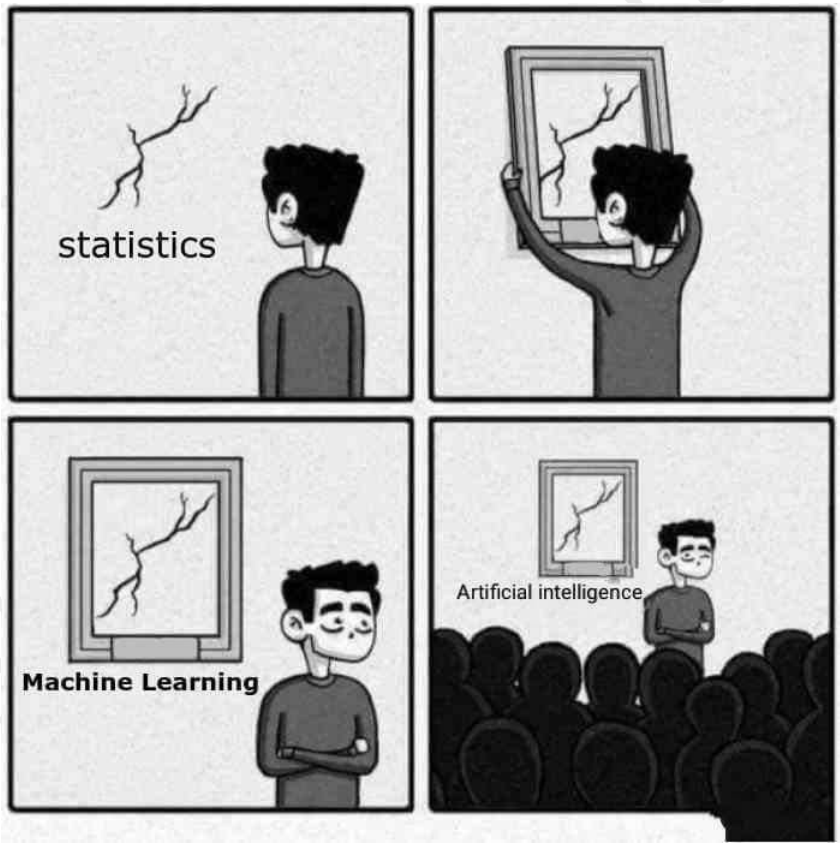
Fantuan's Academia

Fantuan's Math Notes series

Notes on Statistical Learning

Author: Jingxuan Xu



July 13, 2024

Fantuan's Math Notes

Fantuan's Math Notes

Contents

All the Sections with * are hard sections and can be skipped without losing coherence.

This note is referenced on **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** by Trevor Hastie, Robert Tibshirani and Jerome Friedman (The 'big three' in Stanford University!)[3].

Fanwan's Math Notes

Fantuan's Math Notes

Chapter 1

Linear Methods for Regression

1.1 Linear Regression Models

With input vector $X^T = (X_1, X_2, \cdots, X_p)$, we want to predict an output Y using the linear model

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{1.1}$$

1.1.1 Least Squares

Without making any assumptions about the validity of model 1.1, we can use *least squares* to simply find the best linear fit to the data. Pick the coefficients $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ to minimize the *residual sum of squares*

$$\text{RSS}(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{1.2}$$

where $(x_i, y_i), i = 1, \cdots, N$ with $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ are training data. To minimize this, we just simply find the minimum distance of sum of squared residuals from the fitted subplane, as shown in Figure 1.1.1 below.

Fortunately, this minimization has a closed form solution. Denote the input $N \times (p+1)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

and $\mathbf{y} = (y_1, y_2, \cdots, y_N)^T$ the training data output, we have the following solution.

Fantuan's Math Notes

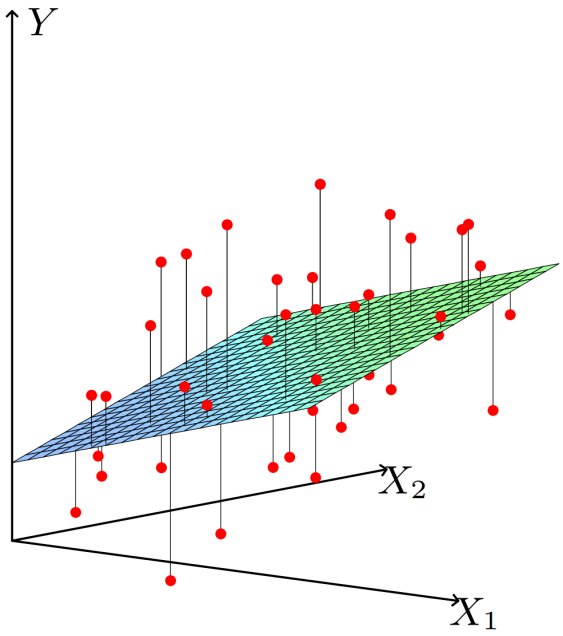


Figure 1.1: Linear least square fitting with $X \in \mathbb{R}^2$

Theorem 1.1.1: Solution to Least Square

If $\mathbf{X}$ is of full rank, the set of coefficients $\hat{\beta}$ that minimizes the residual sum of squares has the form

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

Proof. We can write the residual sum of squares in matrix form

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \tag{1.3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. To minimize this, we let first and second derivatives satisfy

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X} > \mathbf{0}$$

Since $\mathbf{X}$ has full rank, we have the *Hessian matrix* $2\mathbf{X}^T\mathbf{X}$ is positive definite. Therefore, the second derivative condition is naturally meet. For the first derivative, we further expand the equation

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\beta = 0 \quad \Longrightarrow \quad \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$$

Since $\mathbf{X}^T\mathbf{X}$ is positive definite, it is invertible. Therefore, multiply $(\mathbf{X}^T\mathbf{X})^{-1}$ on both sides, we have our

Fantuan's Math Notes

desired solution

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

□

- The predicted values for an input vector x_0 are then $\hat{f}(x_0) = (1 : x_0)^T\hat{\beta}$.
- The fitted values at training inputs are $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the *hat matrix* since it puts a hat on $\mathbf{y}$.

Geometric Interpretation of Least Square Solution:
We minimize RSS by choosing $\hat{\beta}$ so that the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the column subspace of $\mathbf{X}$. This can be seen from the first derivative condition

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

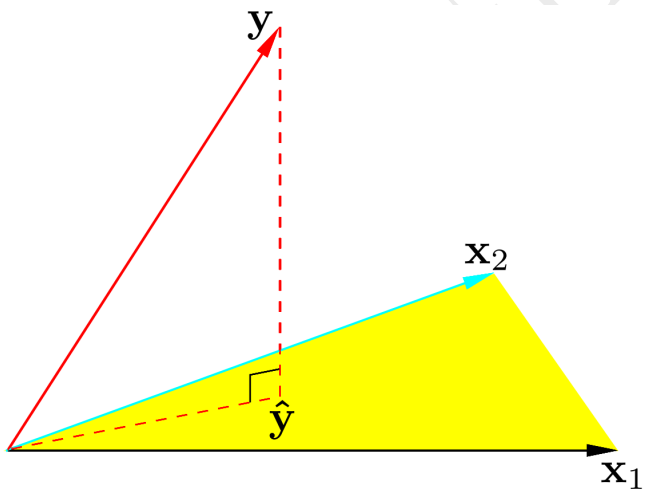


Figure 1.2: Orthogonality of column space of $\mathbf{X}$ and residuals

1.1.2 Inference on Parameters

Up to now we have made no assumptions about the data set. To make inference about the coefficients β, we need the *HEIL Gauss assumptions:*

- **Homogeneity:** y_i has constant variation σ^2.
- **Existance:** $y|x$ is a univariate random variable having a certain probability distribution with finite mean and variance.
- **Independence:** Observations y_i are independent with x_i are fixed.

- **Linear:** The conditional expectation $E(Y|X)$ of Y is linear in $X_1, X_2, \cdots, X_p$.
- **Gaussian:** Deviations of Y around its expectation are additive and Gaussian, i.e.,

$$Y = E(Y|X) + \epsilon = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

With these assumptions, we can pin down the sampling properties of $\hat{\beta}$.

Theorem 1.1.2: Sampling Distribution of $\hat{\beta}$

$$\hat{\beta} \sim N\left(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\right)$$

Proof. We first derive that $\hat{\beta}$ is unbiased. We have

$$E(\hat{\beta}) = E\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta$$

For the variance,

$$\begin{aligned}
\text{var}(\hat{\beta}) &= \text{Var}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \text{Var}(\mathbf{y}) \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\left((\mathbf{X}^T\mathbf{X})^{-1}\right)^T \text{Var}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\left((\mathbf{X}^T\mathbf{X})^T\right)^{-1} \text{Var}(\mathbf{y}) \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \text{Var}(\mathbf{y}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2
\end{aligned}$$

as desired, where we used the fact that $(A^T)^{-1} = (A^{-1})^T$. □

Typically one does not know the true value of σ^2. To make inference on parameters we need to estimate this. The *mean squared error (MSE)* is a typical unbiased estimation of this value.

Proposition 1.1.3: Unbiasedness of Sampling Variance

The mean squared error

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

is unbiased. i.e., $E(\hat{\sigma}^2) = \sigma^2$.

Proof. First note that $\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2$, and we have

$$\begin{aligned}
\mathbf{y} - \mathbf{X}\hat{\beta} &= \mathbf{X}\beta + \epsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}\beta + \epsilon - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) \\
&= \mathbf{X}\beta + \epsilon - \mathbf{X}\beta - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon = \epsilon(\mathbf{I}_N - \mathbf{H})
\end{aligned} \tag{1.4}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix. Before moving on, we first prove two facts about the hat matrix,

$$\mathbf{H}^T = \mathbf{H} \quad \text{and} \quad \mathbf{H}^2 = \mathbf{H}$$

For the first one, we have

$$\mathbf{H}^T = \left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T = (\mathbf{X}^T)^T\left((\mathbf{X}^T\mathbf{X})^{-1}\right)^T\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

For the second one,

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

We call the hat matrix *symmetric and idempotent*. With these two properties hold, along with Equation 1.4, we have

$$\begin{aligned} E\left(\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2\right) &= E\left(\epsilon^T(\mathbf{I}_N - \mathbf{H})^T(\mathbf{I}_N - \mathbf{H})\epsilon\right) = E\left(\epsilon^T(\mathbf{I}_N^T\mathbf{I}_N - \mathbf{I}_N^T\mathbf{H} - \mathbf{H}^T\mathbf{I}_N + \mathbf{H}^T\mathbf{H})\epsilon\right) \\ &= E\left(\epsilon^T(\mathbf{I}_N - \mathbf{H} - \mathbf{H}^T + \mathbf{H}^T\mathbf{H})\epsilon\right) = E\left(\epsilon^T(\mathbf{I}_N - \mathbf{H} - \mathbf{H} + \mathbf{H}^2)\epsilon\right) \\ &= E\left(\epsilon^T(\mathbf{I}_N - \mathbf{H})\epsilon\right) \end{aligned}$$

Notice that

$$\epsilon^T(\mathbf{I}_N - \mathbf{H})\epsilon = \sum_{i,j}\epsilon_i\epsilon_j(\delta_{ij} - H_{ij})$$

where δ_{ij} is the Kronecker delta. Then, since all input are fixed, by linearity of expectation,

$$E\left(\epsilon^T(\mathbf{I}_N - \mathbf{H})\epsilon\right) = \sum_{i,j}(\delta_{ij} - H_{ij})E(\epsilon_i\epsilon_j)$$

Since we assume that observations y_i are independent, we have ϵ_i are independent with expectation 0 and variance σ^2. Therefore,

$$E(\epsilon_i\epsilon_j) = \begin{Bmatrix} E(\epsilon_i)E(\epsilon_j) = 0, & \text{if } i \neq j \\ E(\epsilon_i^2) = \mathrm{Var}(\epsilon_i) + E(\epsilon_i)^2 = \sigma^2 + 0 = \sigma^2, & \text{if } i = j \end{Bmatrix} = \delta_{ij}\sigma^2$$

This finally gives us

$$E\left(\|\mathbf{y} - \mathbf{X}\hat{\beta}\|_2^2\right) = \sigma^2\sum_{i,j}(\delta_{ij} - H_{ij})\delta_{ij} = \sigma^2(N - \mathrm{tr}(\mathbf{H}))$$

Recall that trace has property $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, therefore,

$$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) = \mathrm{tr}\left(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) = \mathrm{tr}(\mathbf{I}_{p+1}) = p + 1$$

Fanyuan's Math Notes

This leads to

$$E\left(\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2^2\right)=\sigma^2(N-p-1)$$

which shows that the mean squared error is unbiased. □

With this unbiased estimation, we are ready to find the pivotal quantity related to β for inference. Our final goal is to derive the Z-score for β, which is the estimation divided by its standard deviation. To do this, we need some lemmas.

Lemma 1.1.4: Distribution of MSE

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

Proof. Here we derive a much more general result:

If $\mathbf{A}$ *is symmetric and idempotent* $n\times n$ *real matrix and* $\mathbf{Z}\sim N(0,\mathbf{I}_n)$ *is a random vector of* n *independent standard normal variables, then* $\mathbf{Z}^T\mathbf{A}\mathbf{Z}\sim\chi^2_r$ *where* r *is the trace of* A.

