

# SEMI-PARAMETRIC GENERALIZED LINEAR MODEL FOR CORRELATED BINARY DATA

ANIKO SZABO

Note: this setup assumes that parts defined in `SPregress.w` are available.

## 1. INTRODUCTION

We extend the Rathouz-Gao semi-parametric generalized linear model to clustered binary outcomes with varying cluster sizes. Treating the number of events  $Y_i$  from a cluster scaled by corresponding cluster size  $n_i$  as the response variable, we use a parametric regression model for the marginal response probability  $Y/n_i$  and assume that for the maximum cluster size  $N$  the probability density distribution of  $Y_i$  is an exponentially tilted version of the reference density distribution.

The conditional mean model is as follows:

$$E\left(\frac{Y_i}{n_i} | \mathbf{Z}_i; \boldsymbol{\beta}\right) = \mu(\mathbf{Z}_i, \boldsymbol{\beta}) \equiv \mu_i = h^{-1}(\mathbf{Z}_i^T \boldsymbol{\beta}) \quad (1)$$

where  $\mathbf{Z}_i$  is a matrix of cluster-level covariates, the function  $h(\cdot)$  is a known strictly increasing link function.

The assumed probability density function for a cluster of maximal size  $N$  is:

$$q_{y,N}(\mathbf{Z}, \boldsymbol{\beta}) \propto q_{y,N}^{(0)} \times \exp[-\omega(\mathbf{Z}, \boldsymbol{\beta}) \times y] \quad (2)$$

where  $q_{y,N}^{(0)}$  is a non-parametric reference distribution, and  $\omega(\mathbf{Z}, \boldsymbol{\beta})$  is the tilting parameter chosen so that the expectation of  $Y/N$  equals to the value defined by the mean model (1), i.e.  $\sum_{y=0}^N \frac{y}{N} q_{y,N}(\mathbf{Z}, \boldsymbol{\beta}) = \mu = h^{-1}(\mathbf{Z}^T \boldsymbol{\beta})$ .

For the purpose of identifiability, the mean of marginal probability  $\{q_{y,N}^{(0)}\}$  is set at an arbitrary fixed value, which can be the mean of the reference group, or the overall marginal mean of the data. We extend the conditional mean and probability density model for cluster sizes smaller than the maximal size  $N$  by assuming marginal compatibility.

## 2. MAIN FUNCTION

```
"../R/SPGLM.R" 1==
```

```
#' Fit semi-parametric GLM
#
#'@rdname spglm
#'@param formula a one-sided formula of the form \code{cbind(r, s) ~ predictors} where \code{r} and \code{s} are
#'@param data an optional matrix or data frame containing the variables in the formula \code{formula}. B
#'@param subset an optional vector specifying a subset of observations to be used.
#'@param weight an optional vector specifying observation weights.
#'@param link a link function for the mean.
#'@param start an optional list with elements named \code{beta}, \code{q}, and/or \code{mu1} giving s
#'@param control a list with parameters controlling the algorithm.
#'@return an object of class \code{spglm} with the fitted model.
#'@export
#' @importFrom stats terms model.matrix

#' Semi-parametric generalized linear model
spglm <- function(formula, data, subset, weights, link="logit", start=NULL, control=list(eps=0.001, maxit=100), fam <- binomial(link=link))
```

---

Date: March 4, 2021.

```

  { Create model matrix from formula and data 2 }
  { Fit model 4 }

  mt <- attr(mf, "terms")
  res <- list(coefficients = beta_new, q=q_new, niter = iter, loglik=logl,
             link = link, call = mc, terms = mt,
             xlevels = .getXlevels(mt, mf),
             data_object=list(model_matrix=mm, resp=Y, n=rowSums(Y), weights=weights),
             model_fun=model_fun)
  class(res) <- "spglm"
  res
}
◇

