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

Name: Chad Zeller

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 questions, provide the R output required to answer the question, in addition to your responses.**
- Your responses must follow the notes. For example, I expect your interpretations to mirror those shown in the notes.

As you know, we are all about interpretation in this course. For each statistic you use in each sentence you write, ask yourself if your boss would understand what in tarnation you are talking about. If not, there is a good chance I will dock points for lack of explanatory clarity.

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.

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.**

   R Regression Score on Price Output:

   Coefficients:

(Intercept)	Price
46.66880	0.05525

   Regression Equation Using MS Equation Editor:

   $$\widehat{Score} = 46.6688 + .05525 * Price$$

   b. **Interpret the regression equation.**

   The *estimated* Consumer Reports score equals 46.6688 points plus 0.05525 times the price of the camera in dollars.

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

   For each *increase* of one dollar in price, the estimated Consumer Reports score *increases* by .05525 points.

2. Continuing with the regression of *Score* on *Price.*

a. Estimate *Score* for a camera costing $70.

R Output calculating a Camera Costing $70:

```
46.6688 + .05525 * 70
[1] 50.5363
```

Estimated Score for a Camera Costing $70:

46.6688 + .05525 * 70 = *50.5363*

The *estimated* Consumer Reports score for a camera costing $70 is 50.5363.

b. State and interpret r^2.

R Output from Summary Statistics for Multiple R-squared:

Multiple R-squared: 0.4668

r^2 Stated and Interpreted**:**

$r^2 = 0.4668$, meaning that 46.68% of the variability in *Score* is determined by *Price*.

c. State and interpret *s*. In your interpretation, make sure to follow the text in bold on page 24 of Video 14.

R Output from Summary Statistics for Residual Standard Error:

Residual standard error: 4.982 on 26 degrees of freedom

s Stated and Interpreted:

The residual standard error is $s = 4.982$, which represents the size of the typical difference between the *predicted value of y* and the *actual observed value of y*. 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". Instead, find out exactly how much higher or lower, using the *residuals* command. Your interpretation should be similar to page 5 in Video 15.**

```
rstandard(reg1)
residuals(reg1)
```

R Output for Standardized Residuals of Each Camera:

1	2	3	4	5	6	7	8
0.2414	1.6955	0.3757	0.8766	1.1010	0.6718	0.4671	1.2641
9	10	11	12	13	14	15	16
1.0585	0.4663	0.0749	-0.2480	-1.1773	0.7160	-0.0521	0.6718
17	18	19	20	21	22	23	24
-2.3575	0.7631	-0.0125	0.2137	-0.9020	-0.7386	-0.1750	0.1912
25	26	27	28				
-0.2284	-1.2866	-1.6025	-2.4364				

R Output for Residuals of Each Camera:

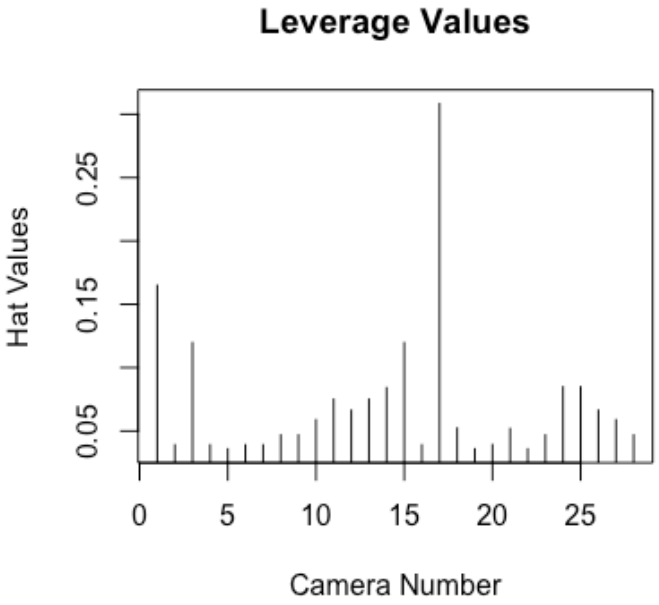
1	2	3	4	5	6	7
1.0990	8.2814	1.7564	4.2814	5.3863	3.2814	2.2814
8	9	10	11	12	13	14
6.1488	5.1488	2.2538.	0.3588	-1.1937	-5.6412	3.4139
15	16	17	18	19	20	21
-0.2436	3.2814	-9.7685	3.7013	-0.0612	1.0438	-4.3761
22	23	24	25	26	27	28
-3.6137	-0.8512	0.9113	-1.0887	-6.1937	-7.7462	-11.8512

Outliers Defined and Interpreted:

Camera numbers 17 and 28 are outliers in the Cameras data set. Camera #17 is an outlier because its *actual Score* of 57 is significantly (9.7685 points) *lower* than the *predicted Score* of 68.7685. Camera #28 is an outlier because its *actual* score of 42 is much (11.8512 points) *lower* than its *predicted Score* of 53.8512.

