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.

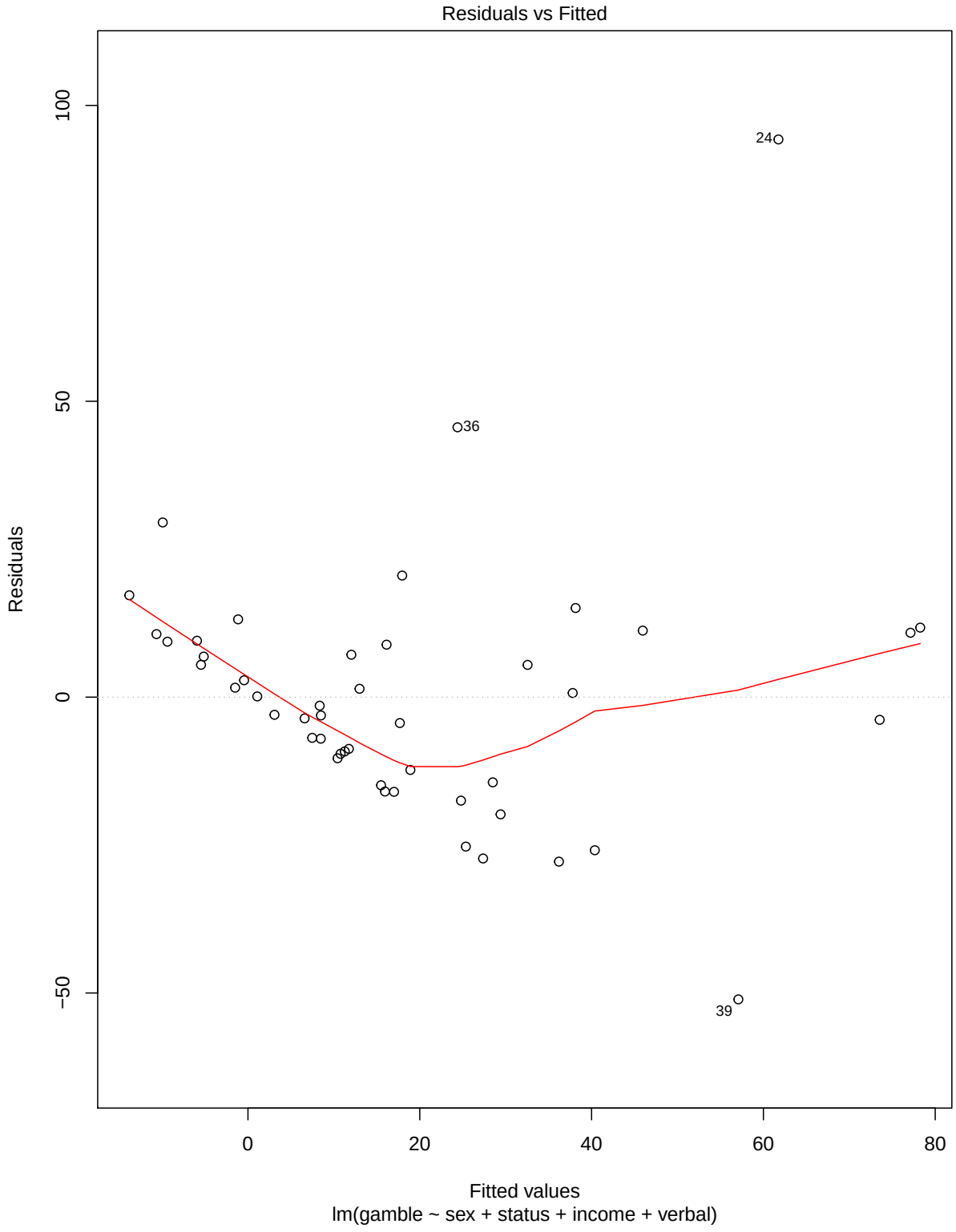
Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

```
# load data
gamble <- (data=teengamb)
# run regression on gamble with specified predictors
```

Answer the following questions:

