IS4303 Week 5

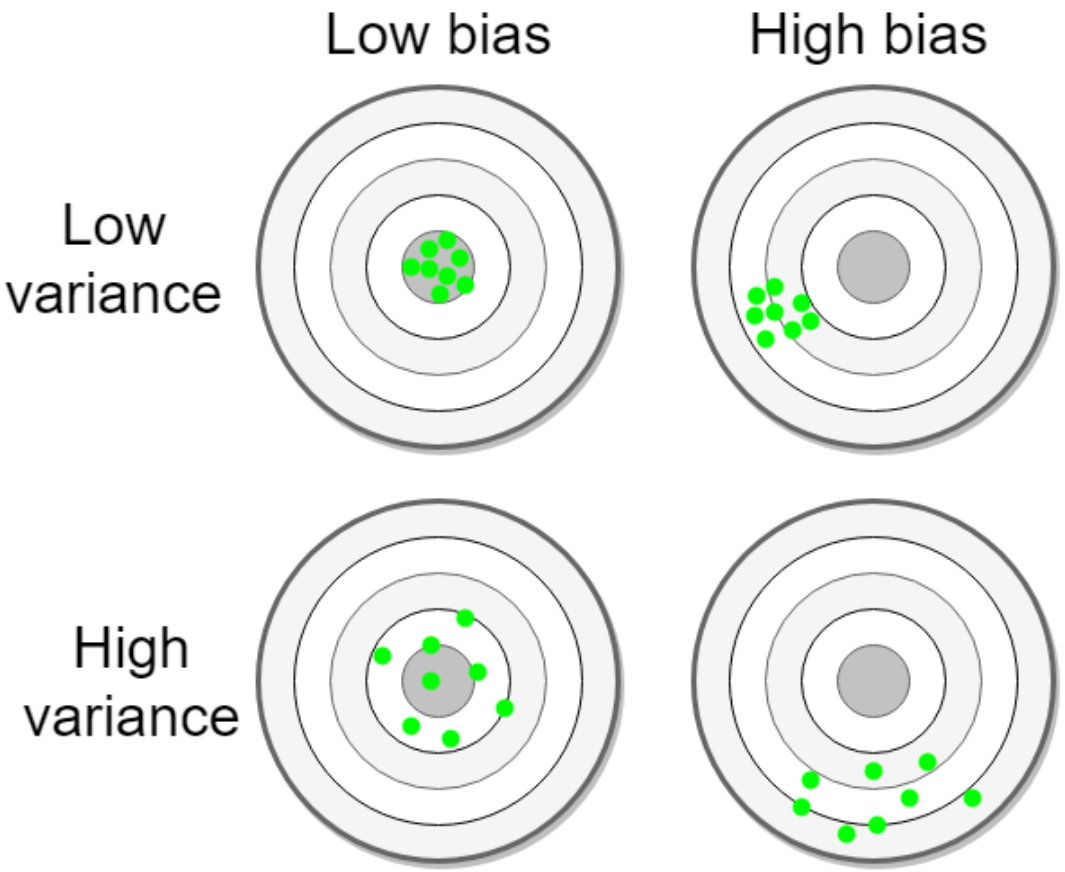
Regularization:
Lasso & Ridge Regression

