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

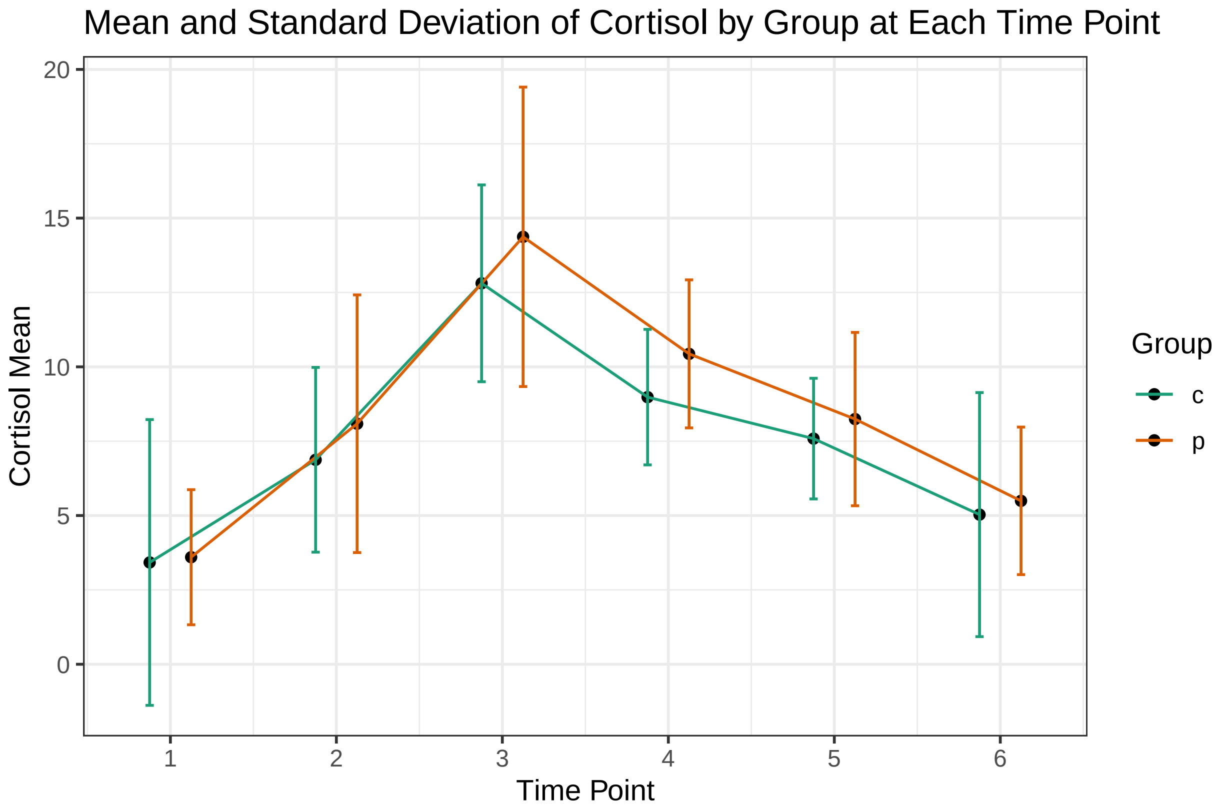
Question 2 Interpret the patterns observed in the means and variances for each group in a few sentences.

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 the 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?



