Homework 6 - STAT 511

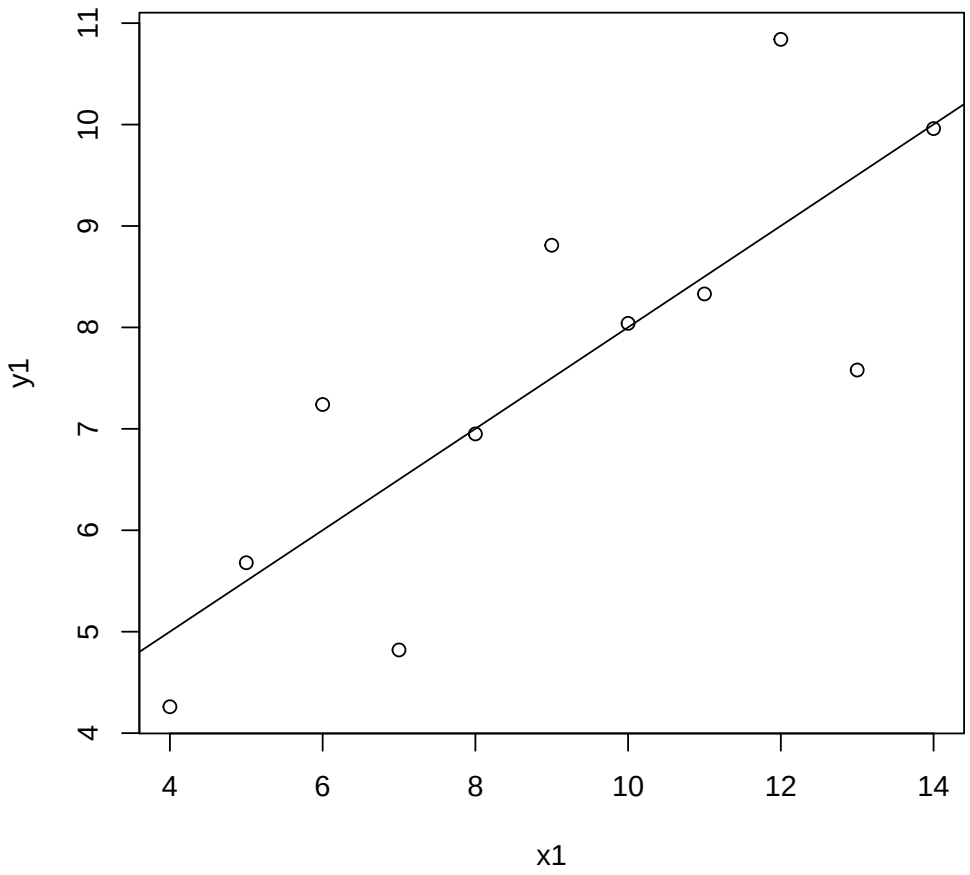
Amal Agarwal

Answer 3

(a) The estimated parameters obtained by fitting the pairs (x_i, y_i) for $i = 1, 2, 3, 4$ are respectively given as

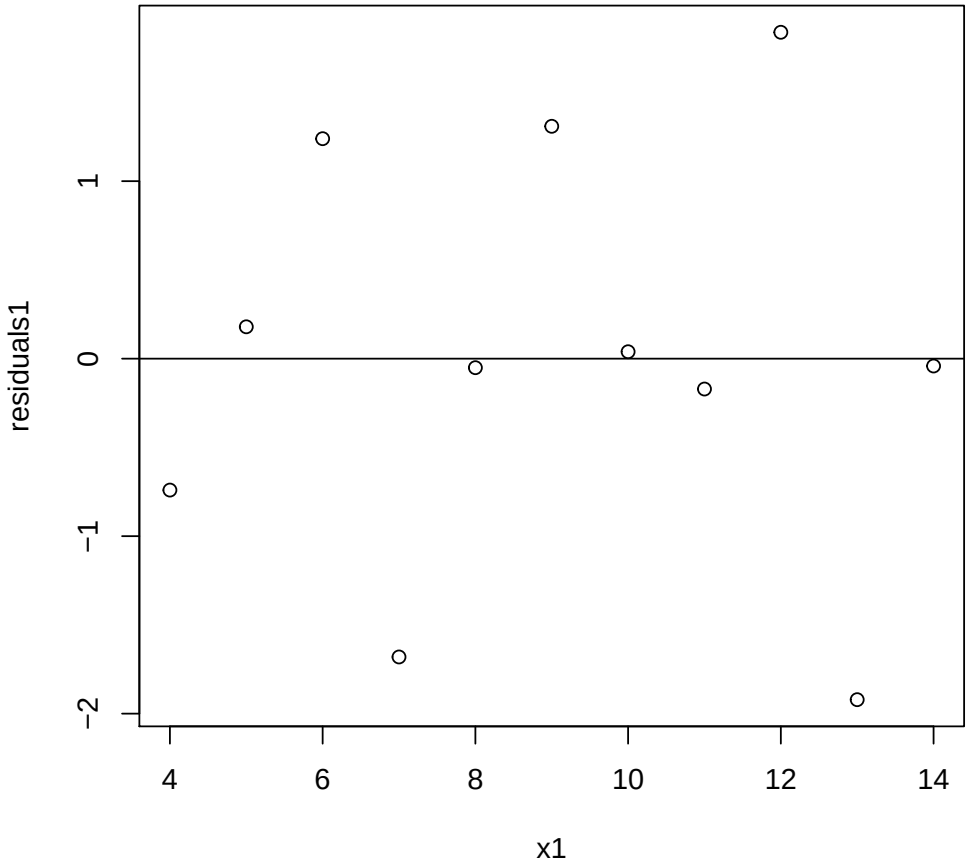
```
(Intercept)          x1
  3.0000909   0.5000909

(Intercept)          x2
   3.000909    0.500000

(Intercept)          x3
  3.0024545   0.4997273

(Intercept)          x4
  3.0017273   0.4999091
```

(b)
- Analysis of (x_1, y_1) within the linear model framework:
  Scatter plot of (x_1, y_1) along the fitted line is given as:



The relationship looks linear from the scatter plot.

The residual plot is given as



The above residual plot shows that the residuals are almost uniformly distributed around 0. Thus our assumption of homoscedascity is verified. There is no positive or negative correlation between successive values of residuals which verifies the assumption of independence. Further there is no identifiable non-linear trend and thus we can also say that our assumption of linearity is verified.

To check for possible outliers, we can do a leave one out analysis to get the studentized residuals and construct the following test.

H_0: Linear Model is correctly specified.
Reject H_0 if $r_0^* > t\left((1-\frac{\alpha}{2}),(n-p-1)\right)$ where $t\left((1-\frac{\alpha}{2}),(n-p-1)\right)$ is the $(1-\frac{\alpha}{2})$ quantile of a t-distribution with $(n-p-1)$ degrees of freedom. This follows

from the fact that $r_i^* = \dfrac{\widehat{\epsilon_{(i)}}}{\widehat{\sigma_{(i)}}\left(\sqrt{1+\underline{X_i}'(X_{(i)}'X_{(i)})^{-1}\underline{X_i}}\right)}$ follows a t-distribution with $(n-p-1)$ degrees of freedom. The studentized residuals are calculated as:

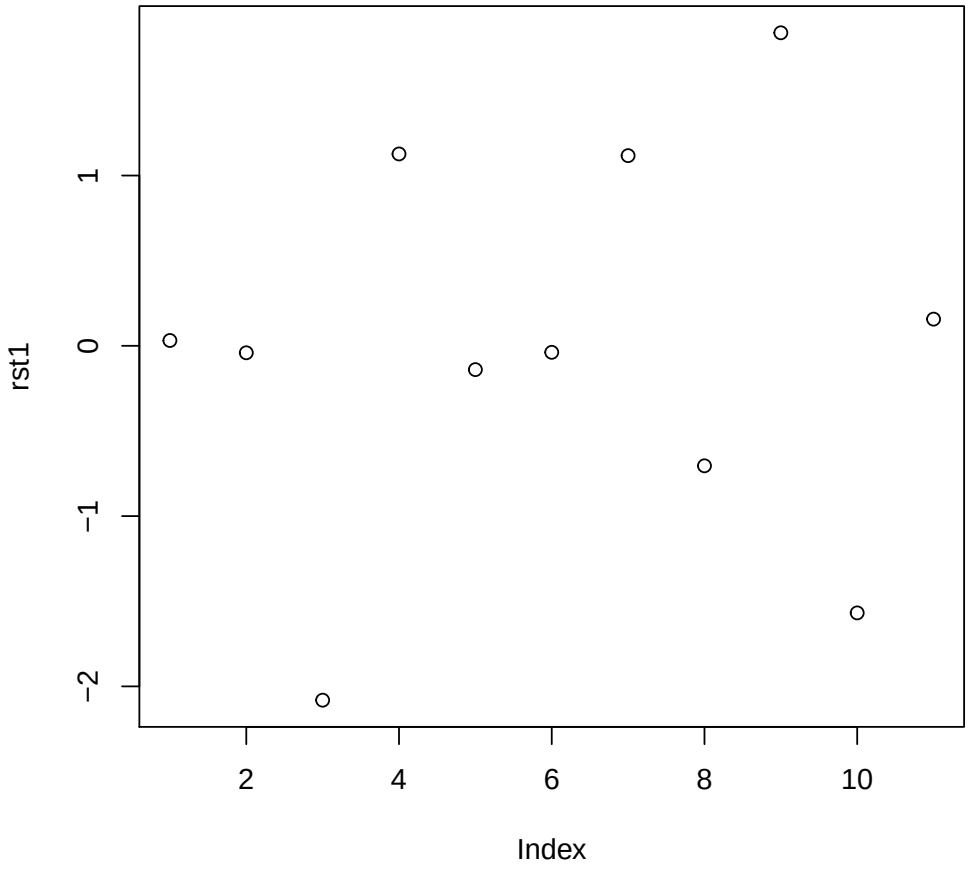
```
> rst1<-rstudent(fit1)
```

The following R code compares the values of the studentized residuals with the $\left(1-\dfrac{\alpha}{2}\right)$ quantile values of the t-distribution.

```
> # level of significance
> alpha<-0.05
> # observation numbers of the outliers
> out1<-rep(0,length(x1))
> rst1<-rstudent(fit1)
> j=1
> for (i in 1:length(x1)){
+   if (rst1[i]>qt((1-(alpha/(2*length(x1)))),(n-p-1))){
+     out1[j]<-i
+     j<-j+1
+   }
+ }
> out1
 [1] 0 0 0 0 0 0 0 0 0 0 0
```

Plotting the studentized residuals

```
> plot(rst1)
> cutoff=qt(.975,df=n-p-1)
> abline(h=cutoff)
> abline(h=-cutoff)
```



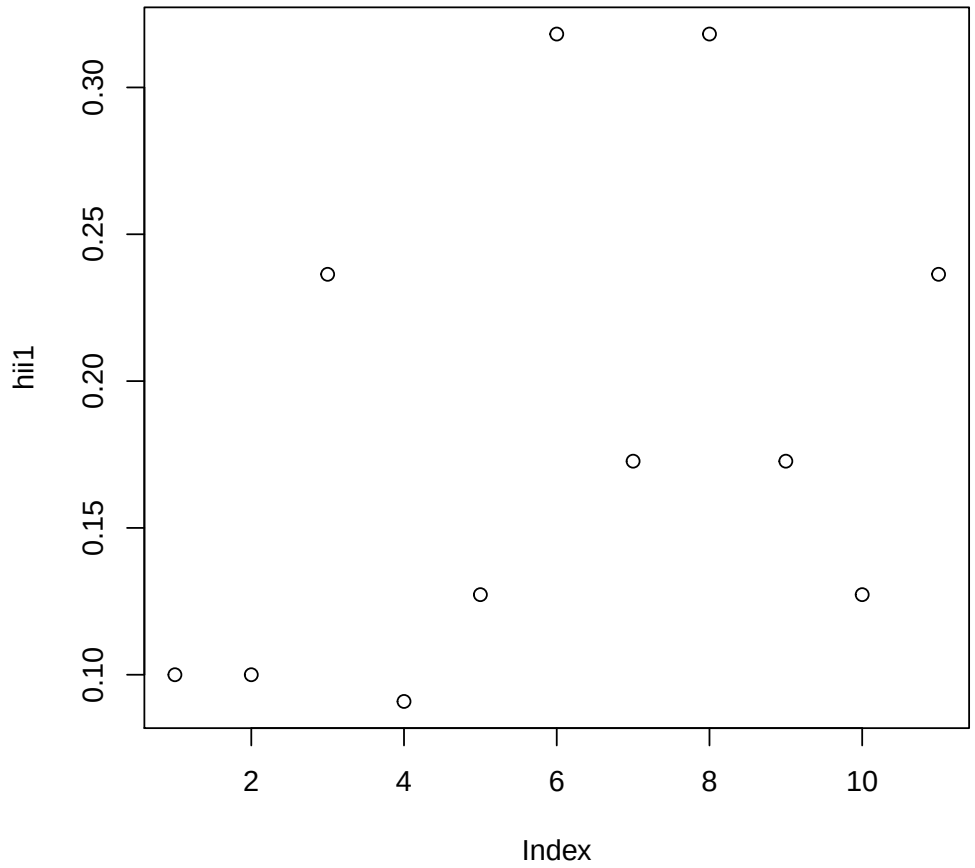
In the above plot, the cutoff lines go out of the bounds. Therefore there are no outliers and hence our null hypothesis is not rejected.
The influential points can be found by looking at various diagnostic measures like leverage i.e. diagonal elements of Hat matrix (h_{ii}) and Cook's distance (CD). We consider an observation likely to be influential if $h_{ii} > \frac{2p}{n}$ or $CD_i > 1$. This is done using the following R code:

```
> hii1<-hatvalues(fit1)
> cd1<-cooks.distance(fit1)
> inf1.1<-rep(0,length(x1))
```

```
> inf1.2<-rep(0,length(x1))
> j=1
> for (i in 1:length(x1)){
+   if (hii1[i]>(2*p/n)){
+     inf1.1[j]<-i
+     j<-j+1
+   }
+ }
> j=1
> for (i in 1:length(x1)){
+   if (cd1[i]>1){
+     inf1.2[j]<-i
+     j<-j+1
+   }
+ }
> inf1.1
 [1] 0 0 0 0 0 0 0 0 0 0 0 0
> inf1.2
 [1] 0 0 0 0 0 0 0 0 0 0 0 0
```

Plotting the diagonal hat values:

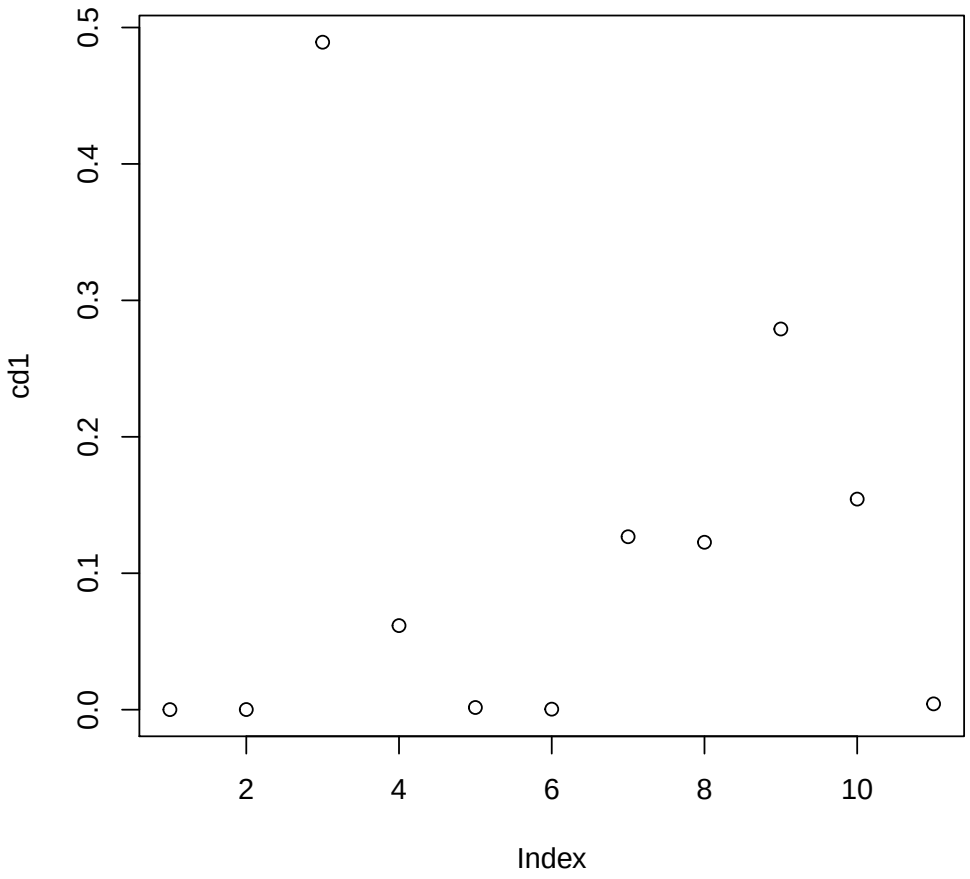
```
> cutoff.h<-2*p/n
> plot(hii1)
> abline(h=cutoff.h)
```



Again the cutoff line in the above graph is out of the bounds.

Plot of Cook's distance values:

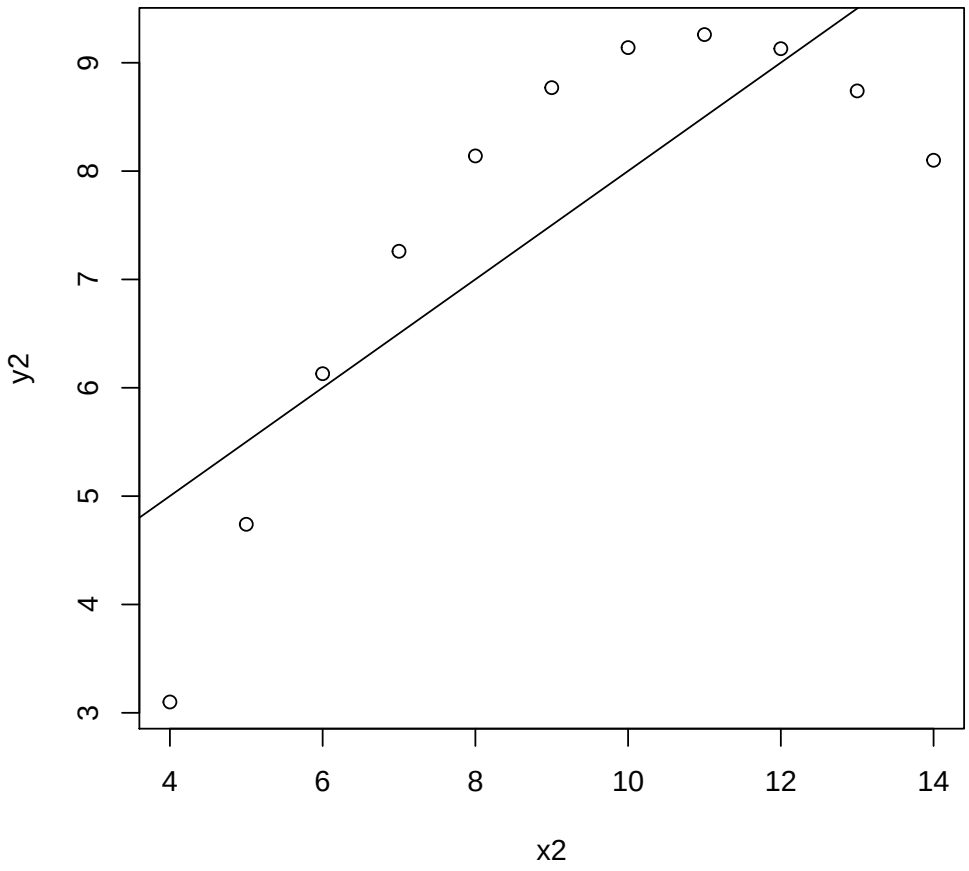
```
> cutoff.cd1<-1
> plot(cd1)
> abline(h=cutoff.cd1)
```



Again the cutoff line in the above graph is out of the bounds.

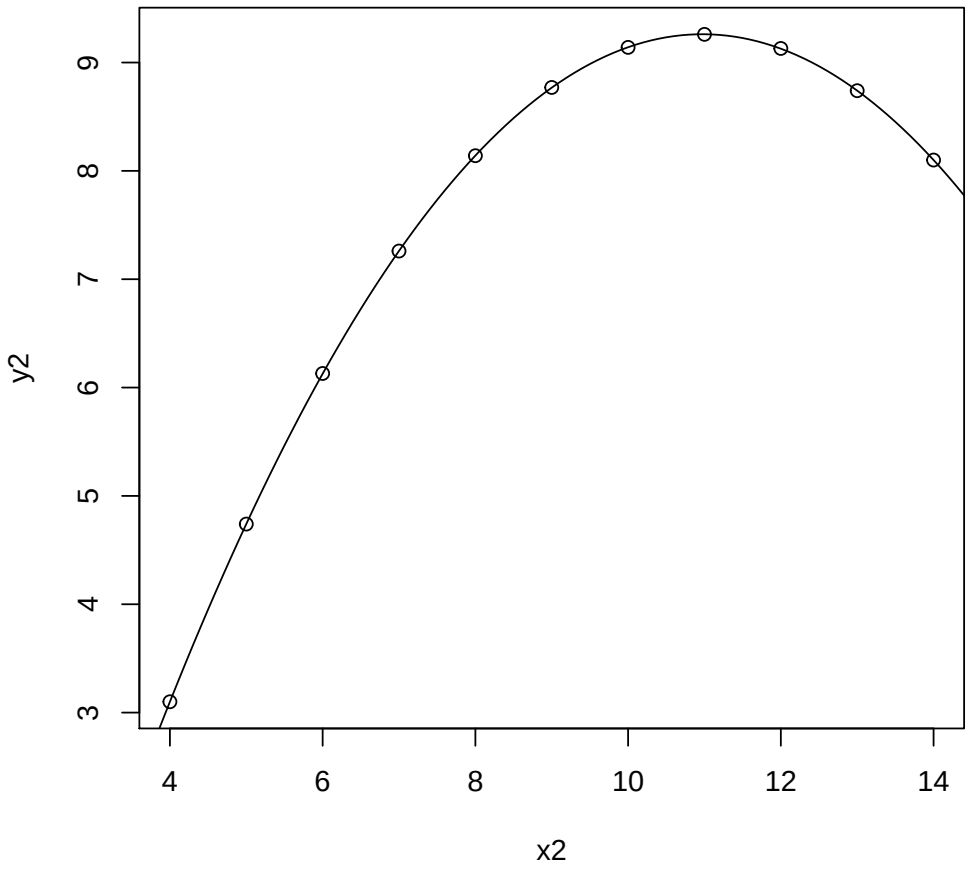
Thus we find no influential points with both the leverage criterion and the Cook's distance criterion.

