Statistics for Business and decision Making

Dr. Ilaria Benedetti

09. Simple regression model

What is an economic model?

- As economists we are usually more interested in studying relationships between variables
  - Economic theory tells us that expenditure on economic goods depends on income
  - Consequently, we call *y* the ‘‘dependent variable’’ and *x* the independent’’ or ‘‘explanatory’’ variable
  - In **econometrics**, we recognize that real-world expenditures are **random variables**, and we want to use data to learn about the relationship

What is an economic model?

- Suppose that we are interested in studying the relationship between **household income** and **expenditure on food**.

- Consider the ‘‘experiment’’ of randomly selecting households from a particular population. The population might consist of households within a particular city, state, province, or country.

  - Suppose that we are interested only in households with an income of $1,000 per week.
  - In this experiment we randomly select a number of households from this population and interview them. We ask the question, ‘‘***How much did you spend per person on food last week***?’’

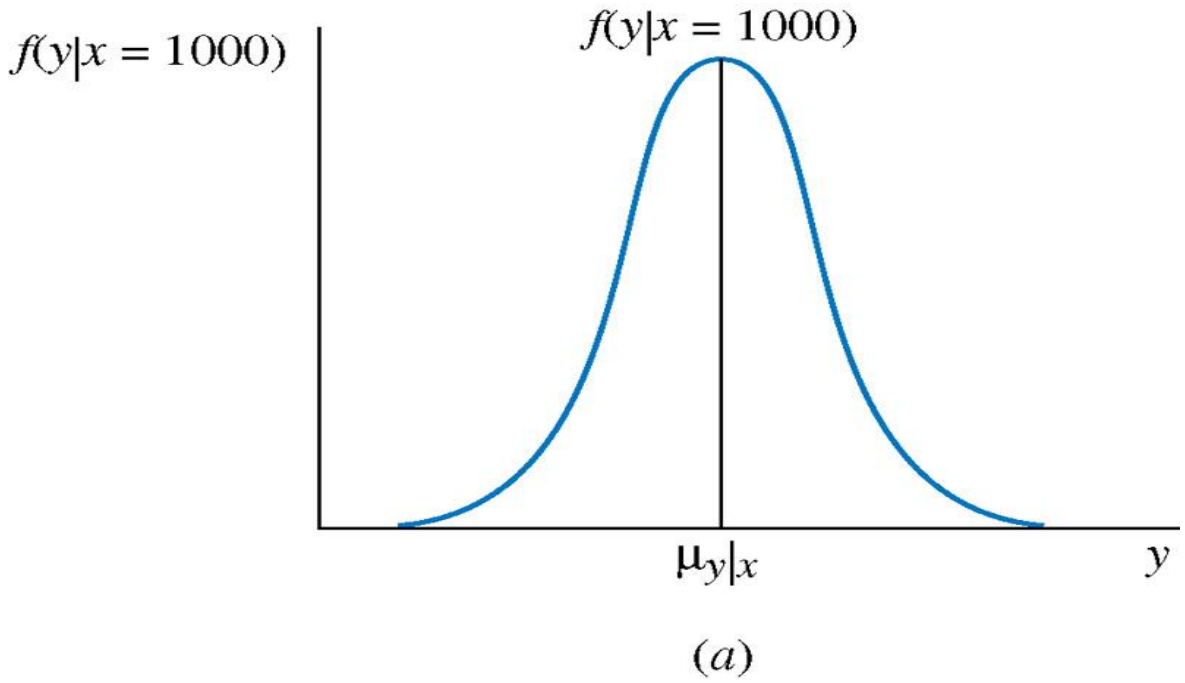
What is an economic model?

- The continuous random variable y (*Weekly food expenditure*) has a probability density function (pdf) that describes the *probabilities* of obtaining various food expenditure values.
- The *pdf* is a conditional probability density function since it is ‘‘conditional’’ upon an *x*

- The **conditional mean**, or expected value, of y is ***E(y|x)***
- The expected value of a random variable is called its ‘‘**mean**’’ value, which is really a contraction of population mean, the center of the probability distribution of the random variable.
- This is not the same as the sample mean, which is the arithmetic average of numerical values.

What is an economic model?

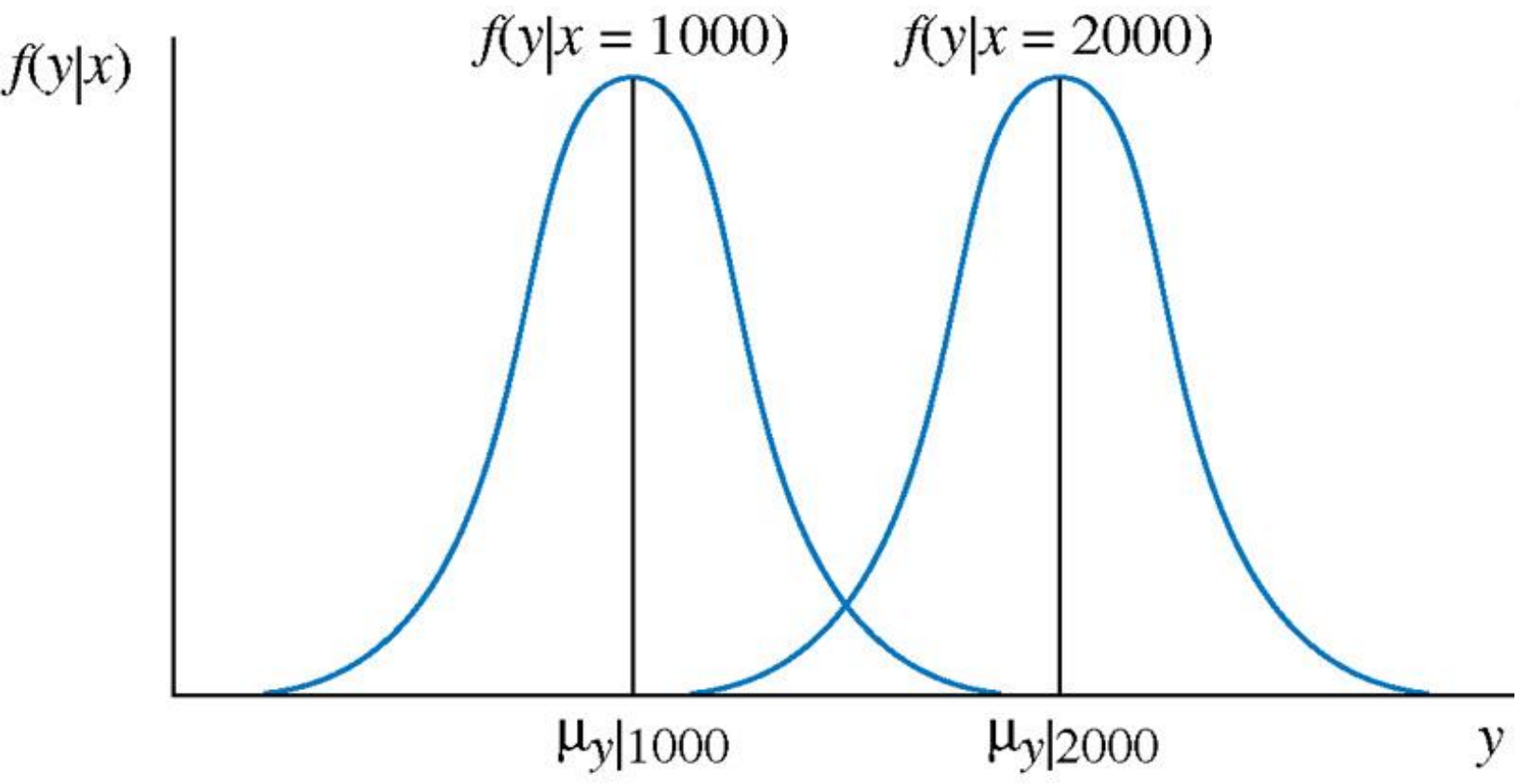
Probability distribution of food expenditure y given income x = $1000



The conditional variance of *y* is σ^2 which measures the dispersion of *y* about its mean $\mu_{y|x}$

The pdf f (y) describes how expenditures are "distributed" over the population

What is an economic model?



(b)

Each conditional pdf shows that expenditures will be distributed about a mean value but the mean expenditure by households with higher income is larger than the mean expenditure by lower income households.

What is an economic model?

