An Economic Model

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK
Gabelli School of Business

An Economic Model, Part I

- Economic theory suggests many relationships between economic variables.
- A regression model is helpful in questions such as the following: If one variable changes in a certain way, by how much will another variable change?
- The regression model is based on assumptions.
- The continuous random variable y has a probability density function (*pdf*).
- The *pdf* is a conditional probability density function because it is ‘‘conditional’’ upon an *x*.

An Economic Model, Part II

- Households with an income of $1,000 per week would have various food expenditure per person for a variety of reasons.
- The *pdf* *f*(y) describes how expenditures are distributed over the population.
- This is a conditional *pdf* because it is “conditional” upon household income.
- The *conditional mean*, or *expected value*, of y is $E(y|x = \$1000) = \mu_{y|x}$ and is our population’s mean weekly food expenditure per person.
- The conditional variance of y is $\text{var}(y|x = \$1000) = \sigma^2$, which measures the dispersion of household expenditures y about their mean.

An Economic Model, Part III

- The parameters $\mu_{y|x}$ and σ^2, if they were known, would give us some valuable information about the population we are considering.
- In order to investigate the relationship between expenditure and income, we must build an economic model and then a corresponding econometric model that forms the basis for a quantitative or empirical economic analysis.
- This econometric model is also called a regression model.

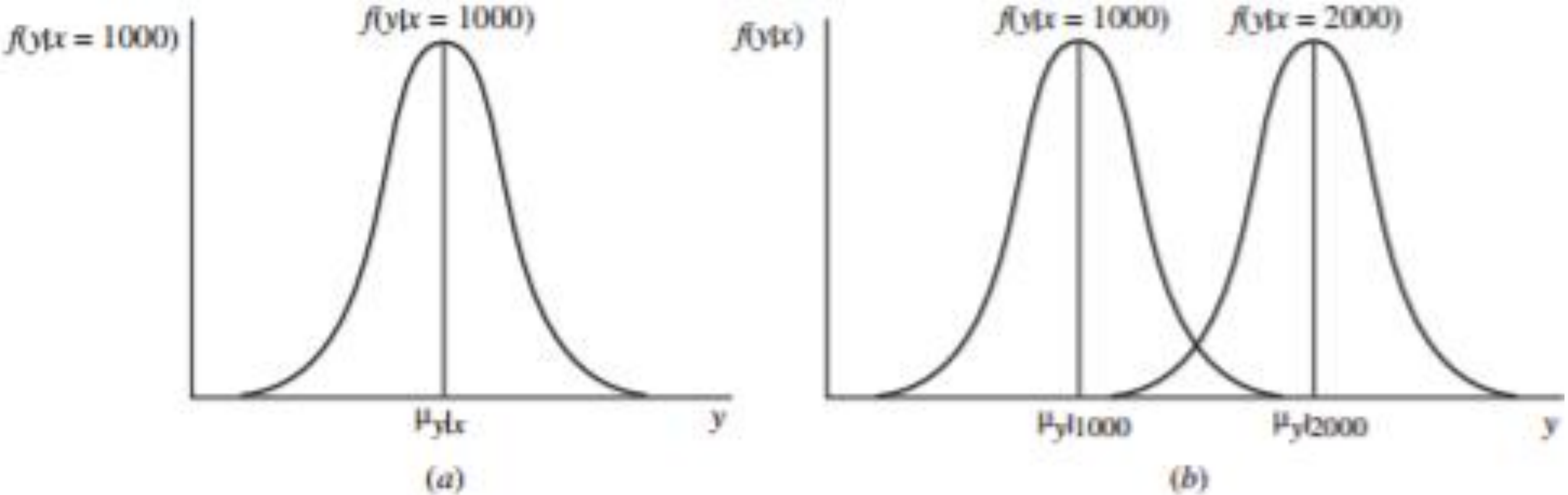

FIGURE 2.1 (*a*) **Probability distribution $f(y|x = 1000)$ of food expenditure y given income x = \$1000. (*b*) Probability distributions of food expenditure y given incomes x = \$1000 and x = \$2000.**

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

An Econometric Model

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

An Econometric Model

- A household spends $80 plus 10 cents of each dollar of income received on food.
- Algebraically, their rule is y = 80 + 0.10x, where y is weekly household food expenditure ($) and x is weekly household income ($).
- Many factors may affect household expenditure on food.
- Let *e* (error term) be everything else affecting food other than income.
- $y = \beta_1 + \beta_2$x + *e*: This equation is the simple regression model.
- A simple linear regression analysis examines the relationship between a y-variable and one x-variable.

Data Generating Process

- For the household food expenditure example, let us assume that we can obtain a sample at a point in time (cross-sectional data).
- The sample consisting of N data pairs is randomly selected from the population. Let (y_i, x_i) denote the *i*th pair.
- The variables $y_i \text{ and } x_i$ are random variables because their values are not known until they are observed. Each observation pair is statistically different from other pairs.
- All pairs drawn from the same population are assumed to follow the same joint pdf and are identically distributed (i.i.d.).

The Random Error and Strict Exogeneity

- The second assumption of the simple regression model concerns the "everything else" term e
- Unlike food expenditure and income, the random error term is not observable; it is unobservable
- The x-variable, income, cannot be used to predict the value of
- $E(e_i|x_i) = 0$ has two implications
    1. $E(e_i|x_i) = 0 \Rightarrow E(e_i) = 0$
    2. $E(e_i|x_i) = 0 \Rightarrow \text{cov}(e_i|x_i) = 0$

The Regression Function

- The conditional expectation $E(y_i|x_i) = \beta_1 + \beta_2 x_i$ is called the regression function.
- This says the population, the average value of the dependent variable for the *i*th observation, conditional on x_i is given by $\beta_1 + \beta_2 x_i$.
- This also says, given a change in x, Δx, the resulting change in $E(y_i|x_i)$ is $\beta_2 \Delta x$ holding all else constant.
- We can say that a change in x leads to, or causes, a change in the expected (population average) value of y given x_i, $E(y_i|x_i)$.

