

Date: June 14, 2020

To: Jeremy Bearimy

CC: Warren Buffett

From: Aida Marchoud

Subject: Profit Predictors

Mr. Buffett would like to see the tools that are being utilized in the Data Mining field could be useful in predicting companies that have profit. He would like to compare the findings with his own to see if they offer any significance in changing his methods. The data provided by Mr. Buffett's team is from April 23, 1990. It will make is easier to compare the predicted and the actual amount. This report will show the models and the significant variables of those models. Mr. Buffett's team will use the information that has been collected below and will give Mr. Buffet the comparisons.

Contents of the report:

- Define Data terms
- Statistic Summary
- Regression Reports
 - Regression Model with Original Data
 - Regression Model with Original Data using Stepwise
 - Regression Model using Cluster Analysis for:
 - 3 clusters
 - Only 3 Clusters
 - 3 Cluster with Original Data
 - 3 Clusters with Original Data and Stepwise (if necessary)
 - 7 clusters
 - Only 7 Clusters
 - 7 Cluster with Original Data
 - 7 Clusters with Original Data and Stepwise (if necessary)
 - 15 clusters
 - Only 15 Clusters
 - 15 Cluster Regression using stepwise
 - 15 Cluster with Original Data
 - 15 Clusters with Original Data and Stepwise (if necessary)
- Summary
- Appendix

DEFINE DATA TERMS:

Type of industry: Drugs, Computer, Oil, Aerospace, Beverages, Soap

Sales (\$M): Sales in millions

Profit (\$)

#emp: number of employees

Profits/emp: Profits divided by number of employees

Assets (\$M): Assets in Millions

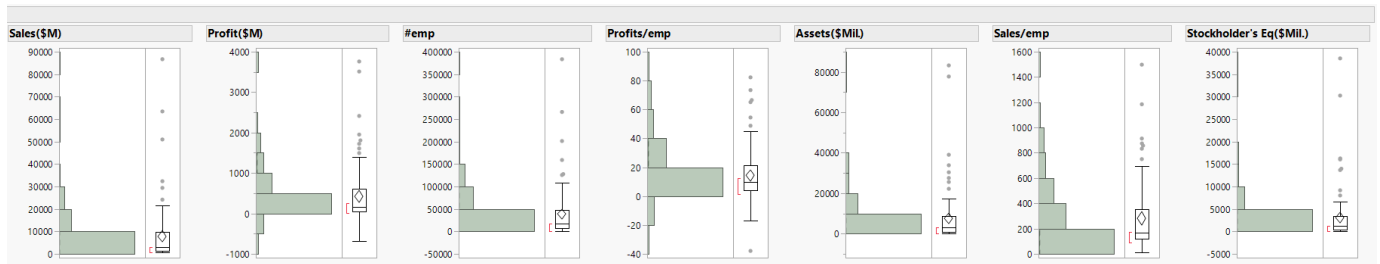
Sales/emp: Sales divided by the number of employees

Stockholder's Eq(\$M): stockholder's equity in millions

STATISTICAL SUMMARY *(See Appendix for the Statistics information)*

When running the statistical summary initially, there looked to be data missing. Before proceeding those missing pieces of data needed to be filled. By implementing a technique, the data set was whole.

The statistical summary was reran as to show the now whole set of data. What is shown is that a number of variables are extremely skewed. To rectify this, the data needed to be displayed over a wide range of values in a compact way by taking the data and using what is called a log formula.



Also it is worth noting, that a lot of the data had significant outliers, which is why the data was so skewed. It might have been worth deleting the data, but because we wanted to compare and contrast with the actual data it made more sense to leave it as is and fix the skew.

REGRESSION REPORTS

There are several regression models that were run, below will be an explanation of the models that were run, why they were run, and how good/bad each of the models are, but please see the appendix for all the models and the notes. Below will be generalized in a way that gives an overview, but will not get into the technicality of the numbers and breakdowns.

Regression Model with Original Data

This model will be the model that all the subsequent models will be compared to as being better or worse off and will influence the decision on what model should be used as a predictor of profit. On the base this model seems like a good model, but a majority of the variables are shown as being insignificant. This model wouldn't be the best model to use to predict profit.

Regression Model with Original Data using Stepwise

Because the Regression Model with all the Original data had some issues, it made more sense to run a new model that only had the variables that would be significant. While the model still seemed good, this model has only the significant variables, which would be better at predicting profit.

Regression Model with Cluster Analysis

Below will be a selection of 3 different clusters to compare whether all or any would be good indicators of profit. A cluster analysis was run for 3-15 clusters. Through the K Means cluster analysis it was shown that 15 clusters is the most optimal cluster choice. Below are the minimum, midpoint, and max (optimal cluster choice). Please see the appendix for the data.

3 Clusters

3 Clusters Only

This model was run to see if the minimum amount of clusters would come up with a significant model. This model overall seems like a not good model, but the variables are pretty significant. It does not seem like enough to have significant variables, if the model overall is okay; especially since the reverse was true in the first model that was run with just the original data. Even more so, after a comparison with this model vs the original data model that was run with stepwise is significant all around.

