Econometrics

Course Calendar	
Week	Main Content
Week 1	**Introduction to Simple Regression**
Week 2	**Simple Regression**
Week 3	**Simple Regression: r^2 & Hands-on-Exercise**
Week 4	**Central Limit Theorem, Probability and Probability Density Function (PDF)**
Week 5	**Hypothesis Testing: Basics**
Week 6	**Simple Regression: Testing of Hypothesis**

Introduction to Simple Regression

Geetha Rani Prakasam,
ICCR Chair Professor,
DBS, UNITECH, PNG.

Outline of Lecture

- Part 1 and Part 2
- Part 1
- Intro to Econometrics
- What are Econometric Models?
- Part 2
- Introduction to Simple Regression

Introduction to Econometrics

- Econometrics deals with the measurement of economic relationships.
- It is an integration of economics, mathematical economics and statistics with an objective to provide numerical values to the parameters of economic relationships.
- The relationships of economic theories are usually expressed in mathematical forms
  and combined with empirical economics.
- The econometrics methods are used to obtain the values of parameters which are essentially the coefficients of the mathematical form of the economic relationships.

Introduction to Econometrics

- The statistical methods which help in explaining the economic phenomenon are adapted as econometric methods.
- The econometric relationships depict the random behaviour of economic relationships which are generally not considered in economics and mathematical formulations.
- Econometric methods can be used in other areas like engineering sciences, biological sciences, medical sciences, geosciences, agricultural sciences etc.

Introduction to Econometrics

- Simply, whenever there is a need of finding the stochastic relationship in mathematical format, the econometric methods and tools help.
- The econometric tools are helpful in explaining the relationships among variables.

What are Econometric Models?

- A model is a simplified representation of a real-world process.
- It should be representative in the sense that it should contain the salient features of the phenomena under study.
- In general, one of the objectives in modeling is to have a simple model to explain a complex phenomenon.
- Such an objective may sometimes lead to oversimplified model and sometimes the assumptions made are unrealistic.
- In practice, generally, all the variables which the experimenter thinks are relevant to explain the phenomenon are included in the model.

What are Econometric Models?

- Rest of the variables are dumped in a basket called “disturbances” where the disturbances are random variables.
- This is the main difference between economic modeling and econometric modeling.
- This is also the main difference between mathematical modeling and statistical modeling.
- The mathematical modeling is exact in nature, whereas the statistical modeling contains a stochastic term also.

What are Econometric Models?

- An economic model is a set of assumptions that describes the behaviour of an economy, or more generally, a phenomenon.
- An econometric model consists of:-
  - a set of equations describing the behaviour. These equations are derived from the economic model and have two parts i.e., observed variables and disturbances.
  - a statement about the errors in the observed values of variables &
  - a specification of the probability distribution of disturbances

Learning Outcomes of Econometrics

- Specify simple and multiple regression equations
- Estimate the parameters and constants and evaluate the statistical significance of the parameters
- Interpret the estimated results and assess the goodness of a fit of regression equations
- Understand the use of dummy variables as explanatory variables and their application in multiple regressions
- Apply the econometric models in testing economic theories.

