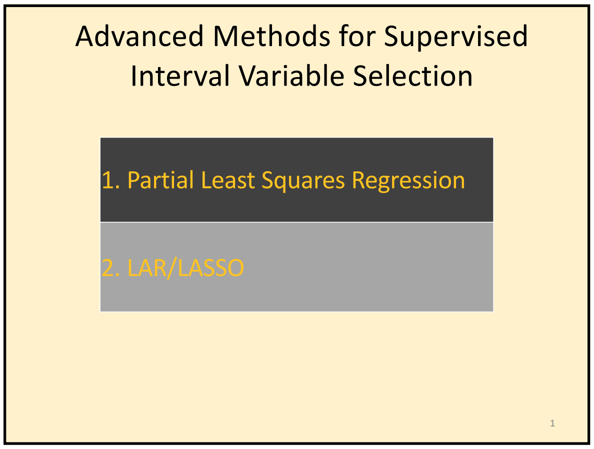
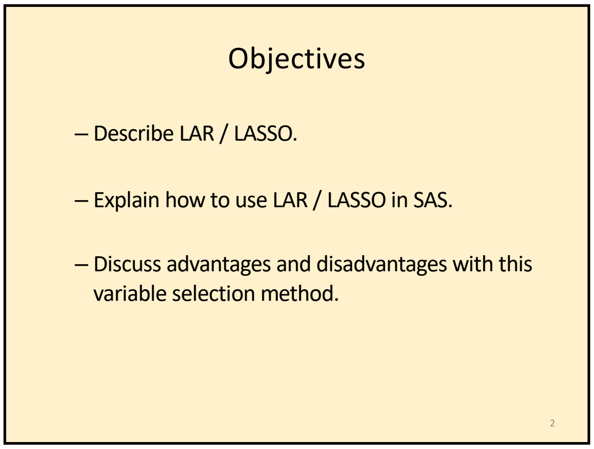


Advanced Methods for Supervised Interval Variable Selection

1. Partial Least Squares Regression
2. LAR/LASSO

1



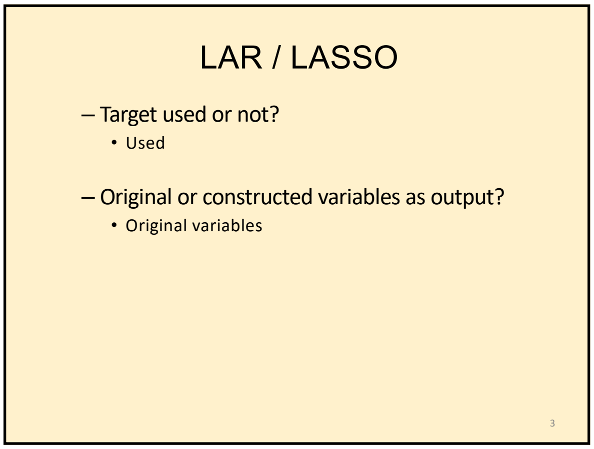
Objectives

– Describe LAR / LASSO.

– Explain how to use LAR / LASSO in SAS.

– Discuss advantages and disadvantages with this variable selection method.

2

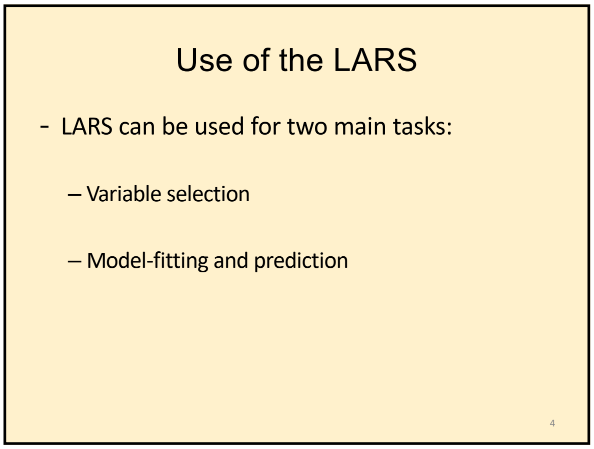


LAR / LASSO

– Target used or not?
  - Used

– Original or constructed variables as output?
  - Original variables

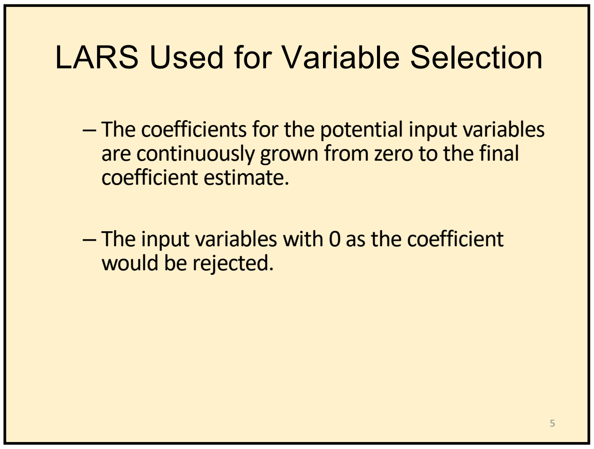
3



Use of the LARS

- LARS can be used for two main tasks:
  - Variable selection
  - Model-fitting and prediction

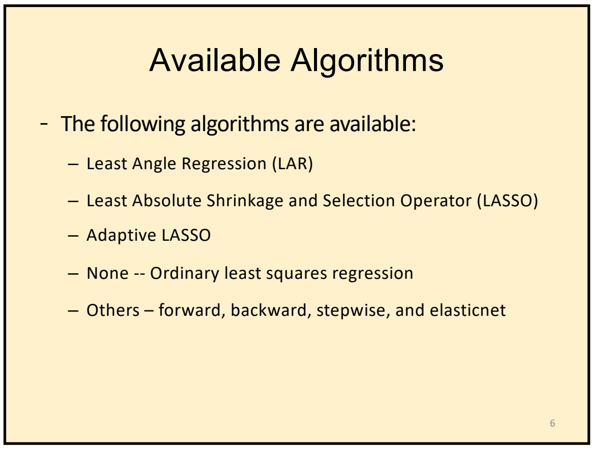
4



LARS Used for Variable Selection

- The coefficients for the potential input variables are continuously grown from zero to the final coefficient estimate.
- The input variables with 0 as the coefficient would be rejected.

