**NAME: ALEXANDROS STEFANIDIS**

**PROGRAMME: eLearning - Advanced Data Analysis using R (Spring 2023)**

TASK 1

a)

```
> lm( Sales ~ Market_Value, data=df )

Call:
lm(formula = Sales ~ Market_Value, data = df)

Coefficients:
 (Intercept)  Market_Value
   2395.6902        0.5452

> reg1 = lm( Sales ~ Market_Value, data=df )
> summary(reg1)

Call:
lm(formula = Sales ~ Market_Value, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4511.8 -2051.5 -1257.2   412.9 13588.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.396e+03  3.945e+02   6.073 4.44e-08 ***
Market_Value 5.452e-01  3.372e-02  16.168  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3366 on 77 degrees of freedom
Multiple R-squared:  0.7725,    Adjusted R-squared:  0.7695
F-statistic: 261.4 on 1 and 77 DF,  p-value: < 2.2e-16
```

b) The model is: *Sales* = 2396 + 0.542 * *Market_Value* + ε , $\varepsilon \sim N(0, 3366^2)$.

c) Explanation of the parameters:

$\beta_0 = 2396$ → The expected value of Sales if the Market_Value is zero, i.e. if the face value of the company is zero then the Annual Sales will be 2396 million dollars.

$\beta_1 = 0.542$ → If we increase the Market_Value by one unit (1 million dollars) the expected value of Sales will increase by 0.542 (half million dollars).

$\hat{\sigma}^2 = 0{,}3366^2$ is the estimated variance of the residuals.

$R^2_{adj} = 0.7695$ , which means that 77% of the total variance of the model is explained by the variable Market_Value.

d) Testing Normality of the residuals:

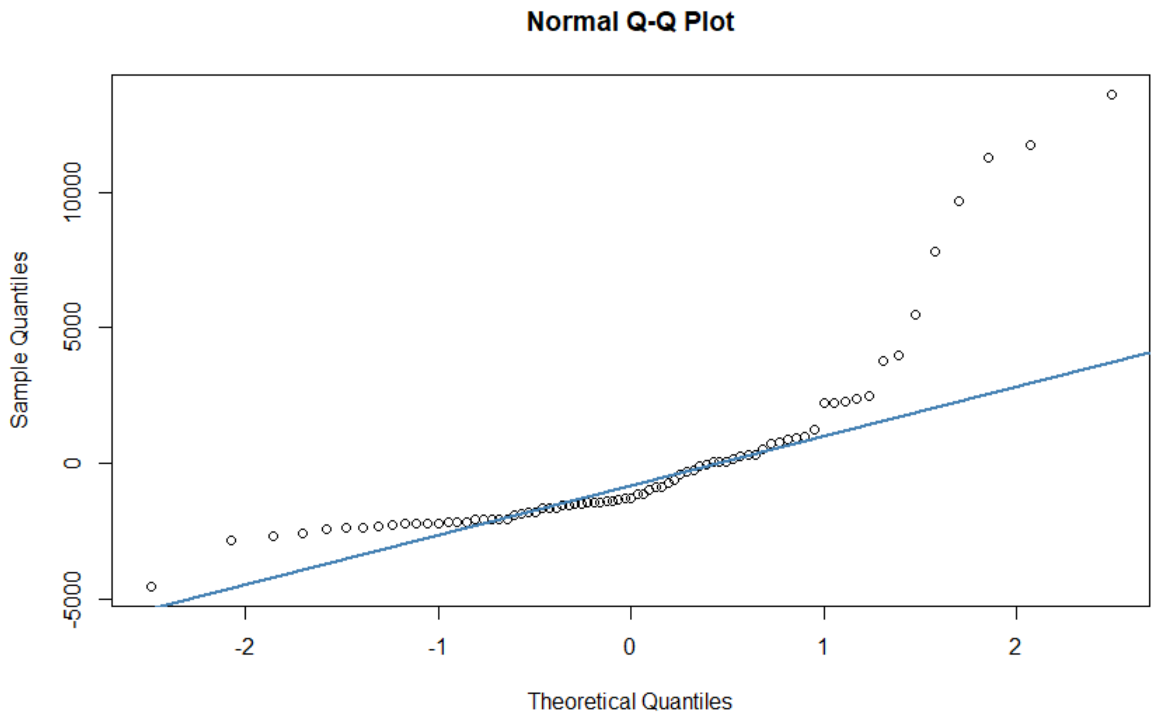


*Figure 1 Normal Q-Q Plot of the residuals*

The qqplot of the residuals shows that they do not come from a Normal distribution.

```
> library(nortest)
> lillie.test(reg1$residuals)

	Lilliefors (Kolmogorov-Smirnov) normality test

data:  reg1$residuals
D = 0.21084, p-value = 2.59e-09

> shapiro.test(reg1$residuals)

	Shapiro-Wilk normality test

data:  reg1$residuals
W = 0.71273, p-value = 3.867e-11
```

