Test 3

JARED YU

1. Consider the multiple regression model for n y-data $y_1, \cdots, y_n$,

$$\mathbf{y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = (y_1, \cdots, y_n)'$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $Var(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are fixed, and each element of the vector $\mathbf{y}$ is normally distributed. $\mathbf{B}_1$ and $\mathbf{B}_2$ are vectors of regression coefficients (i.e., $\mathbf{B}_1$ has at least two regression coefficients, so does $\mathbf{B}_2$) under estimation. Assume that σ^2 and $\mathbf{V}$ are known.

Construct a test statistic for the hypotheses

$$H_0: \mathbf{B}_1 = \mathbf{0} \text{ versus } H_1: \mathbf{B}_1 \neq \mathbf{0}.$$

1) Derive the best linear unbiased estimator for the regression coefficients vector $\mathbf{B} = [\mathbf{B}_1, \mathbf{B}_2]'$ and prove that your estimator is indeed the best linear unbiased estimator.

Ans:

Since $\sigma^2\mathbf{V}$ is the variance-covariance matrix, then $\mathbf{V}$ must be positive definite and nonsingular. There exists an $n \times n$ nonsingular symmetric matrix $\mathbf{K}$ such that $\mathbf{K}^\top\mathbf{K} = \mathbf{K}\mathbf{K} = \mathbf{V}$. We can then define a new set of variables as follows:

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \mathbf{W}_1 = \mathbf{K}^{-1}\mathbf{X}_1, \mathbf{W}_2 = \mathbf{K}^{-1}\mathbf{X}_2, \mathbf{u} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}.$$

The original model is

$$\mathbf{y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \boldsymbol{\varepsilon}$$

and we multiply $\mathbf{K}^{-1}$ from the left on both sides so that it becomes

$$\rightarrow \mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{K}^{-1}\mathbf{X}_2\mathbf{B}_2 + \mathbf{K}^{-1}\boldsymbol{\varepsilon}$$

$$\rightarrow \mathbf{z} = \mathbf{W}_1\mathbf{B}_1 + \mathbf{W}_2\mathbf{B}_2 + \mathbf{u}$$

Next, the expectation and variance for this transformed model will be shown,

$$E(\mathbf{u}) = E(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}E(\boldsymbol{\varepsilon}) = \mathbf{0}$$

$$Var(\mathbf{u}) = Var(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}Var(\boldsymbol{\varepsilon})(\mathbf{K}^{-1})^\top = \mathbf{K}^{-1}(\sigma^2\mathbf{V})\mathbf{K}^{-1} = \sigma^2\mathbf{K}^{-1}(\mathbf{K}\mathbf{K})\mathbf{K}^{-1} = \sigma^2\mathbf{I}$$

The transformation has made it such that the model is mean zero, constant variance, and the response variable is uncorrelated. So, it is now possible to perform least-squares, since it fits the basic required assumptions.

$$S\left(\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{bmatrix}\right) = (\mathbf{z} - \mathbf{W}_1\mathbf{B}_1 - \mathbf{W}_2\mathbf{B}_2)^\top(\mathbf{z} - \mathbf{W}_1\mathbf{B}_1 - \mathbf{W}_2\mathbf{B}_2)$$

$$= (\mathbf{K}^{-1}\mathbf{y} - \mathbf{K}^{-1}\mathbf{X}_1\mathbf{B}_1 - \mathbf{K}^{-1}\mathbf{X}_2\mathbf{B}_2)^\top(\mathbf{K}^{-1}\mathbf{y} - \mathbf{K}^{-1}\mathbf{X}_1\mathbf{B}_1 - \mathbf{K}^{-1}\mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)^\top(\mathbf{K}^{-1})^\top\mathbf{K}^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)^\top\mathbf{K}^{-1}\mathbf{K}^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)^\top(\mathbf{K}\mathbf{K})^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)^\top\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y}^\top - \mathbf{X}_1^\top\mathbf{B}_1^\top - \mathbf{X}_2^\top\mathbf{B}_2^\top)\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= (\mathbf{y}^\top\mathbf{V}^{-1} - \mathbf{X}_1^\top\mathbf{B}_1^\top\mathbf{V}^{-1} - \mathbf{X}_2^\top\mathbf{B}_2^\top\mathbf{V}^{-1})(\mathbf{y} - \mathbf{X}_1\mathbf{B}_1 - \mathbf{X}_2\mathbf{B}_2)$$

$$= \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}_1^\top\mathbf{B}_1^\top\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}_2^\top\mathbf{B}_2^\top\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_1^\top\mathbf{B}_1^\top\mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2^\top\mathbf{B}_2^\top\mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1$$
$$- \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 + \mathbf{X}_1^\top\mathbf{B}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 + \mathbf{X}_2^\top\mathbf{B}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2$$

$$\frac{\partial S}{\partial \mathbf{B}_1} = -2\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + 2\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2$$

$$\frac{\partial S}{\partial \mathbf{B}_2} = -2\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + 2\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2$$

$$\frac{\partial S}{\partial \mathbf{B}_1} \overset{\text{set to}}{=} \mathbf{0}$$

$$\rightarrow -2\mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + 2\mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 = \mathbf{0}$$

$$\rightarrow \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 = \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} \tag{1}$$

$$\frac{\partial S}{\partial \mathbf{B}_2} \overset{\text{set to}}{=} \mathbf{0}$$

$$\rightarrow -2\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} + 2\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + 2\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 = \mathbf{0}$$

$$\rightarrow \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 = \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} \tag{2}$$

The next step is to solve for the systems of equations.

$$\rightarrow \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} \end{bmatrix}$$

$$\rightarrow \boxed{\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} \end{bmatrix}}$$

The above shows the derivation for the least squares estimates of $\mathbf{B}_1$ and $\mathbf{B}_2$. The next step is to show that they have the BLUE properties. To do this, it will first be shown that they are unbiased.

$$E\left(\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}\right) = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}E(\mathbf{y}) \\ \mathbf{X}_2^\top \mathbf{V}^{-1}E(\mathbf{y}) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} E(\mathbf{y})$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} [\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2]$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2\mathbf{B}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$$

The above shows that $\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}$ is an unbiased estimator of $\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$, since $E\left(\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}\right) = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$. The next step will show the variance of this unbiased estimator.

$$Var\left(\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}\right) = Var\left\{\begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} \mathbf{y}\right\}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} Var(\mathbf{y}) \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix}^\top \left(\begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1}\right)^\top$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} (\sigma^2\mathbf{V}) \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix}^\top \left(\begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^\top\right)^{-1}$$

$$= \sigma^2 \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{V} \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{V} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix}^\top \left(\begin{bmatrix} (\mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1)^\top & (\mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2)^\top \\ (\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1)^\top & (\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2)^\top \end{bmatrix}\right)^{-1}$$

$$= \sigma^2 \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{bmatrix} \mathbf{V}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2] \left(\begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 \\ \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \right)^{-1}$$

$$= \sigma^2 \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \left(\begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 \\ \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \right)^{-1}$$

$$= \sigma^2 \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 \\ \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1}$$

It must be shown that $\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}$, has the minimum variance for any linear combination of the estimated coefficients. We already know that they are unbiased. So, it must now be shown that they are the best linear estimator, which means that $\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}$ can minimize the variance for any linear combination of the estimated coefficients.

