Lecture 11: Regularization

STAT 598z: Introduction to computing for statistics

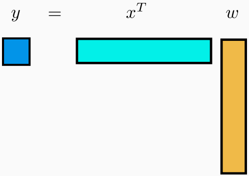
Vinayak Rao

Department of Statistics, Purdue University

February 15, 2017

Consider linear regression:

$$y = \mathbf{x}^\top \mathbf{w} + \epsilon$$



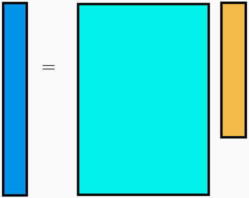
Ordinary least squares

Consider linear regression:

$$y = \mathbf{x}^\top \mathbf{w} + \epsilon$$

In vector notation:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon, \quad \mathbf{y} \in \Re^n, \mathbf{X} \in \Re^{n \times p}$$



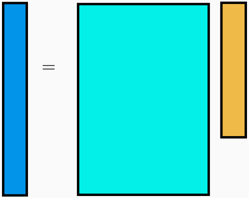
Ordinary least squares

Consider linear regression:

$$y = \mathbf{x}^\top \mathbf{w} + \epsilon$$

In vector notation:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon, \quad \mathbf{y} \in \Re^n, \mathbf{X} \in \Re^{n \times p}$$



$$\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2 = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

Ordinary least squares

Problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2 = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

Solution:

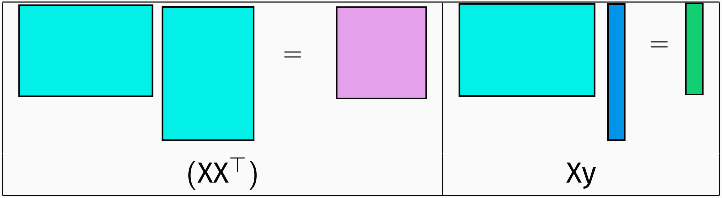
$$\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y} \qquad \text{(correlation in 1-d)}$$

Ordinary least squares

Problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathrm{X}^\top \mathbf{w}\|^2 = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

Solution: $\hat{\mathbf{w}} = (\mathrm{X}\mathrm{X}^\top)^{-1}\mathrm{X}\mathbf{y}$ (correlation in 1-d)



How to do this in `R` (without using `lm`)?

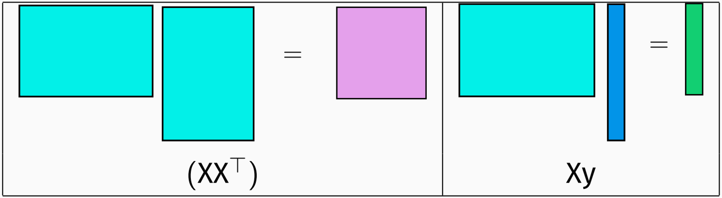
- Do not invert with `solve` and multiply!

Ordinary least squares

Problem:

$$\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2 = \arg\min \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2$$

Solution: $\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ (correlation in 1-d)



How to do this in `R` (without using `lm`)?

- Do not invert with `solve` and multiply!
- Directly solve $(\mathbf{X}\mathbf{X}^\top)\hat{\mathbf{w}} = \mathbf{X}\mathbf{y}$

Prediction error

$\hat{\mathbf{w}}$ is an unbiased estimate of the true $\mathbf{w}$

For a test vector $\mathbf{x}^{test}$ we predict $\mathbf{w}^\top \mathbf{x}^{test}$.

(Squared) prediction error: $PE = \frac{1}{k} \sum_{i=1}^{k} (y_i^{test} - \mathbf{w}^\top \mathbf{x}_i^{test})^2$

Prediction error

$\hat{\mathbf{w}}$ is an unbiased estimate of the true $\mathbf{w}$

For a test vector $\mathbf{x}^{test}$ we predict $\mathbf{w}^\top\mathbf{x}^{test}$.

(Squared) prediction error: $PE = \frac{1}{k}\sum_{i=1}^{k}(y_i^{test} - \mathbf{w}^\top\mathbf{x}_i^{test})^2$

Can show:

- PE is has mean 0
- variance grows with number of features (p)

Prediction error

$\hat{\mathbf{w}}$ is an unbiased estimate of the true $\mathbf{w}$

For a test vector $\mathbf{x}^{test}$ we predict $\mathbf{w}^\top \mathbf{x}^{test}$.

(Squared) prediction error: $PE = \frac{1}{k} \sum_{i=1}^{k} (y_i^{test} - \mathbf{w}^\top \mathbf{x}_i^{test})^2$

Can show:

- PE is has mean 0
- variance grows with number of features (p)

What if $p > n$?

