The Simple Linear Regression Model: Specification and Estimation

- Starting point: Real-world expenditures and income are **random variables**, and we want to build an econometric model and use data to learn about the relationship between them.

- Every econometric model starts with a theory:
  - For example: Economic theory tells us that expenditure on economic goods depends on income
  - Consequently *expenditure* is the "dependent variable" and *income* the independent" or "explanatory" variable.

- To the economic model we will add the random error, to form the econometric model that is the basis for a quantitative or empirical economic analysis

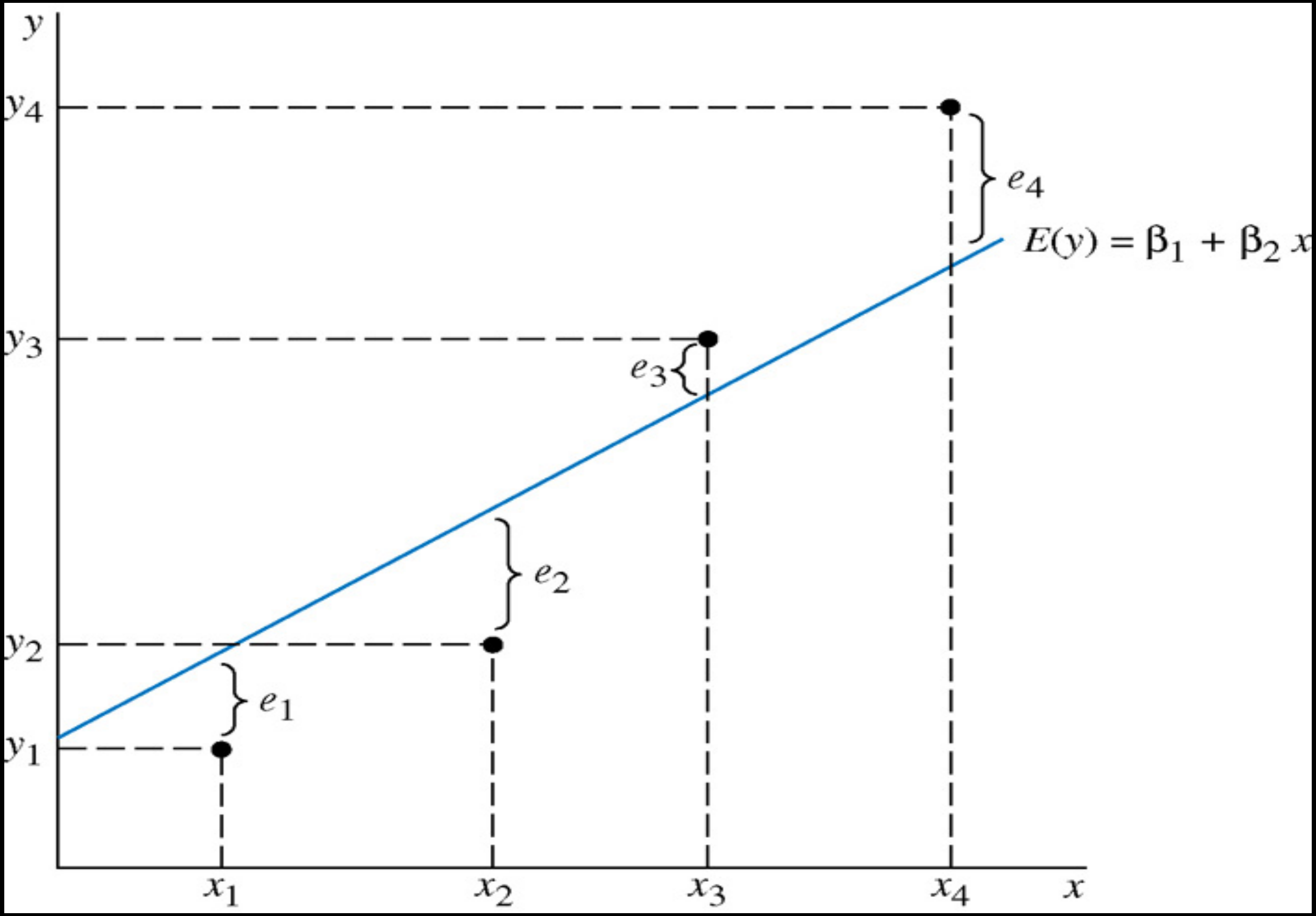
- This most common econometric model is the **Linear Regression Model**

An Econometric Model

- The **simple linear regression** model is a linear model with only one explanatory variable:

where β_1 is the intercept and β_2 is the slope and e is an error term.

The relationship among y, e and the true regression line



- There are several key assumptions underlying the simple linear regression

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

Assumption 1 (*just for convenience*):

The variable x is not random

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

Assumption 2:

The expected value of the random error e is:

This is equivalent to assuming that

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

Assumption 3:

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

The stronger version of this assumption is that the random errors e are statistically independent, (so the values of the dependent variable y are also statistically independent).