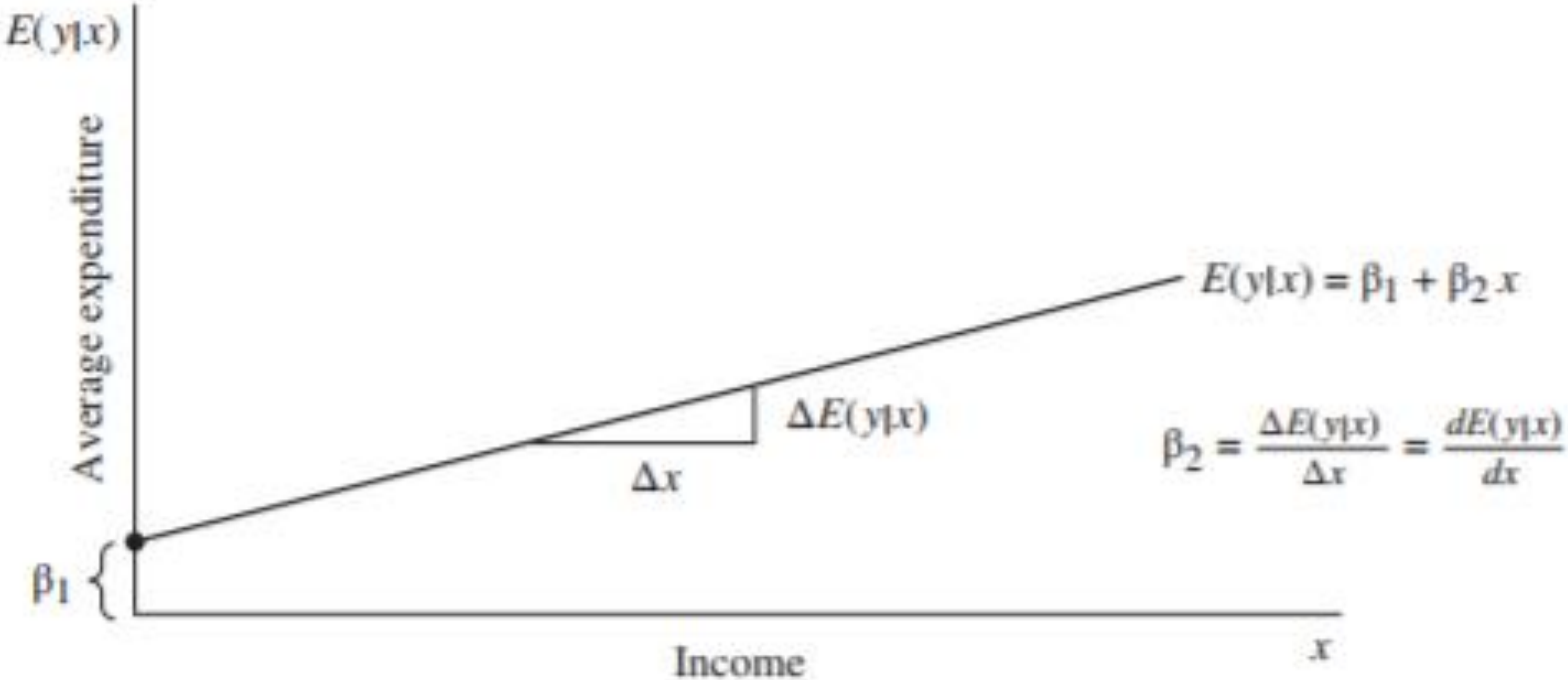

FIGURE 2.2 **The economic model: a linear relationship between average per person food expenditure and income.**

Random Error Variation

- Ideally, the conditional variance of the random error is constant
- $var(e_i|x_i) = \sigma^2$: this is the homoskedasticity assumption
- Assuming the population relationship $y_i = \beta_1 + \beta_2 x_i + e_i$ the conditional variance of the dependent variable is

$$var(e_i|x_i) = var(\beta_1 + \beta_2 x_i + e_i) = var(e_i|x_i) = \sigma^2$$

- If this assumption is violated, and $var(e_i|x_i) \neq \sigma^2$, then the random errors are said to be heteroskedastic

