Sampling

Question 1.

(a) i <u>C</u> ii <u>C</u> iii <u>D</u>

(b) i <u>C</u> ii <u>D</u> iii <u>E</u> iv <u>B</u>

(c) i No, this is not a SRS.

For SRS, each individual should have the same chance of being selected. In this case, 5 friends of Matthew are guaranteed to be selected which means they have 100% chance to be selected. This violates the definition of SRS.

ii Yes, this is a Probability Sample.

For probability sample, individuals in the population can have different chances of being selected; they don't have to be uniform. In this case, every student in Data 100 has a known, non-zero chance of selection. 5 friends of Matthew has 100% probability and each of remaining students has the same known probability to be selected.

Take care of Yourself

Question 2.

(a) Sleep: $\frac{7.5\times20+7\times15+6\times15}{20+15+15}$ Coffee: $\frac{1.5\times20+2\times15+4\times15}{20+15+15}$

$= 6.9$ hours $= 2.4$ cup.

(b) Underclassmen:

experimental value: $\frac{20}{50} = 0.4$

theoretical value: $\frac{400}{1000} = 0.4$

Percent Error $= \frac{|0.4-0.4|}{0.4} \times 100\% = 0\%$

Upperclassmen:

experimental value: $\frac{15}{50} = 0.3$

theoretical value: $\frac{500}{1000} = 0.5$

Percent error $= \frac{|0.3-0.5|}{0.5} \times 100\% = 40\%$

Graduate

experimental value: $\frac{15}{50} = 0.3$

theoretical value $= \frac{100}{1000} = 0.1$

Percent Error $= \frac{|0.3-0.1|}{0.1} \times 100\% = \underline{\underline{200\%}}$

Thus, Graduate Student deviates most.

(c) Sleep:

$0.4 \times 7.5 + 0.5 \times 7 + 0.1 \times 6$

$= 3 + 3.5 + 0.6$

$= 7.1$ hours

Coffee:

$0.4 \times 1.5 + 0.5 \times 2 + 0.1 \times 4$

$= 0.6 + 1 + 0.4$

$= 2$ cups

We assume all samples are representative which means there're no samples that not review for the midterm

(d) Not exactly true. Because it's unlikely that the review-session attendees perfectly represent their group. The ones attending can differ in habits and other personal schedules, making the assumption false.

Properties of a Linear Model with No Constant Term

Question 3

$$\frac{\partial R(\theta)}{\partial \theta} = \frac{\partial \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)^2}{\partial \theta}$$

