Statistics Exam

Please reply to the following questions in an R Markdown, called "surname_name.Rmd" and with title "Surname Name". Produce a pdf document and send both files (rmd and pdf) by mail to veronica.giro@quantide.com within Monday 25 July.

Before starting the exam, load the following packages:

```
require(qdata)
require(dplyr)
require(nortest)
```

Exercise 1

A chemist conducts an experiment to evaluate the efficacy of a solvent to dissolve stains of nail varnish from fabrics. He/She wants to test two types of solvent (`1` and `2`). The experiment consists of immersing 5 stained fabrics into a bowl with a solvent and of measuring the time (in minutes) necessary to dissolve the stain.

```
# Load data
data(varnish)
head(varnish)
```

```
## Source: local data frame [6 x 3]
##
##   Solvent Varnish  Time
##     (int)   (int) (dbl)
## 1       2       3 32.50
## 2       1       3 30.20
## 3       1       3 27.25
## 4       2       3 24.25
## 5       2       2 34.42
## 6       2       2 26.00
```

Consider the following variables:

- `Time` indicates time necessary to dissolve the stain (minutes)
- `Solvent` is a categorical variable with two levels and indicates the solvent type (`1` and `2`)

1. Test the normality of `Time` variable for solvent `1` and for solvent `2`. Comment the results. (Use the command: `tapply(X = varnish$Time, INDEX = varnish$Solvent, ad.test)`).

```
tapply(X = varnish$Time, INDEX = varnish$Solvent, ad.test)
```

```
## $`1`
##
##  Anderson-Darling normality test
##
## data:  X[[i]]
## A = 0.3154, p-value = 0.5082
##
```

```
##
## $`2`
##
##  Anderson-Darling normality test
##
## data:  X[[i]]
## A = 0.35138, p-value = 0.42
```

2. Check the hypothesys that the mean time necessary to dissolve nail varnish is the same for the two types of solvent and comment the results (use `t.test()` function).

```
t.test(varnish$Time ~ varnish$Solvent)
```

```
##
##  Welch Two Sample t-test
##
## data:  varnish$Time by varnish$Solvent
## t = -3.4039, df = 27.995, p-value = 0.002022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.772825 -1.435175
## sample estimates:
## mean in group 1 mean in group 2
##        25.99067        29.59467
```

Exercise 2

The headmaster of a high school is interested in how the number of awards earned this year by each student and the type of program in which he/she was enrolled influence the score obtained on the final math exam.

```
# Load data
data(awards)
head(awards)
```

```
## Source: local data frame [6 x 4]
##
##      id num_awards  prog  math
##   (int)      (int) (int) (int)
## 1    45          0     3    41
## 2   108          0     1    41
## 3    15          0     3    44
## 4    67          0     3    42
## 5   153          0     3    40
## 6    51          0     1    42
```

Consider the following variables:

- `math` represents students' scores on their math final exam
- `num_awards` indicates the number of awards earned by each student in a year
- `prog` is a categorical variable with three levels indicating the type of program in which the students were enrolled. It is coded as 1 = "General", 2 = "Academic" and 3 = "Vocational".

First of all, you have to convert `prog` variable as a factor:

```
awards <- awards %>% mutate(prog =as.factor(prog))
```

1. Fit a linear model to estimate the relation between `math` (as dependent variable) and the variables `prog` and `num_awards` (use `lm()` function). Compute the summary (use `summary.lm()` function) and comment the results. How the model coefficients have to be interpreted?

```
fm <- lm(math ~ prog + num_awards, data=awards)
summary.lm(fm)
```

```
##
## Call:
## lm(formula = math ~ prog + num_awards, data = awards)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7333  -5.6069  -0.3447   5.2676  22.6175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  49.3447     1.1362  43.430  < 2e-16 ***
## prog2         4.0009     1.4214   2.815  0.00538 **
## prog3        -3.7377     1.5589  -2.398  0.01744 *
```

```
## num_awards    3.3878     0.5499   6.161 4.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.586 on 196 degrees of freedom
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3443
## F-statistic: 35.83 on 3 and 196 DF,  p-value: < 2.2e-16
```

