Regression Analysis

Analysis of Variance

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

Basics Concepts

About This Lesson

ANOVA: Analysis of Variance

Population 1: (μ_1, σ_1^2) → Sample 1:$(Y_{1,1},\ldots,Y_{1,n_1})$ → $(\bar{Y}_1, s_1^2)$

Population 2: (μ_2, σ_2^2) → Sample 2: $(Y_{2,1},\ldots,Y_{2,n_2})$ → $(\bar{Y}_2, s_2^2)$

.....

Population k: (μ_k, σ_k^2) → Sample k: $(Y_{k,1},\ldots,Y_{k,n_k})$ → $(\bar{Y}_k, s_k^2)$

ANOVA: Comparing the means of multiple samples

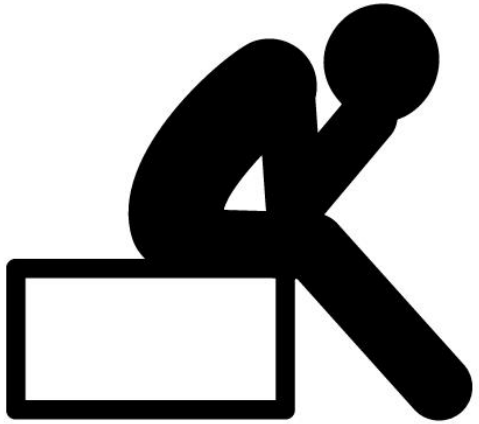
Georgia Tech

ANOVA Example 1: Global Suicide

Data Source:

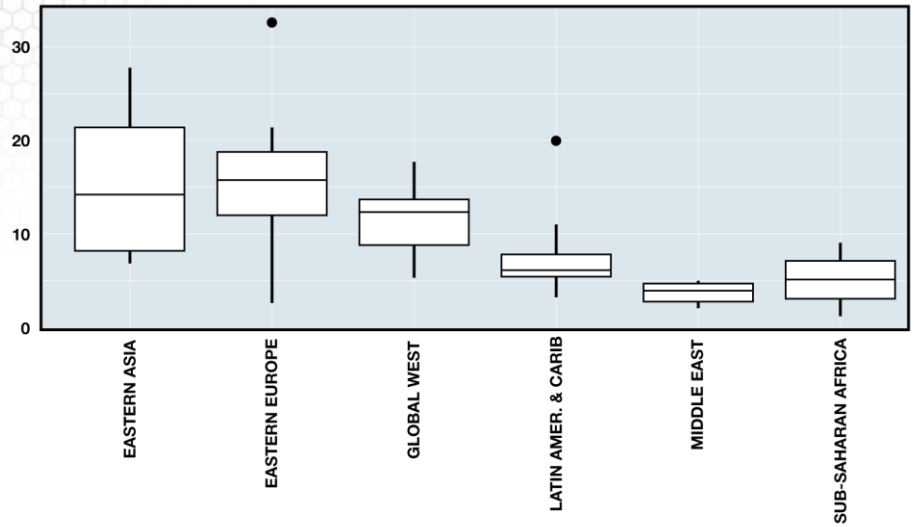
Suicide Rate: Kaggle

https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016



Georgia Tech

ANOVA Example 1: Suicide Rate & Region



1. Is there a difference in the suicide rate by region?

2. Which region has higher suicide rate?

Georgia Tech

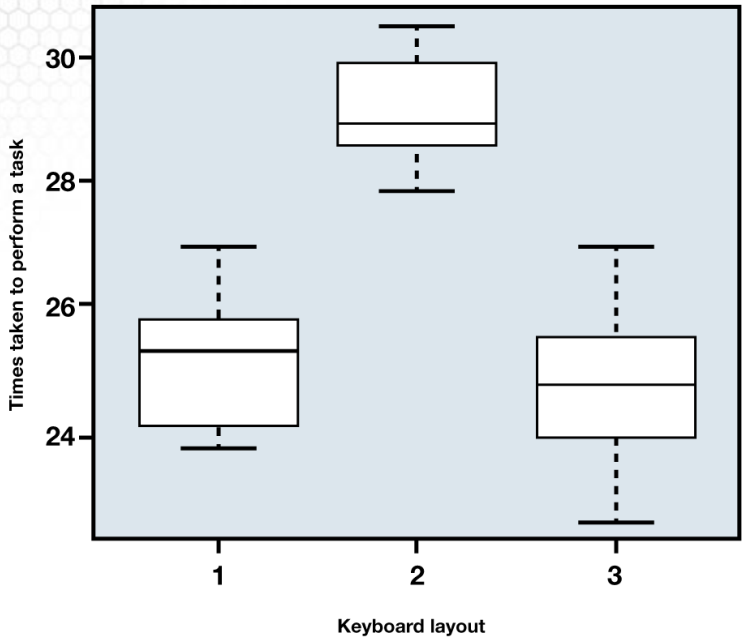
ANOVA Example 2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.

