Index

Sunday, 02 October 2022 8:22

Week	Lectures
Week 1	1.1 Introduction to data visualization 1.2 Defining the message 1.3 Creating Designs Tutorial 1.1: Misleading Visuals Tutorial 1.2: Building a simple dashboard
Week 2	2.1 Probability Distributions 2.2 Business Example 2.3 Guessing the Distribution Identifying and fitting distributions Kurtosis 2.4 Guessing the Distribution of Dataset 1 2.5 Guessing the Distribution of Dataset 2 2.6 Guessing the Distribution of Dataset 3
Week 3	3.1 Determining association between Categorical variables 3.2 Bayes' Rule 3.3 Inferring association between categorical variables - Chi-squared test for Independence 3.4 Chi-squared test of Independence - Implementation in Python 3.5 Chi-squared test of Independence - Implementation in Spreadsheets Formulae
Week 4	4.1 Demand Response Curve 4.2 Elasticity 4.3 Linear Response Curve 4.4 Estimation Problem in Demand Response Curves 4.5 Analysing Linear Demand Response Curve using Simple Linear Regression 4.6 Tutorial- "Building a Simple Linear Regression Model on Python" Formulae
Week 5	5.1 Non-linear Demand Response Curve 5.2 Analysing Constant Elasticity Model using Simple Linear Regression 5.3 Implementing Constant Elasticity Model using Simple Linear Regression 5.4 Optimal Pricing - Revenue Maximization 5.5 Optimal Pricing - Profit Maximization 5.6 Revenue Maximization vs Profit Maximization 5.7 Operations Research: Linear Programming and Duality in Spreadsheets and Python Excel work Primal-Dual 5.8 Implementing Constant Elasticity Model using Simple Linear Regression in Python Formulae
Week 6	6.1 Multiple Linear Regression 6.2 Multiple Linear Regression - Example 6.3 Multiple Linear Regression - Path Diagram 6.4 Multiple Linear Regression - Variance Inflation Factor - Part 1 6.5 Multiple Linear Regression - Variance Inflation Factor - Part 2 6.6 Multiple Linear Regression - Implementation in Python Formulae
Week 7	7.1 Logistic Regression - Predicting the Placements 7.2 Logistic Regression - Working with data 7.3 Logistic Regression - Model Building 7.4 Logistic Regression - Model Evaluation 7.5 Logistic Regression - Interpretation of the Coefficients 7.6 Tutorial- Logistic Regression in Python Formulae
Week 8	8.1 Measuring the Efficiency of a Business Unit 8.2 Efficiency Comparison - Graphical Method 8.3 Optimization Method - Data Envelopment Analysis 8.4 Data Envelopment Analysis - Example with one output and two inputs Formulae
	9.1 Data Envelopment Analysis - Example with two outputs and one input 9.2 Data Envelopment Analysis - Example with multiple outputs and multiple inputs

Week 9	9.3 Data Envelopment Analysis - Prescription for inefficient units (One output and two inputs case) 9.4 Data Envelopment Analysis - Prescription for inefficient units (Two outputs and one input case)
Week 10	10.1 Consumer Choice Models 10.2 Forms of Conjoint Analysis 10.3 Conjoint Problem 10.4 Optimization Formulation of Conjoint Problem 10.5 Problem Specific Notations
Week 11	11.1 Conjoint Problem Formulation Using Linear Programming 11.2 Solving the Conjoint Problem 11.3 Conjoint Analysis using a Statistical Method 11.4 Regression Method for Conjoint Analysis
Week 12	12.1 Epilogue 12.1 Epilogue- Continuation

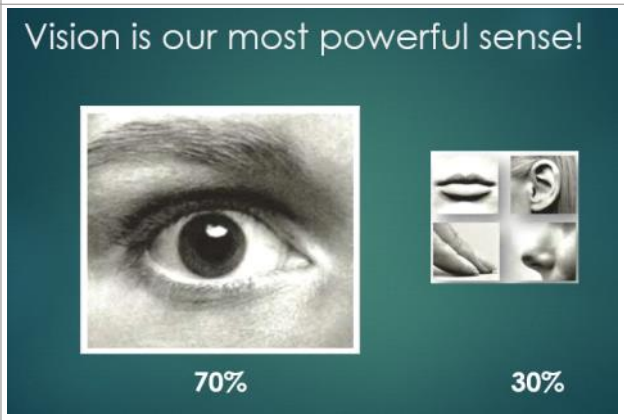
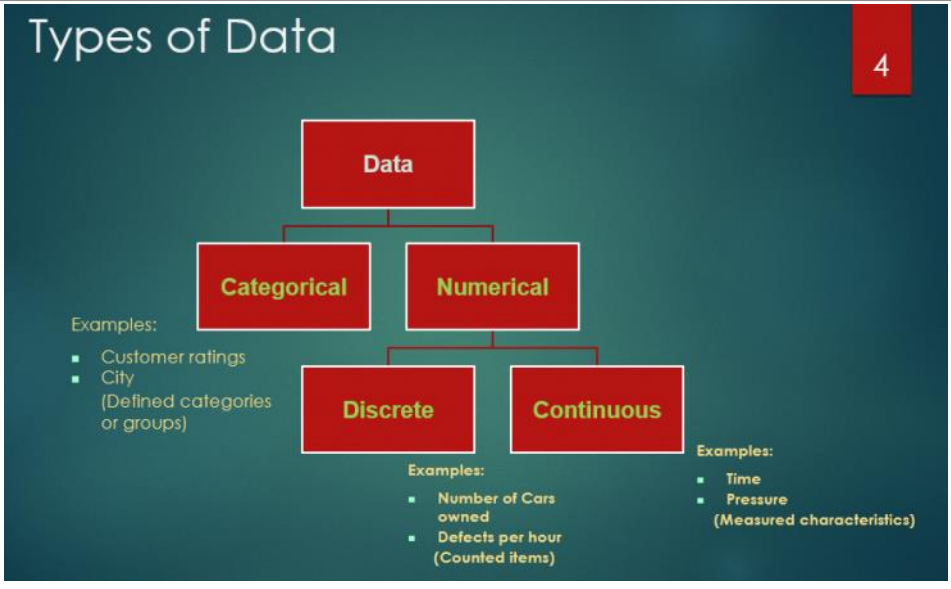
Summary	•

Week 1

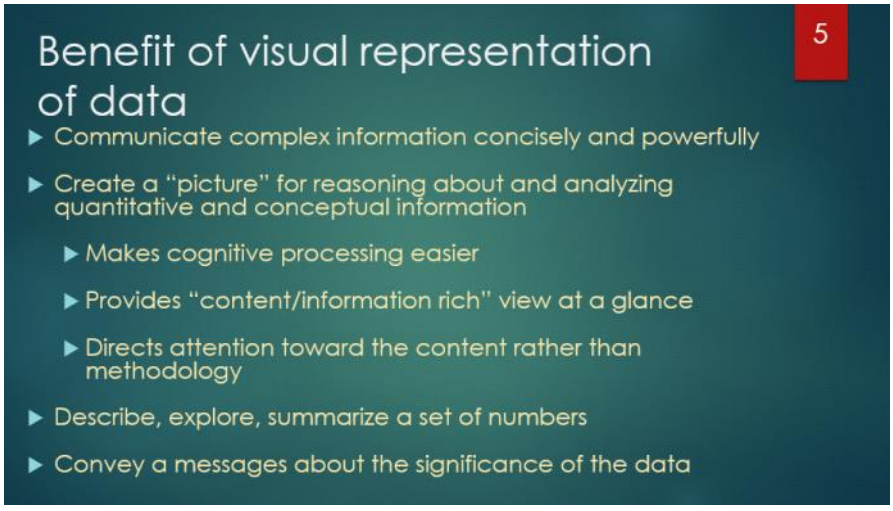
Saturday, 08 October 2022 14:47

1.1 Introduction to data visualization

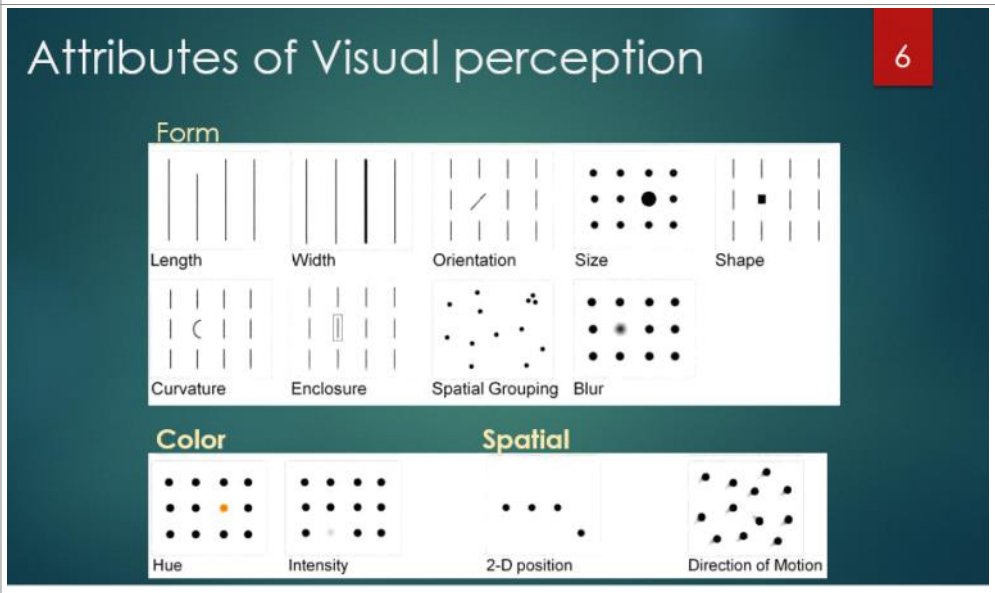
Sunday, 02 October 2022 8:06

Summary	• Types of data • Benefit of visual representation of data • Attributes of visual perception • **Four umbrella principles of effective visualization** • Three-step process of executing an information display
	
	Good Decisions Are Based On An Accurate Understanding Of Good Data 
	 • 70% of the info we consume is visual. So, visual communications is very important. • Visualization is the language of analysts.
• What are the types of data?	 • Technique is more important than visualization tools. • **Trending charts** (*eg,* line charts) **don't make sense on categorical data.** (also, the intervals we make e.g., 0-10, 10-20, etc. also comes in categorical category) • Take some time to contemplate on how you're gonna represent the data/result. It will save you hours of unnecessary debate when you're presenting your result.

- What are the benefits of visual representation?



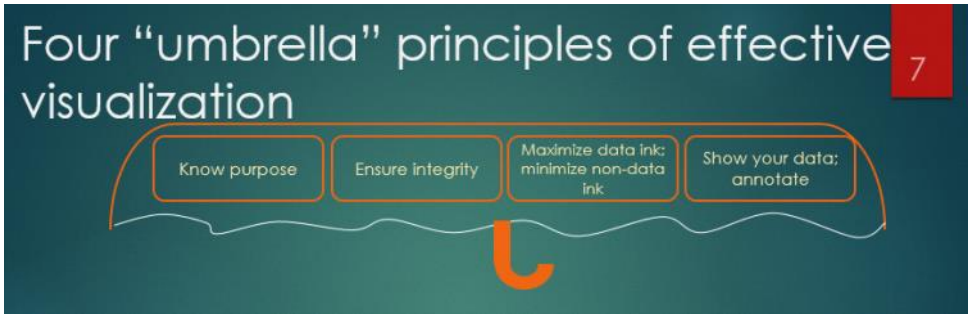
- Job of an analyst: Business has to improve. They do analysis to help business improve. And, they need to communicate it to the stakeholders so that right decisions can be made.
- Sometimes, the visual representation itself are so complicated that people spend more time understanding the visuals than what it is trying to communicate.



How do you help them focus on what is important, and how do you get them not focus on what is not important.

- What are the four umbrella principles of effective visualization?

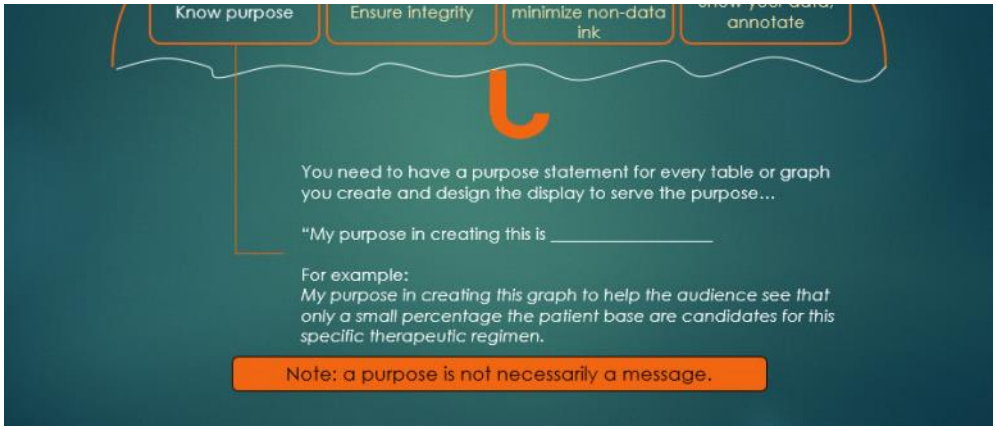
These conditions are necessary but not sufficient for optimal representation of data.



- When you present a chart, these are the four things that you have to absolutely follow.
- Write it down and go through this checklist when creating a visual.

- Describe the four umbrella principles one by one:
  1. Know your purpose

- A purpose is not necessarily a message. How?

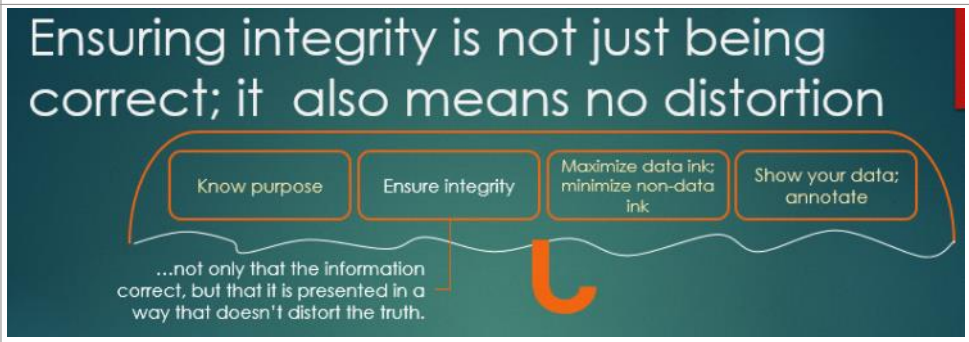


- **Why does this representation has to be here?**
- Once you have a purpose, it will automatically determine the form of the representation.
- Sometimes, a problem can't be solved by using just one visual representation.
  In that case each chart or visual will serve as one step in the larger problem that you're trying to solve.
- When I put a representative graphic together, the purpose of that should say this is going to make my ability to communicate this complicated concept more easier.

- Umbrella principles means it covers everything.
- What is the purpose of having the handle that is shown?
  - The handle emphasizes the concept of the umbrella principles. So, it is there for a reason.
- Literally, everything that's there, there has to be a reason.
  So, if you look at it and say why is this there? What is the purpose of having that particular graphic on the screen?
  You should have a clear purpose.

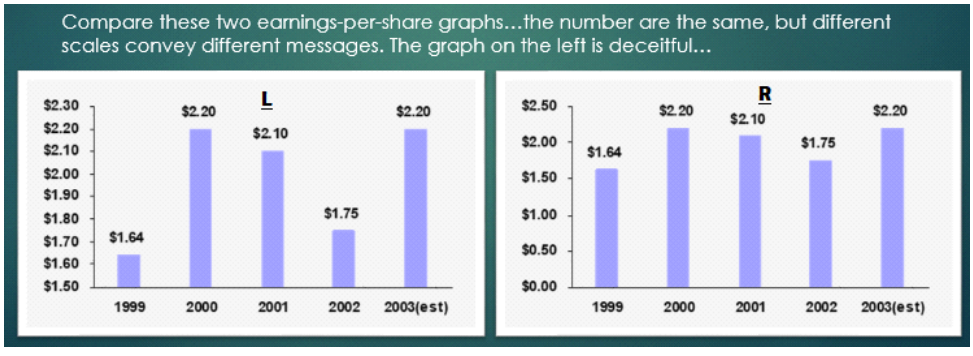
- A purpose is not necessarily a message.
- A message could be communicated over a series of visual representations, and each of those will have a purpose to contribute towards communicating that message.

2. Ensuring integrity



- Typically you will have errors of omission and commission(?) when you're presenting data.
- Even a typo can render the integrity of the presentation questionable.

- Compare these two graphs (on how the truth can be distorted):



- It seems L is so volatile, but in reality it is not.
  When we look 1999 to 2000 graph in L, one would say we grew 6x (we're very quick to measure lengths), but's not the reality.
- L: The y-axis has been cut off.
- The integrity has been lost. The stories can be made biased hugely using these tactics.

3. Maximize data ink; minimize non-data ink

- Spend ink on the item you want to show.
  Do not spend ink on items that are not critical to the thought process.
- Any color ink that that does not serve a purpose and communicate no information to the end user is non-data ink.
- Styling should be very specific for a purpose.

- Compare these two charts on ink and non-data ink

L

	CALLS/DAY	DAYS/YEAR	CALLS/YEAR
PCP	8	200	1,600
NEUROLOGY	5	200	1,000
CV	5.5	200	1,100
ONCOLOGY	5	200	1,000
VIROLOGY	5	200	1,000
TRANPLANT	5	200	1,000
P/T DERM	8	200	1,600

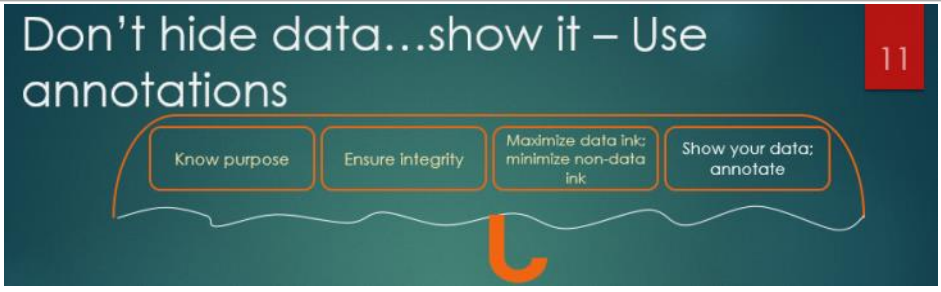
R

Table 1 -- Call Totals by Specialty

	Calls/Day	Days/Year	Calls/Year
PCP	8.0	200	1,600
Neurology	5.0	200	1,000
CV	5.5	200	1,100
Oncology	5.0	200	1,000
Virology	5.0	200	1,000
Transplant	5.0	200	1,000
P/T Derm	8.0	200	1,600

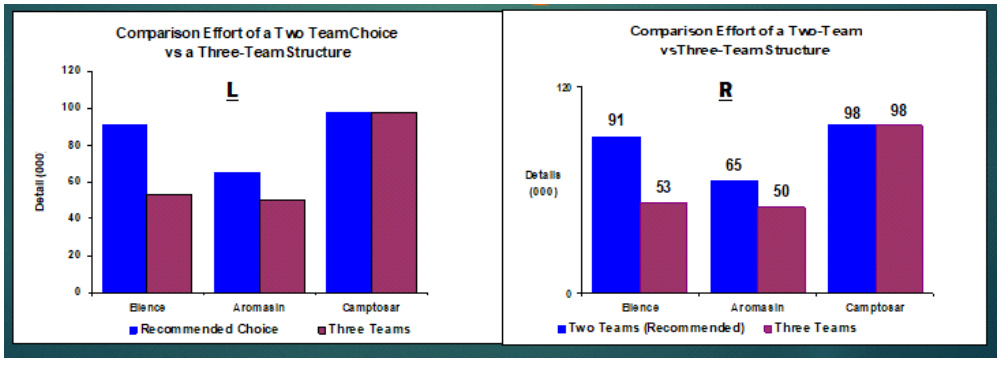
- L: All the gridlines and the yellow colours are all non-data ink. It doesn't convey any additional info.
- R: The title does communicate something. So it is data ink.

4. Annotations

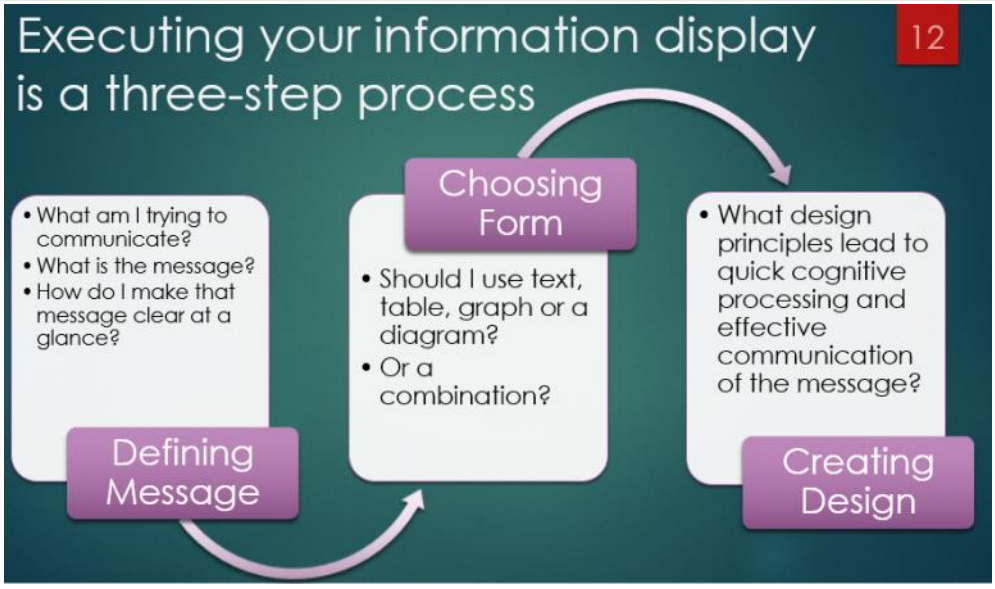


Annotate to help the users.

- Compare these two graphs on annotations.



- L: gives you an axis, and then you have to use a finger or a ruler across this to measure and read off what those bars are.
- R: It's simpler to get rid of the axes tick marks and annotate the data. It visually looks so much easier to read.

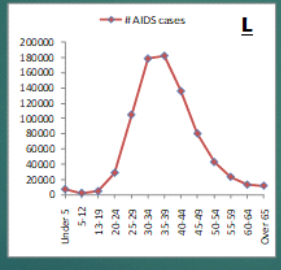
• Is it necessary to annotate every single point? What if when the graphs get crowded?	--- • Now, as you have more data points and the graphs get more crowded, annotation starts to look a bit messy. • In that case, you do not have to annotate every single point. You annotate critical points.
• Executing information display is a three-step process. What are those three steps? Explain one-by-one.	 **Defining Message:** How do you ensure that when the person sees the graph, they get the same message that you're trying to communicate? ○ Somebody is going to pull up that graph without the benefit of having you next to them to explain that graph. ○ It's your responsibility to ensure that anybody who sees it should be able to understand what it says. It should be understood at a glance. **Choosing form:** Sometimes the best way to communicate is to just write a paragraph. There is nothing that says that you have to show it graphically. Choose a form of the visual display that is best aligned to the message you're trying to communicate. **Creating Design:** Design principles come into play to ensure that visual cognitive cognition is easy.
	Choosing the right form is easier because once you get a clear message, the form almost suggests itself. But the design principle is what makes a difference.

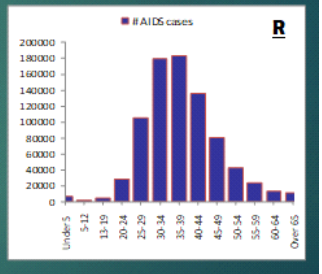
1.2 Defining the message

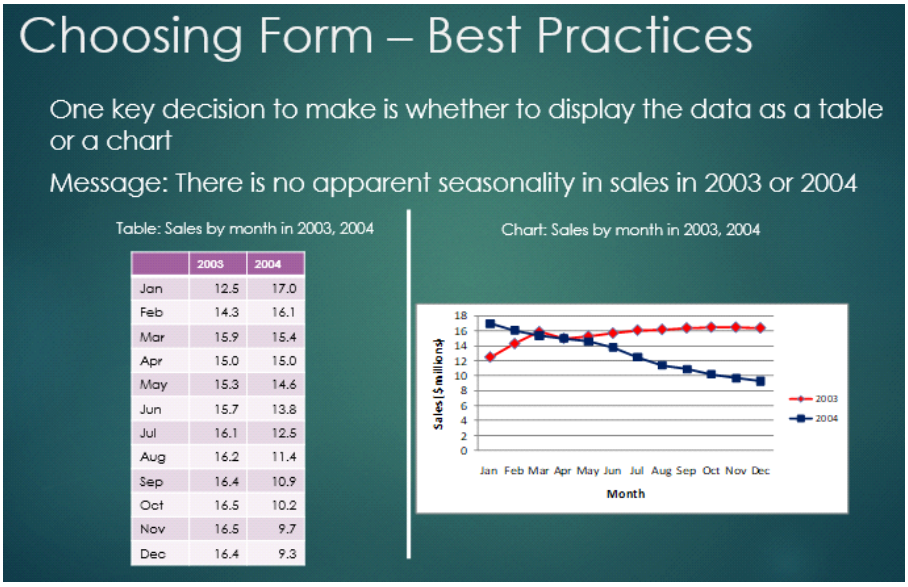
Tuesday, 04 October 2022 4:48

Summary	• Table vs Chart • **Choosing appropriate form** of graphical representation • Examples
	Defining message
	Defining Message You've been asked to produce a data display that provides insight into the population of people with AIDS. To the right is some data about the ages of people with AIDS in the U.S. Given this data, what is the message that you would like to convey? Majority of AIDS cases are in the 25-44 age groups AIDS is most prevalent in the 35-39 age groups
	Choosing form
• For the given data, what form would be more appropriate? Why?	**Choosing Form** Below are three ways to display the data to communicate the message from the distribution of AIDS sufferers. What is the most appropriate form? • R: tells you very clearly where the peak is. • L: Uses a line graph for categorical data, which is absurd.

Age Group	# AIDS cases
Under 5	6,975
5-12	2,099
13-19	4,428
20-24	28,665
25-29	10,5060
30-34	179,164
35-39	182,857
40-44	136,145
45-49	80,242
50-54	42,780
55-59	23,280
60-64	12,898
Over 65	11,555
TOTAL	816,148







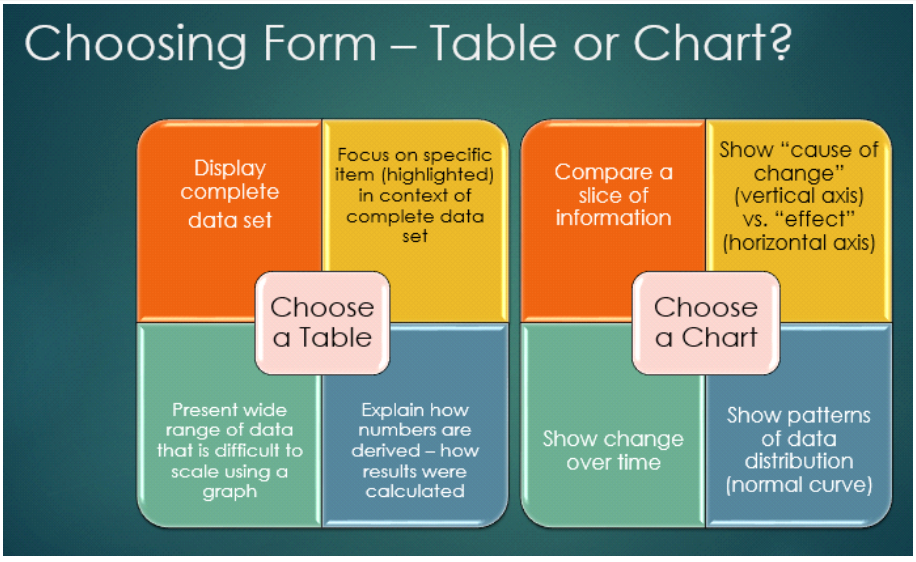
	2003	2004
Jan	12.5	17.0
Feb	14.3	16.1
Mar	15.9	15.4
Apr	15.0	15.0
May	15.3	14.6
Jun	15.7	13.8
Jul	16.1	12.5
Aug	16.2	11.4
Sep	16.4	10.9
Oct	16.5	10.2
Nov	16.5	9.7
Dec	16.4	9.3

- The tabular display can't show that seasonality.
  But the seasonality is very clear in the graphical display.

- To understand the trend in the tabular form takes effort. Much quicker with the graphical representation.

- When to use a table?

- When to use a chart?



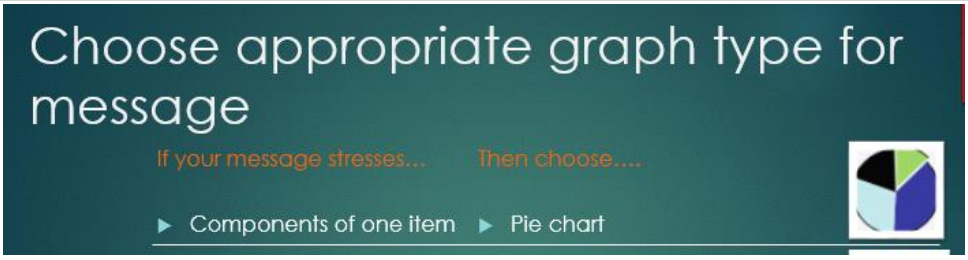
- Table: We're just showing the data, we may not necessarily be inferring something from there.
- I'm trying to factually tell you *e.g.*, what the sales is of every product category. I'm not trying to tell that A is bigger that B, or B is bigger than C. I'm just trying to factually communicate that by stating these are my sales numbers.

- So, **data tables** are used for two things:
  - To just **factually communicating the numbers without drawing inferences.**
  - And **when you want to show the calculations**.
  - Also, if you have **a wide number of columns and a whole number of rows**, a graphical chart display might become overly complicated in which the **simplicity** mandates that you just show it as a table.

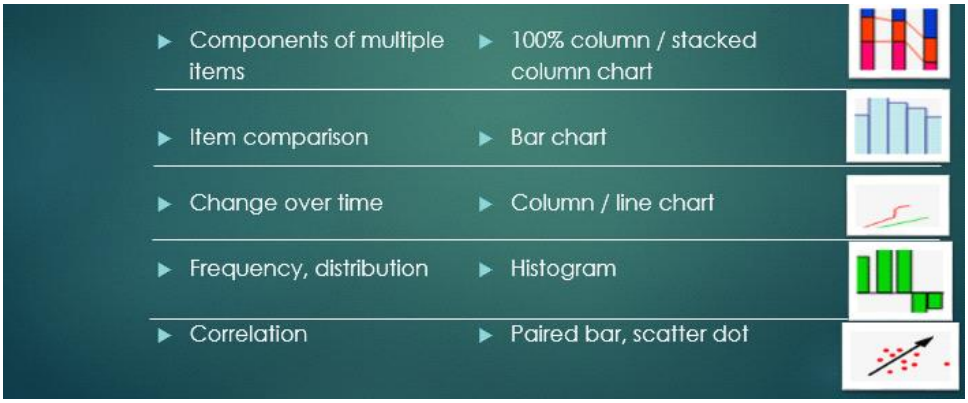
- **Chart**: used when you want **to show patterns, change over time, or to compare two things.**

What form would be appropriate when you want to show:

1. Components of one item
2. Components of multiple items



3. Item comparison (categorical)
4. Change over time
5. Frequency distribution
6. Correlation

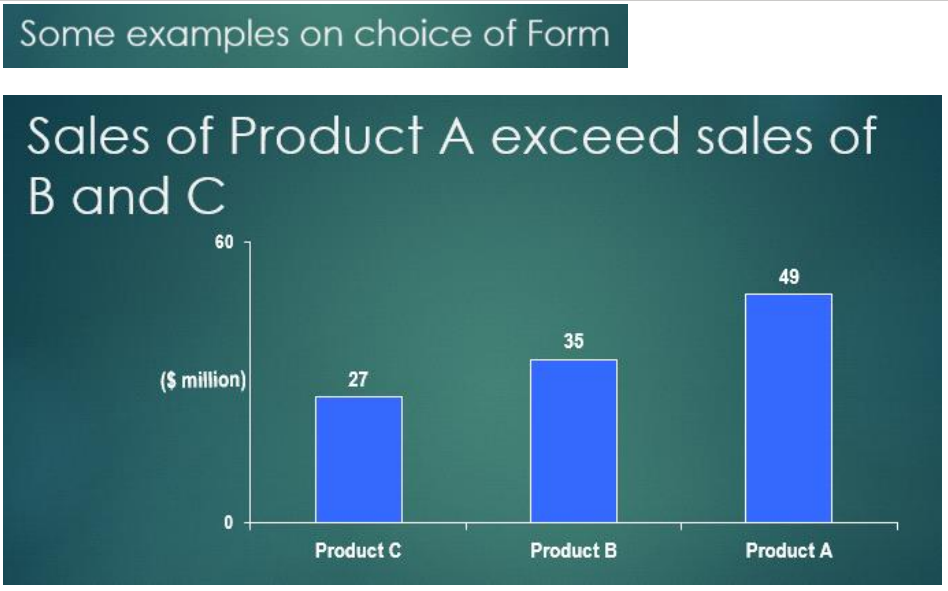


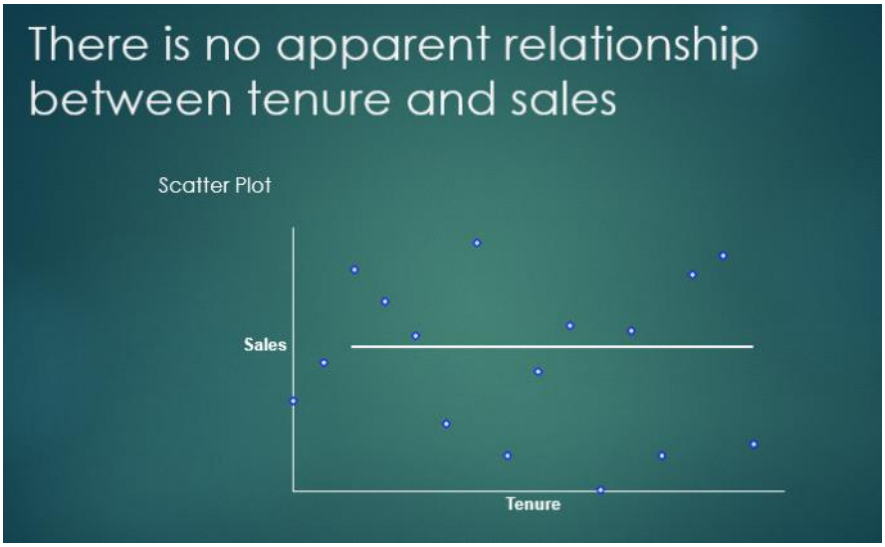
- Why you should avoid using pie charts?

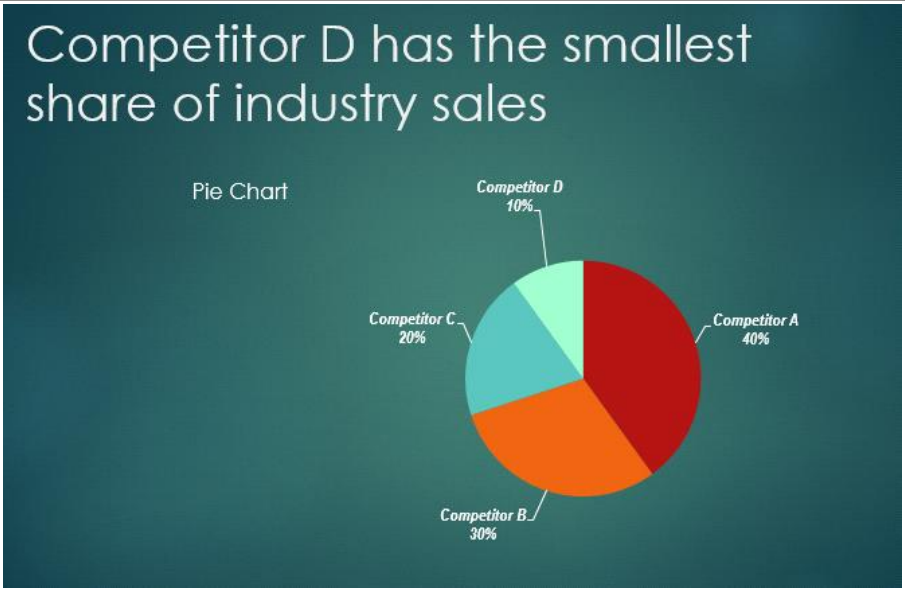
- Pie charts are very difficult to interpret and probably the most grossly misinterpreted type of all charts.

- Perceptually, **we are very good at perceiving relative differences in length. We are not very good at perceiving relative differences in area,** especially when the shapes are different.
- So, if I draw a rectangle and a square and a circle, and I ask you which of these covers a bigger surface area? It is difficult to tell unless you annotate.
- That is what you're asking people to do. You're asking them to compare two different area of slightly different shapes, and trying to determine which has a higher area.
- The same pie chart could be equivalently shown as a column chart, and the message comes across very clearly.

- Once you get the message, the form is very evident.
  - Correlation => Scatter plot
  - Change over time => line chart, etc.

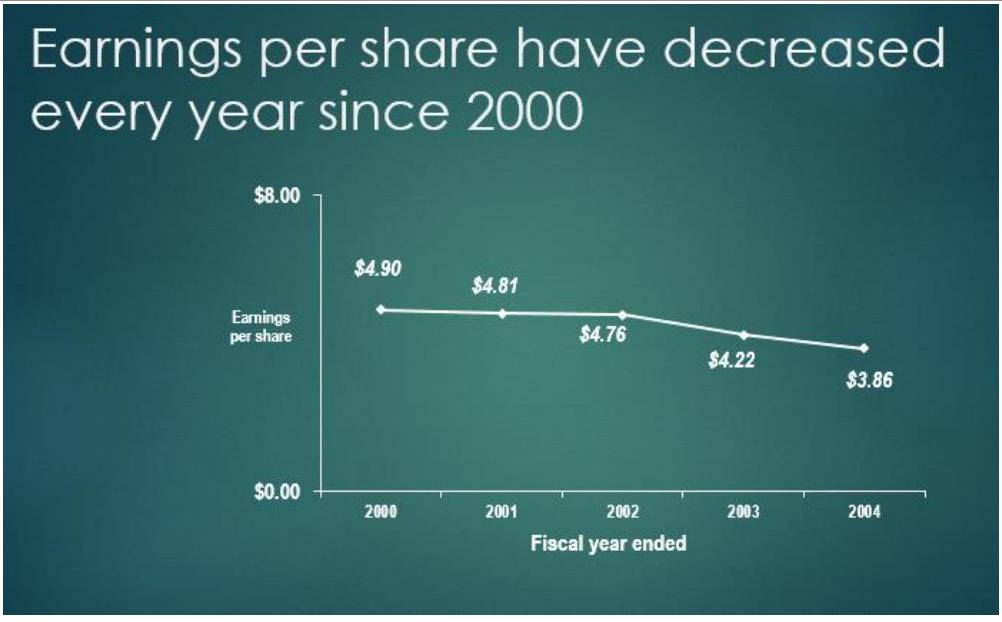


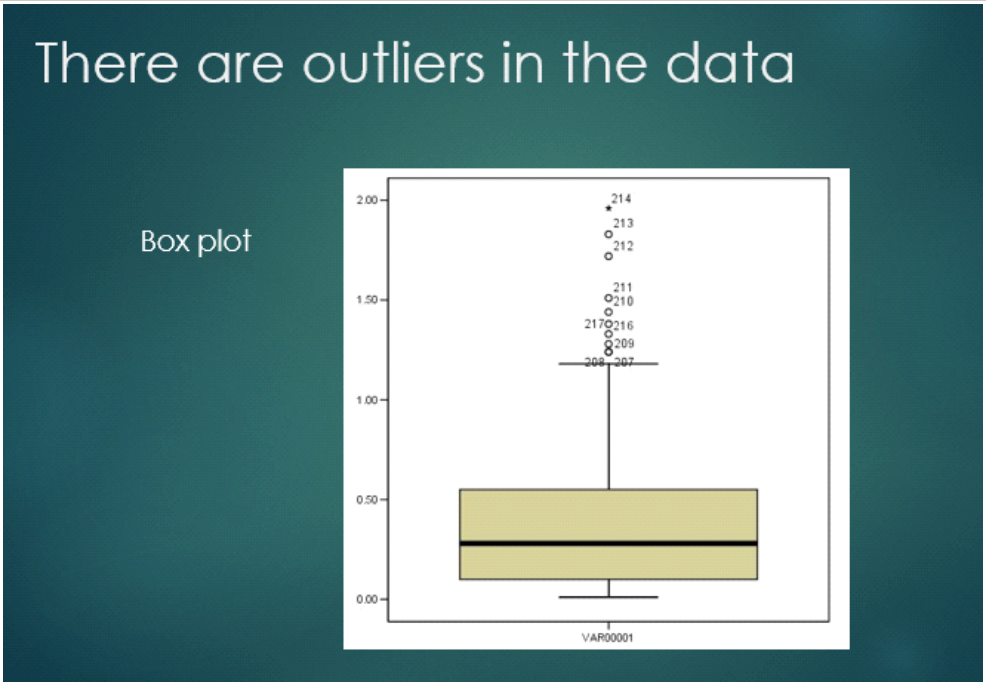




- If you have to use **pie chart, annotation is must.**
- Without it, we wouldn't be able to tell the relative strength between B and C, for example.

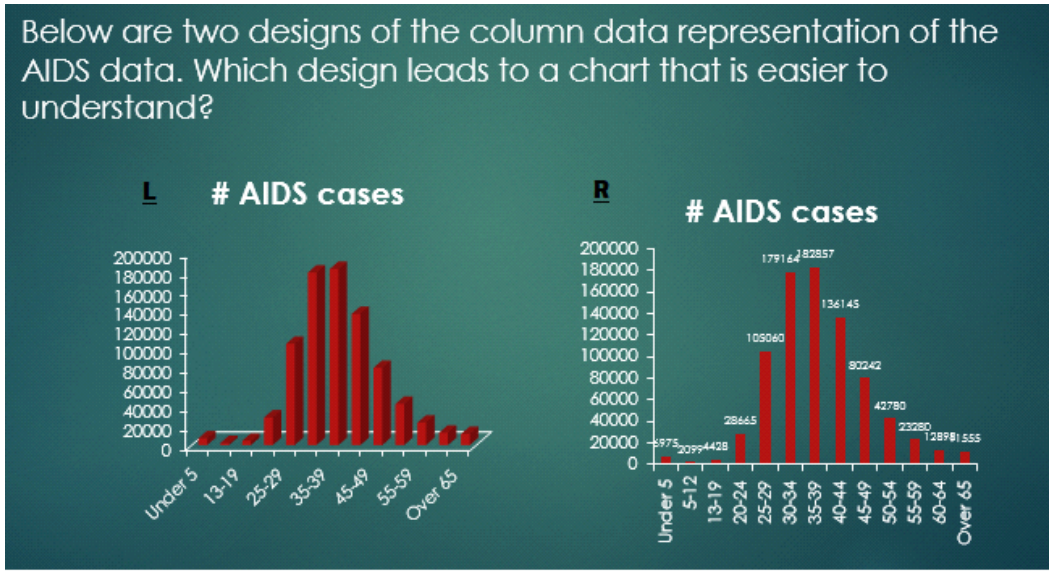
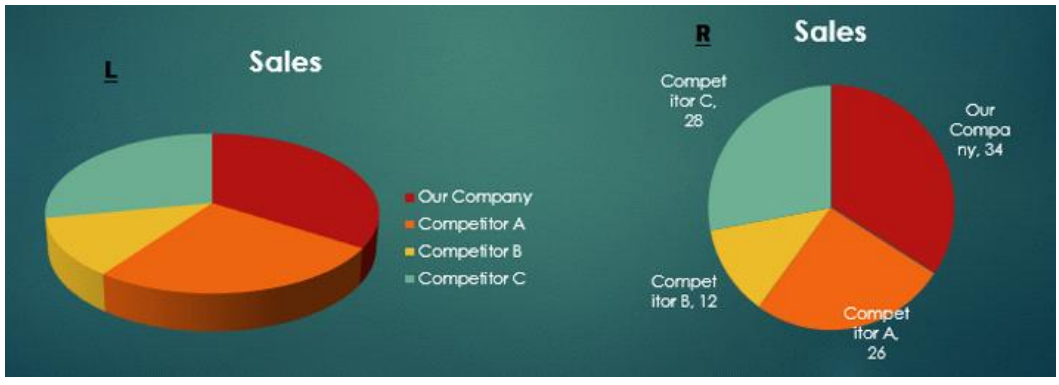
- Here, message plays an important role too. We're showing the D has the smallest share, so pie chart makes sense. But if we were to show the relative strengths between A, B, C and D, in that case pie chart wouldn't be appropriate.



• Recognizing outliers	 • It is very **important to be able to tell outliers and regular events.** • As you enter the workforce, the biggest thing you will see is people will pick up an outlier and make a whole case as if that is a regular event. And they'll say that they have anecdotal information. Ask them to draw a box plot and show where it sticks, and if that anecdote sits within the big yellow bar consider it, otherwise just discard it.

