DBS Business Review 05 | 2023
RESEARCH NOTE | dbsbusinessreview.ie

Considerations for running and interpreting a binary logistic regression analysis – A research note

Emma Beacom
Doctor of Philosophy
University College Cork
Ulster University Business School, Coleraine, Northern Ireland

Abstract

This research note discusses key considerations for analysis of categorical data using a Pearson's chi-square and binary logistic regression. It draws on experience from analysis of a country-level household survey (Northern Ireland Health Survey 2014/15), using SPSSv25, that examined the relationship between household food insecurity status and identified demographic predictors, using Pearson's Chi-Square test to check associations and binary logistic regressions to derive the predictive models. This note presents an overview of the assumptions for both tests which must be satisfied to ensure the tests are appropriate, discusses the usefulness of using Pearson's Chi-Square test as a preliminary test before using binary logistic regression, and presents an overview of how to interpret the output from a binary logistic regression model.

Keywords: Pearson's chi square analysis, binary logistical regression; categorical data; SPSS

Introduction

Binary logistic regression is a useful statistical method which examines and quantifies the predictive association between a range of independent variables and a binary outcome variable. However, if results are to be reliable and applicable, it is important that all assumptions for this test are met before it is applied to the data. This research note provides researchers with a clear overview of how to determine if a binary logistic regression is an appropriate statistical test for your chosen sample (according to satisfied assumptions) and how to interpret the findings from this test. Further, this note considers the use of a preliminary assessment of associations between variables using a Pearson's chi-square test prior to a binary logistic regression, and provides an overview of the relevant assumptions that must be satisfied to produce reliable results from a Pearson's chi-square test.

Preliminary Pearson's-chi square analysis prior to running a binary logistic regression

Studies in various fields, such as medicine and the built environment, have used the approach of checking associations between predictor variables and the dependent variable using preliminary univariate statistical analysis such as a chi-square test, followed by a logistic regression (Wilcox et al., 2016; Chen and Zhang, 2016; Antwi et al., 2017; Gross et al., 2019). This research note draws upon the example of analysing 2231 cases in a health survey data set (Northern Ireland Health Survey 2014/15) (Department of Health, 2018), using SPSSv25, with the purpose of investigating the predictors of food insecurity. This dataset included fifteen demographic variables which had been linked (to varying degrees) to the dependent variable (food insecurity) in the literature, and through prior data collection with stakeholders in an earlier stage of the study. The author conducted a Pearson's chi square test in the first instance to examine which of the independent variables (predictors) were significantly associated with the dependent variable, with the intention to insert only those significantly associated into the binary logistic regression model. Although it is acknowledged that this preliminary analysis approach is not necessary as the binary logistic regression analysis will ultimately indicate significance, it can be useful as a first point to indicate associations, particularly if there has been little prior confirmed association between variables, as it rationalises their inclusion in the regression model. Furthermore, in cases where there are a large number of variables to begin with, or similar variables, it can aid the researcher in choosing which variables should proceed for further analysis.

Pearson's chi-square test assumptions

The ability of the Pearson's Chi-Square test to properly examine associations, and of logistic regression to efficiently assess the contributions of the predictor variables to the dependent variable depends not only on selecting appropriate predictor variables, but also on ensuring assumptions for these tests are satisfied (Stoltzfus, 2011; Rana and Singhal, 2015). Therefore, before the methods for analysis were confirmed it was checked that the assumptions for the Pearson's Chi-Square and for binary logistic regression using all categorical independent variables were met. Assumptions for Pearson's chi-square test relate to (i) the data being collected at random rather than a convenience sample; (ii) the data in the cells being recorded as frequencies as opposed to percentages; (iii) categories within variables being

mutually exclusive, i.e. each respondent was only recorded in one category per variable; (v) independence of study groups; (vi) sufficient sample size (at least 20); and (vii) no cells with expected count of less than five (McHugh, 2013; Rana and Singhal, 2015). Both the dataset itself, as well as the survey methodology (NISRA, 2019) were examined to confirm these assumptions were satisfied.

