Computation of the p-value of the maximum of score tests in the generalized linear model: application to multiple coding

Benoit Liquet, Daniel Commenges

Computation of the p-value of the maximum of score tests in the generalized linear model; application to multiple coding.

LIQUET Benoit[1]

Equipe de Probabilités et de Statistique, CC 051

Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149

Université Montpellier II

Place Eugène Bataillon

34095 Montpellier Cedex 5,

and

COMMENGES Daniel

INSERM E0338, ISPED

146 rue Léo Saignat

33076 Bordeaux cedex, France

Abstract: We propose a method to correct the significance level for a series of tests corresponding to several transformations of an explanatory variable in generalized linear model. Correlation between score test are derived to apply the proposed method.

Key words: Generalized linear model, multiplicity, p-value, Score test

[1]benoit.liquet@isped.u-bordeaux2.fr : Equipe de Probabilités et de Statistique, CC 051, Institut de Mathématiques et de Modélisation de Montpellier, UMR CNRS 5149, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5

1 Introduction

In many applied studies it is common that the model used is not completely fixed in advance and that several possibilities are tried. We consider the case where a regression model belonging to the family of the generalized linear models is used. In epidemiology for instance it is quite common that the study focusses on one particular risk factor; the problem is to analyse whether this risk factor has an influence on the risk of a disease or on a biological trait. To answer the question a regression model is used in which the risk factor will be represented by a continuous variable X, and allowing to adjust on $p-1$ already known risk factors of the studied trait. The analysis focusses on the test of "$\beta = 0$", where β is the coefficient representing the effect of the risk factor of interest. However the form of the effect (or the dose-effect relationship) is not known in advance. While a non-parametric approach to this problem may be useful (Hastie and Tibshirani, 1990), most often a simpler approach is used: several transformations of the original variable $g_k(X)$, $k = 1, \ldots, K$, may be tried. Examples of such transformations are dichotomizations of the original continuous variable, where K cutoff points may be used, and Box-Cox tranformations, where K different powers of the original variable are tried (the Box-Cox family also includes the log transform). For each of the transforms tried a test of "$\beta = 0$" is performed. Generally the most significant test is retained and often given without corrections in publications. Of course this leads to increased type I error risk (Miller, 1981).

For instance in studying the effect of minerals in drinking water on the risk of a disease, threshold effect models have been currently used: higher (or lower) risk is expected for a concentration above a certain level. However this threshold

value is generally unknown, leading the epidemiologist to perform several trials. Multiplying the tests and reporting the most significant one without correction is able to pervert the whole inference process in applied science and particularly in epidemiology.

The p_{value} should be corrected to take account of the multiplicity of the test. The simplest correction is to apply the Bonferoni rule; however this leads to a conservative test if the original tests are positively correlated as is the case here. Efron (1997) proposed a correction taking account of the correlation between two consecutive tests, if there is a natural order between the tests with high correlation between adjacent tests. Liquet and Commenges (2001) and Hashemi and Commenges (2002) proposed a more exact correction taking into account the whole correlation matrix, for score tests obtained in logistic regression and proportional hazards models respectively.

Here, we propose a correction of the p_{value} when multiple transformations of an explanatory variable have been tried in a generalized linear model. We construct K score tests corresponding to the K coding of the explanatory variable. We thus have a vector of test statistics $T = (T_1, \ldots, T_K)$ for the same null hypothesis H_0 which have asymptotically a standard normal distribution. Rejecting H_0 if one of the test T_k is larger than a critical value c is equivalent to rejecting H_0 if $T_{max} > c$, where $T_{max} = max(T_1, \ldots, T_K)$. To cope with the multiplicity problem, we need to compute the probability of Type I error for the statistic T_{max} under the null hypothesis H_0 :

$$p_{value} = P(T_{max} > t_{max}) = 1 - P(T_1 < t_{max}, \ldots, T_{max} < t_{max}).$$

The asymptotic joint distribution of $T_1, \ldots, T_{max}$ is a multivariate normal distribution with zero mean and a certain covariance matrix Σ that we will estimate

in the next section. Then we will be able to compute $P(T_1 < t_{max}, \ldots, T_{max} < t_{max})$ using numerical integration (Genz, 1992).

