Linear Models

Dr. Sancharee Basak

ANOVA - One Way Classification

An Example...

Table 3.2 Fluid flow obtained from the rotary pump head of an Olson heart–lung pump

Observation	rpm	Level	Liters/minute
1	150	5	3.540
2	50	1	1.158
3	50	1	1.128
4	75	2	1.686
5	150	5	3.480
6	150	5	3.510
7	100	3	2.328
8	100	3	2.340
9	100	3	2.298
10	125	4	2.982
11	100	3	2.328
12	50	1	1.140
13	125	4	2.868
14	150	5	3.504
15	100	3	2.340
16	75	2	1.740
17	50	1	1.122
18	50	1	1.128
19	150	5	3.612
20	75	2	1.740

Continuation...

- The experiment was run to determine the effect of the number of revolutions per minute (rpm) of the pump head of an Olson heart-lung pump on the fluid flow rate.
- Five equally spaced levels of rpm were selected, namely, 50, 75, 100, 125 and 150 rpm. These were coded as 1,2,3,4,5 respectively.
- The data is organized into several groups due to one single grouping variable – **ONE-WAY FIXED EFFECT ANOVA**.
- The model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad j = 1, 2, \ldots, n_i; i = 1, 2, \ldots, k.$$

- We want to test

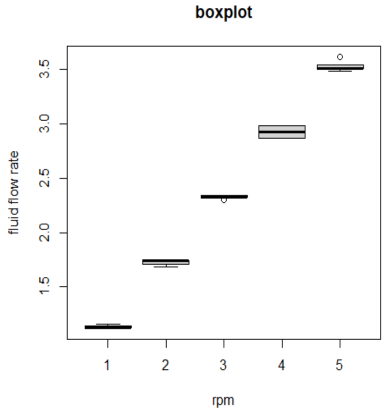
$$H_0 : \tau_1 = \tau_2 = \ldots = \tau_k \quad \text{vs} \quad H_1 : \text{at least two of the } \tau_i' s \text{ differ.}$$

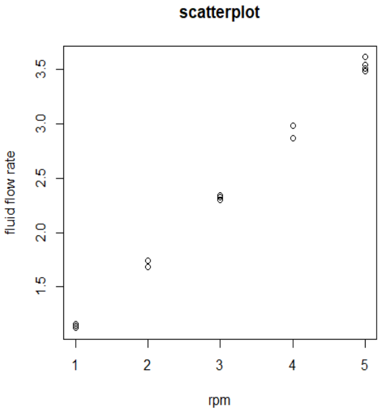
Continuation...

- y denotes the fluid flow rate.
- τ_i denotes the effect of i-th level of rpm.
- x_{ij} takes 1 or 0 value accorsding as i-th level is fixed in case of j-th pump and its fluid flow rate is observed.
- Here we assume ϵ_{ij}'s are mutually independent and $\epsilon_{ij} \sim N(0, \sigma^2)$. Therefore,
- $y_{ij} \sim N(\mu + \tau_i, \sigma^2), j = 1, 2, \ldots, n_i$.

We assume **normality of errors** as well as **homoscedasticity**

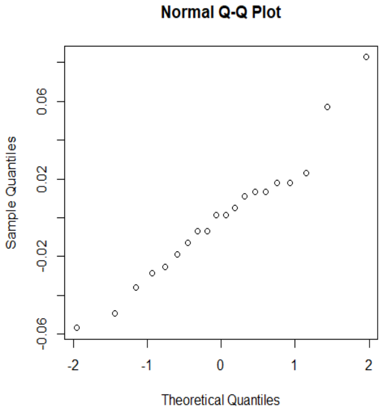
Some graphs...

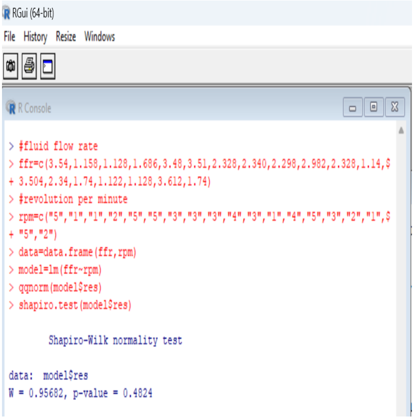




Observations : Means are different w.r.t. different levels whereas variances are almost same in each level.

