Basic Statistical Analysis in R

Get Working Directory

- getwd()

Set Working Directory

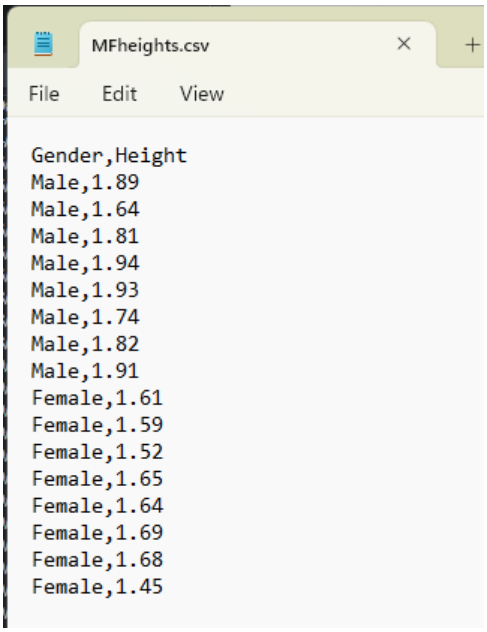
- setwd("C:/Users/andyg/Desktop/R") – for example

1. Descriptive Statistics

Height of Students

	A	B	C
1	Gender	Height	
2	Male	1.89	
3	Male	1.64	
4	Male	1.81	
5	Male	1.94	
6	Male	1.93	
7	Male	1.74	
8	Male	1.82	
9	Male	1.91	
10	Female	1.61	
11	Female	1.59	
12	Female	1.52	
13	Female	1.65	
14	Female	1.64	
15	Female	1.69	
16	Female	1.68	
17	Female	1.45	
18			

Save as “MFheights.csv”



Import into R

- MFheights <- read.csv("MFheights.csv", header = TRUE, colClasses = c("factor", "numeric"))

Look at data

- summary(MFheights)
- summary(MFheights$Height)

```
> summary(MFheights)
  ï..Gender      Height
 Female:8   Min.   :1.590
 Male  :8   1st Qu.:1.688
            Median :1.775
            Mean   :1.776
            3rd Qu.:1.883
            Max.   :1.940
> summary(MFheights$Height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.590   1.688   1.775   1.776   1.883   1.940
```

Can also look individually.

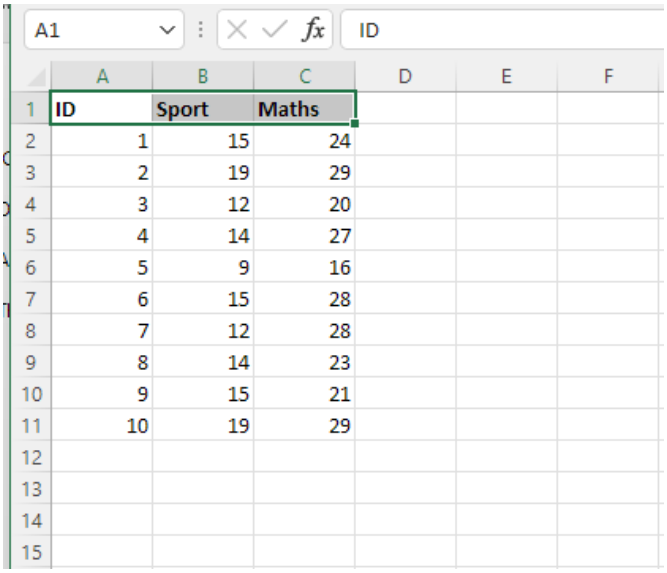
- mean(MFheights$Height)
- median(MFheights$Height)
- sd(MFheights$Height)
- var(MFheights$Height)
- range(MFheights$Height)
- quantile(MFheights$Height)
- IQR(MFheights$Height)
- min(MFheights$Height)
- max(MFheights$Height)

- table(MFheights$Height)

- boxplot(MFheights$Height)
- hist(MFheights$Height)

2. Correlation

Relationship between ability in maths and sports.



	A	B	C	D	E	F
1	ID	Sport	Maths			
2	1	15	24			
3	2	19	29			
4	3	12	20			
5	4	14	27			
6	5	9	16			
7	6	15	28			
8	7	12	28			
9	8	14	23			
10	9	15	21			
11	10	19	29			
12						
13						
14						
15						

Save as “maths-sport.csv”

- Maths.Sport <- read.csv("maths-sport.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric"))

- plot(Maths.Sport$Maths,Maths.Sport$Sport)

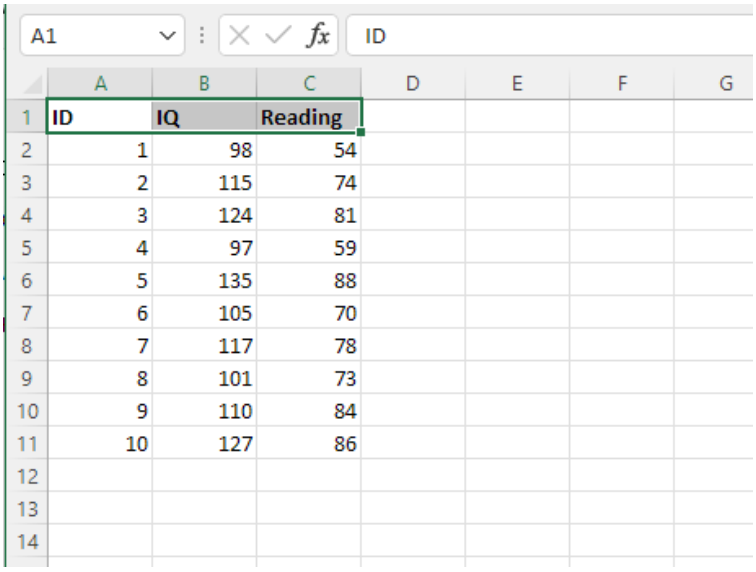
- cor(Maths.Sport$Maths,Maths.Sport$Sport)
- cor.test(Maths.Sport$Maths,Maths.Sport$Sport)

```
> cor(Maths.Sport$Maths,Maths.Sport$Sport)
[1] 0.725103
> cor.test(Maths.Sport$Maths,Maths.Sport$Sport)

        Pearson's product-moment correlation

data:  Maths.Sport$Maths and Maths.Sport$Sport
t = 2.9782, df = 8, p-value = 0.01765
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1756842 0.9300985
sample estimates:
      cor
0.725103
```

Yes, there is a strong significant relationship ($r = 0.725$, $p < 0.05$).

3. Regression

Can we predict reading ability from IQ score? What is the relationship?