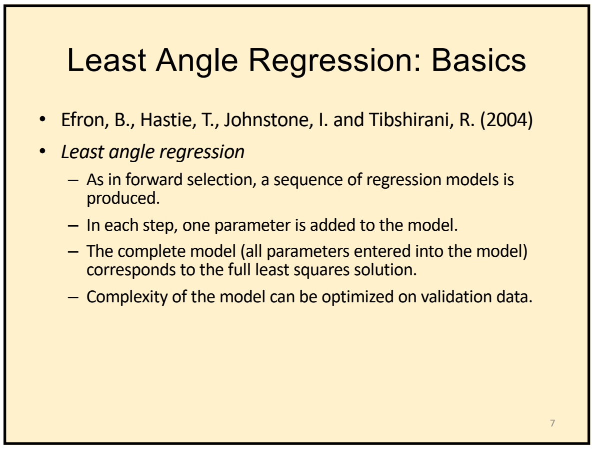
5



Available Algorithms

- The following algorithms are available:
  - Least Angle Regression (LAR)
  - Least Absolute Shrinkage and Selection Operator (LASSO)
  - Adaptive LASSO
  - None -- Ordinary least squares regression
  - Others – forward, backward, stepwise, and elasticnet

6

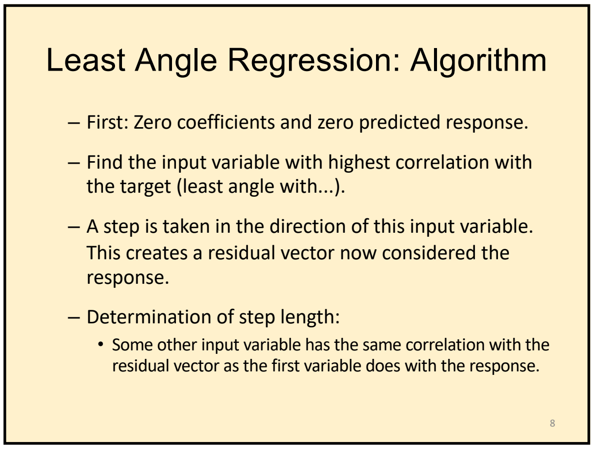


Least Angle Regression: Basics

- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004)
- *Least angle regression*
  - As in forward selection, a sequence of regression models is produced.
  - In each step, one parameter is added to the model.
  - The complete model (all parameters entered into the model) corresponds to the full least squares solution.
  - Complexity of the model can be optimized on validation data.

7

7

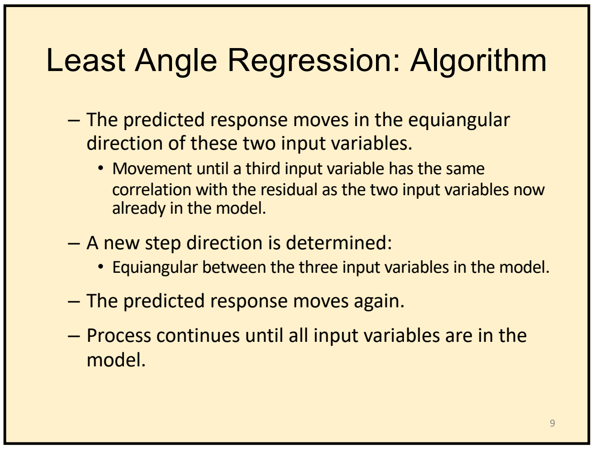


Least Angle Regression: Algorithm

- First: Zero coefficients and zero predicted response.
- Find the input variable with highest correlation with the target (least angle with...).
- A step is taken in the direction of this input variable. This creates a residual vector now considered the response.
- Determination of step length:
  - Some other input variable has the same correlation with the residual vector as the first variable does with the response.

8

8



Least Angle Regression: Algorithm

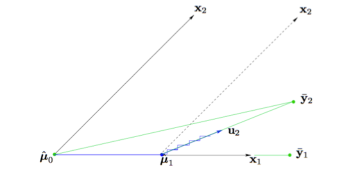
- The predicted response moves in the equiangular direction of these two input variables.
  - Movement until a third input variable has the same correlation with the residual as the two input variables now already in the model.
- A new step direction is determined:
  - Equiangular between the three input variables in the model.
- The predicted response moves again.
- Process continues until all input variables are in the model.

9

9

Least Angle Regression Geometrically

- Two covariates (predictors) x_1 and x_2 and the space $L(x_1, x_2)$ that is spanned by them
- $\mu = E(Y|X)$



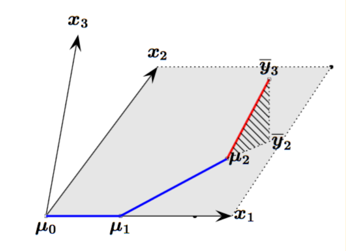
The LAR direction $\mathbf{u}_2$ at step 2 makes an equal angle with $\mathbf{x}_1$ and $\mathbf{x}_2$.

