DATA 511 Intro to Data Science
Course Project 2: SLR and ANOVA
Prof Larose

Name: Danielle Senechal

The project instructions are shown in bold. This is to distinguish the instructions from your work. Your work should be in not-bold.

- Work neatly. Aim for a professional-looking presentation. Make sure that you interpret all of your results, using the language shown in the notes.
- Make sure all graphs and tables fit neatly on the page.
- Neither add nor delete pages.
- No executive summary is required for this project.

For all text output, surround it with a text box. In Word, select the text output, then Insert > Text Box > Draw Text Box.

Apart from this document, which you will save as a pdf and submit, you must submit your R script, containing the code you used to solve the problems. The R script should be neat and easily understandable by people who are not you. It should be well-annotated, describing what you are doing so that anyone could understand it.

This Project is brand new, and may have typos, errors, etc, that I have missed. Please report these to me asap. For this and other reasons, this Project is subject to change at any time (though of course I will be reasonable.)

I realize that this project is not particularly challenging. So ace it. Projects 3 and 4 will be more challenging for you.

Good luck!

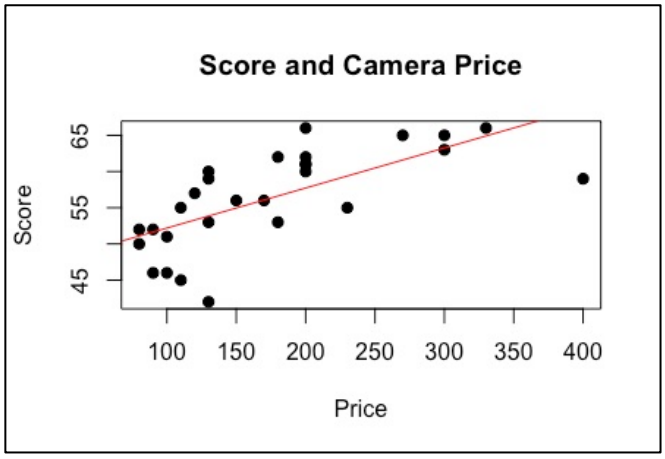
Prof Larose

Part 1: Simple Linear Regression

Import the *cameras* data set. This tiny data set represents a set of 28 cameras. We are using *Price* to estimate *Consumer Reports*' *Score*. The units are dollars and points, respectively.

1. **Regress *Score* on *Price.***
   a. **Provide the regression equation, using MS Equation Editor.**

      The regression equation for Score on Price is $Score = 46.669 + 0.053 * Price$. The plot below is of the camera data, with the red line representing the line of best fit from the regression equation.



   b. **Interpret the regression equation.**

      The estimated camera Score is equivalent to 46.669 points minus 0.053 multiplied by the camera price.

   c. **Interpret the slope coefficient of the regression equation.**

      The slope coefficient is equal to 0.053, and this means that for each unit increase in camera price, the Score will increase by 0.053 points.

2. **Continuing with the regression of *Score* on *Price.***
   a. **Estimate *Score* for a camera costing $70.**

      Using the regression equation above, $Score = 46.669 + 0.053 * 70$, and the estimated Score for a $70 camera is about 51.

   b. **State and interpret r^2.**

      The r^2 value, or the coefficient off determination, for this regression is 0.4668. This is interpreted as that 46.68% of the variability in Score is accounted for by Price.

   c. **State and interpret *s*.**

      The s value, or the residual standard error is 4.982, meaning that the size of our typical prediction error is 4.982 score points.

