Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Summer School on Impact Evaluation

Statistics Boot-Camp

September 9th 2017

Instructor: Professor Gaia Narciso
Email: narcisog@tcd.ie

Irish Aid
An Roinn Gnóthaí Eachtracha agus Trádála
Department of Foreign Affairs and Trade

TIME
TRINITY IMPACT EVALUATION UNIT

TIME: Trinity Impact Evaluation Unit

- ✓ Founded in 2015
- ✓ 8 members
- ✓ Partnership with Irish Aid
- ✓ Projects in various countries, e.g. India, Zambia, Uganda, Kenya, Senegal, Vietnam.

Our vision is to provide strong evidence on what works, so that better investment can be made.

Irish Aid
An Roinn Gnóthaí Eachtracha agus Trádála
Department of Foreign Affairs and Trade

TIME
TRINITY IMPACT EVALUATION UNIT

Summer School Instructors

- **Professor Laura Camfield (University of East Anglia)**
- **Professor Michael King (TCD and TIME)**
- **Professor Tara Mitchell (TCD and TIME)**
- **Professor Gaia Narciso (TCD and TIME)**

Contact Info

Instructor: Gaia Narciso

Email: narcisog@tcd.ie

Teaching Assistant: Margaryta Klymak

Email: klymakm@tcd.ie

Readings

Core:

Gujarati, D. and D. Porter (2009), *Basic Econometrics*, 5/e, McGraw-Hill.

Wooldridge, J. (2009), *Introductory Econometrics: A Modern Approach*, 6/e, Cengage.

Supplementary:

Angrist, J. and Pischke, J. (2009), *Mostly Harmless Econometrics*, Princeton University Press.

Stats Boot Camp - Schedule

9am-9.15am: *Registration*

9.15am-11am: *Topic 1 - Statistical Review*

11am-11.15am: *Coffee Break*

11.15am-12.45pm: *Topic 1 - Statistical Review*

12.45pm-13.30pm: *Lunch Break*

13.30pm-14.30pm: *Topic 2 - Linear Regression Model*

14.30pm-15.30pm: *Topic 3 - Statistical Inference*

15.30pm-15.45pm: *Coffee Break*

15.45pm-17pm: *Lab session* (AP0.12)

Road Map

Topic 1: Statistical Review

i. Random Variables and their Probability Distribution

ii. Joint distributions, Conditional distributions and Independence

iii. Features of Probability Distributions

iv. Features of Joint and Conditional Probability Distributions

v. Populations, Parameters and Random Sampling

vi. Estimators and Estimates

vii. Finite Sample Properties of Estimators

viii. Asymptotic Properties of Estimators

ix. Interval Estimation and Confidence Intervals

x. Hypothesis Testing

Road map

Topic 2: The Linear Regression Model

i. The Simple Regression Model

ii. Ordinary Least Squares (OLS) Estimation

iii. Properties of OLS

iv. Goodness of Fit

v. The Multiple Regression Model

vi. Model Specification

vii. Dummy Variables in Regression Analysis

Topic 3: Statistical inference

Topic 1: Statistical Review

Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

- A random variable is a variable whose value is a numerical outcome of a random phenomenon.

- Denoted by uppercase letters (e.g., X)

- Values of the random variable are denoted by corresponding lowercase letters

- Corresponding values of the random variable:
  $x_1, x_2, x_3, \ldots$

Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

– Random variables may be classified as:

- ***Discrete:*** The random variable assumes a countable number of distinct values

- ***Continuous*:** The random variable is characterized by (infinitely) uncountable values within any interval

– Every random variable is associated with a ***probability distribution*** that describes the variable completely

Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

- A random variable is a variable whose value is a numerical outcome of a random phenomenon.
- A ***discrete*** *random variable* is one which takes a ***finite*** number of values
- All possible outcomes are summarized in what is known as the ***probability distribution***
- *Example*: Suppose X is the number of free throws scored by a basketball player out of two attempts so that $X \in \{0,1,2\}$

  Suppose the ***probability distribution*** of X is given by

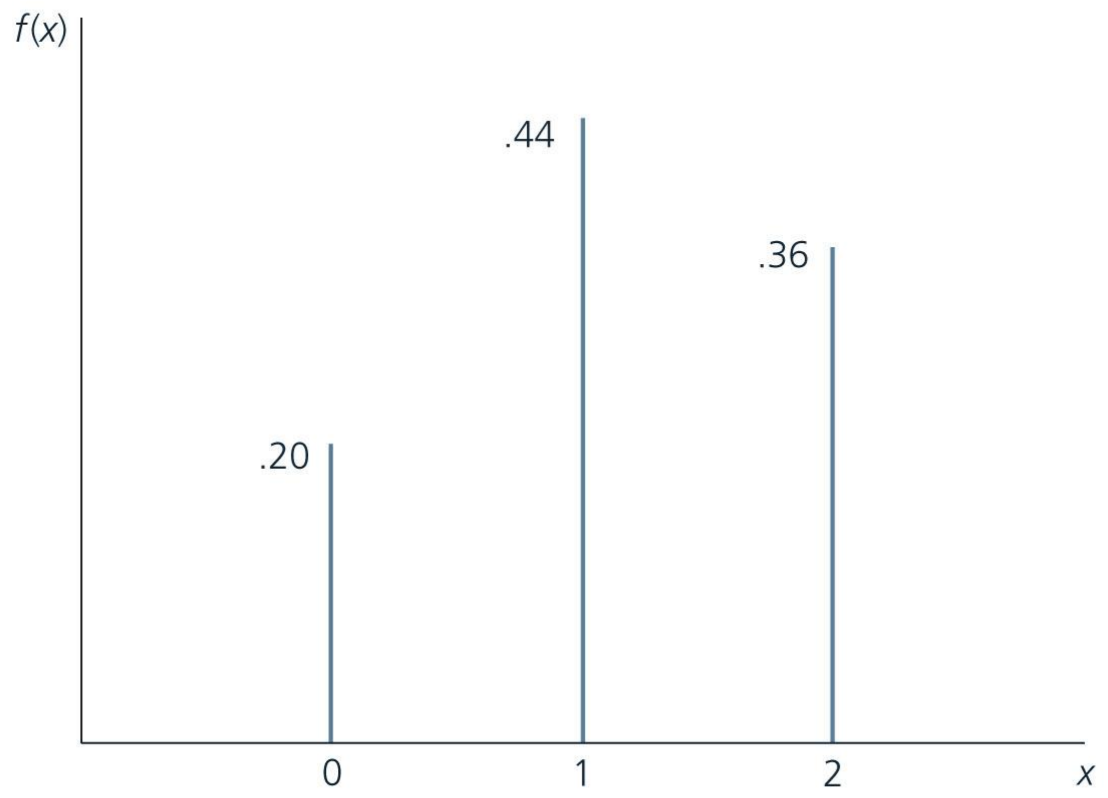
$$f(0) = 0.2,\ f(1) = 0.44,\ f(2) = 0.36$$

a. Calculate the probability that the player makes at least one free throw

b. Draw the pdf of X

FIGURE B.1

The pdf of the number of free throws made out of two attempts.



Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

- A ***continuous*** *random variable* is characterized by (infinitely) uncountable values within any interval
- A ***continuous*** *random variable* takes on so many possible values that it cannot be matched to a positive integer
- Unlike with discrete RVs, continuous RVs have an ***infinite number of potential outcomes***
- With continuous RVs, we deal with intervals, not outcomes: hence ***probability density function*** [PDF]

Topic 1: Statistical Review

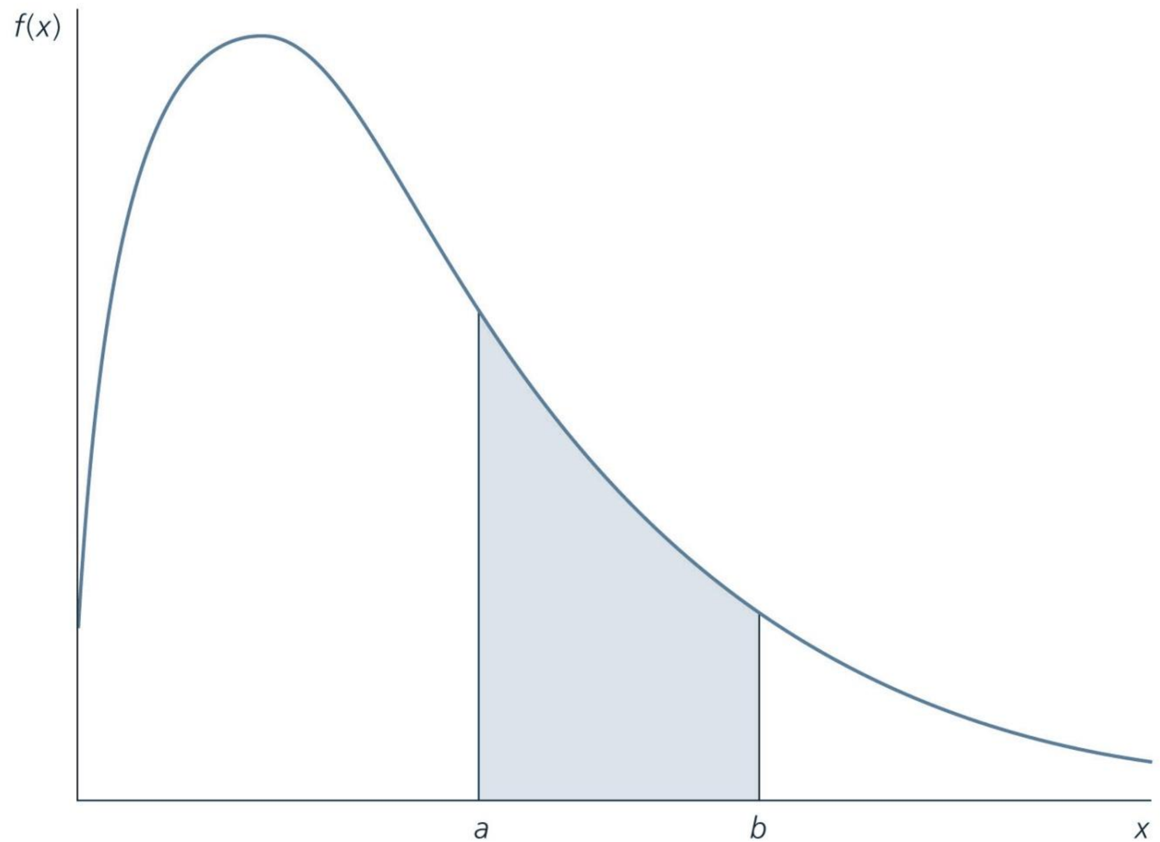
1. Random Variables and their Probability Distribution

- With continuous RVs, we deal with intervals, not outcomes
- Hence probability density function [PDF]
- *A PDF, f(x), of a continuous RV describes the relative likelihood that X assumes a value within a given interval*

- *Example*:
  - **X**: Length of time that a user spends on a webpage before clicking on a link or leaving the page.
  - The probability that **X** lies between *15 seconds* and *30 seconds* is given by the area under the probability density function

FIGURE B.2

The probability that X lies between the points a and b.



Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

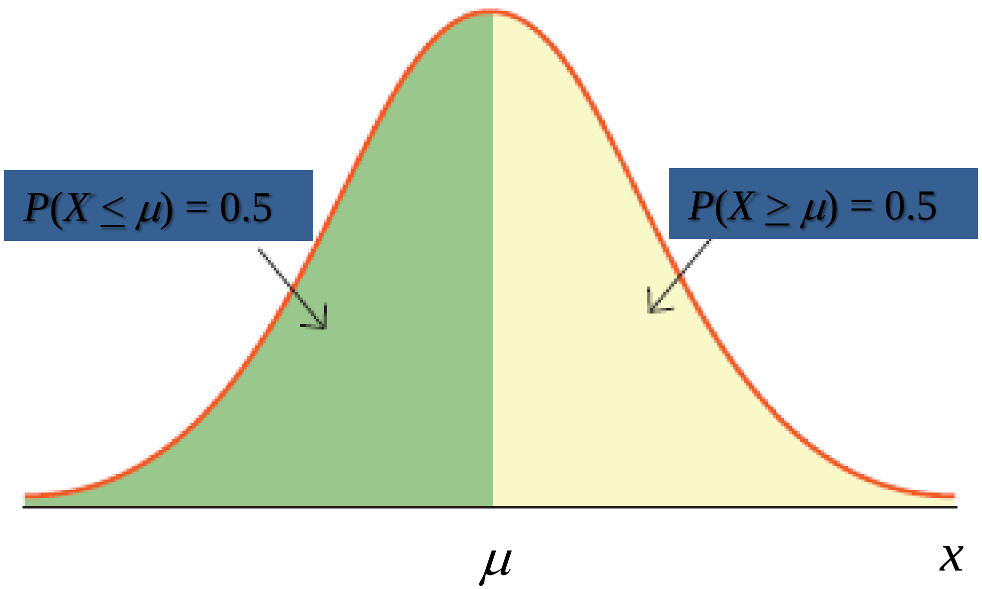
- **“Normal Distribution”** closely approximates the probability distribution of a wide range of real-world RVs – examples include:
  - Rainfall
  - Biology: height, weight, skin area of many animals (often after a ‘log transformation’)
  - Standardized test results (e.g. SAT scores)
  - Financial variables (but this can be contested)
- **The cornerstone of statistical inference**

Topic 1: Statistical Review

1. Random Variables and their Probability Distribution

– The normal distribution is…

- Symmetric
- Bell-shaped and asymptotic: tail gets ever closer to (without touching) the axis



Topic 1: Statistical Review

2. Joint Distributions, Conditional Distributions and Independence

- The *joint probability density function* of two variables (Y,X) can be defined as:

$$f_{Y,X}\left(y,x\right)=P(Y=y,X=x)$$

- In econometrics we are interested in how one random variable is related to another – *conditional distribution of Y given X:*

$$f_{Y|X}\left(y\,|\,x\right)=P\left(Y=y\,|\,X=x\right)$$

- The symbol “ | ” means “given” - in other words, whatever follows “ | ” has already occurred

- If Y and X are independent: $f_{Y|X}\left(y\,|\,x\right)=f_{Y}\left(y\right)$

Topic 1: Statistical Review

3. Features of Probability Distributions

- Expected Value, Population mean
- Variance
- Standard Deviation
- Covariance
- Correlation

Topic 1: Statistical Review

3. Features of Probability Distributions

Expected Value:

