Statistical Decision Theory & Linear Regression

CS 534: Machine Learning

Slides adapted from Lee Cooper, Joydeep Ghosh, Carlos Carvalho, and Ryan Tibshirani

Supervised Learning

- Given a set of variables (features, input, predictors, or independent variables), can we predict the value of one or more outputs (responses, or dependent variables)?

Paired inputs/outputs

$$\{(\mathbf{x}_i, y_i)\}, i = 1, \cdots, N$$

Response Types

- Quantitative — typically continuous valued, natural ordering

- Qualitative — values in a finite set

    - Categorical (discrete): no natural ordering (think object classes)

- Ordered categorical: ordering between values with no distance

    - Example: small, medium, or large

Quantitive Response: Examples

- Predict future C02 level given economic growth data (0-? ppm)

- Predict value of a pixel in a digitized image from the values of neighbor pixels (0-255)

- Predict risk groups of cancer patients given genomic data (poor, moderate, good)

http://news.mit.edu/2016/teaching-machines-to-predict-the-future-0621

Qualitative Response: Examples

- Given height, weight, predict sex {male, female}
- Predicting handwritten digit from image {0,…,9}
- Predict species of iris from petal measurements {Virginica, Setosa, Versacolor} (R.A. Fisher, 1936)



Prediction Tasks

- Regression is used for quantitative responses
- Classification for predicting qualitative responses

- Responses encoded numerically (e.g., {0,1}, {1,2,3,…} or {-1, 1})
- Sometimes referred to as targets

Notation Conventions: Inputs

- X: features or inputs
- N: number of samples
- p: number of features per sample
- *X*: N x p feature matrix
  - Each row is a sample / datapoint
  - Each column is a dimension

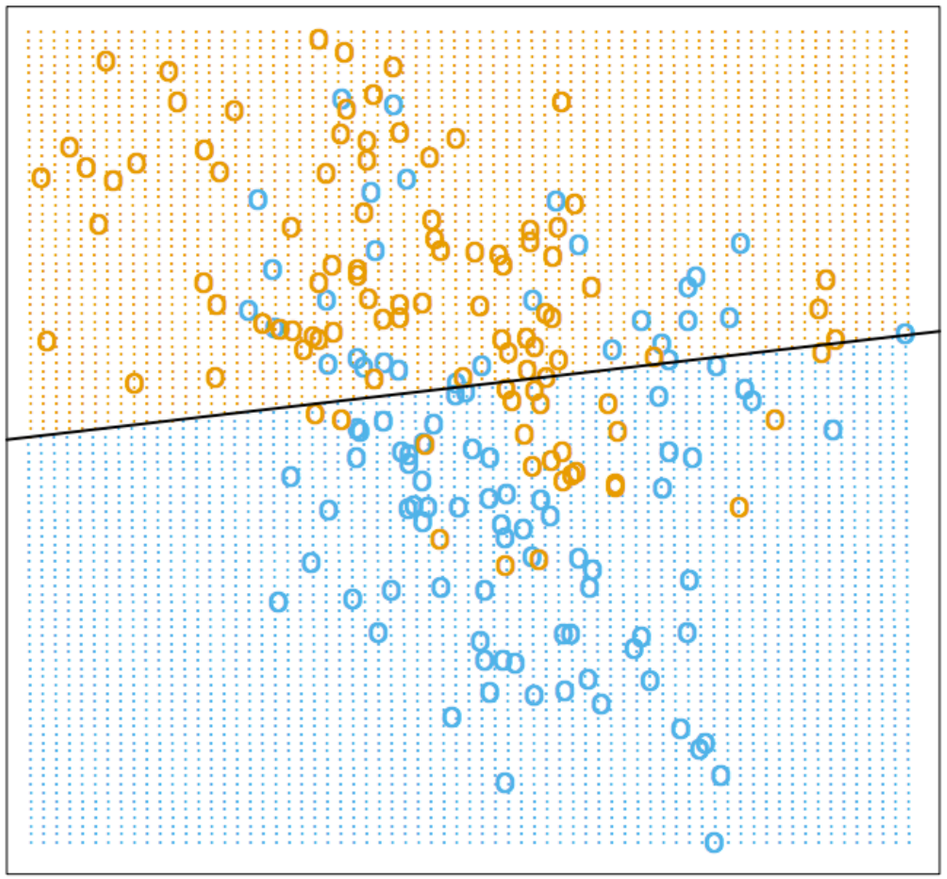
Beware: Not all texts and algorithms use the same notation!

Notation Conventions: Outputs

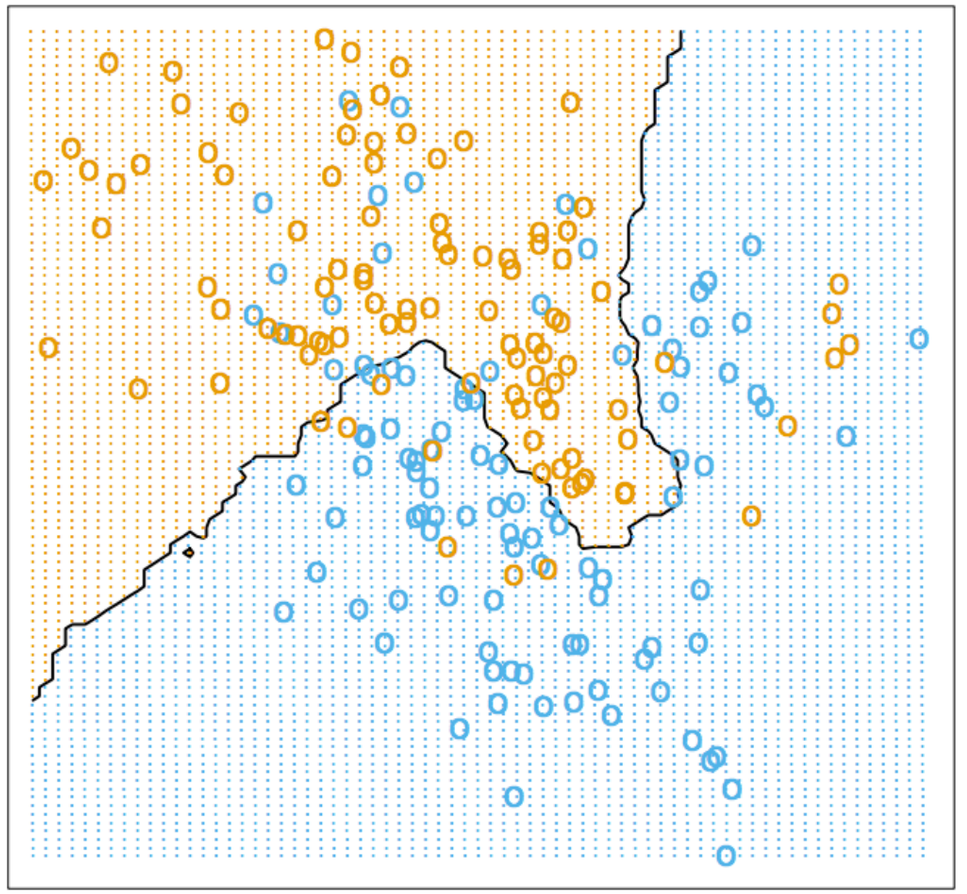
- Y : N x 1 vector with quantitative response
- G : N x 1 vector with qualitative response
- $\hat{Y}, \hat{G}$: Predicted responses
- (x_i, y_i) or (x_i, g_i) : training data

Statistical Decision Theory

Two Basic Predictor Models



Linear Model



k-Nearest Neighbors

Decision Theory

- Can we build a theory of decision making that has foundations in statistical science?

