6. Model Selection

[MS 1] Why?

[MS 2] Criteria

[MS 3] Criterion-based Procedures

[MS 4] Validation and Cross-Validation

[MS 5] Bias-Variance Tradeoff

[MS 6] Shrinkage Method for Model Selection

[MS 1] Why?

[MS 1] Why?

1. Model Interpretability.
   - Often not all predictors are associated with the response.

$$Y = \beta_0 + \beta_A X_A + \beta_I X_I + \epsilon \quad (\beta_A \neq 0,\ \beta_I = 0)$$

     - Discover associated predictors X_A.
     - Remove irrelevant predictors X_I; reduce unnecessary complexity.
   - Occam's razor: Prefer models easier to interpret.
2. Prediction Accuracy.
   - Models are evaluated by prediction accuracy.

[MS 2] Criteria

[Criterion 1] Coefficient of multiple determination R^2

Given a particular model $\mathcal{M}$,

$$R^2(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})}{\text{SST}}$$

- $\text{SSE}(\mathcal{M}) = \|Y - \hat{Y}_{\mathcal{M}}\|^2$: SSE of the model $\mathcal{M}$
- $\text{SST} = \|Y - \bar{Y}\|^2$: sum of squares total

- Always increases as the model size increases.
  - Tends to prefer a larger model.
- Can be used to compare 2 models of the same size.

[Criterion 2] Adjusted R^2

$$R^2_{adj}(\mathcal{M}) = 1 - \frac{\text{SSE}(\mathcal{M})/(n - p_{\mathcal{M}})}{\text{SST}/(n-1)} = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{\text{null}}}$$

where $p_{\mathcal{M}}$ is the number of parameters in the model $\mathcal{M}$

- When $p_{\mathcal{M}}$ increases,
  - $\text{SSE}(\mathcal{M})$ always decreases,
  - $\hat{\sigma}^2_{\mathcal{M}}$ could increase or decrease.
  - $R^2_{adj}(\mathcal{M})$ could increase or decrease.
- Prefer models with larger $R^2_{adj}(\mathcal{M})$
  - A goodness-of-fit measure.

[Criterion 3] Mallow's C_p

$$C_p(\mathcal{M}) = \frac{\text{SSE}(\mathcal{M})}{\hat{\sigma}^2} - n + 2 \times p_{\mathcal{M}}$$

- $\hat{\sigma}^2 = \text{SSE}(\mathcal{F})/df_{\mathcal{F}}$
  - $\mathcal{F}$ denotes the fullest model
  - best estimate of σ^2

- The criterion is motivated from the Model Error (ME).
  - $\text{ME} = \| \text{E}(Y) - \hat{Y} \|^2$
  - $\text{E}(\text{ME}) = \text{E}(\text{SSE}) + \sigma^2(-n + 2p)$
  - See the derivation on Notes.

- Let $a = \mathrm{E}(Y) - \hat{Y}$.

$$\mathrm{E}\,\|a\|^2 = \|\,\mathrm{E}(a)\|^2 + \mathrm{tr}\{\mathrm{var}(a)\}$$

- Thus,

$$\begin{aligned}\mathrm{E}(\mathrm{ME}) =& \|\,\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\|^2 + \mathrm{tr}\{\mathrm{var}(\hat{Y})\} \\ =& \,\mathrm{bias}^2 + \mathrm{variance}\end{aligned}$$

[Criterion 4] AIC: Akaike Information Criterion

- Definition: For a general model $\mathcal{M}$ with parameter θ,

$$\text{AIC}(\mathcal{M}) = -2 \log L_{\mathcal{M}}(\hat{\theta}) + 2 \times p_{\mathcal{M}}$$

  where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- Motivation: Kullback-Leibler discrepancy

$$KL(f, g) = \int \log \frac{f(\boldsymbol{y})}{g(\boldsymbol{y}; \theta)} f(\boldsymbol{y}) d\boldsymbol{y}$$

  - A measure of difference between a true fixed f and various competing models g depending on parameter θ.
    - non-symmetric $KL(f, g) \neq KL(g, f)$.
    - $KL(f, g) \geqslant KL(f, f) = 0$.
    - $KL(f, g) = -\int \log g(\boldsymbol{y}; \theta) f(\boldsymbol{y}) d\boldsymbol{y} + \text{constant}$

AIC under multiple linear models

- if σ^2 is known,

$$\text{AIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + 2p_{\mathcal{M}}.$$

