

Linear Models

Geometric Algorithms

Lecture 24

Introduction

Recap Problem

$$A = \begin{matrix} & \vec{a}_1 & \vec{a}_2 & \vec{a}_3 \\ \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} & & & \end{matrix} \quad \mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}$$

(orthogonal)

Find the \vee projection of \mathbf{b} onto $\text{Col}(A)$.

Hint. $\text{rank}(A) = 2$ $\vec{a}_2 = \vec{a}_1 + \vec{a}_3$

$$\text{Col}([\vec{a}_1 \quad \vec{a}_3]) = \text{Col}(A)$$

Answer

$$A \in \mathbb{R}^{m \times n}$$

$$\hat{b} = A (A^T A)^{-1} A^T \vec{b}$$

but $\text{rank}(A) = n$

\hat{x} : L.S. sol. for $A \vec{x} = \vec{b}$

then $A^T A \hat{x} = (A^T A)^{-1} A^T \vec{b}$

$$A \hat{x} = \vec{b}$$

$$\hat{b} = A \hat{x} = A (A^T A)^{-1} A^T \vec{b}$$

$$\hat{b} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1/2 \\ 8/3 \end{bmatrix} = \dots$$

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix}$$

$$\vec{b} = \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 1 \\ 0 & -1 \\ 1 & -1 \end{bmatrix}$$

$\text{Col}(C) = \text{Col}(A)$

$$C^T = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

$$C^T \vec{b} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}$$

$$(C^T C)^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/3 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$$

Question

Find the matrix which implements orthogonal projection onto the span of $\begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$.

Answer

$$u = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

$$\frac{1}{6} \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

$$\vec{r} \mapsto \text{proj}_{\text{span}\{\vec{u}\}} \vec{r} = \frac{\langle \vec{r}, \vec{u} \rangle}{\langle \vec{u}, \vec{u} \rangle} \vec{u}$$

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \mapsto \frac{\langle e_1, \vec{u} \rangle}{\langle \vec{u}, \vec{u} \rangle} \vec{u} = \frac{[1 \ 0 \ 0] \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}}{[1 \ -1 \ 2] \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \mapsto \frac{1}{6} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \mapsto \frac{2}{6} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix}$$

$$\frac{1}{6} \begin{bmatrix} 1 & -1 & 2 \\ -1 & 1 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

Objectives

1. Use the least square method to build linear *models* of noisy data.
2. Show how we can use linear algebraic methods to model with non-linear models.

Keywords

line of best fit

independent/dependent variables

residuals

prediction

simple least squares regression

multiple regression

polynomial regression

models

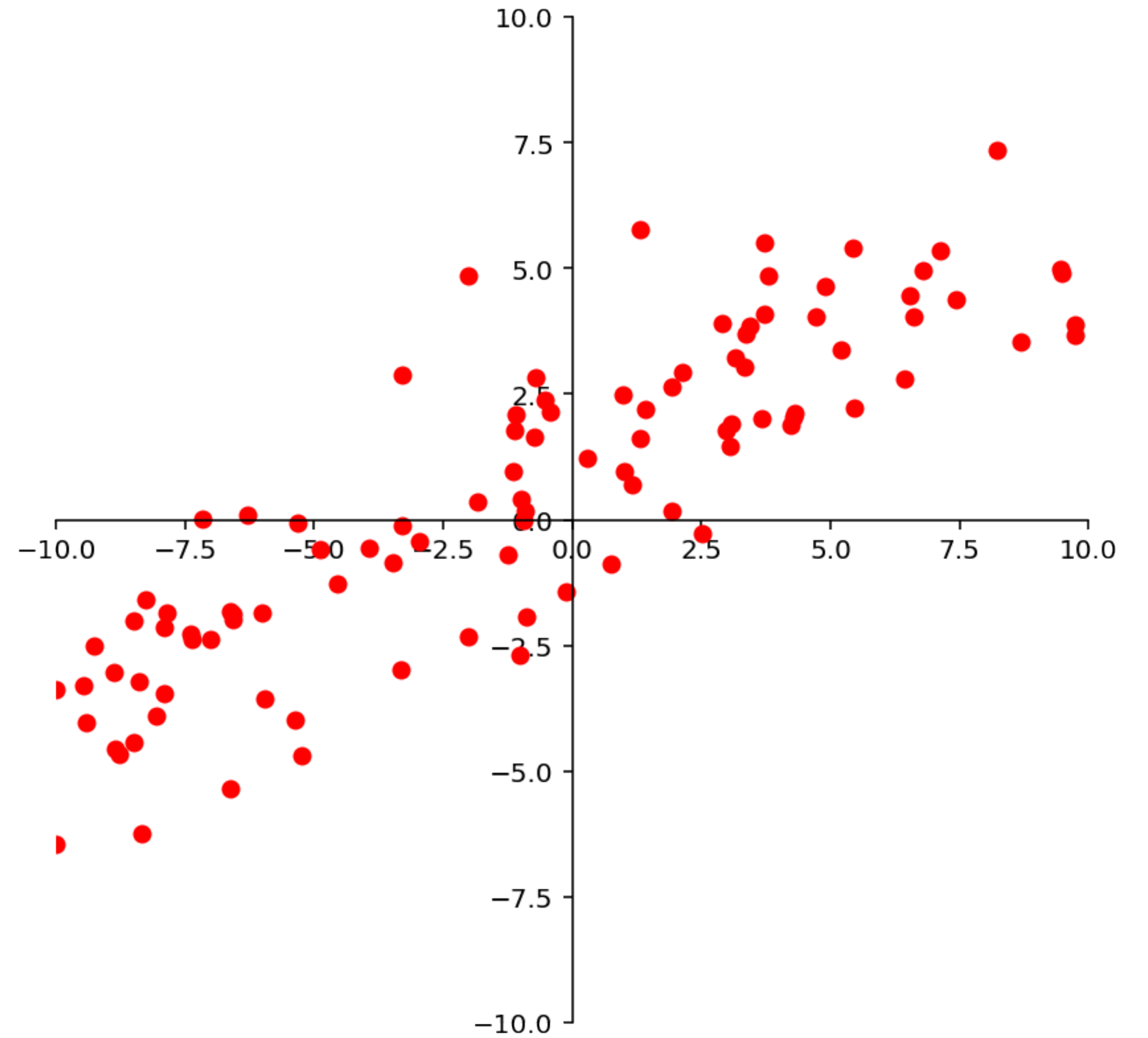
model fitting

model parameters

design matrices

A Warmup: Line of Best Fit

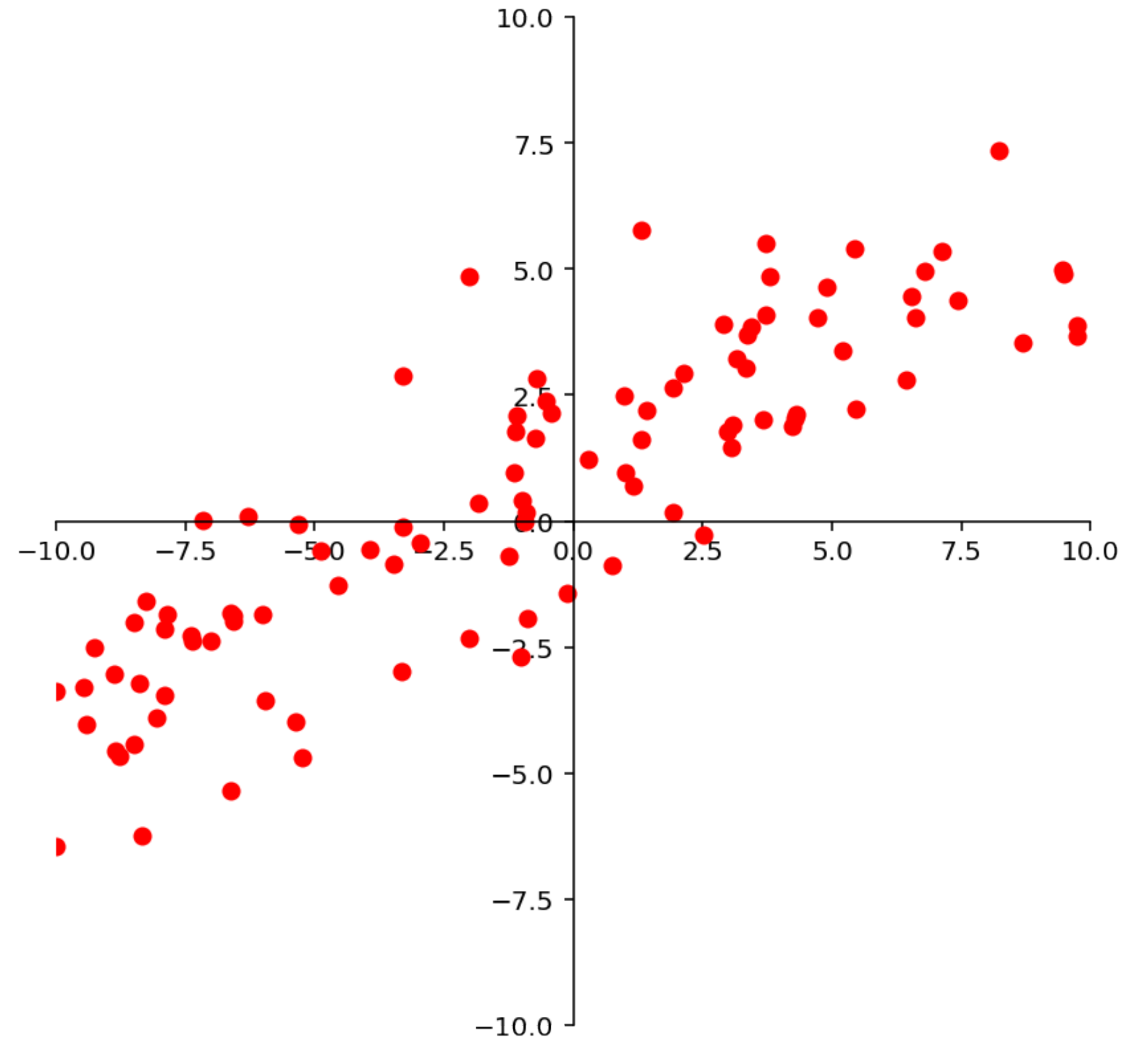
The Setup



The Setup

You're given a set of points in \mathbb{R}^2

$$\{(x_1, y_1), \dots, (x_k, y_k)\}$$

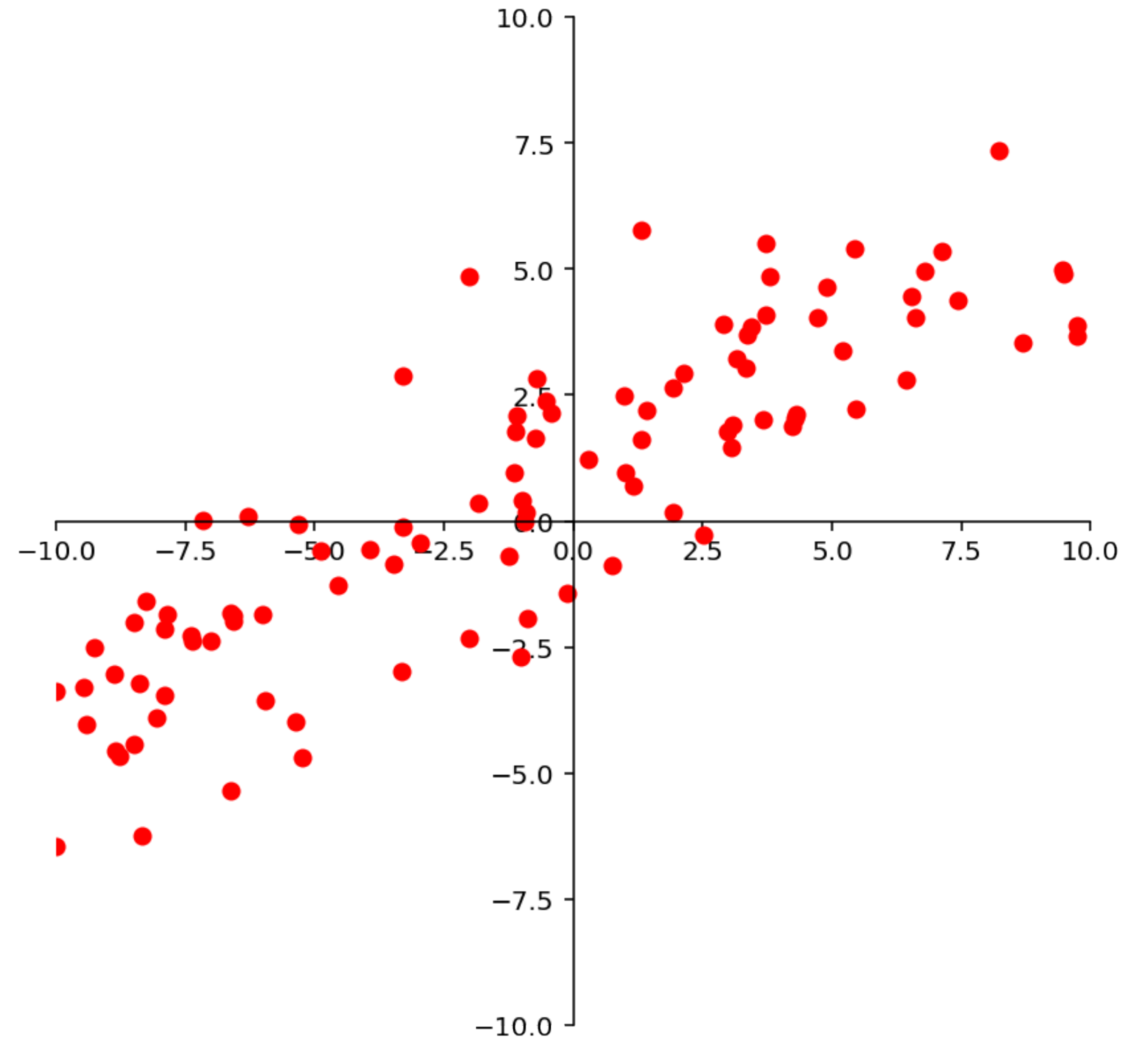


The Setup

You're given a set of points in \mathbb{R}^2

$$\{(x_1, y_1), \dots, (x_k, y_k)\}$$

Example. You collect (height, weight) data for a population.



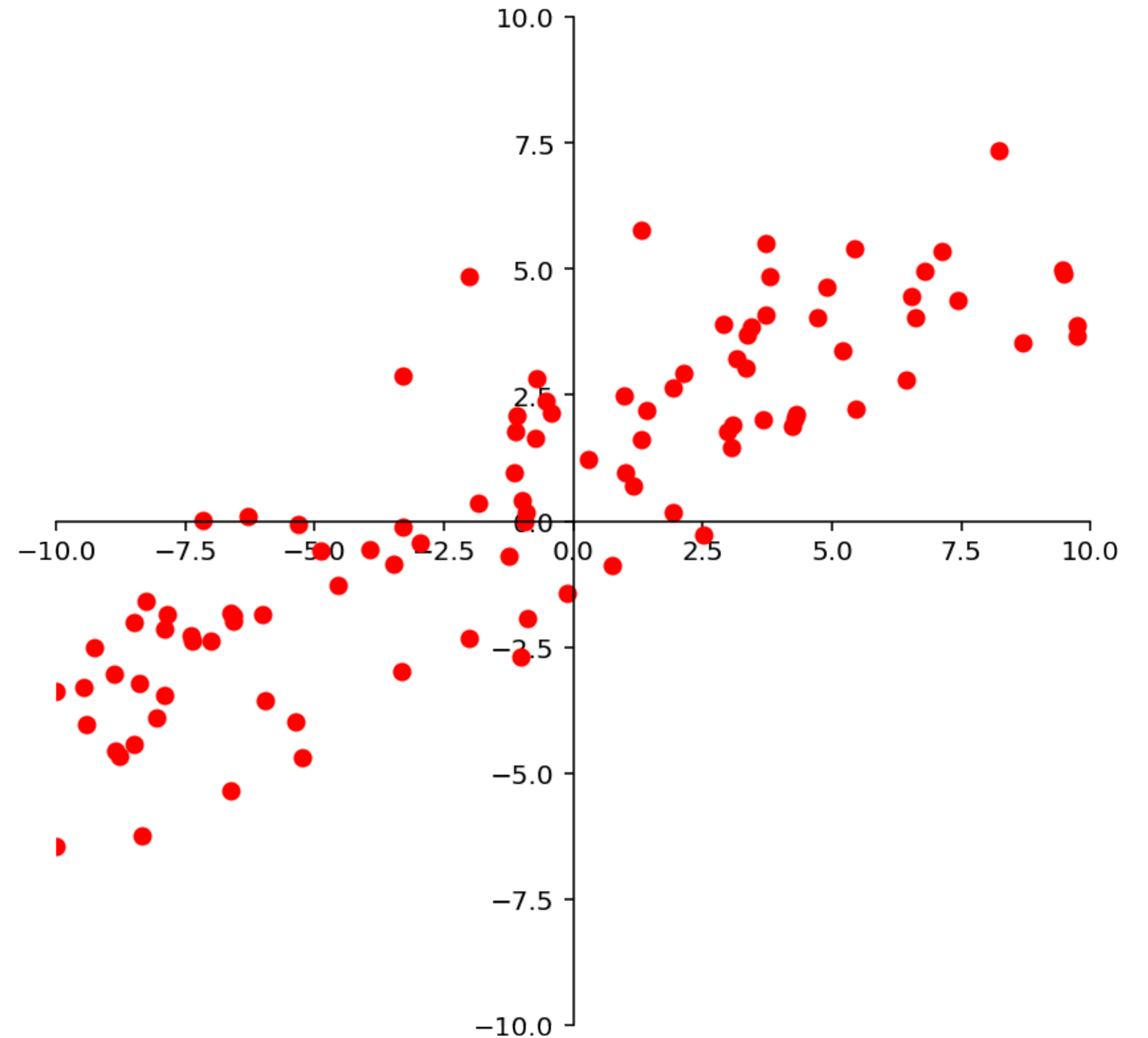
The Setup

You're given a set of points in \mathbb{R}^2

$$\{(x_1, y_1), \dots, (x_k, y_k)\}$$

Example. You collect (height, weight) data for a population.

You notice they *kind of* trend as a line.



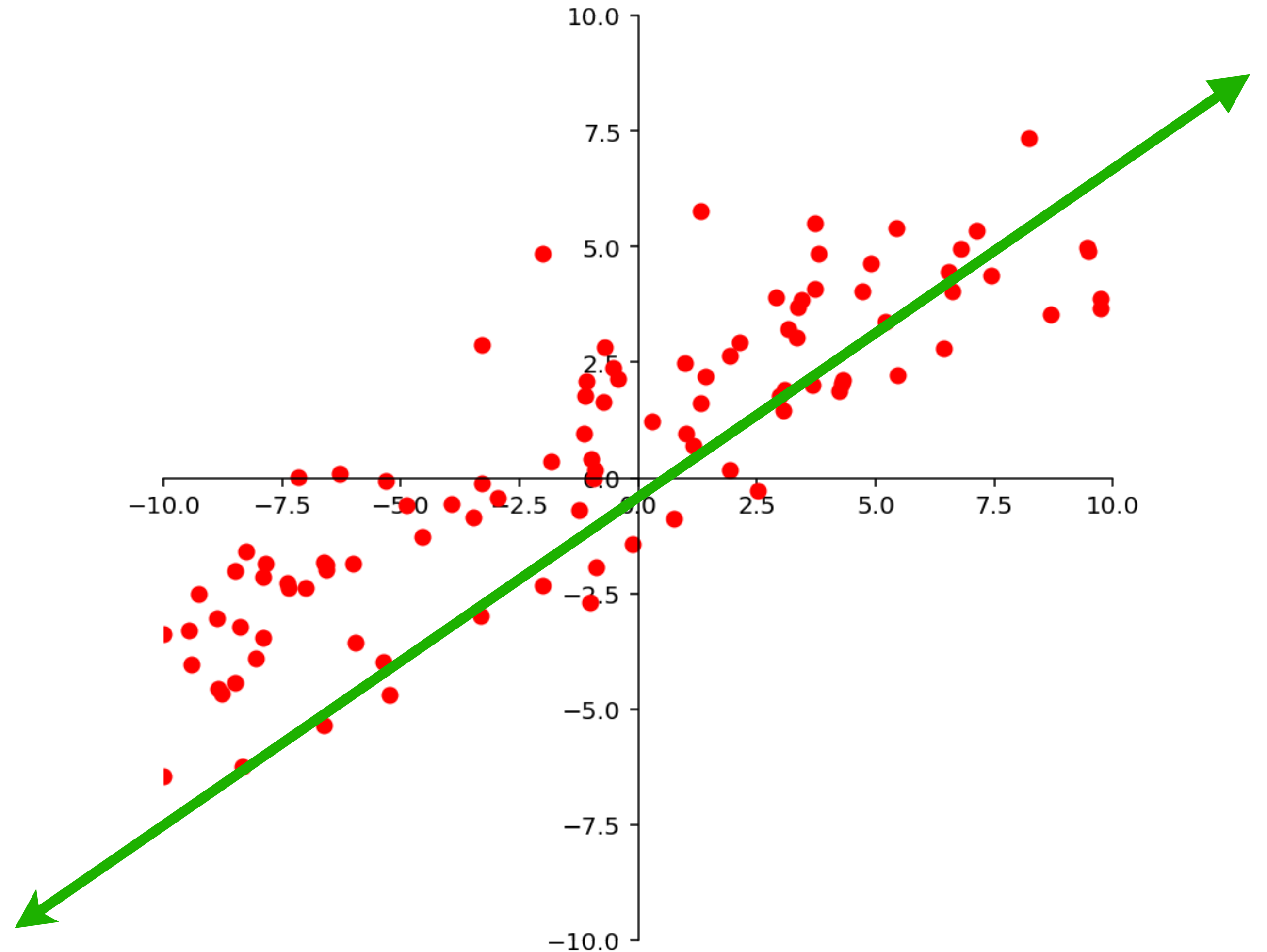
The Setup

You're given a set of points in \mathbb{R}^2

$$\{(x_1, y_1), \dots, (x_k, y_k)\}$$

Example. You collect (height, weight) data for a population.

You notice they *kind of* trend as a line.



