Machine Learning

PATRICK HALL AND LISA SONG

DEPARTMENT OF DECISION SCIENCE

Contemporary Regression Models

- Linear Methods for Regression
- Logistic Methods for Regression
- Generalized Linear Model
- Regularization
  - Penalized Regression: Lasso, Ridge, Elastic Net
- Model Assessment and Selection
  - Model Bias-Variance Tradeoff
- Ensemble Models

Regression: Issues

Requirements	If broken …
Linear relationship between inputs and targets; normality of y and errors	Inappropriate application/unreliable results; use a machine learning technique; use GLM
N > p	Underspecified/unreliable results; use LASSO or Elastic Net penalized regression
No strong multicollinearity	Ill-conditioned/unstable/unreliable results; Use Ridge(L2/Tikhonov)/Elastic Net penalized regression
No influential outliers	Biased predictions, parameters, and statistical tests; use robust methods, i.e. IRLS, Huber loss, investigate/remove outliers
Constant variance/no heteroskedasticity	Lessened predictive accuracy, invalidates statistical tests; use GLM in some cases
Limited correlation between input rows (no autocorrelation)	Invalidates statistical tests; use time-series methods or machine learning technique

Linear Regression

Carl Friedrich Gauss
(1777–1855)

Linear Regression

A linear regression model assumes that the regression function $E(Y|X)$ is linear in the input variables $X_1, X_2, ..., X_p$. Linear models are:

- Simple and often provide an interpretable model
- Sometimes outperform nonlinear models with low signal-to-noise or sparse data
- Can be applied to transformations of the inputs - basis-function methods
- Many nonlinear models are direct generalizations of linear methods

Linear Regression

Let $X^T = (X_1, X_2, ..., X_p)$ be the input vectors for which we want to predict output Y. The linear regression model has the form:

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

where $\beta_j' s$ are unknown parameter coefficients and X_j are input vectors.

Linear Regression

Input vectors, X_j, can be:

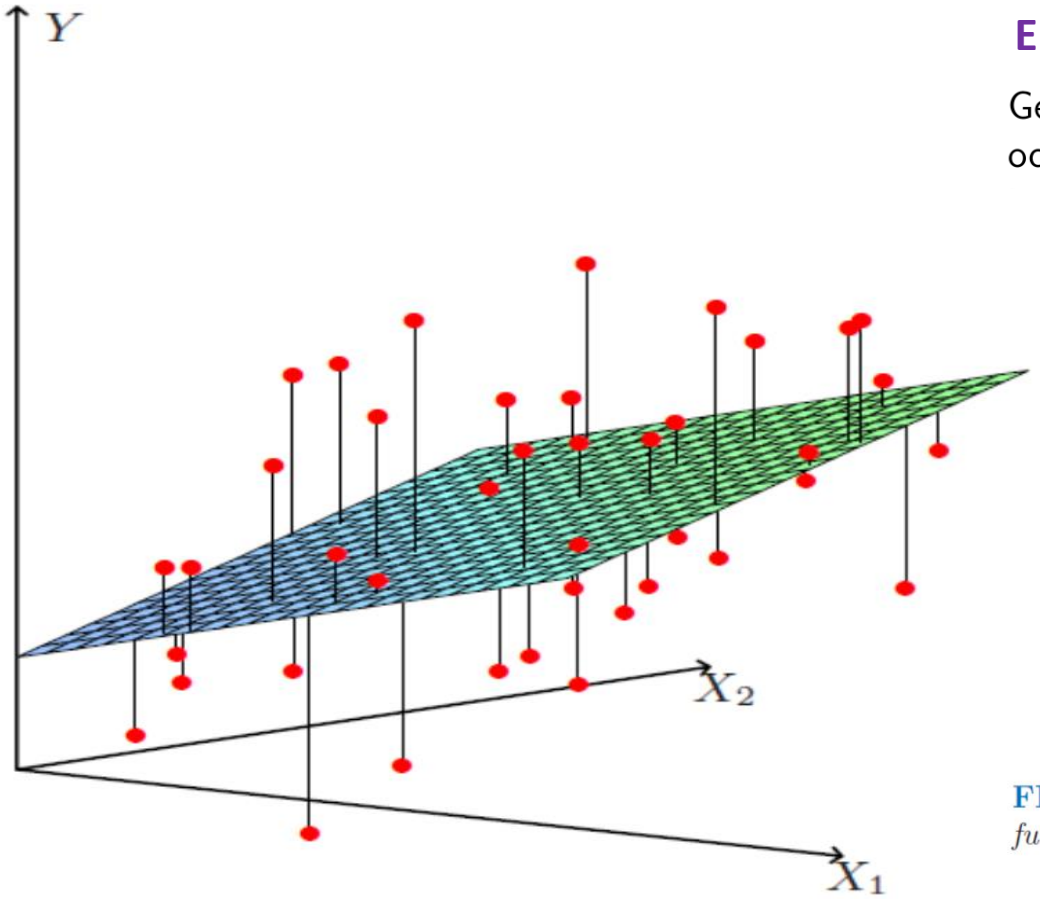
- Quantitative or transformations of quantitative inputs
- Basis expansion that leads to a polynomial representation
- Numeric or dummy coding: level-dependent constants
- Interactions between the input variables

Linear Regression

Regression methods estimates the model parameters $\beta's$ with the training data to minimize the residual sum of squares, $\boldsymbol{RSS(\beta)}$.

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

Regression: Least-Squared Method



Elements of Statistical Learning (pg.45)

Geometry of least-square fitting in the $\mathbb{R}^{p+1}$-dimensional space occupied by the pairs (X, Y).

FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

Linear Regression

Minimizing residual sum of squares, $RSS(\beta)$:

- Express $RSS(\beta) = (y - X\beta)^T(y - X\beta)$ in a quadratic form with $p + 1$ parameters
- Differentiate $RSS(\beta)$ with respect to β
- Assume $\mathbf{X}$ has full column rank and set the first derivative to zero, leads to unique solution:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

Linear Regression

The predicted values are given by:

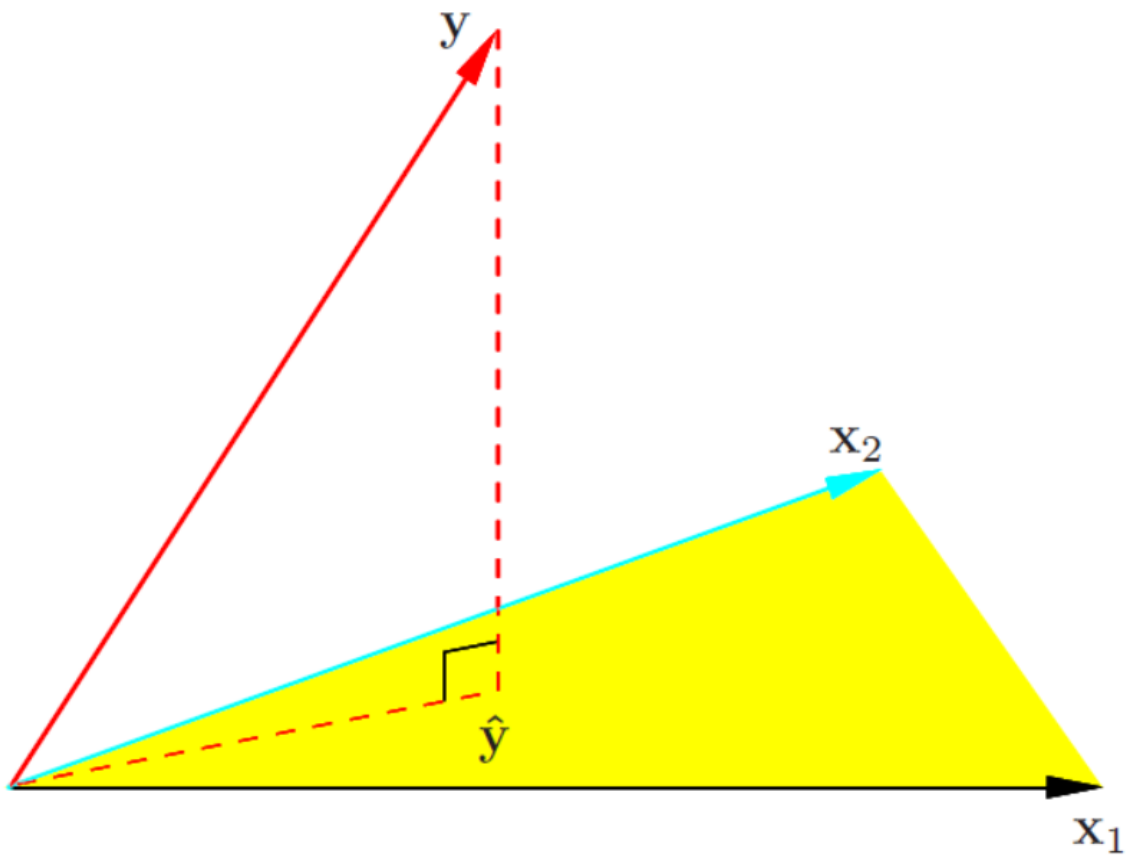
$$\hat{f}(x_0) = (1 : x_0)^T \hat{\beta}$$

and the fitted values applied to the inputs are:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

where $\hat{y}_i = \hat{f}(x_i)$ and $H = X(X^T X)^{-1} X^T$

Regression: Least-Squares Method



Elements of Statistical Learning (pg.45)

Geometrical representation of the least square estimate in $\mathbb{R}^N$. The column vectors of X are denoted as $x_0, x_1, ..., x_p$ with $x_0 = 1$ and these vectors span a subspace of $\mathbb{R}^N$ - i.e. column space of X. We minimize $RSS(\beta) = \|y - X\beta\|^2$ by choosing $\hat{\beta}$ so that the residual vector $y - \hat{y}$ is orthogonal to this subspace. The resulting estimate $\hat{y}$ is called the orthogonal projection of y onto the subspace and H computes the orthogonal projection (hence noted as projection matrix).

FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector* $\mathbf{y}$ *is orthogonally projected onto the hyperplane spanned by the input vectors* $\mathbf{x}_1$ *and* $\mathbf{x}_2$*. The projection* $\hat{\mathbf{y}}$ *represents the vector of the least squares predictions*

Multiple Linear Regression

Multiple Regression: Extension of $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$ with $p > 1$. Suppose that we have a univariate model with no intercept:

$$Y = X\beta + \epsilon$$

Then, the least square estimate and residuals are:

$$\hat{\beta} = \frac{\sum_1^N x_i y_i}{\sum_1^N x_i^2}$$

$$r_i = y_i - x_i \hat{\beta}$$

Multiple Linear Regression

In a vector notation, we can express the *inner product* between x and y as:

$$\hat{\beta} = \frac{< x, y >}{< x, x >}$$

$$r = y - x\hat{\beta}$$

This simple univariate regression provides the building block for multiple linear regression on observational data where orthogonality is almost never preserved.