1.3 Creating Designs

Wednesday, 05 October 2022 12:34

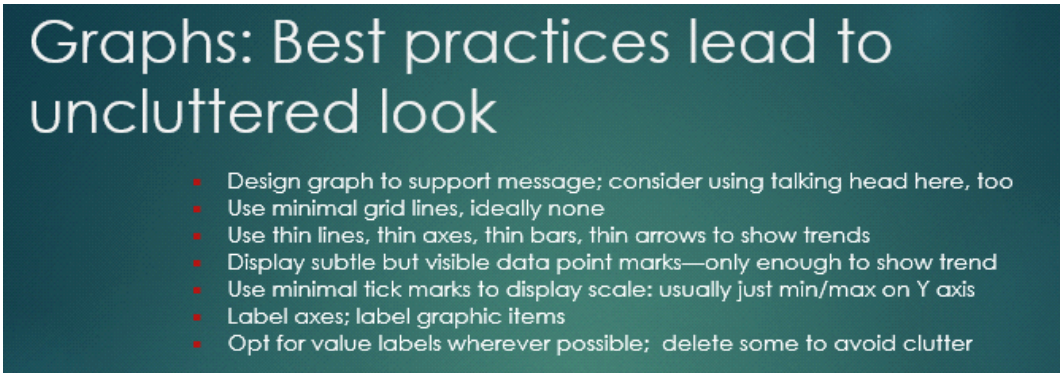
Summary	• **Creating designs - best practices** • **Graphs - best practices** • **Dashboard**
	 • When you visualize you have to be both accurate and precise.
• In the given graphs, which design is easier to understand?	 • L: 30-34 and 34-39 bars, it's difficult to tell whether there's a difference between those two when you use 3D effects because of the shadow. • Also, What purpose does this 3D representation serve? - None. • In some cases, 3D effects can be counter-productive.
• What are the best practices for creating designs?	
• Compare these two charts on malpractices and best practices in creating designs.	 • Consider the slices of Our company, 34 and Competitor A, 26:

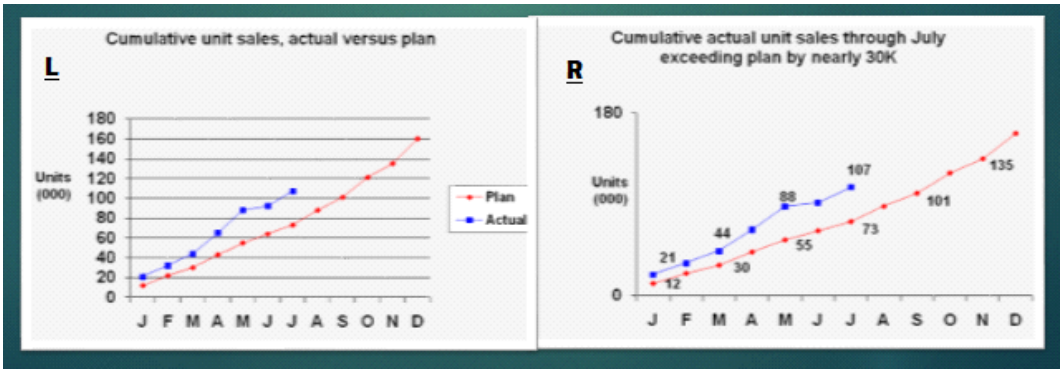
L: Both these slices look same. It essentially says our company and competitor A has same market share.
R: You can see the difference between both of these slices.

- Moreover, if we rotate this pie chart so that Competitor C is at front, it can appear to be same size as our company's.

- 3D effects can be deceitful, more so in pie charts.

- With the help of given graphs, comment on best practices for graphs to get uncluttered look.

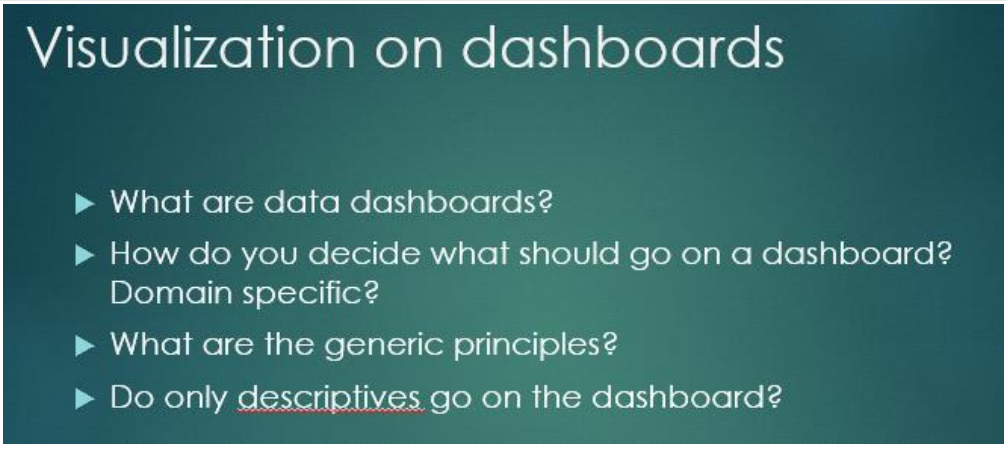


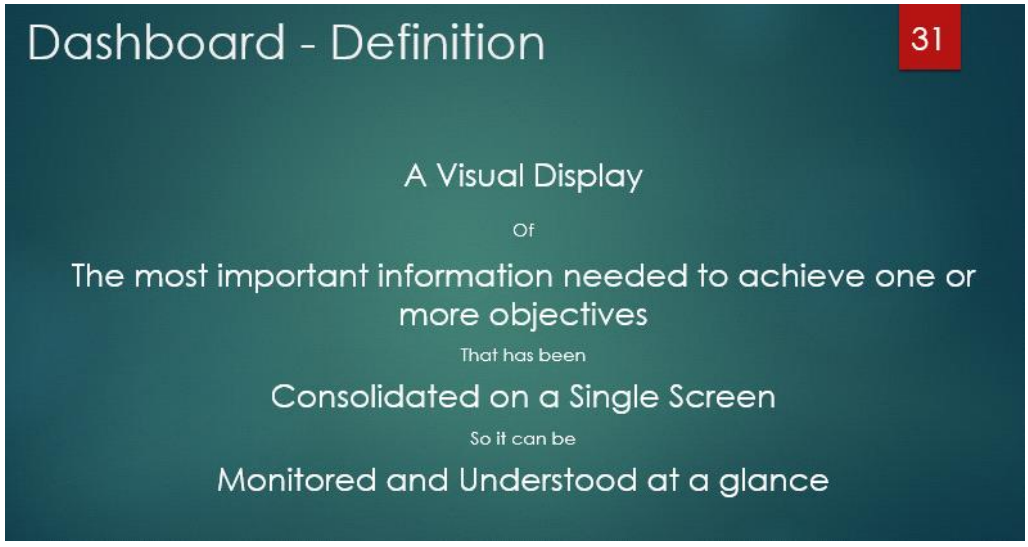


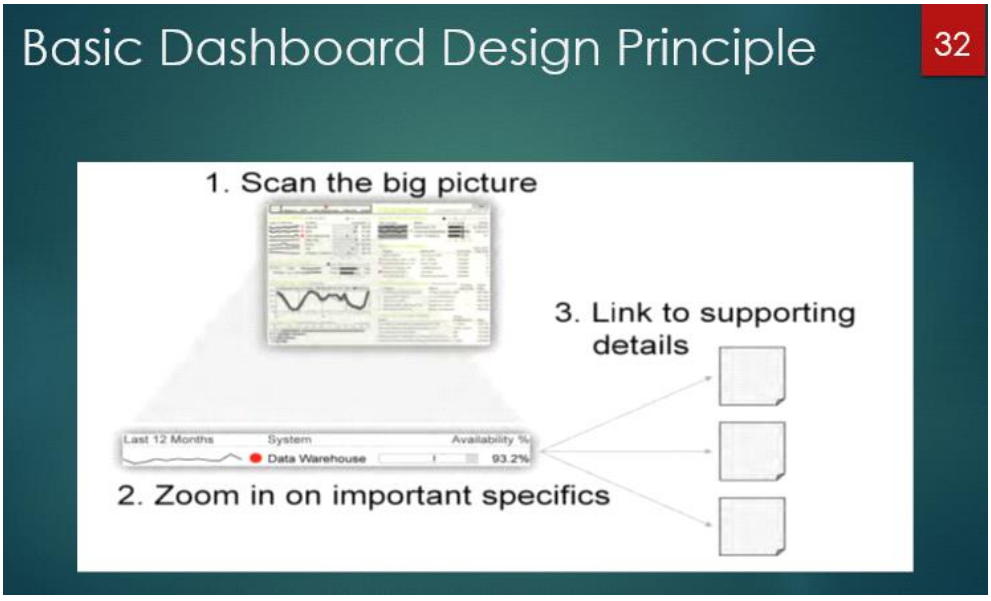
- Compare the titles of both charts. The message is very clear in R.
In L, we have to calculate.

- L: We need to trace the horizontal axis to see what value a dot represents. And, then to compare sales of let's say July, we'll first trace both dots and then calculate the difference.

- Opt for value labels wherever possible and you can delete labels to avoid clutter, as is done in R. Also, since we have annotated in R, no need of grid lines.

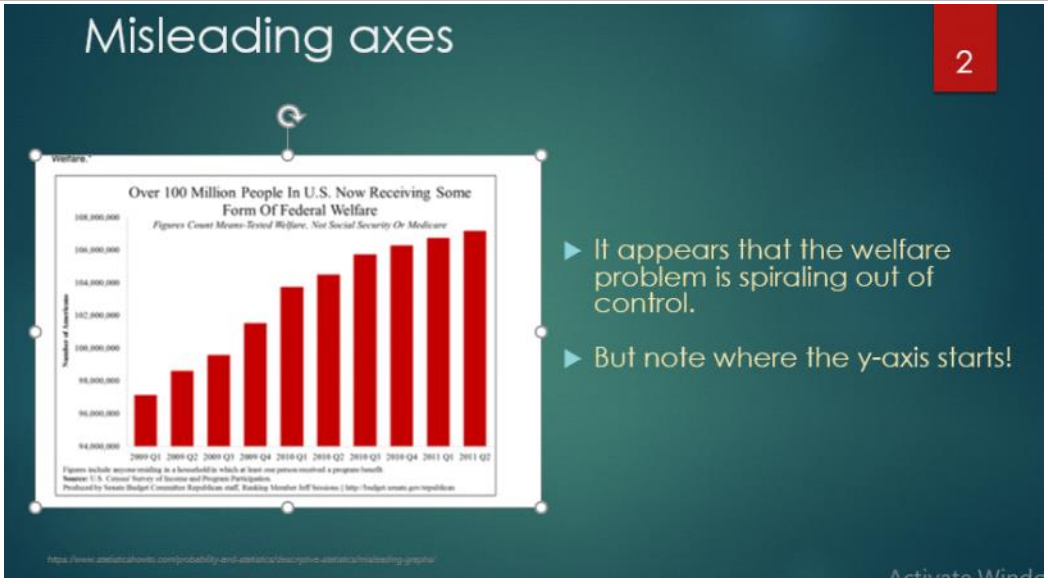


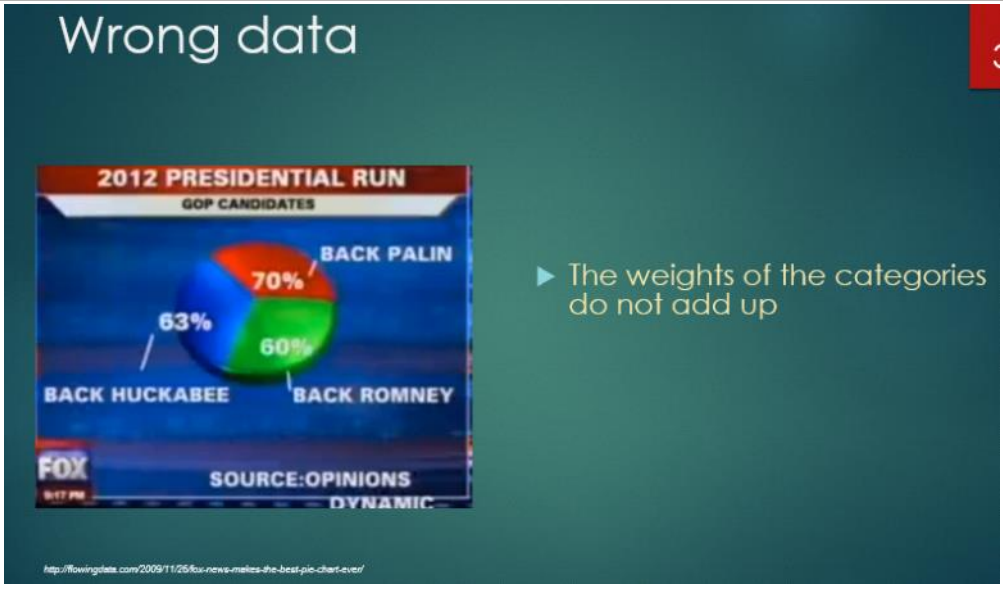
• What is a dashboard?	 • A dashboard is a collection of related displays with some purpose. • **It has to be on a single screen.** You do not want to be scrolling up and down. You look at it and in one glance you should know what's happening. • When designing a dashboard, know what the purpose is for that dashboard. And know what you need to put there to help the user of the dashboard achieve that purpose. If the purpose of the dashboard is to look at one metric only. Then show only that, nothing else.
• Difference between interactive and visual dashboards	• Why a visual display? There's a difference between interactive dashboards and visual dashboards. In interactive dashboards, to see info on something, you need to click on something. A dashboard is supposed to be visual. You don't necessarily have to make it interactive.
	• The word dashboard was borrowed from a car automobile dashboard. Whenever you design a dashboard, you think of a car dashboard. Can you imagine if you're driving a car, and the dashboard requires you to reach in and push a button to look at the fuel consumption, push another button to look at the speed, and the speed display tells you that you're 32% over the average speed limit, is that all relevant? You only need to know how fast you're driving, and that's it. In one glance, from the corner of your eye, you should get the information. And, it's all there in one screen.
• Can a dashboard be real-time?	• Real-time dashboard: A real-time cannot be a dashboard. A dashboard is meant where in one glance you get the information and then you go do what you have to do. A real time is something that you are monitoring. If there's a reason to monitor something in real time, it should not be a dashboard. It should be an entity in itself.
• Tabular displays on dashboard	• The whole idea of dashboards is so that you can put multiple metrics on display at the same time. If you put tabular display where you have to scroll, then by definition, they are not on the same page.


Basic Dashboard Design Principle
32
1. Scan the big picture
3. Link to supporting details
Last 12 Months
System
Availability %
Data Warehouse
93.2%
2. Zoom in on important specifics

Tutorial 1.1: Misleading Visuals

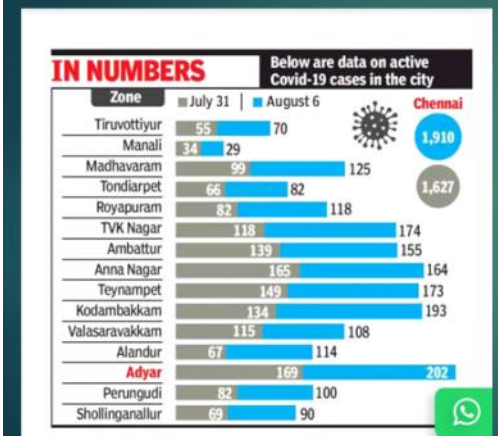
Friday, 07 October 2022 18:49





Inappropriate chart type

4

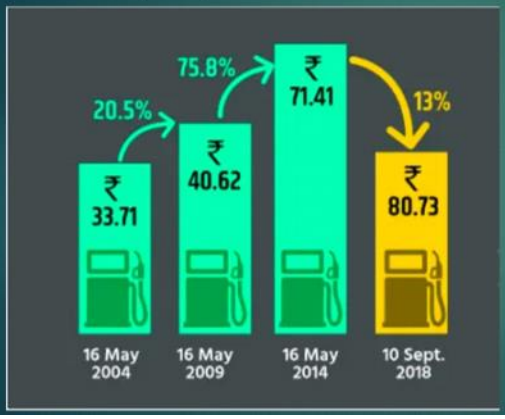


- It is difficult to compare the two datasets using the stacked bar chart.
- Bar chart with two series of columns would be better.

Incorrect bar chart

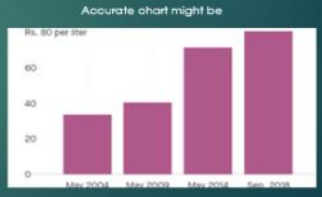
5

%increase in petrol prices



- An arrow points downward to Sept data point, however there is a positive change to the price.
- However, the %increase has reduced for Sept.

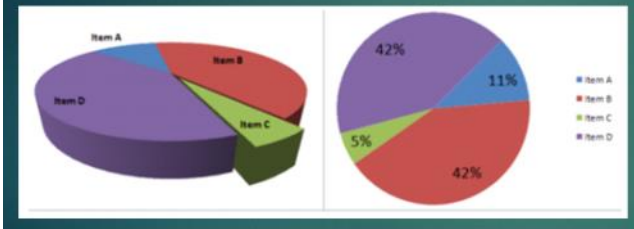
Accurate chart might be



https://qz.com/india/1388030/indias-bjp-peddled-a-prime-example-of-chartjunk-to-its-millions-of-followers/

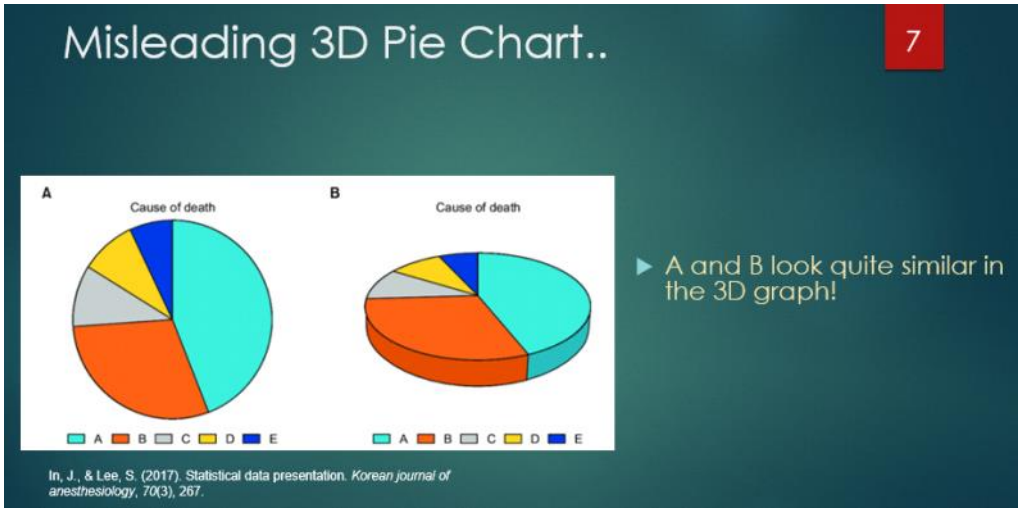
Misleading 3D Pie Chart

6



- The slices closer to you will appear larger than those further away.
- Compare Item A and Item C in both the graph

https://www.noigroup.com/noijam/how-to-spot-a-misleading-graph/


Misleading 3D Pie Chart..
7
A
Cause of death
A B C D E
B
Cause of death
A B C D E
A and B look quite similar in the 3D graph!
In, J., & Lee, S. (2017). Statistical data presentation. Korean journal of anesthesiology, 70(3), 267.

Tutorial 1.2: Building a simple dashboard

Friday, 07 October 2022 22:45

A call centre data is given

Given is time stamp (on hourly basis) and call volumes (number of incoming calls received in that time slot)

(T1)

Time stamp	Call Volume
A	

Objectives

- Show daily call pattern
- Show hourly call pattern
- Get the service level
- Show a sensitivity analysis for service level

1. Daily call pattern

We will show daily pattern month wise, *i.e.*,

$$\text{Average Daily Calls (in a month)} = \frac{\text{Total number of calls in that month}}{\text{Number of days in that month}}$$

Number of calls made in a month

1. Extract months

(T1)

Time stamp	Call Volume	Month
A		=Month(A)

2. Create a separate list of months

(T2)

Month	Total calls made in that month	Average Daily Calls
B	C = [Sum call vol in T1 if month = B]	$\frac{C}{\text{Number of Days in B}}$

3. Plot Daily call pattern with
   Average daily calls vs Month (y vs x)

2. Hourly call pattern

$$\text{Average hourly calls} = \frac{\text{Total number of calls made in a given hour slot}}{\text{Total number of times that hour slot occured in the given data}}$$

Needed variables

1. Extract hour slot in T1

(T1)

Time stamp	Call Volume	Month	Hour slot
A		=Month(A)	=Hour(A)

2. Create a separate list of hour slot (unique values)

(T3)

Hour slot	Total num of calls made in that hour slot	Total number of hour slots	Average hourly calls
D	E = [Sum call vol in T1 if hour slot = D]	F = [Count in T1 if hour slot = D]	$\frac{F}{E}$

3. Plot Average hourly calls vs Hour Slot

3. Service level

$$\text{Service Level} = \frac{\text{Total number of attended calls}}{\text{Total number of incoming calls}} \times 100\ (\%)$$

- **Attended Calls**

$$\text{Total Attended calls} = \text{Total Incoming calls} - \text{Total Unattended calls}$$

- **Unattended Calls**

$$\text{Untattended Calls} = \text{Max}\big((\text{Incoming calls} - \text{Capacity}), 0\big)$$

Reasoning:
- If Incoming < Capacity => Unattended = 0
- If Incoming > Capacity => Unattended = Incoming - Capacity

- **Capacity**

Capacity (in an hour slot)
= Number of agents available in that hour slot × Number of calls one agent can attend in an hour slot

- **Assumptions**

We're assuming that we currently have a total of 11 agents working in 3 shifts of 8 hours each shift

(T4)

	Shift	Number of agents working in this shift

(Shift starts from 8)	8	3
(Shift starts from 10)	10	4
(Shift starts from 12)	12	4

Open hours of the call centre	8 to 20
Work hours per agent	8 hours
Average call duration	5 mins
Number of calls one agent can attend in one hour	$\frac{60}{5} = 12$

- Computing Capacity

(T5)

Hourly Slot	Agents available from shift 8	Agents available from shift 10	Agents available from shift 12	Total number of agents available in this slot	Capacity of this hour slot
H	J = number of agents from T4 If (H>=8 and H<= 8+8) Otherwise 0	K = number of agents from T4 If (H>=10 and H<= 10+8) Otherwise 0	L = number of agents from T4 If (H>=12 and H<= 12+8) Otherwise 0	I = J+K+L	I * 12

- Computing Service Level

1. Extract Call Volume and Hour Slot from (T1)
   (T6)

Call Volume (Incoming)	Hour Slot	Available Capacity	Unattended	Attended
	M	= Vlookup Capacity in T5 where Hour slot = M	Max((Incoming-Capacity), 0)	Incoming - Unattended

2. Compute Service Level

Service level	$\frac{\text{Attended}}{\text{Incoming}} \times 100\ (\%)$

4. Sensitivity Analysis of service level

Here, we'll change the allocation and compute service level for each allocation. That will give us the picture of how sensitive the service level is as we increase number of hired agents (and to some extent in what shift depending on what hourly slot has the maximum load).

- Computation

1. Create a list on allocations
   (T7)

Shift	Allocation 1 (from T4)	Allocation 2 (example)	Allocation 3 (example)
8	3	4	5
10	4	4	4
12	4	4	4

2. Now for each computation, we'll compute service level (as is demonstrated above)
   (T8)

Allocation Number	Service Level
1	
2	
3	

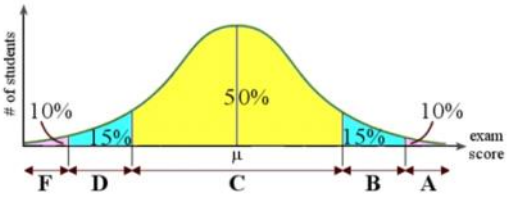
3. Plot a bar graph: Service Level vs Allocation

Week 2

Saturday, 08 October 2022 14:47

2.1 Probability Distributions

Saturday, 08 October 2022 18:39

Summary	• 3 ways to fit distributions to data 1. Trace-driven simulation 2. Theoretical distribution 3. Empirical distribution
	Probability distributions What are distributions? How to identify correct distribution of the data?
	Probability distributions • A statistical model that shows possible outcomes of a particular event or course of action as well as the statistical likelihood of each event. • What do we mean by "Grades in a course follows a normal distribution"? • What do we mean by "Sales for the next month may be uniformly distributed"?  • How do you fit distributions to data?
• In how many ways can we create models once we have the data? ○ Name them. ○ Explain.	**How to go about this?** How do we use the collected business data (sales volume, loan defaulters, Salary hikes in an organization, etc.)? 1. The data values themselves are used directly in the simulation. This is called trace-driven simulation. 2. "Fit" a theoretical distribution to the data (and check whether that "fit" is good!). 3. The data values could be used to define an empirical distribution function in some way. • In point 1, we don't need to fit any distribution. We directly use the data in our analysis. • Theoretical distributions are the distributions we've already studies, eg, normal, uniform distribution, etc. • In point 2, we first fit a distribution in the data. Check how good is the fit. • Point 3: <u>If the data doesn't fit any theoretical distribution, then instead of trying to fit already available distributions, we create our own distributions.</u> Those distributions are called **empirical distributions.** And then we use this distribution in our future analysis.

- What are the essential building blocks of empirical distribution?

What are these empirical distributions?

- Using the data, we build our own distributions.
- How does one build a distribution?

- Essential building blocks:

Define the density/distribution functions.

Estimate the parameters (mean, standard deviation, etc.)

❋ When the professor says distribution function, he means CDF.

Empirical distributions

For ungrouped data:

Let $X_{(i)}$ denote the ith smallest of the X_j's so that: $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$.

$$F(x) = \begin{cases} 0 & \text{if} \quad x < X_{(1)} \\ \dfrac{i-1}{n-1} + \dfrac{x - X_{(i)}}{(n-1)\left(X_{(i+1)} - X_{(i)}\right)} & \text{if} \quad X_{(i)} \le x < X_{(i+1)} \quad \text{for} \quad i = 1,2,\cdots,n-1 \\ 1 & \text{if} \quad X_{(n)} \le x \end{cases}$$

- Given a CDF, we already know how to find the density function.

Empirical distributions

For grouped data:

- Suppose that n X_j's are grouped in k adjacent intervals $[a_0,a_1)$, $[a_1,a_2)$,...$[a_{k-1},a_k)$ so that jth interval contains n_j observations. $n_1 + n_2 + \ldots n_k = n$.
- Let a piecewise linear function G be such that $G(a_0) = 0$, $G(a_j) = (n_1 + n_2 + \ldots n_j)/n$, then:

$$G(x) = \begin{cases} 0 & \text{if } x < a_0 \\ G(a_{j-1}) + \dfrac{x - a_{j-1}}{a_j - a_{j-1}}\left[G(a_j) - G(a_{j-1})\right] & \text{if } a_{j-1} \le x < a_j, j = 1,2,\cdots k \\ 1 & \text{if } a_k \le x. \end{cases}$$

- We have k intervals, and in each interval we have $n_1, n_2, \ldots, n_k$ values.
- $G(a_j)$ is proportional to the observations up to that point/interval.

- When do we use trace driven simulation?
  - What are its drawbacks?

The three approaches...

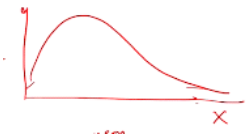
- Approach 1 is used to validate simulation model when comparing model output for an existing system with the corresponding output for the system itself.
- Two drawbacks of approach 1: simulation can only reproduce only what

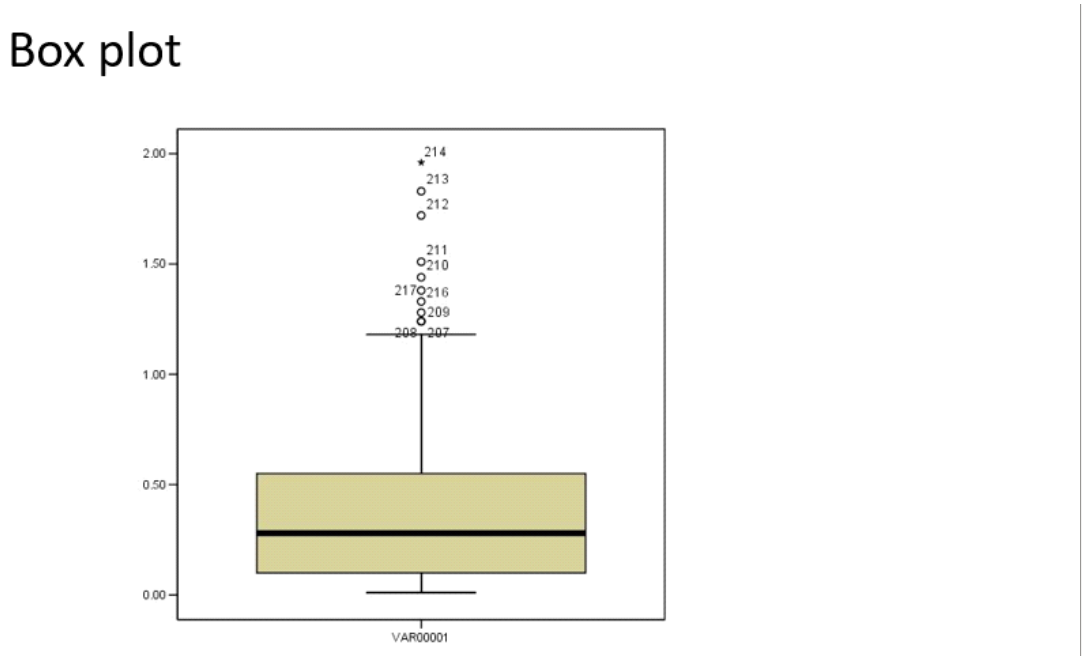
	happened historically; and there is seldom enough data to make all simulation runs. • Approaches 2 and 3 avoid these shortcomings so that any value between minimum and maximum can be generated. So approaches 2 and 3 are preferred over approach 1. • If theoretical distributions can be found that fits the observed data (approach 2), then it is preferred over approach 3.
☐ Approach 1 is not very clear to me	• Approach 1: You have the output, and you want to validate if the output is correct or not. You push the already available data into your model and your model generates an output and you compare that output with the reality (the existing system, what happens in future) and check whether it matches. • So trace-driven simulation is used to validate a model that you already may have built using some approach. • Problem with approach 1 is you're going to test the model only with the data you already have. This may not be enough. • What if the data was collected in a certain circumstance, and as the circumstances change will not give you a fair values.
• What are the drawbacks of empirical approach?	Approach 3 v/s Approach 2 • Empirical distribution may have some irregularities if small number of data points are available. Approach 2 smoothens out the data and may provide information on the overall underlying distribution. • In approach 3, it is usually not possible to generate values outside the range of observed data in the simulation. • If one wants to test the performance of the simulated system under extreme conditions, that can not be done using approach 3. • There may be compelling (physical) reasons in some situations for using a particular theoretical distribution. In that case too, it is better to get empirical support for that distribution from the observed data.

2.2 Business Example

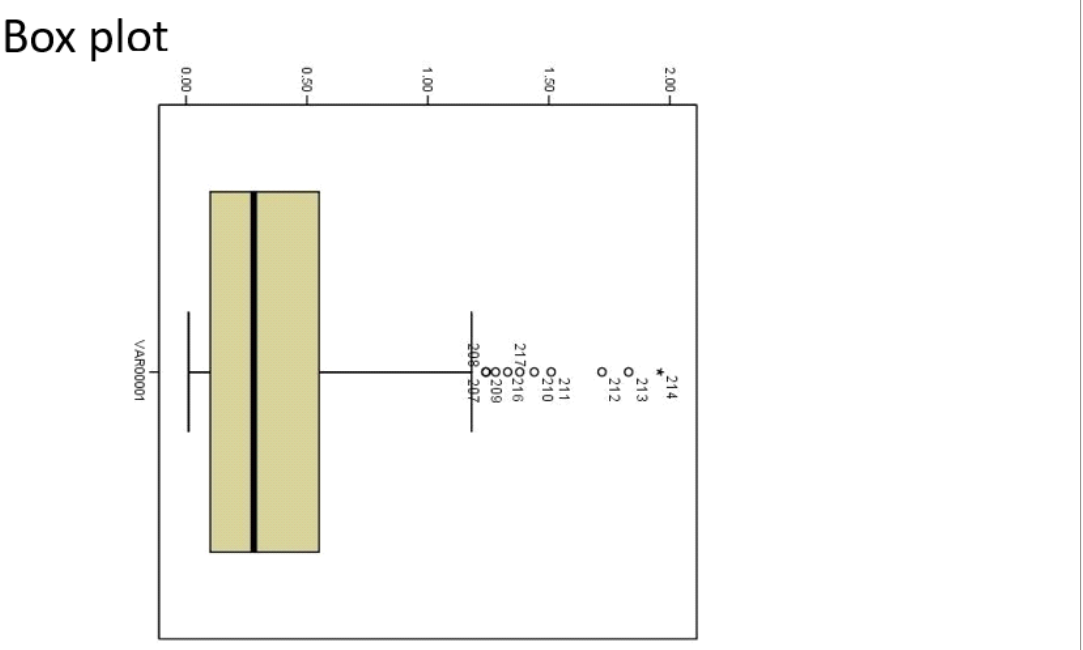
Sunday, 02 October 2022 8:22

Summary	• Interpreting Descriptive Statistics • Goodness-of-fit
	Business example • Data points: 217. • For these data points, we need to fit a probability distribution.

• What is the relation between mean, mode and median for symmetric distributions?
 	
• Try to interpret the given statistics.
 	
• How do you interpret skewness?
 	
  • If skewness is positive, the data is skewed to the _____?
  ○ How will its shape look like?
	Summary statistics
	
	VAR00001
	N — Valid: 217; Missing: 1
	Mean: .4012
	Median: .2800
	Mode: .05[a]
	Std. Deviation: .38093
	Variance: .145
	Skewness: 1.466
	Std. Error of Skewness: .165
	Range: 1.95
	Minimum: .01
	Maximum: 1.96
	Percentiles: 25 .1000; 50 .2800; 75 .5500
	a. Multiple modes exist. The smallest value is shown
	
	• For symmetric distributions: Mean = Mode = Median (or very close to each other) (eg, normal distribution)
	
	• But here, mean, mode and median are different. It means it's not a symmetric distribution. So it rules all the symmetric theoretical distributions for us.
	
	• Looking at the min-max of the data points, it can be observed that none of the data point takes on negative value.
	Rules out all the distributions that go to the negative side of the line.
	
	• Skewness tells us about the symmetry of the distribution.
	The value of skewness is positive, so it's a positive skew, i.e., the data is skewed to the right.
	Right Tail > Left Tail
	
	•
	 
	
	• So consider all the positive skewed distributions as the potential distributions here.

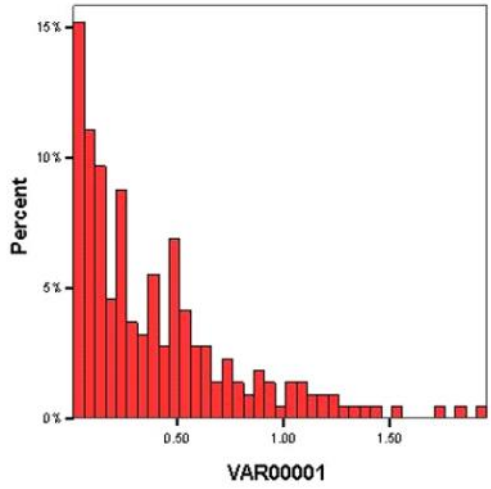


- Below is given the same plot tilted 90 degrees, to get a better picture of the points on x-axis.



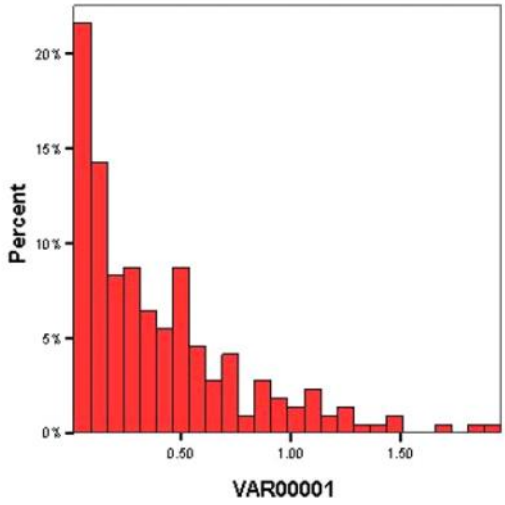
- It's clear from the box plot that this distribution has a positive skew.

Histograms



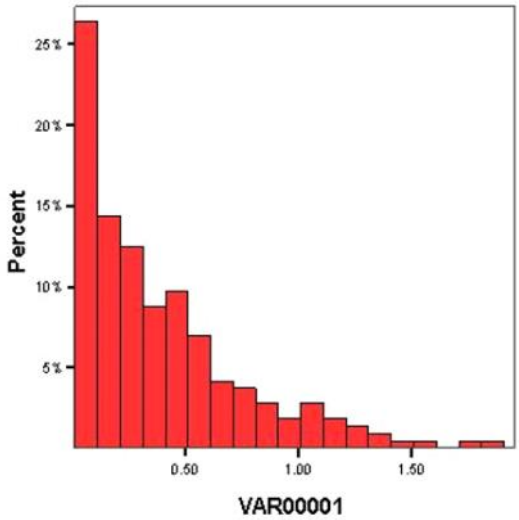
- Large number of values close to 0.
- Very few values more than 1.5.

Histograms



- We increased the width of the bars here.
- Frequency seems to be dropping as you go to the right.

Histograms



- Even thicker bars here.

- What is Coefficient of variation, cv?
  - This statistic is for continuous or discrete distributions?
  - cv = 1 for?
  - cv > 1 for?

Clues from summary statistics

- For the symmetric distributions mean and median should match. In the sample data, if these values are sufficiently close to each other, we can think of a symmetric distribution (e.g. normal).
- Coefficient of variation (*cv*): (ratio of std dev and the mean) for continuous distributions. The *cv = 1* for exponential dist. If the histogram looks like a slightly right-skewed curve with *cv >1*, then lognormal could be better approximation of the distribution.

Note: For many distributions *cv* may not even be properly defined. When? Examples?

- Coefficient of variation, $(\text{cv}) = \frac{\text{Standard Dev}}{\text{Mean}} = \frac{\sigma}{\mu}$
- $cv = 1$ :: Exponential Distribution
  for slightly right-skewed curve with $cv > 1$ :: Lognormal Distribution
- In some cases, *cv* may not be defined.
  Take an example of standard normal distribution where $\mu = 0$.
- Here, $\text{cv} = \frac{0.38093}{0.4012} = 0.9495 \approx 1$

- Lexis ratio
- Skewness (v)
  - $v = 0$ for?
  - $v > 0$?
    - $v = 2$ for?
  - $v < 0$?

Clues from summary statistics

- Lexis ratio: same as *cv* for discrete distributions.
- Skewness (*v*): measure of symmetry of a distribution. For normal dist. *v* = 0. For *v>0,* the distribution is skewed towards right (exponential dist, *v = 2)*. And for *v<0,* the distribution is skewed towards left.

- If Skewness,
  - v = 0 :: Normal distribution
  - v > 0 :: Distribution is skewed towards right
    - v = 2 :: Exponential Distribution
  - v < 0 :: Distribution is skewed towards left
- We'll try to fit exponential distribution here.

Parameter estimation

- Once distribution is guessed, the next step is estimating the parameters of the distribution.

- Each distribution has a set of parameters.
- ✓ Normal distribution has mean and standard deviation
- ✓ Exponential distribution has a "λ".

- Most common method of parameter estimation: MLE (What is this?)

- How can we check the goodness-of-fit of the fitted distribution?
  - Name the methods
  - Explain

Goodness-of-fit

- For the input data we have, we have assumed a probability distribution.
- We also have estimated the parameters for the same.
- How do we know this fitted distribution is "good enough?"

- It can be checked by several methods:
  1. Frequency comparison (a bit technical)
  2. Probability plots (visual tool)
  3. Goodness-of-fit tests (statistical test of goodness. Very widely used).

- We're trying to fit exponential distribution here.

1. In frequency comparison, we can compare the frequency that comes from exponential distribution with the frequency you have observed from the dataset. If the frequencies match, we say that exponential distribution is good fit.
2. Probability plots are visual tools that tell you if the observed frequency/quartiles/percentiles matches with the frequency/quartiles/percentiles of the distribution. If it fits, you'll get a line , if it doesn't fit, you'll be far away from that line.
3. Goodness of fit tests: Many of them uses Chi-Square tests.

2.3 Guessing the Distribution

Sunday, 09 October 2022 12:01

Summary	• Q-Q plot • P-P plot • Chi-Square test
• What are the two probability plots we use to check goodness-of-fit? • How do you interpret Q-Q plot? • What points to we plot in Q-Q plot?	**Probability plots** Q-Q plot: Quantile-quantile plot • Graph of the q_i-quantile of a fitted (model) distribution versus the q_i-quantile of the sample distribution. $x_{q_i}^{M} = \hat{F}^{-1}(q_i)$ $x_{q_i}^{S} = \tilde{F}_n^{-1}(q_i) = X_{(i)}, i = 1,2,..n.$ • If $\hat{F}(x)$ is the correct distribution that is fitted, for a large sample size, then $\hat{F}(x)$ and $F_n(x)$ will be close together and the Q-Q plot will be approximately linear with intercept 0 and slope 1. • For small sample, even if $\hat{F}(x)$ is the correct distribution, there will some departure from the straight line. • $\hat{F}$ :: indicates the distribution that we're trying to fit $\tilde{F}_n$ :: indicates the distribution that come from the sample • If x^M_{qi} that comes from the fitted/model distribution matches with the x^S_{qi} that comes from the sample distribution, then you're going to get a line. •  • You'll plot all the x that comes from X^M, that is model distribution on x-axis, and plot that against the x that comes from sample distribution, X^S. • If x_i from model distribution matches with the sample distribution, you'll get a nice 45° line in Q-Q plot. • This line will have an intercept of 0 and a slope of 1. • Practically we'll get a line that is around this 45° line. And how far away we are from this 45° line determines how good is the model distribution. • If very far from this line, then the model distribution doesn't match the sample distribution.
• How do you interpret P-P plot? • What points do we plot in P-P plot?	**Probability plots** • P-P plot: Probability-Probability plot. A graph of the model probability $\hat{F}(X_{(i)})$ against the sample probability $\tilde{F}_n(X_{(i)}) = q_i, i = 1,2,..n.$ • It is valid for both continuous as well as discrete data sets. • If $\hat{F}(x)$ is the correct distribution that is fitted, for a large sample size, then $\hat{F}(x)$ and $F_n(x)$ will be close together and the P-P plot will be approximately linear with intercept 0 and slope 1. • Here, we compare probability distributions: $\hat{F}$ and $\tilde{F}$ (CDF).

- Q-Q plot amplifies the differences between the _____.
- P-P plot amplifies the differences between the _____.

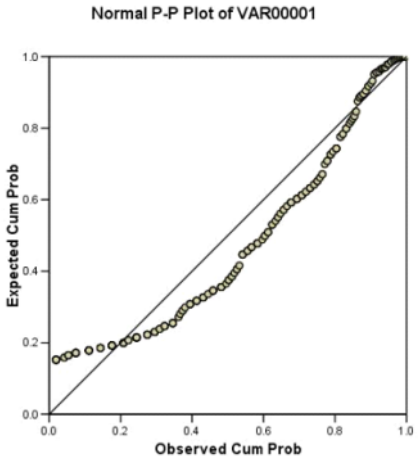
Probability plots

- The *Q-Q* plot will amplify the differences between the tails of the model distribution and the sample distribution.
- Whereas, the *P-P* plot will amplify the differences at the middle portion of the model and sample distribution.

- In general, for both probability plots, what do we plot on the:

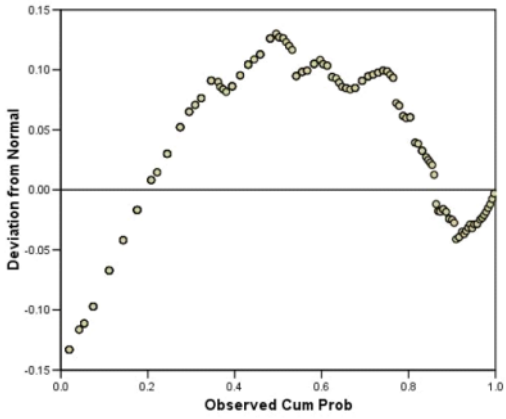
  (1) x-axis?
  (2) y-axis?

Probability plots: Dataset



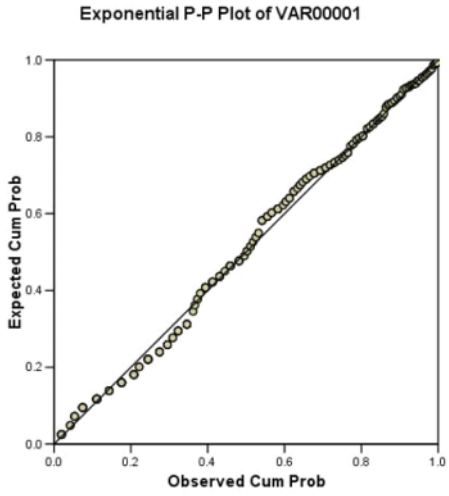
- On the observed data Var1, we've tried to fit Normal distribution using P-P Plot.
- X-axis :: Observed cumulative frequency
  Y-axis :: Expected(model) cumulative frequency
- So, both x- and y-axis will go from 0 to 1.
- Here, observed points don't seem to be close to expected points.