2. Compute model summary using `summary.aov()` function and comment the result. What is the difference between `summary.lm()` and `summary.aov()`?

```
summary.aov(fm)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## prog            2   4002  2001.1   34.77 1.19e-13 ***
## num_awards      1   2184  2184.3   37.96 4.03e-09 ***
## Residuals     196  11279    57.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

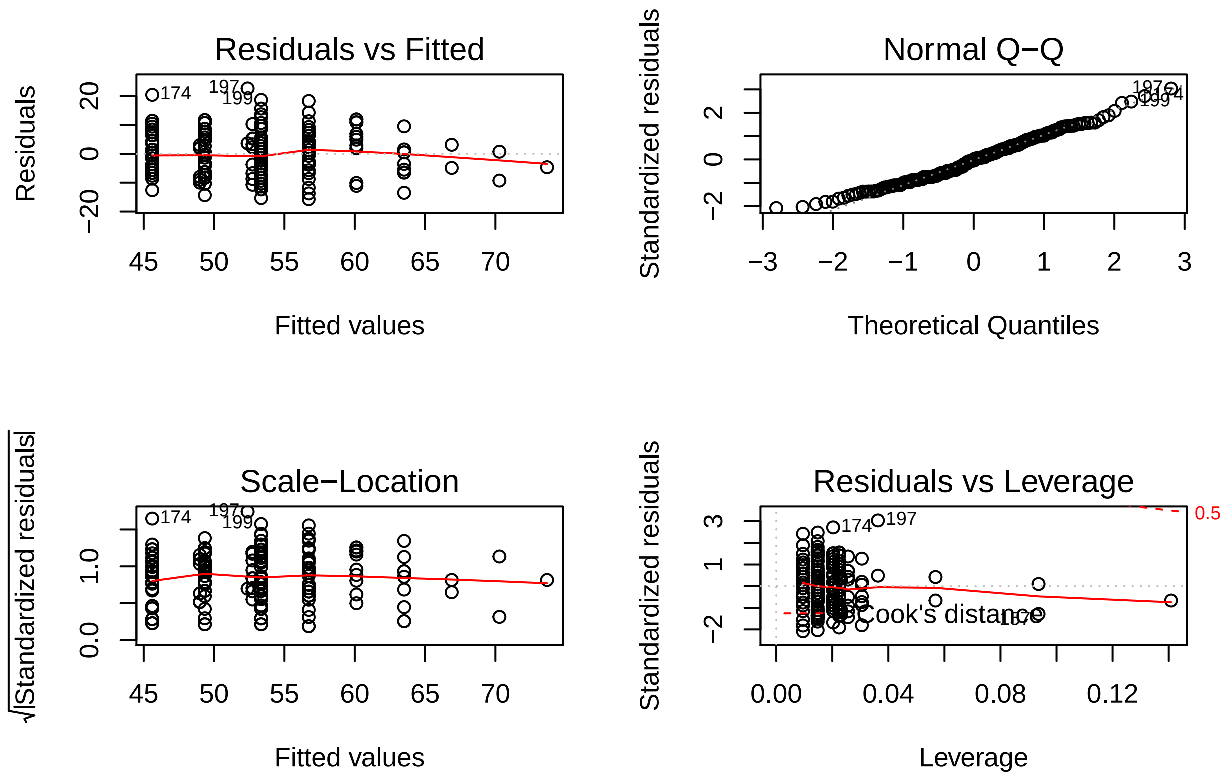
3. Fit the model removing the intercept from the model formula. Compute the summary (use `summary.lm()` function) and comment the results. How the model coefficients have to be interpreted? What is the difference between this model and that estimated at point 1.?

```
fm1 <- lm(math ~ prog + num_awards -1, data=awards)
summary.lm(fm1)
```

```
##
## Call:
## lm(formula = math ~ prog + num_awards - 1, data = awards)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7333  -5.6069  -0.3447   5.2676  22.6175
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## prog1       49.3447     1.1362  43.430  < 2e-16 ***
## prog2       53.3456     0.9222  57.846  < 2e-16 ***
## prog3       45.6069     1.0809  42.193  < 2e-16 ***
## num_awards   3.3878     0.5499   6.161 4.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.586 on 196 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9799
## F-statistic:  2435 on 4 and 196 DF,  p-value: < 2.2e-16
```

4. Perform the residual analysis of the model estimated and comment the results.

```r
op = par(mfrow=c(2,2))
plot(fm)
```




```r
par(op)
```

Exercise 3

A researcher is interested in how GRE (Graduate Record Exam scores) influences admission into graduate school.