3 Cluster with Original Data

This model was run to see if the addition of the original data would make the 3 cluster regression model a good model with significant variables. The answer to that is sort of. It did improve the overall model significantly, but the same problem with the original model came along for the ride as well. Not all the variables were significant. Overall it was okay.

3 Cluster with Original Data and Stepwise

It is worth noting that a regression was run for this, but it will not be in the appendix because it was the exact same as the stepwise done on the regression model with the original data.

7 Clusters

7 Clusters Only

This model was run to see if the median amount of clusters would come up with a significant model. This model overall is worse than the 3 clusters regression model that was run. It does have more variables and a majority of them are significant with a couple being not too terribly insignificant. It is safe to say the original data stepwise regression model is a better model.

7 Cluster with Original Data

This model was run to see if the addition of the original data would make the 7 cluster regression model a good model with significant variables. The answer to that is the same as the 3 cluster with original data. The same issues that arose with the 3 cluster with original data regression model arose with this model as well.

7 Cluster with Original Data and Stepwise

Running this model should show a good model with significant variables. The difference between this model and the original data stepwise model is the addition of variables of Cluster 1 and the log of Sales. This model is good and is similar to the original data stepwise model.

15 Clusters

15 Clusters Only

This model was run to see if the optimal amount of clusters would come up with a significant model. This model overall is better than both the 3 cluster and 7 cluster models. It does have good amount of variables and a majority of them are significant with a few being insignificant. The model overall is okay, but has room for improvement. It is not better than the original data stepwise model.

15 Clusters Stepwise

Because of the number of variables that were insignificant, it makes sense to run a stepwise on the model. It seems that the only variable that was removed was cluster 12, removing any other variable would significantly make the model worse. This model is similar to the 15 clusters only model, so it is also an okay model

15 Cluster with Original Data

This model is to see if the combination of clusters and original data will show a more significant model. The same issue arises with this model as the 3 and 7 cluster with original data models. This model becomes significantly better with the added data, but the variables aren't that significant.

15 Cluster with Original Data and Stepwise

It was similar to the 7 cluster with original data and stepwise, with the exception of cluster 4 replacing cluster 1. The model is a good model overall.

SUMMARY

The clusters didn't really add anything to the models that would make any significant difference. If using clusters is an important piece to predict profit for Mr. Buffett, then using the variables from the 15 Cluster with Original Data and stepwise regression model was the best one out of all the regression models run for the clusters. In conclusion, to predict profit the best model to use is the one with the original data with stepwise.

APPENDIX

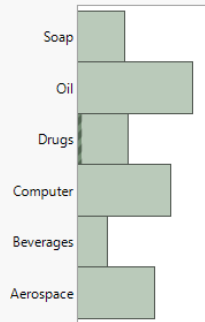
Distribution- Missing Data



Distribution with updated data

Distributions

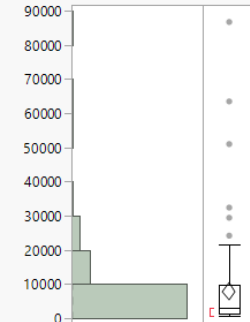
Type



Frequencies

Level	Count	Prob
Aerospace	18	0.18557
Beverages	7	0.07216
Computer	22	0.22680
Drugs	12	0.12371
Oil	27	0.27835
Soap	11	0.11340
Total	97	1.00000
N Missing	0	
6 Levels		

Sales(\$M)



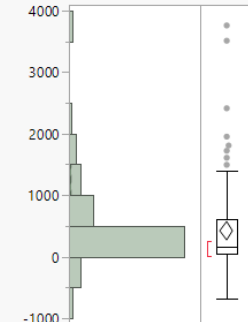
Quantiles

100.0%	maximum	86656
99.5%		86656
97.5%		57830.1
90.0%		19867.6
75.0%	quartile	9662.65
50.0%	median	3122
25.0%	quartile	1421.65
10.0%		815.8
2.5%		610.06
0.5%		576.9
0.0%	minimum	576.9

Summary Statistics

Mean	7940.7082
Std Dev	12784.684
Std Err Mean	1298.088
Upper 95% Mean	10517.393
Lower 95% Mean	5364.0239
N	97
N Missing	0

Profit(\$M)



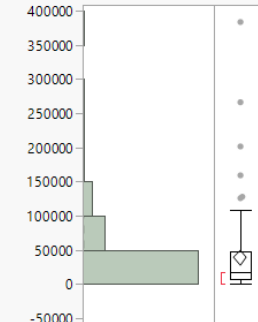
Quantiles

100.0%	maximum	3758
99.5%		3758
97.5%		3016.35
90.0%		1245.8
75.0%	quartile	604.55
50.0%	median	156
25.0%	quartile	43.05
10.0%		1.66
2.5%		-542.55
0.5%		-680.4
0.0%	minimum	-680.4

Summary Statistics

Mean	425.39907
Std Dev	710.03193
Std Err Mean	72.09282
Upper 95% Mean	568.50218
Lower 95% Mean	282.29595
N	97
N Missing	0

#emp



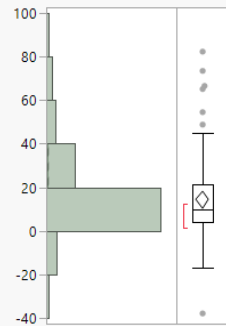
Quantiles

100.0%	maximum	383220
99.5%		383220
97.5%		236930
90.0%		102560
75.0%	quartile	48712
50.0%	median	18000
25.0%	quartile	7732.5
10.0%		3883.6
2.5%		1734
0.5%		560
0.0%	minimum	560

