Linear Regression - Example in R.

August 21, 2014

1 Motivation.

- We are interested in adequately describing the relationship between students exam scores *dependent variable* and their previous homework scores *the explanatory variables.*

2 Exploratory Data Analysis.

2.1 The Data.

The Data being analysed consist of exam results (%) for 91 students, for each student we also have their sex *(Male (n=51),Female (n=40))*, degree program *(Science, Joint, Arts)*, and their previous four homework results *(hw1, hw2,hw3,hw4).* Shown below are the data for the first 6 students and the command in R to obtain them.

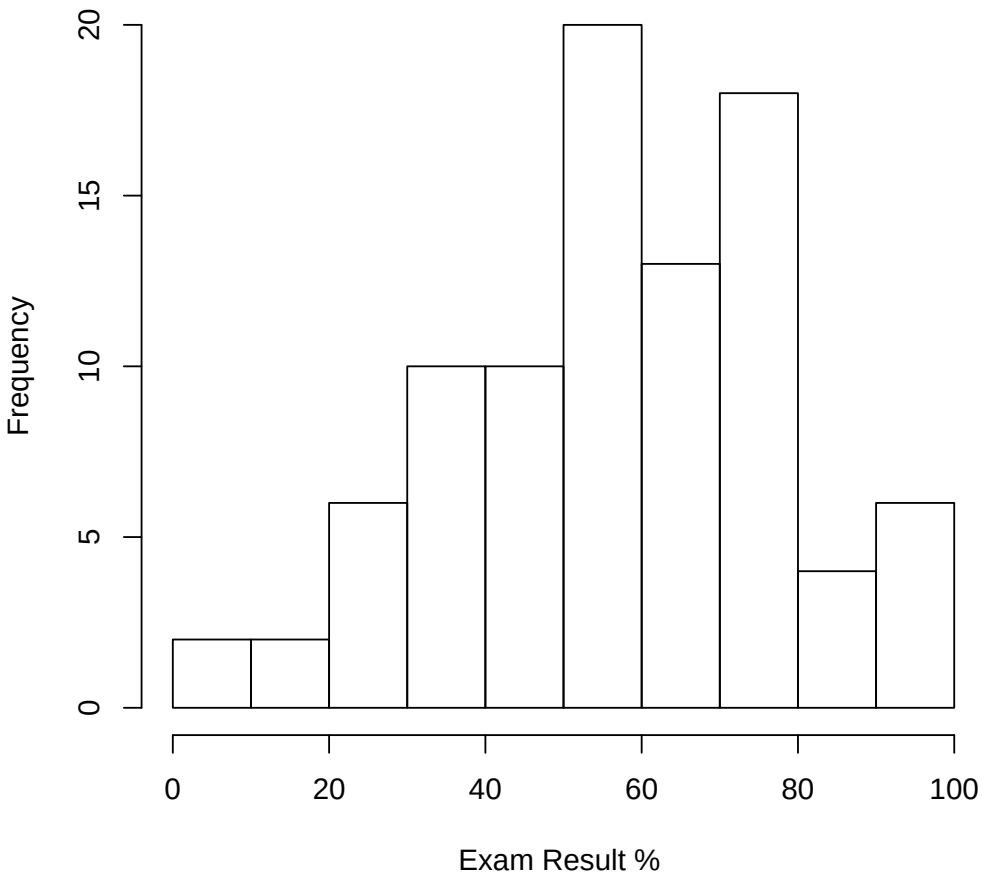
```
> head(data)
```

```
  exam sex hw1 hw2 hw3 hw4  degree high
1   34   F  18  41  27  10    arts    0
2   59   F  35  35  75  75   joint    0
3   35   F  58  51  37  42    arts    0
4   62   F  87  38  78  33   joint    0
5   86   F  90  98  90  92   joint    1
6   29   F  22  38   0  41 Science    0
```