- Let's start with quantitative responses

Decision Theory Preliminaries

- Let X be a real-valued input vector and Y a real valued random output variable

- Let X, Y be jointly distributed Pr(X, Y)

- Find a function f(X) that predicts Y from X

$$f(X) : \mathbb{R}^p \rightarrow \mathbb{R}$$

Decision Theory: Loss Function

- Loss function: definition of performance to penalize errors in prediction

$$L(Y, f(X))$$

- Expected (squared) prediction error (EPE)

$$\begin{aligned}\mathrm{EPE}(f) &= E[Y - f(X)]^2 \\ &= E_X E_{Y|X}[(Y - f(X))^2 | X]\end{aligned}$$

- Optimal solution:

$$\begin{aligned}f(x) &= \mathrm{argmin}_c E_{Y|X}[(Y - c)^2 | X = x] \\ &= E[Y | X = x]\end{aligned}$$

Conditional mean is best predictor

Interpretation of 2 Basic Predictors

- Squared error loss: $L(Y, f(X)) = (Y - f(X))^2$

- kNN takes average over local neighborhood

$$\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x))$$

- Linear regression uses a model-based approach

$$\beta = (E[XX^\top])^{-1} E[XY]$$

- Different approximations for conditional expectations

  - KNN uses locally constant piecewise functions

  - LS uses global function Y = BX

Different Loss Function

- Effect of different loss function

$$L(Y, f(X)) = |Y - f(X)|$$

- Optimal solution:

$$\hat{f}(x) = \text{median}(Y|X = x)$$

- Median vs average!

Linear Regression

Regression Overview

- Most widely used statistical tool for understanding relationships amongst variables
- Conceptually simple method for investigating functional relationships between one or more factors and an outcome of interest
- Relationship is expressed in form of equation or model connecting the dependent variable (response) to one or more explanatory or predictor variables

Regression Examples

- Straight prediction questions
- How much will my house sell for?
- How many runs will the Braves score in 2017?
- What rating will I give this movie?

Regression Examples (2)

- Explanation and understanding
- What is the impact of an MBA on income?
- Does Walmart discriminate against women with regards to salaries?

Regression Formulation

- Given an input vector $\mathbf{x}^T = (x_1, x_2, \ldots, x_p)$, we want to predict the quantitative response Y

- Linear regression form:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

Least Squares

- Minimize sum of square errors (RSS)

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

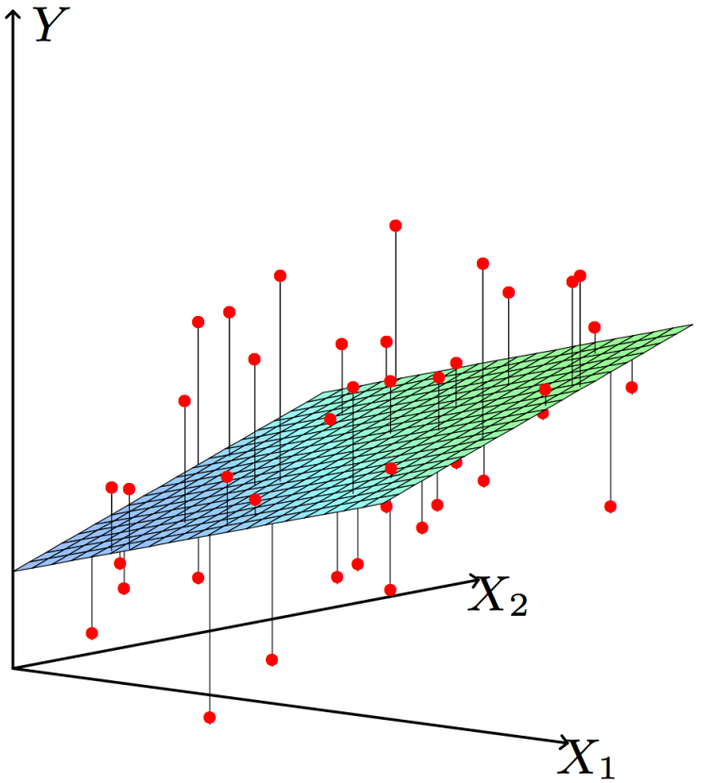


Figure 3.1 (Hastie et al.)

Least Squares Solution

- Differentiate with respect to $\boldsymbol{\beta}$ and set to zero

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

- Show second derivative is positive

$$\frac{\partial^2 \mathrm{RSS}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = 2\mathbf{X}^\top\mathbf{X} > 0$$

- Unique solution:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Geometry of Least Squares

- Outcome vector is orthogonally projected onto hyperplane spanned by input features

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

- Takeaway: Restriction by the choice of features

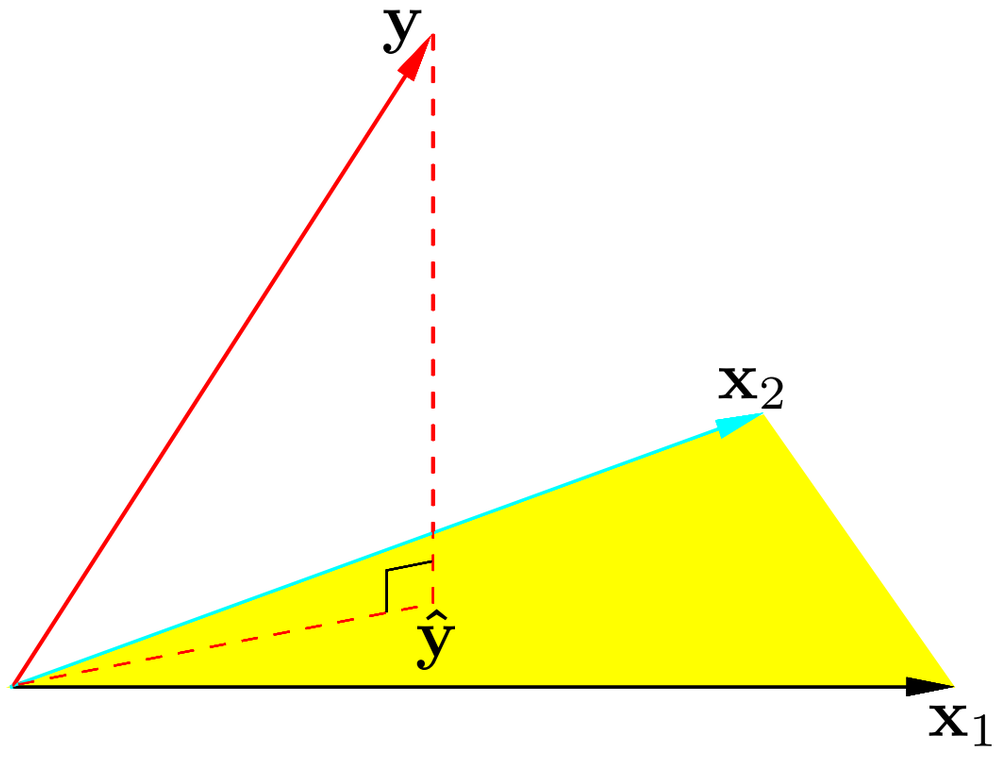
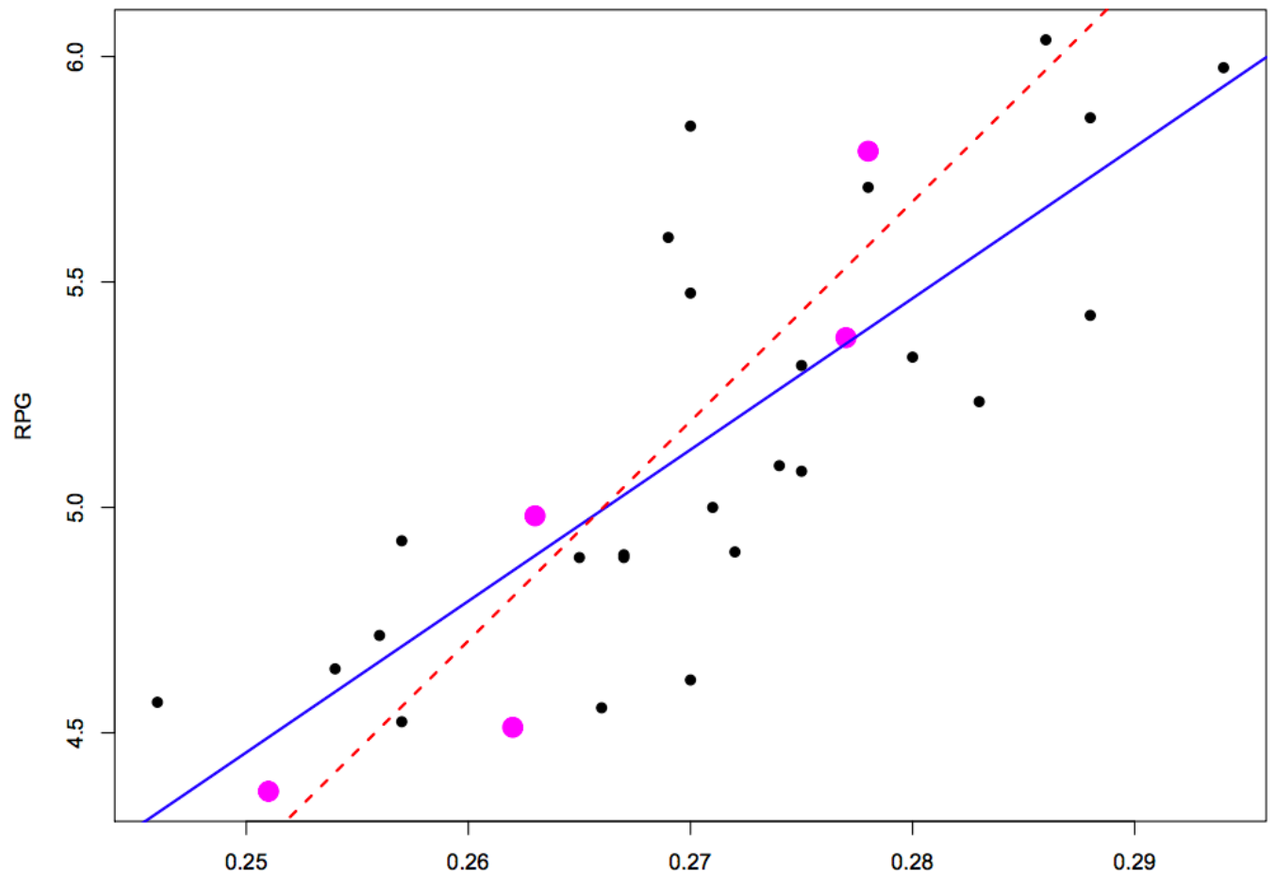


