

Lecture 15

Simple Linear Regression

EECS 398-003: Practical Data Science, Fall 2024

practicaldsc.org • github.com/practicaldsc/fa24

Announcements

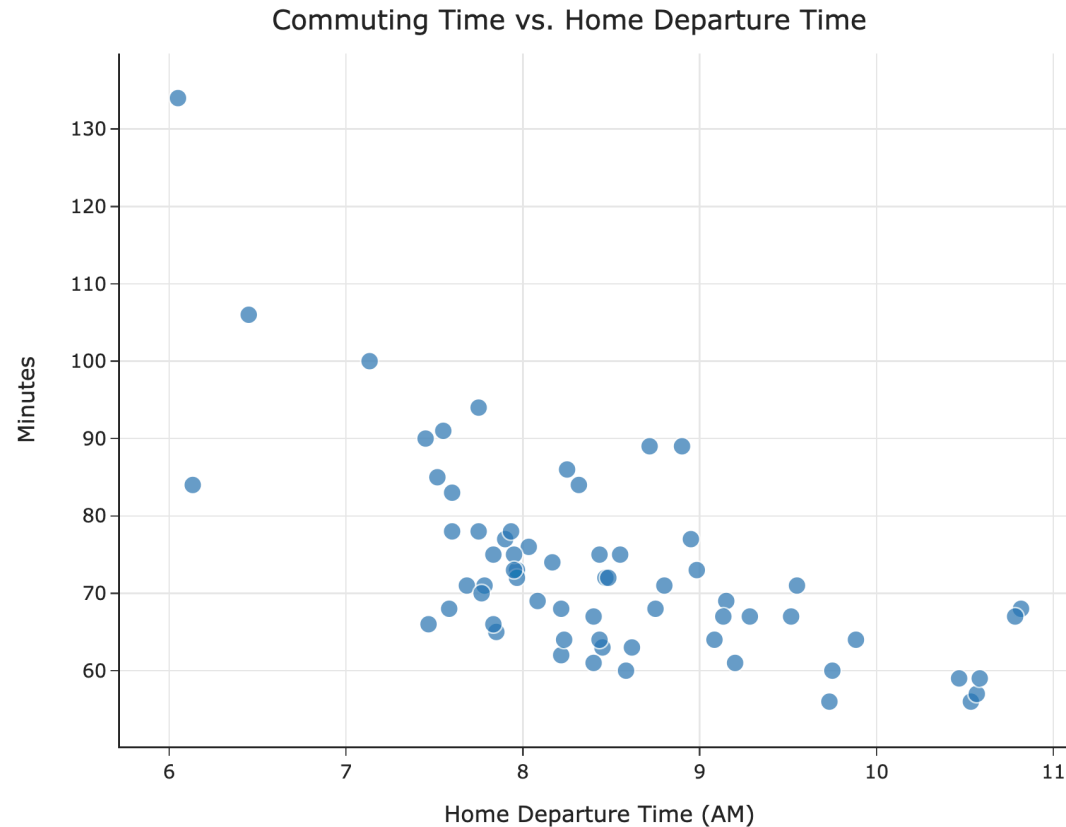
- Homework 7 is due on **Thursday**.
- We've released a Grade Report on Gradescope that has your current overall score in the class, scores on all assignments, and slip day usage so far.
See [#232 on Ed](#) for more details.
- Some updates to the **Syllabus**:
 - You now have 8 slip days instead of 6!
 - The final homework, called the Portfolio Homework, will be an open-ended investigation using the tools from both halves of the semester. Details to come.
 - You'll end up making a website!
 - You can work with a partner, but can't drop it or use slip days on it.
- The IA application is out for next semester! Please consider applying, and let me know if you're interested.

Agenda

- Recap: Models and loss functions.
- Towards simple linear regression.
- Minimizing mean squared error for the simple linear model.
- Correlation.
- Interpreting the formulas.
- Connections to related models.

Recap: Models and loss functions

Overview



- We started by introducing the idea of a hypothesis function, $H(x)$.
- We looked at two possible models:
 - The constant model, $H(x) = h$.
 - The simple linear regression model, $H(x) = w_0 + w_1x$.
- We decided to find the **best constant prediction** to use for predicting commute times, in minutes.

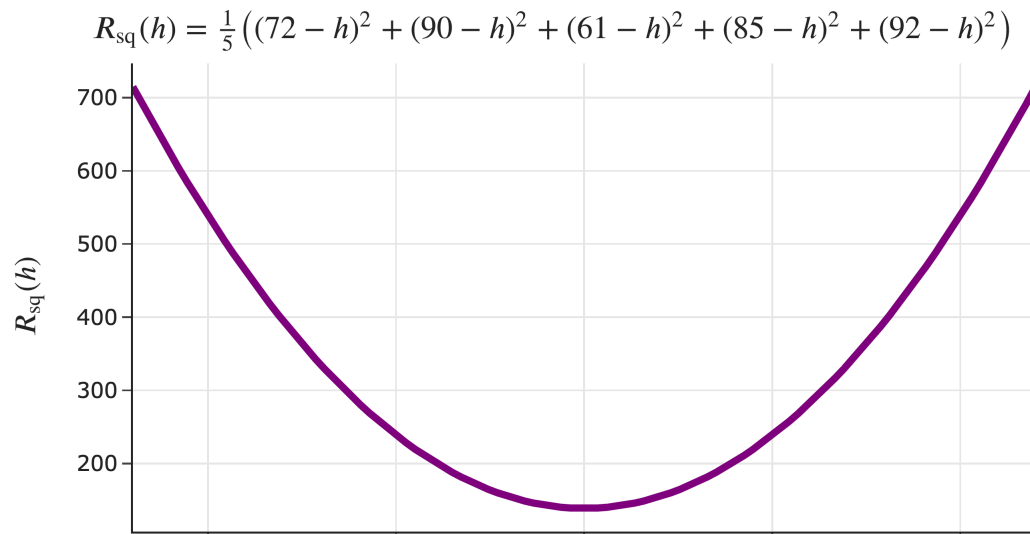
The modeling recipe

- We've now made two full passes through our modeling recipe.
 1. Choose a model.
 2. Choose a loss function.
 3. Minimize average loss to find optimal model parameters.

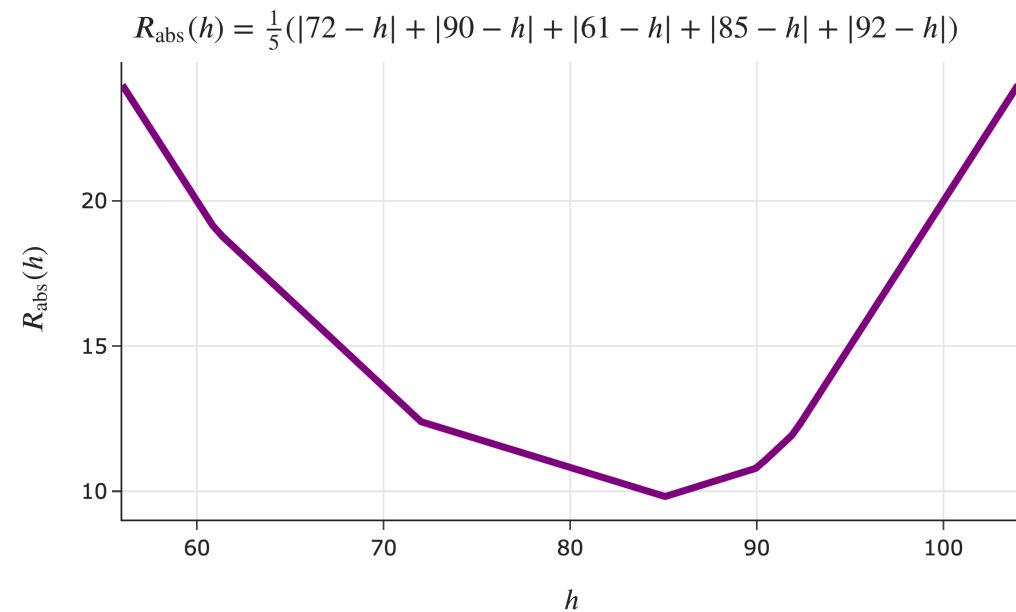
Visualizing average loss

Consider the example dataset 72, 90, 61, 85, 92.

