Chloe Sutter: 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.

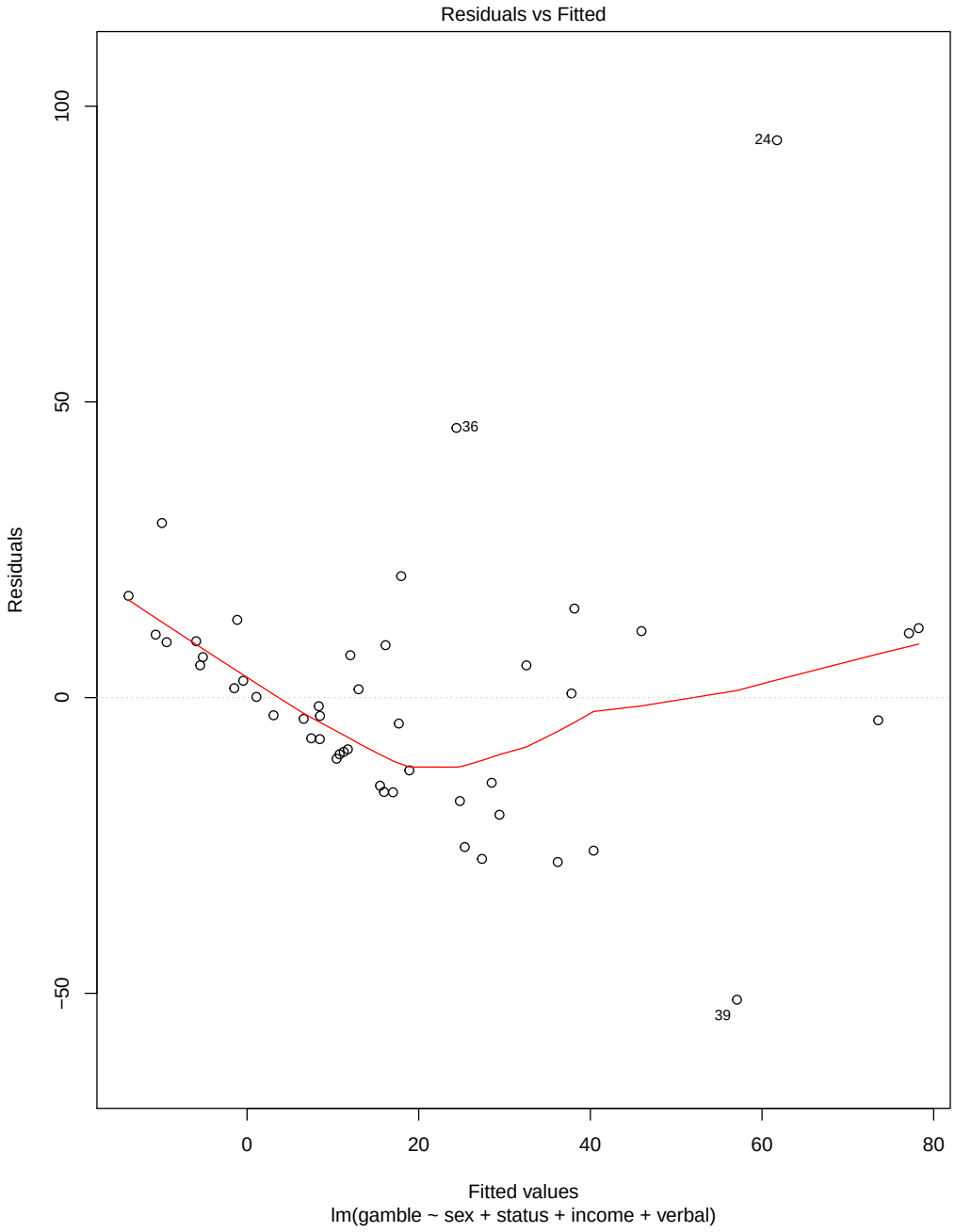
```
# load data
gamble <- (data=teengamb)
# run regression on gamble with specified predictors
model1 <- lm(gamble ~ sex + status + income + verbal, gamble)
model1
# Coefficients:
# (Intercept)          sex       status       income       verbal
    # 22.55565    -22.11833      0.05223      4.96198     -2.95949
```

Answer the following questions:

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

```
plot(model1)
```

Figure 1:

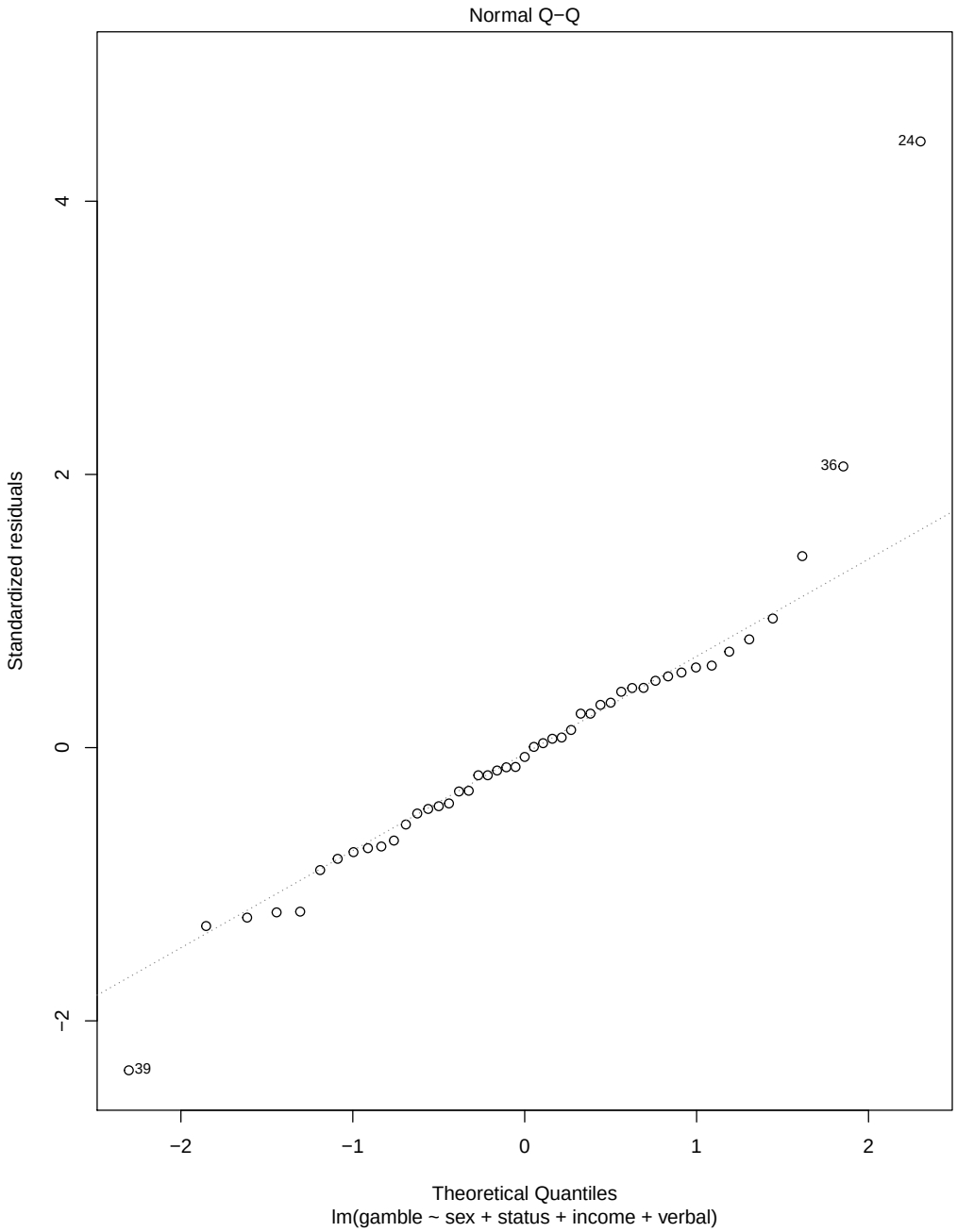


The variance assumption is primarily upheld, as the values of y at each value of x follow (for the most part) a normal distributions. There appear to be outliers at 24, 36, and 39, however.

