Missouri University of Science and Technology

What Gets a Paper Published?

Submitted to:

Dr. V.A. Samaranyake

Submitted by (**GROUP RHD**):

Ramy Khalef

Henry Wong

David Wood

Date:

12/5/19

Introduction

The objective of this experiment was to see if there was a relationship between the word count of articles published in the ASCE (American Society of Civil Engineers) Journal of Construction Engineering and Management and 3 separate factors each with 2 levels (high and low): the time of publication of articles, number of authors on the articles, and the number of downloads for the articles.

Factors

The time of publication was split between volume numbers as volume numbers were incremented as time went on. Volumes 109-126 were low level while volumes 127-145 were high level. The reason volumes before 109 were not evaluated was due to the lack of consistent records ASCE had on volumes for the Journal of Construction Engineering Management before volume 109.

The number of authors was split between 2 or less authors ($n >= 2$) and more than 2 authors ($n < 2$). Articles with less than or equal to 2 authors were low level while articles with more than 2 authors were considered high level. This number was decided by the mean number of authors for the entire population size of papers rounded down.

The number of downloads was split between less than or equal to 400 times and more than 400 times. Less than or equal to 400 times was low level while more than 400 times was high level. This number was decided by the mean number of downloads for the entire population rounded to the nearest hundred.

Brief Description, Randomization, and Output of the Experiment:

The word count of various journal papers in analyzed in relation to changes of time, number of downloads, and number of authors. In attempt to gather the data from the ASCE website, firstly, a python code script was generated to completely randomize (CR) selection, read, and output the results needed from the journal papers of the Journal of Construction Management in ASCE. Randomization occurred in the following order: Volume Number, Issue Number, and then Article Number. Further restrictions were added onto to the code as to not exceed and repeat the treatment combination that were already inputted from the code; as such, in every iteration, the code randomized through a population of papers that were not chosen. Figure 1 below clearly depicts the randomization procedure below.

Figure 1: Code flow of python mining script used to gather data, the chart flows top to bottom and right to left

The script is deemed "done" when all the treatment combinations are filled with recording of the response variable (word count). Accordingly, 8 readings from each replication was made (total 16 experimental run), 2 replications were made; hence our experiment is 2^3 one. A .cvs file was generated from the python script to output the data found from a completely randomized experiment (CRD). They are listed below and labeled in the appendix (table 1).

Moreover, as seen in Figure 1, a list of articles is generated by first randomly generating a volume number from both factor levels for volume. The script then separates the articles within the volume between those with 2 or less authors and more than 2 authors. The list of articles is then filtered again checking the number of downloads for each article. Then from that final list, an article is randomly generated. That article's data is then recorded (factors and response variable). Also, As seen in Figure 1, the entire code process was run until there were at least 2 articles per treatment combination. Technically, the program should only need to run twice, since one list generated a list for every treatment combination. However, in the case of a bug or an error, each treatment combination was checked to ensure it was filled correctly. The program did this by checking for empty values in the Dataframe that was used to store articles. Eventually, the Dataframe was exported to a .csv format which became the final version of the data collected.

The data was collected using a python script utiltizing beautiful soup to grab the raw code from ascelibrary.org. Regex functions were used to isolate specific parts of the website's html code to mine for the factors and the response variable. To get the word count, the function Counter from the collections modeule in the python standard library was used.

ANOVA test:

After attainting the needed result, further analysis was made on JMP to perform the ANOVA. We used the data in the appendix (table 1); as such the low/high levels in the three factors recorded with its according readings is listed below in Table 2.

Time Releases (X1)	Number of Downloads (X2)	Number of Authors (X3)	Word Count (Response or Yield)
0	0	1	4589
1	1	0	10556
1	0	1	6745
0	0	0	7539
0	1	1	7301
1	0	0	11354
1	1	1	7585
0	1	0	10470
1	0	1	10452
0	0	0	6651
1	1	1	7902
1	0	0	8811
0	0	1	6255
0	1	1	4380
0	1	0	5796
1	1	0	6998

Table 2

As a reminder of our different factors and levels, with the allocated low and high levels. Value of 0 will be allocated to low levels; as such, value of 1 will be allocated to that of high levels. Accordingly, our factors and levels are as such:

1- Factor #1 (Time Released):
   - Levels: Volumes 109-126(low),127-145(high)
2- Factor #2 (Number of Downloads):
   - Levels: Less than 400 times (low) / more than or equal 400 times (high)
3- Factor #3 (Number of Authors):
   - Levels: 2 or less (>=) (low) / More than 2 (>) (high)

After outputting those data, we performed an ANOVA test on JMP. After fitting the model, we output a Summary of Fit, Analysis of Variance, and Effect Tests. They are listed below in figure 2 with commentary following.

Summary of Fit

RSquare	0.501744
RSquare Adj	0.06577
Root Mean Square Error	2044.981
Mean of Response	7711.5
Observations (or Sum Wgts)	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	33689820	4812831	1.1509
Error	8	33455584	4181948	**Prob > F**
C. Total	15	67145404		0.4198

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Time Released	1	1	18970380	4.5363	0.0658
Number of Downloads	1	1	123904	0.0296	0.8676
Number of authors	1	1	10507322	2.5125	0.1516
Time Released*Number of Downloads	1	1	3270672	0.7821	0.4023
Time Released*Number of authors	1	1	524176	0.1253	0.7325
Number of Downloads*Number of authors	1	1	7140	0.0017	0.9681
Time Released*Number of Downloads*Number of authors	1	1	286225	0.0684	0.8002

Figure 3

One can see from the Summary of Fit that 16 readings were taken with 8 treatment combinations were made. The latter can be seen in the Analysis of Variance, since the degrees freedom in the model is 7, one less from 8 (df of model is 8-1). We will conduct and perform the ANOVA test on a basis of significance level of 0.05. The ANOVA model gave a To perform the ANOVA test, firstly we must check the Prob>F value in the Analysis of Variance, the value is 0.4198 which is greater than 0.05. Accordingly, we fail to reject the null hypothesis, and the means are zero, and thus insignficant. As such, we don't need to perform the three-way interation, two-way interaction, or main effects, since none will show interaction or significance. Furthemore, the ANOVA gave a value of 0.06577 in the adjusted R-squared, which is relatively low.

Regression Analysis test:

Further analysis should be taken into account after the ANOVA, of which include: Tukey, LSD, or Duncan test. Further, we shall analyze and create a model using regression analysis. However, instead of using the high and low values, we used the actual values from our initial results (in appendix) since our model can be quantitive of nature (shown below in table), they are summarized below in Figure 4. Accordingly, Regression Analysis was done using a code script on SAS, and output are listed below in Figure 5.

Volume number (**X1**)	Number of downloads (**X2**)	Number of authors (**X3**)	Word Count (**Y**)
126	101	3	4589
136	1241	1	10556
134	242	3	6745
115	96	2	7539
114	561	3	7301
141	328	2	11354
139	400	3	7585
124	449	2	10470
134	290	5	10452
118	74	1	6651
138	892	3	7902
143	254	1	8811
110	94	3	6255
122	540	3	4380
116	575	1	5796
132	448	2	6998

Figure 4

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	16
Number of Observations Used	16

Equation 1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	19641604	6547201	1.65	0.2294
Error	12	47503800	3958650		
Corrected Total	15	67145404			

Root MSE	1989.63565	R-Square	0.2925
Dependent Mean	7711.50000	Adj R-Sq	0.1157
Coeff Var	25.80089		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4387.93108	6235.30153	-0.70	0.4950
x1	1	93.42801	49.97982	1.87	0.0862
x2	1	0.96933	1.75143	0.55	0.5901
x3	1	-94.00359	482.11371	-0.19	0.8487

