Chapter 3

Linear Methods for Regression

Overview

Linear regression finds β that minimizes $||y - X\beta||^2$. Put it another way, linear regression tries to find, among all points in the column space of X, the point that has the smallest distance to y. Resorting to Hilbert project theorem, we have $X^T(y - X\hat{\beta}) = 0$, and thus $\hat{\beta} = (X^TX)^{-1}X^Ty$.

to be added

Problem 3.1

The Z-score for the null hypothesis that a coefficient $\beta_j = 0$ is $z_j = \hat{\beta}_j / \sqrt{\widehat{\text{Var}}\left(\hat{\beta}_j\right)}$. It suffices to show that the F-score of dropping the single coefficient β_j from the model is $\hat{\beta}_j^2 / \widehat{\text{Var}}\left(\hat{\beta}_j\right)$.

After applying the Gram-Schmidt procedure sequentially on $x_0, x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p, x_j$, we obtain z_j, which is the component of x_j this is orthogonal to the rest of the feature vectors. Since z_j alone involves x_j, we know that $\langle z_j, y\rangle / \langle z_j, z_j\rangle$ is the multiple regression coefficient of y on x_j, i.e., $\hat{\beta}_j = \langle z_j, y\rangle / \langle z_j, z_j\rangle$.

Then, we can obtain

$$\text{Var}\left(\hat{\beta}_j\right) = \text{Var}\left(\langle z_j, y\rangle / \langle z_j, z_j\rangle\right) = \sigma^2/||z_j||^2. \tag{3.1}$$

Denote the residual sum of square with all the features as RSS_1 and that with all the features except x_j as RSS_0, then it is easy to see that

$$\text{RSS}_0 - \text{RSS}_1 = \hat{\beta}_j^2 ||z_j||^2,$$

By combining the above two equations, we have

$$\begin{aligned} \text{F-score} =& \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1/(N-p-1)} \\ =& \frac{\hat{\beta}_j^2 ||z_j||^2}{\hat{\sigma}^2} = \frac{\hat{\beta}_j^2}{\hat{\sigma}^2/||z_j||^2} = \frac{\hat{\beta}_j^2}{\widehat{\text{Var}}\left(\hat{\beta}_j\right)}, \end{aligned}$$

which completes the proof.

Problem 3.2

In the first approach, for any $f(a) = a^T\beta$, the confidence interval is

$$\left[a^T\beta - \sqrt{\text{Var}\left(a^T\beta\right)}z^{1-\alpha/2}, a^T\beta + \sqrt{\text{Var}\left(a^T\beta\right)}z^{1-\alpha/2}\right],$$

where

$$\text{Var}\left(a^T\beta\right) = \mathbb{E}\left[a^T\beta\beta^T a\right] = a^T\text{Var}\left(\beta\right)a = a^T(X^TX)^{-1}a\sigma^2.$$

In the second approach, for any $f(a) = a^T\beta$, the confidence band is

$$\left[\min\left\{a^T\beta|\beta\in C_\beta^{1-\alpha}\right\}, \max\left\{a^T\beta|\beta\in C_\beta^{1-\alpha}\right\}\right],$$

where

$$C_\beta^{1-\alpha} = \left\{\beta|(\hat\beta-\beta)^TX^TX(\hat\beta-\beta)\le\hat\sigma^2{\chi^2_{p+1}}^{(1-\alpha)}\right\}.$$

The first approach yields *pointwise confidence band*: the confidence interval is defined for each input a, meaning that if we focus on each input a, the confidence band gives us the range that corresponds to $1-\alpha$ confidence. The second approach yields *simultaneous confidence band*: the $1-\alpha$ confidence holds simultaneously for all inputs. The simultaneous confidence band is wider than pointwise counterpart given its global nature.