Normality of errors





Observations : Errors are normally distributed!!

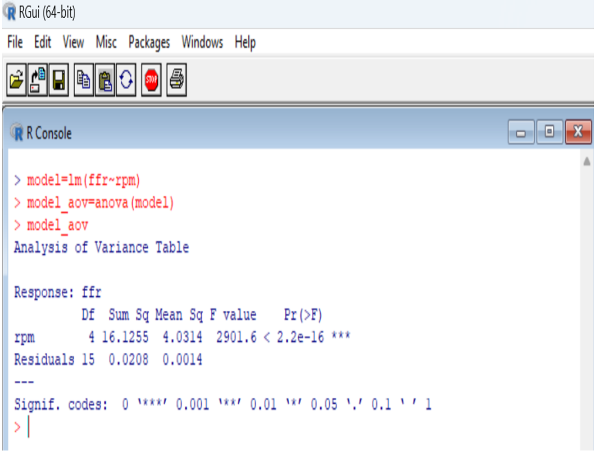
Test for Homoscedasticity

LEVENE's Test :

```
RGui (64-bit)
File Edit View Misc Packages Windows Help

R Console

> #fluid flow rate
> ffr=c(3.54,1.158,1.128,1.686,3.48,3.51,2.328,2.340,2.298,2.982,2.328,1.14,2.868,
+ 3.504,2.34,1.74,1.122,1.128,3.612,1.74)
> #revolution per minute
> rpm=c("5","1","1","2","5","5","3","3","3","4","3","1","4","5","3","2","1","1",
+ "5","2")
> data=data.frame(ffr,rpm)
> library(car)
> leveneTest(ffr~rpm)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4   1.754 0.1907
      15
```

Observation : Homoscedastic!!

ANOVA table




```
> model=lm(ffr~rpm)
> model_aov=anova(model)
> model_aov
Analysis of Variance Table

Response: ffr
          Df  Sum Sq Mean Sq F value    Pr(>F)
rpm        4 16.1255  4.0314  2901.6 < 2.2e-16 ***
Residuals 15  0.0208  0.0014
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

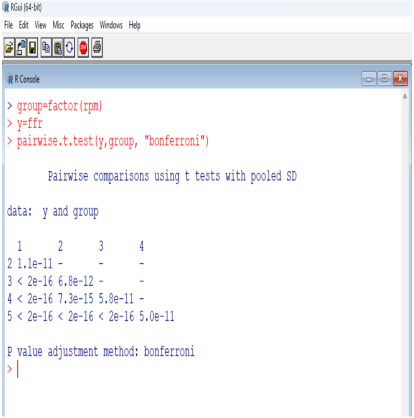
Observation : H_0 is rejected. All treatment means are not equal.

Multiple Comparison

- Whenever, in case of ANOVA model, the decision goes against null, we are generally interested to calculate several confidence intervals for contrasts of treatments.
- A contrast is of the form $\sum_{i=1}^{k} c_i \tau_i$ with $\sum_{i=1}^{n} c_i = 0$.
- Most frequent contrasts are **pairwise comparison of treatments**. It is simply of the form $\tau_j - \tau_l, (j \neq l)$.
- Again sometimes, we want to compare **treatment(s) with control**. It is simply of the form $\tau_j - \tau_1, (j \neq 1)$, where τ_1 represents the effect of control.
- In the above cases, (H_0, H_1) is sub-divided into several alternatives involving two treatments at a time.

Several methods of **multiple comparison** are used to perfom these tests **simultaneously** and the resulting CIs will be termed as **simultaneous CI**.

Multiple Comparison : Bonferroni's Method

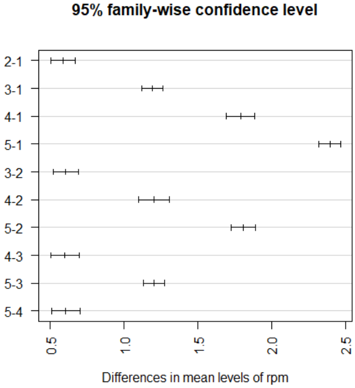


Drawback : Gives wider CI to examine m contrasts if m is large.

