INDIAN INSTITUTE OF TECHNOLOGY KANPUR

Week 11
Artificial Intelligence (AI) for Investments

Lesson 1: Regression Modelling

Introduction

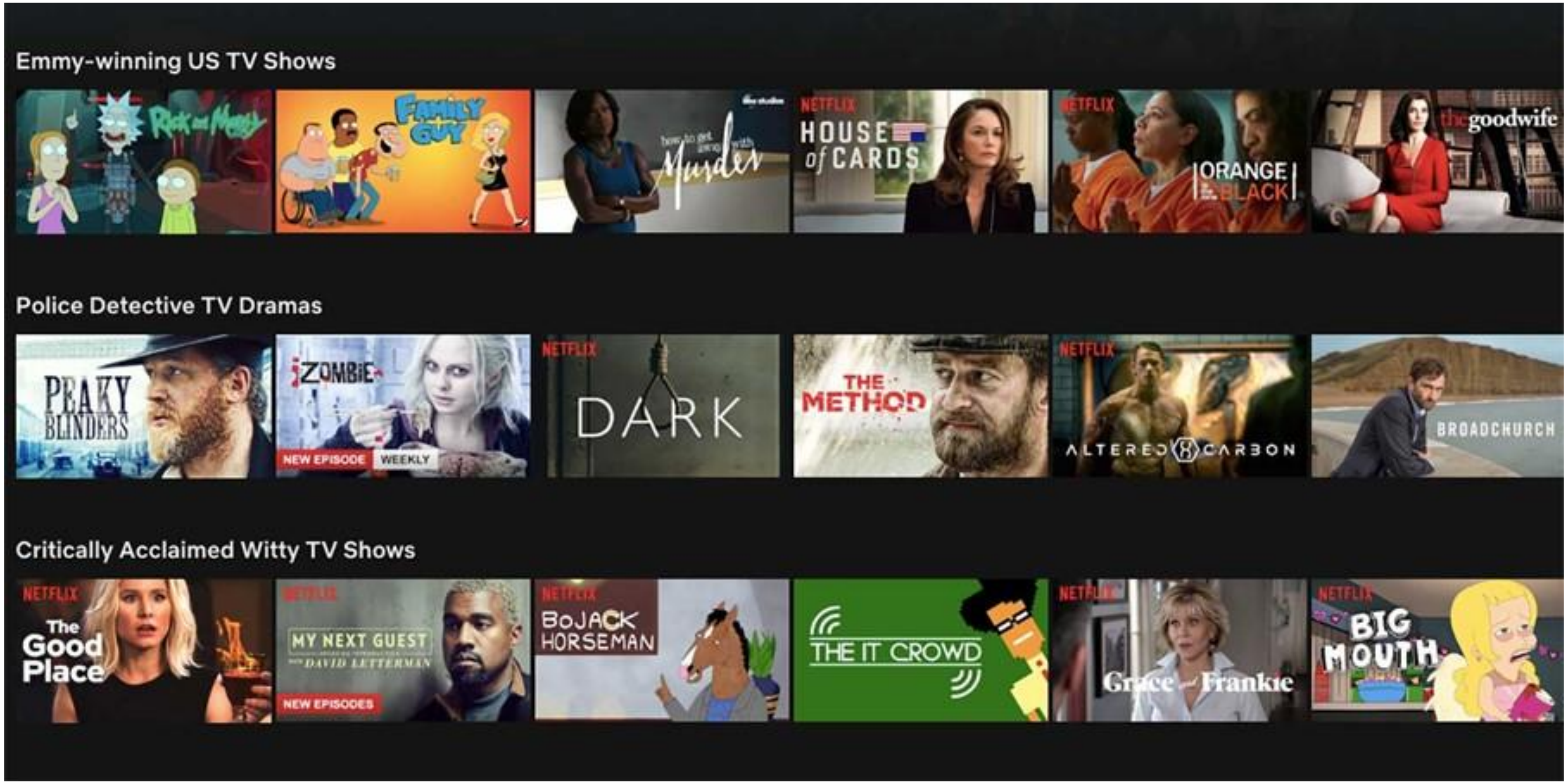
- Regression modelling: background and motivation
- Types of data
- Ordinary least square (OLS) estimation
- Simple linear regression
- Multiple linear regression

Introduction

- Key CLRM assumptions
- Violation of CLRM assumptions
- BLUE properties of OLS estimators
- Hypothesis testing with regression modelling
- Other non-linear functional forms

Background and Motivation

Amazon, Netflix Movie Recommendations



Filtering SPAM from Mail and Messages

Big Data Text Analysis

Headlines

More Headlines

COVID-19 news: See the latest coverage of the coronavirus

India should play a constructive role in Afghanistan: Pak ambassador to China

Hindustan Times · 3 hours ago

- EXCLUSIVE | Haqqani Network Scion Anas Haqqani Says Taliban Won't 'Interfere' in Kashmir, Clarifies Pakista
  News18 · 6 hours ago
- Taliban says Afghanistan 'free nation' as it hails US exit
  Al Jazeera English · 20 hours ago
- India must rethink 'wait and watch' Afghan policy
  The Indian Express · 7 hours ago · Opinion
- Does China really have the upper hand in Afghanistan?
  Moneycontrol.com · 8 hours ago · Opinion

View Full coverage

Pakistani terrorists coming to Afghanistan: Ghani informed Biden on July 23

Hindustan Times · 3 hours ago

- US President Joe Biden says US committed to safe passage for last 100-200 Americans left in Afghanistan
  Times of India · 15 hours ago

View Full coverage

Afghanistan Crisis LIVE: India Announces First Formal Meet With Taliban, Biden Defends US Exit

NDTV · 1 hour ago

- First formal meeting between India-Taliban: Afghanistan should not be used to spread terror in India
  TIMES NOW · 22 hours ago

View Full coverage

Supertech's twin towers case: Take strict action against officials involved in irregularities, says UP CM Yogi

The Indian Express · 3 hours ago

- Supreme Court orders demolition of Supertech Twin Towers | Oneindia News
  Oneindia News · 21 hours ago

View Full coverage

Summary

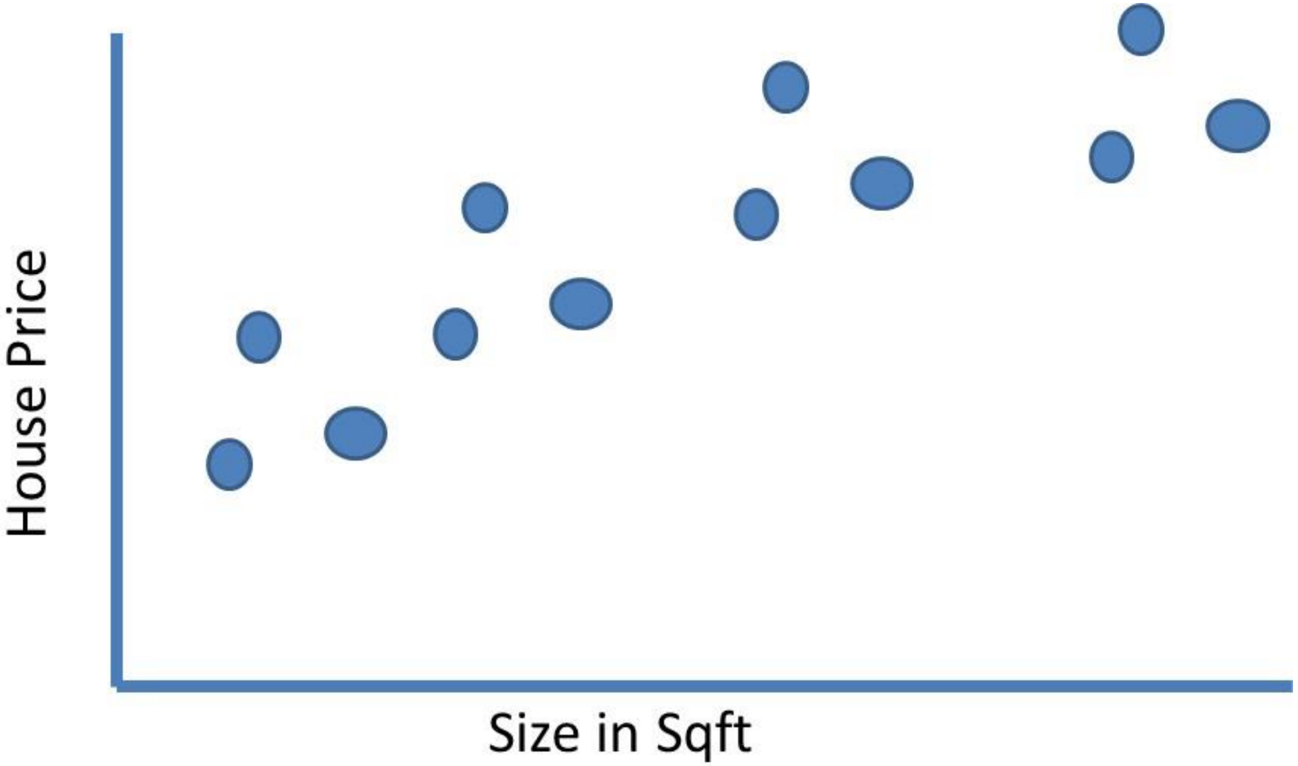
Making the Computers Learn Without Being Explicitly Programmed

- Amazon, Netflix movie recommendations
- Filtering out spams
- Medical prognosis with health records
- Algorithmic trading, credit scoring models
- Making computers think like humans
- Handwriting recognition, natural language processing, web-click data

Machine Learning Algorithms

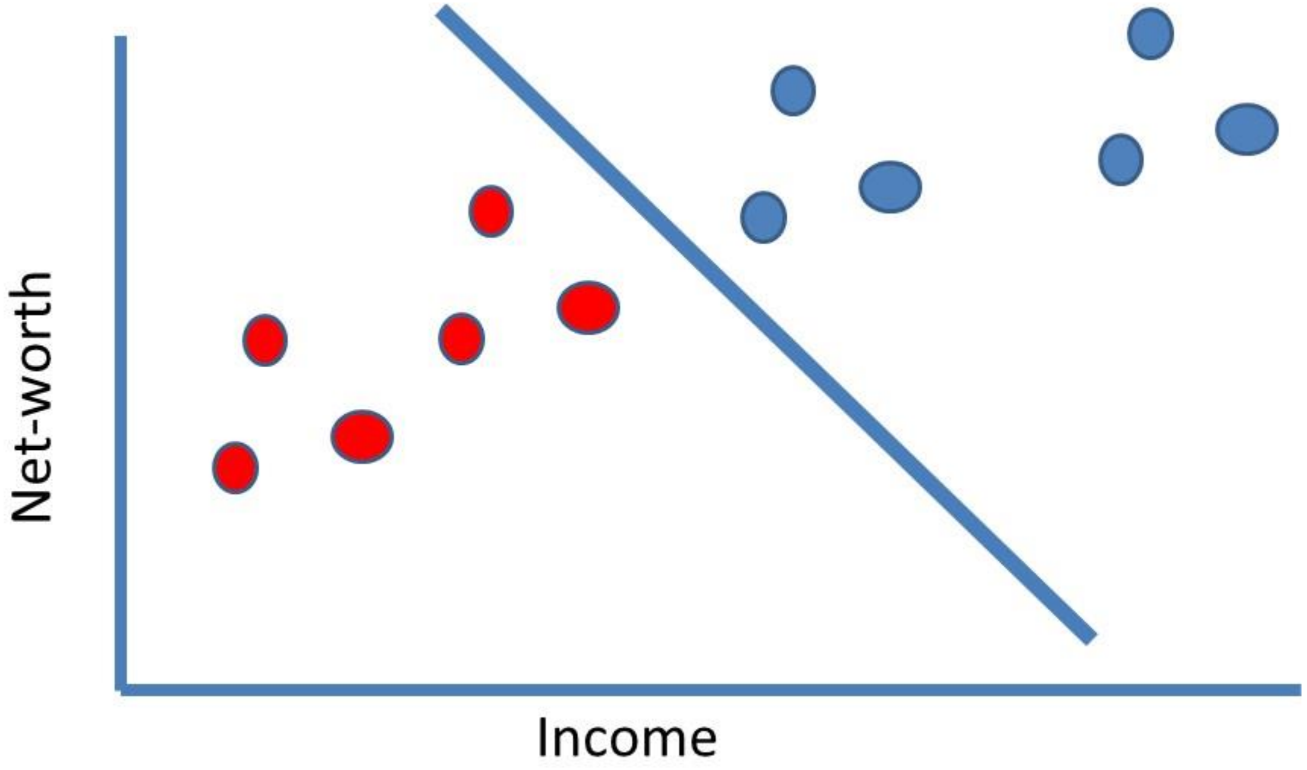
Supervised Learning

House price prediction problem (Regression problem)



