Lecture 2: Simple Linear Regression Model

Shuo Liu

UCLA Summer School Econ 103

June 28, 2017

Outline

General Introduction

- Starting from Lecture 2, we use lower case y and x to denote the economic variables in the regression models (different from Lecture 1), and $\{y_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$ are realized/sample values of those economic variables;
- The economic variable y that we want to explain (e.g. expenditure) is called "**dependent variable**" and the economic variable x that we want to use (e.g. income) to explain the "dependent variable" is called "**independent**" or "**explanatory**" variable;
- In Econometrics, we want to use the data $\{y_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$ to learn about the relationship between y and x.

General Introduction

General Question to Answer: conditional on x (e.g. income), what is your expectation on y (e.g. expenditure)?

$\Longleftrightarrow$

$$E[y|x] =?$$

- **Firstly**, you should have an assumption on the form of $E[y|x]$ (based on data pattern/model selection techniques/just try different forms and etc). The "hypothesized" form is also called the "theoretical regression model";
- **Secondly**, you need to use data $\{y_i\}_{i=1}^n$ and $\{x_i\}_{i=1}^n$ to estimate the "theoretical regression model" to get the "fitted/estimated model";
- **Then**, assess whether each estimated parameter and the fitted model as a whole are good or not.

General Introduction

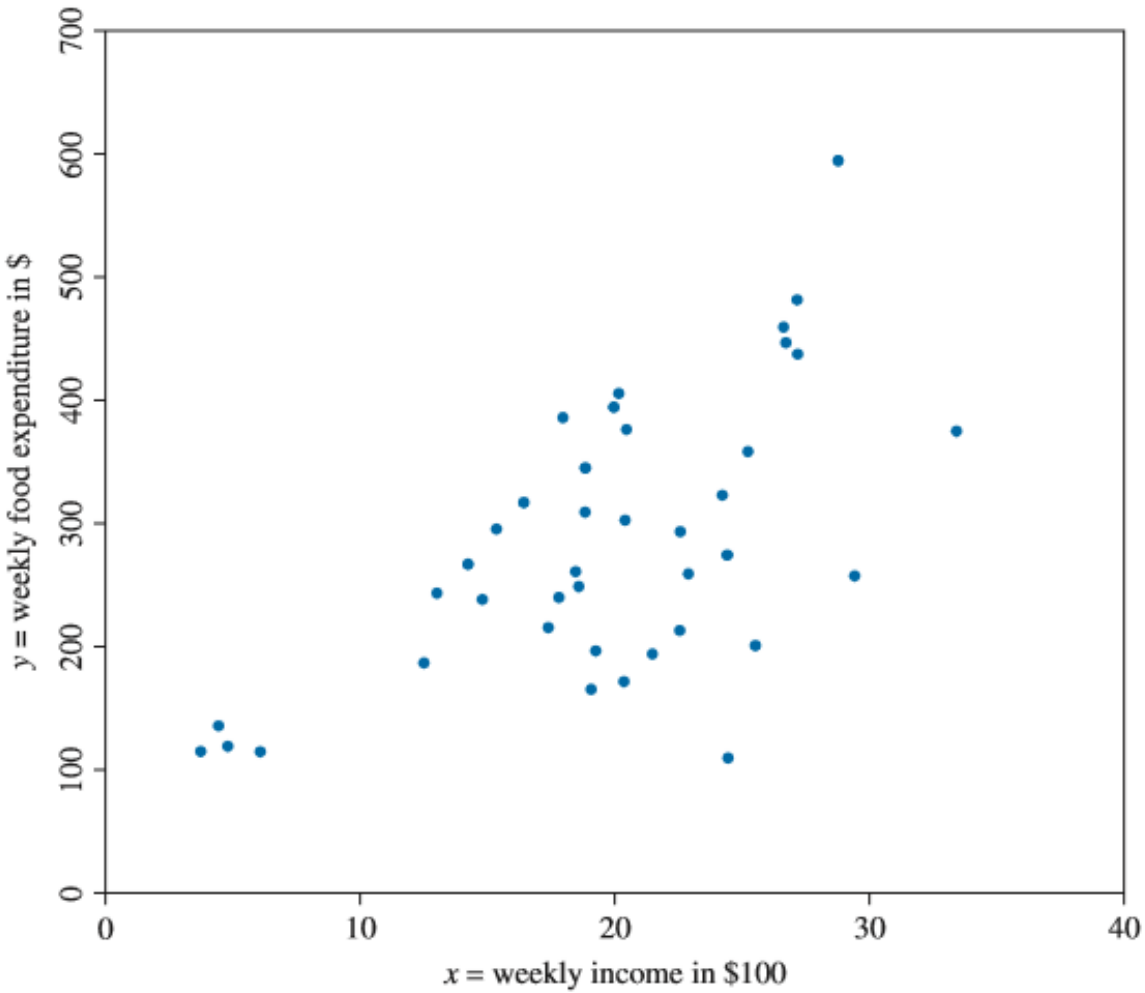

Figure: Data on Food Expenditure and Income

Simple Linear Regression Model

$$E[y|x] = \beta_1 + \beta_2 x \tag{1}$$

where β_1 is the intercept and β_2 is the slope. The model is called "simple linear regression model" because there is only one independent/explanatory variable on the right-hand side and the model is linear in parameters β_1 and β_2. The line in equation (1) is the "theoretical/true line".