To prove this, we first examine the eigenvalues of an idempotent matrix. Note that if $\mathbf{A}$ is idempotent, and suppose λ is an eigenvalue of $\mathbf{A}$, and $\mathbf{v}$ is the corresponding eigenvector, we have

$$\lambda\mathbf{v}=\mathbf{A}\mathbf{v}=\mathbf{A}^2\mathbf{v}=\lambda\mathbf{A}\mathbf{v}=\lambda^2\mathbf{v}$$

since eigenvector $\mathbf{v}\neq 0$, we have $\lambda=\lambda^2$, which indicates that the eigenvector of idempotent matrix could only be 1 or 0. Also, this indicates that the multiplicity of unit eigenvalues equals the rank r of $\mathbf{A}$ (since all nonzero eigenvalues are 1).
By Real Spectral Theorem, a self-adjoint matrix $\mathbf{A}$ (for real matrix, this is just symmetric matrix) has an orthonormal basis consisting eigenvectors of $\mathbf{A}$. Therefore, using eigendecomposition of $\mathbf{A}$, we have

$$\mathbf{A}\mathbf{Q}=\mathbf{Q}\mathbf{\Lambda}\quad\Longrightarrow\quad\mathbf{A}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$$

where $\mathbf{\Lambda}$ is the diagonal matrix with eigenvalues of $\mathbf{A}$ on its diagonal, and $\mathbf{Q}$ is the $n\times n$ matrix with eigenvector corresponding to ith eigenvalue is on the ith column of $\mathbf{Q}$. The inverse equals the transpose since $\mathbf{Q}$ is orthonormal. We can delete from $\mathbf{\Lambda}$ those zero diagonal entries, leaving an $r\times r$ identity matrix. Correspondingly, we delete from $\mathbf{Q}$ those eigenvectors corresponding to null eigenvalues, leaving an $n\times r$ matrix $\mathbf{Q}_{n\times r}$. In this way, $\mathbf{A}$ can be rewritten as

$$\mathbf{A}=\mathbf{Q}_{n\times r}\mathbf{Q}_{n\times r}^T$$

Pantuan's Math Notes

Consider $\mathbf{N} = \mathbf{Q}_{n\times r}^T\mathbf{Z}$. Then, $\mathbf{N}$ is a random vector of r variables having multivariate normal distribution with mean vector 0 and covariance matrix $\mathbf{Q}_{n\times r}^T\mathbf{I}_n\mathbf{Q}_{n\times r}$. However, notice that

$$\mathbf{Q}_{n\times r}^T\mathbf{I}_n\mathbf{Q}_{n\times r} = \mathbf{Q}_{n\times r}^T\mathbf{Q}_{n\times r} = \mathbf{I}_r$$

since it is the multiplication of several orthonormal columns. In multivariate normal distribution, uncorrelated means independent. Therefore, $\mathbf{N}$ is just a multivariate normal random vector consisting of independent normal marginals, this follows that

$$\mathbf{Z}^T\mathbf{\Lambda}\mathbf{Z} = \mathbf{Z}^T\mathbf{Q}_{n\times r}\mathbf{Q}_{n\times r}^T\mathbf{Z} = \mathbf{N}^T\mathbf{N}$$

is a sum of square of r i.i.d. standard normal variables, so it has χ_r^2 distribution.

Now let's come back to our original question. In our case,

$$\frac{(N-p-1)\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2 = \frac{1}{\sigma^2}\epsilon^T(\mathbf{I}_N-\mathbf{H})\epsilon$$

We have shown that $\mathbf{I}_N - \mathbf{H}$ is symmetric and idempotent, and ϵ is a random vector with independent elements of mean 0 and constant variance σ^2. Therefore, ϵ/σ would be a standard normal random vector. We have also shown that $\text{tr}(\mathbf{H}) = p+1$, so the trace $\text{tr}(\mathbf{I}_N - \mathbf{H}) = N-p-1$. Combining all these, with our general result, we have

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

as desired. □

Lemma 1.1.5: Independence between residual and parameters/least square estimates

The *residual* is defined as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_N - \mathbf{H})\mathbf{y}$.

- $\mathbf{e}$ is independent of the least square estimate $\hat{\mathbf{y}}$.
- $\mathbf{e}$ is independent of the parameters $\hat{\beta}$.
- $\hat{\beta}$ is independent of MSE $\hat{\sigma}^2$.

Proof.

- To prove the first one, recall the property of cross-covariance matrix that $\text{Cov}(\mathbf{AX}+a, \mathbf{BY}+b) = \mathbf{A}\text{Cov}(\mathbf{X},\mathbf{Y})\mathbf{B}^T$, we have

$$\text{Cov}(\mathbf{e},\hat{\mathbf{y}}) = \text{Cov}\left((\mathbf{I}_N-\mathbf{H})\mathbf{y}, \mathbf{H}\mathbf{y}\right) = (\mathbf{I}_N-\mathbf{H})\text{Var}(\mathbf{y})\mathbf{H}^T$$

Hao's Math Notes

Since by our assumption, $\mathbf{y}$'s components are independent and having constant variance σ^2, $\text{Var}(\mathbf{y})$ is just a diagonal matrix with σ^2 on the diagonal elements. Therefore,

$$\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = \sigma^2(\mathbf{I}_N - \mathbf{H})\mathbf{H}^T = \sigma^2(\mathbf{H}^T - \mathbf{H}\mathbf{H}^T) = \sigma^2(\mathbf{H} - \mathbf{H}^2) = \sigma^2(\mathbf{H} - \mathbf{H}) = \mathbf{0}$$

Since both $\mathbf{e}$ and $\hat{\mathbf{y}}$ are normally distributed, uncorrelated means independent.

- The second one follows the similar argument.

$$\begin{aligned}\text{Cov}(\mathbf{e}, \hat{\beta}) &= \text{Cov}((\mathbf{I}_N - \mathbf{H})\mathbf{y}, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) = (\mathbf{I}_N - \mathbf{H})\text{Var}(\mathbf{y})\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T \\ &= \sigma^2(\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) \\ &= \sigma^2\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\right) = \mathbf{0}\end{aligned}$$

Since both $\mathbf{e}$ and $\hat{\beta}$ are normally distributed, uncorrelated means independent.

- Since $\hat{\sigma}^2$ is a function of $\mathbf{e}$, i.e.,

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\mathbf{e}^T\mathbf{e}$$

and $\mathbf{e}$ is independent of $\hat{\beta}$, it is also independent of $\hat{\beta}$.

□

With these properties, we can formalize our final test statistics for inference.

Theorem 1.1.6: Hypothesis Test for Single Parameter

The test statistics for the null hypothesis $\beta_j = 0$ is the *Z-score*

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where $v_j = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}$ the jth diagonal element. Under the null hypothesis, $z_j \sim t_{N-p-1}$.

Proof. By Theorem 1.1.2,

$$\hat{\beta} \sim N\left(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\right)$$

Then, the marginal distribution of $\hat{\beta}_j$ is then the linear transformation with $L = (0, 0, \cdots, 0, 1, 0, \cdots, 0)^T$ with a single 1 on the jth component such that

$$\hat{\beta}_j = \mathbf{L}^T\hat{\beta} \sim N(\mathbf{L}^T\beta, \mathbf{L}^T(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2\mathbf{L}) = N\left(\beta_j, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}_{jj}\right)$$

Fan's Math Notes

Therefore, under null hypothesis that $\beta_j = 0$, we have

$$\frac{\hat{\beta}_j}{\sigma\sqrt{v_j}} \sim N(0,1)$$

Therefore, the Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}} = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}\frac{\sigma}{\hat{\sigma}} = \frac{\frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}}{\sqrt{\hat{\sigma}^2/\sigma^2}} = \frac{\frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}}{\sqrt{\frac{(N-p-1)\hat{\sigma}^2/\sigma^2}{N-p-1}}}$$

We have seen that $(N-p-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{N-p-1}$ from Lemma 1.1.4. Moreover, $\frac{\hat{\beta}_j}{\sigma\sqrt{v_j}}$ is independent from $(N-p-1)\hat{\sigma}^2/\sigma^2$ since by Lemma 1.1.5, $\hat{\beta}$ and $\hat{\sigma}^2$ is independent, and their functions must be independent. Therefore, it follows a t-distribution with degree of freedom $\text{df} = N-p-1$. □

Note: If the value of σ is known, then substitute $\hat{\sigma}$ by σ in the formula of z_j, it will simply follow standard normal distribution.

Often we need to test for the significance of groups of coefficients simultaneously. For example, to test if a categorical variable with k levels can be excluded from the model, we need to test whether the coefficients of the dummy variables can all be set to zero. If they are all zero, this forms a *nested model*.

Theorem 1.1.7: Hypothesis Test for Multiple Parameters

The test statistics for the null hypothesis $\mathbf{L}\beta = 0$, where $\mathbf{L}$ is a linear contrast, is the *F-statistic*

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/(p_1 - p_0)}{\text{RSS}_1/(N - p_1 - 1)}$$

where RSS_1 is the residual sum of squares for bigger model with $p_1 + 1$ parameters, and RSS_0 for the smaller one with $p_0 + 1$ parameters. Under the null hypothesis that the smaller model is correct, $F \sim F_{p_1-p_0, N-p_1-1}$.

Proof. Without losing generality, suppose that after reindexing parameters β, our null hypothesis is $\beta_{p_0+1} = \beta_{p_0+2} = \cdots = \beta_{p_1} = 0$. Then, the input for the smaller model is just $\mathbf{X}$ with the last $p_1 - p_0$ columns eliminated, denoted by $\mathbf{X}_{p_0+1}$. The input for the larger model is correspondingly denoted by $\mathbf{X}_{p_1+1}$. Denote their corresponding hat matrix by $\mathbf{H}_{p_0+1}$ and $\mathbf{H}_{p_1+1}$, respectively. Then, as derived in the proof of Proposition 1.1.3,

$$\text{RSS}_0 - \text{RSS}_1 = \epsilon^T(\mathbf{I}_N - \mathbf{H}_{p_0+1})\epsilon - \epsilon^T(\mathbf{I}_N - \mathbf{H}_{p_1+1})\epsilon = \epsilon^T(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})\epsilon$$

Math Notes

We need to prove that $(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})$ is idempotent, i.e., we need to prove that

$$(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})^2 = \mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1} \tag{1.5}$$

Expanding the square of right hand side, and recall that both $\mathbf{H}_{p_1+1}$ and $\mathbf{H}_{p_0+1}$ is symmetric and idempotent, we have

$$\begin{aligned}
&\mathbf{H}_{p_1+1}^2 - \mathbf{H}_{p_1+1}\mathbf{H}_{p_0+1} - \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1} + \mathbf{H}_{p_1+1}^2 = \mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1} \\
\Longrightarrow &\mathbf{H}_{p_1+1} - \mathbf{H}_{p_1+1}\mathbf{H}_{p_0+1} - \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1} + \mathbf{H}_{p_1+1} = \mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1} \\
\Longrightarrow &- \mathbf{H}_{p_1+1}\mathbf{H}_{p_0+1} - \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1} = -2\mathbf{H}_{p_0+1}
\end{aligned}$$

Therefore, we need to show that $\mathbf{H}_{p_1+1}\mathbf{H}_{p_0+1} + \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1} = 2\mathbf{H}_{p_0+1}$. It is tricky in algebra, but from geometric point of view, it is much more easier. Notice that $\mathbf{H}_{p_1+1}$ is a projection from the whole output space to column space of $\mathbf{X}_{p_1+1}$, and $\mathbf{H}_{p_0+1}$ is a projection from the whole output space to column space of $\mathbf{X}_{p_0+1}$. Denote these two column spaces by $\text{col}(\mathbf{X}_{p_1+1})$ and $\text{col}(\mathbf{X}_{p_0+1})$, respectively, and notice that

$$\text{col}(\mathbf{X}_{p_0+1}) \subseteq \text{col}(\mathbf{X}_{p_1+1})$$

since $\mathbf{X}_{p_0+1}$ is the submatrix of $\mathbf{X}_{p_1+1}$. Therefore, if both operator $\mathbf{H}_{p_1+1}$ and $\mathbf{H}_{p_0+1}$ is acted on an object, the object will be projected on the column space of submatrix $\text{col}(\mathbf{X}_{p_0+1})$, which is the same as being acted by the operator $\mathbf{H}_{p_0+1}$. Therefore,

$$\mathbf{H}_{p_1+1}\mathbf{H}_{p_0+1} = \mathbf{H}_{p_0+1} \quad \text{ and } \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1} = \mathbf{H}_{p_0+1}$$

which completes the proof that $(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})$ is idempotent. Note that the trace

$$\text{tr}(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1}) = (p_1 + 1) - (p_0 + 1) = p_1 - p_0$$

By the general result in the proof of Lemma 1.1.4, we have

$$\frac{\text{RSS}_0 - \text{RSS}_1}{\sigma^2} = \frac{\epsilon^T(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})\epsilon}{\sigma^2} \sim \chi^2_{p_1-p_0}$$

We have known that $\text{RSS}_1/\sigma^2 \sim \chi^2_{N-p-1}$, so we are done if we prove that $\text{RSS}_0 - \text{RSS}_1$ and RSS_1 is independent, since then the statistic F would be the form of $\frac{\chi^2_a/a}{\chi^2_b/b}$ with numerator and denominator independent. To do this, note that

$$\begin{aligned}
&\text{Cov}\left((\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})\epsilon, (\mathbf{I}_N - \mathbf{H}_{p_1+1})\epsilon\right) = (\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})(\mathbf{I}_N - \mathbf{H}_{p_1+1})^T\text{Var}(\epsilon) \\
&= \left(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_1+1}\mathbf{H}_{p_1+1}^T - \mathbf{H}_{p_0+1} + \mathbf{H}_{p_0+1}\mathbf{H}_{p_1+1}^T\right)\sigma^2 = \left(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1} + \mathbf{H}_{p_0+1}\right)\sigma^2 = \mathbf{0}
\end{aligned}$$

Fontuan's Math Notes

since they are multivariate normally distributed, uncorrelation implies independence. Since both $(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1}$ and $(\mathbf{I}_N - \mathbf{H}_{p_1+1})$ are idempotent,

$$\text{RSS}_0 - \text{RSS}_1 = \|(\mathbf{H}_{p_1+1} - \mathbf{H}_{p_0+1})\epsilon\|_2^2 \quad \text{and} \quad \text{RSS}_1 = \|(\mathbf{I}_N - \mathbf{H}_{p_1+1})\epsilon\|_2^2$$

are just functions of two independent random vectors. Therefore, they are also independent. □

It can be shown that the z_j are equivalent to the F statistic for dropping the single coefficient β_j from the model.

Proposition 1.1.8: Correspondence between Z-score and F statistic

F statistic for dropping a single coefficient from a model is equal to the square of the corresponding z-score.