Multiple Linear Regression

Now, suppose we have an intercept and a single input x where the least square coefficient of X has the form:

$$\hat{\beta} = \frac{< x - \bar{x}1, y >}{< x - \bar{x}1, x - \bar{x}1 >}$$

where $\bar{x} = \frac{\sum_i x_i}{N}$ and $1 = x_0$, the vector of N ones. That is, regress x on 1 to produce the residual $z = x - \bar{x}1$; and regress y on the residual z to give the coefficient $\hat{\beta}_1$

Multiple Linear Regression

This process is generalized to the case of p inputs:

$$\hat{\beta}_p = \frac{< z_p \cdot y >}{< z_p, z_p >}$$

In summary:

The multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of x_j on y, after x_j has been adjusted for

$x_0, x_1, \ldots, x_{j-1}, x_{j+1}, x_p$

Logistic Regression

Logistic Regression

The logistic regression models the posterior probabilities of **K** classes via a linear function of x, while also ensuring that they sum to one and remain in [0,1]. The model has the form:

$$\log \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} = \beta_{10} + \beta_1^T x$$
$$\log \frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} = \beta_{20} + \beta_2^T x \qquad (4)$$
$$\log \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} = \beta_{(K-1)0} + \beta_{K-1}^T x$$

Logistic Regression

The logistic model is defined in terms of $K-1$ log-odds or logit transformations. Also, to emphasize the dependence on the entire parameter set $\theta = \{b_{10}, B_1^t, ..., b_{(K-1)0}, B_{K-1}^T\}$ and $Pr(G = k|X = x) = p_k(x;\theta)$.

Note, for $K = 2$, the model is reduced to a single linear function with a binary response - widely used in a biostatistical application.

Logistic Regression

Fitting Logistic Regression: Maximum Likelihood

The multinomial distribution is appropriate for the conditional likelihood of G given X. The log-likelihood for N observations is:

$$l(\theta) = \sum_{i=1}^{N} \log p_{gi}(x_i; \theta) \tag{5}$$

where $p_k(x_i; \theta) = Pr(G = k | X = x_i; \theta)$

Logistic Regression

For $K = 2$, an algorithm for finding the maximum log-likelihood - IRLS, *iteratively reweighted least squares*

- Score equations (first derivative of the log-likelihood)
- Hessian matrix (second derivative matrix)
- Adjusted response

Logistic Regression

Summary: Logistic Regression Model

- Used mostly as a data analysis and inference tool
- Understand the role of the input variables in explaining the outcome
- Models are typically fit in a search for a parsimonious model involving a subset of the variables - possibly with some interaction terms

Generalized Linear Model

Generalized Linear Model

Generalized Linear Model:

In GLM, the target y_i is assumed to follow an exponential family of distributions with mean μ_i where μ_i is defined to be some function of $x_i^T \beta$. Here, μ_i is often a nonlinear function of the covariates but considered to be linear because the covariates affect the distribution of y_i only through the linear combination of $x_i^T \beta$.

Generalized Linear Model

GLM Assumptions

- Data Y_i are independently distributed but response variable does not need to be normally distributed
- GLM does not assume linear relationship between response and input, however, does assume linear relationship between the transformed response in terms of the link function and the input variables

Generalized Linear Model

- Input variables can assume nonlinear transformation of the original independent variables
- Homogeneity of the variances does NOT need to be satisfied
- Errors need to be independent but do NOT need to be normally distributed
- Maximum likelihood estimation, thus, relies on large-sample approximation
- Goodness-of-fit measures rely on sufficiently large samples

Generalized Linear Model

Three Components

- Random Component - probability distribution of the response, also called a noise or error model
- Systematic Component – specifies the input variables, specifically the linear predictors of the model
- Link Function - specifies the link between random and systematic component, i.e., how the expected value of the response relates to the linear predictor of the input variables

Anatomy of a GLM: Link Function

Family/distribution defines mean and variance of **Y**

$$E(\boldsymbol{Y}) = \mu = g^{-1}(\boldsymbol{X}\beta)$$

Linear component

Nonlinear link function between linear component and E(**Y**)

$$Var(\boldsymbol{Y}) = V(\mu) = V(g^{-1}(\boldsymbol{X}\beta))$$

Family/Distribution allows for non-constant variance

Family of Distributions

- **Gaussian** distribution, squared error loss, sensitive to outliers
- **Laplace** distribution, absolute error loss, more robust to outliers
- **Huber** loss, hybrid of squared error & absolute error, robust to outliers
- **Poisson** distribution (e.g., number of claims in a time period)
- **Gamma** distribution (e.g., size of insurance claims)
- **Tweedie** distribution (compound Poisson-Gamma)
- **Binomial** distribution, log-loss for binary classification

Subset Selection

Subset Selection

Inadequacy of Least Square Estimates:

- Prediction accuracy - low bias but large variance
- Interpretation - high dimensionality

With subsets selection - retain only a subset of variables and discard the rest. By doing so, accuracy can be improved (bias-variance trade-off) and/or determine smaller subset that exhibits the strongest effects (sacrifice small details). Then least square regression is used to estimate the coefficients of the inputs that are retained.

