Lectures 6-7

one input variable / feature

Simple Linear Regression

DSC 40A, Fall 2024

Agenda

- Simple linear regression.
- Minimizing mean squared error for the simple linear model.
- Correlation.
- Interpreting the formulas.
- Connections to related models.
- What next? Linear algebra.

Groupwork policy enforced starting with groupwork 2

Question 🤔

Answer at q.dsc40a.com

Remember, you can always ask questions at q.dsc40a.com!

If the direct link doesn't work, click the "🤔 Lecture Questions" link in the top right corner of dsc40a.com.

Finding the best linear model

- **Goal:** Out of all **linear** functions $\mathbb{R} \to \mathbb{R}$, find the function H^* with the smallest mean squared error.
  - Linear functions are of the form $H(x) = w_0 + w_1 x$.
  - They are defined by a slope (w_1) and intercept (w_0).
- That is, H^* should be the linear function that minimizes

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - H(x_i) \right)^2$$

- We chose squared loss, since it's the easiest to minimize.

Minimizing mean squared error for the simple linear model

- Our goal is to find the linear hypothesis function $H^*(x)$ that minimizes empirical risk:

$$R_{\text{sq}}(H) = \frac{1}{n} \sum_{i=1}^{n} (y_i - H(x_i))^2$$

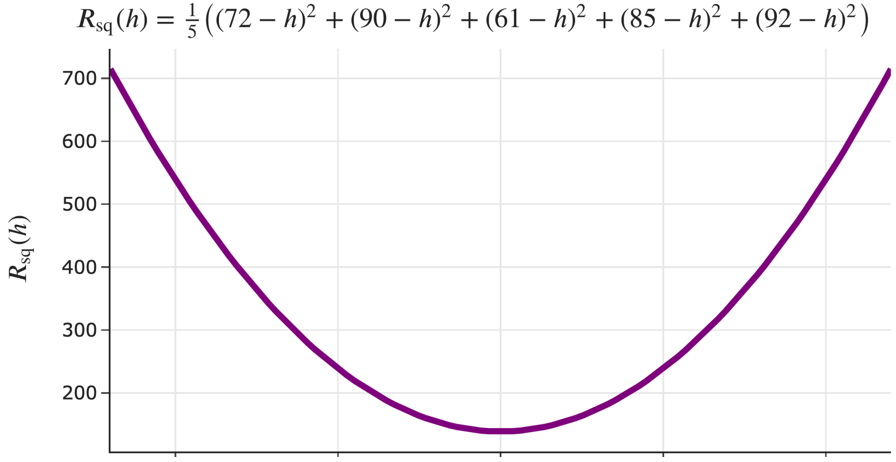
- Plugging in the linear hypothesis $H(x) = w_0 + w_1 x$, we can re-write R_{sq} as a function of w_0 and w_1:

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

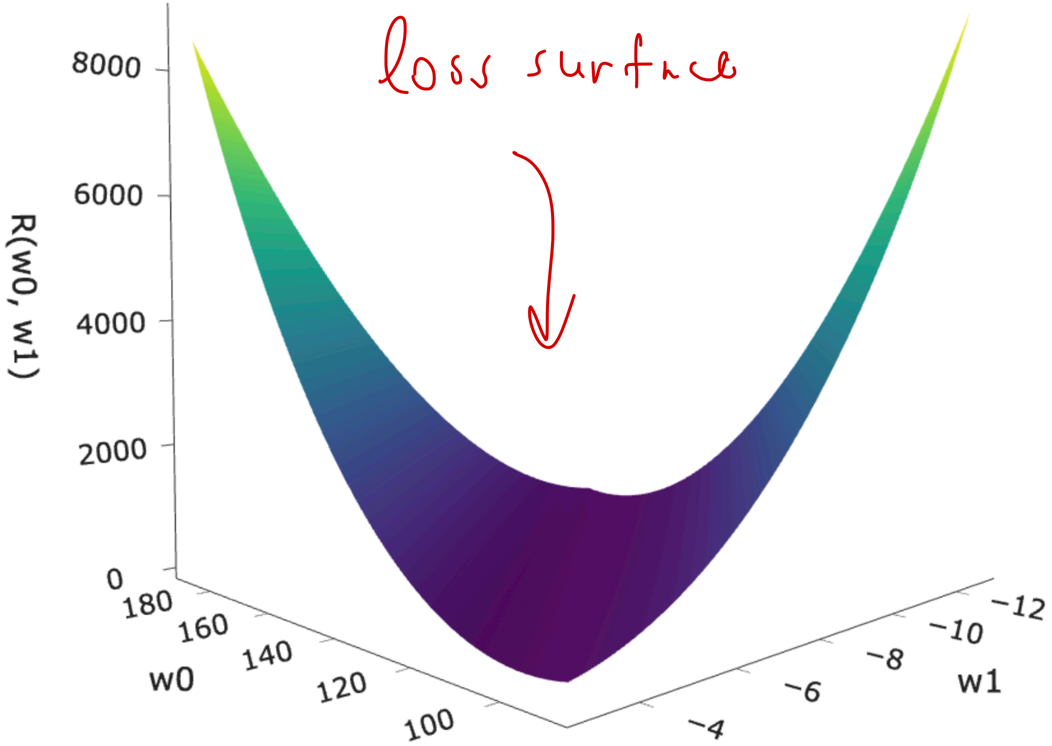
- **How do we find the parameters w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$?**

Loss surface

For the constant model, the graph of $R_{\text{sq}}(h)$ looked like a parabola.



What does the graph of $R_{\text{sq}}(w_0, w_1)$ look like for the simple linear regression model?



Minimizing mean squared error for the simple linear model

Minimizing multivariate functions

- Our goal is to find the parameters w_0^* and w_1^* that minimize mean squared error:

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

- R_{sq} is a function of two variables: w_0 and w_1.
- To minimize a function of multiple variables:
  - Take partial derivatives with respect to each variable. $\frac{\partial R_{sq}}{\partial w_0}, \frac{\partial R_{sq}}{\partial w_1}$
  - Set all partial derivatives to 0. $\frac{\partial R_{sq}}{\partial w_0}(\quad) = 0, \quad \frac{\partial R_{sq}}{\partial w_1}(\quad) = 0$
  - Solve the resulting system of equations.
  - Ensure that you've found a minimum, rather than a maximum or saddle point (using the second derivative test for multivariate functions).

R_{sq} is parabolic, convex → single minima

Example

Find the point (x, y, z) at which the following function is minimized.

$$f(x,y) = \underline{x^2 - 8x} + \underline{y^2 + 6y} - 7$$

complete the square
(no calculus)

$f(x,y) = (x-a)^2 + (y-b)^2 + c$

$f(x,y) = (x-4)^2 - 16 + (y+3)^2 - 9 - 7$

$= \underbrace{(x-4)^2}_{\geq 0} + \underbrace{(y+3)^2}_{\geq 0} \underbrace{- 32}_{\text{const}}$ $\Rightarrow$ $x^* = 4$, $y^* = -3$, $f(x^*, y^*) = -32$

using calculus

$f_x = \frac{\partial f}{\partial x} = 2x - 8$

$f_y = \frac{\partial f}{\partial y} = 2y + 6$

$\Rightarrow$ $x^* = 4$, $y^* = -3$

$f(4,-3) = 16 - 32 + 9 - 18 - 7 = -32$

$\Rightarrow (x^*, y^*) = (4, -3) = \arg\min_{x,y} f(x,y)$ $\quad -32 = \min_{x,y} f(x,y)$

Minimizing mean squared error

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

To find the w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$, we'll:

1. Find $\frac{\partial R_{\text{sq}}}{\partial w_0}$ and set it equal to 0.
2. Find $\frac{\partial R_{\text{sq}}}{\partial w_1}$ and set it equal to 0.
3. Solve the resulting system of equations.