10

10

Least Angle Regression Geometrically

- Three covariates (predictors) x_1, x_2 and x_3 and the space $L(x_1, x_2, x_3)$ that is spanned by them



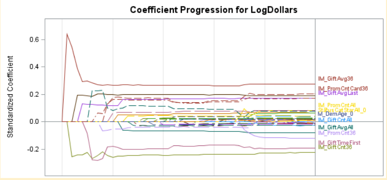
11

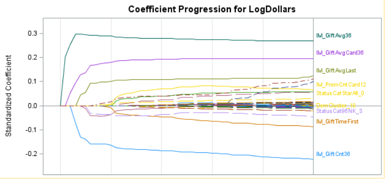
11

Coefficient Traces

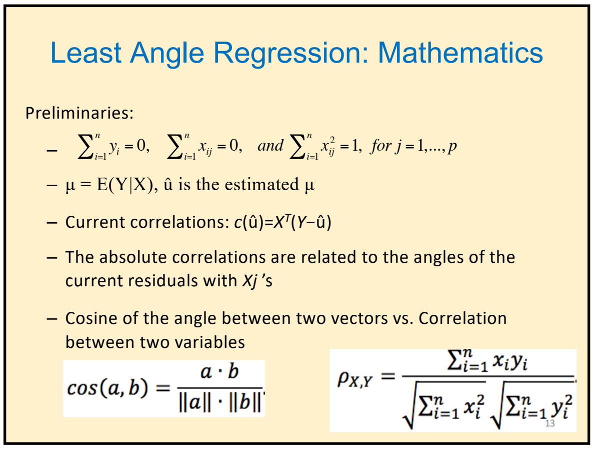
For 50 steps,

Forward Selection →





12



Least Angle Regression: Mathematics

Preliminaries:

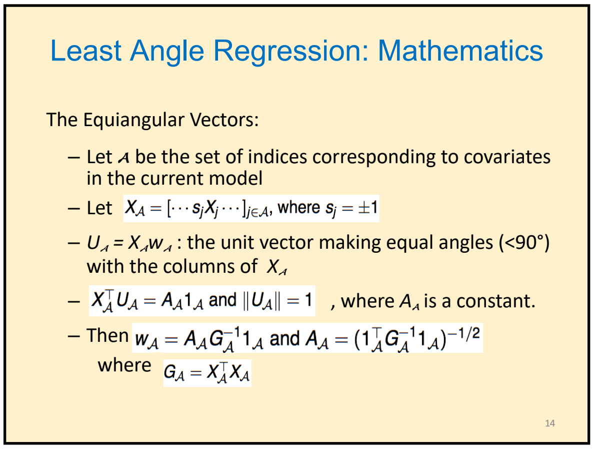
- $\sum_{i=1}^{n} y_i = 0, \quad \sum_{i=1}^{n} x_{ij} = 0, \quad and \sum_{i=1}^{n} x_{ij}^2 = 1, \; for\ j = 1, \ldots, p$
- μ = E(Y|X), û is the estimated μ
- Current correlations: $c(\hat{u}) = X^T(Y - \hat{u})$
- The absolute correlations are related to the angles of the current residuals with Xj 's
- Cosine of the angle between two vectors vs. Correlation between two variables

$$cos(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

13

13



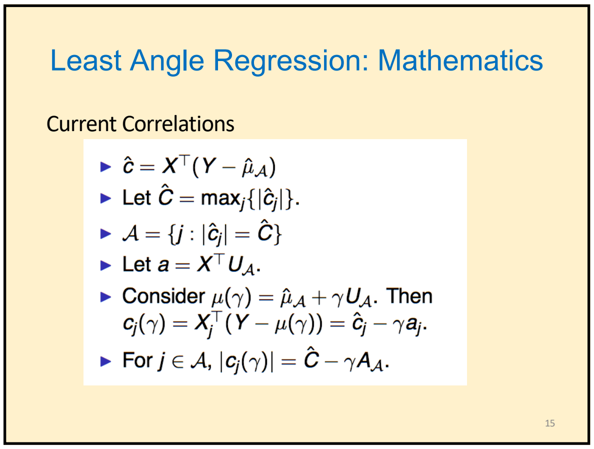
Least Angle Regression: Mathematics

The Equiangular Vectors:

- Let $\mathcal{A}$ be the set of indices corresponding to covariates in the current model
- Let $X_{\mathcal{A}} = [\cdots s_j X_j \cdots]_{j \in \mathcal{A}}$, where $s_j = \pm 1$
- $U_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}$: the unit vector making equal angles (<90°) with the columns of $X_{\mathcal{A}}$
- $X_{\mathcal{A}}^{\top} U_{\mathcal{A}} = A_{\mathcal{A}} 1_{\mathcal{A}}$ and $\|U_{\mathcal{A}}\| = 1$, where $A_{\mathcal{A}}$ is a constant.
- Then $w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} 1_{\mathcal{A}}$ and $A_{\mathcal{A}} = (1_{\mathcal{A}}^{\top} G_{\mathcal{A}}^{-1} 1_{\mathcal{A}})^{-1/2}$ where $G_{\mathcal{A}} = X_{\mathcal{A}}^{\top} X_{\mathcal{A}}$