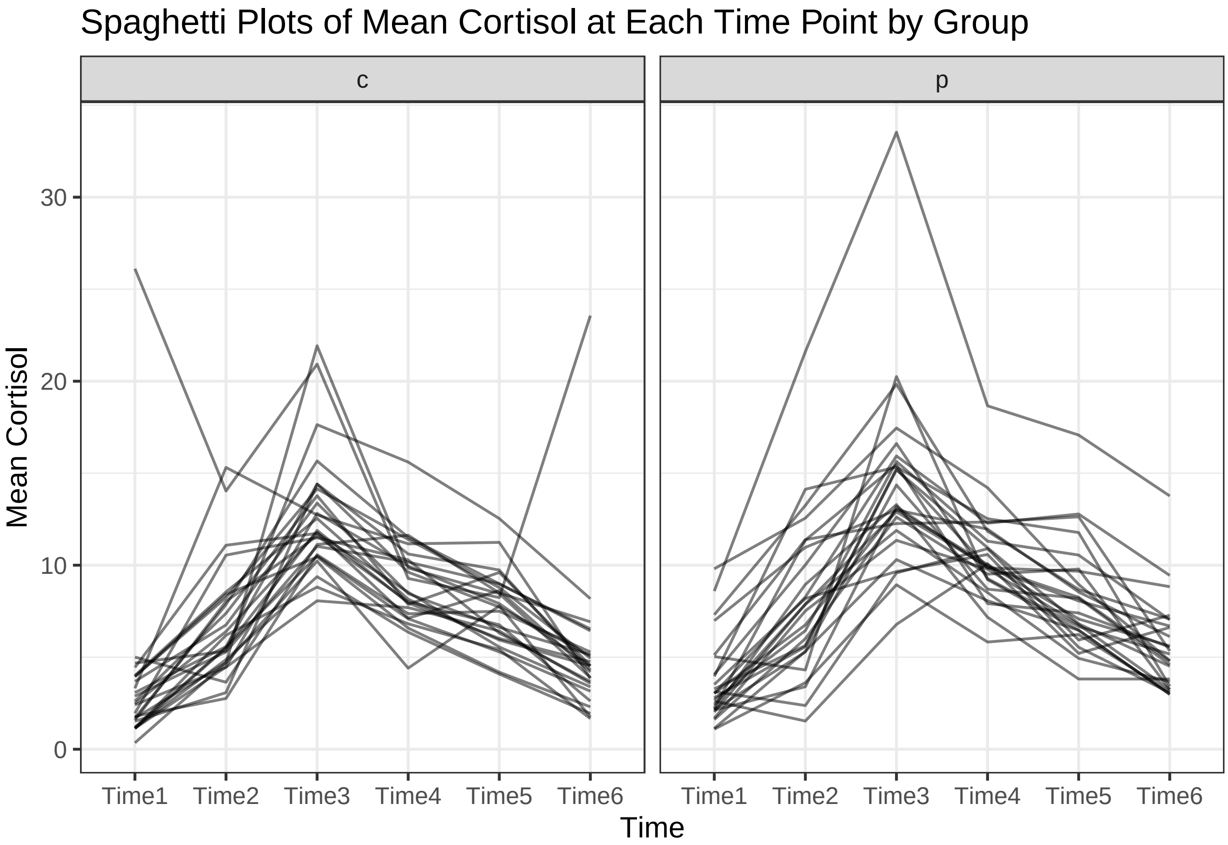
Bars indicate 1 standard deviation. Point offset from time point to avoid overlap

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 possible 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.

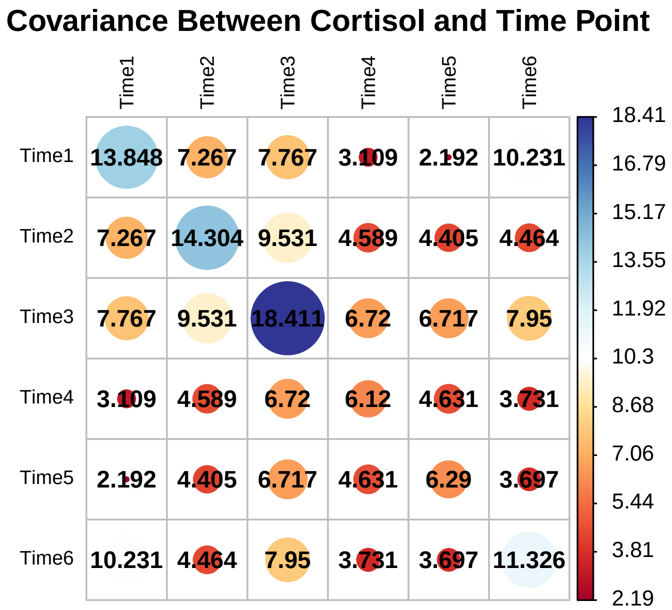
Question 4 Construct a spaghetti plot of each group (one panel for each group). Describe the variation in cortisol levels within a person. Describe the variation in the cortisol levels between people. Which source of variation has more variation? Justify with 1-2 sentences.



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.

