Lecture Assignment 5 - R output analyses

Mary Lesperance

March 20, 2018

Contents

1 Student Performance Data Set

This data set describes student achievement and related variables of students from two Portuguese secondary schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). We will use the Mathematics data set. A description of the variables in in the file called StudentPerformanceData.pdf in the CourseSpaces Assignment5 folder.

Reference: https://archive.ics.uci.edu/ml/datasets/student+performance

1.1 Data preparation and descriptives

```
student=read.table("student-mat.csv",sep=";",header=TRUE)   #Math data set
print(nrow(student))
```

```
## [1] 395
```

```
head(student)
```

```
##   school sex age address famsize Pstatus Medu Fedu     Mjob     Fjob
## 1     GP   F  18       U     GT3       A    4    4  at_home  teacher
## 2     GP   F  17       U     GT3       T    1    1  at_home    other
## 3     GP   F  15       U     LE3       T    1    1  at_home    other
## 4     GP   F  15       U     GT3       T    4    2   health services
## 5     GP   F  16       U     GT3       T    3    3    other    other
## 6     GP   M  16       U     LE3       T    4    3 services    other
##       reason guardian traveltime studytime failures schoolsup famsup paid
## 1     course   mother          2         2        0       yes     no   no
## 2     course   father          1         2        0        no    yes   no
## 3      other   mother          1         2        3       yes     no  yes
## 4       home   mother          1         3        0        no    yes  yes
## 5       home   father          1         2        0        no    yes  yes
## 6 reputation   mother          1         2        0        no    yes  yes
##   activities nursery higher internet romantic famrel freetime goout Dalc
## 1         no     yes    yes       no       no      4        3     4    1
## 2         no      no    yes      yes       no      5        3     3    1
```

```
## 3         no     yes    yes      yes       no      4        3     2    2
## 4        yes     yes    yes      yes      yes      3        2     2    1
## 5         no     yes    yes       no       no      4        3     2    1
## 6        yes     yes    yes      yes       no      5        4     2    1
##   Walc health absences G1 G2 G3
## 1    1      3        6  5  6  6
## 2    1      3        4  5  5  6
## 3    3      3       10  7  8 10
## 4    1      5        2 15 14 15
## 5    2      5        4  6 10 10
## 6    2      5       10 15 15 15
```

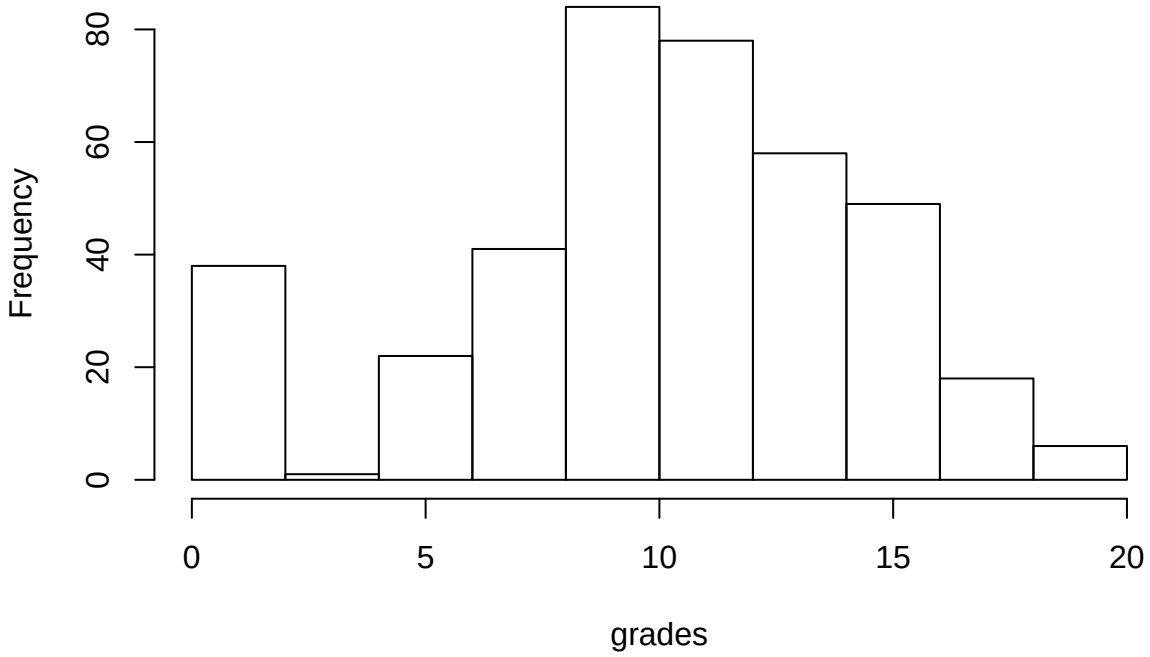
```
summary(student)
```

```
##  school   sex          age       address famsize   Pstatus      Medu
##  GP:349   F:208   Min.   :15.0   R: 88   GT3:281   A: 41   Min.   :0.000
##  MS: 46   M:187   1st Qu.:16.0   U:307   LE3:114   T:354   1st Qu.:2.000
##                   Median :17.0                             Median :3.000
##                   Mean   :16.7                             Mean   :2.749
##                   3rd Qu.:18.0                             3rd Qu.:4.000
##                   Max.   :22.0                             Max.   :4.000
##       Fedu             Mjob           Fjob            reason
##  Min.   :0.000   at_home : 59   at_home : 20   course    :145
##  1st Qu.:2.000   health  : 34   health  : 18   home      :109
##  Median :2.000   other   :141   other   :217   other     : 36
##  Mean   :2.522   services:103   services:111   reputation:105
##  3rd Qu.:3.000   teacher : 58   teacher : 29
##  Max.   :4.000
##    guardian     traveltime      studytime        failures      schoolsup
##  father: 90   Min.   :1.000   Min.   :1.000   Min.   :0.0000   no :344
##  mother:273   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000   yes: 51
##  other : 32   Median :1.000   Median :2.000   Median :0.0000
##               Mean   :1.448   Mean   :2.035   Mean   :0.3342
##               3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##               Max.   :4.000   Max.   :4.000   Max.   :3.0000
##  famsup     paid     activities nursery   higher    internet  romantic
##  no :153   no :214   no :194    no : 81   no : 20   no : 66   no :263
##  yes:242   yes:181   yes:201    yes:314   yes:375   yes:329   yes:132
##
##
##
##
##      famrel         freetime         goout            Dalc
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :4.000   Median :3.000   Median :3.000   Median :1.000
##  Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481
##  3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
##       Walc           health         absences            G1
##  Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00
##  Median :2.000   Median :4.000   Median : 4.000   Median :11.00
##  Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91
##  3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00
```

```
##  Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00
##        G2              G3
##  Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 9.00   1st Qu.: 8.00
##  Median :11.00   Median :11.00
##  Mean   :10.71   Mean   :10.42
##  3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :20.00
```

