Ridge regression

STAT 4710

October 10, 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 not to be too large

Idea: Encourage coefficients not to be too large

First, recall linear regression:

$$\widehat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2.$$

If there are too many features, the coefficients can get a little wild (high variance!).

Idea: Encourage coefficients not to be too large

First, recall linear regression:

$$\widehat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2.$$

If there are too many features, the coefficients can get a little wild (high variance!).

To tame those wild coefficients, we add a penalty to disincentivize large values:

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

for some $\lambda \geq 0$.

Idea: Encourage coefficients not to be too large

First, recall linear regression:

$$\hat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2.$$

If there are too many features, the coefficients can get a little wild (high variance!).

To tame those wild coefficients, we add a penalty to disincentivize large values:

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

for some $\lambda \geq 0$.

Ridge regression is defined even if $p > n$, as long as $\lambda > 0$.

The effect of the penalty parameter λ

The effect of the penalty parameter λ

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

The effect of the penalty parameter λ

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

- The larger λ is, the more of a penalty there is.

The effect of the penalty parameter λ

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

- The larger λ is, the more of a penalty there is.
- For $\lambda = 0$, we get back ordinary least squares.

The effect of the penalty parameter λ

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

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

The effect of the penalty parameter λ

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

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

λ as controls the flexibility of the ridge regression fit, like the degrees of freedom in a spline fit. However, *larger* λ means *fewer* degrees of freedom.

The effect of the penalty parameter λ

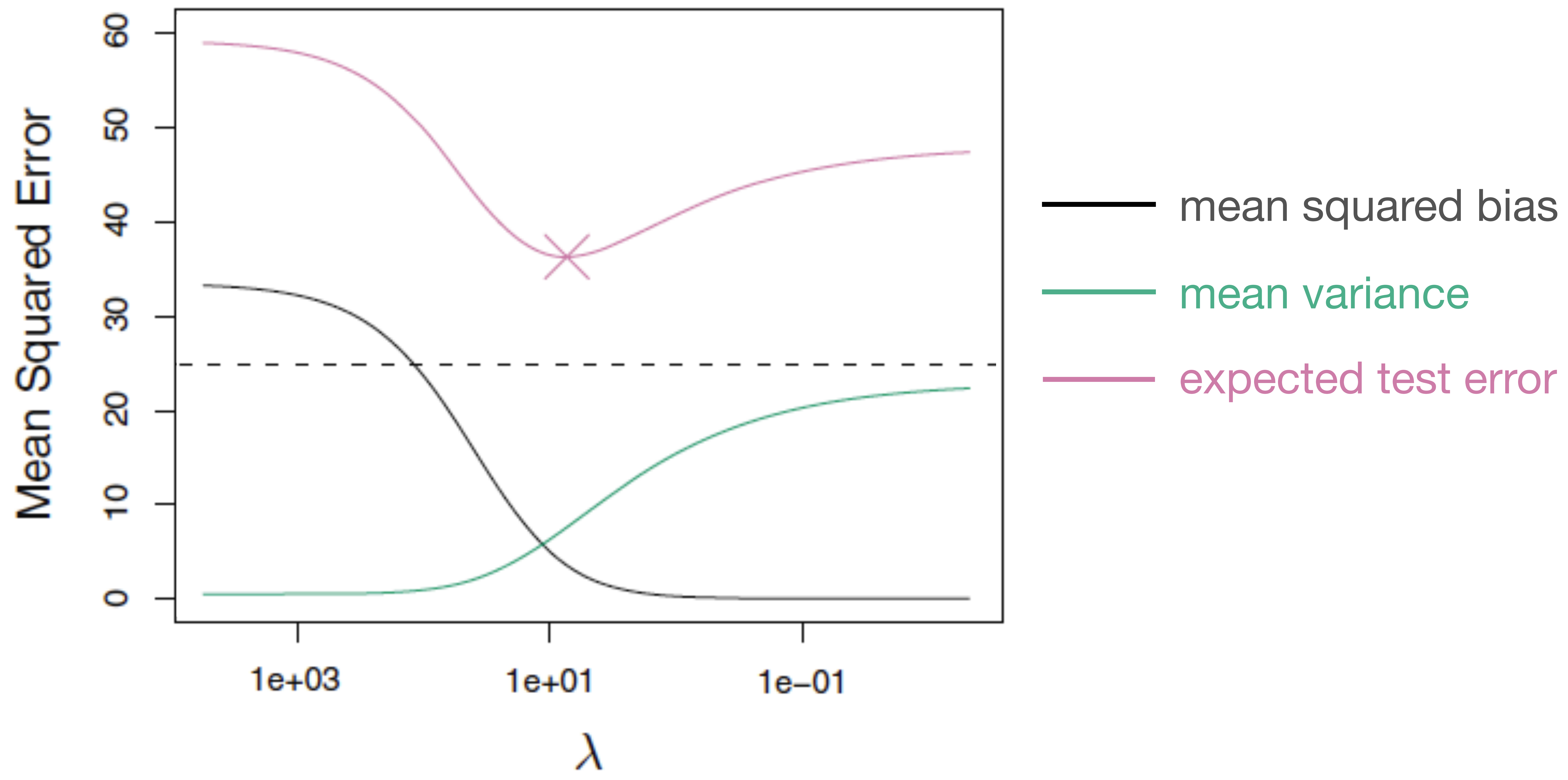
$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2$$