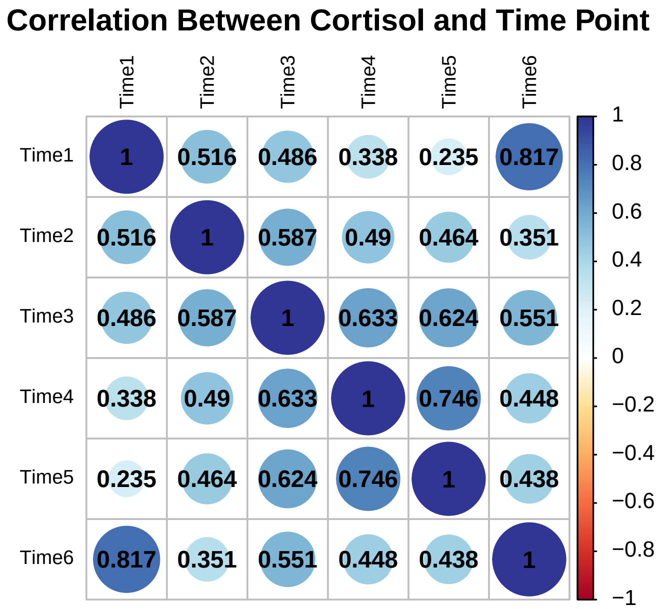
In general there seems to be more variation within an individual versus between individuals. Most individuals tend to follow the same general pattern in mean cortisol levels throughout the day. Many individual subjects experience large changes in mean cortisol level between night and day.

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



Covariance Between Cortisol and Time Point

	Time1	Time2	Time3	Time4	Time5	Time6
Time1	13.848	7.267	7.767	3.109	2.192	10.231
Time2	7.267	14.304	9.531	4.589	4.405	4.464
Time3	7.767	9.531	18.411	6.72	6.717	7.95
Time4	3.109	4.589	6.72	6.12	4.631	3.731
Time5	2.192	4.405	6.717	4.631	6.29	3.697
Time6	10.231	4.464	7.95	3.731	3.697	11.326



Correlation Between Cortisol and Time Point

	Time1	Time2	Time3	Time4	Time5	Time6
Time1	1	0.516	0.486	0.338	0.235	0.817
Time2	0.516	1	0.587	0.49	0.464	0.351
Time3	0.486	0.587	1	0.633	0.624	0.551
Time4	0.338	0.49	0.633	1	0.746	0.448
Time5	0.235	0.464	0.624	0.746	1	0.438
Time6	0.817	0.351	0.551	0.448	0.438	1

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:

$$A = A^T$$

Another way of proving this is showing that $Cov(X,Y) = Cov(Y,X)$, which would hold for all time block covariance relationships.

$$\begin{aligned} Cov(X,Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - XE(Y) - YE(X) + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) = Cov(Y,X) \end{aligned}$$

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).

Question 8 Interpret cell(3,4) and cell(3,2) in the covariance matrix.

- Cell(3,4) of the covariance matrix is the covariance between time block 3 (Time3) and time block 4 (Time4). This covariance is 6.72, meaning that time block 3 and time block 4 tend to move together. In other words this means higher mean cortisol measurements at time 3 tend have higher mean cortisol measurements at time 4.
- Cell(3,2) of the covariance matrix is the covariance between time block 3 (Time3) and time block 2 (Time2). This covariance is 9.531, meaning that time block 3 and 2 tend to move together. In other words this means higher mean cortisol measurements at time 3 tend to have higher mean cortisol measurement at time 2.

Question 9 Interpret cell(3,4) and cell(3,2) in the correlation matrix.

- Cell(3,4) of the correlation matrix is the covariance between time block 3 (Time3) and time block 4 (Time4). This correlation is 0.633, meaning that there is a positive correlation between time block 3 and 4. In other words higher mean cortisol measurements at time 3 tend to have higher mean cortisol measurements at time 4. The correlation of 0.633 is considered moderate-strong.
- Cell(3,2) of the correlation matrix is the correlation between time block 3 (Time3) and time block 2 (Time2). This correlation is 0.587, meaning that there is a positive correlation between time block 3 and 2. In other words higher mean cortisol measurements at time 3 tend to have higher mean cortisol measurements at time 2. The correlation of 0.587 is considered moderate-strong.

Question 10

Derive mathematically the correlations between Time1 and Time 2 using the covariance matrix. In other words, show how the covariance matrix can be used to derive the correlation matrix.

The equation for correlation is as follows:

$$Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Where $Cov(x,y)$ comes from the covariance table at the intersection of the time points, and σ_x and $sigma_y$ are the standard deviation of the two time points. These are found by finding the desired intersection along the diagonal of the covariance matrix, and then taking the square root of the resulting variance. Finding the correlations between Time1 and Time2 using this method:

$$Corr(Time1, Time2) = \frac{Cov(Time1, Time2)}{\sigma_{Time1}\sigma_{Time2}} = \frac{7.267}{\sqrt{13.848}\sqrt{14.304}} = \frac{7.267}{3.721 \times 3.782} = 0.516$$

The resulting correlation between Time1 and Time2 is 0.516. The same method can be used to find the rest of the correlations.