Figure 3.2 (Hastie et al.)

Model Inference and Parameters

- How certain are we about our model and the parameters?

- What if we have only the purple points on the graph (dashed line)? What if we have all the points (solid)? Which line is better?



Need notion of “true line” and a probability distribution

Regression Model Assumptions

- Conditional mean of y is linear in the predictor variables
- Error terms
- Normally distributed (Gaussian)
- IID (constant variance)

$$\mathbf{y} = E[\mathbf{y}|\mathbf{x}_1, \cdots, \mathbf{x}_p] + \varepsilon, \ \ \varepsilon \sim N(0, \sigma^2)$$

$$= \beta_0 + \sum_{j=1}^{p} \mathbf{x}_j \beta_j + \varepsilon$$

Coefficient Interpretation

- jth regression coefficient:

$$\beta_j = \frac{\partial E[\mathbf{y}|\mathbf{x_1}, \cdots, \mathbf{x_p}]}{\partial \mathbf{x_j}}$$

- Interpretation: Holding all other variables constant, β_j is the average change in ***y*** per unit change in $\boldsymbol{x}_j$

Coefficient Distribution

- Sample distribution tells us how close we expect the estimator to be from true value

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2)$$

- Estimator is unbiased $E[\hat{\beta}] = \beta$

- Standard deviation or standard error determines how close estimator is to true value

Model Variance

- Observations uncorrelated and have constant variance σ^2

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top \mathbf{X})^{-1}\sigma^2$$

- Variance estimate (regression standard error)

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- Chi-squared distribution

$$(N-p-1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1}$$

Coefficient Z-score

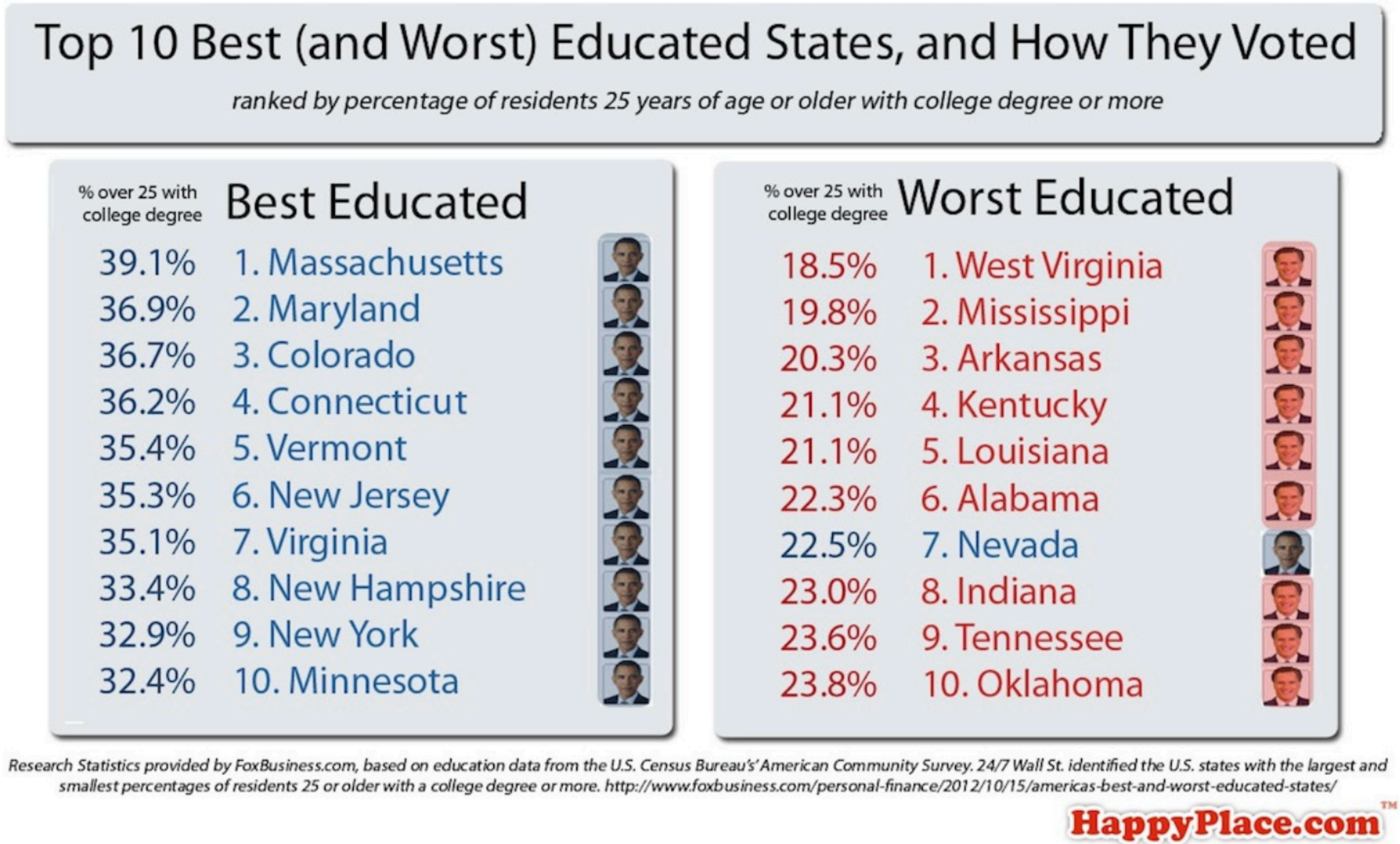
- Impact of feature on our model (null hypothesis coefficient is zero)

- Z-score is normalized coefficient that measures the predictive value of this feature

