Application of Logistic Regression Models in Marketing

ISBN 978-1-943295-24-1

Shaili Singh
Vinod Gupta School of Management, IIT Kharagpur
(shaili@vgsom.iitkharagpur.ac.in)
Aditya Rao
Birla Institute of Technology and Science Pilani
(akraomlr@gmail.com)

Emergence of data science and machine learning have transformed the way marketers today understand client behavior, forecast trends, and optimize marketing campaigns. Among numerous quantitative modelling techniques logistic regression has emerged as a highly valuable analytical approach for addressing modeling and discrimination challenges within the field of marketing. Using an open-source sample dataset from Kaggle this study builds a logistic regression model to train and test for a loyalty program with the objective of optimizing profitability along with market penetration. The study explains the use cases of logistic regression for providing control over targets, measuring market response, and optimizing market spend.

Keywords: Logistic Regression, Decile Analysis, Market Research, Marketing Budget, Quantitative Modeling

1. Background

Data science has truly transformed marketing by allowing firms to extract valuable insights and make data-driven decisions (Boegershausene et.al., 2022). Marketers have traditionally made marketing decisions based on intuition, experience, and limited data according to Gopal Krishnan (2018). However, with the emergence of data science and machine learning, marketers today have access to massive volumes of data as well as advanced analytical approaches to understand customers and markets. A challenging topic for marketing researchers to study is the quantitative modeling of marketing concepts like consumption patterns, consumer behavior, and other behavioral topics (Tayur et. al., 2012). The analysis and prediction of a dichotomous (male/female, yes/no, user/nonuser, satisfied/unsatisfied, etc.) outcome is required for many marketing research problems. Examples of these include whether a product will succeed in the market, as evaluated by Garber (2004) which shows that managers can assess a newly launched product's success rate in the market using the spatial dimension of sales data, whether a consumer is likely to engage in risky buying behaviors as analyzed by DeSarbo and Edwards (1996) which examines the nature of consumers based on the level of compulsive buying as a function of several personality traits and psychological effects, etc.

Upon examining the many types and frequency of utilization of multivariate analytic applications, it becomes evident that regression is often regarded as one of the most commonly employed techniques. According to Hyman and Young (2001), regression is the prevailing multivariate method employed in international marketing serials.

In order to explain potential correlations between variables, biological, behavioral, and social sciences frequently employ linear regression. In economics, linear regression is the predominant empirical tool to draw conclusions. It is used to predict important variables like consumption spending (Deaton, 1992), labor demand and supply (Ehrenberg, 2008). In finance, one of the most important theories of the capital asset pricing model (French, Craig, 2003) is a model used to theoretically calculate the required return rates of any asset in order to create a diversified portfolio using regression. In their literature assessment on export performance, Zou and Stan (1998) found that regression analysis was the predominant analytical method employed by researchers. However, they emphasized the necessity of incorporating more advanced methodologies to enhance conceptual and theoretical progress in this field.

In the present environment, logistic regression emerges as a highly valuable analytical approach for addressing modeling and discrimination challenges within the field of marketing. Due to its incorporation of different data distribution assumptions, this approach yields more suitable and accurate results in relation to model fit and the validity of the study. However, despite its longstanding usage in statistical research, the marketing literature has not given much attention to this technique in comparison to other applications of regression. Two of the most commonly used statistical methods for categorizing outcome variables include Linear Discriminant Analysis and Logistic Regression (Pohar et. al., 2004). When the explanatory variables are not normally distributed, Linear Discriminant Analysis assumptions fail and Logistic Regression has to be used, which provides a decent fit to many types of distribution when the sample size is not small, which makes logistic regression a common choice in marketing, health research, and other sectors.

Logistic regression is a type of regression analysis that is employed when the dependent variable is binary or dichotomous, while the independent variables can be either continuous, categorical, or a combination of both (Stoltzfus, 2011). Multinomial logistic regression is a statistical technique designed to address scenarios when the dependent variable consists of multiple classes. This phenomenon is commonly known as the multivariate case.

The majority of multivariate analytic methods, which involve independent and/or dependent variables, depend on the fundamental premises of continuous data and normality (Sauerbrei et. al., 2020). This requirement is evident throughout the research's data collection and measurement processes.

