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.

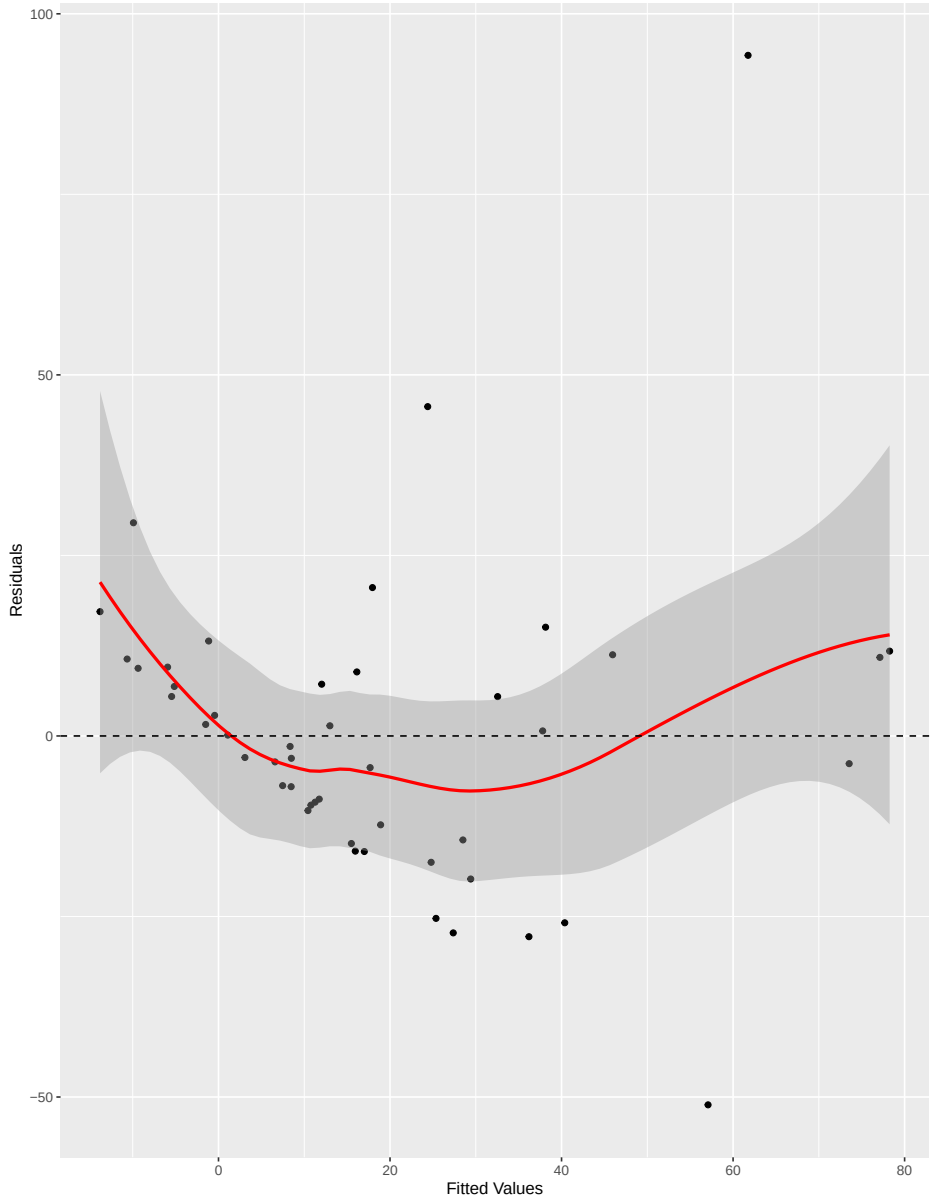
```
library(faraway)
library(tidyverse)

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.

