204: Regression Analysis

Instructor: Prof. Sugata Sen Roy

Report Card
Data Analysis with Regression

Outlier and Influential Observation Detection

Submitted by: Soumarya Basak

Submitted on: June 22, 2022

University of Calcutta

Department of Statistics

Contents

Problem 1

1.1 About the Practical

We have a data on number of person died due to lung cancer and their annual per capita cigarette consumption for different countries. Here our task is to fit a regression model on the date and find the outliers and influential observations.

1.2 Analysis

It is quite obvious that *deaths due to lung cancer* depends on *cigarette consumption* of a person, so for this data, *"per capita cigarette consumption"*(X) is our regressor and *"deaths of the person due to lung cancer"*(y) is our response variable.

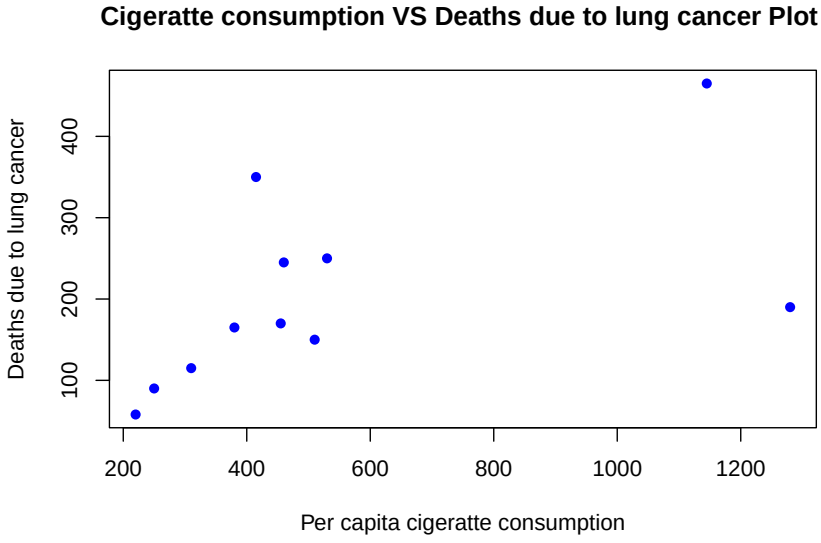
Analysis Required :

1. To fit a regression Model
2. Study the *Standardize* and *Studentize* residuals, and *Leverage*
3. Obtain the DFBETA's , DFFIT's and Cook's Distance

All the analysis required for the data is performed in lab, on suitable software and the result is discussed below.

1.2.1 Fitting Regression Model

Before fitting a model it is always better to have a visualization to the data so we create a scatter plot to observe the pattern in the data.



From the plot it is quite clear that there might be a linear relationship between the variables so now we tried to fit a linear regression model on the data.

Here,
$y\_var =$ "Deaths Due to per to lung cancer (in million)"
$x\_var =$ "per capita cigarette consumption"
Hence the regression model will be,

$$y\_var = \beta_0 + \beta_1(x\_var) + u$$

where, u is the error part of the model with $E(u_i) = 0 \;\; Var(u_i) = \sigma \; \forall \; i = 1, 2, ...n$

On fitting a linear regression model on the data we arrive at the following results. Let give the name to the model as "`model_1`"

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	98.42920	59.30799	1.660	0.1314
x_var	0.19568	0.09343	2.094	0.0657

Residual standard error: 102.7 on 9 degrees of freedom
Multiple R-squared: 0.3277
Adjusted R-squared: 0.253
F-statistic: 4.387 on 1 and 9 DF
p-value: 0.06571

Observation :

From the output of the linear model we see that the **F test is not significant** as the p value is more than 0.05 and so we can conclude that the coefficient are insignificant for the model, so we doesn't need to model the data which quite contradict our assumption or scatter plot that there might be a linear relationship. Also note from the result we can see that the residuals are highly dispersed between -158.90 and 170.36, which is due to bad fit.

Conclusion 1 :

There might be several reason for such result one maybe that the data consists of some outlier value or influential observation which make model insignificant and we need to find out those, and note that the scatter plot shows there are few values which are far away from the other values which causes some influences the model. So we need to detect them. In next sections we will try to detect outliers and the influential observations of the data so that we can fit good model to the data.

1.2.2 Detecting Outliers and Influential Observations

Detecting Residuals and Leverage

Residuals are useful to obtain the outlier in response variable. Note that the high residuals always signifies that the predicted value and the actual value is far away which signifies outlier value.
So we find the residuals for the data by the "`model_1`" and try to observed some higher value within them.