Econometrics and Statistics

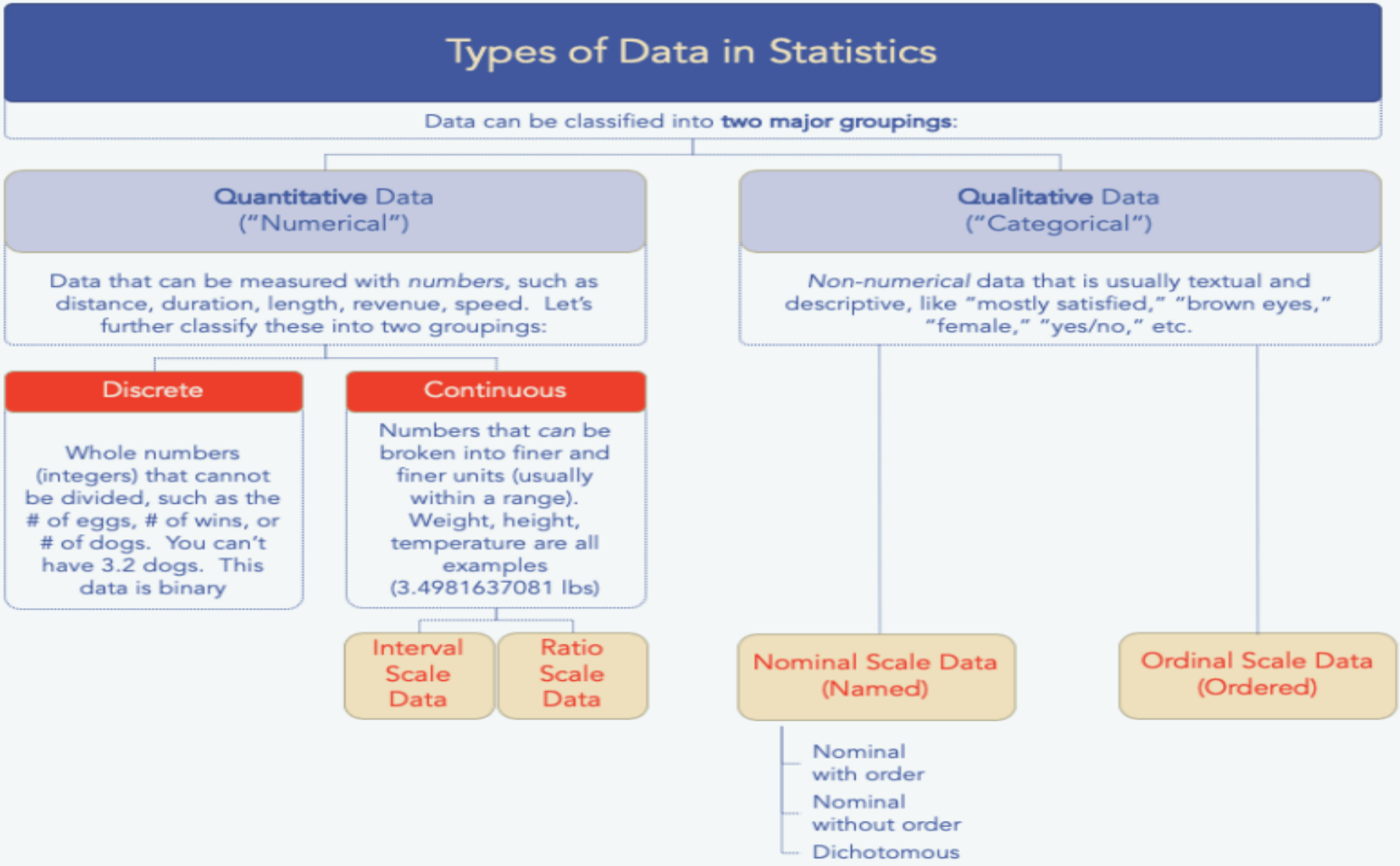
- Econometrics differs both from mathematical statistics and economic statistics.
- In economic statistics, the empirical data is collected recorded, tabulated and used in describing the pattern in their development over time.
- The economic statistics is a descriptive aspect of economics.
- It does not provide either the explanations of the development of various variables or measurement of the parameters of the relationships.

Econometrics and Statistics

- Statistical methods describe the methods of measurement which are developed on the basis of controlled experiments.
- Such methods may not be suitable for the economic phenomenon as they don’t fit in the framework of controlled experiments.
- For example, in real-world experiments, the variables usually change continuously and simultaneously, and so the set up of controlled experiments are not suitable.

Econometrics and Statistics

- Econometrics uses statistical methods after adapting them to the problems of economic life.
- These adopted statistical methods are usually termed as econometric methods.
- Such methods are adjusted so that they become appropriate for the measurement of stochastic relationships.
- These adjustments basically attempt to specify attempts to the stochastic element which operate in real-world data and enters into the determination of observed data.
- This enables the data to be called a random sample which is needed for the application of statistical tools.



