Multiple Linear Regression

Influential Observations

Contents

1. Influential Observation
2. Detecting Influential Observations
3. Cook’s Distance Method
4. DFBETAs
5. Influential Observations in R
6. Influence Plot in R

Outliers in Regression Model

- A regression outlier is an observation that has an unusual value of the dependent variable Y for given X value.

  Here X value may not be unusual.

- A regression outlier will have a large residual.

- The data may have an unusual X value — i.e., it is far from the mean of X.

- Regression outlier or unusual X value may not affect overall model. If both are true simultaneously then it is most likely to influence overall model.

  It is important to develop model which is not influenced by one or few observations.

Influential Observation

- An **influential observation** is an observation whose deletion from the dataset would noticeably change the result of the calculation.

- In regression analysis an influential data point is one whose deletion has a large effect on the parameter estimates or predictions.

Handling of Influential Observations

A single (or a few) observation may have a large influence on the results of regression analysis

It is important to develop a model which is not influenced by just few observations

The problem is encountered more commonly in small sample data

Detecting Influential Observations

Continuing with the “**Performance Index** “ data, we **Model** job performance index (**jpi**) based on aptitude score (**aptitude**), test of language (**tol**), technical knowledge (**technical**) and general information (**general**).

Use two methods to identify the influential observations: **“Cook’s Distance Method” and “DFBETAs”**.

Cook’s Distance Method

Cook's distance **measures the effect of deleting a given observation.**

Let Di be the Cook's distance for observation i.

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \text{ MSE}}$$

$\widehat{Y}_j$ = prediction from the full regression model for observation j

$\widehat{Y}_{j(i)}$ = prediction of j^{th} observation from a refitted model after removing i^{th} observation

MSE = mean square error of the regression model

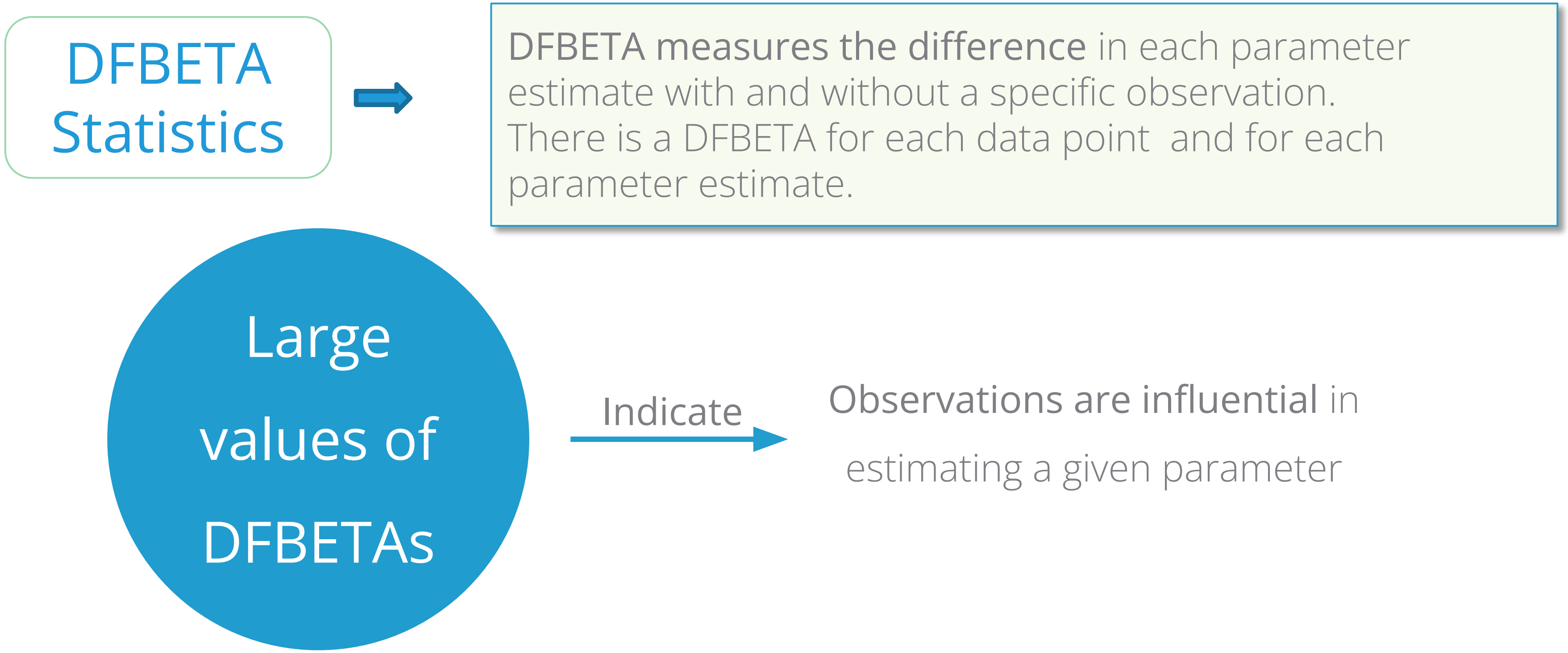
p = number of fitted parameters in the model

