

Lab 3: Simple Linear Regression and Partial Derivatives Solutions

EECS 245, Winter 2026 at the University of Michigan

due by the end of your lab section

Name: _____

username: _____

Each lab worksheet will contain several activities, some of which will involve writing code and others that will involve writing math on paper. To receive credit for a lab, you must complete all activities and show your lab TA by the end of the lab section.

While you must get checked off by your lab TA **individually**, we encourage you to form groups with 1-2 other students to complete the activities together.

Recap: Simple Linear Regression

We've spent all of [Chapter 2](#) learning about the simple linear regression model, $h(x_i) = w_0 + w_1x_i$.

To find the optimal intercept, w_0^* , and slope, w_1^* , we minimized mean squared error:

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1x_i))^2$$

- R_{sq} is a function of w_0 and w_1 , and looks like a bowl in 3D. Since it has two input variables, we found its minimum by taking the partial derivatives of $R_{\text{sq}}(w_0, w_1)$ with respect to w_0 and w_1 , setting both of them equal to 0, and then solving for the resulting w_0^* and w_1^* .
- A partial derivative is defined as the derivative with respect to one variable **while treating all others as constants**.

$$f(x, y) = x^2 + 3xy^2 \implies \frac{\partial f}{\partial x} = 2x + 3y^2$$

- An important fact about the line $h^*(x_i) = w_0^* + w_1^*x_i$ is that it is guaranteed to pass through (\bar{x}, \bar{y}) — in other words, an average input always predicts an average output.
- There are several equivalent ways to write the optimal slope, w_1^* . One of them involves the correlation coefficient, r .

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

average product of x and y , once both are standardized

$$w_1^* = r \frac{\sigma_y}{\sigma_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad w_0^* = \bar{y} - w_1^* \bar{x}$$

Activity 1: The Meaning of Mean Squared Error

Suppose we'd like to predict the number of minutes a delivery will take, y , as a function of distance, x . To do so, we look to our dataset of n deliveries, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, and fit two simple linear models:

- $F(x_i) = a_0 + a_1x_i$, where:

$$a_1 = r \frac{\sigma_y}{\sigma_x}, \quad a_0 = \bar{y} - a_1\bar{x}$$

Here, r is the correlation coefficient between x and y , \bar{x} and \bar{y} are their respective means, and σ_x and σ_y are their respective standard deviations.

- $G(x_i) = b_0 + b_1x_i$, where b_0 and b_1 are chosen such that $G(x_i) = b_0 + b_1x_i$ minimizes **mean absolute error** on the dataset. Assume that no other line minimizes mean absolute error on the dataset, i.e. that the values of b_0 and b_1 are unique.

- a) Fill in the

$$\sum_{i=1}^n (y_i - F(x_i))^2 \quad \text{} \quad \sum_{i=1}^n (y_i - G(x_i))^2$$

- $>$
 \geq
 $=$
 \leq
 $<$
 Impossible to tell

Solution: The quantity on the left hand side is the **total squared error** of model F . By definition, F , is the line that minimizes MSE over all possible linear models.

The quantity on the right hand side is the total squared error of model G . However, G is optimized for **mean absolute error**, not MSE.

The question tells us that F and G are different, so $\sum_{i=1}^n (y_i - F(x_i))^2 < \sum_{i=1}^n (y_i - G(x_i))^2$ must be true.