Let $\boldsymbol{\ell}_1$ and $\boldsymbol{\ell}_2$ be arbitrary vectors of constants, then the linear combination of the estimated coefficients can be denoted as

$$(\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix} = \boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2.$$

It has the following variance,

$$Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) = \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 \\ \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix}$$

Let $\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix}$ be some other estimator of $\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$ and it can be written as,

$$\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix} = \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \right\} \mathbf{y} + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix},$$

where $\mathbf{C}_1$, $\mathbf{C}_2$, are an $r \times n$ and $(p - r) \times n$ matrix while $\mathbf{d}_1$ and $\mathbf{d}_2$, $r \times 1$, and $(p - r) \times 1$ vector respectively. We can next look at the biasedness of this estimator:

$$E\left(\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix} \right) = \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \right\} E(\mathbf{y}) + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

$$= \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \right\} [\mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2] + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} [\mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2] + \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} [\mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2 \mathbf{B}_2] + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \mathbf{B}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 \mathbf{B}_1 + \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \mathbf{X}_1 \mathbf{B}_1 + \mathbf{C}_1 \mathbf{X}_2 \mathbf{B}_2 \\ \mathbf{C}_2 \mathbf{X}_1 \mathbf{B}_1 + \mathbf{C}_2 \mathbf{X}_2 \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \mathbf{X}_1 \mathbf{B}_1 + \mathbf{C}_1 \mathbf{X}_2 \mathbf{B}_2 \\ \mathbf{C}_2 \mathbf{X}_1 \mathbf{B}_1 + \mathbf{C}_2 \mathbf{X}_2 \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1 \mathbf{X}_1 & \mathbf{C}_1 \mathbf{X}_2 \\ \mathbf{C}_2 \mathbf{X}_1 & \mathbf{C}_2 \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$$

It can be seen here that $\begin{bmatrix}\widetilde{\mathbf{B}}_1\\ \widetilde{\mathbf{B}}_2\end{bmatrix}$ is unbiased only in the case that $\mathbf{C}_1\mathbf{X}_1$, $\mathbf{C}_1\mathbf{X}_2$, $\mathbf{C}_2\mathbf{X}_1$, and $\mathbf{C}_2\mathbf{X}_2$ are matrices with only zeroes, while $\mathbf{d}_1$ and $\mathbf{d}_2$ are zero vectors.

We can now consider the case where $\begin{bmatrix}\widetilde{\mathbf{B}}_1\\ \widetilde{\mathbf{B}}_2\end{bmatrix}$ is unbiased. In such a case, the variance will be as follows:

$$
\begin{aligned}
Var\left(\begin{bmatrix}\widetilde{\mathbf{B}}_1\\ \widetilde{\mathbf{B}}_2\end{bmatrix}\right) &= Var\left(\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\right\}\mathbf{y}+\begin{bmatrix}\mathbf{d}_1\\ \mathbf{d}_2\end{bmatrix}\right)\\
&= Var\left(\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\right\}\mathbf{y}\right)\\
&= \left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}\right.\\
&\qquad \left.+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\right\}Var(\mathbf{y})\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\right\}^\top\\
&= \left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\right\}\sigma^2\mathbf{V}\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}^\top\left(\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\right)^\top\right.\\
&\qquad \left.+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}^\top\right\}\\
&= \sigma^2\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}\mathbf{V}\right.\\
&\qquad \left.+\begin{bmatrix}\mathbf{C}_1\\ \mathbf{C}_2\end{bmatrix}\mathbf{V}\right\}\left\{\begin{bmatrix}\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}+\begin{bmatrix}\mathbf{C}_1^\top & \mathbf{C}_2^\top\end{bmatrix}\right\}\\
&= \sigma^2\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{V}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\right.\\
&\qquad +\begin{bmatrix}\mathbf{C}_1\mathbf{V}\\ \mathbf{C}_2\mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\\
&\qquad \left.+\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{V}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{C}_1^\top & \mathbf{C}_2^\top\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\mathbf{V}\\ \mathbf{C}_2\mathbf{V}\end{bmatrix}\begin{bmatrix}\mathbf{C}_1^\top & \mathbf{C}_2^\top\end{bmatrix}\right\}\\
&= \sigma^2\left\{\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\right.\\
&\qquad +\begin{bmatrix}\mathbf{C}_1\mathbf{V}\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{C}_1\mathbf{V}\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{C}_2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{C}_2\mathbf{V}\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\\
&\qquad \left.+\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{C}_1^\top & \mathbf{X}_1^\top\mathbf{C}_2^\top\\ \mathbf{X}_2^\top\mathbf{C}_1^\top & \mathbf{X}_2^\top\mathbf{C}_2^\top\end{bmatrix}+\begin{bmatrix}\mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top\\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top\end{bmatrix}\right\}
\end{aligned}
$$

$$\begin{aligned}
&= \sigma^2 \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{C}_1\mathbf{X}_1 & \mathbf{C}_1\mathbf{X}_2 \\ \mathbf{C}_2\mathbf{X}_1 & \mathbf{C}_2\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \right. \\
&\qquad \left. + \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{C}_1^\top & \mathbf{X}_1^\top \mathbf{C}_2^\top \\ \mathbf{X}_2^\top \mathbf{C}_1^\top & \mathbf{X}_2^\top \mathbf{C}_2^\top \end{bmatrix} + \begin{bmatrix} \mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top \end{bmatrix} \right\} \\
&= \sigma^2 \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{C}_1\mathbf{X}_1 & \mathbf{C}_1\mathbf{X}_2 \\ \mathbf{C}_2\mathbf{X}_1 & \mathbf{C}_2\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \right. \\
&\qquad \left. + \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{C}_1\mathbf{X}_1 & \mathbf{C}_1\mathbf{X}_2 \\ \mathbf{C}_2\mathbf{X}_1 & \mathbf{C}_2\mathbf{X}_2 \end{bmatrix}^\top + \begin{bmatrix} \mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top \end{bmatrix} \right\}
\end{aligned}$$

Now, given that $\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix}$ is unbiased, where $\mathbf{C}_1\mathbf{X}_1$, $\mathbf{C}_1\mathbf{X}_2$, $\mathbf{C}_2\mathbf{X}_1$, and $\mathbf{C}_2\mathbf{X}_2$ are matrices with only zeroes, then we get the following:

$$\sigma^2 \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top \end{bmatrix} \right\}$$

We can now look at the variance of the linear combination of this unbiased estimator:

$$\begin{aligned}
Var\left(\boldsymbol{\ell}_1^\top \widetilde{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widetilde{\mathbf{B}}_2\right) &= Var\left([\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top] \begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix} \right) \\
&= (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) Var\left(\begin{bmatrix} \widetilde{\mathbf{B}}_1 \\ \widetilde{\mathbf{B}}_2 \end{bmatrix} \right) \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix} \\
&= \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} + \begin{bmatrix} \mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top \end{bmatrix} \right\} \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix} \\
&= \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix} + \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \mathbf{C}_1\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_1\mathbf{V}\mathbf{C}_2^\top \\ \mathbf{C}_2\mathbf{V}\mathbf{C}_1^\top & \mathbf{C}_2\mathbf{V}\mathbf{C}_2^\top \end{bmatrix} \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix} \\
&= \ Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \mathbf{V} [\mathbf{C}_1^\top \quad \mathbf{C}_2^\top] \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix}
\end{aligned}$$