	A	B	C	D	E	F	G
1	ID	IQ	Reading				
2	1	98	54				
3	2	115	74				
4	3	124	81				
5	4	97	59				
6	5	135	88				
7	6	105	70				
8	7	117	78				
9	8	101	73				
10	9	110	84				
11	10	127	86				
12							
13							
14							

Save as "IQ-Reading.csv"

- IQ.Reading <- read.csv("IQ-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric"))
- cor.test(IQ.Reading$Reading, IQ.Reading$IQ)
- lm(IQ.Reading$Reading ~ IQ.Reading$IQ)

```
> cor.test(IQ.Reading$Reading,IQ.Reading$IQ)

        Pearson's product-moment correlation

data:  IQ.Reading$Reading and IQ.Reading$IQ
t = 4.626, df = 8, p-value = 0.001697
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4829915 0.9646213
sample estimates:
      cor
0.8531664

> lm(IQ.Reading$Reading ~ IQ.Reading$IQ)

Call:
lm(formula = IQ.Reading$Reading ~ IQ.Reading$IQ)

Coefficients:
  (Intercept)   IQ.Reading$IQ
      -8.8247          0.7398
```

Yes, there is a strong significant relationship (r=0.853, p = 0.0017)

Therefore: Reading = 0.74 x IQ – 8.82

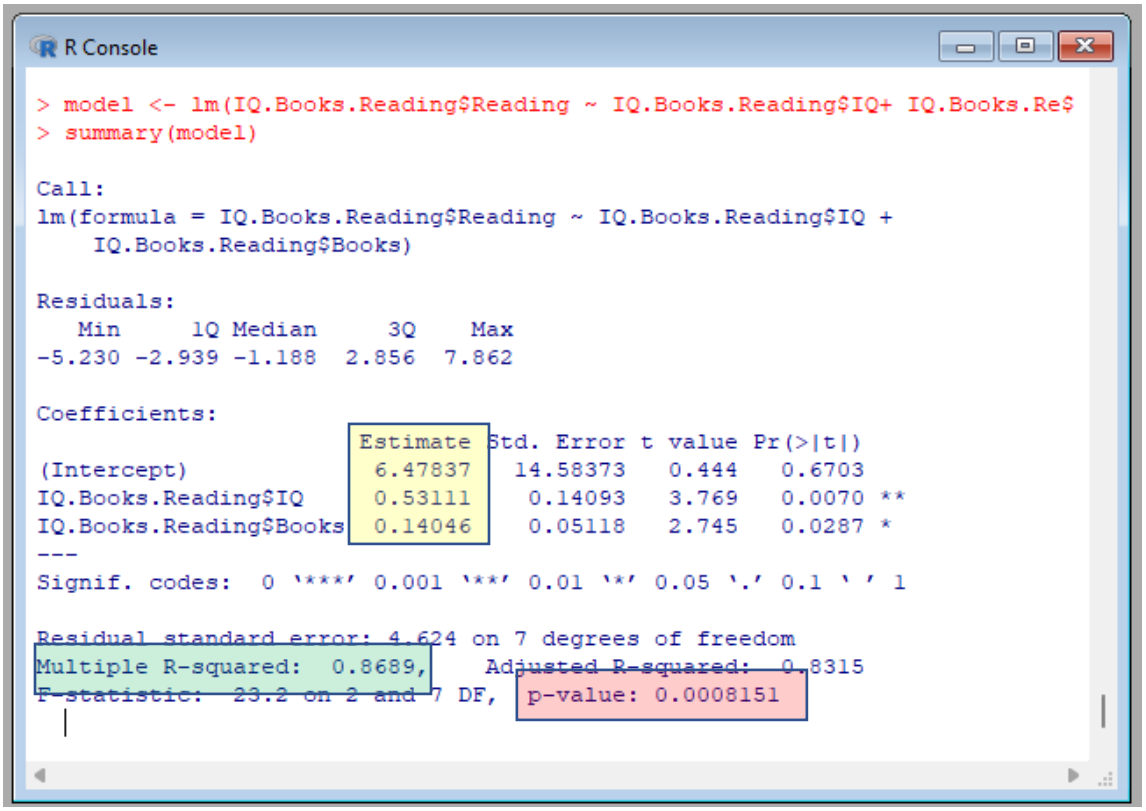
Multiple Regression

Can we predict reading ability from IQ score + number of Books Available? What is the relationship?

	A	B	C	D	E	F
1	ID	IQ	Books	Reading		
2	1	98	5	54		
3	2	115	25	74		
4	3	124	37	81		
5	4	97	10	59		
6	5	135	90	88		
7	6	105	80	70		
8	7	117	88	78		
9	8	101	73	73		
10	9	110	80	84		
11	10	127	100	86		
12						
13						

Save as "IQ-Books-Reading.csv

- IQ.Books.Reading <- read.csv("IQ-Books-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric"))
- model <- lm(IQ.Books.Reading$Reading ~ IQ.Books.Reading$IQ+ IQ.Books.Reading$Books)
- summary(model)



Yes, there is a very strong relationship (R^2=0.869, $p < 0.01$)

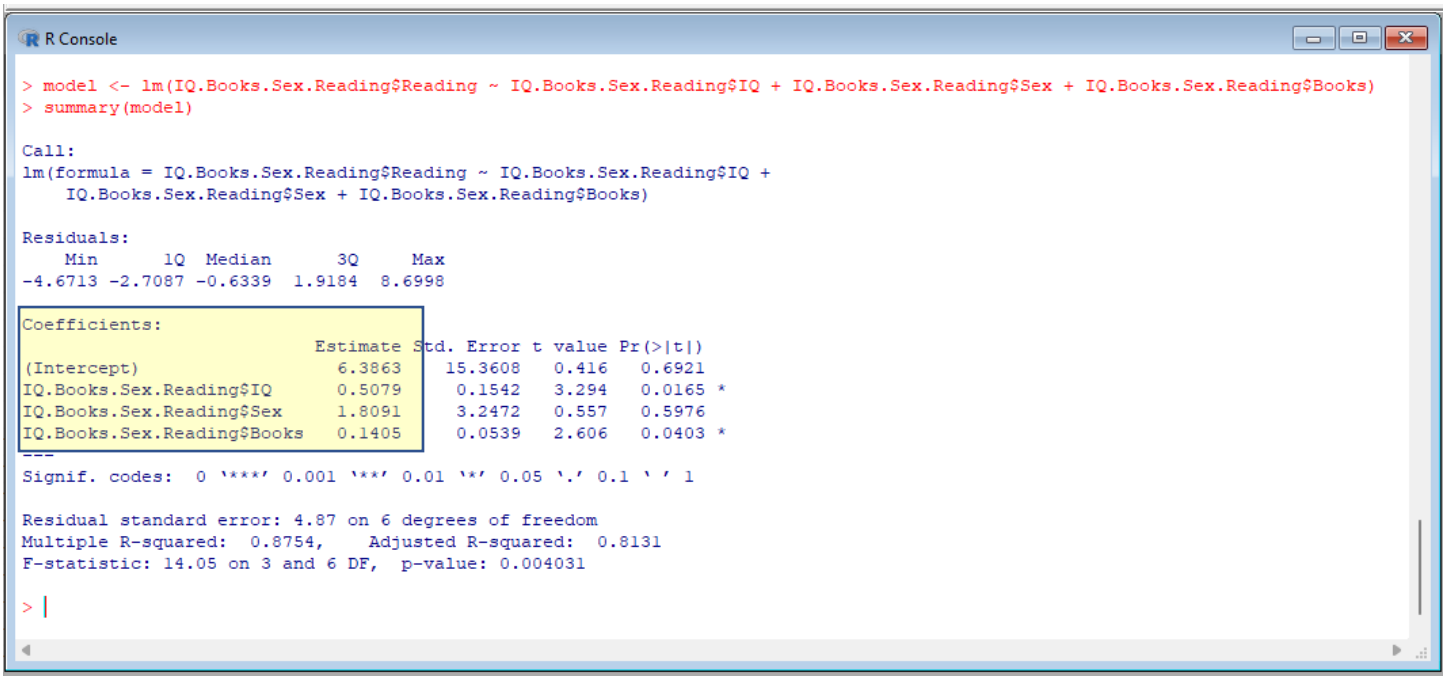
Therefore: Reading =6.478 + .531 x IQ + .140 x Books Available

We can add more variables. For example: Male or Female (1=Male, 2=Female).

