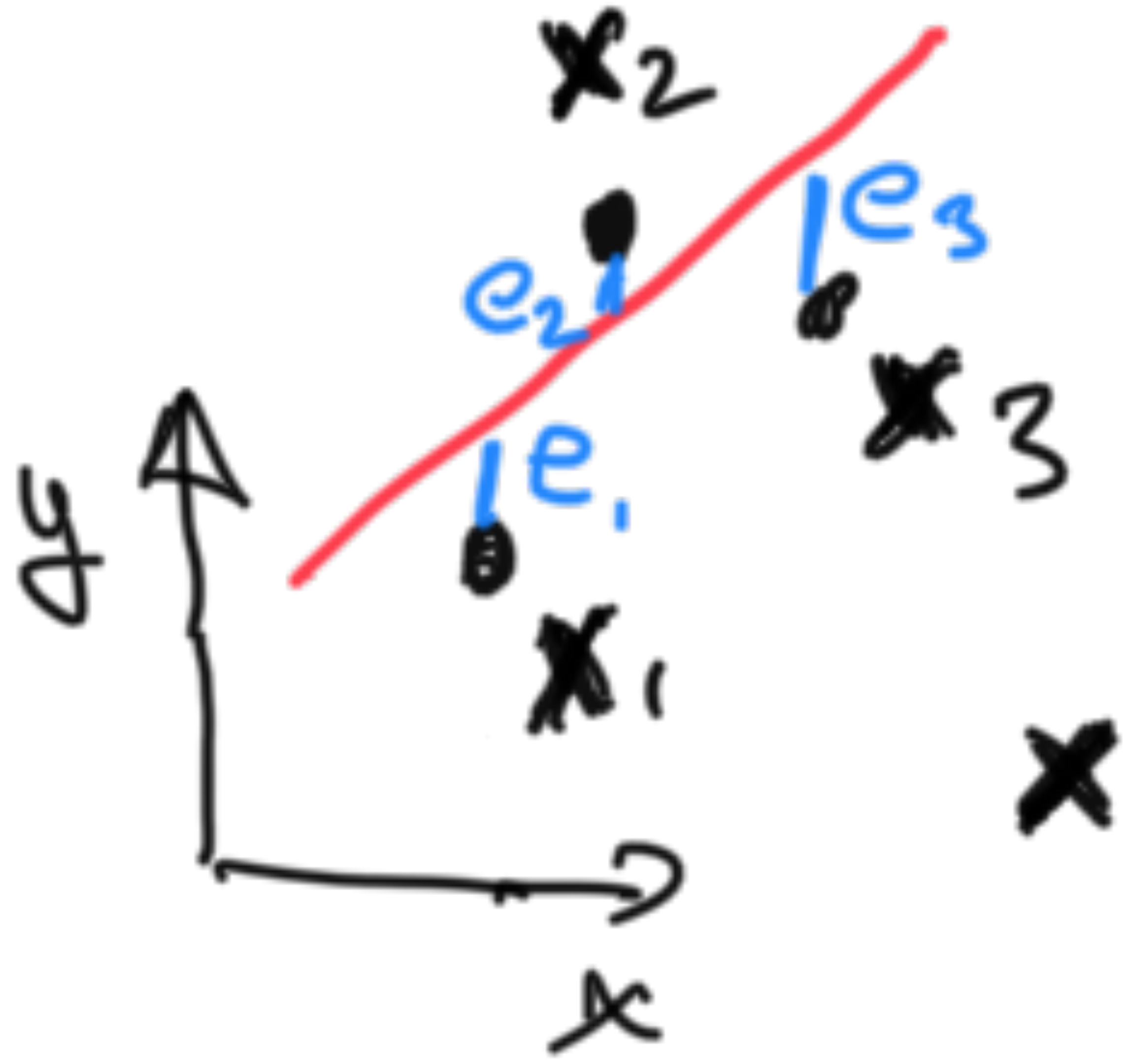
Lecture 4 pre-video

Numpy + OLS setup

https://github.com/COGS118A/demo_notebooks/blob/main/lecture_04_numpy.ipynb

Derive OLS



$$\mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

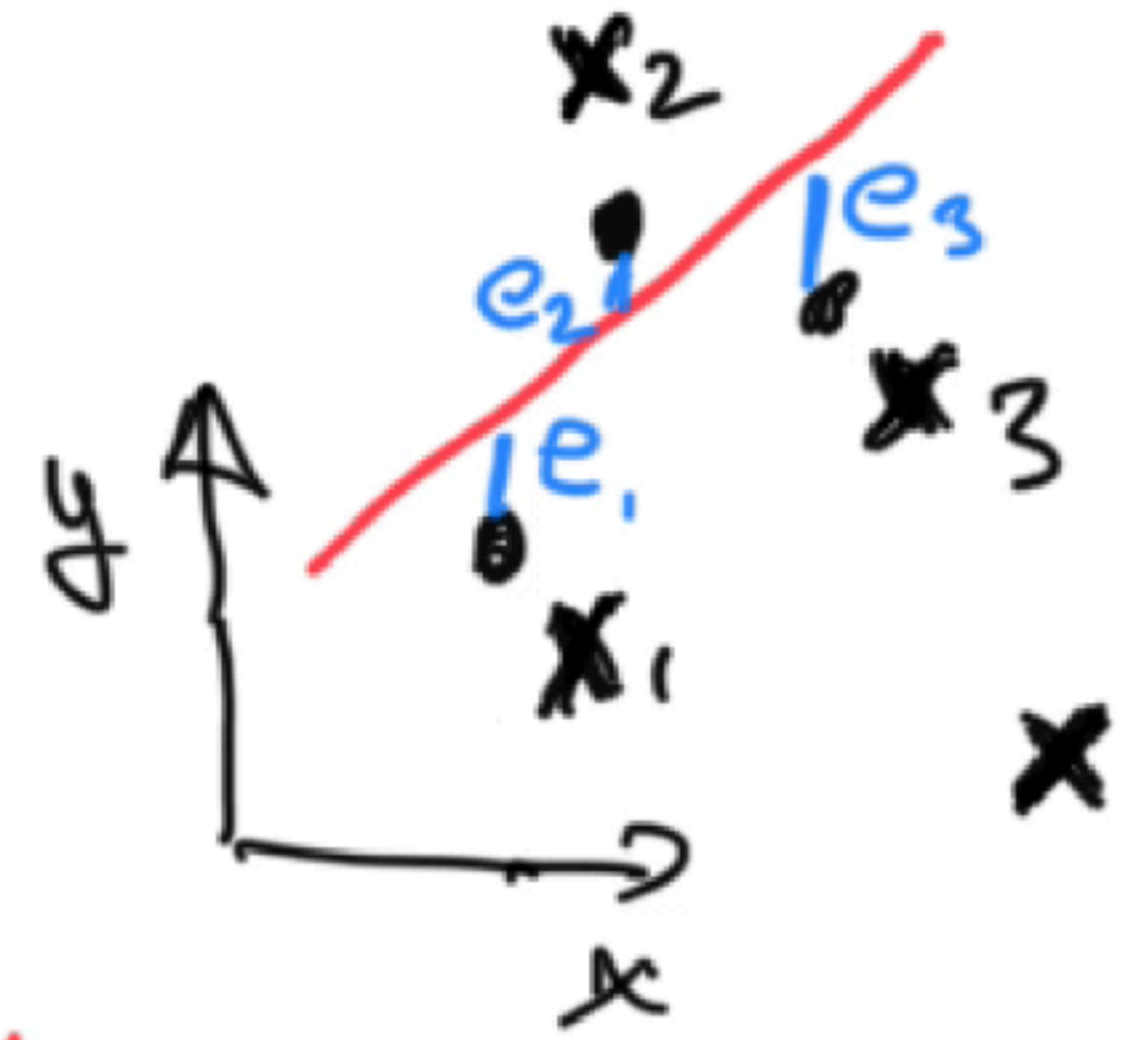
$$\hat{y}_n = b + m x_n \qquad \mathbf{w} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$= \mathbf{x}_n^T \mathbf{w}$$

$$e_n = y_n - \hat{y}_n$$

$$= y_n - \mathbf{x}_n^T \mathbf{w}$$

Derive OLS



$$\mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

$$\hat{y}_n = b + m x_n \qquad \mathbf{w} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$= \mathbf{x}_n^T \mathbf{w}$$

$$e_n = y_n - \hat{y}_n$$

$$= y_n - \mathbf{x}_n^T \mathbf{w}$$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \end{bmatrix}$$

$$\hat{y} = X\mathbf{w} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \end{bmatrix}$$

$$e = y - X\mathbf{w} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \end{bmatrix}$$

RSS · residual sum of squares

$$L(\mathbf{w}) = e^T e = \sum_{n=1}^{N} e_n^2$$

Lecture #3

Linear regression using Ordinary Least Squares

Jason G. Fleischer, Ph.D.
Asst. Teaching Professor
Department of Cognitive Science, UC San Diego