  - Similar to C_p if replace σ^2 by $\hat{\sigma}^2$ (only differ by $-n$)

- if σ^2 is unknown,

$$\text{AIC} = n \log(\text{SSE}_{\mathcal{M}}/n) + 2p_{\mathcal{M}}$$

[Criterion 5] BIC: Bayesian Information Criterion

For a general model $\mathcal{M}$ with parameter θ,

$$\text{BIC}(\mathcal{M}) = -2 \log L_{\mathcal{M}}(\hat{\theta}) + \log(n) \times p_{\mathcal{M}}$$

where $L_{\mathcal{M}}(\hat{\theta})$ denotes the likelihood function of the parameters in the model $\mathcal{M}$ evaluated at the MLE.

- BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC.

BIC is derived under the Bayesian perspective

- Consider the multiple linear model with σ^2 known.
- Suppose we fit a submodel with $\mathbf{X}_p\beta_p$
  - p can be smaller than the total number of covariates
  - Assume β has prior distribution $N_p(\mathbf{m}, \sigma^2 V)$
  - The log posterior distribution of β_p is proportional to

$$\text{BIC} = \frac{\text{SSE}_{\mathcal{M}}}{\sigma^2} + \log(n)p_{\mathcal{M}}$$

- Detailed proof: read Linear Regression Analysis (Lee and Seber) Section 12.3.4

Summary

- R^2: motivated from $\text{corr}^2(\hat{Y}, Y)$ (prefer larger)
- Adjusted R^2: penalizes model complexity (prefer larger)
- C_p: motivated from $\| \text{E}(Y) - \hat{Y} \|^2$ (prefer smaller)
- AIC and BIC: motivated from KL divergence (prefer smaller)

[MS 3] Criterion-based Procedures

3.1 Best Subset Selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here best is defined as having the smallest RSS $=$ SSE, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using C_p, AIC, BIC, or adjusted R^2.

3.2 Stepwise Selection: Forward

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 0, 1, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using C_p, AIC, BIC, or adjusted R^2.

3.3 Stepwise Selection: Backward

1. Let $\mathcal{M}_p$ denote the full model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using C_p, AIC, BIC, or adjusted R^2.

Data example

- 50 states data collected by U.S. Bureau of the Census
- Response: life expectancy

```
#read data and load package
library(faraway)
data(state)
statedata <- data.frame(state.x77,row.names=state.abb)
head(statedata)
```

```
##    Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
## AL       3615   3624        2.1    69.05   15.1    41.3    20  50708
## AK        365   6315        1.5    69.31   11.3    66.7   152 566432
## AZ       2212   4530        1.8    70.55    7.8    58.1    15 113417
## AR       2110   3378        1.9    70.66   10.1    39.9    65  51945
## CA      21198   5114        1.1    71.71   10.3    62.6    20 156361
## CO       2541   4884        0.7    72.06    6.8    63.9   166 103766
```

