**Homework 2-IE 360**

**Sinan DEMİRHAN-2016402330**

**1) Our aim is to build a simple linear regression model where we explain the response variable Y with the explanatory (prediction) variable X. For this, you are supposed to use the following data:**

**x<-scan()**

**186 119 180 137 152 171 169 131 137 105 191 195 118 117 150 186 183**

**176 160 138 109 108 160 150 127 193 182 114 118 129 170 175 115 166**

**136 144 145 185 146 132 117 148 136 154 157 183 189 133 180 141**

**y<-scan()**

**2397 156 1572 339 600 1364 1317 265 288 121 2101 2038 214 132 601 1729**

**1986 1738 757 280 106 112 637 559 195 2227 1689 93 144 273 849 1364 159**

**953 322 357 391 2273 516 229 162 571 338 594 614 2102 2635 278 1960 374**

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

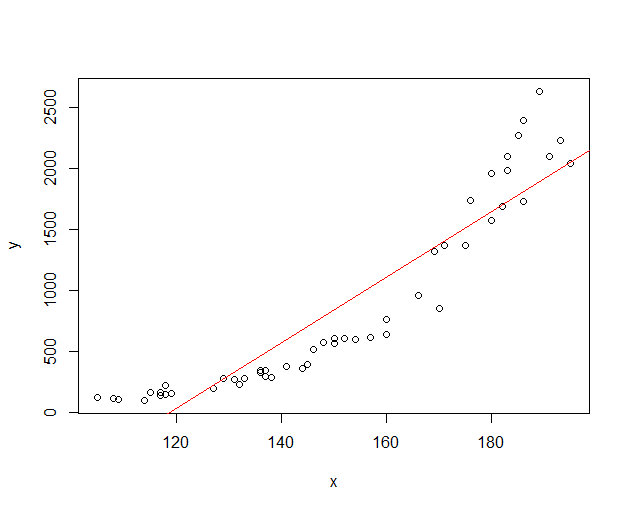
Y = β0+ β1 X+ε

β0 = -3204.156 β1 = 26.949 ε = 296.5

**plot(x,y)**

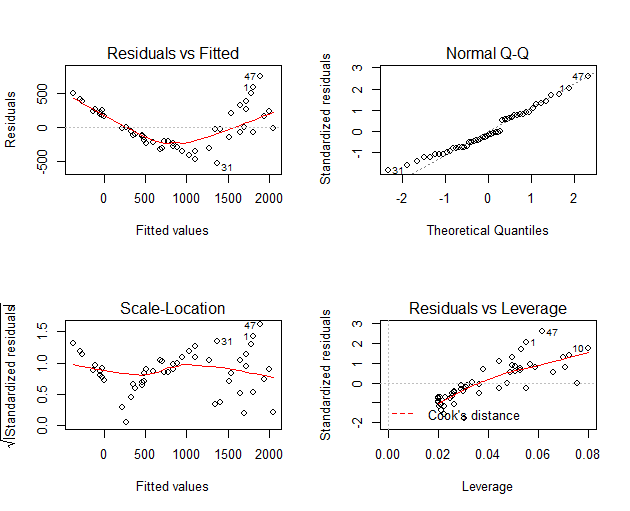
**reg<-lm(y~x)**

**abline(reg,col="red")**



**par(mfrow=c(2,2))**

**plot(reg)**



**gvlma(reg)**

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = reg)

Value p-value Decision

Global Stat 36.3993 2.395e-07 Assumptions NOT satisfied!

Skewness 1.6102 2.045e-01 Assumptions acceptable.

Kurtosis 0.4977 4.805e-01 Assumptions acceptable.

Link Function 33.0563 8.953e-09 Assumptions NOT satisfied!

Heteroscedasticity 1.2351 2.664e-01 Assumptions acceptable.

**residuals versus fitted** **values and Cook's distance plots are not satisfied.**

**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

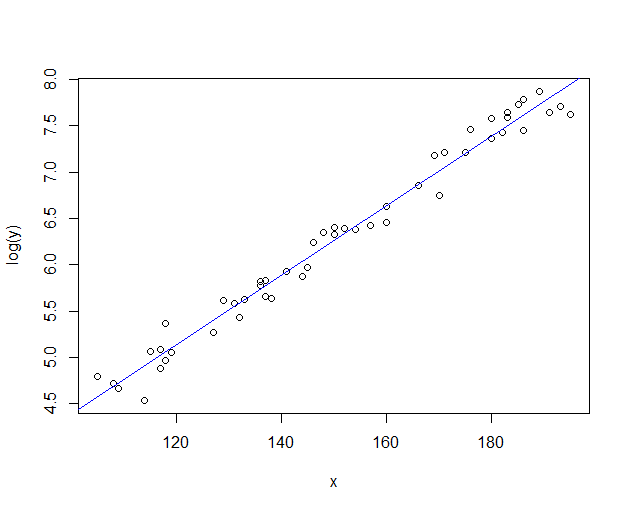
Residual standard error: 296.5 on 48 degrees of freedom