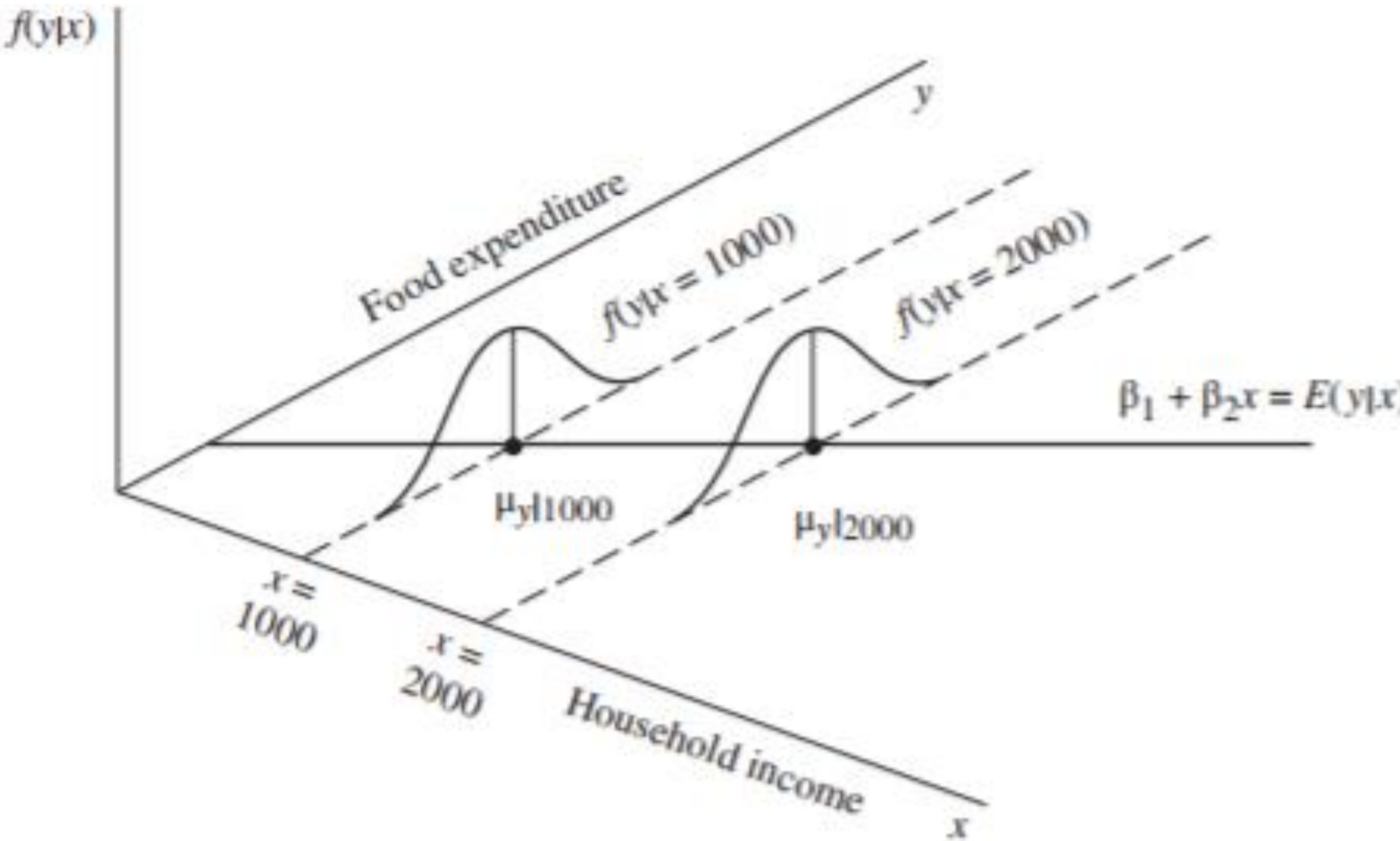

FIGURE 2.5 **The conditional probability density functions for *y*, food expenditure, at two levels of income.**

Variation in x

- In a regression analysis, one of the objectives is to estimate $\beta_2 = \Delta E(y_i|x_i)/\Delta x_i$.
- If we are to hope that a sample of data can be used to estimate the effects of changes in x, then we must observe some different values of the explanatory variable x in the sample.
- The minimum number of x-values in a sample of data that will allow us to proceed is two.

Error Normality

- It is not at all necessary for the random errors to be conditionally normal for regression analysis to “work.”
- When samples are small, it is advantageous for statistical inferences that the random errors, and dependent variable y, given each x-value, are normally distributed.
- Central limit theorem says roughly that collections of many random factors tend toward having a normal distribution.
- It is entirely plausible that the random are normally distributed.

Generalizing the Exogeneity Assumption

- A lack of independence occurs naturally when using financial or macroeconomic time-series data.
- The data series is likely to be correlated across time.
- The assumption that the pairs (y_t, x_t) represent random i.i.d. draws from a probability distribution, is not realistic.
- We cannot predict the random error at time t, e_t, using any of the values of the explanatory variable.

Error Correlation

- It is possible that there are correlations between the random error terms.
- With cross-sectional data, data collected at one point in time, there may be a lack of statistical independence between random errors for individuals who are spatially connected.
- Within a larger sample of data, there may be clusters of observations with correlated errors because of the spatial component.
- The starting point in regression analysis is to assume that there is no error correlation.

Summarizing the Assumptions

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.

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

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

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

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.

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

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Estimating the Regression Parameters

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Estimating the Regression Parameters

- We can use the sample information in Table 2.1, specific values of y_i and x_i, to estimate the unknown regression parameters β_1 and β_2.
- These parameters represent the unknown intercept and slope coefficients for the food expenditure–income relationship.

TABLE 2.1 **Food Expenditure and Income Data**

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

Estimating the Regression Parameters (cont.)

- If we represent the 40 data points as $(y_i, x_i), i = 1, \ldots, N = 40$, and plot them, we obtain the scatter diagram in Figure 2.6.
- Our problem is to estimate the location of the mean expenditure line.
- We would expect this line to be somewhere in the middle of all the data points.

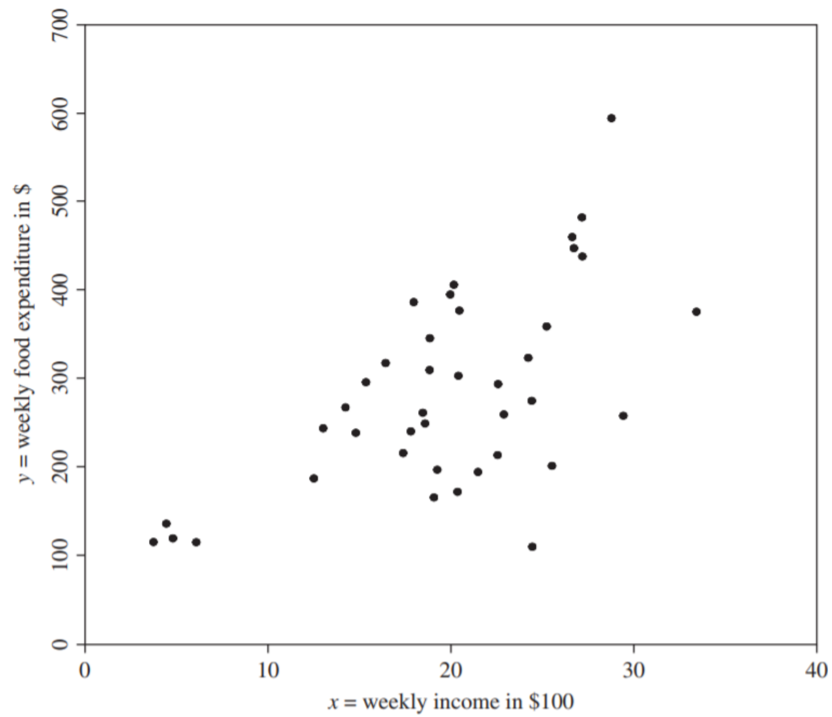

FIGURE 2.6 Data for the food expenditure example.