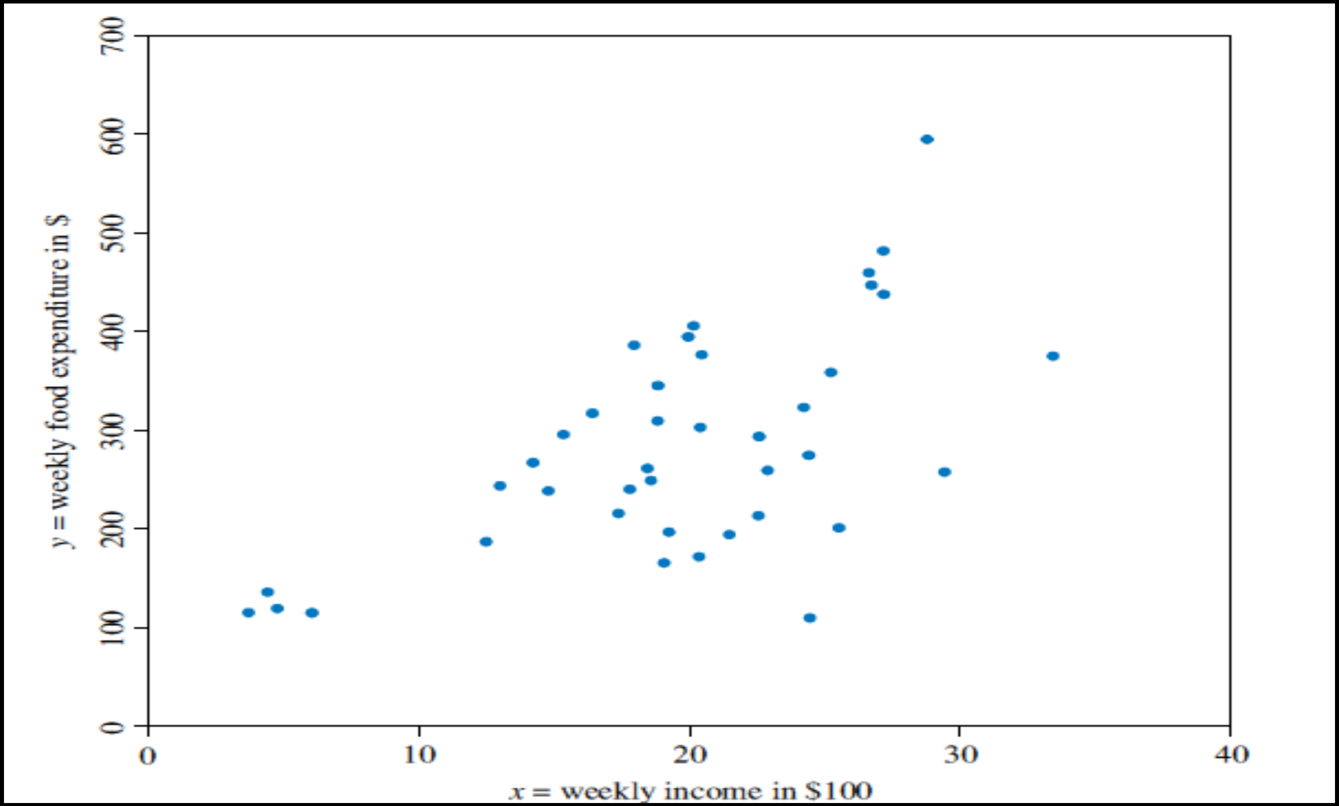
ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

Assumption 4:
The variance of the random error e is:

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

Data on food expenditure and income

Is this a good assumption for our food expenditure model?

