Deloitte.

Comparison of Linear and Logistic Regressions for Segmentation

Debashish Banerjee, Director, Deloitte
Kranthi Ram Nekkalapu, Senior. Consultant, Deloitte

March 15, 2016

Anti-Trust Notice

- The Casualty Actuarial Society (CAS) is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.

- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.

- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Introduction

Insurers often perform segmentation to divide profiles into subsets with common characteristics – then design and implement customized strategies for effective decision making

It's all about Risk Management!

Auto Insurance

- An auto insurance company wants to expand its pricing points significantly beyond its class plan.
- Underwriters would like to segment good vs. poor risks and place them into a wide range of pricing tiers.

Workers' Compensation

- An insurance company providing workers' compensation is trying to predict if there are any misclassifications on a policy.
- The company would like to prioritize policies for manual audit based on its relative chances of misclassification.

Risk Management

- Segmentation is important in different areas of risk management like operational risk, etc.

Credit Risk

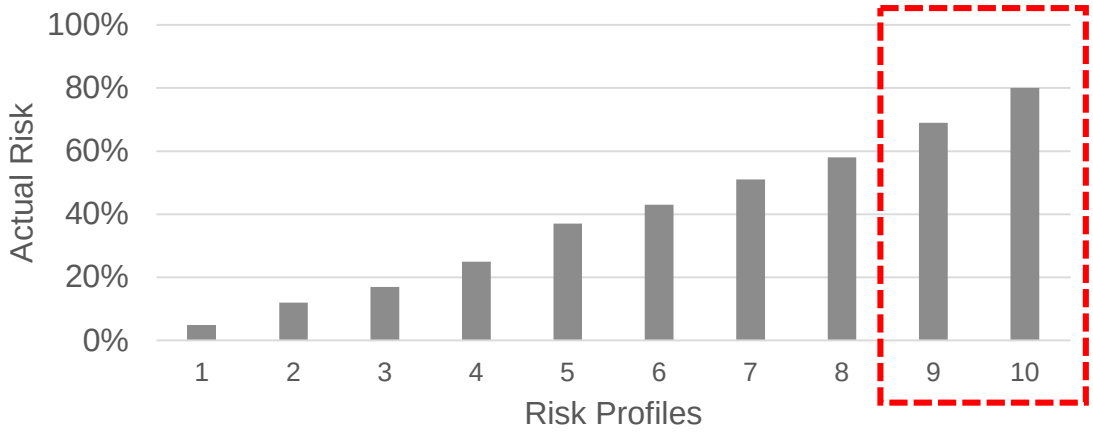
- Applicants are segmented based on credit risk through their credit scores and decisions are made based on the segment they fall in.

Fraud Detection

- Fraud detection is sometimes an integral part of claims modeling where insurance companies segment claims based on their propensity of being a fraud claim.

Putting Predictive Modeling Hat On

Would replacing Logistic Regression by Linear Regression still give the same segmentation results?



Executive Summary

Abstract

- By building models in a constructive manner, we see that the **signs of coefficients will be the same.**
- Also, in most realistic cases, the ranking of observations based on estimated values **will be similar, if not identical**

Empirical Analysis

- **Case Study 1:** For an international personal auto book of business, we show that linear and logistic regressions give the same segmentation results for a claim frequency indicator model.
- **Case Study 2:** For a US commercial auto book of business, we prove that even on an actual dataset where variable selection is done in a constructive manner, linear and logistic regressions give similar segmentation results for a loss ratio indicator model.
- **Case Study 3:** **Simulation Study**
  - We show some limitations when modeling is performed violating a few assumptions.
  - By correcting for the assumptions, we show that linear and logistic regressions give identical segmentation results.

Theoretical Proofs

- We first show that the coefficient in linear and logistic regression have same sign.
- We prove for the case of one independent variable that ranking of observations is identical.
- Under certain assumptions, we extend the proof to multivariate case and show that ranking of observations would be identical.

