Data Mining
Model Selection

Bob Stine
Dept of Statistics, Wharton School
University of Pennsylvania

Wharton
Department of Statistics

From Last Time

- Review from prior class
  - Calibration
  - Missing data procedures

    Missing at random vs. informative missing
  - Problems of greedy model selection

    Problems with stepwise regression.

    So then why be greedy?
- Questions
  - Missing data procedure: Why not impute?

    "Add an indicator" is fast, suited to problems with many missing.
    Imputation more suited to small, well-specified models.

    EG. Suppose every X has missing values. How many imputation models do you need to build, and which cases should you use?

Topics for Today

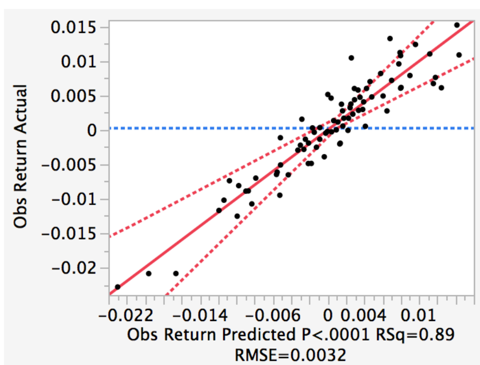
- Over-fitting
    - Model promises more than it delivers
- Model selection procedures
    - Subset selection
    - Regularization (aka, shrinkage)
    - Averaging
- Cross-validation

Model Validation

- Narrow interpretation
  - A predictive model is "valid" if its predictions have the properties advertised by model
  - Calibrated, right on average
  - Correct uncertainty, at least variance

mean
&
variance

- Must know process that selected model
  - Cannot validate a model from a static, "published perspective"
  - Stepwise model for S&P 500 looks okay, but...



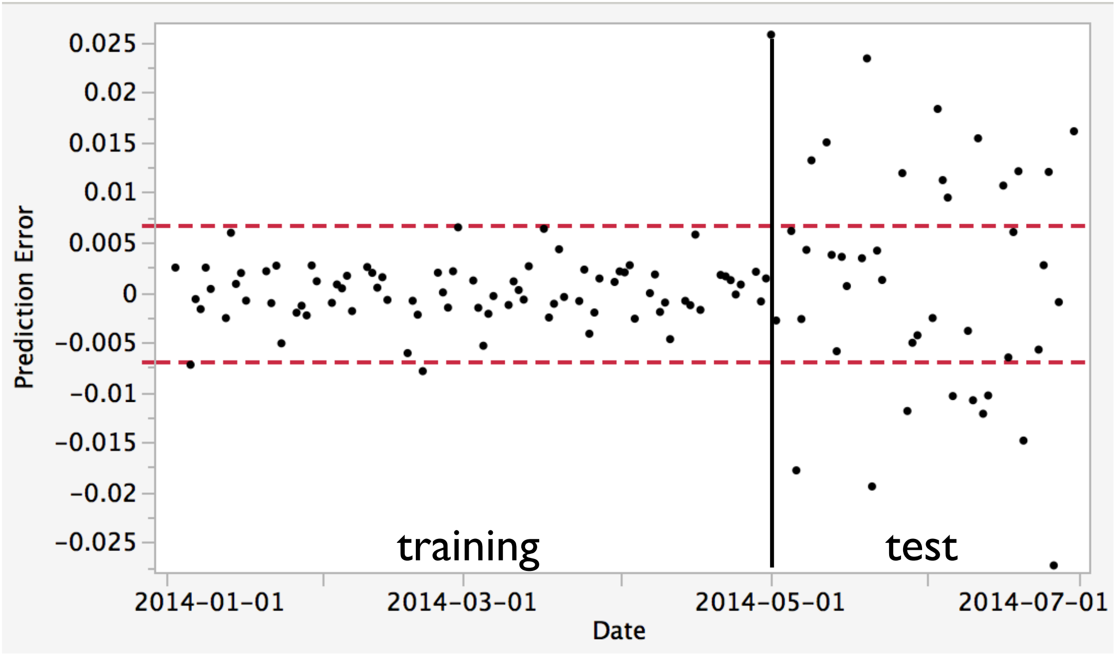
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	29	0.00424379	0.000146	14.1056
Error	52	0.00053947	0.000010	Prob > F
C. Total	81	0.00478325		<.0001*

