

DSC 40A

Lecture 07
Least Squares Regression, pt. II

Last Time

- ▶ **Goal:** Find prediction rule $H(x)$ for predicting salary given years of experience.
- ▶ Minimize mean absolute error?

$$\frac{1}{n} \sum_{i=1}^n |H(x_i) - y_i|$$

- ▶ **Not differentiable**, instead minimize **mean squared error**:

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

- ▶ To avoid **overfitting**, use linear prediction rule:

$$H(x) = w_1 x + w_0$$

Last Time

- ▶ **Goal:** find w_1 and w_0 which minimize MSE:

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

- ▶ **Strategy:** Take derivatives $\partial R_{\text{sq}} / \partial w_1$ and $\partial R_{\text{sq}} / \partial w_0$, set to zero, solve.
- ▶ We found:

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

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

Today

- ▶ Solve these equations to find the **least squares solutions**.
- ▶ See how to easily fit non-linear trends, too.

Strategy

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

1. Solve for w_0 in second equation.
2. Plug solution for w_0 into first equation, solve for w_1 .

Solve for w_0

$$\sum_{i=1}^n (a_i + b_i) = \sum_{i=1}^n a_i + \sum_{i=1}^n b_i$$

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

Multiplying both sides by $\frac{n}{2}$:

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

Breaking the summand into three pieces:

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

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

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

Dividing by n :

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

$$\text{Define } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

Then

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

Solve for w_0

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

Key Fact

- ▶ Define

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

- ▶ Then

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

Solve for w_1

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

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

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

Grouping the things with w_1 :

$$\Rightarrow 0 = \sum [w_1(x_i - \bar{x}) - (y_i - \bar{y})] x_i$$

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

Splitting the summand:

$$\Rightarrow 0 = \sum w_1(x_i - \bar{x}) x_i - \sum (y_i - \bar{y}) x_i$$

Solve for w_1

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

$$\Rightarrow 0 = \sum w_1 (x_i - \bar{x}) x_i - \sum (y_i - \bar{y}) x_i$$

$$\Rightarrow \sum (y_i - \bar{y}) x_i = w_1 \sum (x_i - \bar{x}) x_i$$

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

Solve for w_1

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

This formula works.

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

But this one is better:

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

Show numerators are equal:
(Denominators are too, but left as an exercise)

$$\sum (y_i - \bar{y})(x_i - \bar{x})$$

$$= \sum [(y_i - \bar{y})x_i - (y_i - \bar{y})\bar{x}]$$

$$= \sum (y_i - \bar{y})x_i - \sum (y_i - \bar{y})\bar{x}$$

$$= \sum (y_i - \bar{y})x_i - \bar{x} \sum (y_i - \bar{y})$$

$$= \sum (y_i - \bar{y})x_i$$

Least Squares Solutions

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

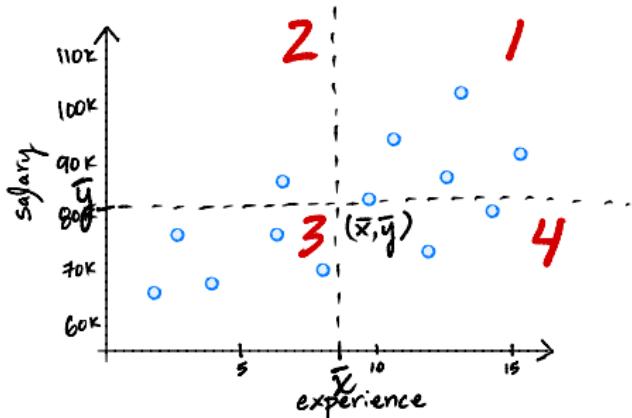
$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad w_0 = \bar{y} - w_1 \bar{x}$$

where

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

Interpretation of Slope

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

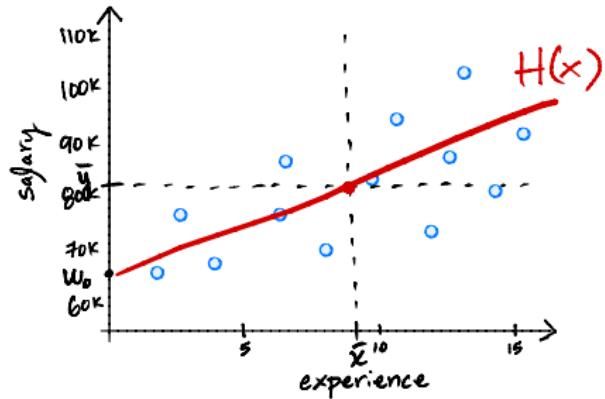


► What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$ when:

- $x_i > \bar{x}$ and $y_i > \bar{y}$? + Q1
- $x_i < \bar{x}$ and $y_i < \bar{y}$? + Q3
- $x_i > \bar{x}$ and $y_i < \bar{y}$? - Q4
- $x_i < \bar{x}$ and $y_i > \bar{y}$? - Q2

Interpretation of Intercept

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



- ▶ What is $H(\bar{x})$?

