Lasso regression

STAT 4710

October 17, 2023

Where we are

Unit 1: R for data mining

Unit 2: Prediction fundamentals

Unit 3: Regression-based methods

Unit 4: Tree-based methods

Unit 5: Deep learning

Lecture 1: Linear and logistic regression

Lecture 2: Regression in high dimensions

Lecture 3: Ridge regression

Lecture 4: Lasso regression

Lecture 5: Unit review and quiz in class

Idea: Encourage coefficients to be exactly zero

Idea: Encourage coefficients to be exactly zero

First, recall ridge regression:

$$\widehat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

Idea: Encourage coefficients to be exactly zero

First, recall ridge regression:

$$\widehat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

The penalty term biases coefficients toward zero, which reduces variance.

Idea: Encourage coefficients to be exactly zero

First, recall ridge regression:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

The penalty term biases coefficients toward zero, which reduces variance.

Another way to reduce variance is to use a different penalty:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

Idea: Encourage coefficients to be exactly zero

First, recall ridge regression:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

The penalty term biases coefficients toward zero, which reduces variance.

Another way to reduce variance is to use a different penalty:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

It turns out that changing the penalty in this way leads to $\hat{\beta}_j^{\text{lasso}} = 0$ for many j.

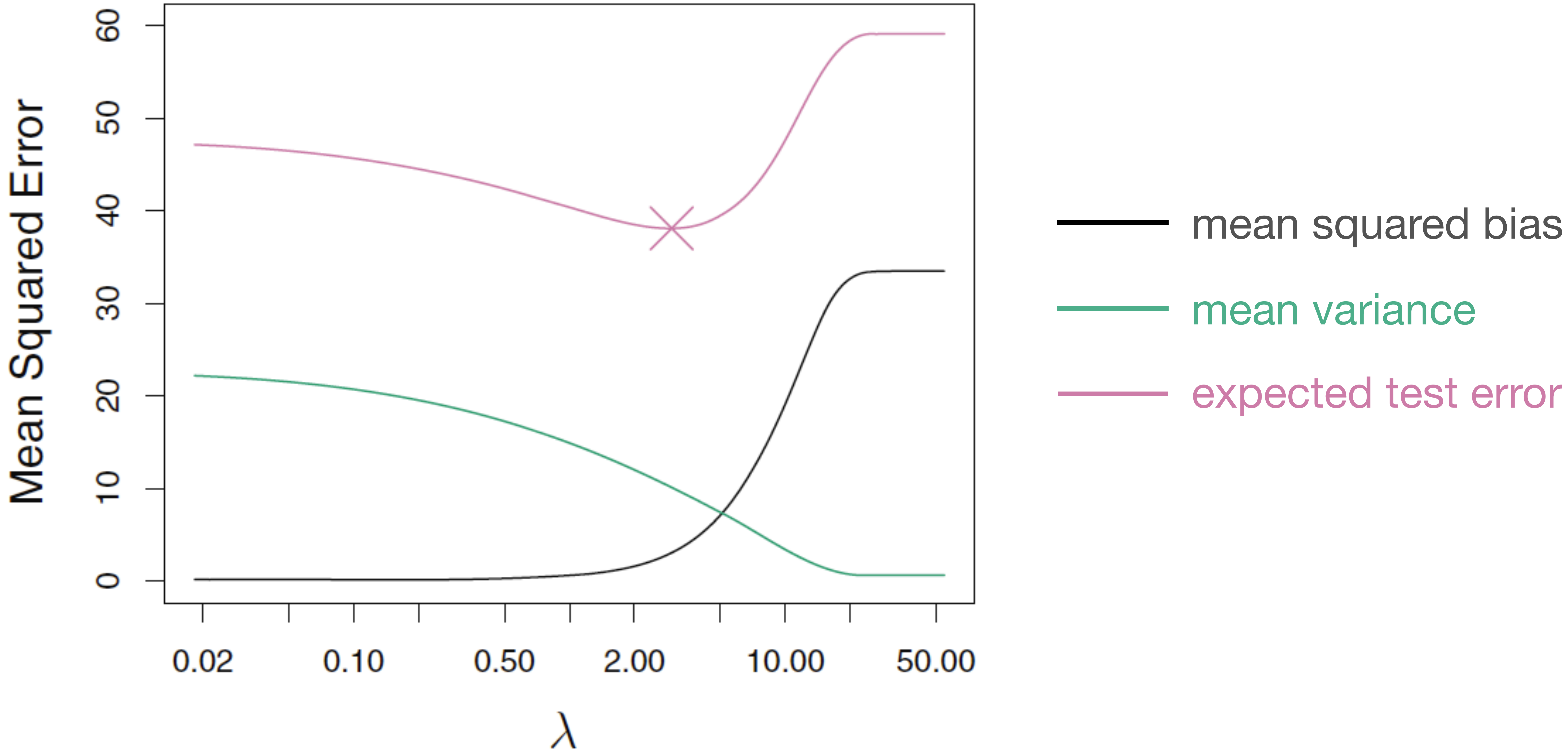
The effect of the penalty parameter λ

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

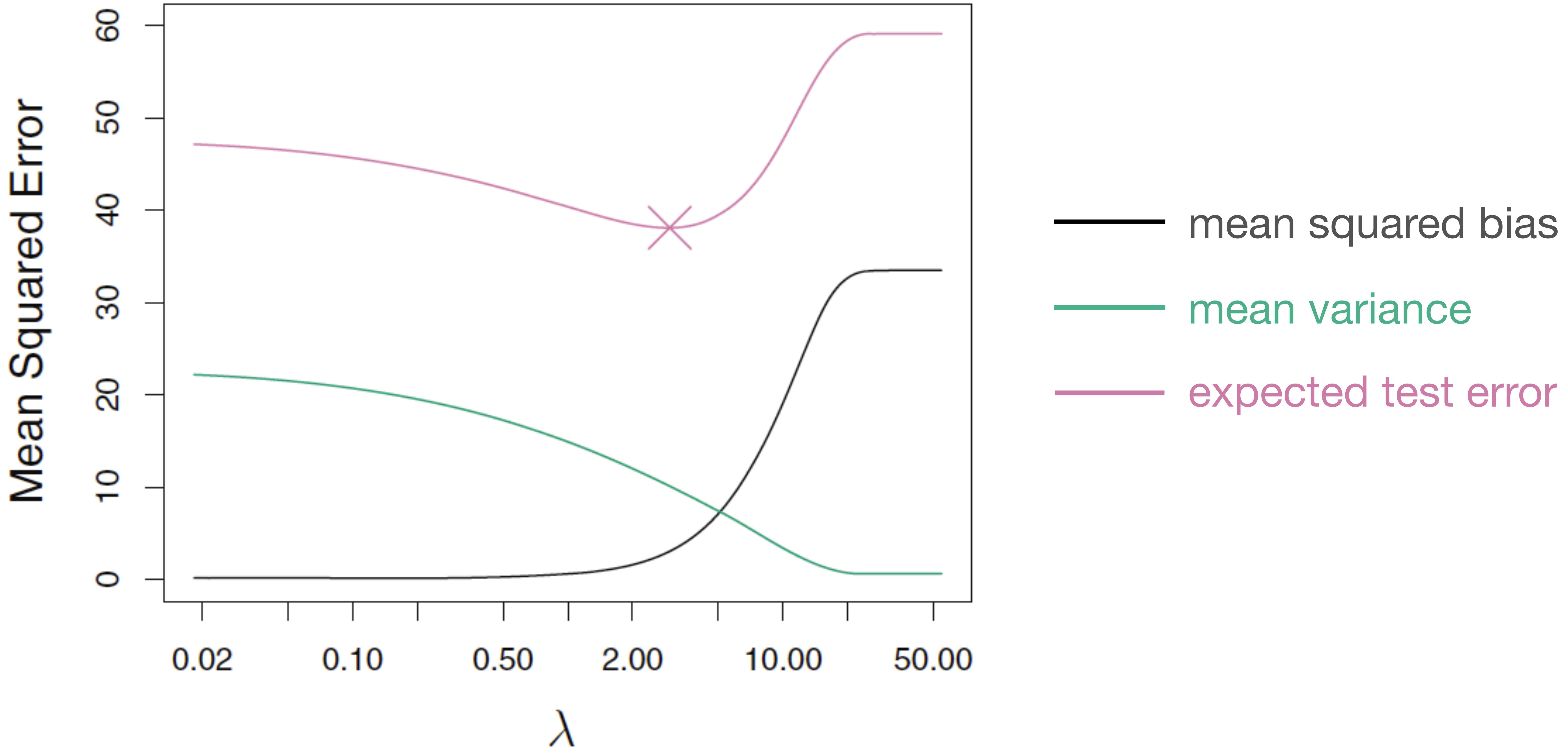
- The larger λ is, the more of a penalty there is.
- For $\lambda = 0$, we get back ordinary least squares (if OLS solution exists)
- For $\lambda = \infty$, we get $\beta_1 = \cdots = \beta_{p-1} = 0$, leaving only the intercept (which is not penalized).

