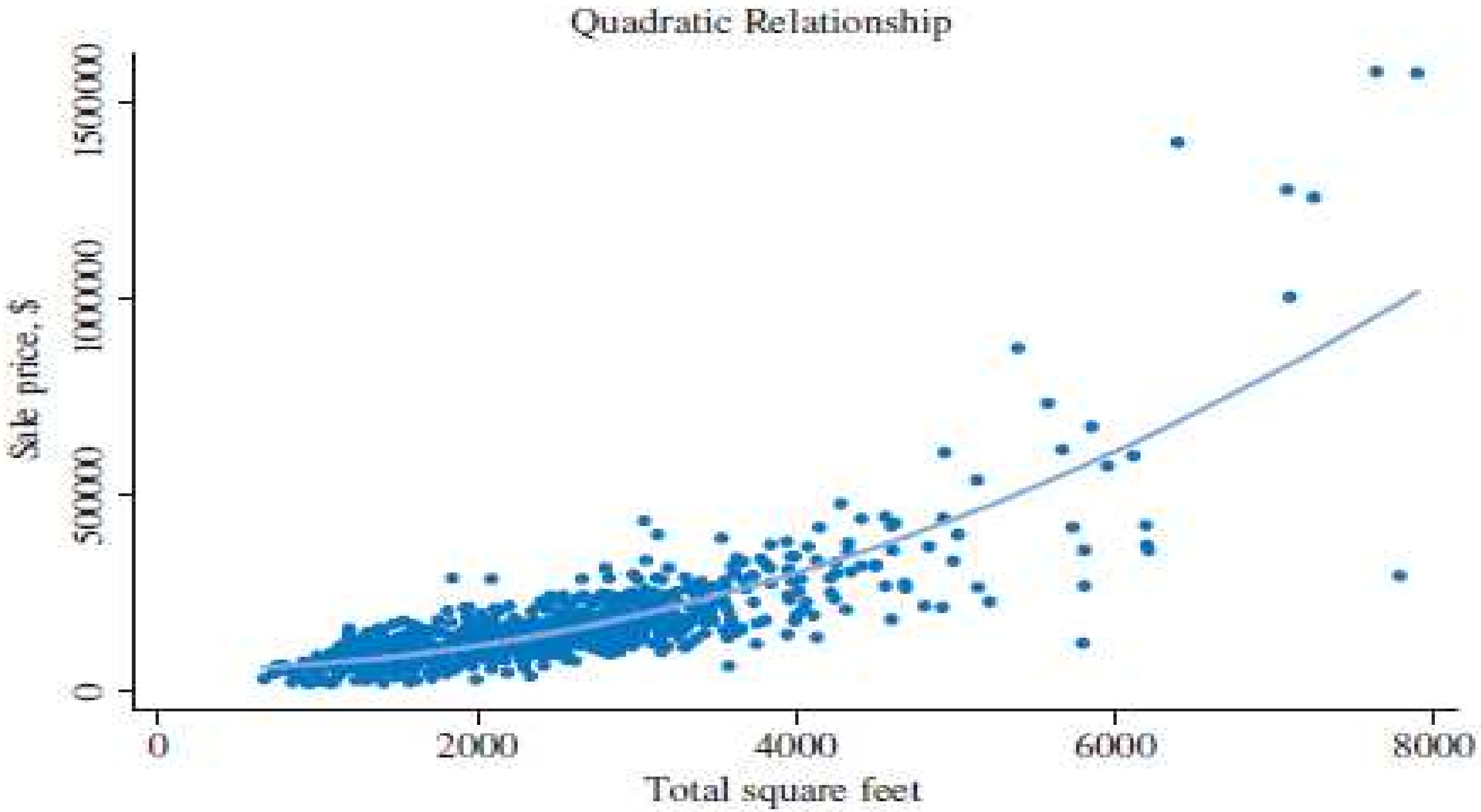
Nonlinear Relationships

A number of issues we must address when building an econometric model (in which *y* depends on *x*):

- Scaling the data
- What does economics say about the relation between *y* and *x*? Is it increasing? Is it linear?
- The **marginal effect =** the slope of the tangent to the curve at a particular point. Does it depend on *x* or *y*?

