arXiv:1003.1002v1 [stat.ME] 4 Mar 2010

Extending The Range of Application of Permutation Tests: the Expected Permutation p-value Approach

Daniel Commenges

INSERM U897, ISPED, 146 rue Léo Saignat, Bordeaux, 33076, France

Tel: (33) 5 57 57 11 82; Fax (33) 5 56 24 00 81

March 4, 2010

Abstract

The limitation of permutation tests is that they assume exchangeability. It is shown that in generalized linear models one can construct permutation tests from score statistics in particular cases. When under the null hypothesis the observations are not exchangeable, a representation in terms of Cox-Snell residuals allows to develop an approach based on an expected permutation p-value (Eppv); this is applied to the logistic regression model. A small simulation stydy and an illustration with real data are given.

Resumé

La limitation des tests de permutation est qu'ils sont basés sur une hypothèse d'échangeabilité. Il est montré que dans les modèles linéaires généralisés on peut construire des tests de permutation par la statistique du score dans des cas particuliers. Quand les observations ne sont pas échangeables sous l'hypothèse nulle, une représentation en terme de résidus de Cox-Snell permet de développer une approche basée sur l'espérance de la p-valeur de permutation; ceci est appliqué au modèle de régression logistique.

Keywords: Exchangeability, Permutation tests, Residuals, Score Test, Logistic regression, p-values.

Version française abrégée

Considérons une statistique $T(Y)$ pour tester une hypothèse H_0. La décision de rejet de H_0 est prise si $T(Y) \geq c_\alpha$, c_α choisi tel que l'erreur de type I est α. La p-valeur est définie comme une variable aléatoire par:

$$pv[T(Y)] = \mathrm{E}\{\mathrm{I}_{\mathrm{T(Y^*)>T(Y)}}|\sigma(\mathrm{Y})\}$$

où Y^* est une variable indépendante de Y mais de même distribution. Les tests de permutation sont basés sur un conditionnement sur les statistiques d'ordre : $Y_{(o)} = Y_{(1)}, \ldots, Y_{(n)}$.

La p-valeur de permutation est:

$$ppv[T(Y)] = \mathrm{E}\{\mathrm{I}_{\mathrm{T(Y^*)>T(Y)}}|\sigma(\mathrm{Y}) \vee \sigma(\mathrm{Y}^*_{(\mathrm{o})} = \mathrm{Y}_{(\mathrm{o})})\}$$

Supposons que nous puissions représenter Y par $Y = g(\varepsilon)$ avec ε échangeable. Une telle représentation a été proposé par Cox et Snell [3]. Alors $T(Y) = T[g(\varepsilon)] = S(\varepsilon)$. Si ε était observé on pourrait utiliser la p-valeur de permutation :

$$pv_{\varepsilon^*_{(o)}=\varepsilon_{(o)}}[S(\varepsilon)] = \mathrm{E}\{\mathrm{I}_{\mathrm{S}(\varepsilon^*)>\mathrm{S}(\varepsilon)}|\sigma(\varepsilon) \vee \sigma(\varepsilon^*_{(\mathrm{o})} = \varepsilon_{(\mathrm{o})})\}.$$

En général ε n'est pas observé. Nous proposons donc de prendre l'espérance:

$$Eppv[T(y)] = \mathrm{E}\{\mathrm{pv}_{\varepsilon^*_{(\mathrm{o})}=\varepsilon_{(\mathrm{o})}}[\mathrm{S}(\varepsilon)]|\sigma(\mathrm{Y})\}.$$

L'espérance peut dépendre de paramètres de nuisance $\gamma \in \Gamma$. Dans ce cas on peut soit les remplacer par les estimateurs du maximum de vraisemblance, soit calculer $\max_{\gamma\in\Gamma} Eppv(\gamma)$. Cette approche est adaptée à un modèle de régression logistique.

