Test 2

JARED YU

1. Prove that in using a regression model analysis to compare the differences of the expected values of the response variable y between the K levels of a categorical regressor x, the sum of squares, SS_T, SS_R, SS_{Res}, will not change regardless of how the $K-1$ indicators of x are coded (Recall any dummy variable D can be coded in many ways, e.g., $D = 0,1$ or $D = -1,1$, or others). [10 points]
   State assumptions in each step of your proof.

Ans:

In this problem we are analyzing the situation of a single regressor that is a categorical variable with K levels. It uses a type of encoding that leads to $K-1$ dummy variables. The type of encoding, D, is arbitrary. To show that the encoding doesn't impact SS_T, SS_R, or SS_{Res}, we can first look at the formulas (based on the Textbook) for the total sum of squares (SS_T), regression sum of squares (SS_R), and residual sum of squares (SS_{Res})

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n},$$
$$SS_R = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n},$$
$$SS_{Res} = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.$$

There are alternate forms for these sums of squares, which I have done in Discussion 4. The same reformulations will be shown below. (*Note: An assumption is that $\widehat{\boldsymbol{\beta}}$ is calculatable.*)

SS_T:

It will be shown that an equivalent formula for SS_T is $(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}})$, where $\bar{\mathbf{y}}$ is a $n \times 1$ vector consisting all of identical $\bar{y} = \sum_{i=1}^{n} y_i$ terms.

$$(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\bar{\mathbf{y}} - \bar{\mathbf{y}}'\mathbf{y} + \bar{\mathbf{y}}'\bar{\mathbf{y}}$$
$$= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\bar{\mathbf{y}} + \bar{\mathbf{y}}'\bar{\mathbf{y}}$$

The previous step holds since $\mathbf{y}'\bar{\mathbf{y}}$ and $\bar{\mathbf{y}}'\mathbf{y}$ are both scalars with the same resulting values. So, the transpose of each other is still the same scalar.

Let's first look at $\bar{\mathbf{y}}'\bar{\mathbf{y}}$. It is the inner product of the same vector which is filled with identical $\bar{y}$ terms. Therefore, the result is $\bar{\mathbf{y}}'\bar{\mathbf{y}} = \sum_{i=1}^{n} \bar{y}^2 = n\bar{y}^2$.

Next, let's look at $-2\mathbf{y}'\bar{\mathbf{y}}$. By first ignoring the constant term -2, it can be seen that $\mathbf{y}'\bar{\mathbf{y}} = \sum_{i=1}^{n} y_i\bar{y} = \bar{y}\sum_{i=1}^{n} y_i = \bar{y}n\bar{y} = n\bar{y}^2$. Therefore, $-2\mathbf{y}'\bar{\mathbf{y}} = -2n\bar{y}^2$.

From this it follows that,

$$(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}}) = \mathbf{y}'\mathbf{y} - 2n\bar{y}^2 + n\bar{y}^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = \mathbf{y}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = SS_T.$$

SS_R:

It will be shown that an equivalent formula for SS_R is $(\hat{\mathbf{y}}-\bar{\mathbf{y}})'(\hat{\mathbf{y}}-\bar{\mathbf{y}})$.

$$(\hat{\mathbf{y}}-\bar{\mathbf{y}})'(\hat{\mathbf{y}}-\bar{\mathbf{y}}) = (\hat{\mathbf{y}}'-\bar{\mathbf{y}}')(\hat{\mathbf{y}}-\bar{\mathbf{y}}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} - \hat{\mathbf{y}}'\bar{\mathbf{y}} - \bar{\mathbf{y}}'\hat{\mathbf{y}} + \bar{\mathbf{y}}'\bar{\mathbf{y}}$$

Starting with $\hat{\mathbf{y}}'\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}}'\hat{\mathbf{y}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.$$

Next $\bar{\mathbf{y}}'\bar{\mathbf{y}} = \frac{(\sum_{i=1}^{n} y_i)^2}{n}$ has already been shown in SS_T.

It can be said that $\hat{\mathbf{y}}'\bar{\mathbf{y}}$ and $\bar{\mathbf{y}}'\hat{\mathbf{y}}$ are both scalars where the same transpose logic can be applied as with $\mathbf{y}'\bar{\mathbf{y}}$ and $\bar{\mathbf{y}}'\mathbf{y}$. So, it will be said first that $\hat{\mathbf{y}}'\bar{\mathbf{y}} = \bar{\mathbf{y}}'\hat{\mathbf{y}}$.

$$\hat{\mathbf{y}}'\bar{\mathbf{y}} = \left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\bar{\mathbf{y}} = (\mathbf{H}\mathbf{y})'\left(\frac{1}{n}\mathbf{J}\mathbf{y}\right)$$

Note that $\mathbf{J}$ is an $n \times n$ matrix consisting entirely of 1's.

$$= \frac{1}{n}\mathbf{y}'\mathbf{H}\mathbf{J}\mathbf{y}$$

Note that the hat matrix, $\mathbf{H}$, is a symmetric matrix.

$$= \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y}$$

Note that $\mathbf{HJ} = \mathbf{J}$, this will first be explained. Let $\mathbf{1}$ be an $n \times 1$ vector consisting of 1's only.

$$\mathbf{HJ} = \mathbf{H}[\mathbf{1} \quad \mathbf{1} \quad \cdots \quad \mathbf{1}] = [\mathbf{H1} \quad \mathbf{H1} \quad \cdots \quad \mathbf{H1}] = [\mathbf{1} \quad \mathbf{1} \quad \cdots \quad \mathbf{1}] = \mathbf{J}$$