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

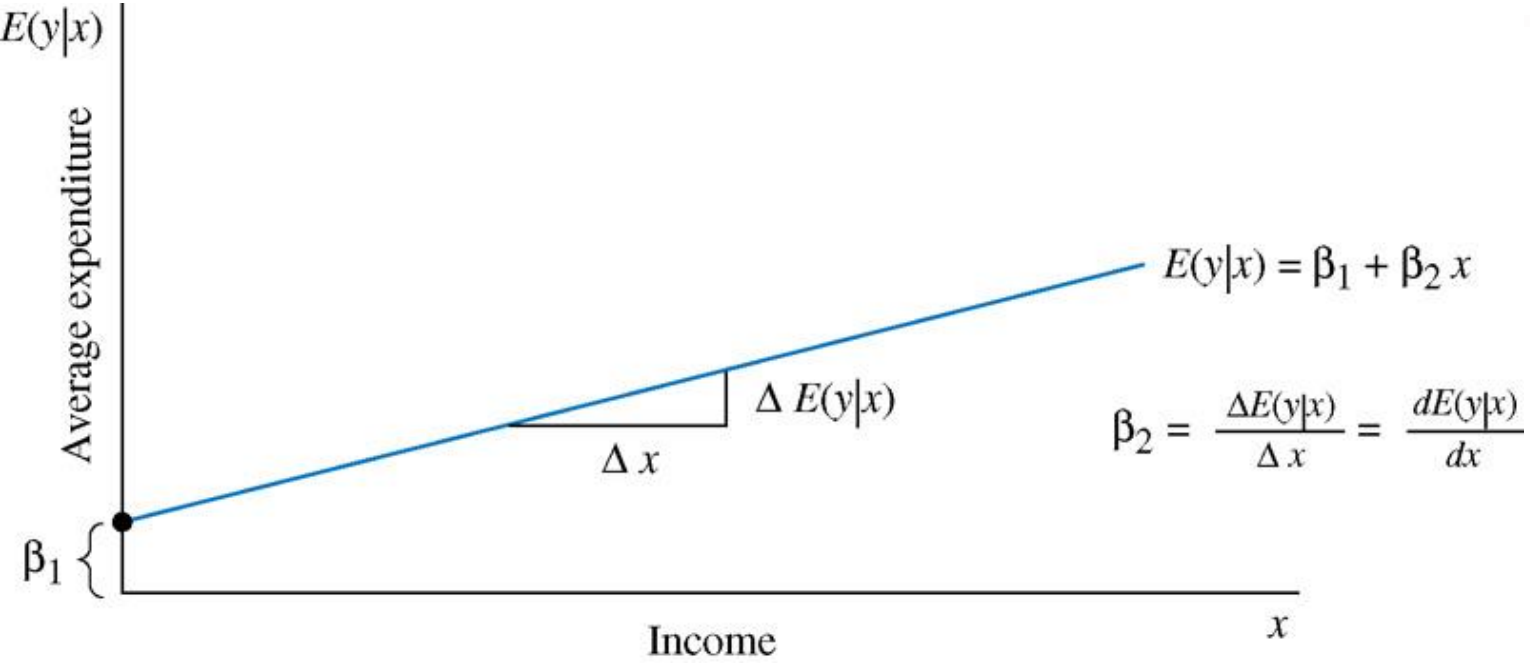
The mathematical representation of our economic model of household food Expenditure is the following:

$$E(y \mid x) = \mu_y = \beta_2 + \beta_2 x$$

where **β_1** is the **intercept** and **β_2** is the **slope**

It is called simple regression not because it is easy, but because there is only one explanatory variable on the right-hand side of the equation.

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

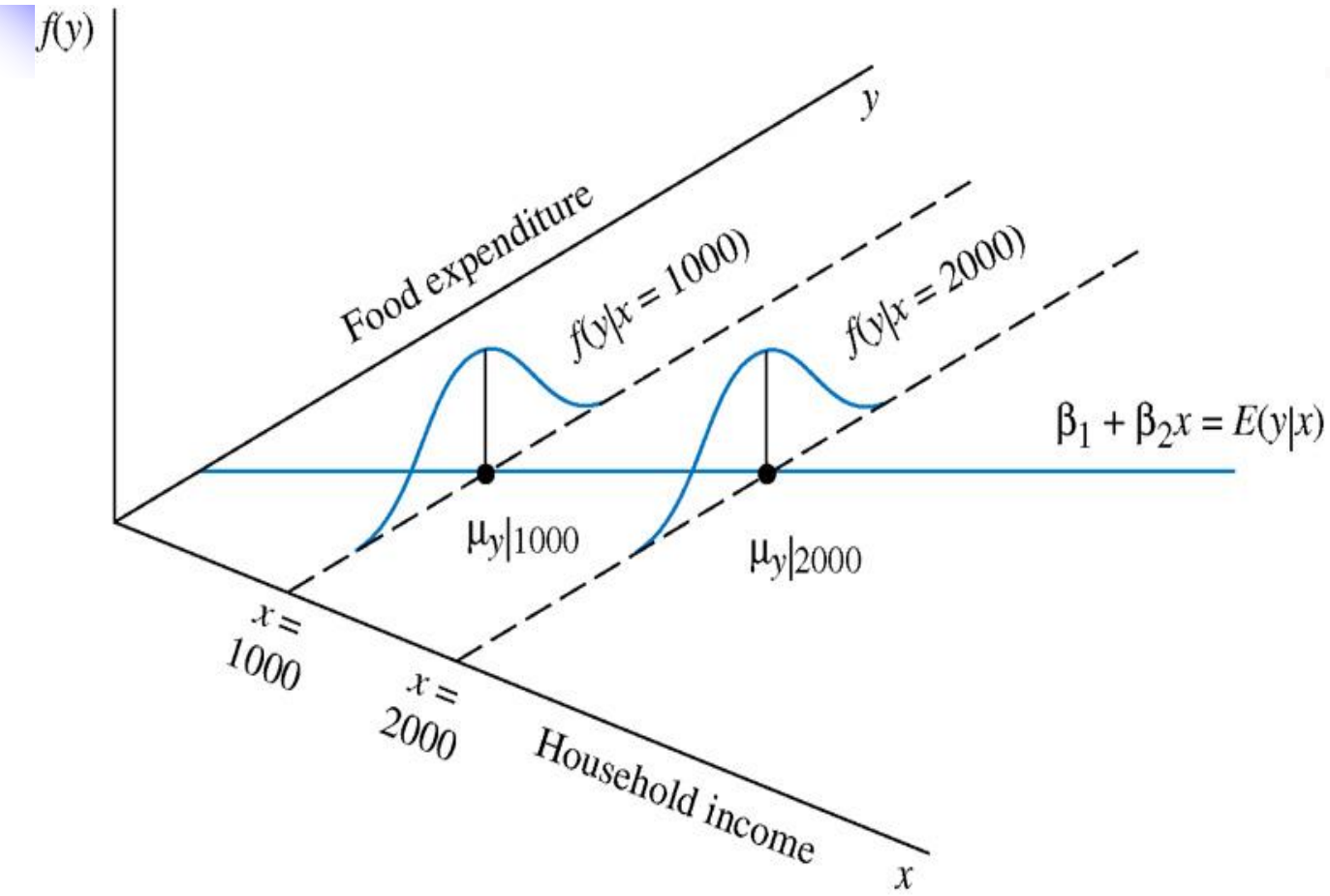


The slope of the regression line can be written as:

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

where "Δ" denotes "change in" and "$dE(y|x)/dx$" denotes the **derivative** of the expected value of y given an x value

An Econometric Model



The probability density functions for y at two levels of income.
This regression function is the foundation of an econometric model for household food expenditure

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL - I

In order to make the econometric model complete we have to make some assumptions

Assumption 1:

The mean value of *y*, for each value of *x*, is given by the *linear regression*

$$E(y \mid x) = \beta_1 + \beta_2 x$$

Assumption 2:

For each value of *x*, the values of *y* are distributed about their mean value, following probability distributions that all have the same variance

$$\text{var}(y \mid x) = \sigma^2$$

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – I

Assumption 3:

The sample values of *y* are all *uncorrelated*, and have zero *covariance*, implying that there is no linear association among them

$$\text{cov}(y_i, y_j) = 0$$

This assumption can be made stronger by assuming that the values of *y* are all statistically independent

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – I

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

Assumption 5:
(*optional*) The values of *y* are *normally distributed* about their mean for each value of *x*

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

THE ERROR TERM

- The essence of regression analysis is that any observation on the dependent variable y can be decomposed into two parts: a systematic component and a random component.

  - The **random error** term is the difference between y and its conditional mean value:

$$e = y - E(y \mid x) = y - \beta_1 - \beta_2 x$$

    - Rearranging gives

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

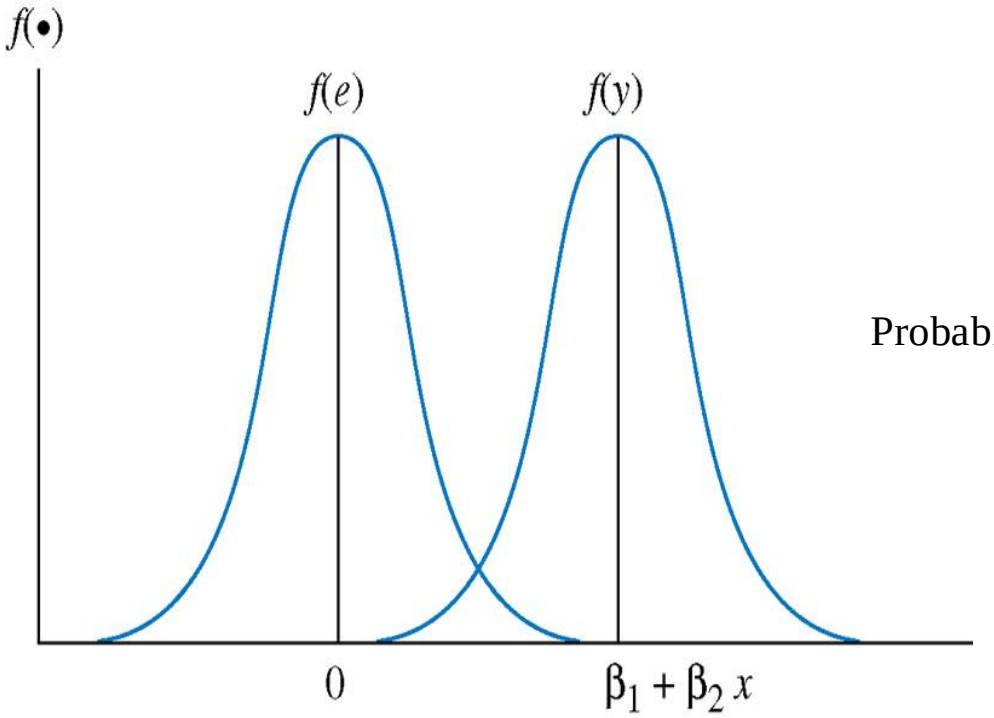
where *y* is the dependent variable and *x* is the independent variable

THE ERROR TERM

- The expected value of the error term, given *x*, is

$$E(e \mid x) = E(y \mid x) - \beta_1 - \beta_2 x = 0$$

The mean value of the error term, given *x*, is zero



Probability density functions for *e* and *y*

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR1:
The value of *y*, for each value of *x*, is: $y = \beta_1 + \beta_2 x + e$

Assumption SR2:
The expected value of the random error *e* is:

$$E(e) = 0$$

This is equivalent to assuming that

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

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR3:

The variance of the random error e is: $\text{var}(e) = \sigma^2 = \text{var}(y)$

- The random variables y and e have the same variance because they differ only by a constant.

