Lab 8: Multiple Linear Regression; The Gradient Vector Solutions

EECS 245, Spring 2026 at the University of Michigan

due for completion at 11:59PM Ann Arbor Time on Wednesday, June 3rd, 2026

Name: _____

uniqname: _____

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 as many of the activities as you can in 2 hours and submit a PDF of your work to Gradescope. We will provide specific instructions on how to submit programming activities (e.g. submitting the notebook or including a screenshot of some output).

Recap: Multiple Linear Regression (Chapter 7.2)

Suppose we have n data points, $(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots (\vec{x}_n, y_n)$, where each $\vec{x}_i$ is a feature vector of d features:

$$\vec{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(d)} \end{bmatrix} \qquad \text{for example, } \vec{x}_i = \begin{bmatrix} \text{height}_i \\ \text{weight}_i \\ \text{shoe size}_i \\ \text{height}_i^2 \\ \cos\left(\text{height}_i \cdot \text{weight}_i\right) \end{bmatrix}$$

This data is stored in the $n \times (d+1)$ matrix X, called the **design matrix**, and observation vector $\vec{y}$.

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \ldots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \ldots & x_2^{(d)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \ldots & x_n^{(d)} \end{bmatrix} = \begin{bmatrix} \text{Aug}(\vec{x}_1)^T \\ \text{Aug}(\vec{x}_2)^T \\ \vdots \\ \text{Aug}(\vec{x}_n)^T \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Our goal is to find the optimal parameter vector, $\vec{w}^*$, which minimizes the mean squared error of our model's predictions on the training data.

$$\text{mean squared error} = R_{\text{sq}}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$$

The optimal $\vec{w}^*$ satisfies the normal equation, $X^T X \vec{w} = X^T \vec{y}$. To make predictions:

- $\vec{p} = X\vec{w}^*$ is a vector containing the prediction for all n observations.
- $h(\vec{x}_i) = \vec{w}^* \cdot \text{Aug}(\vec{x}_i)$ is the prediction for any one observation $\vec{x}_i$.

Activity 1: Multiple Linear Regression

Let X be a **full rank** $n \times 3$ design matrix and $\vec{y} \in \mathbb{R}^n$ be an observation vector. Suppose we have already fit a multiple linear regression model of the form

$$h(\vec{x}_i) = w_0 + w_1 x_i^{(1)} + w_2 x_i^{(2)}$$

Now, suppose we add the feature $(x_i^{(1)} + x_i^{(2)})$ to our design matrix and train a new model.

a) Which of the following are true about the new $n \times 4$ design matrix X_{new} with our added feature? Select all that apply.

- ☐ The columns of X_{new} are linearly independent
- ■ The columns of X_{new} are linearly dependent
- ☐ $\vec{y}$ is orthogonal to all the columns of X_{new}
- ☐ $\vec{y}$ is orthogonal to all the columns of the original design matrix X
- ■ colsp(X) = colsp(X_{new})
- ☐ $X_{\text{new}}^T X_{\text{new}}$ is invertible
- ■ $X_{\text{new}}^T X_{\text{new}}$ is not invertible

Solution: Let's look at the correct options:

- The columns of X_{new} are linearly dependent because the new added column, consisting of values of the form $x_i^{(1)} + x_i^{(2)}$, is a linear combination of columns 1 and 2 of X. By definition, this means the columns of X_{new} are linearly dependent.
- The new added column does not change the set of possible linear combinations of the columns of X, since this added column was already in colsp(X). Therefore, colsp(X) = colsp(X_{new}).
- $X_{\text{new}}^T X_{\text{new}}$ is not a full rank matrix because X_{new}'s columns aren't linearly independent, meaning rank$(X_{\text{new}}) < 4$, and rank$(X_{\text{new}}^T X_{\text{new}})$ = rank(X_{new}), so rank$(X_{\text{new}}^T X_{\text{new}}) < 4$, meaning $X_{\text{new}}^T X_{\text{new}}$ is not invertible.

Note that $\vec{y}$ has no orthogonality relationship to the columns of X or X_{new}. Instead, it's the case that the error vector, $\vec{e} = \vec{y} - X\vec{w}^*$, is orthogonal to the columns of both X and X_{new}.

b) Find a basis for nullsp(X_{new}). (This should be quick!)

