glmperm: A Permutation of Regressor Residuals Test for Inference in Generalized Linear Models

by Wiebke Werft and Axel Benner

Abstract We introduce a new R package called **glmperm** for inference in generalized linear models especially for small and moderate-sized data sets. The inference is based on the permutation of regressor residuals test introduced by Potter (2005). The implementation of **glmperm** outperforms currently available permutation test software as **glmperm** can be applied in situations where more than one covariate is involved.

Introduction

A novel permutation test procedure for inference in logistic regression with small- and moderate-sized datasets was introduced by Potter (2005) and showed good performance in comparison to exact conditional methods. This so-called permutation of regressor residuals (PRR) test is implemented in the R package **logregperm**. However, the application field is limited to logistic regression models. The new **glmperm** package offers an extension of the PRR test to generalized linear models (GLMs) especially for small and moderate-sized data sets. In contrast to existing permutation test software, the **glmperm** package provides a permutation test for situations in which more than one covariate is involved, e.g. if established covariates need to be considered together with the new predictor under test.

Generalized linear models

Let $\boldsymbol{Y}$ be a random response vector whose components are independently distributed with means $\boldsymbol{\mu}$. Furthermore, let the covariates $\boldsymbol{x}_1,...,\boldsymbol{x}_p$ be related to the components of $\boldsymbol{Y}$ via the generalized linear model

$$\begin{aligned} E(Y_i) &= \mu_i, && (1)\\ g(\mu_i) &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, && (2)\\ Var(Y_i) &= \frac{\phi}{w_i}V(\mu_i), && (3) \end{aligned}$$

with $i = 1,...,n$, g a link function and $V(\cdot)$ a known variance function; ϕ is called the dispersion parameter and w_i is a known weight that depends on the underlying observations. The dispersion parameter is constant over observations and might be unknown, e.g. for normal distributions and for binomial and Poisson distributions with over-dispersion (see below).

One is now interested in testing the null hypothesis that the regression coefficient for a covariate of interest, say without loss of generality $\boldsymbol{x}_1$, is zero, i.e. $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Let $\boldsymbol{y}$ be the observed vector of the outcome variable $\boldsymbol{Y}$. Inference is based on the likelihood-ratio test statistic $LR(\boldsymbol{y}; \boldsymbol{x}_1,...,\boldsymbol{x}_p)$, which is defined as the difference in deviances of the models with and without the covariate of interest divided by the dispersion parameter. For simplicity we assume that each variable is represented by a single parameter that describes the contribution of the variable to the model. Therefore, the test statistic has an asymptotic chi-squared distribution with one degree of freedom. (For a deeper look into generalized linear models McCullagh and Nelder (1989) is recommended.)

The PRR test

The basic concept of the PRR test is that it replaces the covariate of interest by its residual $\boldsymbol{r}$ from a linear regression on the remaining covariates $\boldsymbol{x}_2,...,\boldsymbol{x}_p$. This is a simple orthogonal projection of $\boldsymbol{x}_1$ on the space spanned by $\boldsymbol{x}_2,...,\boldsymbol{x}_p$ and ensures that $\boldsymbol{r}$ by its definition is not correlated with $\boldsymbol{x}_2,...,\boldsymbol{x}_p$ while $\boldsymbol{x}_1$ may be correlated. An interesting feature of this projection is that the maximum value of the likelihood for a generalized linear model of $\boldsymbol{y}$ on $\boldsymbol{r},\boldsymbol{x}_2,...,\boldsymbol{x}_p$ is the same as that for $\boldsymbol{y}$ on $\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_p$. Hence, the likelihood-ratio test is the same when using the residuals $\boldsymbol{r}$ instead of the covariate $\boldsymbol{x}_1$. This leads to the idea of using permutations of the residuals $\boldsymbol{r}$ to estimate the null distribution and thus the p-value.

The algorithm for the PRR test to obtain a p-value for testing the null hypothesis $H_0 : \beta_1 = 0$ of the model

$$g(\boldsymbol{\mu}) = \beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + ... + \beta_p\boldsymbol{x}_p \qquad (4)$$

is as follows:

1. Calculate least squares residuals $\boldsymbol{r} = (r_1,...,r_n)$ via

   $$\boldsymbol{r} = \boldsymbol{x_1} - \left(\hat{\gamma}_0 + \hat{\gamma}_1\boldsymbol{x}_2 + ... + \hat{\gamma}_{p-1}\boldsymbol{x}_p\right) \qquad (5)$$

   where $\hat{\gamma}_0, \hat{\gamma}_1, ..., \hat{\gamma}_{p-1}$ are least squares estimates of the coefficients from the linear model

   $$E(\boldsymbol{x}_1) = \gamma_0 + \gamma_1\boldsymbol{x}_2 + ... + \gamma_{p-1}\boldsymbol{x}_p\,. \qquad (6)$$

   Then derive the p-value $\tilde{p}$ of the likelihood-ratio test statistic $LR(\boldsymbol{y}; \boldsymbol{r}, \boldsymbol{x}_2,...,\boldsymbol{x}_p)$ based on $\boldsymbol{r}$ replacing the covariate $\boldsymbol{x}_1$.

