Answer sheet 12

Assignment 1.

(i). This means that all the columns of X_1 are orthogonal to the columns of X_2. In other words $\mathcal{M}(X_1) \perp \mathcal{M}(X_2)$.

(ii). Remember first that

$$X^tX = \begin{pmatrix} X_1^tX_1 & 0 \\ 0 & X_2^tX_2 \end{pmatrix},$$

thus

$$\begin{aligned} H &= (X_1, X_2) \begin{pmatrix} (X_1^tX_1)^{-1} & 0 \\ 0 & (X_2^tX_2)^{-1} \end{pmatrix} (X_1, X_2)^t \\ &= X_1(X_1^tX_1)^{-1}X_1^t + X_2(X_2^tX_2)^{-1}X_2^t = H_1 + H_2. \end{aligned}$$

Moreover as $X_1^tX_2 = 0$, we have $H_1H_2 = 0$. And thus $H_2H_1 = H_2^tH_1^t = (H_1H_2)^t = 0$,

$$HH_1 = (H_1 + H_2)H_1 = H_1^2 = H_1$$

and $H_1H = H_1^tH^t = (HH_1)^t = H_1^t = H_1$.

Interpretation : $H_1H_2 = 0$ comes from the fact that the columns of X_1 et X_2 are orthogonal, hence if one projects on $\mathcal{M}(X_2)$ and then on $\mathcal{M}(X_1)$, will obtain the vector 0 as a result. The interpretation for $H_2H_1 = 0$ is similar. $HH_1 = H_1$ comes from projecting on $\mathcal{M}(X_1)$ and then projecting on $\mathcal{M}(X)$ is equivalent to project uniquely on $\mathcal{M}(X_1)$, as $\mathcal{M}(X_1)$ is a subspace of $\mathcal{M}(X)$. For the same reason, $H_1H = H_1$ because we project on $\mathcal{M}(X)$ and after that on $\mathcal{M}(X_1)$, which is like if we were projecting only on $\mathcal{M}(X_1)$. In tuitively we remark that even if $X_1^tX_2 \neq 0$, we still have $HH_1 = H_1 = H_1H$, but $H_1H_2 \neq 0$ and $H_2H_1 \neq 0$.

(iii). Using the fact that $Hy = (H_1 + H_2)y$,

(a) immediate

(b) follows from $H_2H_1 = 0$;

(c) follows from $H(I - H_1) = H - H_1 = H_2$.

Assignment 2.

(i).

$$(X^tX)^{-1} = \begin{pmatrix} (X_1^tX_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^tX_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^tX_k)^{-1} \end{pmatrix}$$

and

$$(X_L^tX_L)^{-1} = \mathrm{diag}((X_i^tX_i)^{-1} : i \in L).$$

Hence

$$H = X_1(X_1^tX_1)^{-1}X_1^t + \cdots + X_k(X_k^tX_k)^{-1}X_k^t = H_1 + \cdots + H_k$$

and

$$H_L = \sum_{i \in L} X_i(X_i^tX_i)^{-1}X_i^t = \sum_{i \in L} H_i.$$

(ii). If $i = j$, $H_iH_j = H_i^2 = H_i$ and if $i \neq j$, $H_iH_j = X_i(X_i^tX_i)^{-1}X_i^tX_j(X_j^tX_j)^{-1}X_j^t = 0$ so that $X_i^tX_j = 0$.

(iii).

$$\hat{\beta} = (X^tX)^{-1}X^ty = \begin{pmatrix} (X_1^tX_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^tX_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^tX_k)^{-1} \end{pmatrix} \begin{pmatrix} X_1^t \\ X_2^t \\ \vdots \\ X_k^t \end{pmatrix} y = \begin{pmatrix} (X_1^tX_1)^{-1}X_1^ty \\ (X_2^tX_2)^{-1}X_2^ty \\ \vdots \\ (X_k^tX_k)^{-1}X_k^ty \end{pmatrix}.$$

(iv). First of all notice that

$$e_L := y - H_Ly = y - \sum_{i\in L} H_iy$$

and that

$$e_{L\cup\{j\}} := y - H_{L\cup\{j\}}y = y - \sum_{i\in L\cup\{j\}} H_iy.$$

Moreover

$$\begin{aligned}(I - H_{L\cup\{j\}})e_L &= (I - H_{L\cup\{j\}})(I - H_L)y \\ &= (I - H_L - H_{L\cup\{j\}} + H_{L\cup\{j\}}H_L)y \\ &= (I - H_{L\cup\{j\}})y \\ &= e_{L\cup\{j\}}.\end{aligned}$$

Then $e_{L\cup\{j\}}$ is an orthogonal projection of e_L, where $e_L - e_{L\cup\{j\}} \perp e_{L\cup\{j\}}$ and

$$\|e_{L\cup\{j\}}\|^2 + \|e_L - e_{L\cup\{j\}}\|^2 = \|e_L\|^2.$$

Hence

$$RSS_L - RSS_{L\cup\{j\}} = \|e_L\|^2 - \|e_{L\cup\{j\}}\|^2 = ||e_L - e_{L\cup\{j\}}||^2 = ||H_jy||^2$$

is independent from L.