Multiple R-squared: 0.8578, Adjusted R-squared: 0.8548

F-statistic: 289.5 on 1 and 48 DF, p-value: < 2.2e-16

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

**plot(x,log(y))**



**logreg<-lm(log(y)~x)**

**summary(logreg)**

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 \*\*\*

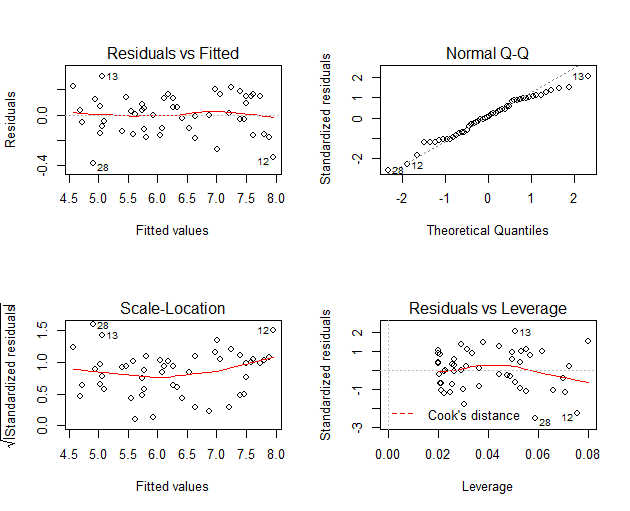
---

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



**gvlma(logreg)**

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = logreg)

Value p-value Decision

Global Stat 2.2065 0.6978 Assumptions acceptable.

Skewness 1.0422 0.3073 Assumptions acceptable.

Kurtosis 0.2773 0.5985 Assumptions acceptable.

Link Function 0.1739 0.6767 Assumptions acceptable.

Heteroscedasticity 0.7131 0.3984 Assumptions acceptable.

By taking logarithm of y values,we obtained a much more satisfied model.Also the new model’s assumptions are much more better and all of them are acceptable.

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

**coefficients(logreg)**

(Intercept) x

0.62498701 0.03760196

β0 = 0.62498701 β1 = 0.03760196

This means that every increasing in x values will be result in incrementation with exp(0.037) in y values.

**confint(logreg,level = 0.95)**

2.5 % 97.5 %

(Intercept) 0.37290407 0.87706996

x 0.03595614 0.03924779

**THE VARİABLES ARE İN THE LOGARİTMİC FORMAT. THEY SHOULD BE CONVERTED BY USİNG EXP() FUNTİON WHEN A PREDİCTİON İS MADE.**

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

**exp(predict(logreg,data.frame(x=125),interval = "confidence"))**

fit lwr upr

1 205.4561 150.0803 281.2641

**exp(predict(logreg,data.frame(x=250),interval = "confidence"))**

fit lwr upr

1 22594.86 15900.85 32106.95

These values are reliable with probability 0.95 between their upper and lower bounds.

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

**exp(predict(logreg,data.frame(x=150),interval="confidence"))**

fit lwr upr

1 525.9897 503.5519 549.4272

**exp(predict(logreg,data.frame(x=150),interval="prediction"))**

fit lwr upr

1 525.9897 385.3353 717.9854

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

Confidence intervals tell you about how well you have determined the mean and prediction intervals tell you where you can expect to see the next data point sampled. The key point is that the prediction interval tells you about the distribution of values, not the uncertainty in determining the population mean.

