Demography India
A Journal of Indian Association of Study of Population
Journal Homepage: https://demographyindia.iasp.ac.in/

A note on application of Logistic Regression Analysis in Demography

Arun Kumar Sharma[1]

Logistic regression analysis is commonly used in descriptive research designs. With the easy availability of computer packages for statistical applications and large data from national surveys like NFHS, NSS, HDS, and LASI, demographers are lured to use multivariate analysis for explanation of demographic events, health outcomes, utilization of welfare services, contraception use, and other qualitative variables, i.e., when the dependent variable is binary, ordinal, or multinomial. In cases of qualitative dependent variables logistic regression analysis is the first choice, although several alternative techniques are also available. It is a non-parametric technique and stands as an alternative to multiple regression analysis and discriminant analysis (Sheskin, 2011). We also receive papers which analyze large data by using logistic regression for explanation of categorical dependent variables. The technique certainly deserves consideration in prediction, categorization, and explanation of qualitative variables, but, like any other advanced technique, it must be applied with some caution. The purpose of this note is to seek attention of young demographers to strengths and weaknesses of logistic regression. Used indiscriminately, logistic analysis may lead to misleading inferences.

Unlike multivariate regression analysis, multiple logistic regression uses maximum likelihood method for the model of probability (p) rather than Ordinary Least Square (OLS) estimation method which uses the concept of minimizing sum of squares of the error terms. Consequently, an iteration method is used for estimation, starting with the best guess of regression coefficients. Wald estimates rather than t values are used for significance of regression coefficients. Square of Wald test follows a chi-square distribution with 1 degree of freedom. Pseudo-R square (Nagelkerke or Cox and Snell) is used as equivalent to R square used in multiple regression model. It is noteworthy that pseudo-R square is a measure of effect size which is interpreted as the relative variation in the dependent variable caused by the predictors. Likelihood ratio (LR) defined as -2 Log_nLL, where LL stands for the ratio of the likelihood of the model without IVs to that with IVs. It has been shown that LR follows a chi-square distribution with degrees of freedom equal to number of parameters. Theory suggests that logistic analysis is a better choice than multiple regression analysis when the observations are not independent and the predictors do not follow a normal distribution (the standard NID assumption behind parametric statistics). Another strength of the logistic analysis is that backward and forward regression techniques to identify an optimum set of predictors are also permitted.

Yet, a researcher must remember that more advanced a statistical application is, more caution has to be exercised in its use. For small samples, collected by the researcher(s), the LR test is considered to be a better (more reliable) choice than the Wald test. Sheskin (2011) argues that Wald test may often be compromised: it tends to inflate standard error of large predictor coefficients, resulting in too low values of chi-square. However, such issues become relatively less significant when the researcher is working with big data.

[1] *Professor of Sociology, IIT Kanpur (Retd)*

Before choosing the logistic regression, with two or more predictor variables, one must ponder on a number of points as follows: risk of overfitting caused by small number of observations; risk of the lack of linear relationship between log odds as dependent variable (DV) and the predictors as independent variables (IDV); possibility of multicollinearity; lack of internal and external validity; and the range of confidence interval. Of course, considering the thumb rule of "ten events per variable" in logistic and Cox regression the problem of overfitting would not exist in large data analysis, e.g. analysis of NFHS, IHDS or NSS datta. The problem of multicollinearity, however, exists when one applies logistic regression or even negative binomial or Poisson regression analysis, a fact commonly ignored in applied demography.

Although logistic regression is more appropriate to analyze non-linear relationships, it cannot take care of all types of non-linear and symmetrical relationships, because it assumes that the relationship between log-odds and independent variables is linear. Logistic regression also assumes that the causal variables are all independently distributed and there is no confounding, i.e., there are no factors which affect both dependent and independent variables. Yet, in cases of confounding, stratification and multiple regression techniques may be used to produce "adjusted" ORs.

Redundancy or multicollinearity (Sperandei, 2014) requires a serious examination, particularly in handling socioeconomic data. For example, one should avoid using both substance use and mental disorders in the same model. There is a common tendency to include all background variables as predictors of any dependent variable. For example, Biswas and Banerjee (2023) included indicators of place of residence, age, education, caste, wealth, household consumption, insurance, and religion to predict maternity care expenditure, based on National Sample Survey data. In such cases regression coefficients suffer from high standard error and are invalid. In general, when religion is found to be a significant predictor, it is explained in terms of socioeconomic differences between different religious communities. If this is true then either religion or socioeconomic variables must be excluded from the list of predictors.