Problem 3.3

(a) Let c^Ty be a linear unbiased estimation of $a^T\beta$. Then we know that

$$\begin{aligned}&\mathbb{E}\left[c^Ty\right] = a^T\beta\\ \Rightarrow & c^TX\beta = a^T\beta\\ \Rightarrow & X^Tc = a.\end{aligned}$$

Since X^T is a fat matrix, the above equation is an under-determined equation for c. The general form of the solution is $c = X(X^TX)^{-1}a + \mathcal{N}(X^T)$ where $\mathcal{N}(.)$ denote the right null space. Then we have

$$\begin{aligned}\text{Var}\left(c^Ty\right) =& \text{Var}\left(a^T(X^TX)^{-1}X^Tn + \mathcal{N}(X^T)^Tn\right)\\ =& \text{Var}\left(a^T(X^TX)^{-1}X^Tn\right) + \text{Var}\left(\mathcal{N}(X^T)^Tn\right) + 2\mathbb{E}\left[a^T(X^TX)^{-1}X^Tnn^T\mathcal{N}(X^T)\right]\\ =& \text{Var}\left(a^T(X^TX)^{-1}X^Tn\right) + \text{Var}\left(\mathcal{N}(X^T)^Tn\right)\\ \ge& \text{Var}\left(a^T(X^TX)^{-1}X^Tn\right).\end{aligned}$$

Note that the third equation holds only when the noise vector n is uncorrelated with equal variance for each element, which is exact the condition of Gauss-Markov theorem.

(b) Similar as the proof of (a), we can show that if Cy is an un-biased estimator of β, then $C = (X^TX)^{-1}X^T + D$ with $DX = 0$.

$$\begin{aligned}\text{Var}\left(((X^TX)^{-1}X^T + D)n\right) =& \text{Var}\left((X^TX)^{-1}X^Tn\right) + \text{Var}\left((Dn\right) + 2\mathbb{E}\left[(X^TX)^{-1}X^Tnn^TD\right]\\ =& \text{Var}\left((X^TX)^{-1}X^Tn\right) + \text{Var}\left((Dn\right) \succeq \text{Var}\left((X^TX)^{-1}X^Tn\right).\end{aligned}$$

Problem 3.4

If QR decomposition of the X is readily available, then

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^Ty = (R^TQ^TQR)^{-1}R^TQy = R^{-1}Qy.\\ \Rightarrow R\hat{\beta} &= Qy.\end{aligned}$$

Backward substitution can be used to calculate $\hat{\beta}$.

Problem 3.6

$$\begin{aligned}\mathbb{P}\left(\beta|y\right) =& \frac{\mathbb{P}\left(y|\beta\right)\mathbb{P}\left(\beta\right)}{\mathbb{P}\left(y\right)} = \frac{1}{\mathbb{P}\left(y\right)}\frac{1}{\sqrt{2\pi K_y}}e^{-\frac{1}{2}(y-X\beta)^TK_y^{-1}(y-X\beta)}\frac{1}{\sqrt{2\pi K_\beta}}e^{-\frac{1}{2}\beta^TK_\beta^{-1}\beta}\\ &\propto e^{-\frac{1}{2\sigma^2}\left(\beta^T(X^TX+\sigma^2/\tau^2I)\beta-2y^TX\beta\right)}\\ &\propto e^{-\frac{1}{2\sigma^2}(\beta-(X^TX+\sigma^2/\tau^2I)^{-1}X^Ty)^T(X^TX+\sigma^2/\tau^2I)(\beta-(X^TX+\sigma^2/\tau^2I)^{-1}X^Ty)}\\ &= \mathcal{N}\left((X^TX+\sigma^2/\tau^2I)^{-1}X^Ty, \sigma^2(X^TX+\sigma^2/\tau^2I)^{-1}\right).\end{aligned}$$

The last equation holds because the $\mathbb{P}\left(\beta|y\right)$ is a valid probability density distribution and should integrate to 1.

$$\mathbb{E}\left[\mathbb{P}\left(\beta|y\right)\right] = \max_\beta \mathbb{P}\left(\beta|y\right) = (X^TX+\sigma^2/\tau^2I)^{-1}X^Ty.$$

Problem 3.7

$$\mathbb{P}\left(\beta_{1\to N}|y;\beta_0\right) = \frac{\mathbb{P}\left(y|\beta_{1\to N};\beta_0\right)\mathbb{P}\left(\beta_{1\to N}\right)}{\mathbb{P}\left(y;\beta_0\right)}.$$

$$\log\left(\mathbb{P}\left(\beta_{1\to N}|y;\beta_0\right)\right) = C + \log\left(\mathbb{P}\left(y|\beta_{1\to N};\beta_0\right)\right) + \log\left(\mathbb{P}\left(\beta_{1\to N}\right)\right).$$

