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.

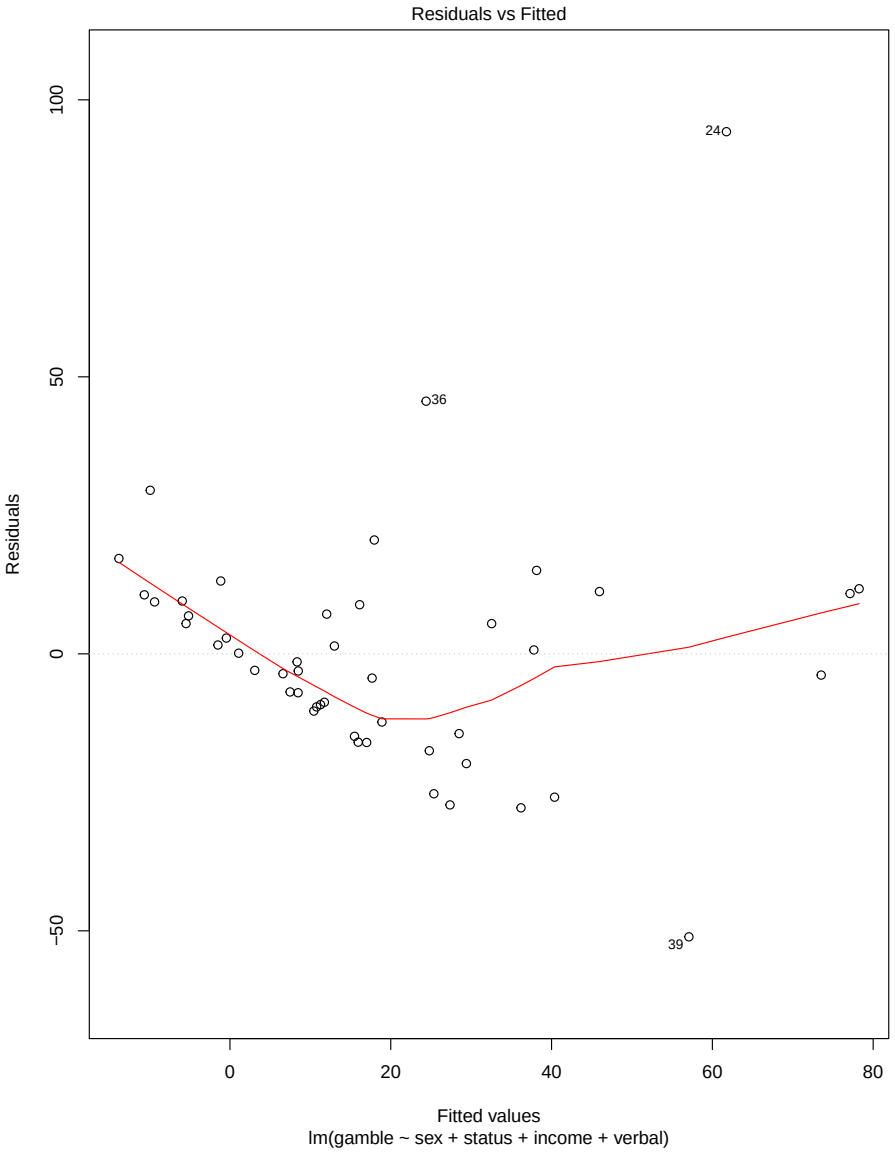
Using the `teengamb` dataset, fit a model with `gamble` as the response and the other variables as predictors.

```
        cex=10*cook/max(cook))
abline(h=c(-2,0,2), lty = 2)
abline(v=c(2,3) * 3/45, lty = 2)
```