Figure 5

We can see from the Regression Analysis, the adjusted R-squared value is greater than that of the ANOVA, it has a value of 0.1157. This means that 11.57% of the data can be explained by the regression model. It is important to note that the adjusted R-squared value is also small. Nevertheless, the regression model produced a model of the following:

$$Y = -4387.93108 + 93.42801(X1) + 0.96933(X2) - 94.00359(X3)$$

Moreover, in performing "virtually" a simple regression test, at a significance level of 0.05, we fail to reject the null hypothesis that the slope parameters ($\beta_{1/2/3}$) for the variables are zero, and thus insignificant (from analysis of variance). Also, with a significance level of 0.05, we fail to reject the null hypothesis that the slope parameter (β_x) for each variable is zero, and thus insignificant (from Pr>|T| values in parameter estimates). Further, as for the intercept parameter in the equation, we fail to reject the null hypothesis that the β_0 is zero; thus insignificant. Also, the model's regression line could not be shown since it would require a 4D illustration since we have 1 Y variable as well as 3 Xs.

Tukey Comparison Test:

In order to conduct a comparison test, the test itself should be significant. In our experiment, the ANOVA result were insignificant in all tests, interactions, as well as main effects. Accordingly one should only conduct the Tukey test in significant interactions; nevertheless, if the three way interaction were significant, our Tukey test shows the below in figure 6.

LSMeans Differences Tukey HSD:
α=0.050 Q=3.9571
LSMean[i] By LSMean[j]

Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	0,0,0	0,0,1	0,1,0	0,1,1	1,0,0	1,0,1	1,1,0	1,1,1
0,0,0	0 0 0 0	1673 2044.98 -6419.2 9765.2	-1038 2044.98 -9130.2 7054.2	1254.5 2044.98 -6837.7 9346.7	-2987.5 2044.98 -11080 5104.7	-1503.5 2044.98 -9595.7 6588.7	-1682 2044.98 -9774.2 6410.2	-648.5 2044.98 -8740.7 7443.7
0,0,1	-1673 2044.98 -9765.2 6419.2	0 0 0 0	-2711 2044.98 -10803 5381.2	-418.5 2044.98 -8510.7 7673.7	-4660.5 2044.98 -12753 3431.7	-3176.5 2044.98 -11269 4915.7	-3355 2044.98 -11447 4737.2	-2321.5 2044.98 -10414 5770.7
0,1,0	1038 2044.98 -7054.2 9130.2	2711 2044.98 -5381.2 10803.2	0 0 0 0	2292.5 2044.98 -5799.7 10384.7	-1949.5 2044.98 -10042 6142.7	-465.5 2044.98 -8557.7 7626.7	-644 2044.98 -8736.2 7448.2	389.5 2044.98 -7702.7 8481.7
0,1,1	-1254.5 2044.98 -9346.7 6837.7	418.5 2044.98 -7673.7 8510.7	-2292.5 2044.98 -10385 5799.7	0 0 0 0	-4242 2044.98 -12334 3850.2	-2758 2044.98 -10850 5334.2	-2936.5 2044.98 -11029 5155.7	-1903 2044.98 -9995.2 6189.2
1,0,0	2987.5 2044.98 -5104.7 11079.7	4660.5 2044.98 -3431.7 12752.7	1949.5 2044.98 -6142.7 10041.7	4242 2044.98 -3850.2 12334.2	0 0 0 0	1484 2044.98 -6608.2 9576.2	1305.5 2044.98 -6786.7 9397.7	2339 2044.98 -5753.2 10431.2
1,0,1	1503.5 2044.98 -6588.7 9595.7	3176.5 2044.98 -4915.7 11268.7	465.5 2044.98 -7626.7 8557.7	2758 2044.98 -5334.2 10850.2	-1484 2044.98 -9576.2 6608.2	0 0 0 0	-178.5 2044.98 -8270.7 7913.7	855 2044.98 -7237.2 8947.2
1,1,0	1682 2044.98 -6410.2 9774.2	3355 2044.98 -4737.2 11447.2	644 2044.98 -7448.2 8736.2	2936.5 2044.98 -5155.7 11028.7	-1305.5 2044.98 -9397.7 6786.7	178.5 2044.98 -7913.7 8270.7	0 0 0 0	1033.5 2044.98 -7058.7 9125.7
1,1,1	648.5 2044.98 -7443.7 8740.7	2321.5 2044.98 -5770.7 10413.7	-389.5 2044.98 -8481.7 7702.7	1903 2044.98 -6189.2 9995.2	-2339 2044.98 -10431 5753.2	-855 2044.98 -8947.2 7237.2	-1033.5 2044.98 -9125.7 7058.7	0 0 0 0

