Lectures 6-7

Simple Linear Regression

DSC 40A, Fall 2024

Agenda

- Simple linear regression.
- Correlation.

Least squares solution

- Interpreting the formulas.
- Connections to related models.

- HW 1 due tonight, HW 2 released
- Submit regrade requests - no need for emails

Least squares solutions

- Our goal was to find the parameters w_0^* and $w_1{}^*$ that minimized:

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

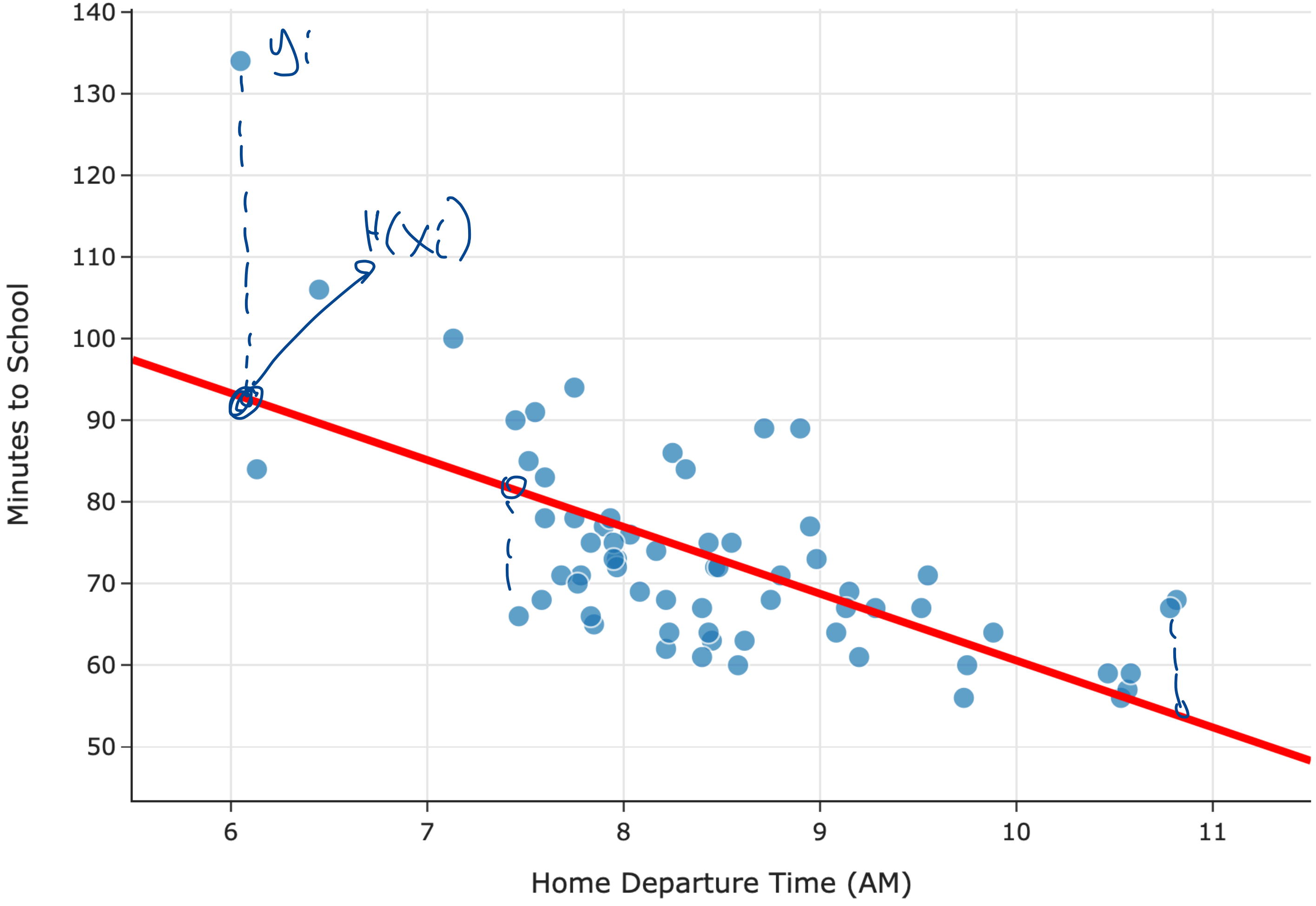
- To do so, we used calculus, and we found that the minimizing values are:

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

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

optimal intercept   optimal slope

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



There is no other <u>line</u> for this dataset with smaller MSE

$$w_0^*, w_1^* = \underset{w_0, w_1}{\arg\min}\, MSE(H)$$

Now what?