Residuals

After fitting the above model on the data we get the following residuals

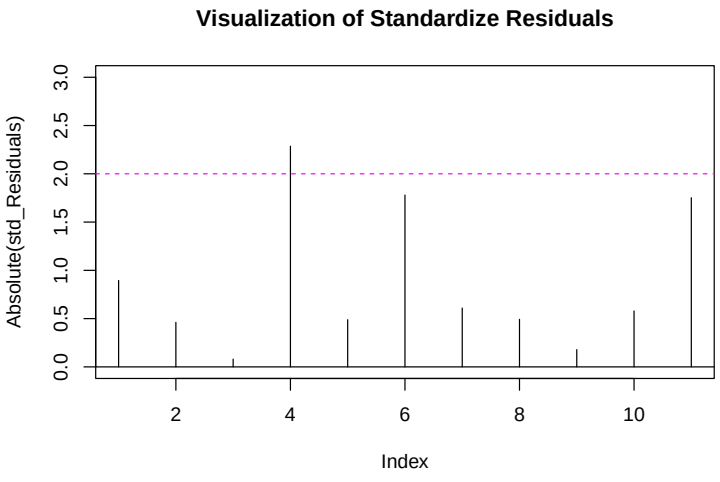
```
##      Residuals
## 1   -83.478964
## 2   -44.090230
## 3    -7.787881
## 4  -158.900543
## 5    47.860008
## 6   142.516357
## 7   -57.349386
## 8   -48.226377
## 9   -17.463937
## 10   56.557660
## 11  170.363293
```

We can see that few of the residuals are very large but it's difficult to say whether the residuals indicates the outlier or not. For that reason we moved to standardized and studentized residuals in the following sections.

Standardize Residuals

This standardized residuals for the data he's shown below

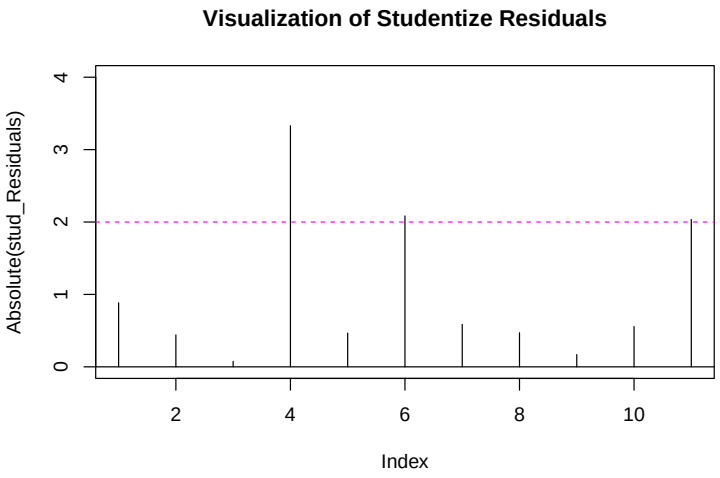
```
##    Standardize_Residuals
## 1             -0.89544653
## 2             -0.46154959
## 3             -0.08047462
## 4             -2.28633046
## 5              0.48868724
## 6              1.77978693
## 7             -0.60956366
## 8             -0.49261973
## 9             -0.17891785
## 10             0.57921016
## 11             1.75221345
```



From the plot it is clear that three residuals values is usually larger than the other value so they can be treated as outlier. But it will be more clear if we look for studentized residuals

Studentize Residuals

```
##    Studentize_Residuals
## 1           -0.88455745
## 2           -0.44039638
## 3           -0.07589952
## 4           -3.32934087
## 5            0.46697601
## 6            2.08444722
## 7           -0.58694596
## 8           -0.47083750
## 9           -0.16898616
## 10           0.55655620
## 11           2.03523182
```



From the plot it is clear that three observations have high studentize residuals so we have to look for them and they are working as outlier in the data set.

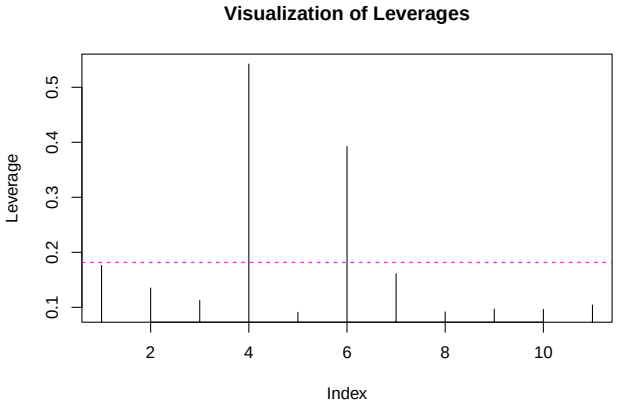
Conclusion 2 :

After finding the high studentized residual value we got that the 4^{th}, 6^{th}, and 11^{th} observations have high studentize residuals which indicates that these observations are the response outliers for the model.

