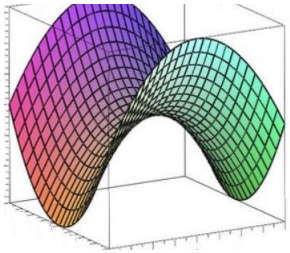
Course Code 009973 (Fall 2021)

Special Topics in Software 2

Statistical Learning



Linear Regression Review

Prof. S. M. Riazul Islam, Dept. of Computer Engineering, Sejong University, Korea

E-mail: riaz@sejong.ac.kr

Linear Regression

Reviews on

- Multiple Linear Regression
  - Estimating the Regression Coefficients

- Linear Regression Demonstration in Python

Linear Regression Review

Multiple Linear Regression

Example: A multiple linear regression model with k predictor variables $X_1, X_2, ..., X_k$ and a response Y, can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_k x_k + \epsilon.$$

As before, the ϵ are the residual terms of the model and the distribution assumption we place on the residuals will allow us later to do inference on the remaining model parameters. Interpret the meaning of the REGRESSION COEFFICIENTS $\beta_0, \beta_1, \beta_2, ..., \beta_k$ in this model.

More complex models may include higher powers of one or more predictor variables, e.g.,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \qquad (1)$$

or interaction effects of two or more variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \qquad (2)$$

Linear Regression Review

Multiple Linear Regression: Model Parameters Estimation

The setup: Consider a multiple linear regression model with k independent predictor variables $x_1, \ldots, x_k$ and one response variable y.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Suppose, we have n observations on the $k + 1$ variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

n should be bigger than k. Why?

Linear Regression Review

Multiple Linear Regression: Model Parameters Estimation

The setup: Consider a multiple linear regression model with k independent predictor variables $x_1, \ldots, x_k$ and one response variable y.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

Suppose, we have n observations on the $k+1$ variables.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

n should be bigger than k. Why?

the sum of squared residuals.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Linear Regression Review

Multiple Linear Regression: Model Parameters Estimation

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad i = 1, \ldots, n$$

Let

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

With this compact notation, the linear regression model can be written in the form

$$\boxed{\mathbf{y} = \mathbf{X}\beta + \epsilon}$$

Linear Regression Review

Multiple Linear Regression: Model Parameters Estimation

In linear algebra terms, the least-squares parameter estimates β are the vectors that minimize

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The vector of fitted values $\hat{\mathbf{y}}$ in a linear regression model can be expressed as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

$$H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The regression residuals can be written in different ways as

$$\epsilon = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Q&A