It is stated in the question that $\sigma^2\mathbf{V}$ is the variance-covariance matrix, therefore it follows that $\mathbf{V}$ must be positive definite and nonsingular. Then it must be that there exists a $n \times n$ nonsingular symmetric matrix $\mathbf{K}$ such that $\mathbf{K}^\top\mathbf{K} = \mathbf{K}\mathbf{K} = \mathbf{V}$. From this it follows that:

$$\begin{aligned}
&Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 (\boldsymbol{\ell}_1^\top \quad \boldsymbol{\ell}_2^\top) \begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix} \mathbf{K}^\top\mathbf{K} [\mathbf{C}_1^\top \quad \mathbf{C}_2^\top] \begin{pmatrix} \boldsymbol{\ell}_1 \\ \boldsymbol{\ell}_2 \end{pmatrix} \\
&= Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 [\boldsymbol{\ell}_1^\top \mathbf{C}_1 + \boldsymbol{\ell}_2^\top \mathbf{C}_2] \mathbf{K}^\top\mathbf{K} [\mathbf{C}_1^\top \boldsymbol{\ell}_1 + \mathbf{C}_2^\top \boldsymbol{\ell}_2] \\
&= Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 [\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}] [\mathbf{K}\mathbf{C}_1^\top \boldsymbol{\ell}_1 + \mathbf{K}\mathbf{C}_2^\top \boldsymbol{\ell}_2] \\
&= Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 [\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}] [\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}]^\top
\end{aligned}$$

From the above, it can be seen that since $[\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}][\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}]^\top \geq 0$ that:

$$\begin{aligned}
Var\left(\boldsymbol{\ell}_1^\top \widetilde{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widetilde{\mathbf{B}}_2\right) = Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right) + \sigma^2 [\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}] [\boldsymbol{\ell}_1^\top \mathbf{C}_1\mathbf{K} + \boldsymbol{\ell}_2^\top \mathbf{C}_2\mathbf{K}]^\top \\
\geq Var\left(\boldsymbol{\ell}_1^\top \widehat{\mathbf{B}}_1 + \boldsymbol{\ell}_2^\top \widehat{\mathbf{B}}_2\right).
\end{aligned}$$

Therefore, the generalized least square estimate of $\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$,

$$\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y} \end{bmatrix}$$

is the best linear unbiased estimator (BLUE). ■

2) Derive the test statistic under the null hypothesis, statistical distribution of the test and its rejection region for H_0.

<u>Ans:</u>

For convenience, some further notation will be added to the problem. Let $\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$ be the vector of regression coefficients partitioned such that $\mathbf{B}_2$ is $(p-r) \times 1$ and $\mathbf{B}_1$ is $r \times 1$. Therefore, $\mathbf{X}_2$ is an $n \times (p-r)$ matrix associated with $\mathbf{B}_2$, and $\mathbf{X}_1$ is an $n \times r$ matrix associated with $\mathbf{B}_1$. Also, let $\mathbf{X}$ be the original $n \times p$ matrix where $\mathbf{X}_1$ and $\mathbf{X}_2$ are from. The model $\mathbf{y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \boldsymbol{\varepsilon}$ represents the *full model*.

(Note: From the textbook, the notation here is applied in reverse, since in the textbook $\mathbf{B}_1$ is $(p \times r) \times 1$ and $\mathbf{B}_2$ is $r \times 1$. *However, since we are testing* $H_0\colon \mathbf{B}_1 = \mathbf{0}$ *rather than* $H_0\colon \mathbf{B}_2 = \mathbf{0}$, *then the notation is being reversed so the textbook method can be applied more smoothly.)*

We are looking to test the following hypothesis,

$$H_0\colon \mathbf{B}_1 = \mathbf{0} \text{ versus } H_1\colon \mathbf{B}_1 \neq \mathbf{0}.$$

From part 1, the full model can be written as follows,

$$\begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top\mathbf{V}^{-1} \\ \mathbf{X}_2^\top\mathbf{V}^{-1} \end{bmatrix} \mathbf{y}.$$

(*Note: Some steps from part 1 will be repeated again for clarity.*) Since $\sigma^2\mathbf{V}$ is the variance-covariance matrix, then $\mathbf{V}$ must be positive definite and nonsingular. Therefore, $\exists \mathbf{K}\ s.t.\ \mathbf{K}^\top\mathbf{K} = \mathbf{K}\mathbf{K} = \mathbf{V}$, where $\mathbf{K}$ is an $n \times n$ nonsingular symmetric matrix. We can then define a new set of variables as follows:

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \mathbf{W}_1 = \mathbf{K}^{-1}\mathbf{X}_1, \mathbf{W}_2 = \mathbf{K}^{-1}\mathbf{X}_2, \mathbf{u} = \mathbf{K}^{-1}\boldsymbol{\varepsilon}.$$

The original model is

$$\mathbf{y} = \mathbf{X}_1\mathbf{B}_1 + \mathbf{X}_2\mathbf{B}_2 + \boldsymbol{\varepsilon}$$

and we multiply $\mathbf{K}^{-1}$ from the left on both sides so that it becomes

$$\rightarrow \mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}_1\mathbf{B}_1 + \mathbf{K}^{-1}\mathbf{X}_2\mathbf{B}_2 + \mathbf{K}^{-1}\boldsymbol{\varepsilon}$$
$$\rightarrow \mathbf{z} = \mathbf{W}_1\mathbf{B}_1 + \mathbf{W}_2\mathbf{B}_2 + \mathbf{u}$$

Next, the expectation and variance for this transformed model will be shown,

$$E(\mathbf{u}) = E(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}E(\boldsymbol{\varepsilon}) = \mathbf{0}$$
$$Var(\mathbf{u}) = Var(\mathbf{K}^{-1}\boldsymbol{\varepsilon}) = \mathbf{K}^{-1}Var(\boldsymbol{\varepsilon})(\mathbf{K}^{-1})^\top = \mathbf{K}^{-1}(\sigma^2\mathbf{V})\mathbf{K}^{-1} = \sigma^2\mathbf{K}^{-1}(\mathbf{K}\mathbf{K})\mathbf{K}^{-1} = \sigma^2\mathbf{I}$$

The transformation has made it such that the model is mean zero, constant variance, and the response variable is uncorrelated, which fulfills the basic assumptions for ordinary least-squares regression. To derive the test statistic, it will then be based on the partial F test.

First the regression sum of squares will be shown:

$$SS_R\left(\begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}\right) = \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top [\mathbf{W}_1 \quad \mathbf{W}_2]^\top \mathbf{z} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{z})^2$$

(Note: $\mathbf{1}_n$ *is a* $n \times 1$ *vector consisting only of one's.)*

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top [\mathbf{K}^{-1}\mathbf{X}_1 \quad \mathbf{K}^{-1}\mathbf{X}_2]^\top \mathbf{K}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2$$

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} (\mathbf{K}^{-1}\mathbf{X}_1)^\top \\ (\mathbf{K}^{-1}\mathbf{X}_2)^\top \end{bmatrix} \mathbf{K}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2$$

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{K}^{-1} \\ \mathbf{X}_2{}^\top \mathbf{K}^{-1} \end{bmatrix} \mathbf{K}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{y})^2$$

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{K}^{-1}\mathbf{K}^{-1} \\ \mathbf{X}_2{}^\top \mathbf{K}^{-1}\mathbf{K}^{-1} \end{bmatrix} \mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{y})^2$$