Model Validation

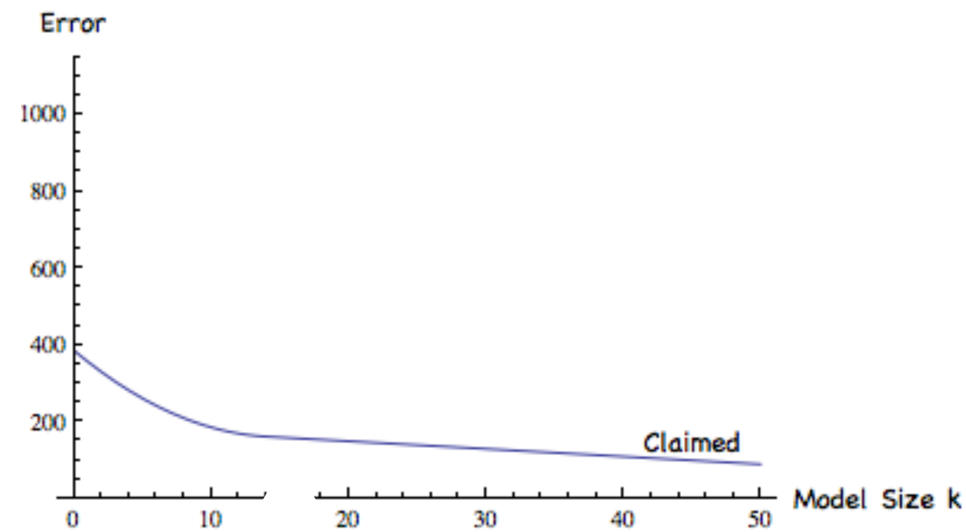
- Fails miserably (as it should) when used to predict future returns
  - Predictors are simply random noise
  - Greedy selection overfits, finding coincidental patterns

±2 RMSE



Over-Fitting

- Critical problem in data mining
  - Caused by an excess of potential explanatory variables (predictors)
- Claimed error rate steadily falls with size of the model
- "Over-confident"
  - Model claims to predict new cases better than it will.



- Challenge
  - Select predictors that produce a model that minimizes the prediction error without over-fitting.

Wharton
Department of Statistics

Multiplicity

- Why is overfitting common?
- Classical model comparison
  - Test statistic, like the usual t-statistic
    Special case of likelihood ratio test
  - Designed for testing one, a priori hypothesis
  - Reject if $|t| > 2$, p-value < 0.05
- Problem of multiple testing (multiplicity)
  - What is the chance that the largest of p z-statistics is greater than 2?

p	P(max |z|>1.96)
1	0.05
5	0.23
25	0.72
100	0.99

Model Selection

- Approaches
  - Find predictive model without overfitting
  - Three broad methods
- Subset selection
  - Greedy L_0 methods like forward stepwise
  - Penalized likelihood (AIC, BIC, RIC)
- Shrinkage
  - Regularized: L_1 (lasso) and L_2 (ridge regression)
  - Bayesian connections, shrink toward prior
- Model averaging

  Next week
  - Don't pick one; rather, average several

Subset Solution

- Bonferroni procedure
  - If testing p hypotheses, then test each at level α/p rather than testing each at level α.
  - $\Pr(\text{Error in p tests}) = \Pr(E_1 \text{ or } E_2 \text{ or } \ldots E_p)$
    $\leq \Sigma \Pr(\text{Error } i^{th} \text{ test})$
  - If test each at level α/p, then
    $\Pr(\text{Error in p tests}) \leq p(\alpha/p) = \alpha$
- Not very popular… easy to see why
  - Loss of power
- Cost of data-driven hypothesis testing

p	Bonferroni z
5	2.6
25	3.1
100	3.5
100000	5.0

Wharton
Department of Statistics

Discussion

- Bonferroni is pretty tight
  - Inequality is almost equality if tests are independent and threshold α/p is small
- Flexible
  - Don't have to test every H_0 at same level
  - Allocate more α to 'interesting' tests

    Split α=0.05 with ½ to p linear terms and ½ to all interactions
- Process matters
  - Look at model for stock market in prior class
  - Many predictors in model pass Bonferroni!