- b) Fill in the

$$\left(\sum_{i=1}^n |y_i - F(x_i)| \right)^2 \quad \text{} \quad \left(\sum_{i=1}^n |y_i - G(x_i)| \right)^2$$

- $>$
 \geq
 $=$
 \leq
 $<$
 Impossible to tell

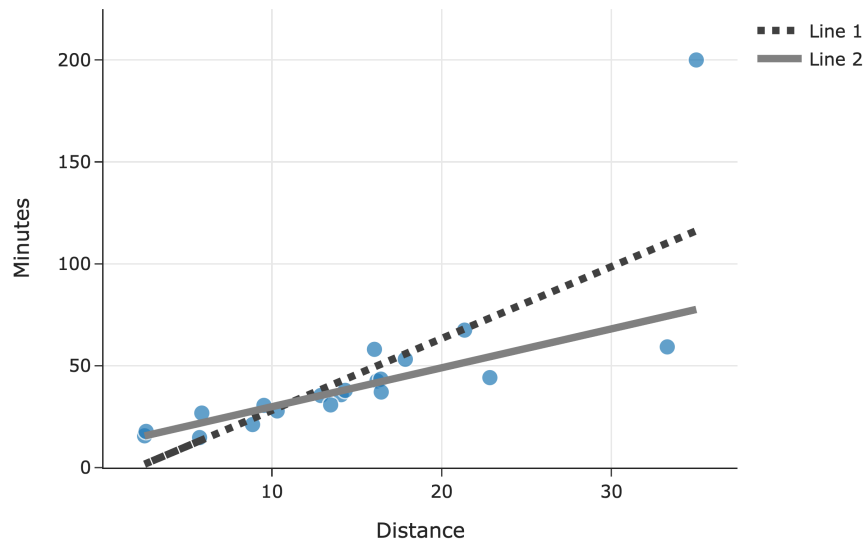
Solution: The quantity on the left hand side is the **squared total absolute error** of model F .

The quantity on the right hand side is the **squared total absolute error** of model G . By definition, G , is the line that minimizes MAE over all possible linear models.

The total absolute error of G must be less than the total absolute error of F , and squaring them doesn't change that relationship because both totals are guaranteed to be positive numbers (sums of absolute values).

Finally, F and G are different, so $\left(\sum_{i=1}^n |y_i - F(x_i)|\right)^2 > \left(\sum_{i=1}^n |y_i - G(x_i)|\right)^2$

c) Below, we've drawn the lines for both F and G along with a scatter plot for the original n deliveries:



Which line corresponds to F ? Line 1 Line 2

Solution: Line 1

The key idea is that models trained with squared loss (MSE) are more sensitive to outliers than models trained with absolute loss (MAE).

Since Line 1 appears to be “pulled up” more strongly by an outlier, it suggests that this line was influenced more heavily by extreme values. This behavior aligns with how MSE-based regression works: outliers have a greater impact on the overall loss because squaring the errors makes large deviations even more significant.

In contrast, MAE-based regression (Line 2) is less sensitive to outliers because absolute differences do not grow as quickly.

Therefore, Line 1 corresponds to F , the MSE-minimizing line.

Activity 2: What Do You Mean?

Suppose we want to fit a simple linear model (using squared loss) that predicts the number of ingredients in a product given its price. We're given that:

- The average cost of a product in our dataset is \$40, i.e. $\bar{x} = 40$
- The average number of ingredients in a product in our dataset is 15, i.e. $\bar{y} = 15$

The intercept and slope of the regression line are $w_0^* = 11$ and $w_1^* = \frac{1}{10}$, respectively.

- a) Suppose Victors' Veil (a skincare product) costs \$40 and has 11 ingredients. What is the squared loss of our model's predicted number of ingredients for Victors' Veil?

Solution: Using the equation of the regression model we have seen in class:

$$h(x_i) = w_0^* + w_1^*x_i$$

Plugging in $w_0^* = 11$, $w_1^* = \frac{1}{10}$, and $x = 40$ gives us:

$$h(x_i) = 11 + \frac{1}{10} \cdot 40 = 15$$

The squared loss is $L = (y_i - h(x_i))^2$, substituting $y = 11$ (actual) and $h(x_i) = 15$ (predicted) gives us:

$$L = (11 - 15)^2 = 16$$

- b) Is it possible to answer part a) above **just** by knowing \bar{x} and \bar{y} , i.e. **without** knowing the values of w_0^* and w_1^* ? Once you select an answer, explain it to your peers.

