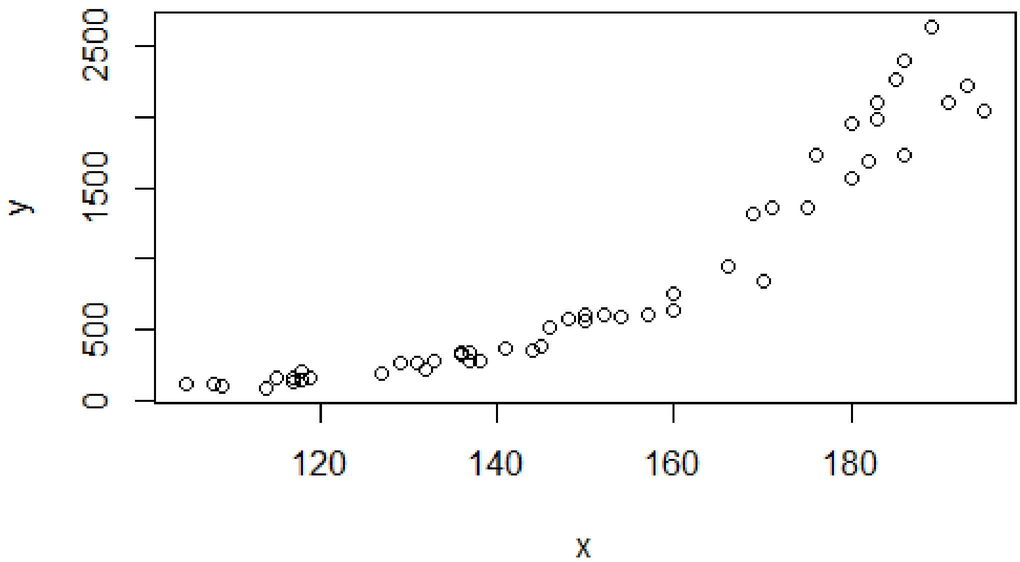
Project 2

1)

a) Construct the standard model Y = β0+ β1 X+ ε and check the model assumptions. Which of them are not met?

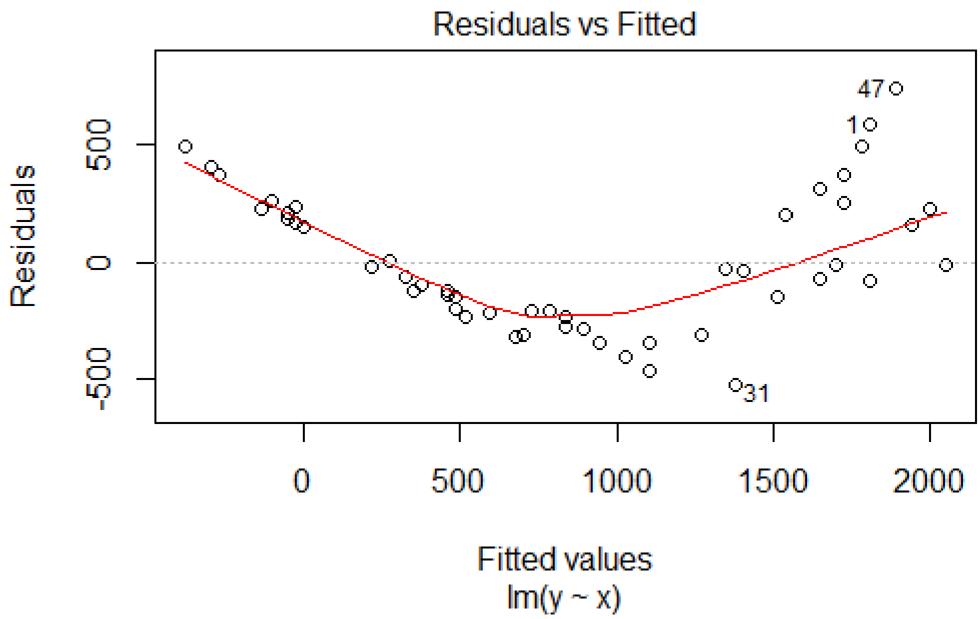



As it can be seen at the scatterplot, the linearity assumption is violated.

```
reg<-lm(formula=y~x)
summary(reg)

# Call:
# lm(formula = y ~ x)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -528.23 -219.89  -50.69  221.97  745.73
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -3204.156    242.602  -13.21   <2e-16 ***
# x              26.949      1.584   17.01   <2e-16 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 296.5 on 48 degrees of freedom
# Multiple R-squared:  0.8578, Adjusted R-squared:  0.8548
# F-statistic: 289.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

Regression indicates that X has significance for response Y. However, adjusted R-squared value can be better.

plot(reg)

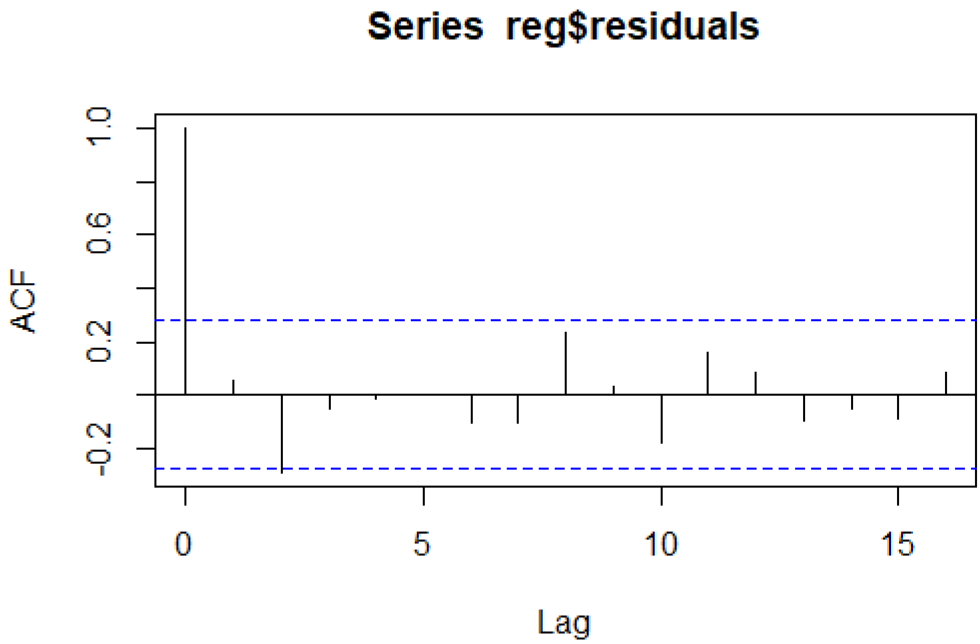



There is a pattern of residuals. It indicates there is a problem in our model.