- The expected value of a *discrete* R.V. Y is the weighted average of all possible values of Y where the weights are determined by the probability distribution.
- The expected value is called the ***population mean***

$$Y \in \left\{ y_1, y_2, y_3, \ldots, y_k \right\}$$

$$E\left(Y\right) = y_1 prob(\mathrm{Y} = y_1) + y_2 prob(\mathrm{Y} = y_2) + \ldots . + y_K prob(\mathrm{Y} = y_k) = \sum_{j=1}^{k} y_j prob(\mathrm{Y} = y_j) = \mu$$

- *Example*: $X \in \left\{0,1,2\right\}$ $prob\left(Y=0\right) = 0.2,\ prob(Y=1) = 0.44,\ prob(Y=2) = 0.36$
- Properties of expectations

Properties of Expectations:

E1: For any constant c, $E(c) = c$

E2: For any constants a and c, $E(aY + c) = aE(Y) + c$

E3: If $\{a_1, a_2, \ldots\ldots, a_n\}$ are constants and $\{Y_1, Y_2, \ldots\ldots, Y_n\}$ are random variables.

$$E(a_1Y_1 + a_2Y_2 + \ldots.. + a_nY_n) = a_1E(Y_1) + a_2E(Y_2) + \ldots. + a_nE(Y_n)$$

$$E\left(\sum_{i=1}^{n} a_iY_i\right) = \sum_{i=1}^{n} a_iE(Y_i)$$

Topic 1: Statistical Review

3. Features of Probability Distributions

<u>Variance</u>:

- Measures how far away a random variable Y is from its population mean:

$$Var(Y) = E\left[\left(Y - E(Y)\right)^2\right] = E\left[\left(Y - \mu\right)^2\right] = \sigma^2$$

- This can also be written as:

$$Var(Y) = E\left(Y^2\right) - \mu^2$$

- Properties of variances

Properties of Variances

V1: For any constant c: $V(c) = 0$

V2: For any constants a and c $V(aY + c) = a^2 V(Y)$

V3: For two random variables X and Y and constants *a* and *b*

$$V(aX + bY) = a^2 V(X) + b^2 V(Y) + 2abCov(X, Y)$$

V4: If $\{a_1, a_2, \ldots\ldots, a_n\}$ are constants and $\{Y_1, Y_2, \ldots\ldots, Y_n\}$ are *uncorrelated* random variables then

$$Var(a_1Y_1 + a_2Y_2 + \ldots\ldots + a_nY_n) = a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + \ldots\ldots a_n^2 Var(Y_n)$$

Topic 1: Statistical Review

3. Features of Probability Distributions

<u>Standard Deviation</u>:

– Positive square root of the variance of the random variable:

$$sd(Y) = \sqrt{Var(Y)} = \sigma$$

Topic 1: Statistical Review

3. Features of Probability Distributions

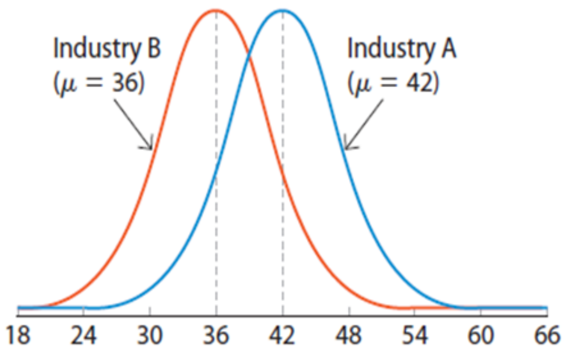
Normal Distribution

- The normal distribution is completely described by two parameters: μ and σ^2
- The population mean describes the distribution's central location
- The population variance describes the distribution's dispersion

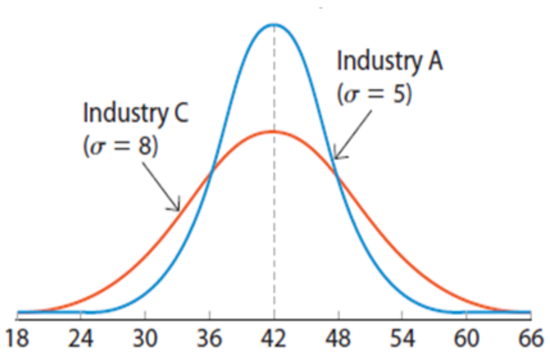
Topic 1: Statistical Review

Example: employee age across three industries

Industry A	Industry B	Industry C
$\mu = 42$ years	$\mu = 36$ years	$\mu = 42$ years
$\sigma = 5$ years	$\sigma = 5$ years	$\sigma = 8$ years



σ is the same, μ is different.



μ is the same, σ is different.

Topic 1: Statistical Review

3. Features of Probability Distributions

<u>Standardizing a Random Variable</u>:

– Given a r.v. Y, a new r.v. can be defined by subtracting the mean and dividing by the standard deviation

$$Z = \frac{Y - \mu}{\sigma}$$

– The ***standard normal distribution*** is a special case, where:

- Mean is equal to zero (E(Z) = 0)
- Standard deviation is equal to one (SD(Z) = 1)

Topic 1: Statistical Review

4. Features of Joint and Conditional Probability Distributions

Covariance:

– Defines the relationship between two random variables, Y and X. It tells us the extent to which the two variables move in the same direction:

$$Cov(Y,X) = E[(Y - E(Y))(X - E(X))]$$
$$= E[(Y - \mu_Y)(X - \mu_X)] = \sigma_{YX}$$

– Properties of Covariance:

COV1: If Y and X are independent then $Cov(Y,X) = 0$

COV2: For any constants a_1, b_1, a_2, b_2 :

$$Cov(a_1 Y + b_1, a_2 X + b_2) = a_1 a_2 Cov(Y,X)$$

Covariance between two r.v.s depends on the units of measurement

COV3: $|Cov(Y,X)| \leq sd(Y) sd(X)$

(Cauchy-Swartz Inequality)

Topic 1: Statistical Review

4. Features of Joint and Conditional Probability Distributions

Correlation:

– Measures the strength of the relationship between two random variables. It does not depend on the units of measurement:

$$Corr(Y,X) = \frac{Cov(Y,X)}{sd(Y)sd(X)} = \frac{\sigma_{YX}}{\sigma_Y \sigma_X} = \rho_{YX}$$

$$-1 \leq \rho_{YX} \leq 1$$

Topic 1: Statistical Review

4. Features of Joint and Conditional Probability Distributions

Conditional Expectation

– Summarises the relationship between Y and X using the *conditional mean* of Y given X

$$E[Y \mid X = x] = \sum_{j=1}^{m} y_j \, f_{Y|X}(y_j \mid X = x)$$

– Weighted average of all possible values of Y, taking account of the fact that X takes on a particular value

– *Reminder:* $f_{Y|X}(y \mid x) = P(Y = y \mid X = x)$ *is the conditional probability distribution*

Topic 1: Statistical Review

4. Features of Joint and Conditional Probability Distributions

- <u>Properties of Conditional Expectations</u>
- CE1: $E[f(Y)|Y] = f(Y)$
- CE2: $E[f(X)Y + g(X)|X] = f(X)E[Y|X] + g(X)$
- CE3: If Y and X are independent then: $E[Y|X] = E[Y]$
- CE4: The Law of Iterated Expectations: $E_X[E[Y|X]] = E[Y]$
- CE5: If $E[Y|X] = E(Y)$ then $Cov(Y,X) = 0$

Topic 1: Statistical Review

5. Populations, Parameters and Random Sampling

- Use statistical inference to learn something about a ***population***
- ***Population***: Complete group of agents, e.g. the population of students studying Economics at TCD
- Typically only observe a ***sample of data***
- *Random sampling*: Drawing random samples from a population
- Know everything about the distribution of the population except for one *parameter*
- Use statistical tools to say something about the unknown parameter
  - Estimation and hypothesis testing

Topic 1: Statistical Review

6. Estimators and Estimates:

Population: consists of all items of interest

– The **Population Parameter** is unknown

Sample: a subset of the population

– The **Sample Statistic** is calculated from sample and used to make inferences about the population (and its parameters)

Topic 1: Statistical Review

6. Estimators and Estimates:

- Given a random sample drawn from a population distribution that depends on an unknown parameter θ, an ***estimator*** of θ is a rule that assigns each possible outcome of the sample a value of θ
- Examples:
  - Estimator for the population mean
  - Estimator for the variance of the population distribution
- An ***estimator*** is given by some function of the RVs
- This yields a (point) ***estimate***
- **Distribution of estimator is the sampling distribution**

Topic 1: Statistical Review

6. Estimators and Estimates:

- **Estimator**: a statistic used to estimate a population parameter; e.g. the *sample mean* is a RV which is an estimator of μ, the population parameter
- **Estimate**: a particular value of the estimator; e.g. the mean of a given sample

Topic 1: Statistical Review

6. Estimators and Estimates

- Each sample drawn from a population produces its own estimate of μ, i.e. its mean
- Take a given sample size, *n*, – each sample of that size will have its own mean
- Therefore the sample mean has its own probability distribution
  - This distribution is called '**the sampling distribution of the mean**'

Topic 1: Statistical Review

7. Properties of Estimators:

A Point Estimator should be…

Unbiased

An estimator is unbiased if its expected value equals the unknown population parameter being estimated

Efficient

An unbiased estimator is efficient if its standard error is lower than that of other unbiased estimators

Consistent

An estimator is consistent if it approaches the unknown population parameter being estimated as the sample size grows larger

Topic 1: Statistical Review

7. Finite Sample Properties of Estimators:

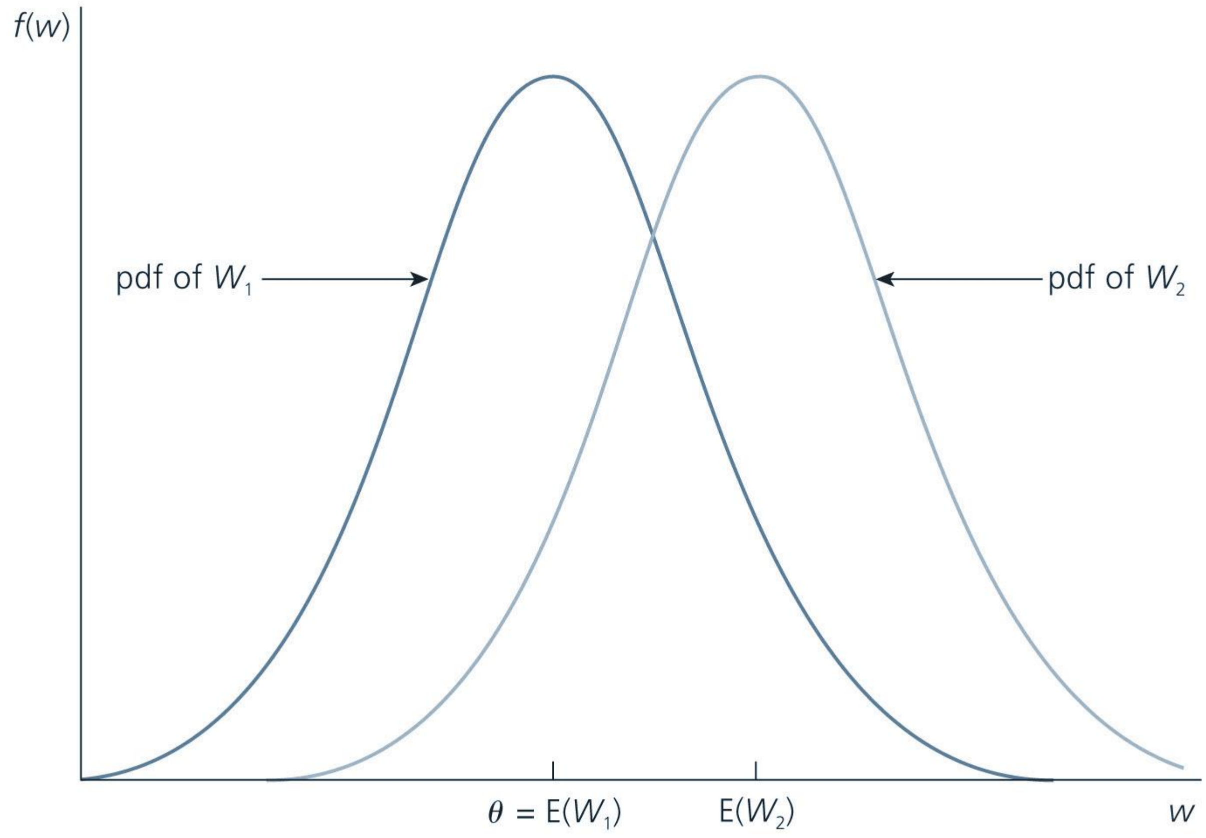
Unbiasedness

An estimator $\hat{\theta}$ of θ is unbiased if $E\left(\hat{\theta}\right)=\theta$ for all values of θ

i.e., on average the estimator is correct

FIGURE C.1

An unbiased estimator, W_1, and an estimator with positive bias, W_2.