All research designs face the problem of threat to validity. There are two forms of validity: internal and external. Internal validity of a model refers to an aspect of the research design that the effects are not due to factors which have not been included in the model and there are no measurement errors. External validity refers to generalizability of results to other settings or cultures. To ensure internal validity, the research needs a good understanding of the definitions, scaling techniques, conceptual framework of the study and parsimony. Use of secondary data requires a good understanding of them, as used in the original data source. Further, parsimonious models with the minimum number of variables must be preferred over more complex models. The external validity may be ensured through external replicability analysis or using the hold out method, i.e., by dividing the data into two parts: one for estimation and another for validation.

For non-extreme events (e.g., probabilities between .25 and .75) application of multiple regression analysis (a parametric technique) is recommended more than the logistic regression (Cabrera, 1994). As a rule, parametric tests are more powerful than the non-parametric tests. However, one must check for the constancy of variance (homoscedasticity) and normality of independent variables. It should not be forgotten that if the conditions meet multiple regression yields BLUE (i.e., unbiased, and efficient) estimates which is not true for logistic regression. Moreover, choice of measurement of dependent and independent variables is of great importance. One need not use the indicators as shown in the published reports and original constructs may be developed using original data file, theory, policy requirements and logic.

Let us explain the above issues in application of logistic regression by using some examples. In general, the following steps are recommended;

1. Assuming that one wants to explain variations in anemia, on the basis of NFHS or other large data. First checkup which independent variables explain anemia.

Selection of criteria for choice of independent variables adds to both theory and policy prescriptions. Check whether variables are dummy or quantitative. Division into mildly anemic, moderately anemic, or severely anemic or any anemia must be done carefully. Dummy variables require fixing a reference category. One has to decide whether one should choose a nominal, ordinal scale or multinomial scale. All the respondents may be divided into one of these categories. One can use binary logistic with two categories, such as anemic and non-anemic. If the degree of anemia is to be explained, the categories mildly anemic, moderately anemic, or severely anemic need to be used. Then the choice is to go for ordinal or multinomial logistic analysis. If there is an order in the categories it is always better to use ordinal logistic analysis. One is not supposed to use independent, binary logistic separately for mildly anemic, moderately anemic, or severely anemic separately as the three categories are not independent. An example of an absurd inference may look as follows: although a specific predictor explains mild anemia, but it does not explain severe anemia. Obviously, if a woman belongs to the category of mildly anemic, she cannot be found in moderately or severely anemic category. Thus, one cannot compare the results of different models treating mild, moderate and severe anemia as dependent variables. Thus, one has to use the categories carefully. In this example, we may choose the ordinal logistic regression or multinomial logistic which is based on cumulative probabilities of response categories rather than individual probabilities.

2. Decision regarding the number of independent variables is very significant. It is better to begin with binary logistic with one explanatory variable at a time, and then use only significant variables in the full model. Yes, for selecting the predictors, one may be less conservative in fixing p value; here a choice of .10 rather than .01 is justified.

3. Choice of reference category for any independent variable is also important because in case of categorical independent variables with three or more categories, odds ratio of any category can only be interpreted with reference to the reference category. Comparison of other categories is technically not sound. This belongs to the same category of problem as the problem of comparing regression coefficients in multiple regression equation. A regression model that can be used for prediction does not always permit the comparison of partial regression coefficients. If the reference category is Hindu, one cannot compare odds ratios between Muslims and Christians. It is better that the dominant category, i.e., the modal category, is taken as the reference category. But theoretical consideration may sometimes demand a different choice. In that case the unconventional choice must be justified. In a paper received for the consideration of publication in *Demography India*, the author(s) used logistic regression analysis for contraceptive use. The data was collected through a primary survey. The independent variables included religion and caste. Religion was categorized as Hindu, Muslim, Christian, Shikh, Buddhist and No Religion. Castes were categorized as SC, ST, OBC, Gen, and No Caste. No Religion and No Caste were used as the reference categories. They were ill defined. This is obviously wrong for two reasons: a. the numbers of cases in the reference categories were too small; and b. estimates for religious and caste categories made no sense when compared with No Religion and No Caste. The paper was not recommended for publication.