Additional benefits of the analysis that contribute to its increasing popularity are derived from the assumptions that logistic regression does not assume a linear relationship between the dependent and independent variables, the independent variables are not required to be of an interval scale, logistic regression does not impose any restrictions on the range of the independent variables and the assumption of normally distributed error terms is not necessary for logistic regression as stated by Stoltzfus (2011).

Using an open-source sample dataset from Kaggle this study builds a logistic regression model to train and test for a loyalty program with the objective of optimizing profitability along with market penetration. The study further explains the use cases of logistic regression for providing control over targets, measuring market response, and optimizing market spend.

2. Literature Review

In recent studies, one area of significant research and development in the marketing domain has been customer relationship management (CRM) (Liu and Ng, 2000). Today, CRM is essential due to the increasing degree of competition among businesses. Furthermore, one should consider the swift shifts in the needs and expectations of clients. CRM is the primary tool used by organizations to take on these difficulties since it may assist them in understanding the various expectations of their clients and, as a result, gain an advantage over their competitors (Anderson, 2002).

Typically, one of the most crucial elements of a successful CRM strategy has been the optimal selection of client targets. To determine the largest number of customers who will purchase a particular product or maintain their business relationship with the company, numerous studies and models have been put forth. Consequently, businesses strive to create predictive models that precisely pinpoint the clients who are most likely to make a purchase.

These models are frequently referred to as a customer prediction system (CPS). The designer of a CPS typically has access to a very large number of features. The design of systems with a relatively small feature set has been incentivized and implemented within the customer recognition community since each feature variable might lead to an increase in the system's cost and operating time. Simultaneously, a sufficient number of characteristics must be included in the model to obtain high recognition rates.

In database marketing, Kim and Street (2004) introduced a novel method for consumer targeting. The potential feature combinations are searched through using a typical Genetic algorithm (GA). Neural networks are then trained using the input features that the genetic algorithm chose. Ahn et. al. (2007) proposed a case-based reasoning method for customer classification that uses the two-dimensional reduction technique to forecast customers' purchasing behavior for a certain product based on demographic variables. Buckinx et al. (2007) used a multivariate linear regression model for feature selection to avoid overfitting to predict customer loyalty. Another simple and straightforward process that starts with an empty set of variables and iteratively includes variables greedily whilst steadily improving model performance for feature selection was used by Kim (2006). Lessmann and Voß (2009) developed a hierarchical reference model for support vector machine (SVM) based classification in real-world CRM applications, with recursive feature elimination proposed as an iterative backward-elimination process for SVM-based feature ranking.

Among the feature selection techniques that have been presented, multiple linear regression models and case-based reasoning do not take nonlinear relations in customer behavior into account. SVMs also face the limitations of kernel selection, considerably higher operation time, and have a low probability of selecting the number of optimal features. Renders and Flasse (1996) claim that because of their shortcomings in using local information and their tendency to employ less a priori knowledge, GAs may experience unduly slow convergence before arriving at an accurate conclusion.

Creating an effective CPS for CRM programs is the main objective of this study. After reviewing the literature, we found that articles on CPSs that are currently available typically take a very complicated and time-consuming approach because it is never easy to satisfy the demands of both high computational speed and good performance. Based on the assessed literature, we found that contrary to our hypothesis, a combination of multiple logistic regression (MLR) and decile segregation approach was not used for the prediction of customer purchasing behavior. Based on this suggested approach, firms can segregate members of their target market according to how likely they are to convert from lead to sale. In that scenario, they may focus exclusively on the most likely customers and avoid wasting the marketing budget on those who had minimal probability of becoming customers. This is the exact way that analytics-enablement can assist any marketing department in maximizing income for the company through scaled-up, wise financial allocation.

3. Objectives

Analytics-enablement analysis can help businesses maximize income by allocating funds wisely. Marketing budgets are set by businesses based on gross sales and marketing proportions. Machine learning can assist marketers in targeting the most likely purchasers while saving money on low-likelihood buyers.

Any firm's marketing budget is determined by the formula: Marketing Budget = Gross Revenue x Marketing%

where Marketing% is defined as the portion of gross revenue that the company decides to invest back into marketing at the end of a sales cycle.