```
RGui (64-bit)
File Edit View Misc Packages Windows Help
R Console
> library(DescTools)
> ScheffeTest(aov(ffr~rpm))

  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$rpm
      diff    lwr.ci    upr.ci    pval
2-1 0.5868 0.4916338 0.6819662 4.3e-11 ***
3-1 1.1916 1.1091836 1.2740164 < 2e-16 ***
4-1 1.7898 1.6807734 1.8988266 < 2e-16 ***
5-1 2.3940 2.3115836 2.4764164 < 2e-16 ***
3-2 0.6048 0.5096338 0.6999662 2.8e-11 ***
4-2 1.2030 1.0840422 1.3219578 3.0e-14 ***
5-2 1.8072 1.7120338 1.9023662 < 2e-16 ***
4-3 0.5982 0.4891734 0.7072266 2.3e-10 ***
5-3 1.2024 1.1199836 1.2848164 < 2e-16 ***
5-4 0.6042 0.4951734 0.7132266 2.0e-10 ***

---
```

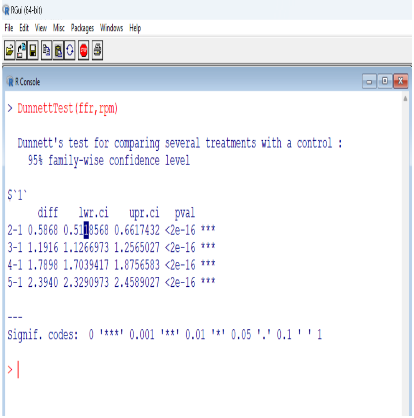
Applicable for any value of m. It has a one-to-one correspondence with ANOVA.

```
R Console
> #fluid flow rate
> ffr=c(3.54,1.158,1.128,1.686,3.48,3.51,2.328,2.340,2.298,2.982,2.328,1.14,$
+ 3.504,2.34,1.74,1.122,1.128,3.612,1.74)
> #revolution per minute
> rpm=c("5","1","1","2","5","5","3","3","3","4","3","1","4","5","3","2","1",$
+ "5","2")
> data=data.frame(ffr,rpm)
> model=lm(ffr~rpm)
> TukeyHSD(aov(ffr~rpm))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = ffr ~ rpm)

$rpm
      diff       lwr       upr p adj
2-1 0.5868 0.5027431 0.6708569     0
3-1 1.1916 1.1188046 1.2643954     0
4-1 1.7898 1.6935007 1.8860993     0
5-1 2.3940 2.3212046 2.4667954     0
3-2 0.6048 0.5207431 0.6888569     0
4-2 1.2030 1.0979289 1.3080711     0
5-2 1.8072 1.7231431 1.8912569     0
4-3 0.5982 0.5019007 0.6944993     0
5-3 1.2024 1.1296046 1.2751954     0
5-4 0.6042 0.5079007 0.7004993     0