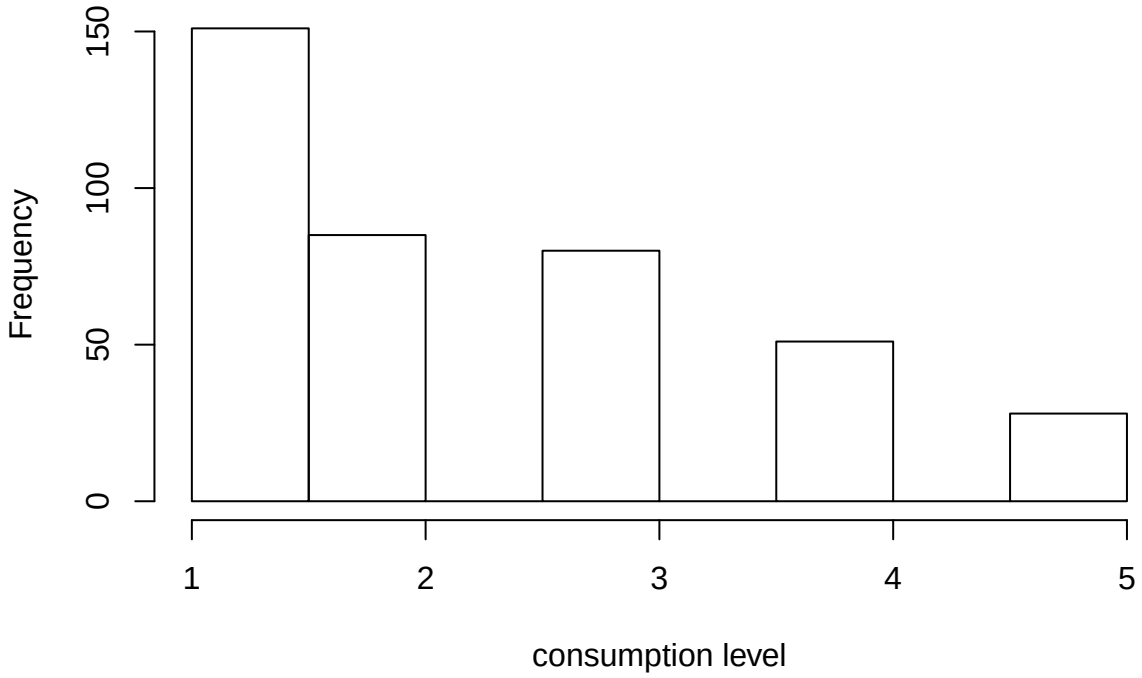
```
hist(student$G3, main='Figure 1: Histogram of student grades', xlab='grades')
```

