

Lasso regression

STAT 471

October 19, 2021

Where we are

✓ **Unit 1:** Intro to modern data mining

✓ **Unit 2:** Tuning predictive models

Unit 3: Regression-based methods

Unit 4: Tree-based methods

Unit 5: Deep learning

Lecture 1: Logistic regression

Lecture 2: Regression in high dimensions

Lecture 3: Ridge regression

[Fall break: No class]

Lecture 4: Lasso regression

Lecture 5: Unit review and quiz in class

Homework 1 due the following **Sunday**.

Midterm exam following **Monday (7-9pm)**.

Idea: Encourage coefficients to be exactly zero

First, recall ridge regression:

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

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

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

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

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

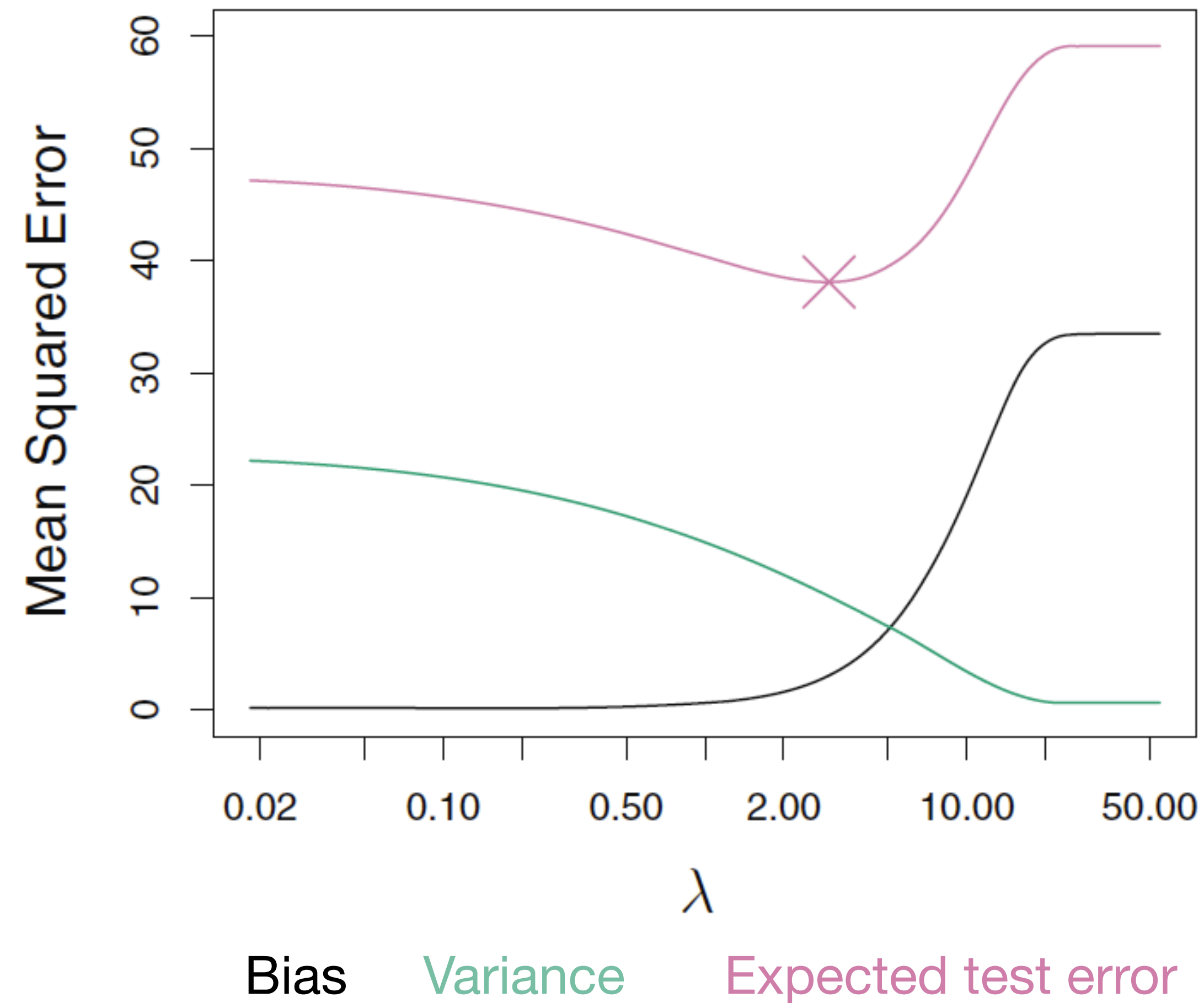
The effect of the penalty parameter λ

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

- 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



Sparsity and interpretability

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

Lasso is therefore a **variable selection** tool.

