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 install the version 0.24 of `qdata` package:

```
install.packages(pkgs = "path-to-package/qdata_0.24.tar.gz", repos = NULL)
```

and 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)`).
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).

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?
2. Compute model summary using `summary.aov()` function and comment the result. What is the difference between `summary.lm()` and `summary.aov()`?
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.?
4. Perform the residual analysis of the model estimated and comment the results.

Exercise 3

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

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