Leverages

Mainly leverage are useful to identify the regressor outlier for a model. The high value of leverage indicates that it is a outlier in regressor variable.

```
##      Leverage
## 1  0.17634086
## 2  0.13518996
## 3  0.11244867
## 4  0.54223194
## 5  0.09101591
## 6  0.39233234
## 7  0.16113488
## 8  0.09172282
## 9  0.09707913
## 10 0.09638538
## 11 0.10411810
```



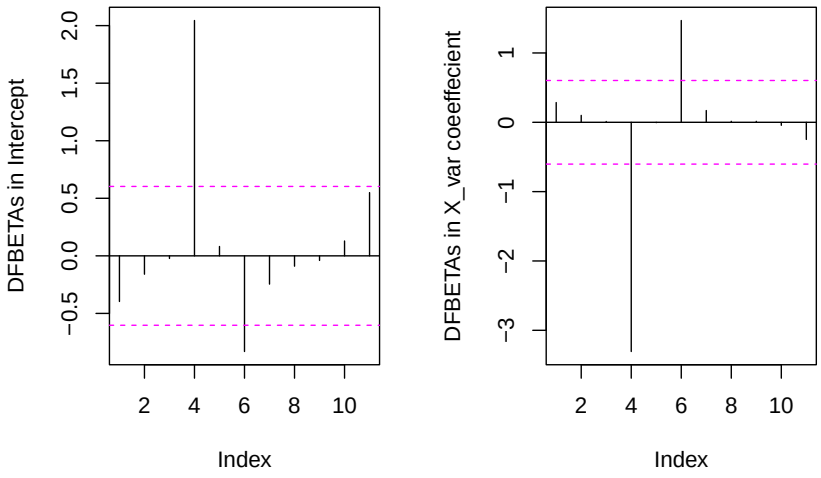
Conclusion 3 :

From the plot of the leverage it is clear that the 4^{th} and 6^{th} observations have high leverages and so they can be treated as the X outlier here.

Calculating DFBETA's

```
##    (Intercept)        x_var
## 1  -0.39641388  0.284880021
## 2  -0.15955137  0.099653248
## 3  -0.02276883  0.011823905
## 4   2.04443096 -3.305819823
## 5   0.08143844 -0.005062287
## 6  -0.83095128  1.468068127
## 7  -0.24573243  0.169824513
## 8  -0.08980784  0.014092928
## 9  -0.03991472  0.013969157
## 10  0.12913805 -0.043327210
## 11  0.54932238 -0.247128933
```

Visuals ofDFBETA's

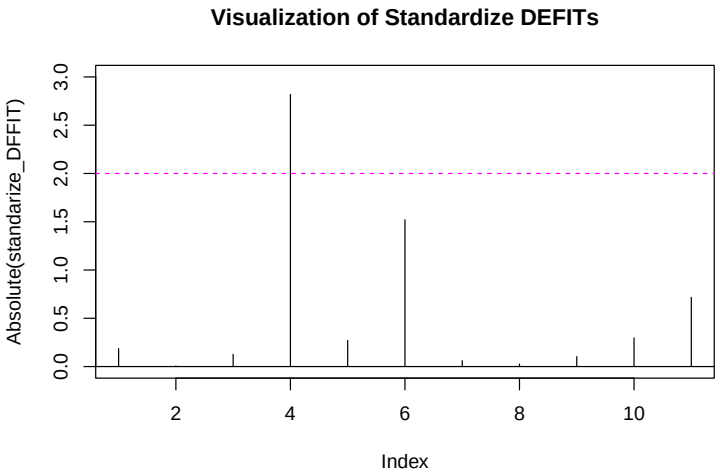


Observation :

Remember that DFBETA is a vector valued and so we have two columns in the output, one for intercept and one for the X variable as the model has two parameters. From the visual plot of the DFBETA, it is clear that two observations has higher DFBETA in both the parameters and so we need to find them. After that we got that it's also come out to be the 4th and the 6th observation.

Calculating DFFIT's

```
##          DEFIT
## 1  -0.40928777
## 2  -0.17412279
## 3  -0.02701590
## 4  -3.62349834
## 5   0.14776620
## 6   1.67488461
## 7  -0.25724494
## 8  -0.14962375
## 9  -0.05541014
## 10  0.18177033
## 11  0.69382767
```

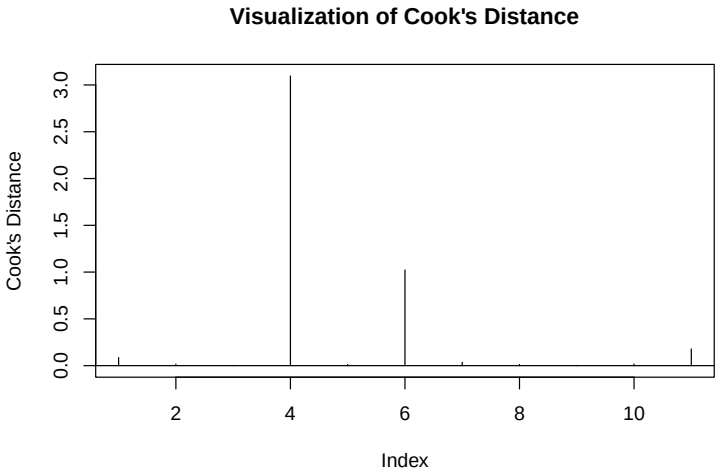


Observation :

Similarly as the previous cases we got that two values 4th and 6th observation have high DFFIT values than others. From the DFFIT values it is difficult to find out the higher value of the lower value and to make our task easier we standardized the DFFIT values and visualize their absolute magnitude.

