IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION

LECTURES 22

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Most popular model
- Idea is to fit a linear relationship between
  - A quantitative outcome variable (Y) and
  - A set of p predictors $\{X_1, X_2, X_3, \ldots, X_p\}$
- Assumption: relationship as expressed in the following model equation holds true for the target population

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where $\beta_0, \ldots, \beta_p$ are coefficients and ε is the noise or unexplained part

MULTIPLE LINEAR REGRESSION

- Objectives:
  - Understanding the relationships between outcome variable and predictors
    - Followed in statistical approach
  - Predicting values of outcome variable for new records
    - Followed in data mining approach
- Applications in data mining
  - Predicting credit card spending, life of an equipment, sales etc.

MULTIPLE LINEAR REGRESSION

- Model Building and Results Interpretation phases differ depending on the objective:
  - Explanatory (predicting the impact of promotional offer on sales)
  - Predictive (predicting sales)
- Selection of suitable data mining techniques depends on the goal itself

MULTIPLE LINEAR REGRESSION

Explanatory Modeling

- Fits the data closely
- Full sample is used to estimate best-fit model
- Performance metrics measure how close model fits the data

Predictive Modeling

- Predicts new records accurately
- Sample is partitioned into training, validation, and test sets and training partition is used to estimate the model
- Performance metrics measure how well model predicts new observations

MULTIPLE LINEAR REGRESSION

Explanatory Modeling

- Model might not have best predictive accuracy
- Statistical techniques with assumed or hypothesized relationships and scarce data (primary data)

Predictive Modeling

- Model might not be best-fit of data
- Machine learning techniques with no assumed structure and large datasets (secondary data)

MULTIPLE LINEAR REGRESSION

- Estimates for target population
  - Coefficients: $\beta_0, \ldots, \beta_p$ and
  - σ, std. deviation of noise (ε)
  - Cannot be measured directly due to unavailability of data on entire population
- Estimation technique:
  - Ordinary least squares (OLS)
    - Computes the sample estimates which minimize the sum of squared deviations between actual values and predicted values

MULTIPLE LINEAR REGRESSION

- Ordinary least squares (OLS)

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

  – Unbiased predictions (on average, closer to actual values)
  – Smallest average squared error

  Given following assumptions hold true
    - Noise follows a normal distribution
    - Linear relationship holds true
    - Observations are independent
    - Homoskedasticity: variability in the outcome variable is same irrespective of the values of the predictors

MULTIPLE LINEAR REGRESSION

- Partitioning in data mining modeling allows relaxation from the first assumption
- In statistical modeling, same sample is used to fit the model and assess its reliability
  - Predictions of new records lack reliability
  - First assumption is required to derive confidence intervals for predictions
- Example: Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…