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:

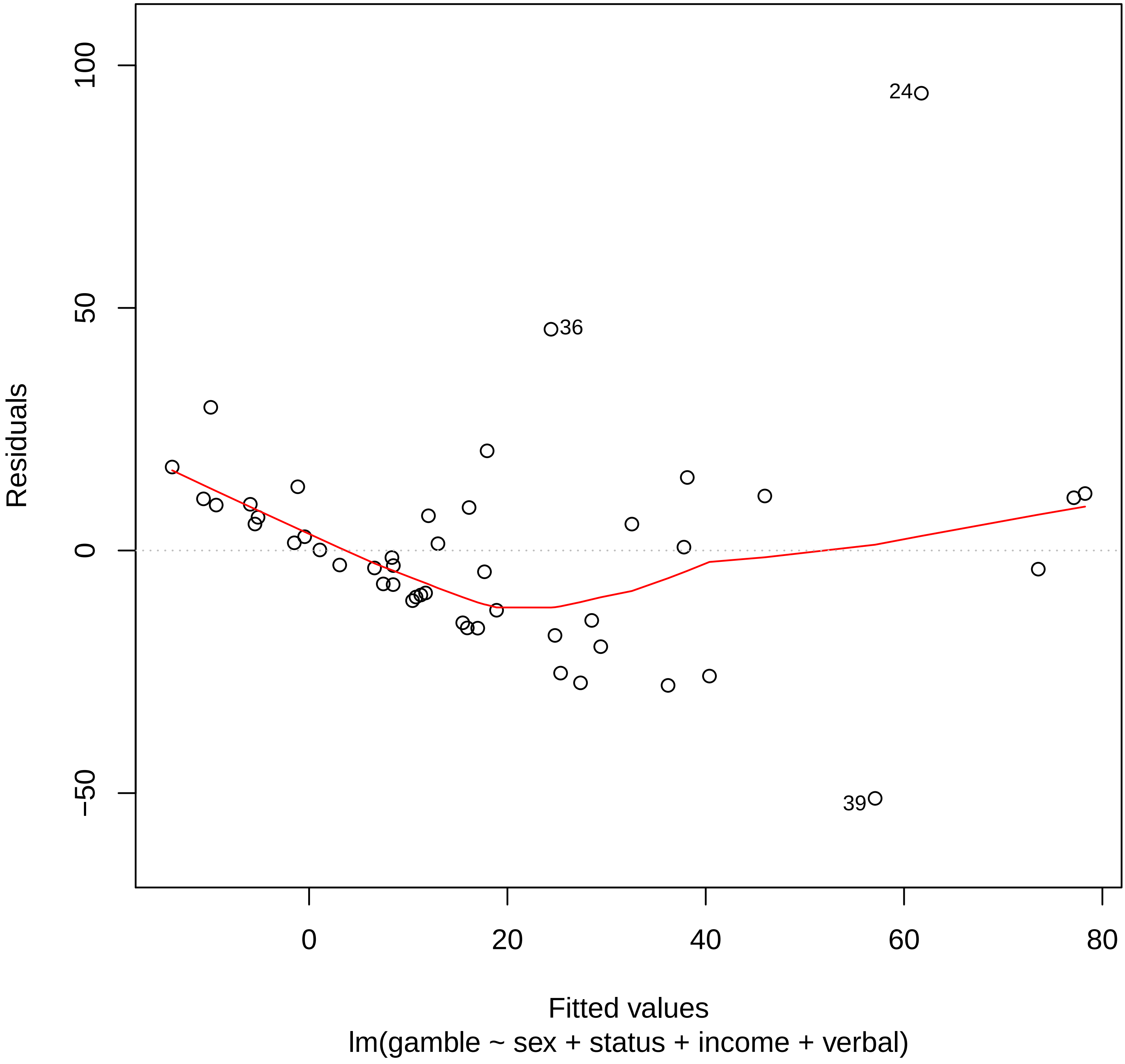
(a) Check the constant variance assumption for the errors by plotting the residuals versus the fitted values.

```r
# a) Check the constant variance assumption for the errors by plotting the residuals
#   versus the fitted values.
model1.resid <- resid(model1)
pdf("ResidFVplot.pdf")
plot(model1, model1.resid, which = 1)
```

```r
dev.off()

# The line of the graph is not very straight, indicating that there might not be a linear
# relationship among residuals. It also appears that residuals aren't necessarily constantly
# variable along the line— they seem to have increasing variance as y increases.
```