We've found the optimal slope and intercept for linear hypothesis functions using squared loss (i.e. for the regression line). Now, we'll:

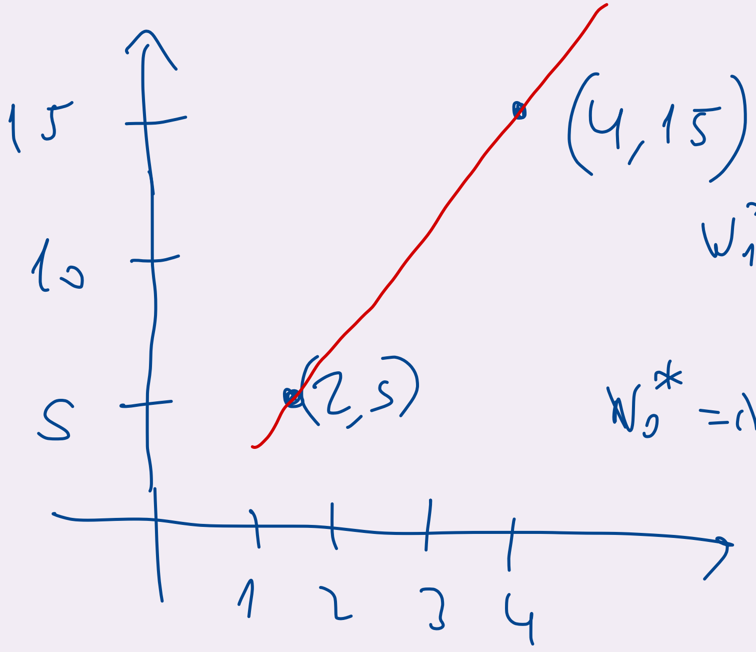
- See how the formulas we just derived connect to the formulas for the slope and intercept of the regression line we saw in DSC 10.
    - They're the same, but we need to do a bit of work to prove that.
- Learn how to interpret the slope of the regression line.
- Understand connections to other related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

Question 🤔

Answer at q.dsc40a.com

Consider a dataset with just two points, $(2, 5)$ and $(4, 15)$. Suppose we want to fit a linear hypothesis function to this dataset using squared loss. What are the values of w_0^* and w_1^* that minimize empirical risk?

- A. $w_0^* = 2$, $w_1^* = 5$
- B. $w_0^* = 3$, $w_1^* = 10$
- C. $w_0^* = -2$, $w_1^* = 5$
- D. $w_0^* = -5$, $w_1^* = 5$



