STAT499 WI25 DRP Final Report

In the Direct Reading Program this quarter, I explored a statistical technique called logistic regression, which is a statistical method primarily used for predicting binary outcomes. This approach is particularly engineered to overcome situations where linear regression is unsuitable due to its tendency to produce predictions outside the probability range of [0,1]. Logistic regression solves this by utilizing a logistic function, ensuring the output is always bounded between 0 and 1.

Unlike linear regression, which models continuous dependent variables, logistic regression is used for binary dependent variables (e.g., success/failure, yes/no), which is widely applicable in medical contexts. The logistic regression model estimates the probability of an seeing a positive response (Y=1) given a set of predictors (X). The logistic model employs a link function, typically the logit function, converting probabilities to log-odds. This allows us to interpret the coefficients as the impact of predictors on the odds of the outcome.

The parameters (beta coefficients) of logistic regression models are estimated using Maximum Likelihood Estimation (MLE). The likelihood function represents the probability of observing the given outcomes in the dataset for some specific parameter values. MLE seeks parameter values (betas) that maximize this likelihood by adjusting coefficients to achieve the highest possible product of probabilities across all observations. For each observation where the outcome is positive (Y=1), the likelihood includes the predicted probability from the model. For each negative observation (Y=0), the likelihood includes the probability that the model predicts Y=0. Therefore, optimizing this likelihood function ensures that the resulting parameters provide the best fit to the observed data.

Throughout the quarter, our practical application involved analyzing a dataset of heart failure clinical records, aiming to identify key predictors of death event. Initially considering multiple predictors, our regression analysis utilized forward and backward selection methods to identify serum creatinine and ejection fraction as significant predictors. The analyses demonstrated that higher serum creatinine levels significantly increased mortality risk, with each unit increase raising the odds of death by approximately 2.23 (log-odds increase of 0.8). Conversely, a higher ejection fraction was associated with reduced mortality risk, decreasing the odds by roughly 5.45 per unit increase (log-odds decrease of -0.056). We also compared the probability graphs for serum creatinine and ejection fraction using linear regression and logistic regression. While the graphs using linear regression often show out of bound probabilities, the graphs using logistic regression appear to be a better fit for both predictors.

Overall, logistic regression provided a robust analytical approach suited for dataset that has a binary outcome, enhancing the interpretability and effectiveness in medical predictive modeling.

References

James, Gareth, et al. An Introduction to Statistical Learning: With Applications in R. Springer, 2013.

McCullagh, P. Generalized Linear Models. CRC Press LLC, 1989. ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/washington/detail.action?docID=5631551. Accessed 22 Jan. 2025.