- $\mathbf{X}\mathbf{X}^\top$ is singular

Regularization

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$

Regularization

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Regularization

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Introducing λ makes problem well-posed, but introduces bias

Regularization

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Introducing λ makes problem well-posed, but introduces bias

- $\lambda = 0$ recovers OLS
- Larger λ causes larger bias

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Introducing λ makes problem well-posed, but introduces bias

- $\lambda = 0$ recovers OLS
- Larger λ causes larger bias
- $\lambda = \infty$?

Regularization

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Introducing λ makes problem well-posed, but introduces bias

- $\lambda = 0$ recovers OLS
- Larger λ causes larger bias
- $\lambda = \infty$? No variance!

$p > n$:

- Cannot invert $\mathbf{X}\mathbf{X}^\top$
- We *can* invert if we add a small λ to the diagonal

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y} \qquad (I \text{ is the identity matrix})$$

Introducing λ makes problem well-posed, but introduces bias

- $\lambda = 0$ recovers OLS
- Larger λ causes larger bias
- $\lambda = \infty$? No variance!

λ trades-off bias and variance

Maybe a nonzero λ is actually good?

Ridge regression (a.k.a. Tikhonov regularization)

Recall $\hat{\mathsf{w}} = (\mathsf{X}\mathsf{X}^\top)^{-1}\mathsf{X}\mathsf{y}$ solves $\hat{\mathsf{w}} = \arg\min \|\mathsf{y} - \mathsf{X}^\top \mathsf{w}\|^2$

Ridge regression (a.k.a. Tikhonov regularization)

Recall $\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ solves $\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top\mathbf{w}\|^2$

$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y}$ solves

$$\hat{\mathbf{w}}_\lambda = \text{argmin}\mathcal{L}_\lambda(\mathbf{w}) := \text{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2$$

Ridge regression (a.k.a. Tikhonov regularization)

Recall $\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ solves $\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top\mathbf{w}\|^2$

$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y}$ solves

$$\hat{\mathbf{w}}_\lambda = \mathrm{argmin}\mathcal{L}_\lambda(\mathbf{w}) := \mathrm{argmin}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2$$

$\|\mathbf{w}\|_2^2 = \sum_{i=1}^{p} w_i^2$ is the squared ℓ_2-norm

$\lambda\|\mathbf{w}\|_2$ is the *shrinkage penalty*.

Ridge regression (a.k.a. Tikhonov regularization)

Recall $\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ solves $\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top\mathbf{w}\|^2$

$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y}$ solves

$$\hat{\mathbf{w}}_\lambda = \operatorname{argmin} \mathcal{L}_\lambda(\mathbf{w}) := \operatorname{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2$$

$\|\mathbf{w}\|_2^2 = \sum_{i=1}^{p} w_i^2$ is the squared ℓ_2-norm

$\lambda\|\mathbf{w}\|_2$ is the *shrinkage penalty*.

Favours $\mathbf{w}$'s with smaller components

Recall $\hat{\mathbf{w}} = (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$ solves $\hat{\mathbf{w}} = \arg\min \|\mathbf{y} - \mathbf{X}^\top \mathbf{w}\|^2$

$\hat{\mathbf{w}}_\lambda = (\mathbf{X}\mathbf{X}^\top + \lambda I)^{-1}\mathbf{X}\mathbf{y}$ solves

$$\hat{\mathbf{w}}_\lambda = \mathrm{argmin}\mathcal{L}_\lambda(\mathbf{w}) := \mathrm{argmin}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda\|\mathbf{w}\|_2^2$$

$\|\mathbf{w}\|_2^2 = \sum_{i=1}^{p} w_i^2$ is the squared ℓ_2-norm

$\lambda\|\mathbf{w}\|_2$ is the *shrinkage penalty.*

Favours $\mathbf{w}$'s with smaller components

λ trades of small training error with 'simple' solutions

Ridge regression (a.k.a. Tikhonov regularization)

Recall $\hat{\mathsf{w}} = (\mathsf{X}\mathsf{X}^\top)^{-1}\mathsf{X}\mathsf{y}$ solves $\hat{\mathsf{w}} = \arg\min \|\mathsf{y} - \mathsf{X}^\top \mathsf{w}\|^2$

$\hat{\mathsf{w}}_\lambda = (\mathsf{X}\mathsf{X}^\top + \lambda I)^{-1}\mathsf{X}\mathsf{y}$ solves

$$\hat{\mathsf{w}}_\lambda = \operatorname{argmin} \mathcal{L}_\lambda(\mathsf{w}) := \operatorname{argmin} \sum_{i=1}^{n} (y_i - \mathsf{x}_i^\top \mathsf{w})^2 + \lambda \|\mathsf{w}\|_2^2$$

$\|\mathsf{w}\|_2^2 = \sum_{i=1}^{p} w_i^2$ is the squared ℓ_2-norm

$\lambda\|\mathsf{w}\|_2$ is the *shrinkage penalty*.