Supervised Learning

Credit default scoring (classification problem)

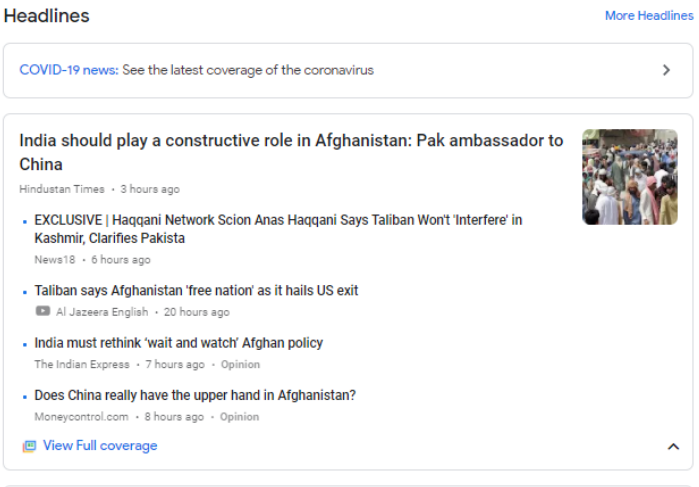


Unsupervised Learning

Clustering problem (clustering problem: market segmentation)

Unsupervised Learning

Clustering problem (clustering problem : news aggregation)



Headlines

More Headlines

COVID-19 news: See the latest coverage of the coronavirus

India should play a constructive role in Afghanistan: Pak ambassador to China

Hindustan Times • 3 hours ago

- EXCLUSIVE | Haqqani Network Scion Anas Haqqani Says Taliban Won't 'Interfere' in Kashmir, Clarifies Pakista

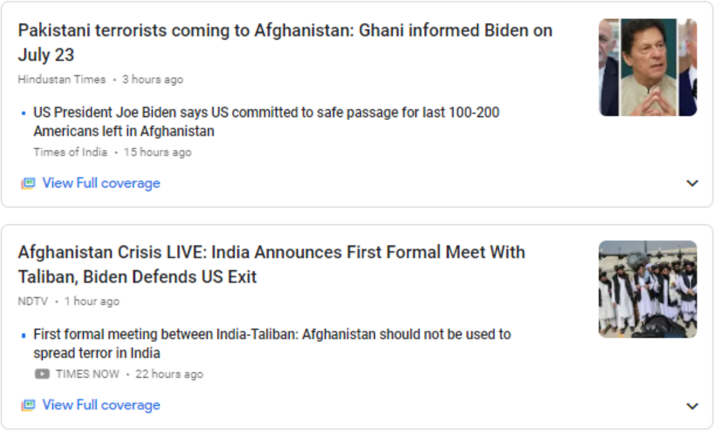
  News18 • 6 hours ago
- Taliban says Afghanistan 'free nation' as it hails US exit

  Al Jazeera English • 20 hours ago
- India must rethink 'wait and watch' Afghan policy

  The Indian Express • 7 hours ago • Opinion
- Does China really have the upper hand in Afghanistan?

  Moneycontrol.com • 8 hours ago • Opinion

View Full coverage



Pakistani terrorists coming to Afghanistan: Ghani informed Biden on July 23

Hindustan Times • 3 hours ago

- US President Joe Biden says US committed to safe passage for last 100-200 Americans left in Afghanistan

  Times of India • 15 hours ago

View Full coverage

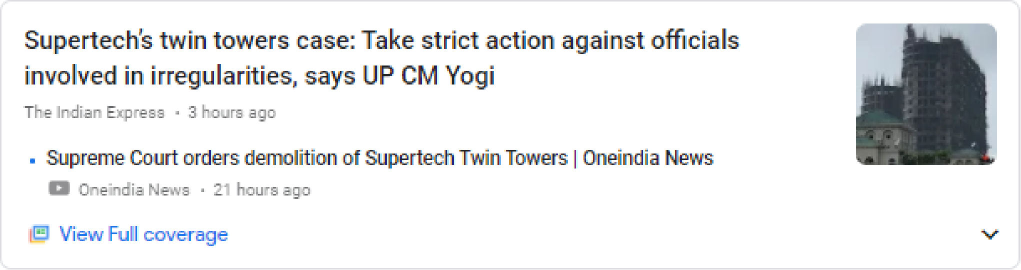
Afghanistan Crisis LIVE: India Announces First Formal Meet With Taliban, Biden Defends US Exit

NDTV • 1 hour ago

- First formal meeting between India-Taliban: Afghanistan should not be used to spread terror in India

  TIMES NOW • 22 hours ago

View Full coverage



Supertech's twin towers case: Take strict action against officials involved in irregularities, says UP CM Yogi

The Indian Express • 3 hours ago

- Supreme Court orders demolition of Supertech Twin Towers | Oneindia News

  Oneindia News • 21 hours ago

View Full coverage

Summary

- Supervised learning algorithm comprise data with features and labels
- The algorithm is trained map the relationship between the features and labels
- Then it makes predictions/create-label on the new unlabeled data based on its features
- The unsupervised learnings algorithms comprise unlabeled data that only carries features
- The data is clustered in groups based on these features

Types of Data

Types of Data

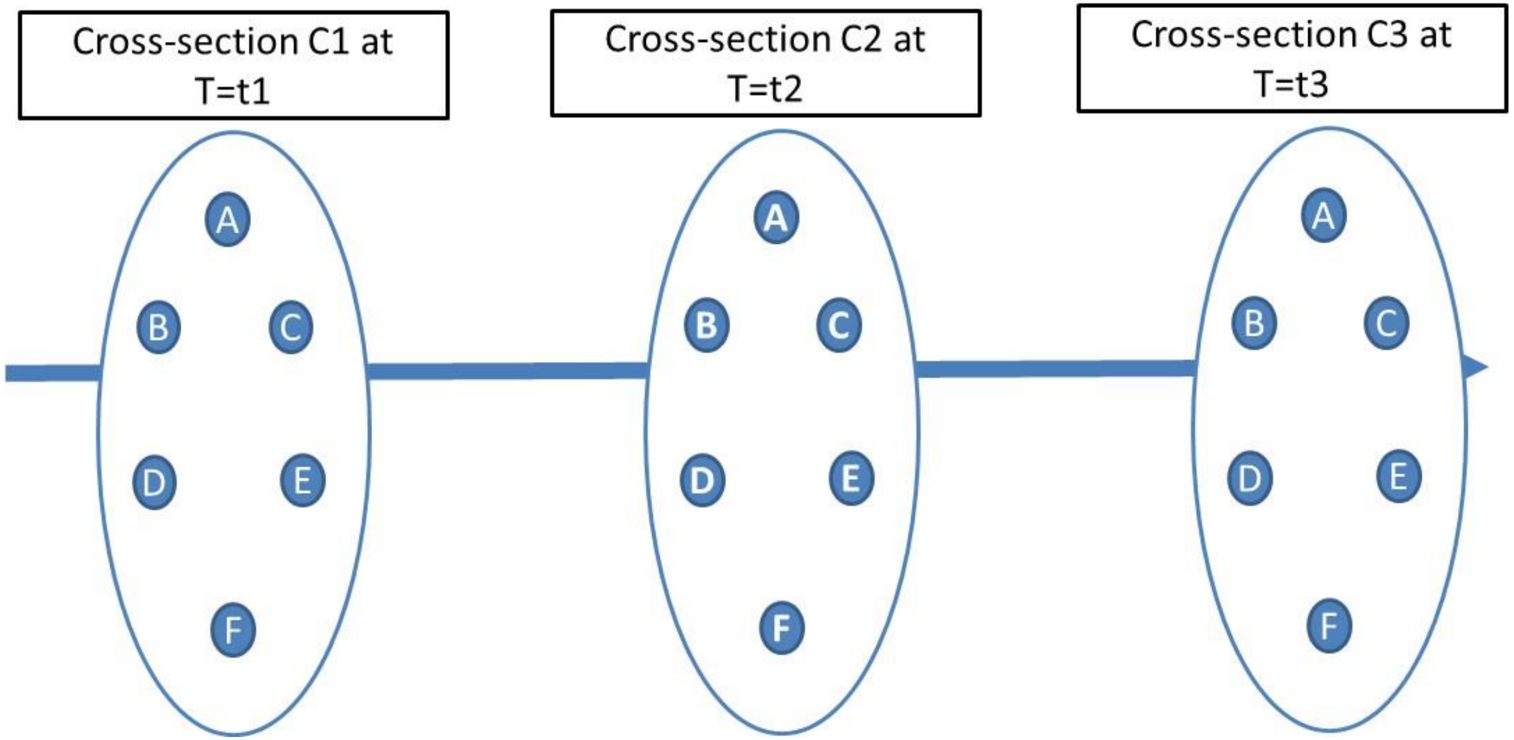
Cross-sectional data: Observations about multiple individuals (units) collected over a single period

Time-series data: Observations of a single individual (unit) collected over multiple periods

Panel or longitudinal data: Observations about multiple individuals (units) collected over various time periods

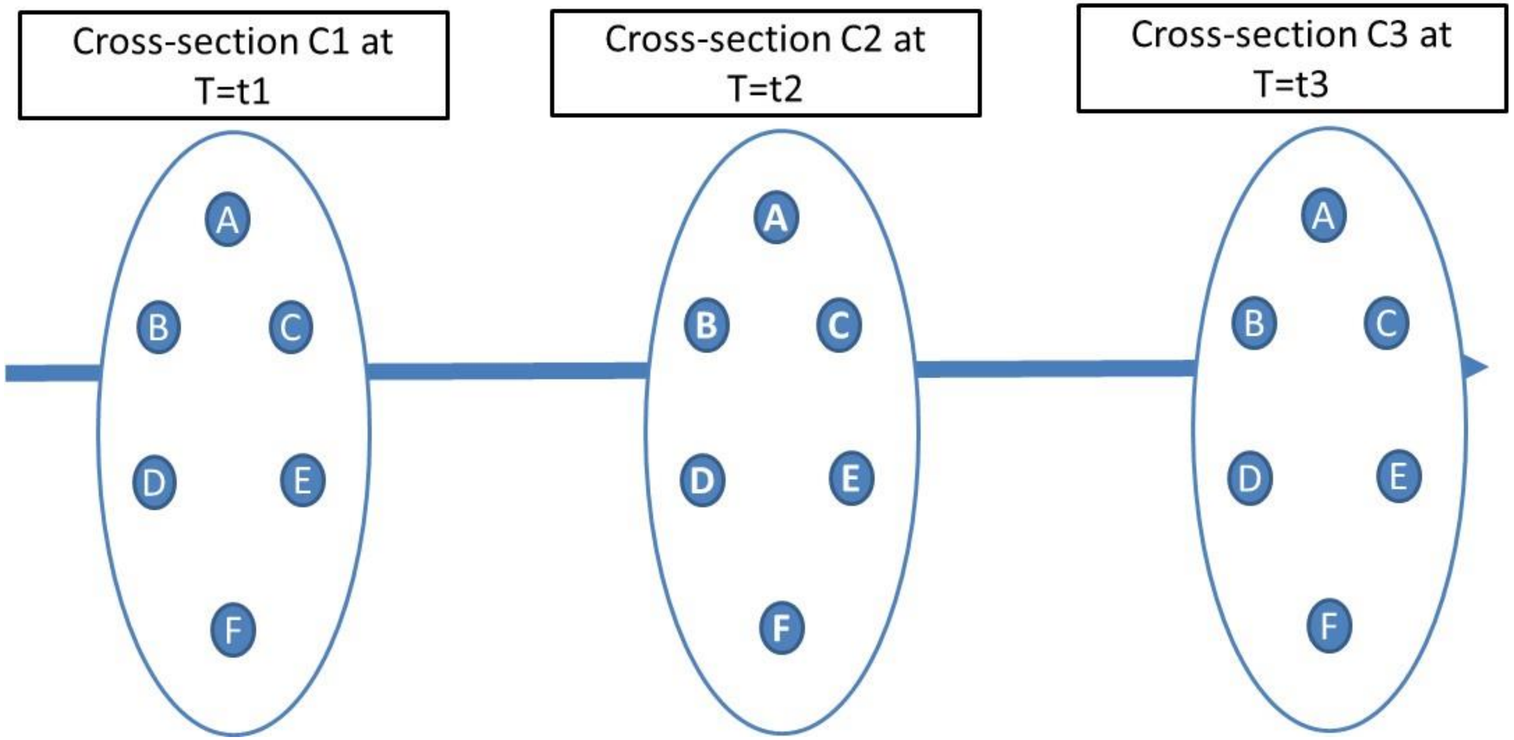
Types of Data

If information about A is collected over times t1, t2, t3 then it is time-series data



