Regression Analysis

Introduction to model selection

Given candidate predictor variables ($\mathbf{X}_1, \ldots, \mathbf{X}_{p-1}$), which variables should we include in our regression model?

First of all, *why should we select?*

- Suppose the "correct" model is a linear model with some set of predictors (subset of $\mathbf{X}_1, \ldots, \mathbf{X}_{p-1}$).
- Let's consider the most general model, the one that includes all of the potential predictors

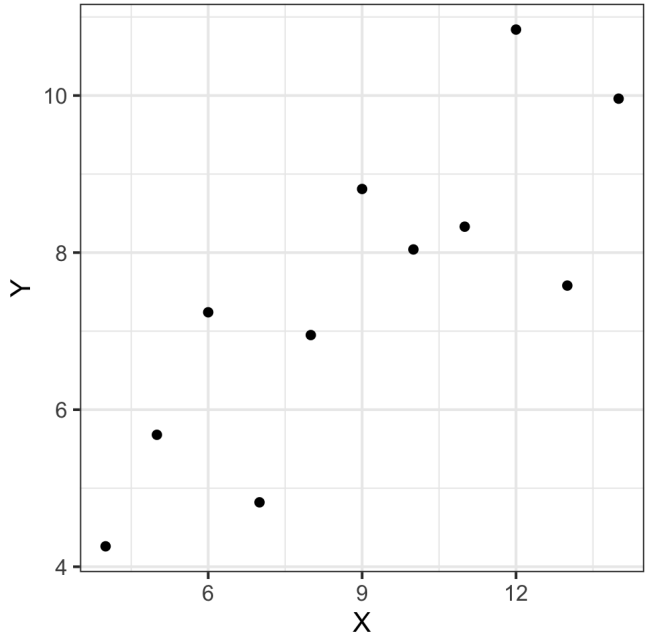
$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-i} + \mathbf{error}_i$$

Introduction to model selection

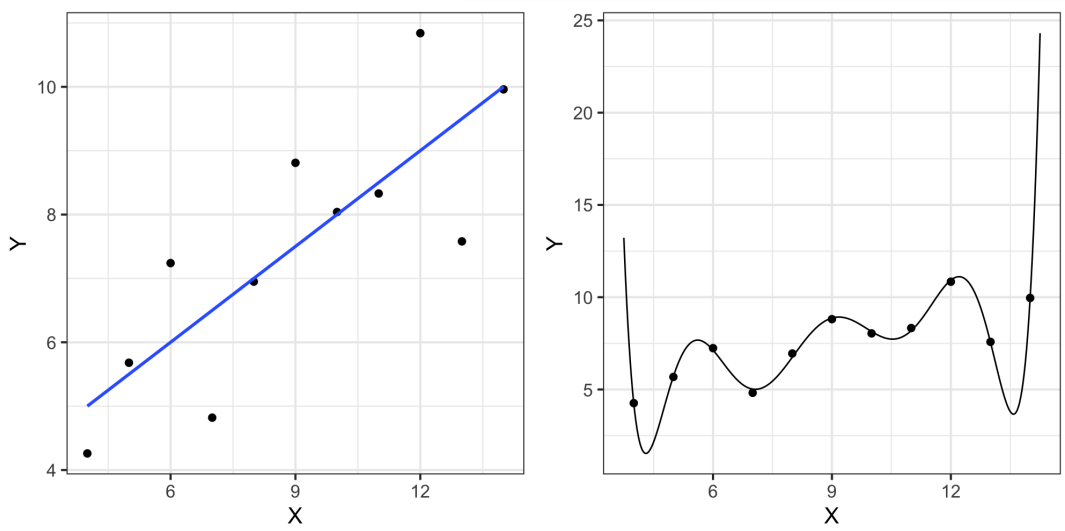
Example A bivariate sample (X, Y) of $n = 11$ observations

M1. $Y = \beta_0 + \beta_1 X + \text{error}$

M2. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_{p-1} X^{p-1} + \text{error}$



Example (cont'd) A bivariate sample (X, Y) of $n = 11$ observations



Left: fit from model 1 $(R^2 = .67)$

Right: fit from model 2 with $p = 9$ $(R^2 = .98)$

Which model is better?

