Magister degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 1 february 2017

Theory

1) Describe what are the aims of Exploratory Data Analysis and present the main numerical summaries for bivariate data.

2) Define the multiple linear regression model and highlight the basic assumptions. Describe the least squares estimators for the regression parameters and define a suitable estimator for the variance parameter. Discuss the usefulness of the fitted regression model for inferential and prediction purposes. Define the confidence intervals for both the regression parameters and the regression line and specify the prediction interval for a future response variable.

Laboratory

3) Consider the `R` commands below, describe what the two codes are intended to do and explain what is being calculated on each line. Here, simulated samples are generated from an exponential distribution with `rate=1/5`.

```
# code no.1
N <- 10000
set.seed(10)
samp <- rexp(N,1/5)
mean(samp)
var(samp)
sd(samp)
# code no.2
set.seed(10)
repl<-10000
n <- 10
sampvar <- NULL
variance <- NULL
for (i in 1:repl){
  sam <- rexp(n,1/5)
  sampvar <- c(sampvar,var(sam))
  variance <- c(variance,var(sam)*9/10)}
mean(sampvar)
```

```
mean(variance)
sd(sampvar)
sd(variance)
hist(sampvar)
hist(variance)
```

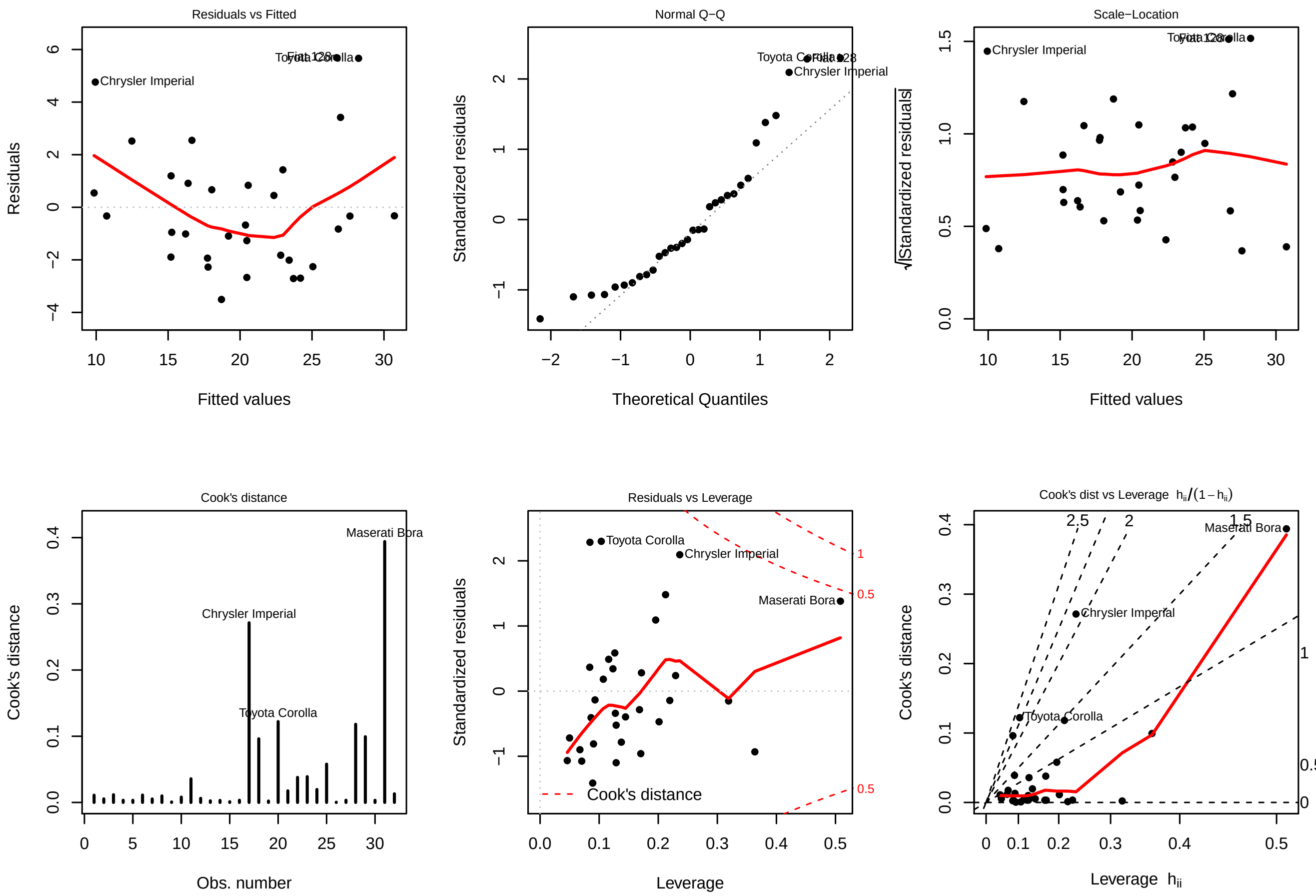
4) Let us consider the dataframe `mtcars`, which comprises the fuel consumption and 10 aspects of design and performance for 32 automobiles (1970s models). The help file is given below

```
A data frame with 32 observations on 11 variables.

[, 1]  mpg  Miles/(US) gallon
[, 2]  cyl  Number of cylinders
[, 3]  disp  Displacement (cu.in.)
[, 4]  hp  Gross horsepower
[, 5]  drat  Rear axle ratio
[, 6]  wt  Weight (1000 lbs)
[, 7]  qsec  1/4 mile time
[, 8]  vs  V/S
[, 9]  am  Transmission (0 = automatic, 1 = manual)
[,10]  gear  Number of forward gears
[,11]  carb  Number of carburetors
```

Describe how to perform a preliminary data analysis on this dataframe, using suitable `R` commands. After fitting the model `fit <- lm(mpg ∼ disp + hp + wt + drat, data=mtcars)`, the following outputs are obtained by the `R` commands `summary(fit)` and `plot(fit)`, respectively.

```
Call:
lm(formula = mpg ~ disp + hp + wt + drat, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5077 -1.9052 -0.5057  0.9821  5.6883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.148738   6.293588   4.631  8.2e-05 ***
disp         0.003815   0.010805   0.353  0.72675
hp          -0.034784   0.011597  -2.999  0.00576 **
wt          -3.479668   1.078371  -3.227  0.00327 **
drat         1.768049   1.319779   1.340  0.19153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.602 on 27 degrees of freedom
Multiple R-squared:  0.8376,Adjusted R-squared:  0.8136
F-statistic: 34.82 on 4 and 27 DF,  p-value: 2.704e-10
```



Describe how to interpret these results, and then suggest how to proceed with further analyses.

Magister degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 13 February 2017

Theory

1) Describe the purpose of an interval estimation procedure. Give the right statistical interpretation of an observed 95% confidence interval for an interest parameter. Present a simple application regarding the estimation of a population mean.

2) Define the multiple linear regression model and highlight the basic assumptions. List some useful steps in the model fitting procedure. Finally, recall the main statistical indices and procedures for model assessment and model selection.

Laboratory

3) Consider the `R` commands below, describe what the two codes are intended to do and explain what is being calculated on each line. Here, the well-known dataset `Advertising` is taken into account.

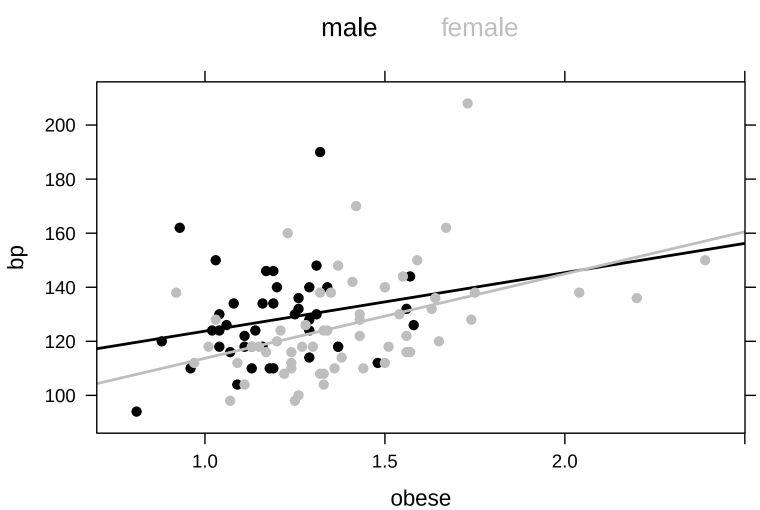
```
mod.adv <- lm(Sales~TV+Radio+Newspaper, Advertising)
summary(mod.adv)
summary(mod.adv)$sigma^2
AIC(mod.adv)
par(mfrow=c(2,2), pty="s", mar=c(3,2,3,2))
plot(mod.adv)
par(mfrow=c(1,1))
#
mod.adv1 <- lm(Sales~TV+Radio+I(TV^2)+TV:Radio, Advertising)
summary(mod.adv1)
summary(mod.adv1)$sigma^2
AIC(mod.adv1)
par(mfrow=c(2,2), pty="s", mar=c(3,2,3,2))
plot(mod.adv1)
par(mfrow=c(1,1))
intc <- predict(mod.adv1, newdata=data.frame(TV=100,Radio=20),
                interval="confidence")
intp <- predict(mod.adv1, newdata=data.frame(TV=100,Radio=20),
                interval="prediction")
```