Subset Selection

Types of Subset Selection

- Best Subset Selection
- Forward- and Backward-Stepwise Selection
- Forward-Stagewise Selection

Subset Selection

Best Subset Regression

- For each $k \in \{0, 1, ..., p\}$ the subset size k that gives the smallest residual sum of squares
- Choosing k - bias-variance tradeoff
- Criteria - typically choose the smallest model that minimizes the expected prediction error estimates
  - Cross-validation to estimate the prediction error and select k
- Alternative method - AIC Criterion

Subset Selection

Forward-Stepwise

- Start with the intercept
- Sequentially add the predictor that most improves the fit via *greedy algorithm*, producing nested sequence of models
  - Computational: Even when $p >> N$
  - Statistical: Lower variance but perhaps more bias

Subset Selection

Backward-Stepwise

- Start with the full model
- Sequentially discard the predictor model that has the least impact (contribution) on the fit
- Criterion: Z-score
- Only used when $N > p$

Subset Selection

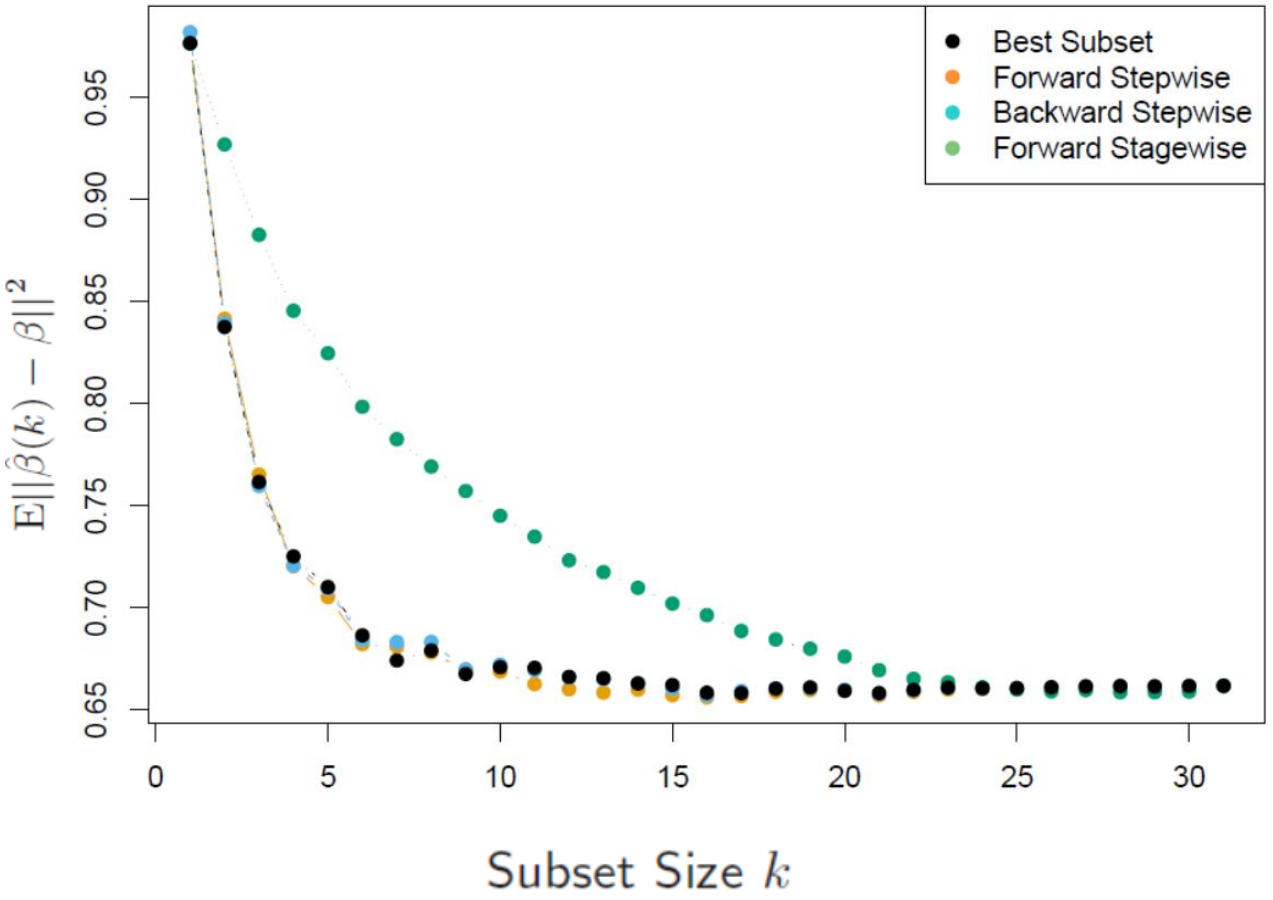
Forward-Stagewise

- Start with an intercept equal to $\bar{y}$ and coefficients of centered predictors are all initially set to zero
- At each step, variable that is most correlated with the current residual is selected.
- Computes the simple linear regression coefficient of the residuals on this chosen variable and adds it to the current coefficient for that variable

Subset Selection

- Continue until none of the variables have correlation with the residuals
- Note, unlike forward-stepwise, none of the other variables are adjusted when a term is added to the model

Subset Selection



Elements of Statistical Learning (pg.59)

Comparison of best-subset with simpler alternative forward- and backward-selection.

Note, forward-stagewise regression takes longer to reach the minimum error.
In this case, it takes over 1000 steps to get all the correlation below 10^{-4}.

FIGURE 3.6. *Comparison of four subset-selection techniques on a simulated linear regression problem* $Y = X^T\beta + \varepsilon$. *There are* $N = 300$ *observations on* $p = 31$ *standard Gaussian variables, with pairwise correlations all equal to* 0.85. *For* 10 *of the variables, the coefficients are drawn at random from a* $N(0, 0.4)$ *distribution; the rest are zero. The noise* $\varepsilon \sim N(0, 6.25)$, *resulting in a signal-to-noise ratio of* 0.64. *Results are averaged over* 50 *simulations. Shown is the mean-squared error of the estimated coefficient* $\hat{\beta}(k)$ *at each step from the true* β.