$H_0$ : The residuals follow the Normal distribution

$H_1$ : The residuals do not follow the Normal distribution

The p-value is $2{,}59 * 10^{-09} < 0.05 = a$ , therefore we reject the null hypothesis of the Lilliefors test. Also, strong evidence to reject the null hypothesis is given by the Shapiro-Wilk normality test.

Testing the Homoscedasticity of the residuals:

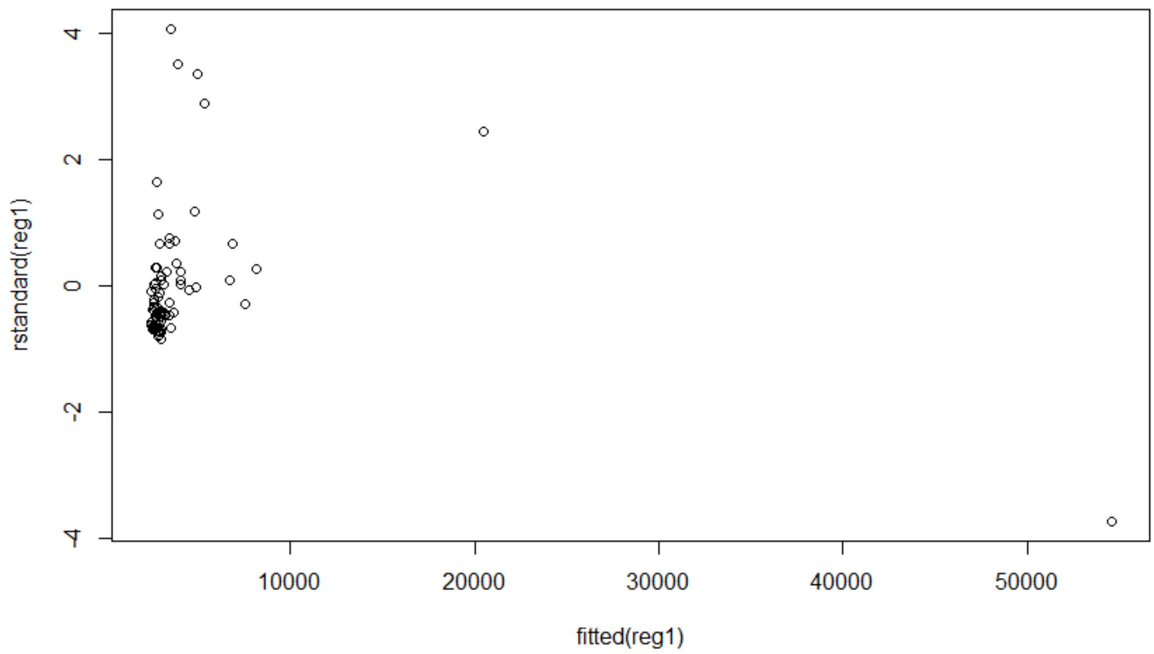


*Figure 2 Fitted values vs Standardized Residuals*

From the above plot we can see that the residuals of the model reg1 do not have constant variance.

Testing randomness of the residuals.

```
> runs.test(reg1$residuals)

	Runs Test

data:  reg1$residuals
statistic = 0, runs = 40, n1 = 39, n2 = 39, n = 78,
p-value = 1
alternative hypothesis: nonrandomness
```

We got a p-value of 1 therefore we accept the null hypothesis of the Runs test which means that the residuals are random.

e) Repeating the same process for the log transformation:

```
> reg2 = lm( log(Sales) ~ log(Market_Value), data=df )
> summary(reg2)

Call:
lm(formula = log(Sales) ~ log(Market_Value), data = df)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-2.42818 -0.46209 -0.06351  0.51314  1.83615 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        2.62837    0.54264   4.844 6.45e-06 ***
log(Market_Value)  0.71122    0.07655   9.291 3.30e-14 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8319 on 77 degrees of freedom
Multiple R-squared:  0.5285,	Adjusted R-squared:  0.5224 
F-statistic: 86.32 on 1 and 77 DF,  p-value: 3.304e-14
```

The model is $\log(Sales) = 0..63 + 71 * \log(MarketValue) + \varepsilon$

$R^2 = 0.52$ means that the variance explained by the variable log(Market_Value) is 50% of the total variance.

