Machine Learning: In the Trenches

Part I: Chris Andrews
University of Michigan

2019 05 30

Advertisement

This talk will cover a couple practical aspects of machine learning:

1. the feasibility of machine learning with correlated outcome data, such as paired eyes
2. assessing variable importance with LASSO in the presence of multi-collinearity among the covariates

Ophthalmology Example

- Response/Target Variable(s): Cystoid Macular Edema (CME) within 90 days after cataract surgery.
  - 13,397 surgeries in 8,560 patients
  - Both right (OD) and left (OS) eye surgeries in 4,837 patients

Ophthalmology Example

- Response/Target Variable(s): CME within 90 days of cataract surgery
- ~100 Predictors/Features, mostly patient-level:
  - Demographics: Age-at-Surgery, Sex, Race, Year-of-Birth, …
  - Geography: Zip Code, Census Tract, Distance to Eye Clinic, …
  - Economics: Insurance, Income, Wealth, …
  - Health: Disease Severity, Blood Test Results, Comorbidities, …
  - Eye Condition: Visual Acuity, Inter Ocular Pressure, Visual Field, …
  - Free Text from Clinic Visit Notes: Physician comments regarding condition, …
  - Treatment: Cumulative Dispersed Energy, Medications, …

First Question

- How does correlated outcome data affect Machine Learning Model Fitting?

Fitting Options

- Local Approach: individual target models
  - Fit model for left eye
  - Fit model for right eye
- Global Approach: multi-target model
  - E.g., Treat outcome as multinomial: (0,0); (1,0); (0,1); (1,1)
- Optimistic Approach:
  - Stack left and right eye data, ignore the dependence, and pretend sample size is “2N”

Fitting Options

- Local Approach: individual target models
  - Fit model for left eye
  - Fit model for right eye
- Global Approach: multi-target model
  - E.g., Treat outcome as multinomial: (0,0); (1,0); (0,1); (1,1)
- Optimistic Approach:
  - Stack data, ignore dependence, and pretend sample size is "2N"
- More

Covariates likely affect right and left eyes equally, but this structure is not utilized

Model assessment may be biased

A Prediction Tool: LASSO

- Using an old tool to demonstrate the point
  - Other (statistical) machine learning tools have same behaviors
- LASSO background:
  - Penalized Regression
  - 1 Tuning Parameter controls penalty
  - OLS estimates shrunk unequally toward 0 (variance/bias tradeoff)
  - Amount of shrinkage controlled by tuning parameter

Machine Learning Model Fitting

- Data are used to
  - estimate coefficients for a given tuning parameter value
  - select the optimal tuning parameter / optimal model
  - assess the quality of the selected model

Machine Learning Model Fitting

- Data are used to
  - estimate coefficients for a given tuning parameter value
  - select the optimal tuning parameter / optimal model
  - assess the performance of selected model
- If the same data are used for all 3 steps, the model may not generalize well to other situations. That is, the performance in other situations may not meet the claimed performance.
  - Train
  - Validate
  - Test

Machine Learning Model Fitting

- Data are used to
  - estimate coefficients for a given tuning parameter value
  - select the optimal tuning parameter / optimal model
  - assess the performance of selected model
- If the same data are used for all 3 steps, the model may not generalize well to other situations. That is, the performance in other situations may not meet the claimed performance.
  - Train
  - Validate
  - Test

"Optimism"

First Question

- How does correlated outcome data affect Machine Learning Model Fitting?
- Easier to address question when the data generating process is known
  - Start with a Simulation Study

Simulation Example

- Outcome measurements on right and left eyes of 10,000 participants together with 100 person-level covariates.
- $n = 10{,}000$
- $p = 100$
- $X_i \sim U(0,1)$
- $\beta_i = \begin{cases} (-1)^i i & i \leq 6 \\ 0 & i > 6 \end{cases}$
- $Z_i \sim Bernoulli_2\left(expit(\beta_i X_i + \beta_u X_u)\right)$
- X_u unmeasured, which induces correlation between Zs.