Proof. By dropping a single coefficient from model, the F statistic will have $F_{1,N-p-1}$ distribution under null hypothesis, and z-score will follow t_{N-p-1} distribution. Therefore, z_j^2 then follows $F_{1,N-p-1}$ distribution. Thus both the z-score and the F statistic test identical hypotheses under identical distributions. Thus they must have the same value in this case. □

Aside from these hypothesis testing, we can also construct $100(1-\alpha\%$ confidence intervals for individual β_j:

$$\left(\hat{\beta}_j - z\left(1-\frac{\alpha}{2}\right)v_j^{\frac{1}{2}}\hat{\sigma}^2, \hat{\beta}_j + z\left(1-\frac{\alpha}{2}\right)v_j^{\frac{1}{2}}\hat{\sigma}^2\right)$$

Or we can construct the confidence band

$$C_\beta = \left\{\beta : (\hat{\beta}-\beta)^T\mathbf{X}^T\mathbf{X}(\hat{\beta}-\beta) \leqslant \hat{\sigma}^2\chi^2_{p+1}(1-\alpha)\right\}$$

1.1.3 The Gauss-Markov Theorem

What is good about least square estimation? The famous *Gauss-Markov Theorem* states that least square estimation of parameters β are the **Best Linear Unbiased Estimation (BLUE)**.

Theorem 1.1.9: Gauss-Markov Theorem

Let $\hat{\beta}$ be the least square estimation. If we have any other linear estimator $\beta^* = \mathbf{C}\mathbf{y}$, then

$$\text{Var}(\hat{\beta}) \leqslant \text{Var}(\beta^*)$$

Fantuan's Math Notes

Proof. Since we can always find a matrix $\mathbf{D}$ such that $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}$, we can show that

$$E(\beta^*) = E(\hat{\beta}) + \mathbf{D}E(\mathbf{y}) = \beta + \mathbf{D}\mathbf{X}\beta$$

since β^* is an unbiased estimator, we must have $\mathbf{D}\mathbf{X} = \mathbf{0}$. The variance of β^* is

$$\begin{aligned}\mathrm{Var}(\beta^*) &= \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)\mathrm{Var}(\mathbf{y})\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)^T = \sigma^2\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)^T \\ &= \sigma^2\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{D}\mathbf{X})^T + \mathbf{D}\mathbf{D}^T\right) \\ &= \sigma^2\left((\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^T\right) > (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2 = \mathrm{Var}(\hat{\beta})\end{aligned}$$

□

However, this theorem just considers unbiased estimators. Sometimes we can consider biased estimator so that the variance may be smaller to achieve a bias-variance trade-off.

1.1.4 Multiple Regression from Simple Regression

Suppose that we have a simple linear regression model with no intercept $Y = X\beta + \epsilon$. Then, the least square estimate and residuals are

$$\hat{\beta} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\langle \mathbf{x}, \mathbf{x}\rangle}, \quad \mathbf{r} = \mathbf{y} - \mathbf{x}\hat{\beta}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_N)^T$ and $\mathbf{y} = (y_1, y_2, \cdots, y_N)^T$. Now, if we consider a multiple linear regression, and denote the columns of the data matrix $\mathbf{X}$ by $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p$, if these columns are orthogonal, i.e., $\langle \mathbf{x}_j, \mathbf{x}_k\rangle = 0$ for all $j \neq k$, then we can show that

$$\hat{\beta}_j = \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\right)_j = (\mathbf{X}^T\mathbf{X})^{-1}_{jj}(\mathbf{X}^T\mathbf{y})_j = \frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{\langle \mathbf{x}_j, \mathbf{x}_j\rangle}$$

is just the univariate estimates, no effects on each other. However, observational data almost never achieves orthogonal columns. So we use *Gram-Schmidt Procedure* to orthogonalize these input columns.

Algorithm 1.1.10: Gram-Schmidt Procedure for Regression Input

1. Initialize $\mathbf{z}_0 = \mathbf{x}_0 = \mathbf{1}$.
2. For $j = 1, 2, \cdots, p$, regress $\mathbf{x}_j$ on $\mathbf{z}_0, \mathbf{z}_1, \cdots, \mathbf{z}_{j-1}$ to produce the coefficient $\hat{\gamma}_{lj} = \langle \mathbf{z}_l, \mathbf{x}_j\rangle / \langle \mathbf{z}_l, \mathbf{z}_l\rangle$, $l = 0, \cdots, j-1$ 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 $\hat{\beta}_p$.

Note that we can rearrange the order of these $\mathbf{x}_j$ so that any one of them can be in the last position. So, the multiple regression coefficient $\hat{\beta}_j$ just represents the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$, after it has been

Yuan's Math Notes

adjusted for all other $\mathbf{x}_k$ with $k \neq j$.

Therefore, if $\mathbf{x}_p$ is highly correlated with some other $\mathbf{x}_k$'s, the residual vector $\mathbf{z}_p$ will be close to zero, and the variance

$$\text{Var}(\hat{\beta}_p) = \frac{\sigma^2}{\|\mathbf{z}_p\|_2^2}$$

is very large. The value of $\hat{\beta}_p$ is then very unstable.

Not only just $\hat{\beta}_p$ can be obtained, but also the entire least square fit. To do this, represent Step 2 in the algorithm in matrix form

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma}$$

where $\mathbf{Z}$ has the column $\mathbf{z}_j$ in order, and $\boldsymbol{\Gamma}$ is the upper-triangular matrix with entries $\hat{\gamma}_{kj}$. Introducing diagonal matrix $\mathbf{D}$ with jth diagonal entry $D_{jj} = \|\mathbf{z}_j\|$. We have

$$\mathbf{X} = \mathbf{Z}\mathbf{D}^{-1}\mathbf{D}\boldsymbol{\Gamma} = \mathbf{Q}\mathbf{R}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{D}^{-1}$ and $\mathbf{R} = \mathbf{D}\boldsymbol{\Gamma}$ is the QR decomposition, where $\mathbf{Q}$ is unitary matrix (columns form orthonormal list), and $\mathbf{R}$ is upper-triangular with positive terms on diagonal. Unitary matrix has property $\mathbf{Q}^T\mathbf{Q} = I$. Then,

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

and

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{Q}\mathbf{R}\mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{Q}\mathbf{Q}^T\mathbf{y}$$

1.2 Subset Selection

1.2.1 Best Subset Selection

This method finds for each $k \in \{0, 1, 2, \cdots, p\}$ the subset of size k that gives smallest residual sum of squares. The best subset curve is necessarily decreasing since more information are integrated when adding more input vectors. Typically we choose smallest model that minimizes an estimate of expected prediction error $E(Y_0 - x_0^T\hat{\beta})^2$.

1.2.2 Forward and Backward Stepwise Selection

These are two *greedy algorithms*.

Fantuan's Math Notes

- **Forward stepwise** starts with the intercept, and then sequentially adds into predictor that most improves the fit.

- **Backward Stepwise** starts with full model, and then sequentially drops predictor that with samllest Z-score.

1.3 Shrinkage Methods

1.3.1 Ridge Regression

Ridge regression is proposed by Hoerl and Kennard (1970[4]). Ridge coefficients minimize a *penalized residual sum of squares*

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

where $\lambda \geqslant 0$ is the *tuning parameter*. An equivalent way to write ridge problem is

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

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

There is a one-to-one correspondence between parameter λ and t. When there are many correlated variables in a linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be canceled by a similarly large negative coefficient on its correlated cousin. By imposing a size constriant, this problem is alleviated.

The ridge solutions are not equivariant under scaling of the inputs, and so one normally standardizes the inputs before solving it.

Proposition 1.3.1: Standardizing inputs in Ridge Problem

The ridge regression problem is equivalent to the problem

$$\hat{\beta}^c = \underset{\beta^c}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c \right)^2 + \lambda \sum_{j=1}^{p} (\beta_j^c)^2 \right\}$$

The intercept is estimated by $\hat{\beta}_0^c = \bar{y}$, and the remaining coefficients get estimated by a ridge regression without intercept, using the centered x_{ij}.

Yuan's Math Notes

Proof. Note that

$$\sum_{i=1}^{N}\left(y_i-\beta_0^c-\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2=\sum_{i=1}^{N}\left(y_i-\beta_0^c-\sum_{j=1}^{p}x_{ij}\beta_j^c+\sum_{j=1}^{p}\bar{x}_j\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2$$
$$=\sum_{i=1}^{N}\left(y_i-\left(\beta_0^c-\sum_{j=1}^{p}\bar{x}_j\beta_j^c\right)-\sum_{j=1}^{p}x_{ij}\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2$$

Comparing above with the ridge problem, if we let $\beta_j^c=\beta_j$ for $j=1,2,\cdots,p$ and $\beta_0^c=\beta_0+\sum_{j=1}^{p}\bar{x}_j\beta_j$, then two equations are equivalent. The solution to the centered ridge problem is

$$\hat{\beta}^c=\underset{\beta^c}{\text{argmin}}\left\{\sum_{i=1}^{N}\left(y_i-\beta_0^c-\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2\right\}$$
$$=\underset{\beta^c}{\text{argmin}}\left\{\sum_{i=1}^{N}\left(\bar{y}-\beta_0^c+y_i-\bar{y}-\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2\right\}$$

We first take derivative w.r.t β_0^c, and it is obvious that the solution making the derivative equal to zero is $\beta_0^c=\bar{y}$. The remaining part is just a ridge problem with both $\mathbf{y}$ and $\mathbf{X}$ centered, with no intercept, i.e.,

$$\hat{\beta}^c=\underset{\beta^c}{\text{argmin}}\left\{\sum_{i=1}^{N}\left(y_i-\bar{y}-\sum_{j=1}^{p}(x_{ij}-\bar{x}_j)\beta_j^c\right)^2+\lambda\sum_{j=1}^{p}(\beta_j^c)^2\right\}$$

□

Due to this consideration, we assume that the centering has been done, so that input matrix $\mathbf{X}$ has p instead of $p+1$ columns. The solution has closed form, and it is straightforward from linear regression.

Theorem 1.3.2: Solution to Ridge Regression

The solution to the ridge regression is

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

Understanding Ridge Regression using SVD

Singular Value Decomposition gives us an additional insight to ridge regression. Since $\mathbf{X}$ has full rank, the SVD of $N\times p$ matrix $\mathbf{X}$ is

$$\mathbf{X}=\mathbf{UDV}^T$$

where $\mathbf{U}$ and $\mathbf{V}$ are $N\times p$ and $p\times p$ orthonormal matrices with columns of $\mathbf{U}$ spanning the column space of $\mathbf{X}$, and the columns of $\mathbf{V}$ spanning the row space. $\mathbf{D}$ is $p\times p$ diagonal matrix with diagonal entries

Fan Zhou's Math Notes

$d_1 \geqslant d_2 \geqslant \cdots \geqslant d_p \geqslant 0$ are singular values of $\mathbf{X}$.

Using SVD for least square fit, we have

$$\begin{aligned}\mathbf{X}\hat{\beta}^{ls} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}^2\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{D}^2\mathbf{V}\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{D}^{-2}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

since unitary matrix has the property $AA^T = I$ and $A^TA = I$. The final $\mathbf{u}_j$ represents the columns of $\mathbf{U}$. The position of $\mathbf{D}$ can be changed since it is a diagonal matrix. Similarly, the ridge solution can be represented as

$$\begin{aligned}\mathbf{X}\hat{\beta}^{\text{ridge}} &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^T(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^T\mathbf{y} = \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{D}\mathbf{U}^T\mathbf{y} \\ &= \sum_{j=1}^{p}\mathbf{u}_j\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

Comparing these two, we can see that ridge solution shrinks the fitted coordinates by the factors $d_j^2/(d_j^2+\lambda)$. This means that with fixed λ, greater amount of shrinkage is applied to the coordiantes of basis vectors with smaller d_j^2.
The square of singular values are eigenvectors of $\mathbf{X}^T\mathbf{X}$, which can be also seen from the eigendecomposition

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$$

The eigenvectors v_j are called the *principal components* directions of $\mathbf{X}$. The first principal component direction v_1 has the property that $\mathbf{z}_1 = \mathbf{X}v_1$ has the largest sample variance amongst all normalized linear combinations of the columns of $\mathbf{X}$. This sample variance is

$$\text{Var}(\mathbf{z}_1) = \text{Var}(\mathbf{X}v_1) = \frac{d_1^2}{N}$$

$\mathbf{z}_1$ is the *first principal component* of $\mathbf{X}$. Subsequent principal components $\mathbf{z}_j$ have maximum variance d_j^2/N, subject to being orthogonal to the earlier ones. Conversely the last principal component has minimum variance. Therefore,

Small singular values d_j corresponds to directions in the column space $\mathbf{X}$ having small variance, and ridge regression shrinks these directions the most.

Fantuan's Main Notes

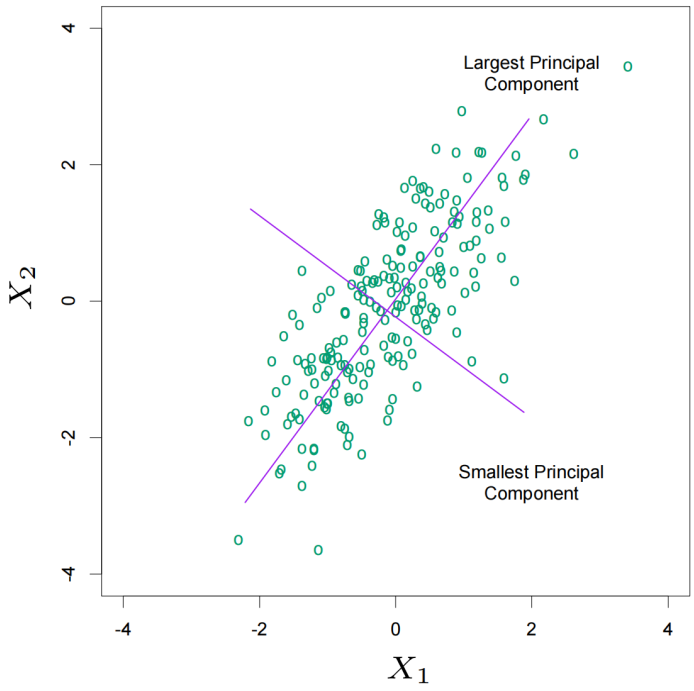


Figure 1.3: Principal components

With orthogonal inputs, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, so the singular values are 1, the ridge estimates are just a scaled version of the least square fit, that is, $\hat{\beta}^{\text{ridge}} = \hat{\beta}/(1+\lambda)$.

