QTM 200: Applied Regression Analysis- Problem Set 5

Farris Sabir

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)
```

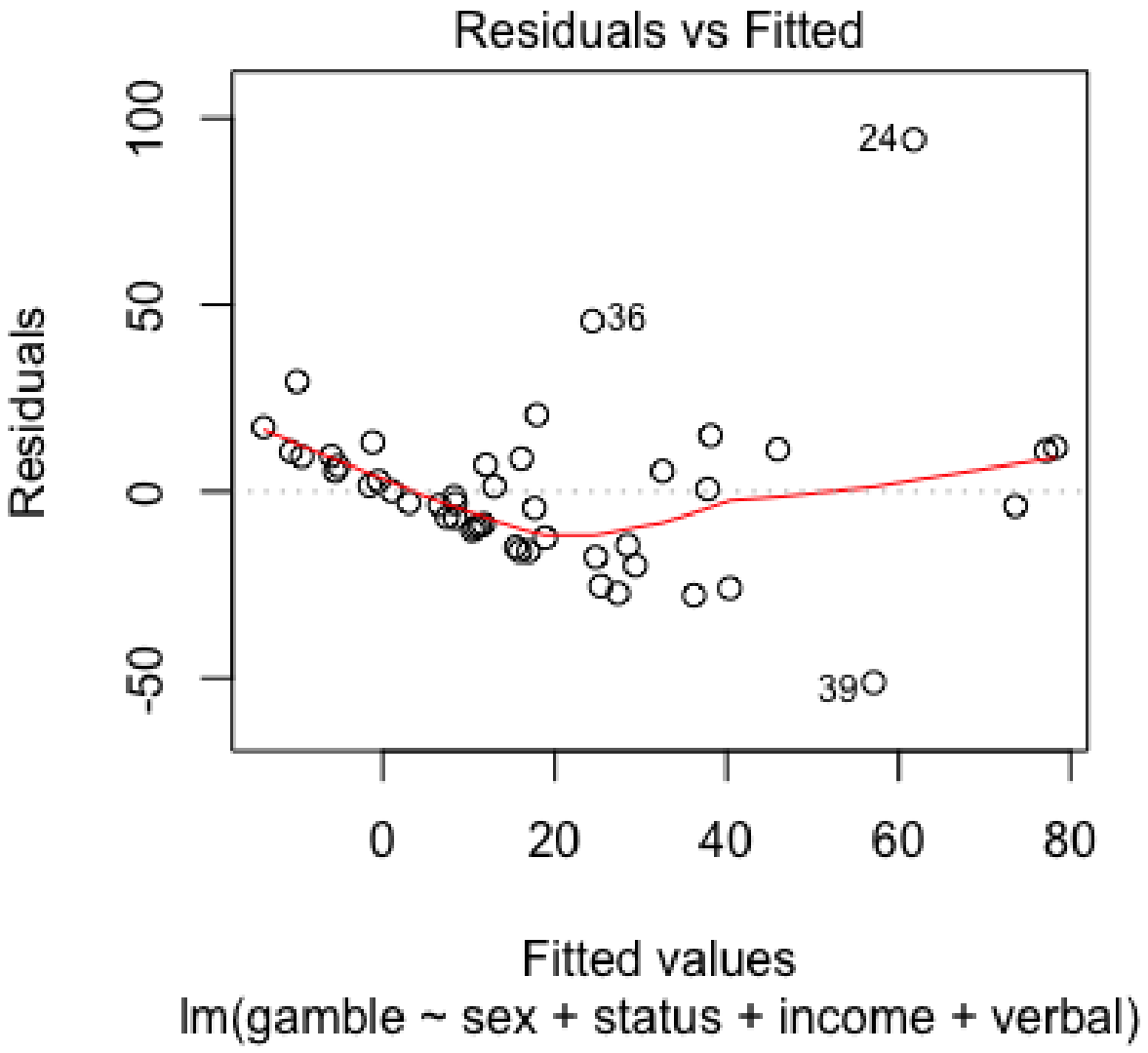
Answer the following questions:

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

