Solutions Sheet 1 (rev 7, 12th May 2021)

Notation: we use $||\cdot||_2$ to denote the l^2-norm.

Question 1

We write the sum of squared residuals as $\mathcal{L}(a,b) = \sum\limits_{i=1}^{n} e_i^2 = \sum\limits_i (w_i - a - bh_i)^2$ The least squares estimates are given by $(\tilde{a},\tilde{b}) = \min\limits_{a,b} \mathcal{L}(a,b)$. We solve

$$\left.\frac{\partial \mathcal{L}}{\partial a}\right|_{\tilde{a},\tilde{b}} = 0, \quad \left.\frac{\partial \mathcal{L}}{\partial b}\right|_{\tilde{a},\tilde{b}} = 0. \tag{1}$$

We then have $\sum_i (w_i - \tilde{a} - \tilde{b}h_i) = 0$, hence $\tilde{a} = \frac{\sum\limits_i w_i - \tilde{b}\sum\limits_i h_i}{n}$. From (1), we also have $\sum_i (w_i - \tilde{a} - \tilde{b}h_i)h_i = 0$. Plugging in $\tilde{a}$, we have

$$\sum_i (w_i h_i - \bar{w}h_i + \tilde{b}\bar{h}h_i - \tilde{b}h_i^2) = 0. \tag{2}$$

Therefore

$$\tilde{b} = \frac{\sum_i w_i h_i - \bar{w}\sum_i h_i}{\sum_i h_i^2 - n\bar{h}^2} = \frac{n\sum_i w_i h_i - (\sum_i h_i)(\sum_i w_i)}{n\sum_i h_i^2 - (\sum_i h_i)^2} \tag{3}$$

To show they are indeed minimisers, one could check the Hessian is positive (semi)-definite.

Question 2

Since we have shifted the mean, we simply replace h_i in the expression of $\tilde{a}$ by g_i. Since $\bar{g} = n^{-1}\sum_i g_i = 0$, we have $\tilde{a}^{(g)} = \bar{w} = n^{-1}\sum_i w_i$. Doing the same for $\tilde{b}$, we obtain

$$\tilde{b}^{(g)} = \sum_{i=1}^{n} g_i w_i / \sum_{i=1}^{n} g_i^2. \tag{4}$$

Writing $g_i = h_i - \bar{h}$, and plugging it into equation (4), we get

$$\tilde{b}^{(g)} = \frac{\sum_i (h_i - \bar{h})w_i}{\sum_i (h_i - \bar{h})^2} = \frac{n\sum_i w_i h_i - (\sum_i h_i)(\sum_i w_i)}{n(\sum_i h_i^2 - 2\bar{h}\sum_i h_i + n\bar{h}^2)} \tag{5}$$

$$= \frac{n\sum_i w_i h_i - (\sum_i h_i)(\sum_i w_i)}{n\sum_i h_i^2 - (\sum_i h_i^2)} = \tilde{b} \tag{6}$$

then it is straightforward to see that

$$\tilde{a} = \tilde{a}^{(g)} - \bar{h}\tilde{b}^{(g)} \tag{7}$$

Question 3

For each j, the fitted value $\hat{w}_j$ is given by $\hat{w}_j = \tilde{a} + \tilde{b}h_j$. Substituting in the values of $\tilde{a}$ and $\tilde{b}$, we obtain

$$\hat{w}_j = \frac{1}{n}(\sum_i w_i - \tilde{b}\sum_i h_i) + \tilde{b}h_j \tag{8}$$

From the expression of $\tilde{b}$ from question 1, we can see that $\tilde{b}$ is a LC of $\{w_i\}_{i=1}^n$. Hence, both $\tilde{b}\sum_i h_i$ and $\tilde{b}h_j$ are LCs of $\{w_i\}_{i=1}^n$. Therefore $\hat{w}_j$ is a LC of $\{w_i\}_{i=1}^n$. [In general, if we have a linear model $Y = X\beta + \varepsilon$, with $X \in \mathbb{R}^{n\times p}$, In Statistical Modelling I, we have seen that $\hat{Y}$ satisfy $\hat{Y} = PY$ where $P := X(X^TX)^{-1}X^T$. Hence, the fitted values $\hat{Y}$ can always be written as a linear combination of the components of Y in OLS.]