Cut off to indicate influential observation,

- Simple operational guideline D_i>1
- Alternative D_i >4/n, where n is the number of observations

* It is recommended to check model performance by excluding highly influential observation

DFBETAs



DFBETA Statistics

DFBETA measures the difference in each parameter estimate with and without a specific observation. There is a DFBETA for each data point and for each parameter estimate.

Large values of DFBETAs

Indicate

Observations are influential in estimating a given parameter

Cut off to indicate influential observation,

- general cut off value recommended is **2**
- size adjusted cut off is taken to be **2/√n**

Influential Observations in R

```
#Importing the Data
```

```
perindex<-read.csv("Performance Index.csv",header=T)
jpimodel<-lm(jpi~aptitude+tol+technical+general,data=perindex)
```

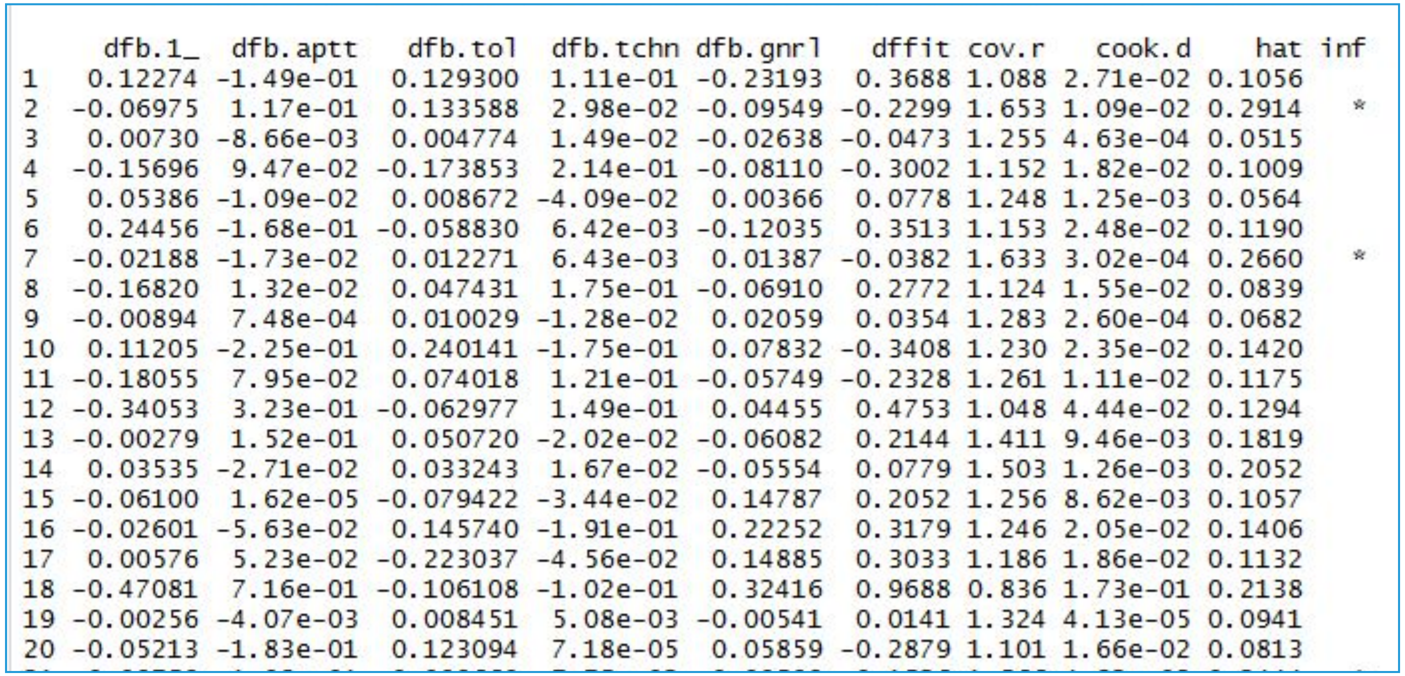
```
#Finding Influential Observations
```

```
influ<-influence.measures(jpimodel)
influ
```