What is shown above is that first the matrix $\mathbf{J}$ is expressed as a vector of vectors, $[\mathbf{1} \quad \mathbf{1} \quad \cdots \quad \mathbf{1}]$. The hat matrix is then multiplied to this to get to $[\mathbf{H1} \quad \mathbf{H1} \quad \cdots \quad \mathbf{H1}]$. This is simply another way to show typical matrix multiplication. The point is that the hat matrix is a projection matrix on the column space of $\mathbf{X}$, the design matrix. As the design matrix, it includes in the first column a vector of 1's, or $\mathbf{1}$. $\mathbf{H}$ times $\mathbf{1}$, i.e. project $\mathbf{1}$ onto the column space of $\mathbf{X}$, is equal to $\mathbf{1}$ since $\mathbf{1}$ is already in the column space of $\mathbf{X}$. $\mathbf{J}$ is a $n \times n$ matrix filled with all 1's. The result then is that $\mathbf{H}$ multiplied by $\mathbf{J}$ is also equal to $\mathbf{J}$.

$$\cdots = \frac{1}{n}\left[\sum_{i=1}^{n} y_i \quad \cdots \quad \sum_{i=1}^{n} y_i\right]\mathbf{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \left(\sum_{i=1}^{n} y_i\right) = \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

Therefore,

$$\hat{\mathbf{y}}'\hat{\mathbf{y}} - \hat{\mathbf{y}}'\bar{\mathbf{y}} - \bar{\mathbf{y}}'\hat{\mathbf{y}} + \bar{\mathbf{y}}'\bar{\mathbf{y}} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - 2\frac{(\sum_{i=1}^{n} y_i)^2}{n} + \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = SS_R$$

SS_{Res}:

It will be shown that an equivalent formula for SS_{Res} is $(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$.

$$(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}$$
$$= \mathbf{y}'\mathbf{y} - 2\widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = SS_{Res}$$

Now, looking at the formulas for each of these sums of squares, it can be seen first that SS_T only contains the y_i terms. Therefore, any sort of encoding will not directly impact it. So, this sum of squares is unaffected. However, it can be seen that SS_R and SS_{Res} both depend on $\hat{\mathbf{y}}$ in addition to the y_i terms. It must be shown then that any type of encoding D will have no impact on the resulting $\hat{\mathbf{y}}$ term. It is worth noting however, that changing the encoding will change both the $\mathbf{X}$ term and $\widehat{\boldsymbol{\beta}}$ term which when multiplied together make up $\hat{\mathbf{y}}$.

To prove then that D does not impact $\hat{\mathbf{y}}$, it will be shown that the hat matrix $\mathbf{H}$, which is the projection matrix that projects the vector $\mathbf{y}$ onto the column space of $\mathbf{X}$, does not change due to changes in the type of encoding, D. The logic is that since $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, if $\mathbf{H}$ is unimpacted, then $\hat{\mathbf{y}}$ is unchanged. From this it would be clear that SS_T, SS_R, SS_{Res} is unimpacted by the encoding D.

We can start by looking at $\mathbf{X}$. Given that it is the design matrix for the regressor, it can be written as follows

$$\mathbf{X} = [\mathbf{1} \quad a_{1,1}\mathbf{1}_{1,1} + a_{1,2}\mathbf{1}_{1,2} \quad \cdots \quad a_{k-1,1}\mathbf{1}_{k-1,1} + a_{k-1,2}\mathbf{1}_{k-1,2}]$$

The meaning is such that the first column, $\mathbf{1}$, is simply the $n \times 1$ vector consisting only of 1's that are used to represent the intercept term in the linear regression model. For each of the $K-1$ categories that are encoded as dummy variables, they are binary in the sense that they have one value in the case that the observation is true for that level and another value otherwise. In the common case, it is 1 if the observation is true for that level and 0 otherwise. It can however be changed to something like 1 if true and -1 otherwise, or any other variety of values that act in such a way.

We can generalize this to something like $a_{1,1}\mathbf{1}_{1,1} + a_{1,2}\mathbf{1}_{1,2}$ in the case of the first level out of the $K-1$ levels. We can think of $a_{1,1}$ and $a_{1,2}$ as the two possible types of encodings for a level. For example, in the case of 1, -1 for the type of encoding, then $a_{1,1} = 1$ and $a_{1,2} = -1$. In the case that the encoding varies per level, then each of the $a_{p,1}$ and $a_{p,2}$ can vary for each of the levels $1, \cdots, K-1$ that p can take on. The $\mathbf{1}_{1,1}$ and $\mathbf{1}_{1,2}$ terms are such that they are $n \times 1$ vectors with binary values 1 and 0. Each of them are 1 in the case that an observation is true for that treatment and 0 otherwise.

To better understand these, we can first look at a simple example. Let's say that there are $K = 3$ levels and so we have 2 dummy variables. The $K = 3$ levels represent treatment A, B, and C respectively. To use $K - 1 = 2$ dummy variables, we drop the level C from representation and say that

$$x_1 = \begin{cases} 1 & \text{if the observation is from treatment } A \\ -1 & \text{otherwise,} \end{cases}$$
$$x_2 = \begin{cases} 1 & \text{if the observation is from treatment } B \\ -1 & \text{otherwise.} \end{cases}$$

