Competencia

Mateo, Miguel, Steward y Camilo

# General methodology

The idea is to use Bayesian model selection using the bic approximation and a lasso regresion to select te important variables, we will try this on 100 different partitions of the data.

After that we will select only the variables that were included on at least the 50% of the partitions. We alse make sure that the distribution of the estmated give us a good estimation interval (not crossing 0).

When we select the variables we estimate a glm with those variables again on 100 different partitions so that we can have an estimation of the global and the specific metric and defined if the proposed model were better than the proposed by other team members.

### Count

For the case the the count case the decision was pretty easy, using BMA we got excelent results in the internal competition. The distribution of the coeefficients were also a good indicator. At the end the fit a poisson model and the prediction was made using the round method (At the begging we tried the floor method as it is the mode of the distribution but the results were not as good as the selected option)

### Continous

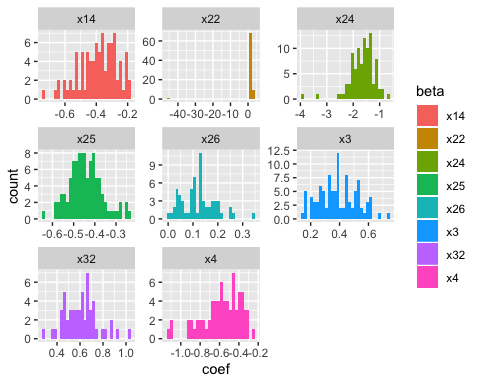
Here we did not include the neither variable x27 nor the x32 even though the were selected using the lasso methodology in more than the 50% of the partitions.

* x32 the distribution crossed 0 too many times.
* x27 and x24 at the end the variables were not significant in the final glm model.

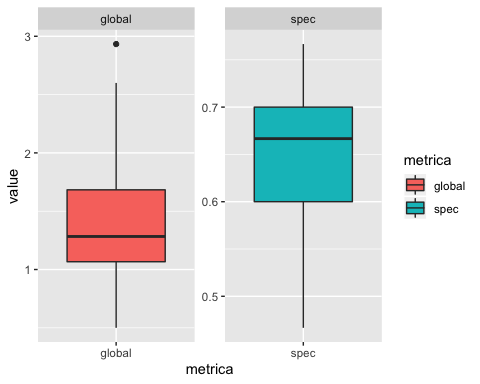
For the final model we performed a ridge regression since we had multicolinearity problems in the selected variables.

# Count

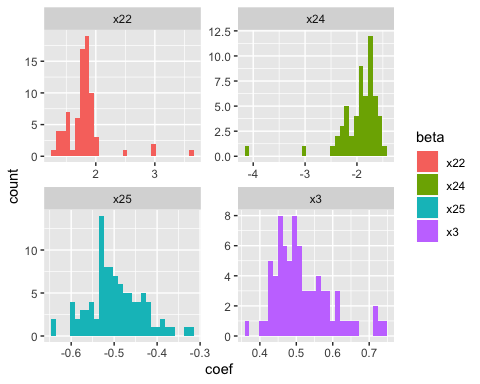
## Using Lasso



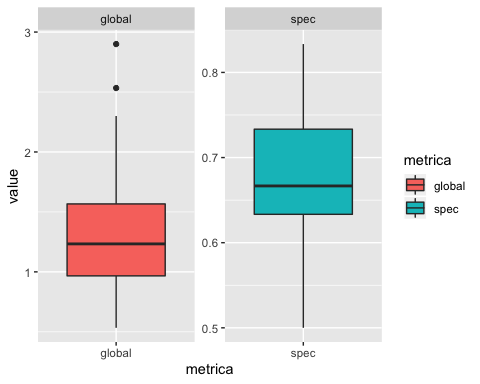
##   
## Call:  
## glm(formula = y ~ ., family = "poisson", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.3275 -0.9845 -0.1807 0.6515 1.7319   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.32376 1.27220 1.827 0.0678 .   
## x14 -0.37633 0.22662 -1.661 0.0968 .   
## x22 1.48403 0.59184 2.507 0.0122 \*   
## x24 -1.69032 0.86096 -1.963 0.0496 \*   
## x25 -0.46308 0.11400 -4.062 4.86e-05 \*\*\*  
## x26 0.09472 0.12990 0.729 0.4659   
## x3 0.31554 0.17115 1.844 0.0652 .   
## x32 0.51057 0.25538 1.999 0.0456 \*   
## x4 -0.44319 0.31389 -1.412 0.1580   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 218.76 on 149 degrees of freedom  
## Residual deviance: 141.62 on 141 degrees of freedom  
## AIC: 407.02  
##   
## Number of Fisher Scoring iterations: 5



