
Solution 1

(a) If B is invertible, then $B^{-1/2}$ exists and the identity is immediate, so if $C = B^{-1/2}AB^{-1/2}$ and $(A+\alpha B)^{-1}Av = \eta v$, where $v \neq 0$, then

$$B^{-1/2}(C+\alpha I)^{-1}B^{-1/2}Av = \eta v \implies (C+\alpha I)^{-1}CB^{1/2}v = \eta B^{1/2}v$$

so η is an eigenvalue of $(C+\alpha I)^{-1}C$ with eigenvector $w = B^{1/2}v$. This implies that

$$Cw = \eta(C+\alpha I)w, \quad \text{so} \quad Cw = \frac{\alpha\eta}{1-\eta}w = \eta' w,$$

say, where $\eta = 1$ would lead to a contradiction. This yields

$$\frac{\alpha\eta}{1-\eta} = \eta' \qquad \implies \qquad \eta = \frac{\eta'}{\alpha+\eta'}.$$

(b) If A is invertible,

$$\begin{aligned} A^{1/2}(A+\alpha B)^{-1}A &= A^{1/2}(A^{1/2}A^{1/2}+\alpha B)^{-1}A \\ &= (I+\alpha A^{-1/2}BA^{-1/2})^{-1}A^{-1/2}A \\ &= (I+\alpha A^{-1/2}BA^{-1/2})^{-1}A^{1/2}, \end{aligned}$$

and then an argument similar to that above gives

$$\eta = \frac{1}{1+\alpha\eta''},$$

where η'' is an eigenvalue of $A^{-1/2}BA^{-1/2}$.

Solution 2

(a) We have $y \sim (X\beta, \sigma^2 I_n) \sim (UD\gamma, \sigma^2 I_n)$, where $\gamma = V^{\mathrm{T}}\beta$, and

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = (VD^{\mathrm{T}}U^{\mathrm{T}}UDV^{\mathrm{T}})^{-1}VD^{\mathrm{T}}U^{\mathrm{T}}y = V(D^{\mathrm{T}}D)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}y,$$

with a similar calculation giving $\widehat{\gamma} = (D^{\mathrm{T}}D)^{-1}D^{\mathrm{T}}U^{\mathrm{T}}y$, so $\widehat{\beta} = V\widehat{\gamma}$ (surprise!)

(b) As $\gamma = V\beta$ and V is orthogonal,

$$Q = (\widehat{\beta}-\beta)^{\mathrm{T}}(\widehat{\beta}-\beta) = (V\widehat{\gamma}-V\gamma)^{\mathrm{T}}(V\widehat{\gamma}-V\gamma) = (\widehat{\gamma}-\gamma)^{\mathrm{T}}V^{\mathrm{T}}V(\widehat{\gamma}-\gamma) = (\widehat{\gamma}-\gamma)^{\mathrm{T}}(\widehat{\gamma}-\gamma),$$

as required. Having $y \sim (UD\gamma, \sigma^2 I_n)$ implies that

$$\widehat{\gamma} = \mathrm{diag}(d_1^{-1}, \ldots, d_p^{-1}, 0, \ldots, 0)U^{\mathrm{T}}y,$$

so

$$\mathrm{var}(\widehat{\gamma}) = \sigma^2\mathrm{diag}(d_1^{-2}, \ldots, d_p^{-2}).$$

This will be large if at least one of the d_r is small, and then there is at least one direction in which γ, i.e., $v^{\mathrm{T}}\beta$ for some $v_{p\times 1}$, is extremely poorly determined.

(c) Under the normal model, $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p$ are independent $\mathcal{N}(\gamma_r, \sigma^2/d_r^2)$ variables, so $\widehat{\gamma}_r - \gamma_r \stackrel{D}{=} \sigma Z_r/d_r$, giving

$$Q = (\widehat{\gamma} - \gamma)^{\mathrm{T}}(\widehat{\gamma} - \gamma) \stackrel{D}{=} \sum_{r=1}^{p} \sigma^2 Z_r^2/d_r^2,$$

and as $\mathrm{E}(Z_r^2) = 1$ and $\mathrm{var}(Z_r^2) = 2$ we get $\mathrm{E}(Q) = \sigma^2 \sum_{r=1}^{p} 1/d_r^2$ and $\mathrm{var}(Q) = 2\sigma^4 \sum_{r=1}^{p} 1/d_r^4$.

Solution 3

(a) We have

$$(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda\beta^{\mathrm{T}}\beta = y^{\mathrm{T}}y - 2y^{\mathrm{T}}X\beta + \beta^{\mathrm{T}}(X^{\mathrm{T}}X + \lambda I_p)\beta,$$

and differentiation with respect to β gives first and second derivatives

$$-2yX^{\mathrm{T}}y + 2(X^{\mathrm{T}}X + \lambda I_p)\beta, \quad 2(X^{\mathrm{T}}X + \lambda I_p).$$

The second derivative matrix is positive definite for any $\lambda > 0$, and setting the first derivative to zero gives

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}y.$$

(b) Setting $X = UDV^{\mathrm{T}}$ gives $X^{\mathrm{T}}y = VD^{\mathrm{T}}U^{\mathrm{T}}y = \sum_{d_j>0} v_j d_j u_j^{\mathrm{T}} y$ and

$$(X^{\mathrm{T}}X + \lambda I_p)^{-1} = (VD^{\mathrm{T}}DV^{\mathrm{T}} + \lambda I_p)^{-1} = \{V(D^{\mathrm{T}}D + \lambda I_p)V^{\mathrm{T}}\}^{-1} = VS_\lambda V^{\mathrm{T}},$$

where $S_\lambda = \mathrm{diag}(d_1^2 + \lambda, \ldots, d_r^2 + \lambda)^{-1}$ exists because all its elements are positive. Hence

$$\widehat{\beta}_\lambda = (X^{\mathrm{T}}X + \lambda I_p)^{-1}X^{\mathrm{T}}y = VS_\lambda V^{\mathrm{T}}(VD^{\mathrm{T}}U^{\mathrm{T}})y = \sum_{d_j>0} \frac{d_j}{d_j^2 + \lambda} u_j^{\mathrm{T}} y \times v_j.$$