Since $\mathbb{P}\left(y|\beta_{1\to N}\right)$ follows $\mathcal{N}(\beta_0 + X\beta_{1\to N}, \sigma^2I)$, and $\mathbb{P}\left(\beta_{1\to N}\right)$ follows $\mathcal{N}(0, \tau^2I)$, we further have

$$\begin{aligned}\log\left(\mathbb{P}\left(\beta_{1\to N}|y;\beta_0\right)\right) =& C' + \frac{1}{2\sigma^2}(y-\beta_0-X\beta_{1\to N})^T(y-\beta_0-X\beta_{1\to N}) + \frac{1}{2\tau^2}\beta_{1\to N}^T\beta_{1\to N}\\ &\propto C'' + (y-\beta_0-X\beta_{1\to N})^T(y-\beta_0-X\beta_{1\to N}) + \frac{\sigma^2}{\tau^2}\beta_{1\to N}^T\beta_{1\to N}.\end{aligned}$$

Problem 3.8

$$\tilde{X}\tilde{X}^T = Q_2RR^TQ_2^T = U\Sigma V^TV\Sigma^TU^T = U\Sigma\Sigma^TU^T.$$

If U and Q_2 are the same up to sign flip (or phase rotation in case of complex space), then we need to have $RR^T = \Sigma\Sigma^T$. With R being an upper-triangular matrix, we can easily prove that RR^T is a diagonal matrix only if R is a diagonal matrix, which implies that the rows in $\tilde{X}$ are orthogonal to each other.

Problem 3.9

If we already have the QR decomposition for $N \times q$ matrix X_1 as QR, and already have the residual $r = (I - X(X^TX)^{-1}X^T)y$, then to establish which one of the features in an $p - q$ matrix X_2, we just need to find the column x in X_2 that has the largest value of $\langle r, x\rangle / \sqrt{\langle x, x\rangle}$.

Problem 3.10

Similar as that in the solution to Problem 3.1, let us denote z_j the component of x_j that is orthogonal to the rest of the features $x_0, x_1, \ldots, x_p$. We have,

$$\text{z-score}(\beta_j) = \frac{\beta_j}{\sqrt{\text{Var}\,(\beta_j)}} = \frac{\beta_j}{\sigma/||z_j||} = \frac{\beta_j||z_j||}{\sigma}.$$

Let us denote the residual sum of square with and without x_j as RSS and $\text{RSS}_{\sim j}$, then we have

$$\text{RSS}_{\sim j} - \text{RSS} = \beta_j^2||z_j||^2 = \text{z-score}^2(\beta_j)/\sigma.$$

Therefore, the variable with the least z-score is the best one to drop in terms of minimize the increment of the residual sum of square.

Problem 3.11

$$\begin{aligned}
\text{BSS}(B, \Sigma) &= \sum_{i=1}^{n}(y_i - B^Tx_i)^T\Sigma^{-1}(y_i - B^Tx_i) \\
&= \sum_{i=1}^{n}\left(y_i^T\Sigma^{-1}y_i + x_i^TB\Sigma^{-1}B^Tx_i - 2x_i^TB\Sigma^{-1}y_i\right) \\
\frac{\partial\text{BSS}(B, \Sigma)}{\partial B} &= \sum_{i=1}^{n}\left(2\Sigma^{-1}Bx_ix_i^T - 2\Sigma^{-1}y_ix_i^T\right) \\
&= \sum_{i=1}^{n}2\Sigma^{-1}\left(Bx_ix_i^T - y_ix_i^T\right) \\
&\overset{(a)}{=} 2\Sigma^{-1}\sum_{i=1}^{n}\left(Bx_ix_i^T - y_ix_i^T\right) = 0 \\
&\Rightarrow B^T = (X^TX)^{-1}X^TY
\end{aligned}$$

If Σ is sample dependent, then Equation (a) no longer holds.

