Regression Analysis

Analysis of Variance

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

Parameter Estimation

Georgia Tech

1

About This Lesson

Georgia Tech

2

ANOVA: Model & Assumptions

Data: Y_{ij} for $j = 1, \cdots, n_i; i = 1, \cdots, k$
Model: $Y_{ij} = \mu_i + \varepsilon_{ij}$ where ε_{ij} = error term

Assumptions:

- ***Constant Variance Assumption:*** Var(ε_{ij}) = σ^2
- ***Independence Assumption:*** $\{\varepsilon_{1j}, \cdots, \varepsilon_{kj}\}$ are independent random variables
- ***Normality Assumption:*** ε_{ij} ~ Normal (0, σ^2)

Georgia Tech

3

ANOVA: Variance Estimation

Comparing means from multiple populations assuming the variances are the same and equal to σ^2:

Pooled Variance Estimator:

$$S^2_{\text{pool}} = \boxed{\frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)}} = \boxed{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2}{N - k}}$$

Where N = total number of samples = (n_1 +…+ n_k)

The degrees of freedom is N-k because we replace $\mu_i \leftarrow \bar{Y}_i$ for i = 1,…,k, thus losing k degrees of freedom

Georgia Tech

4

ANOVA: Variance Estimation (cont'd)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}(n_i - 1)S_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{i_j} - \bar{Y}_i\right)^2}{N - k} = \underline{\textbf{MSE}}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{i_j} - \bar{Y}_i\right)^2 = \text{\underline{S}um of \underline{S}quares of \underline{E}rror} = \underline{\textbf{SSE}}$$

We will use interchangeably Sum of Squared Errors and Sum of Squared Residuals.

5

Mean Squared Error (MSE)

$S_1^2, \ldots, S_k^2$ *The sum of independent chi-square random variables is also chi-square*

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \cdots + \frac{(n_k - 1)S_k^2}{\sigma^2} \sim \chi_\nu^2 \quad \text{where } \nu = N - k$$

The sampling distribution of the pooled variance is a chi-square distribution with N-k degrees of freedom.

6

Estimating Parameters in ANOVA

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

What is the sampling distribution?

If $Y_{i1},\ldots,Y_{in} \sim \mathrm{N}(\mu_i, \sigma^2)$ $\Rightarrow$ $\hat{\mu}_i = \bar{Y}_i = \frac{Y_{i1}+\ldots+Y_{in}}{n_i} \sim \mathrm{N}(\mu_i,\, \sigma^2/n_i)$

But σ^2 is unknown.
So replace σ^2 with the pooled variance estimation: $\sigma^2 \Leftarrow \mathrm{MSE}$

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\mathrm{MSE}/n_i}} \sim \mathrm{t}_{\mathrm{N}-k}$$

Why N – k?

$$\mathrm{MSE} = \hat{\sigma}^2 \sim \chi^2_{\mathrm{N}-k}$$

Georgia Tech

7

Confidence Intervals for the Means

We can use the estimated sample means

$$\hat{\mu}_i = \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \text{ for } i = 1, \cdots, k$$

and the estimated variance

$$\hat{\sigma}^2 = \mathrm{MSE}$$

to calculate (1 - α) confidence intervals for the treatment means:

$$\left(\hat{\mu}_i - \mathrm{t}_{\alpha/2,\,\mathrm{N}-k}\sqrt{\mathrm{MSE}/n_i},\, \hat{\mu}_i + \mathrm{t}_{\alpha/2,\,\mathrm{N}-k}\sqrt{\mathrm{MSE}/n_i}\right)$$

Georgia Tech

8

9