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.

```r
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.

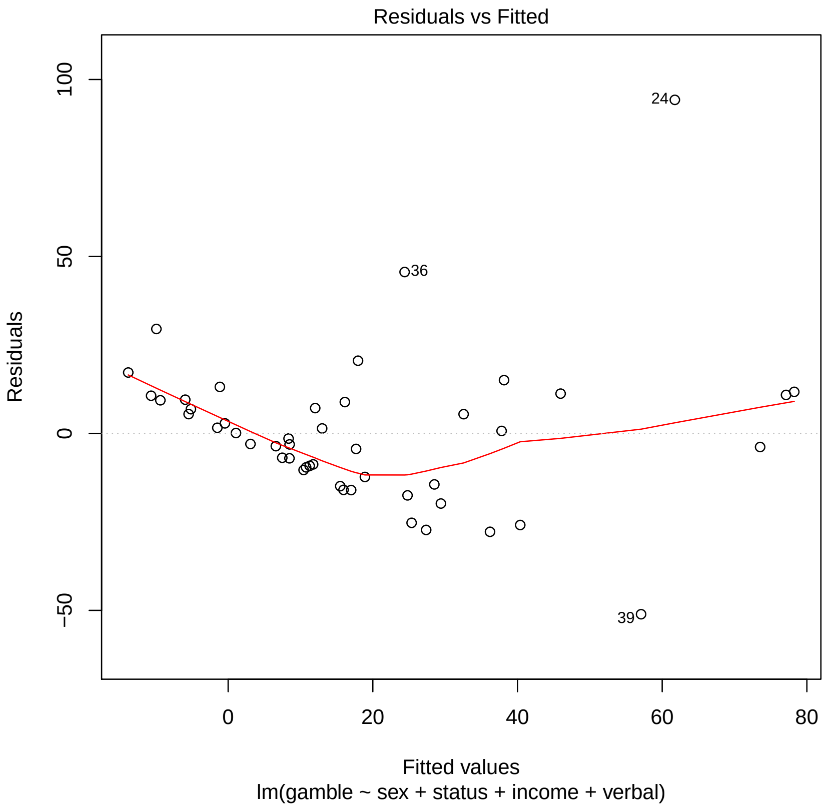

Figure 1: Residuals vs. Fitted Values

```
pdf(”plot1.pdf”)
plot(model1)
abline(h=0,lty=2)
dev.off()
```

Based on the plot of residuals against fitted values, the constant variance assumption doesn’t seem to be met. The residuals fan out as the fitted values increase (especially the outliers after fitted value of 50), and there is no random scatter around the x-axis, some dipping below the horizontal line as fitted values increase.

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

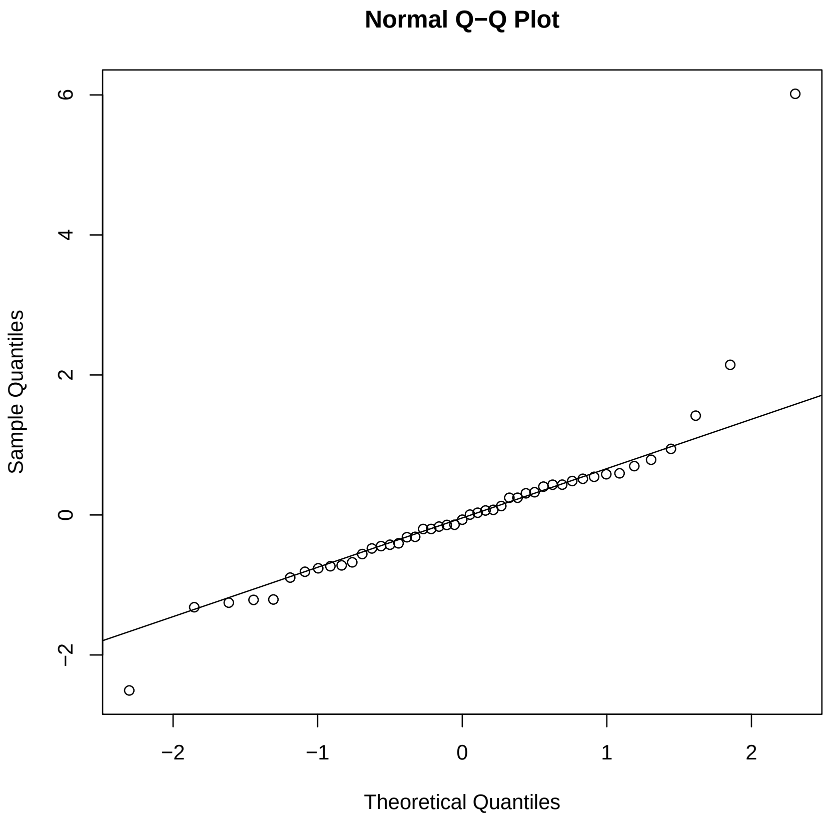

Figure 2: Q-Q Plot of Studentized Residuals