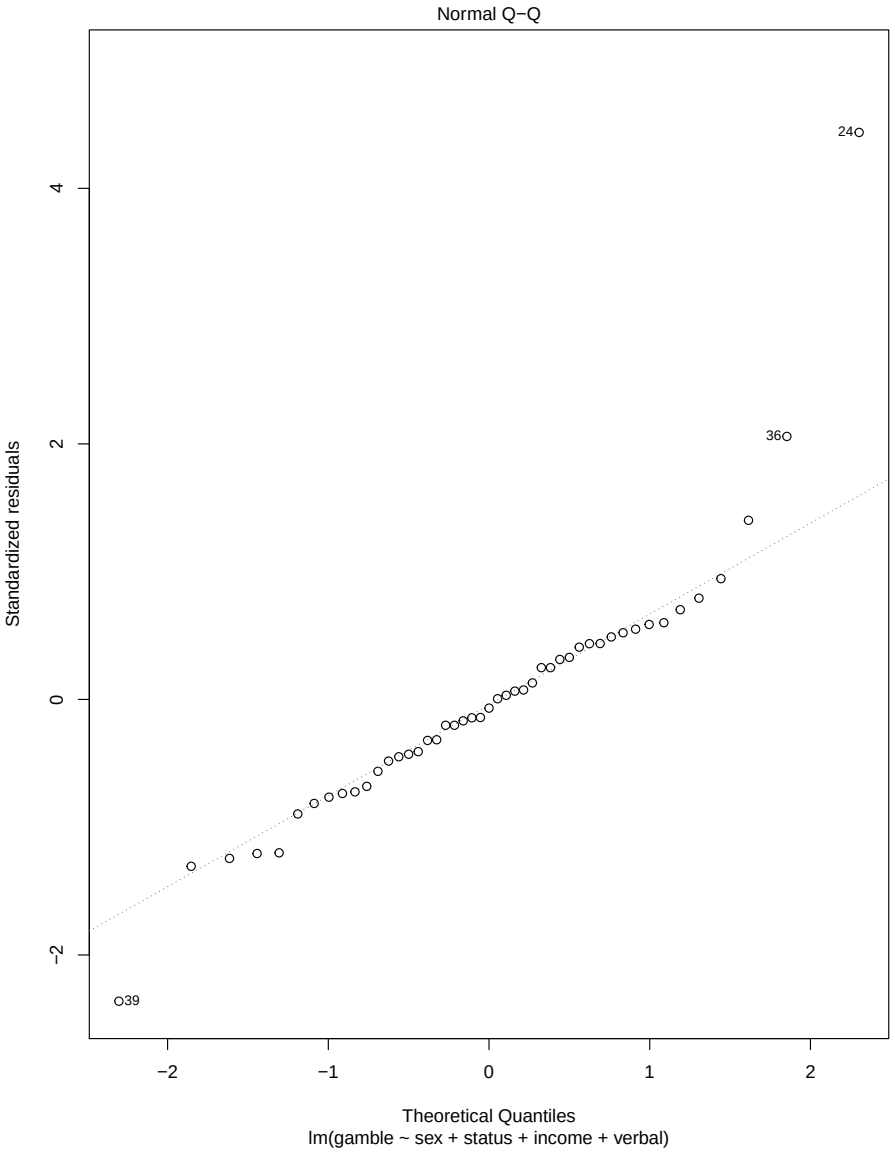
Answer the following questions:

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

Constant variance illustrates the individual error against the predicted value, as the variance of the errors should be constant. Our plot isn't the cleanest, but overall would show a general trend of constant variance for most of the value points. There are some existing outlier values however, as indicated by the label above their data point, and distance from the residual axis of 0.