Probability plots: Dataset



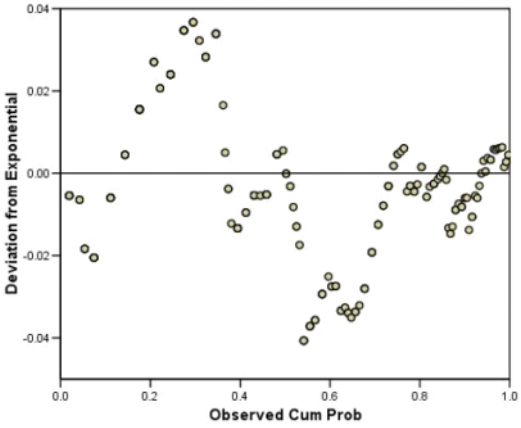
- Here is shown the deviation of observed points from Normal.
- The deviation seems to be high in P-P plot, of the order of 0.15.

Probability plots: Dataset



- Here, we're trying to fit exponential distribution in the dataset using P-P plot.
- The observed points seem to be very close to the 45^{o} line.

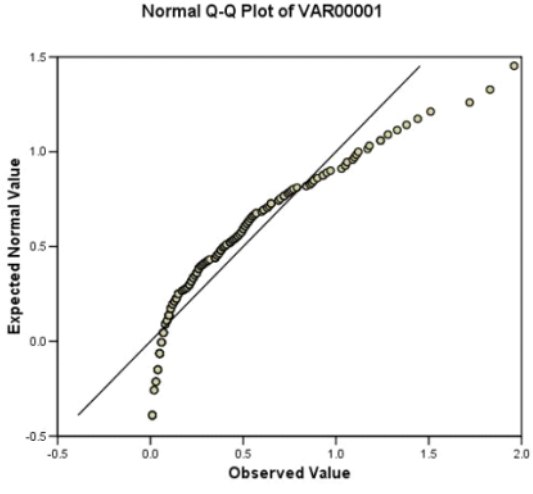
Probability plots: Dataset



- Here, the deviation of the dataset points from the exponential distribution is shown.
- Notice the scale of Y-axis, it's of the order of 0.04, which is small.

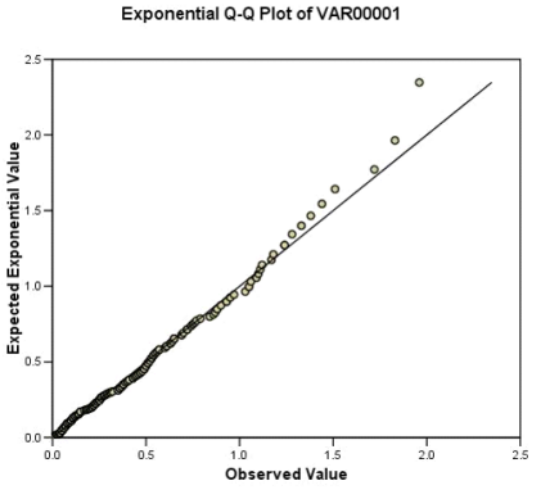
- Conclusion: In P-P plots, the exponential distribution seem to be a better fit than the normal distribution.

Probability plots: Dataset



- In this Q-Q plot, we're trying to fit normal distribution to the dataset Var1.
- X-axis :: X-values from the sample
  Y-axis :: Y-values from the model
- The observed points seem to be deviating from the 45° line.

Probability plots: Dataset



- Q-Q plot with the exponential distribution.
- Seems very close to the 45° line.
  There are some deviations in the upper portion, these deviations are for higher observed values.

- Conclusion: In Q-Q plots as well, the exponential distribution seem to be a better fit than the normal distribution.

Q. Do we now conclude that the exponential distribution is a good fit for the data?
A. No. We also need to look at statistical goodness-of-fit tests.

- Name the two famous statistical goodness-of-fit tests.

Goodness-of-fit tests

- A goodness-of-fit test is a statistical hypothesis test that is used to assess formally whether the observations $X_1, X_2, X_3 \ldots X_n$ are an independent sample from a particular distribution with function $F^{\wedge}$.

H_0: The X_i's are IID random variables with distribution function $F^{\wedge}$.

- Two famous tests:

1. Chi-square test
2. Kolmogorov - Smirnov test

- How do you calculate chi-square test?
  - Look here.

Chi-square test

- Applicable for both, continuous as well as discrete, distributions.
- Method of calculating chi-square test statistic:

1. Divide the entire range of fitted distribution into k adjacent intervals -- $[a_0,a_1), [a_1,a_2), \ldots [a_{k-1},a_k)$, where it could that $a_0 = -\infty$ in which case the first interval is $(-\infty,a_1)$ and/or $a_k = \infty$.

N_j = # of X_i's in the jth interval $[a_{j-1},a_j)$, $j = 1,2 \ldots n$.

2. Next, we compute the expected proportion of X_i's that would fall in the jth interval if we were sampling from fitted distribution

Chi-square test

$$\text{For continuous distributions: } p_j = \int_{a_{j-1}}^{a_j} \hat{f}(x)dx$$

$$\text{For discrete distributions: } p_j = \sum_{a_{j-1} \le x_j < a_j} \hat{p}(x_j).$$

- Finally the test statistic is calculated as:

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}.$$

- Better formula is given here.

- Based on chi-square test statistic when do we:
  - Accept H_o?
  - Reject H_o?

Chi-square test

- This calculated value of the test statistic is compared with the tabulated value of chi-square distribution with $k-1$ df at $1-\alpha$ level of significance.

$$If \quad \chi^2 > \chi^2_{k-1,1-\alpha} \quad \text{Reject } H_0$$

$$If \quad \chi^2 \le \chi^2_{k-1,1-\alpha} \quad \text{Do not Reject } H_0$$

- The data given to us was a time data. Time to get a service in a bank.
- The exponential distribution has a strong association with the queuing theory.

Identifying and fitting distributions

Sunday, 09 October 2022 22:49

• What's the approach we take for identifying and fitting distributions?	Task: Identifying Distributions and fitting distributions Approach: 1. Look at visualizations 2. Use descriptive statistics to further identify distributions 3. Go for tests to confirm judgement
	<u>Detailed Approach</u>
What are the steps in the analysis?	1. Plot Histogram • Whenever you want to find a distribution, the first thing you do is plot Histogram. • When you plot a histogram, you can look at it and tell what family of distributions a dataset may belong to. 2. Do some descriptive Analysis 3. Plot Probability Plots to check goodness of fit i. P-P Plot ii. Q-Q Plot 4. Apply statistical tests to check goodness of fit i. Chi square test ○ For this test, we usually need two kind of frequencies: ▪ Observed frequency: sample frequency ▪ Expected frequency: population frequency, the model distribution where we assume the data is coming from. ○ Output: the chi-square statistic and the $p - value$ ○ Compare the $p - value$ and α (significance level) and, conclude about the null hypothesis. ii. Calculated and tabular chi-square statistic comparison ○ Input: confidence level, degrees of freedom (df) ○ Output: Tabular chi-square statistic ○ Compare tabular and calculated chi-square values and, conclude about the null hypothesis.
Defining Hypotheses	Null Hypothesis: Sample distribution follows model distribution. Alternate Hypothesis: Sample distribution does not follow model distribution.
Chi Square Test • How to calculate expected frequency? `stats.chisquare(obs_freq, expec_freq)` • What does this command return? • Chi-square formula? • What is p-value? • What is α? • Comparing both these values, when do we accept the null hypothesis?	`stats.chisquare(obs_freq, expec_freq)` This command returns two values: 1. Chi-Square Statistic 2. P-value • P-value is the probability of observing this particular sample when the null hypothesis is assumed to be true. • α :: Significance level, that can take a value of 0.05, or 0.10, or 0.15. • If P-value > α :: We accept the null hypothesis.
• Tabulated Chi-Square Statistic `scipy.stats.chi2.ppf(0.95, df=9)` • What arguments does this command take? • How to calculate df?	`scipy.stats.chi2.ppf(0.95, df=9)` Above command gives us the tabulated chi-square statistic, that takes two arguments: 1. First argument is confidence level = 95% we have taken 2. Second argument is degrees of freedom, $df = k - p - 1$ where,

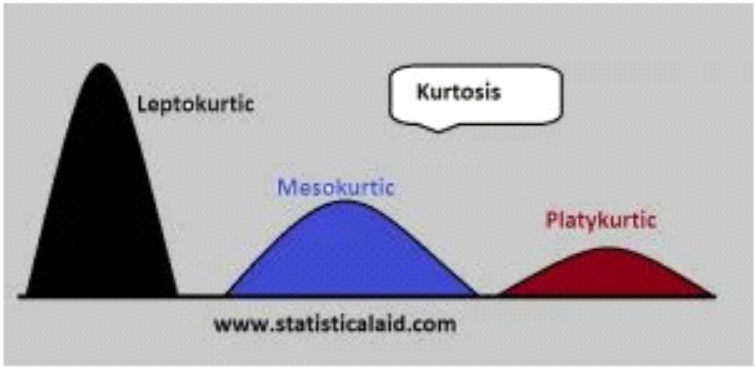
• How do we calculate df in a contingency table?	a. k = number of classes/intervals/buckets b. p = number of parameters we estimate from the sample
• Comparing calculated and tabulated chi-square statistic, when do we accept the null hypothesis?	If Tabulated value ≥ Calculated value » We accept the null hypothesis
Some other observations:	• In Python, whenever we run a code for any of the plots (Q-Q or P-P), by default, it's always comparing it with the **standard normal distribution.** • It converts the data that is fed to it to a standard normal, basically by doing $\frac{x-\mu}{\sigma}$. • It will scale it and then it will compare it with the standard normal(default), and see how the quantiles (or distributions) are fitting. • So, it normalizes the data, and then compare it with the standard normal.
	• Poisson distribution has mean = variance (= λ)
	• Generally for normal distributions the Kurtosis would be around 3.
	• Degree of freedom = k - p - 1 • p = 0 for uniform distribution as there is no parameter. • This point is doubtful. Needs verification
	Question: What range decides if the skewness is very small or very large to consider?

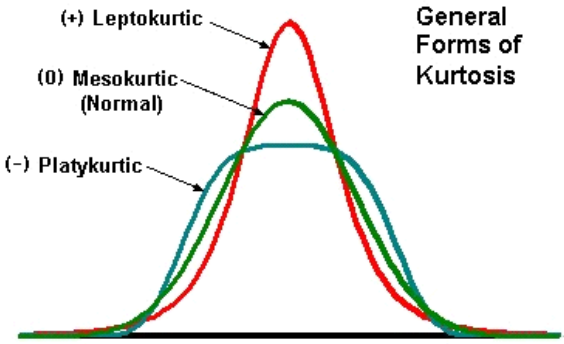
Kurtosis

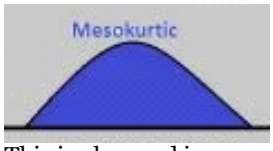
Sunday, 02 October 2022 8:22

Summary	• Kurtosis

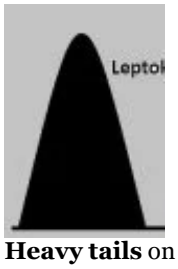
Kurtosis is a statistical measure that defines how heavily the tails of a distribution differ from the tails of a normal distribution. In other words, kurtosis identifies whether the tails of a given distribution contain extreme values.





- The **expected value of kurtosis is 3** (Mesokurtic).
  - 

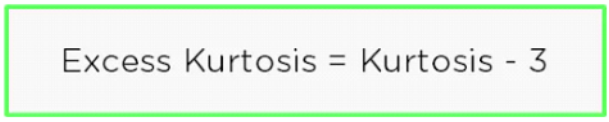
  - This is observed in a symmetric distribution.

- If ***kurtosis* > 3: Positive Kurtosis** (Leptokurtic).
  - 

  - **Heavy tails** on either side, indicating large outliers.
  - In this case, the value of kurtosis will range from 1 to infinity.

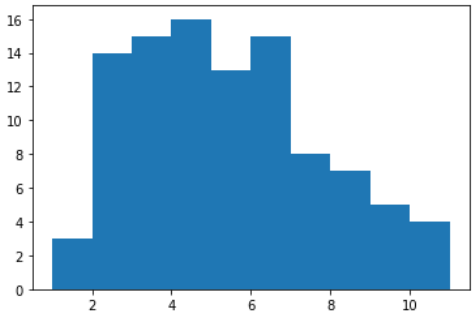
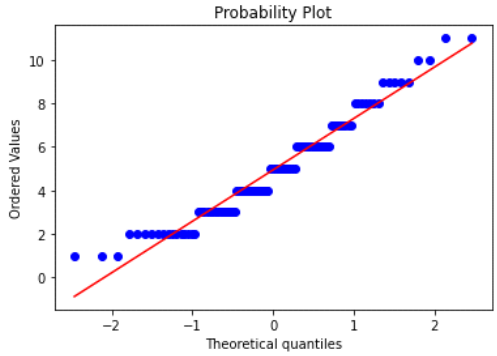
- If ***kurtosis* < 3: Negative kurtosis** (Platykurtic).
  - 

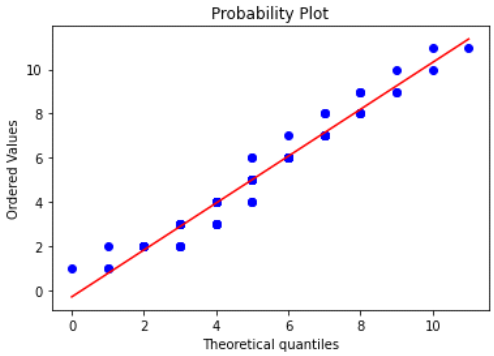
  - **Flat/light tails**, indicating small outliers.

	○ The range of values for a negative kurtosis is from -2 to infinity. The greater the value of kurtosis, the higher the peak. 

2.4 Guessing the Distribution of Dataset 1

Sunday, 09 October 2022 14:09

	Poisson Distribution
Histogram	 • Since there are two peaks, it could be a non-Gaussian distribution. • But, we cannot say whether it's a symmetric or a non-symmetric distribution.
Descriptive Statistics	count 100.000000 mean 4.940000 std 2.381834 min 1.000000 25% 3.000000 50% 5.000000 75% 6.000000 max 11.000000 {('Variance Observed', 5.67), ('Mean Observed', 4.94)}, ('Skew Observed', 0.51), ('Kurt Observed', -0.38) • Variance and mean are close to each other. Think about what distributions have mean and variance close to each other. Ans: Poisson Distribution • Observing the skewness and the kurtosis, we can conclude that we can eliminate certain symmetric distributions such as normal distribution, uniform distribution. • Mean = Median = 5. So, 50% points are lying below 5 and 50% lying above 5. So it's still a symmetric distribution. But, Poisson is not a symmetric distribution.
P-P Plot	 • Points are not lying on the red line. It means the data is not coming from the normal. • Since P-P plot deals with the CDF, and we have data similar to a step function, which also hints that this **distribution could be discrete in nature.**
	P-P w.r.t Poisson distribution with lambda = 4.94



- Now the points are almost around the red line. So, with a certain degree of certainty, we can say that the cumulative distribution from which this data is coming is a Poisson distribution which has a lambda of 4.94.

Statistical tests

obs	frequency	OBS_PROBA	POISSON_PMF	POISSON_FREQ
1	3	0.03	0.035344	3.53
2	14	0.14	0.087299	8.73
3	15	0.15	0.143752	14.38
4	16	0.16	0.177534	17.75
5	13	0.13	0.175404	17.54
6	15	0.15	0.144416	14.44
7	8	0.08	0.101916	10.19
8	7	0.07	0.062933	6.29
9	5	0.05	0.034543	3.45
10	2	0.02	0.017064	1.71
11	2	0.02	0.007663	0.77

- What we first did here, we made a frequency chart (shown in frequency column).
- OBS_PROB = $P(x_i = k) = \frac{frquency\ of\ k}{total\ frquency}$.
  - *e.g.*,
    Total frequency = 100
    $P(x_i = 1) = \frac{3}{100} = 0.03$.
- Then we calculate Poisson pmf
  - $P(x_i = k) = e^{-\lambda}\frac{\lambda^k}{k!}$.
  - Here, $\lambda = 4.94$.
- After this step, we calculate Poisson frequency,
  - $Possion\ freq = Poisson\ pmf \times total\ freq$
    $\Rightarrow \quad = Poisson\ pmf \times 100.$

Observed and Expected frequency

- Observed probability tells you the probability of an observation occurring from the sample.
- The Poisson pmf tells you, the probability of getting the same observation from the population (assuming that the population is Poisson) with the mean of 4.94.
- So now we have both, the observed frequency and the expected frequency (Poisson freq).

Hypothesis

NULL HYPOTHESIS: The given data follows Poisson distribution.

ALTERNATE HYPOTHESIS: The given data does not follow Poisson distribution

Chi-square Test

Calculated chi square statistic = 7.92
p-value = 0.64

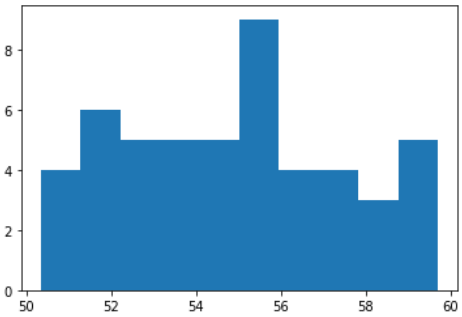
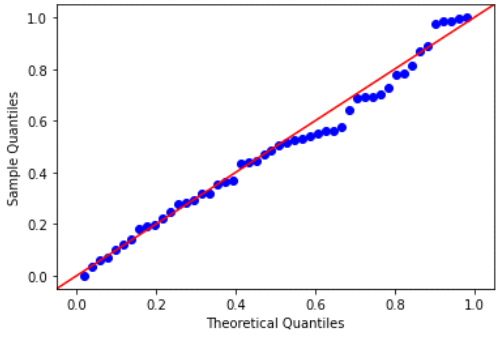
- p-value > α
  This is said to be almost coming from the distribution. So we can accept the null.

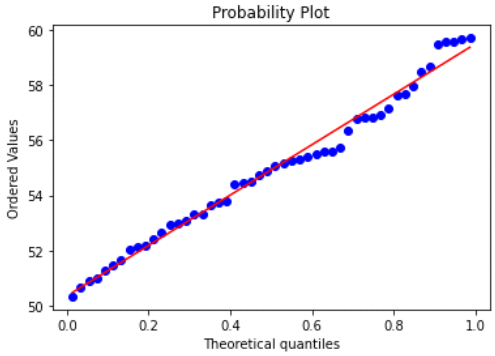
Degrees of freedom	$df = k - p - 1 = 11 - 1 - 1 = 9$. $k = 11 \implies$ total number of classes.
Tabulated Chi-Square	$= 16.92$ • Tabulated value > Calculated value » We accept the null hypothesis
Business Cases	• suppose the given data is from a traffic signal in a city, where the number represents the number of times the signal was violated on a given day, i.e., on day #1 the signal was violated 5 times, on day # 2 the signal was violated 4 times, and so on. This is a count information. • Another example would be that let's say everyday you're manufacturing a thousand products. Out of those thousand products, how many are defective. So, 5 could be the number of defective items on day 1, 4 could be on day #2, and so on. • **Wherever you have count information, those places could be ideal for Poisson distributions to happen.** So, a number of discrete events happening in continuous space is Poisson distribution. • Other examples could include customer arrivals and so on.

•

2.5 Guessing the Distribution of Dataset 2

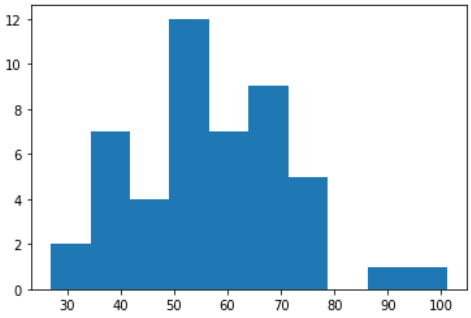
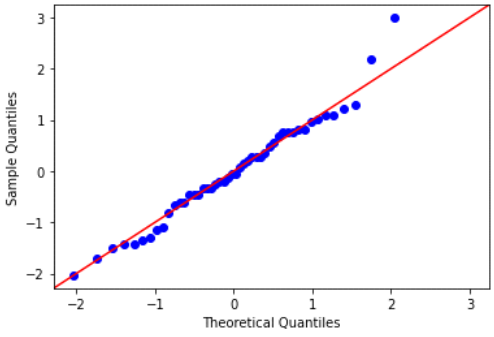
Sunday, 02 October 2022 8:22

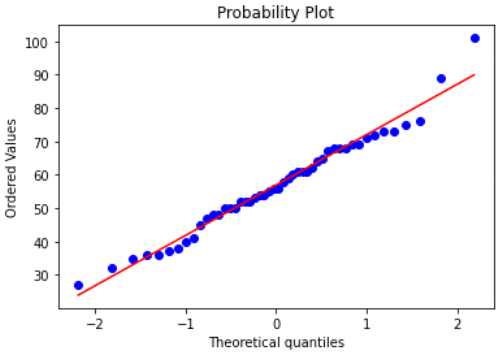
	Uniform Distribution
	 • Peaks on either side, but it's not symmetric in nature, in the sense there is no stepwise decrease on either side. • So, it's a non-Gaussian, somewhat symmetric.
	count 50.000000 mean 54.925080 std 2.659768 min 50.340227 25% 52.965425 50% 54.981739 75% 56.831619 max 59.695597 {('Variance Observed', 7.07), ('Mean Observed', 54.93), ('Skew Observed', 0.18), ('Kurt Observed', -0.87)} • Between the minimum and 25%, 25% to 50%, 50% to 75%, and 75% to maximum, the intervals are around 2. • $\frac{\min + max}{2} = \frac{50+60}{2} = \frac{110}{2} = 55 = mean$. We can assume that this comes from uniform which is between 50 and 60. • Variance or the standard deviation is very small and it is symmetric. • From these numbers, we can conclude that it's a uniform distribution. Because between the quantiles, the split is almost the same, and the standard deviation is very small. • Kurtosis is negative. • Generally for normal distributions the Kurtosis would be around 3. So it cannot be a normal. • Skewness is very low. So, it's not skewed to either side. It is centred.
	Q-Q plot w.r.t Uniform 

	P-P plot w.r.t Uniform  From both plots, we can see that the observed points are very close to the red line. So, it's a uniform distribution.
	NULL HYPOTHESIS: The given data follows Uniform distribution. ALTERNATE HYPOTHESIS: The given data does not follow Uniform distribution
	• Each observation is observed frequency. • Expected frequency for each observation = $mean$ = 54.925080
	Calculated chi square statistic = 6.31 p-value = 0.99 • p-value > α We accept the null.
	$df = k - p - 1 = 50 - 0 - 1 = 49$ $k = 50$: number of observations $p = 0$: for uniform distribution
	Tabulated Chi Square value = 66.34 Tabulated value > Calculated value We accept the null.
Case examples	• Throwing a die. The probability of getting each side is 16 . It's a uniform distribution. • Fuel Efficiency. If you put one litre petrol/diesel, every time you would get a different mileage. It could be 50, 52, 47. In case you're a rash driver, it could be low, say 40, 42.

2.6 Guessing the Distribution of Dataset 3

Sunday, 02 October 2022 8:22

	Normal Distribution
	 • Somewhat Gaussian in nature. • Side bit is sitting between 90 to 100, so it's inclining towards something. It's telling us that there are significant number of observations that are towards one side. Meaning the distribution has a tail which is on the right side. So it could be a Gaussian with the right tail.
	count 48.000000 mean 56.958333 std 14.871018 min 27.000000 25% 48.000000 50% 56.000000 75% 68.000000 max 101.000000 {('Mean Observed', 56.96), ('Variance Observed', 221.15), ('Skew Observed', 0.37), ('Kurt Observed', 0.6)} • The difference between the minimum and the 25th quantile around 20. • Between 25th and 50th is around 10. • 50th to 75th is again around 10. • 75th to the maximum is more than 30, that means it has a tail on the right side. • Also, mean and the median are almost the same. Meaning, it's symmetric. So, huge symmetry with a slight tail. • Positive skew, means that the distribution is right tailed.
	Q-Q Plot  P-P Plot



- Both plots suggest that it's a normal distribution.

NULL HYPOTHESIS: The given data follows Normal distribution.

ALTERNATE HYPOTHESIS: The given data does not follow Normal distribution

We split the data into 6 equal intervals (you can split it in as many intervals you want).

As the area under the curve is one, the area under each bucket will be $\frac{1}{6} = 0.167$.

This the command to do that:

```
n = 1/6
for i in range(1,6):
  prob_intervals = [scipy.stats.norm.ppf(i*n,df['obs'].mean(), df['obs'].std())]
```

```
print(prob_intervals)
```

```
[42.571790240767264]
[50.552980104100826]
[56.958333333333336]
[63.363686562565846]
[71.3448764258994]
```

So the 6 buckets we have now are:

1st bucket : 42.57 and below
2nd bucket : 42.57 to 50.55
3rd bucket : 50.55 to 56.95
4th bucket : 56.95 to 63.36
5th bucket : 63.36 to 71.34
6th bucket : 71.34 and above

Since, we're splitting into six buckets, and we had 48 observations, so the expected frequency under each bucket $\frac{48}{6} = 8$.

- Expected freq = [8, 8, 8, 8, 8, 8]
- Observed freq = [9, 7, 9, 7, 9, 7]

- Observed frequencies can be obtained by first sorting the dataset in excel in ascending order, and then count the number of values fall in each buckets.
  - e.g., here, there are 9 values in the 1st bucket, i.e., 9 values in the dataset below 42.57

Calculated chi square statistic = 0.75
p-value = 0.98

- p-value > α
  We accept the null.

$df = k - p - 1 = 6 - 2 - 1 = 3$

$k = 6$: number of buckets

$p = 2$: for normal distribution

Tabulated Chi Square value = 7.81

Tabulated value > Calculated value
We accept the null.

Business Cases

- Most of our life events can be assumed to be normal.
  - e.g., Our test scores over the years could be a normal distribution.
  - Or test scores of an entire class will be normally distributed.

	• The data that we were working on comes from automotive sales. There were around 48 salesman who have sold tractors on a particular month. So, ideally you will see a couple of people who are performing extremely well, and there will be some salesman who may not be performing well, and in between you will have the majority.

Week 3

Monday, 10 October 2022 11:01

3.1 Determining association between Categorical variables

Monday, 10 October 2022 11:39

Summary	• Contingency table • Joint, marginal and conditional probabilities
• When we infer, it's about the sample data or population?	**Determining and inferring association** • Determining the association: Given a **sample data**, how do you determine that the variables are associated? • Inferring the association: We infer about the **population.** Once you have determined the association, how can you extend that to the population?
• What do we calculate to look if there was a gender discrimination in the admission process?	**Example** • Consider a B-school which shortlisted 1200 candidates (960 men and 240 women) for its post-graduate management program. Out of these, 324 candidates were given offer letters for admission. The data is included here: • After reviewing the record, a women's forum raised the issue of gender discrimination on the basis that 288 male candidates were offered admission against only 36 female candidates. • 1200 is not the population. It's a sample from the entire list of candidates. It's a sample data.
	Example • To this, the B-school management replied that it was not a case of discrimination, but was because of the fact that only 240 women candidates appeared for the examination. Let us review the case using probability. • Let *M* be the event that the candidate is a male. • Let *F* be the event that the candidate is a female. • Let *A* be the event that the candidate is offered admission. • Let A^c be the event that the candidate is not offered admission. (Event A^c is called compliment of event *A*. One can see that $Pr(A) + Pr(A^c) = 1$.)
	Example • Probability that a randomly observed candidate is a male and is offered the admission. $Pr(M \cap A) = 288/1200 = 0.24$ • Probability that randomly observed candidate is a male and is not offered the admission. $Pr(M \cap A^c) = 672/1200 = 0.56$ • Similarly, $Pr(W \cap A) = 36/1200 = 0.03$ $Pr(W \cap A^c) = 204/1200 = 0.17$

	Male	Female	Total
Offers made	288	36	324
Not offered	672	204	876
Total	960	240	1200

Contingency table

- W(woman) = F (female) (It's a typo here)

Example

- In terms of probabilities, the previous table can now be rewritten as:

	Male	Female	Total
Offers made	0.24	0.03	0.27
Not offered	0.56	0.17	0.73
Total	0.8	0.2	1.0

- Joint probabilities (of what?) appear in the main body of the table (e.g. 0.24, 0.03).
- Marginal probabilities (of what?) appear in the margin of the table (e.g. 0.8, 0.2).

Example

- What will be *Pr(A|M)*?
- First of all, what does this mean?
- This conditional probability tells us that we are concerned with admission status of only males!
- We know that out of the 960 male candidates, 288 were offered admission. So probability that a male candidate is offered admission will be 288/960 = 0.3.
- Also observe that:

$$\Pr(A \mid M) = \frac{288}{960} = \frac{288/1200}{960/1200} = \frac{0.24}{0.8} = 0.3.$$

- Relation between conditional, joint and marginal probabilities?

Example

- Now, the numerator, *0.24* is the joint probability of events *A* and *M*. that is *Pr(A∩M) = 0.24*. And 0.8 is the marginal probability of the event *M*, i.e. *Pr(M) = 0.8*.

$$\Pr(A \mid M) = \frac{\Pr(A \cap M)}{\Pr(M)}.$$

- This is, precisely, the definition of conditional probability.

- Back to the problem at hand:

$$\Pr(A \mid W) = \frac{\Pr(A \cap W)}{\Pr(W)} = \frac{0.03}{0.2} = 0.15.$$

- **Conditional Probability = $\frac{\text{Joint Probability}}{\text{Marginal Probability}}$**

$P(A \mid M) = 0.3$
$P(A \mid F) = 0.15$

- From these probabilities, what can you say about the discrimination?

Example: Conclusion

- The probability of admission offer given the candidate is a male is 0.3, twice of 0.15 probability of admission offer given the candidate is a woman.
- Although the use of conditional probability does not, in itself, prove discrimination, there is support for the argument!

3.2 Bayes' Rule

Monday, 10 October 2022 12:56

Summary	• Bayes' Rule
• Prior probability • Posterior probability • Where does Bayes' rule fit into this?	**Baye's rule** • Often we have initial guesses about an event from which we can calculate prior probabilities, using the usual probability theory. • Then, from sources such as data collection, sample, product field tests, we obtain more information about these events. • Given, this new information, we can update our prior beliefs by calculating revised probabilities – this is called the posterior probability. • Baye's rule is used to calculate the posterior probability if we have the initial belief (probability) and the additional sample information.

Baye's rule

- Suppose that a manufacturer receives same raw material from two different suppliers *S1* and *S2*.
- Currently 65% of the raw material comes from *S1* and remaining, 35%, comes from *S2*.
- Also, suppose that from the historical data available with the quality assurance department, we know that *S1* has 98% of the supplied raw material of good quality and *S2* has 95% of the raw material of good quality.
- That is, the probability of a "Good" quality raw material given that the supplier is *S1* is, *Pr(G|S1) = 0.98*. And for the second supplier, this probability is: *Pr(G|S2) = 0.95*.

65% of the raw material comes from S1	35% of the raw material comes from S2

	S1	**S2**	
Good	**98% of S1** P(G | S1)	**95% of S2** P(G | S2)	
Bad	**2% of S1** P(B | S1)	**5% of S2** P(B | S1)	
	P(s1) = **65%**	P(S2) = **35%**	

	S1	**S2**	
Good	0.98×0.65	0.95×0.35	
Bad	0.02×0.65	0.05×0.35	
	0.65	0.35	

Baye's rule

- What is the probability of the raw material being supplied by *S1* and it being good?
- Joint probability, of course!
- This can be calculated using the Baye's formula.

*Pr(S1, G) = Pr(S1 ∩ G) = Pr(S1)*Pr(G|S1) = 0.65*0.98 = 0.637*

*Pr(S2, G) = Pr(S2 ∩ G) = Pr(S2)*Pr(G|S2) = 0.35*0.95 = 0.3325*

- Now, knowing all this information so far, suppose the manufacturer inspects the incoming raw material on receipt and finds a bad quality material.
- He wants to know the supplier who needs to be contacted to complain!

	S1	**S2**	
Good	0.637	0.3325	0.9695
Bad	0.013	0.0175	0.0305
	0.65	0.35	**1**

• State Bayes' Rule

Baye's rule

- We are interested in the posterior probability that a particular supplier is guilty of supplying bad quality product *given that* we have bad quality raw material at our doorstep – *Pr(S1|B)* or *Pr(S2|B)*.
- This is an application of Baye's theorem – finding posterior probability given some initial facts and numbers.
- From Baye's formula we know that:

$$\Pr(S1 \mid B) = \frac{\Pr(S1 \cap B)}{\Pr(B)} = \frac{\Pr(S1) * \Pr(B \mid S1)}{\Pr(B)}.$$

	S1	**S2**	
Good	0.98×0.65 = 0.637 **P(G | S1) × P(S1)**	0.95×0.35 = 0.3325 **P(G | S2) × P(S2)**	0.9695 **P(G)**
Bad	0.02×0.65 = 0.013 **P(B | S1) × P(S1)**	0.05×0.35 = 0.0175 **P(B | S2) × P(S2)**	0.0305 **P(B)**
	0.65 **P(S1)**	0.35 **P(S2)**	**1**

- We have this contingency table, which is prior info.
- Now we found a bad quality material - this is the additional sample info.
- We want to know given a bad quality material (additional info) what is the probability that it came from supplier S1 (or S2)? - Posterior probability
- $P(S1 \mid B) \times P(B) = P(B \mid S1) \times P(S1)$

$$P(S1 \mid B) = \frac{P(B \mid S1) \times P(S1)}{P(B)}$$

Bayes' Rule

$$P(X \mid Y) = \frac{P(Y \mid X) \times P(X)}{P(Y)}$$

Baye's rule

- What is *Pr(B)*?
- That is the probability of receiving a bad quality raw material.
- Now bad quality raw material can from supplies of *S1* or *S2*.
- That is, the event *B* can occur with *S1* or with *S2*.

Pr(B) = Pr(S1 ∩ B) + Pr(S2 ∩ B)

- But *Pr(S1 ∩ B) = Pr(S1)*Pr(B|S1)*, and
- *Pr(S2 ∩ B) = Pr(S2)*Pr(B|S2)*

$$\Pr(S1 \mid B) = \frac{\Pr(S1) * \Pr(B \mid S1)}{\Pr(S1) * \Pr(B \mid S1) + \Pr(S2) * \Pr(B \mid S2)}.$$

- $P(B) = P(S1 \cap B) + P(S2 \cap B)$

$P(B) = P(B \mid S1) \times P(S1) + P(B \mid S2) \times P(S2)$

- Observe how probabilities change with the prior and posterior.

Baye's rule

$$\Pr(S1 \mid B) = \frac{\Pr(S1) * \Pr(B \mid S1)}{\Pr(S1) * \Pr(B \mid S1) + \Pr(S2) * \Pr(B \mid S2)}$$
$$= \frac{0.65 * 0.02}{0.65 * 0.02 + 0.35 * 0.05} = 0.426.$$
$$\Pr(S2 \mid B) = 0.574.$$

- Significance: Find posterior probabilities using prior information.
- Notice that we use *Pr(B|S1)* to find *Pr(S1|B)*.

	S1	**S2**	
Good	0.98×0.65 = 0.637 **P(G | S1) × P(S1)**	0.95×0.35 = 0.3325 **P(G | S2) × P(S2)**	0.9695 **P(G)**
Bad	0.02×0.65 = 0.013 **P(B | S1) × P(S1)**	0.05×0.35 = 0.0175 **P(B | S2) × P(S2)**	0.0305 **P(B)**
	0.65 **P(S1)**	0.35 **P(S2)**	**1**

- So, Prior was:
  - P (a randomly picked item was from S1) = 65%
  - P (a randomly picked item was from S2) = 35%
- Posterior:
  - P (a randomly picked item was from S1 | given that the item was bad) = 42.6%
  - P (a randomly picked item was from S2 | given that the item was bad) = 57.4%

- So we have found association between two categorical variables:

	1. Supplier (s1 & S2), and 2. Quality (good and bad)

3.3 Inferring association between categorical variables - Chi-squared test for Independence

Monday, 10 October 2022 18:00

Summary

- Independence of variables
- Chi-Square distribution
  - Formulae for:
    - Expected frequency
    - Chi-Square statistic
    - df

Inferencing about association

Example: Brand preferences

- Suppose a survey is conducted in Mumbai and Chennai asking respondents their preferences about three brands. The result is summarized below.

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	279	73	225	577
Chennai	165	47	191	403
Total	444	120	416	980

- Independent (explanatory) variable is the city.
- Dependent (response) variable is the brand preference.

- There are two categorical variables here:

1. Brand (A, B, C), and
2. City (Mumbai, Chennai)

- City => Independent (explanatory) variable
- Brand => Dependent (response) variable

Example: Brand preferences

- We know how to summarize the data by calculating the marginal and joint probabilities.
- What are the marginal probabilities? Joint probabilities?
- Now we want to answer the question: "Whether brand preference associated with city?" We use the basis of statistical independence/dependence for this.
- Two categorical variables are statistically independent if the population conditional distributions on one of them is identical to each category of the other.
- In the example, the two conditional distributions are not identical. e.g. Brand A is preferred more in Mumbai than in Chennai.

- Joint and Marginal Probabilities

	Brand A	Brand B	Brand C	
Mumbai	0.28	0.07	0.24	0.59
Chennai	0.17	0.05	0.19	0.41
	0.45	0.12	0.42	1

- Conditional Distribution: P(City | Brand)

	Brand A	Brand B	Brand C	
Mumbai	279 (48%)	73 (13%)	225 (39%)	577 (100%)
Chennai	165 (41%)	47 (12%)	191 (47%)	403 (100%)

- From conditional distribution, we can observe:

- Brand A is preferred more in Mumbai than in Chennai
- Brand B preference is identical in both cities
- Brand C is preferred more in Chennai than in Mumbai

- Since the conditional distributions are not identical, so we conclude that brand preference is associated with city.
- Hence, both categorical variables are dependent on each other.

How to find conditional distribution

Given a joint distribution or contingency table:

	Y1	**Y2**	**Y3**	
X1	*	*	*	**A**
X2	*	*	*	**B**
	K	**L**	**M**	**(total)**

- Find Conditional Distribution

$$\mathbf{P\,(Y \mid X) = \frac{P(X,Y)}{P(X)}}$$

	Y1	Y2	Y3	
X1	***/A**	***/A**	***/A**	**A**
X2	***/B**	***/B**	***/B**	**B**

and

$$\mathbf{P\,(X \mid Y) = \frac{P(X,Y)}{P(Y)}}$$

	Y1	Y2	Y3
X1	***/K**	***/L**	***/M**
X2	***/K**	***/L**	***/M**
	K	**L**	**M**

- Two variables are independent if conditional distributions on one of them is identical to each category of the other.

P (Y | X)

	Y1	Y2	Y3	
X1	***/A**	***/A**	***/A**	**A**
X2	***/B**	***/B**	***/B**	**B**

P (X | Y)

	Y1	Y2	Y3
X1	***/K**	***/L**	***/M**
X2	***/K**	***/L**	***/M**
	K	**L**	**M**

- That is, you find any one of the conditional distributions, and the values in same shaded cells should be equal.
- Then we say both variables are independent of each other.

Example: Brand preferences

- Refer to the same example extended to a third city:

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	440 (44%)	140 (14%)	420 (42%)	1000 (100%)
Chennai	44 (44%)	14 (14%)	42 (42%)	100 (100%)
Delhi	110 (44%)	35 (14%)	105 (42%)	250 (100%)

- Conditional distributions is same across the cities. Hence we can conclude that brand preference is independent of the cities.
- However, statistical independence is a symmetric property between two categorical variables.

- Here, brand preference does not depend on city.

- This is a sample data.
- Statistical independence is a symmetric property, so:
  - If brand preference is independent of city, P(City | Brand) = P(City), then
  - City is also independent of the brand, P(Brand | City) = P(Brand)
    - Proof:

	Brand A	Brand B	Brand C
Mumbai	440 (74%)	140 (74%)	420 (74%)
Chennai	44 (7%)	14 (7%)	42 (7%)
Delhi	110 (19%)	35 (19%)	105 (19%)
Total	594 (100%)	189 (100%)	567 (100%)

- Conclusion:
  - If X is independent of Y, then
  - Y is also independent of X

Example: Brand preferences

- If the conditional distributions within the rows are identical, then so are the distributions within the columns.
- One can verify that the conditional distribution amongst columns equals (74%, 7%, 19%).

- However, the example was a sample data. What about the population?
- Based on this single sample information, can we draw inferences about the population, as we have been doing?
- Answer is in testing our hypothesis, of course!

- Expected frequency = ?
- Can you tell why we assume the variables to be independent in the null hypothesis?

Chi-square distribution

- Null hypothesis –

H_0: The categorical variables are independent.

- Alternate hypothesis –

H_1: The categorical variables are not independent.

Let f_o be the observed frequencies (from the sample)

Let f_e be the expected frequencies, if the variables were independent.

The expected frequency for a cell equals the product of row and column totals for that cell, divided by the total sample size.

- Null hypothesis is always the no effect null hypothesis. Alternate hypothesis says the opposite thing.
- f_e, the expected frequencies, are calculated assuming that the null hypothesis is true.
- Expected frequency, $f_e = \frac{\textbf{Row total} \times \textbf{Column total}}{\textbf{Total Sample size}}$

- Chi-square formula

Example: Brand preference

- Brand preference example, with expected frequencies in brackets for each cell.

	Preferred brand			
City	Brand A	Brand B	Brand C	Total
Mumbai	279 (261.4)	73 (70.7)	225 (244.9)	577
Chennai	165 (182.6)	47 (49.3)	191 (171.1)	403
Total	444	120	416	980

- Chi-squared test statistic:

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}.$$

- One example:
  - $261.4 = \frac{444 \times 577}{980}$

- $\frac{(f_o - f_e)^2}{f_e}$

	Brand A	Brand B	Brand C
Mumbai	1.185	0.075	1.617
Chennai	1.696	0.107	2.314

- Chi-square formula:

- $\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$

$\chi^2 = 6.994 \approx 7$

- How do we calculate df in a contingency table?

Chi-square distribution

- When the H_0 is true, expected and observed frequencies tend to be close for each cell, and the test statistic value is relatively small.
- If H_0 is false, at least some cells have a big gap between expected and observed frequencies, leading to a large test statistic value.
- The larger the χ^2 value, greater is the evidence against the null hypothesis of independence.
- Degrees of freedom for the chi-squared distribution is given by the expression: *df = (r-1)*(c-1)*. *r* and *c* are the # of rows and columns respectively.

- **Degrees of freedom,**

$$Degrees\ of\ freedom = (Number\ of\ rows\ - 1) \times (Number\ of\ columns\ - 1)$$

$$\boldsymbol{df = (r - 1) \times (c - 1)}$$

• Given tabular and calculated chi-squared statistic, when do we accept H_o?	**Chi-square distribution** • For the brand preference example, calculated test statistic value is the $\chi^2 = 7.0$. • Degrees of freedom *df = 2*. So at $\alpha = 0.05$ (95% confidence), the tabular value of test statistic, $\chi^2 = 5.99$. • So we reject the null hypothesis of independence. • However, at $\alpha = 0.01$ (99% confidence), the tabular value of test statistic, $\chi^2 = 9.21$, and we can not reject the null hypothesis. • $df = (2-1) \times (3-1) = 1 \times 2 = 2$ • Calculated Chi-square statistic: $\chi^2 = 7$ • At: $df = 2$ and $\alpha = 0.05$ $(95\% \; confidence)$ ○ Tabular Chi-squared statistic: $\chi^2 = 5.99$ Tabular value < Calculated value » We reject the null hypothesis ○ Conclusion: cities and brand preferences are dependent. • At: $df = 2$ and $\alpha = 0.01$ $(99\% \; confidence)$ ○ Tabular Chi-squared statistic: $\chi^2 = 9.21$ Tabular value ≥ Calculated value » We accept the null hypothesis ○ Conclusion: cities and brand preferences are independent.
	• When we are concluding about hypotheses, we're essentially inferencing about the entire population.

3.4 Chi-squared test of Independence - Implementation in Python

Wednesday, 12 October 2022 12:43

	Later do on your own https://colab.research.google.com/drive/1Oj3HL5e2vYBAnp3_wqOM7jIMMlT5He3A?usp=sharing

3.5 Chi-squared test of Independence - Implementation in Spreadsheets

Wednesday, 12 October 2022 15:15

W3 Formulae

Sunday, 13 November 2022 14:10

$$\text{Conditional Probability} = \frac{\text{Joint Probability}}{\text{Marginal Probability}}$$

Bayes' Rule

$$P(X \mid Y) = \frac{P(Y \mid X) \times P(X)}{P(Y)}$$

- Two variables are independent if conditional distributions on one of them is identical to each category of the other.

P (Y | X)

	Y1	Y2	Y3	
X1	*/A	*/A	*/A	A
X2	*/B	*/B	*/B	B

P (X | Y)

	Y1	Y2	Y3
X1	*/K	*/L	*/M
X2	*/K	*/L	*/M
	K	L	M

- That is, you find any one of the conditional distributions, and the values in same shaded cells should be equal.
- Then we say both variables are independent of each other.