Figure 1: Histogram of student grades




```
hist(student$Walc, main='Figure 2: Histogram of Weekend Alcohol Consumption', xlab='consumption level')
```

Figure 2: Histogram of Weekend Alcohol Consumption



1.2 Analyze the grades for first 10 students

1.2.1 Inferences for the mean, μ

```
y<-head(student$G3,10)
y
```

```
##  [1]  6  6 10 15 10 15 11  6 19 15
```

```
#mean of the first 10 salaries
mean(y)
```

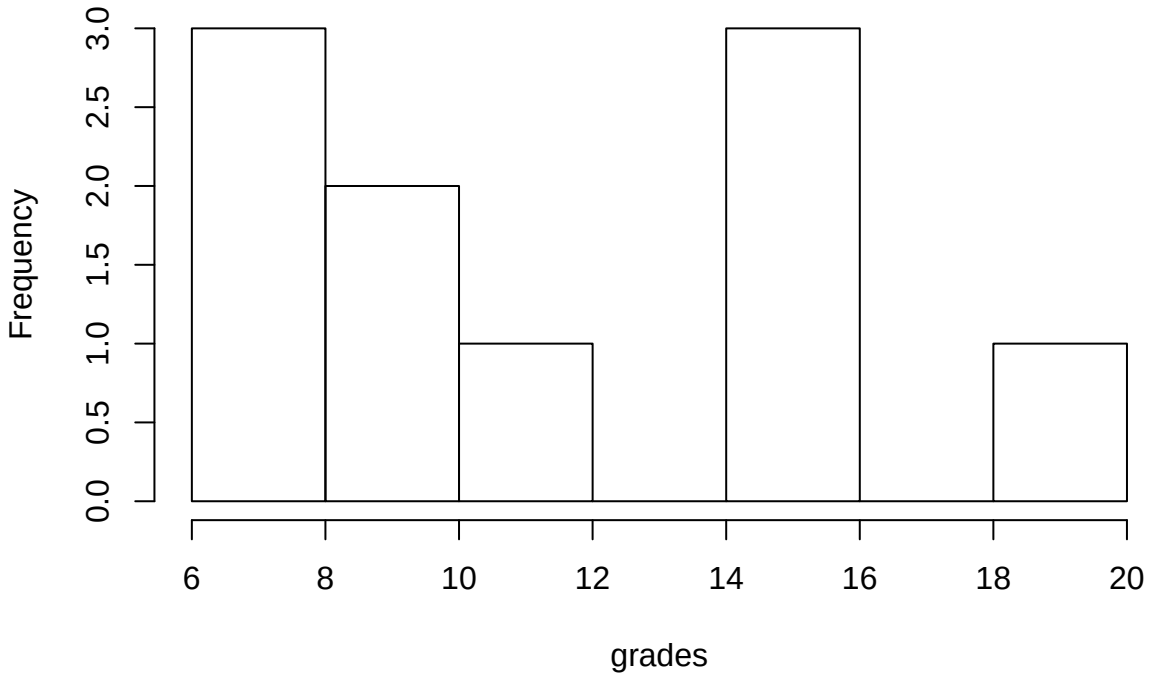
```
## [1] 11.3
```

```
#sd of the first 10 salaries
sd(y)
```

```
## [1] 4.571652
```

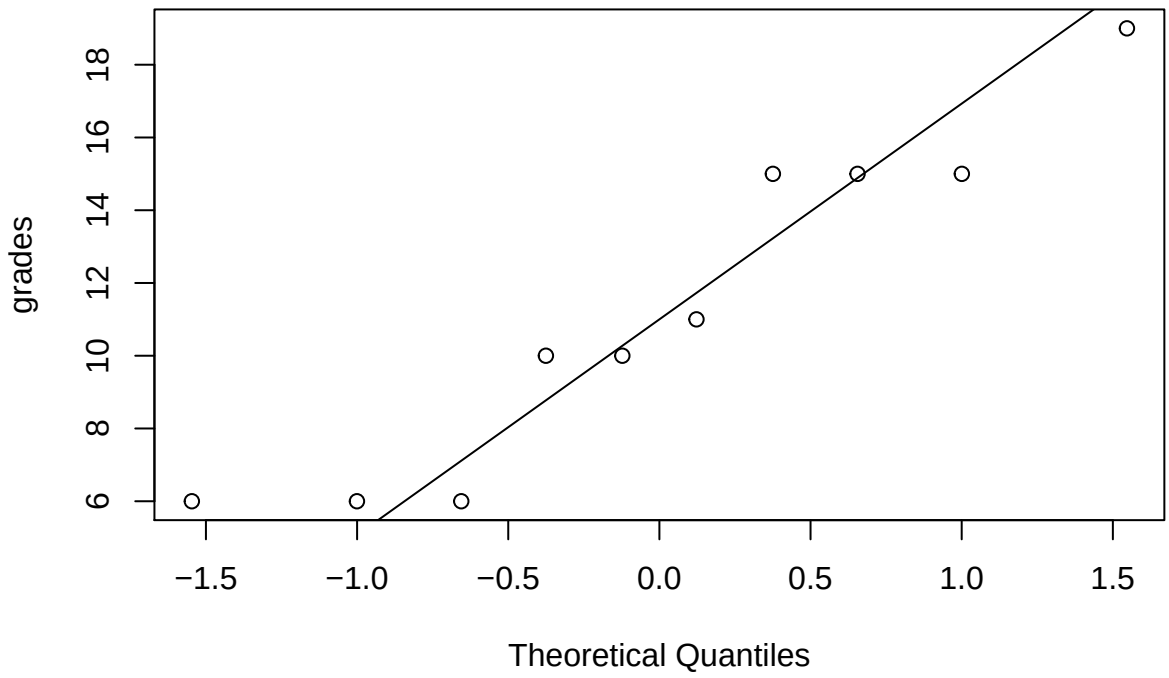
```
hist(y,main='Figure 3: Histogram of 10 student grades', xlab='grades')
```