Mean squared error



Mean absolute error



Empirical risk minimization

- The formal name for the process of minimizing average loss is **empirical risk minimization**.
- Another name for "average loss" is **empirical risk**.
- When we use the squared loss function, $L_{\text{sq}}(y_i, h) = (y_i - h)^2$, the corresponding empirical risk is mean squared error:

$$R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2 \implies h^* = \text{Mean}(y_1, y_2, \dots, y_n)$$

- When we use the absolute loss function, $L_{\text{abs}}(y_i, h) = |y_i - h|$, the corresponding empirical risk is mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h| \implies h^* = \text{Median}(y_1, y_2, \dots, y_n)$$

Empirical risk minimization, in general

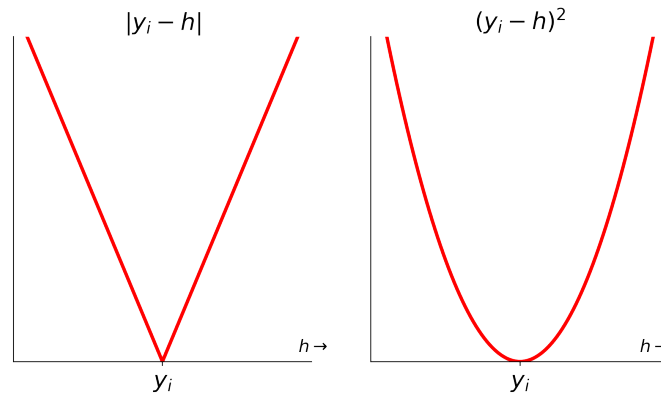
- **Key idea:** If L is **any** loss function, and H is any hypothesis function, the corresponding empirical risk is:

$$R(H) = \frac{1}{n} \sum_{i=1}^n L(y_i, H(x_i))$$

- In Homework 7 (and last week's discussion), you saw several examples in which:
 - You were given a new loss function L .
 - You had to find the optimal parameter h^* for the constant model $H(x_i) = h$.

Choosing a loss function

- For the constant model $H(x) = h$, the **mean** minimizes mean **squared** error.
- For the constant model $H(x) = h$, the **median** minimizes mean **absolute** error.
- In practice, squared loss is the more common choice, as it's easily **differentiable**.



- But how does our choice of loss function impact the resulting optimal prediction?

Comparing the mean and median

- Consider our example dataset of 5 commute times.

$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 92$$

- As of now, the **median is 85** and the **mean is 80**.
- What if we add 200 to the largest commute time, 92?

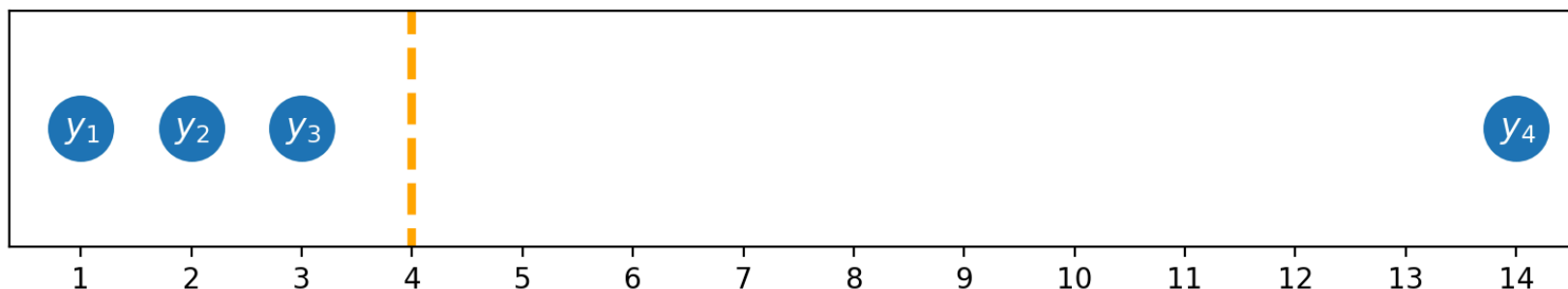
$$y_1 = 72 \qquad y_2 = 90 \qquad y_3 = 61 \qquad y_4 = 85 \qquad y_5 = 292$$

- Now, the median is 10 but the mean is 12.5!
- **Key idea:** The mean is quite **sensitive** to outliers.

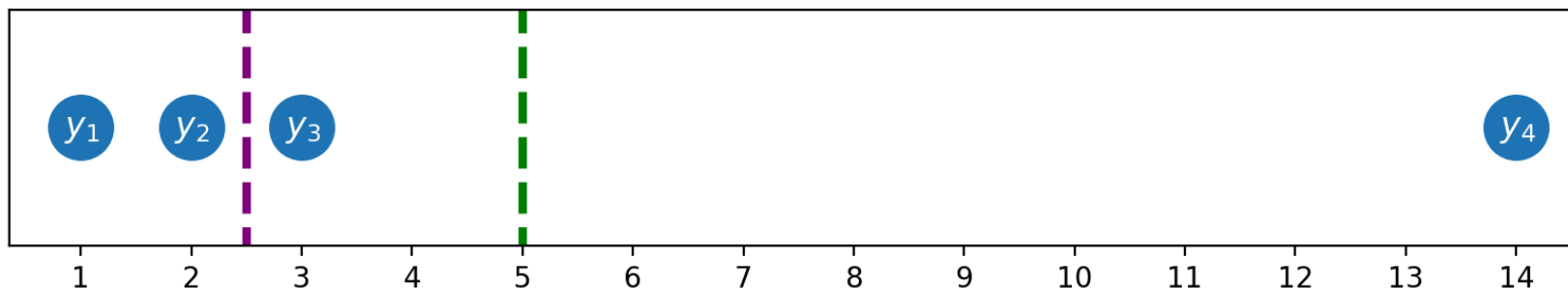
But why?

Outliers

- Below, $|y_4 - h|$ is 10 times as big as $|y_3 - h|$, but $(y_4 - h)^2$ is 100 times $(y_3 - h)^2$.