Question 4

In this question I will write the definitions in terms of a multidimensional linear model where $Y = X\beta + \varepsilon$, with $X \in \mathbb{R}^{n\times p}$
Leverage: Recall that the fitted values $\hat{Y}$ satisfy $\hat{Y} = HY$ where $H = X(X^TX)^{-1}X^T$. The value $h_i := H_{ii}$ is called the *leverage* of the i-th observation. It measures the contribution that Y_i makes to the fitted value $\hat{Y}$. It can be shown that $0 \leq h_i \leq 1$. Since $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1-h_i)$, values of h_i close to 1 force the regression line (or plane) to pass very close to Y_i.
Cook's distance: The Cook's distance D_i of the observation (Y_i, x_i)

$$D_i := \frac{\frac{1}{p}||X(\hat{\beta}_{(-i)} - \hat{\beta})||_2^2}{\tilde{\sigma}^2} \tag{9}$$

where $\hat{\beta}_{(-i)}$ is the OLS estimate of β when omitting observation (Y_i, x_i) and $\tilde{\sigma}^2 := (n-p)^{-1}||Y - X\hat{\beta}||_2^2$.

Cook's distance tries to understand how the fitted values change when you omit a point from the fit and measures the size of that change, normalised by the sampling variance, so the Cook's distances can be seen to be on a normalised scale. Under the usual null hypothesis, Cook's distance has a $F_{p,n-p}$ distribution and the median of that distribution has been used as a cutoff, in that if Cook's distance is greater then the point, i that was omitted is considered to be highly influential. A rule-of-thumb of $D_i > 1$ is another indicator of an influential point.

If a point has an influential Cook's distance then it is worth going back and checking that there were no errors with the original collection and recording of that observation, or otherwise scrutinising that individual for validity.

[So why is $0 \leq h_i \leq 1$? Well, H is symmetric because:

$$H^T = \left\{X(X^TX)^{-1}X^T\right\}^T = X(X^TX)^{-1}X^T = H, \tag{10}$$

and idempotent, which means

$$\begin{aligned} H^2 &= \left\{X(X^TX)^{-1}X^T\right\}\left\{X(X^TX)^{-1}X^T\right\} && (11)\\ &= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T && (12)\\ &= X(X^TX)^{-1}X^T = H. && (13) \end{aligned}$$

Now, let $B = H^2$ and the usual formula for matrix multiplication can be written

$$b_{i,j} = \sum_{k=1}^{n} h_{i,k} h_{k,j}. \tag{14}$$

Hence, $b_{i,i} = \sum_{k=1}^{n} h_{i,k} h_{k,i}$ and since $H = H^2 = B$ this means that

$$\begin{aligned} h_{i,i} &= h_{i,1}h_{1,i} + h_{i,2}h_{2,i} + \cdots + h_{i,n}h_{n,i} && (15) \\ &= h_{i,1}^2 + h_{i,2}^2 + \cdots + h_{i,n}^2 \geq h_{i,i}^2. && (16) \end{aligned}$$

Since $h_{i,i}^2 \leq h_{i,i}$ this means simultaneously that $h_{i,i} \geq h_{i,i}^2 \geq 0$ and, dividing through by $h_{i,i}$ gives $h_{i,i} \leq 1$, as required.]

Question 5

Using suffix notation, we have $w = \beta_i A_{ij} \beta_j, \partial_{\beta_k} w = \delta_{ik} A_{ij} \beta_j + \beta_i A_{ij} \delta_{ik} = A_{kj}\beta_j + \beta_i A_{ik} = 2A_{kj}\beta_j = 2(A\beta)_k$, where we have used that A is symmetric in the penultimate equality.

Or start with the definition of matrix multiplication $\sum_{i=1}^{p} \beta_i \sum_{j=1}^{p} A_{i,j}\beta_j$.

