Answer Key: Problem Set 5

QTM 200: Applied Regression Analysis

Jeffrey Ziegler

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 Monday, February 10, 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
model1 <- lm(gamble ~ sex + status + income + verbal, data=gamble)
```

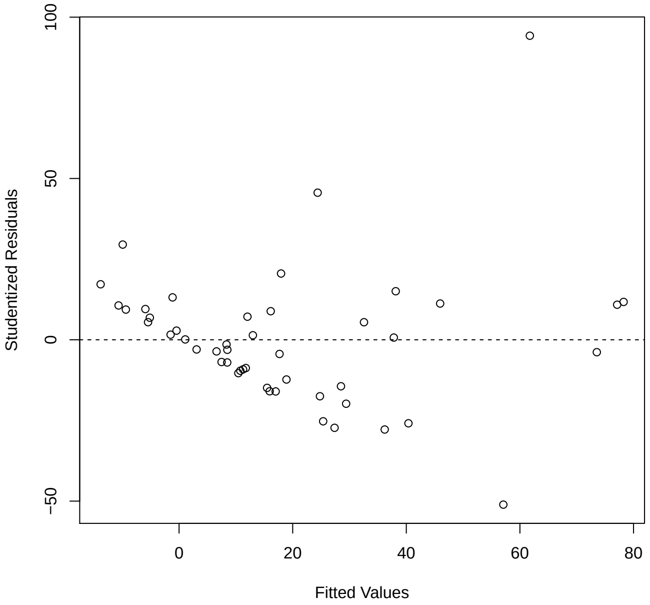
Answer the following questions:

1. Check the constant variance assumption for the errors by plotting the residuals versus the fitted values.

```r
pdf("constant_variance1.pdf")
plot(fitted(model1), resid(model1),
     ylab="Studentized Residuals", xlab="Fitted Values")
abline(0, 0, lty=2)
dev.off()
```

Figure 1 plots differences between the predicted and observed values (studentized residuals) and the fitted values. If the assumption of constant variance for the disturbances was met, the studentized residuals would generally be constant, and we can see there is not much of a linear relationship and they are centered around zero (disturbances seem to average to zero).

Figure 1: Scatter plot of studentized residuals and fitted values from `model1`.

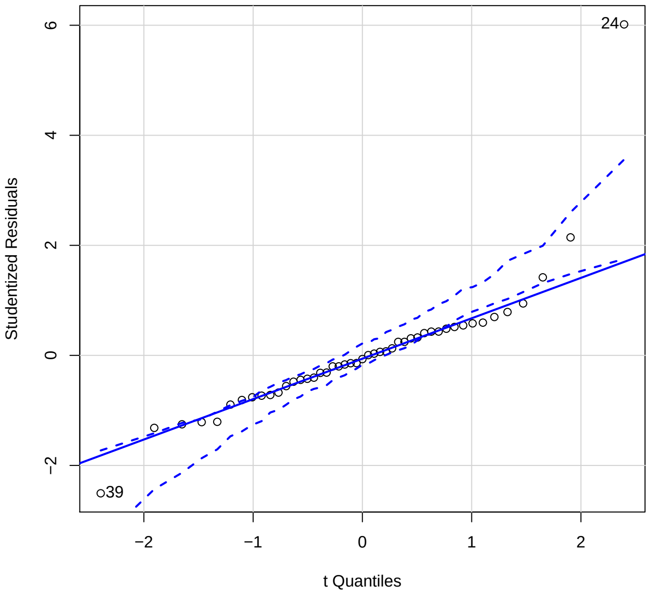


2. Check the normality assumption with a Q-Q plot of the studentized residuals.

```r
pdf("constant_variance1.pdf")
plot(fitted(model1), resid(model1),
     ylab="Studentized Residuals", xlab="Fitted Values")
abline(0, 0, lty=2)
dev.off()
```

Figure 2 depicts the quantile comparison of the studentized residuals. If the assumption of normality for the disturbances was met, the studentized residuals would generally be within the confidence envelopes, but as we can see there is one distinctive outlier as well as other that are just slightly outside the envelopes.

Figure 2: Scatter plot of studentized residuals and fitted values from `model1`.



3. Check for large leverage points by plotting the h values.

Figure 3: Hat values from `model1`.

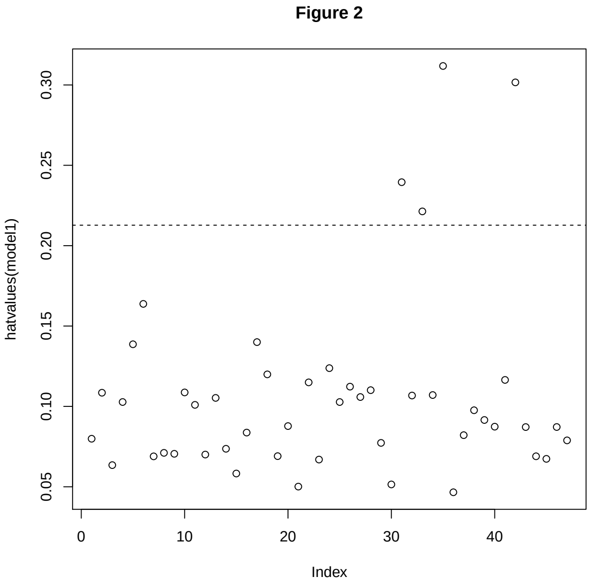


Table 1: Observations with concerning hat values.

Observation	Hat value
31	0.24
33	0.22
35	0.31
42	0.30

To check for leverage points, we can look at the hat values of each observation (remember that hat values of concern are $h_i > 2\bar{h}$). Looking at Figure 4 and Table 1, there are at least 4 observations that may be exerting leverage on our linear estimation.

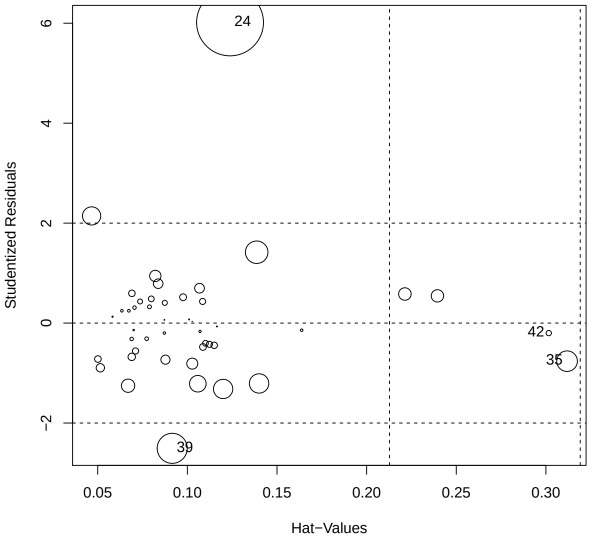
4. Check for outliers by running an `outlierTest`.

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

   We can reject the null hypothesis that observation 24 is not a potentially influential outlier.

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

Figure 4: Bubble plot of hat values and studentized residuals from `model1`.


Studentized Residuals
Hat−Values