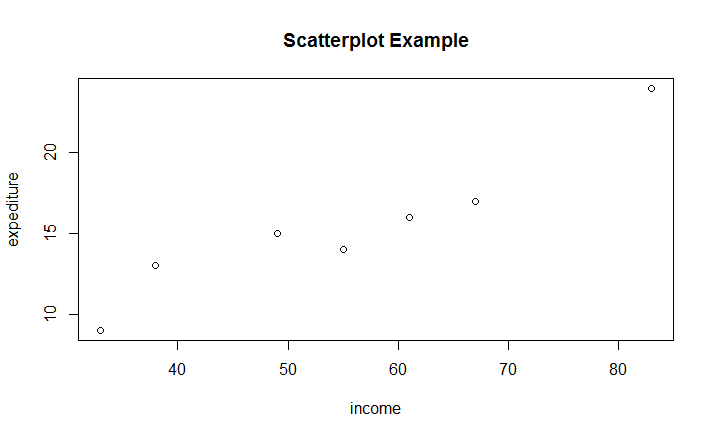
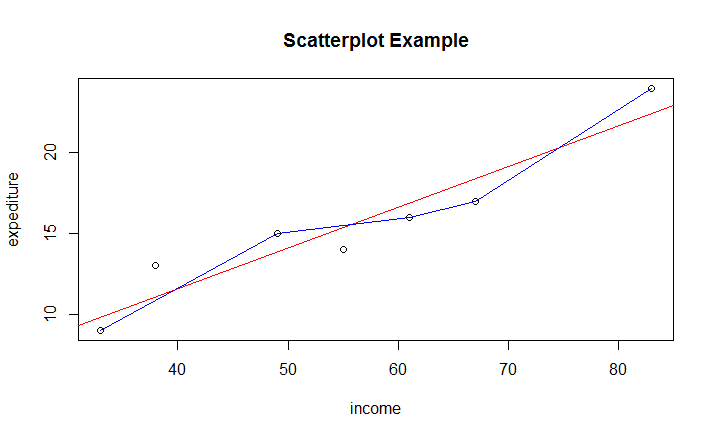
**SIMPLE LINEAR REGRESSION**

|  |
| --- |
| > setwd("C:/Mihaela/R Scripts")  > income.data <- read.csv("income.csv", header=TRUE)  > income.data  income expediture  1 55 14  2 83 24  3 38 13  4 61 16  5 33 9  6 49 15  7 67 17  >plot(income,expediture, main="Scatterplot Example") |
|  |
| |  | | --- | |  | |



> abline(lm(expediture~income), col="red") # regression line (y~x)

> lines(lowess(expediture~income), col="blue") # lowess line (x,y)



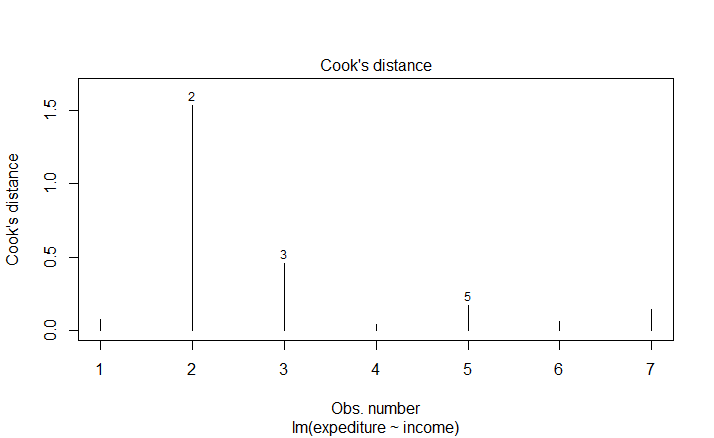
|  |
| --- |
| > library(car)  > fit <- lm(expediture~income) # Fit regression model  > summary(fit)  Call:  lm(formula = expediture ~ income)  Residuals:  1 2 3 4 5 6 7  -1.3925 1.5387 1.8993 -0.9073 -0.8384 1.1222 -1.4220  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1.50733 2.17424 0.693 0.51902  income 0.25246 0.03788 6.664 0.00115 \*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.595 on 5 degrees of freedom  Multiple R-squared: 0.8988, Adjusted R-squared: 0.8786  F-statistic: 44.41 on 1 and 5 DF, p-value: 0.001149  > anova(fit) # anova table  Analysis of Variance Table  Response: expediture  Df Sum Sq Mean Sq F value Pr(>F)  income 1 112.993 112.993 44.41 0.001149 \*\*  Residuals 5 12.721 2.544  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  > |
|  |
| |  | | --- | |  | |