1 Introduction

Permutations tests can be useful as distribution-free tests and also have exact size (as opposed to the asymptotic validity of most conventional tests). However the use of permutation tests in regression problems has been limited because valid permutation tests obtain only if the observations are exchangeable under the null hypothesis. A vector Y has an exchangeable distribution if PY has the same distribution as Y, for any permutation matrix P. If we consider a test statistic $T(Y)$, a permutation test is obtained, if Y is exchangeable, by conditioning on the order statistics $Y_{(o)} = \{Y_{(1)}, \ldots, Y_{(n)}\}$ [6]. The assumption of exchangeability, although a little less stringent than the assumption of identically independently distributed (i.i.d.) observations, is still quite restrictive, and does not hold for instance in regression problems.

The has been many applications of permutation tests; a particularly interesting permutation test was proposed by Mantel [8]. Permutation tests are often based on score tests. For some theory about permutation tests see [1] and for score tests see [2] and [4].

In this paper we propose a new approach, called expected permutation p-value (Eppv), based on permuting an unobserved exchangeable variable. Section 2 presents permutation versions of score tests in generalized linear models. In sectiin 3 some theory about p-values, permutation and conditioning is developed and the Eppv are presented. This approach is then applied to the logistic regression model in section 4. Section 5 presents a short simulation. An illustration with real data is given ins ectiion 6 which concludes.

2 Permutation score tests

Consider a sample of independent random variables Y_i, $i = 1, \dots, n$, and assume a generalized linear model; the contribution of observation i to the likelihood is:

$$f(Y_i;\theta_i,\eta) = \exp\left\{\eta^{-1}\left[\theta_i Y_i - b(\theta_i)\right] + c(Y_i,\eta)\right\}$$

with $E(Y_i) = b'(\theta_i) = \mu_i$ and $\theta_i = Z^i\beta$ where $Z^i = (z_1^i, \dots, z_p^i)$ is a row vector of explanatory variables (considered here as deterministic) and β is a $p \times 1$ vector of regression coefficients; here η denotes the dispersion parameter. Then the score equation obtained by equating to zero the derivative of the loglikelihood L relatively to β is $Z^T\hat{R} = 0$, where Z is the $n \times p$ matrix of explanatory variables z_j^i , and $\hat{R} = (\hat{R}_1, \dots, \hat{R}_n)^T$ is the vector of residuals $\hat{R}_i = Y_i - \mu_i(\hat{\beta})$. Thus the estimated residuals are orthogonal to the space of explanatory variables.

If we consider an explanatory variable indexed by $p+1$, the model becomes $\theta_i = Z^i\beta + z_{p+1}^i\beta_{p+1}$. Lets us denote the parameters $\gamma = (\eta, \beta, \beta_{p+1})$. The score statistic for testing H_0: "$\beta_{p+1} = 0$" has the linear form:

$$S(Y) = \frac{\partial L}{\partial\ \beta_{p+1}}(\beta_{p+1} = 0) = z_{p+1}^T \hat{R}, \tag{1}$$

where $z_{p+1}^T = (z_{p+1}^1, \ldots, z_{p+1}^n)$ is the vector of values for explanatory variable $p+1$ and $\hat{R}$ is the vector of residuals in the model not including variable $p+1$.

A test for H_0: "$\beta_{p+1} = 0$" may be based on the asymptotic distribution of $n^{-1/2}S(Y)$. Let us call $\phi(Y)$ the critical function of the test ($\phi(Y) = 1$: H_0 rejected, $\phi(Y) = 0$: H_0 not rejected); except in simple cases it is not possible to construct exact tests, that is with $\mathrm{E}_\gamma[\phi(Y)] = \alpha$, $\gamma \in \omega$, where ω is the subset of the parameter space corresponding to H_0. For small sample sizes the difference between the nominal and true Type I error rates may be large. In regression models it is tempting to try to construct tests based on permutation of the residuals in the score statistics [10]. Fisher exact test can be shown to be a permutation of the residuals in a score test, in a case where the observations are exchangeable under the null hypothesis. However, generally as soon as there is one explanatory variable under the null hypothesis, neither Y nor $\hat{R}$ are exchangeable; hence, permutation tests cannot be constructed [1].

