Chapter 12: Correlation and Regression

Siddiqua Mazhar Ph.D

Introduction

- Looking to find a relationship between two variables
- If there is a relationship, what is it?
- If there is a relationship, how can we use it?

Words

- Correlation: Tells you that relationship exists or not between two variables.
- Regression: Describes the relationship between two variables (mathematically).

Simple vs. Complex Relationships

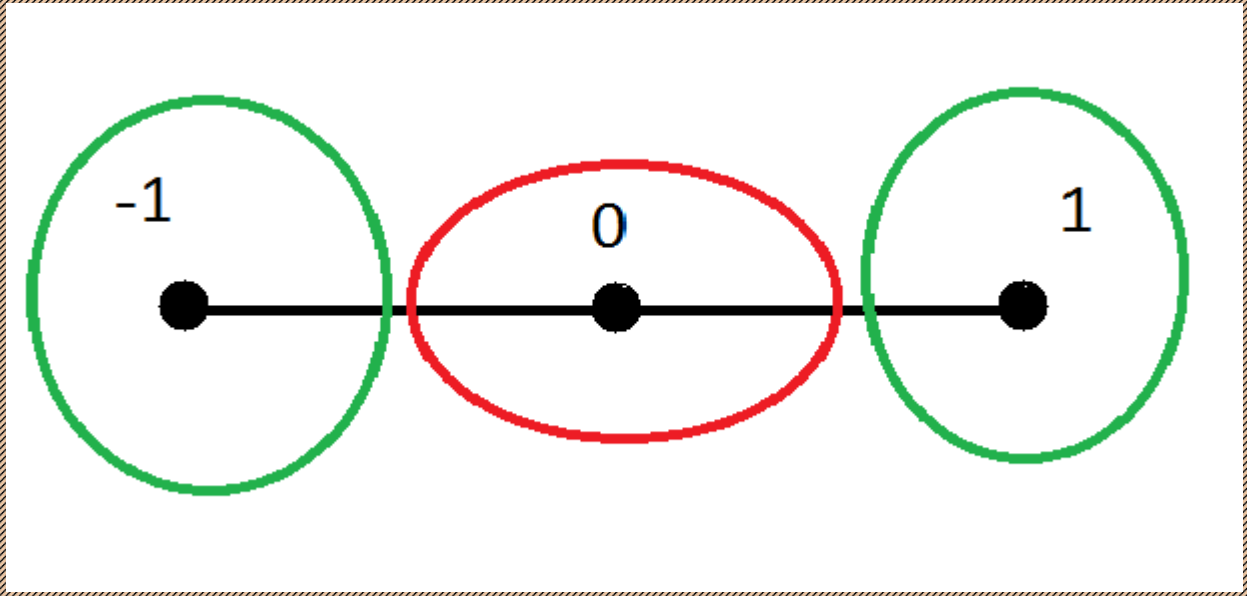
- Simple Regression (one variable cause)
- Multiple Regression (multiple variable cause)

Correlation

- A numerical marker for correlation is called a correlation coefficient
- Pearson correlation coefficient when linear relationship
- Symbol *r*