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

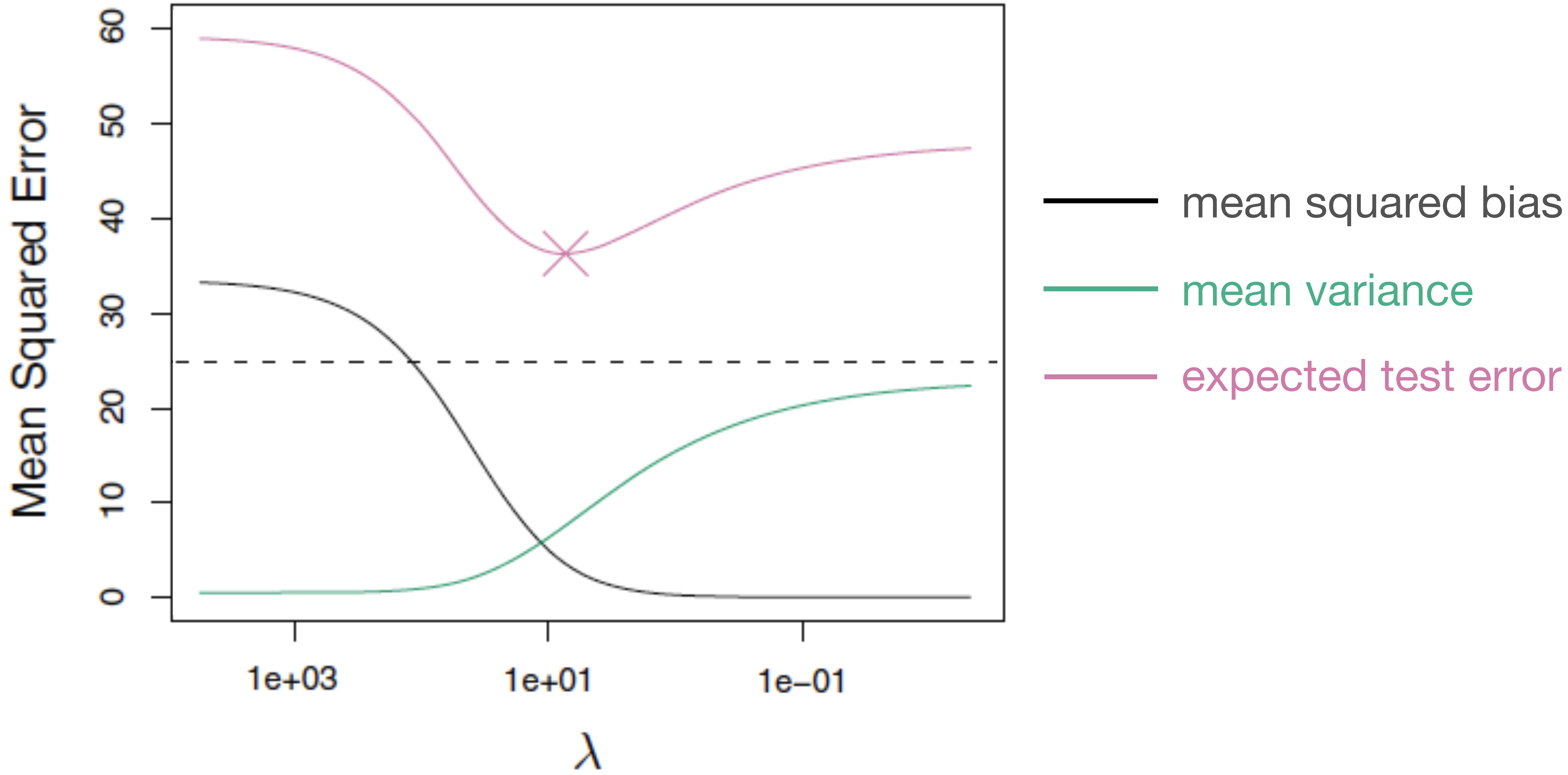
λ as controls the flexibility of the ridge regression fit, like the degrees of freedom in a spline fit. However, *larger* λ means *fewer* degrees of freedom.