Layout 1	Layout 2	Layout 3
23.8	30.2	27.0
25.6	29.9	25.4
24.0	29.1	25.6
25.1	28.8	24.2
25.5	29.1	24.8
26.1	28.6	24.0
23.8	28.3	25.5
25.7	28.7	23.9
24.3	27.9	22.6
26.0	30.5	26.0
24.6	*	23.4
27.0	*	*

Georgia Tech

Operation Time by Keyboard Layout



1. Is there a difference in the time taken to perform a task?

2. Which layout is more effective?

Georgia Tech

ANOVA: Objectives

Primary objectives in ANOVA:

1. Analysis of the variability in the data – the ANOVA table
2. Testing for equal means

$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$

3. Estimation of simultaneous confidence intervals for the mean differences

$$\mu_i - \mu_j \quad \textit{for } i \textit{ and } j = 1, \ldots, k$$

Georgia Tech

Summary

Regression Analysis

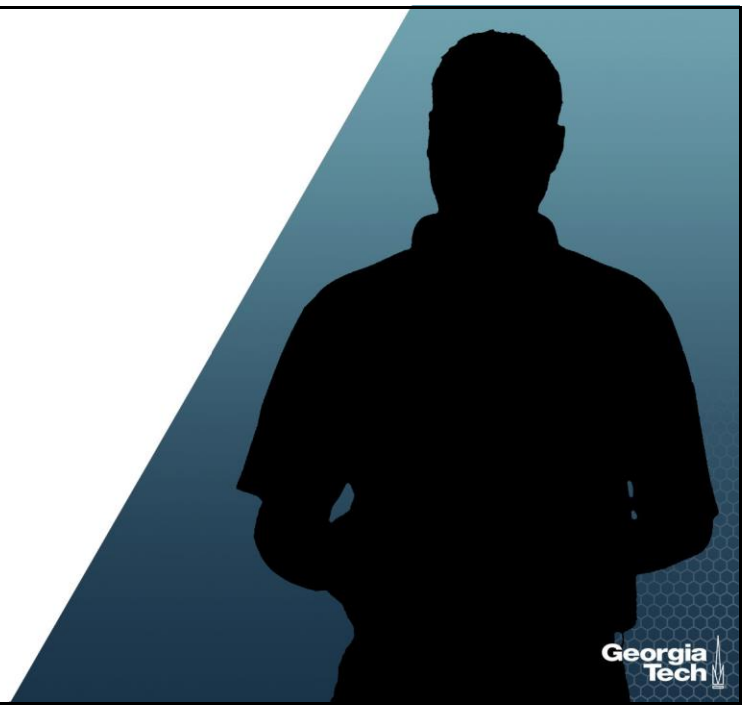
Analysis of Variance

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 $j = 1, \cdots, n_i; i = 1, \cdots, k$
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:

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

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)

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

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

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

Georgia Tech

5

Mean Squared Error (MSE)

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

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

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

Georgia Tech

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

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

Summary

9

Regression Analysis

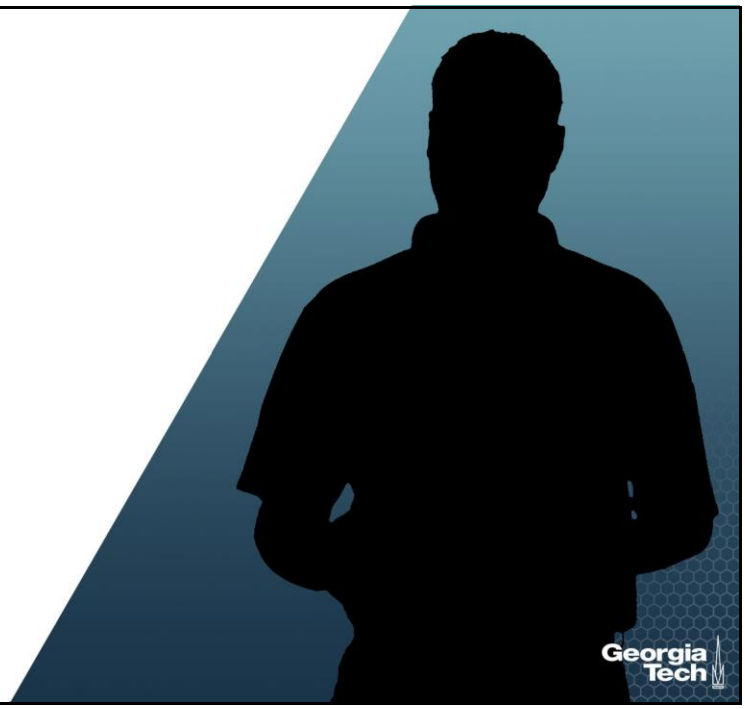
Analysis of Variance

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

Parameter Estimation
Examples

1

About This Lesson



2

Example 1: Global Suicide by Region

What are the estimates for the mean suicide rates for the different regions?

