Chapter 3:
Linear Methods for Regression
The Elements of Statistical Learning

Aaron Smalter

Chapter Outline

- 3.1 Introduction
- 3.2 Linear Regression Models and Least Squares
- 3.3 Multiple Regression from Simple Univariate Regression
- 3.4 Subset Selection and Coefficient Shrinkage
- 3.5 Computational Considerations

3.1 Introduction

- A linear regression model assumes the regression function,

$$E(Y|X)$$

is linear on the inputs

$$X1, ..., Xp$$

- Simple, precomputer model
- Can outperform nonlinear models when low # training cases, low signal-noise ration, sparse
- Can be applied to transformations of the input

3.2 Linear Regression Models and Least Squares

- Vector of inputs: *X = (X1, X2, ..., Xp)*
- Predict real-valued output: *Y*

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \tag{3.1}$$

3.2 Linear Regression Models and Least Squares (cont'd)

- Inputs derived from various sources
  - Quantitative inputs
  - Transformations of inputs (log, square, sqrt)
  - Basis expansions (X2 = X1^2, X3 = X1^3)
  - Numeric coding (map one multi-level input into many Xi's)
  - Interactions between variables (X3 = X1 * X2)

3.2 Linear Regression Models and Least Squares (cont'd)

- Least Squares
  - Pick a set of coefficients,

    B = (B0, B1, ..., Bp)^T

  - In order to minimize the **residual sum of squares** (RSS):

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2. \end{aligned} \tag{3.2}$$

3.2 Linear Regression Models and Least Squares (cont'd)

- How do we pick *B* so that we minimize the RSS?
- Let ***X*** be the *N x (p+1)* matrix
  - Rows are input vectors (1 in first position)
  - Cols are feature vectors
- Let **y** be the *N x 1* vector of outputs

3.2 Linear Regression Models and Least Squares (cont'd)

- Rewrite RSS as,

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta). \tag{3.3}$$

- Differentiate with respect to B,

$$\begin{aligned} \frac{\partial \mathrm{RSS}}{\partial \beta} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^T} &= -2\mathbf{X}^T\mathbf{X}. \end{aligned} \tag{3.4}$$

- Set first derivative to zero (assume **X** has full column rank, **X**^T**X** is positive definite),

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{3.5}$$

- Obtain unique solution:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.6}$$

- Fitted values are:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.7}$$

3.2.2 The Gauss-Markov Theorum

- Asserts: least squares estimates have the **smallest variance** among all linear unbiased estimates.
- Estimate any linear combination of the parameters,

$$\theta = a^T \beta;$$

3.2.2 The Gauss-Markov Theorum (cont'd)

- Least squares estimate is,

$$\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{3.17}$$

- If **X** is fixed, this is a linear function c0^Ty of response vector y
- If linear model is correct, a^TB is unbiased.
- Gauss-Markov Theorum:

The Gauss–Markov theorem states that if we have any other linear estimator $\tilde{\theta} = \mathbf{c}^T\mathbf{y}$ that is unbiased for $a^T\beta$, that is, $\mathrm{E}(\mathbf{c}^T\mathbf{y}) = a^T\beta$, then

$$\mathrm{Var}(a^T\hat{\beta}) \leq \mathrm{Var}(\mathbf{c}^T\mathbf{y}). \tag{3.19}$$

In praise of linear models!

- Despite its simplicity, the linear model has distinct advantages in terms of its *interpretability* and often shows good *predictive performance*.
- Hence we discuss in this lecture some ways in which the simple linear model can be improved, by replacing ordinary least squares fitting with some alternative fitting procedures.

Why consider alternatives to least squares?

- *Prediction Accuracy:* especially when $p > n$, to control the variance.
- *Model Interpretability:* By removing irrelevant features — that is, by setting the corresponding coefficient estimates to zero — we can obtain a model that is more easily interpreted. We will present some approaches for automatically performing *feature selection*.

3.4 Subset Selection and Coefficient Shrinkage

- As mentioned, unbiased estimators are not always the best for prediction.

Three classes of methods

- *Subset Selection*. We identify a subset of the p predictors that we believe to be related to the response. We then fit a model using least squares on the reduced set of variables.
- *Shrinkage*. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as *regularization*) has the effect of reducing variance and can also perform variable selection.
- *Dimension Reduction*. We project the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different *linear combinations*, or *projections*, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares.

3.4.1 Subset Selection

- Improve estimator performance by retaining only a subset of variables.
- Use least squares to estimate coefficients of remaining inputs.
- Several strategies:
  - *Best subset regression*
  - *Forward stepwise selection*
  - *Backwards stepwise selection*
  - *Hybrid stepwise selection*

3.4.1 Subset Selection (cont'd)

- *Best subset regression*
  - For each *k* in {0,1,2,...,p},
  - find subset of size *k* that gives smallest residual sum
- *Leaps and Bounds* procudure (Furnival and Wilson 1974)
- Feasible for *p* up to 30-40.
- Typically choose *k* such that estimate of expected prediction error is minimized.