> plot(TukeyHSD(aov(ffr~rpm), conf.level=.95), las = 2)
```



- Best for **pairwise comparison**.
- Gives **shorter CI** than both Bonferroni's and Schefffe's methods.

Multiple Comparison : Dunnett's Method



Applicable when simultaneous CIs are needed to be calculated for **treatment vs control** contrasts.

ANOVA - Two Way Classification

Battery Experiment

Observations	Duty	Brand	Life per unit cost
1	regular	name	611
2	regular	store	923
3	regular	name	537
4	heavy	store	476
5	regular	name	542
6	regular	name	593
7	regular	store	794
8	heavy	name	445
9	heavy	store	569
10	regular	store	827
11	regular	store	898
12	heavy	name	490
13	heavy	store	480
14	heavy	name	384
15	heavy	store	460
16	heavy	name	413

Continuation...

- The experiment was run to determine the effect of "brand" and "duty" on the lifetime of the batteries.
- Two levels of "Duty" – heavy and regular as well as two levels of "Brand" – name and store are considered here.
- The data is organized into several groups due to two grouping variables – **TWO-WAY FIXED EFFECTS ANOVA**.
- The analysis will be done on the basis of p observations per cell. If $p = 1$ then we can say that TWO=WAY ANOVA ONE OBSERVATION PER CELL. In case of $p = 1$, interaction effect can not be examined.
- The design is **balanced**.
- The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad k = 1(1)p; j = 1(1)n; i = 1(1)m.$$

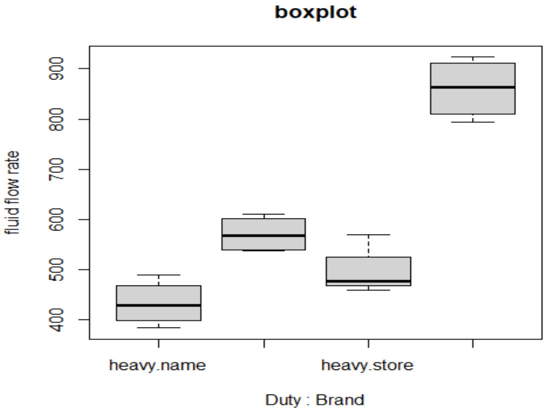
Continuation...

- y denotes the lifetime per unit cost.
- α_i denotes the effect of i-th level of duty; i=1 for regular, 2 for heavy.
- β_j denotes the effect of j-th level of brand; j=1 for name, 2 for store.
- Here we assume ϵ_{ijk}'s are mutually independent and $\epsilon_{ijk} \sim N(0, \sigma^2)$. Therefore,
- $y_{ijk} \sim N(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \sigma^2), k = 1(1)p; j = 1(1)n; i = 1(1)m.$
- We want to test

$$\begin{aligned}
H_{0A} : \alpha_1 = \alpha_2 = \ldots = \alpha_k \quad &\text{vs} \quad H_1 : \text{at least two of the } \alpha_i' s \text{ differ.} \\
