

Package CPMCGLM: Correction of significance level after multiple coding of explanatory variable in generalized linear model.

Jérémie Riou ^{a,b} & Benoît Liquet ^b

^a*Danone Research, Clinical Studies Plateform, Centre Daniel Carasso,
RD 128 - Avenue de la Vauve, 91767 Palaiseau Cedex, FRANCE*

^b*Centre de Recherche INSERM U897, Université Bordeaux 2, ISPED,
146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE*

Plan

① Introduction

Context

② Methods

R function

Statistical context

Correction methods

③ Results

④ Discussion

Introduction

- In epidemiology, a current practice is to transform an explanatory quantitative variable in categorical variable ;
- Either the thresholds are scientifically recognized or they could be defined.
- No Correction of Type-I error rate in the Multiple Testing Context implies an Overestimation of the effect ;
- **How can we correct the Type-I error in this situation ?**

PAQUID Example (1)

- **Aim :** Association study between dementia and rate of HDL cholesterol (Bonarek,2000).
- **Method :** Logistic regression
 - Variable to explain : Dementia (0 :Dement, 1 :No dement) ;
 - Explanatory variable : rate of HDL cholesterol (Hight Density Lipoprotein) (mmol/L) ;
 - Adjustment variables : Age, Sex, Academic level, Wine consumption.

PAQUID Example (2)

- The scientific choice is a transformation of the explanatory variable in categorical variable in order to facilitate the interpretation ;
- **Problem : What is the best coding for the HDL variable ?**

Aims :

- ① Determine the coding which represents the most significance association between outcome and the interest explanatory variable ;
- ② Correct the signification degree of one test in the context of multiple coding of one explanatory variable in a generalized linear model.

The R function

- Creation of only one function :
`CPMCGLM(formula, family, link, data, varcod,
dicho, nb.dicho, categ, nb.categ,
boxcox, nboxcox, N=1000, cutpoint)`
- The arguments of the function are detailed in Methods.

Model

- Consider a generalized linear model (McCullagh Nelder, 1989) with p explanatory variables.
- The canonical parameter θ_i is specified by :

$$\theta_i = Z'_i \gamma + X_i \beta; \quad (1)$$

where X_i is the explanatory variable of interest , Z_i the vector of explanatory adjustment variables.

- For the example the formula of the logistic regression is :

$$\text{logit}[P(Y_i = 1|X_i, Z_i)] = Z'_i \gamma + X_i \beta. \quad (2)$$

Model : Function arguments

- formula :
$$\text{formula} = Y \sim X+Z$$
- family, link :
 - ① Linear regression : family = gaussian, link = identity
 - ② Logistic regression : family = binomial, link = logit
 - ③ Probit regression : family = binomial, link = probit
 - ④ Poisson regression : family = poisson, link = log
- data : Needs to be a data.frame
- varcod : Needs to be a continuous argument

Binary Coding

$$\text{logit}[P(Y_i = 1|X_i, Z_i)] = Z'_i \gamma + X_i(k)\beta^1(k)$$

$$\left\{ \begin{array}{l} \text{if } X_i > c_k \text{ then } X_i(k) = 0 \\ \text{else } X_i(k) = 1 \end{array} \right.$$

- **Tests :**

- Hypotheses for one transformation k :
 $\mathcal{H}_0(k) : \beta^1(k) = 0$, versus $\mathcal{H}_1(k) : \beta^1(k) \neq 0$.
- Score test statistic :

$$T(k) \sim \mathcal{N}(0, 1)$$

- Liquet & Commenges (2005) defined the correlation between two tests $T(k)$ and $T(l)$.

BoxCox transformations

$$X(k) = \begin{cases} \lambda_k^{-1}(X^{\lambda_k} - 1) & \text{if } \lambda_k > 0 \\ \log X & \text{if } \lambda_k = 0, \end{cases}$$

- Hypotheses for one transformation k :
 $\mathcal{H}_0(k) : \beta^1(k) = 0$, versus $\mathcal{H}_1(k) : \beta^1(k) \neq 0$.
- Score test statistic :

$$T(k) \sim \mathcal{N}(0, 1)$$

- The definition of the correlation between tests is valid here.