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

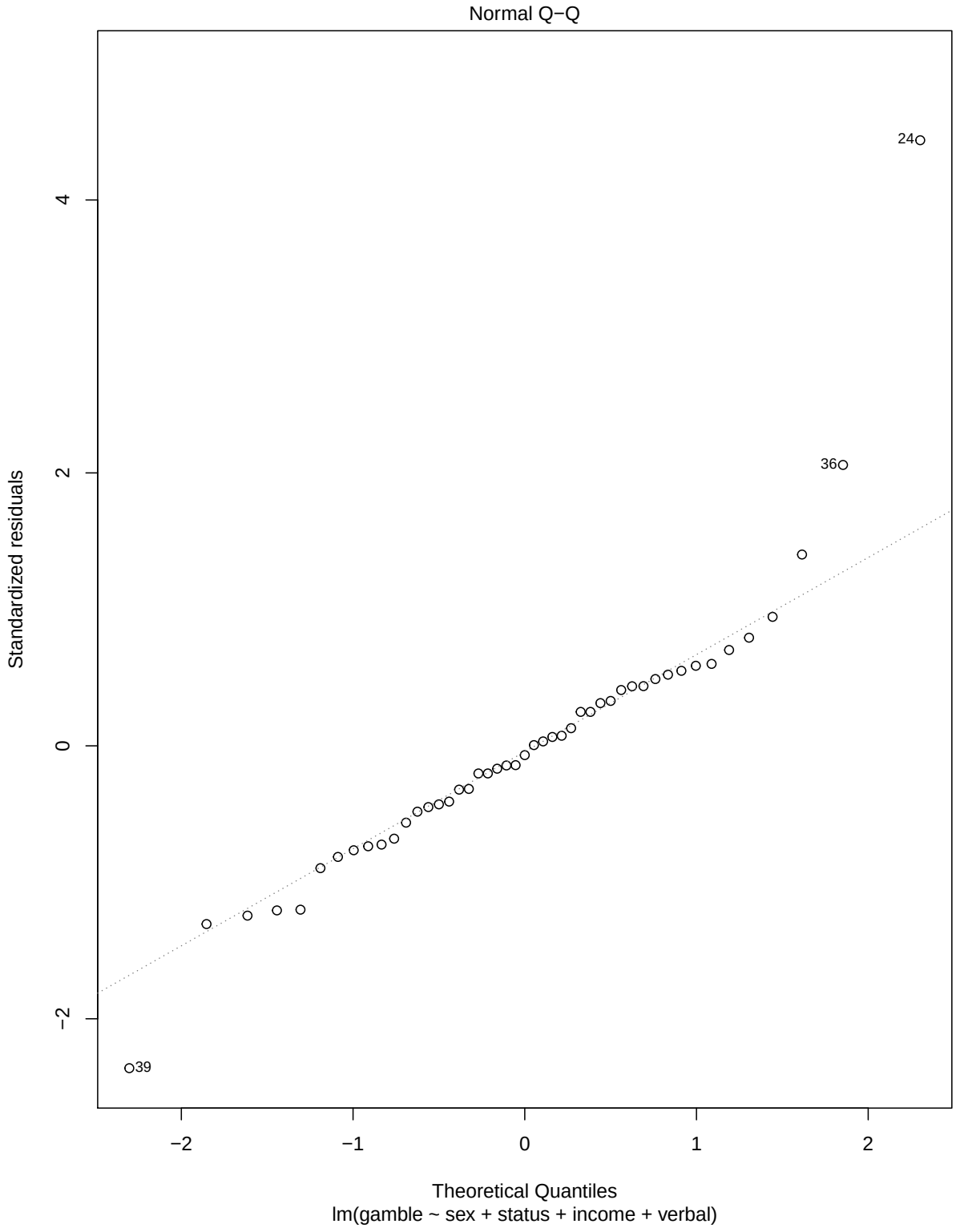
```
plot(model1)
```


Residuals vs Fitted
Residuals
100
50
0
−50
24
36
39
0
20
40
60
80
Fitted values
lm(gamble ~ sex + status + income + verbal)

Interpretation: After plotting the residuals versus the fitted values, we can see that the graph is not very linear and that the constant variance assumption may not necessarily hold for this example.

