Problem Set 5

QTM 200: Applied Regression Analysis

Due: March 4, 2020

Instructions

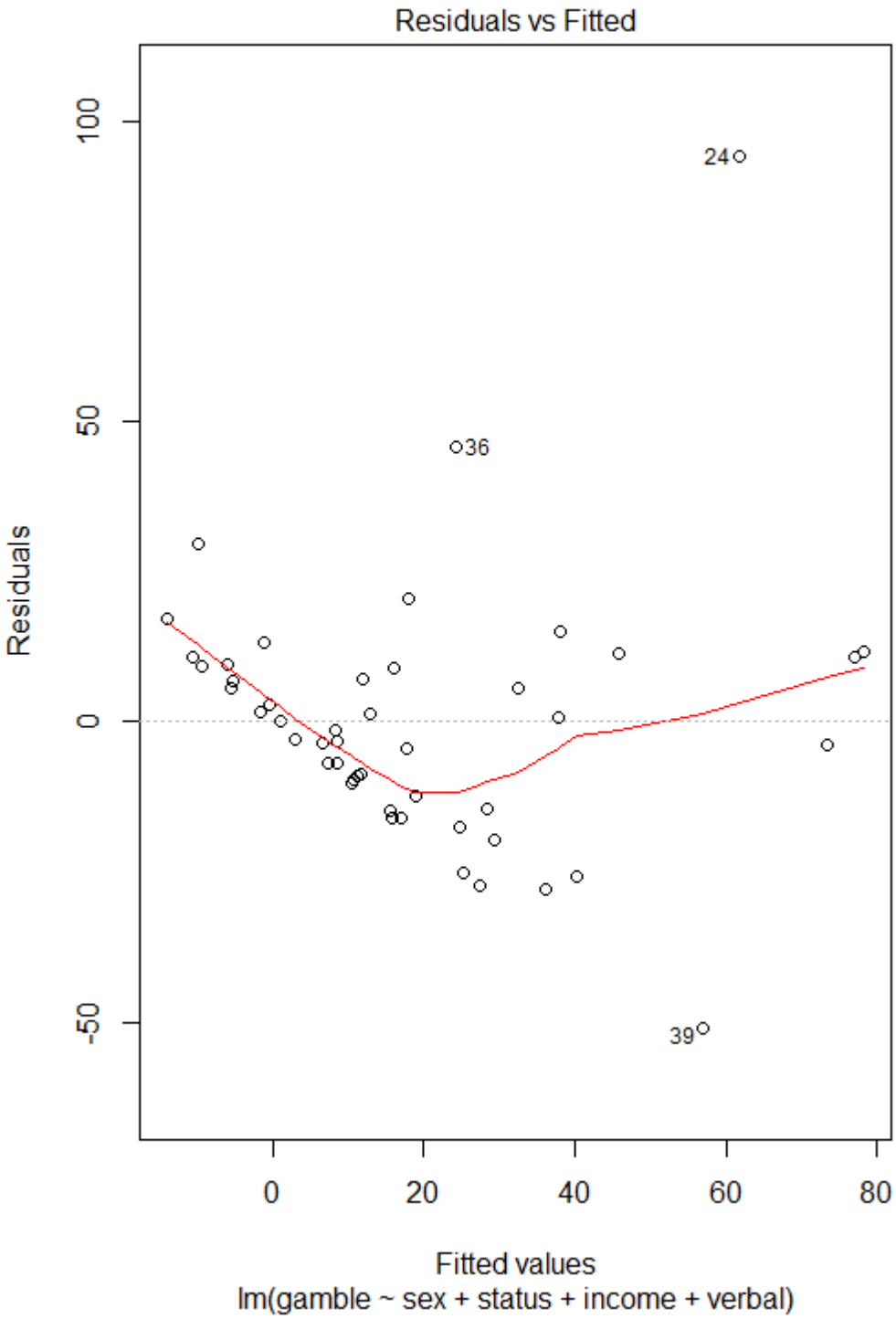
- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular 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)
gamble
View(gamble)
```

Answer the following questions:

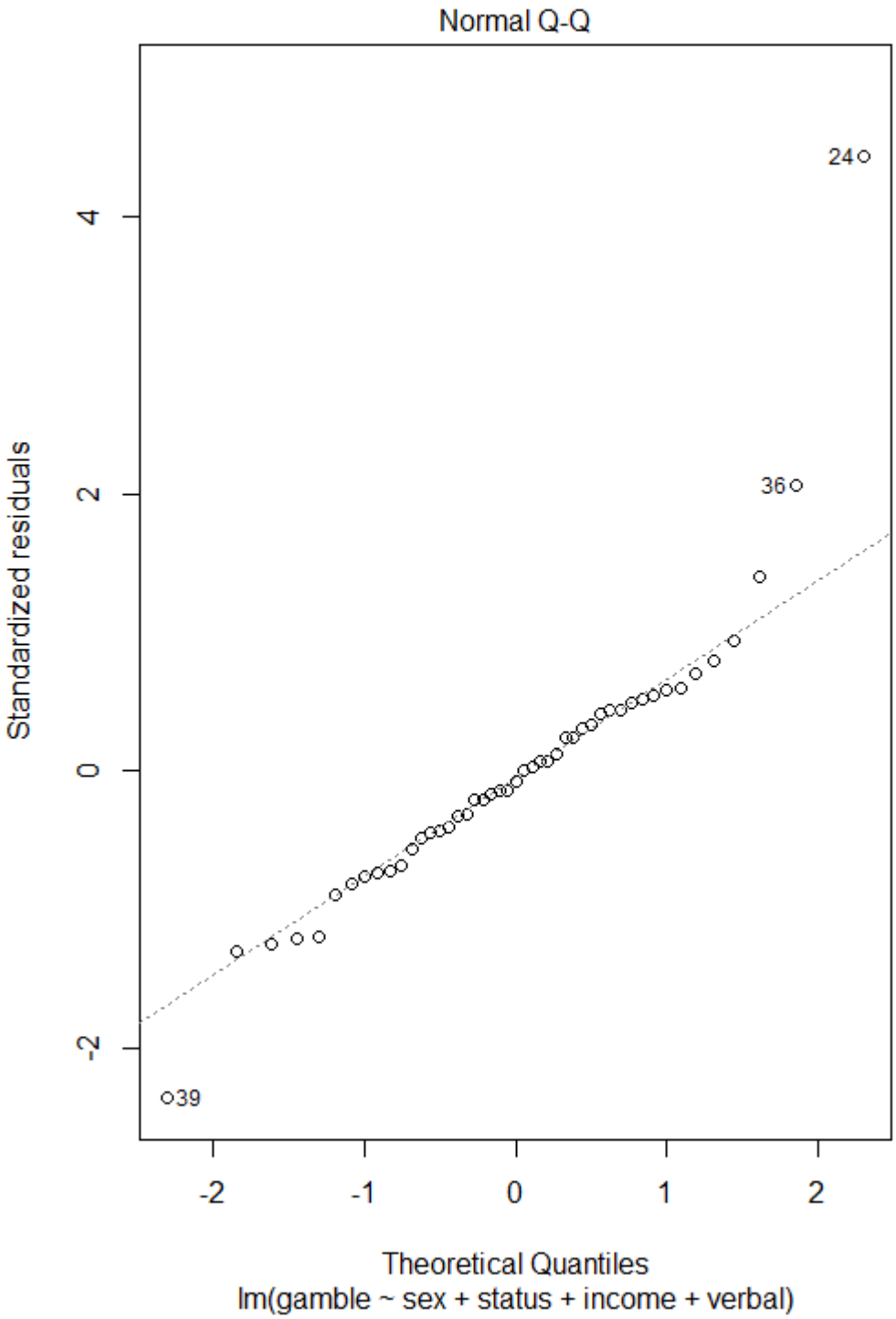
(a) Check the constant variance assumption for the errors by plotting the residuals versus the fitted values.



The constant variance assumption is the individual error against the predicted value, the variance of the error predicted value should be constant. The values of 36 an 24 are shown to not

follow the constant variance assumption while the rest of the residuals do.