Logistic regression assumptions

Assumptions for this test include (i) that the dependent variable is binary in nature, (ii) independence of errors, (iii) absence of multicollinearity, (v) no outliers in the data, and (vi) an adequate number of observations per variable (Stoltzfus, 2011; Pallant, 2016). The dependent variable had previously been recoded to ensure it was binary in nature, thus appropriate for this analysis (Ranganathan et al., 2017). The assumption of independence of errors dictates that there should be no duplicate responses in the dataset (Stoltzfus, 2011). The survey methodology (NISRA, 2019) states responses to be independent of each other and not duplicated, thereby meeting this assumption. An important assumption for logistic regression is the absence of multicollinearity (a high correlation between predictor variables) (Midi et al., 2010; Stoltzfus, 2011; Starkweather and Moske, 2011). Multicollinearity (or collinearity) is a problem because if two variables are highly correlated, the values of one can be predicted by the other, and this will therefore impact upon the regression coefficients and confidence intervals associated with these variables, and subsequently on how they are interpreted (Midi et al., 2010; Johnston et al., 2018). To check this assumption a linear regression was run on the independent variables and the correlation coefficient table (collinearity statistics) (Table 1) was examined to detect any presence of multicollinearity between variables. Collinearity was diagnosed using the Variance Inflation Factors (VIF) test (Miles and Shevlin, 2001; Midi et al., 2010; Thompson et al., 2017). VIF test estimates how much the variance is inflated, and although there is no firm consensus as to the most appropriate cut off point for VIF values (Thompson et al., 2017), generally values above ten are considered to indicate multicollinearity, and this is a common cut off point recommended by statisticians and used in the literature (Lee et al., 2016; Ngema et al., 2018; Johnston et al., 2018). VIF values detected multicollinearity between the Economic Activity and the ILO Employment variable (values of 73 and 135, respectively), therefore the Economic Activity variable was removed from further analysis, and the ILO Employment variable remained as employment status was a more commonly used terminology in the related literature and by relevant stakeholders. The VIF test was then run again without the Economic Activity variable and no multicollinearity was detected. VIF values of lower than three are considered

a conservative estimate of VIF values (Thompson et al., 2017; Johnston et al., 2018), which was the case for all values. Tolerance is a related measure of multicollinearity, and smaller values indicate increased likeliness of multicollinearity (Thompson et al., 2017). Tolerance values were checked, and all met the recommended criteria of being greater than 0.20 (Midi et al., 2010; Chen et al., 2018). There should not be any strongly influential outliers as these would bias predicted outcomes and the overall model (Stoltzfus, 2011; Pallant, 2016). Residual plots were checked, and no influential outliers were detected. A general rule regarding the number of observations per variable is that for every independent variable, there should be no less than ten cases for each binary category (Agresti, 2007; Stoltzfus, 2011). This was checked and it was confirmed that there was an adequate number of observations to avoid problems of an overfit model.

Table 1: Collinearity statistics

	Collinearity Statistics	
	Tolerance	**VIF**
Number of adults in household	0.827	1.209
Number of children in household	0.758	1.319
Gender of respondent	0.954	1.048
Age group of respondent	0.339	2.949
Marital status	0.723	1.382
Tenure (own/rent accommodation)	0.9	1.111
Car/ van available for use by household	0.813	1.229
Highest Qualification level attained	0.528	1.894
General health	0.4	2.502
Carer responsibility for sick/ disabled/ elderly (non-professional)	0.996	1.004

Anxiety/ Depression	0.952	1.05
Employment status	0.007	134.665
Economic Activity	0.014	72.921
Receipt of state benefits	0.885	1.131
Urban/rural location	0.945	1.058

Pearson's Chi-square test of association

Pearson's chi-square tests were used to assess the association between food security status and the predictors identified in the literature and by stakeholders. Pearson's chi-square test is applied to categorical data in order to assess if any observed difference between variables is statistically significant or due to chance (McHugh, 2013; Rana and Singhal, 2015), and has been used in the literature as a first step method of checking associations between predictor variables and the dependent variable before further analysis in a logistic regression model (Chen and Zhang, 2016; Antwi et al., 2017; Gross et al., 2019). This test therefore identified the significance or otherwise of any observed differences between the characteristics of those who were food secure and those who were food insecure. A significance level of $p \leq 0.25$ was chosen for this test, as a more relaxed Type 1 error rate is recommended when determining variables to include in a logistic regression model in order to avoid eliminating potentially important variables (Stoltzfus, 2011; Sperandei, 2013; Antwi et al., 2017). Predictors which were not statistically significant were eliminated and not included in the proceeding logistic regression (Figure 1).