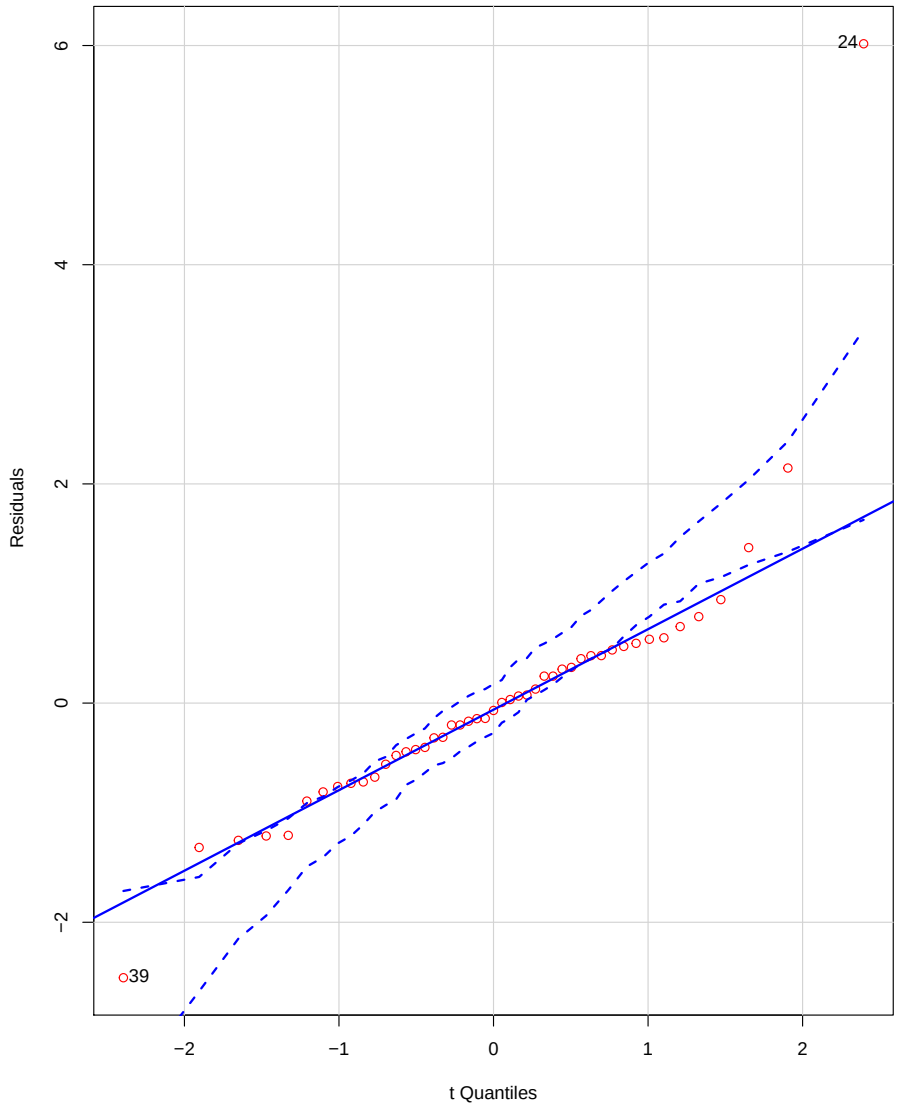
```r
ggplot(model1, aes(.fitted, .resid)) +
  geom_point() +
  stat_smooth(col="red") +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(x="Fitted Values", y="Residuals")
```



The plot has more variance near the center of the distribution of fitted values but their are relatively few outliers compared to the distribution of the data set as a whole.

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

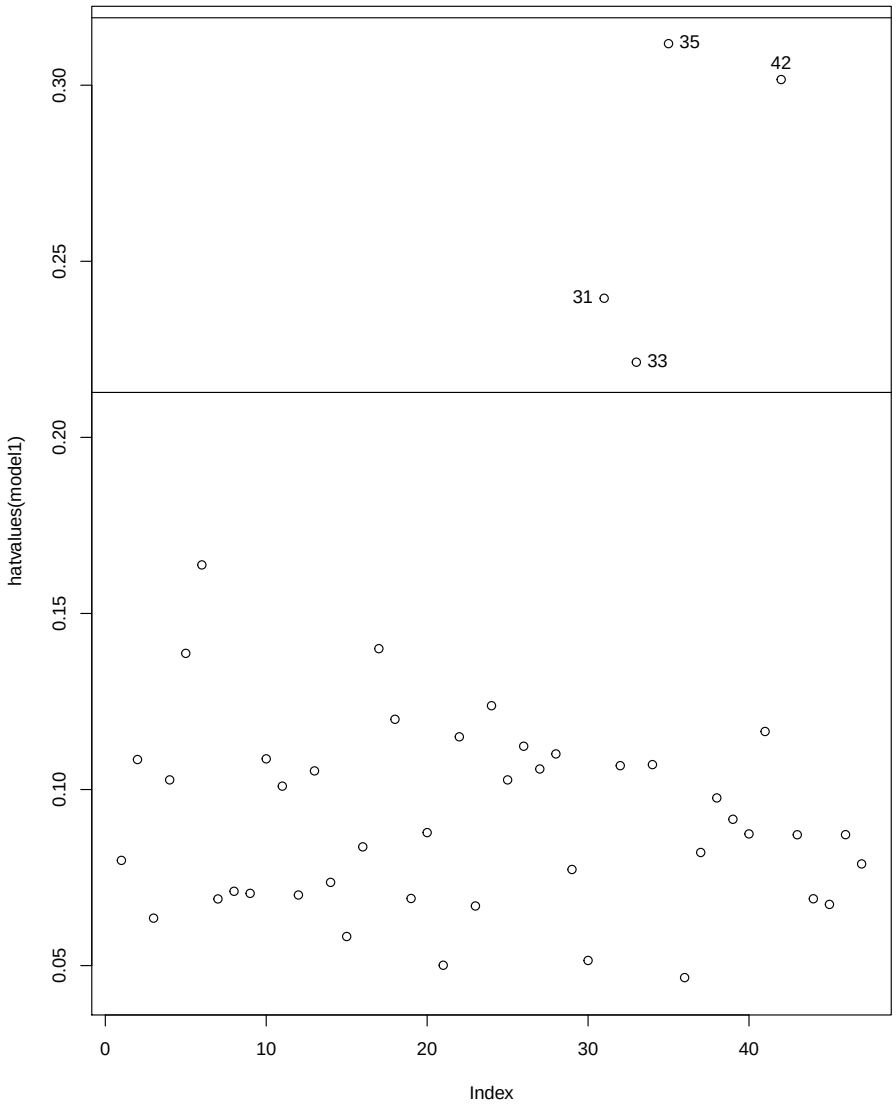
```r
library(car)
qqPlot(model1, ylab=" Residuals", col = "red")
```



Since the ends of the lines diverge from the line of best fit, the data might not be normally distributed due to variation at the extremes.