Likewise

$$\widehat{y}_\lambda = X\widehat{\beta}_\lambda = H_\lambda y = UDS_\lambda D^{\mathrm{T}}U^{\mathrm{T}}y = \sum_{d_j>0} u_j \times \frac{d_j^2}{d_j^2 + \lambda} u_j^{\mathrm{T}} y.$$

Both $\widehat{\beta}_\lambda$ and $\widehat{y}_\lambda$ shrink towards zero as λ increases, with the strongest shrinkage for those vectors v_j and u_j for which d_j is smallest.

(c) We have

$$\mathrm{edf}_\lambda = \mathrm{tr}(H_\lambda) = \mathrm{tr}\{(UDV^{\mathrm{T}})VS_\lambda V^{\mathrm{T}}(UDV^{\mathrm{T}})^{\mathrm{T}}\} = \mathrm{tr}\{U^{\mathrm{T}}UDS_\lambda D^{\mathrm{T}}\} = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

and

$$\begin{aligned}
\mathrm{E}(\widehat{\beta}_\lambda) &= VS_\lambda V^{\mathrm{T}}VD^{\mathrm{T}}U^{\mathrm{T}}UDV^{\mathrm{T}}\beta = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} v_j^{\mathrm{T}}\beta \times v_j, \\
\mathrm{var}(\widehat{\beta}_\lambda) &= VS_\lambda D^{\mathrm{T}}U^{\mathrm{T}}\mathrm{cov}(y)\{VS_\lambda D^{\mathrm{T}}U^{\mathrm{T}}\}^{\mathrm{T}} = \sigma^2 V \mathrm{diag}\left\{\frac{d_1^2}{(d_1^2 + \lambda)^2}, \ldots, \frac{d_p^2}{(d_p^2 + \lambda)^2}\right\} V^{\mathrm{T}},
\end{aligned}$$

so the bias is

$$\mathrm{E}(\widehat{\beta}_\lambda) - \beta = \sum_{r=1}^{p} \frac{d_r^2}{d_r^2 + \lambda} v_r^{\mathrm{T}}\beta \times v_r - \sum_{r=1}^{p} v_r v_r^{\mathrm{T}}\beta = -\sum_{r=1}^{p} \frac{\lambda}{d_r^2 + \lambda} v_r^{\mathrm{T}}\beta \times v_r$$

Solution 4

(a) Let $H = 1_n(1_n^{\mathrm{T}}1_n)^{-1}1_n^{\mathrm{T}}$ correspond to regression on a column of ones, and note that $(I_n - H)1_n = 0$, $Hy = 1_n\overline{y}$ and $HX = 1_n\overline{x}^{\mathrm{T}}$, where $\overline{x}$ contains the means of the columns of X. Then we can set $y_* = (I_n - H)y$ and $X_* = (I_n - H)X$, so

$$y - \beta_0 1_n - X\beta = (I_n - H)y + Hy - \beta_0 1_n - (I_n - H)X\beta + HX\beta = y_* - (\gamma - \overline{y})1_n - X_*\beta,$$

where $\gamma = \beta_0 - \overline{x}^{\mathrm{T}}\beta$. The interpretation of β remains the same; only the intercept γ has changed.

(b) The equality implies that we can write

$$\|y - \beta_0 1_n - X\beta\|_2^2 = \|y_* - (\gamma - \overline{y})1_n - X_*\beta\|_2^2 = \|y_* - X_*\beta\|_2^2 + \|(\gamma - \overline{y})1_n\|_2^2,$$

because

$$(y_* - X_*\beta)^{\mathrm{T}}(\gamma - \overline{y})1_n = (\gamma - \overline{y})(y - X\beta)^{\mathrm{T}}(I - H)^{\mathrm{T}}1_n = (\gamma - \overline{y})(y - X\beta)^{\mathrm{T}}(I - H)1_n = 0.$$

Hence

$$\min_{\beta_0,\beta} \|y - \beta_0 1_n - X\beta\|_2^2 + \lambda p(\beta) = \min_{\gamma,\beta} \|y_* - X_*\beta\|_2^2 + \|(\gamma - \overline{y})1_n\|_2^2 + \lambda p(\beta),$$

which gives $\widehat{\gamma} = \overline{y}$ and $\widehat{\beta}_\lambda$ as the solution to the second minimisation problem, as required. Hence provided y and the columns of X are centered, the intercept need not be included.

(c) Including β_0 in β would mean that as λ increases, $\widehat{\beta}_\lambda \to 0$, i.e., shrinkage would apply also to the intercept, which depends on the units used for measuring y. Hence a change from measuring temperature in $^\circ C$ to $^\circ F$ would lead to different conclusions about the effects of the covariates, which is clearly undesirable.

Expressed in algebra, we would have a column 1_n in X if β contains the intercept, and then if the intercept is the first column of X, we have

$$y - X\beta \mapsto ay + b1_n - X\beta = a(y - X\beta_*), \quad \beta \mapsto \beta_* = \{\beta - (b, 0, \ldots, 0)^{\mathrm{T}}\}/a.$$

In this new parametrisation we have $\widehat{\beta}_{*,\lambda} \to 0$ as $\lambda \to \infty$, corresponding to the estimate of β_0 tending to b rather than to 0, and this would affect all the other parameter estimates.