```
mean(reg$residuals)
# [1] -9.667822e-15
```

Mean of residuals is acceptable.

```
acf(reg$residuals)
```

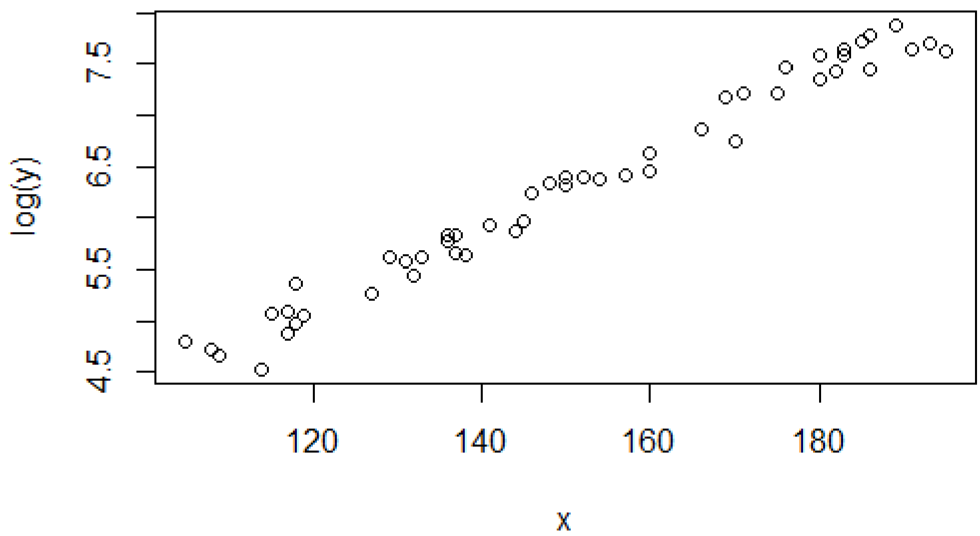



Autocorrelation of residuals indicates that there is not significant correlation within residuals.

However, since there is a pattern of residuals and linearity assumption is violated, this model is not suitable.

b) To fulfill the model assumptions, suggest a better model and re-check the model assumptions.

plot(x,log(y))




This model seems to fulfill the linearity assumption.

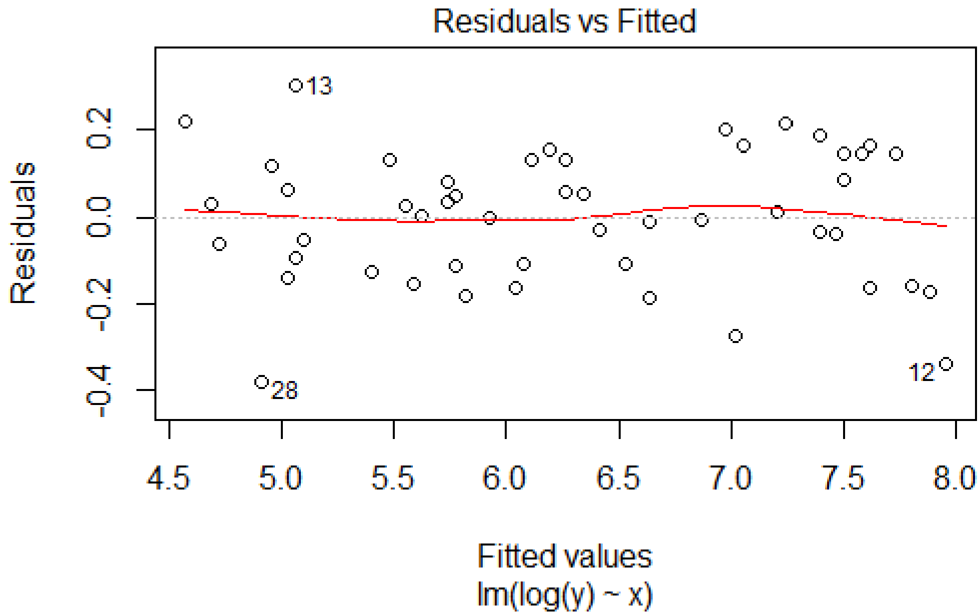
reglog<-lm(formula=log(y)~x)

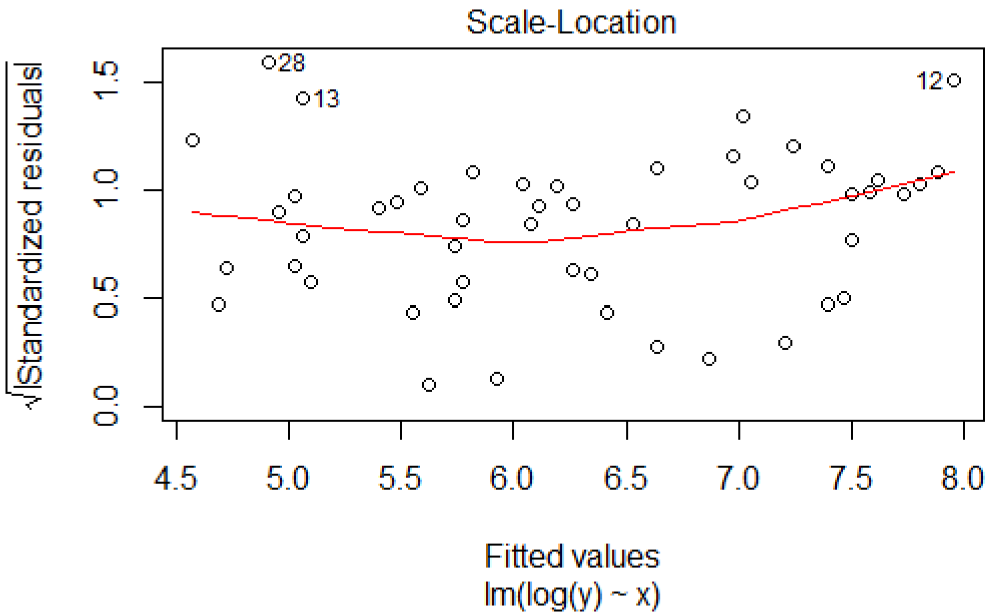
summary(reglog)