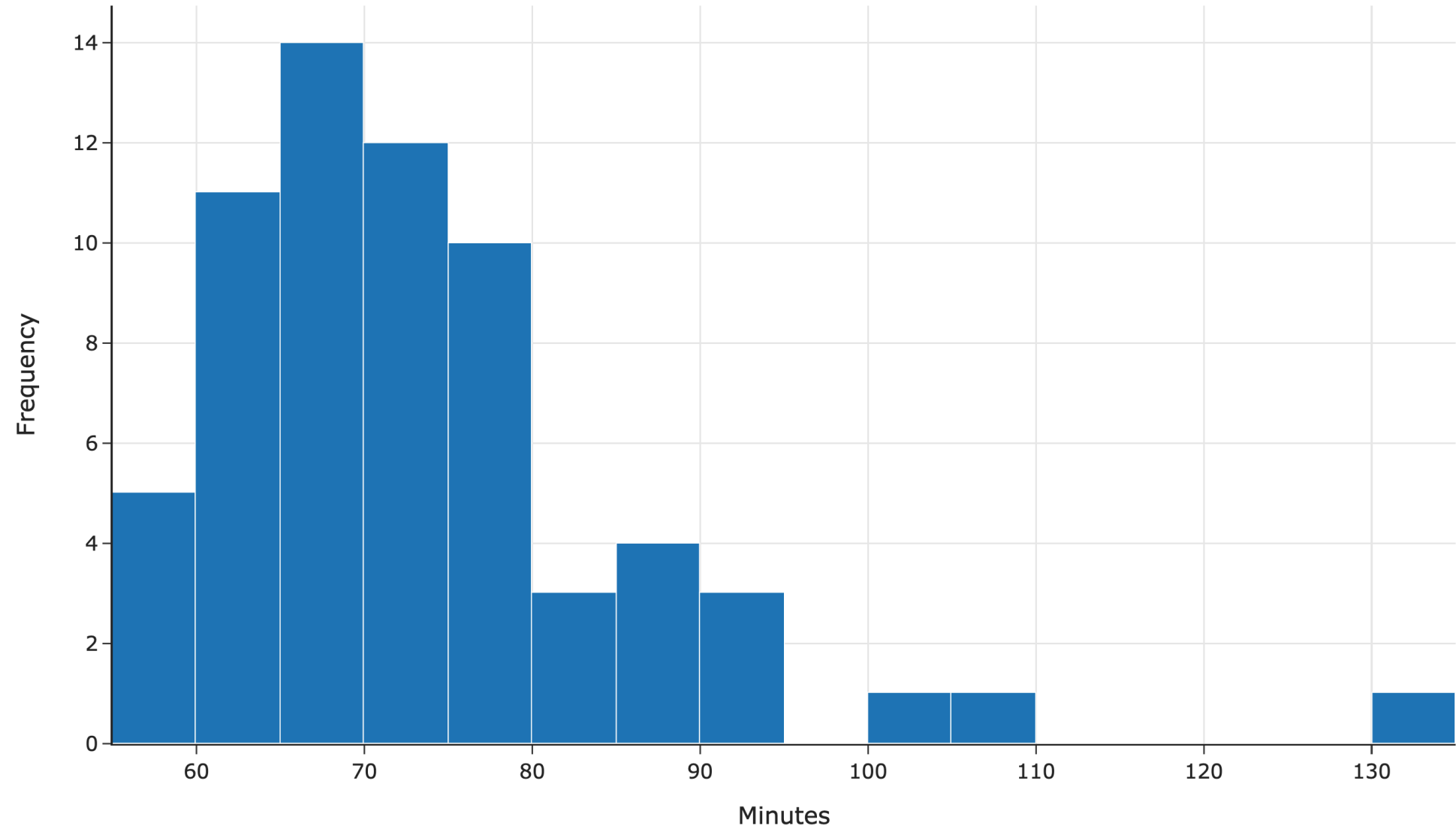


- The result is that the **mean** is "pulled" in the direction of outliers, relative to the **median**.



- As a result, we say the **median** – and absolute loss more generally – is **robust**.

Distribution of Commuting Time

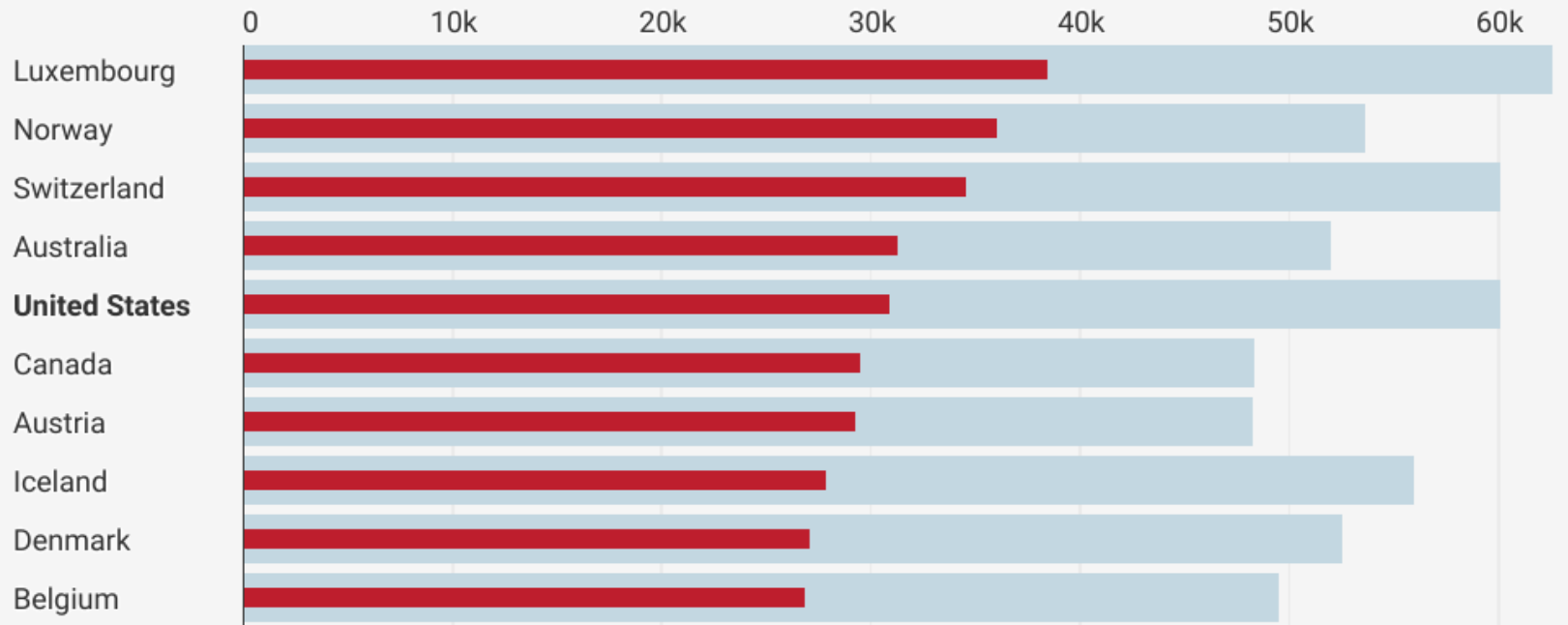


Example: Income inequality

Average vs median income

Median and mean income between 2012 and 2014 in selected OECD countries, in USD; weighted by the currencies' respective [purchasing power](#) (PPP).

■ Average income in USD ■ Median income



Summary: Choosing a loss function

- **Key idea:** Different loss functions lead to different best predictions, h^* !