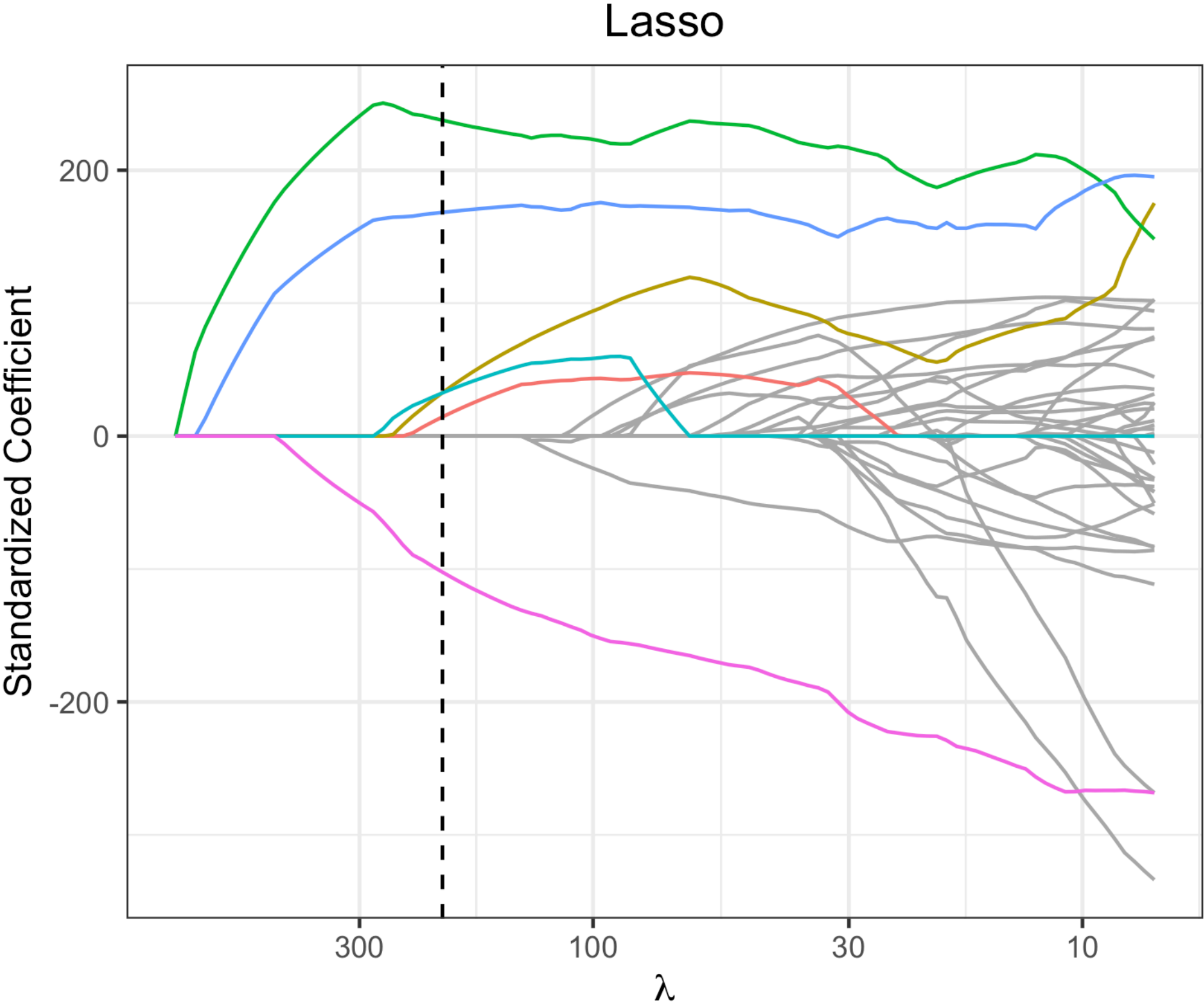
Sparse coefficient vectors are **interpretable**; they suggest which features are important.