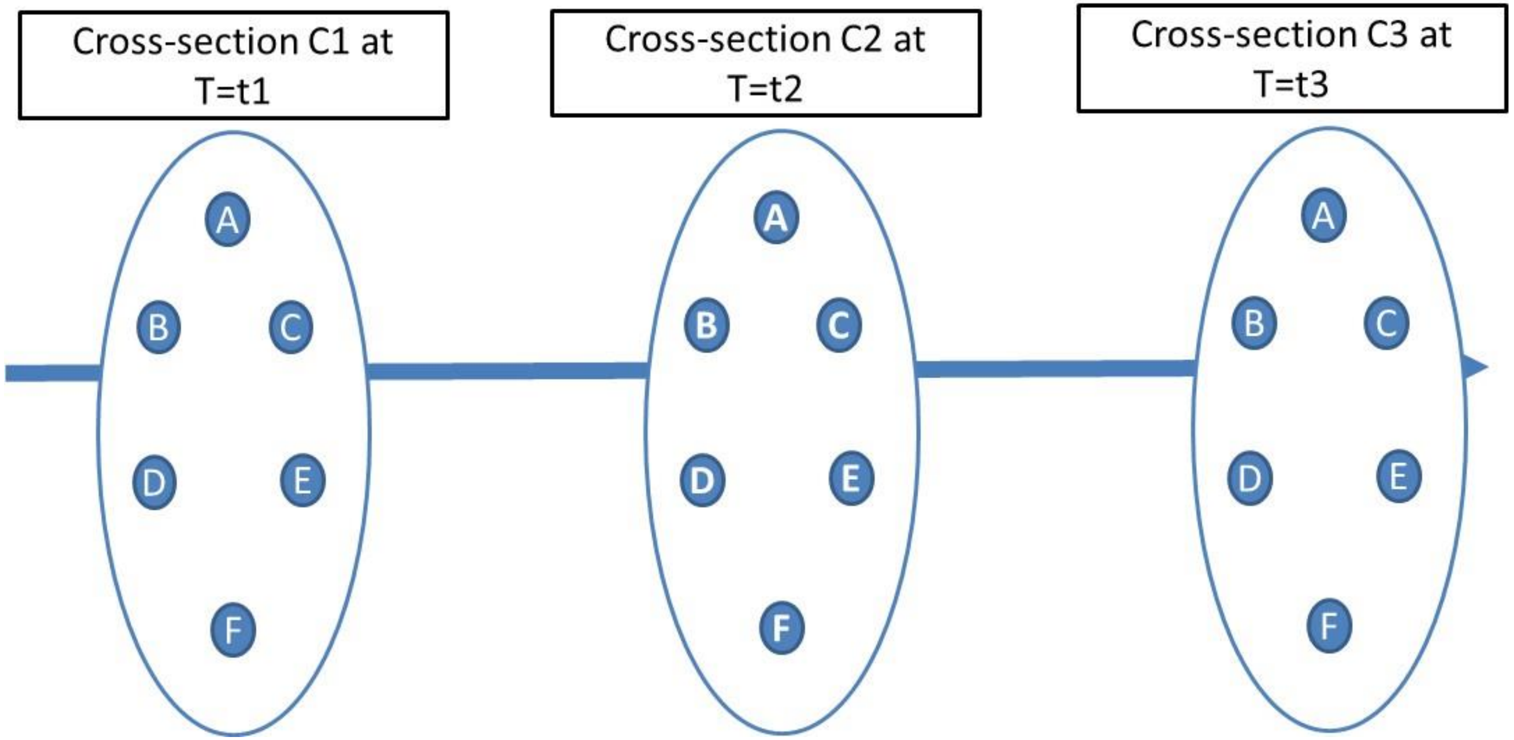
Types of Data

If information about A, B, C, D, E, and F is collected at t1, then it is cross-sectional data

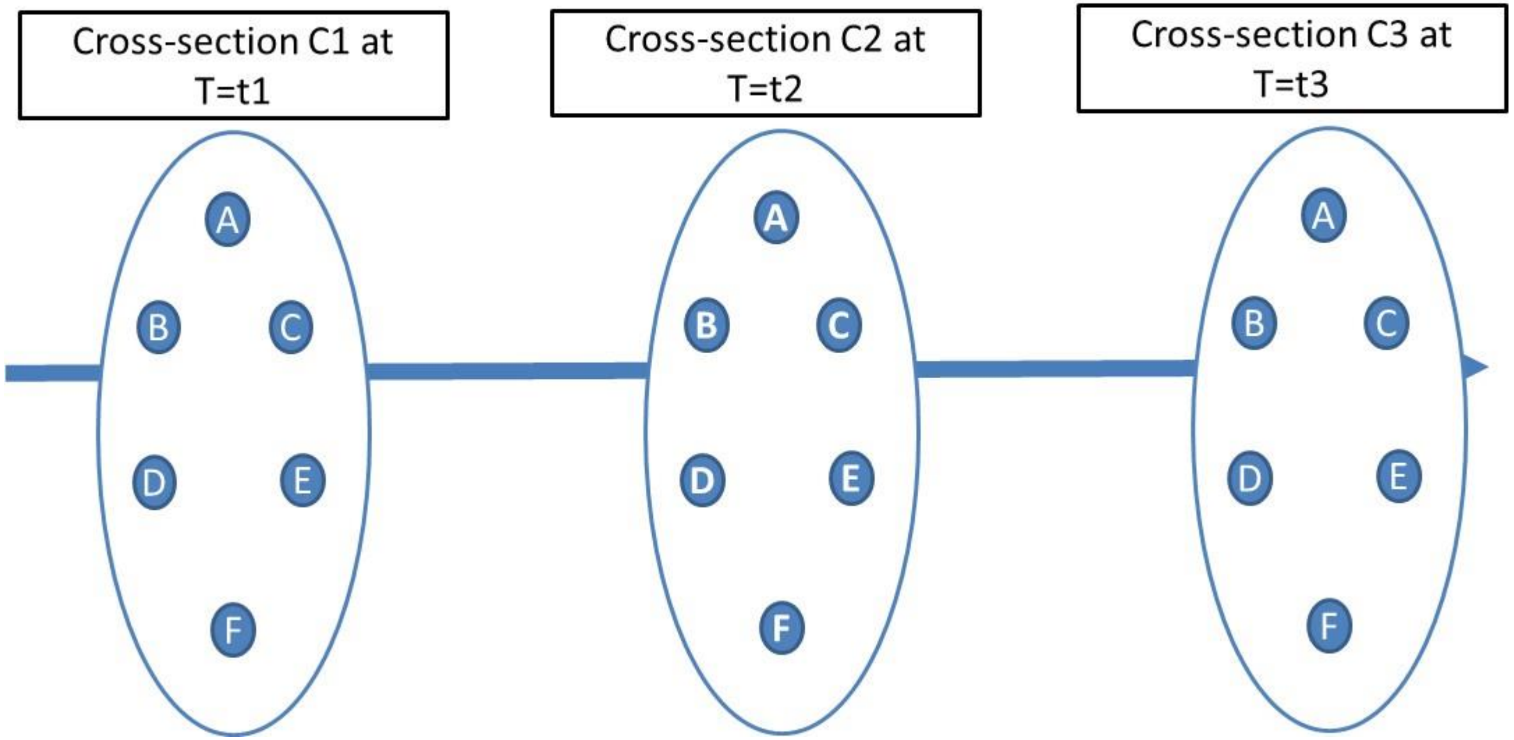


Types of Data

If information about A, B, C, D, E, and F is collected at t1, t2, t3, etc. then it is panel/longitudinal data



Summary


Cross-section C1 at T=t1
Cross-section C2 at T=t2
Cross-section C3 at T=t3
A
B
C
D
E
F
A
B
C
D
E
F
A
B
C
D
E
F

Introduction to Simple Linear Regression

Introduction to Simple Linear Regression

Consider a simple linear regression model provided :

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

- This is also a two-variable linear regression model or bivariate linear regression model
- Here ‘Y’ is the dependent /explained /response /predicted/regressand variable
- Here ‘X’ is the independent/explanatory/predictor/regressor variable

Introduction to Simple Linear Regression

Consider a simple linear regression model provided :

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

- ‘u’ is the error term, residual term or disturbance term that represents unobserved factors other than ‘X’ that affect ‘Y’; since ‘u’ is also the random or stochastic variable it has a probabilistic distribution
- Here, β_0 is the constant term and β_1 is called the slope term (Why?)
- This simple model aims to study the dependence of Y on X

Regression vs. Causation vs. Correlation

While regression deals with the dependence of one variable over another, it does not imply causation

- Regression only establishes the statistical strength of the relation, the causation is established by theory

Example of crop and rain

- A priori theoretical considerations are needed to imply causation
- In regression analysis, dependent variable is considered random or stochastic (i.e., with probability distribution), while explanatory variable is assumed to have fixed values

Regression vs. Causation vs. Correlation

A closely associated concept is correlation, which establishes the degree of linear relationship between the two variables

- In correlation analysis, both the variables are treated in a similar manner and considered to be random

Summary

Expectations Operator

Expectations Operator ‘E’

Any random probabilistic variable is often represented through expectations operator

- Any random variable attains multiple values. For example, a coin-toss can obtain two values with 50% odds for any outcome
- Similarly, in regression any random value is assumed to be probabilistic in nature and its expected value is represented by E(Y)

Expectations Operator ‘E’

For example, if there are ‘n’ possibilities of an event, y1, y2, y3,..,yn each with possibilities p1, p2, p3, p4…pn, then expectations operator is defined as

- E(y)= $p_1 * y_1 + p_2 * y_2 + p_3 * y_3 + \ldots\ldots + p_4 * y_4$
- This is also called probability weighted mean
- If all the probabilities are assumed to be equal then $p_1 = p_2 = \ldots.. = p_n = \frac{1}{n}$
- Then E(y)= $\frac{1}{n}(y_1 + y_2 + y_3 + \ldots. + y_n)$, i.e., simple average of Y’s

Summary

- We discussed the role of expectations operator (E) in the context of stochastic random variable with a probability distribution
- In simple terms, expectations are probability weighted averages of stochastic random variable
- In case there is no a priori probabilities assigned to these variables, then the expectation is simple average of the stochastic random variable

A Simple Example

A Simple Example

Consider a simple example of family income and consumption expenditure shown below

$Y\downarrow$ \ $X\rightarrow$	80	100	120	140	160	180	200	220	240	260
Weekly family consumption expenditure Y,	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1043	966	1211
Conditional means of Y, $E(Y\mid X)$	65	77	89	101	113	125	137	149	161	173

Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

A Simple Example

Here population of 60 families is divided into 10 income (X) groups from 80-260 (independent or fixed variable)

- The corresponding consumption expenditure values (Y) are also shown
- For each given level of income (X), the conditional means E(Y/X) that is mean of Y for a given level of X is also provided