Solution: Let's look at X_{new}:

$$X_{\text{new}} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x}^{(1)} & \vec{x}^{(2)} & \vec{x}^{(1)} + \vec{x}^{(2)} \\ | & | & | & | \end{bmatrix}$$

By construction, the fourth column of X_{new} is the sum of $\vec{x}^{(1)}$ and $\vec{x}^{(2)}$. So, multiplying X_{new} by $\begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}$ — or any scalar multiple of it — will give the zero vector!

$$X_{\text{new}} \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x}^{(1)} & \vec{x}^{(2)} & \vec{x}^{(1)} + \vec{x}^{(2)} \\ | & | & | & | \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix} = \vec{x}^{(1)} + \vec{x}^{(2)} - (\vec{x}^{(1)} + \vec{x}^{(2)}) = \vec{0}$$

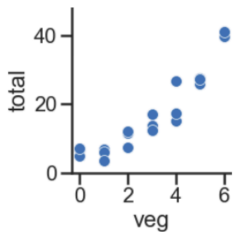
Since $\text{rank}(X) = 3$ (because we were told the original design matrix X was full rank), $\text{rank}(X_{\text{new}}) = 3$ also, meaning $\text{nullsp}(X_{\text{new}})$ is 1-dimensional (from the rank-nullity theorem). So, this vector we've found is a basis for $\text{nullsp}(X_{\text{new}})$.

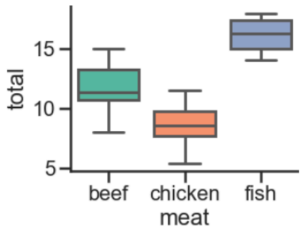
$$\boxed{\text{nullsp}(X_{\text{new}}) = \text{span}\left(\left\{\begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}\right\}\right)}$$

Activity 2: Chicken, Beef, or Fish?

Every week, Lauren goes to her local grocery store and buys exactly one pound of meat (either beef, fish, or chicken) but varying amounts of vegetables. We've collected a dataset containing the pounds of vegetables bought, the type of meat bought, and the total bill. Below we display the first few rows of the dataset and two plots generated using the entire (training) dataset.

veg	meat	total
1	beef	13
3	fish	19
2	beef	16
0	chicken	9





In each part below, we provide you with a model that predicts `total` (her total grocery bill), fit to the dataset by minimizing mean squared error. For each model, determine whether **each optimal parameter w^* is positive, negative or exactly 0**. For example, in part (iv), you'll need to provide 3 answers: one for w_0^*, one for w_1^*, and one for w_2^*.

(i) $h(\vec{x}_i) = w_0$

(ii) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{veg}_i$

(iii) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=chicken}_i$ (one hot encoded feature for chicken)

(iv) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=beef}_i + w_2 \cdot \text{meat=chicken}_i$

(v) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=beef}_i + w_2 \cdot \text{meat=chicken}_i + w_3 \cdot \text{meat=fish}_i$

Solution:

(i) $h(\vec{x}_i) = w_0$

This is the constant model. Since we're minimizing mean squared error, w_0^* is the mean of all total bills in the dataset, which we can tell from the scatter plot is positive.

$$w_0^* = \boxed{\text{positive}}$$

(ii) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{veg}_i$

w_0 is the intercept and w_1 corresponds to pounds of vegetables. As vegetable purchases increase, the total bill increases, so $w_1 > 0$. The intercept looks positive as well, though this is a little less clear, admittedly.

$$w_0^* = \boxed{\text{positive}}, \quad w_1^* = \boxed{\text{positive}}$$

(iii) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=chicken}_i$

Let's think in terms of two cases: Lauren buys chicken, and Lauren doesn't buy chicken.

- Lauren buys chicken: $h(\vec{x}_i) = w_0 + w_1$
- Lauren doesn't buy chicken: $h(\vec{x}_i) = w_0$

Since we're picking w_0^* and w_1^* so that they minimize mean squared error, $w_0^* + w_1^*$ should be the average total bill for purchases involving chicken, and w_0^* should be the average total bill for purchases not involving chicken. Meaning,

$$w_1^* = \text{mean(chicken)} - \text{mean(no chicken)}$$

Both averages are positive, so w_0^* is positive. But, purchases involving chicken tend to be cheaper than purchases not involving chicken (as is evident in the third box plot), so w_1^* is negative.

$$w_0^* = \boxed{\text{positive}}, \quad w_1^* = \boxed{\text{negative}}$$

(iv) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=beef}_i + w_2 \cdot \text{meat=chicken}_i$

Following similar logic to the previous part, w_0 is the mean total for the reference group (fish), w_1 is the difference between the mean of beef and the mean of fish, w_2 is the difference between the mean of chicken and the mean of fish.

$$w_0^* = \boxed{\text{positive}}$$

$w_1^* = \boxed{\text{negative}}$ (beef purchases tend to be less expensive than fish purchases)

$w_2^* = \boxed{\text{negative}}$ (chicken purchases tend to be less expensive than fish purchases)