4) Let us consider the dataframe `bp.obese` of the library `ISwR`, which comprises information about sex, obesity and blood pressure for a random sample of 102 Mexican-American adults in a small California town. The help file and the output of the `str` command are given below

```
This data frame contains the following columns:
sex
a numeric vector code, 0: male, 1: female.
obese
a numeric vector, ratio of actual weight to ideal weight from New York
Metropolitan Life Tables.
bp
a numeric vector, systolic blood pressure (mm Hg).
```

```
'data.frame':	102 obs. of  3 variables:
 $ sex  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ obese: num  1.31 1.31 1.19 1.11 1.34 1.17 1.56 1.18 1.04 1.03 ...
 $ bp   : int  130 148 146 122 140 146 132 110 124 150 ...
```

The aim of the study is to analyze the potential relationship between blood pressure, which is the response variable, and obesity, taking into account also the factor regressor sex. Describe how to perform a preliminary data analysis on this dataframe, using suitable `R` commands and comment the following plot.



After fitting these linear models `fit1 <- lm(bp ~ obese,data=bp.obese)`, `fit2 <- lm(bp ~ obese+sex,data=bp.obese)` and `fit3 <- lm(bp ~ obese*sex,data=bp.obese)`, the following outputs are obtained by the `R` function `summary`.

```
Call:
lm(formula = bp ~ obese, data = bp.obese)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-27.570 -11.241  -2.400   9.116  71.390 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   96.818      8.920   10.86  < 2e-16 ***
obese         23.001      6.667    3.45 0.000822 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.28 on 100 degrees of freedom
Multiple R-squared:  0.1064,Adjusted R-squared:  0.09743 
F-statistic:  11.9 on 1 and 100 DF,  p-value: 0.0008222
```

```
Call:
lm(formula = bp ~ obese + sex, data = bp.obese)

Residuals:
    Min      1Q  Median      3Q     Max 
-24.263 -11.613  -2.057   6.424  72.207 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   93.287      8.937  10.438  < 2e-16 ***
obese         29.038      7.172   4.049 0.000102 ***
sex           -7.730      3.715  -2.081 0.040053 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17 on 99 degrees of freedom
Multiple R-squared:  0.1438,Adjusted R-squared:  0.1265 
F-statistic: 8.314 on 2 and 99 DF,  p-value: 0.0004596
```

```
Call:
lm(formula = bp ~ obese * sex, data = bp.obese)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.645 -11.621  -1.708   6.737  71.500 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  102.112     18.231   5.601 1.95e-07 ***
obese         21.646     15.118   1.432    0.155
sex          -19.596     21.664  -0.905    0.368
obese:sex      9.558     17.191   0.556    0.579
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.05 on 98 degrees of freedom
Multiple R-squared:  0.1465,Adjusted R-squared:  0.1204
F-statistic: 5.607 on 3 and 98 DF,  p-value: 0.001368
```

Describe how to interpret these results, and then suggest how to proceed with further analyses.

Magister degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 15 February 2018

Theory

1) Describe the purpose of a point estimation procedure. List the main property of an estimator and define the standard error. Present a simple application regarding the estimation of a proportion.

2) Define the one-way and the two-way analysis of variance models and highlight the basic assumptions. Describe the statistical tests on the main effects and on the interaction effect of the factors on the mean response.

Laboratory

3) Consider the R commands below, describe what the code is intended to do and explain what is being calculated on each line. Finally, describe the R functions `dbinom`, `pbinom`, `qbinom` and `rbinom`.

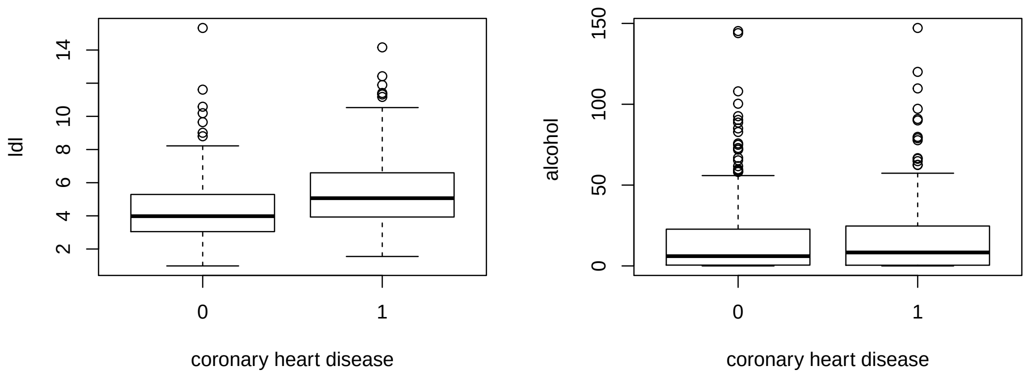
```
par(mfrow=c(2,2))
xx<-seq(0,10,1)
plot(xx,dbinom(xx,10,0.2),pch=19,ylim=c(0,0.5),
     cex.axis=1.5,xlab=" ",ylab=" ",main="A) n=10, p=0.2") # Step 3
segments(0,0,10,0,lwd=2)
#
plot(xx,dbinom(xx,10,0.5),pch=19,ylim=c(0,0.5),lwd=2,
     cex.axis=1.5,xlab=" ",ylab=" ",main="B) n=10, p=0.5") # Step 3
segments(0,0,10,0,lwd=2)
#
plot(xx,dbinom(xx,10,0.8),pch=19,ylim=c(0,0.5),lwd=2,
     cex.axis=1.5,xlab=" ",ylab=" ",main="C) n=10, p=0.8") # Step 3
segments(0,0,10,0,lwd=2)
#
xx<-seq(0,20,1)
plot(xx,dbinom(xx,20,0.5),pch=19,ylim=c(0,0.5),lwd=2,
     cex.axis=1.5,xlab=" ",ylab=" ",main="D) n=20, p=0.5") # Step 3
segments(0,0,20,0,lwd=2)
par(mfrow=c(1,1))
```

4) Let us consider the dataframe `SAheart` of the library `ElemStatLearn`, which comprises information about a retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. The help file and the output of the `str` command are given below

```
A data frame with 462 observations on the following 10 variables.

sbp
  systolic blood pressure
tobacco
  cumulative tobacco (kg)
ldl
  low density lipoprotein cholesterol
adiposity
  a numeric vector
famhist
  family history of heart disease, a factor with levels Absent Present
typea
  type-A behavior
obesity
  a numeric vector
alcohol
  current alcohol consumption
age
  age at onset
chd
  response, coronary heart disease
```

```
'data.frame': 462 obs. of  10 variables:
 $ sbp      : int  160 144 118 170 134 132 142 114 114 132 ...
 $ tobacco  : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
 $ ldl      : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
 $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
 $ famhist  : Factor w/ 2 levels "Absent","Present": 2 1 2 2 2 2 1 2 2 2 ...
 $ typea    : int  49 55 52 51 60 62 59 62 49 69 ...
 $ obesity  : num  25.3 28.9 29.1 32 26 ...
 $ alcohol  : num  97.2 2.06 3.81 24.26 57.34 ...
 $ age      : int  52 63 46 58 49 45 38 58 29 53 ...
 $ chd      : int  1 1 0 1 1 0 0 1 0 1 ...
```

The aim of the study is to analyze the potential relationship between the binary response variable `chd` and the explanatory variables considered in the dataframe. Describe how to perform a preliminary data analysis on this dataframe, using suitable `R` commands, and comment the following plot.



