Viral Pneumonia Pilot Analysis

Dominic LaRoche

Friday, October 17, 2014

## Warning: package 'xlsx' was built under R version 3.1.3

## Warning: package 'rJava' was built under R version 3.1.3

## Warning: package 'xlsxjars' was built under R version 3.1.3

## Warning: package 'xtable' was built under R version 3.1.3

## Warning: package 'ggplot2' was built under R version 3.1.3

## Warning: package 'rtf' was built under R version 3.1.3

## Warning: package 'rms' was built under R version 3.1.3

## Warning: package 'Hmisc' was built under R version 3.1.3

## Warning: package 'Formula' was built under R version 3.1.3

## Warning: package 'gridExtra' was built under R version 3.1.3

## Warning: package 'SparseM' was built under R version 3.1.3

## Warning: package 'pander' was built under R version 3.1.3

## Warning: package 'car' was built under R version 3.1.3

# Introduction

This report summarizes a pilot analysis on the data provided on 11 October 2014. The data is summarized in table 1. I have initially taken two approaches to this problem: 1) regular LASSO, and 2) grouped LASSO. I describe the two methods in the corresponding sections below.

# Standard LASSO Regression

I first fit a model using standard LASSO regression with the "L1Norm" penalty. This method creates a separate indicator variable for every level of every factor. In this data set that creates 40 unique parameters to estimate (including the intercept). The LASSO penalty shrinks the coefficients individually so you may end up with a non-zero coefficient for a single factor level while the other levels have "no effect". The practical implications of this method are that you can then create a single variable in the final model using just the indicator for that particular factor level. E.g., if you had a factor that coded for shirt color with levels 'red', 'blue', and 'green', you may only get en effect for green shirts. In the final model you would only indicate whether the patient was wearing a green shirt or not. The downside of this method is that if the lone factor level does not make sense in the absence of the other levels then this doesn't work. The other thing to keep in mind is that the reference category (the category to which the effect of the level is compared) is very important in this method so careful attention must be paid to making a sensible reference category to which the other categories are all compared. I should also note that this only applies to factors with more than 2 levels, e.g. sex would not be affected.

## Warning: package 'lars' was built under R version 3.1.3

## Warning: package 'ROCR' was built under R version 3.1.3

## Warning: package 'gplots' was built under R version 3.1.3

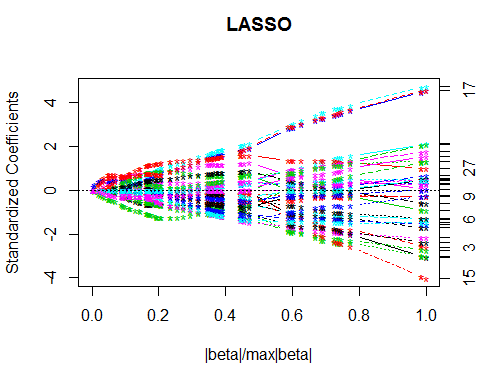


Figure 1. Plot of the regularization paths for the coefficients in the model. This shows how the number of non-zero coefficients decreased with an increasing penalty (right to left)

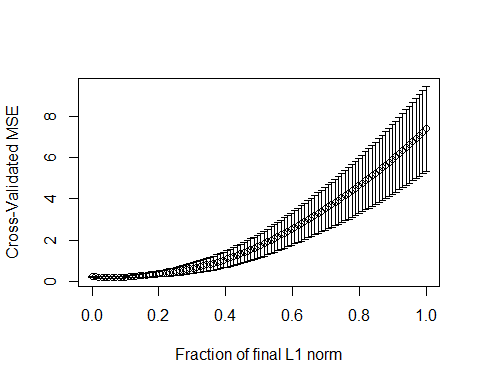


Figure 2. Plot of the mean-squared-prediciton error for each level of the penalty parameter. Standard errors are derived with 30 fold cross-validation (which might be excessive for such a small data-set and lead to overconfidencein the precision of the estimate).

# Grouped LASSO Regression

I also fit a grouped LASSO model for comparison to the regular LASSO. For this model I grouped all levels of a given factor together, but this step is flexible so if there is a more informative way to group variables we can incorporate that in the future. As you can see from figures 3 and 4 the grouped LASSO removed variables at a faster rate than regular LASSO.

## Warning: package 'grplasso' was built under R version 3.1.3

## Warning: package 'gglasso' was built under R version 3.1.3

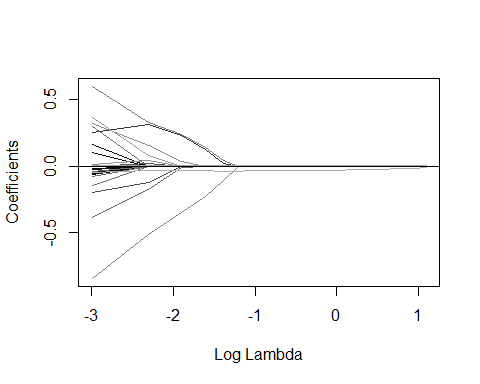


Figure 4. Regularization path for the grouped lasso model. The penalty removes variables faster because of the grouping among factors.