(v) $h(\vec{x}_i) = w_0 + w_1 \cdot \text{meat=beef}_i + w_2 \cdot \text{meat=chicken}_i + w_3 \cdot \text{meat=fish}_i$

This model has a parameter for each meat and an intercept, but since the sum of one hot encoded features for meat is always one, the design matrix is not full rank. Therefore, the optimal solution is not unique, and there are infinitely many optimal parameter vectors $\vec{w}^*$ that minimize mean squared error.

$$w_0^* = \boxed{\text{N/A}} \quad w_1^* = \boxed{\text{N/A}} \quad w_2^* = \boxed{\text{N/A}} \quad w_3^* = \boxed{\text{N/A}}$$

For example, $\vec{w}^* = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$ and $\vec{w}^* = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 3 \end{bmatrix}$ both yield the same predictions. (Don't believe me? Write out all three cases for both $\vec{w}^*$ vectors and see for yourself.)

Activity 3: Gradients and Partial Derivatives

Suppose $\vec{x} \in \mathbb{R}^3$. Let $g(\vec{x}) = (x_1^2 + x_2 - 3)^2 + (x_1 + x_2^2 - 4)^2 + x_3^2$.

a) Find $\nabla g(\vec{x})$. *Hint: Start by finding the partial derivatives of g with respect to x_1, x_2, and x_3.*

Solution:
Let's start by finding the partial derivatives of g with respect to x_1, x_2, and x_3. We'll need to make heavy use of the (regular, scalar-to-scalar) chain rule.

$$\frac{\partial g}{\partial x_1} = 2(x_1^2 + x_2 - 3)(2x_1) + 2(x_1 + x_2^2 - 4)(1) = 4x_1(x_1^2 + x_2 - 3) + 2(x_1 + x_2^2 - 4)$$

$$\frac{\partial g}{\partial x_2} = 2(x_1^2 + x_2 - 3) + 4x_2(x_1 + x_2^2 - 4) \text{ (notice the symmetry with the first case)}$$

$$\frac{\partial g}{\partial x_3} = 2x_3$$

So,

$$\boxed{\nabla g(\vec{x}) = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 3) + 2(x_1 + x_2^2 - 4) \\ 2(x_1^2 + x_2 - 3) + 4x_2(x_1 + x_2^2 - 4) \\ 2x_3 \end{bmatrix}}$$

No need to simplify the expression any further — doing so won't make it any easier to evaluate at a specific point.

b) Evaluate $\nabla g \left(\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \right)$. The result is a vector in $\mathbb{R}^3$. What does it mean?

Solution:

$$\nabla g(\vec{x}) = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 3) + 2(x_1 + x_2^2 - 4) \\ 2(x_1^2 + x_2 - 3) + 4x_2(x_1 + x_2^2 - 4) \\ 2x_3 \end{bmatrix}$$

Notice the shared terms of $x_1^2 + x_2 - 3$ and $x_1 + x_2^2 - 4$ in the first and second components. To make the computation of the gradient at $\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$ easier, let's compute these values first.

$$x_1^2 + x_2 - 3 = 2^2 + 1 - 3 = 2, \quad x_1 + x_2^2 - 4 = 2 + 1^2 - 4 = -1$$

Then,

$$\nabla g \left(\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 4x_1(x_1^2 + x_2 - 3) + 2(x_1 + x_2^2 - 4) \\ 2(x_1^2 + x_2 - 3) + 4x_2(x_1 + x_2^2 - 4) \\ 2x_3 \end{bmatrix} = \begin{bmatrix} 4(2)(2) + 2(-1) \\ 2(2) + 4(1)(-1) \\ 0 \end{bmatrix} = \begin{bmatrix} 14 \\ 0 \\ 0 \end{bmatrix}$$

The vector $\begin{bmatrix} 14 \\ 0 \\ 0 \end{bmatrix}$ describes the direction of steepest ascent at the point $\begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$.

c) Why is it guaranteed that $g(\vec{x})$ **has** a global minimum?

Solution: $g(\vec{x}) = (x_1^2 + x_2 - 3)^2 + (x_1 + x_2^2 - 4)^2 + x_3^2$ is a sum of three squares, each of which is ≥ 0. So, knowing nothing else about what is being squared, we know that $g(\vec{x}) \geq 0$ for all $\vec{x}$, and so $g(\vec{x})$ has a global minimum of **something**, whether it's 0 or some positive number.

Activity 4: The Big Three

In Chapter 8.2, we introduced three key gradient rules for vector-to-scalar functions.