Question 🤔

Answer at q.dsc40a.com

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

Which of the following is equal to $\frac{\partial R_{\text{sq}}}{\partial w_0}$?

- A. $\frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))$
- B. $-\frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))$
- C. $-\frac{2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i)) x_i$
- D. $-\frac{2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))$

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

w_0 = "W naught"

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

$$= \frac{1}{n}\sum_{i=1}^{n}\overbrace{2(y_i - (w_0 + w_1 x_i))\underbrace{\frac{\partial}{\partial w_0}(y_i - (w_0 + w_1 x_i))}_{-1}}^{\text{chain rule}}$$

$$= \frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) \cdot (-1) = \frac{-2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))$$

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

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

$$= \frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))(-x_i)$$

$$= \frac{-2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))x_i$$

Strategy

We have a system of two equations and two unknowns (w_0 and w_1):

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0 \qquad -\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))x_i = 0$$

To proceed, we'll:

1. Solve for w_0 in the first equation.
   The result becomes w_0^*, because it's the "best intercept."
2. Plug w_0^* into the second equation and solve for w_1.
   The result becomes w_1^*, because it's the "best slope."

Goal: isolate w_0

Solving for w_0^*

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0$$

$$\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i)) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} w_0 - \sum_{i=1}^{n} w_1 x_i = 0$$

$$\sum_{i=1}^{n} w_0 = \underbrace{w_0 + w_0 + \dots + w_0}_{n \text{ times}} = n w_0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - n w_0 - w_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - w_1 \sum_{i=1}^{n} x_i = n w_0$$

$$w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - w_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\boxed{w_0^* = \bar{y} - w_1^* \bar{x}}$$

defined

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for w_1^*

$\frac{\partial R_{sq}}{\partial w_1} = 0$

Goal: isolate w_1

$$-\frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))x_i = 0$$

$$\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))x_i = 0$$

$w_0 = \bar{y} - w_1 \bar{x}$

$$\sum_{i=1}^{n}(y_i - (\bar{y} - w_1\bar{x} + w_1 x_i))x_i = 0$$

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i - w_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i = 0$$

$$\sum_{i=1}^{n}(y_i - \bar{y})x_i = w_1 \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

$$\boxed{w_1^* = \frac{\sum_{i=1}^{n}(y_i - \bar{y})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})x_i}}$$

* cannot cancel out x_i in numerator and denominator

Least squares solutions

We've found that the values w_0^* and w_1^* that minimize R_{sq} are:

$$w_1^* = \frac{\displaystyle\sum_{i=1}^{n} (y_i - \bar{y}) x_i}{\displaystyle\sum_{i=1}^{n} (x_i - \bar{x}) x_i} \qquad\qquad w_0^* = \bar{y} - w_1^* \bar{x}$$

where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad\qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

These formulas work, but let's re-write w_1^* to be a little more symmetric.

An equivalent formula for w_1^*

Claim:

$$w_1^* = \frac{\boxed{\sum_{i=1}^{n}(y_i-\bar{y})x_i}}{\boxed{\sum_{i=1}^{n}(x_i-\bar{x})x_i}} = \frac{\boxed{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}}{\boxed{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$(*)$ $\sum_{i=1}^{n}(y_i-\bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - n\frac{1}{n}\sum_{i=1}^{n} y_i = 0$

use $\sum^{n}(x_i-\bar{x}) = 0$ for denominator

$\boxed{(a-b)(c-d) = a(c-d) - b(c-d)}$

Proof:

right numerator

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n} x_i(y_i-\bar{y}) - \overbrace{\sum_{i=1}^{n}\bar{x}(y_i-\bar{y})}^{\text{show } =0} =$$

and we're done

$$= \sum_{i=1}^{n}(y_i-\bar{y})x_i - \bar{x}\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})}_{\overset{(*)}{=}0} = \boxed{\sum_{i=1}^{n}(y_i-\bar{y})x_i}$$

left numerator

need to show

1) $\square = \square \Rightarrow \frac{\square}{\square} = \frac{\square}{\square}$

2) $\square = \square$

Least squares solutions

- The **least squares solutions** for the intercept w_0 and slope w_1 are:

$$\boxed{w_1^* = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad w_0^* = \bar{y} - w_1^*\bar{x}}$$

- We say w_0^* and w_1^* are **optimal parameters**, and the resulting line is called the **regression line**.