> # Cook's D plot

> # identify D values > 4/(n-k-1)

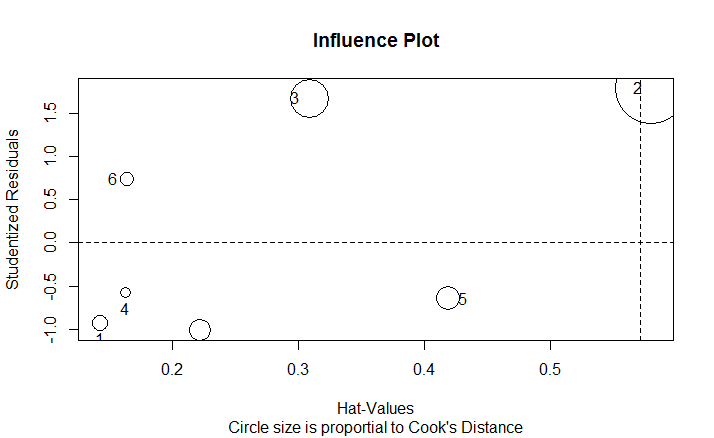
> cutoff <- 4/((nrow(income.data)-length(fit$coefficients)-2))

> plot(fit, which=4, cook.levels=cutoff)



# Influence Plot

influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )

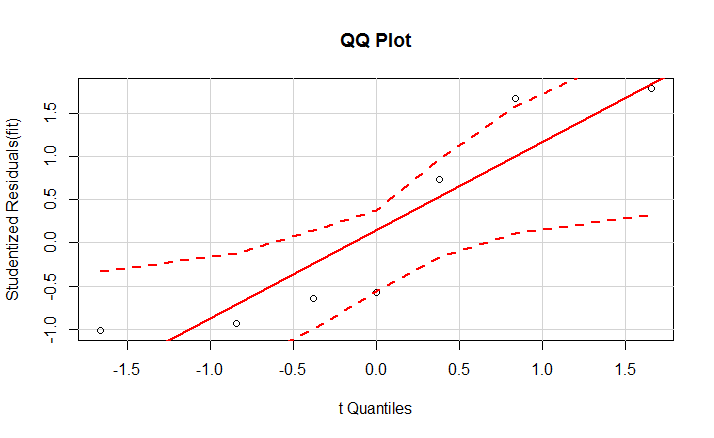


> # Studentizided residuals are the residuals divided by their estimated standard deviation as a way to standardized

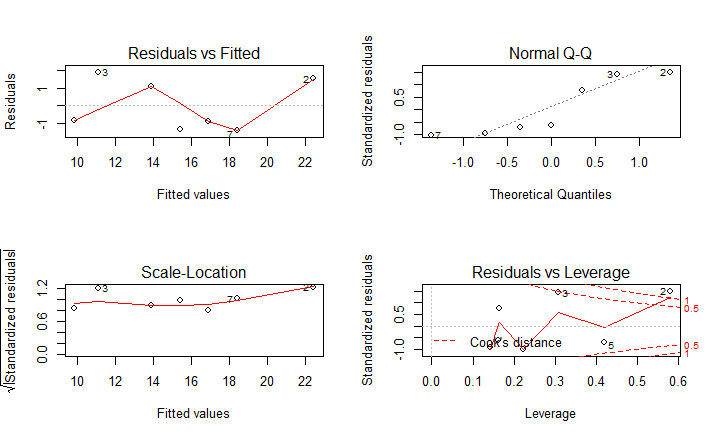
> # Normality of Residuals

> # qq plot for studentized resid

> qqPlot(fit, main="QQ Plot")



|  |
| --- |
| > par(mfrow=c(2,2)) # Change the panel layout to 2 x 2  > plot(fit)  > par(mfrow=c(1,1)) # Change back to 1 x 1  > par(mfrow=c(1,1)) # Change back to 1 x 1 |
|  |
| |  | | --- | | > | |



**Residuals vs Fitted**: If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don’t have non-linear relationship.

**Normal QQ Plot:** This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It’s good if residuals are lined well on the straight dashed line.

**Scale-Location:** This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It’s good if you see a horizontal line with equally (randomly) spread points.

**Residuals vs Leverage:** This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn’t be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don’t really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values.