H_{0B} : \beta_1 = \beta_2 = \ldots = \beta_k \quad &\text{vs} \quad H_1 : \text{at least two of the } \beta_i' s \text{ differ.} \\
H_{0AB} : (\alpha\beta)_{ij} = (\alpha\beta)_{i'j'}, i \neq i'; j \neq j' \quad &\text{vs} \quad H_1 : \text{at least two of the } (\alpha\beta)_{ij}' s \text{ differ.}
\end{aligned}$$

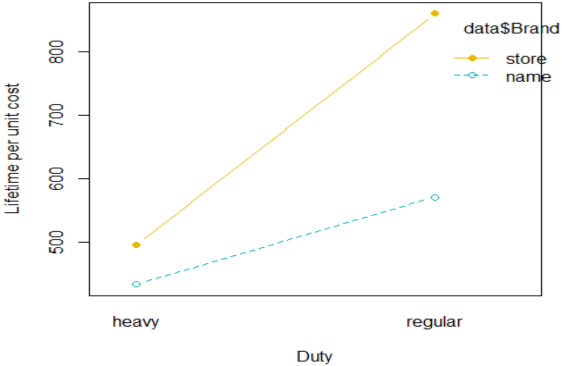
We assume **normality of errors** as well as **homoscedasticity**

Boxplot with two factors

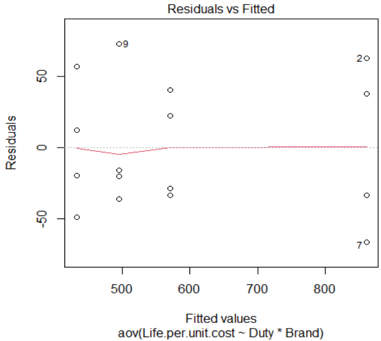


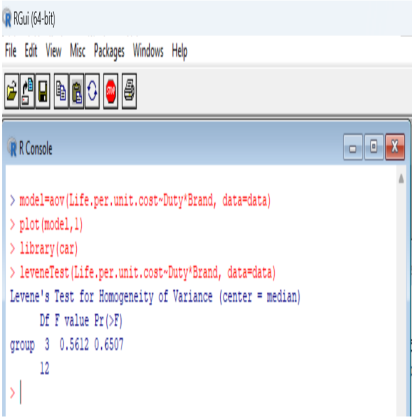
Observation : Treatment means for four treatment combinations are NOT same!!

Interaction plot



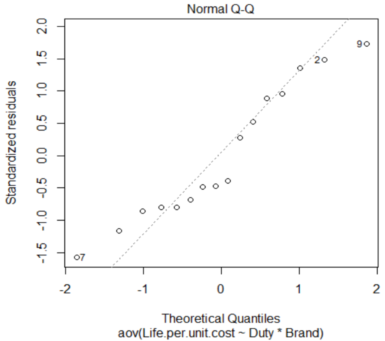
Observation : Interaction effect exists!! Plot with two parallel lines indicates insignificant interaction effect.

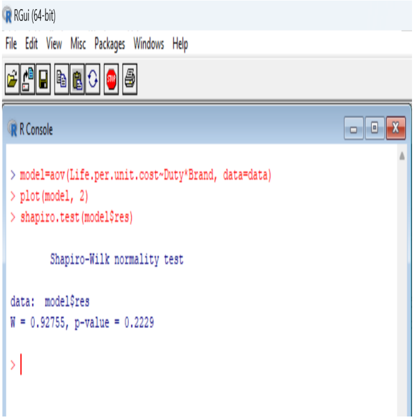




Observation : Homoscedastic!!

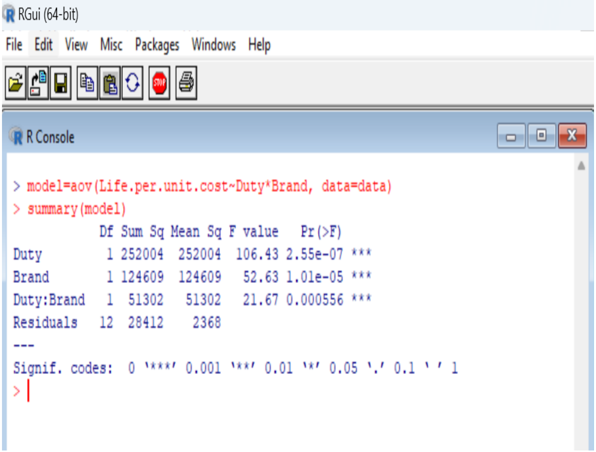
Checking of Normality





Observation : Errors follow normal distribution!!

ANOVA table for FULL model



Observation : H_{0A}, H_{0B}, H_{0AB} all are rejected.

Multiple comparison – Tukey's method

RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
> model=aov(Life.per.unit.cost~Duty*Brand, data=data)
> TukeyHSD(model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Life.per.unit.cost ~ Duty * Brand, data = data)

$Duty
              diff      lwr      upr p adj
regular-heavy  251 197.9905 304.0095 3e-07

$Brand
            diff      lwr      upr    p adj
store-name 176.5 123.4905 229.5095 1.01e-05

$`Duty:Brand`
                             diff        lwr       upr     p adj
regular:name-heavy:name    137.75   35.59848 239.90152 0.0082112
heavy:store-heavy:name      63.25  -38.90152 165.40152 0.3035113
regular:store-heavy:name   427.50  325.34848 529.65152 0.0000002
heavy:store-regular:name   -74.50 -176.65152  27.65152 0.1882188
regular:store-regular:name 289.75  187.59848 391.90152 0.0000115
regular:store-heavy:store  364.25  262.09848 466.40152 0.0000010
```

