Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models.

Benoit Liquet, Jérémie Riou

Liquet and Riou *BMC Medical Research Methodology* 2013, **13**:75
http://www.biomedcentral.com/1471-2288/13/75

BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models

Benoit Liquet[1,2,3] and Jérémie Riou[1,2,4*]

Abstract

Background: In statistical modeling, finding the most favorable coding for an exploratory quantitative variable involves many tests. This process involves multiple testing problems and requires the correction of the significance level.

Methods: For each coding, a test on the nullity of the coefficient associated with the new coded variable is computed. The selected coding corresponds to that associated with the largest statistical test (or equivalently the smallest p_{value}). In the context of the Generalized Linear Model, Liquet and Commenges (Stat Probability Lett,71:33–38,2005) proposed an asymptotic correction of the significance level. This procedure, based on the score test, has been developed for dichotomous and Box-Cox transformations. In this paper, we suggest the use of resampling methods to estimate the significance level for categorical transformations with more than two levels and, by definition those that involve more than one parameter in the model. The categorical transformation is a more flexible way to explore the unknown shape of the effect between an explanatory and a dependent variable.

Results: The simulations we ran in this study showed good performances of the proposed methods. These methods were illustrated using the data from a study of the relationship between cholesterol and dementia.

Conclusion: The algorithms were implemented using R, and the associated CPMCGLM R package is available on the CRAN.

Keywords: Bonferroni procedure, Generalized linear model, Multiple coding, Parametric bootstrap, Permutation, p_{value}, Resampling procedure

Background

In applied studies, the relationship between an explanatory and a dependent variable is routinely measured using a statistical model. For instance, in epidemiology it is quite common that a study focuses on one particular risk factor. The scientific problem is to analyze whether this risk factor has an influence on the risk of occurrence of a disease, a biological trait , or another outcome. To answer to this question, a regression model is often used in which the risk factor will be represented by a continuous X, allowing adjustment on $p-1$ known risk factors of the studied trait. However, the form of the effect (or the dose-effect relationship) is not known in advance, and as such, the continuous variable X is often transformed, typically into categorical variables, by grouping values into two or more categories. An example of this is seen in an *The American Journal of Epidemiology* (October 2009, volume 170, number 8), where four of six papers with continuous exposure used categorization, and only two kept the variable as continuous [1].

*Correspondence: jeremie.riou@isped.u-bordeaux2.fr
[1]University Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France
[2]INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France
Full list of author information is available at the end of the article

Binary coding is often used in epidemiology, either to make interpretation easier, or because a threshold effect is suspected. In a regression model with multiple explanatory variables, the interpretation of the regression coefficient for a binary variable may be easier to understand than a change in one unit of the continuous variable. Dichotomous transformations of a variable X are defined as:

$$X(k) = \begin{cases} 1 & \text{if } X \geq c_k \\ 0 & \text{if } X < c_k \end{cases}$$

Other transformations are also used, in particular Box-Cox transformations which have been defined as:

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

but the choice of the transformation is often subjective. The arbitrariness of the choice of cutpoints may lead to the idea of trying more than one set of values. Hence to analyze data, the statistician may have to use several transformations, and for each the statistician applies a test for "$\beta = 0$" (where β is the coefficient representing the effect of the risk factor of interest). The most favorable transformation is then chosen. The cutpoint giving the minimum p_{value} is often termed "optimal" [2,3]. When testing several codings of a variable, there is a problem with the multiplicity of tests performed, leading to an incorrect p_{value} and possible overestimation of effects [4]. Generally, researchers fail to consider this problem and do not correct the significance level in relation to the number of tests performed [3], which can lead to an increase in the Type-I error [5]. The p_{value} should thus be corrected to take into account the multiplicity of tests.

In many cases, it is now widely recognized that categorization of a continuous variable could introduce major problems to an analysis and interpretation of the associated model [1,3]. It is important to note that the aim of this paper is not to defend this practice, but to improve a practice commonly used by epidemiologists in terms of multiple testing. Furthermore, despite known loss of power following dichotomization in the univariate case, Westfall [6] shown that dichotomizing continuous data can greatly improve the power when multiple comparisons are performed.

Many methods of correction exist, the most simple and well known being the Bonferroni rule. Several authors have improved this method to make it more powerful, however most do not take into account the correlation between the tests [7-11]. If the tests are independent, or moderately dependent, then they provide an upper bound which may be satisfactory. Efron [12] proposed a correction that account for the correlation between two consecutive tests if there is a natural order between the tests, with high correlation between adjacent tests. Liquet and Commenges [13,14] and Hashemi and Commenges [15] proposed a more exact correction, accounting for the whole correlation matrix, for score tests obtained in logistic regression, generalized linear model and proportional hazards models.

Here, we proprose extending these studies to a categorical transformation (with $m > 2$ categories) of the continuous variable by involving more than one parameter in the model; $m-1$ dummy variables are introduced in the model. The categorical transformation is a more flexible way to explore the unknown shape of the effect. In this context, we propose a method and an R program based on resampling approaches to determine the significance level for a series of several transformations (including dichotomous, Box-Cox and categorical transformations) of an explanatory variable in a Generalized Linear Model. The problem of correcting the estimation of the effect will not be examined here.

First, we revisit the example proposed by Liquet and Commenges [14] on the relationship between cholesterol and dementia [16] to provide a framework for our discussion. In section 'Methods: Statistical context', we present the statistical contexts relating to multiple testing; the model, the maximum test and the minimum p_{value} procedure and finally the score tests are exposed. Section 'Methods: Significance level correction' presents the different methods of correction of the Type-I error. A simulation study for the different strategies of coding, and application of the model to the initial example are presented in the section 'Results'. Concluding remarks are given in the two last sections.

