Assignment 1 — due January 28, 2019

1. (Use R to do your work for this question.) The data set `outlier.txt` has three columns, `y`, `x`, and `indic`. The latter was carefully constructed, as you will see.

   (a) Fit a simple linear regression of `y` on `x`; examine the externally studentized residuals in a plot, and conduct a formal outlier test based on the externally studentized residuals.

   (b) Focus on the 9th observation corresponding to $x = 12$ and $y = 26$. Here is one way to test whether it is an outlier.

   i. Delete the questionable point from the data set. This leaves us with an "edited" data set.

   ii. Fit the regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ to the edited data set.

   iii. Use the regression line from the edited data set to *predict* the Y-value that would correspond to $x_* = 12$. Also, based on that edited data set, calculate the standard error for this prediction.

   $$\text{se}(\hat{Y}_{\text{pred}}) = s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

   (You can extract this value from an `lm` object in R, but it might be easier to just calculate it directly from the formula.)

   iv. Test whether the questionable observation is within sampling error of its predicted value:

   $$T = \frac{Y_{\text{obs}} - \hat{Y}_{\text{pred}}}{\text{se}(\hat{Y}_{\text{pred}})}.$$

   where Y_{obs} is the observed Y value for the suspicious observation.

   v. Compute the p-value corresponding to this observed T statistic and apply Bonferroni.

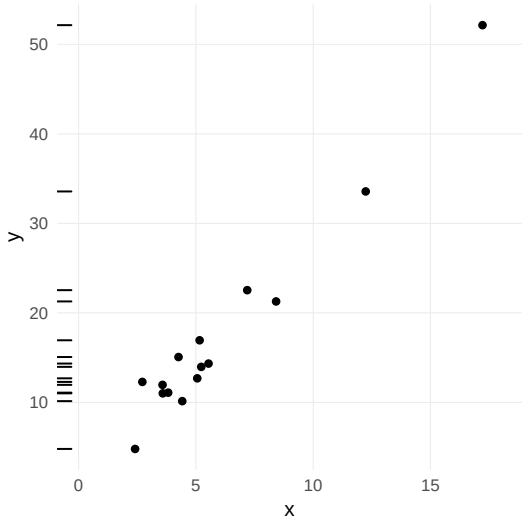
   vi. This T statistic has another name. What is it?

   (c) Now show that your outlier test above can be conducted by performing an appropriately constructed additional sum of squares test.

   i. Fit a multiple regression of `y` on `x` and `indic`. The T-value and p-value for the variable `indic` bear a striking similarity to some of the values that R produces for the outlier test. Describe how.

   ii. For an outlier test, write down formally what hypothesis is being tested.

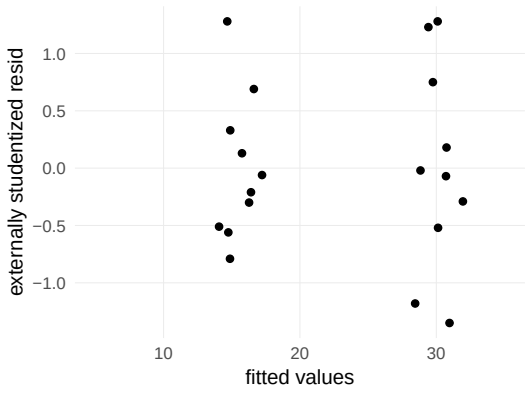
2. (a) The following plot is a plot of Y vs X for a given data set, with a rug plot of Y on the left margin.



   The display of the Y values (shown in the rug plot) is very clearly skewed. *Therefore, in a regression of Y on X, the assumption of normality is clearly violated.*

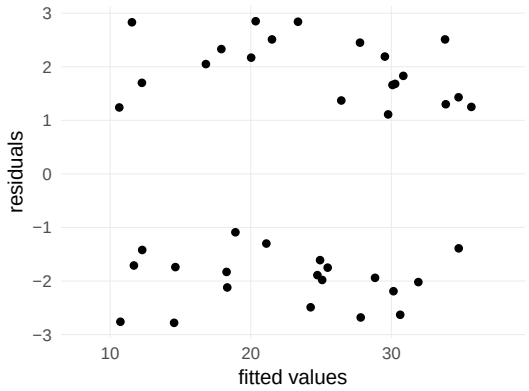
   Indicate whether the preceding italicized statement is True or False, and explain your reasoning.

(b) (10 points) Here is a residual plot of the externally studentized residuals based on a simple linear regression of y on x for a given data set.



The following statement is either True or False. Indicate whether it is True or False and explain your answer. *The data appear in two clumps, indicating a violation of at least one of the assumptions of linear regression.*

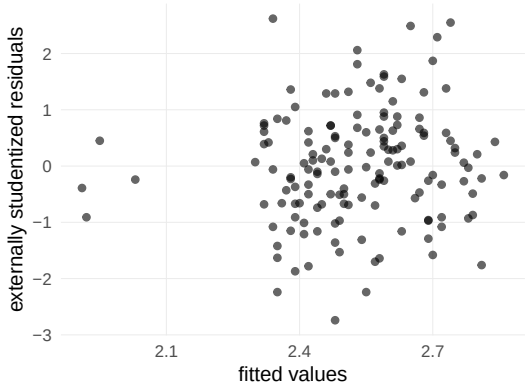
(c) This is a residual plot from fitting a simple linear regression to some data.



Based on the residual plot, the assumption of normality appears to be violated.

Explain why, and provide two distinct mechanisms that could result in a plot of this form.

(d) I fit a linear regression of the form $Y_i = \beta_0 + \beta_1 x_i + e_i$ to a data set with $n = 150$ observations. After fitting the model I made a plot of the externally studentized residuals versus the fitted values, as follows.



Consider the following statement: *When I first look at the residual plot, it seems to have a fan shape indicating unequal variance. However, after some thought, I do not believe that there is any problem with unequal variance.*

Explain why I do not believe that there is any problem with unequal variance.