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{V}^{-1} \\ \mathbf{X}_2{}^\top \mathbf{V}^{-1} \end{bmatrix} \mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{y})^2$$

$$= \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{V}^{-1}\mathbf{y} \\ \mathbf{X}_2{}^\top \mathbf{V}^{-1}\mathbf{y} \end{bmatrix} - \frac{1}{n}(\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{y})^2$$

$$= \widehat{\mathbf{B}}_1^\top \mathbf{X}_1{}^\top \mathbf{V}^{-1}\mathbf{y} + \widehat{\mathbf{B}}_2^\top \mathbf{X}_2{}^\top \mathbf{V}^{-1}\mathbf{y} \; - \frac{1}{n}(\mathbf{1}_n^\top \mathbf{K}^{-1}\mathbf{y})^2$$

Second, the mean residual sum of squares will be shown:

$$MS_{Res} = \frac{SS_{Res}}{n-p} = \frac{\mathbf{z}^\top\mathbf{z} - \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top [\mathbf{W}_1 \quad \mathbf{W}_2]^\top \mathbf{z}}{n-p}$$

$$= \frac{(\mathbf{K}^{-1}\mathbf{y})^\top \mathbf{K}^{-1}\mathbf{y} - \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top [\mathbf{K}^{-1}\mathbf{X}_1 \quad \mathbf{K}^{-1}\mathbf{X}_2]^\top \mathbf{K}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{K}^{-1} \\ \mathbf{X}_2{}^\top \mathbf{K}^{-1} \end{bmatrix} \mathbf{K}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top \begin{bmatrix} \mathbf{X}_1{}^\top \mathbf{V}^{-1} \\ \mathbf{X}_2{}^\top \mathbf{V}^{-1} \end{bmatrix} \mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \widehat{\mathbf{B}}_1^\top \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{y} - \widehat{\mathbf{B}}_2^\top \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \begin{bmatrix} \widehat{\mathbf{B}}_1 \\ \widehat{\mathbf{B}}_2 \end{bmatrix}^\top [\mathbf{X}_1 \quad \mathbf{X}_2]^\top \mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \left\{ \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1} \\ \mathbf{X}_2^\top \mathbf{V}^{-1} \end{bmatrix} \mathbf{y} \right\}^\top [\mathbf{X}_1 \quad \mathbf{X}_2]^\top \mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \mathbf{y}^\top [\mathbf{V}^{-1}\mathbf{X}_1 \quad \mathbf{V}^{-1}\mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top \mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} - \left\{ \mathbf{y}^\top \mathbf{V}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top \right\} \mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top\mathbf{V}^{-1}\left\{\mathbf{y} - [\mathbf{X}_1 \quad \mathbf{X}_2]\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}[\mathbf{X}_1 \quad \mathbf{X}_2]^\top\right\}\mathbf{V}^{-1}\mathbf{y}}{n-p}$$

$$= \frac{\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{V}_D\mathbf{V}^{-1}\mathbf{y}}{n-p}$$

where

$$\mathbf{V}_D = \mathbf{y} - [\mathbf{X}_1 \quad \mathbf{X}_2]\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}[\mathbf{X}_1 \quad \mathbf{X}_2]^\top$$

Under the null hypothesis, $\mathbf{B}_1 = \mathbf{0}$, the *reduced model* can be seen as:

$$\mathbf{y} = \mathbf{X}_2\mathbf{B}_2 + \boldsymbol{\varepsilon}$$
$$\mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}_2\mathbf{B}_2 + \mathbf{K}^{-1}\boldsymbol{\varepsilon}$$
$$\mathbf{z} = \mathbf{W}_2\mathbf{B}_2 + \mathbf{u}$$

The ordinary least-squares estimate of $\mathbf{B}_2$, denoted $\mathbf{B}_2'$ can be seen as:

$$\mathbf{B}_2' = (\mathbf{W}_2^\top\mathbf{W}_2)^{-1}\mathbf{W}_2^\top\mathbf{z}$$
$$= [(\mathbf{K}^{-1}\mathbf{X}_2)^\top\mathbf{K}^{-1}\mathbf{X}_2]^{-1}(\mathbf{K}^{-1}\mathbf{X}_2)^\top\mathbf{K}^{-1}\mathbf{y}$$
$$= [\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{y}$$

Then the regression sum of squares for $\mathbf{B}_2$ can be seen as:

$$SS_R(\mathbf{B}_2) = \mathbf{B}_2'{}^\top\mathbf{W}_2^\top\mathbf{z} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{z})^2$$
$$= \mathbf{B}_2'{}^\top(\mathbf{K}^{-1}\mathbf{X}_2)^\top\mathbf{K}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2$$
$$= \mathbf{B}_2'{}^\top\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2$$

Then it follows that the regression sum of squares due to $\mathbf{B}_1$ given that $\mathbf{B}_2$ is already in the model is:

$$SS_R(\mathbf{B}_1|\mathbf{B}_2) = SS_R\left(\begin{bmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{bmatrix}\right) - SS_R(\mathbf{B}_2)$$
$$= \hat{\mathbf{B}}_1^\top\mathbf{X}_1{}^\top\mathbf{V}^{-1}\mathbf{y} + \hat{\mathbf{B}}_2^\top\mathbf{X}_2{}^\top\mathbf{V}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2 - \left[\mathbf{B}_2'{}^\top\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{y} - \frac{1}{n}(\mathbf{1}_n^\top\mathbf{K}^{-1}\mathbf{y})^2\right]$$
$$= \left(\hat{\mathbf{B}}_1^\top\mathbf{X}_1^\top + \hat{\mathbf{B}}_2^\top\mathbf{X}_2^\top - \mathbf{B}_2'{}^\top\mathbf{X}_2^\top\right)\mathbf{V}^{-1}\mathbf{y}$$
$$= \left(\begin{bmatrix}\hat{\mathbf{B}}_1\\ \hat{\mathbf{B}}_2\end{bmatrix}^\top\begin{bmatrix}\mathbf{X}_1^\top\\ \mathbf{X}_2^\top\end{bmatrix} - \mathbf{B}_2'{}^\top\mathbf{X}_2^\top\right)\mathbf{V}^{-1}\mathbf{y}$$
$$= \left\{\left(\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}\mathbf{y}\right)^\top\begin{bmatrix}\mathbf{X}_1^\top\\ \mathbf{X}_2^\top\end{bmatrix} - \{[\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{y}\}^\top\mathbf{X}_2^\top\right\}\mathbf{V}^{-1}\mathbf{y}$$
$$= \left\{\mathbf{y}^\top\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\end{bmatrix}^\top\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\\ \mathbf{X}_2^\top\end{bmatrix} - \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{X}_2[\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top\right\}\mathbf{V}^{-1}\mathbf{y}$$
$$= \left\{\mathbf{y}^\top[(\mathbf{X}_1^\top\mathbf{V}^{-1})^\top \quad (\mathbf{X}_2^\top\mathbf{V}^{-1})^\top]\begin{bmatrix}\mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top\mathbf{V}^{-1}\mathbf{X}_2\\ \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1^\top\\ \mathbf{X}_2^\top\end{bmatrix}\right.$$
$$\left. - \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{X}_2[\mathbf{X}_2^\top\mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top\right\}\mathbf{V}^{-1}\mathbf{y}$$

$$= \left\{ \mathbf{y}^\top [\mathbf{V}^{-1}\mathbf{X}_1 \quad \mathbf{V}^{-1}\mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{bmatrix} - \mathbf{y}^\top \mathbf{V}^{-1}\mathbf{X}_2 [\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top \right\} \mathbf{V}^{-1}\mathbf{y}$$

$$= \mathbf{y}^\top \mathbf{V}^{-1} \left\{ [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top - \mathbf{X}_2 [\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top \right\} \mathbf{V}^{-1}\mathbf{y}$$

$$= \mathbf{y}^\top \mathbf{V}^{-1}\mathbf{X}_D \mathbf{V}^{-1}\mathbf{y}$$

where

$$\mathbf{X}_D = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top - \mathbf{X}_2 [\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top$$

with $p - (p - r) = r$ degrees of freedom. We can call it the *extra sum of squares due to* $\mathbf{B}_1$.

Then it follows that the partial F test statistic, denoted F_0, can be seen as:

$$F_0 = \frac{SS_R(\mathbf{B}_1 | \mathbf{B}_2)/r}{MS_{Res}}$$

$$\boxed{= \frac{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{X}_D \mathbf{V}^{-1}\mathbf{y}/r}{\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{V}_D \mathbf{V}^{-1}\mathbf{y}/(n-p)}},$$

where

$$\mathbf{X}_D = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top - \mathbf{X}_2 [\mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2]^{-1}\mathbf{X}_2^\top$$

and

$$\mathbf{V}_D = \mathbf{y} - [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_1^\top \mathbf{V}^{-1}\mathbf{X}_2 \\ \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_1 & \mathbf{X}_2^\top \mathbf{V}^{-1}\mathbf{X}_2 \end{bmatrix}^{-1} [\mathbf{X}_1 \quad \mathbf{X}_2]^\top.$$

(*Note: Given that this is the case, it also must be such that* $SS_R(\mathbf{B}_1 | \mathbf{B}_2)$ *and* SS_{Res} *follows a* χ_r^2 *and* χ_{n-p}^2 *distribution respectively.*) In the case that $\mathbf{B}_1 \neq \mathbf{0}$, then F_0 follows a noncentral F distribution with a noncentrality parameter denoted

$$\lambda = \frac{1}{\sigma^2} \mathbf{B}_1' \mathbf{X}_1' [\mathbf{I} - \mathbf{X}_2 (\mathbf{X}_2' \mathbf{X}_2)^{-1} \mathbf{X}_2'] \mathbf{X}_1 \mathbf{B}_1.$$

If after performing the hypothesis test, we see that $F_0 > F_{\alpha,r,n-p}$, then we reject H_0 at the α significance level, drawing the conclusion that at least one of the regressors in $\mathbf{B}_1$ is nonzero.

2. Can **any** linear regression model be checked for model adequacy by statistical testing for lack of fit or goodness of fit? Why or why not? Please provide your answer with detailed justification (i.e., by mathematical proof or by a numerical example).

Ans:

The formal test for a lack of fit test as stated in the textbook is a method for trying to find out if a tentative model adequately describes the data. It comes with certain assumptions which include: normality, independence, constant-variance, and only the first-order or straight-line character of the relationship is in doubt. Another requirement is that there exist replicate observations on the response y for at least one level of x. Such a requirement becomes more difficult in particular when y is a continuous variable. There is also emphasis on the fact that these are true replications, derived from running multiple separate experiments, rather than just have a single experiment that returns multiple measurements at some level x_i.

Furthermore, the formula for partitioning the residual sum of squares into two components can be seen as follows,

$$SS_{Res} = SS_{PE} + SS_{LOF}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(y_{ij} - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2 + \sum_{i=1}^{m} n_i(\bar{y}_i - \hat{y}_i)^2.$$

The term $SS_{PE} = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^2$ is known as the sum of squares due to pure error. The idea behind it is that it is a model-independent estimate of the error, since its calculation doesn't involve the model-dependent term $\hat{y}_i$. The calculation itself will be further clarified.

Let there be a total of n observations in the data. At each level x_i $(i = 1, \cdots, m)$, there are n_i observations. Then let y_{ij} be the jth observation on the response at x_i, $i = 1, \cdots, m$ and $j = 1, \cdots, n_i$. Hence, there are $n = \sum_{i=1}^{m} n_i$ total observations.

Given that the important requirement of replicate observations on the response y for at least one level of x is not being met, the result is that $n_i = 1$ for $i = 1, \cdots, m$. This implies then also that $\bar{y}_i$ will be calculated based off a single observation each time, such that the $\left(y_{ij} - \bar{y}_i\right)$ term within SS_{PE} will zero out after the summation. From this, the test statistic will be incalculable, since it requires deriving $F_0 = \frac{\frac{SS_{LOF}}{m-2}}{\frac{SS_{PE}}{n-m}}$, and so the denominator will lead to a division by zero error. An alternative is to estimate the "pure" error by calculating it based off the "near neighbors." It requires instead looking for levels of x that can be considered "near neighbors" and using their resulting y values to be considered then as replicate values.

Furthermore, there are many other techniques that can be tried, in addition to testing for lack of fit. These other methods for residual analysis provide insight into potential problems with the model, for example outliers, non-normality, etc. These other alternatives may also require certain conditions though, for example $n > p$, where n is the number of observations and p is the total number of regressors (including the intercept).

3. Prove that in selection of p regressors, minimizing $MS_{res}(p)$ will lead to maximizing the adjusted $R^2(p)$.

Ans:

The first step will to be to make the terms and notation more explicit. Let R_p^2 denote the coefficient of multiple determination for a subset regression model that contains within it p terms (where p includes the intercept). It can be written as,

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T},$$

where $SS_R(p)$ and $SS_{Res}(p)$ denote the regression sum of squares and the residual sum of squares for the p-term regression model. Then, let the adjusted $R^2(p)$ be denoted as,

$$R_{Adj,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - R_p^2\right).$$

Furthermore, let $MS_{res}(p)$ be denoted as,

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}.$$

To show that minimizing $MS_{Res}(p)$ leads to maximizing $R^2_{Adj,p}$, let the following be shown:

$$R^2_{Adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p) = 1 - \left(\frac{n-1}{n-p}\right)\left(1 - \left(1 - \frac{SS_{Res}(p)}{SS_T}\right)\right)$$

$$= 1 - \left(\frac{n-1}{n-p}\right)\left(\frac{SS_{Res}(p)}{SS_T}\right) = 1 - \frac{\left(\frac{SS_{res}(p)}{n-p}\right)}{\left(\frac{SS_T}{n-1}\right)} = 1 - \frac{MS_{Res}(p)}{\left(\frac{SS_T}{n-1}\right)}$$

It has been shown above that $R^2_{Adj,p} = 1 - \frac{MS_{Res}(p)}{\left(\frac{SS_T}{n-1}\right)}$. For any given model including a subset of regressors, the denominator term, $\frac{SS_T}{n-1}$, will remain constant. The only term that changes is $MS_{Res}(p)$. Therefore, if in minimizing $MS_{Res}(p)$, the relationship is that $\frac{MS_{Res}(p)}{\left(\frac{SS_T}{n-1}\right)}$ will decrease, and hence $R^2_{Adj,p}$ will increase. Therefore, it follows that minimizing $MS_{Res}(p)$ will at the same time lead to maximizing the adjusted $R^2(p)$. ■

4. Critique the following statement: "For selection of the regressors to include in a linear regression model, the best strategy is always finding the model with the largest possible value of R^2."

Ans:

This methodology is flawed and does not make sense. In the textbook, R^2 is defined as,

$$R^2 = 1 - \frac{SS_{Res}}{SS_T} = \frac{SS_R}{SS_T}.$$

The way that R^2 can be used is that a model with a higher R^2 is considered to be the better model, where $0 \leq R^2 \leq 1$. The reasoning is that larger values (e.g., above 0.9) imply that most of the variability in y is being explained by the model. The issue is that this could work well for simple linear regression with only a single regressor being considered for the model. However, if multiple regressors are added to the model, the value of R^2 can only go up and can never decrease.

The result then is that simply by adding new regressors to the model, the R^2 value can simply keep saying that the new model is an improvement over the previous model. The logic then that R^2 can best choose a model therefore is flawed. To compensate for this behavior where R^2 continually increases for larger models, the adjusted R^2, or R^2_{Adj} can be used. It is denoted as

$$R^2_{Adj} = 1 - \frac{\frac{SS_{Res}}{n-p}}{\frac{SS_T}{n-1}},$$

where n is the total number of observations and p is the number of regressors (including the intercept). Here, the difference is that SS_{Res} and SS_T are being divided by their corresponding degrees of freedom. Therefore, for an improvement to take place in the R^2_{Adj}, there must be an improvement in $MS_{Res} = \frac{SS_{Res}}{n-p}$ (since $\frac{SS_T}{n-1}$ is unaffected by changes in the number of parameters for the model).

There are other methods other than those directly involving R^2 or R^2_{Adj}. These are merely methods for evaluating and comparing subset regression models. Other metrics include residual mean square ($MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}$), Mallow's C_p statistics, along with the Akaike Information Criterion (AIC), and its Bayesian analogues (BIC_{Sch}, BIC_{Sawa}). All these evaluation metrics will analyze the regression model in their unique ways. To say that one method is always ideal is in itself illogical and ignores the fact that each metric gives a different level of insight into the performance of a model.

5. There are multiple methods or criteria for selection of regressors (e.g., forward selection, all possible regressions, R^2, C_p).
    1) Describe all the methods we have learned in this course.

Ans:

The methods are to be separated into two groups. The first is regarding methods of generating the subsets of models. The other group is for finding criteria to see if one subset is preferrable to another. The first group includes the following methods: all possible regressions, forward selection, backward elimination, and stepwise regression. The second group includes the following criterion: coefficient of multiple determination (R^2_p), adjusted R^2 ($R^2_{Adj,p}$), residual mean square ($MS_{Res}(p)$), Mallows' C_p statistic (C_p), Akaike Information Criterion (AIC), and the Bayesian analogues to the AIC (BIC_{Sch}, BIC_{Sawa}). (*Note: It is understood that there are some alternate methods learned at the beginning such as testing for the significance of regression or residual analysis, but they will not be focused on.*)

As stated, the methods for generating the subsets of models includes: all possible regressions, forward selection, backward elimination, and stepwise regression. Each of these methods will be discussed individually, where their methodology and usefulness will be explored.

It is apparent from the course, that the best methodology is not necessarily to simply include all known regressors into a model and use that as a final model. This is made quite apparent in Problem 4, when discussing the R^2 metric. It is ideal then to find out which subset of regressors can possibly be selected for a "final" model. However, with an increasing number of regressors comes an increasing number of subsets to test. A way to analyze them is simply to analyze every single subset of regressors to potentially include in a model. This method is known as "all possible regressions." If a dataset contains K candidate regressors that are being considered for a model, this method will fit an individual model for all 2^K possible subsets. It is worth noting also that these subsets will all include the intercept term by default.

Previously, such a feat was not feasible due to the number of models that would require fitting. However, advanced in computational efficiency has made this method quite practical. After fitting all 2^K possible models, a corresponding set of criteria then is applied to each model to evaluate their effectiveness. For example, for each of the different models, the following can be calculated: $SS_{Res}(p)$, R^2_p, $R^2_{Adj,p}$, $MS_{Res}(p)$, C_p, etc. An idea then is to check which models perform best. It is not the case however that all the criteria will necessarily choose the same model. Furthermore, each of the final set of candidate models must be further explored. This includes model adequacy, residual analysis, multicollinearity, etc.

The other group of methods which include forward selection, backward elimination, and stepwise regression, fall into a group known as *stepwise-type procedures*. These are alternative methods for generating subsets of models. It is the case that sometimes a dataset where K is large, and so performing all possible regressions is computationally expensive. These stepwise-type procedures then will add and/or remove regressors one at a time from a current model.

The forward selection procedure will begin first with an intercept-only model. It then potentially adds regressors from the set of K regressors one at a time to try and find the optimal subset of regressors. To do this, it first tries to find the regressor that has the *largest simple correlation* with y, the response variable. This can be found by calculating $r_{X,Y} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2(Y-\bar{Y})^2}}$. The reason is that the selected regressor then would also have the largest F statistic from the hypothesis test that is checking for the significance of regression. A parameter F_{IN} is included, which is used to determine whether a regressor is to be added. If the corresponding F statistic of that first regressor exceeds F_{IN}, then it is to be added to the base intercept-only model.

For a second regressor to consider for the model, the *partial correlation* can be used. In this case, the idea is to find the next regressor that again will provide the largest partial F statistic, $F = \frac{SS_R(x_j|x_i)}{MS_{Res}(x_i,x_j)}$, where x_i is the first regressor (if added in the previous step) and x_j is the regressor with the highest partial correlation. The interpretation of the F statistic is that it is measuring the increase in the regression um of squares that results from adding the new regressor to the set of already existing regressors. The same requirement is needed, where the resulting partial F statistic must exceed F_{IN} to be added. This process of finding the largest partial F statistic continues until the next one fails to exceed F_{IN}, or all the potential regressors are already added.

A methodology that works in the other direction of forward selection is known as *backward elimination*. Rather than starting with an intercept-only model, it instead starts with the full model that includes all K possible regressors. From this full model, the partial F statistic is calculated for each regressor, treating them as if they were the last regressor to be considered. The regressor with the smallest partial F statistic then is compared to an alternate parameter F_{OUT}, where if it is less than F_{OUT} then the regressor is subsequently deleted. This process repeats until the partial F statistic for a regressor is no longer smaller than the F_{OUT} metric.

It can be apparent that the two methods, forward selection and backward elimination, can conclude with a different set of models. Therefore, the greedy-algorithm aspect of them makes it apparent that a combination of models is possible. The last method, *stepwise regression*, is an alternative that combines elements of both forward selection and backward elimination. It starts in the same fashion as forward selection, where it begins with an intercept-only model and tries to add one regressor at a time from the group of K candidate regressors based on their corresponding F statistics. Again, if the F statistic exceeds the parameter F_{IN}, then it is added to the model. However, at each step the algorithm will also perform a single step from backward elimination. That is, for the current model that includes some subset of regressors, it will then check if any of them need to be removed based on their corresponding partial F statistic. If the smallest partial F statistic of the regressors is below the parameter F_{OUT}, then it is removed from the current subset. This method is useful, since it can be apparent that a model that previously relied on some regressor x_i, will no longer find it helpful in the path towards finding a suitable "final" model.

The next part is to discuss the various criteria that have been discussed for comparing different subsets of models. As stated before, these criteria include: coefficient of multiple determination (R_p^2), adjusted R^2 ($R_{Adj,p}^2$), residual mean square ($MS_{Res}(p)$), Mallows' C_p statistic (C_p), Akaike Information Criterion (AIC), and the Bayesian analogues to the AIC (BIC_{Sch}, BIC_{Sawa}). It is apparent that the stepwise-type procedures did not rely so much on these various criteria for selecting a potential model, however, the all possible regressions method (as described in the textbook) did rely on some of these.