In contingency table,

Expected frequency:

$$f_e = \frac{\text{Row total} \times \text{Column total}}{\text{Total Sample size}}$$

Degree of Freedom,

$$df = (r - 1) \times (c - 1)$$

Chi Square formula,

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

We **accept** the null hypothesis if:

Tabular value ≥ Calculated value

or if:

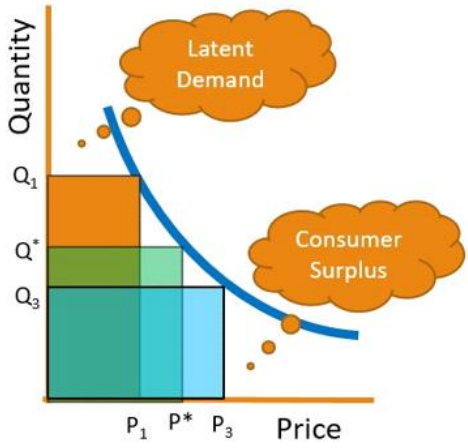
p*-value > *α

Week 4

Wednesday, 12 October 2022 15:29

4.1 Demand Response Curve

Wednesday, 12 October 2022 15:37

Summary	• Demand Response curve • Latent demand and consumer surplus
	Demand response curve Relationship between Price and Demand • Demand response curve: what is the realized demand at a given price for a particular product or service. • How do we estimate the demand response curve?
• Interpret the given curve • What is latent demand? • What is consumer surplus? • How is optimal price decided?	**Basic economics**  • The blue curve shown is the demand response curve. • P^* : Optimal price Q^* : Optimal demand • As P↑ ⇒ Q↓ • Latent Demand: By reducing the price, we can capture more demand. • Consumer Surplus: happens when the price that consumers pay for a product is less than the price they're willing to pay. It's a measure of the additional benefit that consumers receive because they're paying less for something than what they were willing to pay. • So if we increase the price from P^* to P_3, the consumer surplus is going to be eaten away by that much (the light blue region). • (Similar analogy goes for producer surplus) • How is optimal price decided? There are a few ways to go about this: 1. Revenue Maximizing Price 2. Profit maximizing Price Note: Both of the above prices are different.
• What is Demand Response Curve?	**Demand response curve**

• State the properties of demand response curve. • What does it mean for the curve to be downward sloping?	• A Function that describes how demand for a product *D(p)* varies as a function of its price • Similar to the Demand curve in Economics, but for a single seller, in a single market • Four properties 1. Non-negative 2. Continuous 3. Differentiable 4. Downward sloping • Downward sloping: As P↑ ⇒ Q↓ ○ This property sometimes may not hold. For example take an example for rolex watch, a luxury item. Its increase in price sometimes lead to increased demand due to the exclusivity appeal for the product. ○ But these are exception cases, so we will not be considering them.

4.2 Elasticity

Wednesday, 12 October 2022 17:04

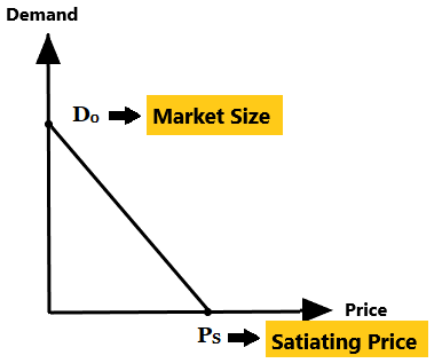
Summary	• Slope • Elasticity ○ Short term and long term elasticity
• What are the two ways to calculate price sensitivity?	There are two ways to calculate Price Sensitivity: 1. Slope 2. Elasticity
• Slope formula • It's value is always: • positive or negative?	**Price sensitivity** Slope – • Measures how demand changes in response to a price change $$\partial(p_1, p_2) = \frac{D(p_2) - D(p_1)}{p_2 - p_1}$$ • $p_1 > p_2$ implies that $D(p_1) < D(p_2)$, hence Slope is always negative. • Slope can be used as a local estimator of demand change for a small change in price. • Slope, $$\boldsymbol{\partial} = \frac{D(p_2) - D(p_1)}{p_2 - p_1}$$ • ∂ is always negative.
• Elasticity formula • Unit of elasticity	**Price sensitivity** Elasticity • Ratio of the percentage change in demand to the percentage change in price $$\epsilon(p_1, p_2) = -\frac{[d(p_2) - d(p_1)]/d(p_1)}{(p_2 - p_1)/p_1}$$ • Unlike slope, elasticity is independent of units • Elasticity of 2 means that a 10% reduction in price will yield a 20% increase in sales • Demand elasticity, $$\varepsilon = \frac{\%\ change\ in\ demand}{\%\ change\ in\ price}$$ $$\Rightarrow \varepsilon = \left| \frac{\frac{D(p_2) - D(p_1)}{D(p_1)}}{\frac{p_2 - p_1}{p_1}} \right|$$ • Elasticity may also depend on time. 1. Short term elasticity

	2. Long term elasticity
• What does high and low value of elasticity mean in terms of short and long term elasticity?	Elasticity | Product	Short term elasticity	Long term elasticity | |---|---|---| | Salt | 0 | 0.1 | | Airline Travel | 0.1 | 2.4 | | Petrol | 0.2 | 0.7 | | Movies | 0.9 | 3.7 | | A two-wheeler | 1.2 | 0.2 | • Elasticity ○ If **high**, means, **alternatives are available**. ▪ Short term: in the short term period ▪ Long term: in the long term period ○ If **low**, means, there is an **urgency and no alternative**. ▪ Short term: in the short term period ▪ Long term: in the long term period

4.3 Linear Response Curve

Wednesday, 12 October 2022 18:07

Summary	• Linear response curve • Constant elasticity curve

• What is linear response curve?
 	
• Define market size
• Define satiating price
 	
• Elasticity formula in terms of linear response curve
 	
• When price = 0
  • Elasticity = ?
 	
• When price → P_S
  • Elasticity = ?
	Linear response curve
	
	• Simplest Price Response Curve:
	
	$$D(p) = D_0 - m * p$$
	where, D_0 is the demand at price = 0 (this is called the market size) and m is the slope.
	• The price at which demand = 0 is called the satiating price, $P_s = \frac{D_0}{m}$.
	• The elasticity of this curve is $\epsilon = \frac{m*p}{D_0 - m*p}$.
	• We see that $\epsilon = 0$ when $p = 0$. $And\ as\ p \to P_s, \epsilon \to \infty$.
	
	 

• Constant Elasticity curve • Revenue formula	**Constant elasticity curve** • After algebraic transition, the constant elasticity curve is given by: $$D = Cp^{-\epsilon}$$ where C is a constant (it is the Demand when price = 1). • It is not guaranteed that the demand is either finite or satiated ($D \to \infty, as\ p \to 0. Also, D \neq 0, for\ any\ p$). • Revenue is $R = p * D = Cp^{(1-\epsilon)}$. • $D(p) = D(1)\ p^{-\varepsilon}$ 

- How to increase revenue for products with:

  (1) Inelastic demand
  (2) Elastic demand

Constant elasticity curve

We notice that:

- When $\epsilon < 1$, (inelastic product demand) the revenue can be increased by simply increasing the prices.
- When $\epsilon > 1$ (elastic demand) the revenue can only be increased by setting price close to zero.

- To increase revenue for products with
  - Inelastic demand: Increase the price
  - Elastic demand: Set the price close to zero
    - Huge demand ⇒ Increased revenue

4.4 Estimation Problem in Demand Response Curves

Wednesday, 12 October 2022 19:03

Summary	• Problem Statement
	Estimation problem • Assume that we conduct an market experiment where we offer different prices and check the realized demand at that time. • So we have the price offered and the corresponding demand values. • Price can be considered as an explanatory variable and the demand can be considered to be the dependent variable. • Can we estimate the slope and the elasticity? • Slope comes into the picture only when we have a linear response curve. • If you think elasticity could be constant, you may want to estimate the elasticity from the given data.

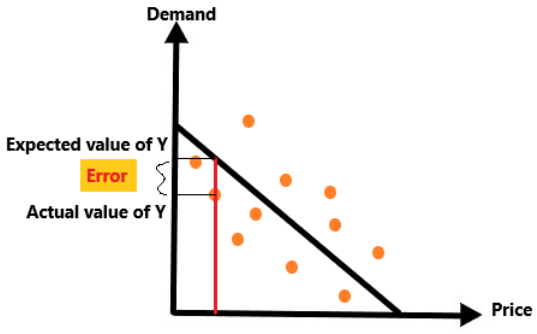
⭐ 4.5 Analysing Linear Demand Response Curve using Simple Linear Regression

Wednesday, 12 October 2022 19:14

Summary	• SLR • Constant elasticity model
• What is the full form of SLR? • What does a SLR model do?	**Linear demand response curve** • A SLR model can help identify D_0 (y-intercept) and m (the slope). • The SLR can tell us if the linear relationship is a good fit for the data available from the market experiment. • Please see the data and the corresponding SLR model. • SLR: Simple Linear Regression • SLR can also help in constant elasticity cases.
• How do you interpret this equation?	Interpreting the equation we got in dataset trend line: $y = -157.7x + 5842.8$ $R^2 = 0.7339$ $y = 5843 - 158x$ • y axis is the demand, and x axis is the price $D = D_0 - m.p,$ $\therefore D_0 = 5843 \rightarrow$ Market Size (total demand) $m = -157.7 \rightarrow$ Slope $R^2 = 0.7339$ • R represents the correlation coefficient between the price and the demand.
• State the SRM equation. • What does it describe?	**Simple Regression Model** Linear on Average ▪ The equation of the SRM describes how the conditional mean of Y depends on X. ▪ The SRM shows that these means lie on a line with intercept β_0 and slope β_1 $\mu_{y\|x} = E(Y\|X = x) = \beta_0 + \beta_1 x$ • Conditional mean: given a X, you find the value of Y. • $E(Y\|X = x)$ is the expected value of Y.
• What is error in SRM? • What is its expected value? • What is the objective of a regression model?	**Simple Regression Model** Deviations from the Mean

- Regression line is also called as _____ ?

- The deviations of responses around $\mu_{y|x}$ are called errors.
- Error, is denoted by ϵ, and $E(\epsilon) = 0$.



- Objective of Regression model is to minimize the Sum of Squared Errors (SSEs).
- The regression line is called as Mean Squared Error line.

- What are the assumptions SRM makes about the error term?

Simple Regression Model

Deviations from the Mean

- The SRM makes three assumptions about the error term

1. Independent. Errors are independent of each other.
2. Equal variance. All errors have the same variance, $Var(\epsilon) = \sigma_\epsilon^2$.
3. Normal. The errors are normally distributed.

- These assumptions have to be checked before performing SRM.

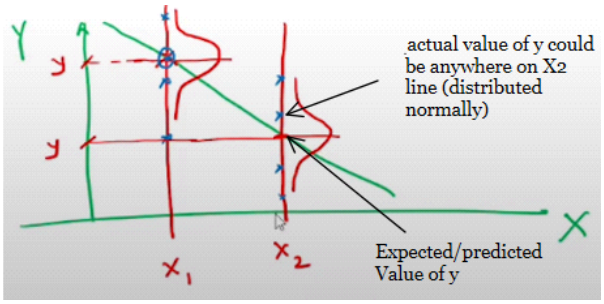
- In SLR, y = ?
- Interpret the given graph

Simple Regression Model

- Observed values of the response Y are linearly related to the values of the explanatory variable X by the equation

$$y = \beta_0 + \beta_1 x + \epsilon \text{ where } \epsilon \sim N(0, \sigma_\epsilon^2)$$

- The observations are independent of one another, have equal variance around the regression line, and are normally distributed around the regression line.



- Interpret given regression statistics
- Multiple R
- R Square
- Standard Error

Interpreting Regression Statistics

Regression Statistics	
Multiple R	0.856678246
R Square	0.733897617
Adjusted R Square	0.724721673
Standard Error	1290.448208
Observations	31

1. Multiple R
   Since we have linear regression here, so it's the **absolute value of correlation coefficient** (strength of correlation). Check here to see what it means in MLR.

2. R Square
   - **Correlation coefficient squared**
   - Another interpretation: **ability of the regression to explain the variability in y.**
     - Using price as an explanatory variable, it can explain 73% variability in demand.
   - It is also called as **Coefficient of Determination**
   - Also,

     $$\boldsymbol{R\ Square\ =\ \frac{Regression\ SS}{Total\ SS}}$$

     (refer anova table)
     $Regression\ SS$: amount explained by the regression
     $Total\ SS$ :: total variability

3. Adjusted R Square
   Makes sense only when we have more than one explanatory variable. Here.

4. Standard Error
   Sample standard deviation of error, s_e.
   s_e is the esimate of σ_ε → Standard Deviation of the error term

5. Observations
   Total number of observations, $n = 31$.

- Interpret given table

Interpreting ANOVA table- tells us whether regression is significant or not.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	133188236.7	133188236.7	79.98060985	7.80618E-10
Residual	29	48292440.76	1665256.578		
Total	30	181480677.4			

- df : $degrees\ of\ freedom$
- SS : $Sum\ of\ Squares$
- MS : $Means\ Squared$

- Regression df
- Residual df
- Total df

df	
Regression	1
Residual	29
Total	30

$y = \beta_0 + \beta_1 x$

- k = number explanatory variables = 1
- n = Total number of observations = 31

- We're predicting 2 parameters here, β_0 and β_1.

1. Regression df = total number of parameters − 1
   $= 2 - 1 = 1$

2. Residual $df = n - k - 1 = 31 - 1 - 1 = 29$.
   (Residual df = Total df - Regression df)

3. Total $df = n - 1 = 31 - 1 = 30$

- Sum of Squares
  - Regression SS

	SS
Regression	133188236.7
Residual	48292440.76
Total	181480677.4

- $Residual\ SS\ =\ SSE.$
- $Regression\ SS\ =\ SSM\ (or)\ SSR.$
  - Sum of the Squares of the model (SSM), or
  - Sum of the Squares of the regression (SSR)

- MS: Mean Squared error
- Standard Error

	MS
Regression	133188236.7
Residual	1665256.578
Total	

$$MS\ =\ \frac{SS}{df}.$$

$Residual\ MS$: MSE : Mean Squared Error

Also, $Standard\ Error\ =\ \sqrt{MSE}$.

- $F\ Test\ Statistic$
- When do we say x has no impact on y?

	F
Regression	79.98060985
Residual	
Total	

$$F-test\ statistic\ =\ \frac{Regression\ MS}{Residual\ MS}.$$

- F : $F\ Test\ Statistic$
  - Usually for hypothesis testing for overall significance of regression.
    - Here, Null hypothesis:
      - H_o : Regression model is not significant.
    - Alternate hypothesis:
      - H_A : Regression model is significant.
      - Regression model:
        $y\ =\ \beta_0\ +\ \beta_1 x.$
      - The question here is when will x have no impact on y?
        - When $\beta_0 = 0,\ \beta_1 = 0.$
        - And then, the entire regression model will collapse.
        - So, we can say:
      - H_o : $\beta_0 = 0, \beta_1 = 0.$
      - H_A : $\beta_0 \neq 0, \beta_1 \neq 0.$

- $p-value$ of F-test statistic

	Significance F
Regression	7.80618E-10
Residual	
Total	

- $Significance\ F\ :\ p-value\ for\ the\ hypothesis\ test.$
  - Here, we've taken confidence level as 95%.
  - So, significance level: Let $\alpha = 0.05$
  - Then $p-value\ <\ \alpha.$
    - So, we reject the null hyposthesis.
    - Conclusion: Regression model is significant. $\beta_0 \neq 0,\ \beta_1 \neq 0.$

Interpret the given table

Interpreting

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	5842.836198	400.6837789	14.58216306	6.94552E-15	5023.345856	6662.326539	5023.345856	6662.326539
Price	-157.7008739	17.63363083	-8.943187902	7.80618E-10	-193.7656983	-121.6360494	-193.7656983	-121.6360494

- Slope and Intercept

	Coefficients
Intercept	5842.836198
Price	-157.7008739

1. y-intercept = 5842 :: estimate of b_0.
2. Slope estimate = -157.7 :: estimate of b_1.

$y = -157.7x + 5842.8$

β_0, β_1 → Population value,
b_0, b_1 → Sample Value (these are the best values you can get from the given sample)

- Standard error of intercept and slope

	Standard Error
Intercept	400.6837789
Price	17.63363083

Standard error: for estimating the value of the intercept (b_0) and the slope (b_1).

- $t - stat$ and $p - value$.

	t Stat	P-value
Intercept	14.58216306	6.94552E-15
Price	-8.943187902	7.80618E-10

We can individually test whether $\beta_0 = 0$, and $\beta_1 = 0$, by running a localised hypothesis test.

- For the intercept and the slope:
  - $H_0 : \beta_0 = 0,\ \beta_1 = 0.$
  - $H_A : \beta_0 \neq 0,\ \beta_1 \neq 0.$
- $t - stat = \frac{sample\ value - hypothecial\ value}{Standard\ error\ of\ estimating\ the\ parameter}$.
- For Intercept: $t - stat = \frac{b_0 - 0}{Standard\ error\ for\ b_o} = \frac{5842.836198}{400.6837789} = 14.58216306$
- For Slope: $t - stat = \frac{b_1 - 0}{Standard\ error\ for\ b_1} = \frac{-157.7008739}{17.63363083} = -8.943187902$
- And for both cases, we can see that $p - value \ll 0.05$,
  - So we reject both null hypothesis.
  - So, we say that 5842.83 is a good estimate for b_0, and -157.70 is a good estimate of b_1.

- $p - value$ of t-stat test of slope and $p - value$ of F-stat test are same.
  Why?

Also, notice that the $p - value$, for the slope here, and $p - value$ of the F test statistic in the ANOVA table are exactly the same. Why?

	P-value
Price	7.80618E-10

	Significance F
Regression	7.80618E-10
Residual	
Total	

  - Because, we are running a Simple Linear Regression model.
  - In a simple linear regression model, regression will not be significant if $\beta_1 = 0$.
  - Here (in the current table), we're directly checking if $\beta_1 = 0$.
  - So, even if the test statistic was different, the overall test is the same.
  - That's why the $p - value$ is the same.

- Lower and upper bound

	Lower 95%	Upper 95%
Intercept	5023.345856	6662.326539
Price	-193.7656983	-121.6360494

with 95% confidence,
- $\beta_1 \in (-193.7, -121.6)$. Expected values = -157.7 (estimate)

	• $\beta_0 \in (5023.3, 6662.3)$. Expected values = 5842.8 (estimate)

4.6 Tutorial- "Building a Simple Linear Regression Model on Python"

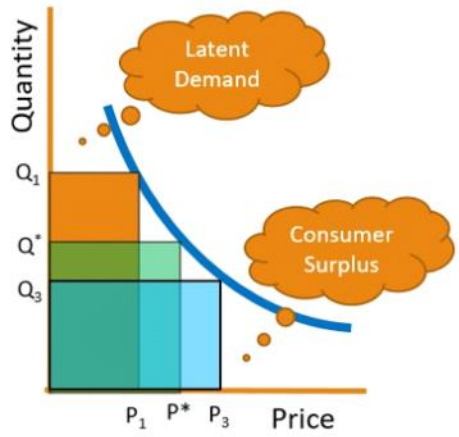
Thursday, 13 October 2022 9:47

	When you have time, watch the video again from Transformation bit (improving R2 value).

W4 Formulae

Sunday, 13 November 2022 15:16

Basic economics



Demand Response Curve

A function that describes how demand $D(p)$ of a product varies as a function of its price (p).

- **Latent Demand:** By reducing the price, we can capture more demand.
- **Consumer Surplus:** when the price that consumers pay for a product or service is less than the price they're willing to pay.
- As P↑ ⇒ Q↓
- 4 properties:
  1. Non-negative
  2. Continuous
  3. Differentiable
  4. Downward sloping

2 ways to decided optimal price:

1. Revenue Maximizing Price
2. Profit maximizing Price

2 ways to calculate Price Sensitivity:

1. Slope
2. Elasticity

$$\text{Slope} = \frac{\text{change in demand}}{\text{change in price}}$$

- ∂ is always negative.

$$\partial = \frac{D(p_2) - D(p_1)}{p_2 - p_1}$$

Demand elasticity,

$$\varepsilon = \frac{\%\text{ change in demand}}{\%\text{ change in price}}$$

$$\varepsilon = \left| \frac{\frac{D(p_2) - D(p_1)}{D(p_1)}}{\frac{p_2 - p_1}{p_1}} \right|$$

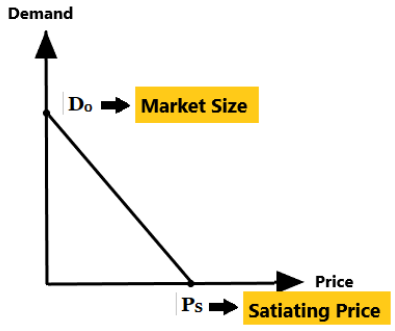
Elasticity:

- If **high**, means, **alternatives are available**.
- If **low**, means, there is an **urgency and no alternative**.

- Short term: in the short term period
- Long term: in the long term period

Linear Response Curve

$$\mathbf{D(p) = D(0) - m.p}$$



$\mathbf{D(0)}$: **Market Size** → Demand at price = 0

$P_S = \frac{\mathbf{D(0)}}{\mathbf{m}}$: **Satiating Price** ⇢ Price at which demand = 0

Elasticity of this curve

$$\varepsilon = \frac{\mathbf{m.p}}{\mathbf{D(p)}} = \frac{\mathbf{m.p}}{\mathbf{D(0) - m.p}}$$

- $\varepsilon = 0$ when $p = 0$
- As $p \longrightarrow P_S$, $\varepsilon \longrightarrow \infty$

Constant Elasticity Curve

$$\mathbf{D(p) = D(1)}\, p^{-\varepsilon}$$

Revenue

$$\mathbf{R = p \times D(p) = D(1)}\, p^{(1-\varepsilon)}$$

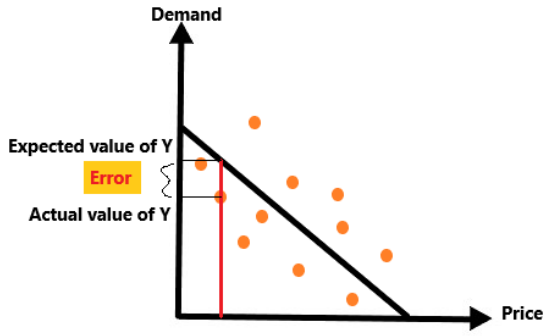
- As $p \longrightarrow 0$, $D \longrightarrow \infty$
- $D \neq 0$ for any p

- To increase revenue for products with
  - Inelastic demand: Increase the price
  - Elastic demand: Set the price close to zero
    - Huge demand ⇒ Increased revenue

Simple Linear Regression Model

The equation of SRM describes how the conditional mean of Y depends on X.

$$\mu_{Y|X} = E(Y \mid X = x) = \beta_0 + \beta_1 x$$



- Regression line is also called Mean Squared Error line.
- $E(\epsilon) = 0$
- We makes 3 assumptions about the error term:
  1. Independent
  2. Equal variance, $Var(\epsilon) = \sigma_\epsilon^2$
  3. Normal. $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

- Observed values of Y are linearly related to the explanatory variable X:

$$\boldsymbol{y = \beta_0 + \beta_1 x + \epsilon,}$$

where, $\boldsymbol{\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)}$

Week 5

Thursday, 13 October 2022 10:33

5.1 Non-linear Demand Response Curve

Tuesday, 18 October 2022 12:18

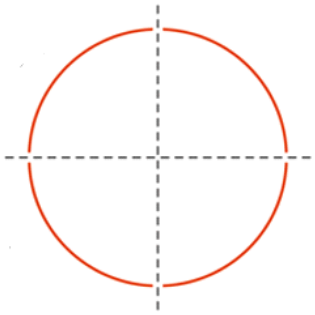
Summary	•
	Non – linear relationship • Price – Demand relationship not linear.... • $D(p) = Cp^{-\epsilon}$. • Given $D(p)$ and p, how do we estimate C and ϵ? • If the given data is non-linear, and is of constant elasticity curve type, can we still use simple linear regression and estimate C and ϵ?

5.2 Analysing Constant Elasticity Model using Simple Linear Regression

Tuesday, 18 October 2022 12:30

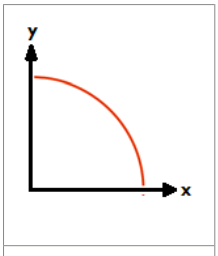
Summary	• Transformations • Transformation on Constant Elasticity Model

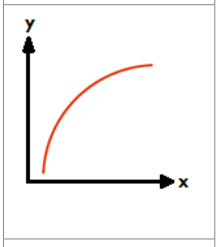
- What transform do we apply on each of these:

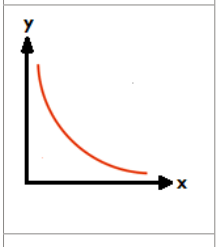


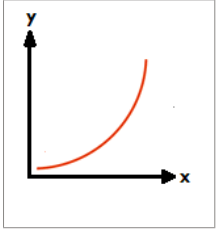
OR:

What transform do we apply on:
(also identify which curve is constant elasticity curve)



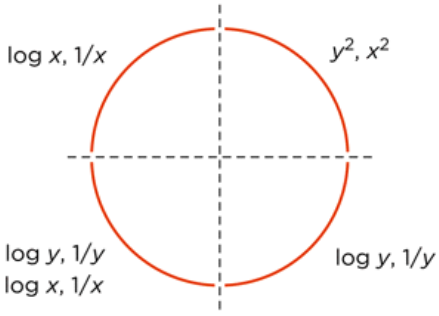






Constant elasticity model

- The constant elasticity model is not linear.
- Transformations allow the use of regression analysis to describe a curved pattern.
- Select the correct transformation --



- We're trying to model constant elasticity model using SLR.
- When we know that the relationship is not linear, we transform the equation, and model the curved pattern.

SLR Equation:

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

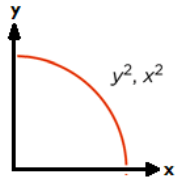
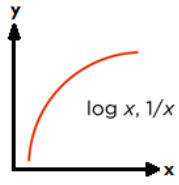
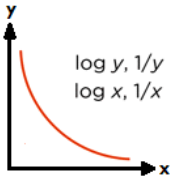
Note: ε → Error term

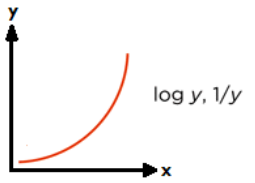
- Constant Elasticity Curve Equation:

$$D(p) = Cp^{-\epsilon}.$$

Note: ϵ → Elasticity

- We transform $D(p)$ $(or\ y)$ and p $(or\ x)$ in such a way that we get a linear relationship between them.

	If the relationship between x and y looks like this you use y^2 , x^2 trasformation.
	If it looks like this than transforming only the explanatory variable to $\log x$ or $\frac{1}{x}$ is sufficient.
	Here, either we can transform: ○ y to $\log y$ or $\frac{1}{y}$ or ○ x to $\log x$ or $\frac{1}{x}$. ○ Or do both.

	○ Also, when we say that the relationship is of $D(p) = Cp^{-\epsilon}$ (constant elasticity) type, we're essentially talking about above kind of relationship.
	If it looks like this than transforming only the response variable to $\log y$ or $\frac{1}{y}$ is sufficient.

- After applying log-log transformation, what equation do we get?
- How do you compare this equation with SLR equation?
- And how do find C and ε from this equation?

Constant elasticity model

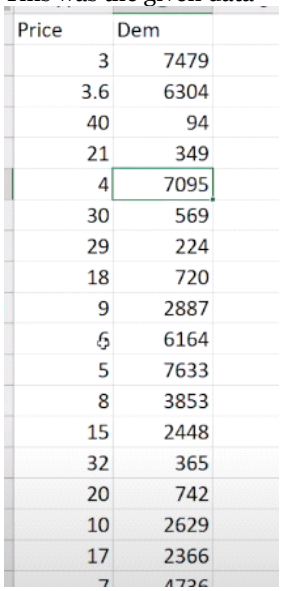
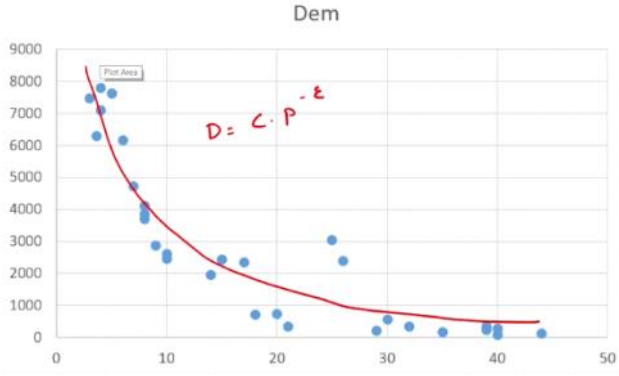
- Log-log transformation can convert the relationship to a linear one. And we get:

$$Log(D) = Log(C) - \epsilon Log(P).$$

- An example of this is available in the Excel sheet.
  - $D(p) = Cp^{-\epsilon}$.
  - Take log on both sides we get $\log(D) = \log(C) - \epsilon \log(p)$.
    - Here, $\log(D)$ → Response variable (y)
    - $\log(p)$ → Explanatory variable (x)
    - and, we've transformed both variables. Now, this relationship looks like a linear relationship. Equivalent to SLR equation. $y = \beta_0 + \beta_1 x + \varepsilon$.
    - $\beta_0 = \log(C)$ → y-intercept or market-size
    - $\beta_1 = -\epsilon$ → Slope
  - Now, we run SLR over this and get the estimate of β_0 and β_1.
    - $\therefore C = antilog(\beta_0) = e^{\beta_0} = D(1)$ → Demand when price = 1.
    - $\varepsilon = -\beta_1$ → Elasticity value.

5.3 Implementing Constant Elasticity Model using Simple Linear Regression

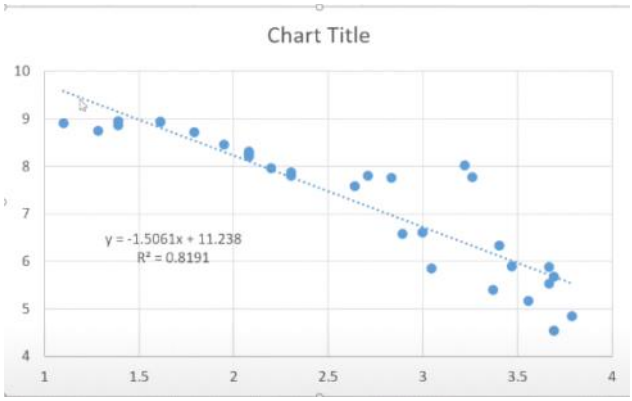
Tuesday, 18 October 2022 13:52

Summary	• Implementing Constant Elasticity Model using SLR on Excel
	This was the given data  The scatter plot of this data looks like this:  The graph looks like constant elasticity. So, we apply log-log transformation on the data.
	We did *ln* on both variables and then plotted the scatter plot

Price	Dem
3	7479
3.6	6304
40	94
21	349
4	7095
30	569
29	224
18	720
9	2887
6	6164
5	7633
8	3853
15	2448
32	365
20	742
10	2629
17	2366

Price	Dem	LN(P)	LN(Dem)
3	7479	1.098612	8.919854
3.6	6304	1.280934	8.74894
40	94	3.688879	4.543295
21	349	3.044522	5.855072
4	7095	1.386294	8.867146
30	569	3.401197	6.34388
29	224	3.367296	5.411646
18	720	2.890372	6.579251
9	2887	2.197225	7.967973
6	6164	1.791759	8.726481
5	7633	1.609438	8.940236
8	3853	2.079442	8.256607
15	2448	2.70805	7.803027
32	365	3.465736	5.899897
20	742	2.995732	6.609349
10	2629	2.302585	7.874359
17	2366	2.833213	7.768956
7	4736	1.94591	8.462948
39	252	3.663562	5.529429

Transformed scatter plot



$R^2 = 0.8191$

- R^2 → Cofficient of Determination
- The given relationship between explanatory variable and response variable is able to explain 81.9% of variation in response variable (demand).
- Therefore, this model looks to be a good model.

- We're given a data demand vs price.
- We make it ln(demand) vs ln(price), and do regression statistics on it.

Link to understand the Regression Statistics

After performing the test we got:

Intercept:
$\beta_0 = 11.238 = \ln(C)$.
$\Rightarrow C = e^{11.23} = 75962.87 = D(1)$.

Slope:
$\beta_1 = -1.506 = -\epsilon$.
$\Rightarrow \epsilon = 1.506$.

- $\epsilon > 1 \Rightarrow$ Highly elastic product.

5.4 Optimal Pricing - Revenue Maximization

Monday, 24 October 2022 10:10

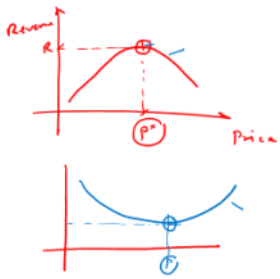
Summary	• Revenue Maximization
	Price Optimization Revenue Maximization Profit Maximization • Price is optimized based on what your objective is.
	Demand response curve • Let us assume a linear relationship between Price and Demand. • From our previous discussion, a linear relationship between Price and Demand is of the type: $$D(p) = D_0 + m * p$$ Where *D(p)* is the demand (as a function of price, *p*), D_0 is the market size (total demand when the price = 0), and *m* is the slope. For the sample we had seen in the last session, this relationship was: $$D(p) = 5842.8 - 157.7 * p$$
• Revenue Maximizing Price	**Sales Revenue function** • Revenue from sales is always calculated as $$Revenue = Demand * Price$$ $$R(p) = D(p) * p$$ $$R(p) = (D_0 + m * p) * p = D_0 * p + m * p^2$$ • For our numerical example, $R(p) = 5842.8 * p - 157.7 * p^2$ • Revenue, $R(p) = D(p) \times p = (D_o + mp) \times p$. • You take derivative of $R(p)$ w.r.t p, set it 0 and get p^*. $R(p) = D(p) \times p = (D_o + mp) \times p$ $= D_o\, p + m\, p^2$ $\frac{\partial R(p)}{\partial p} = 0.$ $D_o + 2mp = 0$ $\boldsymbol{p^* = -\frac{D_o}{2m}}$

Revenue maximization

- We now find the optimal price that maximizes the revenue.
- From the First Order Necessary Condition, we find the partial derivative of the revenue function w.r.to p, and set it to be zero.

$$\frac{\partial R(p)}{\partial p} = 5842.8 - 157.7 * 2 * p$$

$$\frac{\partial R(p)}{\partial p} = 0 \Rightarrow p^* = \frac{5842.8}{2 * 157.7} = 18.52$$



5.5 Optimal Pricing - Profit Maximization

Monday, 24 October 2022 15:35

Summary	• Profit Maximization
• What is marginal cost? • Profit Maximizing Price	

Profit function

- Typically, profit is the difference between the revenue and the cost.
- Assume that the marginal cost of producing the good is c.
- The profit function

$$\pi(p) = Total\ Revenue - Total\ Cost = D(p) * p - D(p) * c$$

$$\pi(p) = D(p) * (p - c) = (D_0 - m * p) * (p - c)$$

For our example, the profit function is, $\pi(p) = (5842.8 - 157.7 * p)(p - c)$

Marginal cost is the cost of producing one unit.
- If d units are produced at the marginal cost of c, then Total cost = $d \times c$.

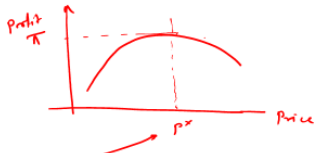
$$Total\ profit = Total\ Revenue - Total\ Cost$$
$$\pi(p) = D(p) \times p - D(p) \times c$$
$$= D(p) \times (p - c)$$
$$= (D_0 - mp) \times (p - c)$$
$$= D_0\, p - mp^2 - D_0\, c + mcp$$
$$\frac{\partial \pi(p)}{\partial p} = 0$$
$$D_0 - 2mp - 0 + mc = 0$$
$$2mp = D_0 + mc$$
$$\boldsymbol{p^* = \frac{D_0 + mc}{2m}}$$

Profit maximization

- To find the optimal price that maximizes profit, we again use the First Order Necessary Condition,

$$\frac{\partial \pi(p)}{\partial p} = D_0 - 2 * m * p - m * c$$

$$\frac{\partial \pi(p)}{\partial p} = 0 \Rightarrow p^* = \frac{D_0 + m * c}{2 * m}$$



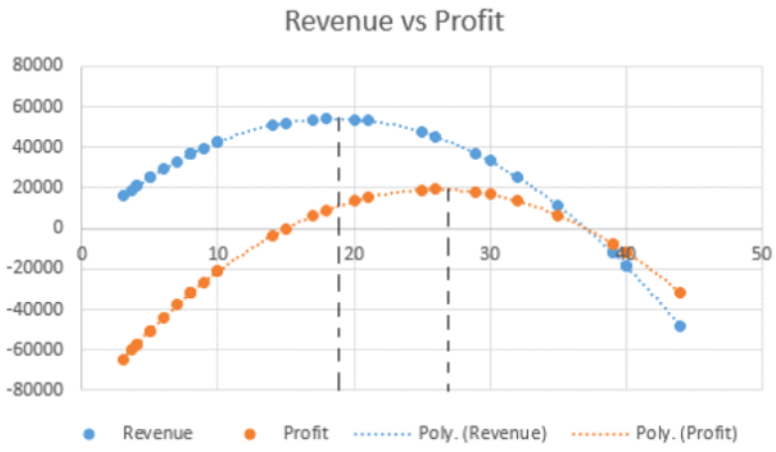
Profit maximization

- For our numerical example, let the marginal cost be 15.
- So the profit maximizing price is,

$$p^* = \frac{5842.8 + 157.7 * 15}{2 * 157.7} = 26.02$$

5.6 Revenue Maximization vs Profit Maximization

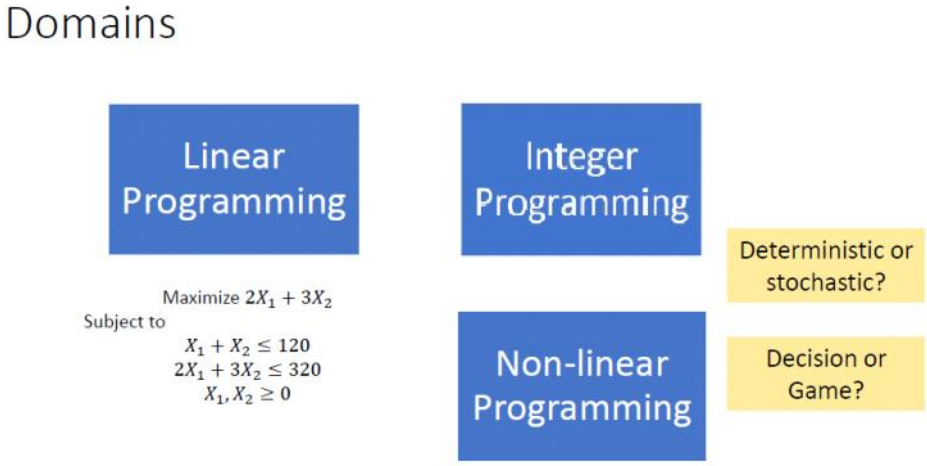
Monday, 24 October 2022 16:15

Summary	•
	$Predicted\ Demand\ =\ Intercept\ +\ Slope\ \times Price = b_0 + b_1 x$ • We've assumed that the $Marginal\ cost, c\ =\ 15$ $Revenue = Predicted\ Demand\ \times\ Price$ $Profit = Predicted\ Demand\ \times (Price\ -marginal\ cost)$ $= D(p)\ \times (p - c)$
• Notice that the revenue maximizing price and the profit maximizing price are different	This is the scatter plot we got in the analysis:  • It is clear that optimal price for revenue and optimal price for profit are not the same.

5.7 Operations Research: Linear Programming and Duality in Spreadsheets and Python

Monday, 24 October 2022 17:07

Summary	• Linear Programming
	

• What is Linear Programming (LP)?
• What is Integer Programming?
• What is Non-linear Programming?
• What assumptions about the variables do we make in LP?
	
	• In **Linear programming,** you have **linear objective function and a linear set of constraints.**
	  ○ Here, the **variables can take continuous values.**
	
	• In **Integer programming, variables can take only integer values.**
	  ○ Integer programming problems are **much more difficult to solve than linear programming problems.**
	  ○ If you're solving a linear programming problem, and you obtain integer solutions, then that is also going to be integer programming optimal solution.
	  ▪ However, if you're obtaining continuous values, then you need to be careful about rounding off. Because rounded off values may not generate an optimal integer programming solution.
	
	• In **non-linear programming, the objective function or the constraints are no longer linear.**
	
	• An assumption in linear programming is that the parameters or variables are **deterministic**.
	• It can sometimes happen that a level of uncertainty kicks in, in that case, you require stochastic programming to solve such problems.
	
	• What we're looking here is a particular firm's problem, for a single person's problem.
	• However, if there exists a scenario where the firm is trying to maximize its profit, given that a competitor firm has decided to produce a certain number of products. In that case, it's no longer a decision, and it becomes a game-theoretic problem.
	• Game theory is a subfield of economics, wherein different agents interact, and each agent tries to maximize his/her profit in relation to other agents.

- Here, we're gonna deal with a linear programming problem, deterministic parameters and variables, and we're going to focus on decision.

Problem Statement:

Linear Programming

Imagine you running an automobile firm which sells cars in three different segments – Hatchback, Sedan and SUV at prices ₹5,00,000, ₹10,00,000 and ₹25,00,000 respectively.

Suppose that the manufacturing of cars primarily requires the following raw materials A and B. The firm has 1,20,000 units of resource A and 1,40,000 units of resource B available. The resource requirements for the manufacturing of each car variant is given below.

Requirements	Resource A	Resource B
Hatchback	15	20
Sedan	20	50
SUV	60	100

How many cars of each type should be produced to maximize revenue?

Decision variables

X_1 - Number of Hatchback cars to be produced

X_2 - Number of Sedan cars to be produced

X_3 - Number of SUV cars to be produced

Objective function

Maximize $500000X_1 + 1000000X_2 + 2500000X_3$

Constraints

Requirements	Resource A	Resource B
Hatchback	15	20
Sedan	20	50
SUV	60	100

- Resource A constraint

$$15X_1 + 20X_2 + 60X_3 \leq 120000$$

- Resource B constraint

$$20X_1 + 50X_2 + 100X_3 \leq 140000$$

- Non-negativity restrictions

$$X_1, X_2, X_3 \geq 0$$

Linear program

$$\text{Maximize } 500000X_1 + 1000000X_2 + 2500000X_3$$

Subject to

$$15X_1 + 20X_2 + 60X_3 \leq 120000$$
$$20X_1 + 50X_2 + 100X_3 \leq 140000$$
$$X_1, X_2, X_3 \geq 0$$

Dual of the linear program

Automobile firm

- Possesses resources A and B
- Manufactures and sells cars
- Aim: Maximize revenue

Dual variables: Shadow price or Marginal price of the resource at the optimum

Primal

$$\text{Maximize } 500000X_1 + 1000000X_2 + 2500000X_3$$
Subject to
$$15X_1 + 20X_2 + 60X_3 \leq 120000$$
$$20X_1 + 50X_2 + 100X_3 \leq 140000$$
$$X_1, X_2, X_3 \geq 0$$

Dual

$$\text{Minimize } 120000Y_1 + 140000Y_2$$
Subject to
$$15Y_1 + 20Y_2 \geq 500000$$
$$20Y_1 + 50Y_2 \geq 1000000$$
$$60Y_1 + 100Y_2 \geq 2500000$$
$$Y_1, Y_2 \geq 0$$

Let Y_1, Y_2 be the costs of resource A and resource B respectively

Buyer

- Purchase resources A and B
- Aim: Minimize total cost

Primal – Dual relationship

Primal

$$\text{Maximize } 500000X_1 + 1000000X_2 + 2500000X_3$$
Subject to
$$15X_1 + 20X_2 + 60X_3 \leq 120000$$
$$20X_1 + 50X_2 + 100X_3 \leq 140000$$
$$X_1, X_2, X_3 \geq 0$$

Dual

$$\text{Minimize } 120000Y_1 + 140000Y_2$$
Subject to
$$15Y_1 + 20Y_2 \geq 500000$$
$$20Y_1 + 50Y_2 \geq 1000000$$
$$60Y_1 + 100Y_2 \geq 2500000$$
$$Y_1, Y_2 \geq 0$$

Primal	Dual
Maximization	Minimization
Number of constraints	Number of variables
Number of variables	Number of constraints
Objective function coefficient	Right hand side in constraints
Right hand side in constraints	Objective function coefficient

How to construct a dual?

Primal

Maximize $500000X_1 + 1000000X_2 + 2500000X_3$

Subject to

$$15X_1 + 20X_2 + 60X_3 \leq 120000$$
$$20X_1 + 50X_2 + 100X_3 \leq 140000$$
$$X_1, X_2, X_3 \geq 0$$

Dual

Let Y_1, Y_2 be the dual variables corresponding to the two constraints

Minimize $120000Y_1 + 140000Y_2$

Subject to

$$15Y_1 + 20Y_2 \geq 500000$$
$$20Y_1 + 50Y_2 \geq 1000000$$
$$60Y_1 + 100Y_2 \geq 2500000$$
$$Y_1, Y_2 \geq 0$$

Dual of the dual is the primal!

