Chapter 3: Linear Methods for Regression

Junrui Di

Contents

1. Introduction

Linear regression assumes that the regression function $E(Y|X)$ is linear in the inputs $X_1, \ldots, X_p$.

2. Linear regression models and least square

[1.] *Linear regression from a least square point of view* (minimal assumption about the distribution)

- Form: $f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j$
- Data: $\{x_i, y_i\}$ $i = 1...N$, each $x_i = (x_{i1}...x_{ip})^T$ is a feature vector, with parameters $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$
- Least square: To minimize $\text{RSS}(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 = \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ or $\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ in matrix form. **LSE makes no assumptions about the validity of the model form**
- LSE: $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- Fitted value: $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}$ is the projector of $\mathbf{y}$ onto the subspace spanned by column space of $\mathbf{X}$.
- Inference on parameters (assuming y_i's are uncorrelated and gave constant variance σ^2, and x_i are fixed)
  - $Var(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$
  - $\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$

[2.] *Linear regression with Gaussian error*

- Model Assumption: $Y = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$

- Distributional properties of model parameters
  - $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$
  - $(N-p-1)\hat{\sigma}^2 \sim \sigma^2\chi^2_{N-p-1}$
  - $\hat{\beta}$ and $\hat{\sigma}^2$ are statistically independent.

- Inference on single parameter β_j

Under $H_o : \beta_j = 0$, $z_j = \frac{\hat{\beta}_k}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{N-p-1}$, and β_j has a $1-2\alpha$ confidence interval of $(\hat{\beta}_j - z^{1-\alpha}\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}, \hat{\beta}_j + z^{1-\alpha}\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}})$

- Nested Model Comparison (test whether the added variables are necessary to the model)

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)} \sim F_{p_1-p_0, N-p_1-1}, \quad \text{where } \mathrm{RSS}_1 \text{ is for the larger model}$$

2.1 The Gauss-Markow Theorem

Least square estimates of β have the smallest variance among all linear unbiased estimates.

The least square estimator to estimate parameters $\theta = \alpha^T\beta$ is $\hat{\theta} = \alpha^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. It is an unbiased estimator, i.e. $E(\alpha^T\hat{\beta}) = \alpha^T\beta$. Gauss-Markow theorem states that $Var(\alpha\hat{\beta})$ has the smallest variance for any unbiased estimator.

We may want to trade a little bias for larger reduction in variance.

2.2 Regression by succesive orthogonolization

Algorithm 3.1 *Regression by Successive Orthogonalization.*

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
2. For $j = 1, 2, \ldots, p$

   Regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_{j-1}$ to produce coefficients $\hat{\gamma}_{\ell j} = \langle \mathbf{z}_\ell, \mathbf{x}_j \rangle / \langle \mathbf{z}_\ell, \mathbf{z}_\ell \rangle$, $\ell = 0, \ldots, j-1$ and residual vector $\mathbf{z}_j = \mathbf{x}_j - \sum_{k=0}^{j-1} \hat{\gamma}_{kj}\mathbf{z}_k$.
3. Regress $\mathbf{y}$ on the residual $\mathbf{z}_p$ to give the estimate $\hat{\beta}_p$.

Figure 1: Gram-Schmidt procedure for multiple regression

2.3 Multiple outcomes

Data: $Y_1 ... Y_K$, with the model $Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j\beta_{jk} + \epsilon_k$, with the matrix form $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$, where $\mathbf{Y}$ is $N \times K$, $\mathbf{X}$ is $N \times p + 1$, and $\mathbf{B}$ us $(p+1) \times K$.

$\mathrm{RSS}(\mathbf{B}) = \sum_k \sum_i (y_{ik} - f_k(x_i))^2 = \mathrm{tr}(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})$ is the RSs with LSE $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$

3. Subset selection

a. Best subset selection

b. Forward and backward selection

4. Shrinkage methods

4.1 Ridge regression

- RSS

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

or in the matrix form

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

with the solution

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

Even if $\mathbf{X}^T\mathbf{X}$ is not of full rank, $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is still nonsingular.

- Degree of freedom $\text{df}(\lambda) = \text{tr}[\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$
- Ridge solutions are not equivariant under scaling of the inputs, and one normally standardizes the inputs before solving for estimation.

4.2 Lasso

- RSS

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

* Shrinkage $s = t/\sum_j |\hat{\beta}_j|$ where $\hat{\beta}_j$ is the least square estimation.

4.3 Subset selection,ridge, and lasso

[1.] *Orthonormal input matrix* $\mathbf{X}$

- Ridge: proportional shrinkage
- LASSO: translate by a constant factor and truncating at zero, i.e soft thresholding
- Best subject: drops all the variables with coefficient smaller than the Mth largest, i.e. hard thresholding

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\mathrm{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