2 Correlation between tests in generalized linear model

2.1 Definitions and notations

Let us consider a generalized linear model (McCullagh and Nelder, 1989) with p explanatory variables where Y_i, $i = 1, \ldots, n$, are independently distributed with probability density function in the exponential family defined as follows :

$$f_Y(Y_i, \theta_i, \phi) = \exp\{[\theta_i Y_i - b(\theta_i)]/a(\phi) + c(Y_i, \phi)\},$$

with $\mathrm{E}[Y_i] = \mu_i = b'(\theta_i)$, $\mathrm{var}[Y_i] = b''(\theta_i)a(\phi)$. We want to test the influence of a variable X^i, adjusted on a vector of explanatory variables Z^i. We consider the case where we do not know the form of the effect of X^i so we may consider K transformations of this variable $X^i(k) = g_k(X^i)$, $k = 1, \ldots, K$. The model for transformation k can be obtained by modeling the canonical parameter θ_i as:

$$\theta_i(k) = Z^i\gamma + X^i(k)\beta^k, \ i = 1, \ldots, n,$$

where $Z^i = (1, Z^i_1, \ldots, Z^i_{p-1})$ and $\gamma = (\gamma_0, \ldots, \gamma_{p-1})^T$ is a $p \times 1$ vector of coefficients. In the sequel,we will denote by $X(k)$ the vector $(X^1(k), \ldots, X^n(k))$; $\mu = (\mu_1, \ldots, \mu_n)^T$ and l^k the log-likelihood of $Y = (Y_1, \ldots, Y_n)^T$ for the model with $X(k)$.

2.2 Score test

For all the K transformations, H_0: "$\beta^k = 0$" is the same null hypothesis, given by $\theta_i(k) = Z^i\gamma$. This defines a unique probability measure under which the

distribution of the score tests (T_k) can be computed as standardized versions of the score statistics $U(k)$ which are asymptotically normal distributed:

$$U(k) = \frac{\partial l^k}{\partial \beta^k}(\beta = 0) \quad = \quad \frac{1}{a(\phi)} X^T(k)[Y - \hat{\mu}] = \frac{1}{a(\phi)} X^T(k)\hat{R},$$

where $\hat{R}$ is the vector of residuals $\hat{R}_i = Y_i - \hat{\mu}_i$ computed under the null hypothesis. Asymptotically, the variance of the score test can be computed as (see Cox and Hinkley, 1979): $\text{var } U(k) = I_{\beta^k\beta^k} - I_{\beta^k\gamma} I_{\gamma\gamma}^{-1} I_{\gamma\beta^k}$, where $I_{\beta^k\beta^k} = -\text{E}\left(\frac{\partial^2 l}{\partial\beta^k\partial\beta^k}\right)$, $I_{\beta^k\gamma} = -\text{E}\left(\frac{\partial^2 l}{\partial\beta^k\partial\gamma}\right)$, $I_{\gamma\gamma} = -\text{E}\left(\frac{\partial^2 l}{\partial\gamma\partial\gamma}\right)$. Another method is to calculate directly $\text{var } U(k)$ by :

$$\text{var } U(k) = \frac{1}{a(\phi)^2} X^T(k) \text{var}[\hat{R}] X(k).$$

The term, $\text{var}[\hat{R}]$, will be also necessary to determine the correlation between score tests.

Estimation of $\text{var}[Y - \hat{\mu}] = \text{var}[\hat{R}]$

A Taylor expansion of $Y - \hat{\mu}$ about its values in γ, the real parameter value, gives

$$Y - \hat{\mu} = Y - \mu - \frac{\partial \mu}{\partial \gamma}^T (\gamma)(\hat{\gamma} - \gamma) + o_p(n^{-1/2}), \tag{1}$$

where $\frac{\partial \mu}{\partial \gamma}^T(\gamma) = \frac{1}{a(\phi)} VZ$ and Z is a $n \times p$ matrix with row $Z^i, i = 1, \ldots, n$; V is the diagonal matrix with diagonal terms $v_{ii} = \text{var } Y_i$. Expanding the first derivatives of the log likelihood around γ yields

$$\frac{\partial l}{\partial \gamma}(\hat{\gamma}) = \frac{\partial l}{\partial \gamma}(\gamma) + \frac{\partial^2 l}{\partial \gamma^2}(\gamma)(\hat{\gamma} - \gamma) + o_p(n^{-1/2}), \tag{2}$$

where $\frac{\partial^2 l}{\partial \gamma^2}(\gamma) = -\frac{1}{a(\phi)^2} Z^T V Z$ and $\frac{\partial l}{\partial \gamma}(\hat{\gamma}) = 0$ since γ is estimated nullifying the score $U_\gamma = \frac{\partial l}{\partial \gamma}(\gamma)$. With (2) we find :

$$\hat{\gamma} - \gamma = a(\phi)^2 (Z^T V Z)^{-1} U_\gamma + o_p(n^{-1/2}),$$

where $U_\gamma = \frac{1}{a(\phi)} Z^T(Y-\mu)$. So, replacing in (1) we find:

$$\begin{aligned} Y - \hat{\mu} &= Y - \mu - (\frac{1}{a(\phi)} VZ) a(\phi)(Z^T V Z)^{-1} Z^T (Y-\mu) + o_p(n^{-1/2}) \\ &= (I-H)(Y-\mu) + o_p(n^{-1/2}), \end{aligned}$$

