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.

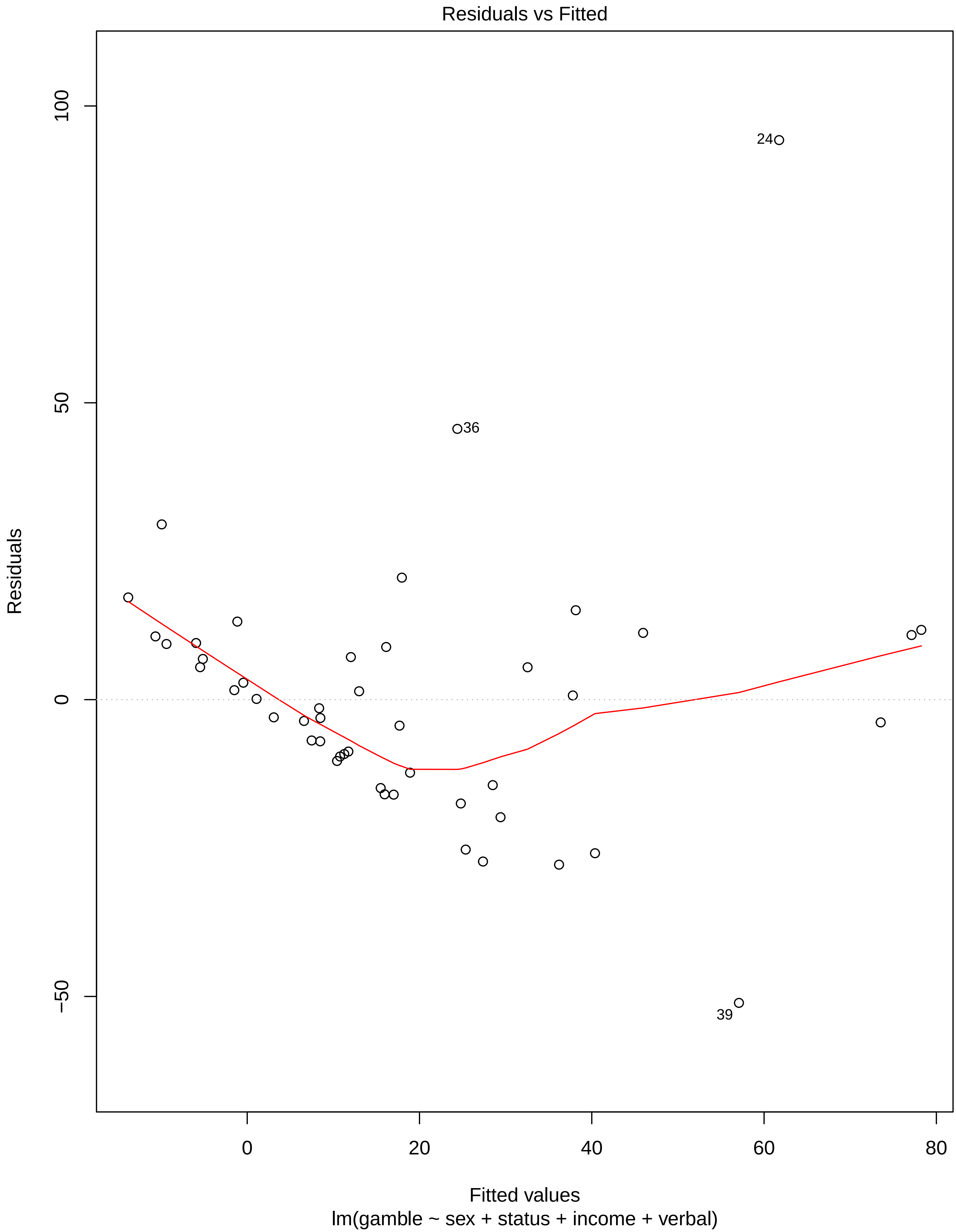
Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

```r
# load data
gamble <- (data=teengamb)
# run regression on gamble with specified predictors
model1 <- lm(gamble ~ sex + status + income + verbal, gamble)
```

Answer the following questions:

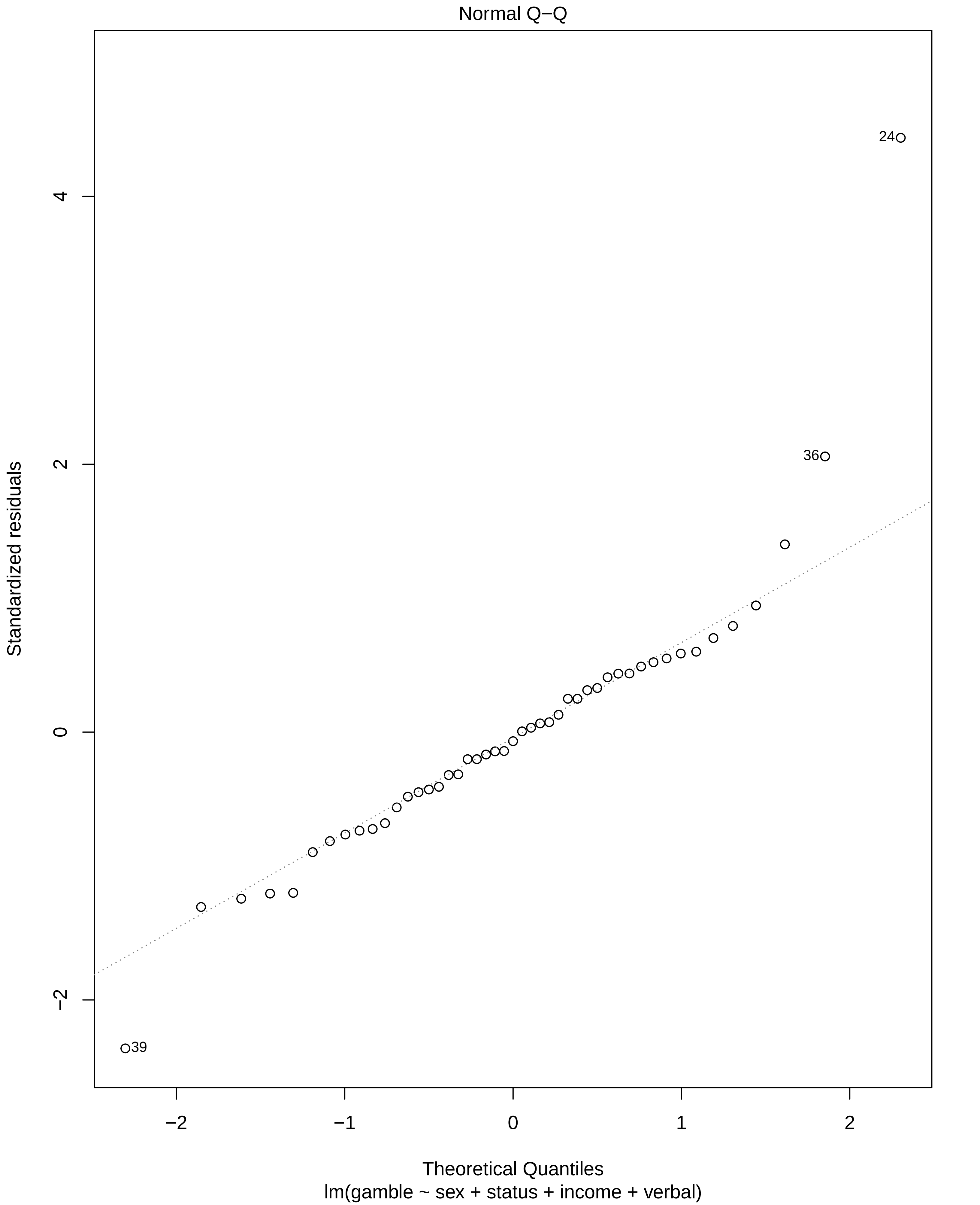
(a) Check the constant variance assumption for the errors by plotting the residuals versus the fitted values.

```r
plot(model1)
#The variance is more or less constant, as evidenced by the fact that the residuals average to about 0 at each fitted value. However, it is not perfect. There also appear to be three notable outliers.
```


