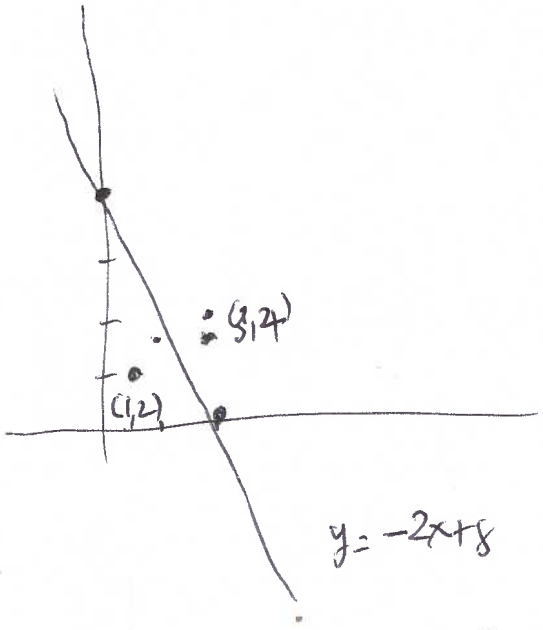
WEEK 4

An introduction to linear regression

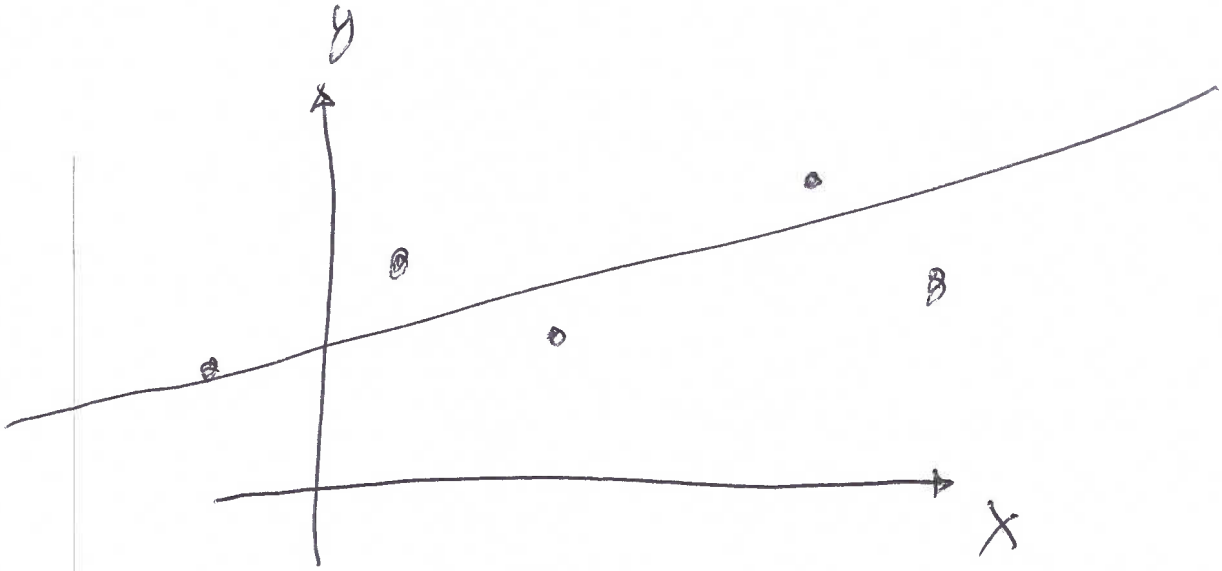


Topics we'll cover

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

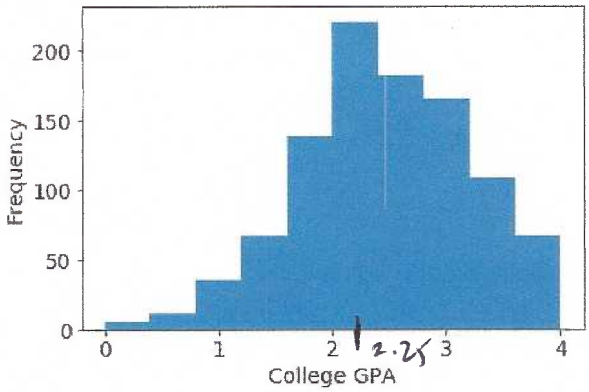
Linear regression

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.

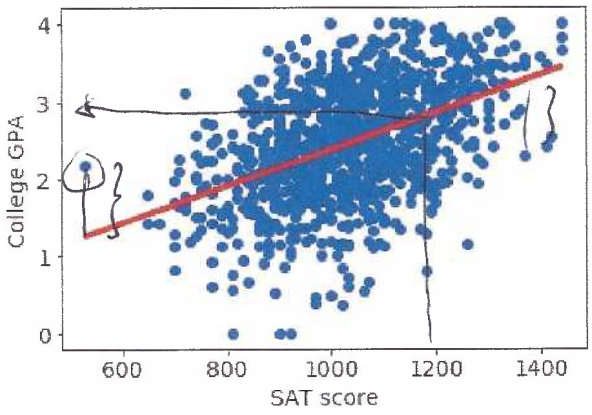
How good is this prediction

Better predictions with more information

freshman

We also have SAT scores of all students.

tilted upward so positive correlation



Mean squared error (MSE) drops to 0.43.

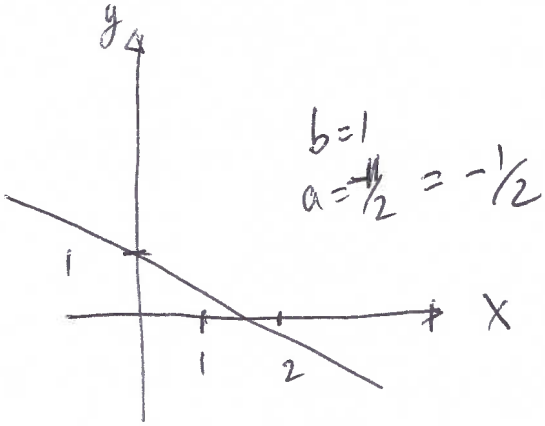
average of squared error → MSE

This is a **regression** problem with:

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

Parametrizing a line

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



$b = 1$

$a = -1/2 = -1/2$

$y = -\frac{1}{2}x + 1$

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,

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

This is the **loss function**.

We are optimizing by the loss function

We want to find the line that gives the least MSE. The line is defined by parameters a & b.

correct value

value we would predict using the line $y = ax + b$

Minimizing the loss function

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

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

To minimize, set $\frac{dL}{da} = \frac{dL}{db} = 0$

$a = \dfrac{\text{Covariance of } x \& y}{\text{Variance of } X}$

$\frac{d}{db} u^2 = 2u\,du$

$$\frac{dL}{db} = \sum_{i=1}^{n} 2\left(y^{(i)} - (ax^{(i)} + b)\right) \cdot (-1) = 0$$

$$\Rightarrow \sum_i y^{(i)} = a \sum_i x^{(i)} + nb$$

$$\Rightarrow b = \frac{1}{n} \sum_i y^{(i)} - a \cdot \frac{1}{n} \sum_i x^{(i)}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x^{(i)}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

$b = \bar{y} - a\bar{x}$, next find a

$$\frac{dL}{da} = 0 \Rightarrow \text{simplify:} \quad a = \frac{\sum_{i=1}^{n} (y^{(i)} - \bar{y})(x^{(i)} - \bar{x})}{\sum (x^{(i)} - \bar{x})^2}$$

$$a = \frac{\sum_{i=1}^{n} (y^{(i)} - \bar{y})(x^{(i)} - \bar{x})}{\sum (x^{(i)} - \bar{x})^2}$$