Example: revisiting the PAQUID cohort example

We revisited the example presented in the article of Liquet and Commenges [13] for a coding of a binary variable in a logistic regression. This example is based on the work of Bonarek et al. [16], who studied the relationship between serum cholesterol levels and dementia. The data came from a nested case-control study of 334 elderly French subjects aged 73 and over who participated in the PAQUID cohort (37 subjects with dementia and 297 controls). The variables age, sex, level of education and wine consumption were considered as adjustment variables. The analysis focused on the influence of HDL-cholesterol(high-density lipoprotein) on the risk of dementia. Bonarek et al. [16] first considered HDL-cholesterol as a continuous variable; then, to ease clinical interpretation, they chose to transform the HDL-cholesterol into a categorical variable with four classes. Finally, as there was no significant difference between the first three quartiles, HDL-cholesterol was split into two categories with a cutpoint at the last quartile. The best p_{value}, 0.007, was obtained in the latter analysis

and was selected for interpretation. However, this p_{value} did not take into account the numerous transformations performed to determine the best representation of the variable of interest. Legitimate questions arising from this include the following: What is the real association between dementia and HDL-cholesterol, with a correction of the Type-I error? Is it really significant? Liquet and Commenges [14] proposed correcting the p_{value} associated with multiple transformation including dichotomous and Box-Cox transformation, however, their method cannot be used with categorical transformation.

Methods

Statistical context

Model

Let us consider a Generalized Linear Model with p explanatory variables [17], where Y_i $(1 \leq i \leq n)$ are independently distributed with probability density function in the exponential family defined as follows:

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}; \quad (1)$$

with $\mathbb{E}[\,Y_i] = \mu_i = b'(\theta_i), \mathbb{V}ar[\,Y_i] = b''(\theta_i)a(\phi)$ and where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known and differentiable functions. $b(\cdot)$ is three times differentiable, and its first derivative $b'(\cdot)$ is invertible. Parameters (θ_i, ϕ) belong to $\Omega \subset \mathbb{R}^2$, where θ_i is the canonical parameter and ϕ is the dispersion parameter.

In this context, we wished to test the association between the outcome Y_i and explanatory variable of interest X_i, adjusted on a vector of explanatory variables $\mathbf{Z_i}$. The form of the effect of X_i is unknown, so we may consider K transformations of this variable $\mathbf{X_i(k)} = g_k(X_i)$ with $k = 1, \ldots, K$.

For example, if we transform the continuous variable in m_k classes, $m_k - 1$ dummy variables are defined from the function $g_k(\cdot)$: $\mathbf{X_i(k)} = g_k(X_i) = (X_i^1(k), \ldots, X_i^{m_k-1}(k))$. Different numbers of level m_k of the categorical transformation are possible.

The model for a transformation k can then be obtained by modeling the canonical parameter θ_i as:

$$\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i} + \boldsymbol{\beta_k}\mathbf{X_i(k)}, \;\; i = 1, \ldots, n;$$

where $\mathbf{Z_i} = (1, Z_i^1, \ldots, Z_i^{p-1})$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_{p-1})^T$ is a $p-1$ vector of coefficients, and $\boldsymbol{\beta_k}$ is the $m_k - 1$ vector of coefficients associated with a categorical transformation k of the variable X_i. For dichotomous or Box-Cox transformations $\boldsymbol{\beta_k}$ reduce to a scalar ($\boldsymbol{\beta_k} \in \mathbb{R}$).

The hypothesis of the test for the transformation k is defined as follows:

$$\mathcal{H}_0(k) : \boldsymbol{\beta_k} = \mathbf{0}_{m_k-1} \;\text{ versus }\; \mathcal{H}_1(k) : \boldsymbol{\beta_k} \neq \mathbf{0}_{m_k-1},$$

where $\mathbf{0}_{m_k-1}$ is a null vector of dimension $m_k - 1$. Under the null hypothesis $\mathcal{H}_0(k)$ we have $\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i}$, which do not depend on k. Thus all the null hypotheses are the same, and denote it by $\mathcal{H}_0$.

Maximum test and minimum P-value procedures

For each coding, k, of the variable X_i, a test statistic T_k is performed on the nullity of the vector $\boldsymbol{\beta_k}$. We then have a vector of test statistics $\mathbf{T} = (T_1, \ldots, T_K)$ for the same null hypothesis (no effect of the risk factor of interest). In the context of dichotomous and Box-Cox transformations, each test statisitic, T_k, has asymptotically, a standard normal distribution. Thus rejecting the null hypothesis if one of the absolute values of the test T_k is larger than a critical value c_α, is equivalent to rejecting the null hypothesis if $T_{max} > c_\alpha$ where $T_{max} = max(|T_1|, \ldots, |T_K|)$. To cope with the multiplicity problem, Liquet and Commenges [13,14] proposed that the probablity of Type-I error for the statistic T_{max} under the null hypothesis be computed as:

$$p_{value} = P(T_{max} \geq t_{max}) = 1 - P(|T_1| < t_{max}, \ldots, |T_{max}| < t_{max},) \quad (2)$$

where t_{max} is the realization of T_{max}.

An equivalent approach is to use a procedure based on the individual p_{value} of each test T_k noted $P_k = P(|T_k| > |t_k|)$ (where t_k is the realization of T_k). The minimum of the K realized p_{value} corresponds to the test k which obtains the highest realization (in absolute values; $k/\ t_{max} = |t_k|$). Then, we have:

$$p_{value} = P(P_{min} \leq p_{min}) \quad (3)$$

where $P_{min} = min(P_1, \ldots, P_K)$ and p_{min} is the realization of P_{min}. The interest of using a procedure based on the p_{value} is the possibility of combining statistical tests which do not follow the same distribution. In the current context, we will combine dichotomous, Box-Cox and categorical transformations with more than two levels.