2.2 Checking Normality.

To use parametric tests such as **t-tests, and ANOVA** we assume that our data come from a Normally distributed population. An *unofficial* way to check this is to plot a simple histogram, as can be seen below the distribution of the exam scores is seemingly normal *(actually unless the data are seriolusly skewed, or bimodal using parametric tests is generally fine, as the ANOVA is quite robust).*

```
> hist(exam,xlab="Exam Result % ",main="")
```



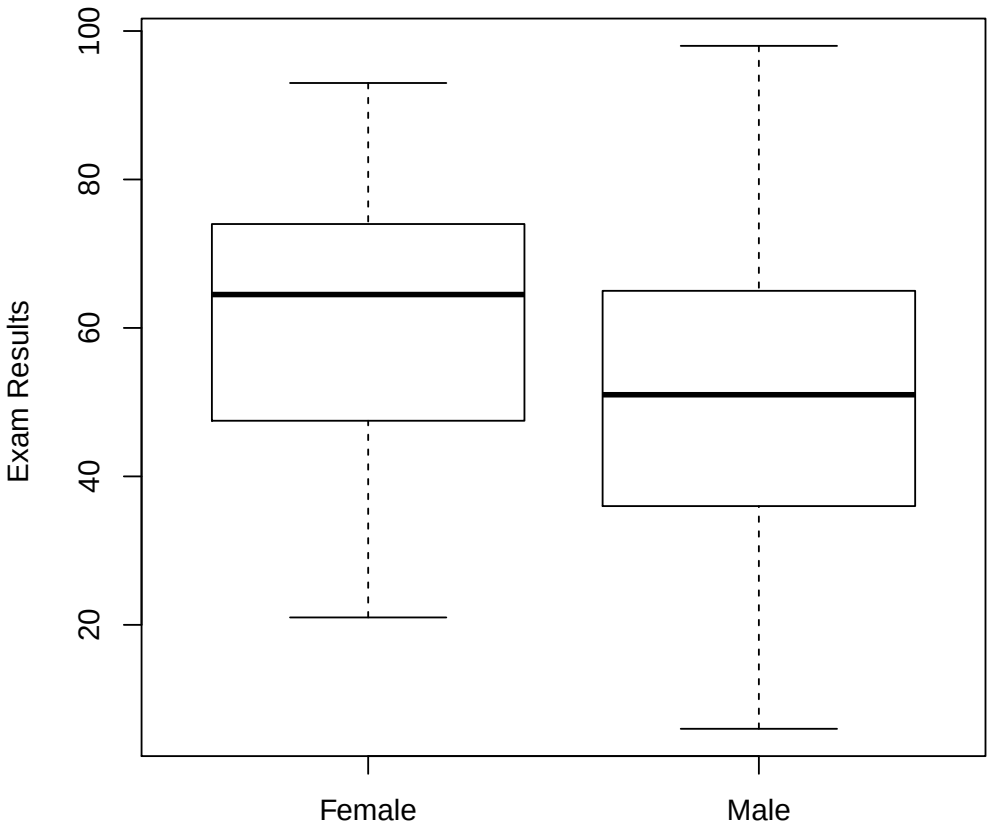

```
> shapiro.test(exam)

        Shapiro-Wilk normality test

data:  exam
W = 0.9844, p-value = 0.3507
```

For an official result we can use the Wilk-Shapiro test of normality, the null hypothesis being that the data are normally distibuted.
From the results above we have a p-value of 0.3507 hence we have no evidence to reject our null hypothesis and are therefore have no evidence against Normality.

