Linear Models

Geometric Algorithms
Lecture 24

CAS CS 132

Practice Problem

$$\begin{array}{cc} & \begin{array}{ccc} a_1 & a_2 & a_3 \end{array} \\ A = & \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \end{array} \qquad \mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}$$

Find the orthogonal *projection of* $\mathbf{b}$ *onto* $\mathrm{Col}(A)$.

hint. $a_3 = a_2 - a_1$

Answer

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & -1 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -4 \end{bmatrix}$$

⓪ Find basis

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

$$B\hat{x} = \hat{b}$$

① Find LS solution

$$B^T B \hat{x} = B^T \vec{b} \quad \text{(normal)}$$

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

② $\hat{b} = B\hat{x}$

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

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

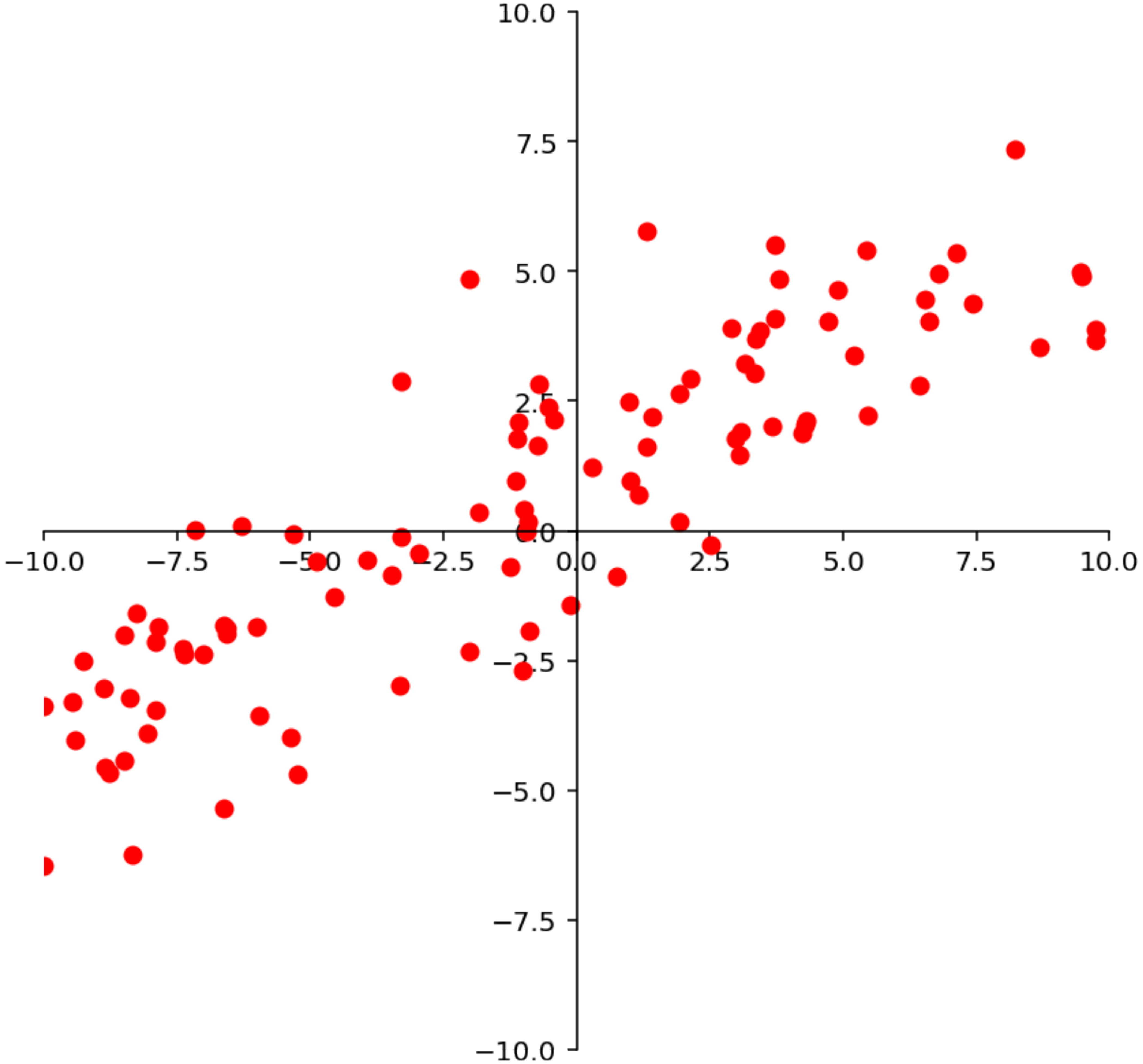
model fitting

model parameters

design matrices

Warm-up: Line of Best Fit

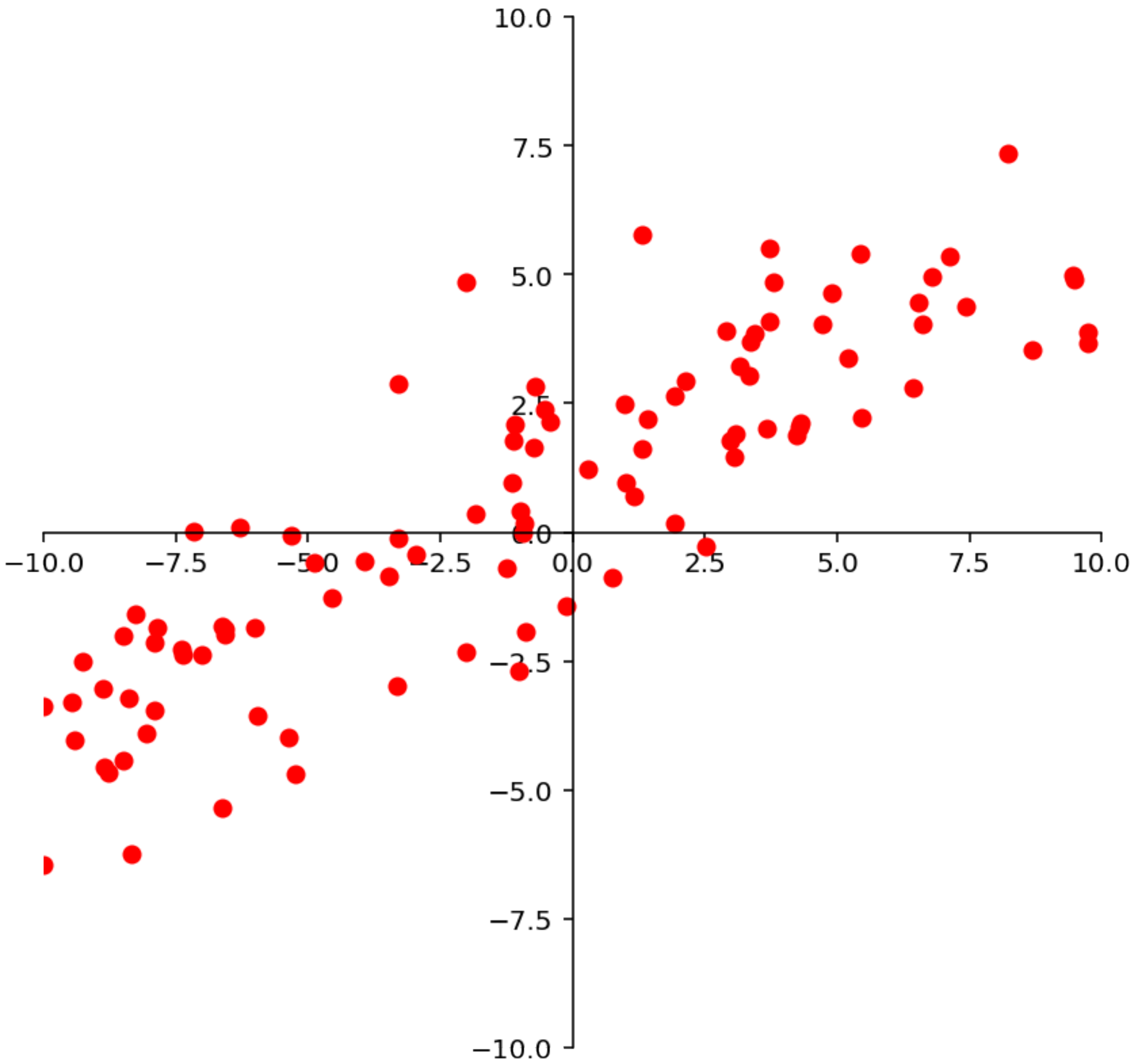
The Setup



The Setup

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

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

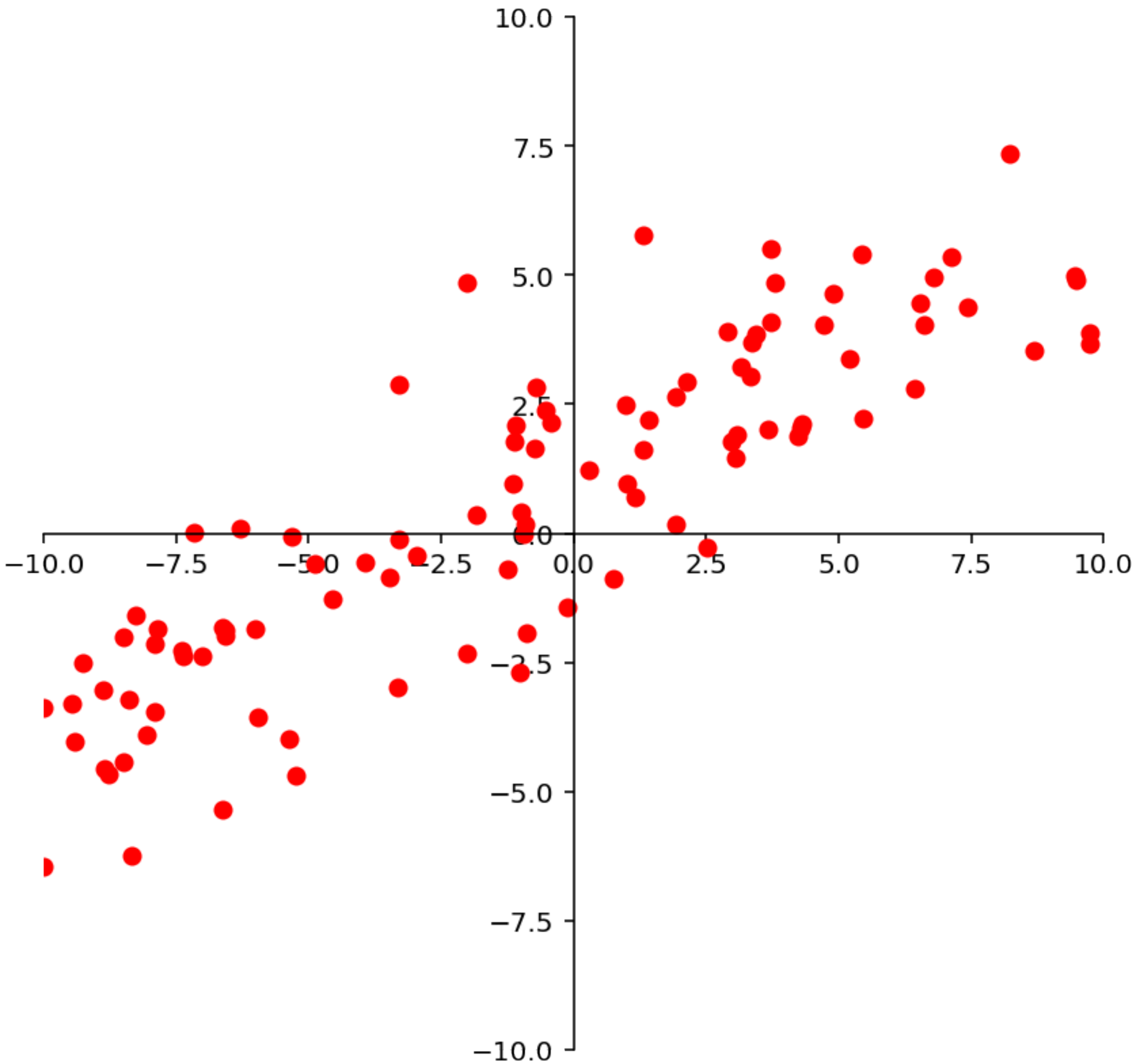


The Setup

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

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

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



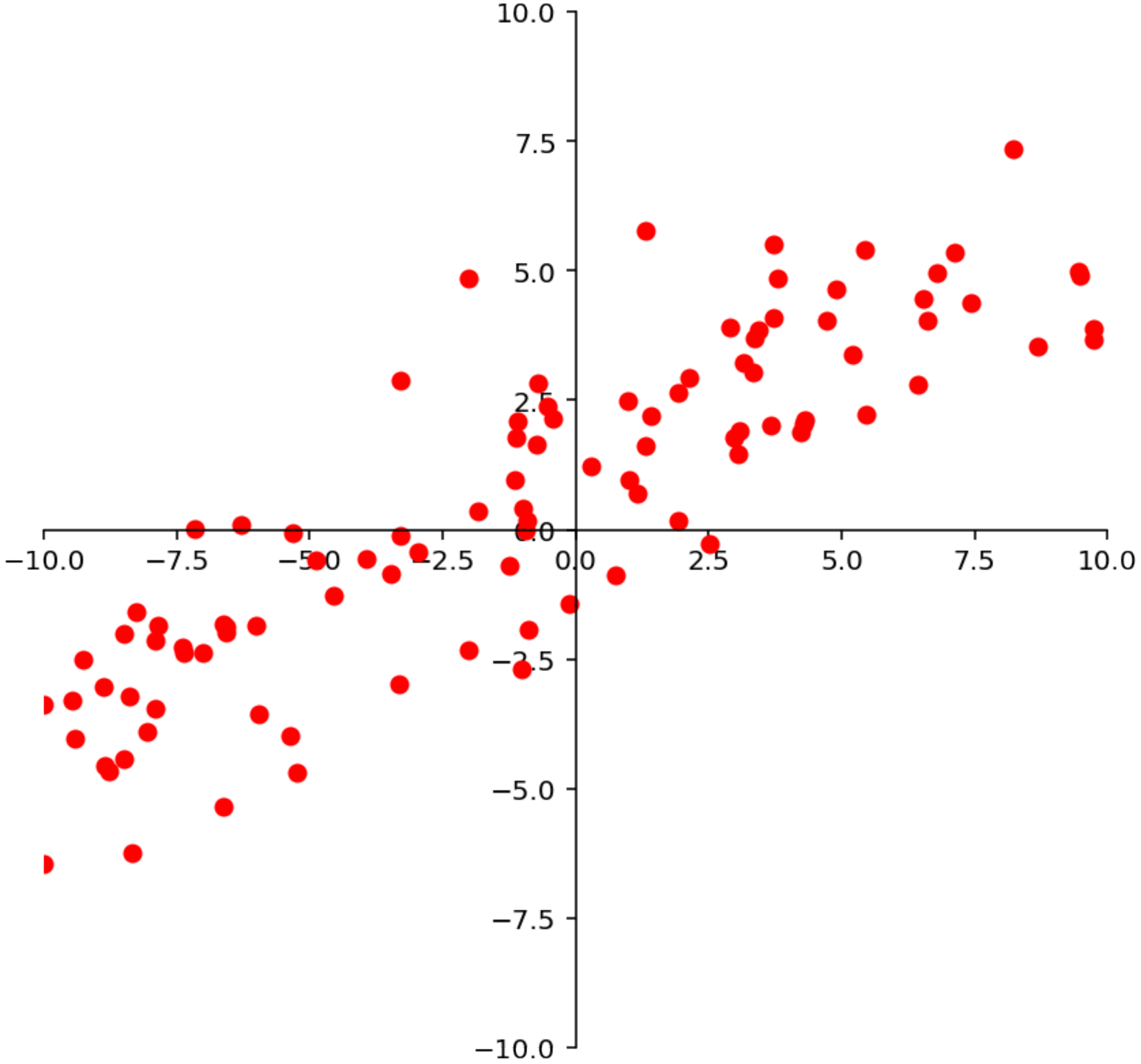
The Setup

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

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

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

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



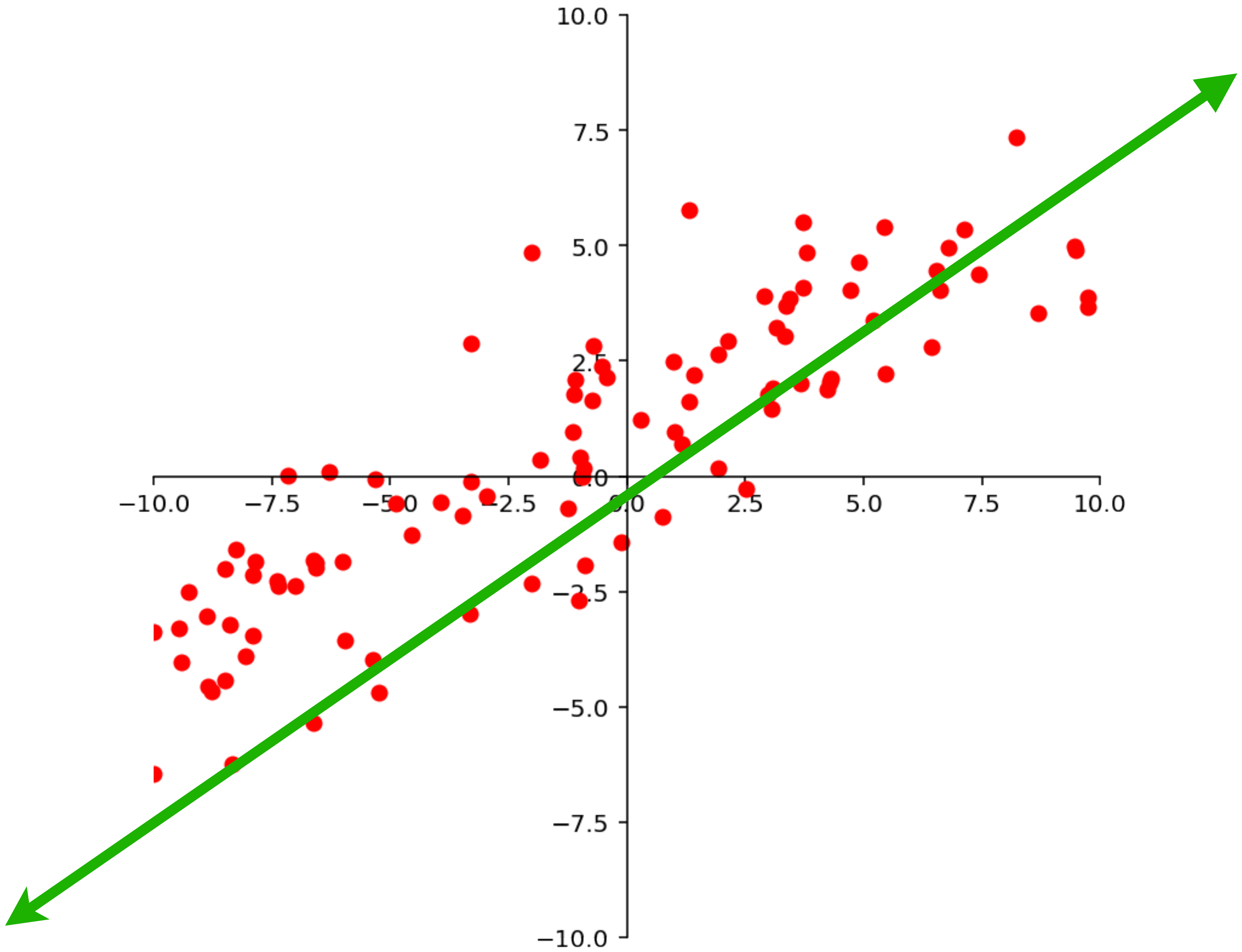
The Setup

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.



