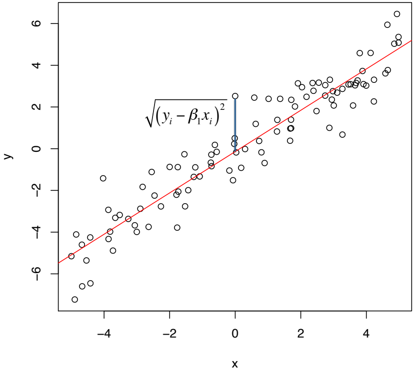
Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Linear Regression

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul
SLIDES ADAPTED FROM LAUREN HANNAH

Fitting a Linear Regression



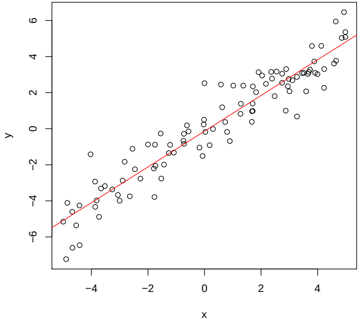
Idea: minimize the Euclidean distance between data and fitted line

$$RSS(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta \mathbf{x}_i)^2$$

How to Find β

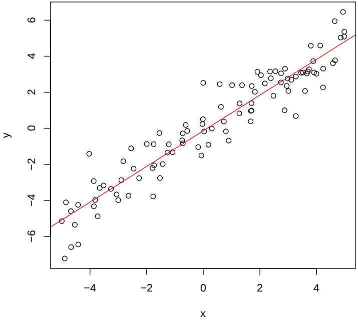
- Use calculus to find the value of β that minimizes the RSS
- The optimal value is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$



Probabilistic Interpretation

- Our analysis so far has not included any probabilities
- Linear regression does have a *probabilisitc* (probability model-based) interpretation

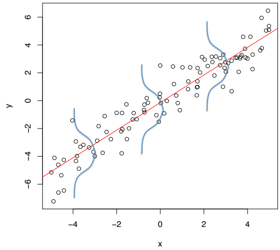


Probabilistic Interpretation

- Linear regression assumes that response values have a Gaussian distribution around the linear mean function,

$$Y_i | \mathbf{x}_i, \beta \sim N(\mathbf{x}_i \beta, \sigma^2)$$

- This is a *discriminative model*, where inputs x are not modeled



- Minimizing RSS is equivalent to maximizing conditional likelihood