$w_1^* = \text{slope} = \frac{15-5}{4-2} = 5$

$w_0^* = \text{intercept}$

$4 \cdot 5 + w_0^* = 15$

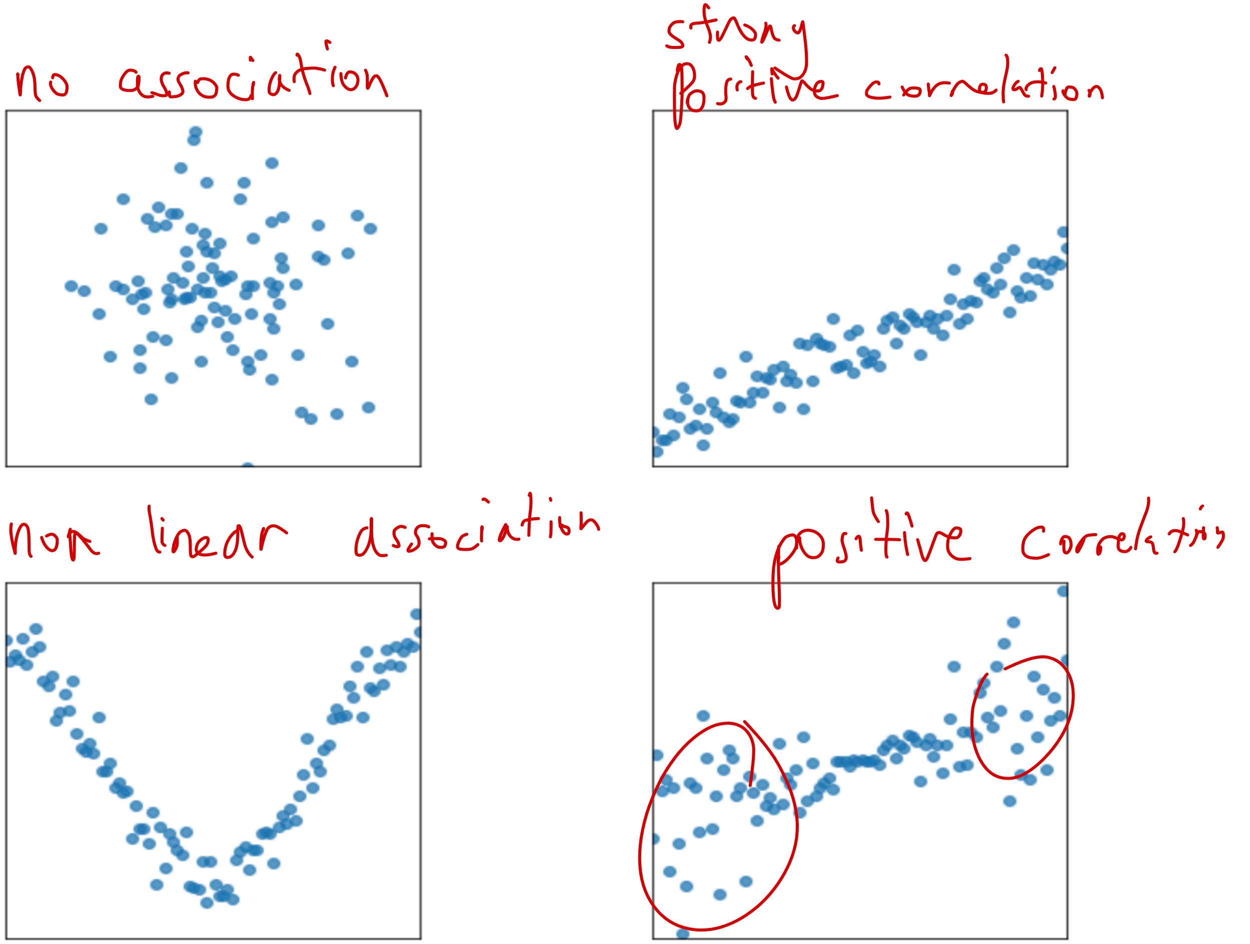
$w_0^* = -5$

Correlation

correlation = linear association
does it look like a line?

Quantifying patterns in scatter plots

- In DSC 10, you were introduced to the idea of the **correlation coefficient**, r.
- It is a measure of the strength of the **linear association** of two variables, x and y.
- Intuitively, it measures how tightly clustered a scatter plot is around a straight line.
- It ranges between -1 and 1.



r negative : negative association
r positive : positive association
the closer r is to ±1, the stronger the correlation!

Pearson's correlation

The correlation coefficient

- The correlation coefficient, r, is defined as the **average of the product of x and y, when both are in standard units**.
- Let σ_x be the standard deviation of the x_is, and $\bar{x}$ be the mean of the x_is.
- x_i in standard units is $\frac{x_i - \bar{x}}{\sigma_x}$. ← mean centering ← std.
- The correlation coefficient, then, is:

$$r = \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{\text{average}} \underbrace{\left(\frac{x_i - \bar{x}}{\sigma_x} \right)}_{x \text{ in standard units}} \underbrace{\left(\frac{y_i - \bar{y}}{\sigma_y} \right)}_{y \text{ in standard units}}$$