(v). The interpretation wrt ANOVA is that in this case, adding one variable X_j does not depend on the variables that are already in the model. **This is not true in general!**

Assignment 3. We know that the ridge regression parameter is a function of the smoothing parameter λ

$$\widehat{\beta}_0 = \overline{y}, \quad \widehat{\gamma}_\lambda = (Z^tZ + \lambda I)^{-1}Z^ty.$$

Let $Z = U_{n\times n}\Sigma_{n\times q}V_{q\times q}^t$ the SVD decomposition of Z with $\Sigma = \operatorname{diag}(\omega_1, \dots, \omega_q)$. A direct computation yields

$$\begin{aligned}\widehat{\gamma}_\lambda &= (Z^tZ + \lambda I)^{-1}Z^ty \\ &= (V\Sigma^t\Sigma V^t + \lambda I)^{-1}V\Sigma^tU^ty \\ &= (V[\Sigma^t\Sigma + \lambda I]V^t)^{-1}V\Sigma^tU^ty \\ &= V(\Sigma^t\Sigma + \lambda I)^{-1}\Sigma^tU^ty.\end{aligned}$$

where

$$\begin{aligned}
\hat{y}_{\text{ridge}} &= X\widehat{\beta}_\lambda \\
&= \widehat{\beta}_0 \mathbf{1} + Z\widehat{\gamma} \\
&= \overline{y}\mathbf{1} + U\left\{\Sigma\left(\Sigma^t\Sigma + \lambda I\right)^{-1}\Sigma^t\right\}U^t y \\
&= \overline{y}\mathbf{1} + \sum_{j=1}^{q} \frac{\omega_j^2}{\omega_j^2+\lambda} u_j(u_j^t y),
\end{aligned}$$

because the matrix between the parenthesis is diagonal $n \times n$ with the q first values equal to $\omega_j^2/(\omega_j^2+\lambda)$ and the $n-q$ remaining vanish.
If $\omega_j \approx 0$ and $\lambda >> \omega_j^2$, and there is much difference between 1 and $\omega_j^2/(\omega_j^2+\lambda) \approx 0$. The parameter λ shrinks the component u_j of $\widehat{y}_{\text{ridge}}$ (which is$\widehat{y}^t_{\text{ridge}} u_j$), and the variance of the fitted values in the direction of u_j is small.

Assignment 4. Since everything is positive $\widehat{\beta}_0 = \overline{y}$ independently on λ, then it is enough to consider $\|\widehat{\gamma}_{\text{ridge}}\|_2^2$. Let $\widehat{\gamma} = \widehat{\gamma}_{\text{ridge}}$.

Let $Z = U\Sigma V^t$ the SVD decomposition of Z. By an argument similar to the one of the previous exercise,

$$\widehat{\gamma} = V(\Sigma^t\Sigma + \lambda I)^{-1}\Sigma^t U^t y = \sum_{j=1}^{q} \frac{\omega_j}{\omega_j^2+\lambda}(u_j^t y)v_j.$$

Since the v_j are orthonormal we find

$$\widehat{\gamma}^t\widehat{\gamma} = \sum_{j=1}^{q}\sum_{i=1}^{q} \frac{\omega_j}{\omega_j^2+\lambda}(u_j^t y)\frac{\omega_i}{\omega_i^2+\lambda}(u_i^t y)v_j^t v_i = \sum_{j=1}^{q}\left(\frac{\omega_j}{\omega_j^2+\lambda}\right)^2 (u_j^t y)^2,$$

which is decreasing in λ.

Assignment 5. (a) Since $\widehat{\beta}_0 = \overline{y}$ (why ?), we have

$$g(\gamma) = \|y - \overline{y}\mathbf{1} - Z\gamma\|_2^2 = \|y^* - Z\gamma\|_2^2 = \sum_{i=1}^{n}\left(y_i^* - \sum_{j=1}^{q} Z_{ij}\gamma_j\right)^2.$$

(b) By the chain rule, we have

$$\frac{\partial g}{\partial \gamma_j}(0) = -\sum_{i=1}^{n} 2\left(y_i^* - \sum_{k=1}^{q} Z_{ik}0\right)Z_{ij} = -2Z_j^T y^* = -2Z_j^T y, \qquad j = 1,\dots,q,$$

since $Z^T\mathbf{1} = 0$.

(c) We have for small t

$$f(te_j) = g(te_j) + \lambda\|te_j\|_1 = g(te_j) + \lambda|t| = g(0) - 2t(Z_j^T y) + \lambda|t| + o(t).$$

If $2Z_j^T y > 0$ then for $t > 0$ small, $f(te_j) < g(0) = f(0)$. If $2Z_j^T y < 0$ then for $t < 0$ small (close to zero), $f(te_j) < f(0)$. In both cases 0 is not a minimiser of f.