where H is the matrix $H = VZ(Z^TVZ)^{-1}Z^T$. Note that this expression is exact for the normal linear model. With an approximation error of order $o_p(n^{-1})$, we have $\mathrm{var}[Y-\hat{\mu}] = \mathrm{var}[(I-H)(Y-\mu)]$. Using the idempotence property of $(I-H)$, it can be seen after some computation that $\mathrm{var}[(I-H)(Y-\mu)] = (I-H)V$. Finally we find :

$$\mathrm{var}\ U(k) = \frac{1}{a(\phi)^2} X^T(k)(I-H)VX(k).$$

In practice, we use an estimator of $\mathrm{var}\ U(k)$ defined by :

$$\widehat{\mathrm{var}}\ U(k) = \frac{1}{\hat{a}(\phi)^2} X^T(k)(I-H)\hat{V}X(k), \tag{3}$$

where $\hat{a}(\phi)$ and $\hat{V}$ are the estimator of $a(\phi)$ and V.

2.3 Correlation between two tests

Let T_k and T_l be two score test statistics associated with the transformations $X(k)$ and $X(l)$:

$$T_k = \frac{U(k)}{\sqrt{\widehat{\mathrm{var}}\ U(k)}}; \quad T_l = \frac{U(l)}{\sqrt{\widehat{\mathrm{var}}\ U(l)}}.$$

Neglecting the covariance between the estimators of the variances of $U(k)$ and $U(l)$, the correlation between the two tests is :

$$\begin{aligned} \rho_{kl} = corr(T_k, T_l) &\simeq \frac{\mathrm{cov}(U(k), U(l))}{\sqrt{\widehat{\mathrm{var}}\ U(l)}\sqrt{\widehat{\mathrm{var}}\ U(l)}} \\ &= \frac{1}{\hat{a}(\phi)^2} \frac{X^T(k)\mathrm{var}[Y-\hat{\mu}]X(l)}{\sqrt{\widehat{\mathrm{var}}\ U(l)}\sqrt{\widehat{\mathrm{var}}\ U(l)}}. \end{aligned}$$

Using (3) we finally obtain an expression which does not depend on $a(\phi)$:

$$\rho_{kl} = \frac{X^T(k)[(I-H)\hat{V}]X(l)}{\sqrt{X^T(k)(I-H)\hat{V}X(k)}\sqrt{X^T(l)(I-H)\hat{V}X(l)}}.$$

Then we can compute $P(T_1 < t_{max}, \ldots, T_K < t_{max})$ for determining the p-value associated to the test T_{max}. This is done by integrating the normal density over a quadrant and this can be done quite accurately for $K < 20$ using the subregion adaptive algorithm proposed by Genz (1992). Note that this result allows a correction of the p-value if one knows the vectors of the transformed variable $X(k), k = 1, \ldots, K$, the matrix of explanatory variables Z and the estimated variances $\hat{V}$; that is, there is the possibility to design a unique program for any model in the family and for any transformation.

3 Simulation

Simulations for studying the effect of the correction on the type I error risk were carried out with a Poisson model ($a(\phi) = 1,\ b(\theta_i) = e^{\theta_i},\ \mu_i = \mathrm{E}[Y_i] = e^{\theta_i}$) consisting of two explanatory variables: Z_1 an adjustment variable and X the variable of interest. We considered the following models

$$\log \mathrm{E}[Y_i] = \theta_i(k) = \gamma_0 + \gamma_1 Z_1^i + \beta X^i(k), \quad i = 1, \ldots, n.$$

where $X^i(k)$ were dichotomized versions of a continuous variable X. For the cut-off points we chose the median for one dichotomous transformation ($K = 1$), the first and the second terciles for two dichotomous transformations ($K = 2$), the quartiles values for three transformations ($K = 3$) and so on. Z_1 and X were generated according to a uniform distribution $U[0, 1]$. The sample size was set to be 100. We used 10,000 replications for the simulation.

3.1 Study of type I error rate

In this simulation, we took $\gamma_0 = 1$, $\gamma_1 = 10$ and $\beta = 0$. For a replication, the rejection criterion of H_0 ($\beta = 0$) was a p_{value} less than 0.05. Thus, for a simulation, the empirical type I error rate was the proportion of p_{value} less than 0.05. Figure 1(a) shows the type I error rate for dichotomous transformations, as a function of the number of coding (K) tried. The naive approach which retains the test with the smallest p-value had, as expected, a type I error rate which increased with the number of cutpoints tried: this is exhibited by the "no correction" curve in Figure 1(a). The error rate calculated by the Bonferroni method decreased with the number of cutpoints, leading to a conservative test. The exact calculation proposed here gave a type I error rate very close to the nominal 0.05 value.

