Variable Selection

Jeff Goldsmith
Department of Biostatistics

August 23, 2018

Overview

- Linear regression
- Variable selection
- Penalties
- Grouped penalties
- Tuning parameters
- Caveats

Multiple linear regression model

- Observe data $(y_i, x_{i1}, \ldots, x_{ip})$ for subjects $1, \ldots, n$. Want to estimate $\beta_0, \beta_1, \ldots, \beta_p$ in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_1 x_{ip} + \epsilon_i; \ \epsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

- Assumptions: residuals have mean zero, constant variance, and are independent
- Estimate parameters using OLS

(This covers a lot of ground – general goodness of linear models, interpretation, inference, unbiasedness, diagnostics, ...)

Multiple linear regression

- Let

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Then we can write the model in a more compact form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Matrix notation

- Matrix notation provides a compact way to discuss regression models and related areas
- $\boldsymbol{X}$ is called the *design matrix*
- In settings where there are many predictors, specifying the design matrix is typically easier than a “formula interface”

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

OLS Estimation

- OLS estimate found by minimizing the RSS:

$$\begin{aligned}
\hat{\beta}_{OLS} &= \arg\min_{\boldsymbol{\beta}} [RSS(\boldsymbol{\beta})] \\
&= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right] \\
(&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y})
\end{aligned}$$

Variable selection

- Often we need to do *variable selection*
  - Sometimes $p > n$
  - Sometimes we have many variables and want to identify the "important" ones
  - Sometimes we have many variables and want to remove "unimportant" ones

Methods for variable selection

- Subset selection
  - More traditional methods
- Penalization / shrinkage
  - More recent methods

Subset selection

- Forward / backward selection
- Best subset selection

- Won’t speak for everyone, but I don’t like these ...
  - Often not feasible
  - Can be unstable
  - Overall uncertainty is hard to assess, so inference is suspect

Penalized regression

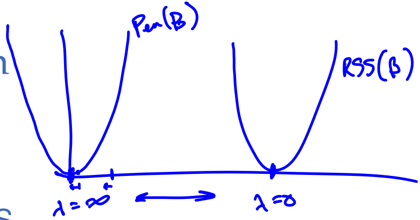
- Trade some bias for lower variance and overall MSE
  - Can outperform OLS in some important ways
- Rather than a subset selection approach, all parameters stay in the model but we restrict their effect
- Penalize the size of the coefficients – "unimportant" variables will have their coefficients shrunk towards to zero

Ridge regression

Start with RSS and add a penalty:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_R &= \arg\min_{\boldsymbol{\beta}} \left[RSS(\boldsymbol{\beta}) + \lambda Pen(\boldsymbol{\beta}) \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda Pen(\boldsymbol{\beta}) \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right]
\end{aligned}
$$

Notes on ridge regression



- Not unbiased
- Lower variance than OLS
- Avoid subset selection; add tuning parameter selection
  - For "small" values of λ, $\hat{\beta}_R \approx \hat{\beta}_{OLS}$
  - For "large" values of λ, $\hat{\beta}_R \approx 0$
- Results in coefficients that are small but non-zero
- Doesn't have to penalize everything
- (Has a closed-form solution)

Standardizing predictors

- OLS estimates are scale equivariant: multiplying a predictor by a constant rescales the coefficient but leaves the model (and the fitted values) unchanged
- In contrast, ridge regression coefficient estimates can change substantially when multiplying a given predictor by a constant
  - Rescaling a predictor changes the impact in of the coefficient in the penalty
- Therefore, it is best to apply ridge regression (and other penalized methods) after standardizing the predictors, using the formula

Tuning parameters

Before considering other penalization methods, a brief digression ...

- Goal is to trade some increase in the bias of $\hat{\beta}$ for a decrease in the variance of $\hat{\beta}$
- Mean squared error formalizes this:

$$
\begin{aligned}
MSE(\hat{\beta}) &= E\left[\left(\hat{\beta} - \beta\right)^2\right] \\
&= E\left[\left(\hat{\beta} - E(\hat{\beta})\right)^2\right] + \left(E(\hat{\beta}) - \beta\right)^2 \\
&= \text{variance}(\hat{\beta}) + \text{bias}^2(\hat{\beta})
\end{aligned}
$$