Yes, it's possible No, it's not possible

Solution: Yes, the values of w_0^* and w_1^* don't impact the answer to part a).

The simple linear model minimizing mean squared error will always go through the point (\bar{x}, \bar{y}) . We're given $\bar{x} = 40$ and $\bar{y} = 15$, meaning that for a product that costs \$40 we will predict that it has 15 ingredients, no matter what the slope and intercept end up being.

Activity 3: Reverse Regression

Suppose we have a dataset of n houses that were recently sold in the Ann Arbor area. For each house, we have its square footage and most recent sale price. The correlation between square footage and price is r .

First, we minimize mean squared error to fit a simple linear model that uses square footage to

predict price. The resulting regression line has an intercept of w_0^* and slope of w_1^* .

$$\text{predicted price}_i = w_0^* + w_1^* \cdot \text{square footage}_i$$

We're now interested in minimizing mean squared error to fit a simple linear model **that uses price to predict square footage** — that is, we're "reversing" the x and y variables. Suppose this new regression line has an intercept of β_0^* and slope of β_1^* .

Find β_1^* . Give your answer in terms of one or more of n, r, w_0^* , and w_1^* .

Solution: Let x represent square footage and y represent price.

We know that $w_1^* = r \frac{\sigma_y}{\sigma_x}$. But what about β_1^* ?

When we take a rule that predicts price from square footage and transform it into a rule that predicts square footage from price, the roles of x and y have swapped; suddenly, square footage is no longer our independent variable, but our dependent variable, and vice versa for price. This means that the altered dataset we work with when using our new prediction rule has σ_x standard deviation for its dependent variable (square footage), and σ_y for its independent variable (price). So, we can write the formula for β_1^* as follows:

$$\beta_1^* = r \frac{\sigma_x}{\sigma_y}$$

In essence, swapping the independent and dependent variables of a dataset changes the slope of the regression line from $r \frac{\sigma_y}{\sigma_x}$ to $r \frac{\sigma_x}{\sigma_y}$. Now, let's simplify to get rid of the σ_x and σ_y :

$$\begin{aligned}\beta_1^* &= r \frac{\sigma_x}{\sigma_y} \\ w_1^* \cdot \beta_1^* &= w_1^* \cdot r \frac{\sigma_x}{\sigma_y} \\ w_1^* \cdot \beta_1^* &= r \frac{\sigma_y}{\sigma_x} \cdot r \frac{\sigma_x}{\sigma_y} \\ w_1^* \cdot \beta_1^* &= r \cdot r \\ \beta_1^* &= \frac{r^2}{w_1^*}\end{aligned}$$

Activity 4: Partial Derivatives and Minimization

Consider the function

$$g(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

- a) Find $\frac{\partial g}{\partial x_1}$ and $\frac{\partial g}{\partial x_2}$, the partial derivatives of g with respect to x_1 and x_2 .

Solution:

$$\begin{aligned}\frac{\partial g}{\partial x_1}g(x_1, x_2) &= \frac{\partial}{\partial x_1} [100(x_2 - x_1^2)^2 + (1 - x_1)^2] \\ &= 200(x_2 - x_1^2) \cdot -2x_1 + 2(1 - x_1) \cdot -1 \\ &= -400x_1(x_2 - x_1^2) - 2(1 - x_1)\end{aligned}$$

$$\begin{aligned}\frac{\partial g}{\partial x_2}g(x_1, x_2) &= \frac{\partial}{\partial x_2} [100(x_2 - x_1^2)^2 + (1 - x_1)^2] \\ &= 200(x_2 - x_1^2)\end{aligned}$$

- b) Find the values of x_1 and x_2 that minimize g . You do not need to use the second derivative test to verify that you've found a minimum. (In fact, "the second derivative test" for functions with multiple input variables is much more complicated, and involves linear algebra.)