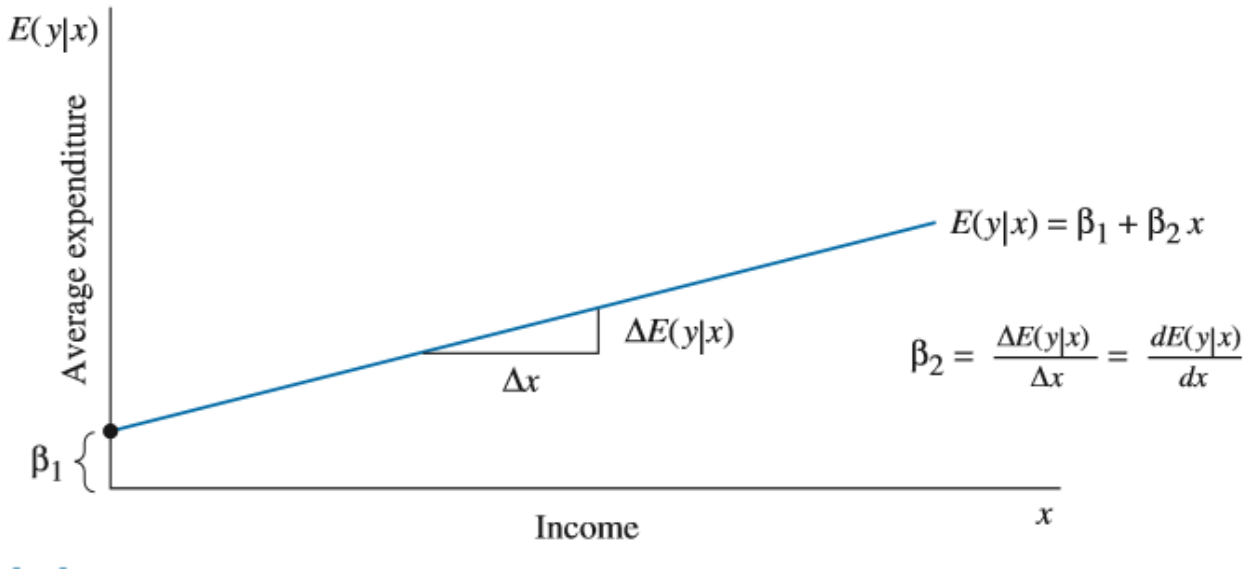

Figure: Linear Relationship between Average Food Expenditure and Income

General Introduction

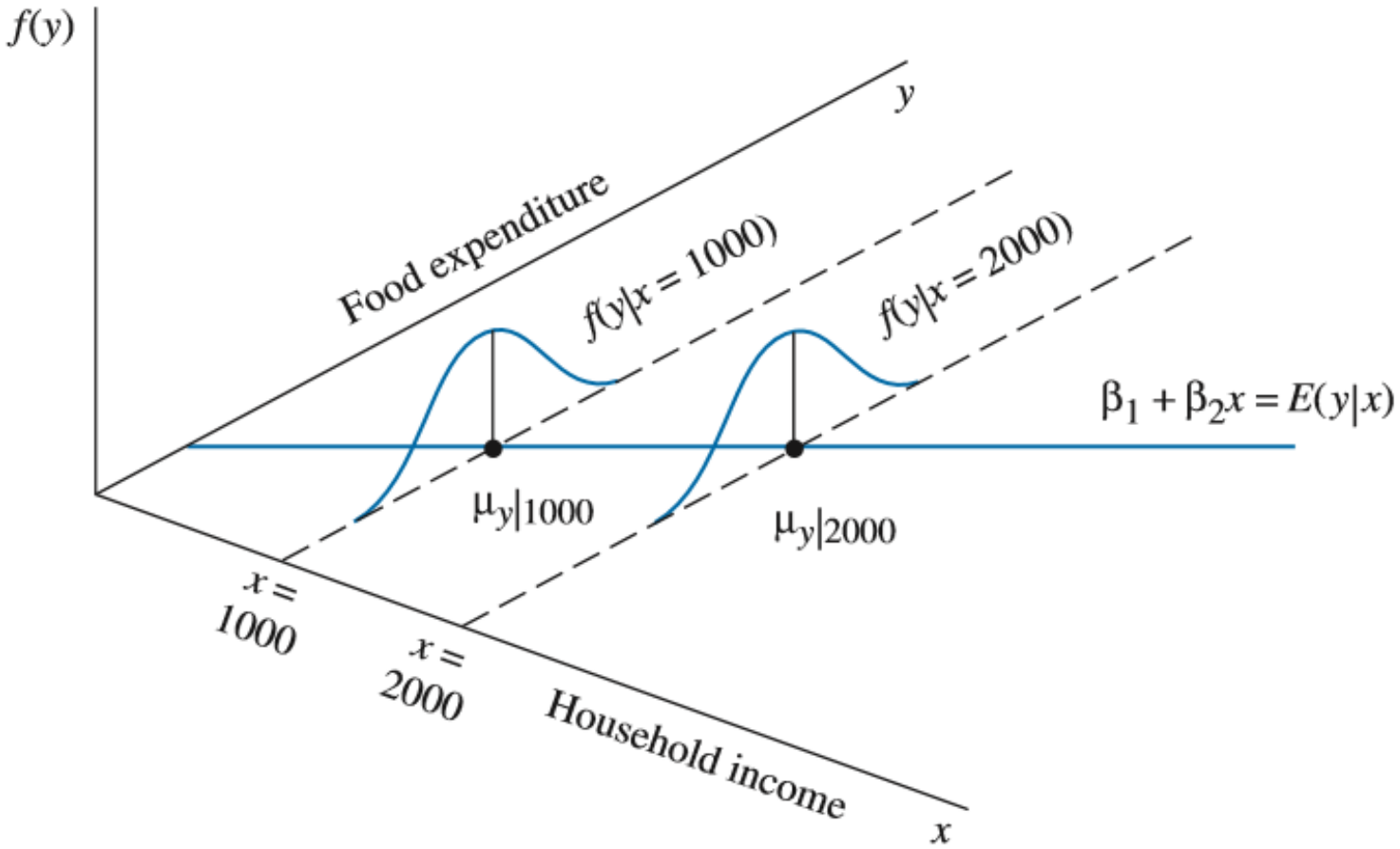

Figure: Conditional pdf's for y at Alternative Levels of Income

Assumptions on the "Theoretical/True" Model (Line)

First version of assumptions on the theoretical/true model:

- **A1—Linearity**: for each value of x, the mean value of y is given by

$$E[y|x] = \beta_1 + \beta_2 x$$

- **A2—Constant variance**: for each value of x, the variance of distribution of y is constant

$$Var(y|x) = \sigma^2$$

- **A3—Uncorrelatedness**: the sample values of y are all uncorrelated.

$$Cov(y_i, y_j) = 0 \quad \text{for all } i \neq j$$

- **A4—Constant x's**: x **is not random** and must take at least two different values (more than one sample points to be used for estimation), y **is random**.
- **A5 (optional)—Normality**: conditional on each value of x, y is normally distributed

$$y|x \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

Assumptions on the "Theoretical/True" Model (Line)

Based on original true model, we define the **random error term** e as:

$$e = y - E[y|x] = y - \beta_1 - \beta_2 x \tag{2}$$

then by rearranging, we have the "second version" of the true model:

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

And by equation (2), it is easy to show that

$$E[e|x] = 0$$

Assumptions on the "Theoretical/True" Model (Line)

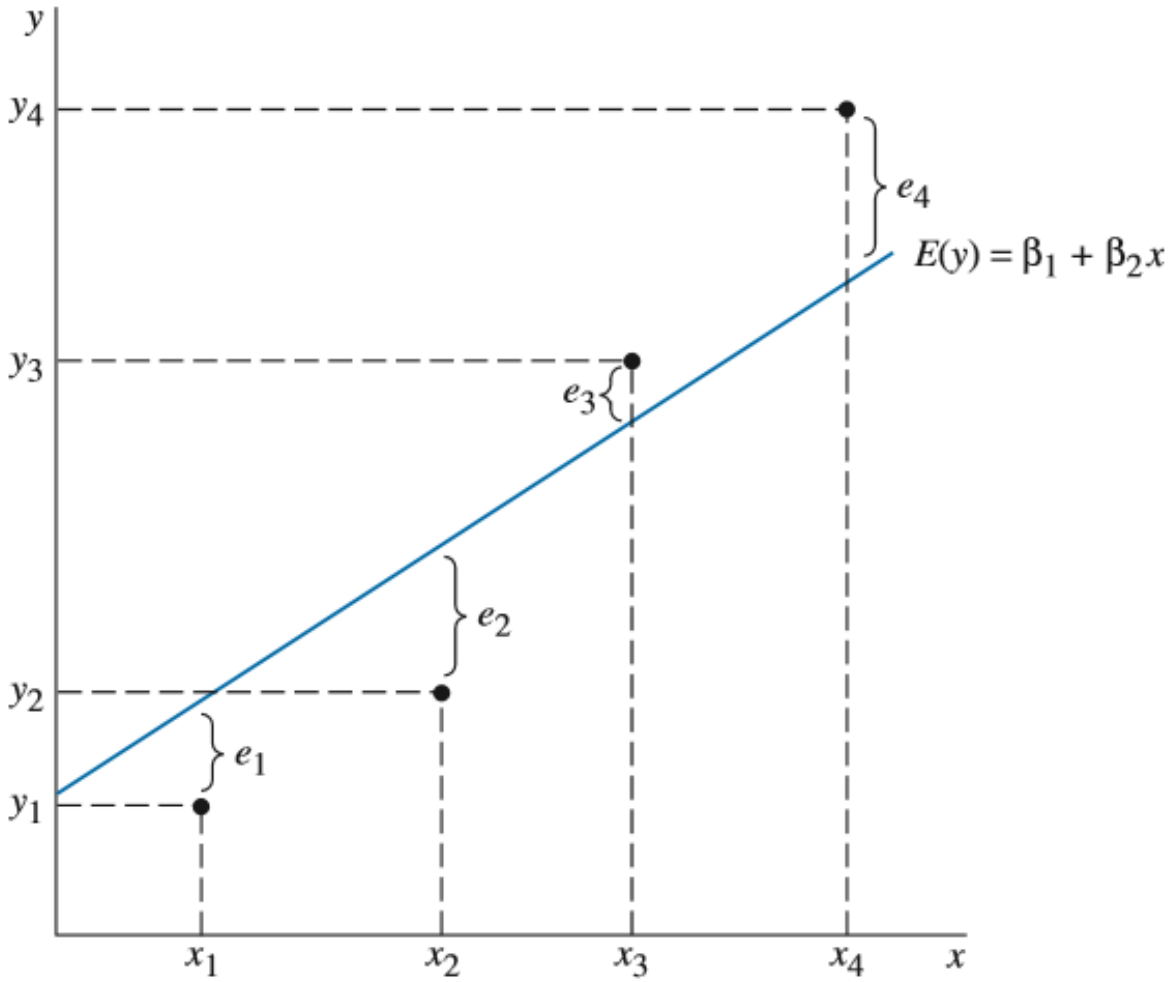

Figure: The Relationship between y, e and the "True" Regression Model

Assumptions on the "Theoretical/True" Model (Line)

Second version of assumptions on the theoretical/true model

- **SR1**: for each value of x, the value of y is

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

- **SR2**: for each value of x,

$$E[e|x] = 0$$

  which is stronger than simply assuming $E(e) = 0$.
- **SR3**: for each value of x, the conditional variance of the random error is

$$Var(e|x) = \sigma^2 \Longrightarrow Var(y|x) = \sigma^2$$

  y and e differ only by a constant due to $SR5$.
- **SR4**: the covariance between any pair of random errors,

$$Cov(e_i, e_j) = 0 \quad \text{for all } i \neq j \Longrightarrow Cov(y_i, y_j) = 0 \quad \text{for all } i \neq j$$

- **SR5**: x **is not random** and must take at least two different values (more than one sample points to be used for estimation), y **is random**. The randomness of y comes from the randomness of e.

Assumptions on the "Theoretical/True" Model (Line)

Second version of assumptions on the theoretical/true model

- **SR6(Optional)**:

$$e \sim N(0, \sigma^2) \Longrightarrow y|x \sim N(\beta_1 + \beta_2 x, \sigma^2)$$

- **SR6** can give us normality, which will be useful to **characterize the distributions** that the statistics/moments associated with estimators follow, and then conduct hypothesis testing.
- If **SR6** does not apply, we are still able to get "normality" through **central limit theorem** if we have large enough sample data.

The "Fitted" Line: OLS Estimation

Notation for "fitted line":

$$\hat{y} = b_1 + b_2 x$$

$\Longleftrightarrow$

$$\hat{y}_i = b_1 + b_2 x_i \quad \text{for any observation } x_i, \text{ the estimated } y \text{ is } \hat{y}_i$$

To measure the "distance" between true observed y_i and estimated $\hat{y}_i$, we define the **residual**:

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

The "Fitted" Line: OLS Estimation

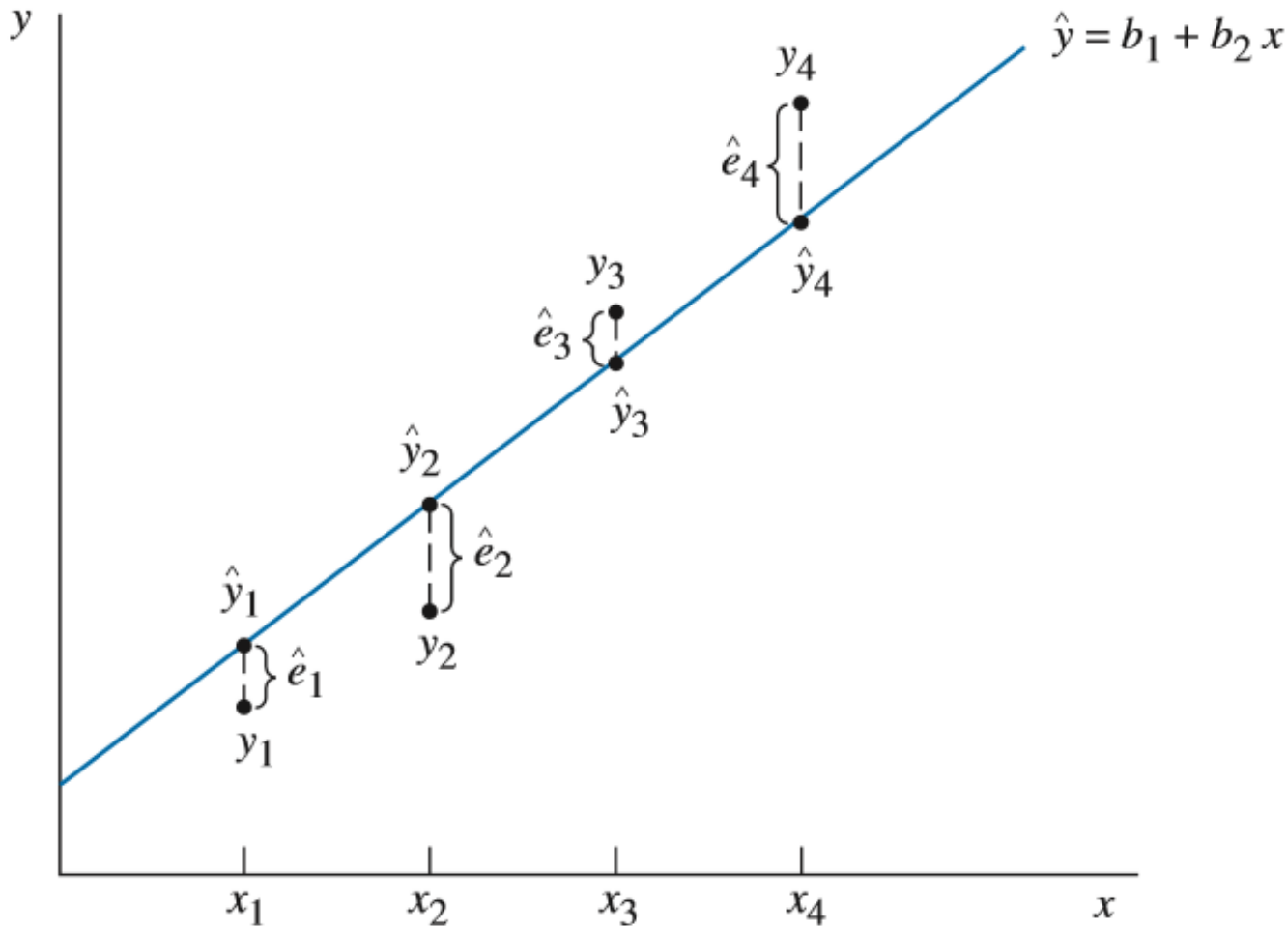

Figure: The Relationship between y, $\hat{e}$ and Fitted Regression Line

The "Fitted" Line: OLS Estimation

How to determine our "fitted/estimated line" is the best one? (as a whole closest to all data points)

We use "sum of squared error" SSE, for one "fitted line", we can calculate residuals $\hat{e}_i$ for all data points, then calculate SSE as:

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2$$

Suppose now we have two fitted lines:

- **Line 1**: $\hat{y}_i^* = b_1^* + b_2^* x_i$, calculate $\hat{e}_i^* = y_i - \hat{y}_i^*$ and $SSE_1 = \sum_{i=1}^{n} \hat{e}_i^{*2}$;
- **Line 2**: $\hat{y}_i^{**} = b_1^{**} + b_2^{**} x_i$, calculate $\hat{e}_i^{**} = y_i - \hat{y}_i^{**}$ and $SSE_2 = \sum_{i=1}^{n} \hat{e}_i^{**2}$.

then Line 1 is better than Line 2 $\Longleftrightarrow SSE_1 < SSE_2$.

