Lecture 16

Regression using Linear Algebra

EECS 398-003: Practical Data Science, Fall 2024

practicaldsc.org • github.com/practicaldsc/fa24

Announcements 📣

- Homework 7 is due **tonight**.
- 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.

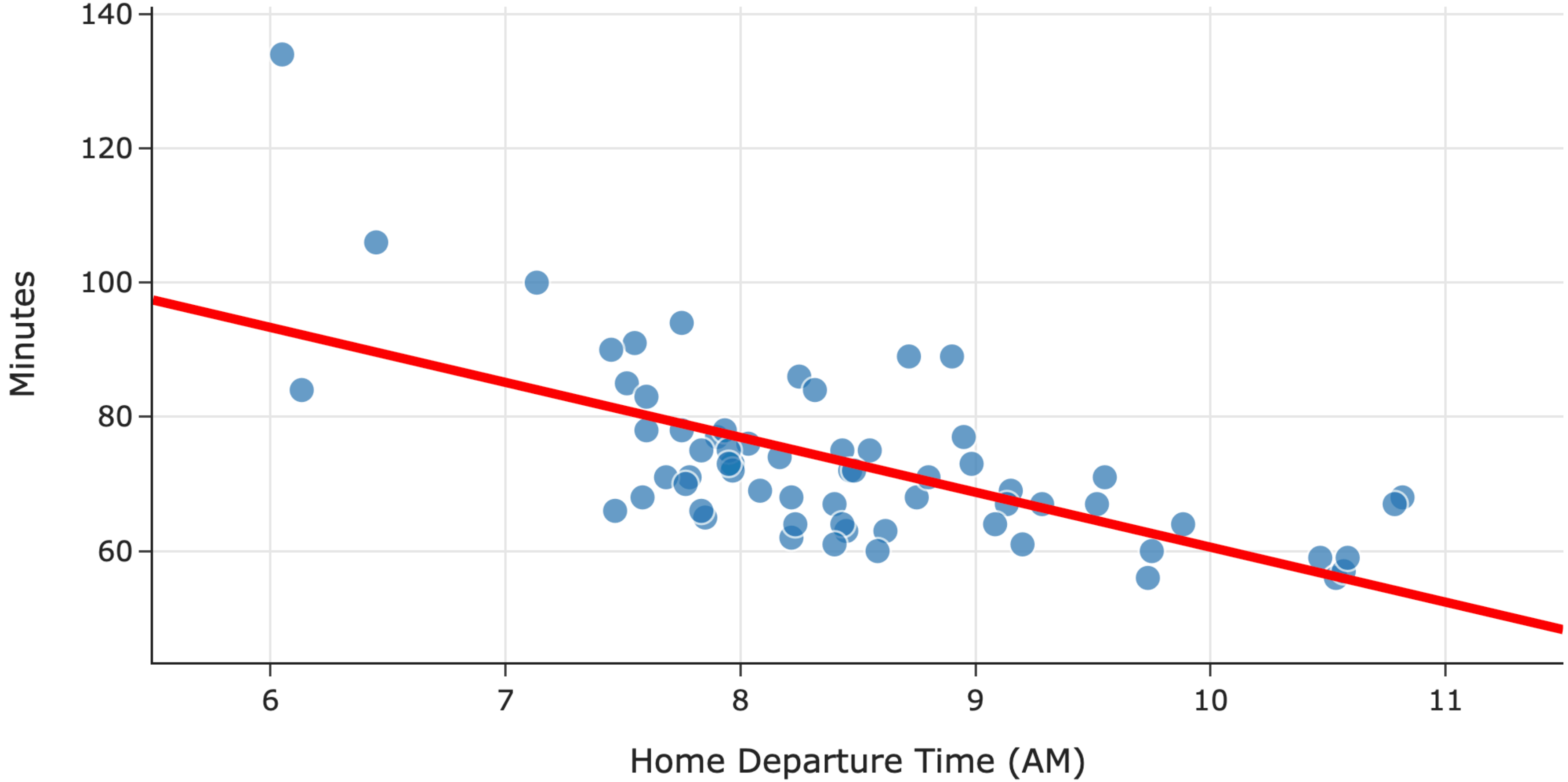
  See **#232 on Ed** for more details.
- Some updates to the **Syllabus**:
  - You now have 8 slip days instead of 6!
  - The final homework, called the Portfolio Homework, will be an open-ended investigation using the tools from both halves of the semester. Details to come.
    - You'll end up making 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! See **#238 on Ed** for more details.

Agenda

- Recap: Simple linear regression.
- Interpreting the formulas.
- Connections to related models.
- Regression and linear algebra.
- Multiple linear regression.

Recap: Simple linear regression

Predicted Commute Time = 142.25 - 8.19 * Departure Hour



Last lecture, we said that the line in **red** is the regression line.

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

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

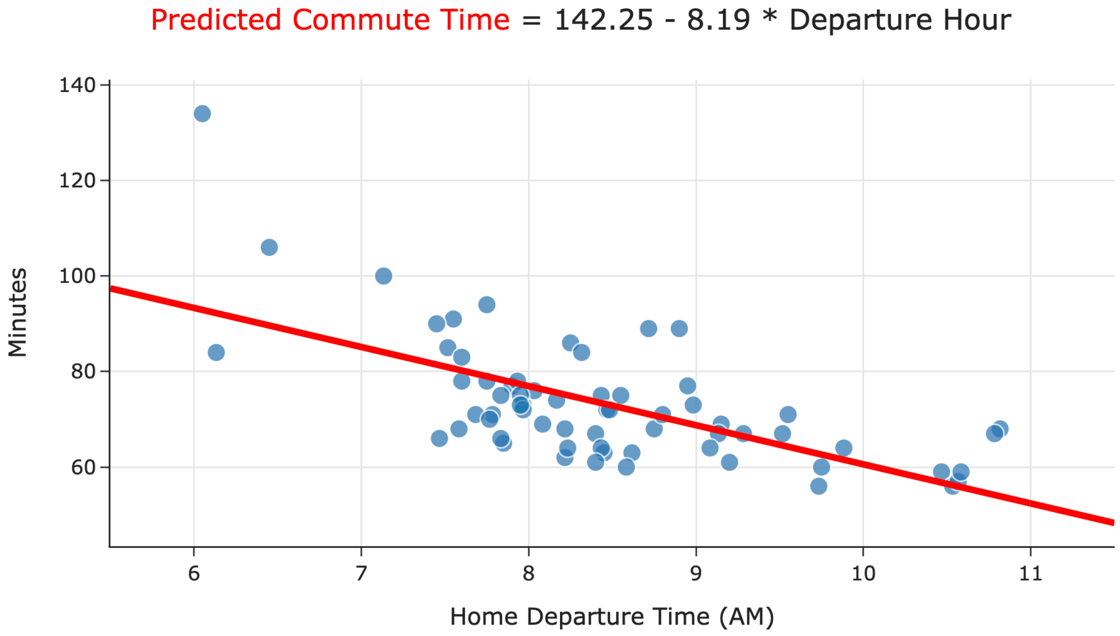
- The resulting line, $H^*(x) = w_0^* + w_1^* x$, is the line that minimizes mean squared error.

It's often called the **(least squares) regression line**, and the **optimal linear predictor**.