Problem 3.12

$$\beta = (X^TX + \lambda I)^{-1}X^Ty.$$

Let

$$\overline{X} \triangleq \begin{bmatrix} X \\ \sqrt{\lambda} I \end{bmatrix} \text{ and } \overline{y} \triangleq \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

Then $\overline{X}^T\overline{X} = X^T X + \lambda I$, and $\beta = (\overline{X}^T\overline{X})^{-1}\overline{X}^T\overline{y}$.

Another way to look at it:

$$\begin{aligned} \text{RSS} &= (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\ &= (y - X\beta)^T(y - X\beta) + \\ &\quad \sqrt{\lambda}\beta^T\sqrt{\lambda}\beta \\ &= (y - X\beta)^T(y - X\beta) + \\ &\quad (0 - \sqrt{\lambda}\beta)^T(0 - \sqrt{\lambda}\beta) \\ &= (\overline{y} - \overline{X}\beta)^T(\overline{y} - \overline{X}\beta). \end{aligned}$$

Problem 3.13

$$\hat{\theta} = ((XV)^T XV)^{-1}(XV)^T y = (V^T X^T XV)^{-1} V^T X^T y = V^{-1}(X^T X)^{-1} X^T y.$$

$$\hat{\beta}^{\text{pcr}} \triangleq V\hat{\theta} = VV^{-1}(X^T X)^{-1} X^T y = (X^T X)^{-1} X^T y = \hat{\beta}^{\text{ls}}.$$

Problem 3.14

To prove that PLS stops at $m = 1$, it suffices to show that $\theta_1 z_1$ already captures the projection of y onto the space spanned by X. More precisely, we want to show

$$\langle y - \theta_1 z_1, X_j \rangle = 0 \text{ for } j = 1, 2, \ldots, p.$$

It is easy to show that with the columns of X being orthogonal to each other, $\theta_1 = 1$. Then, the above equation reduces to

$$\langle y - z_1, X_j \rangle = 0 \text{ for } j = 1, 2, \ldots, p,$$

which holds since $\langle z_1, X_j \rangle = \left\langle \sum_{j=1}^{p} \langle X_j, y \rangle X_j, X_j \right\rangle = \langle X_j, y \rangle$.

Problem 3.15

$$\max_{\alpha, ||\alpha||=1} \text{Var}\,(X\alpha) = \max_{\alpha} \frac{\alpha^T X^T X\alpha}{\alpha^T\alpha} = \max_{\alpha} \frac{\alpha^T V\Sigma^H\Sigma^H V\alpha}{\alpha^T\alpha}.$$

The above objective leads to a α that equals to the first column in V. With the additional constraint that $\alpha^T S v_l = 0, l = 1, \ldots, m-1$, α becomes the m^{th} right eigenvector of X, which is v_m.

$$\max_{\alpha,||\alpha||=1} \text{Corr}^2(y, X\alpha)\text{Var}\,(X\alpha) = \max_{\alpha,||\alpha||=1} \frac{\text{Cov}^2(y, X\alpha)}{\text{Var}\,(y)} = \max_{\alpha} \frac{\alpha^T X^T y y^T X\alpha}{\text{Var}\,(y)\,\alpha^T\alpha}.$$

Without any additional constraint, the solution to the problem above is $\alpha = X^T y/||X^T y||$. With the additional constraint that $\alpha^T S\phi_l = 0, l = 1, \ldots, m-1$, to obtain the solution, we first need to strip off from X the subspace that is spanned by $X\phi_l, l = 1, \ldots, m-1$, which is fulfilled by projecting each column of X into the null space spanned by $X\phi_l, l = 1, \ldots, m-1$ to obtain $\tilde{X}$, and then follow the same reasoning to obtain $\alpha = \tilde{X}^T y/||\tilde{X}^T y||$. The procedure described here is PLS in exact.

Problem 3.16

In the case when the columns in X are orthonormal, we know that multiple regression becomes single univariate regression and $\hat{\beta}_j = X_j^T y$.

