Homework: Multivariate Linear Regression and PCA

Problem 1: Multivariate Linear Regression — Matrix Formulation and Hypothesis Testing

Suppose we observe $n = 60$ individuals and measure:

- Two response variables:

$$\mathbf{Y} = (Y_1, Y_2)$$

- Three predictors:

$$X_1, X_2, X_3$$

The multivariate regression model is

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

where:

- $\mathbf{Y}_{n\times 2}$
- $\mathbf{X}_{n\times 4}$ includes an intercept
- $\mathbf{B}_{4\times 2}$
- $\mathbf{E}_i \sim N_2(\mathbf{0}, \mathbf{\Sigma})$

(a) Estimation

1. Derive the least squares estimator

$$\widehat{\mathbf{B}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

2. Show that

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n-p}\mathbf{E}^\top\mathbf{E}$$

   is an unbiased estimator of $\mathbf{\Sigma}$, where $p = 4$.

(b) SSCP Decomposition

Define:

- Error SSCP:

$$\mathbf{E} = \mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$$

- Hypothesis SSCP for testing $H_0 : \mathbf{CB} = 0$:

$$\mathbf{H} = (\mathbf{C}\widehat{\mathbf{B}})^\top [\mathbf{C}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{C}^\top]^{-1}(\mathbf{C}\widehat{\mathbf{B}})$$

Explain the geometric meaning of $\mathbf{H}$ and $\mathbf{E}$.

(c) Multivariate Test

Test the global hypothesis:

$$H_0 : \beta_{1\cdot} = \beta_{2\cdot} = \beta_{3\cdot} = 0$$

(where each $\beta_{j\cdot}$ is a row vector across both responses).

1. Write down Wilks' Lambda:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|}$$

2. Give the approximate F-distribution used for inference.
3. Explain why separate univariate tests would not control Type I error.

Problem 2: Multivariate Linear Regression — Practical Interpretation

A researcher studies the effect of a training program on two outcomes:

- Y_1: Cognitive score
- Y_2: Physical endurance

Predictors:

- X_1: Hours of training
- X_2: Age
- X_3: Sex (0/1)

The fitted coefficient matrix is:

$$\widehat{\mathbf{B}} = \begin{pmatrix} 50 & 30 \\ 2.5 & 1.2 \\ -0.4 & -0.1 \\ 3.0 & 2.2 \end{pmatrix}$$

Columns correspond to (Y_1, Y_2).

The estimated error covariance matrix is

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} 16 & 8 \\ 8 & 25 \end{pmatrix}.$$

(a) Interpretation

1. Interpret the coefficient for training hours for both outcomes.
2. What does the off-diagonal element of $\widehat{\boldsymbol{\Sigma}}$ imply?

(b) Joint Hypothesis

Test

$$H_0 : \text{Training hours has no effect on either outcome}$$

1. State the null in matrix form.
2. Explain how Wilks' Lambda would be constructed.
3. Explain why the correlation between outcomes affects power.

(c) Comparison to Separate Regressions

Explain what information is lost if two separate regressions are fit instead of the multivariate model.

Problem 3: Principal Components Analysis

You observe $p = 5$ standardized variables with sample correlation matrix:

$$\mathbf{R} = \begin{pmatrix} 1 & .8 & .7 & .1 & .2 \\ .8 & 1 & .75 & .05 & .1 \\ .7 & .75 & 1 & .1 & .15 \\ .1 & .05 & .1 & 1 & .6 \\ .2 & .1 & .15 & .6 & 1 \end{pmatrix}$$

The eigenvalues are:

$$\lambda = (2.6,\ 1.7,\ 0.4,\ 0.2,\ 0.1).$$

(a) Variance Explained

1. Compute the proportion of variance explained by each component.
2. Compute cumulative variance explained.
3. According to the Kaiser rule, how many components should be retained?

(b) Interpretation

Suppose the first eigenvector has large positive weights on variables 1–3 and near-zero weights on 4–5.

1. Interpret PC1.
2. Explain why PC1 captures most of the correlation structure.

(c) Optimal Approximation

1. Explain why retaining the first two PCs gives the best rank-2 approximation.
2. What does “best” mean mathematically?

(d) Conceptual Question

Explain why PCA might fail if the true structure is nonlinear (e.g., data lie on a curved manifold).

Problems from Johnson & Wichern

7.6, 7.9, 7.25