Case Study 1: International Auto

Problem Statement

An international auto book of business wants to rank policies based on their relative risk.

Target Variable: Claim Indicator (1 if a policy has a claim and 0 otherwise)

Approach:

- *Model 1:* Logistic Regression
- *Model 2:* Linear Regression

Results

Comparison of coefficients:

Variable	Linear	Logistic
Female_Ind	-0.013	-0.213
LNWeight	0.039	0.998
NCD	0.001	0.015
AgeCat	0.003	0.040
VAgeCat	-0.008	-0.142

Disruption

	1	2	3	4	5	6	7	8	9	10
1	913	0	0	0	0	0	0	0	0	0
2	27	558	0	0	0	0	0	0	0	0
3	0	0	755	0	0	0	0	0	0	0
4	0	0	0	739	0	0	0	0	0	0
5	0	0	0	0	867	0	0	0	0	0
6	0	0	0	0	0	720	0	0	0	0
7	0	0	0	0	0	17	761	0	0	0
8	0	0	0	0	0	0	0	630	0	0
9	0	0	0	0	0	0	0	0	746	0
10	0	0	0	0	0	0	0	0	0	750

Case Study 2: Commercial Auto

Problem Statement

Segmentation on commercial auto book of business

Target Variable: Loss Ratio Indicator (1 if a policy has a and 0 otherwise)

Approach:

- *Model 1:* Logistic Regression
- *Model 2:* Linear Regression

Results

Disruption

	1	2	3	4	5	6	7	8	9	10
1	7379	396	0	0	0	0	0	0	0	0
2	399	6819	558	0	0	0	0	0	0	0
3	0	558	6312	906	0	0	0	0	0	0
4	0		906	6136	734	0	0	0	0	0
5	0	0	1	722	6192	861	0	0	0	0
6	0	0	0	12	839	6007	918	0	0	0
7	0	0	0	0	11	909	6279	577	0	0
8	0	0	0	0	0	0	583	6484	709	0
9	0	0	0	0	0	0	0	710	6763	303
10	0	0	0	0	0	0	0	0	304	7473