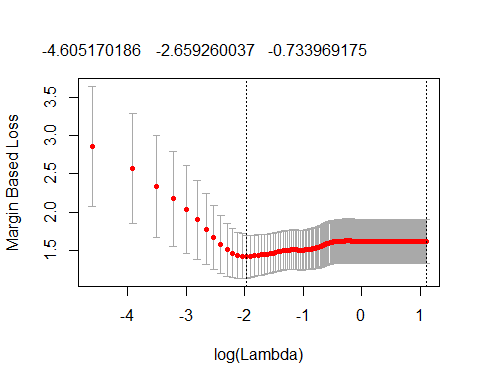


Figure 5. Cross-validated 0-1 loss (a form of prediction error) for each value of the penalty coefficient.

# Important Variables

|  |  |
| --- | --- |
|  | LASSO Coef |
| **(Intercept)** | 0 |
| **rvpADENOVIRUS C, INFLUENZA A 2009 H1N1** | 0 |
| **rvpHMPV** | 0.2311 |
| **rvpINFLUENZA A 2009 H1N1** | 0 |
| **rvpINFLUENZA A H3, INFLUENZA A** | 0 |
| **rvpINFLUENZA A, INFLUENZA A 2009 H1N1** | 0 |
| **rvpRHINOVIRUS** | 0.04161 |
| **rvpRHINOVIRUS, HMPV** | 0 |
| **rvpRSV-A** | 0 |
| **rvpRSV-B** | 0 |
| **other.infect.statYES** | 0.1714 |
| **sexM** | 0 |
| **prev.stayHOME** | 0 |
| **prev.stayHOSPITAL** | 0 |
| **prev.stayICU(ANOTHER INSTITUTION)** | 0 |
| **prev.staySNF** | 0 |
| **immun.comp.statYES** | 0 |
| **lung.dis.statTRUE** | 0 |
| **diabYES** | 0 |
| **curbf1** | 0 |
| **curbf2** | 0 |
| **curbf3** | 0 |
| **curbf4** | 0 |
| **vpdoseYES** | 0 |
| **akiYES** | 0 |
| **pl.effYES** | 0 |
| **dyspYES** | 0 |
| **wheezeYES** | -0.1019 |
| **admitTRANSFER** | 0 |
| **admitWARDS** | 0 |
| **pnaYES** | 0 |
| **mech.ventYES** | 0 |
| **ecmoYES** | 0.2052 |
| **age** | 0 |
| **BMI30YES** | 0 |
| **psi** | 0 |
| **hct** | -0.002343 |

Comparison of coefficients for the two LASSO methods. Variables with non-zero coefficients are determined to be important. It appears that grouped LASSO is more stringent than regular LASSO (continued below)

|  |  |
| --- | --- |
|  | Grouped Coef |
| **(Intercept)** | 0 |
| **rvpADENOVIRUS C, INFLUENZA A 2009 H1N1** | -0.002058 |
| **rvpHMPV** | 0.1318 |
| **rvpINFLUENZA A 2009 H1N1** | -0.05195 |
| **rvpINFLUENZA A H3, INFLUENZA A** | -0.03152 |
| **rvpINFLUENZA A, INFLUENZA A 2009 H1N1** | -0.02106 |
| **rvpRHINOVIRUS** | 0.09209 |
| **rvpRHINOVIRUS, HMPV** | -0.09065 |
| **rvpRSV-A** | -0.04438 |
| **rvpRSV-B** | -0.07947 |
| **other.infect.statYES** | 0.1662 |
| **sexM** | -0.07815 |
| **prev.stayHOME** | 0 |
| **prev.stayHOSPITAL** | 0 |
| **prev.stayICU(ANOTHER INSTITUTION)** | 0 |
| **prev.staySNF** | 0 |
| **immun.comp.statYES** | 0 |
| **lung.dis.statTRUE** | 0 |
| **diabYES** | -0.05454 |
| **curbf1** | -0.0142 |
| **curbf2** | -0.01813 |
| **curbf3** | 0.03183 |
| **curbf4** | -0.01316 |
| **vpdoseYES** | 0 |
| **akiYES** | 0 |
| **pl.effYES** | 0.05589 |
| **dyspYES** | 0.1901 |
| **wheezeYES** | -0.2818 |
| **admitTRANSFER** | 0 |
| **admitWARDS** | 0 |
| **pnaYES** | 0 |
| **mech.ventYES** | 0 |
| **ecmoYES** | 0.1655 |
| **age** | 0 |
| **BMI30YES** | 0 |
| **psi** | 0 |
| **hct** | -0.1136 |

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred