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.

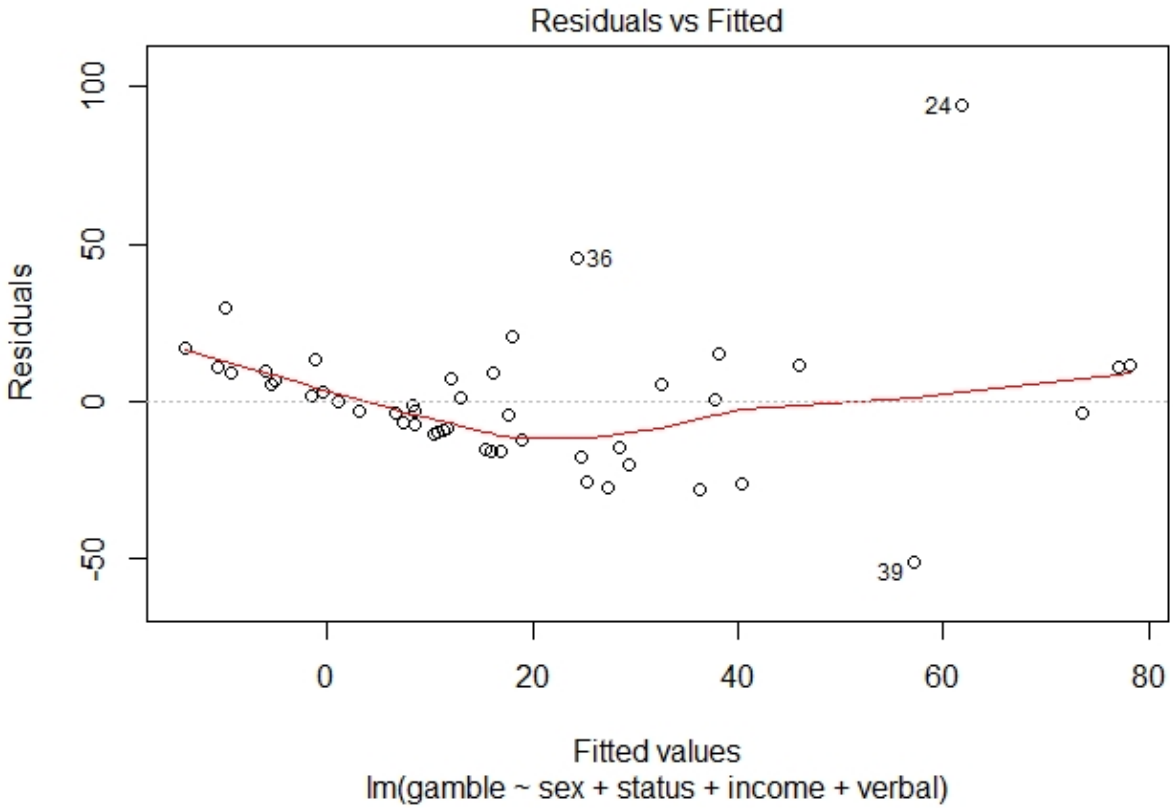
Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