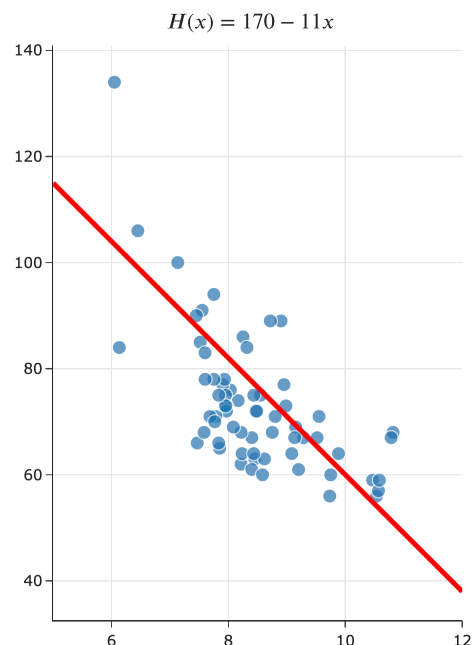
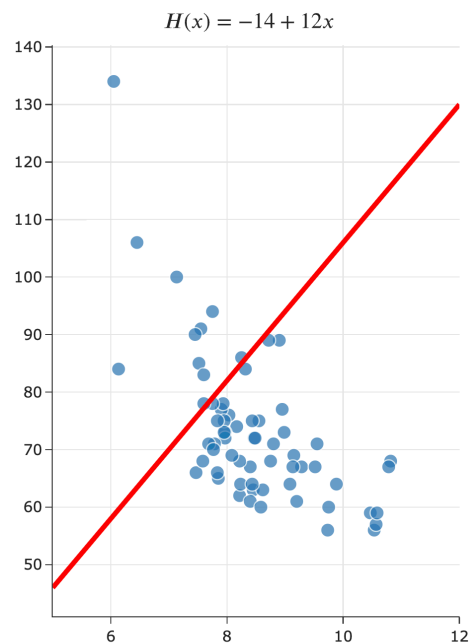
Loss	Minimizer	Always Unique?	Robust to Outliers?	Differentiable?
$L_{\text{sq}}(y_i, h) = (y_i - h)^2$	mean	yes	no	yes
$L_{\text{abs}}(y_i, h) = y_i - h $	median	no	yes	no
$L_{0,1}(y_i, h) = \begin{cases} 0 & y_i = h \\ 1 & y_i \neq h \end{cases}$	mode	no	yes	no
$L_{\infty}(y_i, h)$ See HW 7, Question 5.	???	yes	no	no

- The optimal predictions, h^* , are all **summary statistics** that measure the **center** of the dataset in different ways.

Towards simple linear regression

Recap: Hypothesis functions and parameters

- A hypothesis function, H , takes in an x as input and returns a predicted y .
- **Parameters** define the relationship between the input and output of a hypothesis function.
- **Example:** The simple linear regression model, $H(x) = w_0 + w_1x$, has two parameters: w_0 and w_1 .



The modeling recipe

1. Choose a model.
2. Choose a loss function.
3. Minimize average loss to find optimal model parameters.

Minimizing mean squared error for the simple linear model

- We'll choose squared loss, since it's the easiest to minimize.
- Our goal, then, is to find the linear hypothesis function $H^*(x)$ that minimizes empirical risk:

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

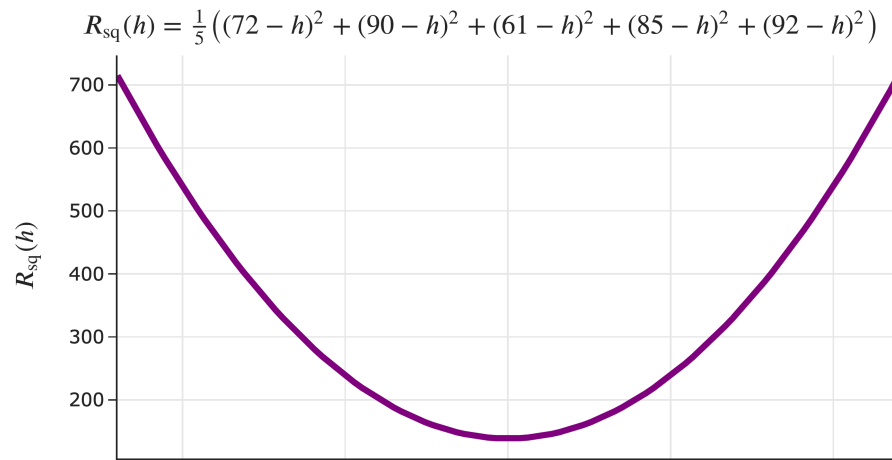
- Since linear hypothesis functions are of the form $H(x) = w_0 + w_1x$, we can re-write R_{sq} as a function of w_0 and w_1 :

$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^n (y_i - (w_0 + w_1x_i))^2$$

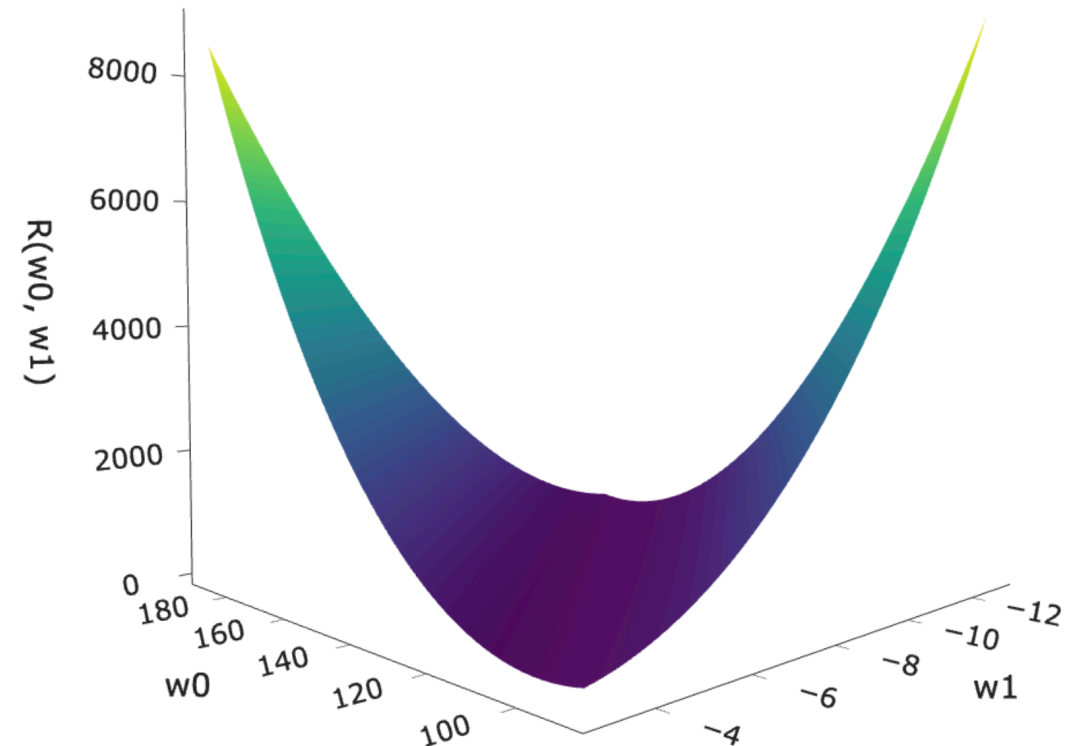
- How do we find the parameters w_0^* and w_1^* that minimize $R_{\text{sq}}(w_0, w_1)$?

