|  |  |
| --- | --- |
| **Module:** | ST2053 |
| **Name:** | Maxim Chopivskyy |
| **Student Number:** | 118364841 |
| **Chapter:** | 3 |

**Maximum 2 pages! Do not delete the page number in the footer.**

(a)

A residual is the difference between an observed data point and the estimated value for what that data point should have been.

injection.df <- read.table("P:\\ST2053\\Previous Exam Datasets\\18-19\\Injection.txt", header = T)

injection.lm <- lm(Conc ~ Time+Dose+Weight+Age, data = injection.df)

e <- resid(injection.lm)

e[1]

It over estimates concentration by 4.335488

(b)

In a simple linear model, case i is said to be an outlier if the corresponding residual is large.

An outlier is detected by finding the studentized residual of the cases. If the absolute value of the ith Studentized residual exceeds 2, case i is said to be an outlier.

I would expect there to be around 15 outliers (300 \* 0.05). This is because the studentized residual[-2, 2] is approximately a 95% confidence interval.

s <- summary(injection.lm)$sigma

r <- e / (s\*(1-h)^0.5)

r[abs(r) > 2]

14 cases are outliers.

(c)

Case i is said to be a case of high leverage if the corresponding x-value, xi, is far from the mean of the x-values. Basically, if the x-value is unusually large or unusually small.

p <- length(coef(injection.lm))

leverage\_cutoff <- (2\*p)/length(injection.df$Conc)

h <- lm.influence(injection.lm)$hat

h[h > leverage\_cutoff]

There are 5 cases with high leverage.

The most extreme case has high leverage because it’s predictor x-value is unusually large.

(d)

d <- (1/p)\*(h/(1-h))\*r^2

sort(d)

Case 22 has highest influence with a Cook’s distance of 0.03812386

It has high influence because it’s an outlier(2.435550 > 2) and it almost has high leverage(0.031134105).

(e)

I would recommend to further investigate the data points with high influence.