Summary Statistics

Mean	38685.854
Std Dev	56705.326
Std Err Mean	5757.5535
Upper 95% Mean	50114.508
Lower 95% Mean	27257.201
N	97
N Missing	0

Profits/emp



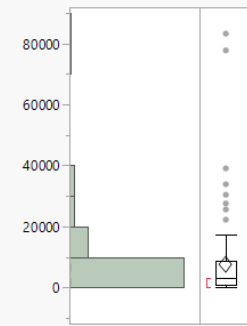
Quantiles

100.0%	maximum	82.242366412
99.5%		82.242366412
97.5%		70.320139509
90.0%		38.774539788
75.0%	quartile	21.348564382
50.0%	median	9.6728971963
25.0%	quartile	3.8880245055
10.0%		1.0093689005
2.5%		-16.70102339
0.5%		-37.8
0.0%	minimum	-37.8

Summary Statistics

Mean	14.414599
Std Dev	18.810574
Std Err Mean	1.9099244
Upper 95% Mean	18.205769
Lower 95% Mean	10.623429
N	97
N Missing	0

Assets(\$Mil)



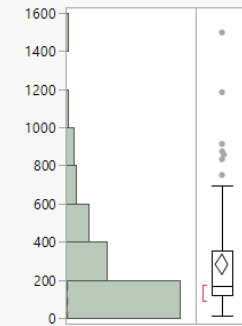
Quantiles

100.0%	maximum	83219
99.5%		83219
97.5%		60339.7
90.0%		16580.8
75.0%	quartile	8856.05
50.0%	median	3332
25.0%	quartile	1058.5
10.0%		474.74
2.5%		329.7
0.5%		325.2
0.0%	minimum	325.2

Summary Statistics

Mean	7708.4959
Std Dev	13025.745
Std Err Mean	1322.5641
Upper 95% Mean	10333.765
Lower 95% Mean	5083.227
N	97
N Missing	0

Sales/emp



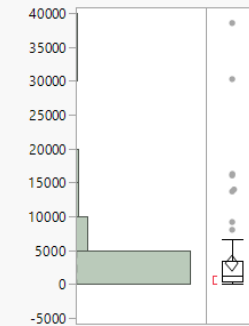
Quantiles

100.0%	maximum	1497.3333333
99.5%		1497.3333333
97.5%		1062.4476717
90.0%		642.8606886
75.0%	quartile	353.90463182
50.0%	median	166.98113208
25.0%	quartile	122.3775036
10.0%		103.57302594
2.5%		71.171381579
0.5%		13.328068044
0.0%	minimum	13.328068044

Summary Statistics

Mean	281.77118
Std Dev	257.1106
Std Err Mean	26.105627
Upper 95% Mean	333.59044
Lower 95% Mean	229.95192
N	97
N Missing	0

Stockholder's Eq(\$Mil.)



Quantiles

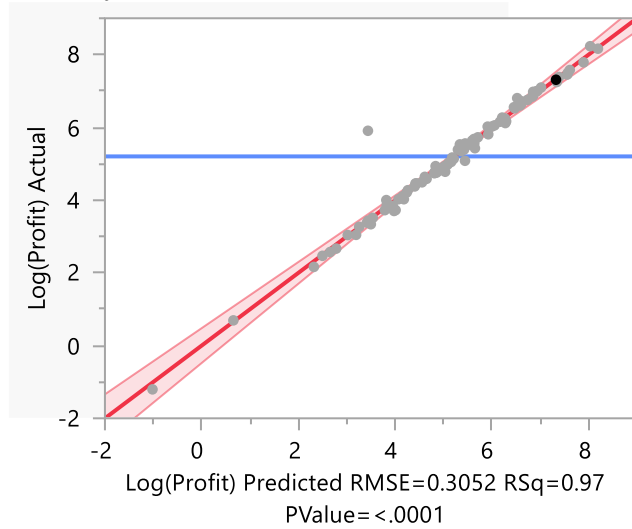
100.0%	maximum	38509
99.5%		38509
97.5%		23957.5
90.0%		6284.4
75.0%	quartile	3448.75
50.0%	median	1176
25.0%	quartile	459
10.0%		197.14
2.5%		122.87
0.5%		120.8
0.0%	minimum	120.8

Summary Statistics

Mean	3104.8894
Std Dev	5622.2514
Std Err Mean	570.85314
Upper 95% Mean	4238.0239
Lower 95% Mean	1971.7549
N	97
N Missing	0