Figure 1 approximately here

3.2 Power

In this simulation, we took $\gamma_0 = 1$, $\gamma_1 = 5$ and $\beta = 5$. We studied the power for a threshold effect model with a cutpoint value at the first tercile ("threshold effect model"). Figure 1(b) gives the power as a function of the number of cutpoints. With any method we expect that the power will be substantially less than that of a test that would be done knowing the cutpoint value: this gives an upper bound for the power (represented by a horizontal line in Figure 1(b)). As expected the exact calculation provides more power than the Bonferroni method. Each time the true cutpoint value appears in the values tried the power increases; however these oscillations decrease with the number of values tried. When the good cutpoint value does not appear in the values tried, the power of the exact

method tends to increase with the number of values tried: the power is slightly higher using deciles (9 values tried) than using quartiles (3 values tried), and this is in contrast with the Bonferroni method.

References

Cox, D.R. and Hinkley, D.V. (1979), *Theoretical Statistics* (Chapman and Hall, New York).

Efron, B. (1997), The length heuristic for simultaneous hypothesis tests, *Biometrika.* **84**(1), 143-157.

Genz, A. (1992), Numerical computation of multivariate normal probabilities, *Journal of Computational and Graphical Statistics.* **47**(1), 141-149.

Hashemi, R. and Commenges, D. (2002), Correction of the p-value after multiple tests in a Cox proportional hazard model. *Lifetime Data Anal.* **8**(4), 335-48.

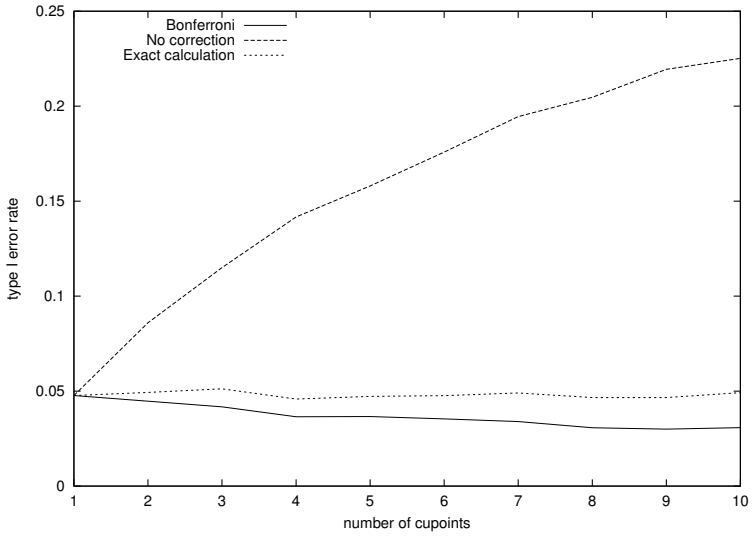
Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models* (Chapman and Hall, London).

Liquet, B. and Commenges, D. (2001), Correction of the P-value after multiple coding of an explanatory variable in logistic regression, *statist. Med.* **20**, 2815-2826.

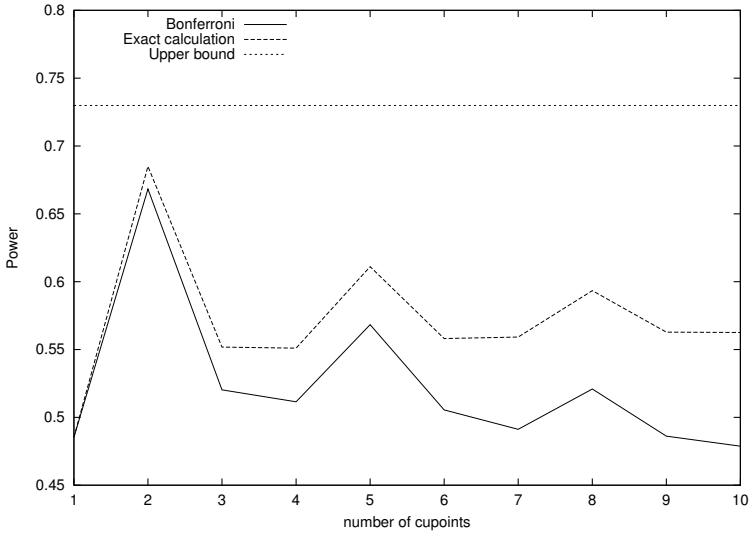
McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models* (Chapman and Hall, New York, 2nd ed.)

Miller, R. (1981), *Simultaneous Statistical Inference* (Springer-Verlag, New York).

LIST OF FIGURE



(a) Type I error rate



(b) Power