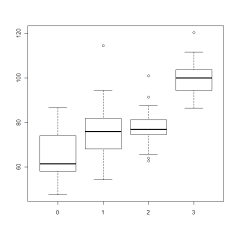
Statistical Computations in R – ANOVA and Regression

Example 1: Test if the average sales of the restaurant after 4 different promotions are equal (use data file restaurant4anova.csv)

No_promo	47.6032	48.9218	86.6995	60.9114	74.0193	58.4517	61.9995	81.6082	58.1103	73.6343
Promo1	83.243	114.432	74.916	70.12	80.039	66.66	70.878	94.427	68.167	82.858
	66.315	70.534	62.506	80.31	75.958	74.315	65.305	91.489	54.258	90.411
	75.991	81.846	81.554	80.38	64.139					
Promo2	79.674	79.314	80.435	80.187	91.444	70.974	75.036	72.623	72.995	75.494
	84.144	75.132	76.865	83.022	82.752	75.295	81.229	65.832	74.65	75.205
	78.09	62.722	64.07	100.902	87.548					
Promo3	98.504	96.154	95.688	111.455	105.033	100.643	98.913	101.709	111.569	86.526
	87.927	103.611	108.959	108.406	100.002	90.109	120.436	90.245	101.11	88.961
	94.453	100.07	91.999	102.255	102.313					


```
y3 <- read.csv("K:/DataMining/Data/restaurant4anova.csv", header=TRUE)
attach(y3)
boxplot(Sales~Promo)
```



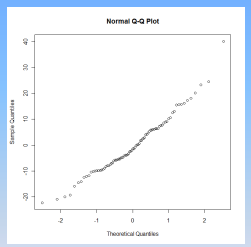

```
> out1 <- aov(Sales~Promo)
> out1
Call:
  aov(formula = Sales ~ Promo)

> summary(out1)
             Df  Sum Sq Mean Sq F value    Pr(>F)
Promo         1  9984.5  9984.5  78.939 1.111e-13 ***
Residuals    83 10498.1   126.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

names(out1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
```



```
qqnorm(out1$residuals)
```



Residuals from the linear model appear to be normally distributed (since data plots along a line)

Example 2: Test if the average job satisfaction score in a company is same for males and females(use data file twoway.csv)

```
> y2 <- read.csv("K:/DataMining/Data/twoway.csv", header=TRUE)
```

Observation	Gender	Position	Score
1	m	S	39
2	m	S	51
3	m	S	54
4	m	S	51
5	m	M	33
6	m	M	36
7	m	M	84
8	m	M	57
9	f	S	60
10	f	S	51
11	f	S	81
12	f	S	57
13	f	M	60
14	f	M	51
15	f	M	69
16	f	M	81

```
> out2 <- aov(Score~Gender*Position)

summary(out2)
                Df  Sum Sq  Mean  Sq F value  Pr(>F)
Gender           1 689.06   689.06 2.9518     0.1115
Position         1  45.56   45.56  0.1952     0.6665
Gender:Position  1   0.56    0.56   0.0024     0.9617
Residuals       12 2801.25  233.44
```

Since the interaction term (Gender:Position) is not significant, it should be dropped and a two-way additive ANOVA model should be used.


```
> out2a <-  aov(Score~Gender+Position)
> summary(out2a)
             Df  Sum Sq  Mean Sq  F value   Pr(>F)
Gender        1 689.06   689.06   3.1971   0.09709 .
Position      1  45.56    45.56    0.2114   0.65326
Residuals    13 2801.81  215.52
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Example 3: The data file football.csv has data on Position, Weight, Time (for 40 yard dash) and player's Rating. Fit a multiple linear regression model to Y = rating as a function of the other 3 variables.