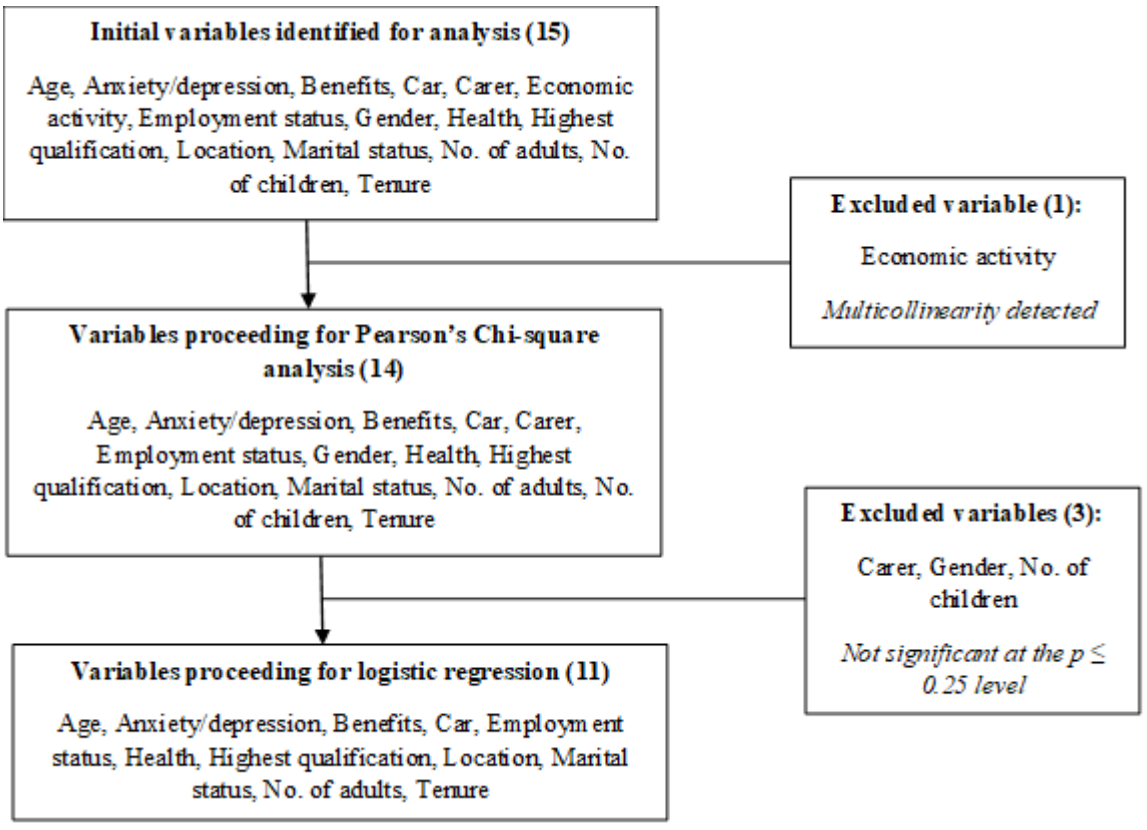

Figure 1: Verification process of the variables for inclusion in regression analysis

Binary logistic regression

Binary logistic regression was carried out to assess significant associations between food security status and the predictors previously identified (Ranganathan et al., 2017). A logistic regression rather than a linear regression was used, as a logistic regression is the appropriate method for analysis of categorical data such as that in this study, while a linear regression is used for continuous data, so was therefore not the appropriate method for this study (Tranmer and Elliot, 2008). Logistic regression is used to produce an 'odds ratio' of a single explanatory variable's (predictor's) effect on the dependent variable in the presence of more than one explanatory variable (Sperandei, 2013), thus creating a framework of a household's odds of being food insecure (or otherwise) according to the predictor variables (Ngema et al., 2018). Although the terms 'odds' and 'probability' are often used interchangeably, odds values are not the same as probabilities, as odds values are not restricted to 0-1, as probability values are (Peng and So, 2002; Tranmer and Elliot, 2008; Sperandei, 2013). Prior to running the analysis, a reference variable for each categorical variable was chosen. This reference variable was either the first or last categorical response in the variable, and was usually the response of null state (e.g. 'no children', 'not in receipt of benefits'), or the response negatively associated with food insecurity in the literature (e.g. own home, good health), to assess how change in variable response affected the dependent variable. Certain variables (number of adults in the household, age group), were recoded prior to analysis to facilitate a certain category in the variable being used as the reference group, in accordance with similar studies in the literature.

Interpreting model output

Results were considered in terms of overall model fit, as well as considering individual variable results. Various inferential tests and descriptive measures were considered when assessing model fit (classification table, likelihood ratio test, Nagelkerke R square, Hosmer and Lemeshow chi-square test, odds ratios).