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:

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

- 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.
 - Set all partial derivatives to 0 and 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).
- To save time, we won't do the derivation live in class, but you are responsible for it!
[Here's a video](#) of me walking through it.

Example

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

$$f(x, y) = x^2 - 8x + y^2 + 6y - 7$$

Minimizing mean squared error

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

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.

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

$$\frac{\partial R_{\text{sq}}}{\partial w_0} =$$

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

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

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."

Solving for w_0^*

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

Solving for w_1^*

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

Least squares solutions

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

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

where:

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

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

An equivalent formula for w_1^*

- Claim:

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

- Proof:

Least squares solutions

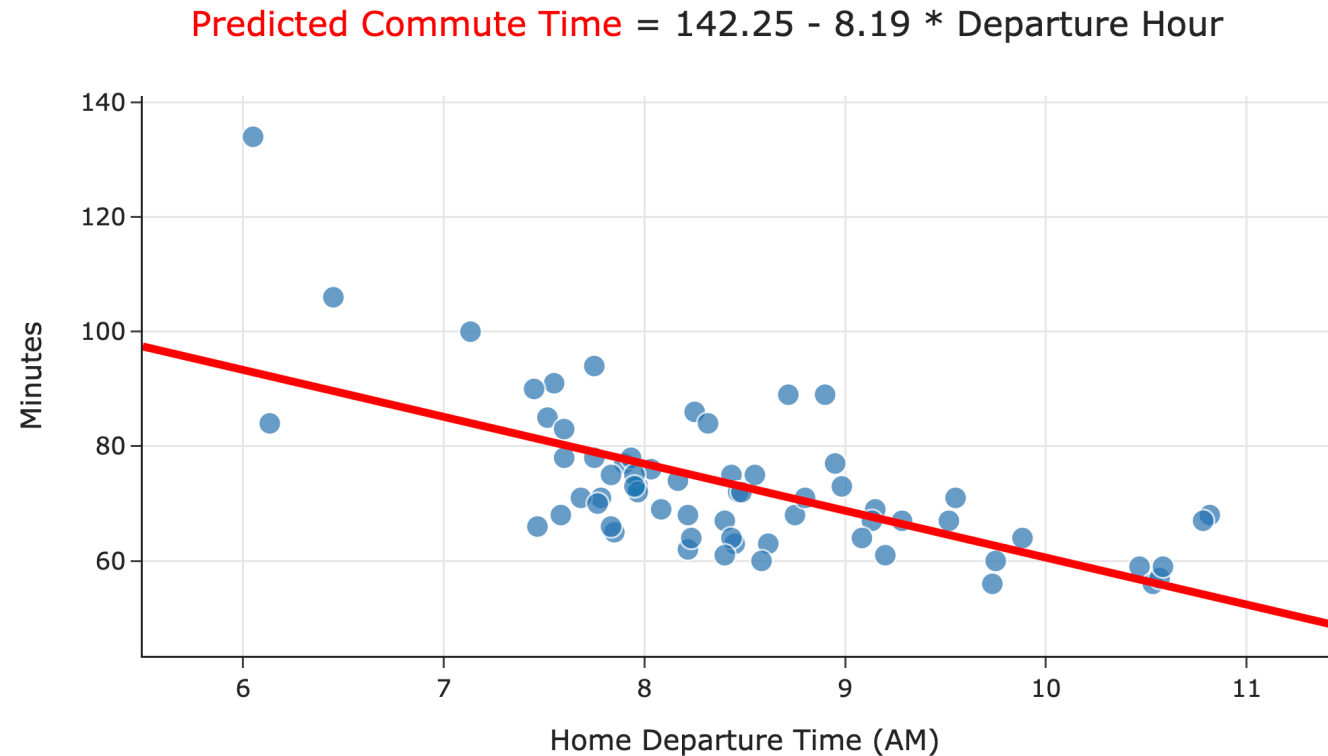
- The **least squares solutions** for the intercept w_0 and slope w_1 are:

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

- We say w_0^* and w_1^* are **optimal parameters**, and the resulting line is called the **regression line**.
- 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 $H^*(x) = w_0^* + w_1^*x$.

Code demo

- Let's test these formulas out in code!

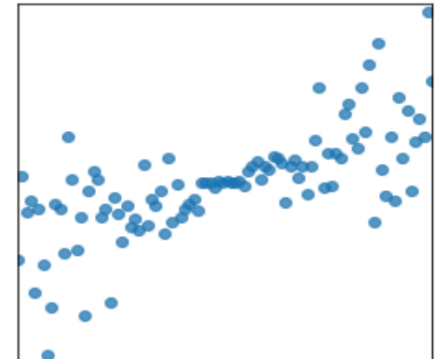
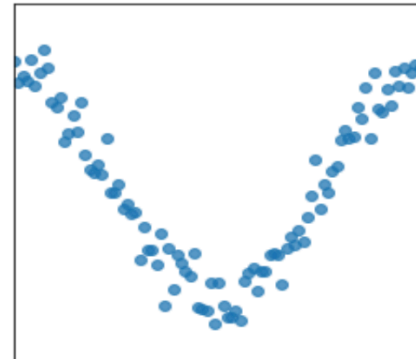
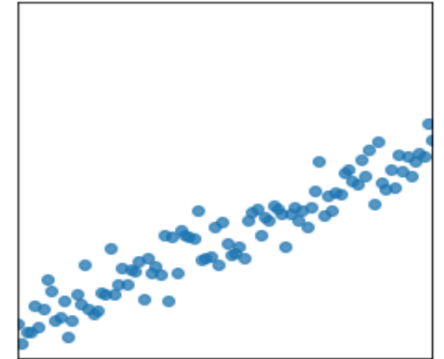
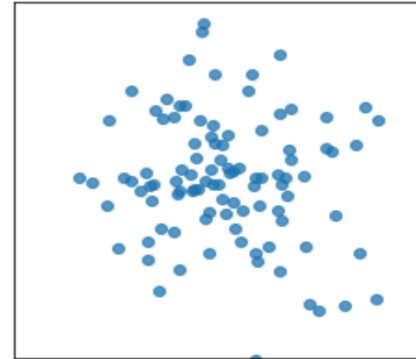


- The supplementary notebook is posted in the usual place on [GitHub](#) and the [course website](#).

Correlation

Quantifying patterns in scatter plots

- The **correlation coefficient**, r , 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.