```
# Call:
# lm(formula = log(y) ~ x)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.37901 -0.11226  0.00721  0.13279  0.30396
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.6249870  0.1253748   4.985 8.49e-06 ***
# x           0.0376020  0.0008186  45.937  < 2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.1532 on 48 degrees of freedom
# Multiple R-squared:  0.9778, Adjusted R-squared:  0.9773
# F-statistic:  2110 on 1 and 48 DF,  p-value: < 2.2e-16
```

Adjusted R-squared value is satisfactory.

plot(reglog)



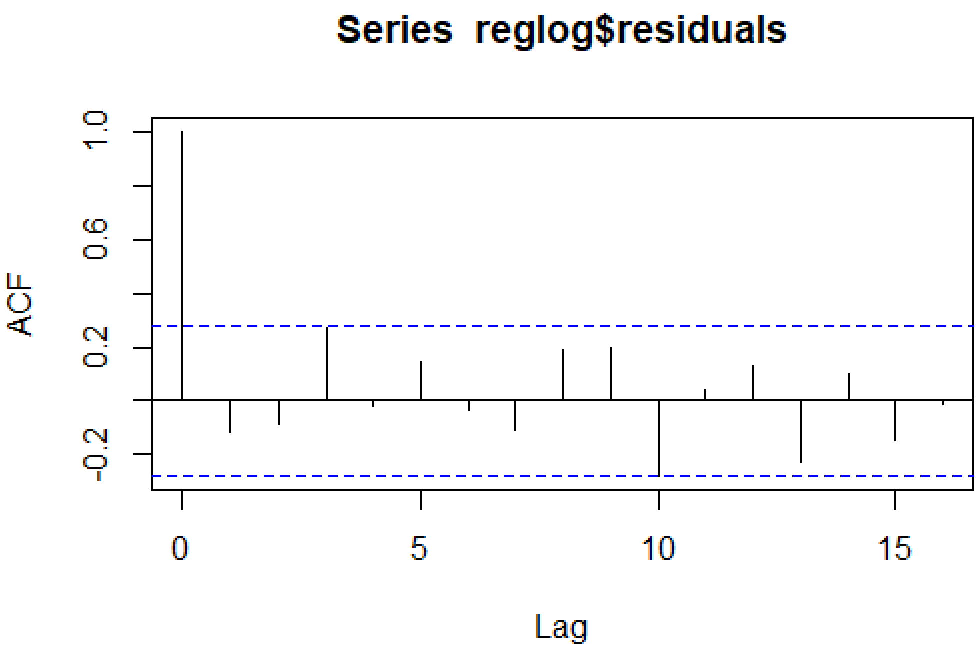




As it can be seen from the graphs, residuals don't have any pattern.

```
mean(reglog$residuals)
# [1] 1.834555e-18
```

Mean of residuals is small enough.

acf(reglog$residuals)




There is not any manipulative correlation within residuals.

c) Interpret the estimated regression coefficient for X and construct a 95% confidence interval for it.

0.0376020 is the regression coefficient for X. So, for 1 increase in the value of X affects the response variable log(Y) as 0.0376020.

```
confint(reglog,parm="x", level = 0.95)
#         2.5 %     97.5 %
# x 0.03595614 0.03924779
```

Confidence interval is (0.03595614, 0.03924779)

d) Using "predict" command, make forecasts for observed values x = 125 and x = 250 of the explanatory variable. Discuss the reliability of these forecasts.

```
xdata<-data.frame(x=c(125,250))
exp(predict(reglog, xdata))
#        1
# 525.9897
```

Since those values are point estimation, probably they are not the exact values of Y for these X values. Interval estimation is more logical and reliable.

e) Construct a 95% confidence interval and prediction interval for x = 150.

```
xdata<-data.frame(x=150)
exp(predict(reglog, xdata, level=0.95, interval="prediction"))
#        fit      lwr      upr
# 1 525.9897 385.3353 717.9854
exp(predict(reglog, xdata, level=0.95, interval="confidence"))
#        fit      lwr      upr
# 1 525.9897 503.5519 549.4272
```

f) How are the confidence and prediction intervals different than each other? Explain the reason of the difference between them.

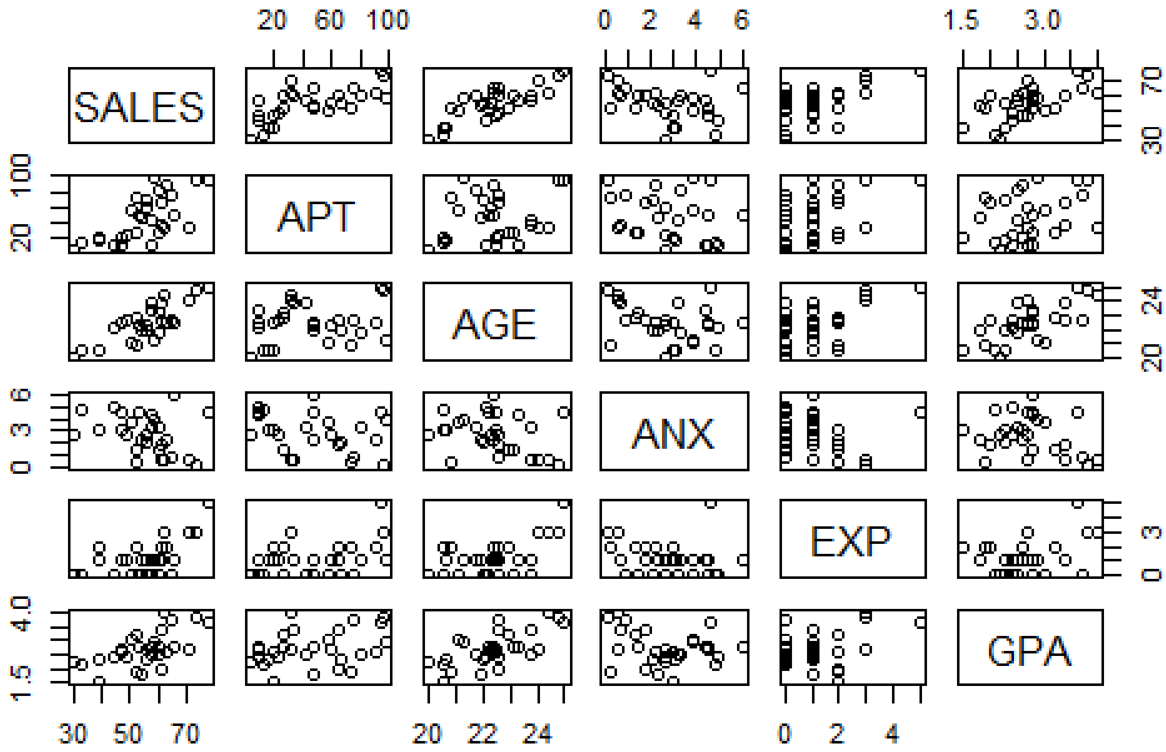
Prediction interval is used with predictions in regression analysis; it is a range of values that predicts the value of Y, based on the model. However, confidence interval is an interval for mean prediction value.

2)

a) Calculate the correlation matrix of all 6 variables and look at all scatter plots between the variables. Which variables do you think are needed to forecast sales values?