    The selection process produces biased estimate of error σ

    Use Bonferroni from the start, not at the end

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0047436	0.000834	5.69	<.0001*
Trading Rule 02	−0.002382	0.000526	−4.53	<.0001*
Trading Rule 06	−0.001643	0.000473	−3.47	0.0010*
Trading Rule 07	−0.002415	0.000501	−4.82	<.0001*
Trading Rule 10	0.0014874	0.000401	3.71	0.0005*

Popular Alternative Rules

- Model selection criteria
  - AIC (Akaike information criterion, C_p)
  - BIC (Bayesian information criterion, SIC)
  - RIC (risk inflation criterion)
- Designed to solve different problems
  - "Equivalent" to varying p-to-enter threshold
  - AIC, C_p: Accept variable if $z^2 > 2$

    Equivalent to putting p-to-enter ≈ 0.16
  - BIC: $z^2 > \log n$

    Aims to identify the "true model"
  - RIC: $z^2 > 2 \log p$ $\approx$ Bonferroni

    The more you consider, the stiffer the penalty

Penalized Likelihood

- Alternative characterization of criteria
- Maximum likelihood in LS regression
  - Find model that minimizes -2 log likelihood
  - Problem: always adds more variables (max R^2)
- Penalized methods
  - Add predictors so long as

    $$-2 \log \text{likelihood} + \lambda \text{ (model size)}$$

    decreases
- Criteria vary in choice of λ
  - 2 for AIC, (log n) for BIC, (2 log p) for RIC

Example

- JMP output
  - Osteo example
- Results
  - Add variables so long as BIC decreases
  - Fit extra then reverts back to best
- AIC vs BIC
  - AIC: less penalty, larger model

Stepwise Fit for ZHIP

Stepwise Regression Control

Stopping Rule: Minimum BIC

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
1171.7563	1217	0.9812358	0.4302	0.4236	75.573847	15	3466.946	3548.36

Current Estimates

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	WEIGHT	Entered	0.0000	455.2374	0.2214	517.71	2	3825.14	3840.47
2	AGE	Entered	0.0000	247.4688	0.3417	249.89	3	3620.31	3640.74
3	FRACTURE?	Entered	0.0000	36.51861	0.3595	212.07	4	3588.63	3614.16
4	OSTASE	Entered	0.0000	26.10392	0.3722	185.61	5	3565.99	3596.62
5	RHEUARTH	Entered	0.0000	26.58858	0.3851	158.62	6	3542.37	3578.09
6	RACE-2	Entered	0.0000	17.46336	0.3936	141.58	7	3527.26	3568.08
7	H_LOS	Entered	0.0002	13.72112	0.4003	128.62	8	3515.66	3561.56
8	PTHI_TM	Entered	0.0013	10.4361	0.4053	119.24	9	3507.22	3558.21
9	HEA	Entered	0.0012	10.45311	0.4104	109.85	10	3498.68	3554.75
10	NTEL_URC	Entered	0.0031	8.651519	0.4146	102.41	11	3491.9	3553.04
11	CAL	Entered	0.0018	9.537956	0.4193	94.013	12	3484.14	3550.36
12	B_1M	Entered	0.0061	7.35131	0.4229	87.998	13	3478.58	3549.87
13	DISC_DO	Entered	0.0048	7.709577	0.4266	81.592	14	3472.6	3548.95
14	MOVE	Entered	0.0058	7.354428	0.4302	75.574	15	3466.95	3548.36
15	CHOL	Entered	0.0085	6.656982	0.4334	70.316	16	3461.98	3548.46
16	PCO_TM	Entered	0.0153	5.633457	0.4362	66.173	17	3458.07	3549.6
17	YR_POST	Entered	0.0157	5.558966	0.4389	62.113	18	3454.21	3550.8
18	HYST	Entered	0.0112	6.101918	0.4418	57.46	19	3449.75	3551.38
19	SUBJ	Entered	0.0101	6.247202	0.4449	52.648	20	3445.09	3551.77
20	IMM_FAM	Entered	0.0261	4.656512	0.4471	49.571	21	3442.13	3553.86
21	EARLYACT	Entered	0.0243	4.755394	0.4494	46.386	22	3439.04	3555.81
22	FRACTRIB	Entered	0.0316	4.320051	0.4515	43.676	23	3436.42	3558.21
23	PLT	Entered	0.0333	4.224109	0.4536	41.071	24	3433.88	3560.71
24	Miss-HAIR	Entered	0.0315	4.297186	0.4557	38.385	25	3431.24	3563.1
25	Best	Specific	.	.	0.4302	75.574	15	3466.95	3548.36

What happens if try either with stock market model?