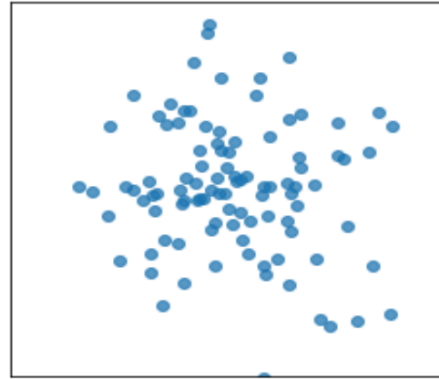
The correlation coefficient

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

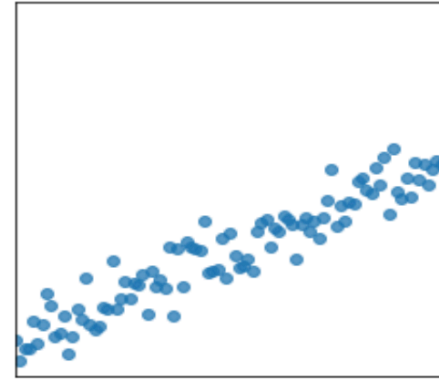
$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

The correlation coefficient, visualized

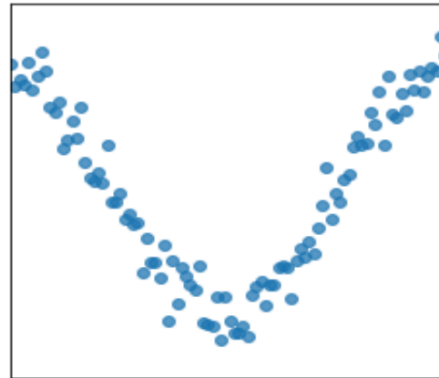
$r = -0.121$



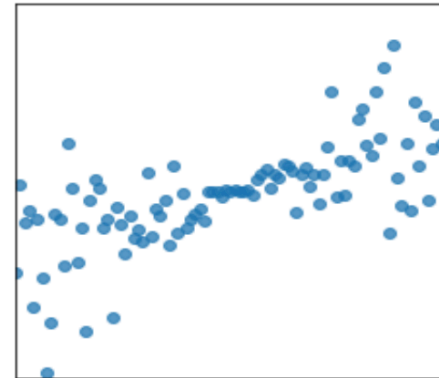
$r = 0.949$



$r = 0.052$



$r = 0.704$



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 !

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

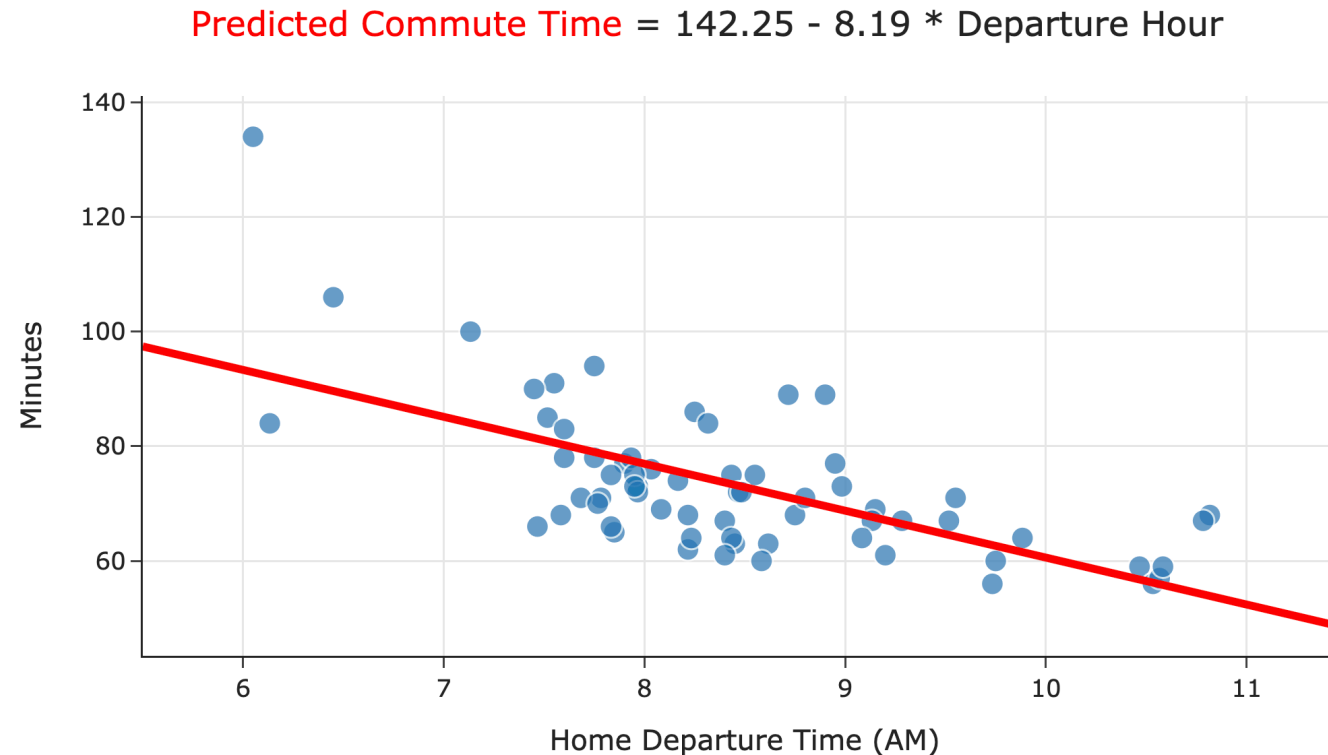
- It's not surprising that r is related to w_1^* , since r is a measure of linear association.
- Concise way of writing w_0^* and w_1^* :

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

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

Code demo

- Let's test these new formulas out in code and see if they match the earlier formulas!

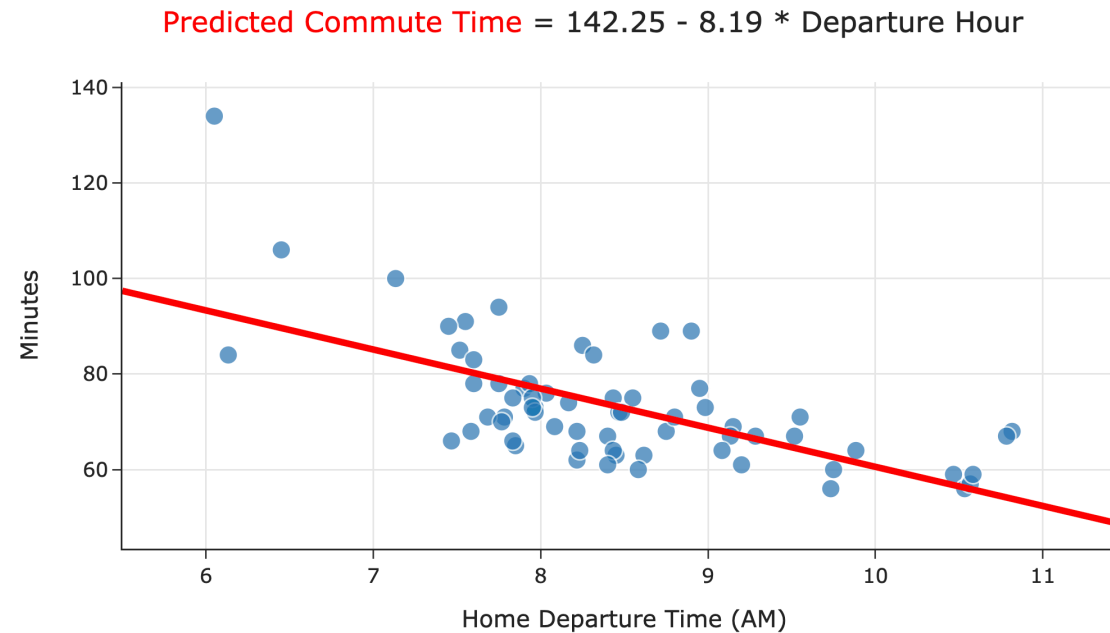


- The supplementary notebook is posted in the usual place on [GitHub](#) and the [course website](#).