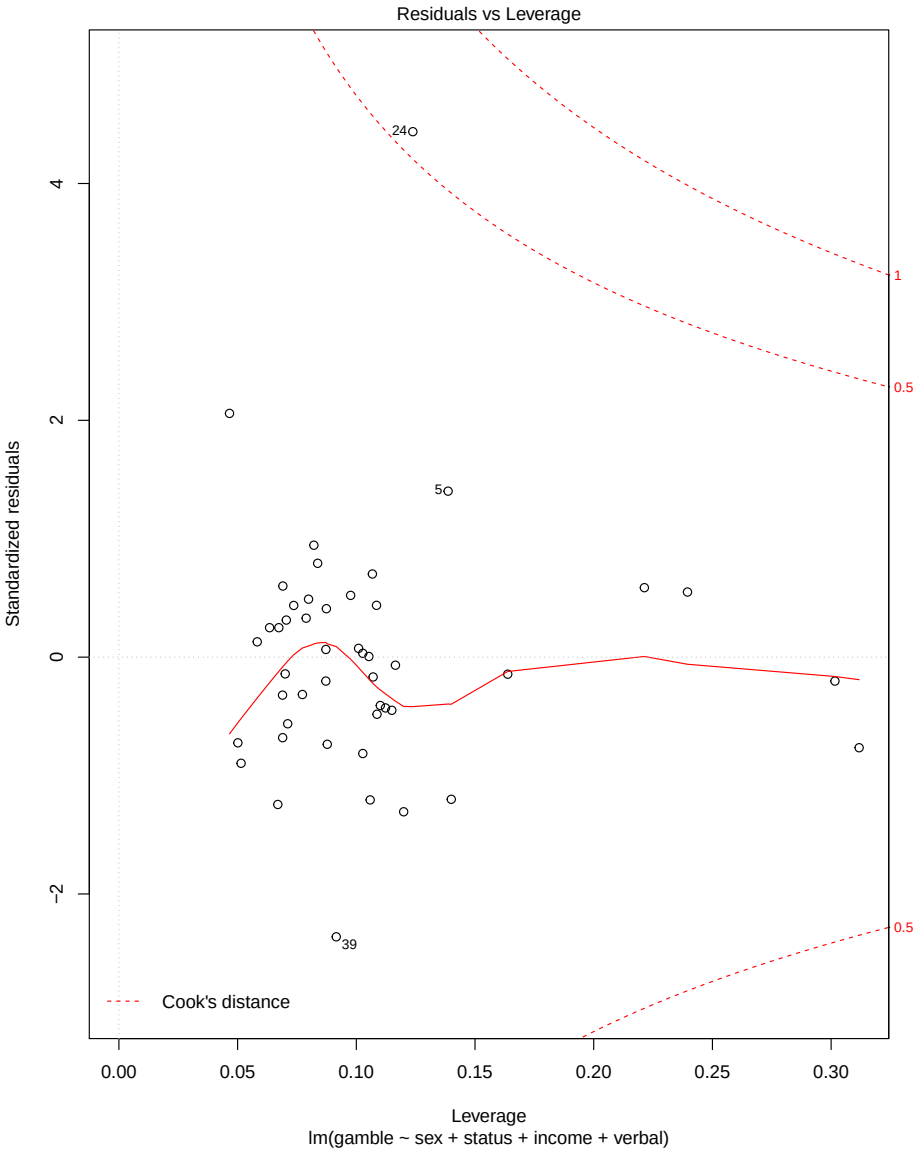


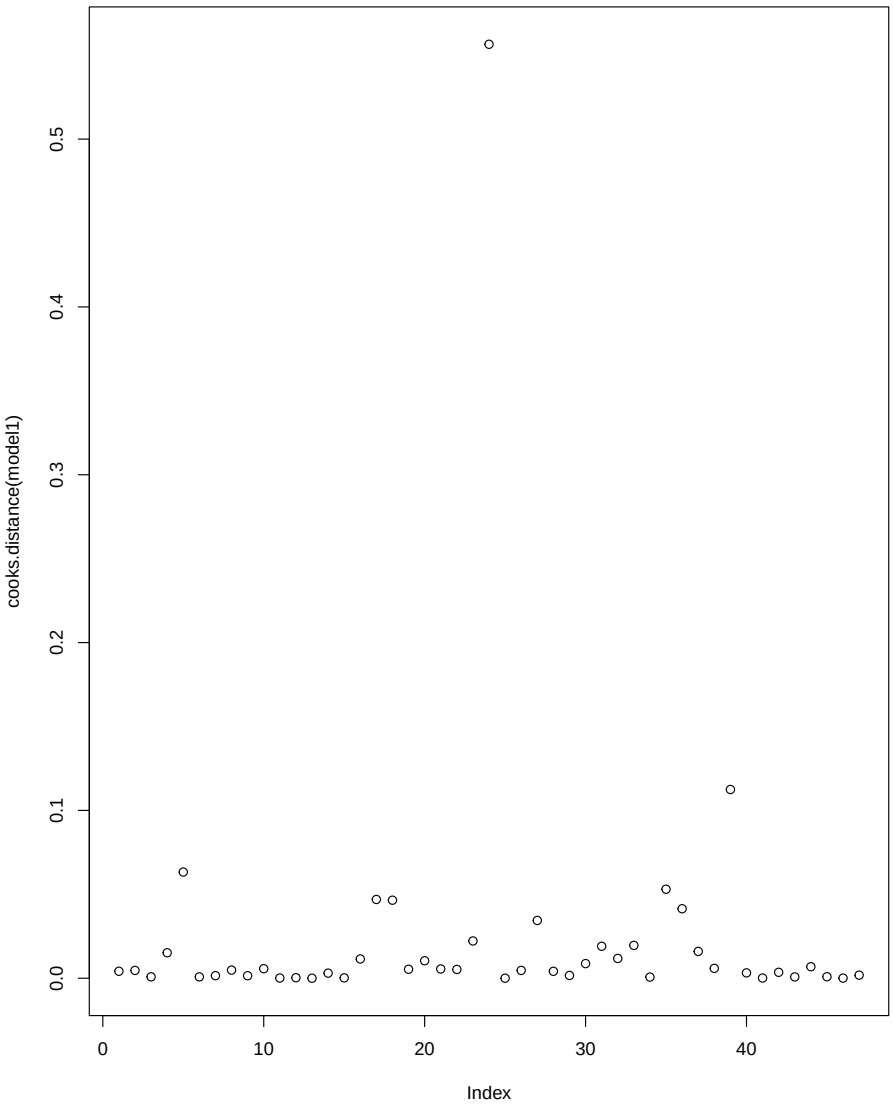
(b) Check the normality assumption with a Q-Q plot of the studentized residuals.



A QQ-plot illustrates the distribution of the data against the expected normal distribution, and normally distributed data will exist on the straight line. On a general trend, we see that there is a relatively approximate straight line with the data points aside from about three data points. A curvature of these data points would indicate that the data is not normal, but at a holistic view, this data seems relatively normal.

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


Residuals vs Leverage
Standardized residuals
Leverage
lm(gamble ~ sex + status + income + verbal)
Cook's distance
24
5
39
1
0.5
0.5
4
2
0
−2
0.00
0.05
0.10
0.15
0.20
0.25
0.30



In these plots, we can see where the outliers exist, especially when observing Cook's distance. Cook's distance shows the influence of each observation on the fitted response values, and can identify an outlier based on it's distance from the axis. There seems to be outlier values through these plots due to this great distance.

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

```
outlierTest(model1)
```

The value at 24, has a large studentized residual above 6, an unadjusted p-value is 4.1041e-07 while Bonferroni p-value is 1.9289e-05. Because it is smaller than a p-value of .05, we reject the null and conclude that in this model, the value 24 in the gamble dataset, is an extreme residual

(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))
```