Score test

We briefly present the score test used for all of the K transformations where the same null hypothesis is tested (*i.e.* $\mathcal{H}_0$: "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k-1}$" given by $\theta_i(k) = \boldsymbol{\gamma}\mathbf{Z_i}$ (with different alternatives)). We present the main results obtained by Liquet and Commenges [14] for the Generalized Linear Model in the context of dichotomous and Box-Cox transformations, and then consider the score test for categorical transformations.

Dichotomous and Box-cox Transformations In the context of dichotomous and Box-Cox transformations, the score test used for testing the effect of the transformed

variable ($\boldsymbol{\beta_k} = 0$ with $\boldsymbol{\beta_k} \in \mathbb{R}$) follows asymptotically a standard normal distribution:

$$T_k = \frac{X(k)^T \hat{R}}{\sqrt{X(k)^T (I - H) V X(k)}}$$

where $\hat{R}$ is the vector of residuals $\hat{R}_i = Y_i - \hat{\mu}_i$ computed under the null hypothesis, V is a diagonal matrix such that $v_{ii} = \hat{Var}(Y_i)$, $H = VZ(Z^T VZ)Z^T$, and Z the $n \times p$ matrix with rows $\mathbf{Z_i}$, $i = 1, \ldots, n$.

The correlation between the different tests has been defined by Liquet and Commenges [14]. Asymptotically, the joint distribution of $T_1, \ldots, T_K$ is a multivariate normal distribution with zero mean and a certain covariance matrix. Thus Liquet and Commenges [14] propose that the p_{value} (associated with the test T_{max}) defined in (2) using numerical integration [18] be calculated. They called their method the "exact method".

Categorical transformations In the context of a categorical transformation in m_k classes, the score test testing $\mathcal{H}_0$: "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k - 1}$" (with $\boldsymbol{\beta_k} \in \mathbb{R}^{m_k - 1}$) follows asymptotically a χ^2 distribution with $m_k - 1$ degrees of freedom and is defined as:

$$T_k = U_k^T I_k^{-1} U_k;$$

where U_k and I_k are respectively the score function and the Fisher information matrix under the null hypothesis [19]. To compute the p_{value} defined in (3), it is necessary to know the joint distribution of $\mathbf{T} = (T_1, \ldots, T_K)$. Some studies have defined the distribution of the multivariate χ^2 [20,21]. However, even though the correlation between the different tests could be easily estimated, it has not been possible, as far as we know, to obtain the joint distribution of $\mathbf{T} = (T_1, \ldots, T_K)$. To overcome this problem, we propose approximating the p_{value} (defined in (3) by the minimum p_{value} procedure) using a resampling method (defined in the next section) which also accounts for the correlation between the test statistics.

Significance level correction

Bonferroni method

One of the most common corrections in multiple testing is the Bonferroni method. It has been described by several authors in various applications [7,11,22]. It allows an upper bound of the significance level of the minimum p_{value} procedure to be computed as:

$$p_{value} = P(P_{min} \leq p_{min}) \leq K \times p_{min}$$

where K is the number of tests. This method is very simple and does not require any assumption about the correlation between the different tests. It can therefore be applied directly to the different possible codings of an explanatory variable. However, this only provides an upper bound of the p_{value}, which may be very conservative if the correlation between tests are high and the number of transformation are large.

Resampling based methods

We propose the use of resampling based methods [23,24] with the aim of building a reference distribution for the test statistics. These procedures have the advantage of taking into account the dependence of the test statistics for evaluating the correct significance level of the minimum p_{value} procedure (or the maximum test procedure). The principle of resampling procedures is to define new samples from the probability measure defined under $\mathcal{H}_0$: "$\boldsymbol{\beta_k} = \mathbf{0}_{m_k - 1}$".

Permutation test procedure Permutation methods can be used to construct tests which control the Type-I error rate [25]. In our context, the algorithm of the permutation procedure is defined as follows:

1. Apply the minimum p_{value} procedure to the original data for the K transformations considered. We note p_{min} the realization of the minimum of the p_{value};
2. As under $\mathcal{H}_0$, the X_i variable has no effect on the response variable, a new dataset is generated by permuting the X_i variable in the initial dataset;
3. Generate B new datasets $s_b^*, b = 1, \ldots, B$ by repeating B times the step 2;
4. For each new dataset, apply the minimum p_{value} procedure for the transfomation under consideration. We note p_{min}^{*b} the smallest p_{value} for each new dataset.
5. The p_{value} defined in (3) is then approximated by:

$$\widehat{p_{value}} = \frac{1}{B} \sum_{b=1}^{B} I_{\left\{p_{min}^{*b} < p_{min}\right\}},$$

where $I_{\{.\}}$ is an indicator function.

However, it is important to note that exchangeability need to be satisfied [25-30]. This condition is much more restrictive than it appears at first sight. In fact, Commenges [29] and Commenges and Liquet [25] showed that the permutation test approach for the score test is robust if the model has only one intercept under the null hypothesis, or if $\mathbf{X_i}$ are independent of Z_i for all i in the context of a linear model and the proportional hazards model. This issue applies in our context. Thus we investigated, the robustness of the permutation method when the exchangeability assumptions is violated.