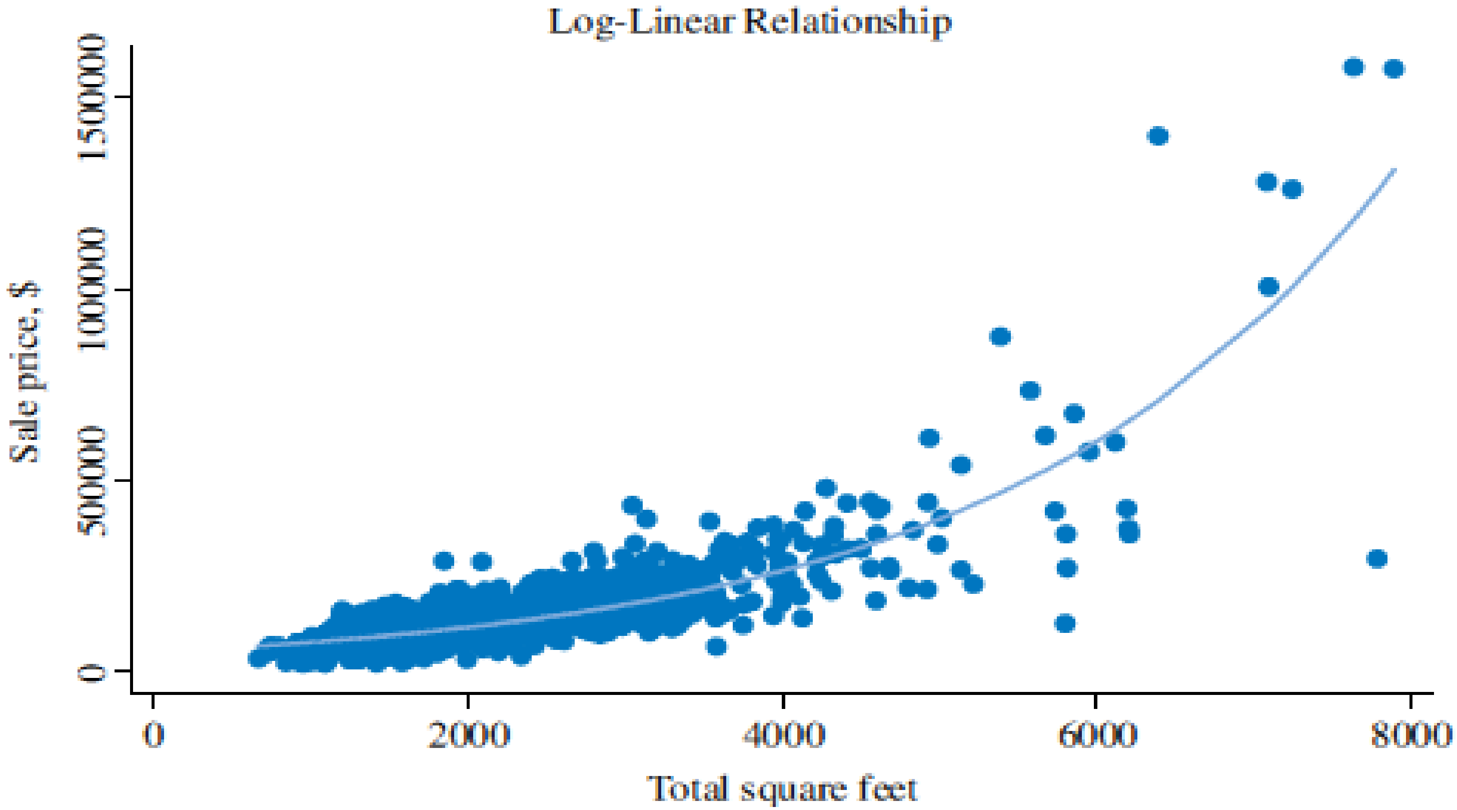

y
Slope at
point y_2, x_2
Slope at
point y_1, x_1
x_1
x_2
x

For a **quadratic model** $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$ the slope is: $\frac{dPRICE}{dSQFT} = 2\alpha_2 SQFT$



