Outliers and Influential Observations

Adam J Sullivan, PhD
02/12/2018

Leverage and Outliers

- We will now move onto leverage points and outliers.
- **Leverage point**: is a value of the predictor that is far from the average of the predictor variables.
- **Outlier points**: is a values of the outcome that is far from the average of the outcome.

Leverage and Outliers

- These two things together help us determine whether certain points have a lot of influence on our regression model.
- For example in Anscombe model 3 it appears that there is one point that not only is an outlier but may be a leverage point as well.
- Instead of trying to parse both of these concepts out we will focus on a plot that helps us consider influential points as a whole.

Cook's D

- Cook's distance attempts to tell us how much $\hat{\beta}$ changes due to the inclusion of the i^{th} observation.

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2}$$

DFFITS

- This quantity measures how much the regression function changes at the i-th case / observation when the i-th case / observation is deleted.
- For small/medium datasets: value of 1 or greater is “suspicious” (RABE). For large dataset: value of $2\sqrt{(p+1)/n}$.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$$

What is h_{ii}?

- In regression we have something we call the hat matrix, for a matrix X:

$$H = X(X^T X)^{-1} X^T$$

- We actually solve regression by performing this operation:

$$\hat{y} = Hy$$

What is h_{ii}?

- This ends up meaning that we have:

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{ii}y_i + \cdots + h_{in}y_n$$

DFBETAS

- This quantity measures how much the coefficients change when the i-th case is deleted.

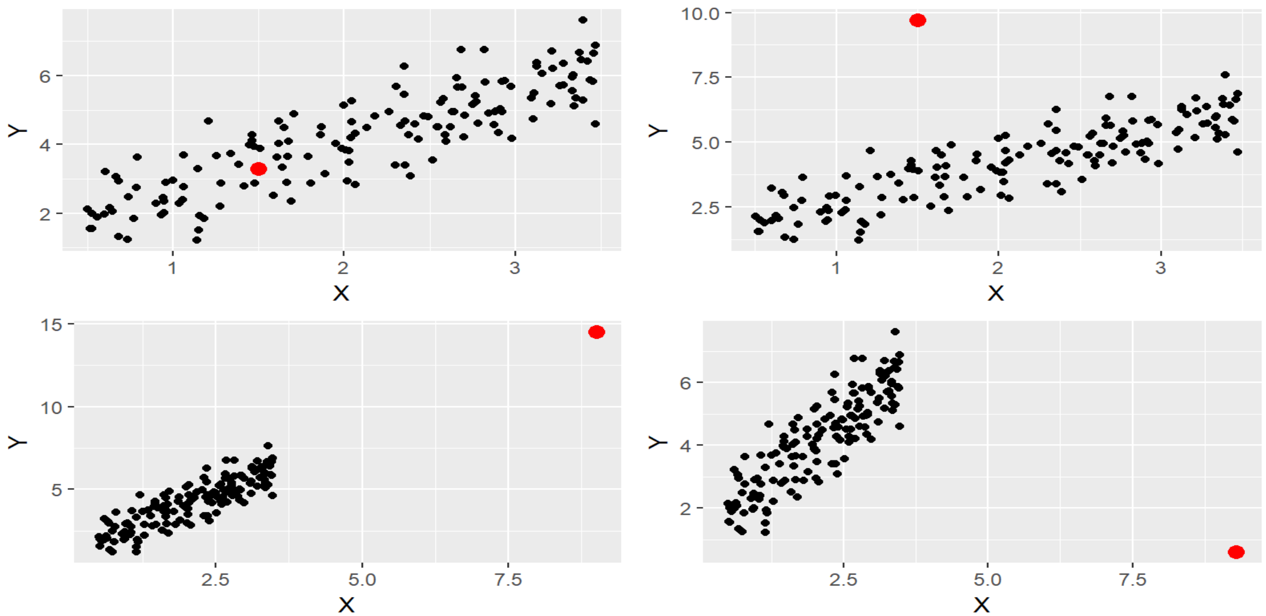
$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2_{(i)} (X^T X)^{-1}_{jj}}}$$

- For small/medium datasets: absolute value of 1 or greater is “suspicious”. For large dataset: absolute value of $2/\sqrt{n}$.