14

14



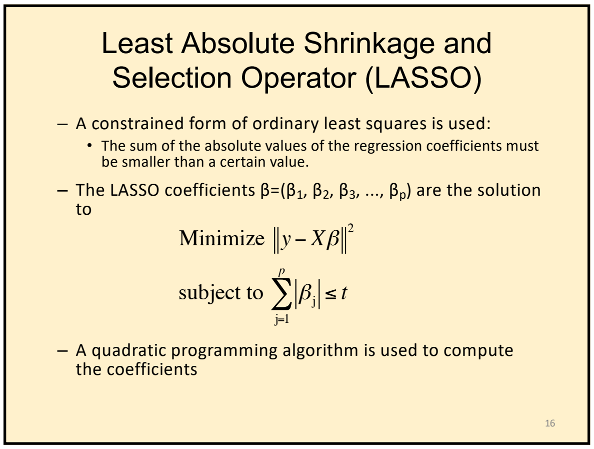
Least Angle Regression: Mathematics

Current Correlations

- ► $\hat{c} = X^{\top}(Y - \hat{\mu}_{\mathcal{A}})$
- ► Let $\hat{C} = \max_j\{|\hat{c}_j|\}$.
- ► $\mathcal{A} = \{j : |\hat{c}_j| = \hat{C}\}$
- ► Let $a = X^{\top} U_{\mathcal{A}}$.
- ► Consider $\mu(\gamma) = \hat{\mu}_{\mathcal{A}} + \gamma U_{\mathcal{A}}$. Then $c_j(\gamma) = X_j^{\top}(Y - \mu(\gamma)) = \hat{c}_j - \gamma a_j$.
- ► For $j \in \mathcal{A}$, $|c_j(\gamma)| = \hat{C} - \gamma A_{\mathcal{A}}$.

15

15



Least Absolute Shrinkage and Selection Operator (LASSO)

- A constrained form of ordinary least squares is used:
  - The sum of the absolute values of the regression coefficients must be smaller than a certain value.
- The LASSO coefficients β=(β_1, β_2, β_3, ..., β_p) are the solution to

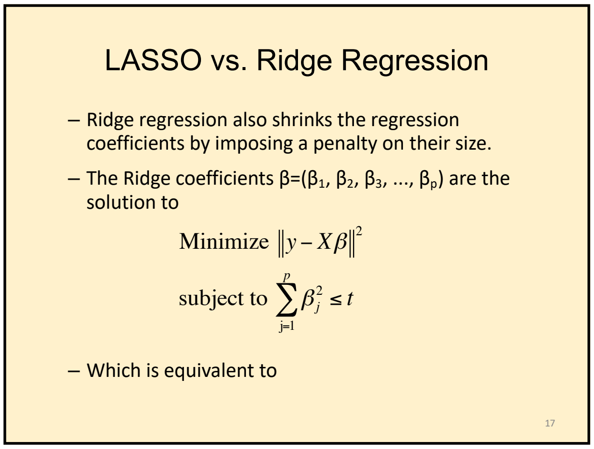
$$\text{Minimize } \|y - X\beta\|^2$$

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \le t$$

- A quadratic programming algorithm is used to compute the coefficients

16

16



LASSO vs. Ridge Regression

- Ridge regression also shrinks the regression coefficients by imposing a penalty on their size.
- The Ridge coefficients β=(β_1, β_2, β_3, ..., β_p) are the solution to

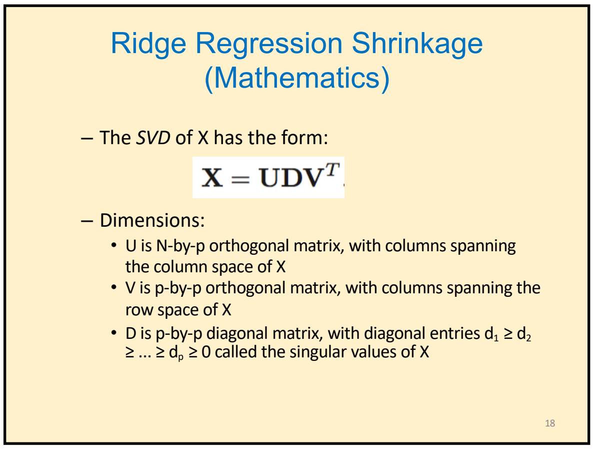
$$\text{Minimize } \|y - X\beta\|^2$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t$$

- Which is equivalent to

17

17



Ridge Regression Shrinkage (Mathematics)

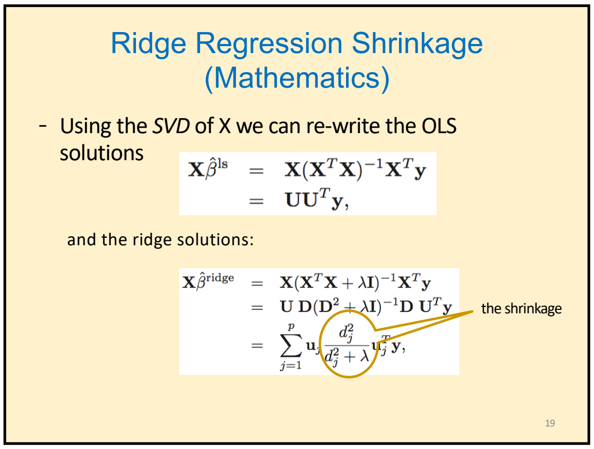
- The *SVD* of X has the form:

$$\mathbf{X} = \mathbf{UDV}^T.$$

- Dimensions:
  - U is N-by-p orthogonal matrix, with columns spanning the column space of X
  - V is p-by-p orthogonal matrix, with columns spanning the row space of X
  - D is p-by-p diagonal matrix, with diagonal entries $d_1 \ge d_2 \ge \ldots \ge d_p \ge 0$ called the singular values of X

18

18



Ridge Regression Shrinkage (Mathematics)

- Using the *SVD* of X we can re-write the OLS solutions

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ls}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \mathbf{U}\mathbf{U}^T\mathbf{y},\end{aligned}$$