Simulation Example: Code Snippets

```
DATA wide long;
    DO i=1 TO &nObs;
    sign  = -1; BetaX = 0;
    /* variables that matter */
    do j=1 to dim(xIn);
        CALL RANUNI(seed,xIn[j]);
        BetaX =BetaX+j*sign*xIn{j};
        sign = -sign;
    END;
    /* variables that don't */
    DO j=1 TO dim(xOut);
        CALL RANUNI(seed,xOut[j]);
    END;
    /* 'unmeasured' variable */
    CALL RANUNI(seed, unm);
        BetaX =BetaX+betau*unm;
        sign = -sign;
    END;
```

Simulation Example: Code Snippets

```
    expitBetaX = 1/(1+exp(-BetaX/20));                  /* Pr(event) */
    CALL RANBIN(seed, 1, expitBetaX, yOD);              /* right */
    CALL RANBIN(seed, 1, expitBetaX, yOS);              /* left */
    yMulti = yOD + 2*yOS;                               /* joint */
    OUTPUT wide;
    eye="OD"; yBernoulli = yOD; OUTPUT long;            /*stacked */
    eye="OS"; yBernoulli = yOS; OUTPUT long;
END; RUN;
```

Simulation Example: Long Data

Obs	xIn1	xOut1	id	eye	yBernoulli
1	0.31487	0.13143	1	OD	0
2	0.31487	0.13143	1	OS	1
3	0.05473	0.13358	2	OD	1
4	0.05473	0.13358	2	OS	0
5	0.76013	0.57699	3	OD	0
6	0.76013	0.57699	3	OS	0

PROC HPGENSELECT

- Implements Logistic LASSO
- Cross-Validation not built-in but can be done ‘by hand’
  - Newer SAS products have PROC GENSELECT with CV options.

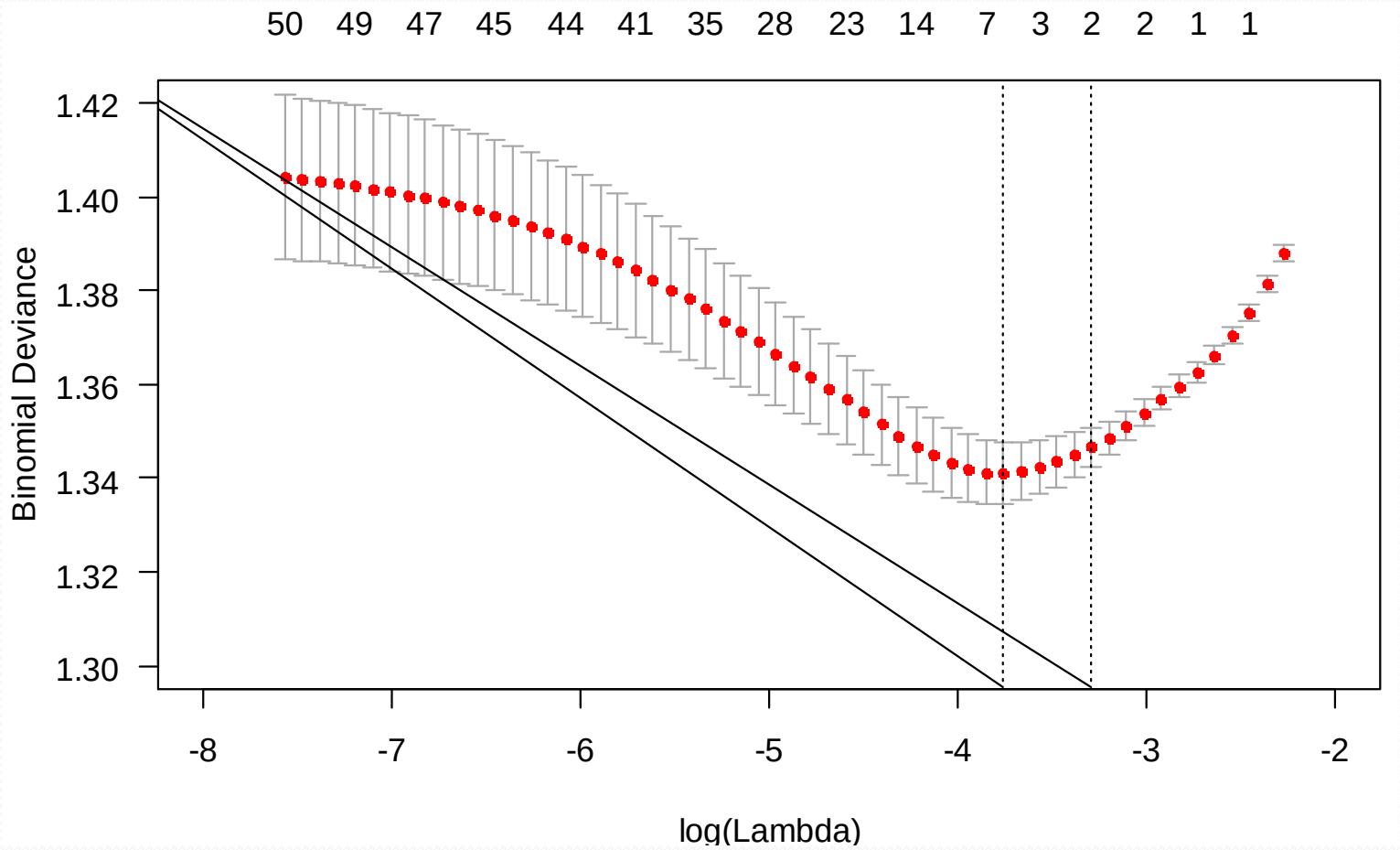
PROC HPGENSELECT

```
PROC HPGENSELECT DATA=wide;
        ODS OUTPUT LASSOSELECTIONDETAILS=lsd_&fold;
        MODEL y(event='1') = x: / DIST=Binary;
        SELECTION METHOD=lasso DETAILS=all;
        PARTITION ROLEVAR=fold(validation="&fold");
RUN;
```