## BMA

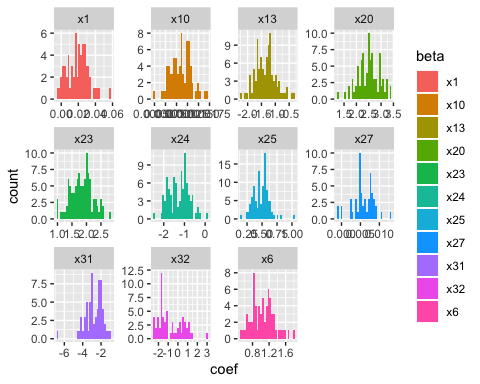


##   
## Call:  
## glm(formula = y ~ ., family = "poisson", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5365 -1.2022 -0.1728 0.6710 2.0295   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.18071 0.54502 5.836 5.35e-09 \*\*\*  
## x22 1.81030 0.58125 3.114 0.00184 \*\*   
## x24 -1.80062 0.82324 -2.187 0.02873 \*   
## x25 -0.50045 0.08696 -5.755 8.67e-09 \*\*\*  
## x3 0.48922 0.15003 3.261 0.00111 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 218.76 on 149 degrees of freedom  
## Residual deviance: 149.99 on 145 degrees of freedom  
## AIC: 407.39  
##   
## Number of Fisher Scoring iterations: 5

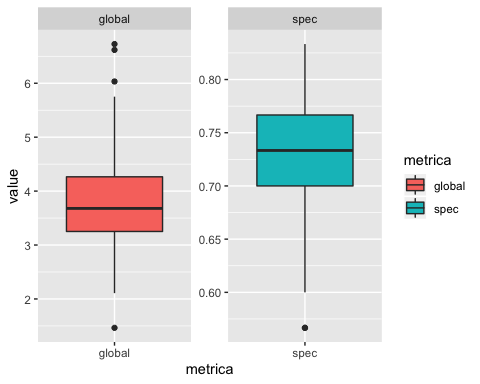


# Continua

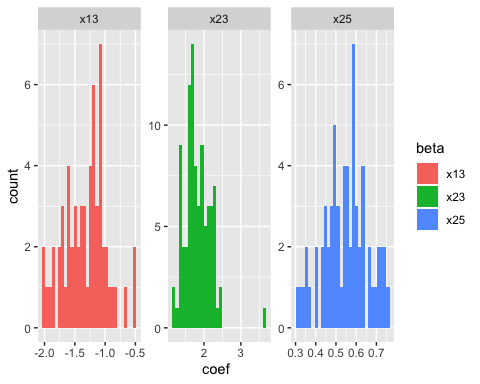
## Lasso



##   
## Call:  
## glm(formula = y ~ ., family = gaussian(), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.4759 -1.2708 -0.0306 1.1462 4.7429   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.3470952 1.5024426 -2.893 0.00442 \*\*   
## x1 0.0281269 0.0165619 1.698 0.09166 .   
## x6 0.9724902 0.5260840 1.849 0.06662 .   
## x10 0.0012507 0.0004084 3.062 0.00263 \*\*   
## x13 -1.3905187 0.4239723 -3.280 0.00131 \*\*   
## x20 2.5456999 0.8396928 3.032 0.00289 \*\*   
## x23 2.0828256 0.4784330 4.353 2.56e-05 \*\*\*  
## x25 0.4968935 0.1775517 2.799 0.00585 \*\*   
## x31 -2.5263406 0.8820860 -2.864 0.00482 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 3.326978)  
##   
## Null deviance: 694.72 on 149 degrees of freedom  
## Residual deviance: 469.10 on 141 degrees of freedom  
## AIC: 616.71  
##   
## Number of Fisher Scoring iterations: 2



## BMA



##   
## Call:  
## glm(formula = y ~ ., family = gaussian(), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.3191 -1.1589 0.0797 1.3678 4.8404   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.1869 1.1959 -1.829 0.069478 .   
## x13 -0.4990 0.4019 -1.241 0.216428   
## x23 1.7693 0.4806 3.681 0.000326 \*\*\*  
## x25 0.3705 0.1830 2.025 0.044734 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 4.071868)  
##   
## Null deviance: 694.72 on 149 degrees of freedom  
## Residual deviance: 594.49 on 146 degrees of freedom  
## AIC: 642.24  
##   
## Number of Fisher Scoring iterations: 2