Calculating Cook's Distance

```
##          Cook_d
## 1  0.0858330926
## 2  0.0166506223
## 3  0.0004102502
## 4  3.0959049540
## 5  0.0119561965
## 6  1.0225722645
## 7  0.0356866335
## 8  0.0122532866
## 9  0.0017208916
## 10 0.0178924695
## 11 0.1784101215
```



Observation :

Cooks distance shows that there are two unusual higher value than the others and those are 4 th and 6 th observation

Conclusion 4 :

All the measures calculated above are mainly useful to detect influential observation in the model.
(I) Firstly higher DFBETA indicates that the estimates of the parameters differs very much from the estimates of the parameters after deleting the i^{th} observation, which indicates that the i^{th} observation may influenced the model and so the high value of DFBETA indicates that i^{th} observation maybe an influential observation,

Our calculation shows that DFBETA for three values — 4th observation, 6th observation, 11th observation are higher. So they may be treated as a influential observation for the model.
(II) As DFBETA is a vector valued measure it is difficult to compare and an improve measure DFFIT is introduced. DFFIT is the difference between the predicted value and the predicted value after removing i^{th} observation.So, high value of DFFIT indicates that predicted value changes very much when we remove one observation which indicates that the observation which was removed may be influential observation.

From our calculation we have only one value more than 2 but there are overall 3 values which are actually larger than the usual DFFIT values. So these 3 values may be influential observations. The 3 values are the 4th observation, 6th observation and 11th observation.
(III) Cook's distance is another measure for detecting influential observation. The the unusual higher value indicates that the observation may be influential observation. From our analysis it is clear that the 4th, 6th observations has higher cook distance.

Up to this stage it comes out that the 4th and 6th observations are highly likely to be influential observations and the 11th observation may also be an influential observation.

1.2.3 Re-modeling the Data

In the very first consequence, we need to deal with these influential observations but as we don't have any knowledge about the genesis of the data we can't manipulate those by cross checking them. The last option we left with is to **remove the 4th and 6th observation** from the data and refit the data. But the fit again comes out insignificant.

Now to overcome this we will next **remove the 11th of observation** as the 11th observation is likely to be an influential observation and remodel the data. Give the model name `"model_2"`.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	-37.7487	48.8416	-0.773	0.46894
x_var	0.4960	0.1207	4.111	0.00628

Residual standard error: 37.79 on 6 degrees of freedom
Multiple R-squared:0.738
Adjusted R-squared: 0.6943
F-statistic: 16.9 on 1 and 6 DF
p-value: 0.00628

Now in the `"model_2"` the F test is significant as the p value is low then 0.05 but the intercept term is statistically insignificant so we need to drop intercept from the model and refit it.

Show the final feet after doing all the necessities is

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
x_var	0.40629	0.03204	12.68	4.39e-06

Residual standard error: 36.69 on 7 degrees of freedom
Multiple R-squared: 0.9583
Adjusted R-squared: 0.9523
F-statistic: 160.8 on 1 and 7 DF
p-value:4.39e-06

Conclusion 5 :

Here the F test is significant and note that they adjusted r squared 0.9523 also high and the residuals are dispersed between -57.21 and 58.10 with residual standard error 36.69, which is low.
But the problem with the model is that the model only use 73% of the information not the total information. Also the degrees of freedom decreases as the data set is small and we remove 3 observations.

1.2.4 Weighted Least Squares

In the above model, as we are deleting 3 observations we are losing nearly 27% of information since the data set is small.
Alternative way to overcome the problem of loss of information, we may use weighted least square method to obtain the estimators.

Since the data consist of influential observations we will give weight to each of the squared errors based on their values. i.e. the large squared error will get less weight.
We may consider the weights as inversely proportional to the cook's distance values, with a the restriction that sum of the weights will be 1.

Now performing a weighted least square estimation method to the data we have the following result. As the model gives a insignificant intercept term we fit the model without the intercept term.