3 Some theory about p-values, permutation and conditioning

3.1 p-values

Consider a test $\phi(Y)$ based on a statistic $T(Y)$. We examine the case where the decision to reject H_0 is taken if $T(Y) \geq c_\alpha$, c_α being chosen such $\mathrm{E}_\gamma[\phi(Y)] = \alpha$. A definition of the p-value which allows to consider it as a random variable (and hence to study its properties) is

$$pv[T(Y)] = \mathrm{E}_\gamma[\mathrm{I}_{\mathrm{T(Y^*)\geq T(Y)}}|\sigma(\mathrm{Y})]$$

where Y^* is a random variable independent from Y but with the same distribution and $\sigma(Y)$ is the sigma-algebra generated by Y. See [11] for properties of the conditional expectation. We can construct a size α test by rejecting H_0 if $pv[T(Y)] \leq \alpha$, that is: $\phi(Y) = I_{pv[T(Y)]\leq\alpha}$.

3.2 Conditional p-values

We may define a p-value conditional on $\mathcal{C}$, where $\mathcal{C} \subset \sigma(Y, Y^*)$ as:

$$pv_{\mathcal{C}}[T(Y)] = \mathrm{E}_\gamma[\mathrm{I}_{\mathrm{T(Y^*)\geq T(Y)}}|\sigma(\mathrm{Y}) \vee \mathcal{C}\}.$$

Conditional tests can be constructed as $\phi(Y) = I_{pv_{\mathcal{C}}[T(Y)]\leq\alpha}$. We have $\mathrm{E}_\gamma[\phi(Y)|\mathcal{C}] = \alpha$; it follows that we also have $\mathrm{E}_\gamma[\phi(Y)] = \alpha$. That is, marginally the test has size α, but the critical regions (and the power) depend on $\mathcal{C}$. The conditional approach has been advocated for two different situations [7].

The first arises if we have a sufficient statistic C for the family of measure $\mathcal{P}^Y = \{P_\gamma, \gamma \in \omega\}$, where $\omega = H \cap K$, the frontier between the sets representing the null (H) and the alternative (K) hypotheses. If $\mathcal{C}$ is the sigma-algebra generated by C, then $pv_{\mathcal{C}}[T(Y)]$ no longer depends on γ, so that we obtain a similar test, $\mathrm{E}_\gamma[\phi(Y)] = \alpha$, $\gamma \in \omega$. Such a test is said to have the Neyman structure relatively to C. As an example consider the case where we observe variables Y_i, $i = 1, \ldots, n$ which are i.i.d. under the family of measures $\mathcal{P}^Y = \{P_\gamma, \gamma \in \omega\}$. Then the order statistic. $Y_{(o)} = Y_{(1)}, \ldots, Y_{(n)}$ is sufficient for γ and if we take $\mathcal{C} = \sigma(\{Y^*_{(o)} = Y_{(o)}\})$ we obtain a permutation test, that is we have $\mathrm{E}[\phi(Y)|Y_{(o)}] = \alpha$. Due to the discrete character of the conditional distribution of $T(Y)$, it is not possible to achieve $\mathrm{E}[\phi(Y)|Y_{(o)}] = \alpha$ for all α, except by resorting to randomisation; we will neglect this problem in the sequel.

The second situation arises in the presence of ancillary statistics Z: here the motivation is to perform the test adapted to the situation fixed by the particular realization of Z. We may also consider S-ancillary statistics whose distribution depends on an unknown parameter ξ, while the distribution of Y given Z does not depend on ξ. While the unconditional p-value depends on both γ and ξ, the p-value conditional on Z does not depend on ξ. As an example consider the case of a regression model where explanatory variables Z^i are associated to response variables Y_i: the regression model specifies the conditional distribution of Y_i given Z^i and depend on γ, while the marginal distribution depends on ξ only. It is natural to consider tests which are conditional on Z; in our formalism, for a test stastic $T(Y, Z)$ we then compute the conditional p-value $pv_{\mathcal{C}}[T(Y, Z)]$ with $\mathcal{C} = \sigma(\{Z^* = Z\})$.

