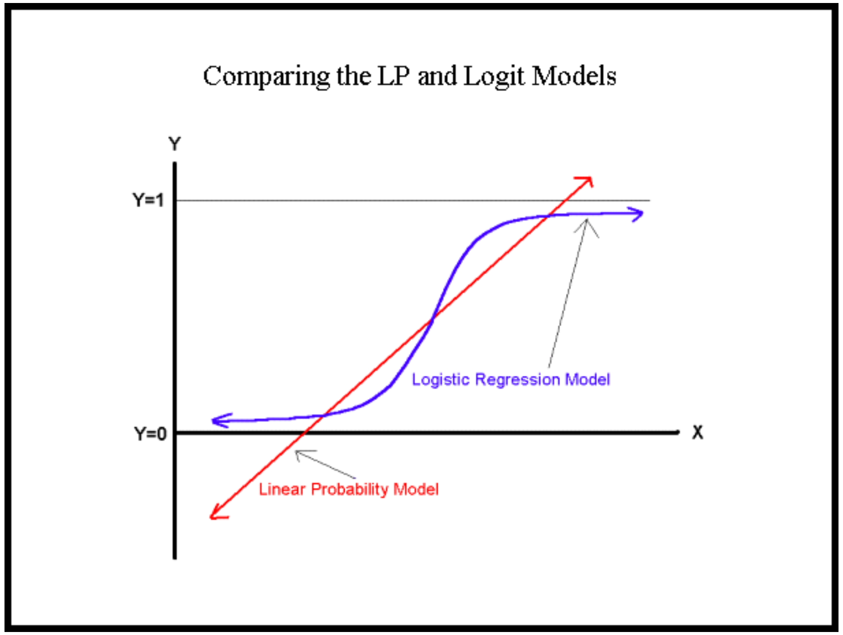
Deloitte.

Comparison of linear and logistic regression for segmentation



Moderator
Peter Wu, Deloitte Consulting LLP

Presenters
Debashish Banerjee
Kranthi Ram Nekkalapu,
Deloitte Consulting LLP

Anti-trust notice

- The Casualty Actuarial Society 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.

Flow of the presentation for today

Introduction

Empirical results

Theoretical proofs

Conclusions

Future scope/next steps

Introduction

What is segmentation?

Segmentation is a methodology that involves dividing a broad market/items/customers into subsets of entities with common characteristics and homogeneous groups — then designing and implementing strategies specific to these segments makes easier decision making.

In **Underwriting**, we are often interested in segmenting policies based on their relative risk of loss to the Insurer's charged premium.

In **Claims Modeling**, Insurance companies would like to segment claims based on their relative severity or time to settlement.

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

Segmentation is used in different areas of **Risk Management** like credit risk, operational risk, reserving and investment among others.

Segmentation is often used for modeling Credit risk. Applicants are segmented based on the estimated credit risk and decisions are made based on the segment in which the applicant falls.