Shrinkage Solution

- Saturated model
  - Rather than pick a subset, consider models that contain <u>all</u> possible features
  - Good start (and maybe finished) if $p << n$

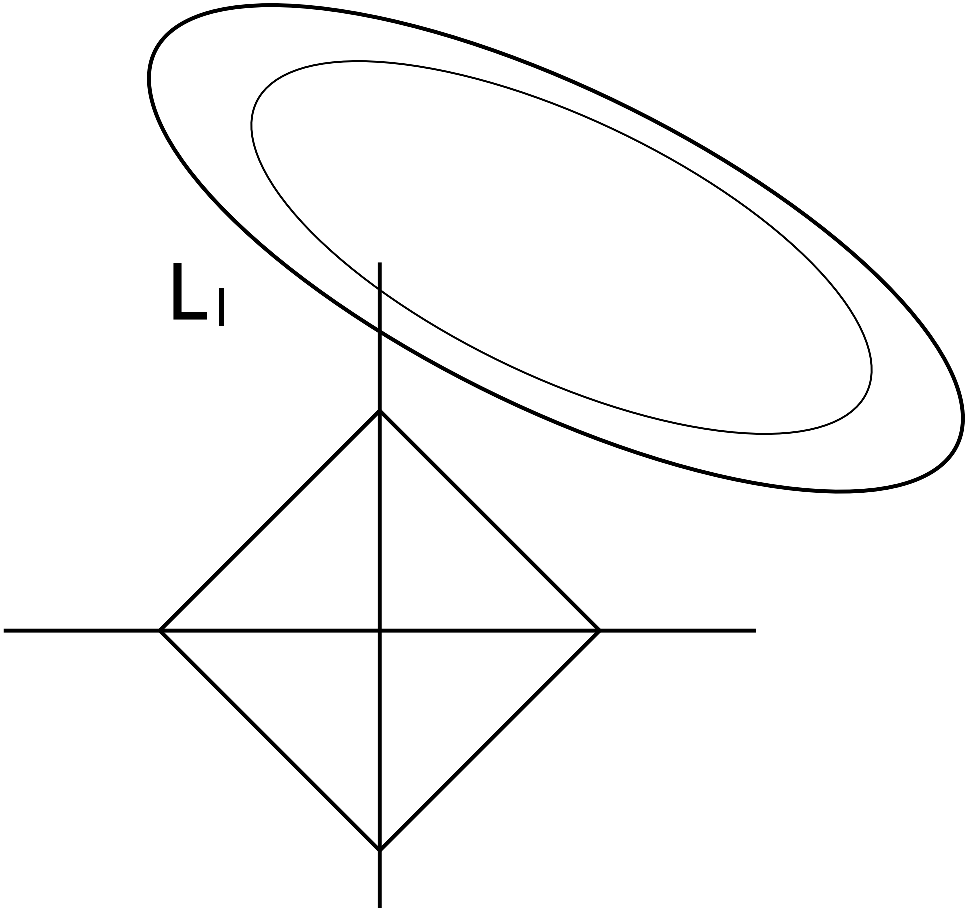
p = # possible Xs

- Shrinkage allows fitting all if $p > n$
- Shrinkage maximizes penalized likelihood
  - Penalize by "size" of the coefficients
  - Fit has to improve by enough (RSS decrease) to compensate for size of coefficients
  - Ridge regression: min RSS + λ_2 b'b
  - LASSO regression: min RSS + λ_1 Σ|bj|