$$= \frac{1}{n}\sum_{i=1}^{n} 2(y_i - \theta x_i)(-x_i)$$

$$= \frac{2}{n}\sum_{i=1}^{n}(x_i^2\theta - x_i y_i)$$

let $R'(\theta) = 0$

then $\cancel{\frac{2}{n}}\sum_{i=1}^{n}(x_i^2\theta - x_i y_i) = 0$

$$\sum_{i=1}^{n} x_i^2 \theta = \sum_{i=1}^{n} x_i y_i$$

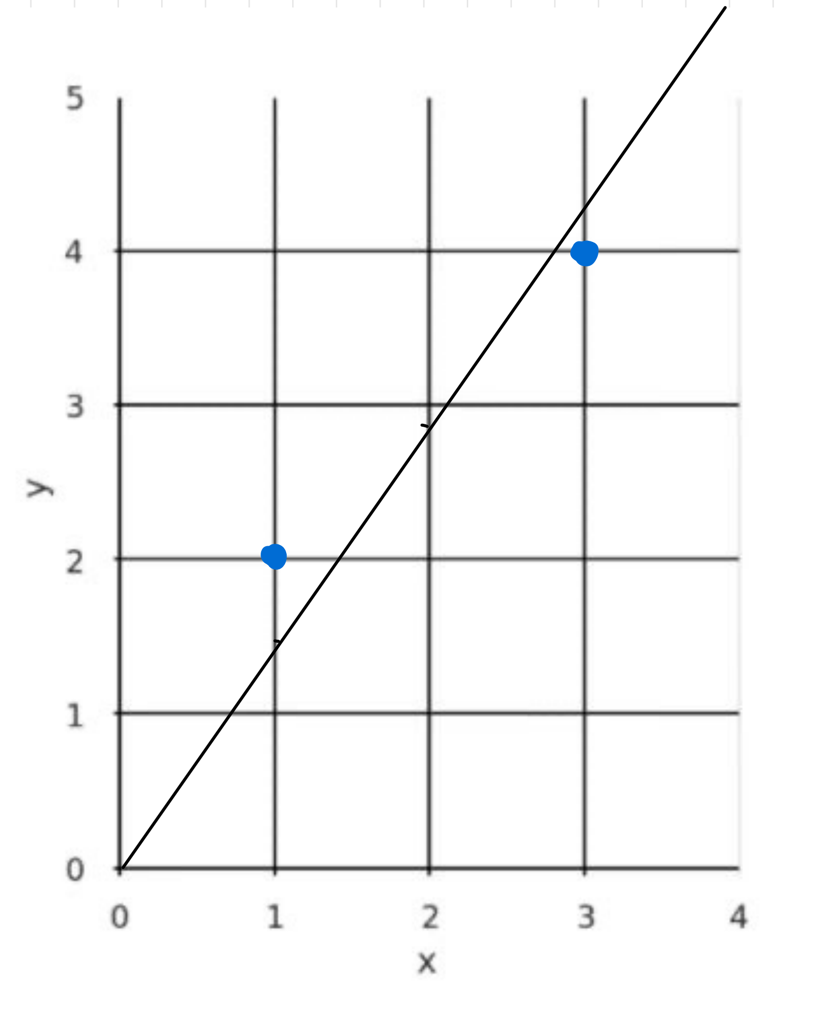
$$\theta \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\theta = \frac{\sum x_i y_i}{\sum x_i^2}$$

Question 4

(a) $\hat{\theta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{1 \times 2 + 3 \times 4}{1 + 9} = 1.4$

