mcprofile package vignette

Daniel Gerhard

Version 0.0-5

Contents

1 Overview

The function `mcpcalc` calculates signed root deviance profiles for objects of class `lm`, `glm`, and `nls`. The parameters of interest can be specified by defining a contrast matrix, allowing also for profiles of parameter linear combinations. The profiles themselves are calculated by conditional optimization for multiple points in the neighbourhood around the parameter estimate. At each of these points the signed root deviance statistic is calculated. To obtain a continuous profile function, interpolation splines are fitted.

Confidence intervals can be calculated by the determination of the intersection between the profile and a critical value, projecting this cutpoint on the scale of the parameter of interest. For simultaneous confidence intervals, controlling the family-wise error rate (FWER) at a specified level α, this value is calculated as a quantile of a multivariate normal- or t-distribution. Multiple tests are calculated analogous by calculating the signed root deviance statistic at the test margin and calculating the probability of locating this statistic under the null hypothesis, assuming a multivariate normal- or t-distribution.

2 GLM profiles

2.1 Generalized linear models

Given a vector of observations y_i with $i = 1, \ldots, N$ as a realization of a random variable Y_i a generalized linear model can be assumed, to capture as much variability in the data by an unknown parameter vector β_j with $j = 1, \ldots, k$. This parameter vector can be linked to the data by a designmatrix of covariates x_{ij}, specifying the parameter layout. Model predictions are obtained by multiplying the design matrix with the unknown parameter vector β_j

$$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j$$

The linear predictor $\boldsymbol{\eta}$ is calculated on a given link function $\tau(\cdot)$, for which the inverse can be used to obtain the predictions on the original scale

$$\eta_i = \tau(\mu_i).$$

Under the assumption that each component of Y_i has a distribution in the exponential family, the density takes the form

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$$

for specific functions $a(\phi), b(\theta), c(y, \phi)$. As the parameters should be estimated conditional on an observation y the log likelihood function can be written as

$$l(\theta, \phi; y) = \log\left(f_Y(y; \theta, \phi)\right).$$

The parameter vector $\hat{\beta}$ can be estimated by maximizing the (log) likelihood $l(\mu_i; y_i)$ or equivalently minimizing the scaled deviance

$$D(\mu_i; y_i) = 2\left(l(y_i; y_i) - l(\mu_i; y_i)\right)$$

with respect to μ_i.

2.2 Parameter linear combinations

If the parameters are not directly of interest, but linear combinations of them, these can be specified by a contrast matrix C_{mj}, with $m = 1, \ldots, M$, and $j = 1, \ldots, k$. Parameter linear combinations are then calculated by

$$\psi_m = \sum_{j=1}^{k} C_{mj}\beta_j$$

The corresponding $(M \times M)$ variance-covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\psi}}$ for these linear combinations is calculated by

$$\boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{C}^T,$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the $(k \times k)$ variance-covariance matrix for the parameter vector $\boldsymbol{\beta} = (\beta_j)$.

2.3 Signed root deviance profiles

To evaluate the deviance near the parameter estimate $\hat{\psi}_m$, multiple points in its neighbourhood $\psi_m^\star$ may be observed. Therefore the deviance $D(\psi_m^\star, \hat{\beta}_j)$ is calculated for different, fixed $\psi_m^\star$ treating the β_i as nuisance parameters. But directly estimating $\hat{\beta}_j$ for a fixed $\psi_m^\star$ is difficult, as $\psi_m^\star$ is defined as a linear combination of the $\hat{\beta}_j$, leading to an overparameterized model.

This problem can be solved by conditioning on a parameter β_j with $C_{mj} \neq 0$, and estimating only $k-1$ parameters $\beta_{j'}$ with $j' \neq j$. The $\hat{\beta}_j$ are therefore estimated with the constraint

$$\psi_m - \sum_{j'=1}^{k-1} c_{mj'}\beta_{j'} = c_{mj}\beta_j.$$

The deviance statistic can be reformulated as a signed root deviance statistic by

$$\hat{r}(\psi_m^\star) = sign\left(\psi_m^\star - \hat{\psi}_m\right)\sqrt{\frac{D(\psi_m^\star, \hat{\beta}_j)}{\phi}},$$

wher ϕ is an additional scaling parameter, e.g. the residual variance when assuming a Gaussian distribution for the observations, or an overdispersion parameter in a quasilikelihood model.