We cannot attach a measure of statistical significance to the selected variables; lasso is still just a prediction method.

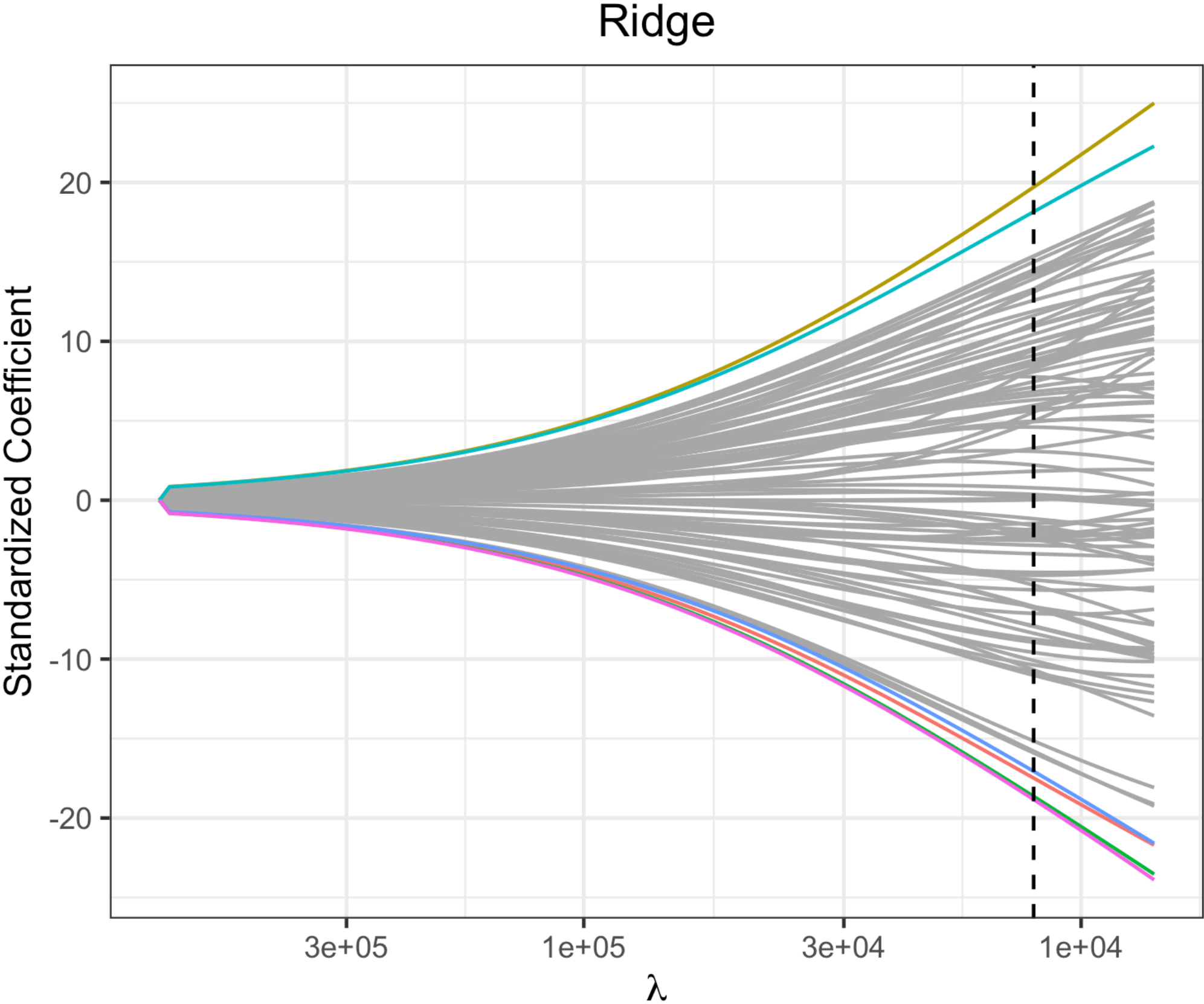
Example: Crime Data

```
# 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)



male.pct.divorce pct.kids.nvrmarried pct.pop.underpov
num.kids.nvrmarried pct.people.dense.hh pct.youngkids2parent



pct.fam2parents pct.kids2parents pct.teens2parents
pct.kids.nvrmarried pct.pop.underpov pct.youngkids2parents

