QDA Lab 5 Part 1

October 2020

This lab is going to show how to run ANOVA and ANCOVA in R. It also includes a comparison to using t-tests when the explanatory variable has two levels only.

```
library(foreign)
library(ggplot2)
```

The data

This data is taken from https://people.bath.ac.uk/pssiw/stats2/page16/page16.html and it contains the IQ scores from three groups of undergraduates of different disciplines as well as their age. (The data is in SPSS format and the foreign library can be used to read this data in, alternatively this data is available as an R data object)

The research question

The research question is **"Is there a difference in mean IQ between the three groups of students?"**

```
# testing the iq data for lab

iq= read.spss("data/iqdata2.sav", to.data.frame=TRUE)
```

Exploring the data

This data has the IQ of students from three different courses and their age Firstly explore the data and visualise it:

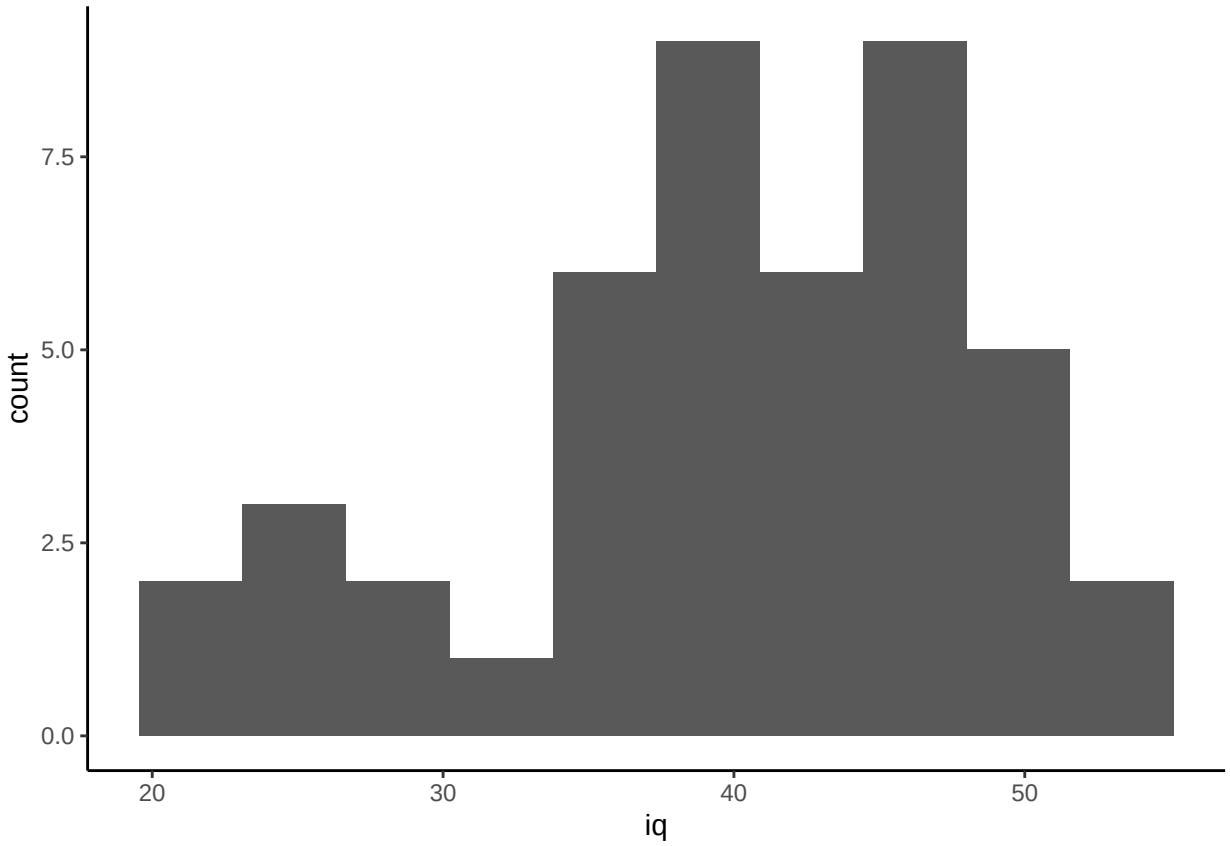
```
summary(iq)
```

```
##                group          iq              age
##  Physics student  :15   Min.   :20.00   Min.   :14.00
##  Maths student    :15   1st Qu.:36.00   1st Qu.:17.00
##  Chemistry student:15   Median :40.00   Median :18.00
##                        Mean   :39.96   Mean   :26.13
##                        3rd Qu.:46.00   3rd Qu.:40.00
##                        Max.   :52.00   Max.   :58.00
```

We can see that we have 45 observations and for each one we have the iq, the age and the group the students are in.