Georgia Tech

3

Parameter Estimation

model = aov(suicidesper100k ~ region, data=reg_data)
model.tables(model, type = "means")

Overall Mean: 10.276

$\hat{\mu}_{easia} = 10.29$	$n_{easia} = 10$
$\hat{\mu}_{wasia} = 0.58$	$n_{wasia} = 1$
$\hat{\mu}_{eeurope} = 17.41$	$n_{eeurope} = 15$
$\hat{\mu}_{weurope} = 12.75$	$n_{weurope} = 3$
$\hat{\mu}_{west} = 11.68$	$n_{west} = 18$
$\hat{\mu}_{lamerica} = 7.86$	$n_{lamerica} = 26$
$\hat{\mu}_{mideast} = 2.46$	$n_{mideast} = 7$
.....	

Georgia Tech

4

Example 2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.

What are the estimates for the mean typing times for the different groups of keyboards?

Layout 1	Layout 2	Layout 3
23.8	30.2	27.0
25.6	29.9	25.4
24.0	29.1	25.6
25.1	28.8	24.2
25.5	29.1	24.8
26.1	28.6	24.0
23.8	28.3	25.5
25.7	28.7	23.9
24.3	27.9	22.6
26.0	30.5	26.0
24.6	*	23.4
27.0	*	*

Georgia Tech

5

Parameter Estimation

model = aov(speed ~ layout) model.tables(model, type = "means")

Tables of means
Grand mean
26.21212
Layout

	1	2	3
	25.12	29.11	24.76
rep	12.00	10.00	11.00

Overall Mean: 26.21212

$\hat{\mu}_{layout1} = 25.12$
$\hat{\mu}_{layout2} = 29.11$
$\hat{\mu}_{layout3} = 24.76$

Georgia Tech

6

Summary

7

Regression Analysis

Analysis of Variance

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

Hypothesis Test for Equal Means

About This Lesson

Hypothesis Test for Equal Means

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

H_A : some means are different

Georgia Tech

Null Hypothesis

- Under the null hypothesis, combine k samples to estimate the overall mean with the overall sample mean (grand mean) $\bar{Y}$:

$$\bar{Y} = \frac{1}{\mathrm{N}} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

- Base the null hypothesis variance estimate S_0^2 on this overall sample mean:

$$\mathrm{S}_0^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2}{\mathrm{N}-1} = \frac{\mathrm{SST}}{\mathrm{N}-1}$$

- **SST** = **S**um of **S**quares **T**otal = $\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}\right)^2$

- Because we only estimate one mean, we lose only 1 df (unlike pooled variance)

$$\frac{(\mathrm{N}-1)\mathrm{S}_0^2}{\sigma^2} = \frac{\mathrm{SST}}{\sigma^2} \sim \chi^2_{\mathrm{N}-1}$$

Georgia Tech

SST Decomposition

We can *partition* SST into two separate parts:

$$\mathbf{SST} = \mathbf{SSE} + \mathbf{SST_R}$$

where $\mathbf{SST_R}$ = **S**um of **S**quares of **Tr**eatments $= \sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y})^2$, and $\bar{Y}_i$ is the ith sample mean.

Recall:

$$\mathbf{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}\right)^2$$

$$\mathbf{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2$$

1. MSE = SSE/(N-k) = *within-group variability*
2. $MSST_R = SST_R/(k-1)$ = *between-group variability*
3. ANOVA: comparing *between* to *within* variability
4. F = between-group variability / within-group variability

Georgia Tech

Testing Equal Variances with F-Test

$$\frac{\mathrm{SST_R}/(k-1)}{\mathrm{SSE}/(\mathrm{N}-k)} \equiv \frac{\mathrm{MST_R}}{\mathrm{MSE}} = \mathrm{F}_0 \sim \mathrm{F}_{k-1,\mathrm{N}-k}$$

if H_o is true

Reject H_o if $\mathbf{F_0 > F_a(k\text{-}1,N\text{-}k)}$, which is the upper α^{th} quantile of the F distribution.