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. That is, is it high leverage? And what does it mean to be high leverage in terms of this particular problem?**

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



The graph displaying hat values clearly shows that camera #17 has by far and away the highest leverage in the data set. This means that camera #17 is the data point *furthest* from the bulk of the data *horizontally*. Camera 17 has a hatvalue of 0.3084, which is *nearly twice the leverage of the next greatest hatvalue in the data set*. Camera 17 has a price of $400, which is extremely large relative to its modest 59 Score.

5. Identify and interpret any influential observations.

R Output of the Median of the F-Distribution:

[1] 0.4679

R Output of the 25th Percentile of the F-Distribution:

[1] 0.1037

R Output of Cook's Distance for Each Camera (Sorted):

19	15	11	23	20
2.907732e-06	1.844601e-04	2.275222e-04	7.521922e-04	9.315879e-04
24	12	25	7	1
1.694298e-03	2.186270e-03	2.418517e-03	4.426320e-03	5.752526e-03
10	6	16	3	22
6.789957e-03	9.157208e-03	9.157208e-03	9.593034e-03	1.013770e-02
4	18	21	5	14
1.558902e-02	1.605785e-02	2.224476e-02	2.252339e-02	2.353657e-02
9	8	13	2	26
2.752206e-02	3.925089e-02	5.625197e-02	5.832554e-02	5.885740e-02
27	28	17		
8.020849e-02	1.458119e-01	1.238879e+00		

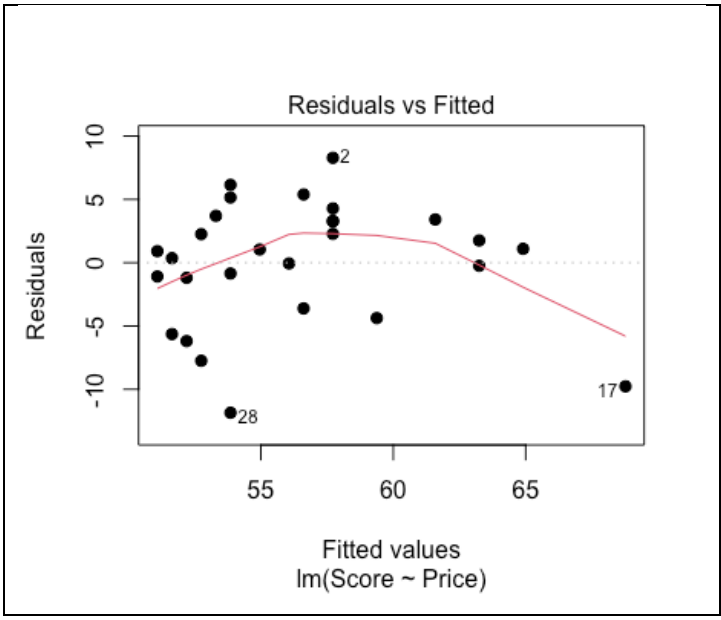
Identification and Interpretation of the Influential Observations:

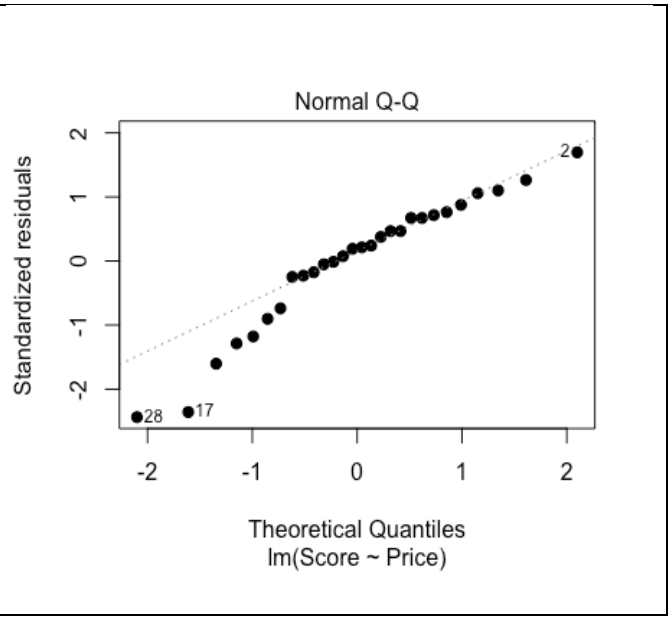
To determine the influence of an observation, we calculate Cook's Distance, which combines *outlier* and *leverage*. To qualify as *definitely influential*, an observation must have a Cook's Distance value *above the median*. This means the F-distribution value must be above the *50th* percentile, which is roughly .4679, calculated by finding the *median* of the F-distribution. In calculating Cook's Distance, Camera #17 is the only point in the data set above the .4679 threshold, and thus the only *definitely influential* data point.

To qualify as "*tending to influential*," an observation must have a value between the *25th* and *50th* percentile. The 25th percentile is roughly .1037, so any observations between .1037 and the median of .4679 is "tending to influential." The only observation that meets the criteria of "*tending to influential*" is Camera #28.

In sum, Camera #17 is *definitely* influential, Camera #28 is "tending to influential," and the remainder of the data sets (all those below the 25th percentile of .1037) are *definitely not* influential.

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! Note: To make your plots pop a bit better, use the option *pch = 19*.**





Based on the residuals vs. fitted plot, the red trend line is *somewhat* smooth, but with some noise (curvature), which appears to be primarily caused by the camera #17 outlier. However, the bulk of the data is consistent enough that we can tentatively accept the zero mean, constant variance, and independence assumptions as being OK.

Based upon the normal Q-Q plot, we can see that the bulk of the residuals lie on or very near the straight line of the plot. There are a few stray data points, notably cameras #17 and 28, which diverge quite a bit, but the bulk of data points are linear. Therefore, we can determine that the normality assumption is validated.

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

Null and Alternative Hypotheses:

$H0: \beta 1 = 0$

The null hypothesis $H0: \beta 1 = 0$ asserts that *no linear relationship exists* between score and price.

$Ha: \beta 1 \neq 0$

The alternative hypothesis $Ha: \beta 1 \neq 0$ asserts that a *linear relationship does exist* between score and price.

R Summary Statistics Output Displaying the P-Value:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	46.66880	2.23844	20.849	< 2e-16 ***
Price	0.05525	0.01158	4.771	6.16e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Analysis of Linear Relationship of Score and Price:

The p-value for the t-test equals less than zero to fifteen decimal places (*2e-16*). Since this p-value is so small (far below the significance level of 0.05), we can reject the null hypothesis and conclude that $\beta 1 = 0$ and that a linear relationship *does* exist between score and price.

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

R Output for the 99% Prediction Interval for a New Camera with $250 Price:

	fit	lwr	upr
1	60.4811	46.1882	74.774

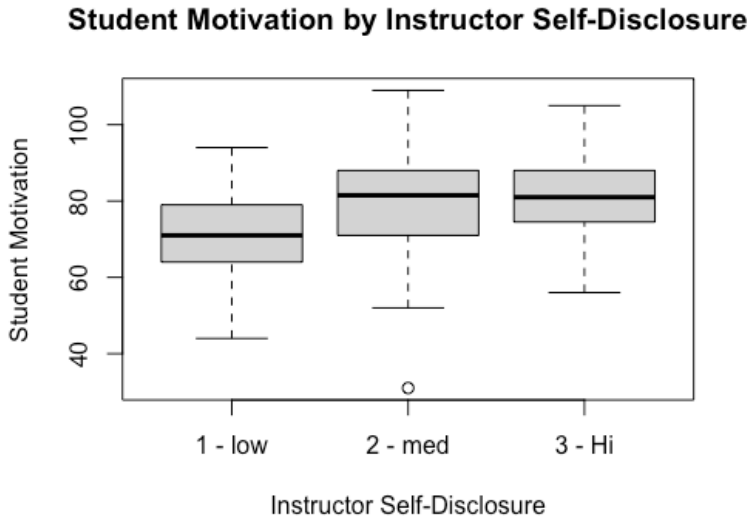
Interpretation of a 99% Prediction Interval for a New Camera with $250 Price:

We are 99% confident that the *score* for a new camera with a price of $250 lies *between* a score of 46.1882 and 74.774.