4. Before the application and interpretation of results, it is important to test multicollinearity, logit linearity, independence of errors, and possibility of the presence of outliers. Commonsense suggests that urban-rural residence, caste, religion, wealth index and education cannot be combined in the same model as predictors because they are not independent. Also, the relationship between a dependent variable

such as anemia and predictors may not always be logit linear as assumed by the decision surface in logistic regression. To give an example of non-linear relationship, in medical use of logistic it is often said that mortality from pneumonia may be higher at both extremes of age; calculating the ORs for age as a predictor of mortality from pneumonia will not give valid results if the ages extended from newborn babies to the elderly. Mallick (2021) uses the following predictors for studying low birth weight among children, based on NFHS 4 data: geographical region, urban-rural residence, household size, wealth index, social group of mother, schooling of mother, mother smoking, chewing tobacco, consuming alcohol, woman deciding for health care, empowered for child rearing, use of oral contraceptive pills, received health check-ups, received food supplementations, received health and nutrition related information, skipped follow-ups at pregnancy, received vaccination at pregnancy, place where ANC received, pregnancy complications, type of delivery, place of delivery, sex of the child, and birth order of the child. The issue is: does it make sense to assume that all of them are independent? Obviously not. Yet, the paper was published in a peer reviewed journal.

5. Estimate the confidence intervals (CI) of OR to identify redundancy. If the confidence interval includes 1 then an OR has no meaning even if p can be defined as significant. In Mallick's paper referred above, food supplementation was shown to be having significant odds ratio, but the confidence interval contained 1. It should be treated as non-significant.

6. Some parts of the population may in fact be outliers. In the national level analysis, some states may be found to be outliers or social groups or communities may be outliers. Likewise, in the analysis of state level data certain districts or sub-districts or blocks may be outliers. Separate analysis of the dominant pattern and outliers is needed to understand the dynamics of complex situation.

7. Betas or regression coefficients are not to be interpreted in the same way as in regression coefficients in multiple regression (measuring change in the DV per unit of change in IDV). In logistic regression one must look for odds ratios (OR) which are not the same things as probabilities or p's (odds are the ratios of p to 1-p). Odds may be high, but the absolute risk may be low. Additionally, odds are obtained as Exp (Beta) and explained as the percentage increase in the probabilities of outcome with a unit change in the independent variable, at a given value of the predictor. Thus, the effect depends on both the regression coefficient as well as the value of the predictor (if it is a numeric variable).

The conclusion of this note is that the logistic regression is a good choice for classification of qualitative criterion variables and for explanation of causal relationships and model fitting but one has to apply it cautiously keeping its assumptions in mind, and choosing the predictors carefully and parsimoniously. One has to avoid temptation to include all possible predictors on which data have been collected. The researcher must also look for the scientific plausibility and meaningfulness of the association. One may find an empirical relationship between domestic violence and fertility, but it may be spurious, unless a plausible theoretical explanation exists to decide the strength and direction of causality. Looked at from this perspective, one cannot overestimate the role of systematic review of literature. The decision regarding whether an empirical relationship is causal or not depends on the conceptual framework of the study, and the conceptual framework comes from a systematic and analytical review of literature.

References

Biswas, Monirujjaman, and Banerjee, Anuradha. (2023). "Investigating various correlates associated with maternity care expenditure in India: Evidence from National Sample Survey data", Demography India, Vol. 52, No. 1, pp. 11-31.

Cabrera, A. F. 1994. "Logistic regression analysis in higher education: An applied

perspective", In John C. Smart (ed.), Higher Education: Handbook of Theory and Research. New York: Agathon Press, pp. 225-256.

Mallick, Akash. 2021. "Prevalence of low birth weight in India and its determinants: Insights from the National Family Health Survey (NFHS), 2015–2016", Anthropol. Anz. J. Biol. Clin. Anthropol, Vol. 78, No. 3, pp. 163–175, Internet (accessed on 6 March 2024).

Sheskin, David J. 2011. Handbook of Parametric and Nonparametric Statistical Procedures (Fifth edition). Boca Raton: Chapman and Hall/CRC.

Sperandei, Sandro. 2014. "Understanding logistic regression analysis", Biochemia Medica, https://hrcak.srce.hr/file/171128 (accessed on 28 February 2024).