Note - eq 3.47 One line isn't so clear, but

$$
\begin{aligned}
(VD^2V^T + \lambda I)(VD^2V^T + \lambda I)^{-1} &= I \\
\Rightarrow (VD^2V^T + \lambda I)(VD^2V^T + \lambda I)^{-1}V &= V \\
\Rightarrow (D^2V^T + \lambda V^T)(VD^2V^T + \lambda I)^{-1}V &= I \\
\Rightarrow (D^2 + \lambda I)V^T(VD^2V^T + \lambda I)^{-1}V &= I
\end{aligned}
$$

and so

$$V^T(VD^2V^T + \lambda I)^{-1}V = (D^2 + \lambda I)^{-1}.$$

3.1 Recall the method of using repeated simple linear regression to do multiple lienar regression: We form an orthogonal spanning set $z_1, \ldots, z_p$ of the column space in such a way that $z_1, \ldots, z_k$ spans the span of the first k columns of X. We finally regress y onto z_p to obtain $\hat{\beta}_p$. But we can do more: If $\gamma_j = \langle y, z_j \rangle$, then $\hat{y} = \sum_j \frac{\gamma_j}{\|z_j\|^2} z_j$ is the orthogonal projection onto the column space of X in the full rank case. What's more, $\sum_{j<p} \frac{\gamma_j}{\|z_j\|^2} z_j$ is the result of using only the first $p-1$ features. Let $r_p = y - \hat{y}$. Recall that $\hat{\beta}_p = \frac{\gamma_p}{\|z_p\|^2}$, and so

$$
\begin{aligned}
\mathrm{RSS}_0 &= \left\| y - \sum_{j<p} \frac{\gamma_j}{\|z_j\|^2} z_j \right\|^2 \\
&= \left\| r_p + \hat{\beta}_p z_p \right\|^2 \\
&= \|r_p\|^2 + \hat{\beta}_p^2 \|z_p\|^2 \\
\mathrm{RSS}_1 &= \|y - \hat{y}\|^2 \\
&= \|r_p\|^2
\end{aligned}
$$

So we can express the formula for the F-score

$$
\begin{aligned}
F &= \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1/(N-p-1)} \\
&= \frac{\hat{\beta}_p^2 \|z_p\|^2}{\hat{\sigma}^2}
\end{aligned}
$$

Now, write $X = Z\Gamma$, where Z's columns are the z_i's and Γ is upper triangular with diagonal entries 1. Then $(X^TX)^{-1} = (\Gamma^T Z^T Z \Gamma)^{-1} = (\Gamma^T D \Gamma)^{-1}$. Here, D is diagonal on $(\|z_1\|^2, \ldots, \|z_p\|^2)$. Since we assume that X^TX is invertible, both Γ and D are, so $(\Gamma^T D \Gamma)^{-1} = \Gamma^{-1} D^{-1} \Gamma^{-T}$. Now, we can compute the lower-right entry

$$
\begin{aligned}
(\Gamma^{-1}D^{-1}\Gamma^{-T})_{pp} &= \sum_{ij}(\Gamma^{-1})_{ip}D^{-1}_{ij}(\Gamma^{-T})_{jp} \\
&= \sum_i \frac{1}{\|z_i\|^2}(\Gamma^{-1})_{ip}(\Gamma^{-T})_{ip} \\
&= \frac{1}{\|z_p\|^2}(\Gamma^{-1}_{pp})^2 \\
&= \frac{1}{\|z_p\|^2}
\end{aligned}
$$

and so we finally find that the F-score can be written as

$$
\frac{\hat{\beta}_p^2}{\hat{\sigma}^2 d_p}
$$

where d_p is the pth diagonal entry of $(X^TX)^{-1}$. By relabelling the features, we find that this also works for any of the features, and so the F-score is the square of the Z-score.

2 I would expect that the pointwise confidence intervals would be narrower. Here is a heuristic argument: Suppose that β_1, β_2 are normally distributed with mean 0 and covariance I. Then one choice of a 95% confidence set is a circle with radius $R_{95} = \log(20) \approx 3$ about the origin. On the other hand, a 95% confidence set for just β_0 is an interval of radius about 1.64, so the 2D set will have much more extreme values of β_0. In paricular, the band for $x_0 = 0$ will be much wider for the 2D confidence region. Another way to see it: If I generate samples for each point individually, I will implicitly be drawing samples from many, many more functions when I make the final plot, and so will have a narrower gap.