Case Study 2: Commercial Auto

Comparison of Coefficients

Similar signs

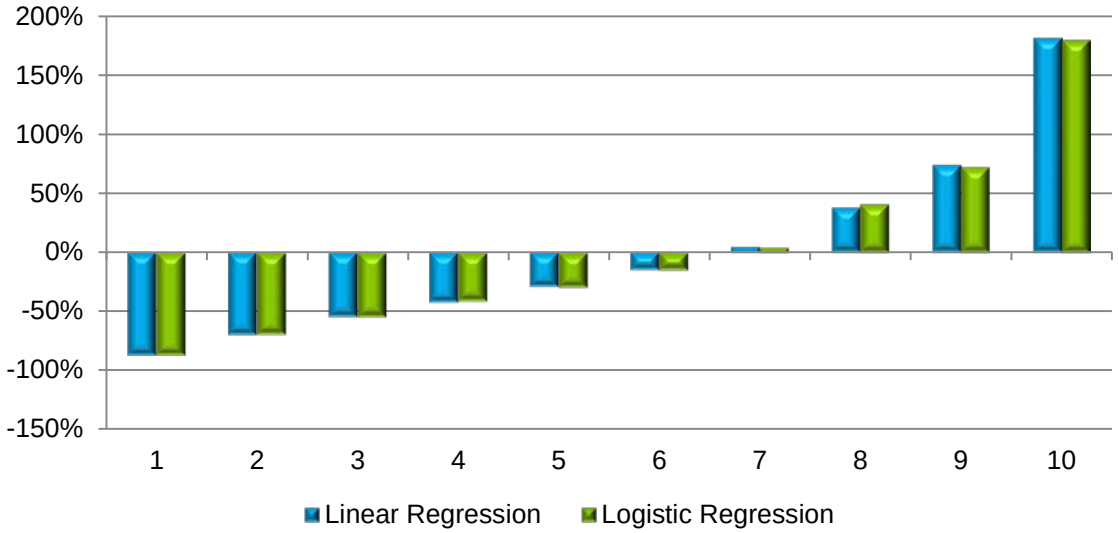
Similar significance

Variable	Logistic Regression	Linear Regression	ProbChiSq	T - value
X1	-5.9622	-0.86904	<.0001	<.0001
X2	1.7444	0.24214	<.0001	<.0001
X3	0.13335	0.01056	<.0001	<.0001
X4	-0.0312	-0.00399	0.0035	0.0063
X5	-0.022	-0.00294	0.0003	0.0005
X6	0.0844	0.01568	<.0001	<.0001
X7	0.066	0.009615	0.0031	0.0038
X8	-0.0396	-0.0312	0.0142	0.0011
X9	-0.1622	-0.01732	<.0001	<.0001
X10	-0.0858	-0.00724	<.0001	0.0004
X11	0.0219	0.00213	0.0267	0.1279
X12	0.1896	0.02286	<.0001	<.0001
X13	-0.0268	-0.00296	0.1528	0.2196
X14	0.0694	0.00888	<.0001	<.0001
X15	0.07575	0.01137	<.0001	<.0001
X16	0.0141	0.00333	0.1726	0.0197
X17	-0.013	-0.00348	0.3303	0.0597
X18	-0.1392	-0.02178	<.0001	<.0001
X19	0.3368	0.07628	<.0001	<.0001

Case Study 2: Commercial Auto

Model Performance

Comparison of Lift Curves



Decile	Linear Regression	Logistic Regression
1	-87%	-87%
2	-70%	-70%
3	-55%	-55%
4	-42%	-41%
5	-29%	-30%
6	-15%	-15%
7	5%	4%
8	38%	41%
9	74%	72%
10	182%	180%

- Both linear and logistic regression give very similar segmentation results.
- There isn't any difference in the lift produced by the two models.

- Logistic denotes the loss ratio Indicator relativity values obtained in Logistic regression.
- Linear denotes the loss ratio Indicator relativity values obtained in Linear regression.
- Loss ratio relativity is obtained by dividing the difference between the number of nonzero loss ratio policies in the decile and the overall average number of nonzero loss ratio policies by the overall number.

Case Study 3: Simulation Study

Data

Data with 10,000 observations has been simulated with six continuous predictive variables (X1 – X6) and one binary target variable (Y).

Target Variable:

Y	Frequency
0	3,500
1	6,500

Independent Variables:

Correlation with Y

Variable	Correlation
X1	0.8
X2	0.6
X3	0.45
X4	0.3
X5	0.1
X6	0

Correlation Matrix

	X1	X2	X3	X4	X5	X6
X1	1.000	0.519	0.274	0.274	0.111	0.003
X2	0.519	1.000	0.904	0.666	0.041	-0.004
X3	0.274	0.904	1.000	0.374	-0.015	0.002
X4	0.274	0.666	0.374	1.000	0.064	-0.011
X5	0.111	0.041	-0.015	0.064	1.000	-0.001
X6	0.003	-0.004	0.002	-0.011	-0.001	1.000

Case Study 3: Continued

Regression with one predictive variable:

- Both linear and logistic regression have been performed with Y as the dependent variable and with **only one** independent variable.
- In all these cases, it has been noticed that the ranking of observations is identical with a rank correlation coefficient of 1 between the predicted vectors.

So, when one variable is picked , it does not matter how well X is able to explain Y. The ranking is always the same.

Case Study 3: Continued

Regression with all predictive variable:

- Coefficients are no longer of the same sign.
- Rank correlation coefficient dropped to 90% with only 50% observations on the diagonal in the disruption report.