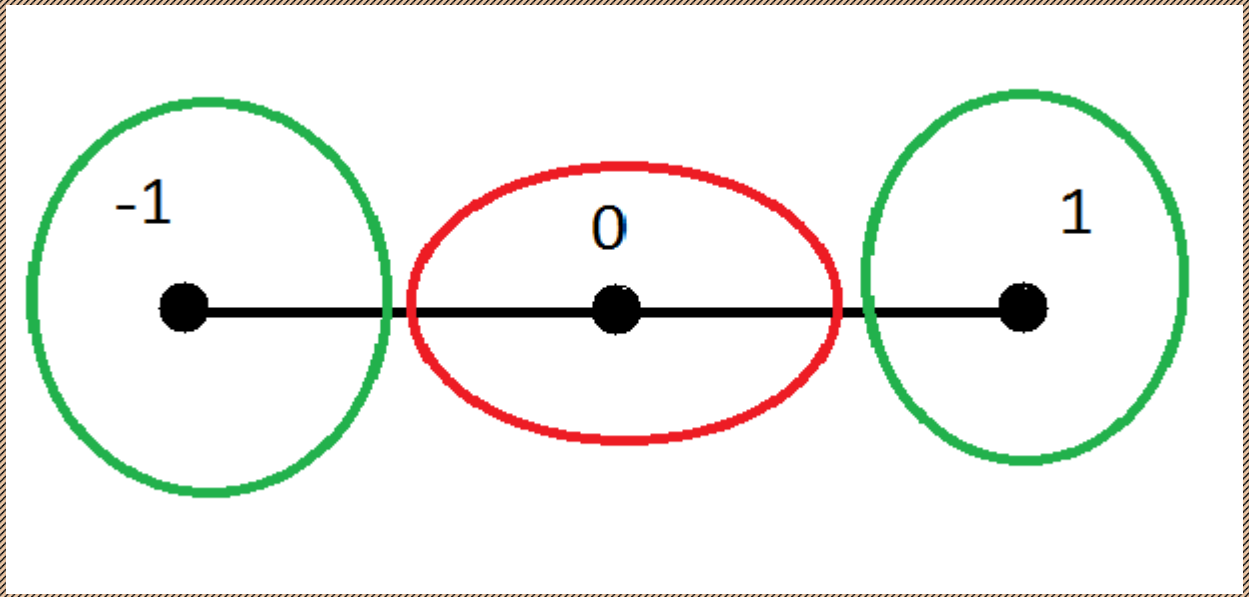
Correlation

- Symbol r $-1 \leq r \leq 1$
- Close to 1 or -1 then Good Fit.
- Close to 0 then Bad fit



Correlation

- Cutoff between Good Fit and Bad fit is _____



Regression

- Regression gives the relationship.
- Computers will make the best relationship it can

Regression

- Least squares regression line gives a line of best fit.

Example 1 – Gun Laws and Crime

- Create a least Squares regression line and give correlation coefficient for the following data.

- The data shows the Brady Scorecard points from 2014 and the murder rates for 2016.
  - Cite: http://www.crimadvisor.com/?page=scorecard
  - Cite: https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/tables/table-4

State	Brady Scorecard points 2014 (x)	Murder Rate 2016 (y)
Wyoming	-28	3.4
Idaho	-19	2.9
Indiana	-14.5	6.5
South Dakota	-9	3.2
New Hampshire	-7	1.4
Michigan	3	6.2
Wisconsin	6	4.0
Pennsylvania	23	5.3
Illinois	40.5	8.3
Maryland	56	8.9

Example 1

- Regression Line:
- Correlation Coefficient:

Example 1

- Predict what the Murder Rate will be if the Brady Points are 3.

- Predict what the Murder Rate will be if the Brady Points are 30.

Limitations

- Causation vs. Correlation
- Extrapolation
- Prediction Range

Correlation ≠ Causation

- Correlation tells us that variables are RELATED (mathematically) or in a PATTERN.

Correlation ≠ Causation

- $X \rightarrow Y$
- $Z \rightarrow \begin{matrix} X \\ Y \end{matrix}$
- $X \; ?? \; Y$

$$Y \rightarrow X$$

$$X_1, X_2, \ldots X_n \rightarrow Y$$

Correlation ≠ Causation

- Use common sense
- Don’t use this to dismiss

Extrapolation

- Extrapolation is predicting outside of known bounds

- For Example if $23 < X < 68$, then should not use $X = 5$ in predictions and not use $X = 72$.

Extrapolation

- Where should we not predict due to extrapolation in the Gun Laws example?

Prediction range

- Range of predictions can vary by problem
  - We predicted Murder rate when Brady points were 3.
  - Michigan actually has 3 for Brady points.

Prediction Range

- There is a formal way of how big of range should be
- “Confidence Interval for predictions”

Example 2 – Money and Placement

- We want to see if there is a correlation between the money spent on a college wrestling program and their final placement in 2017.

- We collect data on the Big 10 programs. We collect the placement in the NCAA wrestling finals along with their expenses on their wrestling teams.

- Cite: http://board.themat.com/index.php?/topic/6277-ncaa-d1-funding-by-the-numbers/
- Cite: https://www.trackwrestling.com/

Institution	Expenses in $ (x)	Placement (y)
Indiana	1,169,695	35
Michigan State	973,026	41
Northwestern	1,445,648	46
Ohio State	2,349,054	2
Pennsylvania State	2,341,696	1
Purdue	1,122,255	50
Rutgers	1,698,342	19
Illinois	1,275,385	11
Iowa	2,270,226	4
Maryland	1,246,160	35
Michigan	1,657,023	10
Minnesota	1,215,970	7
Nebraska	1,599,430	9
Wisconsin	1,531,224	13

Example 2

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 2

- Where would you expect a college to place if a college spent $1,500,000 on their team?

