Explanatory Notes for 6.390

Shaunticlair Ruiz (Current TA)

Fall 2022

Why Not Regularize θ_0?

Key Equation 1

In general, our **regularizer for regression** will be given by **square magnitude** of θ:

$$R(\Theta) = \|\theta\|^2 = \theta \cdot \theta$$

This approach is called **Ridge Regression**.

We'll discuss why it's called "ridge" regression once we find our solution.

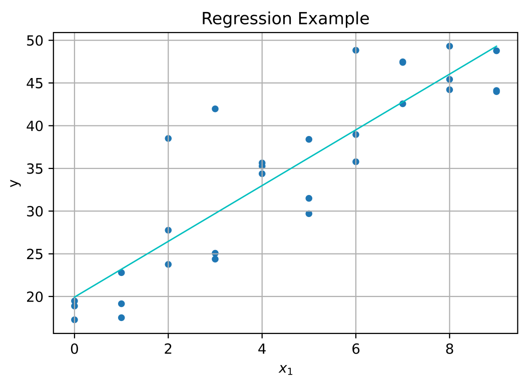
Why not include θ_0?

One thing you might immediately notice is that we used the magnitude of θ instead of Θ: this omits θ_0. Why would we do that?

We'll show that we need to **allow** the **offset** to have whatever value works best, and we shouldn't **punish** it.

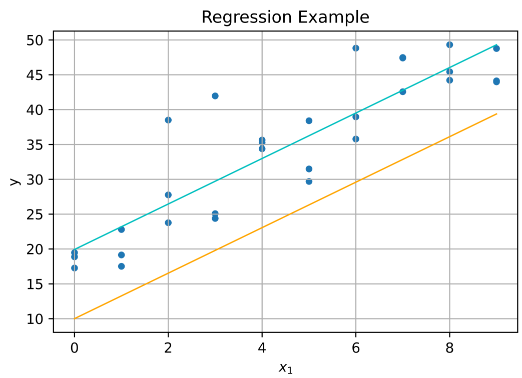
For simplicity, we won't do any regularization here: we can make our point without it.

This is best shown with a **visual** example. Let's take an example with one input x_1. So, we have a **linear** function: $h(x) = \theta_1 x_1 + \theta_0$.



Our regression example.

Let's suppose we **push** for a **much lower** (offset) θ_0 term, while keeping everything else the **same**:



Reducing our offset pulls our line further away from all of our data! That's not helpful.

This shows that we **need** our offset! We use it to **slide** our hyperplane around the space: if all of our data is **far** from $(0, 0)$, we need to be able to **move** our **entire line**.

And regularizing θ_1 wouldn't make this any better: it would just be flatter.

So, we'll keep θ_0 **separate** and **allow** it to take whatever value is **best**.

Concept 2

We **do not regularize** our **offset** term, θ_0.

Instead, we allow θ_0 to **shift** our hyperplane wherever it **needs** to be.

The other terms θ control the **orientation** of the hyperplane: the **direction** it is **facing**. We **regularize** this to push it towards less "complicated" orientations.

This will be discussed more in-depth in the Classification chapter!