A Modified Random Forest Procedure with Survey Applications

Bryan Stanfill
Center for Survey Statistics and Methodology

October 1, 2012

Outline

Plant Breeders and Surveys

Survey Statisticians

- Interested in behavior/beliefs, e.g. PA levels

Plant Breeders

- Interested in traits/phenotypes, e.g. yield

PLANT BREEDERS AND SURVEYS

Survey Statisticians

- Interested in behavior/beliefs, e.g. PA levels
- Ask participant about PA levels

Plant Breeders

- Interested in traits/phenotypes, e.g. yield
- Gather genotype information

Plant Breeders and Surveys

Survey Statisticians

- Interested in behavior/beliefs, e.g. PA levels
- Ask participant about PA levels
- Calibrate survey to infer about true PA level

Plant Breeders

- Interested in traits/phenotypes, e.g. yield
- Gather genotype information
- Combine with field data to understand genome

PLANT BREEDERS AND SURVEYS

Survey Statisticians

- Interested in behavior/beliefs, e.g. PA levels
- Ask participant about PA levels
- Calibrate survey to infer about true PA level
- Estimating usual PA distribution

Plant Breeders

- Interested in traits/phenotypes, e.g. yield
- Gather genotype information
- Combine with field data to understand genome
- Predict future yield

Common Problems and Solutions

Problem

- Highly correlated covariates

Solution

- Only include one prominently

Common Problems and Solutions

Problem

- Highly correlated covariates
- Many unimportant covariates

Solution

- Only include one prominently
- Exclude unnecessary ones

COMMON PROBLEMS AND SOLUTIONS

Problem

- Highly correlated covariates
- Many unimportant covariates
- Lots of noise

Solution

- Only include one prominently
- Exclude unnecessary ones
- Model it

PAMS Objectives

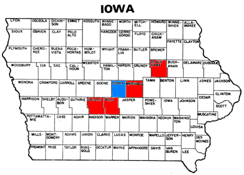
The Physical Activity Measurement Study (PAMS) is a survey designed to obtain information on physical activity patterns of

- Adult women and men (21-70)
- Hispanic and African American populations (limited sample size)
- Rural and non-rural adults

PAMS Objectives, Continued

More specifically,

- Individuals are sampled from four counties in Iowa: Marshall, Black Hawk, Dallas and Polk.
- Goal is 1200 participants at the end of the study, spanning two years.
  - Approximately equal number of males and females.
  - Approximately 10% African American and 10% Hispanic.
  - Minorities over sampled.



Data Collection Process

Data collection was intended to sample individuals uniformly over two years, partitioned into eight quarters.

At the individual level:

- Data were collected on two non-consecutive installments.
- For each installment:
  - Individual wore SenseWear Monitor for 24 hours.
  - 24-hour activity recall administered via phone the following day.
- Individuals filled out physical activity propensity questionnaire (PAPQ): 76 questions resulting in 279 columns per row

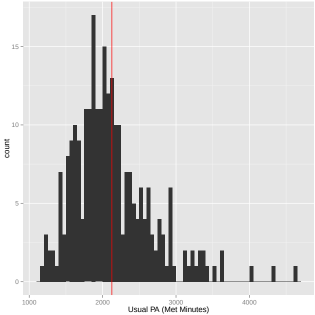
PAMS ANALYSIS

- My goal is to predict usual PA level based on PAPQ
- For each of the proposed models I will use 5-fold cross-validation to asses predictive accuracy
- Leave a random 1/5 of the observations out of the dataset, fit the model to the remaining observations and predict those that were left out
- Compute the root mean square prediction error (RMSPE) and test set correlation (Corr) for the full predicted dataset

Data

Below is a histogram and table summarizing the data I use

Gender	n	PA(Met Min)		BMI		Age	
Both	259	2125.79	(546.86)	29.83	(7.35)	49.14	(13.12)
Female	157	1952.93	(427.28)	30.05	(7.77)	51.46	(11.89)
Male	102	2391.87	(603.31)	29.49	(6.67)	45.57	(14.16)



Linear Models

Usual set-up:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i$$

with $\boldsymbol{y}$ and $\boldsymbol{X}$ known, $\boldsymbol{\beta}$ a parameter vector, $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$ for σ^2 a constant.

