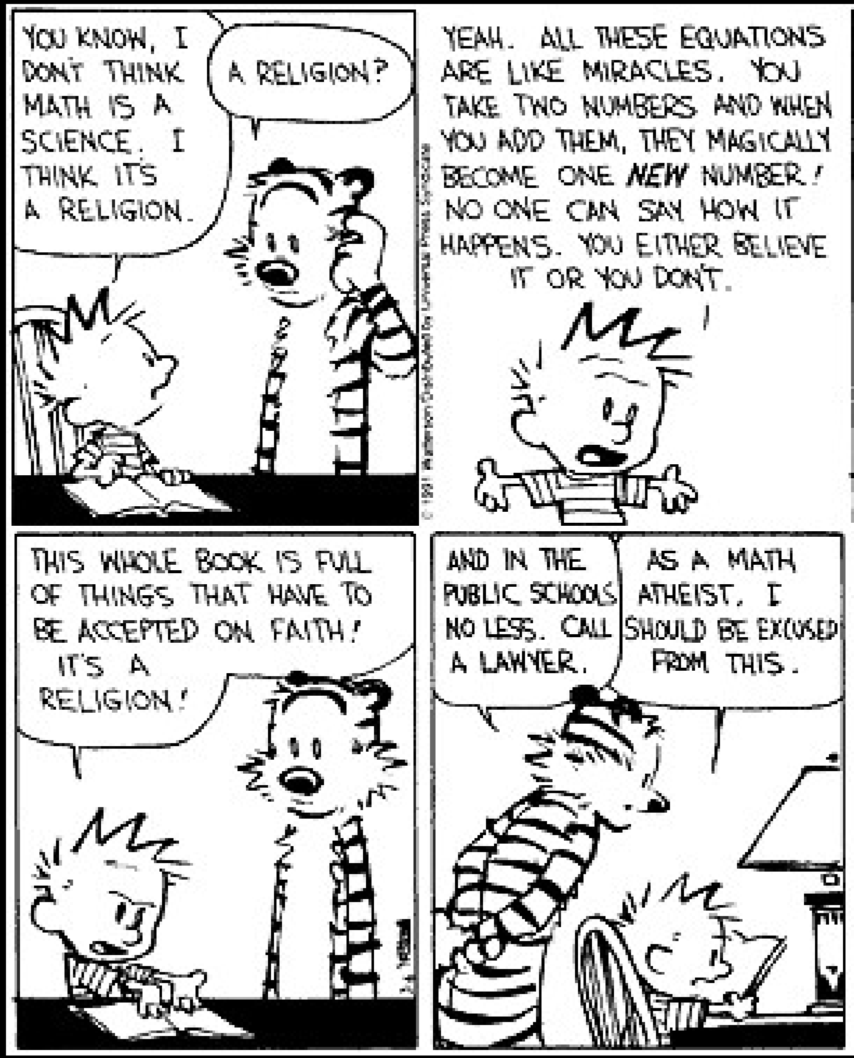


INFO 251: Applied Machine Learning

Regularization

Announcements

- Reminder: Next week's classes on Zoom online
- Assignment 3 due next week
- Quiz 1 scheduled for March 1, first ~40 minutes of class
  - 10-15 multiple choice and short-answer questions
  - See piazza for details on quiz timing

Course Outline

- Causal Inference and Research Design
  - Experimental methods
  - Non-experiment methods
- Machine Learning
  - Design of Machine Learning Experiments
  - **Linear Models and Gradient Descent**
  - Non-linear models
  - Neural models
  - Unsupervised Learning
  - Practicalities, Fairness, Bias
- Special topics

Key Concepts (last lecture)