For least square fit we have degrees of freedom $\text{df} = p + 1$. Note that $\text{tr}(\mathbf{H}) = p + 1$ in this case. Therefore, we can generalize this into ridge regression.

Definition 1.3.3: Effective Degrees of Freedom

The **effective degrees of freedom** in ridge regression is defined as

$$\text{df}(\lambda) = \text{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Note that $\text{df}(\lambda) = p$ when $\lambda = 0$ (no regularization) and $\text{df}(\lambda) \to 0$ as $\lambda \to \infty$.

Understanding Ridge Regression using Bayesian Statistics

We can also see ridge regression from the point of view of Bayesian statistics. It can be derived as the mean or mode of a posterior distribution.

Theorem 1.3.4: Ridge Regression from Bayesian Prior

Assume $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ and the parameters have a Gaussian prior $\beta \sim N(0, \tau^2\mathbf{I})$, independently from each other, and β_0 is not governed by a prior (i.e., the prior distribution is only for β_j with $j = 1, 2, \cdots, p$). Assume σ^2 and τ^2 are known. Then,

- The ridge regression estimate is the mean (and mode) of the posterior distribution of β, with the tuning parameter $\lambda = \sigma^2/\tau^2$.
- The negative log-posterior density of β is proportional to $\sum_{i=1}^{N}(y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$ where $\lambda = \sigma^2/\tau^2$.

Proof. By the assumption, the joint distribution of $(\mathbf{y}, \beta)$ is given by

$$p(\mathbf{y}, \beta) = p(\mathbf{y}|\beta)p(\beta) \propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2}\right)\exp\left(-\frac{\|\beta\|_2^2}{2\tau^2}\right)$$

Hence we have, seeing β as the variable,

$$\begin{aligned}
p(\beta|\mathbf{y}) = \frac{p(\mathbf{y}, \beta)}{p(\mathbf{y})} &\propto p(\mathbf{y}, \beta) \propto \exp\left(-\frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2} - \frac{\|\beta\|_2^2}{2\tau^2}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) - \frac{1}{2\tau^2}\beta^T\beta\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(-\mathbf{y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) - \frac{1}{2\tau^2}\beta^T\beta\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\left(-2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) - \frac{1}{2\tau^2}\beta^T\beta\right) \\
&\propto \exp\left(-\frac{1}{2}\beta^T\left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \frac{1}{\tau^2}\mathbf{I}\right)\beta + \frac{1}{\sigma^2}\beta^T\mathbf{X}^T\mathbf{y}\right)
\end{aligned}$$

This shows that, the posterior distribution of β is normal with mean

$$E(\beta|\mathbf{y}) = (-\frac{b}{2a}) = \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \frac{1}{\tau^2}\mathbf{I}\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y} = \left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

which is the solution of ridge regression with tuning parameter $\lambda = \sigma^2/\tau^2$. Since the posterior is Gaussian, this is also the mode.

Finally, we see that the negative log posterior density of β is

$$-\log(p(\beta|\mathbf{y})) \propto \frac{\|\mathbf{y} - \mathbf{X}\beta\|_2^2}{2\sigma^2} + \frac{\|\beta\|_2^2}{2\tau^2} \propto \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \frac{\sigma^2}{\tau^2}\|\beta\|_2^2$$

equal to the ridge problem. □

Fantuan's Math Notes

1.3.2 Lasso

Lasso (least absolute shrinkage and selection operator) was first proposed by Tibshirani (1996[6]). Instead of using L_2 penalty in ridge regression, lasso use L_1 penalty

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

where $\lambda \geqslant 0$ is the *tuning parameter*. An equivalent way to write lasso problem is

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

$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leqslant t$$

Just as in ridge regression, we can center the input matrix and the solution for $\hat{\beta}_0$ is $\bar{y}$, and thereafter we fit a model without intercept. There is no closed form solution to lasso.

Difference between Subset Selection, Ridge Regression and Lasso

Below is the coefficient profiles for ridge regression (left) and lasso (right). The lasso one is plotted versus the standardized tuning parameter $s = t/\sum_{j=1}^{p} |\hat{\beta}_j^{ls}|$. Notice that ridge regression will not shrink coefficients

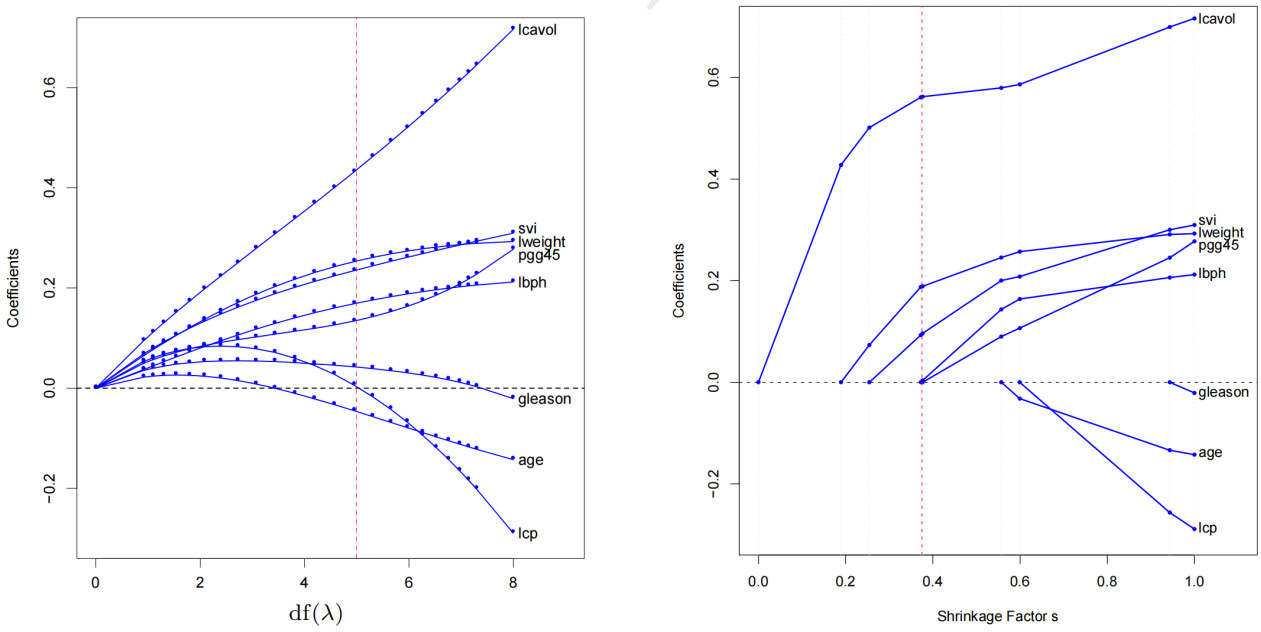


Figure 1.4: Left: Ridge Regression. Right: Lasso. The vertical line is chosen by cross validation

exactly to zero after λ is larger than a value. However, lasso will cause some coefficients to be exactly zero

Main Notes

when t is sufficiently small. Thus lasso does a kind of continuous subset selection. This can be seen from figure below. The residual sum of squares has elliptical contours, centered at the full least square estimate. In 2-dim case shown below, the constraint region for ridge regression is $\beta_1^2 + \beta_2 \leqslant t$, while that for lasso is $|\beta_1| + |\beta_2| \leqslant t$. Both methods find the first point where the elliptical contours hit the constraint region. Since the diamond has corners, it can set som β_j equal to zero.

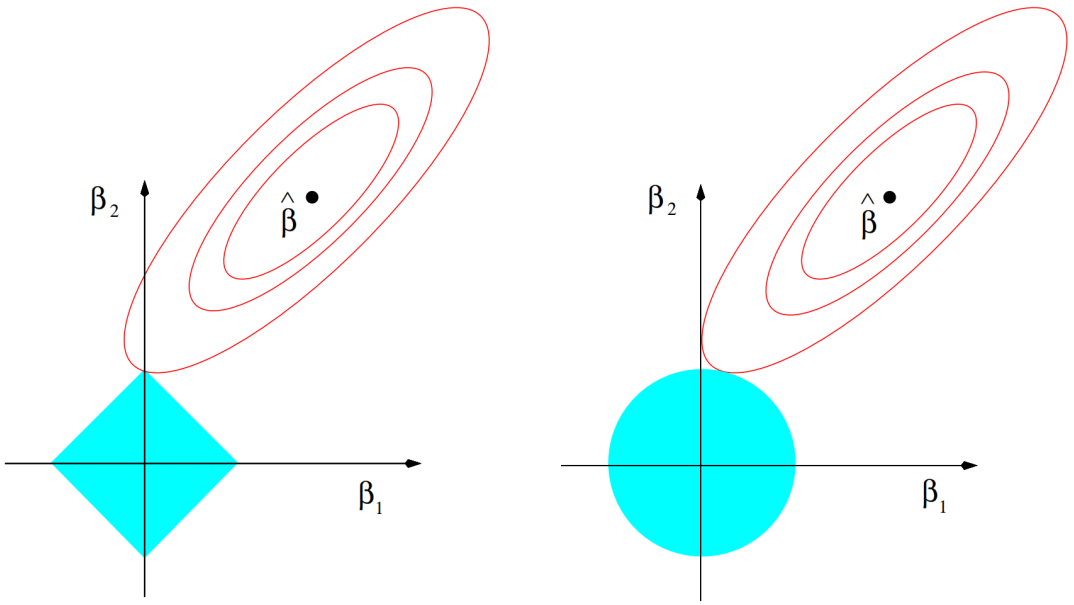


Figure 1.5: Left: lasso penalty. Right: Ridge penalty

If input matrix $\mathbf{X}$ is orthonormal (or just orthogonal), best subset, ridge and lasso just applies a simple transformation to the least square fit $\hat{\beta}_j$, as detailed below in the table.

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	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

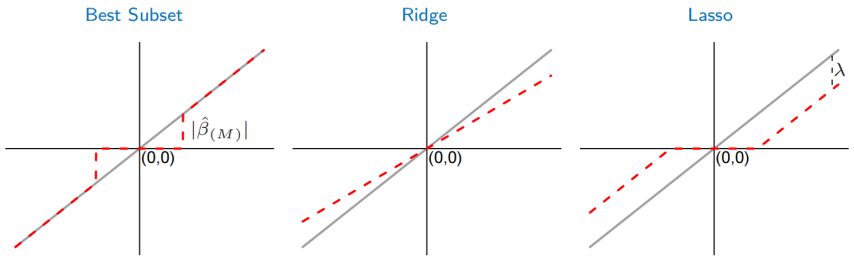


Figure 1.6: Shrinkage parameter when $\mathbf{X}$ is orthogonal

Proof.

- *Best subset with orthogonal input:* For best subset, the QR decomposition of $\mathbf{X}$ is $\mathbf{X} = \mathbf{XI}$. Then, $\hat{\beta} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} = \mathbf{X}^T\mathbf{y}$. For a best subset of size M, we need to choose a sub-matrix $X_S = [X_{j1}, X_{j2}, \cdots, X_{jM}]$ consisting of columns $j_1, j_2, \cdots, j_M$. Use the submatrix to fit the regression line, we have

$$\hat{\beta}^M = \mathbf{X}_S^T\mathbf{y} = [\mathbf{X}_{j1}^T\mathbf{y}, \mathbf{X}_{j2}^T\mathbf{y}, \cdots, \mathbf{X}_{jM}^T\mathbf{y}] = (\hat{\beta}_{j1}, \hat{\beta}_{j2}, \cdots, \hat{\beta}_{jM})$$

To minimize the residual sum of square, one need to minimize

$$\begin{aligned}\left\|\mathbf{y} - \mathbf{X}_S\hat{\beta}^M\right\|_2^2 &= \left\|\mathbf{y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}_S\hat{\beta}^M\right\|_2^2 \\ &= \left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|_2^2 + 2\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)^T\left(\mathbf{X}\hat{\beta} - \mathbf{X}_S\hat{\beta}^M\right) + \left\|\mathbf{X}\hat{\beta} - \mathbf{X}_S\hat{\beta}^M\right\|_2^2 \\ &= \left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|_2^2 + \left\|\mathbf{X}\hat{\beta} - \mathbf{X}_S\hat{\beta}^M\right\|_2^2\end{aligned}$$

where the middle term is 0 since residual is orthogonal to the columns of $\mathbf{X}$. Note that in the last equation, the first term is fixed. So we need to minimize the second term. Note that

$$\left\|\mathbf{X}\hat{\beta} - \mathbf{X}_S\hat{\beta}^M\right\|_2^2 = \left\|\mathbf{X}_K\hat{\beta}^{p-M}\right\|_2^2$$

where $\mathbf{X}_k$ is the submatrix of $\mathbf{X}$ without those columns included in $\mathbf{X}_S$, and $\hat{\beta}^{p-M}$ is those corresponding parameters. Then, since $\mathbf{X}$ is orthogonal, we have

$$\left\|\mathbf{X}_K\hat{\beta}^{p-M}\right\|_2^2 = \left(\hat{\beta}^{p-M}\right)^T\mathbf{X}_K^T\mathbf{X}_K\hat{\beta}^{p-M} = \sum_{k\neq j_1, j_2, \cdots, j_M}\hat{\beta}_k^2$$

Thus, we need to choose the smallest $p - M$ parameters (not considering the intercept) to be excluded from the parameter set. Therefore, we need to choose the largest M ones.

- *Ridge Regression with orthogonal input:*

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

- *Lasso with orthogonal input:* For the lasso penalty with orthogonal matrix, we have

$$\begin{aligned}&\underset{\beta}{\text{argmin}}\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\sum_{j=1}^p|\beta_j|\right\} = \underset{\beta}{\text{argmin}}\left\{\frac{1}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) + \lambda\sum_{j=1}^p|\beta_j|\right\} \\ &= \underset{\beta}{\text{argmin}}\left\{-\mathbf{y}^T\mathbf{X}\beta + \frac{1}{2}\beta^T\beta + \lambda\sum_{j=1}^p|\beta_j|\right\} = \underset{\beta}{\text{argmin}}\left\{\sum_{j=1}^p\left(-\hat{\beta}_j\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j|\right)\right\}\end{aligned}$$

Fantuan's Math Notes

Therefore, we can separately minimize each β_i. Denote the solution by β^*.

