Data 100, Spring 2023

Homework #6

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:

- 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.

1. **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.

Constant predictions

1. (9 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.)

Incorporating L2 loss function into $\hat{R}(\theta_0)$ leads to: $R(\theta_0) = \frac{1}{n} \sum_i (y_i - \theta_0)^2$

To find the best fitting estimate, we take 1st derivative:

$$\frac{dR}{d\theta_0} = \frac{d}{d\theta_0} \left[\frac{1}{n} \sum_i (y_i - \theta_0)^2 \right] = \frac{1}{n} \left[\sum_i (-2 * (y_i - \theta_0)) \right] = \frac{-2}{n} \sum_i (y_i - \theta_0)$$

At best fitting, $\frac{d\hat{R}}{d\theta_0} = 0$, hence $\frac{-2}{n} \sum_i (y_i - \hat{\theta_0}) = 0$, $\Rightarrow \frac{1}{n} \sum_i y_i - \frac{1}{n} \sum_i \hat{\theta_0} = 0$,

In this case, $\hat{\theta_0}$ is the same for every y_i, hence $\hat{\theta_0} = \frac{1}{n} \sum_i y_i = \bar{y}$.

To prove minimum, we take second derivative:

$$\frac{d^2R}{d\theta_0^2} = \frac{d}{d\theta_0} \left[\frac{-2}{n} \sum_i (y_i - \theta_0) \right] = \frac{+2}{n} > 0,$$ hence it's a concave and the critical point acquired is a minimum.

(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.

Incorporating L1 loss function into $\hat{R}(\theta_0)$ leads to: $R(\theta_0) = \frac{1}{n}\sum_i |y_i - \theta_0|$.

By taking 1st order derivative, $\frac{dR(\theta_0)}{d\theta_0} = \frac{d}{d\theta_0}\left\{\frac{1}{n}\sum_i |y_i - \theta_0|\right\} =$

$$\begin{cases} \frac{d}{d\theta_0}\left\{\frac{1}{n}\sum_i (y_i - \theta_0)\right\}, & \text{if } y_i \geq \theta_0 \\ \frac{d}{d\theta_0}\left\{\frac{1}{n}\sum_i (\theta_0 - y_i)\right\}, & \text{if } y_i < \theta_0 \end{cases} \Rightarrow \begin{cases} -1, & \text{if } y_i \geq \theta_0 \\ 1, & \text{if } y_i < \theta_0 \end{cases} \Rightarrow \frac{dR(\theta_0)}{d\theta_0} = \frac{1}{n}\sum_{y_i \geq \theta_0}(-1) + \frac{1}{n}\sum_{y_i < \theta_0}(1)$$

$= 0$,

Hence $\frac{1}{n}\sum_{y_i \geq \theta_0}(1) = \frac{1}{n}\sum_{y_i < \theta_0}(1)$, there are equal # of points left and right of $\hat{\theta_0}$.

$\hat{\theta_0}$ is median of y_i.

(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 incorporate $\hat{R}_{\theta_0} = \frac{1}{n}\sum_i L(y_i, \theta_0)$ with the L0 loss function,

we have: $\frac{d\hat{R}_{\theta_0}}{d\theta_0}$ that is always 0 at both $y \neq \hat{y}$ and $y = \hat{y}$ conditions.

Hence the best-fitting estimate is the sample mode regardless of the y_i.

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.)

If we incorporate $\hat{R}(\theta_0) = \frac{1}{n}\sum_i L(y_i, \theta_0)$ with the loss function, and do 1st order derivative, we will get:

$$\frac{dR(\theta_0)}{d\theta_0} = \begin{cases} \frac{1}{n}\sum_i \frac{d}{d\theta_0}|y_i - \theta_0|, & \text{if } y_i > \theta_0 \\ 0 & \text{if } y_i = \theta_0 \\ \frac{w}{n}\sum_i \frac{d}{d\theta_0}|y_i - \theta_0|, & \text{if } y_i < \theta_0 \end{cases} \Rightarrow \begin{cases} \frac{-1}{n}\sum_i(1), & \text{if } y_i > \theta_0, \\ 0, & \text{if } y_i = \theta_0, \\ \frac{w}{n}\sum_i(1), & \text{if } y_i < \theta_0, \end{cases}$$

At optimum point, $\frac{dR(\theta_0)}{d\theta_0} = 0$, hence $\frac{-1}{n}\sum_{y_i > \theta_0}(1) + 0_{y_i = \theta_0} + \frac{w}{n}\sum_{y_i < \theta_0}(1) = 0$.

Assuming $\frac{1}{1+w}$ of points below $\hat{\theta}_0$, hence $y < \theta_0 \Rightarrow n \times (\frac{1}{1+w})$.

