Problem Set 5

QTM 200: Applied Regression Analysis

Due: March 4, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in R, please include the code you used to get your answers. Please also include the .R file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.

- Your homework should be submitted electronically on the course GitHub page in .pdf form.

- This problem set is due at the beginning of class on Wednesday, March 4, 2020. No late assignments will be accepted.

- Total available points for this homework is 100.

Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

```
1 gamble <- (data=teengamb)
2 # run regression on gamble with specified predictors
3 model1 <- lm(gamble ~ sex + status + income + verbal, gamble)
```

Answer the following questions:

(a) Check the constant variance assumption for the errors by plotting the residuals versus the fitted values.

```
1 # a) Check the constant variance assumption for the errors by plotting the
       residuals
2 #    versus the fitted values.
3 model1.resid <- resid(model1)
4 pdf("ResidFVplot.pdf")
5 plot(model1, model1.resid, which = 1)
```

```
6  dev.off()
7
8  # The line of the graph is not very straight, indicating that there might
       not be a linear
9  # relationship among residuals. It also appears that residuals aren't
       necessarily constantly
10 # variable along the line— they seem to have increasing variance as y
       increases.
```

.

.

(b) Check the normality assumption with a Q-Q plot of the studentized residuals.

```
1 # b) Check the normality assumption with a Q-Q plot of the studentized
      residuals.
2 pdf("QQplot.pdf")
3 plot(model1, which = 2)
4 dev.off()
5
6 # The tails deviate from the diagonal line, having "heavier" values in the
      upper and lower
7 # ends of the line and creating a steeper-looking slope. This suggests
      that our data might
8 # not be distributed normally, and actually have more variation.
```

(c) Check for large leverage points by plotting the h values.

```
1 # c) Check for large leverage points by plotting the h values.
2
3 identify(gamble$sex, gamble$status, gamble$income, gamble$verbal, row.
      names(gamble))
4 hat <- lm.influence(model1)$hat
5 pdf("hvalueplot.pdf")
6 plot(hat)
7 abline(h=2*5/47)
8 abline(h=3*5/47)
9 identify(1:47, hat, row.names(gamble))
10 dev.off()
11
12 # There appears to be four large leverage points above the 3(k +1)/n
      threshold, even though
13 # for some reason they are not labelled with their row name.
```

(d) Check for outliers by running an `outlierTest`.

```
1 # d) Check for outliers by running an outlierTest.
2 outlierTest(model1, data=gamble)
3 # The p-value is 4.1041e-07, which is very low. We can conclude that this
      is an extreme residual.
```

.

(e) Check for influential points by creating a "Bubble plot" with the hat-values and studentized residuals.

```r
# e) Check for influential points by creating a "bubble plot" with the hat
      values and
#    studentized residuals.
pdf("bubbleplot.pdf")
plot(hat, rstudent(model1), type = "n")
cook <- sqrt(cooks.distance(model1))
points(hat, rstudent(model1), cex=10*cook/max(cook))
abline(h=c(2, 0, 2))
abline(v=c(2,3)*5/47)
identify(hat, rstudent(model1), row.names(gamble))
dev.off()

# There are a few points that seem to be influential in the model, but for
      some reason they are
# unlabelled.
```