Assumption SR4:

The covariance between any pair of random errors, e_i and e_j is:

$$\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$$

The stronger version of this assumption is that the random errors e are statistically independent, in which case the values of the dependent variable y are also statistically independent

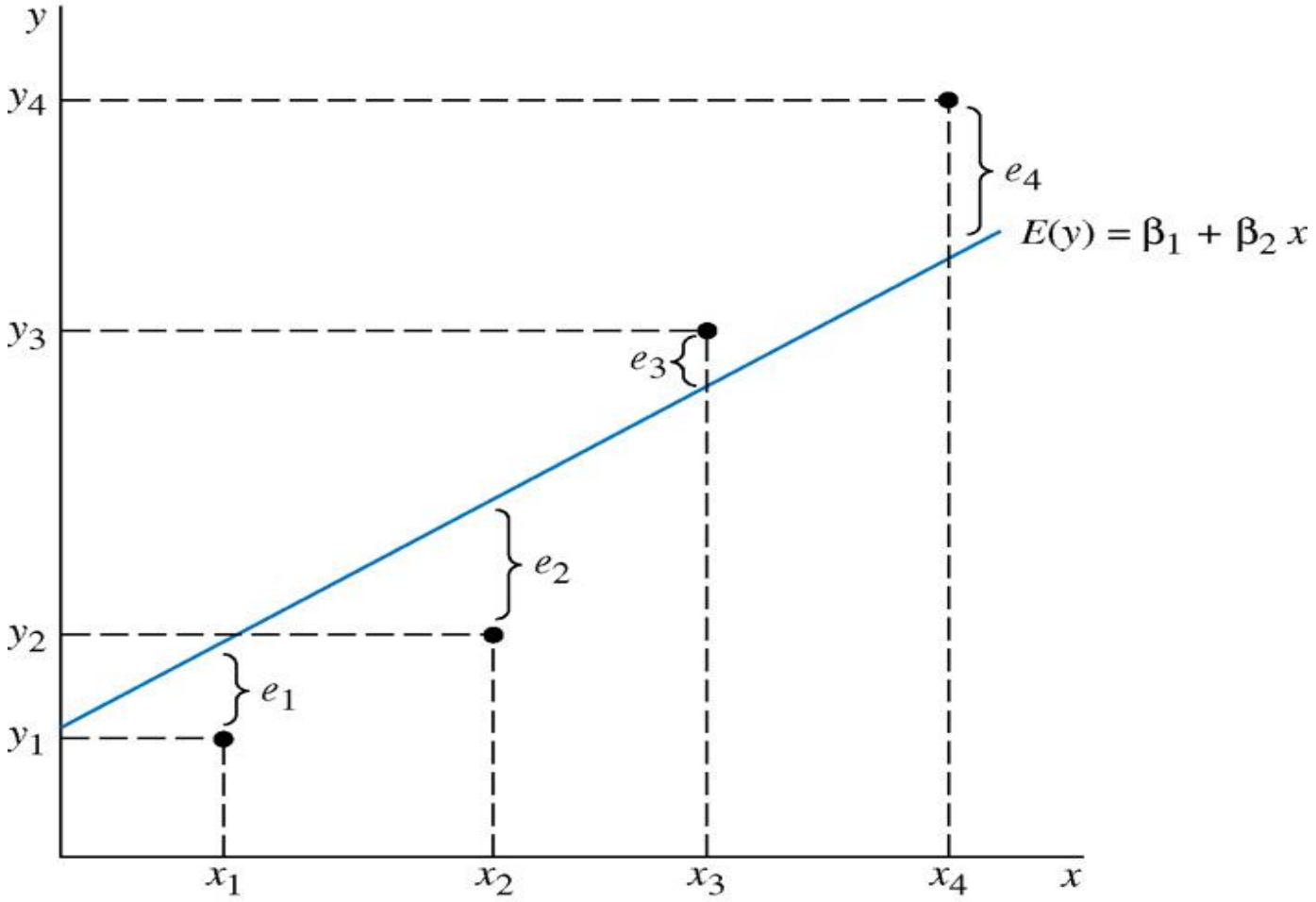
ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL – II

Assumption SR5:
The variable *x* is not random, and must take at least two different values

Assumption SR6:
(*optional*) The values of *e* are *normally distributed* about their mean if the values of *y* are normally distributed, and *vice versa*

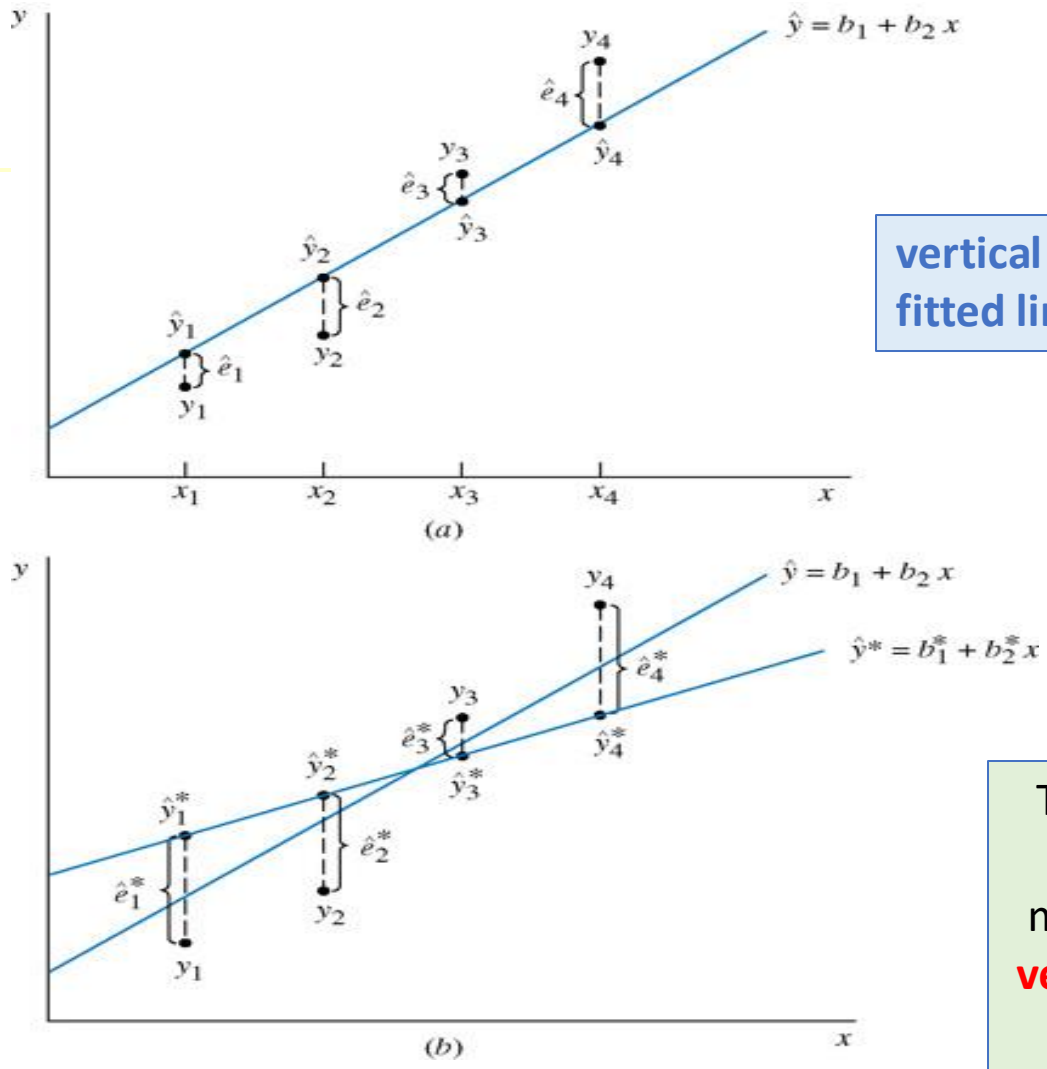
$$e \sim N(0, \sigma^2)$$

The relationship among *y*, *e* and the true regression line



The random error e and the dependent variable y are both random variables, and the properties of one can be determined from the properties of the other.
There is, however, one interesting difference between them: y is ‘‘observable’’ and e is ‘‘unobservable.’

THE LEAST SQUARES PRINCIPLE



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

vertical distances from each point to the fitted line are the least squares residuals

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

The LS principle asserts that to fit a line to the data values we should make **the sum of the squares of the vertical distances from each point to the line as small as possible.**

THE LEAST SQUARES PRINCIPLE

- Suppose we have another fitted line:

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

The least squares line has the smaller sum of squared residuals:

$$\text{if } SSE = \sum_{i=1}^{N} \hat{e}_i^2 \text{ and } SSE^* = \sum_{i=1}^{N} \hat{e}_i^{*2} \text{ then } SSE < SSE^*$$

The problem is to find b1 and b2 in a convenient way.