P-value for the F-test = $P(F > F_0)$, where $F \sim F_{(k-1,N-k)}$

Georgia Tech

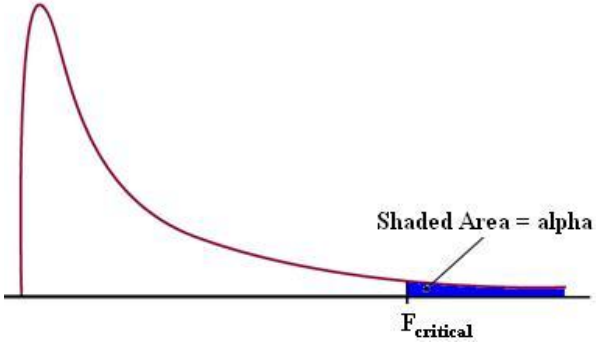
Testing Equal Variances with F-Test

$$\frac{\mathrm{SST_R}/(k-1)}{\mathrm{SSE}/(\mathrm{N}-k)} \equiv \frac{\mathrm{MST_R}}{\mathrm{MSE}} = \mathrm{F}_0 \sim \mathrm{F}_{k-1,\mathrm{N}-k}$$

if H_o is true

Reject H_o if $\mathbf{F_0 > F_a(k\text{-}1,N\text{-}k)}$, which is the upper α^{th} quantile of the F distribution.

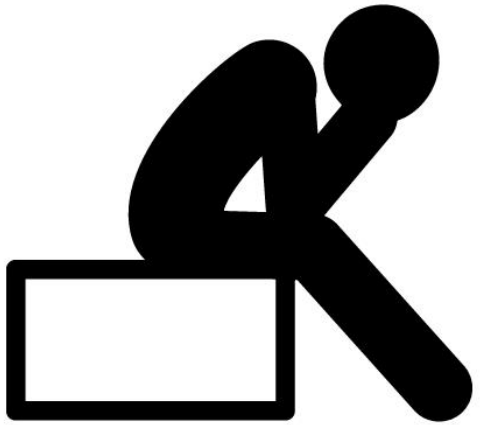
P-value for the F-test = $P(F > F_0)$, where $F \sim F_{(k-1,N-k)}$



Georgia Tech

Example 1: Global Suicide by Region

Are the mean suicide rates equal across the different country regions?



Testing for Equal Means

summary(aov(suicidesper100k ~ region, data=suicide_data))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region	9	1548	172.06	4.767	4.71e-05 ***
Residuals	77	2779	36.09		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SST_R = 1548
k-1 = 9
SSE = 2779
N-k = 77
F-value = 4.767
P-value = 4.71e-05

P-value ≈ 0:
Reject the null hypothesis of equal mean heights

Georgia Tech

Example 2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.

Are the mean typing times for the three keyboard layouts statistically different?

Layout 1	Layout 2	Layout 3
23.8	30.2	27.0
25.6	29.9	25.4
24.0	29.1	25.6
25.1	28.8	24.2
25.5	29.1	24.8
26.1	28.6	24.0
23.8	28.3	25.5
25.7	28.7	23.9
24.3	27.9	22.6
26.0	30.5	26.0
24.6	*	23.4
27.0	*	*

Georgia Tech

Testing for Equal Means

summary(aov(speed ~ layout))

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
layout	2	121.24	60.62	52.84	1.48e-10 ***
Residuals	30	34.42	1.15		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SSTR = 121.24
k-1 = 2
SSE = 34.42
N-k = 30
F-value = 52.84
P-value = 1.48e-10

P-value ≈ 0:
Reject the null hypothesis of equal mean typing times

Georgia Tech

Summary

Regression Analysis

Analysis of Variance

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

Mean Pairwise Comparison

1

About This Lesson

2

Pairwise Comparison of Means

One primary goal of ANOVA might be to determine which treatment means are bigger or smaller. One way to do this is to compare all *k(k-1)/2* pairs of treatments. For a (1 - α) confidence interval for the mean difference $\mu_i - \mu_j$:

$$\left(\hat{\mu}_i - \hat{\mu}_j\right) \pm q_{\alpha,k,N-k}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

- Estimate of difference in means
- α percentile of "studentized range" distribution
- Standard deviation/error of estimator

Difference Between t_α and q_α

Correct for simultaneous inference:

- $q > t$ (at any fixed α and df)
- Intervals are wider to compensate for the fact that we are making simultaneous comparisons (multiplicity correction)

Why?

