Problem Set 5

QTM 200: Applied Regression Analysis

Vanessa Wong

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.

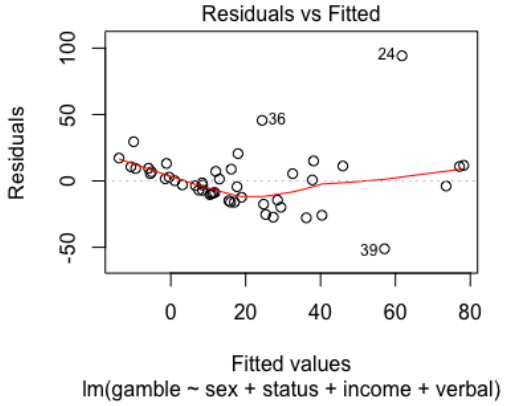
```
######################

# load data
```

Answer the following questions:

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

```
# a. Residuals vs. Fitted Values
par(mfrow=c(2,2)); plot(model1)
plot(model1, which=1)
```

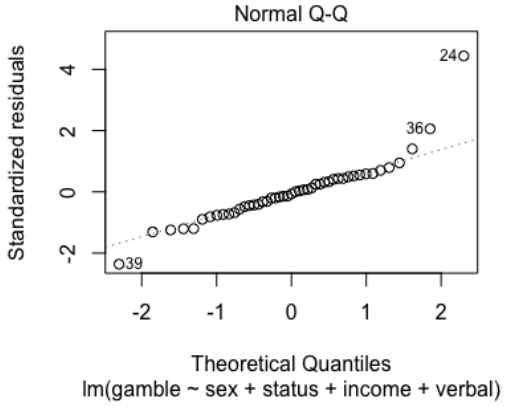
- After fitted value (y hat) equals 45, variance seems to increase/residuals deviate more from 0. variance does not appear to be constant.



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

```
# b. Q-Q plot of studentized residuals
plot(model1, which=2)
```

- The normality assumption appears to be met because almost all points are very tightly clustered along the QQ line.

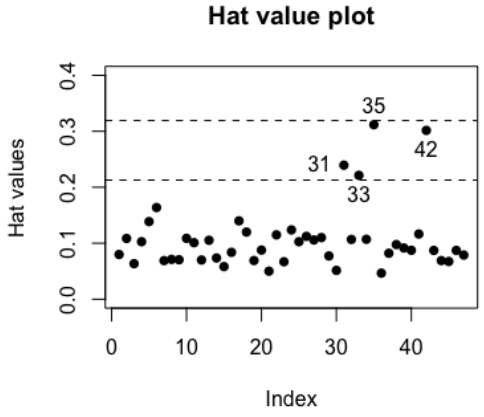


(c) Check for large leverage points by plotting the h values.

```
# c. Plot hat values
plot(hatvalues(model1), pch=16, cex=1, ylab="Hat values", main="Hat value
    plot", ylim=c(0,0.4))
abline(h=(2*5)/47, lty=2)
abline(h=(3*5)/47,lty=2)
identify(1:47, hatvalues(model1), row.names(gamble))
```

- Using the average hat value as a threshold (thresholds = 2 * hbar and 3 * hbar), there are 4 high leverage points in this dataset: observations 31, 33, 35, and 42. These four points

have are potentially influential i.e. have the potential to greatly affect the regression model.



-

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

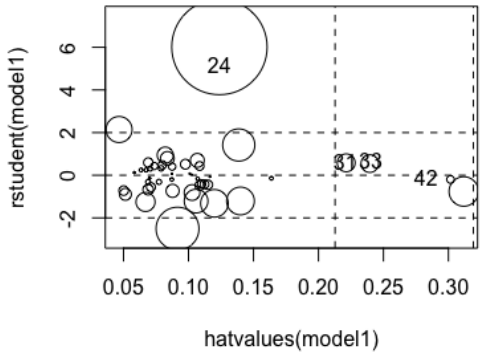
```
outlierTest(model1)
```

- According to outlier test, observation 24 has an adjusted (Bonferroni) p-value 0. therefore, observation 24 is an extreme residual in this model.