Fit Model-Original Data

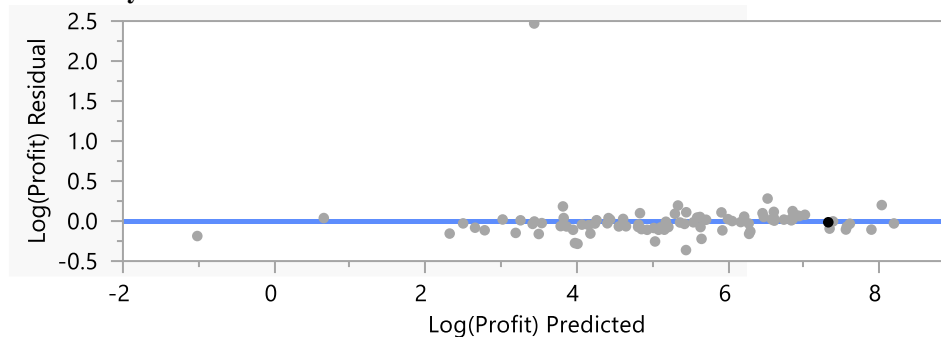
Response Log(Profit) Actual by Predicted Plot



Effect Summary

Source	LogWorth		PValue
log(Profits/emp)	43.633		0.00000
Log(#emp)	2.849		0.00141
Type_Aerospace	0.550		0.28204
log(Assets)	0.521		0.30126
log (Sales)	0.331		0.46653
Log(Stockholder's Eq)	0.325		0.47359
Type_Drugs	0.220		0.60223
Type_Beverages	0.168		0.67910
Type_Computer	0.097		0.80022
Type_Oil	0.093		0.80687
Log(Sales/emp)	0.003		0.99259

Residual by Predicted Plot



Summary of Fit

RSquare	0.970426
RSquare Adj	0.966201
Root Mean Square Error	0.30518

Mean of Response 5.201301
 Observations (or Sum Wgts) 89

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	-6.163997	1.822979	-3.38	0.0011*	.
Type_Aerospace	0.1448871	0.133741	1.08	0.2820	2.5204177
Type_Beverages	0.0661821	0.159374	0.42	0.6791	1.7589088
Type_Computer	0.032403	0.127604	0.25	0.8002	2.4044168
Type_Drugs	0.0748948	0.143104	0.52	0.6022	2.2828197
Type_Oil	0.0352866	0.143844	0.25	0.8069	4.0888277
log (Sales)	0.1713342	0.234135	0.73	0.4665	76.362558
Log(#emp)	0.8447658	0.255094	3.31	0.0014*	99.888586
log(Profits/emp)	1.0020081	0.033263	30.12	<.0001*	1.7296277
log(Assets)	-0.146409	0.14068	-1.04	0.3013	33.50661
Log(Sales/emp)	-0.002386	0.256169	-0.01	0.9926	31.923675
Log(Stockholder's Eq)	0.0740219	0.102782	0.72	0.4736	17.74692

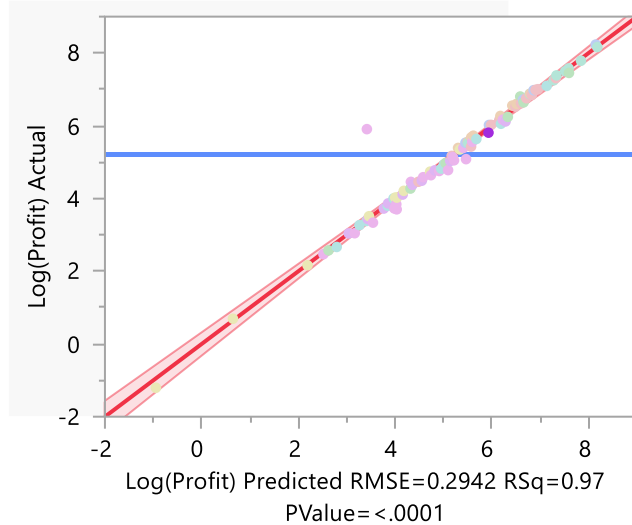
Note: While the R squared adjusted is at .97, RMSE seems low at .31, a majority of the variables do not seem to be significant, and the VIFs are high

Fit Model-Original Data-Stepwise

Fit Group

Response Log(Profit)

Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
log(Profits/emp)	52.222	0.00000
Log(#emp)	23.313	0.00000
log (Sales)	1.374	0.04223
Cluster_4	0.777	0.16696

Summary of Fit

RSquare 0.970016

RSquare Adj	0.968589
Root Mean Square Error	0.294203
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	235.21610	58.8040	679.3807
Error	84	7.27065	0.0866	Prob > F
C. Total	88	242.48675		<.0001*

Parameter Estimates

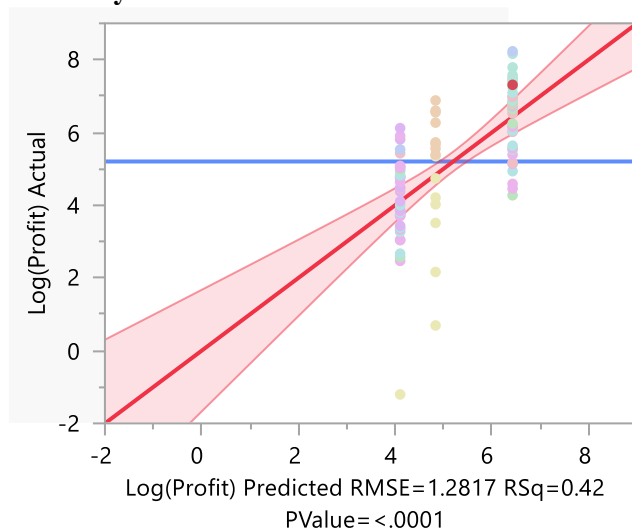
Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	-6.282415	0.267777	-23.46	<.0001*	.
Cluster_4	-0.168745	0.121042	-1.39	0.1670	1.9969944
log (Sales)	0.1429676	0.069312	2.06	0.0422*	7.2008964
Log(#emp)	0.8226749	0.057934	14.20	<.0001*	5.5437819
log(Profits/emp)	0.9946196	0.027533	36.12	<.0001*	1.275094

Note: Out of all 15 clusters, only cluster 4 was identified by stepwise as being significant to keep, the other variables have been repeated by stepwise. The model becomes more significant at .97 like all the stepwise models and the RMSE is at .29. The variables are significant and the VIFs are low.