Topic 1: Statistical Review

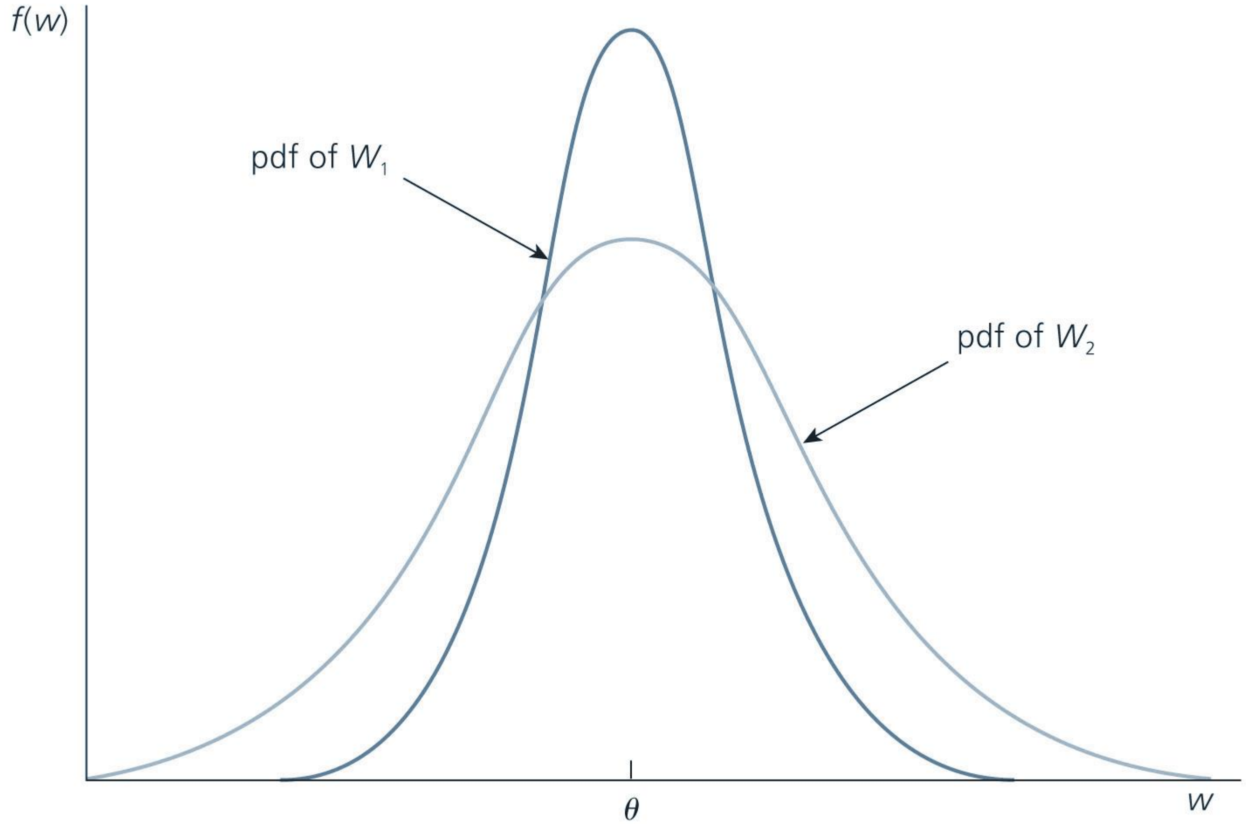
7. Finite Sample Properties of Estimators:

Efficiency

- What about the **dispersion of the distribution of the estimator**? i.e., how likely is it that the estimate is close to the true parameter?
- Useful summary measure for the dispersion in the distribution is the ***sampling variance***.
- ***An efficient estimator is one which has the least amount of dispersion about the mean i.e. the one that has the smallest sampling variance***

FIGURE C.2

The sampling distributions of two unbiased estimators of θ.



Topic 1: Statistical Review

8. Asymptotic Properties of Estimators

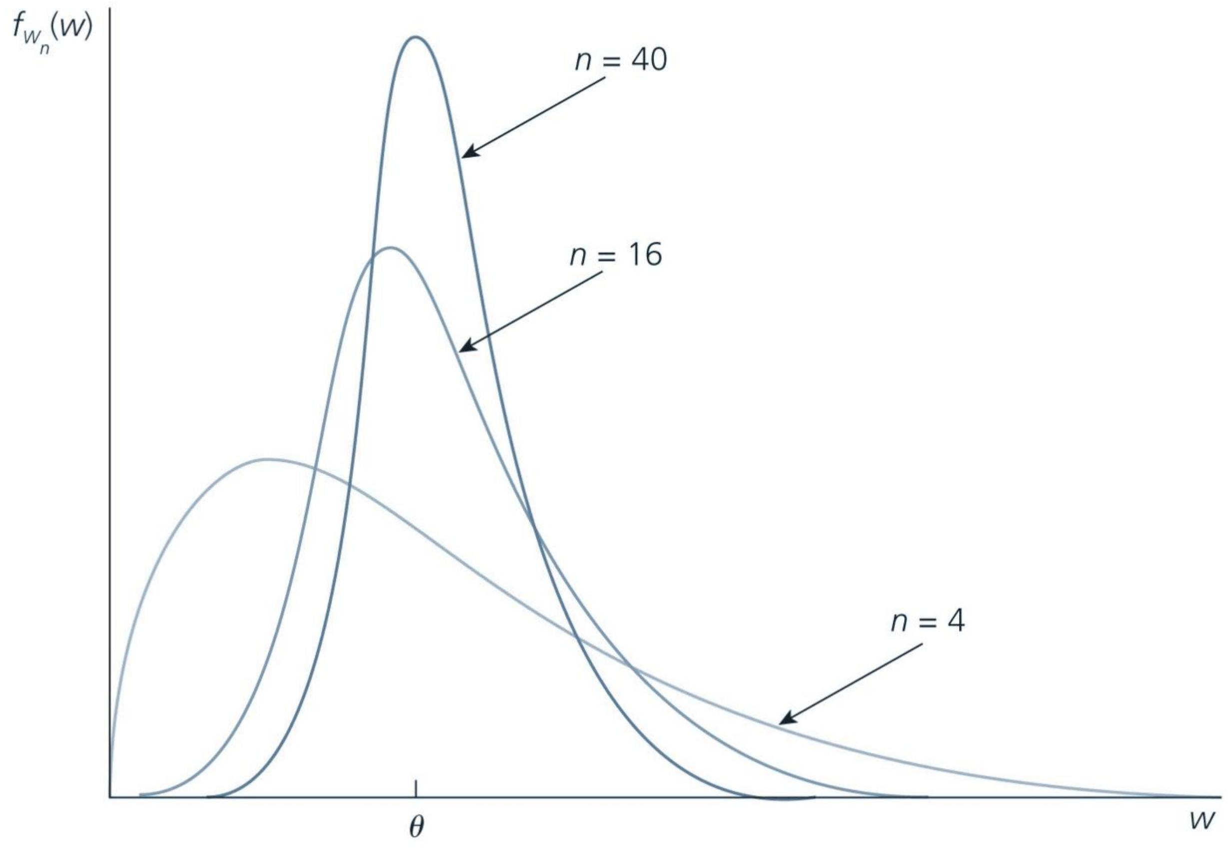
– How do estimators behave if we have very large samples – as *n* increases to infinity?

Consistency

How far is the estimator likely to be from the parameter it is estimating as the sample size increases indefinitely.

FIGURE C.3

The sampling distributions of a consistent estimator for three sample sizes.



Topic 1: Statistical Review

8. Asymptotic Properties of Estimators

– **Asymptotic Normality**

An estimator is said to be asymptotically normally distributed if its sampling distribution tends to approach the normal distribution as the sample size increases indefinitely.

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

- How do we know how accurate an estimate is?

- A **confidence interval** estimates a population parameter within a range of possible values at a specified probability, called the *level of confidence*, using information from a known distribution – the standard normal distribution

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

- A Confidence Interval (CI) provides a range of values that, with a certain level of confidence, contains the population parameter of interest
  - If we took many samples of the same size from a population with mean μ and calculated a confidence interval for each sample, we would find that μ lies within 95% of the intervals
- Also referred to as an "*interval estimate*"
- CIs are constructed around the ***point estimate, ± the margin of error***
- Margin of error accounts for the variability of the estimator and the desired confidence level of the interval

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

- Consider a normally distributed RV, Y
- Two key summary statistics ("moments") are μ, its expected value, and σ, its SD
- Remember, we can convert any normally distributed RV into standard normal

$$Z = \frac{Y - \mu}{\sigma}$$

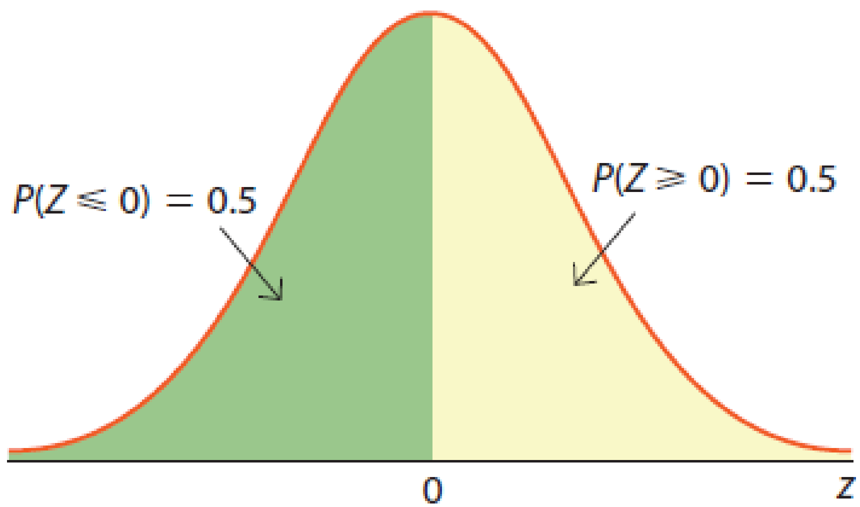
- We would like to build a **confidence interval for μ**
- For now, **we assume that σ is known**

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

The standard normal distribution is a special case, where:

- Mean (μ) is equal to zero ($E(Z) = 0$)
- Standard deviation (σ) is equal to one ($SD(Z) = 1$)



Topic 1: Statistical Review

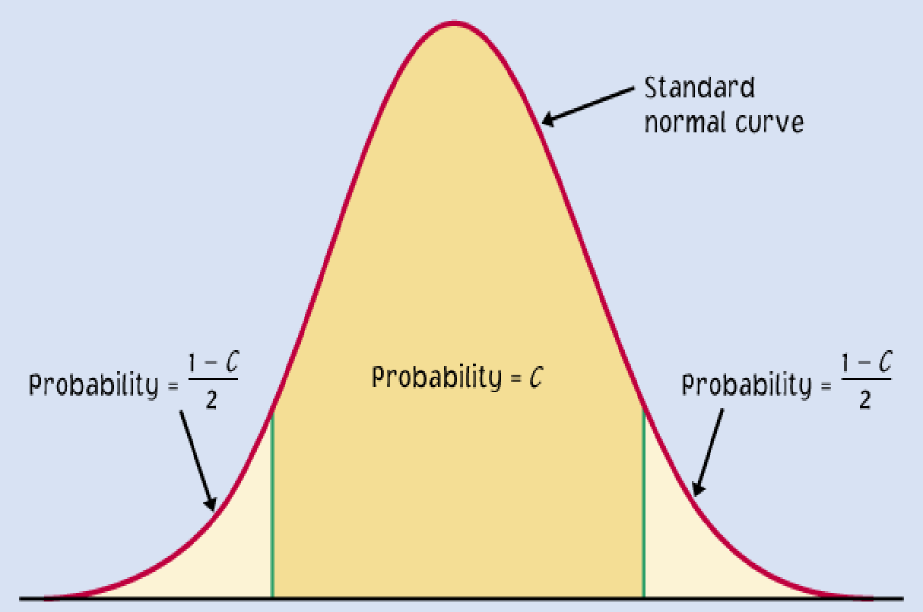
9. Interval Estimation and Confidence Intervals

A **confidence interval** can be expressed as:

- Mean ± *m*
  m is called the **margin of error**
  μ within $\overline{X} \pm m$

A **confidence level *C*** (in %) indicates the probability that the μ falls within the interval.

It represents the area under the normal curve within ± *m* of the center of the curve.



Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

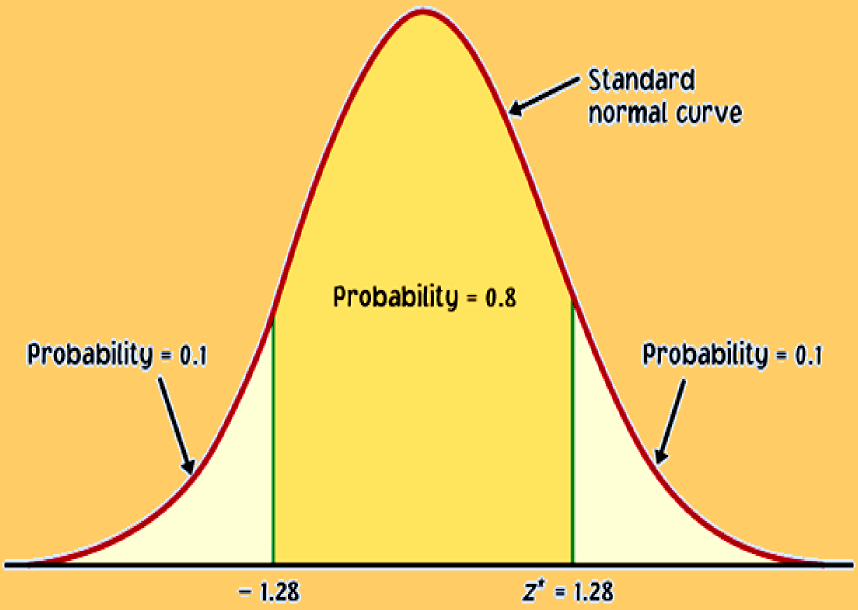
- We will now construct a level *C* confidence interval for the mean μ of a population when the data are a sample of size *n*.
- The construction is based on the ***sampling distribution of the sample mean***
- To construct a level *C* confidence interval we first identify the central *C* area under a Normal curve
- We must find the number z^* such that any Normal distribution has probability *C* within the $\pm z^*$ standard deviations of its mean
- All Normal distributions have the same standardized form. We can obtain everything we need from the same standard Normal curve

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

Practical use of *z*: *z**

- *z** is related to the chosen confidence level *C*.
- *C* is the area under the standard normal curve between $-z^*$ and z^*.



Example: For an 80% confidence level *C*, 80% of the normal curve's area is contained in the interval.

