Data 100: Homework 5

Michael Pham

Spring 2024

Problems

1 Sampling

Problem 1.1ai. In Shiny's survey, which of the following is the population of interest?

A. All UC Berkeley Students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students who fill out Shiny's survey

Solution. C – All students enrolled in Data 100 for this semester (Spring 2024). ■

Problem 1.1aii. Which of the following is the sampling frame?

A. All UC Berkeley Students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students who fill out Shiny's survey

Solution. C – All students enrolled in Data 100 for this semester (Spring 2024). ■

Problem 1.1aiii. Which of the following is the sample?

A. All UC Berkeley Students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students who fill out Shiny's survey

Solution. D – All students who fill out Shiny's survey. ■

Problem 1.1bi. In this sampling scheme, which of the following is the population of interest?

A. UC Berkeley students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students enrolled in Shiny's discussion section

E. All students who fill out Shiny's pre-contest survey

Solution. C – All students enrolled in Data 100 for this semester (Spring 2024). ■

Problem 1.1bii. In this sampling scheme, which of the following is the sampling frame?

A. UC Berkeley students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students enrolled in Shiny's discussion section

E. All students who fill out Shiny's pre-contest survey

Solution. D – All students enrolled in Shiny's discussion section. ■

Problem 1.1biii. Which of the following is the sample?

A. UC Berkeley students

B. All students enrolled in Data 100 across all semesters (Spring 2024 and previous)

C. All students enrolled in Data 100 for this semester (Spring 2024)

D. All students enrolled in Shiny's discussion section

E. All students who fill out Shiny's pre-contest survey

Solution. E – All students who fill out Shiny's pre-contest survey. ■

Problem 1.1biv. Which of the following best characterizes the sample?

A. Simple Random Sample

B. Convenience Sample

C. Probability Sample

Solution. B – Convenience Sample. ■

2 Properties of a Linear Model with No Constant Term

Problem 2.1. Suppose that we don't include an intercept term in our model. That is, our model is now

$$\hat{y} = \theta x,$$

where θ is the single parameter for our model that we need to optimize. (In this equation, x is a scalar, corresponding to a single observation.)

As usual, we are looking to find the value $\hat{\theta}$ that minimizes the average L_2 loss (MSE) across our observed data $\{(x_i, y_i)\}, for\ i \in \{1, \dots, n\}$:

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

The estimating equations derived in the lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model.

Use calculus to find the minimizing $\hat{\theta}$.

That is, simply prove that:

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

Solution. To begin with, we want to minimize the following:

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

We do this by setting the partial derivative to zero. That is, we find the following:

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

So, we proceed as follows:

$$\begin{aligned} \frac{\partial}{\partial \theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta x_i)^2 &= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} (y_i - \theta x_i)^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} 2 (y_i - \theta x_i)(-x_i) \\ &= -\frac{2}{n} \sum_{i=1}^{n} (y_i - \theta x_i) x_i \end{aligned}$$

So, we want to find

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

From here, we observe the following:

$$\begin{aligned}
-\frac{2}{n}\sum_{i=1}^{n}\left(y_i-\theta x_i\right)x_i &= 0\\
\sum_{i=1}^{n}\left(y_i-\theta x_i\right)x_i &= 0\\
\sum_{i=1}^{n}\left(y_i x_i-\theta x_i^2\right) &= 0\\
\sum_{i=1}^{n} y_i x_i-\sum_{i=1}^{n}\theta x_i^2 &= 0\\
\sum_{i=1}^{n} y_i x_i-\theta\sum_{i=1}^{n} x_i^2 &= 0\\
\theta\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i\\
\theta &= \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

Thus, we observe that the value $\hat{\theta}$ that minimizes the average L_2 loss across our observed data is indeed:

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

■

3 MSE "Minimizer"

Recall from calculus that given some function $g(x)$, the x you get from solving $\frac{dg(x)}{dx} = 0$ is called a *critical point* of g – this means it could be a minimizer or a maximizer for g. In this question, we will explore some basic properties and build some intuition on why, for certain loss functions such as squared L_2 loss, the critical point of the empirical risk function (defined as an average loss on the observed data) will always be the minimizer.