The Least Squares Principle

- The fitted regression line is:
  - (2.5) $\hat{y}_i = b_1 + b_2 x_i$
- The least squares residual is:
  - (2.6) $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$

Graphs of the Fitted Regression Line and Least Squares Residual Equations

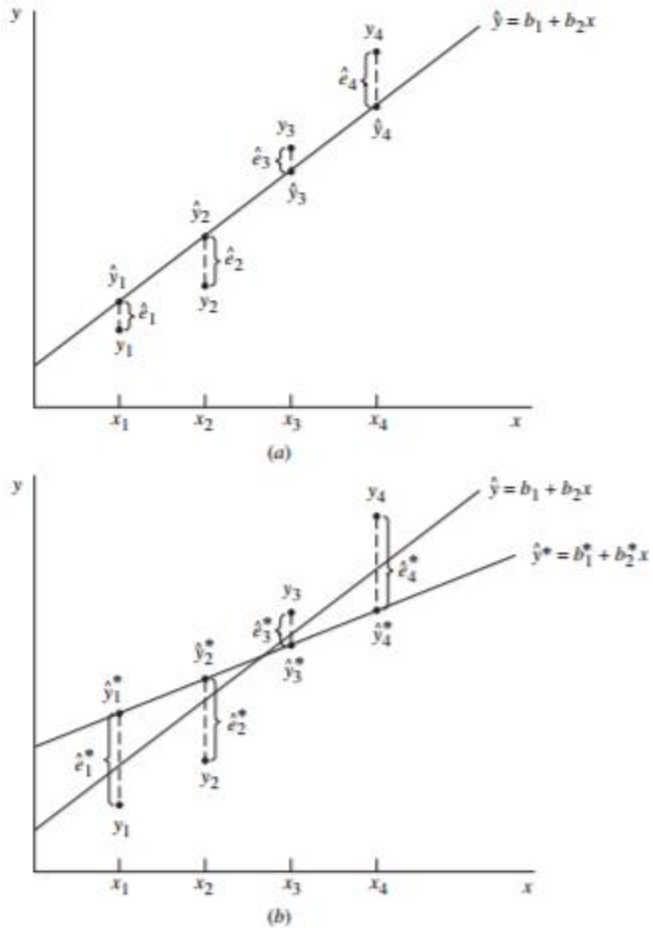

FIGURE 2.7 *(a)* The relationship among y, $\hat{e}$, and the fitted regression line. *(b)* The residuals from another fitted line.

The Ordinary
Least Squares (OLS) Estimators

We will call the estimators b_1 and b_2, given in equations 2.7 and 2.8, the ordinary least squares estimators. “Ordinary least squares” is abbreviated as OLS

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

(2.8) $$b_1 = \bar{y} - b_2 \bar{x}$$

Other Economic Models

- The simple regression model can be applied to estimate the parameters of many relationships in economics, business, and the social sciences.
- Anytime you ask how much a change in one variable will affect another variable, regression analysis is a potential tool.
- Similarly, anytime you wish to predict the value of one variable given the value of another, the least squares regression is a tool to consider.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Assessing the Least Squares Estimators

Gabelli School of Business

Assessing the Least Squares Estimators

We call b_1 and b_2 the least squares estimators. We can investigate the properties of the estimators b_1 and b_2, which are called their sampling properties, and deal with the following important questions:

1. If the least squares estimators are random variables, then what are their expected values, variances, covariances, and probability distributions?
2. How do the least squares estimators compare with other procedures that might be used, and how can we compare alternative estimators?

The Estimator b_2

- The estimator b_2 can be rewritten as:
  - (2.10) $$b_2 = \sum_{i=1}^{N} w_i y_i$$
- Where:
  - (2.11) $$w_i = \frac{x_i - \bar{x}}{\sum (x_i - x)^2}$$
- It could also be written as:
  - (2.12) $$b_2 = \beta_2 + \sum w_i e_i$$

The Expected Values of b_1 and b_2

- We will show that if our model assumptions hold, then $E(b_2) = \beta_2$, which means that the estimator is **unbiased**; we can find the expected value of b_2 using the fact that the expected value of a sum is the sum of the expected values:
  - (2.13)

$$
\begin{aligned}
E(b_2|\mathbf{x}) &= E(\beta_2 + \sum w_i e_i|\mathbf{x}) = E(\beta_2 + w_1 e_1 + w_2 e_2 + \cdots + w_N e_N|\mathbf{x}) \\
&= E(\beta_2) + E(w_1 e_1|\mathbf{x}) + E(w_2 e_2|\mathbf{x}) + \cdots + E(w_N e_N|\mathbf{x}) \\
&= \beta_2 + \sum E(w_i e_i|\mathbf{x}) \\
&= \beta_2 + \sum w_i E(e_i|\mathbf{x}) = \beta_2
\end{aligned}
$$

- Using $E(e_i) = 0$ and $E(w_i e_i) = w_i E(e_i)$

The Expected Values of b_1 and b_2 (cont.)

