STA 414S/2104S: Some notes for HW 1

1. the linear model

   (a)
   $$\ell(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)$$
   - since β is $p \times 1$, so is $\partial\ell/\partial\beta$

   $$\begin{aligned}
   \frac{\partial\ell}{\partial\beta} &= \frac{1}{\sigma^2}X^T(y - X\beta) \\
   \frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)^T(y - X\beta)
   \end{aligned}$$

   $$\left.\frac{\partial\ell}{\partial\beta}\right|_{\hat\beta,\hat\sigma^2} = 0, \qquad \left.\frac{\partial\ell}{\partial\sigma^2}\right|_{\hat\beta,\hat\sigma^2} = 0$$

   $$\begin{aligned}
   \Rightarrow \hat\beta &= (X^TX)^{-1}X^Ty \\
   \hat\sigma^2 &= \frac{1}{n}(y - X\hat\beta)^T(y - X\hat\beta)
   \end{aligned}$$

   Note that the maximum likelihood estimator of σ^2 does not adjust for estimation of β, and is not the usual unbiased estimator, $s^2 = (y - X\beta)^T(y - X\beta)/(n - p)$.

   (b) Assume that the first column of X is a column of 1s. Then, under H_0, $\hat\beta_{(0)} = (1^T1)^{-1}1^Ty = \bar y$ and $\hat\sigma^2_{(0)} = (1/n)\sum(y_i - \bar y)^2$. Then

   $$\begin{aligned}
   \ell(\hat\beta, \hat\sigma^2) - \ell(\hat\beta_{(0)}, \hat\sigma^2_{(0)}) &= -\frac{n}{2}\log\hat\sigma^2 - \frac{n}{2} + \frac{n}{2}\log\hat\sigma^2_{(0)} + \frac{n}{2} \\
   &= \frac{n}{2}\log\left(\frac{\hat\sigma^2_{(0)}}{\hat\sigma^2}\right), \\
   W &= n\log\left(\frac{\hat\sigma^2_{(0)}}{\hat\sigma^2}\right) = n\log\left\{\frac{\frac{1}{n}\sum(y_i - \bar y)^2}{\frac{1}{n}(y - X\hat\beta)^T(y - X\hat\beta)}\right\} \\
   &= n\log\left(\frac{\mathrm{SS}_{regr} + \mathrm{SS}_{resid}}{\mathrm{SS}_{resid}}\right) \\
   &= n\log\left(\frac{\mathrm{SS}_{regr}}{\mathrm{SS}_{resid}} + 1\right) \\
   &= n\log\left(\frac{p}{n - p + 1}F + 1\right)
   \end{aligned}$$

   as
   $$F = \frac{\mathrm{SS}_{regr}/p}{\mathrm{SS}_{resid}/(n - p + 1)}.$$

(c)

$$f(y \mid \beta)f(\beta) = \frac{1}{(\sqrt{2\pi})^{n+1}\sigma^n\tau} \exp -\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \exp -\frac{1}{2\tau}\beta^T\beta$$

The exponent can be expanded to

$$\begin{aligned}
-\frac{1}{2\tau}\beta^T\beta &- \frac{1}{2\sigma^2}(y^Ty + \beta^TX^TX\beta - 2\beta^TX^Ty) \\
= -\frac{1}{2\sigma^2}y^Ty &- \frac{1}{2\sigma^2}\beta^T(\frac{\sigma^2}{\tau^2}I + X^TX)\beta + \frac{2\beta^TX^Ty}{2\sigma^2} \\
= -\frac{1}{2\sigma^2}y^Ty &- \frac{1}{2\sigma^2}\beta^T(\lambda I + X^TX)\beta + \frac{2\beta^T(\lambda I + X^TX)^{-1}(\lambda I + X^TX)X^Ty}{2\sigma^2} \\
= -\frac{1}{2\sigma^2}y^Ty &- \frac{1}{2\sigma^2}\{\beta - (X^TX + \lambda I)^{-1}X^Ty\}^T(X^TX + \lambda I)\{\beta - (X^TX + \lambda I)^{-1}X^Ty\}
\end{aligned}$$

The first term will cancel with the denominator, as it doesn't depend on β. The second term is the exponent for a normal density with

$$\begin{aligned}
\text{mean} &= (X^TX + \lambda I)^{-1}X^Ty \\
\text{var} &= (X^TX + \lambda I)^{-1}\sigma^2.
\end{aligned}$$

A more careful derivation would make sure to get the expression in front of exp correct, which follows from detailed calculation, or from arguing that the resulting conditional density for β given y must integrate to 1.

(d) Actually, this was incorrectly stated, it's the mode/median of the posterior density that leads to the lasso estimator.

2. Thanks to Li Li

(a) Let $(\mathbf{1}, X)$ denote the $N \times (p+1)$ design matrix, $\tilde{X}$ denote the centered input matrix.
As we know, $(\mathbf{1}, X) = QR = (\frac{1}{\sqrt{N}}\mathbf{1}, Q_2)R = (\mathbf{1}, Q_2\tilde{R})$.
Therefore $X = Q_2\tilde{R}$.
X is included in the subspace spanned by Q_2.
Since $\tilde{X} = (x_1 - \bar{x}_p, ..., x_p - \bar{x}_p)$, $\tilde{X}$ is also in the subspace spanned by Q_2.
Moreover, we can find a $p \times p$ matrix P, such that $\tilde{X} = Q_2P$.
On the other hand, the SVD of $\tilde{X}$ has the form $\tilde{X} = UDV^T$.
Therefore $\tilde{X} = Q_2P = UDV^T$.
$Q_2PV = UDV^TV = UD$, since V is a $p \times p$ orthogonal matrix.
We assume that $\tilde{X}$ is nonsingular, so all the diagonal elements of D are greater than 0.
Then we have $Q_2PVD^{-1} = U$. That indicates that U is in the space spanned by Q_2.
Since U is a $N \times p$ orthogonal matrix, $\text{rank}(U) = \text{rank}(Q_2)$.
Therefore Q_2 and U span the same subspace.

(b) When are they the same, up to the sign flips?
$Q_2\tilde{R} - (\bar{x}_1\mathbf{1}, ..., \bar{x}_p\mathbf{1}) = UDV^T$.
if $Q_2 = \pm U$, $U(\tilde{R} \mp DV^T) = (\bar{x}_1\mathbf{1}, ..., \bar{x}_p\mathbf{1})$.
However, by the definition of U, it doesn't depend on $(\bar{x}_1\mathbf{1}, ..., \bar{x}_p\mathbf{1})$.
Therefore, we get two possibilities:
(1) $(\bar{x}_1\mathbf{1}, ..., \bar{x}_p\mathbf{1})$ degenerates to zero.
(2) $\tilde{R} = \pm DV^T$, *i.e.* $r_{ij} = d_i v_{ji}$ (see (*))
It is straightforward that V is a lower triangular matrix.
However, it is also an orthogonal matrix.
Therefore, it must be an diagonal matrix.
Moreover, it is an identity matrix, *i.e.* $V = \pm I$. Since $X^TX = VD^2V^T$, this implies that X^TX is diagonal, i.e. the columns of X are orthogonal.
Expanding on (2):

$$\begin{aligned} \pm U\tilde{R} &= Q_2\tilde{R} = UDV^T + (\bar{x}_1\mathbf{1}, \ldots, \bar{x}_p\mathbf{1}) \\ \Rightarrow \quad & U(\tilde{R} \mp DV^T) = (\bar{x}_1\mathbf{1}, \ldots, \bar{x}_p\mathbf{1}) \\ \Rightarrow \quad & U^TU(\tilde{R} \mp DV^T) = U^T(\bar{x}_1\mathbf{1}, \ldots, \bar{x}_p\mathbf{1}) \qquad (*), \end{aligned}$$

The first term is 0 because $(\bar{x}_1\mathbf{1}, \ldots, \bar{x}_p\mathbf{1})$ is not in the space spanned by U, so we must have

$$\tilde{R} \mp DV^T = 0.$$

3. (a) The code to do this was given in the hints. Note that `set.seed(123)` or something similar will ensure that you can reproduce the results from `sample` each time. Otherwise you will get different training samples each time you run the command `sample(1599, 1000)`.

   (b) Fairly standard, as is

   (c) .

   (d) Most people found test error to be around 0.43, and that best subset regression or sometimes ridge regression worked nearly as well as anything.

   (e) I was worried that with a smallish test set, and most y's in the $(5, 6)$ range, that this would inflate test errors even more than might be expected, but this didn't materialize.

4. Quite straightforward using 1(a).

5. (a) Thanks to Pak Ho: the approximate 95% confidence interval for $\hat{f}(x_0)$ is

$$\begin{aligned} \{(\hat{\beta}_0 &+ \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3) - 1.96\hat{\sigma}_0, \\ (\hat{\beta}_0 &+ \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3) + 1.96\hat{\sigma}_0\} \end{aligned}$$

where

$$\hat{\sigma}_0^2 = \begin{pmatrix} 1 & x_0 & x_0^2 & x_0^3 \end{pmatrix} (X^TX)^{-1} \begin{pmatrix} 1 \\ x_0 \\ x_0^2 \\ x_0^3 \end{pmatrix}.$$

These are called pointwise confidence bands, when plotted as a function of x_0, because the guarantee of 95% coverage is only valid at each single point, and not for the whole function.

(b) The confidence set $C_\beta = \{\beta \mid (\hat{\beta} - \beta)^T X^T X (\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi_4^{2(0.95)}\}$ gives a set of $\beta \in \mathbb{R}^4$ with approximately 95% confidence. Thus for all 4-vectors a, the set of $a^T\beta$ from $\beta \in C_\beta$ is a confidence interval for $a^T\beta$. Since this is true for the choice $a = (1, x_0, x_0^2, x_0^3)$ for any choice of x_0, the resulting confidence band about $\hat{f}(x_0)$ has simultaneous confidence 0.95. The only way I can think of to compute this confidence band is to simulate β's from the normal distribution with mean $\hat{\beta}$ and covariance $\hat{\sigma}^2(X^TX)^{-1}$, and draw the fitted function with these simulated βs. Most people did this with light gray lines, and the band that is eventually filled in is the simultaneous confidence band for $f(x_0)$. This band is wider than the one for (a), because it's guarantee of 95% is for the function, instead of for each function value $f(x_0)$.

The "approximate" 0.95 in (a) and (b) is because σ^2 is estimated, so under normality the exact distribution is t or F, not normal or χ^2.