**2) The file “salesperson.txt” contains a sample data to forecast the SALES (per month) of a person. Using the following variables, we are trying to forecast if a particular applicant will be a good salesperson or not.**

**APT: Selling aptitude test score**

**AGE: Age (in years)**

**ANX: Anxiety test score**

**EXP: Experience (in years)**

**GPA: High school GPA**

**Here, all of these variables may not be needed to forecast sales of a person, so you need to implement stepwise regression to reach a sensible final model.**

**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?**

**cor(salesperson)**

**pairs(salesperson)**

SALES APT AGE ANX EXP GPA

SALES 1.0000000 0.6761204 0.7981406 -0.2958598 0.5498340 0.6217841

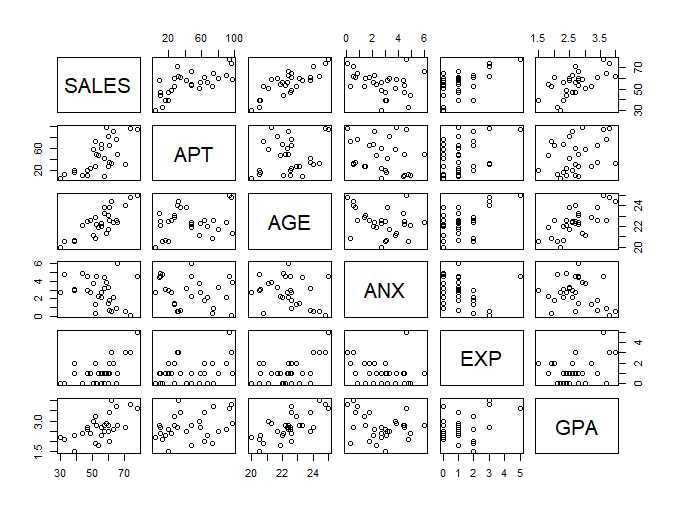
APT 0.6761204 1.0000000 0.2277062 -0.2219880 0.3496392 0.3177716

AGE 0.7981406 0.2277062 1.0000000 -0.2867937 0.5395681 0.6945689

ANX -0.2958598 -0.2219880 -0.2867937 1.0000000 -0.2786892 -0.2443816

EXP 0.5498340 0.3496392 0.5395681 -0.2786892 1.0000000 0.3121288

GPA 0.6217841 0.3177716 0.6945689 -0.2443816 0.3121288 1.0000000



Because of correlations of all values with SALES, AGE and APT variables can be used to predict SALES values.

**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.**

**reg1<-lm(SALES~AGE,data =salesperson)**

**summary(reg1)**

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.**

**summary.aov(lm(SALES~AGE+APT,data =salesperson))**

**Df Sum Sq Mean Sq F value Pr(>F)**

**AGE 1 2303.7 2303.7 163.50 5.72e-13 \*\*\***

**APT 1 932.2 932.2 66.16 9.76e-09 \*\*\***

**Residuals 27 380.4 14.1**

**summary.aov(lm(SALES~AGE+ANX,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 48.032 1.9e-07 \*\*\*

ANX 1 17.7 17.7 0.368 0.549

Residuals 27 1294.9 48.0

**summary.aov(lm(SALES~AGE+EXP,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 50.155 1.29e-07 \*\*\*

EXP 1 72.5 72.5 1.578 0.22

Residuals 27 1240.1 45.9

**summary.aov(lm(SALES~AGE+GPA,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 48.561 1.72e-07 \*\*\*

GPA 1 31.8 31.8 0.669 0.42

Residuals 27 1280.9 47.4

**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 F-value), 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.**

**currentmodel<-lm(SALES~AGE+APT,data =salesperson)**

**reducedmodel<-lm(SALES~APT,data =salesperson)**

**summary(currentmodel)**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 163.50 5.72e-13 \*\*\*

APT 1 932.2 932.2 66.16 9.76e-09 \*\*\*

Residuals 27 380.4 14.1

---

**summary(reducedmodel)**

Df Sum Sq Mean Sq F value Pr(>F)

APT 1 1653 1653.2 23.58 4.11e-05 \*\*\*

Residuals 28 1963 70.1

66>23 : So do not touch currrent model.

**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.)**

**summary.aov(lm(SALES~AGE+APT,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 163.50 5.72e-13 \*\*\*

APT 1 932.2 932.2 66.16 9.76e-09 \*\*\*

Residuals 27 380.4 14.1

**summary.aov(lm(SALES~AGE+APT+GPA,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 158.213 1.47e-12 \*\*\*

APT 1 932.2 932.2 64.021 1.76e-08 \*\*\*

GPA 1 1.8 1.8 0.126 0.725

Residuals 26 378.6 14.6

**summary.aov(lm(SALES~AGE+APT+ANX,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 157.81 1.51e-12 \*\*\*

APT 1 932.2 932.2 63.86 1.81e-08 \*\*\*

ANX 1 0.9 0.9 0.06 0.808

Residuals 26 379.5 14.6

**summary.aov(lm(SALES~AGE+APT+EXP,data =salesperson))**

Df Sum Sq Mean Sq F value Pr(>F)

AGE 1 2303.7 2303.7 157.45 1.55e-12 \*\*\*

APT 1 932.2 932.2 63.71 1.85e-08 \*\*\*

EXP 1 0.0 0.0 0.00 0.986

Residuals 26 380.4 14.6

There is no improvement in the F-values,so current model become the best model for this reggression analysis.