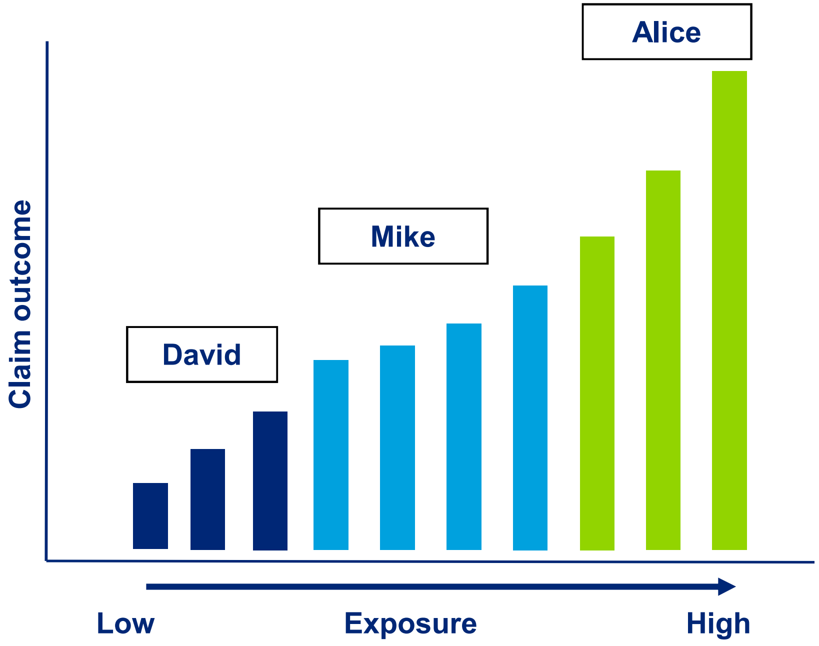
Segmentation — insurance example

Consider the following case of three individuals with different profiles

Alice	David	Mike
Female	Male	Male
38 years old	32 years old	42 years old
Clerk	IT professional	Mechanic
Married	Single	Married
1 prior claim	3 prior claims	No prior claim
Network hospital	Out of network hospital	Network hospital
Lives 40 miles from job	Lives 5 miles from job	Lives 16 miles from job
Employed for 6 years	Employed for 10 years	Employed for 22 years
Working spouse	Lives alone	Unemployed spouse
3 children	No children	2 children
(-) Financial stability	(+) Financial stability	Avg. financial stability

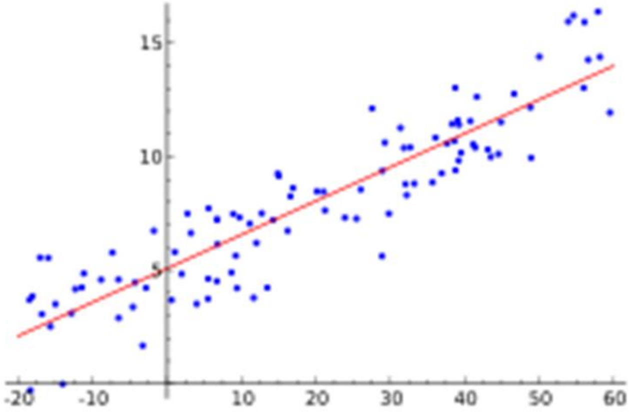
Segmentation — deciles

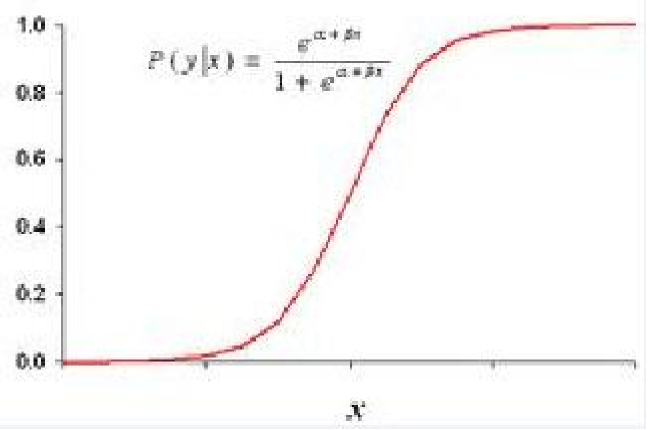
- With Predictive Modeling, a more complete set of data can be automatically assimilated to accurately segment claims into 10 different segments called deciles
- Higher deciles would represent greater risk to the company compared to the lower deciles
- Alice’s claim, which may represent the greatest exposure to an insurance company, will be in the higher deciles
- This will help manage claims more effectively, once we know the propensity to attain a higher severity versus otherwise



Segmentation — predictive modeling approach

- Various characteristics of insured/entities are identified and a predictive model is built by modelling the risk the insured/entity poses to the insurer and the corresponding characteristics
- We often perform regression using Generalised Linear Models (GLM) and create models to predict risk
- Linear Regression, due to its simplicity in interpretation is often used in insurance industry to model risk
- However, we often want to analyse whether an event has occurred or not, such as the occurrence of a fraudulent claim, propensity to attrite etc., i.e., cases where the dependent variables is a binary taking values 0 and 1
- In cases like this, one often uses Logistic Regression to predict the probability of occurrence of an event, such as the probability of occurrence of a fraudulent claim or the probability of a customer attrition/persistency
- In cases like this, Logistic regression has a clear advantage over Linear regression in that Logistic regression predicts the probability of occurrence while Linear regression does not





Segmentation — predictive modeling approach (cont.)