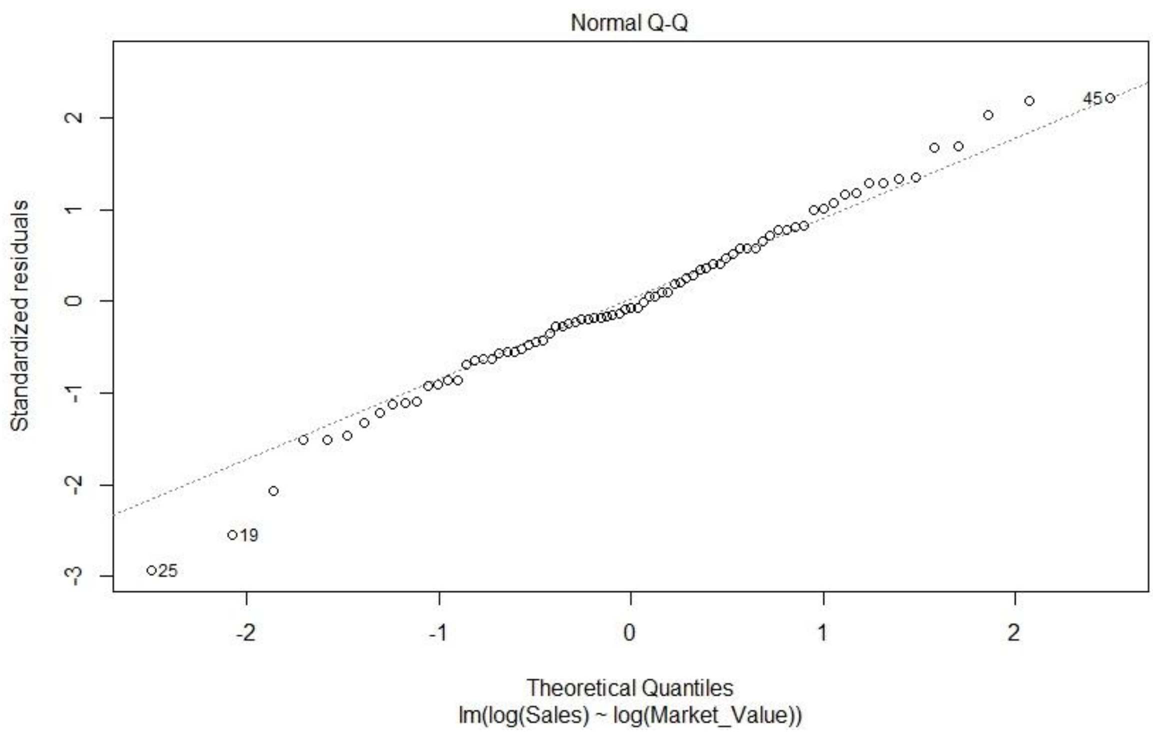


*Figure 3 Normal q-q plot*

The qqplot shows that the residuals come from the Normal distribution.

```
> library(nortest)
> lillie.test(reg2$residuals)

	Lilliefors (Kolmogorov-Smirnov) normality test

data:  reg2$residuals
```

```
D = 0.059249, p-value = 0.7063

> shapiro.test(reg2$residuals)

	Shapiro-Wilk normality test

data:  reg2$residuals
W = 0.98781, p-value = 0.66
```

In both of the Normality tests we figure out that we should accept the null hypothesis, i.e. the $\varepsilon \sim N(\mu, \sigma^2)$.