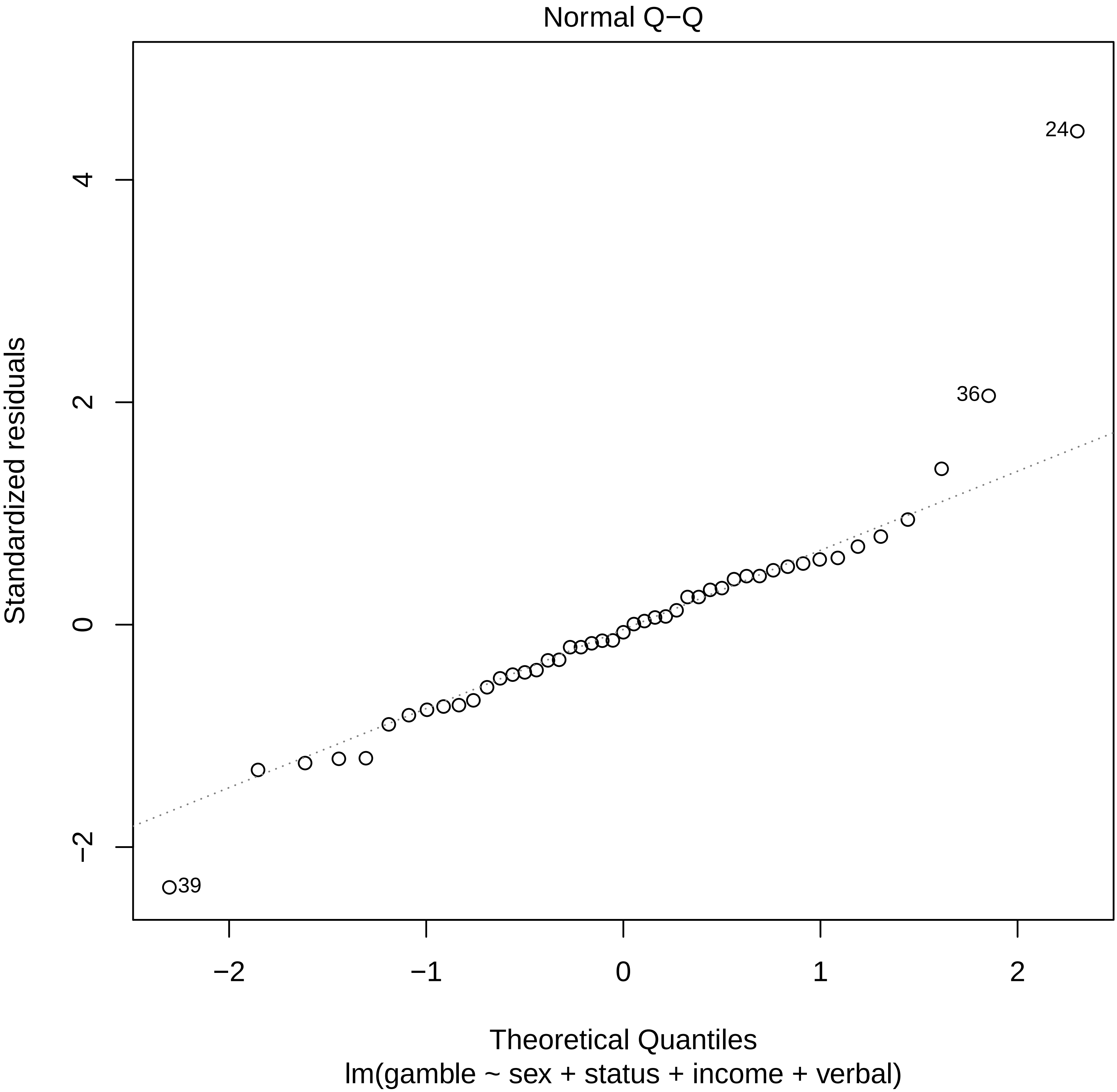
10.6507429929447



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

```
# b) Check the normality assumption with a Q-Q plot of the studentized residuals.
pdf(”QQplot.pdf”)
plot(model1, which = 2)
dev.off()

# The tails deviate from the diagonal line, having ”heavier” values in the upper and lower
# ends of the line and creating a steeper-looking slope. This suggests that our data might
# not be distributed normally, and actually have more variation.
```