3. **The *standardized residuals* represent the residuals standardized so that they may be interpreted somewhat similarly to Z-scores. Suppose we define our outliers to be those cameras whose absolute standardized residual exceeds 2 (this may differ from the notes). Obtain the standardized residuals using the *rstandard* command. Clearly interpret any outliers you find. In your interpretation, don't just say "much higher" or "much lower" like it says in the notes. Instead, find out exactly how much higher or lower, using the *residuals* command.**

```
stand.cam <- rstandard(reg1); stand.cam
        17
-2.35747936
        28
-2.43635192
```

```
residuals(reg1)
        17
 -9.76847975
        28
 -11.85119725
```

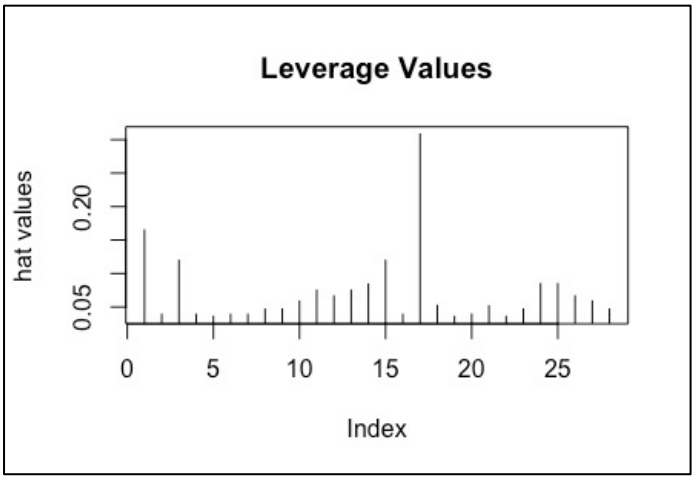
There are two outliers identified in this dataset, camera 17 and camera 28. Camera 17 is an outlier because its actual Score of 59 is considerably lower (about -9.768) than its predicted Score of 68. Camera 28 is also considered an outlier because its actual Score of 42 is much lower (about -11.851) than its predicted value of 54.

4. **The command *hatvalues* provides the leverage values for the regression. (Not to be confused with $\hat{y}$ values.) Provide a plot of the leverage values for the cameras, using something like the following command. Then interpret the camera with the greatest leverage.**

```
plot(hatvalues(reg1), type = "h")
```

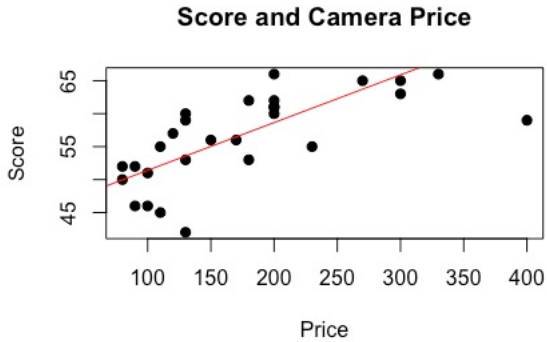
The following plot shows the leverage values on the regression of Score on Price. It can be seen that index 17 is the highest leverage point, meaning that the camera with the highest leverage is camera 17. This camera has a price of $400 but a score of 59. The price is the highest of all cameras and the score is relatively low compared to the other cameras with high prices.

The two side by side plots show the regression line that includes camera 17 (on the left) and the regression line without the camera 17 (on the right). Camera 17 can be seen on the far-right side of the plot and is far from the regression line. When a line of best fit is put onto the same plot of all data points, but without camera 17 in the regression equation, the line does not have a dramatic change from the original plot. The changes of the line's equation can be seen in the regression equations displayed below. This suggests that even though camera 17 has the highest leverage of all points, it is not a very high leverage point overall. Further evaluation to if camera 17 is influential is required.



