Indrajit Choudhury

Hospital Supplier Analysis

Executive Summary:

Using a combination of contingency table investigation, GLM Net Modelling Techniques and general GLM, I have estimated that for the Purchase Order Data set given, which chronicled results during a 36 Month Period, that in the 37th Month, the proportion of Purchase Orders that will be owned by Credo-based suppliers and by Small-Business based suppliers (for the CLPS – R&D Area within Johnson & Johnson) will be 0.069707 and 0.1629539 respectively.

<u>**Introduction:**</u>

We are given Purchase Order Data from Johnson & Johnson, specifically taken from the R&D Area, which is also called CLPS. This data has 907,408 observations consisting of 29 variables. Our variables of interest are 2 of these 29 variables:

1. Credo.Spend.Ind: This is a binary variable that acts as an indicator of whether or not the supplier of one of the observations is determined to be diverse (which essentially means that the business is woman or minority owned) by ARIBA standards. It takes two values: “Y”, which indicates that the supplier is considered diverse and “N”, which indicates that the supplier is not considered diverse. We are interested in those observations that take the value “Y”. We will refer to observations that take the value “Y” as Credo observations.

2. Business.Size.Code: This is a categorical variable that refers to the size of the supplier’s business, as classified by the government. It takes 4 values: “0”, “D”, “L” and “S” where the latter two represent large and small businesses. We are interested in those observations that take the value “S”. We will refer to observations that take the value “S” as Small Business (or SB) observations.

The data gives Purchase Order information over a 36-month period from 1/2010 to 12/2012, which we will call Months 1 to 36 for convenience. The purpose of this month is to use the data given to estimate the proportion of total observations that are Credo observations and the proportion of total observations that are Small Business observations in the 37th month (which would be 1/2013).

The general analysis plan for this project involves the following steps, which will be expanded upon in the results section:

1. Extract the individual dates for each observation in the Purchase Order Data and set it equal to a number between 1 and 36

2. Perform proper investigative techniques within the Purchase Order Data to find several levels of categories which are associated with high proportions of SB and Credo

3. Create covariates for each observation based on the results of Step 2.

4. Create two covariate matrices (one for Credo, one for SB), which will have two components: the covariates chosen in Step 3 and time series components, which will be Credo or SB values during a given month.

5. Using the two covariate matrices, create two smaller matrices which each have 33 months, where each observation has the average value for each covariate for each month, as well as the proportion of the response (SB or Credo) for the last three months.

6. Using the matrices from Step 5, use appropriate GLM Models to estimate the proportion of Credo and SB for the 37th month.

RESULTS

Creation of Time Series Component:

First, I read the PO Data file into R using read.csv. Then, I decided to create an additional variable called poMonth which would indicate which month (out of 36) that each of the observations belonged to. I did this by extracting the month and year from the variable PO.Create.Date using the strsplit function and the ldply function within the plyr package. Then using the year and month, I filled out poMonth to index each observation as belonging to one of the months between 1 and 36.

Next, to get a general idea of what each proportion looked like through the 36 months, I created contingency tables for Month and Credo and Month and Small Business and then plotted each of them by month, also plotting their overall overages (Figures 1 and 2). Additionally, the tables containing these values are Figures 3 and 4.

Figure 1 shows that Credo Percentage for the most part shows a random pattern, although we do note that around the 30th Month there appears to be a relative drop-off where an increase only occurs at the very last month. Additionally, the overall average for Credo Percentage for these 36 months is 0.1006592, though majority of points are somewhat higher or lower than the average.

Figure 2 shows that Small Business Percentage also has a series of ups and downs in a random fashion but unlike Credo Percentages, the variance exhibited by the Small Business Proportion appears to be a bit smaller. Also, the overall average for Small Business percentage is .1919467.

Keeping this information in mind, I decided that to predict both proportions for the 37th month, I would create a GLM Model, that would incorporate time series terms (proportions in previous months) as well as covariates that I would manually extract from the data.

Methodology for Creating New Covariates:

On top of the time series terms, which would be proportion of Credo or Small Business for a particular month, I also wanted to add various other covariates extracted from the Purchase Order Data. I did this with a systematic approach of studying proportion trends for many subgroups.

