Linear regression with distributed data

Estimates and confidence intervals for the parameters of the linear regression model ($Y = X\beta + \varepsilon$) are obtained by sharing summary statistics that allow to exactly recreate 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.

Assuming K nodes and a coordinating node are involved; each node k identifies the local outcome 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 ($N^{(k)} = 3$) and 2 predictors ($p = 2$) and a vector of weights.

newborn_birth_weight	gestational_age	age_admission
4314.84	42	56
3337.88	38	43
3020.90	37	25

weights
10
5
10

The local outcome vector is $Y^{(k)} = \begin{bmatrix} 4314.84 \\ 3337.88 \\ 3020.90 \end{bmatrix}$ and the local predictor matrix $X^{(k)} = \begin{bmatrix} 1 & 42 & 56 \\ 1 & 38 & 43 \\ 1 & 37 & 25 \end{bmatrix}$ and the local weight vector is $W^{(k)} = \begin{bmatrix} 10 \\ 5 \\ 10 \end{bmatrix}$.

Data node

1. Each node computes the three following quantities: ${X^{(k)}}^t \mathrm{diag}(W^{(k)}) X^{(k)}$ (matrix), ${Y^{(k)}}^t \mathrm{diag}(W^{(k)}) Y^{(k)}$ (constant), and ${X^{(k)}}^t \mathrm{diag}(W^{(k)}) Y^{(k)}$ (vector). These quantities and each node's sample size, n_k, are shared to the coordinating node. Each quantity is assigned a column in the exported *csv* files.

Example (continued).

- $X^{(k)^t}\text{diag}(W^{(k)})X^{(k)} = \begin{bmatrix}1 & 1 & 1\\ 42 & 38 & 37\\ 56 & 43 & 25\end{bmatrix}\begin{bmatrix}10 & 0 & 0\\ 0 & 5 & 0\\ 0 & 0 & 10\end{bmatrix}\begin{bmatrix}1 & 42 & 56\\ 1 & 38 & 43\\ 1 & 37 & 25\end{bmatrix}$
$= \begin{bmatrix}25 & 980 & 1\,025\\ 980 & 38\,550 & 40\,940\\ 1\,025 & 40\,940 & 46\,855\end{bmatrix}$

- $Y^{(k)^t}\text{diag}(W^{(k)})Y^{(k)} = [4314.84 \quad 3337.88 \quad 3020.90]\begin{bmatrix}10 & 0 & 0\\ 0 & 5 & 0\\ 0 & 0 & 10\end{bmatrix}\begin{bmatrix}4314.84\\ 3337.88\\ 3020.90\end{bmatrix}$
$= 10 \cdot 4314.84^2 + 5 \cdot 3337.88^2 + 10 \cdot 3020.90^2 = 333\,144\,025$

- $X^{(k)^t}\text{diag}(W^{(k)})Y^{(k)} = \begin{bmatrix}1 & 1 & 1\\ 42 & 38 & 37\\ 56 & 43 & 25\end{bmatrix}\begin{bmatrix}10 & 0 & 0\\ 0 & 5 & 0\\ 0 & 0 & 10\end{bmatrix}\begin{bmatrix}4314.84\\ 3337.88\\ 3020.90\end{bmatrix} = \begin{bmatrix}90\,046.80\\ 3\,564\,163\\ 3\,889\,179.60\end{bmatrix}$

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

$X^{(k)^t}\text{diag}(W^{(k)})X^{(k)} = \begin{bmatrix}25 & 980 & 1\,025\\ 980 & 38\,550 & 40\,940\\ 1\,025 & 40\,940 & 46\,855\end{bmatrix}$, $Y^{(k)^t}\text{diag}(W^{(k)})Y^{(k)} = 333\,144\,025$,
$X^{(k)^t}\text{diag}(W^{(k)})Y^{(k)} = \begin{bmatrix}90\,046.80\\ 3\,564\,163\\ 3\,889\,179.60\end{bmatrix}$ and $n_k = 3$.

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

x^TWx,	y^TWy,	x^TWy	n
25	333 144 025,	90 046.80	3
980,	,	3 564 163	
1 025,	,	3 889 179.60	
980,	,		
38 550,	,		
40 940	,		
1 025	,		
40 940,	,		
46 855	,		

Coordinating node

2. The coordinating node sums the quantities over all nodes:

$$X^t \text{diag}(W) X = \sum_{k=1}^{K} {X^{(k)}}^t \text{diag}\left(W^{(k)}\right) X^{(k)},$$

$$X^t \text{diag}(W) Y = \sum_{k=1}^{K} {X^{(k)}}^t \text{diag}\left(W^{(k)}\right) Y^{(k)},$$

$$Y^t \text{diag}(W) Y = \sum_{k=1}^{K} {Y^{(k)}}^t \text{diag}\left(W^{(k)}\right) Y^{(k)} \text{ and } n = \sum_{k=1}^{K} n_k.$$

3. The parameter estimates are then calculated at the coordinating node, with the following formula (usual Ordinary Least Squares estimator for linear regression when no weights are provided):

$$\hat{\beta} = (X^t \text{diag}(W) X)^{-1} X^t \text{diag}(W) Y$$

$$\text{where } \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^t$$

Lower and upper bounds for the confidence intervals of the model parameters are also calculated at the coordinating node:

$$CI(\beta_j) = [\hat{\beta}_j \pm t_{\frac{\alpha}{2}, n-p-1} \sqrt{\hat{\sigma}^2 [(X^t \text{diag}(W) X)^{-1}]_{j+1, j+1}}]$$

$$\text{where } \hat{\sigma}^2 = \frac{1}{n-p-1} \left(Y^t \text{diag}(W) Y - \hat{\beta}^t X^t \text{diag}(W) Y\right).$$

The outputs of the procedure are the parameters estimates (including intercept), and the upper and lower bounds of the confidence intervals for the model parameters.