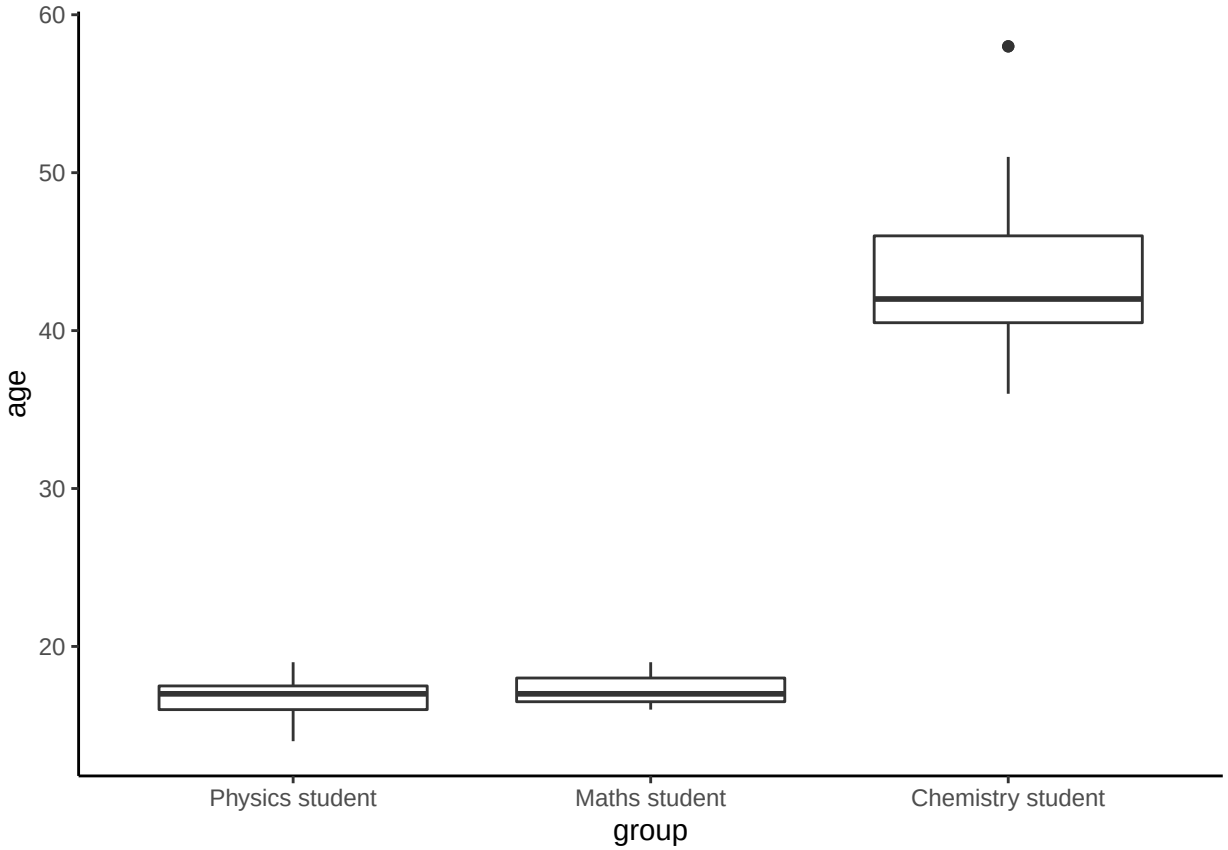
A good way to visualise the distributions is using histograms.

```
ggplot(data=iq, aes(x=iq)) + geom_histogram(bins=10) +theme_classic()
```