  - t-distribution with N - p - 1 degrees of freedom

  - Large values means we can reject null hypothesis

Correlation vs Causation

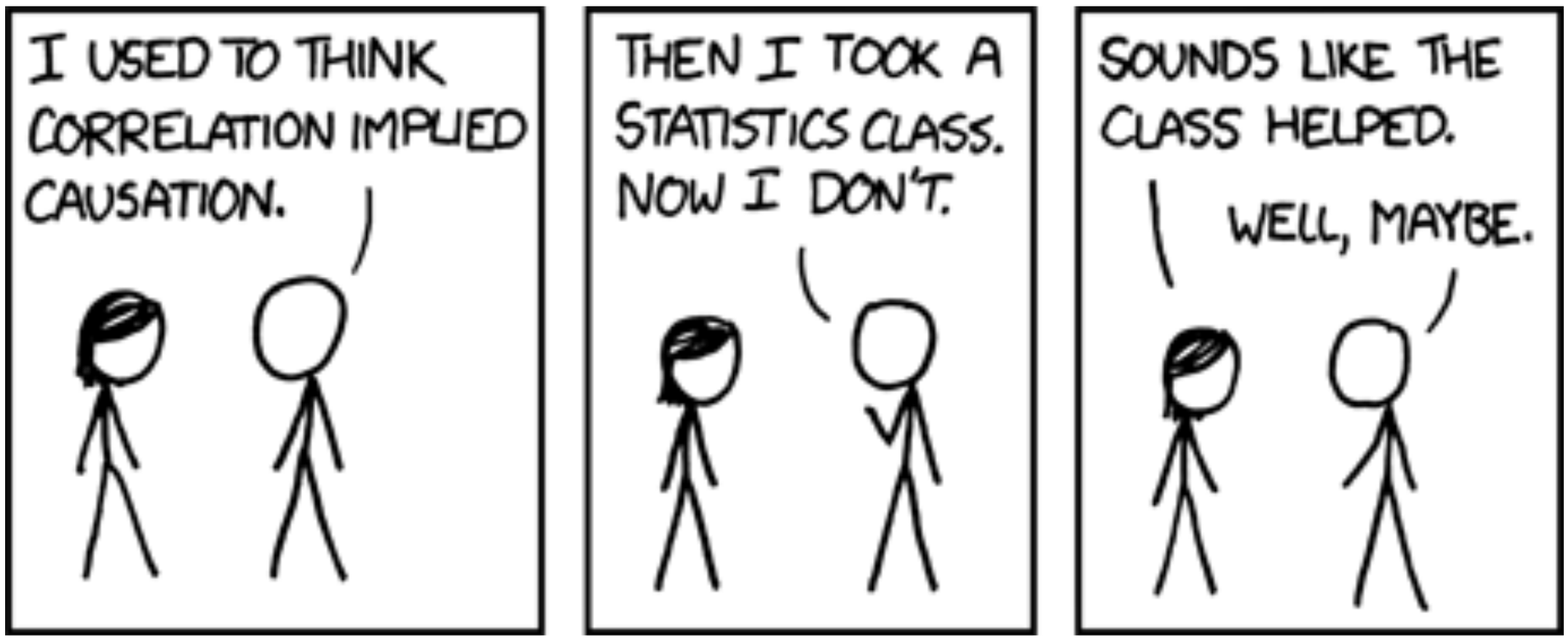


Understanding MLR

- Extremely hard to find “causal” relationships between features and outcome

- Any correlation (association) could be caused by other variables in the background — correlation is NOT causation

- Multivariate regression allows us to control for all important variables by including them in the regression



http://imgs.xkcd.com/comics/correlation.png

Example: Prostate Cancer

Predict prostate-specific antigen (PSA) levels in blood test using the following clinical variables:

- lweight: log(prostate weight)
- age: patient age
- lbph: log(benign prostatic hyperplasia)
- lcavol: log(cancer volume)
- svi: seminal vesicle invasion
- lcp: log(capsular penetration)
- gleason: pathologic grade
- pgg45: % gleason score 4 or 5

Example: Prostate Cancer

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

Table 3.2 (Hastie et al.)

Gauss-Markov Theorem

- One of most famous results in statistics
- Unbiased estimator: $E[\hat{\beta}] = \beta$
- Least squares estimates of the parameters have smallest variance among all linear unbiased estimates
  - AKA best linear unbiased estimator (BLUE)

Linearly Dependent Features

- Solution assumes that columns of X are linearly independent and full rank

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- What happens when two features are correlated?
  - Least squares coefficients not uniquely defined
  - Features should be reduced by filtering or regularization

Regression Coefficients: Formula

- Let $\mathbf{z}_j$ be the residual from regression of $\mathbf{x}_j$ on all other predictors, the nth coefficient can be expressed as

$$\hat{\beta}_j = \frac{< \mathbf{z}_j, \mathbf{y} >}{||\mathbf{z}_j||_2^2}$$

- jth coefficient is the univariate regression coefficient of ***y*** on the residuals after regressing x_j on the others

- If $\mathbf{x}_j$ is highly correlated with the rest, residual $\mathbf{z}_j$ is close to 0 which makes coefficient unstable

Variance Inflation

- Variance of jth multiple regression coefficient

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\mathrm{Var}(< \mathbf{z}_j, \mathbf{y} >)}{||\mathbf{z}_j||_2^4} = \frac{||\mathbf{z}_j||_2^2 \sigma^2}{||\mathbf{z}_j||_2^4} = \frac{\sigma^2}{||\mathbf{z}_j||_2^2}$$

- Correlated predictors inflates the variance of the coefficients

- Regression coefficient of highly correlated value will likely not be significant