alterative

covariance

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

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

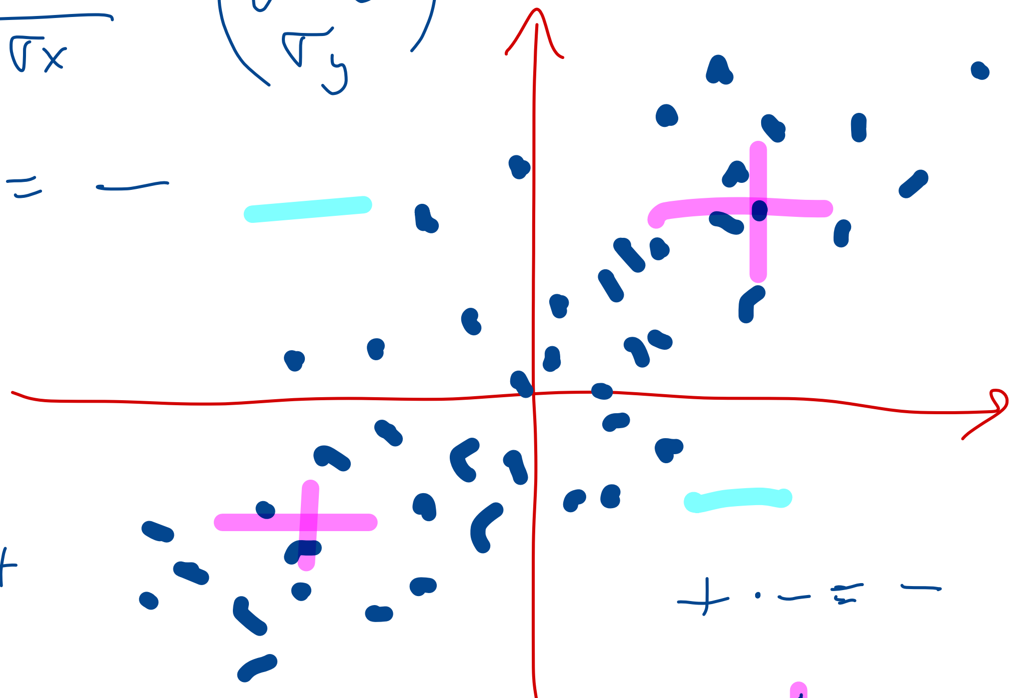
$$\sigma_{xx} = \sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\frac{1}{n}\sum_i \left(\frac{x_i-\bar{x}}{\sigma_x}\right)\cdot\left(\frac{y_i-\bar{y}}{\sigma_y}\right)$$