We should think of λ as controlling the flexibility of the lasso regression fit, like the degrees of freedom in a spline fit. However, larger λ means fewer degrees of freedom.

The bias-variance tradeoff for lasso regression



The bias-variance tradeoff for lasso regression



In practice, λ is chosen by cross-validation.

Sparsity and interpretability

Sparsity and interpretability

Lasso solution $\widehat{\beta}^{\text{lasso}}$ is called sparse
because $\widehat{\beta}_j^{lasso} = 0$ for many j.

Sparsity and interpretability

Lasso solution $\widehat{\beta}^{\text{lasso}}$ is called sparse because $\widehat{\beta}_j^{lasso} = 0$ for many j.

Example: Crime Data

```
# A tibble: 97 x 2
   variable                coefficient
   <chr>                         <dbl>
 1 pct.kids.nvrmarried         85.2
 2 pct.pop.underpov            25.9
 3 male.pct.divorce            22.8
 4 pct.people.dense.hh         10.0
 5 pct.kids2parents            -5.51
 6 pct.youngkids2parents       -0.821
 7 num.kids.nvrmarried          0.00737
 8 population                   0
 9 household.size               0
10 race.pctblack                0
# … with 87 more rows
```

Sparsity and interpretability

Lasso solution $\widehat{\beta}^{\text{lasso}}$ is called sparse because $\widehat{\beta}_j^{\textit{lasso}} = 0$ for many j.

Lasso is therefore a variable selection tool.

Example: Crime Data

```
# A tibble: 97 x 2
   variable                coefficient
   <chr>                         <dbl>
 1 pct.kids.nvrmarried         85.2
 2 pct.pop.underpov            25.9
 3 male.pct.divorce            22.8
 4 pct.people.dense.hh         10.0
 5 pct.kids2parents            -5.51
 6 pct.youngkids2parents       -0.821
 7 num.kids.nvrmarried          0.00737
 8 population                   0
 9 household.size               0
10 race.pctblack                0
# … with 87 more rows
```

Sparsity and interpretability

Lasso solution $\widehat{\beta}^{\text{lasso}}$ is called sparse because $\widehat{\beta}_j^{lasso} = 0$ for many j.

Lasso is therefore a variable selection tool.

Sparse coefficient vectors are interpretable; they suggest which features are important.

Example: Crime Data

```
# A tibble: 97 x 2
   variable                coefficient
   <chr>                         <dbl>
 1 pct.kids.nvrmarried        85.2
 2 pct.pop.underpov           25.9
 3 male.pct.divorce           22.8
 4 pct.people.dense.hh        10.0
 5 pct.kids2parents           -5.51
 6 pct.youngkids2parents      -0.821
 7 num.kids.nvrmarried         0.00737
 8 population                  0
 9 household.size              0
10 race.pctblack               0
# … with 87 more rows
```

Sparsity and interpretability

Lasso solution $\widehat{\beta}^{\text{lasso}}$ is called sparse because $\widehat{\beta}_j^{lasso} = 0$ for many j.

Lasso is therefore a variable selection tool.

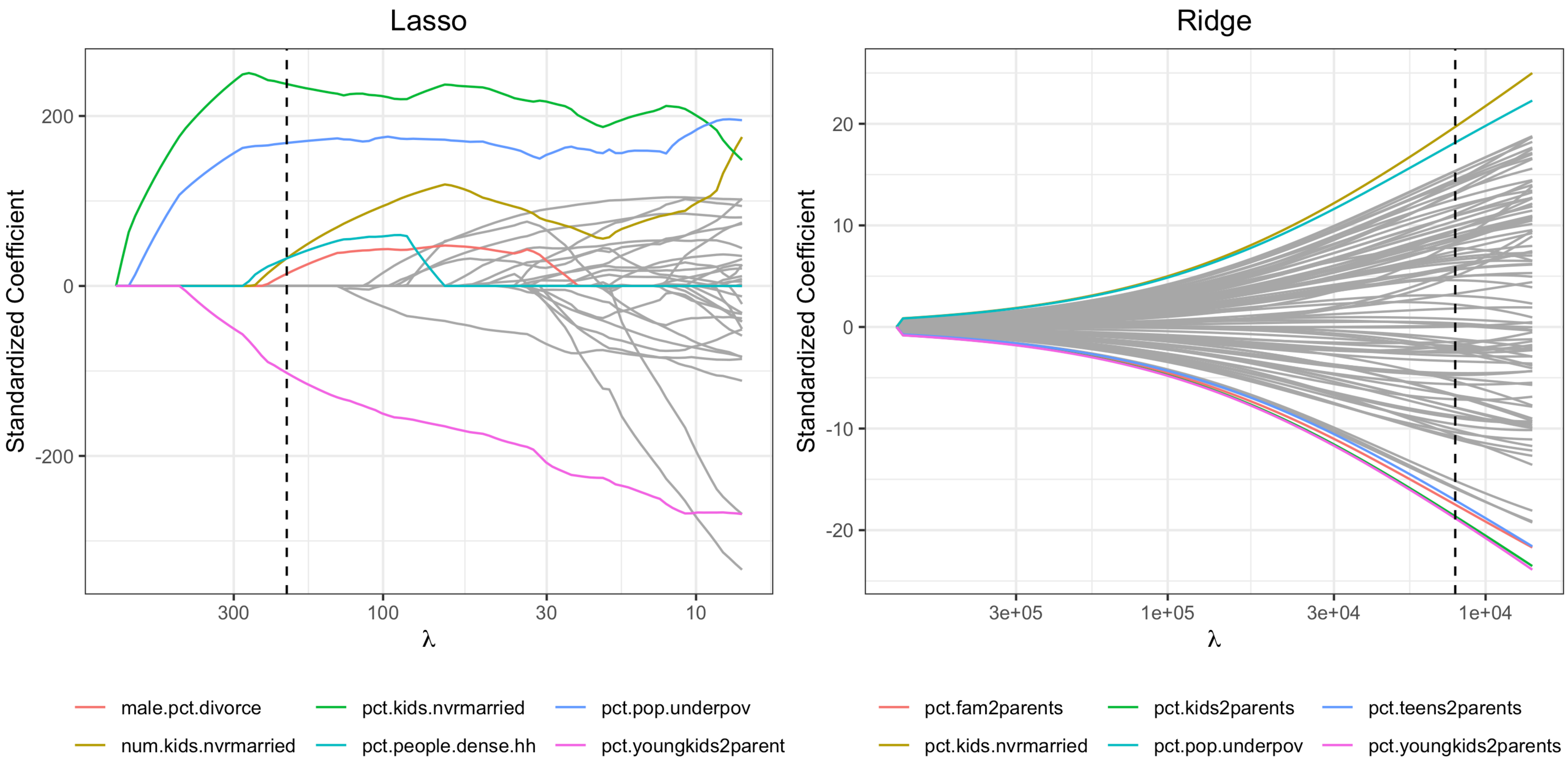
Sparse coefficient vectors are interpretable; they suggest which features are important.

NOTE: Cannot attach a measure of statistical significance to the selected variables.

