Propensity score with distributed data through distributed logistic regression

Estimates and confidence intervals for the parameters of the logistic regression model $P(Y = 1) = \exp(X\beta) / (1 + \exp(X\beta))$ are obtained by making use of the Newton-Raphson algorithm using local gradients and Hessian matrices. This allows for the recreation of the estimates and confidence intervals from the centralised setting. For the procedure, all nodes must have the same number of predictors. The total number of observations (individuals) is N and the number of predictors is p.

After an initial iteration $t = 0$ consisting of averaging the local estimates of each node at the coordination center, the aggregate estimate is updated using the gradients and Hessian matrices of the data nodes. These new estimates allow each node to compute the propensity score of their observations. The process may be repeated until convergence (which yields the propensity scores).

Assuming K nodes and a coordinating node are involved; each node k identifies the local (binary) treatment vector $Y^{(k)}$, the local predictor matrix $X^{(k)}$ and the optional local weight vector $W^{(k)}$. If no weights are provided, a vector of 1s will be used instead, as this represents uniform weights across observations. A column with the quantity 1 for each observation is added in the local predictor matrix to account for the intercept estimate.

Example.
Suppose the following dataset at node k, with 3 observations ($N^{(k)} = 3$), 2 predictors ($p = 2$) and a vector of weights.

treatment	gestational_age	age_admission
0	42	56
0	35	17
1	37	25

weights
10
5
10

The local treatment vector is $Y^{(k)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, the local predictor matrix is $X^{(k)} = \begin{bmatrix} 1 & 42 & 56 \\ 1 & 38 & 43 \\ 1 & 37 & 25 \end{bmatrix}$ and the local weight vector is $W^{(k)} = \begin{bmatrix} 10 \\ 5 \\ 10 \end{bmatrix}$.

Data node (initial phase: iteration $t = 0$)

1. Each node computes its local estimate $\hat{\beta}^{(k)}$ using the standard procedure of maximizing its local likelihood function (e.g. using R's glm function). This estimate is saved as a *csv* file and sent to the coordination node, along with the local sample size.

Example (continued).

The following quantites **<u>are shared</u>** to the coordinating node:

$\hat{\beta}^{(k)} = \begin{bmatrix} -0.2796 \\ 0.385 \\ -0.3195 \end{bmatrix}$

$N^{(k)} = 3$

The exported *csv* will share the following table from node k:

coefs,	n
$-0.2796,$	3
$0.385,$	NA
$-0.3195,$	NA

<u>Coordinating node (initial phase, iteration $t = 0$)</u>

2. The coordinating node averages the local estimates to generate the simple averaging estimate: $\hat{\beta}_{NR,t=0} = \hat{\beta}^{\mathrm{SA}} = \left(\sum_{k=1}^{K} N^{(k)}\, \hat{\beta}^{(k)}\right) / \sum_{k=1}^{K} N^{(k)}$, then sends that quantity back to the data nodes.

<u>Data node (iteration phase, $t = 1, \ldots T$)</u>

3. Each data node computes its local gradient and Hessian evaluated at the latest β estimate, $\hat{\beta}_{NR,t-1}$:

- $D_{NR,t}^{(k)} = {X^{(k)}}^{\mathrm{T}} \operatorname{diag}\left(W^{(k)}\right)\left\{Y^{(k)} - s^{(k)}\left(\hat{\beta}_{NR,t-1}\right)\right\},$
- $V_{NR,t}^{(k)} = {X^{(k)}}^{\mathrm{T}} \operatorname{diag}\left(W^{(k)}\right) P^{(k)}(\hat{\beta}_{NR,t-1}) X^{(k)}.$

Where

- $s^{(k)}(\beta) = \left(\frac{exp(\beta^{\mathrm{T}} x_1^{(k)})}{1+exp(\beta^{\mathrm{T}} x_1^{(k)})}, \ldots, \frac{exp(\beta^{\mathrm{T}} x_{N^{(k)}}^{(k)})}{exp(\beta^{\mathrm{T}} x_{N^{(k)}}^{(k)})}\right)^{\mathrm{T}},$
- $P^{(k)}(\beta) = \operatorname{diag}\left(\left\{\frac{exp(\beta^{\mathrm{T}} x_i^{(k)})}{1+exp(\beta^{\mathrm{T}} x_i^{(k)})}\right\}\left\{1 - \frac{exp(\beta^{\mathrm{T}} x_i^{(k)})}{1+exp(\beta^{\mathrm{T}} x_i^{(k)})}\right\}\right).$

The quantity $s^{(k)}\left(\hat{\beta}_{NR,t-1}\right)$ represents the treatment predictions using the current β estimate and and $P^{(k)}\left(\hat{\beta}_{NR,t-1}\right)$ is a diagonal matrix whose entries can be interpreted as treatment predictions variances.