2. For resampling iterations $b = 1, ..., B$
   - Randomly draw $\boldsymbol{r}^*$ from $\boldsymbol{r} = (r_1, ..., r_n)$ without replacement.
   - Calculate p-values p_b^* of the likelihood-ratio test statistic $LR(\boldsymbol{y}; \boldsymbol{r}^*, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)$.
3. Calculate a permutation p-value for the PRR test according to

$$p_f = \frac{\#(p_b^* \leq f \cdot \tilde{p})}{B}. \qquad (7)$$

Thus, the permutation p-value p_f is the fraction of permutations that have a likelihood-based p-value less than or equal to that for the unpermuted data times a factor f. This factor f is equal or slightly bigger than one, i.e. $f \in \{1; 1.005; 1.01; 1.02; 1.04\}$. It was introduced by Potter (2005) to account for numerical instabilities which might occur when the algorithms to maximize the likelihood do not converge to exactly the same value for different configurations of the data. Note that varying the factor only changes the results by a few per cent or less. In the R package **glmperm** various factors are implemented and are displayed in the summary output. Here, the result presentations will be restricted to factors 1 and 1.02.

The number of resampling iterations B of the algorithm is implemented as `nrep` in the function `prr.test`. By default `nrep=1000`. However, depending on the precision that is desired, the number of iterations could be increased at the cost of longer computation time. We recommend using `nrep=10000`.

In order to provide an estimate of the variance of the result, the permutation p-value p_f could be regarded as a binomial random variable $\mathcal{B}(B, p)$, where B is the number of resampling iterations and p is the unknown value of the true significance level (Good, 2005). As estimate of p the observed p-value $\tilde{p}$ is used multiplied by the factor f. Hence, the standard error for the permutation p-value p_f is provided by $\sqrt{f\tilde{p}(1 - f\tilde{p})/B}$.

For binary outcome variables the PRR test is a competitive approach to exact conditional logistic regression described by Hirji et al. (1987) and Mehta and Patel (1995). The commercial LogXact software has implemented this conditional logistic regression approach. At present, statistical tests employing unconditional permutation methods are not commercially available and the package **glmperm** bridges this gap.

The rationale of exact conditional logistic regression is based on the exact permutation distribution of the sufficient statistic for the parameter of interest, conditional on the sufficient statistics for the specified nuisance parameters. However, its algorithmic constraint is that the distribution of interest will be degenerate if a conditioning variable is continuous. The difference between the two procedures lies in the permutation scheme: While the conditional approach permutes the outcome variable, the PRR test permutes the residuals from the linear model. Note that if one considers regressions with only a single regressor the residuals $\boldsymbol{r}$ are basically equal to $\boldsymbol{x}_1$, and the PRR test reduces to a simple permutation test (shuffle-Z method, see below).

An overview of the methodologic differences between these two and other permutation schemes is provided by Kennedy and Cade (1996). In the context of their work the permutation method used for the PRR test can be viewed as an extension of their shuffle-R permutation scheme for linear models. The variable associated with the parameter under the null hypothesis is regressed on the remaining covariables and replaced by the corresponding residuals; these residuals are then permuted while the response and the covariates are held constant. Freedman and Lane (1983) introduced the shuffle-R method in combination with tests of significance and a detailed discussion of this permutation scheme is provided by ter Braak (1992). The conditional method can be implied with the shuffle-Z method which was first mentioned in the context of multiple regression by Draper and Stoneman (1966). Here the variables associated with the parameters being tested under the null hypothesis are randomized while the response and the covariates are held constant. Kennedy and Cade (1996) discuss the potential pitfalls of the shuffle-Z method and point out that this method violates the ancillarity principle by not holding constant the collinearity between the covariables and the variable associated with the parameter under the null hypothesis. Therefore, Kennedy and Cade (1996) do not recommend the shuffle-Z method unless it employs a pivotal statistic or the hypothesized variable and the remaining covariables are known to be independent.

