Multi-cytokine signature of Trisomy 21

3/26/2020

Model selection procedure

1. Randomly select 80% of the data, within strata of T21 and Sex, to use as the training set.
2. Retain the remaining 20% of observations to serve as a validation set.

Model selection procedure

3. With T21 as the dependent variable, perform one logistic regression for each of 54 cytokines

$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{1}}) + \mathrm{Age} + \mathrm{Sex}$$
$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{2}}) + \mathrm{Age} + \mathrm{Sex}$$
$$\vdots$$
$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{54}}) + \mathrm{Age} + \mathrm{Sex}$$

4. Sort the cytokine p-values in ascending order

Model selection procedure

5. Fit a series of multi-cytokine logistic regression models, sequentially adding one cytokine at a time in order of p-value rank ('forward selection')

$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{1}}) + \mathrm{Age} + \mathrm{Sex}$$

$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{1}}) + \log_2(\mathrm{Concentration}_{\mathbf{2}}) + \mathrm{Age} + \mathrm{Sex}$$

$$\vdots$$

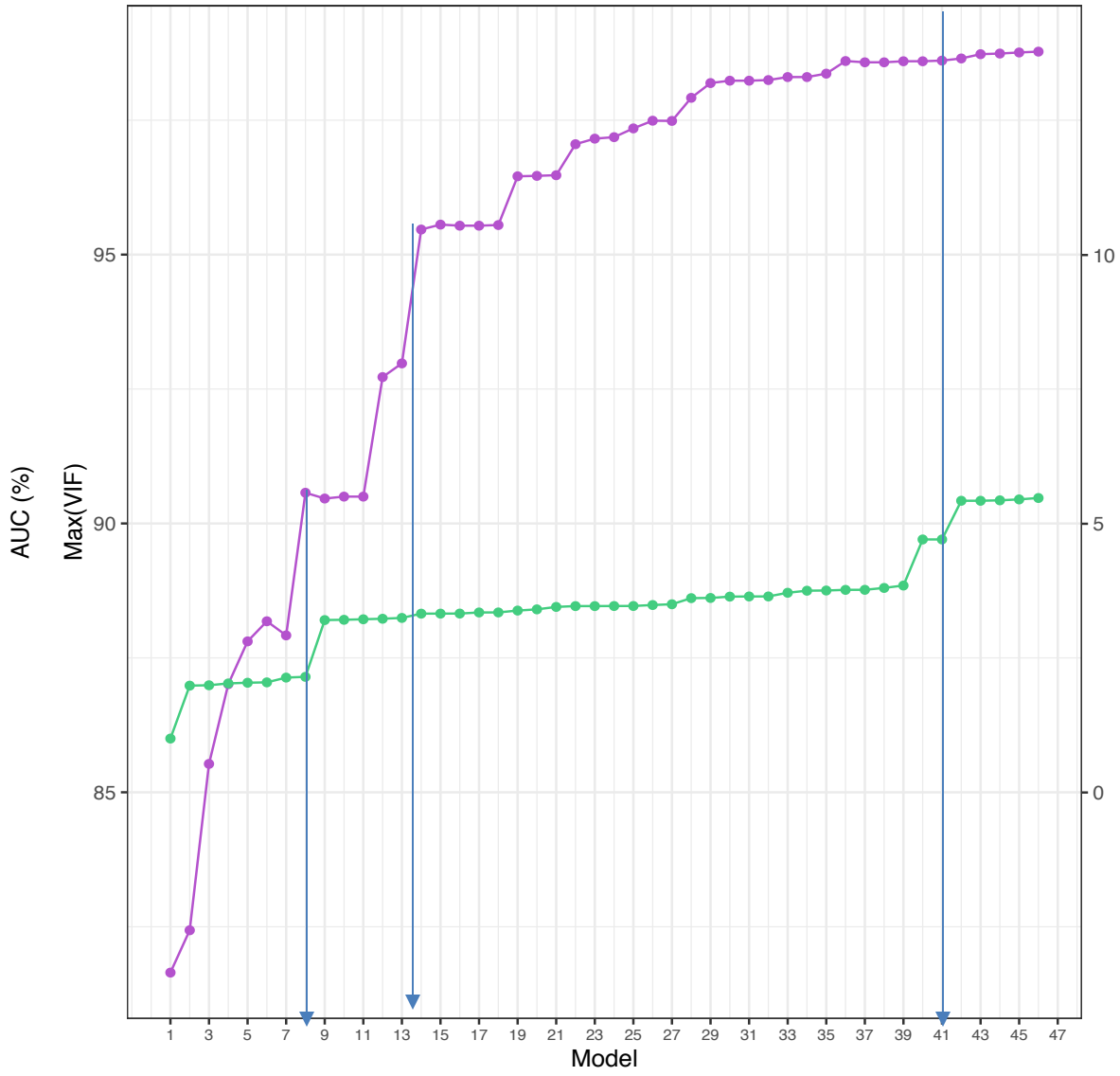
$$\mathrm{T21} \sim \log_2(\mathrm{Concentration}_{\mathbf{1}}) + \log_2(\mathrm{Concentration}_{\mathbf{2}}) + \log_2(\mathrm{Concentration}_{3}) + \cdots + \log_2(\mathrm{Concentration}_{\mathbf{54}}) + \mathrm{Age} + \mathrm{Sex}$$