Interpreting the formulas

Causality

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



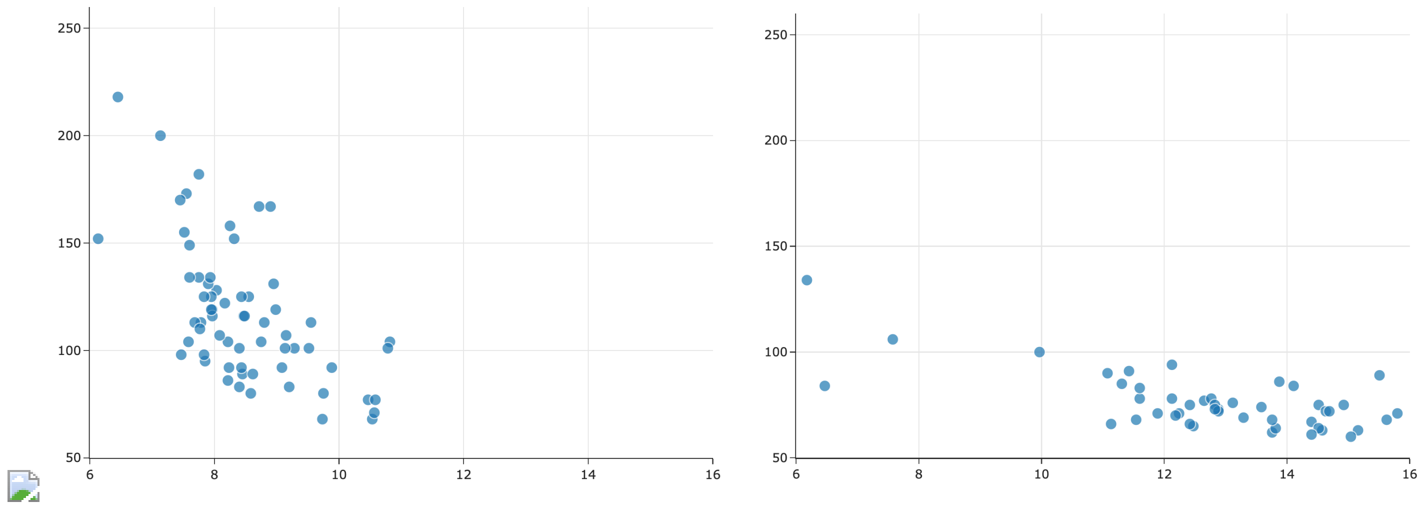
Interpreting the slope

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

- 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

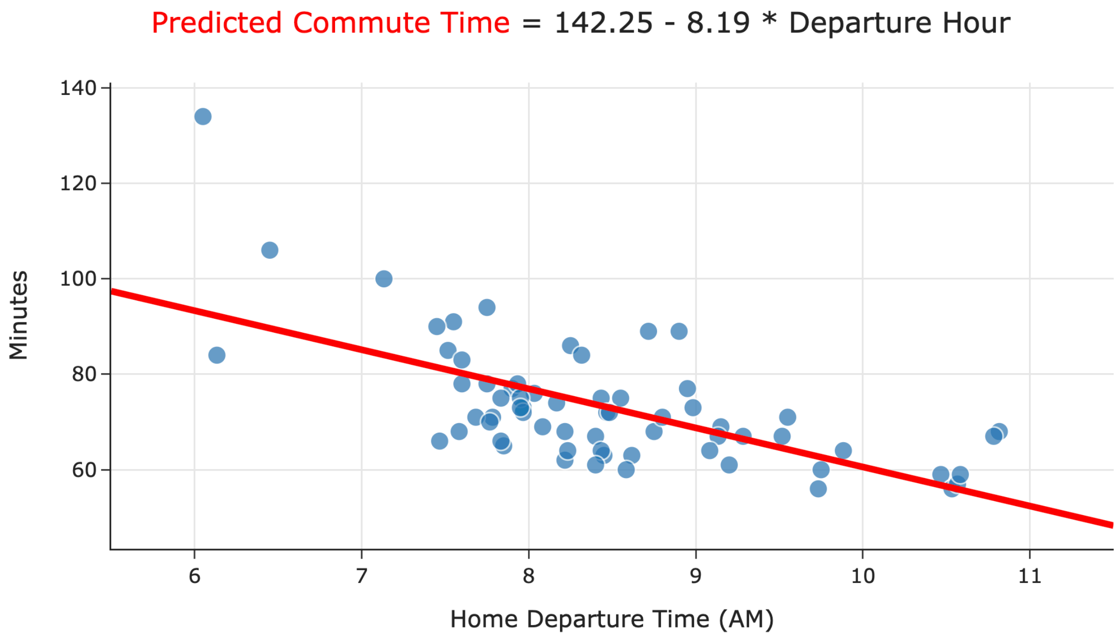
$$w_1^* = r\frac{\sigma_y}{\sigma_x}$$



- 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 $H(x) = w_1 x$ and squared loss.

