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.

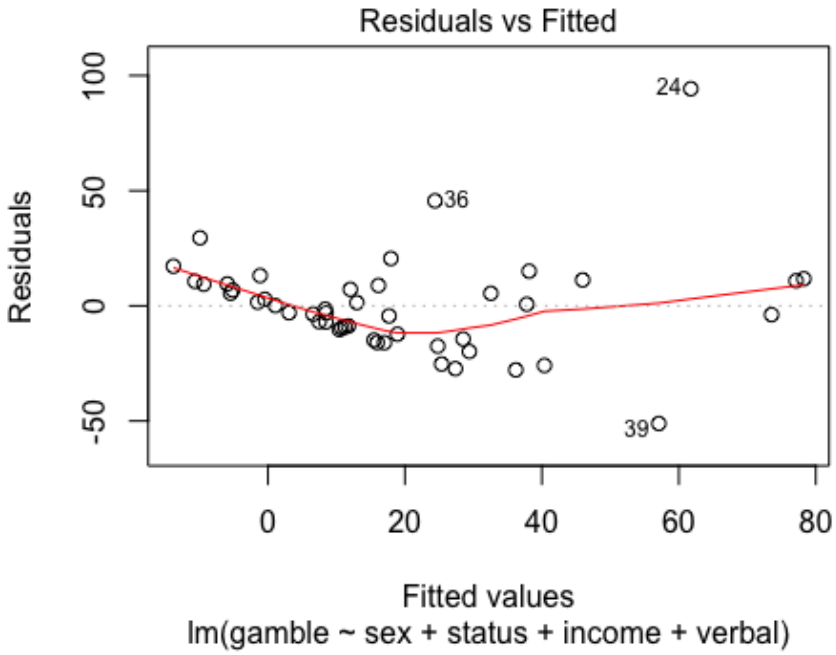
Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

```
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.

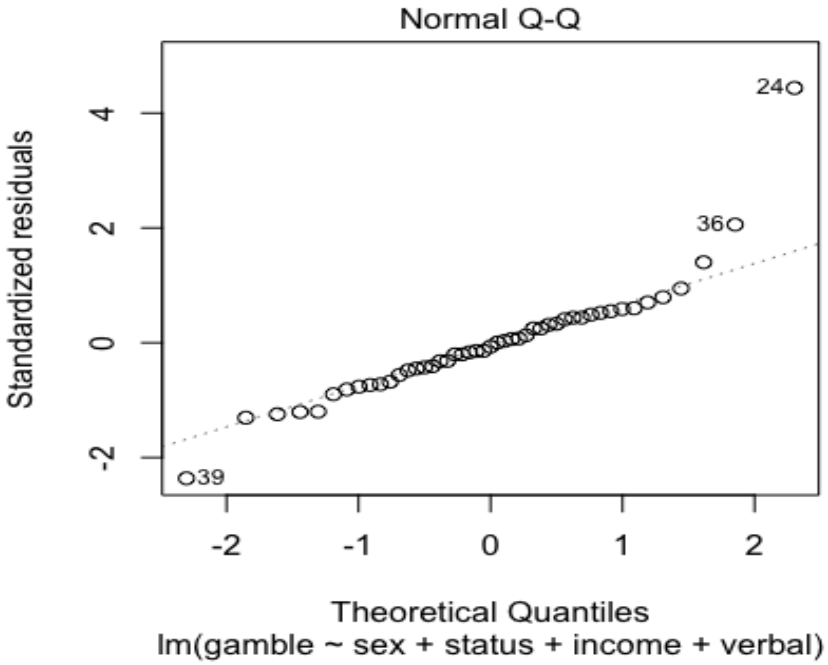
```
# Plot the resitual vs the fitted values
plot(model1, which=1)
```