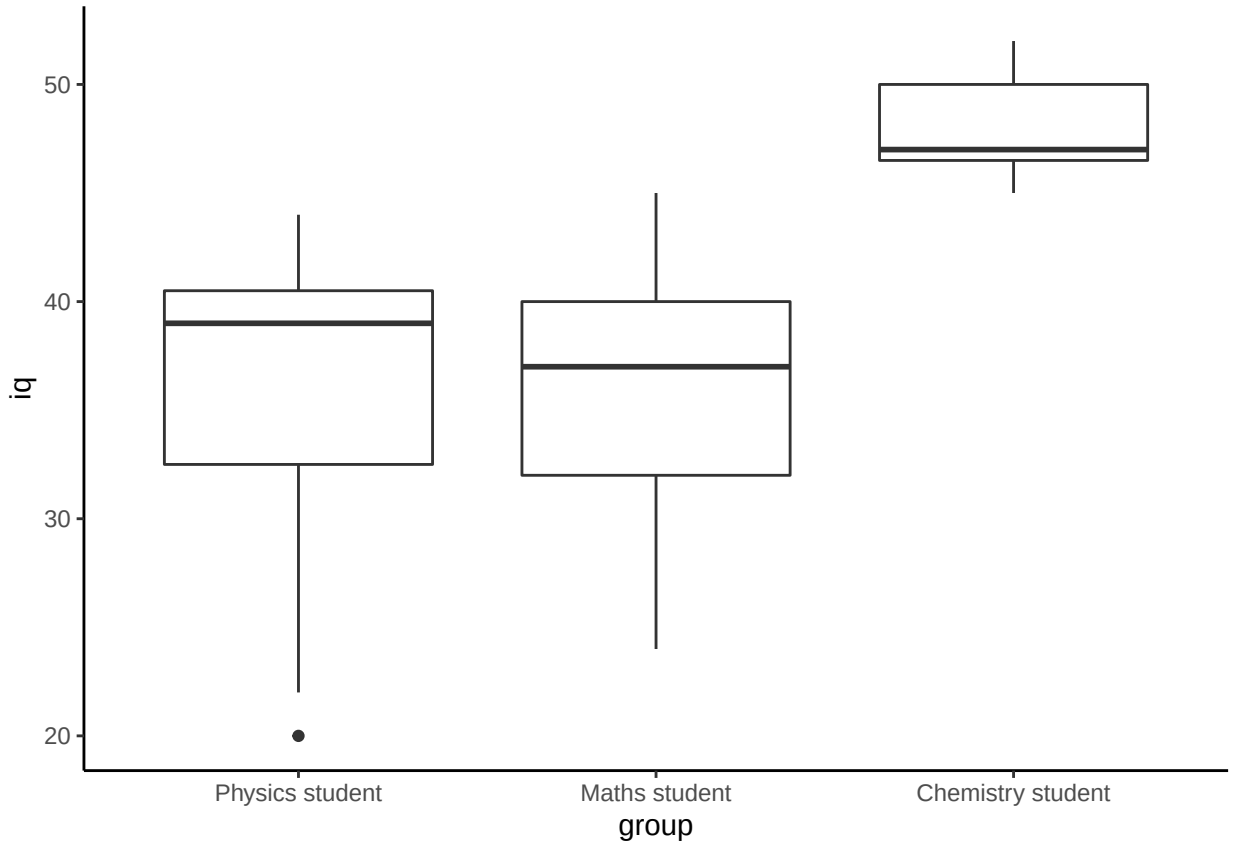
And more importantly we want to see if the mean for iq or age is different between the three groups.

```
ggplot(data = iq, aes(x=group, y=age)) +geom_boxplot() + theme_classic()
```



We can also look at the same plot for iq and group.

```
ggplot(data = iq, aes(x=group, y=iq)) +geom_boxplot() + theme_classic()
```



ANOVA model

Lets build an ANOVA model for the iq with group as the explanatory variable or factor.

```
summary(aov(iq$iq~iq$group))
```