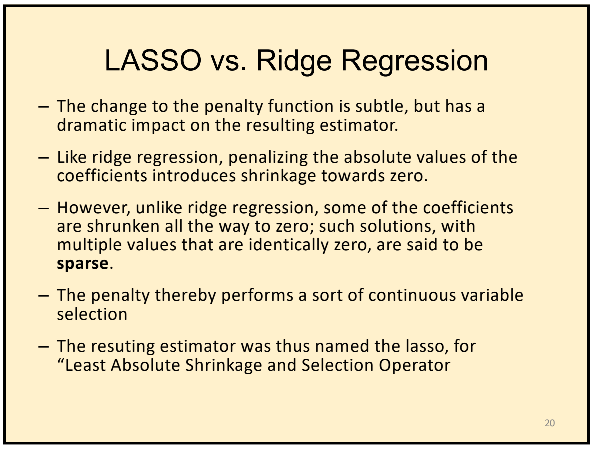
and the ridge solutions:

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\\ &= \mathbf{U}\,\mathbf{D}(\mathbf{D}^2+\lambda\mathbf{I})^{-1}\mathbf{D}\,\mathbf{U}^T\mathbf{y}\\ &= \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y},\end{aligned}$$

the shrinkage

19

19

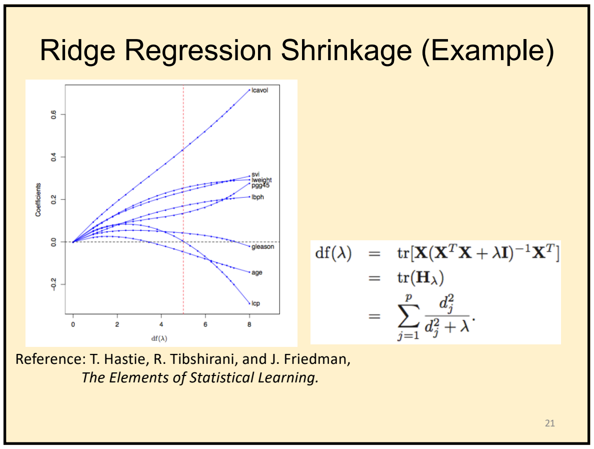


LASSO vs. Ridge Regression

- The change to the penalty function is subtle, but has a dramatic impact on the resulting estimator.
- Like ridge regression, penalizing the absolute values of the coefficients introduces shrinkage towards zero.
- However, unlike ridge regression, some of the coefficients are shrunken all the way to zero; such solutions, with multiple values that are identically zero, are said to be **sparse**.
- The penalty thereby performs a sort of continuous variable selection
- The resuting estimator was thus named the lasso, for "Least Absolute Shrinkage and Selection Operator

20

20



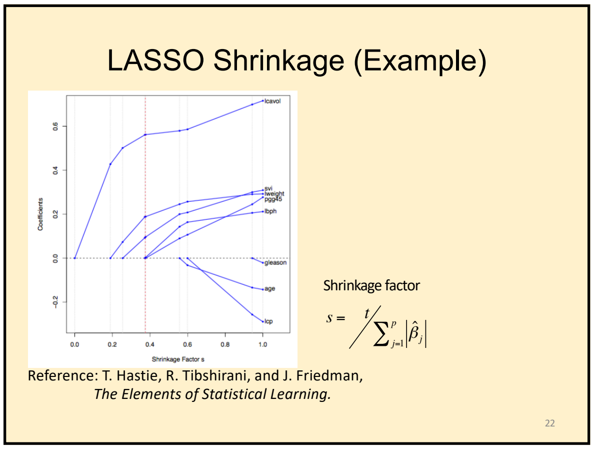
Ridge Regression Shrinkage (Example)

$$\begin{aligned}\text{df}(\lambda) &= \text{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^T]\\ &= \text{tr}(\mathbf{H}_\lambda)\\ &= \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}.\end{aligned}$$

Reference: T. Hastie, R. Tibshirani, and J. Friedman,
The Elements of Statistical Learning.

21

21



LASSO Shrinkage (Example)

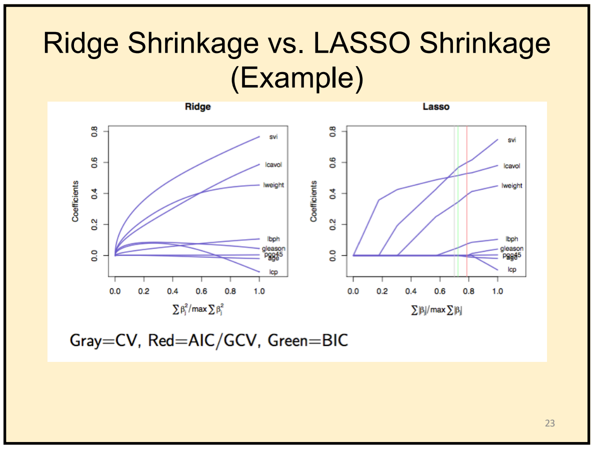
Shrinkage factor

$$s = t \Big/ \sum_{j=1}^{p} \left|\hat{\beta}_j\right|$$

Reference: T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.*

22

22

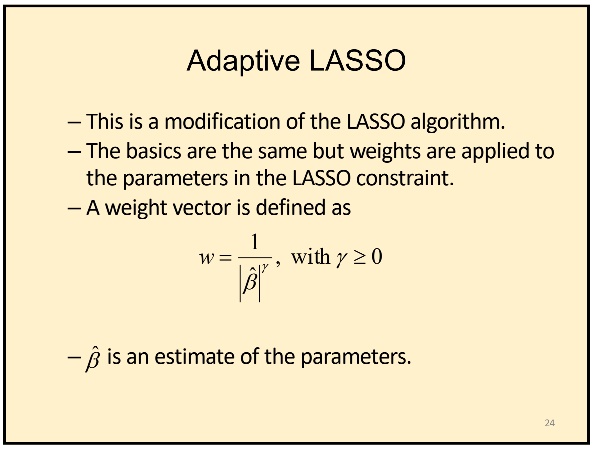


Ridge Shrinkage vs. LASSO Shrinkage (Example)

Gray=CV, Red=AIC/GCV, Green=BIC

23

23



Adaptive LASSO

- This is a modification of the LASSO algorithm.
- The basics are the same but weights are applied to the parameters in the LASSO constraint.
- A weight vector is defined as

$$w = \frac{1}{\left|\hat{\beta}\right|^{\gamma}}, \text{ with } \gamma \geq 0$$

- $\hat{\beta}$ is an estimate of the parameters.