To find subgroups, I took one of the 29 variables and created a contingency table between that variable and the variable of interest (Credo or Small Business). Then, using that table for each level of the variable, I found the count of the number of observations that had that level value and also had the value of the variable of interest. Then, I divided this number by the total number of observations that had that level value to find the proportion of Purchase Orders with the variable of interest. I tabulated this number for each level of the variable and subtracted the overall percentage value from this value. I took the absolute value of this and sorted the observations from highest absolute difference to lowest. If there were a large number of levels, I filtered this table to only show observations with high absolute differences and high total occurrence. Using this sorted table, if one of the levels of a variable exhibited a high deviation from the overall percentage of the variable of interest, I would use it as the basis of a new covariate, which would be a binary variable with value 1 if an observation exhibited that level of that variable and equal to 0 otherwise.

I exhibit this methodology by showing as an example, the process by which I extracted covariates for my Credo models based on values of Company. Looking just at Credo Proportion, I attempt to create new covariates for my GLM Model based on the levels of the variable Company. Company has 13 different levels and using the table command I created a contingency table that calculated the proportion of Credo observations within each of the 13 levels. Additionally, I also calculated the difference between each of these proportions and the overall Credo proportion for the entire dataset, 0.1006592. I found the absolute value of this difference and sorted the table by this value. This table is shown in Figure 5. As we can see, the Company levels HCS, Nutritionals and PR all deviate from average Credo percentage by 10% or more, with Credo proportion values of 0.307004713, 0.262573964 and 0.212037037 respectively. As such, I use this as a basis to create the following 3 variables:

1. $X_1 = \begin{matrix} 1\ if\ Company = HCS \\ 0\ otherwise \end{matrix}$

2. $X_2 = \begin{matrix} 1\ if\ Company = Nutritionals \\ 0\ otherwise \end{matrix}$

3. $X_3 = \begin{matrix} 1\ if\ Company = PR \\ 0\ otherwise \end{matrix}$

I continue this process for various other variables for both Credo and Small Business Proportions.

<u>Summary of New Variables Created:</u>

Credo:

I created contingency tables for Credo Proportion and the variables Company, Category.Name, Subcategory.Name, Business.Size.Code and JNJ.Site.Code in order to find covariates to add to my GLM model for Credo Proportion. The Company table can be seen in Figure 5, the Category Name table can be seen in Figure 6, a subset of the Subcategory Name table can be seen in Figure 7, the Business Size Code table can be seen in Figure 8 and a subset of the JNJ Site Code Table can be seen in Figure 9. Based on these tables, 13 Covariates were created. Each Covariate takes the form $X_i = \begin{matrix} 1\ if\ condition\ fulfilled \\ 0\ otherwise \end{matrix}$. The 13 conditions for the covariates are listed below:

1. Company = HCS

2. Company = Nutritionals

3. Company = PR

4. Category.Name = Consulting- Labor and Professional Services

5. Subcategory.Name = Managed Service Provider

6. Subcategory.Name = Professional Services

7. Subcategory.Name = Temporary Staffing

8. Business.Size.Code = D

9. Business.Size.Code = S

10. JNJ.Site.Code = 171013

11. JNJ.Site.Code = 141018

12. JNJ.Site.Code = 198001

13. JNJ.Site.Code = 165999

Small Business:

I also created contingency tables for Small Business Proportion and the variables Company (Figure 10), Category.Name (which will be omitted since no significant deviations were found), Subcategory.Name (Figure 11), Credo.Spend.Ind (Figure 12) and JNJ.Site.Code (Figure 13) in order to find covariates to add to my GLM model for Small Business Proportion. The Company table can be seen in Figure 5, the Category Name table can be seen in Figure 6, a subset of the Subcategory Name table can be seen in Figure 7, the Business Size Code table can be seen in Figure 8 and a subset of the JNJ Site Code Table can be seen in Figure 9. Based on these tables, 12 Covariates were initially created and 7 more were added later. These covariates take the same form as the Credo covariates and their conditions for a response value of 1 are:

1. Company = PR

2. Company = J&J Medical

3. Company = Global Ortropaedics

4. Subcategory.Name = Clinical - Data Management Technology

5. Subcategory.Name = Clinical - R&D Medical Testing (non-lab)

6. Subcategory.Name = Product Development - Engineering & Testing

7. Subcategory.Name = Professional Services

8. Subcategory.Name = Training and Development

9. Subcategory.Name = R&D Lab Supplies - Equipment and Instrumentation

10. Subcategory.Name = Product Development - Product Design & Prototyping

11. Subcategory.Name = Consulting

12. Subcategory.Name = Clinical - Clinical Lab Services

13. Subcategory.Name = Memberships & Subscriptions

14. Credo.Spend.Ind = Y

15. JNJ.Site.Code = 165999

16. JNJ.Site.Code = 129001

17. JNJ.Site.Code = 620301

18. JNJ.Site.Code = 445501

19. JNJ.Site.Code = 151001

Regression Models Chosen:

Credo:

Using the 13 variables, I created a matrix that consisted of the 907,408 observations, with 15 columns, one column referring to the indexed month, one column referring to the Credo binary value for that observation and the next 13 columns referring to the 13 covariates that we had defined. Next, I created a new matrix with 33 rows and 18 columns. Each of the rows refers to a single month (out of months 34-36), with the next column referring to the Credo Proportion for that month, the next 13 columns referring to the average of each of the covariates for that particular month and the last three columns referring to the Credo Proportions for the prior three months (which is the reason that we start the matrix at the 4^{th} month). For each model, the y response is the Credo Proportion for the current month, while the 13 covariates and the 3 prior Credo proportion terms are where the explanatory terms are chosen from.

I initially fit a GLM Model (specifying family as Gaussian) for the full model using the entirety of the X matrix. Then, using the predict function, I predicted the Credo Proportion of the 37^{th} month, which turned out to be 0.03783186. Then, I plotted the predicted fit using the full model, the actual proportions as well as the predicted value of the 37^{th} month, which can be seen in Figure 14. Additionally, we do the exact same thing, but fitting the GLM Gaussian Model for a reduced model which only includes the 3 time series terms as covariates. This time, we get a predicted Credo Proportion of 0.07285621 and plot the predicted fit (Figure 15). Next, we wish to find the proper subset of covariates to model in a GLM. To do this, we make use of GLM Net, where the cv.glmnet function gives us the appropriate covariates among our full set of covariates to model in a GLM Model. Below, I show the list of covariates chosen by GLM Net:

```
> coef(cvfit, s = "lambda.min")
17 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept)  0.05862258
V1           .
V2           .
V3           .
V4           .
V5           .
V6          -0.03444804
V7           0.15996946
V8           .
V9           .
V10          .
V11          .
V12          .
V13          .
V14          .
V15          0.05176430
V16          .
```

Based on this, the 6th Covariate, 7th Covariate and the 2nd Time Series Covariate are considered appropriate to include in a GLM Model. I decided to use the 6th and 7th Covariates as well as all of the Time Series Covariates (since it would be odd to include one over the other in this instance) in a GLM Model. The results of this are listed below:

```
> sumfit

Call:
glm(formula = y ~ newx, family = gaussian())

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.069707  -0.008214  -0.000319   0.008471   0.031203

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.04497    0.02272   1.980  0.05765 .
newx1       -1.31815    1.00836  -1.307  0.20177
newx2        0.32406    0.10502   3.086  0.00454 **
newx3       -0.52592    0.37025  -1.420  0.16652
newx4        0.18613    0.25445   0.732  0.47054
newx5        0.26283    0.22848   1.150  0.25973
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The predicted Credo Proportion for Month 37 given by this model is 0.069707. Additionally, the actual Credo Proportion Values, predicted Credo Proportion Values and predicted Credo Proportion for the 37th Mont hare plotted and shown in Figure 16. From this graph, we conclude that within the 4-36 Month period, the model appears to predict the value of Credo Proportion somewhat accurately and as a result, we will accept the estimated Credo Proportion Value of 0.069707 as a valid estimate.

Small Business:

Initially, I used 12 covariates as predictors for Small Business, but I wanted to get a more accurate result so I added 7 additional covariates. This ultimately did not make much of an improvement on my fitted model, but decided to keep them since the GLM Net gave the same result anyway.

So, using the 19 covariates, I created a matrix with all of the observations and 21 columns, done in the same way that I did for the Credo variables. Then I created the matrix with data from Months 4-33, where the first column referred to the month, the second column referred to the Small Business Proportion for that month, the next 19 columns referred to the average of the covariates for each of the months, and the last three columns referred to the average Small Business Proportion values for the prior 3 months.

I initially fit a GLM Model (specifying family as Gaussian) for the full model using the entirety of the X matrix. Then, using the predict function, I predicted the Small Business Proportion of the 37th month, which turned out to be 0.09155772. Then, I plotted the predicted fit using the full model, the actual proportions as well as the predicted value of the 37th month, which can be seen in Figure 17. Additionally, we do the exact same thing, but fitting the GLM Gaussian Model for a reduced model which only includes the 3 time series terms as covariates. This time, we get a predicted Small Business

Proportion of 0.1761756 and plot the predicted fit (Figure 18). Next, we wish to find the proper subset of covariates to model in a GLM. To do this, we make use of GLM Net, where the cv.glmnet function gives us the appropriate covariates among our full set of covariates to model in a GLM Model. Below, I show the list of covariates chosen by GLM Net:

```
> coef(cvfit, s = "lambda.min")
23 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept) 0.1769208
V1          1.6307093
V2          .
V3          .
V4          .
V5          .
V6          .
V7          .
V8          .
V9          .
V10         .
V11         .
V12         .
V13         .
V14         .
V15         .
V16         .
V17         .
V18         .
V19         .
V20         .
V21         .
V22         .
```

We note that only the first Covariate is significant. Additionally, when I ran this model with only 12 extra covariates instead of 19, I still got this same result from GLM Net. As a result, I decided to fit my GLM Model as the first Covariate and then the three time series terms. The results of this GLM Model are given below:

```
> sumfit

Call:
glm(formula = y ~ newx, family = gaussian())

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-0.162954  -0.004324   0.005424   0.014409   0.034172

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.12816    0.04366   2.935  0.00646 **
newx1        5.75812    3.49070   1.650  0.10982
newx2        0.03266    4.27625   0.008  0.99396
newx3        0.19404    0.34821   0.557  0.58164
newx4        6.66414    6.59414   1.011  0.32056
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The predicted Small Business Proportion for Month 37 given by this model is 0.1629539. Additionally, the actual Small Business Proportion Values, predicted Small Business Proportion Values and predicted Small Business Proportion for the 37th Month are plotted and shown in Figure 19. From this graph, we conclude that within the 4-36 Month period, the model appears to predict the value of Credo Proportion somewhat accurately and as a result, we will accept the estimated Credo Proportion Value of 0.1629539 as a valid estimate.

Conclusion:

The purpose of this project was to use Purchase Order data during a 36 month period to try to estimate the proportion of Purchase Orders (for the CLPS- R&D area of Johnson & Johnson) in the 37th Month that are filled out by Credo-based suppliers or Small Business-based suppliers. I approached this problem by separating the large dataset by month and predicting each of the two proportions for each of the months by using a GLM model that uses a combination of time series variables and other covariates that are chosen based on subgroups. The time series variable for each month are the proportions for Credo or Small Business for each of the previous three months and the additional covariates were found as indicator variables for subgroups that were discovered by investigating relationships in contingency tables within the data. After using GLM Net to select which models to fit for both Responses, I have estimated that the proportion of Purchase Orders that will be filled by Credo and Small Business owned suppliers will be 0.069707 and 0.1629539 respectively.

Appendix:

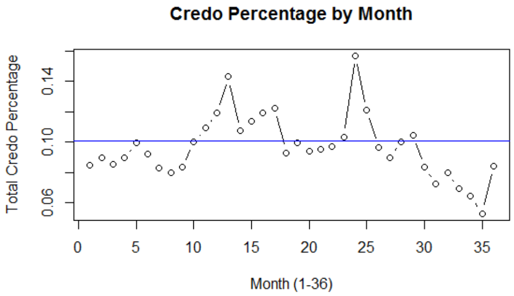

Figure 1

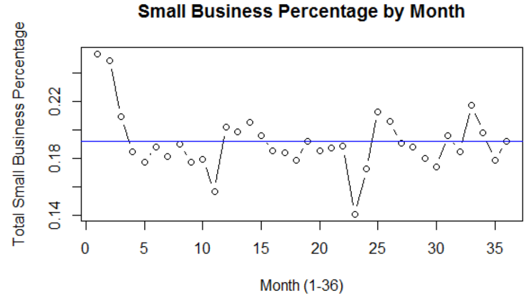

Figure 2

> print(credo_table)