https://www.mymarketresearchmethods.com/data-types-in-statistics/

Type of Data by Sources

- Time series data
- Cross-section data
- Panel data

Introduction to Simple Regression

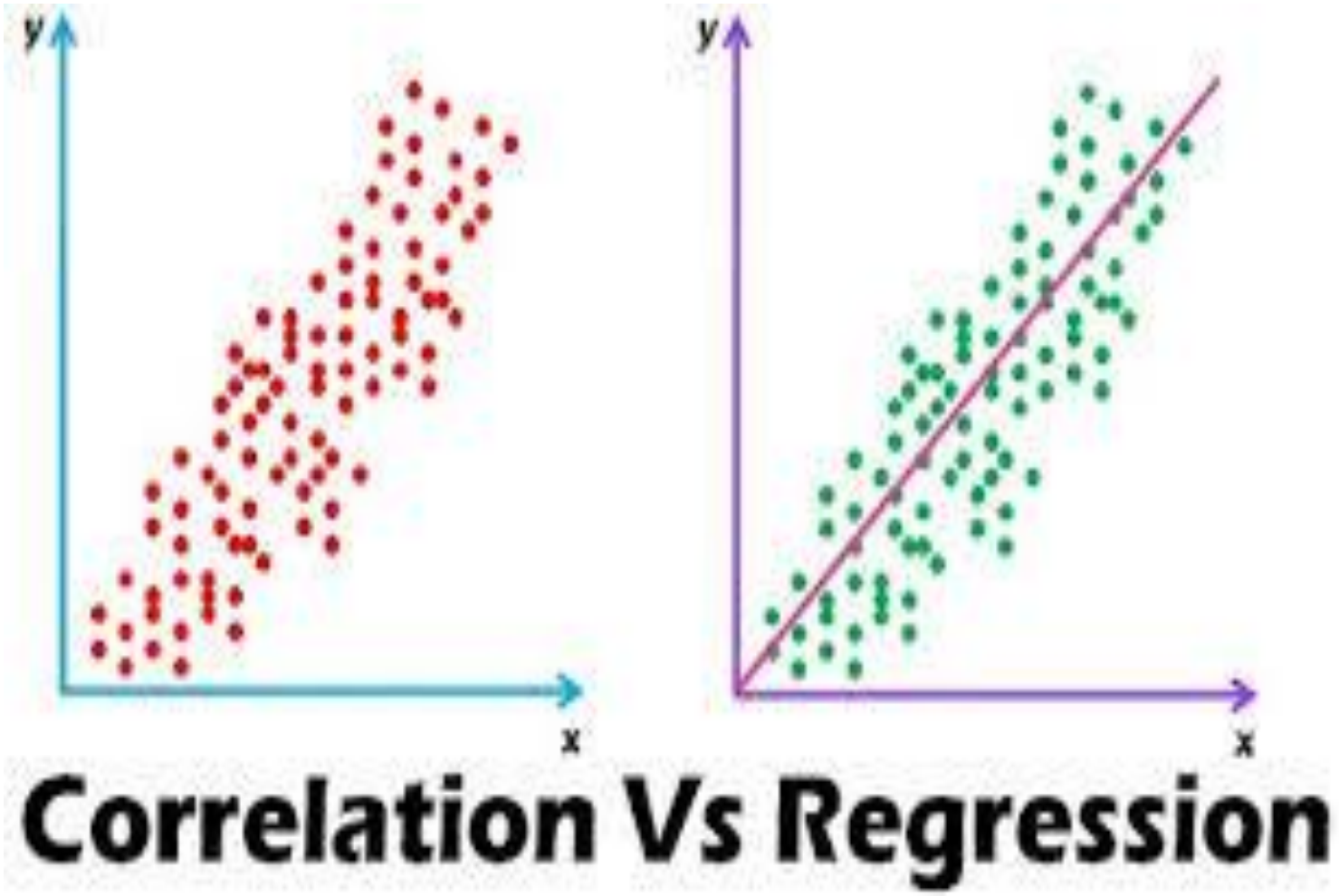
- With the brief introduction in the earlier slides, we now move to learn simple regression.
- We start with the outline of this lecture.

Outline_Part 2 of Lecture

- Regression vs. Correlation
- origin of the term Regression
- Modern Interpretation of Regression Analysis
- Regression vs. Causation
- What is regression?
- Two-variable Regression Analysis: Some basic Ideas
- PRF and SRF
- Summary
- References

Regression vs. Correlation

- Correlation Analysis: the primary objective is to measure the strength or degree of linear association between two variables (both are assumed to be random)
- Regression Analysis: we try to estimate or predict the average value of one variable (dependent, and assumed to be random / stochastic) on the basis of the fixed values of other variables (independent, and non-stochastic)



Source: https://keydifferences.com/difference-between-correlation-and-regression.html.

