Data 100, Summer 2022

Homework #5 Solutions

Total Points: 36

Submission Instructions

You must submit this assignment to Gradescope by **Monday, July 11th at 11:59 PM Pacific**. While Gradescope accepts late submissions, you will not receive **any** credit for a late submission if you do not have prior accommodations (e.g. DSP).

You can work on this assignment in any way you like:

- One way is to download this PDF, print it out, and write directly on these pages (we've provided enough space for you to do so). Alternatively, if you have a tablet, you could save this PDF and write directly on it.
- Another way is to use some form of LaTeX. Overleaf is a great tool; visit the course website for a LaTeX template of this homework.
- You could also write your answers on a blank sheet of paper.

Regardless of what method you choose, the end result needs to end up on Gradescope, as a PDF. If you wrote something on physical paper (like options 1 and 3 above), you will need to use a scanning application (e.g. CamScanner) in order to submit your work.

When submitting on Gradescope, you **must correctly assign pages to each question** (it prompts you to do this after submitting your work). This significantly streamlines the grading process for our tutors. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. If you have any questions about the submission process, please don't hesitate to ask on Piazza.

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.

Properties of Simple Linear Regression

1. (7 points) In lecture, we spent a great deal of time talking about simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a single observation x, our predicted response for this observation is $\hat{y} = \theta_0 + \theta_1 x$. (Note: In this problem we write (θ_0, θ_1) instead of (a, b) to more closely mirror the multiple linear regression model notation.)

   In Lecture 9 we saw that the $\theta_0 = \hat{\theta_0}$ and $\theta_1 = \hat{\theta_1}$ that minimize the average L_2 loss for the simple linear regression model are:

$$\hat{\theta_0} = \bar{y} - \hat{\theta_1}\bar{x}$$
$$\hat{\theta_1} = r\frac{\sigma_y}{\sigma_x}$$

   Or, rearranging terms, our predictions $\hat{y}$ are:

$$\hat{y} = \bar{y} + r\sigma_y \frac{x - \bar{x}}{\sigma_x}$$

   (a) (3 points) As we saw in lecture, a residual e_i is defined to be the difference between a true response y_i and predicted response $\hat{y}_i$. Specifically, $e_i = y_i - \hat{y}_i$. Note that there are n data points, and each data point is denoted by (x_i, y_i).

   Prove, using the equation for $\hat{y}$ above, that $\sum_{i=1}^n e_i = 0$.

   Solution:

$$\begin{aligned}
\sum_{i=1}^n e_i &= \sum_{i=1}^n (y_i - \hat{y}_i) \\
&= \sum_{i=1}^n (y_i - (\bar{y} + r\sigma_y \frac{(x_i - \bar{x})}{\sigma_x}) \\
&= \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y} - r\frac{\sigma_y}{\sigma_x}\sum_{i=1}^n (x_i - \bar{x}) \\
&= n\bar{y} - n\bar{y} - r\frac{\sigma_y}{\sigma_x}[n\bar{x} - n\bar{x}] \\
&= 0
\end{aligned}$$

   (b) (2 points) Using your result from part (a), prove that $\bar{y} = \bar{\hat{y}}$.

Solution:

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= 0 \\
\sum_{i=1}^{n}(y_i - \hat{y}_i) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i &= 0 \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} \hat{y}_i \\
\frac{1}{n}\sum_{i=1}^{n} y_i &= \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \\
\bar{y} &= \bar{\hat{y}}
\end{aligned}$$

(c) (2 points) Prove that $(\bar{x}, \bar{y})$ is on the simple linear regression line.

Solution: Starting with

$$y = \hat{\theta}_0 + \hat{\theta}_1 x$$

we can substitute the definition of $\hat{\theta}_0$ from above as:

$$y = \bar{y} - \hat{\theta}_1 \bar{x} + \hat{\theta}_1 x$$

When we plug $\bar{x}$ in for x, we find the right-hand side becomes

$$\bar{y} - \hat{\theta}_1 \bar{x} + \hat{\theta}_1 \bar{x},$$

which reduces to $\bar{y}$. We see that $(\bar{x}, \bar{y})$ is on the regression line.