Introduction to model selection

A good model should fit the data well, but should also *generalize* well

We often prefer a **simpler** model

- The principle of parsimony
- Trade-off between bias and variance

The principle of parsimony

- Occam's Razor
  - Occam's Razor explained by Lisa: https://youtu.be/Ly0YzGpi63M
  - If two models give similar fits, we should prefer the model with fewer parameters.
  - It is not a foolproof strategy: the true model can be complex
- If the purpose of analysis is description/explanation of a system, parsimonious models are strongly preferred.
  - Collinearity
  - Better interpretability

Bias-variance trade-off

Bias and variance of the Least Squares Estimator (LSE)

Consider the two linear models

M1. $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$
M2. $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$

Two mistakes:

1. Omission of relevant variables (M1 is true, but we estimated M2)
2. Inclusion of irrelevant variables (M2 is true, but we estimated M1)

Let $\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_1^\top, \widehat{\boldsymbol{\beta}}_2^\top]^\top$ be the LSE for M1, and $\widetilde{\boldsymbol{\beta}}_1$ be the LSE for M2

Bias-variance trade-off

Bias and variance of the Least Squares Estimator (LSE)

1. Omission of relevant variables (M1 is true, but we estimated M2)

- The LSE $\widetilde{\boldsymbol{\beta}}_1$ for $\boldsymbol{\beta}_1$ is biased unless $\mathbf{X}_1^\top \mathbf{X}_2 = 0$
- $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) \succcurlyeq \mathrm{Var}(\widetilde{\boldsymbol{\beta}}_1)$,i.e., lower variance than the "full" model

2. Inclusion of irrelevant variables (M2 is true, but we estimated M1)

- The LSEs $\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2$ for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unbiased.
  - $\mathbb{E}[\widehat{\boldsymbol{\beta}}_2] = 0$
- $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_1) \succcurlyeq \mathrm{Var}(\widetilde{\boldsymbol{\beta}}_1)$

Bias-variance trade-off

Bias and variance trade-off on prediction

- Suppose the true model: $Y = f(\mathbf{x}) + \epsilon$, i.i.d., where $\mathrm{Var}(\epsilon) = \sigma^2$
- Consider a model $\hat{g}$, fitted with the data $(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)$.
  - In the case of a linear model, $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ and $\hat{g}(\mathbf{x}) = \mathbf{x}_M^\top \widehat{\boldsymbol{\beta}}_M$ where $\mathbf{x}_M = \{\mathbf{x}_k\}_{k \in M}$
- **Training error:** $\sum_{i=1}^n (Y_i - \hat{g}(\mathbf{x}_i))^2$
- **Expected prediction error (test error)** at $\mathbf{x} = \mathbf{x}_o$: $\mathbb{E}_{Y_o, (Y_1, \ldots, Y_n)}[(Y_o - \hat{g}(\mathbf{x}_o))^2]$ where $Y_o = f(\mathbf{x}_o) + \epsilon_o$

Bias-variance trade-off

Bias and variance trade-off on prediction

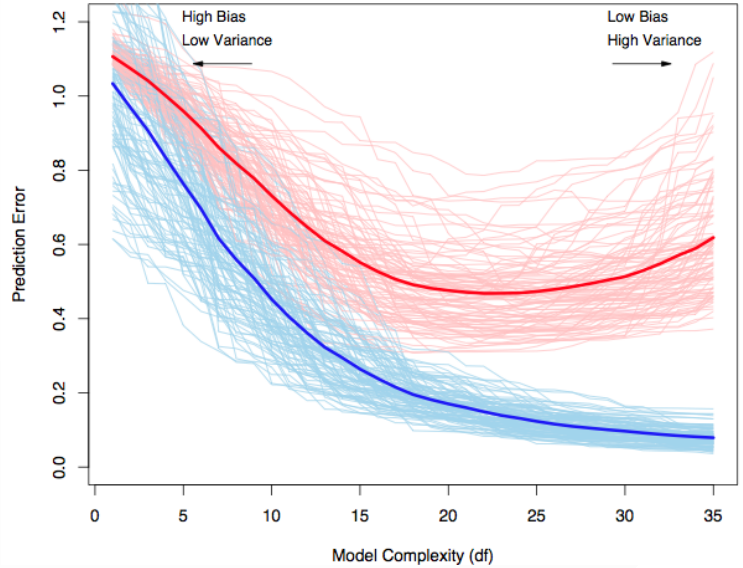
Proposition For any $\mathbf{x} = \mathbf{x}_o$ and a function $\hat{g} : \mathbb{R}^p \to \mathbb{R}$ which depends on the data $Y_1, \ldots, Y_n$, **the expected prediction error** at $\mathbf{x}_o$ can be decomposed as **irreducible error, variance, and bias** of the fitted model $\widehat{g}$.