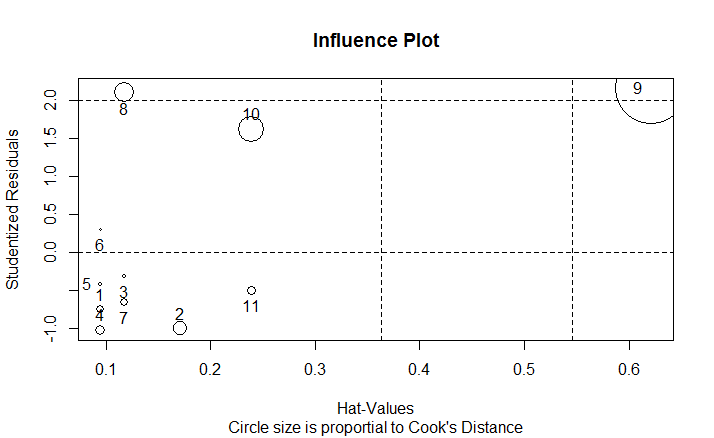
|  |
| --- |
| > # Normality of Residuals  > # qq plot for studentized resid  > # Look for the tails, points should be close to the line or within the confidence intervals.  > qqPlot(fit, main="QQ Plot")  >  > #Test for independence of residuals  > #Null hypothesis that the errors are independent  > library(lmtest)  > dwtest(fit)  Durbin-Watson test  data: fit  DW = 2.1162, p-value = 0.6503  alternative hypothesis: true autocorrelation is greater than 0  >  >  >  > expediture #response variable  [1] 14 24 13 16 9 15 17  > predict(fit)#Predicted values  1 2 3 4 5 6 7  15.392506 22.461322 11.100725 16.907252 9.838437 13.877760 18.421998  > residuals(fit) #Residuals  1 2 3 4 5 6 7  -1.3925060 1.5386785 1.8992748 -0.9072522 -0.8384367 1.1222401 -1.4219984  > stdres(fit) # Standardized residuals  1 2 3 4 5 6 7  -0.9429537 1.4894995 1.4320132 -0.6214080 -0.6898524 0.7695493 -1.0108130  >  >  > # A common rule is to flag any observation whose leverage value, hii, is more than 2 or 3 times larger than the mean leverage value  > mn <-mean(hatvalues(fit))  > mn  [1] 0.2857143  > hatvalues(fit)  1 2 3 4 5 6 7  0.1428687 0.5805802 0.3086221 0.1622079 0.4194198 0.1641418 0.2221595  > hatvalues(fit)/mn  1 2 3 4 5 6 7  0.5000403 2.0320306 1.0801773 0.5677276 1.4679694 0.5744964 0.7775584  >  >  > #Cook’s distance (D) is used as a measure of the influence of a data point in a regression model.  > #The consensus seems to be that a Di value of more that 1 indicates an influential value  > cooks.distance(fit)  1 2 3 4 5 6 7  0.07410377 1.53555010 0.45769431 0.03738174 0.17189731 0.05814730 0.14591034  >  >  >  > estimates = predict(fit, se.fit=TRUE,interval="confidence")  > estimates  $fit  fit lwr upr  1 15.392506 13.842680 16.94233  2 22.461322 19.337073 25.58557  3 11.100725 8.822860 13.37859  4 16.907252 15.255859 18.55865  5 9.838437 7.182982 12.49389  6 13.877760 12.216552 15.53897  7 18.421998 16.489376 20.35462  $se.fit  [1] 0.6029087 1.2153857 0.8861282 0.6424200 1.0330171 0.6462383 0.7518229  $df  [1] 5  $residual.scale  [1] 1.595082  >  >  > newdata = data.frame(income=58)  > newdata  income  1 58  > predict(fit, newdata, interval="predict")  fit lwr upr  1 16.14988 11.75766 20.54209  > #The 95% prediction interval of the expediture for income 58 is between 11.75766 and 20.54209.  >  >  > predict(fit, newdata, interval="confidence")  fit lwr upr  1 16.14988 14.57534 17.72442  > #The 95% confidence interval of the mean expediture for income 58 is between 14.57534 17.72442. |
|  |
| |  | | --- | | > | |