Part 2: Analysis of Variance

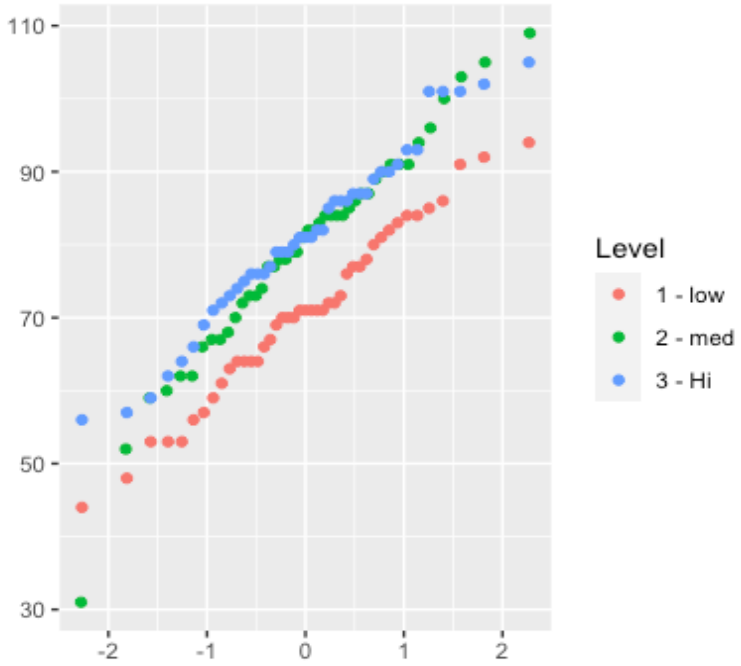
Import the *Facebook* data set. Student motivation (S.M) was measured for high, medium, and low levels (Level) of instructor self-disclosure on Facebook. (Source: *I'll See You on Facebook: The Effects of Computer-Mediated Teacher Self-Disclosure on Student Motivation, Affective Learning, and Classroom 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 *medium* and *high* groups appear to have the same median, but the *low* group has a median quite a bit *lower* than the other groups. The *medium* group has a slightly wider interquartile range than the high group. Precisely how significant the low group diverges from the medium and high groups will require a deeper analysis.

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



In the first test, a quick plot (qplot) was run to check the normality assumption. Given the plots for all three levels (low, medium, high) are approximately linear on the graph, we can conclude that the normality assumption is validated.

R Output for Bartlett's Test of Homogeneity of Variances:

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

Analysis of Bartlett's Test and ANOVA Assumptions:

In the second test, a Bartlett's Test was run to check the equal variances assumption. The null hypothesis is that the variances are *equal* and a *small p-value* for the test rejects the null hypothesis and invalidated the equal variances assumption. Our p-value is *0.2971*, which is quite high. We can therefore conclude there is no evidence that the variances are not equal and that the variance assumption is validated.

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

Null and Alternative Hypotheses:

H0: $\mu 1 = \mu 2 = \mu 3$

The null hypothesis H0: $\mu 1 = \mu 2 = \mu 3$ states that the population means of the low, medium, and high level of instructor self-disclosure are equal.

Ha: one or more of $\mu 1, \mu 2, \mu 3$ not equal

The alternative hypothesis Ha states that at least one of the population means of the low, medium, and high level of instructor self-disclosure is not equal to the other levels.

R Summary Statistics Output:

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

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ANOVA Test Analysis:

Since the p-value is very small (0.00051), far below the significance level of $\alpha = 0.05$, we *reject* the H0.

There appears to be good evidence that not all the population mean *levels* are equal.

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

R Output for Tukey's "Honest Significant Difference" Method:

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

Conclusions from Using Tukey's HSD Method on Which Pairs Differ:

To test for pairwise comparison of means, we perform Tukey's "Honest Significant Difference" method to compare means between the pair of each level of instructor self-disclosure.

The output from for first two pairs, low vs. medium, and low vs. high is *close to zero and well below the level of significance*. The third pair, medium vs. high, has a *very large p-value*.

Therefore, we have the following conclusions:

- The population mean low level student motivation differs significantly from medium level motivation students.
- The population mean low level student motivation differs significantly from high level motivation students.
- The population mean medium level student motivation does not differ significantly from those instructors in the high disclosure category.

==========

Well done!

<u>Deliverables:</u>

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.

This Project is subject to change at any time.