Summary: Logistic Regression and Extensions

Logistic regression, broadly known as the generalized linear models (GLM) is a technique for fitting a regression surface to data in which the dependent variable is a dichotomy. In the following sections, I will outline the critical parts that should be included in a manuscript when logistic regression is conducted.

Checking Data Structure & Statistical Model

The response variable should be dichotomous, multinominal, or ordinal. Normally, response variable of the logistic regression should be based on the true underlying binary phenomenon and the base-rate information should also be provided as it is critical for understanding the sample size and the study design.

The literature supporting design of the study should be reviewed and summarized. Excluding and including variables should be discussed with a background theory. As with any type of statistical modeling, careful examination of each predictor primarily on theoretical grounds should be supported and in the case of categorical predictors, the dummy variables should be stated.

The core elements of the logistic regression model as part of a generalized linear model should be discussed. The anticipated distribution of the response variable (binominal distribution, in this case), the linear component describing how a transformation of the expected value of the response variable can be written as a linear predictor based on a given collection of explanatory variables (logit link function), and also describe the specific choice of a link function specifying the connection between original and the transformed responses.

Sample pre-analysis

The details of the sample should be outlined. Parameter estimation for GLMs, and LR are based on maximum likelihood estimation, a large sample methodology. Thus, as their goodness-of-fit estimates are affected by design features including sample size, they must be outlined prior to the study. In addition, software packages are to be identified, considering the differences across software packages in estimation or statistical testing procedures and in how missing data is treated. Missing data are usually dealt with listwise deletion but at times missing data is a problem when missingness is not at random, and especially when it is related to the dependent variable. Therefore, the degree of unit non-response should be provided with a method used to adjust it. Overdispersion, referred to as extra-binomial variation is quite common in practice. They are often revealed by smaller standard errors for the coefficients than they should be, so adjusting the dispersion factor might be necessary.

Main Analysis

Prior to the analysis of the main model, the model and the data structure should be discussed. Specifically, the impact of outliers and extreme or unusual observations are to be clarified. Also, the linearity in the logit for continuous model predictors, outliers, location of zero-cells or evidence for separation should be outlined and adjusted to develop support for the validity of the model.

After assessing the model, the choice of the hypothesis testing for variable effects should be justified (likelihood ratio tests, score, Wald's $\chi 2$). Although most statistical packages report Wald Chi-square statistics for each predictor variable it could be problematic in small samples with multiple continuous predictors. Therefore, for these cases, Wald Chi-square statistics should be provided supplemented by the likelihood ratio tests in the manuscript.

The result of the analysis should be accompanied by the interpretation of parameter estimates for all variables and the interactions. In Logistic Regression, the unstandardized regression coefficient, is interpreted as the expected change in the log-odds of success given a one-unit change the predictor variable, controlling for other variables in the model. If the predictors are standardized, the choice for standardization of predictors (none, partial, full) should be explicated and justified, along with its limitation.

It is recommended to include multiple summary statistics of model fit beyond general likelihood test, including results from H-L test, model deviance, chi-square difference tests for goodness of fit and comparisons of competing models; and pseudo-*R*2 values. When it comes to reporting the pseudo-*R*2, the reason for choice needs to be explicated. Information criteria such as Akaike's information criterion (AIC) or Schwarz's Bayesian information criterion (BIC), is also reported to provide model fit information through different adjustments.

Classification table and associated statistics should be provided with categorical assessment of model fit. Authors should be cautious about providing the percent of correct predictions as it cannot be the only criteria for classification accuracy. Lastly the authors should explain whether the final model presented is credible or addresses the research questions through data and theory.

Literature Review 1

Cichocka, A., Górska, P., Jost, J. T., Sutton, R. M., & Bilewicz, M. (2018). What inverted U can do for your country: A curvilinear relationship between confidence in the social system and political engagement. *Journal of Personality and Social Psychology*, *115*(5), 883–902. https://doi.org/10.1037/pspp0000168

Cichocka et al., (2017) investigated political engagement and the tendency to justify the sociopolitical system. They hypothesized that the motivation for political engagement should be highest at intermediate levels of system confidence (the sense that the social system is familiar, safe and consensually embraced). In this review, we focused on one of their multiple studies (Study1) to evaluate whether their analysis based on logistic regression was sufficiently discussed in their paper.

