

Homework 1 - BIOS 6643

Dominic Adducci

2023-08-30

Homework 1 Part 1

Question 1 Read the data from the external file and calculate the sample means, standard deviations, and variances at each time point. Make an investigator worthy table of the data.

Table 1: Cortisol Measurements Between Groups

	Time 1	Time 2	Time 3	Time 4	Time 5	Time 6
Mean						
All	3.5107	7.4804	13.5898	9.7086	7.9149	5.2625
Group C	3.4223	6.8739	12.8086	8.9811	7.5864	5.0312
Group P	3.5991	8.0869	14.3710	10.4360	8.2433	5.4938
Standard Deviation						
All	3.7213	3.7820	4.2908	2.4739	2.5079	3.3654
Group C	4.8038	3.1057	3.3092	2.2786	2.0284	4.1036
Group P	2.2710	4.3323	5.0338	2.4885	2.9141	2.4806
Variance						
All	13.8483	14.3036	18.4113	6.1202	6.2898	11.3256
Group C	23.0766	9.6454	10.9506	5.1921	4.1146	16.8394
Group P	5.1576	18.7689	25.3392	6.1925	8.4921	6.1536

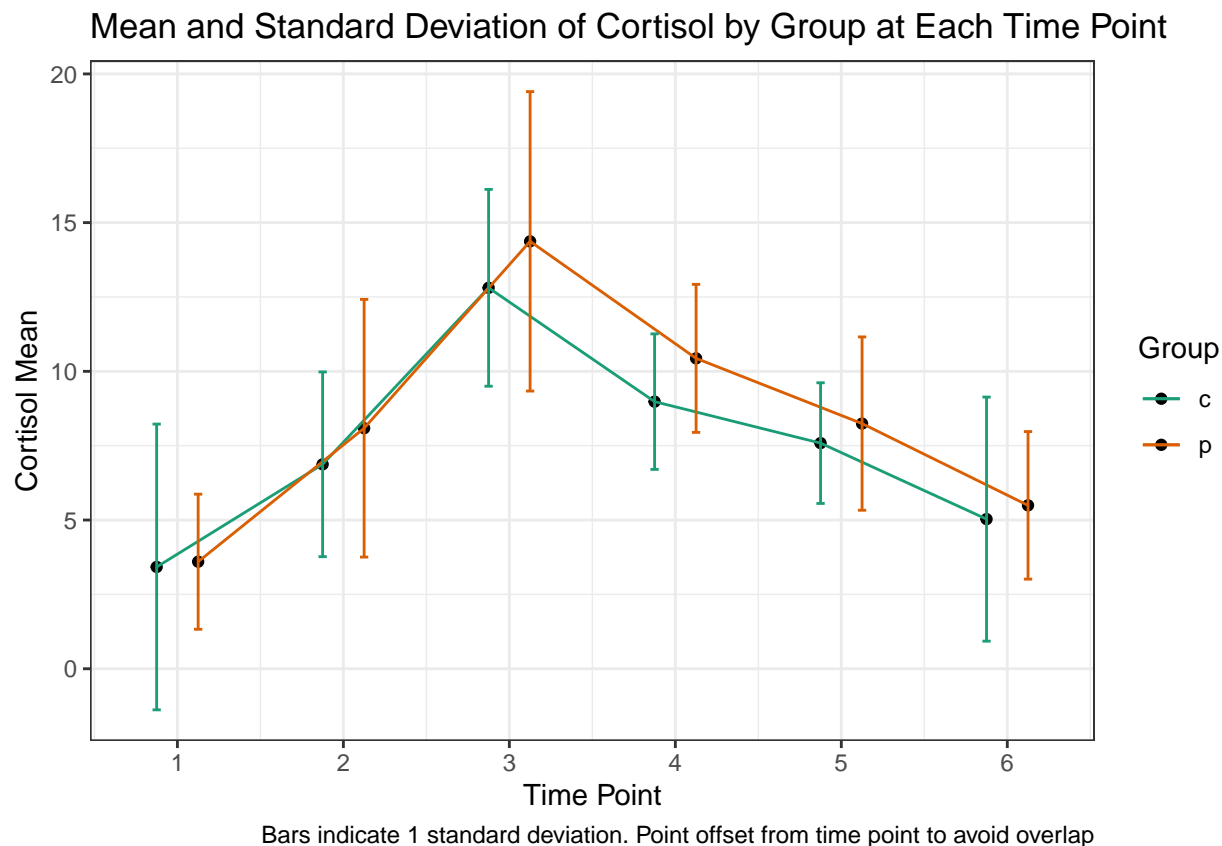
Question 2 Interpret the patterns observed in the means and variances for each group in a few sentences.