Favours w's with smaller components

λ trades of small training error with 'simple' solutions

ℓ_2/ridge/Tikhonov regularization

Ridge regression (solution)

Simple modification of the least-squares solution:

$$\hat{\mathsf{w}}_\lambda = (\mathsf{X}^\top \mathsf{X} + \lambda \mathsf{I}_p)^{-1} \mathsf{X}^\top \mathsf{y}$$

Ridge regression (solution)

Simple modification of the least-squares solution:

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

In the 1-dimensional case,

$$\hat{w}_\lambda = (\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)^{-1} \mathbf{x}^\top \mathbf{y}$$

Ridge regression (solution)

Simple modification of the least-squares solution:

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

In the 1-dimensional case,

$$\begin{aligned} \hat{w}_\lambda &= (\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)^{-1} \mathbf{x}^\top \mathbf{y} \\ &= \frac{\mathbf{x}^\top \mathbf{x}}{(\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)} \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} \end{aligned}$$

Ridge regression (solution)

Simple modification of the least-squares solution:

$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

In the 1-dimensional case,

$$\begin{aligned} \hat{w}_\lambda &= (\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)^{-1} \mathbf{x}^\top \mathbf{y} \\ &= \frac{\mathbf{x}^\top \mathbf{x}}{(\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)} \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} \\ &= \quad c\, \hat{w} \qquad (c < 1) \end{aligned}$$

Ridge regression (solution)

Simple modification of the least-squares solution:

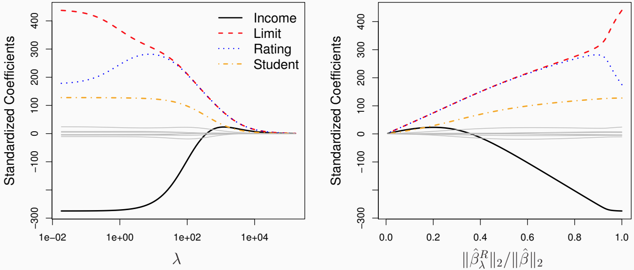
$$\hat{\mathbf{w}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$$

In the 1-dimensional case,

$$\begin{aligned} \hat{w}_\lambda &= (\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)^{-1} \mathbf{x}^\top \mathbf{y} \\ &= \frac{\mathbf{x}^\top \mathbf{x}}{(\mathbf{x}^\top \mathbf{x} + \lambda \mathbf{I}_p)} \frac{\mathbf{x}^\top \mathbf{y}}{\mathbf{x}^\top \mathbf{x}} \\ &= \quad c\, \hat{w} \qquad (c < 1) \end{aligned}$$

Shrinks least-squares solution.

Credit data set (average credit card debt)



James, Witten, Hastic and Tibshirani

How do we choose λ?

Cross-validaton:

How do we choose λ?

Cross-validaton:

- Pick a set of λ's
- For kth fold of cross-validation:

How do we choose λ?

Cross-validaton:

- Pick a set of λ's
- For kth fold of cross-validation:
  - For each λ:
    - Solve the regularized least squares problem on training data.
    - Evaluate estimated $\mathbf{w}$ on held-out data (call this $PE_{\lambda,k}$).

How do we choose λ?

Cross-validaton:

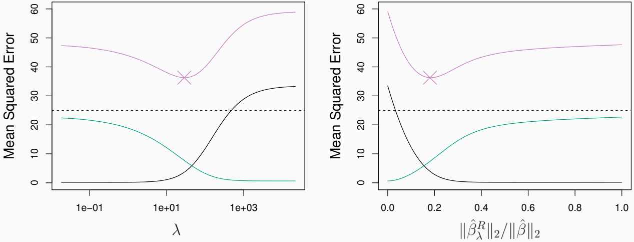
- Pick a set of λ's
- For kth fold of cross-validation:
  - For each λ:
    - Solve the regularized least squares problem on training data.
    - Evaluate estimated $\mathbf{w}$ on held-out data (call this $PE_{\lambda,k}$).
- Pick $\hat{\lambda} = \text{argmin mean}(PE_\lambda)$
  or $\quad (\text{argmin } (\text{mean}(PE_\lambda) + \text{stderr}(PE_\lambda)))$

How do we choose λ?