Mathematical expression for the df of ridge regression is complicated; we skip it.

The bias-variance tradeoff for ridge regression



The bias-variance tradeoff for ridge regression



In practice, λ is chosen by cross-validation.

The importance of feature scaling

The importance of feature scaling

Suppose X_1 is height. Does it matter if it's measured in inches or feet?

The importance of feature scaling

Suppose X_1 is height. Does it matter if it's measured in inches or feet?

For least squares, does not matter. If $X_1 \to 12X_1$, then $\widehat{\beta}_1 \to \frac{1}{12}\widehat{\beta}_1$.

The importance of feature scaling

Suppose X_1 is height. Does it matter if it's measured in inches or feet?

For least squares, does not matter. If $X_1 \to 12X_1$, then $\widehat{\beta}_1 \to \frac{1}{12}\widehat{\beta}_1$.

$$\widehat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2$$

The importance of feature scaling

Suppose X_1 is height. Does it matter if it's measured in inches or feet?

For least squares, does not matter. If $X_1 \to 12X_1$, then $\widehat{\beta}_1 \to \frac{1}{12}\widehat{\beta}_1$.

$$\widehat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2$$

For ridge regression, it does matter! Implicitly, all features assumed on the same scale.

The importance of feature scaling

Suppose X_1 is height. Does it matter if it's measured in inches or feet?

For least squares, does not matter. If $X_1 \to 12X_1$, then $\widehat{\beta}_1 \to \frac{1}{12}\widehat{\beta}_1$.

$$\widehat{\beta}^{\text{least squares}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}))^2$$

For ridge regression, it does matter! Implicitly, all features assumed on the same scale.

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

Feature standardization

Feature standardization

To put features on the same scale, center each feature and divide by its std. dev.:

$$X_{ij}^{\text{std}} = \frac{X_{ij} - \widehat{\mu}_j}{\widehat{\sigma}_j}; \quad \widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}; \quad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2.$$

Feature standardization

To put features on the same scale, center each feature and divide by its std. dev.:

$$X_{ij}^{\text{std}} = \frac{X_{ij} - \widehat{\mu}_j}{\widehat{\sigma}_j}; \quad \widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}; \quad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2.$$

Feature standardization is recommended before applying ridge regression:

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

Feature standardization

To put features on the same scale, center each feature and divide by its std. dev.:

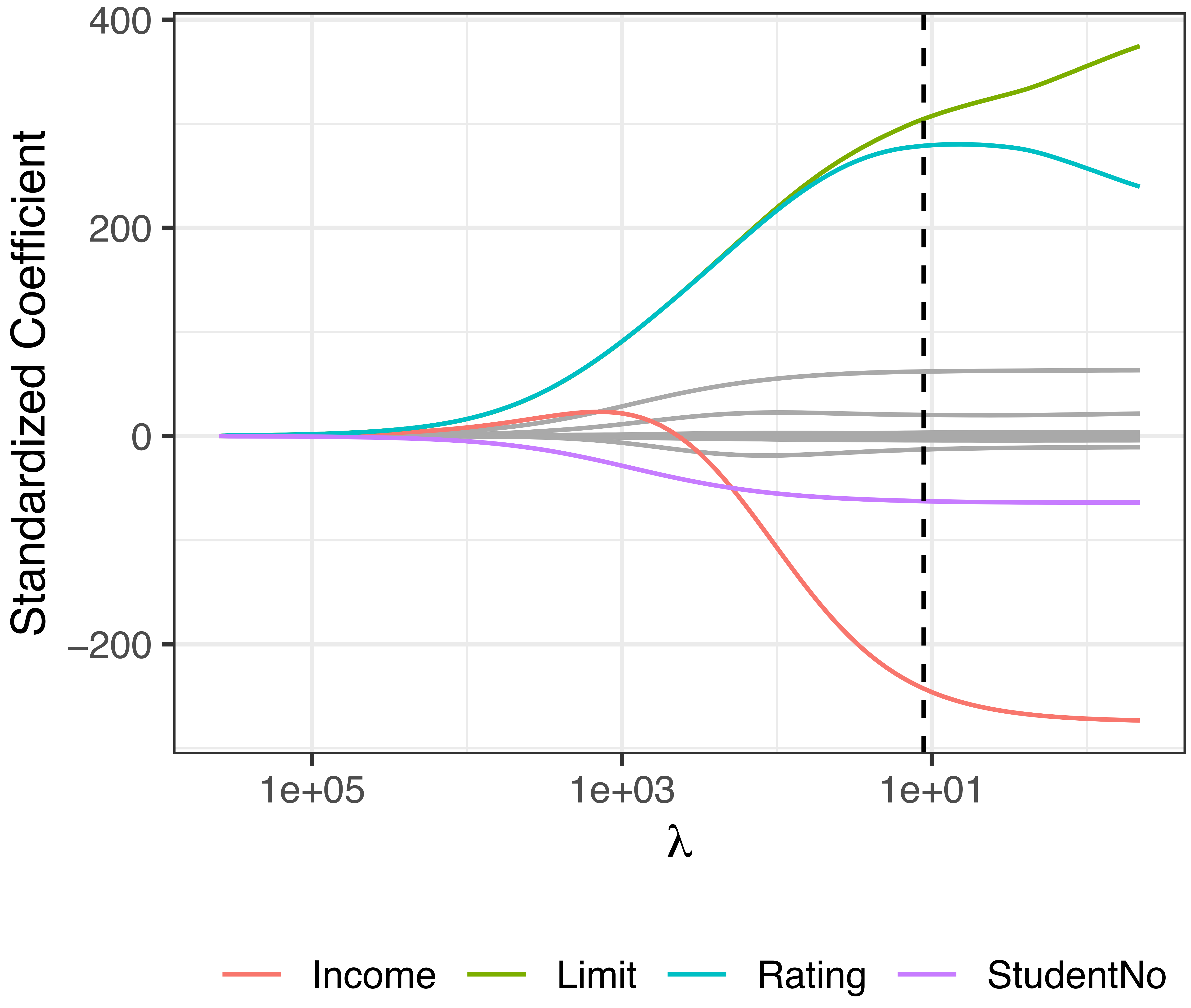
$$X_{ij}^{\text{std}} = \frac{X_{ij} - \widehat{\mu}_j}{\widehat{\sigma}_j}; \quad \widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}; \quad \widehat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2.$$