For **log-linear model** $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$

the slope is: $\frac{dPRICE}{dSQFT} = \gamma_2 PRICE$



Econometric models often employ natural logarithms, because $\Delta\, ln$ = % change

For example, for the **log-linear model**, ln(y) = $\beta_1 + \beta_2 x$:

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

What is the interpretation of 0.09 in a model of wage vs. years of education?

$$\ln(\widehat{WAGE}) = 1.60 + 0.09 \times EDUC$$

For the **linear-log model**:

$$\mathrm{y} = \beta_1 + \beta_2 \ln x + \epsilon$$

the slope is:

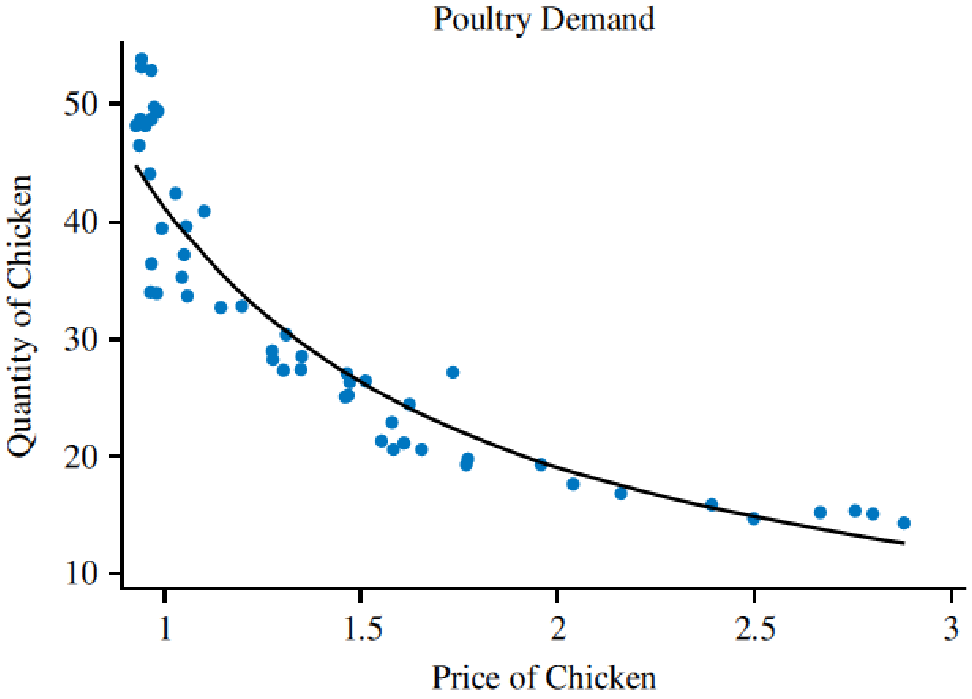
$$\beta_2 = \frac{\Delta y}{\Delta x / x}$$

- The term 100(Δ*x*/*x*) is the percentage change in *x*
- Thus, in the linear-log model we can say that a 1% increase in *x* leads to a β_2 /100 change in *y*