	N	Y	month_credo_splits	credo_above_average
1	28641	2646	0.08457187	0
2	22395	2197	0.08933800	0
3	25530	2378	0.08520854	0
4	22518	2213	0.08948284	0
5	21249	2342	0.09927515	0
6	22470	2271	0.09179095	0
7	19195	1738	0.08302680	0
8	20524	1772	0.07947614	0
9	19569	1777	0.08324745	0
10	20601	2287	0.09992136	0
11	27055	3311	0.10903642	1
12	24933	3369	0.11903752	1
13	35346	5901	0.14306495	1
14	24819	2980	0.10719810	1
15	27604	3540	0.11366555	1
16	22368	3024	0.11909263	1
17	21797	3043	0.12250403	1
18	22566	2309	0.09282412	0
19	18373	2029	0.09945103	0
20	20940	2163	0.09362420	0
21	19476	2042	0.09489730	0
22	19651	2109	0.09692096	0
23	26527	3056	0.10330257	1
24	30808	5726	0.15673072	1
25	39438	5428	0.12098248	1
26	29286	3130	0.09655726	0
27	28975	2855	0.08969526	0
28	22910	2554	0.10029846	0
29	22986	2677	0.10431360	1
30	20530	1870	0.08348214	0
31	16662	1295	0.07211672	0
32	17117	1482	0.07968170	0
33	14216	1059	0.06932897	0
34	14509	996	0.06423734	0
35	13313	741	0.05272520	0
36	11172	1029	0.08433735	0

Figure 3

> print(sb_table)

	O	D	L	S	month_sb_splits	sb_above_average
1	0	143	23212	7932	0.2535238	1
2	1	49	18415	6127	0.2491461	1
3	0	87	21974	5847	0.2095098	1
4	0	128	20034	4569	0.1847479	0
5	0	380	19020	4191	0.1776525	0
6	0	117	19964	4660	0.1883513	0
7	0	133	16999	3801	0.1815793	0
8	3	44	18009	4240	0.1901686	0
9	17	60	17477	3792	0.1776445	0
10	3	142	18638	4105	0.1793516	0
11	6	298	25287	4775	0.1572482	0
12	1	279	22306	5716	0.2019645	1
13	2	440	32608	8197	0.1987296	1
14	4	307	21769	5719	0.2057268	1
15	1	432	24602	6109	0.1961534	1
16	0	248	20438	4706	0.1853340	0
17	7	371	19891	4571	0.1840177	0
18	2	151	20267	4455	0.1790955	0
19	2	300	16172	3928	0.1925301	1
20	0	217	18605	4281	0.1853006	0
21	3	131	17349	4035	0.1875174	0
22	0	110	17546	4104	0.1886029	0
23	1	162	25247	4173	0.1410607	0
24	2	123	30101	6308	0.1726611	0
25	0	543	34772	9551	0.2128783	1
26	0	57	25680	6679	0.2060402	1
27	0	129	25621	6080	0.1910148	0
28	2	225	20449	4788	0.1880302	0
29	0	177	20869	4617	0.1799088	0
30	3	128	18369	3900	0.1741071	0
31	0	101	14332	3524	0.1962466	1
32	2	180	14975	3442	0.1850637	0
33	0	81	11868	3326	0.2177414	1
34	0	139	12296	3070	0.1980006	1
35	0	141	11400	2513	0.1788103	0
36	1	189	9668	2343	0.1920334	1

Figure 4

> print(sorted_credo_company_splits)