- For cases when the dependent variable is binary, Linear regression is not employed because
  - The corresponding estimates obtained from Linear regression can be less than 0 or greater than 1
  - The linear regression assumes the errors to follow normal, which is violated
- In most real life scenarios, we often end of segmenting policies/claims into groups based on the estimated values rather than actually trying to “predict” the accurate value
- Observations are ranked based on the estimated values and then grouped into deciles and business decisions are taken depending on the decile they fall into by observing the key drivers etc. Examples for such applications include tier assignment, company placement, etc.
- If segmentation is the target, then we should not be really worried about the estimates as long as the ranking is preserved

Question:

If one violates all the assumptions and performs Linear regression, how similar would be the segmentation results?

Answer:

Not Significantly Different — in fact , very similar, with hardly a margin of error

Empirical results

Case study 1: international auto

- An international auto book of business is used to compare linear regression and Logistic regression. The exercise is to identify policies with high chance of claim.
- Different predictive variables are regressed against the target variable claim count indicator, that takes value 1 if claim count > 0 and 0 otherwise.
- We experimented and applied both Linear and Logistic regression on the same dataset, with same dependent and independent variables
- When policies are ranked and segmented into deciles with the first 1/10th going into first decile and the next 1/10th into the second and so on, the comparison of the deciles obtained from both the models resulted in the following disruption grid:

	Linear	1	2	3	4	5	6	7	8	9	10	Grand total
Logistic	1	913										913
	2	27	558									585
	3			755								755
	4				739							739
	5					867						867
	6						720					720
	7						17	761				778
	8								630			630
	9									746		746
	10										750	750
Grand total		**940**	**558**	**755**	**739**	**867**	**737**	**761**	**630**	**746**	**750**	**7483**

Case study 1: international auto (cont.)

- It can be seen that most of the policies remained in the same decile while only a few have moved to another decile
- Only 44 out of 7483 moved the decile i.e. 0.6% error and that too a marginal move i.e. to the next decile

Other experiments:

- Just took only one independent variable at a time ; the results yielded exact same ranking (100%) with each variables used at a time i.e., in the form of y=b0+b1x+Error, as long as the sign of the parameter, b1, is the same between the two regression results
- A look at the coefficients obtained in multivariate regression reveals that the signs of coefficients in both the models are the same

Variable	Coefficient linear	Coefficient 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

Case study 2: U.S. commercial auto data

- A similar analysis is carried out using a US Commercial Auto book of data with Loss ration indicator (Indicator takes value 1 if LR > 0 and 0 otherwise) as the target variable and same observations were noticed.
- When multivariate regression is performed using many independent variables, the following coefficients have been observed:

Variable	Logistic regression	Linear regression	ProbChiSq	Probt
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

- The independent variables include driver, vehicle, violation, zip code, agent, and other characteristics.

Case study 2: U.S. commercial auto data (cont.)

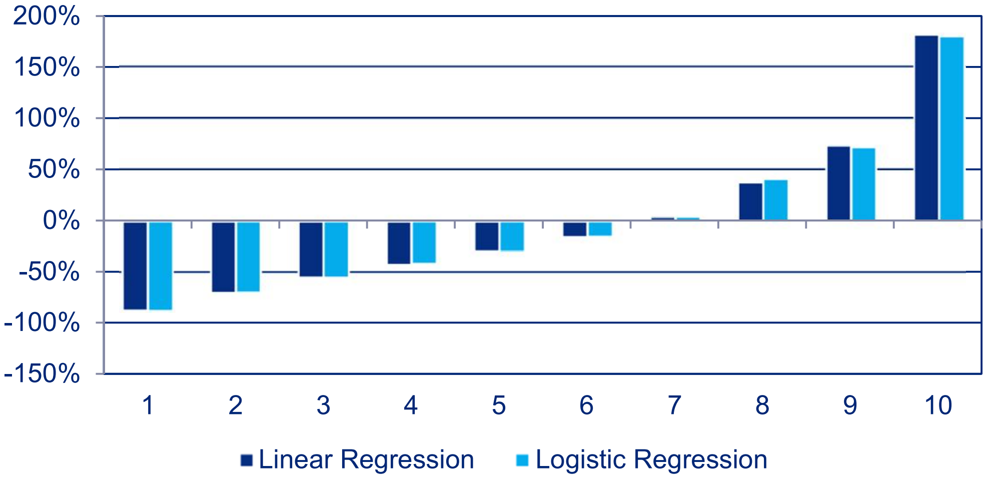
- It can be noticed that the signs of the coefficients have been preserved and the coefficients have very similar significance in both the methods
- When policies are segmented into deciles, the comparison of the deciles obtained from both the models resulted in the following disruption grid:

	Linear	1	2	3	4	5	6	7	8	9	10	Grand total
Logistic	1	7379	396									7775
	2	399	6819	558								7776
	3		558	6312	906							7776
	4			906	6136	734						7776
	5			1	722	6192	861					7776
	6				12	839	6007	918				7776
	7					11	909	6279	577			7776
	8							583	6484	709		7776
	9								710	6763	303	7776
	10									304	7473	7777
Grand total		**7778**	**7773**	**7777**	**7776**	**7776**	**7777**	**7780**	**7771**	**7776**	**7776**	**77760**

- It can be seen that most of the policies remained in the same decile while only a few have moved to adjacent decile
- The rank correlation coefficients is as high as 99.7% suggesting that the ranking of observations is similar using both the models
- Its noted that approximately 15% of the observations moved a decile (+/-1 decile) and 0.03% to +/-2 deciles.
- Similar to earlier data set, we noted that as we added more variables, we see a bit more disruptions. With one variable — its identical, and then as we add variables, it moves a bit.

Case study 2: U.S. commercial auto data (cont.)

Lift curves comparison:



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%

The Lift curves noticed show that the overall performance of both the models is the same

- 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 non-zero loss ratio policies in the decile and the overall average number of non-zero loss ratio policies by the overall number

Case study 3: simulation study

- In order to gather further evidence, a random dataset has been simulated with a binary target variable and multiple independent variables which have correlation among themselves as well as correlation with the dependent variable
- The simulated variables are as follows:
  - A dependent variable Y with 65% success (denoted by 1) and 35% failure (denoted by 0)
  - Independent variables X1, X2, X3, X4, X5 and X6 with correlations of 80%, 60%, 45%, 30%, 10% and 0% with the dependent variable respectively
  - The correlation between the independent variables is as follows:

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

Experiment #1:

- 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. Even when X6 is picked in this case.

Case study 3: simulation study (cont.)

Experiment #2:

- However, when multivariate regression was performed with Y as the dependent variable and all the simulated independent variables, the rank correlation dropped to 90% and the following disruption grid has been observed:

	Linear	1	2	3	4	5	6	7	8	9	10	Grand total
Logistic	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**

- The coefficients from Linear and Logistic are no longer of the same sign
- Although rank correlation coefficient of 90% is good enough to say that both the methods are similar, this is not the ideal case one often encounters in real life scenarios
- 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: simulation study (cont.)

Experiment #3:

- Principal Component Analysis (PCA) has been performed on the independent variables to make them orthogonal and multivariate regression was then performed on the orthogonal variables and the following disruption grid has been noticed:

	Linear	1	2	3	4	5	6	7	8	9	10	Grand total
Logistic	1	995	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**

- Coefficients are now of the same sign in both the models, as we have been observing always.
- 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 you do the linear regression in a constructive and correct way and you to logistic in the correct way, then the results holds true.