$$\left(\frac{x_i \text{ in}}{SD}\right)\left(\frac{y_i \text{ in}}{SD}\right)$$

$+ \cdot +$

$- \cdot + = -$

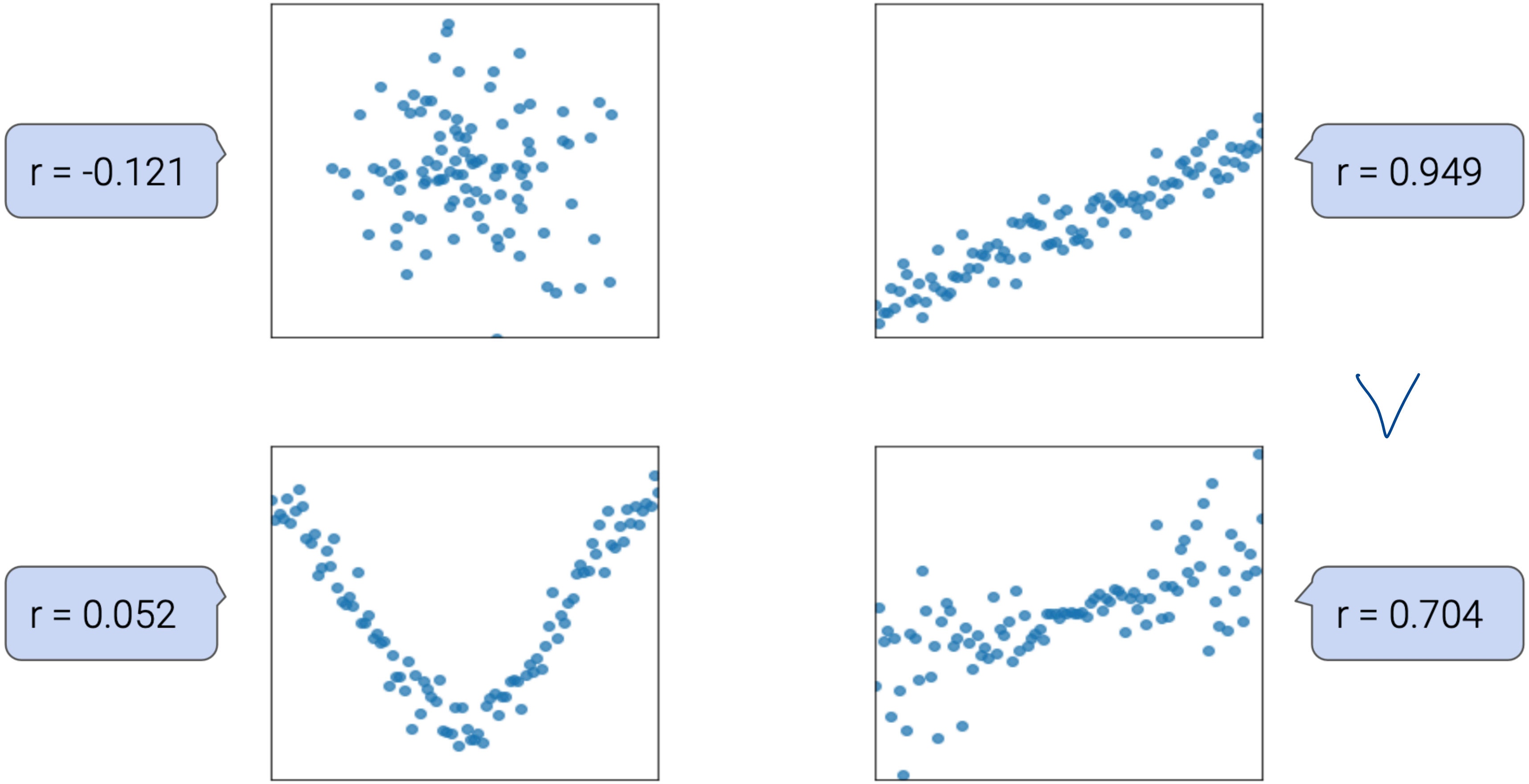


$- \cdot - = +$

$+ \cdot - = -$

$$r = \sum(\;) = + + + + (-) + (-) > 0$$

The correlation coefficient, visualized



Another way to express w_1^*

- It turns out that w_1^*, the optimal slope for the linear hypothesis function when using squared loss (i.e. the regression line), can be written in terms of r!

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

- It's not surprising that r is related to w_1^*, since r is a measure of linear association.
- Concise way of writing w_0^* and w_1^*:

$$w_1^* = r\frac{\sigma_y}{\sigma_x} \qquad w_0^* = \bar{y} - w_1^*\bar{x}$$

Proof that $w_1^* = r\frac{\sigma_y}{\sigma_x}$

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

$$= \frac{\cancel{n} r \cancel{\sigma_x} \sigma_y}{\cancel{n} \sigma_x^{\cancel{2}}} =$$

$$= \frac{r\sigma_y}{\sigma_x}$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma_x}\right)\left(\frac{y_i-\bar{y}}{\sigma_y}\right)$$

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

$$\boxed{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})} = n r \sigma_x \sigma_y$$