Given some linear model $f(x) = \theta x$ for some real scalar θ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\},\ for\ i \in \{1, \dots, n\}$ as the average L_2 loss (MSE):

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

Problem 3.1a. Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\theta) = \frac{1}{n}(y_i - \theta x_i)^2$ for $i \in \{1, \dots, n\}$. In this case, note that the MSE can be written as $\sum_{i=1}^{n} g_i(\theta)$.

Recall from calculus that we can use the 2nd derivative of a function to describe its curvature about a certain point (if it is facing concave up, down, or possibly a point of inflection). You can take the following as a fact: a function is convex if and only if the function's 2nd derivative is non-negative on its domain.

Based on this property, verify that $g_i(\theta)$ is a **convex function**.

Solution. We check that $g_i(\theta)$ is convex by taking the second derivative with respect to its domain. In this case, note that the domain for g_i is θ. So, we proceed as follows:

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{2}{n}\left(y_i - \theta x_i\right)\left(-x_i\right)\right)\\ &= \frac{2x_i^2}{n}\end{aligned}$$

Then, we note that since x_i is being squared, it will always be non-negative. Furthermore, since $n \geq 0$, it follows that it will also always be non-negative as well. Therefore, the second derivative is non-negative, and thus we conclude that $g_i(\theta)$ is a convex function as desired. ■

Problem 3.1bi. Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(\theta)$ is convex if for any two points $(\theta_i, g(\theta_i))$ and $(\theta_j, g(\theta_j))$ on the function,

$$g(c \times \theta_i + (1-c) \times \theta_j) \leq c \times g(\theta_i) + (1-c) \times g(\theta_j)$$

for any real constant $0 \leq c \leq 1$.

Using the definition above, show that if $g(\theta)$ and $h(\theta)$ are both convex functions, their sum $g(\theta) + h(\theta)$ will also be a convex function.

Solution. Using the definition given, we have the following for any two points θ_i, θ_j in the functions g and h's domain, along with any real constant $0 \leq c \leq 1$:

$$\begin{aligned}g(c \times \theta_i + (1-c) \times \theta_j) \leq c \times g(\theta_i) + (1-c) \times g(\theta_j)\\ h(c \times \theta_i + (1-c) \times \theta_j) \leq c \times h(\theta_i) + (1-c) \times h(\theta_j)\end{aligned}$$

Then with this in mind, we observe the following:

$$\begin{aligned}(g+h)(c\times\theta_i+(1-c)\times\theta_j) &= g(c\times\theta_i+(1-c)\times\theta_j)+h(c\times\theta_i+(1-c)\times\theta_j)\\ &\le c\times g(\theta_i)+(1-c)\times g(\theta_j)+h(c\times\theta_i+(1-c)\times\theta_j)\\ &\le c\times g(\theta_i)+(1-c)\times g(\theta_j)+c\times h(\theta_i)+(1-c)\times h(\theta_j)\end{aligned}$$

■

Problem 3.1bii. Based on what you have shown in the previous part, explain intuitively why a (finite) sum of n convex functions is still a convex function when $n > 2$.

Solution. We observe that a function $g = f_1 + f_2 + \ldots + f_n$ where $f_1, f_2, \ldots, f_n$ are convex functions and $n > 2$, when we evaluate it at some point $c\times\theta_i+(1-c)\times\theta_j$, this is equivalent to evaluating the sum of each of the functions f_i where $i \in \{1, 2, \ldots, n\}$ evaluated at that point.

