1) Consider a simple linear regression $Y = X\beta + \epsilon$ with an intercept and one predictor based on a sample of size 4. Or specifically,

$$\begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} + \epsilon. \tag{1}$$

Calculate $(X'X)^{-1}$, $X'Y$, $\hat{\beta}$, $\hat{y}$, and $\hat{\epsilon}$ by hand.

$$Y = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1+1+1+1 & 1+1+0.5+2 \\ 1+1+0.5+2 & 1+1+0.25+4 \end{bmatrix} = \begin{bmatrix} 4 & 4.5 \\ 4.5 & 6.25 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{|4(6.25) - 4.5(4.5)|} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \frac{1}{4.75} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix}$$
$$\approx \begin{bmatrix} 1.3158 & -0.9474 \\ -0.9474 & 0.8421 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 0.5 - 0.5 + 0.3 + 1.2 \\ 0.5 - 0.5 + 0.5 * 0.3 + 2 * 1.2 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix} = \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} \approx \begin{bmatrix} -0.4421 \\ 0.7263 \end{bmatrix}$$

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} = \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} \approx \begin{bmatrix} 0.2842 \\ 0.2842 \\ -0.0789 \\ 1.0105 \end{bmatrix}$$

$$\hat{\epsilon} = (y - \hat{y}) = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} - \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} = \begin{bmatrix} 41/190 \\ -149/190 \\ 36/95 \\ 18/95 \end{bmatrix} \approx \begin{bmatrix} 0.2158 \\ -0.7842 \\ 0.3789 \\ 0.1895 \end{bmatrix}$$

2) Consider the model $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \beta_4\boldsymbol{x_4} + \boldsymbol{\epsilon}$. Give the appreciate $\boldsymbol{C}$ and $\boldsymbol{\theta}_0$ for testing the following hypotheses.

(a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 \equiv \begin{matrix}\beta_1 - \beta_4 = 0 \\ \beta_2 - \beta_4 = 0 \\ \beta_3 - \beta_4 = 0\end{matrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \quad \begin{bmatrix}0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{3x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - \beta_4 \\ \beta_2 - \beta_4 \\ \beta_3 - \beta_4\end{bmatrix}_{3x1} = \begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}_{3x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{3x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}_{3x1}$$

(b) $H_0 : \begin{pmatrix}\beta_1 \\ \beta_3\end{pmatrix} = \begin{pmatrix}\beta_2 + 2 \\ \beta_4\end{pmatrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \quad \begin{bmatrix}0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{2x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - \beta_2 \\ \beta_3 - \beta_4\end{bmatrix}_{2x1} = \begin{bmatrix}2 \\ 0\end{bmatrix}_{2x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{2x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}2 \\ 0\end{bmatrix}_{2x1}$$

(c) $H_0 : \begin{pmatrix}\beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2\end{pmatrix} = \begin{pmatrix}4\beta_3 \\ -6\end{pmatrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \begin{bmatrix}0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0\end{bmatrix}_{2x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - 2\beta_2 - 4\beta_3 \\ \beta_1 + 2\beta_2\end{bmatrix}_{2x1} = \begin{bmatrix}0 \\ -6\end{bmatrix}_{2x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0\end{bmatrix}_{2x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}0 \\ -6\end{bmatrix}_{2x1}$$

3) Consider the model $\boldsymbol{y}_{5x1} = \boldsymbol{X}_{5x3}\boldsymbol{\beta}_{3x1} + \boldsymbol{\varepsilon}_{5x1}$, where

$$\boldsymbol{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \boldsymbol{X} = [\boldsymbol{J} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

(a) Show as rigorously as possible whether $\theta_1 = \beta_2$ is estimable.

$\boldsymbol{X}$ is not full rank. $r(\boldsymbol{X}) = 2 < 3$ $\qquad\qquad$ $\boldsymbol{x_2} = \boldsymbol{J} - \boldsymbol{x_1}$

$\theta_1 = \beta_2$ is estimable if there exists a $\boldsymbol{T}$ matrix for $\boldsymbol{\theta} = \boldsymbol{C\beta} = \boldsymbol{T}E(\boldsymbol{Y})$such that $\boldsymbol{C} = \boldsymbol{TX}$.

$$\theta_1 = \beta_2 \quad \Rightarrow \quad \boldsymbol{\theta} = \boldsymbol{C\beta} \equiv \theta_1 = [0 \quad 0 \quad 1]\boldsymbol{\beta} = \beta_2$$

$$\Rightarrow \quad \boldsymbol{C} = \boldsymbol{TX} \equiv [0 \quad 0 \quad 1] = [t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5]\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \quad [0 \quad 0 \quad 1] = \left[\sum_{i=1}^{5} t_i \quad t_1 + t_2 + t_5 \quad -t_3 - t_4\right]$$

$$\Rightarrow \quad \begin{matrix} t_1 + t_2 + t_3 + t_4 + t_5 = 0 \\ t_1 + t_2 + t_5 = 0 \\ -t_3 - t_4 = 1 \end{matrix} \Rightarrow \begin{matrix} t_3 + t_4 = 0 \\ -t_3 - t_4 = 1 \equiv t_3 + t_4 = -1 \end{matrix}$$

Since $t_3 + t_4 = -1 \neq 0$, there is no $\boldsymbol{T}$ that can satisfy the equation $\boldsymbol{C} = \boldsymbol{TX}$.