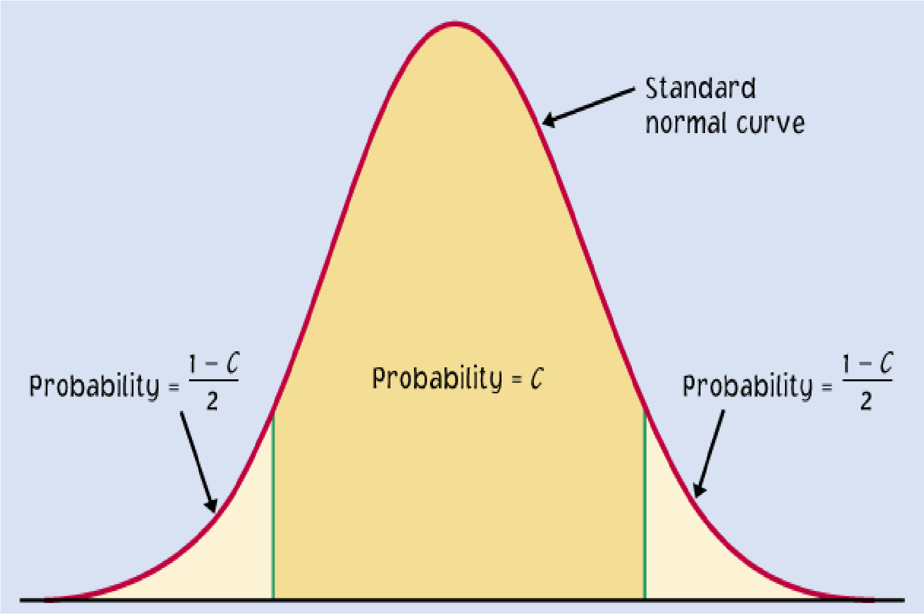
The confidence interval is thus:

$$\bar{x} \pm z^* \sigma / \sqrt{n}$$

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											



Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

- Let $\{Y_1, Y_2, \ldots\ldots, Y_n\}$ be a random sample from a population with a normal distribution with mean μ and variance σ^2: $Y_i \sim N(\mu, \sigma^2)$

  The distribution of the sample average will be: $\bar{Y} \sim N\left(\mu, \sigma^2/n\right)$

- Standardising: $\dfrac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

- Using what we know about the standard normal distribution we can construct a ***95% confidence interval***:

$$\Pr\left(-1.96 \leq \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

Topic 1: Statistical Review

9. Interval Estimation and Confidence Intervals

– Re-arranging:

$$\Pr\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

What if σ unknown?
An unbiased estimator of σ

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right]^{1/2} \qquad \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

95% confidence interval given by:

$$\left[\bar{Y} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{Y} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Topic 1: Statistical Review

– Example:

Given the sample data:

$$\bar{Y} = 40$$

$$s = 10$$

$$n = 36$$

Calculate the 99% confidence interval estimate of the true mean.

$$\left[\bar{Y} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{Y} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} \right]$$

$t_{n-1,\alpha/2}$ is the *critical value* from the t-distribution.

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365
33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307
41.	1.303	1.683	2.020	2.421	2.701	3.301
42.	1.302	1.682	2.018	2.418	2.698	3.296
43.	1.302	1.681	2.017	2.416	2.695	3.291
44.	1.301	1.680	2.015	2.414	2.692	3.286
45.	1.301	1.679	2.014	2.412	2.690	3.281

Topic 1: Statistical Review

10. Hypothesis Testing

- Hypothesis tests resolve conflicts between **two competing hypotheses**
- In any hypothesis test, we need to define:
  - H_0, **the null hypothesis**: the presumed default state of nature or status quo
  - H_A, **the alternative hypothesis**: a contradiction of the default state of nature or status quo
- We conduct hypothesis tests to determine if sample evidence contradicts H_0

10. Hypothesis Testing

On the basis of sample information, we either…

1. **“Reject the null hypothesis”**

– Sample evidence is inconsistent with H_0

2. **“Do not reject the null hypothesis”**

– Sample evidence is not inconsistent with H_0

We do not have enough evidence to “accept” H_0

– This is really important!

Topic 1: Statistical Review

10. Hypothesis Testing

- H_0, the null hypothesis, states the status quo
- H_A, the alternative hypothesis, states whatever we wish to establish, contesting the status quo

In a **two-tailed test**, H_0 can be reject on either size of its hypothesised value

Where the hypothesis test is about the population average (or proportion), this will be:

H_0: $\mu = \mu_0$ versus H_A: $\mu \neq \mu_0$

In a **one-tailed test**, H_0 can only be rejected on one side of the parameter's hypothesized value

Where the hypothesis test is about the population average, this will be:

H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$ (right-tail test)

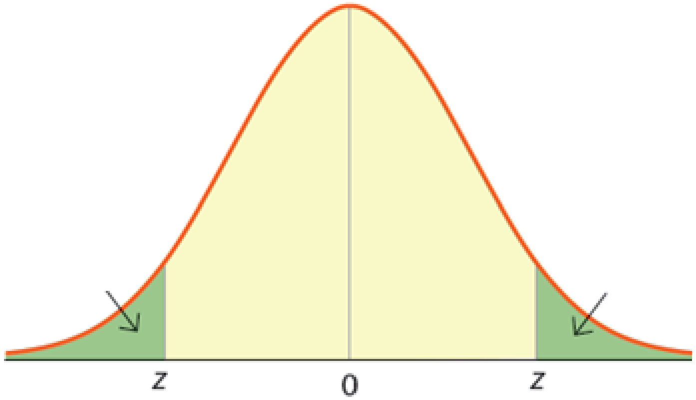
H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$ (left-tail test)

Topic 1: Statistical Review

10. Hypothesis Testing

Two-tail test

The "≠" symbol in H_A indicates that both tail areas of the distribution will be used to make the decision regarding the rejection of H_0



Topic 1: Statistical Review

10. Hypothesis Testing

One-tail test

In a one-tailed test, H_0 can only be rejected on one side of the parameter's hypothesized value

Where the hypothesis test is about the population average, this will be:

H_0: $\mu \leq \mu_0$ versus H_A: $\mu > \mu_0$ (right-tail test)

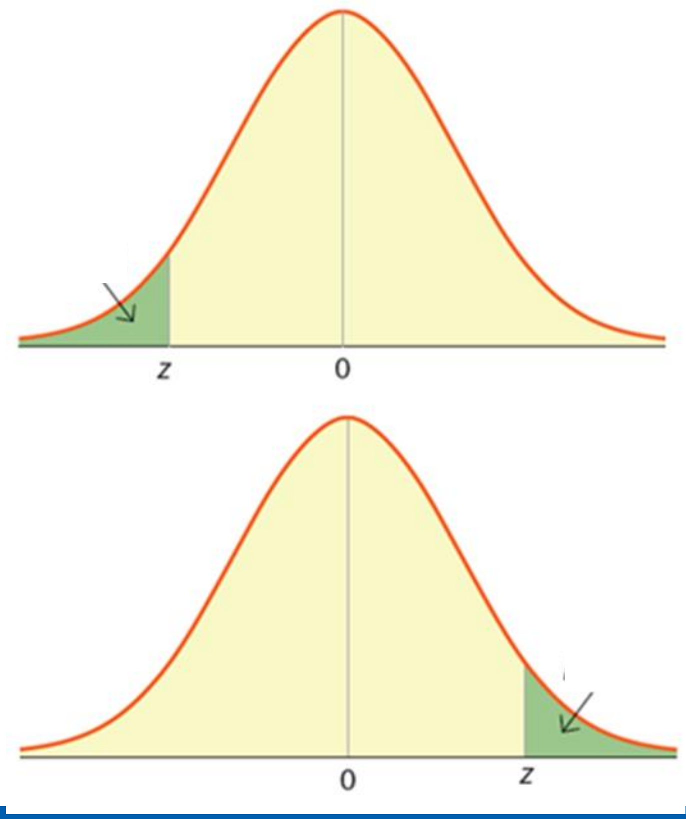
H_0: $\mu \geq \mu_0$ versus H_A: $\mu < \mu_0$ (left-tail test)

Topic 1: Statistical Review

10. Hypothesis Testing

One-tail test

Note that the inequality in H_A determines which tail area will be used to make the decision regarding the rejection of H_0



Topic 1: Statistical Review

10. Hypothesis Testing

A “Type I” error is the significance of the test

- Instances where we reject H_0 even though it is true
- We choose α, the level of significance – therefore we know how often a “Type I” error will occur

A “Type II” error is called the power of the test

- Where we fail to reject H_0 even though it is false
- Occurs with probability β – power of the test is $1-\beta$
- At a given level of significance, beta depends on the standard error ($\sigma/\sqrt{n}$)

Topic 1: Statistical Review

10. Hypothesis Testing

- Hypothesis: statement about a popn. developed for the purpose of testing
- Hypothesis testing: procedure based on sample evidence and probability theory to determine whether the hypothesis is a reasonable statement.
- Steps:

1. State the null (H_0) and alternate (H_A) hypotheses

   Note distinction between one and two-tailed tests

2. State the level of significance

   Probability of rejecting H_0 when it is true (*Type I Error*)

   Note: *Type II Error* – failing to reject H_0 when it is false

   Power of the test: *1-Pr(Type II error)*

3. Select a *test statistic*

   Based on sample information, follows a known distribution

4. Formulate *decision rule*

   Conditions under which null hypothesis is rejected. Based on *critical value* from known probability distribution.

5. Compute the value of the test statistic, make a decision, interpret the results.

Topic 1: Statistical Review

– Example:

Given the sample data:

$\bar{x} = 8.2 \qquad s = 23.9 \qquad n = 36$

Test the null hypothesis that the **population mean is equal to zero**, against an alternative hypothesis that the **population mean is positive**.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

ν	0.10	0.05	0.025	0.01	0.005	0.001
1.	3.078	6.314	12.706	31.821	63.657	318.313
2.	1.886	2.920	4.303	6.965	9.925	22.327
3.	1.638	2.353	3.182	4.541	5.841	10.215
4.	1.533	2.132	2.776	3.747	4.604	7.173
5.	1.476	2.015	2.571	3.365	4.032	5.893
6.	1.440	1.943	2.447	3.143	3.707	5.208
7.	1.415	1.895	2.365	2.998	3.499	4.782
8.	1.397	1.860	2.306	2.896	3.355	4.499
9.	1.383	1.833	2.262	2.821	3.250	4.296
10.	1.372	1.812	2.228	2.764	3.169	4.143
11.	1.363	1.796	2.201	2.718	3.106	4.024
12.	1.356	1.782	2.179	2.681	3.055	3.929
13.	1.350	1.771	2.160	2.650	3.012	3.852
14.	1.345	1.761	2.145	2.624	2.977	3.787
15.	1.341	1.753	2.131	2.602	2.947	3.733
16.	1.337	1.746	2.120	2.583	2.921	3.686
17.	1.333	1.740	2.110	2.567	2.898	3.646
18.	1.330	1.734	2.101	2.552	2.878	3.610
19.	1.328	1.729	2.093	2.539	2.861	3.579
20.	1.325	1.725	2.086	2.528	2.845	3.552
21.	1.323	1.721	2.080	2.518	2.831	3.527
22.	1.321	1.717	2.074	2.508	2.819	3.505
23.	1.319	1.714	2.069	2.500	2.807	3.485
24.	1.318	1.711	2.064	2.492	2.797	3.467
25.	1.316	1.708	2.060	2.485	2.787	3.450
26.	1.315	1.706	2.056	2.479	2.779	3.435
27.	1.314	1.703	2.052	2.473	2.771	3.421
28.	1.313	1.701	2.048	2.467	2.763	3.408
29.	1.311	1.699	2.045	2.462	2.756	3.396
30.	1.310	1.697	2.042	2.457	2.750	3.385
31.	1.309	1.696	2.040	2.453	2.744	3.375
32.	1.309	1.694	2.037	2.449	2.738	3.365
33.	1.308	1.692	2.035	2.445	2.733	3.356
34.	1.307	1.691	2.032	2.441	2.728	3.348
35.	1.306	1.690	2.030	2.438	2.724	3.340
36.	1.306	1.688	2.028	2.434	2.719	3.333
37.	1.305	1.687	2.026	2.431	2.715	3.326
38.	1.304	1.686	2.024	2.429	2.712	3.319
39.	1.304	1.685	2.023	2.426	2.708	3.313
40.	1.303	1.684	2.021	2.423	2.704	3.307
41.	1.303	1.683	2.020	2.421	2.701	3.301
42.	1.302	1.682	2.018	2.418	2.698	3.296
43.	1.302	1.681	2.017	2.416	2.695	3.291
44.	1.301	1.680	2.015	2.414	2.692	3.286
45.	1.301	1.679	2.014	2.412	2.690	3.281

Topic 1: Statistical Review

10. Hypothesis Testing

– P-value:

Alternative means of evaluating decision rule

Probability of observing a sample value as extreme as, or more extreme than the value observed when the null hypothesis is true

- If the p-value is greater than the significance level, H_0 is not rejected
- If the p-value is less than the significance level, H_0 is rejected

If the p-value is less than:

0.10, we have some evidence that H_0 is not true

0.05 we have strong evidence that H_0 is not true

0.01 we have very strong evidence that H_0 is not true

Topic 2:
The Linear Regression Model

Topic 2: The Linear Regression Model

1. Simple Regression Model

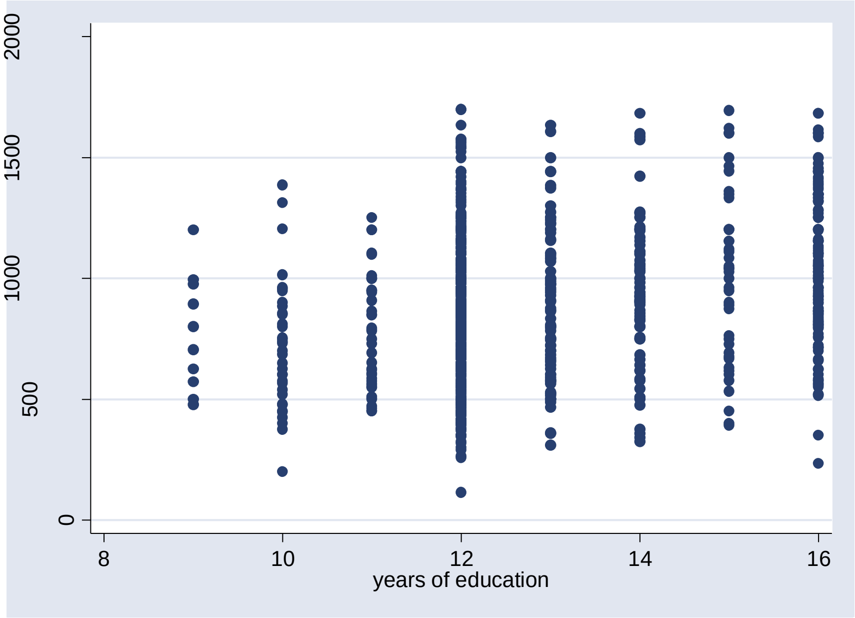
Regression analysis is concerned with the study of the dependence of one variable (*the dependent variable*) on one or more other variables (the explanatory variables) with a view to estimating or predicting the population mean – average value of the dependent variable in terms of the known values of the independent variables.

Bivariate Example: Explaining an individual's average wages given the individual's education level.