	A	B	C	D	E	F
1	ID	IQ	Books	Sex	Reading	
2	1	98	5	1	54	
3	2	115	25	2	74	
4	3	124	37	2	81	
5	4	97	10	1	59	
6	5	135	90	1	88	
7	6	105	80	1	70	
8	7	117	88	2	78	
9	8	101	73	2	73	
10	9	110	80	1	84	
11	10	127	100	2	86	
12						
13						

Save as "IQ-Books-Sex-Reading.csv

- IQ.Books.Sex.Reading <- read.csv("IQ-Books-Sex-Reading.csv", header = TRUE, colClasses = c("numeric", "numeric", "numeric", "numeric", "numeric"))
- model <- lm(IQ.Books.Sex.Reading$Reading ~ IQ.Books.Sex.Reading$IQ+ IQ.Books.Sex.Reading$Books + IQ.Books.Sex.Reading$Sex)
- summary(model)



Again there is a strong significant correlation between the proposed model of reading ability score

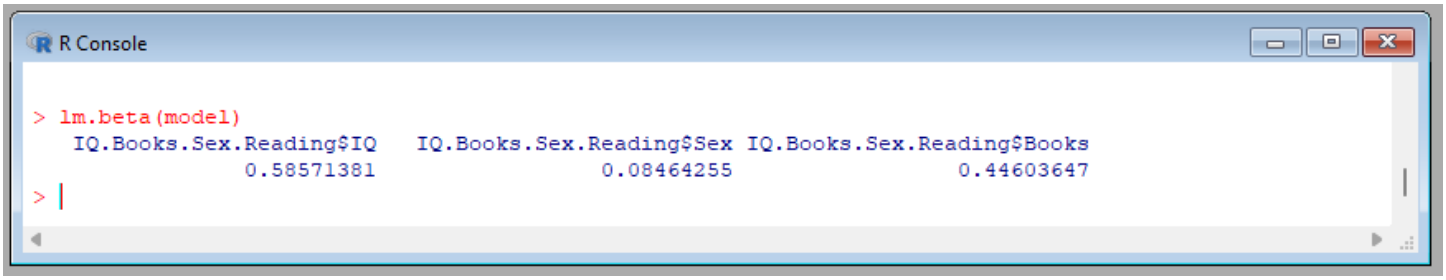
Reading = 6.39 + .5 x IQ + .14 x Books + 1.8 x Sex

Note that Sex is a categorical variable with two levels (Males = 1 & Females = 2).

In this case the regression weighting (1.8) is added to the predicted Reading value. So for females 1.8 (1.8 x 1) would be added and for males 3.6 (1.8 x 2) would be added.

Which variables are important?

Install and load the "QuantPsych" package, and use the function "lm.beta" on your model.



You now have the standardised regression coefficients, which you can now compare.

IQ: 0.58571381

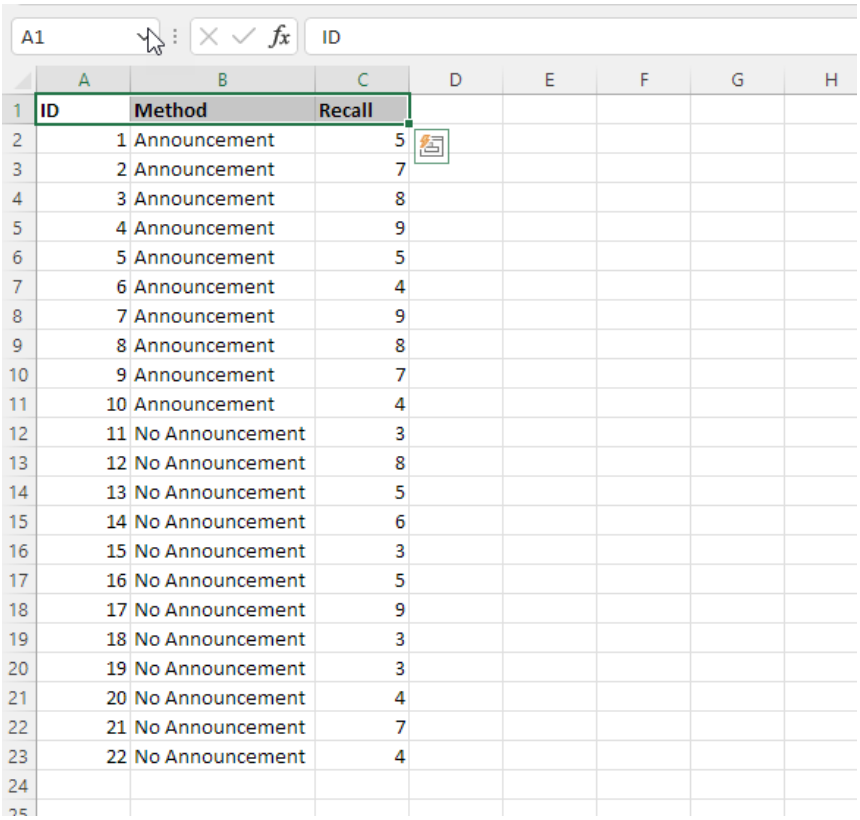
Sex: 0.08464255

Books Available: 0.44603647

And you can see that the most important predictor is IQ, followed closely by Books Available. Sex is not contributing very much.

4. t-Test

Is there a difference between oral announcement of objects and memory?



