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

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

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 $(X^Tc)^T\beta = \alpha^T\beta$ 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

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

then

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

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 $\lambda = \frac{\sigma^2}{\tau^2}$.