Parametric bootstrap procedure In 2000, Good [31] explained: "Permutations test hypotheses concerning distributions; bootstraps test hypotheses concerning parameters. As a result, the bootstrap implies less stringent assumptions". Therefore, an alternative way may be to use

resampling method based on bootstrap [32], which give us an asymptotic reference distribution. This procedure could be defined by the following algorithm:

1. Apply the minimum p_{value} procedure to the original data for the K transformations being considered. We note p_{min} the realization of the minimum of the p_{value};
2. Fit the model under the null hypothesis, using the observed data, and obtain $\hat{\gamma}$, the maximum likelihood estimate (MLE) of γ;
3. Generate a new outcome Y_i^* for each subject from the probability measure defined under $\mathcal{H}_0$. For example, for a logistic model (where $a(\phi) = 1$, $b(\theta_i) = log(1 + e^{\theta_i})$, and $\mu_i = \mathbb{E}(Y_i) = e^{\theta_i}/(1 + e^{\theta_i})$), we generate Y_i^* according to:

$$P(Y_i^* = 1|\mathbf{Z_i}) = \frac{e^{\hat{\gamma}\mathbf{Z_i}}}{1 + e^{\hat{\gamma}\mathbf{Z_i}}}.$$

   Repeat this for all the subjects to obtain a sample noted $s^* = \{Y_i^*, \mathbf{Z_i}, X_i\}$
4. Generate B new datasets $s_b^*, b = 1, \ldots, B$ by repeating B times the step 3;
5. Apply for each new dataset, the minimum p_{value} procedure for the transformation considered. We note p_{min}^{*b} the smallest p_{value} for each new dataset.
6. Then, the p_{value} defined in (3) is then approximated by:

$$\widehat{p_{value}} = \frac{1}{B}\sum_{b=1}^{B} I_{\left\{p_{min}^{*b} < p_{min}\right\}}.$$

Results

Simulation study

The aim of this simulation study was to assess the performance of the two resampling methods to correct the significance level. Three different scenarios of transformations were investigated: dichotomous transformations, categorical transformations with three classes, and categorical transformations with different numbers of classes. To shorten the simulation study section we have not presented the results for the Box-Cox transformations. For each simulation case, the control of the Type-I error and the power of the developed methods were evaluated. For all simulations, the data come from a logistic model (where $a(\phi) = 1$, $b(\theta_i) = log(1 + e^{\theta_i})$, and $\mu_i = \mathbb{E}(Y_i) = e^{\theta_i}/(1 + e^{\theta_i})$) consisting of two explanatory variables: Z, an adjustment variable, and X, the variable of interest. We considered the following models:

$$Logit(P(Y_i = 1|Z_i, X_i(k))) = \theta_i(k) = \gamma_0 + \gamma Z_i + \boldsymbol{\beta}\mathbf{X}_i(k); \quad (4)$$

where Z_i and X_i are independent and were generated according to a standard normal distribution and the vector $\mathbf{X}_i(k)$ was a transformation of a continuous variable X_i. The sample size was set to be 100. We used 1000 replications for each simulation and 1000 samples for the resampling methods.

Dichotomous transformations

We only considered dichotomous transformations to explore a shape effect of the variable of interest. To obtain the best transformation, several cutpoints c_k may be tested. When epidemiological references are not available, a strategy based on the quantile of the continuous variable is most commonly applied. In this simulation we used the median for one dichotomous transformation. For two dichotomous transformations we used the first tercile as the first cutpoint, and the second tercile as the second cutpoint, and so on. This strategy is summarized in Table 1.

Firstly, we investigated the Type-I error rate. For a replication, the rejection criterion of the null hypothesis ($\boldsymbol{\beta_k} = 0$) was a p_{value} less than 0.05. Thus, for a simulation of 1 000 replications, the empirical Type-I error rate was the proportion of tests where the p_{value} was less than 0.05. Figure 1(a) shows the evolution of the Type-I error rate for dichotomous transformations. The *naive* method, without correction of the multiple testing, increases the Type-I error rate with the number of codings tried. For ten codings this error rate reached 0.27. The error rate calculated by the Bonferroni method decreased with the number of cutpoints. This correction was therefore too conservative whereas the exact method and resampling methods gave a Type-I error rate close to the nominal 0.05 value.

When information on the shape of the effect of the explanatory variable was unknown we investigated the power of the methods applied above. We studied the power for *a threshold effect model* with a cutpoint value at the first tercile. Figure 1(b) gives the power as a function of the number of cutpoints tried. The power of the exact and resampling methods are quite similar to one another, and higher than the Bonferroni method. The difference between these methods and Bonferroni method

Table 1 Strategy for dichotomous transformations: values of the cutpoints c_k according to the number of transformations (q_α represents the quantile of order α)

Number of transformations	c_1	c_2	c_3	...	c_9
1	$q_{1/2}$				
2	$q_{1/3}$	$q_{2/3}$			
3	$q_{1/4}$	$q_{2/4}$	$q_{3/4}$		
⋮	⋮	⋮	⋮	⋮	⋮
9	$q_{1/10}$	$q_{2/10}$	$q_{2/10}$	...	$q_{9/10}$

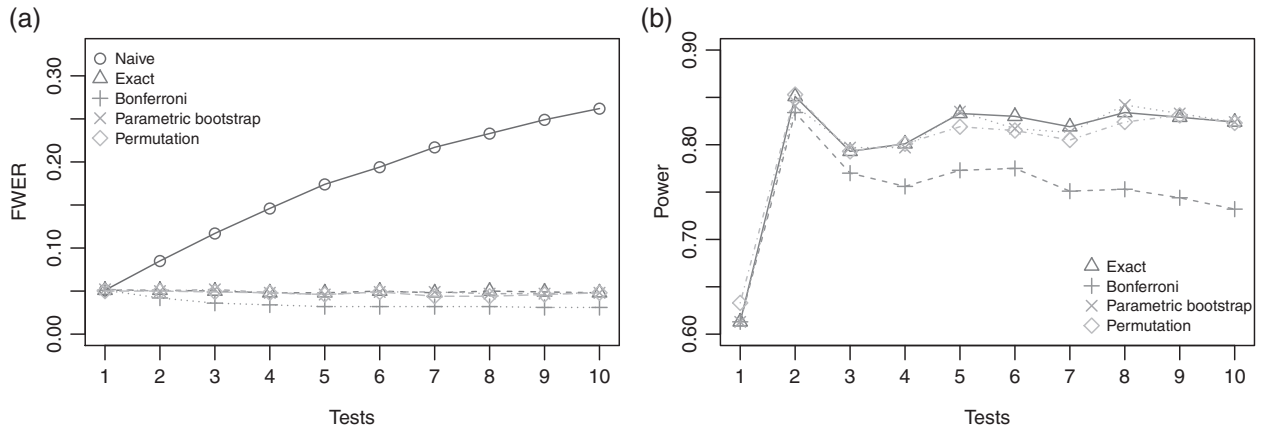

Figure 1 (a) Type-I error rate for various numbers of cutpoints tried according to different methods; (b) power for *a threshold effect model* at the first tercile. ($\gamma_0 = -2.5, \gamma = 1, \beta = 2$).