↓

- Least squares estimates for the unknown parameters β_1 and β_2 are obtained my **minimizing the sum of squares function**:

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

Plot: Food Expenditure and Income Data

scatter diagram

Dataset: food.csv

R- code:

plot(x=food$income, y=food$food_exp, main="Food_exp ~ Income")

Plot: Food Expenditure and Income Data

Dataset: food.csv

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

For illustration we examine typical data on household food expenditure and weekly income from a random sample of 40 households

THE LEAST SQUARES ESTIMATORS

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

(Slope)

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

(Intercept)

Example: Food Expenditure and Income Data

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

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

A convenient way to report the values for b_1 and b_2 is to write out the *estimated* or *fitted* regression line:

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

THE LEAST SQUARES ESTIMATORS

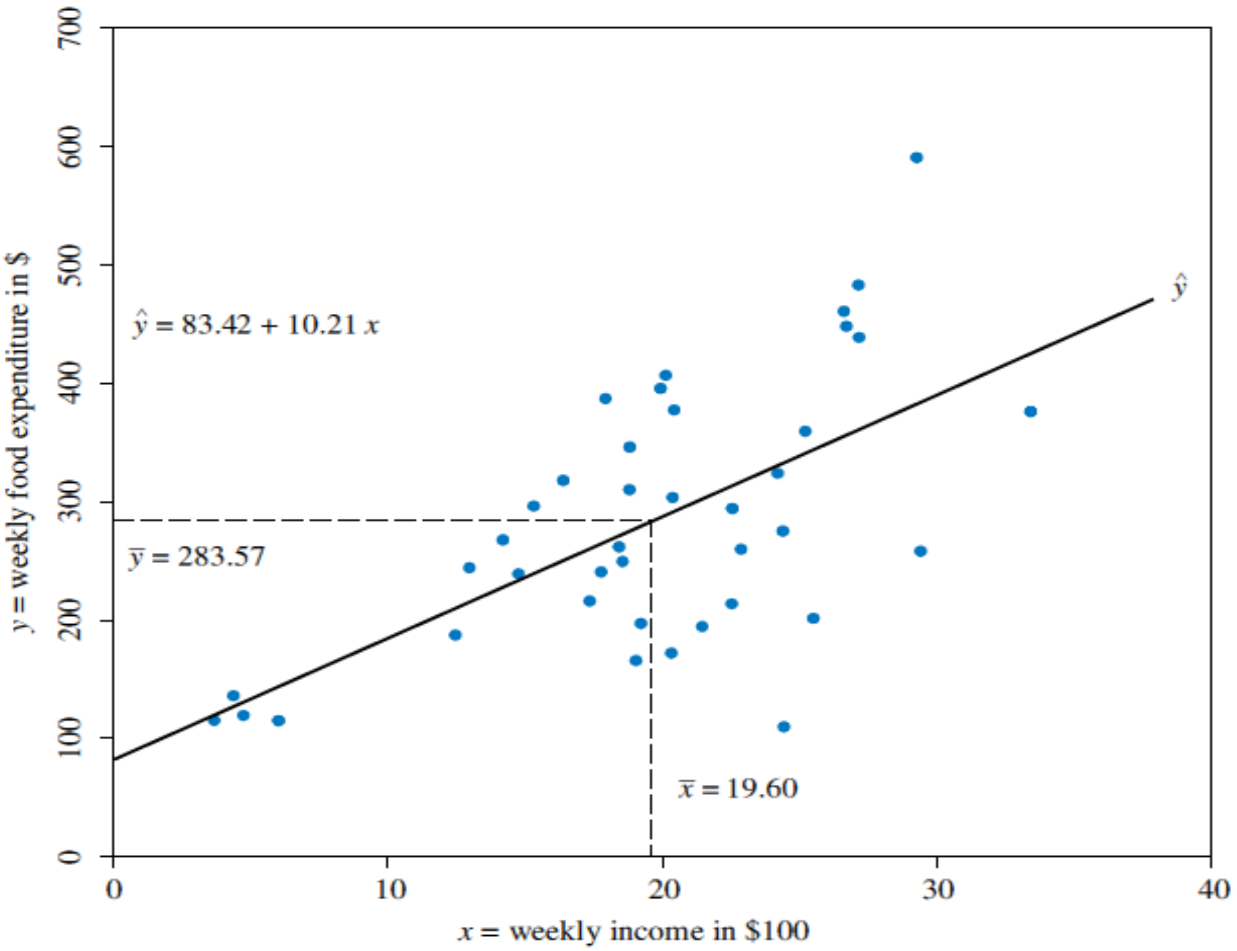
$$b_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

```
> summary(linearMod) # see all the ouput computed by R

Call:
lm(formula = food_exp ~ income, data = food)

Residuals:
     Min       1Q   Median       3Q      Max
-223.025  -50.816   -6.324   67.879  212.044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   83.416     43.410   1.922   0.0622 .
income        10.210      2.093   4.877 1.95e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05
```

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

Example: The fitted regression line



The fitted line based on the least squares parameter estimates passes through the point defined by the sample

$$(\bar{x}; \bar{y}) = (19.60; 283.57)$$

Interpreting the Estimates

The value $b_2 = 10.21$ is an estimate of β_2, the amount by which weekly expenditure on food per household increases when household weekly income increases by \$100.

Thus, we estimate that if income goes up by \$100, expected weekly expenditure on food will increase by approximately \$10.21

Strictly speaking, the intercept estimate $b_1 = 83.42$ is an estimate of the weekly food expenditure on food for a household with zero income.

Elasticities

Income elasticity is a useful way to characterize the responsiveness of consumer expenditure to changes in income.

The elasticity of a variable *y* with respect to another variable *x* is:

$$\varepsilon = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{\Delta y}{\Delta x}\frac{x}{y}$$

Since in the linear economic model is: $\beta_2 = \frac{\Delta E(y)}{\Delta x}$

The ***elasticity of mean expenditure with respect to income*** is:

$$\varepsilon = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x}\frac{x}{E(y)} = \beta_2 \frac{x}{E(y)}$$

Elasticities

To estimate this elasticity we replace β_2 by $\mathbf{b_2 = 10.21}$. We must also replace "x" and "**E(y)**" by something, since in a linear model the elasticity is different on each point upon the regression line.

- A frequently used alternative is to calculate the elasticity at the "***point of the means***" because it is a representative point on the regression line.

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

- This estimated income elasticity takes its usual interpretation.
- We estimate that a 1% increase in weekly household income will lead, on average, to a **0.71% increase in weekly household expenditure on food**, when x and y take their sample mean values, (x; y) = (**19:60; 283:57**).
- Since the estimated income elasticity is less than one, we would classify food as a "**necessity**" rather than a "luxury," which is consistent with what we would expect for an average household.

Prediction

Suppose that we wanted to predict weekly food expenditure for a household with a weekly income of $2000. This prediction is carried out by substituting **x = 20** into our estimated equation to 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

EXERCISE TO DO

A soda vendor at Louisiana State University football games observes that the warmer the temperature at game time the greater the number of sodas that are sold. Based on 32 home games covering five years, the vendor estimates the relationship between soda sales and temperature to be $\hat{y} = -240 + 20x$

where y = the number of sodas she sells and

x = temperature in degrees Fahrenheit.

A. Interpret the estimated slope and intercept. Do the estimates make sense? Why or why not?
B. On a day when the temperature at game time is forecast to be 80◦F, predict how many sodas the vendor will sell.
C. Below what temperature are the predicted sales zero?

Exercise to do after the class

You have the results of a simple linear regression based on state-level data and the District of Columbia, a total of $N = 51$ observations.

(a) The estimated error variance $\hat{\sigma}^2 = 2.04672$. What is the sum of the squared least squares residuals?

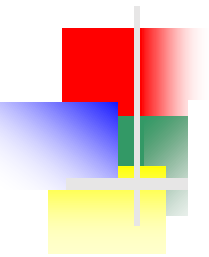
(b) The estimated variance of b_2 is 0.00098. What is the standard error of b_2? What is the value of $\sum(x_i - \overline{x})^2$?

(c) Suppose the dependent variable y_i = the state's mean income (in thousands of dollars) of males who are 18 years of age or older and x_i the percentage of males 18 years or older who are high school graduates. If $b_2 = 0.18$, interpret this result.

(d) Suppose $\overline{x} = 69.139$ and $\overline{y} = 15.187$, what is the estimate of the intercept parameter?

(e) Given the results in (b) and (d), what is $\sum x_i^2$?

(f) For the state of Arkansas the value of $y_i = 12.274$ and the value of $x_i = 58.3$. Compute the least squares residual for Arkansas. (*Hint*: Use the information in parts (c) and (d).).



Assessing the Least Squares Fit

Assessing the Least Squares Fit

Using the food expenditure data we have estimated the parameters of the regression model using the least squares formulas and obtained the least squares estimates .

It is natural, but misguided, to ask the question ‘‘**How good are these estimates?**’’

This question is not answerable since we will never know the true values of the population parameters

Assessing the Least Squares Fit

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?

Assessing the Least Squares Fit

If our model assumptions hold, then $E(b_2) = \beta_2$, which means that the estimator is **unbiased**.

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.

The Variances and Covariances of b_1 and b_2

If the regression model assumptions SR1-SR5 are correct (assumption SR6 is not required), then the variances and covariance of b_1 and b_2 are:

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

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

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

The Variances and Covariances of b_1 and b_2

1. The *larger* the variance term σ^2 , 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 is the sample mean $\bar{x}$, and the covariance has a *sign* opposite to that of $\bar{x}$.

GAUSS-MARKOV THEOREM

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

1. The estimators b_1 and b_2 are "best" when compared to 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**, since 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.

GAUSS-MARKOV THEOREM

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

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) then the least squares estimators are normally distributed:

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

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

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

Estimating the Variance of the Error Term

The variance of the random error e_i is:

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

if the assumption $E(e_i) = 0$ is correct.

