Final Exam

Alan Perez

November 27, 2021

Note: Because this was a final exam, I cannot post the exact questions asked or my answers to those questions online. Instead, This document will contain a brief overview of my findings. This paper was produced using LaTeX.

1 Background Information

This specific paper will analyze the performance of marketing-related Facebook posts by a cosmetics company. Specifically, for each social media post the company made on their official Facebook account, we track the total number of interactions (likes, comments, or shares) that Facebook users had with the post. Here are the variables of interest:

interactions: The total numberof times the Facebook post was liked, commented on, or shared on Facebook.
impressions: The total number of times the post was viewed on Facebook
pagelikes: The number of Facebook users that like/follow the cosmetics company's Facebook page/account
paid: a dummy/indicator variable for whether the post was a paid advertisement on Facebook (=1 for a paid advertising post, =0 for a regular non-paid post)
type: the type of social media post: "Link", "Photo", "Status" (a status update, which is a regular text post), or "video

2 Analysis

Let's begin by checking out our Interactions variable is related to our impressions variable using a simple linear regression.

```
> summary(slr)

Call:
lm(formula = interactions ~ impressions)

Residuals:
Min      1Q  Median      3Q     Max
-767.42  -94.59  -46.53   39.55 1638.73

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.280e+02  1.213e+01   10.55   <2e-16 ***
impressions 3.035e-03  2.765e-04   10.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 228.1 on 487 degrees of freedom
Multiple R-squared:  0.1983,Adjusted R-squared:  0.1966
F-statistic: 120.4 on 1 and 487 DF,  p-value: < 2.2e-16

> # iii
> summary(impressions)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
570    5744    9035   23058   21685  277100
```

From the output above, using a Δx of $\Delta x = 30{,}000$ impressions, we see that for every additional 30,000 impressions on a post, the number of interactions with the post is higher by 91.04 on average. From our intercept we understand that the average number of interactions a post with 0 impressions receives is 128.09. However, this interpretation does not make sense and is not meaningful because it is not possible for a Facebook post to be interacted with without having been viewed. This interpretation is also an extrapolation since the smallest number of impressions received on a post in our data set was 570. By once again looking at our summary of the SLR, we see our p-value is $p = 2.2 \times 10^{-16}$, making the relationship between interactions and impressions statistically significant.

Beyond providing some analysis of the possible relationships, we also want to check that our assumptions of our SLR model are not violated. Below is a scatter plot with our previously determined SLR line of best fit added, Figure 1. The linearity assumption does appear violated here because for posts with more than 150,000 impressions, points are systematically below the line of best fit. To aid our visualization, we add a LOWESS curve (in Red) to a residuals versus impressions plot and find that we have non-linearity, Figure 2. To fix this violation, I suggest we expand our impression variable into a quadratic. After doing that, we see in our residuals versus impressions plot, Figure 3, that our LOWESS curve is much more flat along the horizontal line where E = 0. This is evidence of improvement. We can also make the argument that constant variance assumption also appears violated because the spread of the points about the line of best fit seems to grow as the number of impressions increase.