increases with the number of cutpoints. We also observed that the power was highest at two cutpoints (two transformations). This result, was in fact, expected since we used the first and second terciles respectively as cutpoints for each dichotomous transformation. Power increased again when trying five and eight codings due to the fact that one of these codings corresponded to the first tercile. To conclude, the simulation study with dichotomous transformations showed that the resampling methods provide similar results for the Type-I error rate control and the power as those seen with the exact method.

Categorical transformations with same number of classes

We considered here only categorical transformations with three classes. In this situation, the choices of the two cutpoints (noted c_k^1 and c_k^2) defining the categorical variables into three classes are also subjective. For this simulation study, our strategy was to attempt to find the most favorable transformation into three classes. This consisted of using the tercile of the variable for one transformation with two cutpoints ($c_1^1 = q_{1/3}$ and $c_1^2 = q_{2/3}$); for two transformations we add to the previous choice a transformation with the first quartile and the third quartile for the two cutpoints ($c_2^1 = q_{1/4}$ and $c_2^2 = q_{3/4}$). The global strategy until we obtain 10 transformations in three classes is presented in Table 2.

We investigated the Type-I error rate. Figure 2(a) shows the evolution of the Type-I error rate for categorical transformations in three classes. The results are similar to those we observed for dichotomous transformations. The Bonferroni correction was still too conservative, while resampling methods gave a Type-I error rate close to the nominal 0.05 value.

Next we considered the power of the different methods when the simulated model was specified with a categorical transformation of the continuous variable in three classes defined by cutpoints at the first and third quartile. The two resampling methods gave similar results with a higher power than the Bonferroni method (see Figure 2(b)). The power was highest for two transformations. This result was also expected because, with the strategy presented in Table 2, the transformation into three classes with cutpoints at the first and third quartile is used.

Various categorical transformations

In this last simulation, we presented a more realistic situation where different kinds of transformations were used to investigate the effect of the variable of interest. We proposed trying different categorical transformations and varying the number of classes. The most natural method is to use a dichotomous transformation at the median for one transformation. For two transformations, we added the previous coding and a categorical transformation in

Table 2 Strategy for the categorical transformations in three classes: values of the cutpoints (c_k^1 and c_k^2) for all transformations

Number of transformations	c_k^1	c_k^2
1	$q_{1/3}$	$q_{2/3}$
2	$q_{1/4}$	$q_{3/4}$
3	$q_{1/4}$	$q_{1/2}$
4	$q_{1/2}$	$q_{3/4}$
5	$q_{2/5}$	$q_{4/5}$
6	$q_{1/5}$	$q_{3/5}$
7	$q_{3/5}$	$q_{4/5}$
8	$q_{1/5}$	$q_{2/5}$
9	$q_{1/5}$	$q_{4/5}$
10	$q_{2/5}$	$q_{3/5}$

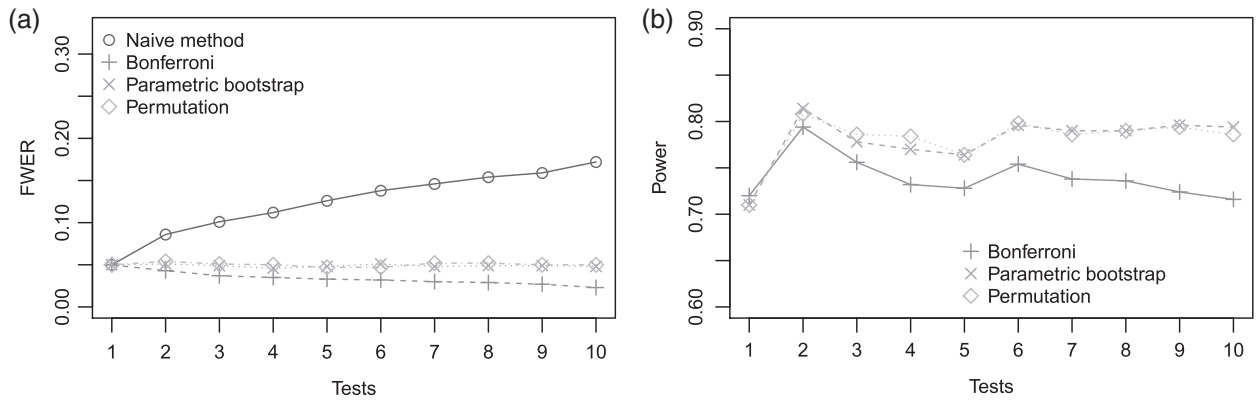

Figure 2 (a) Type-I error rate for various numbers of categorical in three classes according to different methods, (b) and power for an effect of a categorical transformation in three classes defined by cutpoints at the first and third quartile. ($\gamma_0 = -1.25, \gamma_1 = 1, \beta^T = (2, 1.8)$).

three classes based on the tercile. For three transformations, we added the two previous codings and a categorical transformation in four classes based on the quartile, and so on. The strategy proposed in this simulation is presented in Table 3.