(a) If $\hat{\beta}_j \geqslant 0$, then we must have $\beta_j^* \geqslant 0$, otherwise if smaller than zero, we can let $\beta_j^{**} = -\beta_j^*$ where $\epsilon > 0$, and the lasso penalty least square value is smaller (the first term becomes negative, the second and the third term does not change), which contradicts the optimality.

Therefore, we need to solve

$$\underset{\beta_j \geqslant 0}{\operatorname{argmin}} \left\{ \frac{1}{2}\beta_j^2 + (\lambda - \hat{\beta}_j)\beta_j \right\}$$

and it is quadratic in β_j, thus the solution is $(\hat{\beta}_j - \lambda)_+$ (since we need $\beta_j \geqslant 0$).

(b) Similarly, for $\hat{\beta}_j < 0$, we need also $\beta_j^* < 0$, and the situation is the same. We finally get $-(\lambda + \hat{\beta}_j)_+$.

In conclusion, the solution is $\operatorname{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$. □

Generalization of Shrinkage Methods

We can use L_q norm with $q \geqslant 0$ as our shrinkage criterion

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$

Values $q \in (1, 2)$ suggest a compromise betwen lasso and ridge regression. Moreover, Zou and Hastie (2005[8]) introduced the **elastic-net penalty**

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

a different compromise betwen ridge and lasso.

Understanding Generalized Shrinkage Methods from Bayesian Point of View

Thinking of $|\beta_j|^q$ as the log-prior density for β_j, these are also equicontours of the prior distribution of parameters. For example, in Theorem 1.3.4, when $q = 2$ we see the prior as proportional to $\exp\left(\|\beta\|_2^2\right)$. When $q = 1$, the prior is an independent double exponential distribution for each input, with density proportional to $\exp(-|\beta|)$.

In this view, lasso, ridge regression and best subset selection are Bayes estimates with different priors, and the parameter estimates are derived as **posterior modes**, which is the maximizer of posterior. Note that for lasso and best subset, mean does not equal to model.

Pantuan's Math Notes

1.3.3 Least Angle Regression (LAR)

Least Angel Regression (LAR) is first proposed by Efron et al. (2004[2]). It is an ameliorative forward stepwise regression.

Algorithm 1.3.5: Least Angle Regression

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \cdots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue until all p vectors have been entered. After $\min(N-1, p)$ steps, we arrive the full least-squares solution.

There are few things need to be explained in this procedure:

- Suppose $\mathcal{A}_k$ is the active set of variables at the beginning of the kth step, and let $\beta_{\mathcal{A}_k}$ be the coefficient vector for these variables at this step. There will be $k-1$ nonzero values, and the one just entered will be zero. If $\mathbf{r}_k = \mathbf{y} - \mathbf{X}_{\mathcal{A}_k}\beta_{\mathcal{A}_k}$ is the current residual, then the direction of this step is

$$\delta_k = \left(\mathbf{X}_{\mathcal{A}_k}^T \mathbf{X}_{\mathcal{A}_k}\right)^{-1} \mathbf{X}_{\mathcal{A}_k}^T \mathbf{r}_k$$

  The coefficient profile then evolves as

$$\beta_{\mathcal{A}_k}(\alpha) = \beta_{\mathcal{A}_k} + \alpha\delta_k$$

  If the fit vector at the beginning of this step is $\hat{\mathbf{f}}_k$, then it evolves as

$$\hat{\mathbf{f}}_k(\alpha) = \hat{\mathbf{f}}_k + \alpha\mathbf{u}_k, \quad \text{where } \mathbf{u}_k = \mathbf{X}_{\mathcal{A}_k}\delta_k$$

- If $p > N-1$, the LAR algorithm reaches a zero residual solution after $N-1$ steps (the -1 is because we have centered the data).

During Step 4 of LAR algorithm, or with more parameters involved, LAR algorithm actually keep the correlations *tied and decreasing*, when moving from their values to the defined direction.

Fantuan's Ma

Proposition 1.3.6: Tied and Decreasing Correlation of Activated Parameters

The LAR algorithm keeps the activated parameters' correlations with the current residual tied and monotonically decreasing.

Proof. Consider a regression problem with all variables and response having mean zero and standard deviation one. Suppose also that each variable has identical absolute correlation with the response

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y}\rangle| = \lambda, \quad j = 1, 2, \cdots, p$$

Let $\hat{\beta}$ be the least squares coefficient of $\mathbf{y}$ on $\mathbf{X}$, and let $\mathbf{u}(\alpha) = \alpha \mathbf{X}\hat{\beta}$ for $\alpha \in [0, 1]$ be the vector that moves a fraction α toward the least squares fit $\mathbf{u}$. Let RSS be the residual sum of squares from the full least square fit. Then, after progressing towards the direction $\mathbf{u}$, the correlation with the current residual:

$$\frac{1}{N}\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle\right| = \frac{1}{N}\left|\langle \mathbf{x}_j, \alpha(\mathbf{y} - \mathbf{X}\hat{\beta}) + (1-\alpha)\mathbf{y}\rangle\right| = \frac{1}{N}|\alpha\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}\rangle + (1-\alpha)\langle \mathbf{x}_j, \mathbf{y}\rangle|$$

Since each $\mathbf{x}_j$ is orthogonal to the residual $\mathbf{y} - \mathbf{X}\hat{\beta}$, we have $\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}\rangle = 0$ for each j. Hence,

$$\frac{1}{N}\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle\right| = \frac{1-\alpha}{N}|\langle \mathbf{x}_j, \mathbf{y}\rangle| = (1-\alpha)\lambda$$

and hence the correlation of each $\mathbf{x}_j$ with the residuals remain tied in magnitude as we progress toward $\mathbf{u}$.

Now these are not real correlation values. we start with standardized variables and response. However, after moving towards the least square fit, the responses are no longer standard. Therefore, to calculate the correlations, we need to calculate

$$\lambda(\alpha) = \frac{\frac{1}{N}\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle\right|}{\sqrt{\frac{\|\mathbf{x}_j\|^2}{N}}\sqrt{\frac{\|\mathbf{y}-\mathbf{u}(\alpha)\|^2}{N}}} = \frac{\frac{1}{N}\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle\right|}{\sqrt{\frac{\|\mathbf{y}-\mathbf{u}(\alpha)\|^2}{N}}}$$

where the second equation holds since the input matrix is not changed. We can calculate

$$\begin{aligned}
\|\mathbf{y} - \mathbf{u}(\alpha)\|^2 &= \left\|\alpha(\mathbf{y} - \mathbf{X}\hat{\beta}) + (1-\alpha)\mathbf{y}\right\|^2 \\
&= \alpha^2\left\|\mathbf{y} - \mathbf{X}\hat{\beta}\right\|^2 + 2\alpha(1-\alpha)\langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{y}\rangle + (1-\alpha)^2\|\mathbf{y}\|^2 \\
&= \alpha^2\text{RSS} + 2\alpha(1-\alpha)\left[\langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{y} - \mathbf{X}\hat{\beta}\rangle + \langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta}\rangle\right] + (1-\alpha)^2\|\mathbf{y}\|^2 \\
&= \alpha^2\text{RSS} + 2\alpha(1-\alpha)\left[\text{RSS} + \langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta}\rangle\right] + (1-\alpha)^2\|\mathbf{y}\|^2 \qquad (1.6)
\end{aligned}$$

Now notice that

$$\langle \mathbf{y} - \mathbf{X}\hat{\beta}, \mathbf{X}\hat{\beta}\rangle = (\mathbf{y} - \mathbf{X}\hat{\beta})^T\mathbf{X}\hat{\beta} = \mathbf{y}^T\mathbf{X}\hat{\beta} - \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta}$$

Fantuan's Math Notes

$$= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$= \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = 0$$

and

$$\frac{1}{N}\|\mathbf{y}\|^2 = 1$$

since $\mathbf{y}$ is assumed to be standardized. With these two, substitute back to Equation 1.6, we have

$$\frac{1}{N}\|\mathbf{y} - \mathbf{u}(\alpha)\|^2 = \frac{\alpha^2}{N}\text{RSS} + \frac{2\alpha(1-\alpha)}{N}\text{RSS} + (1-\alpha)^2 = \frac{\alpha(2-\alpha)}{N}\text{RSS} + (1-\alpha)^2$$

Therefore, the correlation is then

$$\lambda(\alpha) = \frac{\frac{1}{N}\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{u}(\alpha)\rangle\right|}{\sqrt{\frac{\|\mathbf{y}-\mathbf{u}(\alpha)\|^2}{N}}} = \frac{1-\alpha}{\sqrt{\frac{\alpha(2-\alpha)}{N}\text{RSS} + (1-\alpha)^2}}\lambda$$

which can be assessed that it is a decreasing function w.r.t α, and when $\alpha \to 1$, it converges to 0. □

The name 'least angle' arises from a geometrical interpretation of this process: $\mathbf{u}_k = \mathbf{X}_{\mathcal{A}_k}\delta_k$ makes the smallest (and equal) angle with each of the predictors in $\mathcal{A}_k$.

Proposition 1.3.7: Least Angle Property

the LAR direction $\mathbf{u}_k$ makes an equal angle with each of the predictors in $\mathcal{A}_k$.

Proof. Let λ denote the common correlation at the beginning of step k between the predictors in $\mathcal{A}_k$ and the residual $\mathbf{r}_k$. Then,

$$\langle \mathbf{x}_j, \mathbf{u}_k\rangle = \mathbf{x}_j^T\mathbf{X}_{\mathcal{A}_k}\left(\mathbf{X}_{\mathcal{A}_k}^T\mathbf{X}_{\mathcal{A}_k}\right)^{-1}\mathbf{X}_{\mathcal{A}_k}^T\mathbf{r}_k = \left(\mathbf{X}_{\mathcal{A}_k}^T\mathbf{X}_{\mathcal{A}_k}\left(\mathbf{X}_{\mathcal{A}_k}^T\mathbf{X}_{\mathcal{A}_k}\right)^{-1}\mathbf{X}_{\mathcal{A}_k}^T\mathbf{r}_k\right)_j$$
$$= \left(\mathbf{X}_{\mathcal{A}_k}^T\mathbf{r}_k\right)_j = \lambda$$

for all j. The cosine of the angle between $\mathbf{x}_j$ and $\mathbf{u}_k$ is given by

$$\frac{\langle \mathbf{x}_j, \mathbf{u}_k\rangle}{\|\mathbf{x}_j\|\|\mathbf{u}_k\|} = \frac{\langle \mathbf{x}_j, \mathbf{u}_k\rangle}{\|\mathbf{u}_k\|}$$

which is uniquely determined. Therefore, the angle between $[0, \pi)$ is uniquely determined. □

Till now, we have not talked about the step length α. If we need to take small steps and recheck the correlation every time, the order of computation is extremely large and not efficient. Luckily, *we can work out the exact step length at the beginning of each step.* The idea of this uses the *piecewise linearity* of this algorithm. Note

Fantuan's Math Notes

that in each step, we move β linearly, until another one comes in. Then, after the other one comes in, they continue move linearly together, with probably different slopes.

Proposition 1.3.8: Step Length and Criteria for next Activating Variable

At the beginning of step k of LAR algorithm, let $\mathbf{x}_k$ be the newly added variable, and $\mathbf{x}_b \notin \mathcal{A}_k$ be any variable that has not been activated. Then, either

$$\alpha_{b1} = \frac{\mathbf{x}_k^T\mathbf{r}_k - \mathbf{x}_b^T\mathbf{r}_k}{\mathbf{x}_k^T\mathbf{X}_{\mathcal{A}_k}\delta_k - \mathbf{x}_b^T\mathbf{X}_{\mathcal{A}_k}\delta_k} = \frac{\langle \mathbf{x}_k, \mathbf{y} - \hat{\mathbf{f}}_k\rangle - \langle \mathbf{x}_b, \mathbf{y} - \hat{\mathbf{f}}_k\rangle}{\langle \mathbf{x}_k, \mathbf{u}_k\rangle - \langle \mathbf{x}_b, \mathbf{u}_k\rangle}$$

or

$$\alpha_{b2} = \frac{\mathbf{x}_k^T\mathbf{r}_k + \mathbf{x}_b^T\mathbf{r}_k}{\mathbf{x}_k^T\mathbf{X}_{\mathcal{A}_k}\delta_k + \mathbf{x}_b^T\mathbf{X}_{\mathcal{A}_k}\delta_k} = \frac{\langle \mathbf{x}_k, \mathbf{y} - \hat{\mathbf{f}}_k\rangle + \langle \mathbf{x}_b, \mathbf{y} - \hat{\mathbf{f}}_k\rangle}{\langle \mathbf{x}_k, \mathbf{u}_k\rangle + \langle \mathbf{x}_b, \mathbf{u}_k\rangle}$$

will fall into $(0, 1)$. Choose the one that falls into this region and denote it as α_b. Then, The next variable to enter the active set is the x_b with the smallest corresponding α_b, and the corresponding α_b is the value of the step length α.