Historical origin of the term Regression

- The term REGRESSION was introduced by Francis Galton
- Tendency for tall parents to have tall children and for short parents to have short children, but the average height of children born from parents of a given height tended to move (or regress) toward the average height in the population as a whole (F. Galton, "*Family Likeness in Stature*")

Historical origin of the term Regression

- Galton's Law was confirmed by Karl Pearson: The average height of sons of a group of *tall* fathers < their fathers' height.
- And the average height of sons of a group of *short* fathers > their fathers' height.
- Thus "regressing" tall and short sons alike toward the average height of all men. (K. Pearson and A. Lee, "*On the law of Inheritance*")
- By the words of Galton, this was "*Regression to mediocrity*"

- galton-board-large.mp4

Source: https://www.youtube.com/watch?v=EvHiee7gs9Y

Modern Interpretation of Regression Analysis

- Regression Analysis is concerned with the study of the dependence of one variable (*The Dependent Variable*), on
- one or more other variable(s) (*The Explanatory Variable*),
- with a view to estimating and/or predicting the (population) mean or average value of the former in term of the known or fixed (in repeated sampling) values of the latter.

Statistical vs. Deterministic Relationships

- In regression analysis we are concerned with STATISTICAL DEPENDENCE among variables (not Functional or Deterministic), we essentially deal with RANDOM or STOCHASTIC variables (with the probability distributions)

Regression vs. Causation:

- Regression does not necessarily imply causation.
- A statistical relationship cannot logically imply causation.
- "A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other" (*M.G. Kendal and A. Stuart, "The Advanced Theory of Statistics"*)

Terminology and Notation

Dependent Variable

↓↑

Explained Variable

↓↑

Predictand

↓↑

Regressand

↓↑

Response

↓↑

Endogenous

Explanatory Variable(s)

↓↑

Independent Variable(s)

↓↑

Predictor(s)

↓↑

Regressor(s)

↓↑

Stimulus or control variable(s)

↓↑

Exogenous(es)

What is regression?

- The key idea behind regression analysis is the statistical dependence of one variable on one or more other variable(s)
- The objective of regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable on the basis of the known or fixed values of the explanatory variable(s).
- The success of regression depends on the available and appropriate data .
- The researcher should clearly state the sources of the data used in the analysis, their definitions, their methods of collection, any gaps or omissions and any revisions in the data

Basic Econometrics: Chap 2

- TWO-VARIABLE REGRESSION ANALYSIS: **Some basic Ideas**
- PRF
- SRF

2-1. A Hypothetical Example

- Total population: 60 families
- Y=Weekly family consumption expenditure
- X=Weekly disposable family income
- 60 families were divided into 10 groups of approximately the same income level

  (80, 100, 120, 140, 160, 180, 200, 220, 240, 260)

Table 2-1: Weekly family income X ($), and consumption Y ($)

Y \ X	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
Mean	65	77	89	101	113	125	137	149	161	173

2-1. A Hypothetical Example

- Table 2-1 gives the *conditional distribution* of Y on the given values of X
- Table 2-2 gives the *conditional probabilities* of Y: $p(Y \mid X)$
- *Conditional Mean* (or Expectation): $E(Y \mid X=X_i)$ or simply
- $E(Y \mid X_i)$

Table 2.2: Conditional Probabilities p(Y|Xi) for the Data of Table 2.1

X_i →	Weekly family income									
P(Y / Xi) ↓	80	100	120	140	160	180	200	220	240	260
Conditional Prob. of P(Y/Xi)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	--	1/6	--	1/7	1/6	1/6	--	1/7	1/6	1/7
	--	--	--	1/7	--	--	--	1/7	--	1/7
Conditional Means of Y, E(Y/X)	65	77	89	102	113	125	137	149	161	171