- Analysis of (x_2, y_2) within the linear model framework:
  Scatter plot of (x_2, y_2) along the fitted line is given as:

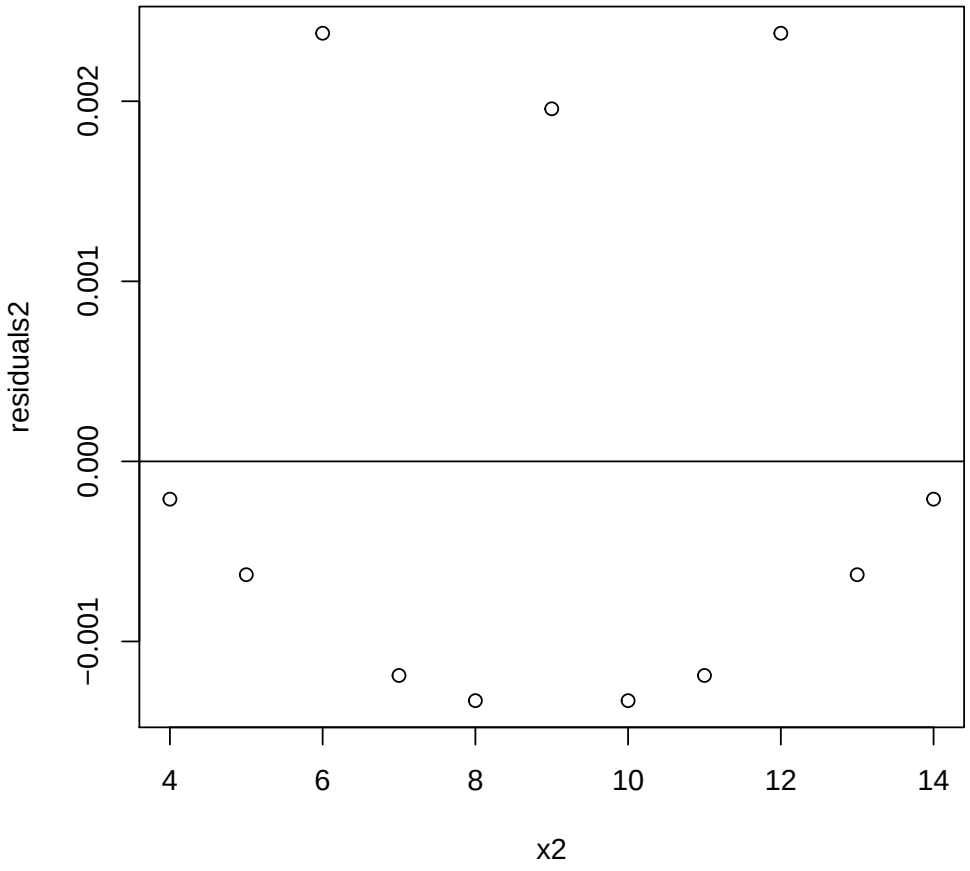


This scatter plot shows a non-linear relationship between the response and predictor. It likes look y depends on x quadratically.

Fitting a polynomial of degree 2 in x2, we obtain the following scatter plot along with the fitted line


y2
3
4
5
6
7
8
9
4
6
8
10
12
14
x2

The residual plot is given as



The above residual plot shows that the residuals are not uniformly distributed around 0. Thus our assumption of homoscedascity is violated. But still there does not seem any positive or negative correlation between successive values of residuals which verifies the assumption of independence. Clearly there is a non-linear trend that has been taken into account already by using polynomial transformation of degree 2.

To check for possible outliers, we calculate the studentized residuals and do the test as before:

```
> rst2<-rstudent(fit2.2)
```

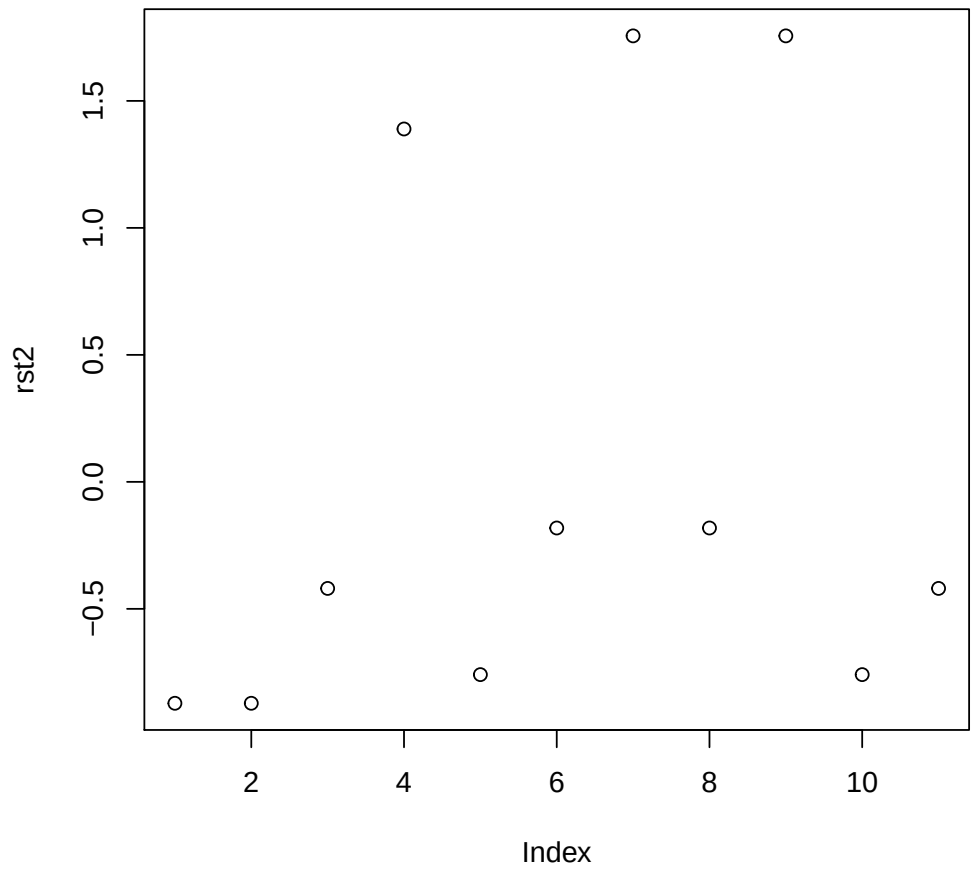
The following R code compares the values of the studentized residuals with the $\left(1-\frac{\alpha}{2}\right)$ quantile values of the t-distribution.

```
> # observation numbers of the outliers
> out2<-rep(0,length(x2))
```

```
> rst2<-rstudent(fit2.2)
> j=1
> for (i in 1:length(x2)){
+   if (rst2[i]>qt((1-(alpha/(2*length(x2)))),(n-p-1))){
+     out2[j]<-i
+     j<-j+1
+   }
+ }
> out2
 [1] 0 0 0 0 0 0 0 0 0 0 0
```

Plotting the studentized residuals:

```
> plot(rst2)
> cutoff<-qt(.975,df=n-p-1)
> abline(h=cutoff)
> abline(h=-cutoff)
```