- The property of unbiasedness is about the average values of b_1 and b_2 if many samples of the same size are drawn from the same population.
  - If we took the averages of estimates from many samples, these averages would approach the true parameter values b_1 and b_2.
  - Unbiasedness does not say that an estimate from any one sample is close to the true parameter value, and thus we cannot say that an estimate is unbiased.
  - We can say that the least squares estimation procedure (or the least squares estimator) is unbiased.

Sampling Variation

To illustrate how the concept of unbiased estimation relates to sampling variation, we present in Table 2.2 least squares estimates of the food expenditure model from 10 hypothetical random samples.

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

The Variances and Covariance of b_1 and b_2

If the regression model assumptions SR1–SR5 are correct, then the variances and covariance of b_1 and b_2 are:

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

- (2.15) $$\operatorname{var}(b_2) = \frac{\sigma^2}{\sum (x_i - x)^2}$$

- (2.16) $$\operatorname{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\overline{x}}{\sum (x_i - x)^2} \right]$$

Major Points About the Variances and Covariances of b_1 and b_2

1. The larger the variance term, the *greater* the uncertainty there is in the statistical model, and the *larger* the variances and covariance of the least squares estimators.
2. The *larger* the sum of squares, $\sum(x_i - x)^2$, the *smaller* the variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters.
3. The larger the sample size *N*, the *smaller* the variances and covariance of the least squares estimators.
4. The larger the term $\sum x_i^2$, the larger the variance of the least squares estimator b_1.
5. The absolute magnitude of the covariance *increases* the larger in magnitude the sample mean $\bar{x}$ is, and the covariance has a *sign* opposite to that of $\bar{x}$.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Gauss–Markov Theorem

The Gauss–Markov Theorem

Given **x** and 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 b_1 and b_2. **They are the best linear unbiased estimators (BLUE)** of b_1 and b_2.

Major Points About the Gauss–Markov Theorem

1. The estimators b_1 and b_2 are "best" when compared with similar estimators, those which are linear and unbiased. The theorem does not say that b_1 and b_2 are the best of all possible estimators.
2. The estimators b_1 and b_2 are best within their class because they have the minimum variance. When comparing two linear and unbiased estimators, we *always* want to use the one with the smaller variance because that estimation rule gives us the higher probability of obtaining an estimate that is close to the true parameter value.
3. In order for the Gauss–Markov theorem to hold, assumptions SR1–SR5 must be true. If any of these assumptions are *not* true, then b_1 and b_2 are *not* the best linear unbiased estimators of β_1 and β_2.

Major Points About the Gauss–Markov Theorem (cont.)

4. The Gauss–Markov theorem does *not* depend on the assumption of normality (assumption SR6).
5. In the simple linear regression model, if we want to use a linear and unbiased estimator, then we have to do no more searching. The estimators b_1 and b_2 are the ones to use. This explains why we are studying these estimators and why they are so widely used in research, not only in economics, but in all social and physical sciences as well.
6. The Gauss–Markov theorem applies to the least squares estimators. It *does not* apply to the least squares *estimates* from a single sample.

The Probability Distributions of the Least Squares Estimators

If we make the normality assumption (assumption SR6 about the error term), and treat **x** as given, then the least squares estimators are normally distributed:

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

A Central Limit Theorem

A central limit theorem: If assumptions SR1–SR5 hold, and if the sample size N is sufficiently large, then the least squares estimators have a distribution that approximates the normal distributions shown in equations 2.17 and 2.18.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Estimating the Variance of the Error Term

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Estimating the Variance of the Error Term

- The variance of the random error e_i is:

$$\text{var}(e_i \mid x) = \sigma^2 = E\{[e_i - E(e_i \mid x)]^2 \mid x\} = E(e_i^2 \mid x)$$

- If the assumption $E(e_i) = 0$ is correct
  - Because the "expectation" is an average value, we might consider estimating σ^2 as the average of the squared errors

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{N}$$

  - Where the error terms are $e_i = y_i - \beta_1 - \beta_2 x_i$

Estimating the Variances and Covariance of the Least Squares Estimators

Replace the unknown error variance σ^2 in equations 2.14–2.16 by $\hat{\sigma}^2$ to obtain:

- (2.20) $\widehat{\operatorname{var}}(b_1|\mathbf{x}) = \hat{\sigma}^2 \left[\dfrac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right]$

- (2.21) $\widehat{\operatorname{var}}(b_2|\mathbf{x}) = \dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$

- (2.22) $\widehat{\operatorname{cov}}(b_1, b_2|\mathbf{x}) = \hat{\sigma}^2 \left[\dfrac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right]$

Estimating the Variances and Covariance of the Least Squares Estimators (cont.)

Replace the unknown error variance σ^2 in equations 2.14–2.16 by $\hat{\sigma}^2$ to obtain:

- (2.23) $se(b_1) = \sqrt{\widehat{\mathrm{var}}(b_1|x)}$
- (2.24) $se(b_2) = \sqrt{\widehat{\mathrm{var}}(b_2|x)}$

Interpreting the Standard Errors, Part I

The standard errors of b_1 and b_2 are measures of the sampling variability of the least squares estimates b_1 and b_2 in repeated samples.

- The estimators are random variables. As such, they have probability distributions, means, and variances.
- In particular, if assumption SR6 holds, and the random error terms e_i are normally distributed, then:

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

Interpreting the Standard Errors, Part II

The estimator variance, var(b_2), or its square root, $\sigma_{b_2} = \sqrt{\text{var}(b_2 \mid x)}$, which we might call the true standard deviation of b_2, measures the sampling variation of the estimates b_2.

- The bigger σ_{b_2} is, the more variation in the least squares estimates b_2 we see from sample to sample. If σ_{b_2} is large, then the estimates might change a great deal from sample to sample.
- If σ_{b_2} is small relative to the parameter b_2, we know that the least squares estimate will fall near b_2 with high probability.