ID	Method	Recall
1	Announcement	5
2	Announcement	7
3	Announcement	8
4	Announcement	9
5	Announcement	5
6	Announcement	4
7	Announcement	9
8	Announcement	8
9	Announcement	7
10	Announcement	4
11	No Announcement	3
12	No Announcement	8
13	No Announcement	5
14	No Announcement	6
15	No Announcement	3
16	No Announcement	5
17	No Announcement	9
18	No Announcement	3
19	No Announcement	3
20	No Announcement	4
21	No Announcement	7
22	No Announcement	4

Save as “Announcement.csv”

- Announcement <- read.csv("Announcement.csv", header = TRUE, colClasses = c("numeric", "factor", "numeric"))
- summary(Announcement)
- boxplot(Announcement$Recall ~ Announcement$Method)

t.test(numeric variable ~ dichotamous variable)
OR t.test(numeric variable1, numeric variable2)

- t.test(Announcement$Recall ~ Announcement$Method)

```
> t.test(Announcement$Recall~Announcement$Method)

        Welch Two Sample t-test

data:  Announcement$Recall by Announcement$Method
t = 1.8526, df = 19.69, p-value = 0.07898
alternative hypothesis: true difference in means between group Announcement and $
95 percent confidence interval:
 -0.2033249  3.4033249
sample estimates:
   mean in group Announcement mean in group No Announcement
                          6.6                           5.0
```

No, there is no significant difference between oral announcement of objects and memory (p = 0.08).

5. ANOVA – One-Way

Is there a difference betweem amount of alcohol consumed and stopping distance?

	A	B	C	D	E	F	G
1	ID	alcohol	stopping				
2	1	No Alcohol	31.2				
3	2	No Alcohol	27.4				
4	3	No Alcohol	29.8				
5	4	No Alcohol	33.5				
6	5	No Alcohol	34				
7	6	One Unit	31.2				
8	7	One Unit	33.7				
9	8	One Unit	34.3				
10	9	One Unit	39.2				
11	10	One Unit	36				
12	11	Two Units	42.1				
13	12	Two Units	40.6				
14	13	Two Units	43.4				
15	14	Two Units	37.9				
16	15	Two Units	39				
17	16	Three Units	46.7				
18	17	Three Units	45.1				
19	18	Three Units	43.2				
20	19	Three Units	41.7				
21	20	Three Units	40.3				
22							

Stopping distance in metres.

Save as"Stopping.csv"

- boxplot(Stopping$stopping ~ Stopping$alcohol)

aov(dependent numerical variable ~ factor)

- anova1 <- aov(Stopping$stopping ~ Stopping$alcohol)
- summary(anova1)

```
> boxplot(Stopping$stopping~Stopping$alcohol)
> anova <- aov(Stopping$stopping ~ Stopping$alcohol)
> summary(anova)
                  Df Sum Sq Mean Sq F value   Pr(>F)
Stopping$alcohol   3  456.1  152.04   21.92 6.52e-06 ***
Residuals         16  111.0    6.94
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Yes ($p < 0.05$). Therefore there is a significant difference in stopping distance according to amount of alcohol drunk.

Where are the differences?

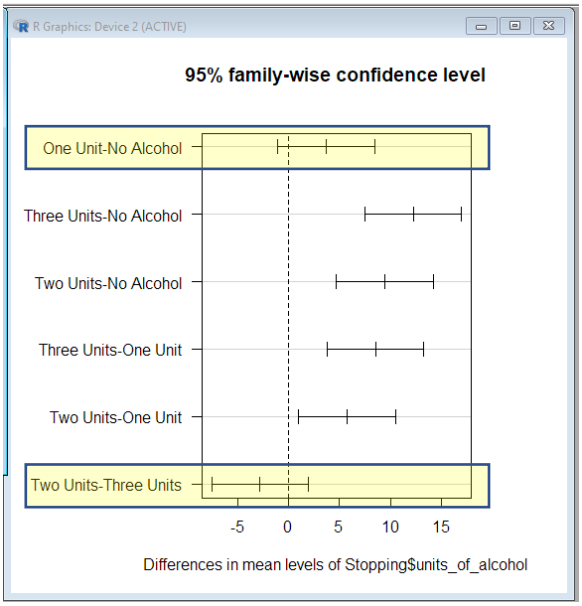
- sum <-TukeyHSD(anova)
- sum

```
> TukeyHSD(anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Stopping$stopping ~ Stopping$alcohol)