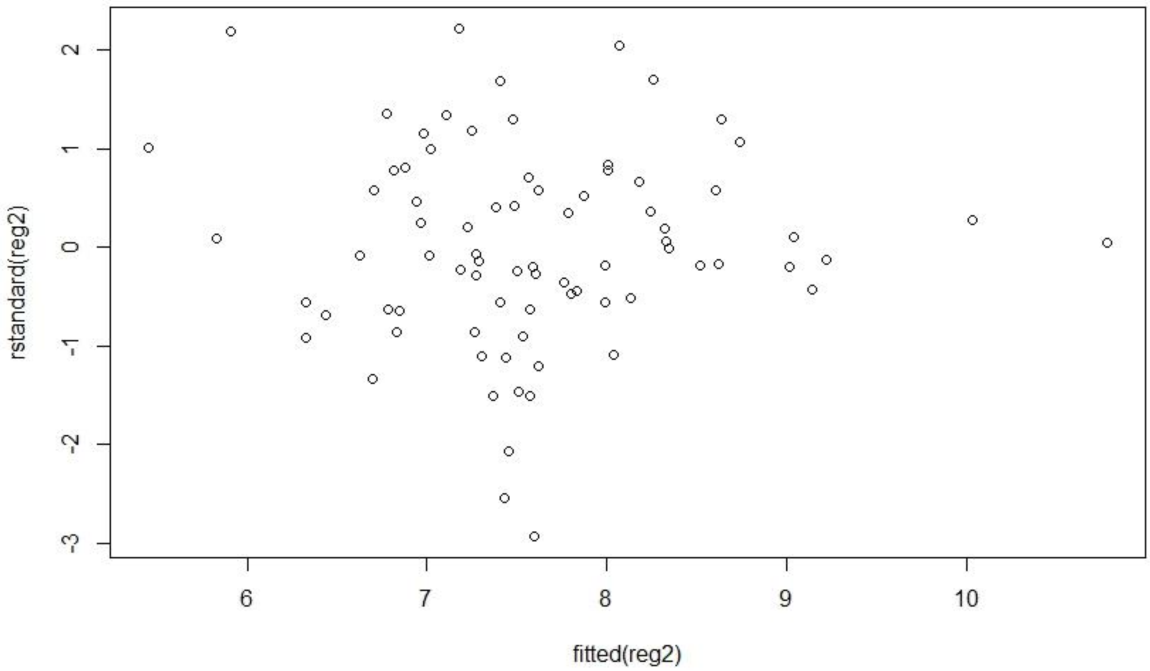


*Figure 4 Standardized residuals vs fitted values*

We have indication that the variance is constant.

```
> library(randtests)
> runs.test(reg2$residuals)

	Runs Test

data:  reg2$residuals
statistic = -1.3676, runs = 34, n1 = 39, n2 = 39, n =
78, p-value = 0.1714
alternative hypothesis: nonrandomness
```

$p - value = 0.17 > 0.05$ , therefore we accept the null hypothesis (the residuals are random).

We thus conclude that the model satisfies all the required assumptions.

f)