Proof. The current correlation with residual at α for $\mathbf{x}_k$ is

$$c_j(\alpha) = \langle \mathbf{x}_k, \mathbf{y} - \hat{\mathbf{f}}_k(\alpha)\rangle = \mathbf{x}_k^T(\mathbf{y} - (\hat{\mathbf{f}}_k - \alpha\mathbf{u}_k)) = \mathbf{x}_k^T\mathbf{r}_k - \alpha\mathbf{x}_k^T\mathbf{u}_k$$

By Proposition 1.3.6 and 1.3.7, this correlation has the same magnitude with other variables that has been activated, and it is decreasing with α monotonically. The criteria for the next variable to add in is to choose α such that the *largest correlation between an variable that is not activated and the residual is equal to this current correlation*. Let $\mathbf{x}_b \notin \mathcal{A}_k$, we want to find the α where

$$|c_j(\alpha)| = \mathbf{x}_b^T\mathbf{r}_k - \alpha\mathbf{x}_b^T\mathbf{u}_k$$

- For $c_j(\alpha) \geqslant 0$, we have

$$\mathbf{x}_k^T\mathbf{r}_k - \alpha\mathbf{x}_k^T\mathbf{u}_k = \mathbf{x}_b^T\mathbf{r}_k - \alpha\mathbf{x}_b^T\mathbf{u}_k \quad \Longrightarrow \quad \alpha = \frac{\langle \mathbf{x}_k, \mathbf{r}_k\rangle - \langle \mathbf{x}_b, \mathbf{r}_k\rangle}{\langle \mathbf{x}_k, \mathbf{u}_k\rangle - \langle \mathbf{x}_b, \mathbf{u}_k\rangle}$$

- For $c_j(\alpha) < 0$, we have

$$-\mathbf{x}_k^T\mathbf{r}_k + \alpha\mathbf{x}_k^T\mathbf{u}_k = \mathbf{x}_b^T\mathbf{r}_k - \alpha\mathbf{x}_b^T\mathbf{u}_k \quad \Longrightarrow \quad \alpha = \frac{\langle \mathbf{x}_k, \mathbf{r}_k\rangle + \langle \mathbf{x}_b, \mathbf{r}_k\rangle}{\langle \mathbf{x}_k, \mathbf{u}_k\rangle + \langle \mathbf{x}_b, \mathbf{u}_k\rangle}$$

Then, the smallest α such that first non-activated variable $\mathbf{x}_b$ achieves this is the step length, and this $\mathbf{x}_b$ is then added to the set $\mathcal{A}_{k+1}$ in the next step. □

Fantuan's Mat

Lasso using Modified algorithm for LAR

Below is the coefficient profile for LAR and Lasso w.r.t the L_1 arc length. The L_1 arc length of a differentiable curve $\beta(s)$ for $s \in [0, S]$ is given by $\mathrm{TV}(\beta, s) = \int_0^S \|\dot{\beta}(s)\|_1 \, \mathrm{d}s$. For piecewise linear LAR coefficeint profile, this amounts to summing the L_1 norms of the changes in coefficients from step to step.

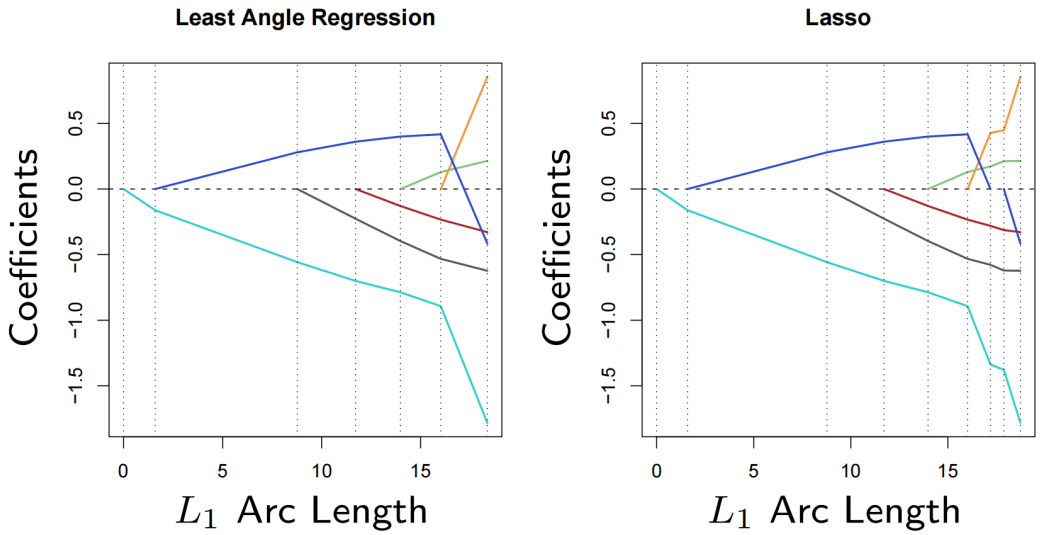


Figure 1.7: Left: LAR profile. Right: Lasso profile

They are almost identical, except when the blue coefficient passes through zero. The least angle is continuous, and lasso is broken. This suggests us the algorithm for computing lasso solutions:

Algorithm 1.3.9: Modified LAR algorithm for Lasso

Based on Algorithm 1.3.5, when moving towards the defined direction, if a non-zero coefficient hits zero, drop its variable from the active set of variables and re-compute the current joint least squares direction.

To see why these algorithms are so similar, suppose $\mathcal{A}$ is the active set of variables at some stage in the algorithm, we can express the tied correlation as

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\beta) = \gamma s_j, \quad \forall j \in \mathcal{A}$$

where $s_j \in \{-1, 1\}$ is the sign of correlation. Also $|\mathbf{x}_k^T(\mathbf{y} - \mathbf{X}\beta)| \leqslant \gamma$ for all $k \notin \mathcal{A}$. Now consider lasso,

$$R(\beta) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

Let $\mathcal{B}$ be the activate set for a given λ. For these variables, $|\hat{\beta}^{\text{lasso}}| > 0$, so $R(\beta)$ is differentiable, and the first

order derivative equals to 0 gives us

$$\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\beta) = \lambda \text{sign}(\beta_j), \quad \forall j \in \mathcal{B}$$

We see that they are identical, unless some passes through zero, and is kicked out from the set $\mathcal{B}$. Also we see that $|\mathbf{x}_k^T(\mathbf{y} - \mathbf{X}\beta)| \leqslant \lambda$ for all $k \notin \mathcal{B}$ is also satisfied for lasso.

1.4 Derived Input Direction Methods

1.4.1 Principal Components Regression (PCR)

Principal component regression forms the derived input columns $\mathbf{z}_m = \mathbf{X}v_m$, where v_m are principle component directions, i.e.,the eigenvectors of $\mathbf{X}^T\mathbf{X}$. $\mathbf{y}$ is regressed on $\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_M$ for some $M \leqslant p$. Sine $\mathbf{z}_m$ are orthogonal, this regression is just a sum of univariate regressions

$$\hat{\mathbf{y}}_{(M)}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{z}_m, \quad \hat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle}$$

Since $\mathbf{z}_m$ are linear combinations of the original $\mathbf{x}_j$, we can express the solution above in terms of $\mathbf{x}_j$ such that

$$\hat{\mathbf{y}}_{(M)}^{\text{pcr}} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathbf{X} v_m = \bar{y}\mathbf{1} + \mathbf{X}\sum_{m=1}^{M} \hat{\theta}_m v_m = \bar{y}\mathbf{1} + \mathbf{X}\hat{\beta}^{\text{pcr}}(M)$$

where $\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$. As with ridge regression, we first standardize the inputs. Note that if $M = p$, we would just get back the usual least squares estimates. This can be seen clearer from SVD. Recall the SVD for $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$. Since $\mathbf{V}$ is orthogonal, we have

$$\mathbf{Z} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D}$$

so we have

$$\mathbf{z}_m = \mathbf{X}v_m = d_m \mathbf{u}_m$$

We can reformulate $\hat{\theta}_m$ as

$$\hat{\theta}_m = \frac{\langle \mathbf{z}_m, \mathbf{y} \rangle}{\langle \mathbf{z}_m, \mathbf{z}_m \rangle} = \frac{d_m \mathbf{u}_m^T \mathbf{y}}{d_m^2 \mathbf{u}_m^T \mathbf{u}_m} = \frac{\langle \mathbf{u}_m, \mathbf{y} \rangle}{d_m}$$

So the β's can be written as

$$\hat{\beta}^{\text{pcr}}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}$$

Using simple algebra, we can also see that

$$\hat{\beta}^{\text{ls}}(p) = \mathbf{V}\mathbf{D}^{-1}\mathbf{U}^T\mathbf{y}$$

Fantuan's Main Notes

as well.

Ridge regression shrinks the coefficients of principal compoenents, whereas PCR discards the $p-M$ smallest eigenvalue components.

1.4.2 Partial Least Squares (PLS)

Partial Least Square is introduced by Wold (1975[7]). It also regress $\mathbf{y}$ on several orthogonal directions $\mathbf{z}_1, \cdots, \mathbf{z}_M$.

Algorithm 1.4.1: 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 $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, 2, \cdots, p$.
2. For $m = 1, 2, \cdots, p$,
   (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 $\mathbf{z}_m$:

$$\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - \left[\frac{\langle \mathbf{z}_m, \mathbf{x}_j^{(m-1)}\rangle}{\langle \mathbf{z}_m, \mathbf{z}_m\rangle}\right]\mathbf{z}_m, \quad j = 1, 2, \cdots, p$$

3. Output the sequence of fitted vectors $\{\hat{\mathbf{y}}^{(m)}\}_{m=1}^{p}$. Since $\{\mathbf{z}_l\}_{l=1}^{m}$ are linear in the original $\mathbf{z}_j$, so is $\hat{\mathbf{y}}^{(m)} = \mathbf{X}\hat{\beta}^{\text{pls}}(m)$. These linear coefficients can be recovered from the squence of PLS transformations.

The mth principal component direction v_m solves

$$\max_{\alpha} \text{Var}(\mathbf{X}\alpha)$$

$$\text{subject to } \|\alpha\| = 1,\ \alpha^T\mathbf{S}v_l = 0, \quad l = 1, 2, \cdots, m-1$$

where $\mathbf{S} = \mathbf{X}^T\mathbf{X}/N$ is the sample covariance matrix (it is N not $N-1$ since $\mathbf{X}$ is standardized, expectation is known to be 0). The condition $\alpha^T\mathbf{S}v_l = 0$ ensures that $\mathbf{z}_m = \mathbf{X}\alpha$ is uncorrelated with all previous linear combinations $\mathbf{z}_l = \mathbf{X}v_l$.

Instead of doing this, partial least squares is a compromise between ordinary regression coefficient and the

principal component directions.

Proposition 1.4.2: Partial Least Square Criterion

The mth PLS direction $\hat{\varphi}_m = (\hat{\varphi}_{m1}, \cdots, \hat{\varphi}_{mp})^T$ solves

$$\max_{\alpha} \text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha)$$

$$\text{subject to } \|\alpha\| = 1,\ \alpha^T \mathbf{S}\hat{\varphi}_l = 0, \quad l = 1, 2, \cdots, m-1$$

which seeks directions that have high variance and have high correlation with the response.

Proof. First note that

$$\text{Corr}^2(\mathbf{y}, \mathbf{X}\alpha)\text{Var}(\mathbf{X}\alpha) = \frac{\text{Cov}^2(\mathbf{y}, \mathbf{X}\alpha)}{\text{Var}(\mathbf{y})\text{Var}(\mathbf{X}\alpha)}\text{Var}(\mathbf{X}\alpha) = \frac{\text{Cov}^2(\mathbf{y}, \mathbf{X}\alpha)}{\text{Var}(\mathbf{y})}$$

where $\text{Var}(\mathbf{y})$ is a fixed value. Therefore, it is equivalent to maximize

$$\text{Cov}^2(\mathbf{y}, \mathbf{X}\alpha) = (\alpha^T \mathbf{X}^T \mathbf{y})^2$$

subject to those required constraints. Due to the constraint that α is orthogonal to all of the $\mathbf{S}\hat{\varphi}_1, \mathbf{S}\hat{\varphi}_2, \cdots, \mathbf{S}\hat{\varphi}_{m-1}$, we have $\alpha^T \text{proj}_{\text{span}(\mathbf{X}\hat{\phi}_1, \cdots, \mathbf{X}\hat{\phi}_{m-1})}\mathbf{X}^T = 0$, so it must be the case that

$$\begin{aligned}
(\alpha^T \mathbf{X}^T \mathbf{y})^2 &= \left(\alpha^T \left(\mathbf{X} - \text{proj}_{\text{span}(\mathbf{X}\hat{\phi}_1, \cdots, \mathbf{X}\hat{\phi}_{m-1})}\mathbf{X}\right)^T \mathbf{y}\right)^2 \\
&= \left(\alpha^T \left(\mathbf{X} - \text{proj}_{\text{span}(\mathbf{z}_1, \cdots, \mathbf{z}_{m-1})}\mathbf{X}\right)^T \mathbf{y}\right)^2 \\
&= \left(\alpha^T (\mathbf{X}^{(m-1)})^T \mathbf{y}\right)^2 = (\alpha^T \hat{\varphi}_m)^2
\end{aligned}$$

The direction which α maximize $(\alpha^T \hat{\varphi}_m)^2$ is just $\hat{\varphi}_m$. Since $\alpha = \hat{\varphi}_m$ satisfies the constraint $\alpha^T \mathbf{S}\hat{\varphi}_l = 0$, the direction $\hat{\varphi}_m$ is the mth PLS direction. □

If the input matrix $\mathbf{X}$ is orthogonal, then partial least squares finds the least squares estimate after $m = 1$ step. This can be seen by that

$$\hat{\theta}_m = \frac{\langle \mathbf{z}_1, \mathbf{y} \rangle}{\langle \mathbf{z}_1, \mathbf{z}_1 \rangle} = \frac{\left\langle \sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle \mathbf{x}_j, \mathbf{y} \right\rangle}{\left\langle \sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle \mathbf{x}_j, \sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle \mathbf{x}_j \right\rangle} = \frac{\sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle^2}{\sum_{j=1}^p \sum_{k=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle \langle \mathbf{x}_k, \mathbf{y} \rangle \langle \mathbf{x}_j, \mathbf{x}_k \rangle}$$

$$\xlongequal{\mathbf{x}_j \text{ are orthogonal}} \frac{\sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle^2}{\sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle^2 \|\mathbf{x}_j\|^2} \xlongequal{\mathbf{x}_j \text{ are standardized}} \frac{\sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle^2}{\sum_{j=1}^p \langle \mathbf{x}_j, \mathbf{y} \rangle^2} = 1 \tag{1.7}$$