Penalized Regression

Model Selection

Issues with Subset Selection Strategies:

- Discrete process - often exhibits high variance without reducing the prediction error of the full model

Shrinkage:

- Continuous
- Does not suffer from high variability

Penalized Regression:

- Ridge Regression
- Lasso Regression
- Elastic Net

Ridge Regression: L2 Penalty

- Impose penalty on the size of the coefficients
- Ridge coefficients minimize a penalized residual sum of squares

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta j^2 \right\} \qquad (6)$$

Here, $\lambda \geq 0$ is the complexity parameter that controls the amount of the shrinkage - larger the value of λ, the greater the amount of shrinkage.

Ridge Regression: L2 Penalty

An equivalent of equation (6) is:

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j)^2$$
$$subject\ to \quad \sum_{j=1}^{p} = \beta_j^2 \leq t \tag{7}$$

which makes explicit the size constraint on the parameters.

Ridge Regression: L2 Penalty

- Intercept β_0 is left out of the penalty term
- If many variables are correlated, then the coefficients can become poorly determined and exhibit high variance
- Imposing a size constraint on the coefficients can alleviate this problem
- Not equivalent under scaling of inputs - normally standardize the inputs before

Lasso Regression: L1 Penalty

The Lasso estimation is defined by:

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j)^2$$
$$subject\ to \quad \sum_{j=1}^{p} = |\beta_j| \leq t \tag{8}$$

Note the similarity to the ridge regression. Here, L2 ridge penalty $\sum_1^p \beta_j^2$ is replaced by L1 lasso penalty $\sum_1^p |\beta_j|$.

Lasso Regression: L1 Penalty

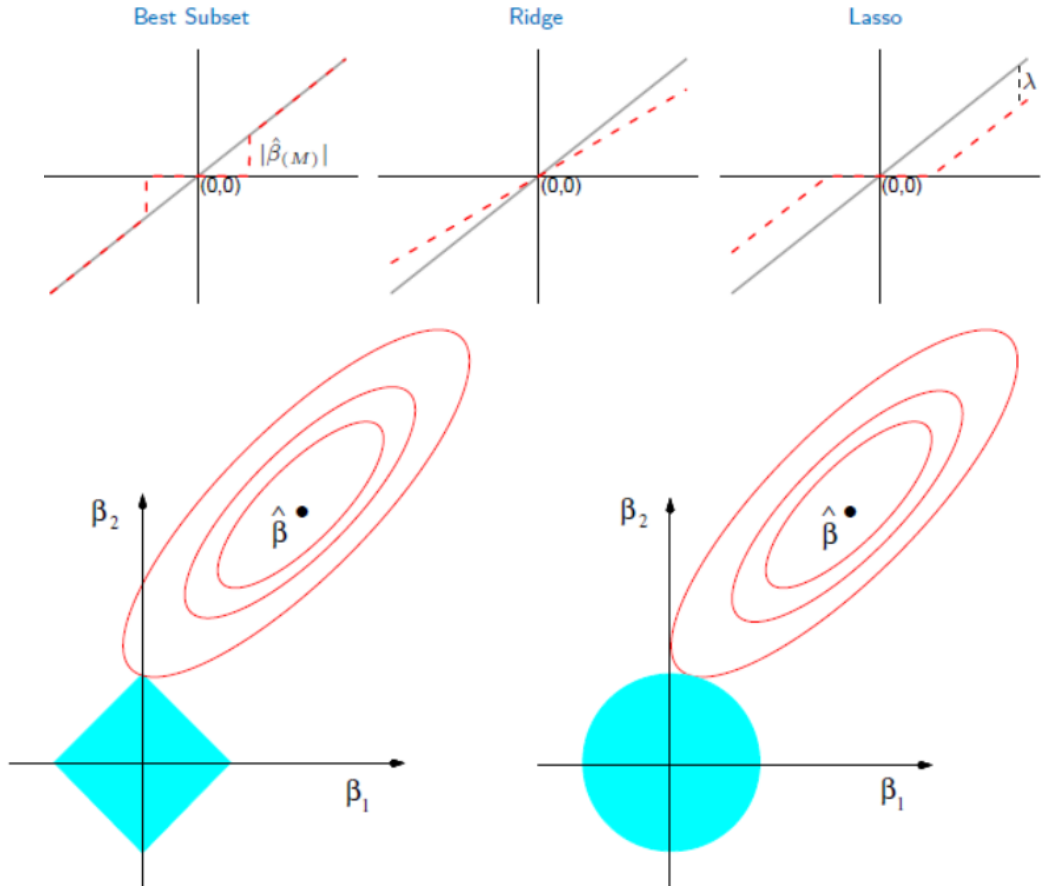
- Due to the nature of the lasso constraint, making it sufficiently small will force some of the coefficients to be exactly zero. Thus, the lasso does a kind of a ‘continuous‘ subset selection.
- It should be adaptively chosen to minimize an estimated expected prediction error

Subset, Ridge, and Lasso

Summary:

- Ridge: proportional shrinkage
- Lasso: variables are selected by their model parameter truncating at zero - ‘soft thresholding‘

Subset, Ridge, and Lasso



Elements of Statistical Learning (pg.71)

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Elastic Net - Modern Approach

Hui Zou and Trevor Hastie
Regularization and variable selection via the elastic net,
Journal of the Royal Statistical Society, 2005

Elastic Net

We can generalize ridge and lasso and view them as Bayes estimates for $q \geq 0$

$$\tilde{\beta} = \underset{\beta}{argmin}\left\{ \sum_{i=1}^{N}(y_i - b_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |b_j|^q \right\} \qquad (9)$$