The two situations have in common the fact that there is a reduction of the number of parameters on which the p-value depends. In the particular case where there is a sufficient statistic for γ, the p-value does not depend on any parameter. However in complex problems this may not be achieved without loosing too much power. One possibility is to replace $pv_{\mathcal{C}}[T(Y);\gamma]$ by $pv_{\mathcal{C}}[T(Y);\hat{\gamma}]$, where $\hat{\gamma}$ is an estimator of γ. We would like to a have a procedure such that $|pv_{\mathcal{C}}[T(Y);\hat{\gamma}] - pv_{\mathcal{C}}[T(Y);\gamma]|$ is as small as possible. Choosing large $\mathcal{C}$ may help to reduce the variance of this random variable. Another way is to apply a minimax argument. If it is known that γ belongs to a compact set Γ, then we may base a test on $\max_{\gamma\in\Gamma} pv_{\mathcal{C}}[T(Y);\gamma]$. This leads to a test of size lower or equal to α.

3.3 The expected conditional p-value

Consider the case where $Y = g(\varepsilon)$, where $g(.)$ is a non-decreasing function; if g is not one-to-one we have $\sigma(Y) \subset \sigma(\varepsilon)$. If we have a statistic $T(Y)$, this defines a statistic $S(\varepsilon) = T(g(Y))$. We may consider the p-value $pv_{\mathcal{C}}[S(\varepsilon)] = \mathrm{E}_\gamma[\mathrm{I}_{\mathrm{S}(\varepsilon^*)\geq \mathrm{S}(\varepsilon)}|\sigma(\varepsilon) \vee \mathcal{C}\}$, where $\mathcal{C} \subset \sigma(\varepsilon^*,\varepsilon)$. Since in general this is not $\sigma(Y)$-measurable, we may consider its expectation $Epv_{\mathcal{C}}[S(\varepsilon)] = \mathrm{E}_\gamma[\mathrm{pv}_{\mathcal{C}}[\mathrm{S}(\varepsilon)]|\sigma(\mathrm{Y})]$. A size-$\alpha$ test can be constructed using this expected conditional p-value as usual.

This approach can in particular be connected with the Cox-Snell family which represents Y as $Y = g(\varepsilon)$, where ε is exchangeable. Such a representation was proposed by Cox and Snell [3] to define residuals. If ε were observed a permutation test could be constructed by conditioning on the order statistic of ε. It is appealing thus to use an expected conditional p-value choosing $\mathcal{C} = \sigma(\varepsilon^*_{(o)} = \varepsilon_{(o)})$. Such a p-value will be called expected permutation p-value (Eppv).

Numerically this method is easy to implement: draw at random ε^* from the distribution of ε conditional on Y; compute the permutation p-value; take the mean of the p-values for a sufficient number of drawings. However the distribution of ε conditional on Y may depend on parameters that may have to be estimated (see sections 3.2 and 4).

4 Applications of the Eppv approach to the logistic model

A logistic regression model is specified by: $\Pr(Y_i = 1) = \pi_i$; $\text{logit}(\pi_i) = z^i\beta$. It can be depicted in terms of latent i.i.d. variables ε_i having a uniform distribution on [0,1]:

$$Y_i = I_{\varepsilon_i \leq \pi_i}$$

A score test for H_0 : "$\beta_{p+1} = 0$" is $T(Y) = S(\varepsilon) = z^T_{p+1}(I_{\varepsilon \leq \pi} - \pi)$ with obvious vectorial notation. For a permutation test only the first part $z^T_{p+1} I_{\varepsilon \leq \pi}$ is needed. However, because $\sum_i I_{\varepsilon_i \leq \pi_i}$ is not constant under permutation of ε, the test is not invariant for a change of origin of z: there is a need to center one of the two vectors involved in this scalar product, a concept also related to that of "clean" form as in [1]. Thus the proposed statistic is $T(Y) = S(\varepsilon) = \sum_i z^i_{p+1}(I_{\varepsilon_i \leq \pi_i} - n^{-1} \sum_i I_{\varepsilon_i \leq \pi_i}) = \sum_i (z^i_{p+1} - \bar{z}_{p+1}) I_{\varepsilon_i \leq \pi_i})$ (where $\bar{z}_{p+1}$ is the mean of z^i_{p+1}), which is invariant.

For computing the Eppv we draw ε from its conditional distribution which is

- $\varepsilon_i \sim U[0, \pi_i]$ if $Y_i = 1$;
- $\varepsilon_i \sim U[\pi_i, 1]$ if $Y_i = 0$.