RSS analogous to -2 log likelihood

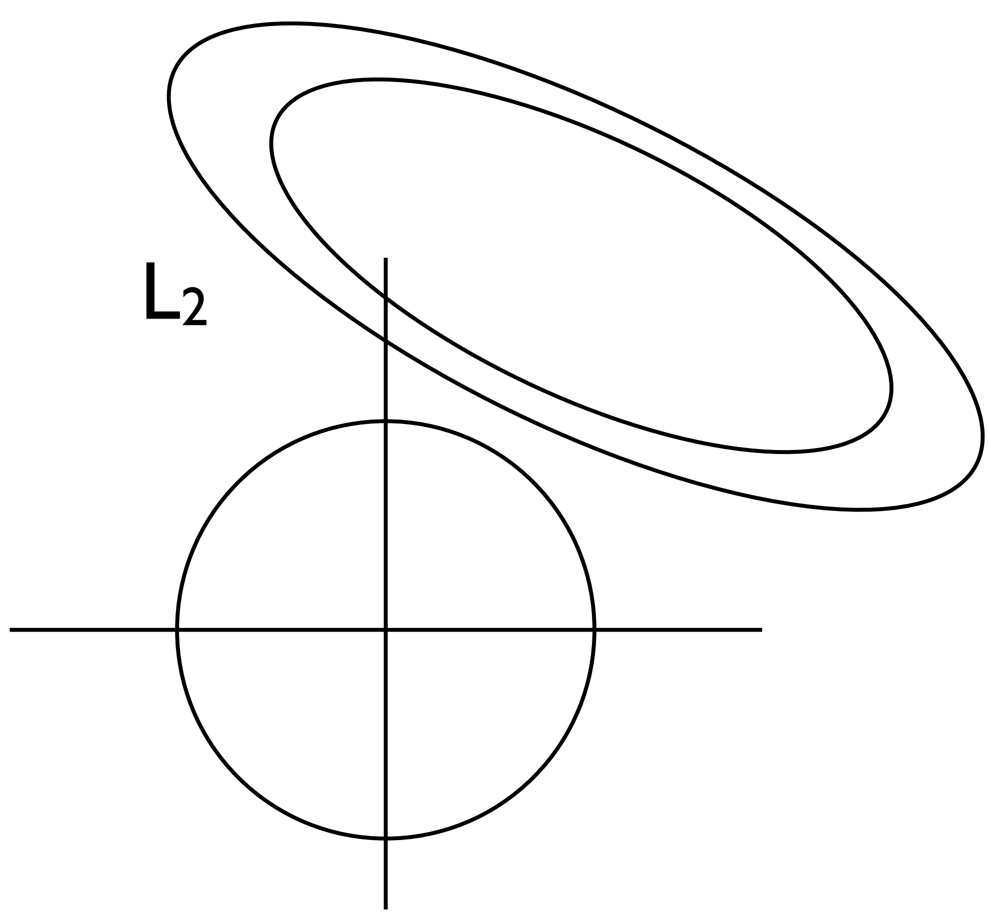
λ = regularization parameter, a tuning parameter that must be chosen

Lasso vs Ridge Regression



min RSS, $\Sigma|b_j|<c$

Corners produce selection

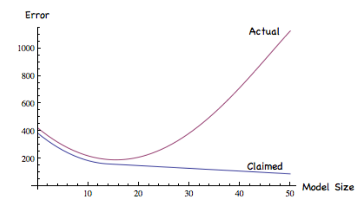


min RSS, $\Sigma b_j^2<c$

Interpret λ as Lagrange multiplier.

Cross-Validation Solution

- Common sense alternative to criteria
  - Apply the model to new data
  - Estimate 'hidden' curve plot of over-fitting



- No free lunches
  - Trade-off

    More data for testing means less for fitting:

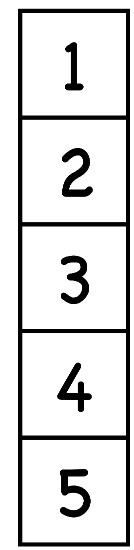
    Good estimate of the fit of a poorly estimated model.

    Poor estimate of the fit of a well estimated model.
  - Highly variable

    Results depend which group was excluded for testing

    Multi-fold cross-validation has become common
  - Optimistic

    Only place I know of a random sample from same population
  - Multi-fold: leave out different subsets