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
x_var	0.41827	0.01443	28.99	5.56e-11

Residual standard error: 5.794 on 10 degrees of freedom
Multiple R-squared: 0.9882
Adjusted R-squared: 0.9871
F-statistic: 840.4 on 1 and 10 DF
p-value:5.562e-11

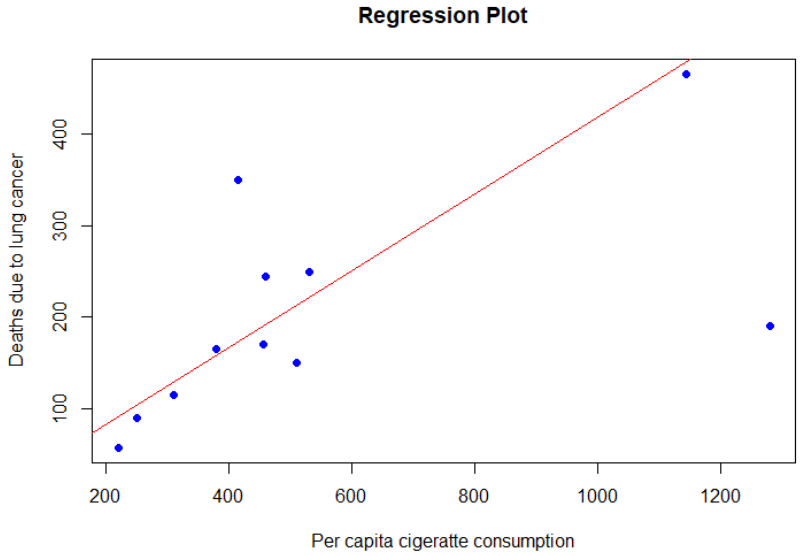
Conclusion 6 :

This model explains 98% variability of response variable with residual standard error 5.794 with 10 degrees of freedom and based on the full data set.
As this model works better then the OLS model after removing 3 observations, we'll stick to this model as we are considering the full data set here.

1.2.5 The Final Model

lung_cancer = (0.41827) $\times$ ***cigarette_consumption***

with *Adjusted R square: 0.98* and *Residual Standard Error: 5.794 with 10 degrees of freedom*



1.3 Summary of Conclusions

1. As we fit the data, the model comes out to be insignificant. There may be several reasons for this but the main reason may be that there are some influential observation as our scatter plot also shows that few values are far away from usual observations.

2. So for looking for influential observation and outliers, we studies the residuals and leverages and it comes out that the 4th, 6th and 11th observations has both the x outliers and y outliers.

3. Outlier doesn't always implies influential observation, so our next task was to identify influential observation and we calculate DFBETA, DFFIT and Cook's distance for that. It comes out that the 4th and 6th observations is highly likely to be influential observations and along with that the 11th observation may also an be influential observation.

4. We first remove the 4th of and 6th observation and refit the model on the data but the model again comes out to be insignificant so we need to drop the 11th observation now and refit the model again and after that we get a significant model without any intercept term.

5. The problem with the model is that now, the model is based on only 73% of the information and the degrees of freedom also decreases as we remove 3 observations from such a small data set. To come out this problem we may use weighted least square method instead of OLS method

6. On considering the weights inversely proportional to the Cook's distance values we came up with model without intercept term that explains 98% of the overall variability and the residual standard error of 5.794 with 10 degrees of freedom.

7. Since the last model is based on the full data set and also works better we will stick to that.

Problem 2

2.1 About the Practical

We have a data on light intensity and surface temperatures of 47 stars, both in logarithm, we have to fit a suitable regression model and find the outliers and influential observations.

2.2 Analysis

Given that light-intensity of star depends on its temperature,i.e. "light-intensity(y)" is the response variable and "surface temperature(x)" is the regressor here.

Analysis Required :

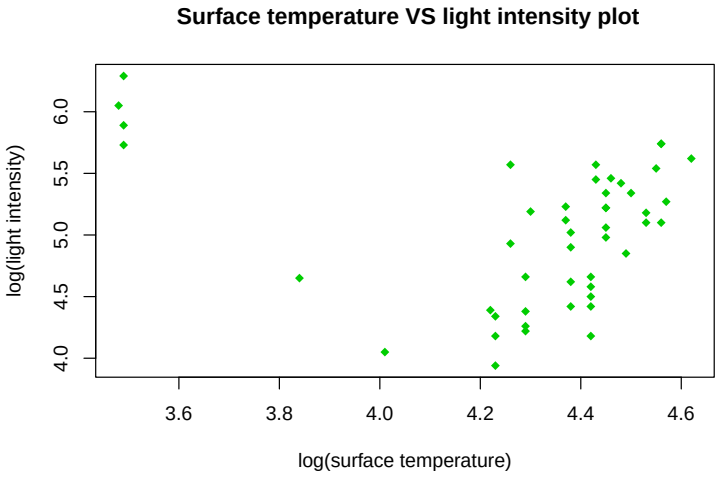
1. To fit a regression Model
2. Study the *Standardize* and *Studentize* residuals, and *Leverage*
3. Obtain the DFBETA's , DFFIT's and Cook's Distance

All the analysis required for the data is performed in lab, on suitable software and the result is discussed below.