So, the objective of marketing-based analytics is to increase gross revenue for the firm to increase the marketing budget for the next cycle. Given the limited marketing budget, marketers can only reach out to a portion of the total target audience, but segregating the target audience on the basis of their buying probability will result in huge savings that would have been used on less probable buyers.

4. Methodology

The analysis for the marketing campaign to account for profitability and market penetration using logistic regression will be done as follows

Using an open-source sample dataset from Kaggle with around 2,20,000 observations having 11 input variables, we would build a logistic regression model using a test set of smaller data size of 10% of the total sample set to predict the probability of buying for the remaining 90% of customers for a loyalty program with the objective of optimizing profitability along with market penetration. The purchase decisions for the 10% of the participants are already known and our objective is to run the marketing campaign only on the most probable buyers among the remaining 90% of customers, keeping in mind the limited marketing budget. The variables are divided into two categories based on the demographic properties and their loyalty properties as shown in Table 1.

Table 1 *The Variables and their Definitions used in the Data*

Variable Name	Variable Description	Significance in model
ID	Customer unique ID	Identity
DemAffl	Demographic affluence on a scale of 1 to 30	Input Variable
Dem Age	Demographic age in years	Input Variable
Dem Cluster Group	Demographic Neighbourhood group (A-F)	Input Variable
DemGender	Demographic Gender (Male (M), Female (F) or Unknown (U))	Input Variable
Dem Reg	Demographic Geographic Region	Input Variable
Dem TV Reg	Demographic Television Region	Input Variable
Loyal Class	Loyalty status (Platinum, Gold, Silver, Tin)	Input Variable
Loyal Spend	Total amount spent	Input Variable
Loyal Time	Total time as a loyalty card user	Input Variable
Target Buy	1 if Customer purchases the product, else 0	Target Variable

We will be using logistic regression, a classification algorithm giving binary outputs due to its various advantages. Based on this model, we will then use the Decile methodology to find out the right portion of these customers that will be the right target for the campaign to focus on, based on the proposed budget and profitability targets. A brief flowchart of the methodology is shown in Figure 1.

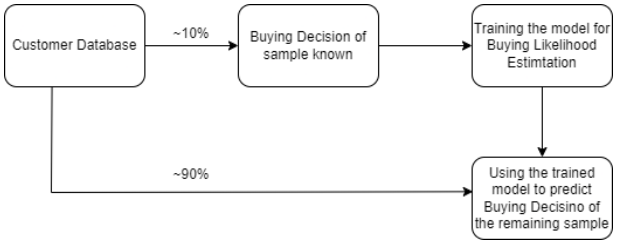

Figure 1 *The Methodology used to Predict the Buying Decisions*

For at least the last 25 years, direct marketing practitioners have been assessing model performance using various measures. The most common metric used is the decile analysis. Each customer's value measure is computed, and the customer base is then sorted by value in descending order, and divided into ten equal parts. The highest 10% of the base is referred to as the first or top decile. The next 10% is in the second decile, and so forth. Judy Bayer (2010) analyzed the generally accepted segments in the telecommunications sector using the decile approach with the end goal of customer retention, efficient business planning, and growth. Another study by Geng Cui et. al. (2015) studied forecasting models considering the resource constraint for profit maximization and proposed a model of partial order-constrained optimization (POCO) with the help of the genetic algorithm using a decile analysis.

By focusing on the decile methodology and its application to logistic regression models in marketing campaigns, this thesis project aims to provide valuable insights into the predictive power of these models and their potential impact on revenue allocation and marketing budget optimization focusing on market penetration and profit maximization using real datasets.

5. Building the Model

Google Colaboratory, or Google Colab for short, is the integrated development environment (IDE) utilized for the procedure. Google Colab is a free Jupyter Notebook environment that operates fully in the cloud and does not require any setup. Google Colab allows us to store and share our analysis, build and run code, and gain free access to powerful computational resources

through the browser. Moreover, it has the ability to swiftly download large datasets straight from servers to Google Drive. After authenticating, we mount the drive folder using Google Colab, and it retrieves the necessary files.