```
plot(model1, which=1)
```

According to Figure 1, the constant variance assumption might be in question because the line of the in the graph is not completely straight, only slightly so. This may indicate variance may vary across different fitted values.

Figure 1: Residuals vs. Fitted

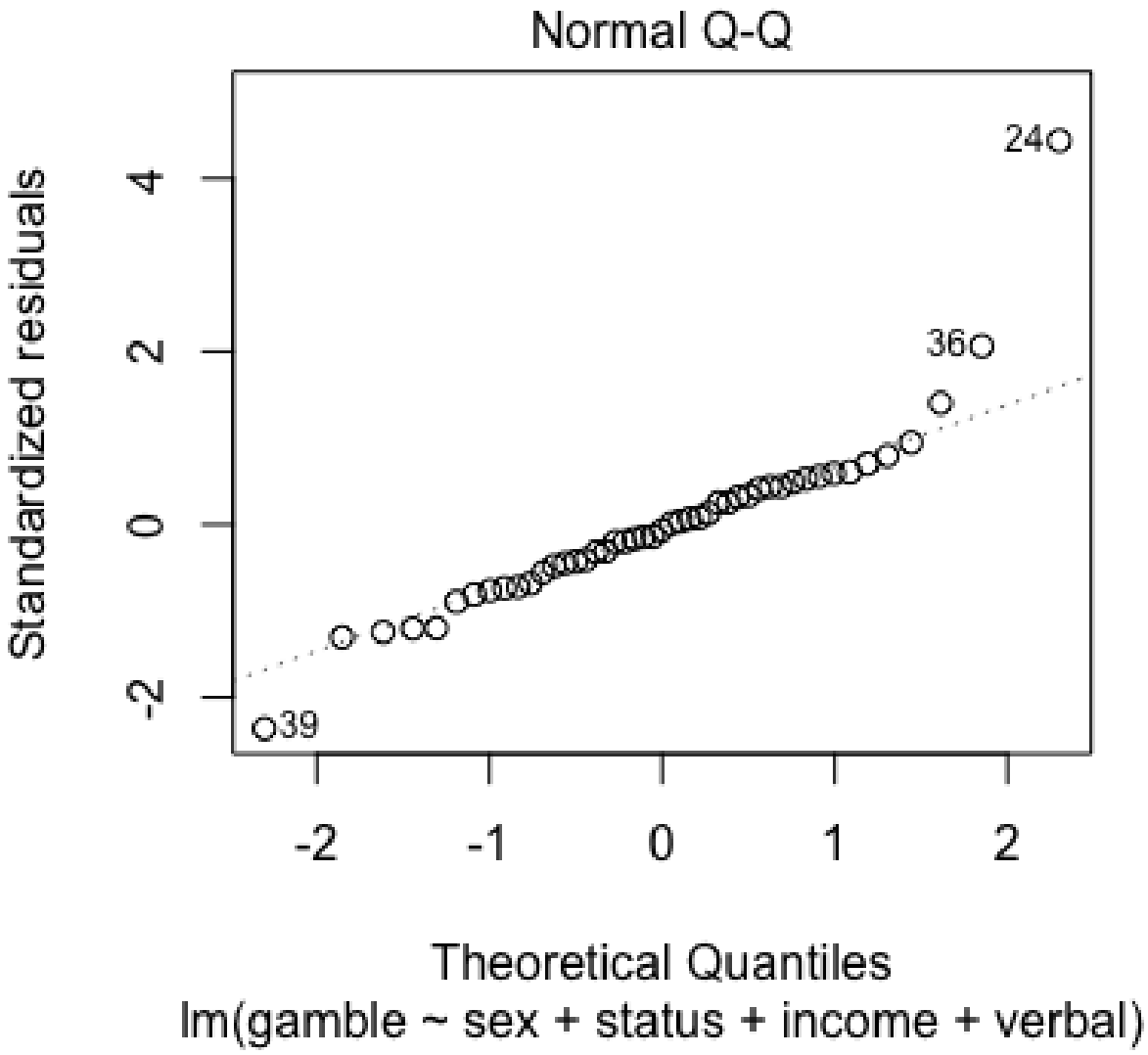


(b) Check the normality assumption with a Q-Q plot of the studentized residuals.