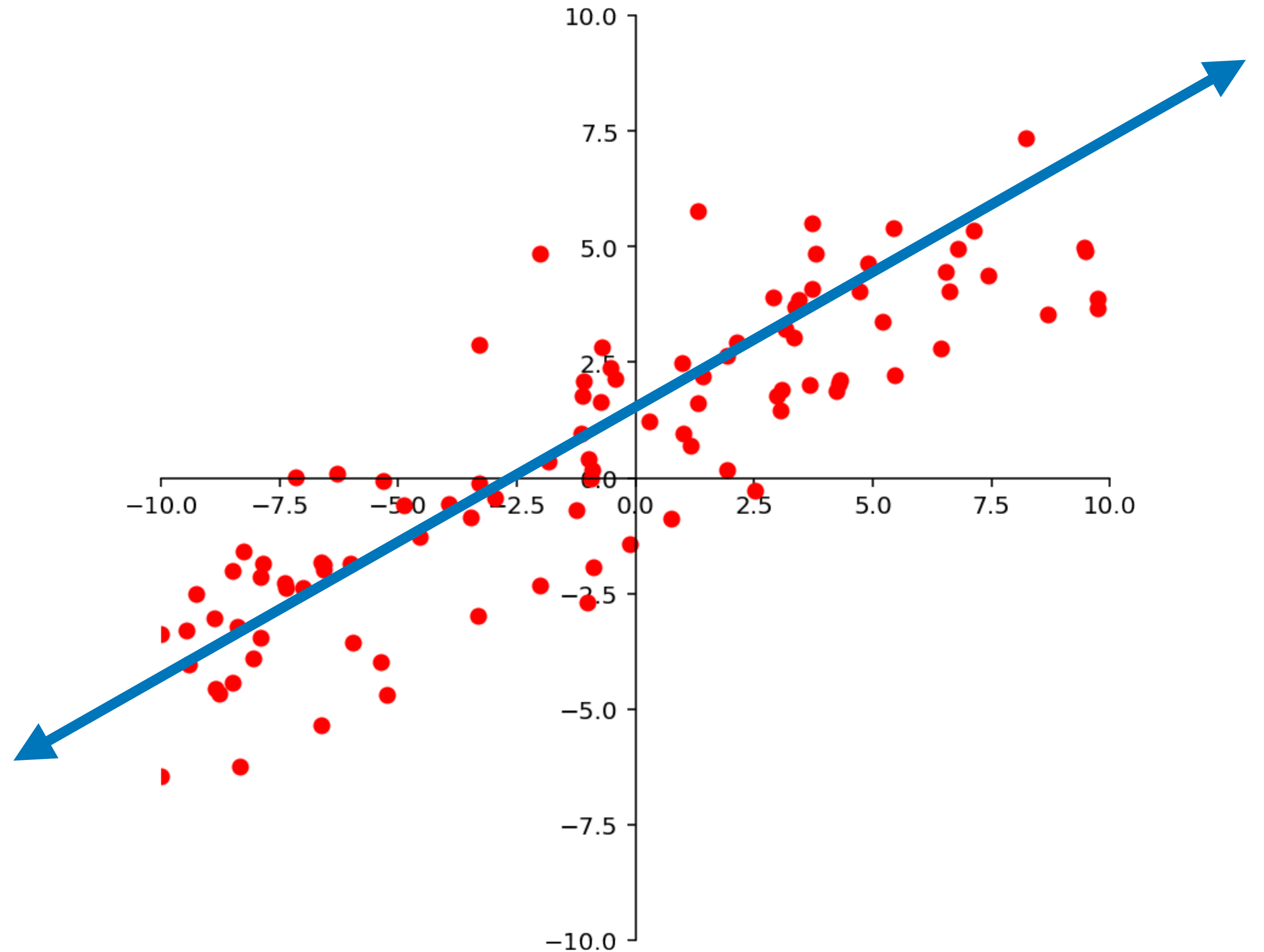
The Setup

You're given a set of points in \mathbb{R}^2

$$\{(x_1, y_1), \dots, (x_k, y_k)\}$$

Example. You collect (height, weight) data for a population.

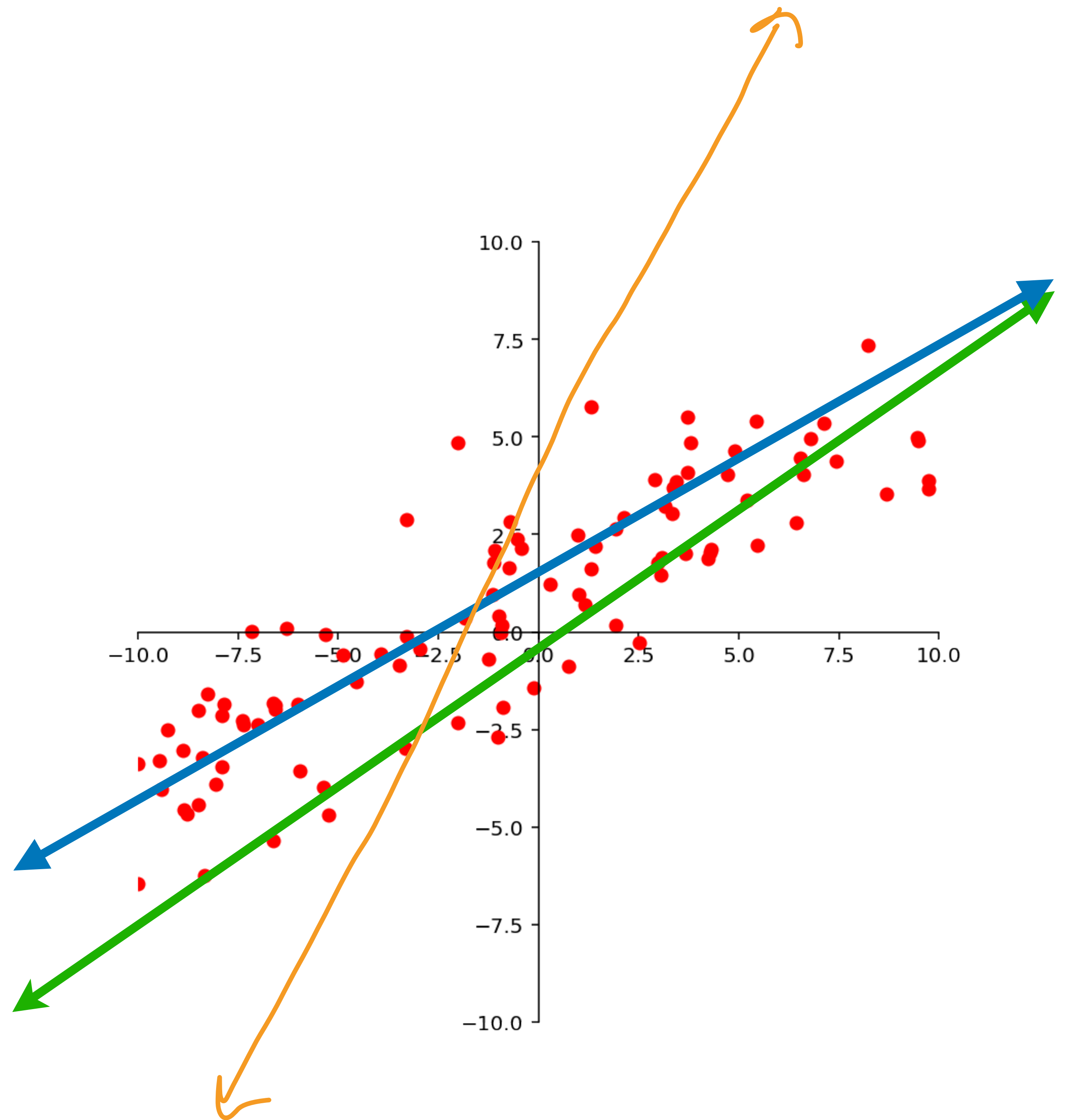
You notice they *kind of* trend as a line.



The Setup

Question. Which line "best" describes the trend of the dataset?

Which one *best models* the dataset?



Two Important Questions

Two Important Questions

1. What is a model?

Two Important Questions

1. What is a model?

We'll come back to this...

Two Important Questions

1. What is a model?

We'll come back to this...

2. What does "best" mean?

Two Important Questions

1. What is a model?

We'll come back to this...

2. What does "best" mean?

This is a make-or-break question.

Least Squares Simple Linear Regression

Problem. Given a set of points $\{(x_1, y_1), \dots, (x_n, y_n)\}$, find the line

$$f(x) = \beta_0 + \beta_1 x$$

which minimizes

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

Least Squares Simple Linear Regression

Problem. Given a set of points $\{(x_1, y_1), \dots, (x_n, y_n)\}$, find the line

$$f(x) = \boxed{\beta_0} + \boxed{\beta_1}x$$

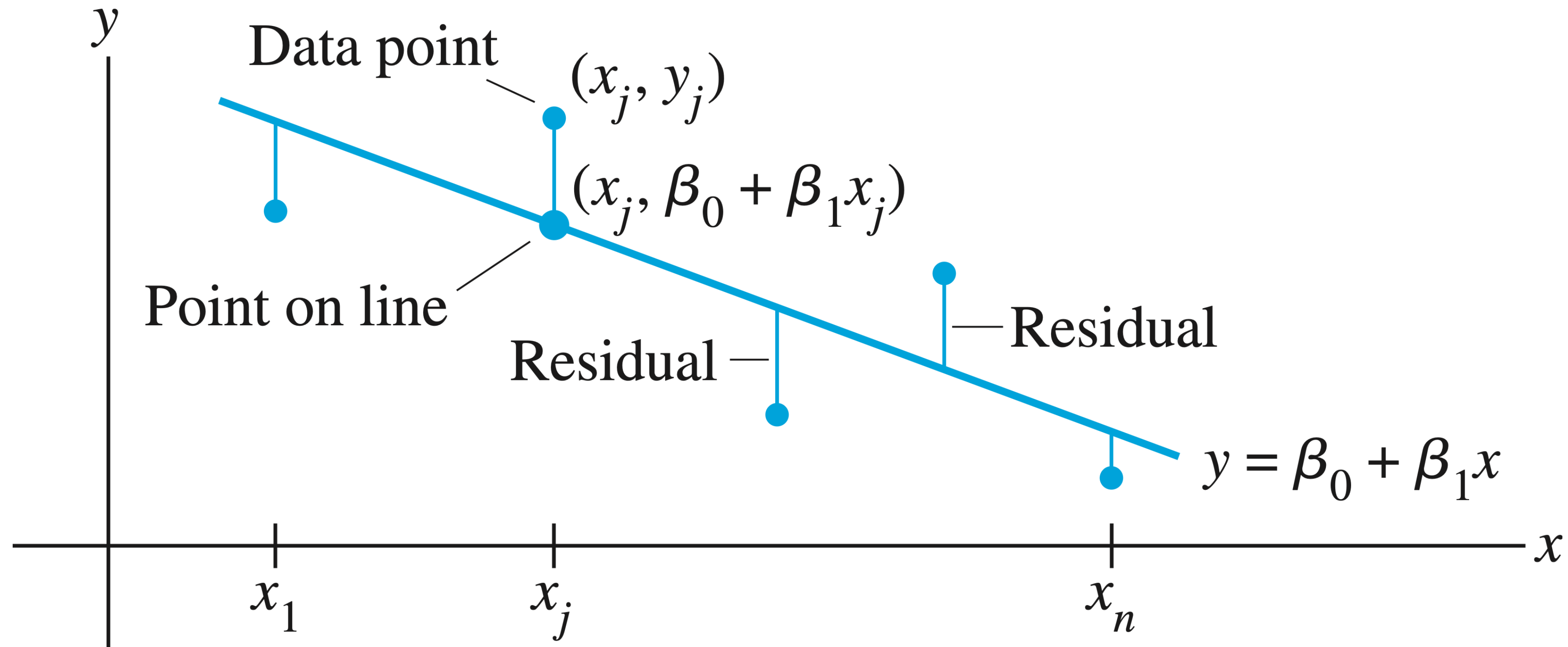
slope
y-intercept

which minimizes

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

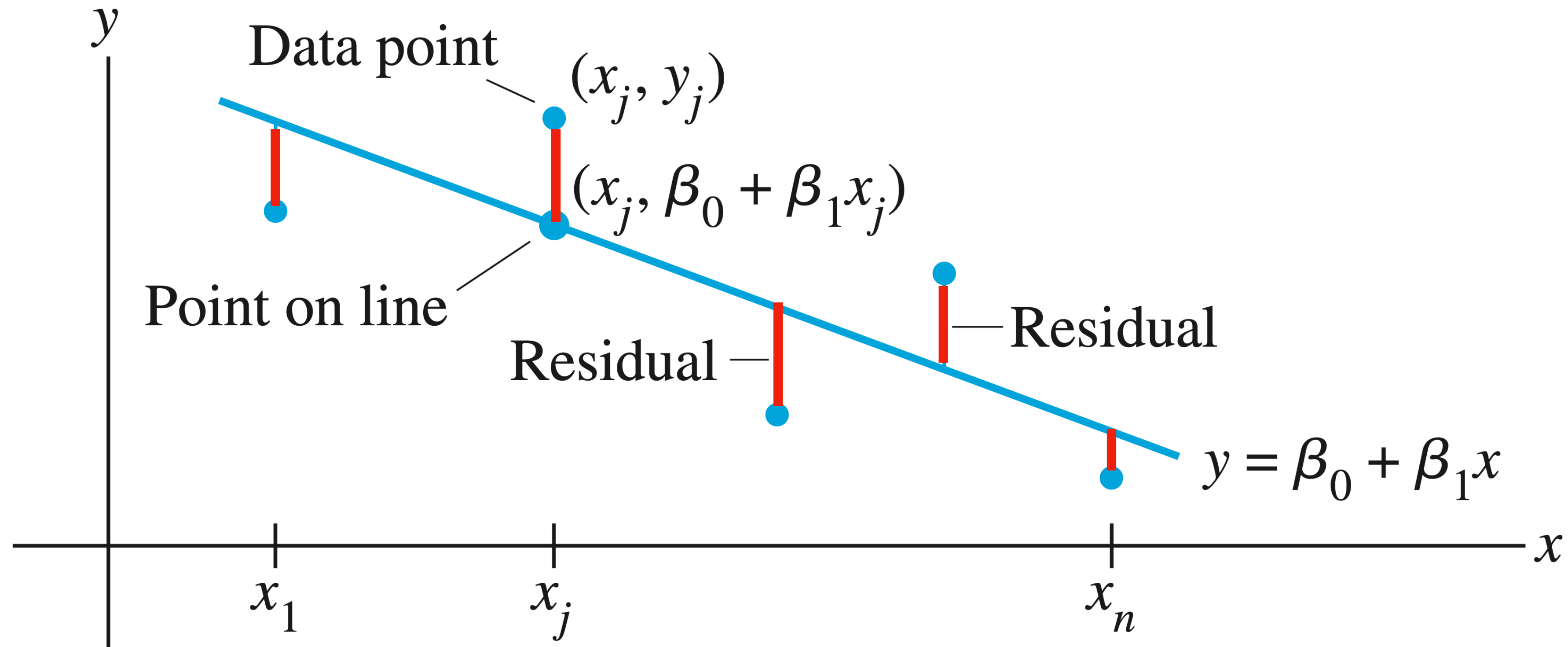
The "best" line minimizes
the *sum of squares of
differences.*

The Picture



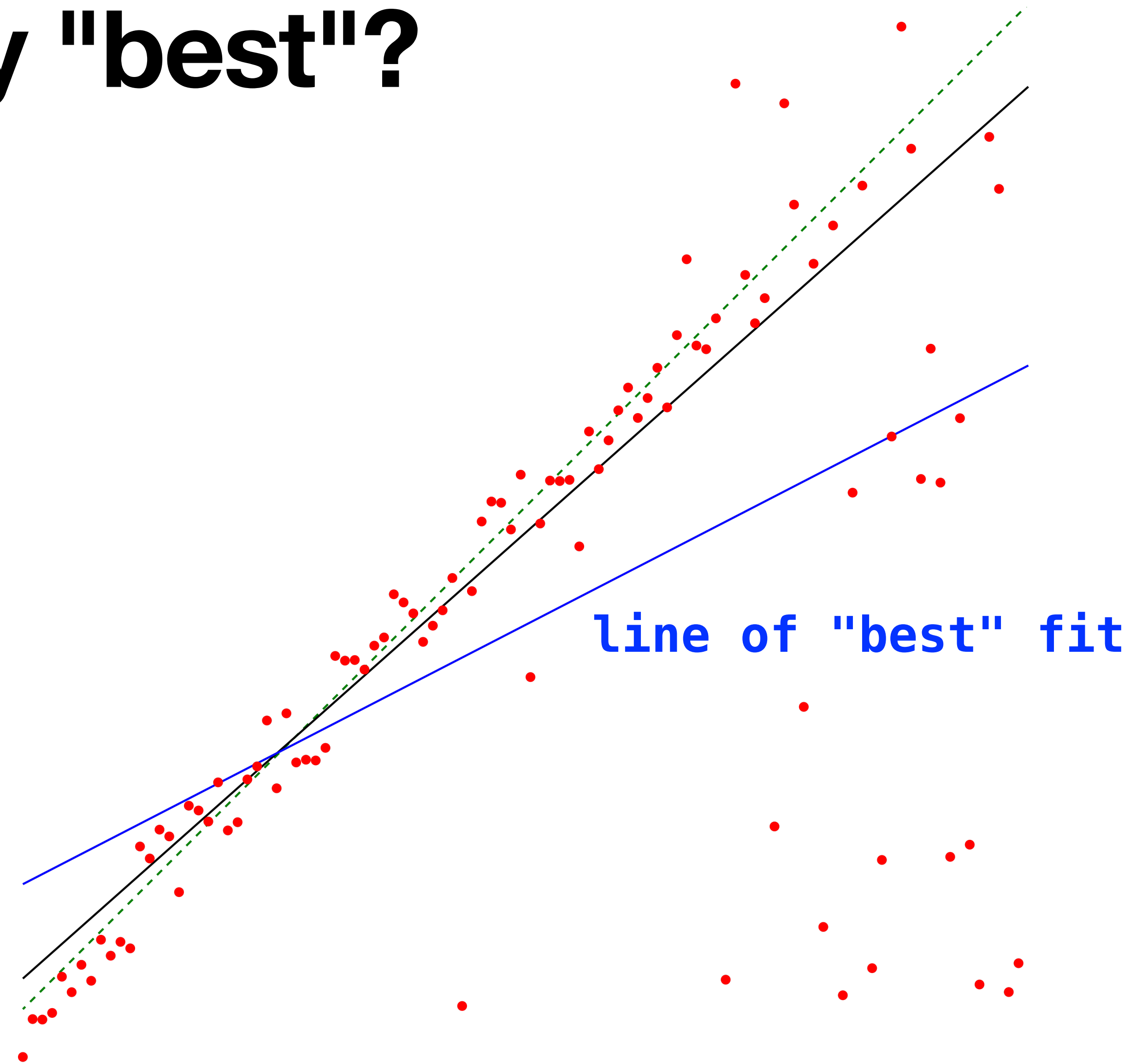
We want to find the line which makes the sum of these differences *as small as possible*.

The Picture



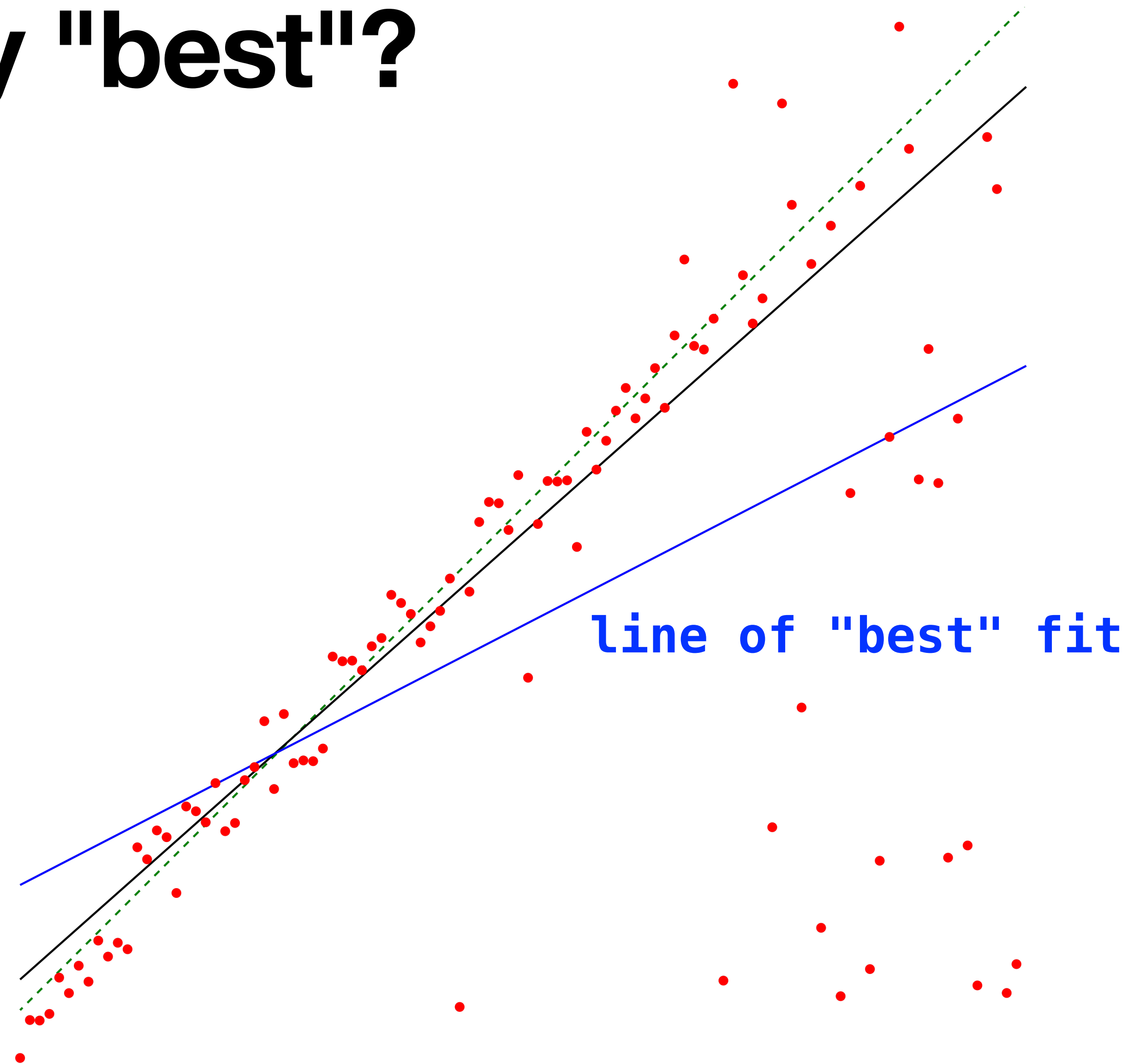
We want to find the line which makes the sum of these differences *as small as possible*.

An Aside: Is this really "best"?



An Aside: Is this really "best"?

