Data 100, Spring 2023

Homework #6

Due Date: Thursday, March 2nd at 11:59 PM Pacific

Total Points: 36

Submission Instructions

You must submit this assignment to Gradescope by the on-time deadline, **Thursday, March 2nd at 11:59 PM Pacific**. 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:

1.
   - 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.
2. **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 for our readers. Failure to do this may result in a score of 0 for untagged questions.

You are responsible for ensuring your submission follows our requirements. 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.

Answer.

Constant predictions

1. (10 points) One model that is even simpler than the linear model is the *constant model* :

$$\hat{y} = \theta_0$$

We predict exactly the same θ_0 for every observation y_i. We might do this if we had no predictor variables. Or, if our predictor variable were categorical (e.g., gender; or treatment vs. control group), we might make a different prediction for each gender, estimating a constant model within each group.

One benefit of studying the constant model is that it is a simple context in which we can build our intuition for how different loss functions differ from each other. For the following question, assume that we observe $y_1, \ldots, y_n$, and we choose θ_0 to minimize the empirical risk of predicting θ_0 for every single y_i:

$$\hat{R}(\theta_0) = \frac{1}{n} \sum_i L(y, \theta_0)$$

(a) (2 points) If we use the L2 loss:

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Show that the best-fitting estimate is the sample mean; i.e., $\hat{\theta}_0 = \bar{y}$. (Note: After finding the critical point, please show or briefly explain why it is the minimum. You should assume this is always necessary, unless otherwise specified.)

First, we will replace variables according to our model.

$$\begin{aligned}
\hat{R}(\theta_0) &= \frac{1}{n} \sum_i L(y, \theta_0) \\
&= \frac{1}{n} \sum_{i=1}^{n} (y - \hat{y})^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} (y - \theta_0)^2
\end{aligned}$$

Next, we will find the critical value by setting the first derivative of the loss function

equal to 0.

$$\begin{aligned}
0 = \frac{d}{d\theta_0}\hat{R}(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\theta_0}(y-\hat{\theta}_0)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(-2)(y-\hat{\theta}_0) \\
0 &= -\frac{2}{n}\left(\sum_{i=1}^{n}y - \sum_{i=1}^{n}\hat{\theta}_0\right) \\
\sum_{i=1}^{n}\hat{\theta}_0 &= \sum_{i=1}^{n}y \\
n\hat{\theta}_0 &= n\bar{y} \\
\hat{\theta}_0 &= \bar{y}
\end{aligned}$$

Now that we have found that $\bar{y}$ is our critical value, we must show that it is a minimum by calculating the concavity of this point using the second derivative.

$$\begin{aligned}
\frac{d^2}{d\theta_0^2}\hat{R}(\theta_0) &= \frac{d}{d\theta_0}\left(-\frac{2}{n}\left(\sum_{i=1}^{n}y - \sum_{i=1}^{n}\hat{\theta}_0\right)\right) \\
&= -\frac{2}{n}\frac{d}{d\theta_0}\left(\sum_{i=1}^{n}y - \sum_{i=1}^{n}\hat{\theta}_0\right) \\
&= -\frac{2}{n}\left(-\sum_{i=1}^{n}\frac{d}{d\theta_0}\hat{\theta}_0\right) \\
&= \frac{2}{n}\left(\sum_{i=1}^{n}1\right) = \frac{2}{n}(n) = 2
\end{aligned}$$

Since the second derivative at θ_0 is greater than 0, the function is concave up and we have a minimum at this critical value. Therefore, $\bar{y}$ is a minimum of the loss function. □

(b) (2 points) If we use the L1 loss:

$$L(y, \hat{y}) = |y - \hat{y}|$$

Show that the best-fitting estimate is the sample median. To simplify the problem, you may assume that n is odd, so the median is well-defined.