Figure 3: Histogram of 10 student grades




```
qqnorm(y,main='Figure 4: Normal QQ plot of 10 student grades', ylab='grades')
qqline(y, lty=1)
```

Figure 4: Normal QQ plot of 10 student grades



Since σ^2 is unknown, we need to use the t-distribution to compute a confidence interval for the mean grade based on our sample of size $n = 10$ observations.

```
#For the small sample of size 10
mean(y)+c(-1,1)*qt(.975,9)*sd(y)/sqrt(length(y))
```

```
## [1]  8.029637 14.570363
```

```
qt(.975,9)
```

```
## [1] 2.262157
```

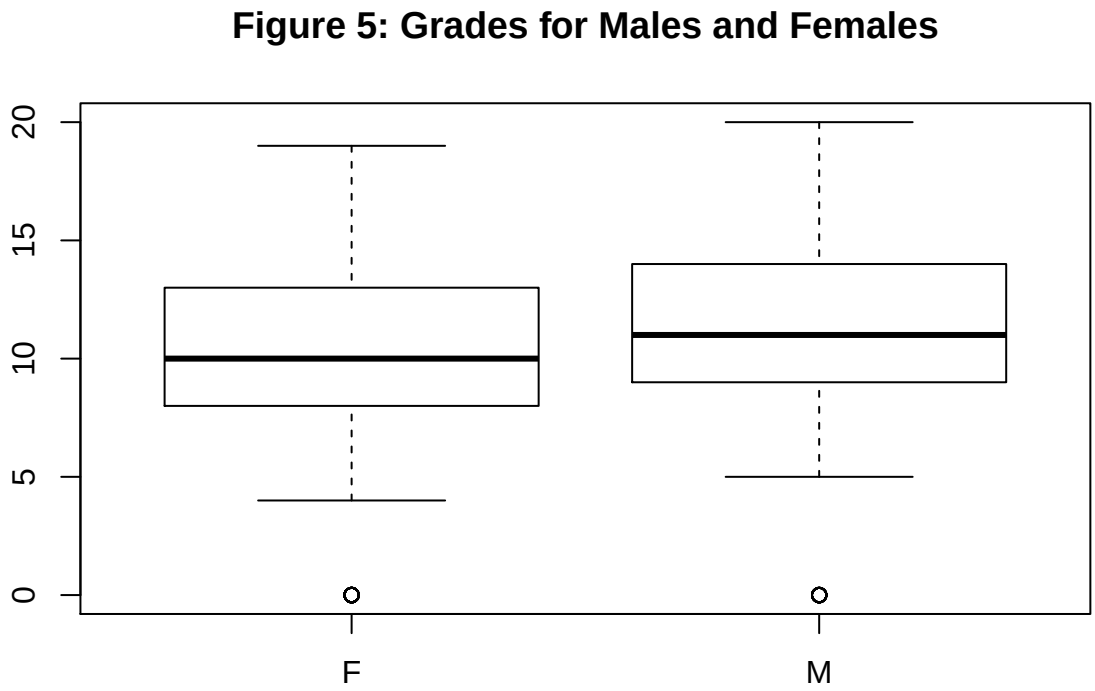
```
t.test(y, mu=10)
```

```
##
##  One Sample t-test
##
## data:  y
## t = 0.89923, df = 9, p-value = 0.392
## alternative hypothesis: true mean is not equal to 10
## 95 percent confidence interval:
##   8.029637 14.570363
## sample estimates:
## mean of x
##      11.3
```

1.3 Compare grades for males and females

1.3.1 Inferences about the differences in the means, $\mu_1 - \mu_2$

```
#Graph the data; side-by-side boxplots are one of my favourites
boxplot(G3~sex,data=student,main='Figure 5: Grades for Males and Females')
```



