Cook’s Distance

Seminar 4

Iliass Tiendrebeogo

Overview

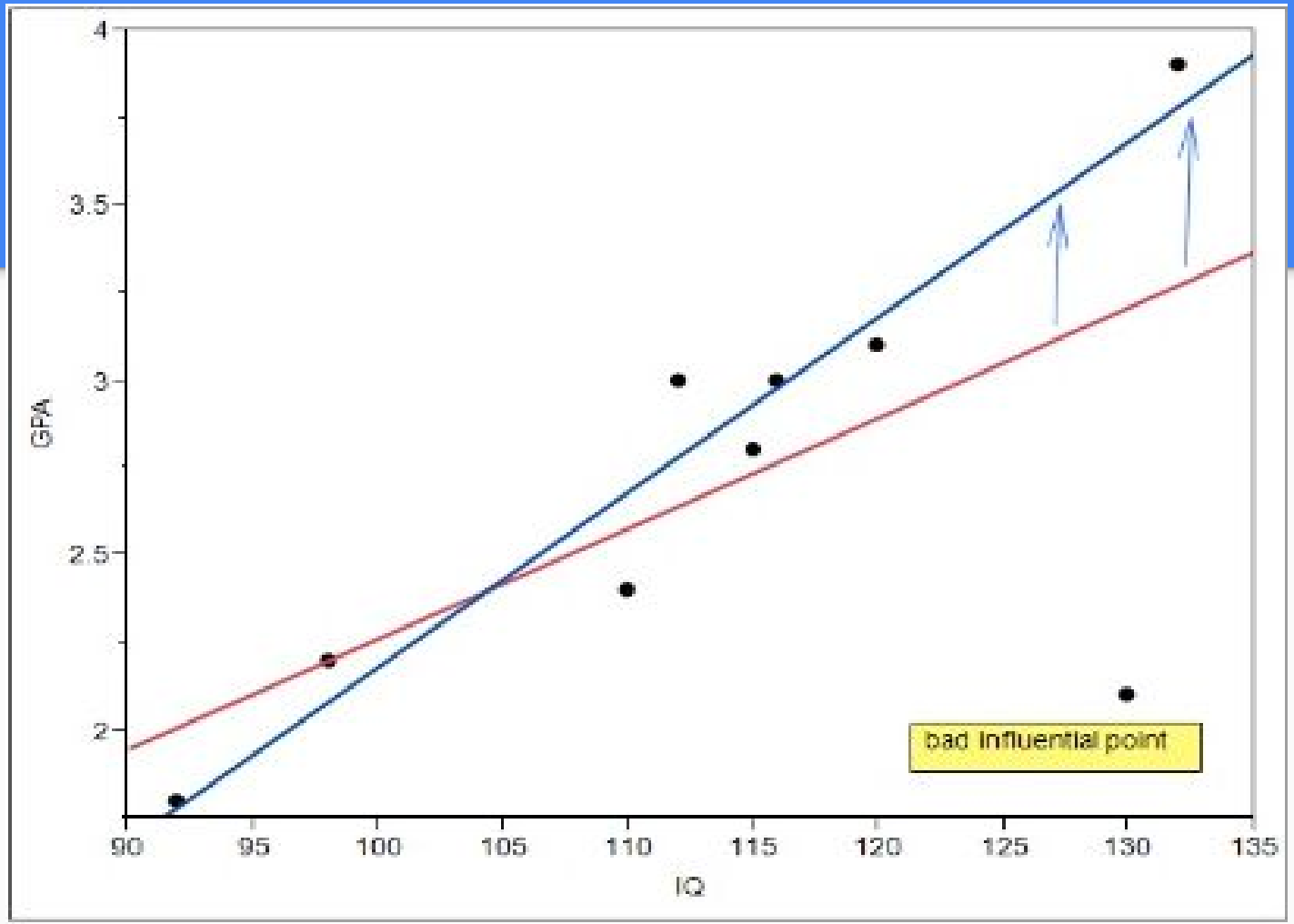
In Regression analysis, Cook's Distance is used to estimate the influence of a data point over the model when performing the ordinary least square regression method.

It was introduced by the American statistician R. Dennis Cook in 1977

Highly Influential data point

A data point is said to be influential if when removed from the calculation change the regression line significantly.

Data point with high leverage can be influential it is an outlier. However, a data point can have an high leverage but not influential, and it goes the same way for an outlier(all outlier are not influential).


4
3.5
3
2.5
2
GPA
90
95
100
105
110
115
120
125
130
135
IQ
bad influential point

Cook’s Distance

Cook's distance measures how much a parameter estimate change when a data-point is removed from the calculation.