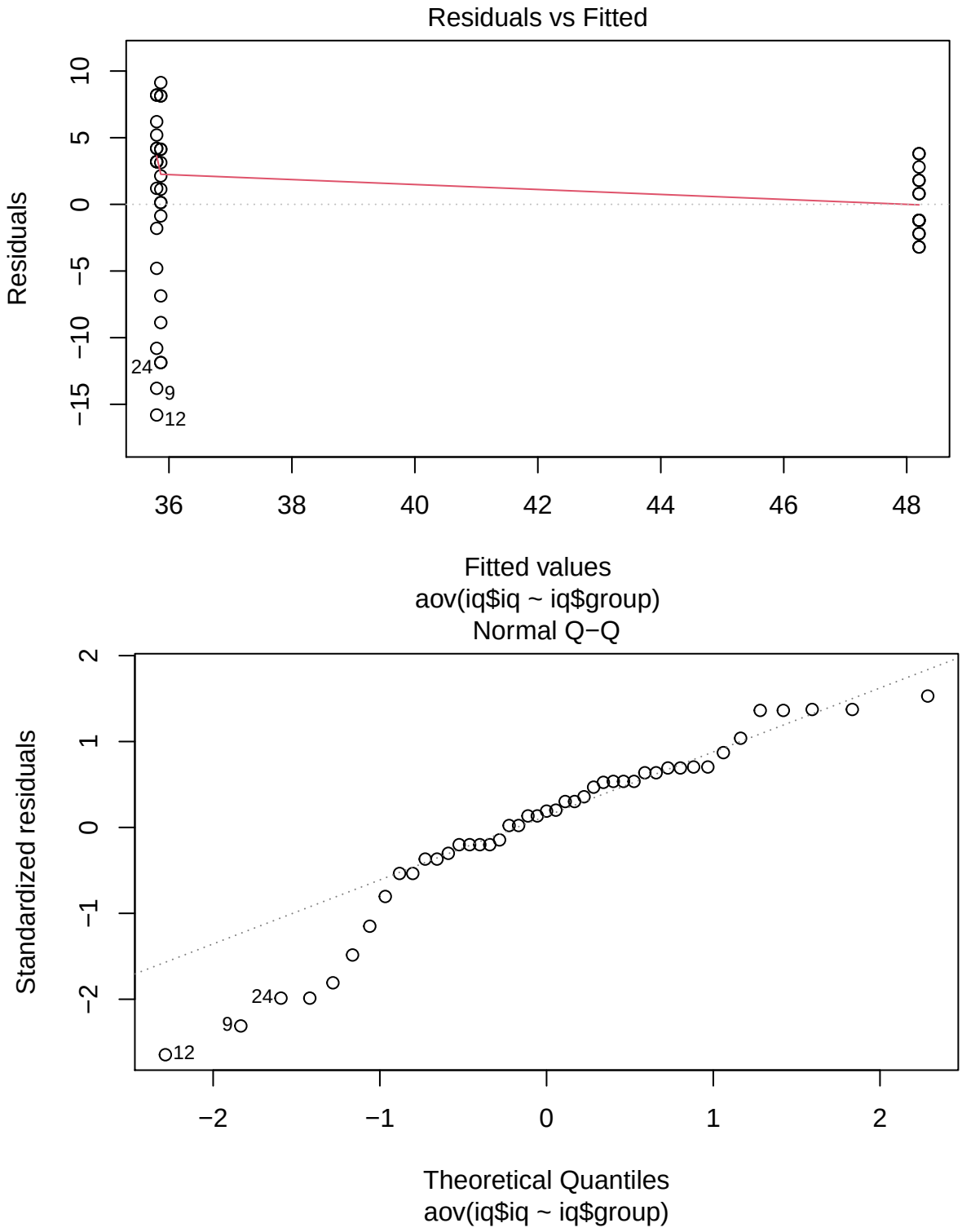
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## iq$group     2   1529   764.7   20.02 7.84e-07 ***
## Residuals   42   1604    38.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