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

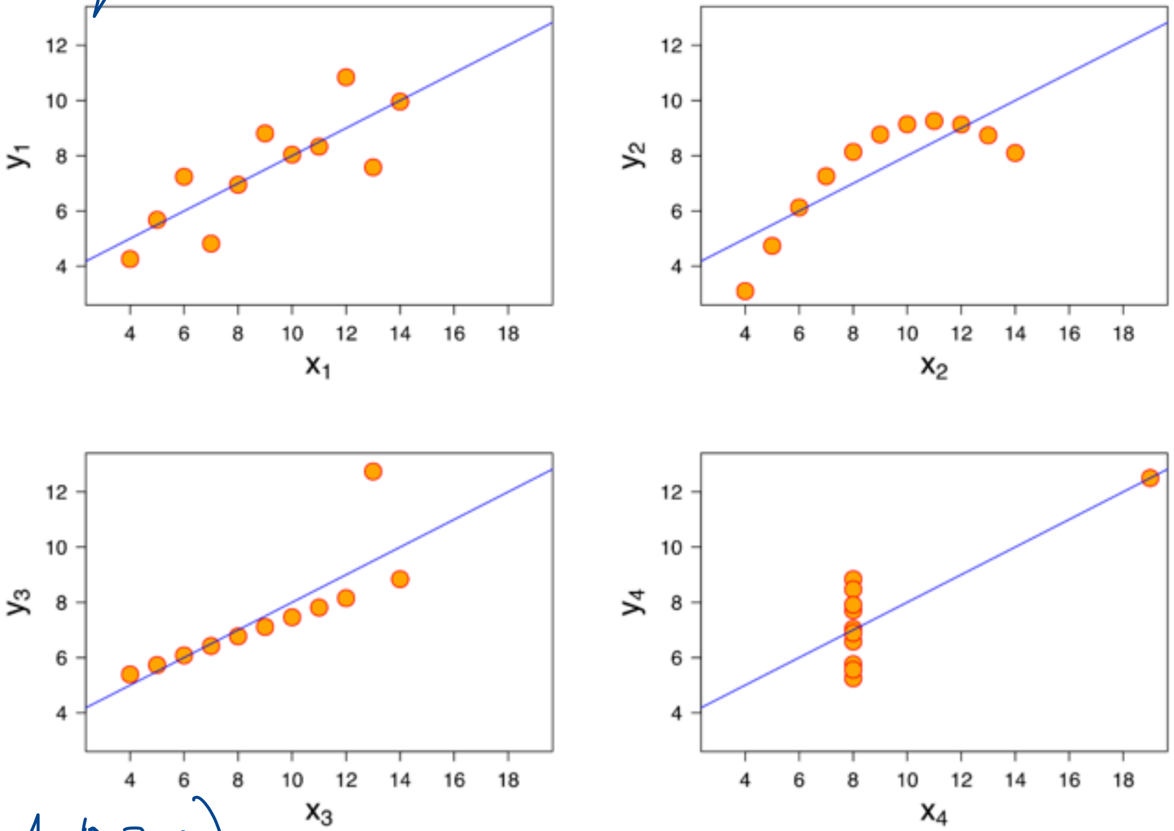
$$\sigma_x^2 = \frac{1}{n}\boxed{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = n\sigma_x^2$$