$$\mathbb{E}_{Y_o,(Y_1,\ldots,Y_n)}[(Y_o - \hat{g}(\mathbf{x}_o))^2] = \sigma^2 + \mathrm{Var}(\hat{g}) + \mathrm{Bias}(\hat{g})^2$$

where $\mathrm{Var}(\hat{g}) = \mathbb{E}[(\hat{g}(\mathbf{x}_o) - \mathbb{E}[\hat{g}(\mathbf{x}_o)])^2]$, $\mathrm{Bias}(\hat{g}) = \mathbb{E}[\hat{g}(\mathbf{x}_o)] - f(\mathbf{x}_o)$

Bias-variance trade-off

Bias and variance trade-off on prediction



Blue curves show the training errors on 100 samples of size 50. Red curves are the corresponding test set errors

Figure credit: ESLR by Friedman, Tibshirani, and Hastie

Model selection

Goal: Find the subset of predictors that gives the “best" model or identify the subset of predictor variables for further study.

- Different “best" subsets serve different purposes (descriptive versus predictive).

Why not compute $t_j = \widehat{\boldsymbol{\beta}}_j / \text{se}(\widehat{\boldsymbol{\beta}}_j)$, $j = 1, \ldots, p$ and drop the predictor variables with large p-values?

- multicollinearity.

Model choice criteria

First, we need some criteria to compare different models

- Coefficient of Determination R^2
- F-statistics for nested models
- PRESS (Prediction Sum of Squares)
- Mallow's C_p
- AIC and BIC

Coefficient of Determination R^2

R^2 measures the proportion of variance in Y explained by the model.

$$R^2 = 1 - \frac{\mathrm{SSErr}}{\mathrm{SSTot}}$$

- SSErr always decreases by adding more predictors to the model
- R^2 can be used for comparing two **non-nested models with the same number of parameters**. However, it is not appropriate for comparing models with different number of parameters.

Adjusted R^2:

$$R^2_{adj} = 1 - \frac{\mathrm{SSErr}/(\mathrm{n} - \mathrm{p})}{\mathrm{SSTot}/(\mathrm{n} - 1)}$$

Nested models

Model 1: Restricted model $\mathbf{Y} = \mathbf{X}_R \boldsymbol{\beta}_R + \boldsymbol{\epsilon}$
Model 2: Full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$

We say Model 1 is nested within Model B if Model 1 can be obtained by putting r constraints on $\boldsymbol{\beta}$ in Model 2.

Test statistic:

$$F = \frac{(\mathrm{SSErr}(\mathrm{Reduced}) - \mathrm{SSErr}(\mathrm{Full}))/r}{\mathrm{SSErr}(\mathrm{Full})/(n-p)}$$

We prefer Model 2 if the observed F is large.

Cross-validation and PRESS

In most cases, we only have one sample (of size n) from the population. How can we know how well a model will predict future observations?

K-fold Cross-validation

- Split data randomly into K roughly equal parts.
- For $k = 1, \ldots, K$, fit the model M using all but the kth part of the data and compute the prediction error sum of squares

$$CV_k = \sum_{i \in \text{kth subset}} (Y_i - \widehat{Y}_{iM}^{(k)})^2$$

where $\widehat{Y}_{iM}^{(k)}$ is the fitted value from the model M using all but the kth part of the data. Then $CV = \sum_{k=1}^{K} CV_k$.