Figure 2-1:population regression line

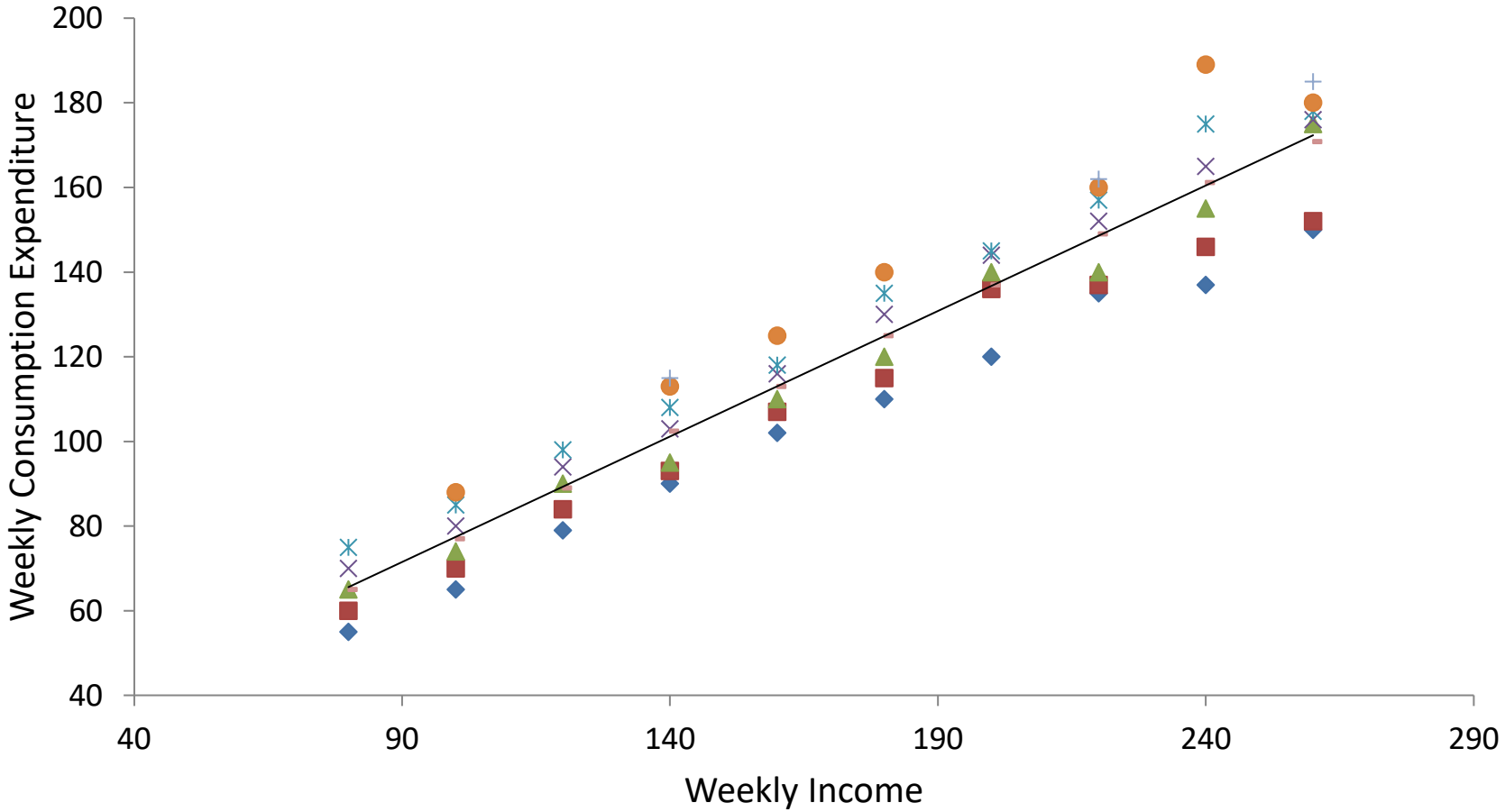


Fig 2-1 & 2.2: population regression line

- Figure 2-1 shows the population regression line (curve). It is the regression of Y on X.
- Geometrically, population regression curve is the locus of the conditional means or expectations of the dependent variable for the fixed values of the explanatory variable (s). This to be drawn as (Fig.2-2).
- Shows that each X (i.e., income level) there is a population of Y values (weekly consumption expenditures) that are spread around the (conditional) mean of those Y values.
- For simplicity, we are assuming that these Y values are distributed symmetrically around their respective (conditional) mean values. And the regression line (or curve) passes through these (conditional) mean values.

2-2. The concepts of (PRF)

- $E(Y \mid X=X_i) = f(X_i)$ --- (2.2.1) is Population Regression Function (PRF) or Population Regression (PR)
- It states merely that the expected value of the distribution of Y given Xi is functionally related to Xi.
- In simple terms, it tells how the mean or average response of Y varies with X.
- In the case of linear function we have linear population regression function (or equation or model)

  $E(Y \mid X=X_i) = f(X_i) = ß_1 + ß_2X_i$ ---- (2.2.2)