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

```
plot(model1)
```

Figure 2:

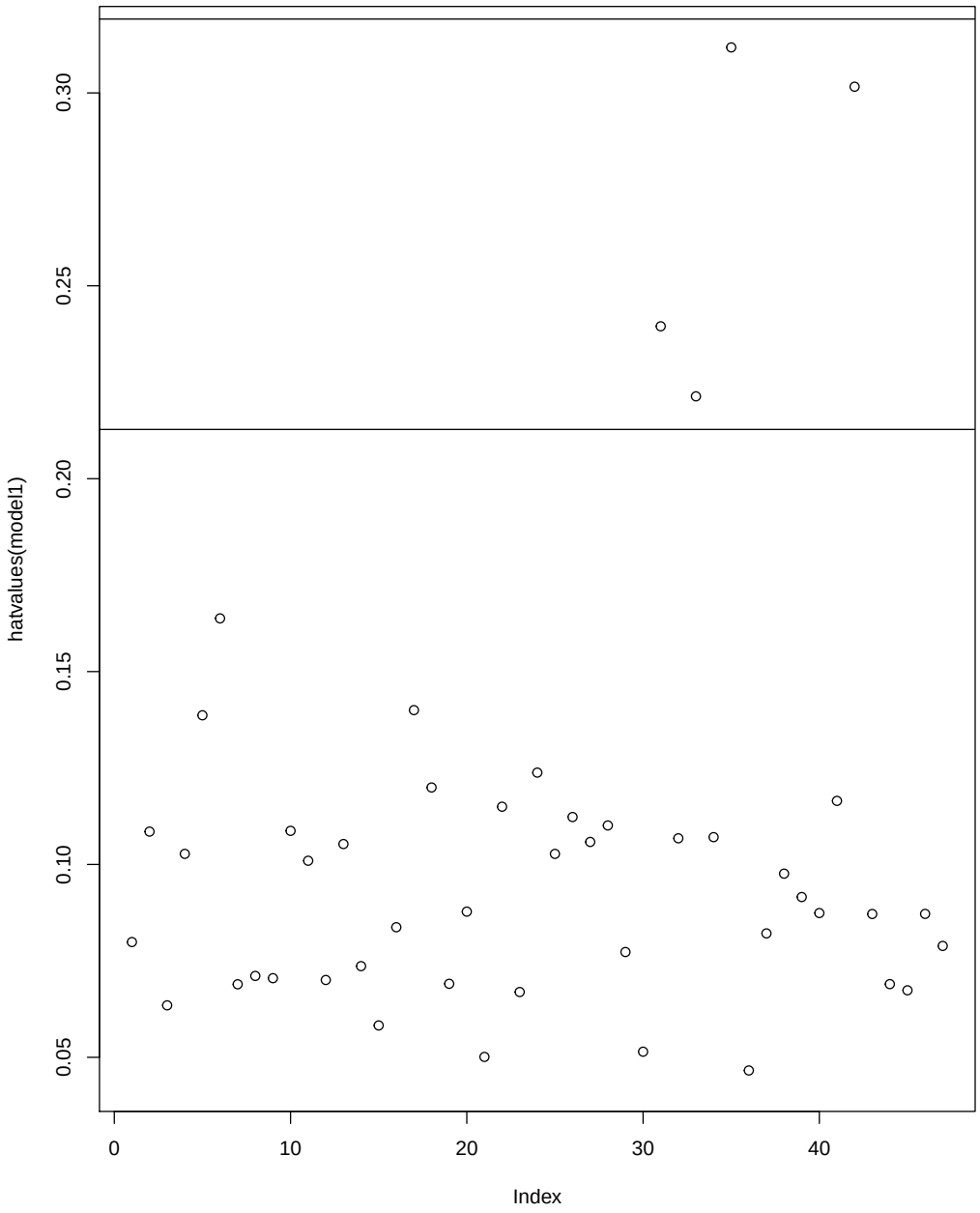


The normality assumption is upheld, as the values of y at each value of x follow a normal distribution. The same outliers still exist, however.

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

```
plot(hatvalues(model1))
abline(h=2*5/47)
abline(h=3*5/47)
```

Figure 3:



Four points demonstrate high leverage (greater than 2*5/47), and thus, they have potential to influence the model.

