Answer sheet 12

Assignment 1. (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

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

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 2. 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

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

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

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

Let $Z = U_{n\times n}\Sigma_{n\times q}V^t_{q\times q}$ the SVD decomposition of Z with $\Sigma = \text{diag}(\omega_1, \ldots, \omega_q)$. A direct computation yields

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

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 in 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.

$$\begin{aligned} f(y;\mu,\nu) &= \frac{1}{\Gamma(\nu)}y^{\nu-1}\left(\frac{\nu}{\mu}\right)^{\nu}\exp(-\nu y/\mu) \\ &= exp\left\{-log(\Gamma(\nu)) + (\nu-1)log(y) + \nu log(\nu) - \nu log(\mu) - \frac{\nu y}{\mu}\right\} \\ &= exp\left\{\nu\left(\frac{-y}{\mu} - log\mu\right) - log(\Gamma(\nu)) + (\nu-1)log(y) + \nu log(\nu)\right\} \end{aligned}$$

Thus $\phi = 1/\nu, \theta = -1/\mu$, $b(\theta) = log\mu = -log(-\theta)$ and $c(y,\phi) = -log(\Gamma(\nu)) + (\nu-1)log(y) + \nu log(\nu)$. The link function is $\eta = \theta = b'^{-1}(\mu) = -1/\mu$, $\mathbb{E}(Y) = b'(\theta) = -1/\theta = \mu$, $\mathrm{Var}(Y) = \phi b''(\theta) = \frac{\mu^2}{\nu}$.

Assignment 6.

$$\begin{aligned} f(y;\lambda,\mu) &= \left(\frac{\lambda}{2\pi y^3}\right)^{1/2}\exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}, \\ &= \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right\} \\ &= \exp\left\{-\frac{\lambda y}{2\mu^2} - \frac{\lambda}{2y} + \frac{\lambda}{\mu} + \frac{1}{2}\log\left(\frac{\lambda}{2\pi y^3}\right)\right\} \\ &= \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \end{aligned}$$

with

$$\theta = 1/2\mu^2, \quad \phi = -1/\lambda, \quad b(\theta) = \sqrt{2\theta}, \quad c(y;\phi) = \frac{1}{2\phi y} + \frac{1}{2}\log\left(-\frac{1}{\phi 2\pi y^3}\right).$$

Thus

$$E(Y) = \frac{\sqrt{2}}{2\sqrt{\theta}} = \mu, \quad \mathrm{Var}(Y) = \phi\frac{-1}{(2\theta)^{3/2}} = \mu^3/\lambda.$$

Assignment 7.

(i).

$$\begin{aligned} f(r;\pi) &= \binom{m}{r}\pi^r(1-\pi)^{m-r} \\ &= exp\left[\log\binom{m}{r} + r\log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi)\right] \\ &= exp\left[m\left\{\frac{r}{m}\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} + \log\binom{m}{r}\right] \\ &= \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y;\phi)\right\}, \end{aligned}$$

with

$$\begin{aligned} y &= \frac{r}{m}, \quad \phi = \frac{1}{m}, \quad c(y;\phi) = \log \binom{m}{r}, \\ \theta &= \log\left(\frac{\pi}{1-\pi}\right), \text{ donc } \quad \pi = e^{\theta}(1-\pi), \quad \pi = \frac{e^{\theta}}{1+e^{\theta}}. \\ b(\theta) &= -\log(1-\pi) = -\log\left\{1 - \frac{e^{\theta}}{1+e^{\theta}}\right\} = -\log\left\{\frac{1}{1+e^{\theta}}\right\} = \log(1+e^{\theta}). \end{aligned}$$

(ii).

$$E(Y) = \mu = b'(\theta) = \frac{e^{\theta}}{1+e^{\theta}} = \pi$$

$$Var(Y) = \phi V(\mu) = \phi b''(\theta) = \frac{e^{\theta}(1+e^{\theta}) - e^{2\theta}}{m(1+e^{\theta})^2} = \frac{e^{\theta}}{m(1+e^{\theta})^2} = \frac{\mu(1-\mu)}{m}.$$

and thus $V(\mu) = \mu(1-\mu)$

Assignment 8.

$$E(Y) = P(X > 0) = 1 - P(X = 0) = 1 - \exp(-\mu) = 1 - \exp\{-\exp(x^T\beta)\}.$$