For group c the mean cortisol measurement increases until time block 3, after which it starts to steadily decrease. This group also has the lowest standard deviation and variance at time block 5. The group trend for standard deviation and variance is that there is a decrease between time block 1 and 2, a slight increase at time block 3, and then a steady decrease between time block 4 and 5, then an increase at time block 6.

For group p the mean cortisol measurement increases until time block 3, after which it starts to steadily decrease. This group also has the lowest standard deviation and variance at time block 1. For this group there is a steady increase in standard deviation and variance from time block 1 to time block 3. There is then a steep decrease at time 4, and then an alternating pattern of increase and decrease in standard deviation and variance between time blocks 4 and 6.

Group c always has a lower mean at each time point compared to group p. Group c has higher standard deviation and variance compared to group p at time block 1 and 6, and lower standard deviation and variance at for all other time blocks.

Question 3 Construct a means and standard deviation graph of the time points with the means connected by a line of a difference color for each group. Interpret the time patterns in the graph. What are your initial thoughts on whether the hypothesis that there are differences in the circadian rhythms between groups will be accepted by your formal analysis?



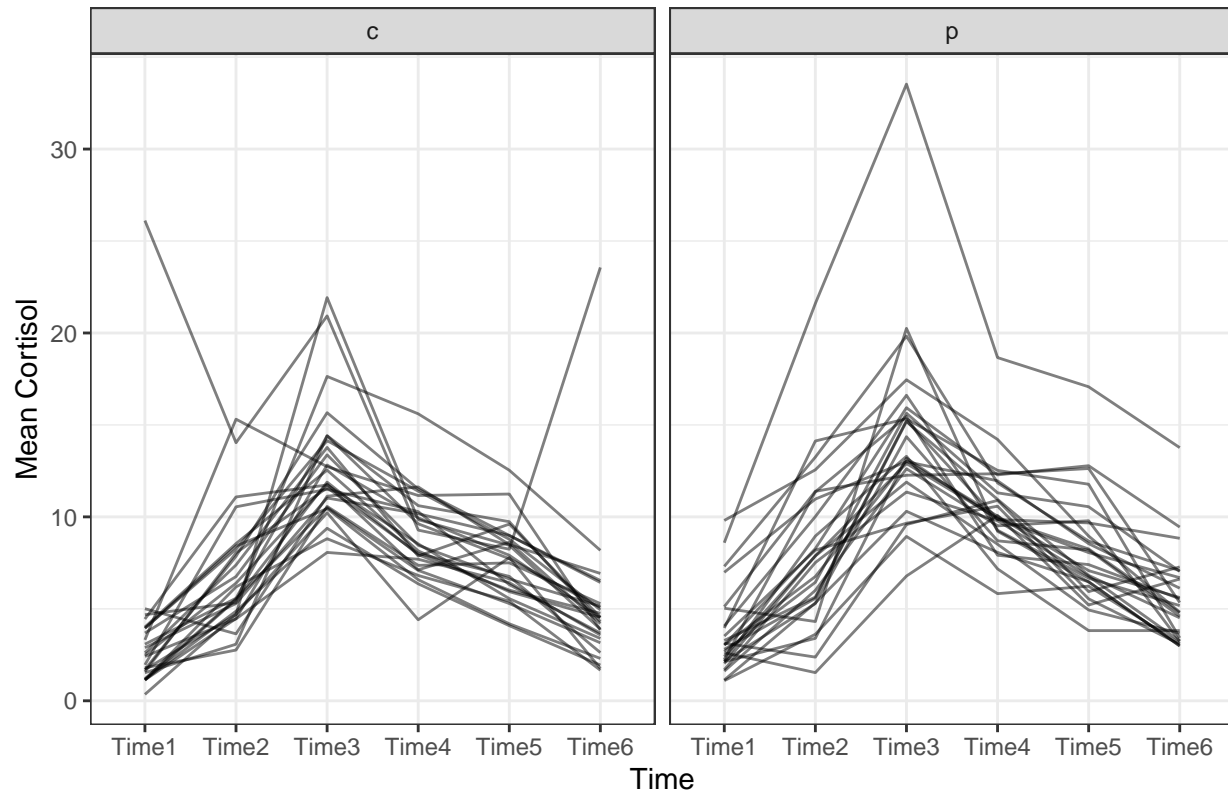
Both groups follow the same general trend, where mean cortisol is lowest at time point 1 (22:00-01:50), and steadily increases until time point 3 (06:00-09:50). Both groups steadily decrease in mean cortisol until the final time point (time point 6, 18:00-21:50). This trend follows the circadian rhythm, where cortisol peaks in the middle of the day, and decreases during night.