Dual

Minimize $120000Y_1 + 140000Y_2$

Subject to

$$15Y_1 + 20Y_2 \geq 500000$$
$$20Y_1 + 50Y_2 \geq 1000000$$
$$60Y_1 + 100Y_2 \geq 2500000$$
$$Y_1, Y_2 \geq 0$$

Standard form

Maximize $-120000Y_1 - 140000Y_2$

Subject to

$$-15Y_1 - 20Y_2 \leq -500000$$
$$-20Y_1 - 50Y_2 \leq -1000000$$
$$-60Y_1 - 100Y_2 \leq -2500000$$
$$Y_1, Y_2 \geq 0$$

Primal

Maximize $500000X_1 + 1000000X_2 + 2500000X_3$

Subject to

$$15X_1 + 20X_2 + 60X_3 \leq 120000$$
$$20X_1 + 50X_2 + 100X_3 \leq 140000$$
$$X_1, X_2, X_3 \geq 0$$

Finding the dual

Minimize $-500000X_1 - 1000000X_2 - 2500000X_3$

Subject to

$$-15X_1 - 20X_2 - 60X_3 \geq -120000$$
$$-20X_1 - 50X_2 - 100X_3 \geq -140000$$
$$X_1, X_2, X_3 \geq 0$$

- Solved on the next page

How to construct a dual?

Primal

Minimize $500X_1 + 100X_2 + 200X_3$

Subject to

$$15X_1 + 20X_2 + 60X_3 \geq 1200$$
$$20X_1 + 50X_2 + 100X_3 \leq 1400$$
$$X_1 \geq 0, X_2 \leq 0, X_3 \text{ unrestricted}$$

Dual

Minimize $-1200Y_1 + 1400Y_2$

Subject to

$$-15Y_1 + 20Y_2 \geq -500$$
$$20Y_1 - 50Y_2 \geq 100$$
$$-60Y_1 + 100Y_2 \geq -200$$
$$60Y_1 - 100Y_2 \geq 200$$
$$Y_1, Y_2 \geq 0$$

Convert to standard form

Define new variables $X_4, X_5, X_6 \geq 0$. Let $X_3 = X_4 - X_5$ and $X_6 = -X_2$.

Minimize $500X_1 - 100X_6 + 200(X_4 - X_5)$

Subject to

$$15X_1 - 20X_6 + 60(X_4 - X_5) \geq 1200$$
$$20X_1 - 50X_6 + 100(X_4 - X_5) \leq 1400$$
$$X_1, X_4, X_5, X_6 \geq 0$$

Convert objective function to maximization

Convert ≥ constraint to ≤ constraint

Maximize $-500X_1 + 100X_6 - 200X_4 + 200X_5$

Subject to

$$-15X_1 + 20X_6 - 60X_4 + 60X_5 \leq -1200$$
$$20X_1 - 50X_6 + 100X_4 - 100X_5 \leq 1400$$
$$X_1, X_4, X_5, X_6 \geq 0$$

- We require a maximization objective.
- And, we require less than or equal to constraints.
- And, we require all the variables to be greater than or equal to 0.

- X_4 and X_5 will handle the unrestricted sign of X_3, as X_3 can take any value.
- X_6 will handle the negative sign of X_2.

- We substitute these new variables in the primal.

- Solved on next page.

Dealing with an equal to constraint:

$$2X_1 + X_2 = 400.$$

We want to convert this constraint to a less than or equal to type.

- First, write it as two constraints

$2X_1 + X_2 \leq 400$
$2X_1 + X_2 \geq 400$

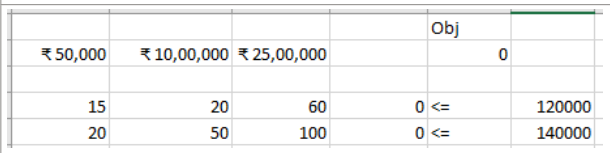
- Now, convert these two into a less than or equal to constraint.

$2X_1 + X_2 \leq 400$
$-2X_1 - X_2 \leq -400$

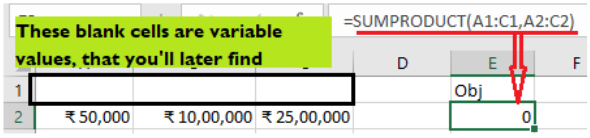
- That's how you do it.

Excel Working

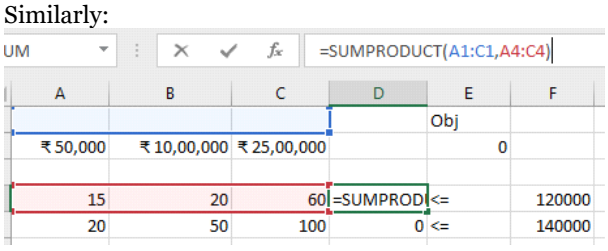
- To solve in Excel, you'll require the Solver add-in.
  - File >> Options >> Add-ins >> Go >> Tick Solver Add-in >> Ok



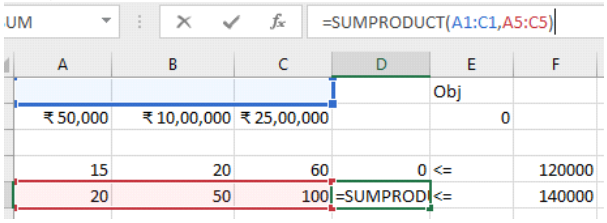
				Obj	
₹ 50,000	₹ 10,00,000	₹ 25,00,000		0	
15	20	60	0	<=	120000
20	50	100	0	<=	140000



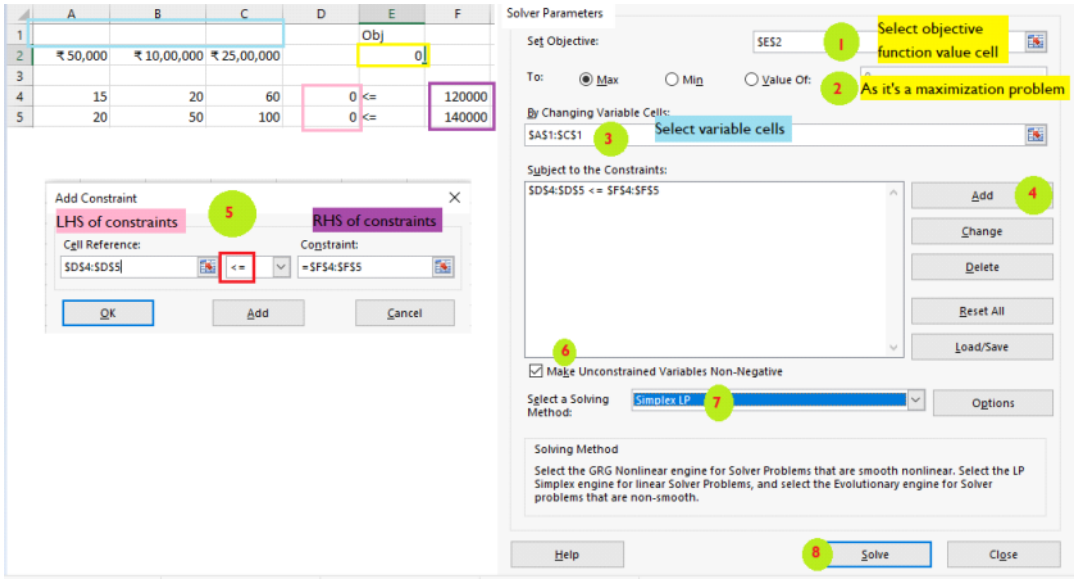
Similarly:

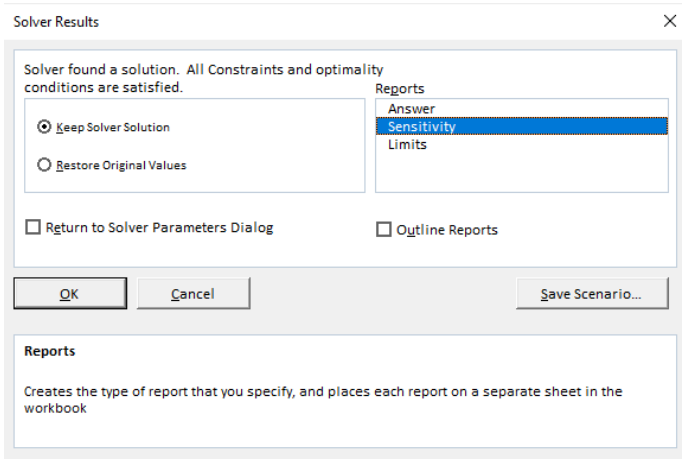


And,



Go to Data >> Solver





This is the solution you get:

0	0	1400		Obj	
₹ 50,000	₹ 10,00,000	₹ 25,00,000		3.5E+09	
15	20	60	84000	<=	120000
20	50	100	140000	<=	140000

Sensitivity report on Primal problem:

Microsoft Excel 16.0 Sensitivity Report
Worksheet: [Book1]Sheet1
Report Created: Tue, 25-Oct-2022 12:58:11

Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$A$1		0	-450000	50000	450000	1E+30
$B$1		0	-250000	1000000	250000	1E+30
$C$1		1400	0	2500000	1E+30	500000

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$D$4		84000	0	120000	1E+30	36000
$D$5		140000	25000	140000	60000	140000

- Allowable increase and decrease tells us in what range the solution will hold.

- Dual variable are also called the Shadow price.
- So, the sensitivity report directly gives you the value of dual the variables (in the final value

column of constraints table).

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$D$4	A	84000	0	120000	1E+30	36000
$D$5	B	140000	25000	140000	60000	140000

Resource B is exhausted

Whereas only 84,00 units of resource A are used
i.e., we have 1,20,000-84,000 = 36,000 units of resource A left

- Resource A is not that much valuable to us.
- If we have a tight constraint, meaning LHS = RHS, that resource is that much valuable.
- If it's not a tight constraint, that resource may not be of that much value to us.
- That's what we observe in the Shadow Price column.
- Let's verify this for solving for dual.

Dual problem

			Obj	
1,20,000	1,40,000		0	
15	20	0	>=	5,00,000
20	50	0	>=	10,00,000
60	100	0	>=	25,00,000

Follow the steps mentioned above and the solution you get:

			Obj	
0	25,000			
1,20,000	1,40,000		3.5E+09	
15	20	5,00,000	>=	5,00,000
20	50	12,50,000	>=	10,00,000
60	100	25,00,000	>=	25,00,000

And, the sensitivity report:

Microsoft Excel 16.0 Sensitivity Report
Worksheet: [Primal and Dual.xlsx]Dual
Report Created: Tue, 25-Oct-2022 13:28:58

Variable Cells

Cell	Name	Final Value	Reduced Cost	Objective Coefficient	Allowable Increase	Allowable Decrease
$A$1		0	15000	120000	1E+30	15000
$B$1		25000	0	140000	20000	140000
$C$1		0	0	0	1E+30	0

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$C$4		500000	7000	500000	1E+30	5.45697E-11
$C$5		1250000	0	1000000	250000	1E+30
$C$6		2500000	0	2500000	2.72848E-10	1E+30

- You can observe that the objective function value of the primal and the dual are the same.

Primal-Dual

Sunday, 13 November 2022 20:03

Rules

Primal	Dual
Maximization	Minimization
Number of constraints	Number of variables
Number of variables	Number of constraints
Objective function coefficient	Right hand side in constraints
Right hand side in constraints	Objective function coefficient

Know this:

In Maximization	you have ≤
In Minimization	you have ≥

It's important to make all the variables ≥ 0, in both min and max problems

Primal	Standard form	Dual
Max	Max with ≤	Min with ≥
Min	Min with ≥	Max with ≤

Dealing with an equal to constraint:

$$2X_1 + X_2 = 400.$$

For max problem, we want ≤:

- First, write it as two constraints

$$2X_1 + X_2 \leq 400$$
$$2X_1 + X_2 \geq 400$$

- Now, convert these two into a less than or equal to constraint.

$$2X_1 + X_2 \leq 400$$
$$-2X_1 - X_2 \leq -400$$

One important thing to know:

Let's say we have a Max optimization problem:

Max _____

subject to

Constraint ❶: _____ ≤ _____

Constraint ❷: _____ ≤ _____

and Decision variables ≥ 0

And we also have the optimal solutions. Let's say the optimal solutions satisfy the constraints:

Constraint ❶ as: LHS = RHS (binding)

Constraint ❷ as: LHS < RHS (not binding)

Binding: A constraint is called "binding" or "active" if it is satisfied as an equality at the optimal solution

---

Now, we already know that in Dual:
variable ⇒ # constraints
constraints ⇒ #variables

If primal variables: X_1, X_2, X_3
and dual variables: Y_1, Y_2
then

The dual variable corresponding to:

- Constraint ❶ will be Y_1.
- Constraint ❷ will be Y_2.

The dual variables corresponding to:

- Binding constraint: will be some positive value.
- Non-binding constraint: will be 0.
  - meaning strict inequality

Conclusion:

Since:

Constraint ❶: LHS = RHS (binding)

Constraint ❷: LHS < RHS (not binding)

∴ Y_1 +ve quantity, and

$\boldsymbol{Y_2 = 0}$

Q1. Convert the given Primal to dual

Min $120000Y_1 + 140000Y_2$

subject to
$15Y_1 + 20Y_2 \geq 500000$
$20Y_1 + 50Y_2 \geq 1000000$
$60Y_1 + 100Y_2 \geq 2500000$
$Y_1, Y_2 \geq 0$

This problem is already in Standard form.

Dual:
1. Min ⇒ Max
2. # variable ⇒ # constraints
3. # constraints ⇒ #variables
4. ≥ ⇒ ≤

∴ Dual:

Max $500000X_1 + 1000000X_2 + 2500000X_3$

subject to
$15X_1 + 20X_2 + 60X_3 \leq 120000$
$20X_1 + 50X_2 + 100X_3 \leq 140000$
$X_1, X_2, X_3 \geq 0$

Q2. Convert the given Primal to dual

Max $500X_1 + 100X_2 + 200X_3$

subject to
$15X_1 + 20X_2 + 60X_3 \geq 1200$
$20X_1 + 50X_2 + 100X_3 \leq 1400$
$X_1 \geq 0, X_2 \leq 0, X_3$ *unrestricted*

Notice, not all variables are ≥ 0

∴ Let, $X_4 - X_5 = X_3$ (∵ X_3 is *unrestriced*)
and,
$X_2 = -X_6$

Now, the problem becomes

Min $500X_1 - 100X_6 + 200X_4 - 200X_5$

subject to
$15X_1 - 20X_6 + 60X_4 - 60X_5 \geq 1200$
$20X_1 - 50X_6 + 100X_4 - 100X_5 \leq 1400$
$X_1, X_4, X_5, X_6 \geq 0$

Convert it to Standard form:
(Max with ≤)

Max $500X_1 - 100X_6 + 200X_4 - 200X_5$

subject to
$-15X_1 + 20X_6 - 60X_4 + 60X_5 \leq -1200$
$20X_1 - 50X_6 + 100X_4 - 100X_5 \leq 1400$
$X_1, X_4, X_5, X_6 \geq 0$

Dual:
1. Max ⇒ Min
2. # variable ⇒ # constraints
3. # constraints ⇒ #variables
4. ≤ ⇒ ≥

∴ Dual:

Min $-1200Y_1 + 1400Y_2$

subject to
$-15Y_1 + 20Y_2 \geq 500$
$20Y_1 - 50Y_2 \geq -100$
$-60Y_1 + 100Y_2 \geq 200$
$60Y_1 - 100Y_2 \geq -200$
$Y_1, Y_2 \geq 0$
(I'm getting constraints' RHS wrong, I don't know why)

Q3. Convert the given Primal to dual

Max $X_1 + 2X_2 + X_3$

subject to
$2X_1 + X_2 - X_3 \leq 2$
$-2X_1 + X_2 - 5X_3 \geq -6$
$4X_1 + X_2 + X_3 \leq 6$
$X_1, X_2, X_3 \geq 0$

Convert it to Standard form:
(Max with ≤)

Max $X_1 + 2X_2 + X_3$

subject to
$2X_1 + X_2 - X_3 \leq 2$
$2X_1 - X_2 + 5X_3 \leq 6$
$4X_1 + X_2 + X_3 \leq 6$
$X_1, X_2, X_3 \geq 0$

Dual:
1. Max ⇒ Min
2. # variable ⇒ # constraints
3. # constraints ⇒ #variables
4. ≤ ⇒ ≥

∴ Dual:

Max $2Y_1 + 6Y_2 + 6Y_3$

subject to
$2Y_1 + 2Y_2 + 4Y_3 \geq 1$
$Y_1 - Y_2 + Y_3 \geq 2$
$-Y_1 + 5Y_2 + Y_3 \geq 1$
$Y_1, Y_2, Y_3 \geq 0$

Q4. Convert the given Primal to dual

Min $2X_2 + 5X_3$

subject to
$X_1 + X_2 \geq 2$
$2X_1 + X_2 + 6X_3 \leq 6$
$X_1 - X_2 + 3X_3 = 4$
$X_1, X_2, X_3 \geq 0$

Dealing with = constraint

$X_1 - X_2 + 3X_3 \geq 4$
$X_1 - X_2 + 3X_3 \leq 4$
So, the problem becomes:

Min $2X_2 + 5X_3$

subject to
$X_1 + X_2 \geq 2$
$2X_1 + X_2 + 6X_3 \leq 6$
$X_1 - X_2 + 3X_3 \geq 4$
$X_1 - X_2 + 3X_3 \leq 4$
$X_1, X_2, X_3 \geq 0$

Convert it to Standard form:
(Min with ≥):

Min $0X_1 + 2X_2 + 5X_3$

subject to
$X_1 + X_2 + 0X_3 \geq 2$
$-2X_1 - X_2 - 6X_3 \geq -6$
$X_1 - X_2 + 3X_3 \geq 4$
$-X_1 + X_2 - 3X_3 \geq -4$
$X_1, X_2, X_3 \geq 0$

Dual:
1. Min ⇒ Max
2. # variable ⇒ # constraints
3. # constraints ⇒ #variables
4. ≥ ⇒ ≤

∴ Dual:

Max $2Y_1 - 6Y_2 + 4Y_3 - 4Y_4$

subject to
$Y_1 - 2Y_2 + Y_3 - Y_4 \leq 0$
$Y_1 - Y_2 - Y_3 + Y_4 \leq 2$
$-6Y_2 + 3Y_3 - 3Y_4 \leq 5$
$Y_1, Y_2, Y_3, Y_4 \geq 0$

5.8 Implementing Constant Elasticity Model using Simple Linear Regression in Python

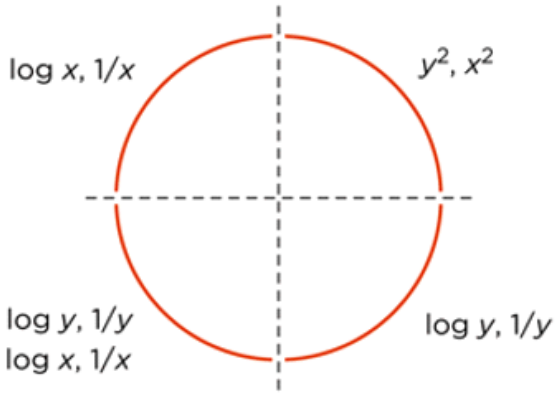
Tuesday, 25 October 2022 18:05

Summary	•
	Data and the python file both are not available.
• Mean Squared Error • RMSE	Mean Squared Error, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ Root Mean Squared Error, $RMSE = \sqrt{MSE}$

W5 Formulae

Sunday, 13 November 2022 21:56

Transformations



We apply log-log transformation in Constant Elasticity Model to convert it into SLR equation.

$$D(p) = Cp^{-\epsilon}$$

Take log

$$\log(D) = \log(C) - \epsilon \log(p)$$

Compare with

$$y = \beta_0 + \beta_1 x$$

then

$y = \log(D)$ → Response variable
$x = \log(p)$ → Explanatory variable

$\beta_0 = \log(C)$ → y-intercept or market-size
$\beta_1 = -\epsilon$ → Slope

and

$\therefore C = e^{\beta_0} = D(1)$ → Demand when price = 1.
$\epsilon = -\beta_1$ → Elasticity value.

Revenue maximizing price	Profit maximizing price
$p^* = -\frac{D(0)}{2m}$	$p^* = \frac{D(0) + mc}{2m}$

Primal-Dual

Mean Squared Error,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error,

$$RMSE = \sqrt{MSE}$$

Week 6

Tuesday, 25 October 2022 18:35

6.1 Multiple Linear Regression

Wednesday, 26 October 2022 13:07

Summary	• Multiple Linear Regression • $\bar{R}^2$ and s_e • Calibration Plot • Marginal and Partial Slopes • Path Diagram and Collinearity • F-statistic • t-statistic • Prediction interval
• What's the basic difference between multiple and simple linear regression?	**Multiple Regression** • In simple linear regression, there's one explanatory variable and one response (dependent) variable. • In multiple linear regression, we have multiple explanatory variables and one response variable.
	The Multiple Regression Model ▪ Use multiple regression to describe the relationship between several explanatory variables and the response. ▪ Multiple regression separates the effects of each explanatory variable on the response and reveals which really matter. • We will study the effect of each of these explanatory variable on the response variable.
• Full form of MLR and MRM. • What is k in MRM?	**The Multiple Regression Model** ▪ Multiple regression model (MRM): model for the association in the population between multiple explanatory variables and a response. ▪ k: the number of explanatory variables in the multiple regression (k = 1 in simple regression). • We can call it either: MLR / MRM. ○ MLR : Multiple Linear Regression ○ MRM: Multiple Regression Model
• Equation of MRM • Assumptions about the error term	**The Multiple Regression Model** The response Y is linearly related to k explanatory variables X_1, X_2, and X_k by the equation $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k + \epsilon$ $\epsilon \sim N(0, \sigma_\epsilon^2)$ ▪ *The unobserved errors in the model* 1. *are independent of one another,* 2. *have equal variance, and* 3. *are normally distributed around the regression equation.* • We will estimate $\beta_0, \beta_1, \ -- - , \beta_k$.

- Difference in the error terms between MLR and SLR

The Multiple Regression Model

- While the SRM bundles all but one explanatory variable into the error term, multiple regression allows for the inclusion of several variables in the model.

- In the MRM, residuals departing from normality may suggest that an important explanatory variable has been omitted.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + --- + \beta_k X_k + \varepsilon$$

- The impact of all other explanatory variables, which was not considered in SLR, will get lumped into the error term of SLR.

- $\bar{R}^2$
- s_e
- How do they behave when an explanatory variable is added?

Interpreting Multiple Regression

- R-squared and s_e

- $\bar{R}^2$ is known as the adjusted *R*-squared. It adjusts for both sample size *n* and model size *k*. It is always smaller than R^2.

- The residual degrees of freedom (*n-k-1*) is the divisor of s_e. $\bar{R}^2$ and s_e move in opposite directions when an explanatory variable is added to the model ($\bar{R}^2$ goes up while s_e goes down).

☐ ? If $\bar{R}^2 < R^2$, and $\bar{R}^2$ keeps increasing as you add more explanatory variables, then does this $\bar{R}^2 < R^2$ relationship always hold? Does it mean that no matter how many explanatory variables you add, you'll never cross R^2?

$Adjusted\ R^2 < R^2$

- Adjusted R^2 gives us a slightly more realistic picture of what is the combined explanatory power of all these explanatory variables.

s_e is the esimate of σ_ε → Standard Deviation of the error term, standard error in estimating the slopes

- In general, we want a large value of $adjusted\ R^2$, and we want a smaller value of s_e ,*i.e.*, if you add explanatory variables to the model:
  - $\bar{R}^2 \uparrow$ and $s_e \downarrow$

- State Adjusted R squared formula

Adjusted R squared formula
(not in the lectures but is asked in the questions)

$$\bar{R}^2 = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{(n - k - 1)}\right]$$

- Define Calibration Plot
- What does *R* represent in:
  - SLR
  - MLR

Interpreting Multiple Regression

- Calibration Plot

- Calibration plot: scatterplot of the response y on the fitted values $\hat{y}$.

- R^2 is the correlation between $\hat{y}$ and y; the tighter data cluster along the diagonal line in the calibration plot, the larger the R^2 value.

- In SLR, *R* represented the correlation between *X* and *Y*.
- In MLR, *R* represents the coefficent of correlation between observed value(y) of the response variable and the fitted/predicted value($\hat{y}$) of the response variable.

- Define
  - Marginal slope
  - Partial slope
- When is Marginal slope = Partial slope?

Interpreting Multiple Regression

- Marginal and Partial Slopes

- Partial slope: slope of an explanatory variable in a multiple regression that statistically excludes the effects of other explanatory variables.

- Marginal slope: slope of an explanatory variable in a simple regression.

- Partial and marginal slopes only agree when the explanatory variables are uncorrelated.

- In SLR: $\beta_1 \rightarrow$ Marginal Slope.
  - Change in Y variable with one unit change in X variable.
- In MLR: $\beta_1 \rightarrow$ Partial Slope.
  - Change in Y variable with one unit change in X variable keeping all the other X variables constant.
- Ideally, we want explanatory variables which are orthogonal to each other, *i.e.*, which are independent of each other.
  - If the variables are truly independent of each other, then the marginal slope and the partial slope will have the same value. But it happens rarely.

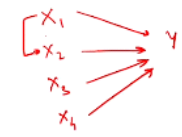
- What is a path diagram?
  - What is the total effect of X on Y?
  - It is represented by the _____ slope?
- Define collinearity

Interpreting Multiple Regression

- Path Diagram

- Path diagram: schematic drawing of the relationships among the explanatory variables and the response.

- Collinearity: very high correlations among the explanatory variables that make the estimates in multiple regression uninterpretable.

- Collinearity: When the explanatory variables are not independent, *i.e.*, when they are correlated. It is a situation that represents very high correlation amongst the explanatory variables. Sometimes it is so severe that it actually makes the estimates of MLR very difficult to interpret.
- Path Diagram:
- 

- We say that X_1 impacts Y in two different ways:
  - There's a direct effect of X_1 on Y, and
  - There's also an indirect effect of X_1 on Y, as it impacts X_2, because they are correlated, that in turn imapcts Y.
  - $\therefore$ $\boldsymbol{Total\ effect\ of\ X_1\ on\ Y\ =\ Direct\ effect + Indirect\ effect}$
- Total effect of X_1 on Y is represented in the Marginal Slope.

Slides that were not covered in lectures

- Errors in MRM satisfy what

Checking Conditions

conditions?

Checking Conditions

Conditions for Inference

- Use the residuals from the fitted MRM to check that the errors in the model

- are independent;
- have equal variance; and
- follow a normal distribution.

- How do you calculate F-Statistic in MRM?
- What null hypothesis do we assume in MRM?

Inference in Multiple Regression

- Inference for the Model: *F*-test

- The *F*-Statistic

$$F = \frac{R^2}{1 - R^2} \frac{n - k - 1}{k}$$

is used to test the null hypothesis that all slopes are equal to zero,

$$H_0: \beta_1 = \beta_2 = 0$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k}$$
$$= \frac{R^2}{1 - R^2} \cdot \frac{\text{Residual df}}{k}$$

- What is the purpose of t-statistic in MRM?
- How is t-statistic calculated?

Inference in Multiple Regression

Inference for One Coefficient

- The *t*-statistic is used to test each slope using the null hypothesis $H_0: \beta_j = 0$.

- The *t*-statistic is calculated as

$$t_j = \frac{b_j - 0}{se(b_j)}$$

$$t_j = \frac{b_j - 0 \text{ (the null hypthesis value)}}{se(b_j)}$$

- An approximate 95% prediction interval is given by?

Inference in Multiple Regression

- Prediction Intervals

- An approximate 95% prediction interval is given by $\hat{y} \pm 2s_e$.

- For example, the 95% prediction interval for price.

6.2 Multiple Linear Regression - Example

Thursday, 27 October 2022 20:54

Summary	• Implementing MLR in Excel

This is the data we're given

	Y	X1	X2
record	GPA at college	Entrance exam	interview
1	9.5	9.8	9.1
2	6.3	7.5	7.1
3	8.2	7.9	7.7
4	9.1	9.5	9.6
5	8.2	9.1	7.5
6	8.32	8.5	8.4
7	9.6	7.54	9.5
8	7.6	8.4	7.8
9	6.5	5.6	7.8
10	8.64	8	8.5
11	9.5	9.8	9.9
12	8.1	8	8.9
13	7.95	7.5	6.9
14	9.99	10	8.9
15	6.87	7.6	7.9

Y: GPA at college
$X1$: Entrance exam
$X2$: Interview

Response variable: Y
Explanatory variables: $X1, X2$

- We want to explain the variation in the response variable using these two explanatory variables.

Correlation Coefficients of Y with $X1, X2$:

	GPA at college
GPA at college	1
Entrance exam	0.74665952
interview	0.763282985

Individual influence of $X1$ (Entrance exam) on Y (GPA)

Y	X1
GPA at college	Entrance exam

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.74665952
R Square	0.557500439
Adjusted R Square	0.523462011
Standard Error	0.78574914
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10.11215109	10.11215109	16.37856019	0.00138404
Residual	13	8.026222245	0.617401711		
Total	14	18.13837333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.301862586	1.493803156	1.540941038	0.147313493	-0.925302932	5.529028103
Entrance exam	0.720234578	0.177965618	4.047043389	0.00138404	0.335763235	1.104705921

- Standard Error: s_e
- $Significance\ F$ values tells us that the regression is significant ($p - value\ <\ 0.05$).
- Coefficient of Entrance exam: $b_1 = 0.72\ \Rightarrow$ Estimate of β_1 (marginal slope).
- This SLR tells us that the entrance exam score is a good explanatory variable for Y (conclude

from the P-value of Entrance exam).

Individual influence of $X2$ (interview) on Y (GPA)

Y	X2
GPA at college	interview

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.763282985
R Square	0.582600915
Adjusted R Square	0.550493293
Standard Error	0.76313828
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	10.56743289	10.56743289	18.14525272	0.00093024
Residual	13	7.570940441	0.582380034		
Total	14	18.13837333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.470298787	1.846585182	0.254685672	0.802950357	-3.519005962	4.459603536
interview	0.934785006	0.219447292	4.259724488	0.00093024	0.460697953	1.408872058

- $Significance\ F$ values tells us that the regression is significant ($p - value\ <\ 0.05$).
- Coefficient of interview: $b_2 = 0.93\ \Rightarrow$ Estimate of β_2 (marginal slope).
- This SLR tells us that the interview score is a good explanatory variable for Y.

Multiple Linear Regression (MLR)

- Interpret:
  - Multiple R
  - R Square
  - Adjusted R Squared
- Regression df?
- Residual df?
- Total df?

Y	X1	X2
GPA at college	Entrance exam	interview

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.860528625
R Square	0.740509514
Adjusted R Square	0.6972611
Standard Error	0.626281041
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	13.43163802	6.715819012	17.12223502	0.000305301
Residual	12	4.70673531	0.392227942		
Total	14	18.13837333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.704401949	1.576544	-0.446801326	0.662975358	-4.139396243	2.730592345
Entrance exam	0.455442321	0.168539069	2.702295228	0.019227553	0.088227236	0.822657406
interview	0.62250322	0.213981085	2.909150685	0.013101725	0.156278487	1.088727953

- Multiple R: $R = 0.86\ \Rightarrow$ correlation coefficient between observed value(y) and the fitted/predicted value($\hat{y}$).
- R Square: explanatory power of this model.
  $R^2 = 0.74 \Rightarrow$ these two explanatory variables are able to explain 74% in the variation of GPA.
- $k = 2$ (two explanatory variables)
- $n = 15$.
- Adjusted R Square: $\bar{R}^2 = 0.69 \Rightarrow$ more realistic view of R^2. The explanatory power of the model is is actually 69% (and not 74%).

	df
Regression	2
Residual	12
Total	14

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

- We're estimating 3 parameters.
  $\therefore$ Regression $df = 3 - 1 = 2$.
- Residual $df = n - k - 1 = 15 - 2 - 1 = 12$.
- Total $df = n - 1 = 15 - 1 = 14$.

- $Significance\ F$ values tells us that the regression is significant ($p - value < 0.05$).

	P-value
Intercept	0.662975358
Entrance exam	0.019227553
interview	0.013101725

$p - value$ for Entrance Exam

H_0: $\beta_1 = 0$: null hypothesis
H_0: $\beta_1 \neq 0$: alternate hypothesis

$\because p - value = 0.019 < 0.05$
null hypothesis is rejected.

$\therefore\ b_1 = 0.455$ is a good estimate of β_1.

$p - value$ for interview

H_0: $\beta_2 = 0$: null hypothesis
H_0: $\beta_2 \neq 0$: alternate hypothesis

$\because p - value = 0.013 < 0.05$
null hypothesis is rejected.

$\therefore\ b_2 = 0.622$ is a good estimate of β_2.

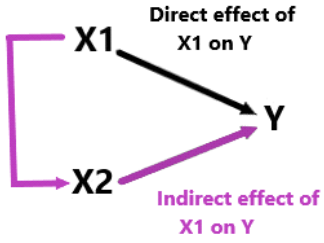
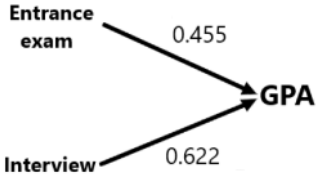
? But the $p - value$ for the intercept is $0.66 > 0.05$, so how are we $\beta_0 = -0.7$ estimating ?

- $\therefore$ Regression Equation:
  $Y = -0.7 + 0.455X_1 + 0.622X_2$.

- Here,0.455 and 0.622 are partial slopes.
  - *e.g.,* keeping constant X_2, one unit increase in X_1 is expected to increase Y by 0.455 units.

6.3 Multiple Linear Regression - Path Diagram

Monday, 07 November 2022 14:51

Summary	• Path Diagram
	From the correlation matrix *GPA at college* *Entrance exam* *interview* GPA at college 1 0.74665952 0.763283 Entrance exam 0.74665952 1 0.540056 interview 0.763282985 0.5400556 1 • We can conclude that the Entrance exam scores as well as the interview scores are highly correlated to the GPA. • Also, notice that the correlation between the Entrance exam and interview variables is 0.54, which also significant.
• Direct and Indirect Effect of X on Y	
• How do you calculate total effect of $X1$ and total effect of $X2$ on Y? • Observe how we form these SLR and MLR equations	*Coefficients* Intercept -0.704401949 Entrance exam 0.455442321 interview 0.62250322 We saw from the MLR analysis that the regression coefficients, $\beta_1 = 0.455$, and $\beta_2 = 0.644$. These are the partial slopes, *i.e.*, these are the direct effects on Y. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ $Y = \beta_0 + 0.455 X_1 + 0.622 X_2$ $GPA = \beta_0 + 0.455\ Entrance\ exam + 0.622\ interview$ 
	Analysing effect of $X1$ (Entrance exam) on Y (GPA)
	Indirect effect We will first analyse the indirect effect of Entrance Exam on GPA. For that we will analyse **effect of the Entrance exam on the interview.** So, we'll treat **Response variable: Interview** **Explanatory variable: Entrance Exam** X Y Entrance exam interview

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	0.811749825
Observations	15

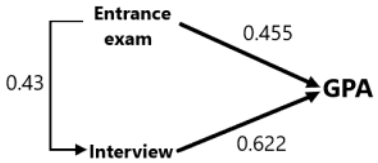
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.527142224	3.527142224	5.352769781	0.037694522
Residual	13	8.566191109	0.658937778		
Total	14	12.09333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.829315637	1.543233569	3.129348489	0.007982787	1.495362205	8.16326907
Entrance exam	0.425366887	0.183854556	2.313605364	0.037694522	0.028173267	0.822560506

$Y = \beta_0 + \beta_1 X$
$Y = \beta_0 + 0.42X$
$interview = \beta_0 + \mathbf{0.43}\ Entrance\ exam$

i.e., one unit change in entrance exam score changes the interview score by 0.42 units.



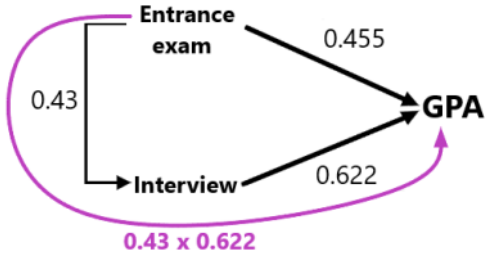
Total effect

$GPA = \beta_0 + 0.455 Entrance\ exam + 0.622 interview$
$interview = \beta_0 + \mathbf{0.43}\ Entrance\ exam$

∴ The indirect effect of Entrance exam on GPA = $0.43 \times 0.622 = 0.27$

Direct effect of Entrance exam on GPA = 0.455

Total effect = Direct effect + Indirect effect
∴ Total effect of Entrance exam on GPA = 0.455 + 0.27 = $\mathbf{0.725}$



This value is same as the marginal slope.

Y	X1
GPA at college	Entrance exam

	Coefficients
Intercept	2.301862586
Entrance exam	0.720234578

Analysing effect of $X2$ (interview) on Y (GPA)

Indirect effect

We will first analyse the indirect effect of interview on GPA.
For that we will analyse **effect of the interview on the Entrance exam.**
So, we'll treat
Response variable: Entrance exam
Explanatory variable: Interview

Y	X
Entrance exam	interview

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	1.030616281
Observations	15

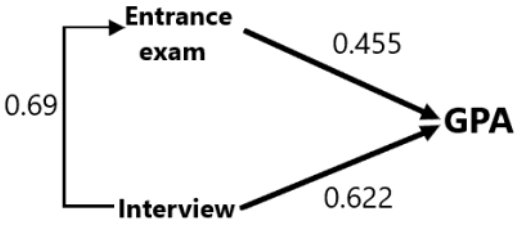
ANOVA

	df	SS	MS	F	Significance F
Regression	1	5.685551047	5.685551047	5.352769781	0.037694522
Residual	13	13.80820895	1.062169919		
Total	14	19.49376			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.579252481	2.493808533	1.034262433	0.319870981	-2.808293309	7.96679827
interview	0.685667034	0.296363003	2.313605364	0.037694522	0.045413691	1.325920377

$Y = \beta_0 + \beta_1 X$
$Y = \beta_0 + 0.69X$
$Entrance\ exam = \beta_0 + \mathbf{0.69}\ interview$

i.e., one unit change in entrance exam score changes the interview score by 0.69 units.



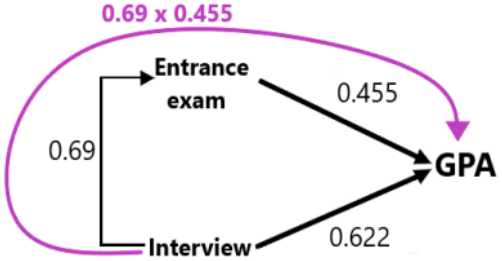
Total effect

$GPA = \beta_0 + 0.455 Entrance\ exam + 0.622 interview$
$Entrance\ exam = \beta_0 + \mathbf{0.69}\ interview$

∴ The indirect effect of interview on GPA
$= 0.69 \times 0.455 = 0.31$

Direct effect of interview on GPA = 0.622

Total effect = Direct effect + Indirect effect
∴ Total effect of interview on GPA = 0.622 + 0.31
= 0.932



This value is same as the marginal slope.

Y	X2
GPA at college	interview

	Coefficients
Intercept	0.470298787
interview	0.934785006

- What if the indirect effect was zero?

If indirect effect were zero, then
Partial slope = Marginal slope

6.4 Multiple Linear Regression - Variance Inflation Factor - Part 1

Monday, 07 November 2022 16:34

Summary	• Variance Inflation Factor

- Define VIF.
  - Formula
  - How do you calculate it?
- What is R_j^2? or How do you calculate it?
- If $R_j^2 \uparrow \Rightarrow$ VIF ?

Collinearity

- Variance Inflation Factor (VIF)
- Variance inflation factor: quantifies the amount of unique variation in each explanatory variable and measures the effect of collinearity.
- The VIF for X_j is $VIF(X_j) = \frac{1}{1-R_j^2}$

where R_j^2 is the Coefficient of Determination in the regression of X_j on ALL of the other explanatory variables.

- R_j^2 is the coefficient of determination on a regression where that particular j^{th} variable is the response variable and all the other explanatory variables are the explanatory variables.
- R_j^2 will be a large value if the regression is significant, that is, if the response variable is fairly correlated with the explanatory variables.
- If R_j^2 is a large value, then VIF will be a large value.
  $R_j^2 \uparrow \Rightarrow$ VIF $\uparrow$
  How?:

Let $R_j^2 = 0.9$	Let $R_j^2 = 0.1$
then $VIF(X_j) = \frac{1}{1-0.9} = \frac{1}{0.1} = 10$	then $VIF(X_j) = \frac{1}{1-0.1} = \frac{1}{0.9} = 1.11$

In MLR:

Y	X1	X2
GPA at college	Entrance exam	interview

	Standard Error
Intercept	1.576544
Entrance exam	0.168539069
interview	0.213981085

- Standard error in estimating b_1 is : $s_e(b_1) = 0.168$
- Standard error in estimating b_2 is : $s_e(b_2) = 0.213$

- Standard error in estimation of partial slopes, $se(b_1)$ =?
- What is s_e?
- What is s_{X_1}?
- With VIF, $se(b_1)$ =?

Why does VIF matter?

- The standard error in estimation of the partial slope gets inflated due to VIF.
- Typically,

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_x}$$

- With VIF

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_x} \times \sqrt{VIF(X_1)}$$

- s_e → Standard error, estimate of σ_ε
- s_{X_1} → Standard deviation in X_1

- If s_{X_1} is quite large, it helps us in understading the variation in Y.
- So if s_{X_1} is large, $se(b_1)$ will be small. Meaning we get high precision in estimating β_1.

- If the explanatory variables are uncorrelated, then VIF =?
- For correlated explanatory variables, VIF?
- Larger the VIF, larger the _____?
- What effect does VIF have on $se(b_1)$?

VIF

- If the explanatory variables are uncorrelated, then $R_j^2 = 0$, and $VIF = 1$.
- However, if the explanatory variables are correlated, then $VIF > 1$. Larger the VIF, larger is collinearity.
- Large VIF also substantially increases the standard error in predicting the partial slopes (*se(b)*). Thereby, making those predictions unreliable.

e.g., Take the example where we treated one explanatory variable as response variable and the other remained explanatory

X	Y
Entrance exam	interview
Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	0.811749825
Observations	15

Y	X
Entrance exam	interview
Regression Statistics	
Multiple R	0.5400556
R Square	0.291660052
Adjusted R Square	0.237172363
Standard Error	1.030616281
Observations	15

	b	R Square	VIF	VIF SQRT
Entrance Exam	0.455442	0.29166	1.411752	1.188172
Interview	0.622503	0.29166	1.411752	1.188172

- Here, since explanatory variables are correlated, $VIF > 1$.
- There is going to be an 18% of increase in $se(b)$.
- Note: Here, b values are taken from MLR, and the R Square is square of coefficient of correlation between both of the explanatory variable.

Inference in Multiple Regression

Inference for One Coefficient

- The *t*-statistic is used to test each slope using the null hypothesis H_0: $\beta_j = 0$.

- The *t*-statistic is calculated as

$$t_j = \frac{b_j - 0}{se(b_j)}$$

- In the example that we considered, $\sqrt{VIF}$ turned out to be very small.
- If $\sqrt{VIF}$ were high, it would inflate the standard errors, and then the t-stats would come down.
- If t-stats is very small, it will impact the $p - value$, and we may not be able to reject the null hypothesis.
- It will mean that that particular explanatory variable may be statistically insignificant for the regression.
- So, we want VIF value small, and it will happen only when the explanatory variables don't have too much correlation.

6.5 Multiple Linear Regression - Variance Inflation Factor - Part 2

Monday, 07 November 2022 18:57

Summary	• An example where explanatory variables on their own are significant (SLR), but when taken together become insignificant (MLR). • Signs and Remedies of Collinearity
	Interpreting Multiple Regression • Example: Estimating the price of a house ▪ Response variable: Price of the house (INR). ▪ Three explanatory variables: Size of the house (in square foot), number of bedrooms and the number of parking lots provided.
	Apartment data is given: • We're trying to predict the price of an apartment. • Price given in the data is in 10 Lakhs. • Area is given in 100 sq ft.
	Here's the correlation coefficient matrix: • Price seems to be affecting by all of the three factors.
	SLR on Area-Price
	Response Variable: Price Explanatory Variable: Area

Area (Sq ft)	# of bedrooms	Parking lot	Price
9.5	2	2	5.68
10	3	2	8.9
8.7	2	1	7.6
10	3	3	10
11.45	2	2	8
20	2	3	9.8
9	2	2	8.1
8.34	2	2	7.1
11	3	2	9.1
13	3	1	5
14.5	4	3	12
16	3	3	11.5
8.19	1	2	6.4
7.9	1	2	7
8.6	2	3	8
10	2	2	7.9
11.45	3	1	9
12	3	3	9.35
15	4	3	10.3
11	3	2	14

	Area (Sq ft)	*# of bedrooms*	*Parking lot*	*Price*
Area (Sq ft)	1			
# of bedrooms	0.503295389	1		
Parking lot	0.434453051	0.274318858	1	
Price	0.467910037	0.605720798	0.501392503	1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.467910037
R Square	0.218939803
Adjusted R Square	0.17554757
Standard Error	1.972955837
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19.64026979	19.64026979	5.045598878	0.037477004
Residual	18	70.06598521	3.892554734		
Total	19	89.706255			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.040157858	1.703678785	2.958396796	0.008412107	1.46086155	8.619454166
Area (Sq ft)	0.327646336	0.145864281	2.246241055	0.037477004	0.021196854	0.634095819

- $Significance\ F$ values tells us that the regression is significant ($p - value\ <\ 0.05$).