Since 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 Variance of the Error Term

The least squares residuals are obtained by replacing the unknown parameters by their least squares estimates:

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

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

There is a simple modification that produces an unbiased estimator, and that is:

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

so that:

$$E\left(\hat{\sigma}^2\right) = \sigma^2$$

Estimating the Variance of the Error Term

Replace the unknown error variance σ^2 by $\hat{\sigma}^2$ to obtain:

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

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

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

Estimating the Variance of the Error Term: food example

x	y	$\hat{y}$	$\hat{e} = y - \hat{y}$
3.69	115.22	121.09	−5.87
4.39	135.98	128.24	7.74
4.75	119.34	131.91	−12.57
6.03	114.96	144.98	−30.02
12.47	187.05	210.73	−23.68

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{304505.2}{38} = 8013.29$$

- 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:

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

The estimator variance, var(b_2), or its square root, $\sigma_{b_2} = \sqrt{\text{var}(b_2)}$ 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

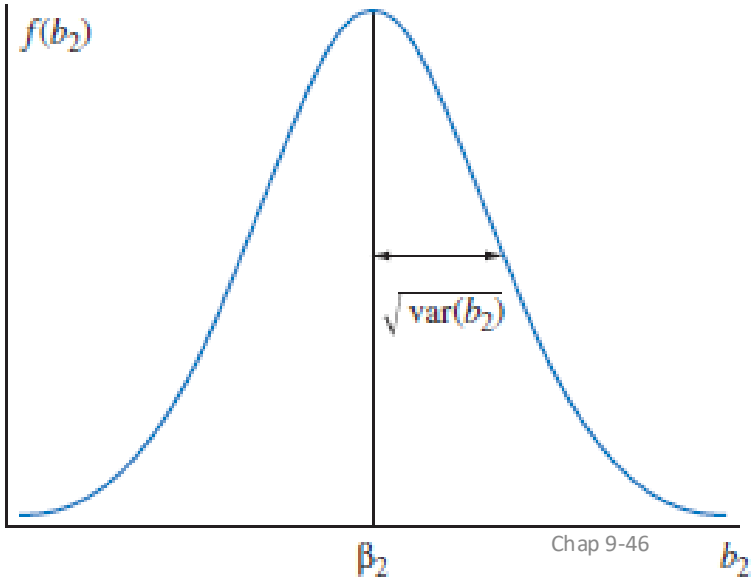
The question we address with the standard error is

"How much variation about their means do the estimates exhibit from sample to sample?"

We estimate σ^2, and then estimate σ_{b_2} using:

$$se(b_2) = \sqrt{var(b_2)} = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$$

The standard error of b_2 is thus an estimate of what the standard deviation of many estimates b_2 would be in a very large number of samples, and is an indicator of the width of the *pdf* of b_2 shown in Fig.



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, because 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.

Including these possibilities means the simple linear regression model can be used to account for nonlinear relationships between variables

Consider the linear model of house prices:

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

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 **SQFT²**
2. a ***log-linear*** equation in which the dependent variable is ***ln(PRICE)***

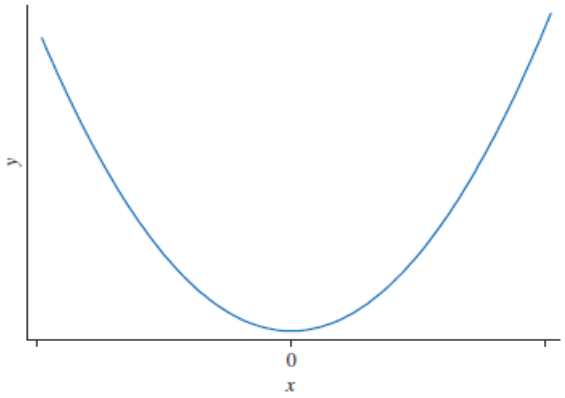
In each case we will find that the slope of the relationship between ***PRICE*** **and *SQFT* is not constant**, but changes from point to point.

$y = \beta_1 + \beta_2 x^2$ | $PRICE = \beta_1 + \beta_2 SQFT^2 + e$

The quadratic function is (parabola). The y-intercept is $\beta 1$.

The shape of the curve is determined by $\beta 2$:

- if $\beta 2 > 0$, then the curve is U-shaped;
- if $\beta 2 < 0$, then the curve has an inverted-U shape.



The **slope** of the function is given by the derivative dy/dx = 2bx, which changes as x changes.

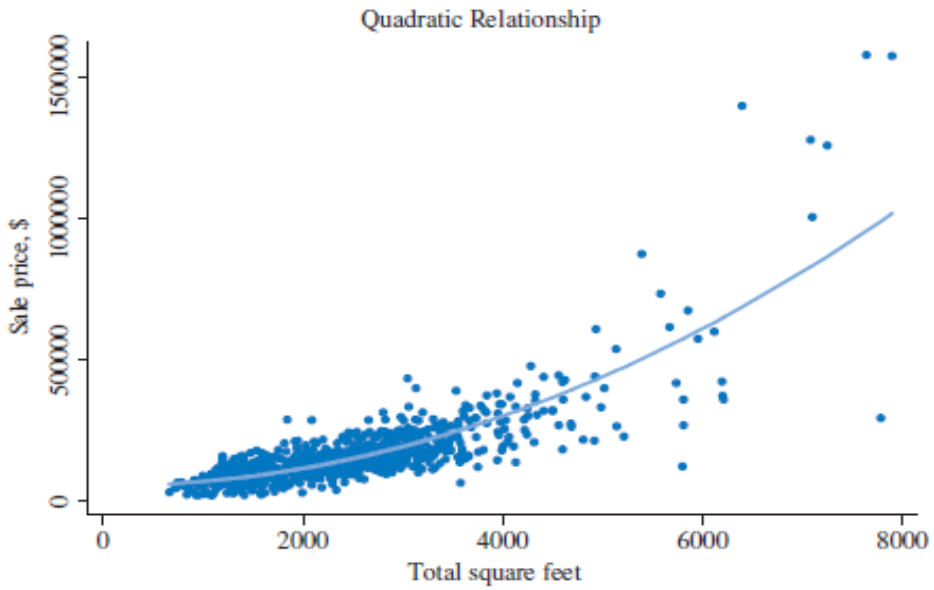
The elasticity, or the percentage change in *y* given a 1% change in *x*, is:

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

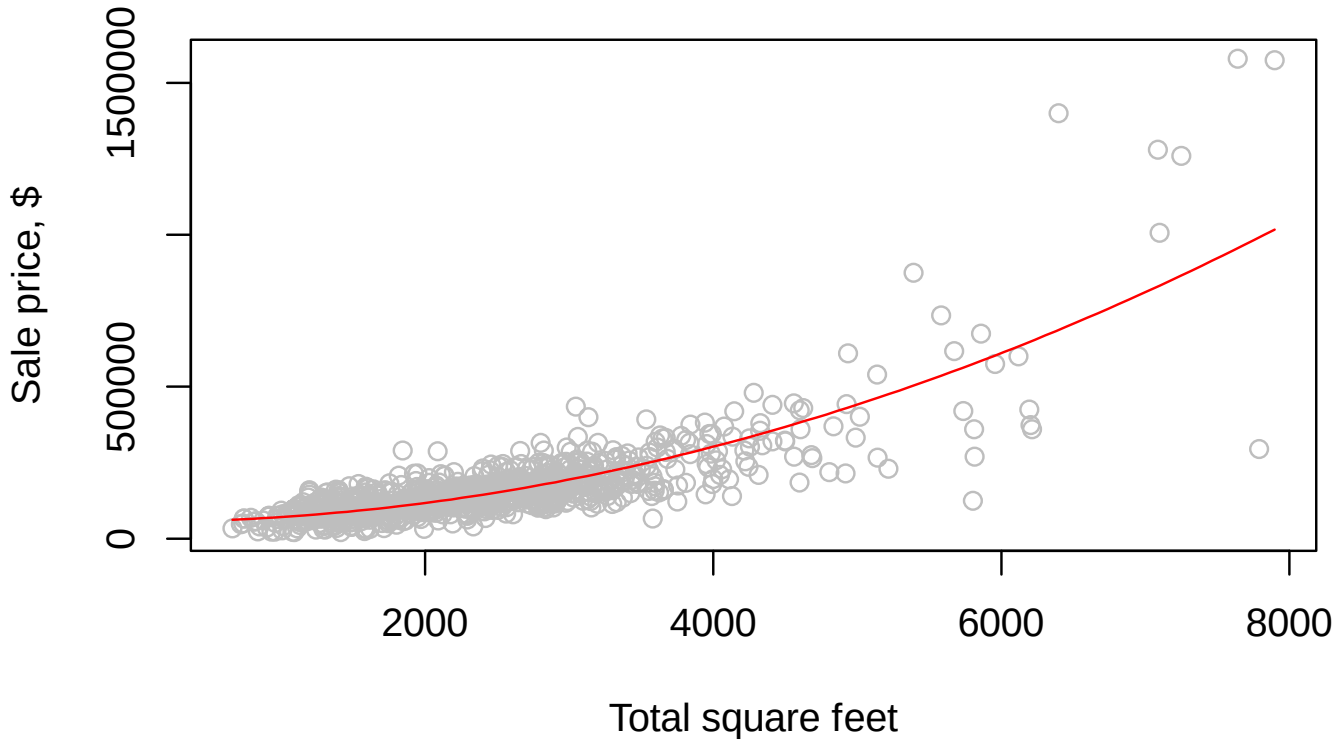
$$PRICE = \beta_1 + \beta_2 SQFT^2 + e$$

The slope is: $$\frac{d(\widehat{PRICE})}{dSQFT} = 2\hat{\beta}_2 SQFT$$

If $\hat{\beta}_2 > 0$, then larger houses will have larger slope, and a larger estimated price per additional square foot



The file br.dat contains data on 1080 houses sold in Baton Rouge, LA during mid-2005.




```
#plot
plot(br$SQFT, br$PRICE, xlab="Total square feet",
    ylab="Sale price, $", col="grey") # xlim=c(0,8000)
#add the quadratic curve to the scatter plot:
curve(b1+b2*x^2, col="red", add=TRUE)
```

EXAMPLE

For 1080 houses sold in Baton Rouge, LA during mid-2005, the estimated quadratic equation is:

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

The estimated **slope** is:

$$\frac{d(\widehat{PRICE})}{dSQFT} = 2\hat{\beta}_2 SQFT \quad \Rightarrow \quad \widehat{slope} = 2(0.0154)SQFT$$

The **elasticity** is:

$$\varepsilon = slope \times \frac{SQFT}{PRICE} \quad \Rightarrow \quad \varepsilon = 2\hat{\beta}_2 SQFT \times \frac{SQFT}{PRICE}$$

The elasticity of house price with respect to house size is the percentage increase in estimated price given a 1% increase in house size. Like the slope, the elasticity changes at each point.