3.4.1 Subset Selection (cont'd)

- *Forward stepwise selection*
  - Searching all subsets is time consuming
  - Instead, find a good path through them.
    - Start with intercept,
    - Sequentially add predictor that most improves the fit.
  - ”Improved fit” based on F-statistic, add predictor that gives largest value of F

$$F = \frac{\mathrm{RSS}(\hat{\beta}) - \mathrm{RSS}(\tilde{\beta})}{\mathrm{RSS}(\tilde{\beta})/(N-k-2)}. \tag{3.40}$$

  - Stop adding when no predictor gives a significantly greater F value.

3.4.1 Subset Selection (cont'd)

- *Backward stepwise selection*
  - Similar to previous procedure, but starts with the full model
  - Sequentially deletes predictors.
  - Drop predictor giving the smallest F value.
  - Stop when dropping any other predictor leads to significanty decrease in F value.

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p. *Why not?*
- Best subset selection may also suffer from statistical problems when p is large: larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data.
- Thus an enormous search space can lead to *overfitting* and high variance of the coefficient estimates.
- For both of these reasons, *stepwise* methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Regularization

- When the number of observations or training examples m is not large enough compared to the number of feature variables n, over-fitting may occur.
- Tends to occur when large weights are found in x.
- What can we do to prevent over-fitting?
- Use L2-regularization
- Regularization :
- Minimize : (Loss Function) + (regularization term)

3.4.3 Shrinkage Methods aka Regularization

- Subset selection is a discrete process (variables retained or not) so may give high variance.
- Shrinkage methods are similar, but use a continuous process to avoid high variance.
- Examine two methods:
  - *Ridge regression (L2 penalty)*
  - *Lasso (L1 penatly)*

L2-Regularization

* Regularization term : $\lambda\|x\|_2^2$
  * $\lambda > 0$ is the regularization parameter
* For LSP, this becomes
  * Minimize $\|Ax - y\|^2 + \|Fx\|_2^2$
    * Regularization term restricts large value components
    * Special case of Tikhonov regularization
    * Can be computed directly ($O(n^3)$)
    * Or can use iterative methods (*e.g.* conjugate gradients method)
* For LRP, this becomes
  * Minimize $l_{avg}(v, x) + \lambda\|x\|_2^2$
  * Smooth and convex, can be solved using gradient descent, steepest descent, Newton, quasi-Newton, truncated Newton, CG methods

3.4.3 Shrinkage Methods (cont'd)

- *Ridge regression*
  - Shrinks regression coefficients by imposing a penalty on their size.
  - Ridge coefficients minimize the penalized sum of squares,

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2, \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \le s, \tag{3.42}$$

    - Where *s* is a complexity parameter controlling the amount of shrinkage.
    - Mitigates high variance produced when many input variables are correlated.

3.4.3 Shrinkage Methods (cont'd)

- *Ridge regression*
  - Can be reparameterized using *centered* inputs, replacing each xij with $x_{ij} - \bar{x}_j$
  - Estimate,

    β_0 by $\bar{y} = \sum_1^N y_i/N$
  - Ridge regression solutions give by,

$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.44}$$

3.4.3 Shrinkage Methods (cont'd)

- *Ridge regression*
  - Assume that response varies most in direction of high variance of the inputs.
  - Shrink the coefficients of the low-variance components more than high-variance ones

L1-Regularization

- Regularization term : $\lambda \|x\|_1$
- LSP : $\|Ax - y\|^2 + \|Fx\|_2^2 + \lambda \|x\|_1$
- LRP : $l_{avg}(v, x) + \lambda \|x\|_1$
- The regularization term penalizes all factors equally
- This makes the *x* ****SPARSE****
  - A sparse *x* means reduced complexity
  - Can be viewed as a selection of relevant/important features
- Non-differentiable -> harder problem
  - Can transform into convex quadratic problem
    - minimize $\|Ax - y\|^2 + \|Fx\|_2^2 + \lambda \sum_{i=1}^{n} u_i$
    - subject to $-u_i \le x_i \le u_i, \quad i = 1, \ldots, n$
  - and use standard convex optimization methods to solve, but these usually cannot handle large practical problems