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{kMSE}$$

Interpretation

Widely used interpretation is : $D > 1$

Several other rules have been defined such as:

$D > 8.5$ if $k > 3$;

$D > 4/n$

Cook’s Distance using R

To build the model:

```
lm_model<-lm(sales ~ adverts + airplay +
attract, data = album2)
```

Call cook’s distance builtin function:

```
cooks.D <- cooks.distance(lm_model)
```

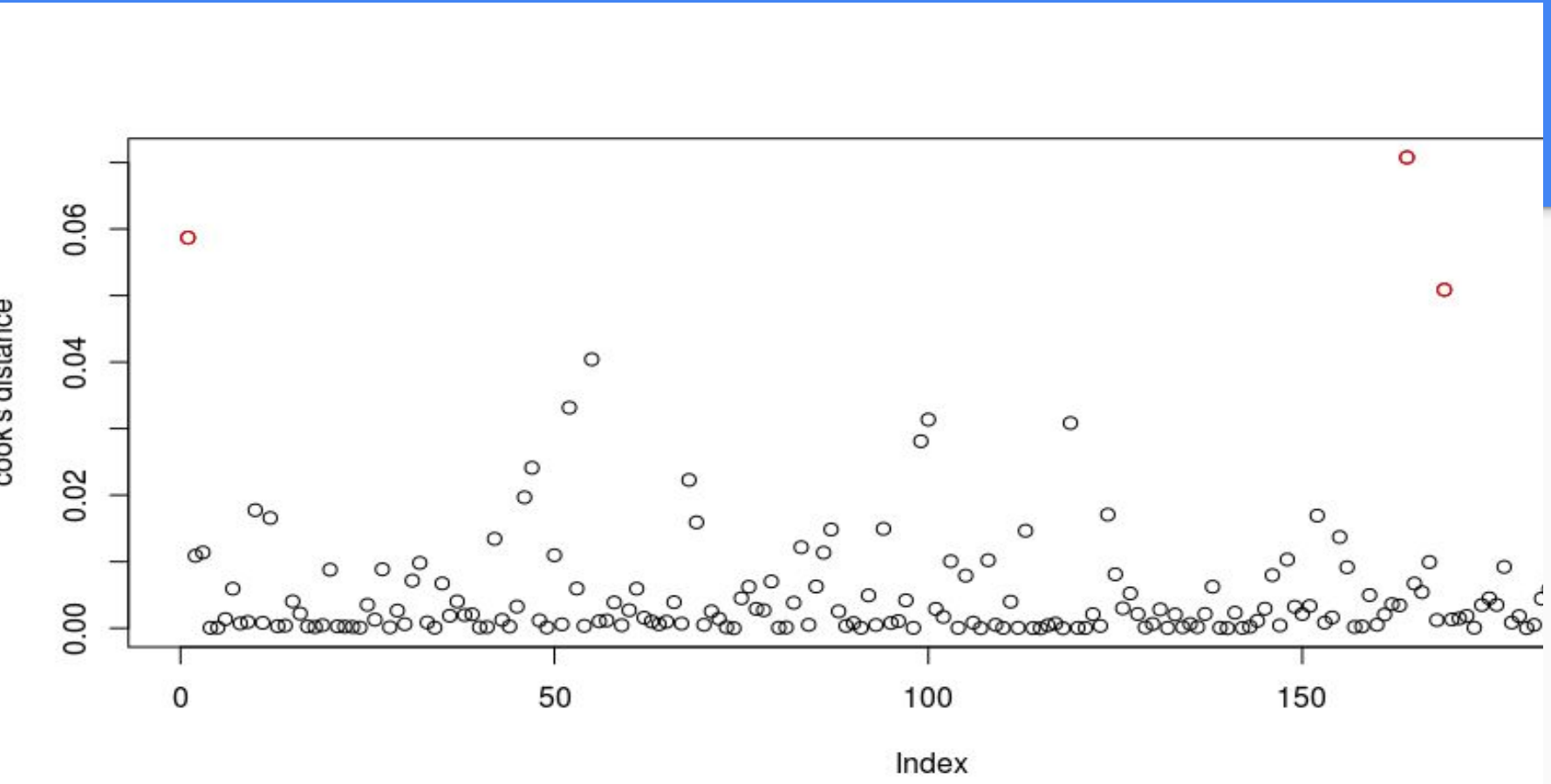
Cook’s Distance using R

determine influential points by applying the threshold of 0.05

```
influential <- cooks.D > 0.05
```

we plot cook's Distance vector

plot(cooks.D, ylab = "cook's distance")


Cook's distance
0.00
0.02
0.04
0.06
0
50
100
150
Index

Discussion

What should we do when a given data-point's D > threshold?:

- Check whether the given point is influential or not?
- Two different reports
- Omit the influential point