With all points

$$Score = 46.669 + 0.053 * Price$$



Excluding Camera 17

$$Score = 44.167 + 0.072 * Price$$

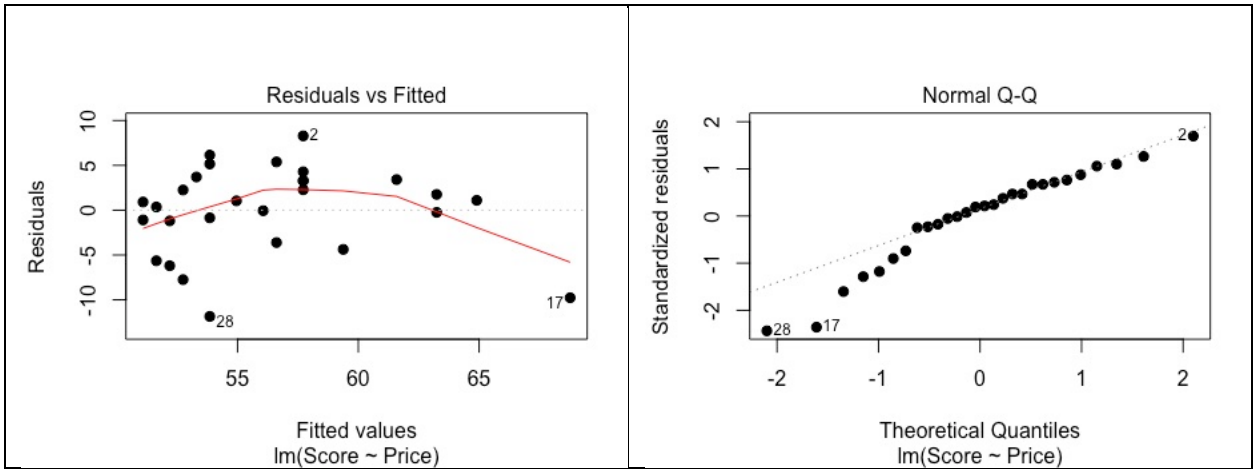
5. Identify and interpret any influential observations.

26	27	28	17
5.885740e-02	8.020849e-02	1.458119e-01	1.238879e+00

F-median: 0.4679148
F-25th Percentile: 0.1037076

The four most influential observations on this regression are cameras 26, 27, 28, and 17. Cameras 26 and 27 do not come close to either the F-median or the F-25th percentile, so therefore are not influential. Camera 28 has a Cook's Distance value of .1458, which is larger than 0.1037076, so it can be said that Camera 28 tends toward influential. For camera 17, which has a Cook's Distance of 1.239, it can be considered highly influential because it is considerably above the F-median of 0.468. This means that camera 17 highly influences the line of best fit that was made in this regression. As discussed in question 4, when looking at side by side comparisons of regression lines including and excluding camera 17, camera 17 does not appear to have high leverage. Although, it is still considered an outlier and highly influential due to its Cook's Distance being much larger than the F-median.

6. **Verify the assumptions for the regression of *Score* on *Price*. Insert the residuals versus fits plot and the normal Q-Q plot in the box provided below. I realize this is a touch cheaty, but I want you to reluctantly accept the assumptions as OK, despite weird Camera 17. Thus, no transformation to linearity is required! (Some extra credit is available for smart peeps who can perform some magic to get all the cameras to line up on the line nicely in the normal Q-Q plot. But really, I wouldn't worry about it.) Note: To make your plots pop a bit better, use the option *pch = 19*.**



7. Test whether a linear relationship between *Score* and *Price* exists, using $\alpha = 0.05$. Provide the null and alternative hypotheses, together with the models defined by each. Use MS Equation Editor for everything mathy. Then complete the test as shown on page 19 of Video 16.

The hypotheses are as follows, where β_1 represents the population:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

Null hypothesis: The null hypothesis $H_0: \beta_1 = 0$ asserts that no linear relationship exists between Score and Price of camera.

Alternative Hypothesis: The alternative hypothesis $H_a: \beta_1 \neq 0$ states that a linear relationship does exist between Score and Price of camera.

The following shows output relating to the slope, with Pr(>|t|) representing the p-value.

Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.66880	2.23844 2	0.849	< 2e-16 ***

Since the p-value is less than the significance value of 5%, we reject the null hypothesis in favor of the alternative, and it can be suggested that a linear relationship does exist between Score and Price of camera.