EXAMPLE 2

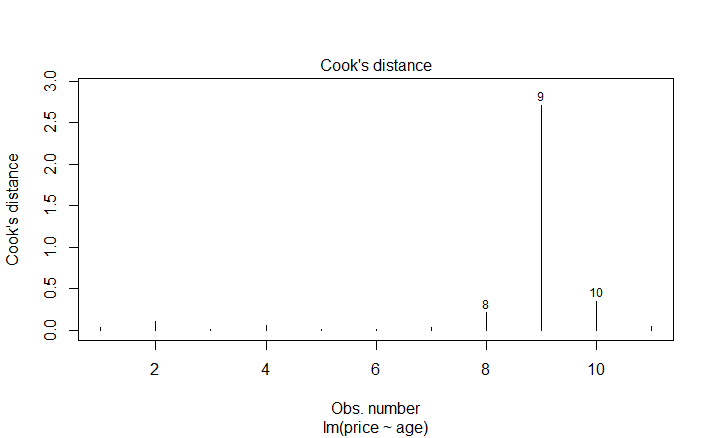
|  |
| --- |
| > setwd("C:/Mihaela/R Scripts")  >  > orion.data <- read.csv("orions.csv", header=TRUE)  > orion.data  age price  1 5 85  2 4 103  3 6 70  4 5 82  5 5 89  6 5 98  7 6 66  8 6 95  9 2 169  10 7 70  11 7 48  >  > attach(orion.data)  > plot(age,price, main="Scatterplot Example")  > abline(lm(price~age), col="red") # regression line (y~x)      > library(car)  > fit1 <- lm(price~age) # Fit regression model  > summary(fit1)  Call:  lm(formula = price ~ age)  Residuals:  Min 1Q Median 3Q Max  -12.162 -8.531 -5.162 8.946 21.099  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 195.47 15.24 12.826 4.36e-07 \*\*\*  age -20.26 2.80 -7.237 4.88e-05 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 12.58 on 9 degrees of freedom  Multiple R-squared: 0.8534, Adjusted R-squared: 0.8371  F-statistic: 52.38 on 1 and 9 DF, p-value: 4.882e-05  > anova(fit1) # anova table  Analysis of Variance Table  Response: price  Df Sum Sq Mean Sq F value Pr(>F)  age 1 8285.0 8285.0 52.38 4.882e-05 \*\*\*  Residuals 9 1423.5 158.2  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | | > | |

> # Influence Plot

> influencePlot(fit1, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook's Distance" )



|  |
| --- |
| > # Cook's D plot  > # Cook's distance measures how much an observation influences the overall model or predicted values  > # identify D values > 4/(n-k-1)  > cutoff <- 4/((nrow(orion.data)-length(fit$coefficients)-2))  > plot(fit1, which=4, cook.levels=cutoff) |
|  |
| |  | | --- | | > | |



> #Run regression without influential point, observation 9

> fit2 <- lm(price~age, data=orion.data[-9,])

> summary(fit2)

Call:

lm(formula = price ~ age, data = orion.data[-9, ])

Residuals:

Min 1Q Median 3Q Max

-12.667 -6.583 -2.262 6.607 20.095

Coefficients:

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

(Intercept) 160.333 20.784 7.714 5.67e-05 \*\*\*

age -14.238 3.663 -3.887 0.00463 \*\*

---

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

Residual standard error: 10.62 on 8 degrees of freedom

Multiple R-squared: 0.6538, Adjusted R-squared: 0.6106

F-statistic: 15.11 on 1 and 8 DF, p-value: 0.004627

>

> # Studentizided residuals are the residuals divided by their estimated standard deviation as a way to standardized

> # distribution of studentized residuals

> library(MASS)

> sresid <- studres(fit1)

>

>

>

> #Diagnostic Plots

> par(mfrow=c(2,2)) # Change the panel layout to 2 x 2

> plot(fit1)

> par(mfrow=c(1,1)) # Change back to 1 x 1

>

>

>

> # Normality of Residuals

> # qq plot for studentized resid

> # Look for the tails, points should be close to the line or within the confidence intervals.

> qqPlot(fit1, main="QQ Plot")

>

> #Test for independence of residuals

> #Null hypothesis that the errors are independent

> library(lmtest)

> dwtest(fit1)

Durbin-Watson test

data: fit1

DW = 1.2488, p-value = 0.1049

alternative hypothesis: true autocorrelation is greater than 0

>

>

>

> expediture #response variable

[1] 14 24 13 16 9 15 17

> predict(fit1)#Predicted values

1 2 3 4 5 6 7 8 9 10

94.16216 114.42342 73.90090 94.16216 94.16216 94.16216 73.90090 73.90090 154.94595 53.63964

11

53.63964

> residuals(fit1) #Residuals

1 2 3 4 5 6 7 8 9

-9.162162 -11.423423 -3.900901 -12.162162 -5.162162 3.837838 -7.900901 21.099099 14.054054

10 11

16.360360 -5.639640

> stdres(fit1) # Standardized residuals

1 2 3 4 5 6 7 8 9

-0.7656215 -0.9977036 -0.3301042 -1.0163118 -0.4313679 0.3207028 -0.6685945 1.7854598 1.8166705

10 11

1.4909522 -0.5139516

>

|  |
| --- |
| > # A common rule is to flag any observation whose leverage value, hii, is more than 2 or 3 times larger than the mean leverage value  > mn <-mean(hatvalues(fit1))  > mn  [1] 0.1818182  > hatvalues(fit1)  1 2 3 4 5 6 7 8 9  0.09459459 0.17117117 0.11711712 0.09459459 0.09459459 0.09459459 0.11711712 0.11711712 0.62162162  10 11  0.23873874 0.23873874  > hatvalues(fit1)/mn  1 2 3 4 5 6 7 8 9 10  0.5202703 0.9414414 0.6441441 0.5202703 0.5202703 0.5202703 0.6441441 0.6441441 3.4189189 1.3130631  11  1.3130631 |
|  |
| |  | | --- | |  | |

>

> #Cook’s distance (D) is used as a measure of the influence of a data point in a regression model.