The "Fitted" Line: OLS Estimation

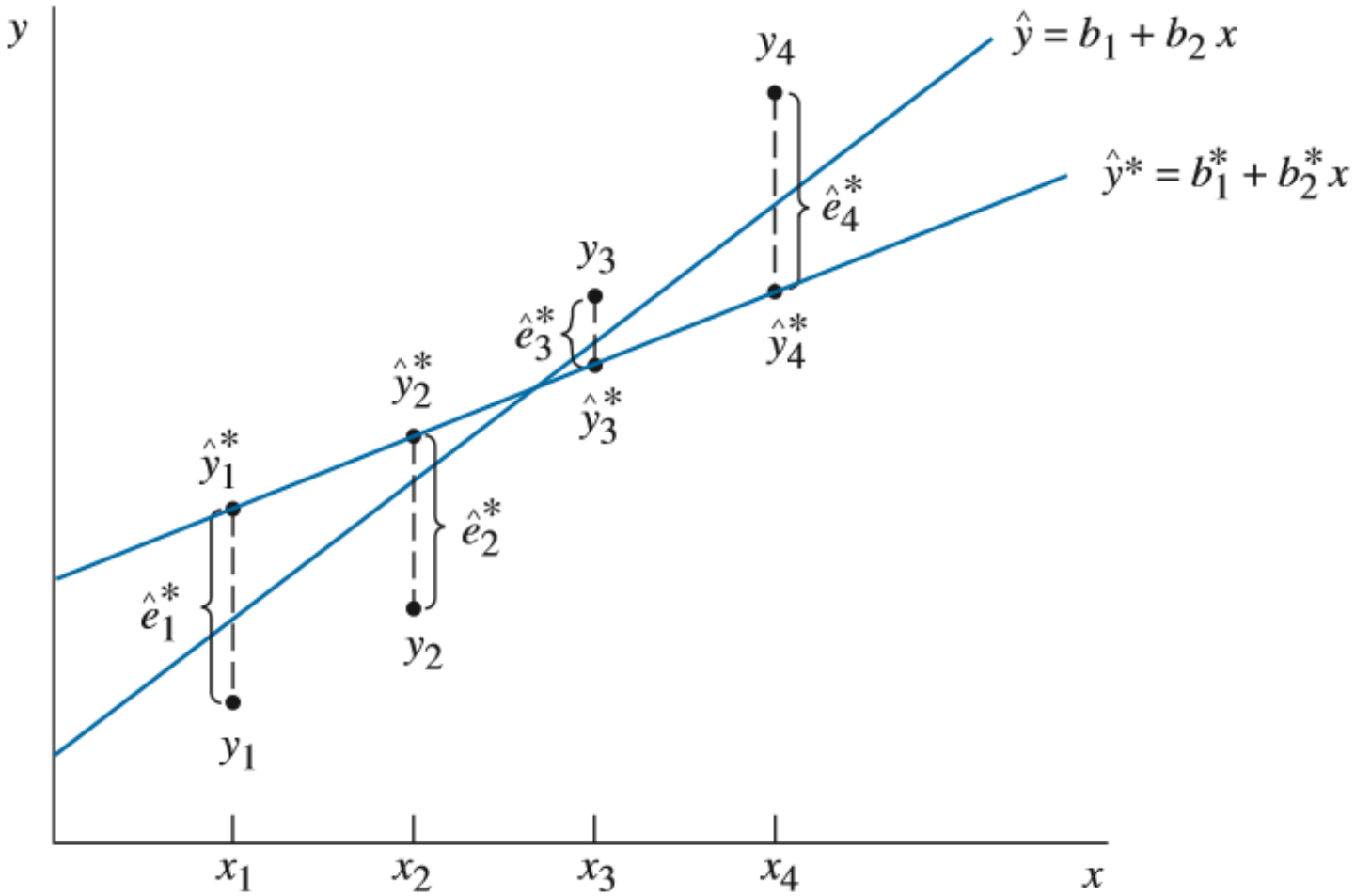

Figure: The Relationship between y, $\hat{e}$ and Alternative Fitted Regression Line

The "Fitted" Line: OLS Estimation

OLS: "Ordinary Least Square" Estimation

Given sample data $\{(x_i, y_i)\}_{i=1}^n$, choose b_1 and b_2 to minimize the "sum of squares function":

$$S(b_1, b_2) = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$$

The optimal (b_1, b_2) satisfies the following necessary conditions:

$$\frac{\partial S(b_1, b_2)}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) = 0 \tag{3}$$

$$\frac{\partial S(b_1, b_2)}{\partial b_2} = -2 \sum_{i=1}^{n} x_i (y_i - b_1 - b_2 x_i) = 0 \tag{4}$$

The "Fitted" Line: OLS Estimation

(3) and (4) $\Longrightarrow$

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

$$b_1 = \bar{y} - b_2\bar{x} \tag{6}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

For example, y is food expenditure (unit: \$1000) and x is personal income (unit: \$1000/*month*), and we estimate $b_2 = 10.2$ and $b_1 = 83.4$, how to interpret?

$$\hat{y}_i = 83.4 + 10.2x_i$$

Alternative Theoretical Model Form

Alternative form: need transformation of original data

$$ln(y) = \alpha + \beta ln(x) + e \tag{7}$$

$$y = \alpha + \beta ln(x) + e \tag{8}$$

$$ln(y) = \alpha + \beta x + e \tag{9}$$

Comments:

- We always firstly take "ln" on original economic data;
- By taking "ln", the coefficient may have more intuitive interpretation. For example, if your model is (7):

$$\beta = \frac{dln(y)}{dln(x)} = \frac{dy/y}{dx/x} = \frac{\triangle y/y}{\triangle x/x} = \frac{\triangle y}{\triangle x}\frac{x}{y} = \eta_{yx} \tag{10}$$

  so that, if estimator of β is b, how to interpret b?
- Alternatively, if your model is $y = \alpha + \beta x + e$, then what if we still want to estimate η_{yx} (elasticity of variable y with respect to x) for different levels of (x, y)? usually we calculate the elasticity at the "point of means" $(\bar{x}, \bar{y})$.

Alternative Theoretical Model Form

- If your model is (9), then how to interpret β?

$$\beta = \frac{dln(y)}{dx} = \frac{dy/y}{dx} = \frac{\triangle y/y}{\triangle x}$$

then how to estimate η_{yx} (elasticity of variable y with respect to x) for different levels of x?

For example, y is house price, x is house square feet, the estimated model is:

$$ln(PRICE) = 10.839 + \underbrace{0.00041}_{b} \times SQFT$$

(1) The estimated **semi-elasticity**: $b = 0.00041$;
(2) The estimated elasticity is:

$$\hat{\eta} = b \times SQFT = 0.00041 \times SQFT$$

(3) For a house with 4000 square feet, the estimated elasticity is 1.645.

Properties of OLS Estimators

After obtaining OLS estimator, next is to **assess the OLS estimators** through the following steps:

- Calculate related "theoretical moments" and characterize "theoretical distributions" of the OLS estimators;
- Use sample data to estimate the "theoretical moments" (if you cannot calculate the theoretical moments directly);
- Provide interval estimation;
- Conduct hypothesis testing, either for each parameter separately or for the whole model, to evaluate whether your fitted model is good or not.

Properties of OLS Estimators

True model and estimated model are:

$$y = \beta_1 + \beta_2 x + e \tag{11}$$

$$\hat{y} = b_1 + b_2 x \tag{12}$$

Distribution and Related Moments (expectations, variances, covariances) of b_1 and b_2:

1. Expectation

- $$b_2 = \sum_{i=1}^{n} w_i (y_i - \bar{y}) = \sum_{i=1}^{n} w_i y_i - \underbrace{\sum_{i=1}^{n} w_i \bar{y}}_{=0} = \beta_2 + \sum_{i=1}^{n} w_i e_i \tag{13}$$

  where $\bar{y} = \beta_1 + \beta_2 \bar{x}$. Then

$$E(b_2) = \beta_2 \quad \text{by } E(e_i) = 0 \tag{14}$$

Properties of OLS Estimators

- $$b_1 = \bar{y} - b_2\bar{x} = \beta_1 + (\beta_2 - b_2)\bar{x} \tag{15}$$

  then

$$E(b_1) = \beta_1 \tag{16}$$

Both estimators b_1 and b_2 are **unbiased** estimator. **How to understand?** If we take the two averages of estimates b_1 and b_2 from many samples (the number of samples converges to infinity), then the two averages will converge to β_1 and β_2.