Simulating Data

```
set.seed(12345)
X = runif(150, .5, 3.5)
beta0 = 1.0
beta1 = 1.5
sigma = 0.7
Y = beta0 + beta1*X + sigma*rnorm(150) # The regular process
# Contaminated data: Four cases
X.suspect1 = 1.5; Y.suspect1 = 3.3
X.suspect2 = 1.5; Y.suspect2 = 9.7
X.suspect3 = 9.0; Y.suspect3 = 14.5
X.suspect4 = 9.3; Y.suspect4 = 0.6
Y.all1 = c(Y, Y.suspect1); X.all1 = c(X, X.suspect1)
Y.all2 = c(Y, Y.suspect2); X.all2 = c(X, X.suspect2)
Y.all3 = c(Y, Y.suspect3); X.all3 = c(X, X.suspect3)
Y.all4 = c(Y, Y.suspect4); X.all4 = c(X, X.suspect4)
```

Plots of Data



Run the 4 Regressions

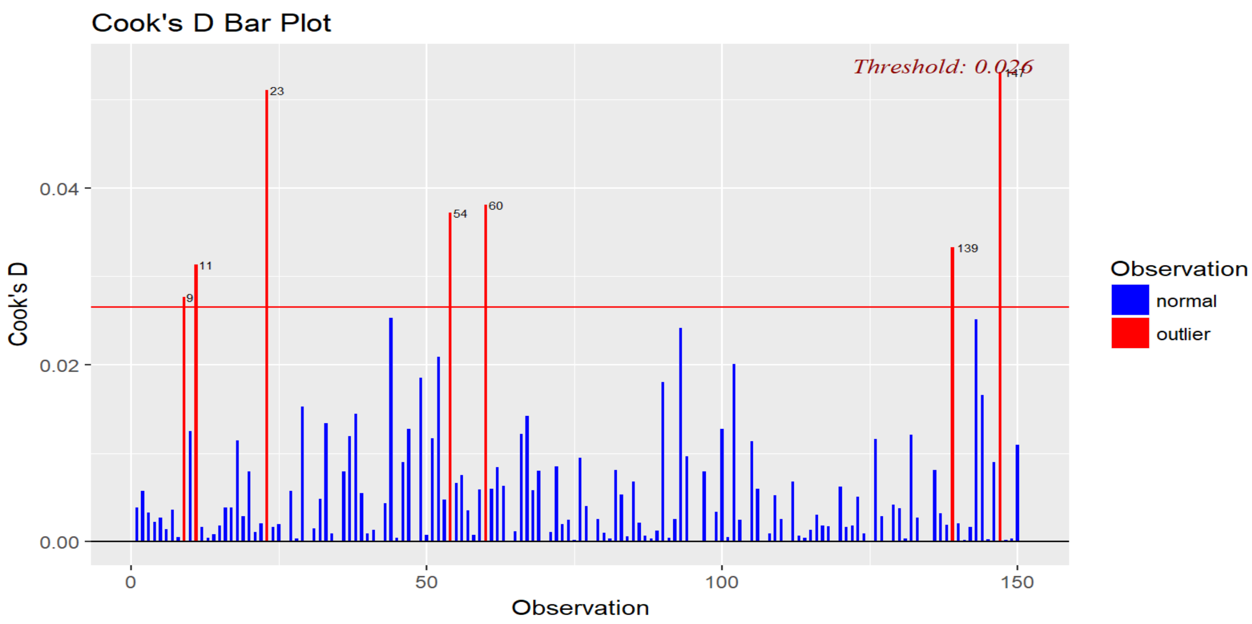
```
out1 <- lm(data=data, Y.all1~X.all1 )
out2 <- lm(data=data, Y.all2~X.all2 )
out3 <- lm(data=data, Y.all3~X.all3 )
out4 <- lm(data=data, Y.all4~X.all4 )
```

Outliers and Influential Points Plots

```
library(olsrr)
ols_cooksd_barplot(out1)
ols_dfbetas_panel(out1)
ols_dffits_plot(out1)
```

