1. Consider a data set provided by the Wisconsin Department of Health and Family Services (DHFS), which involves the following variables:

- **TPY**: total patient years;
- **NUMBED**: number of beds.

The sample size is \( n = 717 \). An analyst proposed to conduct a linear regression of \( \text{LOGTPY} = \log(\text{TPY}) \) on \( \text{LOGNUMBED} = \log(\text{NUMBED}) \), where \( \log(\cdot) \) denotes the natural logarithm. The associated R output is given below.

\[
\text{> summary(lm(LOGTPY ~ LOGNUMBED))}
\]

Coefficients:

|                | Estimate | Std. Error | t value      | Pr(>|t|)  |
|----------------|----------|------------|--------------|-----------|
| (Intercept)    | -0.163315| 0.036045   | -4.531       | 6.88e-06 *** |
| LOGNUMBED      | 1.015739 | 0.008038   | 126.372      | <2e-16 ***  |

---

Residual standard error: 0.1058 on 715 degrees of freedom
Multiple R-squared: 0.9571,
F-statistic: 1.597e+04 on 1 and 715 DF

(a) Provide a possible reasoning on why the logarithmic transform should be used in the linear regression.

(b) Based on the provided information, is it possible to provide a predicted value for \( \text{TPY} \) when \( \text{NUMBED} = 200 \)? If yes, find the predicted value. If no, explain. How about the associated prediction interval?

(c) Suppose you replace \( \text{LOGNUMBED} \) by \( \text{LOG2NUMBED} = \log_2(\text{NUMBED}) \), where \( \log_2(\cdot) \) denotes the logarithm with base 2. Compute the new regression summary by filling the template below. Use XXX to fill entries that you think cannot be computed from the provided information.

\[
\text{> summary(lm(LOGTPY ~ LOG2NUMBED))}
\]

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -------- | ---------- | --------|----------|
| LOG2NUMBED     | -------- | ---------- | --------|----------|

---

Residual standard error: XXX on 715 degrees of freedom
Multiple R-squared: XXX,
F-statistic: XXX on 1 and 715 DF
Residual standard error: _________ on ____ degrees of freedom

Multiple R-squared: _________,

F-statistic: _________ on _________ and _________ DF

(d) Suppose you further replace LOGTPY by LOG2TPY = \log_2(TPY). Compute the new regression summary by filling the template below. Use XXX to fill entries that you think cannot be computed from the provided information.

> summary(lm(LOG2TPY ~ LOG2NUMBED))

Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | _________ | _________ | _________ | _________ |
| LOG2NUMBED | _________ | _________ | _________ | _________ |

Residual standard error: _________ on ____ degrees of freedom

Multiple R-squared: _________,

F-statistic: _________ on _________ and _________ DF

2. Consider the linear regression model \( Y = X\beta + e \), where

\[
Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \quad X = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 2 \\ -1 & 1 & -2 \\ 1 & -1 & 2 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}.
\]

Assume that \( e_i, i = 1, \ldots, 4 \), are independent normal random variables with mean zero and variance \( \sigma^2 > 0 \). Suppose you only observe \( Y \) and \( X \). Let \( a = (a_1, a_2, a_3)^T \) be a column vector (\(^T\) denotes the transpose), we are interested in making statistical inference about the linear combination \( a^T \beta = \sum_{j=1}^{3} a_j \beta_j \).

(a) Is there any problem that you may have when computing the least squares estimate? If yes, then what causes it?

(b) Is it possible to obtain an unbiased estimate of \( \beta_2 \)? If yes, provide the estimate. If not, explain. How about the quantity \( \zeta = \beta_1 + \beta_2 + 2\beta_3 \)? [Hint: The unbiased estimate here does not need to be the BLUE.]
(c) The above system is not identifiable as there exists a nonzero vector $\gamma$ such that $Y = X(\beta + \gamma) + e$ also holds. Find one such $\gamma$.

(d) Is it possible to form a statistical test for the null hypothesis $H_0 : \beta_1 = 0$? If so, provide the test statistic and its distribution under the null. If not, explain.

(e) Is it possible to form a statistical test for the null hypothesis $H_0 : \beta_2 = 0$? If so, provide the test statistics and its distribution under the null. If not, explain.