A2Q3

Influence

Influence is commonly used in outlier detection. When one variate y_u that might be away from the location where most of variates are at, we want to test how large the impact of that variate on a particular attribute is. We have a special way of testing it, we can obtain the **influence** of that variate by removing it, then calculate the difference of attribute with y_u and without y_u:

$$\delta(a,u) = a(y_1, y_2, ..., y_{u-1}, y_u, y_{u+1}, ..., y_N) - a(y_1, y_2, ..., y_{u-1}, y_{u+1}, ..., y_N)$$

Note that a variate has a large influence when it is recorded as an error or it plays an important part in the whole population, even in the whole dataset. When there's no variate with large influence, the population is ideal and this kind of situation is desired.

Example

Data: Wine Quality (Click here)

From wine quality, we can examine total sulfur dioxide:

The standard deviation of the population is $a(y_1, ..., y_N) = \sqrt{\frac{\sum_{u\in\mathcal{P}}(y_u-\bar{y})^2}{N}}$:

```
wine <- read.csv("winequality-red.csv", header = T)
y <- wine$total.sulfur.dioxide
N <- length(y)
ysd <- sd(y)*sqrt((N-1)/N)
ysd
```

```
## [1] 32.88504
```

and $\delta(a, v)$, the influence for a given unit v (we use v to avoid confusion), is:

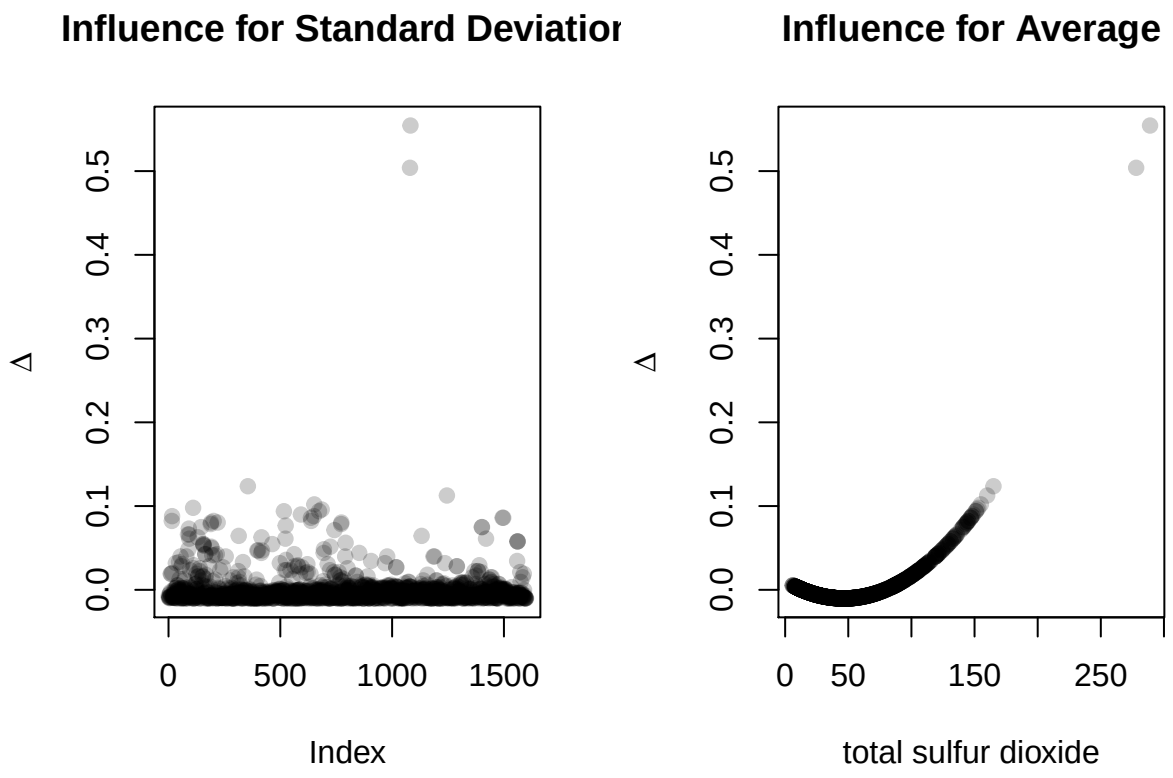
$$\delta(a,v) = \sqrt{\frac{\sum_{u\in\mathcal{P}}(y_u-\bar{y})^2}{N}} - \sqrt{\frac{\sum_{u\in\mathcal{P}}(y_u-\bar{y})^2 - ((y_v-\bar{y})^2)}{N-1}}$$

We can use R loop to calculate the delta value:

```
delta = rep(0, N)
for (i in 1:N) {
  delta[i]=ysd - sd(y[-i])*sqrt((N-2)/(N-1))
}
```

Then, we can plot the influence for every unit v by observation number or by y:

```
par(mfrow= c(1,2))
plot(delta,main ="Influence for Standard Deviation",pch =19,
     col =adjustcolor("black",alpha =0.2),
     xlab ="Index",ylab =bquote(Delta))
plot(y, delta,main ="Influence for Average",pch =19,
     col =adjustcolor("black",alpha =0.2),
     xlab ="total sulfur dioxide",ylab =bquote(Delta))
```



We can see that there are two units that are more influential on the standard deviation than other units

We can extract those two highest influential points using R code:

```
which(delta>0.4)
```

```
## [1] 1080 1082
```

```
y[which(delta>0.4)]
```

```
## [1] 278 289
```

We see that two highest influential points are at index 1080 and 1082 with 278 and 289 total sulfur dioxide respectively. These numbers are the rightmost points in the right graph.