With the command `mod0 <- glm(chd ~ ldl, data = SAheart, family = binomial)`, a simple logistic regression model is defined for describing the potential effect of the level of `ldl` on the probability of coronary heart disease. Comment the model fitting outcomes given by the function `summary`.

```
Call:
glm(formula = chd ~ ldl, family = binomial, data = SAheart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1647  -0.8948  -0.7426   1.2688   1.8637

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.96867    0.27308  -7.209 5.63e-13 ***
ldl          0.27466    0.05164   5.319 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 564.28  on 460  degrees of freedom
AIC: 568.28

Number of Fisher Scoring iterations: 4
```

After fitting these two further logistic regression models `mod1 <- glm(chd ~ ., data = SAheart, family = binomial)` and `mod2 <- glm(chd ~ tobacco + ldl + famhist + typea +  age + ldl:famhist, data = SAheart, family = binomial)`, the following outputs are obtained by the `R` function `summary`.

```
Call:
glm(formula = chd ~ ., family = binomial, data = SAheart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7781  -0.8213  -0.4387   0.8889   2.5435

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
sbp             0.0065040  0.0057304   1.135 0.256374
tobacco         0.0793764  0.0266028   2.984 0.002847 **
ldl             0.1739239  0.0596617   2.915 0.003555 **
adiposity       0.0185866  0.0292894   0.635 0.525700
famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
typea           0.0395950  0.0123202   3.214 0.001310 **
obesity        -0.0629099  0.0442477  -1.422 0.155095
alcohol         0.0001217  0.0044832   0.027 0.978350
age             0.0452253  0.0121298   3.728 0.000193 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 472.14  on 452  degrees of freedom
AIC: 492.14

Number of Fisher Scoring iterations: 5

Call:
glm(formula = chd ~ tobacco + ldl + famhist + typea + age + ldl:famhist,
    family = binomial, data = SAheart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8463  -0.7938  -0.4419   0.9161   2.4956

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -5.79224    0.94625  -6.121 9.28e-10 ***
tobacco             0.08496    0.02628   3.233  0.00122 **
ldl                 0.01758    0.07302   0.241  0.80974
```

```
famhistPresent     -0.77068    0.62341  -1.236  0.21637    
typea               0.03690    0.01240   2.974  0.00294 ** 
age                 0.05140    0.01030   4.990 6.03e-07 ***
ldl:famhistPresent  0.33334    0.11595   2.875  0.00404 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 596.11  on 461  degrees of freedom
Residual deviance: 466.90  on 455  degrees of freedom
AIC: 480.9

Number of Fisher Scoring iterations: 5
```

Describe how to interpret these results, and then suggest how to proceed with further analyses.

Magister degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 11 June 2018

Theory

1) Describe the purpose of a point estimation procedure. List the main property of an estimator and define the standard error. Present a simple application regarding the estimation of a population variance.

2) Introduce and discuss the topic of regression models with non-Gaussian response variables. Consider the case of a Bernoulli distributed response and define the logistic regression model. With regard to a fitted logistic regression model, emphasize the interpretation of the estimated regression parameter and discuss its potential application for predicting a future binary response.

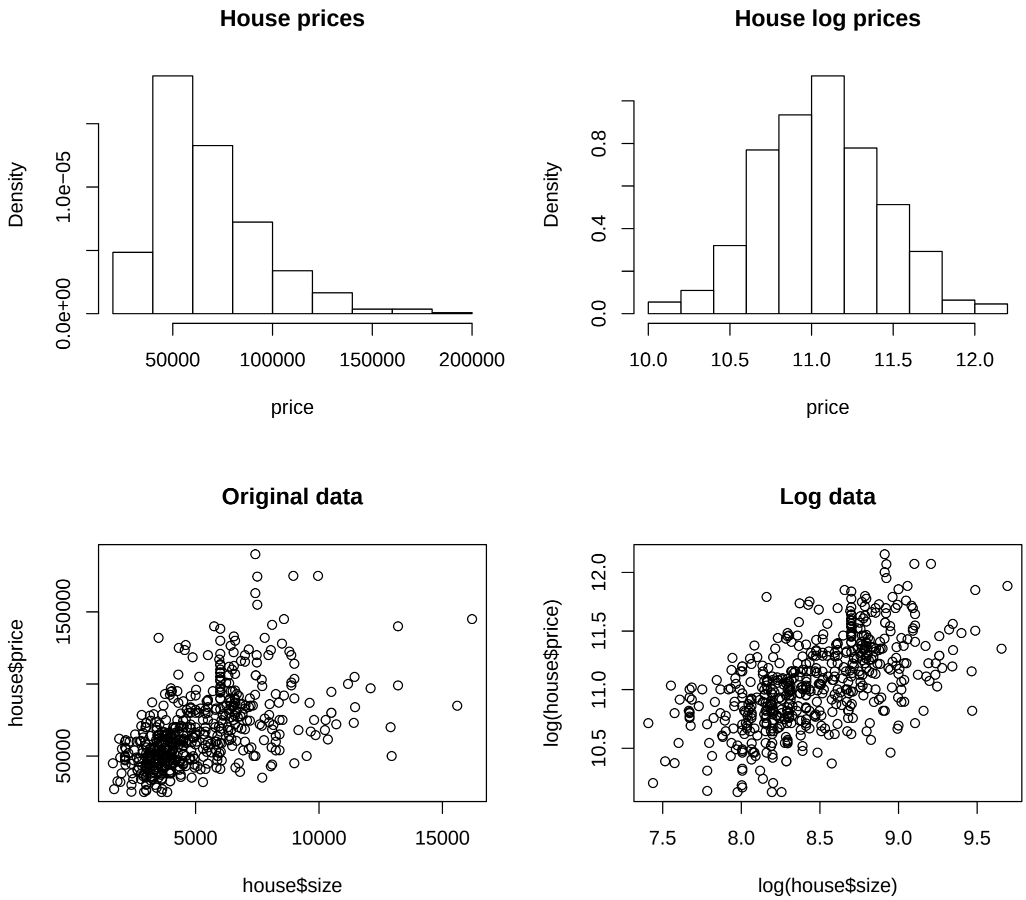
Laboratory

3) Write an `R` code to analyze the behavior of the sampling distribution of the sample variance, as the sample size increases. Consider 1000 simulated random samples of dimension 25, 50, 100 from a normal distribution with `mean=1` and `sd=1`.

4) Let us consider the dataframe `house`, which includes information about the `price`, the `size`, the `floor`, the number of bedrooms (`bed`) and the number of bathrooms (`bath`) of 546 houses. The output of the `str` command is given below

```
'data.frame': 546 obs. of  5 variables:
 $ price: num  42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...
 $ size : int  5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...
 $ bed  : int  3 2 3 3 2 3 3 3 3 3 ...
 $ bath : int  1 1 1 1 1 1 2 1 1 2 ...
 $ floor: int  2 1 1 2 1 1 2 3 1 4 ...
```