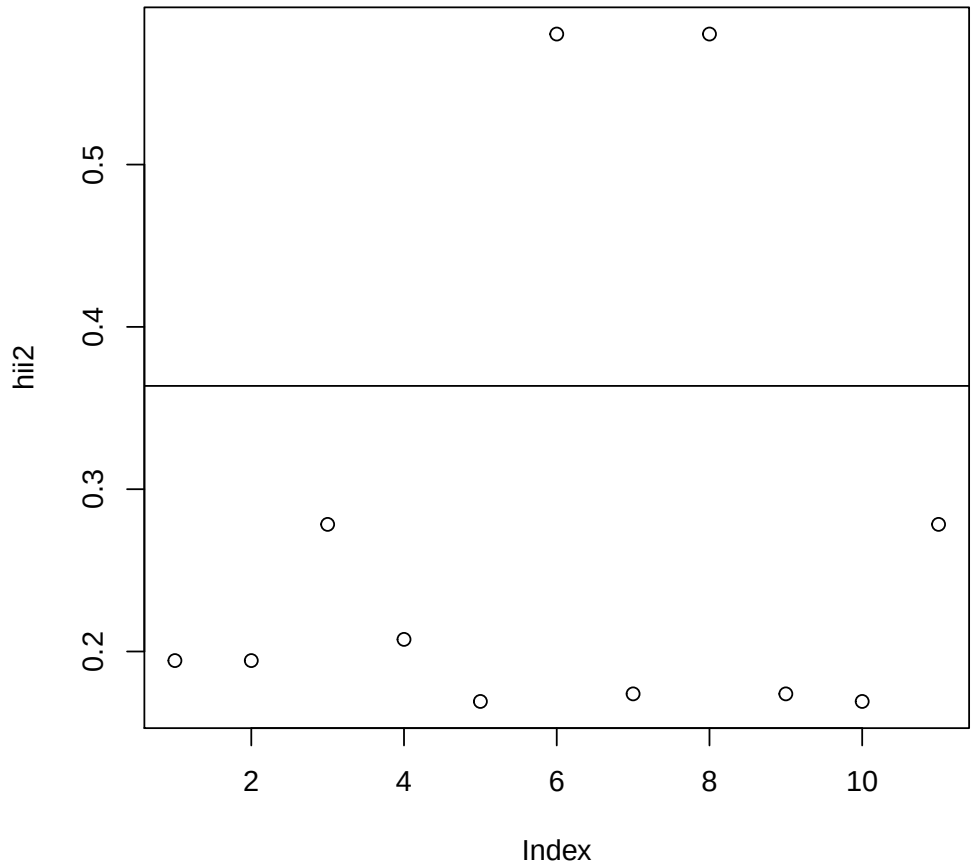
In the above plot, the cutoff lines go out of the bounds. Therefore there are no outliers and hence our null hypothesis is not rejected.

Again influential points were found using the leverage and Cook's distance criteria as follows:

```
> hii2<-hatvalues(fit2.2)
> cd2<-cooks.distance(fit2.2)
> inf2.1<-rep(0,length(x2))
> inf2.2<-rep(0,length(x2))
> j=1
> for (i in 1:length(x2)){
+   if (hii2[i]>(2*p/n)){
+     inf2.1[j]<-i
+     j<-j+1
+   }
+ }
> j=1
> for (i in 1:length(x2)){
+   if (cd2[i]>1){
+     inf2.2[j]<-i
+     j<-j+1
+   }
+ }
> inf2.1
 [1] 6 8 0 0 0 0 0 0 0 0 0
> inf2.2
 [1] 0 0 0 0 0 0 0 0 0 0 0
```

Plotting the diagonal hat values:

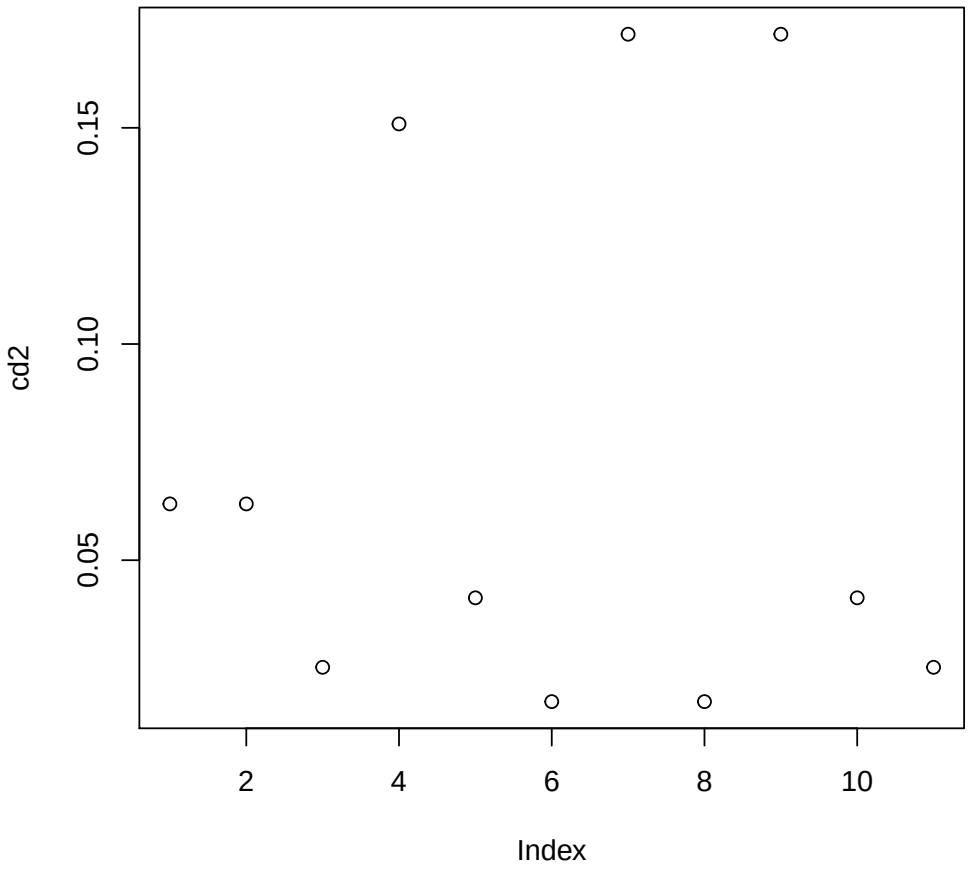
```
> cutoff.h<-2*p/n
> plot(hii2)
> abline(h=cutoff.h)
```



The above plot shows that there are 2 observations that lie out of the cutoff line. From inf2.1 vector these observation numbers are 6 and 8.

Plot of Cook's distance values:

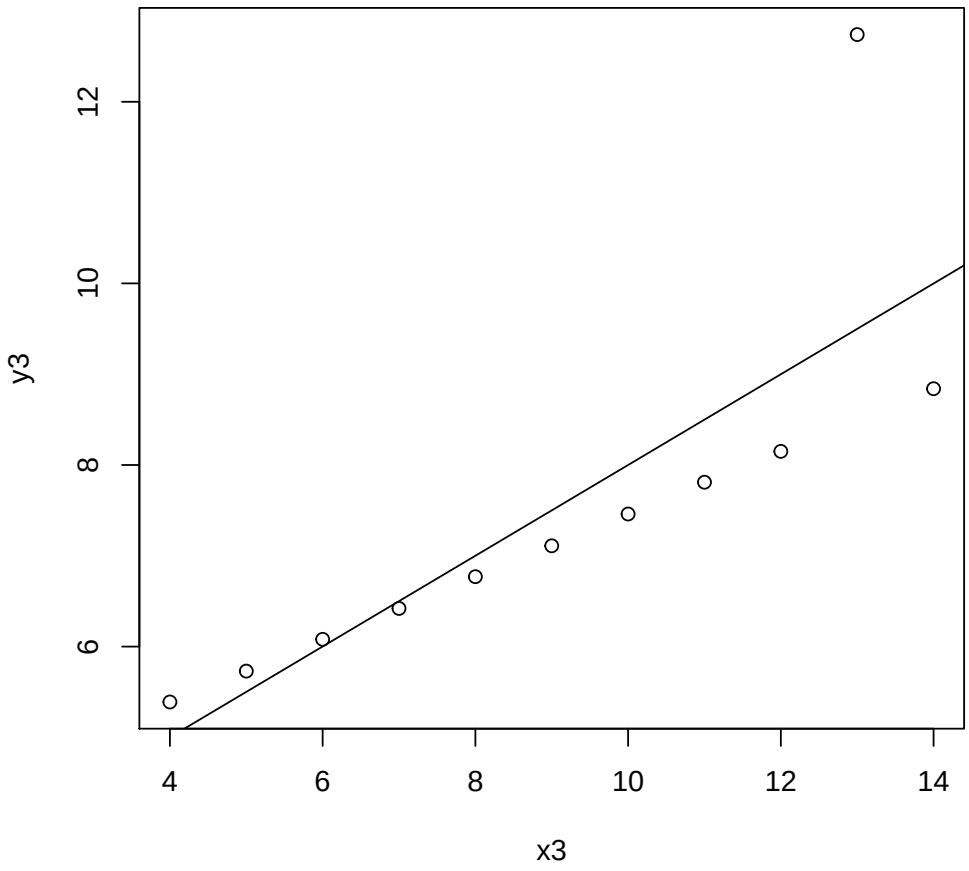
```
> cutoff.cd2<-1
> plot(cd2)
> abline(h=cutoff.cd2)
```