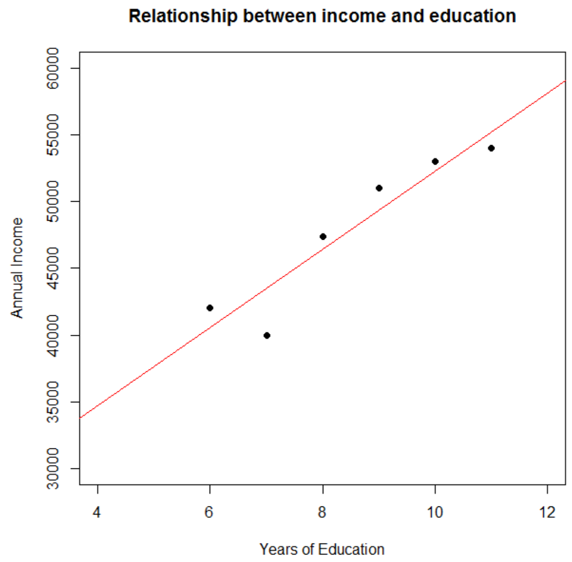
- Cost Functions
- Gradient Descent
- Local and global minima
- Convex functions
- Incremental vs. Batch GD
- Learning rates
- Feature scaling

Outline

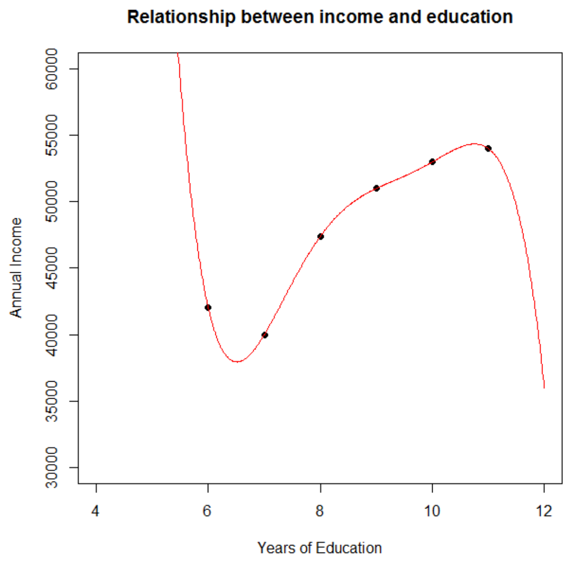
- Overfitting
- Regularization: intution
- Ridge
- Lasso

Overfitting revisited

- Overfitting: If we have too many features, our model may fit the training set very well, but fail to generalize to new examples



$$wages_i = \alpha + \beta * educ_i + error_i$$



$$wages_i = \alpha + \beta_1 * educ_i + \ldots + \beta_5 * educ_i^5 + error_i$$

Overfitting: Solutions

- Later in the course:
  - Feature selection
  - Model selection
  - Dimensionality reduction
- Now: Regularization
  - For instance, ridge regularization: Keep all the features, but reduce magnitude of specific parameters

Regularization: Intuition

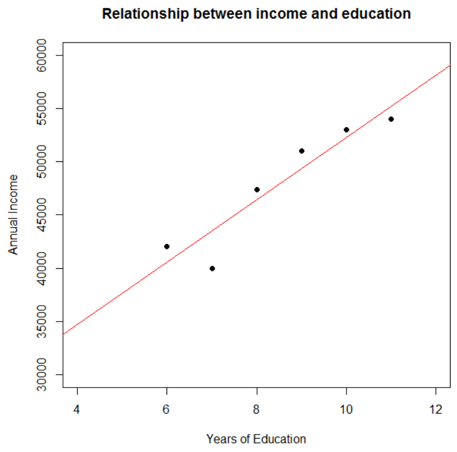
- Occam's Razor
  - A principle of parsimony, economy, or succinctness used in problem-solving. It states that among competing hypotheses, the hypothesis with the fewest assumptions should be selected.

Ockham chooses a razor

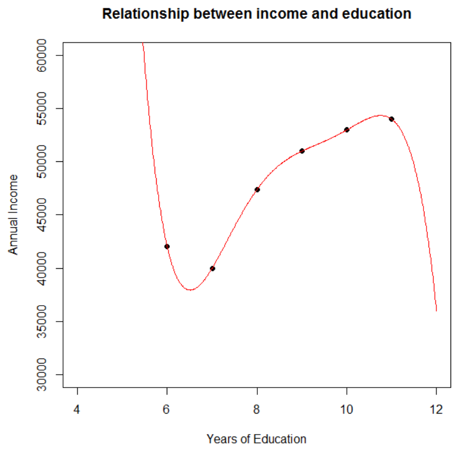
See: Domingos, P. The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery 3* (1999), 409–425.

