

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.

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		

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
```

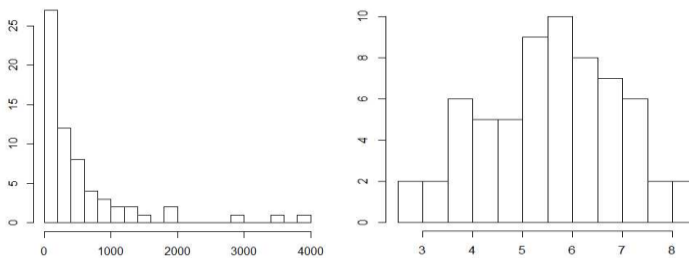
```
survival = cancer_data$Survival
```

```
## Explore the shape of the distribution of the response variable
```

```
hist(survival, xlab="", ylab="Number of Survival Days", main="", nclass=15)
```

```
## T Transform due to skewness of the distribution
```

```
hist(log(survival), xlab="", ylab="Number of Survival Days", main="", nclass=15)
```



5

ANOVA in R

```
## Need to specify Response & Categorical Variables
```

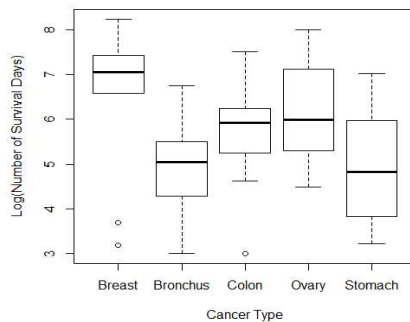
```
survival = log(survival)
```

```
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?*

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

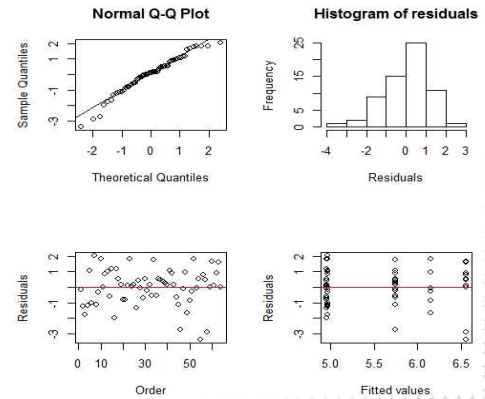


8

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")
```

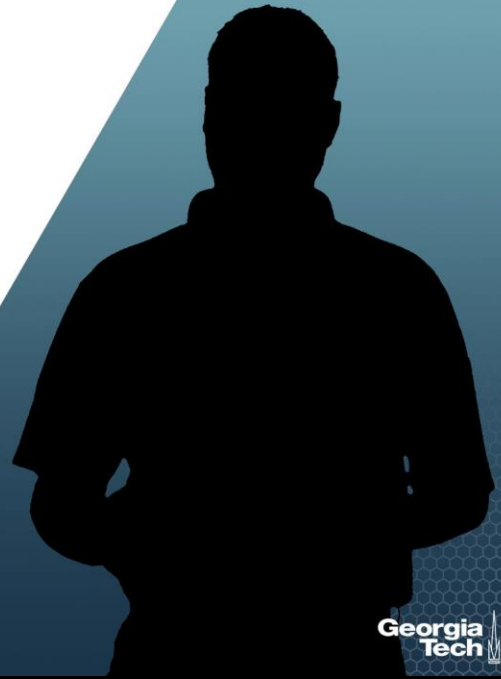
- The quantiles align on the line and the histogram is approx. symmetric thus normality assumption holds
- Residuals are scattered around zero line with no pattern thus both the constant variance and uncorrelated errors hold



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.

Summary



Georgia
Tech