Fit Model-3 Clusters

Response Log(Profit)

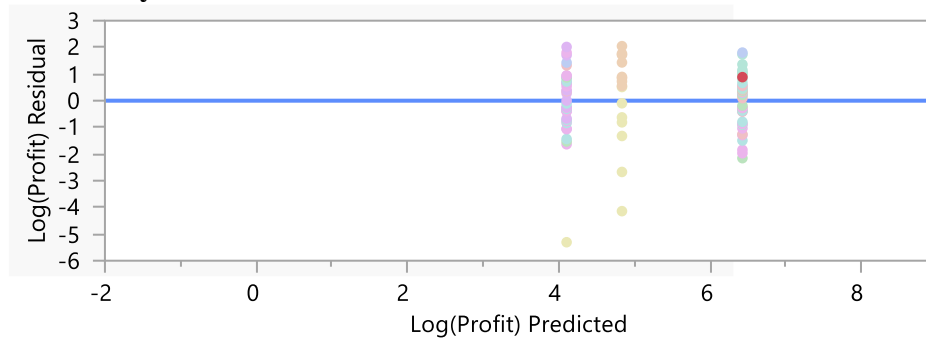
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
Cluster_2	4.112	0.00008
Cluster_1	1.217	0.06066

Residual by Predicted Plot



Summary of Fit

RSquare	0.41735
RSquare Adj	0.4038
Root Mean Square Error	1.281736
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	4.8353976	0.320434	15.09	<.0001*	.
Cluster_1	-0.732061	0.385114	-1.90	0.0607	1.9353933
Cluster_2	1.5924215	0.383509	4.15	<.0001*	1.9353933

Note: R square adjusted is substantially lower than both models that were run with the original data at .40 and a high RMSE at 1.28. The parameters are pretty significant and the VIF is low. Overall this model is not a good one

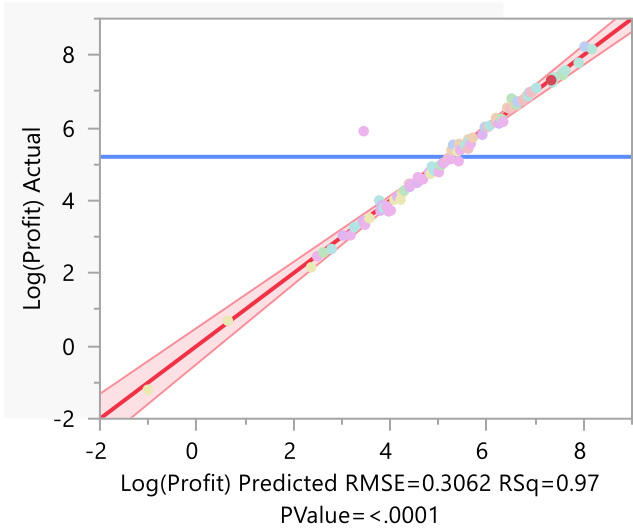
Fit Model-3 Clusters with Original Data

Response Log(Profit)

Singularity Details

Term	Details
Intercept	=Cluster_1 + Cluster_2 + Type_Aerospace

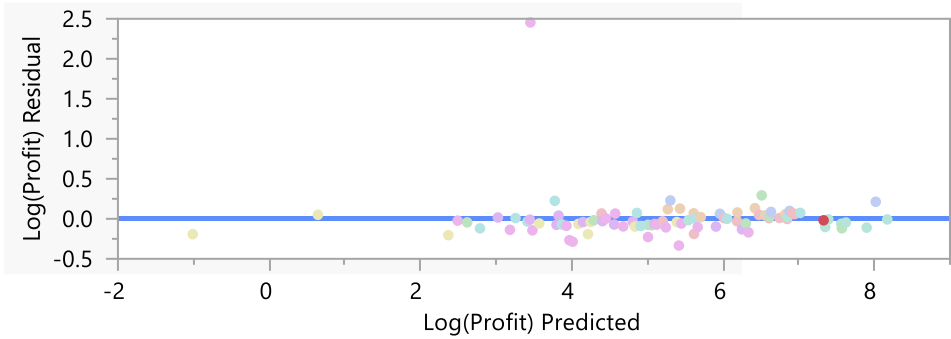
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
log(Profits/emp)	43.125	0.00000
Log(#emp)	2.599	0.00252
log(Assets)	0.556	0.27789
log (Sales)	0.366	0.43074
Log(Stockholder's Eq)	0.325	0.47304
Type_Computer	0.147	0.71364
Type_Drugs	0.138	0.72815
Type_Beverages	0.130	0.74123
Type_Oil	0.100	0.79448
Log(Sales/emp)	0.038	0.91589
Type_Aerospace	.	.
Cluster_2	.	.
Cluster_1	.	.

