Statistics for Engineers (MAT2001)- Lab Experiment-III: Correlation and Regression

1 Regression

Problem:
The following table shows the scores (X) of 10 students on Zoology test and scores (Y) on Botony test .The maximum score in each test was 50.Obtain least square equation of line of regression of X on Y. If it is known that the score of a student in Botony is 28,Estimate his/her score in Zoology.

X	34	37	36	32	32	36	35	34	29
Y	37	37	34	34	33	40	39	37	36

```
x=c(34,37,36,32,32,36,35,34,29,35)
y=c(37,37,34,34,33,40,39,37,36,35)
fit=lm(x~y)
fit

##
## Call:
## lm(formula = x ~ y)
##
## Coefficients:
## (Intercept)            y
##     18.9167       0.4167
```

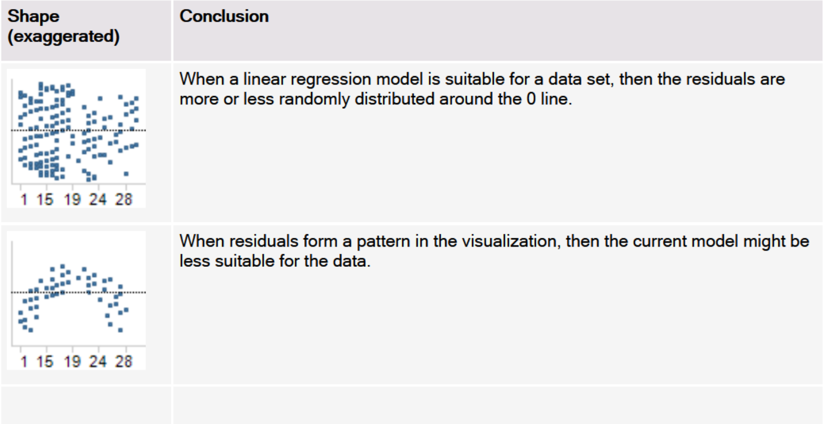
The equation of the line of regression of X and Y is X=18.9167+0.4167Y. The required score of the student in Zoology is 30.58333

```
summary(fit)

##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.9167 -0.5833 -0.2500  1.2292  2.9167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.9167    12.6059   1.501    0.172
## y             0.4167     0.3476   1.199    0.265
##
## Residual standard error: 2.347 on 8 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.04627
## F-statistic: 1.437 on 1 and 8 DF,  p-value: 0.265
```

