Lecture 14

Regression using Linear Algebra

EECS 398: Practical Data Science, Winter 2025

practicaldsc.org • github.com/practicaldsc/wn25 • < See latest announcements here on Ed

Agenda 📅

- Recap: Simple linear regression.
- Interpreting the formulas.
- Regression and linear algebra.
- Multiple linear regression.

Recap: Simple linear regression



But how did we find this line?

Recap: Simple linear regression

• Goal: Use the modeling recipe to find the "best" simple linear hypothesis function.

1. Model:
$$H(x_i) = w_0 + w_1 x_i$$
.
2. Loss function: $L_{sq}(y_i, H(x_i)) = (y_i - H(x_i))^2$.
3. Minimize empirical risk: $R_{sq}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$.

$$\implies w_1^* = rac{\displaystyle\sum_{i=1}^n (x_i - ar{x})(y_i - ar{y})}{\displaystyle\sum_{i=1}^n (x_i - ar{x})^2} = r rac{\sigma_y}{\sigma_x} \qquad w_0^* = ar{y} - w_1^* ar{x}$$

• The resulting line, $H^*(x_i) = w_0^* + w_1^* x_i$, is the unique line that minimizes MSE.

Code demo

• Before we go any further, let's test out our formulas in code.



Predicted Commute Time = 142.25 - 8.19 * Departure Hour

- The supplementary notebook is posted in the usual place on GitHub and the course website.
- Here's another related demo on another website.

Interpreting the formulas

Causality

• Can we conclude that leaving later **causes** you to get to school earlier?



Predicted Commute Time = 142.25 - 8.19 * Departure Hour

Interpreting the slope

$$w_1^* = r rac{\sigma_y}{\sigma_x}$$

- The units of the slope are **units of** *y* **per units of** *x*.
- In our commute times example, in $H^*(x_i) = 142.25 8.19x_i$, our predicted commute time **decreases by 8.19 minutes per hour**.

Interpreting the slope



• Since $\sigma_x \geq 0$ and $\sigma_y \geq 0$, the slope's sign is r's sign.

- As the y values get more spread out, σ_y increases, so the slope gets steeper.
- As the x values get more spread out, σ_x increases, so the slope gets shallower.

Interpreting the intercept

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



$$w_0^*=ar{y}-w_1^*ar{x}$$

• What are the units of the intercept?

• What is the value of $H^*(ar x)$?



Answer at practicaldsc.org/q

We fit a regression line to predict commute times given departure hour. Then, we add 75 minutes to all commute times in our dataset. What happens to the resulting regression line?

- A. Slope increases, intercept increases.
- B. Slope decreases, intercept increases.
- C. Slope stays the same, intercept increases.
- D. Slope stays the same, intercept stays the same.

Regression and linear algebra

Wait... why do we need linear algebra?

- Soon, we'll want to make predictions using more than one feature.
 - Example: Predicting commute times using departure hour and the day of the month.
- Thinking about linear regression in terms of **matrices and vectors** will allow us to find hypothesis functions that:
 - Use multiple features (input variables).

 $\circ~$ Are non-linear in the features, e.g. $H(x_i)=w_0+w_1x_i+w_2x_i^2.$

Simple linear regression, revisited



Predicted Commute Time = 142.25 - 8.19 * Departure Hour

- Model: $H(x_i) = w_0 + w_1 x_i$.
- Loss function: $(y_i H(x_i))^2$.
- To find w_0^* and w_1^* , we minimized empirical risk, i.e. average loss:

$$R_{ ext{sq}}(H) = rac{1}{n}\sum_{i=1}^n \left(y_i - H(x_i)
ight)^2$$

• Observation: $R_{
m sq}(w_0,w_1)$ kind of looks like the formula for the norm of a vector,

$$\|ec{v}\| = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2}.$$

Regression and linear algebra

Let's define a few new terms:

- The observation vector is the vector $\vec{y} \in \mathbb{R}^n$. This is the vector of observed "actual values".
- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- The error vector is the vector $\vec{e} \in \mathbb{R}^n$ with components:

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

Example

Consider
$$H(x_i) = 2 + rac{1}{2} x_i$$
.