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

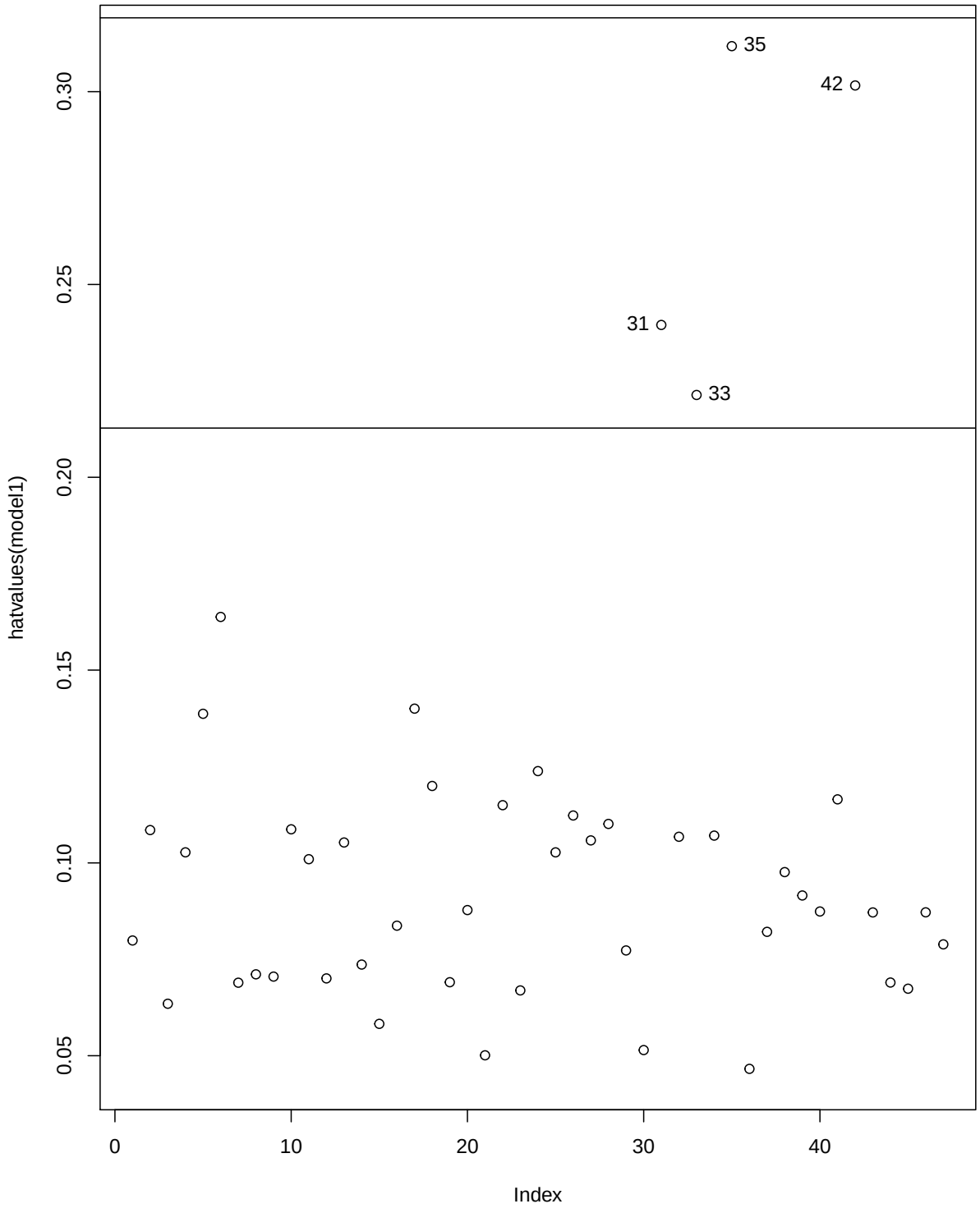
```
plot(model1)
```


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

Interpretation: After plotting a Q-Q plot of the studentized residuals, we can see that normality assumption is met because the vast majority of the observations fall on the line. Those observations that don't fall on the line happen to be at the tail ends where a less amount of data exists.

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

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


35
42
31
33
0.30
0.25
0.20
0.15
0.10
0.05
hatvalues(model1)
0
10
20
30
40
Index

	sex	status	income	verbal	gamble
31	0	18	12.00	2	88.00
33	0	38	15.00	7	90.00
35	0	28	1.50	1	14.10
42	0	61	15.00	9	69.70

Interpretation: Our analysis shows four large leverage points. These points are outlined on the graph above and display the characteristics tabulated above.