This would lead to the following regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where $\varepsilon \sim N(0, \sigma^2)$. In the case that the first observation is from treatment A, and the second is from treatment B, it can be seen as following in the design matrix

$$\begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix},$$

where the first row corresponds to the first observation and the second corresponds with the second. The first column is all 1's to go with the intercept term β_0. The second and third columns correspond with x_1 and x_2 respectively. This is similar to what has been in the textbook so far. However, to do the generalized encoding we can represent the second column as

$$a_{1,1}\mathbf{1}_{1,1} + a_{1,2}\mathbf{1}_{1,2}$$

and the third column as

$$a_{2,1}\mathbf{1}_{2,1} + a_{2,2}\mathbf{1}_{2,2}.$$

We can first focus on the second column, which when looking at the design matrix is $(1 \quad -1)'$. In such a case, we can represent it using a set of two binary vectors multiplied by their corresponding encodings. This can be seen as follows

$$a_{1,1}\mathbf{1}_{1,1} + a_{1,2}\mathbf{1}_{1,2} = 1\begin{pmatrix}1\\0\end{pmatrix} + (-1)\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1\\-1\end{pmatrix}$$

which can also be extended to the third column

$$a_{2,1}\mathbf{1}_{2,1} + a_{2,2}\mathbf{1}_{2,2} = 1\begin{pmatrix}0\\1\end{pmatrix} + (-1)\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}-1\\1\end{pmatrix}$$

As can be seen above, the $a_{p,1}$ and $a_{p,2}$ correspond with the positive and negative encodings of D respectively for the pth treatment level of the categorical variable. The $\mathbf{1}_{p,1}$ and $\mathbf{1}_{p,2}$ are binary vectors that act as Boolean values to help indicate for a certain observation whether it should be positive or negative for a treatment level, after multiplying by the $a_{p,1}$ and $a_{p,2}$ values.

Next, we can try to better frame the argument for how we can compare to hat matrices. We can denote the hat matrix derived from a matrix with one arbitrary type of encoding as $\mathbf{H}_1$ and the hat matrix derived from another matrix with a different arbitrary type of encoding as $\mathbf{H}_2$. Therefore, we would like to know if it's guaranteed that in such a regression model, $\mathbf{H}_1 = \mathbf{H}_2$. (*Note: An assumption is that the encoding, D, consists of real numbers only.*)

Thinking back to $\mathbf{X}$ we have,

$$\mathbf{X} = [\mathbf{1} \quad a_{1,1}\mathbf{1}_{1,1} + a_{1,2}\mathbf{1}_{1,2} \quad \cdots \quad a_{k-1,1}\mathbf{1}_{k-1,1} + a_{k-1,2}\mathbf{1}_{k-1,2}]$$

where the column space of $\mathbf{X}$, denoted $\text{col}(\mathbf{X})$ can be seen as

$$\text{col}(\mathbf{X}) = \left\{ \sum_{i=1}^{k-1} b_i \mathbf{x}_i \,, \qquad b_i \in \mathbb{R} \right\},$$

however, this is in the general sense where $\mathbf{X}$ hasn't yet been given the special $a_{p,1}\mathbf{1}_{p,1} + a_{p,2}\mathbf{1}_{p,2}$ generalization. So, to combine the two we have

$$\text{col}(\mathbf{X}) = \left\{ b_0\mathbf{1} + \sum_{i=1}^{k-1} b_i(a_{i1}\mathbf{1}_{i1} + a_{i2}\mathbf{1}_{i2}) \,, \qquad b_i \in \mathbb{R} \right\}$$

$$= \left\{ b_0\mathbf{1} + \sum_{i=1}^{k-1} (b_i a_{i1})\mathbf{1}_{i1} + \sum_{i=1}^{k-1} (b_i a_{i2})\mathbf{1}_{i2} \,, \qquad b_i \in \mathbb{R} \right\}$$

$$= \left\{ c_0\mathbf{1} + \sum_{i=1}^{k-1} c_{i1}\mathbf{1}_{i1} + \sum_{i=1}^{k-1} c_{i2}\mathbf{1}_{i2} \,, \qquad c_0, c_{i1}, c_{i2} \in \mathbb{R} \right\}$$

This $\text{col}(\mathbf{X})$ can be thought of as the span of all the vectors $\mathbf{1}$, $\mathbf{1}_{i1}$, and $\mathbf{1}_{i2}$ which can be denoted as

$$\text{span}\{\mathbf{1}, \mathbf{1}_{1,1}, \mathbf{1}_{1,2}, \mathbf{1}_{2,1}, \mathbf{1}_{2,2}, \cdots, \mathbf{1}_{k-1,1}, \mathbf{1}_{k-1,2}\}.$$

From this, it can be understood that for any type of encoding used for the design matrix $\mathbf{X}$, the span will be identical.

Then the hat matrix, $\mathbf{H}$, which is the projection matrix of $\mathbf{y}$ onto $\text{col}(\mathbf{X})$, for any type of encoding will lead to the same identical $\mathbf{H}$. To more easily think about it, the hat matrix projects $\mathbf{y}$ vertically onto the column space (in the simple case of $\mathbf{X}$ only have two dimensions). Since the column space is the same for any type of encoding, the resulting projection matrix must also be identical across all encodings, D. It can then be understood that for any two arbitrary types of encoding that $\mathbf{H}_1 = \mathbf{H}_2$.

Therefore, since $\mathbf{H}$ is unimpacted by the type of encoding, then $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ will also be unimpacted by the encoding. It makes sense that not only is SS_T unchanged, but SS_R and SS_{Res} which are dependent on $\hat{\mathbf{y}}$ must also remain unchanged. ■

2. In a study, there are four treatments (labeled as 1, 2, 3, 4) to compare. Assume that there are m subjects per treatment.
    a. Construct an analysis of variance model to compare the four treatments; that is, test whether there is at least one pair of treatments that differ and construct an estimator of every pair of expected treatment difference. [20 points]

Ans:

An analysis of variance model for such a study can be seen below,

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1,2,3,4, \qquad j = 1,2\cdots,m$$

where Y_{ij} is the jth observation for the ith treatment, μ is the grand mean, τ_i is a parameter that represents the effect of the ith treatment, and ε_{ij} is an $NID(0,\sigma^2)$ error component.