Regression: Basis Functions

- Generalize features to basis functions

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{p} \phi_i(x_i)\beta_i$$

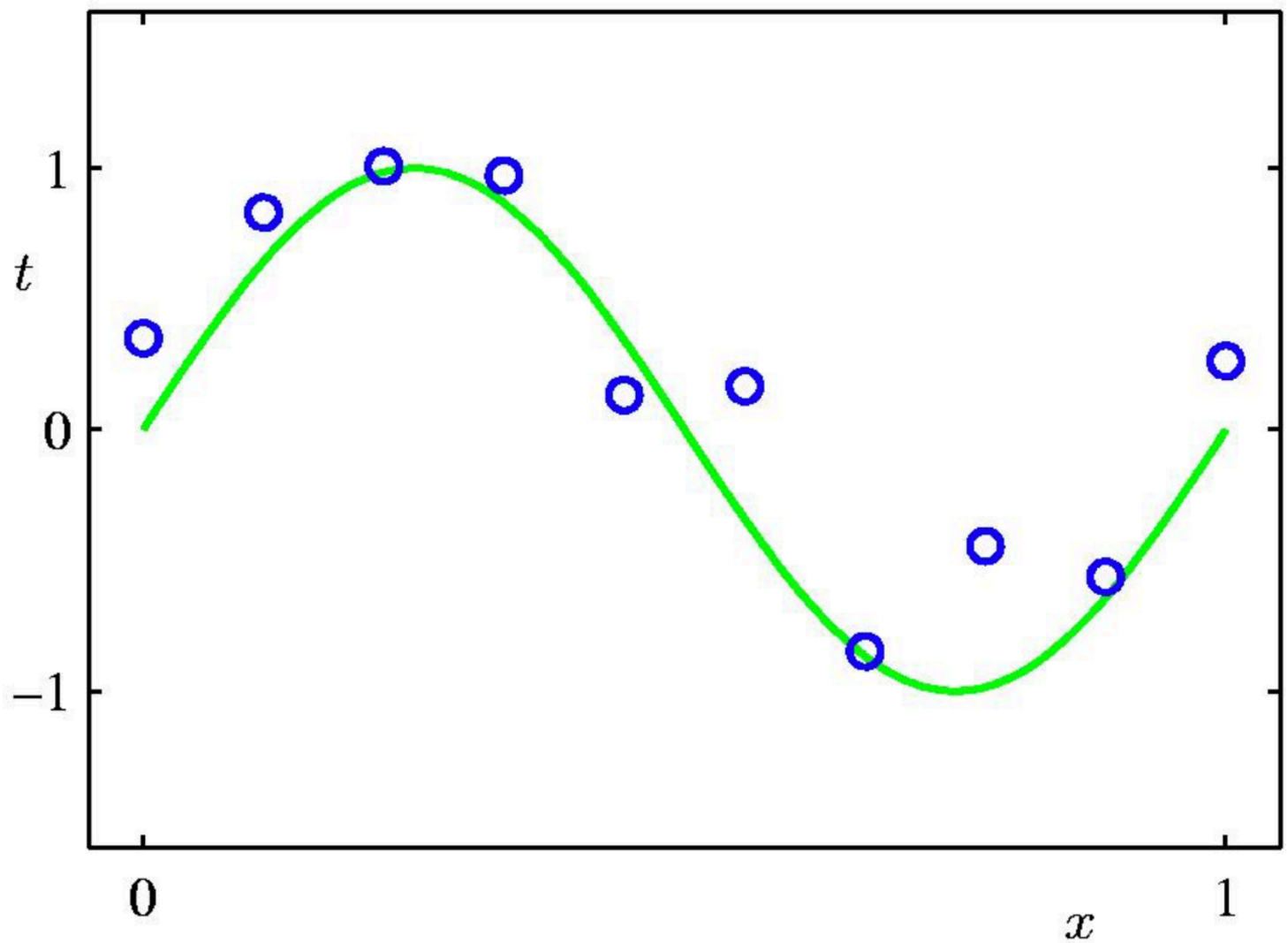
- Special case: linear regression

- Special case: polynomial regression

$$\phi_i(x) = x^i$$

Polynomial Curve Fitting

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_M x^M = \sum_{j=0}^{M} \beta_j x^j$$



LS Estimates: Shortcomings

- Prediction accuracy: Large variance for models
  - Shrink or remove variables to reduce variance
- Interpretation: Large number of predictors makes it hard to understand
  - Sacrifice small details for “big picture”

Feature Selection

Selecting the Best Features

- Brute force method: try all combinations
  - Computationally infeasible for large number of features
- Even with unlimited computational power, what is the optimal number of features?
- How to weigh complexity of the model against the error (RSS)?

Best-Subset Selection

- Finds the subset of size k with the smallest residual sum of squares
  - Leaps and bounds procedure (Furnival and Wilson, 1974) is an efficient algorithm for $p < 40$
- Best-subset curve is necessarily decreasing — cannot be used to select subset size k
  - Tradeoff will be discussed in a few lectures

Stepwise Selection

- Forward: Start with 0 features and sequentially add feature that best improves fit
  - Can be used whenever
- Backward: Start with full model, remove feature that is least detrimental to fit
  - Can only be used when N > p

Feature Selection Comparison

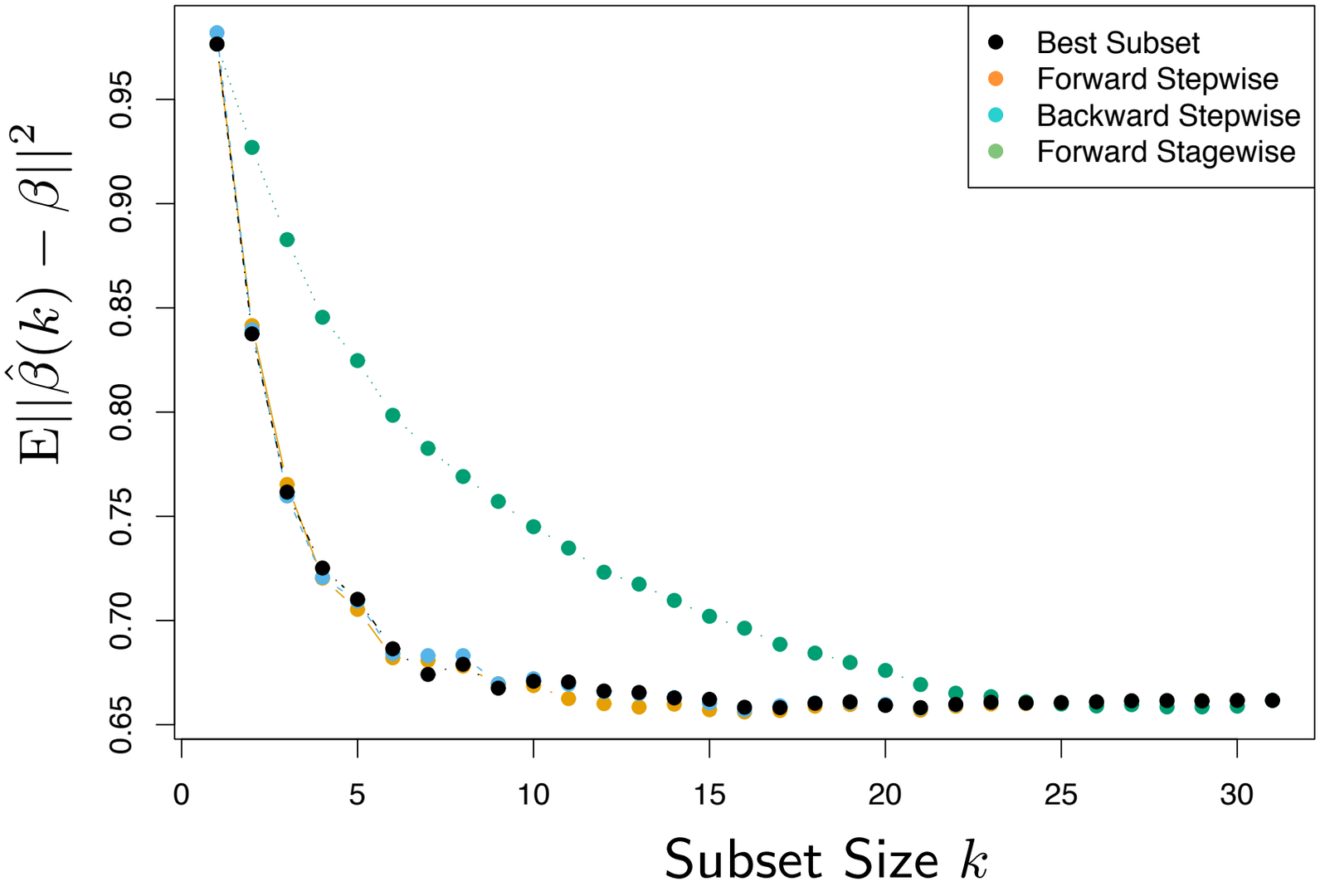


Figure 3.6 (Hastie et al.)

Regularization

Model Regularization

- Basic idea: Add penalty term on model parameters to achieve a more simple model or reduce sensitivity to training data

$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda \text{penalty}(\boldsymbol{\beta})$$

- Reasons:
  - Less prone to overfitting
  - Get the “right” model complexity

Popular Penalties

Name	Penalty function
Ridge	$\|\|\beta\|\|_2$
Lasso	$\|\|\beta\|\|_1$
L_0 regularization	$\|\|\beta\|\|_0$
Elastic net	$\alpha\|\|\beta\|\|_1 + (1-\alpha)\|\|\beta\|\|_2$

Ridge Regularization

- Regularization coefficient controls effective model complexity

$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda ||\boldsymbol{\beta}||_2$$

- Discourage large values

- Also known as shrinkage (statistics) or weight decay (neural nets)

- Closed form solution

$$\hat{\boldsymbol{\beta}} = (\lambda \mathbf{I} + \mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

leads to numerical stability too!