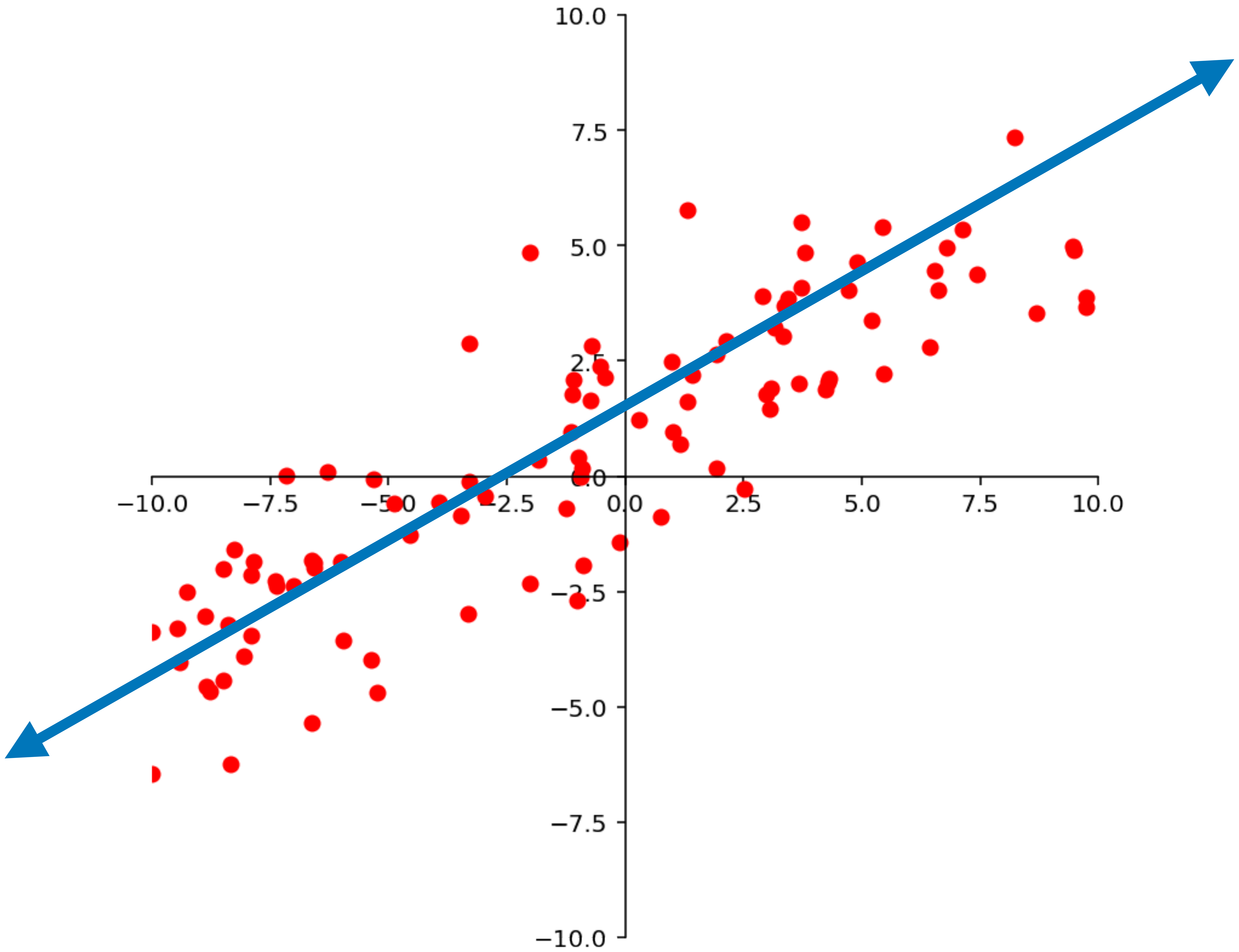
The Setup

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

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

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

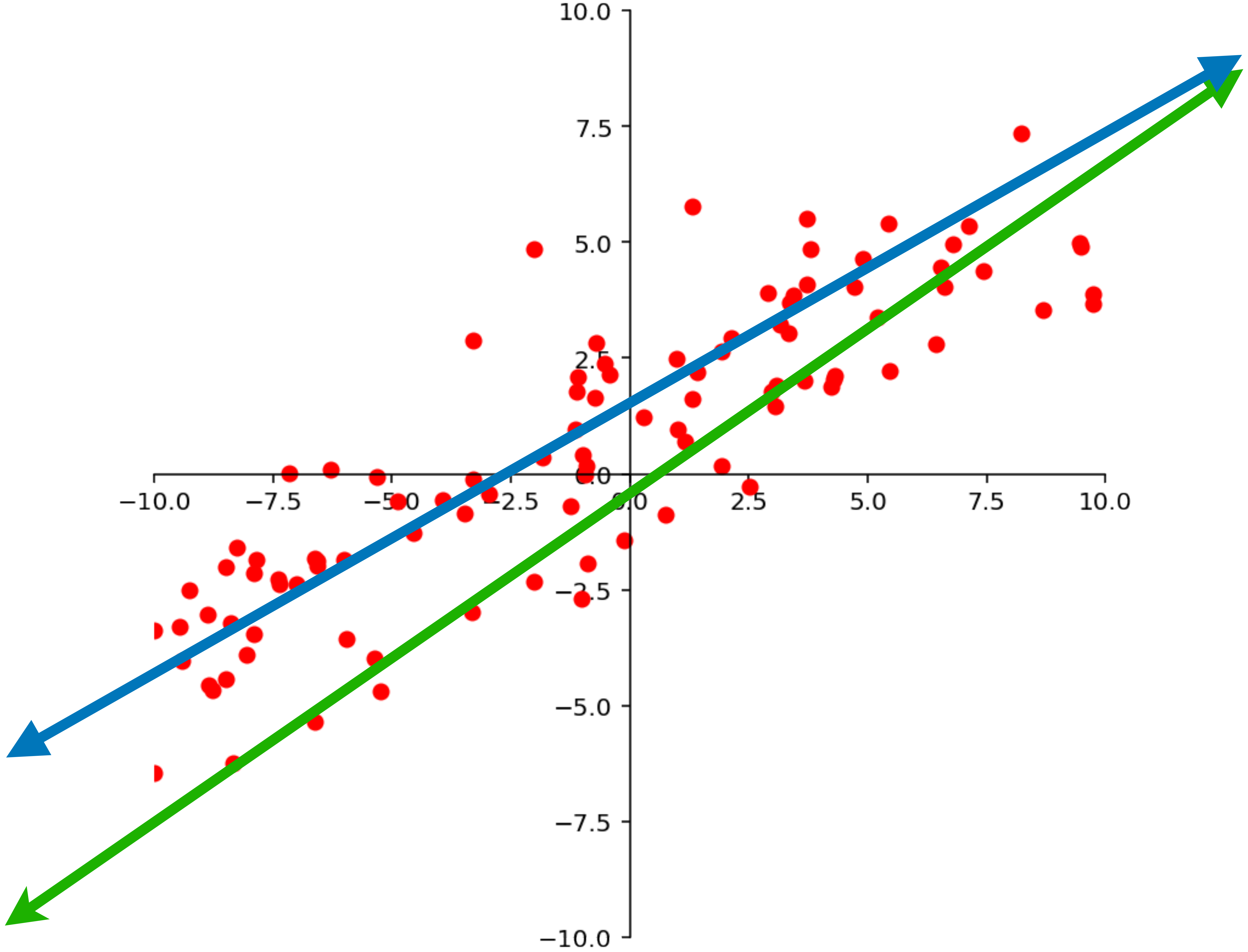
You notice they *kind of* trend as a line.



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 $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, find the line

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

which <u>minimizes</u>

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

Least Squares Simple Linear Regression

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

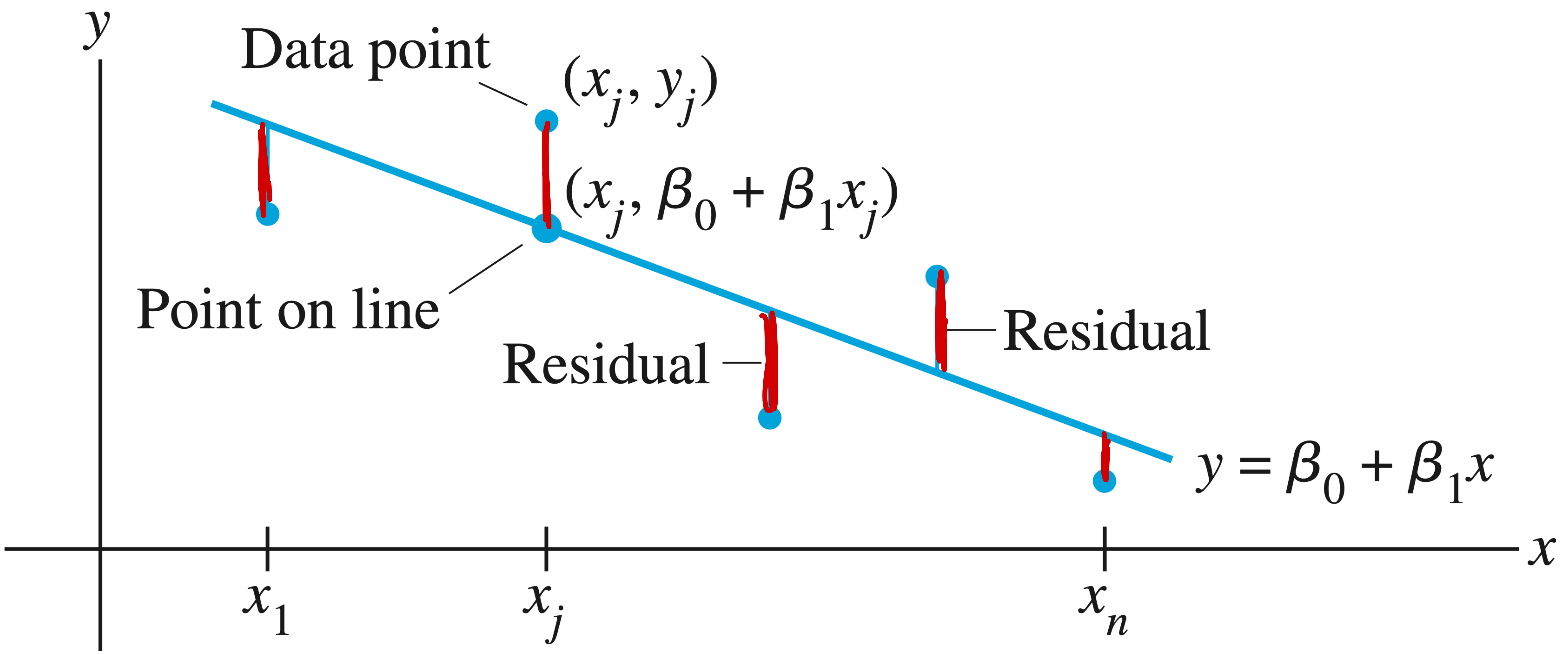
$$f(x) = \beta_0 + \beta_1 x$$

which <u>minimizes</u>

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

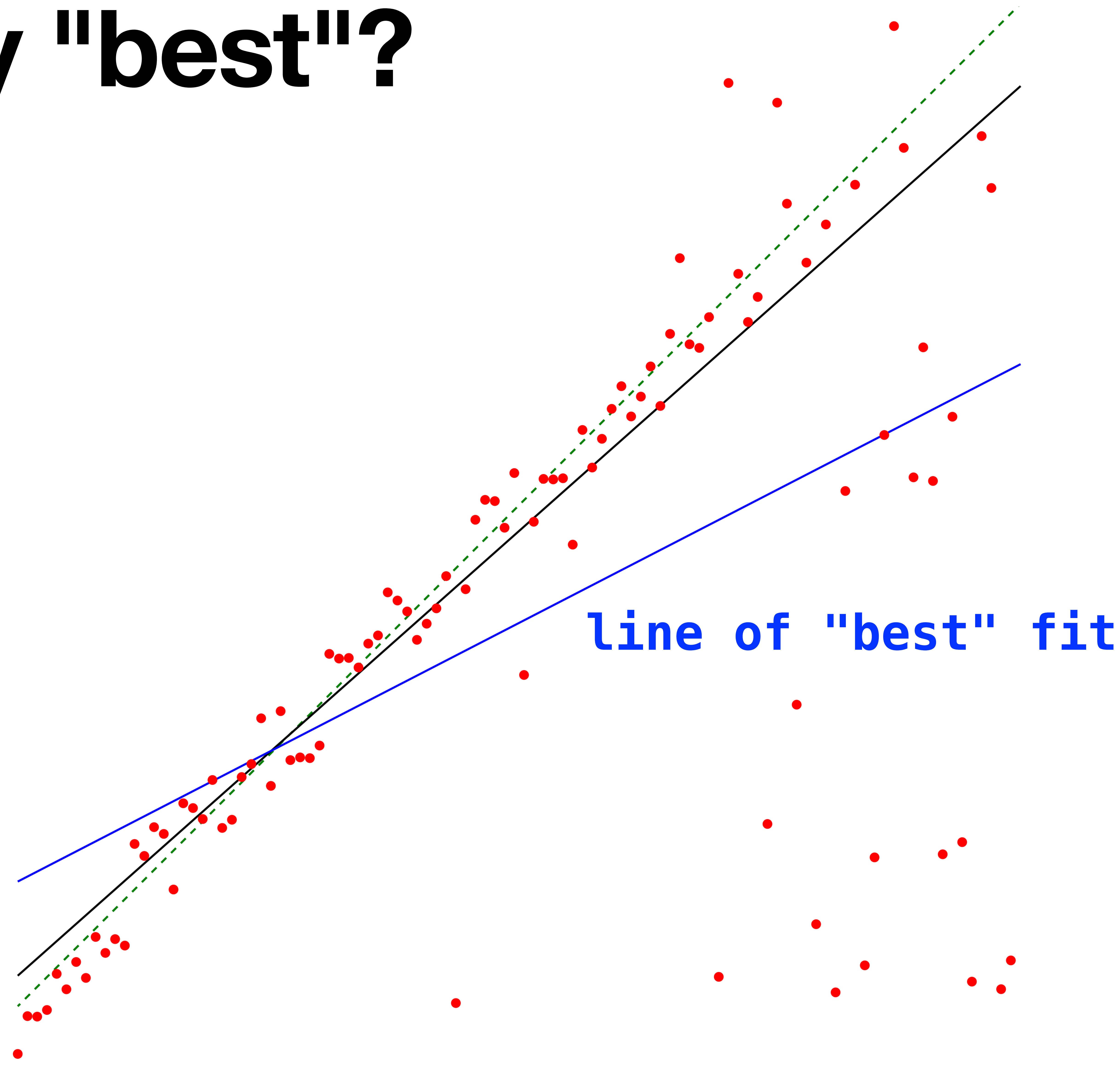
The "best" line minimizes the *sum of squares of differences.*

The Picture



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

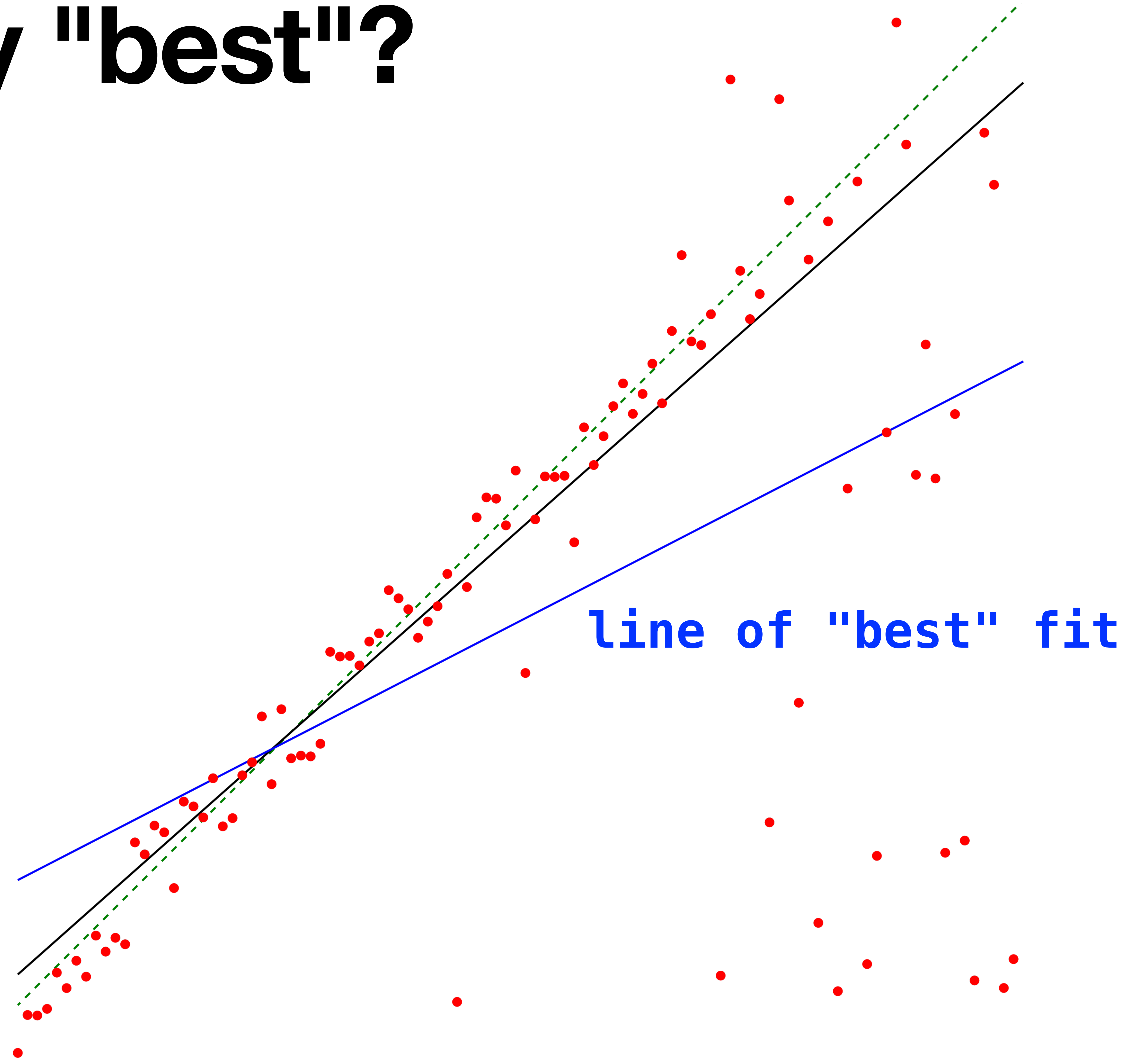
An Aside: Is this really "best"?



https://commons.wikimedia.org/wiki/File:Thiel-Sen_estimator.svg

An Aside: Is this really "best"?