$\therefore \theta_1 = \beta_2$ is not estimable

3) Consider the model $\boldsymbol{y}_{5x1} = \boldsymbol{X}_{5x3}\boldsymbol{\beta}_{3x1} + \boldsymbol{\varepsilon}_{5x1}$, where

$$\boldsymbol{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \boldsymbol{X} = [\boldsymbol{J} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

(b) Show as rigorously as possible whether $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

For $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ to be testable, it must also be estimable with either $\boldsymbol{C}$ or $\boldsymbol{M}$ being full rank.

$$\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix} \qquad \Rightarrow \qquad \boldsymbol{\theta} = \boldsymbol{C\beta} \equiv \boldsymbol{\theta_2} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{bmatrix}$$

$$\Rightarrow \qquad \boldsymbol{C} = \boldsymbol{TX} \equiv \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \qquad \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{5} t_{1i} & t_{11} + t_{12} + t_{15} & -t_{13} - t_{14} \\ \sum_{i=1}^{5} t_{2i} & t_{21} + t_{22} + t_{25} & -t_{23} - t_{24} \end{bmatrix}$$

These equations are satisfied by $\boldsymbol{T} = \begin{bmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$. This is one of many solutions for $\boldsymbol{T}$.

$\therefore \boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is estimable

$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix} \Rightarrow r(\boldsymbol{C}) = 2$ $\therefore \boldsymbol{C}$ is full rank

Since $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is estimable and $\boldsymbol{C}$ is full rank, $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

4) A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights (y = WGHT), average daily exercise time (x = TIME). One of the objectives in this study is to investigate the effect of TIME on weight loss.

(a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete this table.

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	2624.670184/1 = 2624.670184	137.906162 ~F(1,96)	2.86E-20
Error	96	1827.099916	1827.099916/96 = 19.032291		
Corrected Total	97	4451.7701			

(b) State the model assumptions based on which the ANOVA table was computed.

HILE-Gauss:

1. Homogeneity Assumption: We assume each row of ε has same variance σ^2.
2. Independence Assumption: We assume each row of ε is statistically independent of every other row.
3. Linearity Assumption: We assume the expected value of the response are linear functions of the parameter. $E(y) = X\beta$.
4. Existence Assumption: We observe values of random variables with finite variance. $H_0: \sigma^2_{model} = \sigma^2_{error}$
5. The error term follows a Gaussian distribution. $\varepsilon_i \sim N(0, \sigma^2)$

(c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

Yes, daily exercise time is a significant predictor for predicting weight loss.
The test statistic is 137.9062.
The F-test of $\beta_{WGHT} = 0$ for $y = X\beta + \varepsilon$ generates a p-value < 0.0001.
We reject the null hypothesis that $\beta_{WGHT} = 0$ (the average daily exercise time is not significant).
Therefore, daily exercise time appears to be a significant predictor for predicting weight loss.

5) An investigator studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average temperature in degrees Celsius (the predictor). The data *hw2.dat*, is provided on Sakai.

(a) Fit a regression model with the number of high ozone days as the response and the meteorological index as a covariate and provide estimates of β_0, β_1, their standard errors, and their interpretations.

Table 5.1: Regression Model for Number of High Ozone Days

Variable	Parameter Estimate	Standard Error	Interpretation
β_0	-192.98	163.503	The number of high ozone days with a meteorological index temperature of 0.
β_1	15.30	9.421	The change in the number of days with high ozone with every increase in the meteorological index temperature by 1 degree Celsius.

(b) Are all of the β's estimable? Why or why not?

Yes. X is full rank ($r(\boldsymbol{X}) = 2 = p = r$). Therefore, all the β's are estimable.

(c) Report a test of the hypothesis that the number of high ozone days is associated with the meteorological index.

Hypothesis: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

Test Statistic: $t_{obs} = \frac{\widehat{\beta}_1 - \beta_1}{SE_{\beta_1}} = \frac{15.30-0}{9.421} = 1.62 \sim t_{14}$

Degrees of Freedom: $df = 16 - 2 = 14$

P-value: $\Pr(|t| > 1.62) = 2 * (1 - \Pr(t \leq 1.62)) = 0.1267$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that there is an association between the number of high ozone days and meteorological index.

(d) Using the framework of the linear model, report an $\alpha = 0.05$ test of the hypothesis that a 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

Hypothesis: $H_0: \beta_1 = 12$ vs. $H_1: \beta_1 \neq 12$

Test Statistic: $t_{obs} = \frac{\hat{\beta}_1 - \beta_1}{SE_{\beta_1}} = \frac{15.30-12}{9.421} = 0.35 \sim t_{14}$ or
$F = 0.12 \sim F(1,14)$

Degrees of Freedom: $df = 16 - 2 = 14$

Critical Region: $C_\alpha = \left\{t: |t| > t_{v,1-\frac{\alpha}{2}}\right\} \rightarrow C_{0.05} = \{t: |t| > 2.1448\}$

P-value: $\Pr(|t| > 0.35) = 2 * (1 - \Pr(t \leq 0.35)) = 0.7316$ or

$\Pr(F_{1,14} > 0.12) = 1 - \Pr(F_{1,14} \leq 0.12) = 0.7316$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that a 1 degree increase in average temperature is not associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

(e) Calculate the 95% confidence interval and prediction interval for the expected number of days the ozone level exceeded 0.2 ppm when the seasonal meteorological index is 16.

95% Confidence Interval: $(21.7579, 81.7581)$

When the seasonal meteorological index is 16, there is a 95% confidence that the average number of days the ozone level exceeded 0.2 ppm is between 21.76 and 81.76 days.

95% Prediction Interval: $(-7.4416, 110.9576)$

Based on the observed data, there is a 95% chance that a seasonal meteorological index of 16 will result in between 0 and 110.96 days that the ozone level exceeded 0.2ppm.