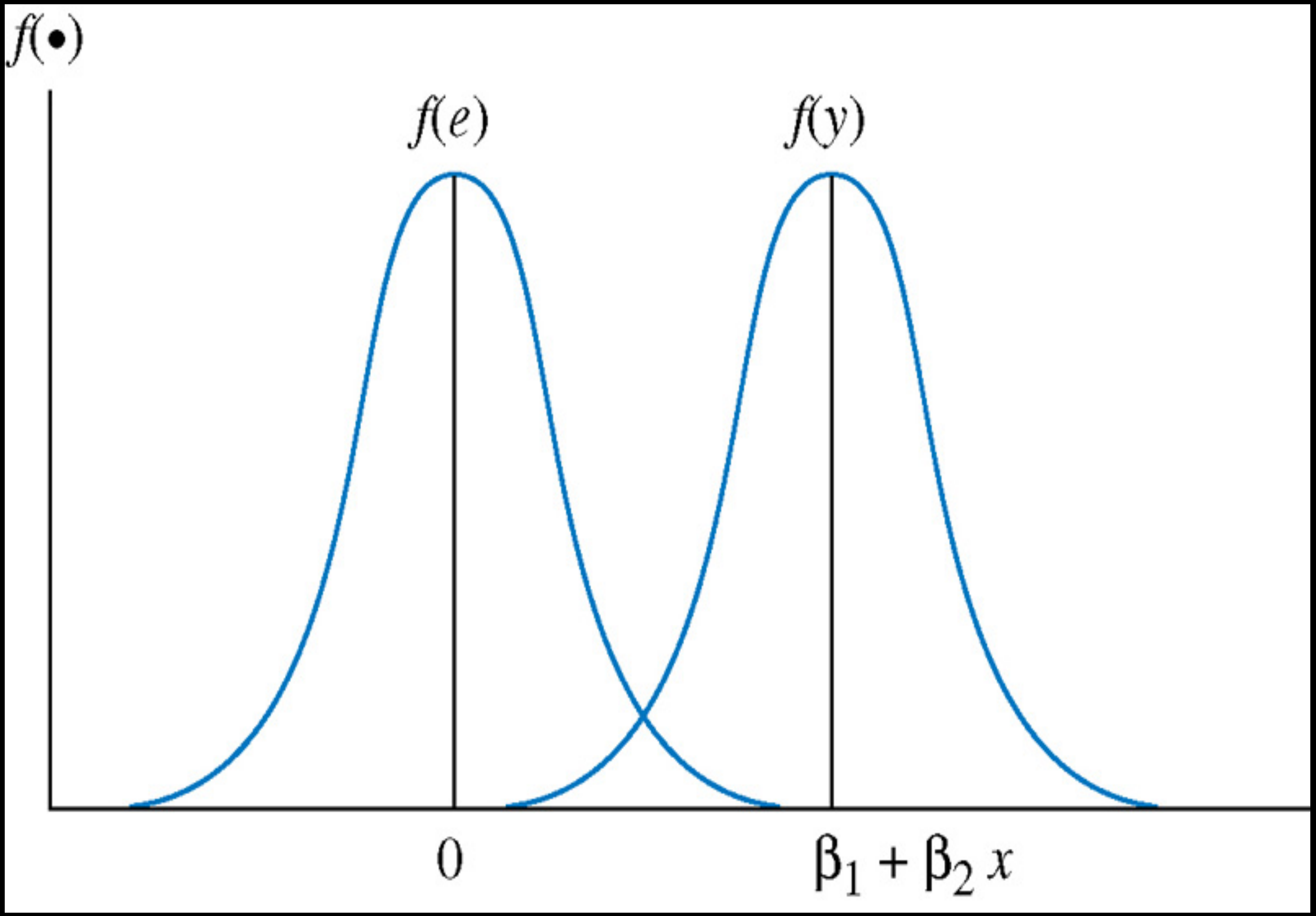


ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL

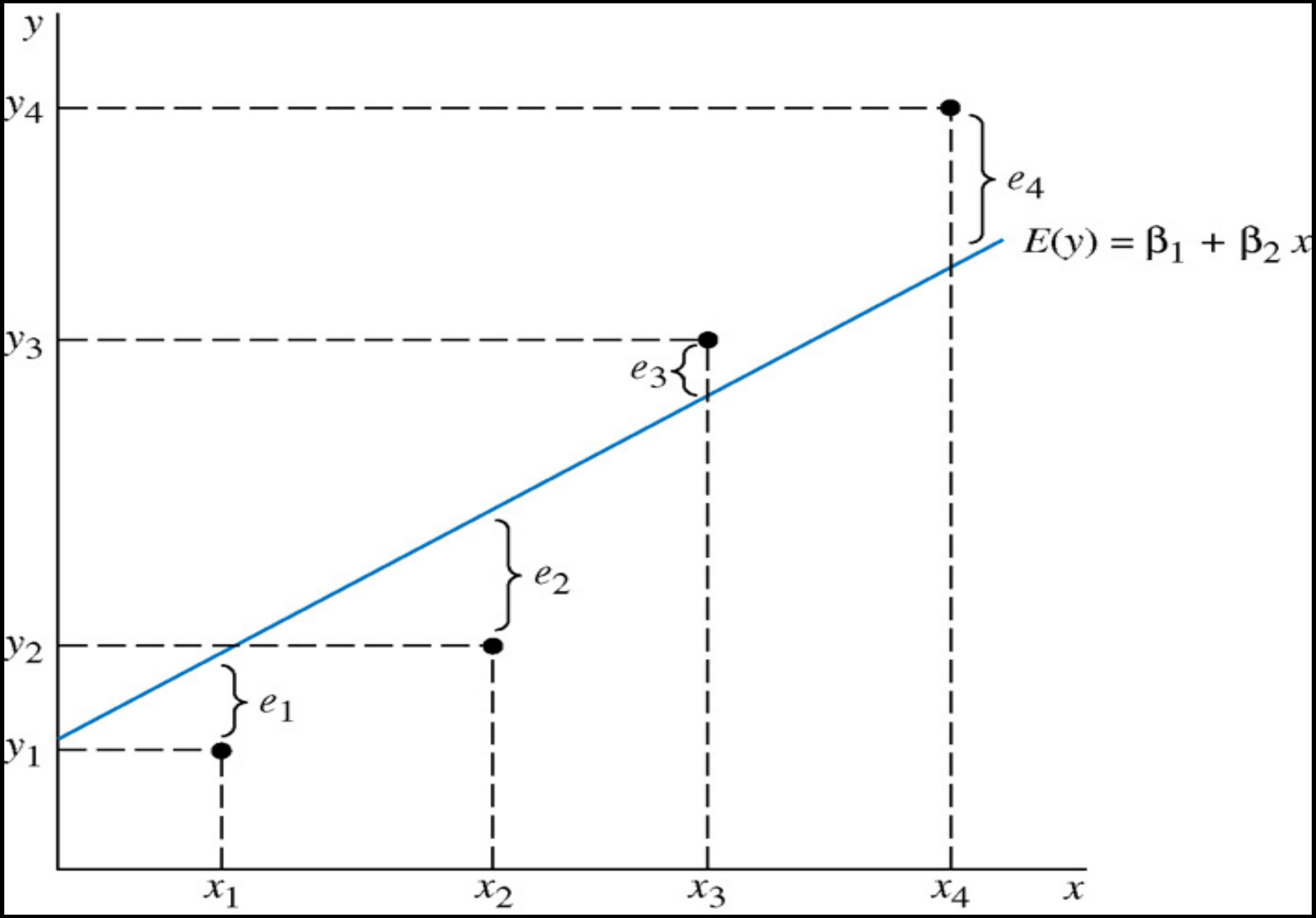
Assumption 5:
(*not crucial*) The values of *e* are *normally distributed* around their mean

This also implies:

Probability density functions for *e* and *y*



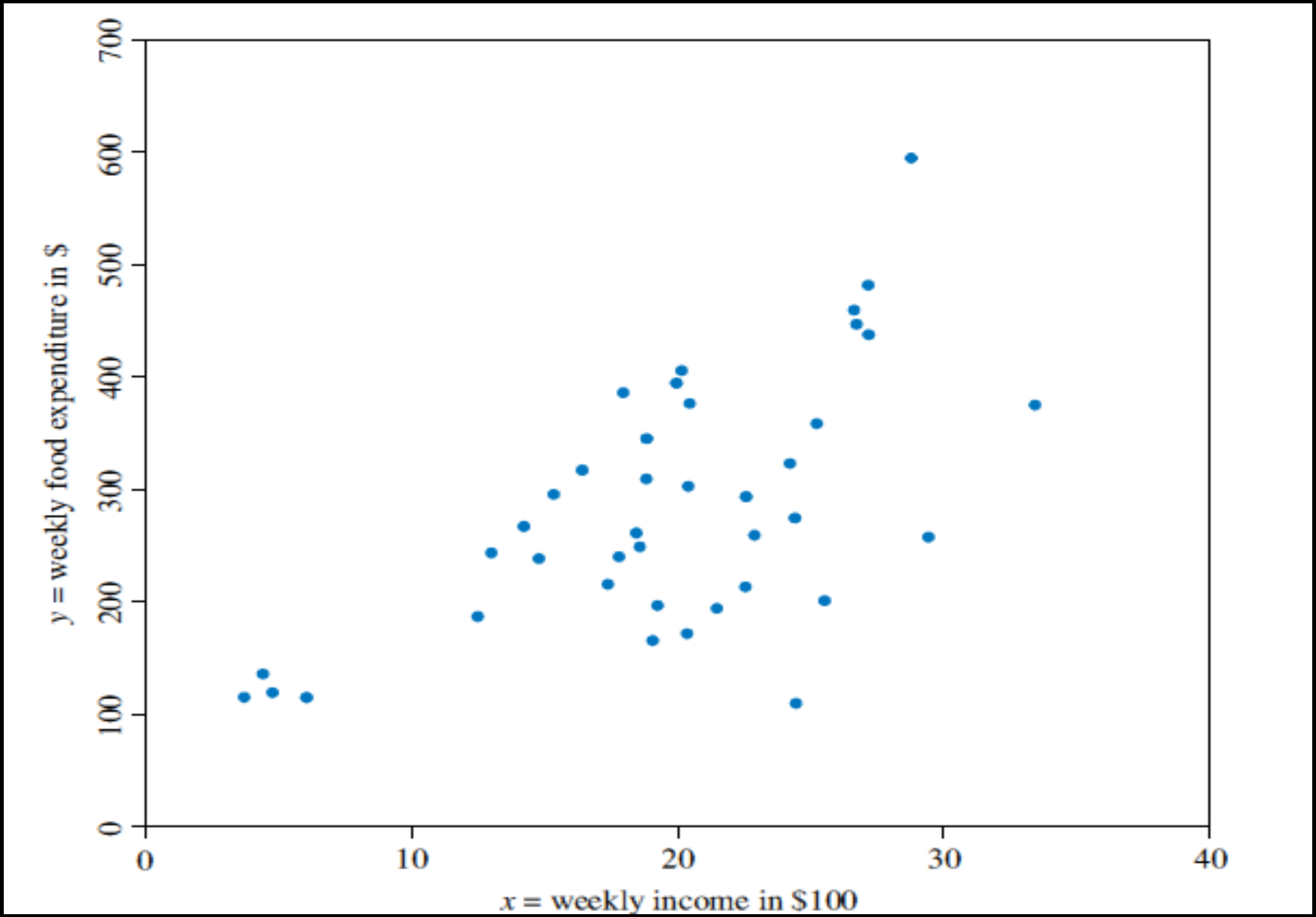
The relationship among *y*, *e* and the true regression line



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

Data on food expenditure and income

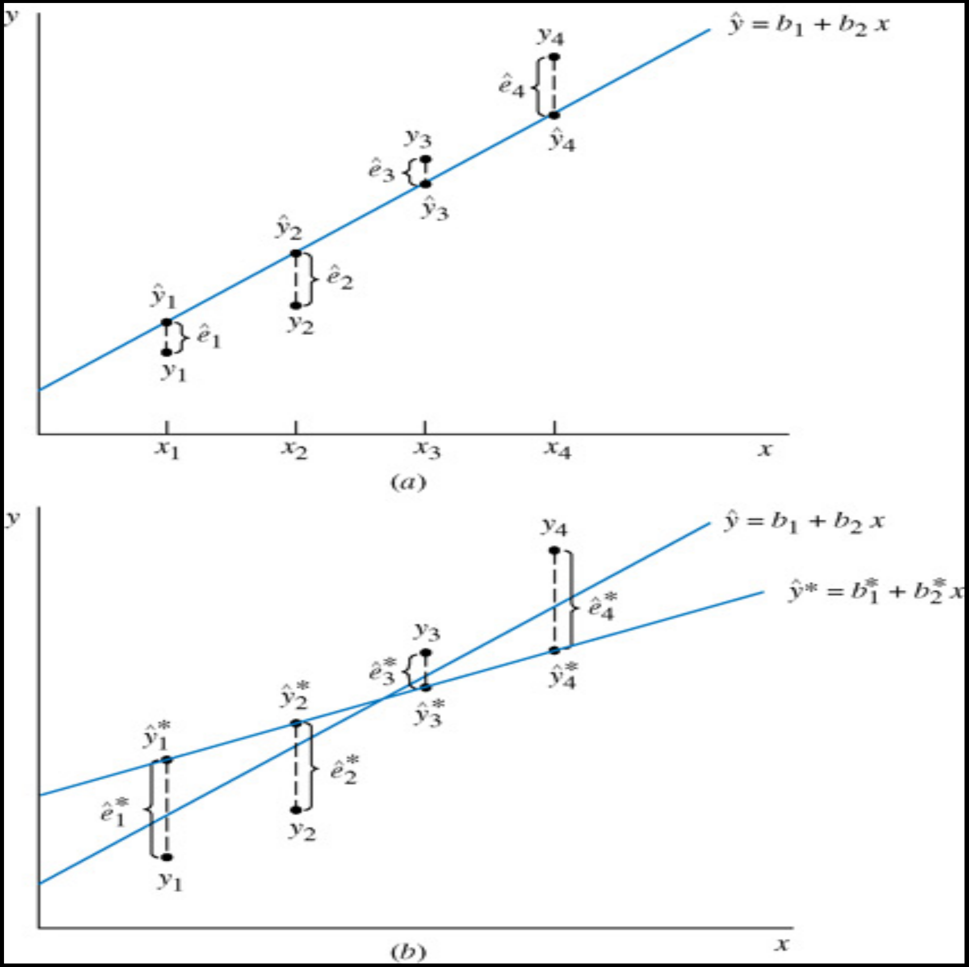


- The fitted regression line is:

The least squares residual is:

The relationship among y, $\hat{e}$ and the fitted regression line

The Least Squares Principle



- Least Squares estimators for the unknown parameters β_1 and β_2 are obtained my minimizing the sum of squared errors:

- Suppose we have another fitted line:

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

Set the derivatives equal to zero to get 2 eqs:

THE LEAST SQUARES ESTIMATORS

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

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

***(OLS)*:**

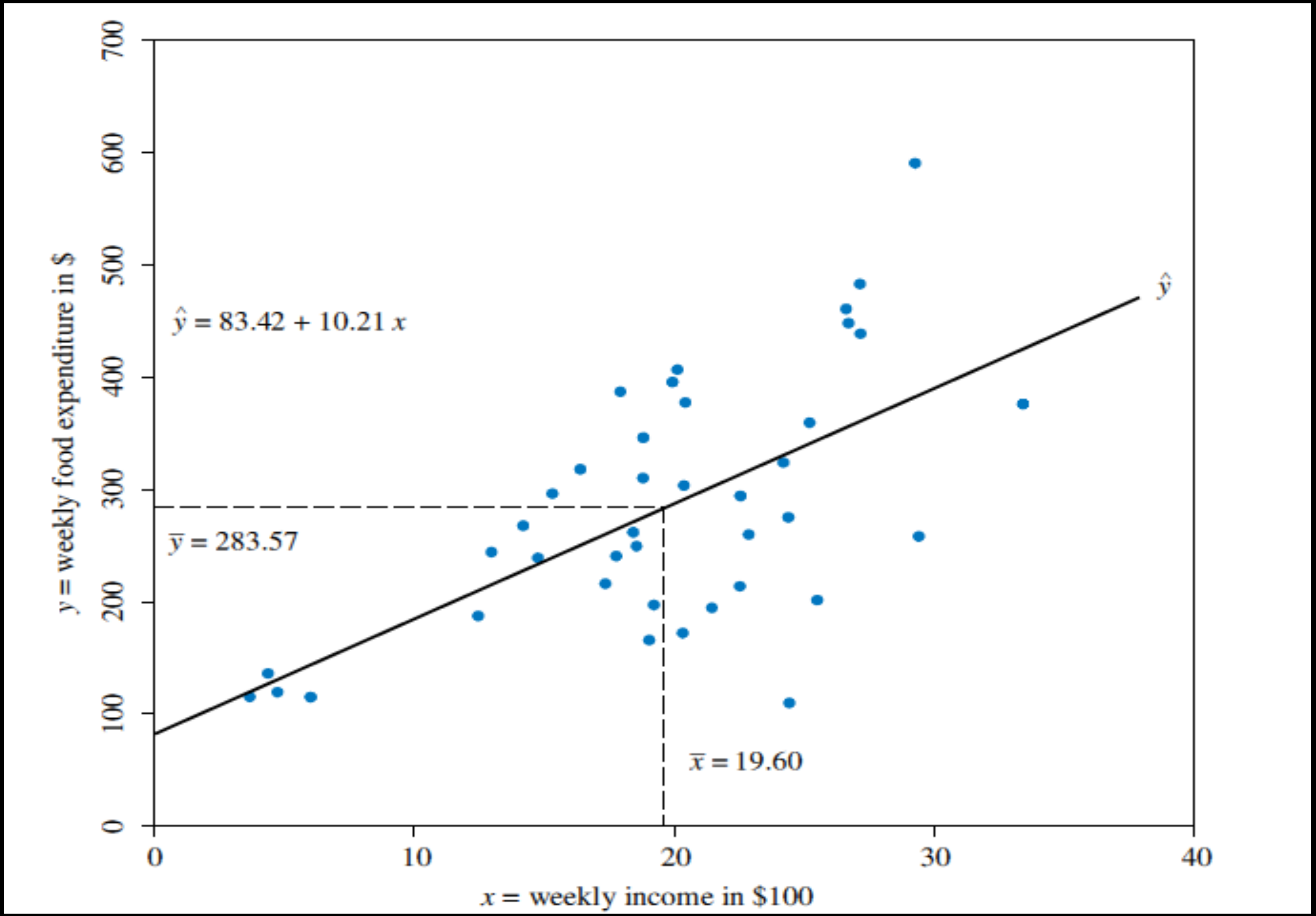
What is the meaning of these formulas?