SLR on Number of Bedroom-Price

Response Variable: Price
Explanatory Variable: Number of Bedrooms

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.605720798
R Square	0.366897685
Adjusted R Square	0.331725334
Standard Error	1.776282599
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	32.91301731	32.91301731	10.43142345	0.004647911
Residual	18	56.79323769	3.155179872		
Total	19	89.706255			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.758615385	1.294091339	3.677186642	0.00172406	2.039830369	7.477400401
# of bedrooms	1.591153846	0.492652153	3.229771424	0.004647911	0.556130079	2.626177613

- $Significance\ F$ values tells us that the regression is significant ($p - value\ <\ 0.05$).

SLR on Parking lot-Price

Response Variable: Price
Explanatory Variable: Parking lot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.501392503
R Square	0.251394442
Adjusted R Square	0.209805244
Standard Error	1.931530785
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	22.55165391	22.55165391	6.04470526	0.024307596
Residual	18	67.15460109	3.730811171		
Total	19	89.706255			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.292065217	1.466039445	3.609770008	0.002003183	2.212030636	8.372099798
Parking lot	1.565652174	0.636806842	2.458598231	0.024307596	0.227770645	2.903533703

- $Significance\ F$ values tells us that the regression is significant ($p - value\ <\ 0.05$).

• So, by themselves, each explanatory variable does help us to predict the prices.

MLR

Response Variable: Price
Explanatory Variable: Area, Number of bedrooms, Parking lot

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702259246
R Square	0.493168049
Adjusted R Square	0.398137058
Standard Error	1.685711946
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	44.24025875	14.74675292	5.189549687	0.010764764
Residual	16	45.46599625	2.841624766		
Total	19	89.706255			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2.667893129	1.67836639	1.589577309	0.131492473	-0.890084674	6.225870932
Area (Sq ft)	0.059750485	0.154380203	0.387034628	0.70383001	-0.267520926	0.387021895
# of bedrooms	1.236821718	0.542442699	2.280096533	0.036650141	0.086894565	2.38674887
Parking lot	1.046580675	0.618622039	1.691793388	0.110065941	-0.264839463	2.358000813

- R^2 indicates that the fitted equation explains % of the variation in price.

• The overall $Significance\ F$ values of the MLR is significant ($p - value\ <\ 0.05$).

• But,

	P-value
Intercept	0.131492473
Area (Sq ft)	0.70383001
# of bedrooms	0.036650141
Parking lot	0.110065941

	Lower 95%	*Upper 95%*
Intercept	-0.890084674	6.225870932
Area (Sq ft)	-0.267520926	0.387021895
# of bedrooms	0.086894565	2.38674887
Parking lot	-0.264839463	2.358000813

$p - value$ for Area

H_0: $\beta_1 = 0$
H_0: $\beta_1 \neq 0$

$\because\ p - value = 0.7 > 0.05$
null hypothesis cannot be rejected.

Also, look at the confidence interval. b_1 can be anywhere between $[-0.26, 0.38]$. It includes 0, we cannot reject it.

$p - value$ for # of bedrooms

H_0: $\beta_2 = 0$
H_0: $\beta_2 \neq 0$

$\because\ p - value = 0.03 < 0.05$
null hypothesis is rejected.

$\therefore\ b_2 = 1.23$ is a good estimate of β_2.

$p - value$ for Parking lot

H_0: $\beta_3 = 0$
H_0: $\beta_3 \neq 0$

$\because\ p - value = 0.11 > 0.05$
null hypothesis cannot be rejected.

Also, the confidence interval for b_3 is $[-0.26, 2.35]$.

• Individually, area and parking lot were significant.

• It happened because there was a strong correlation among the explanatory variables.

	Area (Sq ft)	# of bedrooms	Parking lot	Price
Area (Sq ft)	1			
# of bedrooms	0.503295389	1		
Parking lot	0.434453051	0.274318858	1	
Price	0.467910037	0.605720798	0.501392503	1

- That's what's making partial slopes insignificant even when the marginal slopes were larger.

	Marginal Slope	Partial Slope
Area	0.327	0.059
# of bedroom	1.59	1.23
Parking lot	1.56	1.04

- Also the standard errors for the partial slopes are larger than the marginal slope.

	Marginal Slope	Partial Slope
Area	0.145	0.154
# of bedroom	0.12	0.54
Parking lot	0.63	0.628

VIF

- From our example, the explanatory variables in the MLR are turning to be insignificant.
- The explanatory variables aren't significant once we have taken account of the other explanatory variables.

- Partial slope conveys the unique variation explained by that particular explanatory variable.

- However once you take account of the parking lot and the area, now the number of bedrooms do not offer anything unique that has already not been explained by these two variables.
- Similarly, once you take account of the area and the number of bedrooms, now the parking lot does not offer anything unique that has not already been explained by these two variables.

- This is why the explanatory variables are turning out to be insignificant in MLR, but they are significant in the SLR.
- This is the impact of explanatory variables being correlated.

	b	R-Square	VIF	VIF_SQRT
Area	0.05975	0.348302	1.534452	1.23873
# of bedrooms	1.236822	0.257125	1.346122	1.160225
Parking lot	1.046581	0.192899	1.239002	1.113104

- What signs do these parameters show of collinearity:
  1. R^2
  2. Marginal and partial slopes
  3. F-statistic
  4. Standard errors for partial and marginal slopes
  5. VIF

Collinearity

Signs of Collinearity

- R^2 increases less than we'd expect.
- Slopes of correlated explanatory variables in the model change dramatically.
- The F-statistic is more impressive than individual t-statistics.
- Standard errors for partial slopes are larger than those for marginal slopes.
- Variance inflation factors increase.

- We know that whenever we add an explanatory variable, R^2 is supposed to go up.
- If there's a multi-collinearity, R^2 does not go up drastically, it goes up only fractionally.

- Value of Marginal slopes > Partial slopes
- Standard errors for Partial slopes > Marginal slopes

- What are some remedies for collinearity?

Collinearity

- Remedies for Collinearity
  - Remove redundant explanatory variables.
  - Re-express explanatory variables (e.g., use the average of *Market % Change* and *Dow % Change* as an explanatory variable).
  - Do nothing if the explanatory variables are significant with sensible estimates.

- Re-expressing explanatory variables mean we can combine the correlated explanatory variables to create another explanatory variable.

Removing Explanatory Variables

- Issues
  - After adding several explanatory variables to a model, some of those added and some of those originally present may not be statistically significant.
  - Remove those variables for which both statistics and substance indicate removal (e.g., remove *Dow % Change* rather than *Market % Change*).

6.6 Multiple Linear Regression - Implementation in Python

Monday, 07 November 2022 21:45

Summary	•
	Done in Colab

W6 Formulae

Wednesday, 16 November 2022 20:10

- SLR has one explanatory variable.
- MLR has multiple explanatory variables.

k → number of explanatory variables in MLR

Multiple Linear Regression Model

- Observed values of Y are linearly related to the k explanatory variables as:

$$\boldsymbol{y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,}$$

where, $\boldsymbol{\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)}$

- We makes 3 assumptions about the error term:

1. Independent
2. Equal variance, $Var(\epsilon) = \sigma_\epsilon^2$
3. Normal. $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$

- If you add explanatory variables to the model:
  - $\bar{R}^2 \uparrow$ and $s_e \downarrow$

- R
  - In SLR, R represents the correlation between X and Y.
  - In MLR, R represents the correlation between observed value(y) predicted value($\hat{y}$).

- Calibration plot: Scatterplot between y and $\hat{y}$.

Adjusted R Squared Formula

$$\boldsymbol{\bar{R}^2 = 1 - \left[\frac{(1 - R^2) \times (n - 1)}{(n - k - 1)}\right]}$$

Slopes

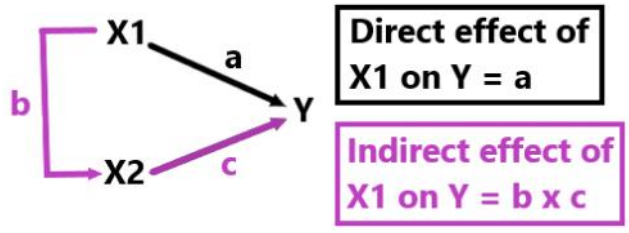
- In SLR: β_1 → Marginal Slope.
  - Change in Y variable with one unit change in X variable.

- In MLR: β_1 → Partial Slope.
  - Change in Y variable with one unit change in X variable keeping all the other X variables constant.

- In MLR, if variables are independent of each other, *i.e.*, correlation = 0, then
  Marginal slope = Partial slope

Path Diagrams

Schematic drawing of the relationships among various X's and Y.



Total effect = Direct effect + Indirect effect

$Y = \beta_0 + aX_1 + cX_2$
$X_2 = \tilde{\beta}_0 + bX_1$

Total effect of $X1$ on $Y = a + b \times c$

Total effect of X_i on Y is represented in the Marginal Slope.

VIF (Variance Inflation Factor)

quantifies the amount of unique variation in each explanatory variable and measures the effect of collinearity.

$$VIF\left(X_j\right) = \frac{1}{1 - R_j^2}$$

- $R_j^2 \uparrow \Rightarrow VIF \uparrow$
- R_j^2 is the coefficient of determination on a regression where that particular j^{th} variable is the response variable and all the other explanatory variables are the explanatory variables.
- If explanatory variables are uncorrelated: $R_j^2 = 0$ and $VIF = 1$
- If they are correlated: $VIF > 1$
- Larger the VIF larger the collinearity.

Typically,

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_{X_1}}$$

With VIF,

$$se(b_1) = \frac{s_e}{\sqrt{n}} \times \frac{1}{s_{X_1}} \times \sqrt{VIF\ (X_1)}$$

- s_e → Standard error, estimate of σ_ε
- s_{X_1} → Standard deviation in X_1
- As $VIF \uparrow \Rightarrow s_e\left(b_1\right) \uparrow$

Collinearity

Occurs when explanatory variables are highly correlated.

- Signs:
  - R^2 does not increase as much on adding explanatory variables.
  - Value of Marginal slopes > Partial slopes
  - Standard errors for Partial slopes > Marginal slopes

- VIF increases

- Remedies
  - Remove redundant explanatory variables
  - Re-express explanatory variables

Week 7

Tuesday, 08 November 2022 11:53

7.1 Logistic Regression - Predicting the Placements

Tuesday, 08 November 2022 12:00

Summary	• Problem Statement: Given the performance of a student (input variable), how likely is the student to get placed (output variable)?
	Will an MBA student get placed? Predicting categorical outcomes B-School: Business School
	Categorical predictions • Placement process B-Schools facilitates the graduates to pick up a job of their choice (amongst the available profiles). • The student attributes (academic performance, prior experience, internships) is expected to have a strong bearing on the outcome of the placement process. • The outcome variable in the example is binary – a student either gets a job or she doesn't! • However, the idea is more generic – the outcome variable could have several categorical values.
	Placement season: Data • Let us consider the following variables that may help explain placement chances of a student: ✓ Academic performance during the undergraduate degree. ✓ Academic performance during the MBA. ✓ Industry experience prior to joining the MBA program. ✓ Participation in the co-curricular and extra-curricular activities. • Let us assume that all these variables are scored on a scale of 10. • Lastly, our response variable is whether the student got placement or not (binary variable).
	Placement season: Data • See the data in the Excel sheet. **Disclaimer**: This is synthetic data. And it is not taken from any academic institute. This is for illustrative purposes only. This dataset should not be associated with any management program anywhere.
	This is the given data:

Student	MBA CGPA	Experience	UG CGPA	Extra-curricular	Day-0 placed
1	9.1	2.3	8.1	8.6	Placed
2	8.9	0	8.7	8.9	Placed
3	7	3.9	8	5.13	Not Placed
4	9.1	1.1	7.8	4.9	Not Placed
5	8.2	0.7	9.3	9.13	Placed
6	6.5	1.5	7.9	4.2	Not Placed
7	7.9	3.1	9.23	6.7	Not Placed
8	5.43	1.2	6.12	7.45	Not Placed
9	8.1	2.1	8.7	7.56	Placed
10	7.89	1.02	7.65	8.9	Placed
11	8.65	2.3	7.98	5.1	Placed
12	9.45	2.5	9.2	6.8	Placed
13	7.8	5.1	8.9	9.1	Placed
14	9.01	0.3	9.2	7.8	Placed
15	6.8	2.3	9.8	9.1	Not Placed
16	7.14	0.4	8.56	6.89	Not Placed
17	8.56	1.3	8.34	7.65	Not Placed
18	7.4	0	8.7	5.6	Not Placed
19	8.23	0	7.8	6.5	Not Placed
20	7.3	0	7.1	8.56	Not Placed
21	9.1	0	9.8	8.9	Placed
22	7.8	1.3	6.7	8.2	Not Placed
23	5.6	2.3	8.7	6.9	Not Placed
24	8.1	1.78	7.28	7.9	Not Placed
25	8.79	3.98	9.11	7	Placed
26	7.19	0.2	8.16	6.7	Not Placed
27	8.18	1.9	8.25	8.9	Not Placed

- We've coded the Day-0 placed column as:
  - Placed: 1
  - Not Placed: 0

Placement season: The Ask

- Given the data, what are we interested in?
- Of course, try and build a model that can help predict the chances of a student for whom the input data is available.
- Towards that, let us define the variables as:

X_1 = Academic performance during the undergraduate degree.

X_2 = Academic performance during the MBA.

X_3 = Industry experience prior to joining the MBA program.

X_4 = Participation in the co-curricular and extra-curricular activities.

$Y = 1$ if the student gets placed, and zero otherwise.

7.2 Logistic Regression - Working with data

Tuesday, 08 November 2022 12:56

Summary	• Logistic Regression
	X_1 = Academic performance during the undergraduate degree. X_2 = Academic performance during the MBA. X_3 = Industry experience prior to joining the MBA program. X_4 = Participation in the co-curricular and extra-curricular activities. $Y = 1$ if the student gets placed, and zero otherwise.
	Predicting the placements • Since the problem has been reduced to predicting the value of Y using X_1, X_2, X_3 and X_4, is this regression? • Can these attributes be used to predict whether a student will pick up a job during the placement process? • Answer is yes! Through "Logistic regression". • However, we need to pay attention to our response variable. • Since the response variable is binary (or generically speaking, categorical), we can't use the regular regression method and expression. • Logistic regression is used to predict the dependent categorical variable.
• How do we solve this problem? • $Odds$ (of success) = ?	**Solution method: Regression** • If this was modeled as a multiple linear regression, we would have $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$ • Since, our Y is binary, assumptions of the regression model won't hold and we won't get good predictions. • Can we try using probabilities? • That is: $Pr\{Y=1\}$ as a predictor. Then, our response variable has values between 0 and 1. • However, if we calculate ODDs, then we can get out of these limits. $$Odds(success\ in\ placements) = \frac{\Pr(Y=1)}{\Pr(Y=0)}$$ • If $P(Y=1) = 0.9$ and $P(Y=0) = 0.1$, then we say the odds of success is $9:1$.

- How do we use $Odds$ in regression equation?
- From there, how do you calculate $P(Y = 1)$?

Solution method: Regression

- More commonly, Log values are used. That is, Log of the odds.
- As a result, we have: (dropping the error term)

$$Log(Odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$Odds = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}$$

$$\Pr(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4}}$$

- Here, we're assuming that the log has a base of e.
- Let Odds $= e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4} = A$

$$\text{Odds} = \frac{\text{P(Y = 1)}}{\text{P(Y = 0)}} = \frac{\text{P(Y = 1)}}{1 - \text{P(Y = 1)}} = \text{A}$$

$$\Rightarrow \text{A} - \text{A P(Y = 1)} = \text{P(Y = 1)}$$

$$\Rightarrow \text{A} = \text{P(Y = 1)}(1 + \text{A})$$

$$\Rightarrow \text{P(Y = 1)} = \frac{\text{A}}{1 + \text{A}} = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2 + \beta_4 X_4}}$$

Solution method: Regression

- Now we can run the regression model and estimate the regression coefficients (the β's).
- The objective function used for this estimation: maximization of the log-likelihood. That is the log of probability of the correct prediction.
- See the Excel sheet.

This is the correlation matrix:

Correlation Matrix

	MBA CGPA	*Experience*	*UG CGPA*	*Extra-curricular*	*Day-0 placed*
MBA CGPA	1				
Experience	-0.038867107	1			
UG CGPA	0.348301526	0.170294352	1		
Extra-curricular	0.16846311	-0.070032269	0.176627455	1	
Day-0 placed	0.599259002	0.166540606	0.424267443	0.354241475	1

- We're interested in Correlation of the response variable (Day-0 placed) with the other explanatory variables.
- Also, if you notice the correlation coefficients among the explanatory variables, the correlations don't seem to be strong except between MBA CGPA and UG CGPA.

This is the working of Logistic Regression

b0	b1	b2	b3	b4								SUM of Log-Likelihood	-7.875726483	
-41.7512	3.2741492	0.5915958	0.83093	0.87624								Cutoff	0.5	
						Observed Y				Likelihood				Predicted Y
Student	MBA CGPA	Experience	UG CGPA	Extra-curricular		Day-0 placed	Logit - Log(odds)	Odds	Prob of Day-0 job	Prob of correct estimate		Log-Likelihood		Classification
1	9.1	2.3	8.1	8.6		1	3.67042506	39.26859	0.975166751	0.975166751		-0.025146796		1
2	8.9	0	8.7	8.9		1	2.41635488	11.20494	0.918065973	0.918065973		-0.085486025		1
3	7	3.9	8	5.13		0	-5.38238078	0.004597	0.00457583	0.99542417		-0.004586331		0
4	9.1	1.1	7.8	4.9		0	-0.5308569	0.588101	0.370317052	0.629682948		-0.462538843		0
5	8.2	0.7	9.3	9.13		1	1.2386607	3.450988	0.775330804	0.775330804		-0.254465497		1
6	6.5	1.5	7.9	4.2		0	-9.3372815	8.81E-05	8.80708E-05	0.999911929		-8.80747E-05		0

No idea how we got the *b* values.
I ran MLR on the data and the values I got are not matching with given values.

	Coefficients
Intercept	-3.058911577
MBA CGPA	0.241765018
Experience	0.06389908
UG CGPA	0.098114914
Extra-curricular	0.08658737

b0	b1	b2	b3	b4			
-41.7512	3.2741492	0.5915958	0.83093	0.87624			
Student	MBA CGPA	Experience	UG CGPA	Extra-curricular		Day-0 placed	Logit - Log(odds)
1	9.1	2.3	8.1	8.6		1	3.67042506
2	8.9	0	8.7	8.9		1	2.41635488
3	7	3.9	8	5.13		0	-5.38238078
4	9.1	1.1	7.8	4.9		0	-0.5308569
5	8.2	0.7	9.3	9.13		1	1.2386607

Logit-Log (odds)
$= b0 + (b1 \times MBA\ CGPA)$
$+ (b2 \times Experience)$
$+ (b3 \times UG\ CGPA)$
$+ (b4 \times Extracurricular)$

=$A$2+SUMPRODUCT($B$2:$E$2, B6:E6)

Basically this is what we're computing here:

$$Log(Odds) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

• Understand the working

Logit - Log(odds)	Odds
3.67042506	39.26859
2.41635488	11.20494
-5.38238078	0.004597
-0.5308569	0.588101

$$Odds = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}$$

• Odds = $e^{Logit-Log(odds)}$

Odds	Prob of Day-0 job
39.26859	0.975166751
11.20494	0.918065973
0.004597	0.00457583
0.588101	0.370317052

$$\Pr(Y=1) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4}}$$

$$\Pr(Y=1) = \frac{\text{Odds}}{1+\text{Odds}}$$

Day-0 placed	Prob of Day-0 job	Prob of correct estimate
1	0.975166751	0.975166751
1	0.918065973	0.918065973
0	0.00457583	0.99542417
0	0.370317052	0.629682948

Prob of correct estimate =
If Day-0 placed = 1
Prob of Day−0 job
Else
1 − Prob of Day−0 job

Likelihood	
Prob of correct estimate	Log-Likelihood
0.975166751	-0.025146796
0.918065973	-0.085486025
0.99542417	-0.004586331
0.629682948	-0.462538843

Log-likelihood = ln(Prob of correct estimate)

SUM of Log-Likelihood	-7.875726483
Cutoff	0.5

Then we sum up all the values in Log-Likelihood column − and that becomes our objective function.

The objective function used for this estimation: maximization of the log-likelihood. That is the log of probability of the correct prediction.

7.3 Logistic Regression - Model Building

Tuesday, 08 November 2022 17:12

Summary	• Calculations: going from probabilities (real numbers) to $0-1$ format.
• How do you go from probabilities to $0-1$ format?	**Regression: Calculations explained** • Notice that this regression will give us the "forecasted" probability that the student will be placed ($Pr\{Y=1\}$). • However, we want the value for our response variable (Y). And not the probability that Y will take on certain value. • Towards that, we define a threshold probability – if the forecasted probability value is above the threshold, we say that $Y = 1$. • On the other hand, if the forecasted probability is below the threshold, we can say that $Y = 0$ (the student won't be placed). • Now we have probabilities in hand, but we want to make the predictions in $0-1$ format. To do that, we set a threshold on probabilities.
	Cutoff 0.5 So, we defined an arbitrary cut-off at 0.5 Prob of Day-0 job / Predicted Y Classification: 0.975166751 → 1 0.918065973 → 1 0.00457583 → 0 0.370317052 → 0 0.775330804 → 1 If Prob of Day-0 job > 0.5 Predicted $Y = 1$ else Predicted $Y = 0$
	That lead us to a result that has some errors: Observed Y / Predicted Y 1 1 1 1 0 0 0 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 1 • This happened because we chose arbitrary value for the threshold.

7.4 Logistic Regression - Model Evaluation

Tuesday, 08 November 2022 17:39

Summary	• Measures to evaluate the performance of the model
	Evaluation of the Model How do we judge whether this logistic regression model is a good? • Typical statistical indicators: (generally based on the Log-likelihood) – deviance, R^2, and information criteria (Akaike, and Baye's). • Some of them have a threshold (often χ^2 based statistic) or sometime it is higher-the-better type (e.g. R^2). • Other performance indicators: (generally based on correct identification) – Accuracy, Precision, Recall. • Obviously, they are all large-the-better type performance indicators.
• What measures do we use in Logistic Regression to evaluate the performance of the model?	**Evaluation of the Model** • ***Accuracy***: Measure of the total number of predictions a model gets right, including both True Positives and True Negatives. • ***Recall:*** Indicates the percentage of the response values (that we are interested in) were actually captured by the model. • ***Precision***: Measures the percentage of the predicted response values (that we are interested in) that were correct.
• To understand the formulae better, see 7.6 page.	**Evaluation of the Model** For the student placement example, the response variable was binary (and we were interested in the chances of student getting placed, $Y = 1$). The performance measures can be interpreted as: • Accuracy: The ratio of the number of times predicted and actual Y values matched (for both $Y = 0$ and $Y = 1$) to the total observations in the sample. • Recall: The ratio of the number of times the prediction for Y was 1, to the total number of instances in the sample where Y was actually 1. • Precision: The ratio of the number of times the actual Y was 1, to the total number of instances where the prediction for Y was 1. • For better understanding, refer here. • We want these values as large as possible to conclude that the model is good.

	Predicted Y		Total
	0	1	
Actual Y 0	14	2	16
1	3	8	11
	17	10	27

At Cut off = 0.5, we get these values of the performance indicators:

Accuracy	81.48%
Recall	72.73%
Precision	80.00%

	Predicted Y		Total
	0	1	
Actual Y 0	14	2	16
1	3	8	11
	17	10	27

$$\text{Accuracy} = \frac{14+8}{27}$$

	Predicted Y		Total
	0	1	
Actual Y 0	14	2	16
1	3	8	11
	17	10	27

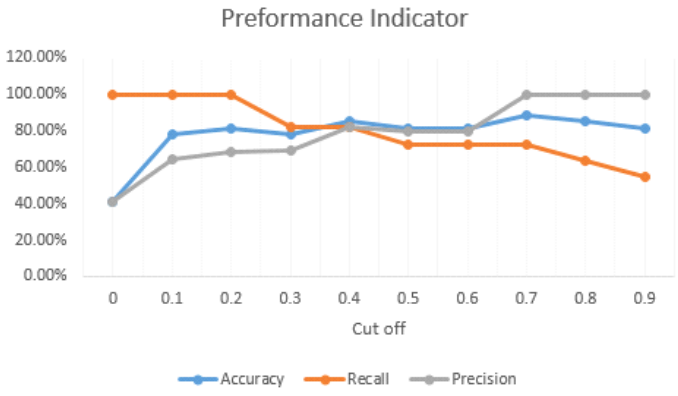
$$\text{Recall} = \frac{8}{11}$$

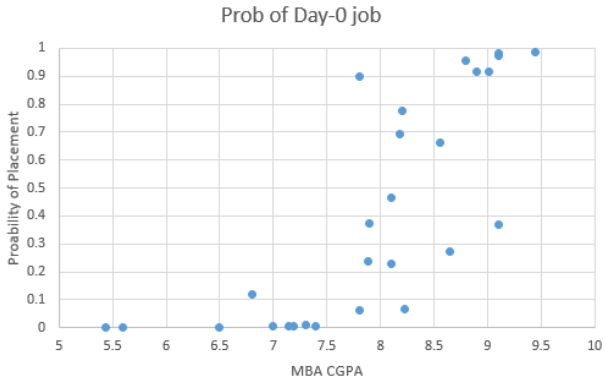
	Predicted Y		Total
	0	1	
Actual Y 0	14	2	16
1	3	8	11
	17	10	27

$$\text{Precision} = \frac{8}{10}$$

So, to predict the appropriate cut-off, we try different numbers and observe which cut off gives us the best values of the performance indicators

Cut off	Accuracy	Recall	Precision
0	40.74%	100.00%	40.74%
0.1	77.78%	100.00%	64.71%
0.2	81.48%	100.00%	68.75%
0.3	77.78%	81.82%	69.23%
0.4	85.19%	81.82%	81.82%
0.5	81.48%	72.73%	80.00%
0.6	81.48%	72.73%	80.00%
0.7	88.89%	72.73%	100.00%
0.8	85.19%	63.64%	100.00%
0.9	81.48%	54.55%	100.00%



	It seems that 0.4 is the best cut-off choice for the given data.
	Here probability vs. MBA CGPA graph is plotted.  It tells something I don’t have the energy to care about. Something about since $b1$ is some value that's why we're seeing this kind of graph.

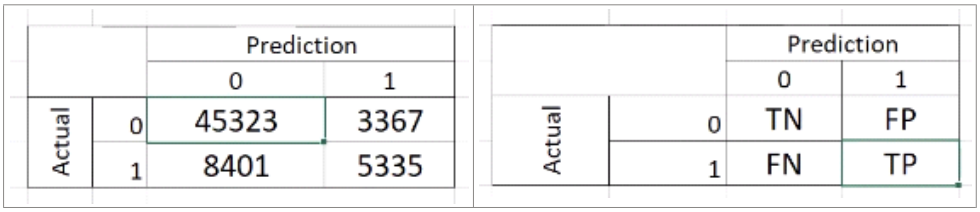
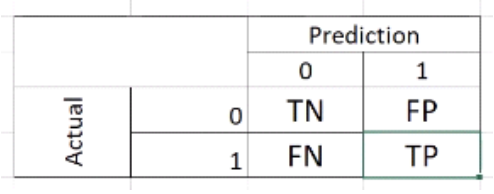
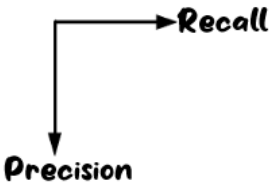
7.5 Logistic Regression - Interpretation of the Coefficients

Wednesday, 09 November 2022 8:58

Summary	• Interpretation of the coefficients in Logistic Regression
• How do you interpret the coefficients in logistic regression? • If an explanatory variable increases by 1 unit, the odds of $Y = 1$ increases by a factor of ?	**Interpretation of the coefficients** • In a multiple linear regression, the regression coefficients (the β's) are the change in the response variable, with a unit change in the corresponding explanatory variable, keeping all the other explanatory variables constant ("partial slopes"). • For Logistic regression, the interpretation is similar, except for the fact that the change is not linear but in terms of log of odds. • Here, a unit change in the explanatory variable brings about a change of β in the log-odds. • So, if the explanatory variable increases by 1 unit, the odds of Y = 1 increases by a factor of 10^{β}. (If we take the natural the, the odds increase by a factor of e^{β}.)
	Interpretation of the coefficients • For the student placement example, the regression coefficient for the explanatory variable MBA CGPA is $(\beta_1 = 3.27)$. • For one unit increase in the CGPA in the MBA program, the odds of the student getting placed increases by $e^{\beta_1} = e^{3.27} = 26.31$. • The probability of student getting placed has definitely increased. However, note that we have interpreted only the increase in odds and not in the actual probability.

7.6 Tutorial- Logistic Regression in Python

Friday, 11 November 2022 16:13

Summary	• Implementing Logistic Regression in Python • Understanding Accuracy, Precision and Recall better
	• The work is done in the Colab.
• Confusion matrix is always _____ matrix. • Define the confusion matrix in terms of TN, TP, FN, FP.	Confusion matrix is always a 2x2 matrix. The confusion matrix that we got in the data set: 
	• 45323: The number of times the prediction was 0 and the actual value was also 0: True Negative (TN). • 5335: The number of times the prediction was 1 and the actual value was also 1: True Positives (TP). • 3367: The number of times the prediction was 1 but the actual value was 0: False Positives (FP). • 8401: The number of times the prediction was 0 and the actual value was 1: False Negatives (FN).
• Define: 1. Accuracy 2. Precision 3. Recall	 $$\textbf{Accuracy} = \frac{\textbf{TN} + \textbf{TP}}{\textbf{Total number of observations}}$$ • We want to predict the number of 1's in our problems. • So, precision is out of all 1 predictions, how many were correct. • Precision for predicting 1: $$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FP}}$$ • Recall focuses on actuals. • Out of all the actual 1's, how many the model is predicting correctly. • Recall for predicting 1: $$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP} + \textbf{FN}}$$ • Tip to remember Precision and Recall: 

	Prediction	
Actual	0	1
0	45323	3367
1	8401	5335

	Prediction	
Actual	0	1
0	TN	FP
1	FN	TP

- Accuracy $= \frac{\text{TN} + \text{FP}}{\text{Total Number of observations}} = \frac{45323 + 5335}{4523 + 3367 + 8401 + 5335} = \frac{50658}{62426} = 0.81$
- Precision $= \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{5335}{5335 + 3367} = \frac{5335}{8702} = 0.61$
- Recall $= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{5335}{5335 + 8401} = \frac{5335}{13736} = 0.39$

```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.84      0.93      0.89     48690
           1       0.61      0.39      0.48     13736

    accuracy                           0.81     62426
   macro avg       0.73      0.66      0.68     62426
weighted avg       0.79      0.81      0.79     62426
```

- F1-score: weighted average between precision and recall.
- Precision and Recall are defined based on the problem that we're trying to address.
- Here, we're trying to address 1, so our precision = 0.61, and recall = 0.39.
- Whereas, the accuracy is always the same, which is 0.81 here.
- Support: number of the data points that we took for the calculations.

W7 Formulae

Thursday, 17 November 2022 22:11

Logistic Regression

Logistic regression is used to predict a dependent categorical variable.

$$\boldsymbol{Odds(success) = \frac{P(Y = 1)}{P(Y = 0)}}$$

In logistic regression, we use log of odds.

$$\boldsymbol{\log(Odds) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}$$

then

$$\boldsymbol{Odds = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k}}$$

$$\boldsymbol{P(Y = 1) = \frac{Odds}{1 + Odds}}$$

- Above expression will give us probabilities. But, we want to make predictions in $0 - 1$ (yes/no) format.
- So, we set a threshold (or cut-off) which is a number.

If $P(Y = 1) >$ cut-off:
$\boldsymbol{\hat{Y} = 1}$

Else:
$\boldsymbol{\hat{Y} = 0}$

- In logistic regression, if an explanatory variable increases by 1 unit, then the odds of $Y = 1$ increases by a factor of e^{β}.

Evaluation of logistic regression model

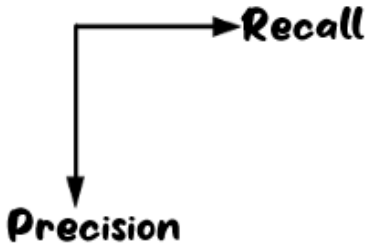
Confusion Matrix

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

$$\text{Accuracy} = \frac{TN + TP}{\text{Total number of observations}}$$

For predicting 1	For predicting 0
$\text{Precision} = \frac{TP}{TP + FP}$	$\text{Precision} = \frac{TN}{TN + FN}$
$\text{Recall} = \frac{TP}{TP + FN}$	$\text{Recall} = \frac{TN}{TN + FP}$

Tip to Remember:

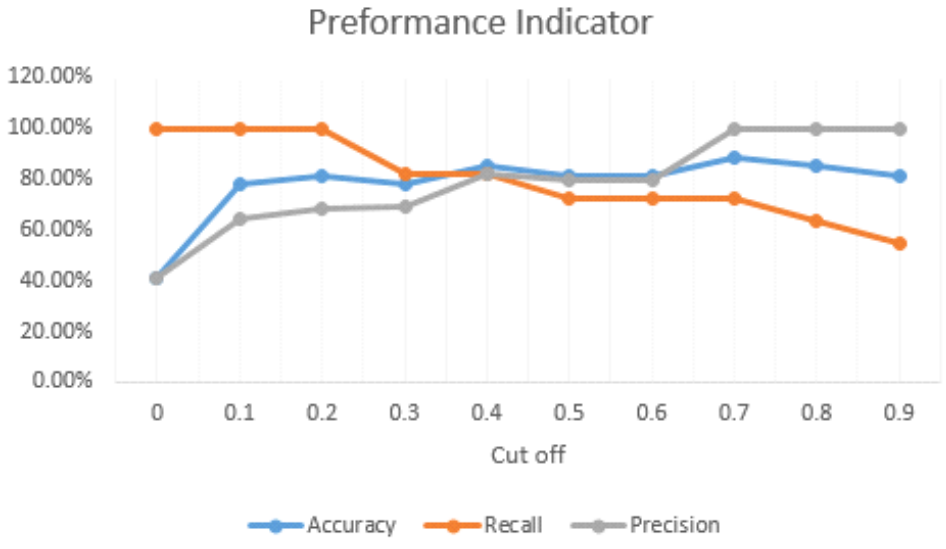


- The parameters are **larger the better** type.

How to find optimal cut-off?

At different cut off-values, we calculate these three performance indicators: accuracy, precision, recall, and then compare the results.

Example:



In the above analysis, optimal cut-off = 0.4

Week 8

Wednesday, 09 November 2022 9:01

8.1 Measuring the Efficiency of a Business Unit

Wednesday, 09 November 2022 9:48

Summary	• Efficiency of an Economic Unit
• Efficiency formula	**Efficiency** Quantifying and comparing the efficiencies of the decision making units • There are two conventional ways people define efficiency: i) Efficiency = $\frac{\text{Actual output}}{\text{Rated output}}$ ii) Efficiency = $\frac{\text{Output}}{\text{Input}}$ • We will be using the (ii) definition in this course.
• What is productive efficiency "frontier"? • The organisations that ○ are, and ○ are not on the frontier are called?	**Productive efficiency (Production efficiency)** • Economics teaches us effective utilization of resources for the maximization of benefits (output). • Productive efficiency is an aspect of economic efficiency focusing on maximizing the output under given constraints (without worrying about optimal allocation, or choice of products, etc.). • The productive efficiency "frontier" are all the combinations of outputs such that the production of one product cannot be increased without sacrificing the output of the other (without any change in the technology). • If the organization (any economic unit) is not on the frontier, it is inefficient. • Frontier is that optimal combination of outputs where you cannot increase output 1 without affecting output 2. • The organisations that run on this frontier are called efficient organisations.
	Efficiency measurement • In the simplest way, efficiency is defined as the ratio of the output to the input.  • However, in reality, this is complex. Why?

- Why efficiency measurement is complex?

Efficiency measurement

- Multiple types of inputs: Labor (white collar, blue collar, etc.); Infrastructure (factory, buildings, land, machinery, etc.); Money (financial assets, loan, etc.)
- Essentially, resources goes as inputs.

- Output can also be in many shapes and forms: customers served/acquired; profits; sales volume, etc.
- How has the organization (or the economic unit) performed is the output.

Immediate questions

- How does the ratio of input to output work in presence of several inputs and several outputs?
- How do we, then, calculate the productive efficiency of an economic unit?
- More importantly, how do we compare several economic units on their efficiency?
- If an economic unit turns out to be inefficient, how can they become efficient?

Common approaches

- Operating ratios: Labor cost per transaction; sales per square feet; runs per innings.
- Problem: Doesn't reflect varying mix of inputs and outputs found in more diverse operations.

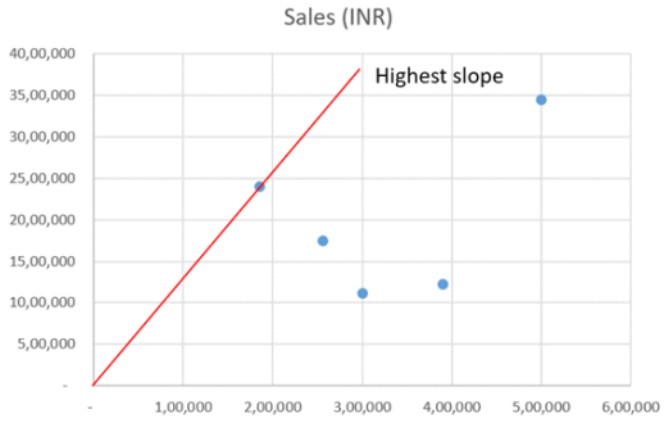
- Financial ratios: Price to earnings ratio (PE); Debt to equity ratio; Earnings per share (EPS).
- Problems?
  - Some inputs/outputs cannot be valued in currency terms.
  - Profitability is not the same as operating efficiency.

8.2 Efficiency Comparison - Graphical Method

Sunday, 02 October 2022 12:15

Summary	• Efficiency comparison using graphical method
	Graphical method Efficiency comparison
	When things are simpler… • When we only have a single input and a single output, a simple ratio of the input to the output is the efficiency. • The economic unit with the highest ratio is the most efficient. • Other economic units need to either increase the output for the same level of input; or reduce the input to achieve the same level of output. • See the data and the graph……
	Single input, single output • Sales: output variable • Budget: input variable • ∴ Efficiency = Sales per INR invested = $\frac{\text{Output}}{\text{Input}} = \frac{\text{Sales}}{\text{Budget}}$ • Sales office 5 has the highest efficiency.
• For a single input, single output, how do you decide the most efficient EU on graph?	**Single input, single output**

Sales office	Budget (INR)	Sales (INR)	Sales per INR invested
1	3,00,000	11,10,000	3.7
2	2,56,000	17,50,000	6.8
3	5,00,000	34,50,000	6.9
4	3,90,000	12,24,000	3.1
5	1,85,000	24,00,000	13.0



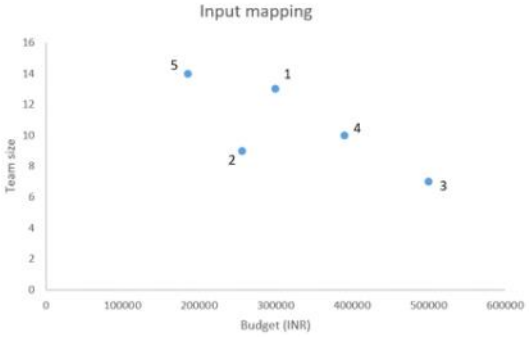
- Sales: y-axis
  Budget: x-axis
- You plot the points and then for every point you draw a line from origin to that point.
- For a single input, single output, the economic unit with the line with the highest slope is considered to be the most efficient.

- For two inputs and a constant output, how do you decide EUs?
- For inputs the frontiers are drawn on the _____ side.

More inputs/outputs

- For two inputs and an output too, things are not difficult.
- Assume that each of the sales office has the same sales target: INR 10,00,000 (output). They have their budgets approved and the respective team sizes (inputs).

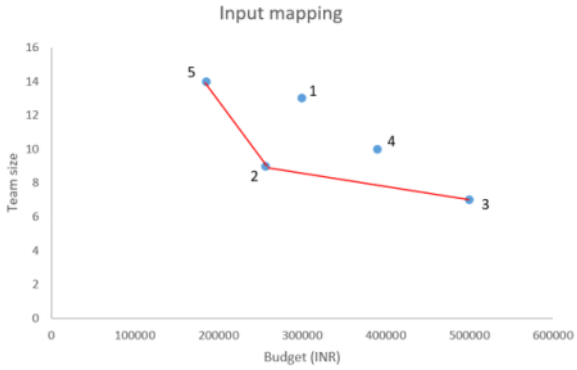
Sales office	Budget (INR)	Team size
1	3,00,000	13
2	2,56,000	9
3	5,00,000	7
4	3,90,000	10
5	1,85,000	14



- In this example we have two inputs: Budget and Team size
- The output (sales target) is same for all the sales offices.
- Here, we're only plotting the inputs.
- The EU's (Economic units) that consume less resources (inputs) are considered efficient.
- 5, 2 and 3 are efficient.

- Observe how the frontier is drawn for the inputs
- _____ the inputs the better.

Two inputs, single output: Efficiency frontier



Sales office	Budget (INR)	Team size
1	3,00,000	13
2	2,56,000	9
3	5,00,000	7
4	3,90,000	10
5	1,85,000	14

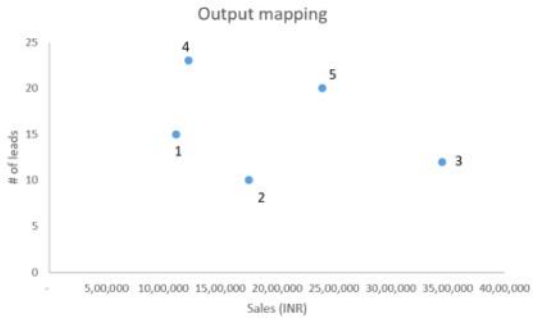
- This is how the efficiency frontier(envelope) would look like.
- For inputs, we draw the frontier on the lower side.
- Lesser the inputs (resources) the better.

- How do you decided EUs for two outputs and a constant input for all?
- For outputs the frontiers are drawn on the _____ side.

One input, two outputs

- Let every sales office be given the same budget (INR 2,00,000). The sales achieved (in INR) and the potential sales leads (potential customers) are the outputs we track.

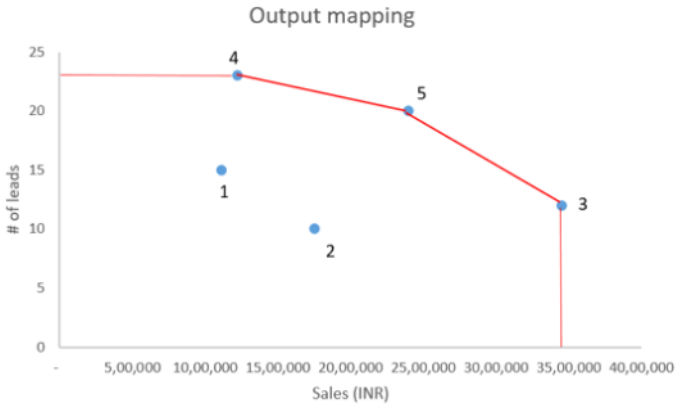
Sales office	Sales (INR)	No of leads
1	11,10,000	15
2	17,50,000	10
3	34,50,000	12
4	12,24,000	23
5	24,00,000	20



- Here, the input (budget) is same for all sales offices.
- And we have two outputs: Sales, No. of leads
- We're only plotting the outputs here.
- Here, 4, 5 and 3 are considered efficient.

- Observe how the frontier is drawn for the outputs.
- _____ the outputs the better.

One input, two outputs: Efficiency Frontier



- For outputs we draw the frontier on the outer side.
- More the outputs the better.

- We looked at two dimensional examples in this lecture. Same analogy will apply to higher dimensions as well.

8.3 Optimization Method - Data Envelopment Analysis

Wednesday, 09 November 2022 12:55

Summary	• Optimization Method - Data Envelopment Analysis
	Optimization method Data Envelopment Analysis
• Full form of DEA. • DEA is used for? • In DEA, an EU is called as?	**Data Envelopment Analysis (DEA)** • A non-parametric mathematical method to find the production frontier. • Can be used to calculate the productive efficiency of an economic unit. • In the DEA terminology, an economic unit is referred to as a "Decision Making Unit (DMU)". • What DEA measures is actually a relative efficiency – Calculate individual efficiencies for each DMU in a set of DMU's. • Formulates an optimization problem for each DMU.
• We solve optimization problem to find ? • What are the constraints we apply on weights?	**DEA logic** • For multiple inputs and multiple outputs, define the weighted ratio. • Since various inputs can't be directly added, define weights for each input. Similar approach for output. • But we don't know the value of each weight? • Solution: Let each DMU choose the input and the output weights to its advantage. • Objective for each DMU: maximize its efficiency by choosing its weights carefully. • Constraints: Choose the weights such that using these weights, they shouldn't get an efficiency more than 1! • You invested 1L rupees in your company and you hired 10 number of workers. You can't directly add 1L and 10. That's why you define weights. • **We solve optimization problem to find optimal weights.** • We solve for each DMU independent of the others in a way that the efficiency for a given DMU is maximized. • So, each DMU will have different weights. • Constraint: Choose weights as such the efficiency does not get more than one. • Also, let's say that sales office 1 has solved the optimization problem and obtained weights s.t. the efficiency is within the constraint. Now, using the weights of sales office 1, the other sales offices should not report an efficiency more than 1. • **That is, your weights should** ○ **not make your efficiency more than 1, and** ○ **not make other's efficiencies more than 1 when they use your weights.** • Given weights, none of the DMU's should get an efficiency more than 1.