- Easy to communicate to quantitatively challenged individuals
- Inference requires some distributional assumption
- May require transformation to achieve this distributional assumption
- Requires tinkering to solve problems on previous slide

Bayesian LASSO & Regression (BLR[1])

General set-up:

$$y_i = \mu + \boldsymbol{x}'_{r_i}\boldsymbol{\beta}_r + \boldsymbol{x}'_{l_i}\boldsymbol{\beta}_l + u_i + \epsilon_i$$

where, $\boldsymbol{X}_r$ a matrix of "fixed" effects, $\boldsymbol{X}_l$ a matrix of "random" effects,

- μ an intercept
- $\boldsymbol{\beta}_r$ a vector of regression coefficients
- $\boldsymbol{\beta}_l$ a vector of LASSO coefficients
- u_i a random person effect
- ϵ_i remaining noise

[1] de los Campos (2009)

BLR, Priors

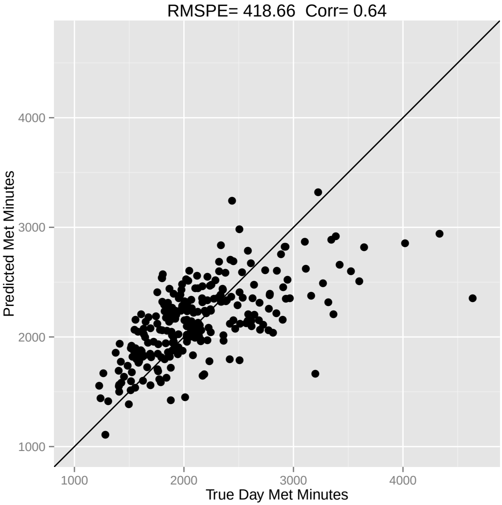
More details,

$$y_i = \mu + \boldsymbol{x}'_{r_i}\boldsymbol{\beta}_r + \boldsymbol{x}'_{l_i}\boldsymbol{\beta}_l + u_i + \epsilon_i$$

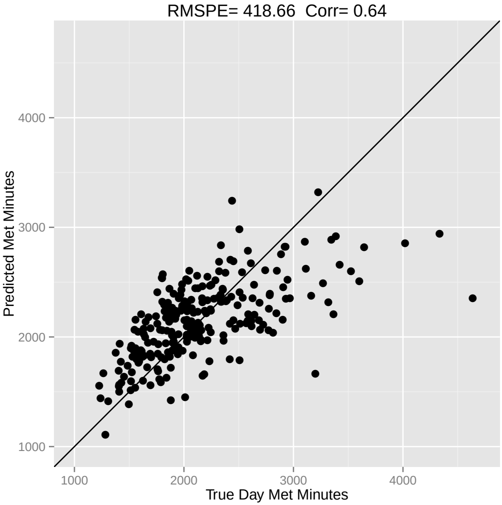
where

$$\begin{aligned}
\mu &\sim N(0, \sigma^2_\mu), \sigma^2_\mu \text{ chosen or modeled} \\
\boldsymbol{\beta}_r &\sim N(0, \boldsymbol{I}\sigma^2_r) \\
\beta_{l_j} &\sim N(0, \sigma^2_\epsilon \tau^2_j), \tau^2_j \sim \text{Exp}(\lambda) \text{ and } \lambda \sim p(\lambda) \\
\boldsymbol{u} &\sim N(0, \boldsymbol{A}\sigma^2_u), \boldsymbol{A} \text{ covariance computed from genealogy} \\
\epsilon_i &\sim N(0, \sigma^2_\epsilon) \\
\sigma^2_m &\sim \chi^{-2}(S_m, df_m) \text{ for } m \in \{\epsilon, u, r\}
\end{aligned}$$