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

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

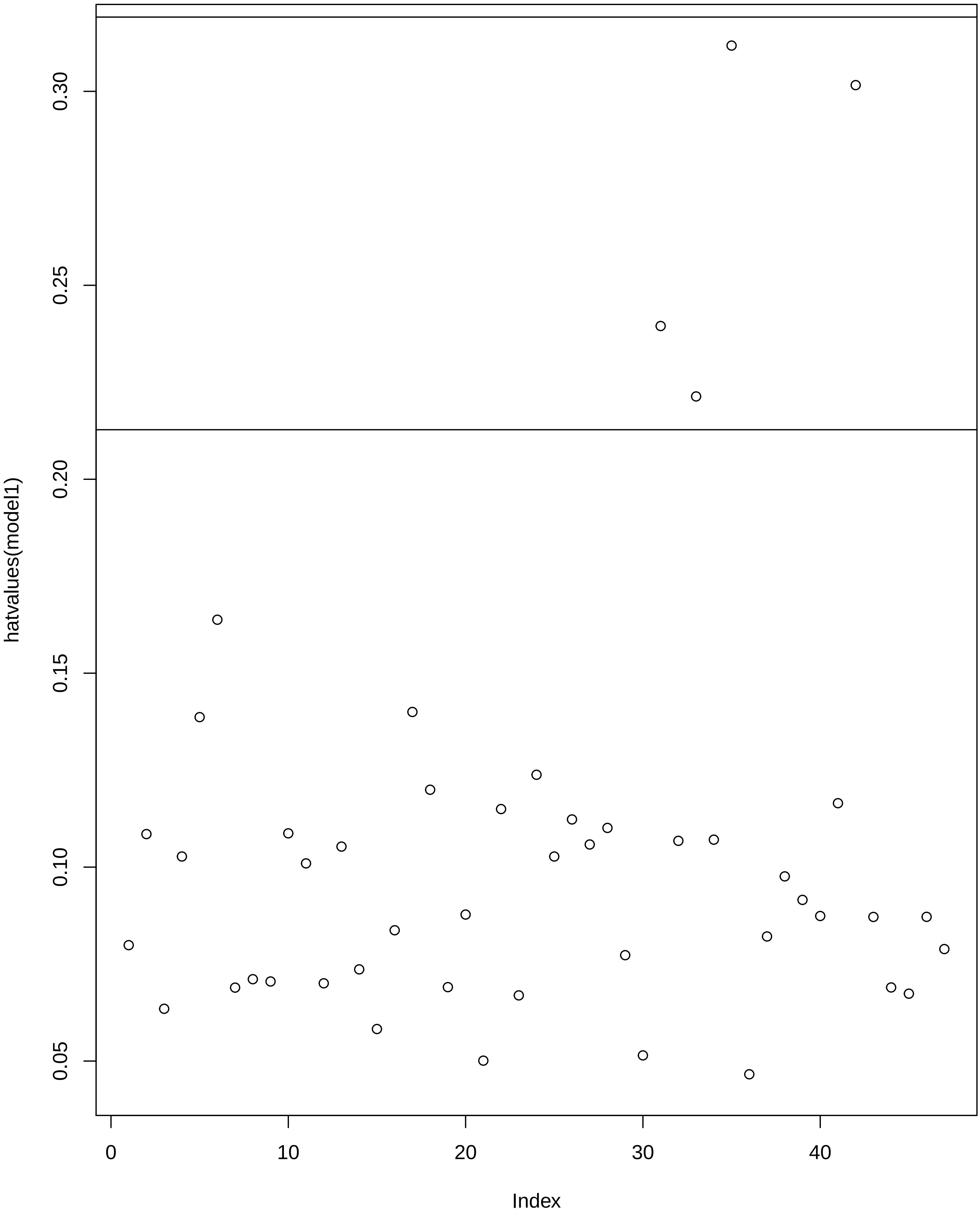
```r
plot(model1)
#The data appears to be generally normally distributed at each value of x, but the same three outliers appear again.
```