Feature standardization is recommended before applying ridge regression:

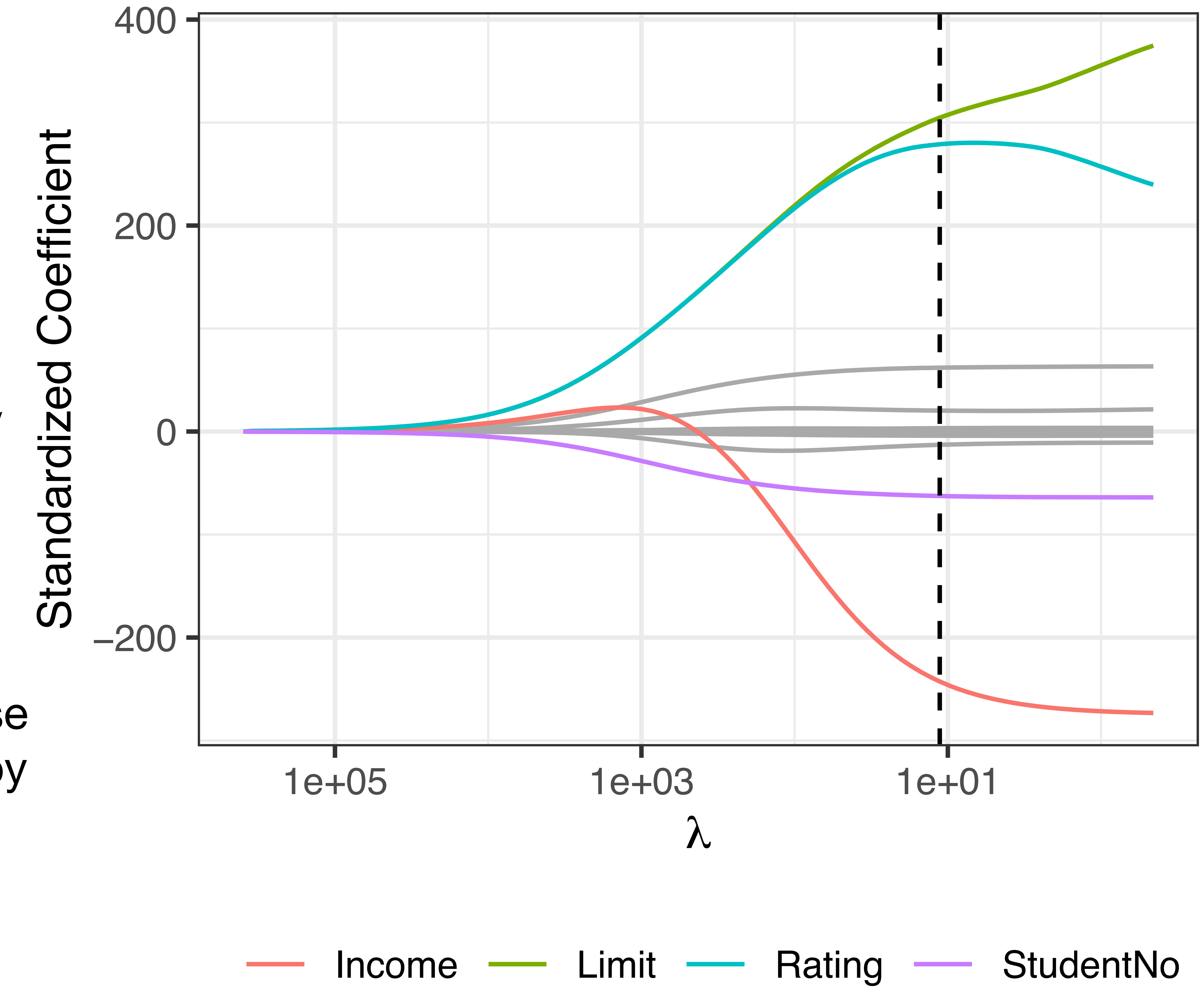
$$\widehat{\beta}^{\text{ridge}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i1}^{\text{std}} + \cdots + \beta_{p-1} X_{i,p-1}^{\text{std}}))^2 + \lambda \sum_{j=1}^{p-1} \beta_j^2$$

Interpretation: Mean response changes by β_j when X_j is increased by a standard deviation.