2.3 Do Males and Females exam scores differ?

Imagine we were interested in statistically testing if there were a significant difference between the average exam scores of Male and Female students. Unofficially we may conclude from the boxplat that Male students on average score lower that Female students. Using a t-test *(as we only have two groups M and F)* we can "officially" test for a difference in means.

```
> boxplot(exam~sex,data=data,ylab="Exam Results ",names=c("Female","Male"))
```



2.3.1 T-test

To "officially" test for a difference in means between exam scores of Male and Female students we use a **two-sample t-test** *(as the samples are independent).*

```
> t.test(exam~sex,data=data)

        Welch Two Sample t-test

data:  exam by sex
t = 2.1433, df = 88.186, p-value = 0.03485
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.6618781 17.5224357
sample estimates:
mean in group F mean in group M
       60.70000        51.60784
```

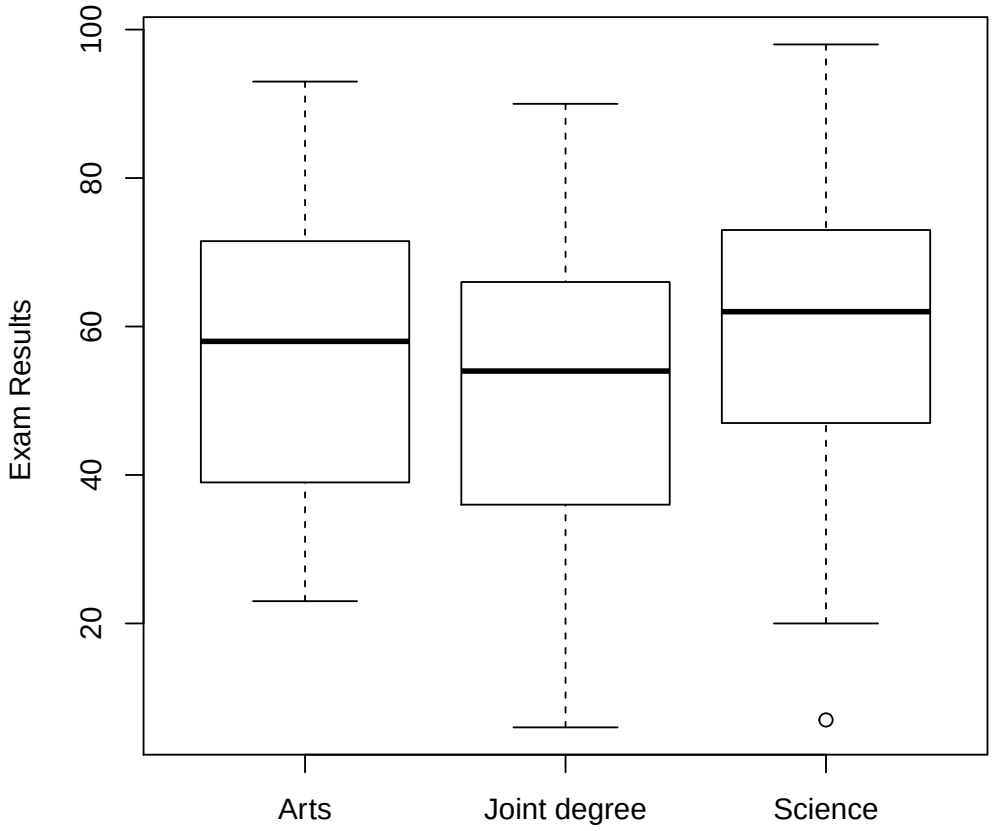
From R we get a p-value of 0.03485, hence as we used a two-sided test we can reject the null hypothesis of the group means being equal *(we have significance at the 4 % level)*, from both the boxplots and the estimated sample means

we are therefore likely to conclude that on average Female students score higher in their exams than Male students *(in our dataset).*

2.4 Do exam scores differ by dregee subject?

If we were interested in statistically testing if there were a significant difference between the average exam scores of students on different degree programs. Unofficially we may conclude from the boxplat that there is seemilgly no difference in the average scores of students taking either a Arts, Science of Joint degree. Using a one-way ANOVA *(as we three groups Arts, Joint and Science)* we can "officially" test for a difference in means.

```
> boxplot(exam~degree,data=data,ylab="Exam Results ",names=c("Arts","Joint degree","Science"))
```