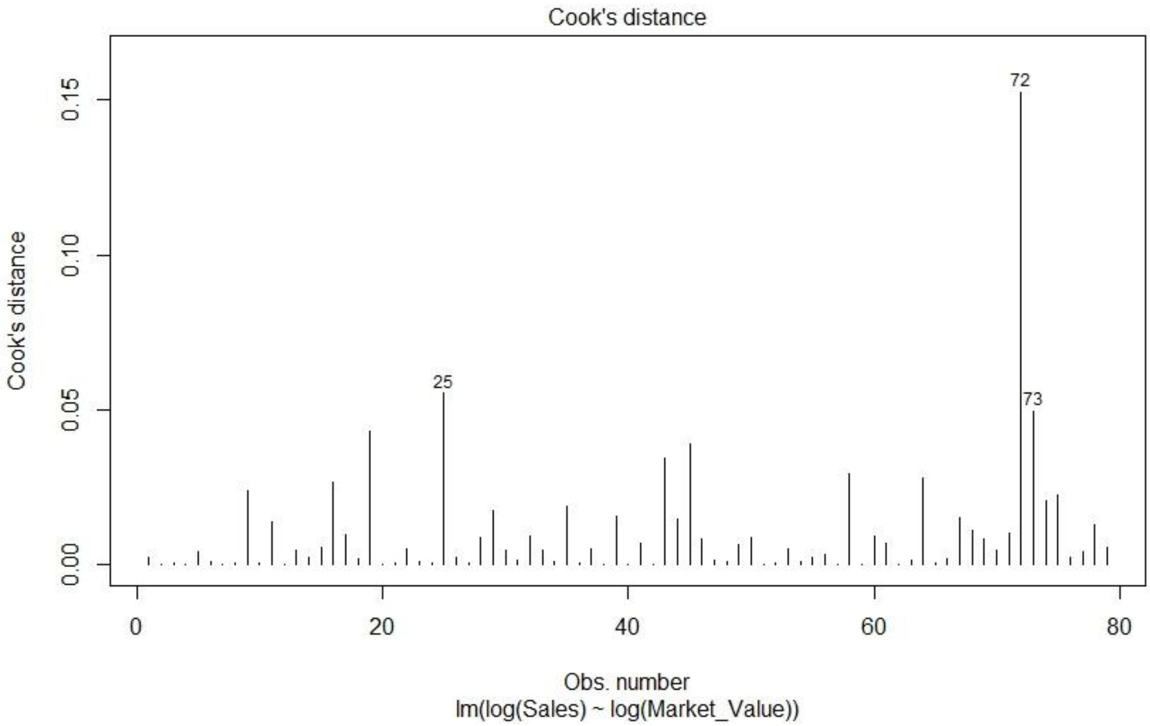


*Figure 5 Cook's distance*

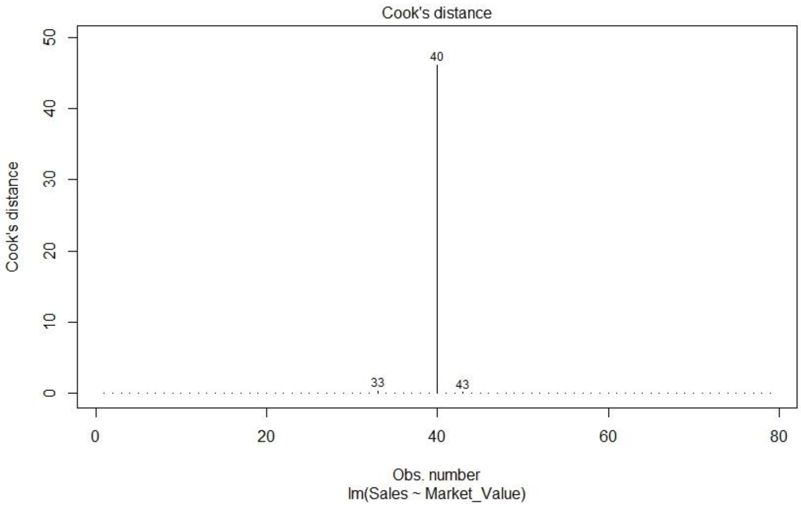


*Figure 6 Cook's distance*

We verify that there is only one outlier in the first model.

General Conclussion: We would prefer the second model because it satisfies all the regression assumptions.

TASK 2

a)

```
> AssetsTr = log(df$Assets)
> SalesTr = log(df$Sales)
> Market_ValueTr = log(df$Market_Value)
> ProfitTr = sign(df$Profit)*log(abs(df$Profit))
```

b)

```
> reg3 = lm( log(Sales) ~ AssetsTr + Market_ValueTr + ProfitTr
, data=df )
> summary(reg3)

Call:
lm(formula = log(Sales) ~ AssetsTr + Market_ValueTr + ProfitTr
,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.88923 -0.51108 -0.00287  0.45364  1.84159

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.33232    0.61575   2.164  0.03367 *
AssetsTr        0.26265    0.08235   3.190  0.00208 **
Market_ValueTr  0.63323    0.08508   7.443 1.36e-10 ***
ProfitTr       -0.06774    0.02909  -2.329  0.02258 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7569 on 75 degrees of freedom
Multiple R-squared:  0.6198,    Adjusted R-squared:  0.6046 
F-statistic: 40.76 on 3 and 75 DF,  p-value: 9.81e-16
```

c)

The model is

$\log(Sales) = 1.33 + 0.263 * AssetsTr + 0.63 * MarkeValueTr - 0.068 * ProfitTr + \varepsilon$ .

$\widehat{\beta_0} = 1{,}33$ means that if the $AssetsTr$, the $MarketValueTr$ and the $ProfitTr$ are zero, then expected value of $\log(Sales)$ is 1,33 million dollars.

$\widehat{\beta_1} = 0{,}263$ means that if we increase by one unit the variable AssetsTr (while all the other variables remain constant) then the expected value of $\log(Sales)$ will increase by 0,023 million dollars.

The same goes for the other constants $\widehat{\beta_2},\ \hat{\beta}_3$

$R_{adj} = 0.605$ means that the variance of the model explained by the variables AssetsTr, Market_ValueTr and ProfiTr is 60%.