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

$$\begin{bmatrix} 2 \\ 4 \\ -2 \end{bmatrix}$$

(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 multiple 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 nxp 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. So you may be able to invert

Question 4 Let Y be an nx1 vector. Let X be an nxp matrix of covariates, let Z be an nxq matrix of additional covariates, let β be an px1 vector of fixed (not random) parameters, and let b be an qx1 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: (nx1)
- Dimensions of X: (nxp)
- Dimensions of β: (px1)

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 (nx1) 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: (nx1)
- Dimensions of Z: (nxq)
- Dimensions of b: (qx1)

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 (nx1) 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}: (nxq)
- Dimensions of Y: (nx1)
- Dimensions of b: (qx1)

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

$$X^TY = X^TX\beta$$

- Dimensions of X^T: (pxn)
- Dimensions of Y: (nx1)
- Dimensions of X: (nxp)
- 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 Xβ 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.

$$Y = X\beta + E \;\; vs \;\; Y = \beta X + E$$

- Dimensions of Y: (nx1)
- Dimensions of X: (nxp)
- 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 name as the number of rows in X. In other words the equation on the right is the correct way of writing a linear regresssion 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 transposing X, (3,6), then multiplying this by X, (6,3). 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 (3x3) that can be inverted (not all square matrices can be inverted). After inverting the subsequent matrix is (3x3) as well.
- $(X^T X)^{-1} X^T$: The expression solved above had dimensions of (3x3), while X^T has dimensions of (3x6). 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 (3x6).
- $(X^T X)^{-1} X^T Y$: The expression solved above has dimensions of (3x6) 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 (3x1).

Because matrix multiplication is associative the order these steps were performed does not matter. The total operation returns the following (3x1) matrix:

$$LSE = \begin{bmatrix} 33.9321 \\ 2.7848 \\ -0.2644 \end{bmatrix}$$

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 (3x3) that can be inverted (not all square matrices can be inverted). After inverting the subsequent matrix is (3x3).
- $X(X^T X)^{-1}$: X has dimensions of (6x3) and the expression above has dimensions of (3x3). 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 (6x3).
- $X(X^T X)^{-1} X^T$: The expression solved above has dimensions of (6x3) and X^T has dimensions of (3x6). 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:

$$H = \begin{bmatrix} 0.2314 & 0.2517 & 0.2118 & 0.1489 & -0.0548 & 0.211 \\ 0.2517 & 0.3124 & 0.0944 & 0.2663 & -0.1479 & 0.2231 \\ 0.2118 & 0.0944 & 0.7044 & -0.3192 & 0.1045 & 0.2041 \\ 0.1489 & 0.2663 & -0.3192 & 0.6143 & 0.1414 & 0.1483 \\ -0.0548 & -0.1479 & 0.1045 & 0.1414 & 0.9404 & 0.0163 \\ 0.211 & 0.2231 & 0.2041 & 0.1483 & 0.0163 & 0.1971 \end{bmatrix}$$

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 estimated coefficients. This returns the estimated outcomes for each subject, $\hat{Y}$.

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

Where X has dimensions of (6x3) and $\hat{\beta}$ has dimensions of (3x1). 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.

$$\begin{bmatrix} 44.6996 \\ 34.23 \\ 76.6374 \\ 29.3299 \\ 91.09 \\ 48.0132 \end{bmatrix}$$

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.

$$E_i = \begin{bmatrix} 2.6996 \\ 1.23 \\ 1.6374 \\ 1.3299 \\ 0.09 \\ -6.9868 \end{bmatrix}$$

Question 8 Assume we have a regression model $Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + E_i, i = 1...n$. Use matrix notation to show how to standardize the outcome (Y_i') and show how to standardize the covariate matrix (X'). We will now have the following regression $Y_i' = \beta_0' + \beta_1' X_{1,i}' + \beta_2' X_{2,i}' + E_i'$. Derive the relationship between $\hat{\beta}$ and $\hat{\beta}'$.

In order to standardize Y_i we can use the following equation:

$$Y_i' = \frac{Y_i - \bar{Y}}{S_Y}$$

Where the equation for S_Y is:

$$S_Y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

Because $\bar{Y}$ is the mean of every subject it is a scalar. In order to do this operation using matrix methods we need to convert this scalar into a vector of $\bar{Y}$ values.

$$\bar{Y}_n = \bar{Y}1_n, \; where \; 1_n = \begin{bmatrix} 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}_n$$

Matrix subtraction can then be used to center the outcome around 0.

$$Y_{centered} = Y - \bar{Y}1_n$$

Where this operation is possible because Y has (nx1) dimensions, and $\bar{Y}1_n$ has (nx1) dimensions. The resulting Y' matrix has (nx1) dimensions.