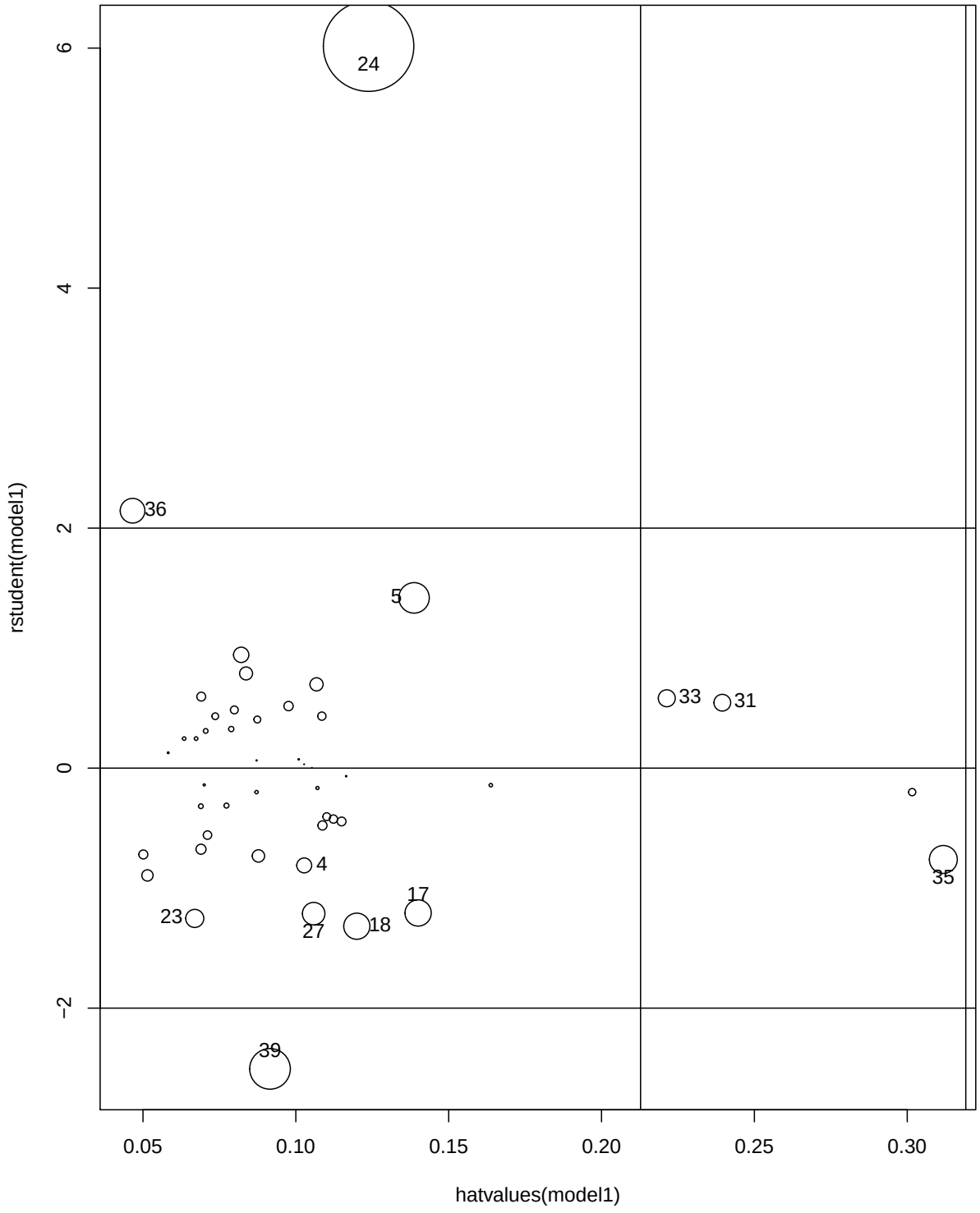
(d) Check for outliers by running an `outlierTest`.

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

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

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

```
plot(hatvalues(model1),rstudent(model1), type = "n")
cooks<-sqrt(cooks.distance(model1))
points(hatvalues(model1),rstudent(model1), cex=10*cooks/max(cooks))
abline(h=c(-2,0,2))
abline(v=c(2,3)*5/47)
identify(hatvalues(model1), rstudent(model1), row.names(gamble))
```


rstudent(model1)
hatvalues(model1)
6
4
2
0
−2
0.05
0.10
0.15
0.20
0.25
0.30
24
36
5
33
31
35
4
17
23
27
18
39

Interpretation: There are several influential points that were discovered after creating a "Bubble plot". They are outlined on the graph above.