A Simple Example

For example, at X= 80, the mean of Y is 65, i.e., E(Y/X=80)=65; these are called conditional expectations or conditional means of Y given the value of X

- As they depend on the conditioning variable X
- The average of all Y's, that is unconditional mean or unconditional expected value E(Y)= 121.2

A Simple Example

This unconditional mean does not account for the level of income (X) and is the prediction of Y (expected value) when there is no knowledge of X

- However, if one has the knowledge of X, then one can improve the prediction by computing conditional mean of Y, i.e., E(Y/X), which is a more accurate prediction of Y

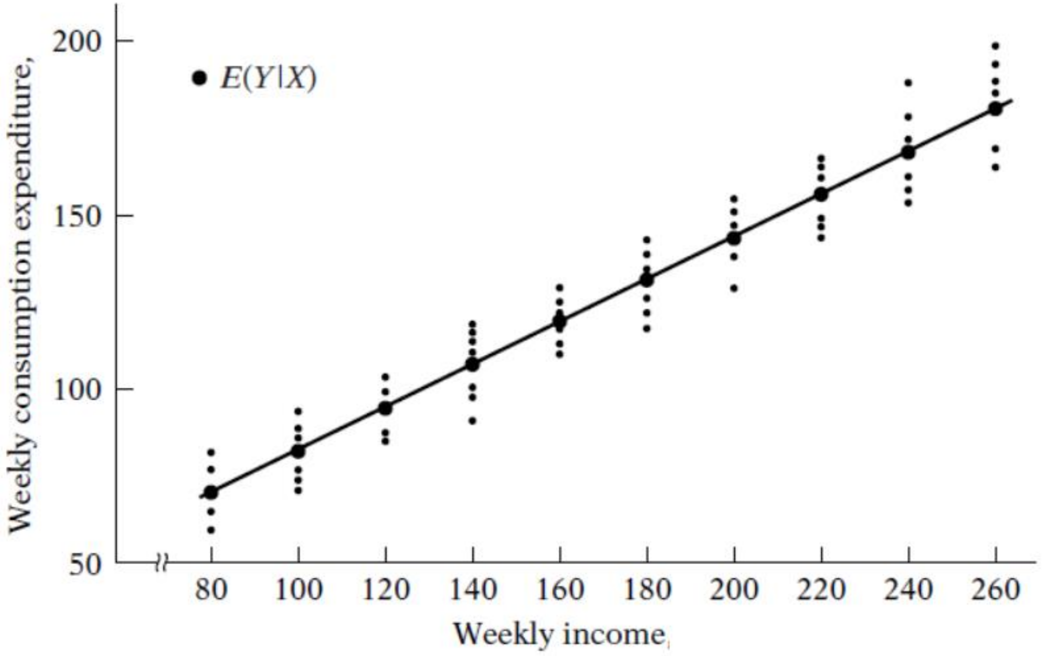
A Simple Example

Que: What is the best (mean) prediction of weekly expenditure of families with a weekly income of X= 140: Y= 101

- Thus the knowledge of the income level may enable us to better predict the mean value of consumption expenditure than if we do not have that knowledge
- This is the essence of regression modelling

A Simple Example

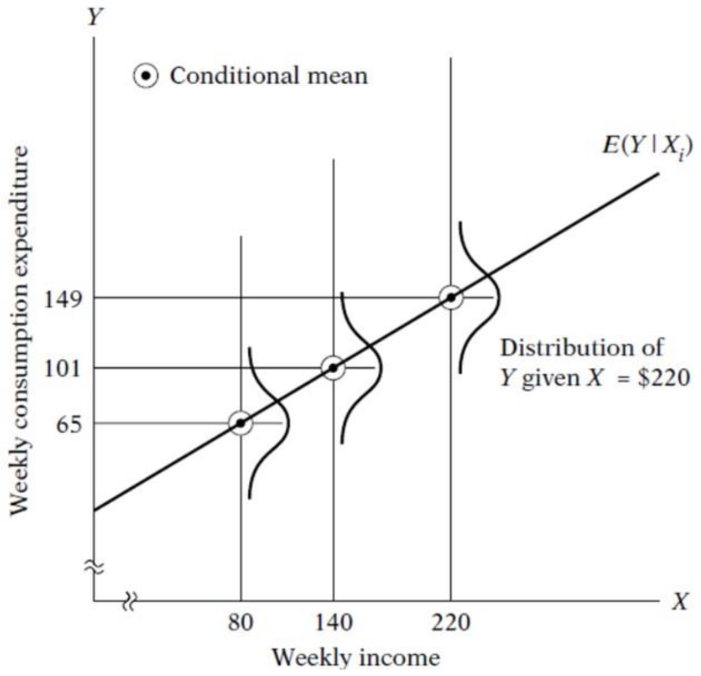
Que: What is the best (mean) prediction of weekly expenditure of families with a weekly income of X= 140: Y= 101



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

A Simple Example

Que: What is the best (mean) prediction of weekly expenditure of families with a weekly income of X= 140: Y= 101



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

Population and Sample Regression Function

Concept of Population Regression Function

If we join these conditional mean values, we obtain what is known as the population regression line (PRL)

- More simply, it is the regression of Y on X
- Geometrically, then, a population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable

Concept of Population Regression Function

Population regression function (PRF)

- $E(Y/X_i)=f(X_i)$
- In this case, $f(X_i)$ is a linear function of X
- The expression is also called population regression function

Concept of Population Regression Function

More generally, for a two variable case: $E(Y/X_i)$=$\beta_0 + \beta_1 X_i$

- Here it is important to note that linearity means linearity in parameters
- $E(Y/X_i)$=$\beta_0 + \beta_1^2 X_i$; this model is non-linear in parameters and will not be handled in linear regression modelling
- $\boldsymbol{E(Y/X_i)}$=$\boldsymbol{\beta_0 + \beta_1\ X_i^2}$, in contrast this model is non-linear in variables and can be handled under linear regression models

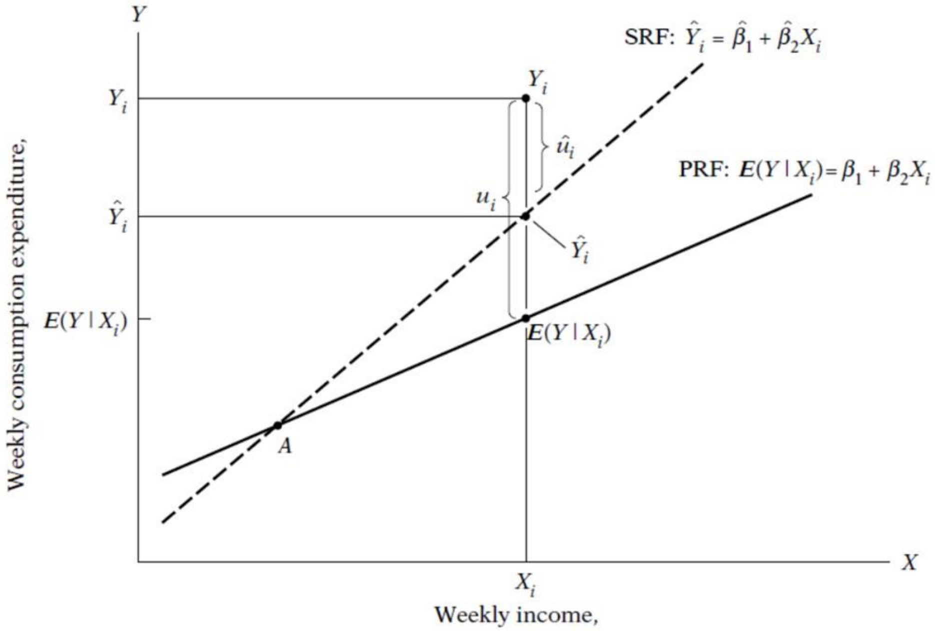
Sample Regression Function (SRF)

Sample regression function is shown by adding “^” hat symbol, indicating the estimated values: $\hat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$

- $\widehat{\beta_0}$ is the estimator of β_0; $\widehat{\beta_1}$ is the estimator of β_1, and $\hat{Y_i}$ is the estimator of y_i
- SRF is only an estimate of PRF
- Thus, SRF can over or underestimate PRF values

Sample Regression Function (SRF)

Sample regression function (SRF): $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

Ordinary Least Square (OLS) Estimation

Method of Ordinary Least Square Estimation (OLS)

Recall the SRF function: $Y_i \;=\widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$; where $\widehat{Y_i} \;=\widehat{\beta_0} + \widehat{\beta_1} X_i$

- Here, $\widehat{\mu_i} = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$
- The line fit should aim to minimize the square this error $\widehat{\mu_i}$
- Concept of OLS suggests that the best cost function to minimize is as follows

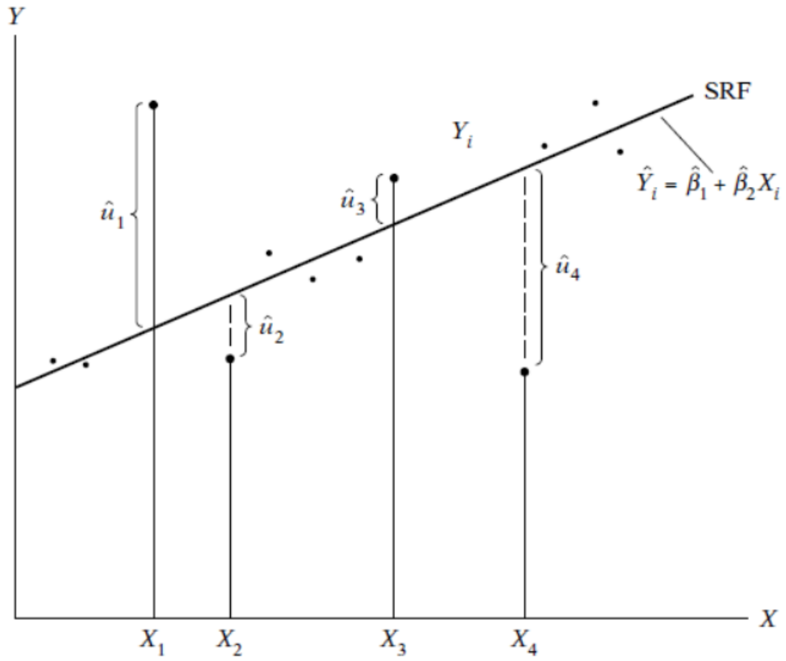
Method of Ordinary Least Square Estimation (OLS)

Recall the SRF function: $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$;   where $\hat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$

- $\sum \widehat{\mu_i^2} = \sum (Y_i - \hat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$
- That is we minimize squared residuals (why not just residuals or absolute residuals)

Method of Ordinary Least Square Estimation (OLS)

- $\sum \widehat{\mu_i^2} = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$: Minimize these squared residuals



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum \left(Y_i - \widehat{Y_i}\right)^2 = \sum \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

- Obvious to note here that $\sum \widehat{\mu_i^2} = f(\widehat{\beta_0}, \widehat{\beta_1})$
- Setting differential of $\sum \widehat{\mu_i^2} = 0$ that satisfies and double differential to positive for minima condition, one obtains the estimates that is, $\widehat{\beta_0}$ and $\widehat{\beta_1}$

Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

- Thus, these estimators are called least square estimators
- The regression model such estimated is also called the Gaussian, standard, or classical linear regression model (CLRM),

Method of Ordinary Least Square Estimation (OLS)

$\sum \widehat{\mu_i^2} = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - \widehat{\beta_0} - \widehat{\beta_1}X_i\right)^2$

- $\frac{\partial\left(\sum \widehat{\mu_i^2}\right)}{\partial \widehat{\beta_0}} = -2\sum\left(\mathrm{Y_i} - \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_i}\right)$=-2$\sum \widehat{\mu}_{\mathrm{i}}$ (partial differential w.r.t. to $\widehat{\beta_0}$)
- $\frac{\partial\left(\sum \widehat{\mu_i^2}\right)}{\partial \widehat{\beta_1}} = -2\sum\left(\mathrm{Y_i} - \widehat{\beta_0} + \widehat{\beta_1}\mathrm{X_i}\right)\mathrm{X_i}$=-2$\sum \widehat{\mu}_{\mathrm{i}}\,\mathrm{X_i}$ (partial differential w.r.t. to $\widehat{\beta_1}$)