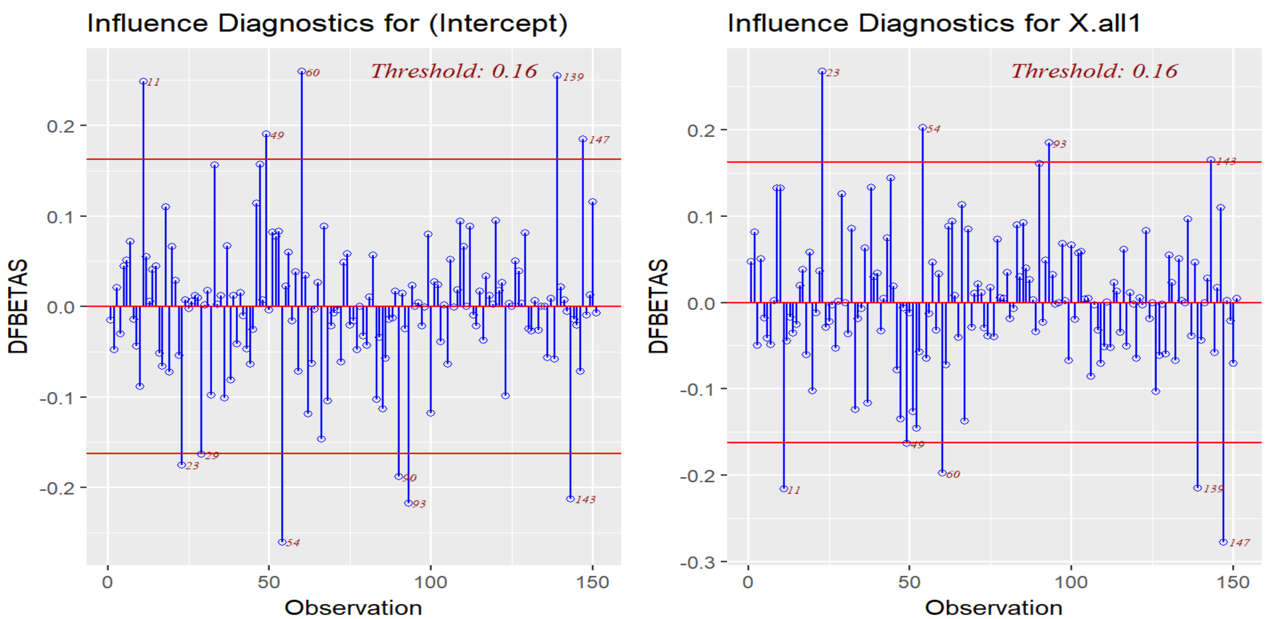
Outliers and Influential Points Plots: Cook's D

```
library(olsrr)
ols_cooksd_barplot(out1)
```



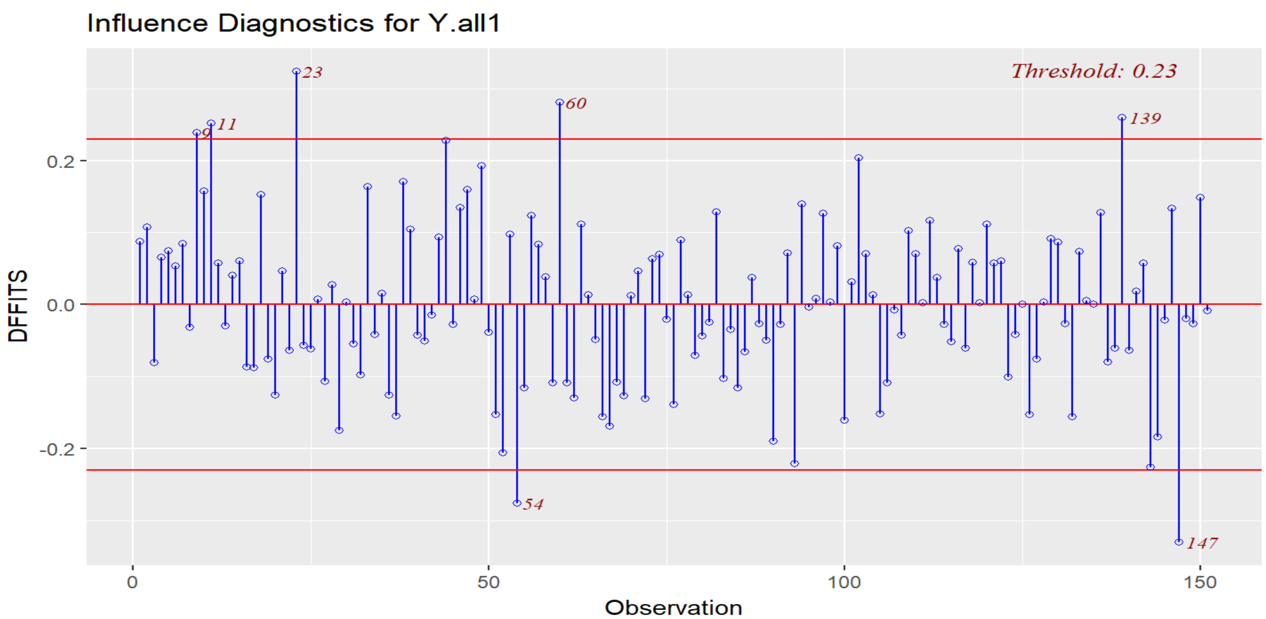
Outliers and Influential Points Plots: DFBETAS

```
library(olsrr)
ols_dfbetas_panel(out1)
```



