

ISyE 6416: Computational Statistics Spring 2017




Lecture 9: Model Selection

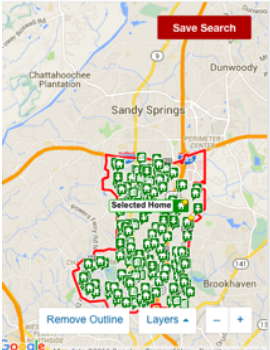
Prof. Yao Xie

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Motivating example: real estate agent

Data was collected on 100 homes recently sold in a city. It consisted of the sale price, house size, the number of bedrooms, the number of bathrooms, the lot size, and the annual real estate tax. Use price as the response variable and determine which of these factors should be included in the regression model.


REDFIN 30342   Filters Tools ▾ Chrystal ▾ 

 Save Search

30342 Real Estate


Showing 234 homes, sorting by [recommended](#) ▾

[Table](#) [Photos](#)







\$599,900
10 Battle Ridge Dr
Atlanta, GA 30342




5 Beds | 4 Baths | 3,742 Sq. Ft.

 Pam O'Connor-Smith saw this home
REDFIN Agent

"Spacious home at the entrance to the community. There is no foyer entry, you are immediately in a sunning area. There is also a s..." [More](#)

HOA \$300 Status Active
\$/Sq. Ft. \$160 On Redfin 54 days
Year Built 1988
Lot Size 0.61 Acres

  [Go Tour It](#)  

Address	Location	Price	Beds	Baths	Sq.Ft.	\$/Sq.Ft.	Days
 10 Battle Ridge Dr	Battle Creek	\$599,900	5	4	3,742	\$160	54
 4650 Windsor Gate Ct	Windsor A...	\$625,000	5	4.5	3,564	\$175	167
 245 Mystic Ridge Hl #245	Mystic Ridge	\$450,000	4	3	—	—	1

Goal of model selection

$$y_i = a_1x_{1i} + \cdots + a_px_{pi} + \epsilon_i, \quad i = 1, \dots, n$$

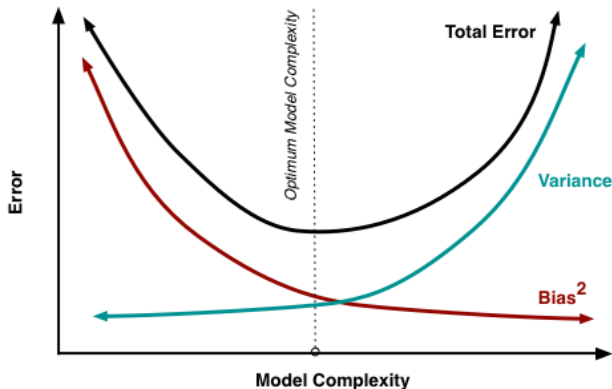
- ▶ When we have p variables (with many possible interactions x_ix_j), it can be difficult to find a good model (a subset of variables to explain response)
- ▶ Which main effect and interaction do we include?
“interpretable model”
- ▶ Essentially this is a combinatoric problem with 2^p possibilities
- ▶ Model selection tries to “simplify” this task
- ▶ The problem of picking out the relevant variables from a larger set is called “model selection” or “variable selection”
- ▶ estimating some coefficients to be exactly 0

Bias-Variance tradeoff

$$y_i = a_1 x_{1i} + \cdots + a_p x_{pi} + \epsilon_i, \quad i = 1, \dots, n$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$\text{MSE} = \text{BIAS}^2 + \text{Variance}$$



Ridge regression

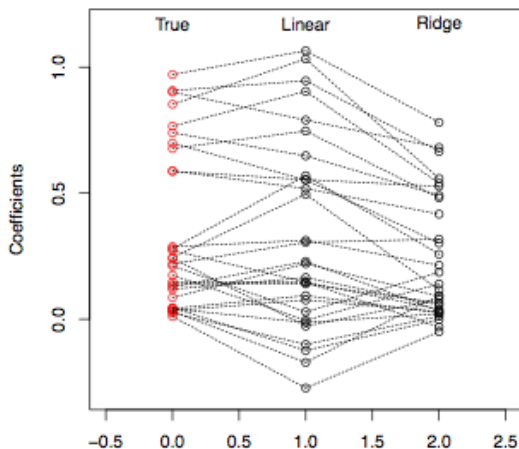
- ▶ Ridge regression shrinks the estimated coefficient towards zero to reduce variance

$$\min_{\beta} \|y - X\beta\|_2^2 + \underbrace{\lambda\|\beta\|_2^2}_{\text{penalty}}$$

- ▶ Solution: $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$
- ▶ Here $\lambda \geq 0$ is a regularization parameter, which is tuned to control the strength of the penalty term
- ▶ $\lambda = 0$: linear regression
- ▶ $\lambda = \infty, \hat{\beta} = 0$
- ▶ for λ in between, we balance the bias and variance of the model

