“Successful Instagram Account”

Stat424: Design Analysis of Experiment Project

Kyungjin Cho | Kcho36@wisc.edu

Table of Contents

List of Tables

List of Plots

1. Project Background

In today's society, according to the statistics, 71% of U.S adults in the age of 18 – 29 uses Instagram (Statista, 2021). Instagram has become the main social channel that has the highest average engagement rate in the world. In this project, I am going to find out the factors that makes Instagram account most successful.

I have implemented three factors in this project which are listed below:

- Variable A: The user has Highlights in his/her profile
- Variable B: The user uploads 3 posts in a week
- Variable C: The user identifies top hashtags

According to the internet, there are several ways to make the Instagram account successful: creating a perfect biography, producing interesting content, posting regularly, utilizing hashtags, taking advantage of highlights and so on. However, in this project I decided to use three of the following bullet-pointed features above.

As mentioned in the proposal, I was going to look at the famous celebrities' Instagram account; however, since they are already considered as successful influencers, I decided to analyze my 24 friends' Instagram account.

Since I am going to implement 3 factors at 2 factorial level, it would be $2^k = 2^3 = 8$ full factorial design. Also, I am replicating the test three times in total, in order to make this test more credible. Here is the table of factors and levels assigned in this experiment:

Factors	Variable	Low (-)	High (+)
Highlights	A	Not Present	Present
Number of Posts in a week	B	1	3
Hashtags	C	Not present	Present

Table 1: Table of Factors and Levels

2. Data

2.1 Data Preparation

In order to reduce the confounding factors, I have chosen 100 accounts that I follow. Those 100 accounts should be categorized as a person's account (not for the business), do not have blue checkmarks (which indicates the influencer's account), and are identified as South Koreans. The reason why I only selected Korean accounts is to reduce the confounding factors in my statistical design. The population difference in each country influences the followers. An individual living in bigger country tends to have more followers than the individual who is living in the small country.

After I selected 100 accounts, I have wrote their names in the excel spreadsheet and used "=SORTBY(SEQUENCE(100),RANDARRAY(100))" in order to create a list of random numbers with no duplicates. Then, from 1, I sort them into those 8 categories that I have assigned until I fill all the design matrix below.

Number	Factors			Replicates			$\bar{y}$
	A(Highlights)	B(Posts)	C(Hashtags)	1	2	3	
1	+	3	+	719	800	946	821.67
2	-	3	+	463	503	484	483.33
3	+	1	+	386	464	371	407.00
4	-	1	+	452	377	344	391.00
5	+	3	-	602	691	522	605.00
6	-	3	-	570	548	632	583.33
7	+	1	-	423	427	485	445.00
8	-	1	-	352	331	359	347.33

Table 2: Design Matrix

2.2 Computing Software

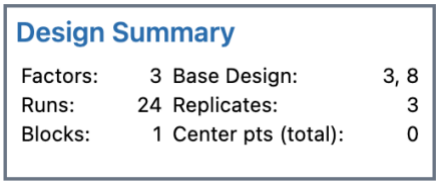
The statistical computing software I have used is Minitab. I chose Minitab because it was easier to implement graphs and ANOVA table than any other statistical software. It also provided the informative plots and tables.

2.3 Designing the Data

Since it is more credible to have Minitab also calculate the replicates than to get $\bar{y}$ alone, I have chosen 3 as the number of replicates for corner points.

Also, from the Options menu, I have checked "Randomize runs" so that I do not have to implement random number generator on my own.

Here is the design summary of the experiment I am going to perform.



As Minitab has assigned the random order for me, I have written down those numbers as "Run Order" in my Planning Matrix table below. Again, I have considered to type all the replicates so that my results will be more precise. Therefore, in total, I have $8 \times 3 = 24$ variables.