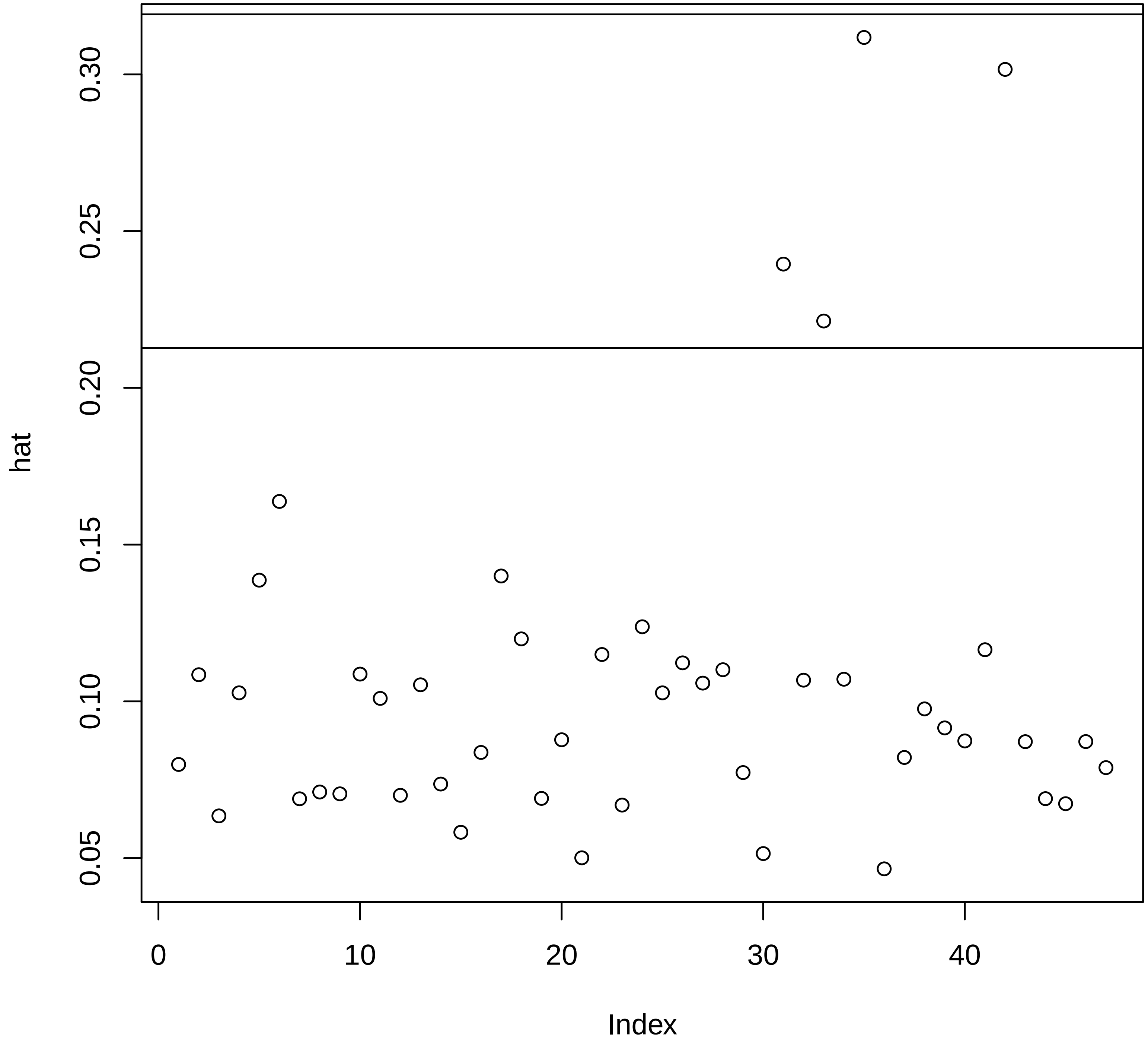

Normal Q−Q
Standardized residuals
4
2
0
−2
24
36
39
−2
−1
0
1
2
Theoretical Quantiles
lm(gamble ~ sex + status + income + verbal)

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

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

identify(gamble$sex, gamble$status, gamble$income, gamble$verbal, row.names(gamble))
hat <- lm.influence(model1)$hat
pdf(”hvalueplot.pdf”)
plot(hat)
abline(h=2*5/47)
abline(h=3*5/47)
identify(1:47, hat, row.names(gamble))
dev.off()