95% CIs for two populations $\Rightarrow$ (.95)(.95) ≈ .90 $\Rightarrow$ The simultaneous or joint confidence level for the two parameters is roughly **90%**.

95% CIs for three populations $\Rightarrow$ (.95)(.95)(.95) ≈ .86 $\Rightarrow$ The simultaneous or joint confidence level for the three parameters is roughly **86%**.

Example1: Global Suicide by Region

Which country regions have different suicide rates?

Georgia Tech

5

Pairwise Comparison

TukeyHSD(aov(suicidesper100k ~ region, data=suicide_data))

Tukey multiple comparisons of means

95% family-wise confidence level

$region

	diff	lwr	upr	p adj
EASTERN EUROPE-ASIA	7.1256986	-0.8654681	15.1168654	0.1218931
GLOBAL WEST-ASIA	1.3948384	-6.3253621	9.1150390	0.9998655
LATIN AMER. & CARIB-ASIA	-2.4242761	-9.7079484	4.8593961	0.9848625
MIDDLE EAST-ASIA	-7.8183246	-17.4646356	1.8279865	0.2171605
NORTHERN AMERICA-ASIA	1.8826591	-18.6470201	22.4123382	0.9999996
OCEANIA-ASIA	-0.6423728	-13.5277421	12.2429965	1.0000000
SUB-SAHARAN AFRICA-ASIA	-4.2457218	-17.1310911	8.6396474	0.9858800
WESTERN ASIA-ASIA	-9.6996143	-30.2292935	10.8300649	0.8717761
WESTERN EUROPE-ASIA	2.4643324	-10.4210369	15.3497016	0.9997844
GLOBAL WEST-EASTERN EUROPE	-5.7308602	-12.5740866	1.1123662	0.1809537
LATIN AMER. & CARIB-EASTERN EUROPE	-9.5499748	-15.8966379	-3.2033117	0.0002123
MIDDLE EAST-EASTERN EUROPE	-14.9440232	-23.9039098	-5.9841367	0.000026

……..

Georgia Tech

6

Pairwise Comparison

TukeyHSD(aov(suicidesper100k ~ region, data=suicide_data))

Tukey multiple comparisons of means

95% family-wise confidence level

$region

	diff	lwr	upr	p adj
EASTERN EUROPE-ASIA	7.1256986	-0.8654681	15.1168654	0.1218931
GLOBAL WEST-ASIA	1.3948384	-6.3253621	9.1150390	0.9998655
LATIN AMER. & CARIB-ASIA	-2.4242761	-9.7079484	4.8593961	0.9848625
MIDDLE EAST-ASIA	-7.8183246	-17.4646356	1.8279865	0.2171605
NORTHERN AMERICA-ASIA	1.8826591	-18.6470201	22.4123382	0.9999996
OCEANIA-ASIA	-0.6423728	-13.5277421	12.2429965	1.0000000
SUB-SAHARAN AFRICA-ASIA	-4.2457218	-17.1310911	8.6396474	0.9858800
WESTERN ASIA-ASIA	-9.6996143	-30.2292935	10.8300649	0.8717761
WESTERN EUROPE-ASIA	2.4643324	-10.4210369	15.3497016	0.9997844
GLOBAL WEST-EASTERN EUROPE	-5.7308602	-12.5740866	1.1123662	0.1809537
LATIN AMER. & CARIB-EASTERN EUROPE	-9.5499748	-15.8966379	-3.2033117	0.0002123
MIDDLE EAST-EASTERN EUROPE	-14.9440232	-23.9039098	-5.9841367	0.000026

- 10 different categories, total of 45 different pairwise comparisons
- Two groups with only one observation and three groups with three observations – not sufficient data for comparison
- Only three pairs have an adjusted p-value smaller than 0.05: Latin America vs Eastern Europe, Middle East vs Eastern Europe and Middle East vs Global West

……..

Georgia Tech

7

ANOVA Example 2: Keyboard Layout

Three different keyboard layouts are being compared in terms of typing speed.

Which mean typing times for the three keyboard layouts are different?

Layout 1	Layout 2	Layout 3
23.8	30.2	27.0
25.6	29.9	25.4
24.0	29.1	25.6
25.1	28.8	24.2
25.5	29.1	24.8
26.1	28.6	24.0
23.8	28.3	25.5
25.7	28.7	23.9
24.3	27.9	22.6
26.0	30.5	26.0
24.6	*	23.4
27.0	*	*

Georgia Tech

8

Pairwise Comparison

TukeyHSD(aov(speed ~ layout))
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = speed ~ layout)

$layout

	diff	lwr	upr	p adj
2-1	3.9850000	2.854395	5.1156053	0.0000000
3-1	-0.3613636	-1.463581	0.7408538	0.7008915
3-2	-4.3463636	-5.500092	-3.1926352	0.0000000

- Keyboard layout 2 has a statistically significantly higher typing time than keyboard layouts 1 and 3, on average.
- It is plausible that keyboard layouts 1 and 3 have similar typing time, on average.

