1.5em

Linear Methods for Regression

Peida Wu

Institute of Mathematics and Science

2025 年 4 月 2 日

Introduction

1. Simple and often provide an adequate and interpretable description of output.
2. Sometimes outperform fancier nonlinear models, especially in situations with small numbers of training cases, low signal-to-noise ratio or sparse data.
3. Can be applied to transformations of the inputs and this considerably expands their scope.

Basic knowledge

Input:

$$X^T = (X_1, \ldots, X_P)$$

The linear regression form:

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

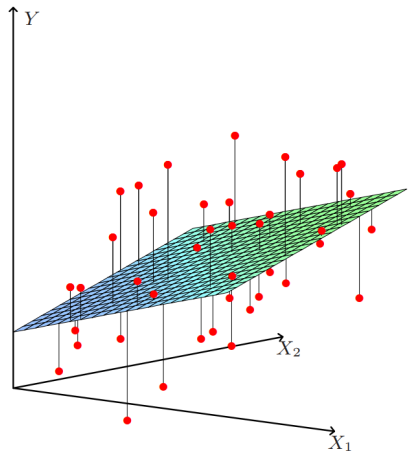
where β_j are unknown parameters. X_j can be:

1. quantitative inputs.
2. transformations of quantitative inputs, such as log, square-root or square.
3. basis expansions, such as $X_2 = X_1^2, X_3 = X_1^3$, which is a polynomial representation.
4. interactions between variables, such as $X_3 = X_1 \cdot X_2$

Least Square

Training data:$(x_1, y_1), \cdots, (x_N, y_N)$,where $x_i = (x_{i1}, \ldots, x_{ip})^T$.
Goal: Minimize the residual sum of squares:

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



Calculation of Least Square

Rewrite in matrix form:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) \tag{3}$$

$$\begin{aligned} \frac{\partial \text{RSS}}{\partial \beta} &= -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} &= 2\mathbf{X}^T\mathbf{X} \end{aligned} \tag{4}$$

Assume $\mathbf{X}$ is full column rank, $\mathbf{X}^T\mathbf{X}$ is positive definite.

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \end{aligned} \tag{5}$$

Geometry view

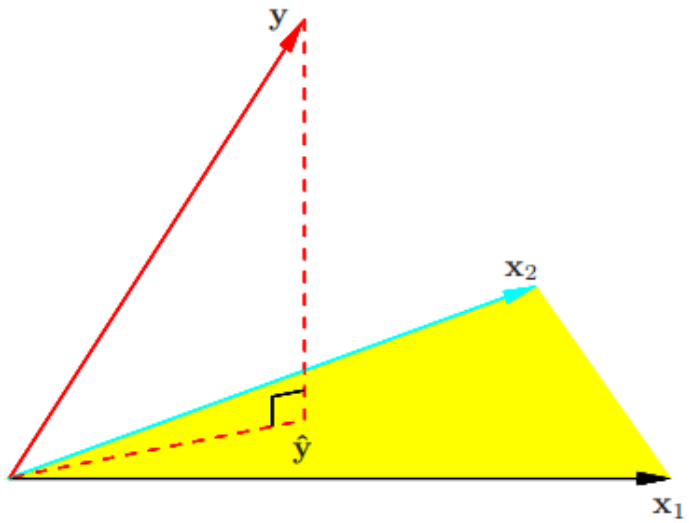


图 2: geometry

$$H = X(X^T X)^{-1} X^T$$

referred as a projection matrix.

Variance of $\hat{\beta}$

Assumption:

1. y_i are not correlated and have constant variance $\hat{\sigma}^2$
2. x_i are fixed

From (5), we get

$$\begin{aligned} \text{Var}(\hat{\beta}) =& (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 \\ \hat{\sigma}^2 =& \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \end{aligned} \tag{6}$$

where $E(\hat{\sigma}^2) = \sigma^2$ is unbiased.

Inference of $\hat{\beta}$

Additional assumption:

1. The linear model is correct for the mean
2. The deviations of Y around $E(Y)$ is Gaussian and additive, that is, $\epsilon \sim N(0, \sigma^2)$

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2) \tag{7}$$

$$(N - p - 1)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{N-p-1} \tag{8}$$

Significance test for groups of coefficients

Goal: To check if some variables can be excluded from model

$$F = \frac{(RSS_0 - RSS_1)/(p_1 - p_0)}{RSS_1/(N - p_1 - 1)} \tag{9}$$

1. RSS_1 residual sum of squares for bigger model
2. RSS_0 residual sum of squares for smaller model

The Gauss-Markov Theorem

The Gauss-Markov Theorem: In linear regression model if the variance satisfies that their mean is zero($E(\epsilon) = 0$), variance uncorrelated and same($var(\epsilon) = \sigma^2 I_n$), then the regression coefficients are the best linear unbiased estimator(BLUE).