Theoretical proofs

Approach

- The coefficients of variables in regression can be estimated either by the least squares estimates or by Maximum Likelihood Estimates (MLE)
- In the case of Linear regression, when errors are normally distributed, least squares and MLE would give the same estimates for the coefficients (A property of Normal distribution)
- For Logistic regression, the errors are not normally distributed and hence least squares and MLE would not give the same estimates for coefficients.
- However, when the independent variables do not have high correlations among themselves (something that we try to have in real life scenarios to account for multi-collinearity issue), we notice that the least squares estimates and MLE would give very similar estimates
- Theoretically, its very difficult to compute the logistics equation and work on them so , we first look at the one variable case and try to make some observations i.e.:
  - When there is only one independent variable, both Least squares estimates and MLE would give the same estimates for the dependent variable for linear and logistics
- Further, for a multivariate equation, we proved:
  - When Least squares estimates are used, the signs of coefficients in both Linear and Logistic regression are the same. i.e., the direction is preserved
  - Similar approach to the above will also prove that : when MLE estimates are used, when there is only one independent variable, the signs of coefficients in both Linear and Logistic regression are the same

Formulation

Consider n variables $X_1, X_2, \ldots, X_n$ 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

Mathematically, this can be represented as follows:

Linear regression: $Y_l = E(Y) = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_n X_n$

Logistic regression: $\log(\frac{E(Y)}{1-E(Y)}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n$

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 MLE estimated are considered for Least squares estimates follows on similar lines

Newton’s method of solving a system of equations

Newton’s method of solving a system of equations is as follows:

Let there be n simultaneous algebraic equations

$$f_1(x_1, \dots , x_n) = f_1(\mathbf{x}) = 0$$

$$\dots\dots\dots$$

$$f_n(x_1, \dots , x_n) = f_n(\mathbf{x}) = 0$$

Where **X** = $[x_1, \dots , x_n]^T$ is an n-dimensional vector. The system of equations can be more concisely represented in vector form as **f(X) = 0**. The Newton – Raphson formula for multivariate problem is:

$$\mathbf{x} <= \mathbf{x} – J_f^{-1}(\mathbf{x})f(\mathbf{x})$$

Where $J_f(\mathbf{x})$ is the Jacobian of function **f(X):**

$$J_f(\mathbf{x}) = \begin{bmatrix} \partial f_1 / \partial x_1 & \dots & \partial f_1 / \partial x_n \\ & \dots\dots & \\ \partial f_n / \partial x_1 & & \partial f_n / \partial x_n \end{bmatrix}$$

Theoretical results — solutions of logistic regression

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

$$\sum_{i=1}^{k}(y_i-(\alpha_0+\alpha_1x_{1i}+\ldots+\alpha_nx_{ni}))^2$$

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 $\alpha_0,\alpha_1,\ldots\alpha_n$

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

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

Theoretical results — solutions of logistic regression (cont.)

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

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

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

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:

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

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

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

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

Theoretical results

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

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:

$$J = \begin{pmatrix} \frac{\partial f_0}{\partial \beta_0} & \cdots & \frac{\partial f_0}{\partial \beta_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial \beta_0} & \cdots & \frac{\partial f_n}{\partial \beta_n} \end{pmatrix}$$

Theoretical results (cont.)

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 $\beta_{0,}\dots,\beta_n$ in the equation

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

Theoretical results (cont.)

Apply the below transformations

$R_2 \rightarrow R_2 - \bar{x}_1 R_1,\ R_3 \rightarrow R_3 - \bar{x}_2 R_1,\ \ldots,\ R_{n+1} \rightarrow R_{n+1} - \bar{x}_n R_1$

This would give the following:

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

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

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

Theoretical results (cont.)

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

$$\begin{bmatrix} k & \sum_{i=1}^{k} x_{1i} & \cdots & \sum_{i=1}^{k} x_{ni} \\ \sum_{i=1}^{k} x_{1i} & \sum_{i=1}^{k} x^2{}_{1i} & \cdots & \sum_{i=1}^{k} x_{1i} x_{ni} \\ & \cdots & \cdots & \\ \sum_{i=1}^{k} x_{ni} & \sum_{i=1}^{k} x_{ni} x_{1i} & & \sum_{i=1}^{k} x^2{}_{ni} \end{bmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \cdots\cdot \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} y_i \\ \sum_{i=1}^{k} x_{1i} y_i \\ \cdots\cdots\cdots \\ \sum_{i=1}^{k} x_{ni} y_i \end{pmatrix}$$