At last the profile function $\hat{r}(\psi_m)$ is estimated by an interpolation spline given multiple estimates for $\hat{r}(\psi_m^\star)$.

3 Simultaneous confidence intervals

Under the assumption that $\psi_m \overset{a}{\sim} N(\psi_m, \widehat{\mathbf{\Sigma}}_{\hat{\psi}})$, M simultaneous confidence intervals can be calculated by

$$\{\psi_m : -z \leq \hat{r}(\psi_m) \leq z\}.$$

An adequate cutoff value z can be computed as the twosided, $1-\alpha$ quantile of a multivariate normal distribution. As the correlation structure of this multivariate normal distribution is unknown, it is approximated by the estimated correlation structure, obtained by standardizing the estimated variance covariance matrix $\widehat{\mathbf{\Sigma}}_{\hat{\psi}}$. If a residual degree of freedom is available, a corresponding multivariate t-distribution may be used instead of calculating the quantile from a Gaussian distribution.

4 A linear model example

In the dataset `cholesterol{multcomp}` the reduction of the cholesterol level is observed for five different treatments. Three different formulations of a drug (20mg once, 10mg twice, and 5mg four times a day), and two competing drugs as control groups were tested. Purpose of the study is to find formulations of the drug, which show a more efficient cholesterol reduction than the control groups.

First, a linear model is used to estimate the marginal means for each treatment:

```
> library(mcprofile)
> data(cholesterol)
> mod <- lm(response ~ trt - 1, data = cholesterol)
```

To specify the comparisons of interest, a contrast matrix has to be defined; this matrix is chosen to reproduce the results in Westfall (1999) comparing all formulations with each other, and comparing each of the two controls separately with the pooled means of the three formulations.

```
> K <- contrMat(table(cholesterol$trt), type = "Tukey")
> Ksub <- rbind(K[c(1, 2, 5), ], `D - test` = c(-1, -1, -1, 3,
+     0)/3, `E - test` = c(-1, -1, -1, 0, 3)/3)
> Ksub
```

Profiles are calculated by

```
> mp <- mcpcalc(mod, Ksub)
```

with corresponding simultaneous confidence intervals:

```
> (ci <- confint(mp, adjust = "single-step"))
```

These are the same as calculated with package `multcomp`, due to the linear model assumptions:

```
> (confint(glht(mod, linfct = Ksub)))
```

Confidence intervals are plotted by

```
> par(mar = c(5, 8, 4, 2) + 0.1)
> plot(ci)
> abline(v = 0, lty = 2)
```

or directly in a profile plot by

```
> print(plot(mp, ci))
```

5 GLM example

In a cell transformation experiment, Balb/c 3T3 cells are treated with different concentrations of a carcinogen. Cells treated with a carcinogen will not stop proliferation; therefore the number of foci (cell accumulations), counted for 10 replicates per concentration level, are a measure of carcinogenicity.

If no specific dose-response relationship between the concentrations of the carcinogen and the foci number should be assumed, a comparison of each concentration level to a negative control might be performed. In a first step a Poisson GLM is used to estimate the concentration means (on a logarithmic scale).

```
> data(cta)
> cta$Conc <- factor(cta$conc, levels = unique(cta$conc))
> gmod <- glm(foci ~ Conc - 1, data = cta, poisson(link = "log"))
```

The parameter differences of interest are specified by a 'Dunnett'-type contrast matrix.

```
> (K <- contrMat(table(cta$Conc), type = "Dunnett"))
```

For these contrast parameters profiles are calculated and plotted.

```
> gmp <- mcpcalc(gmod, K)
> print(plot(gmp, layout = c(4, 2)))
```

Also a fixed parameter range can be defined for the profile calculation.

```
> gmp2 <- mcpcalc(gmod, K, margin = c(-8, 8))
> print(plot(gmp2, layout = c(4, 2)))
```

The function `mcprofileControl` provides several control arguments, which enable for example to compute the profiles for a fixed number of points with fixed distances. Profiles with only four equally distant supporting points on each side are calculated by

```
> gmp3 <- mcpcalc(gmod, K, margin = c(-8, 8), control = mcprofileControl(steps = 4,
+     fixed.range = TRUE))
> print(plot(gmp3, layout = c(4, 2)))
```