```
salesperson<-read.table("C:/Users/gülce/Desktop/3-2/IE360/Assignment 2/sale
sperson.txt", header=TRUE )

salesperson
#    SALES APT  AGE ANX EXP GPA
# 1     44  10 22.1 4.9   0 2.4
# 2     47  19 22.5 3.0   1 2.6
# 3     60  27 23.1 1.5   0 2.8
# 4     71  31 24.0 0.6   3 2.7
# 5     61  64 22.6 1.8   2 2.0
# 6     60  81 21.7 3.3   1 2.5
# 7     58  42 23.8 3.2   0 2.5
# 8     56  67 22.0 2.1   0 2.3
# 9     66  48 22.4 6.0   1 2.8
# 10    61  64 22.6 1.8   1 3.4
# 11    51  57 21.1 3.8   0 3.0
# 12    47  10 22.5 4.5   1 2.7
# 13    53  48 22.2 4.5   0 2.8
# 14    74  96 24.8 0.1   3 3.8
# 15    65  75 22.6 0.9   0 3.7
# 16    33  12 20.5 4.8   0 2.1
# 17    54  47 21.9 2.3   1 1.8
# 18    39  20 20.5 3.0   2 1.5
# 19    52  73 20.8 0.3   2 1.9
# 20    30   4 20.0 2.7   0 2.2
# 21    58   9 23.3 4.4   1 2.8
# 22    59  98 21.3 3.9   1 2.9
# 23    52  27 22.9 1.4   2 3.2
# 24    56  59 22.3 2.7   1 2.7
# 25    49  23 22.6 2.7   1 2.4
# 26    63  90 22.4 2.2   2 2.6
# 27    61  34 23.8 0.7   1 3.4
# 28    39  16 20.6 3.1   1 2.3
# 29    62  32 24.4 0.6   3 4.0
# 30    78  94 25.0 4.6   5 3.6
```

```
summary(salesperson)
#        SALES            APT              AGE              ANX              EXP
#  GPA
#   Min.   :30.0   Min.   : 4.00   Min.   :20.00   Min.   :0.100   Min.   :
#  0.0   Min.   :1.500
#   1st Qu.:49.5   1st Qu.:20.75   1st Qu.:21.75   1st Qu.:1.575   1st Qu.:
#  0.0   1st Qu.:2.325
#   Median :57.0   Median :44.50   Median :22.45   Median :2.700   Median :
#  1.0   Median :2.700
#   Mean   :55.3   Mean   :45.90   Mean   :22.41   Mean   :2.713   Mean   :
#  1.2   Mean   :2.713
#   3rd Qu.:61.0   3rd Qu.:66.25   3rd Qu.:23.05   3rd Qu.:3.875   3rd Qu.:
#  2.0   3rd Qu.:2.975
#   Max.   :78.0   Max.   :98.00   Max.   :25.00   Max.   :6.000   Max.   :
#  5.0   Max.   :4.000
```




As it can be seen in scatter plots, APT and AGE variables seem to be needed to construct the model.

b) Implement stepwise regression by following the steps below and obtain a final regression model.

Step 1: Choose the variable having the highest absolute correlation value. Construct an initial simple linear regression model using this variable and the response.

```
cor(x=salesperson$SALES, y=salesperson)
#      SALES       APT       AGE        ANX      EXP       GPA
# [1,]     1 0.6761204 0.7981406 -0.2958598 0.549834 0.6217841
```

AGE has the highest absolute correlation value.

```
regressionAGE<-lm(formula=SALES~AGE, data=salesperson)
summary(regressionAGE)
#
#Call:
#lm(formula = SALES ~ AGE, data = salesperson)
#
#Residuals:
#    Min      1Q  Median      3Q     Max
#-9.1399 -6.9177  0.6793  4.6449 11.4345
#
#Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -100.853     22.311   -4.52 0.000103 ***
#AGE             6.968      0.994    7.01 1.27e-07 ***
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
#Residual standard error: 6.847 on 28 degrees of freedom
#Multiple R-squared:  0.637,   Adjusted R-squared:  0.6241
#F-statistic: 49.14 on 1 and 28 DF,  p-value: 1.267e-07
```

Step 2: Out of the variables that are not in the model, build a new model by adding one variable into your current model. Use the command anova(currentmodel,newmodel) to test the significance of this new variable with an F-test. Do this for all variables which are not in the current model. Choose the variable that corresponds to largest F-statistic (smallest p-value) and update your current model by adding this variable.

```
anova(regressionAGE)
#Analysis of Variance Table
#
#Response: SALES
#          Df Sum Sq Mean Sq F value    Pr(>F)
#AGE        1 2303.7 2303.69  49.141 1.267e-07 ***
#Residuals 28 1312.6   46.88
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
regressionAGEAPT<-lm(formula=SALES~AGE+APT, data=salesperson)
anova(regressionAGE, regressionAGEAPT)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE
# Model 2: SALES ~ AGE + APT
#   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
# 1     28 1312.61
# 2     27  380.42  1     932.2 66.162 9.757e-09 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

regressionAGEANX<-lm(formula=SALES~AGE+ANX, data=salesperson)
anova(regressionAGE, regressionAGEANX)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE
# Model 2: SALES ~ AGE + ANX
```

```
#    Res.Df     RSS Df Sum of Sq      F Pr(>F)
# 1      28 1312.6
# 2      27 1295.0  1    17.666 0.3683  0.549
```

```
regressionAGEEXP<-lm(formula=SALES~AGE+EXP, data=salesperson)
anova(regressionAGE, regressionAGEEXP)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE
# Model 2: SALES ~ AGE + EXP
#    Res.Df     RSS Df Sum of Sq      F Pr(>F)
# 1      28 1312.6
# 2      27 1240.2  1    72.465 1.5777 0.2199
```

```
regressionAGEGPA<-lm(formula=SALES~AGE+GPA, data=salesperson)
anova(regressionAGE, regressionAGEGPA)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE
# Model 2: SALES ~ AGE + GPA
#    Res.Df     RSS Df Sum of Sq      F Pr(>F)
# 1      28 1312.6
# 2      27 1280.8  1     31.76 0.6695 0.4204
```

As it can be seen from ANOVA tables, APT has largest F-statistic (smallest p-value). So, current model should be updated by adding APT.

Step 3: Once a new variable is added into your current model, build a reduced model by removing one of the variables which was already in your current model (except the last one added in the previous step). Use the command anova(currentmodel,reducedmodel) to test the significance of the removed variable with an F-test. If the p-value of this test is larger than a sensible significance level (if F-statistic is small then critical Fvalue), then update your current equation by removing this variable. Otherwise, do not touch that variable. Do this for all variables in your current model, except the last variable added in the second step.