To test whether at least one pair of treatments differ, it is possible to look at it using the following hypothesis test,

$$H_0: \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0 \text{ vs. } H_1: \tau_i \neq 0 \text{ for at least one } i.$$

The reason is that if every pair of treatments are the same, they must be all equal to zero based on the condition that $\sum_{i=1}^{4} \tau_i = 0$ within the analysis of variance model. The test statistic for such a hypothesis is as follows,

$$F_0 = \frac{MS_{Treatment}}{MS_{Res}} = \frac{\frac{SS_{Treatment}}{4-1}}{\frac{SS_{Res}}{4(m-1)}} = \frac{\frac{m\sum_{i=1}^{4}(\bar{y}_{i.} - \bar{y}_{..})^2}{4-1}}{\frac{\sum_{i=1}^{4}\sum_{j=1}^{m}(y_{ij} - \bar{y}_{i.})^2}{4(m-1)}},$$

where $\bar{y}_{i.} = \frac{\sum_{j=1}^{m} y_{ij}}{m}$ and $\bar{y}_{..} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}}{4m}$. If $F_0 > F_{\alpha,3,4(m-1)}$, then we reject the null hypothesis in favor of the alternative. Otherwise we fail to reject the null hypothesis at significance level α. In the case of rejecting the null hypothesis in favor of the alternative, then we conclude that at least one pair of treatments are different. If we fail to reject the null hypothesis, then the conclusion is that there is no difference across the different treatment levels.

To construct an estimator for every pair of expected treatment difference, there would be $\binom{4}{2} = 6$ total possibilities. The below Table 1 shows all such combinations. *Note: It is possible to show more combinations such as not only Treatment 1 – Treatment 2, but also Treatment 2 – Treatment 1. However, this would be repetitive and not show any new interesting information so it will be skipped.*

Table 1 The below table shows all the different combinations for comparing the difference between treatments.

Compared differences	Difference in expected value	Estimator
Treatment 1 – Treatment 2	$E(Y_{1j}) - E(Y_{2j})$ $= \mu + \tau_1 - (\mu + \tau_2)$ $= \tau_1 - \tau_2$	$\bar{y}_{1.} - \bar{y}_{2.}$ $= \frac{1}{m}\left[\sum_{j=1}^{m} y_{1j} - y_{2j}\right]$
Treatment 1 – Treatment 3	$E(Y_{1j}) - E(Y_{3j})$ $= \mu + \tau_1 - (\mu + \tau_3)$ $= \tau_1 - \tau_3$	$\bar{y}_{1.} - \bar{y}_{3.}$ $= \frac{1}{m}\left[\sum_{j=1}^{m} y_{1j} - y_{3j}\right]$

Treatment 1 – Treatment 4	$E(Y_{1j}) - E(Y_{4j}) = \mu + \tau_1 - (\mu + \tau_4) = \tau_1 - \tau_4$	$\bar{y}_{1.} - \bar{y}_{4.} = \frac{1}{m}\left[\sum_{j=1}^{m} y_{1j} - y_{4j}\right]$
Treatment 2 – Treatment 3	$E(Y_{2j}) - E(Y_{3j}) = \mu + \tau_2 - (\mu + \tau_3) = \tau_2 - \tau_3$	$\bar{y}_{2.} - \bar{y}_{3.} = \frac{1}{m}\left[\sum_{j=1}^{m} y_{2j} - y_{3j}\right]$
Treatment 2 – Treatment 4	$E(Y_{2j}) - E(Y_{4j}) = \mu + \tau_2 - (\mu + \tau_4) = \tau_2 - \tau_4$	$\bar{y}_{2.} - \bar{y}_{4.} = \frac{1}{m}\left[\sum_{j=1}^{m} y_{2j} - y_{4j}\right]$
Treatment 3 – Treatment 4	$E(Y_{3j}) - E(Y_{4j}) = \mu + \tau_3 - (\mu + \tau_4) = \tau_3 - \tau_4$	$\bar{y}_{3.} - \bar{y}_{4.} = \frac{1}{m}\left[\sum_{j=1}^{m} y_{3j} - y_{4j}\right]$

b. Construct a linear regression model such that the regression analysis is equivalent to the analysis of variance in (a). [20 points]

Ans:

The same analysis of variance model in part a) can be rewritten using a regression model. Since there are four treatment levels, it would require three indicator variables defined as follows:

$$x_1 = \begin{cases} 1 & \text{if the observation is from treatment 1} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{if the observation is from treatment 2} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if the observation is from treatment 3} \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding regression model then becomes

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \varepsilon_{ij}, \qquad i = 1,2,3,4, \qquad j = 1,2,\cdots,m$$

where x_{1j} is the value of the indicator variable x_1 for observation j in treatment i, x_{2j} is the value of the indicator variable x_2 for observation j in treatment i, and x_{3j} is the value of the indicator variable x_3 for observation j in treatment i. Furthermore, ε_{ij} is an $NID(0, \sigma^2)$ error component.

