Logistic Regression Experiment Report

Your Name
Department of Computer Science and Engineering
University at Buffalo, Buffalo, NY 14260
`your email`

1 Logistic Regression

[1]The use of binary categorical dependent variables is common in real data. For example, in political science empirical research: voted or not [3], won or lost the electoral contest [4]. For all these situations, a logistic regression is the best suited technique to model the dependent variable's variation given a set of independent variables.

In a logistic regression, the dependent variable only has two categories. Generally, the occurrence of the event is coded as 1 and its absence as 0. To better understand how a logistic regression works, it is necessary to understand the logic of regression analysis as a whole. Let's look at the linear model's classic notation: $y = \mathbf{w}^T\mathbf{x} + b$, where y represents the dependent variable, that is, what we are trying to understand/explain/predict. $\mathbf{x}$ represents the independent variable. The intercept/bias, b, represents the value of y when $\mathbf{x}$ equals zero. The regression coefficient, $\mathbf{w}$, represents the variation observed in y associated with the increase of one unit of $\mathbf{x}$. Technically, it is possible to estimate if there is a linear relationship between a dependent variable y and different independent variables. Moreover, the model allows the observation of the effect magnitude and to test the coefficients' statistical significance (p-value and confidence intervals).

A logistic regression can be interpreted as a particular case of generalized linear models (GLM), in which the dependent variable is dichotomous. This is defined as the following specification:

$$\log \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + b \,.$$

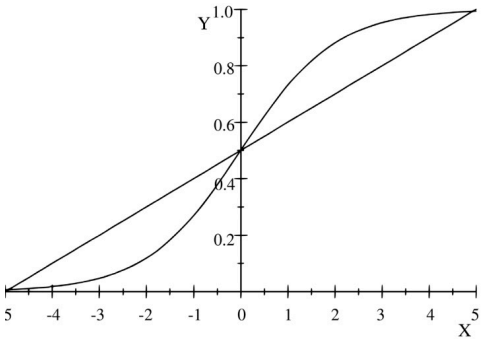

Figure 1: Linear regression line versus logistic curve.

Figure 1 compares the linear and logistic models. Because the dependent variable in the logistic model takes on only two values (0 or 1), the probability predicted by the model must also be limited to that interval. When $\mathbf{x}$ (independent variable) takes on lower values, the probability approaches zero. Conversely, as $\mathbf{x}$ increases, the probability approaches 1. For [5], that logistic functions vary between 0 and 1 explains the model's popularity. Given that the dependent variable's binary nature violates some the linear model's assumptions (homoscedasticity, linearity, normality), using a linear model to analyze binary variables may generate inefficient and biased coefficients.

To train a logistic regression model, we are given a set of training data in terms of $\mathcal{D} \triangleq \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. We use maximum likelihood estimation to define a loss function as

$$\ell(\mathbf{w}, b) = -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w}, b) \,.$$

[1]Adaptive from [2]. Encourage to use your own word in your submission.

Preprint. Under review.

-2log likelihood null	-2log likelihood	Cox & Snell R^2	Nagelkerke R^2	BIC
3,057,559	237,4225	0.229	0.308	301,891

Figure 3: Goodness of fit.

Let $\beta \triangleq (\mathbf{w}; b)$; $\hat{\mathbf{x}} \triangleq (\mathbf{x}; 1)$; $p_1(\hat{x}; \beta) \triangleq p(y = 1|\hat{\mathbf{x}}; \beta)$; $p_0(\hat{x}; \beta) \triangleq p(y = 0|\hat{\mathbf{x}}; \beta)$, the loss can be simplified as

$$\ell(\beta) = \sum_{i=1}^{N} \left((1 - y_i)\beta^T \hat{\mathbf{x}}_i + \log \left(1 + e^{-\beta^T \hat{\mathbf{x}}_i} \right) \right) .$$

Consequently, the gradient of the loss w.r.t. β is equal to

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{m} \hat{\mathbf{x}}_i (1 - y_i - p_0(\hat{\mathbf{x}}_i; \beta)) ,$$

where gradient descent will be applied for optimization.

2 Experiments

To illustrate the application of the logistic regression, we replicated the data from [1] on corruption and reelection. [Description of the data omitted here] Figure 2 summarizes the data.

Variables	Description
Sex (Control)	*Dummy:* Female (0); Male (1)
Age (Control)	Continuous: age at election.
Education (Control)	Categorical ordinal: Read and write (0); Elementary School incomplete (1); Elementary School complete (2); High School incomplete (3); High School complete (4); Tertiary education incomplete (5); Tertiary Education (6).
Poverty (Control)	Continuous: percentage of poor people in the state.
Ideology (Control)	Categorical: Left (0); Center (1); Right (2).
Vote Increase 2006 (Control)	*Dummy*: Increased (1); Lowered (0).
Change (Control)	*Dummy*: Changed parties (1); Did not (0).
Pork (Control)	Continuous: success rate of execution of parliamentary amendments.
Seats per state (Control)	Continuous: number of seats for each state at the Chamber of Deputies.
Expenditures (Control)	Continuous: campaign expenditures
Scandal (IV)	*Dummy:* Involved in a scandal (1); Not involved in a scandal (0).
Reelection (DV)	*Dummy:* Reelected (1); Not-reelected (0).

Figure 2: Data for logistic regression.

We evaluate the quality of the model's fit. Table 3 summarizes some goodness-of-fit measures typically reported in models estimated by the maximum likelihood. We report the value of -2LL to make comparing the models easier. In the null model the -2LL was 3,057,559 and the model with independent variables was 237,4225. In this case, we observe a considerable reduction. This means that the model with the independent variables has a superior fit to the null model. Similarly, the BIC (Bayesian Information Criterion) is another measure based on maximum likelihood. The smaller, the better. The model tested has a BIC of 301.891, while the null model's was 3,066.105.

References

[1] Mônica Maria Machado Ribeiro Nunes de Castro and Felipe Denegri Menegas Nunes. Candidatos corruptos são punidos?: accountability na eleição brasileira de 2006. 2014.

[2] Antônio Alves Tôrres Fernandes, Dalson Britto Figueiredo Filho, Enivaldo Carvalho da Rocha, and Willber Nascimento. Read this paper if you want to learn logistic regression. *Revista de Sociologia e Política*, 2020.

[3] Jairo Nicolau. An analysis of the 2002 presidential elections using logistic regression. *Brazilian Political Science Review*, 1:125–135, 2007.

[4] Vitor De Moraes Peixoto. Financiamento de campanhas: o brasil em perspectiva comparada. 2009.

[5] Matthias Schroder. Logistic regression: A self-learning text. *Technometrics*, 45:109 – 110, 2003.