MSE for predictions

MSE for β isn't feasible in practice

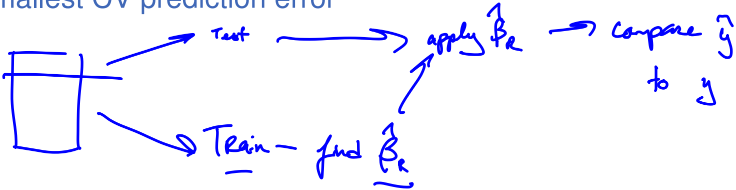
- MSE for predictions is easier, and incorporates dependence on $\hat{\beta}$ through the fitted values

$$MSE(\hat{y}) \quad = \quad E\left[(\hat{y} - y)^2\right]$$

- Can evaluate this using cross-validation

Cross validation

- Set aside some subset of the data as a “test” data set
- Use remaining data to “train” the model (i.e. estimate parameters $\hat{\beta}_R$)
- Compute $E(\hat{y}_{i,\lambda} - y_i)^2$ across all i in the “test” set
- Evaluate many values of λ, and choose the one with the smallest CV prediction error

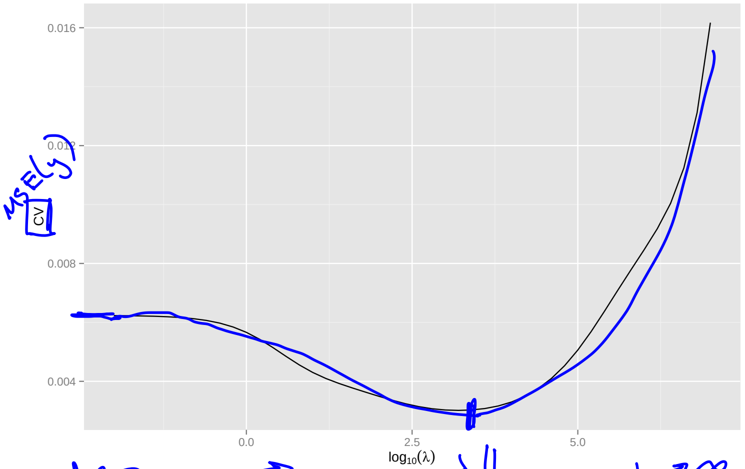


Over and under fitting

Over- and under-fitting are common problems

- Over-fitting means you're fitting the current (training) data too well, and will make bad predictions for future (test) data
- Under-fitting means you're not fitting the current data well enough, and will make bad predictions for future (test) data
- Over-fitting is a problem of high variance; under-fitting is a problem of high bias

Cross validation

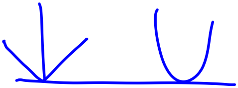


$\lambda \approx 0$ ⟶ λ!! $\lambda \to \infty$

low bias
high var

high bias
low var

Lasso penalization



- Lasso (least absolute shrinkage and selection operator) is a more recent penalized regression estimator
- Basic form is similar to that of ridge regression, but penalty function is different:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_L &= \arg\min_{\boldsymbol{\beta}} \left[RSS(\boldsymbol{\beta}) + \lambda \,||\boldsymbol{\beta}||_1 \right] \\
&= \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]
\end{aligned}$$

- Quite popular – broadly used, many adaptations

Lasso penalization

Some properties of Lasso penalties

- No closed form solution (although there are some computationally useful tricks)
- The different penalty form means Lasso has a tendency to shrink coefficients *all the way* to zero
- Can be useful as an "automated" variable selection approach
- Still have to choose λ; cross validation is a popular tool for this

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Lasso vs ridge – ”the picture”

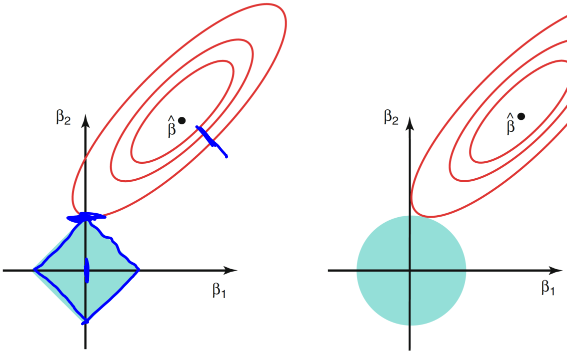


Figure: ISL 6.7, Page 222

The lasso performs ℓ_1 shrinkage, and there are “corners” in the constraint. If the RSS “hits” one of these corners, the coefficient corresponding to the axis is shrunk to zero.