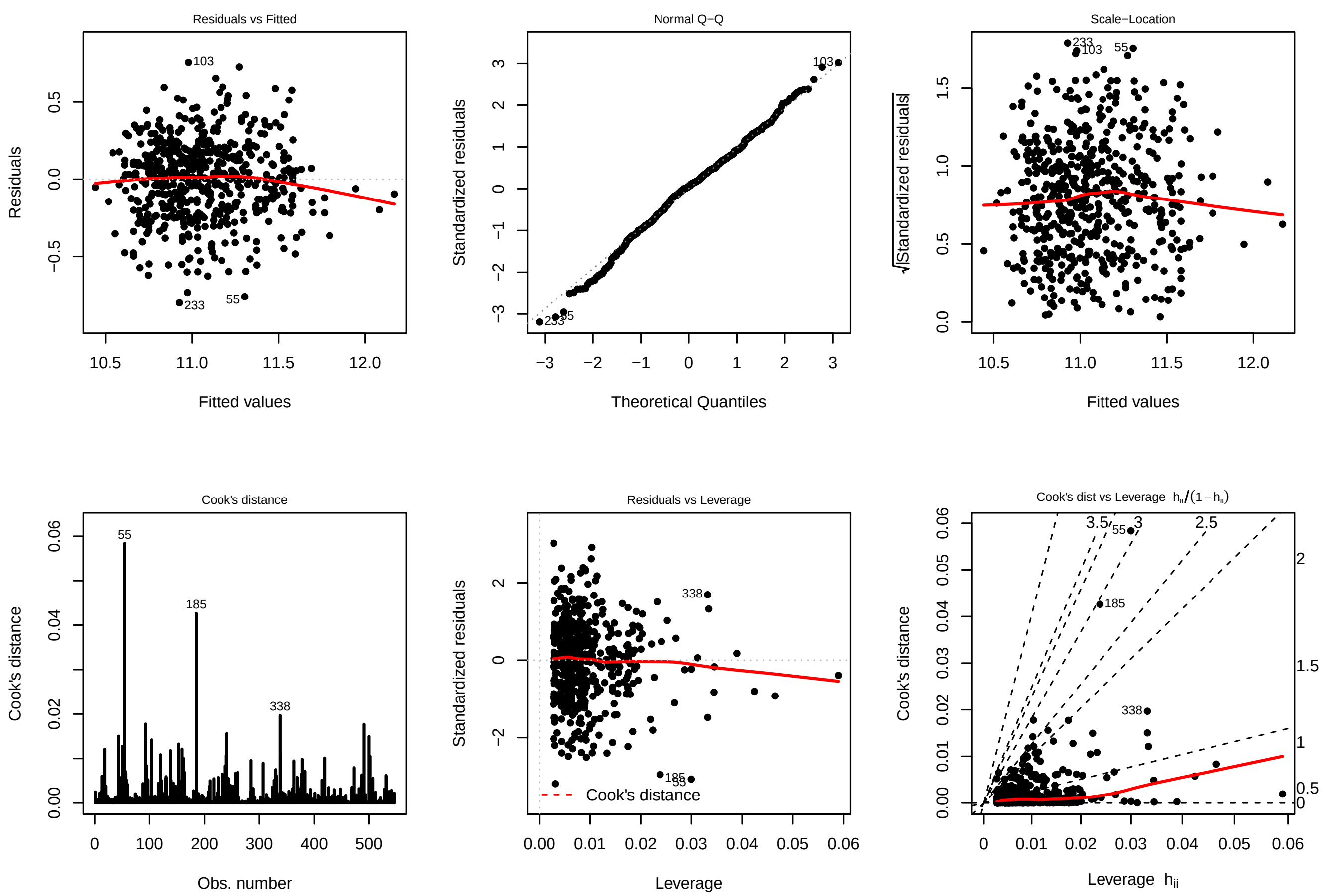
A suitable linear regression model can be defined in order to study the potential relationship between the `price`, which is the response variable, and the explanatory variables considered in the dataframe. Describe how to perform a preliminary data analysis on this dataframe, using suitable `R` commands.
Moreover, consider the following plots and discuss the possibility of measuring the variables `price` and `size` in the logarithmic scale.



After fitting the regression model `fit <- lm(log(price) ~ log(size) + bed + bath + floor, data=house)`, the following outputs are obtained by the `R` commands `summary(fit)` and `plot(fit)`, respectively.

```
Call:
lm(formula = log(price) ~ log(size) + bed + bath + floor, data = house)

Residuals:
     Min       1Q   Median       3Q      Max
-0.80006 -0.16043  0.01391  0.16359  0.75723

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.63539    0.23052  28.785  < 2e-16 ***
log(size)    0.45274    0.02770  16.344  < 2e-16 ***
bed          0.04997    0.01668   2.995  0.00287 **
bath         0.20265    0.02386   8.493  < 2e-16 ***
floor        0.10052    0.01386   7.253 1.42e-12 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2511 on 541 degrees of freedom
Multiple R-squared:  0.5477,Adjusted R-squared:  0.5444
F-statistic: 163.8 on 4 and 541 DF,  p-value: < 2.2e-16
```



Describe how to interpret these results, and then suggest how to proceed with further analyses with particular regard to prediction.

Magister degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 4 February 2019

Theory

1) Define the Gaussian distribution and describe its usefulness in statistical applications..

2) Define the multiple linear regression model and highlight the basic assumptions. Discuss the case in which the explanatory variables are factor, with particular regard to the codification using dummy variables. Finally, consider the situation with both factors and numerical explanatory variables, focusing on the particular case of models admitting different simple regression lines.

Laboratory

3) Consider the R commands below, describe what the three codes are intended to do and explain what is being calculated on each line. Here, the well-known dataset USArrest is taken into account and a Principal Component Analysis procedure is applied.

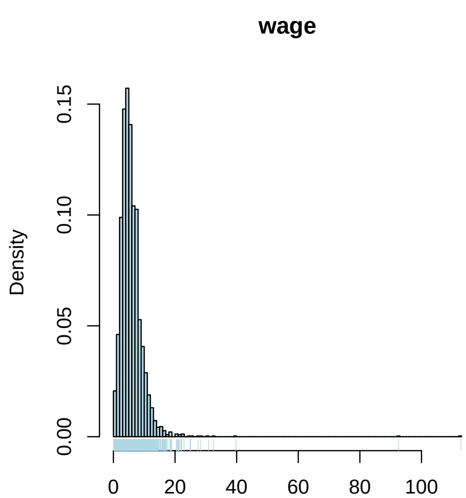
```
# code no.1
obj <- princomp(USArrests, cor=TRUE)
z1 <- -obj$scores[,1]
z2 <- -obj$scores[,2]
phi1<--obj$loadings[,1]
phi2<--obj$loadings[,2]
# code no.2
obj$loadings<--obj$loadings
obj$scores<--obj$scores
biplot(obj, xlab="1st principal component", ylab="2nd principal component",
       xlim=c(-3.5,3.5), col=c(1,2), scale=0)
# code no.3
par(mfrow=c(1,2), pty="s")
plot(obj$sdev^2/4, xlab="Principal component", ylab="PVE", type='b')
plot(cumsum(obj$sdev^2)/4, xlab="Principal component", ylab="Cumulative PVE",
     ylim=c(0,1), type='b')
par(mfrow=c(1,1))
```

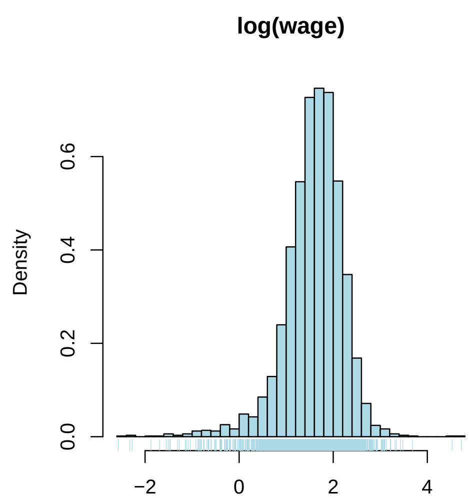
4) Let us consider the dataframe wages, which containes information about 3294 USA working individuals. The data are taken from the National Longitudinal Survey and are related to 1987. The variable as are listed below and the output of the str command is given

```
A data frame with 3294 observations on the following 4 variables.

exper
  experience in years
male
  1 male, 0 female
school
  years of schooling
wage
  wage (in 1980 $) per hour
region
  Center, North, South
```

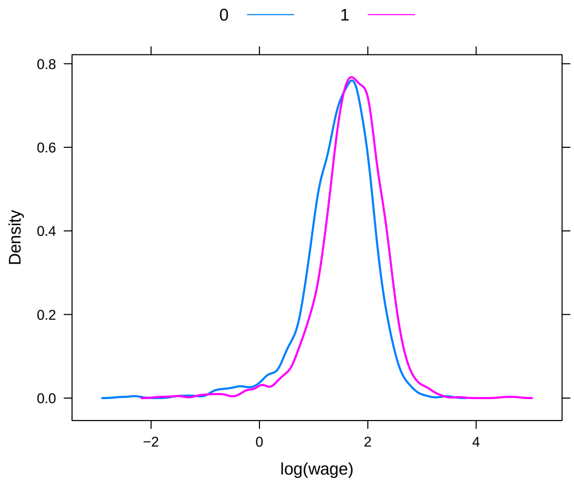
```
'data.frame': 3296 obs. of  5 variables:
 $ exper : int  9 12 11 9 8 9 8 10 12 7 ...
 $ male  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ school: int  13 12 11 14 14 14 12 12 10 12 ...
 $ wage  : num  6.32 5.48 3.64 4.59 2.42 ...
 $ region: Factor w/ 3 levels "Center","North",..: 1 1 3 1 1 1 1 1 1 3 ...
```

The aim of the study is to analyze the potential relationship between the response variable `wage` and the explanatory variables considered in the dataframe. Describe how to perform a preliminary data analysis on this dataframe, using suitable `R` commands. Moreover, consider the following plots and discuss the possibility of measuring the variable `wage` in the logarithmic scale





In order to describe the effect of the factor `male` on the response `log(wage)` we may analize this plot, where the probability distribution of `log(wage)` is represented by considering the kernel density estimates conditioned on the two levels (`1 male`, `0 female`) of the variable `male`



