Indicator (Dummy) Variables – Part III

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 6 – Lecture 6B

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

1-Way ANOVA: One Qualitative Variable in Regression

Regression Approach to Analysis of Variance (ANOVA)

ANOVA is a technique frequently used to analyze data from planned or designed experiments.

1-Way ANOVA: One Qualitative Variable in Regression

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n$$

y_{ij}: jth obs for ith treatment (or factor level)
μ: grand mean
τ_i: the effect of ith treatment
ε_{ij} is NIND$(0, \sigma^2)$

In the balanced case (i.e., equal number of observations per treatment), $\sum_{i=1}^{k} \tau_i = 0$

1-Way ANOVA: One Qualitative Variable in Regression

The mean of the *i*th treatment is

$$\mu_i = \mu + \tau_i \quad , \quad i = 1, \ldots, k$$

The average of *n* observations in the *i*th treatment:

$$\bar{y}_{i.} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}, \quad i = 1, 2, \ldots, k$$

The grand average is

$$\bar{y}_{..} = \frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}$$

1-Way ANOVA: One Qualitative Variable in Regression

How to use indicator variables to perform regression analysis to obtain ANOVA?

Without loss of generality, let us assume that $k = 3$.

Define

$$x_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the observation is from treatment 2} \\ 0 & \text{otherwise} \end{cases}$$

1-Way ANOVA vs. Regression Model

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, \ldots, n$$

For treatment 1:

$$y_{1j} = \beta_0 + \beta_1(1) + \beta_2(0) + \varepsilon_{1j} = \beta_0 + \beta_1 + \varepsilon_{1j}$$

$$\beta_0 + \beta_1 = \mu_1$$

1-Way ANOVA vs. Regression Model

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, 2, \ldots, n$$

For treatment 2:

$$y_{2j} = \beta_0 + \beta_1(0) + \beta_2(1) + \varepsilon_{2j} = \beta_0 + \beta_2 + \varepsilon_{2j}$$

$$\beta_0 + \beta_2 = \mu_2$$

For treatment 3:

$$y_{3j} = \beta_0 + \beta_1(0) + \beta_2(0) + \varepsilon_{3j} = \beta_0 + \varepsilon_{3j}$$

$$\beta_0 = \mu_3$$

1-Way ANOVA vs. Regression Model

Consequently,

$\beta_0 = \mu_3$

$\beta_1 = \mu_1 - \mu_3$

$\beta_2 = \mu_2 - \mu_3$

Assume that

$n = 3$

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix}, \quad \mathbf{X} = \begin{array}{c} \begin{array}{ccc} & x_1 & x_2 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \end{array}$$

Regression Analysis for ANOVA

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

$$\begin{bmatrix} 9 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

$$\hat{\beta}_0 = \bar{y}_{..} - \bar{y}_{1.} - \bar{y}_{2.} = \bar{y}_{3.}, \quad \hat{\beta}_1 = \bar{y}_{1.} - \bar{y}_{3.}, \quad \hat{\beta}_2 = \bar{y}_{2.} - \bar{y}_{3.}$$

Regression Analysis for ANOVA

$$SS_R\left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\right) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = [\bar{y}_{3.}, \bar{y}_{1.} - \bar{y}_{3.}, \bar{y}_{2.} - \bar{y}_{3.}]\begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

$$= y_{..}\bar{y}_{3.} + y_{1.}(\bar{y}_{1.} - \bar{y}_{3.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{3.})$$

$$= (y_{1.} + y_{2.} + y_{3.})\bar{y}_{3.} + y_{1.}(\bar{y}_{1.} - \bar{y}_{3.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{3.})$$

$$= \bar{y}_{1.}y_{1.} + \bar{y}_{2.}y_{2.} + \bar{y}_{3.}y_{3.}$$

$$= \sum_{i=1}^{3} \frac{y_{i.}^2}{3}$$

Regression Analysis for ANOVA

$$SS_{\text{Res}} = \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}^2 - SS_{\text{R}}(\beta_0, \beta_1, \beta_2)$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}^2 - \sum_{i=1}^{3} \frac{y_{i.}^2}{3}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} \left(y_{ij} - \bar{y}_{i.}\right)^2$$

Regression Analysis for ANOVA

To test $H_0: \tau_1 = \tau_2 = \tau_3 = 0$, equivalently,

$\beta_0 = \mu, \quad \beta_1 = 0, \quad \beta_2 = 0$, the reduced model: $y_{ij} = \beta_0 + \varepsilon_{ij}$

$$SS_R(\beta_1, \beta_2 \mid \beta_0) = SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0)$$
$$= \sum_{i=1}^{3} \frac{y_{i.}^2}{3} - \frac{y_{..}^2}{9}$$
$$= 3\sum_{i=1}^{3} (\bar{y}_{i.} - \bar{y}_{..})^2$$

Regression Analysis for ANOVA

The F test is

$$F_0 = \frac{SS_R(\beta_1, \beta_2 | \beta_0)/2}{SS_{Res}/6}$$

$$= \frac{3\sum_{i=1}^{3}(\bar{y}_{i.} - \bar{y}_{..})^2/2}{\sum_{i=1}^{3}\sum_{j=1}^{3}(y_{ij} - \bar{y}_{i.})^2/6}$$

$$= \frac{MS_{Treatments}}{MS_{Res}}$$

Regression Analysis for ANOVA

TABLE 8.4 One-Way Analysis of Variance

Degrees of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Treatments	$n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$	$k-1$	$\frac{SS_{\text{Treatments}}}{k-1}$	$\frac{MS_{\text{Treatments}}}{MS_{\text{Res}}}$
Error	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$	$k(n-1)$	$\frac{SS_{\text{Res}}}{k(n-1)}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2$	$kn-1$		

The regression approach using appropriate dummy variables is identical to 1-way ANOVA

JOHNS HOPKINS

WHITING SCHOOL *of* ENGINEERING