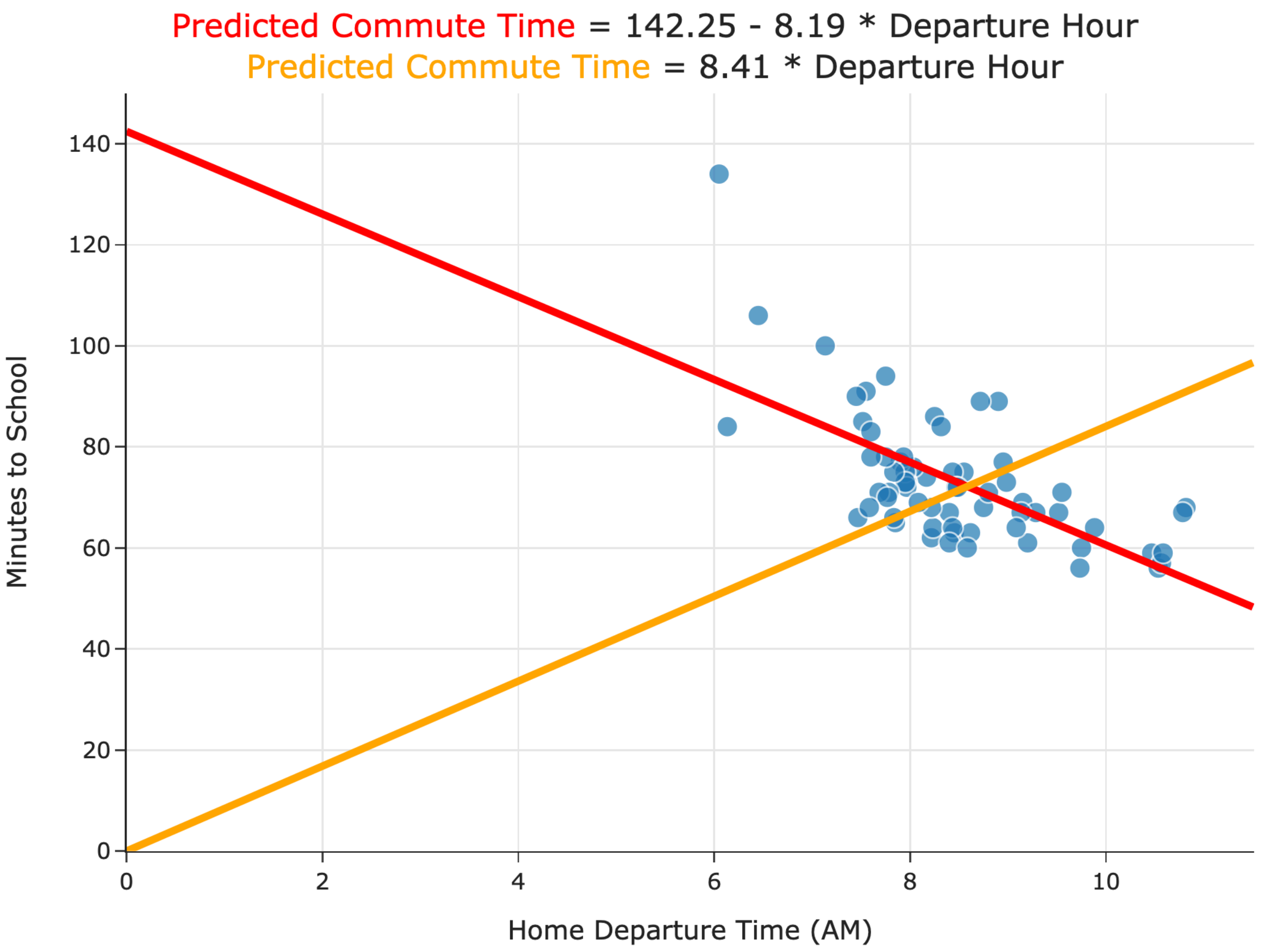
What is the optimal model parameter, w_1^*?