```
regressionAPT<-lm(formula=SALES~APT, data=salesperson)
anova(regressionAGEAPT,regressionAPT)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE + APT
# Model 2: SALES ~ APT
#    Res.Df     RSS Df Sum of Sq      F    Pr(>F)
# 1      27  380.42
# 2      28 1963.15 -1   -1582.7 112.33 4.019e-11 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

F-statistic is high than critical F-value and p-value is so close to 0, so it can be understood that AGE is significant and it shouldn’t be removed.

Step 4: Repeat step 2 and 3 until all possible additions are nonsignificant and all possible deletions are significant. (For this question, do not focus on the model assumptions.)

```
regressionAGEAPTANX<-lm(formula=SALES~AGE+APT+ANX, data=salesperson)
anova(regressionAGEAPT,regressionAGEAPTANX)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE + APT
# Model 2: SALES ~ AGE + APT + ANX
#   Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     27 380.42
# 2     26 379.54  1   0.88082 0.0603 0.8079
```

```
regressionAGEAPTEXP<-lm(formula=SALES~AGE+APT+EXP, data=salesperson)
anova(regressionAGEAPT,regressionAGEAPTEXP)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE + APT
# Model 2: SALES ~ AGE + APT + EXP
#   Res.Df    RSS Df Sum of Sq     F Pr(>F)
# 1     27 380.42
# 2     26 380.41  1 0.0048658 3e-04 0.9856
```

```
regressionAGEAPTGPA<-lm(formula=SALES~AGE+APT+GPA, data=salesperson)
anova(regressionAGEAPT,regressionAGEAPTGPA)
# Analysis of Variance Table
#
# Model 1: SALES ~ AGE + APT
# Model 2: SALES ~ AGE + APT + GPA
#   Res.Df    RSS Df Sum of Sq      F Pr(>F)
# 1     27 380.42
# 2     26 378.58  1    1.8408 0.1264  0.725
```

All other possible additions are nonsignificant as it can be seen from the ANOVA tables. So, the model stays with AGE and APT variables.

c) Write down your estimates for the intercept, coefficient(s) for the variables and residual variance.

```
summary(regressionAGEAPT)

# Call:
# lm(formula = SALES ~ AGE + APT, data = salesperson)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -5.4829 -2.3181 -0.6084  1.1793 10.3399
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -86.79154   12.35276  -7.026 1.49e-07 ***
# AGE           5.93145    0.55964  10.599 4.02e-11 ***
# APT           0.19973    0.02456   8.134 9.76e-09 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 3.754 on 27 degrees of freedom
```

```
# Multiple R-squared:  0.8948, Adjusted R-squared:  0.887
# F-statistic: 114.8 on 2 and 27 DF,  p-value: 6.266e-14
```

```
var(regressionAGEAPT$residuals)
# [1] 13.11787
```

Estimated intercept: -86.79154
Estimated AGE coefficient: 5.93145
Estimated APT coefficient: 0.19973
Residual variance: 13.11787

d) Test if high school GPA of a person has an influence on sales value (Use α = 0.05). State H0, H1 and the p-value of the test.

```
regressionGPA<-lm(formula=SALES~GPA, data=salesperson)

summary(regressionGPA)

# Call:
# lm(formula = SALES ~ GPA, data = salesperson)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -19.4307  -7.4343  -0.3644   6.2064  15.8524
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   24.276      7.561   3.211 0.003315 **
# GPA           11.434      2.722   4.201 0.000245 ***
# ---
# Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#
# Residual standard error: 8.901 on 28 degrees of freedom
# Multiple R-squared:  0.3866, Adjusted R-squared:  0.3647
# F-statistic: 17.65 on 1 and 28 DF,  p-value: 0.0002446
```

HO: coefficient of GPA = 0
H1: coefficient of GPA !=0

p-value=0.000245

p-value is so small which means that GPA has a significant effect on SALES.

3)

a) Build a linear regression model that explains the variability in the PROFIT with the observed information SALES. Use any dummy variables if necessary.

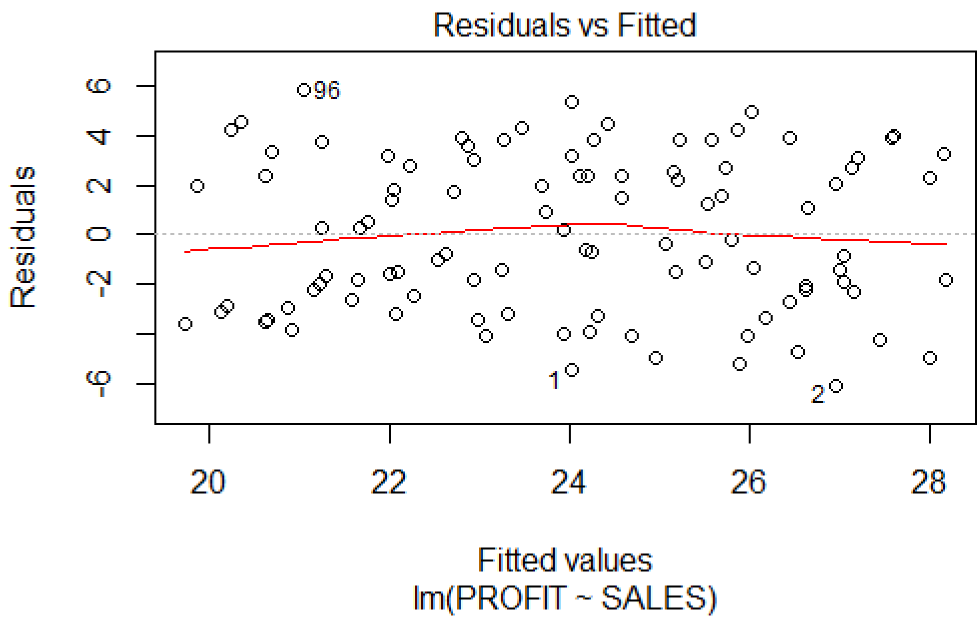
```
sale<-read.table("C:/Users/gülce/Desktop/3-2/IE360/Assignment 2/salesdata.t
xt", header=TRUE )
regsale<-lm(formula=PROFIT~SALES, data=sale)
summary(regsale)

# Call:
```