Example: Crime Data

```
# A tibble: 97 x 2
   variable                coefficient
   <chr>                         <dbl>
 1 pct.kids.nvrmarried         85.2
 2 pct.pop.underpov            25.9
 3 male.pct.divorce            22.8
 4 pct.people.dense.hh         10.0
 5 pct.kids2parents            -5.51
 6 pct.youngkids2parents       -0.821
 7 num.kids.nvrmarried          0.00737
 8 population                   0
 9 household.size               0
10 race.pctblack                0
# … with 87 more rows
```

Lasso trace plot (compared to ridge)



Lasso regression in a simple case

Lasso regression in a simple case

Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

Lasso regression in a simple case

Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X =$

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\arg\min} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$

Lasso regression in a simple case

Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X =$

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\widehat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$

In this simple case, $\widehat{\beta}_j^{\text{OLS}} = Y_j$ and

Lasso regression in a simple case

Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X =$

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\widehat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$

In this simple case, $\widehat{\beta}_j^{\text{OLS}} = Y_j$ and

$$\widehat{\beta}_j^{\text{lasso}} = \begin{cases} Y_j - \lambda/2, & \text{if } Y_j \geq \lambda/2 \\ 0, & \text{if } |Y_j| \leq \lambda/2 \\ Y_j + \lambda/2, & \text{if } Y_j \leq -\lambda/2 \end{cases}$$

Lasso regression in a simple case

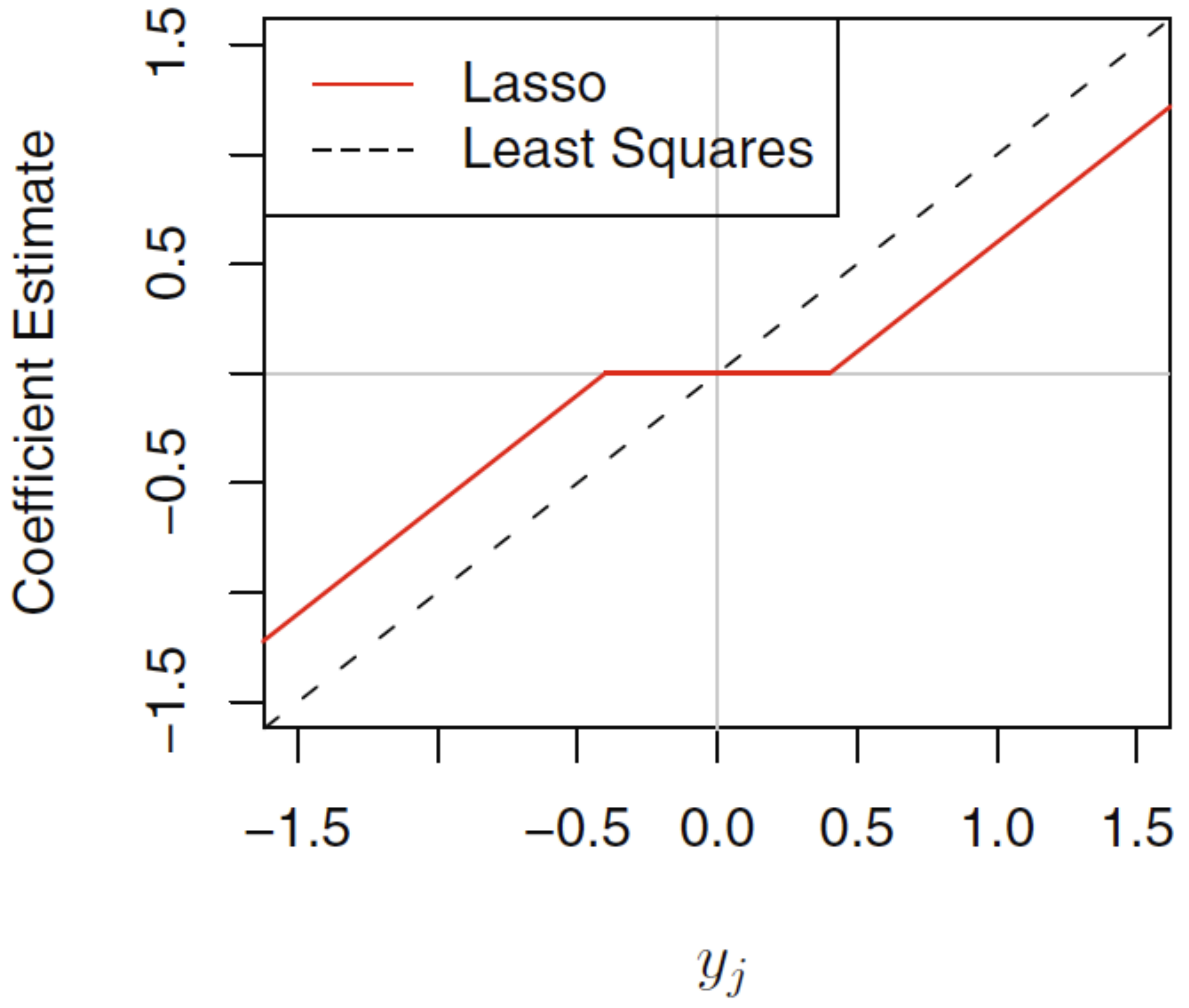
Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X =$

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$



In this simple case, $\hat{\beta}_j^{\text{OLS}} = Y_j$ and

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} Y_j - \lambda/2, & \text{if } Y_j \geq \lambda/2 \\ 0, & \text{if } |Y_j| \leq \lambda/2 \\ Y_j + \lambda/2, & \text{if } Y_j \leq -\lambda/2 \end{cases}$$

Lasso regression in a simple case

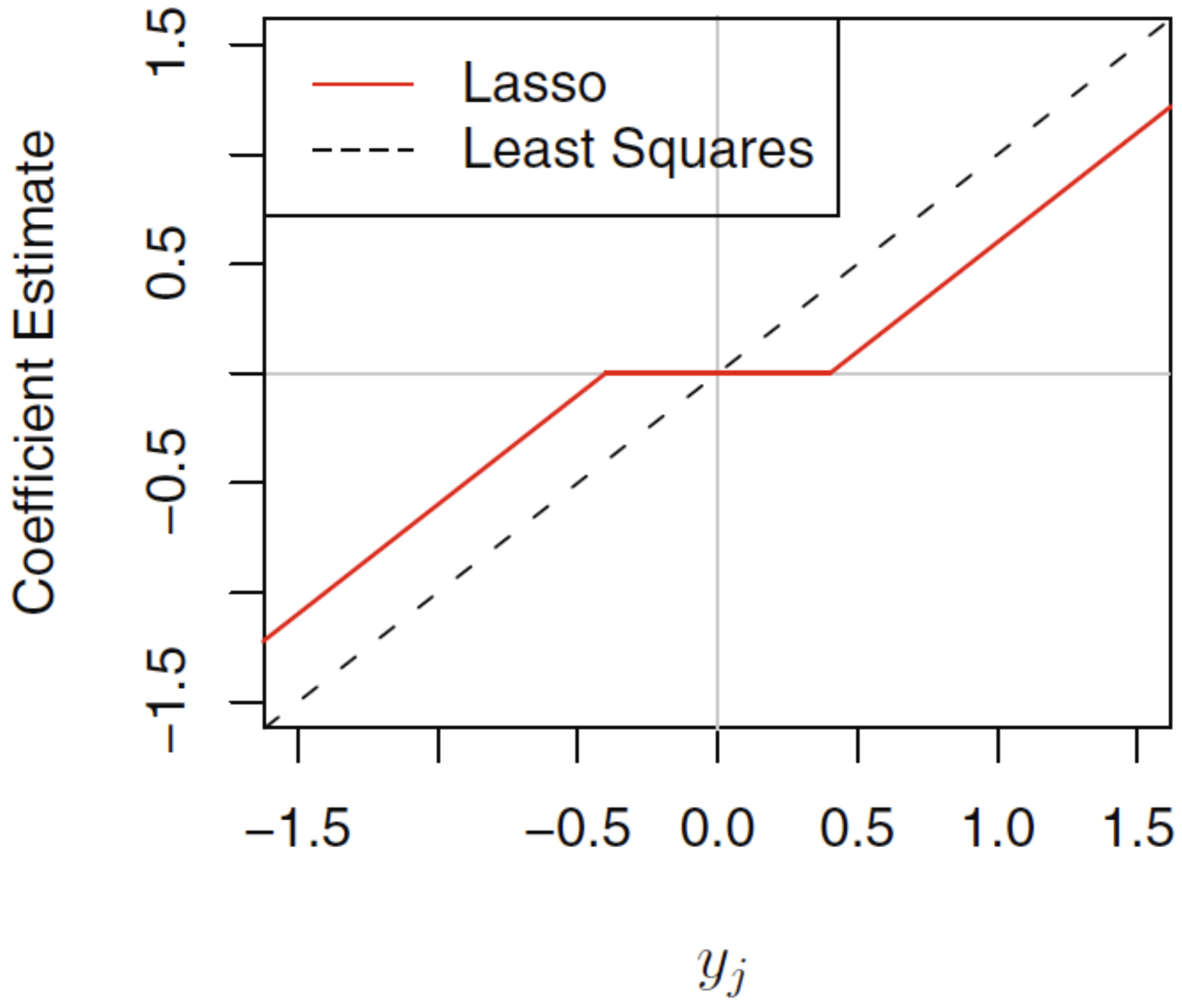
Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X =$

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$



In this simple case, $\hat{\beta}_j^{\text{OLS}} = Y_j$ and

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} Y_j - \lambda/2, & \text{if } Y_j \geq \lambda/2 \\ 0, & \text{if } |Y_j| \leq \lambda/2 \\ Y_j + \lambda/2, & \text{if } Y_j \leq -\lambda/2 \end{cases}$$

$\hat{\beta}^{\text{lasso}}$ obtained by *soft-thresholding* OLS estimate.