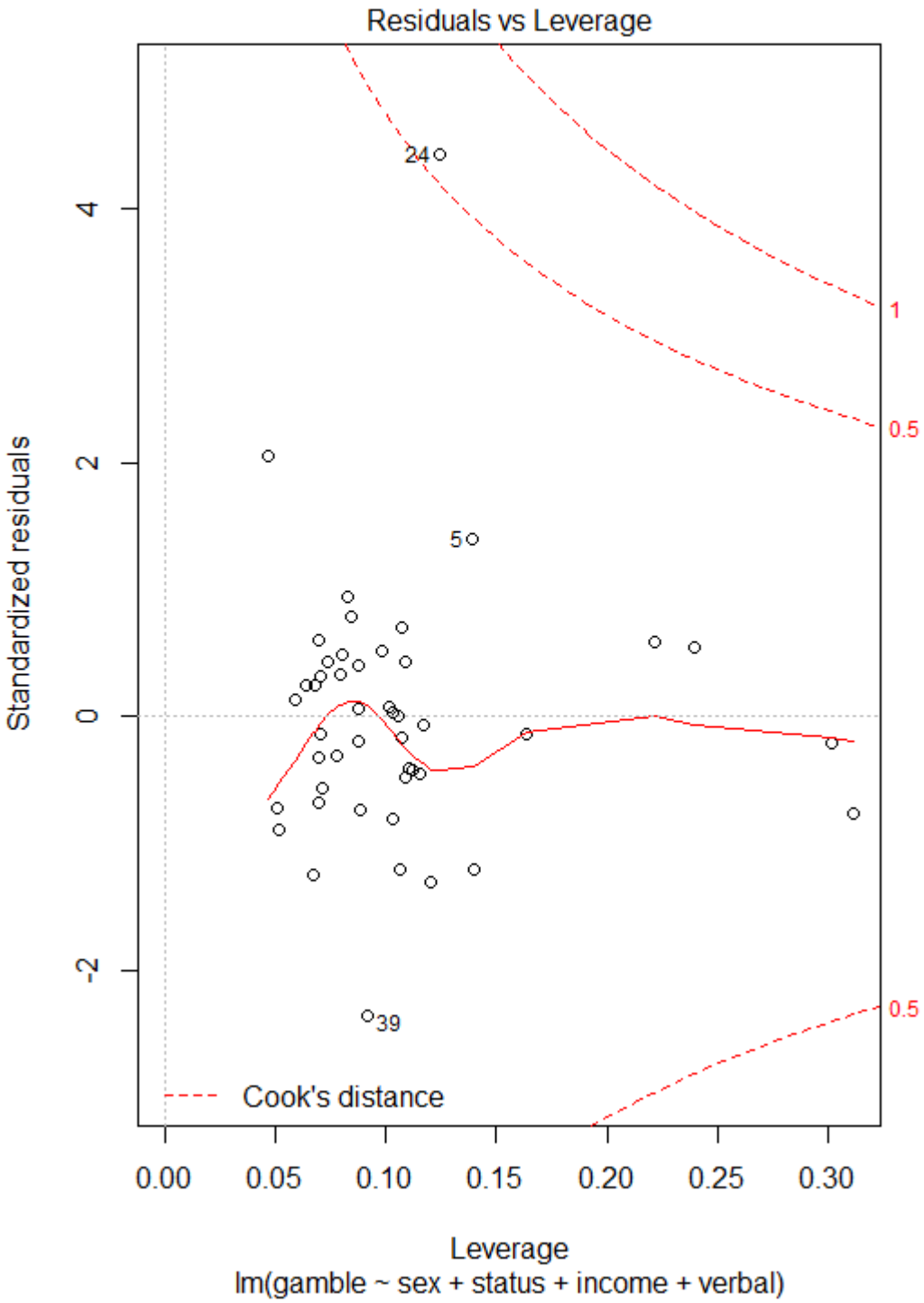
(b) Check the normality assumption with a Q-Q plot of the studentized residuals.



Normality assumption states that the residuals are normally distributed. It can be seen that point 36 and 24 are shown to not follow the normality assumption, while the rest of the resid-

uals do.

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



Cook's distance measures he combination of the observation's leverage and residuals to the regression. The value 24 is shown to be 0.5 cook's distance away, illustrating a high

leverage and high residuals compared to other observations.