Interpreting the formulas

Causality

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



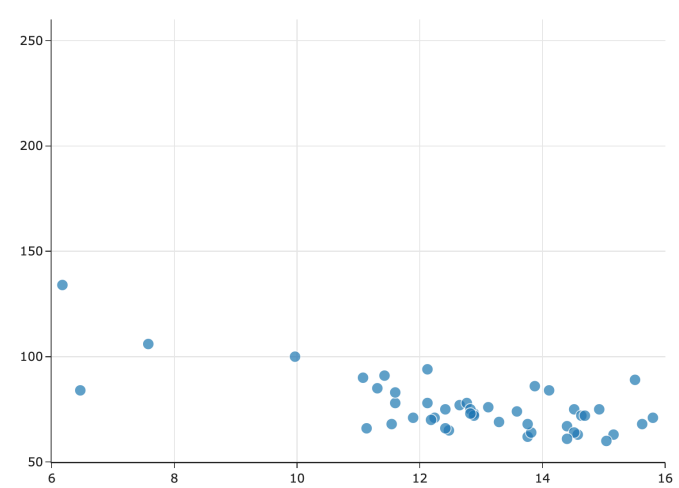
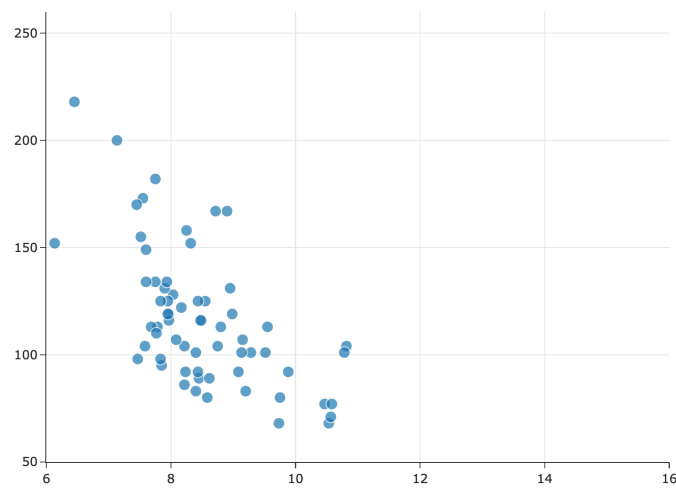
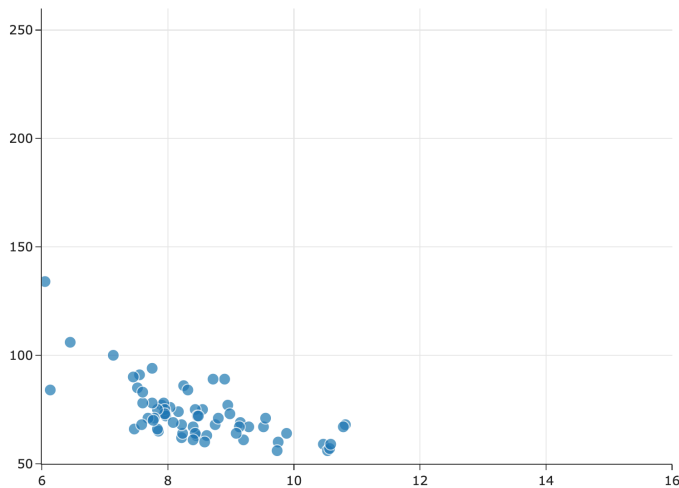
Interpreting the slope

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

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

Interpreting the slope

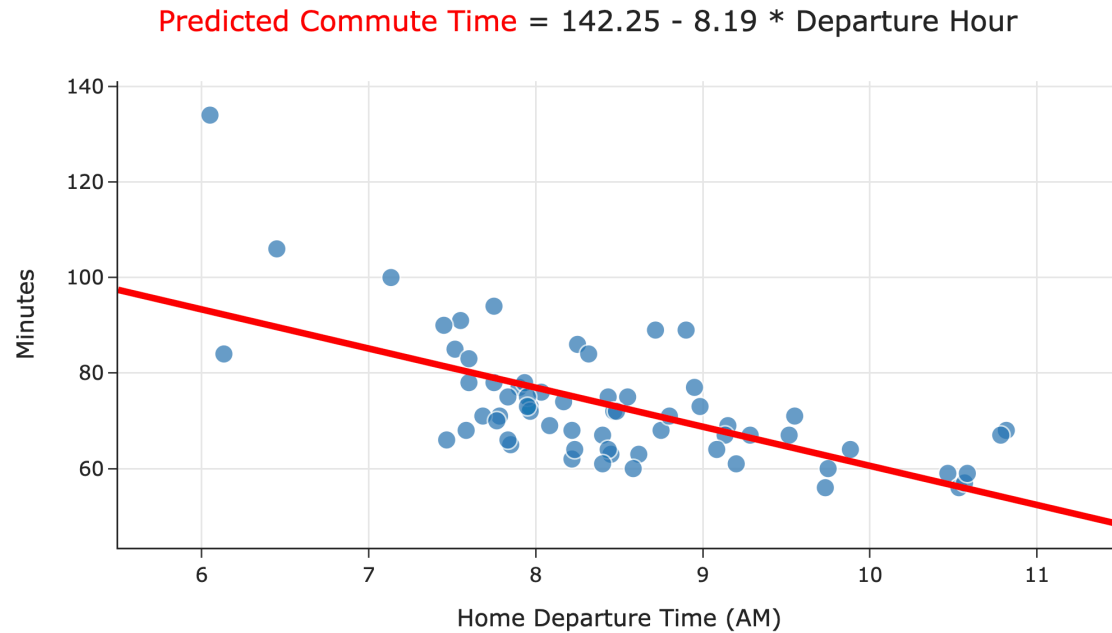
$$w_1^* = r \frac{\sigma_y}{\sigma_x}$$



- 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

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



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

Question 🤔

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.

Question 🤔

Answer at practicaldsc.org/q

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. $w_0^* = -5, w_1^* = 5$

Connections to related models

Question 🤔

Answer at practicaldsc.org/q

Suppose we chose the model $H(x) = w_1x$ and squared loss.

What is the optimal model parameter, w_1^* ?