The cutoff line is out of the bounds in this plot.

Thus observations 6 and 8 are influential points from the leverage criterion while the Cook's distance criterion does not give any influential point.

- Analysis of (x_3, y_3) within the linear model framework:
  Scatter plot of (x_3, y_3) along the fitted line is given as:



The scatter plot shows that there is one outlier that is affecting the regression line. Otherwise the relationship seems to be linear and hence no transformations required. Calculating the studentized residuals and doing the test for identifying outliers confirms this as follows:

```
> rst3<-rstudent(fit3)
```

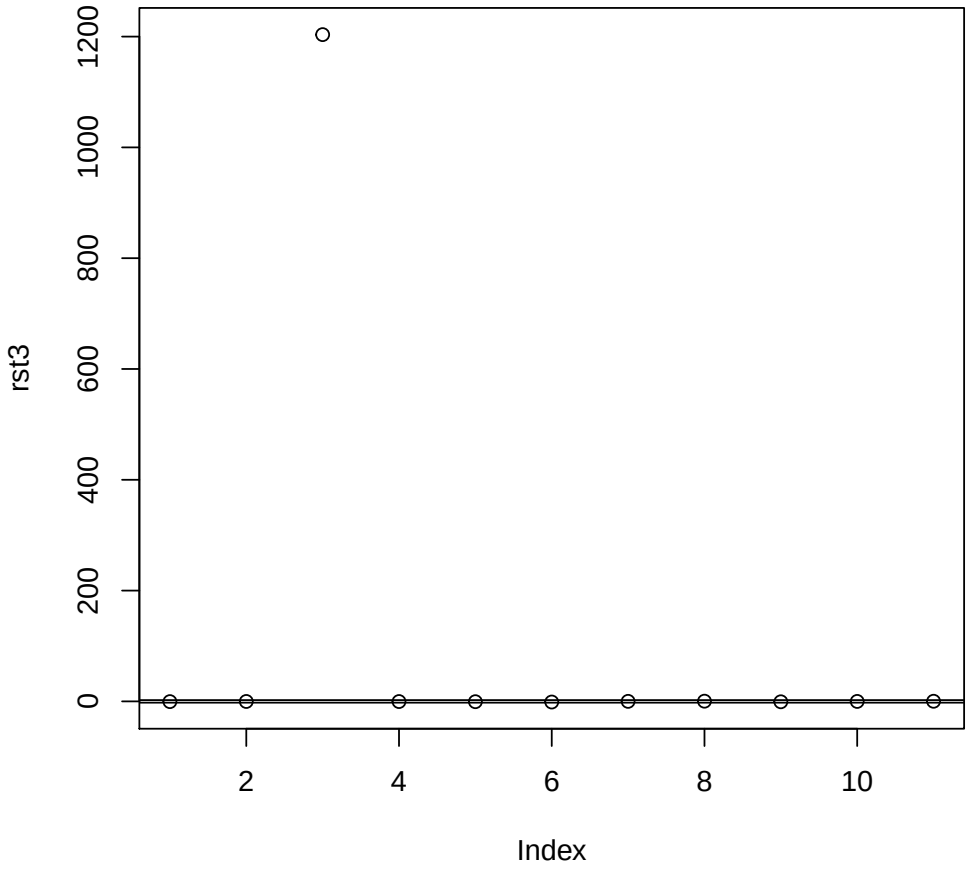
The following R code compares the values of the studentized residuals with the $\left(1 - \frac{\alpha}{2}\right)$ quantile values of the t-distribution.

```
> # observation numbers of the outliers
> out3<-rep(0,length(x3))
> rst3<-rstudent(fit3)
> j=1
```

```
> for (i in 1:length(x3)){
+   if (rst3[i]>qt((1-(alpha/(2*length(x3)))),(n-p-1))){
+     out3[j]<-i
+     j<-j+1
+   }
+ }
> out3
 [1] 3 0 0 0 0 0 0 0 0 0 0
```

Plotting the studentized residuals:

```
> plot(rst3)
> cutoff<-qt(.975,df=n-p-1)
> abline(h=cutoff)
> abline(h=-cutoff)
```