Georgia Tech

9

Summary

Georgia Tech

10

Regression Analysis

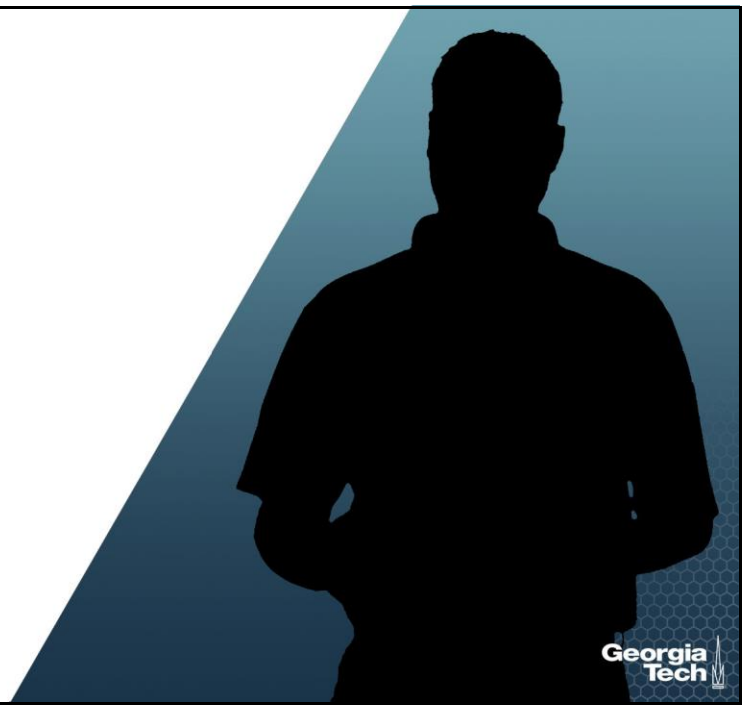
Analysis of Variance

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

Model Fit Assessment

1

About This Lesson



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

Residual Analysis

$Y_{ij} = \mu_i + \varepsilon_{ij}$

- In the model, ε_{ij} is the *error term*. We want ε_{ij} **~ N(0, σ^2).** To check to see if this is true, we examine the residual errors:

$$\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\mu}_i$$

- If the model fit is a good fit, then the residuals should be scattered around zero (randomly).

Georgia Tech

4

Residual Analysis

Residual plots:

- plot $\hat{\varepsilon}_{ij}$ for each treatment group
- plot the quantile-quantile normal plot of $\hat{\varepsilon}_{ij}$
- plot the histogram of $\hat{\varepsilon}_{ij}$

If the scatter of $\hat{\varepsilon}_{ij}$ is <u>not random</u>, it could be that:

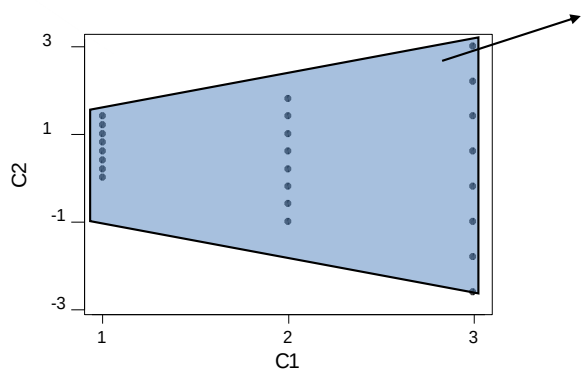
- sample responses are not independent
- variances of responses are not equal

If the quantile-quantile normal plot and the histogram show departure from normality, you may consider a transformation.