Solution: To minimize, set the partial derivatives to 0, then solve the resulting system. The partial derivative with respect to x_2 is a good place to start:

$$\begin{aligned}200(x_2 - x_1^2) &= 0 \\ x_2 - x_1^2 &= 0\end{aligned}$$

We can substitute this into the partial derivative with respect to x_1 :

$$\begin{aligned}-400x_1(x_2 - x_1^2) - 2(1 - x_1) &= 0 \\ -400x_1(0) - 2(1 - x_1) &= 0 \\ -2(1 - x_1) &= 0 \\ x_1 &= 1 \\ x_2 &= 1\end{aligned}$$

So, the values of x_1 and x_2 that minimize g are $x_1 = 1$ and $x_2 = 1$.

Activity 5: Systems of Equations

Next week, we'll start learning about vectors, and various applications of them will involve solv-

ing systems of equations. Here, you'll practice solving systems of equations with three variables.

In each of the following systems of equations, solve for x_1 , x_2 , and x_3 . If you cannot find a unique solution, explain why.

a)

$$\begin{aligned} -4x_1 + 7x_2 - 2x_3 &= 2 \\ x_1 - 2x_2 + x_3 &= 3 \\ 2x_1 - 3x_2 + x_3 &= -4 \end{aligned}$$

Solution: Start by subtracting the second equation from the third to get x_1 in terms of x_2 :

$$\begin{aligned} (2x_1 - 3x_2 + x_3) - (x_1 - 2x_2 + x_3) &= -4 - 3 \\ x_1 - x_2 &= -7 \end{aligned}$$

Substitute $x_1 - x_2 = -7$ into the second equation to get x_3 in terms of x_2 :

$$\begin{aligned} (x_2 - 7) - 2x_2 + x_3 &= 3 \\ -x_2 + x_3 &= 10 \\ x_3 &= x_2 + 10 \end{aligned}$$

Substitute both $x_1 - x_2 = -7$ and $x_3 = x_2 + 10$ into the third equation to solve for x_2 :

$$\begin{aligned} -3(x_2 - 7) + 5x_2 - (x_2 + 10) &= 5 \\ -3x_2 + 21 + 5x_2 - x_2 - 10 &= 5 \\ x_2 + 11 &= 5 \\ x_2 &= -6 \end{aligned}$$

Finally, use x_2 to get both x_1 and x_3 :

$$\begin{aligned} x_1 &= x_2 - 7 = -13 \\ x_3 &= x_2 + 10 = 4 \end{aligned}$$

That leaves us with $x_1 = -13$, $x_2 = -6$, and $x_3 = 4$ as the solution.

b)

$$\begin{aligned} x_1 + 2x_2 - x_3 &= 4 \\ 2x_1 + 4x_2 - 2x_3 &= 8 \\ x_1 - x_2 + 3x_3 &= 1 \end{aligned}$$

Solution: The second equation is just the first equation multiplied by 2, meaning we have 2 different equations and 3 unknowns. In other words, there are infinite solutions for this system, so we cannot uniquely solve it. However, we can still find a relationship between the variables that will satisfy the system:

Use the first equation to express x_1 in terms of x_2 and x_3 :

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 4 \\x_1 &= 4 - 2x_2 + x_3\end{aligned}$$

Substitute this expression for x_1 into the third equation:

$$\begin{aligned}(4 - 2x_2 + x_3) - x_2 + 3x_3 &= 1 \\4 - 3x_2 + 4x_3 &= 1 \\-3x_2 + 4x_3 &= -3\end{aligned}$$

Solve for x_2 in terms of x_3 :

$$\begin{aligned}-3x_2 &= -3 - 4x_3 \\x_2 &= 1 + \frac{4}{3}x_3\end{aligned}$$

Substitute x_2 into the expression for x_1 :

$$\begin{aligned}x_1 &= 4 - 2\left(1 + \frac{4}{3}x_3\right) + x_3 \\&= 2 - \frac{5}{3}x_3\end{aligned}$$