24

24

Adaptive LASSO

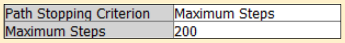
– The adaptive LASSO coefficients β=(β$_1$, β$_2$, β$_3$, …, β$_p$) are the solution to

$$\text{Minimize } \|y - X\beta\|^2$$

$$\text{subject to } \sum_{j=1}^{p} |w_j \beta_j| \le t$$

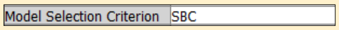
Model Selection

- Two issues are important in the model selection:
  – How many steps should be processed?



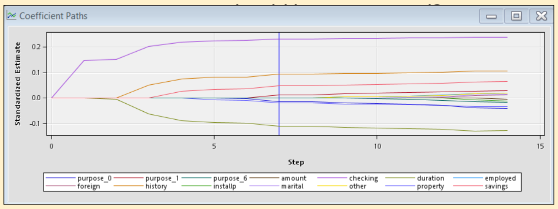
Path Stopping Criterion	Maximum Steps
Maximum Steps	200

  – From which of the steps should the model be selected?



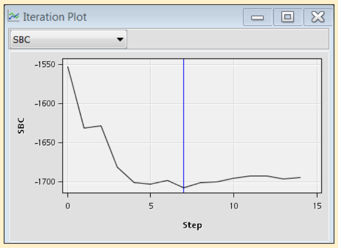
Model Selection Criterion	SBC

How Many Steps Should Be Processed?



From Which of the Steps Should the Model Be Selected?

– How can we select the best model from the sequence of constructed models?



28

From Which of the Steps Should the Model Be Selected?

– The available model selection criteria include
  - SBC – Schwarts Bayesian information criterion
  - BIC – Bayesian information criterion
  - AIC – Akaike information criterion
  - AICC -- Corrected Akaike's information criterion
  - CP – Cp statistic, explained in Regression class
  - Validation – explained in DM I class
  - Cross Validation – explained in PLS slides
  - ADJRSQ, CVEX, PRESS

29

Performance Assessment: Loss Function

- Typical choices for quantitative response *Y*:

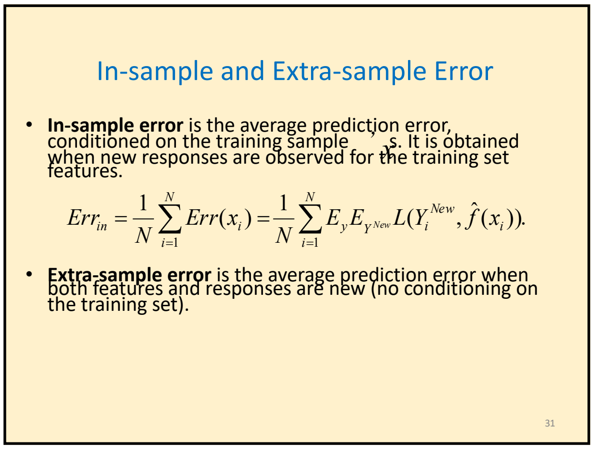
$$L(Y,\hat{f}(X)) = \begin{cases} (Y-\hat{f}(X))^2 & \text{(squared error)} \\ |Y-\hat{f}(X)| & \text{(absolute error)} \end{cases}$$

- Typical choices for categorical response *G*:

$$L(G,\hat{G}(X)) = I(G \neq \hat{G}(X)) \qquad \text{(0-1 loss)}$$

$$\begin{aligned} L(G,\hat{p}(X)) &= -2\sum_{k=1}^{K} I(G=k)\log \hat{p}_k \\ &= -2\log \hat{p}_G(X) \qquad \text{(log-likelihood)} \end{aligned}$$

30



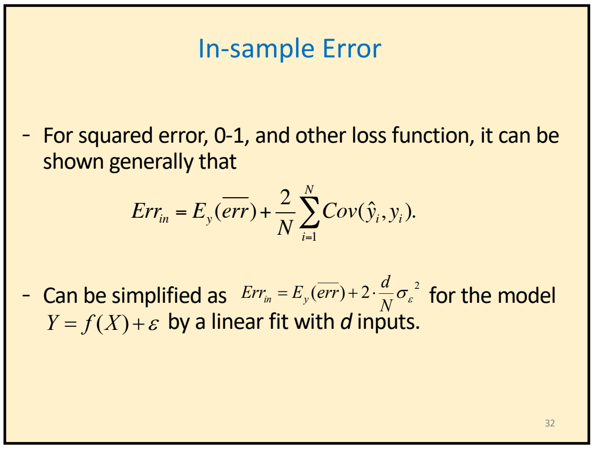
In-sample and Extra-sample Error

- **In-sample error** is the average prediction error, conditioned on the training sample y_i's. It is obtained when new responses are observed for the training set features.

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \frac{1}{N}\sum_{i=1}^{N} E_y E_{Y^{New}} L(Y_i^{New}, \hat{f}(x_i)).$$

- **Extra-sample error** is the average prediction error when both features and responses are new (no conditioning on the training set).