Ridge Regression: Tuning Parameter

- Tuning parameter (λ) controls the strength of the penalty term

- When 0, linear regression estimate

- When infinity, coefficients go to 0

- For in between, balance the fit of the model with shrinking coefficients

Coefficient Path: Ridge

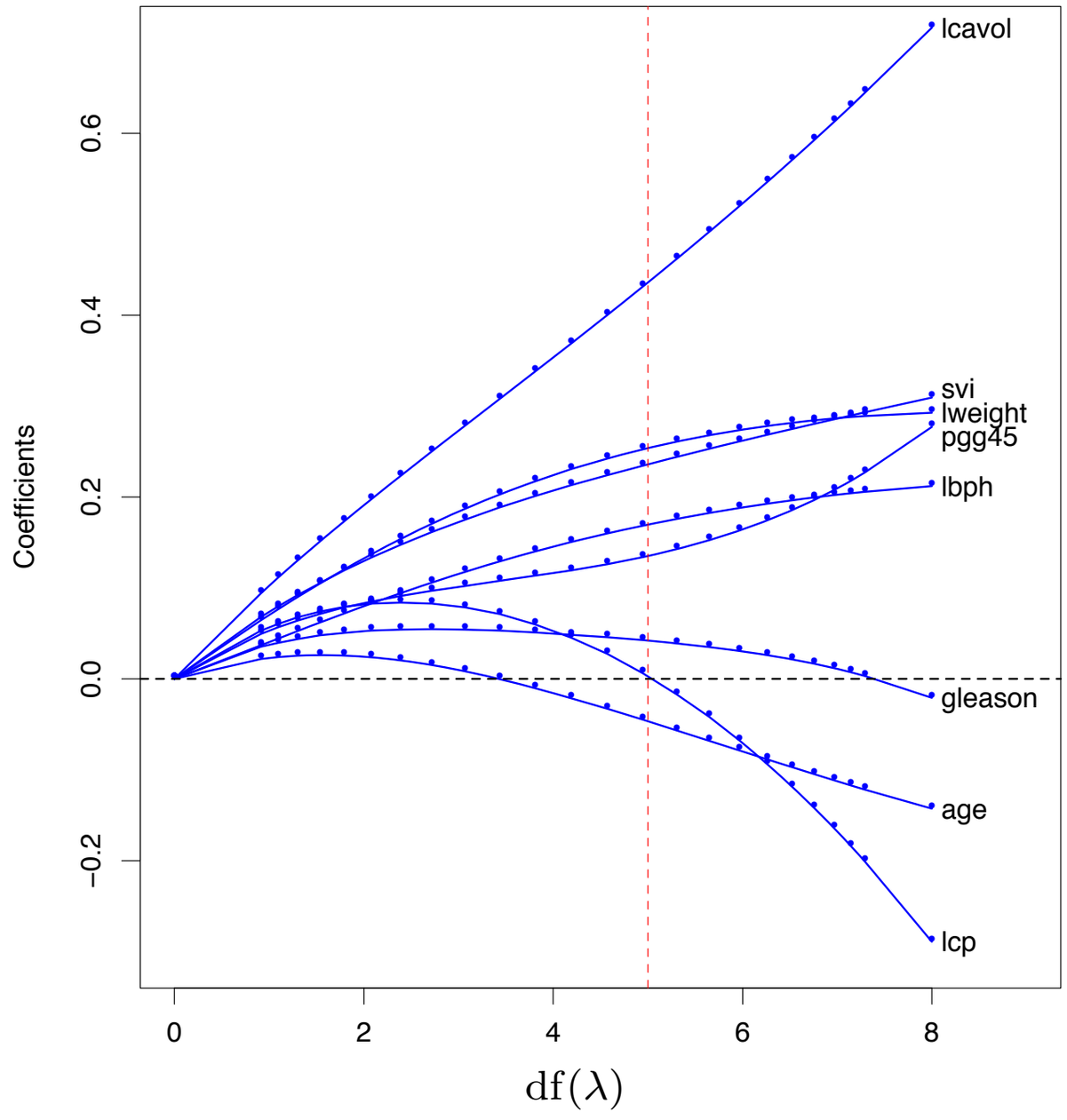


Figure 3.8 (Hastie et al.)

Lasso Regularization

- Subtle difference from the ridge is the use of the 1-norm

$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda ||\boldsymbol{\beta}||_1$$

- Large values drive coefficients to zero (continuous subset selection)

- Also known as basis pursuit in signal processing

- No closed form solution but efficient algorithms exist with approximately same computational cost as ridge