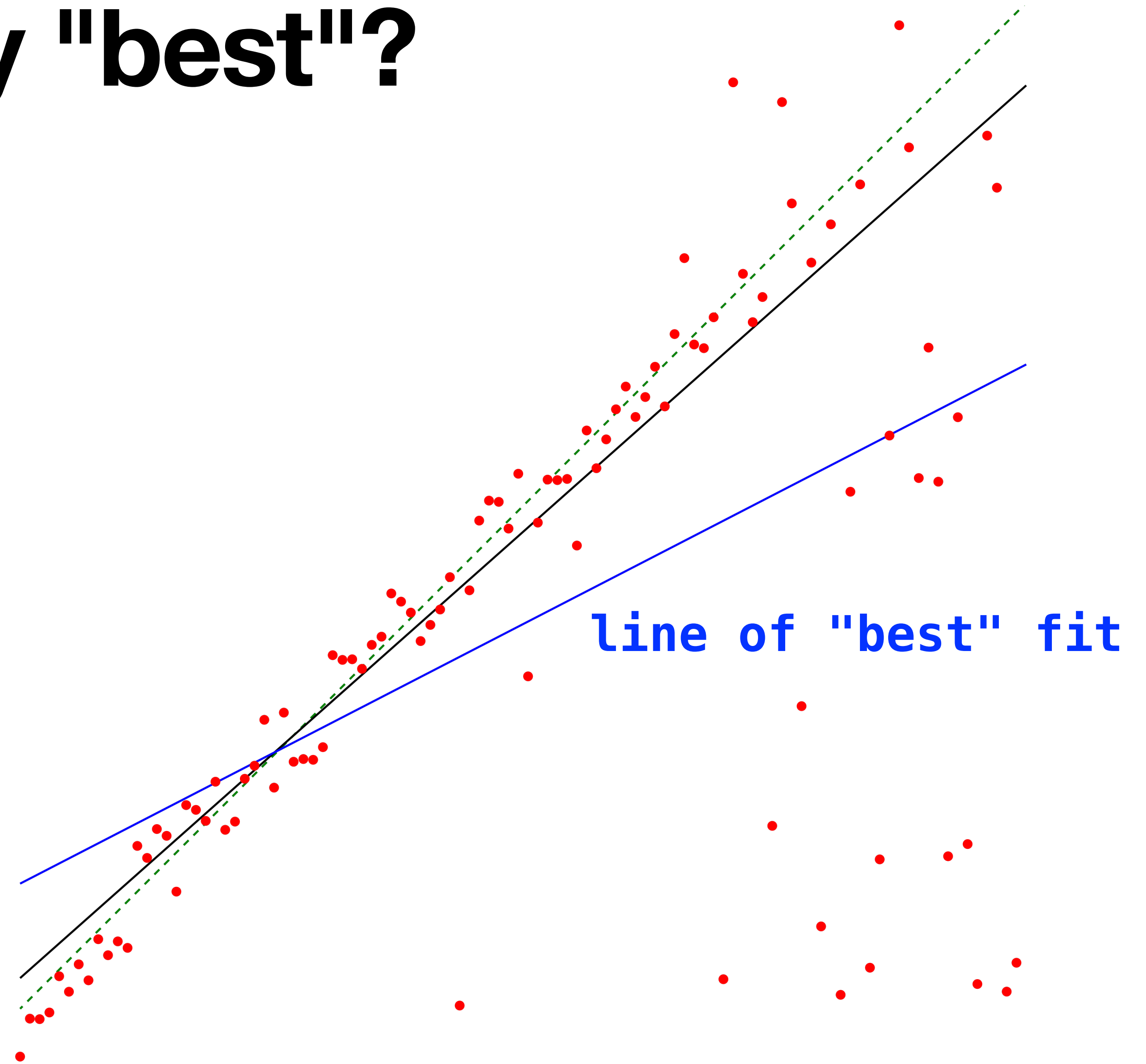
Who's to say...



An Aside: Is this really "best"?

Who's to say...

It depends on the data,
on the application
domain, etc.

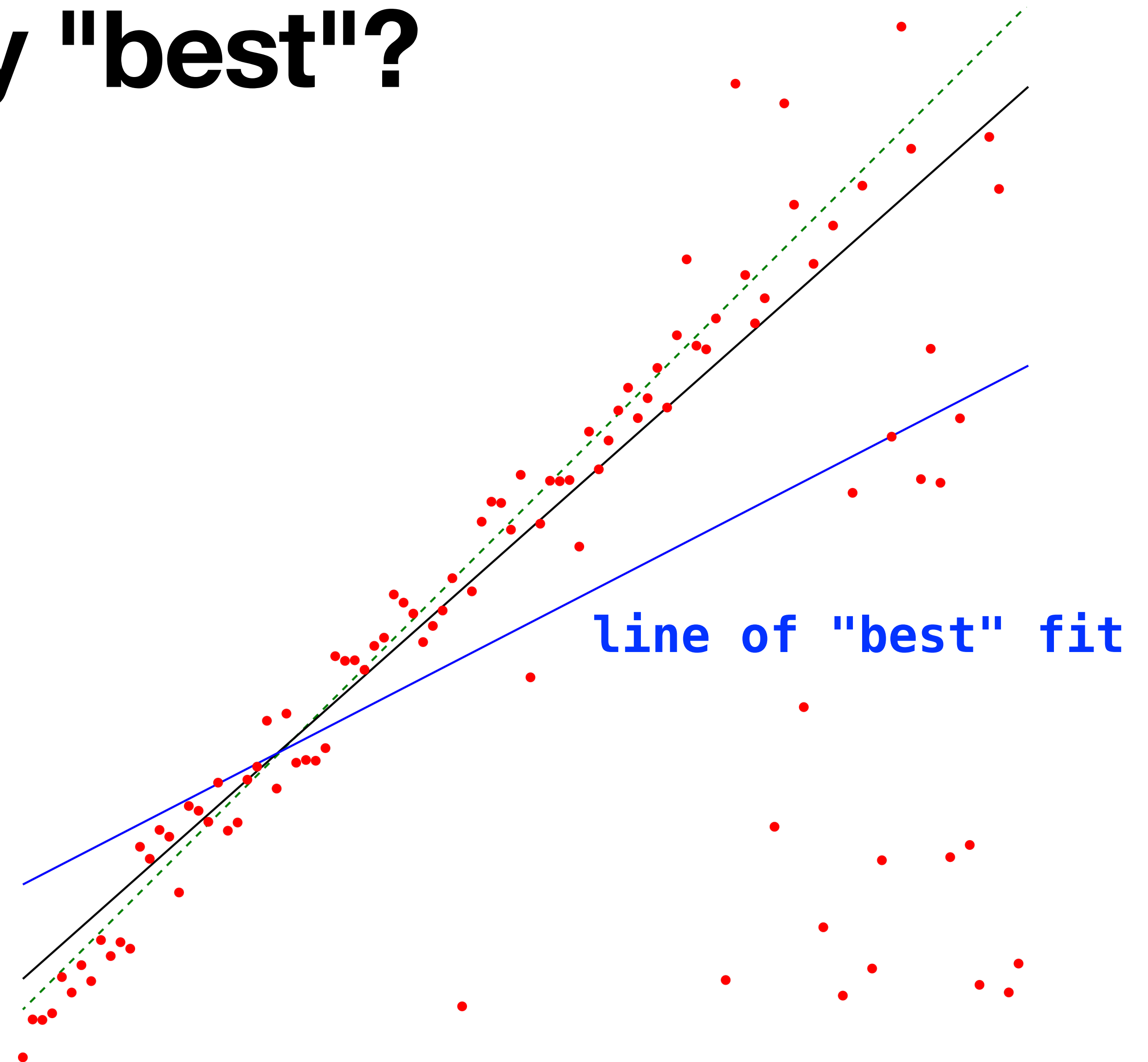


An Aside: Is this really "best"?

Who's to say...

It depends on the data,
on the application
domain, etc.

The point. We fix our
notion of "best" first,
and then we do
calculations and
derivations from there.



Terminology: Datasets

$$\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$$

Terminology: Datasets

$\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$
dataset

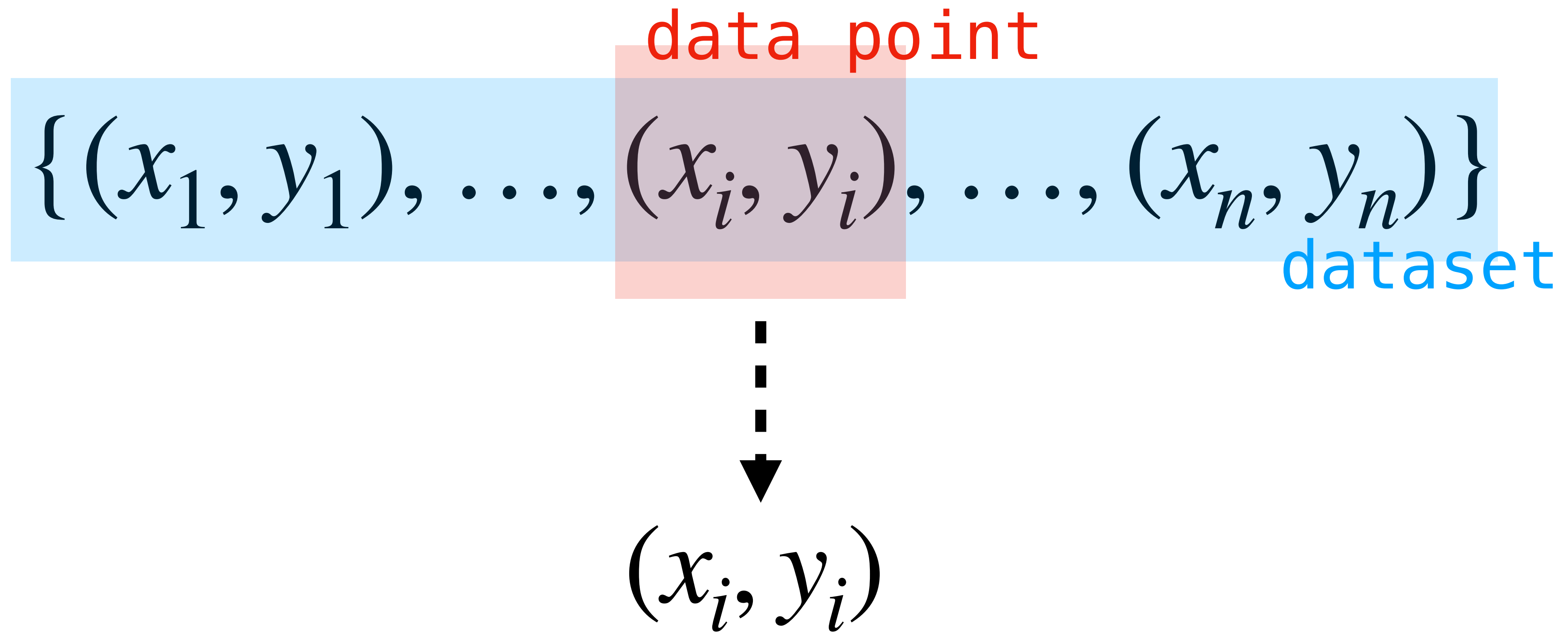
Terminology: Datasets

$\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$

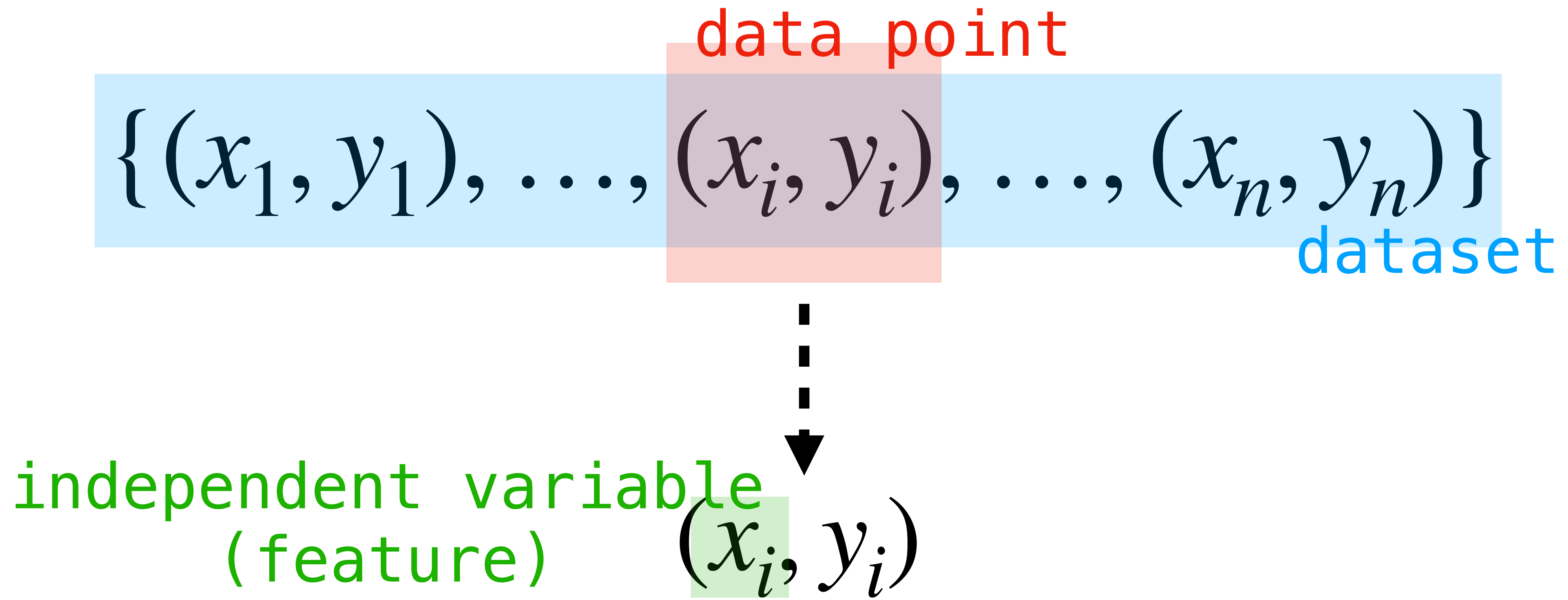
data point

dataset

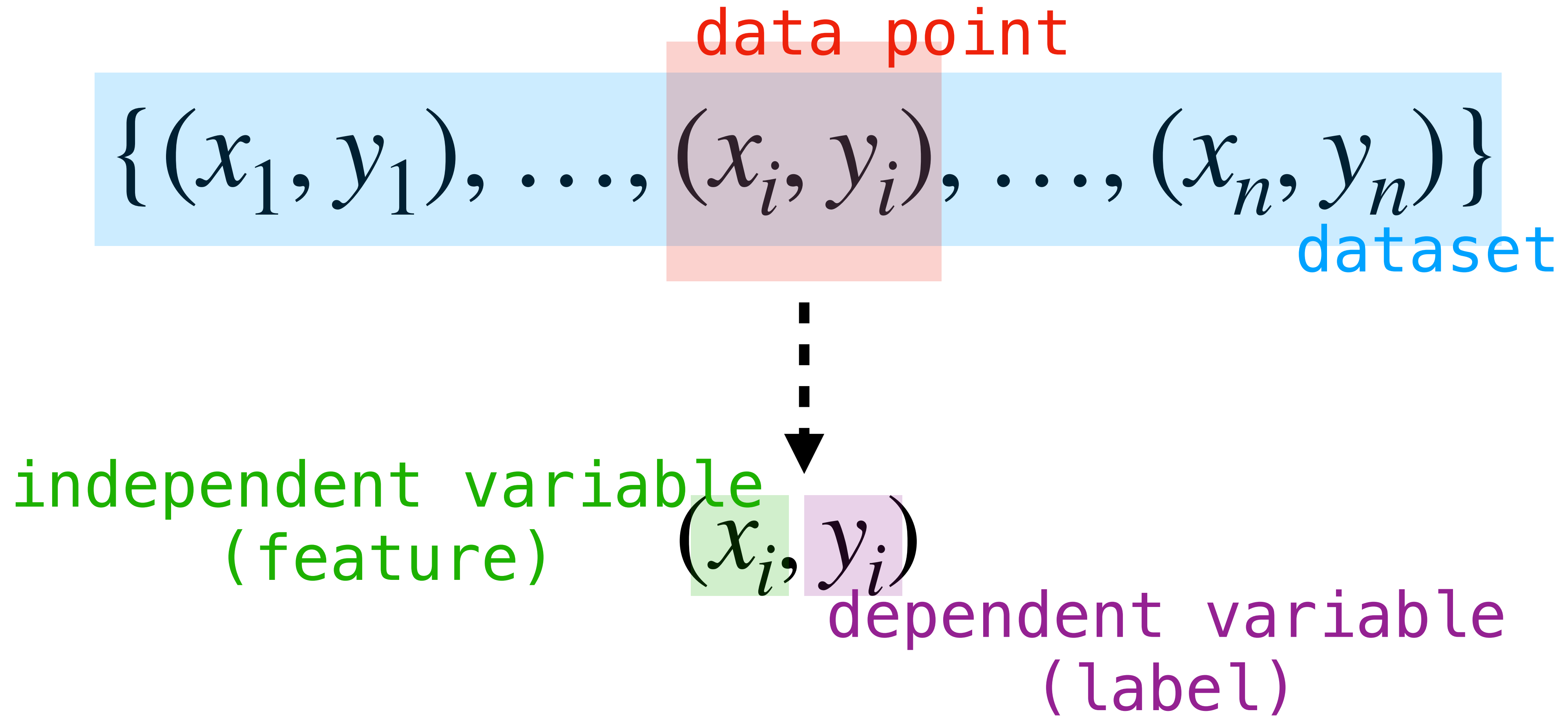
Terminology: Datasets



Terminology: Datasets



Terminology: Datasets



Terminology: Models

$$f(x) = \beta_0 + \beta_1 x$$

Terminology: Models

$$f(x) = \beta_0 + \beta_1 x$$

model

Terminology: Models

model parameters/
regression coefficients

$$f(x) = \beta_0 + \beta_1 x$$

model

Terminology: Least-Squares Error

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

Terminology: Least-Squares Error

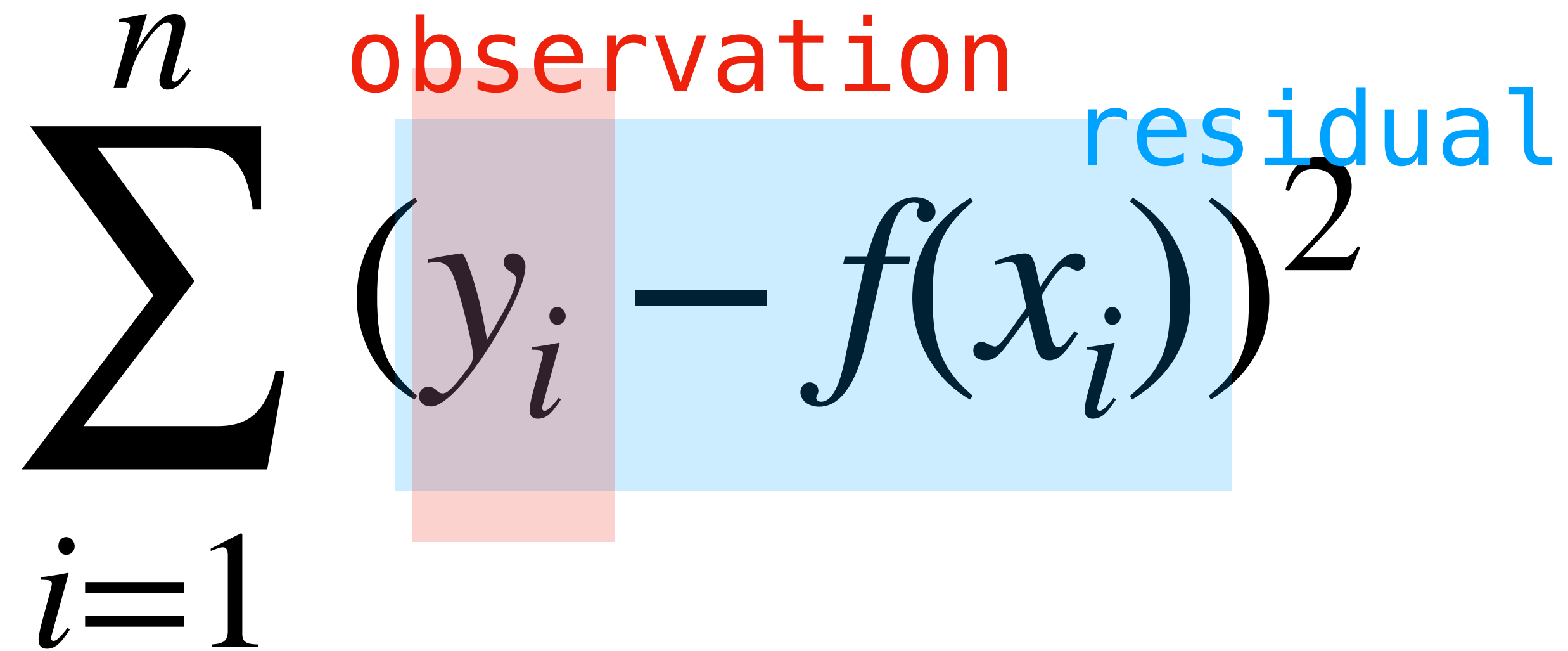
$$\sum_{i=1}^n (y_i - f(x_i))^2$$

residual

Terminology: Least-Squares Error

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

observation residual

The diagram shows the least-squares error formula with two highlighted regions. A light red rectangular box highlights the term y_i , which is labeled "observation" in red text above it. A light blue rectangular box highlights the term $f(x_i)$, which is labeled "residual" in blue text above it. The entire expression is enclosed in large parentheses with a superscript of 2.

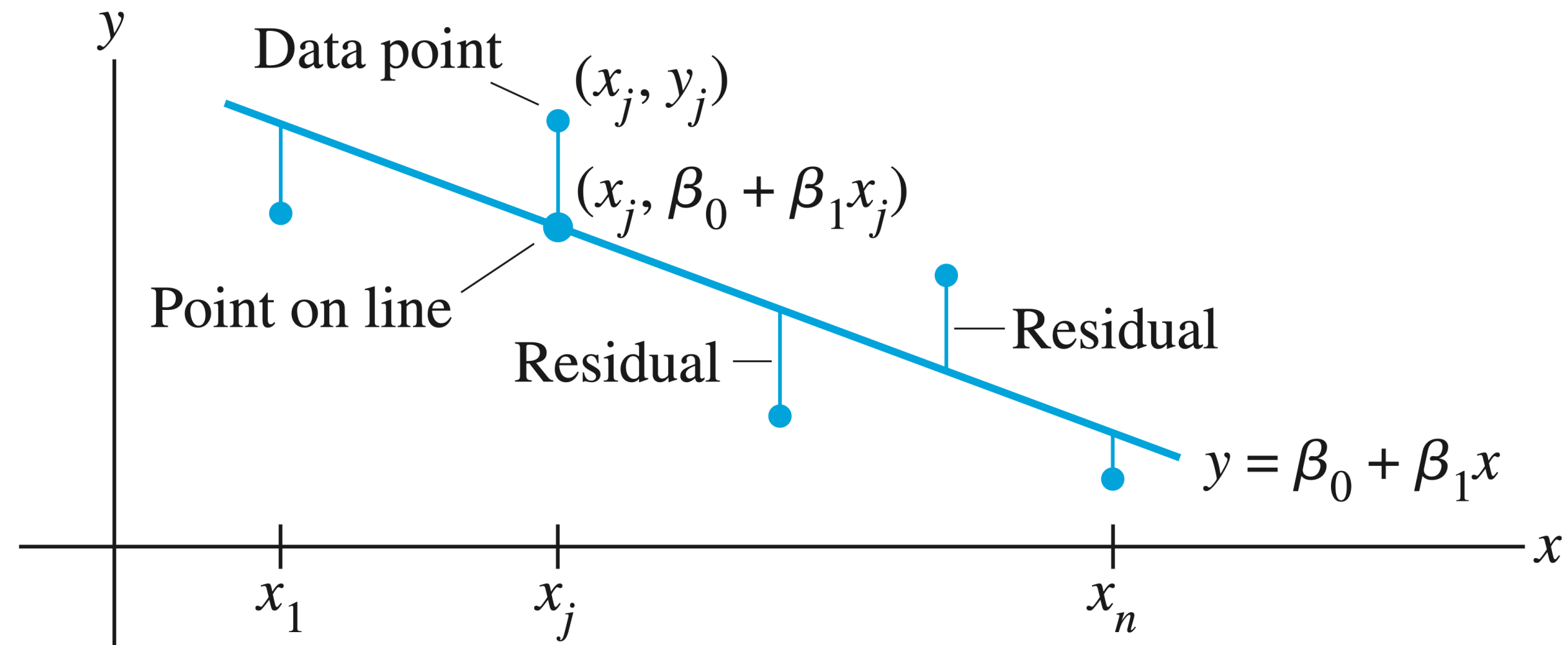
Terminology: Least-Squares Error

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

The diagram illustrates the least-squares error formula with color-coded components:

- observation**: The term y_i is highlighted in a light red box.
- prediction**: The term $f(x_i)$ is highlighted in a light green box.
- residual**: The entire expression $(y_i - f(x_i))^2$ is highlighted in a light blue box.

