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.

```
# load data
gamble <- (data=teengamb)
# run regression on gamble with specified predictors
```

Answer the following questions:

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

```
#check the constant variance, normality assumptions and large leverage
    points
#constant variance by residual vs. fitted & nomality assessment by Q-Q
    plot
plot(model1)
```

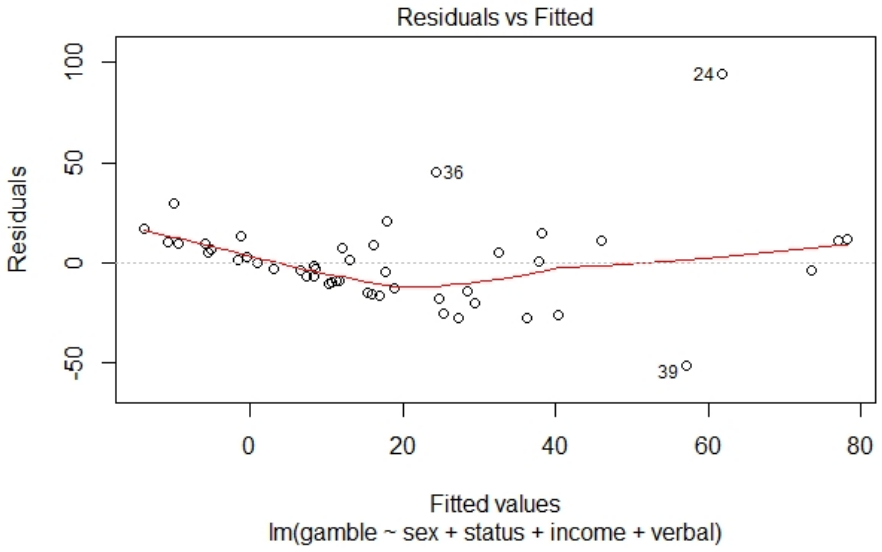

Figure 1: The residual distribution is not very uniform, with variance growing larger for larger fitted values. We therefore cannot safely assume equal variance

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

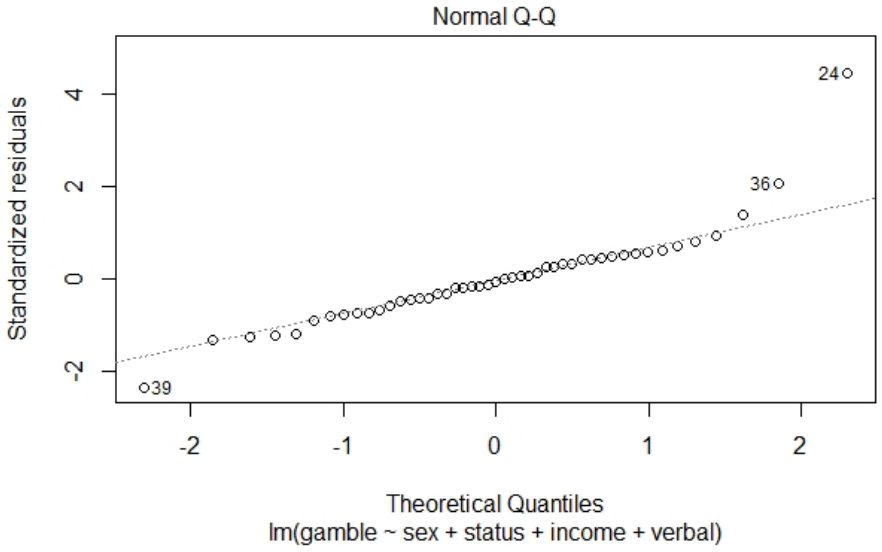

Figure 2: The Q-Q plot for standardized residuals is roughly linear, we can therefore infer that the distributions of all the variables are approximately normal.

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

```
#Plotting hat-values
plot(hatvalues(model1), pch = 16, cex = 1, main = "Hat values plot for
    regression model 'gamble'")
abline(h = 2*4/47, lty = 2)
abline(h= 3*4/47, lty = 2)
identify(1:47, hatvalues(model1), row.names(gamble))
#[1] 31 33 35 42 have high leverage using the thresholds 2(k + 1)/n and 3(
    k + 1)/n
```

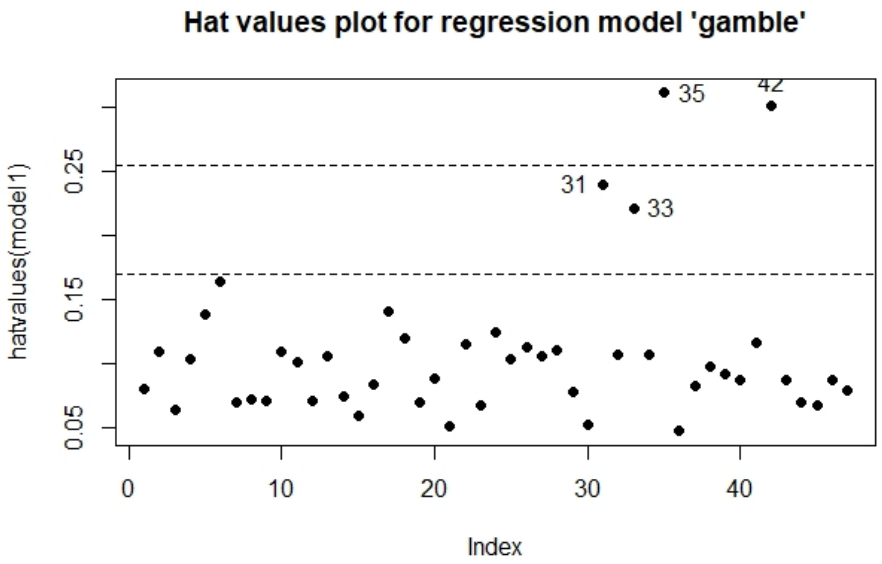

Figure 3: We identified four obervations (row number labeled) as large leverage points for they surpass our thresholds