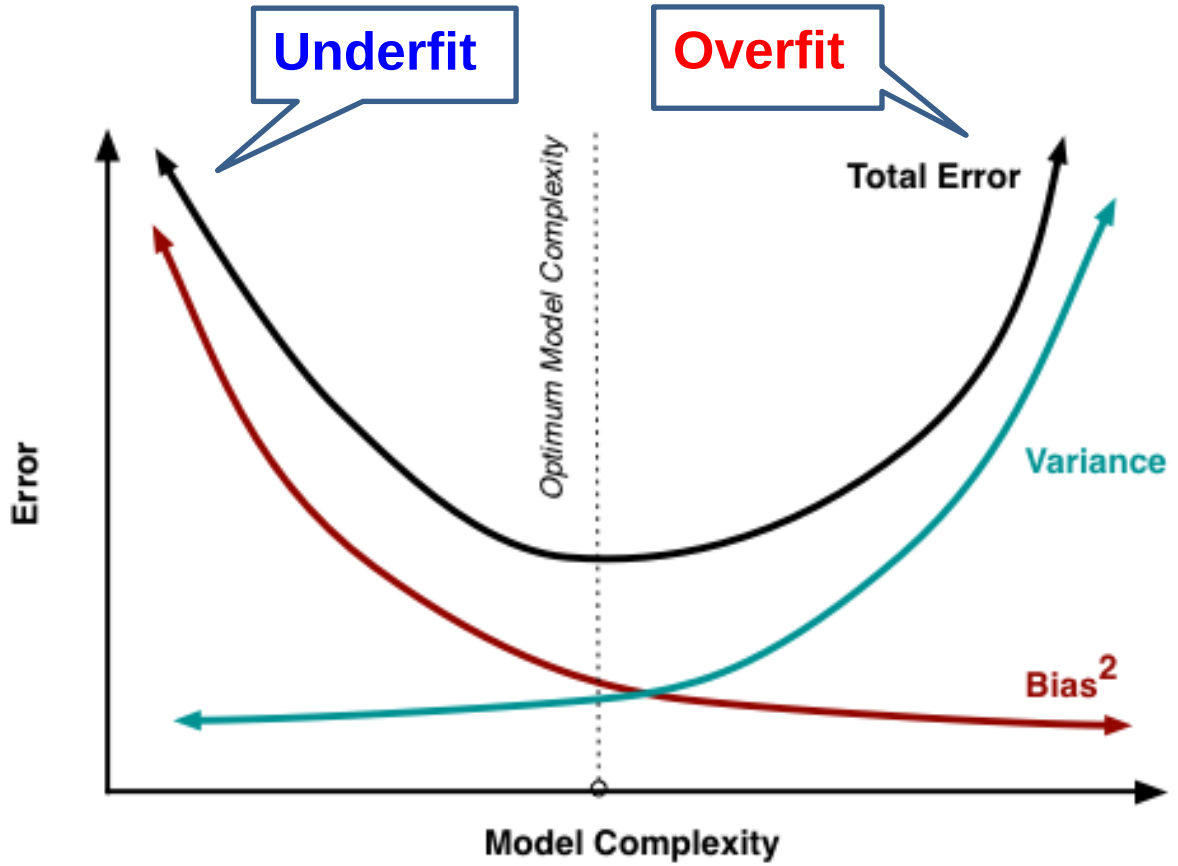
Agenda

- Underfit and Overfit
  - Bias-Variance Tradeoff
- Regularization in Regression
  - Lasso Regression
  - Ridge Regression
- Cross-Validation
  - K-Fold Cross-Validation

Underfit v.s. Overfit

Bias-Variance Tradeoff: $\mathrm{E}\left[\left(y-\hat{f}\left(x\right)\right)^2\right]=\left(\mathrm{Bias}\left[\hat{f}\left(x\right)\right]\right)^2+\mathrm{Var}\left[\hat{f}\left(x\right)\right]+\sigma^2$

