Lecture 12

Multiple Linear Regression

DSC 40A, Fall 2024

Agenda

- Recap: regression and linear algebra
- Multiple linear regression.
- Interpreting parameters.

Recap: Regression and linear algebra

Regression and linear algebra (Solution 1)

- Define the **design matrix** $X \in \mathbb{R}^{n\times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^2$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

intercept

slope

- How do we make the hypothesis vector, $\vec{h} = X\vec{w}$, as close to $\vec{y}$ as possible? Use the parameter vector $\vec{w}^*$:

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

- Solution: We chose $\vec{w}^*$ so that $\vec{h}^* = X\vec{w}^*$ is **the projection of $\vec{y}$ onto the span of the columns of the design matrix,** X and minimized the length of the projection error $||\vec{e}|| = ||\vec{y} - X\vec{w}||$.

Regression and linear algebra (Solution 2)

- Define the **design matrix** $X \in \mathbb{R}^{n\times 2}$, **observation vector** $\vec{y} \in \mathbb{R}^n$, and parameter vector $\vec{w} \in \mathbb{R}^2$ as:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

intercept

slope

- How do we minimize the mean squared error $R_{\text{sq}}(\vec{w}) = \frac{1}{n}||\vec{y} - X\vec{w}||^2$? Using calculus the optimal paramter vector $\vec{w}^*$ is:

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

- Solution: we computed the gradient of $R_{\text{sq}}(\vec{w})$, set it to zero and solved for $\vec{w}$.

$\nabla_{\vec{w}} R_{sq}(\vec{w}) = 0 \rightarrow w^*$

Multiple linear regression

$\hat{y}$ commute time

	departure_hour	day_of_month	minutes
0	10.816667	15	68.0
1	7.750000	16	94.0
2	8.450000	22	63.0
3	7.133333	23	100.0
4	9.150000	30	69.0
...	...	...	...

So far, we've fit **simple** linear regression models, which use only **one** feature (`'departure_hour'`) for making predictions.

Incorporating multiple features

- In the context of the commute times dataset, the simple linear regression model we fit was of the form:

one input variable

$$\begin{aligned}\text{pred. commute} &= H(\text{departure hour}) \\ &= w_0 + w_1 \cdot \text{departure hour}\end{aligned}$$

- Now, we'll try and fit a multiple linear regression model of the form:

two input variables

$$\begin{aligned}\text{pred. commute} &= H(\text{departure hour}) \\ &= w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}\end{aligned}$$

- Linear regression with **multiple** features is called **multiple linear regression**.

- How do we find w_0^*, w_1^*, and w_2^*?

Hypothesis: multiple linear regression model
loss: squared error

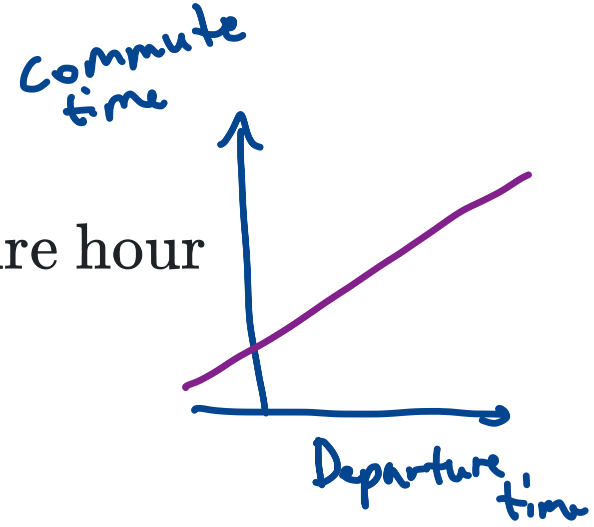
↳ use normal equations

Geometric interpretation

- The hypothesis function:

$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour}$$

  looks like a **line** in 2D.