Dangers of correlation

same mean, std, correlation

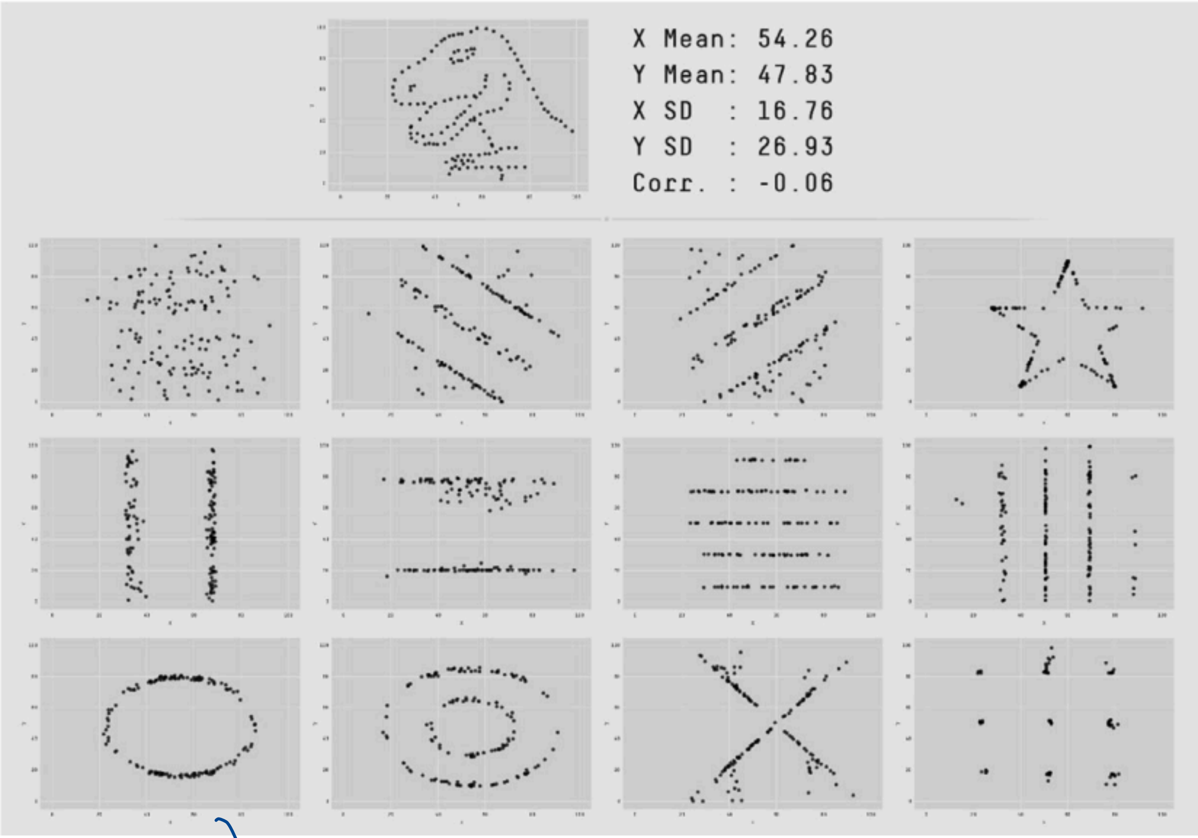
Anscombe's quartet



(Anscombe 1973)

Dangers of correlation

Datasauras dozen



(Matejka et al. 2017)