The results tell us any $\boxed{x_1 = 2 - \frac{5}{3}x_3, \quad x_2 = 1 + \frac{4}{3}x_3, \quad x_3}$ will solve the system. To verify this, let's try with some examples:

Let $x_3 = 3$

$$x_1 = 2 - \frac{5}{3}(3) = -3$$

$$x_2 = 1 + \frac{4}{3}(3) = 5$$

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 4 \\(-3) + 2(5) - (3) &= 4 \\-3 + 10 - 3 &= 4 \quad \checkmark\end{aligned}$$

$$\begin{aligned}2x_1 + 4x_2 - 2x_3 &= 8 \\2(-3) + 4(5) - 2(3) &= 8 \\-6 + 20 - 6 &= 8 \quad \checkmark\end{aligned}$$

$$\begin{aligned}x_1 - x_2 + 3x_3 &= 1 \\(-3) - (5) + 3(3) &= 1 \\-8 + 9 &= 1 \quad \checkmark\end{aligned}$$

Here's another example, this time with $x_3 = 6$

$$\text{Let } x_3 = 6$$

$$x_1 = 2 - \frac{5}{3}(6) = -8$$

$$x_2 = 1 + \frac{4}{3}(6) = 9$$

$$\begin{aligned}x_1 + 2x_2 - x_3 &= 4 \\(-8) + 2(9) - (6) &= 4 \\(-8) + 18 - 6 &= 4 \quad \checkmark\end{aligned}$$

$$\begin{aligned}2x_1 + 4x_2 - 2x_3 &= 8 \\2(-8) + 4(9) - 2(6) &= 8 \\-16 + 36 - 12 &= 8 \quad \checkmark\end{aligned}$$

$$\begin{aligned}x_1 - x_2 + 3x_3 &= 1 \\(-8) - (9) + 3(6) &= 1 \\(-8) - 9 + 18 &= 1 \quad \checkmark\end{aligned}$$

The rest of this worksheet is extra practice (taken from past exams that Suraj wrote). Don't feel pressured to answer all of these problems in lab, but make sure to attempt them at some point.

Activity 6: Transformed Data

Suppose we're given a dataset of n points, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, where \bar{x} is the mean of x_1, x_2, \dots, x_n and \bar{y} is the mean of y_1, y_2, \dots, y_n .

Using this dataset, we create a *transformed* dataset of n points, $(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_n, y'_n)$, where:

$$x'_i = 4x_i - 3 \quad y'_i = y_i + 24$$

So the transformed dataset is of the form

$$(4x_1 - 3, y_1 + 24), (4x_2 - 3, y_2 + 24), \dots, (4x_n - 3, y_n + 24)$$

We decide to fit a simple linear model $h(x'_i) = w_0 + w_1 x'_i$ on the transformed dataset using squared loss. We find that $w_0^* = 7$ and $w_1^* = 2$, so $h^*(x'_i) = 7 + 2x'_i$.

- a) Suppose we were to fit a simple linear model through the original dataset, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, again using squared loss. What would the optimal slope on the original dataset be?

2 4 6 8 11 12 24

Solution: 8

Relative to the dataset with x' , the dataset with x is "compressed" by a factor of 4, so the slope increases by a factor of 4, $2 \cdot 4 = 8$. Concretely, this can be shown by looking at the formula for the new slope $2 = r \frac{\sigma_{y'}}{\sigma_{x'}}$.

$$2 = r \frac{\sigma_{y'}}{\sigma_{x'}}$$

$$2 = r \frac{\sigma_y}{4\sigma_x}$$

$$8 = r \frac{\sigma_y}{\sigma_x} = \text{original slope}$$

$\sigma_{x'} = 4\sigma_x$ because the x values have been stretched apart, leading to an increase in their spread. $\sigma_{y'} = \sigma_y$ because the spread of the y values stays the same, they're just shifting upwards by 24.