R package: leaps

```
library(leaps)
```

regsubsets: R function for model selection

- method: exhaustive search, forward or backward stepwise

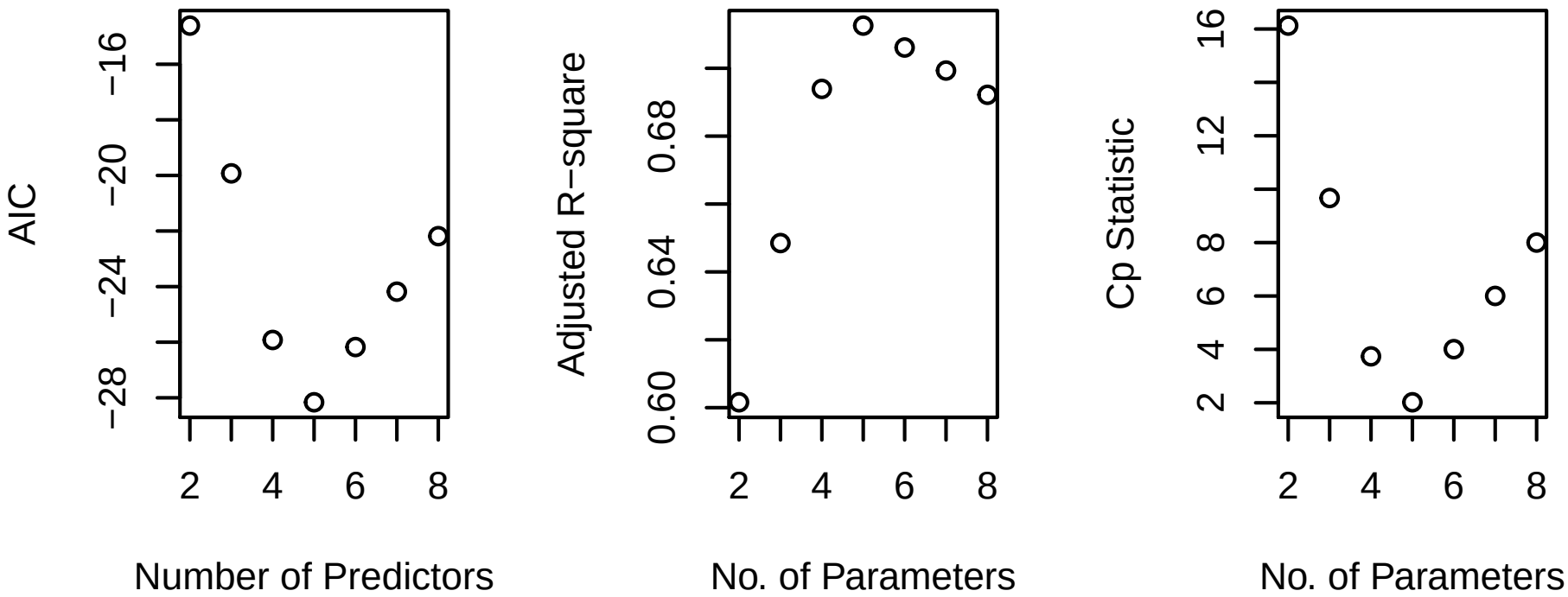
```
b <- regsubsets(Life.Exp~.,data=statedata, method="exhaustive")
rs <- summary(b)
```

- for each size of model p, it finds the variables that produce the minumum RSS .

```
rs$which
```

```
##   (Intercept) Population Income Illiteracy Murder HS.Grad Frost  Area
## 1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
## 2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
## 3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
## 6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
## 7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

```
AIC <- 50*log(rs$rss/50) + (2:8)*2
par(mfrow=c(1,3))
plot(AIC ~ c(2:8), ylab="AIC", xlab="Number of Predictors")
plot(2:8, rs$adjr2, xlab="No. of Parameters", ylab="Adjusted R-square")
plot(2:8,rs$cp,xlab="No. of Parameters",ylab="Cp Statistic")
```



[MS 4] Validation and Cross-Validation

Prediction Error

- Arguably, the goal of a regression analysis is to "build" a model that predicts well.
- One way to measure this is in the expected prediction error of the model.
  - Estimate model parameters $\hat{\beta}$ from training data.
  - Consider future data $(X_{\text{new}}, Y_{\text{new}})$
  - Given X_{new}. Predict Y_{new} by $\hat{Y}_{\text{new}} = X_{\text{new}}\hat{\beta}$.
  - Prediction Error is

$$\text{PE} = \text{E}_{Y_{\text{new}}} \| Y_{\text{new}} - \hat{Y}_{\text{new}} \|^2$$

Model Validation

Model validation refers to checking a selected model against independent data.

1. Collect new data as validation data set.
2. Split data into training and validation set.

- Estimate model by a training set.
- Evaluate Mean Squared Prediction Error by

$$\text{MSPE} = \frac{\sum_{i \in \mathcal{V}} (Y_i - \hat{Y}_i)^2}{|\mathcal{V}|}$$

  - $|\mathcal{V}|$ is the sample size of the validation data set.
  - Y_i is the i th observed response in the validation data set.
  - $\hat{Y}_i$ is the i th predicted response in the validation data set.

Leave-One-Out Cross-Validation

- Suppose we have n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- For $i = 1, \ldots, n$
  - Fit a model with observations exclusing i-th observation.
  - Make a prediciton $\hat{y}_i$ using the fitted model.
  - Define $\text{MSE}_i = (y_i - \hat{y}_i)^2$ (prediction error).
- Define LOOCV estimate

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

k-Fold Cross-Validation

- Split data randomly into K roughly equal parts.
- For $k = 1, \ldots, K$, fit the model using all but the k th part of the data and obtain predicted values $\hat{Y}_{ki}$
- Compute the prediction error mean sum of squares

$$\mathrm{CV}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(Y_{ki} - \hat{Y}_{ki} \right)^2$$

- Compute a K-fold cross-validation estimate

$$\mathrm{CV} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CV}_k$$

Example on LOOCV

```
library(ISLR)
library(boot)
```

- Auto Data: Including MPG, horsepower, and other information for 392 vehicles.
- LOOCV: done by `cv.glm` in the package `boot`.