- **Dot product**: If $f(\vec{x}) = \vec{a} \cdot \vec{x}$, then $\nabla f(\vec{x}) = \vec{a}$.
- **Squared norm**: If $f(\vec{x}) = \|\vec{x}\|^2$, then $\nabla f(\vec{x}) = 2\vec{x}$.
- **Quadratic form**: If $f(\vec{x}) = \vec{x}^T A\vec{x}$, then $\nabla f(\vec{x}) = (A + A^T)\vec{x}$.

In each part below, assume $\vec{x}, \vec{a}, \vec{b} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$, and $c \in \mathbb{R}$.

a) Given $f(\vec{x}) = \vec{x}^T A\vec{x} + \vec{b}^T\vec{x} + c$, find $\nabla f(\vec{x})$.

Solution: The key idea is that the gradient of a sum is a sum of the gradients of the terms, just like with standard derivatives.

$$\nabla f(\vec{x}) = \nabla(\vec{x}^T A\vec{x} + \vec{b}^T\vec{x} + c) = \nabla(\vec{x}^T A\vec{x}) + \nabla(\vec{b}^T\vec{x}) + \nabla(c) = (A + A^T)\vec{x} + \vec{b}$$

Therefore,

$$\boxed{\nabla f(\vec{x}) = (A + A^T)\vec{x} + \vec{b}}$$

b) Given $f(\vec{x}) = \sum_{i=1}^{n} x_i$, find $\nabla f(\vec{x})$.

Solution: Remember that the sum of a vector's components is the same as the dot product of that vector with the vector of all 1's:

$$f(\vec{x}) = \sum_{i=1}^{n} x_i = \vec{x} \cdot \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

So,

$$\boxed{\nabla f(\vec{x}) = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}}$$

c) Given $f(\vec{x}) = \|A\vec{x}\|^2$, find $\nabla f(\vec{x})$. *Hint: Use the fact that* $\|\vec{v}\|^2 = \vec{v}^T\vec{v}$.

Solution:
Let's start by expanding $\|A\vec{x}\|^2$:

$$\begin{aligned} f(\vec{x}) &= \|A\vec{x}\|^2 \\ &= (A\vec{x})^T(A\vec{x}) \\ &= \vec{x}^T A^T A\vec{x} \\ &= \vec{x}^T(A^T A)\vec{x} \end{aligned}$$

$f(\vec{x})$ is a quadratic form, with the matrix $A^T A$. So,

$$\boxed{\nabla f(\vec{x}) = (A^T A + (A^T A)^T)\vec{x} = 2A^T A\vec{x}}$$

Note that $A^T A$ is symmetric, since $(A^T A)^T = A^T A$.

d) Given $f(\vec{x}) = \|\vec{x}\|$, find $\nabla f(\vec{x})$.

Solution:

$$\boxed{\nabla f(\vec{x}) = \frac{\vec{x}}{\|\vec{x}\|}}$$

For the derivation, see the **Norm and Chain Rule example in Chapter 8.2**.

e) Given $f(\vec{x}) = (\vec{a} \cdot \vec{x})^2$, find $\nabla f(\vec{x})$.
Hint: Expand $f(\vec{x})$ so that you can use one of the "big three" rules.

Solution:
Let's follow the hint.

$$\begin{aligned} f(\vec{x}) &= (\vec{a} \cdot \vec{x})^2 \\ &= (\vec{a}^T \vec{x})^2 \\ &= (\vec{x}^T \vec{a})(\vec{a}^T \vec{x}) \\ &= \vec{x}^T (\vec{a}\vec{a}^T)\vec{x} \end{aligned}$$

$f(\vec{x})$ is a quadratic form too, with the matrix $\vec{a}\vec{a}^T$. This is a symmetric matrix ($\vec{a}\vec{a}^T = (\vec{a}\vec{a}^T)^T$). So,

$$\boxed{\nabla f(\vec{x}) = 2(\vec{a}\vec{a}^T)\vec{x} = 2\vec{a}(\vec{a}^T\vec{x}) = 2(\vec{a} \cdot \vec{x})\vec{a}}$$

Activity 5: Quadratic Forms and Symmetry

Suppose $f(\vec{x}) = \vec{x}^T \begin{bmatrix} a & b \\ c & d \end{bmatrix} \vec{x}$, where $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$. (If you'd like, as an example, let $A = \begin{bmatrix} 2 & 3 \\ 7 & -8 \end{bmatrix}$.)

(i) Expand $f(\vec{x})$ so that it doesn't involve matrices or vectors.

(ii) Find $\frac{\partial f}{\partial x_1}$, $\frac{\partial f}{\partial x_2}$, and show that $\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$ satisfies the quadratic form gradient rule.

(iii) Discuss: Why do we typically assume that A is symmetric when defining a quadratic form?

Solution: This problem is the same as the **Quadratic Forms example in Chapter 8.2**.