2.2.1 Fitting Regression Model

Before fitting a model it is always better to have a visualization to the data so we create a scatter plot to observe the pattern in the data.

From the scatter plot we may noticed that there are few values which are far away from the other values, they may be impactful for the model, only by analysing these we can be sure about that.



Now, after fitting a simple linear regression model to the data we will come to the following conclusions,

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	6.7935	1.2365	5.494	1.75e-06
x_var	-0.4133	0.2863	-1.444	0.156

Residual standard error: 0.5646 on 45 degrees of freedom
Multiple R-squared: 0.04427
Adjusted R-squared: 0.02304
F-statistic: 2.085 on 1 and 45 DF
p-value: 0.1557

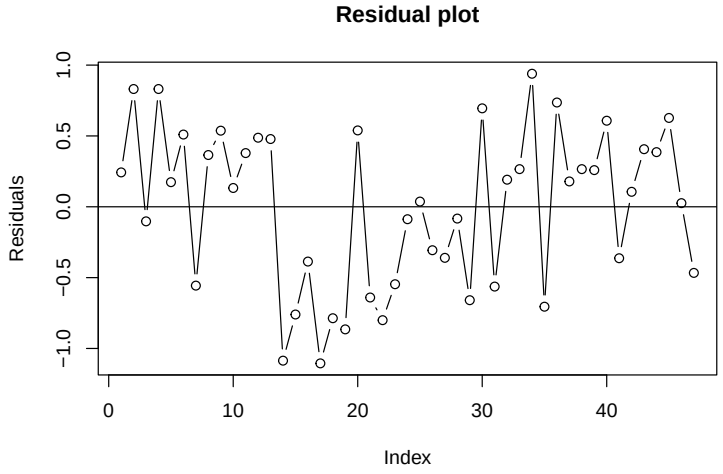
Conclusion 1 :

Note that, the F test is insignificant here as the p value is higher than 0.05. That is the parameters are insignificant which implies that there is no need to model. There are several reasons behind this but the main reason maybe the presence of outlier or influential observations as our scatter plots also shows that. So to fit a better model in the data, we need to identify the influential observations and which is done using the following measures.

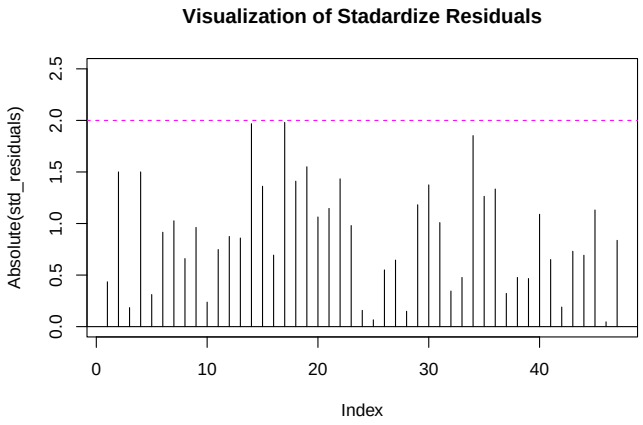
2.2.2 Detecting outlier and Influential Observations

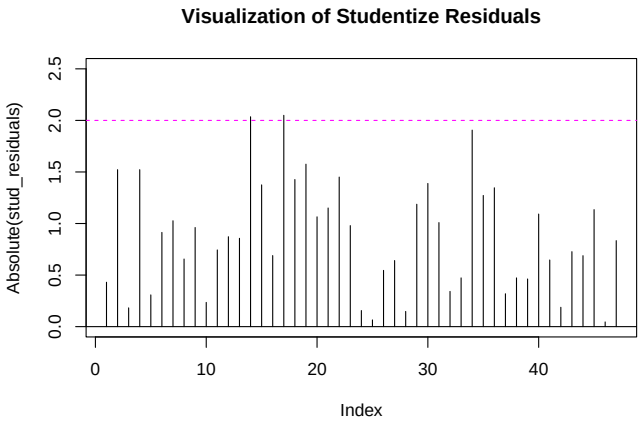
Detecting Residuals

We are proceeding in the next tasks to finding outlier by getting the residual plot of the above fit, as the above fit is insignificant so the residuals might be highly dispersed which displayed in the plot. Although it's quite difficult to study whether the residual causes outlier or not we calculate the standardised and studentize residuals for detecting higher residulas.