Linear -> Logistic	1	2	3	4	5	6	7	8	9	10	Grand Total
1	929	36	13	6	1	14					999
2	70	896	34								1000
3		68	888	44							1000
4			65	884	51						1000
5				66	872	62					1000
6					76	716	19	25	133	31	1000
7						78	99	108	193	522	1000
8						56	378	73	241	252	1000
9						9	419	139	261	172	1000
10						65	85	655	172	24	1001
Grand Total	999	1000	1000	1000	1000	1000	1000	1000	1000	1001	10000

It can be noticed that some of the independent variables here have very high correlation among themselves making the coefficients unstable. So, we are basically violating all assumptions of regression and still achieve about 90% ranking and 50% of points at the diagonal.

Case Study 3: Continued

Regression after constructing principal components:

- Principal components are constructed to make variables orthogonal
- Multivariate regression has been performed on the principal components and noticed that Coefficients are now of the same sign in both the models

Linear -> Logistic	1	2	3	4	5	6	7	8	9	10	Grand Total
1	996	3									999
2	3	983	14								1000
3		14	965	21							1000
4			21	955	24						1000
5				24	947	29					1000
6					29	941	30				1000
7						30	939	31			1000
8							31	943	26		1000
9								26	958	16	1000
10									16	985	1001
Grand Total	999	1000	1000	1000	1000	1000	1000	1000	1000	1001	10000

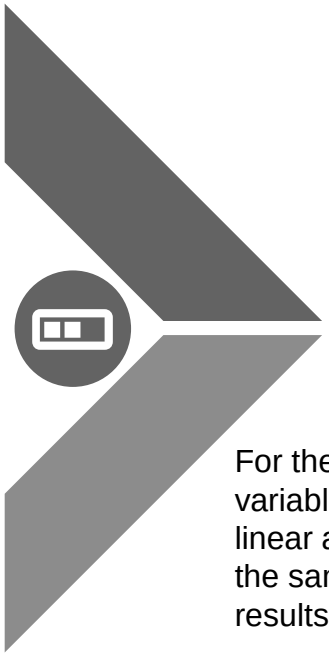
Rank correlation coefficient is now approximately 100% showing that Linear and Logistic regression give very similar segmentation results. Error to the adjacent decile is now <4% again, which suggests that if linear and logistic regressions and performed in a correct and constructive manner, the results hold true.

Theoretical Proofs

Approach

1

Prove that the signs of coefficients are the same when both linear and logistic regression are performed on the same dataset.



2

For the case of one variable, prove that both linear and logistic give the same segmentation results.

3

For the multivariate case, prove that both linear and logistic give similar segmentation results.

For both linear and logistic regressions, least squares estimates will be used to perform regression and compare coefficients and segmentation.

Formulation

Consider n variables $X_1, X_2, ..., X_n$ which are being used to rank observations based on the expected value of a binary or categorical dependent variable Y.

Let there be *K* observations that we are trying to rank. So, all variables are k-variate.

Let Y_l denote the estimated value of Y from Linear regression and Y_L denote the estimated value of Y from Logistic regression.

Mathematical Representation:

Linear regression:

$$Y_l = E(Y) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_n X_n$$

Logistic regression:

$$\log\left(\frac{E(Y)}{1-E(Y)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

Where Y_L = E(Y)

We start with proving that the signs of coefficients of the independent variables in both the models are the same, i.e., α_i and β_i are of the same sign when Least squares estimates are used
The proof when Maximum likelihood estimates (MLE) are considered instead of Least squares estimates follows on similar lines

Solutions of Linear Regression

The coefficients $\alpha_1, \ldots, \alpha_n$ are obtained by minimizing the following function with respect to the parameters $\alpha_0, \alpha_1, \ldots \alpha_n$

$$\sum_{i=1}^{k} (y_i - (\alpha_0 + \alpha_1 x_{1i} + \ldots + \alpha_n x_{ni}))^2$$

The parameters $\alpha_0, \alpha_1, \ldots \alpha_n$ are obtained by partial differentiation with respect to the parameters and equating the obtained coefficients to zero. This method yields the following solutions for $\alpha_0, \alpha_1, \ldots \alpha_n$

$$\alpha_j = \left(\sum_{i=1}^{k} (y_i - \overline{y})(x_{ji} - \overline{x_j})\right) / \sum_{i=1}^{k} (x_{ji} - \overline{x_j})^2$$

The value of α_0 is not important here as we are not really concerned about the sign or magnitude of the intercept term as we are concerned about the signs of coefficients of the parameter.