```

{ Create model matrix from formula and data 2 } ≡

```

if (missing(formula) || (length(formula) != 3L))
  stop("'formula' missing or incorrect")
if (missing(data))
  data <- environment(formula)
mc <- match.call(expand.dots = FALSE)
m <- match(c("formula", "data", "subset", "weights"), names(mc), 0L)
m <- mc[c(1L,m)]
if (is.matrix(eval(m$data, parent.frame()))))
  m$data <- as.data.frame(data)
m[[1L]] <- quote(stats::model.frame)
mf <- eval(m, parent.frame())

# extract and check response variable
Y <- model.response(mf)
if (ncol(Y) != 2) stop("The response variable should be a 2-column matrix")

# create model matrix
mm <- model.matrix(formula, data=mf)

# extract weights
weights <- as.vector(model.weights(mf))
if (!is.null(weights) && !is.numeric(weights))
  stop("'weights' must be a numeric vector")
if (!is.null(weights) && any(weights < 0))
  stop("negative weights not allowed")
◇

```

Fragment referenced in [1](#).

### 3. EM ALGORITHM FOR MODEL FITTING

We will combine an Expectation Maximization with the Newton-Raphson algorithm proposed by Rathouz and Gao (2009) to estimate the model parameters. Recall that the observed data consists of the number of events, cluster size, and cluster-level covariates:  $(y_i, N_i, \mathbf{Z}_i)$ ,  $i = 1, \dots, I$ .

3.0.1. *Observed data log-likelihood.* Under the marginal compatibility assumption, the observed log-likelihood function is:

$$\begin{aligned}\ell(q_{y,N}, \beta) &= \sum_{i=1}^I \log\{q_{y_i, N_i}(\mathbf{Z}_i, \beta)\} \\ &= \sum_{i=1}^I \log \left[ \binom{N_i}{y_i} \sum_{t=y}^{N-N_i+y_i} \frac{\binom{N-N_i}{t-y_i}}{\binom{N}{Y_i}} \times \frac{q_{t,N}^{(0)} \times \exp\{\omega(\mathbf{Z}_i; \beta)t\}}{\sum_{t'=0}^N q_{t',N}^{(0)} \times \exp\{\omega(\mathbf{Z}_i; \beta)t'\}} \right].\end{aligned}\quad (3)$$

It is difficult to utilize the observed log-likelihood directly for parameter estimation. Its score function is complex because there is a logarithm of a sum in (3). Therefore, we implement an EM algorithm for parameter estimation.

3.0.2. *Missing data setup.* Stefanescu et al (2003) have proved that the marginal compatibility assumption is equivalent to assuming that each cluster of size  $N'$  arises from a cluster of maximal cluster size  $N$  with some binary observations missing completely missing at random (MCAR). The missing data setup and the expectation step from EM algorithm are based on this MCAR interpretation.

For the setup of the EM algorithm, the complete data are clusters of size  $N$  with a corresponding number of events  $s_i$ , where  $s_i \in \{y_i, \dots, N\}$ . The missing data correspond to the unobserved outcomes in of the  $(N - N_i)$  cluster components. The missing data can be summarized as  $(s_i - y_i)$  events out of  $(N - N_i)$  elements.

The complete data log-likelihood is:

$$\begin{aligned}\ell_{\text{complete}}(q_{y,N}, \beta \mid N_i, s_i) &= \sum_{i=0}^I \log\{q_{s_i, N}(\mathbf{Z}_i; \beta)\} \\ &= \sum_{i=0}^I \sum_{s=y_i}^N \mathbb{1}(s_i = s) \times \log\{q_{s, N}(\mathbf{Z}_i; \beta)\},\end{aligned}\quad (4)$$

where  $q_{s_i, N}$  is the probability of achieving  $s_i$  events in a cluster of size  $N$  in the complete data set. The  $\mathbb{1}$  denotes the indicator function.