The residual sum of squares can be expressed as

$$\text{RSS} = \left\langle y - X\hat{\beta}, y - X\hat{\beta} \right\rangle = y^T y - \hat{\beta}^T \hat{\beta} = y^T y - \sum_{j=1}^{p} |\hat{\beta}_j|^2.$$

For the best subset selection with size M, we want to pick only M features, or equivalently, drop $p - M$ features so that the residual sum of the squares increase the least. According to the RSS equation above, it is easy to see that the solution is to have the regression coefficient be $\hat{\beta}_j \cdot \mathbf{1}\left(|\beta_j| \geq \left|\hat{\beta}_{(M)}\right|\right)$, with $\hat{\beta}_{(M)}$ representing the M^{th} largest absolute value of the regression coefficients.

With Ridge regression, we have

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

The Lasso estimation can be expressed as

$$\beta^{\text{lasso}} = \arg\min_{\beta} = y^T y + \sum_{j=1}^{p} \beta_j^2 - 2\sum_{j=1}^{p} \beta_j \beta_j^{\text{ls}} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Assume $\beta_j^{\text{lasso}} > 0$, then $\beta_j^{\text{lasso}} = \frac{2\beta_j^{\text{ls}} - \lambda}{2} = \beta_j^{\text{ls}} - \lambda/2$.

Assume $\beta_j^{\text{lasso}} < 0$, then $\beta_j^{\text{lasso}} = \frac{2\beta_j^{\text{ls}} + \lambda}{2} = \beta_j^{\text{ls}} + \lambda/2$.

Therefore $\beta_j^{\text{lasso}} = \text{sign}\left(\beta_j^{\text{ls}}\right)\left(\left|\beta_j^{\text{ls}}\right| - \frac{\lambda}{2}\right)^{+}$.

Problem 3.19

$$\beta^{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y = (V^T)^{-1}(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T U^T y$$

$$||\beta^{\text{ridge}}|| = ||(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T \tilde{y}|| \text{ shrinks as the increase of } \lambda.$$

Problem 3.20

$$\begin{aligned} & \max_{\substack{u^TY^TYu=1 \\ v^TX^TXv=1}} \quad u^TY^TXv \\ \binom{\tilde{u}=(Y^TY)^{1/2}u}{\tilde{v}=(X^TX)^{1/2}v} &= \max_{\substack{\tilde{u}^T\tilde{u}=1 \\ \tilde{v}^T\tilde{v}=1}} \tilde{u}^T(Y^TY)^{-1/2}R_{YX}(X^TX)^{-1/2}\tilde{v} \\ &= \max_{\substack{\|\tilde{u}\|=1 \\ \|\tilde{v}\|=1}} \tilde{u}^TU^*D^*V^{*T}\tilde{v}. \end{aligned}$$

The solution follows by noting that U^* and V^* are unitary matrices and D^* is a diagonal matrix with nonnegative real entries.

Problem 3.23

(a)

$$\left\langle x_j, \alpha X\hat{\beta}\right\rangle = \langle x_j, \alpha X(X^TX)^{-1}X^Ty\rangle = \alpha\left\langle x_j, y\right\rangle,$$

and thus

$$\frac{1}{N}\left|\left\langle x_j, y-\alpha X\hat{\beta}\right\rangle\right| = \frac{1}{N}\left|\langle x_j, y\rangle - \alpha\left\langle x_j, y\right\rangle\right| = (1-\alpha)\lambda,$$

and hence the correlations of each x_j with the residuals remains equal in magnitude as we progress towards u.

(b) to be finished

Problem 3.27

(a) Given the setting that $\beta_j = \beta_j^+ - \beta_j^-$ with β_j^+,$\beta_j^- \geq 0$, we have $|\beta_j| \leq \beta_j^+ + \beta_j^-$, and thus

$$\begin{aligned} & \min_{\beta} L(\beta) + \lambda\sum_j |\beta_j| \\ = & \min_{\beta,\beta^+,\beta^-} L(\beta) + \lambda\sum_j(\beta_j^+ + \beta_j^-) \\ & \text{subject to } \beta_j = \beta_j^+ - \beta_j^-, \beta_j^+ \geq 0, \beta_j^- \geq 0 \end{aligned}$$