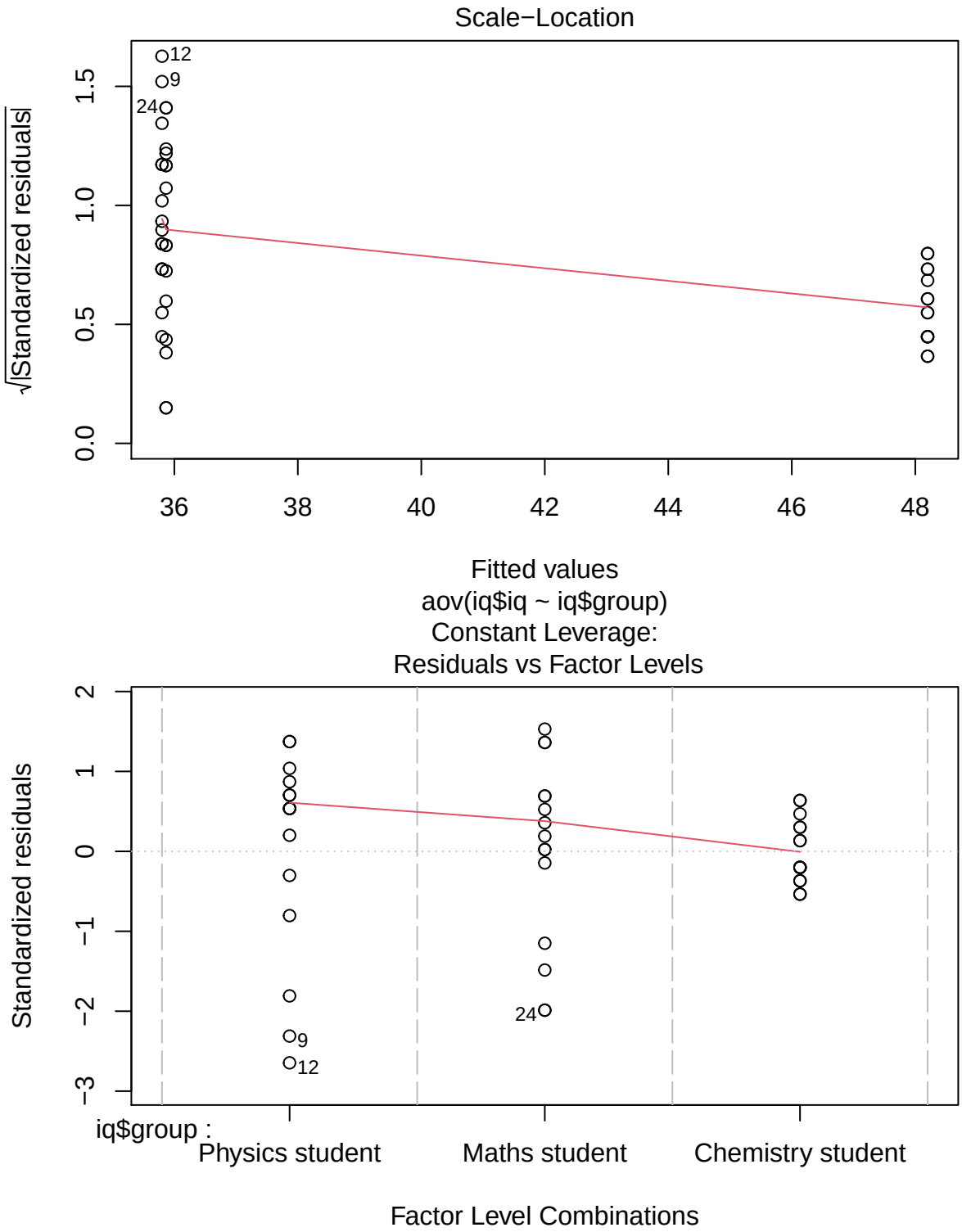
This shows us that with a ratio of F=20.02 the probability that this (or a result more extreme than this) would arise by change alone if the means were the same is extremely small.

This confirms that the groups have different means.

Lets check the diagnostic plots:

```
plot(aov(iq$iq~iq$group))
```


Residuals vs Fitted
Residuals
10
5
0
−5
−10
−15
24
9
12
36
38
40
42
44
46
48
Fitted values
aov(iq$iq ~ iq$group)
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
24
9
12
−2
−1
0
1
2
Theoretical Quantiles
aov(iq$iq ~ iq$group)



The diagnostics for this model do not point to major issues, but there are some outliers (9, 12) that can be considered for further investigation.