Who's to say...

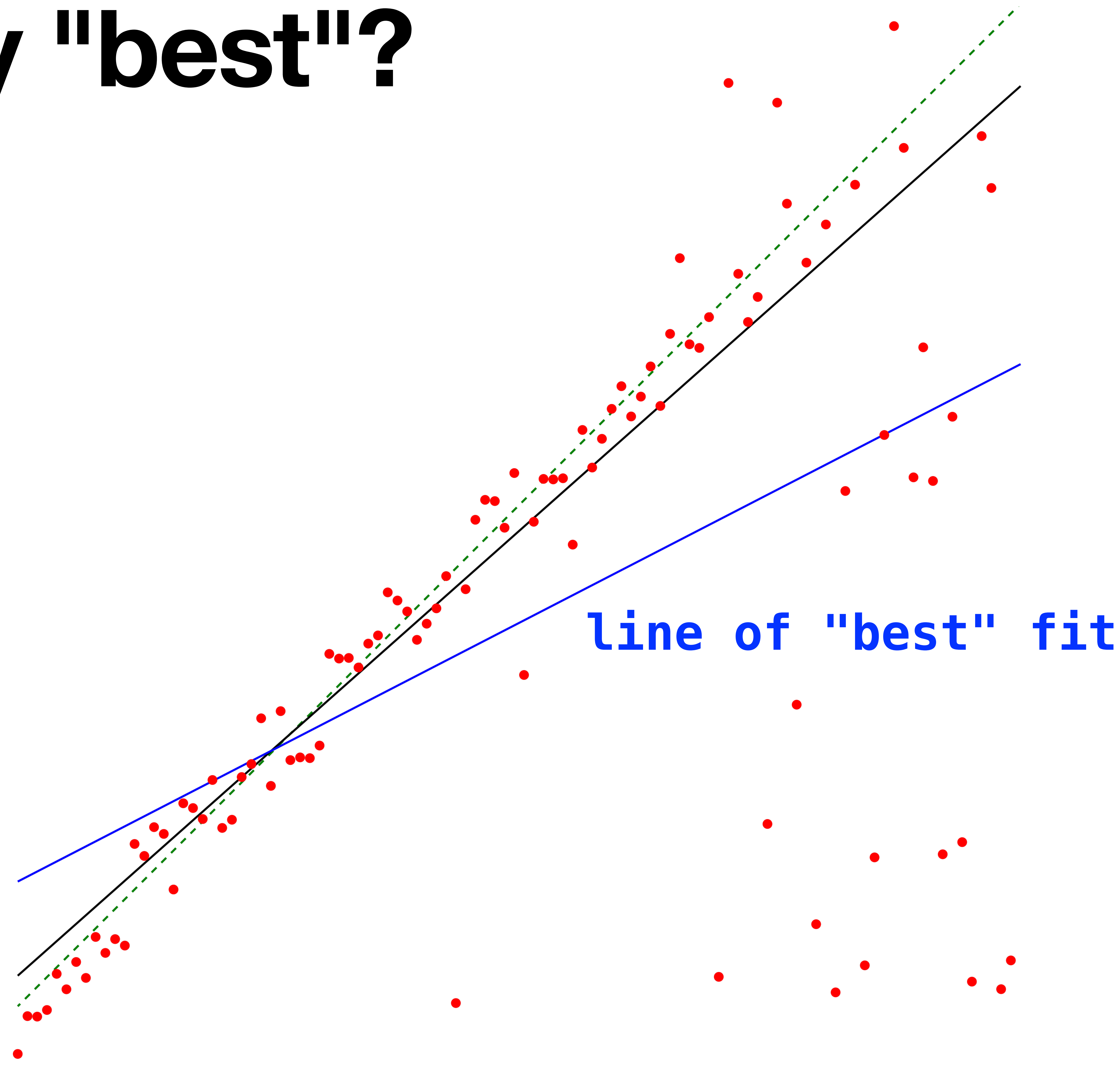


https://commons.wikimedia.org/wiki/File:Thiel-Sen_estimator.svg

An Aside: Is this really "best"?

Who's to say...

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



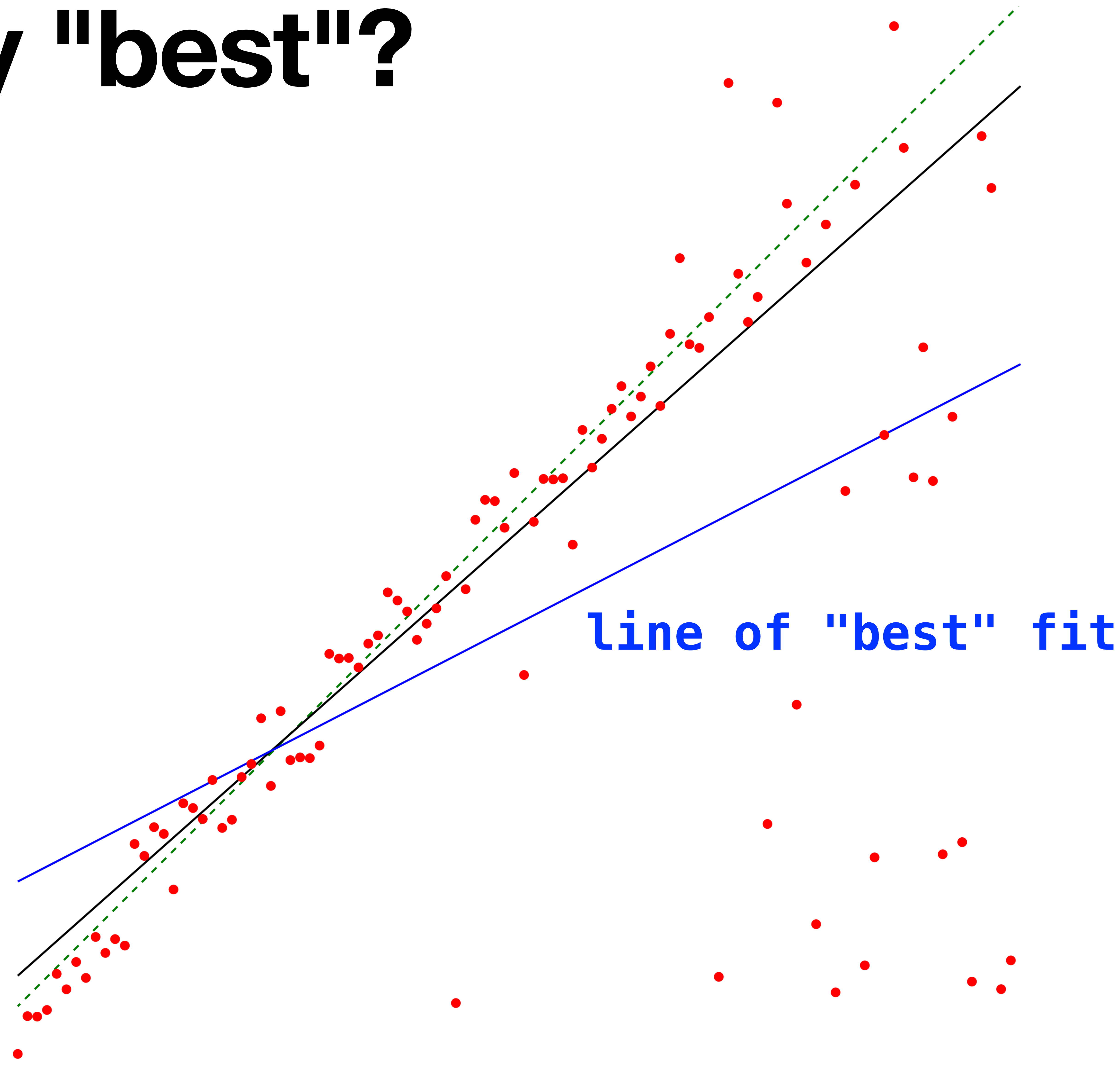
https://commons.wikimedia.org/wiki/File:Thiel-Sen_estimator.svg

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.



https://commons.wikimedia.org/wiki/File:Thiel-Sen_estimator.svg

Terminology: Datasets

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

Terminology: Datasets

$$\underbrace{\{(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_n, y_n)\}}_{\texttt{dataset}}$$

Terminology: Datasets

data point

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

dataset

Terminology: Datasets

data point

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

dataset

$$(x_i, y_i)$$

Terminology: Datasets

data point

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

dataset

independent variable
(feature)

$$(x_i, y_i)$$

Terminology: Datasets

data point

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

dataset

independent variable
(feature)

$$(x_i, y_i)$$

dependent variable
(label)

Terminology: Models

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

Terminology: Models

$$\underbrace{f(x) = \beta_0 + \beta_1 x}_{\texttt{model}}$$

Terminology: Models

model parameters/
regression coefficients

$$\underbrace{f(x) = \beta_0 + \beta_1 x}_{\text{model}}$$

Terminology: Least-Squares Error

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

Terminology: Least-Squares Error

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

Terminology: Least-Squares Error

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

observation

residual

Terminology: Least-Squares Error

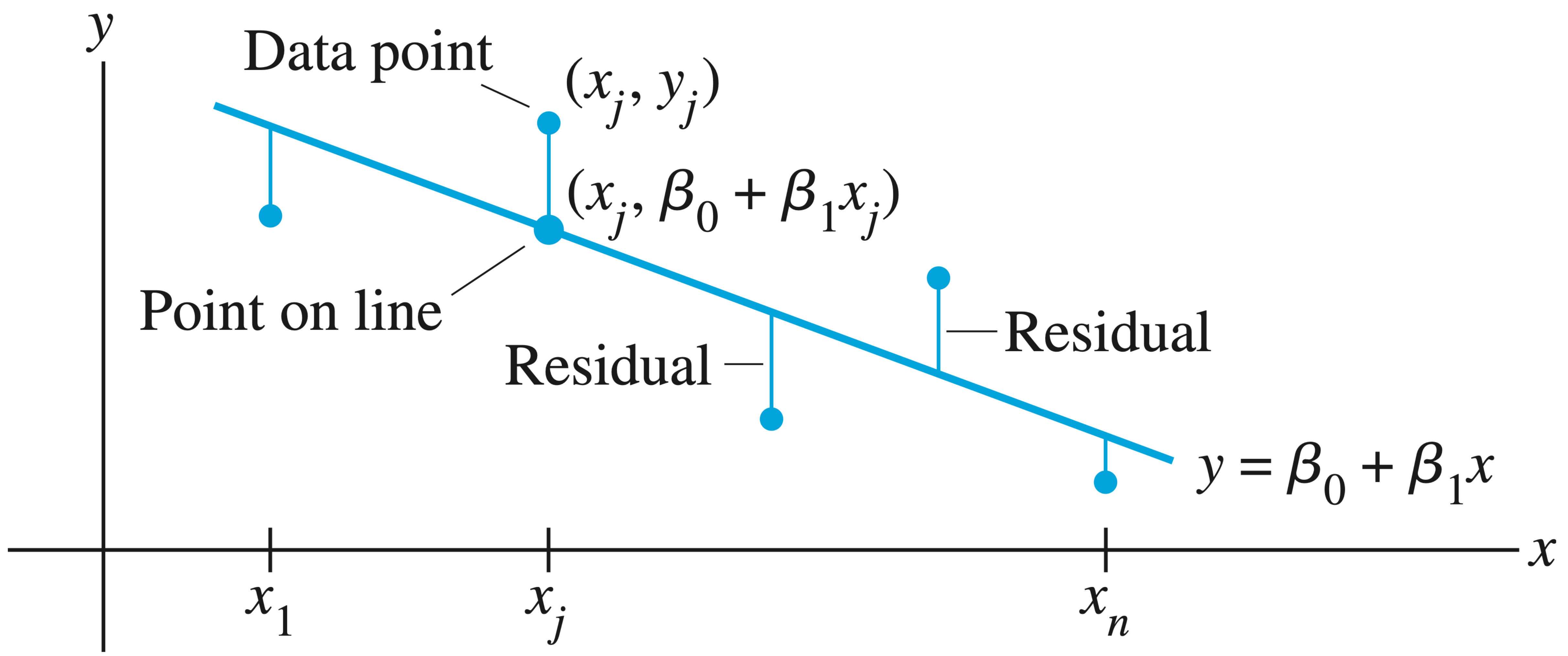
$$\sum_{i=1}^{n} \left(\underbrace{y_i}_{\text{observation}} - \underbrace{f(x_i)}_{\text{prediction}} \right)^2$$

observation

residual

prediction

Terminology



$$\underbrace{\{(x_1, y_1), \ldots, \overbrace{(x_i, y_i)}^{\text{data point}}, \ldots, (x_n, y_n)\}}_{\text{dataset}}$$

$$\underbrace{f(x) = \overbrace{\beta_0 + \beta_1}^{\text{model parameters/ regression coefficients}} x}_{\text{model}}$$

$$(\underbrace{x_i}_{\text{independent variable}}, \underbrace{y_i}_{\text{dependent variable (label)}})$$

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

How to: Finding the Least Squares Line

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

How to: Finding the Least Squares Line

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

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

How to: Finding the Least Squares Line

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

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

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

How to: Finding the Least Squares Line

Don't memorize these.

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

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

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

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

`minimize for least-squares line` `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 <u>design</u> a matrix where finding a least squares solution gives us a least squares line?

A Least Squares Problem

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

A Least Squares Problem

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

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

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

$\boxed{\beta_0 + \beta_1 5.2 = 7.9}$

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.**

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

This is a linear system in the variables β_0 and β_1

A Least Squares Problem

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

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

$$\underset{X}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \overset{\vec{\beta}}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} = \overset{\mathbf{y}}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}$$

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

A Least Squares Problem

$$\begin{array}{ccc} \color{red}{X} & \color{red}{\vec{\beta}} & & \color{red}{\mathbf{y}} \\ \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} & = & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}$$

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

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

A Least Squares Problem

$$\begin{array}{ccc} \mathbf{X} & \vec{\beta} & & \mathbf{y} \\ \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} & \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} & = & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \end{array}$$

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

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.

Recall: The Normal Equations

Recall: The Normal Equations

Theorem. The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is the same as the set of solutions to

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

Recall: The Normal Equations

Theorem. The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is the same as the set of solutions to

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

In particular, this set of solutions is nonempty

Recall: The Normal Equations

Theorem. The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is the same as the set of solutions to

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

In particular, this set of solutions is nonempty

(We just showed that if $\hat{\mathbf{x}}$ is a least squares solution then $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$)

Recall: Unique Least Squares Solutions

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

If A has linearly independent columns, then its unique least squares solution is defined as above.

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

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:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad X^T = \begin{bmatrix} \vec{1}^T \\ \vec{x}^T \end{bmatrix} \qquad X^T X = \begin{bmatrix} 1^T 1 & 1^T x \\ 1^T x & x^T x \end{bmatrix}$$

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

$$X^T X = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = (X^T X)^{-1} X^T \vec{y}$$

$$(X^T X)^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \begin{bmatrix} \cdots \end{bmatrix}$$

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 $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

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 $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

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

Question

Find the line of best fit for the dataset

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

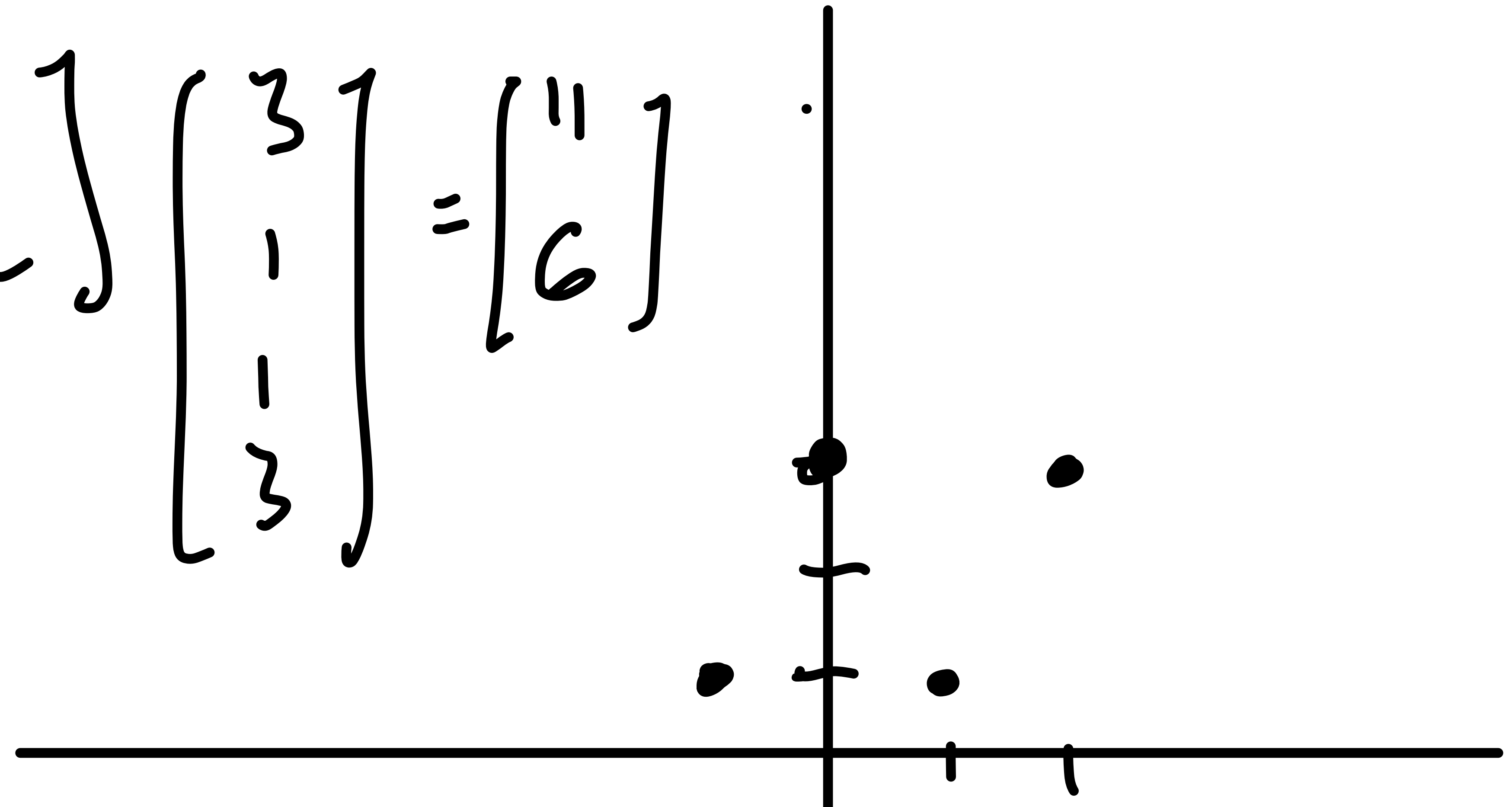
by using the the least-squares method.