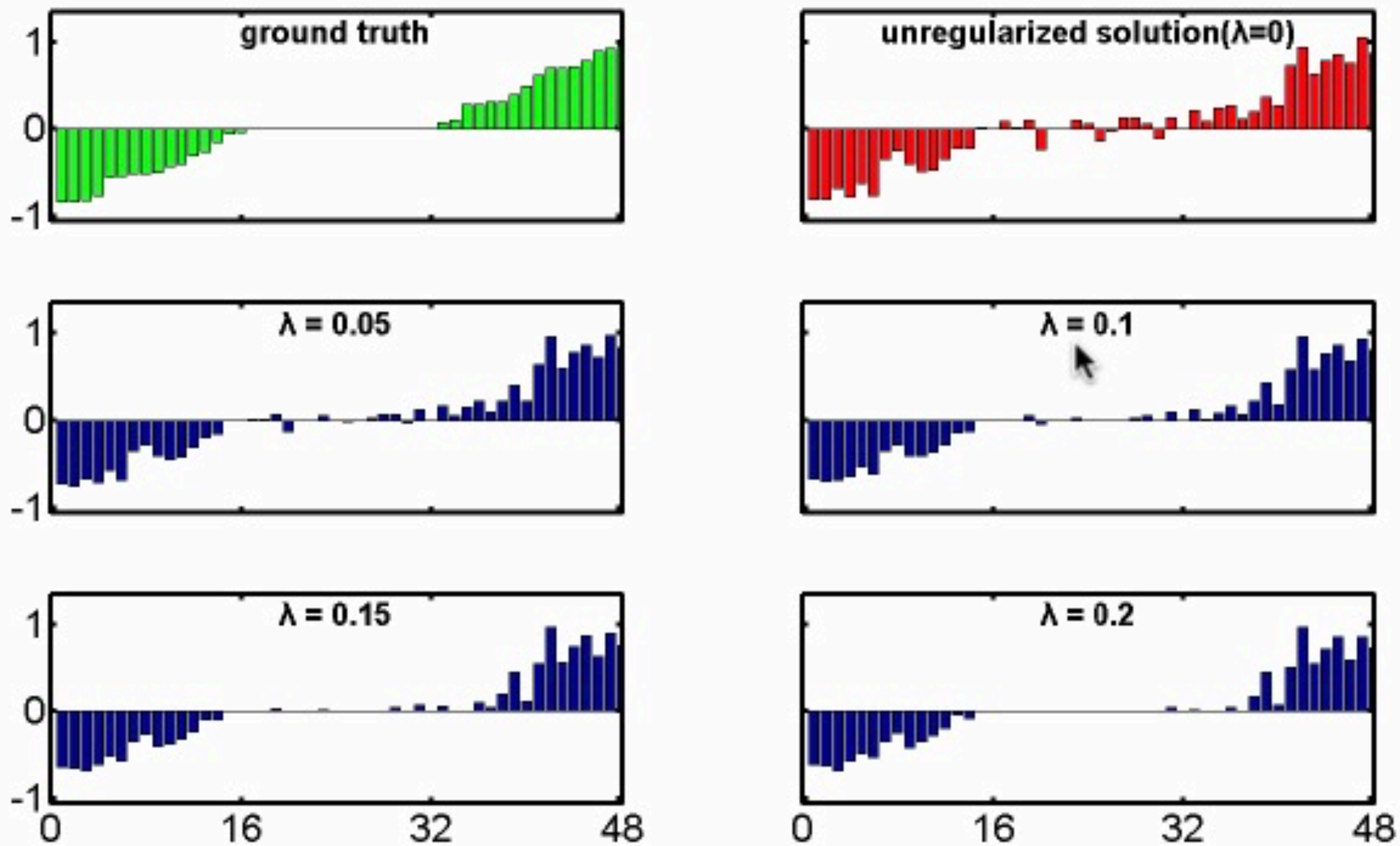
3.4.3 Shrinkage Methods (cont'd)

- *Lasso*
  - Shrinks coefficients, like ridge regression, but has some differences.
  - Estimate is defined by,

$$\begin{aligned}\hat{\beta}^{\text{lasso}} &= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \\ &\quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \le t. \end{aligned} \tag{3.51}$$

  - Agai intercept.
  - Ridge penalty replaced by lasso penalty (L1 penalty)

Effects of L1-Regularization



3.4.6 Multiple Outcome Shrinkage and Selection

- Least squares estimates in multi-output linear model are collection of individual estimates for each output.
- To apply selection and shrinkage we could simply apply a univariate technique
  - individually to each outcome
  - or simultaneously to all outcomes

Summary

- L2-Regression suppresses over-fitting
- L2-Regression does not add too much complexity to existing problems -> easy to calculate
- L1-Regression creates sparse answers, and better approximations in relevant cases
- L1-Regression problems are not differentiable -> need other ways of solving problem (using convex optimization techniques, iterative approaches, etc.)

L1 & L2 Penalty

- ElasticNet regression