Residual by Predicted Plot



Summary of Fit

RSquare 0.970609

RSquare Adj	0.965968
Root Mean Square Error	0.306229
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	Biased	-5.624561	1.890943	-2.97	0.0039*	.
Cluster_1	Biased	-0.197786	0.154669	-1.28	0.2049	5.4689314
Cluster_2	Biased	-0.109067	0.143948	-0.76	0.4510	4.7767528
log (Sales)		0.186975	0.236037	0.79	0.4307	77.077651
Log(#emp)		0.8128571	0.260139	3.12	0.0025*	103.16825
log(Profits/emp)		1.0011574	0.0334	29.97	<.0001*	1.7320017
log(Assets)		-0.15486	0.141697	-1.09	0.2779	33.760309
Log(Sales/emp)		-0.027513	0.259631	-0.11	0.9159	32.568225
Log(Stockholder's Eq)		0.0743758	0.103137	0.72	0.4730	17.747362
Type_Aerospace	Zeroed	0	0	.	.	0
Type_Beverages		0.053362	0.161004	0.33	0.7412	1.7827919
Type_Computer		0.0478863	0.130006	0.37	0.7136	2.4787023
Type_Drugs		0.0514843	0.147573	0.35	0.7281	2.4110201
Type_Oil		0.0377448	0.144383	0.26	0.7945	4.0913335

Note: The model is similar to the first model that was run with the original data with R squared adjusted and RMSE at .97 and .31 respectively. A majority of the variables are not significant and the some of the VIFs are pretty high.

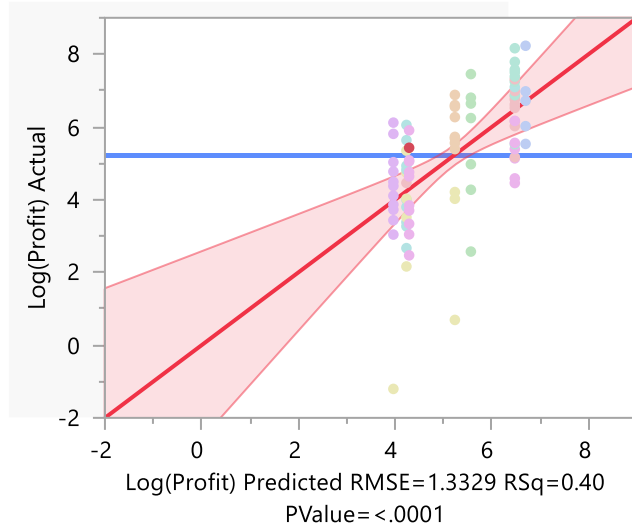
When running the stepwise, the results were the same as the stepwise ran on the original data.

Fit Model-7 Clusters


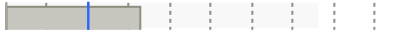
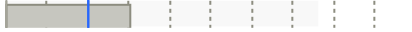
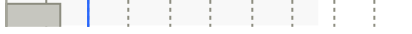
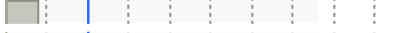
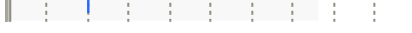
Response Log(Profit)

Whole Model

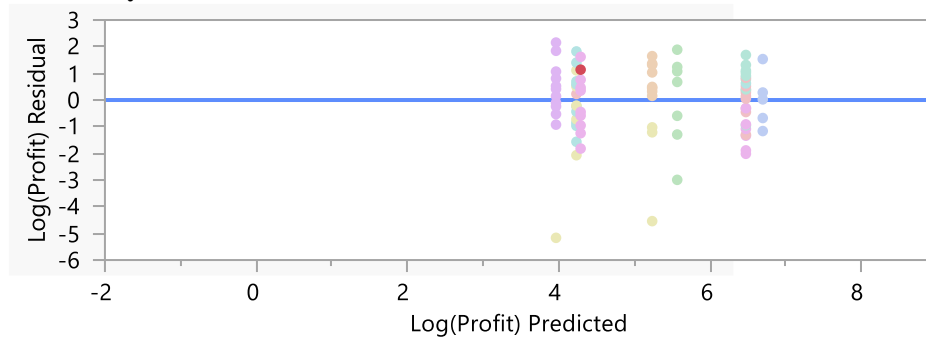
Actual by Predicted Plot



Effect Summary

Source	LogWorth		PValue
Cluster_6	3.636		0.00023
Cluster_5	3.275		0.00053
Cluster_2	3.027		0.00094
Cluster_1	1.344		0.04525
Cluster_4	0.821		0.15117
Cluster_3	0.135		0.73354