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

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

$$\overset{X}{\begin{bmatrix}1&0\\1&1\\1&-1\\1&2\end{bmatrix}}\overset{\vec{\beta}}{\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}} = \overset{\vec{y}}{\begin{bmatrix}3\\1\\1\\3\end{bmatrix}}$$

$$X^T y = \begin{bmatrix}1&1&1&1\\0&1&-1&2\end{bmatrix}\begin{bmatrix}3\\1\\1\\3\end{bmatrix} = \begin{bmatrix}11\\6\end{bmatrix}$$



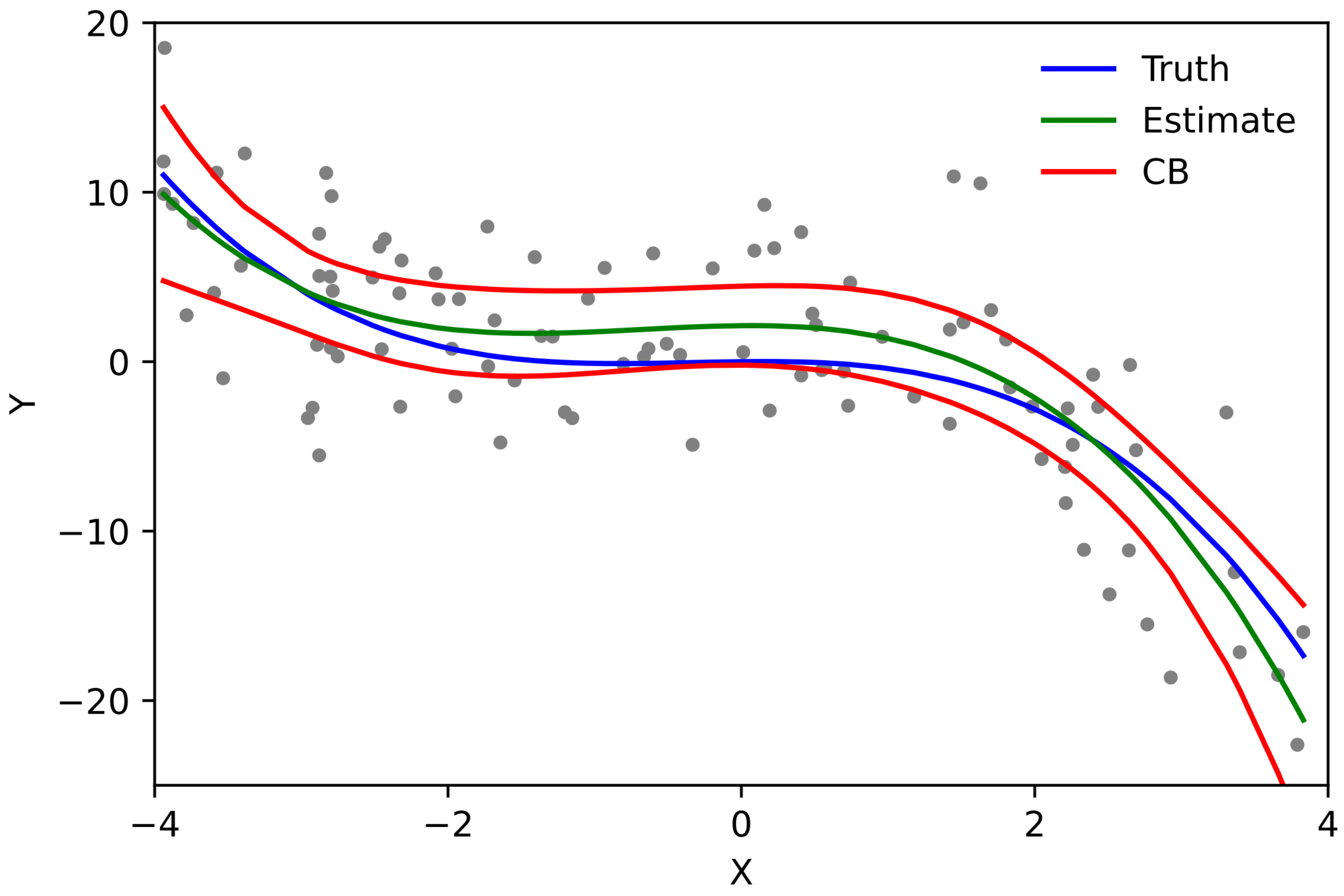
$$\vec{\beta} = \frac{1}{10}\begin{bmatrix}3&-1\\-1&2\end{bmatrix}\begin{bmatrix}11\\6\end{bmatrix} = \frac{1}{10}\begin{bmatrix}33-6\\-11+12\end{bmatrix} = \begin{bmatrix}27\\1\end{bmatrix}\frac{1}{10}$$

$$\vec{\beta} = (X^T X)^{-1} X^T y \qquad \begin{bmatrix}\beta_0\\\beta_1\end{bmatrix} = \begin{bmatrix}2.7\\0.1\end{bmatrix}$$

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

$$(X^T X)^{-1} = \frac{1}{24-4}\begin{bmatrix}6&-2\\-2&4\end{bmatrix} = \frac{1}{10}\begin{bmatrix}3&-1\\-1&2\end{bmatrix}$$

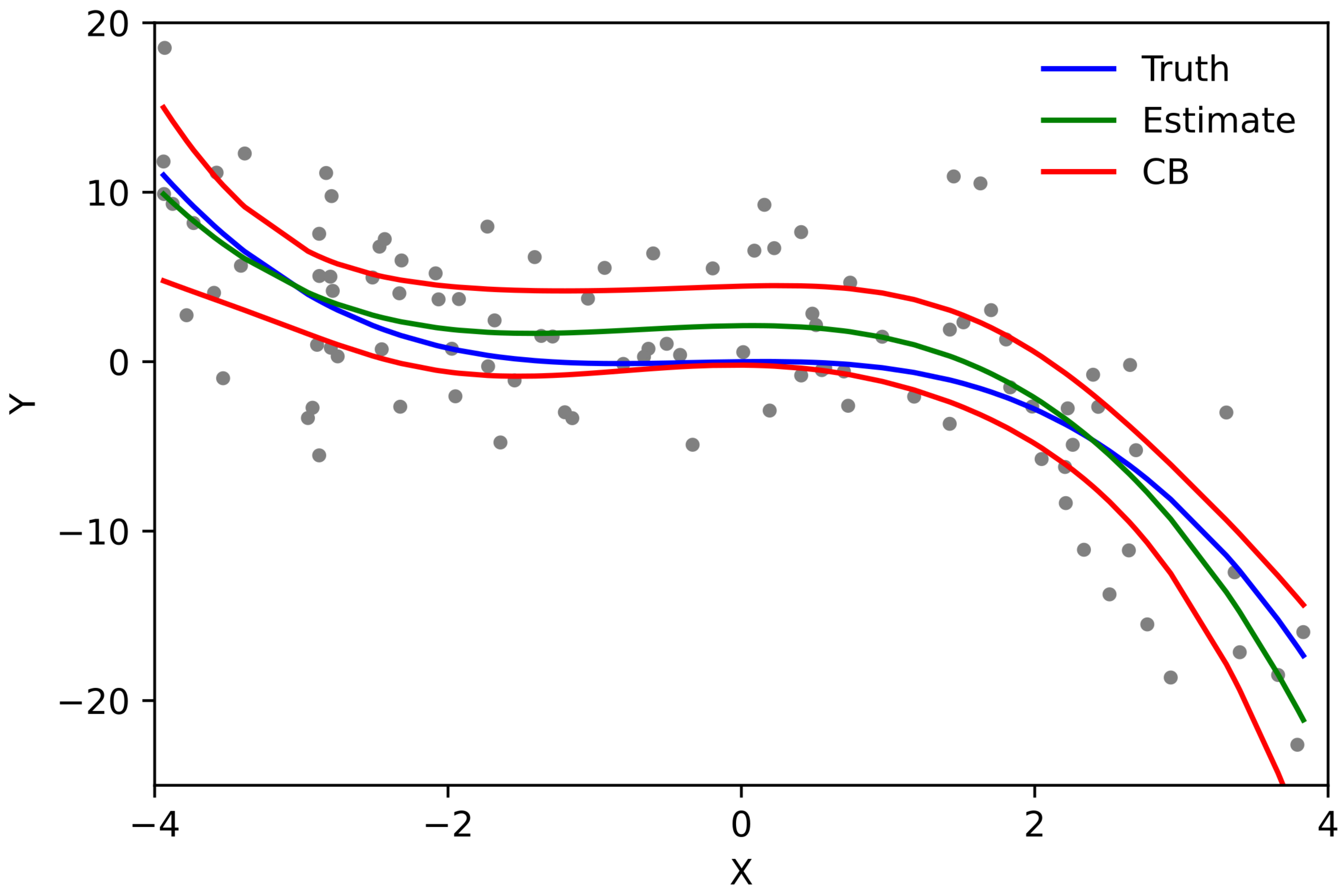
General Regression



https://commons.wikimedia.org/wiki/File:Polyreg_scheffe.svg

General Regression

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

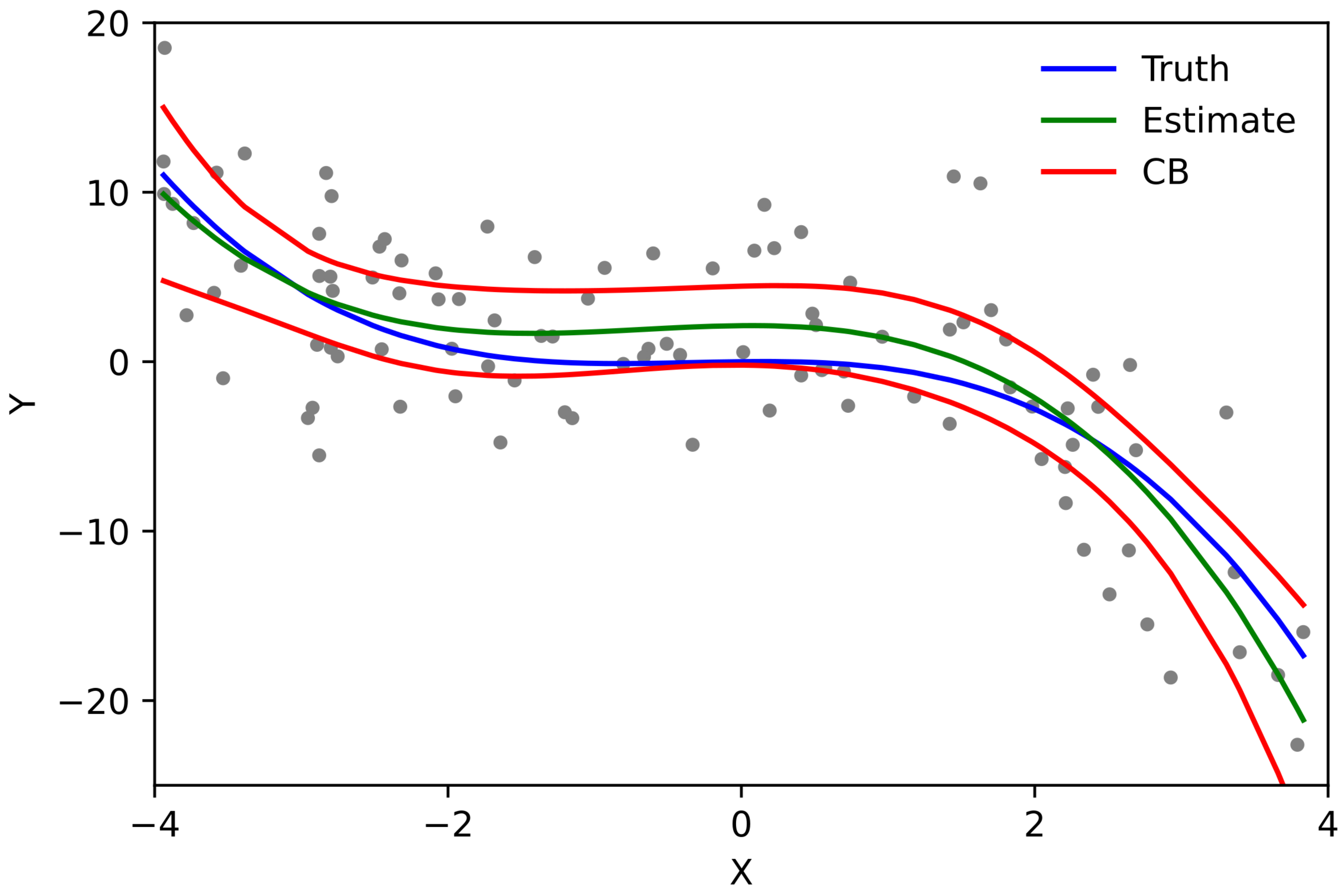


https://commons.wikimedia.org/wiki/File:Polyreg_scheffe.svg

General Regression

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

What we are estimating is a <u>mathematical function</u>



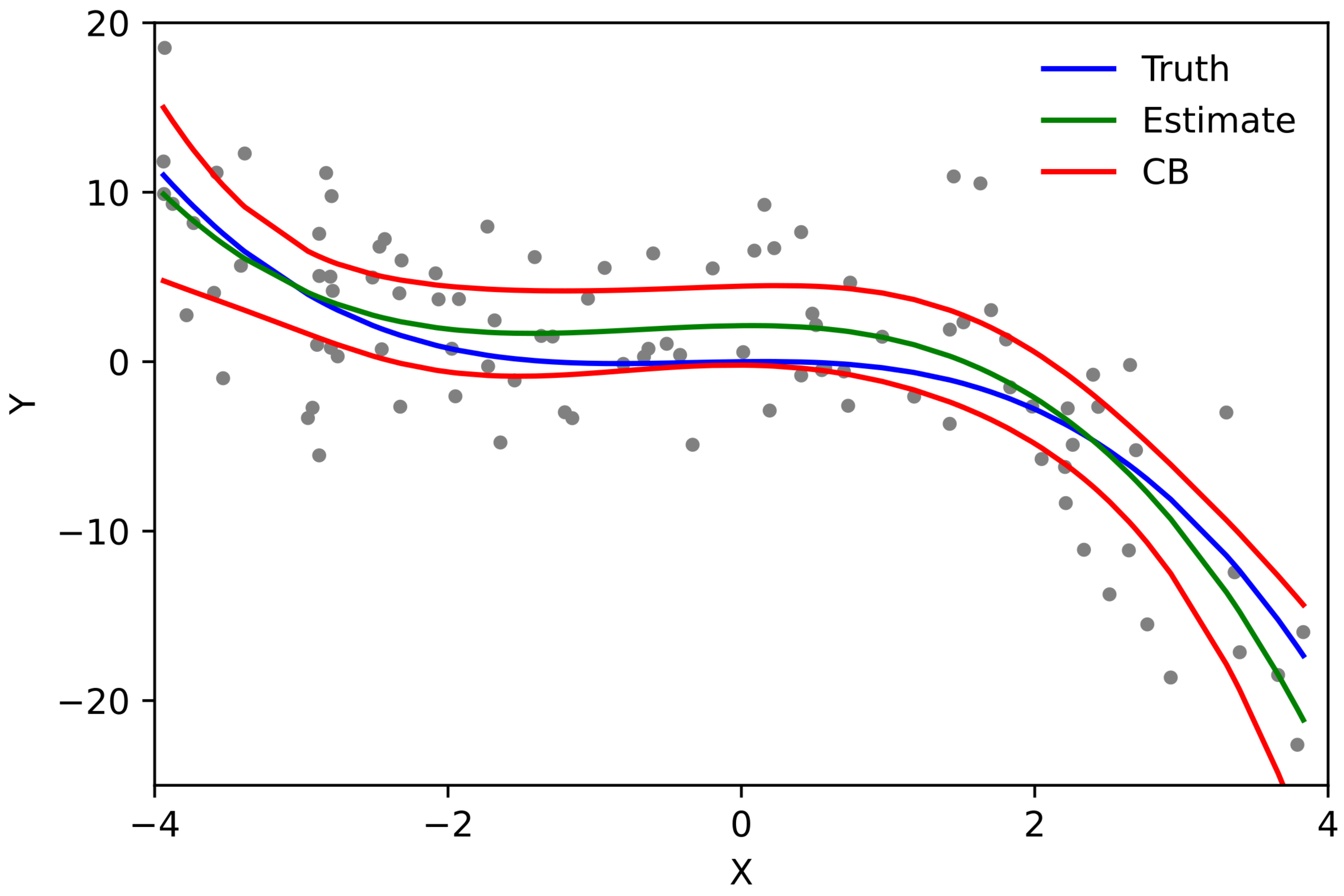
https://commons.wikimedia.org/wiki/File:Polyreg_scheffe.svg

General Regression

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

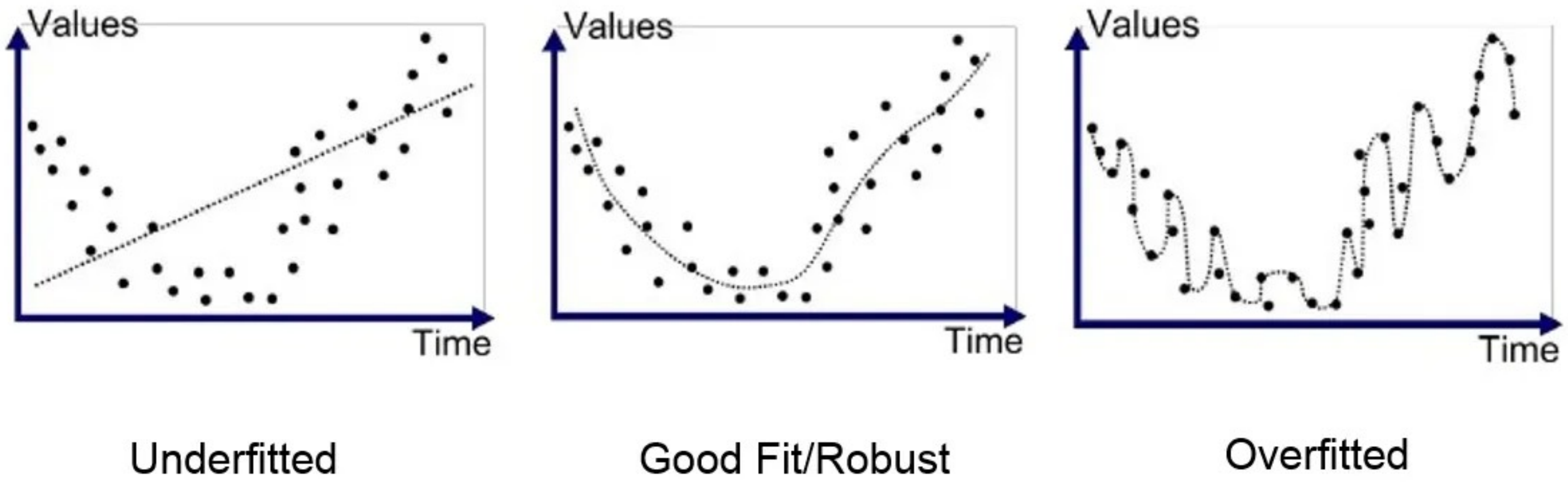
What we are estimating is a <u>mathematical function</u>