8. Provide and interpret a 99% prediction interval for a camera costing $250.

The 99% confidence interval for a camera costing $250 is as follows:

fit	lwr	upr
60.4811	46.18816	74.77404

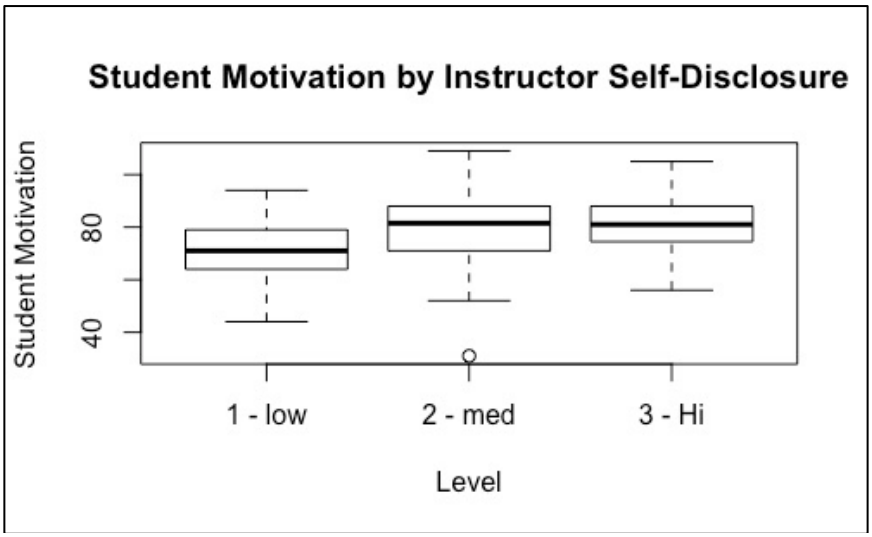
With 99% confidence, it can be said that the Score for a camera costing $250 lies between 46.19 and 74.44.

Part 2: Analysis of Variance

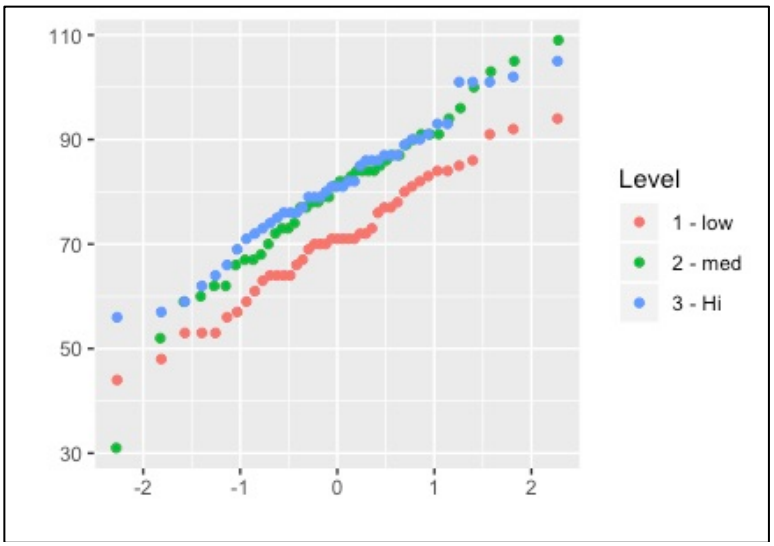
Import the *Facebook* data set. Student motivation (S.M) was measured for high, medium, and low levels (Level) of instructor self-disclosure on Facebook. (Source: *I'll See You on Facebook: The Effects of Computer-Mediated Teacher Self-Disclosure on Student Motivation, Affective Learning, and Classroom Climate*, by Joseph Mazer, Richard Murphy, and Cheri Simonds, Communication Education, Volume 56, 2007.)