Topic 2: The Linear Regression Model

1. Simple Regression Model

Scattergram of distribution of wages corresponding to fixed education levels



Note: Variability in wages for each education level

Despite variability, average wages increase as education level increases

Plotting mean wage for each given education level gives the regression line

Topic 2: The Linear Regression Model

1. Simple Regression Model

The population model

- Mean of Y for a given X is known as the *conditional expected value* *E(Y|X)*

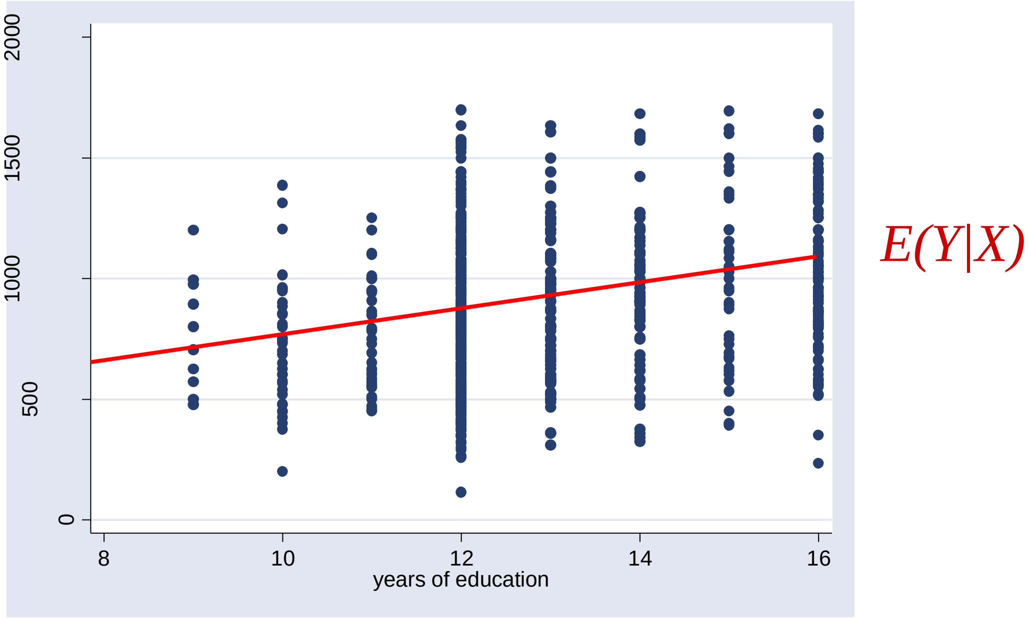
- Note: The *unconditional expected value, E(Y),* is just the mean of the population

- The *population regression is the locus of the conditional means of the dependent variable for the fixed values of the explanatory variables*

Topic 2: The Linear Regression Model

1. Simple Regression Model

Scattergram of distribution of wages corresponding to fixed education levels



Note: Variability in wages for each education level

Despite variability, average wages increase as education level increases

Plotting mean wage for each given education level gives the regression line

Topic 2: The Linear Regression Model

1. Simple Regression Model

The population model

Mean of Y for a given X is known as the *conditional expected value E(Y|X)*

Note: The *unconditional expected value, E(Y),* is just the mean of the population

The population regression is the locus of the conditional means of the dependent variable for the fixed values of the explanatory variables

Population regression function:

$$E(Y|X_i) = f(X_i)$$

Topic 2: The Linear Regression Model

1. Simple Regression Model

The population model

Assume *linear functional form*:

$$E(Y|X_i) = \beta_0 + \beta_1 X_i$$

β_0 : intercept term or constant

β_1: slope coefficient - quantifies the linear relationship between X and Y

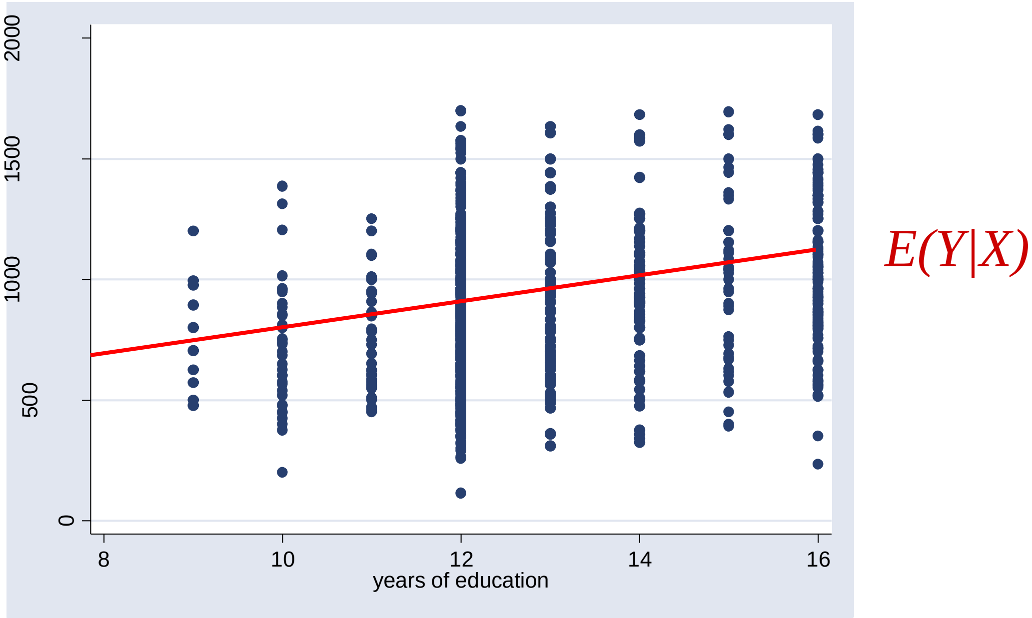
Fixed parameters known as regression coefficients

For each X_i, individual observations will vary around $E(Y|X_i)$

Topic 2: The Linear Regression Model

1. Simple Regression Model

Scattergram of distribution of wages corresponding to fixed education levels



Note: Variability in wages for each education level

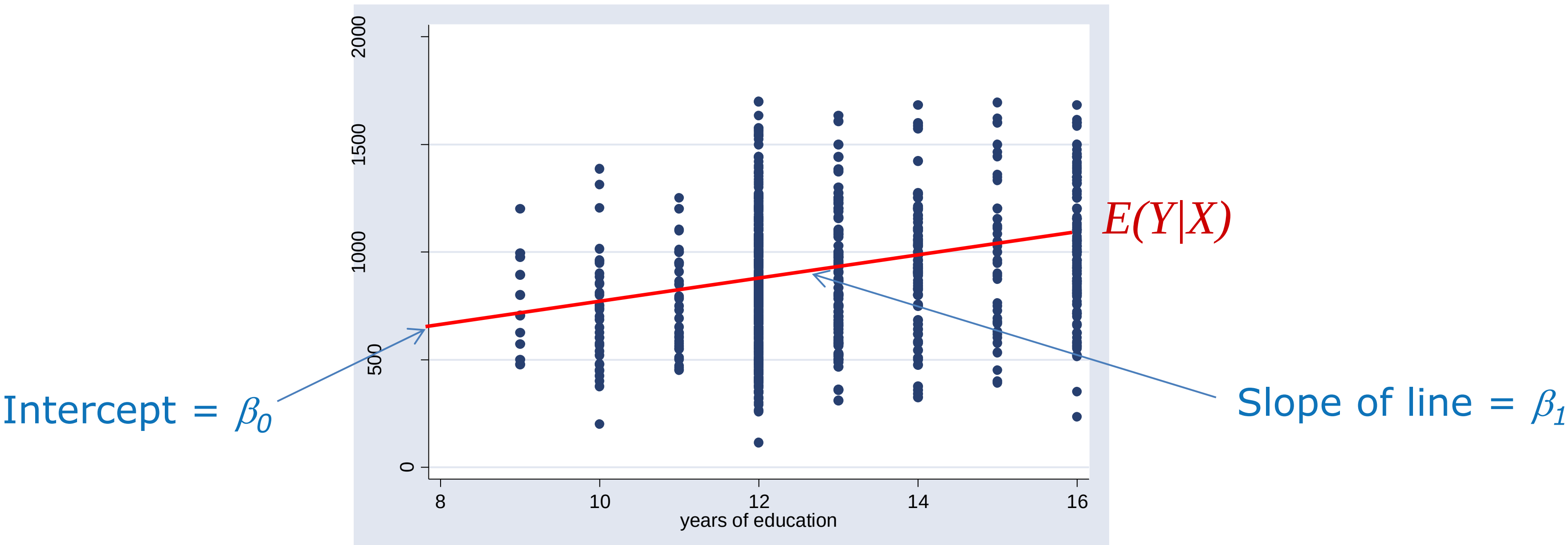
Despite variability, average wages increase as education level increases

Plotting mean wage for each given education level gives the regression line

Topic 2: The Linear Regression Model

1. Simple Regression Model

Scattergram of distribution of wages corresponding to fixed education levels



Note: Variability in wages for each education level

Despite variability, average wages increase as education level increases

Plotting mean wage for each given education level gives the regression line

Topic 2: The Linear Regression Model

1. Simple Regression Model

The population model

Assume linear functional form:

$$E(Y|X_i) = \beta_0+\beta_1X_i$$

β_0: intercept term or constant

β_1: slope coefficient - quantifies the linear relationship between *X* and *Y*

Fixed parameters known as *regression coefficients*

For each X_i, individual observations will vary around $E(Y|X_i)$

Consider *deviation* of any individual observation from conditional mean:

$$u_i = Y_i - E(Y|X_i)$$

u_i : stochastic disturbance/error term – unobservable random deviation of an observation from its conditional mean

Topic 2: The Linear Regression Model

1. Simple Regression Model

The linear regression model

Re-arrange previous equation to get:

$$Y_i = E(Y|X_i) + u_i$$

Each individual observation on Y can be explained in terms of:

- *$E(Y|X_i)$:* mean Y of all individuals with same level of X – systematic or deterministic component of the model – the part of Y explained by X
- *u_i*: random or non-systematic component – includes all omitted variables that can affect Y

Assuming a *linear functional form*:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Topic 2: The Linear Regression Model

1. Simple Regression Model

A note on linearity: Linear in parameters vs. linear in variables

The following is linear in parameters but not in variables:

$$Y_i = \beta_0 + \beta_1 X_i^2 + u_i$$

In some cases transformations are required to make a model linear in parameters

Topic 2: The Linear Regression Model

1. Simple Regression Model

The linear regression model

$$Y_i = \beta_0+\beta_1 X_i + u_i$$

Represents relationship between *Y* and *X* in *population* of data

Using appropriate estimation techniques we **use sample data to estimate values for β_0 and β_1**

β_1: measures *ceteris paribus* effect of *X* on *Y* only if all other factors are fixed and do not change.

Assume u_i fixed so that $\Delta u_i = 0$, then

$$\Delta Y_i = \beta_1 \Delta X_i$$

$$\Delta Y_i / \Delta X_i = \beta_1$$

Unknown u_i – requires assumptions about u_i to estimate *ceteris paribus* relationship

Topic 2: The Linear Regression Model

1. Simple Regression Model

The linear regression model: Assumptions about the error term

- Assume $E(u_i) = 0$: On average the unobservable factors that deviate an individual observation from the mean are zero

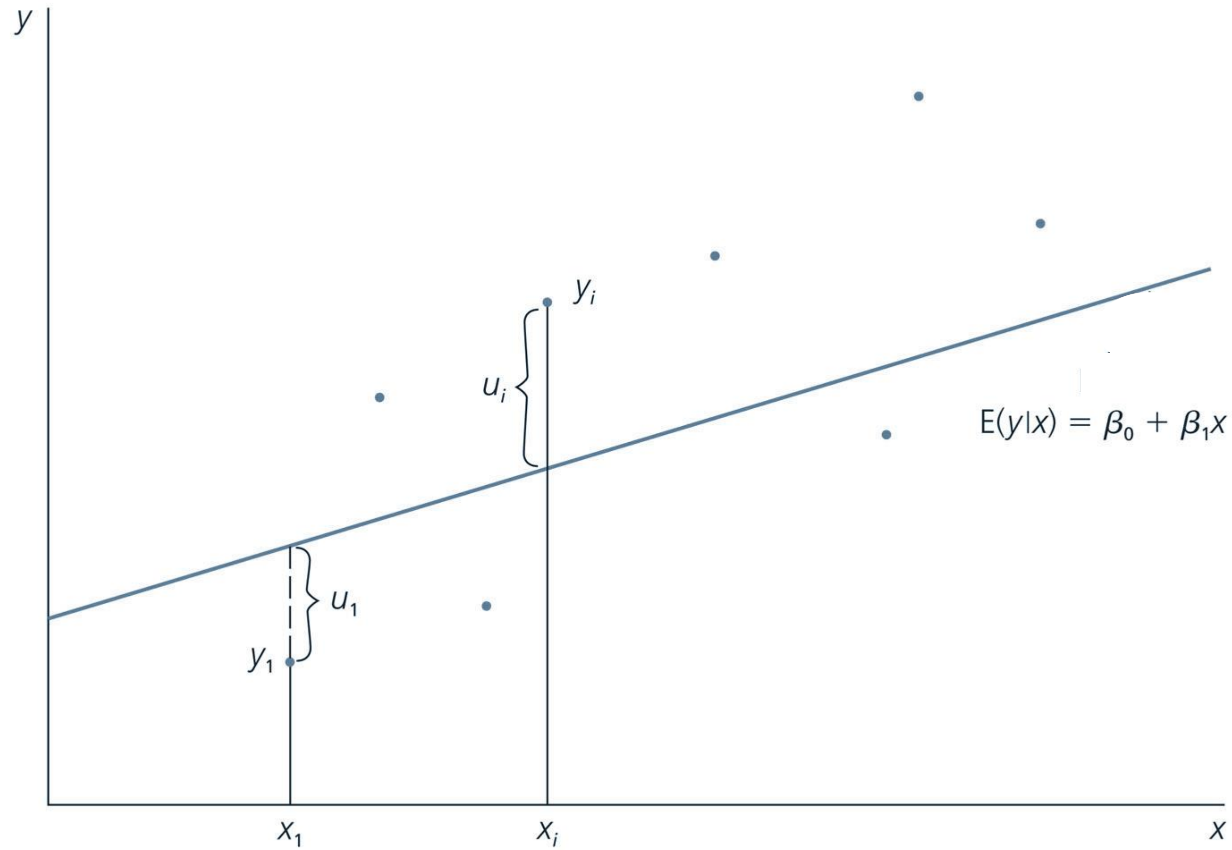
- Assume $E(u_i|X_i) = 0$: mean of u_i conditional on X_i is zero – regardless of what values X_i takes, the unobservables are on average zero

- ***Zero Conditional Mean Assumption***:

$$E(u_i|X_i) = E(u_i) = 0$$

FIGURE 2.7

Graph of $y_i = \beta_0 + \beta_1 x_i + u_i$.