$\frac{w}{1+w}$ of points above $\hat{\theta}_0$, hence $y > \theta_0 \Rightarrow n \times (\frac{w}{1+w})$.

$$\frac{dR(\theta_0)}{d\theta_0} = \frac{-1}{n} \times n \times \left(\frac{w}{1+w}\right)_{y_i > \hat{\theta}_0} + 0_{y_i = \hat{\theta}_0} + \frac{w}{n} \times n \times \left(\frac{1}{1+w}\right)_{y_i < \hat{\theta}_0}$$

$$= 0.$$

Hence the $\frac{1}{1+w}$ percentile is proved.

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: $\vec{e} = \mathbb{Y} - \hat{\mathbb{Y}}$, and $\vec{e} = [e_1, e_2, ..., e_n]^\top$.)

$1^T e = 0$,

$$\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix},$$

In terms of simple linear regression model,

$$\vec{Y} = \hat{\theta}_0 1_n + \hat{\theta}_1 \vec{x} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \end{bmatrix}$$

By putting in 1^T to the left and right of the equation,

$$[1, 1, \cdots 1] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = [1, 1, \cdots, 1] \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \left([1, 1, \cdots 1] \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \end{bmatrix} + [1, 1, \cdots 1] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} \hat{\theta}_1 \end{bmatrix} \right)$$

$$\sum_i e_i = \sum_i y_i - \left(\hat{\theta}_0 + \sum_i x_i \hat{\theta}_1 \right), \quad \text{By definition, } y_i = \theta_0 + x_i \theta_1,$$

Hence $\sum_i e_i = 0$ is proved.

(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)

$$\begin{bmatrix} e_1 x_1 \\ e_2 x_2 \\ \vdots \\ e_n x_n \end{bmatrix} = \begin{bmatrix} (y_1 - \hat{y}_1) x_1 \\ (y_2 - \hat{y}_2) x_2 \\ \vdots \\ (y_n - \hat{y}_n) x_n \end{bmatrix} = \begin{bmatrix} y_1 x_1 \\ y_2 x_2 \\ \vdots \\ y_n x_n \end{bmatrix} - \begin{bmatrix} \hat{y}_1 x_1 \\ \hat{y}_2 x_2 \\ \vdots \\ \hat{y}_n x_n \end{bmatrix} \Rightarrow [1, 1, \cdots 1] \begin{bmatrix} e_1 x_1 \\ e_2 x_2 \\ \vdots \\ e_n x_n \end{bmatrix} =$$

$$[1, 1, \cdots 1] \begin{bmatrix} y_1 x_1 \\ y_2 x_2 \\ \vdots \\ y_n x_n \end{bmatrix} - [1, 1, \cdots 1] \begin{bmatrix} \hat{y}_1 x_1 \\ \hat{y}_2 x_2 \\ \vdots \\ \hat{y}_n x_n \end{bmatrix} \Rightarrow \sum_i e_i x_i = \sum_i y_i x_i - \sum_i \hat{y}_i x_i = \sum_i y_i x_i - \sum_i (\hat{\theta}_0 + x_i \hat{\theta}_1) x_i$$

$$= \sum_i y_i x_i - \sum_i \hat{\theta}_0 x_i - \sum_i x_i^2 \hat{\theta}_1 = \sum_i x_i (y_i - (\hat{\theta}_0 + x_i \hat{\theta}_1)) = 0.$$

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

$\hat{Y}$ is a linear combination of $\vec{X}$, hence it's on span(X). $\vec{e} = Y - \hat{Y}$, meaning $\vec{e}$ is the vector from $\hat{Y}$ the predicted to Y the true evaluation. As an indiction of distance, $\vec{e}$ must be perpendicular to the $\hat{Y}$ plane. Hence it's orthogonal to the $\hat{Y}$ vector and span(X).

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 properties of orthogonality, we can obtain the *normal equations* below:

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

$X^TY - X^TX\theta = 0$, $X^TY = X^TX\theta$, $\theta = (X^TX)^{-1}X^TY$

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

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

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?)

$X^T(Y - X\theta) \Rightarrow \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & & \vdots \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \times \left(\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ & & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \right) =$

$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & & & \vdots \\ \vdots & & x_{ij} & \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \times \begin{bmatrix} y_1 - (\theta_0 + x_{11}\theta_1 + \cdots x_{1p}\theta_p) \\ y_2 - (\theta_0 + x_{21}\theta_1 + \cdots x_{2p}\theta_p) \\ \vdots \\ y_n - (\theta_0 + x_{n1}\theta_1 + \cdots x_{np}\theta_p) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & & & \\ & \ddots & x_{ij} & \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \times \begin{bmatrix} y_1 - \sum_j x_{1j}\theta_j \\ y_2 - \sum_j x_{2j}\theta_j \\ \vdots \\ y_n - \sum_j x_{nj}\theta_j \end{bmatrix}$

By definition, $e = Y - X\hat{\theta}$, hence the above equation can be re-written as

$$X^T(Y - X\theta) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & & & \\ & \ddots & x_{ij} & \\ x_{1p} & \cdots & & x_{np} \end{bmatrix} \times \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = 0, \Rightarrow \sum_i e_i + \sum_i x_{ij} e_i = 0,$$

$p \times n$ $\quad$ $n \times 1$

Hence $\sum_i e_i = 0$ and $\sum_i x_{ij} e_i = 0$.