The estimated variance of the residuals is $\sigma^2 = 0.7569$, which means that 75% of the total variance is explained by the variabels AssetsTr, Market_ValueTr and ProfitTr.

```
d)
> final_model = step(reg3, direction='both')
Start:  AIC=-40.11
log(Sales) ~ AssetsTr + Market_ValueTr + ProfitTr

                 Df Sum of Sq    RSS     AIC
<none>                        42.967 -40.113
- ProfitTr        1     3.106 46.073 -36.598
- AssetsTr        1     5.828 48.795 -32.064
- Market_ValueTr  1    31.735 74.702   1.580
> final_model

Call:
lm(formula = log(Sales) ~ AssetsTr + Market_ValueTr + ProfitTr
,
    data = df)

Coefficients:
   (Intercept)        AssetsTr  Market_ValueTr        ProfitTr
       1.33232         0.26265         0.63323        -0.06774
```

The model is $\log(Sales) = 1{,}33 + 0.262 * AssetsTr + 0{,}63 * MarketValueTr - 0{,}068 * ProfitTr + \varepsilon$.

TASK 4

```
> df = read.table('C:/Users/aleks/Desktop/eLearning_Folder/Adv
anced_Data_Analysis_using_R/Telikh_Ergasia/companies.txt', hea
der=TRUE)
> attach(df)
The following objects are masked from df (pos = 3):

    Assets, Company_Name, Employees, Market_Value,
    Profits, Sales, Sector

The following objects are masked from df (pos = 4):

    Assets, Company_Name, Employees, Market_Value,
    Profits, Sales, Sector

>
> # change of variable
> for (i in 1:79) {
+   if (Profits[i] > 0) {
+     df$Profitable[i] <- 1
+   }
+   else {
+     df$Profitable[i] <- 0
+   }
+ }
> attach(df)
The following objects are masked from df (pos = 3):

    Assets, Company_Name, Employees, Market_Value,
    Profits, Sales, Sector

The following objects are masked from df (pos = 4):

    Assets, Company_Name, Employees, Market_Value,
    Profitable, Profits, Sales, Sector

The following objects are masked from df (pos = 5):

    Assets, Company_Name, Employees, Market_Value,
    Profits, Sales, Sector

>
>
> # Model selection
> model3 <- glm(Profitable ~ Assets + Sales + Market_Value + E
mployees + Sector
+                ,family = binomial)
> final_model3 = step(model3, direction='both')
Start:  AIC=64.46
Profitable ~ Assets + Sales + Market_Value + Employees + Secto
r

               Df Deviance    AIC
- Sector        8   47.242 57.242
- Employees     1   38.475 62.475
- Market_Value  1   38.484 62.484
- Assets        1   38.771 62.771
```

```
- Sales          1   38.885 62.885
<none>               38.458 64.458

Step:  AIC=57.24
Profitable ~ Assets + Sales + Market_Value + Employees

               Df Deviance    AIC
- Employees     1   47.255 55.255
- Sales         1   47.770 55.770
- Assets        1   48.330 56.330
<none>              47.242 57.242
- Market_Value  1   49.888 57.888
+ Sector        8   38.458 64.458

Step:  AIC=55.26
Profitable ~ Assets + Sales + Market_Value

               Df Deviance    AIC
- Assets        1   48.570 54.570
- Sales         1   48.577 54.577
<none>              47.255 55.255
- Market_Value  1   50.229 56.229
+ Employees     1   47.242 57.242
+ Sector        8   38.475 62.475

Step:  AIC=54.57
Profitable ~ Sales + Market_Value

               Df Deviance    AIC
<none>              48.570 54.570
+ Assets        1   47.255 55.255
- Market_Value  1   51.479 55.479
- Sales         1   51.659 55.659
+ Employees     1   48.330 56.330
+ Sector        8   38.785 60.785
> summary(final_model3)

Call:
glm(formula = Profitable ~ Sales + Market_Value, family = bino
mial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2614   0.3783   0.4012   0.4388   1.1566  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   2.554e+00  5.186e-01   4.924 8.48e-07 ***
Sales        -1.651e-04  8.973e-05  -1.840   0.0658 .  
Market_Value  1.544e-04  1.499e-04   1.030   0.3031    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.801  on 78  degrees of freedom
Residual deviance: 48.570  on 76  degrees of freedom
```

```
AIC: 54.57

Number of Fisher Scoring iterations: 7
```

B)