Whether a case is classified into one or the other binary outcome categories is predicted using estimated probabilities and pre-determined cut-off points (Stoltzfus, 2011). The classification table shows the difference in observed and predicted model values, with better model fit being characterised by a smaller difference between observed and predicted model values (Pallant, 2016). This table displays the percentage of correctly predicted outcomes using the model (Peng and So, 2002),

and therefore can be used to evaluate the ability of the model to distinguish between groups (Stoltzfus, 2011).

The likelihood ratio test shows whether the model provides a significant improvement to the null model (Peng and So, 2002), i.e. whether the inclusion of explanatory variables contributes significantly to model fit. A p-value less than 0.05 shows that the Block 1 model is a significant improvement to the Block 0 (null) model.

The Nagelkerke R Square value assesses the variation in the dependent variable and assesses the proportion of the variance explained by the regression model. It is therefore used to measure the success of predicting the dependent variable using the independent variable (Nagelkerke, 1991). A low Nagelkerke R square value suggests that the model may not be a good fit (Ngema *et al*., 2018).

Better model fit is characterised by a p-value greater than 0.05 resulting from the Hosmer and Lemeshow chi-square test. This test assumes that the model is an adequate fit for the data, therefore this null hypothesis is only rejected if $p<0.05$ (Ngema *et al*., 2018).

Individual variable results Exp (β) (odds ratio), Exp (β) confidence intervals, and significance values were also considered. Odds ratios (ORs) and their significance were examined to determine the effect of the independent variables on the outcome variable. When using multiple independent variables in a binary logistic regression, the subsequent OR values represent the odds or likeliness of a change in the binary variable related to the independent variable after controlling for the effects of all other independent variables in the model (Stoltzfus, 2011). An OR value of one would indicate that the odds remain unchanged, an OR value greater than one indicates greater odds, and an OR value of less than one represents a decrease in the odds. Therefore if, for example, one variable had a significant OR value of 1.5, this would mean that for those with this variable characteristic, the odds of being food insecure increase 1.5 times, or by 50% (Stoltzfus, 2011). The confidence level for Exp (β) values (OR values) was set at 95%, therefore there is only a probability of 0.05 or less ($p\leq0.05$) that the value for the OR lies outside of the calculated range. Odds ratio values with a significance level of ≤0.05 were therefore considered significant.

Conclusions

Pearson's chi square test can be used as a preliminary test prior to running a binary logistic regression to check the significance of associations between the dependent and independent variables before they are entered into the predictive model. This preliminary test approach is particularly useful if one is starting off with a large

number of independent variables. Before running a Pearson's chi-square test certain assumptions must be satisfied, namely (i) the data being collected at random rather than a convenience sample; (ii) the data in the cells being recorded as frequencies as opposed to percentages; (iii) categories within variables being mutually exclusive, i.e. each respondent was only recorded in one category per variable; (v) independence of study groups; (vi) sufficient sample size (at least 20); and (vii) no cells with expected count of less than five. Before running a binary logistic regression, assumptions which must be satisfied relate to (i) that the dependent variable is binary in nature, (ii) independence of errors, (iii) absence of multicollinearity, (v) no outliers in the data, and (vi) an adequate number of observations per variable. Logistic regression results can be considered in terms of overall model fit (classification table, likelihood ratio test, Nagelkerke R square, Hosmer and Lemeshow chi-square test), as well as considering individual variable results (odds ratios).

References

Agresti, A. (2007) *An introduction to categorical data analysis*, 2nd ed. John Wiley and Sons, New Jersey.

Antwi, E., Groenwold, R.H.H., Browne, J.L., Franx, A., Agyepong, I.A., Koram, K.A., Klipstein-Grobusch, K., Grobbee, D.E. (2017) 'Development and validation of a prediction model for gestational hypertension in a Ghanaian cohort', *BMJ Open*, 7, e012670. Available at: https://doi.org/10.1136/bmjopen-2016-012670 .

Chen, C. and Zhang, J. (2016) 'Exploring background risk factors for fatigue crashes involving truck drivers on regional roadway networks: a case control study in Jiangxi and Shaanxi, China', *Springer Plus*, 5(582). Available at: https://doi.org/10.1186/s40064-016-2261-y .