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

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

$X^T(Y - X\theta) = 0$

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.)

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

$$= 0,$$

$$\sum_i y_i - n\theta_0 - \theta_1 \sum_i x_{i,1} - \cdots - \theta_p \sum_i x_{i,p} = 0,$$

$$\theta_0 = \frac{1}{n}\sum_i (y_i - \theta_1 x_{i,1} - \cdots \theta_p x_{i,p})$$

And: $\frac{\partial MSE(\theta)}{\partial(\theta_p)} = \frac{-2}{n}\sum_i x_{i,p}(y_i - \theta_0 - \cdots \theta_p x_{i,p}) = 0,$

$$\theta_p = \sum_i x_{i,p}(y_i - \theta_0 - \cdots \theta_p x_{i,p}) \Big/ \sum_i x_{i,p}^2$$

essentially for each θ_p, it can be expressed in the form of $e = Y - X\theta$,

This means $Y - X\theta = 0$. and the normal equation can be solved by $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$ from the MSE partial derivatives.

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 geometric derivation!

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!).

$\bar{y} = (-1+3+4)/3 = 2, \quad \bar{x_1} = (1-2+1)/3 = 0,$

$\sigma_y = \frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{1}{3}\left[(-1-2)^2 + (3-2)^2 + (4-2)^2\right] = \frac{1}{3}\left[9+1+4\right] = 14/3,$

$\sigma_{x_1} = \frac{1}{n}\sum_i (x_{1i} - \bar{x_1})^2 = \frac{1}{3}\left[(1-0)^2 + (-2-0)^2 + (1-0)^2\right] = \frac{1}{3}\left[1+4+1\right] = 2,$

$r = \frac{1}{n}\sum_i \left(\frac{x_{1i}-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right) = \frac{1}{3}\left[\left(\frac{1-0}{2}\times\frac{-1-2}{14/3}\right) + \left(\frac{-2-0}{2}\times\frac{3-2}{14/3}\right) + \left(\frac{1-0}{2}\times\frac{4-2}{14/3}\right)\right]$

$= \frac{1}{3}\left[-\frac{9}{28} - \frac{6}{28} + \frac{6}{28}\right] = -\frac{6}{28}$

Hence $\hat{\theta}_1^S = r\frac{\sigma_y}{\sigma_{x_1}} = -\frac{3}{28}\times\frac{14}{2}\times\frac{1}{2} = -\frac{1}{2}.$

$\hat{\theta}_0^S = \bar{y} - \hat{\theta}_1^S \bar{x} = 2 - (-\frac{1}{4})\times 0 = 2,$

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

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

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

$$= \begin{bmatrix} 1/3 & 0 \\ 0 & 1/6 \end{bmatrix}$$

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

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

The $\hat{\theta}_1$'s acquired here is different from part a.

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

$\hat{y}_1 = 2 \times 1 + (\frac{-1}{2}) = \frac{3}{2}$, $y_1 = -1$, $\Delta y_1 = 5/2$

$\hat{y}_2 = 2 \times 1 + (-2) \times (\frac{-1}{2}) = 3$, $y_2 = 3$, $\Delta y_2 = 0$

$\hat{y}_3 = 2 \times 1 + (\frac{-1}{2}) = \frac{3}{2}$, $y_3 = 4$, $\Delta y_3 = -5/2$,

$MSE = \frac{1}{3} \times (\Delta y_1^2 + \Delta y_2^2 + \Delta y_3^2) = 25/6$.

Sum of residuals is: $\Delta y_1 + \Delta y_2 + \Delta y_3 = 0$.

(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 OLS are the same as SLR in this special example. Check remark at the end of the question to see why this is the case.)

$$X = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} -1 \\ 3 \\ 4 \end{bmatrix}, \quad X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -2 & 1 \\ -1 & 0 & 1 \end{bmatrix},$$

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

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

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

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

$\hat{y}_1 = 2 + (-1/2) \times 1 + (-1) \times 1 = 1/2$, $\Delta y_1 = 3/2$ Sum of residuals

$\hat{y}_2 = 2 + (-1/2) \times (-2) + (0) \times 1 = 3$ $\Delta y_2 = 0$ is 0.

$\hat{y}_3 = 2 + (-1/2) \times 1 + (1) \times 1 = 5/2$ $\Delta y_3 = -3/2$.

$MSE = \frac{1}{3} \times (\Delta y_1^2 + \Delta y_2^2 + \Delta y_3^2) = 9/(2 \times 3) = 3/2$.

y is not part of span(X). Its not orthogonal to span(x) either.

$\hat{y} \in$ span(X). $y - \hat{y}$ is the residual e, $\vec{e}$ is orthogonal to span(X).

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!