Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59PM on the due date. You can use a slip day to extend the deadline by 24 hours; you have four slip days to use in total throughout the quarter.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it. We encourage you type your solutions in LaTeX, using the Overleaf template on the course website.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of 50 points. The point value of each problem or sub-problem is indicated by the number of avocados shown.

- Please remember **assign pages to questions** when you upload your submission to Gradescope. This really helps our graders.

- For all of Problem 5, you'll need to code your answers in Python. More detailed instructions are provided in Problem 5. Note that to submit the homework, you'll have to submit your answers PDF to the Homework 4 assignment on Gradescope, and submit your completed notebook `hw04-code.ipynb` to the Homework 4, Problem 5 autograder on Gradescope.

Problem 1. Reflection and Feedback Form

🥑🥑🥑 Make sure to fill out this Reflection and Feedback Form, linked here for three points on this homework! This form is primarily for your benefit; research shows that reflecting and summarizing knowledge helps you understand and remember it.

Problem 2. Same, but Different

In Lecture 5, we were introduced to one of many formulas for the optimal slope, w_1^*, and optimal intercept, w_0^*, for the simple linear regression model $H(x) = w_0 + w_1 x$ when using squared loss:

$$w_1^* = r \frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^* \bar{x}$$

Then, in Lecture 8, we revisited the simple linear regression model in terms of linear algebra. When $X \in \mathbb{R}^{n \times 2}$ is the design matrix, $\vec{y} \in \mathbb{R}^n$ is the observation vector, and $\vec{w} \in \mathbb{R}^2$ is the parameter vector, we found that the optimal parameter vector $\vec{w}^*$ is one that satisfies the normal equations:

$$X^T X \vec{w} = X^T \vec{y}$$

When $X^T X$ is invertible, $\vec{w}^*$ can be expressed as:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

In this problem, we will prove that both of these formulations are equivalent, for any dataset (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Specifically, we'll show that the vector $\vec{w}^* = (X^T X)^{-1} X^T y$ has two components, the first of which is w_0^* and the second of which is w_1^*. (To do this, we'll need to assume that $(X^T X)^{-1}$ is invertible.)

Note that on first glance, it looks like this problem is quite long, since it has seven subparts. However, the subparts are meant to guide you through the proof. (The problem would take much longer if we just said "prove it!")

a) Express the vector $X^T \vec{y}$ using constants and/or summations involving x_i and/or y_i. Make sure that your answer has the correct dimensions.

b) Express the matrix $X^T X$ using constants and/or summations involving x_i and/or y_i. Make sure that your answer has the correct dimensions.

c) Recall, if $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is a 2×2 matrix, then the inverse of M is given by:

$$M^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Express the matrix $(X^T X)^{-1}$ using constants and/or summations involving x_i and/or y_i.

d) At this point, the expressions you have for $(X^T X)^{-1}$ and $X^T y$ likely involve many summation notations and look... complicated. Let's take a step back and simplify things before we proceed.

Prove that:

$$\sum_{i=1}^{n} x_i^2 = n\sigma_x^2 + n\bar{x}^2$$

Hint: Start with the definition of σ_x^2 and expand the sum.

e) Using your work in (c) and (d), prove that:

$$(X^T X)^{-1} = \frac{1}{n\sigma_x^2} \begin{bmatrix} \sigma_x^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

f) $(X^T X)^{-1}$ is about as simplified as it can be for now. But, before we multiply $(X^T X)^{-1}$ and $X^T y$, we should deal with the fact that at least one of the components in the vector $X^T \vec{y}$ still involves a summation.

Prove that:

$$\sum_{i=1}^{n} x_i y_i = n r \sigma_x \sigma_y + n \bar{x} \bar{y}$$

g) Now, put it all together. That is, prove that:

$$(X^T X)^{-1} X^T \vec{y} = \begin{bmatrix} \bar{y} - r \frac{\sigma_y}{\sigma_x} \bar{x} \\ r \frac{\sigma_y}{\sigma_x} \end{bmatrix}$$

Note that the second component of the vector above is $w_1^* = r\frac{\sigma_y}{\sigma_x}$ and the first component of the vector above is $w_0^* = \bar{y} - r\frac{\sigma_y}{\sigma_x}\bar{x} = \bar{y} - w_1^*\bar{x}$, as we first saw in Lecture 5! This concludes our proof that both formulations of the optimal parameters of the simple linear regression model are equivalent.

Problem 3. Sums of Residuals

In this problem, we will prove that the sum of the residuals of a fit regression model is 0.

We define the ith **residual** to be the difference between the actual and predicted values for individual i in our dataset, when the predictions are made using a regression model whose coefficients w_0^* and w_1^* (or, for multiple linear regression models, w_0^*, w_1^*, w_2^*, ..., w_d^*) are all optimal. In other words, the ith residual e_i is:

$$e_i = y_i - H^*(x_i)$$

We use the letter e for residuals because residuals are also known as errors.