With the commands `mod.0<-lm(log(wage) ~ male,data=wages)` and `mod.1<-lm(log(wage) ~ exper*male, data=wages)`, two regression models are defined for describing the potential effect of `male` and `exper` on the response `log(wage)`. Comment the model fitting outcomes given by the function `summary` (Hint: consider the fact that the average years of experience in the sample is lower for women than for men).

```
Call:
lm(formula = log(wage) ~ male, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0445 -0.3073  0.0544  0.3839  3.0325

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.47475    0.01559   94.59   <2e-16 ***
male1        0.21826    0.02154   10.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6176 on 3294 degrees of freedom
Multiple R-squared:  0.03023,Adjusted R-squared:  0.02994
F-statistic: 102.7 on 1 and 3294 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log(wage) ~ exper * male, data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-4.0906 -0.3050  0.0560  0.3792  3.0468

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.193551   0.057470  20.768  < 2e-16 ***
exper        0.036367   0.007156   5.082 3.94e-07 ***
male1        0.463707   0.079062   5.865 4.93e-09 ***
exper:male1 -0.032071   0.009518  -3.369 0.000762 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6153 on 3292 degrees of freedom
Multiple R-squared:  0.03792,Adjusted R-squared:  0.03704
F-statistic: 43.25 on 3 and 3292 DF,  p-value: < 2.2e-16
```

Finally, a complete regression model is fitted `mod.2<-lm(log(wage) ~., data=wages)` and the following output is obtained by the `R` function `summary`.

```
Call:
lm(formula = log(wage) ~ ., data = wages)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0008 -0.2821  0.0468  0.3673  3.2337

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.279709   0.090321  -3.097  0.00197 **
exper        0.034618   0.004549   7.610 3.55e-14 ***
male1        0.246474   0.020607  11.961  < 2e-16 ***
school       0.122909   0.006278  19.578  < 2e-16 ***
regionNorth  0.051107   0.024505   2.086  0.03709 *
regionSouth  0.047168   0.024969   1.889  0.05898 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5831 on 3290 degrees of freedom
Multiple R-squared:  0.1364,Adjusted R-squared:  0.1351
F-statistic: 103.9 on 5 and 3290 DF,  p-value: < 2.2e-16
```

Describe how to interpret these results, and then suggest how to proceed with further analyses.

Master degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 21 February 2019

Theory

1) Describe the purpose of a point estimation procedure. List the main property of an estimator and define the standard error. Present a simple application regarding the estimation of the difference of the means of two independent populations.

2) Define the one-way and the two-way analysis of variance models and highlight the basic assumptions. Describe the statistical tests on the main effects and on the interaction effect of the factors on the mean response.

Laboratory

3) Describe the `R` functions that can be used for model selection. Furthermore, consider the `R` commands below, describe what the code is intended to do and explain what is being calculated on each line. Here, dataset `trees`, which provides some measurements on felled black cherry trees, is taken into account.

```
cv1 <- 0
cv2 <- 0
n <- length(trees$Volume)
i <-1
for (i in 1:n){
mod1i <- lm(Volume ~ Girth, data = trees[-i,])
mod2i <- lm(Volume ~ Girth + Height, data = trees[-i,])
mu1 <- mod1i$coefficients[1] + mod1i$coefficients[2]*trees$Girth[i]
mu2 <- mod2i$coefficients[1] + mod2i$coefficients[2]*trees$Girth[i] +
            mod2i$coefficients[3]*trees$Height[i]
sd1 <- sqrt(sum(mod1i$residuals^2)/(n-3))
sd2 <- sqrt(sum(mod2i$residuals^2)/(n-4))
cv1 <- cv1 - log(dnorm(trees$Volume[i],mu1,sd1))
cv2 <- cv2 - log(dnorm(trees$Volume[i],mu2,sd2))}
cv1
cv2
```

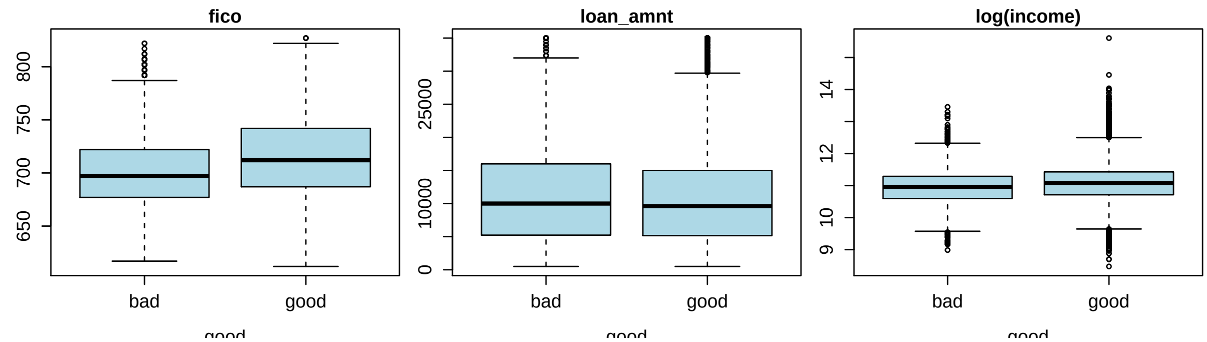
4) Let us consider the data frame `loan`, which contains information about 42,535 loans ranging from 1,000 $ to 35,000 $, issued by a company called Lending Club. The following variables are considered: `good` (the behaviour of the client with values `good` and `bad`), `fico` (the FICO credit score measuring the client credit worthiness), `purpose` (the intended use of the loan, with 8 different categories), `loan_amt` (the credit amount in $) and `income` (the annual income in $ of the client). The variable as are listed below and the output of the `str` command is given

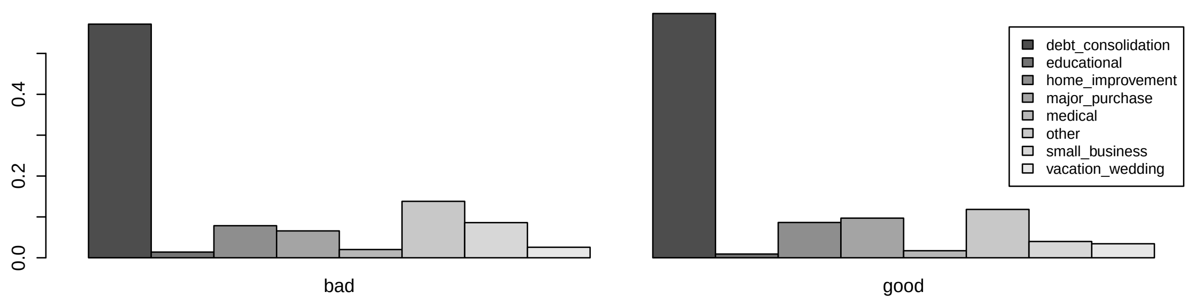
```
'data.frame': 42535 obs. of  5 variables:
 $ good     : Factor w/ 2 levels "bad","good": 2 1 2 2 2 2 2 2 1 1 ...
 $ purpose  : Factor w/ 8 levels "debt_consolidation",..: 1 4 7 6 6 8 1 4 7 6 ...
 $ fico     : int  737 742 737 692 697 732 692 662 677 727 ...
 $ loan_amnt: int  5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...
 $ income   : num  24000 30000 NA 49200 80000 36000 NA 48000 40000 15000 ...
```

Moreover, the output of the command `summary` is also given

```
   good                      purpose             fico      
 bad : 6371   debt_consolidation:25253   Min.   :612.0  
 good:36164   other             : 5160   1st Qu.:687.0  
              major_purchase    : 3926   Median :712.0  
              home_improvement  : 3625   Mean   :715.1  
              small_business    : 1992   3rd Qu.:742.0  
              vacation_wedding  : 1404   Max.   :827.0  
              (Other)           : 1175                  
   loan_amnt         income       
 Min.   :  500   Min.   :   4800  
 1st Qu.: 5200   1st Qu.:  44995  
 Median : 9700   Median :  63000  
 Mean   :11090   Mean   :  75186  
 3rd Qu.:15000   3rd Qu.:  90000  
 Max.   :35000   Max.   :6000000  
                 NA's   :18758    
```

The aim of the study is to analyze the potential relationship between the response variable `good` and the explanatory variables considered in the data frame, in order to evaluate the possible good/bad behaviour of a customer. Describe how to perform a preliminary data analysis on this data frame, using suitable `R` commands. Moreover, consider and discuss the following plots