Simulations in the associated notebook.

3.3.a The variance of c^Ty is

$$
\begin{aligned}
&\mathrm{Var}\left(\sum_i c_i y_i\right) \\
&= \sum_{ij} c_i c_j \,\mathrm{Cov}(y_i, y_j) \\
&= \sum_i c_i^2 \sigma^2 \\
&= \sigma^2 \|c\|^2
\end{aligned}
$$

while the fact that the estimator is unbiased is expressed exactly by the equation $(X^Tc)^T\beta = \alpha^T\beta$. Now, I'm going to make a small assumption. We know that the estimator is not supposed to depend on β, which is unobservable, and I'm going to interpret this mathematically by taking $(X^T c)^T\beta = \alpha^T\beta$ to hold for *all* β. (This is certainly true of the OLS estimator, and it seems that an estimator that did not satisfy this property would be rather useless.)

Write $c = kX(X^TX)^{-1}\alpha + v$ for some constant k and some v orthogonal to $X(X^TX)^{-1}\alpha$, we find $X^Tc = k\alpha + X^Tv$ and taking $\beta = (X^TX)^{-1}\alpha$ gives

$$\begin{aligned}\beta^T(X^Tc) &= ((X^TX)^{-1}\alpha)^T(k\alpha + X^Tv)\\ &= k\alpha^T(X^TX)^{-1}\alpha + \alpha^T(X^TX)^{-1}X^Tv\\ &= k\alpha^T(X^TX)^{-1}\alpha\end{aligned}$$

on one hand, and on the other we know that

$$\begin{aligned}\beta^T(X^Tc) &= \beta^T\alpha\\ &= \alpha^T(X^TX)^{-1}\alpha\end{aligned}$$

and so $k = 1$. This means that

$$\sigma^2\|c\|^2 = \sigma^2\|X(X^TX)^{-1}\alpha\|^2 + \sigma^2\|v\|^2 \geq \sigma^2\alpha^T(X^TX)^{-1}\alpha,$$

which is the variance of the OLS estimator for $\alpha^T\beta$.

3.3.b A matrix A is positive-semidefinite iff $x^TAx \geq 0$ for all $x \in \mathbb{R}^n$. If $a \in \mathbb{R}^n$, then $a^T(\tilde{V} - \hat{V})a = a^T\tilde{V}a - a^T\hat{V}a$. Since $\hat{V}, \tilde{V}$ are the covariance matrices of $\hat{\beta}$ and $\tilde{\beta}$, respectively, we find that

$$\begin{aligned}a^T\tilde{V}a &= \mathrm{Var}(a^T\tilde{\beta})\\ a^T\hat{V}a &= \mathrm{Var}(a^T\hat{\beta}).\end{aligned}$$

So by the above result, $\tilde{V} - \hat{V}$ is positive semidefinite.

3.4 In the full-rank case, a single pass of the Graham-Schmidt procedure expresses $X = QR$ where Q is orthogonal and R is square and upper triangular with 1's along the diagonal. The formula

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

becomes

$$\begin{aligned}&(R^TQ^TQR)^{-1}R^TQ^Ty\\ &= (R^TR)^{-1}R^TQ^Ty\\ &= R^{-1}R^{-T}R^TQ^Ty\\ &= R^{-1}Q^Ty\end{aligned}$$

We can compute R^{-1} during the Graham-Schmidt process as follows: Let Q_k, R_k denote the matrices formed by taking the first k columns of Q, R. We will also store the matrix R_k^{-1}. Given this data, the next step in the Graham-Schmidt process yields Q_{k+1}, R_{k+1}. We update R_{k+1}^{-1} as follows: If

$$R_{k+1} = \left(\begin{array}{c|c} R_k & Z \\ \hline 0 & a_k \end{array}\right),$$

then

$$R_{k+1}^{-1} = \left(\begin{array}{c|c} R_k^{-1} & -a_k R_k^{-1} Z \\ \hline 0 & a_k^{-1} \end{array}\right)$$

So we can store the coefficients of $\hat{\beta}$ in a list that is updated with each newly discovered column of R^{-1} and row of Q^T.