(d) Since g is convex (even if it wasn't we could introduce an $o(\|v\|)$ term)

$$f(v) \geq g(0) + [\nabla g(0)]^T v + \lambda\|v\|_1 \geq g(0) + (\lambda - \|\nabla g(0)\|_\infty)\|v\|_1 = f(0) + (\lambda - \lambda^*)\|v\|_1.$$

As $\lambda \geq \lambda^*$, this shows that f is minimised at 0. If $\lambda > \lambda^*$ then 0 is the only minimiser. It follows from a further assignment that if $\lambda = \lambda^* > 0$, then 0 is the unique minimiser.

Assignment 6. Both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ estimate β_0 by $\overline{y}$ and so $X\widehat{\beta}_1 = \overline{y}\mathbf{1} + Z\widehat{\gamma}_1$ and similarly for $\widehat{\beta}_2$. Therefore we only need to deal with the estimators of γ. Let $y^* = y - \overline{y}\mathbf{1}$.

(a) Assume that $\hat{\gamma}^{(1)}$ and $\hat{\gamma}^{(2)}$ both give an optimal objective value v. Note first that $\|Y - Z\gamma\|_2^2$ is a strictly convex function of $Z\gamma$, and hence for $t \in (0,1)$, we have

$$\|Y - tZ\hat{\gamma}^{(1)} - (1-t)Z\hat{\gamma}^{(2)}\|_2^2 \leq t\|Y - Z\hat{\gamma}^{(1)}\|_2^2 + (1-t)\|Y - Z\hat{\gamma}^{(2)}\|_2^2 \tag{1}$$

with equality if and only if $Z\hat{\gamma}^{(1)} = Z\hat{\gamma}^{(2)}$. Now, by optimality of $\hat{\gamma}^{(1)}$, $\hat{\gamma}^{(2)}$ and convexity of the L^1 norm, we see that

$$\begin{aligned}
v &\leq \|Y - tZ\hat{\gamma}^{(1)} - (1-t)Z\hat{\gamma}^{(2)}\|_2^2 + \lambda\|t\hat{\gamma}^{(1)} + (1-t)\hat{\gamma}^{(2)}\|_1 \\
&\leq t\|Y - Z\hat{\gamma}^{(1)}\|_2^2 + (1-t)\|Y - Z\hat{\gamma}^{(2)}\|_2^2 + \lambda t\|\hat{\gamma}^{(1)}\|_1 + \lambda(1-t)\|\hat{\gamma}^{(2)}\|_1 \\
&= t\{\|Y - Z\hat{\gamma}^{(1)}\|_2^2 + \lambda\|\hat{\gamma}^{(1)}\|_1\} + (1-t)\{\|Y - Z\hat{\gamma}^{(2)}\|_2^2 + \lambda\|\hat{\gamma}^{(2)}\|_1\} \\
&= tv + (1-t)v = v
\end{aligned}$$

by optimality of both $\hat{\gamma}^{(1)}$ and $\hat{\gamma}^{(2)}$. Hence, equality must have been preserved throughout this chain of inequalities, which in particular means that there must have been equality in (1). Thus $Z\hat{\gamma}^{(1)} = Z\hat{\gamma}^{(2)}$, which in turn implies that $X\hat{\beta}_1 = X\hat{\beta}_2$.

(b) We get this directly from (a) :

$$\lambda\|\widehat{\gamma}_1\|_1 = f(\widehat{\gamma}_1) - \|y^* - Z\widehat{\gamma}_1\|_2^2 = f(\widehat{\gamma}_2) - \|y^* - Z\widehat{\gamma}_2\|_2^2 = \lambda\|\widehat{\gamma}_2\|_1.$$

(c) From part (a) we know that the solutions have the form $(\overline{y}, \widehat{\gamma}^T)^T$ and $(\overline{y}, \widehat{\gamma}^T + v^T)^T$, with $Zv = 0$. This means that $v = (-\epsilon, \epsilon)^T$ for some $\epsilon \neq 0$. From part (b) we know that $\|\widehat{\gamma}\|_1 = \|\widehat{\gamma} + v\|_1$. We can find such a nonzero v if and only if $\widehat{\gamma} \neq 0$. (For example, if $\widehat{\gamma}^T = (0, 0.1)$, then any $\epsilon \in [-0.1, 0]$ will do.) So we just need to check that 0 is not a solution. This can be done using a previous assignment ($\lambda = 1 < \lambda^* = 4$) or directly : the objective function in γ is

$$2(1 - \gamma_1 - \gamma_2)^2 + |\gamma_1| + |\gamma_2|.$$

At 0 this equals 2, whereas at $(0,1)^T$ this equals 1. So the optimal $\widehat{\gamma}$ is not zero. Consequetly, there exists an $\epsilon > 0$ for which $\|\widehat{\gamma}\|_1 = \|\widehat{\gamma} + v\|_1$. In fact, a straightforward calculation shows that the set of solutions is

$$\{(\widehat{\gamma}_1, \widehat{\gamma}_2)^T : 0 \leq \widehat{\gamma}_i \quad \text{et} \quad \widehat{\gamma}_1 + \widehat{\gamma}_2 = 3/4\} = \{(3/8, 3/8)^T + (-\epsilon, \epsilon)^T : |\epsilon| \leq 3/8\}.$$