Outliers and Influential Points Plots: DFFITS

```
library(olsrr)
ols_dffits_plot(out1)
```

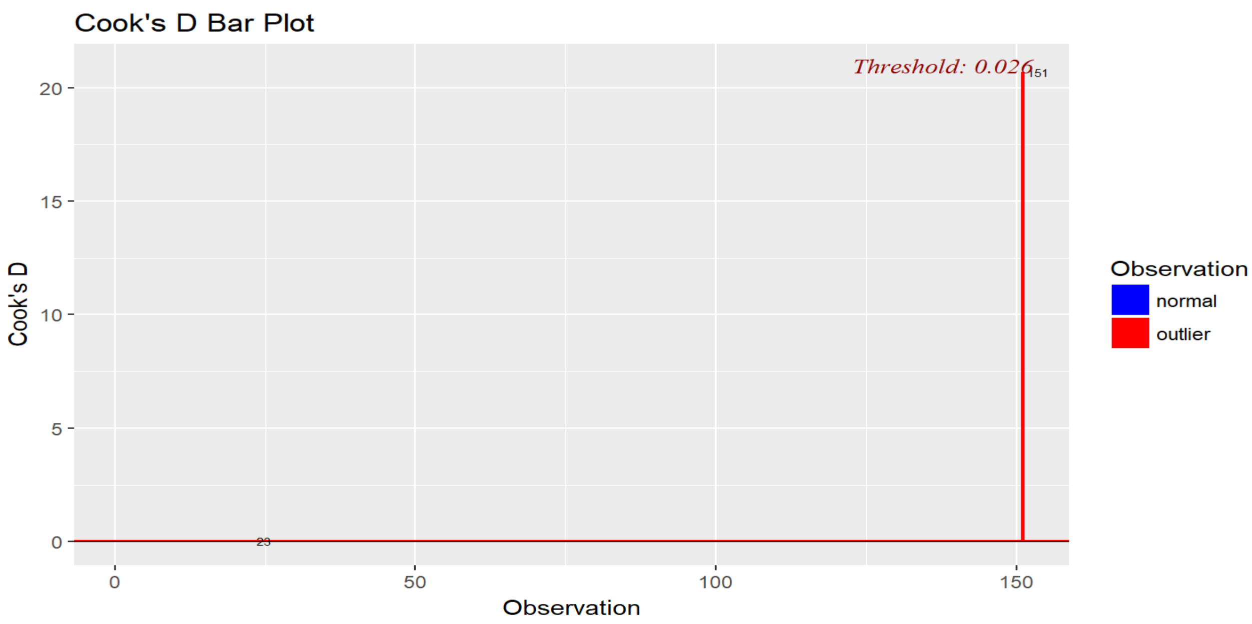


Outliers and Influential Points Plots

```
library(olsrr)
ols_cooksd_barplot(out4)
ols_dfbetas_panel(out4)
ols_dffits_plot(out4)
```