The results for the Type-I error rate were similar to the previous simulation case (not shown here). We then studied the power of the different methods when the simulated model is specified with a categorical transformation of the continuous variable in five classes defined by cutpoints at the quintile. We can see in Figure 3, that, in this situation, the parametric bootstrap method seems slightly more powerful than the permutation method. The resampling methods were also more powerful than the Bonferroni method. Finally, as expected, we can see that the power was highest for four transformations, where one of the transformations used corresponded to a categorical transformation with quintiles as cutpoints.

Robustness of resampling methods

We investigated the robustness of the resampling methods when the exchangeability assumption is violated. The data came from the model defined in (4) with two dependent variables $X_i(k)$ and Z_i. The dependency between $X_i(k)$ and Z_i (formalized by the correlation ratio(η^2)) was specified by the following model:

$$Z_i = \beta^* \mathbf{X}_i(k) + \epsilon_i; \tag{5}$$

where $X_i(k)$ is the binary coding of the X_i variable with a cutpoint at the median. The coefficient β^* was computed according to η^2 and the variance of $X_i(k)$ variable.
We tested three different binary codings with cutpoints at the first, the second and the third tercile. The strategy is used for various values of the correlation ratio (η^2) from 0 to 0.6.

The robustness of the permutation method when the exchangeability assumption is violated was evaluated with respect to the results of the exact method. For different correlation ratios (η^2) we evaluated the control of the Type-I error, the power, the Mean Square Error (MSE) of the estimated p_{value} (p_{value} from the exact method was used as a reference), and the rate of good decision (same decision as for the exact method). These results are presented in Figure 4 and show the good behavior of the permutation method since the Type-I error is controlled at the level 0.05, the power is the same for all the methods, the rate of good decision is always greater than 0.97, and the MSE is very low. Moreover, the distributions of the estimated p_{value} are quite similar for different methods (not shown).

Table 3 Strategy for different categorical transformations: values of the cutpoints for all transformations

Number of transformations	c_k^1	c_k^2	...	c_k^9	c_k^{10}
1	$q_{1/2}$				
2	$q_{1/3}$	$q_{2/3}$			
⋮	⋮	⋮	⋮	⋮	⋮
9	$q_{1/10}$	$q_{2/10}$	...	$q_{9/10}$	
10	$q_{1/11}$	$q_{2/11}$	...	$q_{9/11}$	$q_{10/11}$

Example: revisiting the PAQUID cohort example

In order to find the real association between the two variables of interest in the example described at the end of Background section, we applied our newly developed approach which combined different kinds of transformations. Liquet and Commenges [14] have proposed seven dichotomous and five Box-Cox transformations. However, their method did not allow for categorical transformations. We proposed to add, to the seven dichotomous and

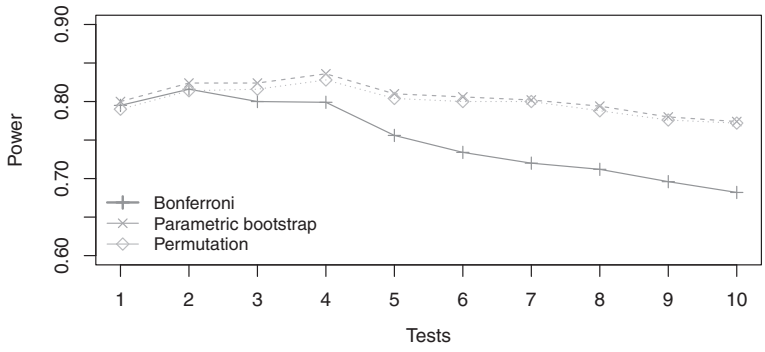

Figure 3 Power for an effect of a categorical transformation in five classes defined by cutpoints at the quintile. ($\gamma_0 = -2.5, \gamma_1 = 1, \beta^T = (-1.3, -0.8, 1.4, 1.7)$).

five Box-Cox transformations for this application, four codings in three classes and four codings in four classes. The best transformation appeared to be the dichotomous transformation of HDL-cholesterol with a cutpoint at the third quartile, as already found by Bonarek et al. [16]. The Bonferroni correction gave a p_{value} equal to 0.140, thus not significant for an α level at 0.05. The p_{value}, which is given by both resampling based methods is 0.038. To conclude, it is important to chose a powerful method of correction, because in this context the p_{value} with no correction given by Bonarek et al. [16] was very optimistic (0.007), and the Bonferroni correction was very conservative, yielding an