3.0.3. *Expectation step.* We take the expectation of complete data log-likelihood (4) given parameters from the previous  $k^{\text{th}}$  iteration, which is shown as follows:

$$\begin{aligned}Q\left(q_{y,N}^{(k)}, \beta^{(k)}\right) &= \mathbb{E} \left\{ \ell_{\text{complete}}(q_{y,N}, \beta \mid N_i, s_i) \mid q_{y,N}^{(k)}, \beta^{(k)}, N_i, y_i \right\} \\ &= \mathbb{E} \left[ \sum_{i=0}^I \sum_{s=0}^N \mathbb{1}(s = s_i) \times \log\{q_{s, N}(\mathbf{Z}_i; \beta)\} \mid q_{y,N}^{(k)}, y_i, N_i \right] \\ &= \sum_{i=0}^I \sum_{s=0}^N \underbrace{\mathbb{E} \{ \mathbb{1}(s = s_i) \mid q_{y,N}^{(k)}, \beta^{(k)}, y_i, N_i \}}_{p_{N_i y_i s_i}^{(k)}} \times \log\{q_{s, N}(\mathbf{Z}_i; \beta)\},\end{aligned}\quad (5)$$

where  $p_{N' y' y} = \mathbf{Pr}_N(Y = y \mid Y' = y', N')$  denotes the conditional probability of achieving  $y$  events in the complete cluster of size  $N$ , given that  $y'$  events have been observed in the original cluster of size  $N'$ . The expression for  $p_{N' y' y}$  can be derived from  $q_{y,N}$  using Bayes theorem and the MCAR interpretation as follows:

$$\begin{aligned}p_{N' y' y} &= \mathbf{Pr}_N(Y = y \mid Y' = y', N') \\ &= \frac{\mathbf{Pr}_{N'}(Y' = y' \mid Y = y) \mathbf{Pr}_N(Y = y)}{\sum_{t=0}^N \mathbf{Pr}_{y'}(Y' = y' \mid Y = t) \mathbf{Pr}_N(Y = t)} \\ &= \frac{\binom{y}{y'} \binom{N-y}{N'-y'} q_{y,N}}{\sum_{t=y'}^{N-N'+y'} \binom{t}{y'} \binom{N-t}{N'-y'} q_{t,N}},\end{aligned}\quad (6)$$

and  $p_{N_i y_i s_i}^{(k)}$  at the  $k$ th step of the iteration can be obtained by using  $q_{y,N}^{(k)}$  on the right-hand side of the expression.

3.0.4. *Maximization step.* Equation (5) can be interpreted as the log-likelihood for the SPGLM with fixed cluster size  $N$  for an expanded data set: each observation with cluster size  $N_i$  and number of responses  $y_i$  is expanded to  $(N + 1)$  replicates with  $0, 1, \dots, N$ , responses, respectively. The probabilities  $p_{N_i y_i s_i}^{(k)}$  serve as cluster weights in the expanded data set. From here we can use a slight generalization of the Newton-Raphson algorithm developed by Wurm and Rathouz (2018) to estimate the baseline density distribution  $\mathbf{q}_N^{(0)}$  and the regression coefficients  $\beta$ .

$\langle \text{Fit model 4} \rangle \equiv$

```

while (iter < control$maxit & difference > control$eps) {
  iter <- iter + 1
  referencef0Pre <- referencef0
  tiltingWsPre <- tiltingWs
  betasPre <- betas
  # passing parameter updates to the next iteration
  if (iter == 1) {
    startUpdates <- list(iter = iter,
                        f0StartValue = NULL,
                        thetaStartValue = NULL,
                        betaStartValue = NULL)
  } else {
    startUpdates <- list(iter = iter,
                        f0StartValue = referencef0Pre,
                        thetaStartValue = tiltingWsPre,
                        betaStartValue = betasPre)
  }
  # EM NR algorithm
  res <- glDrMaximize(DesignMatrix.CONST, CBData, referencef0Pre, tiltingWsPre, startUpdates)
  referencef0 <- res$density.ref
  # notice that this tilting parameter estimates are w.r.t the intermediate dataset, original dataset e
  tiltingWs <- res$tiltParam[(0:(nrow(CBData)-1))*length(referencef0)+1]
  betas <- res$regressionEst
  # monitor parameter estimates for baseline distribution, tilting parameters, regression coefficients
  difference <- sum(abs(referencef0Pre - referencef0)) + sum(abs(betasPre - betas))
}

list(referencef0 = referencef0,
      tiltingWs = tiltingWs,
      betas = betas,
      EMglDrFitRes = res)

```

◇  
Fragment referenced in 1.