- **influence.measures()** produces a class "inf" object
- tabular display showing the DFBETAs for each model variable, DFFITS, covariance ratios, Cook's distances and the diagonal elements of the hat matrix.

Influential Observations in R

```
# Output
```



```
      dfb.1_  dfb.aptt   dfb.tol  dfb.tchn dfb.gnrl   dffit cov.r   cook.d    hat inf
1    0.12274 -1.49e-01  0.129300  1.11e-01 -0.23193  0.3688 1.088 2.71e-02 0.1056
2   -0.06975  1.17e-01  0.133588  2.98e-02 -0.09549 -0.2299 1.653 1.09e-02 0.2914   *
3    0.00730 -8.66e-03  0.004774  1.49e-02 -0.02638 -0.0473 1.255 4.63e-04 0.0515
4   -0.15696  9.47e-02 -0.173853  2.14e-01 -0.08110 -0.3002 1.152 1.82e-02 0.1009
5    0.05386 -1.09e-02  0.008672 -4.09e-02  0.00366  0.0778 1.248 1.25e-03 0.0564
6    0.24456 -1.68e-01 -0.058830  6.42e-03 -0.12035  0.3513 1.153 2.48e-02 0.1190
7   -0.02188 -1.73e-02  0.012271  6.43e-03  0.01387 -0.0382 1.633 3.02e-04 0.2660   *
8   -0.16820  1.32e-02  0.047431  1.75e-01 -0.06910  0.2772 1.124 1.55e-02 0.0839
9   -0.00894  7.48e-04  0.010029 -1.28e-02  0.02059  0.0354 1.283 2.60e-04 0.0682
10   0.11205 -2.25e-01  0.240141 -1.75e-01  0.07832 -0.3408 1.230 2.35e-02 0.1420
11  -0.18055  7.95e-02  0.074018  1.21e-01 -0.05749 -0.2328 1.261 1.11e-02 0.1175
12  -0.34053  3.23e-01 -0.062977  1.49e-01  0.04455  0.4753 1.048 4.44e-02 0.1294
13  -0.00279  1.52e-01  0.050720 -2.02e-02 -0.06082  0.2144 1.411 9.46e-03 0.1819
14   0.03535 -2.71e-02  0.033243  1.67e-02 -0.05554  0.0779 1.503 1.26e-03 0.2052
15  -0.06100  1.62e-05 -0.079422 -3.44e-02  0.14787  0.2052 1.256 8.62e-03 0.1057
16  -0.02601 -5.63e-02  0.145740 -1.91e-01  0.22252  0.3179 1.246 2.05e-02 0.1406
17   0.00576  5.23e-02 -0.223037 -4.56e-02  0.14885  0.3033 1.186 1.86e-02 0.1132
18  -0.47081  7.16e-01 -0.106108 -1.02e-01  0.32416  0.9688 0.836 1.73e-01 0.2138
19  -0.00256 -4.07e-03  0.008451  5.08e-03 -0.00541  0.0141 1.324 4.13e-05 0.0941
20  -0.05213 -1.83e-01  0.123094  7.18e-05  0.05859 -0.2879 1.101 1.66e-02 0.0813
```

Interpretation:
Higher the cook's distance, more is the influence of observation on the model.