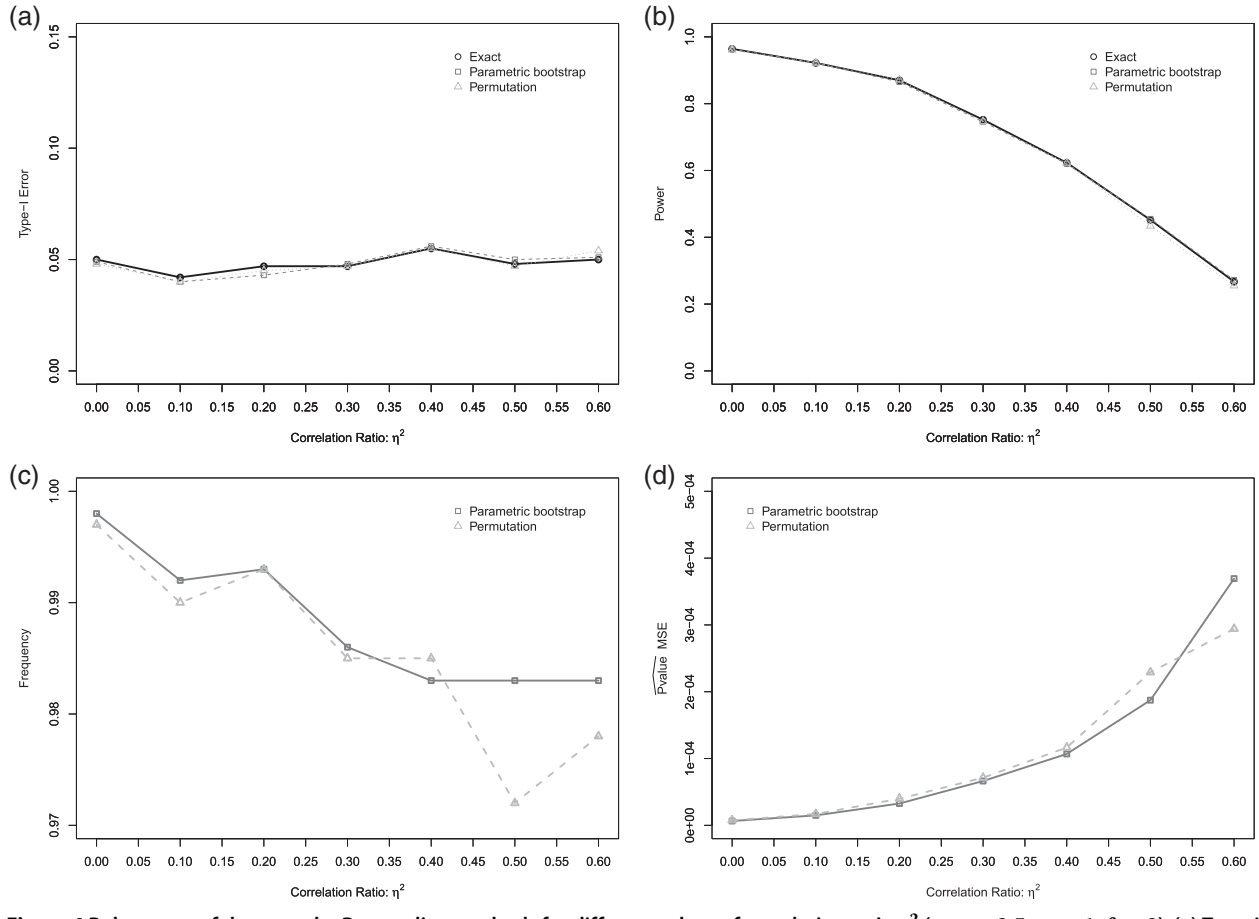

Figure 4 Robustness of the p_{value} by Resampling methods for different values of correlation ratio η^2 ($\gamma_0 = -2.5, \gamma_1 = 1, \beta = 2$): (a) Type-I error; (b) Power; (c) Correct decision rate; (d) Mean Square Error (MSE) of the $\widehat{p_{value}}$.

incorrect conclusion. The proposed approach based on the resampling procedure gave a result which was still significant and more realistic than the uncorrected p_{value}.

Discussion

In this paper, we have considered the problem of correction of significance level for a series of several codings of an explanatory variable in a Generalized Linear Model with several adjusting variables. The methods developed, based on resampling methods, enable us to consider categorical transformations as more flexible in order to explore the unknown shape of the effect between an explanatory and a dependent variable. The simulation studies presented above show, firstly, that the resampling method provides similar results for the Type-I error rate control and the power as those found with the exact method proposed by Liquet and Commenges [14] for dichotomous and Box-Cox transformations. Secondly, in the situation of categorical transformations, these simulations demonstrate the good performance of our proposed approaches. Finally we observed the robustness estimation of the p_{value} by the resampling methods. These methods can be easily generalized to other models, such as the proportional hazards model, and to potentially extend the work of Hashemi and Commenges [15] in the same context.

Conclusion

To conclude, the methods developed, based on resampling, demonstrate good performances, and we have implemented different methods and different strategies of coding in an R package called CPMCGLM M (for Correction of the Pvalue after Multiple Coding in a Generalized Linear Model).

Appendix

The package `CPMCGLM` has been developed in R, an open source statistical software available at http://www.r-project.org. The methods presented in this paper are available in the main function `CPMCGLM()` for Probit, Logit, Linear, and Poisson models. Briefly, the user can specify the transformations tested: Box-Cox, dichotomous or categorical transformations. Two options are possible for defining the cutpoints of the dichotomous and the categorical transformations: the user can either specify them, or the program will automatically use the strategy based on the quantile presented in the simulation study.

The main function provides the best codings according to the maximum test and minimum p_{value} procedures. For this coding, the different methods of correction of the Type-I error rate presented in this paper are provided. We present an illustration of the `CPMCGLM` function on a simulated dataset:

```
data(data_sim)
result<-CPMCGLM(formula=
Weight  Age+as.factor(Sport)+Desease
+Height,
family="gaussian",link="identity",
data=data_sim, varcod="Age",
nb.dicho = 4, nb.categ = 4, nboxcox =
3, N = 10000)

result

Call:
CPMCGLM(formula = Weight   Age +
as.factor(Sport) + Desease +
Height, family = "gaussian", link =
"identity",
data = data_sim,varcod = "Age",
nb.dicho = 4,
nb.categ = 4, nboxcox = 3, N = 10000)

Generalized Linear Model Summary
Family: gaussian
Link: identity
Number of subject: 100
Number of adjustment variable: 4

Resampling
N: 10000

Best coding
Method: Dichotomous transformation
Value of the order quantile cutpoints: 0.6
Value of the quantile cutpoints: 26.4834

Corresponding adjusted pvalue:

           Adjusted pvalue
naive            0.0191
bonferroni       0.2096
bootstrap        0.0686
permutation 0.0656
exact: Correction not available for
these codings
```

Competing interests
Both authors declare that they have no competing interests.

Authors' contributions
BL and JR developed the methodology, the R code, performed the simulation and the analysis on the dataset as well as wrote the manuscript. Both authors read and approved the final manuscript.

