Implementing Binary Regression in R: An Application to Environmental Data

Juhnar Esmeralda, Abdullah Al-Mamun and Evrim Oral*

LSU Health NEW ORLEANS School of Public Health

Louisiana State University Health Sciences Center School of Public Health, Biostatistics and Data Science Program

Contact author: eoral@lsuhsc.edu

Background

The likelihood equations in binary regression cannot be solved explicitly, they require iterative methods such as the Newton-Raphson algorithm to obtain solutions. However, such iterative algorithms present drawbacks, including the risk of converging to local maxima rather than the global maximum, sensitivity to the initial parameter values chosen and computational intensity, particularly for high-dimensional problems. While the iterative methods remain widely employed for optimization problems, it is essential to acknowledge these potential issues and explore alternative algorithms or techniques. As a solution, Tiku and Vaughan proposed an alternative method in 1997.

R-Package BinRegMMLE

We used RStudio to develop our package. We created a project in R Studio using the devtools library, and proceeded to develop R-script files. Utilizing devtools, we built the package and created documentation to outline its properties. The documentation contains specific examples to guide future users of the package we have developed. After that, we installed the package locally using devtools::install() and conducted testing with example data sets. To facilitate sharing with the public and maintain controlled versions, we established a GitHub account and created a repository where the project can be regularly committed and pushed. Finally, we utilized the developed package to analyze environmental data, and compared the results obtained with our package against those derived from SAS.

Methods

Following Tiku and Vaughan's procedure in arriving at covariate estimates ($\hat{\gamma}_0$ and $\hat{\gamma}_1$), and in conjunction with the framework for R and Github specified below (Figure 1), we implemented a function whose iterative algorithm follows the following sequence of derivations:

1. Identification of the distribution associated to the link function
2. Computation of $t_{(i)} = F^{-1}\left(\frac{i}{n+1}\right)$
3. Fitting of a linear regression and obtention of initial estimates following the rationale that $\pi(x_i) = F(z_i) = \int_{-\infty}^{z_i} f(u)\,du \ \ni z_i = \gamma_0 + \gamma_1 x_1$
4. Computation of $z_{(i)} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{[i]}$ using Tiku's MMLE formulas
5. Order the data by $z_{(i)}$.
6. Repeat steps 4 and 5 until difference of $\hat{\gamma}_0$ and $\hat{\gamma}_1$ in their successive values are negligible (tuning of the estimates).

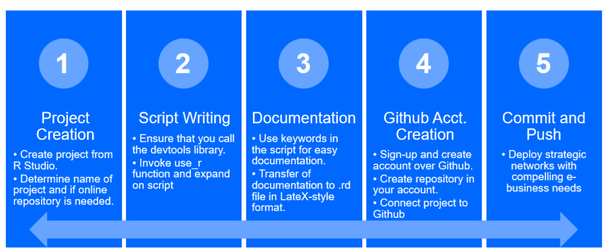

Fig 1. Framework of Package Implementation and Update through Github

Results

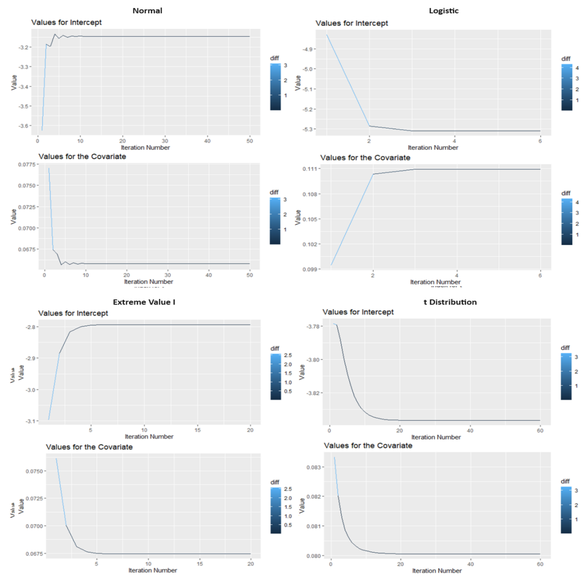

Fig 2. Convergence plots showing progression of covariate values across iterations on the following four Distributions: Normal (upper left), Logistic (upper right pane), Extreme Value I (lower left pane), and t-distribution (lower right pane)

Application: Cancer incidence

We used our R package to estimate the impact of air quality on the cancer incidence rate in Louisiana and compared it with SAS output for validation purposes. To do this, we first categorized cancer incidence rates into two groups based on whether they were above or equal to/below the national average incidence rates. We then used this variable as the outcome, while the air quality index, also grouped into two categories based on median value (high exposed vs low exposed) was used as the predictor.