After importing the required dependencies which include Pandas and NumPy and the training sample with a known buying decision, which accounts for around 10% of the total data, we check for missing values. We can notice that in the sample dataset, all the seven columns except LoyalClass and LoyalSpend, have missing values which need to be imputed to avoid any discrepancy in the model. According to Aljuiaid and Sasi (2016), if less than 5% of the data is missing, mean and mode imputation can be performed since it does not involve other time-taking processes like K-Nearest Neighbours, Hot-Deck, Expectation Maximization, and C5.0 and is relatively easy to perform. Following this, for all continuous variables in the dataset, we have used mean imputation and for all discrete variables, we have used mode imputation.

We then need to convert these discrete variables to a numeric representation and one of the most common techniques is Label Encoding, which comes under Determined techniques. In determined techniques, the encoded values will always stay the same throughout the dataset.

According to Hancock et. al. (2020), simply put, label encoding gives every possible value of a categorical variable an integer value. A categorical variable's encoded counterparts are defined by our arbitrary assignment of integer values to it once we are aware of all of its possible values. Since in the dataset, the possible values of the categorical variables are limited, this technique works well.

According to Buitinck et. al (2013), the Python Scikit-learn toolkit includes an ordinal encoder which can be found in the Category Encoders module, which has been used here. The resulting label encoding which has been used is shown in Figure 2.

```
{'A': 0, 'B': 1, 'C': 2, 'D': 3, 'E': 4, 'F': 5, 'U': 6}
{'F': 0, 'M': 1, 'U': 2}
{'Midlands': 0, 'North': 1, 'Scottish': 2, 'South East': 3, 'South West': 4}
{'Border': 0, 'C Scotland': 1, 'East': 2, 'London': 3, 'Midlands': 4, 'N East': 5, 'N Scot': 6, 'N West': 7, 'S & S East': 8, 'S West':
{'Gold': 0, 'Platinum': 1, 'Silver': 2, 'Tin': 3}
```

Figure 2 *The Label Encoding Scheme used for the Discrete Variables*

We then check the data for multicollinearity. A logistic regression model with highly associated predictor variables is said to exhibit multicollinearity. The variance inflation factor (VIF), which is the reciprocal of tolerance, is the primary tool used to identify multicollinearity, according to Sarkar et. al. (2013). Although there isn't a set cutoff value for multicollinearity, values of VIF more than 10 are frequently taken to be indicative of multicollinearity, as per Allison (2001) and Marquardt (1980). The VIF values calculated for all the model variables are shown in Table 2 and almost all of them are under the cutoff of 10, which indicates that the degree of correlation among the variables is tolerable.

Table 2 *VIF Values for the Model Variables*

Variables	VIF
Dem Affl	6.278630
Dem Age	10.734656
Dem Cluster Group	3.659632
Dem Gender	1.435472
Dem Reg	2.474645
Dem TV Reg	3.752279
Loyal Class	3.851766
Loyal Spend	1.863196
Loyal Time	3.153032

Following this, we perform logistic regression and store the model in Google Drive to predict probabilities for the buying decisions of approximately 90% of the remaining dataset. Owing to resource constraints on Google Collab, we have further divided the 90% dataset into an 80:20 split and applied the model on the 20%, which resulted in 4445 observations in total. This can easily be extended to the remaining dataset to further analyze the model. To test the accuracy of the model generated, we simply calculated the Accuracy score, which compares the values of the predicted and actual buying decision in the model, which comes out to be 80.56% and shows that this model fits the data well. This implies that this logistic regression model can then be applied to the remaining 90% of the data to predict the customer's buying decisions.

	precision	recall	f1-score	support
0	0.82	0.95	0.88	3367
1	0.69	0.36	0.48	1078
accuracy			0.81	4445
macro avg	0.76	0.66	0.68	4445
weighted avg	0.79	0.81	0.78	4445

Figure 3 *Classification Report Statistics*

Further calculating the classification report statistics, we noticed the following values as shown in Figure 3. Here Class 1 represents the number of customers who decided to buy and Class 0 represents the number of customers who decided to not buy. The precision refers to how often the model predicted the event to be positive and it turned out to be true. The recall is the measure of our model correctly identifying True Positives, and the f1-score is the harmonic mean of the precision and recall statistic. A high f1-score for the classification of Class 0 demonstrates that the model can simultaneously achieve high precision and high recall.