- b) Recall, the model $h^*(x'_i) = w_0 + w_1 x'_i$ was fit on the transformed dataset, $(x'_1, y'_1), (x'_2, y'_2), \dots, (x'_n, y'_n)$. $h^*(x'_i)$ happens to pass through the point (\bar{x}, \bar{y}) . What is the value of \bar{x} ? Give your answer as an integer with no variables. *Hint: What else does $h^*(x'_i)$ pass through?*

Solution: $h^*(x'_i)$ is guaranteed to pass through (\bar{x}', \bar{y}') , where \bar{x}' is the mean of the x' values and \bar{y}' is the mean of the y' values. Let's see what that looks like as an equation:

$$\begin{aligned}w_0^* + w_1^* \bar{x}' &= h^*(\bar{x}') \\7 + 2\bar{x}' &= h^*(\bar{x}') \\7 + 2\bar{x}' &= \bar{y}'\end{aligned}$$

Our next step is to write \bar{x}' and \bar{y}' in terms of \bar{x} and \bar{y} instead. To do this, we'll prove that any shifting and scaling transformations to the dataset will also apply to the mean.

Given some dataset a_1, a_2, \dots, a_n and nonzero constants b, c , we'll define a'_1, a'_2, \dots, a'_n as a transformed dataset where $a'_i = a_i \cdot b + c$.

$$\begin{aligned}\bar{a}' &= \frac{1}{n} \sum_{i=1}^n a'_i \\&= \frac{1}{n} \sum_{i=1}^n (a_i \cdot b + c) \\&= \frac{1}{n} \sum_{i=1}^n (a_i \cdot b) + \frac{1}{n} \sum_{i=1}^n c \\&= \frac{1}{n} \sum_{i=1}^n (a_i \cdot b) + c \\&= \frac{b}{n} \sum_{i=1}^n a_i + c \\&= \bar{a} \cdot b + c\end{aligned}$$

Back to the original problem, we can substitute $\bar{x}' = 4\bar{x} - 3$ and $\bar{y}' = \bar{y} + 24$ into our equation:

$$7 + 2(4\bar{x} - 3) = \bar{y} + 24$$

The problem tells us that $h^*(x'_i)$ happens to pass through (\bar{x}, \bar{y}) as well, meaning $2\bar{x} + 7 = \bar{y}$ is also true. Now we have a system of two equations, so we can eliminate \bar{y} to solve for \bar{x} :

$$\begin{aligned}7 + 2(4\bar{x} - 3) &= \bar{y} + 24 \\2\bar{x} + 7 &= \bar{y}\end{aligned}$$

$$\begin{aligned}7 + 2(4\bar{x} - 3) - (2\bar{x} + 7) &= \bar{y} + 24 - \bar{y} \\7 + 8\bar{x} - 6 - 2\bar{x} - 7 &= 24 \\6\bar{x} &= 30 \\ \bar{x} &= 5\end{aligned}$$

Activity 7: A Refresher

Consider a dataset of y_1, y_2, \dots, y_n , all of which are **positive**. We want to fit a constant model, $h(x_i) = w$, to the data.

Let w_p^* be the optimal constant prediction that minimizes average degree- p loss, $R_p(w)$, defined below:

$$R_p(w) = \frac{1}{n} \sum_{i=1}^n |y_i - w|^p$$

For example, w_2^* is the optimal constant prediction that minimizes $R_2(w) = \frac{1}{n} \sum_{i=1}^n |y_i - w|^2$

a) In each of the parts below, determine the value of the quantity provided. By “the data”, we are referring to y_1, y_2, \dots, y_n . The answer choices are as follows; **select one item in each row**.

