IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE LINEAR REGRESSION-PART V
EXHAUSTIVE SEARCH

LECTURE 26

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

MULTIPLE LINEAR REGRESSION

- Steps to reduce the no. of predictors
  - Summary statistics and graphs
  - Statistical methods using computational power
    - Exhaustive search: all possible combinations
    - Partial-iterative search: algorithm based
- Exhaustive Search
  - Large no. of subsets
  - Criteria to compare models
    - Adjusted R^2

MULTIPLE LINEAR REGRESSION

- Adjusted R^2

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

Where R^2 is proportion of explained variability in the model

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(yi - \hat{y}_i)}{\sum_{i=1}^{n}(yi - \bar{y})}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- R^2 is called coefficient of determination

MULTIPLE LINEAR REGRESSION

- R^2 would be equal to squared correlation in a single predictor model, that is how R^2 gets its name

- Adjusted R^2 introduces a penalty on the no. of predictors to trade-off between artificial increase vs. amount of information

- High adjusted R^2 values -> low $\hat{\sigma}^2$

MULTIPLE LINEAR REGRESSION

- Exhaustive Search
  - Criteria to compare models
    - Mallow's C_p
- Mallow's C_p

$$C_p = \frac{SSR}{\widehat{\sigma_f}^2} + 2(p+1) - n$$

Where $\widehat{\sigma_f}^2$ is estimated value of σ^2 in the full model

and SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})$

MULTIPLE LINEAR REGRESSION

- Mallow's C_p
  - Assumption: full model with all predictors is unbiased
    - Predictors elimination would reduce the variability
  - Best subset model would have C_p ~ p+1 and p would be a small value
  - Requires high n value for the training partition relative to p
- Open RStudio

MULTIPLE LINEAR REGRESSION

- Partial-iterative search
  - Computationally cheaper
  - Best subset is not guaranteed
    - Potential of missing “good” sets of predictors
  - Produce close-to-best subsets
  - Preferred approach for large no. of predictors
  - For moderate no. of predictors, exhaustive search is better
- Trade-off between computation cost vs. potential of finding best subset

MULTIPLE LINEAR REGRESSION

- Partial-iterative search algorithms
  - Forward selection
    - Add predictors one by one
    - Strength as a single predictor is used
  - Backward elimination
    - Drop predictors one by one
  - Stepwise regression
    - Add predictors one by one and consider dropping insignificant ones
- 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…