Function arguments

- dicho : A vector with the order of the quantiles which are used for computing the cutpoint of each dichotomous transformation ;
- nb.dicho : The number of dichotomous transformations using quantiles ;
- boxcox : A vector of λ parameters corresponding to each BoxCox transformation ;
- nboxcox : The number of Boxcox transformations. Maximum=5, Strategy : (1,0,2,0.5,1.5).

Coding in more than two classes

$$\text{logit}[P(Y_i = 1|X_i, Z_i)] = Z'_i \gamma + X_i^1(k)\beta^1(k) + \dots + X_i^{m-1}(k)\beta^{(m-1)}(k)$$

- Hypotheses for the k^{th} test :
 $\mathcal{H}_0(k) : \beta^1(k) = \beta^2(k) = \dots = \beta^{m-1}(k) = 0,$
 versus $\mathcal{H}_1(k) : \exists \eta \in \{1, \dots, m-1\} \setminus \beta_\eta(k) \neq 0.$
- Score test statistic : $T(k) \sim \chi_{(m-1) \text{ df}}^2;$
- The correlation between two tests $T(k)$ et $T(l)$ has been defined, here.

Categorical transformations : Function arguments

- nb.categ : the number of categorical transformations using quantiles ;
- categ : a matrix with the order of quantiles which are used for computing the categorical cutpoints of each transformation ;
- cutpoint : a matrix with the values of cutpoints which are used for each transformation.

Example of Matrix Definition :
$$\begin{pmatrix} 7 & 13 & NA \\ 5 & 10 & NA \\ 5 & 10 & 15 \end{pmatrix}$$

Exact method

- **Exact method :**

$$\begin{aligned} p_{value} &= P(T_{max} > t_{max}) \\ &= 1 - P(|T_1| < t_{max}, \dots, |T_{max}| < t_{max}) \end{aligned}$$

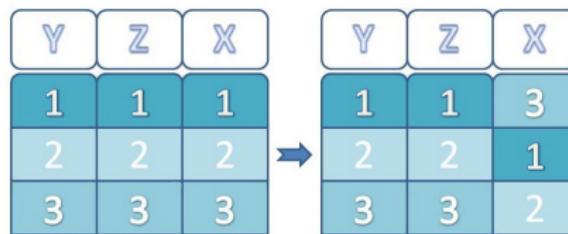
- For the binary coding, the asymptotic distribution of $\{T_1, \dots, T_{max}\}$ follows a $MVN(0, \Sigma)$, where Σ could be estimated. By using a numerical integration, we can estimate asymptotically the p_{value} (`pnorm()` with **R**), (Liquet & Commenges (2005)).
- No solution for a coding in more than two classes, because we don't know the multivariate χ^2 distribution.

Parametric Bootstrap method

- Computation of T_{max}^{init} on the real dataset ;
- Computation of the model coefficients under \mathcal{H}_0 ;
- Resampling for 1 bootstrap sample :
 - Simulation of Y_i^* variable with initial coefficients ;
→ We obtain a new sample for each resampling : (Y_i^*, Z_i, X_i)
- Computation of T_{max}^* of each bootstrap sample ;
- $p_{value} = \text{rate of } T_{max}^* \text{ upper than } T_{max}^{init}$.

Permutation method

- Computation of T_{max}^{init} on the real dataset ;
- Creation of N samples with permutation :



- Computation of T_{max}^* of each permutation sample ;
- p_{value} = rate of T_{max}^* upper than T_{max}^{init} .

Example : Function Outcome(1)

```
Call: CPMCGLM(formula = Dementia ~ Age + Sex + HDL  
+ Acadlevel + Wine, family = "binomial",  
link = "logit", data = datad, varcod = "HDL",  
nb.dicho = 7, categ = matcat, nboxcox = 5, N = 1000)
```

Generalized Linear Model Summary

Family: binomial

Link: logit

Number of subject: 334

Number of adjustment variable: 4

Resampling

N: 1000

Example : Function Outcome(2)

Best coding

Method: Dichotomous transformation

Value of the order quantile cutpoints: 0.75

Value of the quantile cutpoints: 1.6025

Corresponding adjusted pvalue:

Adjusted pvalue

naive 0.007

bonferroni 0.140

bootstrap 0.038

permutation 0.038

exact: Correction not available for these codings

Discussion

- **Key message :** In this context, it is important to correct the p_{value} for the interpretation of the results ;
- **Contribution :** Decrease the bias of the results ;
- **Function :** Easy to use ;
- **Accessibility :** Package available on the CRAN.