2.4.1 One-Way ANOVA

To "officially" test for a difference in means between exam scores of students on different degree programs we use a **One-way ANOVA** *(as the samples are independent).*

```
> anova(lm(exam~degree,data=data))
```

```
Analysis of Variance Table
```

```
Response: exam
          Df Sum Sq Mean Sq F value Pr(>F)
degree     2    453  226.74  0.5165 0.5984
Residuals 88  38634  439.03
```

From R we get a p-value of 0.5984, hence we have no evidence to reject the null hypothesis of the group means being equal, we therefore have no evidence that students on average score higher in their exams dependent on what degree program they're on *(in our dataset)*.

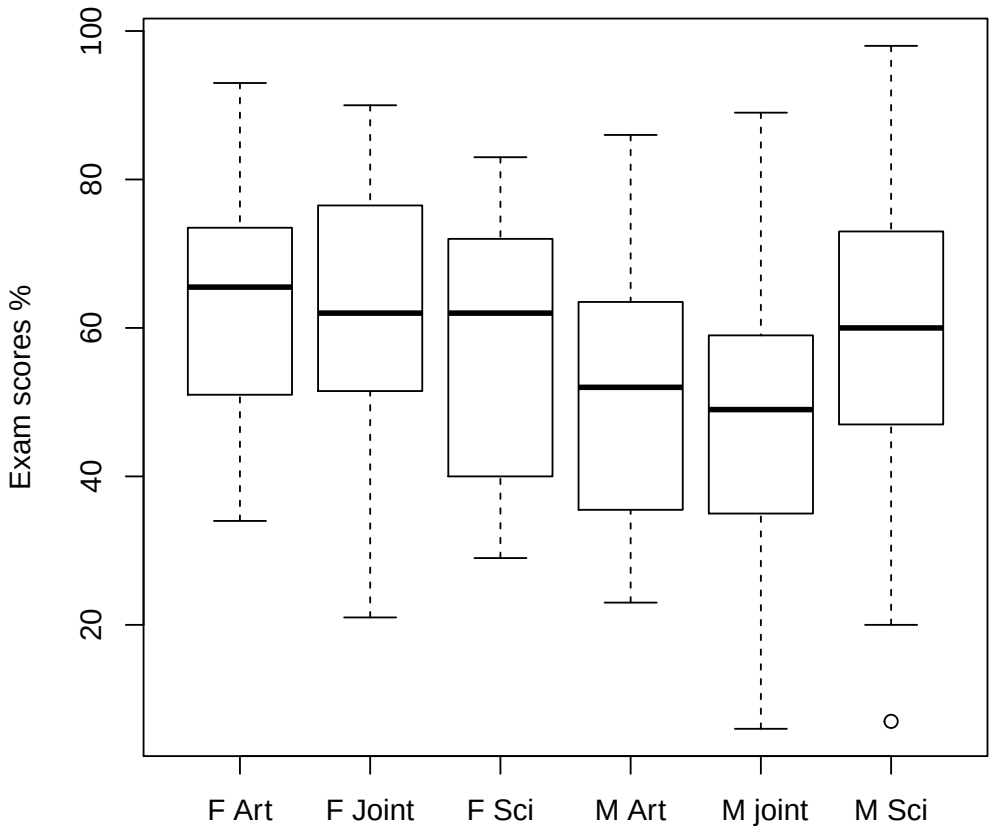
2.4.2 Two way ANOVA

To "officially" test for a difference in means between exam scores of students on different degree programs whilst simultaneously testing for a difference in means between exam scores of Male and Female students we use a **Two-way ANOVA**.

```
> anova(lm(exam~degree*sex,data=data))

Analysis of Variance Table

Response: exam
           Df Sum Sq Mean Sq F value  Pr(>F)
degree      2    453  226.74  0.5360 0.58704
sex         1   1792 1792.39  4.2372 0.04261 *
degree:sex  2    886  442.83  1.0469 0.35552
Residuals  85  35956  423.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From R we get a p-value of 0.58704 relating to degree type, hence we have no evidence to reject the null hypothesis of the group means being equal *(accounting for sex of student)*, we therefore have no evidence that students on average score higher in their exams dependent on what degree program they're on *(in our dataset)*.