Standardize Residuals & Studentize Residuals





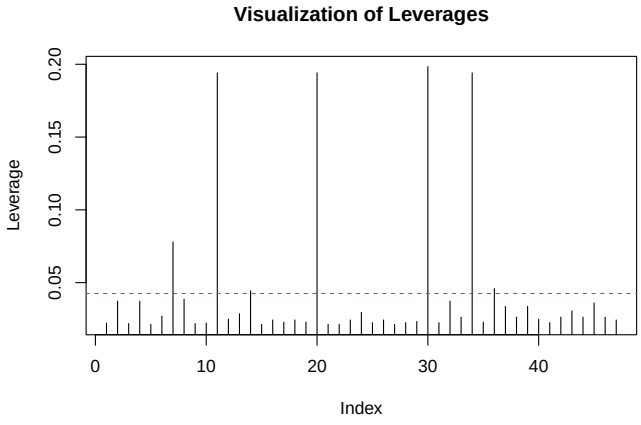
Conclusion 2 :

From the studentize and the standardize residual plot we can notice that there are three values which have higher residuals even that 2 of them also more than 2 hence we can say that this two are y-outliers. Those are the 17th and 14th observation also note that the 34th observation has studentize residual near to 2.

Detecting Leverages

Conclusion 3 :

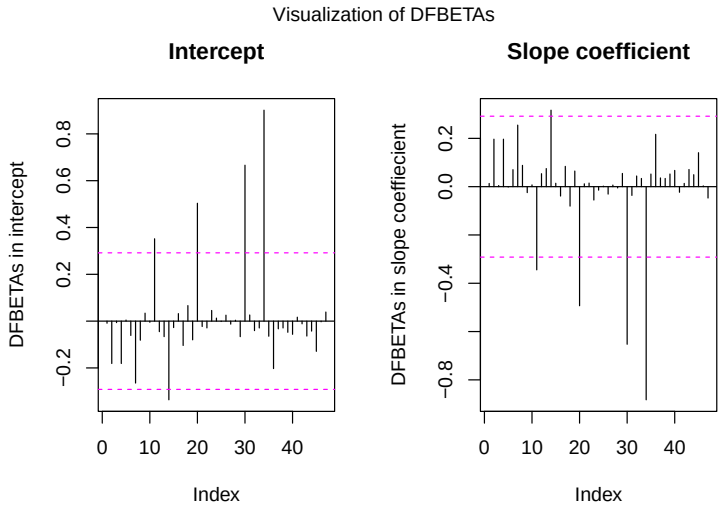
From the plot of the leverage we get 5 observations which has high leverage values even they also exits the required cut of point and so we can treat them as x-outliers. They are 7th, 11th, 20th, 30th, 34th observation



Calculating DFBETA's

Observation :

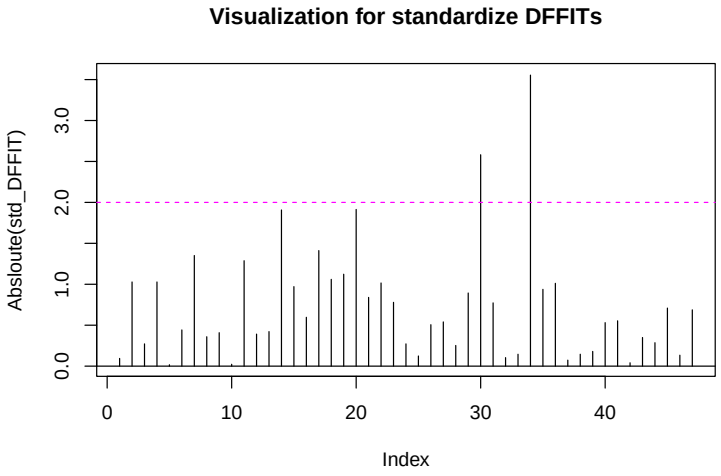
Although DFBETA is vector valued but if we plot them separately for each of the parameters we can notice that 5 of them has high DFBETA to values and so these values may influence the model. Also note that for each of the cases the 5 values are same those are 11th, 14th, 20th, 30th and 34th observations.



Calculating DFFIT's

Observation :

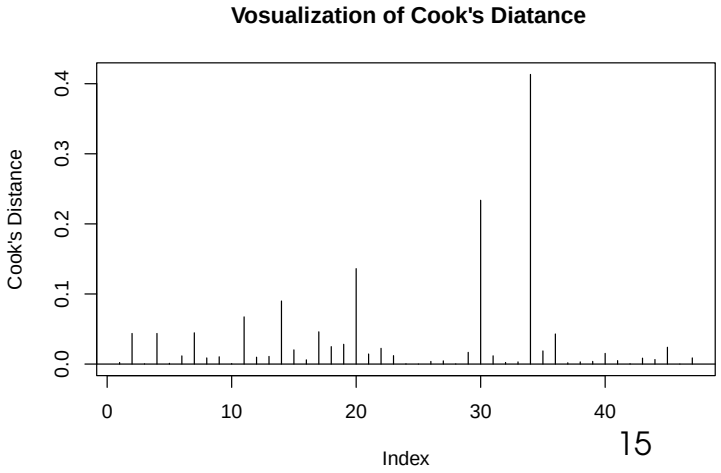
The 34th and 30th observations have higher DFFIT values as they crosses the usual cut off point for standardize DFFIT. They may be influential. Also note that the 20th and 14th observations have higher DFFIT values though they didn't cross the cut off.