> #The consensus seems to be that a Di value of more that 1 indicates an influential value

> cooks.distance(fit1)

1 2 3 4 5 6 7 8

0.030621152 0.102787167 0.007227522 0.053956921 0.009720506 0.005372777 0.029649190 0.211440134

9 10 11

2.710954012 0.348567256 0.041419377

>

> # Assessing Outliers

> outlierTest(fit1) # Bonferonni p-value for most extreme obs

No Studentized residuals with Bonferonni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferonni p

9 2.152261 0.063551 0.69906

> # The linear correlation coefficient, r = −0.92378, suggests a

strong negative linear correlation between age and price of Orions

> cor(orion.data, use="complete.obs", method="pearson")

age price

age 1.0000000 -0.9237821

price -0.9237821 1.0000000

**> ### Analysis without observation 9**

>

> #Run regression without influential point, observation 9

> fit2 <- lm(price~age, data=orion.data[-9,])

> summary(fit2)

Call:

lm(formula = price ~ age, data = orion.data[-9, ])

Residuals:

Min 1Q Median 3Q Max

-12.667 -6.583 -2.262 6.607 20.095

Coefficients:

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

(Intercept) 160.333 20.784 7.714 5.67e-05 \*\*\*

age -14.238 3.663 -3.887 0.00463 \*\*

---

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

Residual standard error: 10.62 on 8 degrees of freedom

Multiple R-squared: 0.6538, Adjusted R-squared: 0.6106

F-statistic: 15.11 on 1 and 8 DF, p-value: 0.004627

>

>

> # Studentizided residuals are the residuals divided by their estimated standard deviation as a way to standardized

> # distribution of studentized residuals

> library(MASS)

> sresid <- studres(fit2)

>

>

>

> #Diagnostic Plots

> par(mfrow=c(2,2)) # Change the panel layout to 2 x 2

> plot(fit2)

> par(mfrow=c(1,1)) # Change back to 1 x 1

>

>

>

> # Normality of Residuals

> # qq plot for studentized resid

> # Look for the tails, points should be close to the line or within the confidence intervals.

> qqPlot(fit2, main="QQ Plot")

>

> #Test for independence of residuals

> #Null hypothesis that the errors are independent

> library(lmtest)

> dwtest(fit2)

Durbin-Watson test

data: fit2

DW = 2.1363, p-value = 0.4975

alternative hypothesis: true autocorrelation is greater than 0

>

>

>

> # A common rule is to flag any observation whose leverage value, hii, is more than 2 or 3 times larger than the mean leverage value

> mn <-mean(hatvalues(fit2))

> mn

[1] 0.2

> hatvalues(fit2)

1 2 3 4 5 6 7 8 10 11

0.1428571 0.4047619 0.1190476 0.1428571 0.1428571 0.1428571 0.1190476 0.1190476 0.3333333 0.3333333

> hatvalues(fit2)/mn

1 2 3 4 5 6 7 8 10 11

0.7142857 2.0238095 0.5952381 0.7142857 0.7142857 0.7142857 0.5952381 0.5952381 1.6666667 1.6666667

>

>

> #Cook’s distance (D) is used as a measure of the influence of a data point in a regression model.

> #The consensus seems to be that a Di value of more that 1 indicates an influential value

> cooks.distance(fit2)

1 2 3 4 5 6 7

1.480738e-02 7.356011e-04 1.637324e-02 4.401718e-02 1.760687e-05 6.768082e-02 5.396886e-02

8 10 11

2.748432e-01 2.898796e-01 5.339108e-01

>

> # Assessing Outliers

> outlierTest(fit2) # Bonferonni p-value for most extreme obs

No Studentized residuals with Bonferonni p < 0.05

Largest |rstudent|:

rstudent unadjusted p-value Bonferonni p

8 2.690902 0.031045 0.31045

>

> cor(orion.data[-9,], use="complete.obs", method="pearson")

age price

age 1.0000000 -0.8086074

price -0.8086074 1.0000000