- A: The standard deviation of the data
- B: The variance of the data
- C: The mean of the data
- D: The median of the data
- E: The midrange of the data, $\frac{y_{\min} + y_{\max}}{2}$
- F: The mode of the data
- G: None of these

		A	B	C	D	E	F	G
(i)	h_0^*	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
(ii)	h_1^*	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
(iii)	$R_1(h_1^*)$	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
(iv)	h_2^*	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
(v)	$R_2(h_2^*)$	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Solution:

- (i) h_0^* is none of these. The original intention was to have R_0 be 0-1 loss, in which case h_0^* would be the mode.
- (ii) h_1^* is the median of the data, since $R_1(w) = \frac{1}{n} \sum_{i=1}^n |y_i - w|$
- (iii) $R_1(h_1^*)$ is the minimum mean absolute error, which is none of these.
- (iv) h_2^* is the mean of the data, since $R_2(w) = \frac{1}{n} \sum_{i=1}^n |y_i - w|^2$ is equivalent to mean squared error.
- (v) $R_2(h_2^*)$ is the variance of the data, or the minimum mean absolute error, shown below:

$$\begin{aligned} R_2(h_2^*) &= \frac{1}{n} \sum_{i=1}^n |y_i - h_2^*|^2 \\ &= \frac{1}{n} \sum_{i=1}^n |y_i - \bar{y}|^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \sigma_y^2 \end{aligned}$$

- b) Now, suppose we want to find the optimal constant prediction, h_U^* , using the "Ultra" loss function, defined below:

$$L_U(y_i, w) = y_i(y_i - w)^2$$

To find h_U^* , we minimize $R_U(w)$, the average Ultra loss. How does h_U^* compare to the mean of the data, M ?

- $h_U^* > M$ $h_U^* \geq M$ $h_U^* = M$ $h_U^* \leq M$ $h_U^* < M$

Solution: Minimizing the average Ultra loss means minimizing the empirical risk:

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

This resembles minimizing mean squared error, except each y_i is given a weight of y_i . All the y_i values are positive, so larger y_i values contribute more to the loss. To reduce their impact of these large y_i values, the minimizer gets pulled higher, causing it to be greater than the mean.

- c) Finally, to find the optimal constant prediction, we will instead minimize **regularized** average Ultra loss, $R_\lambda(w)$, defined below:

$$R_\lambda(w) = \left(\frac{1}{n} \sum_{i=1}^n y_i (y_i - w)^2 \right) + \lambda w^2$$

Here, assume $\lambda > 0$ is some positive constant. (We will cover regularization in more detail later in the term.)

Find w^* , the constant prediction that minimizes $R_\lambda(w)$. Give your answer as an expression in terms of the y_i 's, n , and/or λ .

Solution: To minimize the regularized average Ultra loss, we solve for w by setting $\frac{\partial R}{\partial w} = 0$ and solving for w .

Step 1: Compute the derivative and set to 0.

$$\begin{aligned}\frac{\partial R}{\partial w} R_\lambda(w) &= \frac{\partial}{\partial w} \left[\left(\frac{1}{n} \sum_{i=1}^n y_i (y_i - w)^2 \right) + \lambda w^2 \right] \\ &= -2 \left(\frac{1}{n} \sum_{i=1}^n y_i (y_i - w) \right) + 2\lambda w = 0\end{aligned}$$

Step 2: Expand and simplify.

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

Step 3: Solve for w .

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n y_i^2 &= \frac{1}{n} \sum_{i=1}^n y_i w + \lambda w \\ \frac{1}{n} \sum_{i=1}^n y_i^2 &= w \left(\frac{1}{n} \sum_{i=1}^n y_i + \lambda \right) \\ \frac{\frac{1}{n} \sum_{i=1}^n y_i^2}{\frac{1}{n} \sum_{i=1}^n y_i + \lambda} &= w\end{aligned}$$

Step 4: Multiply by $\frac{n}{n}$

$$w^* = \frac{\sum_{i=1}^n y_i^2}{\sum_{i=1}^n y_i + \lambda}$$