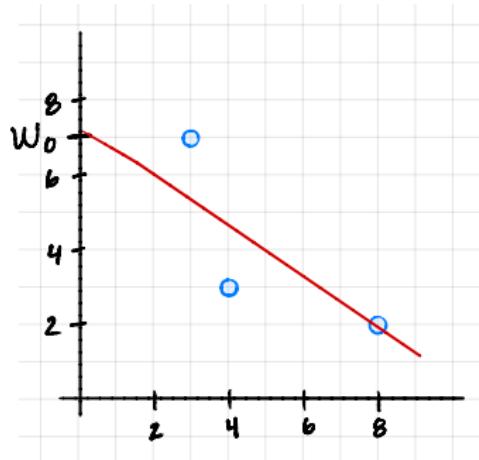
$$\begin{aligned}H(\bar{x}) &= w_1 \bar{x} + w_0 \\&= w_1 \bar{x} + [\bar{y} - w_1 \bar{x}] \\&= \bar{y}\end{aligned}$$

Discussion Question

We fit a linear prediction rule for salary given years of experience. Then everyone gets a \$5,000 raise. Which of these happens?

- a) slope increases, intercept increases
- b) slope decreases, intercept increases
- c) slope stays same, intercept increases
- d) slope stays same, intercept stays same

Example



$$\bar{x} = 5$$

$$\bar{y} = 4$$

$$w_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{-11}{14}$$

$$w_0 = \bar{y} - w_1 \bar{x} = 4 - \left(-\frac{11}{14}\right)5 = 4 + \frac{55}{14}$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
3	7	-2	3	-6	4
4	3	-1	-1	1	1
8	2	3	-2	-6	9
			-11		14

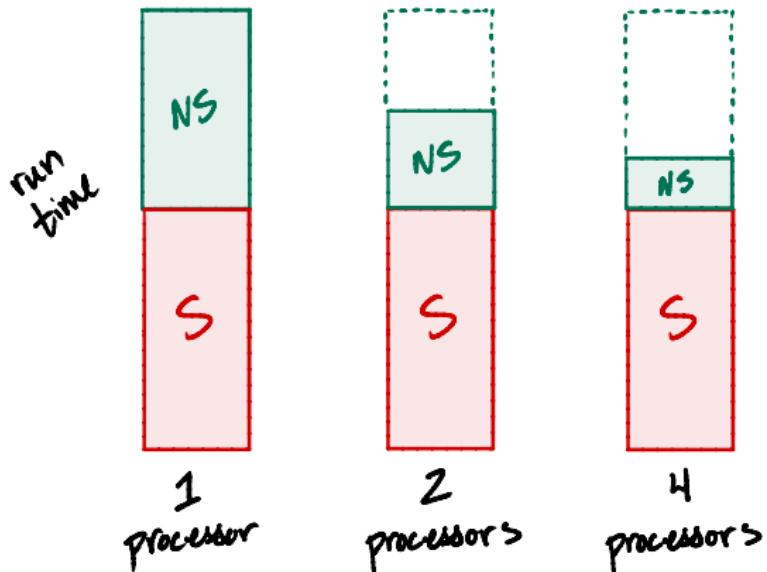
Example: Parallel Processing



Problem

- ▶ Some parts of a program are necessarily **sequential**.
- ▶ E.g., downloading the data must happen before analysis.
- ▶ More processors do not speed up **sequential** code.
- ▶ But they do speed up **non-sequential** code.

Speedup



Amdahl's Law

The time T it takes to run a program on p processors is:

$$T(p) = t_S + \frac{t_{NS}}{p}$$

where t_S and t_{NS} are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Amdahl's Law

The time T it takes to run a program on p processors is:

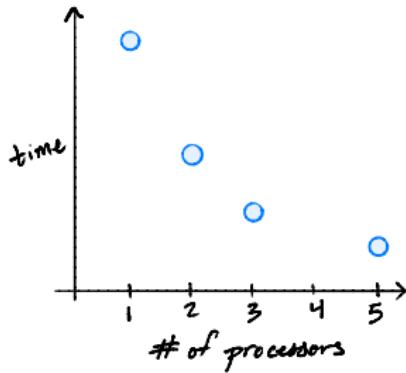
$$T(p) = t_S + \frac{t_{NS}}{p}$$

where t_S and t_{NS} are the time it takes the sequential and non-sequential parts to run on one processor, respectively.