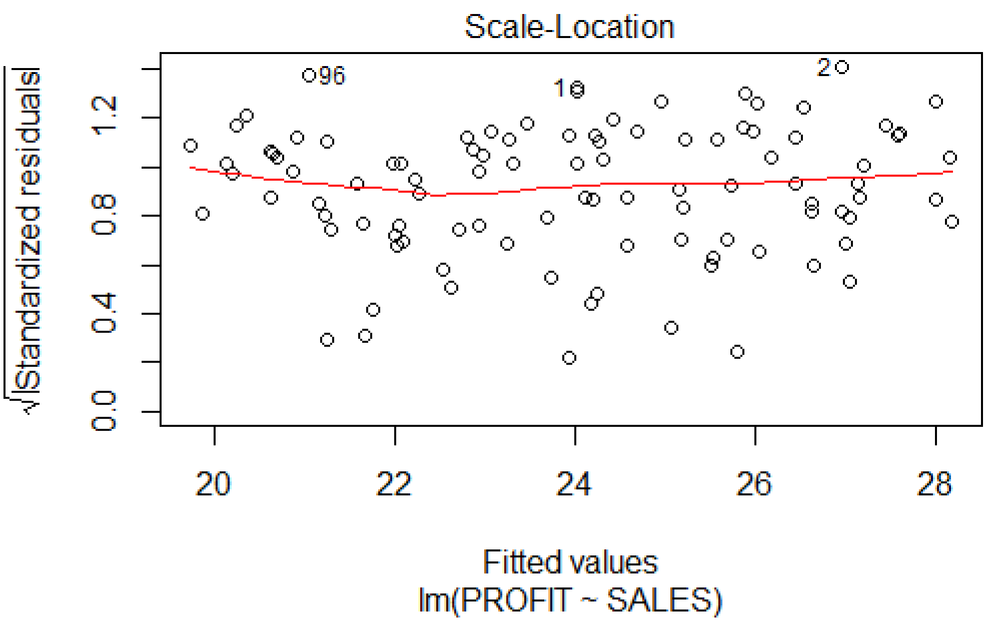
```
# lm(formula = PROFIT ~ SALES, data = sale)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -6.1078 -2.6804 -0.2726  2.7188  5.8453
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) 11.04128    1.77822   6.209 1.29e-08 ***
# SALES        0.43040    0.05813   7.404 4.64e-11 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3.128 on 98 degrees of freedom
# Multiple R-squared:  0.3587, Adjusted R-squared:  0.3522
# F-statistic: 54.82 on 1 and 98 DF,  p-value: 4.637e-11
```

Sales is significant to explain the profit. However, adjusted R-squared is too small. So, sales is not enough to explain the model.

b) Check if the model assumptions are fulfilled or not.

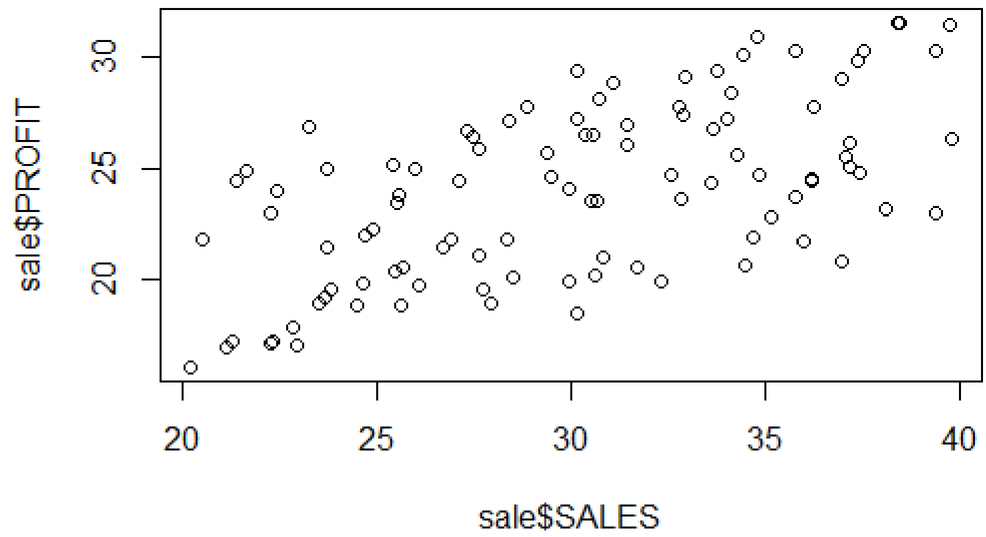
```
plot(regsale)
```






Residuals don't have any pattern.

```
plot(x=sale$SALES,y=sale$PROFIT)
```

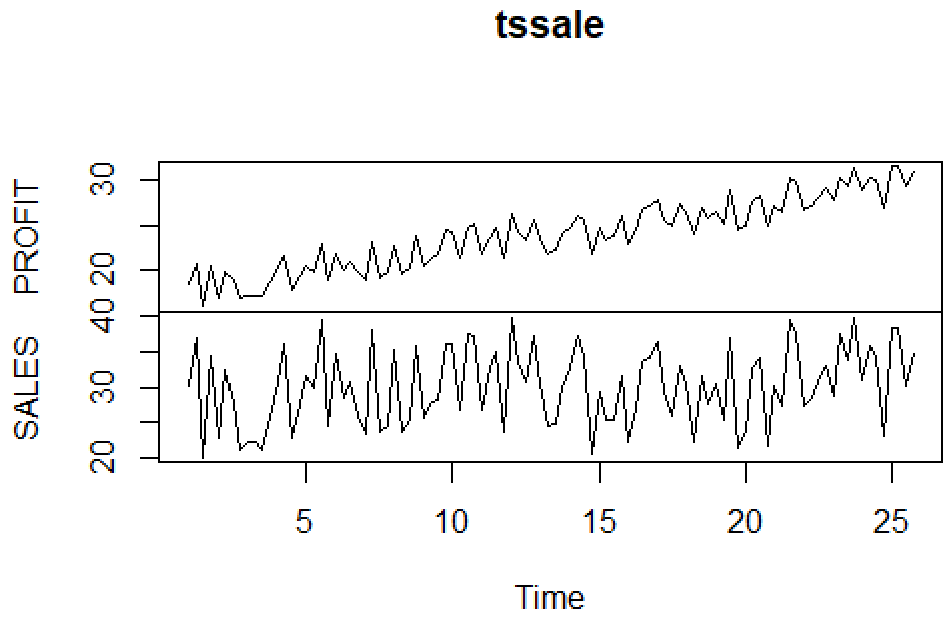



Linearity assumption is violated.