Topic 2: The Linear Regression Model

1. Simple Regression Model

The linear regression model: Notes on the error term

Reasons why an error term will always be required:

- Vagueness of theory
- Unavailability of data
- Measurement error
- Incorrect functional form
- Principle of Parsimony

Topic 2: The Linear Regression Model

1. Simple Regression Model

Regression vs. Correlation

- ***Correlation analysis***: measures the strength or degree of linear association between two random variables
- ***Regression analysis***: estimating the average values of one variable on the basis of the fixed values of the other variables for the purpose of prediction.

Topic 2: The Linear Regression Model

2. Ordinary Least Squares (OLS) Estimation

Estimate the population relationship given by

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

using a random sample of data *i=1,….n*

Least Squares Principle: Minimise the sum of the squared deviations between the actual and the predicted (or fitted) values.

Define the fitted values as $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

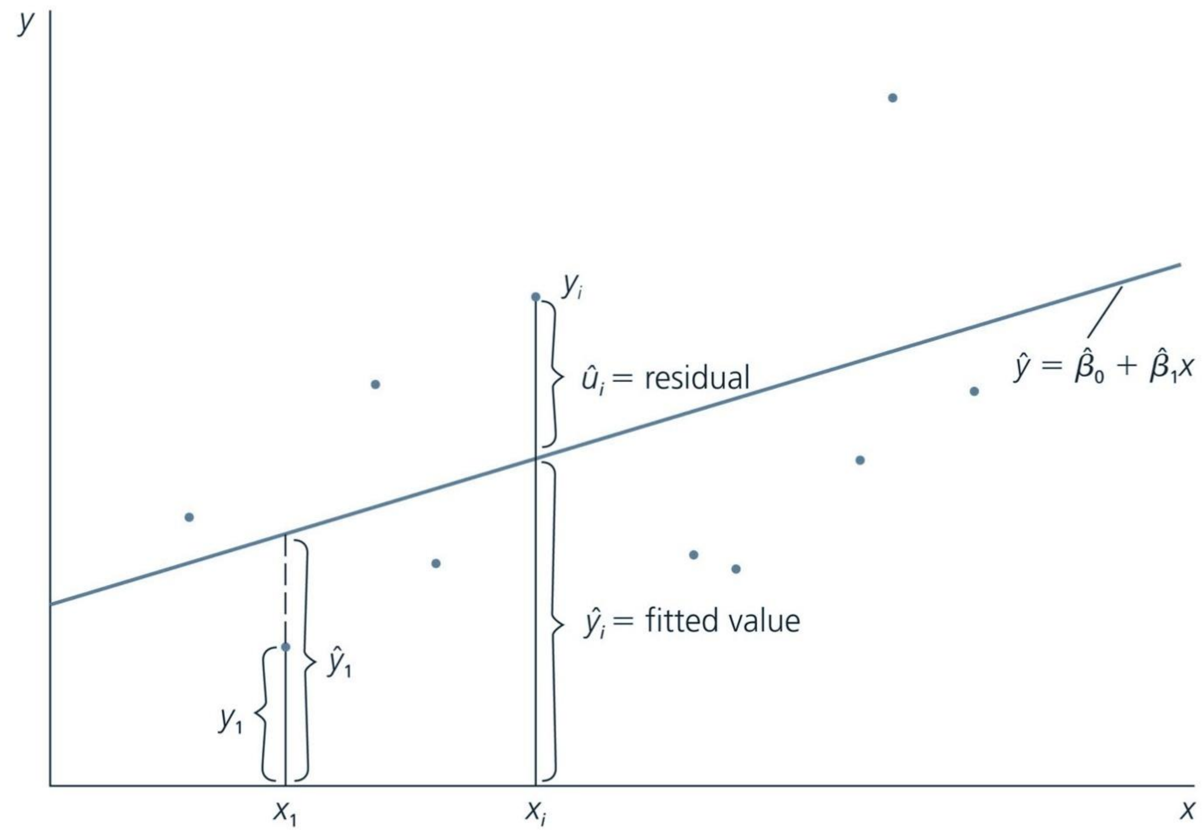
OLS minimises $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$

Solving this optimisation problem yields:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

FIGURE 2.4

Fitted values and residuals.



Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Gauss-Markov Theorem

Under the assumptions of the Classical Linear Regression Model the OLS estimator will be the Best Linear Unbiased Estimator

Linear: estimator is a linear function of a random variable
Unbiased:

$$E\left(\hat{\beta}_0\right) = \beta_0$$

$$E\left(\hat{\beta}_1\right) = \beta_1$$

Best: estimator is most efficient estimator, i.e., estimator has the minimum variance of all linear unbiased estimators

For a robust analysis our estimator must exhibit these properties

What assumptions are required?

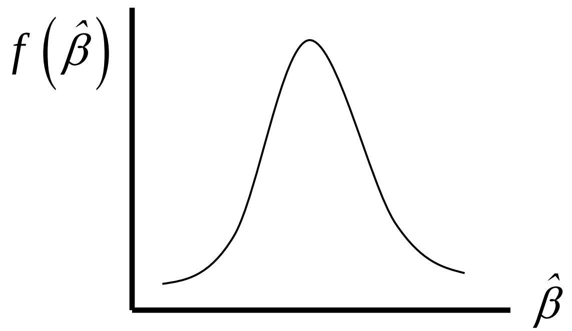
Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

It is important to remember that we use econometrics to estimate population relationships using sample data

For each sample drawn from a population we might expect a different point estimate

The distribution of all possible point estimates is known as the sampling distribution



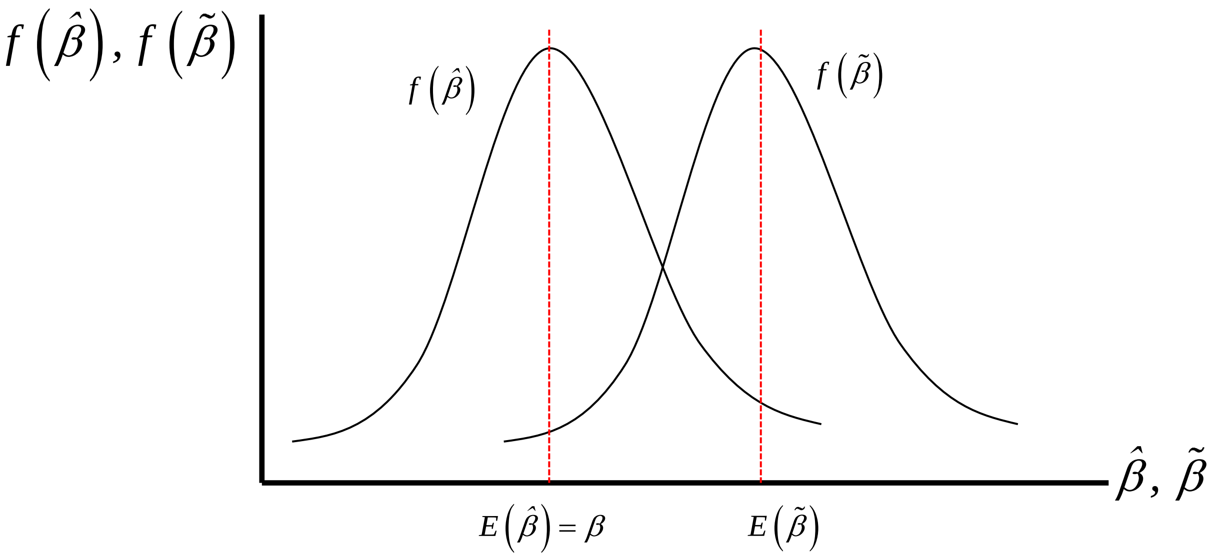
To determine how good an estimator is (e.g. the OLS estimator) we look at moments of the sampling distribution of the estimator (mean and variance)

Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Unbiasedness

An estimator is unbiased if its expected value is equal to its true population value - i.e. on average the estimator is correct



Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Assumptions required to prove unbiasedness:

A1: Regression model is linear in parameters

A2: X are non-stochastic or fixed in repeated sampling

A3: Zero conditional mean

A4: Sample is random

A5: Variability in the Xs

Note: Must be happy to assume that the error term is not correlated with any of the X variables in the model

Topic 2: Regression Models

3. Properties of OLS Estimator

Efficiency

What about the dispersion of the distribution of the estimator?

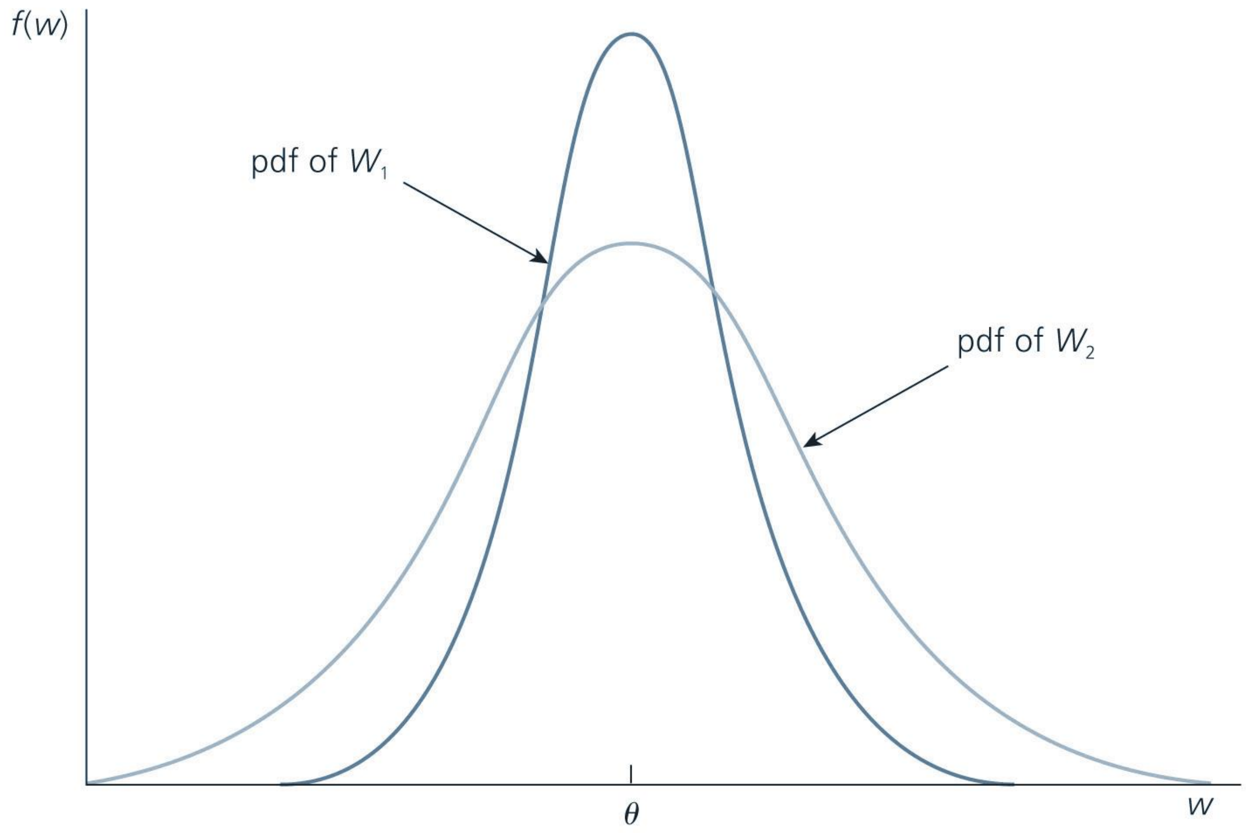
i.e, how likely is it that the estimate is close to the true parameter?

Useful summary measure for the dispersion in the distribution is the *sampling variance*.

An efficient estimator is one which has the least amount of dispersion about its true value i.e. the one that has the smallest sampling variance

FIGURE C.2

The sampling distributions of two unbiased estimators of θ.