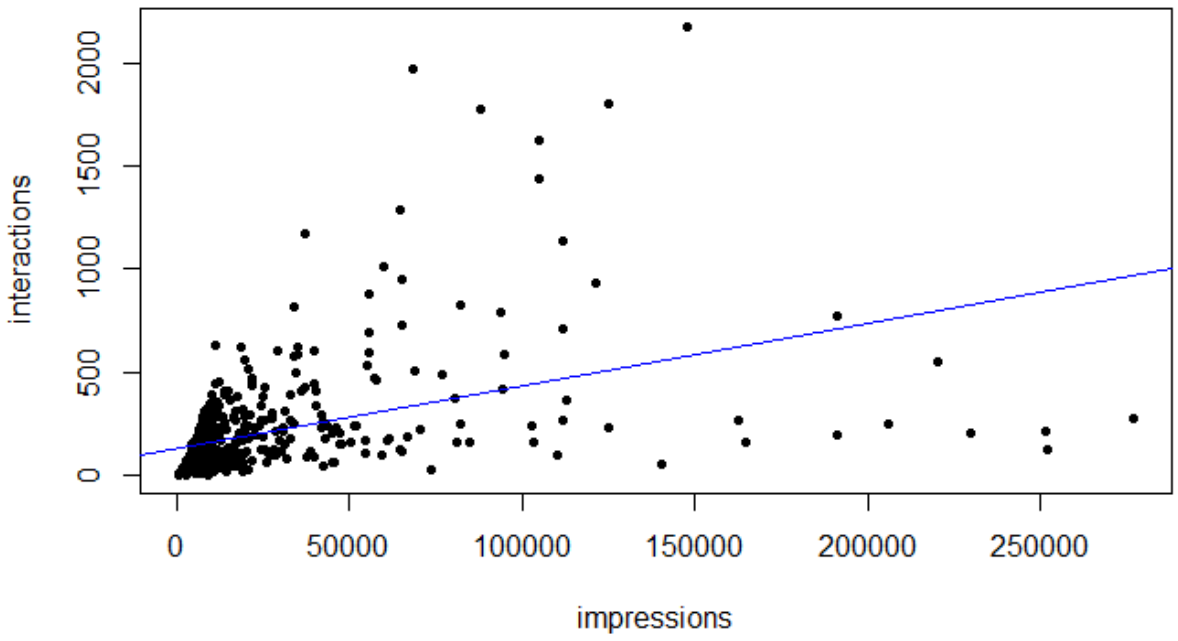


Figure 1: Scatter plot of interactions versus impressions with line of best fit

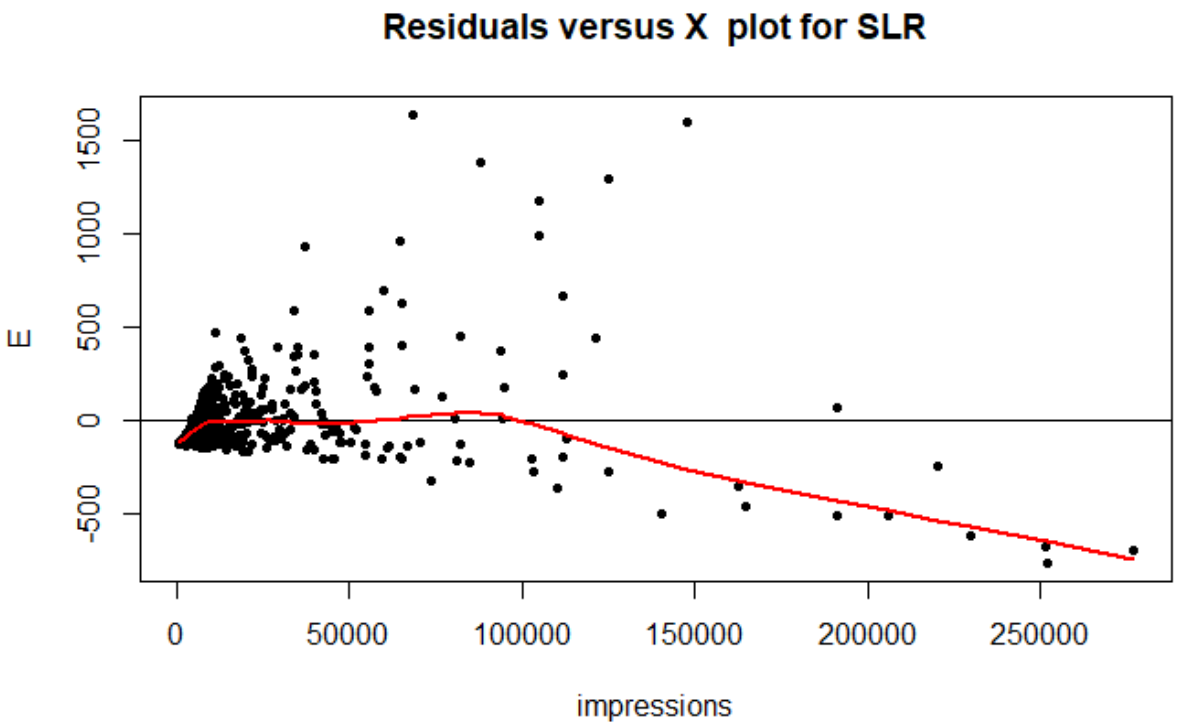


Figure 2: Plot of residuals versus impressions plot for SLR with LOWESS curve

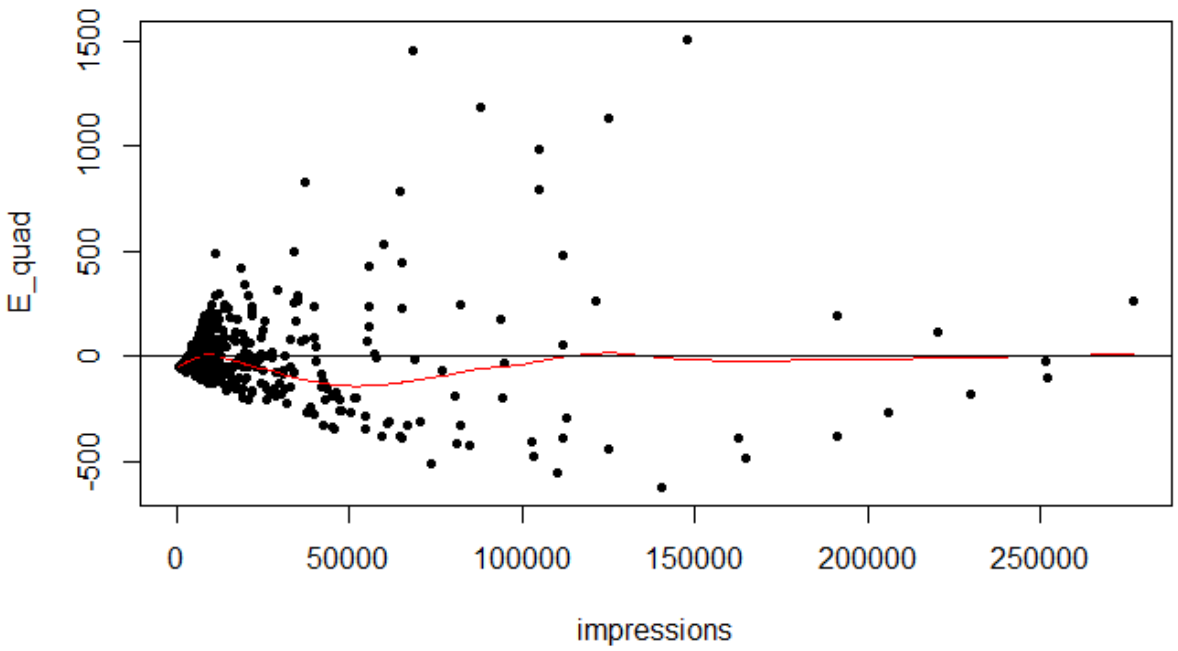


Figure 3: Plot of residuals versus impressions plot for SLR with LOWESS curve

By drawing a histogram of the residuals in Figure 4, we notice a relatively long right tail which indicates a violation of the normality of errors assumption. Specifically, the errors are skewed right. Lastly, I want to discuss the independence assumption. This is slightly more difficult to prove since there are no plots or graphs (or none that I know of at least) that can directly show us a violation of independence. However, I argue that the independence assumption is violated. I suspect that the observations in this analysis are not independent of each other and instead are closely related. It is fair to assume that if someone interacts with a single post from the company, it is likely they will click on their page and interact with several others posts from the company as well.