31



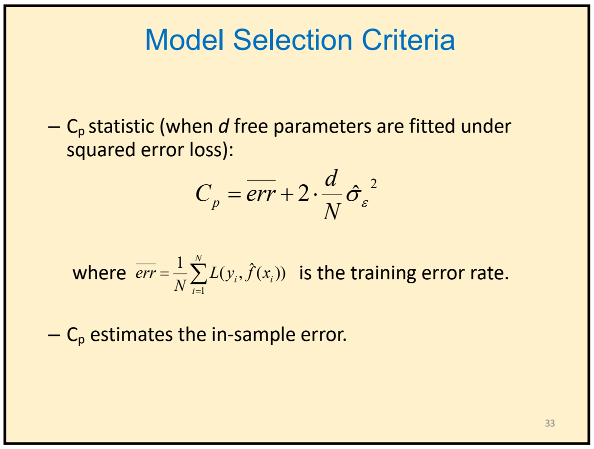
In-sample Error

- For squared error, 0-1, and other loss function, it can be shown generally that

$$Err_{in} = E_y(\overline{err}) + \frac{2}{N}\sum_{i=1}^{N} Cov(\hat{y}_i, y_i).$$

- Can be simplified as $Err_{in} = E_y(\overline{err}) + 2\cdot\frac{d}{N}\sigma_\varepsilon^{2}$ for the model $Y = f(X) + \varepsilon$ by a linear fit with *d* inputs.

32



Model Selection Criteria

– C_p statistic (when *d* free parameters are fitted under squared error loss):

$$C_p = \overline{err} + 2\cdot\frac{d}{N}\hat{\sigma}_\varepsilon^{2}$$

where $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$ is the training error rate.

– C_p estimates the in-sample error.

33

Model Selection Criteria

- AIC (Akaike information criterion), a more generally applicable estimate of the in-sample error when a log-likelihood loss function is used (when $N \to \infty$):

$$-2E[\log\Pr_{\hat{\theta}}(Y)] \approx -\frac{2}{N}E(\sum_{i=1}^{N}\log\Pr_{\hat{\theta}}(y_i)) + 2\frac{d}{N}$$
$$= -\frac{2}{N}\cdot\log lik + 2\frac{d}{N} = \frac{AIC}{N}$$

- General form: $AIC = -2\cdot(\log lik) + 2\cdot d.$
- Choose the model giving smallest AIC over the set of models considered

34

34

Model Selection Criteria

- BIC (Bayesian information criterion) is motivated from Bayesian point of view.
- BIC tends to penalize complex models more heavily, giving preference to simpler models in selection
- Its general form is:

$$BIC = -2\cdot(\log lik) + (\log N)\cdot d.$$

35

35

Bayesian Model Selection

- Suppose we have candidate models $M_m, m = 1,\ldots,M$
- with corresponding model parameters θ_m.
- Prior distribution: $\Pr(\theta_m \mid M_m), m = 1,\ldots,M.$
- Posterior probability:

$$\Pr(M_m \mid Z) \propto \Pr(M_m)\cdot\Pr(Z \mid M_m).$$

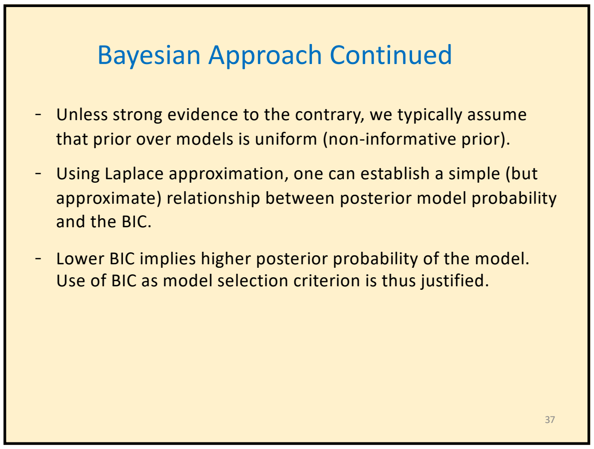
- Compare two models via posterior odds:

$$\frac{\Pr(M_m \mid Z)}{\Pr(M_l \mid Z)} = \frac{\Pr(M_m)}{\Pr(M_l)}\cdot\frac{\Pr(Z \mid M_m)}{\Pr(Z \mid M_l)}$$

- The second factor on the RHS is called the Bayes factor and describes the contribution of the data towards posterior odds.

36

36

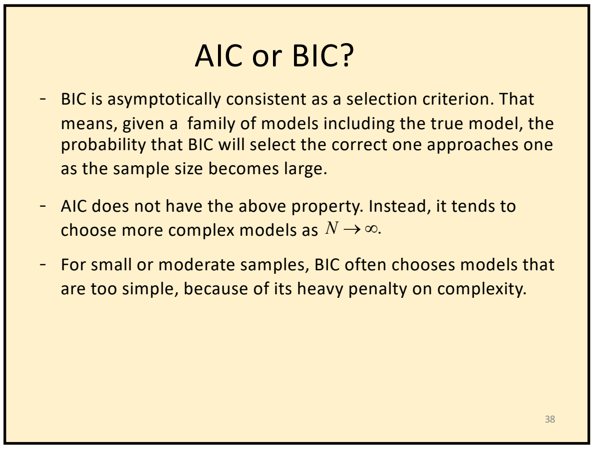


Bayesian Approach Continued

- Unless strong evidence to the contrary, we typically assume that prior over models is uniform (non-informative prior).
- Using Laplace approximation, one can establish a simple (but approximate) relationship between posterior model probability and the BIC.
- Lower BIC implies higher posterior probability of the model. Use of BIC as model selection criterion is thus justified.

