Linear Model Selection & Regularization

Damek Davis
School of ORIE, Cornell University
ORIE 4740 Lec 9–10 (Feb 22, Feb 24)

Recap: A close look at testing error

$$\text{test error} = \text{bias}^2 + \text{variance} + c$$

- As model flexibility increases:

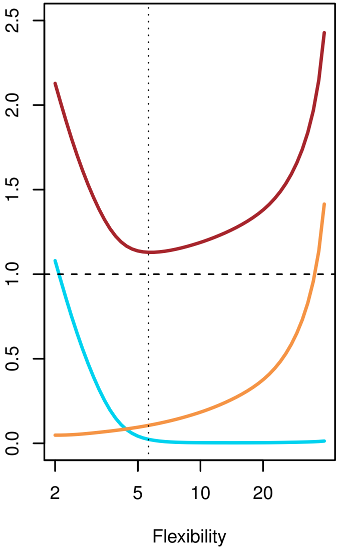
Bias: decreases

Variance: increases

- **Goal:** select model with lowest test error

- Can estimate the test error from data

  E.g., by k-fold cross-validation



Model Selection in Linear Regression (ISLR Sec 6.1)

n data points

1 response Y, p predictor variables $X_1, X_2, \ldots, X_p$

Model Selection in Linear Regression (ISLR Sec 6.1)

n data points

1 response Y, p predictor variables $X_1, X_2, \ldots, X_p$

May not want to use all p predictors:

$$Y \approx X_1 + X_3$$

Model Selection in Linear Regression (ISLR Sec 6.1)

n data points

1 response Y, p predictor variables $X_1, X_2, \ldots, X_p$

May not want to use all p predictors:

$$Y \approx X_1 + X_3$$

Which of the following is **not** a valid reason to use less than p predictors?

A. Some variables may be irrelevant

B. More variables $\Rightarrow$ Harder to interpret the fitted model

C. Less variables $\Rightarrow$ higher bias

D. Extreme case: $p > n \Rightarrow$ Overfit

E. Easier to build larger training sets.

Model Selection in Linear Regression (ISLR Sec 6.1)

n data points

1 response Y, p predictor variables $X_1, X_2, \ldots, X_p$

May not want to use all p predictors:

$$Y \approx X_1 + X_3$$

The following are valid reasons to use less than p predictors:

A. Some variables may be irrelevant

B. More variables $\Rightarrow$ Harder to interpret the fitted model

C. More variables $\Rightarrow$ More flexible $\Rightarrow$ Higher variance

D. Extreme case: $p > n \Rightarrow$ Overfit

E. Easier to build larger training sets.

Model Selection in Linear Regression (ISLR Sec 6.1)

n data points

1 response Y, p predictor variables $X_1, X_2, \ldots, X_p$

May not want to use all p predictors:

$$Y \approx X_1 + X_3$$

The following are valid reasons to use less than p predictors:

A. Some variables may be irrelevant

B. More variables $\Rightarrow$ Harder to interpret the fitted model

C. More variables $\Rightarrow$ More flexible $\Rightarrow$ Higher variance

D. Extreme case: $p > n \Rightarrow$ Overfit

E. Easier to build larger training sets.

Model selection:

- How many variables to use?
- Which variables?

Model Selection in Linear Regression

Model selection: select a subset of variables

- How many variables to use?
- Which variables?

Model Selection in Linear Regression

Model selection: select a subset of variables

- How many variables to use?
- Which variables?

Select variables with small p-values?

Model Selection in Linear Regression

Model selection: select a subset of variables

- How many variables to use?
- Which variables?

Select variables with small p-values?

- Only measures relevance on training data
- Only works well when $n \gg p$

Model Selection in Linear Regression

Model selection: select a subset of variables

- How many variables to use?
- Which variables?

Select variables with small p-values?

- Only measures relevance on training data
- Only works well when $n \gg p$

Other ways to select variables?

Best Subset Selection

(ISLR Sec 6.1.1)

Idea: exhaustive search
Enumerate all possible subsets of variables, select the “best” one

Best Subset Selection

(ISLR Sec 6.1.1)

Idea: exhaustive search

Enumerate all possible subsets of variables, select the "best" one

- Subset of size 0: one model (intercept only)
- Subset of size 1: p models
- Subset of size 2: $p(p-1)/2$ models

$\vdots$

Best Subset Selection

(ISLR Sec 6.1.1)

Idea: exhaustive search

Enumerate all possible subsets of variables, select the "best" one

- Subset of size 0: one model (intercept only) $(\mathcal{M}_0)$
- Subset of size 1: p models $(\mathcal{M}_1)$
- Subset of size 2: $p(p-1)/2$ models $(\mathcal{M}_2)$

  $\vdots$

Algorithm:

1 For $k = 0, 1, 2, \ldots, p$

  a Fit all $\binom{p}{k}$ models with k variables

  b Among them, pick the "best" one; call it $\mathcal{M}_k$

2 Among models $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick a single "best" model

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the "best" one ; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick a single "best" model

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the "best" one ; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick a single "best" model

What would be the ideal way to define "best"?

A. Largest R^2 statistic

B. Smallest R^2 statistic

C. Smallest testing error

D. Smallest training error

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2

Choose one

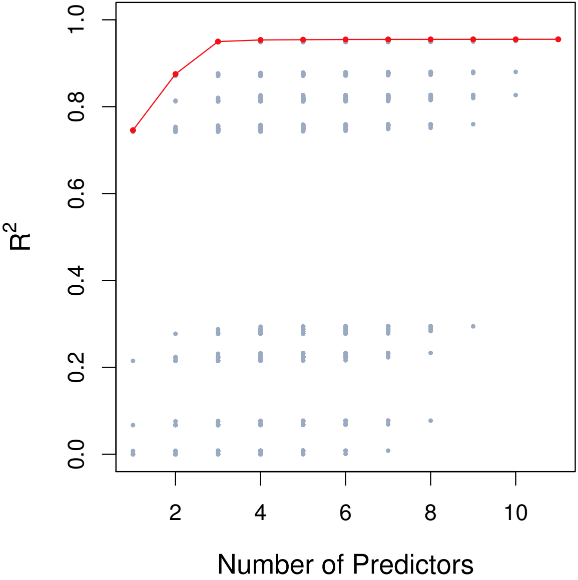
Which models could be returned by this procedure?

A. $\mathcal{M}_0$

B. $\mathcal{M}_k$ for any $k = 1, \ldots, p-1$

C. $\mathcal{M}_p$.

`Credit` Dataset



Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the "best" one ; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick a single "best" model

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick a single "best" model

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2

Problem?

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2

Problem?

- R^2 always increase with more variables
- Will end up selecting all p variables

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - a. Fit all $\binom{p}{k}$ models with k variables
   - b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. ~~Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2~~

Problem?

- R^2 always increase with more variables
- Will end up selecting all p variables

Best Subset Selection

Algorithm:

1 For $k = 0, 1, 2, \ldots, p$

  a Fit all $\binom{p}{k}$ models with k variables

  b Among them, pick the one with largest R^2; call it $\mathcal{M}_k$

2 ~~Among $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$, pick the one with largest R^2~~

Problem?

- R^2 always increase with more variables
- Will end up selecting all p variables

2 Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest <u>estimated</u> test error

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest estimated test error

Best Subset Selection

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest estimated test error

---

How to estimate test error?

- General: Cross-Validation (Expensive!)
- For linear regression: Make appropriate adjustments to the training error or R^2

  Adjusted R^2, AIC, BIC, C_p

Adjusted R^2

Recall:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{explained variability}}{\text{total variablity}}$$

- More predictors $\Rightarrow$ Larger R^2

Adjusted R^2

Recall:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = \frac{\text{explained variability}}{\text{total variablity}}$$

- More predictors $\Rightarrow$ Larger R^2

Using k predictors:

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

- Maximize adjusted $R^2 \Leftrightarrow$ minimize $\text{RSS}/(n-k-1)$
- Penalize large k (number of predictors)

Best Subset Selection Using Adjusted R^2

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

Best Subset Selection Using Adjusted R^2

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - **a** Fit all $\binom{p}{k}$ models with k variables
   - **b** Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

Suppose that we replaced "largest R^2" with "largest adjusted R^2" in part **b**. Would the final model change?

A. Yes

B. No

Best Subset Selection Using Adjusted R^2

Example: `Credit` dataset

- Response: `Balance`
- Predictors: `Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married, Ethnicity` (3 levels)
- $p = 11$, $n = 400$

Best Subset Selection Using Adjusted R^2

Example: `Credit` dataset

- Response: `Balance`
- Predictors: `Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married, Ethnicity` (3 levels)
- $p = 11$, $n = 400$

(see ISLR 6.5.1 for R tutorial)

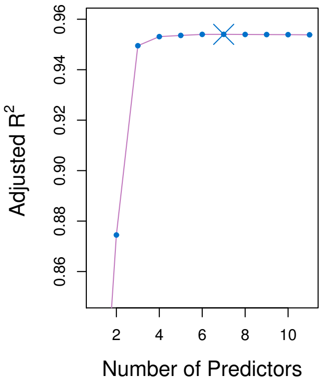
```
> library(leaps)

> regfit.full = regsubsets(Balance~., data=Credit, nvmax=11)

> summary(regfit.full)
```

```
          Income Limit Rating Cards Age Education GenderFemale
1  ( 1 )  " "    " "   "*"    " "   " " " "       " "
2  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
3  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
4  ( 1 )  "*"    "*"   " "    "*"   " " " "       " "
5  ( 1 )  "*"    "*"   "*"    "*"   " " " "       " "
6  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       " "
7  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
8  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
9  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
10  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"
11  ( 1 ) "*"    "*"   "*"    "*"   "*" "*"       "*"
          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian
1  ( 1 )  " "        " "        " "            " "
2  ( 1 )  " "        " "        " "            " "
3  ( 1 )  "*"        " "        " "            " "
4  ( 1 )  "*"        " "        " "            " "
5  ( 1 )  "*"        " "        " "            " "
6  ( 1 )  "*"        " "        " "            " "
7  ( 1 )  "*"        " "        " "            " "
8  ( 1 )  "*"        " "        "*"            " "
9  ( 1 )  "*"        "*"        "*"            " "
10  ( 1 ) "*"        "*"        "*"            "*"
11  ( 1 ) "*"        "*"        "*"            "*"
```

Best Subset Selection Using Adjusted R^2

Example: `Credit` dataset

```
> regfit.full = regsubsets(Balance~., data=Credit, nvmax=11)

> summary(regfit.full)$adjr2
[1] 0.7452098 0.8744888 0.9494991 0.9531099 0.9535789 0.9539961
[7] 0.9540098 0.9539649 0.9539243 0.9538912 0.9538287
```



Best Subset Selection: Summary

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest <u>estimated</u> test error

---

Best Subset Selection: Summary

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   - a. Fit all $\binom{p}{k}$ models with k variables
   - b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest estimated test error

---

Lowest estimated test error:

- General: Lowest CV error
- Linear regression:
  - Highest adjusted R^2
  - Lowest AIC (Akaike information criterion)
  - Lowest C_p estimate
  - Lowest BIC (Bayesian information criterion)
  - Measure how well the model fits training data, while accounting/penalizing for #variables

Best Subset Selection: Summary

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest estimated test error

---

Lowest estimated test error:

- General: Lowest CV error
- Linear regression:
  - Highest adjusted R^2
  - Lowest AIC (Akaike information criterion)
  - Lowest C_p estimate
  - Lowest BIC (Bayesian information criterion)
  - Measure how well the model fits training data, while accounting/penalizing for #variables

Computational issue: enumerate all 2^p subsets of p variables

- $p = 10$, $2^p = 1024$; $p = 20$, $2^p \geq 10^6$; $p = 300$, $2^p \geq 10^{90}$

Best Subset Selection: Summary

Algorithm:

1. For $k = 0, 1, 2, \ldots, p$
   a. Fit all $\binom{p}{k}$ models with k variables
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_k$
2. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with lowest estimated test error

---

Lowest estimated test error:

- General: Lowest CV error
- Linear regression:
  - Highest adjusted R^2
  - Lowest AIC (Akaike information criterion)
  - Lowest C_p estimate
  - Lowest BIC (Bayesian information criterion)
  - Measure how well the model fits training data, while accounting/penalizing for #variables

Computational issue: enumerate all 2^p subsets of p variables

- $p = 10$, $2^p = 1024$; $p = 20$, $2^p \geq 10^6$; $p = 300$, $2^p \geq 10^{90}$
- Age of Earth: $\approx 10^{17}$ seconds.

Stepwise Selection

(ISLR 6.1.2)

Alternative methods for variable selection

- faster
- not guaranteed to find the exact best
- often find a good subset

Stepwise Selection

(ISLR 6.1.2)

Alternative methods for variable selection

- faster
- not guaranteed to find the exact best
- often find a good subset

- Forward selection: greedily add one variable at each step
- Backward selection: greedily remove one variable at each step

Forward Stepwise Selection

Idea: Each step, add the variable giving the greatest additional improvement

Algorithm:

1. $\mathcal{M}_0 =$ model with no variables
2. For $k = 0, 1, 2, \ldots, p - 1$
   - a. Consider all $p - k$ models that add one variable to $\mathcal{M}_k$
   - b. Among them, pick the one with largest R^2; call it $\mathcal{M}_{k+1}$
3. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

Forward Stepwise Selection

Idea: Each step, add the variable giving the greatest additional improvement

Algorithm:

1. $\mathcal{M}_0 =$ model with no variables
2. For $k = 0, 1, 2, \ldots, p-1$
   a. Consider all $p-k$ models that add one variable to $\mathcal{M}_k$
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_{k+1}$
3. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

- Fit 1 null model...
  plus $p + (p-1) + (p-2) + (p-3) + \ldots + 1$ models.

Backward Stepwise Selection

Backward Stepwise Selection

Idea: Each step, remove the least useful variable

Backward Stepwise Selection

Idea: Each step, remove the least useful variable

Algorithm:

1. $\mathcal{M}_p =$ model with all p variables
2. For $k = p, p-1, \ldots, 1$
   - a. Consider all k models that remove one variable from $\mathcal{M}_k$
   - b. Among them, pick the one with largest R^2; call it $\mathcal{M}_{k-1}$
3. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

Backward Stepwise Selection

Idea: Each step, remove the least useful variable

Algorithm:

1. $\mathcal{M}_p$ = model with all p variables
2. For $k = p, p-1, \ldots, 1$
   a. Consider all k models that remove one variable from $\mathcal{M}_k$
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_{k-1}$
3. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

- Fit 1 full model...
  plus $p + (p-1) + (p-2) + (p-3) + \ldots + 1$ models.

Backward Stepwise Selection

Idea: Each step, remove the least useful variable

Algorithm:

1. $\mathcal{M}_p$ = model with all p variables
2. For $k = p, p-1, \ldots, 1$
   a. Consider all k models that remove one variable from $\mathcal{M}_k$
   b. Among them, pick the one with largest R^2; call it $\mathcal{M}_{k-1}$
3. Among $\mathcal{M}_0, \ldots, \mathcal{M}_p$, pick the one with largest adjusted R^2

- Fit 1 full model...
  plus $p + (p-1) + (p-2) + (p-3) + \ldots + 1$ models.
- Require $n > p$: Least squares solution not unique!

Forward Selection Using Adjusted R^2

Example: `Credit` dataset

- $p = 11$ predictors, $n = 400$ data points

(see ISLR 6.5.2 for R tutorial)

```
> regfit.fwd = regsubsets(Balance~., data=Credit,
+ nvmax=11, method = "forward")

> summary(regfit.fwd)
```

```
          Income Limit Rating Cards Age Education GenderFemale
1  ( 1 )  " "    " "   "*"    " "   " " " "       " "
2  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
3  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
4  ( 1 )  "*"    "*"   "*"    " "   " " " "       " "
5  ( 1 )  "*"    "*"   "*"    "*"   " " " "       " "
6  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       " "
7  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
8  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
9  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
10  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"
11  ( 1 ) "*"    "*"   "*"    "*"   "*" "*"       "*"
          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian
1  ( 1 )  " "        " "        " "            " "
2  ( 1 )  " "        " "        " "            " "
3  ( 1 )  "*"        " "        " "            " "
4  ( 1 )  "*"        " "        " "            " "
5  ( 1 )  "*"        " "        " "            " "
6  ( 1 )  "*"        " "        " "            " "
7  ( 1 )  "*"        " "        " "            " "
8  ( 1 )  "*"        " "        "*"            " "
9  ( 1 )  "*"        "*"        "*"            " "
10  ( 1 ) "*"        "*"        "*"            "*"
11  ( 1 ) "*"        "*"        "*"            "*"
```

Backward Selection Using Adjusted R^2

Example: `Credit` dataset

- $p = 11$ predictors, $n = 400$ data points

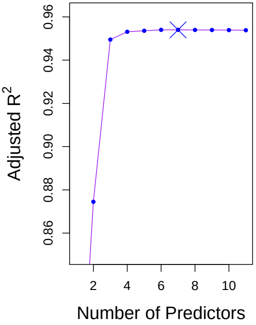
```
> regfit.bwd = regsubsets(Balance∼., data=Credit,
+ nvmax=11, method = "backward")

> summary(regfit.bwd)
```

```
          Income Limit Rating Cards Age Education GenderFemale
1  ( 1 )  " "    " "   "*"    " "   " " " "       " "
2  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
3  ( 1 )  "*"    " "   "*"    " "   " " " "       " "
4  ( 1 )  "*"    "*"   "*"    " "   " " " "       " "
5  ( 1 )  "*"    "*"   "*"    "*"   " " " "       " "
6  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       " "
7  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
8  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
9  ( 1 )  "*"    "*"   "*"    "*"   "*" " "       "*"
10  ( 1 ) "*"    "*"   "*"    "*"   "*" " "       "*"
11  ( 1 ) "*"    "*"   "*"    "*"   "*" "*"       "*"
          StudentYes MarriedYes EthnicityAsian EthnicityCaucasian
1  ( 1 )  " "        " "        " "            " "
2  ( 1 )  " "        " "        " "            " "
3  ( 1 )  "*"        " "        " "            " "
4  ( 1 )  "*"        " "        " "            " "
5  ( 1 )  "*"        " "        " "            " "
6  ( 1 )  "*"        " "        " "            " "
7  ( 1 )  "*"        " "        " "            " "
8  ( 1 )  "*"        " "        "*"            " "
9  ( 1 )  "*"        "*"        "*"            " "
10  ( 1 ) "*"        "*"        "*"            "*"
11  ( 1 ) "*"        "*"        "*"            "*"
```

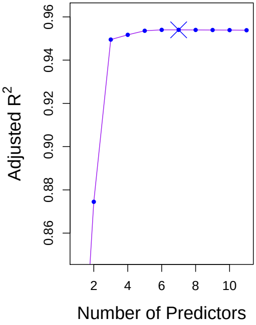
Comparison: Best, Forward and Backward Selection

Example: `Credit` dataset

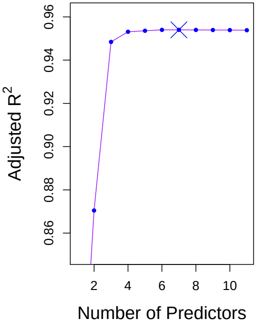
```
> summary(regfit.full)$adjr2
> summary(regfit.fwd)$adjr2
> summary(regfit.bwd)$adjr2
```



Best subset selection



Forward selection



Backward selection

Comparison: Best, Forward and Backward Selection

Example: `Credit` dataset

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income` `student`, `limit`	`rating`, `income`, `student`, `limit`

True or False

Can we apply the subset selection technique to logistic regression.

A. Yes the technique can be applied with small modifications.

B. No the technique only makes sense for linear regression.

We perform best subset, forward stepwise, and backward stepwise selection on the same training data set. In each approach, we obtain $p+1$ models containing $0, 1, 2, \ldots, p$ predictors, and a final model is picked among these $p+1$ models.

Which one of the following is always true?

A The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise.

B The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise.

C The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise.

D The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

Each of the three approach outputs a final model containing some subset of the predictors. Suppose that the *number* of predictors are the same in these three models. (The subsets of predictors are in general different.)

Which one of the following is true about these three final models?

- **A** The *training* RSS of best subset selection is always no higher than the other two.
- **B** The *training* RSS of forward selection is always no higher than those of the other two.
- **C** The *training* RSS of backward selection is always no higher than those of the other two.
- **D** It depends.

Under the setting of the last question, which one of the following is true about the three final models?

A The *test* RSS of best subset selection is always no higher than the other two.

B The *test* RSS of forward selection is always no higher than those of the other two.

C The *test* RSS of backward selection is always no higher than those of the other two.

D It depends.

Suppose you want to perform best subset selection with $p = 20$ variables. How many subsets will you consider?

A. 400

B. n^{20}

C. 1048576

D. n^2

E. 211

Suppose you want to perform forward selection with $p = 20$ variables. How many subsets will you consider?

A. 400

B. n^{20}

C. 1048576

D. n^2

E. 211

Summary

- Model selection for linear regression:
  Select a subset of predictors
  - Better interpretability
  - Not too flexible
  - Especially when p is large

Summary

- Model selection for linear regression:
  Select a subset of predictors
  - Better interpretability
  - Not too flexible
  - Especially when p is large

- Algorithms:
  - Best subset selection: optimal, but slow
  - Forward/backward stepwise selection: fast, but not optimal
  - Generally different outputs

Summary

- Model selection for linear regression:
  Select a subset of predictors
  - Better interpretability
  - Not too flexible
  - Especially when p is large

- Algorithms:
  - Best subset selection: optimal, but slow
  - Forward/backward stepwise selection: fast, but not optimal
  - Generally different outputs

- Do NOT use R^2/RSS to compare models with different #variables
  - Instead use adjusted R^2, AIC, BIC, C_p
  - Or cross-validation

https://dribbble.com/shots/3761660-Cowboy-lasso-smiley

Announcements

Recap: Linear Model Selection

n data points

p variables $X_1, X_2, \ldots, X_p$; p large

Recap: Linear Model Selection

n data points

p variables $X_1, X_2, \ldots, X_p$; p large

Linear model selection: Use $< p$ variables

- Goal: interpretability; controlled flexibility
- Algorithms: pick a subset to optimize adjusted R^2 (or AIC, BIC, CV)
  - Best subset selection: slow, optimal
  - Forward/backward selection: fast, suboptimal, works well

Recap: Linear Model Selection

n data points

p variables $X_1, X_2, \ldots, X_p$; p large

Linear model selection: Use $< p$ variables

- Goal: interpretability; controlled flexibility
- Algorithms: pick a subset to optimize adjusted R^2 (or AIC, BIC, CV)
  - Best subset selection: slow, optimal
  - Forward/backward selection: fast, suboptimal, works well

Today:

Linear model regularization: another way to control flexibility

- Fast, work well, optimal in some cases.

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Two types of regularization:

- Ridge regression: ℓ_2 regularization
- Lasso: ℓ_1 regularization

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Two types of regularization:

- Ridge regression: ℓ_2 regularization
- Lasso: ℓ_1 regularization

Have you used ridge regression?

A. Yes

B. No

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Two types of regularization:

- Ridge regression: ℓ_2 regularization
- Lasso: ℓ_1 regularization

Have you used Lasso?

A. Yes

B. No

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Two types of regularization:

- Ridge regression: ℓ_2 regularization
- Lasso: ℓ_1 regularization

When to use ridge?

A. When all predictors matter

B. When only some predictors matter

C. Always

Regularization in Linear Regression

(ISLR Sec 6.2)

Subset selection: Force some β_j to zero $\Rightarrow$ Hard selection

Regularization: Shrink β_j toward zero $\Rightarrow$ Soft selection

Principle: Smaller β_j $\Rightarrow$ Less flexibility/freedom $\Rightarrow$ Smaller variance

Two types of regularization:

- Ridge regression: ℓ_2 regularization
- Lasso: ℓ_1 regularization

When to use Lasso?

A. When all predictors matter

B. When only some predictors matter

C. Always

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\mathrm{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge regression: minimize

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Note: β_0 not regularized

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge regression: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ: tuning parameter
- Note: β_0 not regularized

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge regression: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ: tuning parameter
- Note: β_0 not regularized
- $\lambda = 0$: Same as least squares (full model)

Which would you expect to be true?

A. $|\hat{\beta}_i|$ tends to increase with λ

B. $|\hat{\beta}_i|$ tends to decrease with λ

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge regression: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ: tuning parameter
- Note: β_0 not regularized
- $\lambda = 0$: Same as least squares (full model)
- $\lambda \to \infty$: All $\beta_j = 0$ for $j \geq 1$ (null model)

Which would you expect to be true?

A. $|\hat{\beta}_i|$ tends to increase with λ

B. $|\hat{\beta}_i|$ tends to decrease with λ

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2$$

Ridge regression: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ: tuning parameter
- Note: β_0 not regularized
- $\lambda = 0$: Same as least squares (full model)
- $\lambda \to \infty$: All $\beta_j = 0$ for $j \geq 1$ (null model)
- Intermediate λ: encourage β_j's to be smaller (than the LS solution)

Which would you expect to be true?

A. variance tends to increase with λ

B. variance tends to decrease with λ

Ridge Regression (ISLR 6.2.1)

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Ridge regression: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

- λ: tuning parameter
- Note: β_0 not regularized
- $\lambda = 0$: Same as least squares (full model)
- $\lambda \to \infty$: All $\beta_j = 0$ for $j \geq 1$ (null model)
- Intermediate λ: encourage β_j's to be smaller (than the LS solution)

Which would you expect to be true?

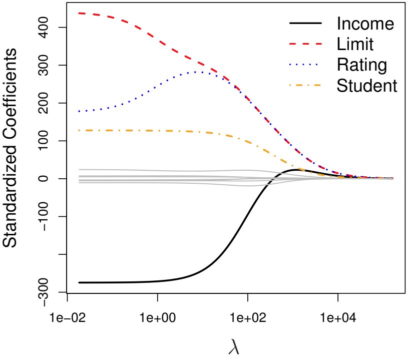
A. bias tends to increase with λ

B. bias tends to decrease with λ

Ridge Regression

Example: `Credit` dataset

$p = 11$ predictors: `Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married, EthnicityAsian, EthnicityCaucasian`



Scaling

Recall: Least squares approach to linear regression: minimize

$$\text{RSS} \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Call minimizer $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Suppose I take the data set and scale the 1st predictor by 2 and refit a least squares model. In other words, suppose I minimize

$$\text{RSS}_2 \triangleq \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 (2x_{i1}) - \sum_{j=2}^{p} \beta_j x_{ij} \right)^2$$

and get estimate $\hat{\beta}'_0, \hat{\beta}'_1, \ldots, \hat{\beta}'_p$, what can I conclude:

A. $\hat{\beta}'_1 = 2\hat{\beta}_1$

B. $\hat{\beta}'_1 = (1/2)\hat{\beta}_1$

C. Neither

Scaling

Recall: Least squares approach to ridge regression: minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

Call minimizer $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Suppose I take the data set and scale the 1st predictor by 2 and refit a ridge regression model. In other words, suppose I minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1(2x_{i1}) + \sum_{j=2}^{p}\beta_j x_{ij}\right)^2 - \lambda\sum_{j=1}^{p}\beta_j^2$$

and get estimate $\hat{\beta}'_0, \hat{\beta}'_1, \ldots, \hat{\beta}'_p$, what can I conclude:

A. $\hat{\beta}'_1 = 2\hat{\beta}_1$

B. $\hat{\beta}'_1 = (1/2)\hat{\beta}_1$

C. Neither

Ridge Regression: Computation

- Ridge regression very sensitive to coefficient scaling!
- Always standardize (center and normalize) the predictors.

(Done automatically using the following commands)

Ridge Regression: Computation

- Ridge regression very sensitive to coefficient scaling!
- Always standardize (center and normalize) the predictors.

(Done automatically using the following commands)

- Apply ridge regression in **R** (cf. ISLR 6.6)

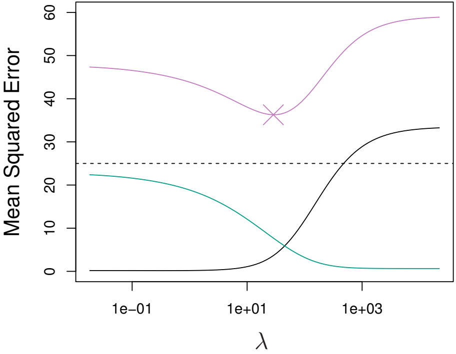
```
> library(glmnet)

> x = model.matrix(Balance~., Credit)

> ridge.fit = glmnet(x, Credit$Balance, alpha = 0, lambda = 0.1)
```

- Computational cost: no more than LS
- Much faster than best subset selection
- Need to choose λ (later)

Ridge Regression vs. Least Squares

Simulated data: $p = 45$, $n = 50$

$$\text{minimize} \;\; \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

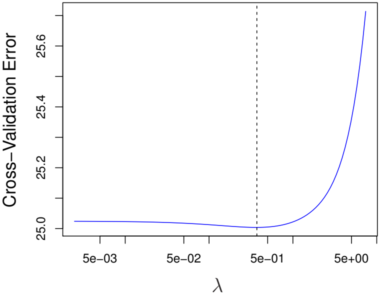


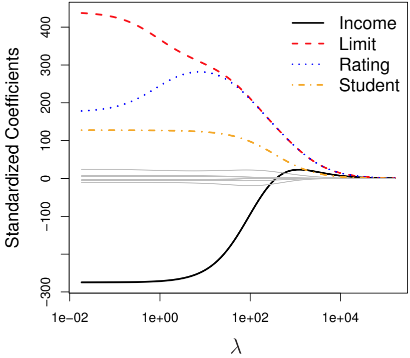
Choosing λ

λ: controls flexibility of ridge regression

- Choose λ to minimize test error
- Estimate test error by cross-validation
- After λ is chosen, refit model using all data.

`Credit` dataset





Lasso (Least Absolute Shrinkage and Selection Operator)

(ISLR 6.2.2)

Lasso (Least Absolute Shrinkage and Selection Operator)

(ISLR 6.2.2)

Recall:

Least squares	Ridge regression
min RSS $:= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$	min RSS $+ \lambda \sum_{j=1}^{p} \beta_j^2$

Lasso (Least Absolute Shrinkage and Selection Operator)

(ISLR 6.2.2)

Recall:

Least squares	Ridge regression
$\min \ \text{RSS} := \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$	$\min \ \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$

Lasso: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso (Least Absolute Shrinkage and Selection Operator)

(ISLR 6.2.2)

Recall:

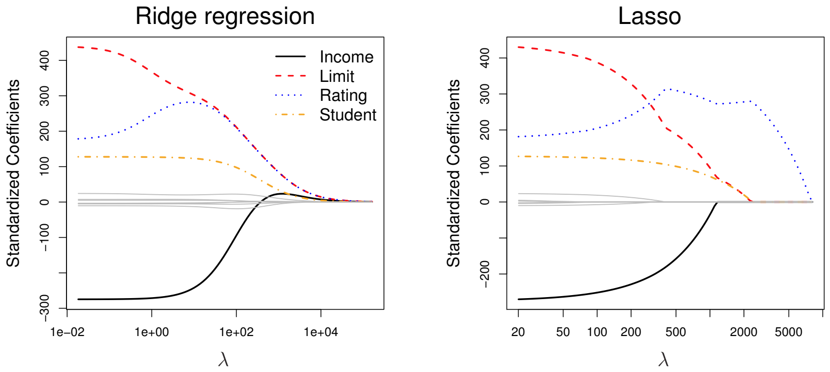
Least squares	Ridge regression
$\min\ \text{RSS} := \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$	$\min\ \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$

Lasso: minimize

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Larger $\lambda \Rightarrow$ more shrinkage of β_j's
- "ℓ_1" instead of "ℓ_2" regularization
- Key property: Some β_j will be shrunk to exactly zero

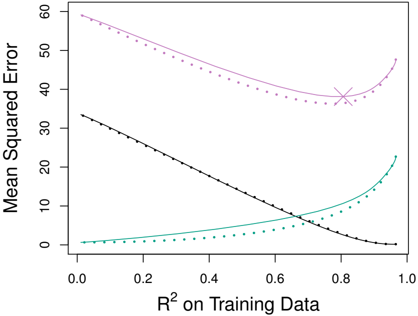
Ridge Regression vs. Lasso



Ridge Regression vs. Lasso

Simulated data 1:

Response depends on all $p = 45$ predictors



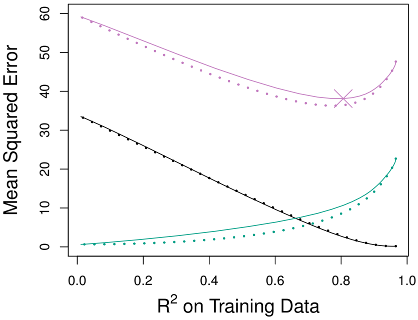
Solid: Lasso   Dash: Ridge

Test MSE   Bias   Variance

Ridge Regression vs. Lasso

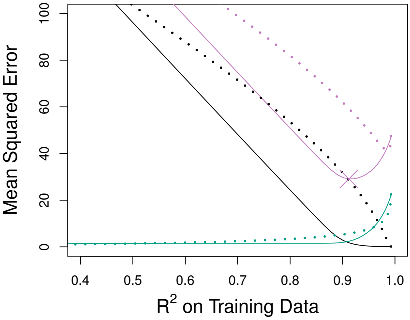
Simulated data 1:

Response depends on all $p = 45$ predictors



Simulated data 2:

Response depends on 2 out of $p = 45$ predictors



Solid: Lasso   Dash: Ridge

Test MSE   Bias   Variance

Lasso: Computation

- Always standardize (center & normalize) the predictors

Done automatically using `glmnet()`

Lasso: Computation

- Always standardize (center & normalize) the predictors

Done automatically using `glmnet()`

- Lasso in **R** (cf. ISLR 6.6)

```
> x = model.matrix(Balance~., Credit)

> ridge.fit = glmnet(x, Credit$Balance, alpha = 1, lambda = 0.5)
```

  - Computational cost: no more than LS
  - Much faster than best subset selection

Summary

Regularization for linear regression:

$$\text{minimize} \quad \text{RSS} + \lambda \times (\text{Regularization Term})$$

- Shrink β_j's towards zero
- Smaller β_j $\Rightarrow$ Less flexibility $\Rightarrow$ Lower variance

Summary

Regularization for linear regression:

$$\text{minimize} \quad \text{RSS} + \lambda \times (\text{Regularization Term})$$

- Shrink β_j's towards zero
- Smaller β_j $\Rightarrow$ Less flexibility $\Rightarrow$ Lower variance

	Ridge regression	Lasso
Regularization:	$\sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \lvert\beta_j\rvert$

Summary

Regularization for linear regression:

$$\text{minimize} \quad \text{RSS} + \lambda \times (\text{Regularization Term})$$

- Shrink β_j's towards zero
- Smaller β_j $\Rightarrow$ Less flexibility $\Rightarrow$ Lower variance

	Ridge regression	Lasso
Regularization:	$\sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \lvert\beta_j\rvert$
Property:	All β_j become smaller	Some β_j will be exactly zero

Summary

Regularization for linear regression:

$$\text{minimize} \quad \text{RSS} + \lambda \times (\text{Regularization Term})$$

- Shrink β_j's towards zero
- Smaller β_j $\Rightarrow$ Less flexibility $\Rightarrow$ Lower variance

	Ridge regression	Lasso
Regularization:	$\sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \lvert\beta_j\rvert$
Property:	All β_j become smaller	Some β_j will be exactly zero
Suitable when y	depends on all predictors	depends on a few predictors

Summary

Regularization for linear regression:

$$\text{minimize} \quad \text{RSS} + \lambda \times (\text{Regularization Term})$$

- Shrink β_j's towards zero
- Smaller $\beta_j \Rightarrow$ Less flexibility $\Rightarrow$ Lower variance

	Ridge regression	Lasso
Regularization:	$\sum_{j=1}^{p} \beta_j^2$	$\sum_{j=1}^{p} \lvert\beta_j\rvert$
Property:	All β_j become smaller	Some β_j will be exactly zero
Suitable when y	depends on all predictors	depends on a few predictors

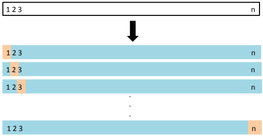
- Choose λ by Cross Validation
- Can easily compute RR (or Lasso) simultaneously for all values of λ

How do we use LOOCV to estimate true test error when training involves regularization?

- Split n data points into:
  - a training set of $n-1$ points
  - a validation set of 1 point
- Consider all n possible ways of splitting

Estimate test error by averaging:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{MSE}_i}_{\text{Error on Sample } i}$$

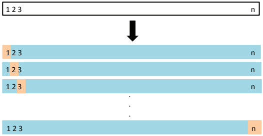


A.

- Split n data points into:
  - a training set of $n-1$ points
  - a validation set of 1 point
- Consider all n possible ways of splitting

Estimate test error by averaging:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{MSE}_i}_{\text{Error on Sample } i} + \frac{\lambda}{n} \times \underbrace{(\text{Regularization Term}_i)}_{\text{Size of Reg on Sample } i}$$

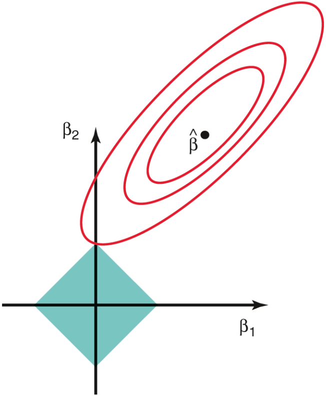


B.

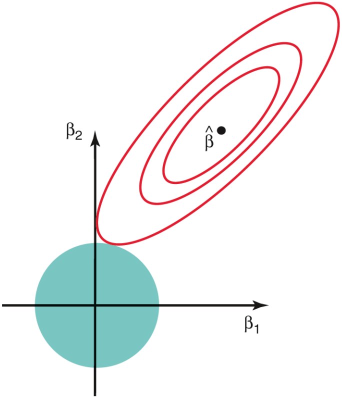
(Optional) The Geometry of Ridge and LASSO

ISLR pp 220 - 227

The Geometry of Ridge and Lasso



$|\beta_1| + |\beta_2| \leq s$



$|\beta_1|^2 + |\beta_2|^2 \leq s$

RSS Contours

Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani
Slides based on Yudong Chen’s slides.