5.1 Confidence intervals

Simultaneous confidence intervals can be calculated, assuming a multivariate normal distribution for the parameters of interest. As the mean differences are calculated on the log link, a transformation back by the exponent results in ratio of mean parameters. The transformation can be conveniently performed by the `exp` function.

```
> (eci <- exp(confint(gmp, adjust = "single-step")))
```

If in this experiment a safety evaluation might be of interest, the multiplicity adjustment should be omitted and a local error rate can be assumed for each interval. When using the confidence interval for equivalence testing, each interval limit is compared with predefined equivalence margins; if both confidence limits are located within the two margins, safety can be concluded. As this decision applies to both, the upper and the lower limit, simultaneously, a corresponding test decision is obtained at an error level of $2 \times \alpha$ for all intervals.

```
> confint(gmp, level = 0.9, adjust = "none")
```

Only increasing carcinogenicity is of interest; therefore onesided confidence intervals can be calculated, reallocating the critical value by shifting some areas with a higher likelihood into the rejection region of a corresponding hypothesis test in favour of areas with a lower likelihood, which are not in the test direction of interest.

```
> confint(gmp, alternative = "greater", adjust = "none")
```

5.2 Multiple hypothesis testing

As the confidence intervals, calculated in the previous section, are only inversions of a multiple hypotheses test, the computation of adjusted p-values is also directly available. These are calculated as the probability of obtaining a more 'extreme' value than the signed root of the deviance statistic under the null hypothesis.

```
> (pvals <- test(gmp, adjust = "single-step"))
> print(plot(gmp, pvals, layout = c(4, 2)))
```

Multiplicity adjustment is performed by calculating the p-values directly from a multivariate distribution, or choosing any adjustment method provided by the function `p.adjust`, e.g. the stepwise procedure of Holm.

```
> test(gmp, adjust = "holm")
```

5.3 Quadratic- and higher order approximations

The Wald-type confidence intervals obtained by the `multcomp` package

```
> (round(exp(confint(glht(gmod, linfct = K))$confint), 2))
```

can be reproduced by recalculating the profiles by substituting the profile deviance statistic with a Wald-type statistic

$$t(\psi_m) = j(\hat{\psi}_m)^{\frac{1}{2}} \left(\hat{\psi}_m - \psi_m \right),$$

where $j(\hat{\psi}_m)$ is the observed Fisher information function.

```
> wmp <- wald(gmp)
> (exp(confint(wmp, adjust = "single-step")))
```

A plot for comparing two different profiles is also available.

The `squared=TRUE` argument changes the scale of the y-axis directly to the deviance statistic instead of the signed root version; hence difference of the profiles to a quadratic function can clearly be seen in this plot.

```
> print(plot(gmp, wmp, squared = TRUE, layout = c(4, 2)))
```

Additionally, profiles based on higher order approximation (Brazzale et al.) can be computed.

```
> hmp <- hoa(gmp)
> (exp(confint(hmp, adjust = "single-step")))
```

These sattlepoint approximations are based on the statistic

$$r^*(\psi_m) = r(\psi_m) + \frac{1}{r(\psi_m)} \log \left(\frac{q(\psi_m)}{r(\psi_m)} \right),$$

with the Wald statistic

$$q(\psi_m) = \rho(\psi_m, \hat{\psi}_m) t(\psi_m)$$

adding a nuisance parameter adjustment, relating the variance estimates of the full model to the conditional estimates by

$$\rho(\psi_m, \hat{\psi}_m) = \left(\frac{\left| j_{\beta\beta} \left(\hat{\psi}_m, \hat{\beta}_i \right) \right|}{\left| j_{\beta\beta} \left(\psi_m, \hat{\beta}_i \right) \right|} \right).$$

6 Ratios of normal means

Instead of investigating the difference of parameters by constructing simple linear combinations, also the ratio of parameter linear combinations may be of interest. Profiling for this problem can be done by optimization conditional to the ratio of parameters defined by separate contrast matrices for the numerator and denominator.

```
> data(Penicillin)
> Penicillin$strain <- as.factor(Penicillin$strain)
> linmod <- lm(diameter ~ strain - 1, data = Penicillin)
> CM <- contrMatRatio(table(Penicillin$strain), type = "Tukey")
> mpr <- mcpcalcRatio(linmod, CM$numC, CM$denC)
> (cir <- confint(mpr, adjust = "single-step"))

> plot(cir)
> abline(v = 1, lty = 2)
```