Lasso vs ridge – "the other picture"

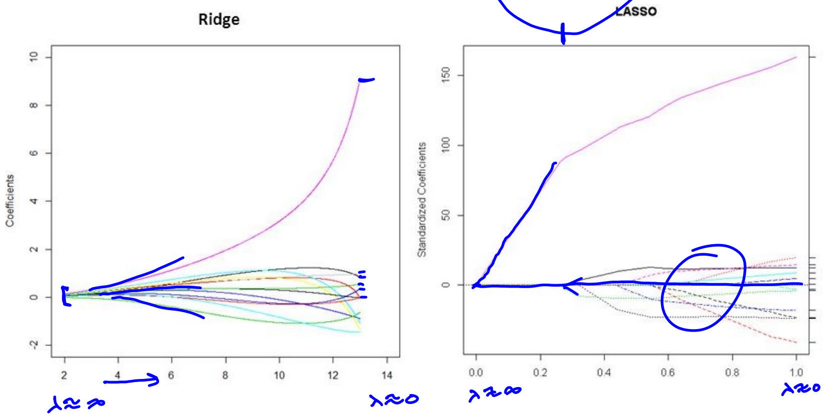


Figure: ISL 6.7, Page 222

Comments on lasso

... especially with respect to MLR:

- Emphasis is on *prediction*, not inference
- Coefficients for selected variables is *not the same* as an MLR including only that subset
- When predictors are correlated, lasso tends to select one element of a group

Bias-reducing penalties

Penalized regression framework is pretty broad:

$$\arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda Pen(\boldsymbol{\beta}) \right]$$

In addition to lasso and ridge regression, other popular alternatives include

- Adaptive lasso
- MCP: minimax concave penalty
- SCAD: smoothly clipped absolute deviation

These all try to give unbiased coefficients for covariates with large effects, but operate under the same general framework

Adaptive lasso

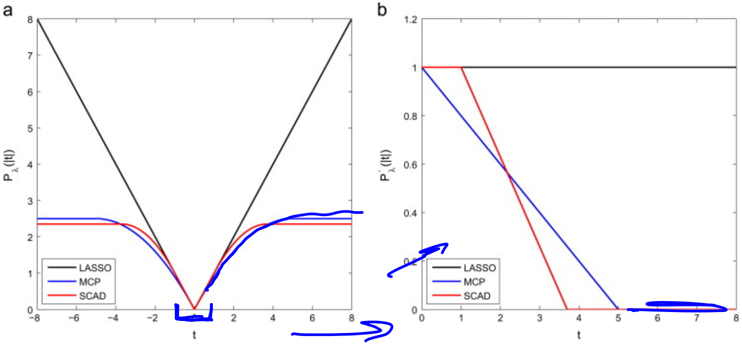
Adaptive lasso adds weights:

$$\arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right]$$

Weights (obtained via e.g. a first-round “standard” lasso) increase or decrease the effect of penalization on individual coefficients.

MCP, SCAD

SCAD and MCP use a different penalty structure:



COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH

Elastic net

- Unbiasedness is one problem with lasso; selection of a single covariate among from a correlated group is another
- The elastic net is one solution to this:

$$\arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \left((1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right) \right]$$

- Combines regularization via the ridge-type penalty with variable selection via the lasso penalty
- More effective to deal with groups of correlated predictors

Group variable selection

Elastic net often "works", but you can make grouping explicit

- Partition covariates into known groups
- Apply a relevant penalty to groups, not individual coefficients

$$\arg\min_{\boldsymbol{\beta}} \left[\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2} + \lambda \sum_{g=1}^{G} Pen(\boldsymbol{\beta}_g) \right]$$

- Group lasso, SCAD, MCP, etc exist
- How you group matters

Closing thoughts

- Inference (and, at least to some extent, interpretation) is challenging in variable selection methods
  - Emphasis is on prediction accuracy
  - Some work of post-selection inference exists, but hasn't yet been widely adopted
- Variable selection can be included in other models
  - Sparse PCA, for example
- Variable selection methods can have a similar goal as principal components regression, but use a very different approach

Other directions

- ~~Variable selection can be included in other models~~
- Other approaches to variable selection (e.g. Bayesian methods ...)
- Consistency across approaches

Sessions's big ideas

- Penalization; tuning parameters; cross validation; group penalties

---

- ISLR 6