40.016: The Analytics Edge
Week 5 Lecture 2

MODEL ASSESSMENT AND MODEL SELECTION:
CROSS VALIDATION AND LASSO

Term 5, 2022

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Announcements

- Exam time: Wednesday 22 June, 2:30 pm to 4:30 pm
- Exam venue:
  - CS01: CC12 (2.406)
  - CS02: CC13 (2.506)
- Week 6 Lecture 1: tutorial
- Week 6 Lecture 2: no class

Outline

- Model assessment and Model selection
- Bias-Variance trade-off
- Subset Selection

- Cross validation
- LASSO

Model assessment and Model selection

GOAL

- Prediction accuracy
- Model interpretability

Bias-Variance trade-off

- Recall the linear regression model fitting problem. The true model is:

$$Y = f(X) + \epsilon$$

  where ϵ is a random error term with mean 0 and variance σ^2.

- Using least squares minimization on training data we find predictor $\hat{f}$ for f.
- (X_0, Y_0): (test) data point.

$$\begin{aligned}
\text{Test MSE} = \mathbb{E}(Y_0 - \hat{f}(X_0))^2 &= \text{Var}(\hat{f}(X_0)) + \mathbb{E}\left[(f(X_0) - \hat{f}(X_0))^2\right] + \sigma^2 \\
&= \text{Variance of estimator} \\
&\quad + \text{Squared Bias} \\
&\qquad + \text{Variance of error term (irreducible errror).}
\end{aligned}$$

- Complex model: typically high variance and low bias.
- Simple model: low variance but high bias.

Subset selection

- We have n observations, and p predictor variables.
- If $n \approx p$ or $n < p$: risk of overfitting.
- Pick a selection of important explanatory variable from the p available.
- Solution space: 2^p subsets.
- BEST SUBSET: Provides the best set of variables but $O(2^p)$ computations.
- FORWARD/BACKWARD STEPWISE ALGORITHM: Provides an approximate best set of variables with $O(p^2)$ computations.

Cross-Validation

- Model assessment technique.
- A model is considered good if it has a low *test set error (TEST MSE)* .
- We often do not have a large test set to validate our model.
- One method of model assessment is *Cross Validation*.

Validation set approach

1. Divide the data randomly into 2 subsets (often roughly of equal size): the *training set* and the *validation set* or *hold-out set*.
2. Use the training set to fit the model, and the validation set to predict and then estimate Mean squared error (MSE).

Potential drawbacks:

1. The method depends on the points chosen, hence different choices may lead to starkly different estimated MSEs.
2. Since we are only using a subset of the available data set, the performance of the model is worse than it would be on a larger data set. And the error estimates tend to be larger.

LOOCV: Leave one out cross validation

Compensates for the drawbacks of the *Validation set approach* yet keeping the same spirit.

1. For every $i \in I = \{1, \ldots, n\}$, train the model on the set $I \setminus \{i\}$.
2. Use this model to predict the ith response, say it is $\hat{y}_i$. and compute $\text{MSE}_i = (y_i - \hat{y}_i)^2$.
3. Compute cross validation error

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i.$$

Advantages:

1. This method has far less bias, since we are fitting the model to $n - 1$ of the points.
2. Does not change depending on the random sample like the validation set method.

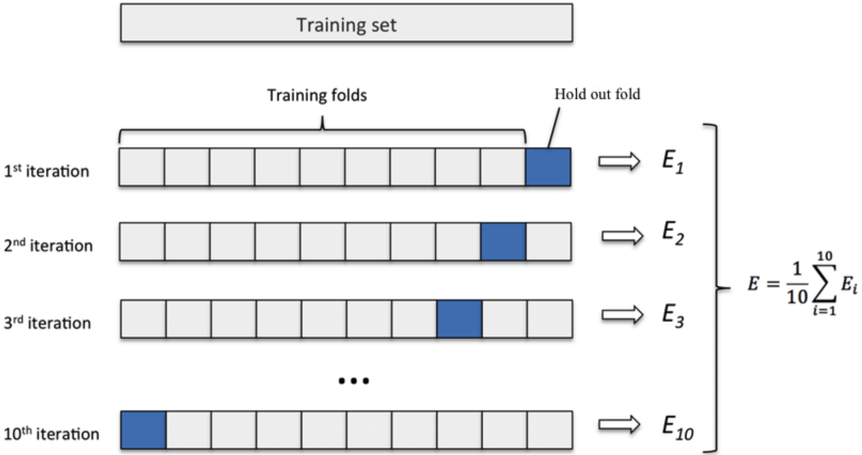
The only potential drawback is that it may be computationally intensive: we need to fit n models.

k-fold cross validation

1. Divide the data randomly into k subsets (folds) of (roughly) equal size.
2. Start with the first fold as a validation set and use the remaining $k-1$ folds to fit the model.
3. Compute the error of the fitted model in the held-out fold.
4. Repeat steps 2 and 3 by using the second, third and so on folds as the hold-out fold with the remaining $k-1$ folds to fit the model.
5. Average the error across all the k fitted models to estimate the cross-validation error.