Lasso regression in a simple case

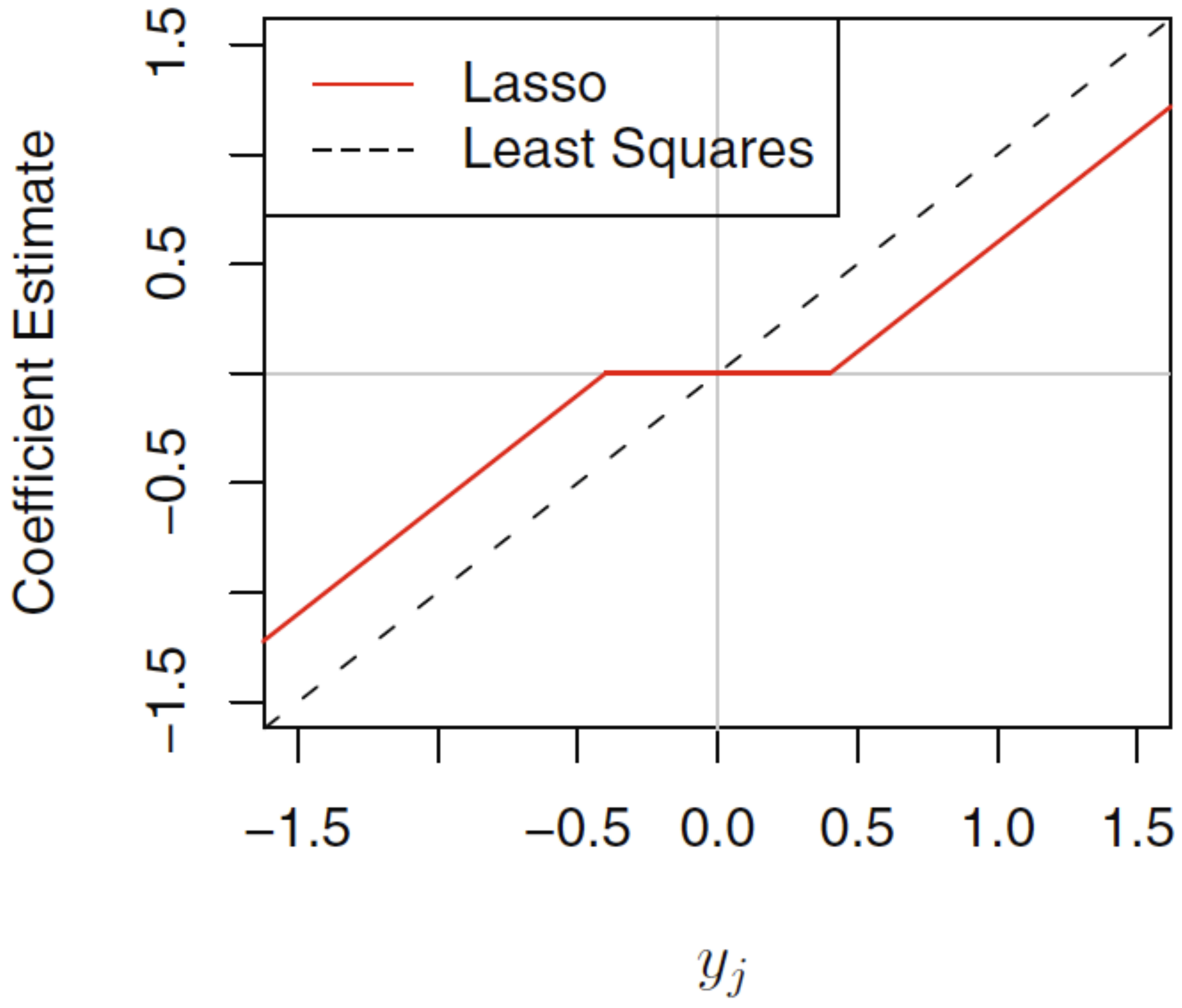
Suppose that $n = p$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$, i.e. $Y_j = \beta_j + \epsilon_j$.

E.g. $X = $

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting lasso regression without intercept:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{j=0}^{p-1} (Y_j - \beta_j)^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}.$$



In this simple case, $\hat{\beta}_j^{\text{OLS}} = Y_j$ and

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} Y_j - \lambda/2, & \text{if } Y_j \geq \lambda/2 \\ 0, & \text{if } |Y_j| \leq \lambda/2 \\ Y_j + \lambda/2, & \text{if } Y_j \leq -\lambda/2 \end{cases}$$

$\hat{\beta}^{\text{lasso}}$ obtained by *soft-thresholding* OLS estimate.

LASSO = Least Angle Shrinkage and Selection Operator.

Feature scaling and standardization

Like for ridge regression, feature scaling matters for the lasso;

Feature standardization is recommended before running the lasso.

Treatment of correlated features

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Lasso coefficients are also unstable for correlated features.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Lasso coefficients are also unstable for correlated features.

For example, consider the linear regression

$$y = \beta_1 X_1 + \beta_2 X_1 + \epsilon,$$

where we've accidentally added the same feature twice.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Lasso coefficients are also unstable for correlated features.

For example, consider the linear regression

$$y = \beta_1 X_1 + \beta_2 X_1 + \epsilon,$$

where we've accidentally added the same feature twice.

- Linear regression is undefined because (β_1, β_2) and $(\beta_1 - c, \beta_2 + c)$ give the same RSS for each c.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Lasso coefficients are also unstable for correlated features.

For example, consider the linear regression

$$y = \beta_1 X_1 + \beta_2 X_1 + \epsilon,$$

where we've accidentally added the same feature twice.

- Linear regression is undefined because (β_1, β_2) and $(\beta_1 - c, \beta_2 + c)$ give the same RSS for each c.
- The lasso penalty does not help "break the tie." In practice, lasso often chooses one of the two features arbitrarily.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Lasso coefficients are also unstable for correlated features.

For example, consider the linear regression

$$y = \beta_1 X_1 + \beta_2 X_1 + \epsilon,$$

where we've accidentally added the same feature twice.

- Linear regression is undefined because (β_1, β_2) and $(\beta_1 - c, \beta_2 + c)$ give the same RSS for each c.
- The lasso penalty does not help "break the tie." In practice, lasso often chooses one of the two features arbitrarily.

Note: Coefficient instability doesn't necessarily translate into prediction instability.

Logistic regression with lasso penalty

Logistic regression with lasso penalty

Logistic regression can be penalized, just like linear regression!

Logistic regression with lasso penalty

Logistic regression can be penalized, just like linear regression!

Recall $\mathscr{L}(\beta)$, the logistic regression likelihood. We can view $-\log \mathscr{L}(\beta)$ as analogous to the linear regression RSS. Continuing the analogy, we can define

$$\widehat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ -\log \mathscr{L}(\beta) + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

Logistic regression with lasso penalty

Logistic regression can be penalized, just like linear regression!

Recall $\mathscr{L}(\beta)$, the logistic regression likelihood. We can view $-\log \mathscr{L}(\beta)$ as analogous to the linear regression RSS. Continuing the analogy, we can define

$$\widehat{\beta}^{\text{lasso}} = \underset{\beta}{\arg\min} \left\{ -\log \mathscr{L}(\beta) + \lambda \sum_{j=1}^{p-1} |\beta_j| \right\}.$$

Subtle point: While $\widehat{\beta}^{\text{lasso}}$ is trained based on a (penalized) log-likelihood, during cross-validation we should choose λ based on whatever measure of test error we care about (e.g. weighted misclassification error).