In order to describe the effect of the factor `fico` on the response `good` we consider a simple logistic regression model fitted using the command `mod.1<-glm(good ∼ fico, data = loan, family = "binomial")`. Comment the model fitting outcomes given by the function `summary` and the output given by the subsequent commands.

```
Call:
glm(formula = good ~ fico, family = "binomial", data = loan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5172   0.4078   0.5306   0.6294   0.9622

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.033068   0.296922  -23.69   <2e-16 ***
fico         0.012358   0.000421   29.35   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35928  on 42534  degrees of freedom
Residual deviance: 34985  on 42533  degrees of freedom
AIC: 34989

Number of Fisher Scoring iterations: 5
```

```
exp(coef(mod1))

 (Intercept)         fico
0.0008822209 1.0124345145
```

```
test <- data.frame(fico=c(700,750))
test$pred <- predict(mod1,test, type="response")
test

  fico      pred
1  700 0.8344391
2  750 0.9033761
```

Two further logistic regression models are fitted using `mod.2<-glm(good ~ fico + loan_amnt, data = loan, family = "binomial")` and `mod.3<-glm(good ~ fico + loan_amnt + income + purpose, data = loan, family = "binomial")`. Comment the corresponding output obtained by the `R` function `summary` and then suggest how to proceed with a further predictive analysis.

```
Call:
glm(formula = good ~ fico + loan_amnt, family = "binomial", data = loan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6261   0.4011   0.5256   0.6261   0.9423

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.367e+00  3.007e-01  -24.50   <2e-16 ***
fico         1.319e-02  4.306e-04   30.62   <2e-16 ***
loan_amnt   -2.229e-05  1.815e-06  -12.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35928  on 42534  degrees of freedom
Residual deviance: 34838  on 42532  degrees of freedom
AIC: 34844

Number of Fisher Scoring iterations: 5
```

```
Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred


Call:
glm(formula = good ~ fico + loan_amnt + income + purpose, family = "binomial",
    data = loan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3659   0.3840   0.5224   0.6320   1.2589

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -7.482e+00  4.119e-01 -18.165  < 2e-16 ***
fico                     1.303e-02  5.906e-04  22.061  < 2e-16 ***
loan_amnt               -3.663e-05  2.580e-06 -14.200  < 2e-16 ***
income                   7.203e-06  5.426e-07  13.275  < 2e-16 ***
purposeeducational      -5.076e-01  2.309e-01  -2.198   0.0279 *
purposehome_improvement -1.077e-01  7.108e-02  -1.515   0.1298
purposemajor_purchase    1.388e-02  7.689e-02   0.180   0.8568
purposemedical          -2.678e-01  1.426e-01  -1.878   0.0604 .
purposeother            -2.988e-01  6.034e-02  -4.952 7.34e-07 ***
purposesmall_business   -9.158e-01  7.016e-02 -13.052  < 2e-16 ***
purposevacation_wedding  1.114e-01  1.126e-01   0.989   0.3226
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20592  on 23776  degrees of freedom
Residual deviance: 19677  on 23766  degrees of freedom
  (18758 observations deleted due to missingness)
AIC: 19699

Number of Fisher Scoring iterations: 5
```

Master degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 28 January 2020

Theory

1) Describe the purpose of a (parametric) hypothesis testing procedure. Define the notions of significance level, critical region and p-value. Present a simple application concerning the testing on the equality of the means of two independent populations.

2) Define the multiple linear regression model and highlight the basic assumptions. Discuss the crucial point of selecting the explanatory variables. Finally, discuss the problem of multicollinearity and consider the potential remedies.

Laboratory

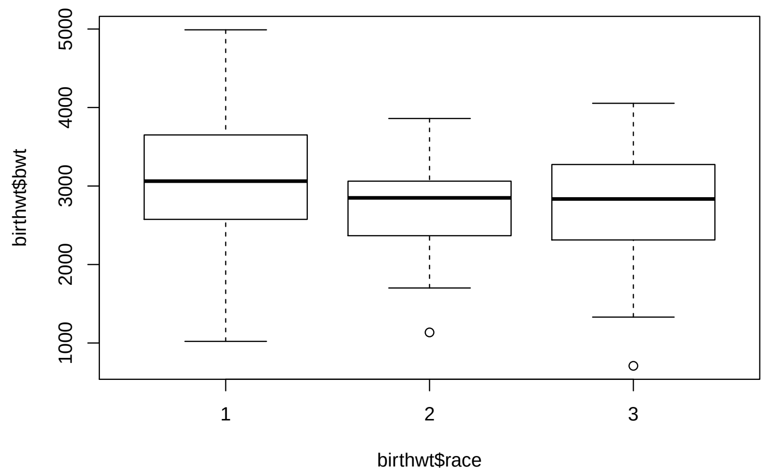
3) Consider the `R` commands below, describe what the code is intended to do and explain what is being calculated on each line.

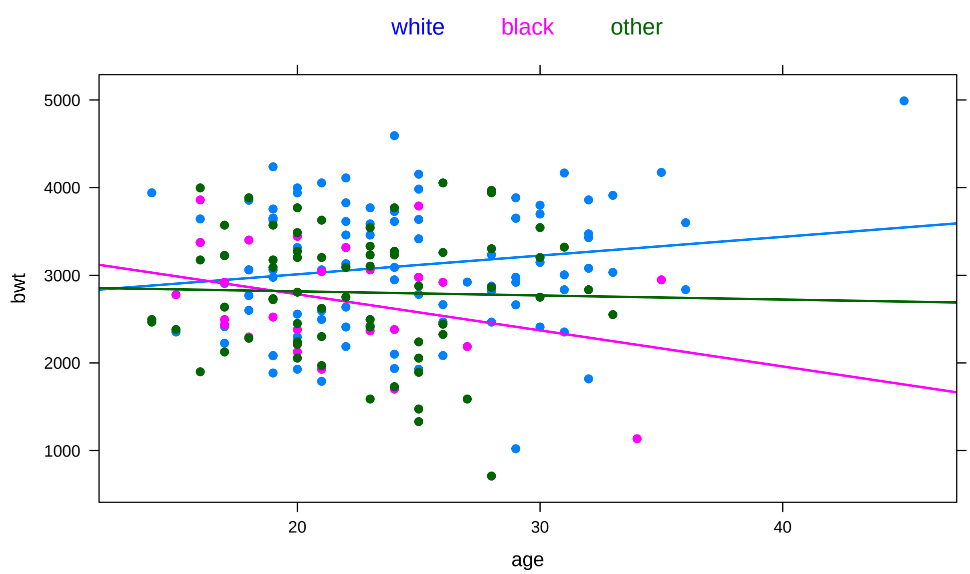
```
set.seed(4)
x <- seq(0,1.5,0.01)
sim1<-rbinom(1000,25,0.25)/25
sim2<-rbinom(1000,50,0.25)/50
sim3<-rbinom(1000,100,0.25)/100
#
par(mfrow=c(1,3),pty="s")
hist(sim1,freq=F,xlab="n=25",ylab=' ',main=' ')
lines(x,dnorm(x,0.25,sqrt(0.25*0.75/10)),lwd=2,col='red')
lines(density(sim1),lwd=2)
hist(sim2,freq=F,xlab="n=50",ylab=' ',main=' ')
lines(x,dnorm(x,0.25,sqrt(0.25*0.75/30)),lwd=2,col='red')
lines(density(sim2),lwd=2)
hist(sim3,freq=F,xlab="n=100",ylab=' ', main=' ')
lines(x,dnorm(x,0.25,sqrt(0.25*0.75/100)),lwd=2,col='red')
lines(density(sim3),lwd=2)
par(mfrow=c(1,1))
```

4) Let us consider the dataframe `birthwt`, which contains data on 189 births at the Baystate Medical Centre, Springfield, Massachusetts during 1986. The focus is on the variables listed below

```
bwt
  birth weight in grams
age
  mother's age in years
race
  mother's race (1 = white, 2 = black, 3 = other)
```

The aim of the study is to analyze the potential relationship between the response variable `bwt` and the explanatory variables `age` and `race`. Describe how to perform a preliminary data analysis on this dataframe using suitable `R` commands and comment the following plots





In order to describe the potential relationship between birth weight and age, taking into account also the factor `race`, we compare the following nested models

```r
bwt.lm1 <- lm(bwt ~ 1 , data = birthwt)
bwt.lm2 <- lm(bwt ~ age, data = birthwt)
bwt.lm3 <- lm(bwt ~ race + age, data = birthwt)
bwt.lm4 <- lm(bwt ~ race*age, data = birthwt)
```