We'll get to the proof soon, but first, a warmup.

a) 🥑🥑 Suppose $\vec{1} \in \mathbb{R}^n$ is a vector containing the value 1 for each element, i.e. $\vec{1} = \begin{bmatrix} 1 \\ 1 \\ ... \\ 1 \end{bmatrix}$.

For any other vector $\vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ ... \\ b_n \end{bmatrix}$, what is the value of $\vec{1}^T\vec{b}$, i.e. what is the dot product of $\vec{1}$ and $\vec{b}$?

b) 🥑🥑🥑 Back to the main problem at hand.

Consider the typical multiple regression scenario where our hypothesis function has an intercept term:

$$H(\vec{x}) = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + ... + w_d x^{(d)}$$

Note that another way of writing the ith residual, $e_i = y_i - H^*(x_i)$, is:

$$e_i = (\vec{y} - X\vec{w}^*)_i$$

Here, X is a $n \times (d+1)$ design matrix, $\vec{y} \in \mathbb{R}^n$ is an observation vector, and $\vec{w} \in \mathbb{R}^{(d+1)}$ is the parameter vector. We'll use $\vec{w}^*$ to denote the optimal parameter vector, or the one that satisfies the normal equations. $(\vec{y} - X\vec{w}^*)_i$ is referring to element i of the vector $\vec{y} - X\vec{w}^*$.

Using facts about $\vec{w}^*$ we learned in Lectures 7 and 8, prove that for multiple linear regression models with an intercept term, the sum of the residuals is 0. That is, prove that:

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

Hint: Refer to the derivation of $\vec{w}^*$ in Lectures 7 and 8. How did we define X? Your proof should not be very long.

c) 🥑🥑🥑🥑 Now suppose our hypothesis function does not have an intercept term, but is otherwise linear with multiple features:

$$H(\vec{x}) = w_1 x^{(1)} + w_2 x^{(2)} + ... + w_d x^{(d)}$$

1. Is it still guaranteed that $\sum_{i=1}^{n} e_i = 0$? Why or why not?

2. Is it still possible that $\sum_{i=1}^{n} e_i = 0$? If you believe the answer is yes, come up with a simple example where a linear hypothesis function without an intercept has residuals that sum to 0. If you believe the answer is no, state why not.

Problem 4. Real Estate

You are given a data set containing information on recently sold houses in San Diego, including

- square footage
- number of bedrooms
- number of bathrooms
- year the house was built
- asking price, or how much the house was originally listed for, before negotiations
- sale price, or how much the house actually sold for, after negotiations

The table below shows the first few rows of the data set. Note that since you don't have the full data set, you cannot answer the questions that follow based on calculations; you must answer conceptually.

House	Square Feet	Bedrooms	Bathrooms	Year	Asking Price	Sale Price
1	1247	3	3	2005	500,000	494,000
2	1670	3	2	1927	1,000,000	985,000
3	716	1	1	1993	335,000	333, 850
4	1600	4	2	1962	830,000	815,000
5	2635	4	3	1993	1,250,000	1,250,000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

a) 🥑🥑🥑 First, suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables. Which feature would you expect to have the largest magnitude weight? Why? (Note that the weight of a feature is the value of w^* for that feature.)

Then, suppose we standardize each variable separately, i.e. we convert each variable to standard units. (Recall, to convert a variable to standard units, we replace each value x_i with $\frac{x_i - \bar{x}}{\sigma_x}$.) Suppose we fit another multiple linear regression model to predict the sale price of a house given all five of the other standardized variables. Now, which feature would you expect to have the largest magnitude weight? Why?

b) 🥑🥑🥑 Suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables in their original, unstandardized form. Suppose the weight for the Year feature is α.

Now, suppose we replace Year with a new feature, Age, which is 0 if the house was built in 2024, 1 if the house was built in 2023, 2 if the house was built in 2022, and so on. If we fit a new multiple linear regression model on all five variables, but using Age instead of Year, what will the weight for the Age feature be, in terms of α?

c) 🥑🥑 Now, suppose we fit a multiple linear regression model to predict the sale price of a house given all five of the other variables, plus a new sixth variable named Rooms, which is the total number of bedrooms and bathrooms in the house. Will our new regression model with an added sixth feature make better predictions than the models we fit in (a) or (b)?

Problem 5. Billy the Waiter

This problem is formatted slightly differently. The entire problem is contained in a supplemental Jupyter Notebook, which you can access **at this link**. This problem is entirely autograded; once you've finished, make sure to submit your `hw04-code.ipynb` notebook to the Homework 4, Problem 5 autograder on Gradescope.

Note that this problem is worth a total of 14 points, split across 6 parts.