$`Stopping$alcohol`
                          diff        lwr       upr     p adj
One Unit-No Alcohol       3.70 -1.0654663  8.465466 0.1596059
Three Units-No Alcohol   12.22  7.4545337 16.985466 0.0000091
Two Units-No Alcohol      9.42  4.6545337 14.185466 0.0001901
Three Units-One Unit      8.52  3.7545337 13.285466 0.0005430
Two Units-One Unit        5.72  0.9545337 10.485466 0.0161514
Two Units-Three Units    -2.80 -7.5654663  1.965466 0.3649771
```

- plot(sum, las=1)

Change margins

- par(mar=c(5,10,5,5))

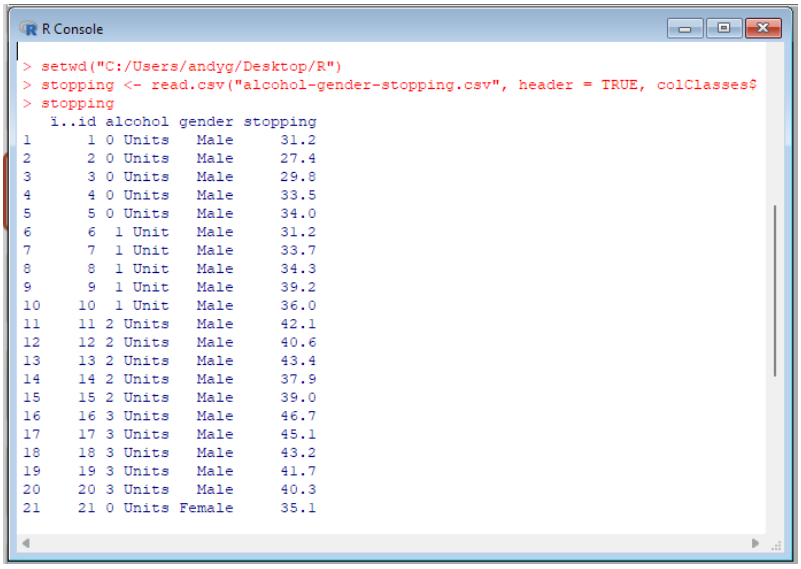


No significant difference when lines cross zero.

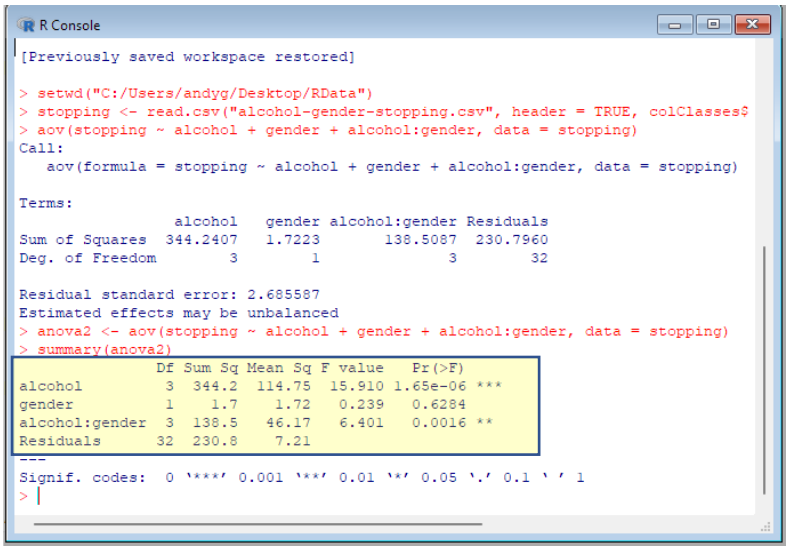
6. ANOVA – Two-Ways

Is the effect of amount of alcohol consumed on stopping distance different for males and females?

- stopping <- read.csv("alcohol-gender-stopping.csv", header = TRUE, colClasses = c("numeric", "factor", "factor", "numeric"))

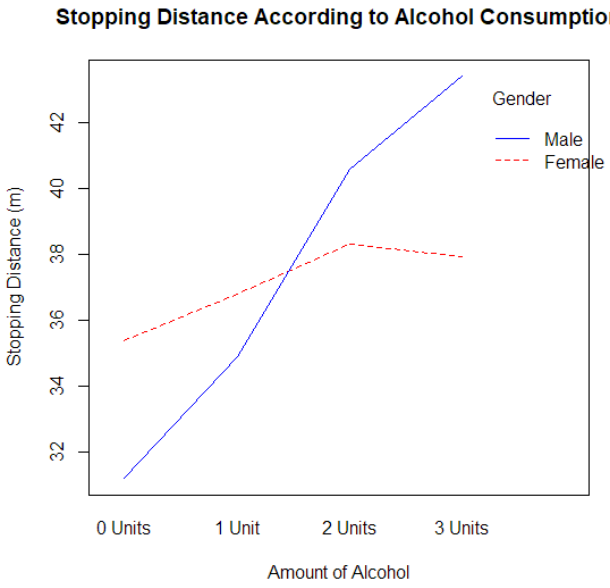


- anova2 = aov(stopping ~ alcohol + gender + alcohol:gender, data = stopping)
- summary(anova2))

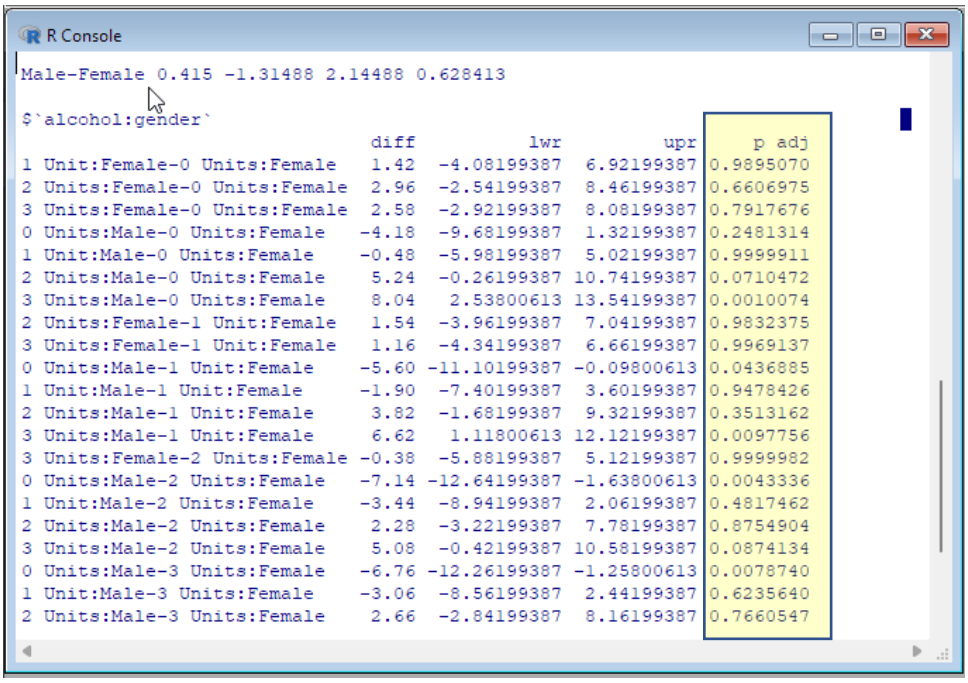


alcohol:gender - Interaction effect = whether the influence of alcohol on stopping distance depends on whether you are male or female – significant (p=0.0016)

- interaction.plot(stopping$alcohol, stopping$gender, stopping$stopping, col=c("red", "blue"), trace.label="Gender", xlab="Amount of Alcohol", ylab="Stopping Distance (m)", main= "Stopping Distance According to Alcohol Consumption")

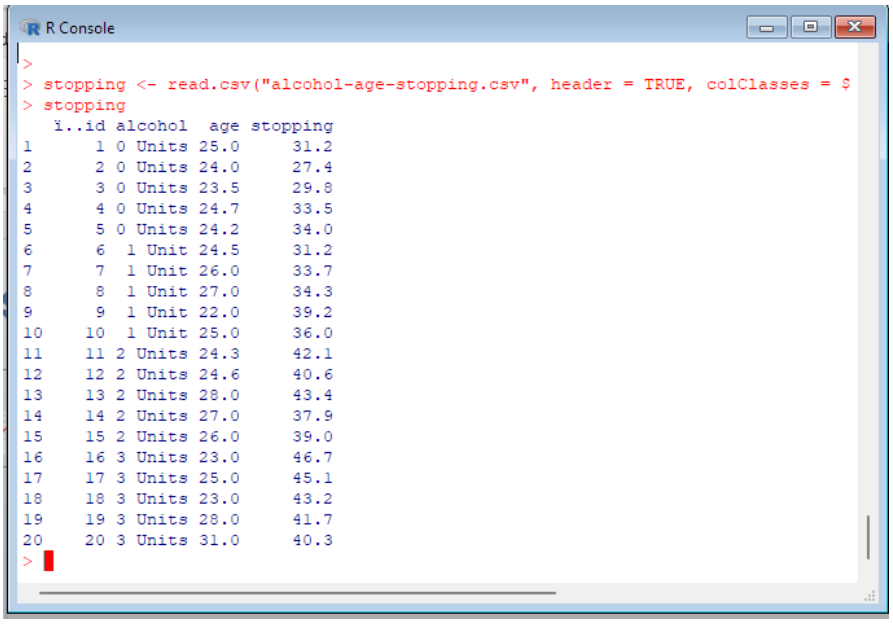


- model_aov <- aov(anova1)
- TukeyHSD(model_aov)

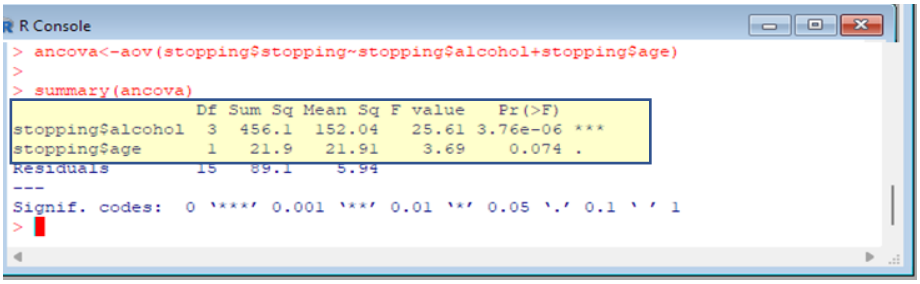


7. ANCOVA

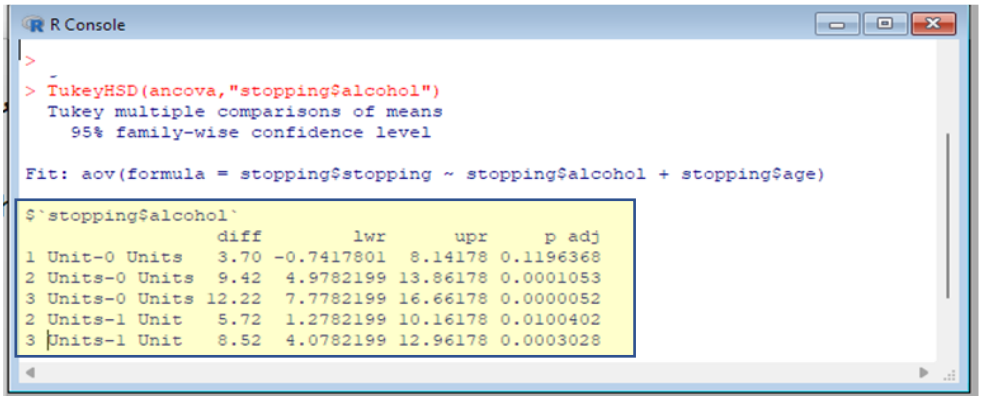
Does age affect the stopping distance according to the amount of alcohol drunk?



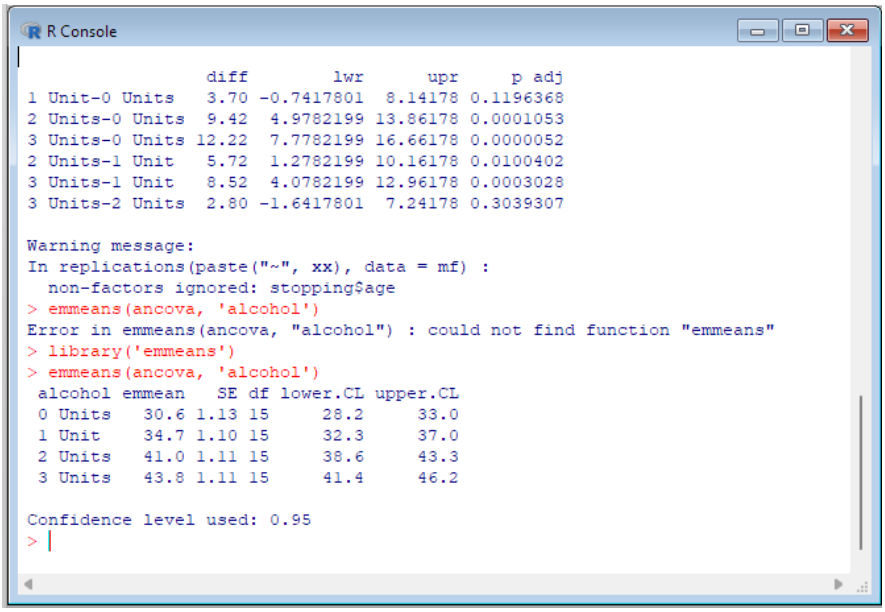
- ancova<-aov(stopping$stopping ~ stopping$alcohol+stopping$age)
- summary(ancova)



- TukeyHSD(ancova,"stopping$alcohol")



- library('emmeans')
- emmeans(ancova, 'alcohol')