Cross-validation and PRESS

An important special case is **PRESS (Prediction Sum of Squares)** where we use all observations other than the ith case to fit the model M.

$$PRESS = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{iM}^{(-i)})^2$$

- also known as Leave-One-Out Cross-Validation (LOOCV)

AIC and BIC Criteria

- Akaike's information criterion (AIC) (Akaike (1973)

$$AIC = -2 \log L(\hat{\theta}) + 2p$$

- Schwarz' Bayesian information criterion (BIC) (Schwarz (1978))

$$BIC = -2 \log L(\hat{\theta}) + (\log n)p$$

where $\hat{\theta}$ is an MLE, p is the number of parameters.

- Better fit = lower value of $-2 \log L$.
- Penalty for higher model complexity = $2p$ or $p \log n$
- Models with lower AIC/BIC are desirable. BIC tends to favor smaller models.

- AIC and BIC are derived from distinct perspectives:
  - AIC intends to minimize the **Kullback-Leibler divergence** between the true distribution and the estimate from a candidate model
  - BIC intends to select a model that maximizes the **posterior model probability**
- Generally speaking, AIC is preferred for prediction tasks, while BIC is preferred for correct model selections [1].
- Likelihood theory indicates that, if we add an unnecessary predictor to a model, then $2 \log L$ will increase by a random amount distributed approximately as χ^2_1.
  - AIC(M_r)-AIC(M_f) = $\{2 \log L(\hat{\theta}_f) - 2 \log L(\hat{\theta}_r)\} - 2$
  - BIC(M_r)-BIC(M_f) = $\{2 \log L(\hat{\theta}_f) - 2 \log L(\hat{\theta}_r)\} - \log n$

[1] Yang, Yuhong. 2005. Can the strengths of AIC and BIC be shared?

Model selection

1. All possible subsets (Best subset regression)

- Fit all possible regression models. pick the “best" model according to the model selection criterion.
- The total number of possible regression models is 2^{p-1} for $p-1$ explanatory variables.

Example: Example with three variables X_1, X_2 and X_3

1. Consider all models:
    1. Models with 1 variable:
    2. Models with 2 variables:
    3. Models with 3 variables:
2. Identify the best model of each size (lowest SSErr or highest R^2)
    1. Best model with 1 variable:
    2. Best model with 2 variable:
    3. Best model with 3 variable:
3. Identify the best overall model (AIC/BIC, CV, or adjusted R^2)

Example: Example with three variables X_1, X_2 and X_3.

Suppose $Y = 2X_1 - 2X_2 + \text{error}$

```
> library(leaps)
> reg_all = regsubsets(Y~X1+X2+X3,data=dat) # from leaps package
> summary(reg_all)

1 subsets of each size up to 3
Selection Algorithm: exhaustive
         X1  X2  X3
1  ( 1 ) " " "*" " "
2  ( 1 ) "*" "*" " "
3  ( 1 ) "*" "*" "*"
```

```
> cbind(Cp=summary(reg_all)$cp,
+       adjr2=summary(reg_all)$adjr2,
+       BIC=summary(reg_all)$bic)
           Cp     adjr2       BIC
[1,] 64.774647 0.4507984 -51.73386
[2,]  2.210466 0.6679596 -98.47524
[3,]  4.000000 0.6652348 -94.08907
```

Example: Example with three variables X_1, X_2 and X_3

Suppose $Y = 2X_1 - 2X_2 + \text{error}$

```
# Leave-One-Out CV
> loocv=function(fit){
+   h= hatvalues(fit)
+   mean((residuals(fit)/(1-h))^2)
+ }

> CVerr=c()
> for(size in 1:3){
+   fit=lm(Y~X[,1:3][,summary(reg_all)$which[size,-1]])
+   CVerr[size]=loocv(fit)
+ }

> CVerr
[1] 50.18586 43.86106 55.57193
```

1. All possible subsets (Best subset regression)

In the case of linear models, the objective is

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_0 \leq k$$

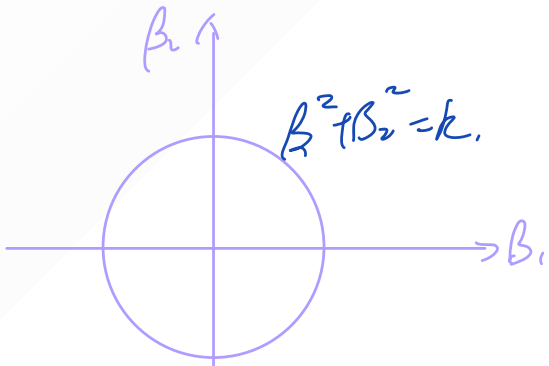
where $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^{p} 1\{\boldsymbol{\beta}_j \neq 0\}$