The selected model is

$$\log\frac{Profitable\_i}{1-Profitable\_i} = 2{,}5 - 0{,}00016 * Sales + 0{,}00015 * Market\_Value$$

$$\text{όπου } Y_i \sim Binomial(Profitable_i, N_i)$$

C)

```
> # (c)
> null = glm(Profitable ~ 1, family = "binomial")
> anova(null, final_model3, test="Chisq")
Analysis of Deviance Table

Model 1: Profitable ~ 1
Model 2: Profitable ~ Sales + Market_Value
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        78     51.801
2        76     48.570  2   3.2309   0.1988


>
```

The model does not differ significantly from the constant model (p-value=0.19>0.5)

D)

```
> # (d)
> newdata = data.frame(Sector = 'A', Assets = mean(Assets), Ma
rket_Value = mean(Market_Value)
+                        , Sales= mean(Sales), Employees = mean(
Employees))
> predict(final_model3, newdata, type="response")
        1
0.9143898
> val1 = predict(final_model3, newdata, type="response")
> p = exp(val1)/(1+exp(val1))
> print(p)
        1
0.7138976
```

The probability for a company of sector A to be profitable is 0.713.

TASK 3

```
> # TASK 3
>
> # a)
> -
> df_4 = read.table('C:/Users/aleks/Desktop/eLearning_Folder/A
dvanced_Data_Analysis_using_R/Telikh_Ergasia/companies.txt', h
eader=TRUE)
> attach(df_4)
The following objects are masked from df_4 (pos = 3):

    Assets, Company_Name, Employees, Market_Value, Profits, Sa
les,
    Sector

The following objects are masked from df_4 (pos = 4):

    Assets, Company_Name, Employees, Market_Value, Profits, Sa
les,
    Sector

> model4 <- glm(Market_Value ~ Assets + Sales + Profits + Empl
oyees + Sector,
+                data = df_4,family = poisson(link=log))
> final_model4 = step(model4, direction='both')
Start:  AIC=34949.87
Market_Value ~ Assets + Sales + Profits + Employees + Sector

            Df Deviance   AIC
<none>            34227 34950
- Sales      1    35597 36318
- Profits    1    35897 36618
- Employees  1    45981 46702
- Assets     1    46160 46881
- Sector     8    68130 68837
```

b)

```
> summary(final_model4)

Call:
glm(formula = Market_Value ~ Assets + Sales + Profits + Employ
ees +
    Sector, family = poisson(link = log), data = df_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-34.348  -17.321   -3.819    8.260   62.548

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.952e+00  7.385e-03 941.350   < 2e-16 ***
```

```
Assets        7.050e-05  6.561e-07 107.457  < 2e-16 ***
Sales        -5.209e-05  1.426e-06 -36.539  < 2e-16 ***
Profits      -1.614e-04  3.825e-06 -42.188  < 2e-16 ***
Employees     9.371e-03  8.954e-05 104.656  < 2e-16 ***
SectorB      -1.380e+00  1.542e-02 -89.500  < 2e-16 ***
SectorC      -4.470e-01  1.569e-02 -28.485  < 2e-16 ***
SectorD       7.190e-01  1.086e-02  66.207  < 2e-16 ***
SectorE      -1.349e-01  1.103e-02 -12.228  < 2e-16 ***
SectorF       7.068e-01  1.268e-02  55.738  < 2e-16 ***
SectorG      -7.551e-02  1.592e-02  -4.744 2.09e-06 ***
SectorH      -1.613e-01  1.174e-02 -13.743  < 2e-16 ***
SectorI       1.153e-01  1.147e-02  10.049  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 764400  on 78  degrees of freedom
Residual deviance:  34227  on 66  degrees of freedom
AIC: 34950

Number of Fisher Scoring iterations: 5
```

The final model is

$$\log(\lambda_i) = 6.9 + 7.06 \times 10^{-5} * Assets_i + \ -5.209 \times 10^{-5} * Sales_i - 1.614 \times 10^{-4} * Profits_i + 9.371 \times 10^{-3} * Employees_i - 1.380 * SectorB_i - 0.4 * SectorC_i + 0.7 \times SectorD_i - 0.13 * SectorE_i + 0.7 * SectorF_i - 0.07 * SectorG_i - 0.16 * SectorH_i - 0.115 * SectorI_i$$

Interpretation of the parameters:

$e^{\widehat{\beta_0}} = e^{6.9} = 992.7$ millions USD, is the expected relative change of the Market_Value when all the other covariates are constant.

$e^{\widehat{\beta_1}} = e^{7.06\times10^{-5}} = 1.000071$ is the expected relative change of $Market\_Values$ when $Assets_i$ is increased by one unit. That means that the face value of the company will increase by (1.000071-1) * (100%) = 0.007 % when the property owned by the company will increase by one million USD.