Fantuan's Math Notes

Thus,

$$\hat{\mathbf{y}^{(1)}} = \bar{y}\mathbf{1} + \hat{\theta}_1\mathbf{z}_1 = \bar{y}\mathbf{1} + \sum_{j=1}^{p}\langle \mathbf{x}_j, \mathbf{y}\rangle \mathbf{x}_j \overset{\mathbf{x}_j \text{ are standardized}}{=\!=\!=} \bar{y}\mathbf{1} + \frac{\sum_{j=1}^{p}\langle \mathbf{x}_j, \mathbf{y}\rangle}{\langle \mathbf{x}_j, \mathbf{x}_j\rangle}\mathbf{x}_j$$

is just the least square fit under the condition of orthogonal input (Section 1.1.4). Subsequent steps have no effect since the $\hat{\varphi}_{mj}$ are zero for $m > 1$. This can be seen if we examine $\hat{\varphi}_{2k}$ for some $k = 1, 2, \cdots, p$,

$$\begin{aligned}
\hat{\varphi}_{2k} = \langle \mathbf{x}_k^{(1)}, \mathbf{y}\rangle &= \left\langle \mathbf{x}_k - \frac{\langle \mathbf{z}_1, \mathbf{x}_k\rangle}{\langle \mathbf{z}_1, \mathbf{z}_1\rangle}\mathbf{z}_1, \mathbf{y} \right\rangle = \langle \mathbf{x}_k, \mathbf{y}\rangle - \frac{\langle \mathbf{z}_1, \mathbf{x}_k\rangle}{\langle \mathbf{z}_1, \mathbf{z}_1\rangle}\langle \mathbf{z}_1, \mathbf{y}\rangle \\
&\overset{1.7}{=\!=} \langle \mathbf{x}_k, \mathbf{y}\rangle - \langle \mathbf{z}_1, \mathbf{x}_k\rangle = \langle \mathbf{x}_k, \mathbf{y}\rangle - \left\langle \sum_{j=1}^{p}\langle \mathbf{x}_j, \mathbf{y}\rangle \mathbf{x}_j, \mathbf{x}_k \right\rangle \\
&\overset{\mathbf{x}_j \text{ are orthogonal}}{=\!=\!=} \langle \mathbf{x}_k, \mathbf{y}\rangle - \langle \mathbf{x}_k, \mathbf{y}\rangle \langle \mathbf{x}_k, \mathbf{x}_k\rangle = 0
\end{aligned}$$

1.5 Multiple Outcomes

1.5.1 Multiple Outcome Regression

Suppose we have multiple outputs $Y_1, Y_2, \cdots, Y_K$ that we wish to predict from inputs $X_0, X_1, \cdots, X_p$.

Uncorrelated Error

We assume a linear model for each output

$$Y_k = \beta_{0k} + \sum_{j=1}^{p} X_j\beta_{jk} + \epsilon_k = f_k(X) + \epsilon_k$$

With N training cases we can write the model in matrix notation

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where $\mathbf{Y}$ is the $N \times K$ response matrix, with ikth entry y_{ik}, $\mathbf{X}$ is the $N \times (p+1)$ input matrix, $\mathbf{B}$ is the $(p+1) \times K$ matrix of parameters and $\mathbf{E}$ is the $N \times K$ matrix of errors. Then, generalized from Univariate Least Square,

$$\text{RSS}(\mathbf{B}) = \sum_{k=1}^{K}\sum_{i=1}^{N}(y_{ik} - f_k(x_i))^2 = \text{tr}\left[(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\right]$$

(trace: adding up k's). The least square then have exactly the same form as before

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Multiple outcomes do not affect one another's least squares estimates.

Fantuan's Math Notes

Correlated Error

If errors $\epsilon = (\epsilon_1, \cdots, \epsilon_K)$ is correlated with $\text{Cov}(\epsilon) = \mathbf{\Sigma}$, then the *multivariate weighted criterion*

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) = \sum_{i=1}^{N} (y_i - f(x_i))^T \mathbf{\Sigma}^{-1} (y_i - f(x_i))$$

However, here comes the surprising result, that this does not effect the solution of least squares estimation.

Proposition 1.5.1: Correlated Error Multiple Outcome Linear Regression

The correlation of errors does not effect the solution of multiple outcomes least square estimation. i.e., the solution to minimize

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) = \sum_{i=1}^{N} (y_i - f(x_i))^T \mathbf{\Sigma}^{-1} (y_i - f(x_i))$$

is still

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Proof. We need to minimize

$$\begin{aligned}
\text{RSS}(\mathbf{B}; \mathbf{\Sigma}) &= \sum_{i=1}^{N} \left(y_i - \mathbf{B}^T x_i\right)^T \mathbf{\Sigma}^{-1} \left(y_i - \mathbf{B}^T x_i\right) = \sum_{i=1}^{N} \text{tr}\left(\left(y_i - \mathbf{B}^T x_i\right)^T \mathbf{\Sigma}^{-1} \left(y_i - \mathbf{B}^T x_i\right)\right) \\
&= \sum_{i=1}^{N} \text{tr}\left(\mathbf{\Sigma}^{-1} \left(y_i - \mathbf{B}^T x_i\right) \left(y_i - \mathbf{B}^T x_i\right)^T\right) = \text{tr}\left(\mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB})\right) \\
&= \text{tr}\left((\mathbf{Y}^* - \mathbf{XB}^*)^T (\mathbf{Y}^* - \mathbf{XB}^*)\right)
\end{aligned}$$

where $\mathbf{Y}^* = \mathbf{Y}\mathbf{\Sigma}^{-1/2}$, and $\mathbf{B}^* = \mathbf{B}\mathbf{\Sigma}^{-1/2}$. The least square estimation is then

$$\hat{\mathbf{B}}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^*$$

Hence,

$$\hat{\mathbf{B}} = \hat{\mathbf{B}}^* \mathbf{\Sigma}^{1/2} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}^* \mathbf{\Sigma}^{1/2} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

which goes back to least square estimate. □

Correlation Varying among Observations

If the $\mathbf{\Sigma}_i$ vary among observations, then it is no longer the case, and the solution for $\mathbf{B}$ no longer decouples. We need to minimize

$$\text{RSS}(\mathbf{B}; \mathbf{\Sigma}_i) = \sum_{i=1}^{N} (y_i - \mathbf{B}^T x_i)^T \mathbf{\Sigma}_i^{-1} (y_i - \mathbf{B}^T x_i)$$

Fantuan's Math Notes

We need first derivative to be zero, which indicates the solution need to satisfy that

$$\sum_{i=1}^{N} x_i x_i^T \hat{\mathbf{B}} \mathbf{\Sigma}_i^{-1} = \sum_{i=1}^{N} x_i y_i^T \mathbf{\Sigma}_i^{-1}$$

1.5.2 Multiple Outcome Shrinkage and Selection

Find Directions: Canonical Correlation Analysis (CCA)

CCA finds a sequence of uncorrelated linear combination $\mathbf{X}v_m$, $m = 1, 2, \cdots, M$ of the $\mathbf{x}_j$, and a corresponding sequence of uncorrelated linear combinations $\mathbf{Y}u_m$ of the responses $\mathbf{y}_k$, such that the correlation

$$\text{Corr}^2(\mathbf{Y}u_m, \mathbf{X}v_m)$$

are successively maximized. At most $M = \min(K, p)$ directions can be found. The leading canonical response variates are those linear combinations best predicted by the $\mathbf{x}_j$, and vise versa for those trailing ones. The solution can be computed using a generalized SVD of sample cross-covariance matrix $\mathbf{Y}^T\mathbf{X}/N$, *assume* $\mathbf{X}$ *and* $\mathbf{Y}$ *are centered.*

Proposition 1.5.2: Canonical Correlation Analysis

The leading pair of canonical variates u_1 and v_1 solve the problem

$$\max_{\substack{u^T(\mathbf{Y}^T\mathbf{Y})u=1 \\ v^T(\mathbf{X}^T\mathbf{X})v=1}} u^T(\mathbf{Y}^T\mathbf{X})v$$

a generalized SVD problem. The solution is given by $u_m = (\mathbf{Y}^T\mathbf{Y})^{-1/2}u_m^*$ and $v_m = (\mathbf{X}^T\mathbf{X})^{-1/2}v_m^*$, where u_m^* and v_m^*, $m = 1, 2, \cdots, \min(K, p)$ are the mth left and right singular vectors in

$$(\mathbf{Y}^T\mathbf{Y})^{-1/2}(\mathbf{Y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1/2} = \mathbf{U}^*\mathbf{D}^*(\mathbf{V}^*)^T$$

Proof. 1. First, the correlation need to be maximized is given by

$$\text{Corr}^2(\mathbf{Y}u_m, \mathbf{X}v_m) = \frac{\text{Cov}^2(\mathbf{Y}u_m, \mathbf{X}v_m)}{\text{Var}(\mathbf{Y}u_m)\text{Var}(\mathbf{X}v_m)} = \frac{\left(u_m^T(\mathbf{Y}^T\mathbf{X})v_m\right)^2}{u_m^T(\mathbf{Y}^T\mathbf{Y})u_m v_m^T(\mathbf{X}^T\mathbf{X})v_m}$$

Under the condition that $v^T(\mathbf{X}^T\mathbf{X})v = 1$ and $u^T(\mathbf{Y}^T\mathbf{Y})u = 1$, the objective function is thus

$$\text{Corr}^2(\mathbf{Y}u_m, \mathbf{X}v_m) = \left(u_m^T(\mathbf{Y}^T\mathbf{X})v_m\right)^2$$

since the leading pair of canonical variates maximize the correlation, we know that they are indeed the solution to the generalized SVD problem.

Fantuan Math Notes

2. To find the solutions u_1 and v_1, we write the objective function in Lagrangian multiplier form

$$L(u, v, \lambda_1, \lambda_2) = u^T(\mathbf{Y}^T\mathbf{X})v - \frac{\lambda_1}{2}\left(u^T(\mathbf{Y}^T\mathbf{Y})u - 1\right) - \frac{\lambda_2}{2}\left(v^T(\mathbf{X}^T\mathbf{X})v - 1\right)$$

Taking derivatives and setting them to zero yield

$$\frac{\partial L}{\partial u} = (\mathbf{Y}^T\mathbf{X})v - \lambda_1(\mathbf{Y}^T\mathbf{Y})u = 0 \tag{1.8}$$

$$\frac{\partial L}{\partial v} = (\mathbf{X}^T\mathbf{Y})u - \lambda_2(\mathbf{X}^T\mathbf{X})u = 0 \tag{1.9}$$

Multiplying the first equation by u^T and the second by v^T, and noting that the constraints said $u^T(\mathbf{Y}^T\mathbf{Y})u = 1$ and $v^T(\mathbf{X}^T\mathbf{X})v = 1$, we have

$$u^T\mathbf{Y}^T\mathbf{X}v = \lambda_1 \tag{1.10}$$

$$v^T\mathbf{X}^T\mathbf{Y}u = \lambda_2 \tag{1.11}$$

We see that $\lambda_1 = \lambda_2 = u^T\mathbf{Y}^T\mathbf{X}v$, and we denote $\lambda = \lambda_1 = \lambda_2$. Denote

$$\mathbf{M} = (\mathbf{Y}^T\mathbf{Y})^{-1/2}(\mathbf{Y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1/2}$$

We need to find the relation between $\mathbf{M}$ and u, v. Recall that u_1^* and v_1^* are the leading left and right singular vectors of $\mathbf{M}$. Then, $u_1 = (\mathbf{Y}^T\mathbf{Y})^{-1/2}u_1^*$ and $v_1 = (\mathbf{X}^T\mathbf{X})^{-1/2}v_1^*$ solves the equations 1.10 and 1.11, since

$$u_1^T\mathbf{Y}^T\mathbf{X}v_1 = (u_1^*)^T(\mathbf{Y}^T\mathbf{Y})^{-1/2}(\mathbf{Y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1/2}v_1^* = (u_1^*)^T\mathbf{M}v_1^* = d_1^* = \lambda$$

where d_1^* is the first singular value, and similarly we can have the same result for 1.11.

3. Now for successive u_m and v_m, we need to consider the problem

$$\max_{\substack{u^T(\mathbf{Y}^T\mathbf{Y})u=1 \\ v^T(\mathbf{X}^T\mathbf{X})v=1 \\ u^Tu_j=0, \forall j<m \\ v^Tv_j=0, \forall j<m}} u^T(\mathbf{Y}^T\mathbf{X})v$$

The Lagrangian multiplier form now becomes

$$\begin{aligned} L(u, v, \lambda_1, \lambda_2) = {} & u^T(\mathbf{Y}^T\mathbf{X})v - \frac{\lambda_1}{2}\left(u^T(\mathbf{Y}^T\mathbf{Y})u - 1\right) - \frac{\lambda_2}{2}\left(v^T(\mathbf{X}^T\mathbf{X})v - 1\right) \\ & - \sum_{j<m}\alpha_j u^Tu_j - \sum_{j<m}\beta_j v^Tv_j \end{aligned}$$

Fantuan's Math Notes

where α_j's and β_j's are newly added Lagrangian multipliers. Again, take the derivatives, and we have

$$\frac{\partial L}{\partial u} = (\mathbf{Y}^T\mathbf{X})v - \lambda_1(\mathbf{Y}^T\mathbf{Y})u - \sum_{j<m}\alpha_j u_j = 0$$

$$\frac{\partial L}{\partial v} = (\mathbf{X}^T\mathbf{Y})u - \lambda_2(\mathbf{X}^T\mathbf{X})u - \sum_{j<m}\beta_j v_j = 0$$

We multiply first equation by u^T, and the second by v^T. Now we see that, since $u^T u_j = v^T v_j = 0$ by constraint, the newly added multiplier terms become

$$-u^T\sum_{j<m}\alpha_j u_j = 0, \quad \text{and} \quad -v^T\sum_{j<m}\beta_j v_j = 0$$

and it goes back to the original Lagrangian multiplier equations 1.8 and 1.9. This indicates that the entire sequence $u_m, v_m, m = 1, 2, \cdots, \min(K, p)$ is also given by

$$u_m = (\mathbf{Y}^T\mathbf{Y})^{-1/2}u_m^* \quad \text{and} \quad v_m = (\mathbf{X}^T\mathbf{X})^{-1/2}v_m^*$$

□

Dimension Reduction: Reduced-rank Regression

Reduced-rank Regression was introduced by Izenman (1975[5]). Given an error covariance $\text{Cov}(\epsilon) = \mathbf{\Sigma}$, we solve the following restricted multivariate regression problem:

$$\hat{\mathbf{B}}^{\text{rr}}(m) = \underset{\text{rank}(\mathbf{B})=m}{\text{argmin}} \sum_{i=1}^{N}(y_i - \mathbf{B}^T x_i)^T\mathbf{\Sigma}^{-1}(y_i - \mathbf{B}^T x_i)$$

With suitable estimation of $\mathbf{\Sigma}$, we can write closed form solution.