To compute an estimate, we must select values for *SQFT* and *PRICE*

- A common approach is to choose a point on the fitted relationship
  - That is, we choose a value for *SQFT* and choose for price the corresponding fitted value

For houses of 2000, 4000 and 6000 square feet, the estimated prices are:

$$\hat{Y} = 55776.56 + 0.0154(2000)^2 = 117{,}461.77\ \$$$

$$\hat{Y} = 55776.56 + 0.0154(4000)^2 = 302{,}517.39\ \$$$

$$\hat{Y} = 55776.56 + 0.0154(6000)^2 = 610{,}943.42\ \$$$

To compute an estimate, we must select values for *SQFT* and *PRICE*

- A common approach is to choose a point on the fitted relationship
  - That is, we choose a value for *SQFT* and choose for price the corresponding fitted value

For houses of 2000, 4000 and 6000 square feet, the estimated elasticities are:

1.05 using $\hat{PRICE} = \$117{,}461.77$
1.63 using $\hat{PRICE} = \$302{,}517.39$
1.82 using $\hat{PRICE} = \$610{,}943.42$

Interpretation: For a 2000-square-foot house, we estimate that a 1% increase in house size will increase price by 1.05%

```
125 sqftx=c(2000, 4000, 6000) #given values for sqft
126 pricex=b1+b2*sqftx^2 #prices corresponding to given sqft
127 DpriceDsqft <- 2*b2*sqftx # marginal effect of sqft on price
128 elasticity=DpriceDsqft*sqftx/pricex
129 b1; b2; DpriceDsqft; elasticity #prints results
130
```

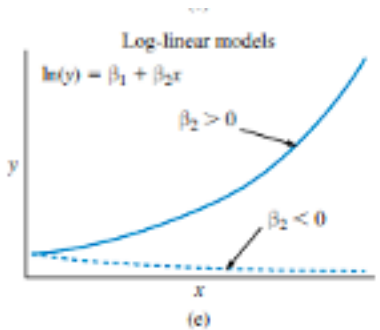
```
31:35  # (Untitled)
```

```
onsole  Terminal ×  Jobs ×
/Dropbox/CORSO SAPIENZA/SBDM 2021_2022/dataset/
] 55776.57
] 0.0154213
]  61.68521 123.37041 185.05562
] 1.050303 1.631251 1.817408
```

Interpretation: For a 2000-square-foot house, we estimate that a 1% increase in house size will increase price by 1.05%

Log – linear model

The log-linear model, $\ln(y) = \beta_1 + \beta_2 x$, has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side



β_2/y	Slope $y > 0$
$\beta_2 > 0$	Increasing marginal effect
$\beta_2 < 0$	function decreases at decreasing rates

Both i y change at each point and are the same sign as β_2.

Using the ... $\exp[\ln(y)] = \exp[\beta 1 + \beta 2]$

In the log-linear model, a one-unit increase in *x* leads, approximately, to a 100 β_2% change in *y*

A 1-unit increase in x leads approximately, to a 100xβ_2% change in y

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

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 the slope expression, we can solve for a **semi-elasticity**, which tells us the percentage change in y given a 1-unit increase in x:*

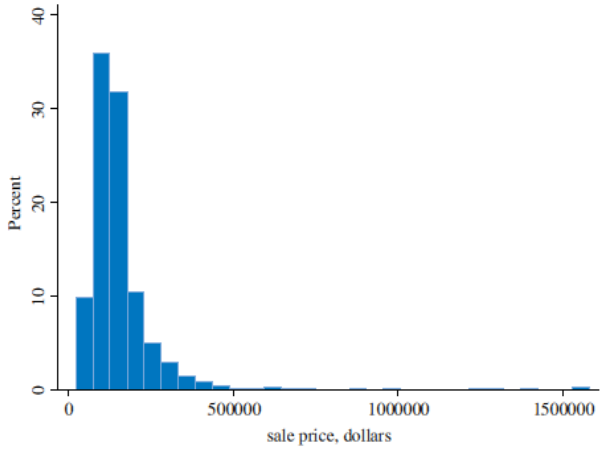
$$\eta = \frac{100(dy/dx)}{dx} = 100b$$

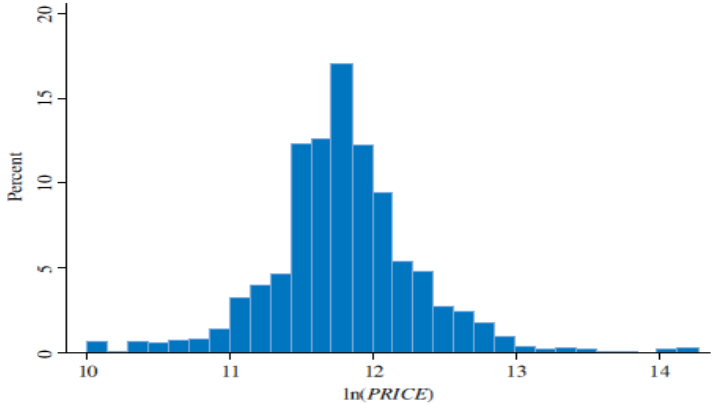
EXAMPLE

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

$$\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$$

This logarithmic transformation can regularize data that is skewed with a long tail to the right





The extremely skewed distribution of PRICE becomes more symmetric, if not bell-shaped, after taking the logarithm.

Many economic variables, including prices, incomes, and wages, have skewed distributions, and the use of logarithms in models for such variables is common.

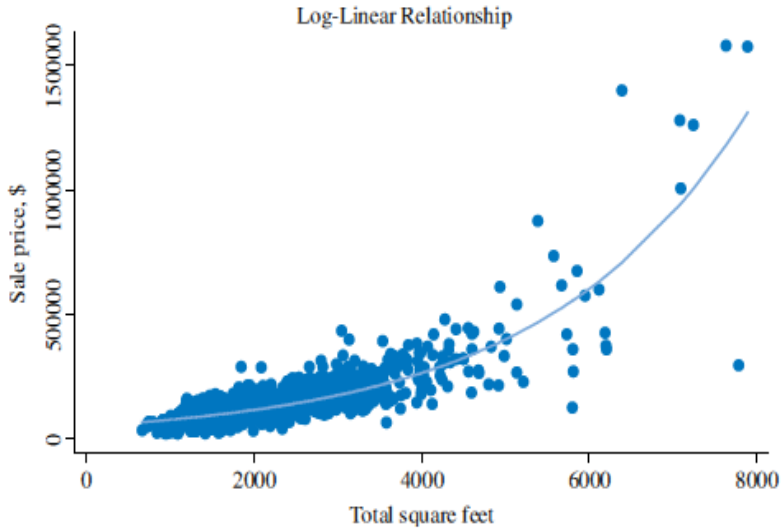
Using the Baton Rouge data, the fitted log-linear model is:

$$\ln(\widehat{PRICE}) = 10.8386 + 0.0004113SQFT$$

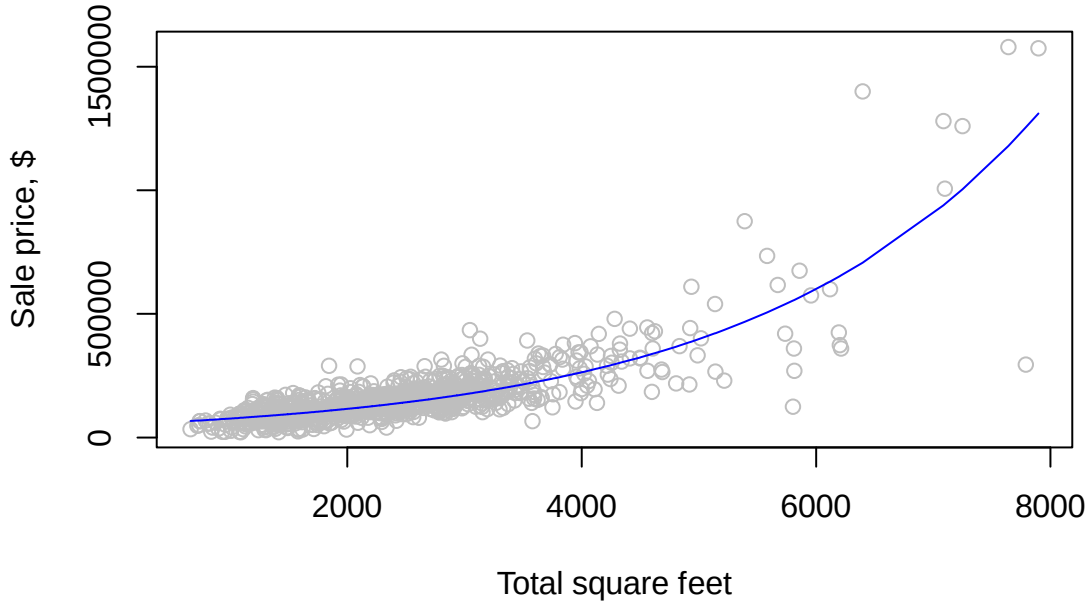
To obtain predicted price take the **anti-logarithm**, which is the exponential function:

$$\widehat{PRICE} = \exp\left[\ln(\widehat{PRICE})\right] = \exp(10.8386 + 0.0004113SQFT)$$

$$\widehat{PRICE} = \exp(10.8386 + 0.0004113 * (2000) = \mathbf{115,975.5\$}$$



```
# graphical representation
ordat <- br[order(br$SQFT), ] #order the dataset
mod4 <- lm(log(PRICE)~SQFT, data=ordat)
plot(br$SQFT, br$PRICE, xlab="Total square feet",
    ylab="Sale price, $", col="grey") # xlim=c(0,8000)
lines(exp(fitted(mod4))~ordat$SQFT, col="blue", main="Log-linear Model")
```