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

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

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

#There are four points with high hat values that have high leverage and thus potential to influence the model
```

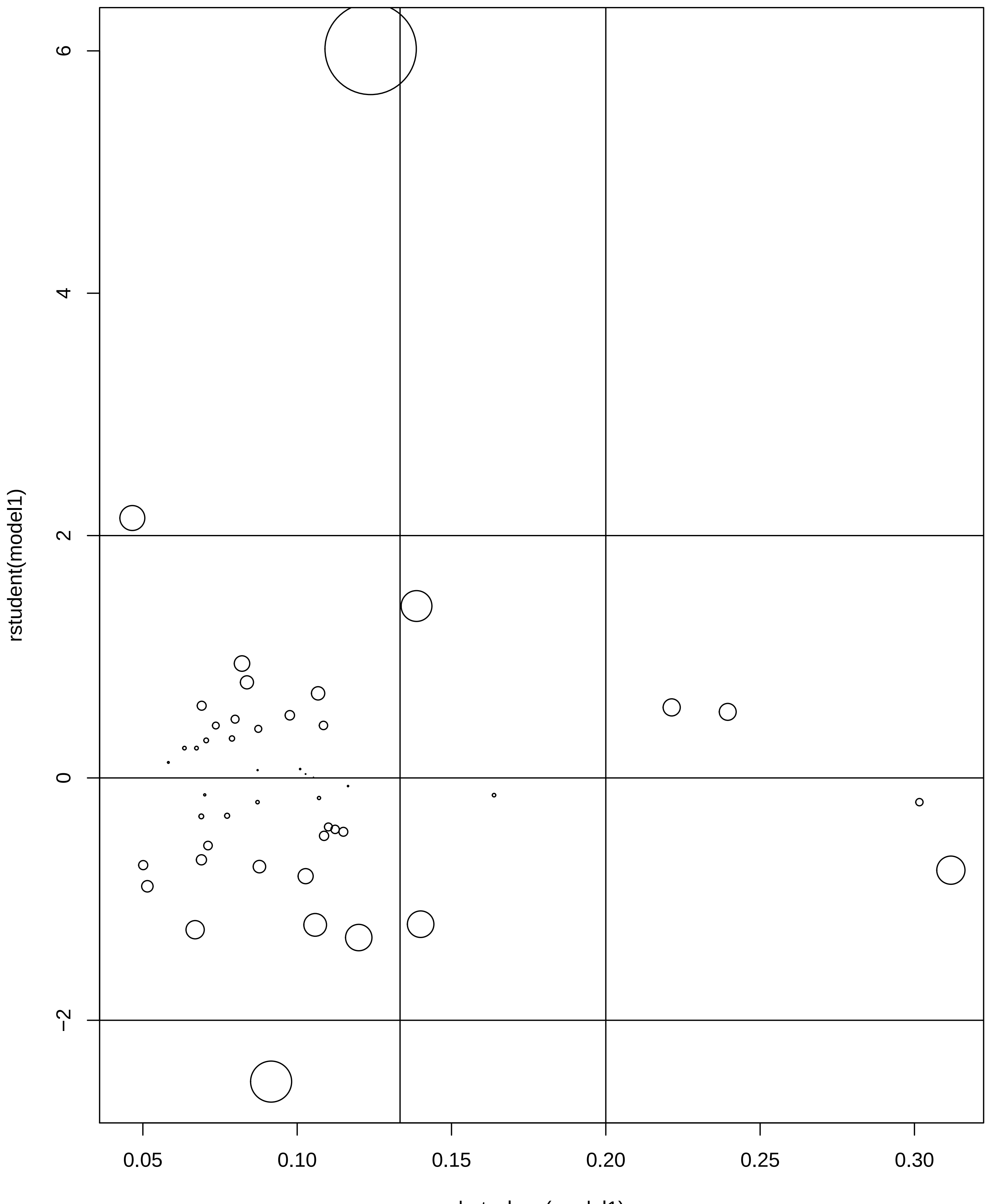

hatvalues(model1)
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
Index

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

```
outlierTest(model1)
#Given the very small Bonferroni p value (1.9289*10^-5), it we reject the
    null hypothesis that there are no outliers, because the probability of
    getting these results if there were no outliers is extremely low
```

(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))
abline(v=c(2,3)*3/45)
#There is point with a very large large Cook's distance and studentized
    residual, indicating that despite its relatively unremarkable hat value
     it is quite influential. Otherwise, there are several other points
    that have either a large studentized residual or a large hat value but
    never both (and usually a fairly reasonable Cook's distance),
    indicating that none of these outliers is highly influential.
```


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