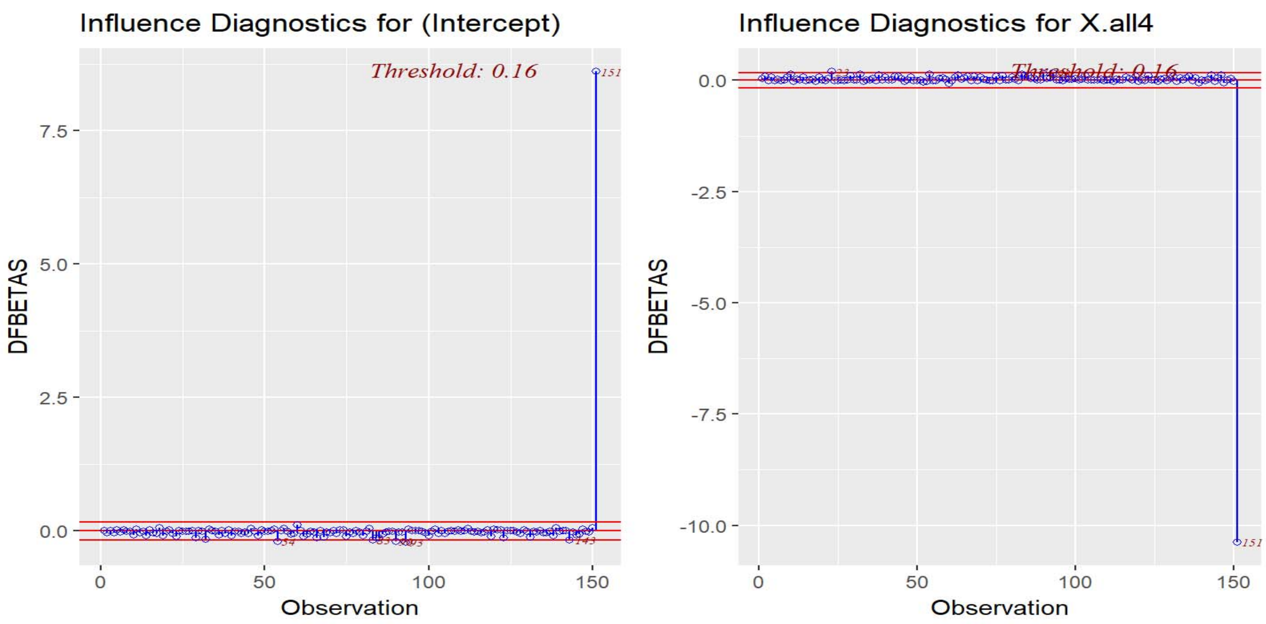
Outliers and Influential Points Plots: Cook's D

```
library(olsrr)
ols_cooksd_barplot(out4)
```



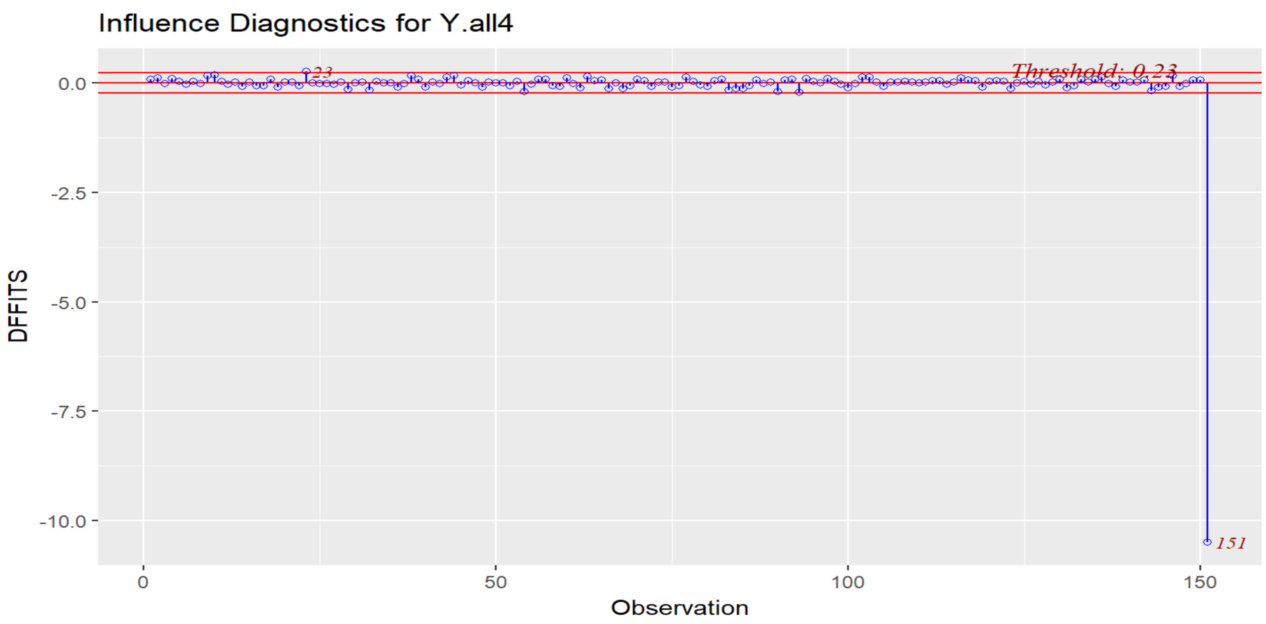
Outliers and Influential Points Plots: DFBETAS

```
library(olsrr)
ols_dfbetas_panel(out4)
```



