**R Code Manual: Model Estimation**

**Manufacturer Sampling Model:**

Steps to execute the code:

* Open Rstudio and open the files ‘loglikelihood\_sampling.R’ and ‘mcmc\_sampling.R’
* Put the features file, called ‘manu\_features.csv’ in the working directory (to check the working directory, type getwd() in R)
* Click ‘source’ in the R console, or type in: source(‘~/mcmc\_sampling.R’)
* When prompted, type in the column numbers from ‘manu\_features.csv’ to choose the features to include in the model, separated by a comma (e.g. 4,6,7,12,13,15,16,21,22,23,24,25,26,27)
* Type in the column number of the sampling fraction
* Type in the column number of the refusal fraction (choose intentional or unintentional)
* Type in the column number of the number of shipments
* Type in the column number of the number of sampled shipments
* Type in the ‘burnin’ parameter value, this is the number of pre-iteration of the model (e.g. 1000)
* Type in the ‘nsamples’ parameter value, this is the number of iteration of the model (e.g. 1000)

Steps to interpret the results:

* In order to find significant features (at 95% confidence level), type in q\_significant\_features for sampling, and p\_significant\_features for refusals
* In order to test out-of-sample prediction, we use Spearman Correlation coefficient that compares the predicted rank and the actual rank of sampling fractions and risk scores. To find the value of the correlation coefficient, type in cor\_q for sampling and cor\_p for refusals. In order to test whether the correlation is significantly greater than 0, type in cor\_test\_q for sampling and cor\_test\_p for refusals
* In order to visualize the distribution of predicted scores, type in plot(sort(q\_mean)) for sampling scores, and plot(sort(p\_mean)) for risk scores
* The posterior distributions of all model parameters are saved in 'posterior samples.csv' in the working directory
* The predicted scores are saved, in addition to the features, in the file 'inspection\_features\_scores.csv' in the working directory

Remark: In order to validate the model using ROC curves, we need the response variables (sampling fraction and refusal fraction) to be binary. Since this is not the case, we cannot use ROC.

**Consignee/Importer Inspection Model:**

Steps to execute the code:

* Open Rstudio and open the files ‘loglikelihood\_inspection.R’ and ‘mcmc\_inspection.R’
* Put the features file, called ‘consignee\_features.csv’ or ‘importer\_features.csv’ in the working directory (to check the working directory, type getwd() in R)
* Click ‘source’ in the R console, or type in: source(‘~/mcmc\_inspection.R’)
* When prompted, type in the column numbers from the features file to choose the features to include in the model, separated by a comma (e.g. 3,5,8,9,10,11,12,13,14,15,16)
* Type in the column number of the inspection status
* Type in the column number of the inspection outcome
* Type in the ‘burnin’ parameter value, this is the number of pre-iteration of the model (e.g. 1000)
* Type in the ‘nsamples’ parameter value, this is the number of iteration of the model (e.g. 1000)

Steps to interpret the results:

* In order to find significant features (at 95% confidence level), type in q\_significant\_features for sampling, and p\_significant\_features for refusals
* In order to test out-of-sample prediction, we use ROC analysis. Type q\_AUC for the inspection AUC and p\_AUC for the inspection outcome AUC
* In order to visualize the distribution of predicted scores, type in plot(sort(q\_mean)) for sampling scores, and plot(sort(p\_mean)) for risk scores
* The posterior distributions of all model parameters are saved in 'posterior samples.csv' in the working directory
* The predicted scores are saved, in addition to the features, in the file 'inspection\_features\_scores.csv' in the working directory