Georgia Tech

5

Residual Plot Example 1

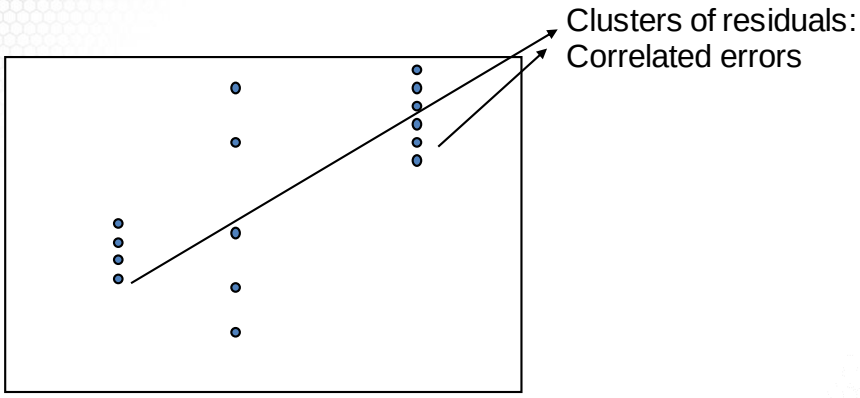


Non-constant variance

Georgia Tech

6

Residual Plot Example 2



Georgia Tech

7

Example1: Global Suicide by Region

Is the ANOVA model a good fit?

Georgia Tech

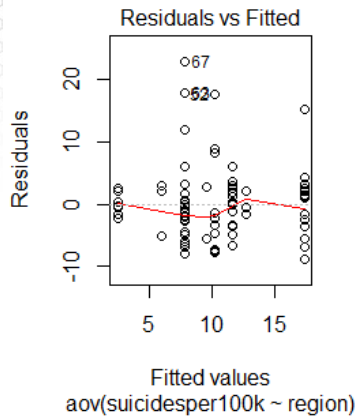
8

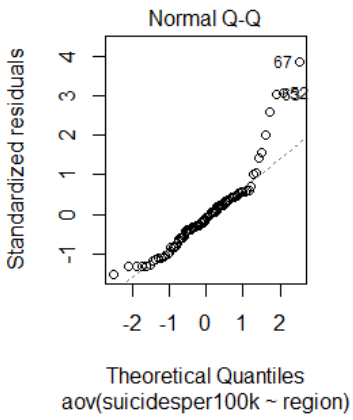
Residual Analysis

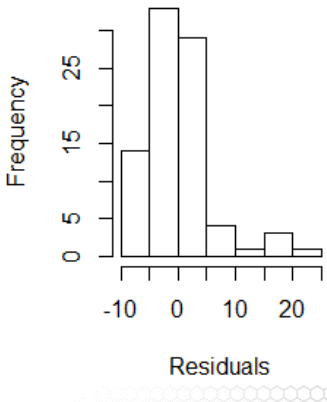
```
## Discard groups with 1 observation
region = suicide_data$region
groups.1 = c(which(region=="NORTHERN AMERICA"),which(region=="WESTERN ASIA"))
suicide_data = suicide_data[-groups.1,]
model.1 = aov(suicidesper100k ~ region, data=suicide_data)

## Diagnostic plots
plot(model.1)
resid.model.1=residuals(model.1)
```

9

Residual Analysis







10

Summary

11

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

Regression Analysis

Analysis of Variance

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

ANOVA R Example

1

About This Lesson

2

Cancer Survival

Reference:
Cameron, E. and Pauling, L. (1978) Supplemental ascorbate in the supportive treatment of cancer: re-evaluation of prolongation of survival times in terminal human cancer. Proceedings of the National Academy of Science USA, 75, 4538-4542.

Georgia Tech

3

ANOVA Example Data

Response Variable:
Y_{ij} = The number of survival days for the j^{th} patient with i^{th} type of cancer

Categories:
Cancer type i for i = 1,2,3,4,5

Stomach	Bronchus	Colon	Ovary	Breast
124	81	248	1234	1235
42	461	377	89	24
25	20	189	201	1581
45	450	1843	356	1166
412	246	180	2970	40
51	166	537	456	727
1112	63	519		3808
46	64	455		791
103	155	406		1804
876	859	365		3460
146	151	942		719
340	166	776		
396	37	372		
	223	163		
	138	101		
	72	20		
	245	283		

Georgia Tech

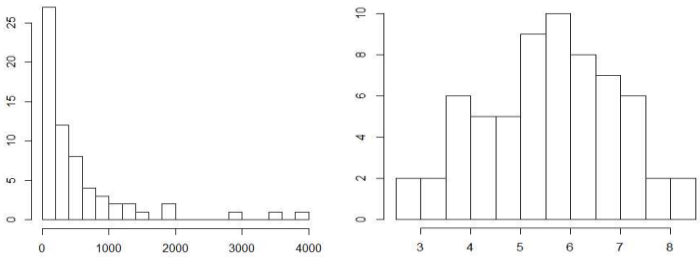
4

Exploratory Data Analysis in R

Read data with 'read.table' R command for reading ASCII files
cancer_data = read.table("CancerStudy.txt", header=T)
Response Variable
survival = cancer_data$Survival

Explore the shape of the distribution of the response variable
hist(survival, xlab="", ylab="Number of Survival Days", main="", nclass=15)