```
glm.fit = glm(mpg ~ horsepower, data =Auto)
```

- glm gives the same fit as lm but can be input for cv.glm

```
cv.err = cv.glm( Auto, glm.fit)
```

```
cv.err$delta[1] #LOOCV estimate
```

```
## [1] 24.23151
```

Example on K-fold CV

- Set K option in cv.glm

```
cv.glm( Auto, glm.fit, K=10)$delta[1]
```

```
## [1] 24.12499
```

- Similar value to LOOCV.
- K-fold CV can be less computationally demanding compared to LOOCV under general models.

[MS 5] Bias-Variance Tradeoff

Decomposing PE

- Expected prediction error/ Mean squared error

$$\text{MSE} = \text{E}(\text{PE}) = \text{E}\,\|Y_{\text{new}} - \hat{Y}_{\text{new}}\|^2$$

- We have

$$\begin{aligned}\text{MSE} =& \|\,\text{E}(Y_{\text{new}}) - \text{E}(\hat{Y}_{\text{new}})\|^2 + \text{tr}\{\text{var}(Y_{\text{new}} - \hat{Y}_{\text{new}})\} \\ =& \,\text{bias}^2 + \text{variance}\end{aligned}$$

  - $\hat{Y}_{\text{new}}$ is from old (training) data.
  - Y_{new} is from new data.
    - When independent, variance $= \text{tr}\{\text{var}(\epsilon_{\text{new}}) + \text{var}(\hat{Y}_{\text{new}})\}$
    - $\text{tr}\{\text{var}(\epsilon_{\text{new}})\}$ is the irreducible variane while $\text{tr}\{\text{var}(\hat{Y}_{\text{new}})\}$ depends on model.

Bias-variance trade-off

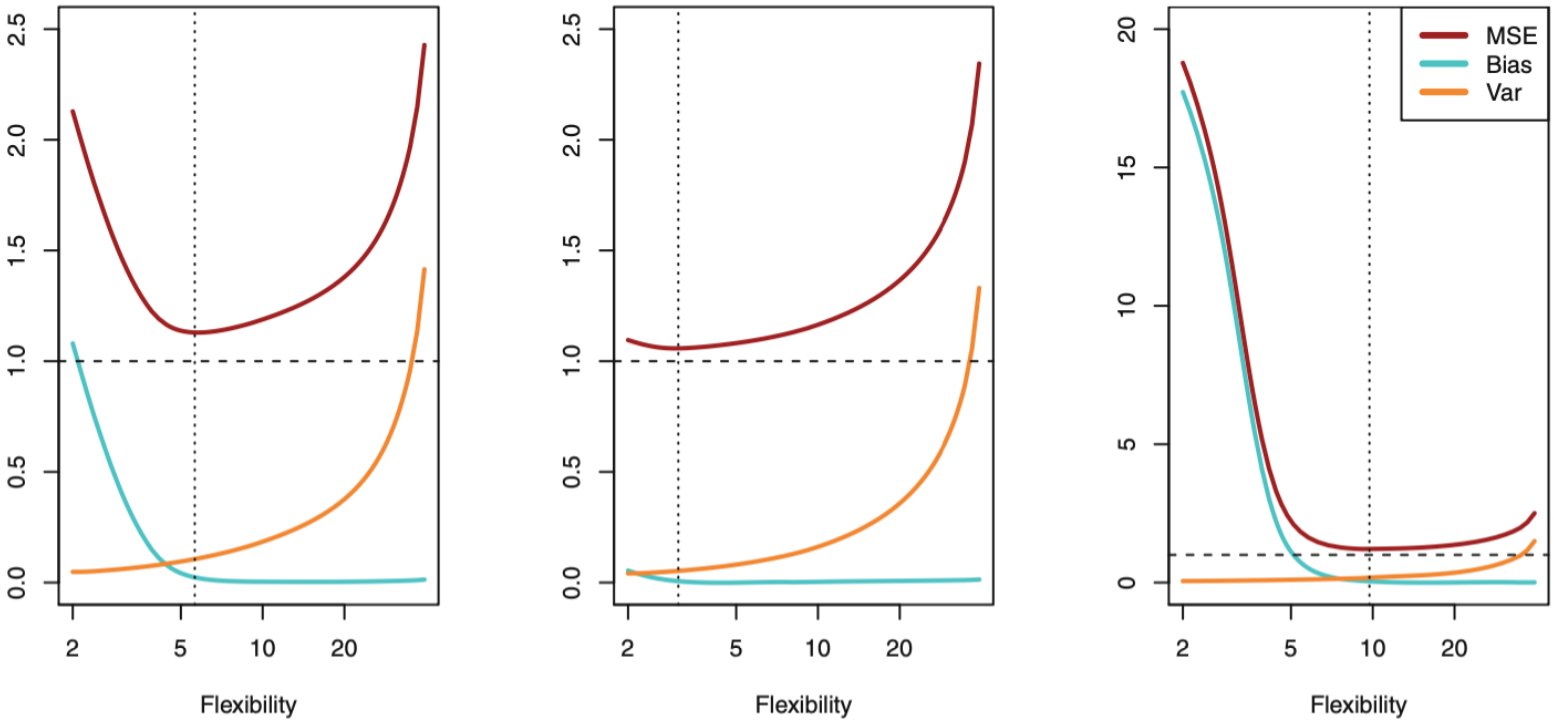


Figure 1: Figure from "An Introduction to Statistical Learning".

- It is possible to find a model with lower MSE than an unbiased model!
- Bias-variance trade-off is "generic" in statistics: almost always introducing some bias yields a decrease in MSE.

Stein Shrinkage

1. Suppose $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_p)$.
2. An obvious estiamte of $\boldsymbol{\mu}$ is $\mathbf{Z}$.

- Unbiased estiamte.
