|  |  |  |  |
| --- | --- | --- | --- |
| DNSC 6213-10 Statistics for Analytics II | Lab 4 | GWlogoBlue | Student ID**:**  **G****27279521** |

**(This assignment is to be completed individually. While you may interact with others to gain mastery of the generic course content, all work associated with this assignment must be strictly and exclusively yours, including creating the R code and determining the correct answers/completing the template; the only collaboration permitted is with the instructor.)**

Ergonomic Systems Inc. (ESI) wishes to identify key variables which are related to sales. They have collected the following variables from a random sample of customers: sales, age, income, education, and tenure. “sales” represents the amount of purchases by the customer in the past year. “tenure” represents the number of years that the customer has been listed in ESI’s database. One additional variable, ID, has been collected: this represents the unique identification number of the customer.

ESI wishes to begin by determining which observations are outliers and which are influential.

1. From the gvlma function, can we be reasonably certain that Sales is non-linearly related to the set of independent variables? (Yes or No) No Why/Why not? the p-value is bigger than 0.05 which means that failed to reject the null hypothesis. Global Stat is the test for the dependent variable's linear or not. So, Dependent variable is a linear combination of independent variables
2. From the gvlma function, can we be reasonably certain that the residuals in the population are skewed? (Yes or No) No Why/Why not? From gvlma function, the p-value of Skewness is bigger than 0.05 which means that failed to reject the null hypothesis. The null hypothesis is the residuals are not skewed in the population.
3. From the gvlma function, can we be reasonably certain that the residuals in the population deviate from a normal distribution? (Yes or No) No Why/Why not? the p-value is bigger than 0.05. Failed to reject null hypothesis which is the distribution of the residuals in not kurtotic in the population.
4. From the gvlma function, can we be reasonably certain that the residuals in the population are heteroscedastic? (Yes or No) No Why/Why not? the p-value is bigger than 0.05. Failed to reject Null hypothesis which is residuals is homoscedastic in the population
5. Please list here the ID numbers of all observations (if any) which are X-outliers (based on the hat values) 5(9921), 38(10663),85(10305) are outliers.
6. Please list here the ID numbers of all observations (if any) which are Y-outliers (based on “rstudent”) 25(10166),28(10805),37(10231),50 (10195),55(10566),76(9670),79(10375),82 (9716),86(9776),89(10126),96 (11636),99(7905) are the outliers
7. Please list here the ID numbers of all observations (if any) which are influential with respect to the intercept (based on dfbetas): 15(10632),42(10011),76(9670),89(10126)
8. Please list here the ID numbers of all observations (if any) which are influential with respect to the slope associated with age (based on dfbetas): 42(10011),50(10195),82(9716),99(7905)
9. Please list here the ID numbers of all observations (if any) which are influential with respect to the slope associated with income (based on dfbetas): 76(9670)
10. Please list here the ID numbers of all observations (if any) which are influential with respect to the slope associated with education (based on dfbetas): 42(10011),50(10195),76(9670),89(10126)
11. Please list here the ID numbers of all observations (if any) which are influential with respect to the slope associated with tenure (based on dfbetas): 50(),76(9670),82(9716),89(10126),99(7905)
12. Please list here the ID numbers of all observations (if any) which are influential with respect to the set of independent variable slopes (based on Cook’s D): There is not any ID numbers that has influential with respoect to the set of indepent variable slopes.

Paste your full R script immediately below this line:

# Lab4

# install.packages("gvlma")

library(gvlma)

head(a4,10)

tail(a4,10)

a4.lm <- lm(sales~age+income+education+tenure, data = a4)

summary(a4.lm)

# question 1~ question 4

a4.gvl <- gvlma(a4.lm)

summary(a4.gvl)

# questino 5

hatvalues(a4.lm)

hv <- as.data.frame(hatvalues(a4.lm))

colnames(hv) <- c("hatvalues")

mn <- mean(hv$hatvalues)

hv$warn <- ifelse(hv$hatvalues >3\*mn, 'x3',

ifelse(hv$hatvalues > 2\*mn, 'x2', '-'))

subset(hv, warn %in% c("x2","x3"))

hv[order(hv$hatvalues),]

# question 6

rstudent(a4.lm)

rs <- as.data.frame(rstudent(a4.lm))

colnames(rs) <- "rstudent"

critval <- qt(.95, nrow(rs)-2-1)

rs$warn <- ifelse(abs(rs$rstudent)>critval,

'Warn', '-')

subset(rs, warn == "Warn")

rs[order(rs$rstudent),]

# question 7 ~ question 11

dfbetas(a4.lm)

dfb <- as.data.frame(dfbetas(a4.lm))

critval <- 2/sqrt(nrow(dfb))

dfb$Warn <- ifelse(abs(dfb)>critval, "Warn", "-")

subset(dfb, Warn[,1] == "Warn" | Warn[,2] == "Warn")

# question 12

cooks.distance(a4.lm)

cd <- as.data.frame(cooks.distance(a4.lm))

colnames(cd) <- "CooksD"

critival <- qf(.50,2,nrow(cd)-2)

cd$warn <- ifelse(abs(cd$CooksD)>critival,

'Warn', '-')

subset(cd, warn=="Warn")

cd[order(cd$CooksD), ]