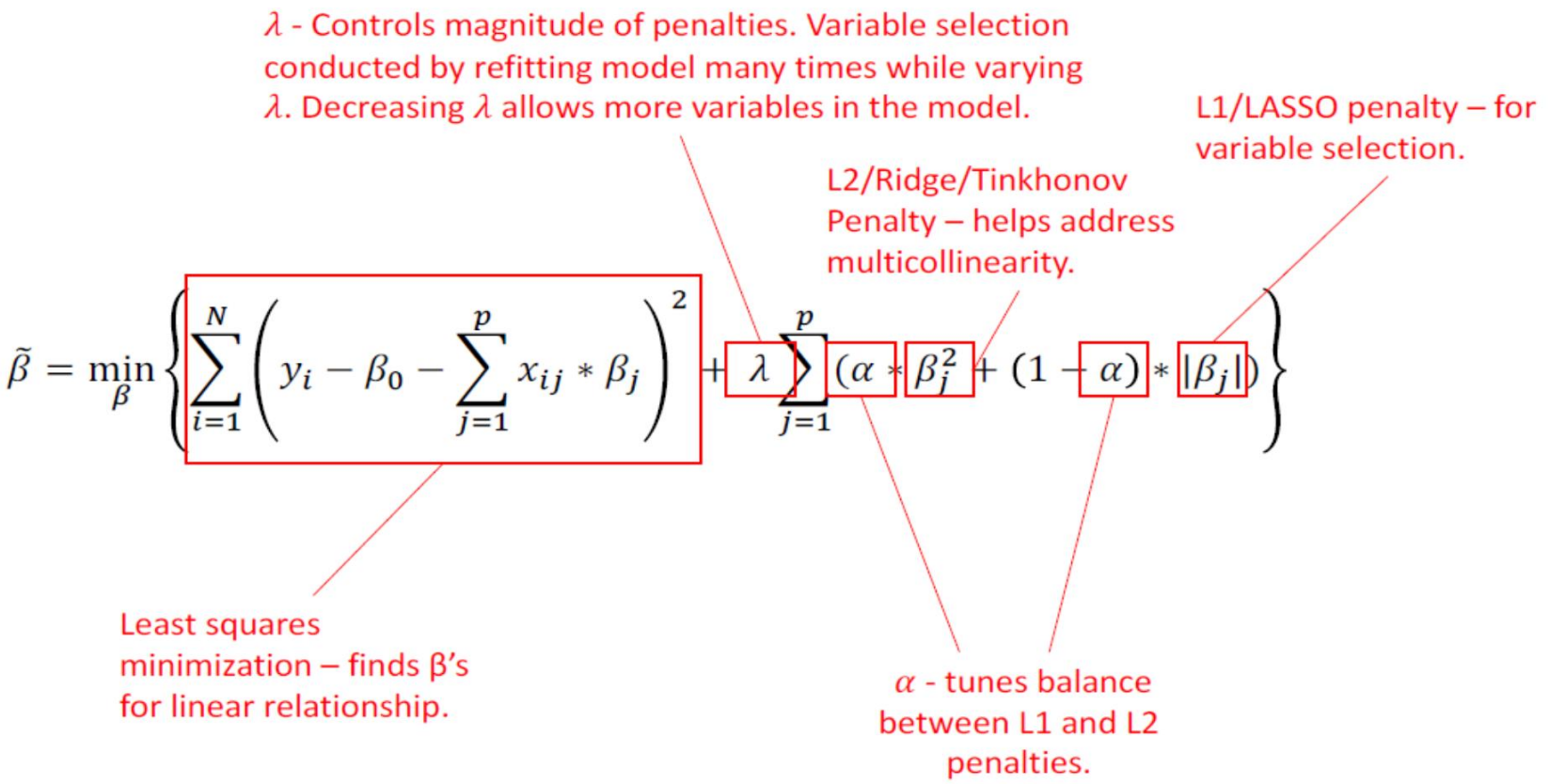
The value $q = 0$ corresponds to subset selection; $q = 1$ corresponds to lasso; and $q = 2$ corresponds to ridge.

Elastic Net

For $q \in (1, 2)$ suggest a compromise between the lasso and ridge, but does not share the ability of lasso for setting coefficients exactly to zero. Partly for this reason and for computational tractability, Zou and Hastie introduced the **elastic net penalty**:

$$\lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1 - \alpha)|\beta_j|) \tag{10}$$

Anatomy of Elastic Net: L1 & L2 Penalty



Elastic Net with Iteratively Reweighted Least Squares

Iteratively Reweighted Least Square complements fitting methods in the presence of the outliers by:

- Initially giving all observations equal weight then…
  - Train the model to estimate the β's and find a linear relationship/equation

$$\tilde{\beta} = \min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} * \beta_j \right)^2 \right\}$$