- But $\|\mathbf{Z}\|^2$ tends to be too large.
  - $\mathrm{E}(\|\mathbf{Z}\|^2) = p\sigma^2 + \|\boldsymbol{\mu}\|^2$
  - $> \|\boldsymbol{\mu}\|^2$. Intuitively, at least some of the elements of the estimate are too large.

3. Another estimator $c\mathbf{Z}$ with a constant $c \in (0, 1)$.
   - Biased.
   - But by bias-variance trade-off, we can choose an appropriate c so that mean squared error $\mathrm{E}(\|c\mathbf{Z} - \boldsymbol{\mu}\|^2)$ is small.

Shrinkage and Penalty

- Corresponds to

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 + \lambda \times \|\boldsymbol{\mu}\|^2$$

- This is also Lagrange form of the "constrained" minimization.

$$\text{minimize}_{\boldsymbol{\mu}} \|\mathbf{Z} - \boldsymbol{\mu}\|^2 \qquad \text{subject to } \|\boldsymbol{\mu}\|^2 \leqslant C$$

  - For any λ, there is some C such that the solutions of two problems are the same, and vice versa.
  - Intuitively, constrains $\|$ minimizer $\|^2$ not too large.
    - If $C = \infty$ or $\lambda = 0$, solution is OLS.
    - As C gets smaller, λ gets larger, find solution subject to the constraint $\|\boldsymbol{\mu}\|^2 \leqslant C$.

[MS 6] Shrinkage Method for Model Selection

Ridge Regression

Motivation: Suppose $\mathbf{X}^\top \mathbf{X} = n\mathbf{I}_p$.

- $\sqrt{n}(\hat{\beta} - \beta) \sim N_p(0, \sigma^2 \mathbf{I}_p)$.
- $\hat{\beta}$ has a shrinkaged version $\tau\hat{\beta}$ with smaller MSE.
- Ridge Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2$$

  - Also corresponds to an $\|\beta\|^2$ constrained optimization.
  - Solution: $\hat{\beta}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}$

- Ridge Regression will include all p predictors in the final model.
- The penalty $\lambda\|\beta\|^2$
  - will shrink all of the coefficients towards zero
  - but it will not set any of them exactly to zero (unless $\lambda = \infty$)
  - may not be a problem for prediction accuracy
  - can create a challenge in model interpretation if p is too large

Lasso Regression

Lasso Regression:

$$\min_\beta \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

- $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$
- Also corresponds to an $\|\boldsymbol{\beta}\|_1$ constrained optimization.
- Lasso can zero some coefficients.
  - If $\mathbf{X}^\top\mathbf{X} = \mathbf{I}_p$ and $\lambda = 2\gamma$, lasso solution

$$\tilde{\beta}_j = \begin{cases} \text{sign}(\hat{\beta}_j) \times (|\hat{\beta}_j| - \gamma), & \gamma \leq |\hat{\beta}_j|, \\ 0, & \text{otherwise} \end{cases}$$

Graph Illustration

- Consider $p = 2$.
- The solid blue areas are the constraint regions $|\beta_1|^2 + |\beta_2|^2 \leqslant C$ and $|\beta_1| + |\beta_2| \leqslant C$
- The red ellipses given regions of constant RSS.

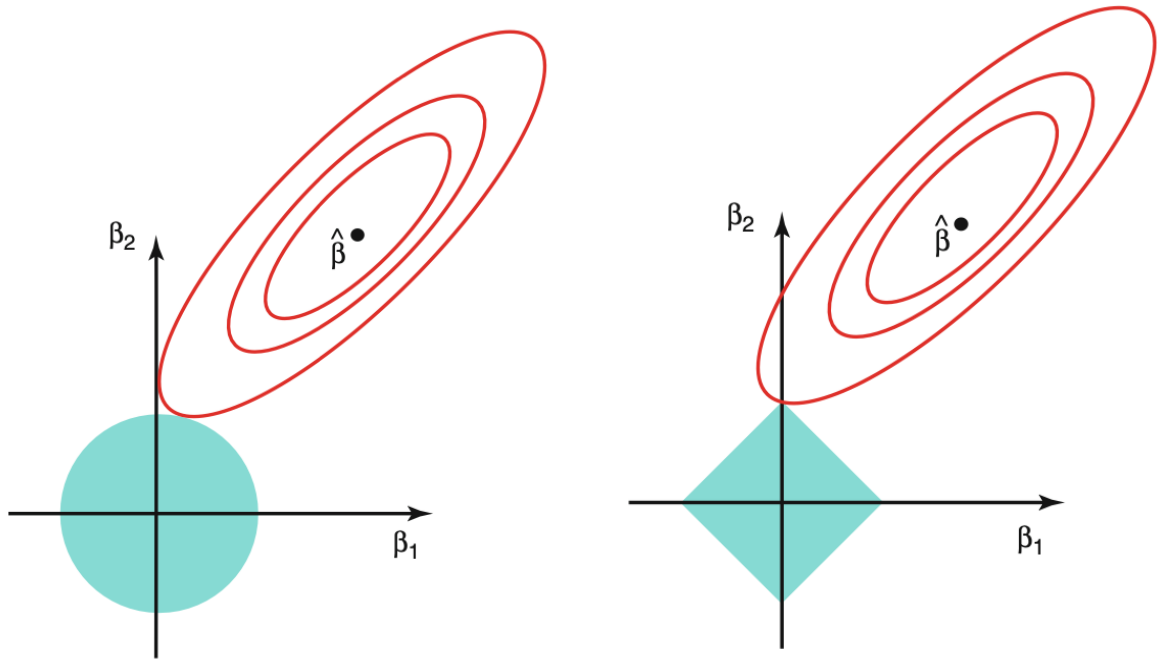


Figure 2: From "An Introduction to Statistical Learning".

Comparison

- Neither ridge regression nor the lasso will universally dominate the other.
- In general, one might expect
  - lasso to perform better: a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero.
  - Ridge regression will perform better: the response is a function of many predictors, all with coefficients of roughly equal size.
- The number of predictors that is related to the response is never known a priori for real data sets.
- Cross-validation can be used in order to determine which approach is better on a particular data set and also choose λ.

Example

- Hitters Data: Records and salaries for baseball players.