Sketch of proof

1. unbiased estimator of β.
2. calculate Var($\hat{\beta}$)
3. prove the best linear unbiased estimator by contradiction

Multiple Output

Linear model for the output is:

$$\begin{aligned} Y_k =& \beta_{0k} + \sum_{j=1}^{p} X_j \beta_{jk} + \epsilon_k \\ =& f_k(\mathbf{X}) + \epsilon_k \end{aligned} \tag{10}$$

Matrix form **Y=XB+E**

$$\begin{aligned} \text{RSS}(B) =& \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \\ =& tr[(\mathbf{(Y\text{-}XB)}^T \mathbf{(Y\text{-}XB)})] \end{aligned} \tag{11}$$

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{XY} \tag{12}$$

Suppose $\text{Cov}(\epsilon) = \Sigma$,

$$\text{RSS}(\mathbf{B}, \Sigma) = \sum^{N} (y_i - f(x_i))^T \Sigma^{-1} (y_i - f(x_i)) \tag{13}$$

Drawback of least square

Drawback of least squares estimates:

1. prediction accuracy.
2. interpretation.

Best-Subset Selection

Subset selection: This approach involves identifying 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.
How to choose k:

1. cross-validation to estimate prediction error
2. AIC criterion

residue criteria

$$C_p = \frac{1}{n}\left(\mathrm{RSS} + 2p\hat{\sigma}^2\right)$$

$$\mathrm{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2p\hat{\sigma}^2\right)$$

$$\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\hat{\sigma}^2\right)$$

$$\text{Adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{TSS}/(n-1)}$$

Best-Subset Selection

Algorithm 1 Best subset selection

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots, p$:
   1. Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   2. Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Forward-Stepwise Selection

Forward stepwise selection starts with the intercept, and then sequentially adds into the model the predictor that most improves the fit.
Forward stepwise selection is a greedy algorithm, which may be sub-optimal compared to best- subset selection.

1. computational($p >> N$)
2. statistical(low variance but perhaps more bias)

Forward-Stepwise Selection

Algorithm 2 Forward stepwise selection

1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   1. Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   2. Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Backward stepwise selection

Backward-stepwise selection starts with the full model, and sequentially deletes the predictor that has the least impact on the fit. The candidate for dropping is the variable with the smallest prediction error.

Backward selection can only be used when $N > p$, while forward stepwise can always be used

Backward stepwise selection

Algorithm 3 Backward stepwise selection

1. Let $\mathcal{M}_0$ denote the full model, which contains p predictors.
2. For $k = p, \ldots, 1$:
   1. Consider all k models that augment the predictors in $\mathcal{M}_k$ for a total of $k-1$ predictors.
   2. Choose the best among these k models, and call it $\mathcal{M}_{k-1}$. Here best is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Shinkage methods

Drawback of subset slection: Since it's a discrete process, no variables retianed, leading to high variance.
Benefit of shinkage method: Continuous and doesn't suffer high variability.

Ridge regression

Imposing a penalty on their size.

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} \beta_j^2\} \quad (14)$$

The larger the value of λ, the coeffecients are shrunk toward zero. An equavalent way:

$$\begin{aligned} \hat{\beta}^{ridge} =& \text{argmin}_{\beta} \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j)^2 \\ & \text{subject to} \sum_{i=1}^{p} \beta_j^2 \leq t \end{aligned} \quad (15)$$

Ridge regression

Matrix form:

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta \tag{16}$$

The solution in matrix form is:

$$\hat{\beta}^{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{17}$$

1. The solution is a linear function of $\mathbf{y}$
2. Make the problem nonsingular

Another insight of ridge regression

Input $X \in \mathbf{R}^{N \times p}$, $X = UDV^T$.
Least square case:

$$\begin{aligned} X\hat{\beta}^{ls} &= X(X^TX)^{-1}X^Ty \\ &= UU^Ty \end{aligned} \tag{18}$$

Ridge regression case:

$$\begin{aligned} X\hat{\beta}^{ridge} &= X(X^TX + \lambda I)^{-1}X^Ty \\ &= UD(D^2 + \lambda I)^{-1}DU^Ty \\ &= \sum_{j=1}^{p} u_j \frac{d_j^2}{d_j^2 + \lambda} u_j^Ty \end{aligned} \tag{19}$$