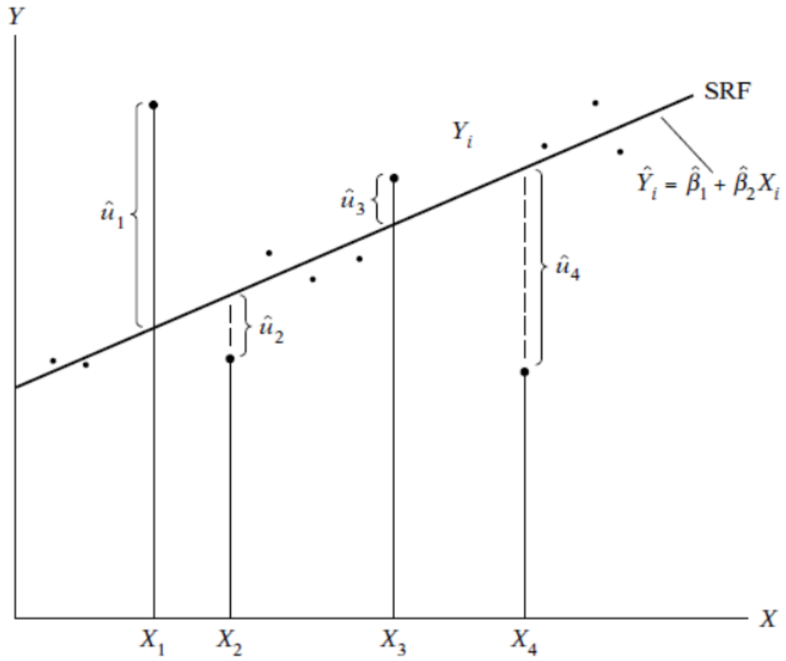
Method of Ordinary Least Square Estimation (OLS)

$$\sum \widehat{\mu_i^2} = \sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$$

- $\widehat{\boldsymbol{\beta_1}} = \frac{\sum(\boldsymbol{X_i} - \overline{\boldsymbol{X}})(\boldsymbol{Y_i} - \overline{\boldsymbol{Y}})}{(\boldsymbol{X_i} - \overline{\boldsymbol{X}})^2}$ and $\widehat{\boldsymbol{\beta_0}} = \overline{\boldsymbol{Y}} - \widehat{\boldsymbol{\beta_1}} \overline{\boldsymbol{X}}$
- However, to achieve this closed form solution CLRM-OLS makes certain assumptions

Summary

- $\sum \widehat{\mu_i^2} = \sum (Y_i - \widehat{Y_i})^2 = \sum (Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i)^2$: Minimize these squared residuals



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Introduction to Multiple Linear Regression

Introduction to Multiple Linear Regression

We can generalize the two variable problem into multiple linear regression as $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u}$

- $\boldsymbol{X_i' s}$ represent the explanatory variables
- Here the coefficients $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_n}$ are called the partial regression coefficients

Introduction to Multiple Linear Regression

Multiple linear regression $\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} \ldots. + \boldsymbol{\beta_n X_n} + \boldsymbol{u}$

- Other aspects of the regression remain the same, including the properties of the error term, that is, $\boldsymbol{u}$
- Zero conditional mean of error: $E(u_i|X_{1i}, X_{2i}, \ldots, X_{ni}) = 0$ for each 'i'
- No serial correlation: $cov(u_i, u_j) = 0$; Homoscedasticity: $var(u_i) = \sigma^2$

Introduction to Multiple Linear Regression

Multiple linear regression $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n + u}$

- Zero correlation (or covariance) between u_i and X: $cov(u_i, X_1)$=$cov(u_i, X_2)$=$cov(u_i, X_n)$=0
- The model is correctly specified

Introduction to Multiple Linear Regression

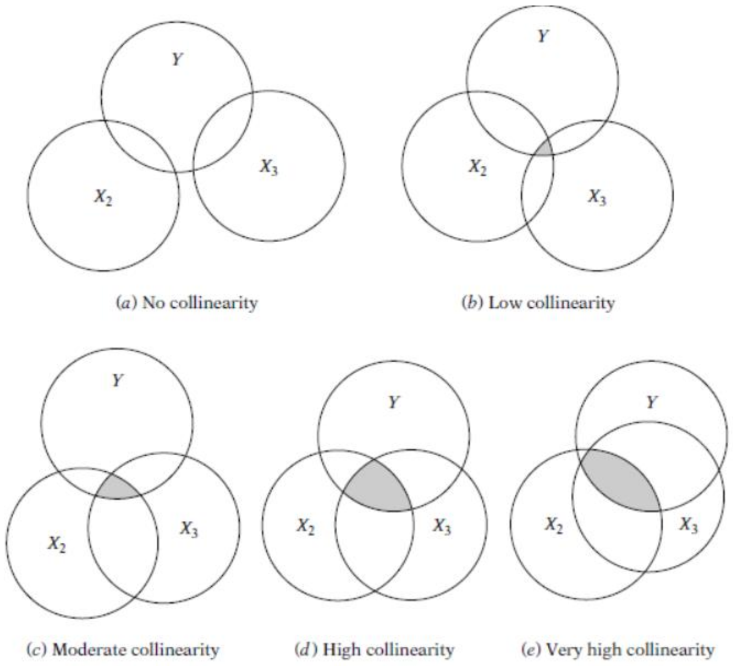
Multiple linear regression $\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots. + \beta_n X_n + u}$

- Lastly, one more condition is added; that is no exact linear relationship between X_i and X_js $(X_1, X_2 \ldots X_n)$: $\alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots \alpha_n X_n \neq 0$
- If such a relationship exists, the model will be affected by the problem of perfect multicollinearity, and will not run (i.e., indeterminate)

Introduction to Multiple Linear regression

However, there may be instances of less than perfect collinearity across variables and can affect the estimation

- If the multicollinearity is not perfect but high, then the estimators have large variances (standard errors of estimate)

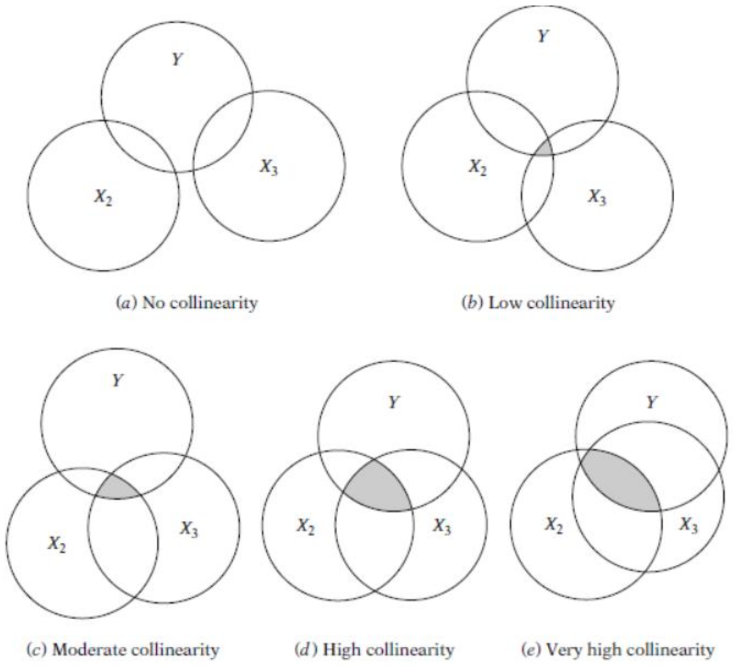


Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Introduction to Multiple Linear regression

However, there may be instances of less than perfect collinearity across variables and can affect the estimation

- This makes the 't-values' low and high chances of failure to reject the null-hypothesis (wider confidence intervals), even though the R^2 may be high



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Summary

- We discussed multiple linear regression model
- All the properties and discussions on simple linear regression model apply to multiple linear regression model
- Some important properties of simple and multiple linear regression included: (a) zero conditional mean of the error term; (b) error term should not be serially correlated; (c) variance of the error term should be constant: Homoscedasticity; (d) no correlation between the error term and the independent variable; (e) model should be correctly specified; (f) multicollinearity should be low

Interpreting the Multiple Linear Regression

Interpreting the Multiple Linear Regression

Similar to the two variable regression, the following expression

$$E(Y|X_1 \dots . X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots . + \beta_n X_n$$

- Represents the conditional mean or expected value of Y given the fixed values of all the $X_i's$
- The partial coefficient β_1 is the effect of X_1 on Y, net of any effect from other explanatory variables ($X_i's$), or in other words, keeping all the $X_i's$ constant

Interpreting the Multiple Linear Regression

Similar to the two variable regression, the following expression

$E(Y|X_1 \ldots .X_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots . + \beta_n X_n$

- The definition of $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$, which is same as earlier
- One also calculates adjusted-$R^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{(MSS\ of\ RSS)}{(MSS\ of\ TSS)}$; remember the dfs?

Interpreting the Multiple Linear Regression

$$\text{Adjusted-}R^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{(MSS\ of\ RSS)}{(MSS\ of\ TSS)}$$

- Or Adjusted-R^2=1-(1-R^2)*(n-1)/(n-k)
- Adjusted-R^2 penalizes addition of more variables. So if the R^2 is inflated just by adding the number of variables, rather than their quality, then Adjusted-R^2 can identify the same

Interpreting the Multiple Linear Regression

In the OLS estimation each parameter $(\widehat{\beta_0}, \widehat{\beta_1})$ is estimated with some error

- The square-root of the variance of the estimated parameter indicates that error in estimation or the precession of the estimate

Summary

- The interpretation of multiple linear regression model broadly remains similar to the bivariate regression model
- The coefficients are partial coefficients that measure the impact of independent variable on dependent variable, keeping other variables constant
- The explanatory power of the model is measured using the R^2 measure
- An improvement of over the R^2 measure is adjusted- R^2 measure which penalizes the addition of variables in the model
- Lower standard errors of the coefficients increases the power and efficiency of the mode
- OLS estimators are the best estimators in the class of linear estimators

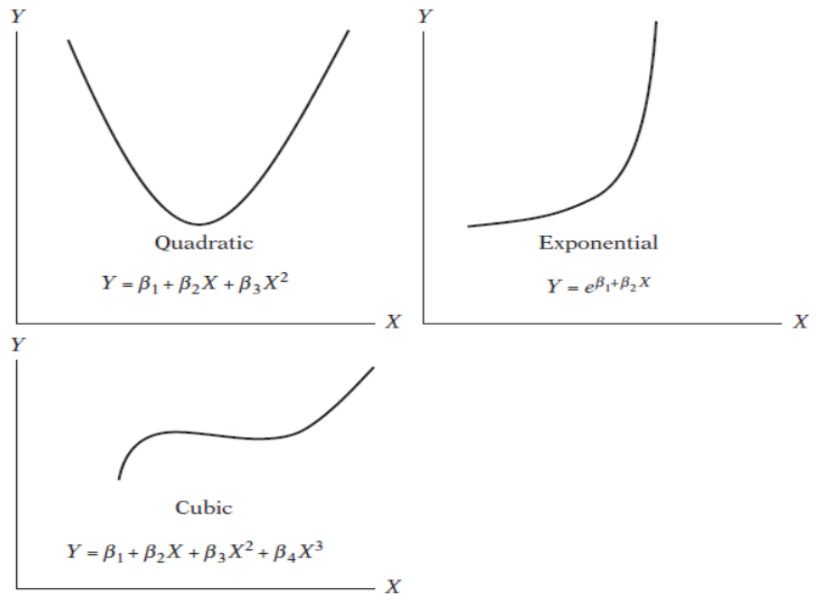
Key CLRM Assumptions

Key CLRM Assumptions

The Gaussian, standard, or classical linear regression model (CLRM), makes 10 key assumptions

- **Assumption 1:** The regression model is linear in the parameters $(\widehat{\beta_0}, \widehat{\beta_1}, \ldots.)$