Differentiating the loss function with $L1$ loss and setting it equal to 0 will give us the equation

$$\sum_{\theta_0 < y_i} 1 = \sum_{\theta_0 > y_i} 1.$$

The best parameter that minimizes MAE must satisfy the above equation, which is the median since there's an equal amount of data less than and greater than the median. In addition, notice that the derivative changes from negative to positive at the median.

$$\begin{aligned}
\frac{\partial}{\partial \theta_0} = \hat{R}(\theta_0) &= \frac{\partial}{\partial \theta_0} \frac{1}{n} \sum_i L(y, \hat{y}) \\
&= \frac{\partial}{\partial \theta_0} \frac{1}{n} \sum_i |y - \theta_0| \\
&= \frac{1}{n} \frac{\partial}{\partial \theta_0} \left(\sum_{y > \theta_0} (y - \theta_0) + \sum_{y < \theta_0} (\theta_0 - y) \right) \\
&= \frac{1}{n} \left(\sum_{y > \theta_0} (-1) + \sum_{y < \theta_0} (1) \right)
\end{aligned}$$

Therefore, we have confirmed that the median is a minimum and not a maximum. □

(c) (2 points) Another option is to use what we might call the L0 loss

$$L(y, \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{if } y = \hat{y} \end{cases}$$

Show that the best-fitting estimate is the sample mode. (Written explanation is sufficient, no equations required.)

If we set $\hat{y}$ equal to the mode, that means $\hat{y}$ would obtain the value of the most frequent data point. This also means that there would be less data points that are not equal to the mode, so it would minimize the loss. □

Note: This loss is interesting because it doesn't care the least bit how far y is from $\hat{y}$, only whether y is perfectly predicted or not. This is a natural loss to use if y is a categorical feature with no ordinal structure.

(d) (3 points) Consider a weighted version of the L1 loss:

$$L(y, \hat{y}) = \begin{cases} |y - \hat{y}| & \text{if } y > \hat{y} \\ 0 & \text{if } y = \hat{y} \\ w \cdot |y - \hat{y}| & \text{if } y < \hat{y} \end{cases}$$

where $w > 0$ is a nonzero weight that tells us how much more costly overestimates are vs underestimates.

Show that the optimal choice of $\hat{y} = \theta_0$ is where $\frac{1}{1+w}$ of the data points is below $\hat{\theta}_0$ and $\frac{w}{1+w}$ of the data points is above $\hat{\theta}_0$. (Note: This point is a summary statistic known as the $\frac{1}{1+w}$ percentile.)

$$\begin{aligned} \hat{R}(\theta_0) &= \frac{1}{n} \sum_i L(y, \theta_0) \\ &= \frac{1}{n} \sum_{y>\theta_0} |y - \theta_0| + \frac{1}{n} \sum_{y<\theta_0} w * |y - \theta_0| \\ &= \frac{1}{n} \sum_{y>\theta_0} (y - \theta_0) + \frac{1}{n} \sum_{y<\theta_0} w * (\theta_0 - y) \end{aligned}$$

$$\begin{aligned} 0 = \frac{\partial}{\partial \theta_0} \hat{R}(\theta_0) &= \frac{1}{n} \sum_{y>\theta_0} \frac{\partial}{\partial \theta_0} (y - \theta_0) + \frac{1}{n} \sum_{y<\theta_0} \frac{\partial}{\partial \theta_0} w * (\theta_0 - y) \\ 0 &= -\sum_{y>\theta_0} 1 + \sum_{y<\theta_0} w \\ \sum_{y>\theta_0} 1 &= \sum_{y<\theta_0} w \end{aligned}$$

Let's say there are m values greater than θ_0 and n values less than θ_0. Then the last equation is equivalent to $m = nw$. Observe that there are $m + n = n + nw = n(1 + w)$ total y-values. This means that there are $\frac{n}{n(1+w)} = \frac{1}{1+w}$ values below θ_0 and $\frac{nw}{n(1+w)} = \frac{w}{1+w}$. Therefore, the most optimal θ_0 that minimizes the loss is where $\frac{1}{1+w}$ of the values are below it and $\frac{w}{1+w}$ of the values are above it. Also observe that for the y values less than θ_0, the derivative is -1, while for the y values greater than θ_0, the derivative is $w > 0$. Thus, there is a local extrema at θ_0 and it is in fact a minimizer because it changes from negative to positive as the y values increase. □

Geometric Perspective of Simple Linear Regression

2. (8 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

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

   This means we can write our predicted response vector as $\hat{\mathbb{Y}} = \mathbb{X}\begin{bmatrix} \hat{\theta_0} \\ \hat{\theta_1} \end{bmatrix} = \hat{\theta_0}\mathbb{1}_n + \hat{\theta_1}\vec{x}$.

   In this problem, $\mathbb{1}_n$ is the n-vector of all 1s and $\vec{x}$ refers to the n-length vector $[x_1, x_2, ..., x_n]^\top$. Note, $\vec{x}$ 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: $e = \mathbb{Y} - \hat{\mathbb{Y}}$, and $e = [e_1, e_2, ..., e_n]^\top$.)

   Observe that the variable e represents a the residual vector between $\mathbb{Y}$ and $\hat{\mathbb{Y}}$, two $n \times 1$ matrices. We can also see that $\hat{\mathbb{Y}} \in$ span($\mathbb{X}$) because all vectors of $\mathbb{X}\theta$ are just the two vectors of $\mathbb{X}$ scaled by $\hat{\theta_0}$ and $\hat{\theta_1}$ respectively. By definition, the vectors e_i are orthogonal to the span of $\mathbb{X}$ and $\mathbb{X}^T(e) = \vec{0}$. Notice that the dot product of the first column of $\mathbb{X}$ (i.e the first row of $\mathbb{X}^T$) and e is equal to the sum of all e_i. Since the definition states that $\mathbb{X}^T(e)$ equals the zero vector, the sum of all e_i must equal 0. Therefore, $\sum_{i=1}^{n} e_i = 0$. □

   (b) (3 points) Similarly, show that $\sum_{i=1}^{n} e_i x_i = 0$ using a geometric property. (Hint: Your answer should be very similar to the above)

Following the above proof, we will now perform dot product with the second column of $\mathbb{X}$ (i.e the second row of $\mathbb{X}^T$) and e. This gives us $\sum_{i=1}^{n} x_i e_i$. Recall in part (a) that $\mathbb{X}^T(e) = \vec{0}$, so each element in $\mathbb{X}^T(e)$ must evaluate to 0. Therefore, $\sum_{i=1}^{n} x_i e_i = 0$. □

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

Recall that $e = \mathbb{Y} - \hat{\mathbb{Y}}$ and e is orthogonal to span($\mathbb{X}$). Since $\hat{\mathbb{Y}} \in$ span($\mathbb{X}$) and e is orthogonal to the span of $\mathbb{X}$, e must also be orthogonal to $\hat{\mathbb{Y}}$. □

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

Calculus Perspective of Normal Equations

3. (7 points) In the lecture, we discussed a geometric argument to get the least squares estimator. Based on the orthogonality principle, we can obtain the *normal equations* below:

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

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:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} | & | & | & & | \\ \mathbb{1} & x_1 & x_2 & \cdots & x_p \\ | & | & | & & | \end{bmatrix}$$

where $\mathbb{1}$ is the vector of all 1s of length n and x_j is the n-vector $\begin{bmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{bmatrix}$. In other words, it is the jth feature vector.

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 $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$ should average to zero, and the residual vector e should be orthogonal to X_j for every j. That is, show that the matrix form of normal equation can be written as:

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

and

$$x_j^\top e = \sum_i x_{i,j} e_i = 0$$

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?)

Observe that the residual vector e is represented by $\mathbb{Y} - \mathbb{X}\hat{\theta}$. Then the normal equation becomes $\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\hat{\theta}) = \mathbb{X}^T(e) = 0$, and expanding this, we get

$$\begin{bmatrix} 1 & \ldots & 1 \\ x_{1,1} & \ldots & x_{n,1} \\ \vdots & \ddots & \vdots \\ x_{1,p} & \ldots & x_{n,p} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} e_1 + \ldots + e_n \\ e_1 x_{1,1} + \ldots + e_n x_{n,1} \\ \vdots \\ e_1 x_{1,p} + \ldots + e_n x_{n,p} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

As we can see, the first element $e_1 + \ldots + e_n$ must equal 0, so $\sum_{i=1}^{n} e_i = 0 \Rightarrow \bar{e}\ \frac{1}{n}\sum_{i=1}^{n} = 0$, as desired. Similarly, each of the other elements in the vector must also evaluate to zero, so $\sum_{i=1}^{n} e_i x_{i,j} = 0$ for each row i. This clearly shows that $x_j^T e = \sum_i x_{i,j} e_j = 0$, as desired. □

(b) (4 points) Remember that the (empirical) 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 Q 3a.)

We will find the partial derivative of each θ_i.

$$\begin{aligned}
\frac{\partial}{\partial\theta_0}\text{MSE}(\theta) &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_{i,1} - \cdots - \theta_p x_{i,p}) = 0 \\
&= -\frac{2}{n}\sum_{i=1}^{n}(y_i - x_{i,j}\theta_i) = \sum_{i=1}^{n} e_i = 0 \\
\frac{\partial}{\partial\theta_1}\text{MSE}(\theta) &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_{i,1} - \cdots - \theta_p x_{i,p})x_{i,1} = \sum_{i=1}^{n} e_i x_{i,1} = 0 \\
&\vdots \\
\frac{\partial}{\partial\theta_p}\text{MSE}(\theta) &= -\frac{2}{n}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_{i,1} - \cdots - \theta_p x_{i,p})x_{i,p} = \sum_{i=1}^{n} e_i x_{i,p} = 0
\end{aligned}$$

Observe that each partial derivative that we set equal to zero represents the dot product of each row in the design matrix, which was derived in (3a) from the normal equations. As we can see, the values where all θ_i are equal to 0 (local minimum of MSE) are equivalent to the values derived from the normal equations. Thus, any minimizing θ vector must solve the normal equations. □

Remark: The two subparts above again together 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 true for other models and loss types. We hope these exercises clear up some mysteries about the orthogonality principle!

A Special Case of Linear Regression

4. (12 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 a Ordinary Least Square (OLS) with two features, respectively.

The data are given below:

y	bias	x_1	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 $\hat{\theta}_0^S = \bar{y} - \hat{\theta}_1^S \bar{x}$). Specify which x you are using and show all steps. You may find it helpful to keep intermediate steps in the square root (they cancel out nicely at the end!).

We will be using the x_1 feature. Solving for $\hat{\theta}_1^S$ first, we get

$$\begin{aligned}
\hat{\theta}_1^S = r\frac{\sigma_y}{\sigma_x} &= \frac{\sigma_y}{\sigma_x}\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \\
&= \frac{1}{3\sigma_x^2}\sum(x_i - \bar{x})(y_i - \bar{y}) \\
&= \frac{(1)(-1-2) + (-2)(3-2) + (1)(4-2)}{3\left(\sqrt{\frac{1}{3}\sum(x_1 - \bar{x})^2}\right)^2} \\
&= \frac{-3-2+2}{3\left(\frac{1}{3}((1)^2 + (-2)^2 + (1)^2)\right)} \\
&= -\frac{3}{1+4+1} = -\frac{1}{2}
\end{aligned}$$