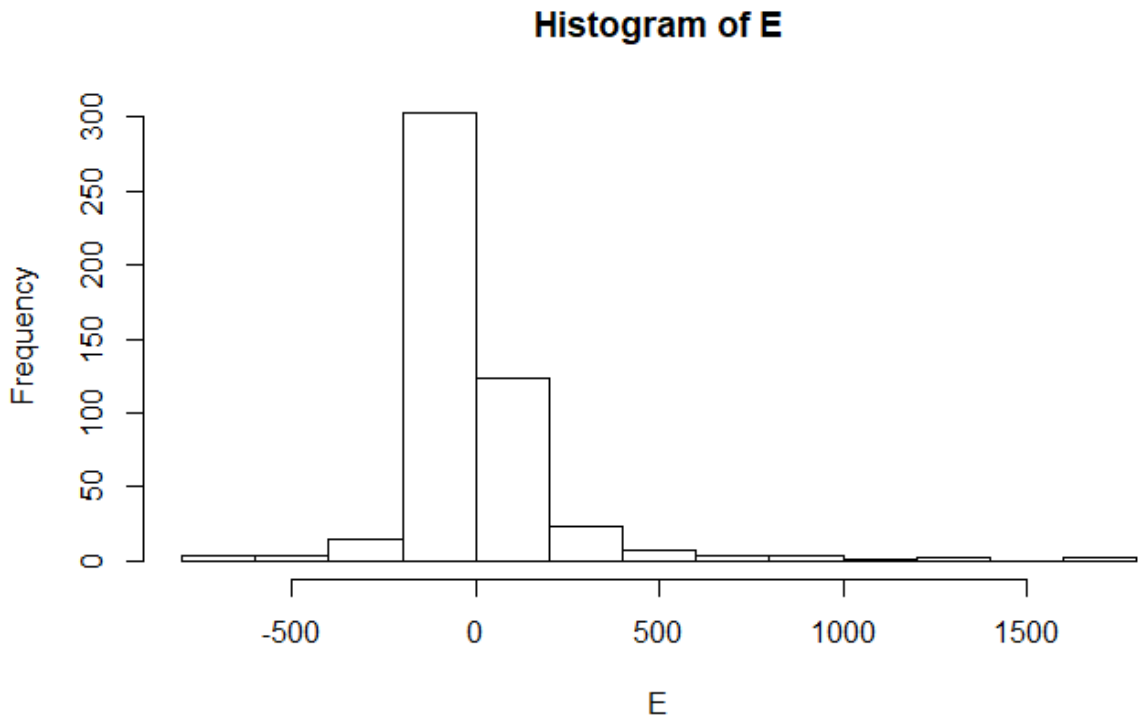
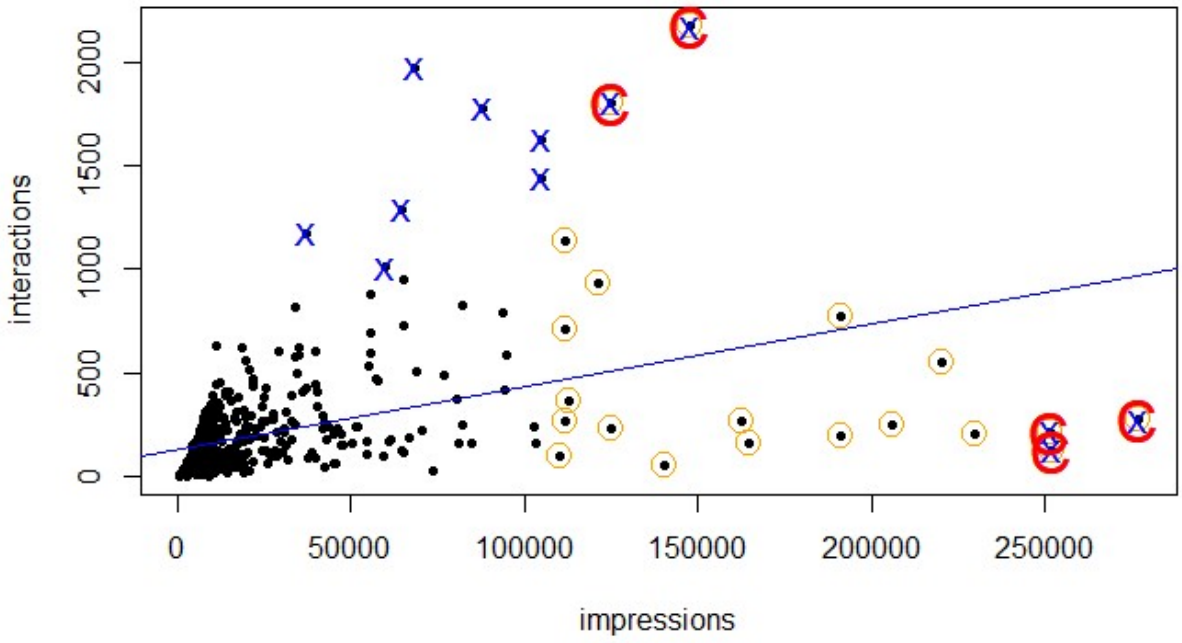


Figure 4: Histogram of the residuals

```
> plot(impressions, interactions, pch=20)
> abline(slr, col='Blue')
> # ii
> lines(lowess(impressions, interactions), col="RED") # lowess creates x and y atributes
> lines(lowess(impressions, interactions), col="RED") # lowess creates x and y atributes
> E <- residuals(slr)
> plot(impressions, E, pch=20, main="Residuals versus X  plot for SLR")
> abline(h=0)  # a flat line
> # add a LOWESS curve for visuals reference:
> lines(lowess(impressions, E), col="RED", lw=2)
> E <- residuals(slr)
> plot(impressions, E, pch=20, main="Residuals versus impressions plot for SLR")
> abline(h=0)  # a flat line
> # add a LOWESS curve for visuals reference:
> lines(lowess(impressions, E), col="RED", lw=2)
```