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



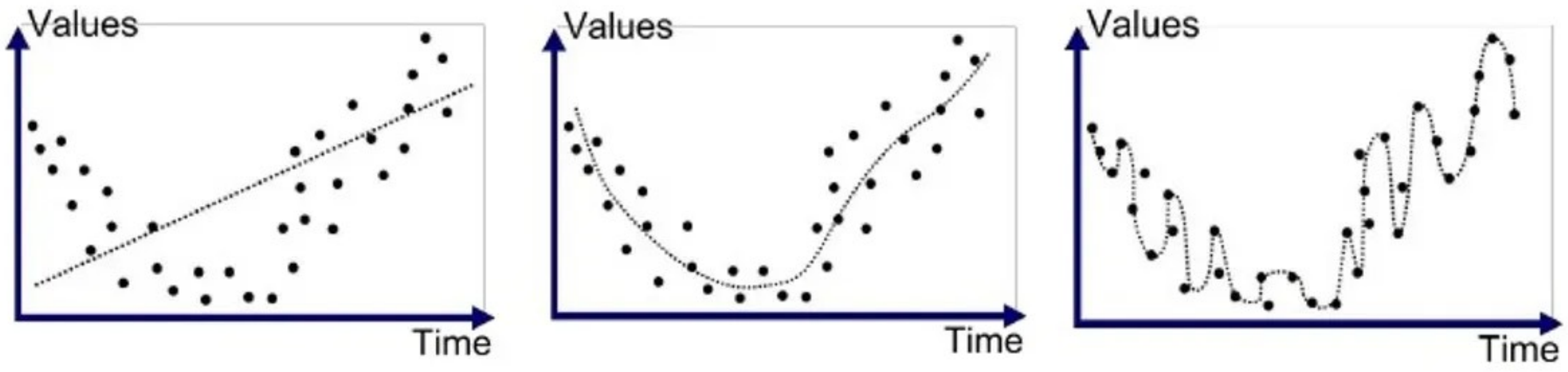
https://commons.wikimedia.org/wiki/File:Polyreg_scheffe.svg

Models



https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Models

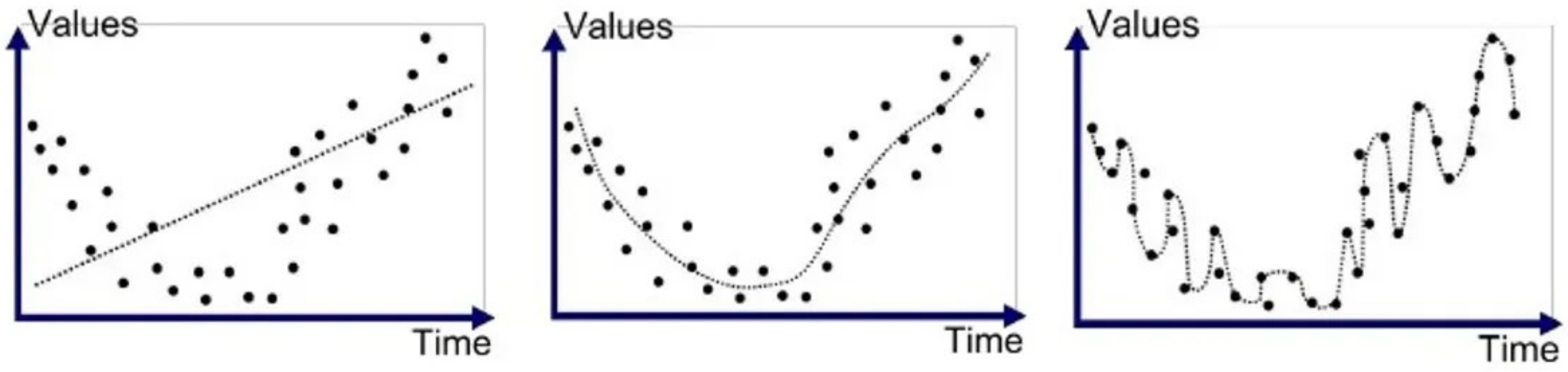


Underfitted Good Fit/Robust Overfitted

Therefore, a *model* is a mathematical function.

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Models



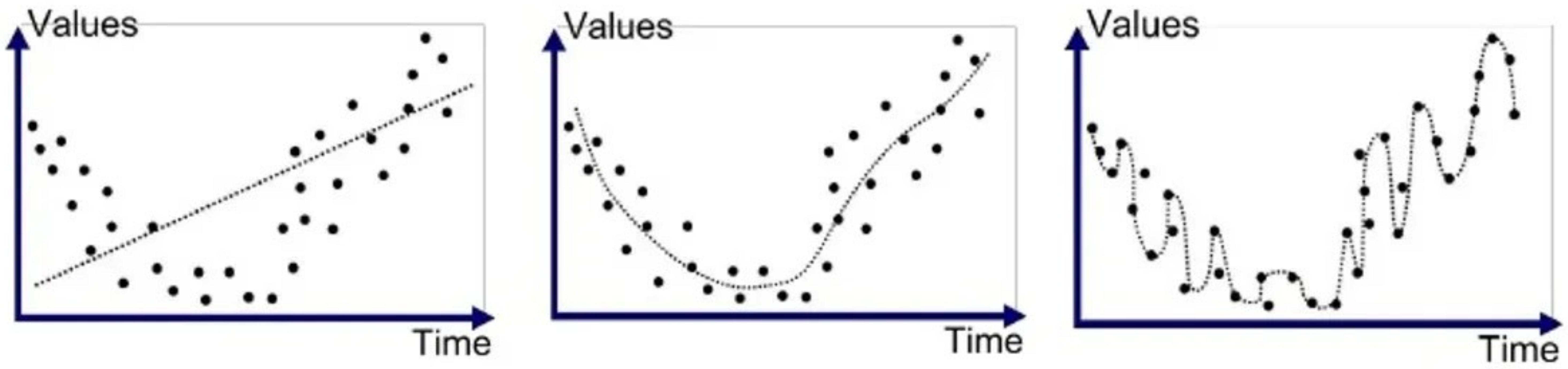
Underfitted Good Fit/Robust Overfitted

Therefore, a *model* is a mathematical function.

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

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Models



Underfitted Good Fit/Robust Overfitted

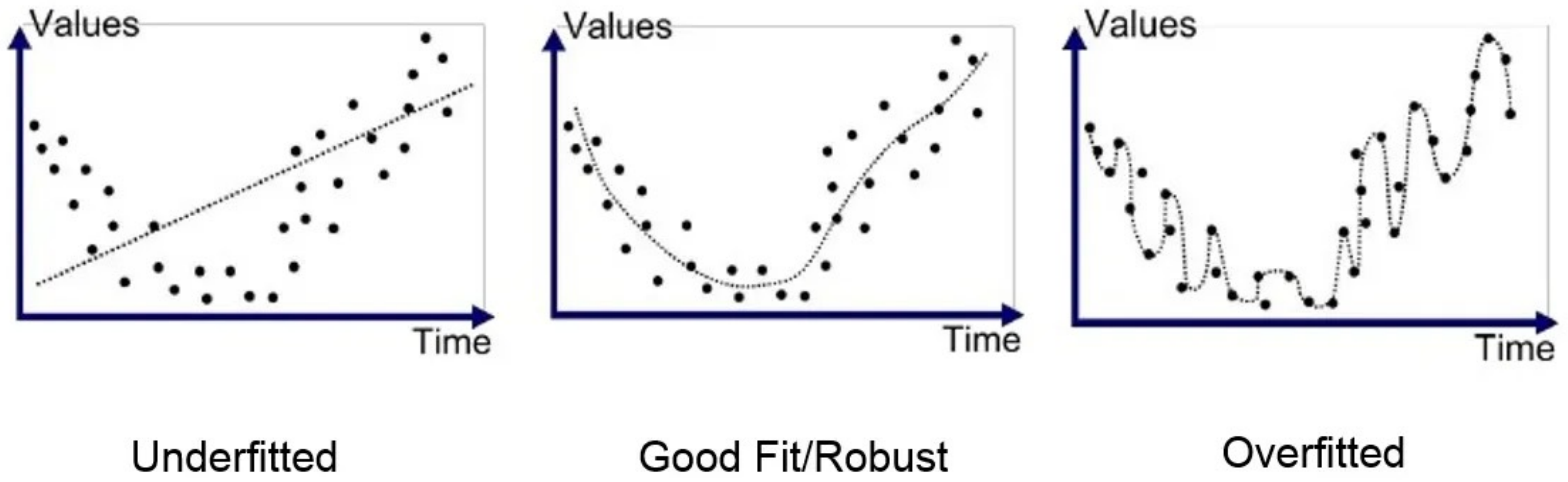
Therefore, a *model* is a mathematical function.

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

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

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

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 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.*

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

How To: Prediction

How To: Prediction

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

How To: Prediction

Problem. Given the data $\{(x_1, y_1), \ldots, (x_k, y_k)\}$ 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 $f(x')$.

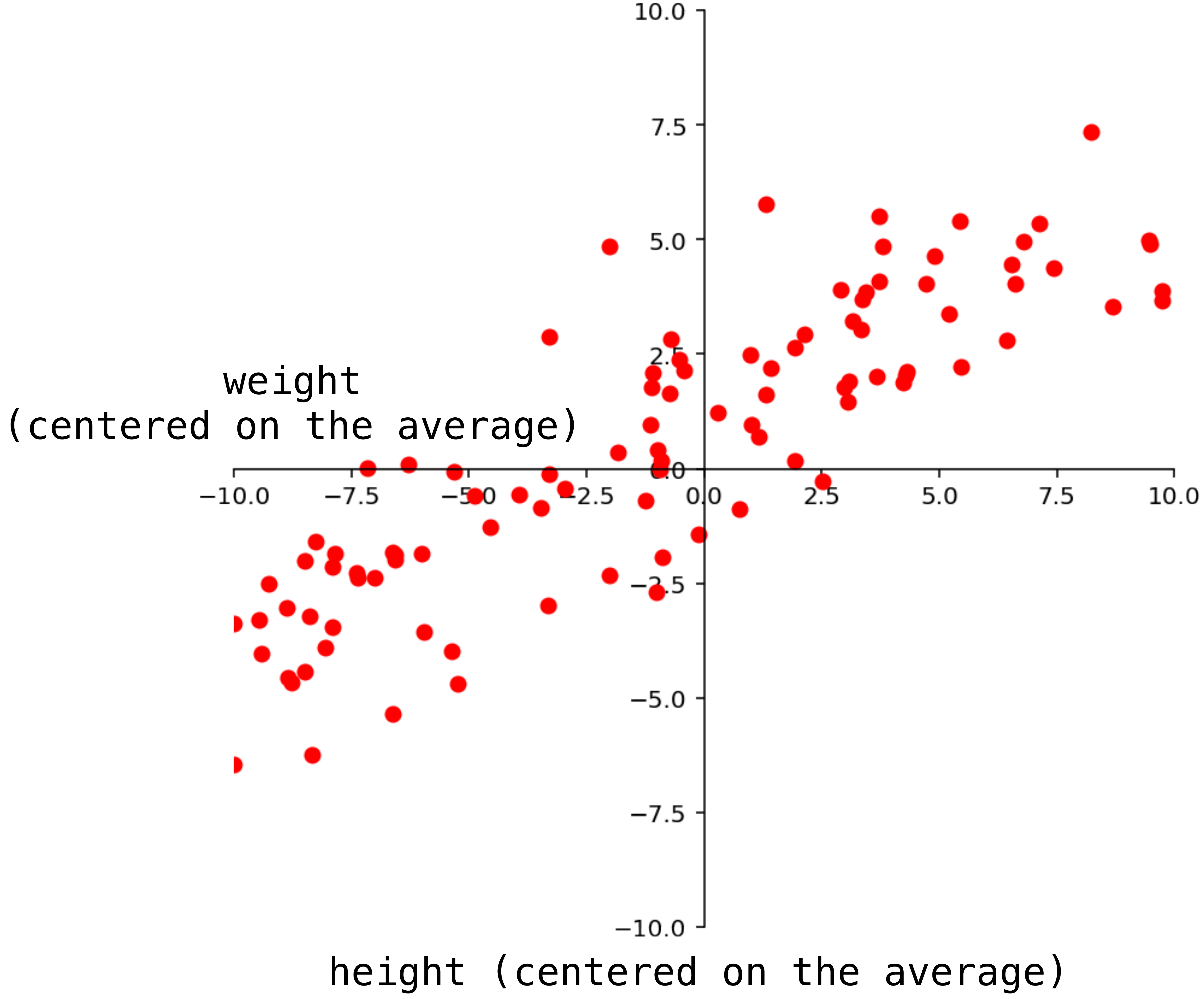
How To: Prediction

Problem. Given the data $\{(x_1, y_1), \ldots, (x_k, y_k)\}$ 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 $f(x')$.

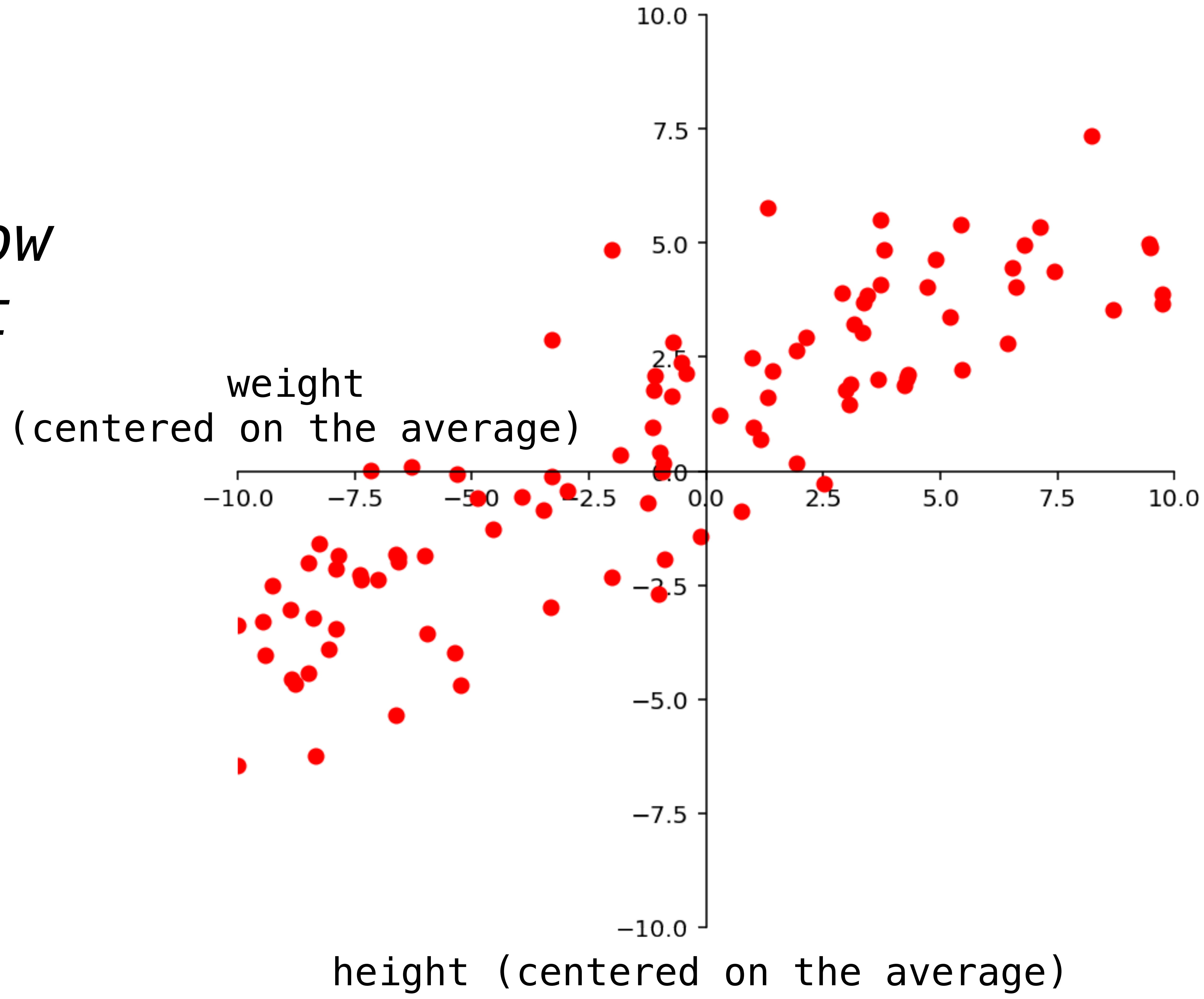
This generalizes to any model fitting problem