From R we get a p-value 0.04261 *(we have significance at the 5 % level)* relating to sex of student, hence we can reject the null hypothesis of the group means being equal *(accounting for degree type of student)*. From the boxplots we are likely to conclude that on average Female students score higher in theis exams than Male students *(in our dataset)*.
From R we get a p-value of 0.35552 relating to the interaction term *(i.e. the term asking if the relationship between degree type and exam score is different dependent on sex)*, hence we have no evidence to reject the null hypothesis of the group means being equal.

```
> boxplot(exam~degree+sex,data=data,ylab="Exam scores %",
+         names=c("F Art","F Joint", "F Sci", "M Art", " M joint", "M Sci"))
```



3 The Linear Regression Model.

3.1 The fitted line.

Any basic statistical package will easily give us the *best fitting line* i.e. the *best* estimate for the mathematical formula describing the relationship between the dependent variable and the explanatory ones. We want to use the students previous homework scores to explain the relationship to their exam scores, as well as their sex and degree program, this is termed *multiple linear regression*. Recall the formula is,

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_p x_{pi} + \epsilon_i$$

where;

- The **intercept**, β_0 of the line, gives the expected value of Y when all $x's = 0$.
- The **slopes**, $\beta_j, j = 1...p$ where p is the number of explanatory variables, give the increase (or decrease) in Y for a **unit** increase in each $x_j i$, given the other explanatory variables in the model.

```
> mod<-lm(exam~hw1 + hw2 + hw3 + hw4 + degree*sex,data=data)
> summary(mod)
```

```
Call:
lm(formula = exam ~ hw1 + hw2 + hw3 + hw4 + degree * sex, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.021  -7.154   0.257   6.696  27.160