Modifications for the extension to GLMs

Three main modifications have been implemented in the **glmperm** package in comparison to the **logregperm** package. Note that the new package **glmperm** includes all features of the package **logregperm**. At present, the **logregperm** package is still available on CRAN. In general, the **glmperm** package could replace the **logregperm** package.

1. The extension of the `prr.test` function in the **glmperm** package provides several new arguments compared to the version in **logregperm**. The input is now organised as a formula expression equivalent to fitting a generalized linear model with `glm` (R package **stats**). Hence, for easier usage the syntax of `prr.test` is adapted from `glm`. The covariable of inter-

est about which inference is to be made is to be included as argument `var='x1'`. An argument `seed` is provided to allow for reproducibility.

2. The implicit function `glm.perm` is extended to calculate not only the deviances for the different resampling iterations but also the dispersion parameter for each permutation separately. For Poisson and logistic regression models the dispersion parameters are pre-defined as $\phi = 1$ for all iterations; the likelihood-ratio test statistic is then simply the difference of deviances. For all other GLM the dispersion parameters will be estimated for each resampling iteration based on the underlying data. This ensures that the p-value of the likelihood-ratio test statistic for this precise resampling iteration is accordingly specified given the data.

3. A new feature of the package is that it includes a summary function to view the main results of `prr.test`. It summarises the permutation of regressor residual-based p-value p_f for the various factors f, the observed likelihood-ratio test statistic and the observed p-value $\tilde{p}$ based on the chi-squared distribution with one degree of freedom. For Poisson and logistic regression models a warning occurs in case of possible overdispersion ($\phi > 1.5$) or underdispersion ($\phi < 0.5$) and recommends use of `family=quasibinomial()` or `quasipoisson()` instead.

An example session

To illustrate the usage of the **glmperm** package for a GLM, an example session with simulated data is presented. First, we simulated data for three independent variables with $n = 20$ samples and binary and discrete response variables for the logistic and Poisson regression model, respectively.

```
# binary response variable
n <- 20
set.seed(4278)
x1 <- rnorm(n)
x0 <- rnorm(n)+x1
y1 <- ifelse(x0+x1+2*rnorm(n)>0,1,0)
test1 <- prr.test(y1~x0+x1,
          var="x0", family=binomial())
x2 <- rbinom(n,1,0.6)
y2 <- ifelse(x1+x2+rnorm(n)>0,1,0)
test2 <- prr.test(y2~x1+x2, var="x1",
          nrep=10000,family=binomial())

# Poisson response variable
set.seed(4278)
x1 <- rnorm(n)
x0 <- rnorm(n) + x1
nu <- rgamma(n, shape = 2, scale = 1)
y <- rpois(n, lambda = exp(2) * nu)
test3 <- prr.test(y~x0+x1,
          var="x0", family=poisson())
test4 <- prr.test(y~x0, var="x0",
          nrep=1000,family=poisson())
```

A condensed version of the displayed result summary of `test2` (only factors $f = 1$ and $f = 1.02$ are shown) is given by:

```
> summary(test2)
-----------------------------------------
Results based on chi-squared distribution
-----------------------------------------
observed p-value: 0.0332
--------------------
Results based on PRR
--------------------
permutation p-value for simulated p-values <=
observed p-value: 0.0522 (Std.err: 0.0018)
permutation p-value for simulated p-values <=
1.02 observed p-value: 0.0531 (Std.err: 0.0018)
```

For the above example `test2` the exact conditional logistic regression p-value calculated via LogXact-4 is 0.0526, whereas the p-value of the PRR test is 0.0522 for factor $f = 1$, and based on the chi-squared distribution it is 0.0332. The example demonstrates that the p-value obtained via PRR test (or LogXact) leads to a different rejection decision than a p-value calculated via an approximation by the chi-squared distribution. Hence, when small sample sizes are considered the PRR test should be preferred over an approximation via chi-squared distribution.

Special case: Overdispersion

For the computation of GLM the dispersion parameter ϕ for the binomial and Poisson distribution is set equal to one. However, there exist cases of data distribution which violate this assumption. The variance in (3) can then be better described when using a dispersion parameter $\phi \neq 1$. The case of $\phi > 1$ is known as overdispersion as the variance is larger than expected under the assumption of binomial or Poisson distribution. In practice, one can still use the algorithms for generalized linear models if the estimation of the variance takes account of the dispersion parameter $\phi > 1$. As a consequence of overdispersion the residual deviance is then divided by the estimation $\hat{\phi}$ of the dispersion factor instead of $\phi = 1$. Hence, this has direct influence on the likelihood-ratio test statistic which is the difference in deviances and therefore is also scaled by $\hat{\phi}$. The corresponding p-values differ if overdispersion is considered or not, i.e. if $\phi = 1$ or $\phi = \hat{\phi}$ is used. In the PRR test one can account for overdispersion when using `family=quasipoisson()` or quasibinomial() instead of `family=poisson()` or binomial().