Interpreting the Standard Errors, Part III

The question we address with the standard error is
“How much variation about their means do the estimates exhibit from sample to sample?”

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK
Gabelli School of Business

Estimating Nonlinear Relationships

FORDHAM | Gabelli School of Business
THE JESUIT UNIVERSITY OF NEW YORK

Estimating Nonlinear Relationships

- Economic variables are not always related by straight-line relationships; in fact, many economic relationships are represented by curved lines, and are said to display *curvilinear forms*.
- Fortunately, the simple linear regression model $y = \beta_1 + \beta_2 + e$ is much more flexible than it looks at first glance.
- The variables *y* and *x* can be transformations, involving logarithms, squares, cubes, or reciprocals of the basic economic variables, or they can be indicator variables that take only the values zero and one.

Nonlinear Relationships
House Price Example

- Consider the linear model of house prices
- Where *SQFT* is the square footage
  - It may be reasonable to assume that larger and more expensive homes have a higher value for an additional square foot of living area than smaller, less expensive homes
- We can build this into our model in two ways
  1. A quadratic equation in which the explanatory variable is $SQFT^2$
  2. A log-linear equation in which the dependent variable is ln(*PRICE*)

Quadratic Functions

- The quadratic function $y = \beta_1 + \beta_2 x^2$ is a parabola
- The elasticity, or the percentage change in y given a 1% change in x, is:

$$\varepsilon = \text{slope} \times x/y = 2bx^2/y$$

Using a Quadratic Model

- A quadratic model for house prices includes the squared value of *SQFT*, giving:
  - (2.26) $$PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$$
- The slope is:
  - (2.27) $$\frac{d\left(PRICE\right)}{dSQFT} = 2\hat{\alpha}_2 SQFT$$
- If $\hat{\alpha}_2 > 0$, then larger houses will have larger slope, and a larger estimated price per additional square foot

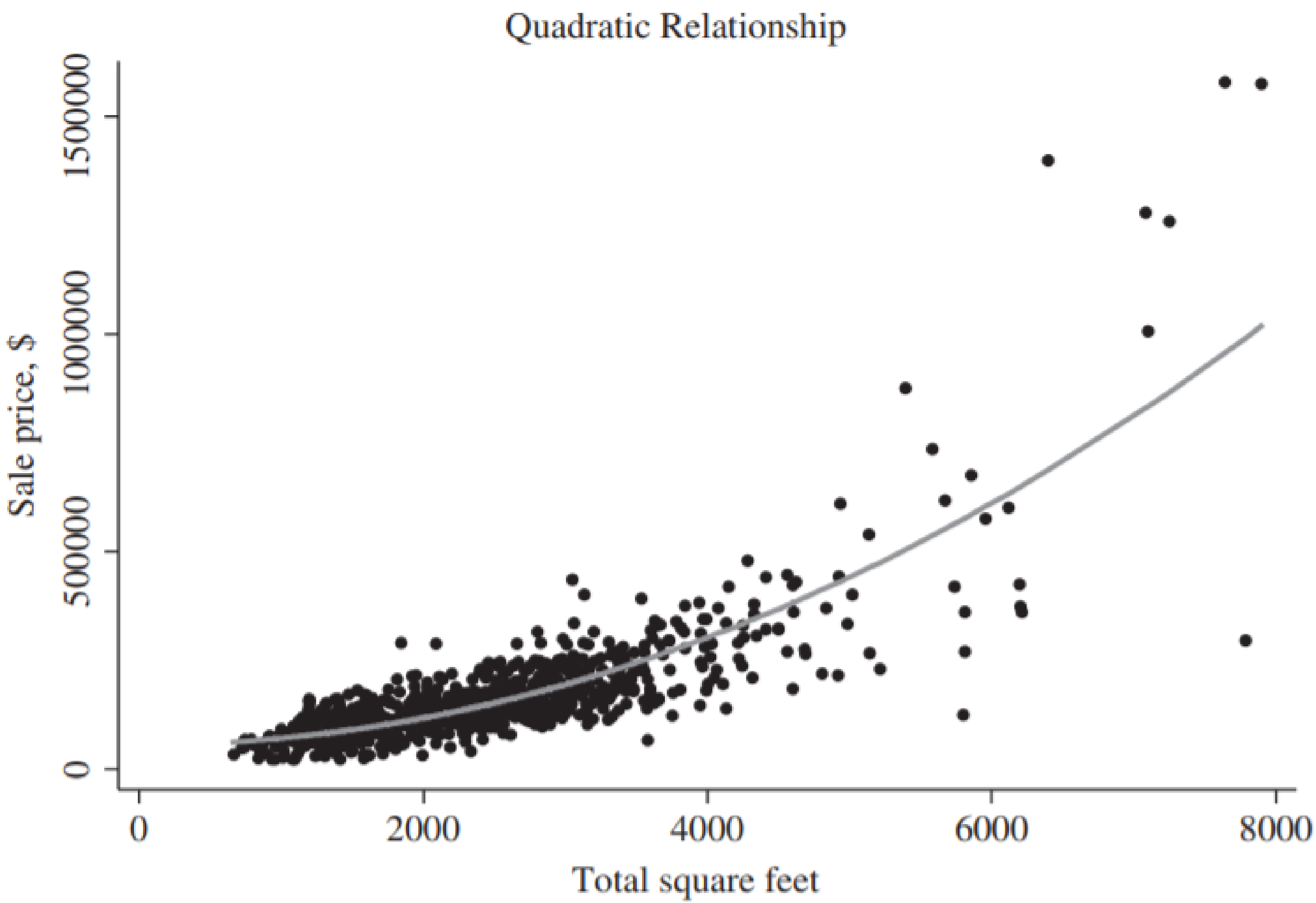

FIGURE 2.14 **A fitted quadratic relationship.**

A Log-Linear Function

The log-linear equation $\ln(y) = a + bx$ has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side