9. Provide and comment on a boxplot of student motivation, by level of instructor self-disclosure.

The boxplot below shows that the medians of the medium and high-level groups are similar. The two boxplots are very similar, which may suggest similar means as well, even though the medium-level has a slightly larger range than the high-level. The low-level group appears to have a lower median than the other two groups. At this point it is undetermined if the low-level group is lower by a meaningful amount. Further investigation to determine if the population means differ significantly is required.



10. Verify the ANOVA assumptions. Provide the two outputs required, and comment on each. What is your conclusion?



Bartlett test of homogeneity of variances

data: S.M by Level
Bartlett's K-squared = 2.4272, df = 2, p-value = 0.2971

Since the graph shows all three levels in relatively straight lines, the normality assumption is validated for this ANOVA.
A Bartlett's test is used to check for equal variance. The null hypothesis states that the variances are equal. The p-value for this test is 0.2971, which is higher than 0.05, therefore there is not enough evidence to show that the variances are unequal. The equal variance assumption is validated.
These two tests lead to the conclusion that all ANOVA assumptions have been validated, and an appropriate analysis of variance can be performed.

11. Perform the appropriate analysis of variance, using $\alpha = 0.05$. Provide the hypotheses, explaining what the symbols mean. Then complete the ANOVA similar to page 18 in Video 18.

The hypotheses are as follows, where $\mu_1, \mu_2,$ and μ_3 represent the population means for student motivation:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \text{At least one } \mu \text{ is different, not all are equal}$$

Null hypothesis: The null hypothesis, $H_0: \mu_1 = \mu_2 = \mu_3$, asserts that all population means are equal.

Alternative Hypothesis: The alternative hypothesis states that at least one of the population means for student motivation is different than the others, or that one or more of them is not equal.

The following shows output relating to the ANOVA, with Pr(>F) representing the p-value.

summary(ANOVAfit)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Level	2	2712	1355.9	8.052	0.00051 ***
Residuals	127	21386	168.4		

Since the p-value of 0.00051 is less than the significance value of 5%, we reject the null hypothesis in favor of the alternative, and it can be suggested that all of the population means for student motivations are not equal, and at least one differs from the other. Tukey's method can be used to determine which are not equal.

12. Determine which pairs of population means differ significantly, providing clear conclusions for each pair.

The Tukey's method evaluated each possible pairwise comparison of the population means. These include:

- 2-Med vs 1-Low
- 3-Hi vs 1-Low
- 3-Hi vs 2-Med

The results from this evaluation can be seen below.

It can be observed that the p-values are all low, except for the for the third comparison which has a large p-value of 0.8086271.

Using these p-values the following conclusions about which pairs of population means differ significantly can be made:

- The population mean for student motivation of low levels of instructor self-disclosure on Facebook differs significantly from medium levels of instructor self-disclosure.
- The population mean for student motivation of low levels of instructor self-disclosure on Facebook differs significantly from high levels of instructor self-disclosure.
- The population mean for student motivation of medium levels of instructor self-disclosure on Facebook does not differ significantly from high levels of instructor self-disclosure.

```
TukeyHSD(ANOVAfit)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = S.M ~ Level, data = facebook)

$Level
                           diff        lwr        upr     p adj
2 - med-1 - low  8.735729  2.136620 15.334839 0.0059306
3 - Hi-1 - low  10.465116  3.828190 17.102043 0.0008073
3 - Hi-2 - med   1.729387 -4.869722  8.328496 0.8086271
```

==========

Well done!

Deliverables:

1. **Save your completed Word document as a pdf file, named *Doe_Jane_Project2* (if your name is Jane Doe, with last name first!). Because of virus issues, no Word documents will be accepted.**
2. **Your well-annotated R script, named *Doe_Jane_Project2_RScript.***

Do NOT zip these two files together. Rather, make two separate submissions using the Project Submissions Tool.