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

- B. $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

- C. $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

- D. $\frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i}$

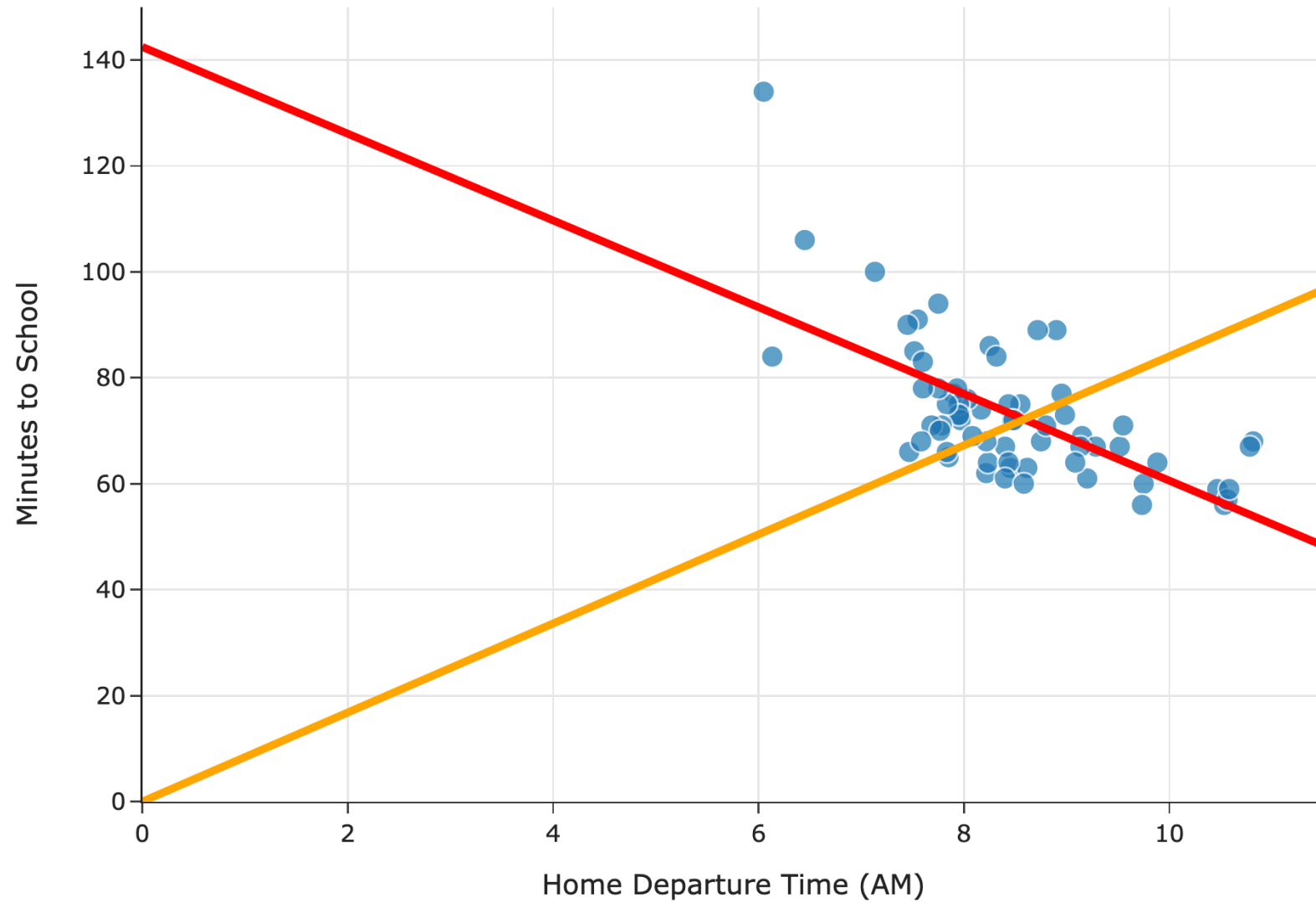
Exercise

Suppose we chose the model $H(x) = w_1 x$ and squared loss.

What is the optimal model parameter, w_1^* ?

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

Predicted Commute Time = $8.41 * \text{Departure Hour}$



Exercise

Suppose we choose the model $H(x) = w_0$ and squared loss.

What is the optimal model parameter, w_0^* ?

Comparing mean squared errors

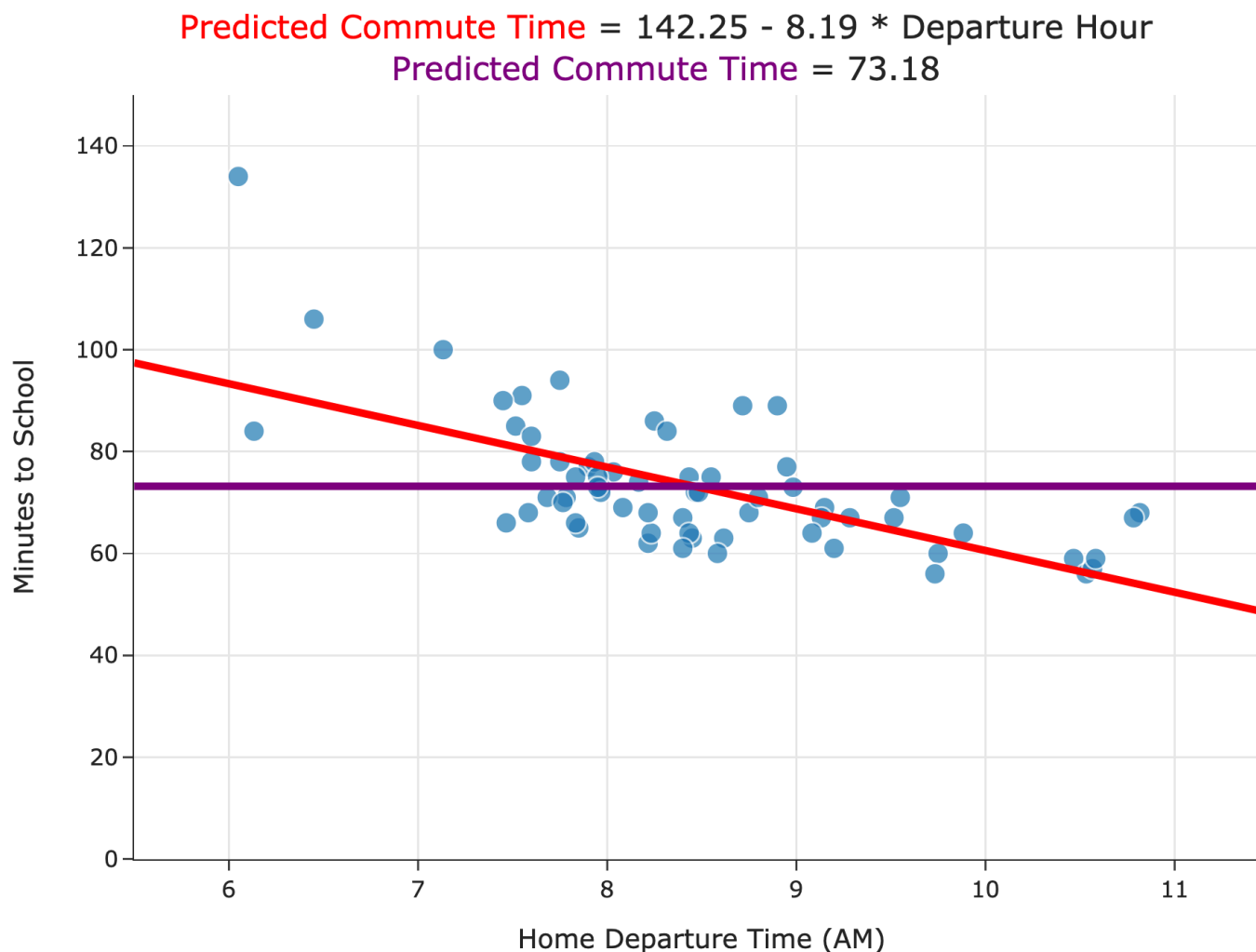
- With both:
 - the constant model, $H(x) = h$, and
 - the simple linear regression model, $H(x) = w_0 + w_1x$,

when we chose squared loss, we minimized mean squared error to find optimal parameters:

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

- Which model minimizes mean squared error more?

Comparing mean squared errors



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

- The MSE of the best **simple linear regression model** is ≈ 97 .
- The MSE of the best **constant model** is ≈ 167 .
- The **simple linear regression model** is a more flexible version of the **constant model**.