Ridge regression trace plot



Ridge regression trace plot



Change in mean response when feature increases by one standard deviation.

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

Ridge 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 ridge regression without intercept:

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

Ridge 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 ridge regression without intercept:

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

In this simple case, $\widehat{\beta}_j^{\text{OLS}} = Y_j$ and $\widehat{\beta}_j^{\text{ridge}} = Y_j/(1+\lambda)$
(OLS stands for ordinary least squares).

Ridge 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. $Y_j = \beta_j + \epsilon_j$.

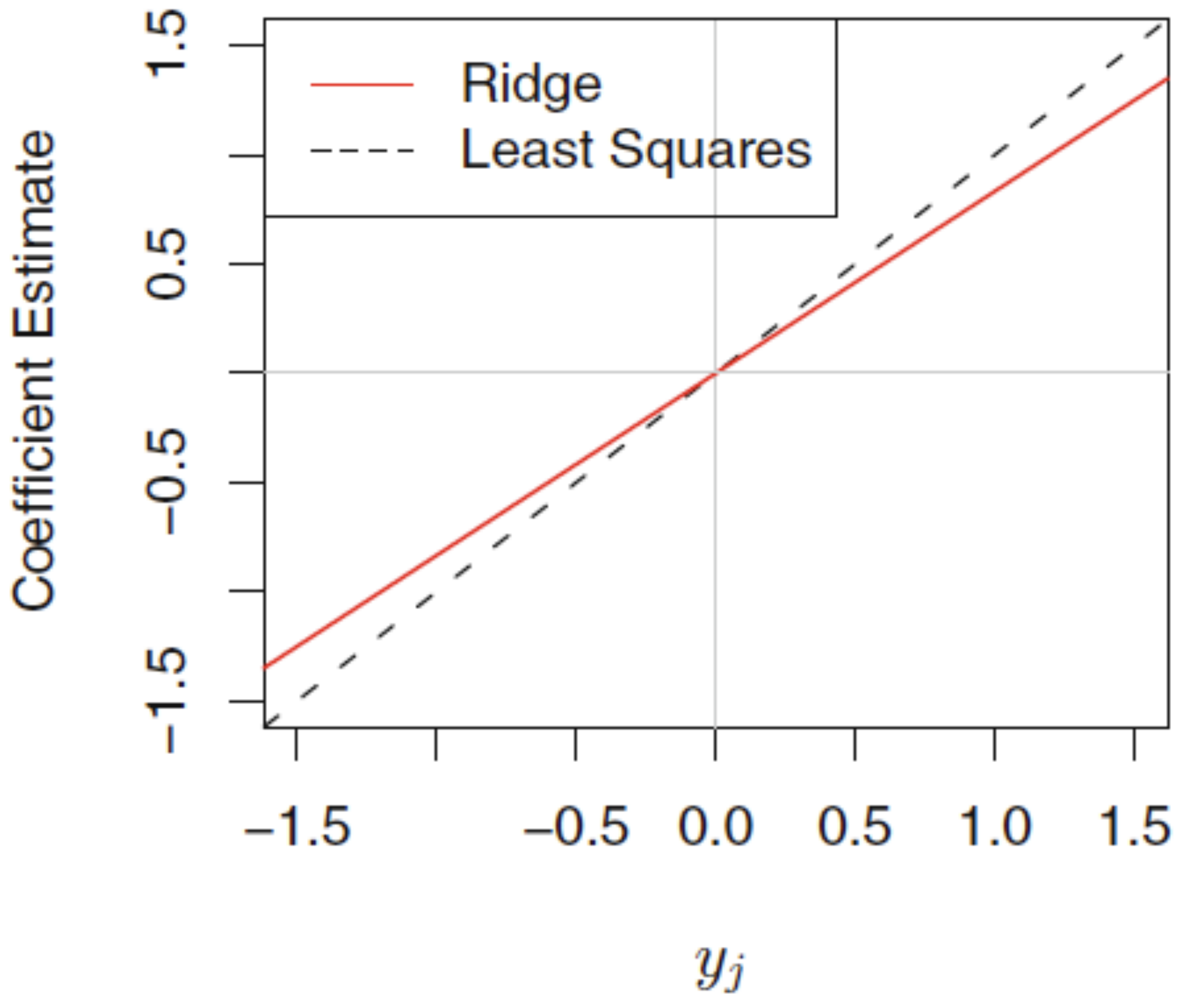
E.g. $X = $

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

Consider fitting ridge regression without intercept:

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

In this simple case, $\hat{\beta}_j^{\text{OLS}} = Y_j$ and $\hat{\beta}_j^{\text{ridge}} = Y_j/(1+\lambda)$ (OLS stands for ordinary least squares).



Ridge 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 = $

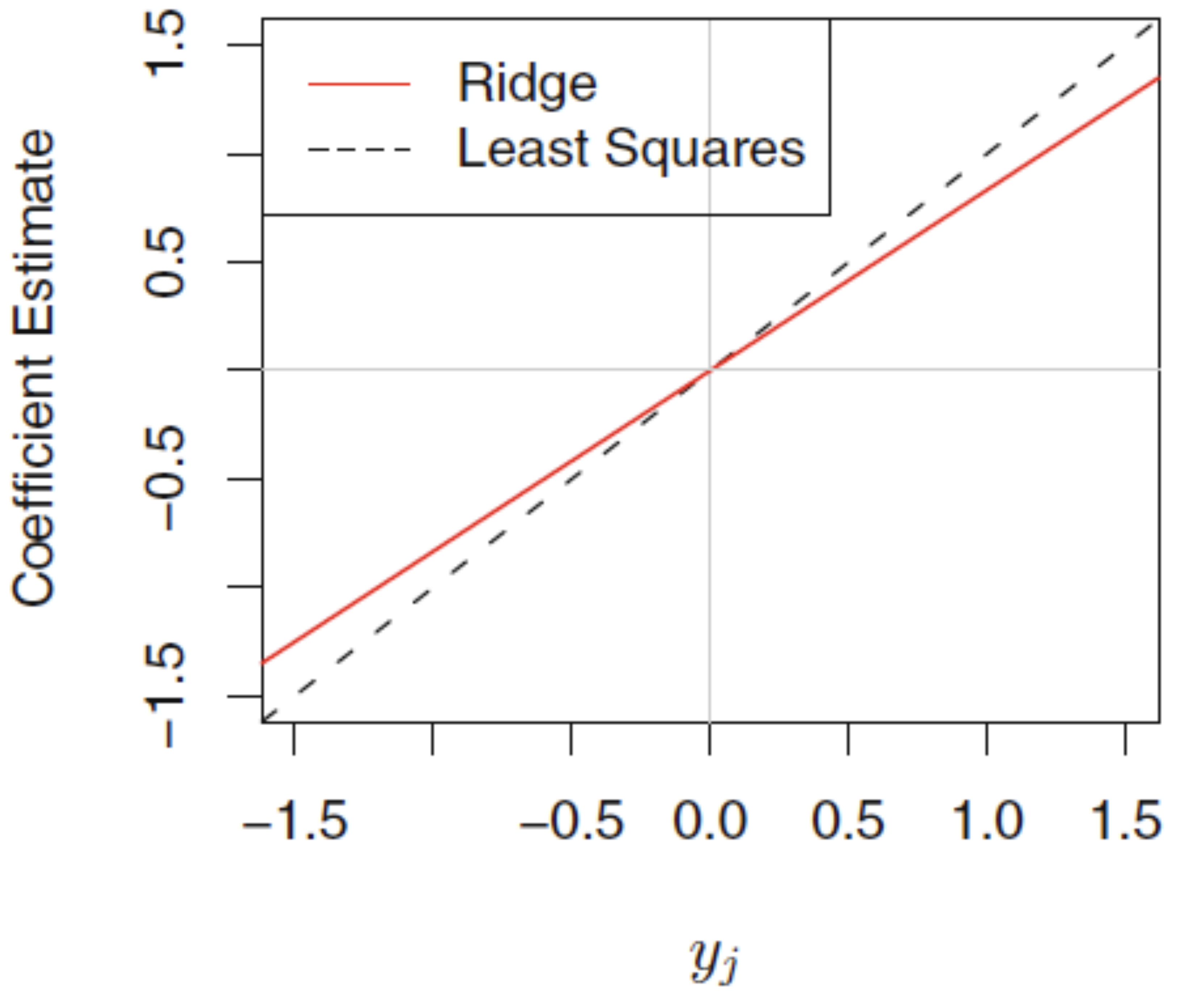
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Consider fitting ridge regression without intercept:

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

In this simple case, $\hat{\beta}_j^{\text{OLS}} = Y_j$ and $\hat{\beta}_j^{\text{ridge}} = Y_j/(1+\lambda)$ (OLS stands for ordinary least squares).

So $\hat{\beta}^{\text{ridge}} = \dfrac{1}{1+\lambda}\hat{\beta}^{\text{OLS}}$, i.e. the ridge estimate is obtained by *shrinking* the OLS estimate by a factor of $1+\lambda$.



Treatment of correlated features

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Ridge regression is more stable, “splitting the credit” among correlated features.

Treatment of correlated features

Linear regression coefficients for correlated features tend to be unstable.

Ridge regression is more stable, “splitting the credit” among 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.

Ridge regression is more stable, "splitting the credit" among 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.

Ridge regression is more stable, “splitting the credit” among 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.
- Ridge regression will obtain $\widehat{\beta}$ from $y = \beta X_1 + \epsilon$, and set $\widehat{\beta}_1 = \widehat{\beta}_2 = \frac{1}{2}\widehat{\beta}$.

Logistic regression with ridge penalty

Logistic regression with ridge penalty

Logistic regression can be penalized, just like linear regression!

Logistic regression with ridge 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

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\arg\min} \left\{ -\log \mathscr{L}(\beta) + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

Logistic regression with ridge 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

$$\widehat{\beta}^{\text{ridge}} = \underset{\beta}{\arg\min} \left\{ -\log \mathscr{L}(\beta) + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}.$$

Subtle point: While $\widehat{\beta}^{\text{ridge}}$ 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).

Summary: Ridge Regression

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.
- The penalty parameter λ controls the degrees of freedom of the fit, with *larger* λ giving *fewer* degrees of freedom.

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.
- The penalty parameter λ controls the degrees of freedom of the fit, with *larger* λ giving *fewer* degrees of freedom.
- Bias-variance trade-off: larger λ gives higher bias but lower variance.

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.
- The penalty parameter λ controls the degrees of freedom of the fit, with *larger* λ giving *fewer* degrees of freedom.
- Bias-variance trade-off: larger λ gives higher bias but lower variance.
- Features need to be standardized prior to the application of ridge regression.

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.
- The penalty parameter λ controls the degrees of freedom of the fit, with *larger* λ giving *fewer* degrees of freedom.
- Bias-variance trade-off: larger λ gives higher bias but lower variance.
- Features need to be standardized prior to the application of ridge regression.
- Ridge regression tends to split the credit among correlated features.

Summary: Ridge Regression

- Penalized regression method encouraging coefficients not to be too large.
- The penalty parameter λ controls the degrees of freedom of the fit, with *larger* λ giving *fewer* degrees of freedom.
- Bias-variance trade-off: larger λ gives higher bias but lower variance.
- Features need to be standardized prior to the application of ridge regression.
- Ridge regression tends to split the credit among correlated features.
- Ridge penalization can be applied to logistic regression as well.