The slope of the log-linear model is:

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = \hat{\gamma}_2 \ \widehat{PRICE} = 0.0004113 \ \widehat{PRICE}$$

$$(\widehat{\gamma 2}) = \textbf{Slope= b2*y}$$

0.0004113*100000=41.13

0.0004113*500000=205.63

Interpretation of the coefficient:

For a house with a predicted *PRICE* of $100,000, the estimated increase in *PRICE* for an additional square foot of house area is ***$41.13***, and

for a house with a predicted *PRICE* of $500,000, the estimated increase in *PRICE* for an additional square foot of house area is ***$205.63***

- The estimated elasticity is:

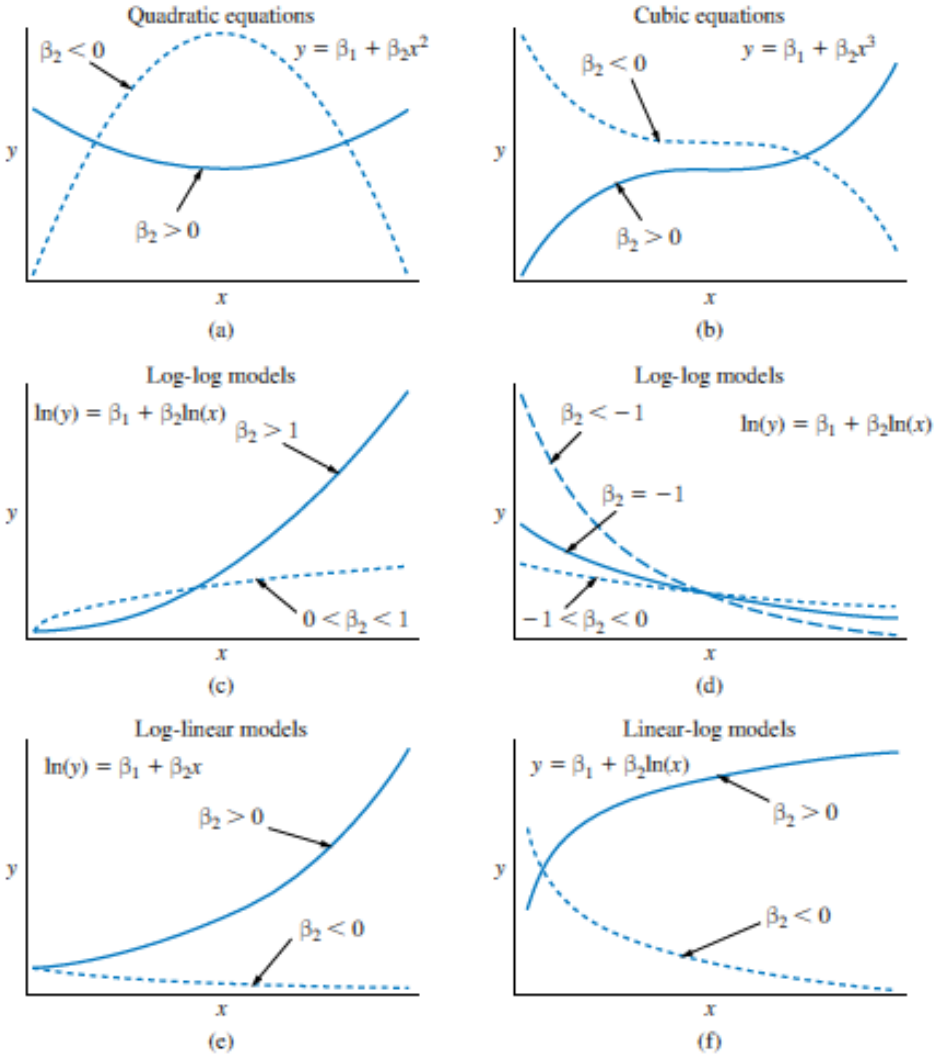
$\hat{\varepsilon} = b2 * x$

$$\hat{\varepsilon} = \hat{\gamma}_2 SQFT = 0.0004113 SQFT$$

- For a house with 2000-square-feet, the estimated elasticity is 0.823:
  - ***A 1% increase in house size is estimated to increase selling price by 0.823%***
- For a house with 4000 square feet, the estimated elasticity is 1.645:
  - ***A 1% increase in house size is estimated to increase selling price by 1.645%***

  - Using the ‘‘semi-elasticity’’ we can say that, for a one-square-foot increase in size, we estimate a price increase of 0.04%
  - ***Or, perhaps more usefully, we estimate that a 100-square-foot increase will increase price by approximately 4%.***

Choosing functional form



Choosing functional form

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	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1 unit change in x leads to (approximately) a 100 β_2% change in y		
Linear-Log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

We should do our best to choose a functional form that is:

- consistent with economic theory
- that fits the data well
- that is such that the assumptions of the regression model are satisfied

- ✓ In real-world problems it is sometimes difficult to achieve all these goals;
- ✓ Furthermore, we will never truly know the correct functional relationship, no matter how many years we study econometrics
- ✓ The truth is out there, but we will never know it
- ✓ In applications of econometrics we must simply do the best we can to choose a satisfactory functional form

Regression with Indicator Variables

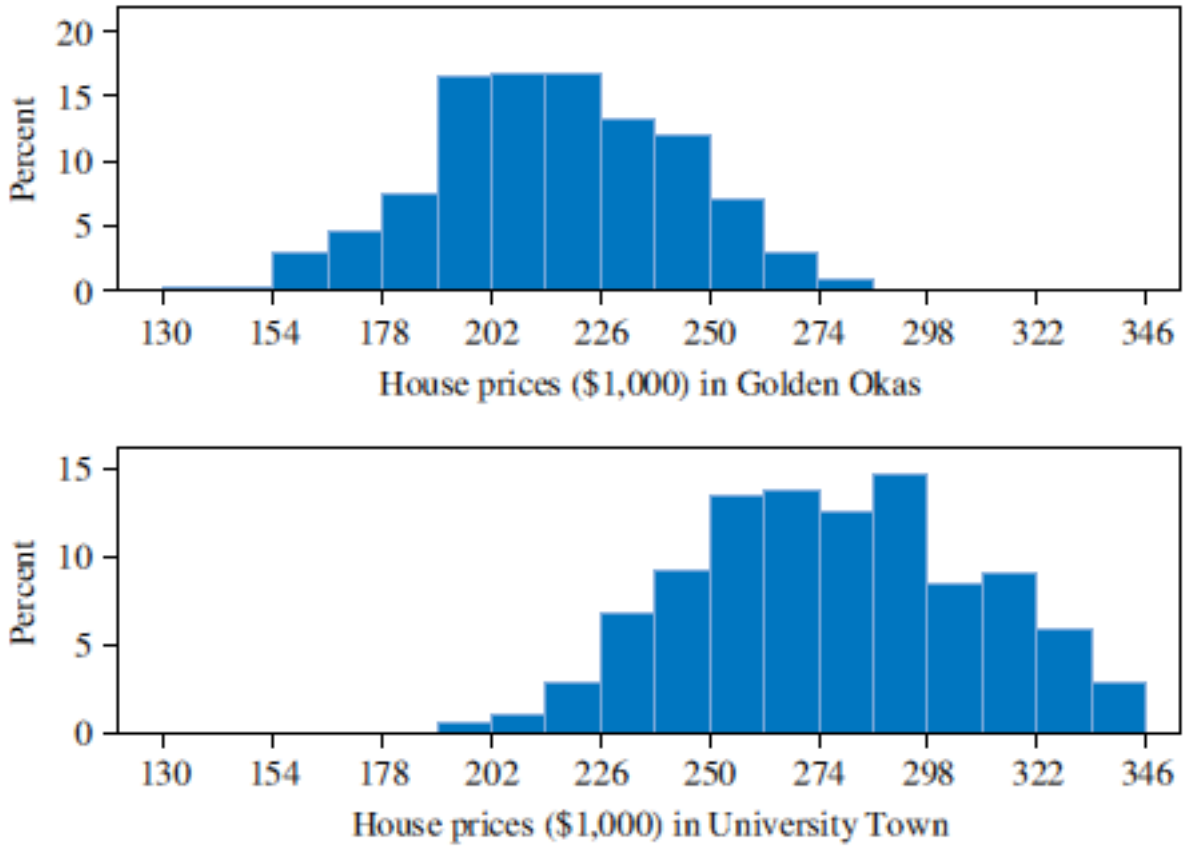
An indicator variable is a binary variable that takes the values zero or one; it is used to represent a nonquantitative characteristic, such as gender, race, or location…

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

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

How do we model this?

Distributions of the house prices



The mean of the distribution of house prices in University Town appears to be larger than the mean of the distribution of house prices from Golden Oaks.

The sample mean of the 519 house prices in University Town is 277.2416, whereas the sample mean of the 481 Golden Oaks houses is 215.7325.

If we include UTOWN in a regression model as an explanatory variable, what do we have?

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

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

- The estimated regression is:

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

- The least squares estimators b_1 and b_2 in this indicator variable regression can be shown to be:

$$b_1 = \overline{PRICE}_{\text{Golden Oaks}}$$

$$b_2 = \overline{PRICE}_{\text{University Town}} - \overline{PRICE}_{\text{Golden Oaks}}$$

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

How do we account for location, which is a qualitative variable?