Question 6

The eigenvalues are $1 + \lambda + p$ and $1 + \lambda - p$. The corresponding eigenvectors are $(-1, 1)^T$ and $(1, 1)^T$.

Question 7

Note that the first column of the design matrix is $\mathbf{1}$, and the first component of β is β_0.

(A) For random variables Y, β Bayes theorem states:

$$p(\beta|Y) = \frac{p(Y|\beta)p(\beta)}{p(Y)}. \tag{17}$$

However, a "standard trick"/device in Bayesian statistics is to realize that we will eventually be interested in learning about the form of posterior density of $p(\beta|Y)$ as a function of β. You're not interested in anything that might be a constant multiple of it, because you know that since $p(\beta|Y)$ is a density you always know that

$$\int_{\beta} p(\beta|Y) d\beta = 1, \tag{18}$$

and whatever bit of the density does not directly involve β is part of the normalizing constant and we don't need to know what it is. In particular, $p(Y)$ is part of the normalising constant, because it does not involve β, so we can temporarily ignore it and focus on learning:

$$p(\beta|Y) \propto p(Y|\beta)p(\beta). \tag{19}$$

(B) So, let's figure out what $p(\beta)$ and $p(Y|\beta)$ are. The question says that $\beta_i \sim N(0, \tau^2)$. So, you will remember the formula for the normal distribution, which means the prior density for β_i is

$$p(\beta_i) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\beta_i^2}{2\tau^2}\right), \tag{20}$$

for $\beta \in (-\infty, \infty)$. Now, remember we are only interested in anything that directly involves β, so we can simplify our interest in (20) to get

$$p(\beta_i) \propto \exp\left(-\frac{\beta_i^2}{2\tau^2}\right). \tag{21}$$

Now, we need the prior density for the *vector* β, not just the ith component. We assume that the $\{\beta_i\}$ are independent and so the multivariate prior for the β is just the product of the individual densities:

$$p(\beta) = \prod_{i=1}^{p} p(\beta_i) \propto \prod_{i=1}^{p} \exp\left(-\frac{\beta_i^2}{2\tau^2}\right) = \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{p}\beta_i^2\right), \tag{22}$$

using the $\exp(a)\exp(b) = \exp(a+b)$ facility of exp. We know $||\beta||_2^2 = \sum_{i=1}^{p}\beta_i^2$, which is just the squared 2-norm of β, so we can write

$$p(\beta) \propto \exp\left(-\frac{||\beta||_2^2}{2\tau^2}\right). \tag{23}$$

The question also states that $Y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$, which is just the regression model and enforcing that the errors are Gaussian with variance of σ^2. We follow a similar procedure that we did immediately above for β, so:

$$p(Y_i|\beta) \propto \exp\left\{-\frac{(Y_i - \beta_0 - x_i^T\beta)^2}{2\sigma^2}\right\}, \tag{24}$$

is from the univariate normal distribution formula and, by exactly the same reasoning as in the previous paragraph, this becomes

$$p(Y|\beta) \propto \exp\left(-\frac{||Y - X\beta||_2^2}{2\sigma^2}\right), \tag{25}$$

for the multivariate $Y = (Y_1, \ldots, Y_n)^T$, incorporating the β_0 into the β vector and the associated vector of 1s into the design matrix X.

(C) Now to get the essential part of the posterior density $p(\beta|Y)$ we use (19) to obtain:

$$\begin{aligned}
p(\beta|Y) &\propto p(Y|\beta)p(\beta) && (26)\\
&= \exp\left(-\frac{||Y - X\beta||_2^2}{2\sigma^2}\right)\exp\left(-\frac{||\beta||_2^2}{2\tau^2}\right) && (27)\\
&= \exp\left\{-\frac{||Y - X\beta||_2^2 + (\sigma^2/\tau^2)||\beta||_2^2}{2\sigma^2}\right\} && (28)\\
&= \exp\left\{-\frac{||Y - X\beta||_2^2 + \lambda||\beta||_2^2}{2\sigma^2}\right\}, && (29)
\end{aligned}$$

where $\lambda = \sigma^2/\tau^2$.