Ridge versus lasso

Ridge versus lasso

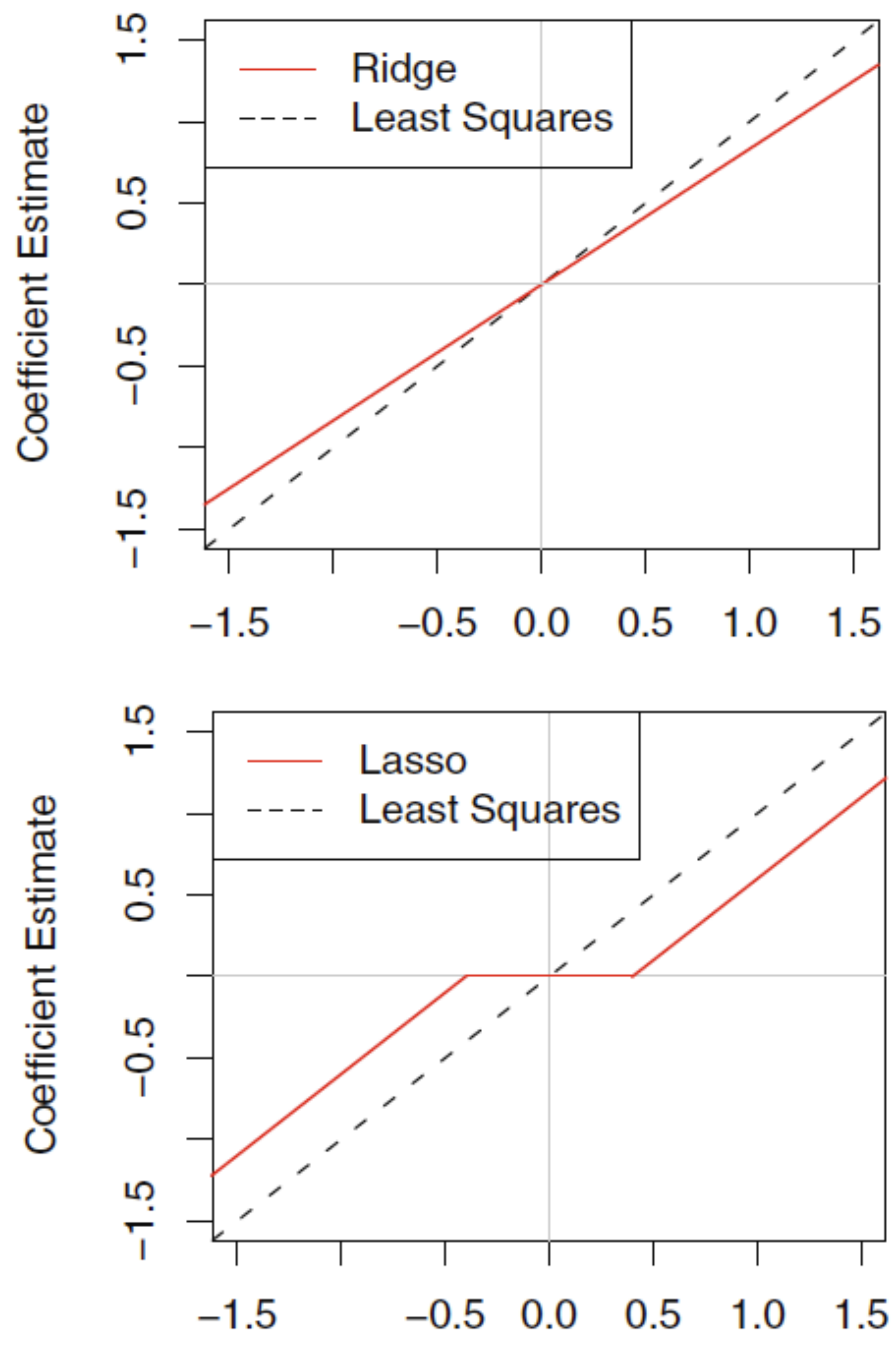
	Least squares	Ridge	Lasso

Ridge versus lasso

	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert\beta_j\rvert$

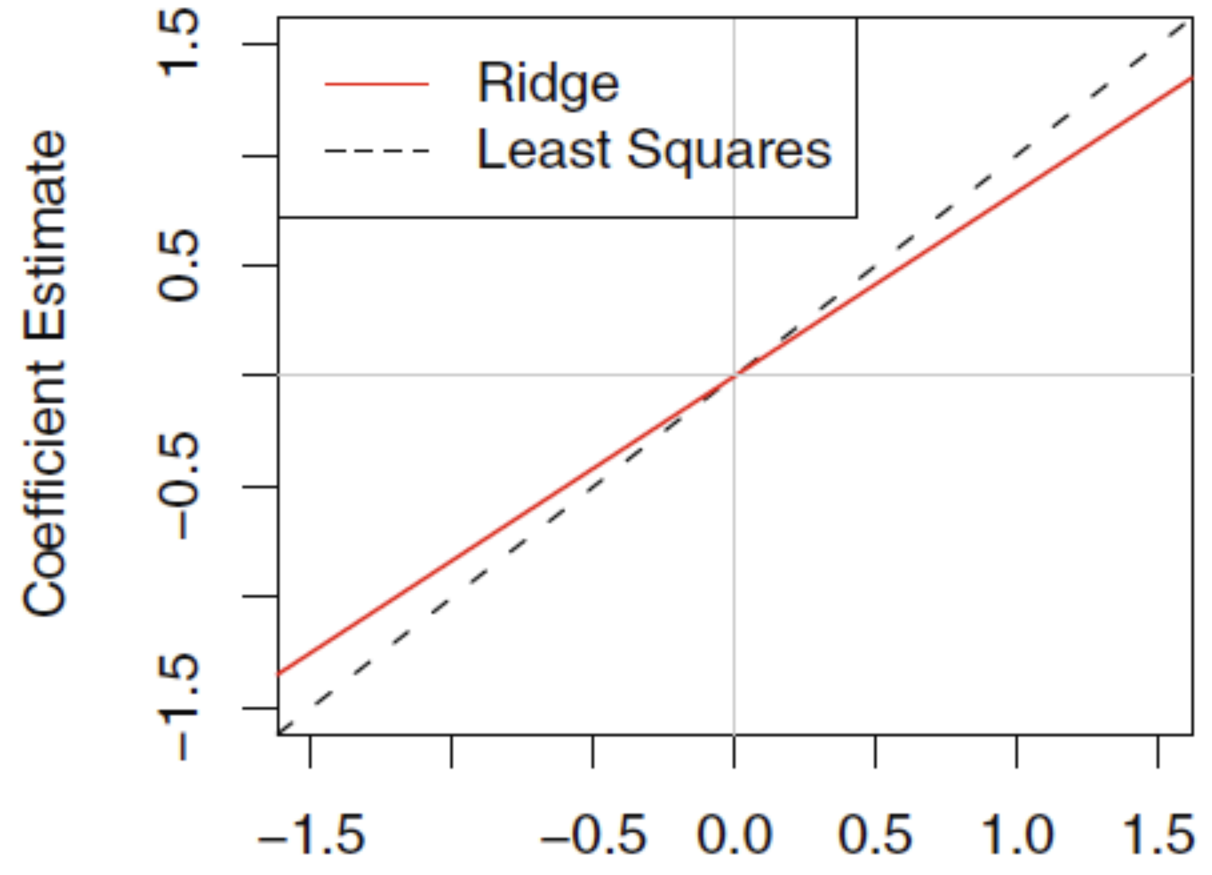
Ridge versus lasso

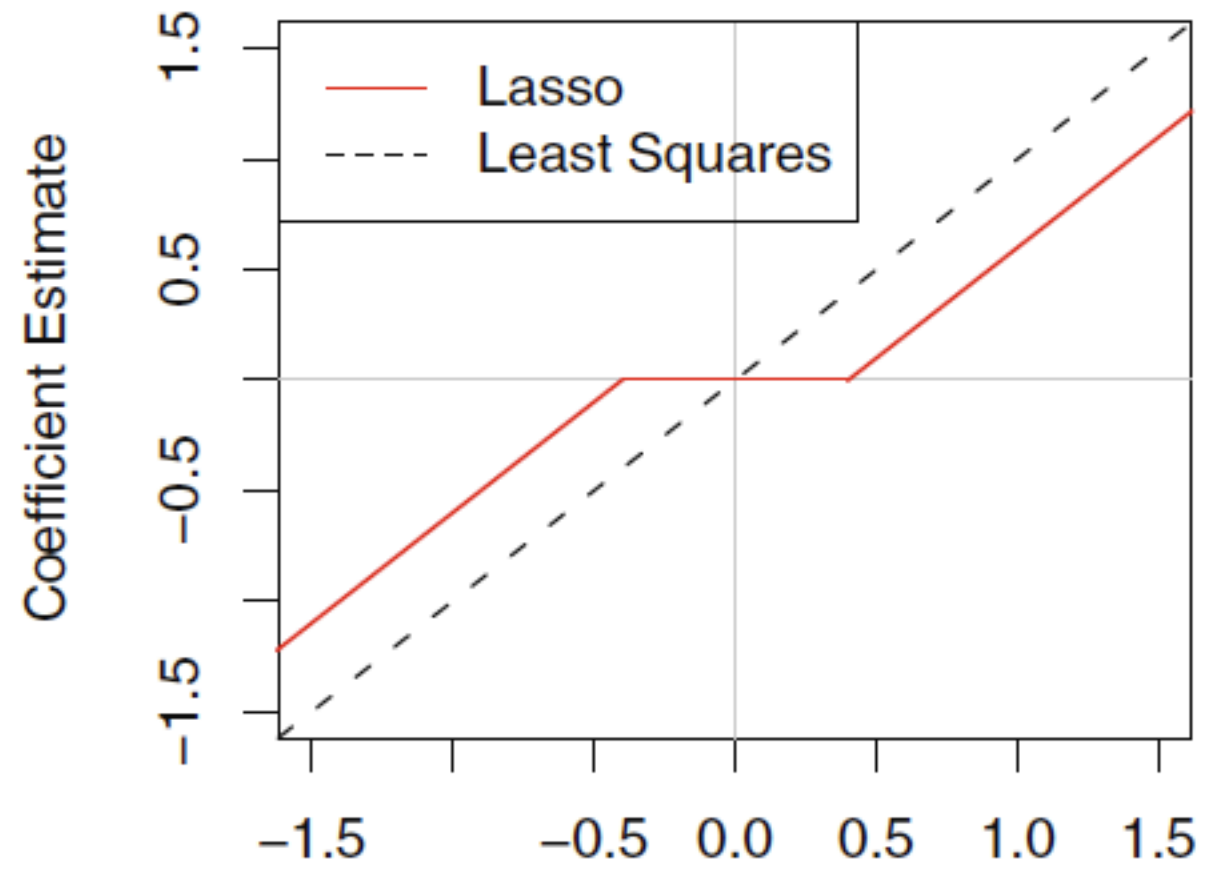