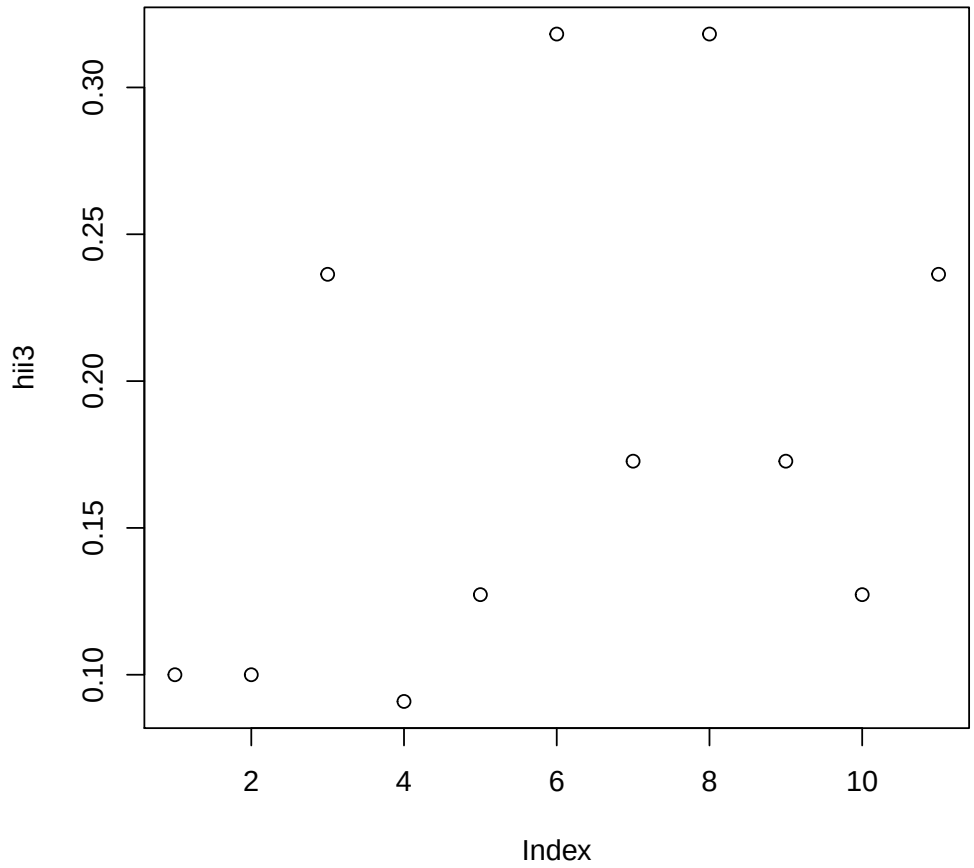


Thus we see that observation 3 is an outlier. Again influential points were found using the leverage and Cook's distance criteria as follows:

```
> hii3<-hatvalues(fit3)
> cd3<-cooks.distance(fit3)
> inf3.1<-rep(0,length(x3))
> inf3.2<-rep(0,length(x3))
> j=1
> for (i in 1:length(x3)){
+   if (hii3[i]>(2*p/n)){
+     inf3.1[j]<-i
+     j<-j+1
+   }
+ }
> j=1
> for (i in 1:length(x2)){
+   if (cd3[i]>1){
+     inf3.2[j]<-i
+     j<-j+1
+   }
+ }
> inf3.1
 [1] 0 0 0 0 0 0 0 0 0 0 0
> inf3.2
 [1] 3 0 0 0 0 0 0 0 0 0 0
```

Plotting the diagonal hat values:

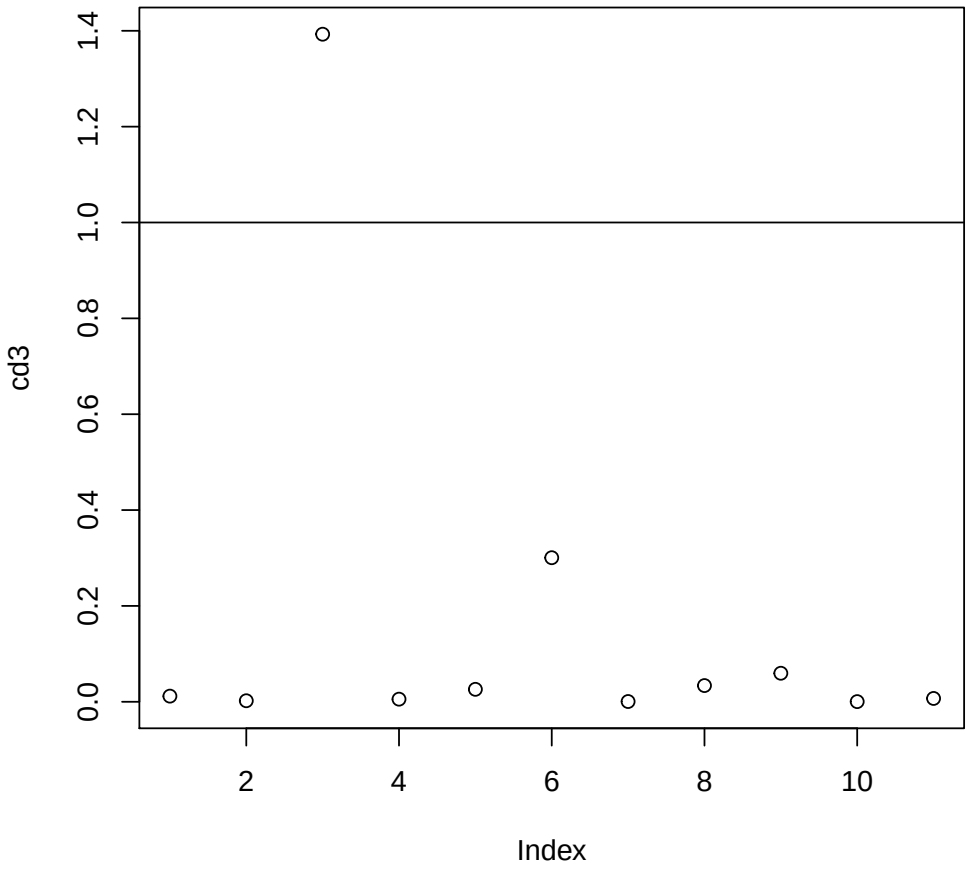
```
> cutoff.h<-2*p/n
> plot(hii3)
> abline(h=cutoff.h)
```



The above plot shows that the cutoff lines are out of bounds.

Plot of Cook's distance values:

```
> cutoff.cd3<-1
> plot(cd3)
> abline(h=cutoff.cd3)
```

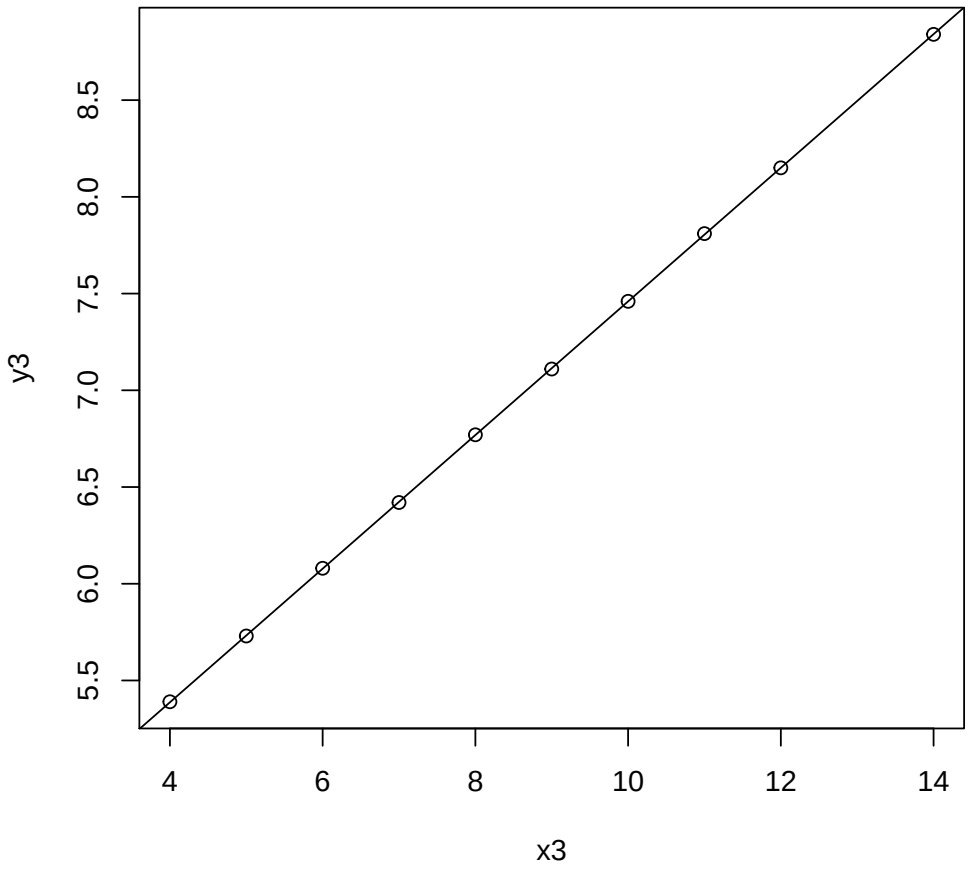


The above plot shows that there is one observation that lies out of the cutoff line. From inf3.2 vector this is observation number 3.

Thus there are no influential points according to the leverage criterion while the Cook's distance criterion gives observation 3 as an influential point.

Since according to Cook's distance, the outlier observation number 3 is also influential it may be wise to drop it.