Now the formula for the multivariate normal distribution, where $V \sim N_p(\mu, \Sigma)$ and V is a p-dimensional random vector, μ is the p-dimensional mean vector and Σ is the $p \times p$ variance-covariance matrix with $\mathrm{cov}(V_i, V_j) = \Sigma_{i,j}$, is given by

$$f(V|\mu, \Sigma) \propto \exp\left\{-\frac{1}{2}(V - \mu)^T\Sigma^{-1}(V - \mu)\right\}. \tag{30}$$

Although we haven't done it yet, we are thinking of equating β to V and then showing that β has a normal distribution.

So, can we rewrite (29) in the form of (30)? Let us write out the numerator of (29) inside the exponential:

$$\begin{aligned}
||Y - X\beta||_2^2 + \lambda||\beta||_2^2 &= (Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta & (31)\\
&= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta & (32)\\
&= \beta^T(X^TX + \lambda I_p)\beta - 2\beta^TX^TY + Y^TY. & (33)
\end{aligned}$$

Now similarly expand the similar quantity inside the exponential (30) to get

$$(V - \mu)^T\Sigma^{-1}(V - \mu) = V^T\Sigma^{-1}V - 2V^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu. \qquad (34)$$

The form of (33) and (34) are similar. Look at the quadratic forms in each line first: then you can see that we can set $V = \beta$ and $\Sigma^{-1} = (X^TX + \lambda I_p)$. We also set $\Sigma^{-1}\mu = X^TY$. Then, we can rewrite (33) as

$$\begin{aligned}
(33) &= \beta^T\Sigma^{-1}\beta - 2\beta^T\Sigma^{-1}\mu + Y^TY & (35)\\
&= \beta^T\Sigma^{-1}\beta - 2\beta^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\mu + Y^TY & (36)\\
&= \beta^T\Sigma^{-1}\beta - 2\beta^T\Sigma^{-1}\mu + \mu^T\Sigma^{-1}\mu + \text{const not depending on } \beta & (37)\\
&= (\beta - \mu)^T\Sigma^{-1}(\beta - \mu) + \text{const not depending on } \beta. & (38)
\end{aligned}$$

Using our Bayesian 'constant not depending on β' not relevant when considering the distribution (using the Bayesian device/trick mentioned earlier), we can ignore the constant. The remainder is precisely the form of the term inside the exponential for the multivariate normal distribution. So, the posterior distribution of $\beta|Y \sim N_p(\mu, \Sigma)$.

(D) For the normal distribution the mean is the mode. So the posterior mean (and mode) of $\beta|Y$ is given by

$$\mu = \Sigma X^TY = (X^TX + \lambda I_p)^{-1}X^TY, \qquad (39)$$

and this answers the question: the posterior mean is precisely the ridge regression estimator as given on slide 10 of Lecture 4.

(E) Note that this derivation does not mention the objective function (directly), nor maximisation, but it's based purely on a Bayesian argument from prior and likelihood through the posterior. The ridge parameter $\lambda = \sigma^2/\tau^2$ is entirely controlled by the prior variance for β — so you can think of λ and τ are equivalent quantities. For example, if your prior knowledge is very strong about β, then you'd expect to consider a small prior variance, which translates to a large λ. It's also noticeable that the Bayesian approach does not really consider the condition number or how invertible X^TX is directly.

The above solution is written in a lot of detail. It's acceptable to be brief and here is a more compact solution, which starts at (C) and replaces the argument after point (C).

Since β and Y are normally distributed, $p(\beta) \propto \exp(-\frac{||\beta||_2^2}{2\tau^2})$. Therefore, $p(\beta|Y) \propto \exp\left(-\frac{||Y-X\beta||_2^2+(\sigma^2/\tau^2)||\beta||_2^2}{2\sigma^2}\right)$, which is also normally distributed (which can be immediately recognised by the form, esp. the quadratic and linear bits in β). Since the distribution is Gaussian, the mean is obtained by maximising the density function. Observe that inside the exponential it has the same form (up to a constant) as a ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. The mode and the mean of this distribution is