Transform due to skewness of the distribution
hist(log(survival), xlab="", ylab="Number of Survival Days", main="", nclass=15)

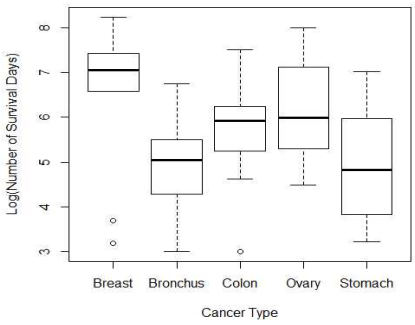


Georgia Tech

5

ANOVA in R

Need to specify Response & Categorical Variables
survival = log(survival)
cancertype = cancer_data$Organ
Convert into categorical variable in R
cancertype = as.factor(cancertype)
boxplot(survival~cancertype, xlab = "Cancer Type", ylab = "Log(Number of Survival Days)")….



- **Within-variability** – some groups have higher variability than others
- **Between-variability** – there is some variability between the means of the five groups
- *Is the between-variability significantly larger than the within-variability?*

Georgia Tech

6

ANOVA in R (cont'd)

```
## ANOVA in R: Is the between-variability significantly larger than within-variability
model = aov(survival ~ cancertype)
summary(model)
             Df    Sum Sq   Mean Sq   F value    Pr(>F)
cancertype    4     24.49     6.122     4.286   0.00412 **
Residuals    59     84.27     1.428
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Obtain estimated means
model.tables(model, type="means")
Tables of means
Grand mean

5.555785

cancertype
        Breast   Bronchus   Colon   Ovary   Stomach
        6.559    4.953      5.749   6.151   4.968
rep     11.000   17.000     17.000  6.000   13.000
```

7

Pairwise Comparison in R

```
## Which means are statistically significantly different? Pairwise Comparison
TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = survival ~ cancertype)

$cancertype
                     diff        lwr         upr       p adj
Bronchus-Breast -1.60543320 -2.906741 -0.3041254  0.0083352
Colon-Breast    -0.80948110 -2.110789  0.4918267  0.4119156
Ovary-Breast    -0.40798703 -2.114754  1.2987803  0.9615409
Stomach-Breast  -1.59068365 -2.968399 -0.2129685  0.0158132
Colon-Bronchus   0.79595210 -0.357534  1.9494382  0.3072938
Ovary-Bronchus   1.19744617 -0.399483  2.7943753  0.2296079
Stomach-Bronchus 0.01474955 -1.224293  1.2537924  0.9999997
Ovary-Colon      0.40149407 -1.195435  1.9984232  0.9540004
Stomach-Colon   -0.78120255 -2.020245  0.4578403  0.3981146
Stomach-Ovary   -1.18269662 -2.842480  0.4770864  0.2763506
```

Statistically significant:

$\log(\hat{\mu}_{Bronchus}) - \log(\hat{\mu}_{Breast})$

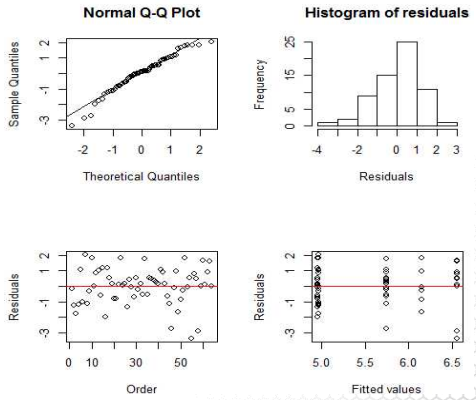
$\log(\hat{\mu}_{Stomach}) - \log(\hat{\mu}_{Breast})$

8

Residual Analysis in R

```
par(mfrow=c(2,2))
qqnorm(residuals(model))
qqline(residuals(model))
hist(residuals(model), main="Histogram of residuals",
xlab="Residuals")
plot(residuals(model), xlab="Order", ylab="Residuals")
abline(0, 0, lty=1, col="red")
plot(fitted(model), residuals(model), xlab="Fitted values",
ylab="Residuals")
abline(0, 0, lty=1, col="red")
```

- The quantiles align on the line and the histogram is approx. symmetric thus normality assumption holds
- Residuals are scattered around zero line with no pattern thus both the constant variance and uncorrelated errors hold



Georgia Tech

9

Cancer Survival: Findings

- There is strong evidence for the difference in the survival time across the five different types of cancer;
- Survival time: Breast cancer vs. Bronchus or Stomach cancer.

Georgia Tech

10

Summary

11