Influential Plot in R

```
#Influence Plot
```

```r
install.packages("car")
library(car)

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

influencePlot() creates a “bubble” plot of Studentized residuals by hat values, with the areas of the circles representing the observations proportional to Cook’s distances.

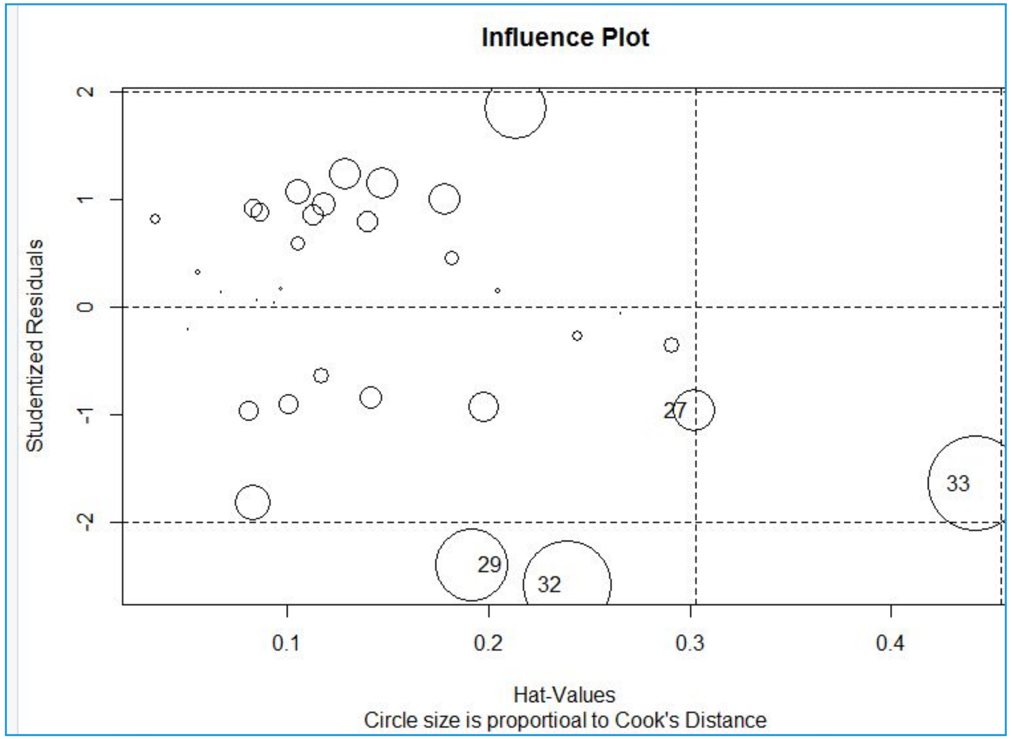
id.method="identify**"** enables interactive point identification.

main = Title for plot

sub = X axis label

Influence Plot in R

```
# Output
```



Interpretation:
The data points 27, 29, 32, 33 are detected as influential observations.

Quick Recap

In this session, we learnt what are **influential observations in regression analysis:**

Influential Observations

- Having a few observations influence the results of regression analysis is not desirable

How to Calculate Influential Observations

- Such influential observations can be calculated via two most widely used methods. Cook's Distance and DFBetas

Cook's Distance

- Cook's distance measures the effect of deleting a given observation.

DFBetas

- DFBETA measures the difference in each parameter estimate with and without the influential point.

Influential Observations in R

- **`influence.measures()`** produces object giving influential observations by different measures
- **`influencePlot()`** creates a "bubble" plot of Studentized residuals by hat values, with the areas of the circles representing the observations proportional to Cook's distances