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

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



The numbered points in the graph have high leverage due to their large hat values.

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

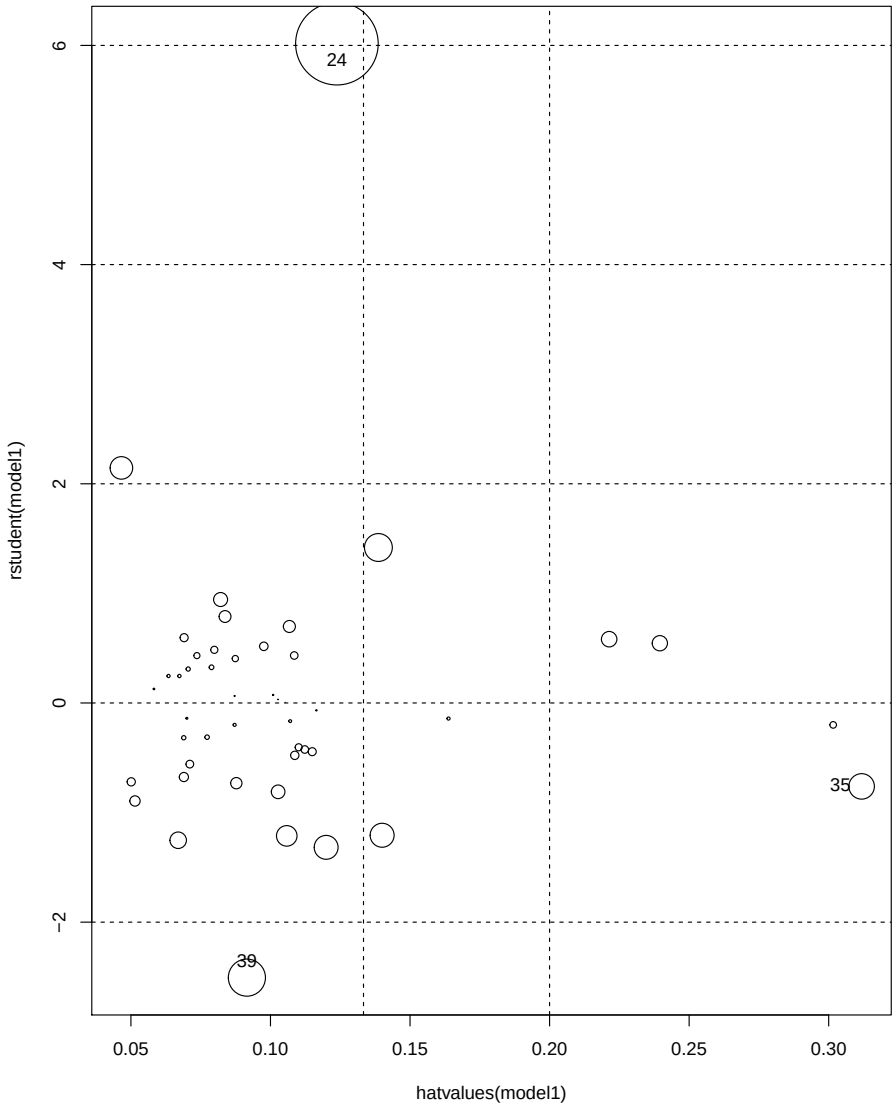
```
outlierTest(model1)
```

```
 rstudent unadjusted p-value Bonferroni p
24 6.016116         4.1041e-07   1.9289e-05
```

The p value from the outlier test is less than 0.05, so I 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, 4, 6), lty = 2)
abline(v = c(2,3)*3/45, lty = 2)
identify(hatvalues(model1), rstudent(model1), row.names(gamble))
```



The bubble labeled 24 has large cook's distance and studentized residual values. This means that this point has a large amount of inlfluence on the model. Point 35, which has a large hat value, but a small studentized residual, shows less influence.