Example: Height from Weight



Example: Height from Weight

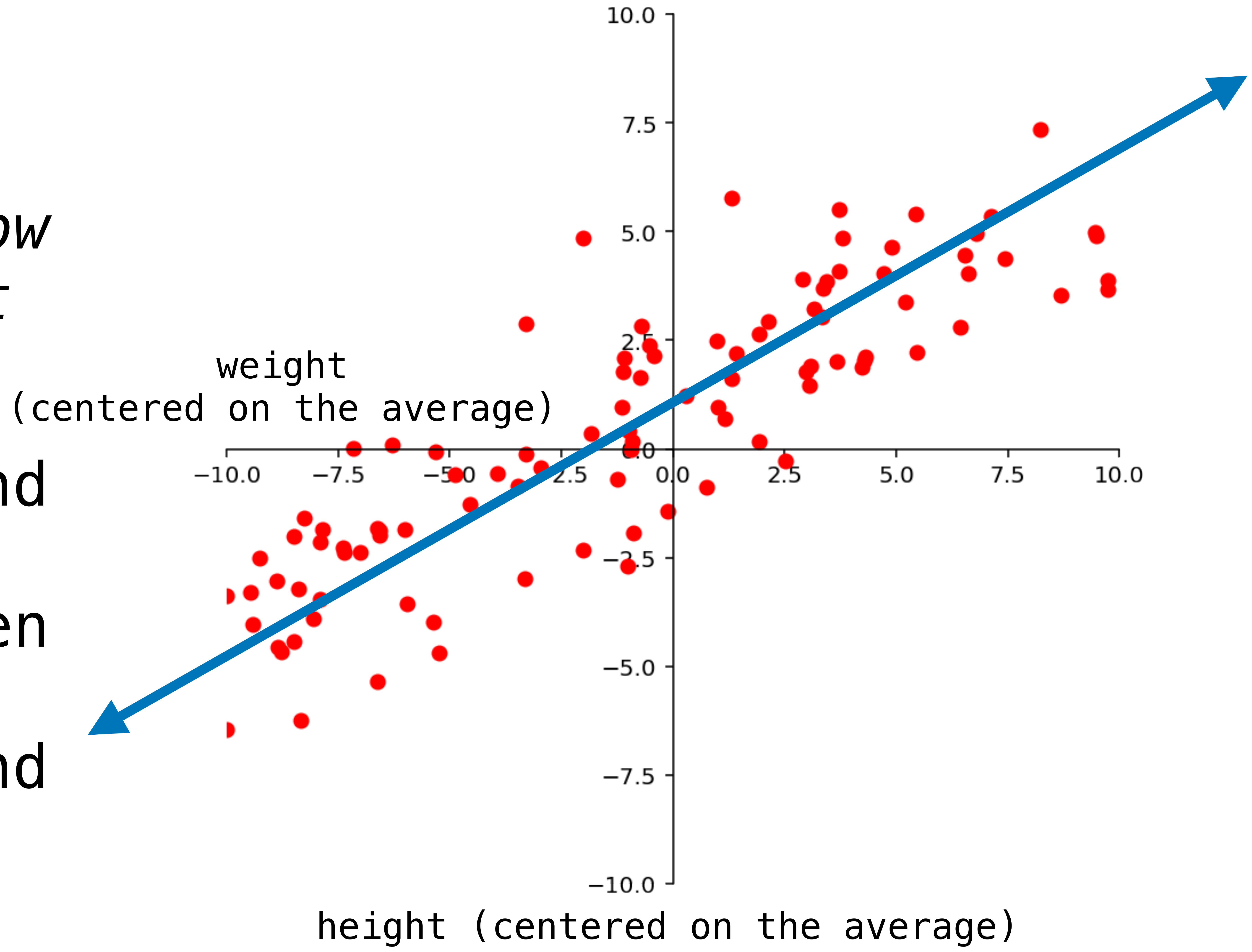
Suppose we know that person X weighs 150lb. *How would we guess the height of person X?*



Example: Height from Weight

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

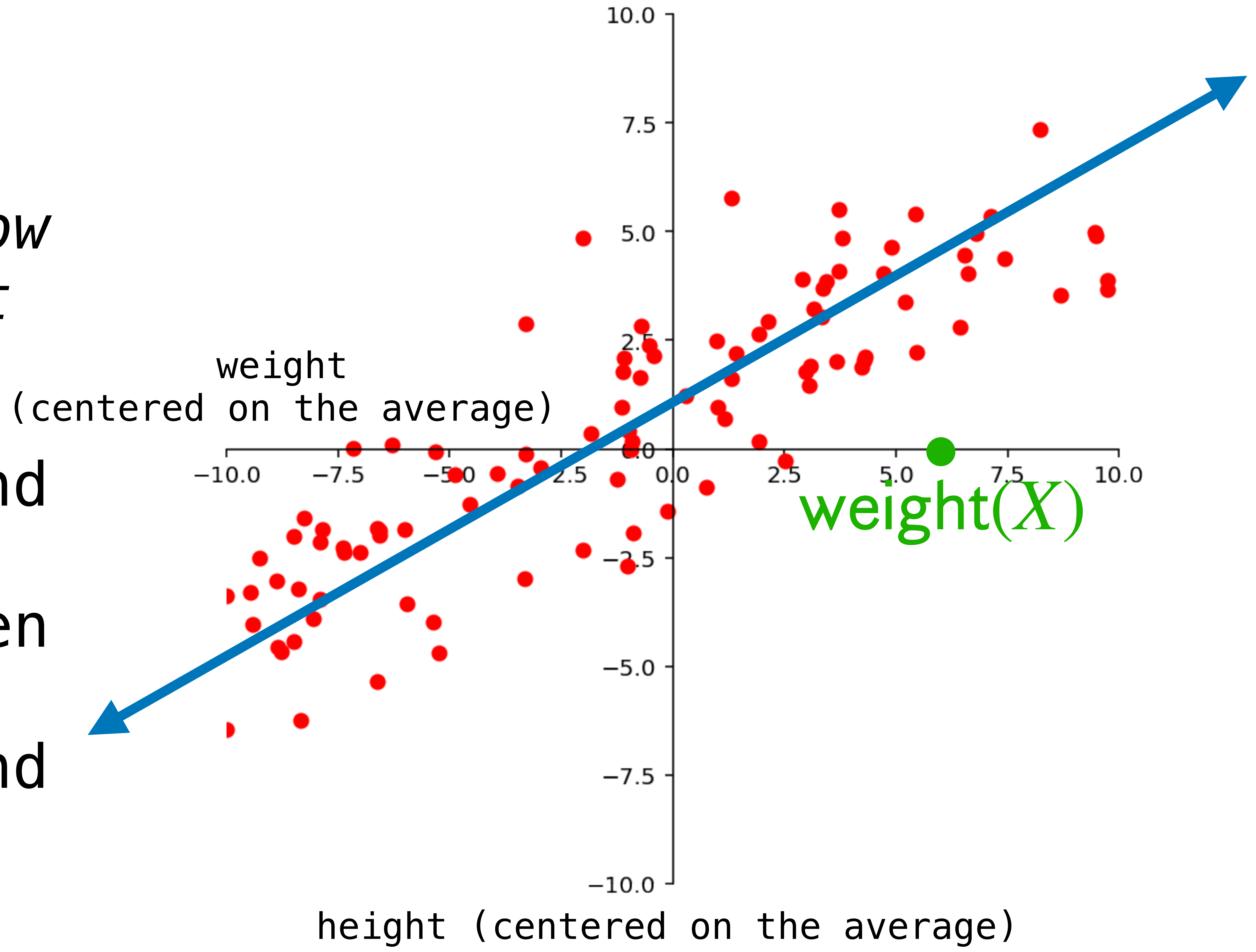
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



Example: Height from Weight

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

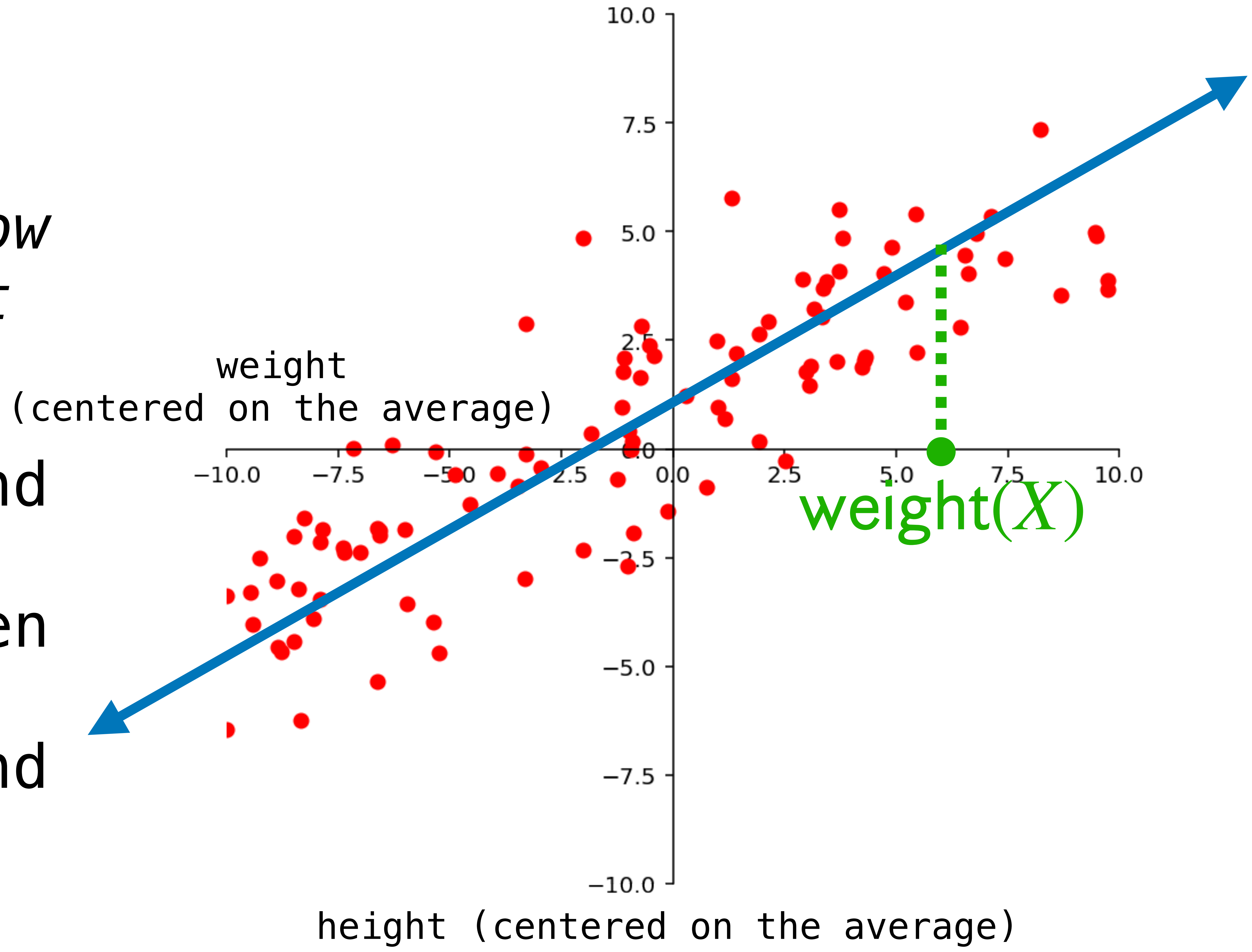
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



Example: Height from Weight

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

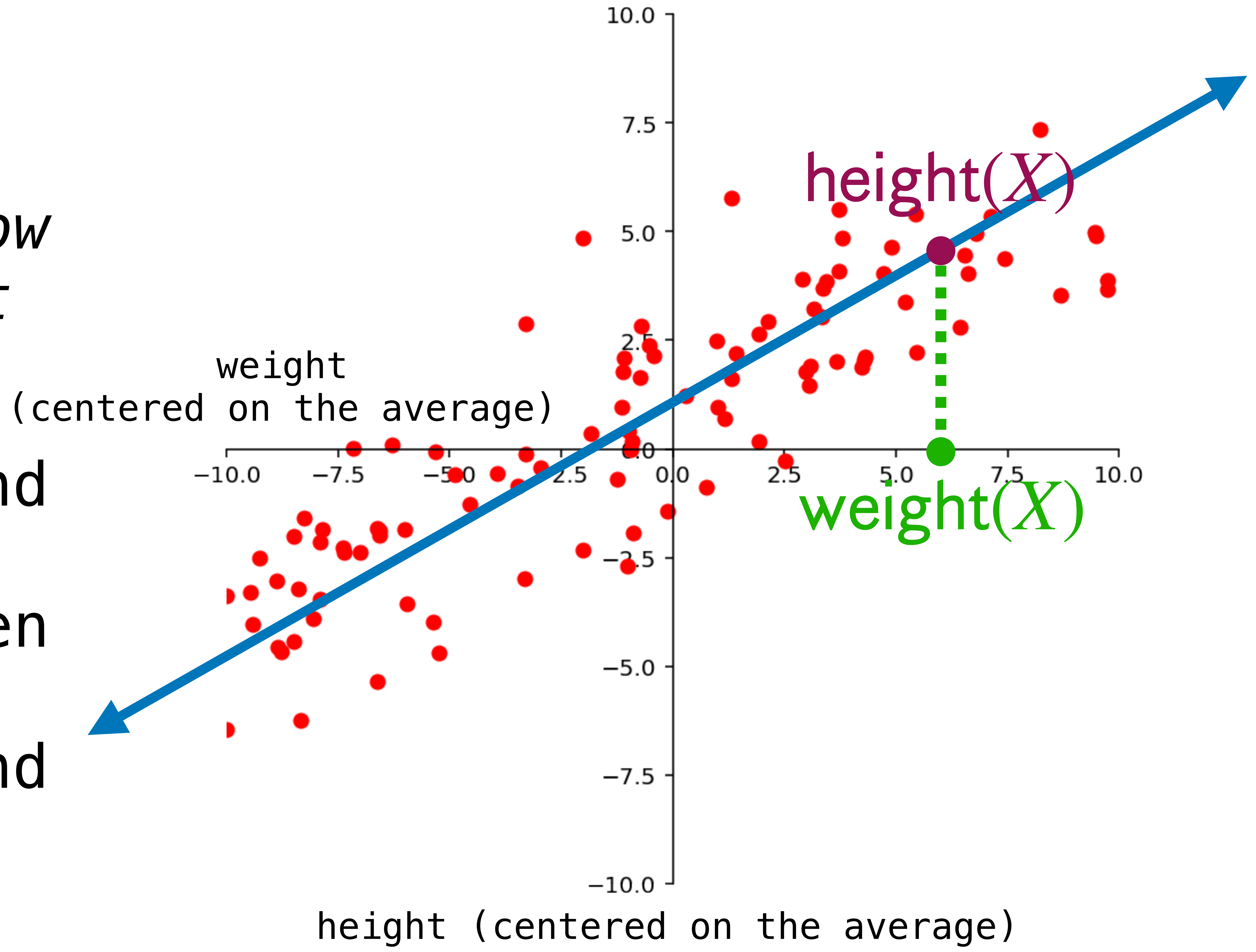
If we know the heights and weights of a population (from which X comes), then we can **find the line of best fit for that data** and then use that function.



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.



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

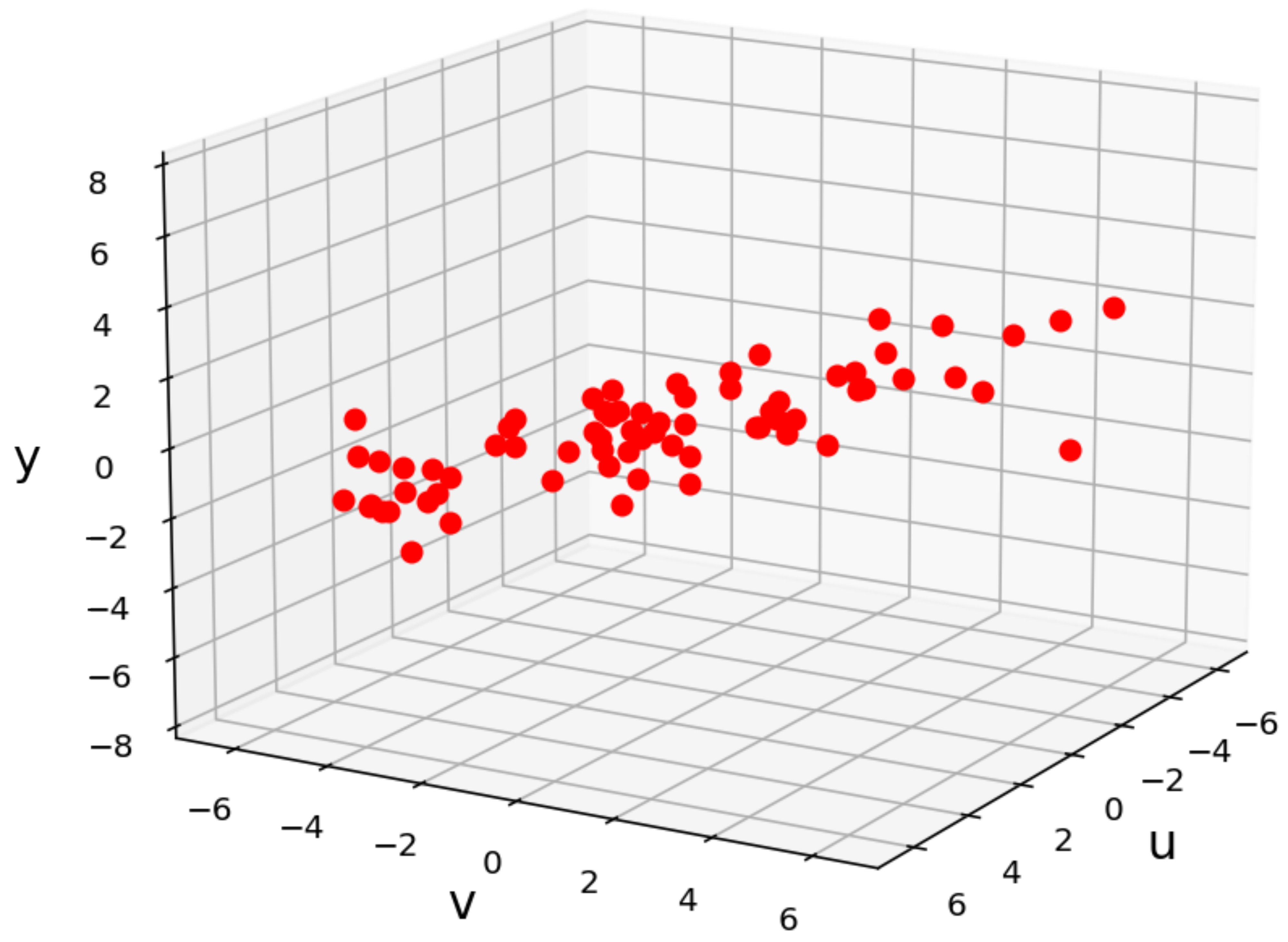
Example: Terrain Data

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

Problem: Find the <u>plane</u> which "best" fits the data.