Checking the correctness of the result can also be done by using step function in R.

**Fitall<-lm(SALES~.,data = salesperson)**

**Fitstart<-lm(SALES~AGE,data = salesperson)**

**step(Fitstart,direction = "forward",scope = formula(Fitall))**

Start: AIC=117.36

SALES ~ AGE

Df Sum of Sq RSS AIC

+ APT 1 932.20 380.42 82.202

<none> 1312.61 117.357

+ EXP 1 72.46 1240.15 117.654

+ GPA 1 31.76 1280.85 118.623

+ ANX 1 17.67 1294.95 118.951

Step: AIC=82.2

SALES ~ AGE + APT

Df Sum of Sq RSS AIC

<none> 380.42 82.202

+ GPA 1 1.84081 378.58 84.057

+ ANX 1 0.88082 379.54 84.133

+ EXP 1 0.00487 380.41 84.202

Call:

lm(formula = SALES ~ AGE + APT, data = salesperson)

Coefficients:

(Intercept) AGE APT

-86.7915 5.9314 0.1997

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

reglast<-lm(SALES~AGE+APT,data =salesperson)

summary(reglast)

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

**Estimates**

**(Intercept) -86.79154**

**AGE 5.93145**

**APT 0.19973**

**Residual Variance = sqrt(3.754) =14.0925**

**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.**

**summary(lm(SALES~GPA,data = salesperson))**

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

H0 : Estimation for GPA =0

H1 : Estimation for GPA **≠ 0**

p-value for GPA = 0.000245 < α = 0.05

Because p-value is less than 0.05, we can say that GPA of a person has an influence on sales value.

**3. The file “salesdata.txt" contains quarterly PROFIT of a company (in thousand dollars) with the quarterly SALES of their product (in tons) starting from the first quarter of 1988.**

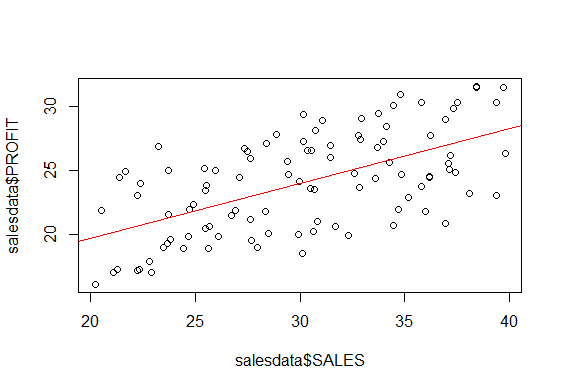
**a) Build a linear regression model that explains the variability in the PROFIT with the observed information SALES. Use any dummy variables if necessary.**

**salesdata**

**sales<-lm(PROFİT~SALES,data=salesdata)**

**plot(salesdata$SALES,salesdata$PROFIT)**

**abline(sales,col="red")**



**summary(sales)**

Call:

lm(formula = PROFIT ~ SALES, data = salesdata)

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

There are high variability in the predictions by only looking at the SALES data.

So we should use dummy variables in predictions to increase the R squared value.

**len<-length(salesdata[,1])**

**newsales<-data.frame(PROFIT=salesdata$PROFIT,SALES=salesdata$SALES,TIME=1:len)**

**newsales[1:4,]**

PROFIT SALES TIME

1 18.51422 30.12393 1

2 20.83479 36.94505 2

3 16.09410 20.20793 3

4 20.64832 34.47554 4

**regsales<-lm(PROFİT~SALES+TIME,data=newsales)**

**summary(regsales)**

Call:

lm(formula = PROFIT ~ SALES + TIME, data = newsales)

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 \*\*\*

TIME 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

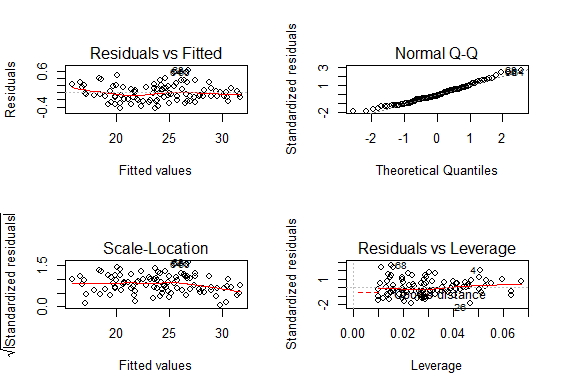
Multiple R-squared: 0.9963, Adjusted R-squared: 0.9962

F-statistic: 1.31e+04 on 2 and 97 DF, p-value: < 2.2e-16

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

**par(mfrow=c(2,2))**

**plot(regsales)**



All model assumptions are fulfilled.We can see this by using gvlma function with level of significance 0.05.