 $ec{y}= ec{h}= ec{h}= ec{q} \ ec{e}=ec{y}-ec{h}= ec{v}$

$$egin{aligned} R_{ ext{sq}}(H) &= rac{1}{n}\sum_{i=1}^n \left(egin{smallmatrix} oldsymbol{y}_i - H(x_i)
ight)^2 \ &= \end{aligned}$$

Regression and linear algebra

Let's define a few new terms:

- The observation vector is the vector $\vec{y} \in \mathbb{R}^n$. This is the vector of observed "actual values".
- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- The error vector is the vector $\vec{e} \in \mathbb{R}^n$ with components:

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

• Key idea: We can rewrite the mean squared error of *H* as:

$$R_{ ext{sq}}(H) = rac{1}{n}\sum_{i=1}^n \left(oldsymbol{y}_i - H(x_i)
ight)^2 = rac{1}{n} \|ec{oldsymbol{e}}\|^2 = rac{1}{n} \|ec{oldsymbol{y}} - ec{oldsymbol{h}}\|^2$$

The hypothesis vector

- The **hypothesis vector** is the vector $\vec{h} \in \mathbb{R}^n$ with components $H(x_i)$. This is the vector of predicted values.
- For the linear hypothesis function $H(x) = w_0 + w_1 x$, the hypothesis vector can be written:

$$ec{h} = egin{bmatrix} w_0 + w_1 x_1 \ w_0 + w_1 x_2 \ dots \ dots \ w_0 + w_1 x_n \end{bmatrix} = \ ec{w}_0 + w_1 x_n \end{bmatrix}$$

Rewriting the mean squared error

• Define the **design matrix** $X \in \mathbb{R}^{n \times 2}$ as:

$$X = egin{bmatrix} 1 & x_1 \ 1 & x_2 \ dots & dots \ dots & dots \ 1 & x_n \end{bmatrix}$$

- Define the parameter vector $ec w \in \mathbb{R}^2$ to be $ec w = egin{bmatrix} w_0 \\ w_1 \end{bmatrix}$.
- Then, $\vec{h} = X\vec{w}$, so the mean squared error becomes:

$$R_{ ext{sq}}(H) = rac{1}{n} \|ec{m{y}} - ec{m{h}}\|^2 \implies \left| R_{ ext{sq}}(ec{w}) = rac{1}{n} \|ec{m{y}} - m{X}ec{w}\|^2
ight|^2$$

Minimizing mean squared error, again

• To find the optimal model parameters for simple linear regression, w_0^* and w_1^* , we previously minimized:

$$R_{ ext{sq}}(w_0,w_1) = rac{1}{n}\sum_{i=1}^n (m{y_i} - (w_0 + w_1m{x_i}))^2$$

• Now that we've reframed the simple linear regression problem in terms of linear algebra, we can find w_0^* and w_1^* by finding the $\vec{w}^* = \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix}$ that minimizes:

$$egin{aligned} R_{ ext{sq}}(ec{w}) &= rac{1}{n} \|ec{y} - oldsymbol{X}ec{w}\|^2 \end{aligned}$$

• Do we already know the $ec{w}^*$ that minimizes $R_{
m sq}(ec{w})$?

Minimizing mean squared error, using projections?

- X and \vec{y} are fixed: they come from our data.
- Our goal is to pick the \vec{w}^* that minimizes:

$$R_{ ext{sq}}(ec{w}) = rac{1}{n} \|ec{y} - oldsymbol{X}ec{w}\|^2$$

• This is equivalent to picking the \vec{w}^* that minimizes:

 $\|ec{y} - Xec{w}\|^2$

- This is equivalent to finding the w_0^* and w_1^* so that $X\vec{w}^*$ is as "close" to \vec{y} as possible.
- Solution: Find the orthogonal projection of \vec{y} onto span(X)!
- We already did this in Linear Algebra Guide 4, which you're reviewing in Homework 6, Question 6!