	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert \beta_j \rvert$
Penalty effect	N/A	Shrinkage	Shrinkage and selection



Ridge versus lasso

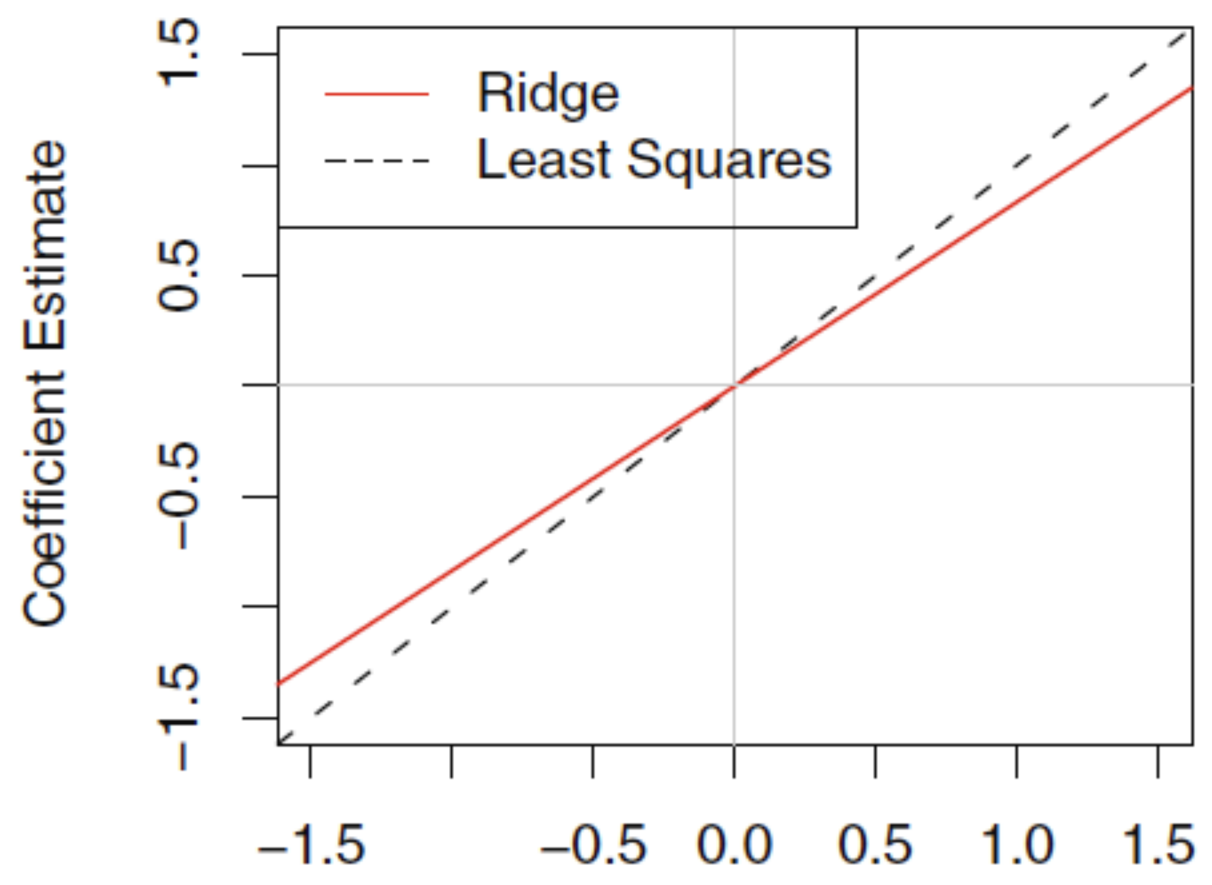
	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert\beta_j\rvert$
Penalty effect	N/A	Shrinkage	Shrinkage and selection
Sparsity	No	No	Yes