Lasso: Least Absolute Selection and Shrinkage Operator

Coefficient Path: Lasso

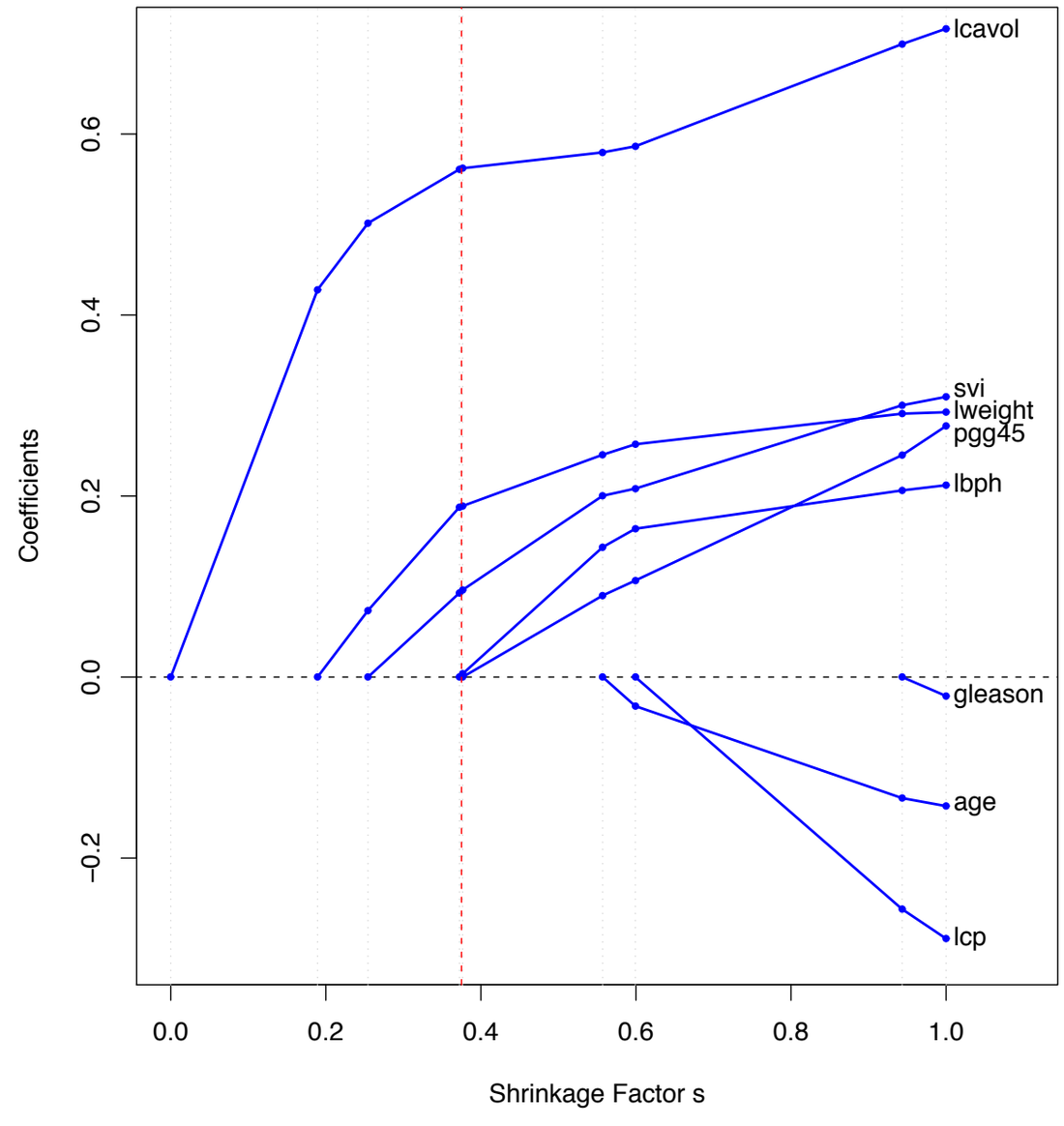
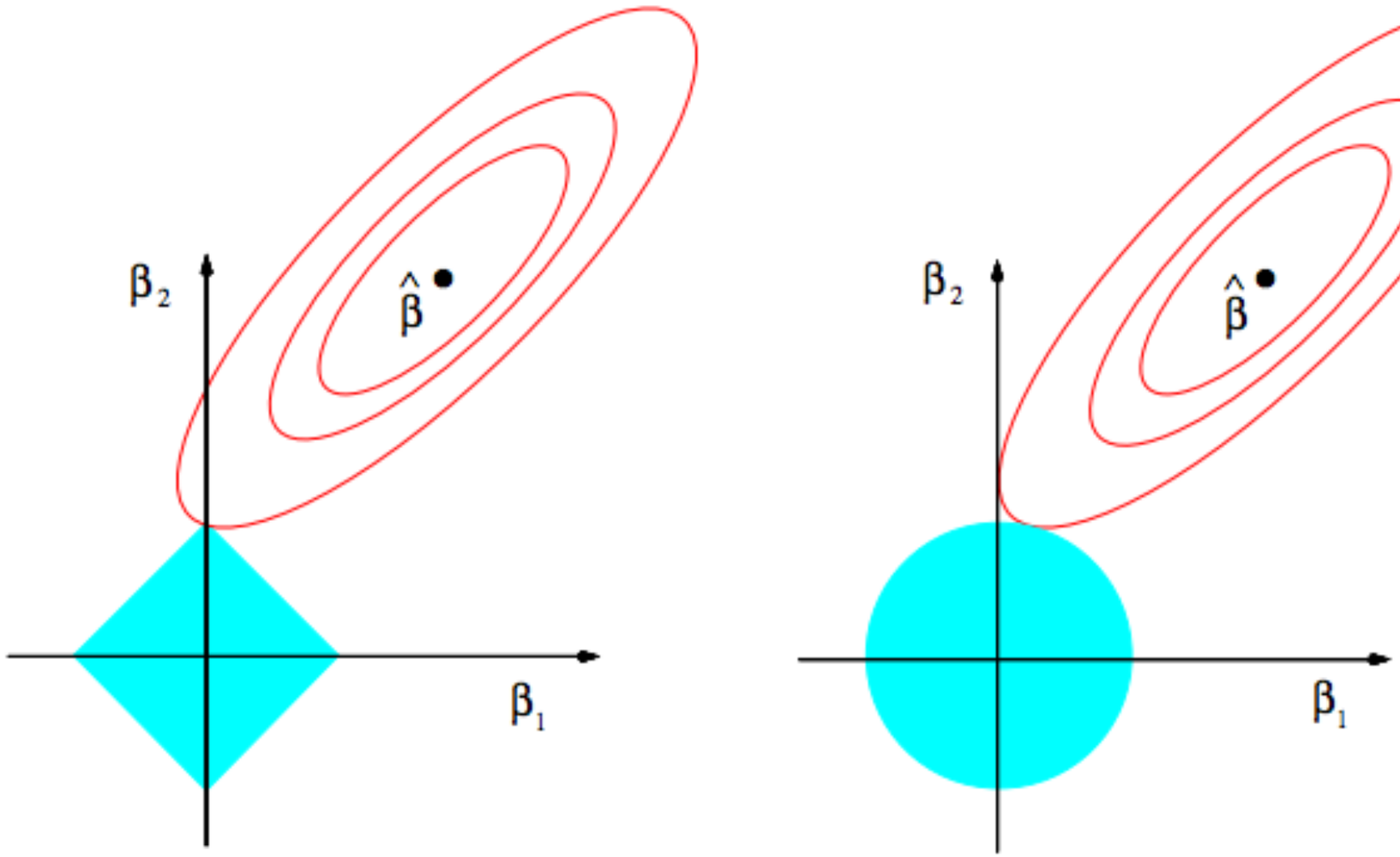


Figure 3.10 (Hastie et al.)

Ridge & Lasso Regularization: Notes

- If intercept term is included in regression, this coefficient is left unpenalized
  - Usually center the columns of X to exclude intercept
- Penalty term can be unfair if predictors are on different scales
  - Scale columns of X to have same sample variance

Ridge and Lasso Comparison



Lasso

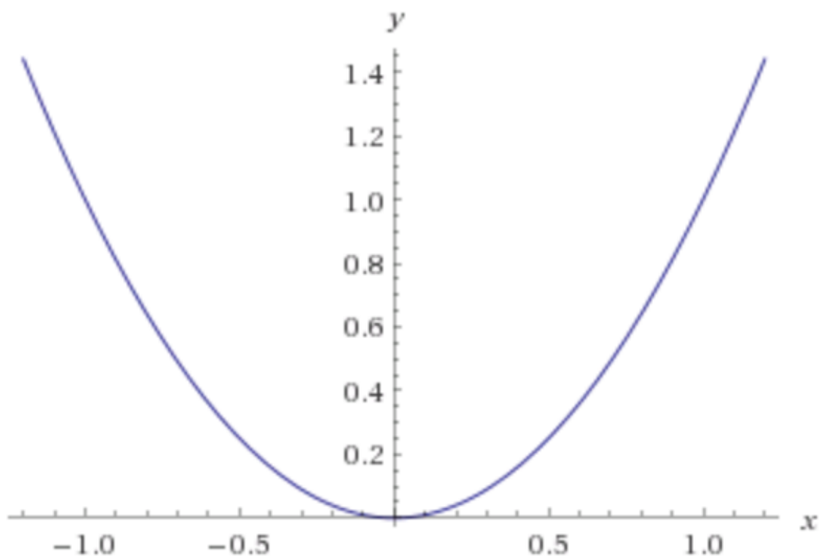
Ridge

Figure 3.11 (Hastie et al.)

Intuition: Ridge and Lasso

What happens for gradient descent?

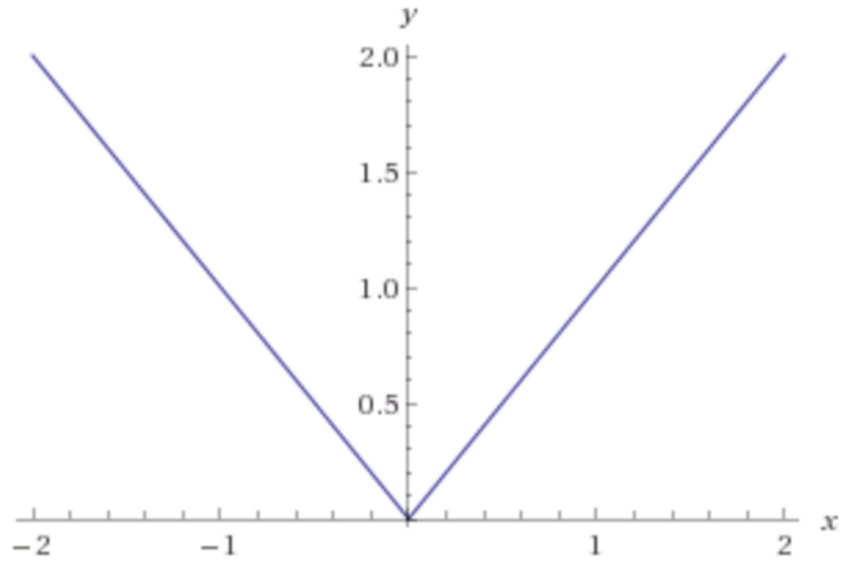
Ridge



$$\frac{\partial}{\partial x}\lambda||x||_2 = \pm\lambda x$$

Push towards 0 gets weaker as x gets smaller

Lasso



$$\frac{\partial}{\partial x}\lambda||x||_1 = \pm\lambda$$

Always pushes elements towards 0

What You Should Be Thinking

- How to choose an appropriate value of ?
  - Hard question that will be discussed a bit later
- What happens if none of the coefficients are small?
  - Regularization may still help as it greatly reduces the variance of our prediction while introducing some bias
  - This will be discussed further in a few lectures

Effect of Selection on Coefficients

Term	LS	Best Subset	Ridge	Lasso
`Intercept`	2.465	2.477	2.452	2.468
`lcavol`	0.680	0.740	0.420	0.533
`lweight`	0.263	0.316	0.238	0.169
`age`	−0.141		−0.046	
`lbph`	0.210		0.162	0.002
`svi`	0.305		0.227	0.094
`lcp`	−0.288		0.000	
`gleason`	−0.021		0.040	
`pgg45`	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

Table 3.3 (Hastie et al.)