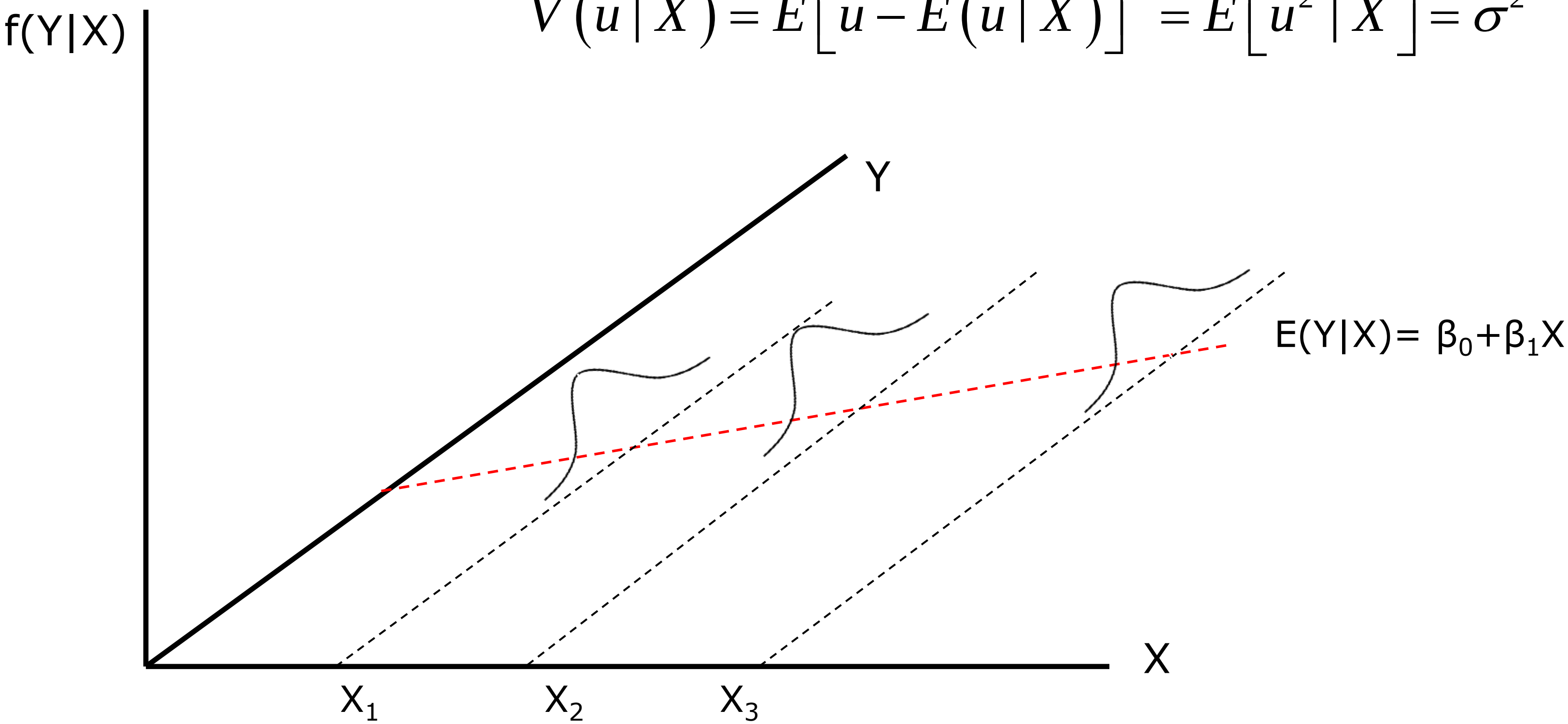
Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Assumptions required to prove efficiency:

A6: Homoscedasticity

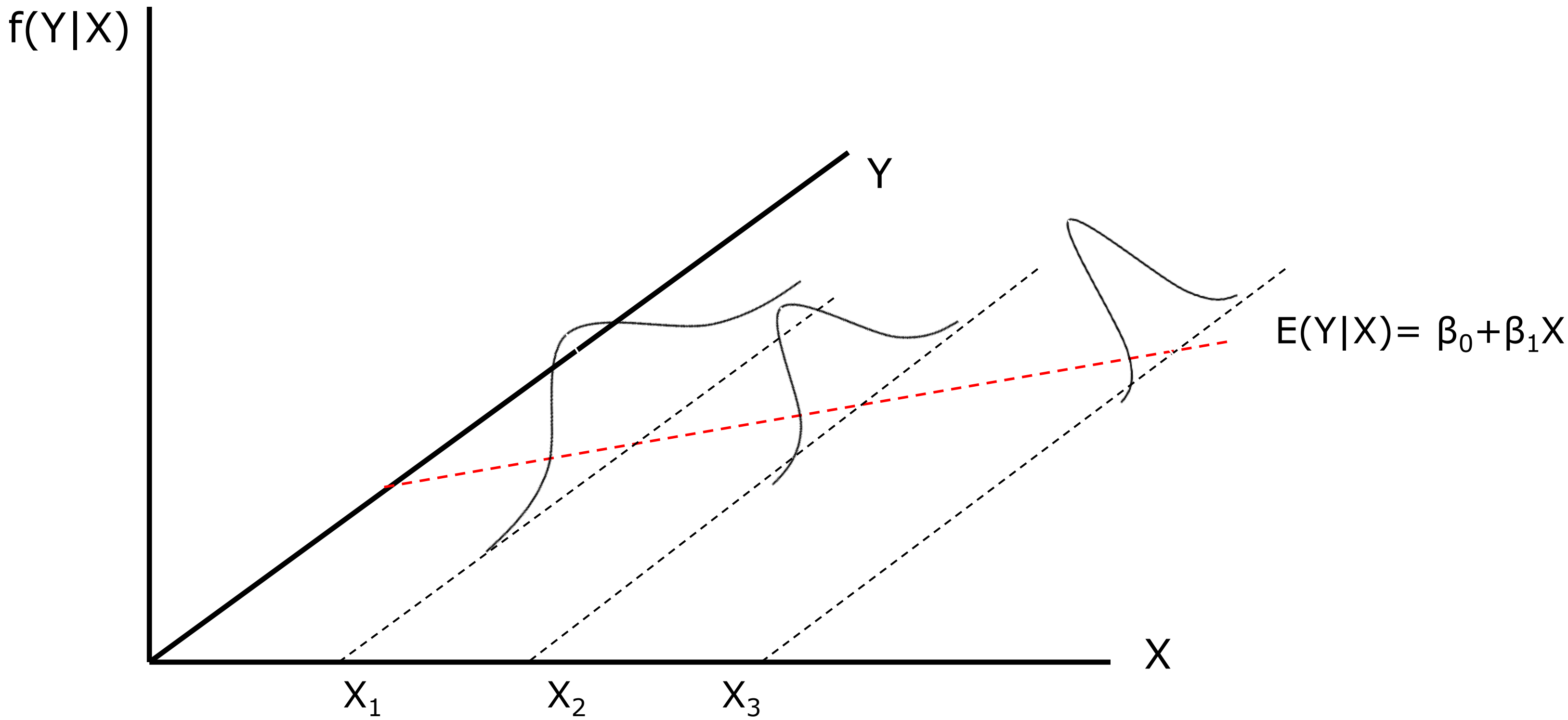
$$V\left(u \mid X\right) = E\left[u - E\left(u \mid X\right)\right]^2 = E\left[u^2 \mid X\right] = \sigma^2$$



Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Heteroscedasticity: $V(u \mid X) = \sigma_i^2$



Topic 2: The Linear Regression Model

3. Properties of OLS Estimator

Assumptions required to prove efficiency:

A6: Homoscedasticity

A7: No autocorrelation or spatial correlation

$$Cov\left(u_i, u_j \mid X_i, X_j\right) = E\left(\left[u_i - E\left(u_i\right) \mid X_i\right]\left[u_j - E\left(u_j\right) \mid X_j\right]\right) = E\left(\left[u_i \mid X_i\right]\left[u_j \mid X_j\right]\right) = 0$$

Topic 2: The Linear Regression Model

4. Goodness of Fit

How well does regression line 'fit' the observations?

– R^2 (coefficient of determination) measures the **proportion of the sample variance of Y_i explained by the model** where variation is measured as squared deviation from sample mean.

– This measure will be bound by zero and one where there is an intercept in the model

Topic 2: The Linear Regression Model

4. Goodness of Fit

How well does regression line ‘fit’ the observations?

Total Sum of Squares: $\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$

Explained Sum of Squares: $\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$

Residual Sum of Squares: $\sum_{i=1}^{n}\hat{u}_i^2$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

Topic 2: The Linear Regression Model

5. Interpretation of coefficients & units of measurement

Estimate impact that average return on equity (%) has on salary of CEOs (in thousands of euros)

$$\hat{salary}_i = 963.191 + 18.501 ROE_i$$

β_0 = 963.191 $\Rightarrow$ when *ROE* = 0, predicted *salary* = 963.191

Interpret as €963,161

β_1 = 18.501 $\Rightarrow$ when ΔROE = 1, Δ predicted *salary* = 18.501

Interpret as €18,501

$$\hat{salary}_i = 963.191 + 18.501(20) = 1{,}333.191$$

Use equation to compared predicted salaries for different *ROE*s, e.g. if *ROE* = 20:

Interpret as €1,333,191

Note: Importance of units of measurement in interpretation of results

Topic 2: The Linear Regression Model

6. Regression model with two independent variables

Say we have information on more variables that may influence *Y*:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

β_0 : measures the average value of *Y* when X_1 and X_2 are zero

β_1 and β_2 are the partial regression coefficients/slope coefficients which measure the ceteris paribus effect of X_1 and X_2 on *Y*, respectively

Key assumption: $E\left(u \mid X_1, X_2\right) = 0$

This can be extended to any number of independent variables as long as the number of observations in the sample exceeds the number of variables

Note – for accuracy of the estimator ***the number of observations should greatly exceed the number of variables***!

Model can be estimated using OLS in the same way as the simple regression case

Topic 2: The Linear Regression Model

6. Regression model with two independent variables

OLS slope coefficients depend on the relationship between each of the individual variables and *Y* and on the relationship between the *X*'s

$$productivity = \beta_0 + \beta_1 education\_head + \beta_2 land\_quality + u$$

Where k=2, $\hat{\beta}_1$ gives the pure effect of X_1 on Y, netting out the effect of X_2.

For example:

$\hat{\beta}_1$ is the effect of education on productivity holding the quality of land constant

Topic 2: The Linear Regression Model

7. Functional Form

Incorporate non-linearity into the model

Regress productivity (measured in kg per ha) on years of schooling:

$$\hat{prod}_i = 0.413 + 0.025 educ_i$$

$\Rightarrow$ same return of β_1 = 0.025 (0.025 kg per ha) for each additional year of schooling

$$ln\left(\hat{prod}_i\right) = -1.02 + 0.074 educ_i$$

Regress ln($prod_i$) on years of schooling:

$$\%\,\Delta\, prod_i \approx (100 * 0.074) \Delta educ_i$$

Topic 2: The Linear Regression Model

7. Functional Form

In general:

1) If we estimate

$$lnY = \beta_0 + \beta_1 X + u$$

100* Δ ln$Y/\Delta X$ = 100* $\hat{\beta}_1$

Percentage change in Y as a result of a one unit change in X

2) If we estimate $\ln Y_i = \beta_0 + \beta_1 \ln X_i + u$

Δ lnY/Δln$X$ = $\hat{\beta}_1$

Percentage change in Y as a result of a one unit change in X

Topic 2: The Linear Regression Model

8. Dummy Variables

- Dummy variables assume 0 and 1 values and are used to indicate the presence of an attribute
- For example: male or female
- Categorical variables have more than one category – for example region (north, south, east, west) or gender (male, female)
- If a qualitative variables has m categories introduce m-1 dummy variables
- The excluded category is called the *base* category

<u>Example:</u>

$$prod = \beta_0 + \beta_1 female + \beta_2 educ + u$$

$$\beta_1 = E\left(prod \mid female, educ\right) - E\left(prod \mid male, educ\right)$$

8. Dummy Variables

Interacting dummy variables

Interacting dummy variables is a very powerful way of understanding relevant variables while controlling for underlying characteristics

Example:

$$prod = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 female * married + \beta_4 educ + u$$

Holding Education constant:

Female = 0, Married = 0: average productivity of single men $\hat{\beta}_0$

Female = 0, Married = 1: average productivity of married men $\hat{\beta}_0 + \hat{\beta}_2$

Female = 1, Married = 0: average productivity of single females $\hat{\beta}_0 + \hat{\beta}_1$

Female = 1, Married = 1: average productivity of married females $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Topic 2: The Linear Regression Model

9. Model specification

Inclusion of irrelevant variables:

- OLS estimator unbiased but with higher variance if *X*'s correlated

Exclusion of relevant variables:

- Omitted variable bias if variables correlated with variables included in the estimated model
- Biased and inconsistent estimates prevents causal relationship from being identified

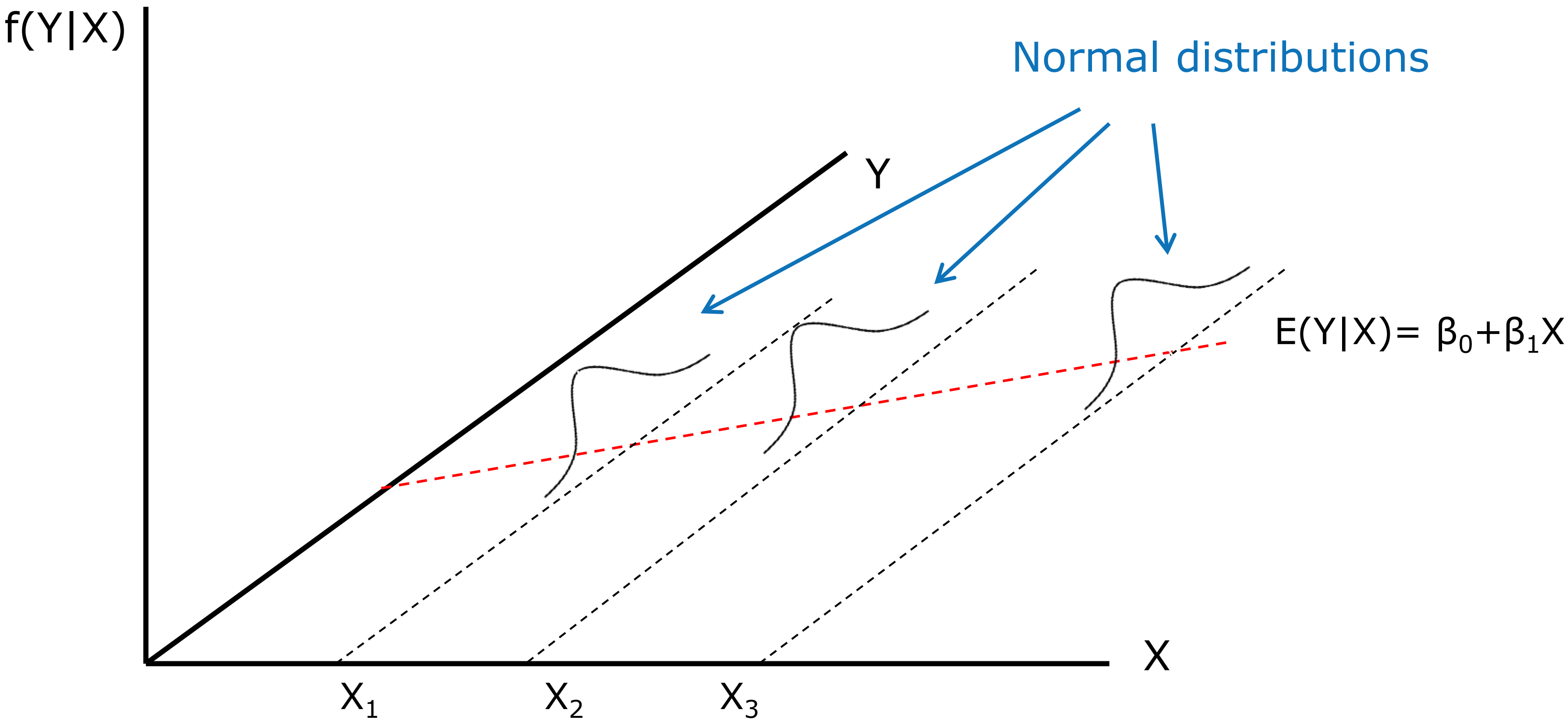
Topic 3: Statistical Inference

Topic3: Statistical Inference

Assuming *u* normally distributed we can say that the sampling distribution of $\hat{\beta}$ will also be normally distributed