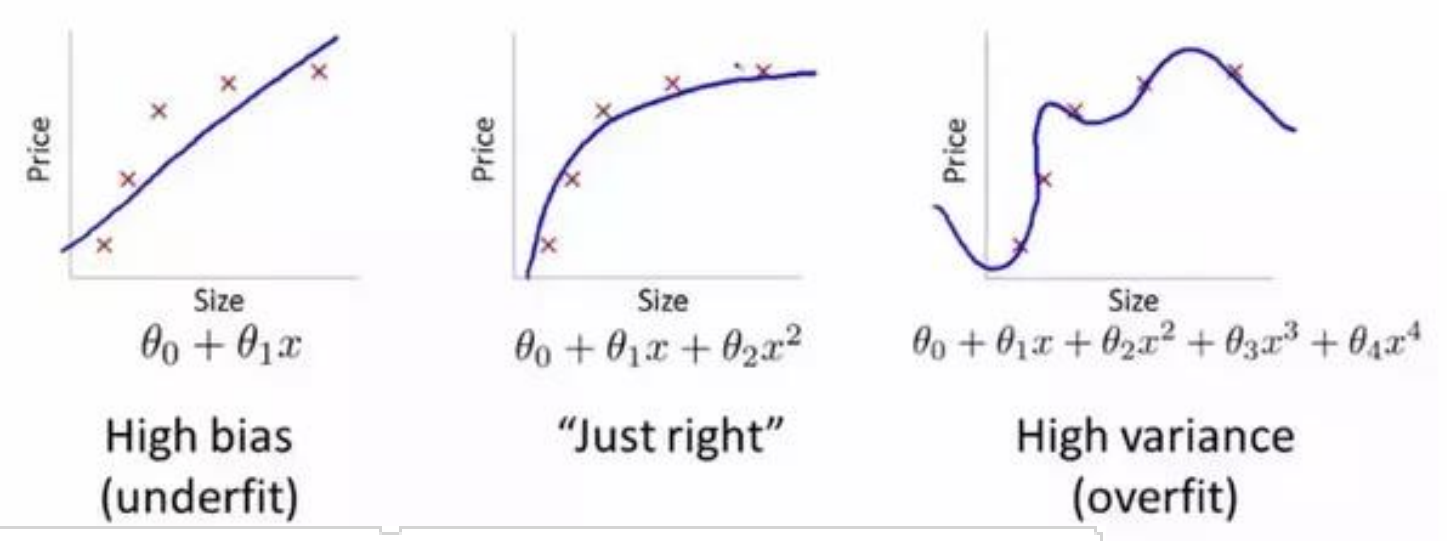




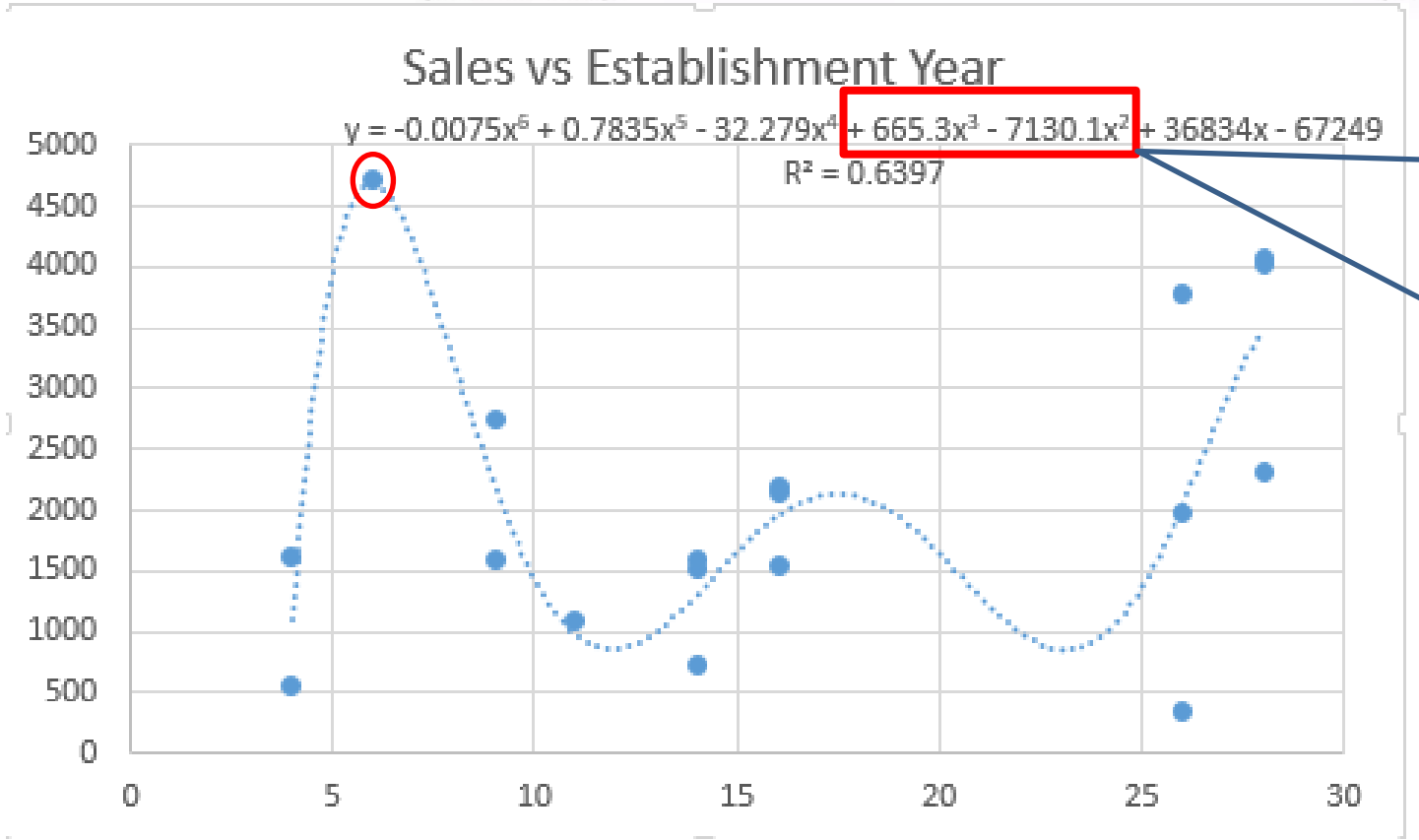
https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