Continuation...

- The main effects of two levels of “Duty” on the lifetime per unit cost are significantly different, i.e., lifetime is markedly different of batteries due to heavy duty than regular duty.
- The main effects of two levelsof “Brand” on the lifetime per unit cost are significantly different i.e., lifetimes per unit cost, on an average, are markedly different in case of two brand of batteries.
- In case of interaction effect,
  - for heavy duty batteries, effect of brand on lifetime is insignificant.
  - the effect of ‘heavy’ battery of ‘store’ brand insignificantly differs from that of ‘regular’ battery of ‘name‘ brand.

Analysis of Covariance – ANCOVA

Scenario...

- Suppose, in an experiment, along with response variable y, there is another variable, say x which is linearly related to y. The variable can not be controlled by the expeimenter but observed during or prior the experiment. This x is called covariate or concomitant variable.
- ANCOVA basically adjusts the observed response variable for the effect of concomitant variable; otherwise it would inflate the error SS resulting in making true differences in the response due to treatments harder to select.
- **ANCOVA is basically a combination of ANOVA and Regression.**

An Example...

Table 14-8 Breaking Strength Data (y = strength in pounds and x = diameter in 10^{-3} inches)

Machine 1		Machine 2		Machine 3	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15
207	126	216	130	180	106

ANCOVA can be used here to remove the effect of thickness on strength of fiber when testing for differences in strength between machines.

Continuation...

- **Model :**

$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \epsilon_{ij}; i = 1(1)k; j = 1(1)n.$$

- μ : overall effect
  y_{ij} : breaking strength of j-th fiber manufactured by i-th machine
  x_{ij} : thickness of j-th fiber manufactured by i-th machine
  α_i : additional effect due to i-th machine
  β : regression coefficient
  ϵ_{ij} : random error associated with y_{ij}
- **Assumptions :**
  - $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$
  - True relationship between x and y is linear
  - The regression coefficients for each treatment are identical, i.e., the interaction is insignificant
  - The concomitant variable is independent of the treatment.

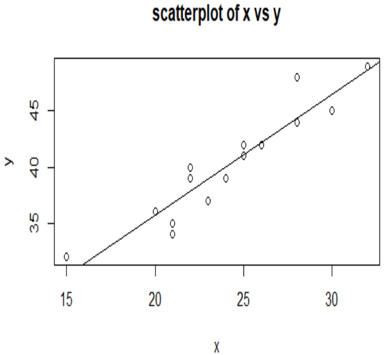
Continuation...

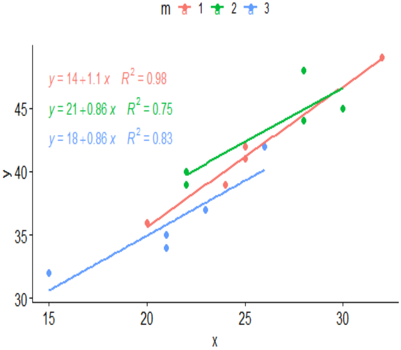
Set of Hypotheses:

- $H_0 : \alpha_i = 0, i = 1(1)k$ vs $H_1 : \text{not} \;\; H_0$.
- $H_0' : \beta = 0$ vs $H_1' : \beta \neq 0$.

If one fails to reject H_0', the model reduces to one-way ANOVA. If one fails to reject H_0, the model reduces to regression model.

Checking Linearity





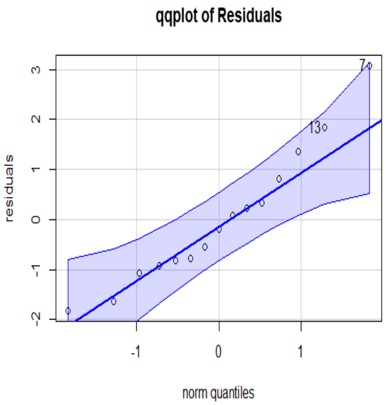
Observation : Breaking strength of fiber manufactured by each machine is linearly dependent on thickness.

Checking Independence between Covariate and Treatment

```
Console   Terminal   Background Jobs
R 4.3.0 · ~/
> #relation between covariate and factor
> mod1=lm(x~m)
> anova(mod1)
Analysis of Variance Table

Response: x
          Df  Sum Sq Mean Sq F value Pr(>F)
m          2  66.133  33.067  2.0286 0.1742
Residuals 12 195.600  16.300
```

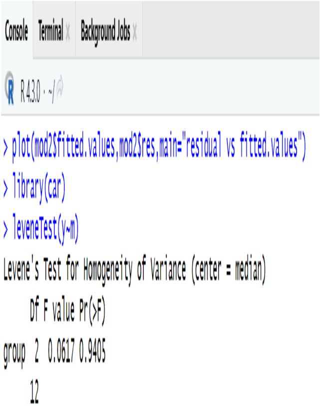
Observation : Cofactor and Treatment are independent of each other. So no reason to assume that machines produce fibers of different thickness.