```
x <- read.csv("M:/DataMining/Data/football.csv")
```

	Position	Weight	Time	Rating	Dguard
1	Guard	322	5.38	7.4	1
2	Guard	303	5.18	7.0	1
3	Guard	317	5.34	6.8	1
4	Guard	330	5.46	6.7	1
5	Guard	334	5.18	6.3	1
6	Guard	308	5.32	6.1	1
7	Guard	310	5.28	6.0	1
8	Guard	318	5.37	6.0	1
9	Guard	321	5.25	6.0	1
10	Guard	295	5.34	5.8	1
11	Guard	328	5.31	5.3	1
12	Guard	320	5.64	5.0	1
13	Guard	304	5.20	5.0	1
14	Offensive tackle	325	4.95	8.5	0
15	Offensive tackle	361	5.50	8.0	0
16	Offensive tackle	315	5.39	7.8	0
17	Offensive tackle	307	4.98	7.6	0
18	Offensive tackle	326	5.20	7.3	0
19	Offensive tackle	320	5.36	7.1	0
20	Offensive tackle	287	5.05	6.8	0
21	Offensive tackle	332	5.26	6.8	0
22	Offensive tackle	334	5.55	6.4	0
23	Offensive tackle	312	5.15	6.3	0
24	Offensive tackle	299	5.35	6.1	0
25	Offensive tackle	333	5.59	6.0	0

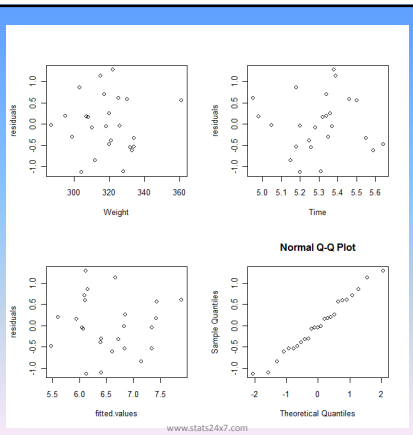

```
> attach(x)
> names(x)
[1] "Position" "Weight"   "Time"     "Rating"   "Dguard"

 outreg2 <- lm(Rating~Weight+Time+Dguard)
>summary(outreg2)
>Call:
lm(formula = Rating ~ Weight + Time + Dguard)
Residuals:
     Min       1Q   Median       3Q      Max
-1.12702 -0.48001 -0.04766  0.55881  1.28344
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.95563    4.49988   2.657   0.0148 *
Weight       0.02219    0.01039   2.135   0.0447 *
Time        -2.27755    0.92895  -2.452   0.0231 *
Dguard      -0.73237    0.28935  -2.531   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.6936 on 21 degrees of freedom
Multiple R-squared: 0.4755,     Adjusted R-squared: 0.4005
F-statistic: 6.345 on 3 and 21 DF,  p-value: 0.003118
```



```
> layout(1:2)
> names(outreg2)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
> attach(outreg2)
```

```
> nf <- layout(matrix(c(1, 2, 3, 4),2,2,byrow=TRUE))
> layout.show(nf)
> plot(residuals~Weight)
> plot(residuals~Time)
> plot(residuals~fitted.values)
> qqnorm(residuals)
```

Variance Inflation Factor in R

Install package HH, then:

- library(HH)
- out <- lm(Rating~Weight+Time+Dguard)
- > vif(out)
- Weight Time Dguard
- 1.296048 1.291599 1.086100


```
> summary(out)

Call:
lm(formula = Rating ~ Weight + Time + Dguard)

Residuals:
    Min      1Q  Median      3Q     Max
-1.12702 -0.48001 -0.04766  0.55881  1.28344

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.95563    4.49988   2.657   0.0148 *
Weight       0.02219    0.01039   2.135   0.0447 *
Time        -2.27755    0.92895  -2.452   0.0231 *
Dguard      -0.73237    0.28935  -2.531   0.0194 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6936 on 21 degrees of freedom
Multiple R-squared: 0.4755,     Adjusted R-squared: 0.4005
F-statistic: 6.345 on 3 and 21 DF,  p-value: 0.003118
```