Sensitivity and Specificity

Derek Sonderegger

2019-02-20

# library(devtools)   
# install\_github('dereksonderegger/BurkPx')   
library(BurkPx)  
library(ggplot2)

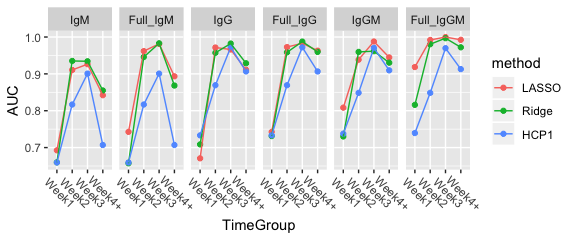
# Human ROC analysis on Human models

We first split the patients into test/training sets and then fit all the various models (IgG, IgM, IgGM, etc) using only patients from the training set. Because we have 100 Meliod patients, then 50 of those patients get assigned to the test group and 50 to the training group. Likewise of the 400 controls, 200 get assigned to the test set and 200 to the training.

Once the patients have been assigned to either the test or training set, all of the patients serologies are included in the set. This means that a single patient with many serologies might have an oversized effect. But we did this to try to keep our sample sizes as high as possible.

## Model Selection

To assess how well the models work, we will look at the Area Under the Curve (AUC) for the Reciever-Operator Curve (ROC). To generate this by week, We take the human data and split it into Healthy, Week 1, Week 2, etc. For each of the Weeks, we calculate probability of infection for the the Healthy and Infected groups using the six different human models.

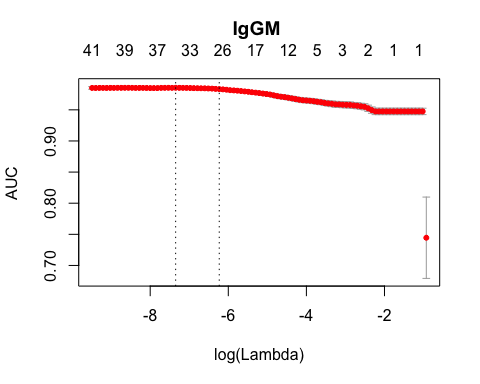


From this analysis, it is clear that for the first week, we need both the IgG and IgM sereologies. It also seems that the LASSO is working better than Ridge Regression and that IgG is working better than IgM. However the best performance is by the IgGM data which uses both IgG and IgM observation values. The “Full” graphs are where I used the model fit on both the test and training data to predict the test set.

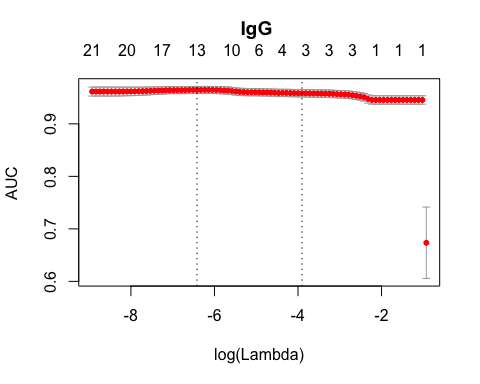
## How many covariates are used?