Solutions of Linear Regression - Continued

Revisiting the solution for linear regression, the system of equations for solving for the coefficients would be of the form:

$$\begin{bmatrix} k & \sum_{i=1}^{k} x_{1i} & \cdots & \sum_{i=1}^{k} x_{ni} \\ \sum_{i=1}^{k} x_{1i} & \sum_{i=1}^{k} x^2_{1i} & \cdots & \sum_{i=1}^{k} x_{1i}x_{ni} \\ & \cdots & \cdots & \\ \sum_{i=1}^{k} x_{ni} & \sum_{i=1}^{k} x_{ni}x_{1i} & & \sum_{i=1}^{k} x^2_{ni} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} y_i \\ \sum_{i=1}^{k} x_{1i}y_i \\ \cdots\cdots\cdots \\ \sum_{i=1}^{k} x_{ni}y_i \end{pmatrix}$$

Applying the transformations: $R_2 \rightarrow R_2 - \bar{x}_1 R_1$, $R_3 \rightarrow R_3 - \bar{x}_2 R_1$, … , $R_{n+1} \rightarrow R_{n+1} - \bar{x}_n R_1$ would give the following for solving $\beta_1, \beta_2, \cdots, \beta_n$

$$\begin{bmatrix} \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)^2 & \cdots & \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)(x_{ni} - \bar{x}_n) \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{k} (x_{ni} - \bar{x}_n)(x_{1i} - \bar{x}_1) & \cdots & \sum_{i=1}^{k} (x_{ni} - \bar{x}_n)^2 \end{bmatrix} \begin{pmatrix} \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \\ \vdots \\ \sum_{i=1}^{k} (x_{ni} - \bar{x}_1)(y_i - \bar{y}) \end{pmatrix}$$

Solutions of Logistic Regression

The coefficients $\beta_0, \beta_1, ..., \beta_n$ are obtained by minimizing the equation

$$\sum_{i=1}^{k} (y_i - (\exp(\beta_0 + \beta_1 x_1 + ... + \beta_n x_n) / (1 + \exp(\beta_0 + \beta_1 x_1 + ... \beta_n x_n))))^2$$

With respect to $\beta_0, \beta_1, ..., \beta_n$

These coefficients can be obtained by solving the equations obtained after partial differentiating the above equation w.r.t each of the coefficients. The system of equations thus obtained would be:

$$\sum_{i=1}^{k} \frac{y_i \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni})}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2} - \sum_{i=1}^{k} \frac{(\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^3} = 0$$

$$\sum_{i=1}^{k} \frac{y_i x_{1i} \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni})}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2} - \sum_{i=1}^{k} \frac{x_{1i}(\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^3} = 0$$

$$\sum_{i=1}^{k} \frac{y_i x_{ni} \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni})}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2} - \sum_{i=1}^{k} \frac{x_{ni}(\exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^2}{(1 + \exp(\beta_0 + \beta_1 x_{1i} + ... + \beta_n x_{ni}))^3} = 0$$

Solutions of Logistic Regression - Continued

The solution to the above system cannot be obtained directly but can be solved numerically using multivariate Newton Raphson method. In order to apply the Newton Raphson method, we will start with an initial solution of

$$\tilde{\beta} = (\beta_0, \beta_1, ...\beta_n) = (0, 0, ..., 0)$$

We will denote the initial solution by β^0, the solution after first iteration by β^1 ,and so on.