$$\arg\min_{\beta} ||Y - X\beta||_2^2 + (\sigma^2/\tau^2)||\beta||_2^2 \tag{40}$$

which is indeed the ridge regression estimate.
For the variance, we know that $p(\beta|Y)$ is has a normal distribution and its mean is given by the ridge estimator $\hat{\beta}_\lambda^R$. Using Σ for the covariance matrix, so that $p(\beta|Y) \sim \mathcal{N}(\hat{\beta}_\lambda^R, \Sigma)$. The exponent is given $-\frac{1}{2}(\beta - \hat{\beta}_\lambda^R)^T\Sigma^{-1}(\beta - \hat{\beta}_\lambda^R) = -\frac{1}{2}(\beta^T\Sigma^{-1}\beta - 2\beta^T\Sigma^{-1}\hat{\beta}_\lambda^R + (\hat{\beta}_\lambda^R)^T\Sigma^{-1}\hat{\beta}_\lambda^R)$. Find Σ^{-1} by equating the second order β term from the $p(\beta|Y)$ exponent:

$$\beta^T\Sigma^{-1}\beta = \frac{1}{\sigma^2}\beta^T X^T X\beta + \frac{1}{\tau^2}\beta^T\beta = \frac{1}{\sigma^2}\beta^T(X^TX + \frac{\sigma^2}{\tau^2}I)\beta \tag{41}$$

Hence the $\Sigma^{-1} = \frac{1}{\sigma^2}(X^TX + \frac{\sigma^2}{\tau^2}I)$. The covariance matrix is given by its inverse.

Question 8

Centring is given in Lecture 3 , page 6 as

$$x_{i,j}^* = x_{i,j} - \bar{x}_j, \tag{42}$$

where $\bar{x}_j$ is the mean of variable $j = 1, \dots, p$, i.e. $\bar{x}_j = n^{-1}\sum_{i=1}^n x_{i,j}$. We can construct the vector of the p $\bar{x}_j$ means by notation:

$$n^{-1}\mathbf{1}^T X = \bar{X}^T. \tag{43}$$

I.e. $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)^T$. To form a centred data matrix we want to subtract the jth mean off the jth column of X, which we can do by

$$X_C = X - \mathbf{1}\bar{X}^T = X - \mathbf{1}\mathbf{1}^T X/n = (I_n - \mathbf{1}\mathbf{1}^T/n)X, \tag{44}$$

as required.

Define the centring matrix by $C = I_n - \mathbf{1}\mathbf{1}^T/n$. Observe that $\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{n\times n}$ is a matrix with all entries 1. $C^2 = (I_n - \frac{\mathbf{1}\mathbf{1}^T}{n})(I_n - \frac{\mathbf{1}\mathbf{1}^T}{n}) = I_n - 2\frac{\mathbf{1}\mathbf{1}^T}{n} + n\frac{\mathbf{1}\mathbf{1}^T}{n^2} = I_n - \frac{\mathbf{1}\mathbf{1}^T}{n} = C$. Therefore $C^k = C$ for all $k \in \mathbb{N}$, so it is idempotent and hence a projection matrix.

Since $C^T = C$, it is an orthogonal projection matrix. Further, Cv is a projection of v onto the $n - 1$ dimensional subspace that is orthogonal to the $\mathbf{1}$, which is the subspace of all n-vectors whose components sum to zero.

Question 9

Recall that the Ridge estimator is $\hat{\beta}_\lambda^R = (X^TX + \lambda I)^{-1}X^TY$. Therefore the bias is given by $\mathbb{E}\hat{\beta}_\lambda^R - \beta = (X^TX + \lambda I)^{-1}X^TX\beta - \beta$. Plugging in $X^TX = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and writing $\beta = (\beta_0, \beta_1)^T$, the bias is then given by $\left(\frac{\lambda(\beta_0 - \beta_1\rho + \beta_0\lambda)}{\rho^2 - (1+\lambda^2)}, \frac{\lambda(\beta_1 - \beta_0\rho + \beta_1\lambda)}{\rho^2 - (1+\lambda^2)}\right)^T$. As an exercise, one could show that $X^TX + \lambda I$ is always invertible, even when X doesn't not have full rank. One implication of this is that we can always find the ridge estimator when we cannot find the OLS estimator.

