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 $Y^{(k)}$ and the local predictor matrix $X^{(k)}$. 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 ($p = 2$).

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

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

Data node

1. Each node computes the three following quantities: ${X^{(k)}}^t X^{(k)}$ (matrix), ${Y^{(k)}}^t Y^{(k)}$ (constant), and ${X^{(k)}}^t Y^{(k)}$ (vector). The quantities are shared to the coordinating node. Each quantity is assigned a column in the exported *csv* files.

Example (continued).

- $\boldsymbol{{X^{(k)}}^t X^{(k)}} = \begin{bmatrix} 1 & 1 & 1 \\ 42 & 38 & 37 \\ 56 & 43 & 25 \end{bmatrix} \begin{bmatrix} 1 & 42 & 56 \\ 1 & 38 & 43 \\ 1 & 37 & 25 \end{bmatrix} = \begin{bmatrix} 3 & 117 & 124 \\ 117 & 4577 & 4911 \\ 124 & 4911 & 5610 \end{bmatrix}$

- ${Y^{(k)}}^t Y^{(k)} = \begin{bmatrix} 4314.84 & 3337.88 & 3020.90 \end{bmatrix} \begin{bmatrix} 4314.84 \\ 3337.88 \\ 3020.90 \end{bmatrix}$
$= 4314.84^2 + 3337.88^2 + 3020.90^2 = 38\,885\,123.93$

- ${X^{(k)}}^t Y^{(k)} = \begin{bmatrix} 1 & 1 & 1 \\ 42 & 38 & 37 \\ 56 & 43 & 25 \end{bmatrix} \begin{bmatrix} 4314.84 \\ 3337.88 \\ 3020.90 \end{bmatrix} = \begin{bmatrix} 10\,673.62 \\ 419\,836.02 \\ 460\,682.38 \end{bmatrix}$

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

$$X^{(k)^t}X^{(k)} = \begin{bmatrix} 3 & 117 & 124 \\ 117 & 4577 & 4911 \\ 124 & 4911 & 5610 \end{bmatrix},\ Y^{(k)^t}Y^{(k)} = 38\ 885\ 123.93 \text{ and } X^{(k)^t}Y^{(k)} = \begin{bmatrix} 10\ 673.62 \\ 419\ 836.02 \\ 460\ 682.38 \end{bmatrix}.$$

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

$\mathbf{x^Tx}$,	$\mathbf{y^Ty}$,	$\mathbf{x^Ty}$
3,	38 885 123.93,	10 673.62
117,	,	419 836.02
124,	,	460 682.38
117,	,	
4577,	,	
4911,	,	
124,	,	
4911,	,	
5610,	,	

<u>Coordinating node</u>

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

$$X^tX = \sum_{k=1}^{K} X^{(k)^t}X^{(k)},\ X^tY = \sum_{k=1}^{K} X^{(k)^t}Y^{(k)} \text{ and } Y^tY = \sum_{k=1}^{K} Y^{(k)^t}Y^{(k)}.$$

3. The parameter estimates are then calculated at the coordinating node, with the following formula (usual Ordinary Least Squares estimator for linear regression):

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

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

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

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

$$\text{where } \hat{\sigma}^2 = \frac{1}{N-p-1}\left(Y^tY - \hat{\beta}^tX^tY\right).$$

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.