BLR Results

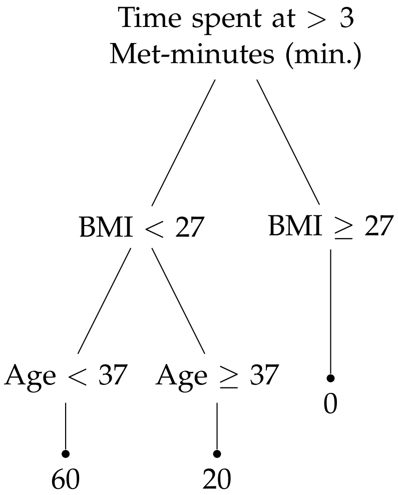

RMSPE= 418.66 Corr= 0.64
Predicted Met Minutes
True Day Met Minutes
1000
2000
3000
4000
1000
2000
3000
4000

BLR Results



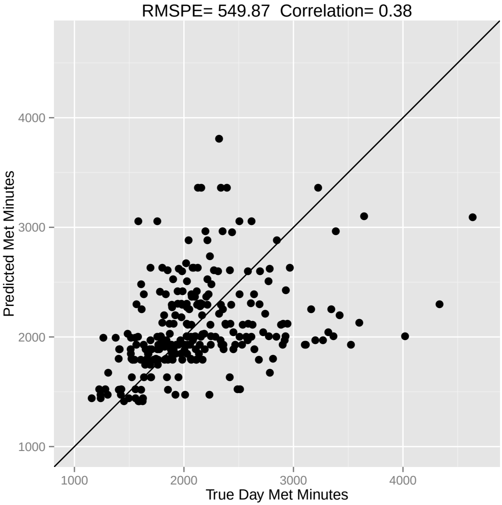
Drawbacks: prior specification, explicability

Trees

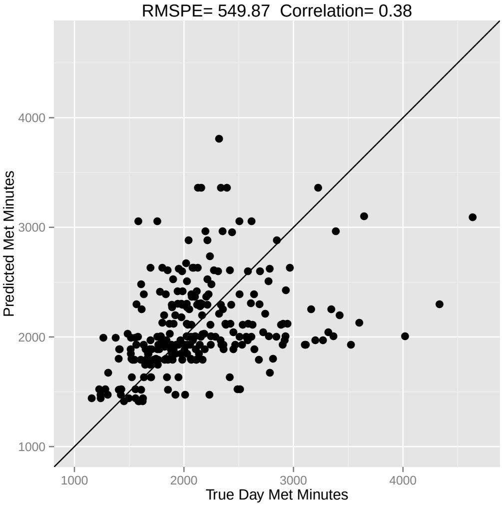


- Classification tree for discrete response
- Regression tree for continuous response
- Use covariate information to predict individual response
- Estimates and predictions are binned

TREE RESULTS



TREE RESULTS



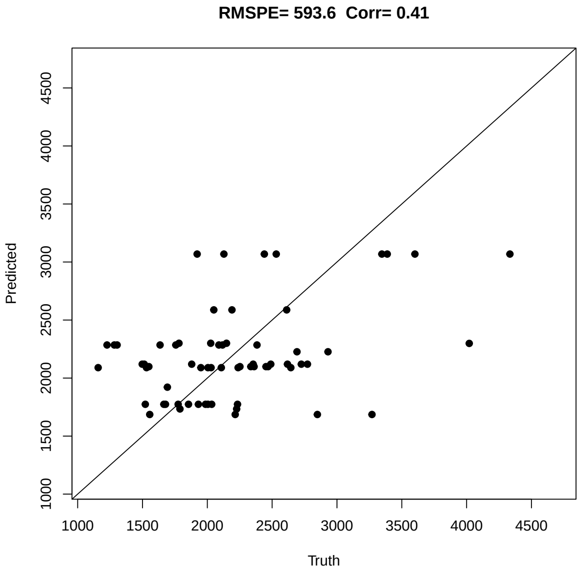
Drawbacks: predictive accuracy, binned predictions