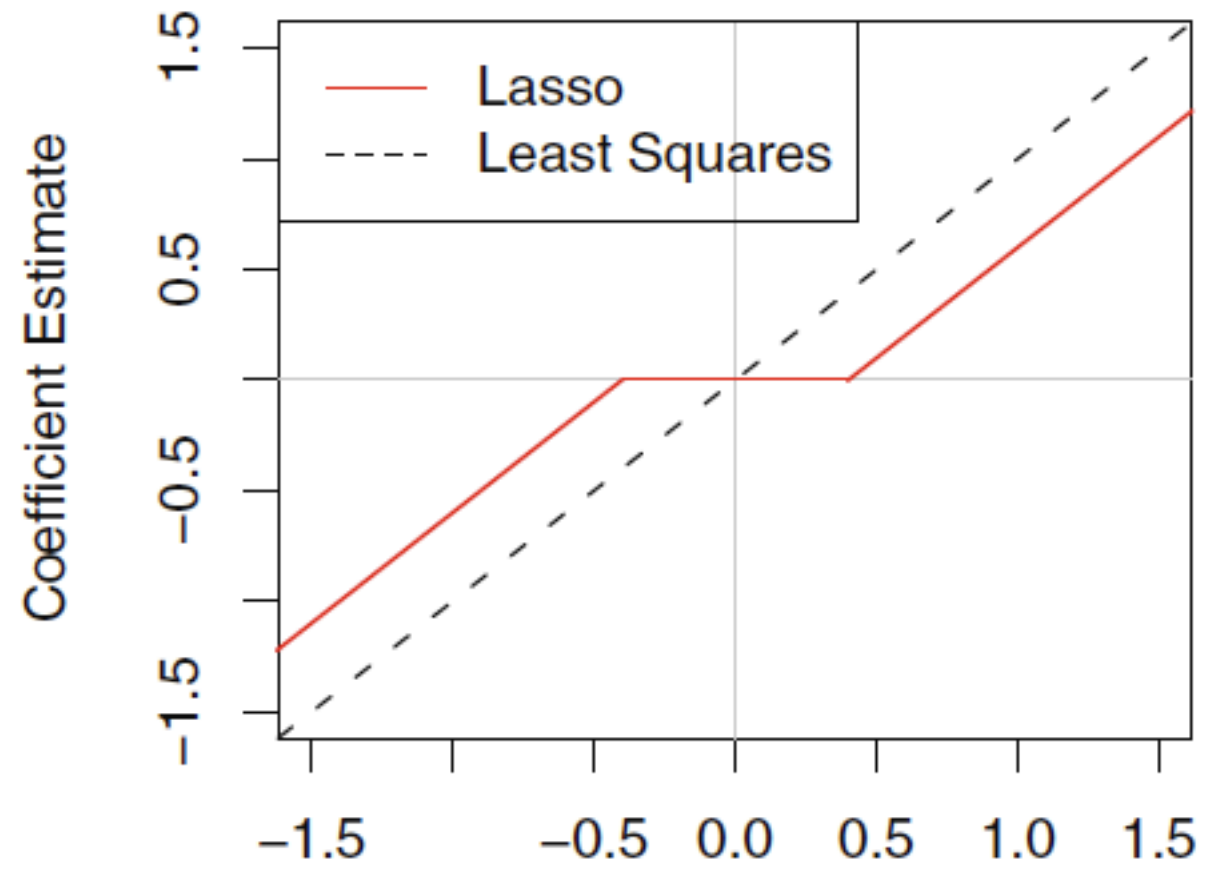




Ridge versus lasso

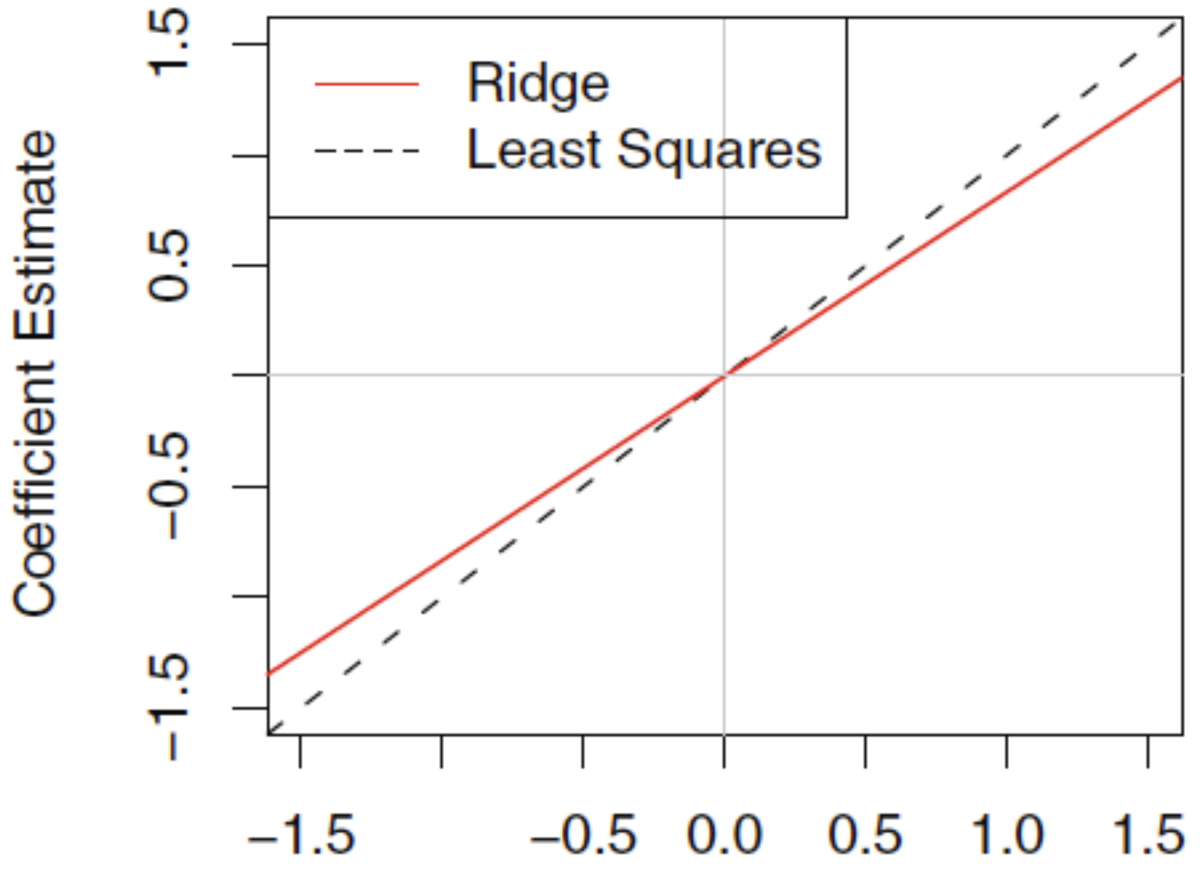
	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert \beta_j \rvert$
Penalty effect	N/A	Shrinkage	Shrinkage and selection
Sparsity	No	No	Yes
Correlated features	(Unstable)	Splits the credit (stable)	Chooses one arbitrarily (unstable)