The lasso

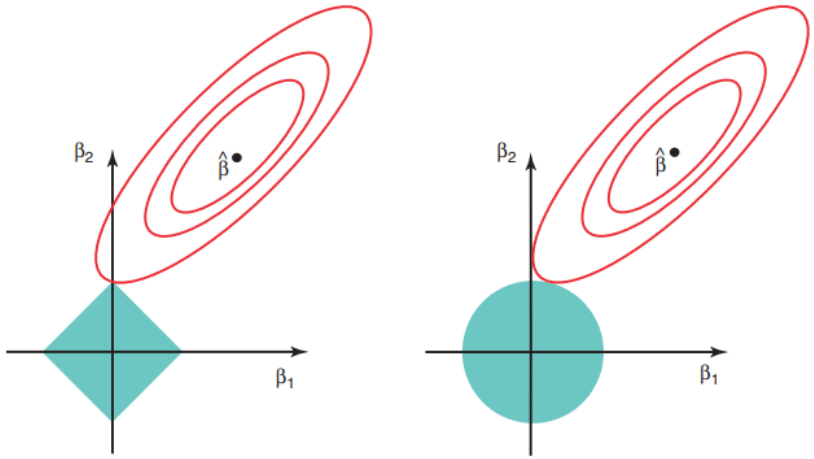
The lasso regression:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{20}$$

The equivalent case:

$$\begin{aligned} \hat{\beta}^{lasso} =& \text{argmin}_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ & \text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t \end{aligned} \tag{21}$$

The lasso



Simplified cases:

$$\textit{minimize} \sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{22}$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2}, & y_j > \frac{\lambda}{2} \\ y_j + \frac{\lambda}{2}, & y_j < \frac{\lambda}{2} \\ 0, & |y_j| \leq \frac{\lambda}{2} \end{cases} \tag{23}$$

Bayesian view

Suppose $\beta = (\beta_0, \ldots, \beta_p)^T$ has prior distribution $p(\beta)$.

$$p(\beta|X, Y) \propto f(Y|X, \beta)p(\beta|X) = F(Y|X, \beta)p(\beta) \qquad (24)$$

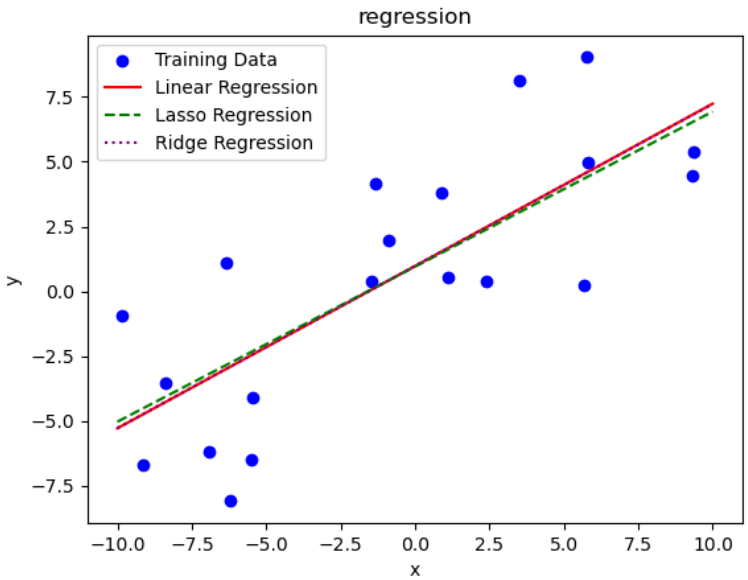
Suppose the linear model errors are independent and are normal distributed.
Assume $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$, g is density function.

1. g is gaussian distribution with 0 mean and λ standard deviation. The posterior mode for β is given by the ridge regression.
2. g is a double-exponential distribution with mean zero and λ parameter, the posterior mode for β is lasso solution.

Numerical experiment

100 samples: 20 train, 80 test.

$$y = 0.5x + 1 + \epsilon, \epsilon \sim N(0, 3)$$



Numerical experiment

Linear Regression Coefficient: $y = 0.6257x + 0.9761$
Lasso Regression Coefficient: $y = 0.5980x + 0.9516$
Ridge Regression Coefficient: $y = 0.6248x + 0.9753$

表 1: Comparison of Train and Test MSE

Model	Train MSE	Test MSE
Linear Regression	9.6471	14.2372
Lasso Regression	9.6748	13.7189
Ridge Regression	9.6471	14.2200