Level		Least Sq Mean
1,0,0	A	10082.500
1,1,0	A	8777.000
1,0,1	A	8598.500
0,1,0	A	8133.000
1,1,1	A	7743.500
0,0,0	A	7095.000
0,1,1	A	5840.500
0,0,1	A	5422.000

Figure 6

Levels not connected by same letter are significantly different. Consequently, in figures 6 all levels are connected by the same letter; hence all levels are insignificant as it is important to note this test is not appropriate because none of our tests showed significance; and thus, no proper conclusion or comparison can be carried out.

Test for multicollinearity:

Despite the test holding to be insignificant from all aspects, multicollinearity test was conducted to test for evidence of multicollinearity in the model. After running the vin collin function on SAS, the SAS output was presented below in figure 7:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-4387.93108	6235.30153	-0.70	0.4950	0
x1	1	93.42801	49.97982	1.87	0.0862	1.10665
x2	1	0.96933	1.75143	0.55	0.5901	1.14168
x3	1	-94.00359	482.11371	-0.19	0.8487	1.04220

Figure 7

It is important to note that for a variable to be involved in multicollinearity it must have a VIF value of greater than 10. By analyzing figure 7 above, all of the variable in the model (x1, x2, and x3) have VIF values that are less than 10; as such, none of the variables are involved in multicollinearity. Since the model consists those models; hence the model itself as a whole does not contain any evidence to multicollinearity.

Selecting best model:

On attempting to select a model that is better performing, selection criteria using the ADJRSQ method was performed by outputting the BEST 7 models (hence, all applicable models). Accordingly, here are the results from the SAS output presented in figure 8:

Number in Model	Adjusted R-Square	R-Square	C(p)	MSE	Variables in Model
1	0.2165	0.2688	0.4029	3507062	x1
2	0.1811	0.2903	2.0380	3665715	x1 x2
2	0.1628	0.2745	2.3063	3747413	x1 x3
3	0.1157	0.2925	4.0000	3958650	x1 x2 x3
1	0.0212	0.0865	3.4947	4381292	x2
2	-.0540	0.0865	5.4943	4718204	x2 x3
1	-.0679	0.0033	4.9065	4780483	x3

Figure 8

It is apparent from the results above the following:

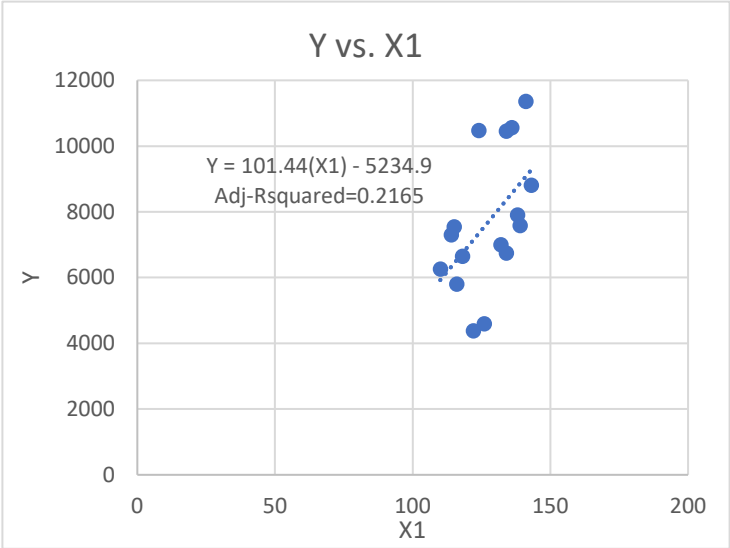
- After conducting the F-test for all the latter models, we conducted the following:

Model # (same orientation as in the past table)	Pr>|T|
1	0.0397
2	0.1077
3	0.1242
4	0.2294
5	0.2689
6	0.5554
7	0.8337

Figure 9

  Accordingly only the first model is significant, with a Pr>|T| value of 0.0397, we reject the null hypothesis that the means of x1 are zero; thus they significant and at least one of the means is different than zero.
- The model with the highest adjusted R-squared value is the first model, which had an R-squared value of 0.2165, and 1 variable [Volume Number (x1)].
- The model with the lowest MSE is also the first model, with MSE=3507062
- Models 1, and 3 only satisfy the condition of Cp being less or equal to p+1, noting that model 1 has the less Cp value.

Accordingly, and due to the observations listed above, model 1 is the best model in selection for the model. A regression line was fitted with its according equation to show how this model performs linearly:



Limitations:

Within the scope of our work, we are constrained with a certain knowledge, it is certain that further analysis could be done. Furthermore, this data is also subject to change throughout time, since the number of downloads will change through time, this may show significance latter to that aspect, but nothing is certain until testing occurs through time. Moreover, no prediction of intervals were made since the model itself is insignificant from all aspects; hence, no proper reliance on conclusions to such were appropriate to be made.

Conclusion and recommendations:

It is apparent that this test is not a significant one, nor does the better selection of the model with X1 represent a better model since it has a very low Adjusted R-squared value; meaning , the formulation and model do not represent the entire population appropriately. Accordingly, a better set of factors and levels need to be determined for further testing; further, more reading should be made to accurately draw conclusion for the new model with factor that more appropriately represent the yield or outcome.

References used:

https://ascelibrary.org/journal/jcemd4

Appendix

	Randomization within the below							Slots taken for treatment combination (2 replications)			
reading	volume (**X1**)	issue	article	downloads (**X2**)	authors (**X3**)	title	response (**Y**)	Time Releases	Number of Downloads	Number of Authors	Replication
1	126	1	3	101	3	Design/Build Methods for Eletrical Contracting Industry	4589	0	0	1	1
2	136	4	3	1241	1	PPP Experiences in Indian Cities: Barries, Enablers, and the Way Forward	10556	1	1	0	2
3	134	7	10	242	3	Modeling Time-Constraints in Construction Operations through Simulations	6745	1	0	1	1
4	115	1	3	96	2	Application of Robotics in Bridge Deck Fabrication	7539	0	0	0	2
5	114	3	1	561	3	Artificial Intellignence Techniques for Generating Construction Project Plans	7301	0	1	1	1
6	141	3	2	328	2	Synthetic Cash Flow Model with Singularity Functions. II: Feasible Prompt Payment Discount Scenarios	11354	1	0	0	2
7	139	5	5	400	3	Scheduling Model for Rehabilitation of Distribution Networks Using MINLP	7585	1	1	1	1
8	124	5	3	449	2	Incentive/Disincentive Contracts: Perceptions of Owners and Contractors	10470	0	1	0	2
9	134	12	1	290	5	Analysis of Techniques Leading to Radical Reduction in Proejct Cycle Time	10452	1	0	1	2
10	118	3	1	74	1	A Challenge for Research	6651	0	0	0	1
11	138	10	9	892	3	Using Pajek and Centrality Analysis to Identify a Social Network fo Construction Trades	7902	1	1	1	2
12	143	4	9	254	1	Metrics for Management of Asphalt Plant Sustainability	8811	1	0	0	1
13	110	2	3	94	3	Settlement of Construction Jurisdictional Disputes	6255	0	0	1	2
14	122	4	4	540	3	Holistic Appraisal of Value Engineering in Construction in United States	4380	0	1	1	2
15	116	4	4	575	1	Decision-Support System for Modeling Bid/No-Bid Decision Problem	5796	0	1	0	1
16	132	4	7	448	2	Constructability Analysis of the Bridge Superstructure Rotation Construction Method in China	6998	1	1	0	1