2. Variance and Covariance

If assumptions $SR1 - SR5$ are correct, we have

$$Var(b_1) = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{17}$$

Properties of OLS Estimators

$$Var(b_2) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{18}$$

$$Cov(b_1, b_2) = \sigma^2 \frac{-\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{19}$$

Comments:

- the larger the variance term σ^2, the greater the uncertainty in the statistical model, and the larger the variances and covariance of the OLS estimators;
- the larger the sum of squares of variation of the independent variables $\sum_{i=1}^{n}(x_i - \bar{x})^2$, the smaller the variances of the least squares estimators and the more precisely we can estimate the unknown parameters.

Properties of OLS Estimators

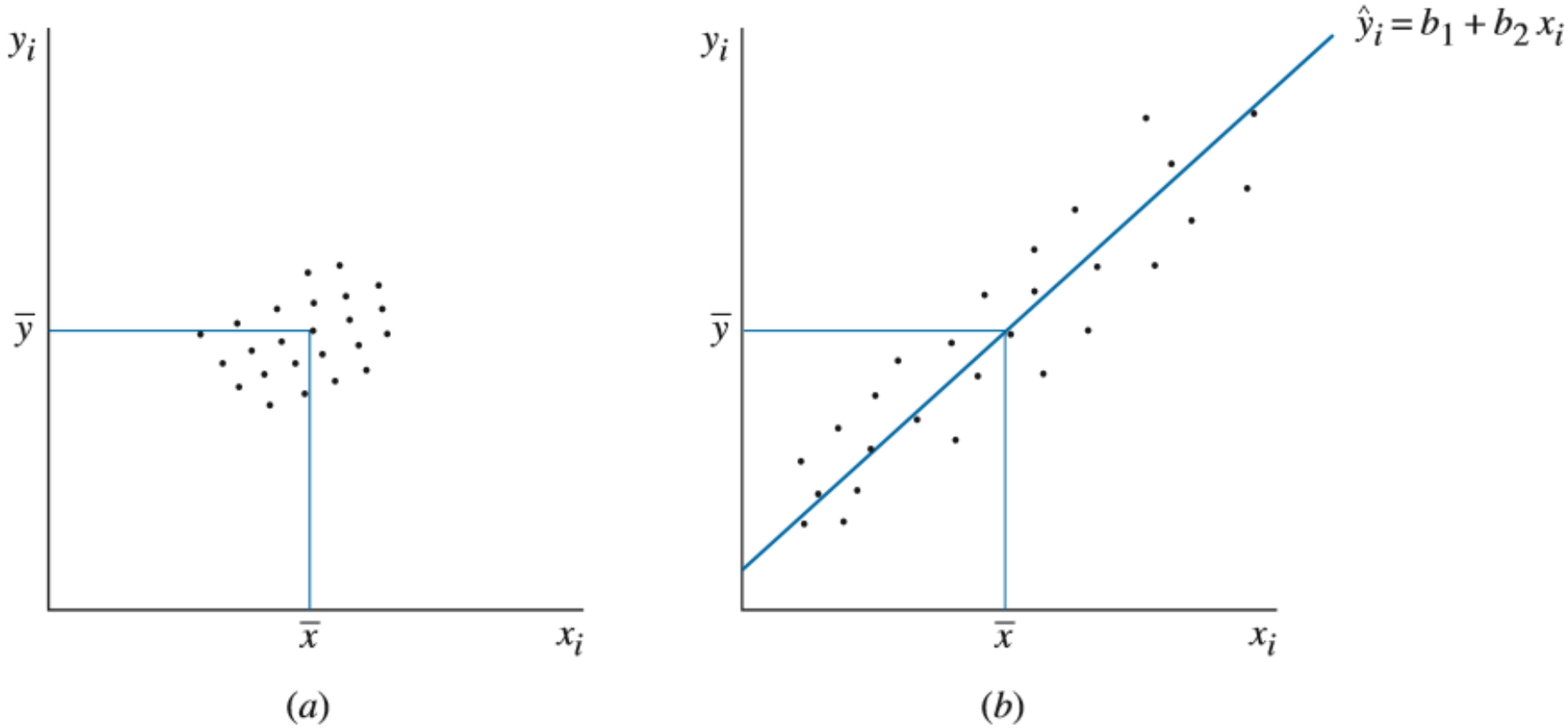

FIGURE 2.11 The influence of variation in the explanatory variable x on precision of estimation: (a) low x variation, low precision: (b) high x variation, high precision.

3. Probability Distribution of the OLS Estimators b_1 and b_2

- **Normality Assumption**: if we make the normality assumption (SR6) about the error term $e \sim N(0, \sigma^2)$, then the OLS estimators are normally distributed:

$$b_1 \sim N\left(\beta_1, \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{20}$$

$$b_2 \sim N\left(\beta_2, \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{21}$$

- **Without Normality Assumption**: if assumptions SR1-SR5 hold and the sample size n is sufficiently large, then OLS estimators b_1 and b_2 have distributions that **approximate** the normal distributions shown in (20) and (21), by **Central Limit Theorem**.

Properties of OLS Estimators

Why do we want to characterize the probability distributions of b_1 and b_2?

Because we want to (1) do hypothesis testing, e.g. if we get estimator b_2, we want to test **whether the corresponding parameter β_2 is significantly different from zero**; (2) provide interval estimation, i.e. confidence interval of β_2, not only point estimator b_2.

Example of hypothesis testing (if (21) applies and $Var(b_2) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is known):

$$H_0 : \beta_2 = 0;$$

$$H_1 : \beta_2 \neq 0; \text{or} \beta_2 > 0; \text{or} \beta_2 < 0.$$

4. Estimate Theoretical Moments (mainly variances and covariances of OLS estimators)

In last example, we know the value σ^2 thus know $Var(b_2)$, but usually we don't know. **Then what should we do**?

- Estimate $Var(e) = \sigma^2$:
  since $Var(e) = E(e^2) - E(e)^2 = E(e^2)$, we use residuals $\hat{e}_i$,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 \tag{22}$$

  and it can be proved $E[\hat{\sigma}^2] = \sigma^2$. (22) is for simple regression model, **what if we have K parameters to estimate in multiple regression model**?