1.1 Diagnostic Visualizations

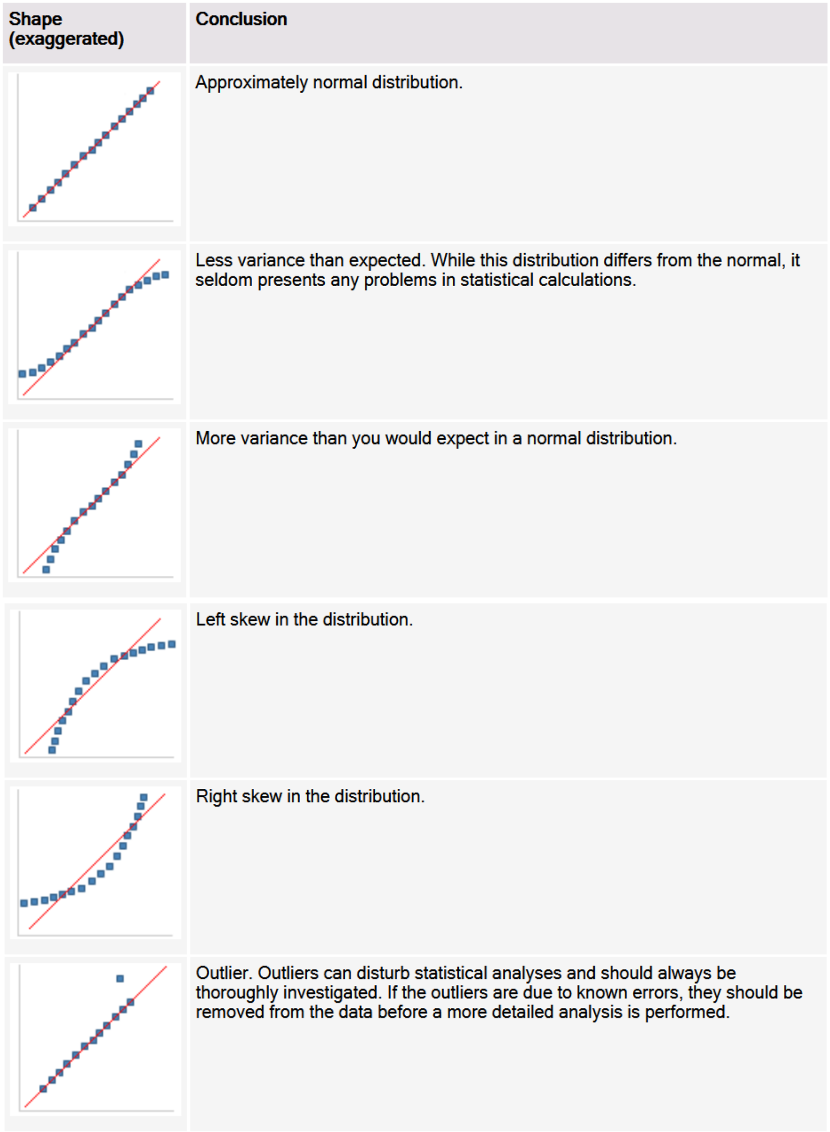
Residuals vs. Fitted: The residuals vs. fitted visualization is a scatter plot showing the residuals on the Y-axis and the fitted values on the X-axis. You can compare it to doing a linear fit and then flipping the fitted line so that it becomes horizontal. Values that have the residual 0 are those that would end up directly on the estimated regression line. The residuals vs fit plot is commonly used to detect non-linearity, unequal error variances and outliers.


1 15 19 24 28
1 15 19 24 28

Shape (exaggerated)	Conclusion
	When a linear regression model is suitable for a data set, then the residuals are more or less randomly distributed around the 0 line.
	When residuals form a pattern in the visualization, then the current model might be less suitable for the data.

Normal Quantile-Quantile

The normal quantile-quantile visualization calculates the normal quantiles of all values in a column. The values (Y-axis) are then plotted against the normal quantiles (X-axis).



Shape (exaggerated)	Conclusion
	Approximately normal distribution.
	Less variance than expected. While this distribution differs from the normal, it seldom presents any problems in statistical calculations.
	More variance than you would expect in a normal distribution.
	Left skew in the distribution.
	Right skew in the distribution.
	Outlier. Outliers can disturb statistical analyses and should always be thoroughly investigated. If the outliers are due to known errors, they should be removed from the data before a more detailed analysis is performed.

Scale – Location: The scale – location plot is similar to the residuals vs fit plot, but instead of linear residuals it uses the square root of the residuals. It is used to reveal trends in the magnitudes of residuals. For a good model, the values should be more or less randomly distributed.

Cook's Distance: Cook's distance is a statistic which tries to identify those values which have more influence than others on the estimated coefficients.

Problem:

The following data pertain to the resistance in (ohms) and the failure times (minutes) of 24 overloaded resistors.