Next, solving for $\hat{\theta}_0^S$, we get

$$\hat{\theta}_0^S = \bar{y} - \hat{\theta}_1^S \bar{x} = 2 + \frac{1}{2} * 0 = 2$$

□

(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: $\hat{\theta}^S = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top y$. 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: $\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}^{-1} = \begin{bmatrix} 1/a & 0 \\ 0 & 1/b \end{bmatrix}$)

$$
\begin{aligned}
\hat{\theta}^S &= (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T y \\
&= \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} \\
&= \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ -3 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{6} \end{bmatrix} \begin{bmatrix} 6 \\ -3 \end{bmatrix} = \begin{bmatrix} 2 \\ -\frac{1}{2} \end{bmatrix}
\end{aligned}
$$

(c) (2 points) Find the MSE for the SLR model above. (As a sanity check, sum of residuals should be 0.)

$$
\begin{aligned}
\text{MSE}(\theta) &= \frac{1}{n} \|\mathbb{Y} - \mathbb{X}\theta\|_2^2 \\
&= \frac{1}{3} \left\| \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & -2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ -\frac{1}{2} \end{bmatrix} \right\|^2 \\
&= \frac{1}{3} \left\| \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix} - \begin{bmatrix} 3/2 \\ 3 \\ 3/2 \end{bmatrix} \right\|^2 \\
&= \frac{1}{3} \left\| \begin{bmatrix} -5/2 \\ 0 \\ 5/2 \end{bmatrix} \right\|^2 = \frac{1}{3} \left(\frac{25}{4} + 0 + \frac{25}{4} \right) = \frac{25}{6}
\end{aligned}
$$

□

(d) (2 points) 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: $\hat{\theta}^O = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top y$ Explicitly write out the matrix $\mathbb{X}$ for this problem and show all steps. (Hint: The intercept and coefficient of x_1 for MLR are the same as SLR in this special example. Check remark at the end of the question to see why this is the case.)

$$\begin{aligned}\hat{\theta}^O &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y \\ &= \left(\begin{bmatrix}1 & 1 & 1\\ 1 & -2 & 1\\ -1 & 0 & 1\end{bmatrix}\begin{bmatrix}1 & 1 & -1\\ 1 & -2 & 0\\ 1 & 1 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & 1 & 1\\ 1 & -2 & 1\\ -1 & 0 & 1\end{bmatrix}\begin{bmatrix}-1\\ 3\\ 4\end{bmatrix} \\ &= \begin{bmatrix}3 & 0 & 0\\ 0 & 6 & 0\\ 0 & 0 & 2\end{bmatrix}^{-1}\begin{bmatrix}6\\ -3\\ 5\end{bmatrix} = \begin{bmatrix}\frac{1}{3} & 0 & 0\\ 0 & \frac{1}{6} & 0\\ 0 & 0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}6\\ -3\\ 5\end{bmatrix} = \begin{bmatrix}2\\ -\frac{1}{2}\\ \frac{5}{2}\end{bmatrix}\end{aligned}$$

□

(e) (3 points) Show that MSE for the MLR is 0. What is the relationship between y and span($\mathbb{X}$). (As a sanity check, sum of residuals should be 0.)

$$\begin{aligned}\text{MSE}(\theta) &= \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\theta\|_2^2 \\ &= \frac{1}{3}\left\|\begin{bmatrix}-1\\ 3\\ 4\end{bmatrix} - \begin{bmatrix}1 & 1 & -1\\ 1 & -2 & 0\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}2\\ -\frac{1}{2}\\ \frac{5}{2}\end{bmatrix}\right\|_2^2 \\ &= \frac{1}{3}\left\|\begin{bmatrix}-1\\ 3\\ 4\end{bmatrix} - \begin{bmatrix}-1\\ 3\\ 4\end{bmatrix}\right\|^2 = \frac{1}{3}\left\|\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}\right\|^2 = 0\end{aligned}$$

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-colinearity, in the future. Don't worry if you don't understand it yet!