```
Hitters=na.omit(Hitters)
head(Hitters,2)
```

```
##                AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
## -Alan Ashby      315   81     7   24  38    39    14   3449   835      69   321
## -Alvin Davis     479  130    18   66  72    76     3   1624   457      63   224
##                CRBI CWalks League Division PutOuts Assists Errors Salary
## -Alan Ashby     414    375      N        W     632      43     10    475
## -Alvin Davis    266    263      A        W     880      82     14    480
##                NewLeague
## -Alan Ashby            N
## -Alvin Davis           A
```

```
x=model.matrix(Salary ~ ., Hitters)[,-1]
y=Hitters$Salary
```

- In glmnet() function: alpha option determines the model type.
  - alpha $= 0$ ridge; alpha $= 1$ lasso.

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod=glmnet(x, y, alpha=0, lambda=grid)
```

- Read reuslts for the 60th λ

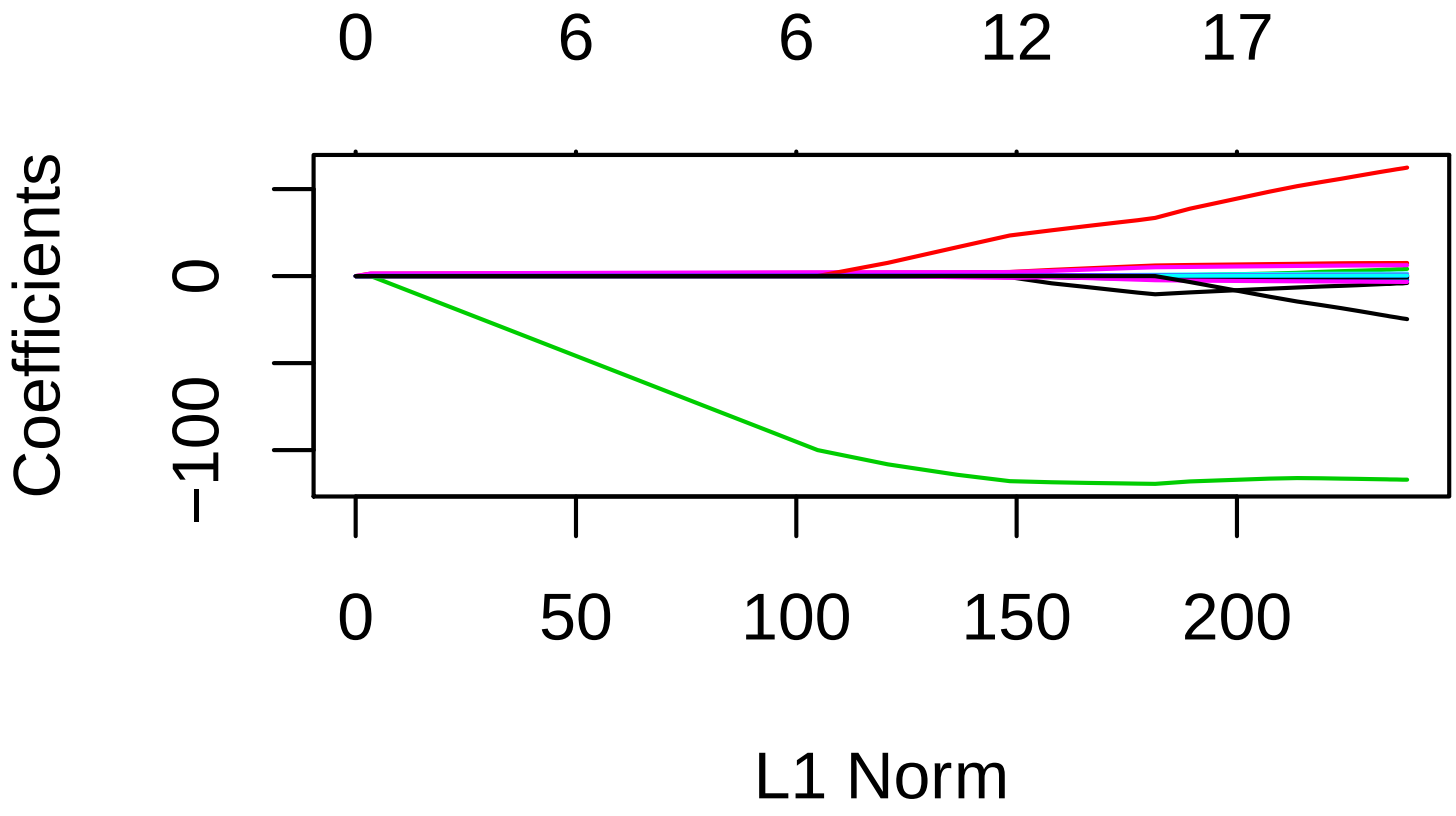
```
ridge.mod$lambda[60] #||beta||^2
```

```
## [1] 705.4802
```

```
coef(ridge.mod)[1:5,60]
```

```
## (Intercept)       AtBat        Hits       HmRun        Runs
##  54.3251995   0.1121111   0.6562241   1.1798091   0.9376971
```