We experienced a stable performance of the PRR test for overdispersed data. The following treepipit data (Müller and Hothorn, 2004) provides an example of overdispersed data. We show the results of the chi-squared approximation of the likelihood-ratio test statistic as well as the results of the PRR test for the usage of `family=poisson()` and `family=quasipoisson()`, respectively.

```
# example with family=poisson()
data(treepipit, package="coin")
test5<-prr.test(counts~cbpiles+coverstorey
 +coniferous+coverregen,data=treepipit,
 var="cbpiles",family=poisson())
summary(test5)
-----------------------------------------
Results based on chi-squared distribution
-----------------------------------------
observed p-value: 0.0037
--------------------
Results based on PRR
--------------------
permutation p-value for simulated p-values <=
observed p-value: 0.083 (Std.err: 0.0019)
permutation p-value for simulated p-values <=
1.02 observed p-value: 0.084 (Std.err: 0.0019)

# example with family=quasipoisson()
test6<-prr.test(counts~cbpiles+coverstorey
 +coniferous+coverregen,data=treepipit,
 var="cbpiles",family=quasipoisson())
summary(test6)
-----------------------------------------
Results based on chi-squared distribution
-----------------------------------------
observed p-value: 0.0651
--------------------
Results based on PRR
--------------------
permutation p-value for simulated p-values <=
observed p-value: 0.07 (Std.err: 0.0078)
permutation p-value for simulated p-values <=
1.02 observed p-value: 0.071 (Std.err: 0.0079)
```

The p-values based on the chi-squared distribution of the likelihood-ratio test statistic are $p = 0.0037$ and $p = 0.0651$ when using `family=poisson()` and `family=quasipoisson()`, respectively. Hence, a different test decision is made whereas the test decision for the PRR test is the same for both cases ($p = 0.083$ and $p = 0.07$).

Summary

The **glmperm** package provides a permutation of regressor residuals test for generalized linear models. This version of a permutation test for inference in GLMs is especially suitable for situations in which more than one covariate is involved in the model. The key input feature of the PRR test is to use the orthogonal projection of the variable of interest on the space spanned by all other covariates instead of the variable of interest itself. This feature provides a reasonable amendment to existing permutation test software which do not incorporate situations with more than one covariate. Applications to logistic and Poisson models show good performance when compared to gold standards. For the special case of data with overdispersion the PRR test is more robust compared to methods based on approximations of the test statistic distribution.

Acknowledgements

We would like to acknowledge the many valuable suggestions made by two anonymous referees.

Bibliography

C.J.F. ter Braak. Permutation versus bootstrap significance tests in multiple regression and anova. In K.H. Jöckel, G. Rothe and W. Sendler, editors. *Bootstraping and Related Techniques*, 79-85, Springer, 1992.

N.R. Draper and D.M. Stoneman. Testing for the inclusion of variables in linear regression by a randomization technique. *Technometrics*, **8**: 695-698, 1966.

D. Freedman and D. Lane. A nonstochastic interpretation of reported significance levels. *Journal of Business and Economic Statistics*, **1**:292-298, 1983.

P. Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses, 3rd edn.* Springer series in statistics, 2005. ISBN 0-387-20279-X.

K.F. Hirji, C.R. Mehta and N.R. Patel. Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, **82**:1110-1117, 1987.

P.E. Kennedy and B.S. Cade. Randomization tests for multiple regression. *Communications in Statistics - Simulation and Computation*, **25(4)**:923-936, 1996.

D.M. Potter. A permutation test for inference in logistic regression with small- and moderate-sized datasets, *Statistics in Medicine*, **24**:693-708, 2005.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, 2nd edn.* Chapman and Hall, London, 1989. ISBN 0-412-31760-5.

C.R. Mehta and N.R. Patel Exact logistic regression: theory and examples. *Statistics in Medicine*, **14**:2143-2160, 1995.

J. Müller and T. Hothorn. Maximally selected two-sample statistics as a new tool for the identification and assessment of habitat factors with an application to breeding bird communities in oak forests, *European Journal of Forest Research*, **123**:219-228, 2004.

Wiebke Werft
German Cancer Research Center
Im Neuenheimer Feld 280, 69120 Heidelberg
Germany
`w.werft@dkfz.de`

Axel Benner
German Cancer Research Center
Im Neuenheimer Feld 280, 69120 Heidelberg
Germany
`benner@dkfz.de`