1.3.2 Inferences for the differences: Assume variances equal and unknown

```
#this uses pooled estimate of variance for test that H0: mu_F - mu_M = 0
t.test(G3~sex, data=student,var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  G3 by sex
## t = -2.062, df = 393, p-value = 0.03987
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.85205632 -0.04412838
## sample estimates:
## mean in group F mean in group M
##        9.966346       10.914439
```

1.3.3 Inferences for the differences: Do not assume variances are equal

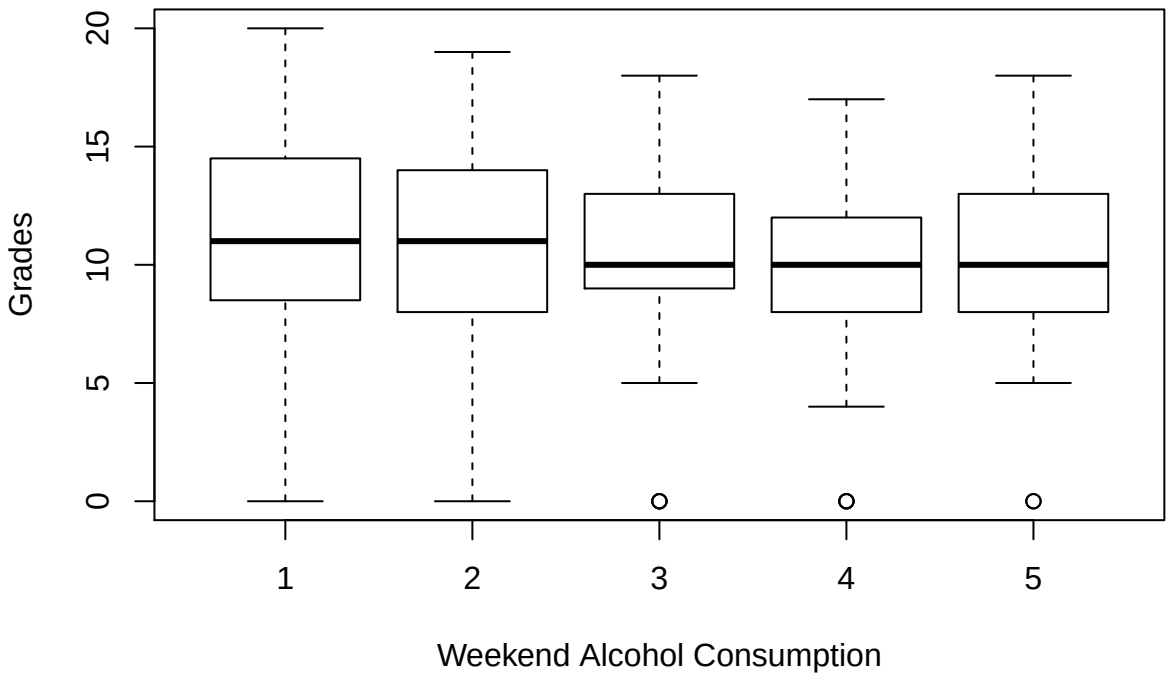
```
#this does NOT use pooled estimate of variance for test that H0: mu_F - mu_M = 0
t.test(G3~sex, data=student,var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  G3 by sex
## t = -2.0651, df = 390.57, p-value = 0.03958
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.85073226 -0.04545244
## sample estimates:
## mean in group F mean in group M
##        9.966346       10.914439
```