Properties of OLS Estimators

- Estimate $Var(b_1), Var(b_2), Cov(b_1, b_2)$:

$$\widehat{Var}(b_1) = \hat{\sigma}^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{23}$$

$$\widehat{Var}(b_2) = \hat{\sigma}^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{24}$$

$$\widehat{Cov}(b_1, b_2) = \hat{\sigma}^2 \frac{-\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{25}$$

- The square roots of the estimated variances are the "standard errors" of b_1 and b_2:

$$Se(b_1) = \sqrt{\widehat{Var}(b_1)}$$

$$Se(b_2) = \sqrt{\widehat{Var}(b_2)}$$

- Estimated Variance-Covariance Matrix

$$\begin{bmatrix} \widehat{Var}(b_1) & \widehat{Cov}(b_1, b_2) \\ \widehat{Cov}(b_1, b_2) & \widehat{Var}(b_2) \end{bmatrix}$$

Properties of OLS Estimators

But if we replace σ^2 with $\hat{\sigma}^2$, will "normality" still apply?

- If using σ^2, normality applies:

$$\frac{b_2 - \beta_2}{\sqrt{Var(b_2)}} \sim N(0,1) \tag{26}$$

- If using $\hat{\sigma}^2$, "normality" will no longer apply:

$$\frac{b_2 - \beta_2}{\sqrt{\widehat{Var}(b_1)}} \sim t_{n-2} \tag{27}$$

Properties of OLS Estimators

To evaluate whether OLS estimators are good under some specific cases, we have **Gauss-Markov Theorem**:

Under the assumptions SR1-SR5 of the linear regression model, the OLS estimators b_1 and b_2 are the **Best Linear Unbiased Estimators (BLUE)** for β_1 and β_2. That is, b_1 and b_2 have the smallest variance **within** all linear and unbiased estimators of β_1 and β_2.

Comments:

- The theorem does not say that b_1 and b_2 are the best of all possible estimators;
- When comparing two linear and unbiased estimators, we always want to use the one with the smaller variance, since that estimation rule gives us the higher probability of obtaining an estimate that is close to the true parameter value;
- Note that Gauss Markov Theorem does not depend on SR6 (normality assumption).

Estimation of Nonlinear Relationship

More about nonlinear relationship: the simple linear regression model can be used to account for nonlinear relationship between variables.

1. y and x can be transformations of the basic economic variables, involing logarithms, squares, cubes or reciprocals, and etc.

Example: we want to research how house price is correlated/determined by square feet of the house?

- Originally we may assume:

$$PRICE = \beta_1 + \beta_2 SQFT + e \tag{28}$$

But this may not be consistent with the data pattern and also 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, that is, the slope β_2 may vary from point to point.

Estimation of Nonlinear Relationship

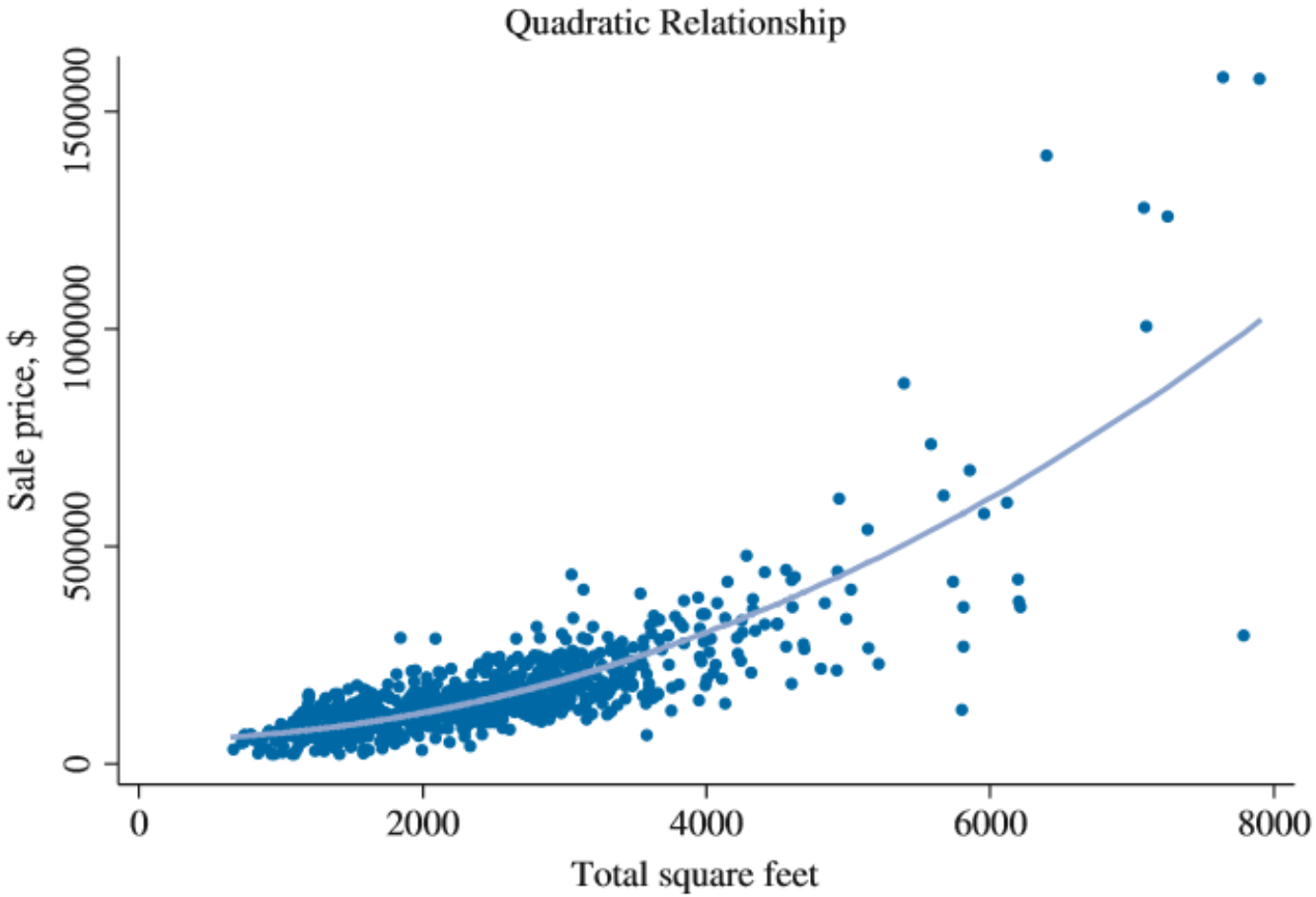

Figure: Quadratic Relationship

Estimation of Nonlinear Relationship

We can build this pattern into our model in two ways:

- a quadratic equation in which the explanatory variable is $SQFT^2$

$$PRICE = \beta_1 + \beta_2 SQFT^2 + e \tag{29}$$

- a log-linear equation in which the dependent variable is $ln(PRICE)$

$$ln(PRICE) = \beta_1 + \beta_2 SQFT + e \tag{30}$$

For (29) and (30), how to interpret β_2, and how to get elasticity of $PRICE$ w.t. $SQFT$ at specific level $(SQFT, PRICE)$?