Chen, W., Zhang, S., Li. R. and Shahabi, H. (2018) 'Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling', *Science of the Total Environment*, 644,1006-1018. Available at: https://doi.org/10.1016/j.scitotenv.2018.06.389

Department of Health (2018) *Health Survey Northern Ireland, 2014-2015*. [data collection]. Northern Ireland: UK Data Service. SN:8347. Available at: http://doi.org/10.5255/UKDA-SN-8347-1

Gross, S.M., Biehl, E., Marshall, B., Paige, D.M. and Mmari, K. (2019) 'Role of the elementary school cafeteria environment in fruit, vegetable, and whole-grain consumption by 6- to 8-year-old students', *Journal of Nutrition Education and Behavior*, 51,41-47. Available at: https://doi.org/10.1016/j.jneb.2018.07.002

Johnston, R., Jones, K. and Manley, D. (2018) 'Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour', *Quality & Quantity*, 52,1957-1976. Available at: https://doi.org/10.1007/s11135-017-0584-6

Lee, C.C., Liang, C.M., Chen, J.Z. and Tung, C.H. (2016) 'Effects of the housing price to income ratio on tenure choice in Taiwan: forecasting performance of the hierarchical generalized linear model and traditional binary logistic regression model', *Journal of Housing and the Built Environment*, 33,675-694. Available at: https://doi.org/10.1007/s10901-017-9572-3

McHugh, M.L. (2013) 'The Chi-square test of independence', *Biochemia Medica*, 23,143-149. Available at: https://dx.doi.org/10.11613%2FBM.2013.018

Midi, H., Sarkar, S.K. and Rana, S. (2010) 'Collinearity diagnostics of binary logistic regression model', *Journal of Interdisciplinary Mathematics*, 13,253-267. Available at: https://doi.org/10.1080/09720502.2010.10700699

Miles, J. and Shevlin, M. (2001) *Applying regression and correlation: a guide for students and researchers*. London: SAGE Publications.

Nagelkerke, N.J.D. (1991) 'A note on a general definition of the coefficient of determination', *Biometrika*, 78,691-692. Available at: https://doi.org/10.1093/biomet/78.3.691

Ngema, P.Z., Sibanda, M. and Musemwa, L. (2018) 'Household food security status and its determinants in Maphumulo local municipality, South Africa', *Sustainability*, 10,1-23. Available at: https://doi.org/10.3390/su10093307 .

NISRA (2019) *Health survey Northern Ireland*. Belfast: NISRA. Available at: https://www.nisra.gov.uk/statistics/health-survey-northern-ireland (Accessed: 28 February 2023).

Pallant J. (2016) *SPSS Survival Manual*. 6th ed. London: Open University Press.

Peng, J. & So, T.S.H. (2002) 'Logistic regression analysis and reporting: a primer.' *Understanding statistics*, 1,31-70. Available at: https://doi.org/10.1207/S15328031US0101_04

Rana, R. and Singhal, R. (2015) 'Chi-square Test and its application in hypothesis testing', *Journal of the Practice of Cardiovascular Sciences*, 1,69-71. Available at: https://doi.org/10.4103/2395-5414.157577

Ranganathan, P., Pramesh, C.S. and Aggarwal, R. (2017) 'Common pitfalls in statistical analysis: logistic regression', *Perspectives in Clinical Research*, 8,148-151. Available at: https://doi.org/10.4103/picr.picr_87_17

Sperandei, S. (2013) 'Understanding logistic regression analysis', *Biochemica Medica*, 24,12-18. Available at: https://doi.org/10.11613/BM.2014.003

Starkweather, J. and Moske, A.K. (2011) *Multinominal logistic regression*. Available at: https://it.unt.edu/sites/default/files/mlr_jds_aug2011.pdf (Accessed: 28 February 2023).

Stoltzfus, J.C. (2011) 'Logistic regression: a brief primer', *Academic Emergency Medicine*, 18,1099-1104. Available at: https://doi.org/10.1111/j.1553-2712.2011.01185.x

Thompson, C.G., Kim, R.S., Aloe, A.M. and Becker, B.J. (2017) 'Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results', *Basic and Applied Social Psychology*, 39,81-90. Available at: https://doi.org/10.1080/01973533.2016.1277529

Tranmer, M. and Elliot, M. (2008) *Binary logistic regression*. Available at: http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2008/2008-20-binary-logistic-regression.pdf (Accessed: 28 February 2023).

Wilcox, A., Levi, E.E. and Garrett, J.M. (2016) 'Predictors of non-attendance to the postpartum follow-up visit', *Maternal and Child Health Journal*, 20,S22-S27. Available at: https://doi.org/10.1007/s10995-016-2184-9 .