```
# The values of y conditioned on each value of x do not appear to have the
    same standard deviation at each x value. There is much greater
    variance around 60. As a result, the constant variance assumption
    appears to be violated.
```

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

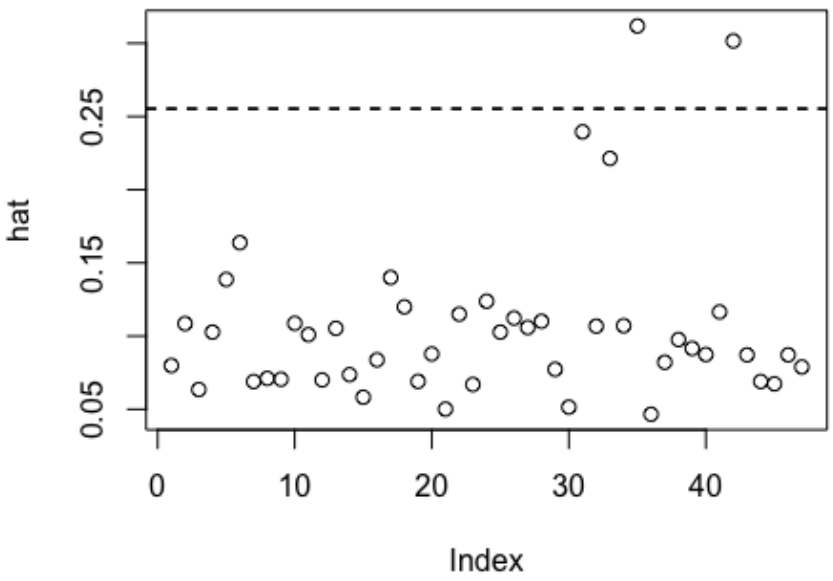
```
# Use a qqplot to graph the distribution of the studentized residuals
    conditional on the fitted values of the regression model1
plot(model1, which=2)
```




```
#Since the values do not all follow the same linear trend, there is
    evidence that the values of the studentized residuals conditioned on
    the fitted values do not all follow a normal distribution. Specifically
    , values 24, 36, and 39 do not appear to follow a normal distribution
```

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

```
# Calculate the hat values
hat <- lm.influence(model1)$hat
# Plot the hat values
plot(hat)
# Use the sum function to determine the number of predictors in "hat."
    This is the k value
sum(hat)
# The k value is 5. There are 47 observations. Use 2(5+1)/47 and 3(5+1)/47
    as a guide for high leverage
abline(h=2*(5+1)/47, lty=2)
abline(h=2*(5+1)/47, lty=2)
```



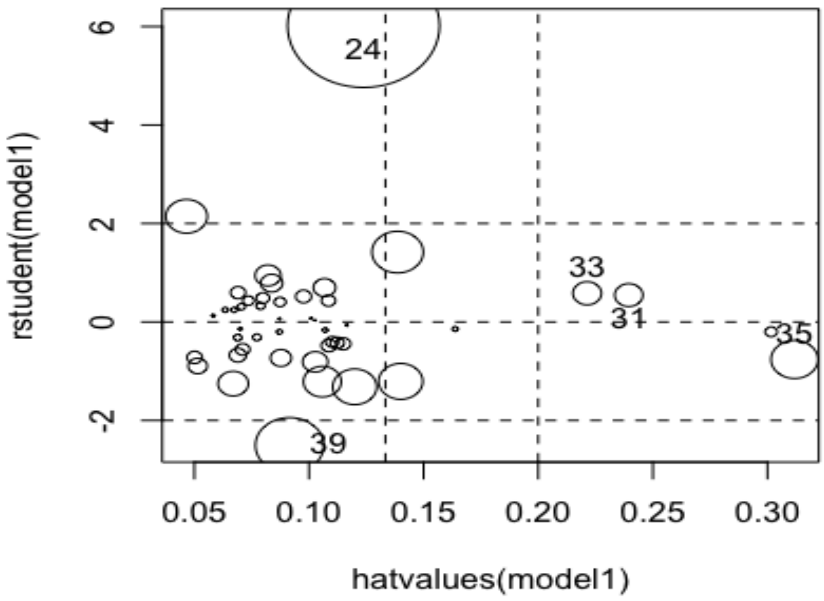

```
# Two points have high leverage using the threshold 2(5+1)/47 and 3(5+1)/
    47. They have potential to greatly influence the fitted model.
```

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

```
outlierTest(model1)
rstudent unadjusted p-value Bonferroni p
24 6.016116          4.1041e-07   1.9289e-05
# The p-value is smaller than the significance level of 0.05, so reject
    the null hypothesis that there are no outliers.
```

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

```
plot(hatvalues(model1),rstudent(model1), 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)*3/45, lty=2)
identify(hatvalues(model1),rstudent(model1), row.names(gamble))
```




```
# Point 24 has the largest influence on the dataset because of its large
    Cook's distance and standardized residuals.
```