The first criterion is coefficient of multiple determination or R_p^2. The formula is as follows,

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T},$$

where $SS_R(p)$, $SS_{Res}(p)$ and SS_T denote the regression sum of squares, residual sum of squares, and the total sum of squares for a model that contains p terms (*Note: The SS_T remains constant despite the number of regressors in the model.*). It has been stated already in Problem 3, that maximizing R_p^2 is not useful, since it will increase merely by adding regressors to the model.

However, an alternative is to plot R_p^2 vs. p, and look for a "knee" in the plot where additional regressors no longer greatly improve the model. Another alternative is to look for an R^2-*adequate* (α) *subset*. The idea is to determine a subset of models with an R^2 value that is not significantly different from the R^2 for the "full" model containing all K regressors. The metric used is

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{\alpha,n,K}),$$

where

$$d_{\alpha,n,K} = \frac{KF_{\alpha,K,n-K-1}}{n-K-1}.$$

Here, the R^2 for the "full" model is R_{K+1}^2. Any model with an R^2 larger than R_0^2 is said to belong to the R^2-adequate (α) subset. When analyzing R_p^2, large values closer to 1 are preferred.

Also mentioned previously in Problem 4 is the adjusted R^2 or $R_{Adj,p}^2$. The formula for $R_{Adj,p}^2$ again is as follows,

$$R_{Adj,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2).$$

This is actually the same formula as seen previously in Problem 4, except that it includes in it the R_p^2 term just discussed. As mentioned before, $R_{Adj,p}^2$ will only increase if $MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}$ also increases (a.k.a., the residual mean square for a model with p regressors).

The next criterion to analyze is the residual mean square or $MS_{Res}(p)$. This same criterion was analyzed briefly in Problem 3, alongside $R_{Adj,p}^2$. The formula for $MS_{Res}(p)$ is as follows,

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p},$$

which is the residual sum of squares ($SS_{Res}(p)$) divided by its degrees of freedom. The issue with $SS_{Res}(p)$ is that as seen in R_p^2, it will steadily decrease as regressors are added to the model. By looking instead at the residual mean square, it is possible to see a different behavior as p increases. Hypothetically, it will decrease, then start to increase again as p gets past a certain number. This makes sense as eventually, the decrease in $SS_{Res}(p)$ will be outweighed by the increase in p. When using $MS_{Res}(p)$, then the options are to look at the following: the minimum $MS_{Res}(p)$, the p for which $MS_{Res}(p)$ is close to the $MS_{Res}(p)$ for the full model, and the p nearest to where $MS_{Res}(p)$ stops decreasing and begins to increase.

The next metric is Mallows' C_p statistic, or C_p. The formula for C_p is as follows,

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p.$$

The formula is slightly more complicated, as it is based on the mean square error for a fitted value of a model with p regressors in it. This formula involves the squared bias and variance of the fitted values. A requirement however is that $\hat{\sigma}^2$ is a suitable estimate for σ^2. Therefore, several options can be used to estimate σ^2. The way that it is utilized is that smaller values are generally preferrable. Another possibility is to plot C_p against p, to see how close the values are to the line $C_p = p$. Models with lower bias are closer to this line. It is a method to see how much of a tradeoff can be made between the bias of a model and the reduction in the average error of prediction.

The last set of criteria to be analyzed are Akaike Information Criterion (AIC), and the Bayesian analogues to the AIC (BIC_{Sch}, BIC_{Sawa}). We will first start with Akaike Information Criterion, or AIC. This metric has the goal of maximizing the expected entropy for a certain model, where entropy is defined as the measure of expected information. The information is understood to be the Kullback-Leibler information measure. The formula is as follows,

$$AIC = n \ln\left(\frac{SS_{Res}}{n}\right) + 2p.$$

It can be seen here that like the other metrics, AIC includes the SS_{Res} term. Therefore, it penalizes the behavior of only increasing when additional regressors are added by including a penalty term. Two alternatives are also possible which are similar to AIC. They are both called the Bayesian information criterion, or BIC. However, one was developed by Schwartz (BIC_{Sch}) and another was developed by Sawa (BIC_{Sawa}). They are slightly different, and the textbook provides the following formula for BIC_{Sch},

$$BIC_{Sch} = n \ln\left(\frac{SS_{Res}}{n}\right) + p \ln n.$$

2) Is there the best method or criterion?

Ans:

None of the above mentions methods or criterion claim to be the best or lead to the best results. Furthermore, based on their descriptions and similar to the argument made previously in Problem 4, there is no single suitable method or criterion that is superior to every other. A way to look at the problem is to see that each method or criterion employs some idea to try and analyze the regression model in some specific way. This way of “seeing” a regression model is different through the lens of each method. It would be convenient if it were such the case that for example if one method and one criterion *always* led to the “best” regression model. However, there is in fact no such thing as a “best” regression model. There is often the case that a group of models are equally suitable for a problem.

The data and its underlying behavior is in general too complex for it to be said that any one of these is always the best choice. In fact, it makes more sense to apply as many of these as is feasible such that their collective observations can be gathered and analyzed as a whole. For example, with the all possible regressions method, a useful idea is to see which of the final models is chosen (or “voted” for) by a group of several criteria. Furthermore, it is possible to compare the different models generated by all possible regressions along with the stepwise-type procedures. Another point is that after having narrowed down on a potential group of models, there is next the important step of doing careful analysis on each of these models, checking to see if for example there an issue of multicollinearity or problems concerning residual analysis. After doing that, a next step could be to repeat the same process, going through another series of subset selection and model comparison via various criteria. It is also worth mentioning that domain knowledge can play an important role. For example, if someone knows something specific about a set of variables, that can prove to be more useful than merely selecting a model based off certain analytical tools.

3) Create a dataset (for a model with response variable y, two regressors x_1 and x_2, and its interaction x_1x_2) to illustrate how to apply each method or criterion to selection of regressors. Provide details about how you get the dataset. Do not use any data in the Textbook.

Ans:

The approach for this problem is to first generate the data. Then, each of the variable selection methods will be used, that is: all possible regressions, forward selection, backward elimination, and stepwise regression. From here, the various criteria discussed will be used for model comparison. Within all possible regressions, the different criteria can be used for each of the generated models. For the stepwise-type procedures, the criteria can be used to compare each of the final models.

To generate the dataset, it is required to first generate two variables, x_1 and x_2. Then, their interaction term needs to be created by multiplying the two vectors together row-wise to create x_1x_2. To generate x_1 and x_2, the idea was to randomly sample them from two different distributions. The distributions are both a mixture of two normal distributions, but each mixture is different. For x_1, it is sampled from the mixture $0.4N(0,1) + 0.6N(2.5,2^2)$. For x_2, it is sampled from the mixture $0.9N(1,2^2) + 0.1N(3,5^2)$. From these, 1,000 observations are randomly sampled each. To create the response variable y, an arbitrary relationship is created, where $y = 1.5 + 2x_1 + 3x_2 + 0.5x_1x_2 + \varepsilon$, where ε are randomly generated from the $N(0,1)$ distribution. A density plot for each of the generated variables can be seen below in Figure 1.