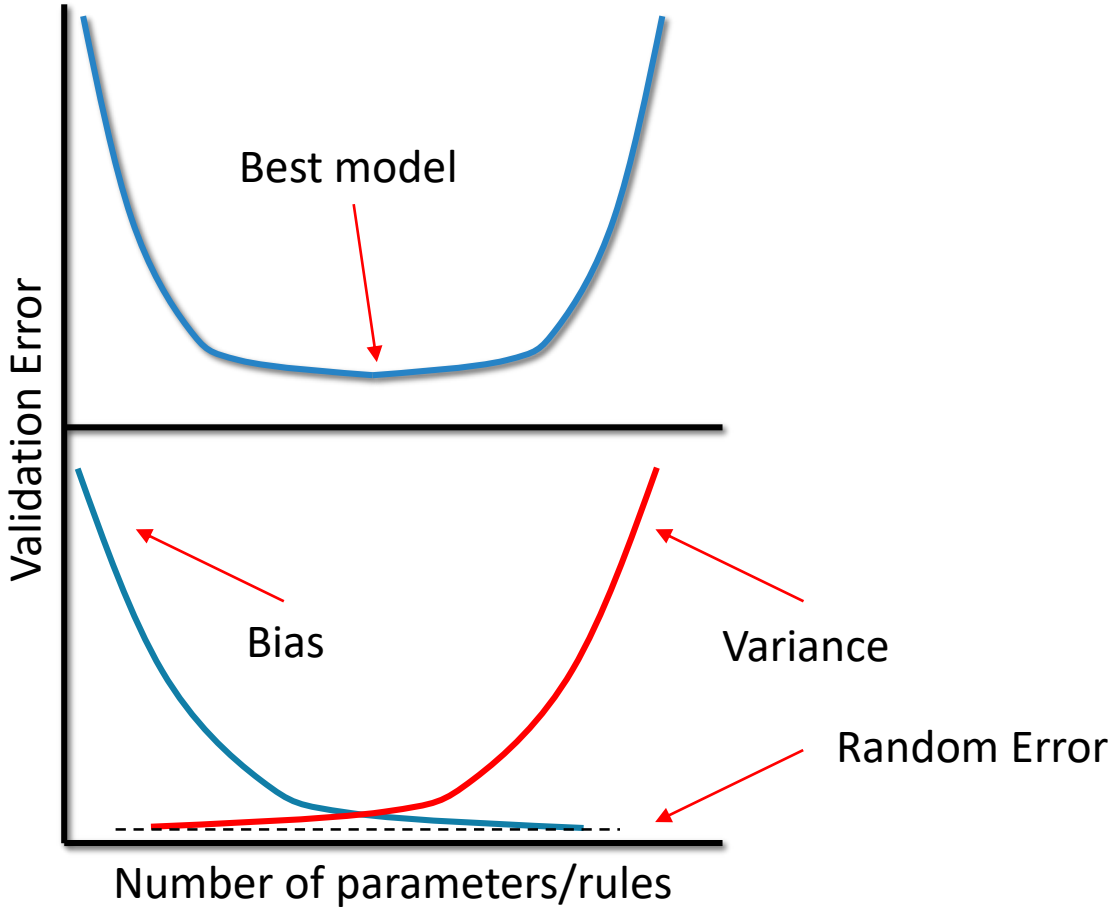
"Inner Loop"

  - Calculate the residuals given these β's/ linear equation
  - Re-weight observations that cause high residuals to have a lower impact in the train model
  - Re-train to find new β's/linear equation
  - Continue calculating residuals, re-weighting observations, and re-training until β's become stable and weighted residuals are small…

"Outer Loop"