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.22788    4.80236   0.464  0.64395
hw1                 0.32375    0.05089   6.362 1.11e-08 ***
hw2                 0.16230    0.05502   2.950  0.00416 **
hw3                 0.22001    0.05444   4.042  0.00012 ***
hw4                 0.18622    0.06011   3.098  0.00268 **
degreejoint        -0.10499    4.09184  -0.026  0.97959
degreeScience      -4.46427    4.27262  -1.045  0.29920
sexM                2.72409    4.40265   0.619  0.53782
degreejoint:sexM   -0.54350    5.68821  -0.096  0.92412
degreeScience:sexM  5.15083    5.81793   0.885  0.37860
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.22 on 81 degrees of freedom
Multiple R-squared:  0.7835,        Adjusted R-squared:  0.7594
F-statistic: 32.56 on 9 and 81 DF,  p-value: < 2.2e-16
```

Recall the assumptions.

- The explanatory variables are related **linearly** to the response.
- The errors have constant variance.
- The errors are independent.
- The errors are Normally distributed.

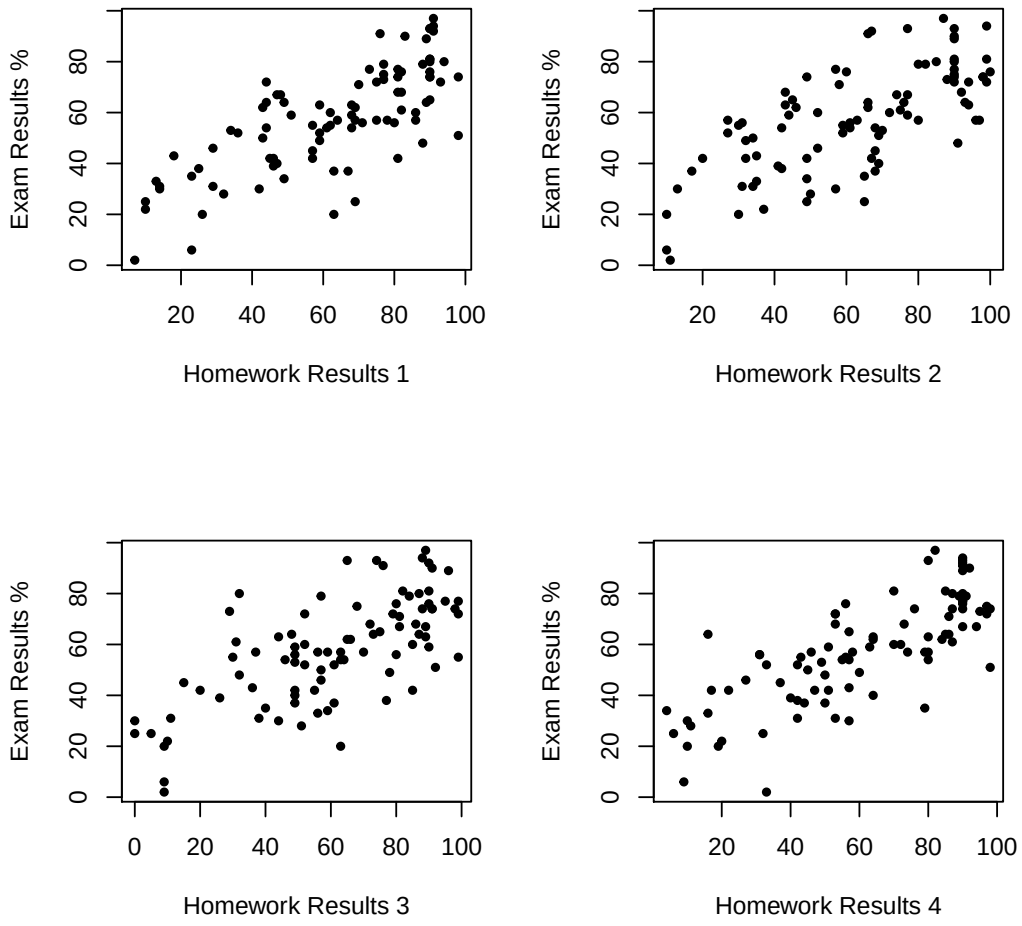
3.2 Checking the assumptions.

3.2.1 How to check the assumptions.