1.4 Regression analysis of Grades as a function of Weekend Alcohol Consumption

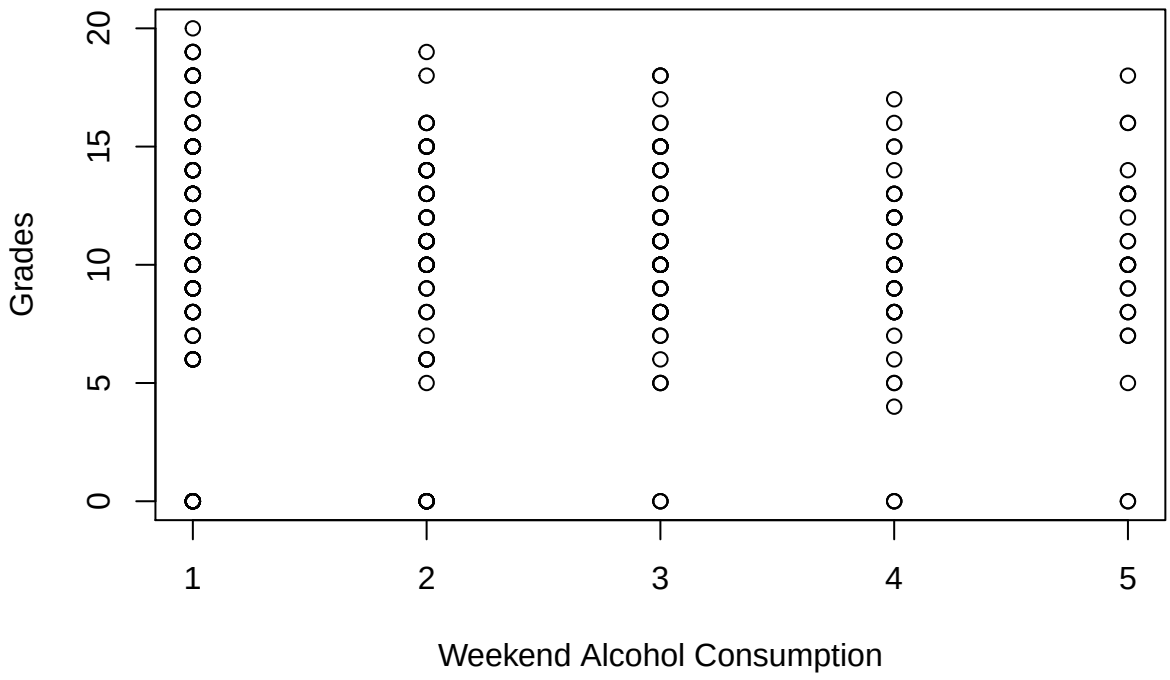
```
#Graph the data; side-by-side boxplots are one of my favourites

boxplot(student$G3~student$Walc,
    xlab='Weekend Alcohol Consumption',ylab='Grades',
     main='Figure 6: Grades versus Weekend Alcohol Consumption')
```

Figure 6: Grades versus Weekend Alcohol Consumption




```
plot(student$G3~student$Walc,
     xlab='Weekend Alcohol Consumption',ylab='Grades',
     main='Figure 7: Grades versus Weekend Alcohol Consumption')
```

Figure 7: Grades versus Weekend Alcohol Consumption

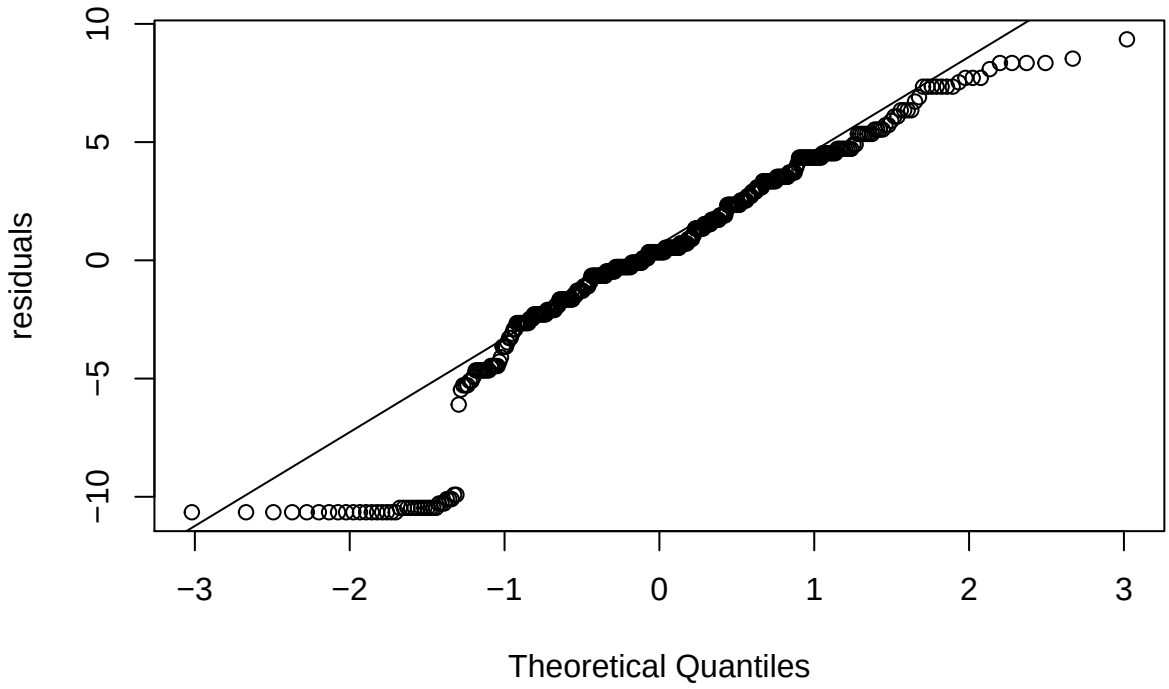



```
#Fit the regression model
student.lm<-lm(G3~Walc, data=student)
summary(student.lm)
```

```
##
## Call:
## lm(formula = G3 ~ Walc, data = student)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6537  -2.0071   0.3463   3.3463   9.3463
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8385     0.4708  23.019   <2e-16 ***
## Walc         -0.1848     0.1792  -1.031    0.303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.581 on 393 degrees of freedom
## Multiple R-squared:  0.002698,   Adjusted R-squared:  0.00016
## F-statistic: 1.063 on 1 and 393 DF,  p-value: 0.3032
```