First, the structure of the study, response variable, predictors and the theoretical basis was reviewed. Specifically, the authors included the nominal choice (three ordinal options including, "I would definitely participate", "I do not know", or "I would definitely not participate") as the response variable. Although this variable (political engagement) is not categorical in nature, voting behaviors occurs in real life and tend to be categorized. However, they did not include this explanation when discussing their specific response measure for the logistic regression. Nevertheless, the authors operationalized this voting intention of the parliamentary elections held the in real life to measure to political engagement, which I think has a great external validity. Also, they clearly stated with sufficient theoretical analysis why they wanted to examine the system confidence as the predictor variable (measured by "In general, our society is fair" etc.) for the dependent variable, political engagement. Additional control variable such as age, gender and education and political conservatism were mentioned to be included in the analysis, also with prior theoretical support. As the authors wanted to control these variables and ultimately see the unique effect of system confidence on political engagement, they included it in their model of analysis. Unfortunately, the coding method for the categorical variable was not clearly outlined for gender and the nominal response variable. Furthermore, although the researchers mentioned that they used a multinomial logistic regression, also assessing quadratic effects of system confidence on participants' intention to vote, they did not refer to the logit link function or the anticipated distribution. However, they still clarified why they used ordinal logistic regression for the ordinal response variable with detail added in the supplement material. In addition, they mentioned the data did not meet the parallel slopes assumption, leading them to employ a multinomial logistic regression instead.

Next, I aim to review whether there was critical analysis about the sample details. With specificity the authors outlined how they used existing domestic survey involving a large, nationwide, statistically representative sample of the Dutch referendum (979 participants in total). This satisfied their criteria in which they aimed to analyze data sets at least 173 participants, which would provide 80% power to detect small to medium effect size. However, it would have been better if base-rate information should also be provided as it is critical for understanding the sample size and the study design. Although additional information about software packages were left empty and did not mention how they treated missing data. Additional discussion on whether there was an overdispersion of the collected sample, which is quite common in practice was not included in the manuscript.

Lastly, in this section I aim to overview the main analysis conducted via logistic regression. Before the main analysis, outliers and multicollinearity should be checked and reported. As, multicollinearity among predictors inflate the standard errors for the estimated regression coefficients, it tends to affect the validity of statistical tests of these estimates. Even though, there were no specific analysis of multicollinearity, they provided the correlation matrix including the control variables. Also, they mentioned to mean-center the predictor variable prior to the analysis. Outliers and influential cases were said to be checked by using Cook’s distance. Other discussion about the logit linearity were left blank. Nevertheless, the specific method of hypothesis testing, interpretation of parameter, and the method for analyzing final and multiple models were provided. First, the parameters were all described along with their interpretation by the authors. However, it would have been better if the interpretation of the odd ratios was provided for the predictor variable. In addition, graphs of predicted probabilities along the level of a continuous independent variable was provided by the authors, thus enhancing the understanding of the readers. This is usually recommended for logistic regression with multiple qualitative independent variables.

The Results section should include interpretation of all effects identified as statistically or substantively significant, and the nature of interactive or polynomial effects should be identified and clarified. As the authors aimed to examine a curvilinear hypothesis in a nationally representative sample of Polish citizens, they used logistic regression to assess the linear and the quadratic relationship. Indeed, the authors highlighted that the analysis revealed a significant linear effect and the significant quadratic effects of system confidence on intentions to vote. Also, the authors followed the recommendation and helped the understanding of the logistic model using graphical methods to highlight he quadratic effect. Furthermore, the authors provided how they included variables in their final model and provided effect size such as Nagelkerke R-squared and -2 log likelihood to indicate model significance. Normally, researchers should be explicit in identifying the pseudo-*R*2 presented in applied manuscripts along with the reason. Even though the authors mentioned their choice of pseudo-R2, they did not justify the selection. Finally, the classification table and associated statistics were provided with categorical assessment of model fit. Along with the percent of correct predictions as it cannot be the only criteria for classification accuracy, they provided two models. Taken this altogether, the authors explained that the final model presented was credible through data and theory.