Outliers and Influential Points Plots: DFFITS

```
library(olsrr)
ols_dffits_plot(out4)
```



What can we do with this point?

- We can decide to remove the point and re-run the regression.

```
library(broom)
out4a <- lm(data=data[-151,], Y.all4~X.all4 )

tidy4a <- tidy(out4a, conf.int = T)
tidy4 <- tidy(out4, conf.int = T)
knitr::kable(bind_rows(tidy4, tidy4a)[-c(3,4)])

glance4a <- glance(out4a)
glance4 <- glance(out4)
knitr::kable(bind_rows(glance4, glance4a))
```

What can we do with this point?

term	estimate	p.value	conf.low	conf.high
(Intercept)	2.3977868	0	1.9682861	2.827287
X.all4	0.8313212	0	0.6518475	1.010795
(Intercept)	1.2495210	0	0.9491648	1.549877
X.all4	1.4092337	0	1.2773737	1.541094

What can we do with this point?

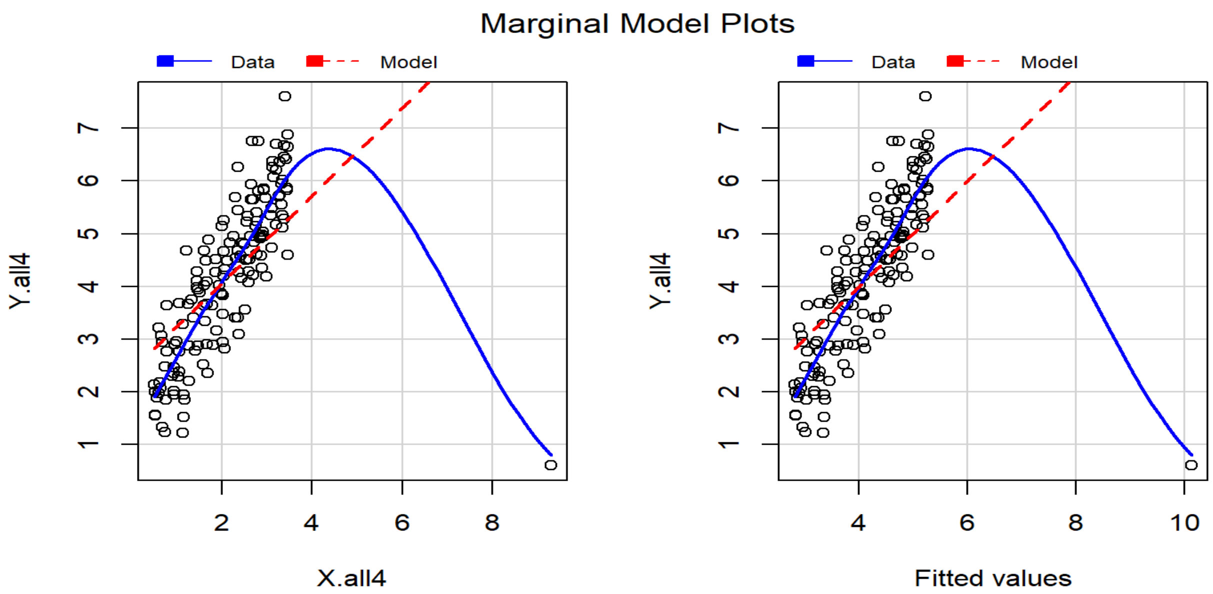
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.resi
0.3598977	0.3556017	1.1853180	83.7753	0	2	-238.9247	483.8494	492.9013	209.34182	
0.7508560	0.7491726	0.7272015	446.0339	0	2	-164.0513	334.1026	343.1345	78.26566	

Marginal Model Plots

- We will consider the next level of plots called Marginal Model Plots.
- The aim of these plots is to show how well out model fits the data.

```
  library(car)
mmps(out4)
```

Marginal Model Plots



What Can we See?