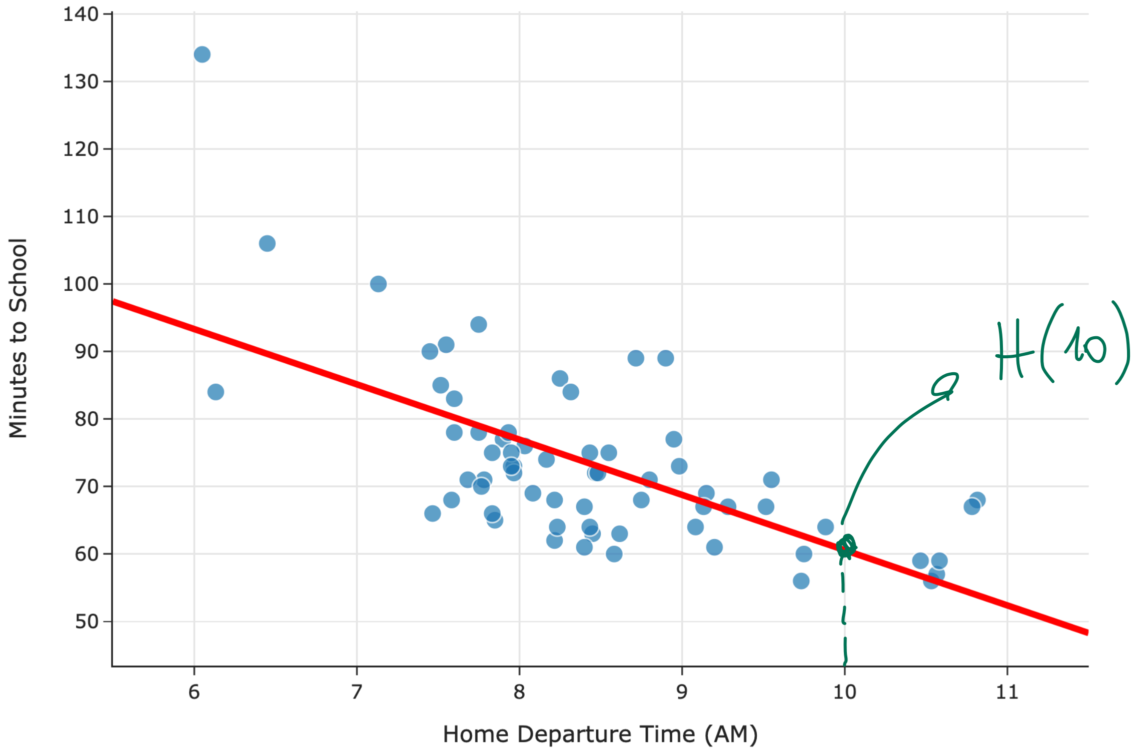
when using squared loss

- The process of minimizing empirical risk to find optimal parameters is also called "**fitting to the data.**"
- To make predictions about the future, we use $\boxed{H^*(x) = w_0^* + w_1^* x}$.

Causality

w_0^* w_1^*

Predicted Commute Time = $\underbrace{142.25}$ - $\underbrace{8.19}$ * Departure Hour



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

No!

This is just a pattern!

What's next?

We now know how to find the optimal slope and intercept for linear hypothesis functions. Next, we'll:

- See how the formulas we just derived connect to the formulas for the slope and intercept of the regression line we saw in DSC 10.
  - They're the same, but we need to do a bit of work to prove that.
- Learn how to interpret the slope of the regression line.
- Discuss *causality*.
- Learn how to build regression models with **multiple inputs**.
  - To do this, we'll need linear algebra!