Regularization: Intuition

- Idea: Add a cost penalty for additional complexity in the model
- Example: polynomial regression
  - Model: $Y_i = \theta_0 + \theta_1 X_i + \ \ldots + \theta_k X_i^k$
  - Parameters: $\theta_0, \ldots, \theta_k$
  - Original "Cost": $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \ldots + \theta_k X_i^k - Y_i\right)^2$



$wages_i = \alpha + \beta * educ_i + error_i$



$wages_i = \alpha + \beta_1 * educ_i + \ldots \ + \beta_5 * educ_i^5 + error_i$

Regularization: Intuition

- Original Cost
  - $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \dots + \theta_k X_i^k - Y_i\right)^2$
- Intuitive Goal
  - $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \dots + \theta_k X_i^k - Y_i\right)^2 + C(\theta_1, \dots, \theta_k)$
- Penalized (Regularized) Cost
  - $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \dots + \theta_k X_i^k - Y_i\right)^2 + \boxed{\boxed{\lambda}\sum_{j=1}^{k}\left|\theta_j\right|^{\boxed{2}}}$

penalty

"Ridge" coefficient

Regularization parameter

Regularization and Linear Regression

- Original Gradient Descent
  - Repeat until convergence:

    α <- α – R $\frac{\partial}{\partial \alpha} J(\alpha, \beta)$

    β <- β – R $\frac{\partial}{\partial \beta} J(\alpha, \beta)$

  - Original derivative of J (in linear regression, $Y_i = \alpha + \beta X_i$)

    α <- α – R$\frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta X_i - Y_i)$

    β <- β – R$\frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta X_i - Y_i) \quad X_i$

- Regularized version has new partial derivatives:

  β <- β – R$\left[\frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta X_i - Y_i) \quad X_i \; + \frac{\lambda}{N}\beta\right]$

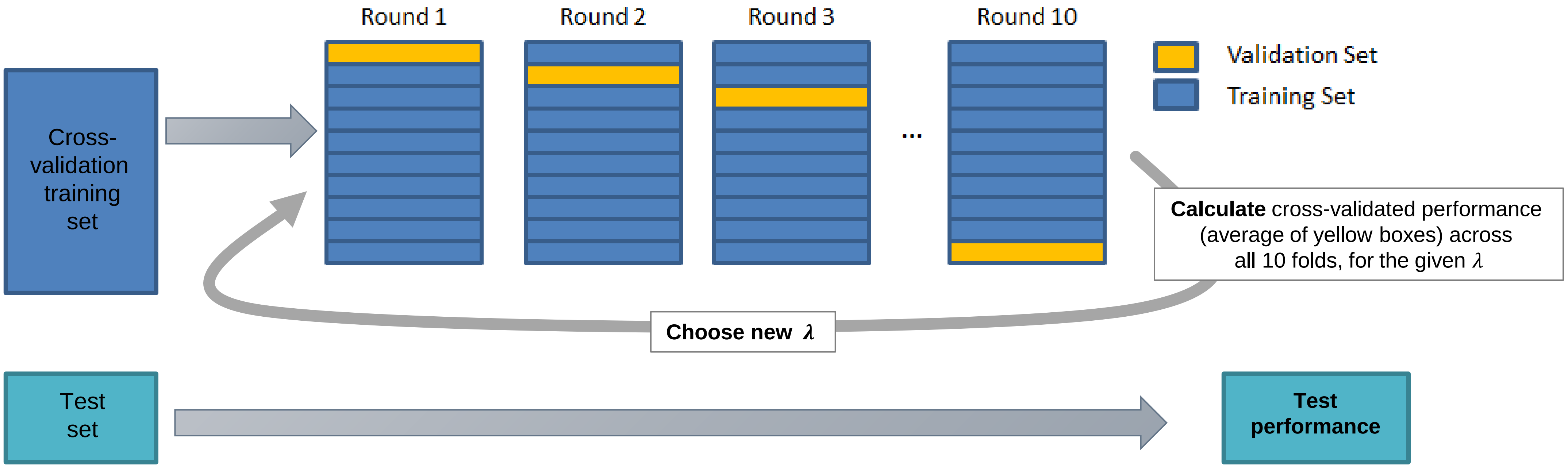
- Rewritten:

  β <- $\beta\left(1 - R\frac{\lambda}{N}\right) - R\frac{1}{N}\sum_{i=1}^{N}(\alpha + \beta X_i - Y_i) \quad X_i$

Regularization: Some notes

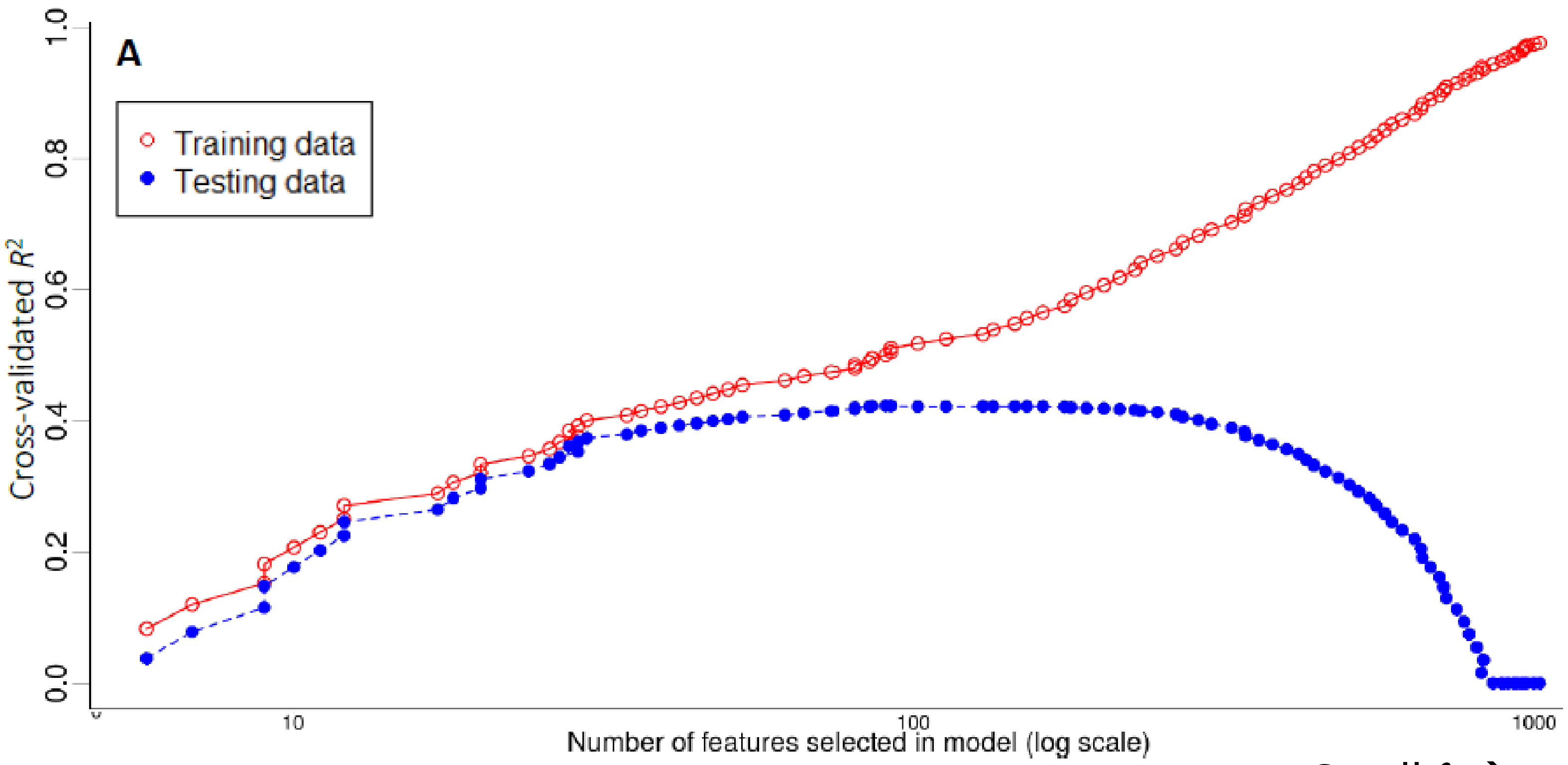
- How to select λ?
  - Cross validation!
  - Choose λ that minimizes cross-validated performance (yellow boxes)
    - i.e., repeat dark blue process for a variety of candidate values of λ