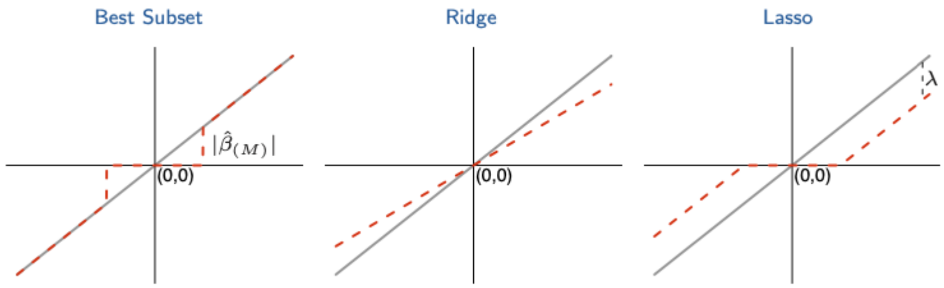

Figure 2: Gram-Schmidt procedure for multiple regression

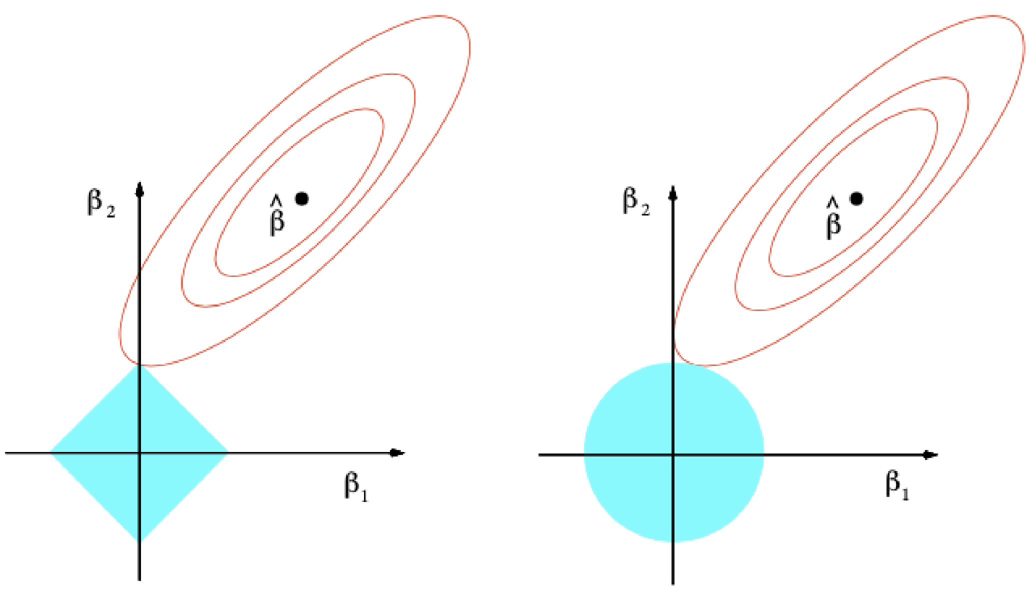

Figure 3: Gram-Schmidt procedure for multiple regression

[2.] *Nonorthogonal case*

Elastic net

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

4.4 Least angle regression

5. Methods using derived input directions

5.1 Principal components regression

PC regression forms the derived input columns $\mathbf{z}_m = \mathbf{X} v_m$ and then regresses $\mathbf{y}$ on $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_M$. Since they are orthogonal, each parameter is simply $\hat{\theta}_n = \frac{<\mathbf{z}_m, \mathbf{y}>}{<\mathbf{z}_m, \mathbf{z}_m>}$. It can be converted back to $\hat{\beta}_M^{pcr} = \sum_{m=1}^{M} \hat{\theta}_m v_m$.

The mth principal component direction v_m solveS:

$$\begin{aligned}
&\max_\alpha \mathrm{Var}(\mathbf{X}\alpha) \\
&\text{subject to } ||\alpha| = 1, \alpha^T \mathbf{S} v_l = 0, \quad l = 1, ...n-1
\end{aligned}$$

where $\mathbf{S}$ is the sample covariance

5.2 Partial least square

Algorithm 3.3 *Partial Least Squares.*

1. Standardize each $\mathbf{x}_j$ to have mean zero and variance one. Set $\hat{\mathbf{y}}^{(0)} = \bar{y}\mathbf{1}$, and $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, \ldots, p$.
2. For $m = 1, 2, \ldots, p$
   (a) $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} \mathbf{x}_j^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle \mathbf{x}_j^{(m-1)}, \mathbf{y} \rangle$.
   (b) $\hat{\theta}_m = \langle \mathbf{z}_m, \mathbf{y} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle$.
   (c) $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.
   (d) Orthogonalize each $\mathbf{x}_j^{(m-1)}$ with respect to $\mathbf{z}_m$: $\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - [\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)} \rangle / \langle \mathbf{z}_m, \mathbf{z}_m \rangle] \mathbf{z}_m$, $j = 1, 2, \ldots, p$.
3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_1^p$. Since the $\{\mathbf{z}_\ell\}_1^m$ are linear in the original $\mathbf{x}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\mathrm{pls}}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Figure 4: Gram-Schmidt procedure for multiple regression

The mth PLS direction $\hat{\psi}_m$ solves:

$$\begin{aligned}&\max_{\alpha}\mathrm{Corr}^2(\mathbf{y},\mathbf{X}\alpha)\mathrm{Var}(\mathbf{X}\alpha)\\&\text{subject to } ||\alpha| = 1, \alpha^T\mathbf{S}\hat{\psi}_l = 0, \quad l = 1, ...n-1\end{aligned}$$