Calculating Cook's Distance

Observation :

From the visualization of Cook's distance it is clear that the 11th, 14th, 20th, 30th and 34th observation has high Cook's distance which implies for these values the model doesn't fit well.



Conclusion 4 :

As we study DFBETAs, DFFITs to and cooks distance, the most common cases that the 11th, 14th, 20th, 30th and 34th observations influence the model so they might be influential observation for the model and need to diagnosed. Also note that view of this observations working as outlier observations too.

2.2.3 Diagnosing the Influential Observations

As we don't have the knowledge about the data so we can't cross verify these values and only left with nothing but dropping those values from the data and refitting the model.

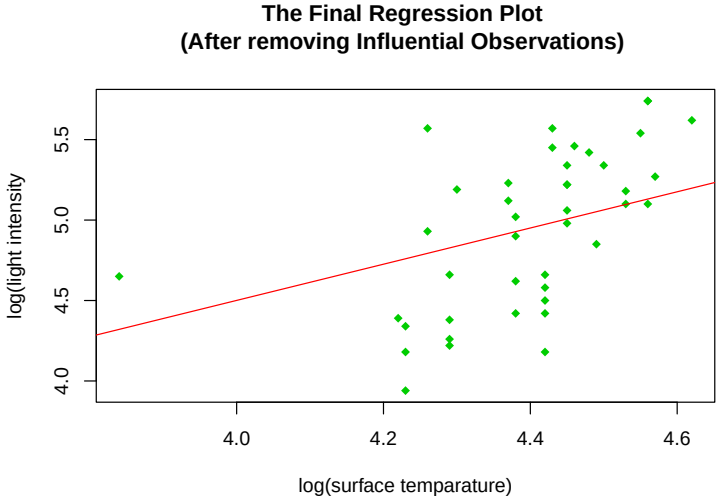
Dropping one of the influential observations at once, keep the model insignificant or perhaps even dropping any pair of this influential observations or any 3 of these influential observations keep the model insignificant so we drop the 11th, 14th, 20th, 30th and 34th observation from the data and refit the model.

Deleting this influential observation from the model gives a fit with significant F-statistic but the intercept is statistically insignificant so after ignoring the intercept term the final fitting becomes as following,

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
x_var	1.12516	0.01485	75.79	<2e-16

Residual standard error: 0.4231 on 41 degrees of freedom
Multiple R-squared: 0.9929
Adjusted R-squared: 0.9927
F-statistic: 5744 on 1 and 41 DF
p-value: < 2.2e-16



Conclusion 5 :

We'll stick to this model as the F-test is significant and this slope parameter is also significant. The residual standard error decreases from the first model and they adjusted at square increases to 0.99 which indicates a good fit, so you prefer to go with this model.

2.2.4 Weighted Least Squares

Since we are deleting 5 observations we are nearly removing 10% of the information. So another approach in such a situation is to use weighted least squares instead of ordinary least squares. i.e. we would give less importance to the higher errors values. As a choice we can choose weights in such a way that it would be inversely proportional to *cook's distance* values and the sum of the weights will be 1.

On performing Weighted least square method, with such weights we unfortunately end up with a very bad model, in which the explanatory variable has an insignificant coefficient and the model has a very bad Adjusted R^2.

Conclusion 6 :

So, in this case we avoid the weighted least square and took the OLS model after removing 5 influential observations.

2.2.5 The Final Model

log(light_intensity) = (1.12516) $\times$ log(surface_temperature)

with *Adjusted R square: 0.99* and *Residual Standard Error: 0.4231*

2.3 Summary of Conclusions

1. As we fit the data the model comes out to be insignificant. There may be several reasons for this but the main reason may be that there are some influential observation as our scatter plot also shows that few values are far away from usual observations.

2. So for looking for influential observation and outliers we studies the residuals and leverages and it comes out that the 14th, 17th and 34th observations has y-outliers and 7th, 11th, 20th, 30th, 34th observations are x-outliers .

3. Outlier doesn't always implies influential observation, so our next task was to identify influential observation and we calculate DFBETA, DFFIT and Cook's distance for that. It comes out that the11th, 14th, 20th, 30th and 34th observations is highly likely to be influential observations.

4. We first remove the influential observations each at a time, a pair at a time or even three of those influential observation at a time and refit the model on the data but the models again come out to be insignificant so we need to drop all the influential

observation and refit the model again and after that we get a significant model without any intercept term.

5. Here comes our final conclusion that, a data which contains influential observation may fit a model bad but if we detect influential observation and can remove them it will be a better fit.

Appendix

The overall analysis is performed in R software, and the necessary output is discussed here. The data set and the codes can be found in the following Github repository.

Source code : https://github.com/SoumaryaBasak/Regression_Analysis_1.git