Figure 23.1

Terrain Data for Multiple Regression



Example: Terrain Data

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

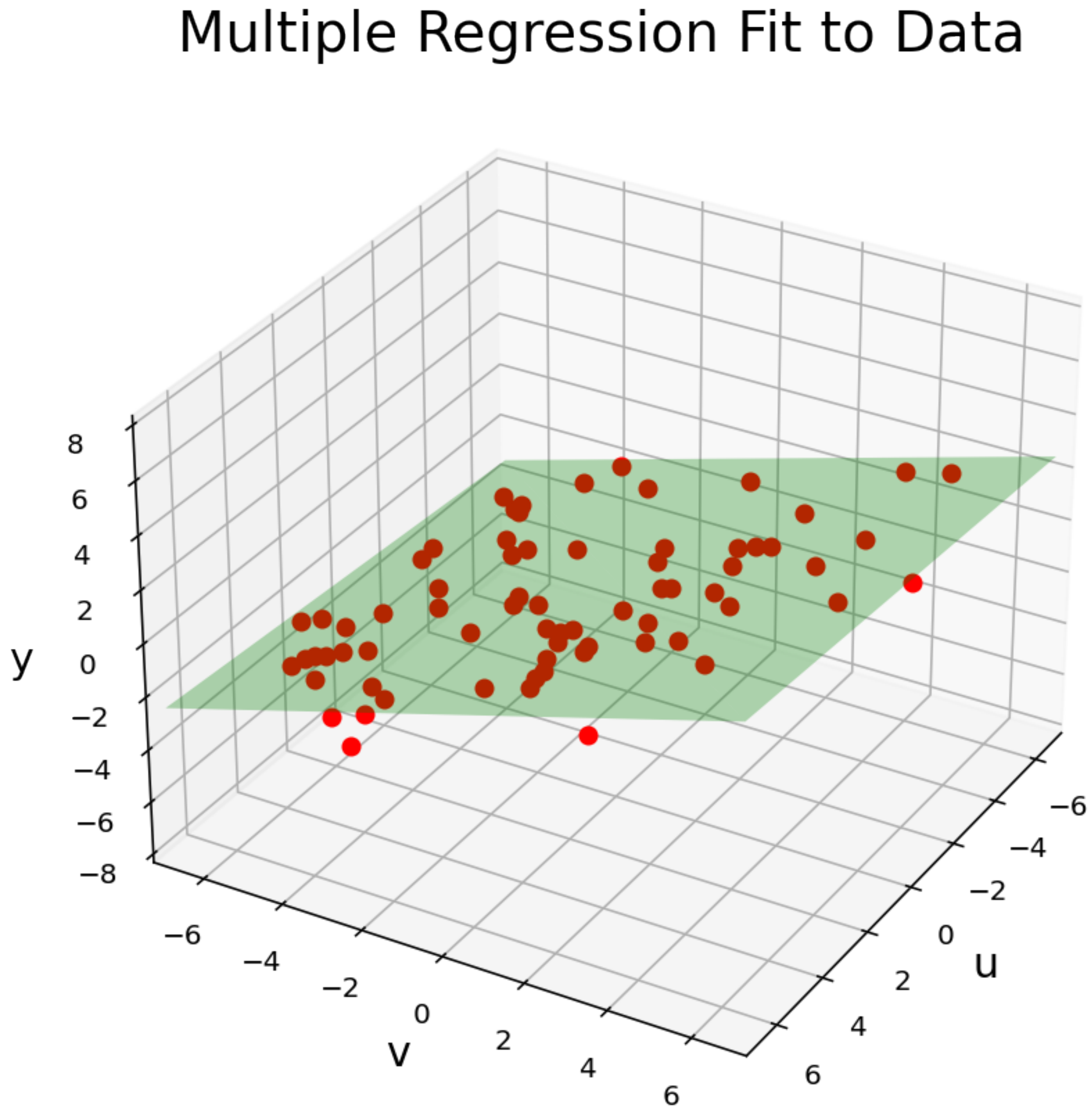
Problem: Find $\beta_0, \beta_1, \beta_2$ such that

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

which minimizes

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

Figure 23.2



Example: Terrain Data

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

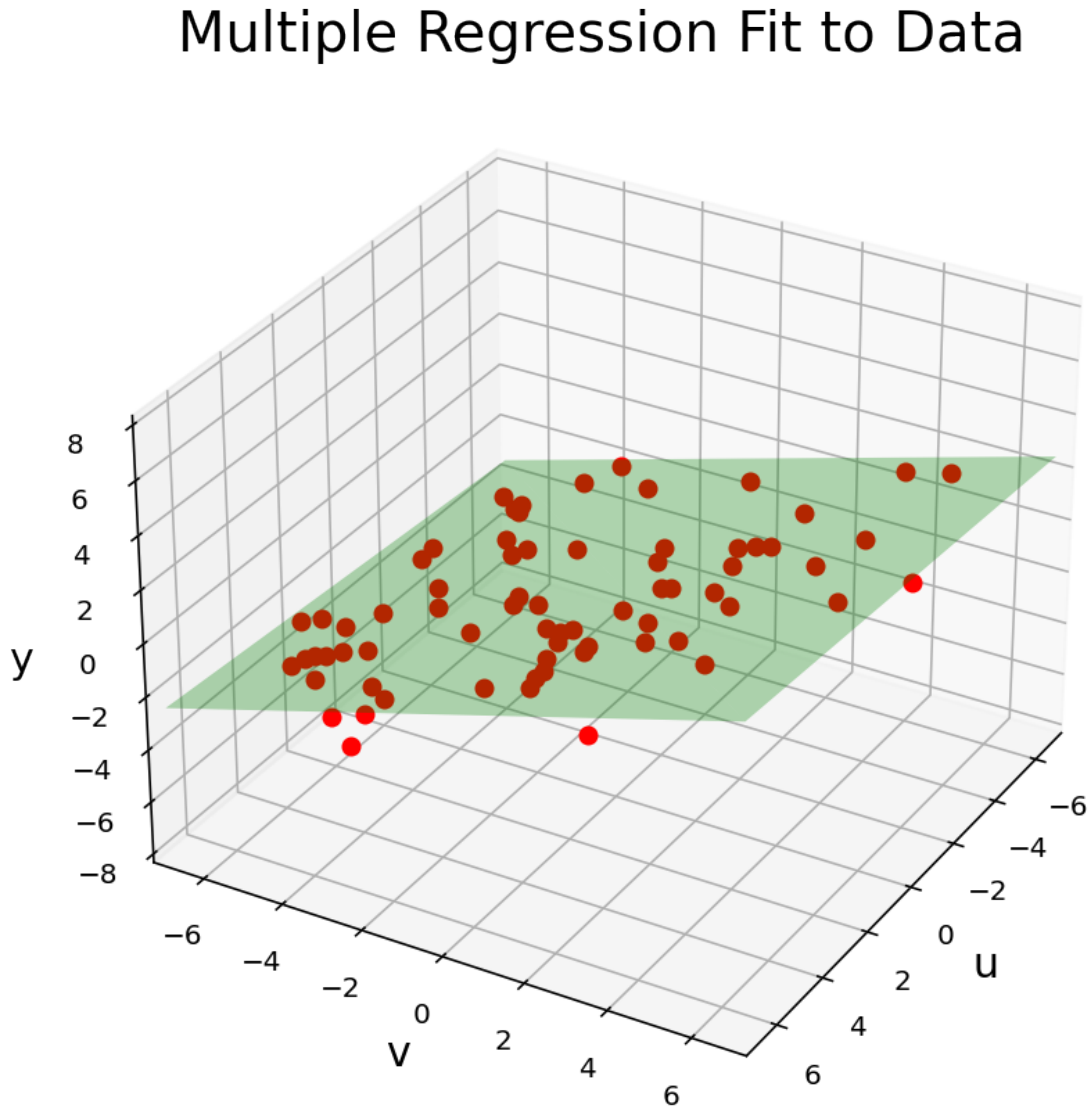
Problem: Find $\beta_0, \beta_1, \beta_2$ such that

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

which minimizes

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

Figure 23.2



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

Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)\}$
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

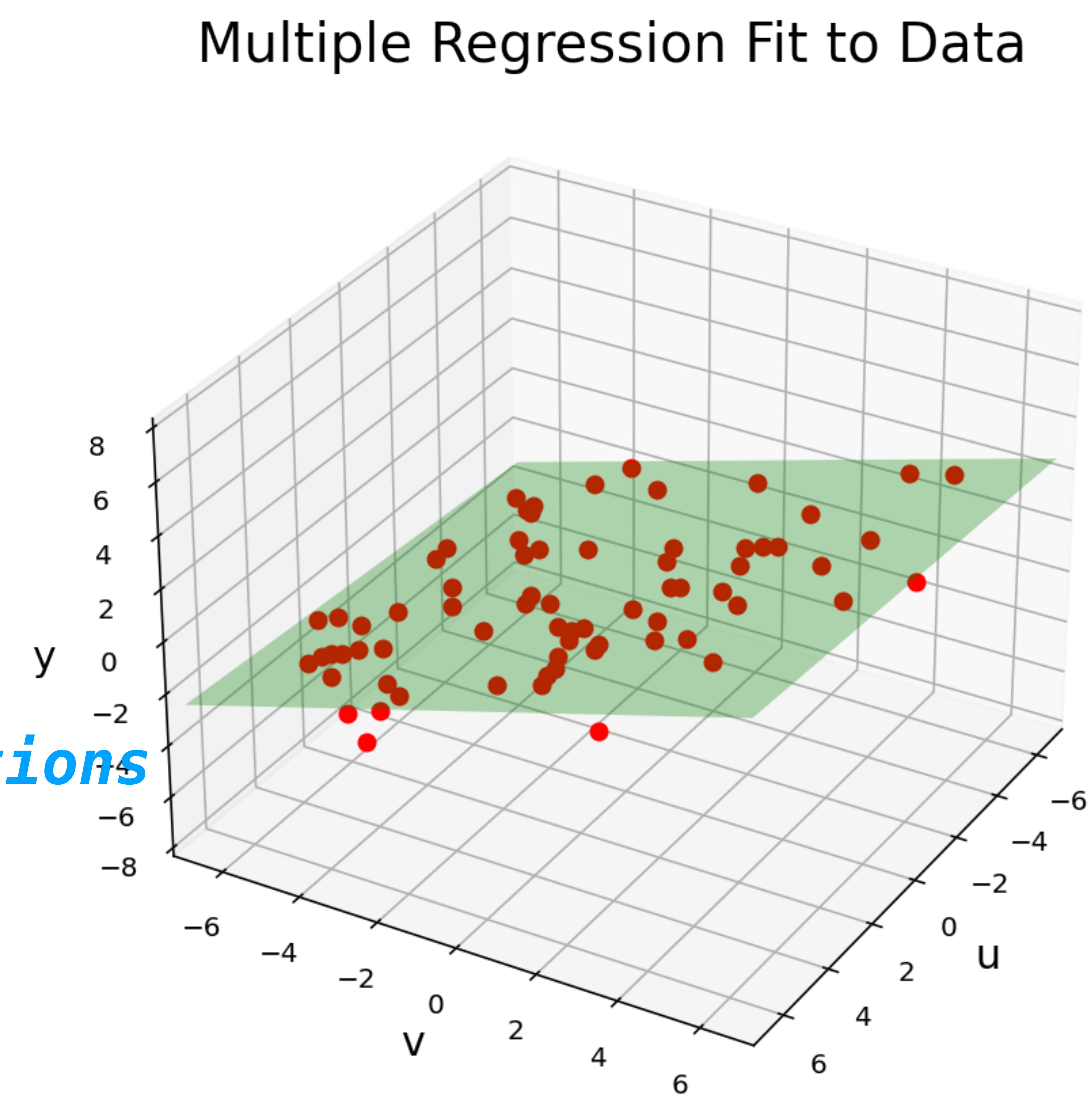
$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y$$

recall: planes are given by linear equations

which minimizes

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

Figure 23.2



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

Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)\}$ 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

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

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

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: $\{(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)\}$ 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: $\{(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)\}$
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

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

$$\underbrace{\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & y_k \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}_{\vec{\beta}} = \underbrace{\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix}}_{\mathbf{z}}$$

Step 2: Rewrite the system as a matrix equation.

Example: Terrain Data

Dataset: $\{(x_1, y_1, z_1), \ldots, (x_k, y_k, z_k)\}$
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

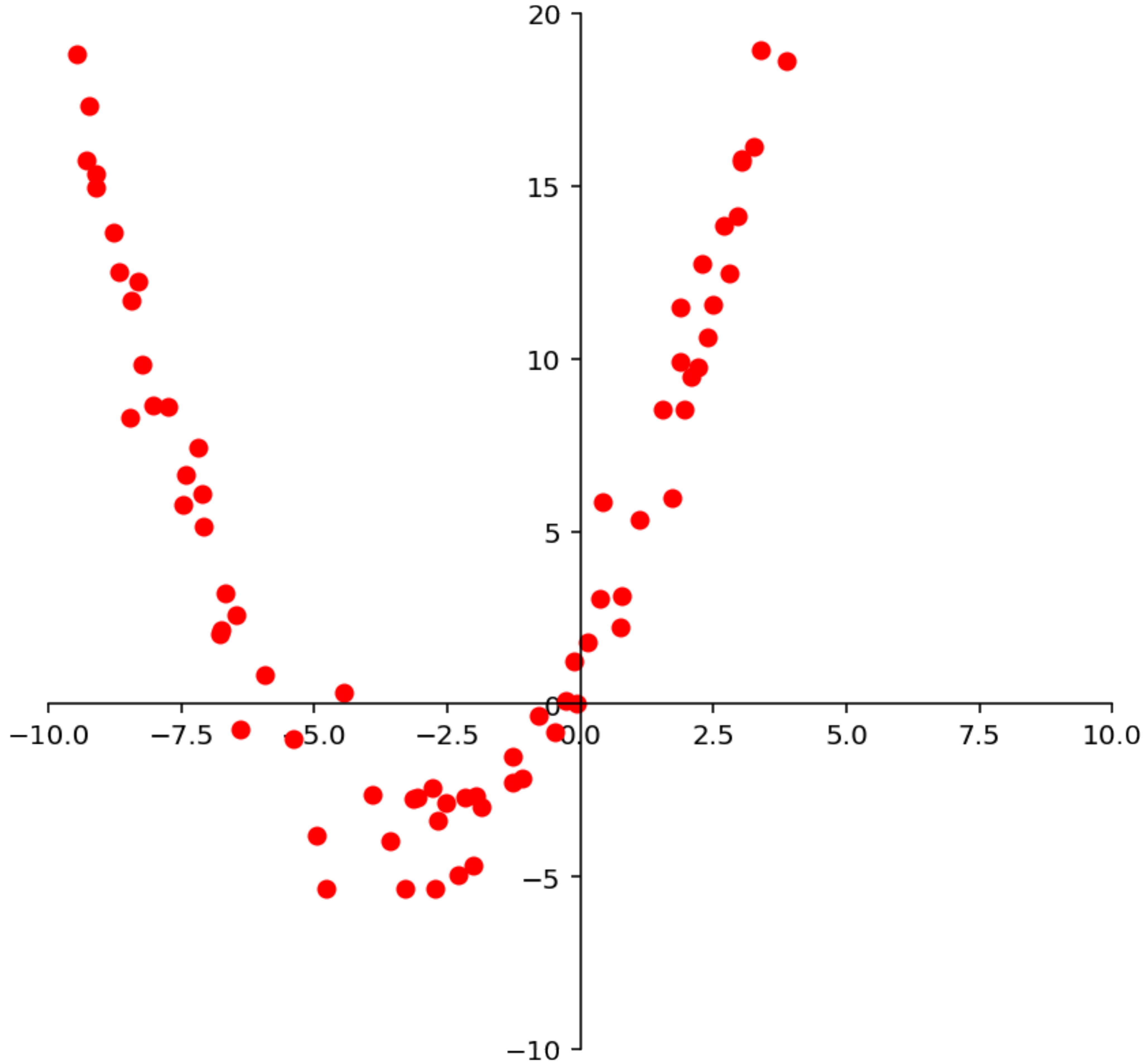
1. What if we have *more than one* independent value?