We also want to consider the size and direction of the effects that each group have on the dependent variable (iq)

```r
summary.lm(aov(iq$iq~iq$group))
```

```
##
## Call:
## aov(formula = iq$iq ~ iq$group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.800  -2.200   1.133   3.800   9.133
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                35.80000    1.59589  22.433  < 2e-16 ***
## iq$groupMaths student       0.06667    2.25694   0.030    0.977
## iq$groupChemistry student 12.40000    2.25694   5.494 2.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.181 on 42 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4636
## F-statistic: 20.02 on 2 and 42 DF,  p-value: 7.843e-07
```

This output can help us understand what effect each level of group has on the estimate for mean iq. The intercept also corresponds to the physics group (first one in the data file). The coefficients show the difference in the other groups. In this case the Maths students do not have a significant difference in mean IQ when compared to the physics students, however the chemistry students' mean IQ is significantly different (12.4 higher).

```r
aggregate(iq~group, data=iq, FUN="mean")
```

```
##               group       iq
## 1   Physics student 35.80000
## 2     Maths student 35.86667
## 3 Chemistry student 48.20000
```

If we simply look at the mean for each group (above) it can be seen that the mean for the physics students is the same as the intercept. And adding the coefficients from the summary.lm table to the physics students means gives us the two other means.

We may want to consider joining two categories

If we treat Physics and Maths students as one group:

```r
iq$two.groups<-ifelse(iq$group=="Chemistry student", "Chemistry", "Maths+Physics")
```

Lets see what this new attribute looks like:

```r
table(iq$two.groups)
```

```
##
##     Chemistry Maths+Physics
##            15            30
```

If we forgot about ANOVA and wanted to test the hypothesis H_0 mean(chemistry IQ) = mean (maths+physics IQ) vs H_1 Chemistry IQ is Higher.

We can do this with a t-test, proceeded by a variance test

```r
var.test(iq$iq~iq$two.groups)
```

```
##
##  F test to compare two variances
##
## data:  iq$iq by iq$two.groups
## F = 0.10927, num df = 14, denom df = 29, p-value = 8.269e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.04639089 0.29942300
## sample estimates:
## ratio of variances
##          0.1092681
```

Now we can run a t test

```r
t.test(iq$iq~iq$two.groups)
```

```
##
##  Welch Two Sample t-test
##
## data:  iq$iq by iq$two.groups
## t = 8.464, df = 39.184, p-value = 2.216e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.411784 15.321549
## sample estimates:
##     mean in group Chemistry mean in group Maths+Physics
##                    48.20000                    35.83333
```

This confirms that the difference is significant as the p-value of this t-test is very small.

Now lets do the same using ANOVA

```r
summary.lm(aov(iq$iq~iq$two.groups))
```

```
##
## Call:
## aov(formula = iq$iq ~ iq$two.groups)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.833  -2.200   1.167   3.800   9.167
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   48.200      1.577  30.560  < 2e-16 ***
## iq$two.groupsMaths+Physics  -12.367      1.932  -6.402 9.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.109 on 43 degrees of freedom
## Multiple R-squared:  0.488,  Adjusted R-squared:  0.4761
## F-statistic: 40.98 on 1 and 43 DF,  p-value: 9.51e-08
```

This has also found a significant difference , and the coefficient estimates show us that the mean iq for maths+physics is 12.367 less than that for Chemistry.