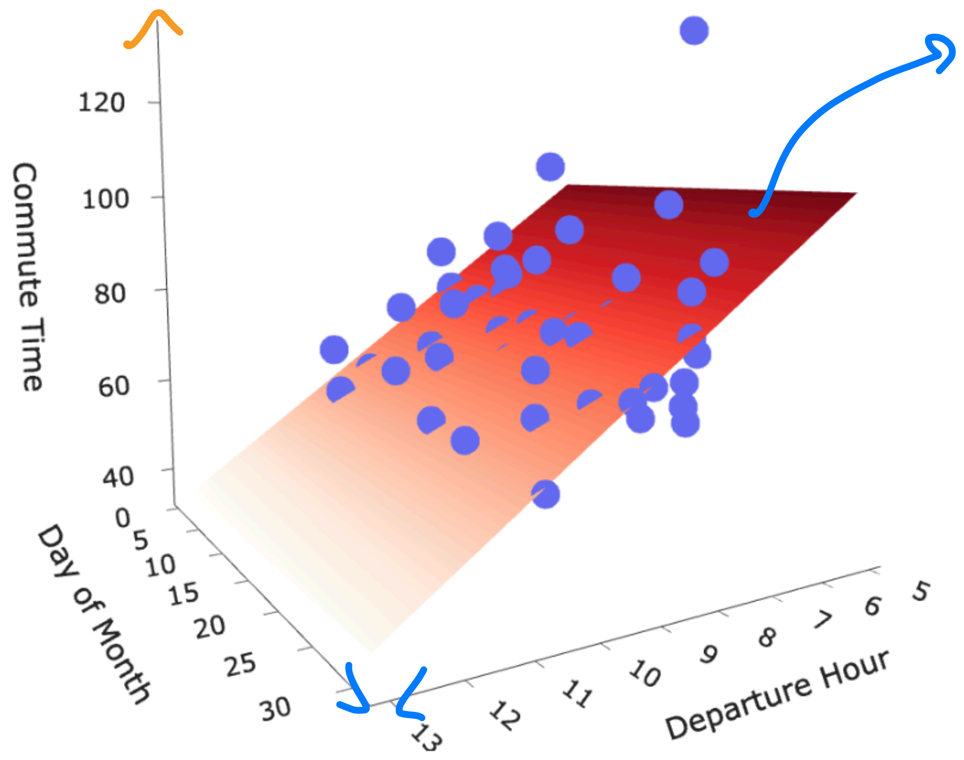


- **Questions**:
  - How many dimensions do we need to graph the hypothesis function:
    $H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$
  - What is the shape of the hypothesis function?

$z = ax + by + c$

$\Rightarrow$ plane

Commute Time vs. Departure Hour and Day of Month



H (Departure, Day of month)

Our new hypothesis function is a **plane** in 3D!

Our goal is to find the **plane** of best fit that pierces through the cloud of points.

The setup

- Suppose we have the following dataset.

y commute time

	departure_hour	day_of_month	minutes
row			
1	8.45	22	63.0
2	8.90	28	89.0
3	8.72	18	89.0

- We can represent each day with a **feature vector**, $\vec{x}$:

$$\vec{x}_1 = \begin{bmatrix} 8.45 \\ 22 \end{bmatrix} \qquad \vec{x}_2 = \begin{bmatrix} 8.90 \\ 28 \end{bmatrix} \qquad \vec{x}_3 = \begin{bmatrix} 8.72 \\ 18 \end{bmatrix}$$

$$H(\vec{x}_i) \approx y_i$$

The hypothesis vector

$\vec{h}$ = X $\vec{w}$

prediction   design matrix   parameter vector

- When our hypothesis function is of the form:

$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour} + w_2 \cdot \text{day of month}$$

the hypothesis vector $\vec{h} \in \mathbb{R}^n$ can be written as:

$$\vec{h} = \begin{bmatrix} H(\text{departure hour}_1, \text{day}_1) \\ H(\text{departure hour}_2, \text{day}_2) \\ \dots \\ H(\text{departure hour}_n, \text{day}_n) \end{bmatrix} = \begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

$\in \mathbb{R}^{n \times 1}$ $\vec{x}_1$ $n \times 3$ 3×1

X

design matrix

$H(\text{departure hour}_2, \text{day}_2) =$

$$[1, \text{departure hour}_2, \text{day}_2] \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = w_0 + w_1 \, \text{departure hour}_2 + w_2 \cdot \text{day}_2$$