Elastic Net Regularization

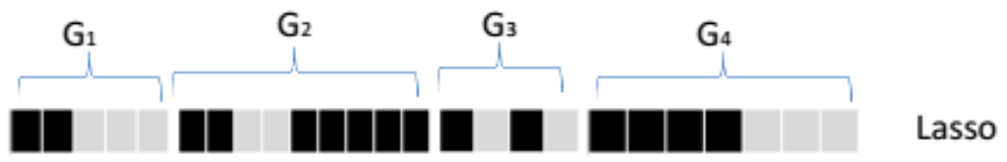
- Compromise between ridge and lasso

$$\min_{\beta} L(\mathbf{X}\boldsymbol{\beta}, \mathbf{y}) + \lambda(\alpha||\boldsymbol{\beta}||_2 + (1-\alpha)||\boldsymbol{\beta}||_1)$$

- Selects variables like lasso
- Shrinks coefficients of correlated predictions like ridge
- Computational advantages over general L_q penalties

Group Lasso for Sparse Learning

- SLEP package (http://www.yelab.net/software/SLEP/)
- Variety of methods to shrink parameters
  - Group Lasso
  - Sparse Group Lasso
  - Overlapping Group Lasso
  - Tree Structured Group Lasso



https://turbosnu.wordpress.com/2016/01/20/note-down-feature-selection-for-adc/

Multilevel Models

Multilevel Models (MLM)

- "Multilevel modeling is a generalization of generalized linear modeling" (Gelman, 2005)
- Multilevel model also known as
  - Hierarchical model
  - Mixed effect model
  - …
- "Level" in multilevel refers to hierarchy of parameters

Why MLM?

- MLM is useful when:
  - Insufficient data for lowest level models while higher level models are too coarse
  - Desire to get similar results for individuals within a group
- MLM models entities at the lowest level but “borrows strength” from higher levels

Classical Regression vs. MLM

- When there is very little group-level variation, multilevel modeling reduces to classical regression with no group indicators

- When group-level coefficients vary greatly, multilevel modeling reduces to classical regression with group indicators (group dummy codes)

- Advantage occurs between these two cases

Example: Alcohol Abuse

- Study alcohol abuse among young people (Dominici, 2005)
- A person is a member of a family and a resident of a state
- Level 1 (person): person's ability to metabolize alcohol
- Level 2 (family): alcohol abuse in the family
- Level 3 (state): state laws

Example: Alcohol Abuse Data

- 3 years of longitudinall data
- 82 adolescents beginning at age 14
- Covariates:
  - COA: indicator variable whether adolescent is a child of alcoholic parent
  - PEER: 8-point scale that shoes the proportion of their friends who drink alcohol
  - Time: 0, 1, 2

Example: Proposed MLM

Alcohol use in individual i at time t

PEER of individual i at time t

COA of individual i at time t

$$y_{it} = \beta_{0i} + \beta_p p_{it} + \beta_c c_{it} + \beta_{1i} t + \epsilon_{it}$$

$$\beta_{0i} = \beta_{00} + b_{0i}$$

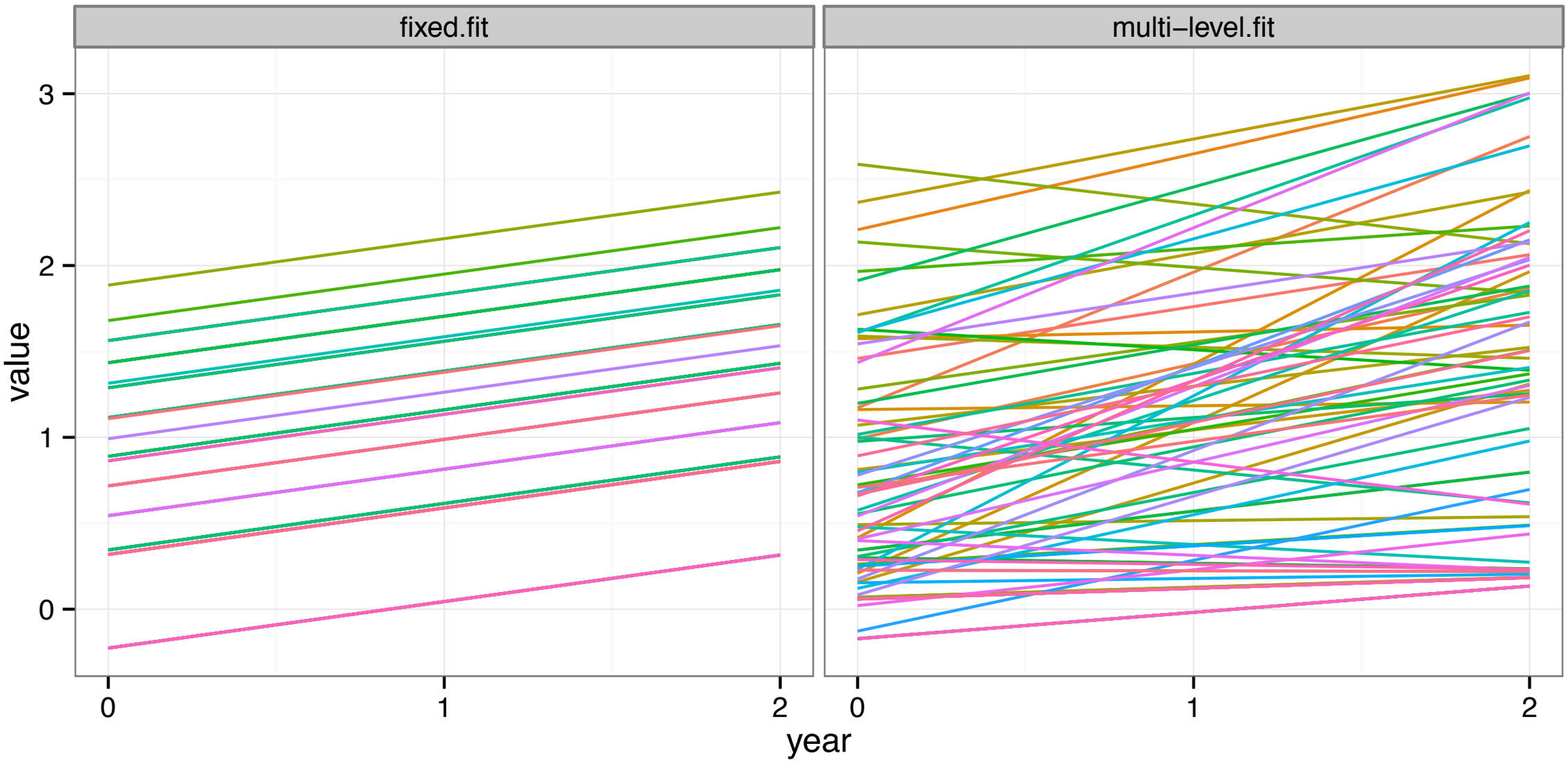
$$\beta_{1i} = \beta_{10} + b_{1i}$$

Noise effects

MLM Computational

- Maximum Likelihood Estimation
  - Finds it with respect to all the parameters through EM iterations
- Restricted Maximum Likelihood Estimation
  - Focuses only on D (random-effect covariance) through EM iterations (Harville, 1976)
- Bayes Posterior Estimation
  - Gibbs sampling (Gelman and Hill, 2007)

Example: Fitted Results



Example: Fitted Results (2)