6. Decile Analysis and Results

After running the derived model on the test data, which includes 90% of the sample, we have obtained the probability of buying for each customer which can be used for further analysis by the firms.

Applying the decile methodology to this test data, we first sort the probabilities of buying for all the customers in the decreasing likelihood of buying to create the deciles. Since we have 4445 observations here in total, we further create decile sets of around 445 each. This also helps us find the most probable buyers for each decile along with the probability threshold for each decile.

The following table is the analysis of each decile.

Table 3 *Decile Analysis*

Decile	Total	Prob Threshold	Good	% Good	Bad	Cumm Good	Cumm Bad	% Cumm Good	% Cumm Bad	% Cumm Bad Avoided
1	445	54.84%	330	74%	115	330	115	31%	3%	97%
2	445	40.07%	216	49%	229	546	344	51%	10%	90%
3	445	31.10%	128	29%	317	674	661	63%	20%	80%
4	445	24.34%	107	24%	338	781	999	72%	30%	70%
5	445	19.19%	75	17%	370	856	1369	79%	41%	59%
6	445	14.92%	79	18%	366	935	1735	87%	52%	48%
7	445	11.11%	50	11%	395	985	2130	91%	63%	37%
8	445	8.06%	40	9%	405	1025	2535	95%	75%	25%
9	445	4.86%	34	8%	411	1059	2946	98%	87%	13%
10	440	0.44%	19	4%	421	1078	3367	100%	100%	0%

The *Total* column above is the number of observations per decile. As discussed earlier, the *Prob Threshold* column is the value of the last observation in each decile. The *Good* column indicates the count of actual outcomes in the decile who decided to buy. The *Bad* column indicates the number of customers who did not actually buy the product, which is *Total – Good. Cumm Good* and *Cumm Bad* indicate the cumulative sum of *Good* and *Buy* respectively, and *% Cumm Bad Avoided* is essentially 1 - *% Cumm Bad*. Based on the Decile Analysis, we now have a clear read on the customers' behavior and the firm's strategy going forward based on the profitability and market penetration targets. For example, if the firm is opting to be very conservative in its approach, it would make sense to only target the top decile, which would essentially mean that we will be targeting 31% of the customers who are likely to buy. In this approach, 97% of bad customers who are unlikely to buy will be avoided. This also obviously means that we will be missing out on the remaining 69%, which could be considered as the cost paid for being more conservative. Including more deciles in the approach would obviously increase the proportion of customers likely to buy but also expose it to more bad customers who are unlikely to buy, thus increasing market penetration at the cost of profitability, and this trade-off can be decided by the firm's current financial conditions.

7. Conclusions and Future Scope of Work

Several business decisions and conclusions can be drawn from the aforementioned research:

Control Over Targets: In certain cases, the client has a business requirement that a minimum number of clients be consistently pursued. In these situations, we can have more precise control over extra records and bigger targets by taking into account the top 3 deciles rather than just the top 2.

Measuring Market Response: Measuring the market response and the post-recommendation analysis are both simple tasks. We may evaluate the success of the current marketing campaign by simply adjusting the probability threshold and counting the number of good clients.

Optimizing Marketing Spend: Companies can save time, money, and resources by concentrating on the top deciles rather than wasting them on non-responders or the wrong kind of clientele.

We focused on decile analysis in this work, but by including intervals smaller than 10%, such as 5% or 1%, decile analysis can be made even more comprehensive. Although this generalization to smaller intervals can yield business advice at even lower levels, smaller intervals demand greater computational and data processing power. Complex or non-linear data cannot be accurately analyzed by Logistic Regression, which generates linear models. Instead, sophisticated machine learning methods such as Neural Networks can be used to handle such data.