- 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 * Departure Hour
Predicted Commute Time = 8.41 * Departure Hour
Minutes to School
Home Departure Time (AM)
0
20
40
60
80
100
120
140
0
2
4
6
8
10

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

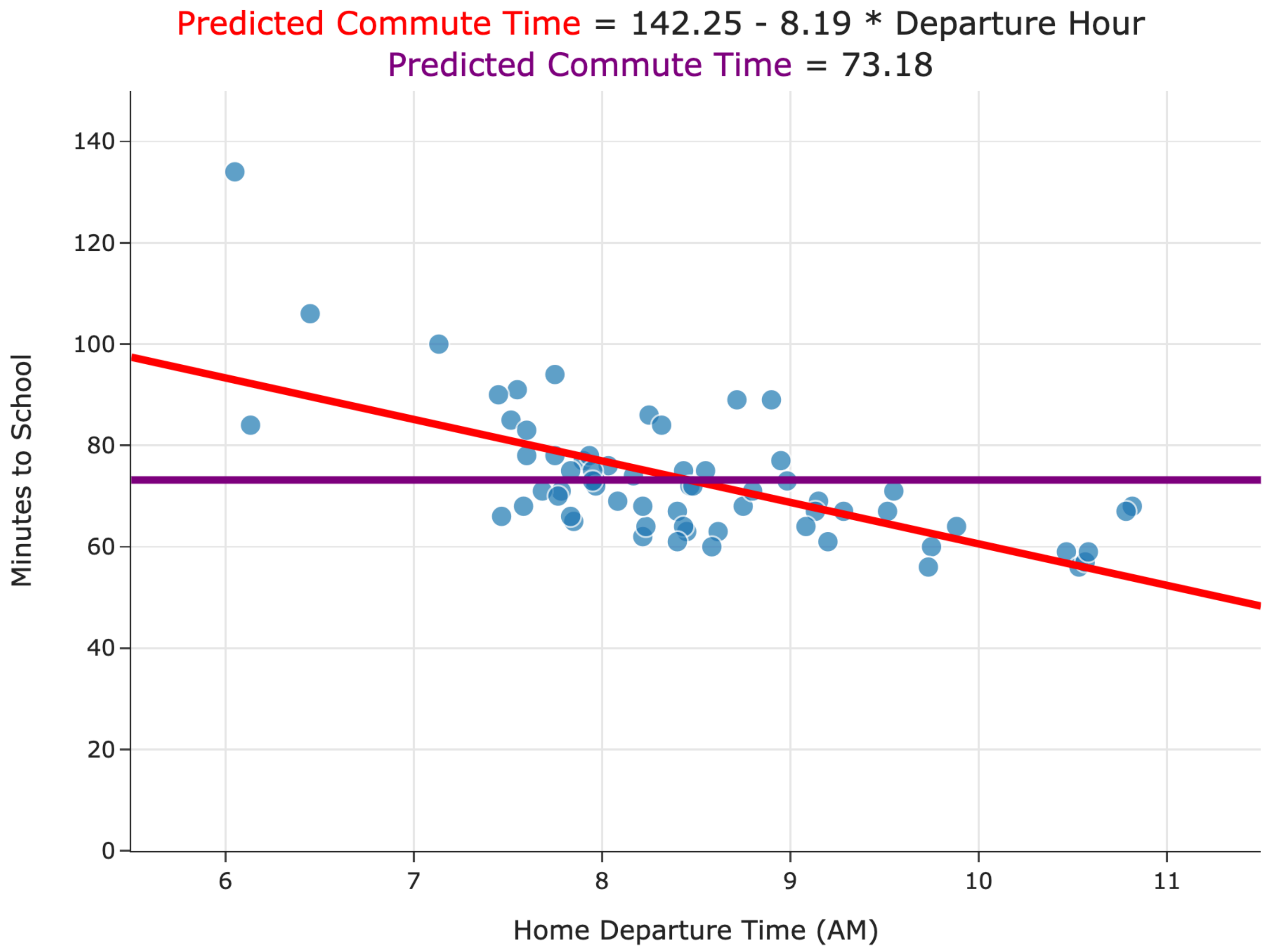
- With both:
  - the constant model, $H(x) = h$, and
  - the simple linear regression model, $H(x) = w_0 + w_1 x$,