If the π_i are known, an exact permutation test follows. In practice one may replace π_i by an estimator $\hat{\pi}_i$, the maximum likelihood estimator of π_i under H_0, leading to an approximate test. It is conjectured that the type I error probability is $\alpha + O_p(n^{-1/2})$, similar as when using the asymptotic distribution of the standardized score statistic. However for small sample size the Eppv approach may have better performance because of the non-standard conditioning. Another possibility is to apply the minimax approach. Consider the case $p = 1$ and it is known that $\beta_1 \in [a, b]$. One can find $\max_{\beta_1 \in [a,b]} Eppv(\beta_1)$ and this leads to a test with type I error probability lower or equal to α. In practice the maximum can be found numerically.

It is interesting to note that when there is no explanatory variable under the null hypothesis, the Eppv test reduces to Fischer's exact test; this happens because for all i, $\hat{\pi}_i = \bar{Y}$ so that permuting ε is identical to permuting Y.

5 Simulation study

We have simulated a Logistic regression model given by:

$$logit(\pi_i) = \beta_0 + \beta_1 z_1^i + \beta_2 z_2^i$$

with $\beta_0 = 0$; $\beta_1 = 1$; $z_1^i = w_1^i - 1$; $z_2^i = (w_2^i - 1)(z_1^i)^d$, where w_1^i and w_2^i are independent with exponential distributions. The values $d = 0$, where z_1^i and z_2^i were independent, and $d = 1$ and $d = -1$, producing two different cases of non-linear dependencies between z_1^i and z_2^i, were tried. Samples of sizes 30 and 15 were generated from this model. The problem was: testing H_0 : "$\beta_2 = 0$" at size $\alpha = 0.05$. The empirical sizes (for $\beta_2 = 0$) and powers (for $\beta_2 = 1$ for $n = 30$ and $\beta_2 = 2$ for $n = 15$) of the likelihood ratio (LR) test, the Wald test, a score test based on permutation of residuals (PR) and the Eppv test have been estimated by simulation using 10000 replicates. We have also tried a Bootstrap test: among several possibilities we have chosen the one which seemed the most natural that is a non-parametric bootstrap of the Wald test; the guidelines given in [5], that is resampling $(\beta_2^* - \hat{\beta}_2)/\sigma^*_{\beta_2}$ (where β_2^* is the maximum likelihood estimate of β_2 for a resample and $\sigma^*_{\beta_2}$ is the estimated standard deviation of β_2^*), have been applied; this time-consuming test (using 499 resamples) has been studied on only 1000 replicates. For simplicity, for all the tests, only marginal probabilities were estimated, that is we regenerated the z_1^i and z_2^i at each replicate.

The results appear in Table 1 (with β_2 simply denoted β). It is clear that the Wald test tends to be conservative while the LR test tends to be anti-conservative. These behaviours are more marked for $n = 15$ than for $n = 30$. The tests based on permutation better respect the size of the tests with a tendency to conservative for $d = 1$; the Eppv test has a better stability than permutation of residuals. The bootstrap Wald test is not really practical for $n = 15$ because many configurations generated by resampling are too particular and lead to failure of convergence of the algorithm; so the results of this test are not displayed in Table 1. For $n = 30$ it is strongly anti-conservative: the estimated type I error risks are 0.088, 0.097, 0.14 for $d = 0$, 1 and -1 respectively.

The power of the Eppv test is always higher than that of the Wald test and of the test based on permutation of residuals; it is sometimes lower than that of the likelihood ratio test but the latter is not very reliable in the situations considered. In conclusion when working with small samples and when we can suspect a dependency between the factor studied and the other explanatory variables, the Eppv test seems the most reliable among the tests considered here.