```
plot(model1, which=2)
```

This Q-Q plot in Figure 2 supports the normality assumption because most values fall on the line. Values falling on the line indicate that variation in those x-values are normally distributed. More error occurs at the end, but that is expected, especially since there are less data for these end values.

Figure 2: Normal Q-Q



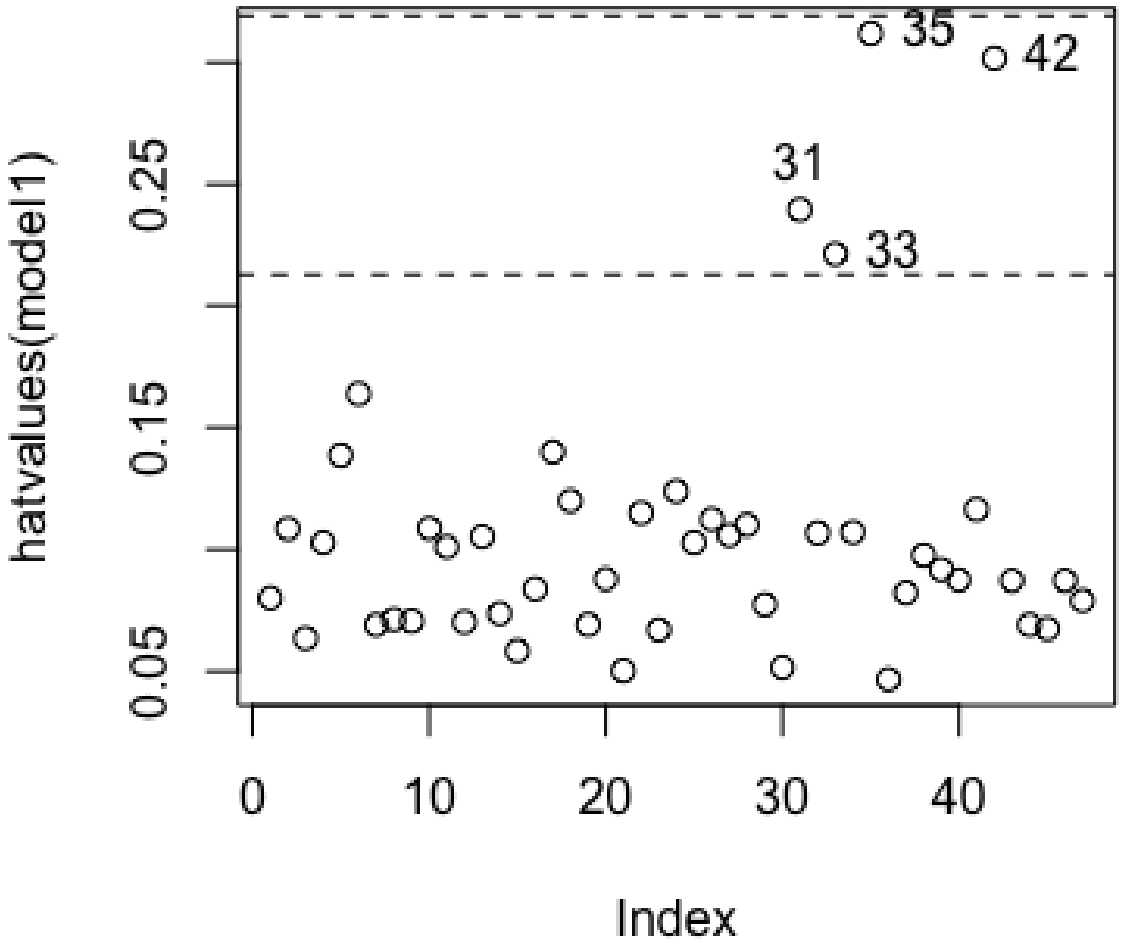
(c) Check for large leverage points by plotting the h values.

```
plot(hatvalues(model1))
abline(h=2*5/47, lty=2)
abline(h=3*5/47, lty=2)
identify(1:47, hatvalues(model1), row.names(gamble))
```

There are four points with high leverage using threshold of $\frac{2(k+1)}{n}$, but not threshold of

$\frac{3(k+1)}{n}$ as shown in the Figure 3.

Figure 3: Hat Values Plot



These points have the potential to greatly influence the fitted model:

```
gamble[c(31, 33, 35, 42), ]
```

```
sex status income verbal gamble
31   0     18   12.0      2   88.0
```

```
33   0     38   15.0      7   90.0
35   0     28    1.5      1   14.1
42   0     61   15.0      9   69.7
```

(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
```

Because the adjusted p-value for the largest studentized residual is less than 0.05, this 24th observation has an extreme residual or is an outlier.

(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)*5/47, lty=2)
identify(hatvalues(model1) , rstudent(model1), row.names (gamble))
```

Figure 4 helps track which points possibly and actually are influential.

The same four points from part (c) have high leverage under threshold of $\frac{2(k+1)}{n}$, but not threshold of $\frac{3(k+1)}{n}$.

Two points have some of the largest regression residuals, the 24th and 39th observation.

```
   sex status income verbal gamble
24   0     27     10      4    156
39   0     51     10      6      6
```

The 24th observation likely has the largest influence on the model. It has a large regression residual, as verified by part d, and large Cook's distance.

Figure 4: Bubble Plot