IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION- PART 2

LECTURE 47

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

This function is called *logistic response function*

LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1-P} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1-P}$$

  - Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
    - This metric is popular in sports, horse racing, gambling, and many other areas

LOGISTIC REGRESSION

- Logit
  - Previous equation can be rewritten as

$$odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

    - Range is now (0, ∞)
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

    - Standard logistic model
    - Now, LHS and RHS both have same range (-∞, ∞)
  - Log(odds) is called logit
    - Logit is used as the outcome variable in the model instead of categorical Y

LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio

- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task

LOGISTIC REGRESSION

- Estimation Technique
  - Least squares method used in multiple linear regression cannot be used
    - Non-linear formulation of logistic regression
  - Maximum likelihood method is used
    - Estimates are optimized in order to maximize the likelihood of obtaining the observations used in training the model
    - Less robust than estimation techniques used in linear regression
    - Reliability of estimates
      - Outcome variable categories should have adequate proportion
      - Adequate sample size w.r.t no. of estimates

LOGISTIC REGRESSION

- Estimation Technique
  - Maximum likelihood method is used
    - Collinearity issues similar to linear regression
- Interpretation of Results
  - Logit model
    - Additive factor (β)
      - If $\beta < 0$, increase in x => decrease in logit values
      - If $\beta > 0$, increase in x => increase in logit values
    - For any value of x, interpretative statements of results are same

Thanks…