

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.

```
1 gamble <- (data=teengamb)
2 # run regression on gamble with specified predictors
3 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.
- (b) Check the normality assumption with a Q-Q plot of the studentized residuals.

```

1 #answered the first two questions together with explanation of the 4 plot
2 # run regression on gamble with specified predictors
3 #Constant variance assumption for the errors use plot residuals vs. the
  fitted values
4 modell <- lm(gamble ~ sex + status + income + verbal, gamble)
5 #The residuals vs. fitted values is a simple scatterplot between residuals
  and predicted values, which looks random with outliers in the up right
  corner.
6 #Check the normality assumption use Q-Q plot of the studentized residuals
7 #The QQ plot has a straight line with points 24, 36 and 39 deviate from
  the straight line.
8 plot(modell)
9 par(mfrow=c(2,2), plot(modell))
10 summary(modell)
11 #The Scale-Location plot, like the the first, looks random.
12 #The Cooks distance plot tells us which points have the greatest influence
  on the regression. And we see that points 24, 5 and 39 have great
  influence on the model.

```

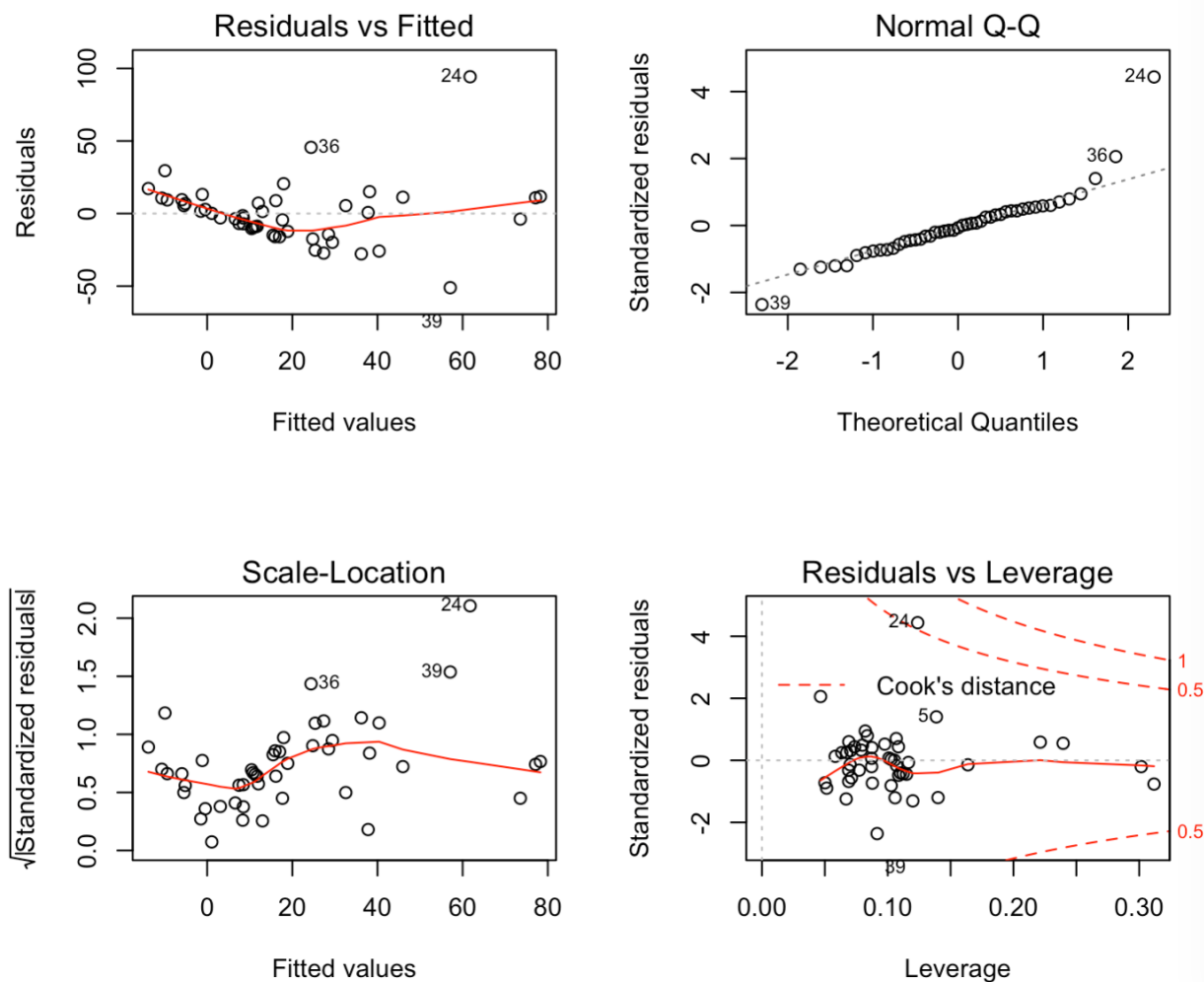


Figure 1: Diagnose of model.

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

