SNS COLLEGE OF TECHNOLOGY

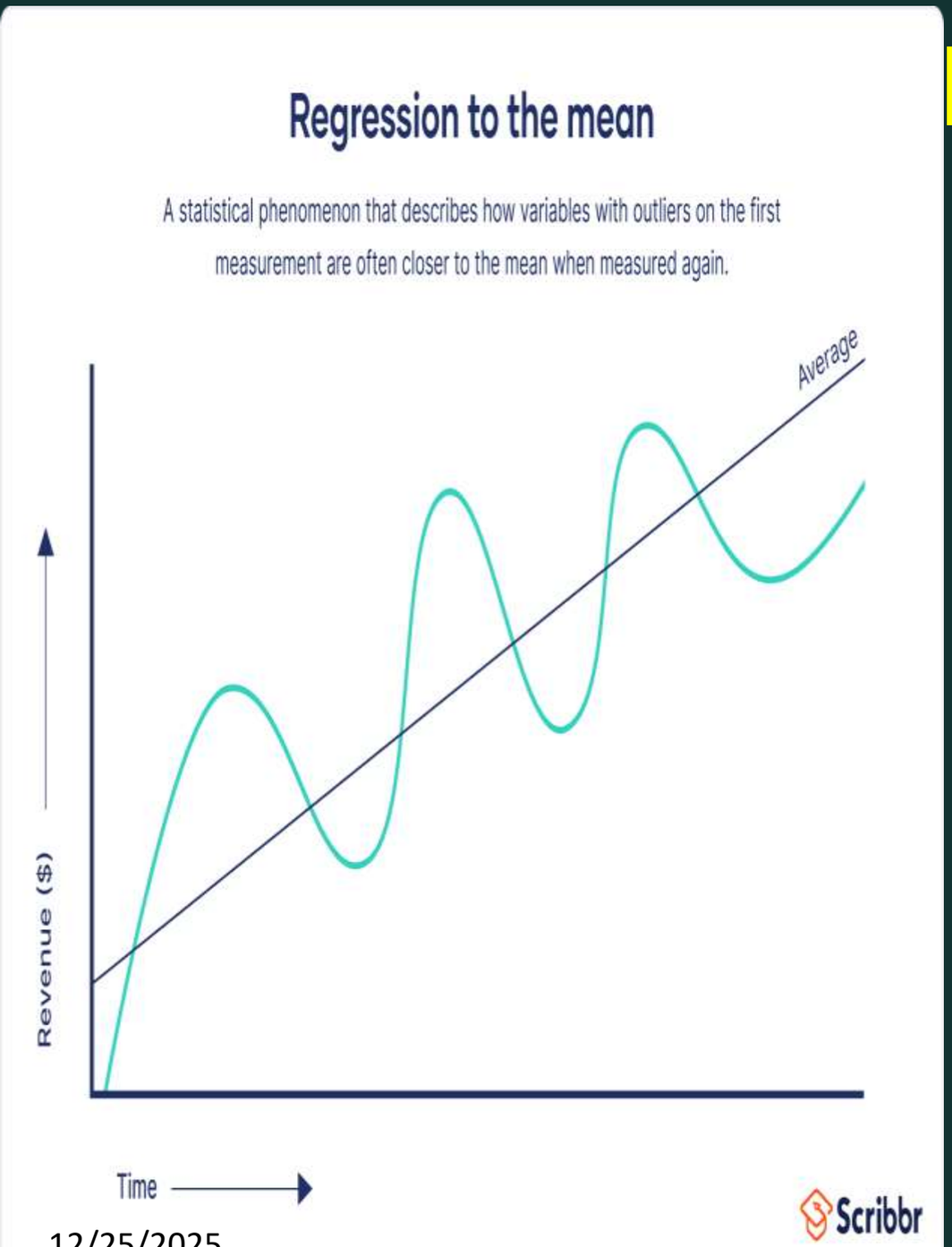
An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT II – Regression toward the mean.



EMPATHY:

Regression to the Mean | Definition & Examples

- Regression to the mean (RTM) is a statistical phenomenon describing how variables much higher or lower than the mean are often much closer to the mean when measured a second time.
- Regression to the mean is due to natural variation or chance.
- It can be observed in everyday life, particularly in research that intentionally focuses on the most extreme cases or events.
- It is sometimes also called regression toward the mean.

DEFINE:

Example: Regression to the mean

- A basketball player scores unusually high in one game but returns to their average scoring in the subsequent match.
- The idea can also apply to entire seasons because you can count a season as a sample of games.
- A rookie can have an unusually good first season and then experience the “sophomore slump,” where performance declines toward the average.
- Conversely, baseball players with subpar batting averages one season tend to improve toward the mean the next season.

IDEATE:

What is regression to the mean?

- Regression to the mean occurs when a nonrandom sample is selected from a population and you measure two imperfectly correlated variables, such as two consecutive blood pressure measurements.
- The smaller the correlation between the two variables, the larger the effect of RTM.
- The more extreme the value from the population mean, the more room there is to regress to the mean.
- Regression to the mean can be explained by considering, for example, that skill and performance are imperfectly correlated due to the role of luck.
- This may lead you to find a causal relationship where there isn't one.

PROTOTYPING:

Why is regression to the mean a problem?

- Regression to the mean can prove problematic particularly in research studies that measure the effectiveness of an intervention, program, or policy.
- It can mislead researchers to believe that an intervention is the cause of an observed change, when in reality it is due to chance.
- This is particularly evident when researchers focus on measurements of people, cases, or organizations at the extremes, such as the worst-performing, the best-educated, or the unhealthiest.
- RTM shows us that, statistically, the lowest cases are likely to improve the second time, while those at their peak will likely perform worse even without the intervention.
- Because it can distort results, you need to take regression to the mean into account when designing your research as well as when analyzing your findings.

TESTING:

Calculating the percent of regression to the mean

- Alternatively, you can calculate the percent of regression to the mean during your data analysis. You can use the formula below to calculate regression to the mean.

Prm = 100(1 – r)wherePrm = percent of regression to the meanr = correlation coefficient

- When there is perfect correlation between your variables, then r = 1

- Using the formula above, Prm = 100(1 – 1) = 0
- In other words, if your variables are perfectly correlated, they will not regress to the mean.

- On the contrary, when there is no correlation (r = 0), there is 100% regression to the mean.

- When there is imperfect correlation between your variables, then r is between –1 and +1.

If your example r = 0.2, there is 80% regression to the mean:

Prm = 100(1 – 0.2)

Prm = 100(0.8)

Prm = 80%