- Plotting the residuals against the explanatory variable will indicate if the wrong model has been fitted (i.e. higher order terms are needed) or if there is some dependence on some other explanatory variable. If this is the case some obvious patterning will be visible in the plot.
- Plotting the residuals in order, any trend visible may indicate *seasonal pattern* or *autocorrelation*.
- Plotting the residuals against the fitted values will indicate if there is non-constant error variance, i.e. if the variance increases with the mean the residuals will fan out as the fitted value increases. Usually transforming the data, or using another distribution will help.
- A Normal probability plot, histogram of the residuals or say a Wilk-Shapiro test will indicate if the normality assumption is valid, however high non-normality should have been picked up from exploring the data initially.

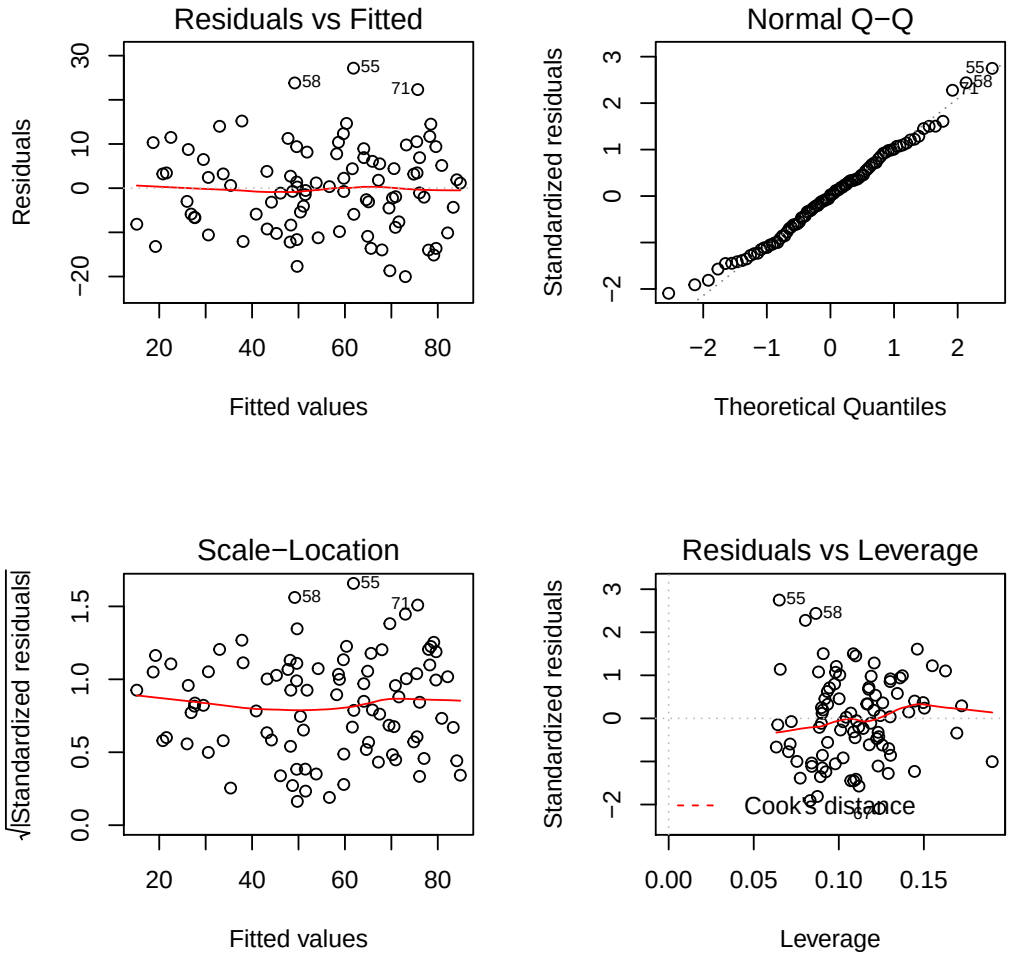
To check the linearity assumption a simple plot of each continious covariate against the response will suffice. As we can see there is no cause for concern as each covariate *(hw1,hw2,hw3,hw4)* clearly has a linear relationship with the response *(exam)*