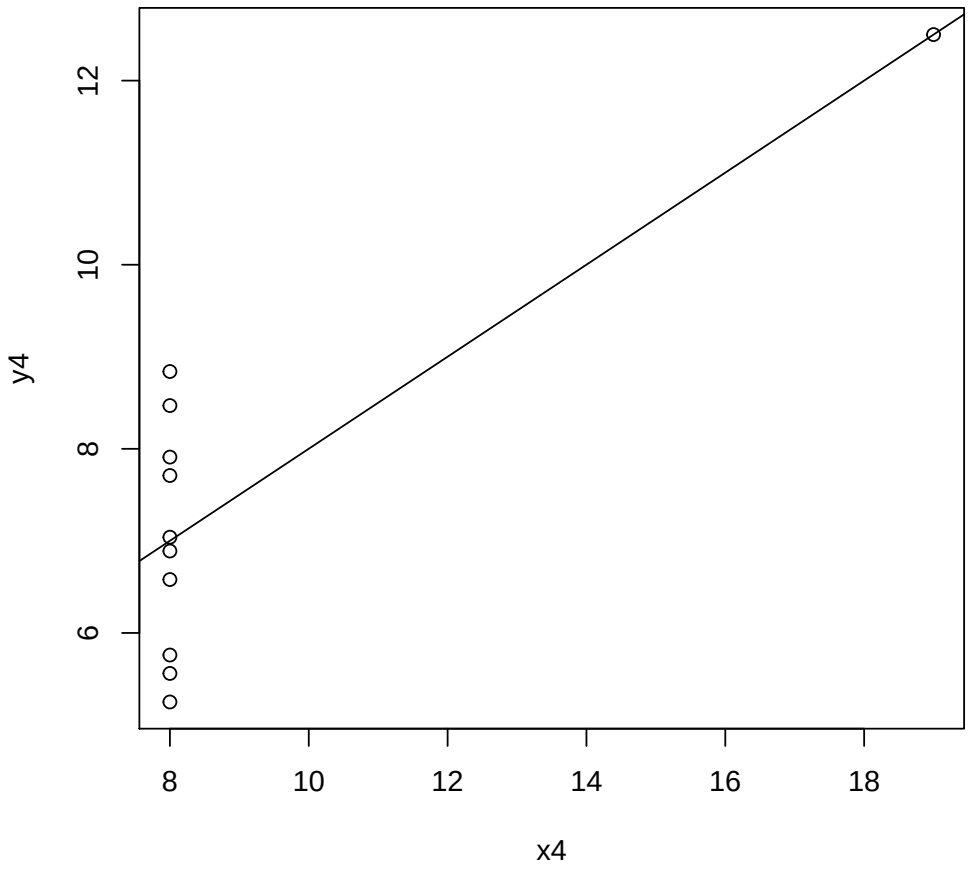
Removing this outlier and fitting the model again we get



which looks like a perfect fit. The new estimated parameters are:

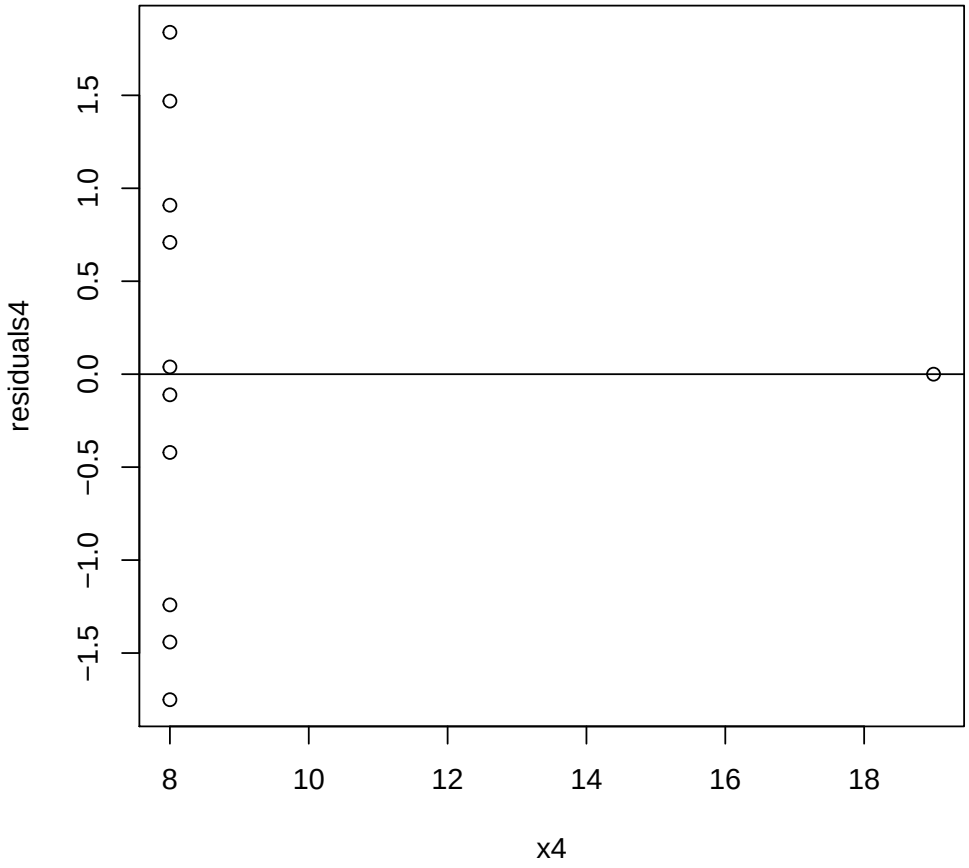
```
(Intercept)          x3
  4.0056494   0.3453896
```

- Analysis of (x_4, y_4) within the linear model framework:
  Scatter plot of (x_4, y_4) along the fitted line is given as:



In this case, the scatter plot clearly shows that there is an outlier. However it seems that the experimenter who took the data measured the response many times on one value of the predictor and only once for another value of the predictor. The scatter in different values of response may be a result of instrumental errors. The relationship between x and y still seems to be linear with some slope and there is no reason to neglect the outlier.

The residual plot is given as



The above residual plot shows that the residuals are uniformly distributed around 0. Thus our assumption of homoscedascity is verified. Further there does not seem any positive or negative correlation between successive values of residuals which verifies the assumption of independence. It may appear at the first glance that there is a non-linear trend. However this is merely an artifact since we only have two data points effectively corresponding to two different predictor values.

(c) Predicting $E(y)$ and 95% prediction intervals at $x = 13$ for all the fitted models:

- For the first model, the predicted value of y at x=13 along with the 95% confidence interval are given as

```
> newdata1=data.frame(x1=13)
> predict(fit1, newdata1, interval="predict")
```

```
       fit      lwr      upr
1 9.501273 6.390801 12.61174
```

The modeling assumptions for the fitted model are satisfied in this case and so 95% prediction intervals can be trusted.

- For the second model, the predicted value of y at x=13 along with the 95% confidence interval are given as

```
> newdata2=data.frame(x2=13)
> predict(fit2.2, newdata2, interval="predict")
       fit      lwr     upr
1 8.740629 8.736269 8.74499
```

The assumption of homoscedascity from the residual plot does not seem to hold good even for the polynomial model and so the 95% prediction interval can't be trusted.

- For the third model, the predicted value of y at x=13 along with the 95% confidence interval are given as

```
> newdata3=data.frame(x3=13)
> predict(fit3.2, newdata3, interval="predict")
       fit      lwr      upr
1 8.495714 8.487582 8.503846
```

The model fitted without the outlier looks very good and so 95% prediction intervals based on this model can be trusted.

- For the last model, the predicted value of y at x=13 along with the 95% confidence interval are given as

```
> newdata4=data.frame(x4=13)
> predict(fit4, newdata4, interval="predict")
       fit      lwr      upr
1 9.500545 6.392357 12.60873
```

The 95% prediction intervals based on this model is solely based on the data point corresponding to the 2nd value of the predictor. Already the interval is very wide pointing to a high uncertainity. Thus it can be trusted but with a pinch of salt.