Regression

- Problems:



- Example:



If you do not restrict the magnitude of coefficient size, your model will perform well on train data, but "overfit"

Overfit

- How to control overfit issue:
  - **Regularization: Penalized regression**
  - **Cross-Validation**
  - Feature Selection (Week4 Tutorial)
  - Pruning in decision tree (e.g., early stopping; Week5 Lecture)
  - Ensemble Learning methods (Week6 Tutorial)
  - Dropout in deep learning and neural network
  - More…

Regularization in Regression

- Regularization refers to the process of introducing additional constraints/penalties into the optimization problem to get a more sensible solution, and to control the problem of overfitting.

- Consider the standard OLS method (Ordinary Least Squares) to obtain regression coefficients (in Week 4):

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i$$

$$\min_{\beta_0,\ldots,\beta_j} \varepsilon^T\varepsilon = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Lasso Regression

- LASSO (least absolute shrinkage and selection operator):

$$\min_{\beta_0,\ldots,\beta_j} RSS(\beta) = \varepsilon^T \varepsilon = \sum_{i=1}^{N} (y_i - f(X_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$subject\ to \sum_{j=1}^{p} |\beta_j| \le threshold$$

- Lagrange optimization:

$$\min_{\beta_0,\ldots,\beta_j} \mathrm{L}(\beta) = RSS(\beta) + \lambda \sum_{j=1}^{p} |\beta_j| = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

L1 Penalty on the magnitude of coefficient size

λ: Regularization strength

Ridge Regression

- Ridge Regression:

$$\min_{\beta_0,\ldots,\beta_j} RSS(\beta) = \varepsilon^T \varepsilon = \sum_{i=1}^{N} (y_i - f(X_i))^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$$subject\ to \sum_{j=1}^{p} \beta_j^2 \le threshold$$

- Lagrange optimization:

$$\min_{\beta_0,\ldots,\beta_j} \mathrm{L}(\beta) = RSS(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