Lasso regression in a simple case

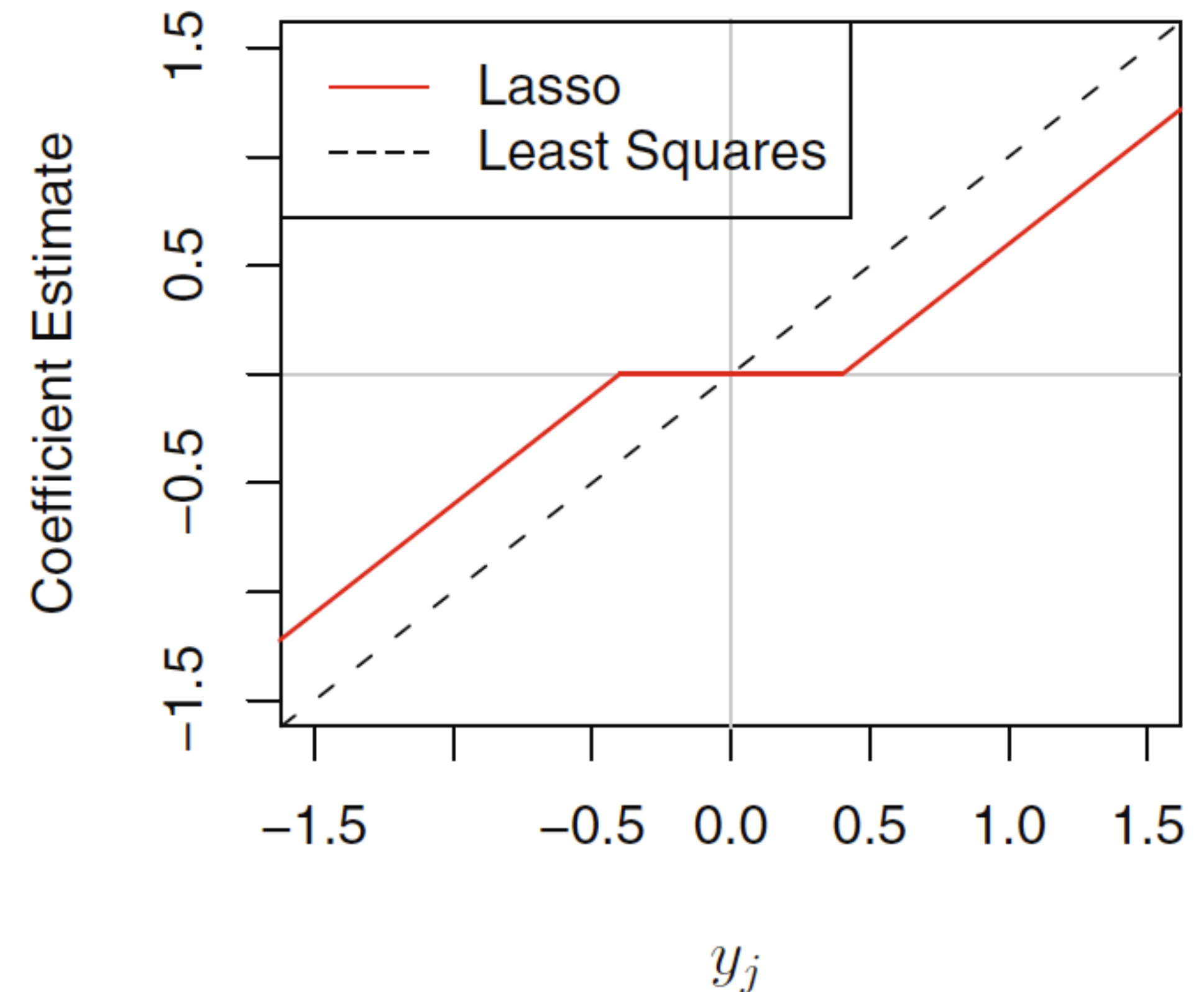
Suppose that $n = p - 1$ and $X_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

Consider fitting lasso regression without intercept:

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

Then,

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



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

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

Lasso regression can be penalized, just like linear regression!