Residual by Predicted Plot



Summary of Fit

RSquare	0.399182
RSquare Adj	0.35522
Root Mean Square Error	1.332933
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	6	96.79645	16.1327	9.0801
Error	82	145.69030	1.7767	Prob > F
C. Total	88	242.48675		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	6.696665	0.596106	11.23	<.0001*	.
Cluster_1	-1.461859	0.718931	-2.03	0.0452*	2.8044944
Cluster_2	-2.407721	0.701436	-3.43	0.0009*	3.0741573
Cluster_3	-0.223057	0.653001	-0.34	0.7335	4.3146067
Cluster_4	-1.130871	0.780486	-1.45	0.1512	2.211236
Cluster_5	-2.463429	0.682925	-3.61	0.0005*	3.4449438
Cluster_6	-2.733268	0.709508	-3.85	0.0002*	2.941573

Note: R square adjusted is substantially lower than both models that were run with the original data at .36 and a high RMSE at 1.33. The parameters are pretty significant and the VIF is low. Overall this model is not a good one

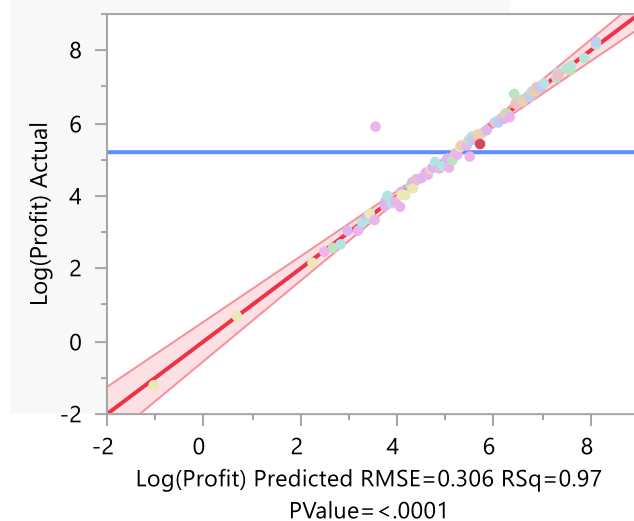
Fit Model-7 Clusters with Original Data

Response Log(Profit)

Singularity Details

Term	Details
Cluster_4	=Type_Beverages=Intercept - Cluster_1 - Cluster_2 - Cluster_3 - Cluster_5 - Type_Computer

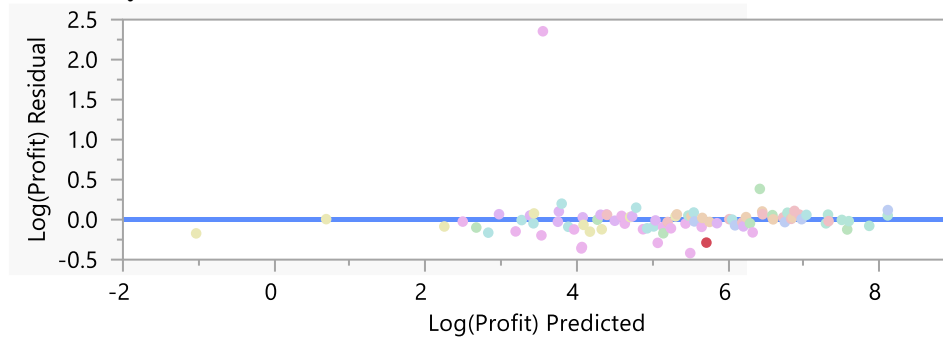
Actual by Predicted Plot



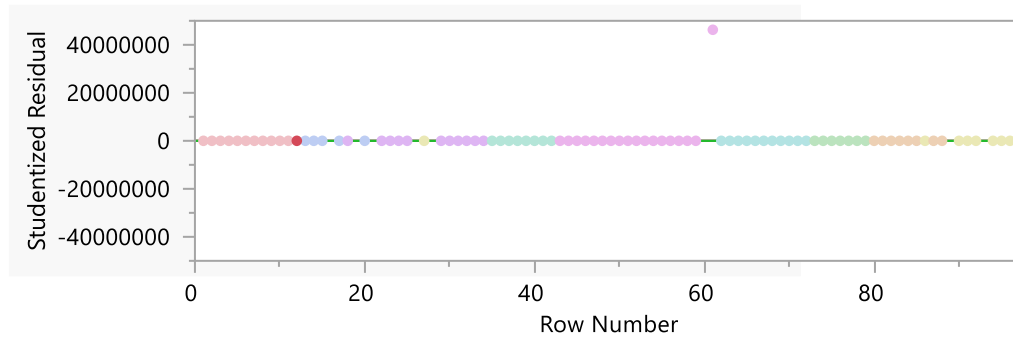
Effect Summary

Source	LogWorth	PValue
log(Profits/emp)	41.514	0.00000
Log(#emp)	2.889	0.00129
Cluster_6	0.971	0.10697
log(Assets)	0.380	0.41708
log (Sales)	0.201	0.62996
Type_Aerospace	0.169	0.67826
Log(Stockholder's Eq)	0.127	0.74593
Log(Sales/emp)	0.101	0.79247
Type_Oil	0.087	0.81902
Type_Drugs	0.055	0.88185
Type_Computer	.	.
Type_Beverages	.	.
Cluster_5	.	.
Cluster_4	.	.
Cluster_3	.	.
Cluster_2	.	.
Cluster_1	.	.

Residual by Predicted Plot



Studentized Residuals



Externally studentized residuals with 95% simultaneous limits (Bonferroni) in red, individual limits in green.