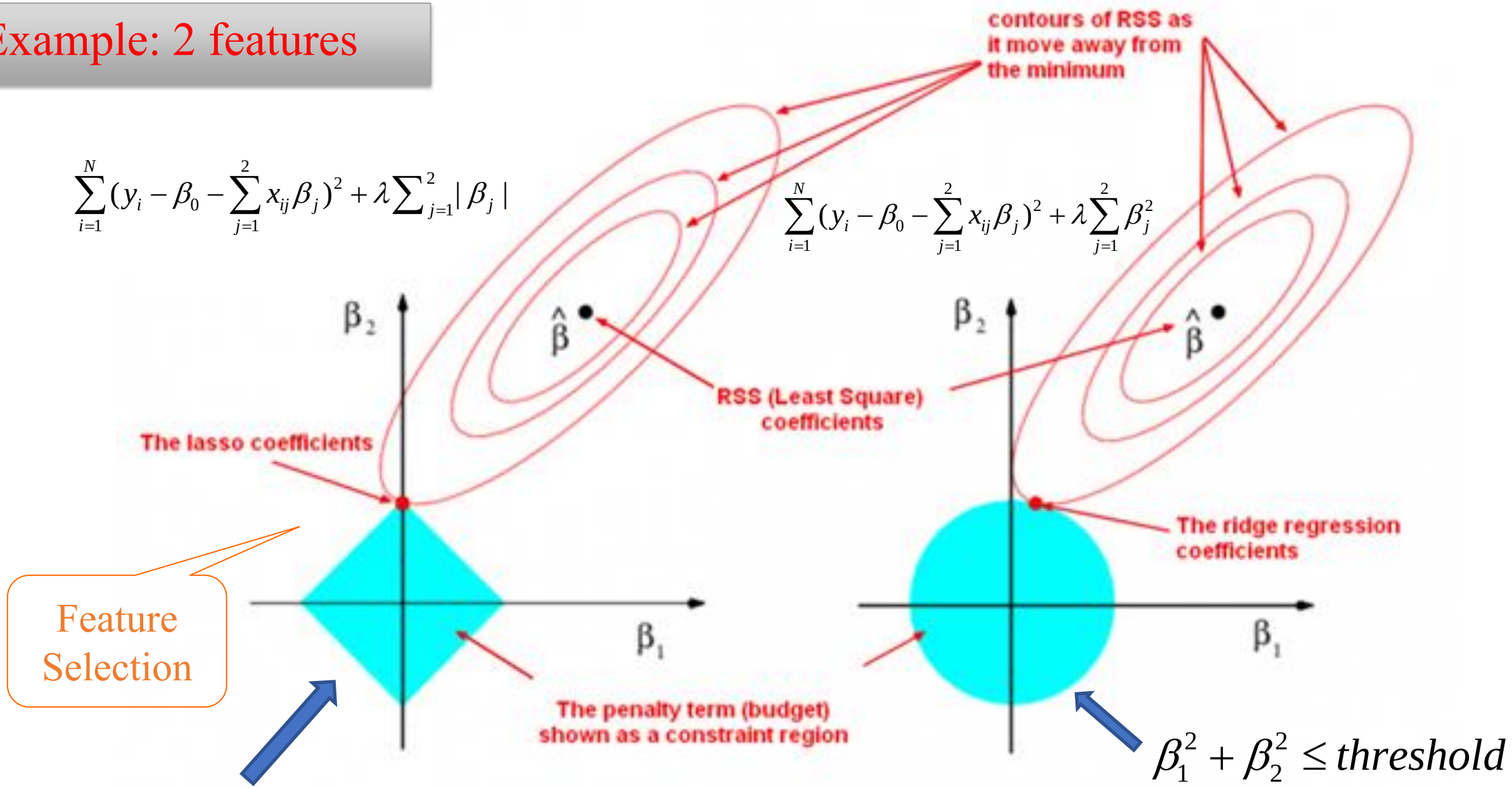
L2 Penalty on the magnitude of coefficient size

λ: Regularization strength

Example: 2 features

$$\sum_{i=1}^{N}(y_i-\beta_0-\sum_{j=1}^{2}x_{ij}\beta_j)^2+\lambda\sum_{j=1}^{2}|\beta_j|$$

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



Feature Selection

$|\beta_1|+|\beta_2|\leq threshold$

LASSO

$\beta_1^2+\beta_2^2\leq threshold$

RIDGE REGRESSION

Regression Coefficient Shrinkage

- References of Math Proof:
  - http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf
  - http://math.bu.edu/people/cgineste/classes/ma575/p/w14_1.pdf

Simplified Proof I

A two-feature model without intercept: $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

OLS: $\min\limits_{\beta_1,\beta_2} \text{RSS}(\beta_1, \beta_2) = \varepsilon^T \varepsilon = (y - \beta_1 x_1 - \beta_2 x_2)^T (y - \beta_1 x_1 - \beta_2 x_2)$

(1) Lasso: $\min\limits_{\beta_1,\beta_2} L(\beta_1, \beta_2) = \text{RSS}(\beta_1, \beta_2) + \lambda \mid \beta_1 \mid + \lambda \mid \beta_2 \mid$

For β_1: If $\beta_1 > 0 \Rightarrow FD: -y^T x_1 - x_1^T y + 2\beta_1 x_1^T x_1 + \beta_2 x_1^T x_2 + \beta_2 x_2^T x_1 + \lambda = 0$

$$\Rightarrow \beta_1 = \frac{(2x_1^T y - 2\beta_2 x_1^T x_2) - \lambda}{2x_1^T x_1} \Rightarrow \exists \lambda: \ \beta_1 = 0$$

If $\beta_1 < 0 \Rightarrow$ DIY...