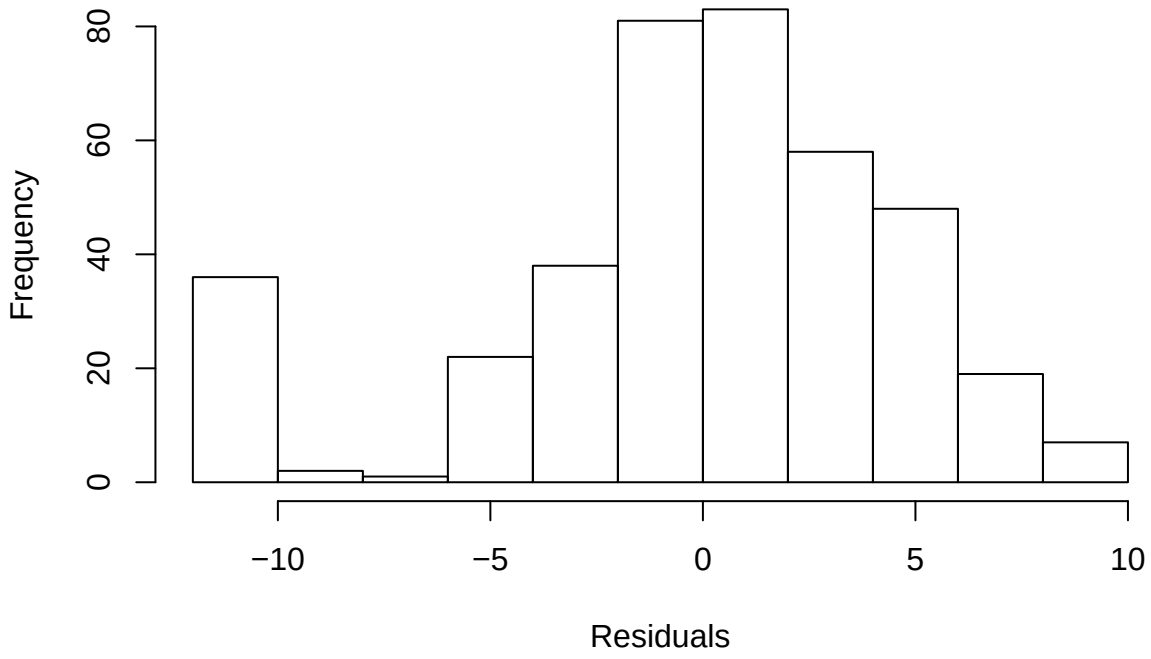
```
qqnorm(resid(student.lm), main="Figure 8: QQ plot of residuals from regression", ylab="residuals")
qqline(resid(student.lm),lty=1)
```

Figure 8: QQ plot of residuals from regression



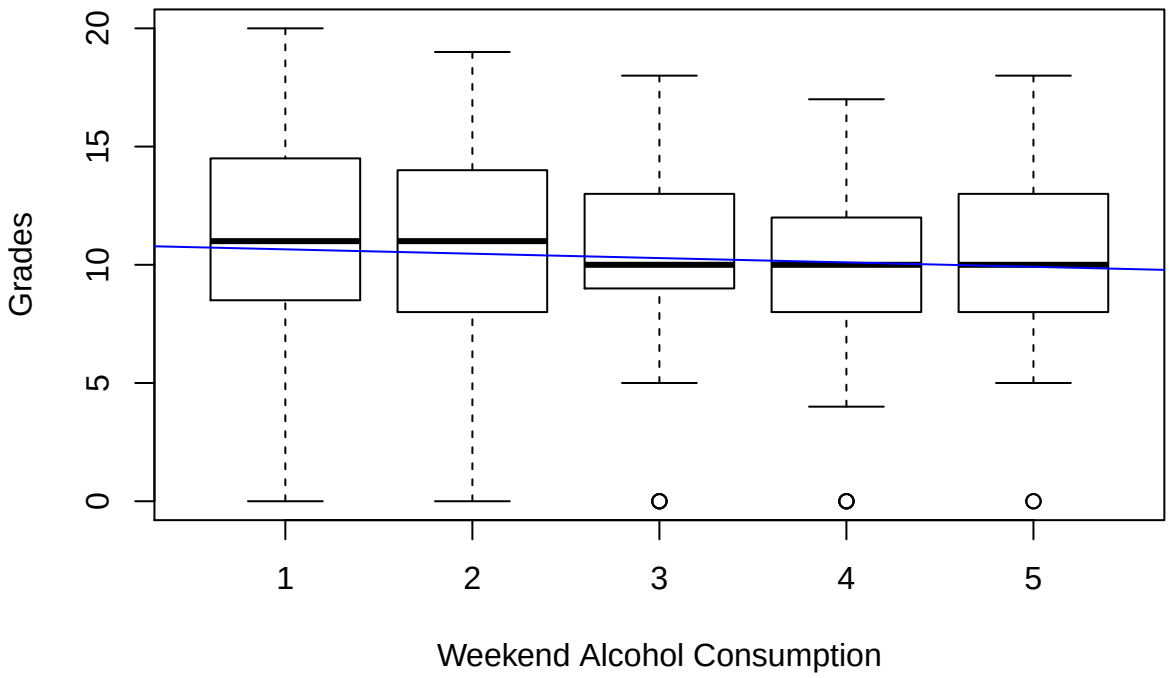

```
hist(resid(student.lm),main='Figure 9: Histogram of Residuals',xlab='Residuals')
```