After this we can divide by the standard deviation of Y to standardize the outcome:

$$Y' = \frac{Y - \bar{Y}1_n}{S_Y}$$

Where this operation is possible because S_Y is a scalar.

Standardizing the covariate matrix is similar, but needs an additional matrix for the denominator. For the mean estimates of the coefficients we have a vector with (2x1) dimensions, which can be transposed to get a vector with (1x2) dimensions.

$$\bar{X}^T = \begin{bmatrix} \bar{X}_1, & \bar{X}_2 \end{bmatrix}$$

First we need to turn this into an (nx2) matrix for the, which can be done by multiplying a matrix of ones with (nx1) dimensions with the $\bar{X}$ vector with (1x2) dimensions.

$$1_n\bar{X}^T_{2\times 1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_n \begin{bmatrix} \bar{X}_1, & \bar{X}_2 \end{bmatrix} = \begin{bmatrix} \bar{X}_1 & \bar{X}_2 \\ \bar{X}_1 & \bar{X}_2 \\ \vdots & \vdots \\ \bar{X}_1 & \bar{X}_2 \end{bmatrix}$$

We can then use the above operation to start the standardization of the covariate matrix, X.

$$X - 1_n\bar{X}^T_{2\times 1} = \begin{bmatrix} X_{1,1} & X_{2,1} \\ X_{1,2} & X_{2,2} \\ \vdots & \vdots \\ X_{1,n} & X_{2,n} \end{bmatrix} - \begin{bmatrix} \bar{X}_1 & \bar{X}_2 \\ \bar{X}_1 & \bar{X}_2 \\ \vdots & \vdots \\ \bar{X}_1 & \bar{X}_2 \end{bmatrix}$$

Where the resulting matrix has (nx2) dimensions.

The standard deviation of each element can be found with the following equation:

$$S_{X,j} = \sqrt{\frac{\sum(x_{i,j} - \bar{X}_j)^2}{n-1}}$$

From here we can get a vector of the standard deviation for the covariates, which will have dimensions of (nx2) after transposing.

$$S_X^T = \begin{bmatrix} S_{X,1}, & S_{X,2} \end{bmatrix}$$

Because we need to use these value to divide the centered covariates the easiest way to do this would be to make another vector with the inverted standard deviations. Note, cannot invert this matrix because it is not square.

$$S_{X,inv}^T = \begin{bmatrix} \frac{1}{S_{X,1}}, & \frac{1}{S_{X,2}} \end{bmatrix}$$

We then need to convert this into a matrix with (nx2) dimensions in order to multiply this by the centered covariates. In order to do this we can use a matrix of ones which has dimensions (nx1).

$$1_p S_{X,inv}^T = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} \frac{1}{S_{X,1}}, & \frac{1}{S_{X,2}} \end{bmatrix} = \begin{bmatrix} \frac{1}{S_{X,1}}, & \frac{1}{S_{X,2}} \\ \frac{1}{S_{X,1}}, & \frac{1}{S_{X,2}} \end{bmatrix}$$

Where the resulting matrix has (2x2) dimensions.

Putting these together will standardize the covariate matrix:

$$X' = (X - 1_n \bar{X}_{2\times 1}^T) 1_n S_{X,inv}^T$$

Where this operation works because the first matrix on the right hand side of the equivalence has (nx2) dimensions, and the second has (2x2) dimensions, resulting in a matrix with (nx2) dimensions.

Putting all of this together results in the following:

$$\frac{Y_i - \bar{Y}}{S_Y} = \beta_0' + \beta_1' \frac{X_{1,i} - \bar{X}_1}{S_{X,1}} + \beta_2 \frac{X_{2,i} - \bar{X}_2}{S_{X,2}} + E_i$$

Which can be simplified as:

$$Y_i' = \beta_0' + \beta_1' X_{1,i}' + \beta_2' X_{2,i}' + E_i'$$

Note that β_0' is 0 because the covariate matrix was centered.