3.5 The original ridge objective is

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Rearranging it, we get

$$\sum_{i=1}^{N}\left(y_i - [\beta_0 + \sum_{j=1}^{p} \overline{x}_j\beta_j] + \sum_{j=1}^{p}(x_{ij} - \overline{x}_j)\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

So the minimum $\hat{\beta}_\lambda^{ridge}$ to this will give a minimum to

$$\sum_{i=1}^{N}\left(y_i - \beta_0^c + \sum_{j=1}^{p}(x_{ij} - \overline{x}_j)\beta_j^c\right)^2 + \lambda \sum_{i=1}^{p} (\beta_j^c)^2,$$

where we simply set $\beta_0^c := \beta_0 + \sum_{j=1}^{p} \overline{x}_j\beta_j$ and $\beta_j^c := \hat{\beta}_j$ if $j \geq 1$. The lasso condition is exactly the same. In both cases, the objective function is differentiable with respect to β_0, and the derivative in the lasso and ridge case is

$$\begin{aligned} &\sum_{i=1}^{n} 2(y_i - \beta_0^2 + \sum_{j=1}^{p}(x_{ij} - \overline{x}_j)\beta_j^c) \\ =&N\overline{y} - N\beta_0^c, \end{aligned}$$

and so $\beta_0^c = \overline{y}$ at the minimum in each case.

3.6 Let f_B be $N(0, \tau I)$. Then

$$f_{B|Y}(\beta|y) = \frac{f_{BY}(\beta, y)}{f_Y(y)} = \frac{f_{Y|B}(y|\beta)f_B(\beta)}{f_Y(y)}.$$

We are working under the assumption that y is Gaussian noise added onto $X\beta$, so

$$f_{Y|B}(y|\beta) = \frac{1}{(2\pi\sigma)^{\frac{p}{2}}} e^{-\frac{1}{2\sigma^2}\|y-X\beta\|^2},$$

while

$$f_B(\beta) = \frac{1}{(2\pi\tau)^{\frac{p}{2}}} e^{-\frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2}.$$

and so

$$f_{Y|B}(y|\beta)f_B(\beta) = \frac{1}{2\pi(\tau\sigma)^{\frac{p}{2}}} \exp\left\{\frac{-1}{2}\left(\frac{\|y-X\beta\|^2}{\sigma^2} + \frac{\|\beta\|^2}{\tau^2}\right)\right\}.$$

Since this is quadratic in β, we find that this posterior distribution is also Gaussian, so the median and mode match. We'll find the value by taking differentiating with repsect to β_K:

$$f_{Y|B}(y|\beta)f_B(\beta)\left[\frac{1}{\sigma^2}\sum_i (y_i - \sum_j x_{ij}\beta_j)(x_{iK}) - \frac{1}{\tau^2}\beta_K\right]$$

Setting this equal to zero for each component gives the vector equation

$$\frac{1}{\sigma^2}X^T(y - X\beta) = \frac{1}{\tau^2}\beta,$$

and solving that yields

$$\beta = (X^TX + \frac{\sigma^2}{\tau^2}I)^{-1}X^Ty,$$

which is the ridge regression solution for which $\lambda = \frac{\sigma^2}{\tau^2}$.

3.7*** I'm not sure what this problem is asking about. If it's the same scenario as the previous problem, we reuse that work to find that the posterior density is

$$\frac{C}{2\pi(\tau\sigma)^{\frac{p}{2}}} \exp\left\{\frac{-1}{2}\left(\frac{\|y-X\beta\|^2}{\sigma^2} + \frac{\|\beta\|^2}{\tau^2}\right)\right\},$$

with C chosen to make the expression integrate to 1 as a function of y. We take the log and get something that's is proportional to the desired quantity, but with an extra constant added.

3.8 Let $\mathbf{1}, X_1, \ldots, X_p$ denote the columns of the unnormalized matrix X. The matrix Q in the QR-decomposition can be found via the Graham-Schmidt process. Since the first column of X is all 1's, the GS process starts by making the first column of Q $\frac{1}{\sqrt{N}}\mathbf{1}$. The other columns are orthogonalized with respect to this vector:

$$X_i \mapsto \tilde{X}_i = X_i - \langle X_i, \frac{1}{\sqrt{N}}\mathbf{1}\frac{1}{\sqrt{N}}\mathbf{1}\rangle = X_i - \overline{X_i}\mathbf{1}.$$

Since the remaining columns of Q are formed by finding an orthonomal basis of the span of these vectors, which are none other than $\widetilde{X}$'s columns, and the columns of the matrix U also form an orthogonal basis of that span, we know that the span of Q_2 is the same as the span of U.