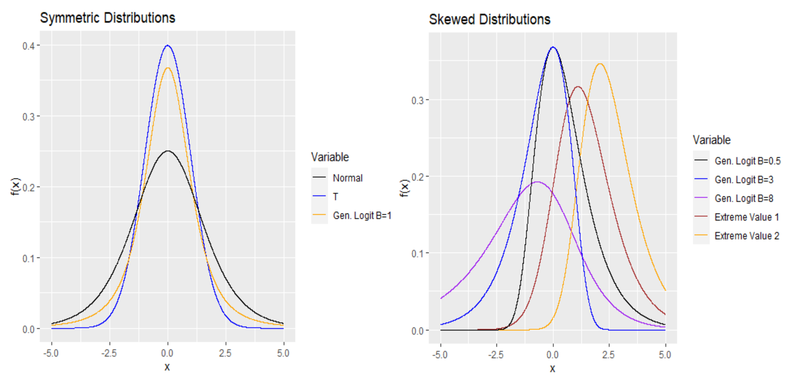

Figure 3. Distribution of different link functions included in the R package. The package covers not only symmetric distributions (left) but also skewed distributions (right).

Application (Cont)

We considered several different link functions and compared the R package results with the results obtained from SAS.

The results are provided below. Realize that SAS does not have available link functions other than logit, probit and complementary log log.

Type	Distribution	Link Function Name	Parameter	SAS Estimates (SE)	MMLEs from R-package (SE)	p-value from MMLE
SYMMETRIC	Logistic	Logit (inverse CDF of Logistic dist.)	Intercept	0.9731 (0.103)	0.9731 (0.103)	.
			Slope	-0.28 (0.1452)	-0.28 (0.1452)	0.0539
	Normal	Probit (inverse CDF of stand. Normal dist.)	Intercept	0.6 (0.0615)	0.6 (0.0615)	.
			Slope	-0.1692 (0.0877)	-0.1692 (0.0877)	0.0537
	Student's t	NA (Inverse CDF of t dist.) df, v = 3	Intercept	NA	0.6741 (0.0739)	.
			Slope	NA	-0.1983 (0.1029)	0.0534
SKEWED	Extreme Value I	Loglog (Inverse CDF of Extreme Value I dist.)	Intercept	NA	1.1377 (0.0881)	.
			Slope	NA	-0.2349 (0.1219)	0.0534
	Extreme Value II	cloglog (Inverse CDF of Extreme Value II dist.)	Intercept	0.2575 (0.0578)	0.2575 (0.0578)	.
			Slope	-0.1634 (0.0848)	-0.1634 (0.0848)	0.0540
	Generalized Logistic	NA (Inverse CDF of Generalized Logistic dist.) Shape parameter, b = 3	Intercept	NA	2.1824 (0.0929)	.
			Slope	NA	-0.2494 (0.1293)	0.0537
	Generalized Logistic	NA (Inverse CDF of Generalized Logistic dist.) Shape parameter, b = 0.56	Intercept	NA	0.258 (0.1157)	.
			Slope	NA	-0.3189 (0.1655)	0.0540
	Generalized Logistic	NA (Inverse CDF of Generalized Logistic dist.) Shape parameter, b = 8	Intercept	NA	3.197 (0.0899)	.
			Slope	NA	-0.2403 (0.1246)	0.0534

Discussions and Recommendations

The R package developed made use of the coronary heart disease data as referenced by Tiku and Vaughan (1997) to examine the convergence of covariates, β_{1i} and β_{2i}. The resulting values appear to approach rapidly to the true value in no more than 5 iterations in each of the four distributions.

The advantage of our R package compared to its rivals lies in its availability to include non-traditional link functions. Researchers can use the Deviance value to select the best fit to their model using several different link functions.

Conclusions and Future Work

We developed a new R-package designed to handle binary outcomes, offering greater flexibility with available link functions compared to some existing software. We demonstrated its utility through a real-life example from environmental data.

Our next step is extending the package to the case where there are multiple covariates in the model.

References

Tiku, M.L. and Vaughan, D.C. (1997), Logistic and Nonlogistic Density Functions in Binary Regression with Nonstochastic Covariates. Biom. J., 39: 883-898. https://doi.org/10.1002/bimj.4710390802