If the original system of equations to be solved for the equations is denoted by

$$\tilde{f} = \begin{pmatrix} f_0 \\ f_1 \\ \cdots \\ f_n \end{pmatrix}$$

Then the solution after p iterations would be $\tilde{\beta}_p = \tilde{\beta}_{p-1} - J^{-1} f_{\tilde{\beta}=\tilde{\beta}_{p-1}}$

Where J is the Jacobian of $\tilde{f}$ at $\tilde{\beta}_{p-1}$ and can be denoted by: J = $\begin{pmatrix} \frac{\partial f_0}{\partial \beta_0} & \cdots & \frac{\partial f_0}{\partial \beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial \beta_0} & \cdots & \frac{\partial f_n}{\partial \beta_n} \end{pmatrix}$

Solutions of Logistic Regression - Continued

Solving for the above matrix at the initial solution (0,…,0) would give the following:

$$J_{\beta^0} = \begin{bmatrix} \frac{-1}{16} & \frac{-\sum_{i=1}^{k} x_{1i}}{16} & \cdots & \frac{-\sum_{i=1}^{k} x_{ni}}{16} \\ \frac{-\sum_{i=1}^{k} x_{1i}}{16} & \frac{-\sum_{i=1}^{k} x^2{}_{1i}}{16} & \cdots & \frac{-\sum_{i=1}^{k} x_{1i}x_{ni}}{16} \\ & \cdots & \cdots & \\ \frac{-\sum_{i=1}^{k} x_{ni}}{16} & \frac{-\sum_{i=1}^{k} x_{1i}x_{ni}}{16} & & \frac{-\sum_{i=1}^{k} x^2{}_{ni}}{16} \end{bmatrix}$$

The solution after first iteration β^1 would be obtained by solving for $\beta_0, \ldots, \beta_n$ in the equation

$$\begin{bmatrix} \frac{1}{16} & \frac{\sum_{i=1}^{k} x_{1i}}{16} & \cdots & \frac{\sum_{i=1}^{k} x_{ni}}{16} \\ \frac{\sum_{i=1}^{k} x_{1i}}{16} & \frac{\sum_{i=1}^{k} x^2{}_{1i}}{16} & \cdots & \frac{\sum_{i=1}^{k} x_{1i}x_{ni}}{16} \\ & \cdots & \cdots & \\ \frac{\sum_{i=1}^{k} x_{ni}}{16} & \frac{\sum_{i=1}^{k} x_{1i}x_{ni}}{16} & & \frac{\sum_{i=1}^{k} x^2{}_{ni}}{16} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \frac{\sum_{i=1}^{k} y_i}{4} - \frac{k}{8} \\ \frac{\sum_{i=1}^{k} x_{1i}y_i}{4} - \frac{\sum_{i=1}^{k} x_{1i}}{8} \\ \cdots \\ \frac{\sum_{i=1}^{k} x_{ni}y_i}{4} - \frac{\sum_{i=1}^{k} x_{ni}}{8} \end{pmatrix}$$

Solutions of Logistic Regression - Continued

Applying the below transformations

$$R_2 \rightarrow R_2 - R_1,\ R_3 \rightarrow R_3 - R_1,\ \ldots,\ R_{n+1} \rightarrow R_{n+1} - R_1$$

$$\frac{1}{16}\begin{bmatrix} k & k\bar{x}_1 & \cdots & k\bar{x}_n \\ 0 & \sum_{i=1}^{k}(x_{1i}-\bar{x}_1)^2 & \cdots & \sum_{i=1}^{k}(x_{1i}-\bar{x}_1)(x_{ni}-\bar{x}_n) \\ & \cdots & \cdots & \\ 0 & \sum_{i=1}^{k}(x_{ni}-\bar{x}_n)(x_{1i}-\bar{x}_1) & & \sum_{i=1}^{k}(x_{ni}-\bar{x}_n)^2 \end{bmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} k\bar{y} - \frac{k}{8} \\ \frac{1}{4}\sum_{i=1}^{k}(x_{1i}-\bar{x}_1)(y_i-\bar{y}) \\ \vdots \\ \frac{1}{4}\sum_{i=1}^{k}(x_{ni}-\bar{x}_1)(y_i-\bar{y}) \end{pmatrix}$$