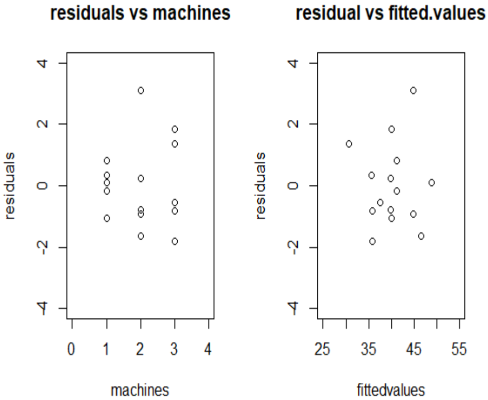
```
Console  Terminal  Background Jobs
R 4.3.0 · ~/
> #ancova model
> mod2=lm(y~x*m)
> library(car)
> qqPlot(mod2$res, main="qqplot of Residuals", ylab="residuals")
[1]  7 13
> shapiro.test(mod2$res)

        Shapiro-Wilk normality test

data:  mod2$res
W = 0.94633, p-value = 0.4686
```



Observation : Errors follow normal distribution!!

Checking Homoscedasticity...





Observation : Homoscedastic!!

Model fitting...

```
Console   Terminal   Background Jobs
R 4.3.0 · ~/
> #ancova model
> mod2=lm(y~x*m)
> anova(mod2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
x          1 305.130 305.130 108.7648 2.52e-06 ***
m          2  13.284   6.642   2.3675   0.1492
x:m        2   2.737   1.369   0.4878   0.6293
Residuals  9  25.249   2.805
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observation : No significant interaction effect. After removing the effect of covariate, treatment effects eventually becomes insignificant.

Interaction part...

We can perform goodness-of-fit test for both interactive and additive models.

Console | Terminal | Background Jobs

R 4.3.0 · ~/

```
> #model with interaction
> mod2=lm(y~x*m)
> #model without interaction
> mod4=lm(y~x+m)
> AIC(mod2,mod4)
     df      AIC
mod2  7 64.37903
mod4  5 61.92291
> anova(mod2,mod4,test="Chisq")
Analysis of Variance Table

Model 1: y ~ x * m
Model 2: y ~ x + m
  Res.Df    RSS Df Sum of Sq Pr(>Chi)
1      9 25.249
2     11 27.986 -2   -2.7372    0.614
> 
```

Observation : Interactive model has slight worse AIC along with NO significant reduction in residual sum of square.

Additive Model

```
Console  Terminal  Background Jobs
R 4.3.0 · ~/
> mod4=lm(y~x+m)
> anova(mod4)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value   Pr(>F)
x          1 305.130 305.130 119.9330 2.96e-07 ***
m          2  13.284   6.642   2.6106   0.1181
Residuals 11  27.986   2.544
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Observation : After adjustment for thickness, the breaking strength remains uniform irrespective of machine type.

Ignoring Covariate – ANOVA model

Console Terminal Background Jobs

R 4.3.0 · ~/

```
> #ANOVA model
> mod5=lm(y~m)
> anova(mod5)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
m          2  140.4  70.200  4.0893 0.04423 *
Residuals 12  206.0  17.167
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observation : Getting exactly opposite result what we have got using ANCOVA model!!!

Multiple Comparison

```
Console   Terminal ×   Background Jobs ×
R 4.3.0 · ~/
> library(emmeans)
> library(rstatix)
> pwc <- data %>%
+   emmeans_test(
+     y ~ m, covariate = x,
+     p.adjust.method = "bonferroni"
+   )
> pwc
# A tibble: 3 × 9
  term  .y.   group1 group2    df statistic      p p.adj p.adj.signif
* <chr> <chr> <chr>  <chr>  <dbl>     <dbl>  <dbl> <dbl> <chr>
1 x*m   y     1      2         11     -1.02 0.328  0.984 ns
2 x*m   y     1      3         11      1.43 0.180  0.541 ns
3 x*m   y     2      3         11      2.28 0.0433 0.130 ns
> get_emmeans(pwc)
# A tibble: 3 × 8
      x m     emmean    se    df conf.low conf.high method
  <dbl> <fct>  <dbl> <dbl> <dbl>    <dbl>     <dbl> <chr>
1  24.1 1       40.4 0.724    11     38.8      42.0 Emmeans test
2  24.1 2       41.4 0.744    11     39.8      43.1 Emmeans test
3  24.1 3       38.8 0.788    11     37.1      40.5 Emmeans test
> 
```

Observation : The difference between each pair is statistically insignificant!!

Summary of ANCOVA model

Console | Terminal | Background Jobs

R 4.3.0 · ~/