**gvlma(regsales)**

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = regsales)

Value p-value Decision

Global Stat 5.4956 0.24012 Assumptions acceptable.

Skewness 3.7924 0.05149 Assumptions acceptable.

Kurtosis 0.1614 0.68785 Assumptions acceptable.

Link Function 0.3692 0.54344 Assumptions acceptable.

Heteroscedasticity 1.1725 0.27888 Assumptions acceptable.

**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?**

We can add the model the seosanality effect with quarters.

**x1<-1:len%%4 ==1**

**x2<-1:len%%4 ==2**

**x3<-1:len%%4 ==3**

**x4<-1:len%%4 ==0**

**powerdata<-data.frame(PROFIT=salesdata$PROFIT,SALES=salesdata$SALES,TIME=1:len,S1=x1\*1,S2=x2\*1,S3=x3\*1,S4=x4\*1)**

**powerdata[1:8,]**

PROFIT SALES TIME S1 S2 S3 S4

1 18.51422 30.12393 1 1 0 0 0

2 20.83479 36.94505 2 0 1 0 0

3 16.09410 20.20793 3 0 0 1 0

4 20.64832 34.47554 4 0 0 0 1

**fit<-lm(PROFIT~SALES+TIME+S2+S3+S4,data=powerdata)**

**summary(fit)**

Call:

lm(formula = PROFIT ~ SALES + TIME + S1 + S2 + S3 + S4, data = powerdata)

Residuals:

Min 1Q Median 3Q Max

-0.39727 -0.14822 -0.02437 0.17304 0.47302

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.7420639 0.1256262 77.548 < 2e-16 \*\*\*

SALES 0.2947736 0.0040791 72.264 < 2e-16 \*\*\*

TIME 0.1098060 0.0007599 144.507 < 2e-16 \*\*\*

S1 -0.2114063 0.0603577 -3.503 0.000707 \*\*\*

S2 -0.3050027 0.0604004 -5.050 2.16e-06 \*\*\*

S3 -0.1313686 0.0603554 -2.177 0.032017 \*

S4 NA NA NA NA

---

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

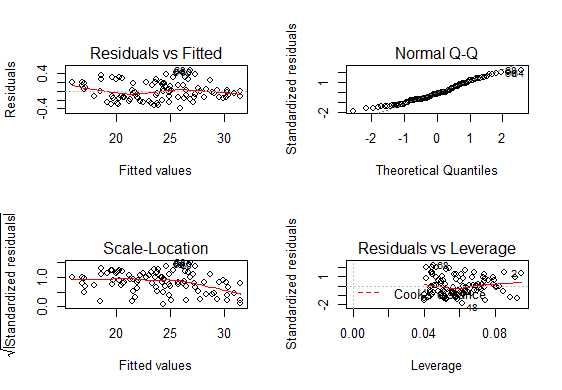
Residual standard error: 0.2131 on 94 degrees of freedom

Multiple R-squared: 0.9971, Adjusted R-squared: 0.997

F-statistic: 6567 on 5 and 94 DF, p-value: < 2.2e-16

**par(mfrow=c(2,2))**

**plot(fit)**



Again,all model assumptions are fulfilled and we can see this by using gvlma function with level of significance 0.05.

**gvlma(fit)**

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS

USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:

Level of Significance = 0.05

Call:

gvlma(x = fit)

Value p-value Decision

Global Stat 6.372 0.1730 Assumptions acceptable.

Skewness 2.177 0.1401 Assumptions acceptable.

Kurtosis 2.459 0.1169 Assumptions acceptable.

Link Function 0.131 0.7174 Assumptions acceptable.

Heteroscedasticity 1.606 0.2051 Assumptions acceptable.

**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?**

**Predictions by using founded two models;**

**predict(regsales,data.frame(SALES=30,TIME=101),interval = "prediction")**

fit lwr upr

1 29.52249 29.03921 30.00577

**predict(fit,data.frame(SALES=30,TIME=101,S1=1,S2=0,S3=0,S4=0),interval = "prediction")**

fit lwr upr

1 29.46427 29.02564 29.90291