An optimization problem we've seen before

• The optimal parameter vector, $ec{w}^* = \begin{bmatrix} w_0^* & w_1^* \end{bmatrix}^T$, is the one that minimizes:

$$R_{ ext{sq}}(ec{w}) = rac{1}{n} \|ec{y} - oldsymbol{X}ec{w}\|^2$$

• In LARDS Section 8 (and your linear algebra class), we showed that the \vec{w}^* that minimizes the length of the error vector, $\|\vec{e}\| = \|\vec{y} - X\vec{w}\|$, is the one that satisifes the **normal equations**:

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

• The minimizer of $\|ec{e}\|$ is the same as the minimizer of $R_{
m sq}(ec{w}).$

$$\frac{1}{n} \|\vec{e}\|^2 = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

• Key idea: The \vec{w}^* that solves the normal equations also minimizes $R_{
m sq}(\vec{w})!$

The normal equations

• The normal equations are the system of 2 equations and 2 unknowns defined by:

$$egin{array}{ll} X^T X ec{w}^* = X^T ec{y} \end{array}$$

- Why are they called the **normal** equations?
- If $X^T X$ is invertible, there is a unique solution to the normal equations:

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

• If $X^T X$ is not invertible, then there are infinitely many solutions to the normal equations. We will explore this idea as the semester progresses.

The optimal parameter vector, $ec{w}^*$

- To find the optimal model parameters for simple linear regression, w_0^* and w_1^* , we previously minimized $R_{
 m sq}(w_0,w_1)=rac{1}{n}\sum_{i=1}^n(y_i-(w_0+w_1x_i))^2$.
 - We found, using calculus, that:

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

• $w_0^* = \bar{y} - w_1^* \bar{x}.$

• Another way of finding optimal model parameters for simple linear regression is to find the \vec{w}^* that minimizes $R_{
m sq}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$.

 $\circ~$ The minimizer, if $X^T X$ is invertible, is the vector $\left|ec{w}^* = (X^T X)^{-1} X^T ec{y}
ight|$

• These formulas are equivalent!

Code demo

• To give us a break from math, we'll switch to a notebook, showing that both formulas – that is, (1) the formulas for w_1^* and w_0^* we found using calculus, and (2) the formula for \vec{w}^* we found using linear algebra – give the same results.

• You'll prove this in Homework 7 \cong .

- We'll use the same supplementary notebook as earlier, posted in the usual place on GitHub and the course website.
- Then, we'll use our new linear algebraic formulation of regression to incorporate **multiple features** in our prediction process.

Summary: Regression and linear algebra

• 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:

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

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

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

• 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.

Multiple linear regression

	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:

 $ext{pred. commute} = H(ext{departure hour}_i) \ = w_0 + w_1 \cdot ext{departure hour}_i$

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

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

- Linear regression with multiple features is called multiple linear regression.
- How do we find w_0^* , w_1^* , and w_2^* ?

Geometric interpretation

• The hypothesis function:

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

looks like a line in 2D.

• Questions:

 \circ How many dimensions do we need to graph the hypothesis function:

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

• What is the shape of the hypothesis function?

Commute Time vs. Departure Hour and 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 hypothesis vector

• When our hypothesis function is of the form:

 $H(ext{departure hour}_i, ext{day of month}_i) = w_0 + w_1 \cdot ext{departure hour}_i + w_2 \cdot ext{day of month}_i$ the hypothesis vector $ec{h} \in \mathbb{R}^n$ can be written as:

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

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 once again:

 $X^T X ec{w}^* = X^T ec{y}$

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

$$ec{w}^* = (X^T X)^{-1} X^T ec{y}$$
 36

Code demo

- Let's switch back to the notebook and use what we've just learned to find the w_0^*, w_1^* , and w_2^* that minimize mean squared error for the following hypothesis function: $H(\text{departure hour}_i, \text{day of month}_i) = w_0 + w_1 \cdot \text{departure hour}_i + w_2 \cdot \text{day of month}_i$
 - We'll use the same supplementary notebook as earlier, posted in the usual place on GitHub and the course website.
 - Next class, we'll present a more general formulation of multiple linear regression and see how it can be used to incorporate (many) more sophisticated features.
 - Then, we'll start discussing the nature of **how we choose which features to use**, and why more isn't always better.