Linear in Parameters



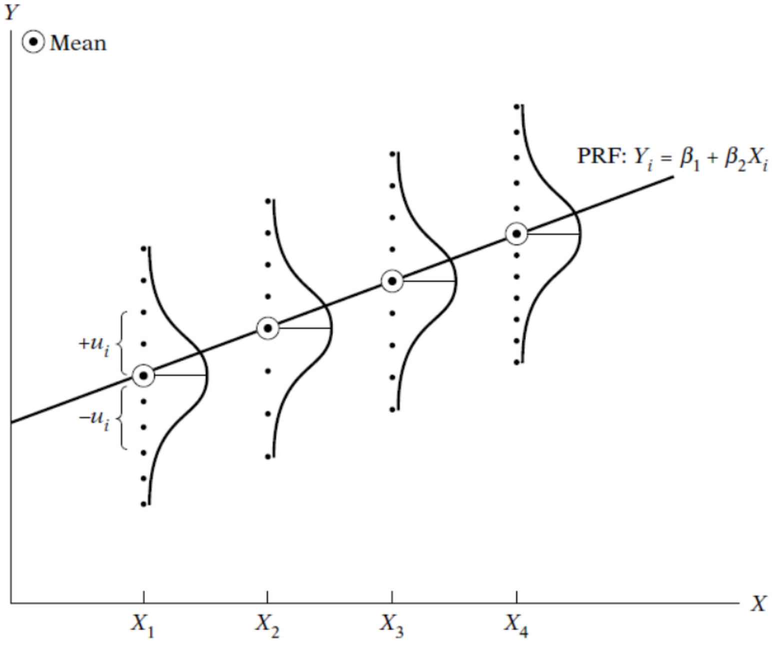
Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 2)

Key CLRM Assumptions

Assumption 2: Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be non-stochastic

Key CLRM Assumptions

Assumption 3: Zero conditional mean of disturbance ($\boldsymbol{u_i}$): given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. $E(u_i/X_i)$=0

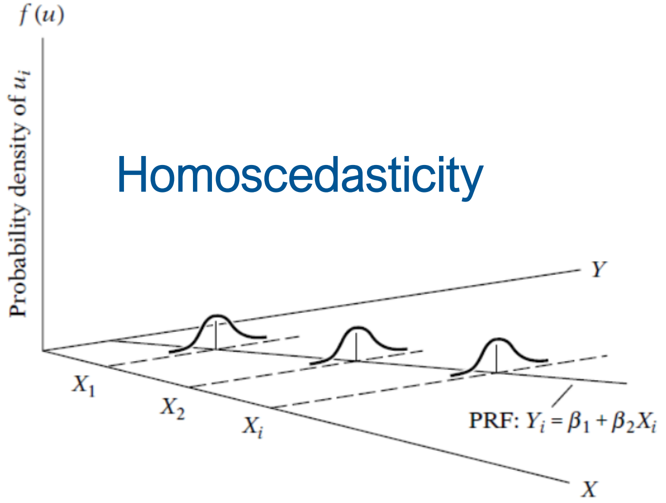


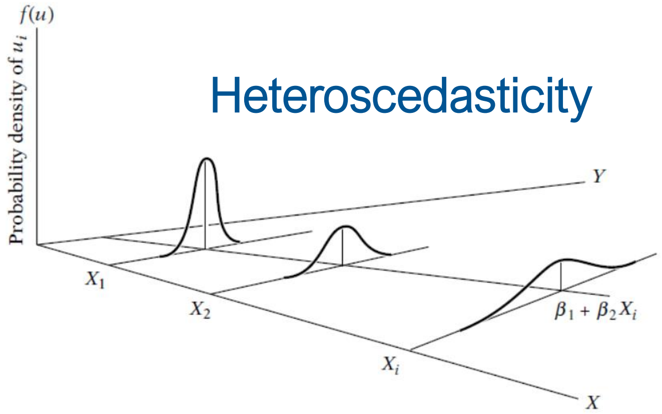
Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Key CLRM Assumptions

Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of u_i are identical. $var(u_i/x_i)$=$E[u_i - E(u_i|X_i)]^2 = constant = \sigma^2$

- Heteroscedastic variance= $var(u_i/x_i)$=σ_i^2





Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Key CLRM Assumptions

Assumption 5: No autocorrelation between the disturbances. Given any two X values, Xi and Xj (i ≠ j), the correlation between any two u_i and u_j (i ≠ j) is zero.

Symbolically, $Cov(u_i, u_j | X_i, X_j) = E\left[[u_i - E(u_i)|X_i][u_j - E(u_j)|X_j]\right]^2$

$= E[(u_i|X_i)(u_j|X_j)] = 0$

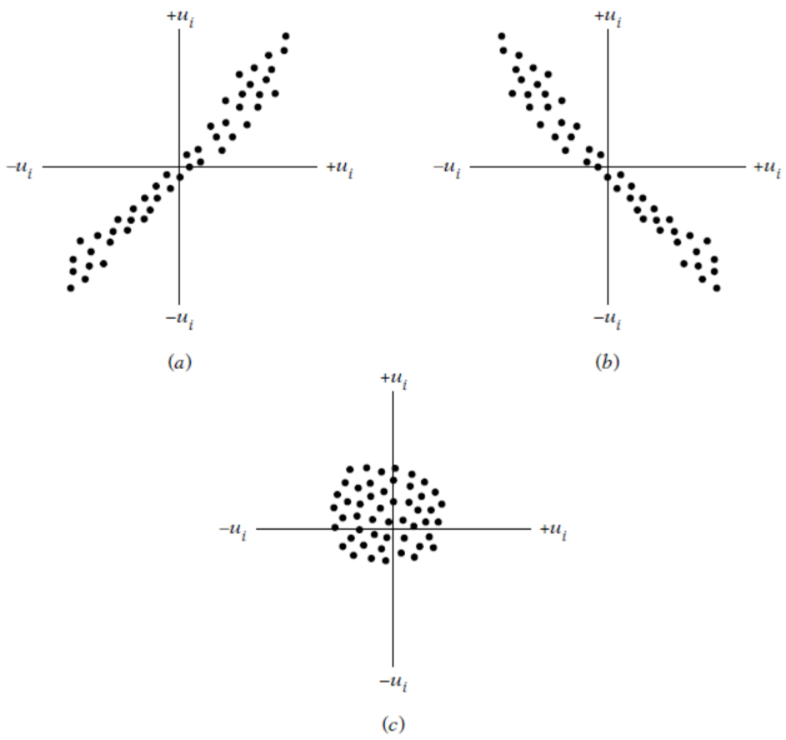
Key CLRM Assumptions

Assumption 5: No autocorrelation between the disturbances

(a) Positive autocorrelation

(b) negative autocorrelation

(c) No autocorrelation



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 3)

Key CLRM Assumptions

Assumption 6: Zero covariance between u_i and X_i, or $E(u_i X_i) = 0$.

- $Cov(u_i, X_i) = E[(u_i - E(u_i))(X_i - E(X_i)]$
- By definition: $E(u_i)=0$; $E(u_i X_i) - E(u_i)E(X_i)$
- $Cov(u_i, X_i) = E(u_i X_i) = 0$
- That is, u_i and X_i are not correlated

Key CLRM Assumptions

- Assumption 7: The number of observations must be greater than the number of parameters to be estimated
- Assumption 8: The X values (independent variable) must have some finite variance
- Assumption 9: The regression model is correctly specified
- Assumption 10: There is no perfect multicollinearity, i.e., no perfect linear relationships among the explanatory variables

