Agri Articles
(e-Magazine for Agricultural Articles)
Volume: 03, Issue: 02 (MAR-APR, 2023)
Available online at http://www.agriarticles.com
©Agri Articles, ISSN: 2582-9882

Time Series Analysis for Logistic Regression Model

(*Rakesh Kumar Meena[1] and Vishnu Shankar Meena[2])
[1]Research Scholar (Agril. Statistics), IAS, BHU, Varanasi-221005
[2]Assistant Professor (Agril. Economics), COA, Bharatpur
*Corresponding Author's email: rakeshdeepbsjpr@gmail.com

Logistic regression analysis (LRA) extends the techniques of multiple regression analysis to research situations in which the outcome variable is categorical. Logistic regression models the probabilities for classification problems with two possible outcomes. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative distribution function of logistic distribution.

It's an extension of the linear regression model for classification problems. Logistic Regression is one of the most commonly used machine learning algorithms that is used to model a binary variable that takes only 2 values – 0 and 1. Logistic Regression is to develop a mathematical equation that can give us a score in the range of 0 to 1.

The fundamental model underlying multiple regression analysis (MRA) posits that a continuous outcome variable is, in theory, a linear combination of a set of predictors and error. Thus, for an outcome variable, Y, and a set of p predictor variables, X1,...,Xp, the MRA model is of the form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon = \alpha + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

Where is the Y-intercept (i.e., the expected value of Y when all X's are set to 0), β_j is a multiple (partial) regression coefficient (i.e., the expected change in Y per unit change in X_j assuming all other X's are held constant) and e is the error of prediction. If error is omitted, the resulting model represents the expected, or predicted, value of Y:

$$E(Y|X_1,\ldots,X_p) = Y' = \alpha + \sum_{j=1}^{p} \beta_j X_j$$

Note that Y=Y' +e. Thus, we can interpret the MRA model as follows: each observed score, Y, is made up of an expected, or predictable component, Y', that is a function of the predictor variables X1,...,Xp, and an error, or unpredictable component, e , that represents error of measurement (i.e., unreliability) and/or error in the selection of the model (i.e., mis-specification).

Types of Logistic Regression: 1. Binary Logistic Regression: The categorical response has only two 2 possible outcomes. Example: Spam or Not 2. Multinomial Logistic Regression: Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan) 3. Ordinal Logistic Regression: Three or more categories with ordering. Example: Movie rating from 1 to 5

When and why binary logistic regression? When the dependent variable is non-parametric and we don't have homoscedasticity (Variance of dependent variable and independent variable are not equal).Used when the dependent variable has only two levels.(yes / No ,male /female)**.**If multivariate normality is suspected.If we don't have linearity.

Assumptions of Logistic regression: The logistic regression method assumes that:The outcome is a binary or dichotomous variable like yes or no, positive or negative, 1 or 0.There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is logit (p) = log (p / (1-p)), where p is the probabilities of the outcome. There are no influential values (extreme values or outliers) in the continuous predictors. There are no high intercorrelations (i.e. multicollinearity) among the predictors.

Diagnostics and model checking for logistic regression: Assessment of model fit – model deviance: The deviance of a fitted model compares the log-likelihood of the fitted model to the log-likelihood of a model with n parameters that fits the n observations perfectly. It can be shown that the likelihood of this saturated model is equal to 1 yielding a log-likelihood equal to 0. Therefore, the deviance for the logistic regression model is

$$\text{DAV} = -2 \sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i) + (1 - Y_i) \log (1- \hat{\pi}_i)$$

Where, $\hat{\pi}_i$ is the fitted values for the ith observation. The smaller the deviance, the closer the fitted value is to the saturated model. The larger the deviance, the poorer the fit.

Hosmer-Lemeshow goodness of fit test:

$$\text{For this test, } H_0\text{: E[Y]} = \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$$

$$H_a : \text{E[Y]} \neq \frac{\exp(X'\beta)}{1+\exp(X'\beta)}$$

To calculate the test statistic: Order the fitted values, Group the fitted values in to c classes (c is between 6 and 10) of roughly equal size, Calculate the observed and expected number in each group, Perform a χ 2 goodness of fit test.

Multicollinearity: Multicollinearity (or collinearity) occurs when two or more independent variables in the model are approximately determined by a linear combination of other independent variables in the model. For example, we would have a problem with multicollinearity if we had both height measured in inches and height measured in feet in the same model. The degree of multicollinearity can vary and can have different effects on the model. When perfect collinearity occurs, that is, when one independent variable is a perfect linear combination of the others, it is impossible to obtain a unique estimate of regression coefficients with all the independent variables in the model. Moderate multicollinearity is fairly common since any correlation among the independent variables is an indication of collinearity. When severe multicollinearity occurs, the standard errors for the coefficients tend to be very large (inflated), and sometimes the estimated logistic regression coefficients can be highly unreliable.

Residuals of logistic regression model: Residuals can be useful for identifying potential outliers (observations not well fit by the model) or misspecified models. Two types of residuals (i) Deviance residuals (ii) Partial residuals

$$Dev_i = \pm \{[-2 \sum_{i=1}^{n} [Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)]\}^{1/2}$$

Where the sign is positive when $Y_i \geq \hat{\pi}_i$ and negative otherwise.

Partial residuals: The partial residual is useful for assessing how the predictors should be transformed. For the ith observation, the partial residual for the jth predictor is

$$r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_{i-\hat{\pi}_i}}{\hat{\pi}_i \ (1-\hat{\pi}_i)}$$

This approach assumes additivity of predictors.

Advantages and Disadvantages of logistic regression: Logistic regression has been widely used by many different people, but it struggles with its restrictive expressiveness (e.g. interactions must be added manually) and other models may have better predictive performance. Another disadvantage of the logistic regression model is that the interpretation is more difficult because the interpretation of the weights is multiplicative and not additive. Logistic regression can suffer from complete separation. If there is a feature that would perfectly separate the two classes, the logistic regression model can no longer be trained. This is because the weight for that feature would not converge, because the optimal weight would be infinite. This is really a bit unfortunate, because such a feature is really useful. The problem of complete separation can be solved by introducing penalization of the weights or defining a prior probability distribution of weights. On the good side, the logistic regression model is not only a classification model, but also gives you probabilities. This is a big advantage over models that can only provide the final classification. Knowing that an instance has a 99% probability for a class compared to 51% makes a big difference. Logistic regression can also be extended from binary classification to multi-class classification.

References

1. Danny A, JOLIS Journal 2008, Logistic Regression Analysis USA
2. Hosmer, D. and Stanley, L. (1989). Applied Logistic Regression, John Wiley and Sons, Inc.
3. Tjur, Tue (2009). "Coefficients of determination in logistic regression models". *American Statistician*: 366–372.
4. Allison, Paul D. "Measures of fit for logistic regression".
5. Pohar, Maja; Blas, Mateja; Turk, Sandra (2004). "Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study".
6. Hosmer, David (2013). *Applied logistic regression*. Hoboken, New Jersey: Wiley.