Then, we can iteratively apply the definition for a function to be convex onto each of the functions f_i and eventually see that g itself will satisfy the convex condition. ■

Problem 3.1c. Remember from part (a) that the MSE can be written as:

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

We solve for its critical point by taking the gradient with respect to parameter θ and setting that expression to 0. Explain why this solution is guaranteed to minimize the MSE.

Solution. We observe that each of the g_i's are convex functions. As such, their sum will also be a convex function. From here, we see that taking the gradient with respect to θ and setting it equal to zero will thus find the minimum value of the function. ■

4 Geometric Perspective of Simple Linear Regression

In Lecture 12, we viewed both the simple linear regression model and the multiple linear regression model through the lens of linear algebra. The key geometric insight was that if we train a model on some design matrix $\mathbb{X}$ and true response vector $\mathbb{Y}$, our predicted response $\hat{\mathbb{Y}} = \mathbb{X}\hat{\theta}$ is the vector in span($\mathbb{X}$) that is closest to $\mathbb{Y}$.

In the simple linear regression case, our optimal vector θ is $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]$, and our design matrix is

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} | & | \\ 1_n & \mathbb{X}_{:,1} \\ | & | \end{bmatrix}$$

This means we can write our predicted response vector as $\hat{\mathbb{Y}} = \mathbb{X} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \hat{\theta}_0 1_n + \hat{\theta}_1 \mathbb{X}_{:,1}$.

In this problem, 1_n is the n-vector of all 1's and $\mathbb{X}_{:,1}$ refers to the n-length vector

$$[x_1, x_2, ..., x_n]^\top$$

.

Note, $\mathbb{X}_{:,1}$ is a feature, not an observation.

For this problem, assume we are working with the **simple linear regression model**, though the properties we establish here hold for any linear regression model that contains an intercept term.

Problem 4.1a. Explain why $\sum_{i=1}^{n} e_i = 0$ using a geometric property.

Solution. Recall that $\vec{e}$ is orthogonal to the span of $\mathbb{X}$. Furthermore, we note that $\mathbb{X}$ is defined to be:

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Then, we see that $1_n \in \operatorname{span} \mathbb{X}$.

Next, recall that $\vec{e}$ is defined to be

$$\vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Then, we note that because $\vec{e}$ is orthogonal to $\operatorname{span} \mathbb{X}$, it follows then that $\vec{e}$ is orthogonal to 1_n; in other words, we have:

$$\begin{aligned} \vec{e} \cdot 1_n &= 0 \\ \vec{e}^\mathsf{T} 1_n &= 0 \\ \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} &= 0 \\ \sum_{i=1}^{n} e_i &= 0 \end{aligned}$$

■

Problem 4.1b. Similarly, explain why $\sum_{i=1}^{n} e_i x_i = 0$ using a geometric property.

Solution. As stated in the previous question, $\vec{e}$ is orthogonal to $\operatorname{span} \mathbb{X}$. Then, we note that

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Then, we see that $\vec{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^\mathsf{T}$ is in $\operatorname{span} \mathbb{X}$; in other words, $\vec{e}$ and $\vec{x}$ are orthogonal to each other. Then, by definition, their dot product will be equal to zero, so we see that:

$$\begin{aligned} \vec{e} \cdot x &= 0 \\ \vec{e}^\mathsf{T} x &= 0 \\ \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} &= 0 \\ \sum_{i=1}^{n} e_i x_i &= 0 \end{aligned}$$

■

Problem 4.1c. Briefly explain why the vector $\hat{\mathbb{Y}}$ must also be orthogonal to the residual vector $\vec{e}$.

Solution. Finally, we observe that by definition, we have that $\hat{\mathbb{Y}} = \mathbb{X} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \hat{\theta}_0 1_n + \hat{\theta}_1 \mathbb{X}_{:,1}$.