The quantities $D_{NR,t}^{(k)}$ and $V_{NR,t}^{(k)}$ are sent to the coordinating center.

Example (continued).

Suppose the coordinating node shared $\hat{\beta}_{NR,t=0} = \hat{\beta}^{\text{SA}} = \begin{bmatrix} -2 \\ 1 \\ -1 \end{bmatrix}$.

The following quantities **<u>are shared</u>** to the coordinating node (iteration $t = 1$):

$D_{NR,t=1}^{(k)} = \begin{bmatrix} -0.1192 \\ -4.5297 \\ -5.1257 \end{bmatrix}$, $V_{NR,t}^{(k)} = \begin{bmatrix} 0.1050 & 3.9898 & 4.5147 \\ 3.9898 & 151.6107 & 171.5595 \\ 4.5147 & 171.5595 & 194.1331 \end{bmatrix}$.

The exported *csv* will share the following table from node k:

gradient,	hessian_intercept	hessian_pred1	hessian_pred2
−0.1192,	0.1050,	3.9898,	4.5147
−4.5297,	3.9898,	151.6107,	171.5595
−5.1257,	4.5147,	171.5595,	194.1331

<u>Coordinating node (iteration phase, $t = 1, \ldots T - 1$)</u>

4. The coordinating center computes the global gradient and Hessian

- $\bar{D}_{NR,t} = \sum_{k=1}^{K} D_{NR,t}^{(k)}$,
- $\bar{V}_{NR,t} = \sum_{k=1}^{K} V_{NR,t}^{(k)}$.

It then computes the Newton-Raphson iteration $\hat{\beta}_{NR,t} = \hat{\beta}_{NR,t-1} + \left(\bar{V}_{NR,t}\right)^{-1} \bar{D}_{NR,t}$ and sends the new estimate to the data nodes.

Steps **3.** and **4.** are repeated until convergence or a fixed number of times (T).

<u>Coordinating node (last iteration T)</u>

5. The coordinating center computes the global gradient and Hessian

- $\bar{D}_{NR,T} = \sum_{k=1}^{K} D_{NR,T}^{(k)}$,
- $\bar{V}_{NR,T} = \sum_{k=1}^{K} V_{NR,T}^{(k)}$.

It then computes the Newton-Raphson iteration $\hat{\beta}_{NR,T} = \hat{\beta}_{NR,T-1} + \left(\bar{V}_{NR,T}\right)^{-1} \bar{D}_{NR,T}$.

Each node then computes their propensity scores, which is the treatment predictions using the last β estimate: $s^{(k)}\left(\hat{\beta}_{NR,T}\right)$. The inverse probability weights $IPW^{(k)}$ are then obtained from the propensity scores.

- $IPW_i^{(k)} = \frac{Y_i^{(k)}}{s_i^{(k)}\left(\hat{\beta}_{NR,T}\right)} + \frac{1 - Y_i^{(k)}}{1 - s_i^{(k)}\left(\hat{\beta}_{NR,T}\right)}$,

where subindex i denotes the i^{th} observation in the node's dataset.

Example (continued).

Suppose the coordinating node shared $\hat{\beta}_{NR,t=T} = \begin{bmatrix} -0.2 \\ 0.4 \\ -0.5 \end{bmatrix}$. Also recall that $Y^{(k)} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

The propensity scores for this node are:

$$s^{(k)}\left(\hat{\beta}_{NR,t=T}\right) = \left(\frac{\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_1^{(k)}\right)}{1+\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_1^{(k)}\right)}, \frac{\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_2^{(k)}\right)}{1+\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_2^{(k)}\right)}, \frac{\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_3^{(k)}\right)}{1+\exp\left(\hat{\beta}_{NR,t=T}{}^{\mathrm{T}} x_3^{(k)}\right)} \right)^{\mathrm{T}}$$
$$= (0.00001, 0.99503, 0.89090)^{\mathrm{T}}.$$

The inverted probability weights for this node are:

$$IPW^{(k)} = (1.00001, 201.33681, 1.12246)^{\mathrm{T}}.$$

Note: optional parameter *threshold* L, which may take values in $[0, 0.5]$, will make sure that the propensity scores are within the boundaries $[L, 1-L]$. Setting $L = 0$ will not change any of the propensity scores.

Example (continued).

Suppose that all node agreed on using a threshold of $L = 0.05$.

The propensity scores for this node will then be truncated to:

$$(0.05, 0.95, 0.89090)^{\mathrm{T}}.$$

This yields new inverted probability weights for this node, which are:

$$IPW^{(k)} = (1.05263, 20, 1.12246)^{\mathrm{T}}.$$