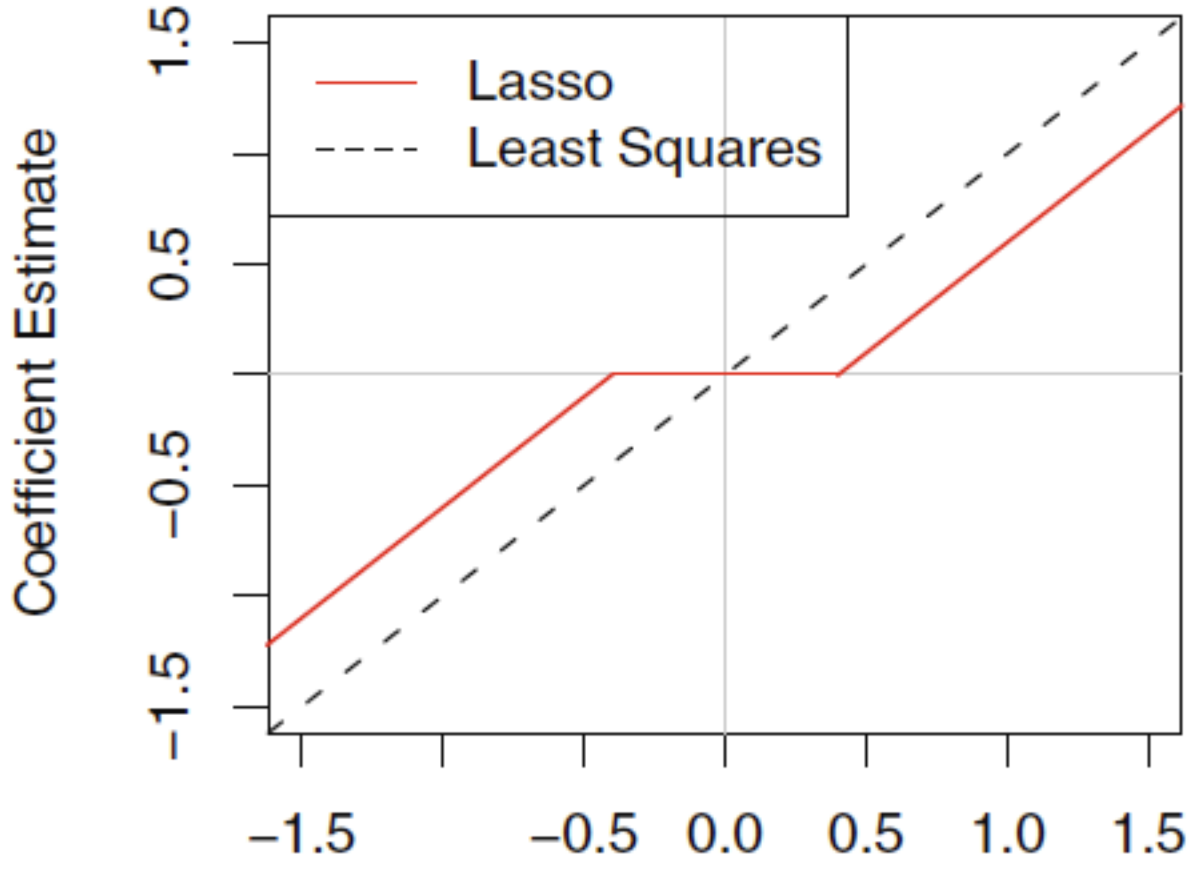




Ridge versus lasso

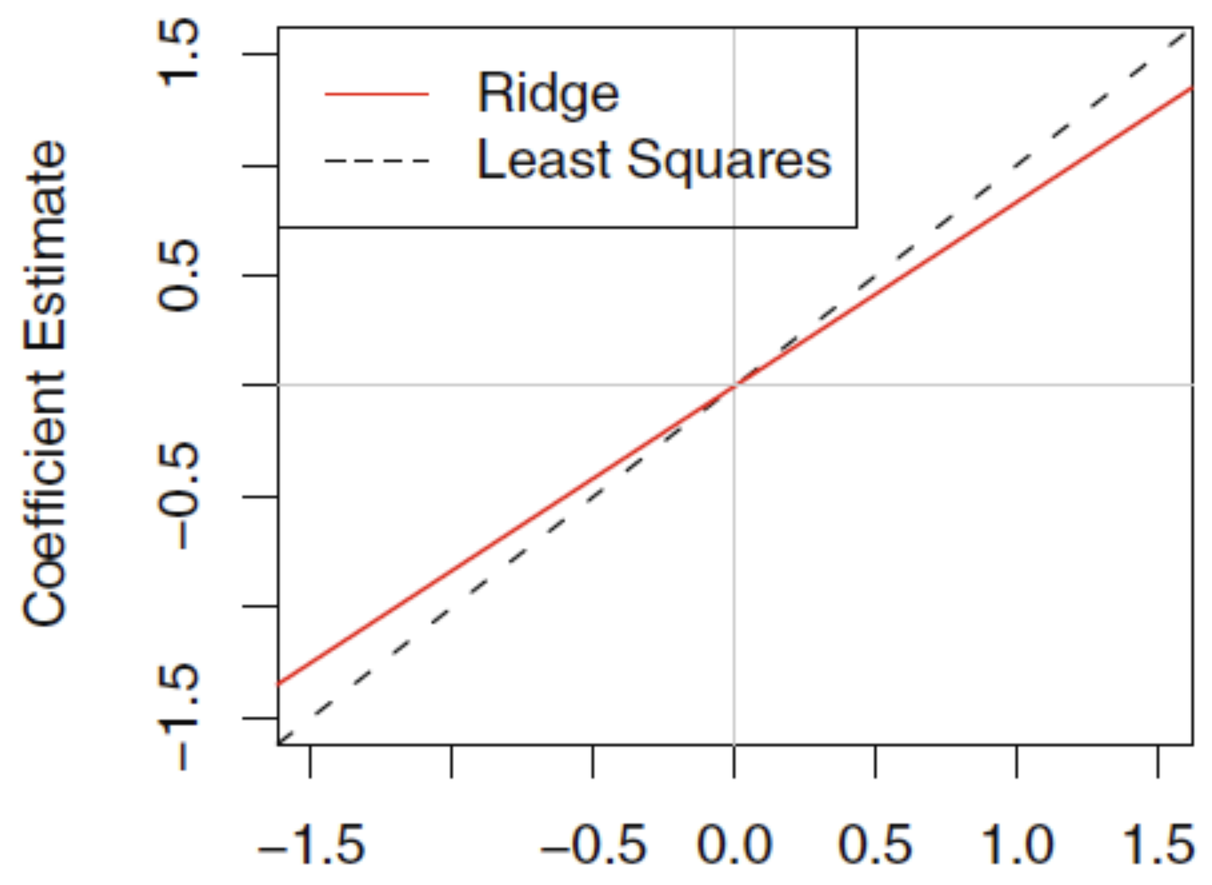
	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert \beta_j \rvert$
Penalty effect	N/A	Shrinkage	Shrinkage and selection
Sparsity	No	No	Yes
Correlated features	(Unstable)	Splits the credit (stable)	Chooses one arbitrarily (unstable)
Performs better when	n/p is large	Many features have small effects	Few features have large effects

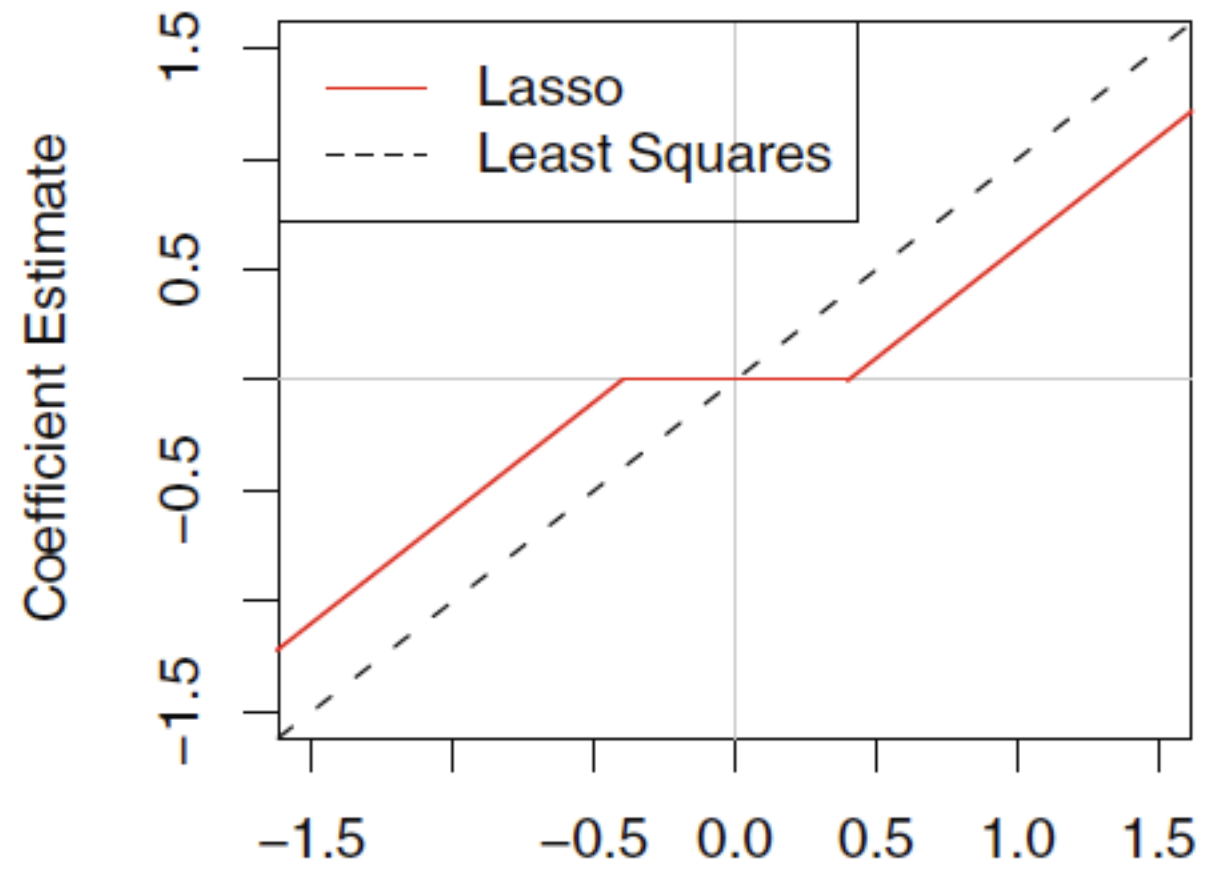




Ridge versus lasso

	Least squares	Ridge	Lasso
Penalty	None	$\sum_{j=1}^{p-1} \beta_j^2$	$\sum_{j=1}^{p-1} \lvert\beta_j\rvert$
Penalty effect	N/A	Shrinkage	Shrinkage and selection
Sparsity	No	No	Yes
Correlated features	(Unstable)	Splits the credit (stable)	Chooses one arbitrarily (unstable)
Performs better when	n/p is large	Many features have small effects	Few features have large effects
Works when $p > n$	No	Yes	Yes





Elastic net regression

Get the benefits of ridge and lasso regression by combining the two penalties:

$$\text{Penalty} = (1-\alpha)\sum_{j=1}^{p}\beta_j^2 + \alpha\sum_{j=1}^{p}|\beta_j|$$

- When $\alpha = 0$, we get ridge regression
- When $\alpha = 1$, we get lasso regression
- When $0 < \alpha < 1$, we get ridge-like shrinkage as well as lasso-like selection

Elastic net gives sparse solutions as long as $\alpha > 0$.

How to choose α? Can cross-validate over α and λ: First choose α to minimize CV error, then choose λ according to the one-standard-error rule.

Summary: Lasso Regression

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.
  - Can be applied to logistic regression as well.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.
  - Can be applied to logistic regression as well.
- Unlike ridge:

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.
  - Can be applied to logistic regression as well.
- Unlike ridge:
  - Sets some coefficients exactly to zero, achieving variable selection.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.
  - Can be applied to logistic regression as well.
- Unlike ridge:
  - Sets some coefficients exactly to zero, achieving variable selection.
  - Coefficient estimates unstable in the presence of correlated features.

Summary: Lasso Regression

- Penalized regression method encouraging coefficients to be sparse.
- Like ridge:
  - λ controls the degrees of freedom; *larger* λ gives *fewer* degrees of freedom.
  - Bias-variance trade-off: larger λ gives higher bias but lower variance.
  - Features need to be standardized.
  - Can be applied to logistic regression as well.
- Unlike ridge:
  - Sets some coefficients exactly to zero, achieving variable selection.
  - Coefficient estimates unstable in the presence of correlated features.
- Can be combined with ridge (elastic net).