After adjusting for age, a significant difference (F(3, 15) = 25.61, p <.001) and was found between the amount of alcohol drunk and the stopping distance. Post hoc comparisons using the Tukey test were carried out. No significant difference was found between stopping distances with No Alcohol and One Unit of alcohol (p = 0.12), with only a 3.42 difference in mean stopping distances, but the other differences were significant, with the mean difference in stopping distance ranging from 5.7 to 12.2 meters

An ANCOVA is a form of analysis that tells you whether your groups differ on a dependent variable when you have partialled out (removed the effects of) another variable called the covariate that has a relationship with the dependent variable. ANCOVA therefore adjusts the means on the covariate so that the mean covariate score is the same for all groups and therefore any changes in the DV are solely due to the treatment variable.

It is mostly used in pre-test-post-test designs where researchers want to partial out (remove or hold constant) the effect of a confounding variable. Using simple difference scores between pre and post-test does not achieve this. ANCOVA then answers the question, What would the means of each group be on the DV if the means of the different groups had the same mean on the pre-test score?", i.e. it adjusts the DV means so that we can investigate the mean differences for post-test score only.

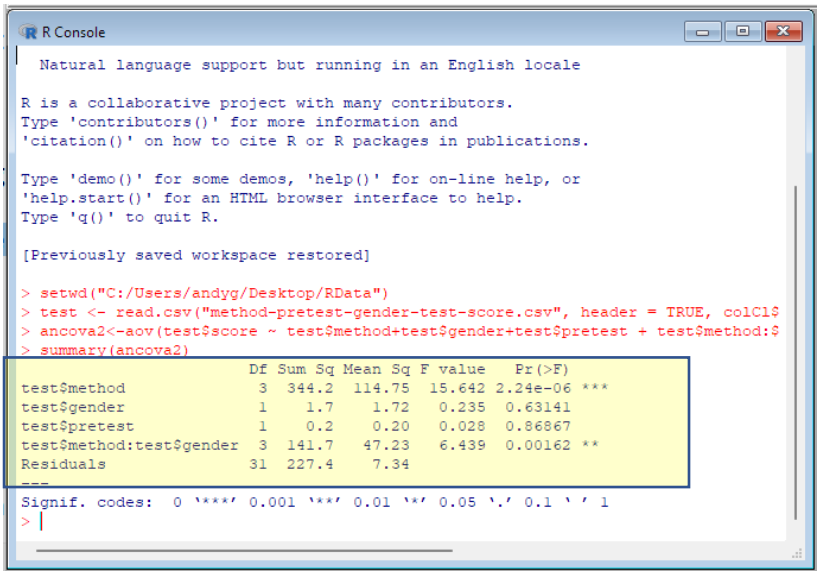
.

8. ANCOVA – Two-Way

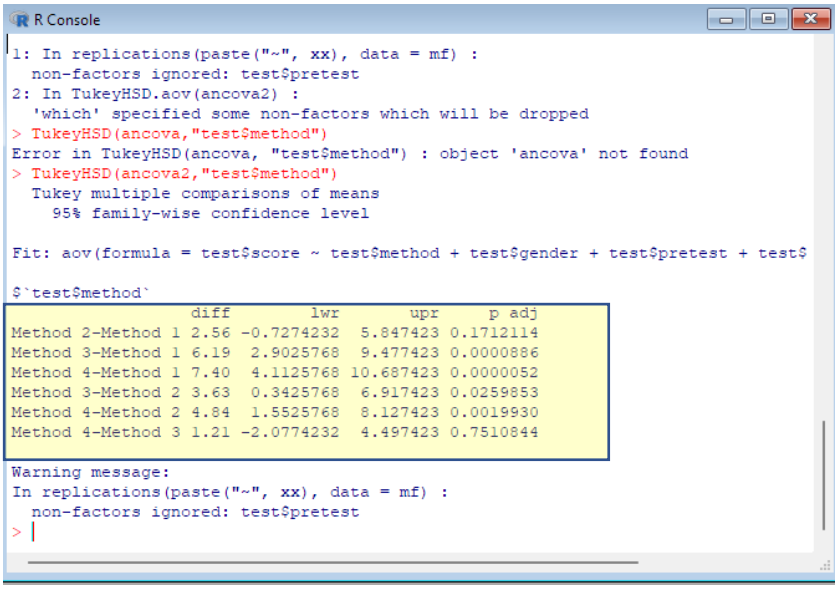
A sample of male and female adults was selected and randomly allocated to four teaching groups using different teaching methods, after which they were given a test.
Group 1: Method 1
Group 2: Method 2
Group 3: Method 3
Group 4: Method 4
Two categorical variables – Method & Gender
Two continuous variables – Pre-Test Score & Final Test Score

- test <- read.csv("method-pretest-gender-test-score.csv", header = TRUE, colClasses = c("factor","factor","numeric", "numeric"))