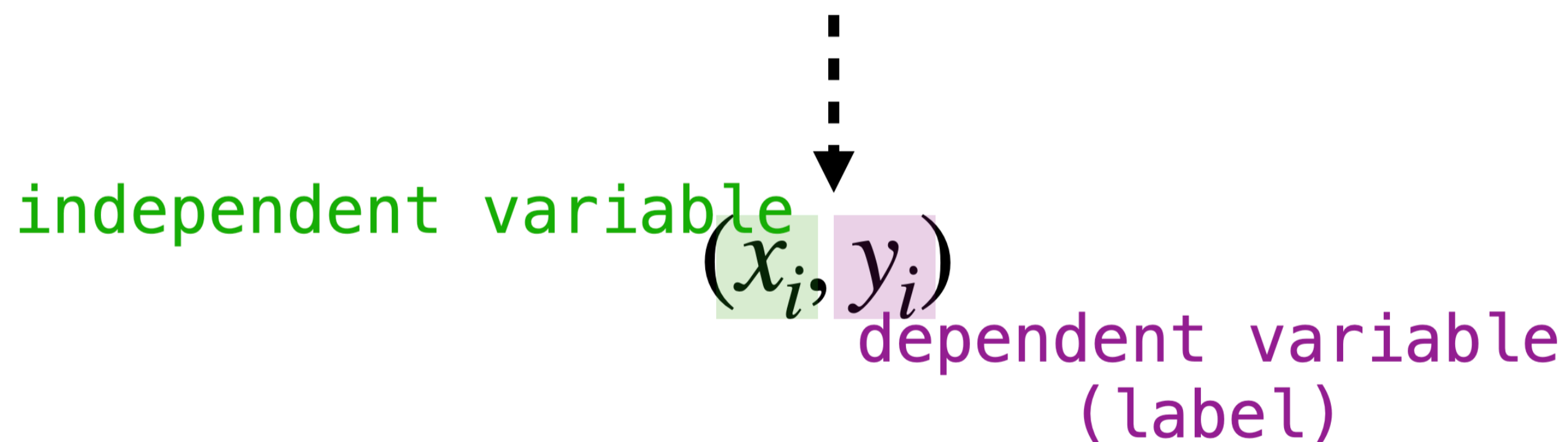
Terminology



$$\{(x_1, y_1), \dots, \overset{\text{data point}}{(x_i, y_i)}, \dots, (x_n, y_n)\}$$

dataset

$$f(x) = \overset{\text{model parameters/ regression coefficients}}{\beta_0} + \overset{\text{model}}{\beta_1 x}$$



$$\sum_{i=1}^n \overset{\text{observation}}{(y_i - \overset{\text{prediction}}{f(x_i)})^2}$$

residual

How to: Finding the Least Squares Line

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$$

How to: Finding the Least Squares Line

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$$

Problem. Find the least squares line for the dataset $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

How to: Finding the Least Squares Line

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad \beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$$

Problem. Find the least squares line for the dataset $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

Solution (First attempt). Use these equations...

How to: Finding the Least Squares Line

Don't memorize these.

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \quad \beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$$

Problem. Find the least squares line for the dataset $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

Solution (First attempt). Use these equations...

An Observation

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_{i=1}^n ((A\mathbf{x})_i - \mathbf{b}_i)^2$$

An Observation

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

minimize for least-squares line

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_{i=1}^n ((A\mathbf{x})_i - \mathbf{b}_i)^2$$

minimize for least-squares method

An Observation

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

minimize for least-squares line

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_{i=1}^n ((A\mathbf{x})_i - \mathbf{b}_i)^2$$

minimize for least-squares method

These expressions look very similar.

An Observation

$$\sum_{i=1}^n (y_i - f(x_i))^2$$

minimize for least-squares line

$$\|A\mathbf{x} - \mathbf{b}\|^2 = \sum_{i=1}^n ((A\mathbf{x})_i - \mathbf{b}_i)^2$$

minimize for least-squares method

These expressions look very similar.

Can we design a matrix where finding a least squares solution gives us a least squares line?

A Least Squares Problem

$$\beta_0 + \beta_1 x_1 = y_1$$

$$\beta_0 + \beta_1 x_2 = y_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_n = y_n$$

A Least Squares Problem

In the "ideal" world, we could find parameters β_0 and β_1 such that all of these equations hold.

$$\beta_0 + \beta_1 x_1 = y_1$$

$$\beta_0 + \beta_1 x_2 = y_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_n = y_n$$

A Least Squares Problem

In the "ideal" world, we could find parameters β_0 and β_1 such that all of these equations hold.

This would mean **all the points already lie on a single line.**

$$\beta_0 + \beta_1 x_1 = y_1$$

$$\beta_0 + \beta_1 x_2 = y_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_n = y_n$$

A Least Squares Problem

In the "ideal" world, we could find parameters β_0 and β_1 such that all of these equations hold.

This would mean **all the points already lie on a single line.**

This is a linear system in the variables β_0 and β_1

$$\begin{aligned}\beta_0 + \beta_1 x_1 &= y_1 \\ \beta_0 + \beta_1 x_2 &= y_2 \\ &\vdots \\ \beta_0 + \beta_1 x_n &= y_n\end{aligned}$$

A Least Squares Problem

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A Least Squares Problem

In the "ideal" world,
*this matrix equation
has a solution.*

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A Least Squares Problem

In the "ideal" world,
*this matrix equation
has a solution.*

In reality this system
is unlikely to have a
solution, **but maybe we
can find an
approximate solution.**

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

A Least Squares Problem

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\|X\vec{\beta} - \mathbf{y}\|^2 = \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2$$

A Least Squares Problem

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\|X\vec{\beta} - \mathbf{y}\|^2 = \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2$$

The sum of squares of residuals is the squared distances between $X\beta$ and \mathbf{y} .

A Least Squares Problem

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\|X\vec{\beta} - \mathbf{y}\|^2 = \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2$$

The sum of squares of residuals is the squared distances between $X\beta$ and \mathbf{y} .

Least squares solutions to this system give us parameters for least squares lines.

Just for Fun

$$\beta_1 = \frac{n \sum_i x_i y_i - \left(\sum_i x_i \right) \left(\sum_i y_i \right)}{n \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$

Let's derive it:

$$X = \begin{bmatrix} \vec{1} & \vec{x} \end{bmatrix}$$

$$X^T X = \begin{bmatrix} \vec{1}^T \\ \vec{x}^T \end{bmatrix} \begin{bmatrix} \vec{1} & \vec{x} \end{bmatrix} = \begin{bmatrix} 1 \cdot 1 & 1 \cdot x \\ x \cdot 1 & x \cdot x \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{(1 \cdot 1)(x \cdot x) - (x \cdot 1)^2} \begin{bmatrix} x \cdot x & -1 \cdot x \\ -x \cdot 1 & 1 \cdot 1 \end{bmatrix}$$

$$X^T \vec{y} = \begin{bmatrix} \vec{1}^T \\ \vec{x}^T \end{bmatrix} \vec{y} = \begin{bmatrix} 1 \cdot y \\ x \cdot y \end{bmatrix} \frac{1}{n(x \cdot x) - (x \cdot 1)^2} \begin{bmatrix} \dots \end{bmatrix} \begin{bmatrix} 1 \cdot y \\ x \cdot y \end{bmatrix}$$

How To: Least Squares Line

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

How To: Least Squares Line

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Problem. Find the least squares line for the dataset $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

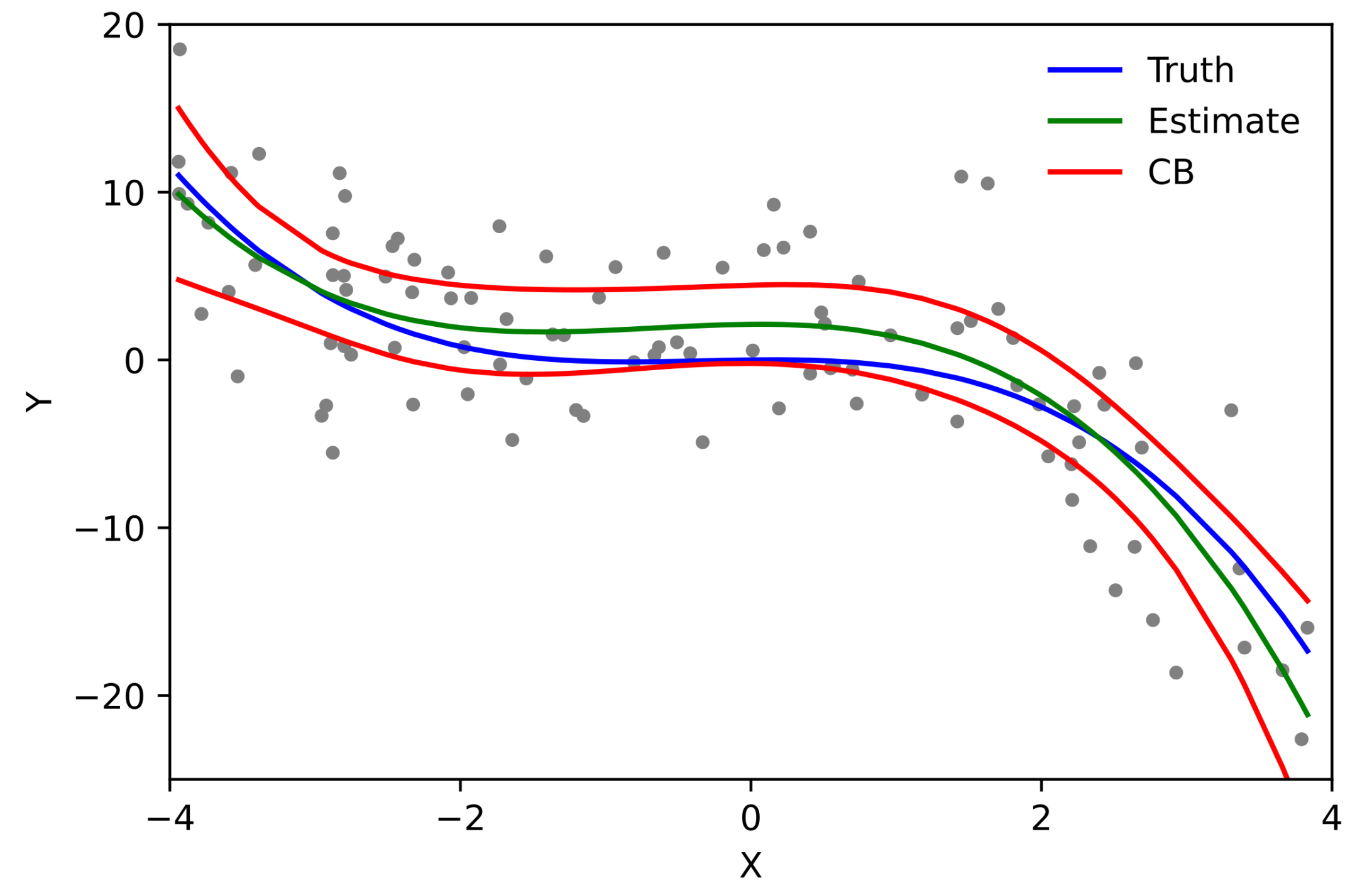
How To: Least Squares Line

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Problem. Find the least squares line for the dataset $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

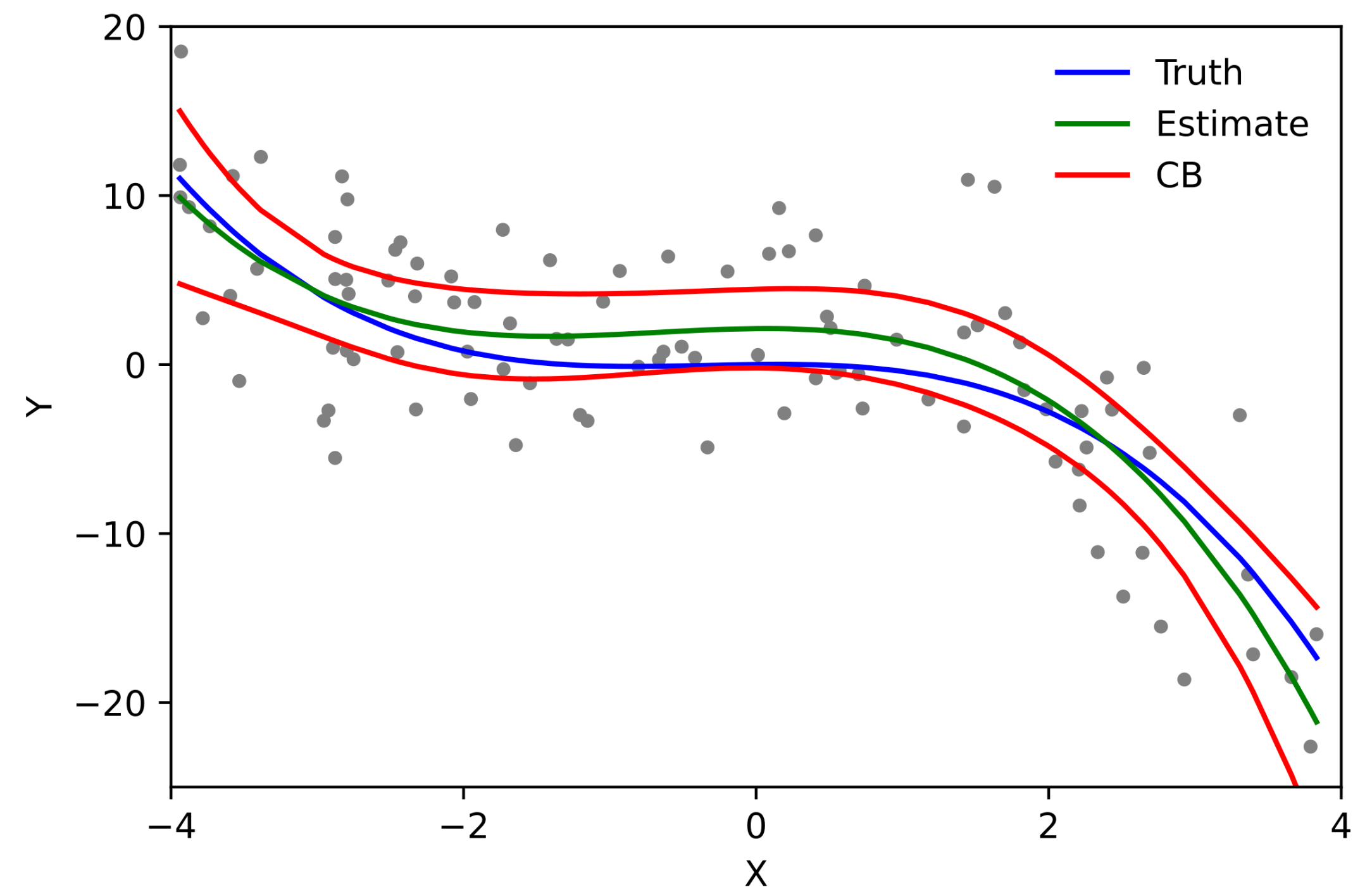
Solution. Find the least squares solution to the above equation.

General Regression



General Regression

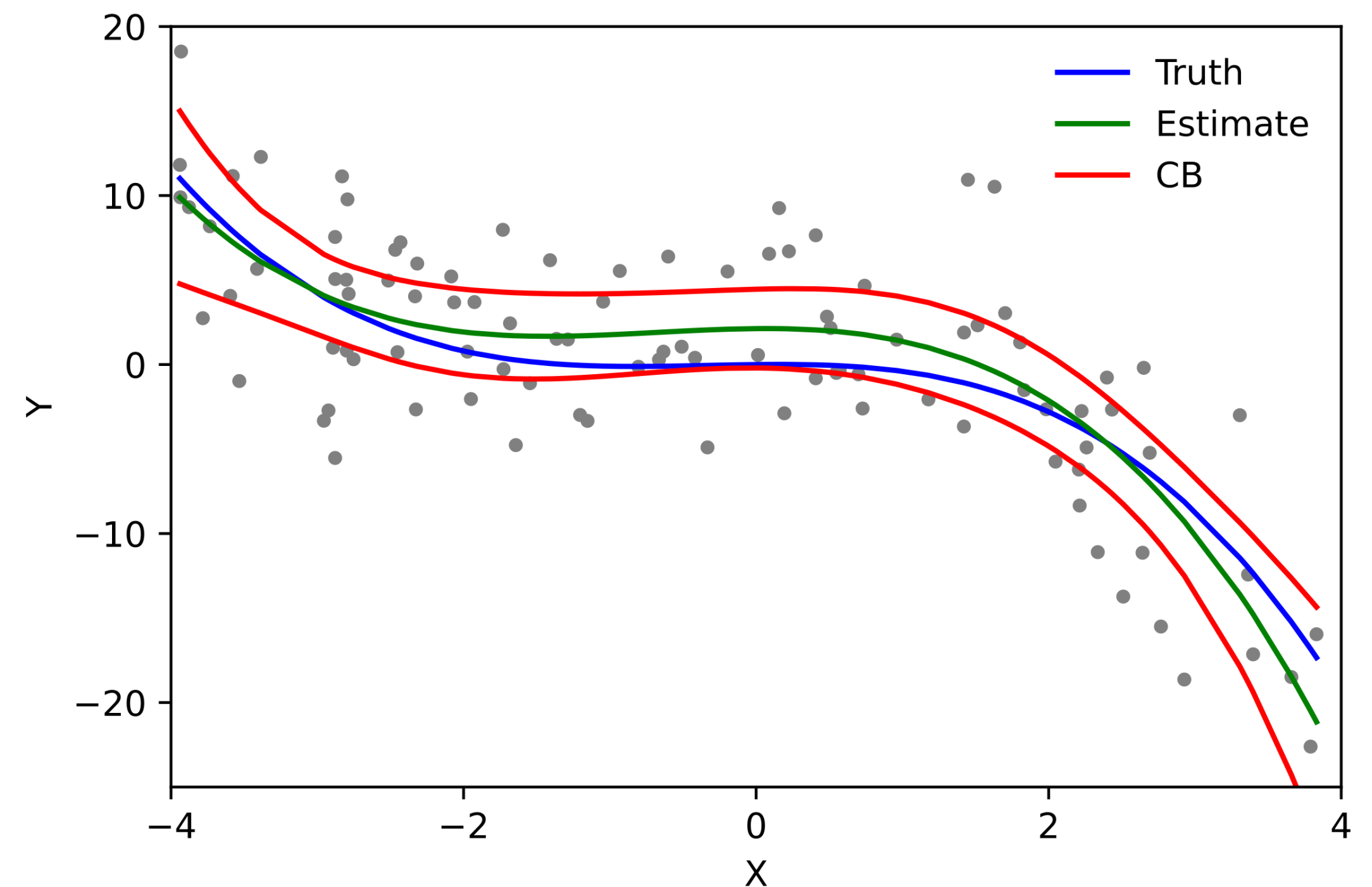
Regression is the process of estimating the relationships independent and dependent variables in a dataset.



General Regression

Regression is the process of estimating the relationships independent and dependent variables in a dataset.

What we are estimating is a mathematical function

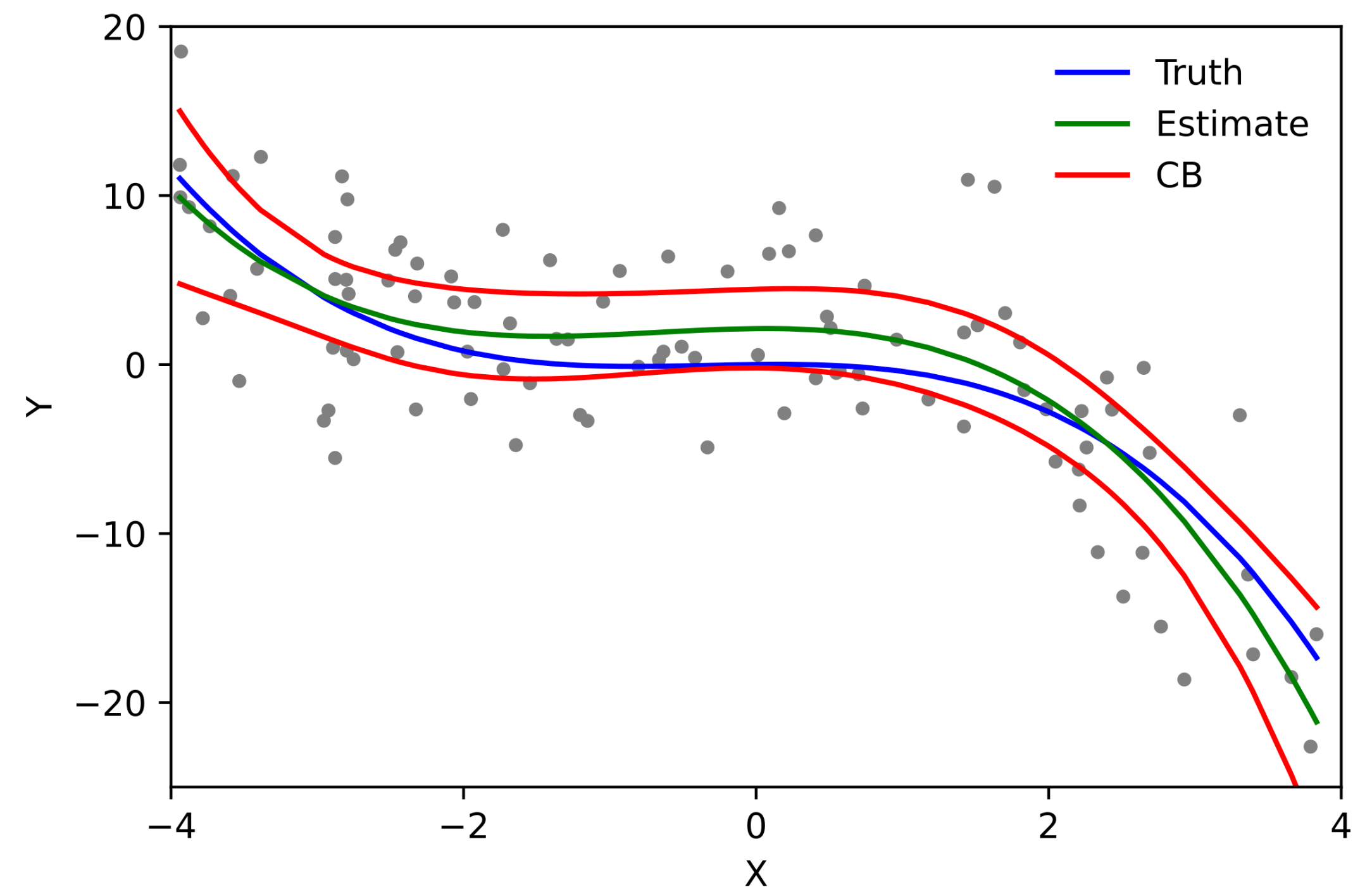


General Regression

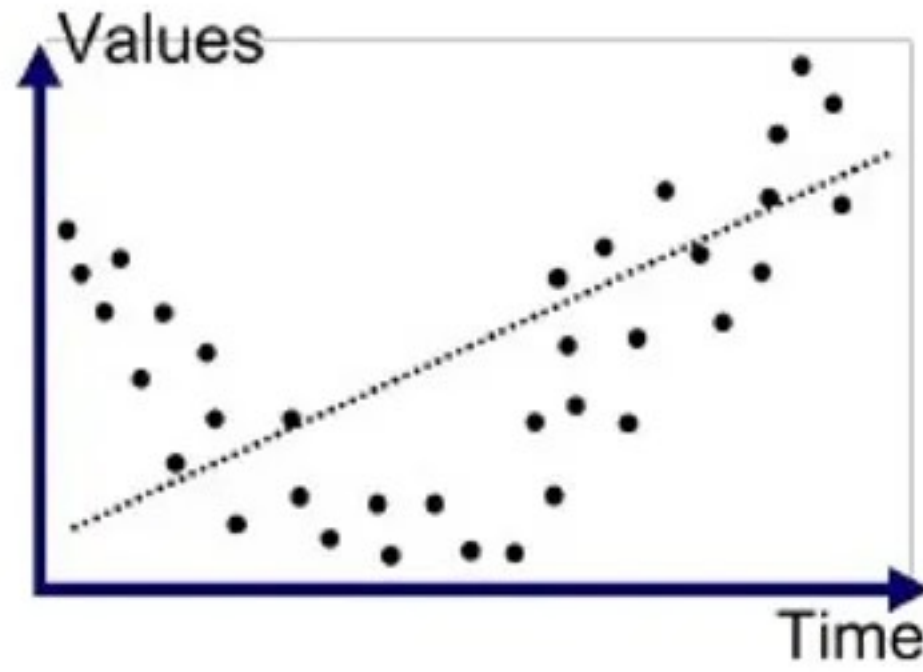
Regression is the process of estimating the relationships independent and dependent variables in a dataset.