The lagrangian function of the above problem is

$$\mathcal{L}(\beta^+,\beta^-,\lambda^+,\lambda^-) = L(\beta_j^+ - \beta_j^-) + \lambda\sum_j(\beta_j^+ + \beta_j^-) - \sum_j \lambda_j^+\beta_j^+ - \sum_j \lambda_j^-\beta_j^-.$$

KKT condition implies the following

$$\begin{aligned}\text{stationarity: } & \nabla L(\beta)_j \frac{\partial \beta_j}{\partial \beta_j^+} + \lambda - \lambda_j^+ = 0\\ & \nabla L(\beta)_j \frac{\partial \beta_j}{\partial \beta_j^-} + \lambda - \lambda_j^- = 0\\ \text{primal feasibility: } & \beta_j^+ \geq 0, \beta_j^- \geq 0 \text{ for all } j\\ \text{dual feasibility: } & \lambda_j^+ \geq 0, \lambda_j^- \geq 0 \text{ for all } j\\ \text{complementary slackness: } & \lambda_j^+\beta_j^+ = 0, \lambda_j^-\beta_j^- = 0 \text{ for all } j\end{aligned}$$

(b) By combining the stationarity condition and the dual feasibility condition, we have $|\nabla L(\beta)_j| \leq \lambda, \forall j$, which implies that $\nabla L(\beta)_j = 0 \forall j$ if $\lambda = 0$.

If $\beta_j^+ > 0$ and $\lambda > 0$, then from complementary slackness, we know that $\lambda_j^+ = 0$ and subsequenty $\nabla L(\beta)_j = -\lambda$, $\lambda_j^- > 0$, and $\beta_j^- = 0$.

If $\beta_j^- > 0$ and $\lambda > 0$, then from complementary slackness, we know that $\lambda_j^- = 0$ and subsequenty $\nabla L(\beta)_j = \lambda$, $\lambda_j^+ > 0$, and $\beta_j^+ = 0$.

Then it is easier to say for any “active” predicator having $\beta_j \neq 0$, we must have $\nabla L(\beta)_j = -\lambda$ if $\beta_j > 0$ and $\nabla L(\beta)_j = \lambda$ if $\beta_j < 0$.

$$\begin{aligned}\nabla L(\beta)_j =& \frac{\partial L(\beta)}{\partial \beta_j} = \frac{\partial \sum_{n=1}^N \left(y - \sum_{i=1}^P x_{ni}\beta_i\right)^2}{\partial \beta_j}\\ =& \sum_{n=1}^N 2\left(y - \sum_{i=1}^P x_{ni}\beta_i\right) x_{nj} = 2\left\langle y - X\beta, X_j\right\rangle\\ =& 2 \times \text{ the correlation between the } j\text{th predictor and the current residual.}\end{aligned}$$

This says that if the coefficient in the L1 normal regularization term is λ, then all the predictors that have non-zero regression coefficient will have its correlation with the residual equal to $\pm\lambda/2$.

(c) Suppose that the set of active predictors is unchanged from $\lambda_0 \geq \lambda \geq \lambda_1$, and let us denote the set of activate predictor as J, then from the result in (b) we know that

$$\begin{aligned}2\left\langle y - X\beta^{\lambda_0}, X_j\right\rangle = \lambda_0, \forall j \in J\\ 2\left\langle y - X\beta^{\lambda_1}, X_j\right\rangle = \lambda_1, \forall j \in J,\end{aligned}$$

where β^{λ_0} and β^{λ_1} are the optimal solution with $\lambda = \lambda_0$ and $\lambda = \lambda_1$, respectively. The convex combination between the two becomes

$$\begin{aligned}2\alpha\left\langle y - X\beta^{\lambda_0}, X_j\right\rangle + 2(1-\alpha)\left\langle y - X\beta^{\lambda_1}, X_j\right\rangle = \alpha\lambda_0 + (1-\alpha)\lambda_1, \forall j \in J\\ \Rightarrow 2\left\langle y - X\left(\alpha\beta^{\lambda_0} + (1-\alpha)\beta^{\lambda_1}\right), X_j\right\rangle = \alpha\lambda_0 + (1-\alpha)\lambda_1, \forall j \in J.\end{aligned}$$

This means that $\alpha\beta^{\lambda_0} + (1-\alpha)\beta^{\lambda_1}$ is the solution to the regression problem with the L1 regularization coefficient of $\alpha\lambda_0 + (1-\alpha)\lambda_1$.