When will the two match? If they match, we can express $\widetilde{X} = Q_2 R_2$ and also $\widetilde{X} = UDV^T$. Assuming $Q_2 = U$, multiplying both of these expressions by U^T will yield $R_2 = DV^T$, which means that $V = R_2^T D^{-1}$ is a lower triangular orthonormal matrix, which means it's diagonal with entries ± 1. This, in turn, means that $R_2 = VD$ is diagonal, hence the columns of $\widetilde{X}$ are already orthogonal and simply need to be rescaled to become orthonormal.

Conversely, if the columns of $\widetilde{X}$ are orthogonal, then scaling each of them to have unit norm gives an orthonomal matrix U. Letting D be the diagonal matrix whose entries are the lengths of the columns of $\tilde{X}$, we have $\tilde{X} = UD$ is both the QR and SVD decomposition of $\tilde{X}$, so the first matrix in each decomposition is the same.

3.9 Recall that if $X_1 = QR$ is the QR-decomposition, then the predictions are given by $\hat{y} = QQ^T y$. Letting $Q_1, \ldots, Q_q$ denote the columns of Q, we find

$$\begin{aligned}
\hat{y}_i &= \sum_{ab} Q_{ia} Q^T_{ab} y_b \\
&= \sum_a Q_{ia} \langle Q_a, y \rangle
\end{aligned}$$

so

$$\hat{y} = \sum_a Q_a \langle Q_a, y \rangle.$$

Now, suppose we wish to consider the effect of adding a new column vector to X_1. The new Q-matrix in the QR-decomposition of X_1 with this new column is found by simply doing another stage of the Graham-Schmidt process. That is, if C_j is the column vector corresponding to the feature j to be added, the new column in the QR-decomp will found by taking

$$\overline{Q}_j = C_j - \sum_{a=1}^{q} Q_a \langle Q_a, C_j \rangle,$$

and then $Q_j = \frac{\overline{Q}_j}{\|\overline{Q}_j\|}$. The new prediction will then become

$$\hat{y}^{(j)} = Q_j \langle Q_j, y \rangle + \sum_a Q_a \langle Q_a, y \rangle$$

which gives an RSS of

$$\begin{aligned}
\|\hat{y}^{(j)} - y\|^2 &= \|Q_j \langle Q_j, y \rangle + \hat{y} - y\|^2 \\
&= \|Q_j \langle Q_j, r + \hat{y} \rangle - r\|^2 \\
&= \|Q_j \langle Q_j, r \rangle - r\|^2,
\end{aligned}$$

where the last line follows because $\hat{y}$ is in the span of $Q_1, \ldots, Q_q$, which are orthogonal to Q_j.

This means that to find the feature that reduces the RSS the most, we simply have to do the following for each new feature:

- Calculate the orthogonal projection of C_j onto the span of $Q_1, \ldots, Q_q$ and normalize it to obtain the vector Q_j.
- Calculate the square norm of the residual of r projected onto Q_j

Then choose the feature giving the smallest value for the second step.

Remark Looking online at some discussions of forward stepwise regression, many say that they pick the feature that has the highest t-score or F-score. Recall that exercise 1 shows the square of the t-score is the F-score, and the F-score is defined by

$$F = \left(\frac{\text{RSS}_{small}}{\text{RSS}_{small+1}} - 1 \right) (N - (q+1)),$$

and so maximizing this or the t-score is equivalent to minimizing the new RSS.

3.10 Since exercise 1 establishes the square of the z-score is the F-score for one feature and

$$F = \left(\frac{\text{RSS}_{small} - \text{RSS}_{big}}{\text{RSS}_{big}} \right) (N - p),$$

we find that the feature with the smallest F-score will have the least impact on RSS.

3.11 We have the following matrices of the given dimensions:

- Y - $N \times K$
- X - $N \times p$
- B - $p \times K$
- Σ - $K \times K$

and we want to minimize the quantity

$$\sum_i (y_i - f(x_i))\Sigma^{-1}(y_i - f(x_i))^T,$$

where y_i and x_i are the ith rows of Y and X, respectively. If A is a matrix, let A_{*b} and A_{a*} denote the bth column and ath row. We have $f(x_i) = (XB)_{i*} =$