37

37

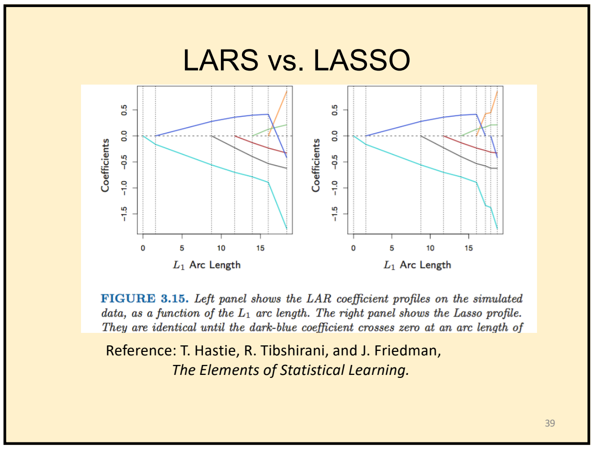


AIC or BIC?

- BIC is asymptotically consistent as a selection criterion. That means, given a family of models including the true model, the probability that BIC will select the correct one approaches one as the sample size becomes large.
- AIC does not have the above property. Instead, it tends to choose more complex models as $N \to \infty$.
- For small or moderate samples, BIC often chooses models that are too simple, because of its heavy penalty on complexity.

38

38



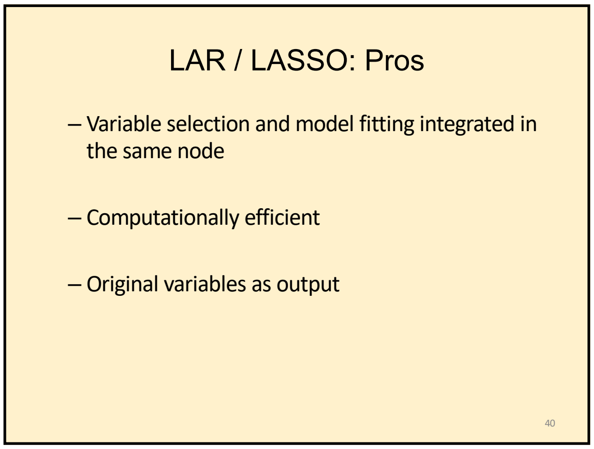
LARS vs. LASSO

FIGURE 3.15. *Left panel shows the LAR coefficient profiles on the simulated data, as a function of the L_1 arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of*

Reference: T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.*

39

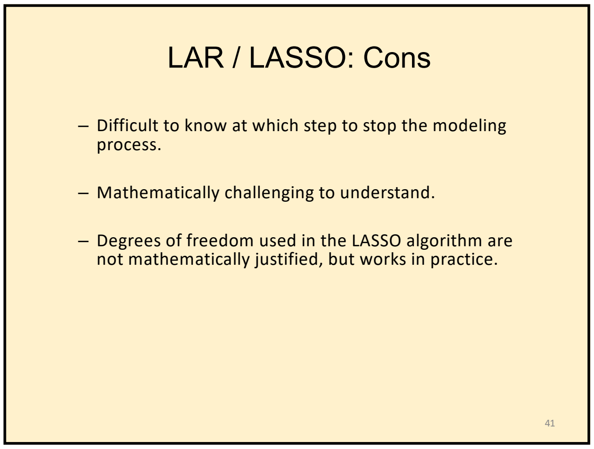
39



LAR / LASSO: Pros

- Variable selection and model fitting integrated in the same node
- Computationally efficient
- Original variables as output

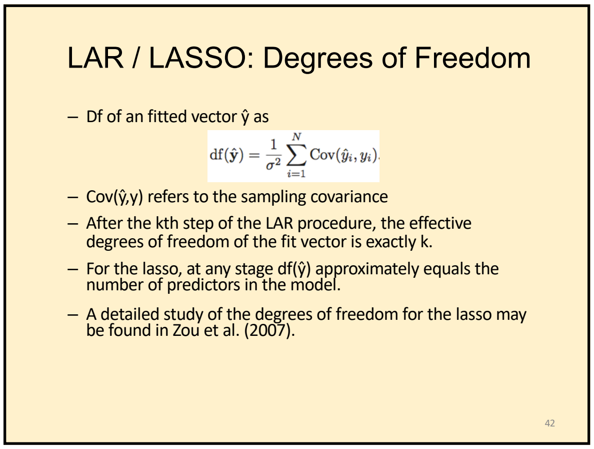
40



LAR / LASSO: Cons

- Difficult to know at which step to stop the modeling process.
- Mathematically challenging to understand.
- Degrees of freedom used in the LASSO algorithm are not mathematically justified, but works in practice.

41



LAR / LASSO: Degrees of Freedom

- Df of an fitted vector ŷ as

$$\mathrm{df}(\hat{\mathbf{y}}) = \frac{1}{\sigma^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i).$$

- Cov(ŷ,y) refers to the sampling covariance
- After the kth step of the LAR procedure, the effective degrees of freedom of the fit vector is exactly k.
- For the lasso, at any stage df(ŷ) approximately equals the number of predictors in the model.
- A detailed study of the degrees of freedom for the lasso may be found in Zou et al. (2007).

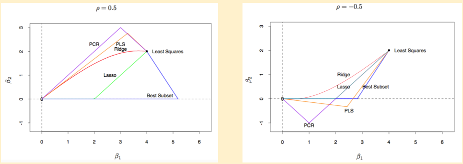
42

LAR / LASSO for Variable Selection

- This demonstration illustrates how to use the LARS to perform variable selection.
- Variations of the LARS node/procedure are shown.

A Comparison of the Selection and Shrinkage

– Consider a simulated example with two correlated inputs X1 and X2, with correlation ρ. We assume that the true regression coefficients are $\beta_1 = 4$ and $\beta_2 = 2$.



– A full study: Frank and Friedman (1993). These authors conclude that for minimizing the prediction error, ridge regression is generally preferable to variable subset selection, PC regression, and PLS

Questions?