$e^{\widehat{\beta_2}} = e^{-5.209\times10^{-5}} = 0.999$ is the expected relative change of $Market\_Values$ when $Sales_i$ is increased by one unit. That means that the face value of the company will decrease by (1-0.999) * (100%) = 0.1 % when the annual sales of the company will increase by one million USD.

Similar interpretation goes with the rest of the parameters.

c)

```
> null_4 <- glm(Market_Value~1, data = df_4, family=poisson)
> anova (final_model4, null_4, test = "Chisq")
Analysis of Deviance Table

Model 1: Market_Value ~ Assets + Sales + Profits + Employees +
Sector
Model 2: Market_Value ~ 1
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1        66      34227
2        78     764400 -12  -730173 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

For the above test we got that p-value=2.2*10^(-16) < 0.05, therefore our final model differs from the constant model, significantly.

d)

```
> newdata4 = data.frame(Sector = 'A', Assets = mean(Assets),
+                        Sales= mean(Sales), Profits = mean(Pro
fits), Employees = mean(Employees))
> exp(predict(final_model4, newdata = newdata4))
       1
1756.807
```

We conclude that the predicted face value of the company for the sector A, when all the other covariates are equal to their mean value, will be 1756,8 million dollars.

TASK 5

a)

```
> library(MASS)
> model5 <- lda(Sector ~ Assets + Sales +
+                  Market_Value + Profits +Employees, data = df
)
> model52 <- predict(model5, data = df)
> t <- table(model52$class, df$Sector)
> t

     A  B  C  D  E  F  G  H  I
  A 14  5  4  2  7  1  3  2  6
  B  1 12  0  0  0  0  0  0  0
  C  0  0  0  0  0  0  0  0  0
  D  0  0  0  2  0  0  0  0  0
  E  0  0  1  2  2  0  1  3  0
  F  0  0  0  0  0  1  0  0  0
  G  0  0  0  0  0  0  0  0  0
  H  0  0  0  0  1  0  0  5  0
  I  0  0  1  2  0  0  0  0  1
> sum(diag(t))/sum(t)
[1] 0.4683544
```

The percentage of correct fitted values is 46%.

c)

```
> # c)
> library(class)
> df5 <- df[ , ! names(df) %in% c("Company_Name", "Sector")]
> model5c <- knn(train = df5, test = df5, cl = Profitable , k=
4)
> t2 <- table(model5c, Profitable)
> t2
       Profitable
model5c  0  1
      0  3  0
      1  5 71
> sum(diag(t2))/sum(t2)
[1] 0.9367089
```

The percentage of the correct fitted values is 93.7%

d)

```
install.packages("tree")

library("tree")

fit1 <- tree(as.factor(Profitable) ~ Assets + Sales +

        Market_Value +Employees, data = df5)

summary(fit1)

plot(fit1, lwd=2, col="blue")

text(fit1, cex = 1.3)
```

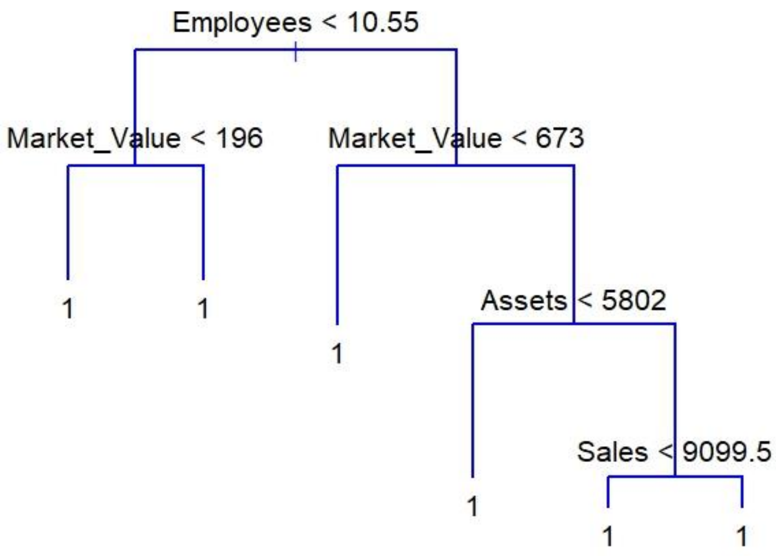


*Figure 7 Decision Tree*