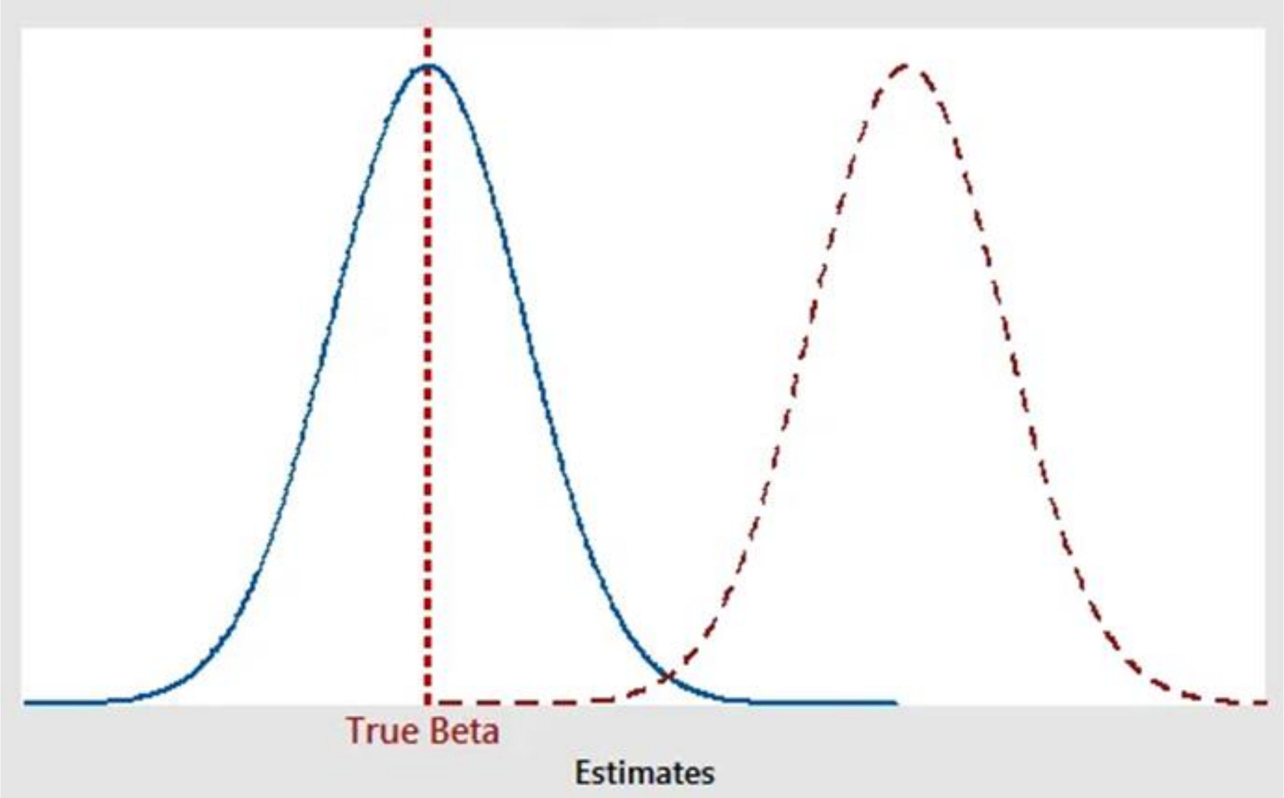
Summary

- In this video we reviewed and summarized the ten (10) key CLRM assumptions

BLUE Properties of OLS Estimators

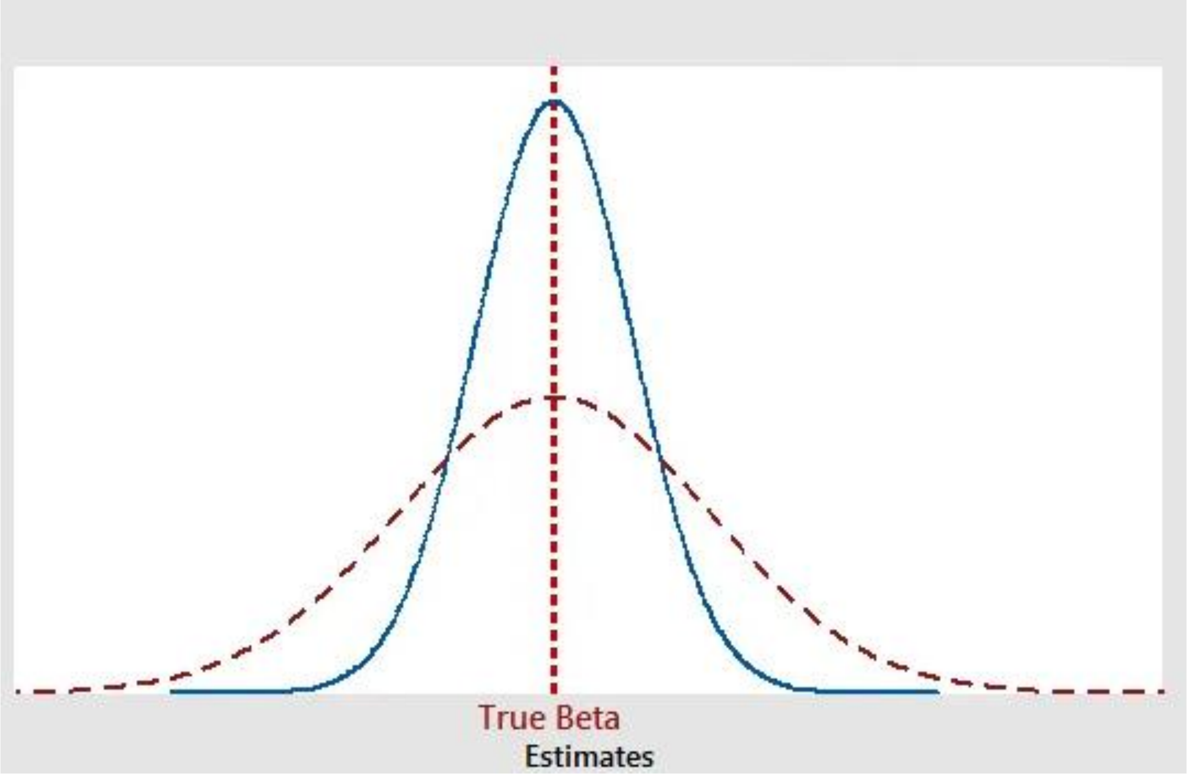
BLUE Properties of OLS Estimators

Biased and Unbiased estimators



BLUE Properties of OLS Estimators

Efficient and inefficient estimators

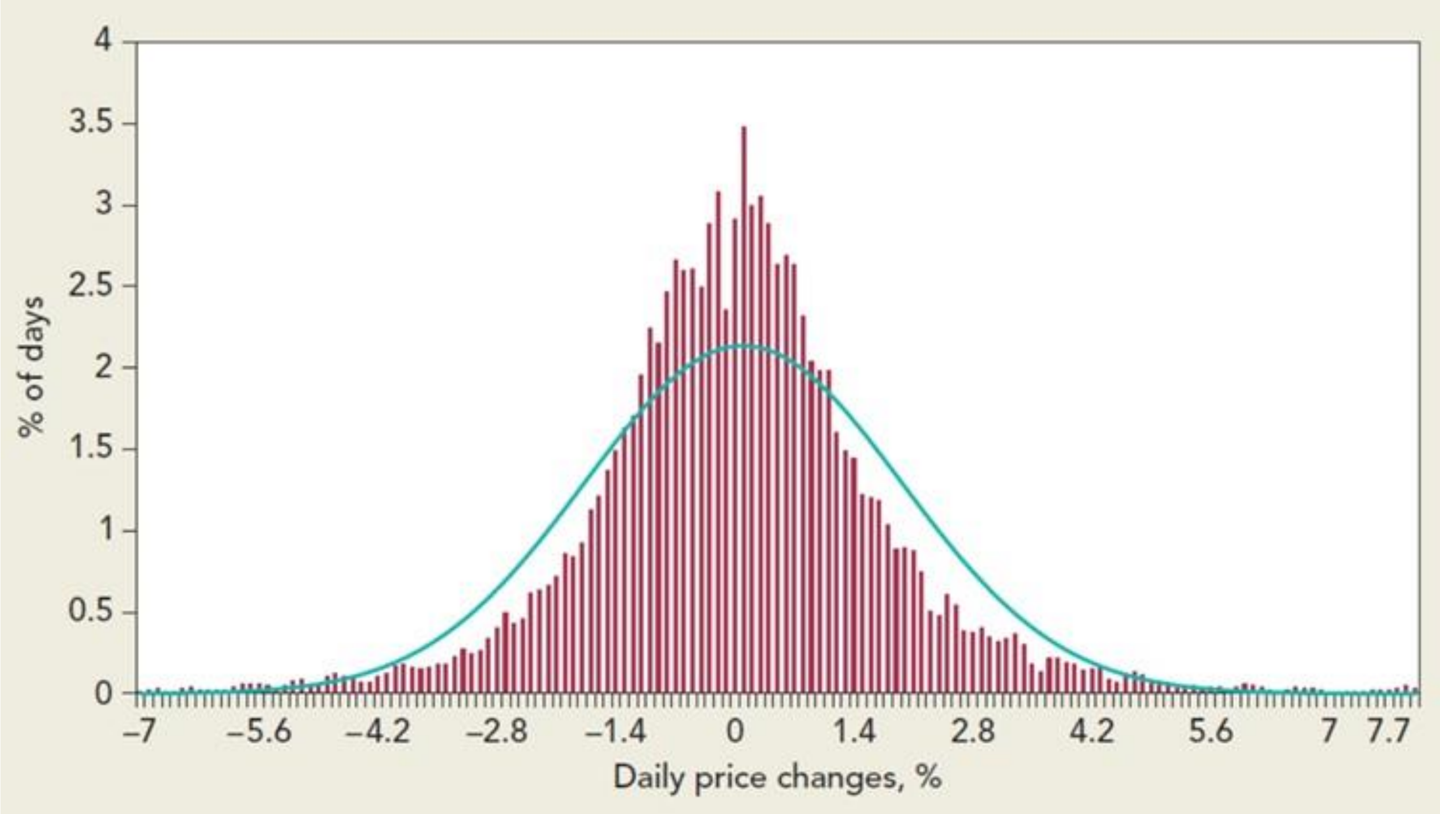


Summary

- OLS estimators are BEST in linear class of estimators
- They are best, linear, unbiased, and efficient estimators
- Thus, they are also consistent estimators: for large samples, OLS estimators converge to true population parameters

Classical Normal Linear Regression Model (CNLRM) and Hypothesis testing I

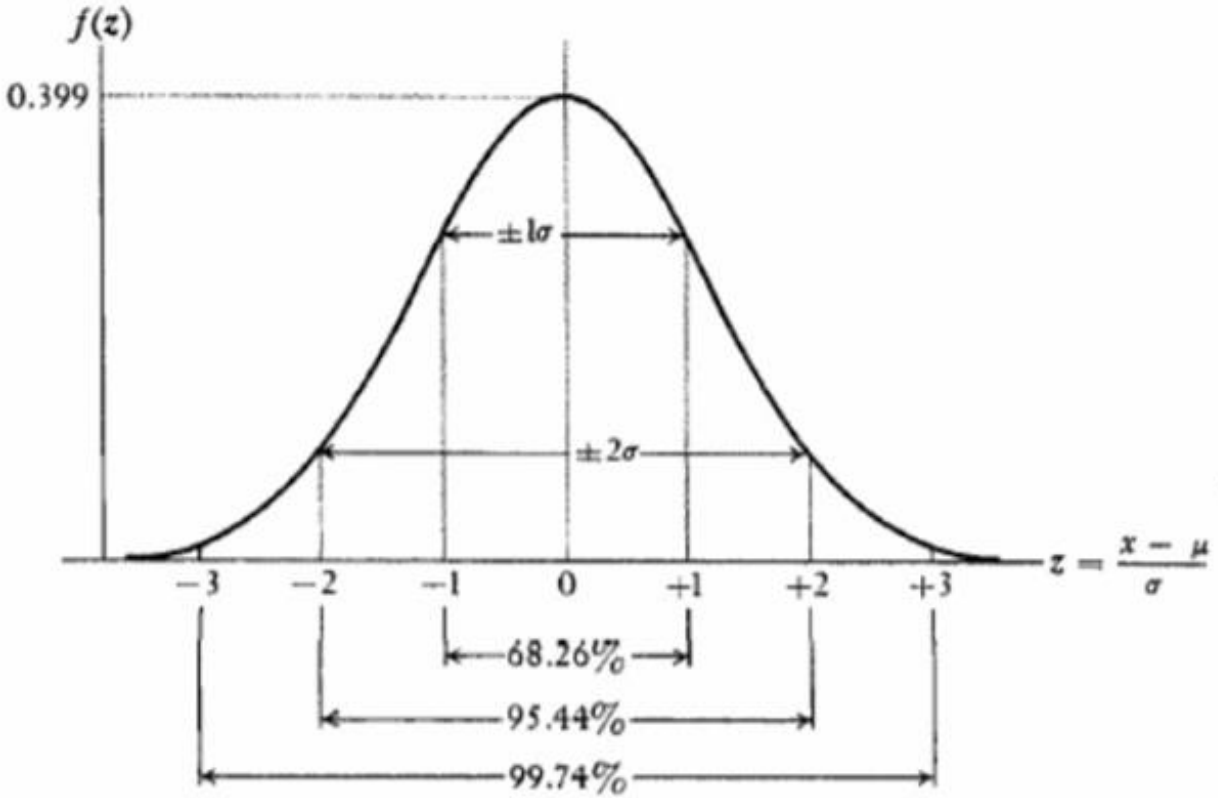
A Few Words on Normal Distribution



Brealey, Myers and Allen; Principles of Corporate Finance. 10th, 11th, or 12th editions. Chapter 8

A Few Words on Normal Distribution

Standard Normal Distribution



Classical Normal Linear Regression Model (CNLRM)

The estimation of sample parameters is not complete without hypothesis testing $(\widehat{\beta_0}, \widehat{\beta_1})$

- It is important to draw inferences about population parameters using sample estimates, more clearly, we would like the estimated parameters to be as close as possible to population parameters
- It must be noted that the randomness in the beta (coefficient) estimates is introduced by μ_i (error term): How?
- Thus, these sample coefficient estimates also have a probability distribution [as one take different samples from population, one gets different estimates]

The Normality Assumption of the Error Term μ_i

To make any inference about the probability distribution of the estimate, we need to make some assumption about the distribution of the error term μ_i

- The CNLRM assumes that μ_i is distributed normally with the following:
- $Mean = E(u_i) = 0$
- $Variance = E[u_i - E(u_i)]^2 = E(u_i)^2 = \sigma^2$
- $Covariance = E\left[[u_i - E(u_i)][u_j - E\left(u_j\right)]\right] = E\left(u_i, u_j\right) = 0;\ i \neq j$
- These assumptions are summarised as $\boldsymbol{u_i \sim N(0, \sigma^2)}$

Properties of OLS Estimators under Normality

Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

- Under the normality assumption, OLS estimates are unbiased, efficient, consistent (estimates converge to their to population values as sample size increases) $Y_i \; = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{\mu_i}$;   where $\hat{Y_i} \; = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- Mean: $E(\widehat{\beta_1}) = \beta_1$; Variance = $var\left(\widehat{\beta_1}\right) = \sigma^2_{\widehat{\beta_1}}$; then $\boldsymbol{\widehat{\beta_1}} \sim \boldsymbol{N(\beta_1, \sigma^2_{\widehat{\beta_1}})}$

Properties of OLS Estimators under Normality

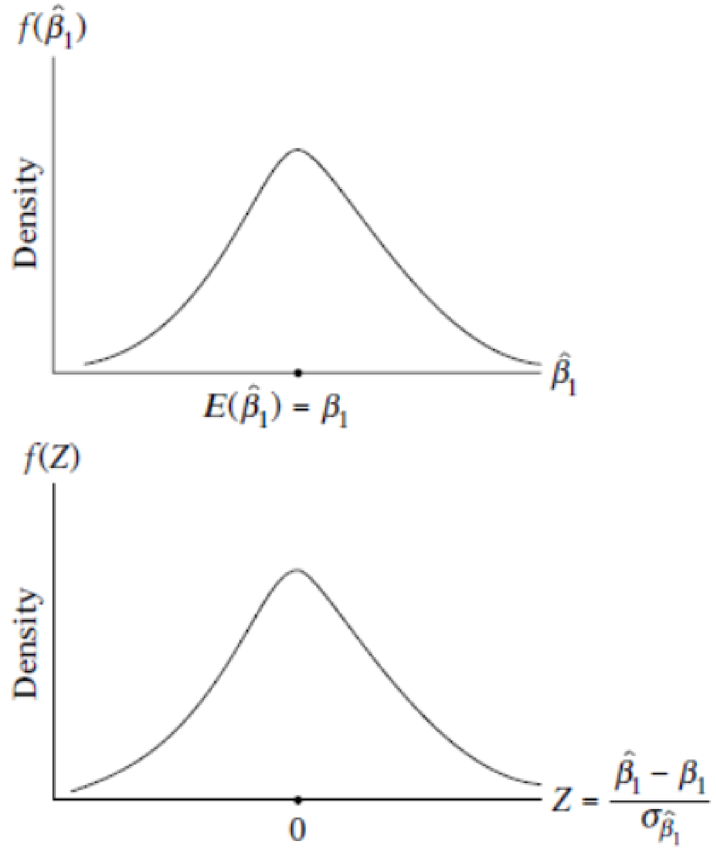
Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