Between groups the mean cortisol levels are nearly the same for time points 1 and 6. For all other time points in the measurement period group “c” has lower mean cortisol levels compared to group p. The standard deviation for each group varies between time points. For time points 1 and 6 group “c” has a much larger standard deviation compared to group “p”. For time points 2, 3, and 5 group “p” has a larger standard deviation. Time point 4 has similar standard deviations between groups.

Mean cortisol levels in general, with the exception of the time point 1 and 6, are lower for group “c” compared to group “p”. While this initially seems to suggest that there is a difference in cortisol between women who suffer from depression and those who do not, the standard deviation vary between groups at each time point enough to possibly be obscuring what is happening when only looking at means cortisol levels for each time block. Possible explanations for this behavior could be additional variables which may be affecting cortisol levels beyond a diagnosis of depression.

Question 4 Construct a spaghetti plot of each group (one panel for each group). Describe the variation in cortisol levels within a person. Describe the variation in the cortisol levels between people. Which source of variation has more variation? Justify with 1-2 sentences.

Spaghetti Plots of Mean Cortisol at Each Time Point by Group

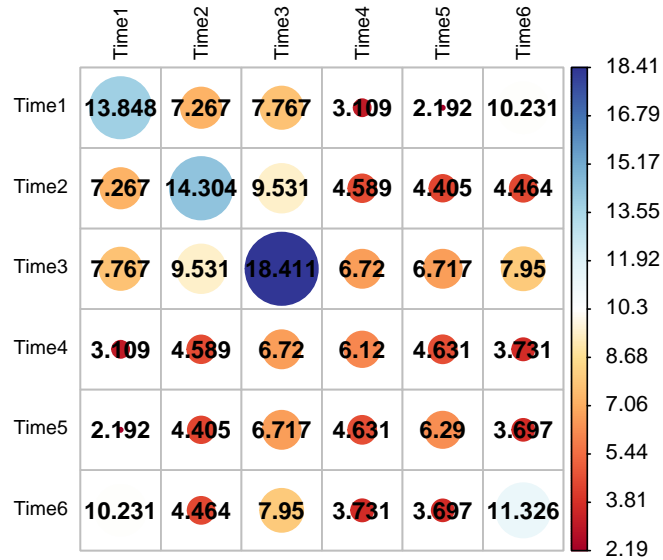


In general the pattern of variation within individuals is the same for both groups, with the middle portion of the day having the highest mean cortisol levels for most individuals, and night/early morning having the lowest. The variation between individuals also changes depending on the time point. For group “c” the general pattern is that the first and last measurements have the least amount of variation between individuals, and for group “p” the variation between individuals is generally lowest at times 1 and 6, with spreading and coalescing of values throughout the measurement period. Additionally, there are a few outliers who may be strongly influencing variance measurements at each time point.

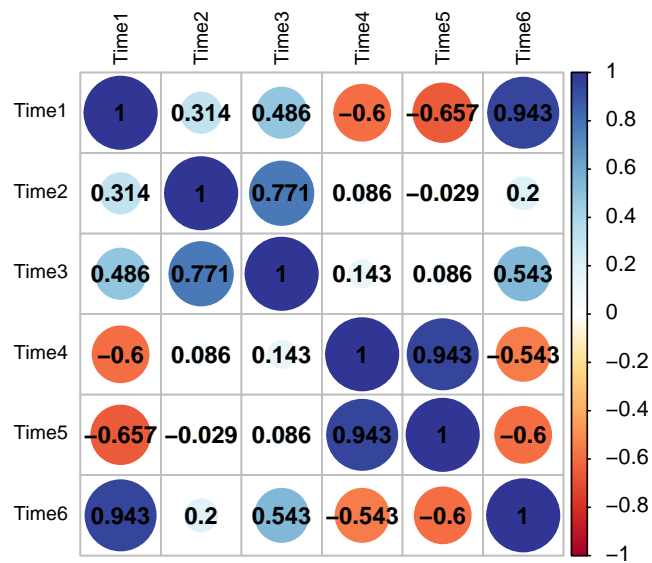
In general there seems to be more variation between subjects compared to within an individual. This is somewhat obscured by the fact that the individuals tend to follow the same trend in mean cortisol levels throughout the day, but overall there are wide variances between individuals during the middle portion of the experimental period.