8. References

1. Hyman, M.R. and Young, Z. (2001), "International marketing serials: a retrospective", International Marketing Review, Vol. 18 No. 6, pp. 667-716.
2. Zou, S. and Stan, S. (1998), "The determinants of export performance: a review of the empirical literature between 1987 and 1997", International Marketing Review, Vol. 15 No. 5, pp. 333-56.
3. Gopal Krishnan (2018), "Marketers, Big Data and Intuition – Implications for Strategy and Decision-Making".
4. Boegershausen, J., Datta, H., Borah, A., & Stephen, A. T. (2022). Fields of Gold: Scraping Web Data for Marketing Insights. In Journal of Marketing (Vol. 86, Issue 5, pp. 1–20). SAGE Publications. https://doi.org/10.1177/00222429221100750
5. Tayur, S., Ganeshan, R., & Magazine, M. (Eds.). (2012). *Quantitative models for supply chain management* (Vol. 17). Springer Science & Business Media.
6. Garber, T., Goldenberg, J., Libai, B., & Muller, E. (2004). From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, *23*(3), 419-428.
7. DeSarbo, W. S., & Edwards, E. A. (1996). Typologies of compulsive buying behavior: A constrained cluster wise regression approach. *Journal of consumer psychology*, *5*(3), 231-262.
8. Hughs, A. (1996) The Complete Database Marketer. New York: McGraw Hill.
9. Bayer, J. (2010). Customer segmentation in the telecommunications industry. *Journal of Database marketing & customer strategy management*, *17*, 247-256.
10. Cui, G., Wong, M. L., & Wan, X. (2015). Targeting high value customers while under resource constraint: partial order constrained optimization with genetic algorithm. *Journal of Interactive Marketing*, *29*(1), 27-37.
11. Stoltzfus, J. C. (2011). Logistic regression: a brief primer. *Academic emergency medicine*, *18*(10), 1099-1104.
12. Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., ... & TG2 of the STRATOS initiative Michal Abrahamowicz Heiko Becher Harald Binder Daniela Dunkler Frank Harrell Georg Heinze Aris Perperoglou Geraldine Rauch Patrick Royston Willi Sauerbrei. (2020). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and prognostic research*, *4*, 1-18.
13. Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, *1*(1), 143.
14. Deaton, Angus (1992). *Understanding Consumption*. Oxford University Press.
15. Ehrenberg; Smith (2008). *Modern Labor Economics* (10th international ed.). London: Addison-Wesley.
16. French, Craig W. (2003). "The Treynor Capital Asset Pricing Model". *Journal of Investment Management*. **1** (2): 60–72.
17. K.S. Ng and H. Liu (2000). Customer retention via data mining, *Artif Intell Rev* 14(6) 569–590.
18. E.T. Anderson (2002). Sharing the wealth: when should firms treat customers as partners? *Manage Sci* 48(8) 955–971.
19. Y.S.Kim (2006). Toward a successful CRM: variable selection, sampling, and ensemble, *Decis Support Syst* 41 542–553.
20. Y.S. Kim and W.N. Street (2004). An intelligent system for customer targeting: a data mining approach, *Decis Support Syst* 215–228.
21. H. Ahn, K. Kim and I. Han (2007). A case-based reasoning system with the two-dimensional reduction technique for customer classification, *Expert Syst Appl* 32.
22. W. Buckinx, G. Verstraeten and D.V. Poel (2007). Predicting customer loyalty using the internal transactional database.
23. Y.S.Kim, (2006). Toward a successful CRM: variable selection, sampling, and ensemble.
24. S. Lessmann and S. Voß (2009). A reference model for customer-centric data mining with support vector machines.
25. J.M. Renders and S.P. Flasse, (1996). Hybrid methods using genetic algorithms for global optimization.
26. Aljuaid, T., and Sasi, S. (2016). Proper imputation techniques for missing values in data sets. In 2016 International Conference on Data Science and Engineering (ICDSE) (pp. 1-5). IEEE.
27. Hancock, J. T., and Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 1-41.
28. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, et al. (2013). API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD workshop: languages for data mining and machine learning. pp. 108–22.
29. Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of interdisciplinary mathematics*, *13*(3), 253-267.
30. P. D. Allison (2001). Logistic Regression Using the SAS system: Theory and Applications, Cary, NC: SAS Institute Inc.
31. Marquardt, D. W. (1980). You should standardize the predictor variables in your regression models. Journal of the American Statistical Association 75: 74–103.