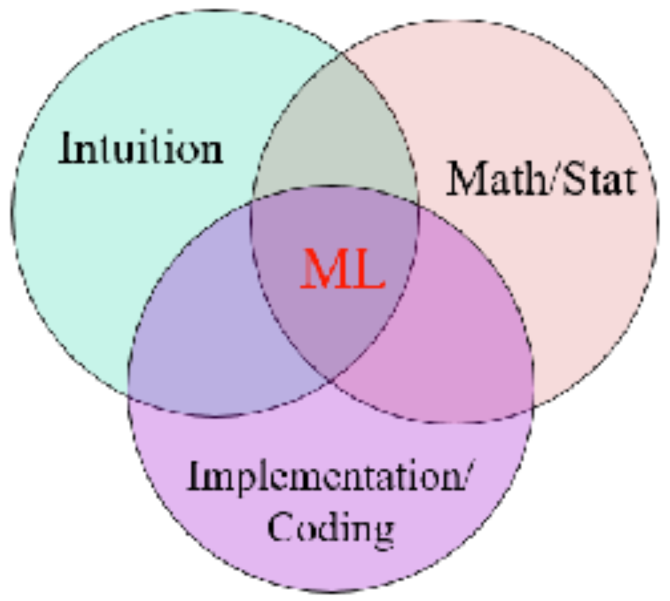
jfleischer@ucsd.edu

@jasongfleischer

https://jgfleischer.com

Logistics

- Please fill out the lecture participation survey every lecture day!
- Discussion section this week
    - Questions and time to work on D1 & A1
- To hand in this coming week
    - D1: Basic NumPy <— Fri 11:59pm on Datahub
    - A1: Basic math   <— Mon 11:59pm on Gradescope
- Give me feedback about the lectures!
    - Email jfleischer@ucsd.edu, subject line COGS118A lecture feedback
    - or the anonymous course feedback form

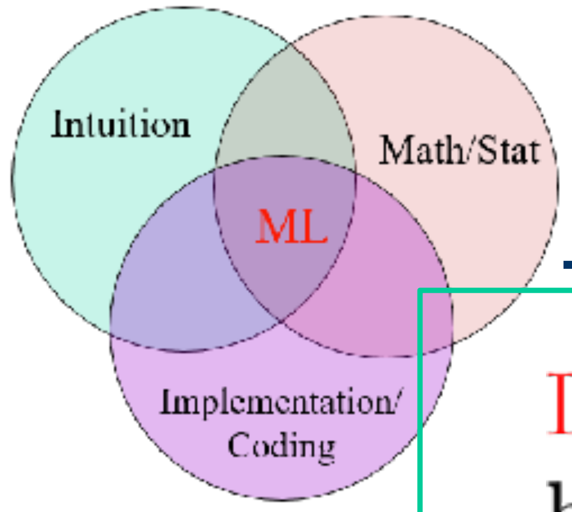


Implementation

NumPy basics

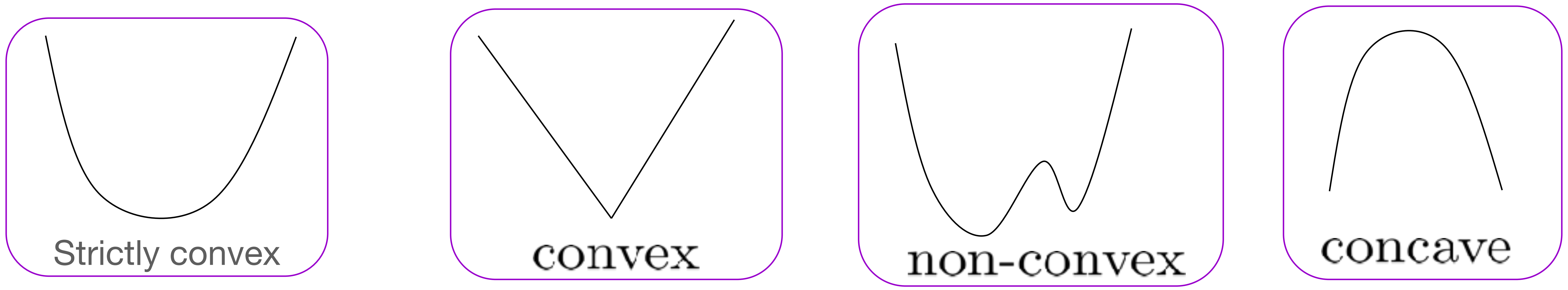
https://colab.research.google.com/github/COGS118A/demo_notebooks/blob/main/lecture_04_pre_numpy.ipynb

https://github.com/COGS118A/demo_notebooks.git



Recap: Convexity

Intuition: Understanding the convexity of the estimation functions allows us to better design the learning algorithms and allows us to judge the quality (global vs. local optimal) of the learned models.