- By the properties of standard normal distribution $Z = \frac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}}$
- Where $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{1})$: Z is normally distributed with mean of 0, and SD=1

Properties of OLS Estimators under Normality: Summary

Normal distributions are very easily defined with just two parameters, i.e., mean and variance of the population

Mean: $\boldsymbol{E(\widehat{\beta_1})=\beta_1}$; Variance = $\boldsymbol{var(\widehat{\beta_1}) = \sigma^2_{\widehat{\beta_1}}}$; then $\boldsymbol{\widehat{\beta_1} \sim N(\beta_1, \sigma^2_{\widehat{\beta_1}})}$



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 4)

Classical Normal Linear Regression Model (CNLRM) and Hypothesis testing II

Interval Estimation and Hypothesis Testing

While in repeated sampling the point estimate $\widehat{\beta_1}$ converges to true population parameter, i.e., $E(\widehat{\beta_1})$=β_1, but the accuracy of this point estimate is important: How reliable is this estimate

- This is so because the single estimate differs from true value; this reliability of the estimate is measured by its standard error

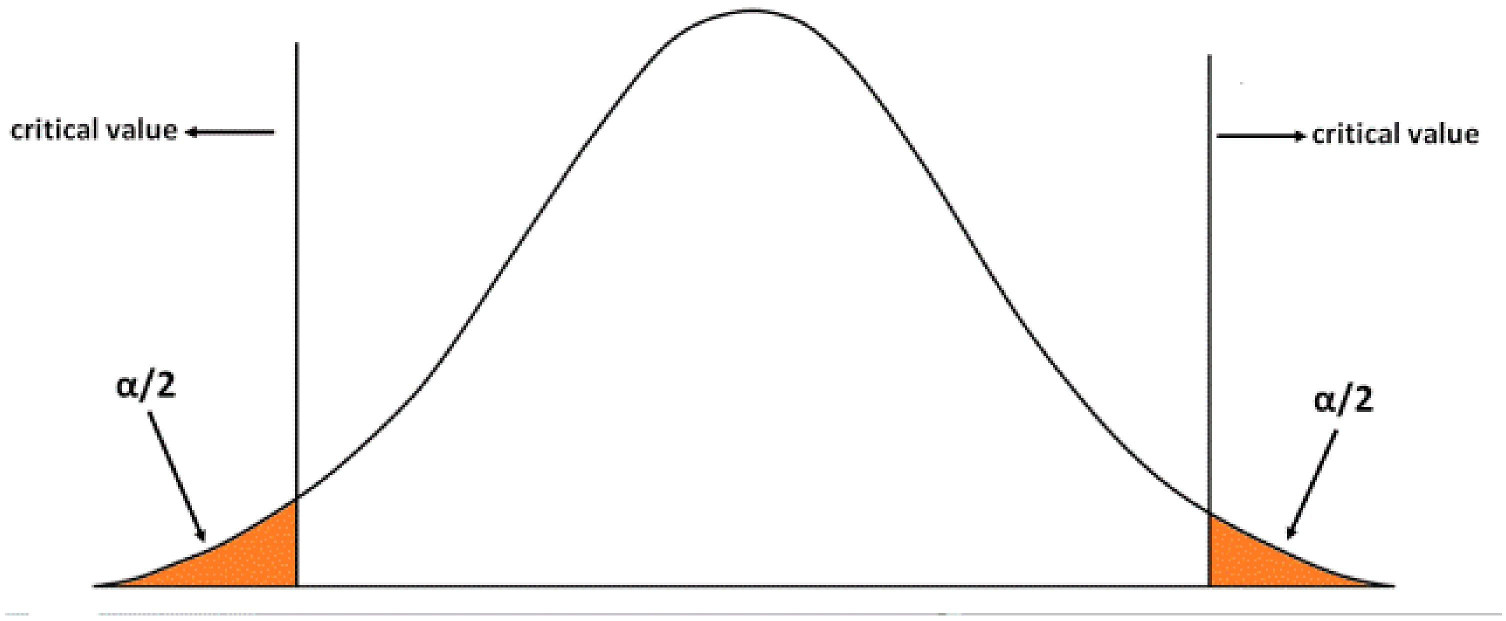
Interval Estimation and Hypothesis Testing

In the OLS estimation each parameter $(\widehat{\beta_0}, \widehat{\beta_1})$ is estimated with some error

- The square-root of the variance of the estimated parameter indicates that error in estimation or the precession of the estimate

Interval Estimation and Hypothesis Testing

In statistics we configure the confidence interval around the estimate



Interval Estimation and Hypothesis Testing

For example, if you hypothesize that the population parameter = β_1; then you set-up a confidence interval [1- α] around the estimate β_1

- If the estimate does not fall in this interval, then you can reject your hypothesis at 5% significance level
- Practically, you hypothesize that coefficient is zero. That is, the X variable does not have any impact on the Y variable. Then you set-up a confidence interval around that zero value

Interval Estimation and Hypothesis Testing

Then that range can be written as ($\pm t_{\alpha/2}$); if the estimated value falls outside this value, then with a given level of confidence (1- α generally 90%, 95%, or 99%) or significance level 10%, 5%, or 1% you reject the hypothesis and state that the variable has a significant relationship

Interval Estimation and Hypothesis Testing

Null hypothesis: H0: The true population parameter is β_1 (= ‘0’ in most cases)

- Alternate hypothesis H1: The true population parameter is not β_1
- Decision rule: Construct [1- α] confidence interval for the population parameter β_1; if the estimate falls outside this value, you reject the null H0 (don’t say you accept the null hypothesis). If the estimated parameter falls inside this range, you can not reject the null

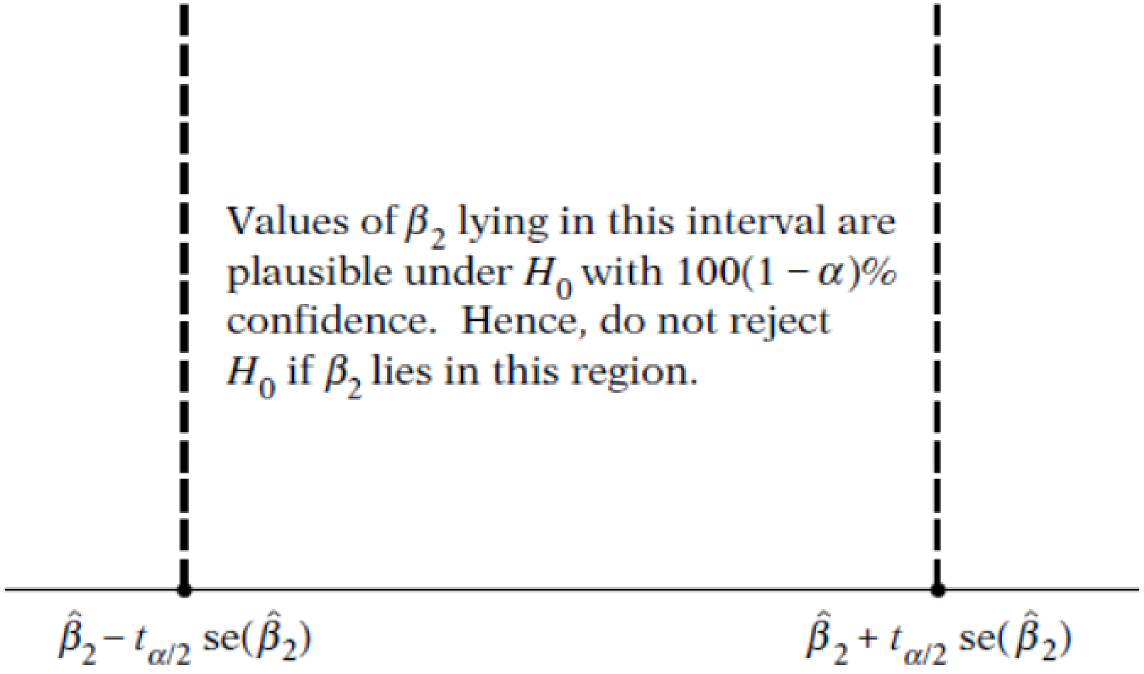
Interval Estimation and Hypothesis Testing

So if you hypothesized that β_1=0 (i.e., no impact of X on Y) and the estimate falls in the confidence interval [this is checked by looking at the t-value of the estimate] then you say that you fail to reject the null and there is no relationship between X and Y

- What if it falls outside

Interval Estimation and Hypothesis Testing

Interpretation: (1-α)% of the times the true parameter will fall within the confidence interval



Damodar N. Gujarati, *Basic Econometrics*, 4th edition onwards (Chapter 5)

Other Functional Forms and Non-linear Transformations

Other Functional Forms and Non-linear Transformations

- Log-linear or log-log model: $Y_i = \beta_1 X_i^{\beta_2} * e^{u_i}$; take natural log and transform the model as below
- $\ln(Y_i) = \ln(\beta_1) + \beta_2 \ln(X_i) + u_i$ or alternatively
- $Y_i' = \alpha + \beta_2 X_i' + u_i$: The model is now linear in parameters α and β_2
- The interpretation goes as follows: β_2 measures percentage change in Y_i for a given percentage change in X_i

Other Functional Forms and Non-linear Transformations

- Log-lin model: $Y_t = Y_0(1+r)^t$; take natural log and transform the model as below
- $\ln(Y_t) = \beta_1 + t\beta_2$
- This is a semi-log model, and β_2 measures proportional change in Y_t for a given absolute change in t
- Vice-versa interpretation goes for Lin-log model below (absolute change in $\boldsymbol{Y_t}$ for a % or relative change in $\boldsymbol{X_t}$.
- $\boldsymbol{Y_t = \beta_1 + ln(X_t)\,\beta_2}$

Summary and Concluding Remarks

Summary and Concluding Remarks

- Among supervised learning algorithms, regression algorithm is a very important tool employed in the finance domain for applications such as forecasting security prices or credit scoring
- Regression algorithms can be run with only two variables (one independent and one independent) : simple regression or with more than two variables: Multiple regression
- They key variables in a regression include a dependent variable, one or more independent variables, coefficients of these variables, and an error term

Summary and Concluding Remarks

- The error term accounts for the variation in the dependent variable that can not be explained by the model (independent variables)
- While regression analysis can provide the statistical significance of the relationship, the direction of causality should come a priori from the theoretical underpinnings (rain vs. crop example)
- OLS is the most often employed method to estimate a regression model, which involves minimizing residual sum of squares

Summary and Concluding Remarks

- OLS estimation of regression involves 10 key assumptions
- The most important assumptions include linearity in parameters, exogeneity of independent variables, zero conditional mean of the error (residual) term, homoscedasticity of error variances, absence of multicollinearity, no autocorrelation across error terms, no correlation between error and dependent variables
- If these assumptions are held then OLS estimators are referred to as BLUE, that is best linear unbiased and efficient estimates

Summary and Concluding Remarks

- The statistical significance of OLS estimators is determined through hypothesis testing of coefficients individually
- This requires normality assumption of the error (residuals)
- Very often the model is not linear and may require some kind of transformation to make it linear, which can be subsequently estimated through OLS
- However, the interpretation of coefficients also change with such transformations

Thanks!