- Both its slope and elasticity change at each point and are the same sign as *b*
  - The slope is $dy/dx = by$
- The elasticity, the percentage change in *y* given a 1% increase in *x*, at a point on this curve is:

$$\varepsilon = slope \times x/y = bx$$

Using a Log-Linear Model

Consider again the model for the price of a house as a function of the square footage, but now written in semi-log form:

- (2.29) $\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$
- This logarithmic transformation can regularize data that are skewed with a long tail to the right

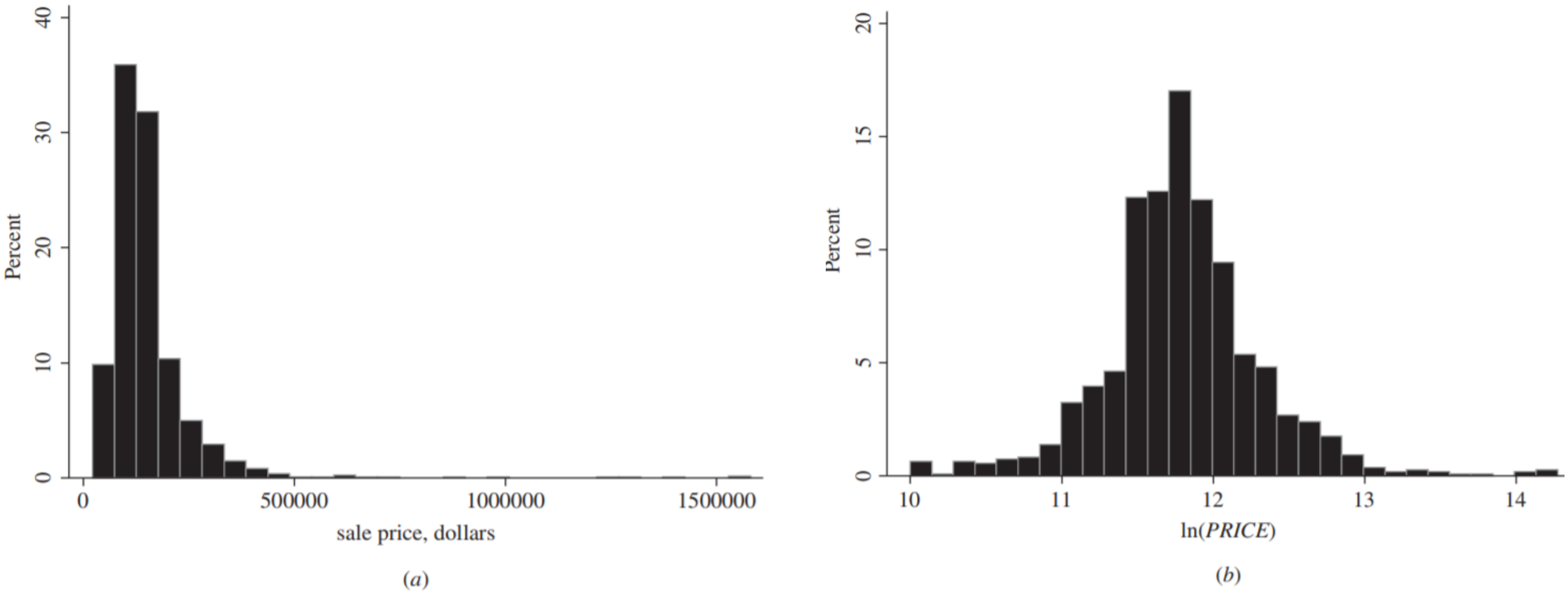

Figure 2.16 (a) Histogram of PRICE, (b) histogram of ln(PRICE)