$x_i B$. Taking the partial derivative with respect to B_{ab}, we obtain

$$\begin{aligned}
&\sum_i (y_i - x_i B)\Sigma^{-1}(-(0,\ldots,0,\underbrace{X_{ia}}_{\text{index } b},0,\ldots,0)^T) \\
&+(-(0,\ldots,0,\underbrace{X_{ia}}_{\text{index } b},0,\ldots,0)\Sigma^{-1}(y_i - x_i B)^T \\
&= -\sum_i (y_i - x_i B)(\Sigma^{-1})_{*b} X_{ia} + X_{ia}(\Sigma^{-1})_{b*}(y_i - x_i B)^T \\
&= -\left(\sum_i (y_i - x_i B) X_{ia}\right)(\Sigma^{-1})_{*b} - (\Sigma^{-1})_{b*}\left(\sum_i X_{ia}(y_i - x_i B)^T\right) \\
&= -\left(X_{*a}^T (Y - XB)\right)(\Sigma^{-1})_{*b} - (\Sigma^{-1})_{b*}\left((Y - XB)^T X_{*a}\right) \\
&= -2X_{*a}^T (Y - XB)(\Sigma^{-1})_{*b}
\end{aligned}$$

If this quantity is zero for all a, b, then

$$0 = X^T (Y - XB)\Sigma^{-1},$$

and we may cancel Σ and solve as usual to find the solution $B = (X^T X)^{-1} X^T Y$.

Furthermore, the value of $\frac{\partial^2}{\partial B_{ab}\partial B_{st}}$ is $2(X^T X)_{as}(\Sigma^{-1})_{bt}$. Hence if $B = \sum_{cd} B_{cd} E_{cd}$ is a test "vector", we find that $v^T H v$ has the value

$$\begin{aligned}
&\sum_{abst} B_{ab} 2(X^T X)_{as}(\Sigma^{-1})_{bt} B_{st} \\
=&2\sum_{as}(X^T X)_{as}\sum_{bt} B_{ab}(\Sigma^{-1})_{bt} B_{st} \\
=&2\mathrm{trace}(X^T X(B^T \Sigma^{-1} B)) \\
=&2\mathrm{trace}(X B^T \Sigma^{-1} B X^T) \\
=&2\sum_a \sum_{ij} (XB^T)_{ai}(\Sigma^{-1})_{ij}(BX^T)_{ja} \\
>&0,
\end{aligned}$$

since Σ^{-1} is positive definite.

As for what happens when Σ is not constant, let's take the following simple case: we have observations $(x_1, y_1)\ldots,(x_N, y_N)$ forming data matrix X and observation matrix Y. We then run another round of observations on the same inputs to obtain $(x_1, y_1'),\ldots,(x_N, y_N')$, yielding X, Y'. On the first round of observations, we had Σ_1 for the error correlations, and Σ_2 on the second round. We find that the condition for the partial derivatives to vanish is

$$0 = X^T (Y - XB)\Sigma_1^{-1} + X^T (Y' - XB)\Sigma_2^{-1},$$

which we can solve to obtain

$$B = (X^T X)^{-1} X^T (Y\Sigma_1^{-1} + Y'\Sigma_2^{-1})(\Sigma_1^{-1} + \Sigma_1^{-1})^{-1}.$$

Meanwhile, the formula from the first part of the problem uses a data matrix with two copies of X stacked on top of each other and gives

$$B = (2X^T X)^{-1} X^T (Y + Y'),$$

and so the two will match iff (after clearing the leading $X^T X$ term and multiplying by $\Sigma_1^{-1} + \Sigma_2^{-1}$)

$$\frac{1}{2} X^T (Y + Y')(\Sigma_1^{-1} + \Sigma_2^{-1}) = X^T (Y\Sigma_1^{-1} + Y'\Sigma_2^{-1}),$$

which is equivalent to

$$X^T (Y'\Sigma_1^{-1} + Y\Sigma_2^{-1}) = X^T (Y\Sigma_1^{-1} + Y'\Sigma_2^{-1}).$$

For a concrete example, we take

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}, \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3} \end{bmatrix},$$

$$X = [1]$$

$$Y = \begin{bmatrix} 1 & 0 \end{bmatrix}, Y' = \begin{bmatrix} 0 & 1 \end{bmatrix},$$

and the two sides of the equation are $\begin{bmatrix} 1 & 2 \end{bmatrix}$ and $\begin{bmatrix} 1 & 3 \end{bmatrix}$. Indeed, the coefficients that ignore Σ give $B = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix}$, simply averaging the observatinos, but taking the varying Σ into account gives $B = \begin{bmatrix} \frac{1}{2} & \frac{3}{5} \end{bmatrix}$, a result that puts more weight on the observation with smaller variance.