Assumptions about the error term can be summarized in:

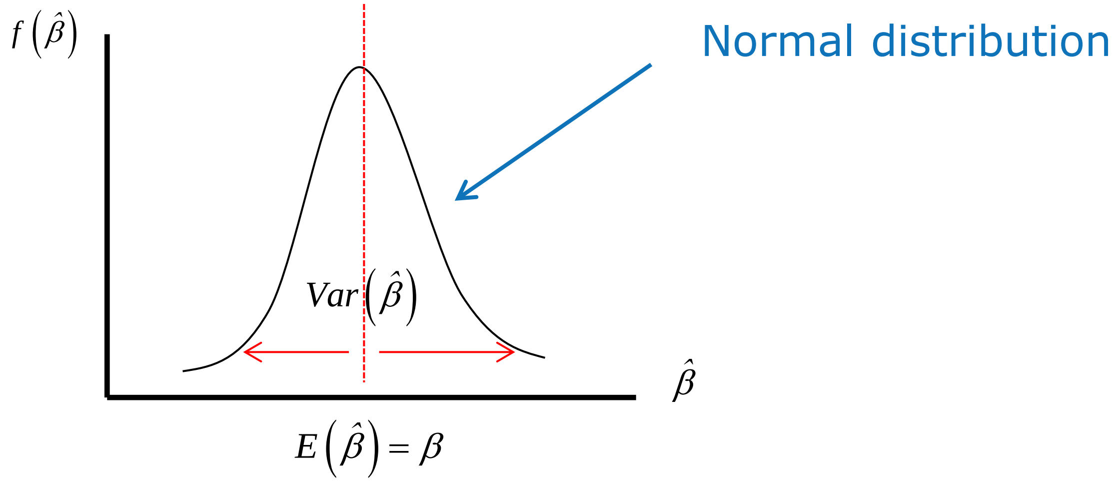
$$u \sim N\left(0, \sigma^2\right)$$



Topic 3: Statistical Inference

The exact sampling distribution of $\hat{\beta}$ will be

$$\hat{\beta} \sim N\left(\beta, Var\left(\hat{\beta}\right)\right)$$



Once we know the exact sampling distribution we can standardize to get a statistic which we know follows a standard normal distribution:

$$\frac{\hat{\beta} - \beta}{\sqrt{Var\left(\hat{\beta}\right)}} \sim N\left(0,1\right)$$

Topic 3: Statistical Inference

However, we need to know dispersion (variance) of sampling distribution of OLS estimator in order to perform statistical tests

In multiple regression model: $V\left(\hat{\beta}_k\right)=\dfrac{\sigma^2}{\sum\left(X_i-\bar{X}\right)^2\left(1-R_k^2\right)}$

Depends on:

a) σ^2: the error variance (reduces accuracy of estimates)

b) $\sum\left(X_i-\bar{X}\right)^2$: variation in X (increases accuracy of estimates)

c) $R^2{}_k$: the coefficient of determination from a regression of X_k on all other independent variables (degree of *multicollinearity* reduces accuracy of estimates)

What about the variance of the error terms σ^2?

Estimate using:

$$\hat{\sigma}^2=\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2$$

Test statistic become t-test statistic: $\dfrac{\hat{\beta}-\beta}{se\left(\hat{\beta}\right)}\sim t_{n-k-1}$

Topic 3: Statistical Inference

Hypothesis testing about a single population parameter

- Assume the following population model follows all CLM assumptions

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + u$$

- OLS produces unbiased estimates but how accurate are they?
- *Test by constructing hypotheses about population parameters and using sample estimates and statistical theory to test whether hypotheses are true*
- In particular, we are interested in testing whether population parameters significantly differ from zero: $H_0 : \beta_k = 0$

Statistical theory tells us that the statistic: $\dfrac{\hat{\beta}_k - \beta_k}{se\left(\hat{\beta}_k\right)}$ follows a *t distribution*

Which under the null is:

$$t_{\hat{\beta}_k} = \frac{\hat{\beta}_k - 0}{se\left(\hat{\beta}_k\right)} = \frac{\hat{\beta}_k}{se\left(\hat{\beta}_k\right)} \sim t_{n-k-1}$$

Topic 3: Statistical Inference

Hypothesis testing about a single population parameter

Two-sided alternative hypothesis

$$H_A : \beta_k \neq 0$$

Large positive and negative values of computed test statistic inconsistent with null

Reject null if $\left| t_{\hat{\beta}_k} \right| > c$

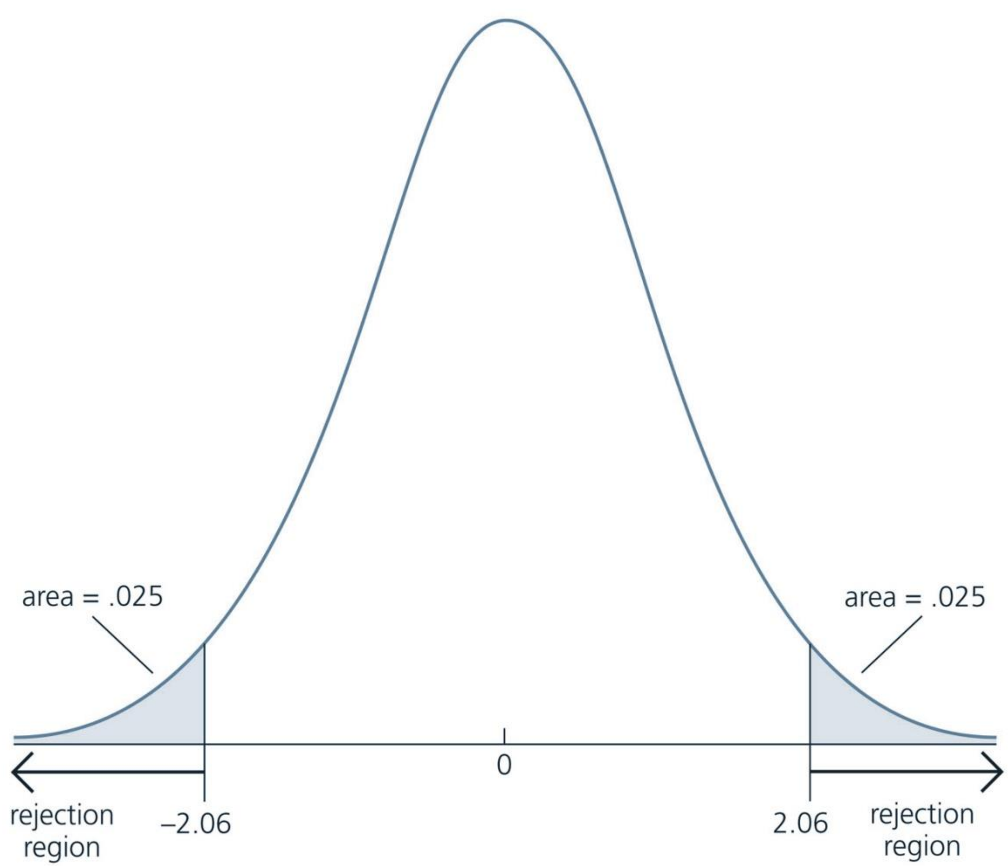
<u>Example:</u>

$$H_0 : \beta_k = 0$$
$$H_A : \beta_k \neq 0$$
$$df = 25 \qquad \alpha = 0.05$$
$$\text{threshold is anywhere above or below}$$
$$\mathrm{c} = 2.06$$

Note: ***If null rejected variable is said to be 'statistically significant' at the chosen significance level***

FIGURE 4.4

5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.



Topic 3: Statistical Inference

Hypothesis testing about a single population parameter

P-value approach:

Given the computed t-statistic, what is the smallest significance level at which the null hypothesis would be rejected?

P-values below 0.05 provide strong evidence against the null

For two sided alternative p-value is given by:

$$P\left(|T| > \left|t_{\hat{\beta}_k}\right|\right) = 2 * P\left(T > \left|t_{\hat{\beta}_k}\right|\right)$$

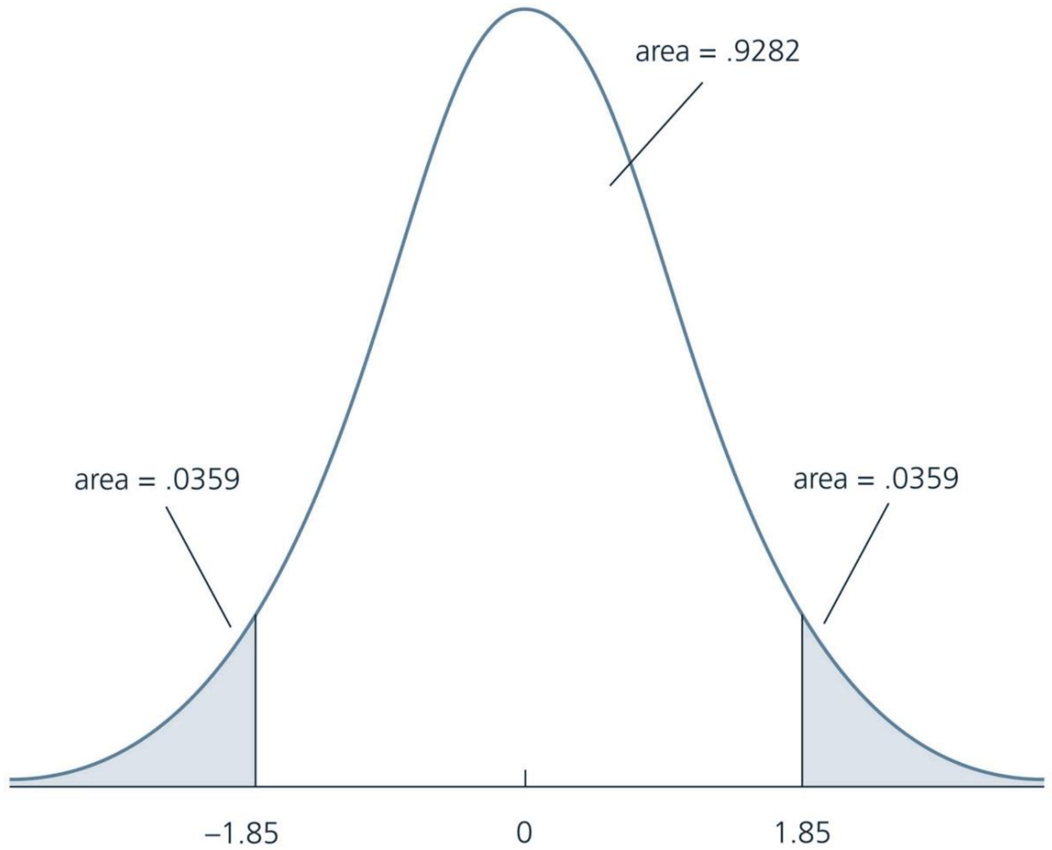
<u>Example</u>

$$H_0 : \beta_k = 0$$
$$H_A : \beta_k \neq 0$$
$$df = 40 \qquad t_{\hat{\beta}_k} = 1.85$$
$$P\left(|T| > \left|t_{\hat{\beta}_k}\right|\right) = P\left(|T| > 1.85\right) = 2 * P\left(T > 1.85\right)$$
$$= 2 * \left(0.0359\right) = 0.0718$$

This means that if the null hypothesis is true, we will observe an absolute value of the t statistic as large as 1.85 about 7.2% of the time.

FIGURE 4.6

Obtaining the p-value against a two-sided alternative, when $t = 1.85$ and $df = 40$.



Topic 3: Statistical Inference
Testing hypothesis about multiple linear restrictions

Consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u$$

We wish to test whether X_3 , X_4 and X_5 should be excluded:

$$H_0 : \beta_3 = 0, \beta_4 = 0, \beta_5 = 0$$
$$H_A : \text{ Not } H_0$$

Approach:

Estimate unrestricted and restricted model

Compare $SSR = \sum \hat{u}_i^2$ or R^2

$$F \equiv \frac{(SSR_r - SSR_{ur})/J}{SSR_{ur}/(n-k-1)} \sim F_{J,n-k-1}$$

$$F \equiv \frac{(R_{ur}^2 - R_r^2)/J}{(1-R_{ur}^2)/(n-k-1)} \sim F_{J,n-k-1}$$

Large values inconsistent with null

Topic 3: Statistical Inference

Testing hypothesis about multiple linear restrictions

Decision rule:

Compare to critical value from F distribution with J and n-k-1 degrees of freedom.

Reject null if $F > F_{J,n\text{-}k\text{-}1}$

P-value:

Smallest significance level at which the null hypothesis would be rejected.

$$P\left\{F > F_{J,n-k-1}\right\}$$

The smaller the p-value the more evidence we have against the null hypothesis

Topic 3: Statistical Inference

Overall test for significance of the Regression

General model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

Test of null hypothesis that all variables except intercept insignificant:

$$H_0 : \beta_1 = 0, \beta_2 = 0, ..., \beta_k = 0$$

Test statistic:

$$F \equiv \frac{\left(R^2\right)/k}{\left(1 - R^2\right)/\left(n - k - 1\right)} \sim F_{k,n-k-1}$$

Large values inconsistent with null

$$R^2 = R^2_{ur} \qquad R^2_r = 0$$

Topic 3: Statistical Inference

- Statistical inference requires that the distributional assumptions about the error terms hold
- This is needed to make sure that the standard errors of the OLS estimator are computed correctly
- Recall the assumptions required to prove efficiency:

A6: Homoscedasticity

$$V\left(u \mid X\right) = E\left[u - E\left(u \mid X\right)\right]^2 = E\left[u^2 \mid X\right] = \sigma^2$$

A7: No autocorrelation or spatial correlation

$$Cov\left(u_i, u_j \mid X_i, X_j\right) = E\left(\left[u_i - E\left(u_i\right) \mid X_i\right]\left[u_j - E\left(u_j\right) \mid X_j\right]\right) = E\left(\left[u_i \mid X_i\right]\left[u_j \mid X_j\right]\right) = 0$$

Contact details

- Do not hesitate to contact me in case you need further information/clarification.

- Email: narcisog@tcd.ie

Lab session

The lab session will take place in room AP0.12

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Thank you!