Interpreting the formulas

Interpreting the slope

no units

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

units of y

units of x

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

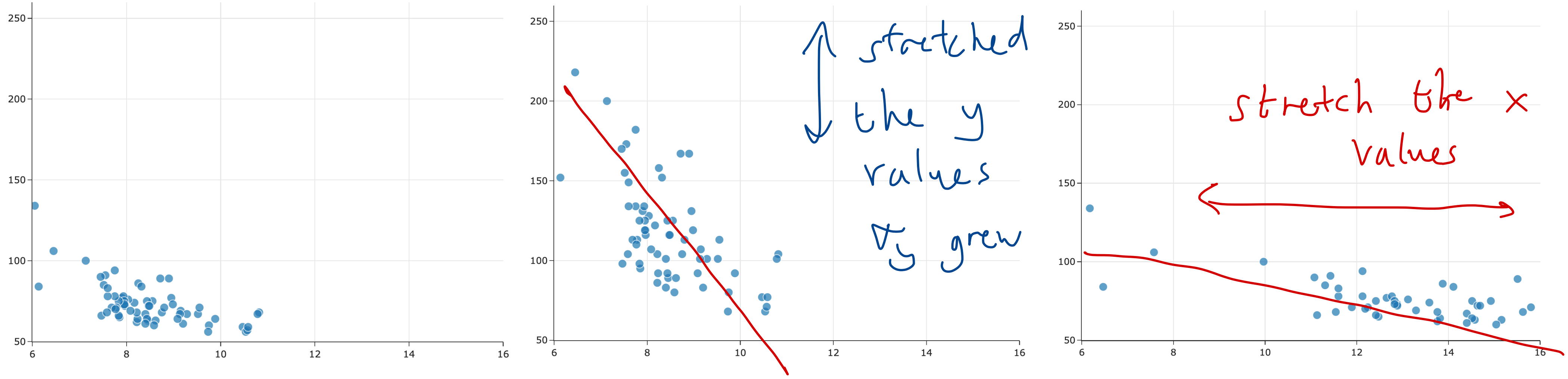
x departure time in hours

y commute time in minutes

r is the same in all plots

Interpreting the slope

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



- 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

reminder

$\bar{x} = \text{mean}\{x_i\} = \frac{1}{n}\sum_i x_i$

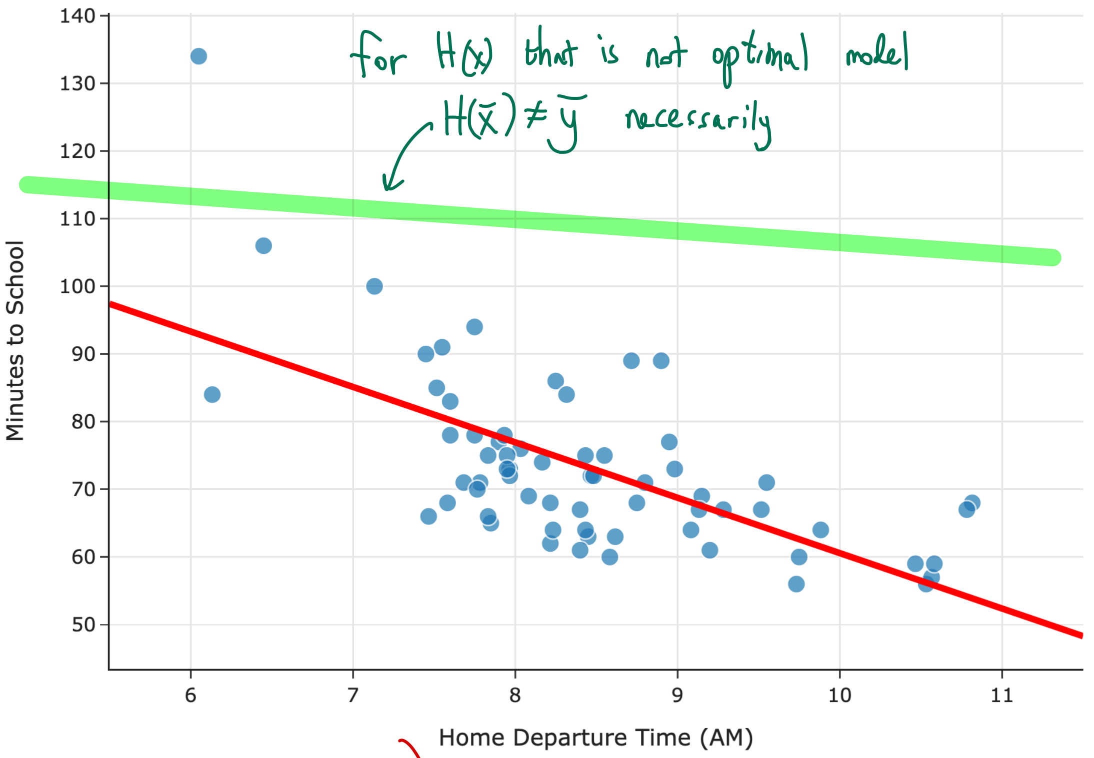
$\bar{y} = \text{mean}\{y_i\} = \frac{1}{n}\sum_i y_i$

units of y

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

$\frac{\text{units of } y}{\text{units of } x} \cdot (\text{units of } x) = \text{units of } y$

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



$H(x=0)$ = intercept

= predicted commute time @ midnight

- What are the units of the intercept?

units of y

- What is the value of $H^*(\bar{x})$?

$H^*(x_i) = \underbrace{w_0^*}_{} + w_1^* x_i$

$= \bar{y} - w_1^* \bar{x} + w_1^* x_i$

$= \bar{y} - w_1^* (\bar{x} - x_i)$

$H^*(\bar{x}) = \bar{y} - w_1^* (\underbrace{\cancel{\bar{x} - \bar{x}}}_{=0}) = \bar{y}$

$\boxed{H^*(\bar{x}) = \bar{y}}$

Question 🤔

Answer at [q.dsc40a.com](q.dsc40a.com)

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.