Regression Analysis

Analysis of Variance

Nicoleta Serban, Ph.D.
Professor
School of Industrial and Systems Engineering

ANOVA vs Simple Linear Regression

About This Lesson

ANOVA & Linear Regression

Simple Linear Regression:
Data: $\{(x_i, y_i), \ldots, (x_n, y_n)\}$
Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i;\ i = 1, \cdots, n$

ANOVA: A linear regression model where the predicting factor is a categorical variable.

ANOVA:
Data: Y_{ij} for $j = 1, \cdots, n;\ i = 1, \cdots, k$
Model: $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$ where $\sum_{i=1}^{k} \tau_i = 0$ and
$\mu_i = i^{\text{th}}$ group mean decomposed into $\mu_i = \mu + \tau_i$

Georgia Tech

ANOVA & Linear Regression (cont'd)

ANOVA:

Data: Y_{ij} for $j = 1, \cdots, \mathrm{n};\ i = 1, \cdots, \mathrm{k}$

Model: $Y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \tau_i + \varepsilon_{ij}$ where $\sum_{i=1}^{k} \tau_i = 0$ and

$\mu_i = i^{\text{th}}$ group mean decomposed into $\mu_i = \mu + \tau_i$

Define Y to be the response variable (as a single column vector):

$$Y = \left(Y_{11}, \cdots, Y_{1\mathrm{n}_1}, Y_{21}, \cdots, Y_{2\mathrm{n}_2}, \cdots, Y_{\mathrm{k}1}, \cdots, Y_{\mathrm{kn}_\mathrm{k}}\right)$$

Define L to be the label/categorical variable (as a single column vector):

$$L = \left(L_{11}, \cdots, L_{1\mathrm{n}_1}, L_{21}, \cdots, L_{2\mathrm{n}_2}, \cdots, L_{\mathrm{k}1}, \cdots, L_{\mathrm{kn}_\mathrm{k}}\right)$$

Linear Regression: Y ~ L

Georgia Tech

ANOVA & Linear Regression (cont'd)

Categorical Variables in Linear Regression:

- Transform categories into column vector dummy variables

  $X_1 = (1,..,1,0,...,0)$; $X_2 = (0,...,0,1,..,1,0,...,0)$; ... ; $X_k = (0,...,0,1,..,1)$

  Each X_i has length N where $N = n_1 + n_2 + ... + n_k$. Each 1 indicates a Y value with the corresponding L label.a

- Let r index the rows of the column vectors X_is and Y. If intercept in model, define k-1 dummy variables because of linear dependence: $(1,...,1) = X_1 + X_2 + ... + X_k$

  Model: $Y_r = \beta_0 + \beta_1 X_{1_r} + \beta_2 X_{2_r} + \cdots + \beta_{k-1} X_{k-1_r} + \varepsilon_r;\ r = 1, \cdots, N$

- If no intercept in the model, define all k dummy variables

  Model: $Y_r = \beta_1 X_{1_r} + \beta_2 X_{2_r} + \cdots + \beta_k X_{k_r} + \varepsilon_r;\ r = 1, \cdots, N$

ANOVA: A linear regression model with multiple predictors Multiple Linear Regression

Georgia Tech

Summary