Then, we observe the following:

$$\begin{aligned}
\vec{e} \cdot \hat{\mathbb{Y}} &= \vec{e}^{\intercal} \hat{\mathbb{Y}} \\
&= \vec{e} \cdot \left(\mathbb{X} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} \right) \\
&= \vec{e} \cdot \left(\hat{\theta}_0 1_n + \hat{\theta}_1 \mathbb{X}_{:,1} \right) \\
&= \vec{e} \cdot \left(\hat{\theta}_0 1_n \right) + \vec{e} \cdot \left(\hat{\theta}_1 \mathbb{X}_{:,1} \right) \\
&= \hat{\theta}_0 \left(\vec{e} \cdot 1_n \right) + \hat{\theta}_1 \left(\vec{e} \cdot \mathbb{X}_{:,1} \right) \\
&= \hat{\theta}_0 (0) + \hat{\theta}_1 (0) \\
&= 0
\end{aligned}$$

And since their dot product is equal to zero, it follows that the two must be orthogonal as desired. ■

5 A Special Case of Linear Regression

In this question, we fit two models:

$$\begin{aligned} y^S &= \theta_0^S + \theta_1^S x_1 \\ y^O &= \theta_0^O + \theta_1^O x_1 + \theta_2^O x_2 \end{aligned}$$

using L2 loss. The superscript S is to denote a Simple Linear Regression (SLR) and O is used to denote an Ordinary Least Square (OLS) with two features, respectively.

The data are given below:

$\mathbb{Y}$	bias	$\mathbb{X}_{:,1}$	$\mathbb{X}_{:,2}$
-1	1	1	-1
3	1	-2	0
4	1	1	1

Problem 5.1a. Find $\hat{\theta_0^S}$ and $\hat{\theta_1^S}$ using the formulas derived in lecture 10 ($\hat{\theta}_1^S = r\frac{\sigma_y}{\sigma_x}$ and $\hat{\theta}_0^S = \bar{y} - \hat{\theta}_1^S \bar{x}$). Show all steps.

Solution. Before we start, we calculate some important values. First, we observe that

$$\begin{aligned} \bar{x}_1 &= (1+1-2)/3 \\ &= 0 \\ \bar{y} &= (-1+3+4)/3 \\ &= 2 \\ \sigma_x &= \sqrt{\frac{(1^2+(-2)^2+1^2)}{3}} \\ &= \sqrt{2} \\ \sigma_y &= \sqrt{\frac{((-1-2)^2+(3-2)^2+(4-2)^2)}{3}} \\ &= \sqrt{14/3} \\ r &= \frac{1}{3}\left(\left(\frac{1}{\sqrt{2}}\right)\left(\frac{-1-2}{\sqrt{14/3}}\right)+\left(\frac{-2}{\sqrt{2}}\right)\left(\frac{3-2}{\sqrt{14/3}}\right)+\left(\frac{1}{\sqrt{2}}\right)\left(\frac{4-2}{\sqrt{14/3}}\right)\right) \\ &= \frac{1}{3}\left(\frac{-3}{\sqrt{2}\sqrt{14/3}}\right) \\ &= -\frac{1}{\sqrt{2}\sqrt{14/3}} \end{aligned}$$

Then, from here, we observe that:

$$\begin{aligned}
\hat{\theta}_1^S &= r\frac{\sigma_y}{\sigma_x} \\
&= -\frac{1}{\sqrt{2}\sqrt{14/3}} \cdot \frac{\sqrt{14/3}}{\sqrt{2}} \\
&= -\frac{1}{2} \\
\hat{\theta}_0^S &= \bar{y} - \hat{\theta}_1^S \bar{x} \\
&= 2 + \frac{1}{2}(0) \\
&= 2
\end{aligned}$$

■

Problem 5.1b. Find $\hat{\theta}^S = \begin{bmatrix} \hat{\theta}_0^S \\ \hat{\theta}_1^S \end{bmatrix}$ using the formula derived in lecture 12: $\hat{\theta}^S = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}$. Explicitly write out the matrix $\mathbb{X}$ for this problem and show all steps. How does it compare to your answer to part a)?