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

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

- **Which model minimizes mean squared error more?**

Comparing mean squared errors



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

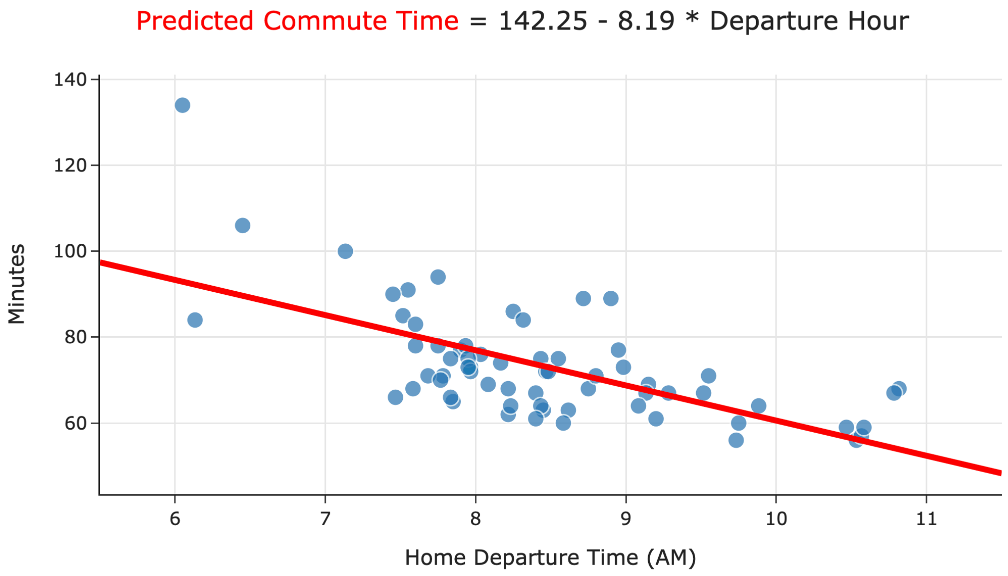
- The MSE of the best simple linear regression model is ≈ 97.
- The MSE of the best constant model is ≈ 167.
- The simple linear regression model is a more flexible version of the constant model.

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).
    - Are non-linear in the features, e.g. $H(x) = w_0 + w_1 x + w_2 x^2$.

Simple linear regression, revisited



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

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

- **Observation**: $R_{\text{sq}}(w_0, w_1)$ *kind of* looks like the formula for the norm of a vector,