Estimation of Nonlinear Relationship

If the estimated quadratic equation is:

$$PRICE = 55776.56 + 0.0154SQFT^2 \tag{31}$$

the estimated slope is:

$$\frac{d(PRICE)}{d(SQFT)} = 2 \times 0.0154 \times SQFT \tag{32}$$

the elasticity is:

$$\hat{\eta} = slope \times \frac{SQFT}{PRICE} = 2 \times 0.0154 \times SQFT \times \frac{SQFT}{PRICE} \tag{33}$$

Use one data point ($SQFT = 2000, PRICE = \$117461.77$), the elasticity is 1.05, meaning a 1% in house size will increase house price by 1.05%.

Take care of units of economic variables, in this case, unit of PRICE is \$1, but it may also be $\$1,000$.

Estimation of Nonlinear Relationship

2. y and/or x can be indicator variables that only take values zero and one.

- Indicator variable is usually used to represent qualitative characteristic, such as gender (male or female), race, or location;
- For example

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

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

Estimation of Nonlinear Relationship

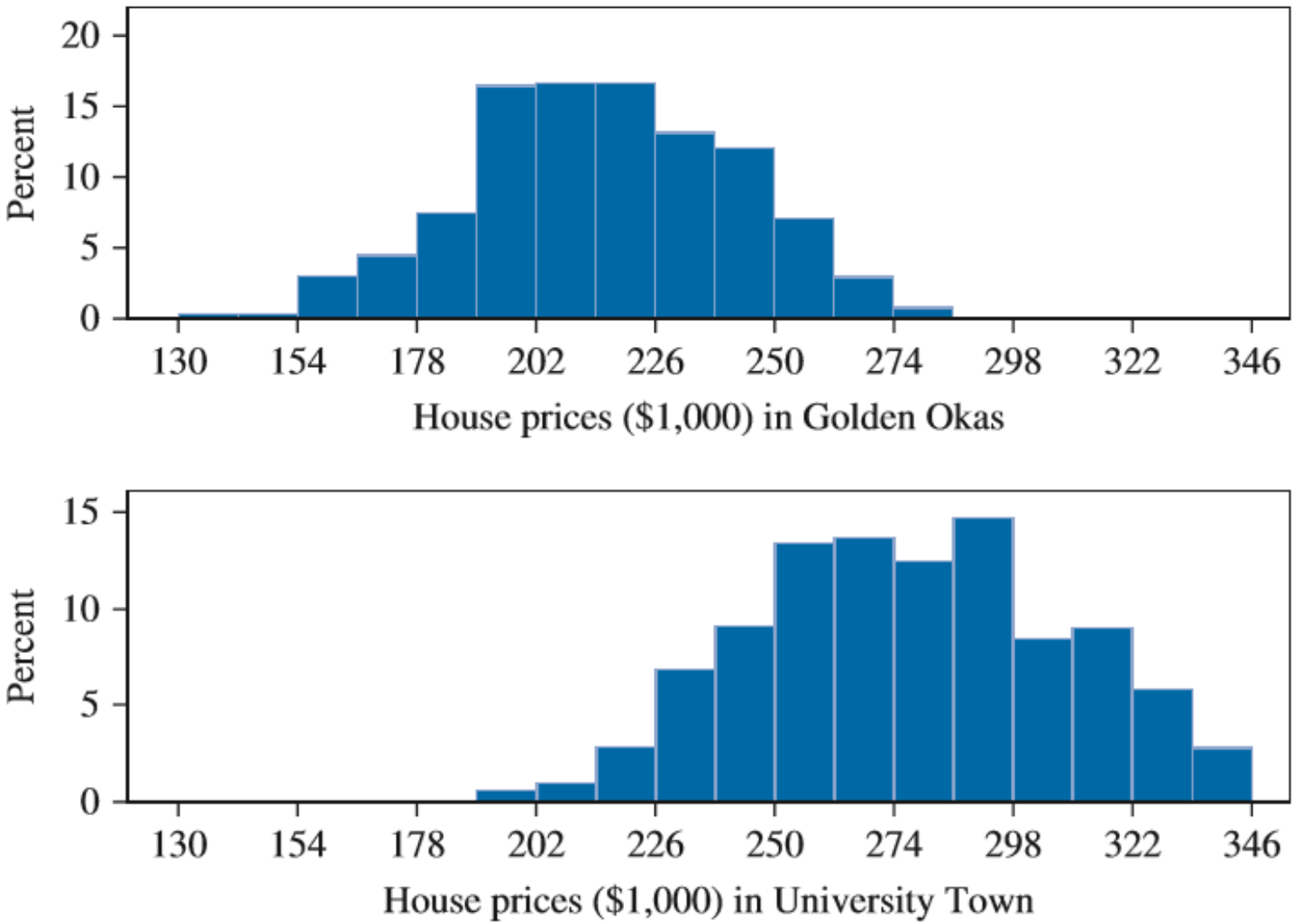

Figure: Distributions of House Prices

Estimation of Nonlinear Relationship

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

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

- The estimated regression is:

$$\begin{aligned} \widehat{PRICE} &= b_1 + b_2 UTOWN \\ &= 215.733 + 61.51 \times UTOWN \\ &= \begin{cases} 277.242 & \text{if } UTOWN = 1 \\ 215.733 & \text{if } UTOWN = 0 \end{cases} \end{aligned}$$

- How to interpret β_2 and b_2:

$$\beta_2 = E[PRICE]_{Universitytown} - E[PRICE]_{Goldenoaks} \tag{34}$$

$$b_2 = \overline{PRICE}_{Universitytown} - \overline{PRICE}_{Goldenoaks} \tag{35}$$