Solution. From the data given, we observe that $\mathbb{X}$ is equal to

$$\mathbb{X} := \begin{bmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{bmatrix}$$

Then, we see that

$$\mathbb{X}^\intercal = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{bmatrix}$$

Also, we have

$$\mathbb{Y} = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

Then, we see the following:

$$\begin{aligned}
\mathbb{X}^\intercal \mathbb{X} &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1(1) + 1(1) + 1(1) & 1(1) + 1(-2) + 1(1) \\ 1(1) + -2(1) + 1(1) & 1(1) + -2(-2) + 1(1) \end{bmatrix} \\
&= \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix} \\
(\mathbb{X}^\intercal \mathbb{X})^{-1} &= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/6 \end{bmatrix}
\end{aligned}$$

And we also observe that

$$\begin{aligned}\mathbb{X}^\intercal\mathbb{Y} &= \begin{bmatrix}1 & 1 & 1\\ 1 & -2 & 1\end{bmatrix}\begin{bmatrix}-1\\ 3\\ 4\end{bmatrix}\\ &= \begin{bmatrix}1(-1)+1(3)+1(4)\\ 1(-1)+-2(3)+1(4)\end{bmatrix}\\ &= \begin{bmatrix}6\\ -3\end{bmatrix}\end{aligned}$$

Putting this all together, we see then that

$$\begin{aligned}\hat{\theta}^S &= (\mathbb{X}^\intercal\mathbb{X})^{-1}\mathbb{X}^\intercal\mathbb{Y}\\ &= \begin{bmatrix}1/3 & 0\\ 0 & 1/6\end{bmatrix}\begin{bmatrix}6\\ -3\end{bmatrix}\\ &= \begin{bmatrix}2\\ -\frac{1}{2}\end{bmatrix}\end{aligned}$$

We observe that the entries in $\hat{\theta}^S = \begin{bmatrix}\hat{\theta}_0^S\\ \hat{\theta}_1^S\end{bmatrix}$ coincides with the values we got for $\hat{\theta}_0^S$ and $\hat{\theta}_1^S$ found in the previous question. ■

> **Problem 5.1c.** Find the MSE for the SLR model above.

Solution. First, note that we have

$$y^S = 2 - \frac{1}{2}x_1$$

So, we observe the following:

$$\begin{aligned}y_1^S &= 2-(1/2)(1) = 3/2\\ y_2^S &= 2-(1/2)(-2) = 3\\ y_3^S &= 2-(1/2)(1) = 3/2\end{aligned}$$

Then, the MSE is:

$$\begin{aligned}\frac{1}{3}\left((-1-(3/2))^2+(3-3)^2+(4-(3/2))^2\right) &= \frac{1}{3}(25/4+0+25/4)\\ &= \frac{1}{3}(25/2)\\ &= \frac{25}{6}\end{aligned}$$

■

> **Problem 5.1d.** Find $\hat{\theta}^O = \begin{bmatrix}\hat{\theta}_0^O\\ \hat{\theta}_1^O\\ \hat{\theta}_2^O\end{bmatrix}$ using the formula derived in lecture 12: $\hat{\theta}^O = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}$. Explicitly write out the matrix $\mathbb{X}$ for this problem and **show all steps**.

Solution. We observe the following:

$$\mathbb{X} = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

So, we have

$$\mathbb{X}^\intercal = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

And we also have

$$\mathbb{Y} = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

With this in mind, we see:

$$\begin{aligned} \mathbb{X}^\intercal \mathbb{X} &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1(1)+1(1)+1(1) & 1(1)+1(-2)+1(1) & 1(-1)+1(0)+1(1) \\ 1(1)+-2(1)+1(1) & 1(1)+-2(-2)+1(1) & 1(-1)+-2(0)+1(1) \\ -1(1)+0(1)+1(1) & -1(1)+0(-2)+1(1) & -1(-1)+0(0)+1(1) \end{bmatrix} \\ &= \begin{bmatrix} 3 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 2 \end{bmatrix} \end{aligned}$$