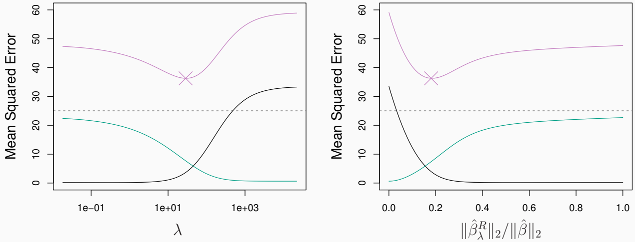
Cross-validaton:

- Pick a set of λ's
- For kth fold of cross-validation:
  - For each λ:
    - Solve the regularized least squares problem on training data.
    - Evaluate estimated $\mathbf{w}$ on held-out data (call this $PE_{\lambda,k}$).
- Pick $\hat{\lambda} = \text{argmin mean}(PE_\lambda)$
  or $\quad\quad (\text{argmin } (\text{mean}(PE_\lambda) + \text{stderr}(PE_\lambda)))$
- Having chosen $\hat{\lambda}$ solve regularized least square on all data

Does this work?

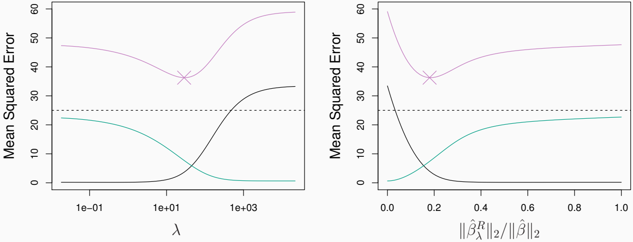


Does this work?



Ridge regression improves performance by reducing variance

Does this work?



Ridge regression improves performance by reducing variance

- does not perform feature selection
- just shrinks components of $\mathbf{w}$ towards 0

For the former: Lasso

Regularization as constrained optimization

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2$ is equivalent to

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \ \|\mathbf{w}\|_2^2 \leq \gamma$

(Note: γ will depend on data)

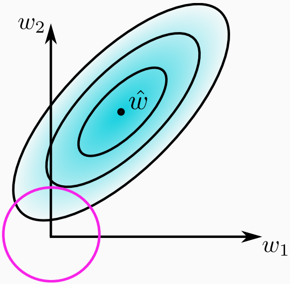
Regularization as constrained optimization

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2$ is equivalent to

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \; \|\mathbf{w}\|_2^2 \leq \gamma$

(Note: γ will depend on data)

First problem: regularized optimization
Second problem: constrained optimization

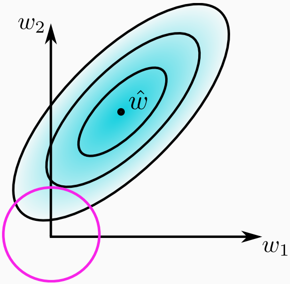


$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2$ is equivalent to

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \ \|\mathbf{w}\|_2^2 \leq \gamma$

(Note: γ will depend on data)

First problem: regularized optimization
Second problem: constrained optimization



Lasso

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \ \|\mathbf{w}\|_1 \leq \gamma$

$\|\mathbf{w}\|_1 = \sum_{i=1}^{p} |w|_i$ is the ℓ_1-norm.

$\operatorname{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t.\ \|\mathbf{w}\|_1 \leq \gamma$

$\|\mathbf{w}\|_1 = \sum_{i=1}^{p} |w|_i$ is the ℓ_1-norm.

Lasso: least absolute shrinkage and selection operator.

$$\hat{\mathbf{w}} = \operatorname{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_1$$

Lasso

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \ \|\mathbf{w}\|_1 \leq \gamma$

$\|\mathbf{w}\|_1 = \sum_{i=1}^{p} |w|_i$ is the ℓ_1-norm.

Lasso: least absolute shrinkage and selection operator.

$$\hat{\mathbf{w}} = \text{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_1$$

- Penalizes small w_i's more than ridge regression.
- Tolerates larger w_i's more than ridge regression.

Lasso

$\text{argmin}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^2 \quad s.t. \ \|\mathbf{w}\|_1 \leq \gamma$

$\|\mathbf{w}\|_1 = \sum_{i=1}^{p} |w|_i$ is the ℓ_1-norm.

Lasso: least absolute shrinkage and selection operator.

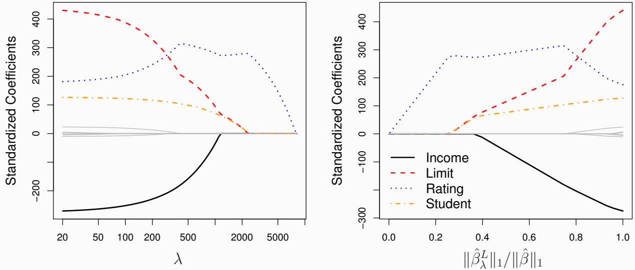
$$\hat{\mathbf{w}} = \text{argmin} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_1$$

- Penalizes small w_i's more than ridge regression.
- Tolerates larger w_i's more than ridge regression.

Result:

- $\hat{\mathbf{w}}_{LASSO}$ has some components *exactly* equal to zero.
- Performs feature selection.

Credit data set (average credit card debt)



James, Witten, Hastic and Tibshirani