Question 10

We want to minimise

$$M_j = -\hat{\beta}_j^{\text{ls}}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j| \tag{45}$$

When $\hat{\beta}_j^{\text{ls}} < 0$, using the same argument as in lectures, $\beta_j \leq 0$. Taking derivative of M_j and set equal to zero with $\beta_j \leq 0$.

$$\frac{\partial M_j}{\partial \beta_j} = -\hat{\beta}_j^{\text{ls}} + \hat{\beta}_j^{\text{lasso}} - \lambda \tag{46}$$

So the solution $\hat{\beta}_j^{\text{lasso}} = \hat{\beta}_j^{\text{ls}} + \lambda$. This quantity is only feasible if < 0, so the solution is given by

$$-(-\hat{\beta}_j^{\text{ls}} - \lambda)^+ = \text{sgn}(\hat{\beta}_j^{\text{ls}})(|\hat{\beta}_j^{\text{ls}}| - \lambda)^+.$$

An alternative method of doing this is via the *Karush-Kuhn-Tucker conditions.* A reference of the KKT conditions can be found in section 2.24 at `http://www.statslab.cam.ac.uk/~rds37/teaching/modern_stat_methods/notes_MSM.pdf` Note that

$$\frac{1}{2}||Y - X\beta||_2^2 = \sum_{j=1}^{p} \frac{1}{2}(\hat{\beta}_k^{\text{OLS}} - \beta_k)^2 + \frac{1}{2}||Y - X\hat{\beta}^{\text{OLS}}||_2^2. \tag{47}$$

where $\hat{\beta}_j^{\text{OLS}} = \hat{\beta}_j^{\text{ls}}$. Hence finding the Lasso estimator amounts to finding the minimiser of

$$\frac{1}{2}(\hat{\beta}_k^{\text{OLS}} - \beta_k)^2 + \lambda|\beta_k|. \tag{48}$$

We write $\hat{\beta}$ for $\hat{\beta}_\lambda^L$ for simplicity. One could show that $|\hat{\beta}_k|$ is unique. By the KKT conditions,

$$\hat{\beta}_k^{\text{OLS}} - \hat{\beta}_k = \lambda\hat{v}_k \tag{49}$$

where $|\hat{v}_k| \leq 1$ and $\hat{v}_k = \text{sgn}(\hat{\beta}_k)$ if $\hat{\beta}_k \neq 0$. Thus $\hat{\beta}_k = 0$ when $|\hat{\beta}^{\text{OLS}}| \leq \lambda$. If $\hat{\beta}_k^{\text{OLS}} > \lambda$, $\hat{\beta}_k = \hat{\beta}_k^{\text{OLS}} - \lambda$. If $\hat{\beta}_k^{\text{OLS}} < -\lambda, \hat{\beta}_k = \hat{\beta}_k^{\text{OLS}} + \lambda$.

Question 11&12

Please see https://www.r-bloggers.com/ridge-regression-and-the-lasso/ for some R codes and explanations. See the following sample code

```
# Make swiss easier to access and type
swiss <- datasets::swiss

# Load ridge/lasso/lar packages
library(glmnet)
library(lars)

# Make x  the model matrix for glmnet
x <- model.matrix(Fertility~., swiss)[,-1]
```

```
# Get the response
y <- swiss$Fertility

lambda.value=5 # Choose a lambda value

# Fit the ridge and lasso models
ridge.fit = glmnet(x, y,alpha=0,lambda=lambda.value)
lasso.fit = glmnet(x, y,alpha=1,lambda=lambda.value)

# Do prediction with them
y.lasso<-predict(lasso.fit,newx=x,s=lambda.value)
y.ridge<-predict(ridge.fit,newx=x,s=lambda.value)

# Work out the MSE
lasso.MSE=mean((y.lasso-y)^2)
ridge.MSE=mean((y.ridge-y)^2)

# Fit the LAR and plot it
lars.fit=lars(x, y,type=c("lasso"))
plot(lars.fit)
```