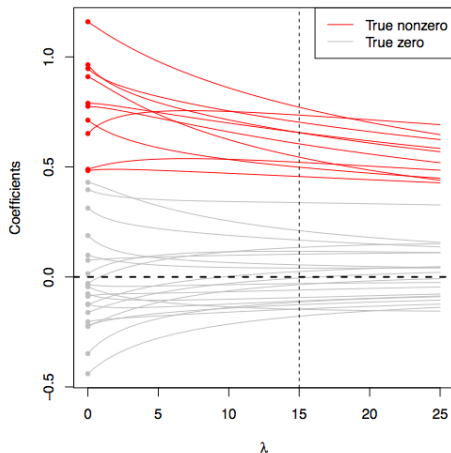
Numerical example

$$n = 50, p = 30, \sigma^2 = 1, \lambda = 25$$



Ridge regression doesn't perform model selection

- ▶ Now if we vary λ to get different ridge regression coefficients, the larger the λ the more shrunken
- ▶ Note that gray coefficient paths are not *exactly* zero, they are shrunken, but still nonzero



Lasso

The lasso estimate

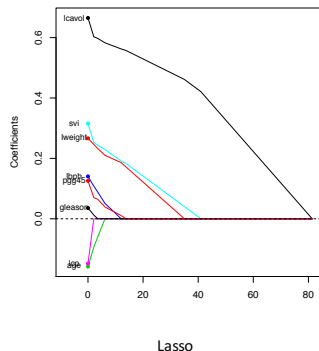
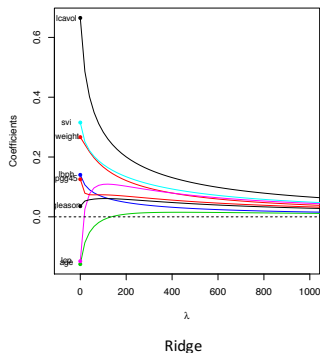
$$\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

- ▶ The penalty term $\|\beta\|_2^2$ is replaced by ℓ_1 norm $\|\beta\|_1$
- ▶ λ controls the strength of the penalty
- ▶ lasso is able to perform model selection in the linear model
- ▶ As λ increases, more coefficients are set to 0 (less variables are selected)
- ▶ Among the nonzero coefficients, more shrinkage is employed

R. Tibshirani, 1996, Regression shrinkage and selection via the lasso.

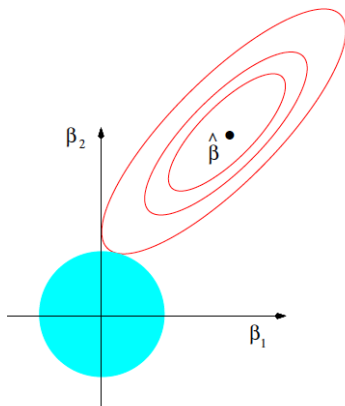
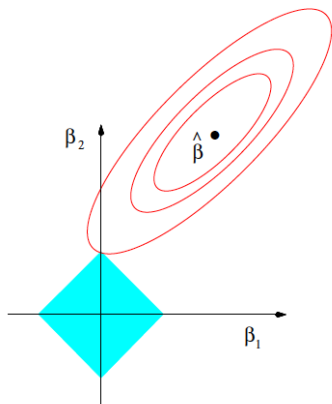
Example: prostate data

- ▶ We are interested in the level of prostate-specific antigen (PSA) elevated in men who has prostate cancer.
- ▶ Measure PSA on $n = 97$ patients, $p = 8$ clinical variables



- ▶ If we want the 3 leading factors, we report “cancer volume”, “seminal sesicle invasion”, “prostate weight”

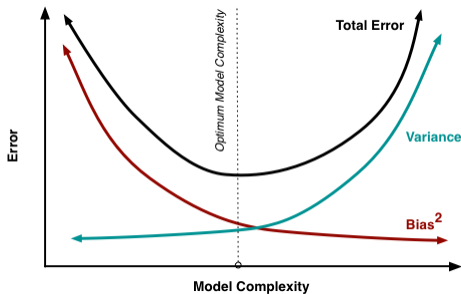
Why does lasso gives zero coefficients?



Choose a value for λ

- ▶ the value of λ controls how many variables are selected
- ▶ larger λ : more emphasis on regularization term, less emphasis on data fit
- ▶ How to choose λ to achieve a good bias-variance tradeoff?