Cross-Validation & Parameter Tuning

The Bias / Variance Trade-off

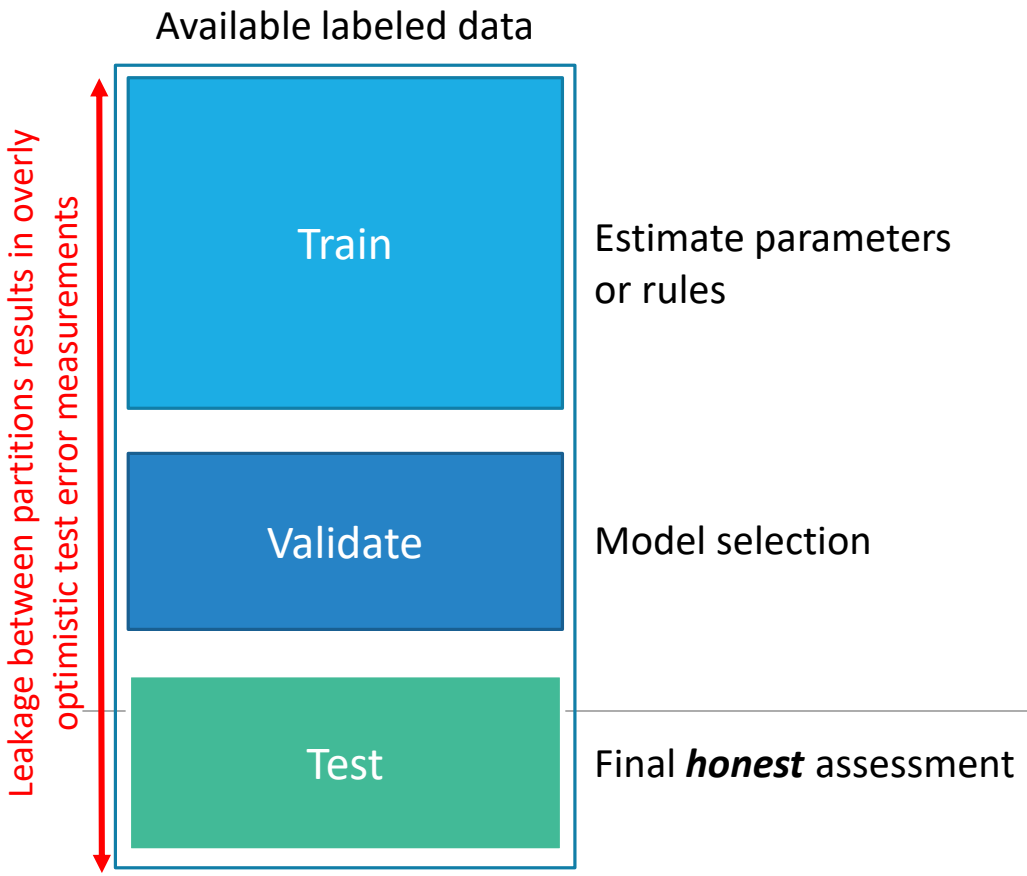


Total Error = Bias + Variance + Random Error = $(\hat{f}(x) - f(x))^2$

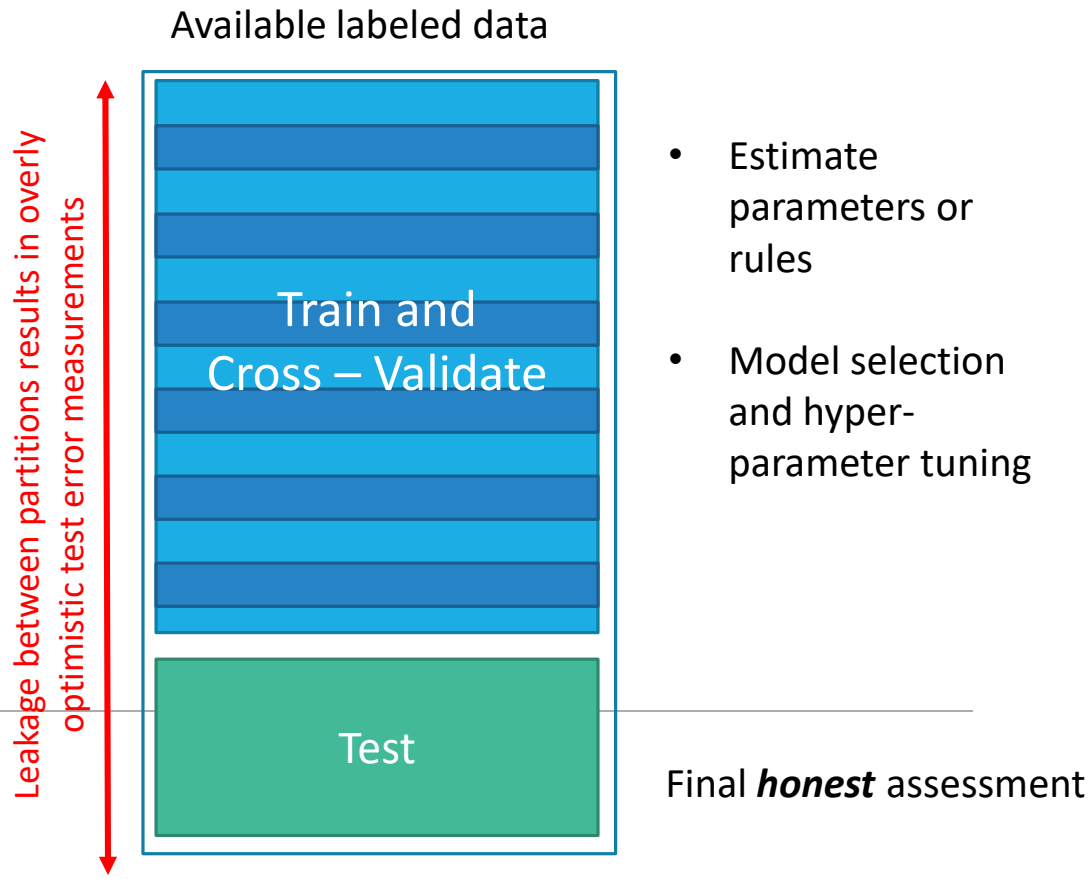
Bias = $E[\hat{f}(x)] - f(x)$ or the error that arises from a model's inability to replicate the fundamental phenomena represented by a data set.

Variance = $(\hat{f}(x) - E[\hat{f}(x)])^2$ or the error that arises from a model's ability to produce differing predictions from the values in a new data set.

Bias/Variance Trade-off in Practice: Honest assessment



Best suited for big data or linear models using traditional forward, backward, or stepwise selection.



- Estimate parameters or rules
- Model selection and hyper-parameter tuning

Nearly always a more generalizable approach, but computationally intensive. Best suited for complex models with many hyper-parameters and small to medium sized data.

Ensemble Models

Ensemble Models

- **Ensemble models** – methods for combining the posterior probability or predictions of two or more predictive models to create a potentially more accurate or stable model

- Ensemble Methods
  - **Simple averaging**
  - **Top-*t* ensemble selection**
  - **Hill climbing ensemble selection**
  - **Weighted averaging**
  - **Stacking**

Ensemble Models

- **Simple Averaging** (simple soft voting) – takes the average of the posterior probability for each response level across the models and then classifies the observation based on the level that has the maximum average probability
- **Top-*t* Ensemble Selection** – take the top *t* models out the M that are generated when the models are ranked by an accuracy measure and uses validation data to determine the best value for *t*
  - Similar to weighted averaging , but, equal weights are assigned to a subset of the available models
- **Hill-climbing Ensemble Selection** – improvement in the accuracy of adding any given model, i.e. model that most improves the misclassification rate in the validation set. The final ensemble is selected based on the misclassification rate in the test set
  - Similar to weighted averaging, but here, different weights are assigned to each model depending on how many times a particular model is included in the ensemble

Ensemble Models

- **Weighted Averaging** – weighted average of the posterior probability for each response level is considered with a model-specific weight applied
- **Stacking** – uses posterior probability from the various models that are used as inputs and the original response (target) variable that is used as the response
  - Can use linear regression model to generate the weights for a weighted averaging ensemble
  - Can also implement penalized linear regression – LASSO, Ridge, and Elastic Net
  - Other models can also be implemented – decision tree and random forest