Proposition 1.5.3: Reduced Rank Regression

With $\mathbf{\Sigma}$ replaced by the estimate $\mathbf{Y}^T\mathbf{Y}/N$ or $(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/(N - pK)$, where $\hat{\mathbf{B}}$ is the least squares fit, the solution of reduced-rank regression problem is given by a CCA of $\mathbf{Y}$ and $\mathbf{X}$

$$\hat{\mathbf{B}}^{\text{rr}}(m) = \hat{\mathbf{B}}\mathbf{U}_m\mathbf{U}_m^-$$

where $\mathbf{U}_m$ is the $K \times m$ sub-matrix of $\mathbf{U}$ consisting of the first m columns, and $\mathbf{U}$ is the $K \times M$ matrix of left canonical vectors $u_1, u_2, \cdots, u_M$. $\mathbf{U}_m^-$ is its generalized inverse.

Proof. To solve the reduced rank regression problem, let $\mathbf{Y}^* = \mathbf{Y}\mathbf{\Sigma}^{-1/2}$ and $\mathbf{B}^* = \mathbf{B}\mathbf{\Sigma}^{-1/2}$, we first solve

Tian's Math Notes

when $\mathbf{\Sigma}$ is estimated by $\mathbf{Y}^T\mathbf{Y}/N$. Then it is equivalent to solve

$$\begin{aligned}
&\sum_{i=1}^{N}(y_i - \mathbf{B}^T x_i)^T\mathbf{\Sigma}^{-1}(y_i - \mathbf{B}^T x_i) = \text{tr}[(\mathbf{Y} - \mathbf{XB})\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})^T] \\
&\overset{\mathbf{\Sigma} \text{ is symmetric}}{=\!=\!=} \text{tr}\left[(\mathbf{Y}\mathbf{\Sigma}^{-1/2} - \mathbf{XB}\mathbf{\Sigma}^{-1/2})(\mathbf{Y}\mathbf{\Sigma}^{-1/2} - \mathbf{XB}\mathbf{\Sigma}^{-1/2})^T\right] \\
&= \text{tr}\left[(\mathbf{Y}^* - \mathbf{XB}^*)(\mathbf{Y}^* - \mathbf{XB}^*)^T\right] = \text{tr}\left[(\mathbf{Y}^* - \mathbf{XB}^*)^T(\mathbf{Y}^* - \mathbf{XB}^*)\right] \\
&= \|\mathbf{Y}^* - \mathbf{XB}^*\|_F^2
\end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius Norm for matrix. We need to minimize this quantity with the constraint that rank of $\mathbf{B}$ is m. Now let $\hat{\mathbf{B}}^*$ be the least square estimate of $\mathbf{Y}^*$, that is, $\hat{\mathbf{B}}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}^*$, then

$$\begin{aligned}
\|\mathbf{Y}^* - \mathbf{XB}^*\|_F^2 &= \|\mathbf{Y}^* - \mathbf{X}\hat{\mathbf{B}}^* + \mathbf{X}\hat{\mathbf{B}}^* - \mathbf{XB}^*\|_F^2 \\
&= \|\mathbf{Y}^* - \mathbf{X}\hat{\mathbf{B}}^*\|_F^2 + \|\mathbf{X}\hat{\mathbf{B}}^* - \mathbf{XB}^*\|_F^2
\end{aligned}$$

where the middle term is gone since residuals are orthogonal to the input matrix. Now we see that first term is not dependent on $\mathbf{B}$, and matrix $\mathbf{B}$ and $\mathbf{B}^*$ has the same rank (the rank of a matrix does not change when we multiply it by a full-rank matrix), so the minimization problem becomes

$$\begin{aligned}
&\underset{\text{rank}(\mathbf{B}^*)=m}{\text{argmin}} \|\mathbf{X}\hat{\mathbf{B}}^* - \mathbf{XB}^*\|_F^2 \\
&= \underset{\text{rank}(\mathbf{B}^*)=m}{\text{argmin}} \text{tr}\left[(\hat{\mathbf{B}}^{*T}\mathbf{X}^T - \mathbf{B}^{*T}\mathbf{X}^T)(\mathbf{X}\hat{\mathbf{B}}^* - \mathbf{XB}^*)\right] \\
&= \underset{\text{rank}(\mathbf{B}^*)=m}{\text{argmin}} n\,\text{tr}\left[(\hat{\mathbf{B}}^{*T} - \mathbf{B}^{*T})\frac{\mathbf{X}^T\mathbf{X}}{n}(\hat{\mathbf{B}}^* - \mathbf{B}^*)\right] \\
&= \underset{\text{rank}(\mathbf{B}^*)=m}{\text{argmin}} \text{tr}\left[(\hat{\mathbf{B}}^{*T} - \mathbf{B}^{*T})\mathbf{\Sigma}_X(\hat{\mathbf{B}}^* - \mathbf{B}^*)\right]
\end{aligned}$$

where $\mathbf{\Sigma}_X = (\mathbf{X}^T\mathbf{X})/n$ the covariance matrix estimation of $\mathbf{X}$. If we let $\hat{\mathbf{B}}^{**} = \mathbf{\Sigma}_X^{1/2}\hat{\mathbf{B}}^*$, and $\mathbf{B}^{**} = \mathbf{\Sigma}_X^{1/2}\mathbf{B}^*$, since $\mathbf{B}^{**}$ and $\mathbf{B}^*$ still have the same rank, we have

$$\begin{aligned}
&\underset{\text{rank}(\mathbf{B}^*)=m}{\text{argmin}} \text{tr}\left[(\hat{\mathbf{B}}^{*T}\mathbf{\Sigma}_X^{1/2} - \mathbf{B}^{*T}\mathbf{\Sigma}_X^{1/2})(\mathbf{\Sigma}_X^{1/2}\hat{\mathbf{B}}^* - \mathbf{\Sigma}_X^{1/2}\mathbf{B}^*)\right] \\
&= \underset{\text{rank}(\mathbf{B}^{**})=m}{\text{argmin}} \|\hat{\mathbf{B}}^{**} - \mathbf{B}^{**}\|_F^2 = \underset{\text{rank}(\mathbf{B}^{**})=m}{\text{argmin}} \|\hat{\mathbf{B}}^{**T} - \mathbf{B}^{**T}\|_F^2
\end{aligned}$$

where the last equation holds since the Frobenius norm of a matrix is the same with its transpose. This is then a **reduced-rank approximation problem**. By **Eckart–Young–Mirsky theorem**, the solution is the first m set of SVD of $\hat{\mathbf{B}}^{**}$, that is, if $\hat{\mathbf{B}}^{**} = \mathbf{V}^*\mathbf{D}^*\mathbf{U}^{*T}$, the solution to $\mathbf{B}^{**}$ is

$$\mathbf{B}^{**}(m) = \mathbf{V}_m^*\mathbf{D}^*\mathbf{U}_m^{*T}$$

Fantuan's Math Notes

the first m set of SVD. The reson why we write the singular value decomposition as 'VDU' rather than 'UDV' is due to the structure of matrix $\hat{\mathbf{B}}^{**}$,

$$\begin{aligned}\hat{\mathbf{B}}^{**} &= \mathbf{\Sigma}_X^{1/2}\hat{\mathbf{B}}^{*} = \mathbf{\Sigma}_X^{1/2}\hat{\mathbf{B}}\mathbf{\Sigma}^{-1/2} \\ &= \mathbf{\Sigma}_X^{1/2}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}_X^{-1/2}\left(\mathbf{X}^T\mathbf{Y}\right)\mathbf{\Sigma}^{-1/2}\end{aligned}$$

which is the transpose of our canonical matrix $(\mathbf{Y}^T\mathbf{Y})^{-1/2}(\mathbf{Y}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1/2}$, so the left canonical variates are actually right canonical variates in this case. We can see from the second equation that the ordinary least square estimation can be represented as

$$\hat{\mathbf{B}} = \mathbf{\Sigma}_X^{-1/2}\hat{\mathbf{B}}^{**}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}_X^{-1/2}\mathbf{V}^*\mathbf{D}^*\mathbf{U}^{*T}\mathbf{\Sigma}^{1/2}$$

Since $\mathbf{B}^{**}(m) = \mathbf{\Sigma}_X^{1/2}\hat{\mathbf{B}}^{\mathrm{rr}}(m)\mathbf{\Sigma}^{-1/2}$, the final solution can be represented as

$$\begin{aligned}\hat{\mathbf{B}}^{\mathrm{rr}}(m) &= \mathbf{\Sigma}_X^{-1/2}\mathbf{V}_m^*\mathbf{D}^*\mathbf{U}_m^{*T}\mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}_X^{-1/2}\mathbf{V}^*\mathbf{D}^*\mathbf{U}^{*T}\mathbf{U}_m^*\mathbf{U}_m^{*T}\mathbf{\Sigma}^{1/2} \\ &= \mathbf{\Sigma}_X^{-1/2}\mathbf{\Sigma}_X^{-1/2}(\mathbf{X}^T\mathbf{Y})\mathbf{\Sigma}^{-1/2}\mathbf{U}_m^*\mathbf{U}_m^{*T}\mathbf{\Sigma}^{1/2} \\ &= \mathbf{\Sigma}_X^{-1}\mathbf{X}^T\mathbf{Y}\left(\mathbf{\Sigma}^{-1/2}\mathbf{U}_m^*\right)\left(\mathbf{U}_m^{*T}\mathbf{\Sigma}^{1/2}\right) \\ &= \hat{\mathbf{B}}\mathbf{U}_m\mathbf{U}_m^-\end{aligned}$$

where $\mathbf{U}_m = \mathbf{\Sigma}^{-1/2}\mathbf{U}_m^*$ is the first m set of left canonical variates, and $\mathbf{U}_m^- = \mathbf{U}_m^{*T}\mathbf{\Sigma}^{1/2}$ is a generalized inverse (since $\mathbf{U}_m^*$ is orthogonal).

It can be also shown that the estimation $(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}})/(N-pK)$ also works. □

Writing the solution as

$$\hat{\mathbf{B}}^{\mathrm{rr}}(m) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}(\mathbf{Y}\mathbf{U}_m)\mathbf{U}_m^-$$

we see that reduced-rank regression performs a linear regression on the pooled response matrix $\mathbf{Y}\mathbf{U}_m$, and then maps the coefficients back to the original space. The fits are given by

$$\hat{\mathbf{Y}}^{\mathrm{rr}}(m) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}\mathbf{U}_m\mathbf{U}_m^- = \mathbf{H}\mathbf{Y}\mathbf{P}_m$$

where $\mathbf{H}$ is the hat matrix, and $\mathbf{P}_m$ is the rank-m CCA response projection operator.

Shrinkage of Reduced-rank Regression:

Breiman and Friedman (1997[1]) proposed some shrinkage in canonical variates. It has the form

$$\hat{\mathbf{B}}^{c+w} = \hat{\mathbf{B}}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$$

Fantuan's Math Notes

where $\mathbf{\Lambda}$ is a diagonal shrinkage matrix. Based on optimal prediction in the population setting, the diagonal entries are

$$\lambda_m = \frac{c_m^2}{c_m^2 + \frac{p}{N}(1 - c_m^2)}, m = 1, 2, \cdots, M$$

where c_m is the mth canonical correlation coefficient. Note that this converges to 1 when $p/N \to 0$. The fitted response is then $\hat{\mathbf{Y}}^{c+w} = \mathbf{HYS}^{c+w}$, where $\mathbf{S}^{c+w} = \mathbf{U\Lambda U}^{-1}$ is the response shrinkage operator.

They also suggested shrinking in both Y and X spaces. This leads to hybrid shrinkage method of the form

$$\hat{\mathbf{Y}}^{\text{ridge},c+w} = \mathbf{A}_\lambda \mathbf{YS}^{c+w}$$

where $\mathbf{A}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$ is the ridge shrinkage operator.

More linear methods for regression can be found in appendix A.

Fantuan's Math Notes

Chapter 2

Linear Methods for Classification

Fantuan's Math Notes

Fantuan's Math Notes

Appendices

Fantuan's Math Notes

Fantuan's Math Notes

Appendix A

Other Linear Methods for Regression

A.1 Forward Stagewise Regression (FS)

It starts like forward-stepwise regression, with an intercept equal to $\bar{y}$, and centered predictors with coefficients initially all 0. At each step the algorithm identifies the variable most correlated with the current residual. It then computes the simple linear regression coefficient of the residual on this chosen variable, and then adds it to the current coefficient for that variable. This is continued till none of the variables have correlation with the residuals—i.e. the least-squares fit when $N > p$.

A.2 Incremental Forward Stagewise Regression

Algorithm A.2.1: Incremental Forward Stagewise Regression FS_ϵ

1. Start with residual $\mathbf{r}$ equal to $\mathbf{y}$ and $\beta_1, \beta_2, \cdots, \beta_p = 0$. All predictors are standardized to have mean zero and unit norm.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Update $\beta_j \leftarrow \beta_j + \delta_j$, where $\delta_j = \epsilon \cdot \text{sign}[\langle \mathbf{x}_j, \mathbf{r} \rangle]$ and $\epsilon > 0$ is a small step size. Set $\mathbf{r} \leftarrow \mathbf{r} - \delta_j \mathbf{x}_j$.
4. Repeat steps 2 and 3 many times, until the residuals are uncorrelated with all the predictors.

If $\delta_j = \langle \mathbf{x}_j, \mathbf{r} \rangle / \langle \mathbf{x}_j, \mathbf{x}_j \rangle = \langle \mathbf{x}_j, \mathbf{r} \rangle$ (the least-squares coefficient of the residual on jth predictor), then this is exactly the usual forward stagewise procedure (FS).

Fall '17 Math Notes

Fantuan's Math Notes

Bibliography

[1] Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(1):3–54.

[2] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression.

[3] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[4] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

[5] Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264.

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

[7] Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142.

[8] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Fantuan's Math Notes