β_j coefficients can be found with the following equation:

$$\beta_j = \frac{Cov(X_j, Y)}{Var(X_j)}$$

Expanding out the numerator:

$$\begin{aligned} Cov(X_j, Y) \frac{S_{X,j} S_y}{S_{X,j} S_y} &= E[(Y - \bar{Y})(X_j - \bar{X}_j)] \frac{S_{X,j} S_y}{S_{X,j} S_y} \\ &= E\left[\frac{(Y - \bar{Y})(X_j - \bar{X}_j)}{S_{X,j} S_y} \right] S_{X,j} S_y \\ &= Cov(X_j', Y') S_{X,j} S_y \end{aligned}$$

Expanding out the denominator:

$$\begin{aligned} Var(X_j) \frac{S_{x,j}^2}{S_{x,j}^2} &= E[(X_j - \bar{X}_j)^2] \frac{S_{x,j}^2}{S_{x,j}^2} \\ &= E\left[\frac{(X_j - \bar{X}_j)^2}{S_{x,j}^2} \right] S_{x,j}^2 \\ &= Var(X_j') S_{x,j}^2 \end{aligned}$$

Putting these together gives the following relationship:

$$\begin{aligned} \beta_j &= \frac{Cov(X_j', Y') S_{X,j} S_Y}{Var(X_j') S_{X,j}^2} \\ &= \beta_j' \frac{S_Y}{S_{X,j}} \end{aligned}$$