Simplified Proof II

A two-feature model without intercept: $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

OLS: $\min_{\beta_1, \beta_2} \text{RSS}(\beta_1, \beta_2) = \varepsilon^T \varepsilon = (y - \beta_1 x_1 - \beta_2 x_2)^T (y - \beta_1 x_1 - \beta_2 x_2)$

(2) Ridge: $\min_{\beta_1, \beta_2} L(\beta_1, \beta_2) = \text{RSS}(\beta_1, \beta_2) + \lambda \beta_1^2 + \lambda \beta_2^2$

For β_1: $\Rightarrow FD: -y^T x_1 - x_1^T y + 2\beta_1 x_1^T x_1 + \beta_2 x_1^T x_2 + \beta_2 x_2^T x_1 + 2\lambda \beta_1 = 0$

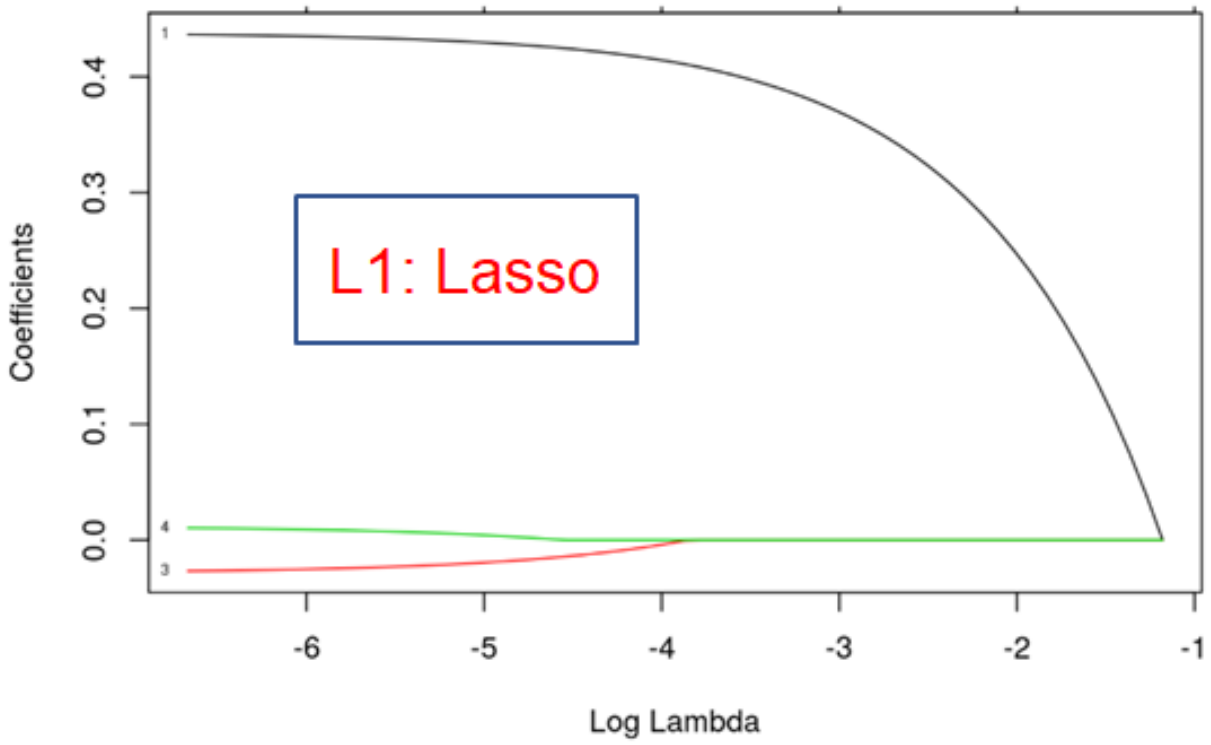
$\Rightarrow \beta_1 = \dfrac{x_1^T y - \beta_2 x_1^T x_2}{x_1^T x_1 + \lambda} \Rightarrow \forall \lambda: \ \beta_1 \neq 0$ (numerator is very unlikely to be exactly 0)