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

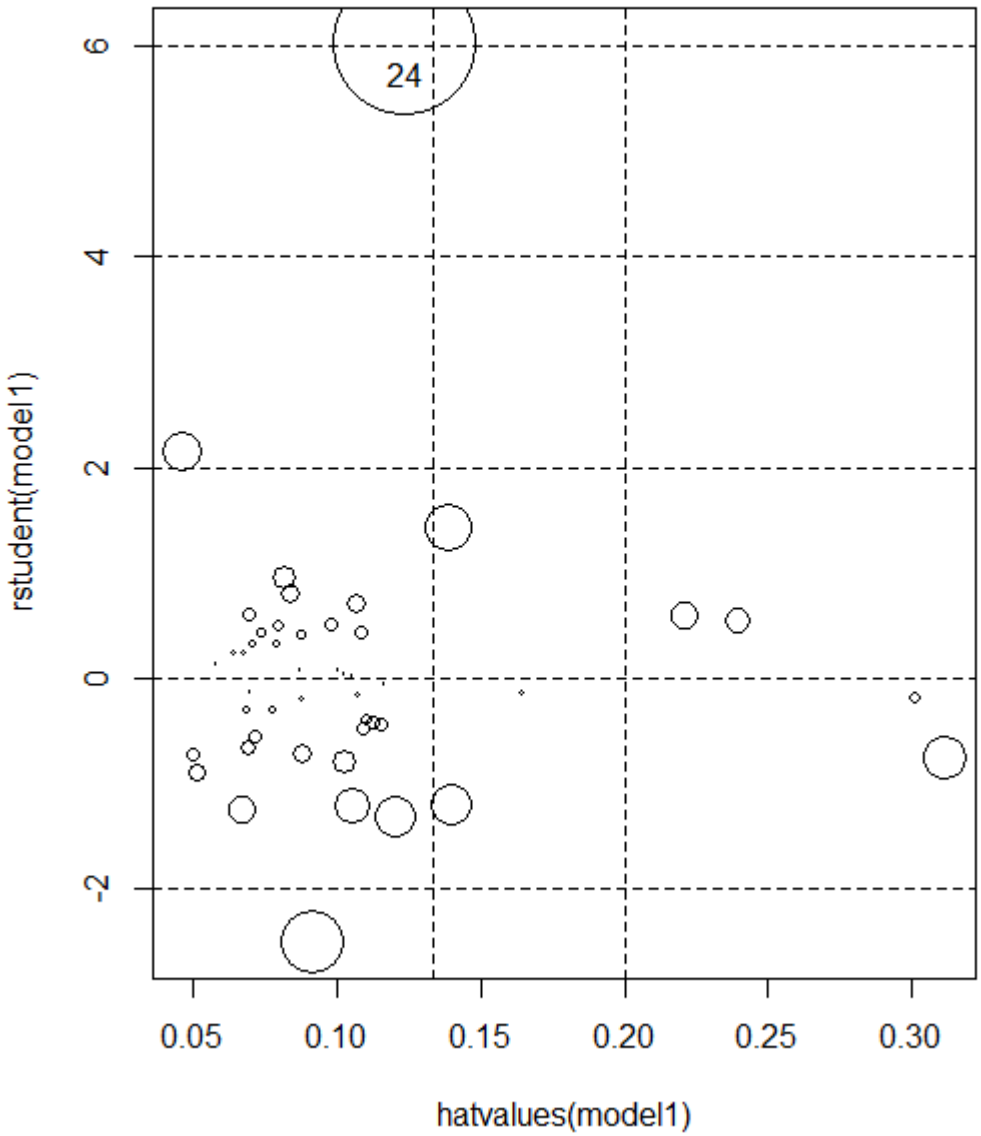
```
library(car)
outlierTest(model1)
```

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

If we assume that the alpha is 0.05, the adjusted p value (Bonferroni p) is 1.93×10^{-5}, which is lower than the alpha, illustrating that the observation 24 has extreme residuals which is both high leverage and influential.

(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))
```



It is seen that the most influential point is 24 as it has the largest regression residuals and the largest cook's distance (shown by the size).