For the **log-log model**

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

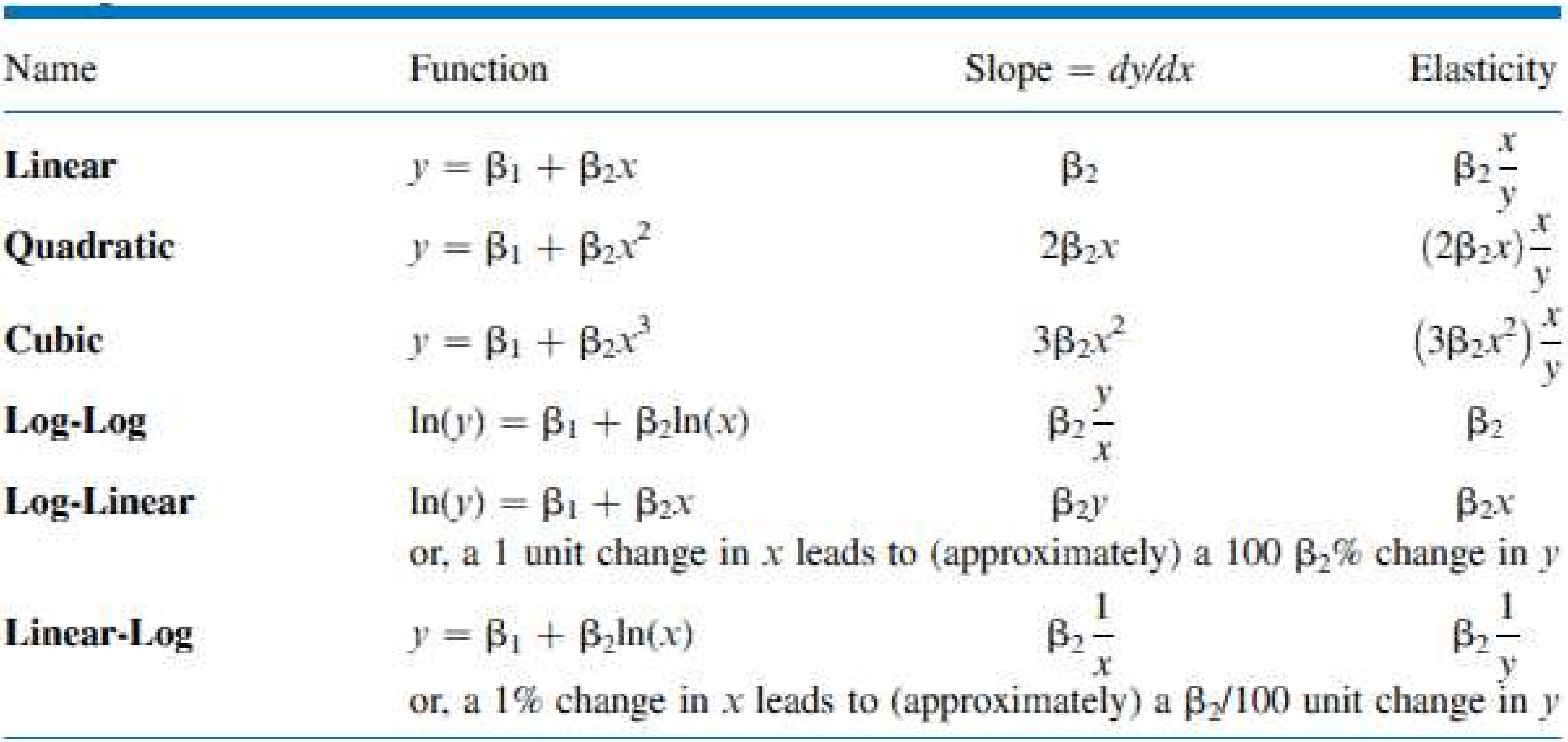
β_2 is interpreter as elasticity



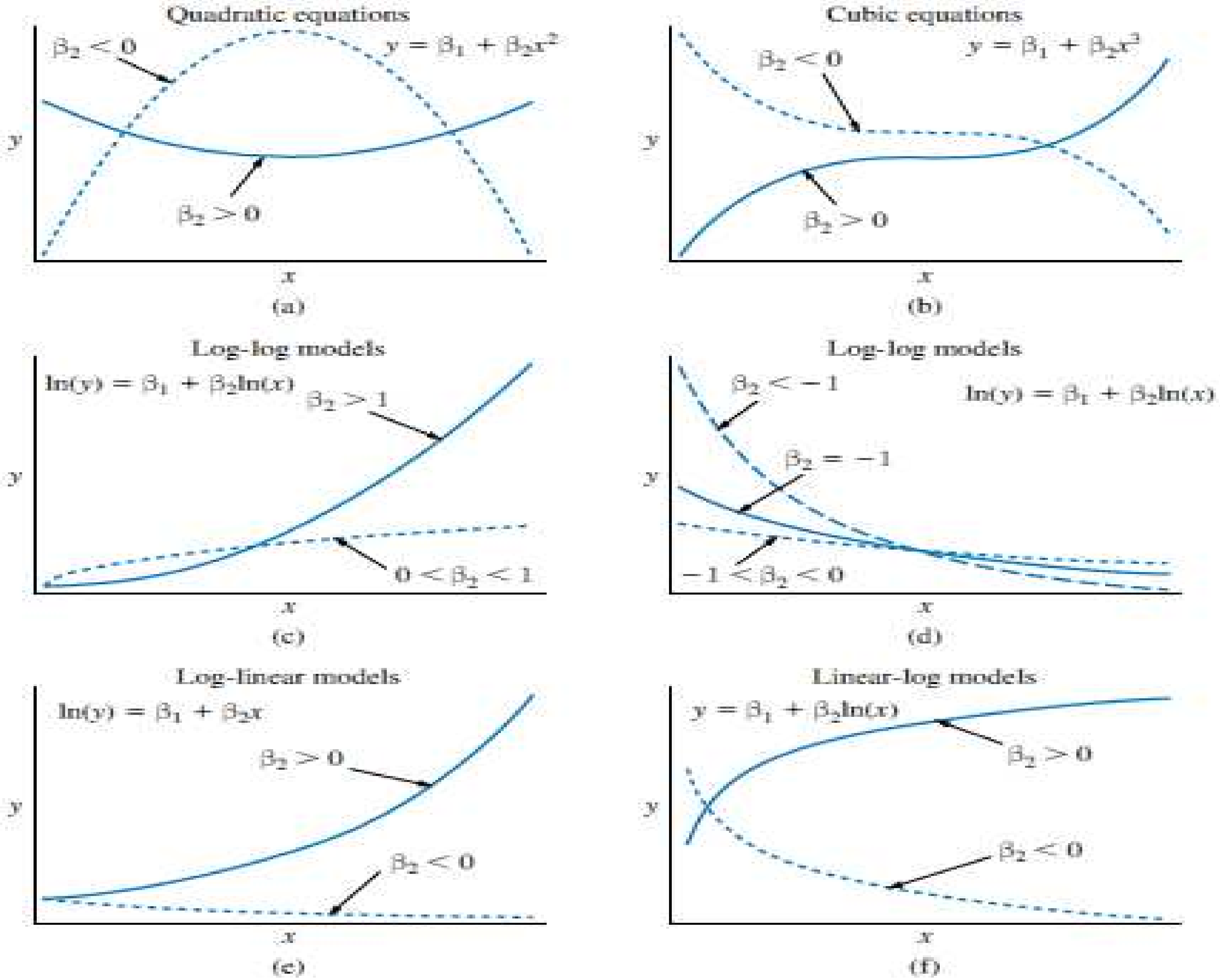
The estimated model is:

$$\widehat{\ln(Q)} = 3.72 - 1.21\ln(P)$$

- The price elasticity of demand is 1.121: a 1% increase in real price is estimated to reduce quantity consumed by 1.121%