6 Illustration on real data

Even in a large study very small numbers may occur in some categories of the sample which are of interest. The small problem treated here for illustration is taken from a real study on the effect of wine consumption of the risk of developing dementia [9]. In this study, 2273 non-demented subjects were followed up during three years. Subjects were classified according to their wine consumption as: no drinkers, mild drinkers moderate or heavy drinkers. During the follow-up 99 cases of dementia developed. Potentially important confounding factors were age, gender and educational level (here coded as a binary variable: no primary diploma vs primary diploma or above). Globally it appeared from a logistic regression analysis that moderate wine consumption was a protective factor against dementia. However if we try to analyze the data separately by gender (which is legitimate because both the curse of dementia and drinking habits are different among genders) very small numbers occur. In particular, there were 28 dementia cases among 811 non-drinking women and 0 cases among 44 moderate or heavy drinking women.

Table 1: Simulation results based on 10000 replicates of a logistic regression model comparing the Wald test, the likelihood ratio test (LR), the test based on permutation of residuals (PR) and the Eppv test; the theoretical size of the tests is 0.05.

		Wald	LR	PR	Eppv
n = 30					
	$d = 0$	0.044	0.063	0.051	0.052
$\beta = 0$	$d = 1$	0.025	0.069	0.015	0.025
(Type I error)	$d = -1$	0.020	0.080	0.062	0.046
	$d = 0$	0.45	0.53	0.47	0.48
$\beta = 1$	$d = 1$	0.17	0.31	0.14	0.17
(Power)	$d = -1$	0.81	0.91	0.85	0.88
n = 15					
	$d = 0$	0.020	0.072	0.049	0.049
$\beta = 0$	$d = 1$	0.009	0.094	0.018	0.020
(Type I error)	$d = -1$	0.007	0.094	0.066	0.041
	$d = 0$	0.22	0.52	0.57	0.58
$\beta = 2$	$d = 1$	0.10	0.41	0.19	0.22
(Power)	$d = -1$	0.16	0.65	0.75	0.79

With such figures, a logistic regression with wine consumption as an explanatory variable fails to converge so that it is not possible to use a Wald test and a likelihood ratio test is probably not very reliable. For one-sided alternative, Fisher's exact test gave a p-value equal to 0.21 and when adjusting on age and educational level, we obtained p-values equal to 0.18 and 0.13 with the PR and Eppv tests respectively; on the basis of these data, taking into account possible confounding factors, the hypothesis that consumption of wine has no effect on risk of dementia among women cannot be rejected.

In conclusion the Eppv approach extends permutation tests ideas to complex problems. Bootstrap was also in part motivated by such an extension but unlike bootstrap, the Eppv approach keeps the idea of conditioning on the order statistic of an exchangeable vector.

References

[1] Commenges, D. (2003), Transformations which preserve exchangeability with application to permutation tests, *Journal of Nonparametric Statistics* **59**, 171-195.

[2] Commenges, D. and Jacqmin–Gadda, H. (1997), Generalized score test of homogeneity based on correlated random effect models, *J Roy. Statist. Soc. B* **59**, 157-171.

[3] Cox, D. R. and Snell, E. J. (1968), A general definition of residuals, *J Roy. Statist. Soc. B* **30**, 248-75.

[4] Drylewicz, J., Commenges, D. and Thiébaut (2010), Score tests for exploring complex models: application to HIV dynamics model, *Biometrical Journal*, **52**, 10-21.

[5] Hall, P. and Wilson, S.R. (1991), Two guidelines for bootstrap hypothesis testing. *Biometrics*, **47**, 757-762.

[6] Kalbfleisch, J.D. (1978), Likelihood methods and nonparametric tests. *J. Amer. Statist. Assoc.* **73**, 167-170 .

[7] Lehmann, E.L. (1986), *Testing Statistical Hypotheseses* New-York: Wiley.

[8] Mantel, N. (1967), The detection of disease clustering and a generalized regression approach, *Cancer Research* **27** Part 1, 209-220.

[9] Orgogozo, JM., Dartigues, JF., Lafont, S., Letenneur, L., Commenges D., Salamon R., Renaud, S. and Breteler, M. (1997), "Wine consumption and dementia in the elderly: a prospective community study in the Bordeaux area," *Revue Neurologique* 153, 185-192.

[10] Schmoyer, R. L. (1994), Permutation tests for correlation in regression errors. *Journal of the American Statistical Association* **89**, 1507-1516.

[11] Williams, D. (1991), *Probability with martingales*, Cambridge: Cambridge University Press.