   multiple regression, (hyper)plane of best fit

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

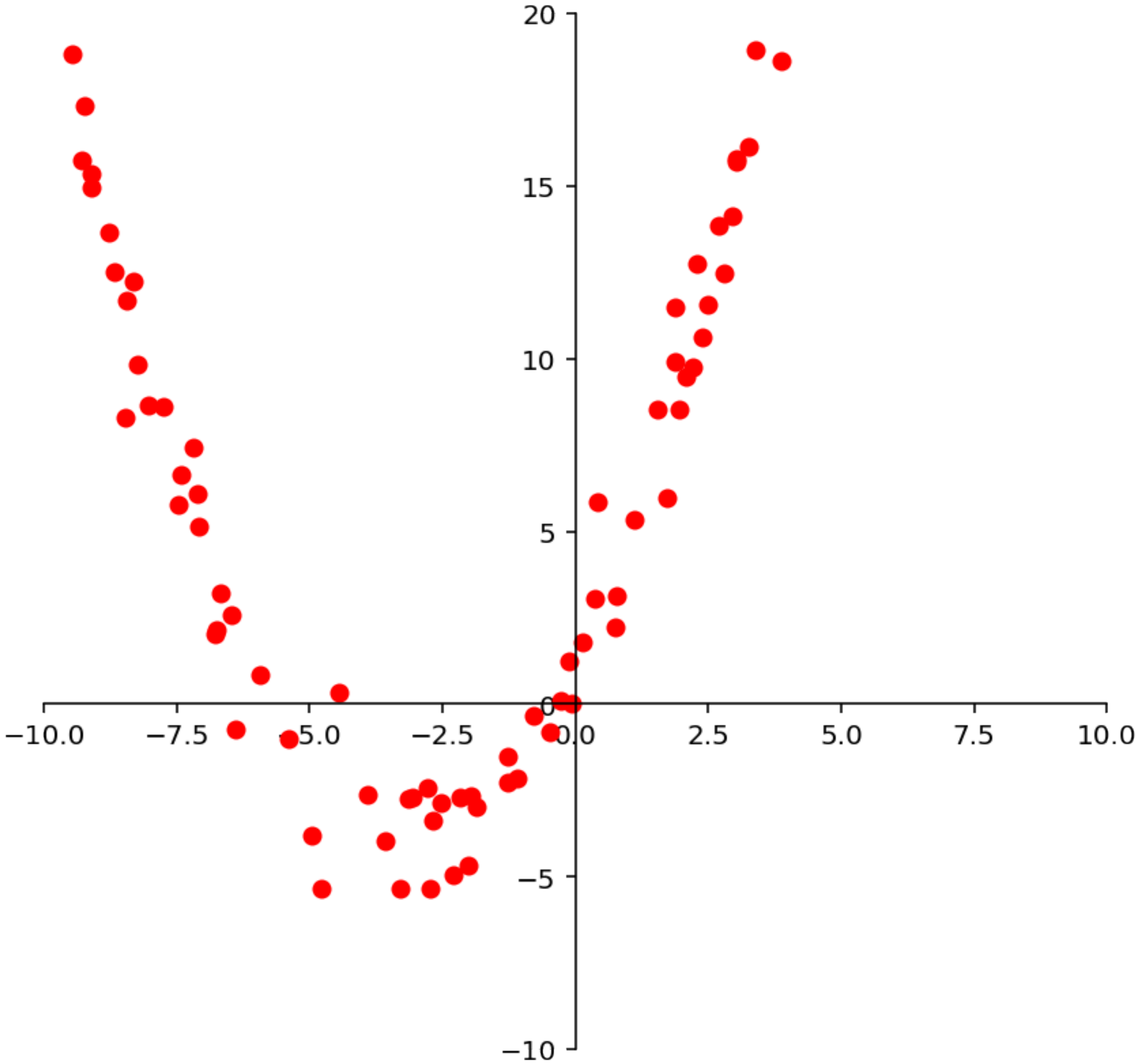
   e.g., polynomial regression

Example: Best Fit Quadratic



Example: Best Fit Quadratic

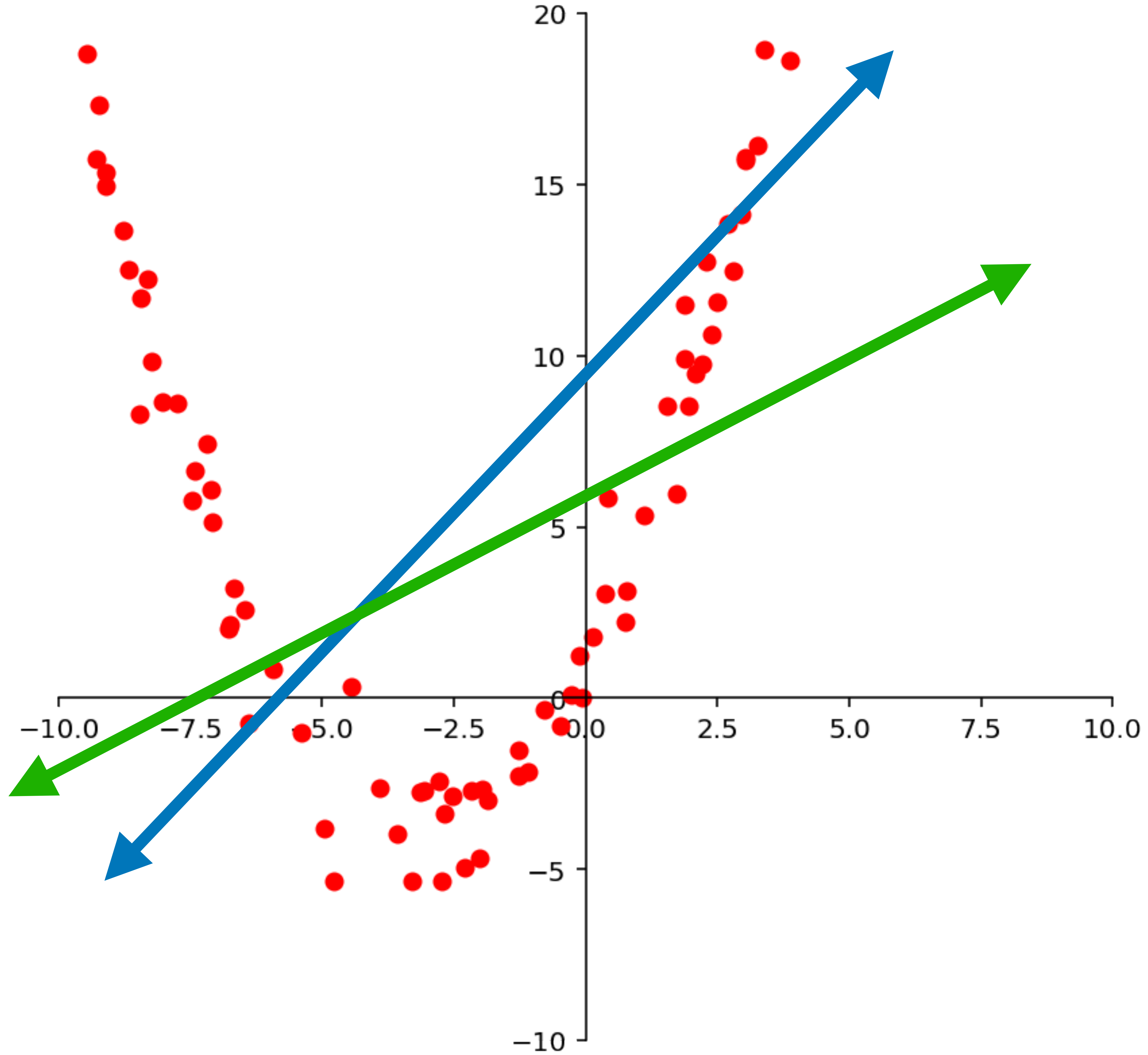
Dataset: $\{(x_1, y_1), \ldots, (x_k, y_k)\}$



Example: Best Fit Quadratic

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

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

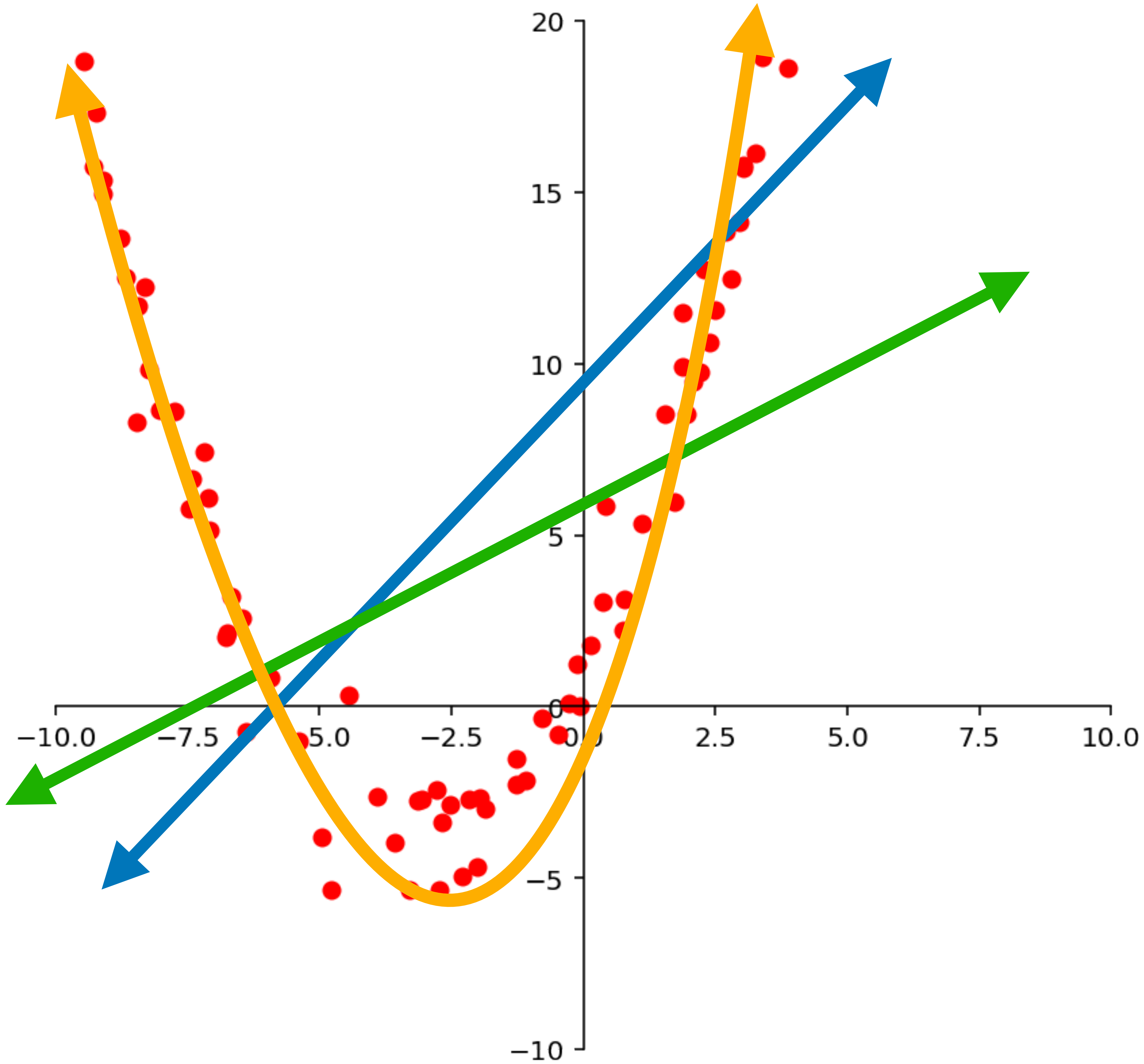


Example: Best Fit Quadratic

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

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

What about a parabola?



Example: Best Fit Quadratic

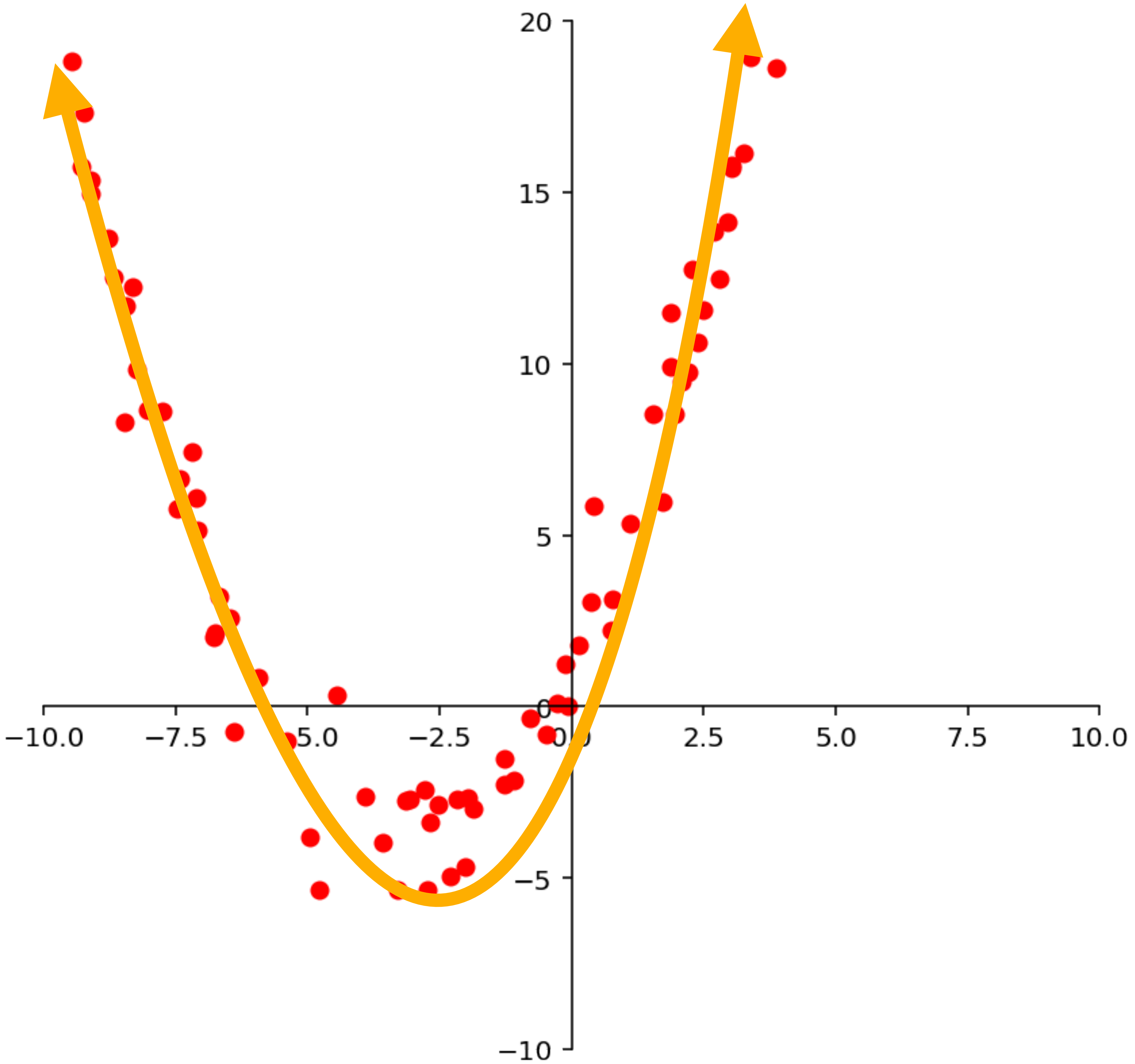
Dataset: $\{(x_1, y_1), \ldots, (x_k, y_k)\}$

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

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

minimizes

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



Example: Best Fit Quadratic

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

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

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

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

minimizes

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

$$\underline{\beta_0} + \underline{\beta_1} x_1 + \underline{\beta_2} x_1^2 = y_1$$

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

$$\vdots$$

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

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

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

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

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

minimizes

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

$$\underbrace{\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_k & x_k^2 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}}_{\vec{\beta}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}}_{\mathbf{y}}$$

Step 2: Rewrite the system as a matrix equation.

Example: Best Fit Quadratic

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

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.

The Takeaway

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

Linear in Parameters

Definition. A function $f: \mathbb{R}^n \to \mathbb{R}$ is **linear in the parameters** $\beta_1, \ldots, \beta_k$ if it can be written as

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

for functions $\phi_1, \ldots, \phi_k : \mathbb{R}^n \to \mathbb{R}$

Example: $f(x, y, z) = \beta_0 \, 2^{x+y} \log z + \beta_1 (x^2 + y^2) + \beta_2$

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

General Linear Regression

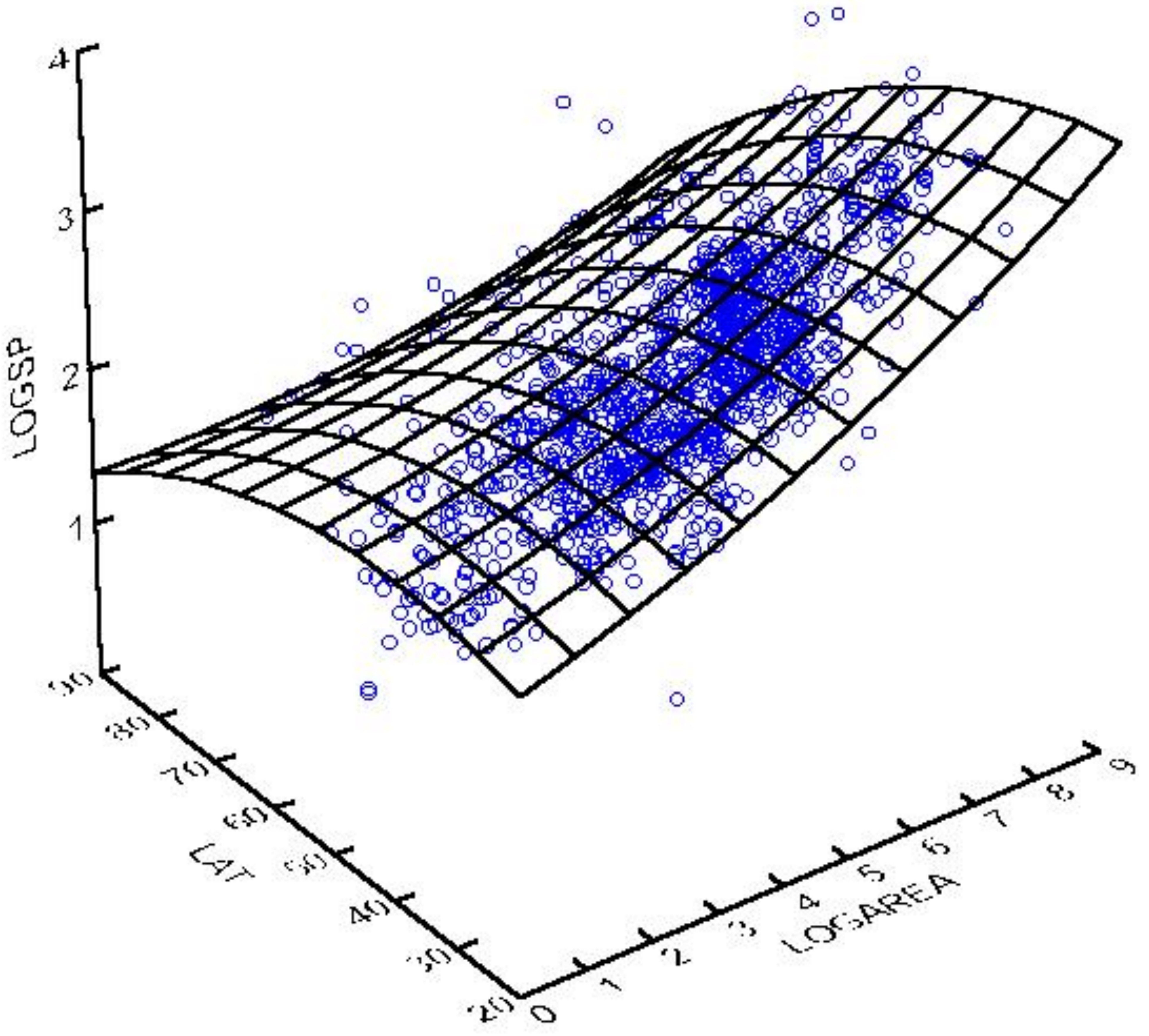
dataset: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$

Problem. Given a function

$$f_{\beta_1,\ldots,\beta_k} : \mathbb{R}^n \to \mathbb{R}$$

which is *linear in the parameters* $\beta_1, \ldots \beta_k$, find values for $\beta_1, \ldots, \beta_k$ which minimize

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



https://ordination.okstate.edu/MULTIPLE.htm

General Linear Regression

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

Problem. Given a function

$$f_{\beta_1,\ldots,\beta_k} : \mathbb{R}^n \to \mathbb{R}$$

which is *linear in the parameters* $\beta_1, \ldots \beta_k$, find values for $\beta_1, \ldots, \beta_k$ which minimize

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

$$\begin{aligned}
\beta_1\phi_1(\mathbf{x}_1) + \ldots + \beta_k\phi_k(\mathbf{x}_1) &= y_1 \\
\beta_1\phi_1(\mathbf{x}_2) + \ldots + \beta_k\phi_k(\mathbf{x}_2) &= y_2 \\
&\vdots \\
\beta_1\phi_1(\mathbf{x}_2) + \ldots + \beta_k\phi_k(\mathbf{x}_2) &= y_2
\end{aligned}$$

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

General Linear Regression

This is still linear in the β's

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

Problem. Given a function

$$f_{\beta_1,\ldots,\beta_k} : \mathbb{R}^n \to \mathbb{R}$$

which is *linear in the parameters* $\beta_1, \ldots \beta_k$, find values for $\beta_1, \ldots, \beta_k$ which minimize

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

$$\beta_1\phi_1(\mathbf{x}_1) + \ldots + \beta_k\phi_k(\mathbf{x}_1) = y_1$$
$$\beta_1\phi_1(\mathbf{x}_2) + \ldots + \beta_k\phi_k(\mathbf{x}_2) = y_2$$
$$\vdots$$
$$\beta_1\phi_1(\mathbf{x}_2) + \ldots + \beta_k\phi_k(\mathbf{x}_2) = y_2$$

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

General Linear Regression

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

Problem. Given a function

$$f_{\beta_1,\ldots,\beta_k} : \mathbb{R}^n \to \mathbb{R}$$

which is *linear in the parameters* $\beta_1, \ldots \beta_k$, find values for $\beta_1, \ldots, \beta_k$ which minimize

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

design matrix

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

Step 2: Rewrite the system as a matrix equation.

General Linear Regression

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

Problem. Given a function

$$f_{\beta_1,\ldots,\beta_k} : \mathbb{R}^n \to \mathbb{R}$$

which is *linear in the parameters* $\beta_1, \ldots \beta_k$, find values for $\beta_1, \ldots, \beta_k$ which minimize

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

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

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 $\beta_1, \beta_2, \ldots, \beta_k$ given the dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$.

How To: Design Matrices

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

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

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

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

$$\mathbf{x}_1 = (0,0,0) \qquad y_1 = 5$$
$$\mathbf{x}_2 = (\pi,3,1) \qquad y_2 = 3$$

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

$$\beta_1 \cos x + \beta_2 e^{-x_1 x_2} - \beta_1 x_3 + \beta_3$$

$$\beta_1 \boxed{(\cos x - x_3)} + \beta_2 \boxed{e^{-x_1 x_2}} + \beta_3 \boxed{}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ -2 & e^{-3\pi} & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \end{bmatrix}$$

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.