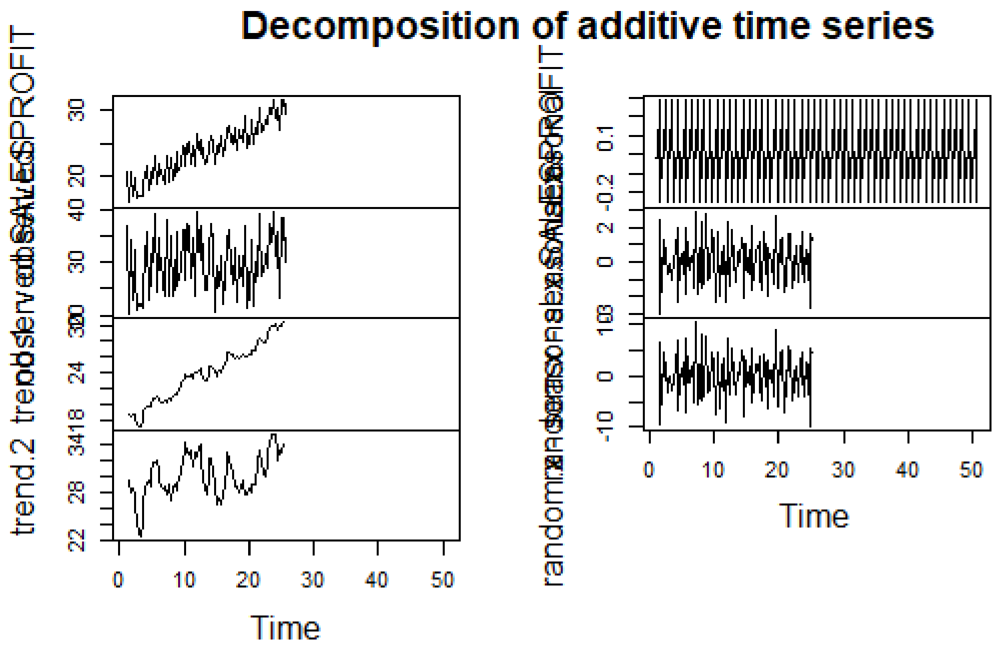
The model is not suitable to explain Profit.

c) Find a way to meet all model assumptions. Build a new model. Again, use any dummy variables if necessary. Check the model assumptions for this new model. Is this new model reliable?

```
tssale<-ts(sale, freq=4)
plot(tssale)
```




```
tssaledecompose<-decompose(tssale)
plot(tssaledecompose)
```



As it can be seen, there is a trend in the profit and sales is weak at explaining it. So, dummy variable should be included to explain the trend in the model.

```
trend<-c(1:100)
```

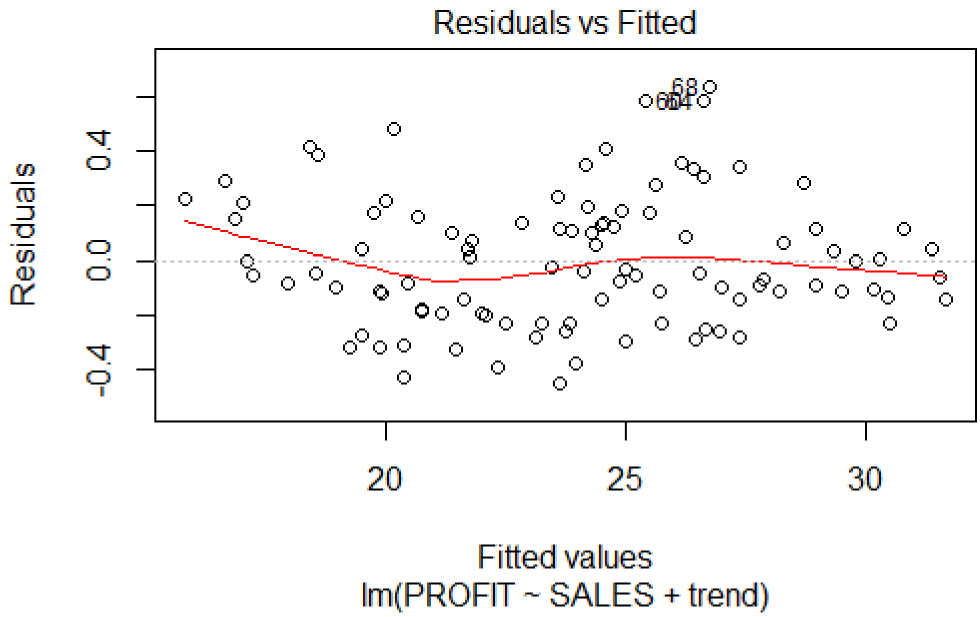
```
regdummy<-lm(formula=PROFIT~SALES+trend, data=sale)
```