Cross-validation



Prediction errors

- ▶ regression model $y_i = \mathbf{x}_i^T \beta + \epsilon_i$
- ▶ using **training data** (x_i, y_i) , $i = 1, \dots, n$
fit model by finding $\hat{\beta}$
- ▶ average prediction error over *another set of observations*
 $y'_i = \mathbf{x}'_i{}^T \beta + \epsilon'_i$ that are *independent* of training data

$$\text{PE}(\hat{\beta}) = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (y'_i - \mathbf{x}'_i{}^T \hat{\beta})^2 \right]$$

- ▶ prediction error is a function of λ since $\hat{\beta}$ depends on λ
- ▶ Goal: find λ to minimize $\text{PE}(\lambda)$

Test error

- ▶ When it's not easy to compute the expectation, we use *test error* or *Residual Sum of Square* (RSS) to estimate prediction error
- ▶ **test data** (\mathbf{x}'_i, y'_i)

$$\text{RSS}(\lambda) = \frac{1}{n} \sum_{i=1}^n (y'_i - \mathbf{x}'_i{}^T \beta)^2$$

- ▶ we do not really have “test” data when algorithm is developed
- ▶ idea: use part of training data for training, the remaining training data for estimating testing error, called **cross-validation**

K -fold cross validation

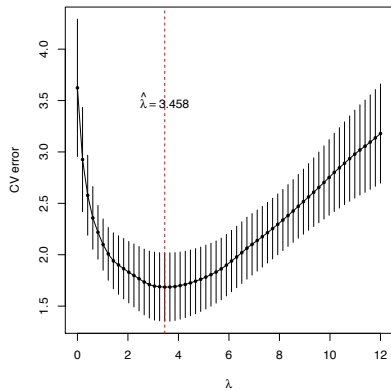
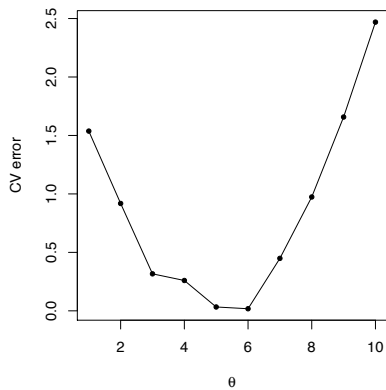
- ▶ For a number K , split the training samples into K batches (commonly $K = 5$ or $K = 10$)
- ▶ training on all but the k th part, and then validating on the k th part, iterating over $k = 1, \dots, K$

$$\text{RSS}(\lambda) = \frac{1}{K} \sum_{i=1}^K \text{RSS}_i(\lambda)$$

- ▶ When $K = n$, called **leave-one-out** cross-validation, because we leave out one data point at a time

1	2	3	4	5
Train	Train	Validation	Train	Train

CV error



Leave-one-out short cut

- ▶ for leave-one-out CV, test error can be calculated in close form (saves computation, without having to average CV errors)

$$\text{RSS}(\lambda) = \frac{1}{n} \sum_{i=1}^n \left[\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{1 - S_{ii}} \right]^2$$

$$S = X(X^T X + \lambda I)^{-1} X^T$$

Cross-validation alternatives

Cross-validation is a highly popular tool, but it is not the only way to choose λ . There are other ways

- ▶ Information criterion like AIC, BIC
- ▶ SURE (Stein's Unbiased Risk Estimate)
- ▶ Theoretically-guided choices (problem specific)

RSS is biased

- ▶ one can show that the training error is a downward-biased estimate of PE, and the bias is

$$\mathbb{E}(\text{RSS}) - \text{PE} = -2 \sum_{i=1}^n \text{cov}(\hat{y}_i, y_i)$$

- ▶ for linear estimation $\hat{y}_i = Ay_i$ for some A

$$\mathbb{E}(\text{RSS}) - \text{PE} = -2\sigma^2 \text{Tr}(A)$$

i.e.

$$\text{RSS} + 2\sigma^2 \text{Tr}(A)$$

is an unbiased estimator of the prediction risk.

Cp statistic

- ▶ Given some model $S \in \{1, \dots, p\}$

$$\hat{y}_i = Ay_i, \quad A = X_S(X_S^T X_S)^{-1} X_S^T$$

where X_S : the submatrix of X only with the columns corresponding to the selected variables

$$\begin{aligned} \text{Tr}(A) &= \text{Tr}(X_S(X_S^T X_S)^{-1} X_S^T) \\ &= \text{Tr}((X_S^T X_S)^{-1} X_S^T X_S) = \text{Tr}(I_S) = |S| \end{aligned}$$

- ▶ Unbiased estimate of the prediction error: C_p statistic

$$\underbrace{\text{RSS}}_{\text{data fit error}} + \underbrace{2|S|\hat{\sigma}^2}_{\text{model complexity}}$$

- ▶ C_p statistic has an equivalent form: **Akaike Information Criterion (AIC)**

$$C_p = \frac{RSS}{\hat{\sigma}^2} - n + 2|S|$$

- ▶ **Bayesian Information Criterion (BIC)**

$$C_p = \frac{RSS}{\hat{\sigma}^2} - n + |S| \log n$$