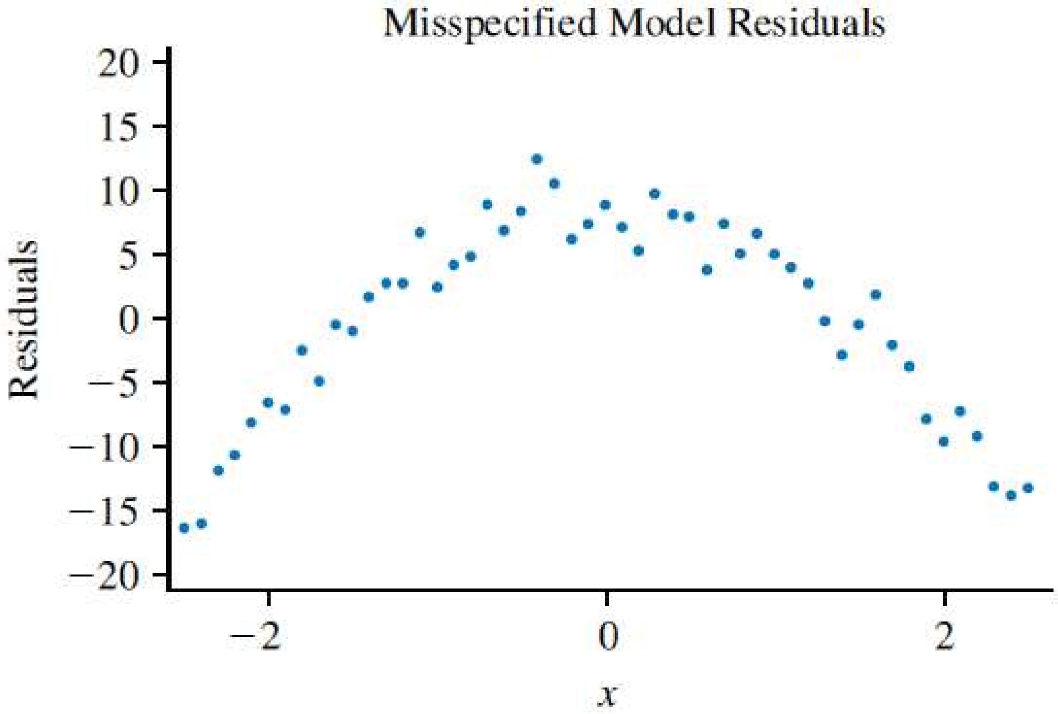


Name	Function	Slope $= dy/dx$	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-Log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2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


Quadratic equations
$\beta_2 < 0$
$y = \beta_1 + \beta_2 x^2$
y
$\beta_2 > 0$
x
(a)
Cubic equations
$\beta_2 < 0$
$y = \beta_1 + \beta_2 x^3$
y
$\beta_2 > 0$
x
(b)
Log-log models
$\ln(y) = \beta_1 + \beta_2 \ln(x)$
$\beta_2 > 1$
y
$0 < \beta_2 < 1$
x
(c)
Log-log models
$\beta_2 < -1$
$\ln(y) = \beta_1 + \beta_2 \ln(x)$
$\beta_2 = -1$
y
$-1 < \beta_2 < 0$
x
(d)
Log-linear models
$\ln(y) = \beta_1 + \beta_2 x$
$\beta_2 > 0$
y
$\beta_2 < 0$
x
(e)
Linear-log models
$y = \beta_1 + \beta_2 \ln(x)$
$\beta_2 > 0$
y
$\beta_2 < 0$
x
(f)

How to check whether the functional form of a mmodel is well specified?

- Formal tests (e.g. RESET)
- Graph of residuals



Logarithms

Suppose that the variable *y* has a normal distribution, with mean μ and variance σ^2

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

- It can shown that:

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

Hence, for a log-linear model ln(y) = β_1+ $\beta_2 x$ + e with $e \sim N(0, \sigma^2)$:

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

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

$$= e^{\beta_1+\beta_2 x_i} e^{\sigma^2/2}$$

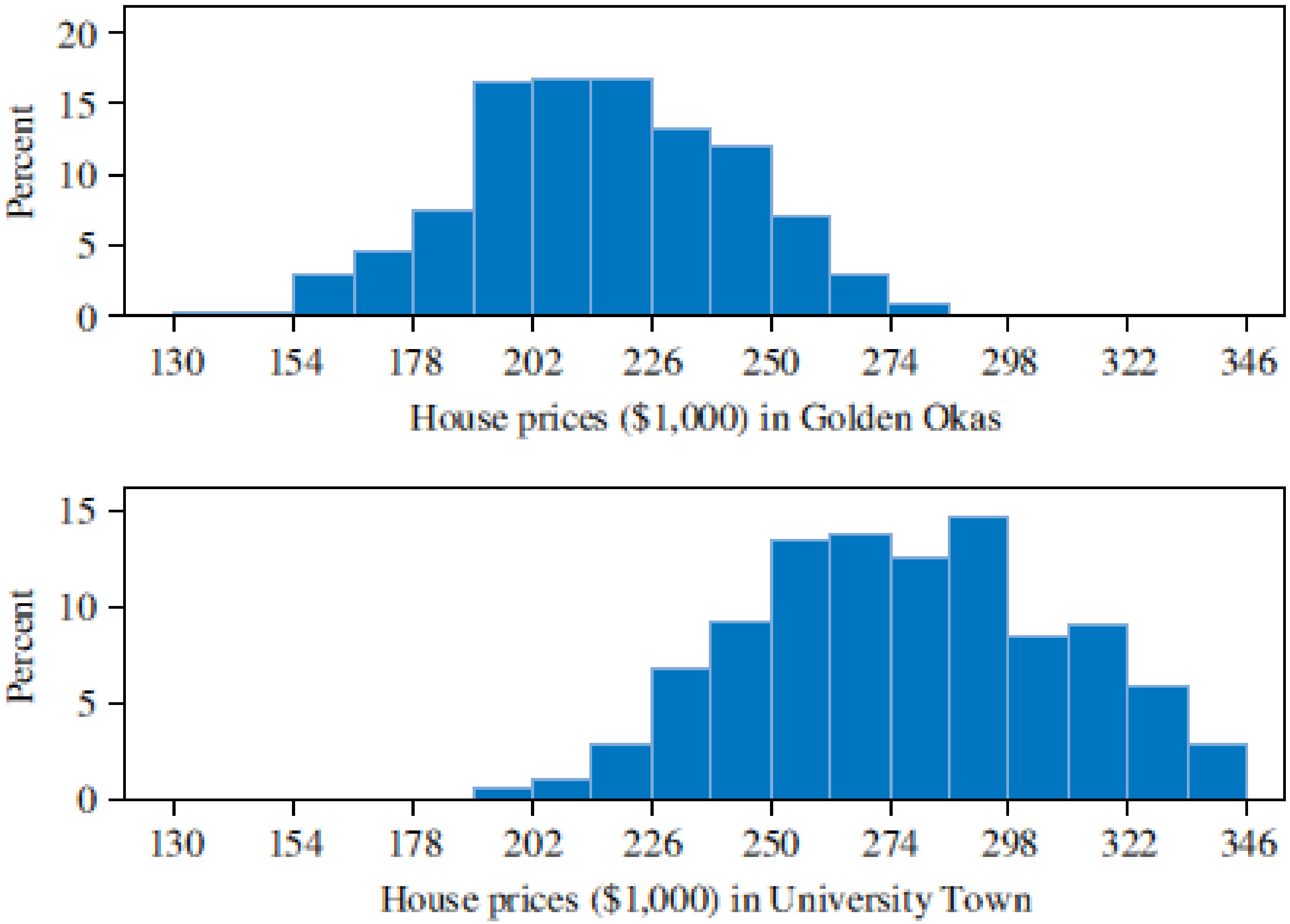
$$= e^{\beta_1+\beta_2 x_i+\sigma^2/2}$$

Regression with Indicator or Interaction Variables

An indicator variable is a binary variable that takes the values zero or one. It is used to represent a qualitative (nonquantitative) characteristic, such as gender, race, or location

$$UTOWN = \begin{cases} 1 & \text{house is in University Town} \\ 0 & \text{house is in Golden Oaks} \end{cases}$$

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


Percent
20
15
10
5
0
130
154
178
202
226
250
274
298
322
346
House prices ($1,000) in Golden Okas
Percent
15
10
5
0
130
154
178
202
226
250
274
298
322
346
House prices ($1,000) in University Town

Theoretical values in a model with the indicator variable:

$$E(PRICE) = \begin{cases} \beta_1 + \beta_2 & \text{if } UTOWN = 1 \\ \beta_1 & \text{if } UTOWN = 0 \end{cases}$$

$$\begin{aligned} \widehat{PRICE} &= b_1 + b_2 UTOWN \\ &= 215.7325 + 61.5091 UTOWN \\ &= \begin{cases} 277.2416 & \text{if } UTOWN = 1 \\ 215.7325 & \text{if } UTOWN = 0 \end{cases} \end{aligned}$$

Consider a model in which ln(*WAGE*) depends on years of education (*EDUC*) and years of experience (*EXPER*):

$$\ln\left(WAGE\right) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + e$$

If we believe the effect of an extra year of experience on wages will depend on the level of education. This can be done by including an **interaction variable:**

$$\ln\left(WAGE\right) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 \left(EDUC \times EXPER\right) + e$$

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 (EDUC \times EXPER) + e$$

The effect of another year of experience, holding education constant, is:

$$\left.\frac{\Delta \ln(WAGE)}{\Delta EXPER}\right|_{EDUC \text{ fixed}} = \beta_3 + \beta_4 EDUC$$

The approximate percentage change in wage given a one-year increase in experience is

$$100(\beta_3 + \beta_4 EDUC)\%$$