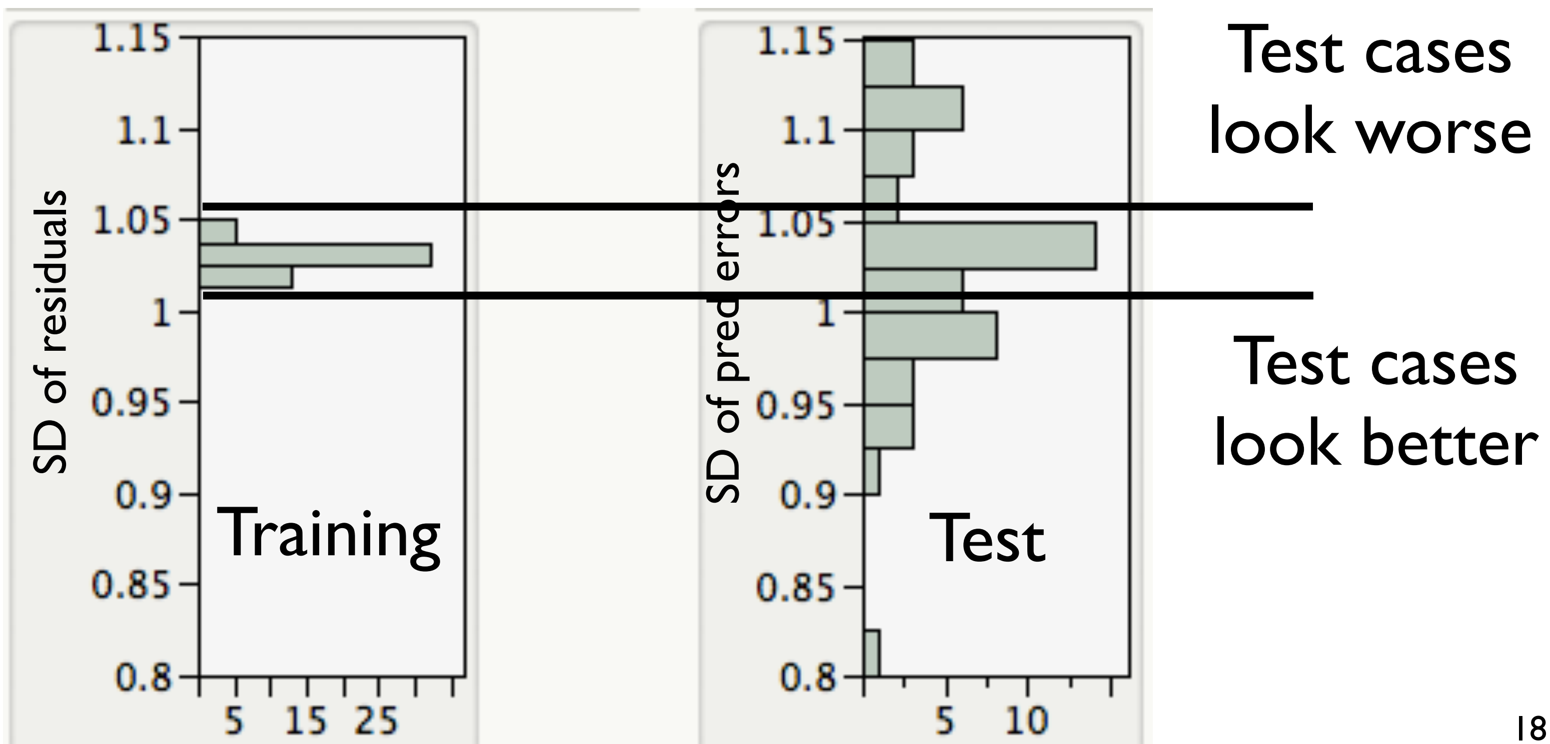
Variability of CV

- Example
  - Compare ‘simple’ and ‘complex’ osteo models

    Need to fit both to the same CV samples… Not so easy in JMP
  - Evaluate one model
- Method of validation
  - Exclude some of the cases
  - Fit the model to others
  - Predict the held-back cases
  - Repeat, allowing missing data to affect results
  - Compare out-of-sample errors to model claims
- Is assessment correct?
  - Under what conditions?

Osteo Example

- CV 50 times, split sample
- Variability
  - If only did one CV sample, might think model would be 20% better or 15% worse than claimed!



CV in Data Mining

- DM methods often require a three-way CV
  - Training sample to fit model
  - Tuning sample to pick special constants
  - Test sample to see how well final model does
- Methods without tuning sample have advantage
  - Use all of the data to pick the model, without having to reserve a portion for the choice of constants
  - Example: method that has "honest" p-values, akin to regression model with Bonferroni
- Caution
  - Software not always clear how the CV is done
  - Be sure CV includes the choice of form of model

Lasso

- Regularized regression model
  - Find regression that minimizes

    $$\text{Residual SS} + \lambda\, \Sigma|\beta_i|$$

    where λ is a tuning constant
  - Bayesian: double exponential prior on β
  - Scaling issues

    What happens if the β's are not on a common scale?
- L_1 shrinkage
  - Shrink estimated parameters toward zero
  - Penalty determines amount of shrinkage

    Larger penalty (λ), fewer variable effects in model
  - Equivalent to constrained optimization

Lasso Example

- How to set the tuning parameter λ?
- Empirical: Vary λ to see how fit changes
  - Cross-validation, typically 10-fold CV
  - Large values of λ lead to very sparse models

    Shrinks all the way back to zero
  - Small values of λ produce dense models
  - CV compares prediction errors for choices
- Implementations
  - Generalized regression in JMP Pro
  - glmnet package in R (See James et al, Ch 6)