For any $\lambda_0 \geq \lambda \geq \lambda_1$, we have

$$\begin{aligned}\lambda =& \alpha\lambda_0 + (1-\alpha)\lambda_1 = \lambda_0 + (1-\alpha)\left(\lambda_1 - \lambda_0\right)\\ \beta^{\lambda} =& \alpha\beta^{\lambda_0} + (1-\alpha)\beta^{\lambda_1} = \beta^{\lambda_0} + (1-\alpha)\left(\beta^{\lambda_1} - \beta^{\lambda_0}\right)\\ \Rightarrow \beta^{\lambda} =& \beta^{\lambda_0} - (\lambda - \lambda_0)\frac{\beta^{\lambda_1} - \beta^{\lambda_0}}{\lambda_0 - \lambda_1}.\end{aligned}$$

Problem 3.28

It is easy to prove by contradiction that $\hat{\beta}_j$ and $\hat{\beta}_j^*$ should satisfy $\hat{\beta}_j + \hat{\beta}_j^* = a$ and $\hat{\beta}_j\hat{\beta}_j^* \geq 0$.

Problem 3.29

$$a = (X^TX + \lambda)^{-1}X^Ty.$$

Let us denote $x \triangleq X^TX$ and $z \triangleq X^Ty$, then the above equation reduces to $a = \frac{z}{x+\lambda}$. Let us further denote $X_m \triangleq [X, X, \ldots, X]_{N\times m}$ and $\mathbf{1}_{m\times 1} \triangleq [1, 1, \ldots, 1]^T$, then the coefficients after refitting the ridge regression become

$$\begin{aligned}A =& (X_m^TX_m + \lambda I_m)^{-1}X_m^Ty\\ =& (x\mathbf{1}_{m\times 1}\mathbf{1}_{m\times 1}^T + \lambda I_m)^{-1}z\mathbf{1}_{m\times 1}\\ \left(\begin{matrix}\text{Sherman-Morrison}\\ \text{formula}\end{matrix}\right) =& \left(\frac{1}{\lambda}I_m - \frac{\frac{x}{\lambda^2}\mathbf{1}_{m\times 1}\mathbf{1}_{m\times 1}^T}{1 + \frac{mx}{\lambda}}\right)z\mathbf{1}_{m\times 1}\\ =& \left(\frac{z}{\lambda} - \frac{mxz/\lambda}{1 + mx/\lambda}\right)\mathbf{1}_{m\times 1}\\ =& \frac{z}{\lambda + mx}\mathbf{1}_{m\times 1} = a\frac{\lambda + x}{\lambda + mx}\mathbf{1}_{m\times 1}.\end{aligned}$$

In the case when $x = X^TX = 1$, we have $A = a\frac{\lambda+1}{\lambda+m}\mathbf{1}_{m\times 1}$.

Problem 3.30

Similarly as Problem 3.12, we can rewrite the regularized loss function as

$$\begin{aligned}& ||y - X\beta||^2 + \lambda[\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1]\\ =& (y - X\beta)^T(y - X\beta) + (0 - \sqrt{\lambda\alpha}\beta)^T(0 - \sqrt{\lambda\alpha}\beta) + \lambda(1-\alpha)||\beta||_1\\ =& (\overline{y} - \overline{X}\beta)^T(\overline{y} - \overline{X}\beta) + \lambda(1-\alpha)||\beta||_1,\end{aligned}$$

where $\overline{y} = [y^T, 0_p]^T$ and $\overline{X} = \left[X^T, \sqrt{\alpha\lambda}I_{p\times p}\right]^T$. Since the regularization term only has absolute-value norm of β, it leads to a lasso problem.

Unfinished Problems

Problem 3.5

- Problem 3.15
- Problem 3.21
- Problem 3.22
- Problem 3.23
- Problem 3.24
- Problem 3.25
- Problem 3.26