Comparing the regression model to the analysis of variance model, with a total of four treatments in the study, the regression coefficients can be understood as follows

$$\beta_0 = \mu_4, \beta_1 = \mu_1 - \mu_4, \beta_2 = \mu_2 - \mu_4, \beta_3 = \mu_3 - \mu_4$$

The above formulation comes from the following basic example of treatment 1, where

$$x_{1j} = 1 \text{ and } x_{ij} = 0 \text{ for } i \neq 1$$

Applying the above example of the jth observation of treatment 1, it leads to

$$y_{1j} = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0) + \varepsilon_{1j} = \beta_0 + \beta_1 + \varepsilon_{1j}$$

Then, looking towards the analysis of variance model we have,

$$y_{1j} = \mu + \tau_1 + \varepsilon_{1j} = \mu_1 + \varepsilon_{1j}$$

Comparing these two results we have

$$\beta_0 + \beta_1 + \varepsilon_{1j} = \mu_1 + \varepsilon_{1j}$$
$$\rightarrow \beta_1 = \mu_1 - \beta_0$$

Likewise, for any treatment $i = 1, \cdots, 3$, we have

$$\beta_i = \mu_i - \beta_0$$

Lastly, for treatment 4, the resulting regression model is

$$y_{4j} = \beta_0 + \varepsilon_{4j}$$

while the analysis-of-variance model is

$$y_{4j} = \mu + \tau_4 = \mu_4 + \varepsilon_{4j}$$

Therefore, for treatment 4,

$$\beta_0 = \mu_4$$

From the above derivation, it follows then that

$$\beta_0 = \mu_4, \beta_1 = \mu_1 - \mu_4, \beta_2 = \mu_2 - \mu_4, \beta_3 = \mu_3 - \mu_4$$

as stated previously.

From here, to test the null hypothesis, we can instead of

$$\mu_1 = \mu_2 = \mu_3 = \mu_4$$

we can test

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_1: \beta_i \neq 0 \text{ for at least one } i.$$

The reason that we can use such a hypothesis test is that if $\beta_i = 0$ for $i = 1, 2, 3$, then $\mu_i = \mu_4$ for $i = 1, 2, 3$ (based on the above derivation for β_i, $i = 0, \cdots, 4$). Such a hypothesis then is equivalent to testing $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

To test the above hypothesis test we can use the following test statistic,

$$F_0 = \frac{\frac{SS_R(\beta_1, \beta_2, \beta_3 | \beta_0)}{4 - 1}}{\frac{SS_{Res}}{4m - 4}}.$$

To understand the test statistic, first the coefficients will be stated:

$$\hat{\beta}_0 = \bar{y}_{..} - \bar{y}_{1.} - \bar{y}_{2.} - \bar{y}_{3.} = \bar{y}_{4.}$$
$$\hat{\beta}_1 = \bar{y}_{1.} - \bar{y}_{4.}$$
$$\hat{\beta}_2 = \bar{y}_{2.} - \bar{y}_{4.}$$
$$\hat{\beta}_3 = \bar{y}_{3.} - \bar{y}_{4.}$$

The full model of the regression sum of squares is as follows

$$SS_R(\beta_0, \beta_1, \beta_2, \beta_3) = \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}\right)^2}{4m}$$

$$= [\bar{y}_{4.} \quad \bar{y}_{1.} - \bar{y}_{4.} \quad \bar{y}_{2.} - \bar{y}_{4.} \quad \bar{y}_{3.} - \bar{y}_{4.}] \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \\ y_{3.} \end{bmatrix} - \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}\right)^2}{4m}$$

$$= \bar{y}_{4.} y_{..} + y_{1.}(\bar{y}_{1.} - \bar{y}_{4.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{4.}) + y_{3.}(\bar{y}_{3.} - \bar{y}_{4.}) - \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}\right)^2}{4m}$$

$$= \bar{y}_{4.}(y_{1.} + y_{2.} + y_{3.} + y_{4.}) + y_{1.}(\bar{y}_{1.} - \bar{y}_{4.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{4.}) + y_{3.}(\bar{y}_{3.} - \bar{y}_{4.}) - \frac{\left(\sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}\right)^2}{4m}$$

$$= \sum_{i=1}^{4} y_{i.}\bar{y}_{i.} - \frac{(y_{..})^2}{4m} = \sum_{i=1}^{4} y_{i.}\left(\frac{y_{i.}}{m}\right) - \frac{(y_{..})^2}{4m} = \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - \frac{(y_{..})^2}{4m}$$

The reduced model regression sum of squares is as follows

$$SS_R(\beta_0) = \hat{\beta}_0 \mathbf{1}'_{4m}\mathbf{y} - \frac{(y_{..})^2}{4m} = \bar{y}_{..}y_{..} - \frac{(y_{..})^2}{4m} = \frac{(y_{..})^2}{4m} - \frac{(y_{..})^2}{4m} = 0$$

From here it follows that the partial regression sum of squares is as follows

$$SS_R(\beta_1, \beta_2, \beta_3 | \beta_0) = SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_R(\beta_0)$$

$$= \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - \frac{(y_{..})^2}{4m} - 0 = m\sum_{i=1}^{4}(\bar{y}_{i.} - \bar{y}_{..})^2$$

This last step will be proven below

$$m\sum_{i=1}^{4}(\bar{y}_{i.} - \bar{y}_{..})^2 = m\sum_{i=1}^{4}(\bar{y}_{i.}^2 - 2\bar{y}_{i.}\bar{y}_{..} + \bar{y}_{..}^2) = m\left(\sum_{i=1}^{4}\bar{y}_{i.}^2 - 2\bar{y}_{..}\sum_{i=1}^{4}\bar{y}_{i.} + 4\bar{y}_{..}^2\right)$$

$$= m\left[\sum_{i=1}^{4}\left(\frac{y_{i.}}{m}\right)^2 - 2\left(\frac{y_{..}}{4m}\right)\sum_{i=1}^{4}\left(\frac{y_{i.}}{m}\right) + 4\left(\frac{y_{..}}{4m}\right)^2\right] = \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - \frac{y_{..}}{2}\left(\frac{\sum_{i=1}^{4} y_{i.}}{m}\right) + \frac{y_{..}^2}{4m}$$

$$= \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - 2y_{..}\left(\frac{\sum_{i=1}^{4} y_{i.}}{4m}\right) + \frac{y_{..}^2}{4m} = \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - 2\left(\frac{y_{..}^2}{m}\right) + \frac{y_{..}^2}{4m} = \frac{\sum_{i=1}^{4} y_{i.}^2}{m} - \frac{y_{..}^2}{4m}$$