# There appears to be four large leverage points above the 3(k +1)/n threshold, even though
# for some reason they are not labelled with their row name.
```


hat
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
Index

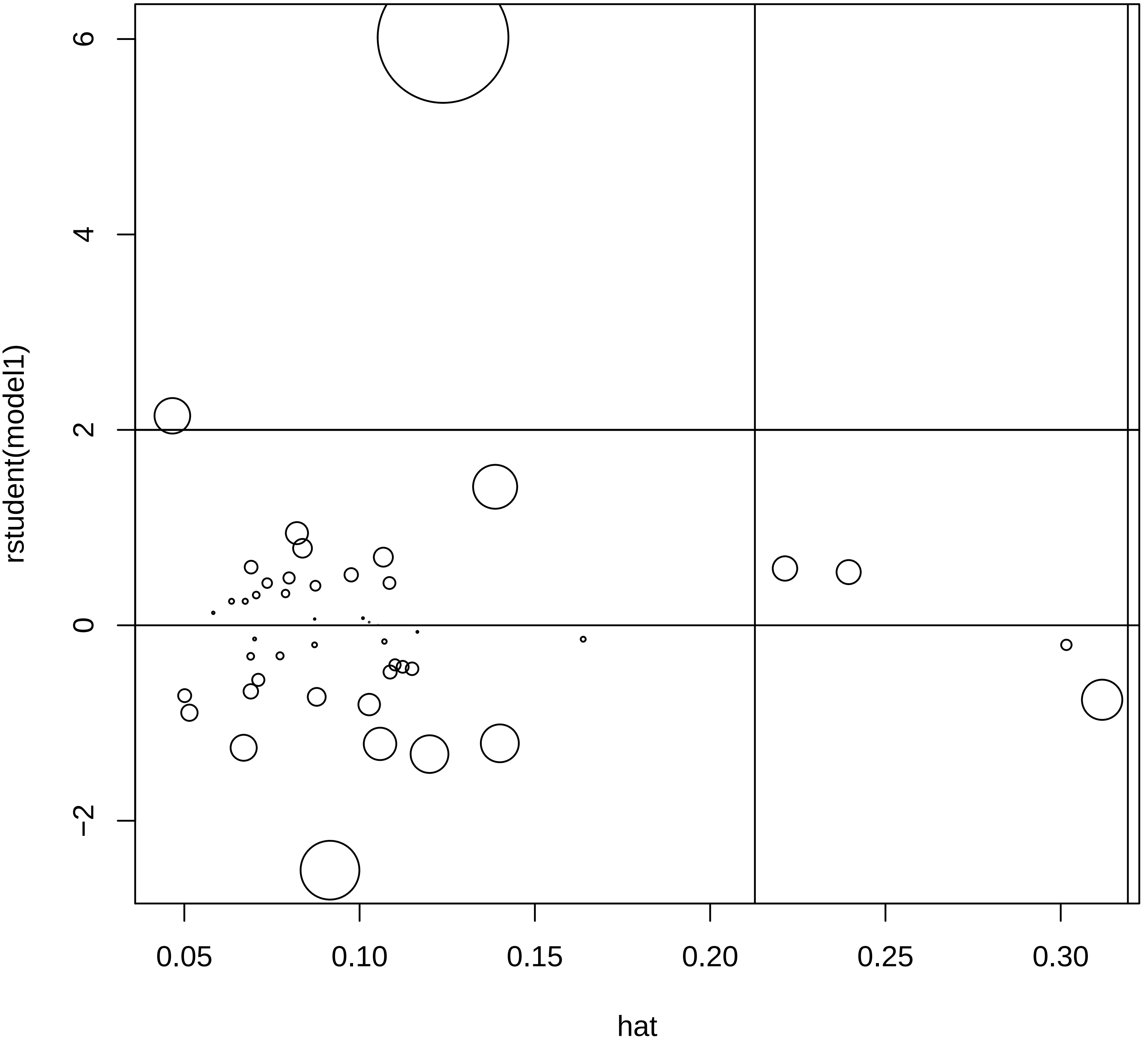
(d) Check for outliers by running an `outlierTest`.

```
# d) Check for outliers by running an outlierTest.
outlierTest(model1, data=gamble)
# The p-value is 4.1041e-07, which is very low. We can conclude that this is an extreme residual.
```

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

```
# e) Check for influential points by creating a ”bubble plot” with the hat values and
#   studentized residuals.
pdf(”bubbleplot.pdf”)
plot(hat, rstudent(model1), type = ”n”)
cook <- sqrt(cooks.distance(model1))
points(hat, rstudent(model1), cex=10*cook/max(cook))
abline(h=c(-2, 0, 2))
abline(v=c(2,3)*5/47)
identify(hat, rstudent(model1), row.names(gamble))
dev.off()

# There are a few points that seem to be influential in the model, but for some reason they are
# unlabelled.
```


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