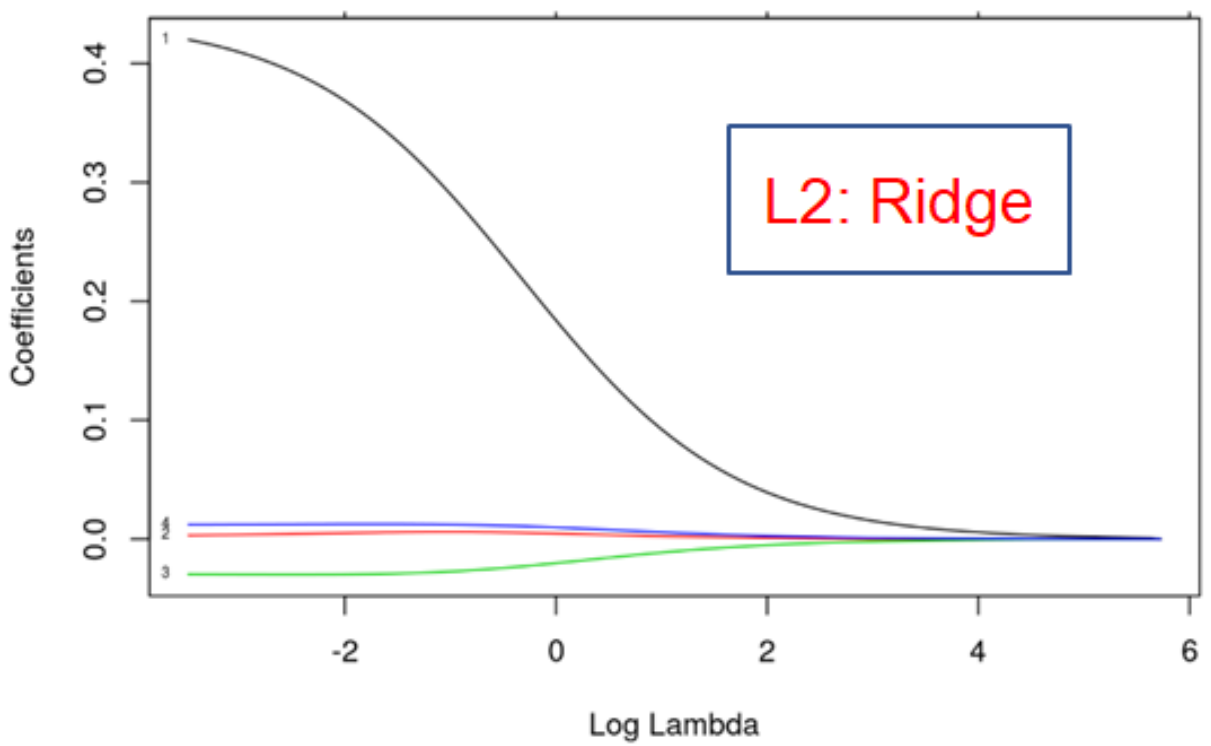
Regression Coefficient Shrinkage

$$\min \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

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

Comparison between L1 (Lasso) and L2 (Ridge) regularization





Cross-Validation

In practice…

- The cross-validation procedure is repeated K times
- K random partitions of the original sample
- The K results are again averaged (or otherwise combined) to produce a single estimation.
- You can use cross validation for both binary classification and multi-class classification problems

Cross-Validation

K-Fold Cross Validation (e.g., K=10)

- Model Evaluation
- Model Comparison
- Model Tuning

All observations are used for both training and validation, and each observation is used for validation exactly once.

Cross-Validation

Why do we need cross-validation:

- Model Evaluation
  - Assess how the prediction performance of one specific model will generalize to other new independent datasets
- Model Comparison
  - Compare prediction performance of several candidate models across different datasets
- Model Tuning
  - Find the best hyperparameters for one specific model (e.g., regularization strength λ for Lasso/Ridge)

Any Questions ?

Thank you!