Geometric Perspective of Least Squares

2. (5 points) In Lecture 11, 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 $\hat{\theta} = [\hat{\theta}_0, \hat{\theta}_1]^T$, and our design matrix is

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

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

   Note, in this problem, $\vec{x}$ refers to the n-length vector $[x_1, x_2, ..., x_n]^T$. In other words, it 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) Using the geometric properties from lecture, prove that $\sum\limits_{i=1}^{n} e_i = 0$.

   Hint: Recall, we define the residual vector as $e = \mathbb{Y} - \hat{\mathbb{Y}}$, and $e = [e_1, e_2, ..., e_n]^T$.

Solution: We can think of $\sum_{i=1}^{n} e_i$ as the dot product of the residual vector, e and the one vector, $\mathbb{1}$. That is,

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

The predicted $\hat{\mathbb{Y}}$ is the vector closest to $\mathbb{Y}$ in the span($\{\mathbb{1}, \vec{x}\}$). In other words, $\hat{\mathbb{Y}}$ is the projection of $\mathbb{Y}$ into the span, and the difference $\mathbb{Y} - \hat{\mathbb{Y}}$ is orthogonal to any vector in the span($\{\mathbb{1}, \vec{x}\}$). So, $(\mathbb{Y} - \hat{\mathbb{Y}})$ is orthogonal to $\mathbb{1}$. Orthogonality means that the dot product between $\mathbb{Y} - \hat{\mathbb{Y}}$ and any vector in the span is 0. In particular,

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

Since $\vec{e} = \mathbb{Y} - \hat{\mathbb{Y}}$, we have shown that

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

This same argument can be used to establish that $\vec{x} \cdot e = 0$ because $\vec{x}$ is also in the span$(\{\mathbb{1}, \vec{x}\})$.

And, again, since $\hat{\mathbb{Y}}$ is in this span (specifically, $\hat{\mathbb{Y}} = \hat{\theta}_0 \mathbb{1} + \hat{\theta}_1 \vec{x}$) it also follows that $\hat{\mathbb{Y}} \cdot e = 0$. In other words, we have just given explanations for all parts of this problem.

(b) (2 points) Explain why the vectors $\vec{x}$ (as defined in the problem) and $\hat{\mathbb{Y}}$ are both orthogonal to the residual vector e. *Hint: Two vectors are orthogonal if their dot product is 0.*

Solution: In the previous problem, we explained why $\vec{e}$ is orthogonal to any vector in the span$(\{\mathbb{1}, \vec{x}\})$. Therefore it follows that the dot product of x and $\hat{\mathbb{Y}} \in$ span$(\{1, \})$ with e is 0.

Properties of a Linear Model With No Constant Term

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

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

where γ 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 $\hat{\gamma}$ that minimizes the average L_2 loss (mean squared error) across our observed data $\{(x_i, y_i)\}, i = 1, \dots, n$:

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

The normal equations derived in 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, and will also explore whether or not our properties from the previous problem still hold.

3. (4 points) Use calculus to find the minimizing $\hat{\gamma}$. That is, prove that

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

Note: This is the slope of our regression line, analogous to $\hat{\theta}_1$ from our simple linear regression model.

Solution:

As in lecture, the value of γ that minimizes $\frac{1}{n}\sum_{i=1}^{n}(y_i - \gamma x_i)^2$ is the same value that minimizes $\sum_{i=1}^{n}(y_i - \gamma x_i)^2$.

Differentiate the sum of squares with respect to γ to find:

$$-2\sum_{i=1}^{n}(y_i - \gamma x_i)x_1$$

Set the derivative to 0 and solve for the minimizing $\hat{\gamma}$

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(y_i - \hat{\gamma} x_i)x_1 \\ &= \sum_{i=1}^{n} y_i x_i - \hat{\gamma}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

Rearrange terms

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

to find,

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

4. (8 points) For our new simplified model, our design matrix $\mathbb{X}$ is:

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

Therefore our predicted response vector $\hat{\mathbb{Y}}$ can be expressed as $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$. ($\vec{x}$ here is defined the same way it was in Question 2.)