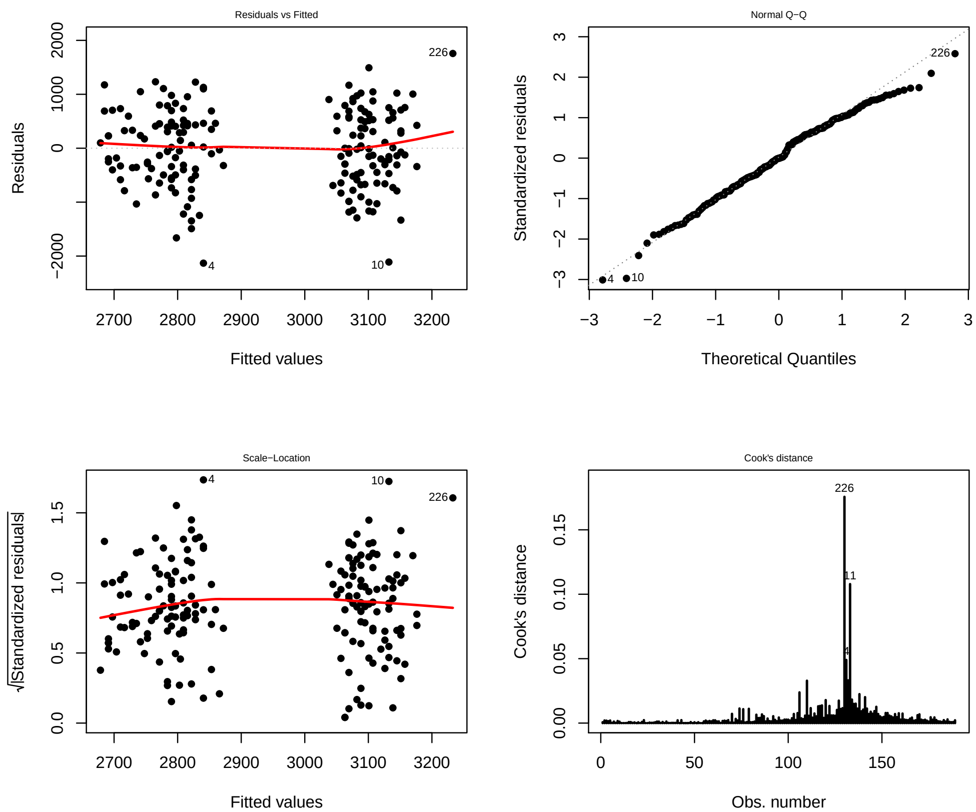
Describe the four models and comment the results given by the Analysis of Variance Table, reported below. Moreover, propose some alternative model selection procedures.

```
anova(bwt.lm1, bwt.lm2, bwt.lm3, bwt.lm4)

Analysis of Variance Table

Model 1: bwt ~ 1
Model 2: bwt ~ age
Model 3: bwt ~ race + age
Model 4: bwt ~ race * age
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1    188 99969656
2    187 99154173  1    815483 1.6145 0.20547
3    185 94754346  2   4399826 4.3555 0.01419 *
4    183 92431148  2   2323199 2.2998 0.10317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

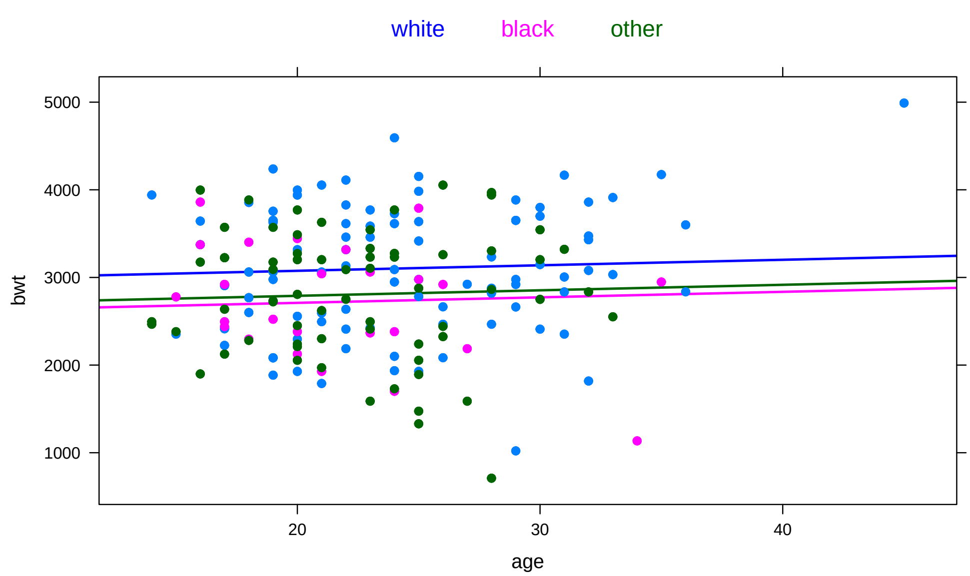
Let us consider Model 3 and comment the output obtained by the `R` functions `summary` and `plot`.

```
Call:
lm(formula = bwt ~ race + age, data = birthwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2131.57  -488.02    -1.16   521.87  1757.07

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2949.979    255.352  11.553   <2e-16 ***
race2       -365.715    160.636  -2.277   0.0240 *
race3       -285.466    115.531  -2.471   0.0144 *
age            6.288     10.073   0.624   0.5332
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 715.7 on 185 degrees of freedom
```

```
Multiple R-squared:  0.05217,Adjusted R-squared:  0.0368
F-statistic: 3.394 on 3 and 185 DF,  p-value: 0.01909
```



Finally, discuss the following graphical output and then suggest how to proceed with further analyses.



Master degree course in Computer Science - University of Udine
Corso di laurea magistrale in Informatica - Università degli Studi di Udine

Applied Statistics and Data Analysis

Written exam - 18 February 2020

Theory

1) Describe what are the aims of Exploratory Data Analysis and present the main graphical summaries for describing the relationship between different types (namely, categorical and numerical) of variables.

2) Define the simple linear regression model and recall the t test on the nullity of the slope parameter, discussing its role in evaluating the model adequacy. Define the one-way analysis of variance model and describe the statistical test on the effect of the factor on the mean response. Finally, compare the regression model and the ANOVA model when the levels of the factor are quantitative.

Laboratory

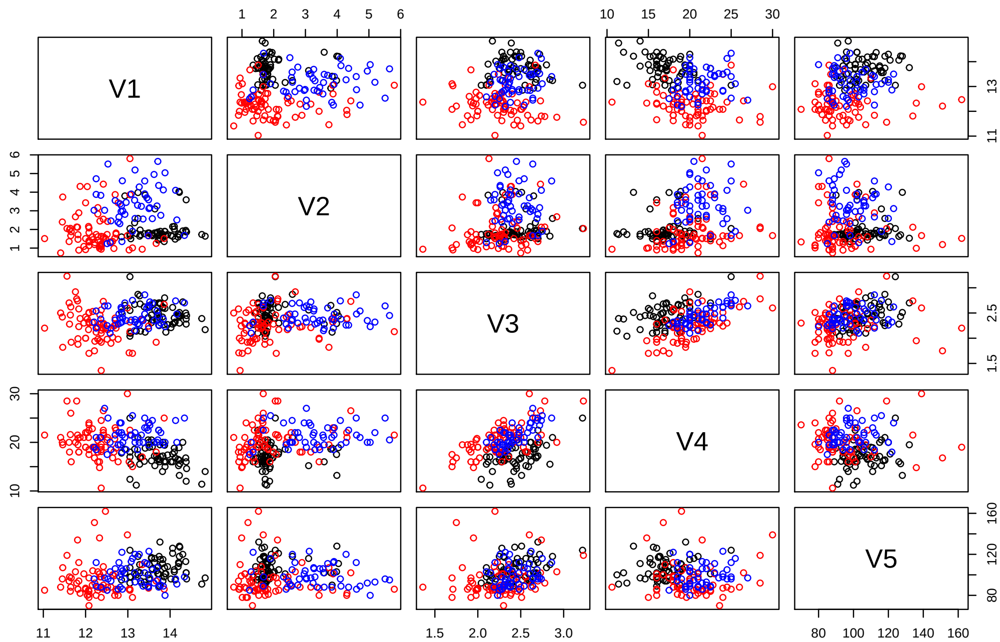
3) Describe the `R` functions that can be used for model selection. Furthermore, consider the `R` commands below, describe what the code is intended to do and explain what is being calculated on each line. Here, dataset `trees`, which provides some measurements on felled black cherry trees, is taken into account.

