

# Regression Analysis

## Analysis of Variance

**Nicoleta Serban, Ph.D.**

*Professor*

School of Industrial and Systems Engineering

Parameter Estimation



1

## About This Lesson



2

# ANOVA: Model & Assumptions

**Data:**  $Y_{ij}$  for  $j = 1, \dots, n_i; i = 1, \dots, k$

**Model:**  $Y_{ij} = \mu_i + \varepsilon_{ij}$  where  $\varepsilon_{ij}$  = error term

## Assumptions:

- **Constant Variance Assumption:**  $\text{Var}(\varepsilon_{ij}) = \sigma^2$
- **Independence Assumption:**  $\{\varepsilon_{1j}, \dots, \varepsilon_{kj}\}$  are independent random variables
- **Normality Assumption:**  $\varepsilon_{ij} \sim \text{Normal}(0, \sigma^2)$



3

# ANOVA: Variance Estimation

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



*Pooled Variance Estimator:*

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

Where  $N$  = total number of samples =  $(n_1 + \dots + n_k)$

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



4

## ANOVA: Variance Estimation (cont'd)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^k (n_i - 1) S_i^2}{\sum_{i=1}^k (n_i - 1)} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2}{N - k} = \text{MSE}$$

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 = \text{Sum of Squares of Error} = \text{SSE}$$

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



5

## Mean Squared Error (MSE)

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

$$\frac{\text{SSE}}{\sigma^2} = \frac{(n_1 - 1) S_1^2}{\sigma^2} + \dots + \frac{(n_k - 1) S_k^2}{\sigma^2} \sim \chi_v^2 \text{ where } v = N - k$$

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

But  $\sigma^2$  is unknown.

So replace  $\sigma^2$  with the pooled variance estimation:

$$\sigma^2 \leftarrow \text{MSE}$$

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\text{MSE}/n_i}} \sim t_{N-k}$$

Why  $N - k$ ?

$$\text{MSE} = \hat{\sigma}^2 \sim \chi_{N-k}^2$$



7

## Confidence Intervals for the Means

We can use the estimated sample means

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

and the estimated variance

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

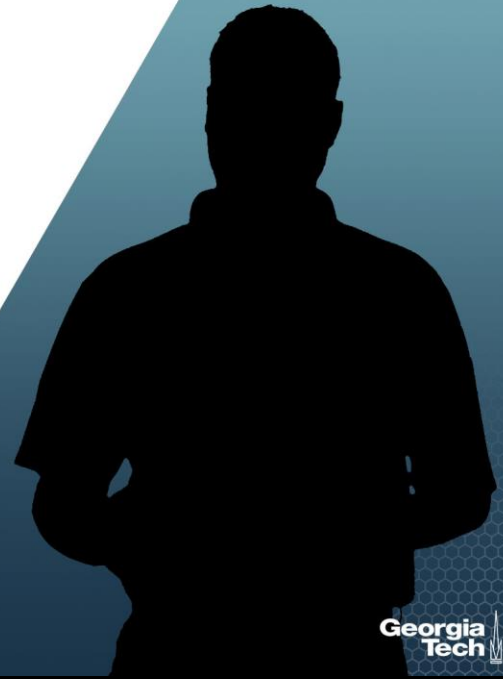
to calculate  $(1 - \alpha)$  confidence intervals for the treatment means:

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



8

# Summary



Georgia  
Tech