- Where would you expect a college to place if a college spent $2,000,000 on their team?

Example 2

- Circle all values where you should not predict due to extrapolation.

$400,000 $800,000 $1,250,000

$1,700,000 $2,250,000 $2,750,000

Example 3 – Money and Points

- We want to see if there is a correlation between the money spent on a college wrestling program and their final points earned in 2017.

- We collect data on the Big 10 programs. Points scored in the NCAA wrestling finals and total expenses on wrestling teams are displayed.

- Cite: http://board.themat.com/index.php?/topic/6277-ncaa-d1-funding-by-the-numbers/
- Cite: https://www.trackwrestling.com/

Institution	Expenses $ (x)	Points scored (y)
Indiana	1169695	8.5
Michigan State	973026	4.5
Northwestern	1445648	3.5
Ohio State	2349054	110
Pennsylvania State	2341696	146.5
Purdue	1122255	2.5
Rutgers	1698342	24.5
Illinois	1275385	43.5
Iowa	2270226	97
Maryland	1246160	8.5
Michigan	1657023	47.5
Minnesota	1215970	62.5
Nebraska	1599430	59.5
Wisconsin	1531224	39.5

Example 3

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 3

- How many points would you expect a college to place if a college spent $1,500,000 on their team?

- How many points would you expect a college to place if a college spent $2,000,000 on their team?

Example 4 – Big Mac and Happiness

- We want to see if there is a correlation between a country's Big Mac Index and their overall level of happiness.
- That data was collected for a sample of 11 countries.
- Note: The Cantril ladder asks those to evaluate the quality of life on a scale from 0 to 10.

- Cite: http://www.globalprice.info/en/?p=statistics/bigmac
- Cite: https://s3.amazonaws.com/happiness-report/2018/WHR_web.pdf

Country	Big Mac Index (x)	Cantril ladder (y)
Australia	4.33	7.272
Canada	5.25	7.328
Czech Republic	3.79	6.711
Denmark	4.61	7.555
Hungary	3.02	5.620
Japan	3.32	5.915
Mexico	2.57	6.488
Norway	5.95	7.594
Switzerland	6.55	7.487
United States	5.58	6.886
United Kingdom	4.12	6.814

Example 4

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 4

- Use the line to predict the Happiness level if the Big Mac index is $3.06.

Example 5 – McD's and Happiness

- We want to see if there is a correlation between the number of McDonalds restaurants per million people in a country and their overall level of happiness.
- That data was collected for a sample of 11 countries.
  - Cite: https://s3.amazonaws.com/happiness-report/2018/WHR_web.pdf
  - Cite: https://en.wikipedia.org/wiki/List_of_countries_with_McDonald%27s_restaurants

Country	McDonalds per Million (x)	Cantril ladder score (y)
Australia	32.2	7.272
Canada	33.9	7.328
Czech Republic	5.9	6.711
Denmark	17.9	7.555
Hungary	7.6	5.620
Japan	28.4	5.915
Mexico	1.8	6.488
Norway	11.7	7.594
Switzerland	15.6	7.487
United States	40.9	6.886
United Kingdom	17.8	6.814

Example 5

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 5

- Use the line to predict the Happiness level if the Number of McDonalds per Million people is 18.4.

A way to cheat…

- I want to show how we can paint a narrative if we selectively ignore data that we do not like and selectively focus on data we do like.

- Focus on subsets of the data from number of McDonalds example for the next two examples.

Example 6 – Number of McDonalds

Country	McDonalds per Million (x)	Cantril ladder score (y)
Japan	28.4	5.915
Norway	11.7	7.594
Switzerland	15.6	7.487
United Kingdom	17.8	6.814

Example 6

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 6

- Use the line to predict the Happiness level if the Number of McDonalds per Million people is 18.4.

Example 6

- Circle all values where you should not predict due to extrapolation.

0 5 10 15 20 25 30 35 40

Example 7 – Number of McDonalds

Country	McDonalds per Million	Cantril ladder score (y)
United Kingdom	17.8	6.814
Mexico	1.8	6.488
Australia	32.2	7.272
Canada	33.9	7.328

Example 7

- Find The Least Squares regression line
- Find The Correlation Coefficient
- Is this line a good fit?

Example 7

- Use the line to predict the Happiness level if the Number of McDonalds per Million people is 18.4.

Example 7

- Circle all values where you should not predict due to extrapolation.

0 5 10 15 20 25 30 35 40

Samples

- Don’t forget that samples should be representative of population
- Randomly picked