```
> par(mfrow=c(2,2))
> plot(hw1,exam,pch=20,main="",ylab="Exam Results %",xlab="Homework Results 1")
> plot(hw2,exam,pch=20,main="",ylab="Exam Results %",xlab="Homework Results 2")
> plot(hw3,exam,pch=20,main="",ylab="Exam Results %",xlab="Homework Results 3")
> plot(hw4,exam,pch=20,main="",ylab="Exam Results %",xlab="Homework Results 4")
```



From the plots below we can see from the **Normal Q-Q** plot that it is reasonable to assume that our erros are normally distributed *(the residuals follow the 45‘ line)*. From the **Fitted values vs Residuals** plot we can see that we can assume constant error variance *(there is a seemingly random scatter of points)*. From the **Residuals vs covariats *(hw1,hw2,hw3,hw4)*** plots we can see that we can legitametely assume constant error variance *(there is a seemingly random scatter of points)*. From the **Ordered Residuals** plot we can see that we can legitametely assume independence *(there is a seemingly random scatter of points)*.

```
> par(mfrow=c(2,2))
> plot(mod)
> plot(residuals(mod),hw1,xlab="residuals")
> plot(residuals(mod),hw2,xlab="residuals")
> plot(residuals(mod),hw3,xlab="residuals")
> plot(residuals(mod),hw4,xlab="residuals")
> plot(order(residuals(mod)),ylab="Residuals in Order")
```



3.3 Interpreting the model.

From the summary output above essentially we are testing;

- $H_0 : \beta_0 = 0$ verses $H_1 : \beta_0 \neq 0$ and;
- $H_0 : \beta_1 = 0$ verses $H_1 : \beta_1 \neq 0$

P-values indicate that all previous homework assignments and sex of students are important in predeicitng final exam scores *(in our dataset)*, whereas degree type of student seemingly isn't important neither are the interaction terms. The **intercept** estimate of the line, gives the expected value of exam scores when all covariates are zero, i.e. for our model on average the expected exam score increases by 2.22788 if all homework assingments scored zero taking into account the baseline *(i.e (Female))* . The **estimate** column in the **summary** output tells us that there is on average an 0.32375 *significant* increase in exam score for an **unit** increase in score for the first homework *(hw1)*, given the other explanatory variables in the model.

3.4 Model Comparison.

We further fit an additional model discarding the considered *insignificant* covariates, degree type and sex as well as the intercet term, syntax below.

```
> mod2<-lm(exam~0 + hw1 + hw2 + hw3 + hw4,data=data)
> summary(mod2)

Call:
lm(formula = exam ~ 0 + hw1 + hw2 + hw3 + hw4, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.1260  -5.9442  -0.2331   7.2985  27.6988

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
hw1  0.30989    0.04707   6.583 3.34e-09 ***
hw2  0.20879    0.05017   4.162 7.39e-05 ***
hw3  0.24866    0.05100   4.876 4.84e-06 ***
hw4  0.16802    0.05618   2.991  0.00362 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.24 on 87 degrees of freedom
Multiple R-squared:  0.9715,        Adjusted R-squared:  0.9702
F-statistic:   742 on 4 and 87 DF,  p-value: < 2.2e-16
```

As before the **estimate** column in the **summary** output tells us that there is on average an 0.30989 *significant* increase in exam score for an **unit** increase in score for the first homework *(hw1)*, given the other explanatory variables in the model, a 0.20879 *significant* increase in exam score for an **unit** increase in score for the second homework *(hw2)*, a 0.24866 *significant* increase in exam score for an **unit** increase in score for the third homework *(hw3)*, a 0.16802 *significant* increase in exam score for an **unit** increase in score for the fourth homework *(hw4)*.

3.5 The *"best" fitting model.*

For a simple linear model we can look at the R^2 values, *(or the Adjusted R^2)*, which is measure of fit statistic telling us how much of the variation in the data the model is explaining. The initial model with all the possible covariates has an adjusted R^2 value of **0.7594**, telling us that the model successfully explains **75**% of the variation in the data. Our second reduced model has an adjusted R^2 value of **0.9702**, telling us that the model successfully explains **97**% of the variation in the data.
Clearly therefore the second model is preferable, and we can conclude that only students previous homework scores are important in predicting their exam scores.

3.6 Conclusions.

- From the **two sample t-test** we can conclude that there is a significant difference between exam scores of Male and Female students.
- From the **one-way ANOVA** we have no evidence to suggest that exam scores of Arts, Science or Joint degree students differ.
- From the **two-way ANOVA** we have no evidence to suggest that exam scores of Arts, Science or Joint degree students differ, but can conclude that there is still a significant difference in exam scores of Male and Female students accounting for degree type. We can also conclude that there is no evidence to suggest a significant difference in the relationship between degree type and exam score dependent on sex of student.
- From the regression analysis we note that only previous homework scores are important in predicting a students exam score, once their homework scores have been included the sex of the student is no longer considered an important factor in predicting exam score.