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 $\beta_1, \beta_2, \cdots, \beta_n$

$$\left[\begin{pmatrix} \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)^2 & \cdots & \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)(x_{ni} - \bar{x}_n) \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{k} (x_{ni} - \bar{x}_n)(x_{1i} - \bar{x}_1) & \cdots & \sum_{i=1}^{k} (x_{ni} - \bar{x}_n)^2 \end{pmatrix}\right] \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \cdots\cdot \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \\ \vdots \\ \sum_{i=1}^{k} (x_{ni} - \bar{x}_1)(y_i - \bar{y}) \end{pmatrix}$$

Theoretical results (cont.)

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 the two 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 => $\alpha_1(x_{1i} - x_{1j}) + \cdots + \alpha_n(x_{ni} - x_{nj})$ > 0

${y^L}_i - {y^L}_j$ > 0 => $\beta_1(x_{1i} - x_{1j}) + \cdots + \beta_n(x_{ni} - x_{nj})$ > 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:

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

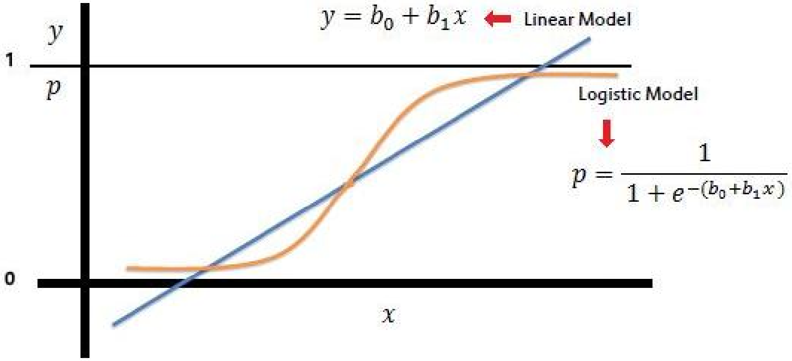
- We have already proved that the coefficients in both the regressions would be of the same sign. This proves that the above expressions will be of the same sign.
- From this, we can conclude that for the case when there is only one independent variable, the ranking of observations would be the same. 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.

For MLE case:

- If MLE estimates are considered instead of Least Squares estimates, for cases when there are many independent variables, the signs of coefficients will not be the same but when there is only one independent variable, the signs of coefficients will be the same
- The proof for this follows on the same lines as the above proof

Conclusions

Conclusions



- Case of one independent variable:
  - Same signs of coefficients
  - Same ranking of observations based on the estimated values
  - Least Squares Estimated and Maximum Likelihood Estimates would both yield similar results
- Case of more independent variables:
  - Same signs for coefficients if Least squares Estimates are used for both the models
  - Signs could be different if Maximum Likelihood Estimates are used instead of Least Squares Estimates
  - If collinearity is treated, the signs would be consistent even for MLEs
- Real life scenario:
  - Modeling is performed on uncorrelated variables
  - If correlations are noticed, variables are treated accordingly by considering Principal Components etc.,
  - Linear Regression and Logistic Regression would both give similar segmentation results
- Our prior experience:
  - Model segmentation results are not only insensitive to the distribution assumption, but also insensitive to variable format, such as discrete or continuous, linear or more complex form.
  - Example: More the accidents and violation, the worse the results.
  - Ranking model result will not change much if we use a linear format, a discrete format, or other more complex format

Future scope/next steps

Future scope

- We noticed that when Maximum Likelihood Estimates are used for prediction, the coefficients are not always of the same sign because of several reasons. We cited the case of multi-collinearity which is making the coefficients unstable and showed that when this issues is resolved, the coefficients will be of the same sign. Mathematically, there is no bound on the correlations that can be allowed and it would be interesting to derive upper bounds for the correlation coefficients so that both Linear and Logistic regression give the same direction with respect to the independent variables

- We showed that in real life scenarios, the rank correlation coefficient is very high. However, we haven't 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

- 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