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

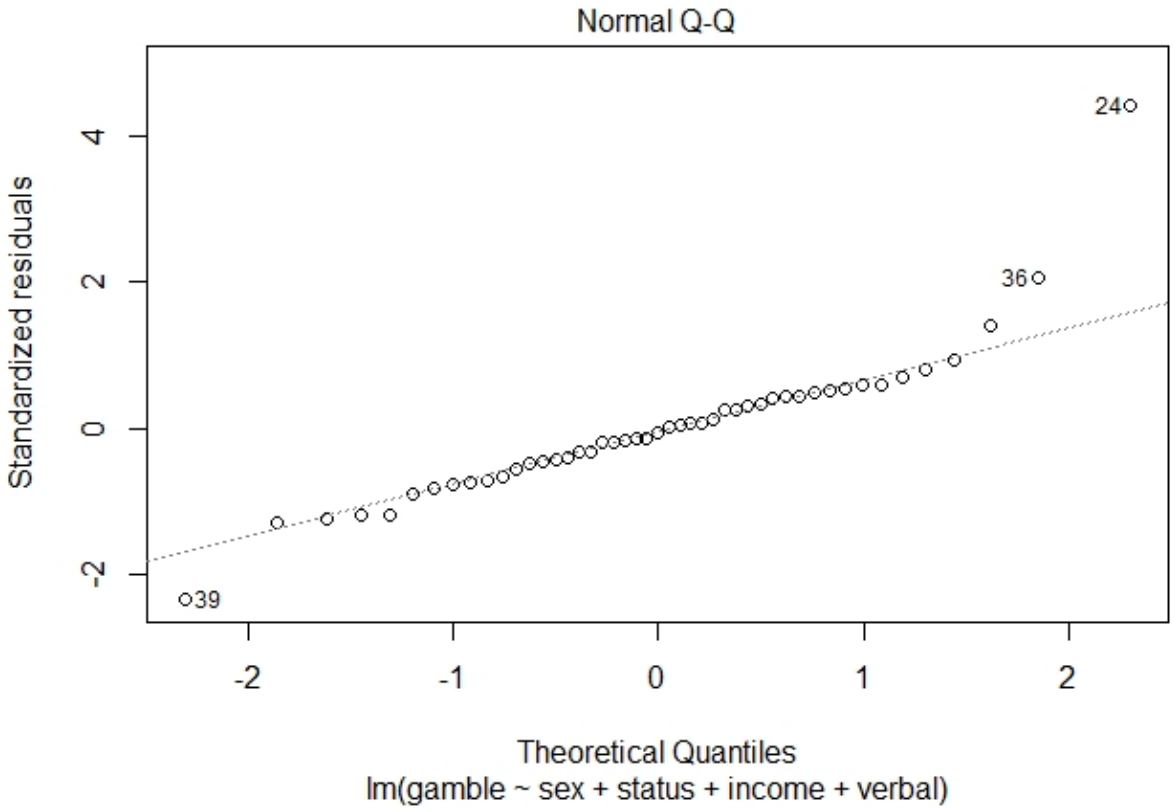
```r
plot(model1) #first plot
```



Overall, the plot looks good and the constant variance assumption seems to be satisfied. There is a larger amount of variance in the middle of the plot around 30 on the x-axis compared to the edges of the plot but it does not seem drastic.

