Machine Learning Framework (for supporting information)

Data pre-processing

Cesarean sections (observations) with twin births, c-sections with arrival to delivery taking less than 5 or more than 90 minutes, or c-sections where blood loss was higher than 2000 ml. were removed. Additionally, based on expert knowledge observations with impossible predictor variable combinations (indicating erroneous observations) were dropped. This resulted in 349 of the 2892 observations were dropped from the analysis.

For the predictor variables, each categorical variable was split into binary variables with one for each possible categorical item and a missing item (if the categorical variable had missing values). This commonly called One-Hot or One-out-of-K encoding. Variables relating to the experience of each team member (midwife, surgeon, and anesthesiologist) were measured in years, but were transformed into ordinal variables with the following intervals, 0-1 year, 1-5 years, and 5+ years of experience. These intervals are represented by ordered integers, i.e. 0, 1, 2. For BMI, the real valued scale was transformed into underweight (<18.5), normal weight (18.5-25), overweight (25-30), class I obesity (30-35), class II obesity (35-40), and class 3 obesity (>40). Again an ordinal representation was used and the intervals are represented by integers from 0 to 5.

After preprocessing, the resulting data had 2543 observation and 92 predictor variables, which also included indicator variables for the CS grade 1, 2, 3, or 4. The 134 observations from CS grade 4 where not further analyzed. The data had 0.01% missing values, these values were imputed (median, single imputation). Although single imputation is often ill advised, in this case, the percentage of missing data is so low the effects are negligible.
Data Analysis

The preprocessed data was split 80-20 into a development and test set. The development set is used to establish the model parameters of the ElasticNet model\textsuperscript{1,2} (including the L1/L2 ratio hyperparameter), the test set is untouched and only used to establish performance of the final model. The number of observations is 1919 in the development set and 490 in the test set.

The analysis was performed on different subsets of the data, these subsets are defined as CS grade 1, 2 and 3 separately, and CS grade 1 and CS grade 2 combined. For these scenarios, the number of observations in the development set are 140, 809, 970, and 949, respectively. Similarly, their corresponding test sets have sizes 25, 213, 252, and 238.

For the CS-grade defined subset of development and test set, the following analysis was done for the outcome Arrival to Birth. For a development set consisting of N observations, the following procedure to determine the final ElasticNet model.

1. Standardized continuous variables to have zero mean and unit variance on the development set.

2. Jackknifing / Leave-one-out cross-validation\textsuperscript{3,4}, with N-1 observation as training data and 1 observation as validation set
   2.1. An inner 5-fold cross-validation is then used on training data (N-1 observations) to determine the optimal L1/L2 ratio for the elastic-net regularization.
   2.2. The optimal L1/L2 ratio is then used when predicting the outcome of the validation set (here a single observation).

The setup above allows the estimation of the validation error and its standard deviation over held-out data. The Jackknife samples from the loop above (a single jackknife sample consists of the parameter estimates when holding out a single observation) are used to identify predictor variable importance and their uncertainties.
Significant predictor variables are identified from the Jackknife samples using a two-sided t-test with alpha=0.01 and Bonferroni corrected\(^5\). For visualization, the weights of significant predictor variables are scaled such that their absolute value sum to 100. These weights are denoted as the percentage significant predictor weight (PSFW).

Finally, for prediction on the unseen test data, the final model is trained on the full (subset) development set using the median L1/L2 ratio determined above. The prediction error is used to assess model generalization.

**References**


