Data 100, Fall 2023

Homework #6

Total Points: 28

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, October 12th at 11:59 PM**. Please read the syllabus for **the grace period policy**. No late submissions beyond the grace period will be accepted. While course staff is happy to help you if you encounter difficulties with submission, we may not be able to respond to last-minute requests for assistance (TAs need to sleep, after all!). **We strongly encourage you to plan to submit your work to Gradescope several hours before the stated deadline.** This way, you will have ample time to reach out to staff for submission support.

This assignment is entirely on paper. Your submission (a single PDF) can be generated as follows:

- You can type your answers. We recommend LaTeX, the math typesetting language. Overleaf is a great tool to type in LaTeX.
- Download this PDF, print it out, and write directly on these pages. If you have a tablet, you may save this PDF and write directly on it.
- Write your answers on a blank sheet of physical or digital paper.
- Note: If you write your answers on physical paper, use a scanning application (e.g., CamScanner, Apple Notes) to generate a PDF.

Important: When submitting on Gradescope, you **must tag pages to each question correctly** (it prompts you to do this after submitting your work). This significantly streamlines the grading process and allows us to release grades more quickly.

Your work will NOT be graded if you do not select pages on Gradescope. We will not be granting regrade requests nor extensions to submissions that don't follow instructions.

If you encounter any difficulties with submission, please don't hesitate to reach out to staff prior to the deadline.

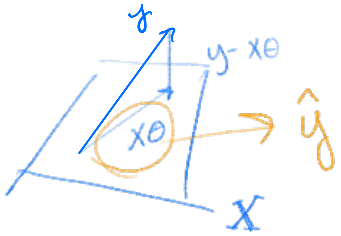
Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others, please include their names at the top of your submission.

Geometric Perspective of Simple Linear Regression

1. (7 points) 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



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

$$X = \begin{bmatrix} 1 & x_1 \end{bmatrix}$$

$$X\vec{\hat{\theta}} =$$

$$\text{or} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

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

   In this problem, $\mathbb{1}_n$ 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.

   (a) (3 points) Recall in the last assignment, we showed that $\sum_{i=1}^{n} e_i = 0$ algebraically. In this question, explain why $\sum_{i=1}^{n} e_i = 0$ using a geometric property. (Hint: $\vec{e} = \mathbb{Y} - \hat{\mathbb{Y}}$, and $\vec{e} = [e_1, e_2, ..., e_n]^\top$.)

1a

① $\sum_{i=1}^{n} e_i$ is the summation of e_i's from 1 to n by definition.

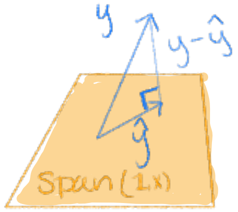
Say we instead think of $\sum_{i=1}^{n} e_i$ as the dot product between the residual vector and $\mathbb{1}_n$ the n-vector of all 1s :

$$\sum_{i=1}^{n} e_i = \mathbb{1} \cdot \vec{e} = [1_1, 1_2 \ldots 1_n] \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1 + e_2 + \ldots + e_n$$

② Recall, we define the residual vector as $\vec{e} = \mathbb{Y} - \hat{\mathbb{Y}}$

Where $\hat{\mathbb{Y}}$ is the predicted outcome, closest to $\mathbb{Y}$ in Span $(\mathbb{X})$.

The difference between true and predicted outcomes, $\hat{\mathbb{Y}} - \mathbb{Y}$, is orthogonal to any vector in span $(\{\mathbb{1}, \mathbb{X})$



③ Therefore, $\mathbb{Y} - \hat{\mathbb{Y}}$ is orthogonal to $\mathbb{1}$, meaning the dot product between the two vectors is 0

$$(\mathbb{Y} - \hat{\mathbb{Y}}) \cdot \mathbb{1} = 0$$

④ Since $e = \mathbb{Y} - \hat{\mathbb{Y}}$,

$$\underbrace{e \cdot \mathbb{1} = 0}$$

$$\downarrow$$

$$\mathbb{1} \cdot \vec{e} = \sum_{i=1}^{n} e_i = 0$$

(b) (2 points) Similarly, explain why $\sum_{i=1}^{n} e_i x_i = 0$ using a geometric property. (Hint: Your answer should be very similar to the above)

(c) (2 points) Briefly explain why the vector $\hat{\mathbb{Y}}$ must also be orthogonal to the residual vector $\vec{e}$.

Remark: Solving the minimum L2 loss solution is equivalent to the geometric perspective.

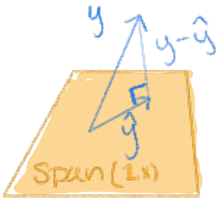
1b

① again, it would be helpful to think of $\sum_{i=1}^{n} e_i x_i = 0$ as

$$\vec{e}\,\mathbb{X} \cdot \mathbb{1} \quad \text{where} \quad \mathbb{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

② As stated in ② for the answer above,

The difference between true and predicted outcomes, $\hat{\mathbb{Y}} - \mathbb{Y}$, is orthogonal to any vector in span($\{\mathbb{1}, \mathbb{X}\}$)



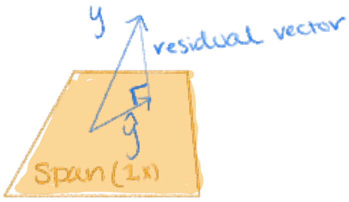
$\mathbb{X}$ is in our span($\{1, \mathbb{X}\}$),

③ Therefore, $\mathbb{Y} - \hat{\mathbb{Y}}$ is orthogonal to $\mathbb{X}$, meaning the dot product between the two vectors is 0

$$\vec{e} \cdot \mathbb{X} = 0$$

1c

The vector $\hat{\mathbb{Y}}$ is the projection of y onto the Span ($\mathbb{X}$), therefore, $\mathbb{Y}$ is, by definition within Span ($\mathbb{X}$). So it follows that the residual vector is orthogonal to it.



Calculus Perspective of Normal Equations

2. (7 points) During Lecture 12, we discussed a geometric argument to get the least squares estimator. Based on the properties of orthogonality, we can obtain the *normal equations* below:

$$\mathbb{X}^\top(\mathbb{Y} - \mathbb{X}\hat{\theta}) = \vec{0}.$$

We can rearrange the equation to solve for θ when $\mathbb{X}$ is full column rank.

$$\hat{\theta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}.$$

Here, we are using $\mathbb{X}$ to denote the design matrix:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} | & | & | & & | \\ \mathbb{1} & \mathbb{X}_{:,1} & \mathbb{X}_{:,2} & \cdots & \mathbb{X}_{:,p} \\ | & | & | & & | \end{bmatrix}$$

where $\mathbb{1}$ is the vector of all 1's of length n and $\mathbb{X}_{:,j}$ is the n-vector $\begin{bmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{bmatrix}$. In other words, it is the j-th feature vector. This notation is more general

To build intuition for these equations and relate them to the SLR estimating equations, we will derive them algebraically using calculus.

(a) (3 points) Show that finding the optimal estimator $\hat{\theta}$ by solving the normal equations is equivalent to requiring that the residual vector $\vec{e} = \mathbb{Y} - \mathbb{X}\hat{\theta}$ should average to zero, and the residual vector $\vec{e}$ should be orthogonal to $\mathbb{X}_{:,j}$ for every j. That is, show that the matrix form of the normal equation can be written as:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

and

$$\mathbb{X}_{:,j}^\top \vec{e} = \sum_i x_{i,j} e_i = 0$$

for all $j = 1, \ldots, p$. (Hint: Expand the normal equation above and perform matrix multiplication for the first few terms. Can you find a pattern?)

1 As given by the normal equation:

$$X^T(y - X\hat{\theta}) = \vec{0}$$

2 Using the definition of the residuals to be the difference between the actual and predicted response values, we can substitute and expand with matrix multiplication:

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1j} \\ 1 & X_{21} & X_{22} & X_{23} & \cdots & X_{2j} \\ \vdots & & & & & \\ 1 & X_{i1} & X_{i2} & X_{i3} & \cdots & X_{ij} \end{bmatrix}$$

$$X^T e = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{1,1} & X_{21} & & X_{i,1} \\ \vdots & & \ddots & \vdots \\ X_{1,j} & \cdots & & X_{i,j} \end{bmatrix} \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} e_1 + e_2 + \cdots + e_n \\ e_1 X_{1,1} + e_2 X_{21} + \cdots + e_n X_{i,1} \\ \vdots \\ e_1 X_{1j} + \cdots + e_n X_{ij} \end{bmatrix} = \vec{0}$$

3 note the pattern where each row is the sum of

$$\begin{bmatrix} \sum_{i=1}^{n} e_i \\ \sum_{i=1}^{n} x_{i,1} e_i \\ \sum_{i=1}^{n} x_{i,2} e_i \\ \vdots \\ \sum_{i=1}^{n} x_{i,j} e_i \end{bmatrix} = \vec{0}$$

4 Because this vector is equal to the zero vector, each row must sum to zero. Thus, it can be further simplified into the equations

From the first row of the resulting matrix:

$$\sum_{i=1}^{n} e_i = 0 \Rightarrow \frac{1}{n} \sum_{i}^{n} e_i = \bar{e} = 0$$

From the subsequent rows:

$$X_{:,j}^T \vec{e} = \sum_{i} x_{ij} e_i = 0$$

(b) (4 points) Remember that the MSE for multiple linear regression is

$$\text{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_{i,1} - \cdots - \theta_p x_{i,p})^2$$

Use calculus to show that any $\theta = [\theta_0, \theta_1, ..., \theta_p]^\top$ that minimizes the MSE must solve the normal equations.

(Hint: Recall that, at a minimum of MSE, the partial derivatives of MSE with respect to every θ_i must all be zero. Find these partial derivatives and compare them to your answer in Question 2a.)

Remark: The two sub-parts above again show that the geometric perspective is equivalent to the calculus approach of solving derivative and setting it to 0 for OLS. This is a desirable property of a linear model with L2 loss, and it generally does not hold for other models and loss types. We hope these exercises clear up some mysteries about geometric derivation!

1) Begin by taking the partial derivatives of the MSE equation w/ respect to each θ_i value for $i \in \mathbb{Z}^+$

$$MSE(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_1 - \theta_0 - \theta_1 x_{i,1} - \theta_2 x_{i,2} - \cdots - \theta_n x_{i,p})^2$$

$$\frac{d}{d\theta_0}[MSE\ \theta] = \frac{2}{n}\sum_{i=1}^{n}(y_1 - \theta_0 - \theta_1 x_{i,1} - \ldots - \theta_n x_{ip})(-1)$$

$$\frac{d}{d\theta_1}[MSE\ \theta] = \frac{2}{n}\sum_{i=1}^{n}(y_1 - \theta_0 - \theta_1 x_{i1} - \ldots - \theta_n x_{ip})(-x_{i,1})$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

2) note the pattern where each derivative of the MSE equation is:

$$\frac{d}{d\theta_n}[MSE\ \theta] = \frac{2}{n}\sum_{i=1}^{n}(y_1 - \theta_0 - \theta_1 x_{i1} - \ldots - \theta_n x_{ip})(-x_{i,p})$$

3) We can rewrite this in matrix form as

$$\frac{d}{d\theta_n}[MSE(\theta)] = \frac{-2}{n}\sum_{i=1}^{n}(x_{i,p})\left(y - \begin{bmatrix}1 & x_{i,1} & x_{i2} & \cdots & x_{ip}\end{bmatrix}\begin{bmatrix}\theta_0 \\ \theta_1 \\ \vdots \\ \theta_n\end{bmatrix}\right)$$

4) and simplify with the previously proved equations:

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

$$= \frac{-2}{n}\sum_{i=1}^{n}(x_{ip})(y_i - \hat{y}_i)$$

$$= \frac{-2}{n}\sum_{i=1}^{n}(x_{ip})e_i$$

$$\frac{d}{d\theta_n}[MSE(\theta)] = \sum_{i=1}^{n}(x_{ip})e_i = 0$$

Thus, because every partial derivative w/ respect to all and any $\theta = [\theta_0\ \theta_1 \cdots \theta_p]^T$ is zero (as proved in 2a) MSE is minimized and θ_i solves the normal equations.

A Special Case of Linear Regression

3. (14 points) In this question, we fit two models:

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

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

(a) (3 points) Find θ_0^S and θ_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. You may find it helpful to keep intermediate steps in the square root (they cancel out nicely at the end!).

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) \qquad \sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(j_i-\bar{j})^2}{n}}$$

① Take the given equations for the correlation coefficient and SD. plug in the given values to solve for variables

$$\bar{y} = \frac{-1+3+4}{3} = 2 \quad , \quad \sigma_y = \sqrt{\frac{(-1-2)^2+(3-2)^2+(4-2)^2}{3}} = \sqrt{\frac{14}{3}}$$

$$\bar{x} = \frac{1+-2+1}{3} = 0 \qquad \sigma_x = \sqrt{\frac{(1-0)^2+(-2-0)^2+(1-0)^2}{3}} = \sqrt{2}$$

$$r = \frac{1}{3}\left[\left(\frac{-1-2}{\sqrt{\frac{14}{3}}}\right)\left(\frac{1-0}{\sqrt{2}}\right) + \left(\frac{3-2}{\sqrt{\frac{14}{3}}}\right)\left(\frac{-2-0}{\sqrt{2}}\right) + \left(\frac{4-2}{\sqrt{\frac{14}{3}}}\right)\left(\frac{1-0}{\sqrt{2}}\right)\right]$$

$$= \frac{1}{3}\left[\frac{-3}{\sqrt{\frac{28}{3}}} + \frac{-2}{\sqrt{\frac{28}{3}}} + \frac{2}{\sqrt{\frac{28}{3}}}\right]$$

$$= \frac{1}{3}\left[\frac{-3}{\sqrt{\frac{28}{3}}}\right]$$

$$r = \frac{-1}{\sqrt{\frac{28}{3}}} = -\sqrt{\frac{3}{28}}$$

② Using the correlation coefficient and other variables we can now Solve for θ_0^s and θ_1^s

$$\hat{\theta}_1^s = r\frac{\sigma_y}{\sigma_x} = -\sqrt{\frac{3}{28}}\ \frac{\sqrt{\frac{14}{3}}}{\sqrt{2}}$$

$$= -\sqrt{\frac{3}{28}} \cdot \sqrt{\frac{7}{3}}$$

$$= -\sqrt{\frac{21}{84}}$$

$$= -\sqrt{\frac{1}{4}}$$

$$\boxed{\hat{\theta}_1^s = -\frac{1}{2}}$$

$$\hat{\theta}_0^s = \bar{y} - \hat{\theta}_1^s\bar{x} = 2 - \left(-\frac{1}{2}\cdot 0\right)$$

$$\boxed{\hat{\theta}_0^s = 2}$$

(b) (2 points) 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)? (Hint: $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1} = \begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix}$)

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

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

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

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

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

$$(X^T X)^{-1} X^T Y = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/6 & -1/3 & 1/6 \end{bmatrix}\begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ -1/2 \end{bmatrix} \qquad \boxed{\hat{\theta}^S = \begin{bmatrix} 2 \\ -1/2 \end{bmatrix}}$$

(c) (1 point) Find the MSE for the SLR model above. (As a sanity check, the sum of residuals should be 0.)

$$R(\theta) = \frac{1}{n}\|y - X\theta\|_2^2$$

$$X\theta = \begin{bmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 2 \\ -1/2 \end{bmatrix} = \begin{bmatrix} 3/2 \\ 3 \\ 3/2 \end{bmatrix}$$

$$y - X\theta = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} 3/2 \\ 3 \\ 3/2 \end{bmatrix} = \begin{bmatrix} -5/2 \\ 0 \\ 5/2 \end{bmatrix}$$

$$R(\theta) = \frac{1}{3}\left(\sqrt{(-5/2)^2 + 0^2 + (5/2)^2}\right)^2 = \frac{1}{3}\left(\sqrt{\frac{25}{2}}\right)^2 = \frac{1}{3}\left(\frac{25}{2}\right) = \boxed{\frac{25}{6}}$$

(d) (3 points) Find $\hat{\theta}^O = \begin{bmatrix} \hat{\theta}^O_0 \\ \hat{\theta}^O_1 \\ \hat{\theta}^O_2 \end{bmatrix}$ 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**. (Hint: The intercept and coefficient of $\mathbb{X}_{:,1}$ for OLS are the same as SLR in this special example. Check the remark at the end of the question to see why this is the case.)

$$\hat{\Theta}^o = (X^T X)^{-1} X^T Y$$

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

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

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

$$(X^T X)^{-1} X^T y = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/6 & -1/3 & 1/6 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1/2 \\ 5/2 \end{bmatrix}$$

$$\boxed{\hat{\Theta}^o = \begin{bmatrix} 2 \\ -1/2 \\ 5/2 \end{bmatrix}}$$

(e) (2 points) Show that MSE for the OLS is 0. What is the relationship between $\mathbb{Y}$ and span($\mathbb{X}$)? (As a sanity check, the sum of residuals should be 0.)

$$R(\Theta) = \frac{1}{n} \| y - X\Theta \|_2^2$$

$$X\Theta = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -1/2 \\ 5/2 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}$$

$$y - X\theta = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$R(\theta) = \frac{1}{3}\sqrt{0^2 + 0^2 + 0^2} = \frac{1}{3}(0) = 0$$

$$\boxed{R(\theta) = 0}$$

There are pivots in every column which indicates y is a linear combination of the vectors in X, and therefore in the span(X).

Span(X) & y:

$$\left[\begin{array}{ccc|c} 1 & 1 & -1 & -1 \\ 1 & -2 & 0 & 3 \\ 1 & 1 & 1 & 4 \end{array}\right] = \left[\begin{array}{ccc|c} \boxed{1} & 1 & -1 & -1 \\ 0 & \boxed{-3} & 1 & 4 \\ 0 & 0 & \boxed{2} & 5 \end{array}\right]$$

(f) (3 points) 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**.

$$X_{:2}^2 = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

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

To put into reduced echelon form:

$$\begin{bmatrix} 3 & 0 & 2 \\ 0 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix} \to \begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 1 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 1 \\ 0 & 3 & 1 \end{bmatrix} \to \begin{bmatrix} 3 & 0 & 2 \\ 0 & 3 & 1 \\ 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & 0 & 2/3 \\ 0 & \boxed{1} & 1/3 \\ 0 & 0 & 0 \end{bmatrix}$$

Because there are only 2 pivot points, $X^T X$ is not full rank and is therefore not invertible nor an optimal unique solution.

Remark: This question intends to give you some practice with SLR and OLS with actual numbers. It is important to note that the coefficients corresponding to the same variable in different linear models are usually not the same. They are only identical in this problem because we have carefully constructed the matrix such that features are orthogonal to each other to simplify the calculations. We will discuss the opposite case, multi-collinearity, in the future. Don't worry if you don't understand it yet!