Earlier in this homework, we established several properties that held true for the simple linear regression model that contained an intercept term. For each of the following four properties, state whether or not they still hold true even when there isn't an intercept term. Be sure to justify your answer.

(a) (2 points) $\sum_{i=1}^{n} e_i = 0$.

Solution: This property does not hold anymore. Note that our proof for question 3 requires $\mathbb{1}$ to be one of the columns in our design matrix. Without an intercept term, the feature matrix might not have $\mathbb{1}$ as one of its columns. Hence, we do not know for sure that $\mathbb{1} \cdot e = 0$.

Intuitively speaking, without an intercept term our regression line is forced to pass through the origin, so we can't "position" it in order to guarantee that the residuals sum to 0.

For a concrete counterexample, consider the points $\{(-1,2),(1,3)\}$. Then, $\hat{\gamma} = \frac{-1\cdot 2+1\cdot 3}{(-1)^2+1^2} = \frac{1}{2}$. Then, $\hat{y_1} = \hat{\gamma}x_1 = -\frac{1}{2}$ and $\hat{y_2} = \hat{\gamma}x_2 = \frac{1}{2}$, making the residuals $e_1 = 2 - (-\frac{1}{2}) = \frac{5}{2}$ and $e_2 = 3 - \frac{1}{2} = \frac{5}{2}$; clearly $e_1 + e_2 = 5 \neq 0$.

(b) (2 points) The column vector $\vec{x}$ and the residual vector e are orthogonal.

Solution: This property still holds. We know that $\vec{x}$ is the only column in our design matrix. Since the residuals are orthogonal to the column space of our feature matrix, we know that $\vec{x} \cdot e = 0$.

(c) (2 points) The predicted response vector $\hat{\mathbb{Y}}$ and the residual vector e are orthogonal.

Solution: This property still holds. Note that our design matrix in this scenario is just one column, which is $\vec{x}$. Since the residuals are orthogonal to the column space of our feature matrix, we know that they are orthogonal to anything in the span of $\vec{x}$. Note that $\hat{\mathbb{Y}} = \hat{\gamma}\vec{x}$, which means that $\hat{\mathbb{Y}} \in \text{span}(\{\vec{x}\})$. Hence $\hat{\mathbb{Y}}$ is orthogonal to the residuals, which means that $\hat{\mathbb{Y}} \cdot e = 0$.

(d) (2 points) $(\bar{x}, \bar{y})$ is on the regression line.

Solution: This property does not hold anymore. Note that our derivation in question 1 relied on the value of $\hat{\theta_0}$. However, $\hat{\theta_0}$ does not exist anymore since we don't have an intercept term, which means that $\hat{\gamma}\bar{x}$ is not necessarily equal to $\bar{y}$.

For concreteness, consider the counterexample in the solutions for 4a. There, $\bar{x} = 0$, $\bar{y} = \frac{5}{2}$ and $\hat{\gamma} = \frac{1}{2}$, and so $\hat{\gamma}\bar{x} = 0 \neq \frac{5}{2}$. Thus, in this case, $(\bar{x}, \bar{y})$ is not on the regression line.

MSE "Minimizer"

5. (10 points) 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 average loss on the observed data) will always be the minimizer.

   Given some linear model $f(x) = \gamma x$ for some real scalar γ, we can write the empirical risk of the model f given the observed data $\{x_i, y_i\}, i = 1, \ldots, n$ as the average L_2 loss, also known as mean squared error (MSE):

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

   (a) (3 points) Let's investigate one of the n functions in the summation in the MSE. Define $g_i(\gamma) = \frac{1}{n}(y_i - \gamma x_i)^2$ for $i = 1, \ldots, n$. In this case, note that the MSE can be written as $\sum_{i=1}^{n} g_i(\gamma)$.