The equations for solving $\beta_1, \beta_2, \cdots, \beta_n$ would effectively be

$$\frac{1}{16}\begin{bmatrix} \sum_{i=1}^{k}(x_{1i}-\bar{x}_1)^2 & \cdots & \sum_{i=1}^{k}(x_{1i}-\bar{x}_1)(x_{ni}-\bar{x}_n) \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{k}(x_{ni}-\bar{x}_n)(x_{1i}-\bar{x}_1) & \cdots & \sum_{i=1}^{k}(x_{ni}-\bar{x}_n)^2 \end{bmatrix}\begin{pmatrix} \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \frac{1}{4}\sum_{i=1}^{k}(x_{1i}-\bar{x}_1)(y_i-\bar{y}) \\ \vdots \\ \frac{1}{4}\sum_{i=1}^{k}(x_{ni}-\bar{x}_1)(y_i-\bar{y}) \end{pmatrix}$$

Comparison of Solutions

It can be observed that the solution of equations in logistic regression after first iteration is very similar to the solution obtained in linear regression except for a positive constant multiplier. Since we started with an initial solution of (0,0,…0) for the coefficients, the sign in the first iteration denotes the final sign, provided solution exists.

This proves that the coefficients of variables in linear regression and logistic regression would always be of the same sign.

Comparison of Methods for Segmentation

Our next aim is to inspect the ranking of observations obtained from both the methods. Consider the expected values of y_i and y_j from Linear and Logistic regressions.

Let y^l_i and y^l_j denote the estimated values from linear regression and y^L_i and y^L_j denote the estimated values in Logistic regression. Then,

$$y^l_i = \alpha_0 + \alpha_1 x_{1i} + \cdots \alpha_n x_{ni}$$

$$y^l_j = \alpha_0 + \alpha_1 x_{1j} + \cdots \alpha_n x_{nj}$$

$$y^L_j = \exp(\beta_0 + \beta_1 x_{1i} + \cdots \beta_n x_{ni}) / (1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots \beta_n x_{ni}))$$

$$y^L_j = \exp(\beta_0 + \beta_1 x_{1j} + \cdots \beta_n x_{nj}) / (1 + \exp(\beta_0 + \beta_1 x_{1j} + \cdots \beta_n x_{nj}))$$

$$y^l_i - y^l_j > 0 \Rightarrow \alpha_1(x_{1i} - x_{1j}) + \cdots + \alpha_n(x_{ni} - x_{nj}) > 0$$

$$y^L_i - y^L_j > 0 \Rightarrow \beta_1(x_{1i} - x_{1j}) + \cdots + \beta_n(x_{ni} - x_{nj}) > 0$$

If observations are ranked based on the estimated values, then the ranking of observations would be the same for both the methods if for all i and j, the above two expressions on the left side of the inequality have the same sign.

Case of One Independent Variable

- When there is only one independent variable, to prove that the ranking of observations based on the estimated values of dependent variables would be the same if the following expressions have the same sign:

$$\alpha_1(x_{1i} - x_{1j}) \text{ and } \beta_1(x_{1i} - x_{1j})$$

- We have already proved that the coefficients in both the regressions would be of the same sign. This proves that the above expressions will be of the same sign.

From this, we can conclude that for the case one independent variable, the ranking of observations would be the same. So, during segmentation exercise, in the case of one independent variable, it does not matter which method one employs between linear and logistic regression. The final solution would be the same.