- Given a specific sample we can now derive our *least squares **estimates***

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

The fitted regression line

Estimates for the Food Expenditure Function

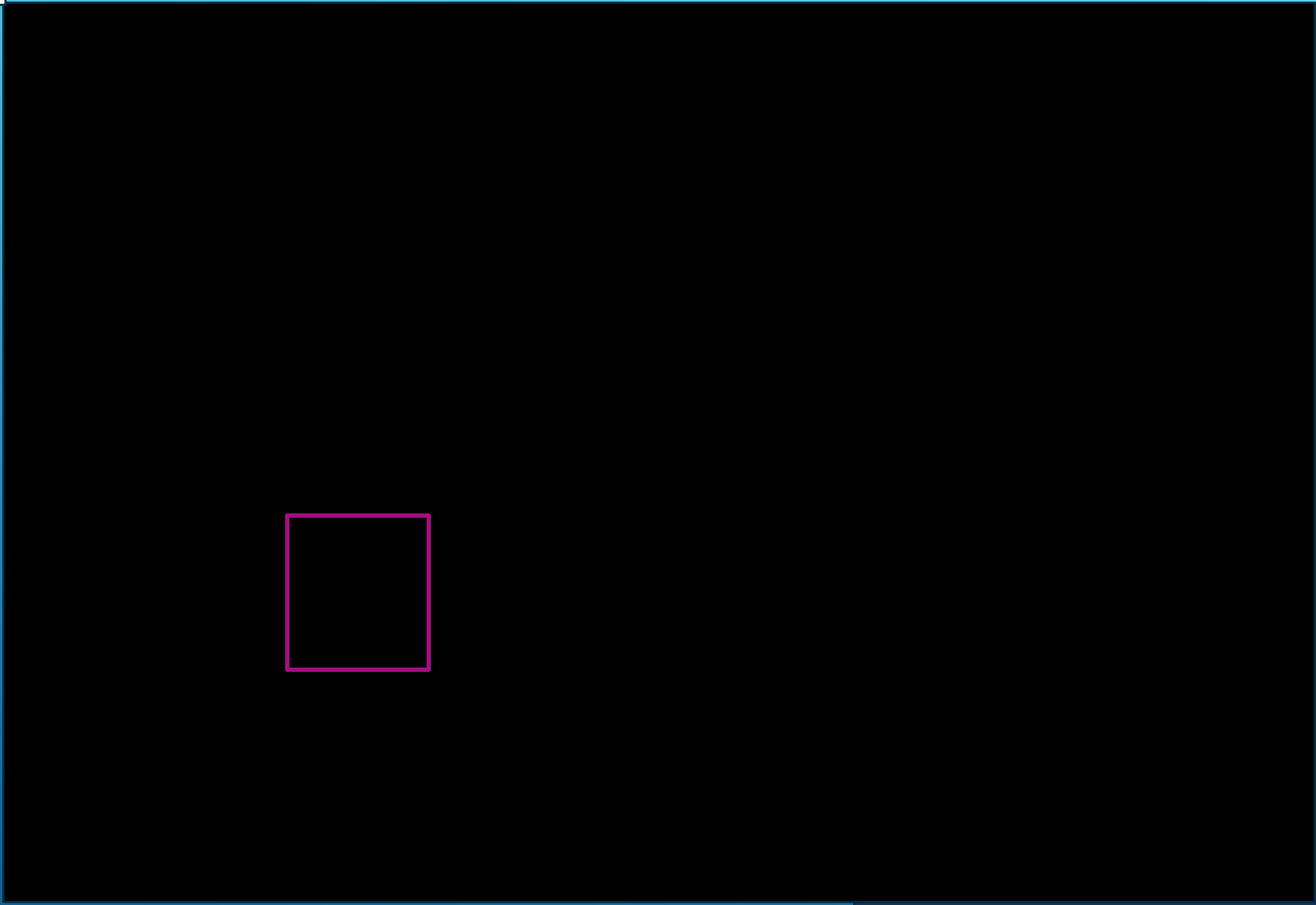


- The value $b_2 = 10.21$, thus, we estimate that if income goes up by \$100, expected weekly expenditure on food will increase by approximately \$10.21

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

Regression Output

Computer Output



- 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:

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

- 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:

We typically calculate the elasticity at the "point of the means" because it is a representative point on the regression line.

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.

THE LEAST SQUARES ESTIMATORS

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

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

Least Squares estimators are random variables!

- How good are the Least Squares **estimators**?
  - What good properties are we looking for?
    - Unbiasedness -> $E(b_2) = \beta_2$
    - Minimal variance (efficiency)
    - Consistency
- How do the Least Squares estimators compare to alternative estimators?

Unbiasedness

- The property of unbiasedness means that if we took the averages of estimates from many samples, these averages would approach the true parameter.

- Unbiasedness does not mean that an estimate from any **one** sample is close to the true parameter.

Estimates from 10 Samples

Sample	b_1	b_2
1	131.69	6.48
2	57.25	10.88
3	103.91	8.14
4	46.50	11.90
5	84.23	9.29
6	26.63	13.55
7	64.21	10.93
8	79.66	9.76
9	97.30	8.05
10	95.96	7.77

Unbiasedness

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

- Let's prove it!

The estimator b_2

can be also written as:

$\sum(\quad)x_i =$ $\sum(x_i-\bar{x})(x_i-\bar{x})$

$= \sum(x_i-\bar{x})^2$

Let's denote

So:

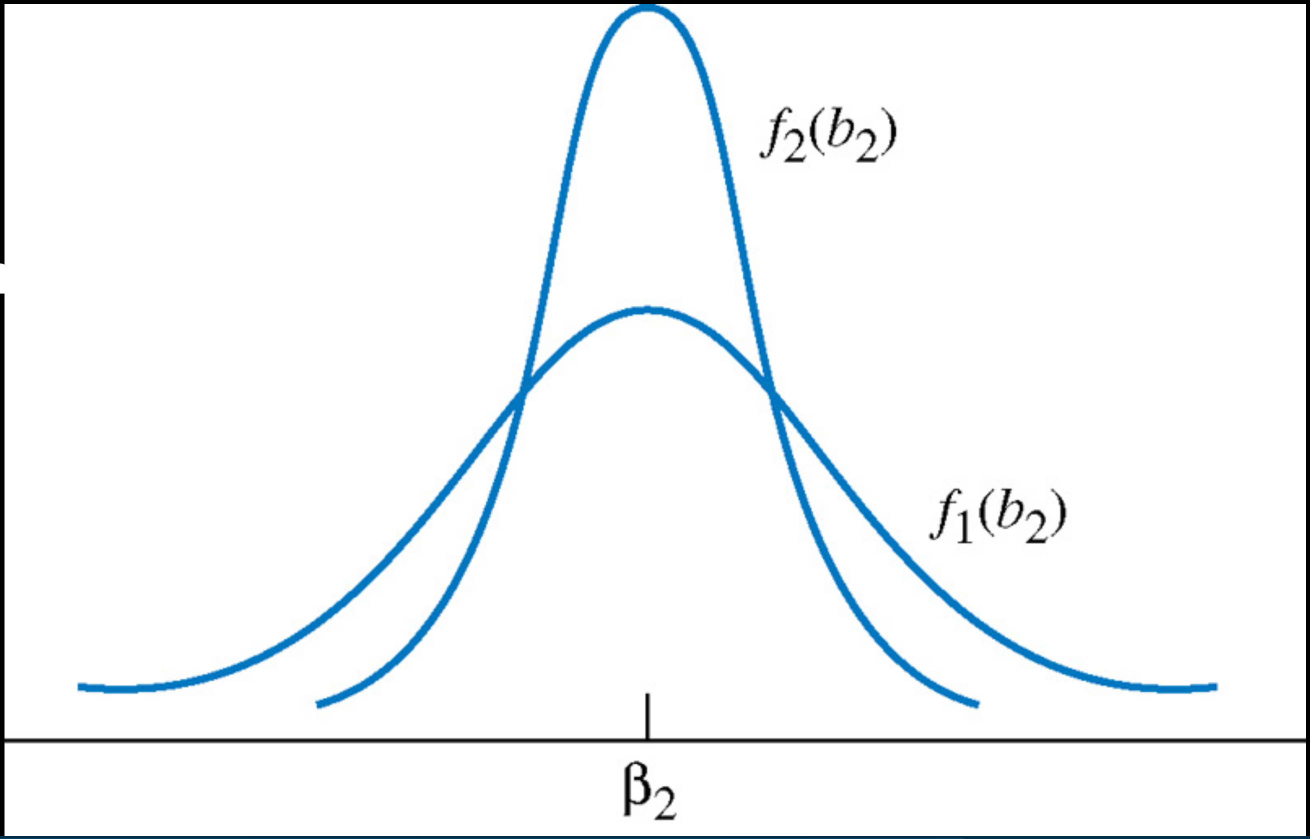
The Estimator b_2

- $E(b_2) = E(\beta_2 + \Sigma w_i e_i) = \beta_2 + \Sigma w_i E(e_i)$
  $= \beta_2$

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

Variance

E
m



Efficiency

- If you have two candidate estimators where one estimator has lower variance than the other, you will say that the estimator with the lower variance is more efficient.

- When comparing between two candidates for 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**.

The Variance of b_2

The Variance of b_1

MAJOR POINTS ABOUT THE VARIANCE OF b_1 AND b_2

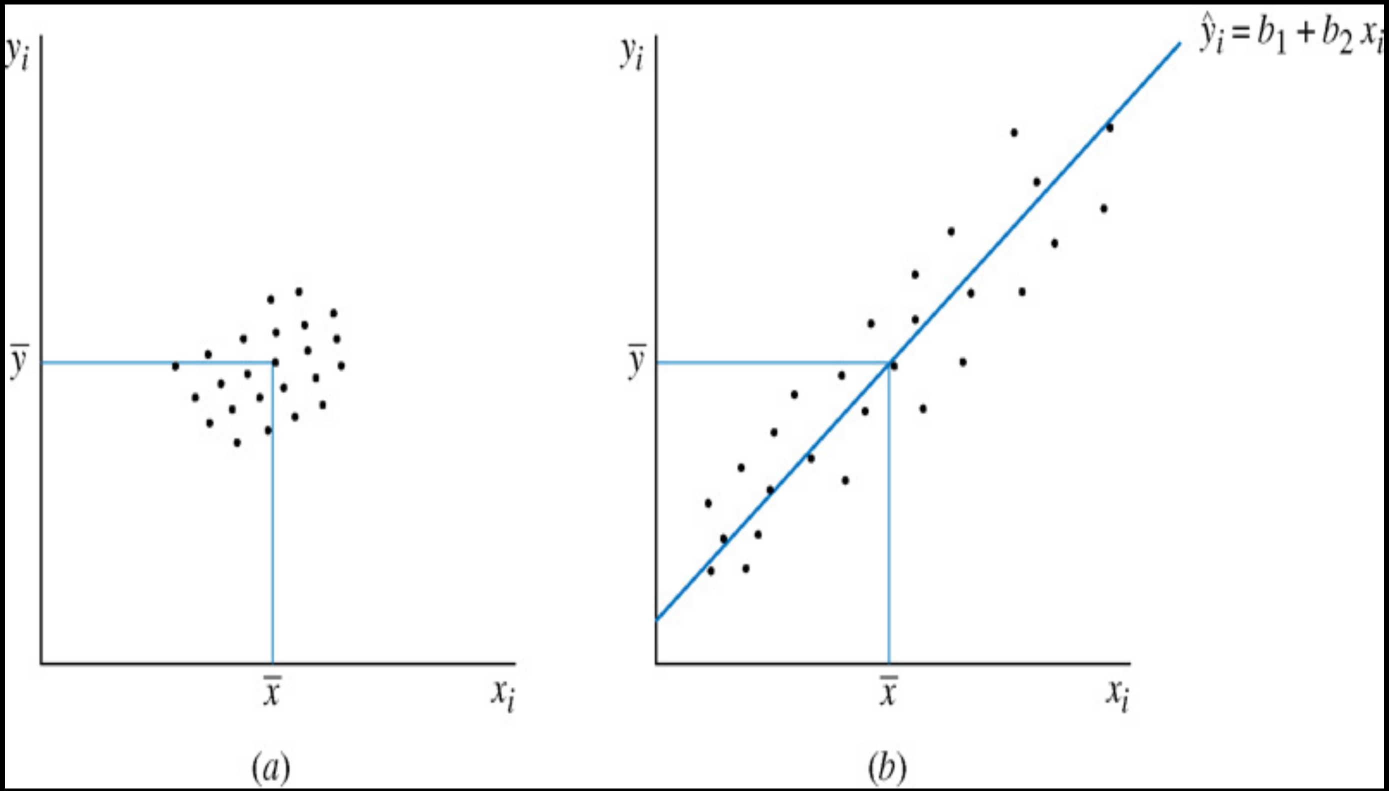
1. The *larger* the variance term σ^2 , the *greater* the uncertainty there is in the statistical model, and the *larger* the variances of the least squares estimators.

2. The *larger* , the *smaller* the variances of the least squares estimators and the more *precisely* we can estimate the unknown parameters.