Summary of Fit

RSquare	0.971804
RSquare Adj	0.96601
Root Mean Square Error	0.306039
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	Biased	-5.95716	1.792743	-3.32	0.0014*	.
Cluster_1	Biased	-0.092267	0.243268	-0.38	0.7056	6.0913779
Cluster_2	Biased	-0.230469	0.325701	-0.71	0.4814	12.573384
Cluster_3	Biased	-0.188701	0.275655	-0.68	0.4958	14.585086
Cluster_4	Biased	-0.187544	0.194503	-0.96	0.3381	2.6050687
Cluster_5	Biased	-0.315719	0.215237	-1.47	0.1467	6.4913374
Cluster_6		-0.335692	0.205685	-1.63	0.1070	4.6895696
Type_Aerospace		0.0718636	0.172538	0.42	0.6783	4.1712889
Type_Beverages	Zeroed	0	0	.	.	0
Type_Computer	Zeroed	0	0	.	.	0
Type_Drugs		0.0354317	0.237557	0.15	0.8818	6.2555469
Type_Oil		-0.059962	0.261123	-0.23	0.8190	13.398639
log (Sales)		0.116926	0.241673	0.48	0.6300	80.90299
Log(#emp)		0.8639228	0.258072	3.35	0.0013*	101.66109
log(Profits/emp)		1.0043239	0.034118	29.44	<.0001*	1.8094564
log(Assets)		-0.119391	0.146289	-0.82	0.4171	36.02876

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Log(Sales/emp)	0.0694097	0.26284	0.26	0.7925	33.419553
Log(Stockholder's Eq)	0.0347333	0.106794	0.33	0.7459	19.051994

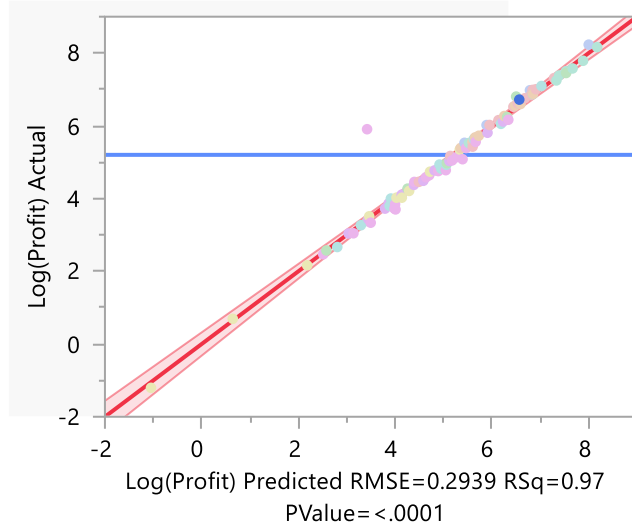
Note: The model is similar to the first model that was run with the original data with R squared adjusted and RMSE at .97 and .31 respectively. A majority of the variables are not significant and the some of the VIFs are pretty high.

Fit Model-7 Clusters with Original Data-Stepwise

Fit Group

Response Log(Profit)

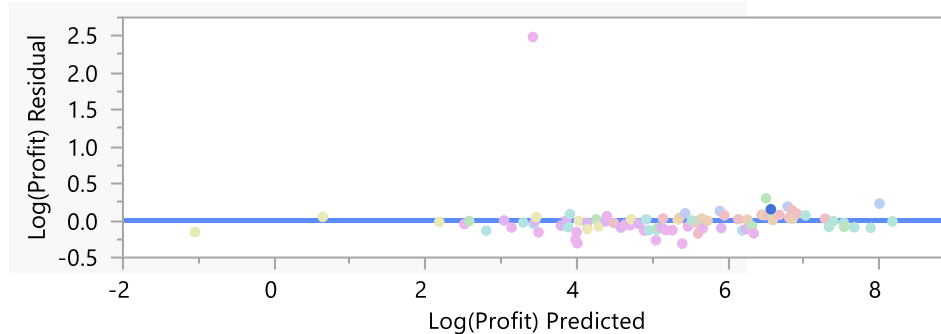
Actual by Predicted Plot



Effect Summary

Source	LogWorth	PValue
log(Profits/emp)	50.890	0.00000
Log(#emp)	27.437	0.00000
log (Sales)	0.941	0.11463
Cluster_1	0.819	0.15167

Residual by Predicted Plot



Summary of Fit

RSquare	0.970068
RSquare Adj	0.968643
Root Mean Square Error	0.293947
Mean of Response	5.201301
Observations (or Sum Wgts)	89

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	235.22875	58.8072	680.6010
Error	84	7.25800	0.0864	Prob > F
C. Total	88	242.48675		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	-6.090009	0.27073	-22.49	<.0001*	.
Cluster_1	0.1648297	0.113926	1.45	0.1517	1.4481218
log (Sales)	0.0824183	0.051697	1.59	0.1146	4.012825
Log(#emp)	0.8455546	0.051102	16.55	<.0001*	4.3207764
log(Profits/emp)	1.0115231	0.029112	34.75	<.0001*	1.4280857

Note: R squared Adjusted is .97 like all the other good models that were run. This has a low RMSE at .29. The variables are significant and the VIFs are under 5