```

1 #Check for large leverage points and plot h values
2 hatvalues(modell)
3 plot(hatvalues(modell) , pch=16)
4 #Plot the hat values with thresholds
5 abline (h=2*3/47 , lty =2)
6 abline (h=3*3/47 , lty =2)
7 identify(1:47, hatvalues(modell))

```

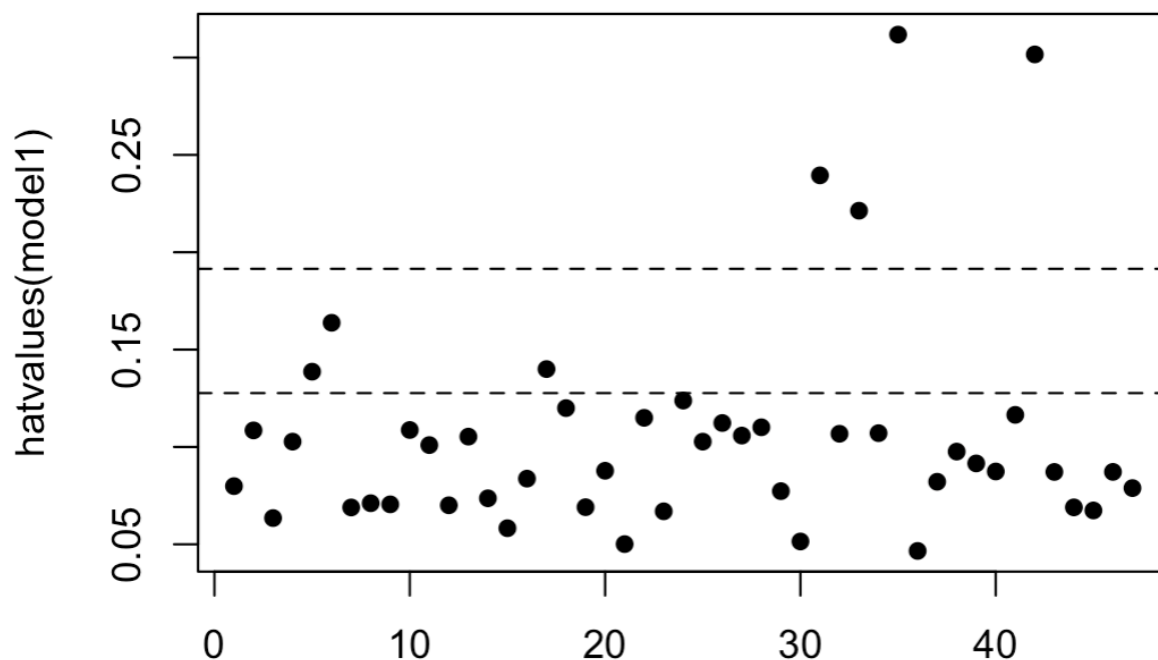


Figure 2: h value plot.

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