- "non-convex" problem due to the constraints
- Traditionally, the problem becomes quickly intractable as the number of candidate models increases exponentially with the increase in p
- Recent advances in mixed integer optimization algorithms allow more scalable r best subset selection regression (Bertsimas et al. (2016))

2. Sequential search procedures

- Forward Stepwise Selection
- Backward Elimination
- Stepwise Selection

Model selection



3. Ridge regression

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_2^2 \le k$$

- Shrink all coefficients toward zero (bias-variance tradeoff)
- We include all predictors in the model **(no selection!)**, while alleviating inflated variation in $\widehat{\boldsymbol{\beta}}$ due to collinearity
- Equivalent to solve

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \underset{\beta\in\mathbb{R}^p}{\arg\min}\{\frac{1}{n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}\boldsymbol{\beta}_j^2\}$$

where λ depends on k

Model selection

3. Ridge regression

- An analytic solution is available for Ridge regression:

$$\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = (\mathbf{X}^\top \mathbf{X} + (n\lambda)\mathbf{I}_n)^{-1}\mathbf{X}^\top \mathbf{Y}$$

- The penalty λ trades off between the bias and variance. Large λ leads to a larger bias but less variance
  - when $n\lambda \to 0$, $\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) \to \widehat{\boldsymbol{\beta}}_{LS}(\lambda)$
  - when $n\lambda \to \infty$, $\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) \to 0$

Model selection

4. LASSO (least absolute shrinkage and selection operator) regression

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq k$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\boldsymbol{\beta}_j|$

- A convex relaxation of the best-subset regression problem.
- Equivalent to solve

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\beta}\{\|\frac{1}{n}\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}|\theta_j|\}$$

where λ is a lagrange multiplier.

Model selection

4. LASSO (least absolute shrinkage and selection operator) regression

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\beta}\{\frac{1}{n}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}|\theta_j|\}$$

- LASSO performs both estimation and variable selection
  - $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ tends to contain zeros!
- Higher λ = larger penalty = more sparse solution
- LASSO has become a popular approach in high-dimensional regression problems where the number of predictors p can grow with the sample size n, and p can be potentially larger than n.

4. LASSO (least absolute shrinkage and selection operator) regression

- One of the key assumptions is sparsity
  - the number of $\boldsymbol{\beta}_j$ such that $\boldsymbol{\beta}_j \neq 0$, i.e., $s := |\{j; \boldsymbol{\beta}_j \neq 0\}|$, $s \ll n, p$
- Suppose predictors are not "too correlated". With a "good" choice of λ, LASSO correctly identifies non-zero elements of $\boldsymbol{\beta}$ and with high probability,

$$\|\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) - \boldsymbol{\beta}\|_2^2 \leq C\sigma^2 \frac{s \log p}{n}$$

- If there is a group of highly correlated variables, the LASSO tends to select only one variable from the group.

Model selection

5. Elastic Net regression

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \arg\min_{\beta}\{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\boldsymbol{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\boldsymbol{\beta}_j|\right)\}$$