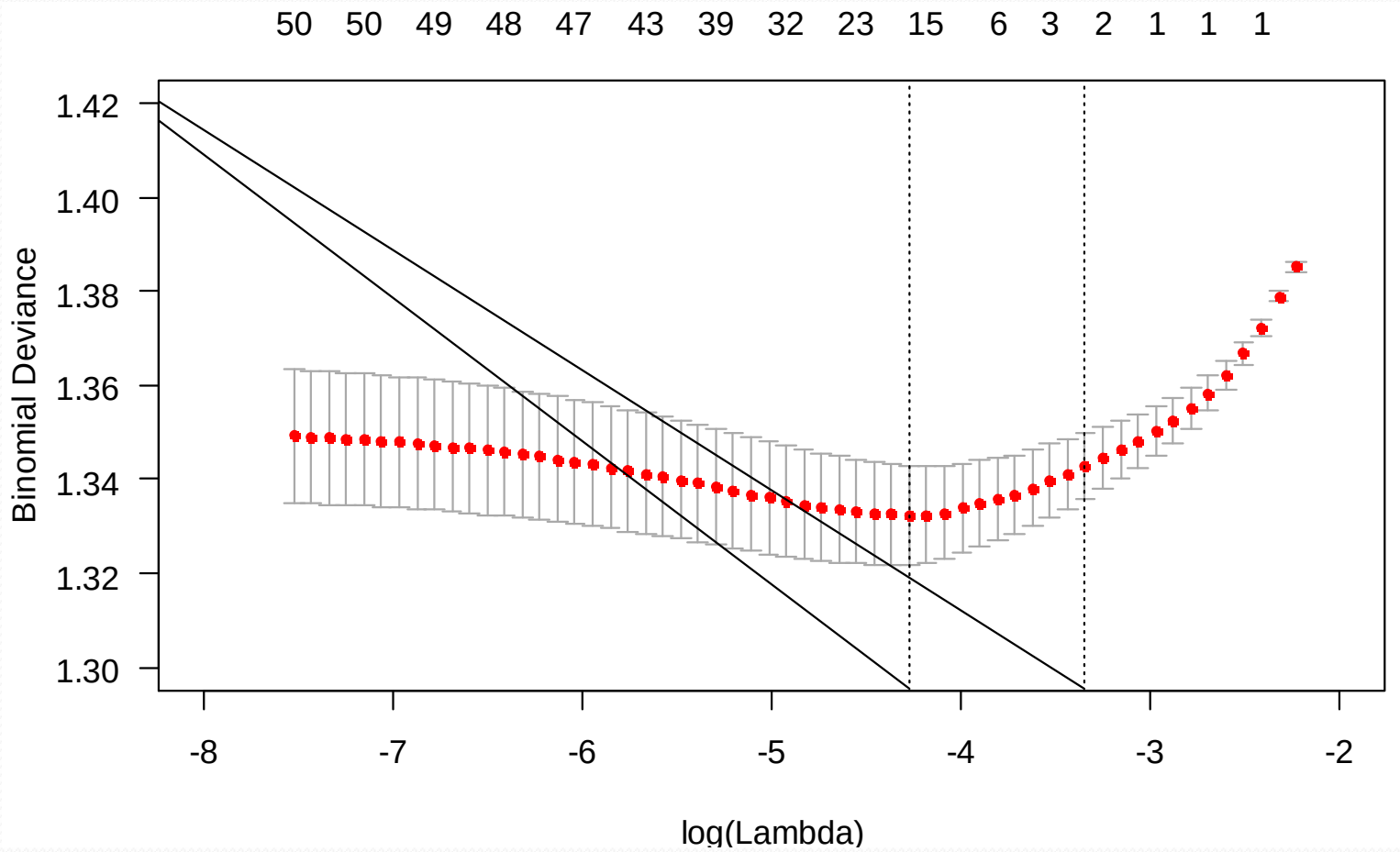
Put in a %DO fold=1 %TO 5 loop.