2-2. The concepts of (PRF)

$$E(Y \mid X=X_i) = f(X_i) = ß_1 + ß_2X_i$$

- $ß_1$ and $ß_2$ are regression coefficients, $ß_1$is intercept and $ß_2$ is slope coefficient
- Linearity in the Variables
- Linearity in the Parameters: *mean a regression that is linear in the parameters; the ß's (that is, the parameters) are raised to the first power only.*
- *It may or may not be linear in the explanatory variables, the X's.*

2-4. Stochastic Specification of PRF

- $U_i = Y - E(Y \mid X=X_i)$ or $Y_i = E(Y \mid X=X_i) + U_i$ ---(2.4.1)
- $Yi = ß_1 + ß_2X_i + U_i$ ---- (2.4.2)
- U_i = Stochastic disturbance or stochastic error term. It is nonsystematic component
- Component $E(Y \mid X=X_i)$ is systematic or deterministic.
- It is the mean consumption expenditure of all the families with the same level of income
- $Y1 = 55 = ß_1 + ß_2(80) + U_1$ ---- (2.4.3)
- $Y2 = 60 = ß_1 + ß_2(80) + U_2$ ---- (2.4.3)
- $Y5 = 75 = ß_1 + ß_2(80) + U_5$ ---- (2.4.3)

Stochastic PRF

- If we take the expected value of (2.4.1) on both sides,
- $E(Y_i \mid X_i) = E[E(Y_i \mid X_i] + E(u_i \mid X_i)$
- $= E(Y_i \mid X_i) + u_i X_i)$ ------(2.4.4)
- The assumption that the regression line passes through the conditional means of Y implies that $E(U_i \mid X_i) = 0$ ----(2.4.5)
- This specification in (2.4.5)has the advantage – that there are other variables besides income that affect consumption expenditure
- An individual family's consumption expenditure cannot be fully explained by the variable(s) included in the regression model.

2-5. The Significance of the Stochastic Disturbance Term

- U_i = Stochastic Disturbance Term is a surrogate for all variables that are omitted from the model but they collectively affect Y
- Many reasons why not include such variables into the model as follows:

2-5. The Significance of the Stochastic Disturbance Term

Why not include as many as variable into the model (or the reasons for using u_i)

+ *Vagueness of theory*
+ *Unavailability of Data*
+ *Core Variables vs. Peripheral Variables*
+ *Intrinsic randomness in human behavior*
+ *Poor proxy variables*
+ *Principle of parsimony*
+ *Wrong functional form*

2-6. The Sample Regression Function (SRF)

- Task now is to estimate the PRF on the basis of sample information.
- Can we estimate the PRF from the sample data?
- We may not be able to estimate the PRF ‘accurately’ because of sampling fluctuations.

2-6. The Sample Regression Function (SRF)

Table 2-4: A random sample from the population

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Table 2-5: Another random sample from the population

Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

2-6. The Sample Regression Function (SRF)

- Which of the two regression lines represents the ‘true’ population regression line?
- Does any of them represent PR line / curve?
- Regression lines in figure 2.3 are known as sample regression lines.
- Approximation of the true PR
- N different SRFs for N different samples and these SRFs are not likely to be the same.

2-6. The Sample Regression Function (SRF)

- Fig.2-3: SRF1 and SRF 2
- $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ --------(2.6.1)
- $\hat{Y}_i$ = estimator of $E(Y \mid X_i)$
- $\hat{\beta}_1$ = estimator of β_1
- $\hat{\beta}_2$ = estimator of β_2
- Estimate = A particular numerical value obtained by the estimator in an application
- SRF in stochastic form: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_I$ ----(2.6.2)
  or $Y_i = \hat{Y}_i + \hat{u}_i$ -------(2.6.3)

2-6. The Sample Regression Function (SRF)

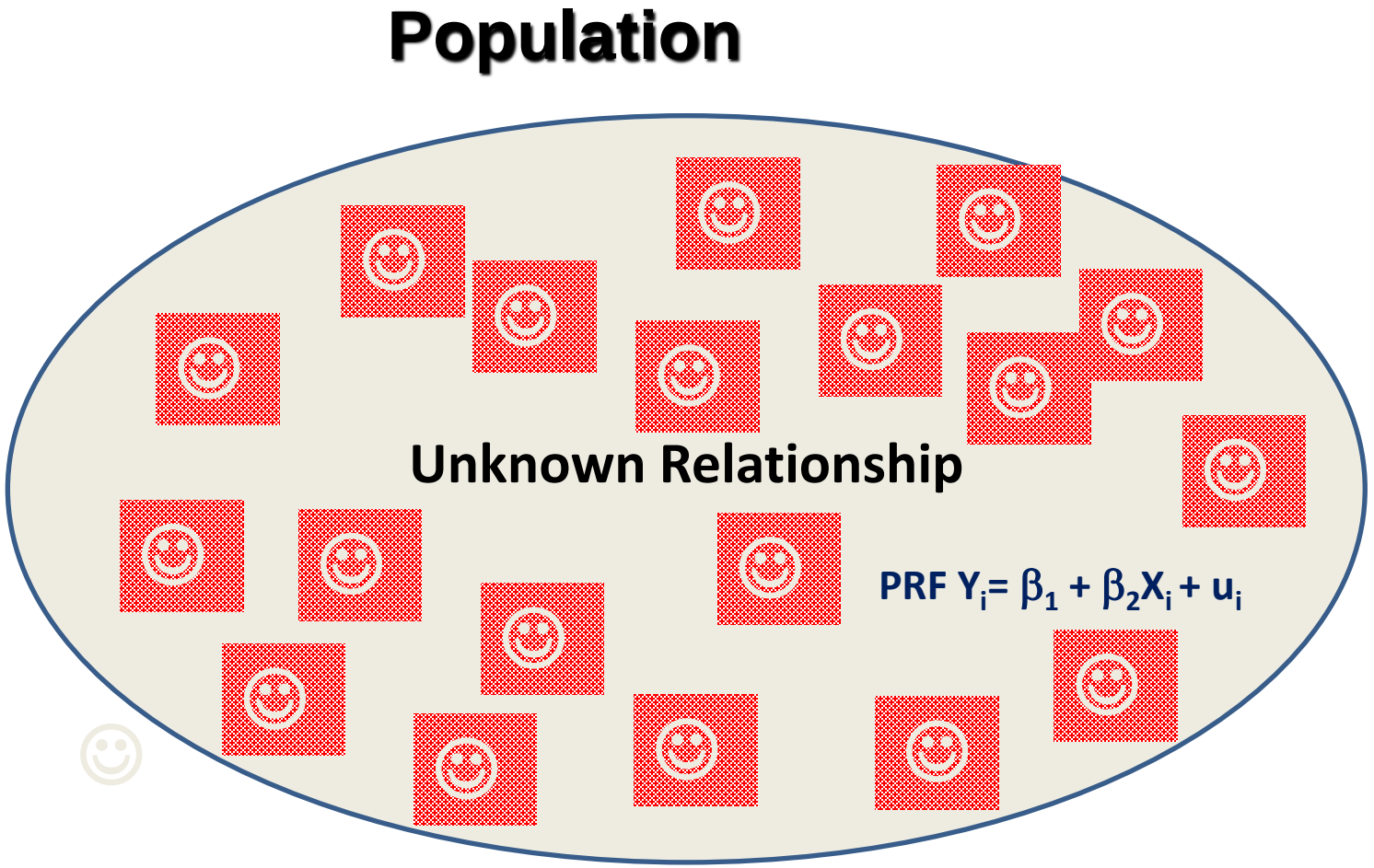
- **Primary objective in regression analysis is to estimate the**
- **PRF $Y_i = \beta_1 + \beta_2 X_i + u_i$**
  - **on the basis of the**
- **SRF $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$**
- **and how to construct SRF so that $\hat{\beta}_1$ close to β_1 and $\hat{\beta}_2$ close to β_2 as much as possible.**

2-6. The Sample Regression Function (SRF)