Random Forest

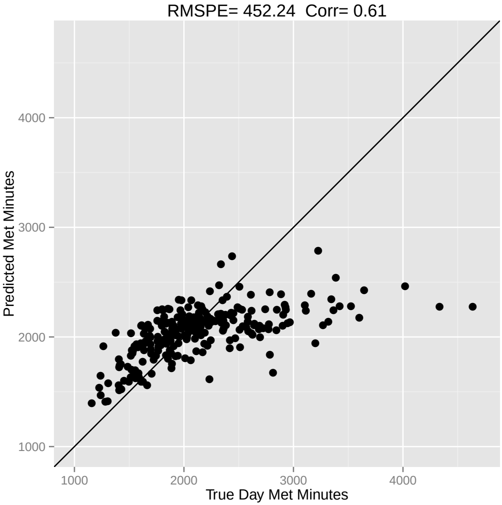
Assume a sample of size N with M covariates. To build each tree:

1. select n observations from N without replacement, called "in-bag" observations
2. at each node randomly select m covariates from which to split
3. with the complete tree classify the $N - n$ "out-of-bag" observations

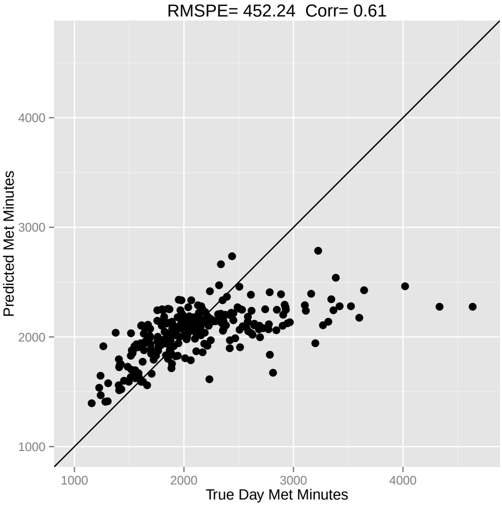
A forest's predictive accuracy is measured by it's "out-of-bag" (oob) error rate


RMSPE= 593.6 Corr= 0.41
Predicted
1000
1500
2000
2500
3000
3500
4000
4500
Truth
1000
1500
2000
2500
3000
3500
4000
4500

Random Forest Results



RANDOM FOREST RESULTS



Drawbacks: interpretability, parsimony, explicability

MOTIVATION

In plant breeding context we may be able to improve on basic random forest

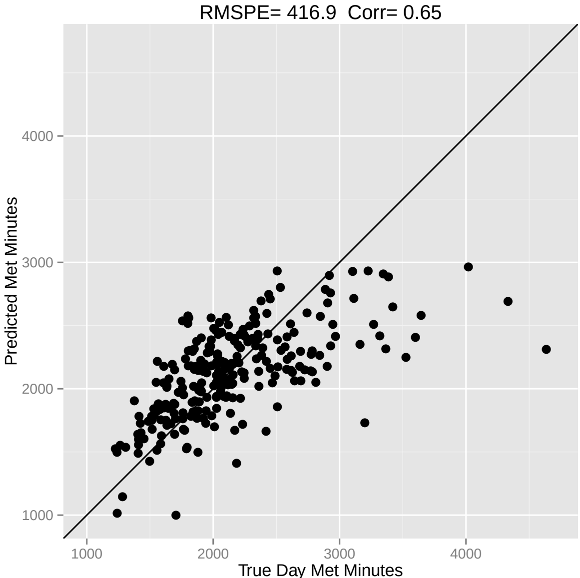
- Have 10's of covariates associated with design/pedigree
- Have 1000's of highly correlated or uninformative markers
- Why not use BLR concept with random forest predictive power?

Details

Assume a sample of size N with M covariates, M_1 which are known to be informative and M_2 which are highly correlated and largely uninformative. To build each tree:

1. select n observations from N without replacement, called "in-bag" observations
2. at each node split based on $\{M_1, m_2\}$ covariates where m_2 are chosen randomly from M_2
3. with the complete tree classify the $N - n$ "out-of-bag" observations

MODIFIED RF RESULTS



Discussion

- When there are lots of covariates that hold little to no information it's better to ignore a lot of them
- When all covariates have atleast some unique information this method is not as useful
- Here it seems that each individual question has something to contribute

Thanks! Questions?