Number	Factors			y	Run Order
	A(Highlights)	B(Posts)	C(Hashtags)		
1	-	1	-	359	1
2	+	1	-	485	22
3	-	3	-	632	23
4	+	3	-	522	18
5	-	1	+	344	21
6	+	1	+	371	8
7	-	3	+	484	20
8	+	3	+	946	4
9	-	1	-	331	5
10	+	1	-	427	12
11	-	3	-	548	13
12	+	3	-	691	14
13	-	1	+	377	24
14	+	1	+	464	17
15	-	3	+	503	11
16	+	3	+	800	9
17	-	1	-	352	7
18	+	1	-	423	19
19	-	3	-	570	3
20	+	3	-	602	10
21	-	1	+	452	6
22	+	1	+	386	16
23	-	3	+	463	15
24	+	3	+	719	2

Table 3: Planning Matrix

3. Data Analysis

<Estimated Effects and Coded Coefficients>

Term	Effect	Coef	SE Coef	T-Value	P-Value	VIF
Constant		510.5	12.4	41.08	0.000	
Highlights	118.4	59.2	12.4	4.76	0.000	1.00
Posts	225.7	112.9	12.4	9.08	0.000	1.00
Hashtags	30.6	15.3	12.4	1.23	0.236	1.00
Highlights*Posts	61.6	30.8	12.4	2.48	0.025	1.00
Highlights*Hashtags	58.8	29.4	12.4	2.36	0.031	1.00
Posts*Hashtags	27.7	13.9	12.4	1.12	0.281	1.00
Highlights*Posts*Hashtags	99.6	49.8	12.4	4.01	0.001	1.00

*Significant factors at 95% confidence interval are colored red

Table 4: Estimated Effects and Coefficients

According to the Coded Coefficients table in Minitab analysis, I was able to find that at 95% confidence interval, Highlights, Posts, Highlights*Posts, Highlights*Hashtags, Highlights*Posts*Hashtags were the significant factors.

At 99% confidence interval, Highlights, Posts, Highlights*Posts*Hashtags were the significant factors.

<Analysis of Variance>

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Model	7	503111	71873	19.39	0.000
Linear	3	395525	131842	35.57	0.000
Highlights	1	84135	84135	22.70	0.000
Posts	1	305778	305778	82.50	0.000
Hashtags	1	5612	5612	1.51	0.236
2-Way Interactions	3	48085	16028	4.32	0.021
Highlights*Posts	1	22755	22755	6.14	0.025
Highlights*Hashtags	1	20709	20709	5.59	0.031
Posts*Hashtags	1	4620	4620	1.25	0.281
3-Way Interactions	1	59501	59501	16.05	0.001
Highlights*Posts*Hashtags	1	59501	59501	16.05	0.001
Error	16	59303	3706		
Total	23	562414			

Table 5: Analysis of Variance

ANOVA table is shown above. The 2-way interactions and 3-way interactions were considered as significant factor at 95% confidence interval.

<Model Summary>

S	R-sq	R-sq(adj)	R-sq(pred)
60.8803	89.46%	84.84%	76.28%

Table 6: Model Summary

The model summary indicates that the adjusted R^2 equals to 84.84%, which means that 84.84% of the data fit the regression model. I believe that 84.84% is approximately a good fit for the model.

4. Regression Modeling

The regression equation of the data would be $y = \mu + \left(\frac{A}{2}\right)x_1 + \left(\frac{B}{2}\right)x_2 + \left(\frac{C}{2}\right)x_3 + \left(\frac{AB}{2}\right)x_1x_2 + \left(\frac{AC}{2}\right)x_1x_3 + \left(\frac{BC}{2}\right)x_2x_3 + \left(\frac{ABC}{2}\right)x_1x_2x_3 + \varepsilon$ where A, B, C equals to the following:

Factor	Name
A	Highlights
B	Posts
C	Hashtags

Also, $x_1x_2x_3$ equals to the factor variables for A, B, C respectively.
Using the Minitab, the regression equation for 2^3 factorial design is

$$y = 284.7 - 2.4(A) + 112.9(B) - 12.5(C) + 30.8(A)(B) - 70.2(A)(C) + 13.9(B)(C) + 49.8(A)(B)(C)$$

Which is in uncoded units,

$$Followers = 284.7 - 2.4(Highlights) + 112.9(Posts) - 12.5(Hashtags) + 30.8(Highlights)(Posts) - 70.2(Highlights)(Hashtags) + 13.9(Posts)(Hashtags) + 49.8(Highlights)(Posts)(Hashtags)$$

For example, a person who utilizes highlights, posts 3 times a week, and identifies top hashtags will have approximately $y = 284.7 - 2.4(1) + 112.9(3) - 12.5(1) + 30.8(1)(3) - 70.2(1)(1) + 13.9(3)(1) + 49.8(1)(3)(1) = 821.8 \approx 822$ followers on his/her Instagram.