	N	Y	credo_company_splits	credo_company_diff
HCS	4264	1889	0.307004713	0.2063454728
Nutritionals	997	355	0.262573964	0.1619147241
PR	2553	687	0.212037037	0.1113777967
J&J Medical	4028	8	0.001982161	-0.0986770798
OTC	31589	6183	0.163692682	0.0630334420
Corporate	44562	8409	0.158747239	0.0580879987
Pharm R&D	259667	11894	0.043798631	-0.0568606095
Global Surgery	98717	17332	0.149350705	0.0486914649
Consumer Products	90254	15198	0.144122444	0.0434632039
JSC	81339	7333	0.082698033	-0.0179612072
Pharm Commercial	34334	3109	0.083032877	-0.0176263637
Global Ortropaedics	46706	5936	0.112761673	0.0121024328
Global Medical Solution	117059	13006	0.099996156	-0.0006630846

	credo_company_abs_diff
HCS	0.2063454728
Nutritionals	0.1619147241
PR	0.1113777967
J&J Medical	0.0986770798
OTC	0.0630334420
Corporate	0.0580879987
Pharm R&D	0.0568606095
Global Surgery	0.0486914649
Consumer Products	0.0434632039
JSC	0.0179612072
Pharm Commercial	0.0176263637
Global Ortropaedics	0.0121024328
Global Medical Solution	0.0006630846

Figure 5

> print(categ_credo_table)

	N	Y	categ_credo_splits
Consulting- Labor and Professional Services	264139	69723	0.20883778
Research & Development (Products & Packaging)	551930	21616	0.03768835

	categ_credo_diff
Consulting- Labor and Professional Services	0.10817854
Research & Development (Products & Packaging)	-0.06297089

Figure 6

```
> head(sorted_subcateg_credo_table)
                                                 N     Y subcateg_credo_splits
Preclinical - Anatomical Matls & Testing (Human)    21    12             0.3636364
Managed Service Provider                          2442  1264             0.3410685
Professional Services                             2625  1194             0.3126473
Temporary Staffing                              140450 46203             0.2475342
Consulting                                       47542 15551             0.2464774
Benefits (Employees)                               544     0             0.0000000
                                                subcateg_credo_diff
Preclinical - Anatomical Matls & Testing (Human)           0.2629771
Managed Service Provider                                   0.2404093
Professional Services                                      0.2119880
Temporary Staffing                                         0.1468750
Consulting                                                 0.1458182
Benefits (Employees)                                      -0.1006592
                                                subcateg_credo_absdiff
Preclinical - Anatomical Matls & Testing (Human)              0.2629771
Managed Service Provider                                      0.2404093
Professional Services                                         0.2119880
Temporary Staffing                                            0.1468750
Consulting                                                    0.1458182
Benefits (Employees)                                          0.1006592
```

Figure 7

```
> print(sorted_bus_credo_table)
       N     Y bus_credo_splits bus_credo_diff bus_credo_absdiff
D      5  6937       0.99927975     0.89862051        0.89862051
S 128856 45318       0.26018809     0.15952885        0.15952885
O     63     0       0.00000000    -0.10065924        0.10065924
L 687145 39084       0.05381774    -0.04684151        0.04684151
```

Figure 8

```
> head(sorted_site_credo_table)
          N    Y site_credo_splits site_credo_diff site_credo_absdiff large_group
171013  762 1051         0.5797022       0.4790429          0.4790429           1
141018 2563 1415         0.3557064       0.2550471          0.2550471           1
198001  748  386         0.3403880       0.2397288          0.2397288           1
165999 1527  745         0.3279049       0.2272457          0.2272457           1
198011  956  416         0.3032070       0.2025478          0.2025478           1
138239 3036 1276         0.2959184       0.1952591          0.1952591           1
```

Figure 9

```
> print(sorted_sb_company_table)
                        O    D      L     S sb_company_splits sb_company_diff
PR                      0    0   1902  1338       0.412962963     0.221016231
J&J Medical             0    0   4013    23       0.005698712    -0.186248020
Global Ortropaedics    11  115  37243 15273       0.290129554     0.098182823
Global Surgery          9 4055  82413 29572       0.254823394     0.062876662
Pharm Commercial        2   39  32237  5165       0.137943007    -0.054003725
Nutritionals            0   10   1011   331       0.244822485     0.052875753
JSC                     0  208  75871 12593       0.142017773    -0.049928958
OTC                     0  107  28745  8920       0.236153765     0.044207033
Global Medical Solution 8  333 107017 22707       0.174581940    -0.017364792
Pharm R&D              18 1106 222168 48269       0.177746436    -0.014200295
Corporate              15   64  43267  9625       0.181703196    -0.010243536
Consumer Products       0  286  85946 19220       0.182263020    -0.009683712
HCS                     0  619   4396  1138       0.184950431    -0.006996301
                        sb_company_abs_diff
PR                              0.221016231
J&J Medical                     0.186248020
Global Ortropaedics             0.098182823
Global Surgery                  0.062876662
Pharm Commercial                0.054003725
Nutritionals                    0.052875753
JSC                             0.049928958
OTC                             0.044207033
Global Medical Solution         0.017364792
Pharm R&D                       0.014200295
Corporate                       0.010243536
Consumer Products               0.009683712
HCS                             0.006996301
```