Code Appendix

```
# Loading necessary libraries into R
library(dplyr)
library(tidyr)
library(ggplot2)
library(kableExtra)
library(RColorBrewer)
library(corrplot)
library(Hmisc)
library(knitr)

opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

# Laptop Data: 'C:/Biostatistics Masters Program/Fall
# 2023/BIOS 6643 - Analysis of Longitudinal
# Data/BIOS6643_ALD/Homework 1/DataRaw/cortdata.csv'

# Desktop Data: 'C:/Users/domin/Documents/Biostatistics
# Masters Program/Fall 2023/BIOS 6643 - Analysis of
# Longitudinal Data/BIOS6643_ALD/Homework
# 1/DataRaw/cortdata.csv'

# Loading data set (cortdata.csv) into R.
cort_data <- read.csv("C:/Users/domin/Documents/Biostatistics Masters Program/Fall 2023/BIOS 6643 - Anal


# START PART 1 QUESTION 1 CODE

time_block_metrics <- function(data_file, time_block) {
    # data_file: Refers to the data set.  time_point:
    # Refers to the specific time blocks (character
    # string).
    mean <- mean(data_file[, time_block])
    sd <- sd(data_file[, time_block])
    var <- var(data_file[, time_block])
    metrics <- data.frame(cbind(mean, sd, var))
    colnames(metrics) <- c("mean", "sd", "var")
    return(metrics)
}

# Selecting out subjects who are grouped as time 'c'.
cort_data_c <- cort_data[cort_data$casecontrol == "c", ]
cort_data_p <- cort_data[cort_data$casecontrol == "p", ]

# Initializing time block column names to input into above
# function.
time_blocks <- c("Time1", "Time2", "Time3", "Time4", "Time5",
    "Time6")

# Calculating mean, standard deviation, and variance for
# all subjects and by group.
all_metrics <- matrix(unlist(sapply(time_blocks, function(x) time_block_metrics(cort_data,
    x))), ncol = 6)
colnames(all_metrics) <- c("Time1", "Time2", "Time3", "Time4",
    "Time5", "Time6")
rownames(all_metrics) <- c("Mean", "SD", "Var")

c_metrics <- matrix(unlist(sapply(time_blocks, function(x) time_block_metrics(cort_data_c,
    x))), ncol = 6)
colnames(c_metrics) <- c("Time1", "Time2", "Time3", "Time4",
    "Time5", "Time6")
rownames(c_metrics) <- c("Mean", "SD", "Var")

p_metrics <- matrix(unlist(sapply(time_blocks, function(x) time_block_metrics(cort_data_p,
    x))), ncol = 6)
colnames(p_metrics) <- c("Time1", "Time2", "Time3", "Time4",
    "Time5", "Time6")
rownames(p_metrics) <- c("Mean", "SD", "Var")

# Combining outcomes together.
metrics_raw <- rbind(all_metrics, c_metrics, p_metrics)

metrics <- metrics_raw[c(1, 4, 7, 2, 5, 8, 3, 6, 9), ]

rownames(metrics) <- c("All", "Group C", "Group P", "All", "Group C",
    "Group P", "All", "Group C", "Group P")


# Making a table to illustrate the mean, standard
# deviation, and variance of each time block.
time_metric_table <- kbl(metrics, caption = "Cortisol Measurements Between Groups",
    col.names = c("Time 1", "Time 2", "Time 3", "Time 4", "Time 5",
        "Time 6"), booktabs = T, align = "cc", digits = 4, linesep = c("",
        "\\addlinespace")) %>%
    kable_styling(full_width = F, latex_options = "HOLD_position") %>%
    pack_rows("Mean", 1, 3) %>%
    pack_rows("Standard Deviation", 4, 6) %>%
    pack_rows("Variance", 7, 9)

time_metric_table

# END PART 1 QUESTION 1 CODE


# START Part 1 QUESTION 3 CODE

# Making a data frame for plotting the mean and standard
# deviation for group c.
c_group_metrics <- data.frame(t(c_metrics[1:2, ]))

# Calculating both upper and lower standard deviation for c
# group.
c_group_metrics$lower_SD <- c_group_metrics$Mean - c_group_metrics$SD
c_group_metrics$upper_SD <- c_group_metrics$Mean + c_group_metrics$SD
```

```r
# Adding in timepoints for each measurement.
c_group_metrics$time_point <- c(1, 2, 3, 4, 5, 6)

# Adding in group designation.
c_group_metrics$Group <- "c"

# Making a data frame for plotting the mean and standard
# deviation for group p.
p_group_metrics <- data.frame(t(p_metrics[1:2, ]))

# Calculating both upper and lower standard deviation for p
# group.
p_group_metrics$lower_SD <- p_group_metrics$Mean - p_group_metrics$SD
p_group_metrics$upper_SD <- p_group_metrics$Mean + p_group_metrics$SD

# Adding in time points for each measurement.
p_group_metrics$time_point <- c(1, 2, 3, 4, 5, 6)

# Adding in group designation.
p_group_metrics$Group <- "p"


# Combining both groups together into a single data frame.
groups_metric <- rbind(c_group_metrics, p_group_metrics)


# Making a plot that shows both the group means and
# standard deviation at each individual time point, with
# colored lines connecting each group.

group_metric_plot <- ggplot(groups_metric, aes(x = time_point,
    y = Mean, fill = Group)) + geom_point(position = position_dodge(width = 0.5),
    colour = "black") + geom_line(aes(color = Group), position = position_dodge(width = 0.5)) +
    geom_errorbar(aes(ymin = lower_SD, ymax = upper_SD, color = Group),
        width = 0.1, position = position_dodge(width = 0.5),
        ) + labs(title = "Mean and Standard Deviation of Cortisol by Group at Each Time Point",