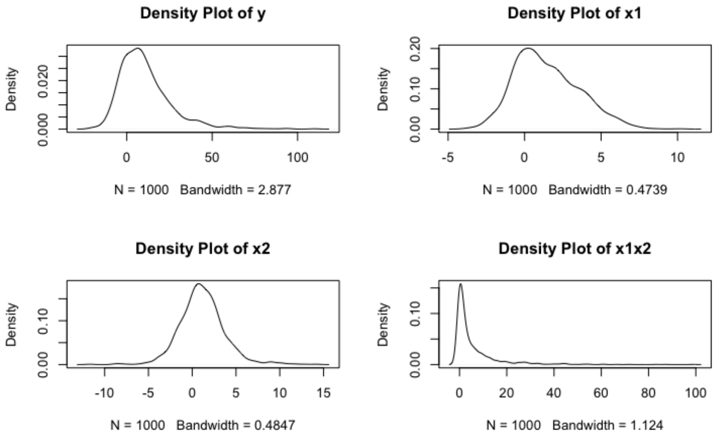

Figure 1 The above figure shows a density plot for each of the generated variables. Clockwise from top-left are: y, x_1, x_1x_2, and x_2.

Once the data is generated, the goal is to implement the variable selection methods combined with the various criteria discussed. Using R's "leap" package to implement possible regressions with the "regsubsets()" function, it will determine a "best" model for each case when the model contains 1, 2, and 3 regressors (selected from the set of 3 total regressors, where the last case is somewhat redundant since it is merely analyzing all possible regressors). The "best" is determined by which has the lowest SS_{Res}. From here, three potential models are output. The chosen models are shown below in Table 1.

Table 1 The below table shows the "best" model for when the model contains 1, 2, and 3 predictors.

Number of Predictors	Best Predictors to Use
1	x_2
2	x_2, x_1x_2
3	x_1, x_2, x_1x_2

From these three models, their corresponding R_p^2, $R_{Adj,p}^2$, $MS_{Res}(p)$, C_p, AIC, and BIC_{Sch} are calculated (*Note: R's leap package doesn't calculate BIC_{Sawa}, as mentioned in the textbook. Furthermore, the textbook doesn't provide a formula from which it can be calculated. Therefore, it is left out of this analysis.*). Below in Table 2 are the metrics, their corresponding values, and the model they selected relative to those values. It is apparent that every criterion has unanimously "voted" for the "full" model containing all three regressors.

Table 2 The below table shows the "best" model from Table 1 chosen by each criterion and their corresponding values.

Criterion	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p	BIC	AIC
Criterion Value	0.9955	0.9955	1.0833	4	-5383.595	84.0043
Number of Predictors	3	3	3	3	3	3

The next step is to use the stepwise-type procedures. The textbook mentions using the (default, pre-installed) "MASS" package and the "step()" function. However, the "step()" function doesn't seem to exist in my current (up-to-date) version of the package (nor does it seem to exist after searching for it online), while “stepAIC()” does exist. Therefore, this function will be used as a replacement. From reading the documentation, this function will perform the stepwise selection methods, while using the already mentioned AIC criterion for choosing models rather than the partial F test.

The tables below show the results of using “stepAIC()” for forward selection (Table 3), backward selection (Table 4), and stepwise regression (Table 3) (*Note: The output from forward selection and stepwise regression are identical and so their output is shown in the same table.*). In Table 3, it can be seen that it starts from the intercept-only model, then adds a regressor at each step. It first adds x_2, then x_1x_2, and finally x_1. The AIC remains in the thousands, before dropping down sharply to 84 in the last step, indicating a large improvement in the model. This makes sense, given that from $y = 1.5 + 2x_1 + 3x_2 + 0.5x_1x_2 + \varepsilon$, all the regressors (and thus their corresponding distributions) are influential towards the response variable. With regard to backward elimination, it is also worth making the point that in Table 3, no regressors are ever removed in the “backward elimination step” of stepwise regression.

Looking at Table 4, it is apparent that starting from the “full” model containing all possible regressors, it never makes a deletion of a regressor and so it simply stays at the “full” model and exits the algorithm. This coincides with the behavior of the stepwise regression algorithm at the last step in Table 3. Furthermore, it is not surprising, given the difference in AIC between the fourth step and third step seen in Table 3.

Table 3 The table below shows the output from using forward selection on the randomly generated dataset in R.

Step	Model	AIC
1	$y \sim 1$	5,489.23
2	$y \sim x_2$	3,461.67
3	$y \sim x_2 + x_1x_2$	2,056.47
4	$y \sim x_1 + x_2 + x_1x_2$	84

Table 4 The table below shows the output from using backward elimination on the randomly generated dataset in R.

Step	Model	AIC
1	$y \sim x_1 + x_2 + x_1x_2$	84

It is apparent that all three stepwise-type procedures have settled on the same “final” model. The AIC dramatically changes as it goes from the full model to the other models. Furthermore, since the stepwise-type procedures have also unanimously chosen the full model, there is no need to re-calculate the various criterion for this model, since it has been done already in Table 2. Based on the variable selection and model comparison methods, they have unanimously chosen that the “best” model is the full model containing all the regressors. In all of the comparisons, no other model was preferred over this one. From this, it can be concluded that the best “final” model is the “full” model containing all three regressors, x_1, x_2, and x_1x_2. It has been decided unanimously by all variable selection methods and model comparison criteria.

Code Appendix

```
# Generate data
N <- 1e3
# Reference: https://stats.stackexchange.com/questions/70855/generating-random-variables-from-a-mixture-of-normal-distributions
set.seed(1); u_sample <- runif(n = N)
set.seed(1)
x1 <- sapply(1:N, function(x) {
  if (u_sample[x] > 0.6) {
    return(rnorm(n = 1, mean = 0, sd = 1))
  } else {
    return(rnorm(n = 1, mean = 2.5, sd = 2))
  }
})
set.seed(1)
x2 <- sapply(1:N, function(x) {
  if (u_sample[x] > 0.1) {
    return(rnorm(n = 1, mean = 1, sd = 2))
  } else {
    return(rnorm(n = 1, mean = 3, sd = 5))
  }
})
x1x2 <- x1*x2

y <- 1.5 + 2 * x1 + 3 * x2 + 0.5 * x1x2 + rnorm(n = N, mean = 0, sd = 1)
par(mfrow=c(2,2))
plot(density(y));plot(density(x1));plot(density(x2));plot(density(x1x2))

# full model
df <- as.data.frame(cbind(y, x1, x2, x1x2))
full_model <- lm(formula = y~., data = df)
intercept_only <- lm(y~1, data = df)
summary(full_model)

# all possible regressors
best <- regsubsets(x = y~., data = df, nvmax = 3)
res.sum <- summary(best)
p.m <- 2:4
aic <- N * log(res.sum$rss / N) + 2 * p.m
MS_Res <- res.sum$rss / (N - p.m)
data.frame(
  R2 = which.max(res.sum$rsq),
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$cp),
  BIC = which.min(res.sum$bic),
  AIC = which.min(aic),
  MSRes = which.min(MS_Res)
)

data.frame(
  R2 = round(res.sum$rsq[3],4),
  Adj.R2 = round(res.sum$adjr2[3],4),
  CP = round(res.sum$cp[3],4),
  BIC = round(res.sum$bic[3],4),
  AIC = round(aic[3],4),
  MSRes = round(MS_Res[3],4)
)

# forward selection
forward <- MASS::stepAIC(intercept_only,
                         scope = list(upper=full_model, lower=intercept_only),
                         direction = c('forward'))

# backward elimination
backward <- MASS::stepAIC(full_model,
                          # scope = list(upper=full_model, lower=intercept_only),
                          direction = c('backward'))
```

```
# stepwise regression
stepwise <- MASS::stepAIC(intercept_only,
                       scope = list(upper=full_model, lower=intercept_only),
                       direction = c('both'))
```