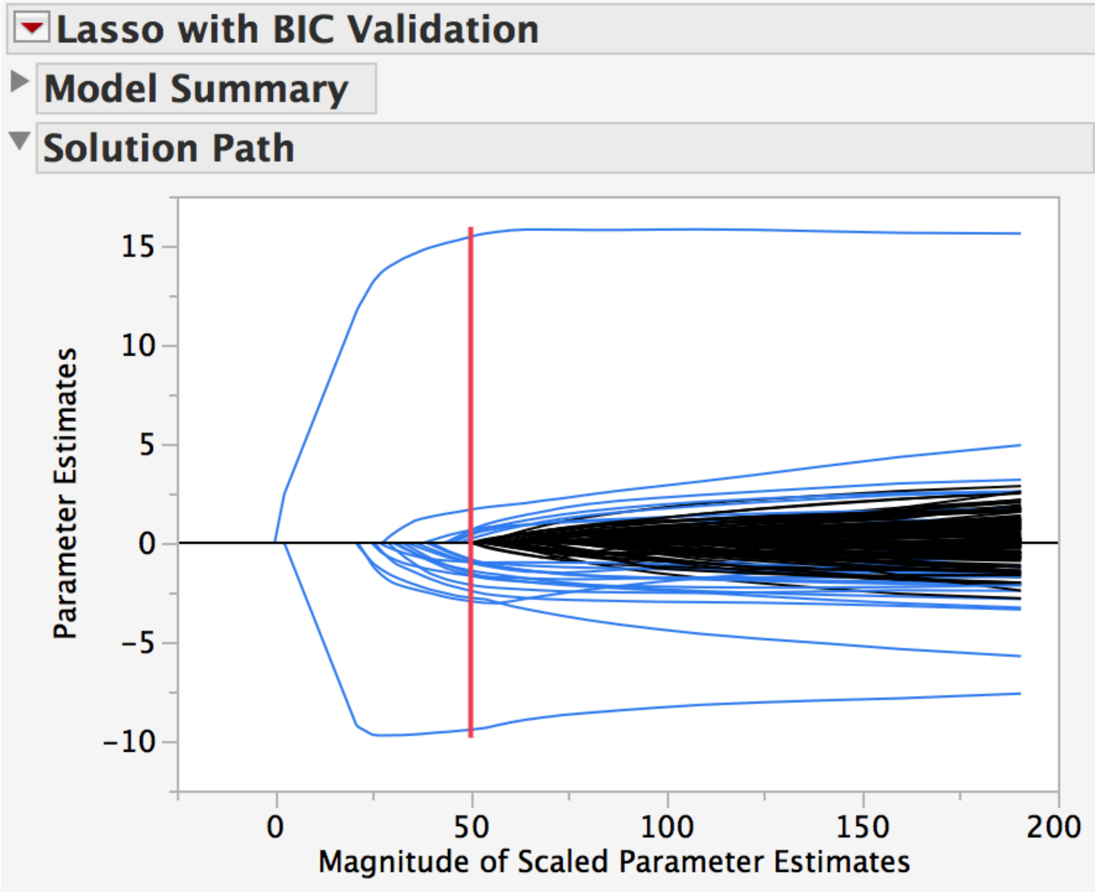
    More "naked" software than JMP or Stata

Lasso Example

- Fit L_1 regression, Lasso
  - Plot estimated coefficients as relax penalty
  - Implemented in JMP as "generalized regression"

osteo model

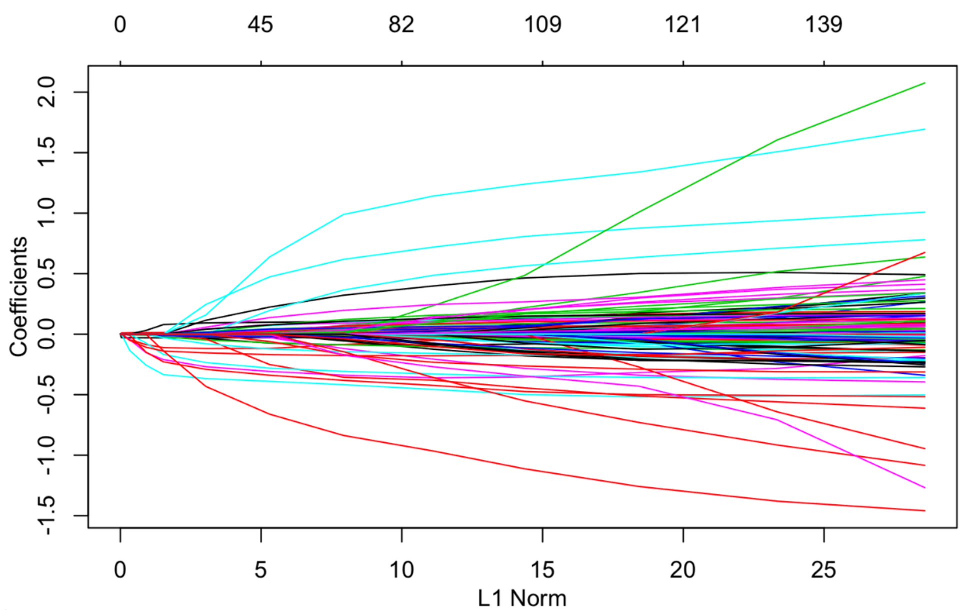
Where to stop adding features?