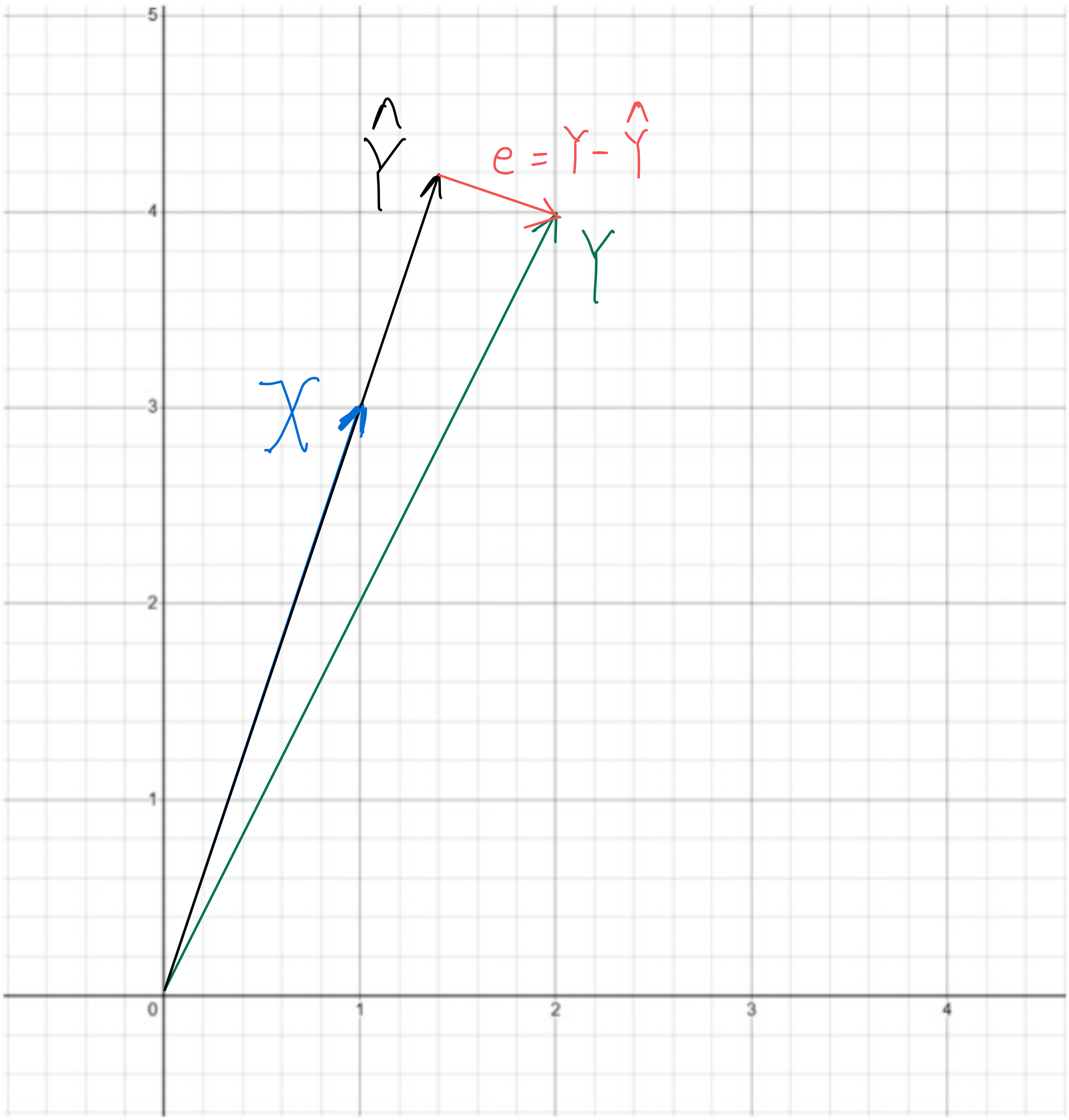
$\hat{y} = 1.4x$



(b) $X = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ $Y = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$

$$\hat{Y} = \hat{\theta} X = 1.4 \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 1.4 \\ 4.2 \end{bmatrix}$$

(c)



(d)

$$\text{proj}_X Y = \frac{\begin{bmatrix} 2 \\ 4 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 3 \end{bmatrix}}{10} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$= 1.4 \times \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1.4 \\ 4.2 \end{bmatrix} = \hat{Y}$$

The vector $\hat{Y}$ is equal to the projection

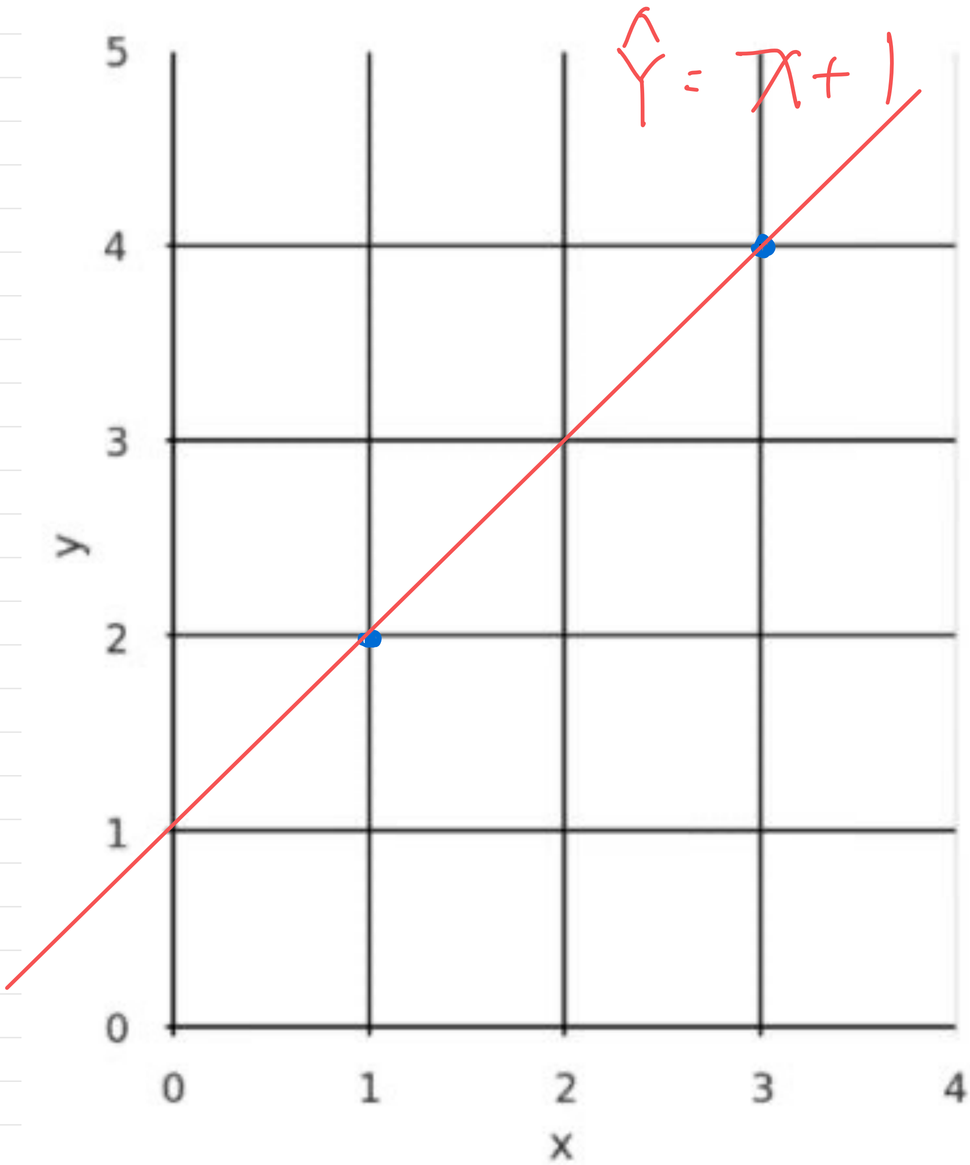
(e) Formula in Q3 = $\hat{\theta} = \frac{\sum x_i y_i}{\sum x_i^2}$