Problem: we don't know t_S and t_{NS} .

Fitting Amdahl's Law

- ▶ **Solution:** we will learn t_S and t_{NS} from data.
- ▶ Run with varying number of processors, record total time:



- ▶ Find decision rule $H(p) = \frac{t_{NS}}{p} + t_S$ by minimizing MSE.
~~prediction~~

General Problem

- ▶ Given data $(x_1, y_1), \dots, (x_n, y_n)$.
- ▶ Fit a **non-linear** rule $H(x) = w_1 \cdot \frac{1}{x} + w_0$ by minimizing MSE:

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

Using definition of H :

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

Minimizing MSE

- ▶ Take derivatives, you'll find:

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

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

Minimizing MSE

- ▶ Set to zero, solve. You'll find:

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

$$w_0 = \bar{y} - w_1 \cdot \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$$

Minimizing MSE

- ▶ Set to zero, solve. You'll find:

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

- ▶ Define $z_i = \frac{1}{x_i}$, $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}$. Then:

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2}$$
$$w_0 = \bar{y} - w_1 \bar{z}$$

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot \frac{1}{x} + w_0$:

1. Create a new data set $(z_1, y_1), \dots, (z_n, y_n)$, where $z_i = \frac{1}{x_i}$.
2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \quad w_0 = \bar{y} - w_1 \cdot \bar{z}$$

3. Use w_1 and w_0 in original decision rule, $H(x)$.

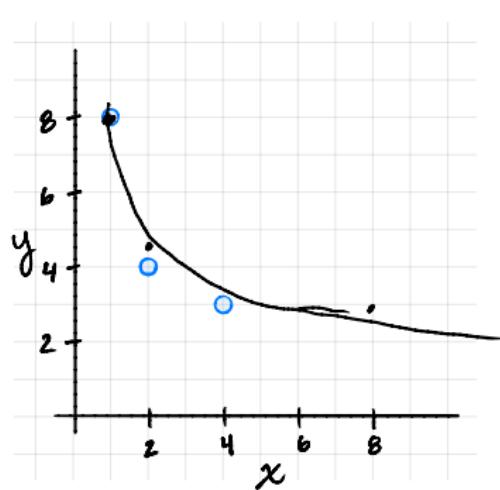
Example: Amdahl's Law

- ▶ We have timed our program:

Processors	Time (Hours)
1	8
2	4
4	3

- ▶ Fit prediction rule: $H(p) = \frac{t_{NS}}{p} + t_S$

Example: fitting $H(x) = w_1 \cdot \frac{1}{x_i} + x_0$



$$\bar{z} = \frac{7}{4} \cdot \frac{1}{3} = \frac{7}{12}$$

$$\bar{y} = 5$$

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} = 2 \cdot \frac{144}{42} = \frac{144}{21}$$

$$w_0 = \bar{y} - w_1 \bar{z} \quad 5 - \frac{144}{21} \cdot \frac{7}{12} = 5 - \frac{12}{3} = 1$$

x_i	z_i	y_i	$(z_i - \bar{z})$	$(y_i - \bar{y})$	$(z_i - \bar{z})(y_i - \bar{y})$	$(z_i - \bar{z})^2$
1	1	8	5/12	3	15/12	25/144
2	1/2	4	-1/12	-1	1/12	1/144
4	1/4	3	-4/12	-2	8/12	16/144
					2	42/144

Example: Amdahl's Law

- ▶ We found: $t_{NS} = \frac{48}{7} \approx 6.88$, $t_S = 1$
- ▶ Our prediction rule:

$$H(p) = \frac{t_{NS}}{p} + t_S$$

$$= \frac{6.88}{p} + 1$$

Fitting Non-Linear Trends

To fit a prediction rule of the form $H(x) = w_1 \cdot f(x) + w_0$:

1. Create a new data set $(z_1, y_1), \dots, (z_n, y_n)$, where $z_i = f(x_i)$.
2. Fit $H(z) = w_1 z + w_0$ using familiar least squares solutions:

$$w_1 = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})^2} \quad w_0 = \bar{y} - w_1 \cdot \bar{z}$$

3. Use w_1 and w_0 in original decision rule, $H(x)$.

Fitting Non-Linear Trends

- We can fit rules like:

$$w_1x + w_0 \quad w_1 \cdot \frac{1}{x} + w_0 \quad w_1x^2 + w_0 \quad w_1e^x + w_0$$

- We can't fit rules like:

$$w_0e^{w_1x} \quad \sin(w_1x + w_0)$$

- Can fit as long as **linear** function of w_1, w_0 .

What's Left?

- ▶ How do we make predictions with lots of features?
- ▶ E.g., experience, age, GPA, number of internships, etc.