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 $z_1, \dots, z_p$ of the column space in such a way that $z_1, \dots, z_k$ 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 $(\|z_1\|^2, \dots, \|z_p\|^2)$. 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 $(X^T X)^{-1}$. 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.

3.3.a The variance of $c^T y$ 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 $(X^T c)^T \beta = \alpha^T \beta$. 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 $c = kX(X^T X)^{-1}\alpha + v$ for some constant k and some v orthogonal to $X(X^T X)^{-1}\alpha$, we find $X^T c = k\alpha + X^T v$ and taking $\beta = (X^T X)^{-1}\alpha$ gives

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

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 $\alpha^T\beta$.