Question 5 Calculate the 6x6 covariance and correlation matrices for the six time points of cortisol levels.

Covariance Between Cortisol and Time Point



Correlation Between Cortisol and Time Point



Spearman correlation was used due to significant outliers.

Question 6 The covariance matrix is symmetric. Explain why in 1-2 sentences with a mathematical justification.

The covariance matrix is symmetric because the transposition of this matrix is the same as the original matrix. In mathematical terms, using “A” to refer to the covariance matrix:

$$A = A^T$$

Question 7 Verify that the diagonal elements of the covariance matrix are the variances of each time point by comparing your variance-covariance matrix to the table created in problem 1.

The diagonal elements of the covariance matrix (Covariance Between Cortisol and Time Point) are the variances of each time point in the table (Table 1: Cortisol Measurements Between Groups).

Homework 1 Part 2

Question 1 For the following matrices and vectors write down the dimensions in rxc notation (rows x columns). ' and T both mean transpose for this problem.

Part A

(3x1) matrix.

Part B

$$(2, 4, -2)'$$

Transposing this vector would result in a vector with (3x1) dimensions.

Part C

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}$$

(3x3) matrix.

Part D

$$\begin{bmatrix} 2 & 1 \\ 4 & -6 \\ -2 & 7 \end{bmatrix}$$

(3x2) matrix.

Part E

$$\begin{bmatrix} 2 & 1 \\ 4 & -6 \\ -2 & 7 \end{bmatrix}^T = \begin{bmatrix} 2 & 4 & -2 \\ 1 & -6 & 7 \end{bmatrix}$$

After transposing the resulting matrix is (2x3).

Question 2 Can we multiply these matrices? If no, explain how you figured out the computation was not possible. If yes, what will be the dimensions in rxc notation.

Part A

$$\begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix} \begin{bmatrix} 1 & 4 & 10 \end{bmatrix}$$

Multiplying matrices with dimensions of (3×1) and (1×3) is possible because the number of columns in the first matrix is the same as the number of rows in the second matrix. The resulting matrix is (3×3)

Part B

$$\begin{bmatrix} 1 & 4 & 10 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix}$$

Multiplying matrices with dimensions orientation of (1×3) and (3×1) is possible because the number of columns in the first matrix is the same as the number of rows in the second matrix. The resulting matrix is (1×1)

Part C

$$\begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 4 & -6 \\ -2 & 7 \end{bmatrix}$$

Multiplying matrices with the dimensions orientation of (3×3) and (3×2) is possible because the number of columns in the first matrix is the same as the number of rows in the second matrix. The resulting matrix is (3×2) .

Part D

$$\begin{bmatrix} 2 & 1 \\ 4 & -6 \\ -2 & 7 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}$$

Multiplying matrices with the dimensions orientation of (3×2) and (3×3) is not possible because the number of columns in the first matrix is not the same as the number of rows.

Part E

$$\begin{bmatrix} 2 & 1 \\ 4 & -6 \\ -2 & 7 \end{bmatrix}^T \begin{bmatrix} 2 & 1 & 1 \\ 4 & -6 & 0 \\ -2 & 7 & 2 \end{bmatrix}$$

Multiplying matrices with the dimensions orientation of $(3 \times 2)^T = (2 \times 3)$ and (3×3) is possible because the number of columns in the first matrix is the same as the second matrix. The resulting matrix is (2×3) .

Question 3 Let X be an $n \times p$ matrix of covariate data. Can I invert X ? Explain.

The covariance matrix is always square, but also need to be either positive definite or positive semi-definite in order to invert.

Question 4 Let Y be an $n \times 1$ vector. Let X be an $n \times p$ matrix of covariates, let Z be an $n \times q$ matrix of additional covariates, let β be an $p \times 1$ vector of fixed (not random) parameters, and let b be an $q \times 1$ vector. Using an analysis of the dimensions determine whether the following operations are possible or equations make sense. When no, explain the issue. When yes, provide a dimension evaluation showing the resulting dimensions of the dimensions match on both sides of an equation.