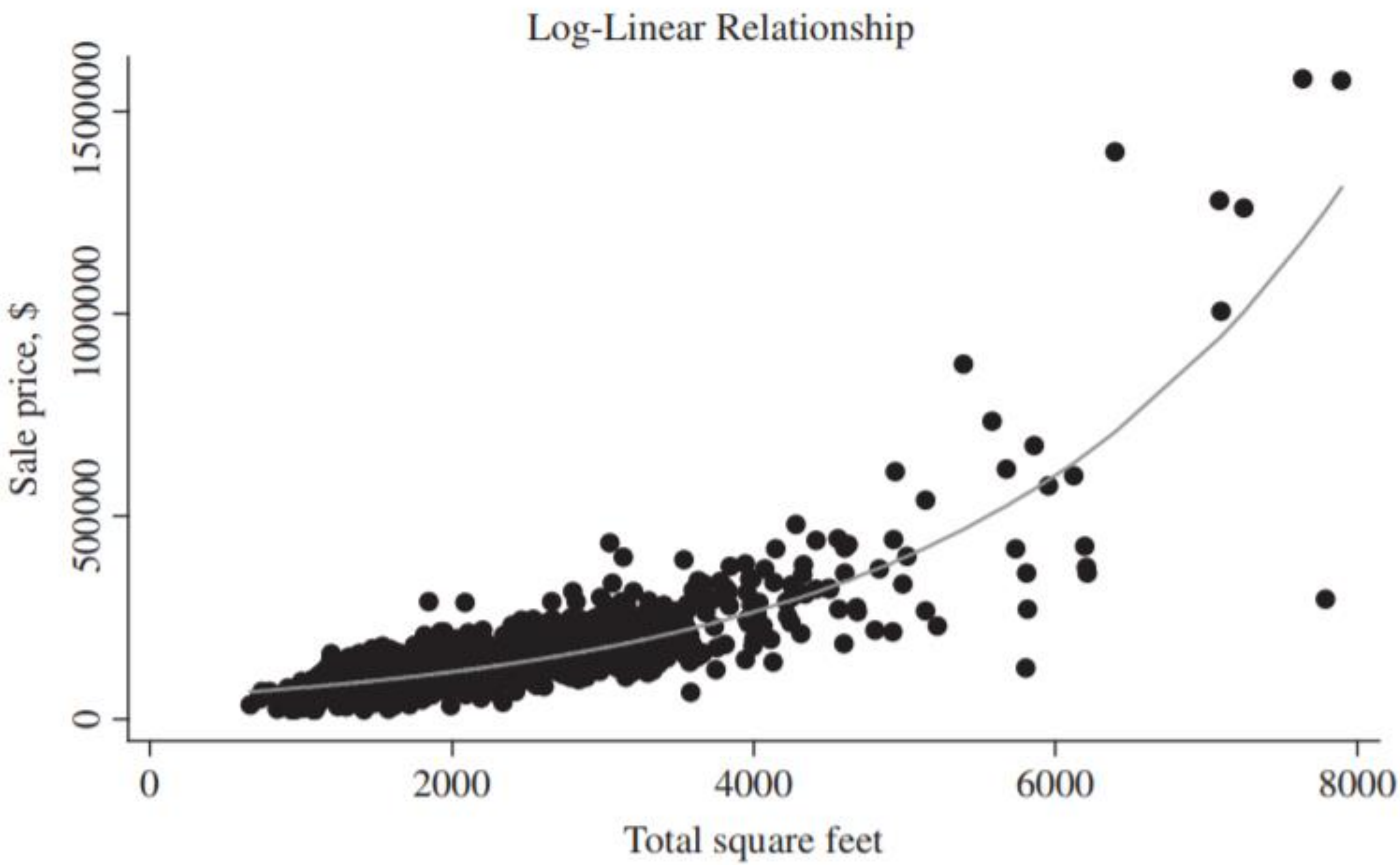

FIGURE 2.17 **The fitted log-linear model.**

Choosing a Functional Form

- We should do our best to choose a functional form:
  - That is consistent with economic theory
  - That fits the data well
  - Such that the assumptions of the regression model are satisfied
- In real-world problems, it is sometimes difficult to achieve all these goals
  - In applications of econometrics, we must simply do the best we can to choose a satisfactory functional form

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Indicator Variables

Regression With Indicator Variables

- An indicator variable is a binary variable that takes the values zero or one; it is used to represent a nonquantitative characteristic, such as gender, race, or location

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

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

  - How do we model this?

Regression With Indicator Variables (cont.)

- When an indicator variable is used in a regression, it is important to write the regression function for the different values of the indicator variable

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

- In the simple regression model, an indicator variable on the right-hand side gives us a way to estimate the differences between population means

The Independent Variable

This section contains a more advanced discussion of the assumptions of the simple regression model.

- We say more about different possible DGPs.
- We explore their implications for the assumptions of the simple regression model.
- We investigate how the properties of the least squares estimator change, if at all, when we no longer condition on x.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

The Independent Variable

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Random and Independent X

Assumptions of the Independent Random-*x* Linear Regression Model

IRX1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i, i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

IRX2: The random error has mean zero, $E(e_i) = 0$.

IRX3: The random error has constant variance, $\text{var}(e_i) = \sigma^2$.

IRX4: The random errors e_i and e_j for any two observations are uncorrelated, $\text{cov}(e_i, e_j) = 0$.

IRX5: The random errors $e_1, e_2, \ldots, e_N$ are statistically independent of $x_1, \ldots, x_N$, and x_i takes at least two different values.

IRX6: $e_i \sim N(0, \sigma^2)$.

Random and Strictly Exogenous X

- Statistical independence between x_i and x_j for all values of *i* and *j* is a very strong assumption and most likely only suitable in experimental situations
- A weaker assumption is that the explanatory variable x is strictly exogenous
- The implications of strict exogeneity
  - Implication 1: $E(e_i) = 0$. The "average" of all factors omitted from the regression model is zero.
  - Implication 2: $cov(x_i, e_j) = 0$. There is no correlation between the omitted factors associated with observation *j* and the value of the explanatory variable for observation.

Random Sampling

- Survey methodology is an important area of statistics.
- The idea is to collect data pairs (y_i, x_i) in such a way that the ith pair is statistically independent of the pair.
- This ensures that x_j is statistically independent of e_j.

Assumptions of the Simple Linear Regression Model Under Random Sampling

RS1: The observable variables y and x are related by $y_i = \beta_1 + \beta_2 x_i + e_i$, $i = 1, \ldots, N$, where β_1 and β_2 are unknown population parameters and e_i is a random error term.

RS2: The data pairs (y_i, x_i) are statistically independent of all other data pairs and have the same joint distribution $f(y_i, x_i)$. They are independent and identically distributed.

RS3: $E(e_i|x_i) = 0$ for $i = 1, \ldots, N$; x is strictly exogenous.

RS4: The random error has constant conditional variance, $\text{var}(e_i|x_i) = \sigma^2$.

RS5: x_i takes at least two different values.

RS6: $e_i \sim N(0, \sigma^2)$.

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Key Words

- Assumptions
- Asymptotic
- Biased estimator
- BLUE
- Degrees of freedom
- Dependent variable
- Deviation from the mean form
- Econometric model
- Economic model
- Elasticity
- Exogenous variable
- Gauss–Markov theorem
- Heteroskedastic
- Homoskedastic
- Independent variable
- Indicator variable
- Least squares estimates
- Least squares estimators
- Least squares principle
- Linear estimator
- Log-linear model
- Nonlinear relationship
- Prediction
- Quadratic model
- Random error term
- Random-x
- Regression model
- Regression parameters
- Repeated sampling
- Sampling precision
- Sampling properties
- Simple linear regression analysis
- Simple linear regression
- Specification error
- Strictly exogenous
- Unbiased estimator

Copyright

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business