Iliass Tiendrebeogo
March 2, 2016

# Seminar 4 - Cook's Distance
# Math 567: Winter 2016

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 R. Dennis Cook in 1977. [Cook, 1977]

## 1 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).

The **leverage** define how far apart is a given data point from the average(mean/median). Points with high leverage tend to pull the regression line toward themselves and have impact on the slop of the regression line hence **influential**.

## 2 Cook's Distance

Cook's distance measures how much a parameter estimate change when a datapoint is removed from the calculation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance needs closer examination in the analysis.

The algebraic expression.

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

Where:
$\hat{Y}_j$ is the prediction from the full regression model for observation $j$;
$\hat{Y}_{j(i)}$ is The prediction for observation $j$ from a refitted regression model in which observation $i$ has been omitted;
$k$ is the number of prediction parameter of the regression model;
$MSE$ is the mean square error.

# 3 Interpretation of Cook's Distance

There are several rules when interpreting **cook's distance**. The widely used criterion is that a point is considered to be highly influential if $D_i > 1$ [Stanford Weisberg, 1982]

Different rules have been defined such as: $D_i > 0.85$ if $p > 3$ [McDonal, 2002] where $p$ is the number of regression parameter. [Bollen, 1990] declares a data-point to be influential when $D_i > \frac{4}{n}$ where $n$ is the number of observation.

# 4 Cook's Distance using R

In this example we use album sales2 data from the book "Discovering Statistics Using R" [Andy Field, 2012]
First we build the fitted regression model after loading the data

```
album2<-read.delim("Album Sales 2.dat", header = TRUE)
# run the multiple regression model
lm_model<-lm(sales   adverts + airplay + attract, data = album2)
# analyse the cook's D by calling the built-in function cooks.distance
cooks.D <- cooks.distance(lm_model)
# determine influential points by applying the threshold of 0.05
influential <- cooks.D > 0.05
we managed to isolate points at positions 1, 164, 169
```

Now we plot cook's Distance vector

```
plot(cooks.D, ylab = "cook's distance")
# point with high D_i will be shown red
points(1, cooks.D[1], col = 'red')
points(164, cooks.D[164], col = 'red')
points(169, cooks.D[169], col = 'red')
```

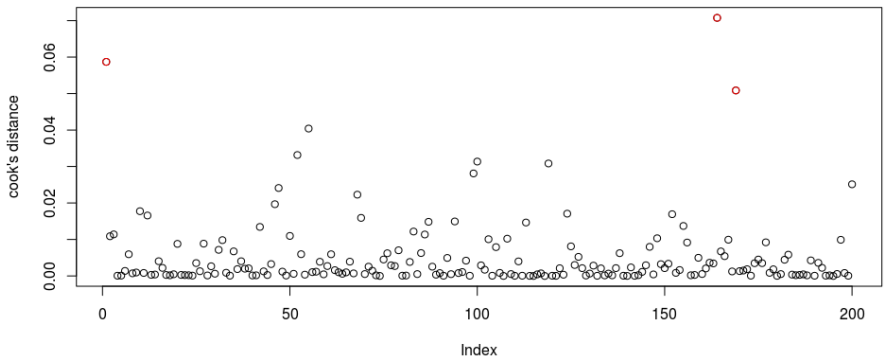


Figure 1: Cook's Distance on 200 album sales data

# 5 Discussion

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

First we should make sure that the observation is recorded correctly.
Do the outcome change when the data is removed from the calculation? if the outcome do not change, it's not a big deal, but if the outcome change then this a problem that we need to fix. We don't want our statistical model relaying on a single data-point.

In that case we could reports two different results , one whit the influential data and another one without the the influential data.

Or we could restrict the analysis to values of $X$ for which the relationship holds. Which is usually not a good idea since we are omitting a data-point even though we would report that a outlier has been removed from the calculation.

# References

Zo Field Andy Field, Jeremy Miles. *DISCOVERING STATISTICS USING R*. Sage Publucation, Ltd., 2012. URL `http://studysites.sagepub.com/dsur/main.htm`.

Kenneth; Robert W. Bollen. *Modern Methods of Data Analysis*. Newbury Park, 1990.

R. Dennis Cook. *Detection of Influential Observation in Linear Regression*. Taylor and Francis, Ltd., 1977. URL http://www.jstor.org/stable/1268249.

Barry McDonal. A teaching note on cooks distance, 2002.

R. Dennis Cook Stanford Weisberg. Residuals and influence in regression, 1982. URL https://books.google.com/books?id=MVSqAAAAIAAJ.