$X\vec{e} = \vec{0}$

Finding the optimal parameters

- To find the optimal parameter vector, $\vec{w}^*$, we can use the **design matrix** $X \in \mathbb{R}^{n \times 3}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

$$X = \begin{bmatrix} 1 & \text{departure hour}_1 & \text{day}_1 \\ 1 & \text{departure hour}_2 & \text{day}_2 \\ \dots & \dots & \dots \\ 1 & \text{departure hour}_n & \text{day}_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} \text{commute time}_1 \\ \text{commute time}_2 \\ \vdots \\ \text{commute time}_n \end{bmatrix}$$

- Then, all we need to do is solve the **normal equations**:

$$X^T X \vec{w}^* = X^T \vec{y} \qquad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

If $X^T X$ is invertible, we know the solution is:

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

x_1 - 1st data point

Notation for multiple linear regression

$x^{(1)}$ - 1st feature in my data

- We will need to keep track of multiple features for every individual in our dataset.
  - In practice, we could have hundreds or thousands of features!
- As before, subscripts distinguish between individuals in our dataset. We have n individuals, also called **training examples**.
- Superscripts distinguish between **features**. We have d features. $x^{(1)}, x^{(2)}, \dots, x^{(d)}$

$$\text{departure hour: } x^{(1)} \in \mathbb{R}^n$$

$$\text{day of month: } x^{(2)} \in \mathbb{R}^n$$

Think of $x^{(1)}$, $x^{(2)}$, ... as new variable names, like new letters.

(column of X)

not exponent

$x_4^{(7)}$ the value of the 7th feature for the 4th data point (row of X)

Augmented feature vectors

- The **augmented feature vector** $\text{Aug}(\vec{x})$ is the vector obtained by adding a 1 to the front of feature vector $\vec{x}$:

$$\vec{x} = \begin{array}{c} \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ \vdots \\ x^{(d)} \end{bmatrix} \end{array} \begin{array}{l} \to \text{departure time} \\ \to \text{day of the month} \\ \to \text{temperature} \\ \\ \to \text{hours slept} \end{array} \qquad \text{Aug}(\vec{x}) = \underbrace{\begin{bmatrix} 1 \\ x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{bmatrix}}_{(d+1)\times 1} \qquad \vec{w} = \underbrace{\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}}_{(d+1)\times 1}$$

- Then, our hypothesis function is:

$$\begin{aligned} H(\vec{x}) &= w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)} \\ &= \underbrace{\vec{w}}_{d+1} \cdot \underbrace{\text{Aug}(\vec{x})}_{d+1} \quad \to \text{prediction} \end{aligned}$$

The general problem

- 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:

ith data point

$$\vec{x}_i = \begin{bmatrix} x_i^{(1)} \\ x_i^{(2)} \\ \vdots \\ x_i^{(d)} \end{bmatrix} \in \mathbb{R}^d$$

Simple linear regression

dataset (x, y)

$(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$

$x_i \in \mathbb{R}$ scalars

- We want to find a good linear hypothesis function:

$$\begin{aligned} H(\vec{x}) &= w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \ldots + w_d x^{(d)} \\ &= \vec{w} \cdot \text{Aug}(\vec{x}) \end{aligned}$$

How do we find $w_0^*, w_1^*, w_2^*, \ldots, w_d^*$?

The general solution

- Define the **design matrix** $X \in \mathbb{R}^{n \times (d+1)}$ and **observation vector** $\vec{y} \in \mathbb{R}^n$:

datapoint i

$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(d)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(d)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} & \dots & x_n^{(d)} \end{bmatrix}_{n \times (d+1)} = \begin{bmatrix} \text{Aug}(\vec{x}_1)^T \\ \text{Aug}(\vec{x}_2)^T \\ \vdots \\ \text{Aug}(\vec{x}_n)^T \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

- Then, solve the **normal equations** to find the optimal parameter vector, $\vec{w}^*$:

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

feature for all n datapoints

Terminology for parameters

- With d features, $\vec{w}$ has $d + 1$ entries.
- w_0 is the **bias**, also known as the **intercept**.
- $w_1, w_2, \ldots, w_d$ each give the **weight**, or **coefficient**, or **slope**, of a feature.