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \dots + \theta_k X_i^k - Y_i\right)^2 + \lambda \sum_{j=1}^{k} \theta_j^2$$



Regularization: Some notes

- Example



← Large λ Small λ →

Regularization: Some notes

Polynomial regression example: $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \ldots + \theta_k X_i^k - Y_i\right)^2 + \lambda \sum_{j=1}^{k} \theta_j^2$

Wages/education example: $J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(\alpha + \beta X_i + \gamma Z_i - Y_i)^2 + \lambda \sum_{j=1}^{k} \beta^2 + \gamma^2$

- What happens in regularization if features are in different units?
  - Penalty on different scales
  - One solution: Normalize features
- Do we penalize the intercept?
  - Typically, no. The intercept is typically not a sign of overfitting
  - Or: center the data around zero (Y is mean zero), regularize all coefficients

Outline

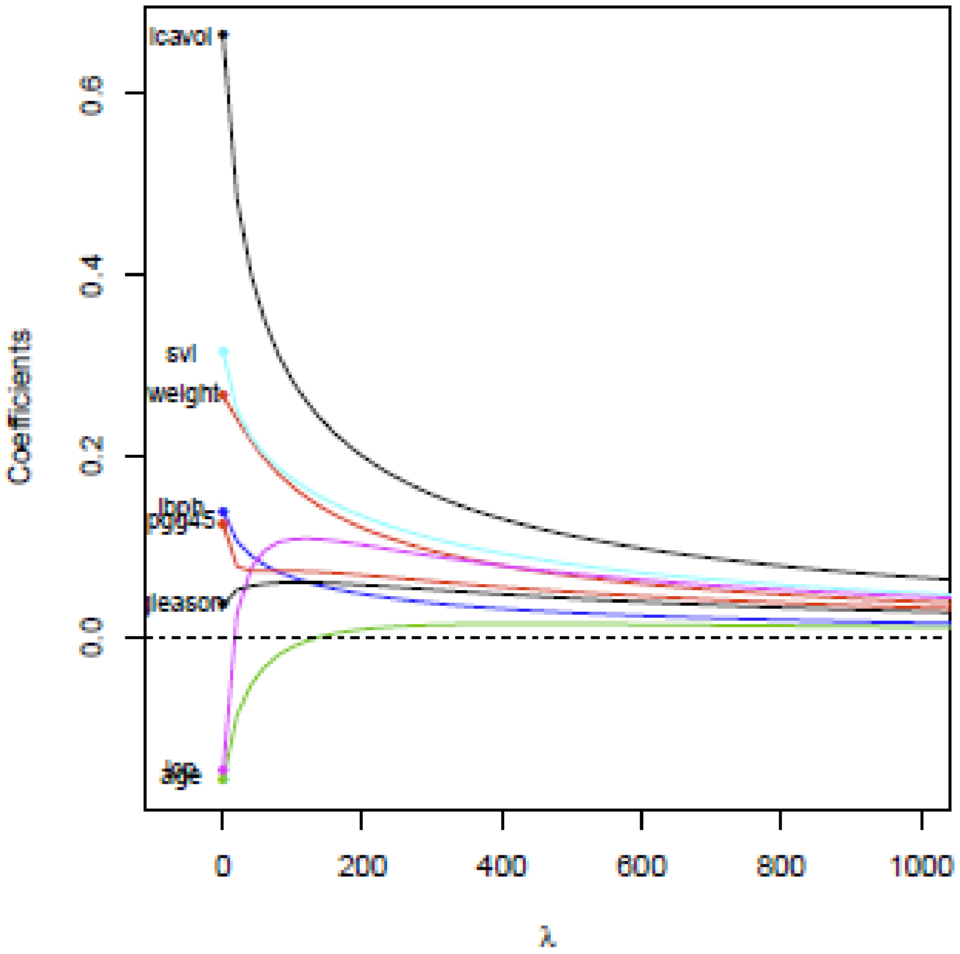
- Regularization
- **Ridge and Lasso**
- Logistic regression (inference)
- Logistic regression (prediction and gradient descent)
- Support vector machines
- Kernels