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

```
plot(hatvalues(model1), rstudent(model1), ylim=c(-3,7.5), type="n")
cook <- sqrt(cooks.distance(model1))
points(hatvalues(model1), rstudent(model1), cex=10*cook/max(cook))
abline(h=c(-2,0,2), lty=2)
abline(v=c(2,3)*5/47, lty=2)
identify(hatvalues(model1), rstudent(model1), row.names(gamble))
```

- Observations 31, 33, 35, and 42 are influential points because all 4 are above at least one threshold for studentized residuals and hat values. This means all 4 points are both high leverage and are regression outliers, making them influential points (they greatly affect the regression model).Observation 24 was also identified because while it did not pass any of the hat value thresholds, it has an unusally large Cook's distance.



-

```
#########################
# load libraries
# set wd
# clear global .envir
#########################

# remove objects
rm(list=ls())
# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices
    ", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1,
    TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package,
    character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg,  dependencies = TRUE)
  sapply(pkg,  require,  character.only = TRUE)
}

# here is where you load any necessary packages
install.packages(car)
library(car)
library(boot)
# ex: stringr
# lapply(c("stringr"),  pkgTest)

lapply(c("faraway"),  pkgTest)
```

```
34
35 # set working directory
36 setwd("~/GitHub/QTM200Spring2020/problem_sets/PS5")
37
38
39 ######################
40 # Problem 1
41 ######################
42
43 # load data
44 gamble <- (data=teengamb)
45 # run regression on gamble with specified predictors
46 model1 <- lm(gamble ~ sex + status + income + verbal, gamble)
47 summary(model1)
48
49 # a. Residuals vs. Fitted Values
50 par(mfrow=c(2,2)); plot(model1)
51 plot(model1, which=1)
52 # after fitted value of y hat equals ~ 45, variance seems to increase/
       residuals deviate more from 0. variance does not appear to be constant.
53
54 # b. Q-Q plot of studentized residuals
55 plot(model1, which=2)
56 # the normality assumption appears to be met because almost all points are
       very tightly clustered along the QQ line.
57
58 # c. Plot hat values
59 plot(hatvalues(model1), pch=16, cex=1, ylab="Hat values", main="Hat value plot
       ", ylim=c(0,0.4))
60 abline(h=(2*5)/47, lty=2)
61 abline(h=(3*5)/47,lty=2)
62 identify(1:47, hatvalues(model1), row.names(gamble))
63 # using the average hat value as a threshold (thresholds = 2 * hbar and 3 *
       hbar), there are 4 high leverage points in this dataset: observations 31,
       33, 35, and 42.
64 # these four points have are potentially influential i.e. have the potential
       to greatly affect the regression model.
65
66 # d. Outlier test
67 outlierTest(model1)
68 # according to outlier test, observation 24 has an adjusted (Bonferroni) p-
       value ~ 0. therefore, observation 24 is an extreme residual in this model.
69
70 # e. Bubble plot
71 plot(hatvalues(model1), rstudent(model1), ylim=c(-3,7.5), type="n")
72 cook <- sqrt(cooks.distance(model1))
73 points(hatvalues(model1), rstudent(model1), cex=10*cook/max(cook))
74 abline(h=c(-2,0,2), lty=2)
75 abline(v=c(2,3)*5/47, lty=2)
76 identify(hatvalues(model1), rstudent(model1), row.names(gamble))
77 # observations 31, 33, 35, and 42 are influential points because all 4 are
```

```
above at least one threshold for studentized residuals and hat values.
This means all 4 points are both high leverage and are regression outliers
, making them influential points (they greatly affect the regression model
).Observation 24 was also identified because while it did not pass any of
the hat value thresholds, it has an unusally large Cook's distance.
```