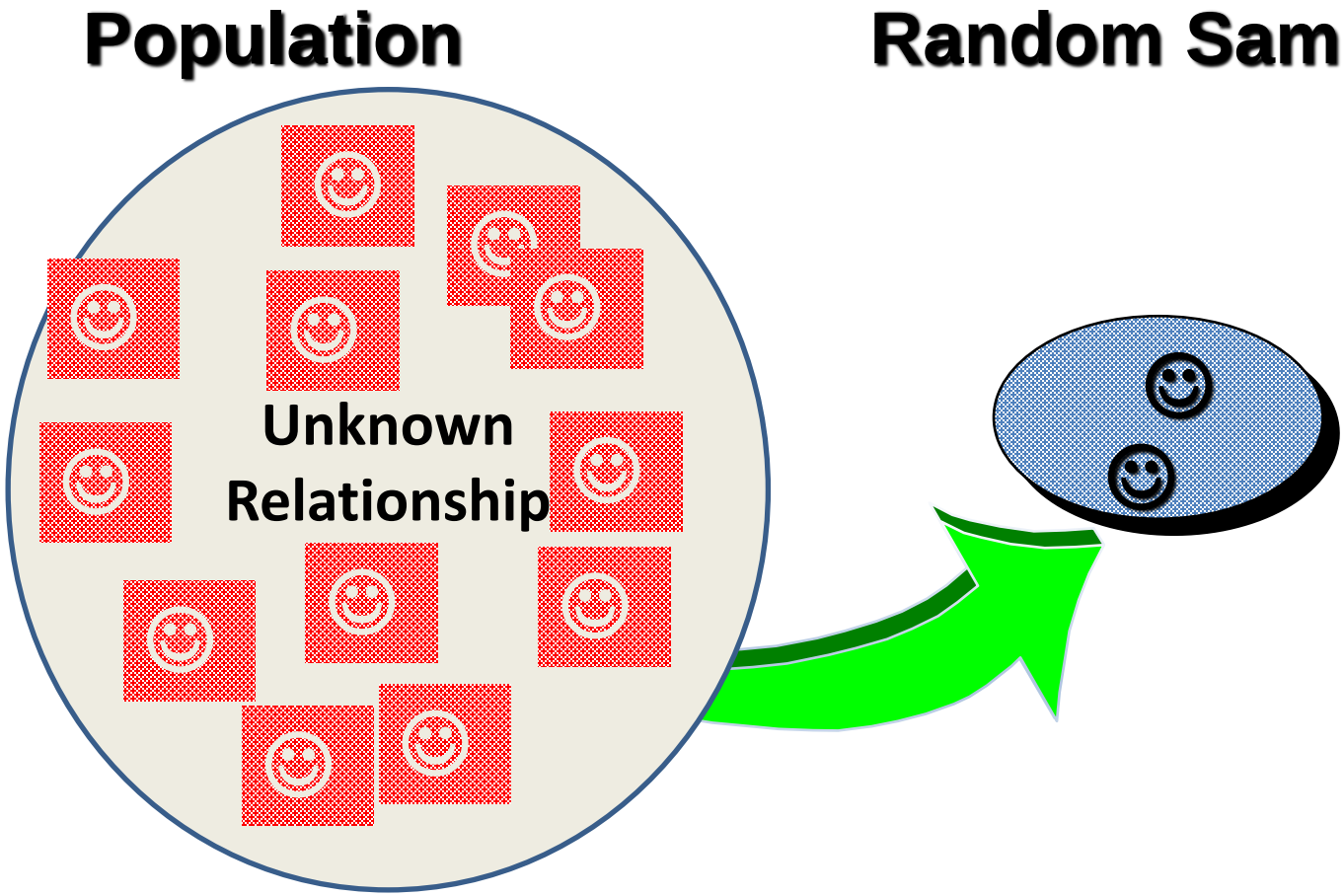
- Population Regression Function PRF
- Linearity in the parameters
- Stochastic PRF
- Stochastic Disturbance Term u_i plays a critical role in estimating the PRF
- Sample of observations from population
- Stochastic Sample Regression Function SRF is used to estimate the PRF

Figure 2.3: Regression line based on two samples

Population & Sample Regression Functions



Population & Sample Regression Functions



2-7. Summary and Conclusions

- The key concept underlying regression analysis is the concept of the population regression function (PRF).
- This book Basic Econometrics by D. Guajarati (many basic Econometrics books) deals with linear PRFs: linear in the unknown parameters.
- They may or may not be linear in the variables.

2-7. Summary and Conclusions

- For empirical purposes, it is the stochastic PRF that matters. The stochastic disturbance term u_i plays a critical role in estimating the PRF.
- The PRF is an idealized concept, since in practice one rarely has access to the entire population of interest.
- Generally, one has a sample of observations from population and use the stochastic sample regression (SRF) to estimate the PRF.

References

- Econometrics, Chapter 1, Introduction to Econometrics Shalabh, IIT Kanpur, India
- Gujarati Damodar N, and Dawn C. Porter, (2003), Basic Econometrics 5th Edition, Chapter 1 and 2.
- Surbhi, S, (2021), Difference between Correlation and Regression, available at https://keydifferences.com/difference-between-correlation-and-regression.html.
- Galton Board video, available at https://www.youtube.com/watch?v=EvHiee7gs9Y