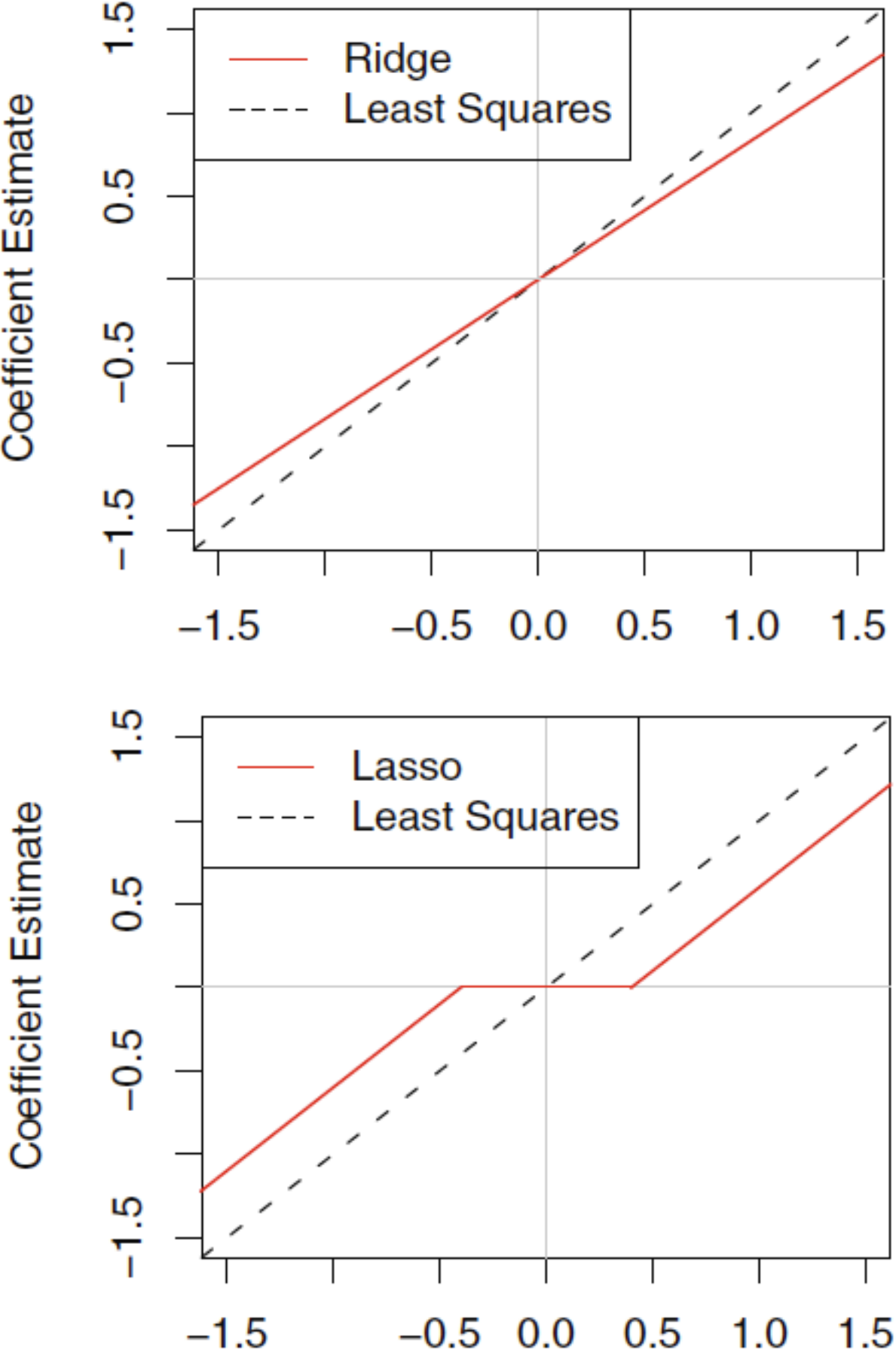
Recall $\mathcal{L}(\beta)$, the logistic regression likelihood. We can view $-\log \mathcal{L}(\beta)$ as analogous to the linear regression RSS. Continuing the analogy, we can define

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

Subtle point: While $\hat{\beta}^{\text{lasso}}$ 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. misclassification error).

Ridge versus lasso

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