|  |  |
| --- | --- |
| **Module:** | ST2053 |
| **Name:** | Cian Linehan |
| **Student Number:** | 119301381 |
| **Chapter:** | 3 |

**(a)**

> exec.lm = lm(Salary ~ Experience + Education + Profits + Sales, data=exec.df)

> exec.resid = resid(exec.lm)

> exec.resid[1]

1

8899.441

A residual is the difference between the observed value and fitted value (on the regression line) for the dependent variable (Y), residual = observed - fitted. Here = 8899.441 = observed - fitted. This means the regression model under-estimated the salary of case 1 by €8899.441.

**(b)**

An outlier is detected by calculating the studentized residuals for each individual case as residuals do not have the same variance. The formula for this is . If the absolute value of the studentized residual for the i-th case is greater than 2, then we can say this case is an outlier. This is reasonable because the errors and therefore residuals are normally distributed, since we have scaled them to have a variance of 1, they now follow the standard normal distribution. 2 is a reasonable cut off point as it is roughly the cut-off point for the 95% confidence interval (1.96) of which we can expect most values (95%) to be reasonably located within. A value of greater than 2 suggests it is an extreme case. I would expect to find 5% of the cases to be outliers; .05 x 166 = 8.3, so roughly 8 outliers. There was found to be 7 outliers as found below.

> h = lm.influence(exec.lm)$hat

> s = summary(exec.lm)$sigma

> r = (exec.resid)/(s\*(1-h)^.5)

> counter = 0

> for(res in r)

+ {if(res > 2 | res < -2)

+ counter = counter + 1}

> counter

[1] 7

**(c)**

A case is said to be of high leverage if the corresponding x-value is far from the mean of the x-values, i.e. if the x-value is unusually larger or smaller than the mean of x-values. In multiple regression, it measures the distance between the vector of x-values for the i-th case and the vector of mean values of each variable.

> plot(h, type='h')

> identify(h, n=1)

[1] 43

> exec.df[43,]

Salary Experience Education Profits Sales BoardMember

43 156500 24 12 8 5 No

> summary(exec.df$Experience)

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 8.00 13.00 13.22 20.00 26.00

> summary(exec.df$Education)

Min. 1st Qu. Median Mean 3rd Qu. Max.

12.00 14.00 16.00 15.88 18.00 20.00

> summary(exec.df$Profits)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 6.00 8.00 7.59 9.00 11.00

> summary(exec.df$Sales)

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.000 6.000 8.000 7.596 9.000 12.000

Highest leverage was found to be in case 43. This is likely because, as shown above, experience is close to the max of 26 years whilst education is the absolute minimum of the data set which are two opposite extrema of x-values. Profits and sales are both equal to the median here and are not a factor in making this a case of high leverage. See figure 1 for plot.

**(d)**

A case is a case of high influence if omitting the case from the data set would cause large changes in the intercept and slope, i.e. it has a disproportionate effect on slope and intercept. Case #43 has the highest influence with a cooks distance of .08987667, this is likely because it has the highest leverage (of h = .115 as shown in (c)) but also a studentized residual of 1.85 which is not above the cutoff point of 2 but is nevertheless quite high. The cutoff point for high leverage in this model is .06 = (2\*5/166), which again supports the fact that #43 has high leverage as it is over the cutoff. See figure 2 for plot.

> p = length(coef(exec.lm))

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

> plot(d,type='h')

> identify(d,n=1)

[1] 43

> tableD.df = data.frame(exec.df[,2],r,h,d)

> tableD.df[43,]

exec.df...2. r h d

43 24 1.858357 0.1151417 0.08987667

**(e)**

I would recommend that case 43 is investigated thoroughly as it has been shown it is of high leverage, high influence and very close to being an outlier. If it is discovered that a mistake has been made in the data, then it should be omitted. Otherwise the model should be ran both with and without case 43 and analysis taken from both, but if no error can be found in the data then the case should not be omitted.

![Chart, histogram

Description automatically generated]() ![Chart, histogram

Description automatically generated]()

Figure 1 Figure 2