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

```
#Plotting cook's distances for outliers
summary(model1)
plot (cooks.distance(model1), pch = 16, cex = 1)
abline(h=4/(47-4-1), lty = 2) #showing any cook's distance > 4/(n-k-1)
identify(1:47, cooks.distance(model1), row.names(gamble)) #finding which
    observations are the outliers
#Outlier test
outlierTest(model1, row.names(model1))
```

Output:

```
No Studentized residuals with Bonferroni p <
Largest |rstudent|:
rstudent unadjusted p-value Bonferroni p
24 6.016116         4.1041e-07   1.9289e-05
```

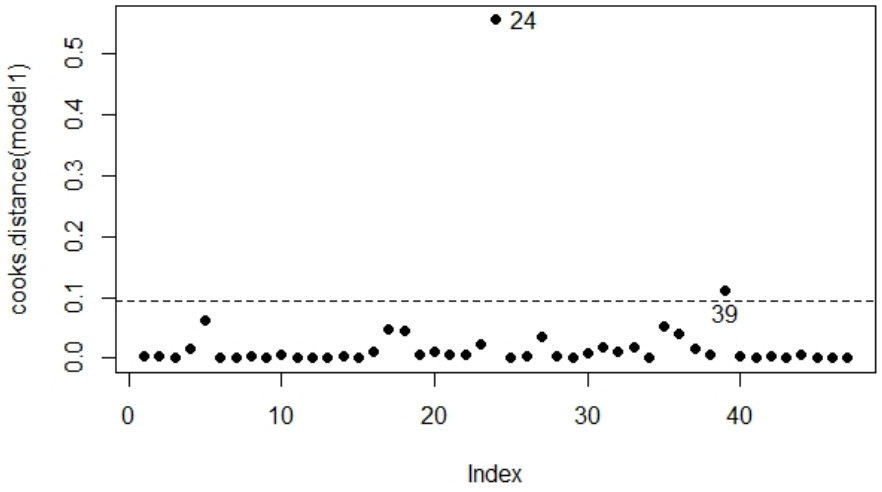

Figure 4: Plot of Cook's distance for regression model "gamble", This suggest that the 24th observation in the dataset has the largest studentized residual and is a significant outlier ($P < 0.001$)

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

```r
#Creating a bubble plot
plot(hatvalues(model1), rstudent(model1), type = "n", main = "Individual
    influences in regression model 'gamble'",
      xlab = "h values", ylab = "studentized residuals")
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)*4/47, lty = 2)
identify(hatvalues(model1), rstudent(model1), row.names (gamble))
```

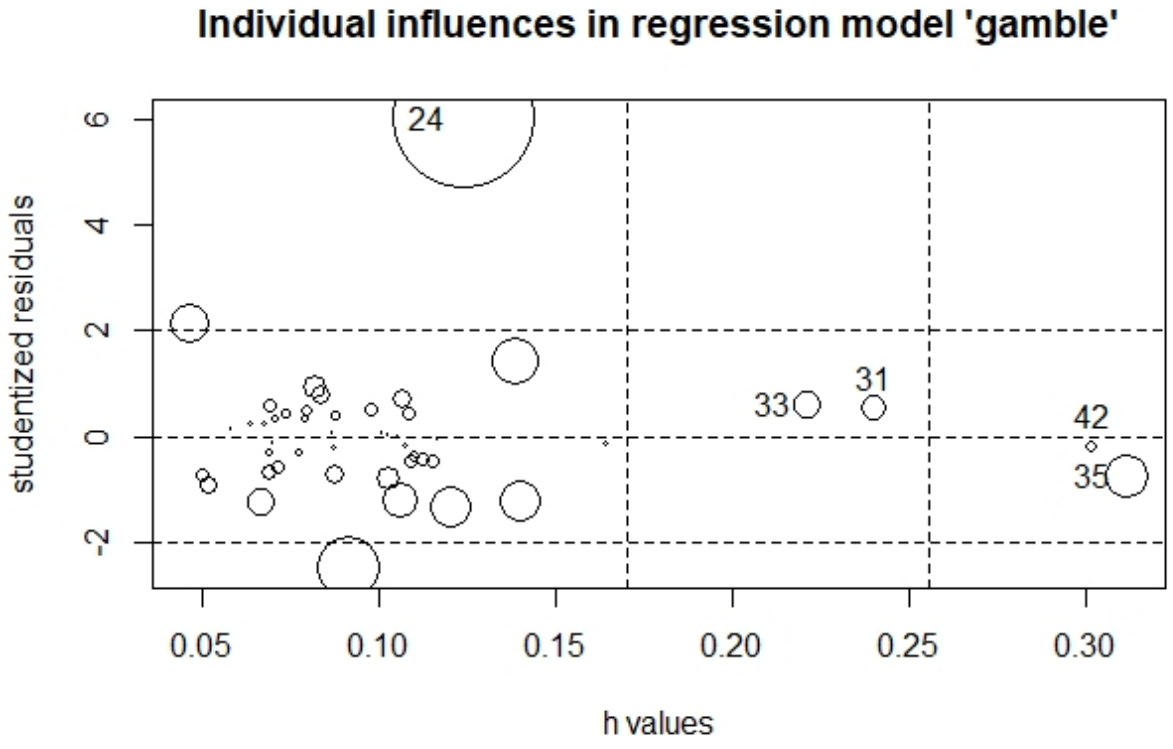

Figure 5: Bubble plot of influence measures for regression model "gamble". We can see the four large leverage points as well as the outlier "24", which has the biggest influence on the model