- ancova2<-aov(test$score ~ test$method+test$gender+test$pretest + test$method:test$gender)
- summary(ancova2)

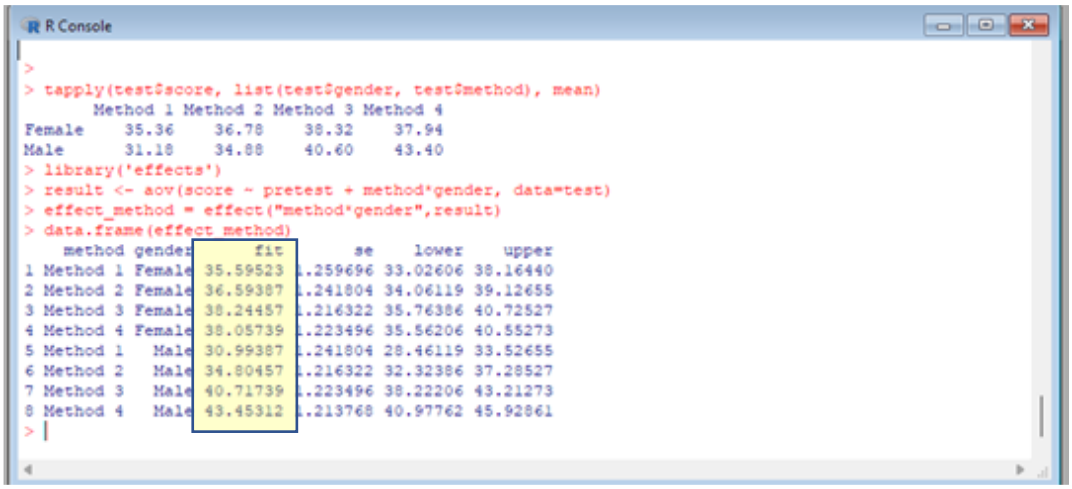


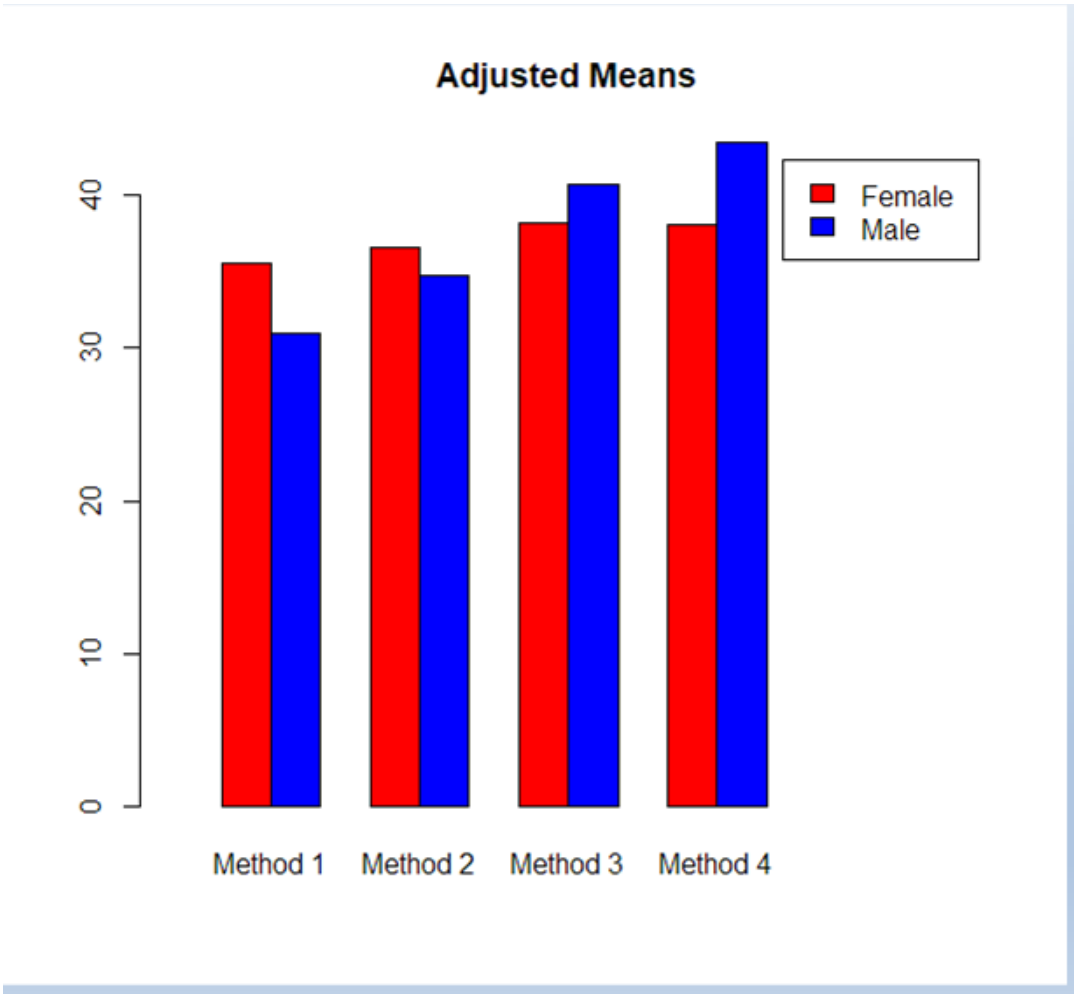
- TukeyHSD(ancova2,"test$method")



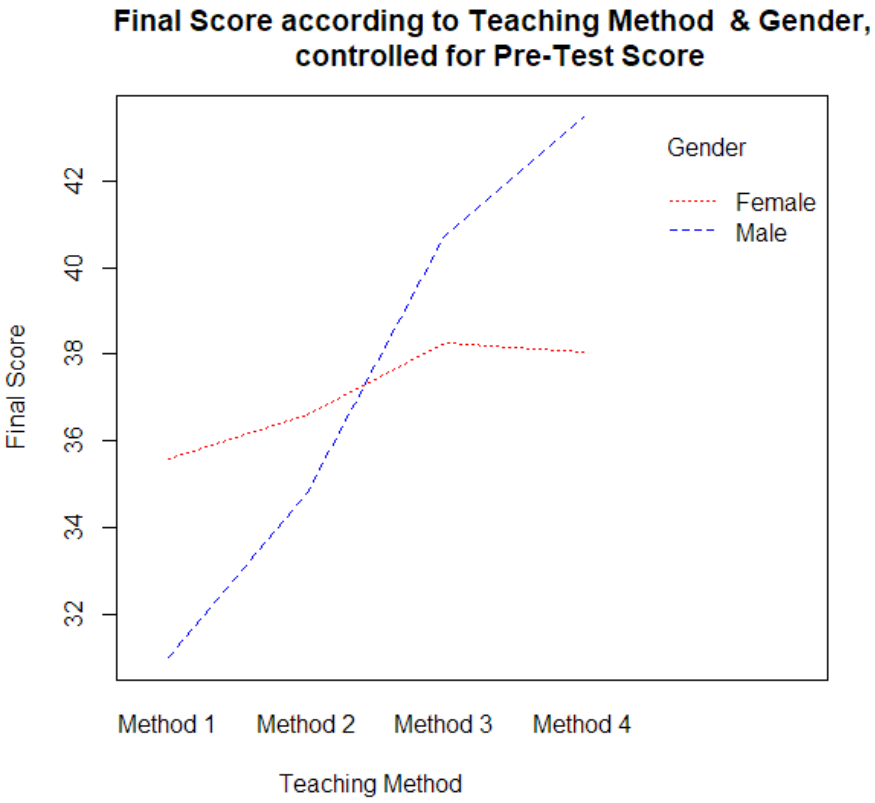
Using the “effects” package.

- library("effects")
- result <- aov(score ~ pretest + method*gender, data=test)
- effect_method = effect("method*gender",result)
- data.frame(effect_method)





- interaction.plot(means$method, means$gender, means$fit, col=c("red", "blue"), trace.label="Gender", xlab="Teaching Method", ylab="Final Score", main= "Final Score according to Teaching Method & Gender, \n controlled for Pre-Test Score")



A 2 by 2 between-groups analysis of covariance was conducted to assess the effectiveness of four methods of teaching English for male and female participants. The independent variables were the teaching method (method 1, 2, 3 or 4) and gender. The dependent variable was scores on the final test, administered following completion of the intervention programs. Scores on the pre-test administered prior to the commencement of the programs were used as a covariate to control for individual differences.
After adjusting for pre-test scores, there was a significant interaction effect: $F(1, 31) = 6.439$, $p < .001$. These results suggest that males and females respond differently to the four methods. Males showed a more substantial increase on the final test score after participation the class using methods 1 and 2 than methods 3 & 4, whereas females did not differ so much.

9. **Chi-Squared (χ^2)** (Test for independence}

Is there a significant difference between the proportion of males and females who smoke?

	Smoke	Not Smoke
Male	18	24
Female	35	14

Input data

- smokers <- matrix(c(18,35,24,14), ncol=2)
- rownames(smokers) <- c("Male", "Female")
- colnames(smokers) <- c("Smoke", "Not Smoke")

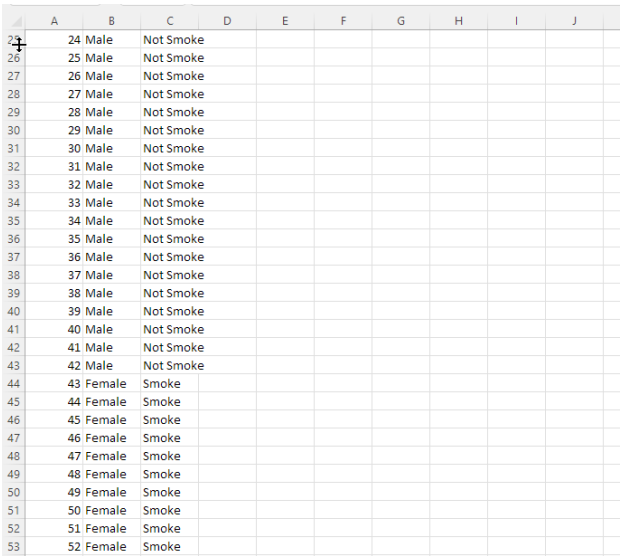
- smokers