    x = "Time Point", y = "Cortisol Mean", caption = "Bars indicate 1 standard deviation. Point offset
    scale_color_brewer(palette = "Dark2") + scale_x_continuous(breaks = seq(1,
    6, 1)) + theme_bw()

group_metric_plot

# END Part 1 QUESTION 3 CODE.


# START Part 1 QUESTION 4 CODE

# Transforming original data frame from wide to long
# format.
cort_data_long <- cort_data %>%
    pivot_longer(cols = "Time1":"Time6", names_to = "time", values_to = "mean_cortisol")

# Making spaghetti plots.
```

```
cort_spaghetti <- ggplot(cort_data_long, aes(x = factor(time),
    y = mean_cortisol)) + geom_line(alpha = 0.5, aes(group = SubjectId)) +
    facet_grid(. ~ casecontrol) + labs(title = "Spaghetti Plots of Mean Cortisol at Each Time Point by C
    x = "Time", y = "Mean Cortisol") + theme_bw()


cort_spaghetti

# END Part 1 QUESTION 4 CODE


# START PART 1 QUESTION 5

# calculating covariance matrix of the six time points for
# cortisol levels.  Using 'corrplot' function from the
# 'corrplot' library to plot both the covariance and
# correlation matrix.

par(mfrow = c(2, 1))

# Selecting out necessary columns for each time point to
# calculate the covariance and correlation matrices.
time_cort <- cort_data %>%
    select(Time1:Time6)

# Calculating covariance matrix and plotting.
time_cort_cov <- cov(time_cort)

time_cort_cov_plot <- corrplot(time_cort_cov, method = "circle",
    tl.col = "black", addCoef.col = "black", number.cex = 0.9,
    tl.cex = 0.75, col = COL2("RdYlBu"), is.corr = F, tl.pos = "lt",
    number.digits = 3, title = "Covariance Between Cortisol and Time Point",
    cl.align.text = "l", mar = c(0, 0, 2, 0))


# Calculating correlation matrix and plotting. spearman
# correlation was chosen due to the outliers in the data.
time_cor_corr <- cor(time_cort)


time_cort_corr_plot <- corrplot(time_cor_corr, method = "circle",
    tl.col = "black", addCoef.col = "black", number.cex = 0.9,
    tl.cex = 0.75, col = COL2("RdYlBu"), tl.pos = "lt", number.digits = 3,
    title = "Correlation Between Cortisol and Time Point", cl.align.text = "l",
    mar = c(0, 0, 2, 0))

par(mfrow = c(1, 1))

# END PART 1 QUESTION 5


# START PART 2 CODE
```

```r
# Function for writing a matrix in LaTeX from a r matrix.
# Code found at
# https://stackoverflow.com/questions/45591286/
# for-r-markdown-how-do-i-display-a-matrix-from-r-variable

write_matrix <- function(x) {
    begin <- "\\begin{bmatrix}"
    end <- "\\end{bmatrix}"
    X <- apply(x, 1, function(x) {
        paste(paste(x, collapse = "&"), "\\\\")
    })
    paste(c(begin, X, end), collapse = "")
}


# START PART 2 QUESTION 1 CODE

# Making matrices.
q1_pa_mat <- matrix(c(2, 4, -2), nrow = 3)

q1_pc_mat <- matrix(c(2, 1, 1, 4, -6, 0, -2, 7, 2), nrow = 3,
    byrow = TRUE)

q1_pd_mat <- matrix(c(2, 1, 4, -6, -2, 7), nrow = 3, byrow = TRUE)

p1_pe_mat_1 <- matrix(c(2, 1, 4, -6, -2, 7), nrow = 3, byrow = TRUE)

p1_pe_mat_2 <- matrix(c(2, 4, -2, 1, -6, 7), nrow = 2, byrow = TRUE)

# END PART 2 QUESTION 1 CODE


# START PART 2 QUESTION 2 CODE

# Making matrices.

q2_pa_mat_1 <- matrix(c(2, 4, -2), nrow = 3, byrow = TRUE)

q2_pa_mat_2 <- matrix(c(1, 4, 10), nrow = 1, byrow = TRUE)

# END PART 2 QUESTION 2 CODE


# START PART 2 QUESTION 6 CODE

q6_mat <- matrix(c(1, 2, 3, 2, 5, -6, 3, -6, -1), nrow = 3, byrow = TRUE)

# END PART 2 QUESTION 6 CODE


# START PART 2 QUESTION 7 CODE

# Initializing vectors.
```

```
i <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 1, byrow = TRUE)
x_i1 <- matrix(c(7, 4, 16, 3, 21, 8), nrow = 1, byrow = TRUE)
x_i2 <- matrix(c(33, 41, 7, 49, 5, 31), nrow = 1, byrow = TRUE)
y_i <- matrix(c(42, 33, 75, 28, 91, 55), nrow = 1, byrow = TRUE)

# Creating X matrix and Y matrix so they can be used in
# operations correctly.  Adding a column to the X matrix to
# account for the intercept.
intercept_x <- matrix(c(1, 1, 1, 1, 1, 1), nrow = 6)
x <- cbind(intercept_x, t(rbind(x_i1, x_i2)))
y <- t(y_i)

# Combining vectors and naming rows and columns.
p2_q7_df <- data.frame(rbind(x_i1, x_i2, y_i))
colnames(p2_q7_df) <- c("1", "2", "3", "4", "5", "6")
p2_q7_df$i <- c("X1i", "X2i", "Yi")
p2_q7_df <- p2_q7_df %>%
    relocate(i, .before = 1)

# Making a table showing the vectors.
p2_q7_table <- kbl(p2_q7_df, booktabs = T, align = "cc") %>%
    kable_styling(latex_options = "HOLD_position")

p2_q7_table


# START Part 2 Question 7 Part A CODE

least_square_estimates <- solve(t(x) %*% x) %*% t(x) %*% y

# END PART 2 QUESTION 7 PART A CODE


# START PART 2 QUESTION 7 PART B CODE

hat_mat <- x %*% solve((t(x) %*% x)) %*% t(x)

# END PART 2 QUESTION 7 PART B CODE


# START PART 2 QUESTION 7 PART C CODE

y_hat <- x %*% least_square_estimates

# END PART 2 QUESTION 7 PART C CODE


# START PART 2 QUESTION 7 PART D CODE

e_hat <- y_hat - y

# END PART 2 QUESTION 7 PART D CODE
```

```r
# END PART 2 QUESTION 7 CODE


# START PART 2 QUESTION 8 CODE

# making y_bar matrix.
y_n_mat <- matrix(c(1, ".", ".", ".", 1), ncol = 1, byrow = TRUE)
```