Formula in Q4 : $\to \hat{Y} = \theta X$

Projection formula: $\to \text{proj}_X Y = \frac{Y \cdot X}{\|X\|_2^2} \cdot X$

The prediction vector $\hat{Y}$ is exactly the projection of Y onto X. The best θ that minimize the error is exactly the coefficient used in the projection formula.

(f)



MSE Minimizer

Question 5

(a) $g_i(\theta) = \frac{1}{n}(y_i - \theta x_i)^2$

$$\frac{d g_i(\theta)}{d\theta} = \frac{2}{n}(y_i - \theta x_i)(-x_i)$$

$$= \frac{2}{n}(\theta x_i^2 - x_i y_i)$$

$$\frac{d}{d\theta}\frac{d g_i(\theta)}{d\theta} = \frac{2}{n} x_i^2$$

As $n \geq 0$, $x_i^2 \geq 0$, then $\frac{d}{d\theta}\frac{d g_i(\theta)}{d\theta}$ is guaranteed to be non-negative. We can verify $g_i(\theta)$ is a convex function.

(b) i if $g(\theta)$ & $h(\theta)$ is convex, then

$$\begin{cases} g(c\times\theta_i + (1-c)\times\theta_j) \le c\times g(\theta_i) + (1-c)\,g(\theta_j) & ① \\ h(c\times\theta_i + (1-c)\times\theta_j) \le c\times h(\theta_i) + (1-c)\,h(\theta_j) & ② \end{cases}$$

Consider $f(\theta) = g(\theta) + h(\theta)$, if $f(\theta)$ is convex

$$f(c\theta_i + (1-c)\theta_j) \le c f(\theta_i) + (1-c) f(\theta_j)$$

Based on ① + ②, we have

$$g(c\theta_i + (1-c)\theta_j) + h(c\theta_i + (1-c)\theta_j)$$

$$\le c g(\theta_i) + (1-c) g(\theta_j) + c h(\theta_i) + (1-c) h(\theta_j)$$

Rearrange right hand side

$$\le c g(\theta_i) + c h(\theta_i) + (1-c) g(\theta_j) + (1-c) h(\theta_j)$$

which is $c f(\theta_i) + (1-c) f(\theta_j)$

Then we get

$$g(c\theta_i + (1-c)\theta_j) + h(c\theta_i + (1-c)\theta_j) \le c f(\theta_i) + (1-c) f(\theta_j)$$

$$\Rightarrow f(c\theta_i + (1-c)\theta_j) \le c f(\theta_i) + (1-c) f(\theta_j)$$

So, $f(\theta)$ is convex.

ii We can consider the sum of n convex functions as the the summation of the previous sum and the newly added convex function, it will always be 2 sums.

initial sum = $f_1(\theta) + f_2(\theta)$

then

```
for i in range(3, n+1):
    sum = sum + fi(θ)
```

cc) We've taken the second derivative of $\sum_{i=1}^{n} g_i(\theta)$ which is ≥ 0 It indicates the convexity.