```
pdf("plot2.pdf")
qqnorm(rstudent(model1))
qqline(rstudent(model1))
dev.off()
```

Based on the Q-Q plot of the studentized residuals, the normality assumption is satisfied for the most part except for the outliers that occur as the theoretical quantiles increase.

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

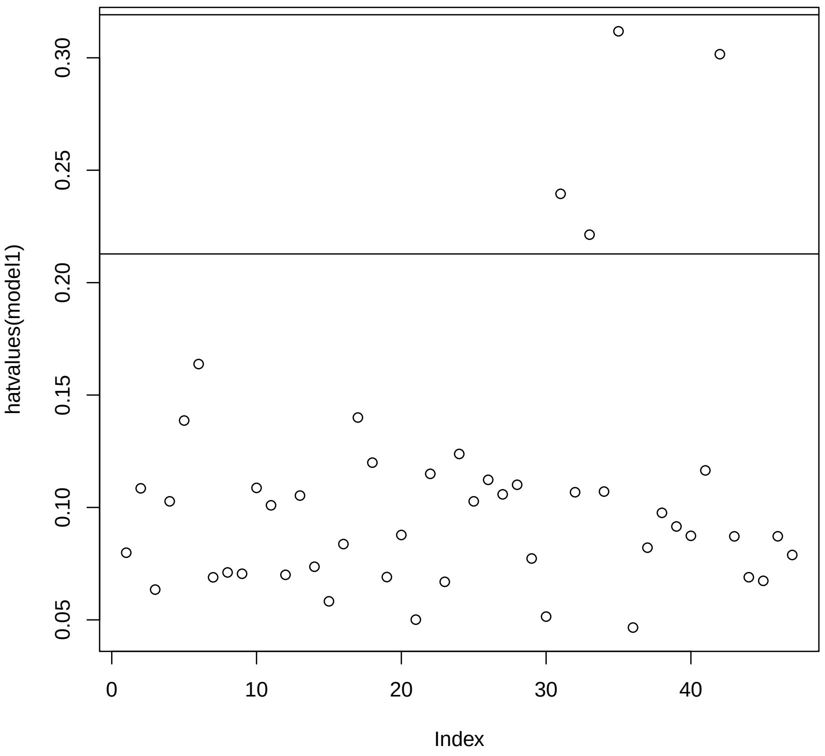

Figure 3: Plot of Hat Values

```r
pdf("plot3.pdf")
plot(hatvalues(model1))
abline(h = 2 * 5 / 47) # Based on the thresholds and k (number of predictors) = 4
abline(h = 3 * 5 / 47)
dev.off()
```

The four points with higher hat values above the drawn line have high leverage, meaning that they have the potential to influence the fitted model. The lines were drawn based on the thresholds $2(k + 1) = n$ and $3(k + 1) = n$, where $k =$ number of predictors and $n =$ sample size.

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

```r
outlierTest(model1, row.names(gamble))
```

Since the adjusted p-value (Bonferroni p) for the largest studentized residual is larger than 0.05 with a value of 0.76464, we would conclude that this model doesn't have any extreme residuals.

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

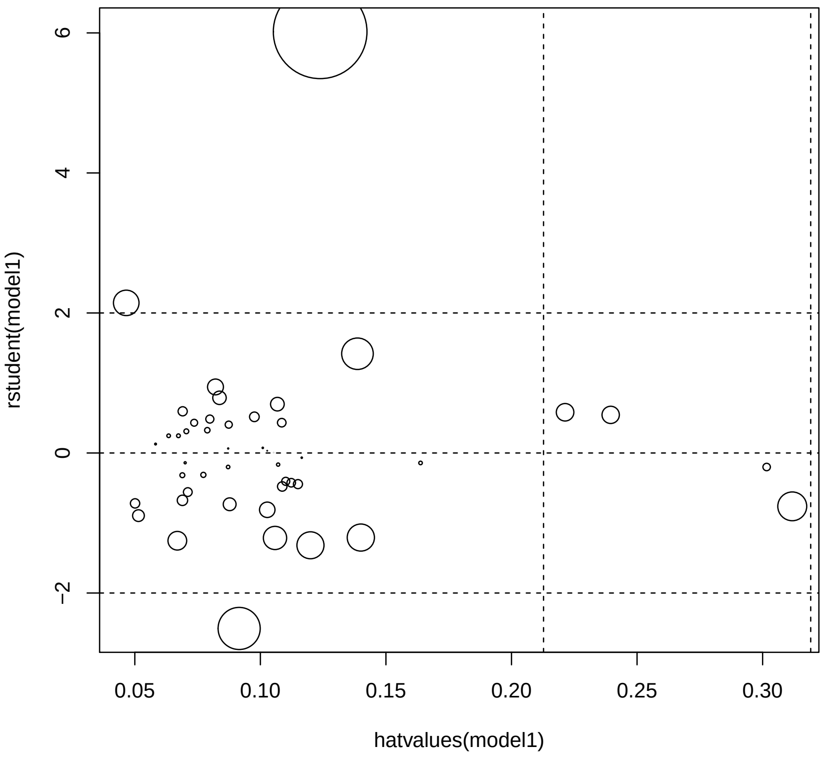

Figure 4: Bubble Plot

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

```r
pdf("plot4.pdf")
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))
dev.off()
```

Based on the Bubble plot, there's an observation with a large studentized residual and a high Cook's distance, but with a low leverage. There are few observations with higher hat values/leverage, but with a small studentized residual/Cook's distance. The bubble plot shows that we do not have extreme outliers, since no observation has both a large studentized residual and high leverage/hat value.