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

```
# install car package
outlierTest(model1)
#     rstudent unadjusted p-value Bonferroni p
#   24 6.016116         4.1041e-07   1.9289e-05
```

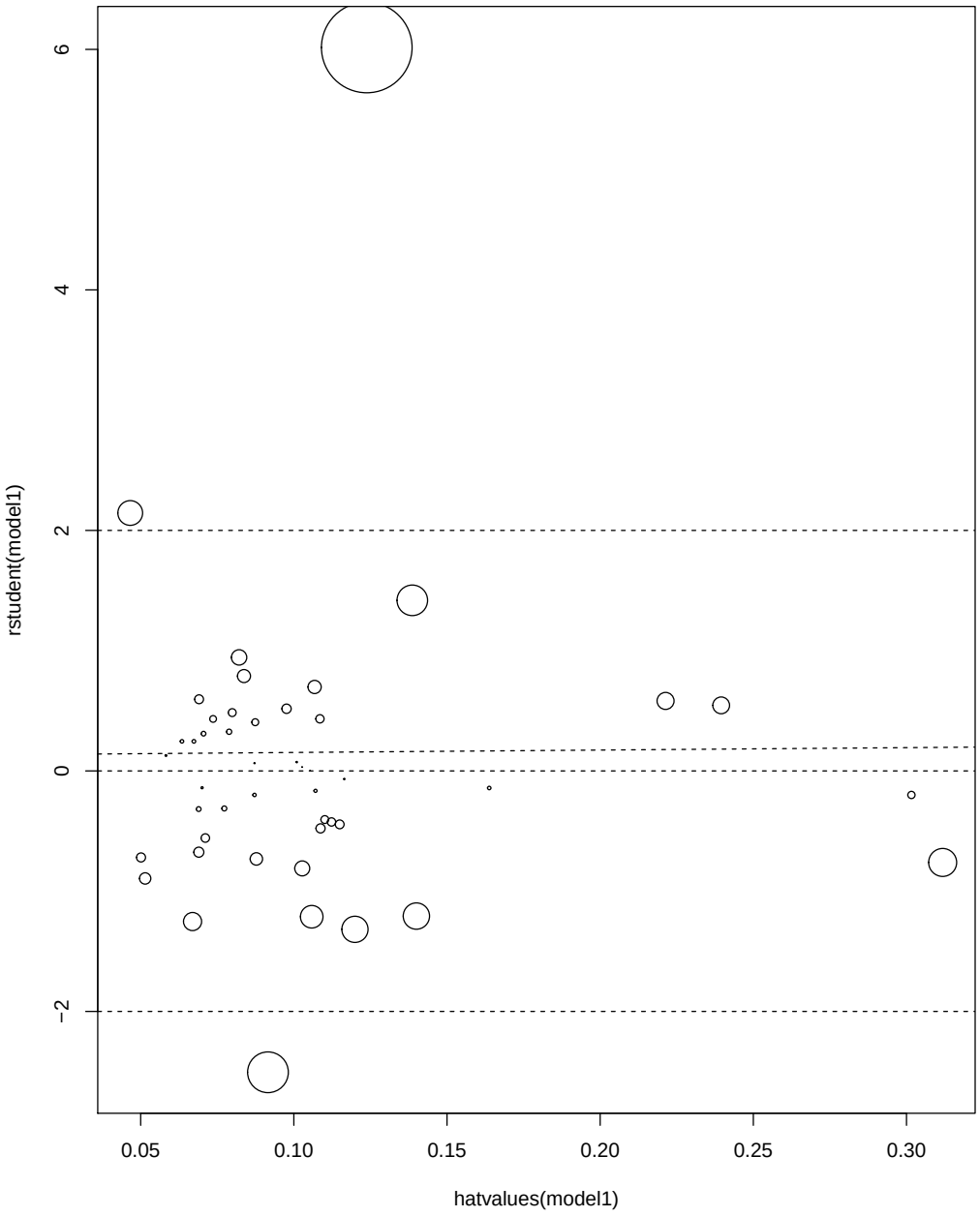
Since the adjusted p value for the model is substantially smaller than 0.05, we would reject the null hypothesis and conclude that the model does display certain extreme residuals.

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

```
plot(hatvalues(model1), rstudent(model1), type = "n")
cook <- sqrt(cooks.distance(model1))
points(hatvalues(model1), rstudent(model1), cex=10*cook/max(cook))
abline(h=c(-2,0,2), lty=2)
abline(c=c(2,3)*3/45, lty=2)
identify(hatvalues(model1), rstudent(model1))
```

One point (24) has both large leverage and large regression residuals, so it has the highest influence on the model.

Figure 4:


rstudent(model1)
6
4
2
0
−2
0.05
0.10
0.15
0.20
0.25
0.30
hatvalues(model1)