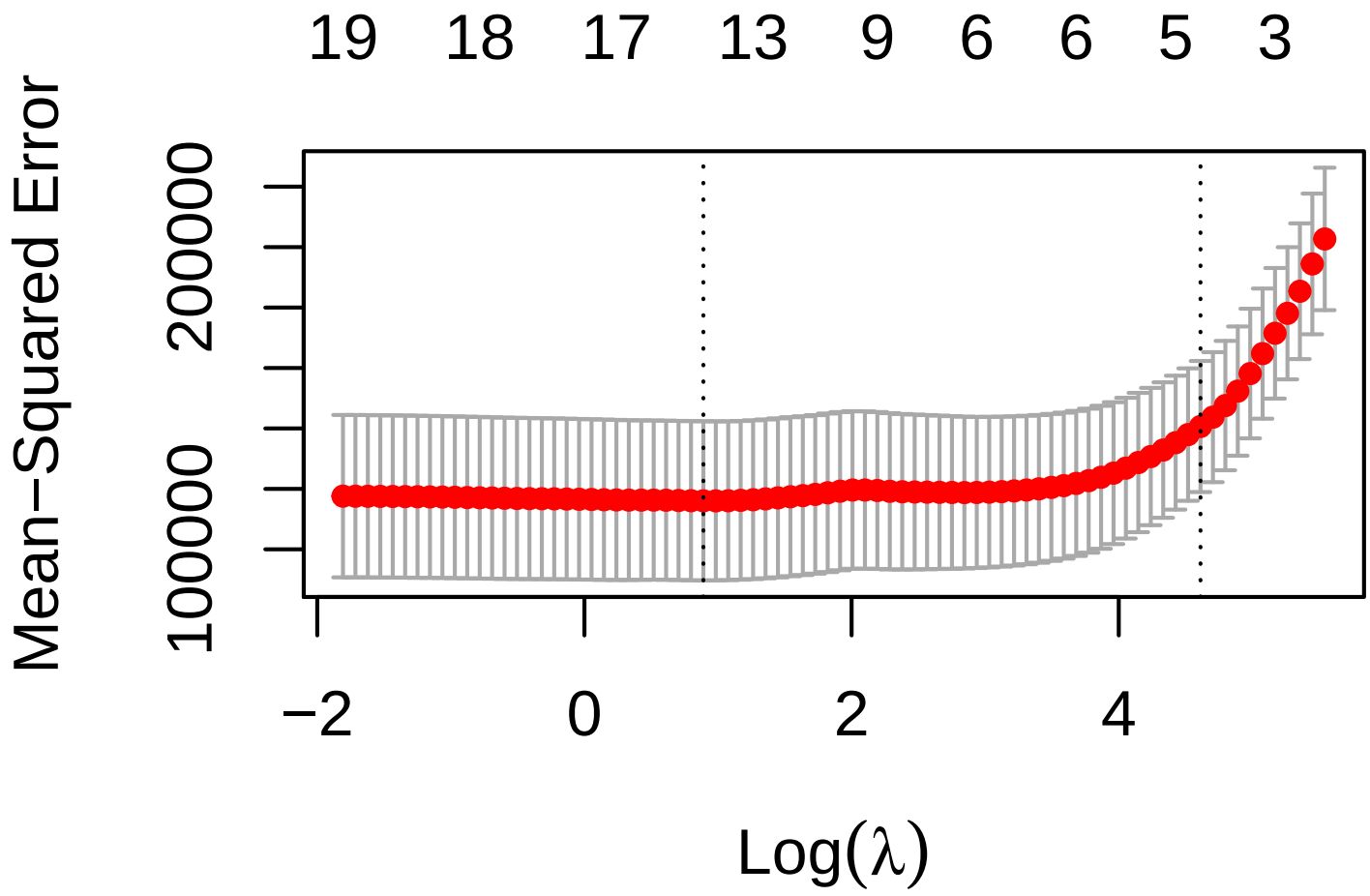
```
lasso.mod=glmnet(x,y,alpha=1,lambda=grid)
plot(lasso.mod)
```



- Each curve corresponds to a variable.
- It shows the path of its coefficient against the $\|\hat{\beta}\|_1$.
- The axis above indicates # of nonzero coefficients at the current λ.

Cross validaiton

```
cv.out <- cv.glmnet(x, y, alpha=1) #default # of folds is 10
plot(cv.out)
```




```
cv.out$lambda.min
```

```
## [1] 2.436791
```