What we are estimating is a mathematical function

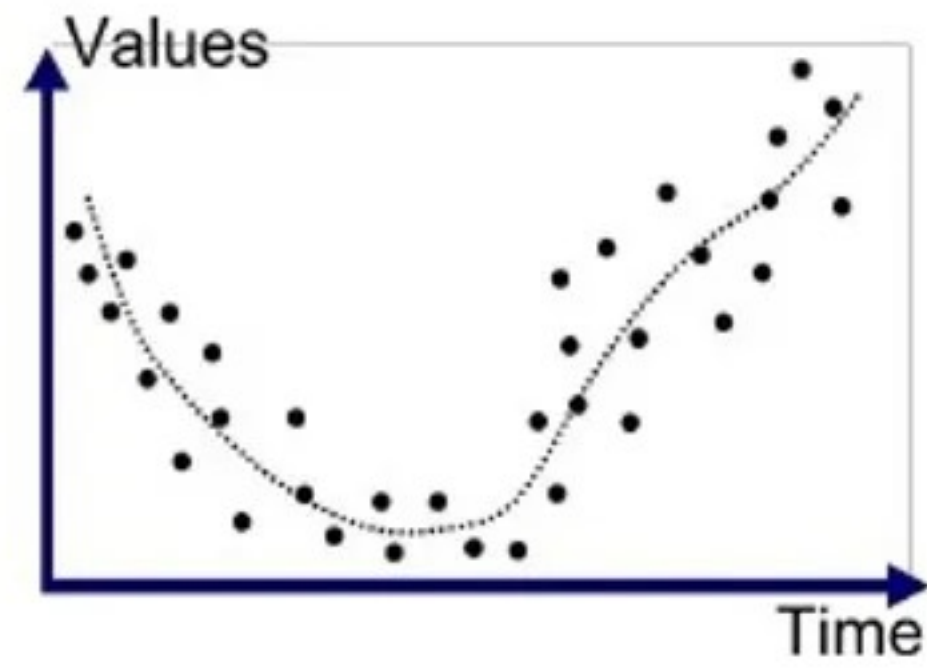
We think of the environment has providing us a function from our independent variables to our dependent variables.



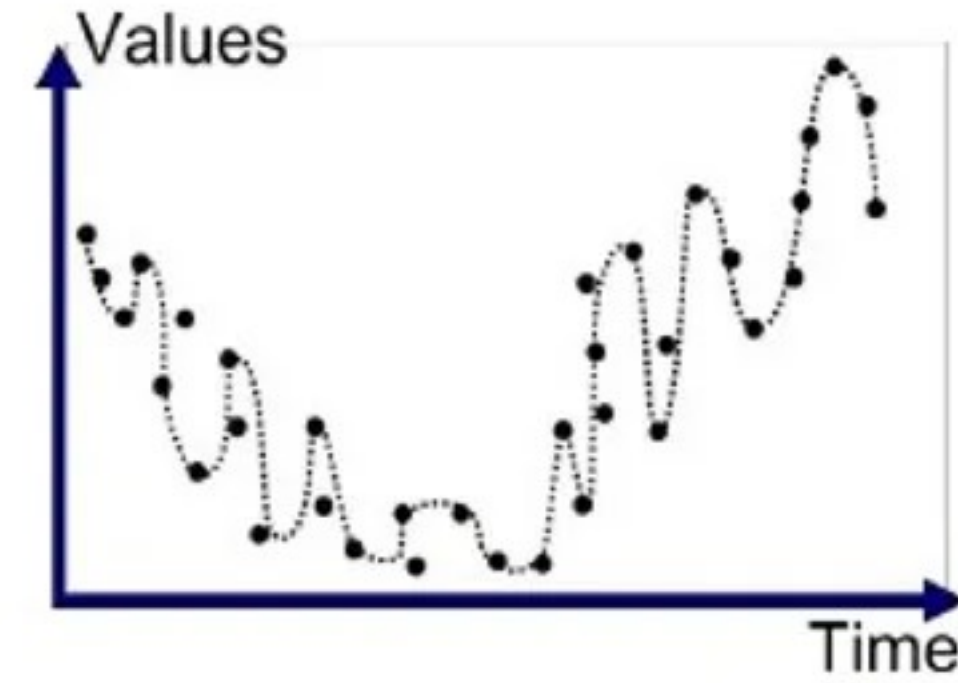
Models



Underfitted

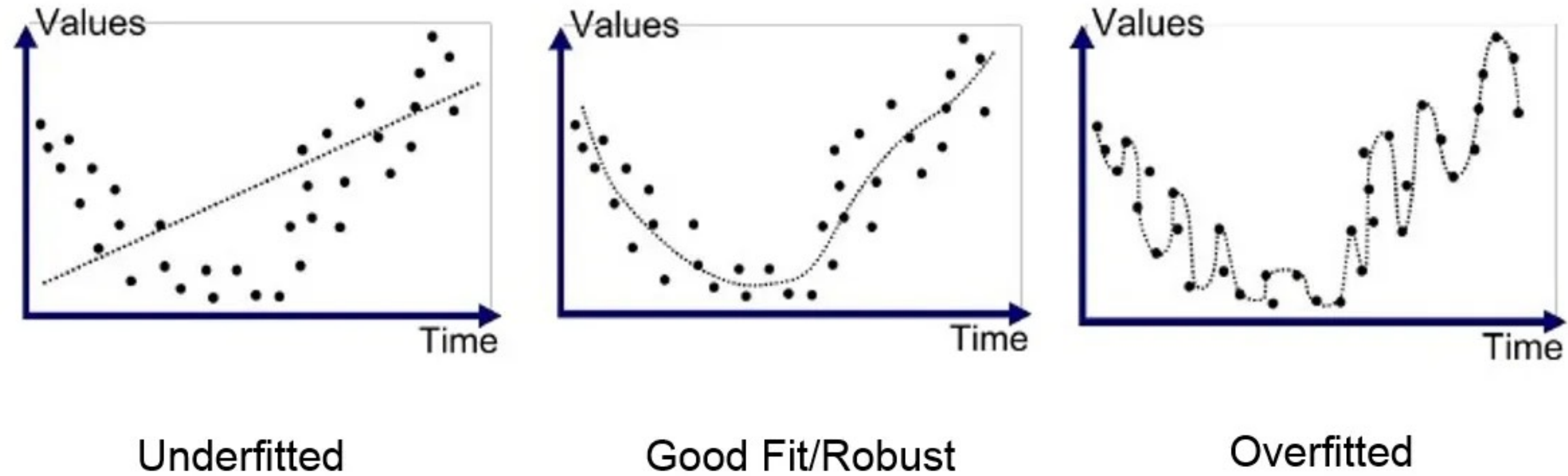


Good Fit/Robust



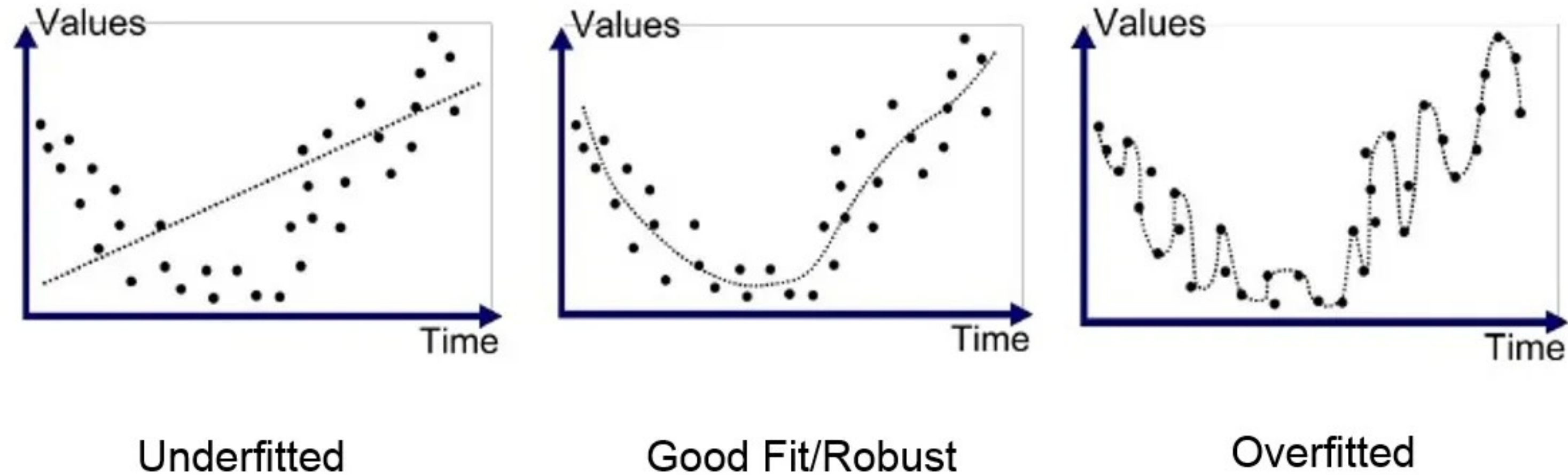
Overfitted

Models



Therefore, a *model* is a mathematical function.

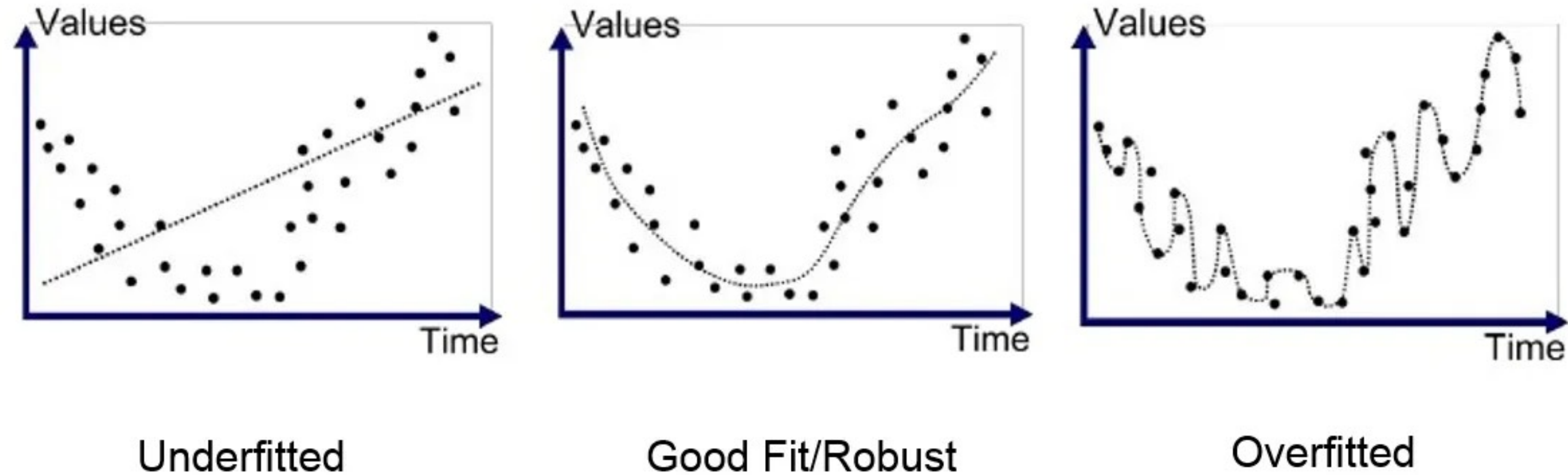
Models



Therefore, a *model* is a mathematical function.

We're interested in finding mathematical functions that "correctly" model the data we've seen.

Models

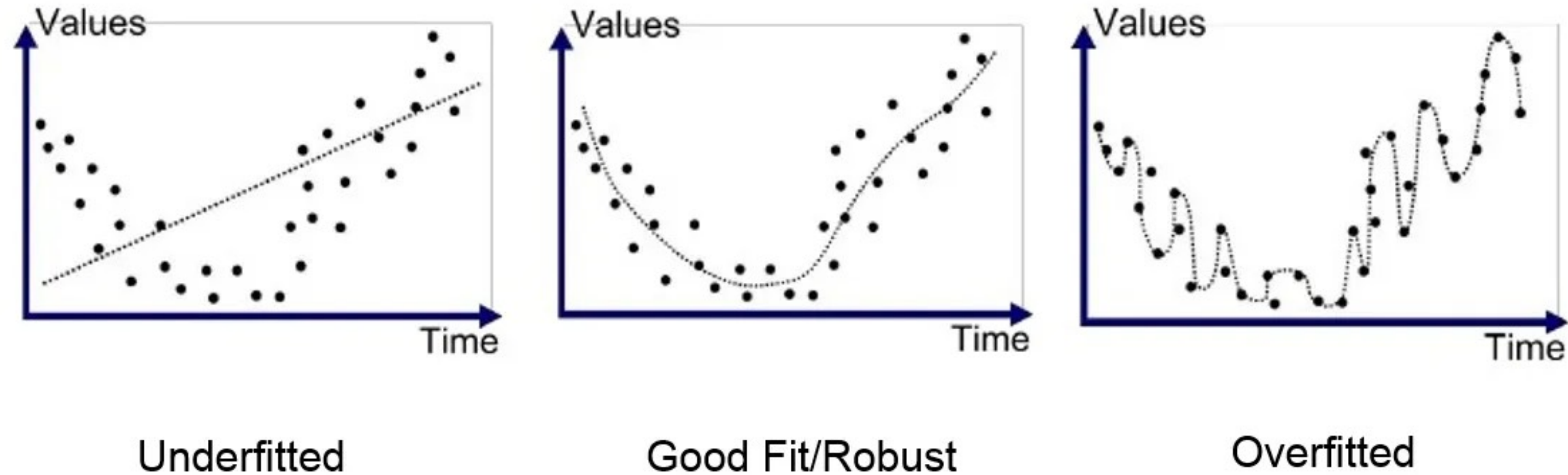


Therefore, a *model* is a mathematical function.

We're interested in finding mathematical functions that "correctly" model the data we've seen.

But this would be a bit boring if we *just* wanted to model data we've seen.

Models



Therefore, a *model* is a mathematical function.

We're interested in finding mathematical functions that "correctly" model the data we've seen.

But this would be a bit boring if we *just* wanted to model data we've seen.

(Advanced) We pick models from weaker classes of functions so that they are more robust when we **predict** values using the model.

How To: Prediction

How To: Prediction

Problem. Given the data $\{(x_1, y_1), \dots, (x_k, y_k)\}$ use the line of best fit to predict the value of y' for the input x' .

How To: Prediction

Problem. Given the data $\{(x_1, y_1), \dots, (x_k, y_k)\}$ use the line of best fit to predict the value of y' for the input x' .

Solution. Find the best fit line $f(x) = \beta_0 + \beta_1 x$.
The predicted value of x' is $f(x')$.

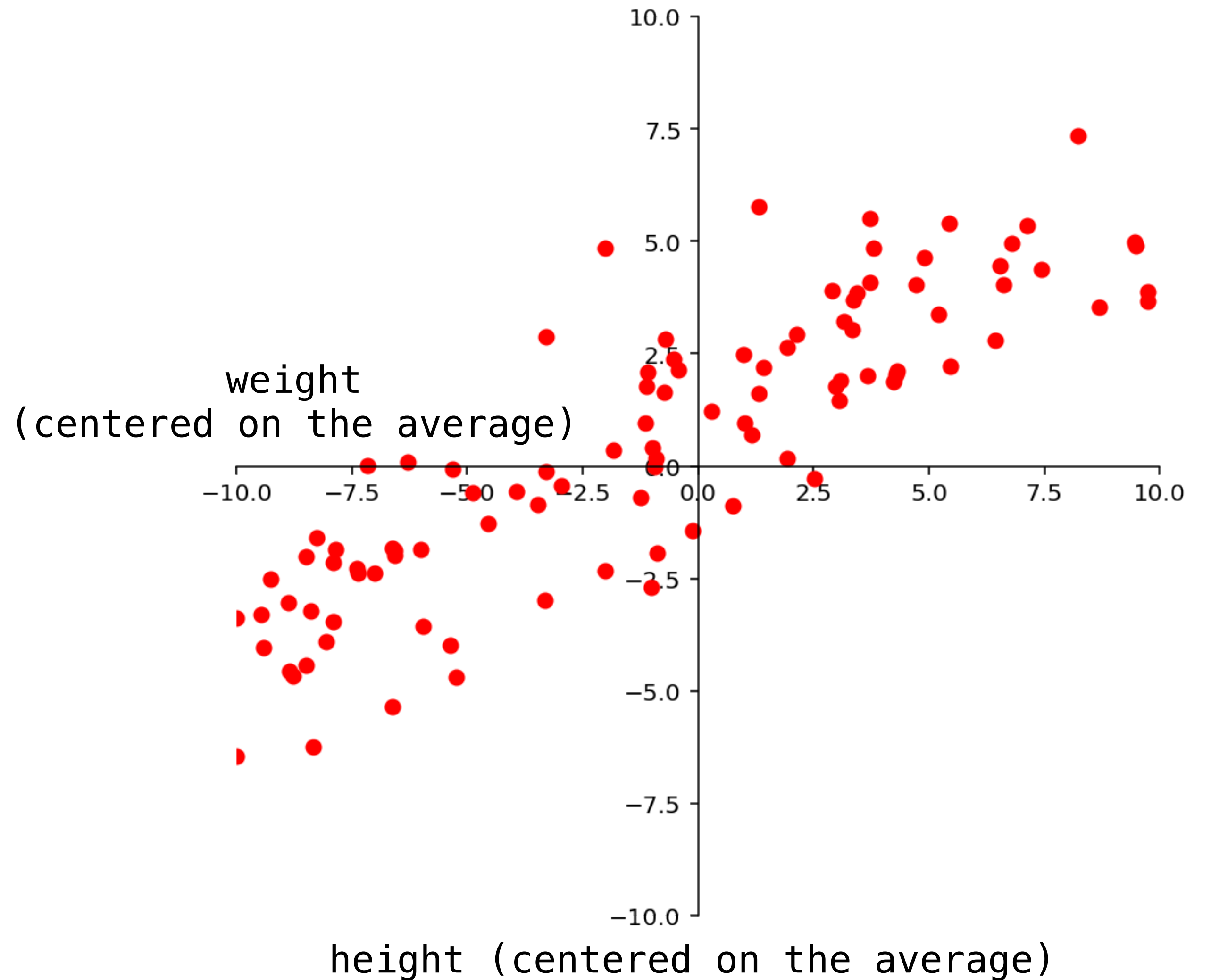
How To: Prediction

Problem. Given the data $\{(x_1, y_1), \dots, (x_k, y_k)\}$ use the line of best fit to predict the value of y' for the input x' .

Solution. Find the best fit line $f(x) = \beta_0 + \beta_1 x$.
The predicted value of x' is $f(x')$.

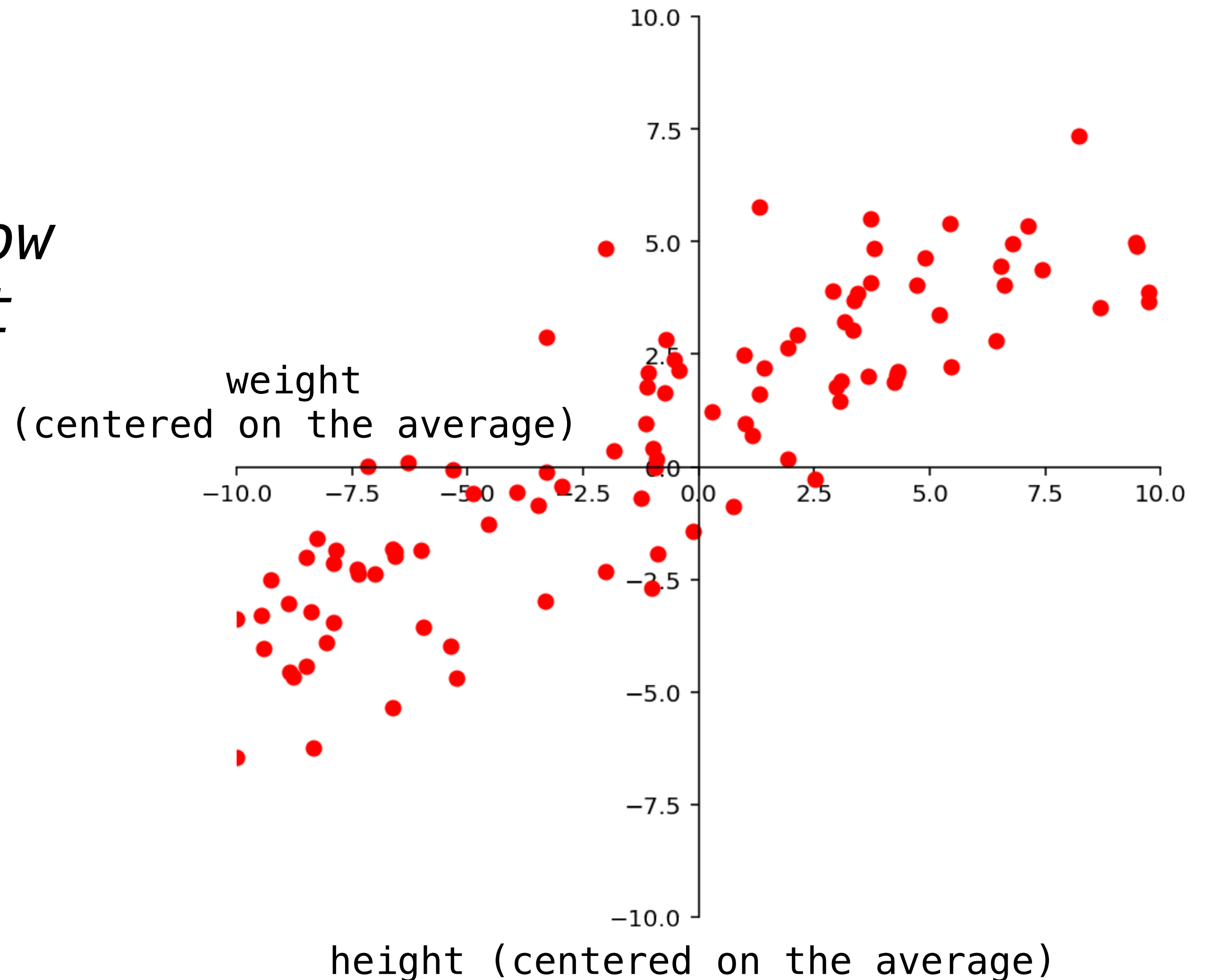
**This generalizes to any
model fitting problem**