## Antigen Coef  
## 1 BPSS1498\_HCP1.B 2.762212e-05  
## 2 LPSA 2.741923e-04  
## 3 LPSB 1.046623e-04  
## Antigen Coef  
## 1 IgG\_BPSL1201\_IMPS 1.140828e-04  
## 2 IgG\_BPSL1404\_ClpX 3.717731e-04  
## 3 IgG\_BPSL2827\_DNAK -1.484080e-04  
## 4 IgG\_BPSS0135 1.482746e-04  
## 5 IgG\_BPSS0476\_GroS 2.811126e-05  
## 6 IgG\_BPSS0530 1.310555e-04  
## 7 IgG\_BPSS1498\_HCP1.B 4.732412e-04  
## 8 IgG\_BPSS1652 5.337745e-05  
## 9 IgG\_CPS 1.980401e-05  
## 10 IgG\_LPSA 3.982208e-04  
## 11 IgG\_LPSB 1.628721e-04  
## 12 IgM\_BPSL1743\_Arg 1.067014e-04  
## 13 IgM\_BPSL2522\_OmpA -2.603365e-03  
## 14 IgM\_BPSL2827\_DNAK 7.608634e-04  
## 15 IgM\_BPSL3222\_rpIL -6.812789e-04  
## 16 IgM\_BPSL3396\_AtpD -6.829135e-04  
## 17 IgM\_BPSS0135 -1.329298e-04  
## 18 IgM\_BPSS0476\_GroS -1.112135e-04  
## 19 IgM\_BPSS0530 2.396669e-03  
## 20 IgM\_BPSS1652 1.202640e-03  
## 21 IgM\_BPSS1769\_NADH 8.247896e-04  
## 22 IgM\_BPSS1850 -3.822299e-04  
## 23 IgM\_CPS 5.421209e-05  
## 24 IgM\_LPSA 1.156942e-05  
## 25 IgM\_LPSB -7.014006e-04  
## 26 IgM\_MSHR5855.WCL -6.903194e-04  
## # A tibble: 18 x 2  
## # Groups: Antigen [18]  
## Antigen n  
## <chr> <int>  
## 1 BPSL1201\_IMPS 1  
## 2 BPSL1404\_ClpX 1  
## 3 BPSL1743\_Arg 1  
## 4 BPSL2522\_OmpA 1  
## 5 BPSL2827\_DNAK 2  
## 6 BPSL3222\_rpIL 1  
## 7 BPSL3396\_AtpD 1  
## 8 BPSS0135 2  
## 9 BPSS0476\_GroS 2  
## 10 BPSS0530 2  
## 11 BPSS1498\_HCP1.B 1  
## 12 BPSS1652 2  
## 13 BPSS1769\_NADH 1  
## 14 BPSS1850 1  
## 15 CPS 2  
## 16 LPSA 2  
## 17 LPSB 2  
## 18 MSHR5855.WCL 1

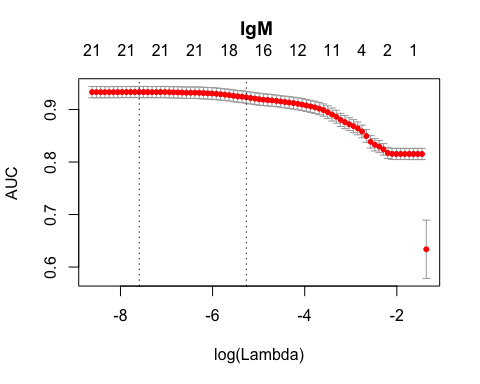
## What happens to the IgGM model as we decrease the number of covariates.



## What happens to the IgG model as we decrease the number of covariates.



## What happens to the IgM model as we decrease the number of covariates.



I am a little concerned with how large the IgGM model is, but considering our sample sizes..

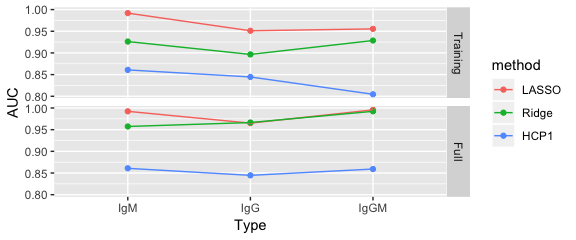
## # A tibble: 2 x 3  
## Status n perc  
## <fct> <int> <dbl>  
## 1 Negative 399 0.808  
## 2 Melioid 95 0.192

maybe this isn’t a big deal. The data consists of nearly 500 subjects split 80% / 20% between Healthy and Melioid.

# Nonhuman Primate

The nonhuman primate data was collected from both the batelle and Tulane information. When I created the testing and training sets, we split the Batelle individual subjects equally into the test and training sets and likewise split the Tulane negative individuals equally into the test and training sets.

## # A tibble: 6 x 4  
## # Groups: Origin, Status [3]  
## Origin Status n Type   
## <chr> <fct> <int> <chr>   
## 1 Battele Negative 6 Training  
## 2 Battele Melioid 12 Training  
## 3 Tulane Negative 24 Training  
## 4 Battele Negative 6 Test   
## 5 Battele Melioid 12 Test   
## 6 Tulane Negative 23 Test



For the Non-human primates, we see that the models that were trained on only the training data still did quite well in predicting the testing set with AUC values in the 90s. However the models trained with both the test and training sets (aka “Full” data), did a better job predicting the testing set, as we would expect. Notably, there isn’t a huge difference in prediction capabilities and, unlike the human data, we see that the IgM serum is better at prediction than the IgG.