Figure 10

```
> sorted_subcateg_sb_table[1:10,]
                                                 O    D     L     S
Clinical - Data Management Technology            0    0   207   942
Clinical - R&D Medical Testing (non-lab)         0    0   394  1766
Product Development - Engineering & Testing      0    0   472   942
Professional Services                            0    0  1746  2073
Training and Development                         2    3  9899 11342
R&D Lab Supplies - Equipment and Instrumentation 0 1586 73800 67082
Product Development - Product Design & Prototyping 2    0   813   689
Consulting                                      18   38 34862 28175
Clinical - Clinical Lab Services                 0    0  4601  3689
Memberships & Subscriptions                      0    0  2438  1908
                                                 subcateg_sb_splits subcateg_sb_diff
Clinical - Data Management Technology                     0.8198433        0.6278966
Clinical - R&D Medical Testing (non-lab)                  0.8175926        0.6256459
Product Development - Engineering & Testing               0.6661952        0.4742485
Professional Services                                     0.5428123        0.3508655
Training and Development                                  0.5338417        0.3418949
R&D Lab Supplies - Equipment and Instrumentation          0.4708566        0.2789099
Product Development - Product Design & Prototyping        0.4581117        0.2661650
Consulting                                                0.4465630        0.2546163
Clinical - Clinical Lab Services                          0.4449940        0.2530472
Memberships & Subscriptions                               0.4390244        0.2470777
                                                 subcateg_sb_absdiff large_group
Clinical - Data Management Technology                      0.6278966           1
Clinical - R&D Medical Testing (non-lab)                   0.6256459           1
Product Development - Engineering & Testing                0.4742485           1
Professional Services                                      0.3508655           1
Training and Development                                   0.3418949           1
R&D Lab Supplies - Equipment and Instrumentation           0.2789099           1
Product Development - Product Design & Prototyping         0.2661650           1
Consulting                                                 0.2546163           1
Clinical - Clinical Lab Services                           0.2530472           1
Memberships & Subscriptions                                0.2470777           1
```

Figure 11

```
> print(sorted_credo_sb_table)
   O    D      L      S credo_sb_splits credo_sb_diff credo_sb_absdiff
Y  0 6937  39084  45318       0.4961517    0.30420497       0.30420497
N 63    5 687145 128856       0.1578984   -0.03404832       0.03404832
```

Figure 12

```
> head(sorted_site_sb_table)
       O  D     L    S site_sb_splits site_sb_diff site_sb_absdiff large_group
165999 0  1  1169 1102      0.4850352    0.2930885       0.2930885           1
129001 0  0   683  640      0.4837491    0.2918023       0.2918023           1
620301 0  0   653  569      0.4656301    0.2736834       0.2736834           1
445501 0  0   770  640      0.4539007    0.2619540       0.2619540           1
151001 4 31 10696 7575      0.4137988    0.2218520       0.2218520           1
94501  2  4  3400 2284      0.4014060    0.2094592       0.2094592           1
```