Figure 9: Histogram of Residuals



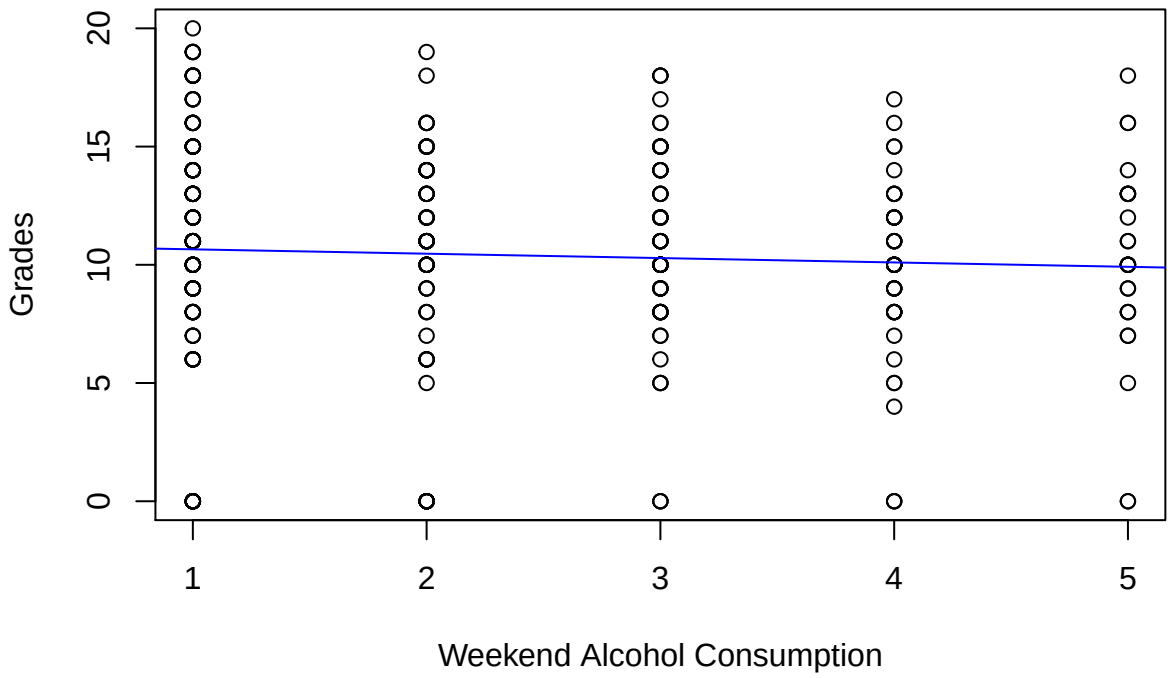

```
boxplot(student$G3~student$Walc,
    xlab='Weekend Alcohol Consumption',ylab='Grades',
     main='Figure 10: Grades versus Weekend Alcohol Consumption')
abline(student.lm,col=4)
```

Figure 10: Grades versus Weekend Alcohol Consumption




```
plot(student$G3~student$Walc,
     xlab='Weekend Alcohol Consumption',ylab='Grades',
     main='Figure 11: Grades versus Weekend Alcohol Consumption')
abline(student.lm,col=4)
```

Figure 11: Grades versus Weekend Alcohol Consumption




```
confint(student.lm)
```

```
##                  2.5 %     97.5 %
## (Intercept)  9.9128106 11.7642096
## Walc        -0.5370749  0.1675468
```

```
anova(student.lm)
```

```
## Analysis of Variance Table
##
## Response: G3
##            Df Sum Sq Mean Sq F value Pr(>F)
## Walc        1   22.3  22.310  1.0631 0.3032
## Residuals 393 8247.6  20.986
```