```
> smokers <- matrix(c(18,35,24,14), ncol=2)
> rownames(smokers) <- c("Male", "Female")
> colnames(smokers) <- c("Smoke", "Not Smoke")
> smokers
       Smoke Not Smoke
Male      18         24
Female    35         14
```

- chisq.test(smokers)

```
> chisq.test(smokers)

        Pearson's Chi-squared test with Yates' continuity correction

data:  smokers
X-squared = 6.4615, df = 1, p-value = 0.01102
```

OR



	A	B	C
25	24	Male	Not Smoke
26	25	Male	Not Smoke
27	26	Male	Not Smoke
28	27	Male	Not Smoke
29	28	Male	Not Smoke
30	29	Male	Not Smoke
31	30	Male	Not Smoke
32	31	Male	Not Smoke
33	32	Male	Not Smoke
34	33	Male	Not Smoke
35	34	Male	Not Smoke
36	35	Male	Not Smoke
37	36	Male	Not Smoke
38	37	Male	Not Smoke
39	38	Male	Not Smoke
40	39	Male	Not Smoke
41	40	Male	Not Smoke
42	41	Male	Not Smoke
43	42	Male	Not Smoke
44	43	Female	Smoke
45	44	Female	Smoke
46	45	Female	Smoke
47	46	Female	Smoke
48	47	Female	Smoke
49	48	Female	Smoke
50	49	Female	Smoke
51	50	Female	Smoke
52	51	Female	Smoke
53	52	Female	Smoke

Save as "smoke.csv"

Import data into R

- smoke <- read.csv("smoke.csv", header = TRUE, colClasses = c("numeric", "factor","factor"))
- MFsmoke <- table(smoke$Sex, smoke$Smoke)
- chisq.test(MFsmoke)

```
> MFsmoke <- table(smoke$Sex, smoke$Smoke)
> MFsmoke

        Not Smoke Smoke
  Female        14    35
  Male          24    18
> chisq.test(MFsmoke)

        Pearson's Chi-squared test with Yates' continuity correction

data:  MFsmoke
X-squared = 6.4615, df = 1, p-value = 0.01102
```

p <0.05 - we can conclude that there a significant difference between males and females in their smoking activity.