We will now explore outlying and influential observations for the SLR model. We first calculate the average leverage to be $\hat{h} = 0.0041$. The points that have more than *three times* the average leverage are represented on the scatter plot by orange circles. We then calculated the studentized residuals E^* and placed blue X's on all points with studentized residuals larger than 3 in absolute value. Lastly, we calculated the Cook's Distance values and drew red letter C's on all points with Cook's Distance values greater than 0.28. In the end, there were 5 points flagged as being influential by Cook's Distance. Each of which were high leverage points and high residual points indicated by both the orange circle and the blue X.




```
> h <- hatvalues(slr)
> hbar <- mean(h)
> hbar
[1] 0.00408998
> high_lev <- which(h > 3*hbar)
> length(high_lev)
[1] 20
> points(impressions[high_lev], interactions[high_lev], col="orange", cex=2)
> # iii.
> E_star <- rstudent(slr)
> y_outlier <- which(abs(E_star) > 3)
> length(y_outlier)
[1] 12
> points(impressions[y_outlier], interactions[y_outlier], col="BLUE", pch="X")
> # iv.
> D <- cooks.distance(slr)
> high_cook <- which(D > 0.28)
> length(high_cook)
[1] 5
> points(impressions[high_cook], interactions[high_cook], col="RED", pch="C", cex = 2)
```

Our next goal was to learn under what conditions people tend to interact with our posts a lot. Here, we will not use the number of impressions as a predictor since this variable is likely causally affected by the same variables that drive interactions. So we will use three predictor variables **pagelikes, paid,** and **type** in order to predict **interactions**. This is done by fitting an MLR with **paid** as a dummy/indicator variable and **type** as a categorical variable encoded as a character string variable in R. Here, the **type** variable will be included in the model using **"Link"** as the reference group.

A summary of our MLR can be seen below. We can make the following observations:

- On posts of fixed advertisement status (whether it was an advertisement post or a regular post) and post type, every additional 15,000 Facebook users that like/follow the cosmetics company's account is associated with about 15 (15.12) more interactions on average.
- Among posts of fixed page likes/follows and post type, advertised posts receive approximately 49 (49.53) more interactions than regular non-paid posts, on average.
- Among posts of fixed page likes/follows and advertisement status, photos have on average 104 (104.2523) more interactions than link posts.
- Assuming the number of users that like/follow the cosmetics company's Facebook page is fixed and the advertisement status is fixed, it appears that video posts get the most interactions on average. Under the same conditions, it seems that link posts get the least amount of interaction.

We are also interested on whether *any* of the three X's we used actually relate to the count of interactions a post receives. We can address this problem by performing an overall F-test. We start by writing out our null and alternative hypotheses both in equations and in words.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$; None of the variables pagelikes, advertisement status (paid), and post type are related to post interactions
$H_A : \beta_j \neq 0$ for j = 1, 2, 3, 4, or 5; At least one of the variables pagelikes, advertisement status (paid), or post type are related to post interactions

From the summary of our MLR, our F test statistic is $F_{ts} = 2.22$ and our p-value is $p = 0.0512$. Since our p-value is a little larger than our significance level of 0.05, we do not have enough evidence to reject H_0. That is, there is not enough evidence to conclude at least one of the variables pagelikes, advertisement status (paid), or post type are related to post interactions. Had our significance level been 0.06, then we would have had enough evidence to reject the null hypothesis and conclude that at least one of our variables are related to the count of interactions a post receives. Since we were only a bit off from having a p-value less than $\alpha = 0.05$ I argue that we have enough evidence to say at least one of the three X's are related to interactions

```
> mlr <- lm(interactions ~ pagelikes + paid + factor(type))
> summary(mlr)

Call:
lm(formula = interactions ~ pagelikes + paid + factor(type))

Residuals:
Min      1Q  Median      3Q     Max
-250.15 -125.47  -65.70   29.56 1952.46

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)        -41.806456 100.923998  -0.414   0.6789
pagelikes            0.001008   0.000731   1.379   0.1684
paid                49.530764  25.569973   1.937   0.0533 .
factor(type)Photo  104.252284  55.530266   1.877   0.0611 .
factor(type)Status 114.079492  66.888092   1.706   0.0887 .
factor(type)Video  173.207973 110.871460   1.562   0.1189
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 253 on 483 degrees of freedom
Multiple R-squared:  0.02247,Adjusted R-squared:  0.01235
F-statistic:  2.22 on 5 and 483 DF,  p-value: 0.0512
```