3. The larger the term , the larger the variance of the least squares estimator b_1 (but note that not b_2)

4. The larger the sample size *N*, the *smaller* the variances of the least squares estimators. The least squares estimator is thus also a **consistent estimator**

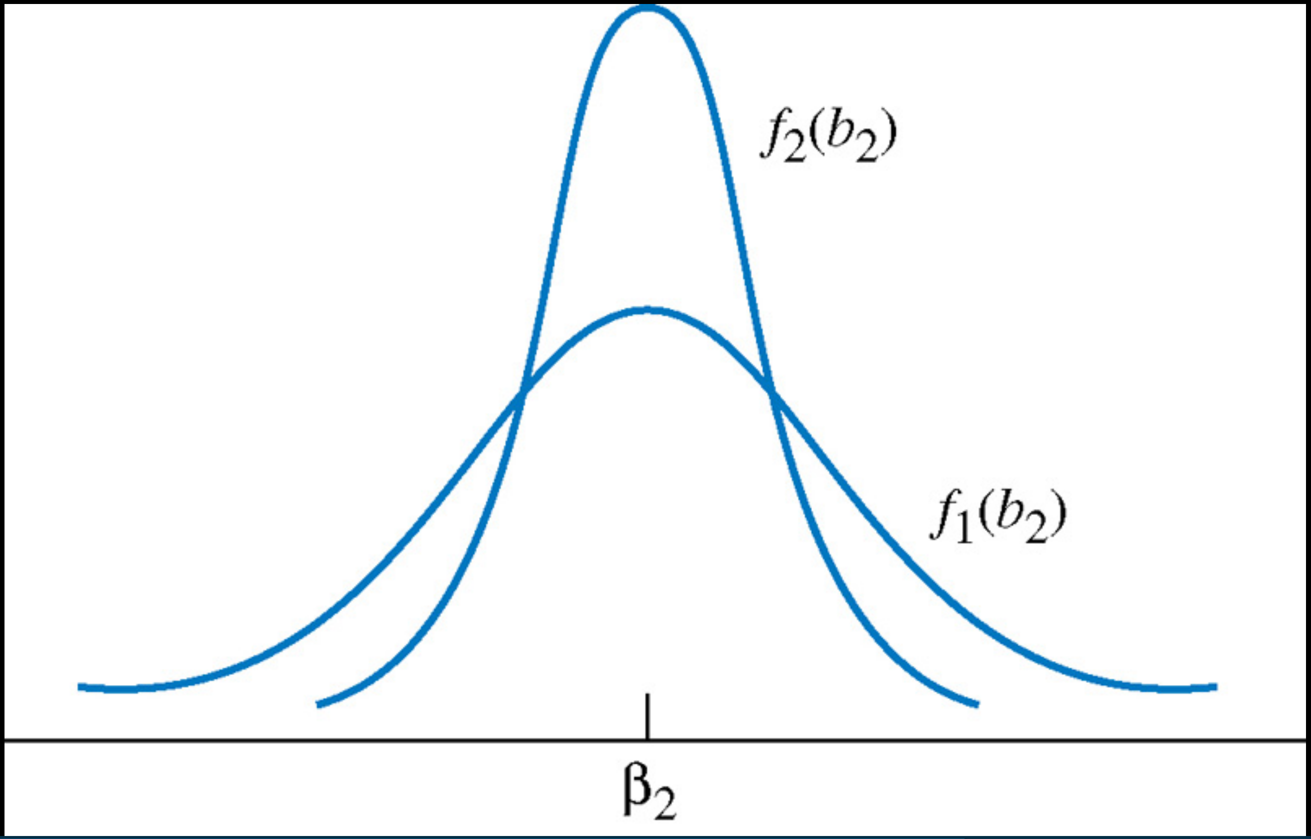
The Variances of b_1 and b_2



Consistency

- An estimator is **consistent** if, as the sample size increases, the estimates (produced by the estimator) "converge" to the true value of the parameter being estimated.
- To be slightly more precise - consistency means that, as the sample size increases, the sampling distribution of the estimator becomes increasingly concentrated at the true parameter value.

Variance of the estimator is decreasing with sample size



Consistency VS. Unbiasedness

- These two are not equivalent!
- **Unbiasedness** is a statement about the expected value of the sampling distribution of the estimator.
- **Consistency** is a statement about "where the sampling distribution of the estimator is going" as the sample size increases.

Consistency VS. Unbiasedness

An Example

- Consider a random variable X that is normally distributed $N(\mu,\sigma^2)$.
- We can randomly draw from this population to create a sample $x_1,\dots,x_n$
- Suppose you're estimating μ.
- X_1 is an unbiased estimator of μ
    - since $E(X_1)=\mu$
- But the variance of X_1 is always σ^2 so it is not consistent.
- Similarly there are examples for estimators that are consistent but still biased.

GAUSS-MARKOV THEOREM

Under the assumptions 2-4 of the linear regression model, the estimators b_1 and b_2 have the smallest variance of all **linear and unbiased estimators** of β_1 and β_2. They are the **Best Linear Unbiased Estimators (BLUE)** of β_1 and β_2

MAJOR POINTS ABOUT THE GAUSS-MARKOV THEOREM

Why Best?

The estimators b_1 and b_2 are **best** within their class because they have the minimum variance. In other words, among the class of linear and unbiased estimators, the least squares estimators are the most **efficient** ones.

MAJOR POINTS ABOUT THE GAUSS-MARKOV THEOREM

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.

MAJOR POINTS ABOUT THE GAUSS-MARKOV THEOREM

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. In order for the Gauss-Markov Theorem to hold, assumptions 2-4 must hold. If any of these assumptions does ***not*** hold, then b_1 and b_2 are ***not*** the best linear unbiased estimators of β_1 and β_2.

3. Gauss-Markov Theorem does *not* depend on the assumption of normality (assumption 5) or the assumption of x's being random variables (assumption 1)

4. The Gauss-Markov theorem applies to the least squares *estimators*. It *does not* apply to the least squares *estimates* from a **single** sample.

A CENTRAL LIMIT THEOREM

If assumptions 2-4 hold, and if the sample size N is ***sufficiently large***, then the least squares estimators have a distribution that approximates the normal distributions.

- We can make the normality assumption that the least squares estimators are normally distributed:

Estimating the Variance of the Least Squares (OLS) Estimator

- We can make the normality assumption that the least squares estimators are normally distributed:

Estimating the Variance of the Error Term

- The variance of the random error e_i is:

- The average is an intuitive average for the mean:

  where :

- This estimator turned out to be consistent but biased.

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

so that:

- Replace the unknown error variance σ^2 by :

Estimating the variance of the Least Squares Estimators

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

Standard Errors in the Regression Output

Computer Output