```
> #ancova model
> mod2=lm(y~x*m)
> summary(mod2)

Call:
lm(formula = y ~ x * m)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8272 -0.8707 -0.1791  0.5816  3.0857 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.5722     4.9375   2.749 0.022520 *  
x             1.1043     0.1937   5.702 0.000294 ***
m2            7.3421     7.6684   0.957 0.363355    
m3            4.1068     6.6631   0.616 0.552932    
x:m2         -0.2471     0.2960  -0.835 0.425337    
x:m3         -0.2401     0.2843  -0.845 0.420215    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.675 on 9 degrees of freedom
Multiple R-squared:  0.9271,	Adjusted R-squared:  0.8866 
F-statistic:  22.9 on 5 and 9 DF,  p-value: 7.191e-05

> 
```

Interpretation of interactive model

- Here machine I is the reference group.
  - If a fiber is manufactured by machine I, then its estimated breaking strength will be 13.5722+1.1043*x.
  - If a fiber is manufactured by machine II, then its estimated breaking strength will be (13.5722+7.3421)+(1.1043-0.2471)*x.
  - If a fiber is manufactured by machine III, then its estimated breaking strength will be (13.5722+4.1068)+(1.1043-0.2401)*x.
- Almost 89% of total variance has been explained by this model.
- Estimated intercept as well as regression coefficient are significantly non-zero, while others are not.
- Due to insignificant interaction, consideration of additive model is necessary.

Summary of ANOVA model

Console Terminal × Background Jobs ×

R 4.3.0 · ~/

```
> mod4=lm(y~x+m)
> summary(mod4)

Call:
lm(formula = y ~ x + m)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.0160 -0.9586 -0.3841  0.9518  2.8920 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   17.360      2.961   5.862 0.000109 ***
x              0.954      0.114   8.365 4.26e-06 ***
m2             1.037      1.013   1.024 0.328012    
m3            -1.584      1.107  -1.431 0.180292    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.595 on 11 degrees of freedom
Multiple R-squared:  0.9192,	Adjusted R-squared:  0.8972 
F-statistic: 41.72 on 3 and 11 DF,  p-value: 2.665e-06
```

Interpretation of interactive model

- Here also, machine I is the reference group.
  - If a fiber is manufactured by machine I, then its estimated breaking strength will be 17.360+0.954*x.
  - If a fiber is manufactured by machine II, then its estimated breaking strength will be 1.037 unit more than that of fiber manufactured by machine I.
  - If a fiber is manufactured by machine III, then its estimated breaking strength will be 1.584 unit less than that of fiber manufactured by machine I.
- Almost 90% of total variance has been explained by this model.
- Estimated intercept as well as regression coefficient are significantly non-zero, while others are not.

Some Specific Cases

The basic assumptions of ANOVA are **independence, normality & homoscedasticity.**

What to do if these assumptions are violated?

Kruskal-Wallis One-way ANOVA test

Checking homogeneity

- Go for Bartlett's test.
  Advantage : Best when normality assumption is satisfied.
  Drawback : Quite sensitive to normality.
- Go for Levene's test.
  Advantage : More robust than Bartlett's test. Applicable for non-normal case.
  Drawback : Use absolute deviations of y-values from group means.
- Go for Brown-Forsythe test.
  Advantage : Applicable for non-normal case. More robust than Levene's test in case of skewed/heavy-tailed distribution as it uses absolute deviations of y-values from group trimmed means/ group medians.

R Codes:
library(vGWAS); brown.forsythe.test(y, group, data)

ANOVA for heteroscedastic group

- **Welch's F-test :**

  R Code:
  library(rstatix)
  welch (underscore) anova (underscore) test(data, formula)
  OR
  library(misty); test.welch(formula, data, posthoc = TRUE)
  [*Here Welch's ANOVA including Games-Howell post hoc test for multiple comparison is performed.*]
  OR
  library(onewaytests); welch.test(formula, data, rate = 0)
  [*for outlier case mainly*]

- **Brown Forsythe test for equality of means :**

  R Code:
  library(onewaytests); bf.test(formula, data)

ANOVA for paired groups

- If the data are normally distributed go for Repeated Measures ANOVA.
- If the data are normally distributed go for Friedman's test.