Case of Many Independent Variables

Let us revisit the expression:

$$\alpha_1(x_{1i}-x_{1j})+\cdots+\alpha_n(x_{ni}-x_{nj}) \qquad \beta_1(x_{1i}-x_{1j})+\cdots+\beta_n(x_{ni}-x_{nj})$$

Ranking would be identical if both the above expressions have the same sign, i.e., their product should be positive

$$\Rightarrow \quad (\alpha_1,\ldots\alpha_n)\begin{pmatrix} z_1' \\ \ldots \\ z_n' \end{pmatrix}(z_1,\ldots,z_n)(\beta_1,\ldots,\beta_n) > 0$$

where, $z_k = x_{ki} - x_{kj}$

Since α_i and β_i have the same sign, let $\beta_i = k_i\alpha_i$ where $k_i > 0$

The above inequality can be rewritten as:

$$(\alpha_1,\ldots\alpha_n)\underbrace{\begin{pmatrix} z_1'z_1 & \ldots & z_1'z_n \\ \ldots & \ldots & \ldots \\ z_n'z_n & \ldots & z_n'z_n \end{pmatrix}}_{A} Diag(k_1,\ldots k_n)(\alpha_1,\ldots,\alpha_n) > 0$$

Case of Many Independent Variables

Ranking would be identical if the matrix A below is positive definite (P.D)

$$A = \begin{pmatrix} z_1'z_1 & \dots & z_1'z_n \\ \dots & \dots & \dots \\ z_n'z_n & \dots & z_n'z_n \end{pmatrix} Diag(k_1, \dots k_n)$$

In an ideal case when all the independent variables are orthogonal, the above matrix transforms to:

$$A = Diag(z_1'z_1, \dots z_n'z_n) Diag(k_1, \dots k_n)$$

$$= Diag(D_1'D_1, \dots D_n'D_n)$$

Where $D_i = \sqrt{k_i} z_i$

=> Matrix A is positive Definite and therefore,

$$(\alpha_1, \dots \alpha_n) \begin{pmatrix} z_1'z_1 & \dots & z_1'z_n \\ \dots & \dots & \dots \\ z_n'z_n & \dots & z_n'z_n \end{pmatrix} Diag(k_1, \dots k_n)(\alpha_1, \dots, \alpha_n) > 0$$

For the case when there are many independent variables, ranking based on both linear and logistic regression would be identical if the independent variables are mutually orthogonal (ideal case of regression)

Conclusions

Summary of Findings

Signs of Coefficients: The signs of coefficients would remain the same in both the methods if Least Squares estimates are used for prediction.

Case of MLE: When MLE is used for prediction, and if there is only one independent variable, the signs of coefficients would be the same for both methods. However, the same cannot be said for multivariate case.

Case of One Independent Variable: The ranking of variables based on estimated values would be exactly the same for both linear and logistic regression approaches.

Multivariate Case: In an ideal scenario where independent variables are orthogonal, both linear and logistic regressions produce identical segmentation results. When the condition is violated, correlations between independent variables impact segmentation.

Other Findings

- **Risk Profiles:** Very often, when the observations are ranked, they would be divided into deciles/centiles to generate lift curves and to check the goodness of fit. It is also observed that these risk profiles (deciles/centiles) are significantly similar.
- **Our Experience:** Our experience further indicates that such ranking model results are not only insensitive to the distribution assumptions, but also insensitive to variable format, such as discrete or continuous, linear or more complex form. For example, we know that the more the number of accidents and violations, the worse the results. Then, a ranking model result will not change much if we use a linear format, a discrete format, or other more complex format.

Further Scope

Bounds for Correlation Coefficient: We showed that in “real life” scenarios, the rank correlation coefficient is very high. However, we have not derived any bounds for the coefficient and this depends on several other factors taken into consideration. It would be interesting to derive bounds for this correlation coefficient for different cases.

Other Distributions: Another topic of interest for future research would be to prove these results on a wider class of exponential family of distributions. We would like to test these results on different distributions that belong to exponential family and generalize the results further.

Questions & Answers