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

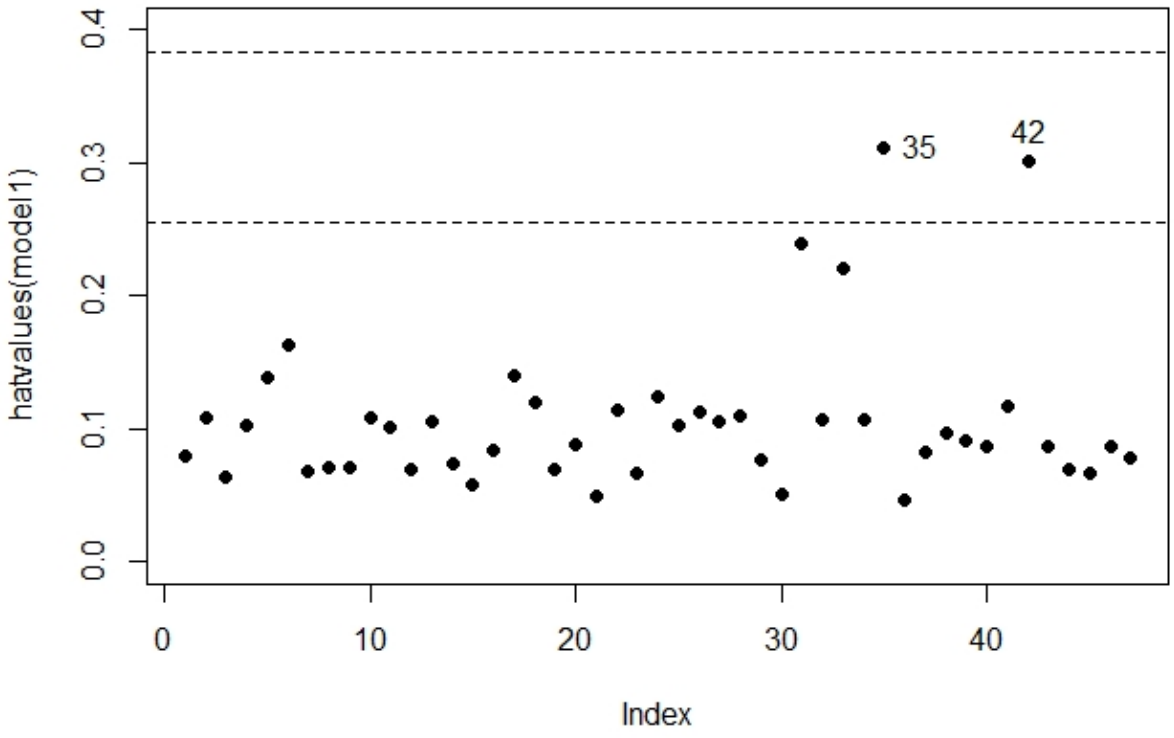
```r
plot(model1) #second plot
```



The normality assumption appears to be satisfied because the points fall along a straight line on average with the exception of 24, 36, and 39.

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

```
plot(hatvalues(model1), pch=16, cex=1, ylim=c(0,0.4))
abline(h=2*(5+1)/47, lty=2)
abline(h=3*(5+1)/47, lty=2)
identify(1:47, hatvalues(model1), row.names(gamble))
```



There are two large leverage points: 35 and 42.

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

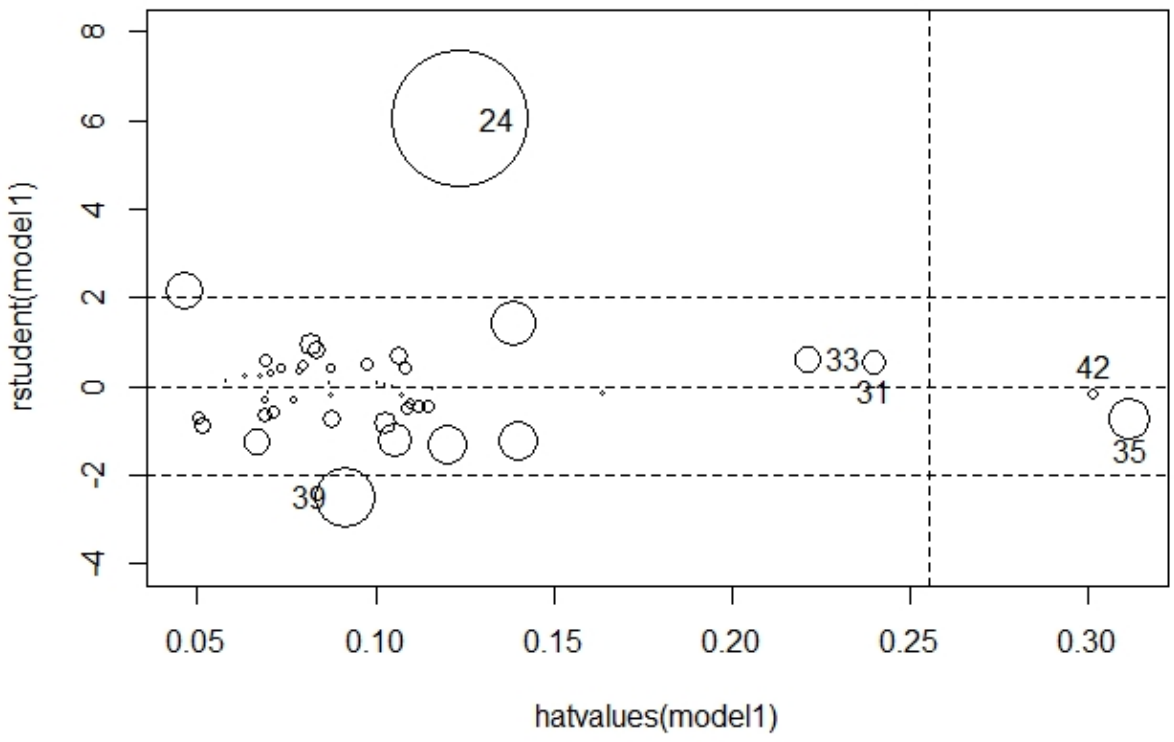
```
library(car)
outlierTest(model1, row.names(gamble))
```

The adjusted p-value for the largest error hat*i is larger than 0.05, we conclude that this

model does not have any extreme residuals.

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

```
plot(hatvalues(model1), rstudent(model1), type="n", ylim=c(-4,8))
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)*6/47, lty=2)
identify(hatvalues(model1), rstudent(model1), row.names(gamble))
influence.measures(model1)
```



The influential points are 24, 31, 33, 35, 39, and 42.