In a convex function, any critical point where the gradient is 0 corresponding to a global minimum. rather than a saddle. That's why this solution is guaranteed to minimize the MSE.

Geometry perspective of SLR

Question 6.

(a) The OLS requires that the residual error vector $\vec{e} = Y - \hat{Y}$ be orthogonal to every column of X. 1_n is one of these columns, then $1_n^T \cdot \vec{e} = 0$

As 1_n^T is a vector, we can represent it

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1}$$

The $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n\times 1} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$, based on dot product calculation

it will be $1 \times e_1 + 1 \times e_2 + \cdots + 1 \times e_n = \sum_{i=1}^{n} e_i = 0$

Thus, we can derive why $\sum_{i=1}^{n} e_i = 0$

(b) Similar idea, $X_{:,1}$ is another one of these colums

then $X_{:,1}^T \cdot \vec{e} = 0$, where $X_{:,1}^T = \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{bmatrix}$

$$X_{:,1}^T \vec{e} = \begin{bmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{bmatrix}_{n\times 1} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{n\times 1}$$

$$= x_1 \cdot e_1 + x_2 \cdot e_2 + \cdots x_n \cdot e_n$$

$$= \sum_{i=1}^{n} x_i e_i = 0$$

(c) The $\hat{Y} = X\theta$ is restricted to lie in the subspace spanned by the columns of X. Geometrically, vector $\hat{Y}$ is ensured to be the closest possible point to Y by projecting Y onto X space. The "closest distance" is measured by residual vector $\vec{e}$, when $\vec{e}$ is orthoganal to X space, it reaches the closest. And $\hat{Y}$ lies on this space, so $\hat{Y}$ must be orthogonal to residual vector $\vec{e}$

A special Case of Linear Regression

Question 7

(a) $X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & -1 \end{bmatrix}$ $\quad X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix}$ $\quad Y = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$

$$\hat{\theta}^0 = (X^T X)^{-1} X^T Y$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & -1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{6} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & 0 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ -\frac{1}{2} \\ -\frac{5}{2} \end{bmatrix}$$

Thus, $\hat{\theta}^0 = \begin{bmatrix} 2 \\ -0.5 \\ -2.5 \end{bmatrix}$

(b) $MSE = \frac{1}{n}(\|Y - \hat{Y}\|_2)^2$

$$= \frac{1}{n}(\|Y - X\hat{\theta}^o\|_2)^2$$

$$Y = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & -1 \end{bmatrix} \quad \hat{\theta}^o = \begin{bmatrix} 2 \\ -0.5 \\ -2.5 \end{bmatrix}$$

Plug into MSE,

we get $\frac{1}{n}\left(\left\|\begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}\right\|_2\right)^2$

$$= \frac{1}{n} \cdot 0$$

$$= 0$$

Explanation: the OLS solution perfectly fits the data

Data points lie exactly on the plane defined by the model, and based on previous questions, residual vector is orthogonal to the column space of X, and also zero.

(c)

$$X_{new} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \qquad X_{new}^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\hat{\theta}_{new} = (X^T X)^{-1} X^T Y$$

$$= \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$= \begin{bmatrix} 3 & 0 & 2 \\ 0 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$\det\left(\begin{bmatrix} 3 & 0 & 2 \\ 0 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix} \right) = 3 \times \begin{vmatrix} 6 & 2 \\ 2 & 2 \end{vmatrix} + 2 \times \begin{vmatrix} 0 & 6 \\ 2 & 2 \end{vmatrix}$$

$$= 24 - 24 = 0$$

As det $= 0$, the matrix is non-invertible which means it's impossible to find a unique optimal solution