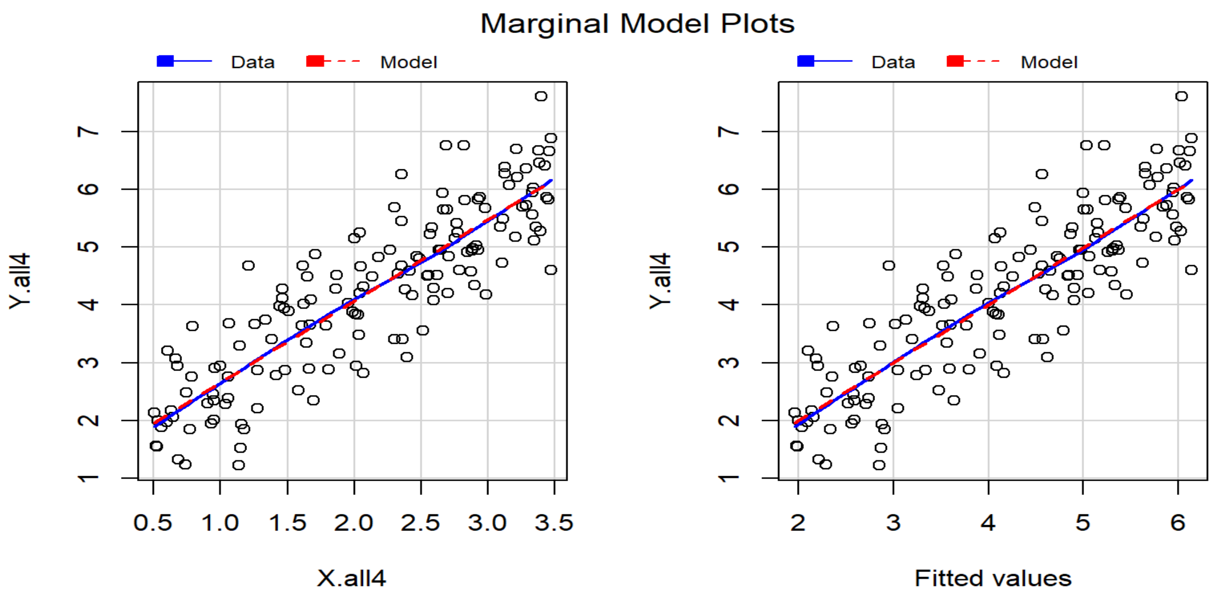
- From the figure we can see that the blue line represents a loess(smoothing) line for the data and the dashed line represents the model which R fitted.
- We can see that our data is very skewed by the outlier
- Also we can see that the loess line is more curved than our data model.

What happens when we Delete points?

- When we remove the point the difference is drastic

```
mmps(out4a)
```

What happens when we Delete Points??



Outlier Treatment

- Once the outliers are identified and you have decided to make amends as per the nature of the problem, you may consider one of the following approaches.
  1. Imputation
  2. Capping
  3. Prediction

Imputation

We can impute the value by replacing it with:

- mean
- median
- mode
- Other Regression techniques

We will consider this further in missing Data.

Capping

- For missing values that lie outside the 1.5*IQR limits, we could cap it by replacing those observations outside the lower limit with the value of 5th %ile and those that lie above the upper limit, with the value of 95th %ile.
- Below is a sample code that achieves this.

```
x <- dataframe$variable_of)interest
qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
caps <- quantile(x, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(x, na.rm = T)
x[x < (qnt[1] - H)] <- caps[1]
x[x > (qnt[2] + H)] <- caps[2]
```

Prediction

- In yet another approach, the outliers can be replaced with missing values (NA) and then can be predicted by considering them as a response variable.
- We will discuss this when considering missing data.