IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

# MULTIPLE LINEAR REGRESSION Part-3

## LECTURES 24

**DR. GAURAV DIXIT**

DEPARTMENT OF MANAGEMENT STUDIES

# 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

# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Availability of large no. of variables for selecting a set of predictors
  - Main idea is to select most useful set of predictors for a given outcome variable of interest
  - Selecting all the variables in the model is not recommended
    - Data collection issues in future
    - Measurement accuracy issues for some variables
    - Missing values
    - Parsimony

# MULTIPLE LINEAR REGRESSION

- Variable Selection
  - Selecting all the variables in the model is not recommended
    - Multicollinearity: two or more predictors sharing the same linear relationship with the outcome variable
    - Sample size issues: Rule of thumb

$$n > 5*(p+2)$$

Where n=no. of observations

And p=no. of predictors

    - Variance of predictions might increase due to inclusion of predictors which are uncorrelated with the outcome variable
    - Average error of predictions might increase due to exclusion of predictors which are correlated with the outcome variable

# 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…