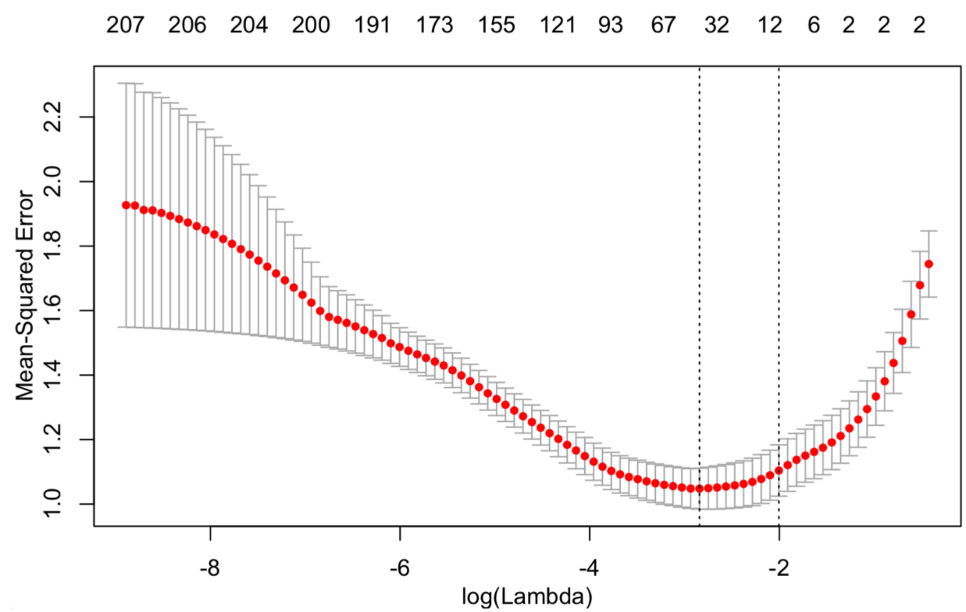


Wharton
Department of Statistics

Lasso Example in R

- Follow script from James
  - See on-line document “Glmnet Vignette”
- Similar output
  - Less formatting, but more accessible details





Repeated 10-fold CV

Discussion of CV

- Use in model selection vs model validation
  - Shrinkage methods use CV to pick model
  - Validation reserves data to test final model
- Comments on use in validation
  - Cannot do selection and validation at same time
  - Flexible: models do not have to be nested
  - Optimistic

    Splits in CV are samples from one "population"

    Real test in practice often collected later than training data
  - Population drift

    Populations often change over time; CV considers a shapshot
- Alternatives?
  - Bootstrap methods

Take-Aways

- Overfitting
  - Increased model complexity often claims to produce a better fit, but in fact it got worse
- Model selection methods
  - Criteria such as AIC or p-value thresholds
  - Shrinkage methods such as lasso
- Cross validation
  - Multiple roles: validation vs model selection
  - Flexible and intuitive, but highly variable

Some questions to ponder...

- If you fit a regression model with 10 coefficients, what's the chance that one is statistically significant by chance alone?
  - How can you avoid this problem?
- If you have a coefficient in your model that has a $t \approx 2$, what is going to happen to its significance if you apply split-sample CV?
- Why is cross-validation used to pick lasso models?
- Is further CV needed to validation a lasso fit?

Next Time

- Thursday   Newberry Lab
  - Hands-on time with JMP, R, and data
  - Fit models to the ANES data

    You can come to class, but I won't be here!

- Friday   July 4th holiday