Example: Height from Weight



Example: Height from Weight

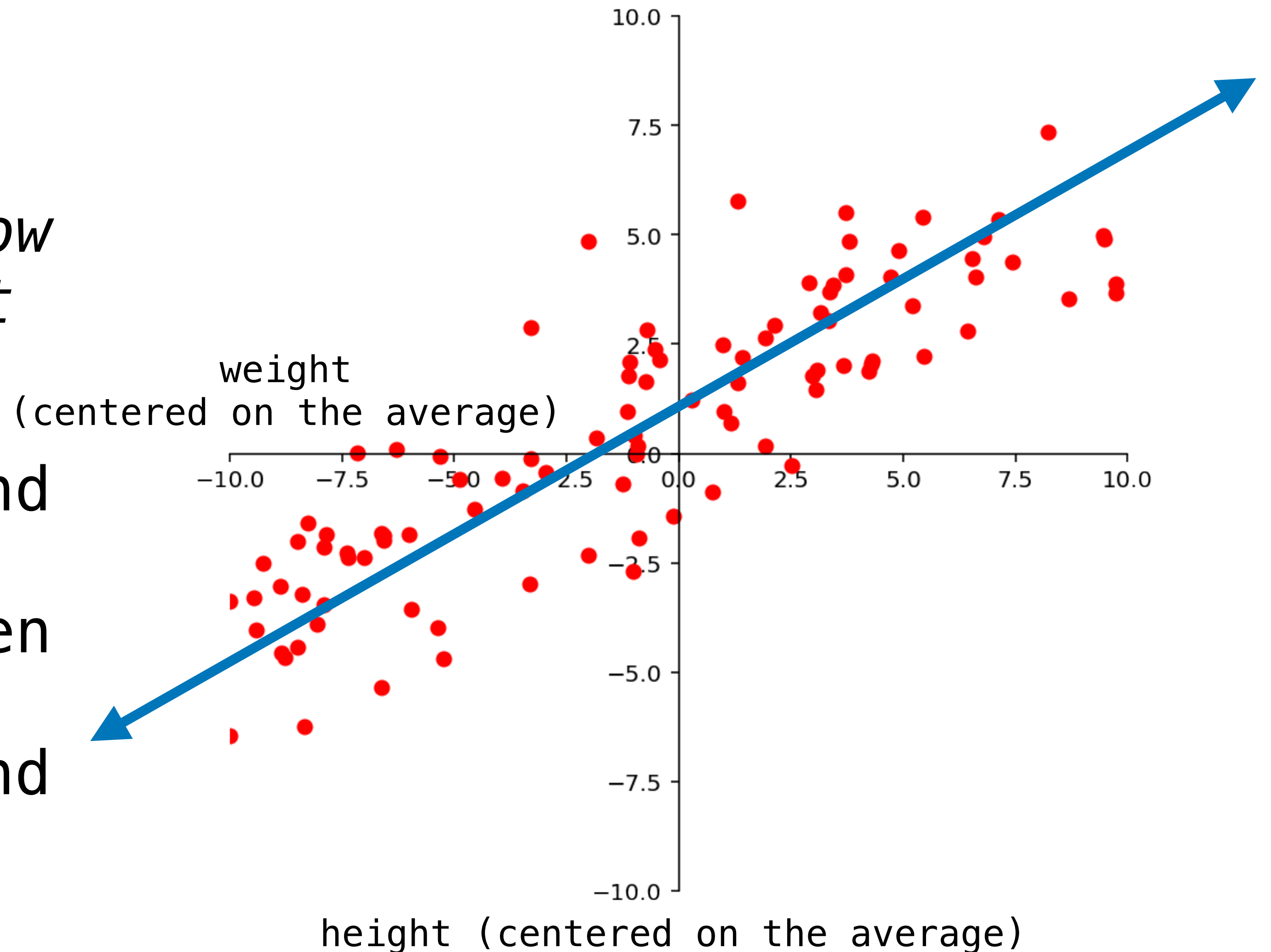
Suppose we know that person X weighs 150lb. *How would we guess the height of person X ?*



Example: Height from Weight

Suppose we know that person X weighs 150lb. *How would we guess the height of person X ?*

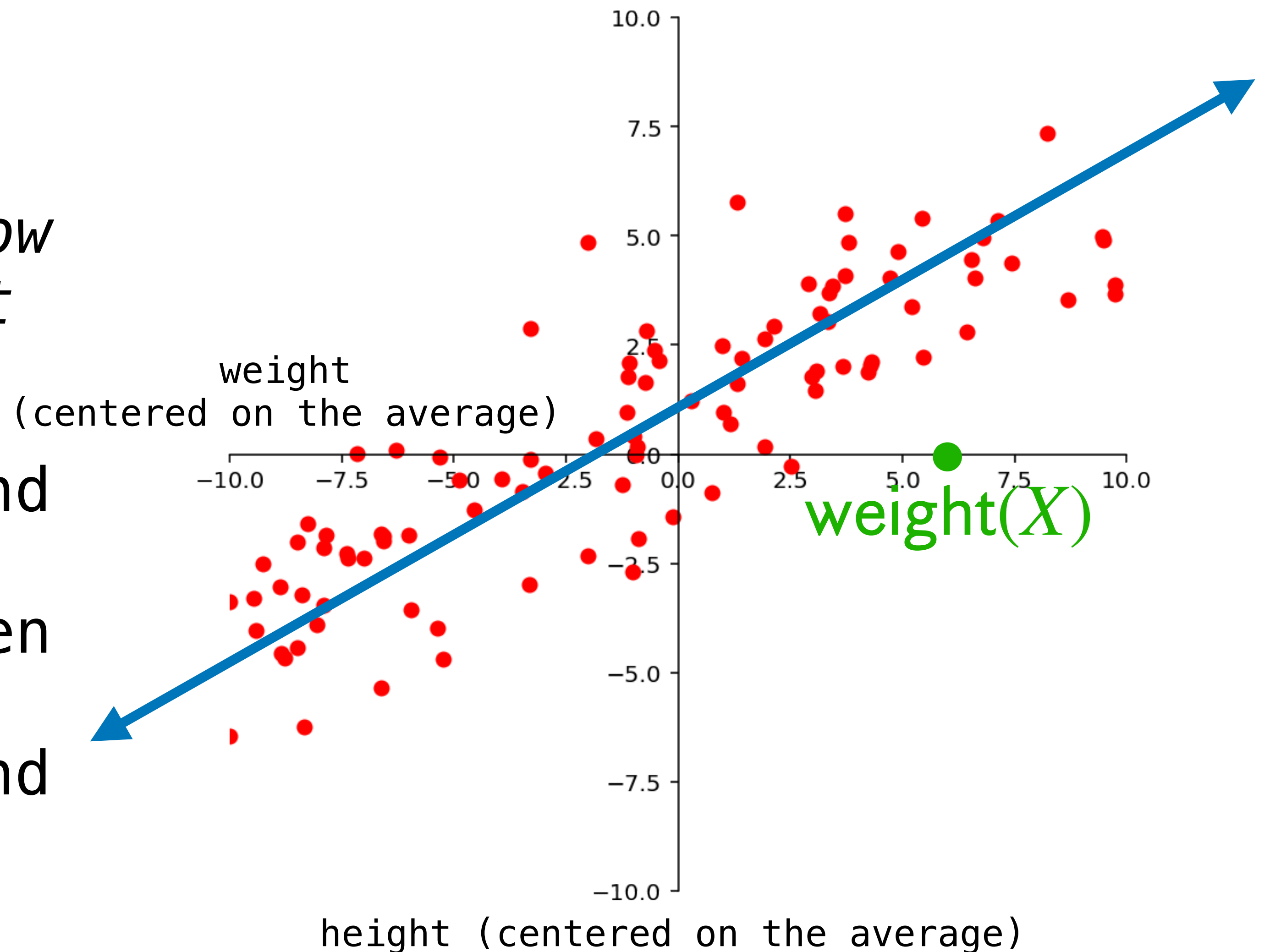
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



Example: Height from Weight

Suppose we know that person X weighs 150lb. *How would we guess the height of person X ?*

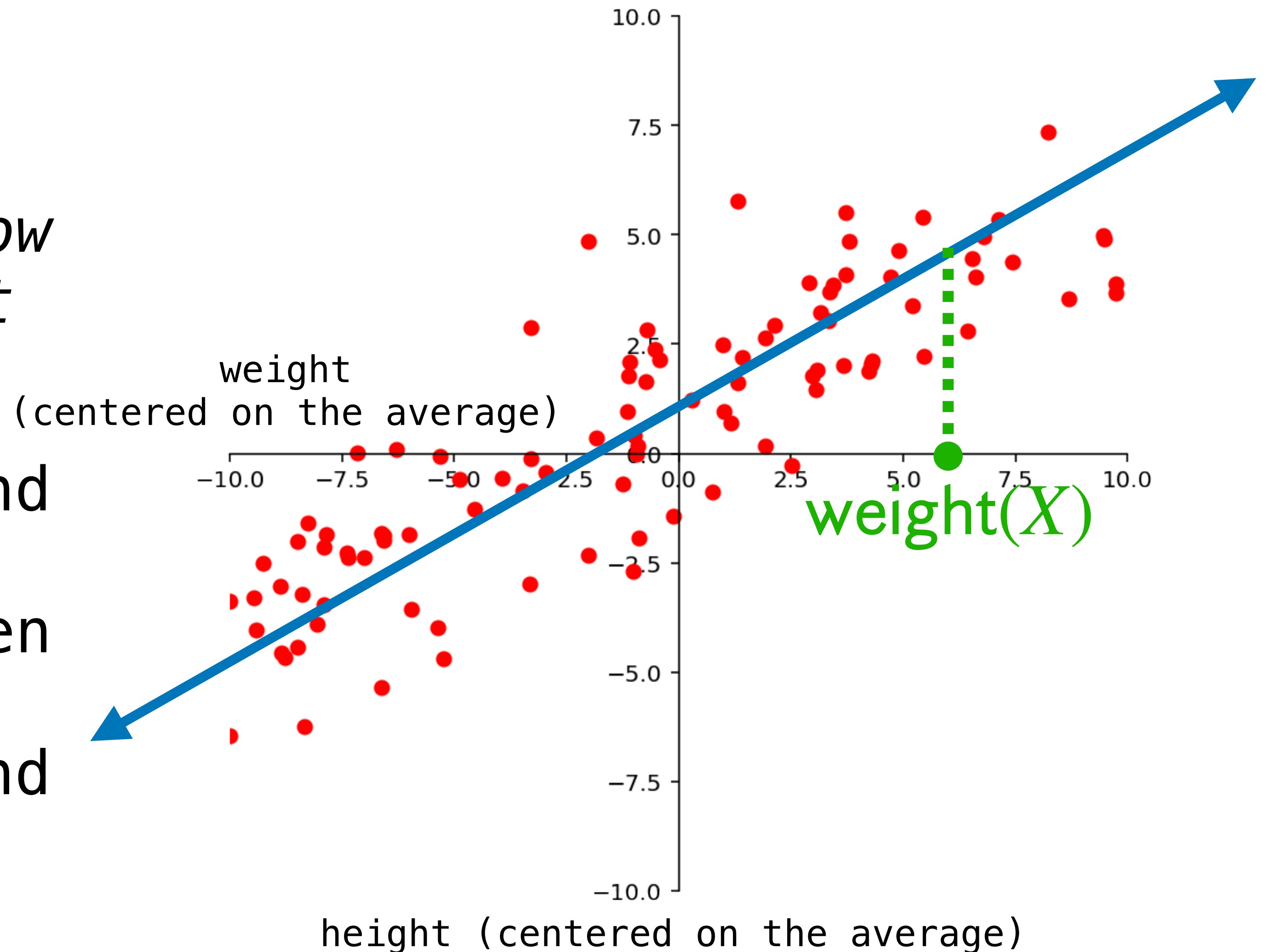
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



Example: Height from Weight

Suppose we know that person X weighs 150lb. *How would we guess the height of person X ?*

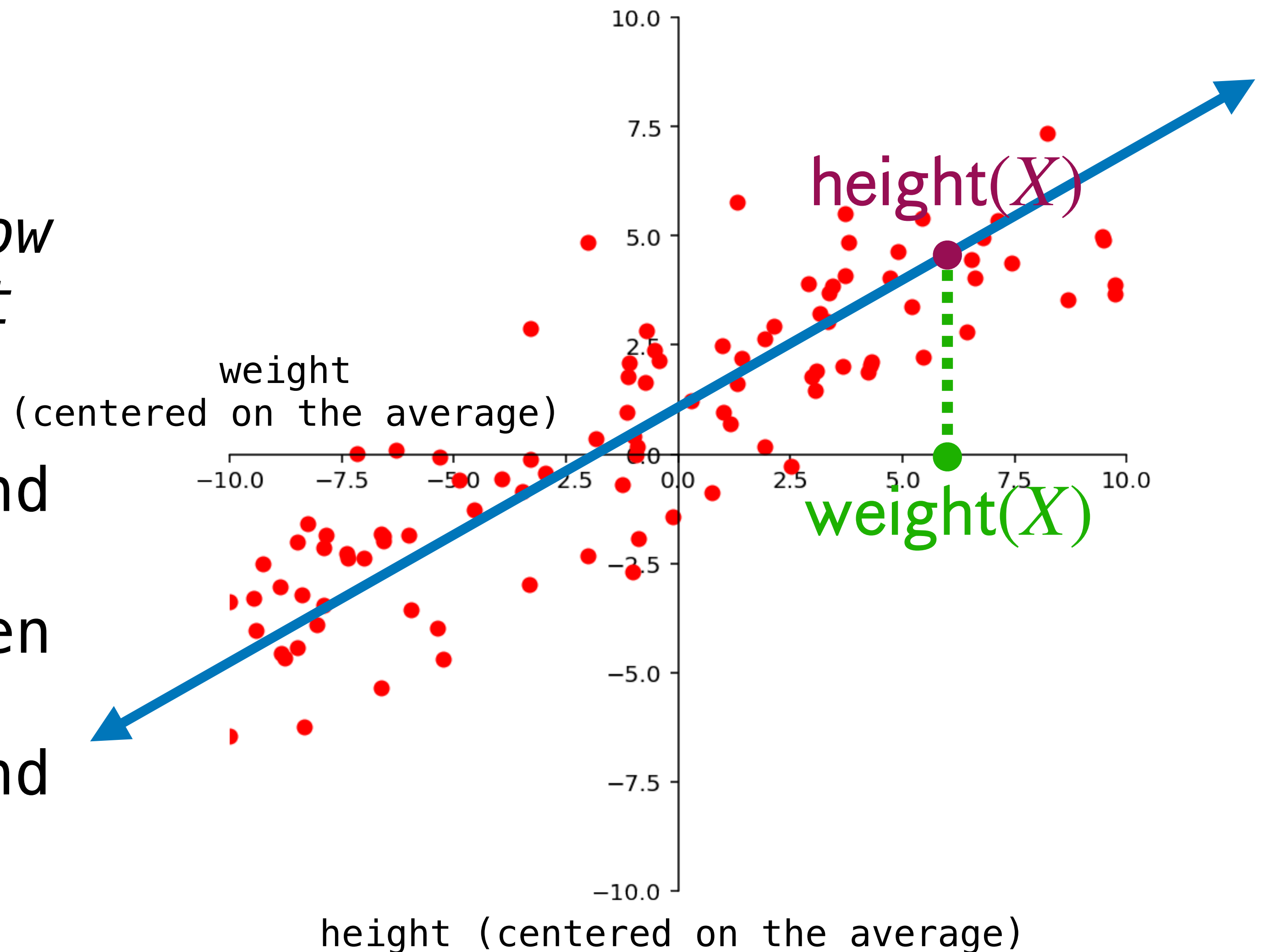
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



Example: Height from Weight

Suppose we know that person X weighs 150lb. *How would we guess the height of person X ?*

If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit** for that data and then use that function.



Question

Find the line of best fit for the dataset

$$\{(0,3), (1,1), (-1,1), (2,3)\}$$

If you have time, graph your result and use it to "predict" the corresponding value for the input 4.

$\{(0,3), (1,1), (-1,1), (2,3)\}$

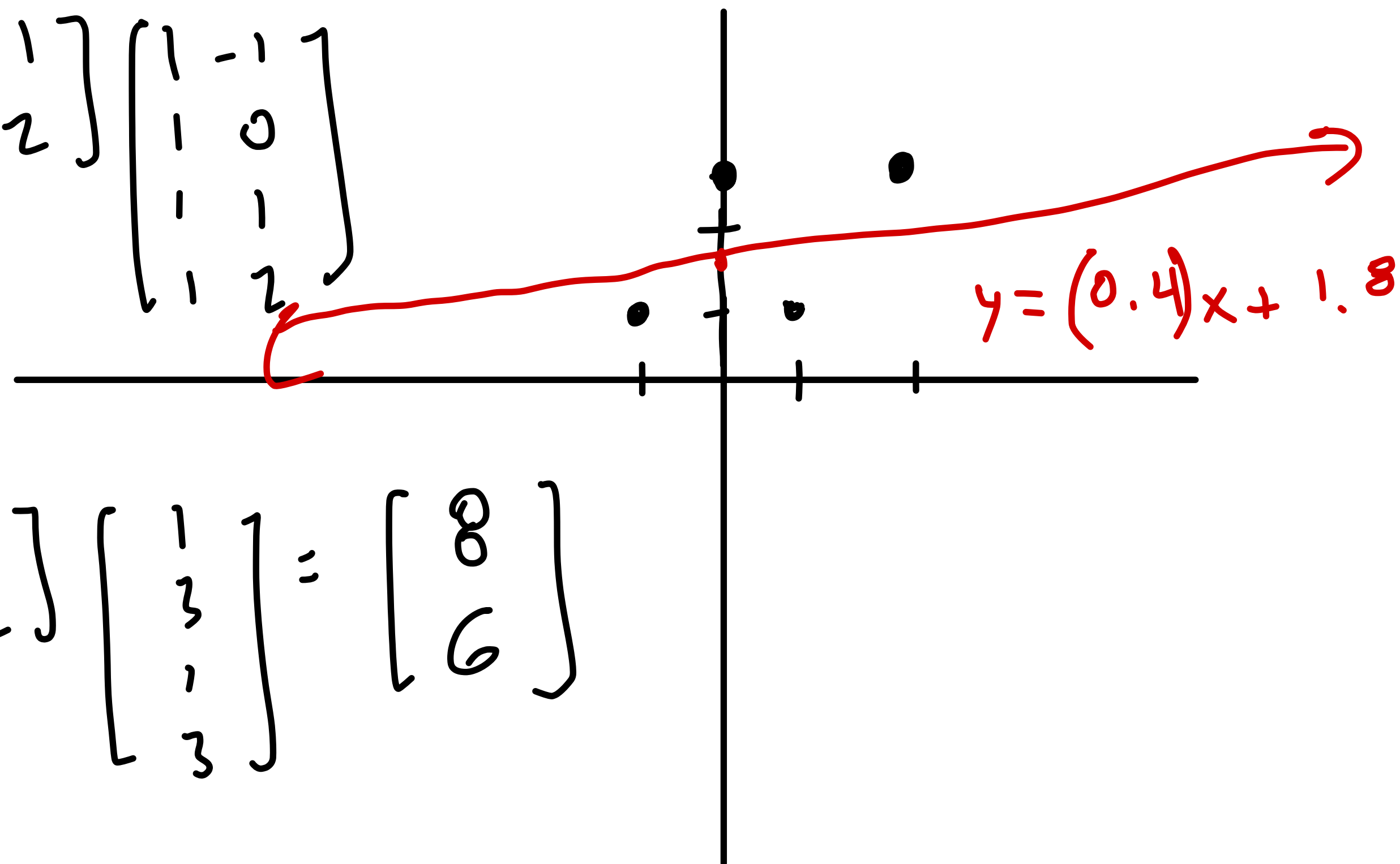
Answer

$$X = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & -1 \\ 1 & 2 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & -1 \\ 1 & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 4 & 2 \\ 2 & 6 \end{bmatrix}$$

$$\vec{y} = \begin{bmatrix} 1 \\ 3 \\ 1 \\ 3 \end{bmatrix}$$

$$X^T \vec{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 8 \\ 6 \end{bmatrix}$$



$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \vec{y} = \frac{1}{20} \begin{bmatrix} 6 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 8 \\ 6 \end{bmatrix}$$
$$= \frac{1}{5} \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} = \begin{bmatrix} 9/5 \\ 2/5 \end{bmatrix} = \begin{bmatrix} 1.8 \\ 0.4 \end{bmatrix}$$

Linear Models and Least Squares Regression

"Vectors" of Generalization

"Vectors" of Generalization

1. What if we have *more than one* independent value?

"Vectors" of Generalization

1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

"Vectors" of Generalization

1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

2. What if our data is not *exactly* linear.

"Vectors" of Generalization

1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

2. What if our data is not *exactly* linear.

e.g., polynomial regression

"Vectors" of Generalization

1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

2. What if our data is not *exactly* linear.

e.g., polynomial regression

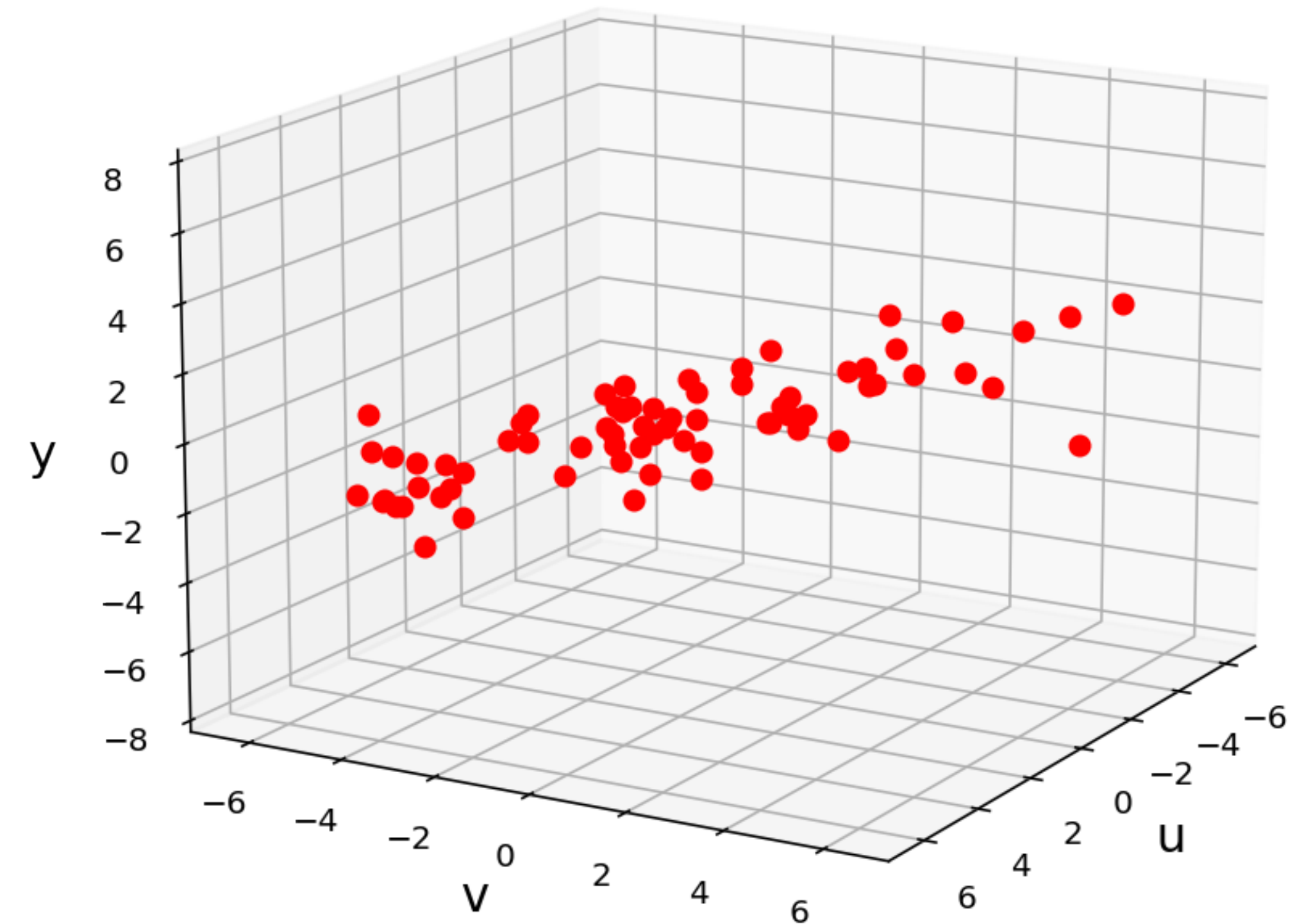
Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude
and latitude and z_i is an
altitude.

