1. [16 marks] Define vectors $v_1$, $v_2$, and $v_3$ as follows:

$$v_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}.$$ 

Let $S$ be the subspace of $\mathbb{R}^5$ spanned by these vectors.

Use $R$ to do the following:

(a) Find the orthogonal projection matrix for $S$. Also find the orthogonal projection matrix for $S^\perp$ (the orthogonal complement of $S$). You need to write down both projection matrices and present the R-code you used to produce them.

(b) Consider the following three vectors:

$$w_1 = \begin{bmatrix} 4 \\ -1 \\ 3 \\ -1 \\ -6 \end{bmatrix}, \quad w_2 = \begin{bmatrix} -3 \\ 0 \\ -2 \\ 6 \\ 9 \end{bmatrix}, \quad w_3 = \begin{bmatrix} -3 \\ 4 \\ 0 \\ 0 \\ -1 \end{bmatrix}.$$ 

For each of these vectors, determine whether the vector is in $S$, is in $S^\perp$ or is not in either $S$ or $S^\perp$. Explain your answers.

(c) Suppose we further decompose $S$ into two smaller subspaces: $S_1$ which consists of the subspace spanned by $v_1$ and $S_2$ which consists of all elements of $S$ that are orthogonal to $S_1$.

i. Is $S_2$ the same as the subspace of $\mathbb{R}^5$ spanned by $v_2$ and $v_3$? Explain your answer.

ii. Find the projection matrix for $S_2$.

2. [14 marks] Weighted least squares can be used in situations when the values of the response variable do not all have the same variance. Suppose that we have the following situation:

$$Y = \mu_Y + \epsilon \quad \text{where} \quad \mu_Y = X\beta \quad \text{and} \quad \epsilon \sim N(0, \Sigma).$$ 

Now define a matrix $W$ as $W = \Sigma^{-1}$. Then the weighted least squares estimates for $\beta$ and $\mu_Y$ are given by:

$$\hat{\beta} = (X'WX)^{-1}X'WY$$

$$\hat{\mu}_Y = X\hat{\beta} = X(X'WX)^{-1}X'WY$$
(a) Find $E(\hat{\beta})$ and the covariance matrix for $\hat{\beta}$. Show your work.

(b) Find $E(\hat{\mu_Y})$ and the covariance matrix for $\hat{\mu_Y}$. Show your work.

3. [20 marks] This question uses data collected by the Australian Institute of Sport. If you look at “2001 Assignments” you will find the data on the STATS 330 webpage under the “Data Sets” link as the “Institute of Sport Data.” You should look at this data set to get a feel for the data BUT you do NOT need to use it directly. For this question we are only going to use the measurements for 66 of the female athletes (athletics, basketball, rowing and swimming). Suppose we want to create a “parallel lines” model for the relationship between percent body fat (Bfat) and body mass index (BMI which is equal to weight/height^2) – that is the relationship is linear for each sport with a common slope but the intercepts may differ.

Suppose we use the baseline model to create indicator variables for the qualitative factor sport as follows:

$$
\text{sport} \quad B \quad R \quad S \\
\text{BBall} \quad 1 \quad 0 \quad 0 \\
\text{Row} \quad 0 \quad 1 \quad 0 \\
\text{Swim} \quad 0 \quad 0 \quad 1 \\
\text{Athletics} \quad 0 \quad 0 \quad 0
$$

Write the parallel lines model as: $\text{Bfat} = \beta_0 + \beta_B B + \beta_R R + \beta_S S + \beta_{BMI} \text{BMI}$.

Then the fitted coefficients and their covariance matrix is:

$$
\hat{\beta} = \begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_B \\
\hat{\beta}_R \\
\hat{\beta}_S \\
\hat{\beta}_{BMI}
\end{bmatrix} = \begin{bmatrix}
-9.965 \\
6.947 \\
5.015 \\
0.043 \\
1.087
\end{bmatrix} \quad \hat{\Sigma}_\beta = \begin{bmatrix}
8.257 & -0.625 & -0.147 & -0.452 & -0.356 \\
-0.625 & 0.874 & 0.314 & 0.325 & 0.014 \\
-0.147 & 0.314 & 0.644 & 0.317 & -0.008 \\
-0.452 & 0.325 & 0.317 & 1.105 & 0.006 \\
-0.356 & 0.014 & -0.008 & 0.006 & 0.016
\end{bmatrix}
$$

(a) Use this information to:

i. Find a 95% confidence interval for the expected percent body fat for a female swimmer who has BMI = 15.2.

ii. Test the hypothesis that the intercept for basketball players is the same as that for rowers.

(b) There are many ways that the three indicator columns for the factor sport could be created other than the set given above. Suppose we define the first two indicator variables as:

$$
\text{sport} \quad V_1 \quad V_2 \quad V_3 \\
\text{BBall} \quad 1 \quad 1 \quad ? \\
\text{Row} \quad 1 \quad 0 \quad ? \\
\text{Swim} \quad 0 \quad 0 \quad ? \\
\text{Athletics} \quad 0 \quad 0 \quad ?
$$

i. Fill in values for V3 to create a valid set of indicator variables.

ii. Find $\hat{\beta}$ and $\hat{\Sigma}_\beta$ if this new set of indicator variables is used.