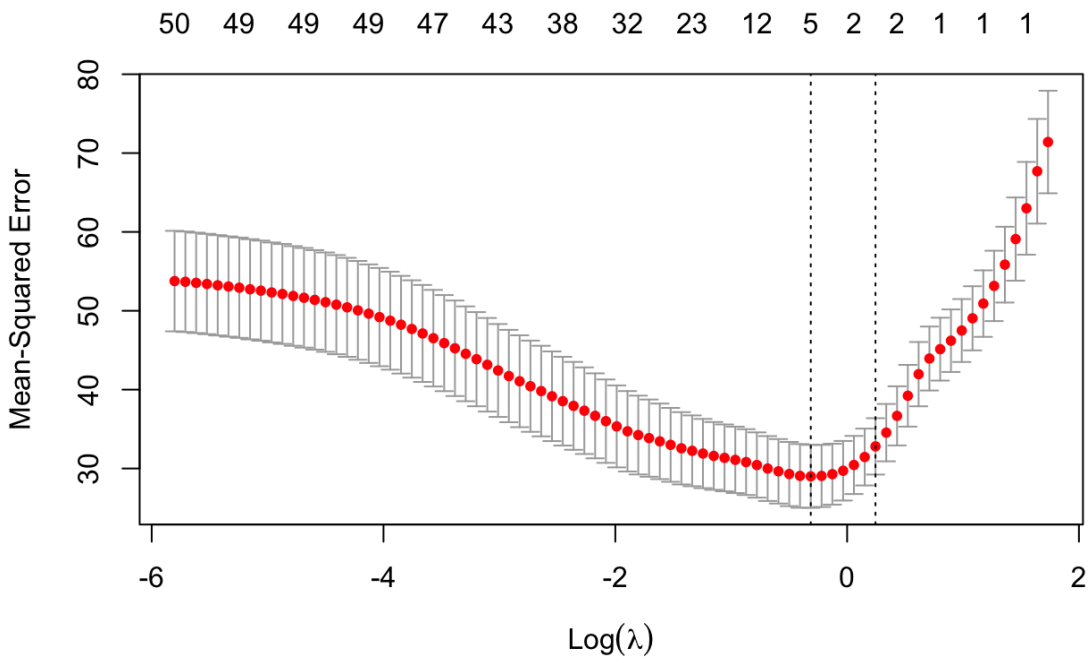
- Elastic net combines L1 and L2 (Lasso and Ridge) approaches
- Designed to overcome limitations of LASSO
  - In particualr, EN alleviates inconsistency problem in LASSO estimation in the case of highly correlated variables, leading to better performance.

Example: Example with 50 variables $X_1, X_2, \ldots, X_{50}$ and $n = 100$. Suppose $Y = 2X_1 - 2X_2 + \text{error}$.

- The package `glmnet` in `R` creates a grid of 100 λ values and fit $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ (`alpha=1`) over the grid of λ values
- `cv.glmnet` function does k-fold cross-validation to select best λ values

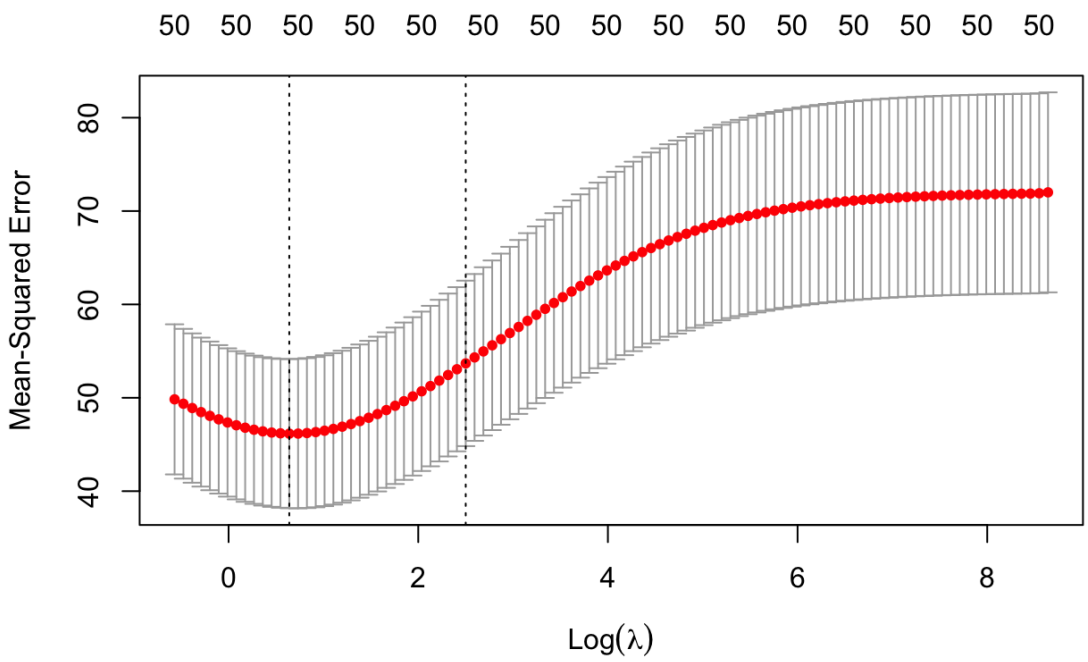
```
> library(glmnet)
> cv.fit=cv.glmnet(X,Y,nfolds = 10, alpha=1) # LASSO (alpha=1) with 10-fold CV
> as.numeric(coef(cv.fit,s = cv.fit$lambda.1se))
 [1]  0.09089487  0.71958945 -2.29377198  0.00000000  0.00000000  0.00000000  0.00000000
 [8]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[15]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[22]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[29]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[36]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[43]  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
[50]  0.00000000  0.00000000
```

```
> plot(cv.fit)
```