Problem: Find the plane
which "best" fits the
data.

Figure 23.1

Terrain Data for Multiple Regression



Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

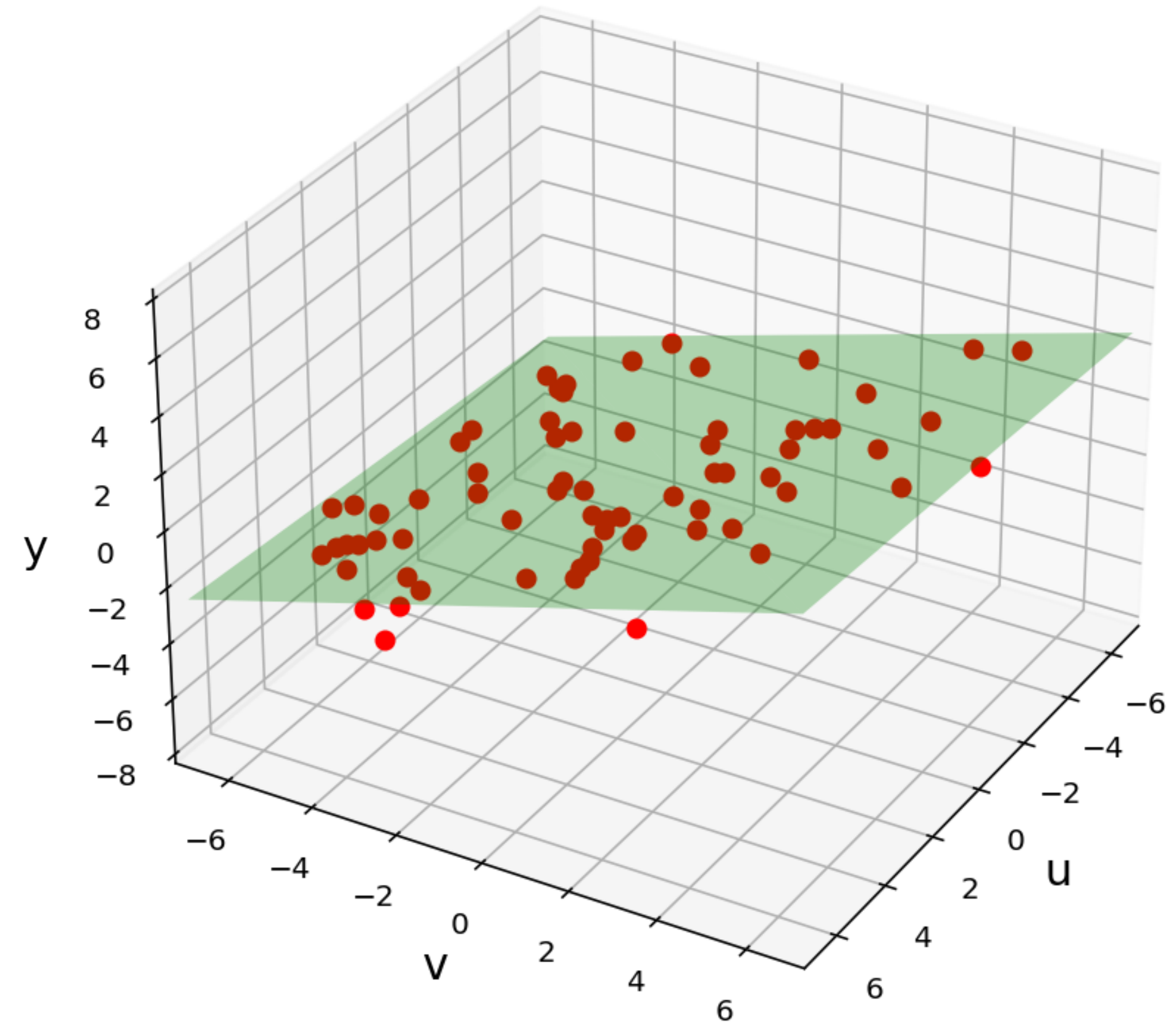
$$f(x, y) = \beta_0 + \beta_1x + \beta_2y$$

which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

Figure 23.2

Multiple Regression Fit to Data



Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

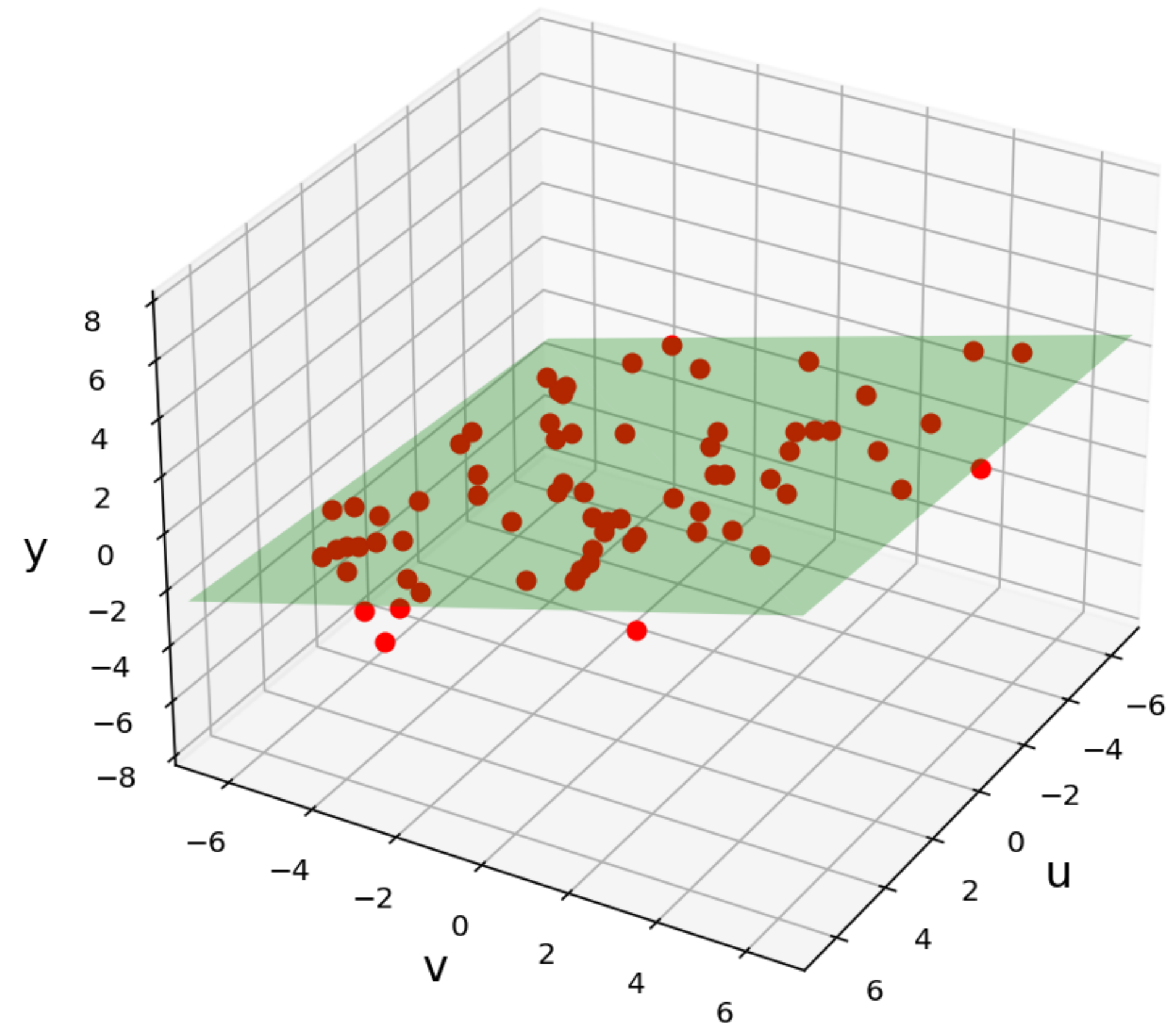
which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$f(x, y)$ is a good approximation of the altitude.

Figure 23.2

Multiple Regression Fit to Data



Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

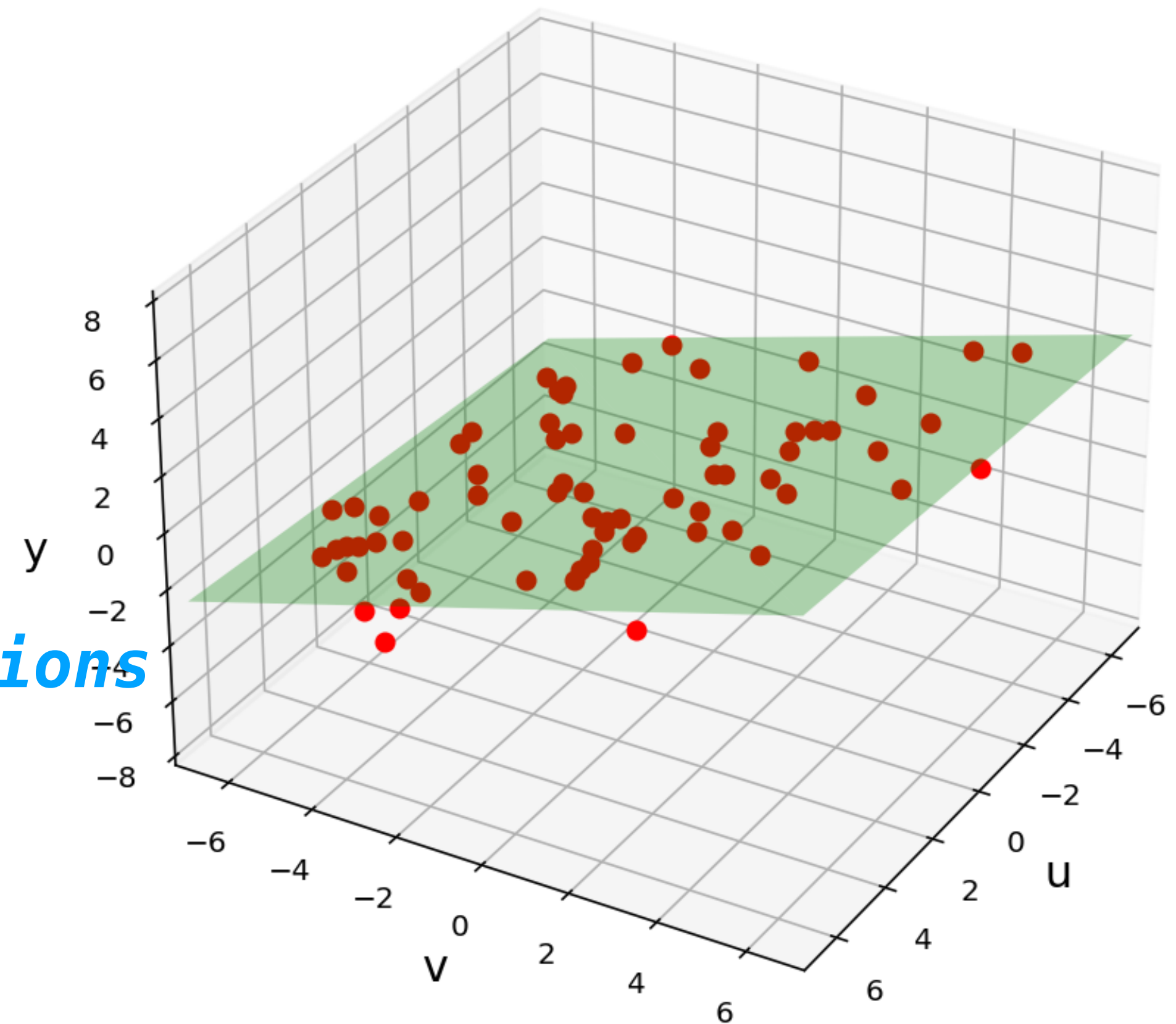
recall: planes are given by linear equations
which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$f(x, y)$ is a good approximation of the altitude.

Figure 23.2

Multiple Regression Fit to Data



Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$$\beta_0 + \beta_1 x_1 + \beta_2 y_1 = z_1$$

$$\beta_0 + \beta_1 x_2 + \beta_2 y_2 = z_2$$

\vdots

$$\beta_0 + \beta_1 x_k + \beta_2 y_k = z_k$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables $\beta_0, \beta_1, \beta_2$

Example: Terrain Data

This is still linear in the β 's

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$$\beta_0 + \beta_1 x_1 + \beta_2 y_1 = z_1$$

$$\beta_0 + \beta_1 x_2 + \beta_2 y_2 = z_2$$

\vdots

$$\beta_0 + \beta_1 x_k + \beta_2 y_k = z_k$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables $\beta_0, \beta_1, \beta_2$

Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & y_k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

Step 2: Rewrite the system as a matrix equation.

Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \dots, (x_k, y_k, z_k)\}$
where (x_i, y_i) is an longitude and latitude and z_i is an altitude.

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

which minimizes

$$\sum_{i=1}^k (f(x_i, y_i) - z_i)^2$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbf{z}$$

Step 3: Find the least squares solution of this system and use as the parameters of your model.

An Aside: Unique Least Squares

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & y_k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}$$

Question (Conceptual). *Why can almost always assume that the columns of this matrix are linearly independent?*

Answer

Answer

If the columns were linearly dependent, then one of our independent variables can be computed in terms of the others.

Answer

If the columns were linearly dependent, then one of our independent variables can be computed in terms of the others.

First off, this is very unlikely.

Answer

If the columns were linearly dependent, then one of our independent variables can be computed in terms of the others.

First off, this is very unlikely.

Second, this variable could be then be thought of as a *dependent* variable.

Answer


If the columns were linearly dependent, then one of our independent variables can be computed in terms of the others.

First off, this is very unlikely.

Second, this variable could be then be thought of as a *dependent* variable.

It wouldn't contribute anything when using the least squares method.

"Vectors" of Generalization


- 
1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

2. What if our data is not *exactly* linear.

e.g., polynomial regression

"Vectors" of Generalization

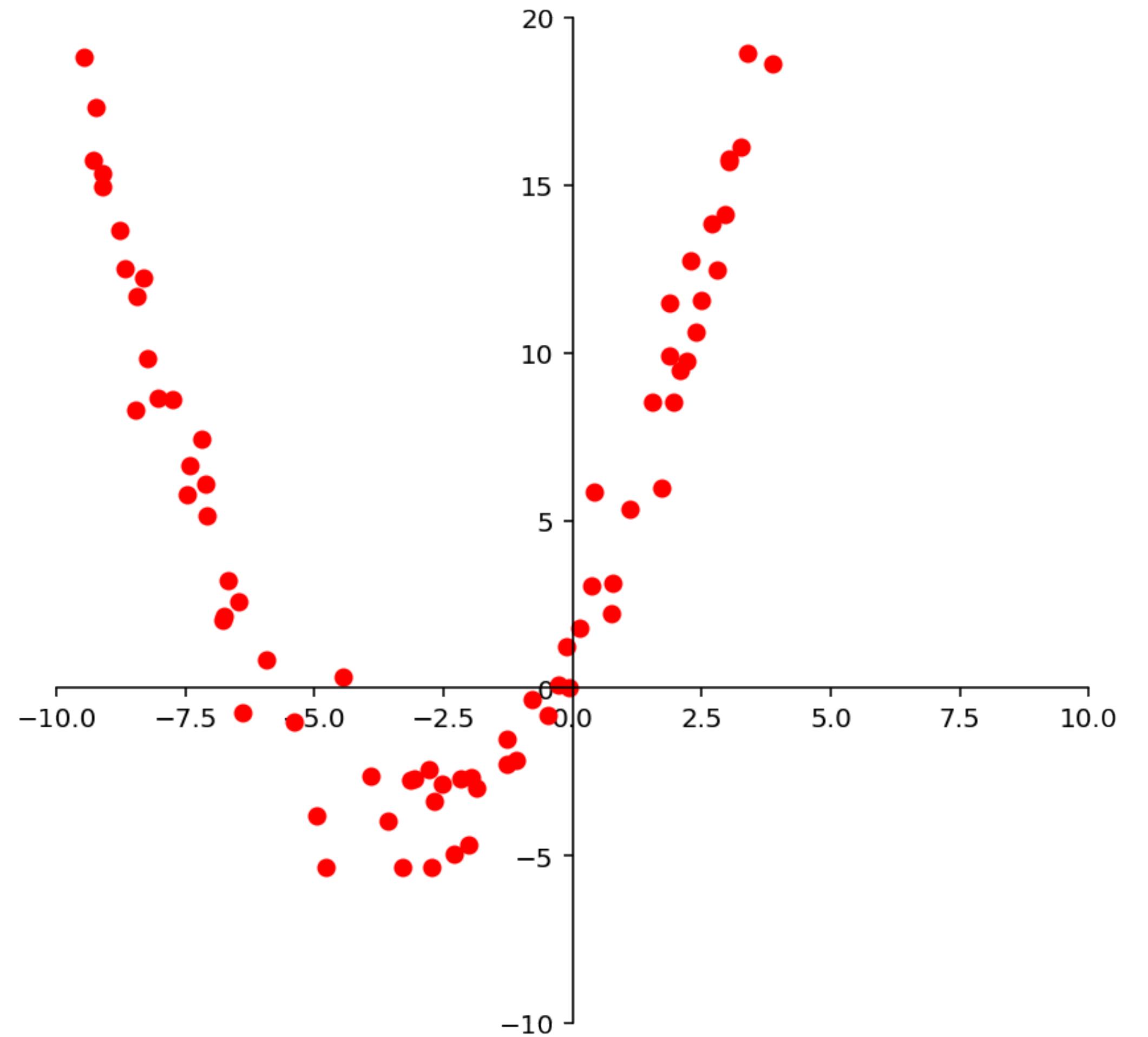
- 
1. What if we have *more than one* independent value?

multiple regression, (hyper)plane of best fit

2. What if our data is not *exactly* linear.

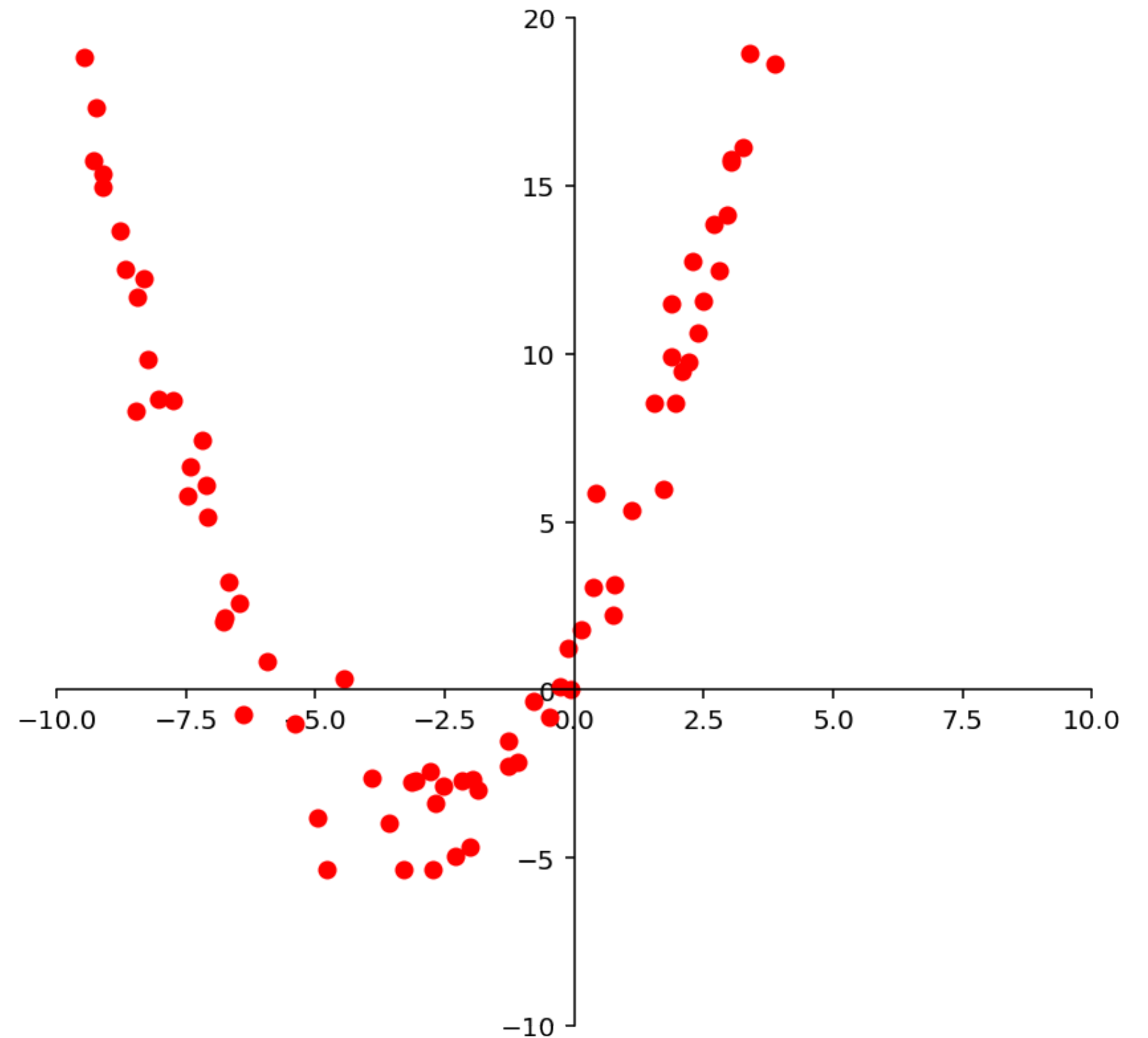
e.g., polynomial regression

Example: Best Fit Quadratic



Example: Best Fit Quadratic

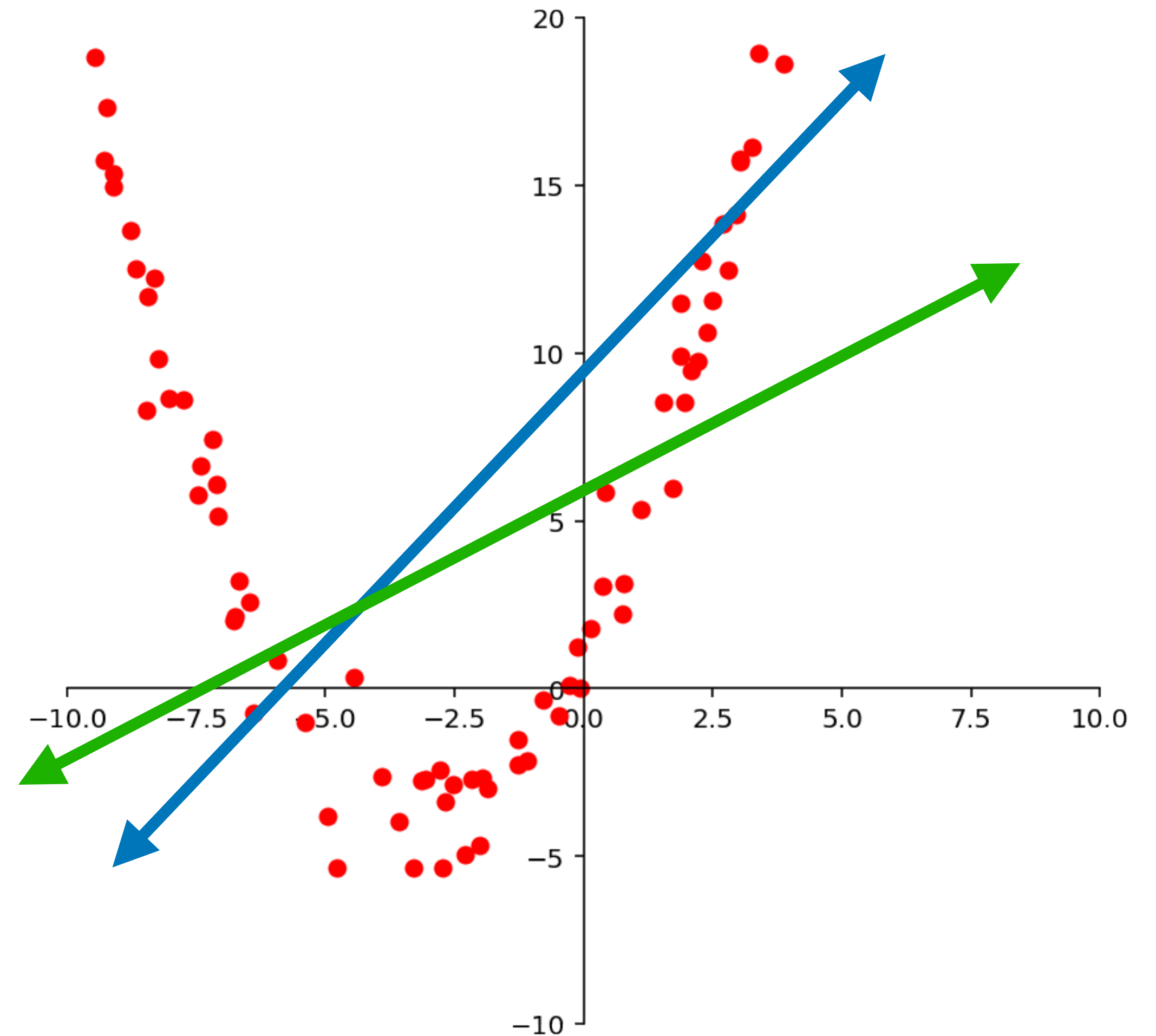
Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$



Example: Best Fit Quadratic

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

The issue: There is no good line to approximate this data.