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

Interpreting parameters

Example: Predicting sales

- For each of 27 stores, we have:
  - net sales, y
  - square feet,
  - inventory,
  - advertising expenditure,
  - district size, and
  - number of competing stores. X
- **Goal**: Predict net sales given the other five features.
- To begin, we'll start trying to fit the hypothesis function to predict sales:

$$H(\text{square feet}, \text{competitors}) = w_0 + w_1 \cdot \text{square feet} + w_2 \cdot \text{competitors}$$

$n = 27$

$d = 5$

$d = 2$

Question 🤔

Answer at q.dsc40a.com

$$H(\text{square feet}, \text{competitors}) = w_0 + w_1 \cdot \text{square feet} + w_2 \cdot \text{competitors}$$

What will be the signs of w_1^* and w_2^*?

- A. w_1^* $+$ w_2^* $+$
- B. w_1^* $+$ w_2^* $-$
- A. w_1^* $-$ w_2^* $+$
- A. w_1^* $-$ w_2^* $-$

Let's find out! Follow along in this notebook.

Question 🤔

Answer at q.dsc40a.com

Which feature is most "important"?

- A. square feet: $w_1^* = 16.202$
- B. competitors: $w_2^* = -5.311$
- C. inventory: $w_2^* = 0.175$
- D. advertising: $w_3^* = 11.526$
- E. district size: $w_4^* = 13.580$

$5\ (100 \text{ dollars}) \approx \frac{5}{\cancel{157}}\ (100 \cdot \cancel{157} \text{ yen})$

Which features are most "important"?

- The most important feature is **not necessarily** the feature with largest magnitude weight.
- Features are measured in different units, i.e. different scales.
  - Suppose I fit one hypothesis function, H_1, with sales in US dollars, and another hypothesis function, H_2, with sales in Japanese yen (1 USD $\approx$ 157 yen).
  - Sales is just as important in both hypothesis functions.
  - But the weight of sales in H_1 will be 157 times larger than the weight of sales in H_2.
- **Solution**: If you care about the interpretability of the resulting weights, **standardize** each feature before performing regression, i.e. convert each feature to standard units.

Standard units

- Recall: to convert a feature $x_1, x_2, \ldots, x_n$ to standard units, we use the formula:

$$x_{i \text{ (su)}} = \frac{x_i - \bar{x}}{\sigma_x}$$

- Example: 1, 7, 7, 9.
  - Mean: $\frac{1+7+7+9}{4} = \frac{24}{4} = 6$.
  - Standard deviation:

$$\text{SD} = \sqrt{\frac{1}{4}((1-6)^2 + (7-6)^2 + (7-6)^2 + (9-6)^2)} = \sqrt{\frac{1}{4} \cdot 36} = 3$$

  - Standardized data:

$$1 \mapsto \frac{1-6}{3} = \boxed{-\frac{5}{3}} \qquad 7 \mapsto \frac{7-6}{3} = \boxed{\frac{1}{3}} \qquad 7 \mapsto \boxed{\frac{1}{3}} \qquad 9 \mapsto \frac{9-6}{3} = \boxed{1}$$

Standard units for multiple linear regression

- The result of standardizing each feature (separately!) is that the units of each feature are on the same scale.
    - There's no need to standardize the outcome (net sales), since it's not being compared to anything.
    - Also, we can't standardize the column of all 1s.
- Then, solve the normal equations. The resulting $w_0^*, w_1^*, \ldots, w_d^*$ are called the **standardized regression coefficients**.
- Standardized regression coefficients can be directly compared to one another.
- Note that standardizing each feature **does not** change the MSE of the resulting hypothesis function!

Once again, let's try it out! Follow along in this notebook.

Summary

- The normal equations can be used to solve multiple linear regression problems.
- Interpret the parameters as weights. Signs give meaningful information. Can only compare weight magnitude if data is standardized.
- On Friday: nonlinear features!