50 49 49 49 47 43 38 32 23 12 5 2 2 1 1 1
Mean-Squared Error
30 40 50 60 70 80
-6 -4 -2 0 2
Log(λ)

Ridge regression with 10-fold CV

```
> cv.ridgefit=cv.glmnet(X,Y,nfolds = 10,alpha=0) # Ridge (alpha=0) with 10-fold CV
> plot(cv.ridgefit)
```



Best Subset Selection, Ridge Regression and the Lasso

In the case of an orthogonal $\mathbf{X}$, all procedures have explicit solutions.

- Each method applies a simple transformation to the least squares estimate $\widehat{\boldsymbol{\beta}}$

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I\left(\left\|\hat{\beta}_j\right\| \geq \left\|\hat{\beta}_{(M)}\right\|\right)$
Ridge	$\hat{\beta}_j / (1+\lambda)$
Lasso	$\operatorname{sign}\left(\hat{\beta}_j\right)\left(\left\|\hat{\beta}_j\right\| - \lambda\right)_+$

Subset Selection, Ridge Regression and the Lasso

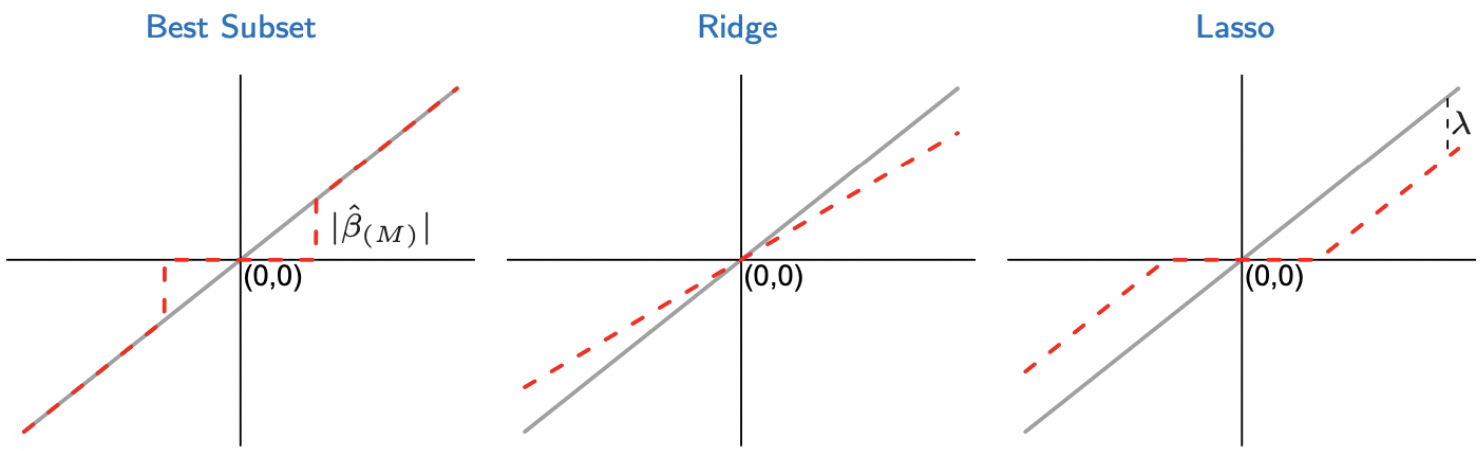


Figure credit: ESLR by Friedman, Tibshirani, and Hastie