Then for the denominator, SS_{Res} can be seen as follows

$$SS_{Res} = \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} = \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - \frac{\sum_{i=1}^{4} y_{i.}^2}{m} = \sum_{i=1}^{4}\sum_{j=1}^{m}\left(y_{ij}^2 - \bar{y}_{i.}\right)^2$$

The proof for the last step will be shown below

$$\sum_{i=1}^{4}\sum_{j=1}^{m}\left(y_{ij}^2 - \bar{y}_{i.}\right)^2 = \sum_{i=1}^{4}\sum_{j=1}^{m}\left(y_{ij}^2 - 2y_{ij}\bar{y}_{i.} + \bar{y}_{i.}^2\right)$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - 2\sum_{i=1}^{4}\bar{y}_{i.}\left(\sum_{j=1}^{m} y_{ij}\right) + m\sum_{i=1}^{4}\bar{y}_{i.}^2 = \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - 2\sum_{i=1}^{4}\left[\left(\frac{y_{i.}}{m}\right)y_{i.}\right] + m\sum_{i=1}^{4}\left(\frac{y_{i.}}{m}\right)^2$$

$$= \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - 2\frac{\sum_{i=1}^{4} y_{i.}^2}{m} + \frac{\sum_{i=1}^{4} y_{i.}^2}{m} = \sum_{i=1}^{4}\sum_{j=1}^{m} y_{ij}^2 - \frac{\sum_{i=1}^{4} y_{i.}^2}{m}$$

From this it follows that the test statistic becomes

$$F_0 = \frac{\frac{SS_R(\beta_1, \beta_2, \beta_3 | \beta_0)}{4-1}}{\frac{SS_{Res}}{4m-4}} = \frac{\frac{m\sum_{i=1}^{4}(\bar{y}_{i.} - \bar{y}_{..})^2}{4-1}}{\frac{\sum_{i=1}^{4}\sum_{j=1}^{m}\left(y_{ij}^2 - \bar{y}_{i.}\right)^2}{4m-4}} = \frac{MS_{Treatment}}{MS_{Res}}$$

This derived test statistic is the same as the one in the previous part. It likewise follows the $F_{3,4m-4}$ distribution. ■

State assumptions in each step of your proof or derivation in a) and b).

Do not use any math/stat software for calculation for Problem 3, except for obtaining the percentile of standard normal, t, chi-square, or F distribution, matrix operations, or basic mathematical calculations.

3. Ten observations on the response variable y associated with two regressor variables x_1 and x_2 are given in the following table. The model fitted to these observations is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma x_{1i} x_{2i} + \varepsilon_i, \qquad i = 1, \cdots, n, \tag{1}$$

where ε's are identically and independently distributed as normal random variable with mean zero and a known variance $\sigma^2 = 4$.

   a. Test the null hypothesis "there is no difference between the y-intercept for $x_2 = 1$ and the y-intercept for $x_2 = -1$ and there is no difference between the slope for $x_2 = 1$ and the slope for $x_2 = -1$" at a statistical significance level of 0.05. [20 pts]

Ans:

In the case of $x_2 = 1$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 + \gamma x_{1i} + \varepsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \gamma) x_{1i} + \varepsilon_i \end{aligned}$$

In the case of $x_2 = -1$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} - \beta_2 - \gamma x_{1i} + \varepsilon_i \\ &= (\beta_0 - \beta_2) + (\beta_1 - \gamma) x_{1i} + \varepsilon_i \end{aligned}$$

Then the null hypothesis is stating that

$$H_0: \begin{cases} \beta_0 + \beta_2 = \beta_0 - \beta_2 \\ \beta_1 + \gamma = \beta_1 - \gamma \end{cases}$$

$$\rightarrow H_0: \begin{cases} \beta_2 = 0 \\ \gamma = 0 \end{cases}$$

Therefore, the null can be hypothesis stated as

$$H_0: \beta_2 = \gamma = 0$$

Turning this into a model, we can state both the null and the alterative as,

$$H_0: y = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \text{ vs. } H_1: y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma x_{i1} x_{i2} + \varepsilon_i$$

The above is a partial F test, where we are comparing a reduced model to the full model. The test statistic then can be seen as follows,

$$F_0 = \frac{SS_R(\beta_2, \gamma | \beta_0, \beta_1)/2}{MS_{Res}}$$

where F_0 follows a $F_{2,10-4}$ distribution.