Part A:

$$Y = X\beta$$

- Dimensions of Y : ($n \times 1$)
- Dimensions of X : ($n \times p$)
- Dimensions of β : ($p \times 1$)

The operation on the right hand side of the equivalence is possible because the number of columns of X are the same as the number of rows of β , and this operation results in a matrix with ($n \times 1$) dimensions, the same as vector Y . The operation makes sense because multiplying the matrix of covariate values (X) for the subjects by the vector of coefficients (β) returns the outcome values for the subjects (Y).

Part B

$$Y = Zb$$

- Dimensions of Y : ($n \times 1$)
- Dimensions of Z : ($n \times q$)
- Dimensions of b : ($q \times 1$)

The operation on the right hand side of the equivalence is possible because the number of columns of Z are the same as the number of rows of b , and this operation results in a matrix with ($n \times 1$) dimensions, the same as vector Y . The operation makes sense because multiplying the matrix for subject-specific/random effects (Z) by the vector of random effects (b) returns the outcome values for the subjects (Y). This is a model with no fixed effects.

Part C

$$Z^{-1}Y = b$$

- Dimensions of Z^{-1} : ($n \times q$)
- Dimensions of Y : ($n \times 1$)
- Dimensions of b : ($q \times 1$)

Inverting Z is only possible if Z is a square matrix (and even then it may not be possible). Additionally, multiplying Z^{-1} by Y is not possible because the number rows of Z^{-1} is not the same as the number of columns in Y . The equation is attempting to return the vector of random effects from the subject outcome (Y) and the matrix for subject-specific/random effects (Z).

Part D

$$X^T Y = X^T X \beta$$

- Dimensions of X^T : (pxn)
- Dimensions of Y : (nx1)
- Dimensions of X : (npx)
- Dimensions of β : (px1)

The operations on both side of the equivalence are possible because the number of columns in X^T is the same as the number of rows in Y , and the number of columns in X^T are the same as the number of rows in X and the number of columns in X is the same as the number of rows in β . Matrix multiplication is associative so order of these multiplications does not matter. The equivalence is possible because $X\beta$ is equivalent to Y . The resulting matrices of both sides of the equivalence have dimensions of (px1).

Question 5 Let Y be a nx1 vector of outcomes. Let X be an nxp matrix of covariate data. Let β be a px1 vector of regression coefficients. Let E be an nx1 vector of residuals. Explain why the two equations below are not equivalent ways of writing a linear regression in matrix notation and which is mathematically correct.

$$Y = X\beta + E \text{ vs } Y = \beta X + E$$

- Dimensions of Y : (nx1)
- Dimensions of X : (npx)
- Dimensions of β : (px1)
- Dimensions of E : (nx1)

These two equations are not equivalent because order matters when multiplying two matrices together. The equation on the left is possible because the number of columns in X are the same as the number of rows in β . In contrast, the equation on the left is not possible because the number of columns in β is not the same as the number of rows in X . In other words the equation on the right is the correct way of writing a linear regression in matrix notation.

$$Y = X\beta + E$$

Question 6 Can the following matrix be a variance-covariance matrix? Why or why not?

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 5 & -6 \\ 3 & -6 & -1 \end{bmatrix}$$

There are several requirements which a matrix must have in order to possible be a variance-covariance matrix.

- Square matrix: This matrix is square.
- Symmetric matrix: This matrix is symmetric.
- Variances a positive: This is violated, the lower right diagonal has a negative value.

Because there is a negative variance this could not be a variance-covariance matrix.

Question 7 Assume we have a regression model $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + E_i$ and we observe the following data:

i	1	2	3	4	5	6
X1i	7	4	16	3	21	8
X2i	33	41	7	49	5	31
Yi	42	33	75	28	91	55

Using matrix methods (which you can program in R-bust must show the matrix calculations in R), obtain (a) $\hat{\beta} = (X^T X)^{-1} X^T Y$, the least squares estimate; (b) estimate the hat matrix $H = X(X^T X)^{-1} X^T$; (c) estimate \hat{Y} ; (d) residual vector \hat{E}_i .

All R code used for calculating the following are found in the Appendix at the end of this file.