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

Regression and linear algebra

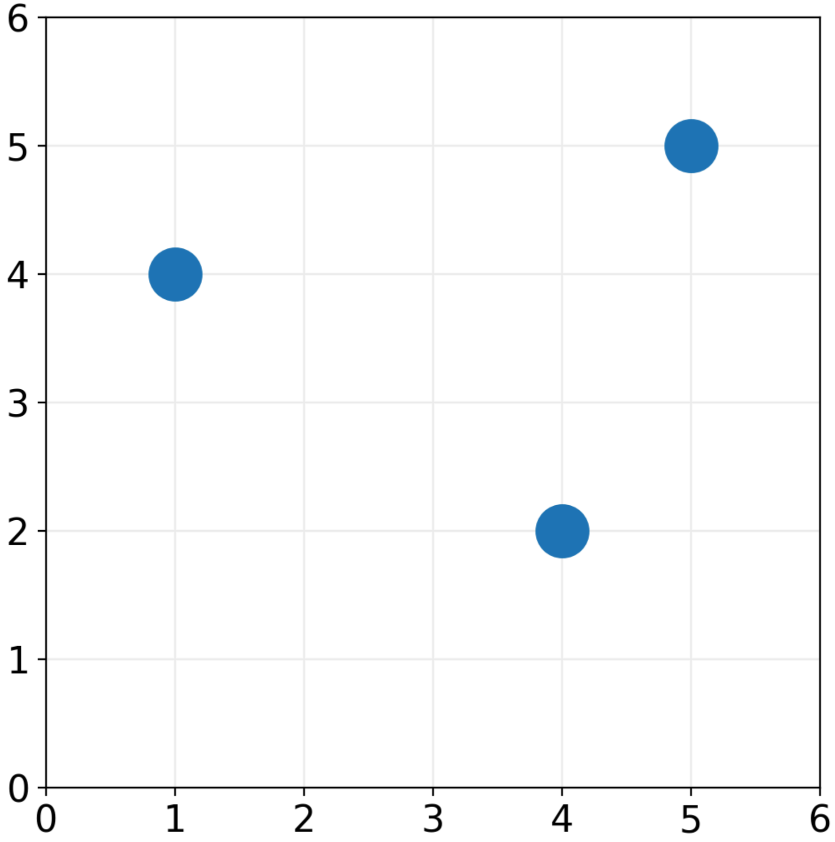
Let's define a few new terms:

- 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 $H(x) = 2 + \frac{1}{2}x$.



$\vec{y} =$

$\vec{h} =$

$\vec{e} = \vec{y} - \vec{h} =$

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

$$=$$

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:

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

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

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

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:

$$\vec{h} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix} =$$

Rewriting the mean squared error

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

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

- Define the **parameter vector** $\vec{w} \in \mathbb{R}^2$ to be $\vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$.

- Then, $\vec{h} = X\vec{w}$, so the mean squared error becomes:

$$R_{\text{sq}}(H) = \frac{1}{n} \|\vec{y} - \vec{h}\|^2 \implies \boxed{R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2}$$

Minimizing mean squared error, again

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

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

- 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:

$$\boxed{R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2}$$

- Do we already know the $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w})$?

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:

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

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

$$\|\vec{y} - X\vec{w}\|^2$$

- 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 $\text{span}(X)$!
- **We already did this in LARDS, Section 8!**

An optimization problem we've seen before

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

$$R_{\text{sq}}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

- 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 $\|\vec{e}\|$ is the same as the minimizer of $R_{\text{sq}}(\vec{w})$.

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

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

The normal equations

- The normal equation**s** are the system of 2 equations and 2 unknowns defined by:

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

- 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, $\vec{w}^*$

- To find the optimal model parameters for simple linear regression, w_0^* and w_1^*, we previously minimized $R_{\text{sq}}(w_0, w_1) = \frac{1}{n}\sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$.
  - We found, using calculus, that:
    - $\boxed{w_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r\frac{\sigma_y}{\sigma_x}}$.
    - $\boxed{w_0^* = \bar{y} - w_1^* \bar{x}}$.
- Another way of finding optimal model parameters for simple linear regression is to find the $\vec{w}^*$ that minimizes $R_{\text{sq}}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$.
  - The minimizer, if X^TX is invertible, is the vector $\boxed{\vec{w}^* = (X^TX)^{-1}X^T\vec{y}}$.
- 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 8 😄.
- The supplementary notebook is 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:

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

- 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, $\vec{w}^*$:

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

- 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:

$$\begin{aligned}\text{pred. commute} &= H(\text{departure hour}) \\ &= w_0 + w_1 \cdot \text{departure hour}\end{aligned}$$

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

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

- 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:

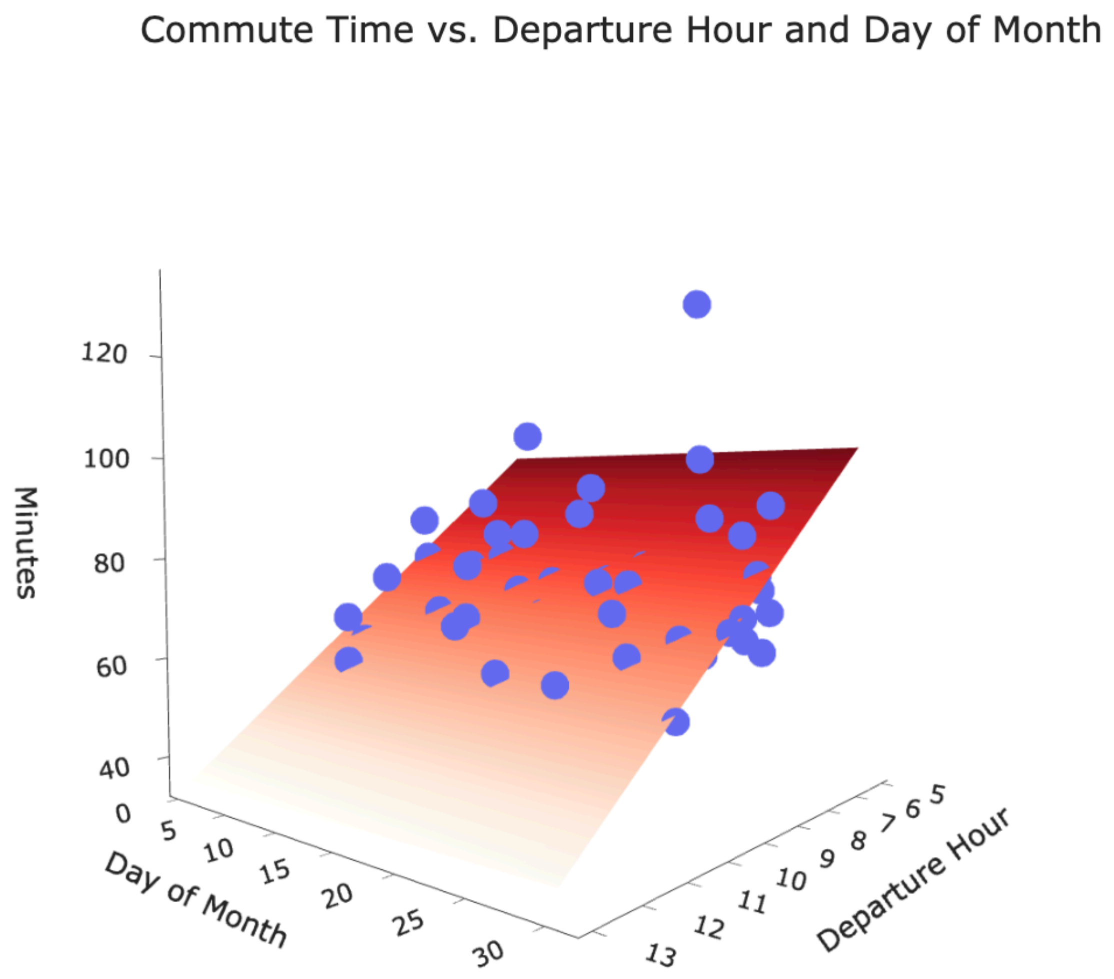
$$H(\text{departure hour}) = w_0 + w_1 \cdot \text{departure hour}$$

looks like a **line** in 2D.

- **Questions**:
    - How many dimensions do we need to graph the hypothesis function:

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

    - What is the shape of the hypothesis function?



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:

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

the hypothesis vector $\vec{h} \in \mathbb{R}^n$ can be written as:

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

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:

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

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

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

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:

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

- The supplementary notebook is 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.