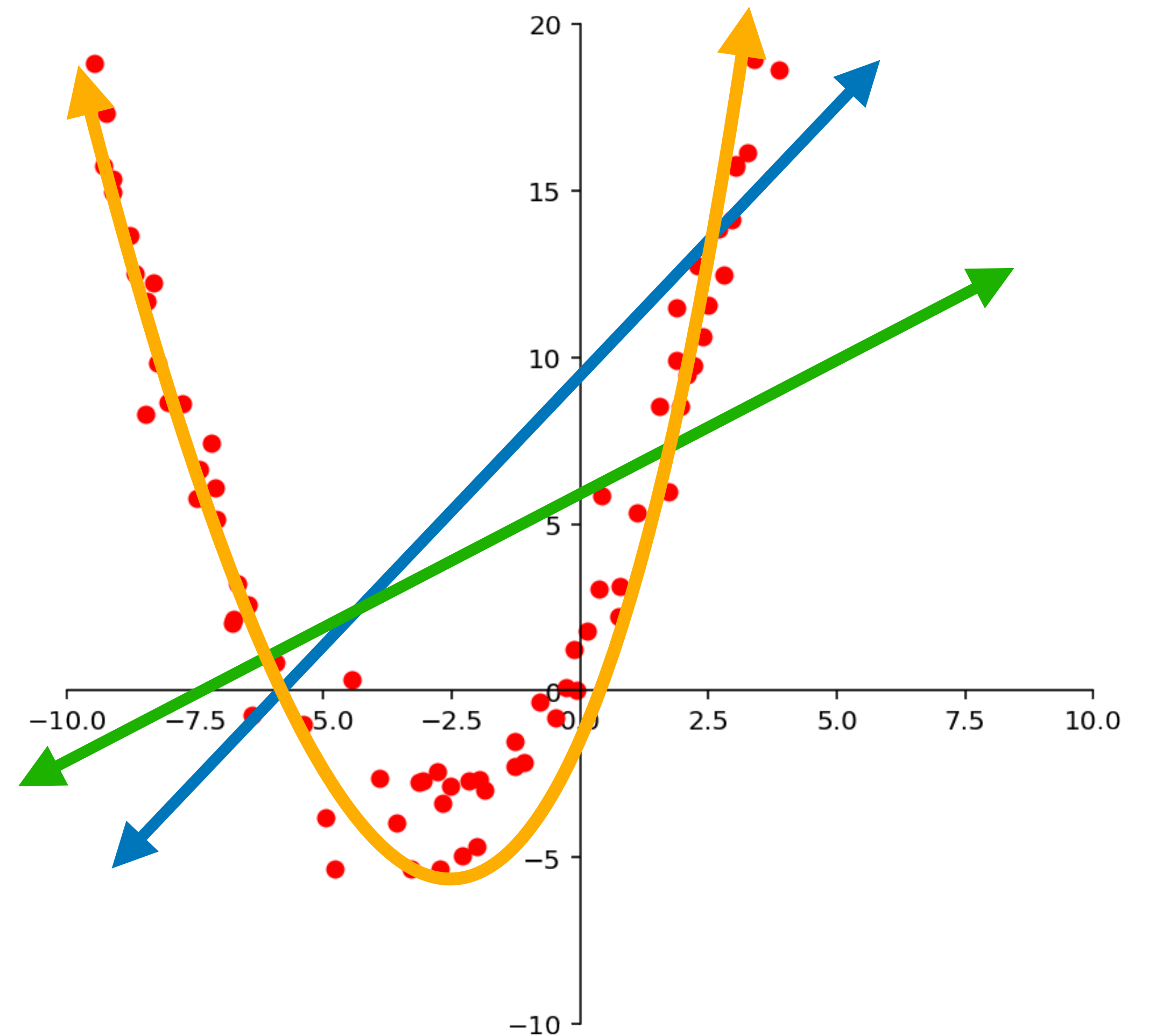


Example: Best Fit Quadratic

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

The issue: There is no good line to approximate this data.

What about a parabola?



Example: Best Fit Quadratic

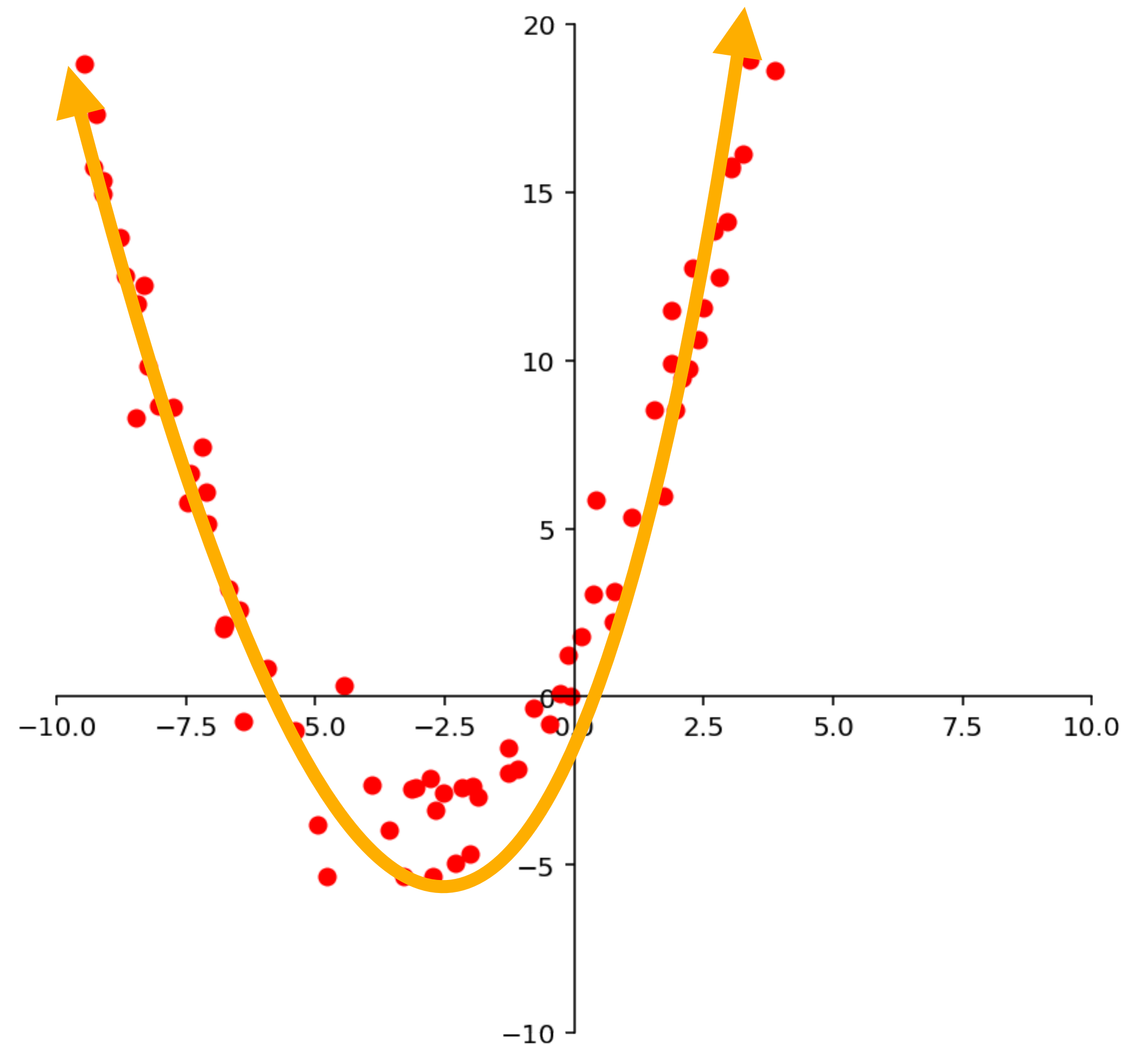
Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

minimizes

$$\sum_{i=1}^k (f(x_i) - y_i)^2$$



Example: Best Fit Quadratic

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

minimizes

$$\sum_{i=1}^k (f(x_i) - y_i)^2$$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 = y_1$$

$$\beta_0 + \beta_1 x_2 + \beta_2 x_2^2 = y_2$$

\vdots

$$\beta_0 + \beta_1 x_k + \beta_2 x_k^2 = y_k$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables $\beta_0, \beta_1, \beta_2$

Example: Best Fit Quadratic

This is still linear in the β 's

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

$$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2 = y_1$$

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$\beta_0 + \beta_1 x_2 + \beta_2 x_2^2 = y_2$$

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

\vdots

minimizes

$$\beta_0 + \beta_1 x_k + \beta_2 x_k^2 = y_k$$

$$\sum_{i=1}^k (f(x_i) - y_i)^2$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables $\beta_0, \beta_1, \beta_2$

Example: Best Fit Quadratic

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

minimizes

$$\sum_{i=1}^k (f(x_i) - y_i)^2$$

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & x_k^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}$$

Step 2: Rewrite the system as a matrix equation.

Example: Best Fit Quadratic

Dataset: $\{(x_1, y_1), \dots, (x_k, y_k)\}$

Problem: Find $\beta_0, \beta_1, \beta_2$ such that

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

minimizes

$$\sum_{i=1}^k (f(x_i) - y_i)^2$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

Step 3: Find the least squares solution of this system and use as the parameters of your model.

Question

Find the parabola of best fit for the dataset

$$\{(0,3), (1,1), (-1,1)\}$$

Hint. Plot it

Answer

$$\{(0,3), (1,1), (-1,1)\}$$

The Takeaway

We can use non-linear modeling functions as long as they are linear in the parameters.

Linear in Parameters

non examples:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto \beta_1 \beta_2 x_1$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \mapsto e^{\beta_1 x_1}$$

Definition. A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is **linear in the parameters** β_1, \dots, β_k if it can be written as

$$f(\mathbf{x}) = \beta_1 \phi_1(\mathbf{x}) + \beta_2 \phi_2(\mathbf{x}) + \dots + \beta_k \phi_k(\mathbf{x})$$

not necessarily linear

for functions $\phi_1, \dots, \phi_k: \mathbb{R}^n \rightarrow \mathbb{R}$

Example: $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \mapsto \beta_2 e^{x_1} + \beta_3 \log(x_3^{x_1})$

An Aside: Statistical Models (Another view)

$$\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$$

An Aside: Statistical Models (Another view)

$$\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$$

So far, we have been considering *inconsistent* systems of the form $\mathbf{y} = X\vec{\beta}$.

An Aside: Statistical Models (Another view)

$$\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$$

So far, we have been considering *inconsistent* systems of the form $\mathbf{y} = X\vec{\beta}$.

It is also common to *make the system consistent* by adding error terms (the ϵ 's).

An Aside: Statistical Models (Another view)

$$\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$$

So far, we have been considering *inconsistent* systems of the form $\mathbf{y} = X\vec{\beta}$.

It is also common to *make the system consistent* by adding error terms (the ϵ 's).

(We won't use this view, this is mostly for your personal betterment, and because the notes use this notation occasionally.)

An Aside: Statistical Models (Another view)

design matrix

$$\mathbf{y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$$

So far, we have been considering *inconsistent* systems of the form $\mathbf{y} = \mathbf{X}\vec{\beta}$.

It is also common to *make the system consistent* by adding error terms (the ϵ 's).

(We won't use this view, this is mostly for your personal betterment, and because the notes use this notation occasionally.)

We can build design matrices for function which are linear in their parameters.

General Linear Regression

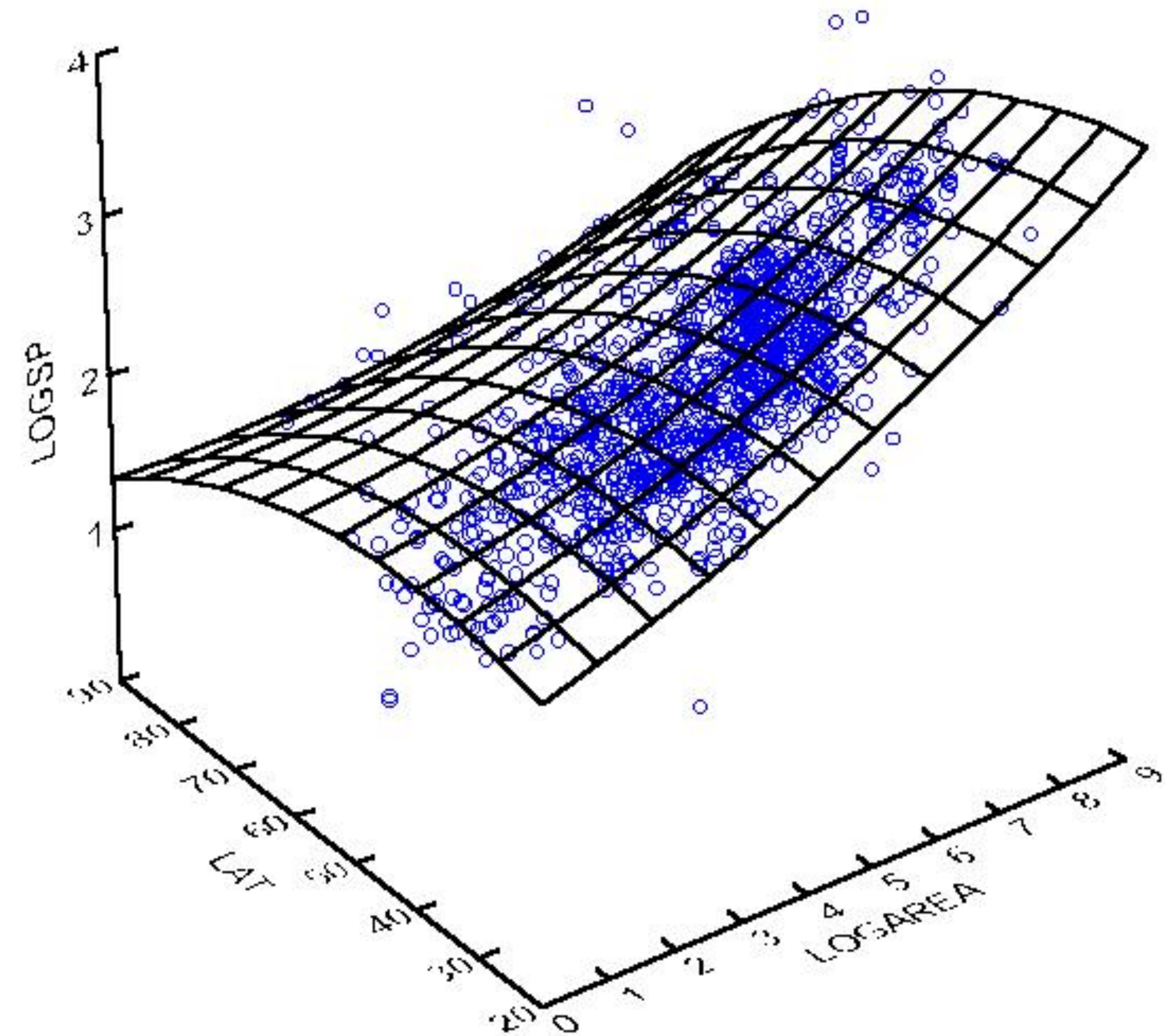
dataset: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1, \dots, \beta_k} : \mathbb{R}^n \rightarrow \mathbb{R}$$

which is *linear in the parameters* β_1, \dots, β_k , find values for β_1, \dots, β_k which minimize

$$\sum_{i=1}^k (f_{\beta_1, \dots, \beta_k}(\mathbf{x}_i) - y_i)^2$$



General Linear Regression

dataset: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1, \dots, \beta_k} : \mathbb{R}^n \rightarrow \mathbb{R}$$

which is *linear in the parameters* β_1, \dots, β_k , find values for β_1, \dots, β_k which minimize

$$\sum_{i=1}^k (f_{\beta_1, \dots, \beta_k}(\mathbf{x}_i) - y_i)^2$$

$$\beta_1 \phi_1(\mathbf{x}_1) + \dots + \beta_k \phi_k(\mathbf{x}_1) = y_1$$

$$\beta_1 \phi_1(\mathbf{x}_2) + \dots + \beta_k \phi_k(\mathbf{x}_2) = y_2$$

\vdots

$$\beta_1 \phi_1(\mathbf{x}_2) + \dots + \beta_k \phi_k(\mathbf{x}_2) = y_2$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables β_1, \dots, β_k

General Linear Regression

This is still linear in the β 's

dataset: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1, \dots, \beta_k} : \mathbb{R}^n \rightarrow \mathbb{R}$$

which is *linear in the parameters* β_1, \dots, β_k , find values for β_1, \dots, β_k which minimize

$$\sum_{i=1}^k (f_{\beta_1, \dots, \beta_k}(\mathbf{x}_i) - y_i)^2$$

$$\beta_1 \phi_1(\mathbf{x}_1) + \dots + \beta_k \phi_k(\mathbf{x}_1) = y_1$$

$$\beta_1 \phi_1(\mathbf{x}_2) + \dots + \beta_k \phi_k(\mathbf{x}_2) = y_2$$

\vdots

$$\beta_1 \phi_1(\mathbf{x}_2) + \dots + \beta_k \phi_k(\mathbf{x}_2) = y_2$$

Step 1: Set up an (almost assuredly inconsistent) system of linear equations in terms of the variables β_1, \dots, β_k

General Linear Regression

dataset: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1, \dots, \beta_k} : \mathbb{R}^n \rightarrow \mathbb{R}$$

which is *linear in the parameters* β_1, \dots, β_k , find values for β_1, \dots, β_k which minimize

$$\sum_{i=1}^k (f_{\beta_1, \dots, \beta_k}(\mathbf{x}_i) - y_i)^2$$

design matrix

$$\begin{matrix} & \text{design matrix} \\ & X \\ \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_k(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots & \phi_k(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & \dots & \phi_k(\mathbf{x}_m) \end{bmatrix} & \begin{bmatrix} \vec{\beta} \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} & = & \begin{bmatrix} \mathbf{y} \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \end{matrix}$$

Step 2: Rewrite the system as a matrix equation.

General Linear Regression

dataset: $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1, \dots, \beta_k} : \mathbb{R}^n \rightarrow \mathbb{R}$$

which is *linear in the parameters* β_1, \dots, β_k , find values for β_1, \dots, β_k which minimize

$$\sum_{i=1}^k (f_{\beta_1, \dots, \beta_k}(\mathbf{x}_i) - y_i)^2$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

Step 3: Find the least squares solution of this system and use as the parameters of your model.

How To: Design Matrices

How To: Design Matrices

Problem. Find the design matrix for least squares regression with the function f in terms of the parameters $\beta_1, \beta_2, \dots, \beta_k$ given the dataset $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$.

How To: Design Matrices

Problem. Find the design matrix for least squares regression with the function f in terms of the parameters $\beta_1, \beta_2, \dots, \beta_k$ given the dataset $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$.

Solution. First write $f(\mathbf{x})$ as $\beta_1\phi_1(\mathbf{x}) + \dots + \beta_k\phi_k(\mathbf{x})$ where ϕ_1, \dots, ϕ_k are potentially non-linear functions. Then build the matrix:

$$\begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_k(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots & \phi_k(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_m) & \phi_2(\mathbf{x}_m) & \dots & \phi_k(\mathbf{x}_m) \end{bmatrix}$$

Question

Find the design matrix for the least squares regression with the function

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \mapsto \beta_1 \cos(x_1) + \beta_2 e^{-x_1 x_2} - \beta_1 x_3 + \beta_3$$

for the dataset

$$\mathbf{x}_1 = (0, 0, 0) \quad y_1 = 5$$

$$\mathbf{x}_2 = (\pi, 3, 1) \quad y_2 = 3$$

Answer: $\begin{bmatrix} 1 & 1 & 1 \\ -2 & e^{-3\pi} & 1 \end{bmatrix}$

Practical Considerations

Practical Considerations

Many functions require large design matrices, e.g. multivariate polynomials have *a lot* of possible terms.

Practical Considerations

Many functions require large design matrices, e.g. multivariate polynomials have *a lot* of possible terms.

We haven't actually talked about *which* modeling functions to use.

Practical Considerations

Many functions require large design matrices, e.g. multivariate polynomials have *a lot* of possible terms.

We haven't actually talked about *which* modeling functions to use.

Again, is least-squares error really what we want? What if we want to minimize something else?

Practical Considerations

Many functions require large design matrices, e.g. multivariate polynomials have *a lot* of possible terms.

We haven't actually talked about *which* modeling functions to use.

Again, is least-squares error really what we want? What if we want to minimize something else?

Concerns for another class.

One Last Thing

Please through the last section of the notes
"Multiple Regression in Practice"

It will be useful for Homework 12.