- Given weights, how do you decide efficient and inefficient DMUs?

DEA logic

- Using a DMU's weights, if it can't achieve an efficiency of 1, then it is truly inefficient.
- Using a DMU's weight, if an other DMU gets an efficiency of 1, then that other DMU is really good!
- Note that we are referring to only relative efficiency.

- Variables in DEA:
  - $K = ?$
  - $N = ?$
  - $M = ?$
  - $I_{ik} = ?$
  - $O_{jk} = ?$
  - $x_{ik} = ?$
  - $y_{jk} = ?$
  - $E_k = ?$

DEA – Mathematical formulation

- K = # of DMU's considered in the dataset.
- N = # of inputs considered for the DMU's
- M = # of outputs considered for the DMU's.
- I_{ik} = Recorded value of input i for the DMU k. (i = 1, 2, … N, k = 1, 2, …, K).
- O_{jk} = Recorded value of output j for the DMU k. (j = 1, 2, … M, k = 1, 2, …, K).
- x_{ik} = Weight assigned to input i by the DMU k. (i = 1, 2, … N, k = 1, 2, …, K).
- y_{jk} = Weight assigned to output j by the DMU k. (j = 1, 2, … M, k = 1, 2, …, K).
- E_k = Efficiency of the DMU k. (k = 1, 2, …, K)

Example:

Inputs

Sales office	Budget (INR)	Team size
1	3,00,000	13
2	2,56,000	9
3	5,00,000	7
4	3,90,000	10
5	1,85,000	14

- Input #2 for sales office 3, $I_{23} = 7$
- Input #1 for sales office 4, $I_{14} = 3{,}90{,}000$

Corresponding weights will be:
- x_{23}
- x_{14}

Outputs

Sales office	Sales (INR)	No of leads
1	11,10,000	15
2	17,50,000	10
3	34,50,000	12
4	12,24,000	23
5	24,00,000	20

- Output #1 for sales office 5, $O_{15} = 24{,}00{,}000$
- Output #2 for sales office 1, $O_{21} = 15$

- y_{15}
- y_{21}

- Efficiency = ?
- For a particular DMU k,
  - $E_k = ?$

DEA – Mathematical formulation

- Efficiency is defined as a ratio of weighted outputs to the weighted inputs.

$$Efficiency = \frac{Weighted\ Output}{Weighted\ Input}$$

- Each DMU defines its own efficiency using the weights they want to assign to their inputs and outputs.
- For a particular DMU k, the efficiency is:

$$E_k = \frac{y_{1k}O_{1k}+y_{2k}O_{2k}+y_{3k}O_{3k}\ldots+y_{Mk}O_{Mk}}{x_{1k}I_{1k}+x_{2k}I_{2k}+x_{3k}I_{3k}\ldots+x_{Nk}I_{Nk}}$$

- Each DMU tries to maximize ?
- The optimization problem for

DEA – Optimization problem

each DMU k is ?
- Decision variables are ?

- Each DMU tries to maximize their own efficiency by adjusting the weights assigned to the inputs and the outputs.
- Only constraint on this: using these weights, none of the DMU's should get an efficiency more than 1!
- For each DMU k, the optimization problem is:

$$Max\ E_k$$

$$subject\ to\ \ E_k \leq 1, \qquad k = 1, 2, \ldots K$$

$$Decision\ variables:\ x_{ik}, y_{jk} \geq 0, \forall i, \forall j.$$

- Issues with DEA Optimization problem given above.

DEA – Optimization problem

Complexities

- Remember that the efficiency was defined as a ratio of weighted inputs to weighted outputs.
- Weights are the decision variables.
- Hence the objective function and the constraints of this optimization problem are ratios of decision variables.
- That is, they are NON LINEAR!

- Is there a way to linearize the problem?
- A linear optimization problem is much easier to solve, of course.

- How do you linearize DEA optimization problem?

DEA – Optimization problem

- To linearize, we maximize the numerator of the efficiency equation for the DMU k.
- And normalize the denominator to 1.
- For the constraints, we rearrange the efficiency terms to make it linear.
- So the revised formulation is:

$$\text{Efficiency} = \frac{\text{Weighted output}}{\text{Weighted input}} \leq 1$$

To linearize:

$$\text{Weighted output} \leq \text{Weighted input}$$

- Write DEA optimization problem after linearization.

DEA – Optimization problem

- For a DMU k,

$$Max\ y_{1k}O_{1k} + y_{2k}O_{2k} + y_{3k}O_{3k} \ldots + y_{Mk}O_{Mk}$$

$$subject\ to$$

$$x_{1k}I_{1k} + x_{2k}I_{2k} + x_{3k}I_{3k} \ldots + x_{Nk}I_{Nk} = 1$$

$$y_{1k}O_{11} + y_{2k}O_{21} \ldots + y_{Mk}O_{M1} \leq x_{1k}I_{11} + x_{2k}I_{21} \ldots + x_{Nk}I_{N1}$$

$$y_{1k}O_{12} + y_{2k}O_{22} \ldots + y_{Mk}O_{M2} \leq x_{1k}I_{12} + x_{2k}I_{22} \ldots + x_{Nk}I_{N2}$$

$$\vdots$$

$$y_{1k}O_{1k} + y_{2k}O_{2K} \ldots + y_{Mk}O_{MK} \leq x_{1k}I_{1K} + x_{2k}I_{2K} \ldots + x_{Nk}I_{NK}$$

$$Decision\ variables:\ x_{ik}, y_{jk} \geq 0, \forall i, \forall j.$$

- We want weights that keep efficiencies of all DMU $\leq$ 1

- Output weights for DMU k : $y_{1k}, y_{2k}, \dots, y_{Mk}$
- Input weights for DMU k : $x_{1k}, x_{2k}, \dots, x_{Nk}$

Objective: Max Weighted Output of DMU k
subject to

Weighted Input of DMU k = 1,

and **using the weights of DMU *k***:

DMU1 weighted output $\leq$ DMU1 weighted input
DMU2 weighted output $\leq$ DMU2 weighted input
⋮
DMUK weighted output $\leq$ DMUK weighted input

Decision variables: $x_{ik}, y_{jk} \geq 0, \quad \forall i, \forall j$

8.4 Data Envelopment Analysis - Example with one output and two inputs

Wednesday, 09 November 2022 16:06

Summary	• DEA Example - One output and two inputs

• Formulate DEA – LP optimization problem for each DMU

DEA – Linear programming

- Let us revisit the sales example where we had one output and two inputs.
- Each sales office has the same sales target: INR 10,00,000 (output). They have their budgets approved and the respective team sizes (inputs).

Sales office	Budget (INR)	Team size
1	3,00,000	13
2	2,56,000	9
3	5,00,000	7
4	3,90,000	10
5	1,85,000	14

- Let us formulate linear programs to calculate the efficiency of each sales office.

DEA – Linear programming

- Notice that we will have to formulate an optimization for each of the sales offices independently.

Let us start with office # 1. For this office,

- The only Output $O_{jk} = O_{11} = 10{,}00{,}000$.
- The two inputs: $(I_{ik})\ Budget,\ I_{11} = 3{,}00{,}000;\ Team\ size,\ I_{21} = 13$.
- We need one output weight (y_{11}), and two input weights (x_{11}, x_{21}).
- We expect the optimization problem to tell us the optimal values of the weights.

DEA – LP: Sales office 1

$$Max\ y_{11} * 1000000$$

$$subject\ to\ \ x_{11} * 300000 + x_{21} * 13 = 1$$

$$y_{11} * 1000000 \leq x_{11} * 300000 + x_{21} * 13$$

$$y_{11} * 1000000 \leq x_{11} * 256000 + x_{21} * 9$$

$$y_{11} * 1000000 \leq x_{11} * 500000 + x_{21} * 7$$

$$y_{11} * 1000000 \leq x_{11} * 390000 + x_{21} * 10$$

$$y_{11} * 1000000 \leq x_{11} * 185000 + x_{21} * 14$$

$$Decision\ variables:\ x_{11}, x_{21}, y_{11} \geq 0$$

DEA – LP: Sales office 2

$$Max\ y_{12} * 1000000$$

$$subject\ to\ \ x_{12} * 256000 + x_{22} * 9 = 1$$
$$y_{12} * 1000000 \leq x_{12} * 300000 + x_{22} * 13$$
$$y_{12} * 1000000 \leq x_{12} * 256000 + x_{22} * 9$$
$$y_{12} * 1000000 \leq x_{12} * 500000 + x_{22} * 7$$
$$y_{12} * 1000000 \leq x_{12} * 390000 + x_{22} * 10$$
$$y_{12} * 1000000 \leq x_{12} * 185000 + x_{22} * 14$$

$$Decision\ variables: x_{12}, x_{22}, y_{12} \geq 0$$

DEA – LP: Sales office 3

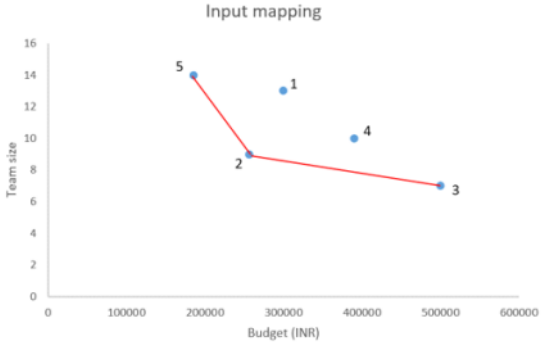
$$Max\ y_{13} * 1000000$$

$$subject\ to\ \ x_{13} * 500000 + x_{23} * 7 = 1$$
$$y_{13} * 1000000 \leq x_{13} * 300000 + x_{23} * 13$$
$$y_{13} * 1000000 \leq x_{13} * 256000 + x_{23} * 9$$
$$y_{13} * 1000000 \leq x_{13} * 500000 + x_{23} * 7$$
$$y_{13} * 1000000 \leq x_{13} * 390000 + x_{23} * 10$$
$$y_{13} * 1000000 \leq x_{13} * 185000 + x_{23} * 14$$

$$Decision\ variables: x_{13}, x_{23}, y_{13} \geq 0$$

DEA LP – Sales offices

- Similarly, the LP can be formulated for the other offices.
- See Excel sheet for the solution of the LP. Did we have the same solution in the graphical method?



Working has been done in the Excel sheet.

Results:

Sales office k	inp1 wt x1k	inp2 wt x2k	out wt y1k	Efficien
1	2.06E-06	0.029302518	7.91993E-07	0.79
2	2.61E-06	0.036998437	0.000001	1
3	7.39E-07	0.090103397	0.000001	1
4	6.21E-07	0.075776398	8.40994E-07	0.84
5	2.61E-06	0.036998437	0.000001	1

W8 Formulae

Thursday, 17 November 2022 23:01

• The organisations that run on the productive efficiency "frontier" are called efficient organisations. • Frontier is that optimal combination of outputs where you cannot increase output 1 without sacrificing output 2.
(not completed yet)

Week 9

Thursday, 01 December 2022 19:50

9.1 Data Envelopment Analysis - Example with two outputs and one input

Thursday, 01 December 2022 20:10

Summary	• DEA Example - Two outputs and one input

DEA – LP: Two output case

- Remember the sales problem where we had two output and a single input?
- Every sales office be given the same budget (INR 2,00,000). The sales achieved (in INR) and the potential sales leads (potential customers) are the outputs.

Sales office	Sales (INR)	No of leads
1	11,10,000	15
2	17,50,000	10
3	34,50,000	12
4	12,24,000	23
5	24,00,000	20

DEA – LP: Two output case

- Like before, we need to formulate an optimization problem for each of the sales office.
- Here, the only input is (I_{ik}) is the budget to run the office.
- The outputs are (O_{jk}): O_{1k} is the sales achieved; and O_{2k} is the potential leads.
- Corresponding to this, we would need one input weight (x_{1k}) and two output weights (y_{1k}, y_{2k}).

DEA – LP: Two output case – Sales office 1

$$Max\ y_{11} * 1110000 + y_{21} * 15$$

$$subject\ to\ \ x_{11} * 200000 = 1$$

$$y_{11} * 1110000 + y_{21} * 15 \leq x_{11} * 200000$$

$$y_{11} * 1750000 + y_{21} * 10 \leq x_{11} * 200000$$

$$y_{11} * 3450000 + y_{21} * 12 \leq x_{11} * 200000$$

$$y_{11} * 1224000 + y_{21} * 23 \leq x_{11} * 200000$$

$$y_{11} * 2400000 + y_{21} * 20 \leq x_{11} * 200000$$

$$Decision\ variables:\ x_{11}, y_{11}, y_{21} \geq 0$$

DEA – LP: Two output case – Sales office 2

$$Max\ y_{12} * 1750000 + y_{22} * 10$$

$$subject\ to\ \ x_{12} * 200000 = 1$$

$$y_{12} * 1110000 + y_{22} * 15 \leq x_{11} * 200000$$

$$y_{12} * 1750000 + y_{22} * 10 \leq x_{11} * 200000$$

$$y_{12} * 3450000 + y_{22} * 12 \leq x_{11} * 200000$$

$$y_{12} * 1224000 + y_{22} * 23 \leq x_{11} * 200000$$

$$y_{12} * 2400000 + y_{22} * 20 \leq x_{11} * 200000$$

$$Decision\ variables: x_{12}, y_{12}, y_{22} \geq 0$$

DEA – LP: Two output case

- Similarly, LP can be formulated for each of the sale office.
- If the objective function for the LP of a particular sales office is 1, then the sales office is efficient. Else, it is not.
- The results should match our graphical output (even that graphical output is DEA).
- See Excel sheet for the solutions.

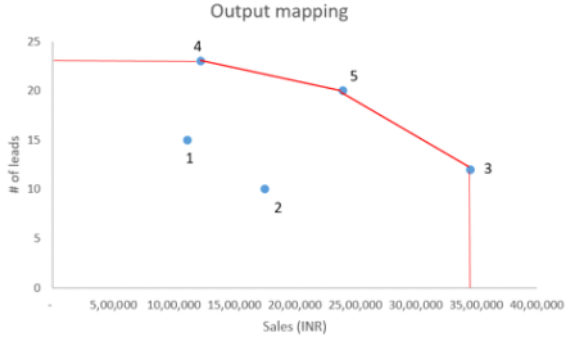
From Excel:

Results:

Sales office k	inp1 wt x1k	out1 wt y1k	out2 wt y2k	Efficiency
1	0.000005	9.76563E-08	0.03828125	0.68
2	0.000005	1.99005E-07	0.026119403	0.61
3	0.000005	2.89855E-07	0	1
4	0.000005	0	0.043478261	1
5	0.000005	1.99005E-07	0.026119403	1

DEA – LP: Two output case

- Graphical method had...



- From the graph as well as from the excel working, we can conclude that sales office 3, 4 and 5 are efficient DMUs.

9.2 Data Envelopment Analysis - Example with multiple outputs and multiple inputs

Friday, 02 December 2022 9:26

Summary	• DEA Example - multiple outputs and multiple inputs

DEA: LP – a Generic formulation

- Let us consider a case of multiple inputs and multiple outputs.
- For our comparison of the performance of the various sales offices, let us consider the full data.

	Inputs		Outputs	
Sales office	Budget (INR)	Team size	Sales (INR)	No of leads
1	3,00,000	13	11,10,000	15
2	2,56,000	9	17,50,000	10
3	5,00,000	7	34,50,000	12
4	3,90,000	10	12,24,000	23
5	1,85,000	14	24,00,000	20

- Plotting this data is not possible. In previous examples, we plotted only either the input or the output neglecting the other.

DEA: LP – a Generic formulation

- Two inputs and two outputs – need to define the variables accordingly.
- Difficulty to imagine this on a 2D plot.
- Optimization problem for each sales office ...
- Formulation and solution in Excel...

DEA: LP – Sales office 1

$$Max\ y_{11} * 1110000 + y_{21} * 15$$

$$\begin{aligned} subject\ to \quad & x_{11} * 300000 + x_{21} * 13 = 1 \\ & y_{11} * 1110000 + y_{21} * 15 \le x_{11} * 300000 + x_{21} * 13 \\ & y_{11} * 1750000 + y_{21} * 10 \le x_{11} * 256000 + x_{21} * 9 \\ & y_{11} * 3450000 + y_{21} * 12 \le x_{11} * 500000 + x_{21} * 7 \\ & y_{11} * 1224000 + y_{21} * 23 \le x_{11} * 390000 + x_{21} * 10 \\ & y_{11} * 2400000 + y_{21} * 20 \le x_{11} * 185000 + x_{21} * 14 \end{aligned}$$

$$Decision\ variables:\ x_{11}, x_{21}, y_{11}, y_{21} \ge 0$$

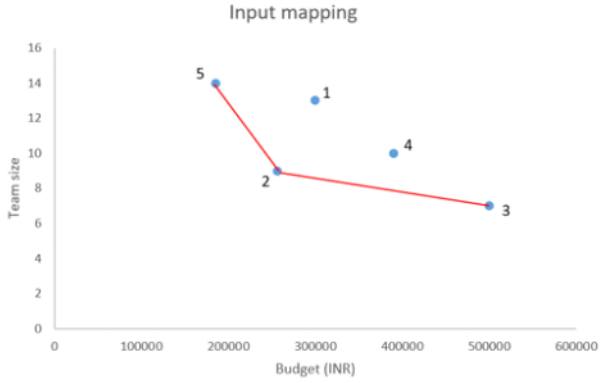
Excel Result:

Results:

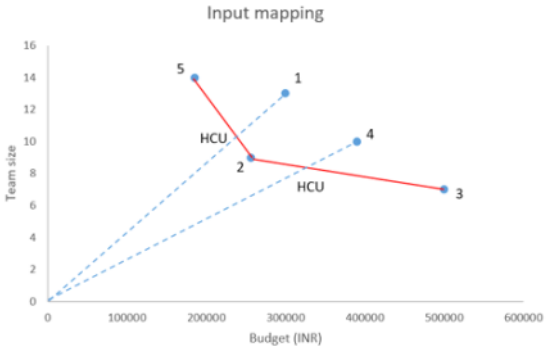
Sales office	inp1 wt	inp2 wt	out1 wt	out2 wt	
k	x1k	x2k	y1k	y2k	Efficiency
1	1.4755E-06	0.04287356	0	0.04365967	0.65
2	2.401E-06	0.04281713	4.3484E-07	0	0.76
3	1.6004E-06	0.02854094	2.8986E-07	0	1
4	1.4693E-06	0.04269541	0	0.04347826	1
5	2.3006E-06	0.04102761	4.1667E-07	0	1

9.3 Data Envelopment Analysis - Prescription for inefficient units (One output and two inputs case)

Friday, 02 December 2022 12:09

Summary	• DEA - Prescription for inefficient units (One output and two inputs case) • HCU • Graphical and analytical method to find HCU
	Inefficient DMU's Identification and prescription
	Inefficient DMU's • The graphical output and the optimization problem solution identifies the efficient and the inefficient DMU's. • The inefficient DMU's are NOT on the efficiency frontier. • How can the inefficient DMU's become efficient? • Of course, by moving towards the efficiency frontier (or the `envelope'). • However, the move to the envelope is neither vertical nor horizontal. • Important economic concepts: the types of efficiencies ("scale" vs "technical"), and disposability of inputs/outputs ("constant" vs "variable" return-to-scale). • CRS: Constant return-to-scale • VRS: Variable return-to-scale ○ IRS: Increasing return-to-scale ○ DRS: Decreasing return-to-scale
	Graphical method for inefficient DMU's • Let us consider the two-input, one-output example again. 
	Two input example • DMU's 1 and 4 are inefficient. • These DMU's need to move towards the frontier. • When they hit the frontier, that point is called "Hypothetical Composite Unit (HCU)". • For a 2D graph, easy to find the coordinates of the HCU.

Two input example



- The points where these blue dotted lines intersect the envelope are called **HCU (Hypothetical Composite Unit).**
- So it is recommended that DMU 1 moves to HCU_1 and DMU 4 moves to HCU_4.
  - Hypothetical because it is not present in the data as of now. The actual coordinates of the DMU 1 and 4 are not at their corresponding HCU points.
  - Composed because you can say, for example that HCU_4 is composed of some elements of DMU 2 and some elements of DMU 3.
- HCU coordinates can be identified by finding the intersection points of blue line and red line.
  - Or you can find the values using the dual variable values (shown below).

Graphical calculation of HCU

- Coordinates of DMU 1 are (3,00,000, 13). To generate sales of INR 10,00,000, it uses a budget of INR 3,00,000 and a team size of 13.
- The HCU is created as a combination of two efficient units as reference.
- For inefficient DMU 1, DMUs 2 and 5 are the reference.
- The HCU for 1 is found as a combination of 2 and 5.
- From simple geometry, we can find the HCU for 1 to be (2,37,540, 10.3).
- Conclusion: For the DMU 1 to be called efficient, it needs to reduce its input: from the current budget of INR 3,00,000 it needs to spend only INR 2,37,540; and reduce the team size from 13 to 10.3.

- For 10.3 team size you can understand as 10 members will be working full time and the 11th member will be working only for 30% of the time.

Analytical calculation of HCU

- From the solution of the optimization problem for DMU 1, we find that the efficiency is 0.792 (inefficient!).
- From the sensitivity report in Excel, we find that only two Dual variables are non-zero ("shadow prices"). Those variables correspond to the constraints for DMU 2 and 5. Hence, we conclude that these are the reference units for DMU 1.
- The dual variable values are 0.74 and 0.26.

- Sensitivity report from Excel:

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$5	Equality cons	1	0.791993397	1	1E+30	1
$G$7	DMU 1	0.791993397	0	0	1E+30	0.208006603
$G$8	DMU 2	0.791993397	0.740817169	0	0.174011299	0.236666667
$G$9	DMU 3	0.791993397	0	0	1E+30	0.444903013
$G$10	DMU 4	0.791993397	0	0	1E+30	0.305819232
$G$11	DMU 5	0.791993397	0.259182831	0	0.236666667	0.285790032

- Shadow prices are the values of the dual variable.
- From the table we can conclude that only 3 dual variables have non-zero values.
  - The dual variable corresponding to the equality constraint, whose value equals the value of the objective function value, and
  - The dual variable of constraints corresponding to DMU 2 and 5.

- Now, since the non-zero dual variables correspond to only DMU 2 and 5 for Linear Programming of DMU 1,
  ⇒ we say that for DMU 1, DMU 2 and 5 are the reference units.

- Value of the dual variable corresponding to DMU 2 = 0.74
- Value of the dual variable corresponding to DMU 5 = 0.26

This is the data that we had

Sales Target 10,00,000

Sales office	Budget (INR)	Team size
1	3,00,000	13
2	2,56,000	9
3	5,00,000	7
4	3,90,000	10
5	1,85,000	14

And efficiencies:

Results:

Sales office	inp1 wt	inp2 wt	out wt	
k	x1k	x2k	y1k	Efficiency
1	2.06E-06	0.029302518	7.91993E-07	0.79
2	2.61E-06	0.036998437	0.000001	1
3	7.39E-07	0.090103397	0.000001	1
4	6.21E-07	0.075776398	8.40994E-07	0.84
5	2.61E-06	0.036998437	0.000001	1

Analytical calculation of HCU for DMU 1

	Reference units		HCU for 1	Actual values	Excess inputs used
	2 (2,56,000,9)	5 (1,85,000, 14)			
Dual variable	0.74	0.26			
Sales	10,00,000*0.74 +	10,00,000*0.26	10,00,000	10,00,000	-
Budget	2,56,000*0.74 +	1,85,000*0.26	2,37,540	3,00,000	62,460
Team size	9*0.74 +	14*0.26	10.3	13	2.7

Sensitivity report of LP4:

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$5	Equality cons	1	0.840993789	1	1E+30	1
$G$7	DMU 1	0.840993789	0	0	1E+30	0.330434783
$G$8	DMU 2	0.840993789	0.704968944	0	0.183333333	0.2
$G$9	DMU 3	0.840993789	0.295031056	0	0.2	0.405263158
$G$10	DMU 4	0.840993789	0	0	1E+30	0.159006211
$G$11	DMU 5	0.840993789	0	0	1E+30	0.334782609

- Non-zero dual variables for LP4 corresponds to DMU 2 and DMU 3 constraints, so these DMUs are the reference units for the inefficient DMU 4.

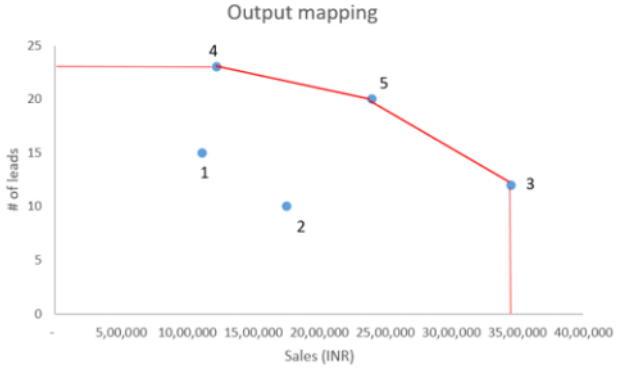
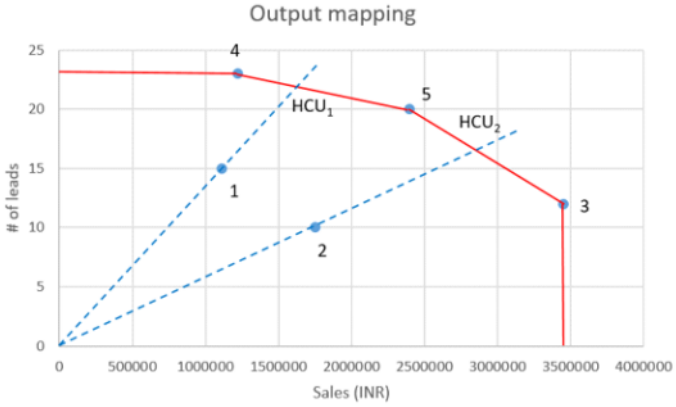
- Value of the dual variable corresponding to DMU 2 = 0.705
- Value of the dual variable corresponding to DMU 3 = 0.295

Analytical calculation of HCU for DMU 4

	Reference units		HCU for 1	Actual values	Excess inputs used
	2 (2,56,000,9)	3 (5,00,000, 7)			
Dual variable	0.705	0.295			
Sales	10,00,000*0.705	+ 10,00,000*0.295	10,00,000	10,00,000	-
Budget	2,56,000*0.705	+ 5,00,000*0.295	3,27,980	3,90,000	62,020
Team size	9*0.705	+ 7*0.295	8.41	10	1.6

9.4 Data Envelopment Analysis - Prescription for inefficient units (Two outputs and one input case)

Friday, 02 December 2022 14:59

Summary	• DEA - Prescription for inefficient units (Two outputs and one input case) • Graphical and analytical method to find HCU
	Inefficient DMU's • Let us consider the two-output, one-input example. 
	Two-output example • DMU's 1 and 2 are inefficient. • These DMU's need to move towards the frontier. • When they hit the frontier, that point is called "Hypothetical Composite Unit (HCU)". • For a 2D graph, easy to find the coordinates of the HCU.
	Two-output example 
	This is the data:

Budget 2,00,000

Sales office	Sales (INR)	No of leads
1	11,10,000	15
2	17,50,000	10
3	34,50,000	12
4	12,24,000	23
5	24,00,000	20

And the efficiencies:

Results:

Sales office k	inp1 wt x1k	out1 wt y1k	out2 wt y2k	Efficiency
1	0.000005	9.76563E-08	0.03828125	0.68
2	0.000005	1.99005E-07	0.026119403	0.61
3	0.000005	2.89855E-07	0	1
4	0.000005	0	0.043478261	1
5	0.000005	1.99005E-07	0.026119403	1

Graphical calculation of HCU

- Coordinates of DMU 1 are (11,10,000, 15). With a budget of INR 2,00,000, the DMU 1 generates a sales of INR 11,10,000 and number of sales leads of 15.
- The HCU is created as a combination of two efficient units as reference.
- For inefficient DMU 1, DMUs 4 and 5 are the reference.
- From simple geometry, we can find the HCU for 1 to be (16,26,000, 21.97).
- Conclusion: For the DMU 1 to be called efficient, it needs to increase its output: current sales of INR 11,10,000 needs to be increased to INR 16,26,000; and number of sales leads needs to go up from 15 to 21.97.

Analytical calculation of HCU

- From the solution of the optimization problem for DMU 1, we find that the efficiency is 0.682 (inefficient!).
- From the sensitivity report in Excel, we find that the dual variable values are 0.449 and 0.233, and these are from DMU 4 and 5, respectively.
- However, the dual variables don't add up to 1! (as they did in case of two-input example.)
- To achieve that, we divide the shadow prices with the efficiency value and get the final weights to be 0.658 and 0.342 for the reference DMU's 4 and 5 respectively.

Sensitivity report for LP1:

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$5	Equality cons	1	0.682617188	1	1E+30	1
$G$7	DMU 1	0.682617188	0	0	1E+30	0.317382813
$G$8	DMU 2	0.553710938	0	0	1E+30	0.446289063
$G$9	DMU 3	0.796289063	0	0	1E+30	0.203710938
$G$10	DMU 4	1	0.44921875	0	0.15	0.155671642
$G$11	DMU 5	1	0.233398438	0	0.09678018	0.130434783

- Value of the dual variable corresponding to DMU 4 = 0.449
- Value of the dual variable corresponding to DMU 5 = 0.233
- But $0.449 + 0.233 = 0.682 \neq 1$
- So, we normalize these values to: $\frac{0.449}{0.682} = 0.658$, and $\frac{0.233}{0.682} = 0.342$.

- So, modified value of the dual variable corresponding to DMU 4 = 0.658
- Modified value of the dual variable corresponding to DMU 5 = 0.342

Analytical calculation of HCU for DMU 1

	Reference units		HCU for 1	Actual values	Need to increase output by
	4 (12,24,000,23)	5 (24,00,000, 20)			
Dual variable	0.658	0.342			
Budget	2,00,000*0.658	+ 2,00,000*0.658	2,00,000	2,00,000	-
Sales	12,24,000*0.658	+ 24,00,000*0.658	16,26,094	11,10,000	5,16,094
# of leads	23*0.658	+ 20*0.26	21.97	15	6.97

Sensitivity report for LP2:

Constraints

Cell	Name	Final Value	Shadow Price	Constraint R.H. Side	Allowable Increase	Allowable Decrease
$G$5	Equality cons	1	0.609452736	1	1E+30	1
$G$7	DMU 1	0.612686567	0	0	1E+30	0.387313433
$G$8	DMU 2	0.609452736	0	0	1E+30	0.390547264
$G$9	DMU 3	1	0.273631841	0	0.4375	0.203710938
$G$10	DMU 4	0.844328358	0	0	1E+30	0.155671642
$G$11	DMU 5	1	0.335820896	0	0.09678018	0.304347826

- Value of the dual variable corresponding to DMU 3 = 0.274
- Value of the dual variable corresponding to DMU 5 = 0.336
- But again $0.274 + 0.336 = 0.61 \neq 1$
- So, we normalize these values to: $\frac{0.274}{0.61} = 0.449$, and $\frac{0.336}{0.61} = 0.551$.

- So, modified value of the dual variable corresponding to DMU 3 = 0.449
- Modified value of the dual variable corresponding to DMU 5 = 0.551

Analytical calculation of HCU for DMU 2

	Reference units		HCU for 1	Actual values	Need to increase output by
	3 (34,50,000,12)	5 (24,00,000, 20)			
Dual variable	0.449	0.551			
Budget	2,00,000*0.449	+ 2,00,000*0.551	2,00,000	2,00,000	-
Sales	34,50,000*0.449	+ 24,00,000*0.551	28,71,450	17,50,000	11,21,429
# of leads	12*0.449	+ 20*0.551	16.4	10	16.4

Week 10

Friday, 02 December 2022 15:10

10.1 Consumer Choice Models

Saturday, 03 December 2022 10:36

Summary	•

Consumer choice models

Conjoint analysis

- What we want to understand is the choices made by the consumers and how they're making it.

A consumer thinking process

- Consider a consumer comparing four products.
- We have data for two attributes on these variants.
- We also have consumer choices data available for us.

We wish to know:

- How important each attribute is to the consumer?
- What are the attribute values of the "ideal" product (from the perspective of the consumer)?

Illustrative example

	Attribute 1	Attribute 2
Product 1	1.5	12
Product 2	10	8
Product 3	2.3	4
Product 4	1	7

Pairs	Pairwise preference data	
	Prefers	Over
(1,2)	1	2
(1,3)	1	3
(1,4)	4	1
(2,3)	2	3
(2,4)	2	4
(3,4)	4	3

A consumer thinking process

- The commodity could be potato chips; and the two attributes could be crispiness and quantity in a pack.
- The consumer is comparing four different products available in the market.
- Consumer provides her choices – the choices may be pairwise, or just preferences scores for each of the brands/products.

- The marketer would like to know how important crispiness and quantity are, to this consumer (weights attached by the consumer while providing the choices).
- Using this information, can the marketer build the "ideal" chips packet that would be most preferred by this consumer?

Conjoint analysis

- Literally, conjoint analysis means an analysis of features considered jointly.

- These set of techniques have been around of a long time now.
- Conjoint analysis has its origins from a research article published in the Journal of Mathematical Psychology in 1964.

Conjoint analysis

- Family of techniques that model choice by decomposing overall preference or evaluation in terms of the relative values of the components or attributes to respondents.
- Conjoint analysis, in that sense, constructs a value system by asking about preferences on a small subset of products and then using the system to make predictions about the relative choices.
- Conjoint analysis, in the sense of optimization, can also be used to arrive at the "best product" – a product that has all the attributes at a level desirable (preferable) to the customer.

10.2 Forms of Conjoint Analysis

Saturday, 03 December 2022 11:06

Summary	•
	Forms of conjoint analysis • Choice-Based Conjoint (CBC) analysis – most commonly used form of the conjoint analysis. The customer chooses their most preferred full-profile product amongst a set of 3-4 options provided. • Adaptive Conjoint Analysis (ACA) – Each customer asked different set of questions which are dynamically decided based on their responses. • Full-profile Conjoint Analysis – Full suite of options are presented to the consumer, and their preference is sought on these. • Menu-based conjoint analysis – The customer is shown a list of attributes (and their levels) with associated prices. The customer then chooses what they want in their ideal product. They also need to pay attention to the price in their decision. • The problem with CBC analysis is that it's a fixed set of questions that we keep on asking consumers, and depending on the choices we want to offer to the consumers, the number of questions may get too many. • In ACA analysis, we adapt the questions for a given set in such a way that we can get all the necessary information from the consumers probably in less time and less number of questions. • That helps us get the responses faster before the consumer gets cognitively tired. • In menu-based conjoint analysis, we ask the consumers to build the product by giving them the list of attributes with their associated prices. • They're asked what value do they prefer for a given attribute. • In this chapter, we'll focus more on the full-profile conjoint analysis.
	Applications of conjoint analysis Marketing • Once the attributes most preferred by consumers are known, these can be highlighted in all the communication channels (such as advertising, promotion, etc.) • Consumers may differ in their choices of preferred attributes, and hence conjoint analysis can help in segmenting the market.
	Applications of conjoint analysis Product development • Once the preferred attributes are known, the product development team can focus on refining these attributes and developing something that the consumers would like. • Even at the initial development stage, the choice of attributes to focus on can be narrowed down using conjoint analysis on the products available in the market.

Applications of conjoint analysis

Pricing

- Through conjoint analysis, the most preferred attributes are highlighted.
- The organization can then decide to price the product based on the level of attribute present in that variant (if the consumer prefers motorbikes with high powered engine, the motorbike with higher "horse-power" may be sold at a premium).
- Conjoint analysis may also reveal consumer's willingness-to-pay (WTP) for particular attributes.

The process

- By defining products as collections of attributes and having the individual consumer react to a number of alternatives…
- One can infer each attribute's

a) importance, and

b) most desired level for each customer.

10.3 Conjoint Problem

Saturday, 03 December 2022 12:27

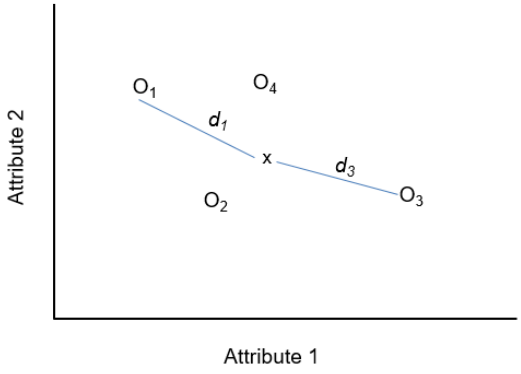
Summary	•

Conjoint analysis: Optimization method

Ref: Srinivasan V. and A.D. Shocker (1973), "Linear programming techniques for multidimensional analysis of preferences," *Psychometrika*, 38: 337 – 369.

- The article mentioned in the reference is commonly referred to as LINMAP.

Geometric explanation

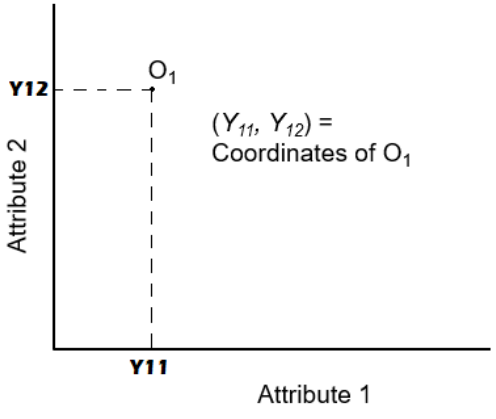


(X_1, X_2) = Coordinates of the Ideal option

(Y_{11}, Y_{12}) = Coordinates of O_1

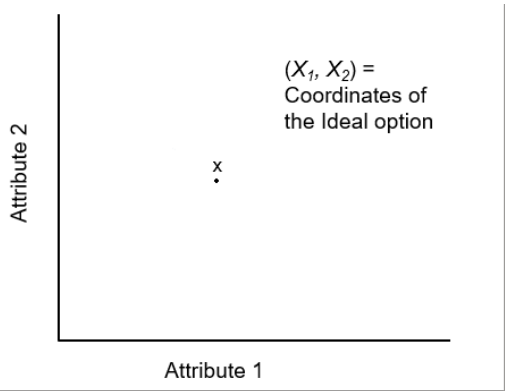
d_1 = Distance between O_1 and ideal point

Y_{11} : For O_1 attribute 1 value
Y_{12} : For O_1 attribute 2 value



x : is the ideal product that the consumers want

X_1, X_2 are the coordinates of x that we need to find out.
(we don't know where x is located)

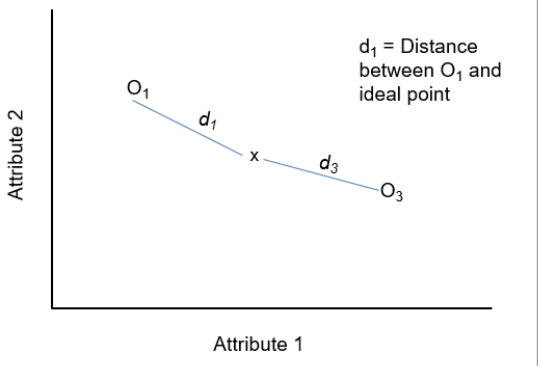


Consumers have an ideal point in their mind, and based on the relative distances between the ideal point and the given choices, they make their preferences.

d_1: (weighted) distance between the ideal point and O_1
d_3: (weighted) distance between the ideal point and O_3

Let's say in pairwise data, consumers prefer O_1 over O_3, it means that in consumer's mind, O_1 is closer to their ideal than O_3.

$\Rightarrow \quad d_1 < d_3$



- We know the coordinates of O_1 and O_3, but as we don't know the coordinates of the ideal point, we don't know the numeric value of d_1 and d_3.
- But we can find the mathematical expressions for the distances in terms of X_1 and X_2.
- And that's how we can figure out the ranking of these distances.

- We also have to figure that when the consumer is giving their preferences, are they looking at the attribute 1 value more or the attribute 2 value? And based on that we decide weights of these attributes.

- So do these distances depend on attribute 1 more or attribute 2 more? That's why we calculate weighted distances.

So, what we know:

- The coordinates of each of the product.
- Consumer's preferences, and based on that we can find raking of the distances.

In linear programming we find out:

- Coordinates of x
- The weights associated with each attribute

10.4 Optimization Formulation of Conjoint Problem

Saturday, 03 December 2022 13:30

Summary

•

The logic

- A methodology for analyzing individual differences in preference judgements with regard to a set of options.
- The product options are represented as points in a multi-attribute space.
- Different subjects correspond to different "ideal points" that denote their "most-preferred" location.
- Given two options, the customer is supposed to prefer that option which is "closer" to his/her ideal point.
- As a measure of distance, normally either the Euclidean metric or the weighted Euclidean metric is used.

- Subjects means consumers.
- Sometimes their preferences may not be rational. As consumers are evaluating the products based on multi-attributes, it sometimes may so happen that they say that they prefer, let's say, product 1 over 3, but it may not be true.

Illustrative example

	Attribute 1	Attribute 2
Product 1	1.5	12
Product 2	10	8
Product 3	2.3	4
Product 4	1	7

Pairs	Pairwise preference data	
	Prefers	Over
(1,2)	1	2
(1,3)	1	3
(1,4)	4	1
(2,3)	2	3
(2,4)	2	4
(3,4)	4	3

- Total possible pairs given n products $= \frac{n(n-1)}{2}$.
- Here $n = 4$, so possible pairs $\frac{4 \times 3}{2} = 6$.

LP model using pairwise judgements: Notations

- Set of options on which the preference judgement is made: $J = \{1, 2, \dots, n\}$.
- The n options are described in terms of t dimensions, $P = \{1, 2, \dots, t\}$.
- The pre-specified location of the jth option in the t-dimensional space is denoted by Y_j. That is $Y_j = \{y_{j,p}\}\ p \in P$.
- The ideal point of the subject is $X = \{x_p\}\ p \in P$, that is, the product location most preferred by the individual. (x_p can be positive, negative, or zero).

- Y_{jp}: j denotes the option number, and p denotes the dimension (or attribute number).

	Problem specific notations • We have four variants to be compared: $n = 4$. $j = 1, 2, 3, 4$. • The *4 (n = 4)* options are described on *2* dimensions. $t = 2$. $P = \{1, 2\}$. • Specified location of the j^{th} option in the 2-dimensional space is denoted by Y_j. This has two coordinates: $Y_j = \{Y_{j1}, Y_{j2}\}$. We have $Y_1, Y_2, Y_3,$ and Y_4. • Ideal point for this consumer is $X = \{X_1, X_2\}$.

10.5 Problem Specific Notations

Saturday, 03 December 2022 13:49

Summary	•

LP model using pairwise judgements: Notations

- Then the unweighted and weighted distance of the *j*th option from the ideal point, respectively, is given by:

$$d_j^u = \left[\sum_{p\in P}(y_{j,p} - x_p)^2\right]^{1/2}, \forall j \in J$$

$$d_j^w = \left[\sum_{p\in P} w_p(y_{j,p} - x_p)^2\right]^{1/2}, \forall j \in J$$

- Weights are non-negative for all the attributes.
- Moreover, the squared distance is $s_j = \left(d_j^w\right)^2$.

	Attribute 1	Attribute 2
Product 1	1.5	12
Product 2	10	8
Product 3	2.3	4
Product 4	1	7

In the given example,

$Y_j = (Y_{j1}, Y_{j2})$ for $j = \{1, 2, 3, 4\}$
$X = (X_1, X_2)$

Unweighted distance between Y_j and X:

$$d_j^u = \sqrt{(Y_{j1} - X_1)^2 + (Y_{j2} - X_2)^2}$$

And weighted distance between Y_j and X:

$$d_j^w = \sqrt{w_1.(Y_{j1} - X_1)^2 + w_2.(Y_{j2} - X_2)^2}$$

- Also, $w_p \geq 0 \quad \forall p \in P$
- Note that, we don't know the coordinates of X and the weights here.

So,

$$\boldsymbol{d_j^u = \sqrt{\sum_{p\in P}(Y_{jp} - X_p)^2}, \quad \forall j \in J}$$

And,

$$\boldsymbol{d_j^w = \sqrt{\sum_{p\in P} w_p.(Y_{jp} - X_p)^2}, \quad \forall j \in J}$$

And the squared distance is:

$$\boldsymbol{s_j = (d_j^w)^2}$$

LP model using pairwise judgements: Notations

- Let $\Omega = \{j, k\}$ denote the set of ordered pairs *(j, k)* where *j* designates the preferred product on a forced choice basis resulting from a paired comparison involving *j* and *k*.

- In this set-up, the only variables are weights (w_p) and the ideal point (x_p). All the others are parameters and are known apriori.
- Any optimization problem for which *(W, X)* is the solution, will satisfy the (distance) inequality:

$$s_k \geq s_j$$

Ordered pair: (j, k), where j is preferred over k.
$\Rightarrow d_j \leq d_k$
$\Rightarrow s_j \leq s_k$

- We want to find the weights and the x coordinates.

- $W = [w_1, w_2, \dots, w_t]$ → weight vector
- $X = [x_1, x_2, \dots, x_t]$ → coordinates of ideal product (vector)

- So, if (W, X) is the solution of any optimization problem, then the set of constraints given below should be satisfied
  $s_j \leq s_k$
- But it's possible that sometimes this constraint gets violated for some ordered pairs, but we want to keep this violation to the minimum.