Literature Review 2

van Prooijen, J.-W., & Krouwel, A. P. M. (2020). Overclaiming Knowledge Predicts Anti-establishment Voting. *Social Psychological and Personality Science*, *11*(3), 356–363. https://doi.org/10.1177/1948550619862260

People often vote against the political establishment, and to find out why **Van Prooijen & Krouwel (2020)** investigated how overclaiming one's own knowledge predicts anti-establishment voting. In the context of a Dutch referendum on a European Union treaty with a clear pro- versus anti-establishment voting option, the authors conducted a logistic regression with a hierarchical procedure.

First, the structure of the study, response variable, predictors and the theoretical basis was reviewed. Specifically, the authors included the dichotomous choice (pro vs anti-establishment voting option) as the response variable. Because this variable is dichotomous in nature and occur in real life (political votes), it seemed to have a reasonable construct and was also possible for logistic regression analysis. Also, they clearly stated with sufficient theoretical analysis why they wanted to examine the knowledge overclaiming as the predictor variable (measured by self-perceived knowledge and actual knowledge) for the response variable, anti-establishment voting (in Step 2). Control variable such as age, gender and education were mentioned to be included in the analysis (in Step 1) based on prior studies. As the authors wanted to control the anti-establishment sentiment on the voting outcome, and find out the effect of knowledge overclaiming variable, they included it in the Step 1 control variable. In the research paper, the coding method for the categorical variable was clearly outlined such as gender (1 = man, 2 = woman), and the response variable (coded as 1 = voted against the treaty and 0 = voted in favor of the treaty). Furthermore, the researchers mentioned that they used a binary logistic regression with binomial distribution as the anticipated distribution of the response variable. Although they refer to the logit link function, the graphs provided the logistic regression slope.

Next, I aim to review whether there was critical analysis about the sample details. With specificity the authors outlined how they collected samples using the Dutch referendum (5,568 participants in total). Overall, the authors described how their large sample size provides more than 99% power for even very small effect sizes and therefore provided base for setting their level of significance (at .001). Although additional information about software packages were left empty, they tried to find out why certain participants dropped out in the second wave of data collection with an effort to avoid missing data not at random, by additional analysis in the supplemental material. There was no reference to the overdispersion of the sample, which is quite common in practice.

Lastly, this paragraph targets to overview the main analysis conducted by Van Prooijen & Krouwel (2020). Before the main analysis, outliers and multicollinearity should be checked and reported. As, multicollinearity among predictors inflate the standard errors for the estimated regression coefficients, it tends to affect the validity of statistical tests of these estimates. However, there were no specific analysis of multicollinearity, even though two independent variables were correlated with each other. Despite that, the specific method of hypothesis testing, interpretation of parameter, and the method for analyzing final and multiple models were provided. Primarily, parameters were all described along with their interpretation by the authors. The interpretation of the odd ratios was provided for each variable, thus enhancing comprehensibility of the research. The exponentiated regression coefficients can normally be interpreted as the odds ratio for the variable. In this research for example, the odds ratio of the main predictor variable (self-perceived understanding) indicated that for each point increase in anti-establishment sentiments, the chance for an anti-establishment vote becomes 1.62 times more likely. In addition, graphs of predicted probabilities along the level of a continuous independent variable was provided by the authors. This is usually recommended for logistic regression with multiple qualitative independent variables. Furthermore, the authors provided how they included variables in their final model and provided effect size such as Nagelkerke R-squared and -2 log likelihood to indicate model significance. Normally, researchers should be identifying the reason for the selection of the pseudo-*R*2 presented but they failed to justify the selection. Finally, the classification table and associated statistics were provided with categorical assessment of model fit. Along with the percent of correct predictions as it cannot be the only criteria for classification accuracy, they provided two models. Taken this altogether, the authors explained that the final model with coherent theory and supporting data analysis.