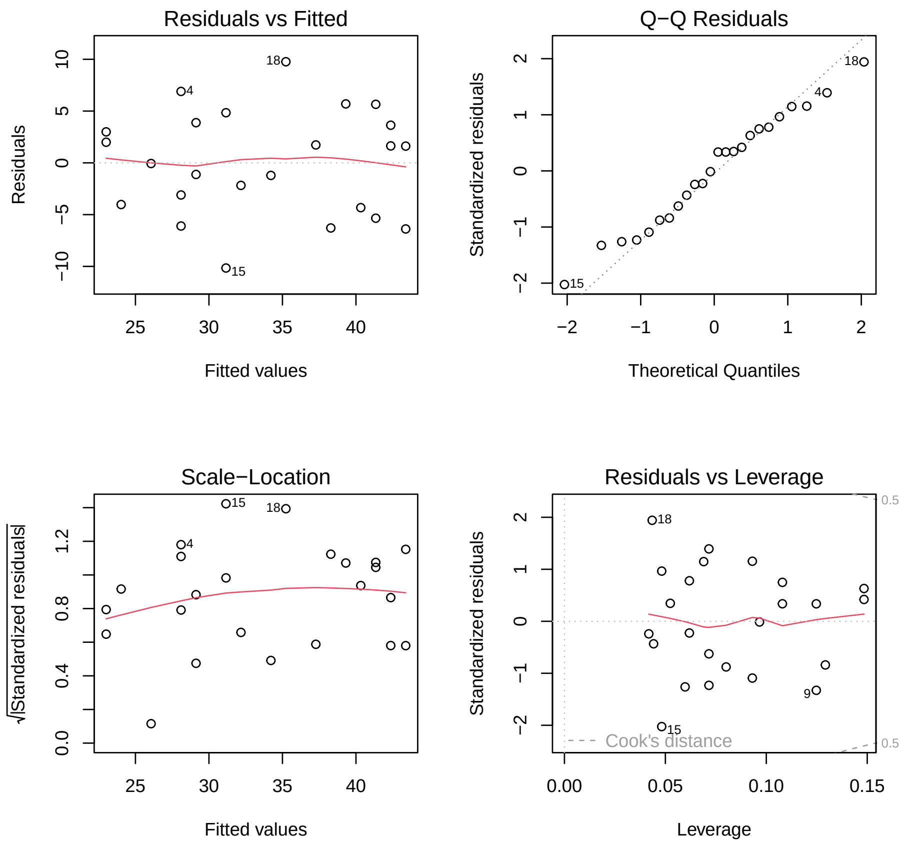
Resistance(x)	43	29	44	33	33	47	34	31	48	34	46	37	36	39
	36	47	28	40	42	33	46	28	48	45				
Failure time(y)	32	20	45	35	22	46	28	26	37	33	47	30	36	33
	21	44	26	45	39	25	36	25	45	36				

```
res <- c(43,29,44,33,33,47,34 ,
         31,48,34,46,37,36,39,36 ,
         47,28,40,42,33,46,28,48,45 )
fail <- c(32,20,45,35,22,46,28,
          26,37,33,47,30,36,33,21,44,
          26,45,39,25,36,25,45,36       )
fit=lm(fail~res)
fit

##
## Call:
## lm(formula = fail ~ res)
##
## Coefficients:
## (Intercept)          res
##      -5.518        1.019
```

```
summary(fit)

##
## Call:
## lm(formula = fail ~ res)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1590  -4.1026   0.7752   3.6954   9.7658
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.5175     6.1961  -0.890    0.383
## res           1.0188     0.1581   6.444 1.75e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.142 on 22 degrees of freedom
## Multiple R-squared:  0.6537, Adjusted R-squared:  0.6379
## F-statistic: 41.53 on 1 and 22 DF,  p-value: 1.751e-06
```

Diagnostic Visualizations

```
par(mfrow=c(2,2));
plot(fit)
```



1.2 Multiple Regression

Problem:
The sale of a Product in lakhs of rupees(Y) is expected to be influenced by two variables namely the advertising expenditure X1 (in'OOO Rs) and the number of sales persons(X2) in a region. Sample data on 8 Regions of a state has given the following results

```
Y=c(110,80,70,120,150,90,70,120)
X1=c(30,40,20,50,60,40,20,60)
X2=c(11,10,7,15,19,12,8,14)
```

Area	y	Xl	X2
1	110	30	11
2	80	40	10
3	70	20	7
4	120	50	15
5	150	60	19
6	90	40	12
7	70	20	8
8	120	60	14

```
input_data=data.frame(Y,X1,X2)
input_data

##     Y X1 X2
## 1 110 30 11
## 2  80 40 10
## 3  70 20  7
## 4 120 50 15
## 5 150 60 19
## 6  90 40 12
## 7  70 20  8
## 8 120 60 14
```