5. Plots

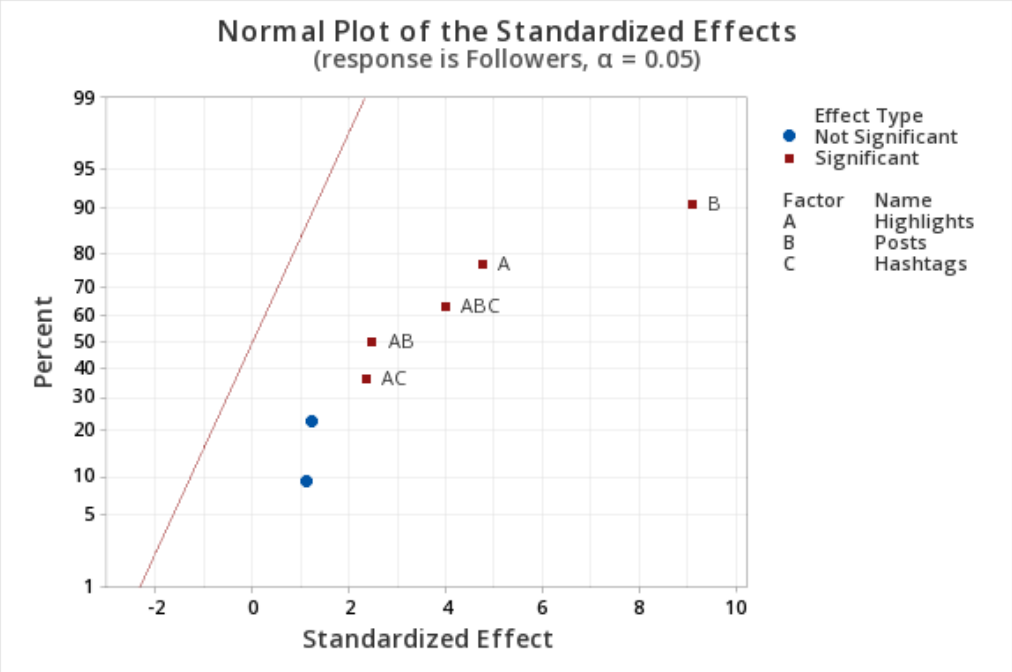


Figure 1: Normal Plot of Standarized Effects

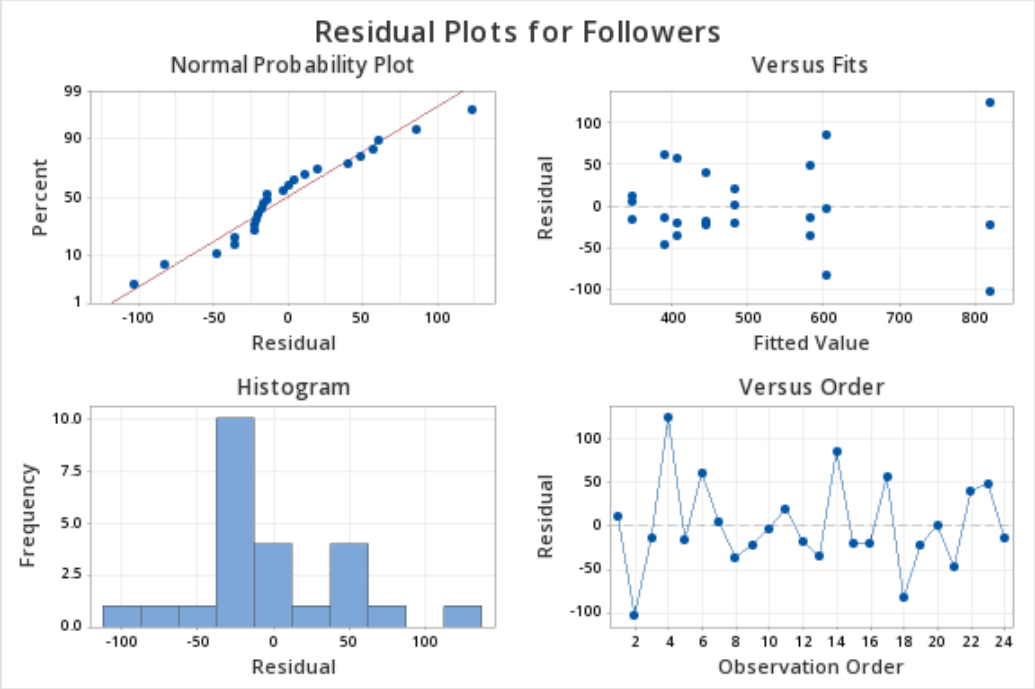


Figure 2: Residual Plot

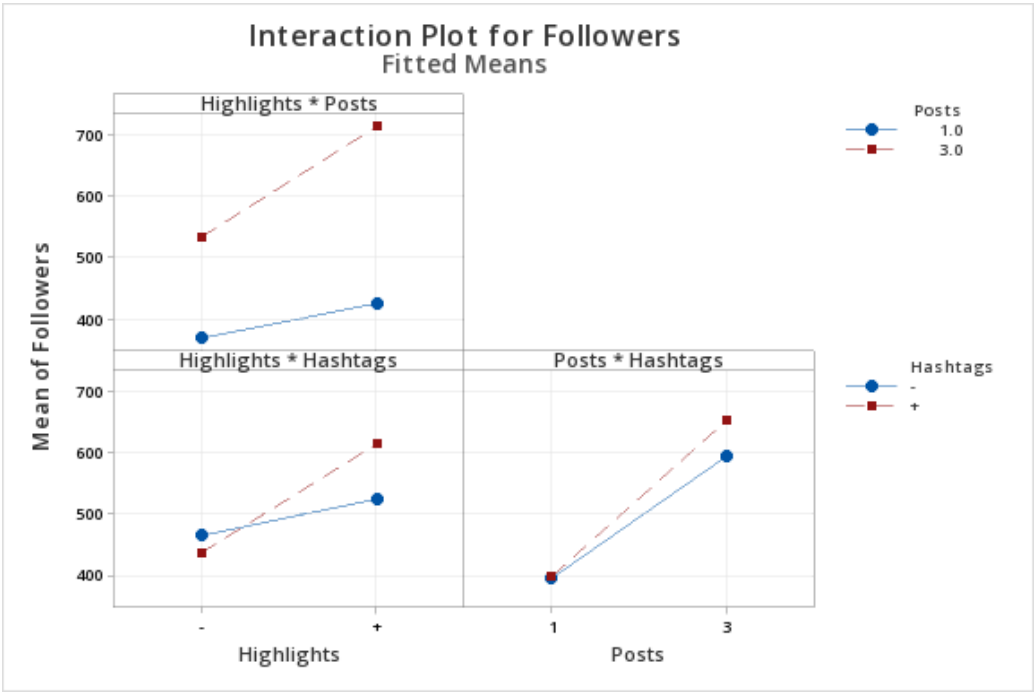


Figure 3: Interaction Plot

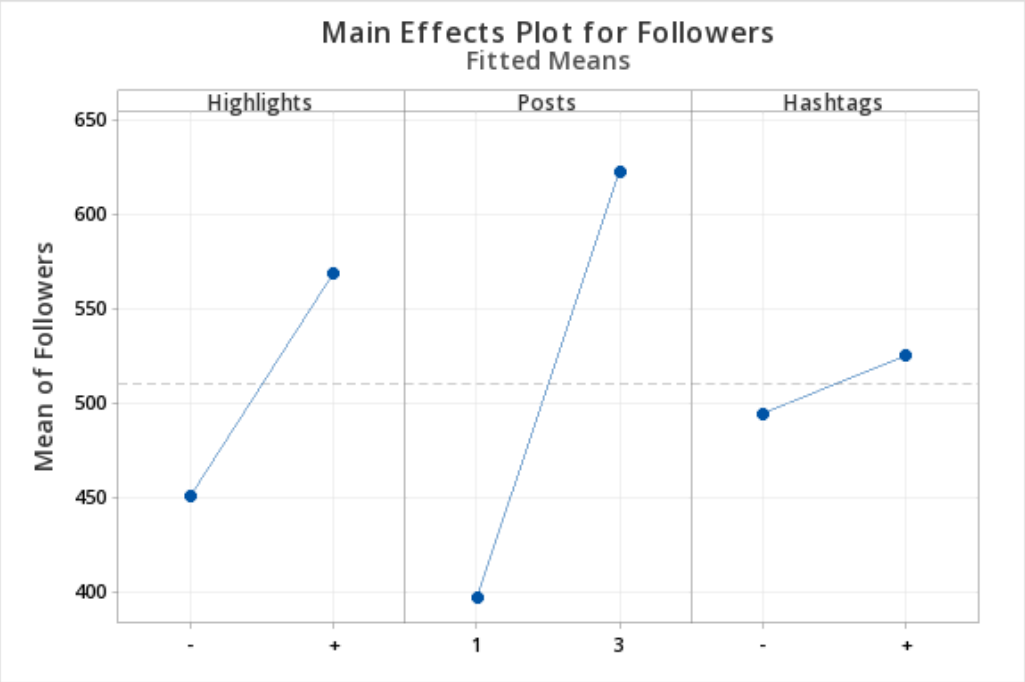


Figure 4: Main Effects Plot

6. Scientific Inference

According to the plots shown above, here is the important features ranked from one to five:

I. Posts
II. Highlights
III. Posts*Highlights*Hashtags
IV. Highlights*Posts
V. Highlights*Hashtags

Uploading posts regularly (here, I defined regularly as three times a week) was the most important feature of making one's Instagram profile successful. By successful, it means that the user has lot of followers.

By looking through the residual plot shown above, I believe that model was well explained by the linear regression. Although it is not perfectly linear, I would assume that the variables are randomly dispersed around the horizontal line. From the residual vs. fits graph, although the data are spreaded randomly, we notice the heteroscedasticity in this graph because as the fitted values become larger, more spread the data is.

According to the interaction plots, I could interpret that Highlights*Hashtags and Highlights*Posts have interaction effects, as the p-value of Highlights*Hashtags and Highlights*Posts confirms. The interaction plots of both indicates non-parallel lines. Although Hashtags*Posts also have non-parallel lines, it is not as non-parallel as the other twos.

According to the main effects plots, it is clear that all three of the graphs are not horizontal. However, since Factor C, Hashtags, is close to the horizontal and its p-value from ANOVA indicates that it is not significant, we could ignore Hashtags from the inference. Posts, just like it

was proven in the first plot, is the most significant factor of making one's Instagram successful. The slope of the line is almost 1, which signifies that posting regularly is in fact the most important guide to having a lot of followers. The line of Highlights also underscores the importance of utilizing highlights features in Instagram, but it is not as significant as updating your feed regularly.

Although 89.46% of variation is explained by the independent variables in this model, there are some shortages to my findings. First, this data does not explain on behalf of all the models. In this experiment, I only had three replicates; however, if more data were collected from the bigger population, this model could be more credible. Second, Although I have removed some confounding variables, there are still a lot of factors that could affect my findings. For example, the date that the user created Instagram account could also affect the number of followers since older user will have more time to gather one's followers than someone who is new to this social media.

So far, with the data that I have collected, I conclude that to make your Instagram account successful, it is important to post regularly and utilize the highlights function.

7. Sources:

- https://www.statista.com/statistics/246199/share-of-us-internet-users-who-use-instagram-by-age-group/
- https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/interaction-plot/interpret-the-results/
- https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/main-effects-plot/interpret-the-results/