Ex 3.23

Ex. 3.23 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:

$$\frac{1}{N}|\langle \mathbf{x}_j, \mathbf{y}\rangle| = \lambda, \ j = 1, \ldots, p.$$

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 squares fit.

(a) Show that

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

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

(b) Show that these correlations are all equal to

$$\lambda(\alpha) = \frac{(1-\alpha)}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N}\cdot RSS}}\cdot\lambda,$$

and hence they decrease monotonically to zero.

(c) Use these results to show that the LAR algorithm in Section 3.4.4 keeps the correlations tied and monotonically decreasing, as claimed in (3.55).

PAPERTALK

Ex 3.23 (a) LSE: $SSR(\hat\beta) = \|Y - X\hat\beta\|^2 = (Y-X\hat\beta)^T(Y-X\hat\beta)$

$\frac{\partial SSR}{\partial \beta} = -2X^T(Y-X\hat\beta) = 0 \Rightarrow X^T(Y-X\hat\beta) = 0$ $\quad \underset{(p\times N)}{X^T} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_p^T \end{bmatrix}$ $\quad X_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(N)} \end{bmatrix}$

Which means $X_j^T(Y-X\hat\beta) = 0$ $\quad j = 1, 2, \ldots, p$

$\Rightarrow \langle X_j^T, \underline{Y - X\hat\beta}\rangle = 0$

$\frac{1}{N}\langle X_j^T, Y - \alpha X\hat\beta\rangle = \frac{1}{N}\langle X_j^T, \alpha Y + (1-\alpha)Y - \alpha X\hat\beta\rangle = \langle X_j^T, \alpha(Y-X\hat\beta) + (1-\alpha)Y\rangle$

$= \frac{1}{N}\langle X_j^T, \alpha(Y-X\hat\beta)\rangle + \frac{1}{N}\langle X_j^T, (1-\alpha)Y\rangle$

$= \frac{1}{N}\langle X_j^T, (1-\alpha)Y\rangle = (1-\alpha)\cdot\frac{1}{N}\langle X_j^T, Y\rangle$

$= (1-\alpha)\lambda$

(b) Correlation:

$$\frac{\langle X_j, Y - \mu(\alpha)\rangle / N}{\sqrt{\frac{\langle X_j, X_j\rangle}{N}}\cdot\sqrt{\frac{\langle Y-\mu(\alpha), Y-\mu(\alpha)\rangle}{N}}} \quad (b.1)$$

HY $\quad H^T = H$

$(Y^T X\hat\beta)^T = Y^T H^T Y = Y^T H Y$

$\mu(\alpha) = \alpha X\hat\beta$

$\cancel{Y^T(Y-\alpha X\hat\beta)}$ $\quad \alpha\hat\beta X^T(\alpha X\hat\beta - Y)$

where $\langle Y-\mu(\alpha), Y-\mu(\alpha)\rangle = Y^TY - \alpha Y^TX\hat\beta - \underline{\alpha\hat\beta^TX^TY + \alpha^2\hat\beta^TX^TX\hat\beta}$

As we know from (a) $X^T(Y-X\hat\beta) = 0$ $\quad X^TY = X^TX\hat\beta$

$\langle Y-\mu(\alpha), Y-\mu(\alpha)\rangle = Y^TY - \alpha Y^TX\hat\beta - \alpha\hat\beta^TX^TY + \alpha^2\hat\beta^TX^TY$

$= Y^TY - 2\alpha Y^TX\hat\beta + \alpha^2 Y^TX\hat\beta$

HY

$= Y^TY + \alpha(\alpha-2)Y^TX\hat\beta \quad (b.2)$

$RSS = \|Y - X\hat\beta\|^2 = [(I-H)Y]^T(I-H)Y = Y^T(I-H)Y = Y^TY - Y^TX\hat\beta$

$\Rightarrow Y^TX\hat\beta = Y^TY - RSS$

Plug in (b.2) $\Rightarrow Y^TY + \alpha(\alpha-2)(Y^TY - RSS) = (1-\alpha)^2Y^TY + \alpha(2-\alpha)RSS$

Since $y \sim N(0,1)$ $\quad \frac{1}{N}Y^TY = 1$

$\Rightarrow$ plug in (b.1) $\Rightarrow \dfrac{(1-\alpha)\lambda}{\sqrt{\underbrace{\frac{1}{N}Y^TY}_{=1}(1-\alpha)^2 + \frac{1}{N}\alpha(2-\alpha)RSS}}$ proved.

(c) From LAR Algorithm,

the moving direction is $\underline{\delta_k = (X_{A_k}^TX_{A_k})^{-1}X_{A_k}^TY_k}$

$\downarrow$

$X_{A_k}\delta_k = HY_k = X_k\hat\beta$

Then $\mu(\alpha) = \alpha X_k\hat\beta = \alpha X_{A_k}\delta_k$

keep tied?

monotonically decreasing: from (b), as α increases, correlation decreases

Ex. 3.24

Ex. 3.24 *LAR directions.* Using the notation around equation (3.55) on page 74, show that the LAR direction makes an equal angle with each of the predictors in $\mathcal{A}_k$.

Ex. 3.24 $\quad \delta_k = (X_{A_k}^TX_{A_k})^{-1}X_{A_k}^Tr_k \quad (3.55)$

LAR direction: $X_{A_k}\delta_k = \mu_k = X_{A_k}(X_{A_k}^TX_{A_k})^{-1}X_{A_k}^Tr_k$

since $\lambda = \frac{\langle X_i, r_k\rangle}{\|X_i\|\cdot\|r_k\|}$ where $X_i \in A_k$

absolute value since the coefficients can be negative $\quad \alpha X_i + (1-\alpha)X_i$

Then $\dfrac{|\langle X_i, r_k - \alpha\mu_k\rangle|}{\|X_i\|\|r_k - \alpha\mu_k\|} = \dfrac{1}{\|r_k - \alpha\mu_k\|}\cdot\dfrac{|\langle X_i, (1-\alpha)r_k + \alpha(r_k-\mu_k)\rangle|}{\|X_i\|}$

$= \dfrac{1}{\|r_k-\alpha\mu_k\|}\cdot\dfrac{|\alpha\langle X_i, r_k-\mu_k\rangle + (1-\alpha)\langle X_i, r_k\rangle|}{\|X_i\|}$

$= \dfrac{(1-\alpha)\lambda\|r_k\|}{\|r_k - \alpha\mu_k\|}$

Since $\langle X_i, r_k - \mu_k\rangle = 0$ as the covariates are orthogonal to the residual vector.

Proved.