```

1 #check for outliers and run outlierTest
2 library(car)
3 sort(rstudent(model1))
4 outlierTest(model1, row.names(gamble))
5 # rstudent unadjusted p-value Bonferroni p
6 #24 6.016116      4.1041e-07    1.9289e-05
7 #39 -2.5060898    1.6269e-02    7.6464e-01
8 #36 2.1448259     3.7942e-02           NA
9 #three outliers listed from most to least

```

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

```

1 #Check for influential points by creating "Bubble plot" with the hat-
  values and studentized residuals
2 plot(hatvalues(model1), rstudent(model1), type = "n")

```

```

3 cook<-sqrt(cooks.distance(model1))
4 #bring together leverage, studentized residuals, and cooks distance
5 points(hatvalues(model1), rstudent(model1), cex=10*cook/max(cook))
6 abline(h=c(-2,0,2), lty=2)
7 abline(v=c(2,3)*3/47, lty=2)
8 identify(hatvalues(model1), rstudent(model1), row.names(gamble))

```

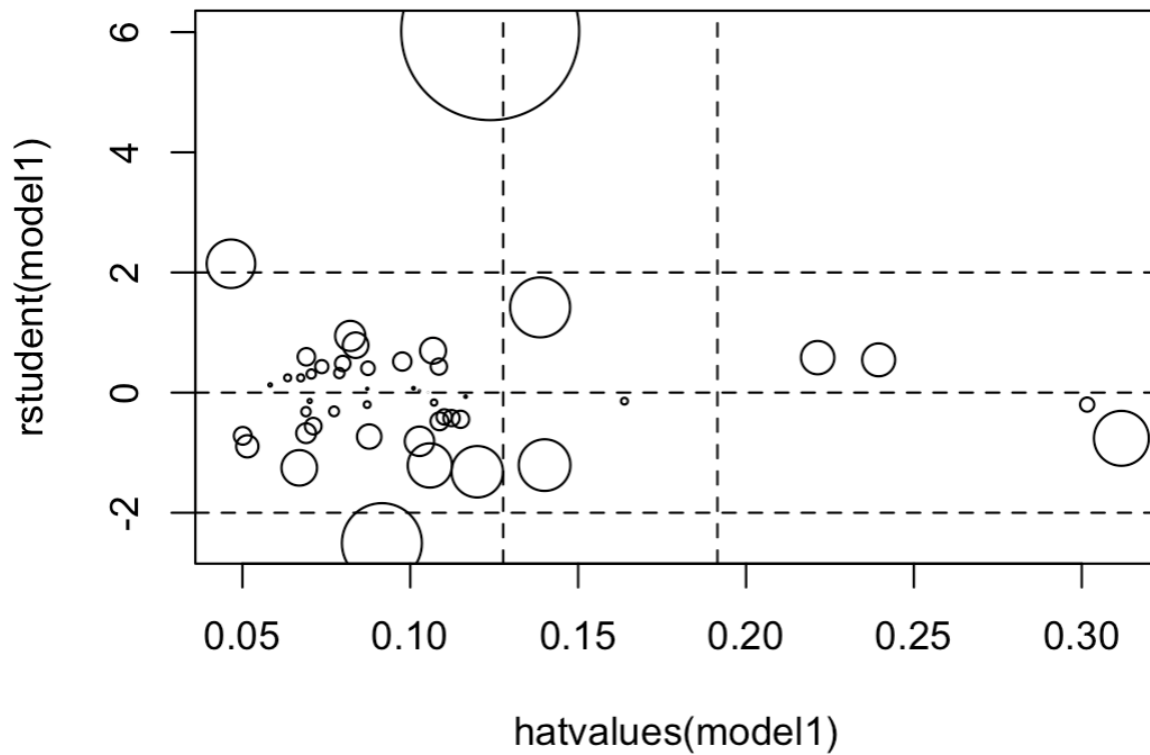


Figure 3: Bubble plot.