Project1

1 Description of the problem

1.1 Type of data collection

y_a1	y_a2	...	y_a24	y_b1	y_b2	...	y_b24	W	H	D	S
0.12	0.45	...	0.78	0.23	0.56	...	0.89	Mon	3	15	1.2
0.34	0.67	...	0.91	0.45	0.12	...	0.34	Tue	2	22	0.8
0.56	0.23	...	0.45	0.78	0.34	...	0.56	Wed	4	7	1.5
...	...	...	...	...	...	...	...	...	...	...	...

where - y_{ai} is the power comsumption in the ith hour within the ath day, $i = 1, ..., 24$

- y_{bj} is the the power comsumption in the jth hour within the bth weekday,$j = 1, .., 24$
- W:weekday,H:Household,D:Day,S:Seasonal

1.2 Variables Description

- type and size of the sample
- Sample Space $\{W, H\}$

where W:weekday, factors with levels 1,2,3...7, for example W=1 is monday,

- H:household,facotr with levels 1,...,33
- D:day, e.g D: 1st January
- S:sinusoidal seasonal term,$\sin\Big(\frac{2\pi}{T}(t-\varphi)\Big)$,

where $T = 365$,

$\varphi = 31 + 28 + 21 = 80$, that we considered the first three months as a cycle,then repeated this period in the inusoidal seasonal function, which is $\sin\left(\frac{D-31-28-21}{365} \cdot 2\pi\right)$

2 Statistical methods

2.1 Missing Data

2.1.1 Linear Interpolation

2.2 The way of smoothing

2.3 Modeling method

2.3.1 Multiivariate Linear Model

Sample Space:$y_a = cbind(y_{a1}, ..., y_{a24})$,$y_a = cbind(y_{b1}, ..., y_{b24})$

Additive model without interactions:

$$y_a \sim W + H + S$$

,$Y = \{y_a, y_b\} \in \mathcal{R}^{365\times 96}$

Multication with interactions:

$$y_a \sim W * H * S$$

,$Y = \{y_a, y_b\} \in \mathcal{R}^{365\times 96}$

2.3.2 Parameter Estimation

Assume that we have the sample $y_n = (y_{a1_n}, ..., y_{a24_n})^T$=$y_n = (y_n^{(1)}, ..., y_n^{(24)})^T$,

where n is the weekday of nth day, $n = 1, ..., N = 365 * 30$

We can estimate the parameter by OLS,

$$\hat{\theta}_i = (X^T X)^{-1} X^T y^{(i)}$$

2.4 Multivariate Time Series Model

2.4.1 Data Representation

The observation vector y consists of N time points with 24-dimensional observations:

$$y = \begin{pmatrix} y_1^T \\ \vdots \\ y_N^T \end{pmatrix} = \left(y^{(1)}, \dots, y^{(24)} \right)$$

2.4.2 Model Structure

$$y = X \cdot \Theta + Z$$

where: - **Parameter space**: $\Theta = \begin{pmatrix} \theta_1 & \cdots & \theta_{24} \end{pmatrix}$ - **Design matrix**: X $(N \times p)$ - **Noise term**: $Z = \left(z^{(1)}, \dots, z^{(24)} \right)$

2.4.3 Component Form

For each time point t:

$$y_t = X_t \cdot \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_{24} \end{pmatrix} + z_t$$

2.4.4 Notation Guide

Symbol	Dimension	Description
$y^{(k)}$	$N \times 1$	Observation sequence for feature k
θ_k	$p \times 1$	Parameter vector for feature k
$z^{(k)}$	$N \times 1$	Noise term for feature k

2.5 Statistic Test

2.5.1 MANOVA

Before we use multilevel statistic test. there are some assumptions needed to be checked

1. **Multivariate normality**

$$E_i \sim N_p(0, \Sigma), \quad i = 1, \dots, n$$

Each residual vector is drawn from the same p-variate normal distribution.

2. **Independence**
   The residuals E_i are mutually independent and identically distributed.
3. **Homoscedasticity**
   The covariance matrix (Σ) is constant across all observations and under both the full and reduced models.
4. **Full-rank design matrix**
   ($\text{rank}(X) = q$), ensuring the parameter matrix (B) is identifiable and ($\widehat{\Sigma}$) is invertible.
5. **Correct model specification**
   The true relationship follows

$$Y = X\Theta + Z,$$

   and any linear constraints under (H_0) are correctly specified.

2.6 Multivariate Linear Model, MLM

2.6.1 test statistics

This part describes the transition from univariate hypothesis testing to multivariate hypothesis testing, and demonstrates how to apply multivariate test statistics (Roy's Largest Root, Lawley–Hotelling Trace, Pillai's Trace, Wilks' Lambda) to a dataset with 48 response variables. The goal is to perform significance testing based on these statistics.

2.6.2 Data Structure and Notation

- Let (N) denote the sample size.
- The 48 response variables are organized as: $Y = [, y_{a1}, \cdots, y_{a24}, ; y_{b1}, \cdots, y_{b24},] \in \mathbb{R}^{N\times 48}$
- The covariates (including intercept) form the design matrix: $X = [\mathbf{1}, ; W, ; H, ; D, ; S] \in \mathbb{R}^{N\times p}, \quad p = 5.$
- The reduced model includes only the intercept: $X_0 = \mathbf{1}_N \in \mathbb{R}^{N\times 1}$.

2.7 Detailed Methods

2.7.1 Univariate F-Test

1. **Model:** $H_0:; y = X_0\beta_0 + \varepsilon \quad \text{vs.} \quad H_1:; y = X\beta + \varepsilon, $ where $y \in \mathbb{R}^N$.
2. **Projection matrices:** $P_0 = X_0(X_0 \top X_0)-1X_0^\top, \quad P = X(X \top X)-1X^\top$.
3. **Sum of Squares:** $\text{SSR} = y^\top(P - P_0)y, \quad \text{SSE} = y^\top(I - P)y$.
4. **F-statistic:** $F = \frac{\text{SSR}/(p-1)}{\text{SSE}/(N-p)} \sim F_{p-1,,N-p}$.

2.7.2 Multivariate Hypothesis Testing

2.7.2.1 Projection and Sum of Squares Matrices

- Using the same projection matrices (P) and (P_0) from above.
- Define:$H = Y^\top(P - P_0)Y, \quad E = Y^\top(I - P)Y$.

2.7.2.2 Eigenvalue Decomposition

- Compute the matrix: $A = HE^{-1}$.
- Obtain eigenvalues $\{\lambda_i\}_{i=1}^r$ where$r = \min(p - 1, 48)$

2.7.2.3 Multivariate Test Statistics

- **Roy's Largest Root:** Focuses on the strongest signal direction.
- **Lawley–Hotelling Trace:** Aggregates the overall multivariate signal strength.
- **Pillai's Trace:** Sums the explained proportion of each dimension.
- **Wilks' Lambda:** Measures the proportion of unexplained variation.

2.7.3 Connection to Univariate Testing

- When (q=1), (H) and (E) reduce to scalars, and all four statistics simplify to the classical univariate F-statistic.

2.7.4 AIC