The numerator can be understood as follows

$$SS_R(\beta_2, \gamma | \beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2, \gamma) - SS_R(\beta_0, \beta_1)$$

We can first look at the reduced model $SS_R(\beta_0, \beta_1)$

$$SS_R(\beta_0, \beta_1) = \widehat{\boldsymbol{\beta}}_1 \mathbf{X}_1' \mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

where $\widehat{\boldsymbol{\beta}}_1$ are the set of coefficients estimates pertaining to the reduced model, and $\mathbf{X}_1$ is the reduced dataset used to generate that set of coefficient estimators. The full model regression sum of squares can be seen as follows

$$SS_R(\beta_0, \beta_1, \beta_2, \gamma) = \widehat{\boldsymbol{\beta}} \mathbf{X}' \mathbf{y} - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

Therefore,

$$SS_R(\beta_2, \gamma|\beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2, \gamma) - SS_R(\beta_0, \beta_1)$$
$$= \widehat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n} - \left[\widehat{\boldsymbol{\beta}}_1\mathbf{X}_1'\mathbf{y} - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right] = (\widehat{\boldsymbol{\beta}}\mathbf{X}' - \widehat{\boldsymbol{\beta}}_1\mathbf{X}_1')\mathbf{y}$$

From this, the test statistic can be calculated to be $F_0 \approx 4.4918$, which has a corresponding p-value of ≈ 0.0642. This p-value is larger than the significance level $\alpha = 0.05$, therefore the decision is that we fail to reject the null hypothesis. The conclusion then is that there is not enough evidence to reject the claim that there is no difference between the y-intercept for $x_2 = 1$ and the y-intercept for $x_2 = -1$ and there is no difference between the slope for $x_2 = 1$ and the slope for $x_2 = -1$

b. Estimate the difference, $E(y|x_1 = 5, x_2 = 1) - E(y|x_1 = 5, x_2 = -1)$, and calculate its 95% confidence interval. [10 pts]

Ans:

(*Note: Here the full model in part a) is being used, rather than the reduced model.*) We can estimate first try to estimate the difference by finding out the formula for the expected differences between the two points.

$$E(y|x_1 = 5, x_2 = 1) = \beta_0 + \beta_1(5) + \beta_2(1) + \gamma(5)(1)$$
$$= \beta_0 + 5\beta_1 + \beta_2 + 5\gamma$$

$$E(y|x_1 = 5, x_2 = -1) = \beta_0 + \beta_1(5) + \beta_2(-1) + \gamma(5)(-1)$$
$$= \beta_0 + 5\beta_1 - \beta_2 - 5\gamma$$

$$\rightarrow E(y|x_1 = 5, x_2 = 1) - E(y|x_1 = 5, x_2 = -1)$$
$$= \beta_0 + 5\beta_1 + \beta_2 + 5\gamma - (\beta_0 + 5\beta_1 - \beta_2 - 5\gamma)$$
$$= 2\beta_2 + 10\gamma$$

So, we can estimate the difference by using the estimated coefficients $\hat{\beta}_2$ and $\hat{\gamma}$. The resulting $2\hat{\beta}_2 + 10\hat{\gamma} = 2(-1.5353) + 10(-0.1315)$ $\boxed{\approx -4.3855}$.

Let $\hat{y}_{01} = E(y|x_1 = 5, x_2 = 1)$ and $\hat{y}_{02} = E(y|x_1 = 5, x_2 = -1)$ be the estimated values given $\mathbf{x}_{01}$ and $\mathbf{x}_{02}$ respectively, where $\mathbf{x}_{01} = (5,1)^{\mathsf{T}}$ and $\mathbf{x}_{02} = (5,-1)^{\mathsf{T}}$. To find the confidence interval of this difference, we must first find the variance of the difference. This will be shown as follows

$$Var(\hat{y}_{01} - \hat{y}_{02}) = Var\left[(\mathbf{x}_{01} - \mathbf{x}_{02})'\widehat{\boldsymbol{\beta}}\right]$$
$$= (\mathbf{x}_{01} - \mathbf{x}_{02})'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02}) = \sigma^2(\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02})$$

From this, we can try to normalize the difference

$$\frac{\hat{y}_{01} - \hat{y}_{02} - (y_{01} - y_{02})}{\sqrt{\sigma^2(\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02})}} \sim N(0,1)$$

So, we see that the confidence interval is

$$\hat{y}_{01} - \hat{y}_{02} \pm z_{\frac{\alpha}{2}}\sqrt{\sigma^2(\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02})}$$

Therefore, it follows that the 95% C.I. is

$$(4.3837 - 8.7692) \pm 2.5158$$
$$\boxed{\approx [-6.9013, -1.8697]}$$

c. Predict the difference in y value at $x_1 = 5$ between $x_2 = 1$ and $x_2 = -1$. [10 pts]

Ans:

We can think of a predicted observation as $\hat{y}_0 = \mathbf{x}_0'\widehat{\boldsymbol{\beta}}$. For a new example separate from the dataset, we can denote it as $\hat{y}_0^{(new)} = \mathbf{x}_0'\widehat{\boldsymbol{\beta}} + \varepsilon^{(new)}$. Then for the two new observations $\mathbf{x}_{01}$ and $\mathbf{x}_{02}$, we have $\hat{y}_{01}^{(new)} = \mathbf{x}_{01}'\widehat{\boldsymbol{\beta}} + \varepsilon_1^{(new)}$ and $\hat{y}_{02}^{(new)} = \mathbf{x}_{02}'\widehat{\boldsymbol{\beta}} + \varepsilon_2^{(new)}$. We must then find out the variance for the difference between the two. (*Note: An assumption is that the random errors have the same distribution as the random errors from the sample data.*)

$$Var\left(\hat{y}_{01}^{(new)} - \hat{y}_{02}^{(new)}\right) = Var\left[(\mathbf{x}_{01} - \mathbf{x}_{02})'\widehat{\boldsymbol{\beta}}\right] + Var\left(\varepsilon_1^{(new)}\right) + Var\left(\varepsilon_2^{(new)}\right)$$
$$= \sigma^2(\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02}) + \sigma^2 + \sigma^2$$
$$= \sigma^2[2 + (\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02})]$$

So, it follows that a $100(1-\alpha)\%$ P.I. for $\hat{y}_{01}^{(new)} - \hat{y}_{02}^{(new)}$ is as follows

$$\hat{y}_{01} - \hat{y}_{02} \pm z_{\frac{\alpha}{2}}\sqrt{\sigma^2[2 + (\mathbf{x}_{01} - \mathbf{x}_{02})'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{x}_{01} - \mathbf{x}_{02})]}$$