   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 is a **convex function**.

Solution: First, we have

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

Then we take the 2nd derivative of $g_i(\gamma)$ with respect to γ.

$$\begin{aligned}\frac{dg_i(\gamma)}{d\gamma} &= \frac{1}{n}(-2y_i x_i + 2x_i^2\gamma)\\ \frac{d^2 g_i(\gamma)}{d\gamma^2} &= \frac{2}{n}x_i^2\end{aligned}$$

Since n is a positive number, $\frac{2}{n} > 0$. In addition, $x_i^2 \geq 0$. Thus, $\frac{2}{n}x_i^2$ is always non-negative, which means g_i is a convex function. In terms of g_i's curvature, we can see that for all γ, the function either faces concave up or is a constant, which occurs when $x_i = 0$.

   (b) (2 points) Briefly explain intuitively in words why given a convex function $g(x)$, the critical point we get by solving $\frac{dg(x)}{dx} = 0$ minimizes g. You can assume that $\frac{dg(x)}{dx}$ is a function of x (and not a constant).

Solution: For a convex function $g(x)$, its 2nd derivative is always non-negative. This means that as x increases, the slope $\frac{dg(x)}{dx}$ is either increasing or staying the same. When $\frac{dg(x)}{dx} = 0$, it's at a point where the slope is turning from negative to 0. We know that when the slope is negative, $g(x)$ is decreasing and when the slope is positive, $g(x)$ increasing. This means the before the function gets to the point where the slope is 0, the function has been decreasing and as it hits this point, the function has stopped decreasing. Hence this point must be the minimum and the function can only stay the same or increase as x increases after getting to this point.

(c) (3 points) Now that we have shown that each term in the summation of the MSE is a convex function, one might wonder if the entire summation is convex given that it is a sum of convex functions.

Let's look at the formal definition of a **convex function**. Algebraically speaking, a function $g(x)$ is convex if for any two points $(x_1, g(x_1))$ and $(x_2, g(x_2))$ on the function,

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$

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

Intuitively, the above definition says that, given the plot of a convex function $g(x)$, if you connect 2 randomly chosen points on the function, the line segment will always lie on or above $g(x)$ (try this with the graph of $y = x^2$).

i. (2 points) Using the definition above, show that if $g(x)$ and $h(x)$ are both convex functions, their sum $g(x) + h(x)$ will also be a convex function.

Solution: By definition, since $g(x)$ and $h(x)$ are both convex, we have that for all $0 \leq c \leq 1$:

$$g(cx_1 + (1-c)x_2) \leq cg(x_1) + (1-c)g(x_2)$$
$$h(cx_1 + (1-c)x_2) \leq ch(x_1) + (1-c)h(x_2)$$

Adding the 2 inequalities and combining the terms, we have:

$$(g+h)(cx_1 + (1-c)x_2) \leq c(g+h)(x_1) + (1-c)(g+h)(x_2)$$

Therefore, by definition, we have shown that $(g+h)(x) = g(x) + h(x)$ is also a convex function.

ii. (1 point) Based on what you have shown in the previous part, explain intuitively why the sum of n convex functions is still a convex function when $n > 2$.

> **Solution:** We can repeatedly apply what we have shown in the previous part on the first 2 convex functions out of all the convex functions, and eventually, we will reduce the sum of any convex functions to a sum of two convex functions, which is a convex function.

(d) (1 point) Finally, explain why in our case that, when we solve for the critical point of the MSE by taking the gradient with respect to the parameter and setting the expression to 0, it is guranteed that the solution we find will minimize the MSE.

> **Solution:** The MSE is a summation of n convex functions, which as we saw in the previous part makes the entire MSE a convex function. Since for a convex function, its critical point is a minimizer, the critical point of the MSE is the minimizer of the function.

Closing note: In this question, we have discussed only the simple linear model with no constant term—a single-variable function. However, the above properties extend more generally to all multivariable linear regression models; this proof is beyond the scope of this course and is left to a future you.

Congratulations! You have finished Homework 5!