Acknowledgements
We would like to thank Luc Letenneur (ISPED, University of Bordeaux) for making the data available and the Danone Research Clinical Study Platform.

Author details
[1]University Bordeaux, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France. [2]INSERM, ISPED, Centre INSERM U-897-Epidemiologie-Biostatistique, Bordeaux, F-33000, France. [3]MRC Biostatistics Unit, Institute of Public Health, Cambridge, CB2 0SR, UK. [4]Danone Research, Avenue de la Vauve, Route départementale 128, Palaiseau Cedex 91767, France.

Received: 9 January 2013 Accepted: 17 May 2013
Published: 8 June 2013

References

1. Bennette C, Vickers A: **Against Quantiles: categorization of continuous variables in epidemiologic research, and its discontents.** *BMC Med Res Methodol* 2012, **12:**21–25.
2. Altman D, Lausen B, Sauerbrei W, Schumacher M: **Dangers of using optimal cutpoints in the evaluation of prognostic factors.** *J Natl Cancer Inst* 1994, **86**(11):829–835.
3. Royston P, Altman D, Sauerbrei W: **Dichotomizing continuous predictors in multiple regression: a bad idea.** *Stat Med* 2006, **25:**127–141.
4. Harrell FE, Lee KL, Mark DB: **Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.** *Stat Med* 1996, **15**(4):361–387.
5. Miller RG: *Simultaneous statistical inference. 2nd ed*. New York - Heidelberg: Berlin: Springer- Verlag. XVI 299; 1981. figs. DM 44.00.
6. Westfall PH: **Improving power by dichotomizing (even under normality).** *Stat Biopharm Res* 2011, **3**(2):353–362.
7. Simes R: **An improved bonferroni procedure for multiple tests of significance.** *Biometrika* 1986, **73**(3):751–754.
8. Sidak Z: **Rectangular confidence regions for the means of multivariate normal distributions.** *J Am Stat Assoc* 1967, **62:**626–633.
9. Holm S: **A simple sequentially rejective multiple test procedure.** *Scand J Stat* 1979, **6:**65–70.
10. Hommel G: **A stagewise rejective multiple test procedure based on a modified Bonferroni test.** *Biometrika* 1988, **75:**383–386.
11. Hochberg Y: **A sharper Bonferroni procedure for multiple test procedure.** *Biometrika* 1988, **75:**800–802.
12. Efron B: **The length heuristic for simultaneous hypothesis tests.** *Biometrika* 1997, **84:**143–157.
13. Liquet B, Commenges D: **Correction of the P-value after multiple coding of an explanatory variable in logistic regression.** *Stat Med* 2001, **20:**2815–2826.
14. Liquet B, Commenges D: **Computation of the p-value of the minimum of score tests in the generalized linear model, application to multiple coding.** *Stat Probability Lett* 2005, **71:**33–38.
15. Hashemi R, Commenges D: **Correction of the p-value after multiple tests in a Cox proportional hazard model.** *Lifetime Data Anal* 2002, **8:**335–348.
16. Bonarek M, Barberger-Gateau P, Letenneur L, Deschamps V, Iron A, Dubroca B, Dartigues J: **between cholesterol, apolipoprotein E polymorphism and dementia: a cross-sectional analysis from the PAQUID study.** *Neuroepidemiology* 2000, **19:**141–48.
17. McCullagh P, Nelder J: *Genaralized Linear Models*. 2edition. New York: Chapman & Hall; 1989.
18. Genz A: **Numerical computation of multivariate normal probabilities.** *J Comput Graphical Stat* 1992, **1:**141–149.
19. Cox D, Hinkley D: *Theoretical Statistics*. London: Chapman & Hall; 1994.
20. Royen T: **Expansions for the multivariate chi-Square distribution.** *J Multivariate Anal* 1991, **38:**213–232.
21. Dagupsta N, Spurrier J: **A class of multivariate χ^2 distributions with applications to comparison with a control.** *Commun Stat- Theory Methods* 1997, **26:**1559–1573.
22. Worsley K: **An improved Bonferroni inequality and applications.** *Biometrika* 1982, **69:**297–302.
23. Westfall PH, Young S: *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics*. New York: NY Wiley; 1992. xvii, 340 p.
24. Yu K, Liang F, Ciampa J, Chatterjee N: **Efficient *p*-value evaluation for resampling-based tests.** *Biostatistics* 2011, **12**(3):582–593.
25. Commenges D, Liquet B: **Asymptotic distribution of score statistics for spatial cluster detection with censored data.** *Biometrics* 2008, **64**(4):1287–1289.
26. Romano J: **On the behavior of randomization tests without a group invariance assumption.** *J Am Stat Assoc* 1990, **85**(411–412):686.
27. Xu H, Hsu J: **Applying the generalized partitioning principle to control the generalized familywise error rate.** *Biom J* 2007, **49:**52–67.
28. Kaizar E, Li Y, Hsu J: **Permutation multiple tests of binary features do not uniformly control error rates.** *J Am Stat Assoc* 2011, **106**(495):1067–1074.
29. Commenges D: **Transformations which preserve exchangeability and application to permutation tests.** *J Nonparametric Stat* 2003, **15**(2):171–185.
30. Westfall PH, Troendle JF: **Multiple testing with minimal assumptions.** *Biom J* 2008, **50**(5):745–755.
31. Good P: *Permutation Tests: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. New-York: Springer-Verlag; 2000.
32. Efron B, Tibshirani R: *An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. London: Chapman and Hall/CRC; 1994.

doi:10.1186/1471-2288-13-75
Cite this article as: Liquet and Riou: Correction of the significance level when attempting multiple transformations of an explanatory variable in generalized linear models. *BMC Medical Research Methodology* 2013 **13**:75.