Then, we see that the inverse is simply

$$(\mathbb{X}^\intercal \mathbb{X})^{-1} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/6 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$$

Furthermore, we observe that

$$\begin{aligned} \mathbb{X}^\intercal \mathbb{Y} &= \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} \\ &= \begin{bmatrix} 1(-1)+1(3)+1(4) \\ 1(-1)+-2(3)+1(4) \\ -1(-1)+0(3)+1(4) \end{bmatrix} \\ &= \begin{bmatrix} 6 \\ -3 \\ 5 \end{bmatrix} \end{aligned}$$

Putting this all together, we get

$$\begin{aligned} (\mathbb{X}^\intercal \mathbb{X})^{-1} \mathbb{X}^\intercal \mathbb{Y} &= \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/6 & 0 \\ 0 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 6 \\ -3 \\ 5 \end{bmatrix} \\ &= \begin{bmatrix} 2 \\ -1/2 \\ 5/2 \end{bmatrix} \end{aligned}$$

■

Problem 5.1e. Show that MSE for the OLS is 0. What is the relationship between $\mathbb{Y}$ and span$(\mathbb{X})$?

Solution. From the previous question, we see then that we have the following:

$$y^O = \theta_0^O + \theta_1^O x_1 + \theta_2^O x_2$$

And we see that $\theta_0^O = 2, \theta_1^O = -1/2, \theta_2^O = 5/2$.
Then, with this in mind, we see the following:

$$\begin{aligned} y_1^O &= 2 - (1/2)(1) + (5/2)(-1) = -1 \\ y_2^O &= 2 - (1/2)(-2) + (5/2)(0) = 3 \\ y_3^O &= 2 - (1/2)(1) + (5/2)(1) = 4 \end{aligned}$$

Then, to find the MSE, we do:

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

Thus, we see that the MSE for the OLS is indeed 0 as desired. Since the MSE is equal to zero, it means then that $\mathbb{Y}$ is in fact in the span of $\mathbb{X}$, and that our model used is perfect. ■

Problem 5.1f. Instead of using $\mathbb{X}_{:,2}$ as a feature in our second model, we decided to transform it and use $\mathbb{X}_{:,2}^2$ instead. That is, the dataset we use is modified as follows:

$\mathbb{Y}$	bias	$\mathbb{X}_{:,1}$	$\mathbb{X}_{:,2}^2$
-1	1	1	$(-1)^2 = 1$
3	1	-2	$0^2 = 0$
4	1	1	$1^2 = 1$

Accordingly, we calculate a single prediction using the new model as specified below:

$$y^{new} = \theta_0^{new} + \theta_1^{new} x_1 + \theta_2^{new} x_2^2$$

Is it possible to find a unique optimal solution in this case? If so, compute $\hat{\theta}^{new}$ and the corresponding value of MSE. If not, explain why this is not possible. Regardless of which way you answer, similar to part d), explicitly write out the matrix $\mathbb{X}_{new}$ for this problem and **show all steps**.

Solution. No, it is not possible. We note that $\hat{\theta}^{new} = (\mathbb{X}^\intercal \mathbb{X})^{-1} \mathbb{X}^\intercal \mathbb{Y}$.
Then, with the given definition, we see that $\mathbb{X}$ will be equal to:

$$\mathbb{X} := \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Then, we note that because $\mathbb{X}$ is not full column rank (note that $-2c_1 + 3c_3$, where c_i is the i^{th} column of $\mathbb{X}$), we note that it is not invertible. Then, $\mathbb{X}^\intercal$ isn't either, and thus we observe that we can't get $(\mathbb{X}^\intercal \mathbb{X})^{-1}$. Thus, we see that $\hat{\theta}^{new}$ doesn't exist. ■