- Indicator variables are used to account for qualitative factors in econometric models
- They are often called **dummy, binary or dichotomous** variables, because they take just two values, usually one or zero, to indicate the presence or absence of a characteristic or to indicate whether a condition is true or false
- They are also called **dummy variables**, to indicate that we are creating a numeric variable for a qualitative, non-numeric characteristic
- We use the terms indicator variable and dummy variable interchangeably.

Generally, we define an indicator variable D as:

$$D = \begin{cases} 1 & \text{if characteristic is present} \\ 0 & \text{if characteristic is not present} \end{cases}$$

- So, to account for location, a qualitative variable, we would have:

$$D = \begin{cases} 1 & \text{if property is in the desirable neighborhood} \\ 0 & \text{if property is not in the desirable neighborhood} \end{cases}$$

Adding our indicator variable to our model:

$$PRICE = \beta_1 + \delta D + e$$

If our model is correctly specified, then:

$$E(PRICE) = \begin{cases} (\beta_1 + \delta) & when\ D = 1 \\ \beta_1 & when\ D = 0 \end{cases}$$

Adding an indicator variable causes a **parallel shift** in the relationship by the amount **δ**

An indicator variable like D that is incorporated into a regression model to capture a shift in the intercept as the result of some qualitative factor is called an intercept indicator variable, or an intercept dummy variable

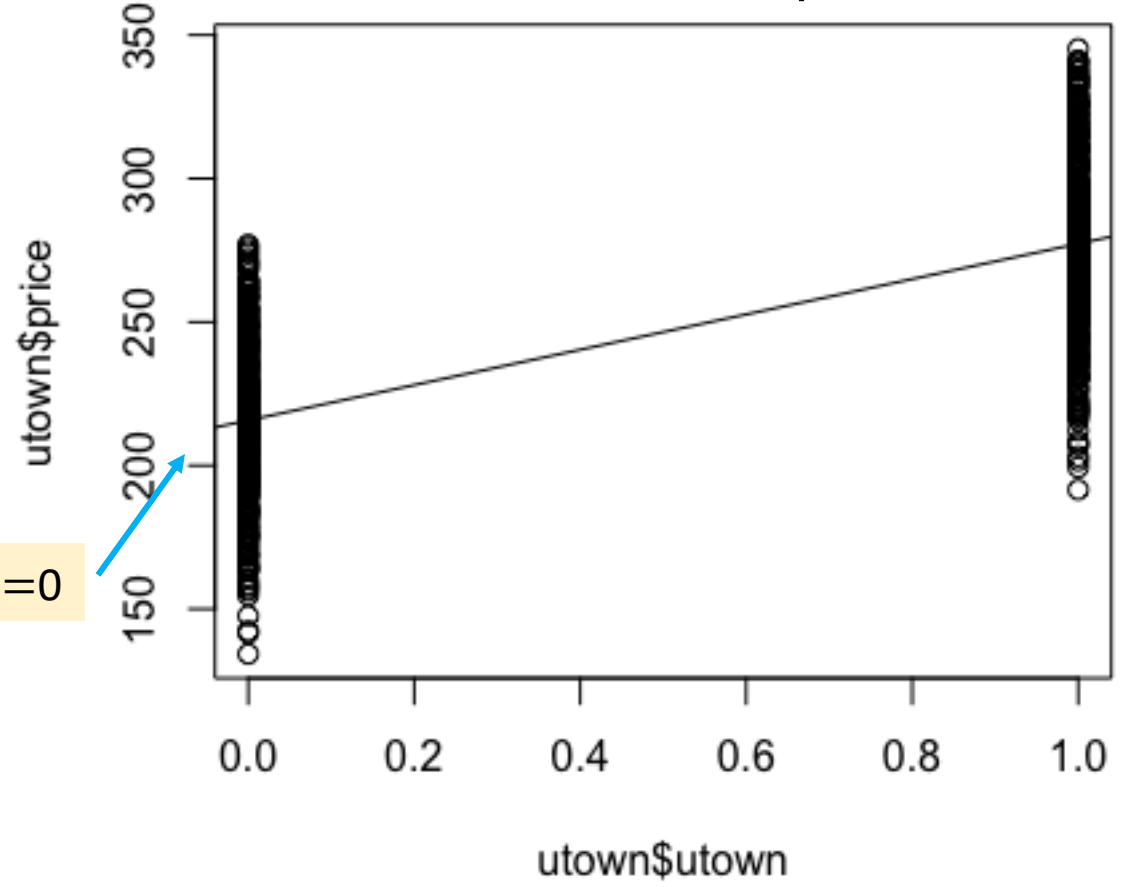
The least squares estimator's properties are not affected by the fact that one of the explanatory variables consists only of zeros and ones

D is treated as any other explanatory variable.

We can construct an interval estimate for D, or we can test the significance of its least squares estimate

$$PRICE = \beta_1 + \delta D + e$$

$$PRICE = 215.73 + 61.51D + e$$



$\beta_1(+\delta)$ $when\ D = 1$

$\beta_1\ when\ D$ =0

The value **D = 0** defines the **reference group**, or **base group**

The value **D = 0** defines the **reference group**, or **base group**

We could pick any base. For example:

$$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$

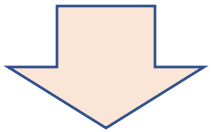
Suppose we included both *D* and *LD*:

$$PRICE = \beta_1 + \delta D + \lambda LD + e$$

The variables *D* and *LD* are such that ***D + LD = 1***

Since the intercept variable $x_1 = 1$, we have created a model with ***exact collinearity***

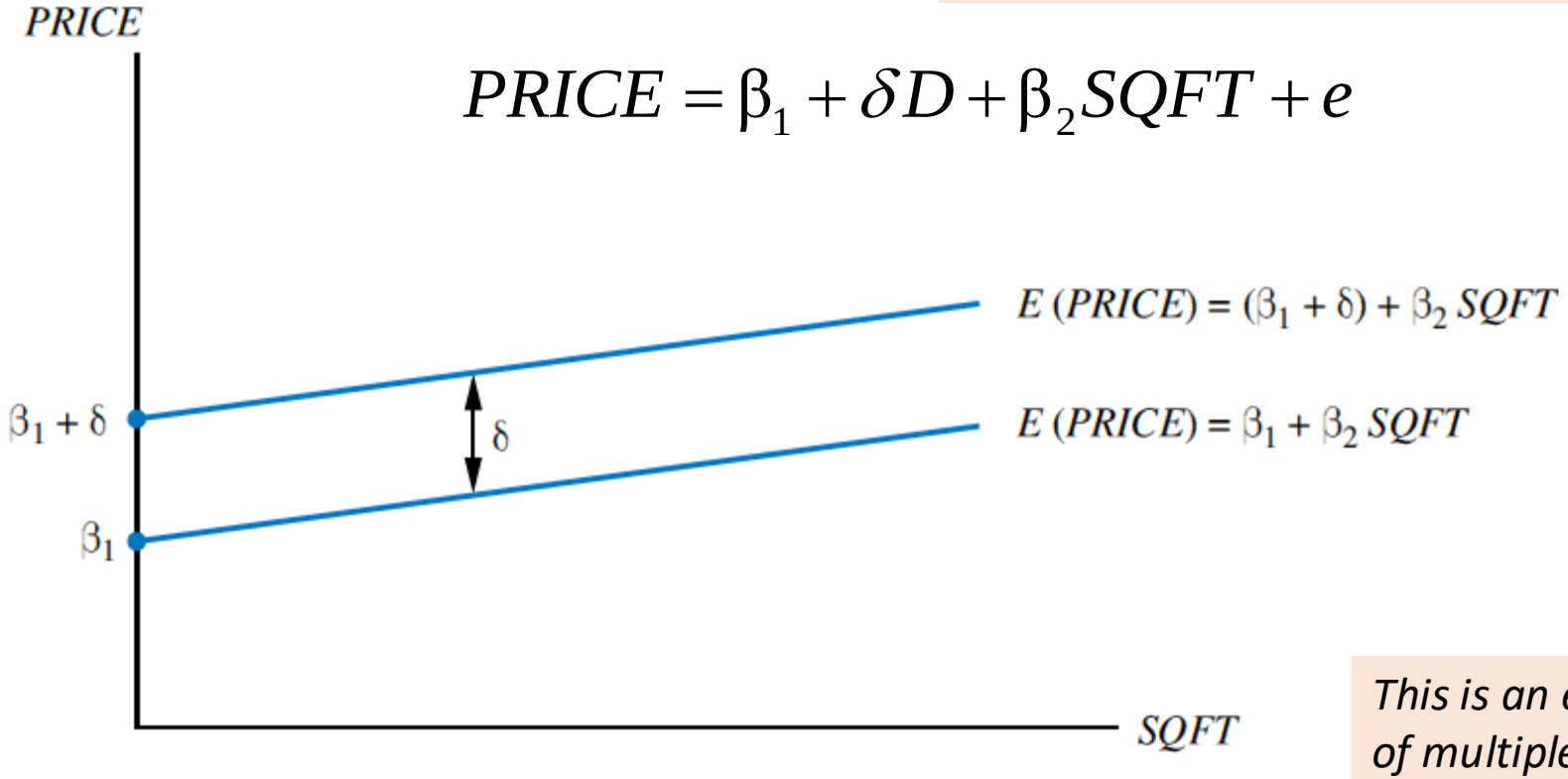
We have fallen into the ***dummy variable trap***.



By including only one of the indicator variables the omitted variable defines the reference group and we avoid the problem

Graphical representation of dummy variable

$$PRICE = \beta_1 + \delta D + \beta_2 SQFT + e$$



This is an example of multiple regression….

The value **D = 0** defines the **reference group**, or **base group**

We could pick any base. For example:

$$LD = \begin{cases} 1 & \text{if property is not in the desirable neighborhood} \\ 0 & \text{if property is in the desirable neighborhood} \end{cases}$$