Assignment 2 Question 1

Sheen Thusoo

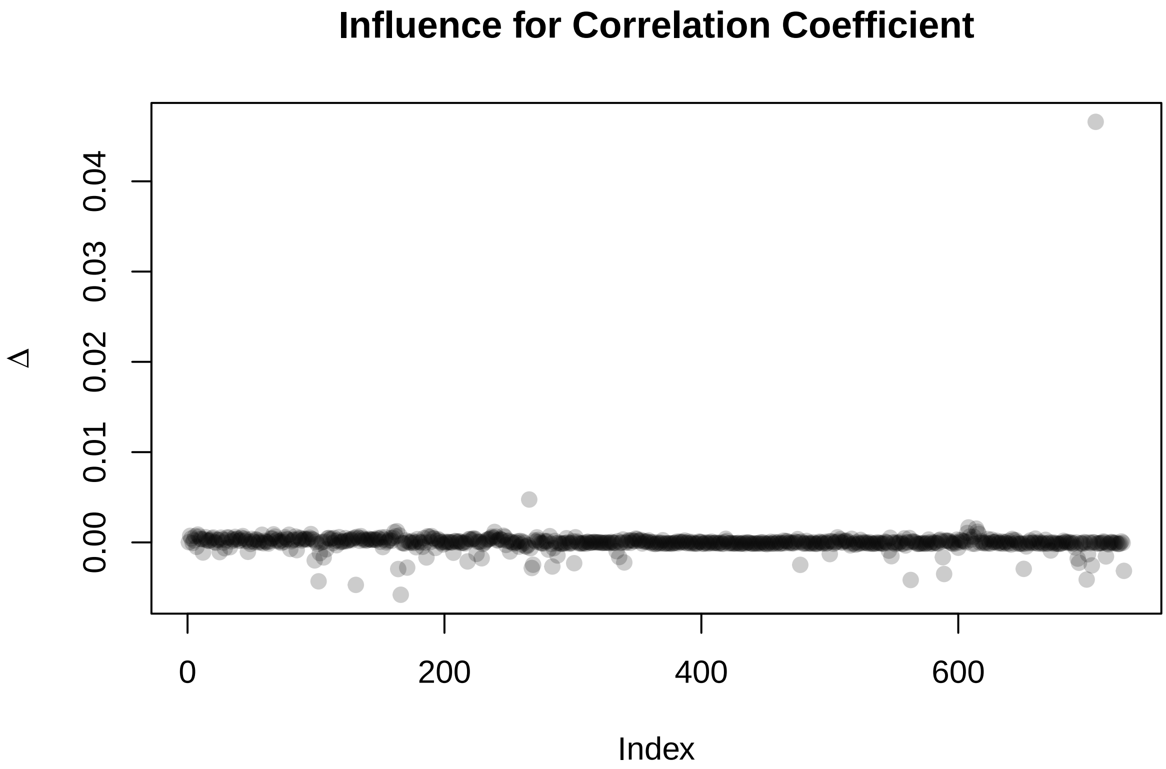
Part d)

```r
data <- read.csv("EconomicMobility.csv")
```

```r
correlation_coefficient <- function(pop_y, commute_y) {
  cor(pop_y, commute_y, method = "pearson", use='complete.obs')
}

delta = numeric(length(data$Population))
for (i in 1:length(data$Population)) {
  ## y[-i] removes the ith element from a vector
  delta[i] = correlation_coefficient(data$Population, data$Commute) -
    correlation_coefficient(data$Population[-i], data$Commute[-i])
}
```

```r
plot(delta, main="Influence for Correlation Coefficient", ylab=bquote(Delta), pch=19,
     col=adjustcolor("black", alpha = 0.2))
```



Are there any influential points? If so, determine if there is anything interesting about them.

```
data[which(delta > 0.04),]
```

```
##            Name  Mobility State Population Commute Longitude Latitude
## 707 Los Angeles 0.0960902    CA   16393360   0.225 -116.2887 34.07517
```

```
delta[which(delta > 0.04)]
```

```
## [1] 0.04658742
```

Are there any influential points? If so, determine if there is anything interesting about them. There is an influential point in the data which has an influence value that is much greater than other points (around 0.046). From the code chunk above, this point is in Los Angeles where the *Population* variate is 16393360 and the *Commute* variate is 0.225. However, compared to other cities in the dataset with a similar population value, the Commute variate is much lower; this is an outlier since it is much higher than what is expected. Thus, the correlation coefficient is much different than the others and the influence is much greater for this data point.