LP model using pairwise judgements: Formulation

- Given option locations $Y_j = \{y_{j,p}\}$ and the set of ordered pairs Ω, determine the solution *(W, X)* such that

(a) weights are non-negative and

(b) distance inequality is violated as less as possible.

- The objective function can be formulated as a minimization function. And the objective function can be defined as the "poorness of fit":

$$B = poorness\ of\ fit = \sum_{(j,k)\in\Omega} (s_j - s_k)^+$$

$(x)^+ = \max(0, x)$
$\therefore (s_j - s_k)^+ = \max(0, (s_j - s_k))$

- Now, in ordered pair (j, k), we know that j is preferred over k
  $\Rightarrow s_j \leq s_k$
  $\Rightarrow (s_j - s_k) < 0$
  $\Rightarrow \therefore (s_j - s_k)^+ = \max(0, (s_j - s_k)) = 0$

- But when there's a violation
  $\Rightarrow s_j > s_k$
  $\Rightarrow (s_j - s_k) > 0$
  $\Rightarrow (s_j - s_k)^+ = (s_j - s_k) > 0$
- And it represents the "poorness of fit"
- We want this value to be minimum, that is, minimum violations.

Final formulation

Let,

$$a_{jkp} = y_{kp}^2 - y_{jp}^2, \forall (j,k) \in \Omega \text{ and p} \in P$$

$$b_{jkp} = -2(y_{kp} - y_{jp}), \forall (j,k) \in \Omega \text{ and p} \in P$$

$$\text{V} = \{\text{v}_\text{p}\} = \{\text{w}_\text{p}\text{x}_\text{p}\}, \text{p} \in P$$

$$z_{jk} = max\left[0, -\left[\sum_{p \in P} w_p a_{jkp} + \sum_{p \in P} v_p b_{jkp}\right]\right]$$

	Attribute 1	Attribute 2
Product 1	1.5	12
Product 2	10	8
Product 3	2.3	4
Product 4	1	7

Pairs	Pairwise preference data	
	Prefers	Over
(1,2)	1	2
(1,3)	1	3
(1,4)	4	1
(2,3)	2	3
(2,4)	2	4
(3,4)	4	3

Example:
$y_{11} = 1.5, y_{31} = 2.3$

$a_{131} = y_{31}^2 - y_{11}^2 = 2.3^2 - 1.5^2$

$b_{131} = -2(y_{31} - y_{11}) = -2(2.3 - 1.5) = -2(0.8) = -1.6$

Here, $\Omega = \{(1,2), (1,3), (4,1), (2,3), (2,4), (4,3)\}$
And
$p = \{1, 2\}$: number of attributes

$$\forall (j,k) \in \Omega \text{ and } p \in P$$

$$\boldsymbol{a_{jkp} = y_{kp}^2 - y_{jp}^2}$$

$$\boldsymbol{b_{jkp} = -2(y_{kp} - y_{jp})}$$

$$\boldsymbol{V = \{v_p\} = \{w_p x_p\}}$$

$$\boldsymbol{z_{jk} = \max\left[0, -\left[\sum_{p \in P} w_p a_{jkp} + \sum_{p \in P} v_p b_{jkp}\right]\right]}$$

Final formulation

$$A_p = \sum_{(j,k)\in\Omega} a_{jkp} \quad for\ p \in P$$

$$D_p = \sum_{(j,k)\in\Omega} b_{jkp} \quad for\ p \in P$$

Here, $\Omega = \{(1,2), (1,3), (4,1), (2,3), (2,4), (4,3)\}$
And
$p = \{1, 2\}$: number of attributes

So,

$A_1 = a_{121} + a_{131} + a_{411} + a_{231} + a_{241} + a_{431}$
$A_2 = a_{122} + a_{132} + a_{412} + a_{232} + a_{242} + a_{432}$

Similarly we can find D_1 and D_2.

$for\ p \in P$

$$\boldsymbol{A_p = \sum_{(j,k)\in\Omega} a_{jkp}}$$

$$\boldsymbol{D_p = \sum_{(j,k)\in\Omega} b_{jkp}}$$

Final formulation: LP

$$Min \sum_{(j,k)\in\Omega} z_{jk}$$

Subject to:

$$\sum_{p\in P} w_p a_{jkp} + \sum_{p\in P} v_p b_{jkp} + z_{jk} \geq 0 \ for\ (j,k) \in \Omega$$

$$\sum_{p\in P} w_p A_p + \sum_{p\in P} v_p D_p = 1$$

$$w_p \geq 0 \ and\ v_p\ unrestricted\ for\ p \in P$$

$$z_{jk} \geq 0 \ for\ (j,k) \in \Omega$$

Here, unknown variables are: w_p , v_p and z_{jk}.

LP Formulation

$$\boldsymbol{\min \sum_{(j,k)\in\Omega} z_{jk}}$$

Subject to

$$\forall\ (j,k) \in \Omega, \qquad p \in P$$

$$\sum_{p \in P} w_p a_{jkp} + \sum_{p \in P} v_p b_{jkp} + z_{jk} \geq 0$$

$$\sum_{p \in P} w_p A_p + \sum_{p \in P} v_p D_p = 1$$

Decision variables:

$$w_p \geq 0 \text{ and } v_p \text{ unrestricted}$$

$$z_{jk} \geq 0$$

Once you solve this problem, you've found w_p and you can find x coordinates as:

$$x_p = \frac{v_p}{w_p}$$

W10 Formulae

Saturday, 03 December 2022 19:46

Unweighted distance,

$$d_j^u = \sqrt{\sum_{p \in P} (Y_{jp} - X_p)^2} \quad , \qquad \forall\, j \in J$$

Weighted distance,

$$d_j^w = \sqrt{\sum_{p \in P} w_p . (Y_{jp} - X_p)^2} \quad , \qquad \forall\, j \in J$$

Square distance.

$$s_j = (d_j^w)^2$$

$$\forall (j,k) \in \Omega \text{ and } p \in P$$

$$a_{jkp} = y_{kp}^2 - y_{jp}^2$$

$$b_{jkp} = -2(y_{kp} - y_{jp})$$

$$V = \{v_p\} = \{w_p x_p\}$$

$$z_{jk} = \max\left[0, -\left[\sum_{p \in P} w_p a_{jkp} + \sum_{p \in P} v_p b_{jkp}\right]\right]$$

$for\ p \in P$

$$A_p = \sum_{(j,k)\in\Omega} a_{jkp}$$

$$D_p = \sum_{(j,k)\in\Omega} b_{jkp}$$

LP Formulation

$$\min \sum_{(j,k)\in\Omega} z_{jk}$$

Subject to

$$\forall\,(j,k) \in \Omega, \qquad p \in P$$

$$\sum_{p\in P} w_p a_{jkp} + \sum_{p\in P} v_p b_{jkp} + z_{jk} \geq 0$$

$$\sum_{p\in P} w_p A_p + \sum_{p\in P} v_p D_p = 1$$

Decision variables:

$$w_p \geq 0 \ \text{and} \ v_p \ \text{unrestricted}$$

$$z_{jk} \geq 0$$

you can find x coordinates as:

$$x_p = \frac{v_p}{w_p}$$

Week 11

Friday, 02 December 2022 15:15

11.1 Conjoint Problem Formulation Using Linear Programming

Saturday, 03 December 2022 21:07

Summary	•

This is the data given to us:

Product	Attribute 1	Attribute 2
1	1.5	12
2	10	8
3	2.3	4
4	1	7
5	9	1

Preference Table

Consumer choices	j		k
Prefers	1	over	2
Prefers	3	over	1
Prefers	4	over	1
Prefers	5	over	1
Prefers	2	over	3
Prefers	2	over	4
Prefers	2	over	5
Prefers	4	over	3
Prefers	3	over	5
Prefers	4	over	5

we can make ordered pair table like:

(j,k)
(1,2)
(3,1)
(4,1)
(5,1)
(2,3)
(2,4)
(2,5)
(4,3)
(3,5)
(4,5)

Now, calculate a_{jkp} and b_{jkp}: (here $p = \{1,2\}$)

(j,k)	a_{jk1}	a_{jk2}	b_{jk1}	b_{jk2}
(1,2)	a_{121}	a_{122}	b_{121}	b_{122}
(3,1)	a_{311}	a_{312}	b_{311}	b_{432}
(4,1)	a_{411}	a_{412}	b_{411}	b_{412}
(5,1)	a_{511}	a_{512}	b_{511}	b_{512}
(2,3)	a_{231}	a_{232}	b_{231}	b_{232}
(2,4)	a_{241}	a_{242}	b_{241}	b_{242}
(2,5)	a_{251}	a_{252}	b_{251}	b_{252}
(4,3)	a_{431}	a_{432}	b_{431}	b_{432}
(3,5)	a_{351}	a_{352}	b_{351}	b_{352}
(4,5)	a_{451}	a_{452}	b_{451}	b_{452}
Sum up:	A_1	A_2	D_1	D_2

And we calculate these values using formulae:

$$a_{jkp} = y_{kp}^2 - y_{jp}^2$$

$$b_{jkp} = -2(y_{kp} - y_{jp})$$

$$A_p = \sum_{(j,k)\in\Omega} a_{jkp}$$

$$D_p = \sum_{(j,k)\in\Omega} b_{jkp}$$

In excel, we get these values:

	ajk1	ajk2	bjk1	Bjk2
(1,2)	97.75	-80	-17	8
(3,1)	-3.04	128	1.6	-16
(4,1)	1.25	95	-1	-10
(5,1)	-78.75	143	15	-22
(2,3)	-94.71	-48	15.4	8
(2,4)	-99	-15	18	2
(2,5)	-19	-63	2	14
(4,3)	4.29	-33	-2.6	6
(3,5)	75.71	-15	-13.4	6
(4,5)	80	-48	-16	12
Sum	-35.5	64	2	8

LP Formulation

$$\min \sum_{(j,k)\in\Omega} z_{jk}$$

Subject to

$$\forall\,(j,k) \in \Omega, \qquad p \in P$$

$$\sum_{p\in P} w_p a_{jkp} + \sum_{p\in P} v_p b_{jkp} + z_{jk} \geq 0$$

$$\sum_{p\in P} w_p A_p + \sum_{p\in P} v_p D_p = 1$$

Decision variables:

$$w_p \geq 0 \text{ and } v_p \text{ unrestricted}$$

$$z_{jk} \geq 0$$

Now, our decision variables are:
$w_1, w_2 \geq 0$, v_1, v_2 unrestricted
And $z_{jk} \geq 0 \;\; \forall\,(j,k)\; pairs$

Now, the first constraint can be rewritten as:

$$\sum_{p\in P} w_p a_{jkp} + \sum_{p\in P} v_p b_{jkp} + \mathbf{1}.z_{jk} \geq 0$$

So ,
z_{jk} can only take one value that is specific for that (j,k) pair. So we can say that it's coefficient is 1 for that particular z_{jk} value.

So we, can set the coefficients like:

	w_1	w_2	v_1	v_2	z_{12}	z_{31}	z_{41}	z_{51}	z_{23}	z_{24}	z_{25}	z_{43}	z_{35}	z_{45}
	-	-	-	-	-	LP	soln	comes	here	-	-	-	-	-
(j,k)	a_{jk1}	a_{jk2}	b_{jk1}	b_{jk2}										
(1,2)	a_{121}	a_{122}	b_{121}	b_{122}	1	0	0	0	0	0	0	0	0	0
(3,1)	a_{311}	a_{312}	b_{311}	b_{432}	0	1	0	0	0	0	0	0	0	0
(4,1)	a_{411}	a_{412}	b_{411}	b_{412}	0	0	1	0	0	0	0	0	0	0
(5,1)	a_{511}	a_{512}	b_{511}	b_{512}	0	0	0	1	0	0	0	0	0	0
(2,3)	a_{231}	a_{232}	b_{231}	b_{232}	0	0	0	0	1	0	0	0	0	0
(2,4)	a_{241}	a_{242}	b_{241}	b_{242}	0	0	0	0	0	1	0	0	0	0
(2,5)	a_{251}	a_{252}	b_{251}	b_{252}	0	0	0	0	0	0	1	0	0	0

(4,3)	a_{431}	a_{432}	b_{431}	b_{432}	0	0	0	0	0	0	0	1	0	0
(3,5)	a_{351}	a_{352}	b_{351}	b_{352}	0	0	0	0	0	0	0	0	1	0
(4,5)	a_{451}	a_{452}	b_{451}	b_{452}	0	0	0	0	0	0	0	0	0	1
Sum up:	A_1	A_2	D_1	D_2										

The way we set the first constraint:

$$\sum_{p \in P} w_p a_{jkp} + \sum_{p \in P} v_p b_{jkp} + z_{jk} \geq 0$$

	Take sum product of decision variables row with		
	w_1 w_2 v_1 v_2 z_{12} z_{31} z_{41} z_{51} z_{23} z_{24} z_{25} z_{43} z_{35} z_{45} - - - - - LP soln comes here - - - - -		
Constraint 1	Row 1 (1,2) a_{121} a_{122} b_{121} b_{122} 1 0 0 0 0 0 0 0 0 0	≥	0
Constraint 2	Row 2 (3,1) a_{311} a_{312} b_{311} b_{432} 0 1 0 0 0 0 0 0 0 0	≥	0
...	...		
Constraint 10	Row 10 (4,5) a_{451} a_{452} b_{451} b_{452} 0 0 0 0 0 0 0 0 0 1	≥	0

Second Constraint:

$$\sum_{p \in P} w_p A_p + \sum_{p \in P} v_p D_p = 1$$

	Take sum product of:		
	w_1 w_2 v_1 v_2 - - - -		
Constraint 11	With A_1 A_2 D_1 D_2	=	1

And the objective:

$$\min \sum_{(j,k) \in \Omega} z_{jk}$$

Min	Sum of all these values:
	z_{12} z_{31} z_{41} z_{51} z_{23} z_{24} z_{25} z_{43} z_{35} z_{45} - LP soln comes here - - - - -

Decision variables constraints:

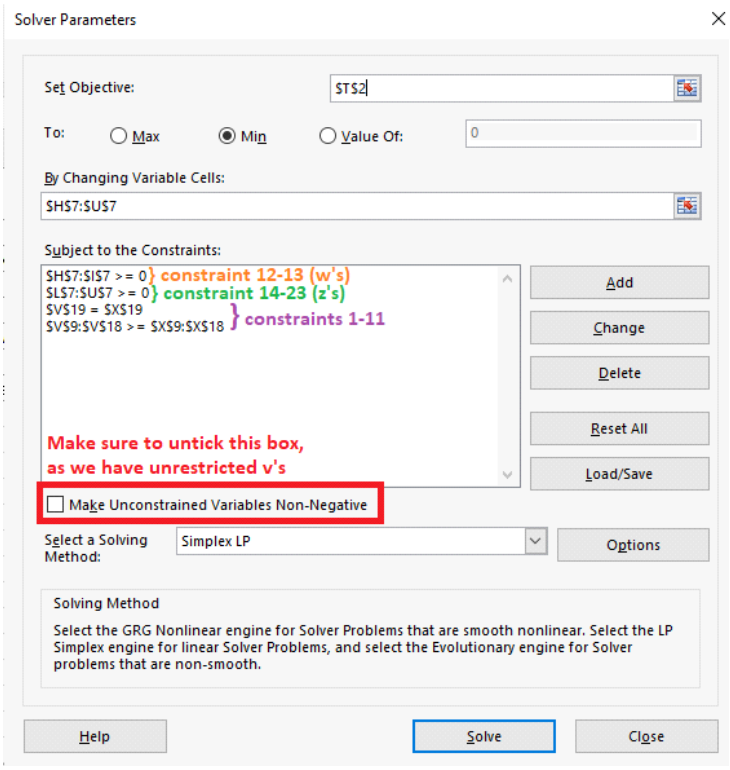
$$w_p \geq 0 \text{ and } v_p \text{ unrestricted}$$

Constraint 12	$w_1 \geq 0$
Constraint 13	$w_2 \geq 0$

$$z_{jk} \geq 0$$

Constraint 14	$z_{12} \geq 0$
Constraint 15	$z_{31} \geq 0$
...	...
Constraint 23	$z_{45} \geq 0$

And a few things to notice in the Solver:



And this is the result we get in the end:

										Objective function		Min		0.25				
	x1	x2																
	5.63367409	6.67860491																
	w1	w2	v1	v2	z12	z31	z41	z51	z23	z24	z25	z43	z35	z45				
	0.00676769	0.00991238	0.038127	0.066201	0.25	0	0	0	0	0	0	0	0	0				
	ajk1	ajk2	bjk1	Bjk2														
(1,2)	97.75	-80	-17	8	1	0	0	0	0	0	0	0	0	0	-1.1E-15	>=	0	
(3,1)	-3.04	128	1.6	-16	0	1	0	0	0	0	0	0	0	0	0.25	>=	0	
(4,1)	1.25	95	-1	-10	0	0	1	0	0	0	0	0	0	0	0.25	>=	0	
(5,1)	-78.75	143	15	-22	0	0	0	1	0	0	0	0	0	0	6.66E-16	>=	0	
(2,3)	-94.71	-48	15.4	8	0	0	0	0	1	0	0	0	0	0	1.11E-15	>=	0	
(2,4)	-99	-15	18	2	0	0	0	0	0	1	0	0	0	0	6.66E-16	>=	0	
(2,5)	-19	-63	2	14	0	0	0	0	0	0	1	0	0	0	0.25	>=	0	
(4,3)	4.29	-33	-2.6	6	0	0	0	0	0	0	0	1	0	0	4.44E-16	>=	0	
(3,5)	75.71	-15	-13.4	6	0	0	0	0	0	0	0	0	1	0	0.25	>=	0	
(4,5)	80	-48	-16	12	0	0	0	0	0	0	0	0	0	1	0.25	>=	0	
Sum	-35.5	64	2	8											1	=	1	

We can find x coordinates using:

$$x_p = \frac{v_p}{w_p}$$

These are the decision variables values:

w1	w2	v1	v2	z12	z31	z41	z51	z23	z24	z25	z43	z35	z45
0.00676769	0.00991238	0.038127	0.066201	0.25	0	0	0	0	0	0	0	0	0

Objective function value:

Objective function	Min	0.25

And, x coordinates:

x1	x2
5.63367409	6.67860491

11.2 Solving the Conjoint Problem

Sunday, 04 December 2022 10:50

Summary	•
	Now, if we look at the weights w1: 0.00676769 | w2: 0.00991238 We see that consumer gives more importance to w_2 $\frac{w_2}{w_1} = 1.46$ *i.e.,* attribute 2 is 46% more important than attribute 1.
	x coordinates: x1: 5.63367409 | x2: 6.67860491 It means that for an ideal product attribute 1 value should be 5.63 and attribute 2 value should be 6.68.
	Objective function value Objective function | Min | 0.25 z12: 0.25 | z31: 0 | z41: 0 | z51: 0 | z23: 0 | z24: 0 | z25: 0 | z43: 0 | z35: 0 | z45: 0 We ideally want minimum value to be 0. The factor contributing to the objective value is z_{12}. z values in some sense represent the violation of the pairwise choices provided by the consumer. It means that with the given weights, all the pair choices are correct except for the choice $(1,2)$, which is a violation. The consumer should have ideally chosen 2 over 1.

11.3 Conjoint Analysis using a Statistical Method

Sunday, 04 December 2022 11:20

Summary	•
	Conjoint analysis: Statistical method – Linear regression • If you have **continuous attribute data** (like we had in the previous problem), and if you have collected consumer's data in **pairwise** approach, you should go with the **optimization method.** • On the other hand, if the attribute data is **categorical**, and if you have collected only the preference **ranking (or rating)** data, you should go with the **regression** method. ○ For ranking: lower the rank number, better the product ○ For rating: higher the rating, better the product ○ We'll be considering rating data here.
	Introduction: How? • A traditional conjoint analysis is really just a multiple regression problem. • The respondent's ratings for the product concepts form the *dependent* variable. • The characteristics of the product (the attribute levels) are the *independent* (predictor) variables. • The estimated *betas* associated with the independent variables are the *utilities* (preference scores) for the levels. • The *R-Square* for the regression characterizes the internal consistency of the respondent. • y: ratings (dependent/response variable) • $x's$: attribute data (explanatory variables) • $\beta's$: regression coefficients, here are called **utilities or part worth**
	Example Consumer choice process for a cell phone purchase. Three attributes – Brand, Battery size, and the Camera resolution. • We have three brands to choose from; two options for battery, and three choices for the camera resolution. • We assume that this is a full-factorial design. Therefore, the respondent (customer, buyer) is shown all the 18 combinations available. And they provide a preference rank for each of these combinations. • See the Excel file... • 3 brand choices. 2 battery choices and 3 camera resolution choices • So, total number of combinations that can be formed = $3 \times 2 \times 3 = 18$ • So, consumers will be asked to rate 18 options. • Rating is on a scale of 10.
	Here, the data that was provided to the customers:

Options	Manufacturer	Battery	Front camera
1	Samsung	4500	20 MP
2	Samsung	4500	13 MP
3	Samsung	4500	8 MP
4	Samsung	6000	20 MP
5	Samsung	6000	13 MP
6	Samsung	6000	8 MP
7	Vivo	4500	20 MP
8	Vivo	4500	13 MP
9	Vivo	4500	8 MP
10	Vivo	6000	20 MP
11	Vivo	6000	13 MP
12	Vivo	6000	8 MP
13	Xiomi	4500	20 MP
14	Xiomi	4500	13 MP
15	Xiomi	4500	8 MP
16	Xiomi	6000	20 MP
17	Xiomi	6000	13 MP
18	Xiomi	6000	8 MP

And the customers were asked to rate these products on a scale of 10:

Options	Manufacturer	Battery	Front camera	Preference rating
1	Samsung	4500	20 MP	7
2	Samsung	4500	13 MP	3
3	Samsung	4500	8 MP	1
4	Samsung	6000	20 MP	9
5	Samsung	6000	13 MP	6
6	Samsung	6000	8 MP	3
7	Vivo	4500	20 MP	7
8	Vivo	4500	13 MP	5
9	Vivo	4500	8 MP	2
10	Vivo	6000	20 MP	10
11	Vivo	6000	13 MP	7
12	Vivo	6000	8 MP	5
13	Xiomi	4500	20 MP	7
14	Xiomi	4500	13 MP	6
15	Xiomi	4500	8 MP	4
16	Xiomi	6000	20 MP	10
17	Xiomi	6000	13 MP	7
18	Xiomi	6000	8 MP	6

Here,
Response variable = preference rating
Explanatory variables: Brand, battery and camera

To run a regression on this problem we need to first codify this categorical data into numbers:

Brand

Samsung	1
Vivo	2
Xiomi	3

Battery

4500 mAh	1
6500 mAh	2

Camera

20 MP	1
13 MP	2
8 MP	3

After codifying the data looks like this:

Options	Manufacturer	Battery	Front camera	Preference
1	1	1	1	7
2	1	1	2	3
3	1	1	3	1
4	1	2	1	9
5	1	2	2	6
6	1	2	3	3
7	2	1	1	7
8	2	1	2	5
9	2	1	3	2
10	2	2	1	10
11	2	2	2	7
12	2	2	3	5
13	3	1	1	7
14	3	1	2	6
15	3	1	3	4
16	3	2	1	10
17	3	2	2	7
18	3	2	3	6

But we will not go run the regression on this data also. We'll further codify the data into binary variables.

		Samsung	Vivo	Xiomi			4500	6000		20 MP	13 MP	8 MP	
Options	Manufacturer	X1	X2	X3	Battery	X4	X5		Front camera	X6	X7	X8	Preference
1	1	1	0	0	1	1	0		1	1	0	0	7
2	1	1	0	0	1	1	0		2	0	1	0	3
3	1	1	0	0	1	1	0		3	0	0	1	1
4	1	1	0	0	2	0	1		1	1	0	0	9
5	1	1	0	0	2	0	1		2	0	1	0	6
6	1	1	0	0	2	0	1		3	0	0	1	3
7	2	0	1	0	1	1	0		1	1	0	0	7
8	2	0	1	0	1	1	0		2	0	1	0	5
9	2	0	1	0	1	1	0		3	0	0	1	2
10	2	0	1	0	2	0	1		1	1	0	0	10
11	2	0	1	0	2	0	1		2	0	1	0	7
12	2	0	1	0	2	0	1		3	0	0	1	5
13	3	0	0	1	1	1	0		1	1	0	0	7
14	3	0	0	1	1	1	0		2	0	1	0	6
15	3	0	0	1	1	1	0		3	0	0	1	4
16	3	0	0	1	2	0	1		1	1	0	0	10
17	3	0	0	1	2	0	1		2	0	1	0	7
18	3	0	0	1	2	0	1		3	0	0	1	6

We've defined 3 explanatory variables for brand

$$X_1 = \begin{cases} 1, & if\ brand = Samsung \\ 0, & otherwise \end{cases}$$

$$X_2 = \begin{cases} 1, & if\ brand = Vivo \\ 0, & otherwise \end{cases}$$

$$X_3 = \begin{cases} 1, & if\ brand = Xiomi \\ 0, & otherwise \end{cases}$$

Similarly we define explanatory variables for various battery and camera choices:

$$X_4 = \begin{cases} 1, & if\ battery = 4500\ mAH \\ 0, & otherwise \end{cases}$$

$$X_5 = \begin{cases} 1, & if\ battery = 60000\ mAH \\ 0, & otherwise \end{cases}$$

$$X_6 = \begin{cases} 1, & if\ camera = 20\ MP \\ 0, & otherwise \end{cases}$$

$$X_7 = \begin{cases} 1, & if\ camera = 13\ MP \\ 0, & otherwise \end{cases}$$

$$X_8 = \begin{cases} 1, & if\ camera = 8\ MP \\ 0, & otherwise \end{cases}$$

This is how the codified data looks like:

	Manufacturer			Battery		Camera			
	Samsung	Vivo	Xiomi	4500	6000	20 MP	13 MP	8 MP	

	Manufacturer			Battery		Camera			
	Samsung	Vivo	Xiomi	4500	6000	20 MP	13 MP	8 MP	
Options	X1	X2	X3	X4	X5	X6	X7	X8	Preference
1	1	0	0	1	0	1	0	0	7
2	1	0	0	1	0	0	1	0	3
3	1	0	0	1	0	0	0	1	1
4	1	0	0	0	1	1	0	0	9
5	1	0	0	0	1	0	1	0	6
6	1	0	0	0	1	0	0	1	3
7	0	1	0	1	0	1	0	0	7
8	0	1	0	1	0	0	1	0	5
9	0	1	0	1	0	0	0	1	2
10	0	1	0	0	1	1	0	0	10
11	0	1	0	0	1	0	1	0	7
12	0	1	0	0	1	0	0	1	5
13	0	0	1	1	0	1	0	0	7
14	0	0	1	1	0	0	1	0	6
15	0	0	1	1	0	0	0	1	4
16	0	0	1	0	1	1	0	0	10
17	0	0	1	0	1	0	1	0	7
18	0	0	1	0	1	0	0	1	6

We do not want a situation of multi-collinearity in regression. Multi-collinearity occurs when the explanatory variables are correlated.

- In this data, let's say we know that $X_2 = 0$, and $X_3 = 0$, *i.e.*, we know that the given product is neither Vivo nor Xiomi, then it's obvious that the product brand in Samsung, *i.e.*, $X_1 = 1$.
- Knowing two variables value, gives us the value of third variable.

- So, there's a strong correlation between brand variables.
- Similarly, we can show a strong correlation between battery and camera variables.

Q. How do you remove this multi-collinearity?
A. From each of these three groups, we will remove one of the variables.
  - Let's say we remove, Samsung brand (X_1) from brands, 4500mAh(X_4) from batteries, and 8MP (X_8) from camera.
  - Removing these variables will not change the regression result for us.

	Manufacturer			Battery		Camera			
	Samsung	Vivo	Xiomi	4500	6000	20 MP	13 MP	8 MP	
Options	X1	X2	X3	X4	X5	X6	X7	X8	Preference
1	1	0	0	1	0	1	0	0	7
2	1	0	0	1	0	0	1	0	3
3	1	0	0	1	0	0	0	1	1
4	1	0	0	0	1	1	0	0	9
5	1	0	0	0	1	0	1	0	6
6	1	0	0	0	1	0	0	1	3
7	0	1	0	1	0	1	0	0	7
8	0	1	0	1	0	0	1	0	5
9	0	1	0	1	0	0	0	1	2
10	0	1	0	0	1	1	0	0	10
11	0	1	0	0	1	0	1	0	7
12	0	1	0	0	1	0	0	1	5
13	0	0	1	1	0	1	0	0	7
14	0	0	1	1	0	0	1	0	6
15	0	0	1	1	0	0	0	1	4
16	0	0	1	0	1	1	0	0	10
17	0	0	1	0	1	0	1	0	7
18	0	0	1	0	1	0	0	1	6

- So the removed variable will be coded as giving a utility of 0. They become the base values or base utilities.
- And you calculate other utilities as deviation from these utilities.
- In that sense, we don't get the direct utility values, we only get the relative utility values.

So, this is how the final codified, ready to use for regression, data looks like:

	X2	X3	X5	X6	X7	
Options	Vivo	Xiomi		6000	20 MP	13 MP	Preference
1	0	0	0	1	0	7
2	0	0	0	0	1	3
3	0	0	0	0	0	1
4	0	0	1	1	0	9
5	0	0	1	0	1	6
6	0	0	1	0	0	3
7	1	0	0	1	0	7
8	1	0	0	0	1	5
9	1	0	0	0	0	2

Options	X2 Vivo	X3 Xiomi	X5 6000	X6 20 MP	X7 13 MP	Preference
1	0	0	0	1	0	7
2	0	0	0	0	1	3
3	0	0	0	0	0	1
4	0	0	1	1	0	9
5	0	0	1	0	1	6
6	0	0	1	0	0	3
7	1	0	0	1	0	7
8	1	0	0	0	1	5
9	1	0	0	0	0	2
10	1	0	1	1	0	10
11	1	0	1	0	1	7
12	1	0	1	0	0	5
13	0	1	0	1	0	7
14	0	1	0	0	1	6
15	0	1	0	0	0	4
16	0	1	1	1	0	10
17	0	1	1	0	1	7
18	0	1	1	0	0	6

11.4 Regression Method for Conjoint Analysis

Sunday, 04 December 2022 14:20

Summary	•

Options	X2 Vivo	X3 Xiomi	X5 6000	X6 20 MP	X7 13 MP	Preference
1	0	0	0	1	0	7
2	0	0	0	0	1	3
3	0	0	0	0	0	1
4	0	0	1	1	0	9
5	0	0	1	0	1	6
6	0	0	1	0	0	3
7	1	0	0	1	0	7
8	1	0	0	0	1	5
9	1	0	0	0	0	2
10	1	0	1	1	0	10
11	1	0	1	0	1	7
12	1	0	1	0	0	5
13	0	1	0	1	0	7
14	0	1	0	0	1	6
15	0	1	0	0	0	4
16	0	1	1	1	0	10
17	0	1	1	0	1	7
18	0	1	1	0	0	6

After running the regression, this is the output we get

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.975568829
R Square	0.95173454
Adjusted R Square	0.931623932
Standard Error	0.666666667
Observations	18

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	105.1666667	21.03333333	47.325	1.73991E-07
Residual	12	5.333333333	0.444444444		
Total	17	110.5			

			Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
	Intercept		1.333333333	0.384900179	3.464101615	0.004681605	0.494707884	2.171958782
X2	Vivo		1.166666667	0.384900179	3.031088913	0.010445429	0.328041218	2.005292116
X3	Xiomi		1.833333333	0.384900179	4.763139721	0.000461644	0.994707884	2.671958782
X5		6000	2.333333333	0.314269681	7.424621202	8.00423E-06	1.648598521	3.018068145
X6	20 MP		4.833333333	0.384900179	12.55736835	2.91134E-08	3.994707884	5.671958782
X7	13 MP		2.166666667	0.384900179	5.629165125	0.000110881	1.328041218	3.005292116

Example: Regression statistics

- The regression equation is:

$$Y = 1.334 + 1.167 * VIVO + 1.83 * XIOMI + 2.334 * (6000mAh) + 4.83 * (20MP) + 2.16 * (13MP)$$

- The regression coefficients are called “part-worth”.
- The regression is significant.
- Each of the explanatory variables also significant.

Partworths

- These are the level utilities for the attributes.
- The total worth of the product (option) is calculated from multiple attributes and multiple levels of attributes together.
- Utility values for the separate parts of the product (assigned to the attributes) are the part worths.

- Partworth for each level of each attribute:

Brand/Manufacturer	
Samsung	0
Vivo	1.167
Xiomi	1.834

Battery	
4500 mAh	0
6000 mAh	2.334

Camera resolution	
8 MP	0
13 MP	2.167
20 MP	4.834

- You can think of the total part worth as the rating the customer has given to a product.
- Here, we get the relative part worths, not the absolute part worths.

- Samsung is the base level for brands. So, it gets a part worth (or utility) of 0.
- Vivo gets a part worth of 1.167 relative to Samsung.
- Xiomi gives an increment of 1.834 part worth over the base level.

- Also, we can rank brands, or battery, or camera preferences, based on their part worths.
  - *e.g.*, for brand attribute: Xiomi > Vivo > Samsung

Importance of each attribute

Attribute	Partworth range
Brand	1.834 – 0 = 1.834
Battery	2.334 – 0 = 2.334
Camera	4.834 – 0 = 4.834

Total of ranges = 1.834 + 2.334 + 4.834 = 9.002

Attribute	Importance
Brand	1.834/9 = 20.37%
Battery	2.334/9 = 25.92%
Camera	4.834/9 = 53.7%

- $Range = \max\ value\ - \min value$

- So, we basically normalize the ranges and find the percentages of importance for each attribute.

- And the ideal products are those that got a rating of 10.

Week 12

Friday, 02 December 2022 15:15

12.1 Epilogue

Sunday, 04 December 2022 14:56

Summary	•
	# Epilogue
	## The course • So far, we have seen some applications of analytics in various business examples. • Clearly, this is only an introduction. • As is true for any academic program, the course contents introduce you to the ideas of the field. • The real test is in the scenarios presented in reality. • In the field, no business problems comes as a "regression problem" or an "optimization problem". • They are just daily challenges to the managers, viz. "marketing effectiveness", "financial options", "operational efficiency", etc.
	## Course contents • Topics for the course were chosen to expose important principles of data analytics in businesses. • In-depth discussions in subsequent courses.... • Coming shortly: More Business-Analytics Electives!
	## Way forward Today any manager is expected to 1. Either, Manage traditional businesses using advanced analytics. 2. Or, Manage newer data-driven businesses. • One may argue that every business today is data-driven? • Hence, analytical skills are mandatory.

Purpose of analytics

Four stages of analytics

- Descriptive analytics

"What Happened?"

- Diagnostic analytics

"Why did this happen?"

- Predictive analytics

"What might happen in the future?"

- Prescriptive analytics

"What should we do next?"

- Not just to observe patterns and trends.
- But, to take better decisions.

	Good decisions	Bad decisions
Good analytics	Ideal.	Are you listening to your data scientists?
Bad analytics	Do you need analytics?	Karma.

Have the organizations changed?

- Not really!
- Our core objectives are still the same: "Make more profits", "Reach out to more customers", "Keep your employees happy", "Secure your infrastructure", etc.
- Decision processes are now different.
- Today's large organizations mimic those decision processes through analytics.

HR and Analytics

- Earlier, an employer would know her employees well (personal rapport, small scale, family ties, etc.).

Now –

- Employee profile mining (Naukri.com; Monster.com; LinkedIn, etc.)
- Employee behavior pattern analysis (attendance analytics, performance review, websites visited, etc.).
- Employee turnover analytics (When would an employee leave an organization? What is average duration of stay? Why do they leave?)

Marketing and analytics

- Earlier, a seller would know her customers well (geographical proximity, strong networks, smaller supply chains, personal connects, etc.)

Now –

- Customer identification and segmentation (logins, saved cards, preferred stores, demographic data, IP addresses, etc.).
- Customer buying behavior (order history, demographic data, etc.).
- Customer retentions (loyalty programs, customized offers).

Finance and analytics

- The financial investment alternatives were few and non-complex, lenders knew their borrowers.

Today –

- We have complex financial alternatives (e.g. financial derivatives).
- Financial product design is data intensive.
- Investment decisions are time-critical (algorithmic trading, micro-seconds time phasing, etc.)
- Very complex and open debt market.
- Analyzing financial performance and health of a firm is complex acvitiy.

Operations and analytics

- Today's supply chains are complex, with multiple suppliers, and geographical spread and multi-modal logistics.
- JIT revolution has necessitated time optimization of the logistics and supply chain operations.
- Number and criticality of the resources in the manufacturing requires sensorization of the equipment for condition monitoring.
- Volume and variety of products produced has increased. Due to this, there is a need to track the production and movement of these products to the customers.

- JIT- Just In Time

12.1 Epilogue- Continuation

Sunday, 04 December 2022 16:19

Summary	•
	Development of appropriate techniques
	Financial engineering • Intersection of mathematics, computer science, data science, and economic theory. • Mathematical finance; Computational finance (generally considered as subfields of financial engineering). • Misnomer: this does NOT belong to any traditional engineering domains (though many of its students may come from that background!). • "Quants" (aka, Quantitative Analysts) – a person who works in this fields. A person who applies quantitative techniques to finance. • Specialized field on its own (e.g. One can get a masters degree in financial engineering).
	Marketing engineering • Origin of the term: "The age of marketing engineering," authored by Lilien, Rangaswamy and Matanovich back in 1998. • "Using (such) computer decision models for making marketing decisions is known as `marketing engineering'." Another definition: "A systematic approach to harness data and knowledge to drive effective marketing decision making and implementation through a technology-enabled and mode-supported decision process." (Rangaswamy, de Bruyn, 2013).
	Marketing engineering • Another quote from that the classic article: "Emerging corporations don't have any need for classical marketing education. What they had a need for is understanding customers… how to analyze databases – supplier databases, demographic, and geographic databases." • Representative tools and techniques: choice modelling, conjoint analysis, yield management, Bass forecasting, Customer lifetime value, etc.

Operational excellence

- "Executing a business strategy most efficiently."
- Collection and analysis of operational data -> defining performance metrics -> model building -> Recommendations -> Implementation.
- Industry 4.0 is enabling automation via data-driven decisions.
- Applications: Supply chain efficiency; Quality measurements (Six Sigma); Resource utilization (machine uptime); Logistics operations.
- Supply Chain Analytics is a domain in itself!
- "Operations Forensics" Richard Lai (2013).

Decision making <-> Analytics

- Three V's of Big Data: Volume, Variety, Velocity.
- Advantages of analytics driven decision making: Deeper insights, faster!
- Using a lot of data points into analysis – hopefully provides better insights.
- Efficient (and correct) algorithms – derive the insights faster.

- Traditional decision making processes would have slowed down in the face of today's scale.

Where do we go from here?

- Only an introduction to business analytics.
- More analytics-centric management electives coming!

- Today, unimaginable to have business decisions taken without considering the data!
- So, plenty of cases available. Try to access them and study.

- But always remember, the techniques (data analysis) and the context (business problem) are both important!

{Contact}

Sunday, 13 November 2022 10:48

If you find any error or have a suggestion, you can write me at:


sheetalnotes@outlook.com

Wednesday, 07 December 2022 12:35

Say a data is distributed as “Normal” with a right tail. If it is compared with an “Symmetric Normal” distribution, then which of the following states is/ are true (choose all that is applicable)

Options :

6406531194254. ✖ In a “P-P plot”, the data will fall on a line which is indicate at 45-degree to the X-axis

6406531194255. ✔ In a “Q-Q plot”, the data will fall on a line which is indicate at 45-degree to the X-axis

6406531194256. ✔ In a “P-P plot”, the data will not entirely fall on a line which is indicate at 45-degree to the X-axis

6406531194257. ✖ In a “Q-Q plot”, the data will not entirely fall on a line which is indicate at 45-degree to the X-axis

6406531194258. ✖ Cannot use P-P plot or Q-Q plot as the assumed distribution is discrete

Product	Sales of a Product in a City for a given Year								
	City-1 (1990)	City-2 (1990)	City-3 (1990)	City-1 (1991)	City-2 (1991)	City-3 (1991)	City-1 (1992)	City-2 (1992)	City-3 (1992)
A	100	90	250	120	50	120	140	20	500
B	145	300	500	175	250	250	195	230	1000
C	90	180	30	100	110	15	110	95	58
D	130	220	132	140	200	61	150	180	270

Table-1

Say you want to see if the distribution of sales of Product-A in the Table-1 follows a uniform distribution within the range of 0 to 300 when split into bins as specified in Table-2. Then what is the expected frequency in any given bin (round your answer to two decimal places)?

Bin Number	Bin Range
Bin-1	Sales value less than or equal to 50
Bin-2	Sales value greater than 50 but less than or equal to 100
Bin-3	Sales value greater than 100 but less than or equal to 150
Bin-4	Sales value greater than 150 but less than or equal to 200
Bin-5	Sales value greater than 200 but less than or equal to 250
Bin-6	Sales value greater than 250 but less than or equal to 300

Table-2

1.5

If a Chi-Square Goodness-Of-Fit Test to check if the data for Product-A in Table-1 follows a Uniform Distribution with bins as specified in Table-2, then what is the value of the Test statistic (round your answer to two decimal places)?

{Hint: $Chi\text{-}square = \sum_k \frac{(observed_k - Expected_k)^2}{Expected_k}$}

3.50 to 3.70

A Chi-squared Goodness-Of-Fit test with the bins as specified in Table-2 is going to be carried out to check if the data on sales (whole data in Table-1) is normal or not. Then what is the degrees of freedom for the test?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

3

Say you want to see if the distribution of sales of Product-D in the Table-1 follows a uniform distribution within the range of 0 to 300 when split into bins as specified in Table-2. Then what is the expected frequency in any given bin (round your answer to two decimal places)?

Bin Number	Bin Range
Bin-1	Sales value less than or equal to 50
Bin-2	Sales value greater than 50 but less than or equal to 100
Bin-3	Sales value greater than 100 but less than or equal to 150
Bin-4	Sales value greater than 150 but less than or equal to 200
Bin-5	Sales value greater than 200 but less than or equal to 250
Bin-6	Sales value greater than 250 but less than or equal to 300

Table-2

1.5

Question Label : Short Answer Question

If a Chi-Square Goodness-Of-Fit Test to check if the data for Product-D in Table-1 follows a Uniform Distribution with bins as specified in Table-2, then what is the value of the Test statistic (round your answer to two decimal places)?

{Hint: Chi-square $= \sum_k \frac{(observed_k - Expected_k)^2}{Expected_k}$}

Possible Answers :

6.20 to 6.40

A Chi-squared Goodness-Of-Fit test with the bins as specified in Table-2 is going to be carried out to check if the data on sales (whole data in Table-1) is normal or not. Then what is the degrees of freedom for the test?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

3

Question Label : Comprehension

The daily sales of choco-chip ice-creams in an ice-cream parlour is assumed to follow a Normal distribution with a mean of 5 and a variance of 4. To check the assumption of normal distribution, the sales of choco-chip ice-creams at the ice-cream parlour for 7 days was collected (Table-1). On all 7 days, the total number of ice-creams (of all flavours) sold per day was 100. With this data, answer the given subquestions.

{Note: Round all your calculations (intermediate and final) to one decimal place}

Day	The actual number of choco-chip ice-creams sold
Day-1	12
Day-2	14
Day-3	12
Day-4	13
Day-5	12
Day-6	11
Day-7	12

Table-1

How many degrees of freedom does that Chi-Square Goodness of Fit Test statistic have?

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

4

Which of the following is the **correct null hypothesis** for the Goodness of fit test for this problem (with data in Table-1)?

Options :

A. ✔ The distribution of sale of choco-chip ice-creams at the ice-cream parlour is normal

B. ✖ The distribution of sale of choco-chip ice-creams is normal

C. ✖ The distribution of choco-chip ice-creams at the ice-cream parlour is normal

D. ✖ The distribution of choco-chip ice-creams is normal

Question Label : Comprehension

The health parameters of 100 patients are used to develop a logistic model that predicts if a given patient has CORONA (Y=1) or not (Y=0). It is seen that the model correctly predicts the presence of CORONA in 23 patients and correctly predicted the absence of CORONA in 11 patients. The threshold set to obtain these results was 0.6, and the variable of focus was to identify the patients with CORONA (Y=1). Then answer the given subquestions

Sub questions

What is the accuracy (in percentage) of predicting CORONA patients correctly? (round to two decimal places)

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

16.90 to 17.01

The number of data points that contribute to the "Type-1 Error" of the model is _____

NOTE: Enter your answer to the nearest integer.

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Equal

Text Areas : PlainText

Possible Answers :

77

With what precision (in percentage) are patients without CORONA identified by the model? (round to two decimal places)

Response Type : Numeric

Evaluation Required For SA : Yes

Show Word Count : Yes

Answers Type : Range

Text Areas : PlainText

Possible Answers :

12.5 to 13.5

Say a data is distributed as “Normal” with a right tail. If it is compared with an “Symmetric Normal” distribution, then which of the following states is/ are true (choose all that is applicable)

Options :

6406531194254. ✖ In a “P-P plot”, the data will fall on a line which is indicate at 45-degree to the X-axis

6406531194255. ✔ In a “Q-Q plot”, the data will fall on a line which is indicate at 45-degree to the X-axis

6406531194256. ✔ In a “P-P plot”, the data will not entirely fall on a line which is indicate at 45-degree to the X-axis

6406531194257. ✖ In a “Q-Q plot”, the data will not entirely fall on a line which is indicate at 45-degree to the X-axis

6406531194258. ✖ Cannot use P-P plot or Q-Q plot as the assumed distribution is discrete