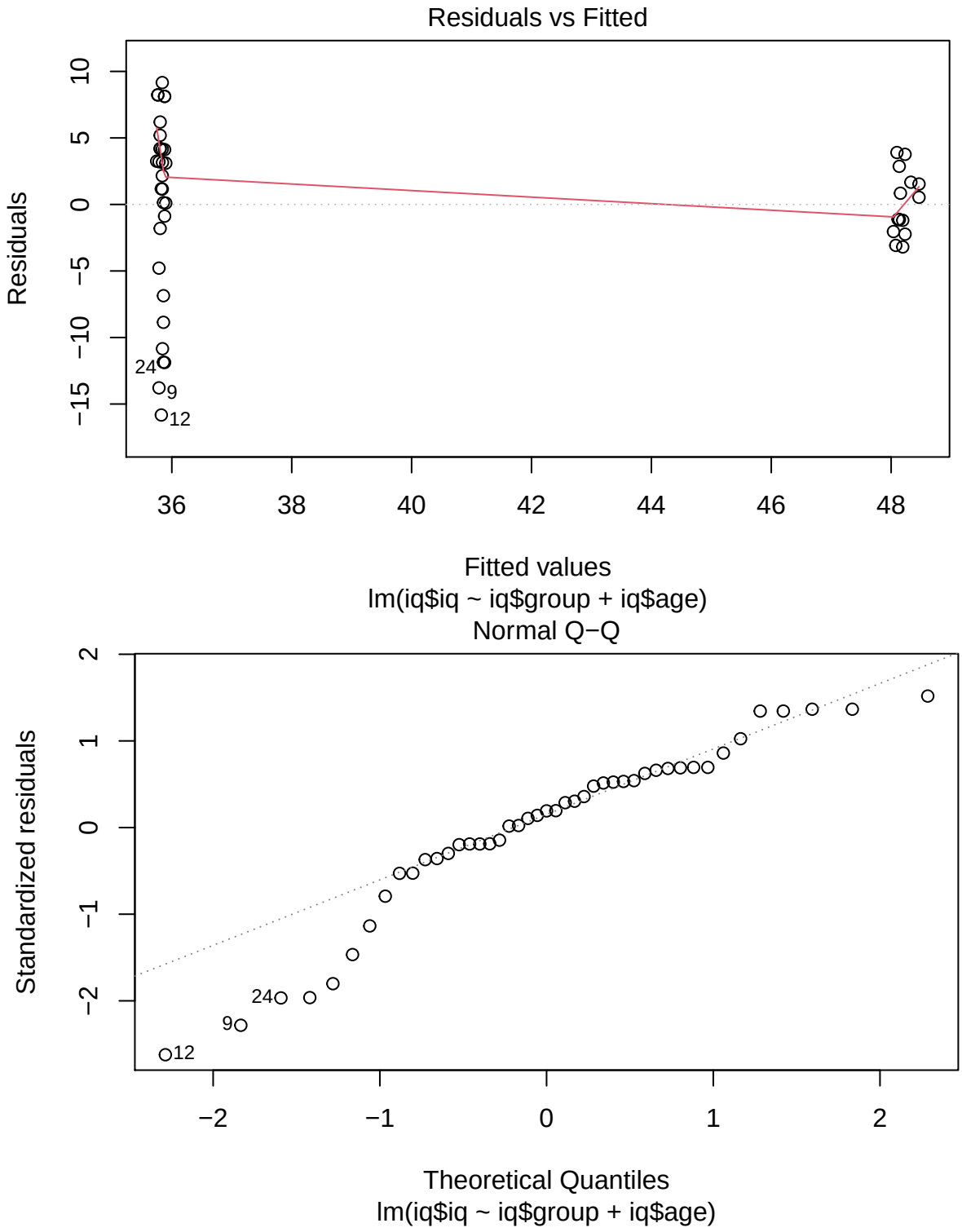
ANCOVA

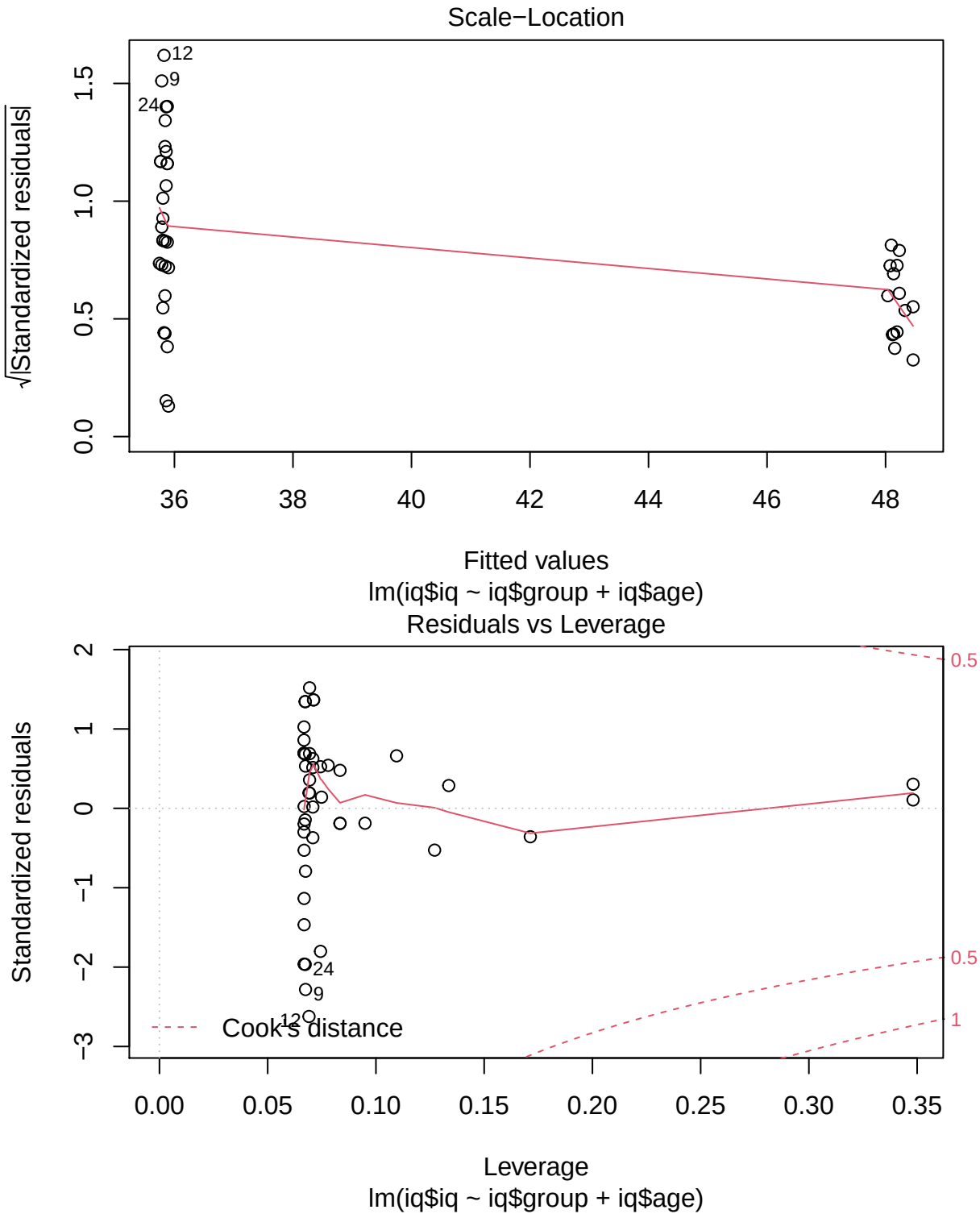
Using the same data set lets move to ANCOVA, where we also introduce another covariate (explanatory variable) that is continuous - in this case we will use age.

```
ancova.iq<-lm(iq$iq~iq$group+iq$age)
summary(ancova.iq)
```

```
##
## Call:
## lm(formula = iq$iq ~ iq$group + iq$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.825  -2.038   1.159   3.903   9.159
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                35.47555    4.37252   8.113 4.64e-10 ***
## iq$groupMaths student       0.05503    2.28876   0.024    0.981
## iq$groupChemistry student 11.86486    7.08054   1.676    0.101
## iq$age                      0.01939    0.24283   0.080    0.937
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.255 on 41 degrees of freedom
## Multiple R-squared:  0.4881, Adjusted R-squared:  0.4506
## F-statistic: 13.03 on 3 and 41 DF,  p-value: 4.066e-06
```

```
plot(ancova.iq)
```


Residuals vs Fitted
Residuals
10
5
0
−5
−10
−15
24
9
12
36
38
40
42
44
46
48
Fitted values
lm(iq$iq ~ iq$group + iq$age)
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
24
9
12
−2
−1
0
1
2
Theoretical Quantiles
lm(iq$iq ~ iq$group + iq$age)



We can see from the results of the model that age is not a significant coefficient, and the r^2 is less than 50%. This does not seems like a model that is more useful than the ANOVA using only the group as the explanatory variable.