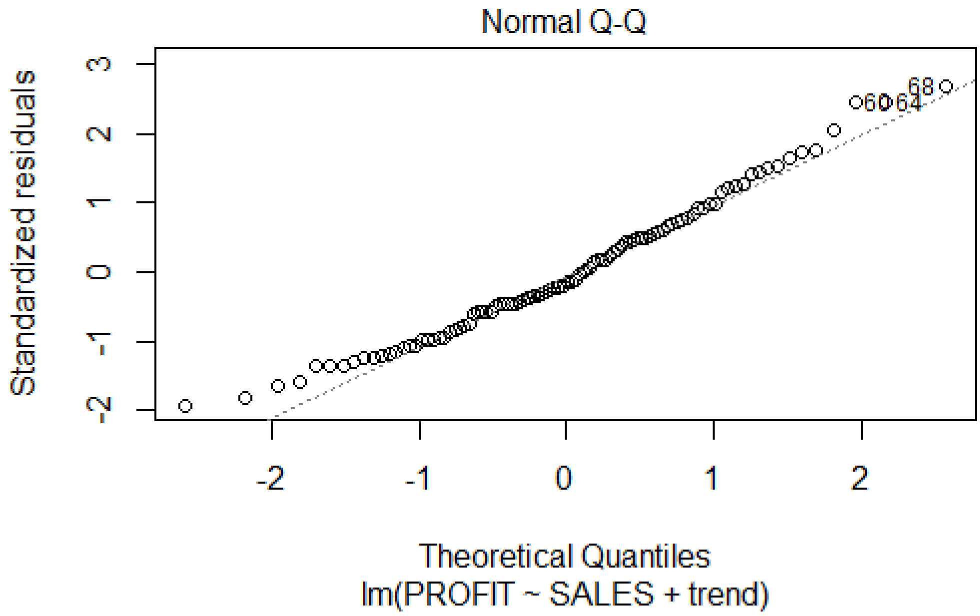
```
summary(regdummy)
```

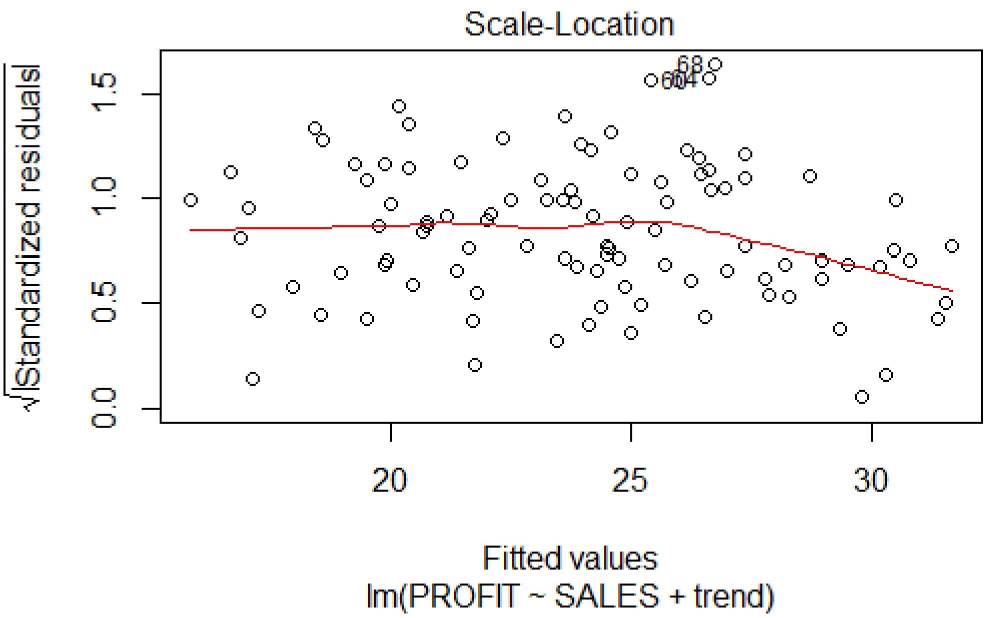
```
# Call:
# lm(formula = PROFIT ~ SALES + trend, data = sale)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -0.44876 -0.17948 -0.04109  0.14313  0.63518
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) 9.6060923  0.1360058   70.63   <2e-16 ***
# SALES       0.2936245  0.0045556   64.45   <2e-16 ***
# trend       0.1099769  0.0008493  129.49   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.2385 on 97 degrees of freedom
# Multiple R-squared:  0.9963, Adjusted R-squared:  0.9962
# F-statistic: 1.31e+04 on 2 and 97 DF,  p-value: < 2.2e-16
```

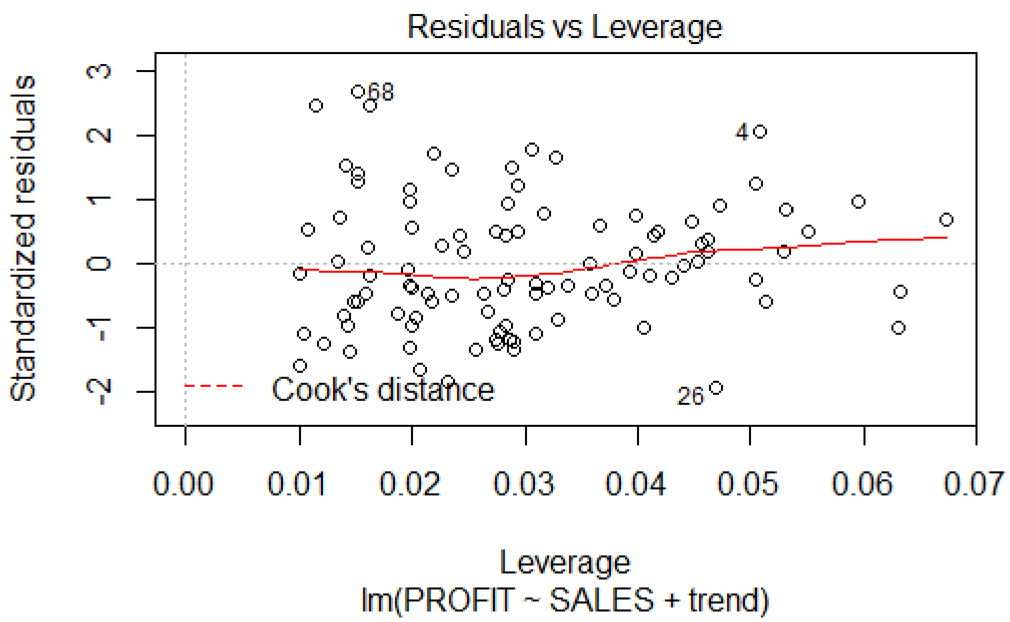
In the new model, Sales and trend has significant and adjusted R-squared is high a lot. So, the model explains the profit very well.

```
plot(regdummy)
```




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(PROFIT ~ SALES + trend)
68
60
64


Scale-Location
√|Standardized residuals|
Fitted values
lm(PROFIT ~ SALES + trend)
68
60
64




There is not any pattern or any problem in the residuals.

d) Your expected sales in the first quarter of 2013 is 30 tons. According to your model in (c), what is your forecast for the profit in this quarter?

```
predict(regdummy,data.frame(SALES=30,trend=101))
#        1
# 29.52249
```