Combine the ODS files and plot to determine good penalty.

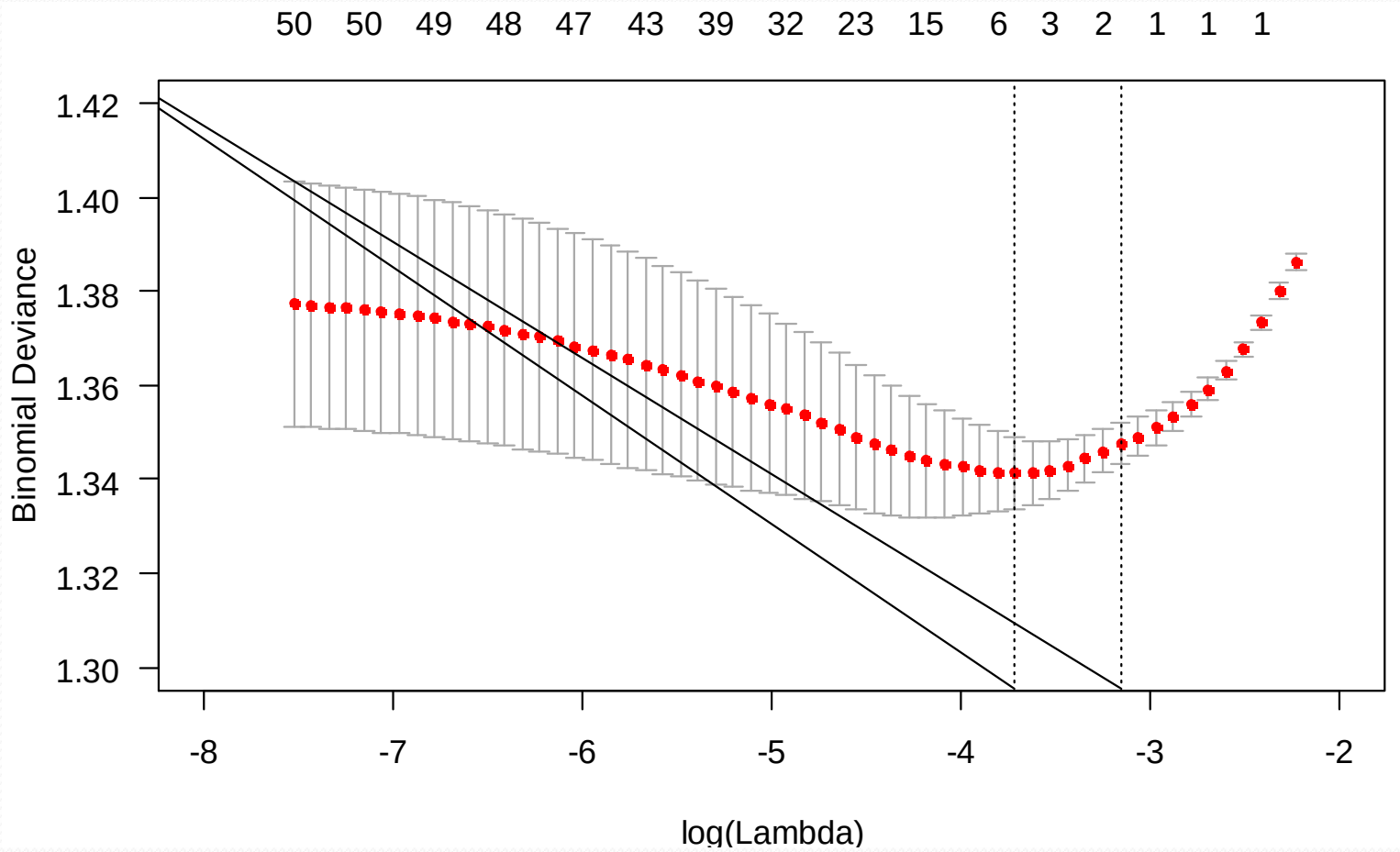
Optimal Penalty: One Eye



Optimal Penalty? Eye-Level Folds



Optimal Penalty: Person-Level Folds



Leakage

- Each eye in the validation fold is likely to have an eye in the training folds
- Those eyes have the same Xs (person-level covariates)
- The Ys from the same person are correlated beyond the measured covariates
- Information from training folds has leaked into validation fold
- Performance estimates will be too high (“Optimistic”)