Some of the commonly used choices for k are 5 or 10.
When $k=n$, this reduces to LOOCV.

k-fold cross validation



LASSO

TWO OBJECTIVES:

- Minimize sum of squared errors in the training set.

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \, .$$

- Penalize complexity for the model. Minimize $\sum_{j=i}^{p} |\beta_j|$.

LASSO

- LASSO: Least absolute shrinkage and selection operator.
- For a tuning parameter $\lambda \geq 0$:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Balance data fit (first term) with model complexity (second term)

1. When $\lambda = 0$, LASSO reduces to standard linear regression.
2. When $\lambda \uparrow \infty$, the second term dominates and LASSO will make all the beta coefficients for the predictor variables go to zero.

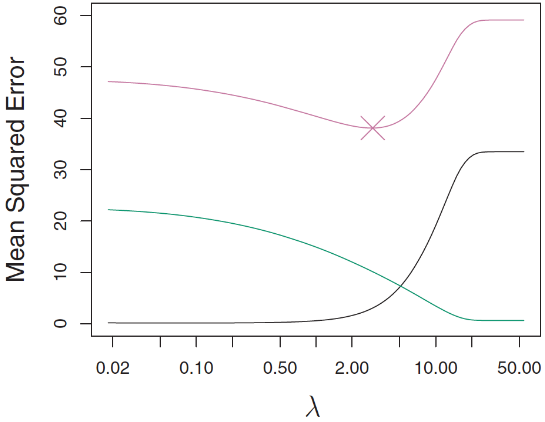
LASSO

1. Proposed in the paper *Regression Shrinkage and Selection via the Lasso*, JRSS B, 1996, by Robert Tibshirani.
2. Around 47000 citations as of June 2022.
3. The objective function in LASSO is convex and tries to roughly promote sparsity.
4. Advantage of LASSO is that since it is convex, the local optimum is the global optimum.
5. Unfortunately, objective function is not differentiable unlike standard linear regression. But there are efficient ways to solve the problem to optimality.

Choice of λ

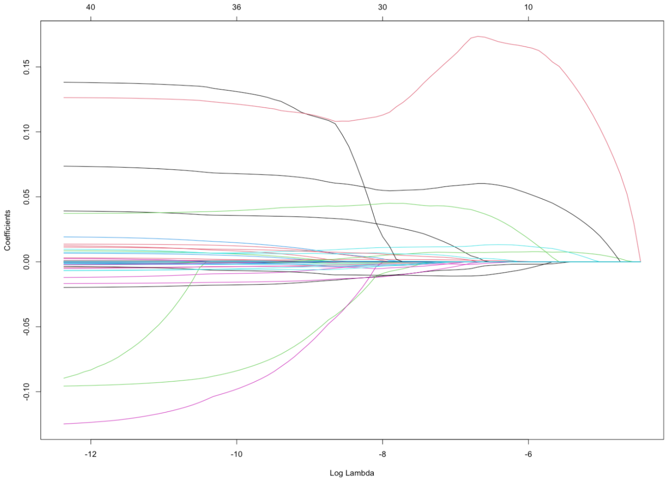
1. Use a grid of possible values and compute the cross-validation error for each value of λ.
2. Choose the λ with the smallest cross-validation error.
3. Finally refit the final model using all the observations for the selected value of λ.

LASSO



- Black line: Squared Bias
- Green line: Variance
- Purple line: Test MSE

LASSO


40
36
30
10
0.15
0.10
0.05
0.00
-0.05
-0.10
Coefficients
-12
-10
-8
-6
Log Lambda

Alternatives to LASSO

- LASSO:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Ridge Regression:

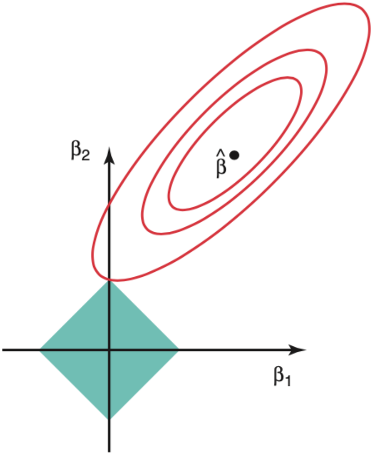
$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

- Elastic Net:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2.$$

– combine ridge regression and LASSO penalty.

LASSO vs Ridge regression



Econometrics: Cross-country growth regression

- Understand factors (economic, political, social) that affect rate of economic growth.
- For example: GDP, degree of capitalism, population growth, equipment investment.
- Many such variables have been proposed. Little guidance from economic theory on choice.
- Why not use subset selection from linear regression.
- We use dataset from *I just ran two million regressions* by Sala-I-Martin and *Model uncertainty in cross country growth regression* by Fernandez et. al.
- 41 possible explanatory variables with 72 countries.
- Note that if you try all 2^{41} possible combinations, it leads to around 2 trillion possibilities.