Part A $\hat{\beta} = (X^T X)^{-1} X^T Y$, the least squares estimate.

Solved in one step by R (code found in appendix), but one way steps can be broken down as follows:

- $(X^T X)^{-1}$: Solved by inverting X then multiplying this by X. Because the number of columns of X^T is the same as the number of rows of X this operation can occur. This results in a square matrix with dimensions (2x2) that can be inverted (not all square matrices can be inverted). After inverting the subsequent matrix is (2x2).
- $(X^T X)^{-1} X^T$: The expression solved above had dimensions of (2x2), while X^T has dimensions of (2x6). Because the number of columns of $(X^T X)^{-1}$ is equal to the number of rows of X^T this operation can occur, and returns a matrix with dimensions of (2x6).
- $(X^T X)^{-1} X^T Y$: The expression solved above has dimensions of (2x6) and Y has dimensions of (6x1). Because the number of columns of $(X^T X)^{-1} X^T$ is the same as the number of rows in Y this operation can occur, and returns a matrix with dimensions of (2x1).

Because matrix multiplication is associative the order these steps were performed does not matter. The total operation returns the following (2x1) matrix:

$$LSE = \begin{bmatrix} 4.4048 \\ 0.3803 \end{bmatrix}$$

Part B estimate the hat matrix $H = X(X^T X)^{-1} X^T$

Solved in one step by R (code found in appendix), but one way steps can be broken down as follows:

- $(X^T X)^{-1}$: Same as in the above part, solved by inverting X then multiplying this by X. Because the number of columns of X^T is the same as the number of rows of X this can be down. This results in a square matrix with dimensions (2x2) that can be inverted (not all square matrices can be inverted). After inverting the subsequent matrix is (2x2).
- $X(X^T X)^{-1}$: X has dimensions of (6x2) and the expression above has dimensions of (2x2). Because the number of columns of X is the same as the number of rows of $(X^T X)^{-1}$ this operation can occur. The resulting matrix has dimensions of (6x2).
- $X(X^T X)^{-1} X^T$: The expression solved above has dimensions of (6x2) and X^T has dimensions of (2x6). Because the number of columns of $X(X^T X)^{-1}$ is the same as the number of rows of X^T this operation can occur. The resulting matrix has dimensions of (6x6).

Because matrix multiplication is associative the order these steps were performed does not matter. The total operation returns the following (6x6) matrix:

$$H = \begin{bmatrix} 0.1795 & 0.2115 & 0.0736 & 0.2484 & 0.076 & 0.1721 \\ 0.2115 & 0.2813 & -0.0125 & 0.3433 & -0.0467 & 0.1931 \\ 0.0736 & -0.0125 & 0.3369 & -0.0543 & 0.4525 & 0.1007 \\ 0.2484 & 0.3433 & -0.0543 & 0.4234 & -0.1093 & 0.2229 \\ 0.076 & -0.0467 & 0.4525 & -0.1093 & 0.6109 & 0.1143 \\ 0.1721 & 0.1931 & 0.1007 & 0.2229 & 0.1143 & 0.168 \end{bmatrix}$$

Part C estimate \hat{Y}

Solved in one step by R, code found in appendix. \hat{Y} is found by multiplying X, the subjects covariate matrix, by $\hat{\beta}$, the vector of coefficients. This returns the estimated outcomes for each subject, \hat{Y} .

$$\hat{Y} = X\hat{\beta}$$

Where X has dimensions of (6x2) and $\hat{\beta}$ has dimensions of (2x1). Because the number of columns in X is the same as the number of rows in $\hat{\beta}$ this operation can occur, resulting in the (6x1) matrix shown below.

$$\begin{bmatrix} 43.3847 \\ 33.213 \\ 73.139 \\ 31.8509 \\ 94.4023 \\ 47.0288 \end{bmatrix}$$

Part D residual vector \hat{E}_i

The residual vector \hat{E}_i is found by subtracting the actual outcomes by the predicted outcomes:

$$E_i = \hat{Y}_i - Y_i$$

\hat{Y}_i has dimensions of (6x1) and Y has dimensions of (6x1). Because they have matching dimensions the above operation can occur, which results in a (6x1) matrix, shown below.

$$E_i = \begin{bmatrix} 1.3847 \\ 0.213 \\ -1.861 \\ 3.8509 \\ 3.4023 \\ -7.9712 \end{bmatrix}$$