Statistics for model selection

6. Output and plot statistics to evaluate model performance

Statistic	Purpose	Interpretation
Area under the ROC curve (AUC)	Quantify model's ability to discriminate between groups of the dependent variable.	Higher is better.
Variance inflation factor (VIF)	Indicate the presence of multi-collinearity (high correlation between predictors; over-fitting)	Lower is better. VIF>5 indicates possible multi-collinearity.
Residual deviance	Estimate of variability not explained by predictors	Lower is better.
Akaike Information Criteria (AIC)	Provide estimate of *relative* model quality, as compared to other models.	Lower is better. Prioritizes the minimization of false negatives.
Bayesian Information Criteria (AIC)	Provide estimate of *relative* model quality, as compared to other models.	Lower is better. Prioritizes parsimony.

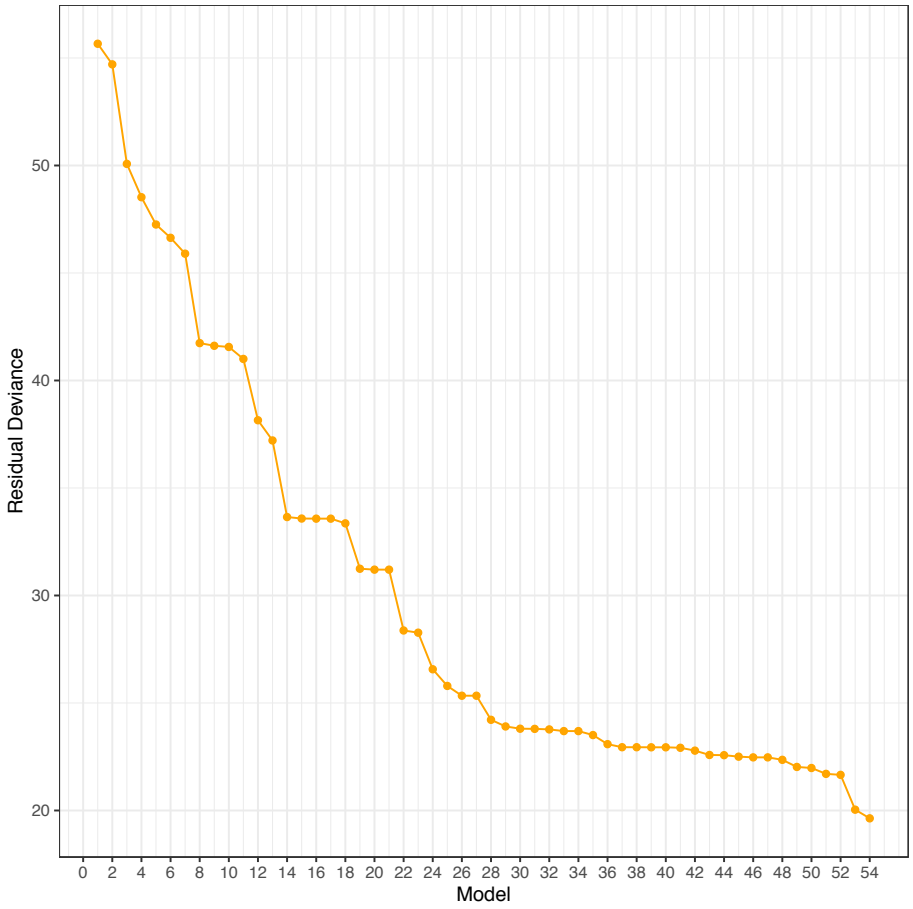
Model fit statistics in training data

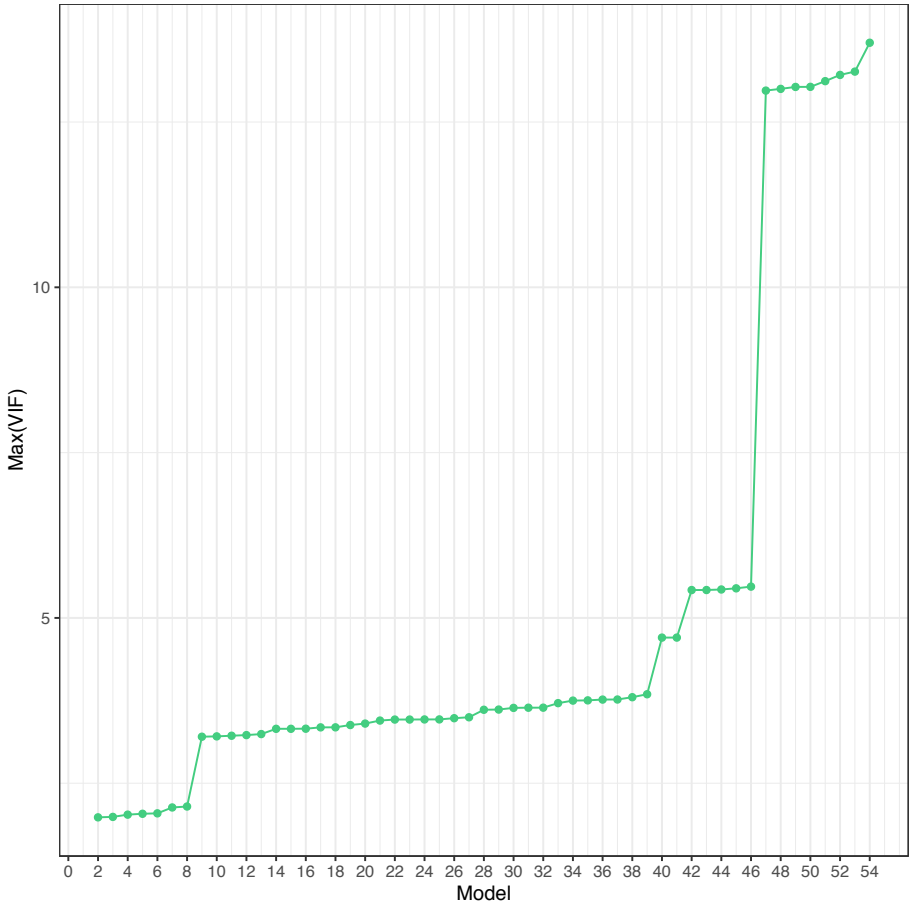


Top candidate models
(AUC, VIF, # cytokines)

Model 8 (>90%, <5, 8)
Model 14 (>95%, <5, 14)
Model 41 (>95%, <5, 41)

Model fit statistics in training data





Model fit statistics in training data