```r
# Load data
data(admission)
head(admission)
```

```
## Source: local data frame [6 x 4]
##
##   admit   gre   gpa  rank
##   (int) (int) (dbl) (int)
## 1     0   380  3.61     3
## 2     1   660  3.67     3
## 3     1   800  4.00     1
## 4     1   640  3.19     4
## 5     0   520  2.93     4
## 6     1   760  3.00     2
```

Consider the following variables:

- `admit` is a binary variable (`0` (Not admitted) and `1` (Admitted)) and represents admission into graduate school
- `gre` represents Graduate Record Exam scores

1. Fit a logistic regression model between `admit` (as dependent variable) and `gre` (as independent variable) (use `glm()` function and specify the `family` parameter as `"binomial"`) and compute the summary of the fitted model. Comment the results, explaining the coefficients meaning.

```r
fm1 <- glm(admit ~ gre, data = admission, family = "binomial")
summary(fm1)
```

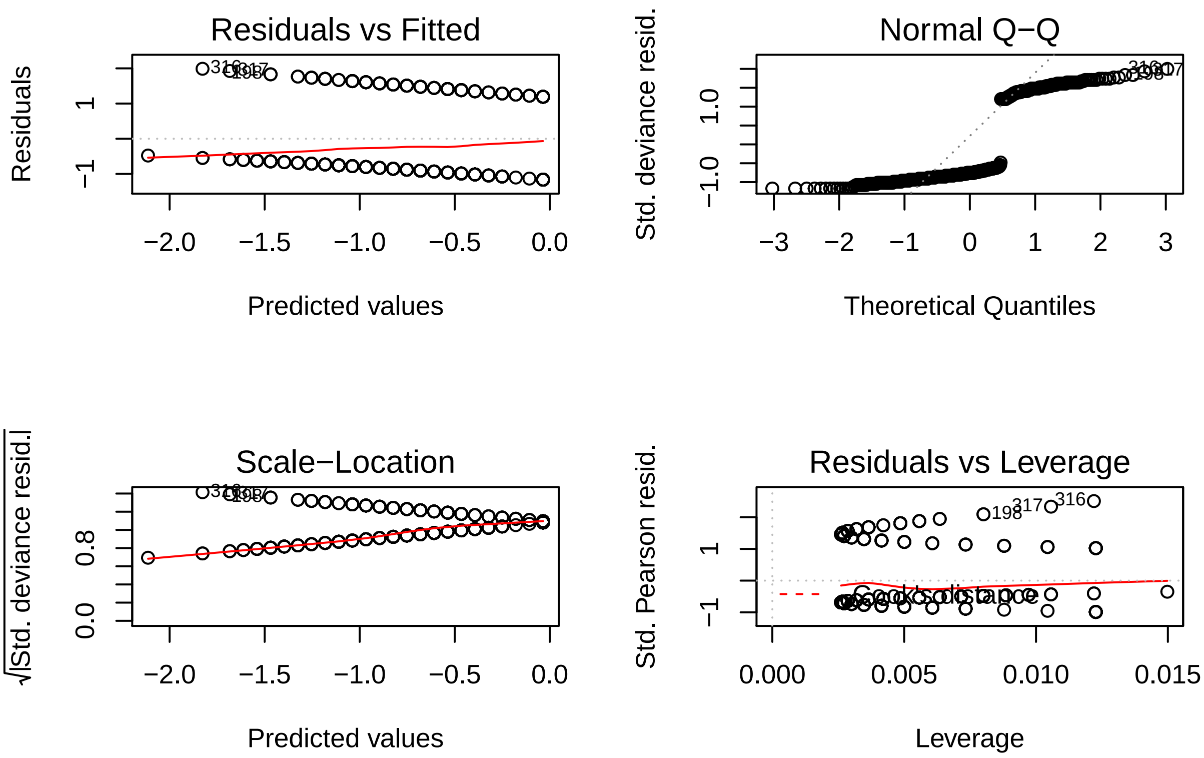
```
##
## Call:
## glm(formula = admit ~ gre, family = "binomial", data = admission)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1623  -0.9052  -0.7547   1.3486   1.9879
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.901344   0.606038  -4.787 1.69e-06 ***
## gre          0.003582   0.000986   3.633  0.00028 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 486.06  on 398  degrees of freedom
```

```
## AIC: 490.06
##
## Number of Fisher Scoring iterations: 4
```

```r
# For every one unit change in gre, the log odds of admission (versus non-admission) increases by 0.003
```

2. Perform the residual analysis of the model estimated and comment the results.

```r
op <- par(mfrow = c(2,2))
plot(fm1)
```




```r
par(op)
# The diagnostic graphs are not really nice, but similar configurations of points is not infrequent, wh
```