Figure 13

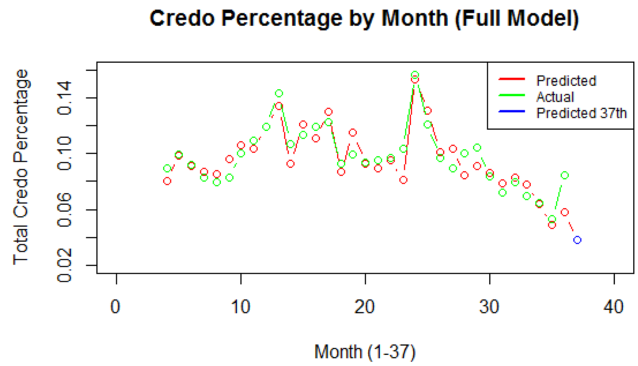

Figure 14

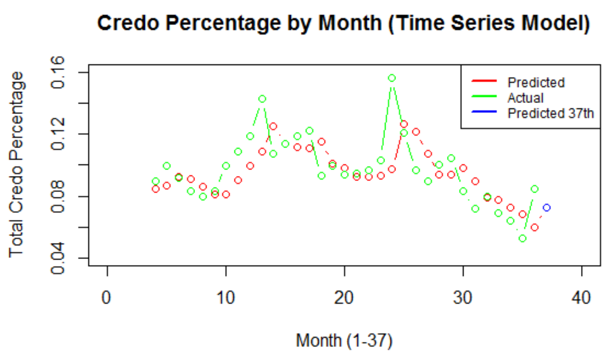

Figure 15

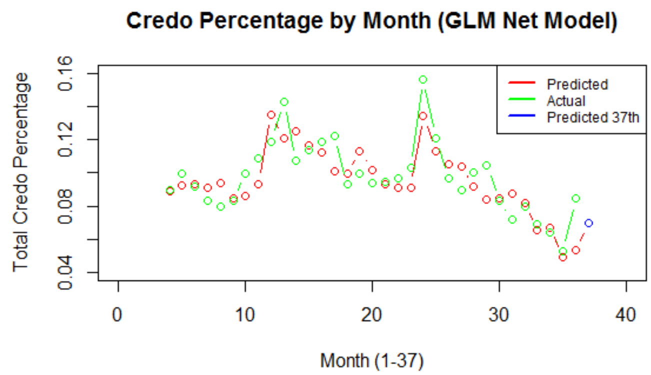

Figure 16

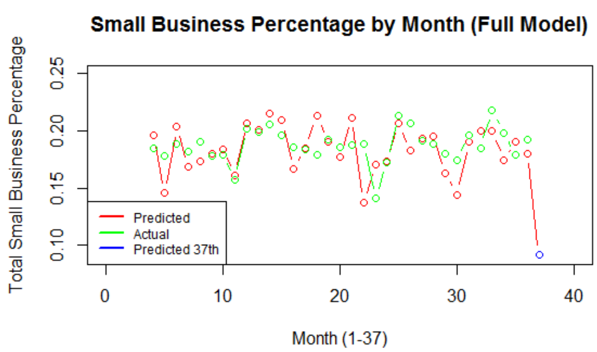

Figure 17

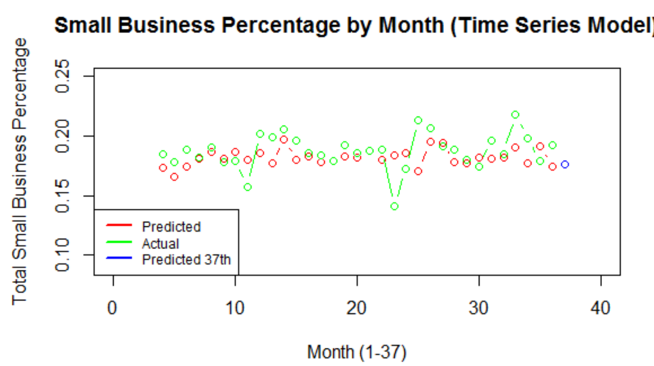

Figure 18

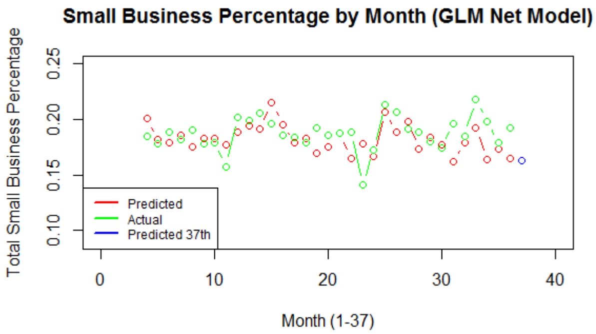
Small Business Percentage by Month (GLM Net Model)
Total Small Business Percentage
0.10
0.15
0.20
0.25
Predicted
Actual
Predicted 37th
0
10
20
30
40
Month (1-37)

Figure 19