"Ridge"

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \dots + \theta_k X_i^k - Y_i\right)^2 + \lambda \sum_{j=1}^{k} \boxed{\theta_j^2}$$

- L_2 norm (ridge regression): penalty proportional to θ^2
  - Works best when a subset of the true coefficients are small
  - Will never set coefficients to zero exactly
  - Cannot perform variable selection in the linear model
  - Coefficients harder to interpret

Ridge: Coefficient plot



Source: Ryan Tibshirani

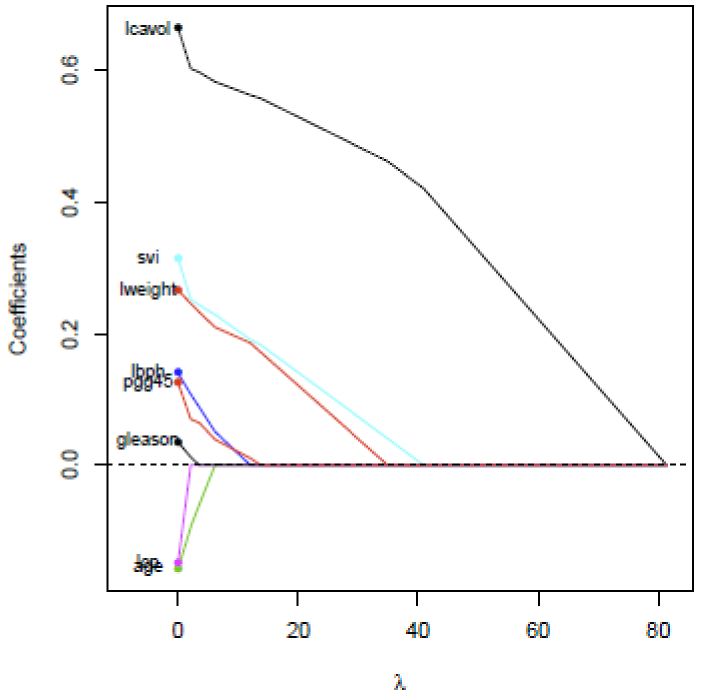
LASSO

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}\left(\theta_0 + \theta_1 X_i + \ \ldots + \theta_k X_i^k - Y_i\right)^2 + \lambda \sum_{j=1}^{k} |\theta_j|$$

- L_1 norm (lasso regression): penalty proportional to θ
  - Selects more relevant features and discards the others, vs. Ridge regression which reduces parameters but doesn't drive to zero
    - See ESL pp. 68; Andrew et al (2007). "Scalable training of L_1-regularized log-linear models".
  - Not differentiable
  - Coefficients still difficult to interpret, though "post-lasso" versions can reduce bias (e.g., Belloni & Chernozhukov)

LASSO: Coefficient plot

- Least Absolute Selection and Shrinkage Operator
  - See ESL section 3.4
  - Tibshirani (1996), “Regression Shrinkage and Selection via the Lasso"



Other forms of Regularization

Model	Fit measure	Entropy measure[4][5]
AIC/BIC	$\|Y - X\beta\|_2$	$\|\beta\|_0$
Ridge regression	$\|Y - X\beta\|_2$	$\|\beta\|_2$
Lasso[6]	$\|Y - X\beta\|_2$	$\|\beta\|_1$
Basis pursuit denoising	$\|Y - X\beta\|_2$	$\lambda\|\beta\|_1$
Rudin-Osher-Fatemi model (TV)	$\|Y - X\beta\|_2$	$\lambda\|\nabla\beta\|_1$
Potts model	$\|Y - X\beta\|_2$	$\lambda\|\nabla\beta\|_0$
RLAD[7]	$\|Y - X\beta\|_1$	$\|\beta\|_1$
Dantzig Selector[8]	$\|X^\top(Y - X\beta)\|_\infty$	$\|\beta\|_1$
SLOPE[9]	$\|Y - X\beta\|_2$	$\sum_{i=1}^{p} \lambda_i \|\beta\|_{(i)}$

A linear combination of the LASSO and ridge regression methods is elastic net regularization.