```
cv1 <- 0
cv2 <- 0
n <- length(trees$Volume)
i <-1
for (i in 1:n){
mod1i <- lm(Volume ~ Girth, data = trees[-i,])
mod2i <- lm(Volume ~ Girth + Height, data = trees[-i,])
mu1 <- mod1i$coefficients[1] + mod1i$coefficients[2]*trees$Girth[i]
mu2 <- mod2i$coefficients[1] + mod2i$coefficients[2]*trees$Girth[i] +
            mod2i$coefficients[3]*trees$Height[i]
sd1 <- sqrt(sum(mod1i$residuals^2)/(n-3))
sd2 <- sqrt(sum(mod2i$residuals^2)/(n-4))
cv1 <- cv1 - log(dnorm(trees$Volume[i],mu1,sd1))
cv2 <- cv2 - log(dnorm(trees$Volume[i],mu2,sd2))}
cv1
cv2
```

4) Let us consider the dataframe `wines`, which contains information about 178 samples of wines grown in the same region in Italy. The cultivar of each wine sample is observed (variable `cultivar`, with labels `1`, `2`, `3`), together with the concentration of the 13 different chemicals (variables `V1-V13`). Describe how to perform a preliminary data analysis on this dataframe using suitable `R` commands and comment the following outputs.

```
summary(wine)

 cultivar       V1              V2              V3       
 1:59     Min.   :11.03   Min.   :0.740   Min.   :1.360  
 2:71     1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210  
 3:48     Median :13.05   Median :1.865   Median :2.360  
          Mean   :13.00   Mean   :2.336   Mean   :2.367  
          3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558  
          Max.   :14.83   Max.   :5.800   Max.   :3.230  
       V4              V5               V6              V7       
 Min.   :10.60   Min.   : 70.00   Min.   :0.980   Min.   :0.340  
 1st Qu.:17.20   1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205  
 Median :19.50   Median : 98.00   Median :2.355   Median :2.135  
 Mean   :19.49   Mean   : 99.74   Mean   :2.295   Mean   :2.029  
 3rd Qu.:21.50   3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875  
 Max.   :30.00   Max.   :162.00   Max.   :3.880   Max.   :5.080  
       V8               V9              V10              V11        
 Min.   :0.1300   Min.   :0.410   Min.   : 1.280   Min.   :0.4800  
 1st Qu.:0.2700   1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825  
 Median :0.3400   Median :1.555   Median : 4.690   Median :0.9650  
 Mean   :0.3619   Mean   :1.591   Mean   : 5.058   Mean   :0.9574  
 3rd Qu.:0.4375   3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200  
 Max.   :0.6600   Max.   :3.580   Max.   :13.000   Max.   :1.7100  
      V12             V13        
 Min.   :1.270   Min.   : 278.0  
 1st Qu.:1.938   1st Qu.: 500.5  
 Median :2.780   Median : 673.5  
 Mean   :2.612   Mean   : 746.9  
 3rd Qu.:3.170   3rd Qu.: 985.0  
 Max.   :4.000   Max.   :1680.0  

sapply(wine[2:14],sd)

         V1          V2          V3          V4          V5 
  0.8118265   1.1171461   0.2743440   3.3395638  14.2824835 
         V6          V7          V8          V9         V10 
  0.6258510   0.9988587   0.1244533   0.5723589   2.3182859 
        V11         V12         V13 
  0.2285716   0.7099904 314.9074743 
```

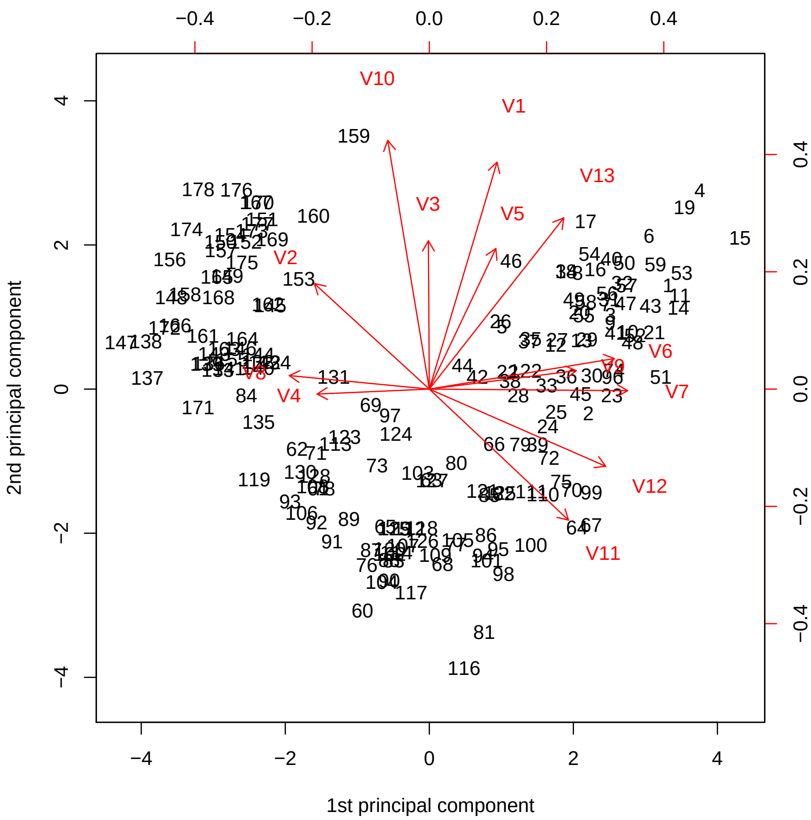
Moreover, discuss the results given by the scatterplot matrix considered below, which considers the first 5 numerical variables, with colours indicating **cultivar 1**, cultivar 2 and cultivar 3.

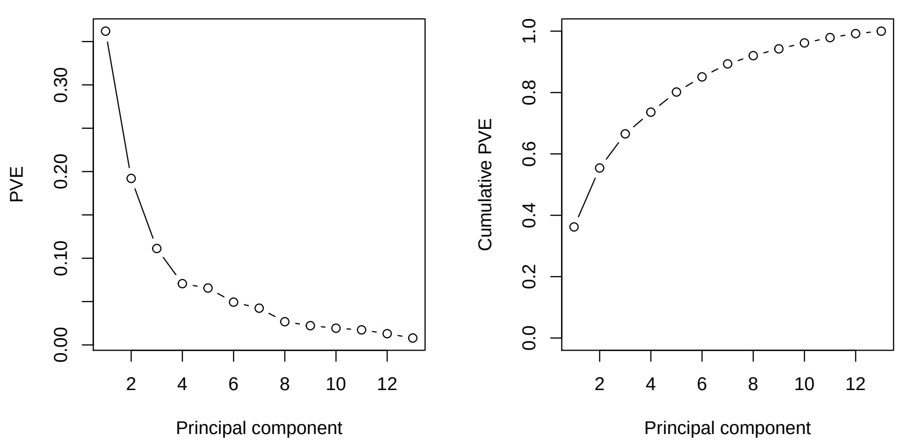


The aim of the study is to adequately synthesize the information given by the original variables `V1-V13`, in order to capture as much of the information as possible. A further objective is to use some of these new derived variable for distinguishing the three different cultivars.
The Principal Components Analysis procedure is applied. Present the main features of this statistical procedure, describe the arguments specified below in the function `princomp` and discuss the output of the function `loadings`.

```
wine.pca<-princomp(wine[2:14], cor=TRUE)
loadings(wine.pca)[,1:4]

          Comp.1       Comp.2      Comp.3      Comp.4
V1   0.144329395  0.483651548  0.20738262  0.01785630
V2  -0.245187580  0.224930935 -0.08901289 -0.53689028
V3  -0.002051061  0.316068814 -0.62622390  0.21417556
V4  -0.239320405 -0.010590502 -0.61208035 -0.06085941
V5   0.141992042  0.299634003 -0.13075693  0.35179658
V6   0.394660845  0.065039512 -0.14617896 -0.19806835
V7   0.422934297 -0.003359812 -0.15068190 -0.15229479
V8  -0.298533103  0.028779488 -0.17036816  0.20330102
V9   0.313429488  0.039301722 -0.14945431 -0.39905653
V10 -0.088616705  0.529995672  0.13730621 -0.06592568
V11  0.296714564 -0.279235148 -0.08522192  0.42777141
V12  0.376167411 -0.164496193 -0.16600459 -0.18412074
V13  0.286752227  0.364902832  0.12674592  0.23207086
```

Moreover, discuss the following graphical outputs





Finally, comment this last plot, with particular concern to the aim of characterizing the three different cultivars.