An introduction to linear regression

Topics we'll cover

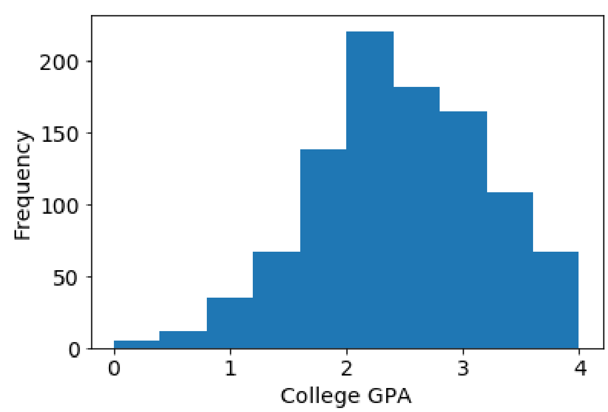
1. The regression problem in one dimension
2. Predictor and response variables
3. A loss function formulation
4. Deriving the optimal solution

Linear regression

Fitting a line to a bunch of points.

Example: college GPAs

Distribution of GPAs of students at a certain Ivy League university.

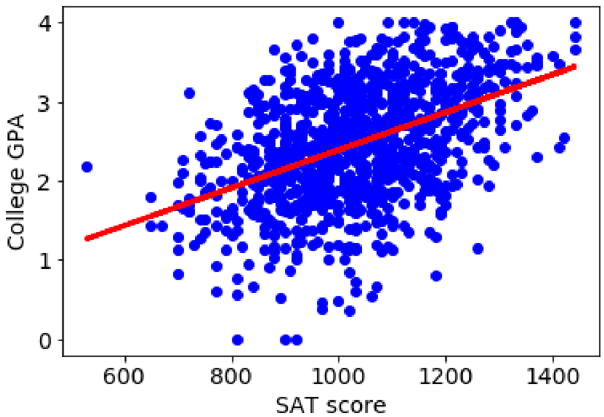


What GPA to predict for a random student from this group?

- Without further information, predict the **mean**, 2.47.
- What is the average squared error of this prediction?
  That is, $\mathbb{E}[((\text{student's GPA}) - (\text{predicted GPA}))^2]$?
  The **variance** of the distribution, 0.55.

Better predictions with more information

We also have SAT scores of all students.



Mean squared error (MSE) drops to 0.43.

This is a **regression** problem with:

- **Predictor variable**: SAT score
- **Response variable**: College GPA

Parametrizing a line

A line can be parameterized as $y = ax + b$ (a: slope, b: intercept).

The line fitting problem

Pick a line (parameters a, b) suited to the data, $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R} \times \mathbb{R}$

- $x^{(i)}, y^{(i)}$ are predictor and response variables, e.g. SAT score, GPA of ith student.
- Minimize the mean squared error,

$$\mathrm{MSE}(a, b) = \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - (ax^{(i)} + b))^2.$$

This is the **loss function**.

Minimizing the loss function

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$, minimize

$$L(a, b) = \sum_{i=1}^{n} (y^{(i)} - (ax^{(i)} + b))^2.$$