Fitting Comparison

	One Eye	Eye-Level Folds	Person-Level Folds
Estimated Average Squared Error	0.241	0.237	0.241
Holdout Average Squared Error	0.254	0.258	0.254
“Optimism”	5.4%	8.2%	5.4%

Fitting Comparison

	One Eye	Eye-Level Folds	Person-Level Folds
Missed Strong Predictor	1%	0%	0%
Missed Weak Predictor	28%	6%	23%
# Extraneous Variables Selected	4.9	17.5	5.1

Cystoid Macular Edema Example

- A Machine Learning approach outperformed a logistic regression analysis
  - The predictive accuracies of the Logistic Regression Model and the Machine Learning Model applied to the hold-out data were 0.88 and 0.96
  - The false positive rates were 0.09 and 0.01
  - The positive predictive values were 0.11 and 0.48
  - The AUCs (c-statistics) were 0.64 and 0.75

Cystoid Macular Edema Example

- A Machine Learning approach outperformed a logistic regression analysis
  - The predictive accuracies of the Logistic Regression Model and the Machine Learning Model applied to the hold-out data were 0.88 and 0.96
  - The false positive rates were 0.09 and 0.01
  - The positive predictive values were 0.11 and 0.48
  - The AUCs (c-statistics) were 0.64 and 0.75
- *The most predictive model inputs were* the surgeon performing the operation, the duration of the surgery, and the patient's age at surgery, birth month, body mass index, and sex.

Second Question

- Assessing variable importance when predictors are correlated

Variable Importance: Prediction Accuracy

- One method of estimating the importance of a variable in prediction accuracy is to determine how much the accuracy degrades when that variable is excluded.
- The most predictive variable is scaled to 100% importance.
- Importance of other variables are measured as percentages of that.

Variable Importance: Coefficient Size

- In today's simulation, all explanatory variables have the same variance.
- Thus the magnitudes of the regression coefficients are a measures importance
  - If X variances are not equal, the regression coefficients cannot be used to rank variables in this way.
  - If the Machine Learning algorithm doesn't have coefficients, this approach can't be used.

Correlated Predictors: Simulation

Both Predictors In Model

- Coefficients of xIn20 and xRepeat are smaller than expected.

xIn16	1	0.745750
xIn17	1	-0.770222
xIn18	1	0.903520
xIn19	1	-0.878764
xIn20	1	0.430800
xRepeat	1	0.430800

One Predictor In Model

- Coefficient of xIn20 large, as expected.

xIn16	1	0.745761
xIn17	1	-0.770240
xIn18	1	0.903540
xIn19	1	-0.878782
xIn20	1	0.861609

Correlated Predictors: CME Example

- Age-at-surgery is a valuable predictor of CME.
- Year-of-birth is a valuable predictor of CME.
- As the dataset covers a relatively short calendar period (2013-2017), these are highly correlated ($r>0.9$)
- Their importance would have been artificially low due to the correlation if both were available for prediction.

Correlated Predictors: CME Context

Age and Birth Year Both in Model

- Importance of Age and Birth Year are smaller than expected.

Variable	Importance
Best	100%
…	…
Age	60%
Birth Year	55%
…	…

Age Only In Model

- Importance of Age as expected.

Variable	Importance
Age	100%
…	…
Others	…
…	…
…	…

Morals of the Stories

- Incorrect analysis of correlated outcomes may
  - Increase apparent prediction accuracy but actually decrease prediction accuracy on new data
  - Lead to overly complicated models
- Attention to the cross-validation scheme may lead to a stronger model identification

Morals of the Stories

- Incorrect analysis of correlated outcomes may
  - Increase apparent prediction accuracy but actually decrease prediction accuracy on new data
  - Lead to over complicated model
- Attention to the cross-validation scheme may lead to a stronger model identification
- Measuring the importance of individual predictors when predictors are correlated is a problem in data science just like in statistics.

Thank you

- Chris Andrews
- chrisaa@umich.edu