Therefore, it follows that a 95% P.I. is

$$(4.3837 - 8.7692) \pm 4.6578$$
$$\boxed{\approx [-9.0433, 0.2723]}$$

The question however doesn't specifically ask for a prediction interval. It asks to predict the difference between the two predictions for the two $\mathbf{x}_{01}$ and $\mathbf{x}_{02}$ observations. In such a case, it would be the same as in the previous part where we are looking at the estimated difference $\boxed{\approx -4.3855}$. However, from the Textbook so far, in the situations of discussing predictions, it has only been the case that we are looking at prediction intervals rather than simply the predictions themselves.

d. Now fit Model (2): $y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$ to the 10 observations. Calculate the residual for the observation #8 and its variance. [10 pts]

Ans:

The Model 2 can be fit using linear regression. It leads to the following estimated coefficients $\hat{\beta}_0 = 7.1$ and $\hat{\beta}_1 = -1.9$. The residual for the 8th observation is $y_8 - \hat{y}_8 = 10 - 9$ $\boxed{= 1}$. The variance for this residual can be calculated as follows

$$e = y - \hat{y} = \mathbf{y} - \mathbf{x}'\widehat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$
$$Var(e) = Var\big((\mathbf{I} - \mathbf{H})\mathbf{y}\big) = (\mathbf{I} - \mathbf{H})Var(\mathbf{y})(\mathbf{I} - \mathbf{H})'$$
$$= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' = \sigma^2(\mathbf{I} - \mathbf{H})^2 = \sigma^2(\mathbf{I} - \mathbf{H})$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. So, it follows that the variance for the 8th observation is the value in row 8, column 8 of $\sigma^2(\mathbf{I} - \mathbf{H})$. Calculating this in RStudio leads to a value of $\boxed{1.6}$.

State assumptions in your derivations and calculations in a), b), c), d).

Code Appendix

```
### Problem 3
df <- data.frame(
  y=c(7,8,5,4,2,10,9,10,8,8),
  x1=c(9,6,10,8,5,7,6,5,5,4),
  x2=rep(c(1,-1), each=5)
)
y <- df[,1]
n <- nrow(df)
ones <- rep(1, n)
```

```
x1 <- df[,2]; x2 <- df[,3]
X <- cbind(ones, x1, x2, x1*x2)

### part (a)
beta_hat_calc <- function(X, y) {
  beta_hat <- solve(t(X) %*% X) %*% t(X) %*% y
  return(beta_hat)
}
H_calc <- function(X) {
  H <- X %*% solve(t(X) %*% X) %*% t(X)
  return(H)
}
y_hat_calc <- function(H, y) {
  y_hat <- H %*% y
  return(y_hat)
}
e_calc <- function(y, y_hat) {
  e <- y - y_hat
  return(e)
}

X_red <- cbind(ones, x1)
beta_hat_full <- beta_hat_calc(X = X, y = y)
beta_hat_red <- beta_hat_calc(X = X_red, y = y)

SS_R_calc <- function(beta_hat, X, y) {
  n <- length(y)
  SS_R <- (t(beta_hat) %*% t(X) %*% y) - ((sum(y)^2) / n)
  return(SS_R)
}
SS_Res_calc <- function(y, beta_hat, X) {
  SS_Res <- (t(y) %*% y) - (t(beta_hat) %*% t(X) %*% y)
  return(SS_Res)
}

SS_Res_full <- SS_Res_calc(y = y, beta_hat = beta_hat, X = X)
k <- 3; p <- k + 1
r <- 2
MS_Res <- SS_Res_full / (n - p)
SS_R_full <- SS_R_calc(beta_hat = beta_hat, X = X, y = y)
SS_R_red <- SS_R_calc(beta_hat = beta_hat_red, X = X_red, y = y)

F_0 <- ((SS_R_full - SS_R_red) / r) / MS_Res
alpha <- 0.05
qf(p = (1 - alpha), df1 = k, df2 = (n - k - 1))
pf(q = F_0, df1 = r, df2 = (n - k - 1), lower.tail = FALSE)

### part (b)
2 * beta_hat[3] + 10 * beta_hat[4] # -4.38551

z_value <- qnorm(alpha/2, lower.tail = FALSE)
x_01 <- c(1, 5, 1, 5)
x_02 <- c(1, 5, -1, -5)
y_hat_01 <- t(x_01) %*% beta_hat
y_hat_02 <- t(x_02) %*% beta_hat
sigma_squared <- 2

CI_bound <- z_value * sqrt(sigma_squared * t(x_01 - x_02) %*%
                           solve(t(X) %*% X) %*%
                           (x_01 - x_02))

(y_hat_01 - y_hat_02) + CI_bound
(y_hat_01 - y_hat_02) - CI_bound

### part (c)
PI_bound <- z_value * sqrt(sigma_squared *
```

```
                    (2 +
                      t(x_01 - x_02) %*%
                      solve(t(X) %*% X) %*%
                      (x_01 - x_02)))

(y_hat_01 - y_hat_02) + PI_bound
(y_hat_01 - y_hat_02) - PI_bound

### part (d)
X_d <- X[,c(1,3)]
beta_hat_d <- beta_hat_calc(X = X_d, y = y)
H_d <- H_calc(X = X_d)
y_hat_d <- y_hat_calc(H = H_d, y = y)
e_d <- e_calc(y = y, y_hat = y_hat_d)

e_d[8]

residual_variance <- sigma_squared * (diag(n) - H_d)
residual_variance[8,8]
```