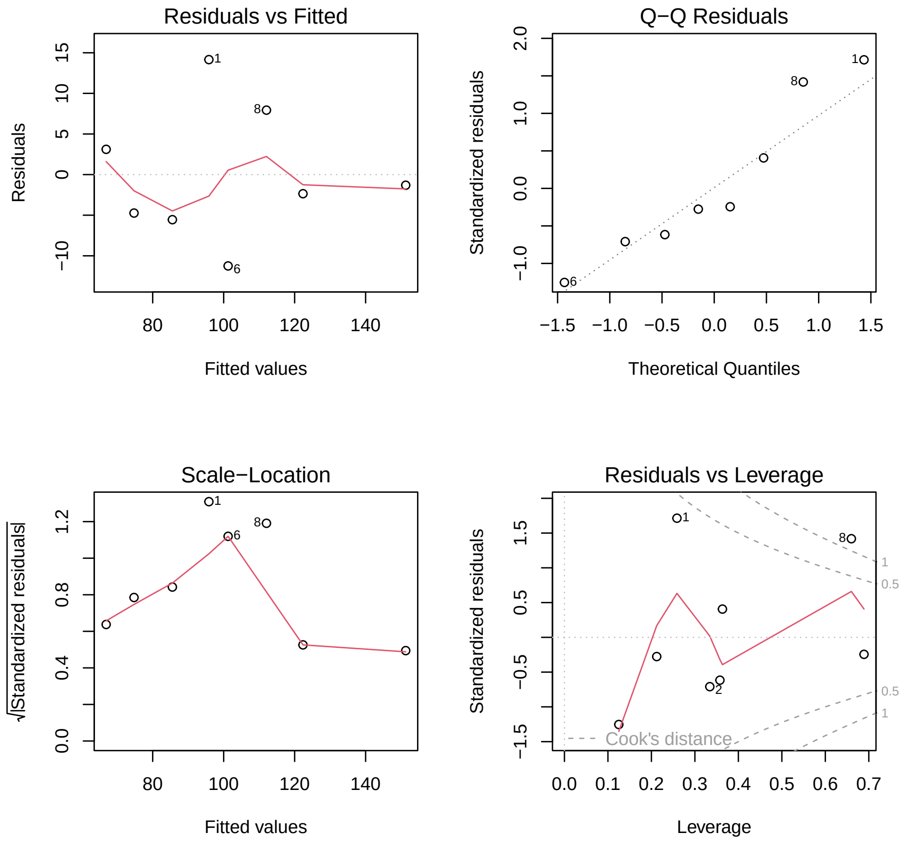
```
RegModel <- lm(Y~X1+X2, data=input_data)
RegModel

##
## Call:
## lm(formula = Y ~ X1 + X2, data = input_data)
##
## Coefficients:
## (Intercept)           X1           X2
##     16.8314      -0.2442       7.8488
```

```
summary(RegModel)

##
## Call:
## lm(formula = Y ~ X1 + X2, data = input_data)
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  14.157  -5.552   3.110  -2.355  -1.308 -11.250  -4.738   7.936
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.8314    11.8290   1.423   0.2140
## X1           -0.2442     0.5375  -0.454   0.6687
## X2            7.8488     2.1945   3.577   0.0159 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.593 on 5 degrees of freedom
## Multiple R-squared:  0.9191, Adjusted R-squared:  0.8867
## F-statistic:  28.4 on 2 and 5 DF,  p-value: 0.001862
```

Interpretation : The regression model is Y =16.834-0.2442*X1+7.8488*X2

```
par(mfrow=c(2,2));
plot(RegModel)
```



Experiment

1. We want to predict the sales of a store based on advertising budget (in thousands of dollars). We have data for 15 different observations. What will be sales for 29000.

Observation	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Budget ($k)	5	8	10	6	7	4	9	12	11	3	13	15	14	6	7
Sales ($k)	45	52	60	48	50	40	58	72	68	35	76	85	80	47	53

2.We want to predict the expenditure of individuals based on their stock holdings (in thousands of dollars) and savings (in thousands of dollars). Give the plots.

Stock ($k)	15	20	18	10	22	12	16	24	30	13	14	21	19	25	28
Savings ($k)	10	12	14	8	18	6	13	20	25	7	10	15	11	17	22
Expenditure ($k)	25	30	28	18	38	20	27	42	50	21	23	35	32	40	45

3. Find the regression of tree's age based on circumference for the Orange dataset.