Python Data Products

Course 3: Making Meaningful Predictions from Data

Lecture: Motivation behind the MSE

Learning objectives

In this lecture we will...

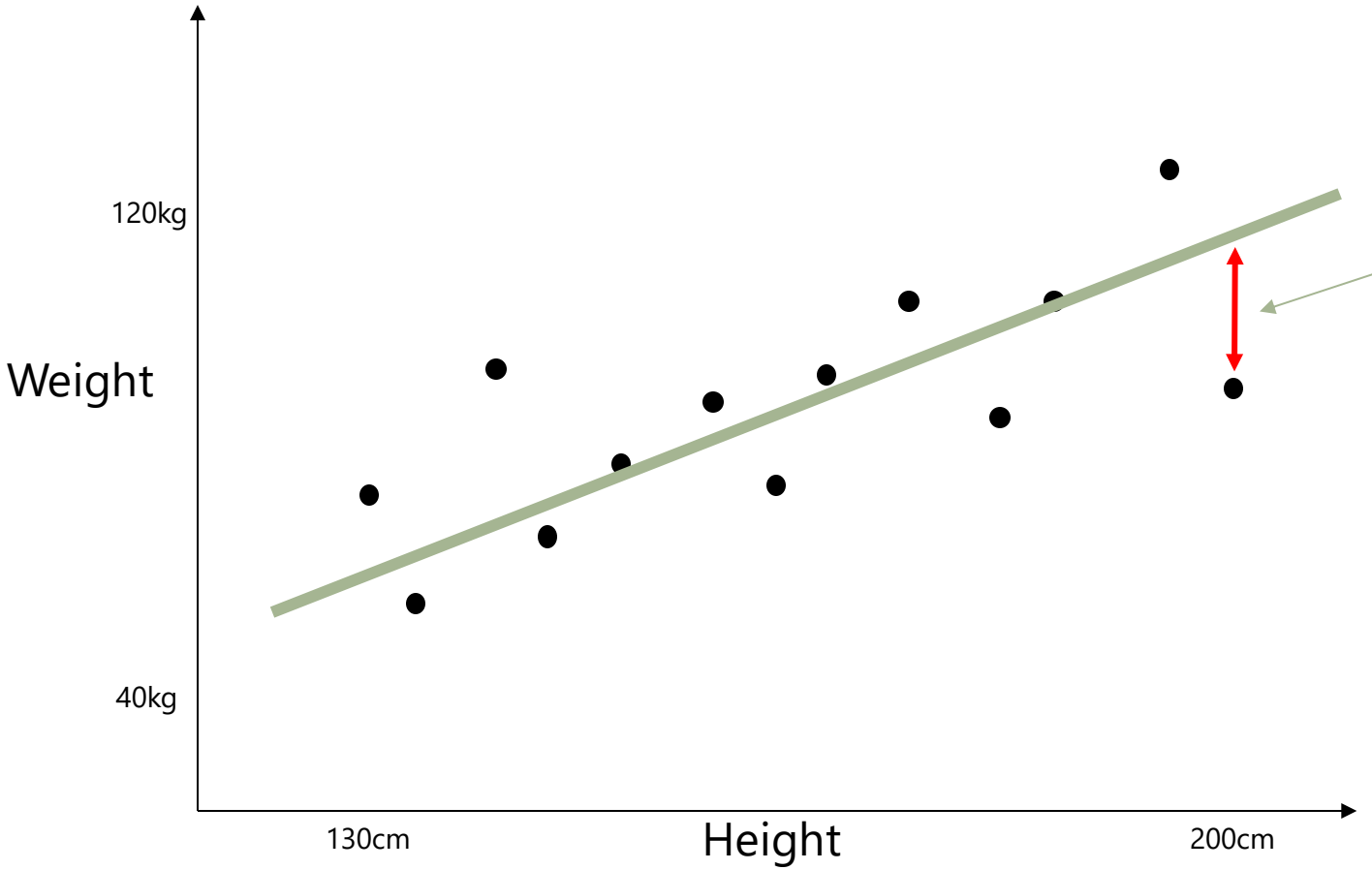
- Present the most common evaluation measures used for regression models (the Mean Squared Error)
- Motivate the choice of this particular error measure using statistics and probability

Regression diagnostics

Q: How should we evaluate our regression model?

Regression diagnostics

Q: Can we find a line that (approximately) fits the data)?



Error between correct value and prediction

$$y_i - X_i \cdot \theta$$

Concept: Mean Squared Error

Mean-squared error (MSE)

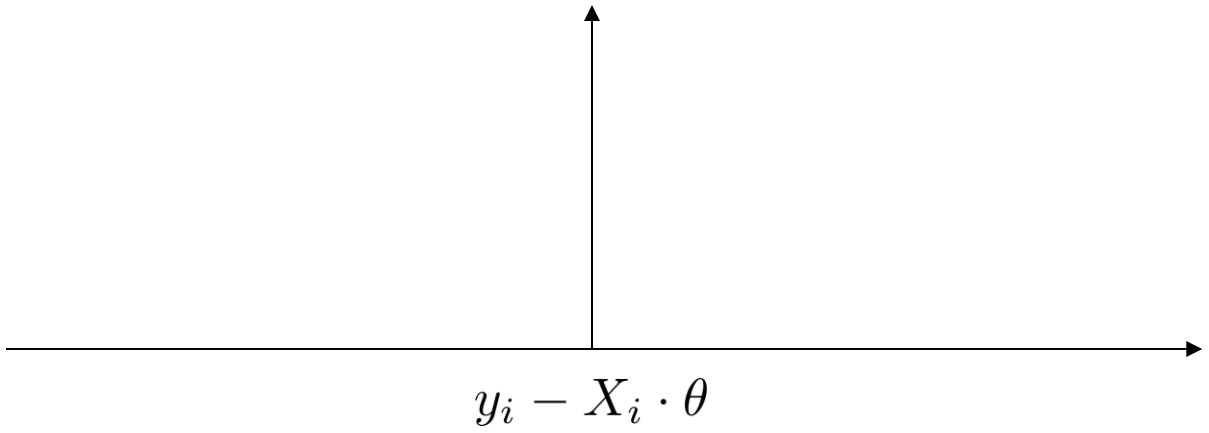
$$\frac{1}{N}\|y - X\theta\|_2^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i \cdot \theta)^2$$

Regression diagnostics

Q: Why MSE (and not mean-absolute-error or something else)

Regression diagnostics



label = prediction + error

$$y_i = X_i \cdot \theta + \mathcal{N}(0, \sigma)$$

Regression diagnostics

$$p_\theta(y|X) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y_i - X_i \cdot \theta)^2}{2\sigma^2}}$$

$$\max_\theta p_\theta(y|X) = \max_\theta \prod_i e^{-(y_i - X_i \cdot \theta)^2}$$

$$= \min_\theta \sum_i (y_i - X_i \cdot \theta)^2$$

Summary of concepts

- Understand the motivation behind the MSE in terms of probability
- Understand the notion of "error distributions"
- (at a high level) understand the relationship between likelihood (probability) and error (prediction)

On your own...

- Compute MSE and related statistics (like Mean **Absolute** Error) and compare cases where the errors are high and low