Math:

$$\forall w_0, w_1, a \in [0, 1]$$

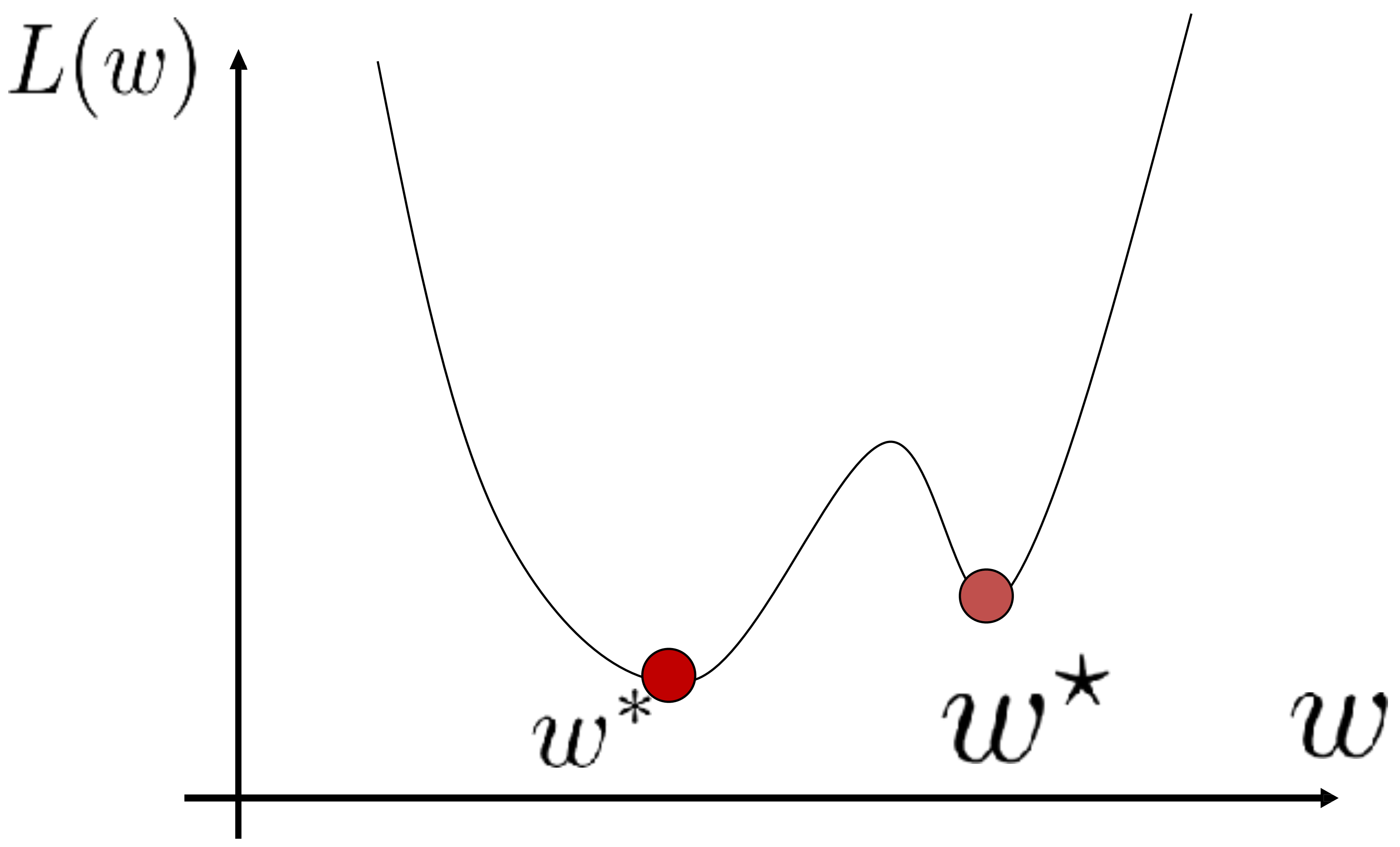
$$aL(w_0) + (1-a)L(w_1) \geq L(aw_0 + (1-a)w_1)$$

or

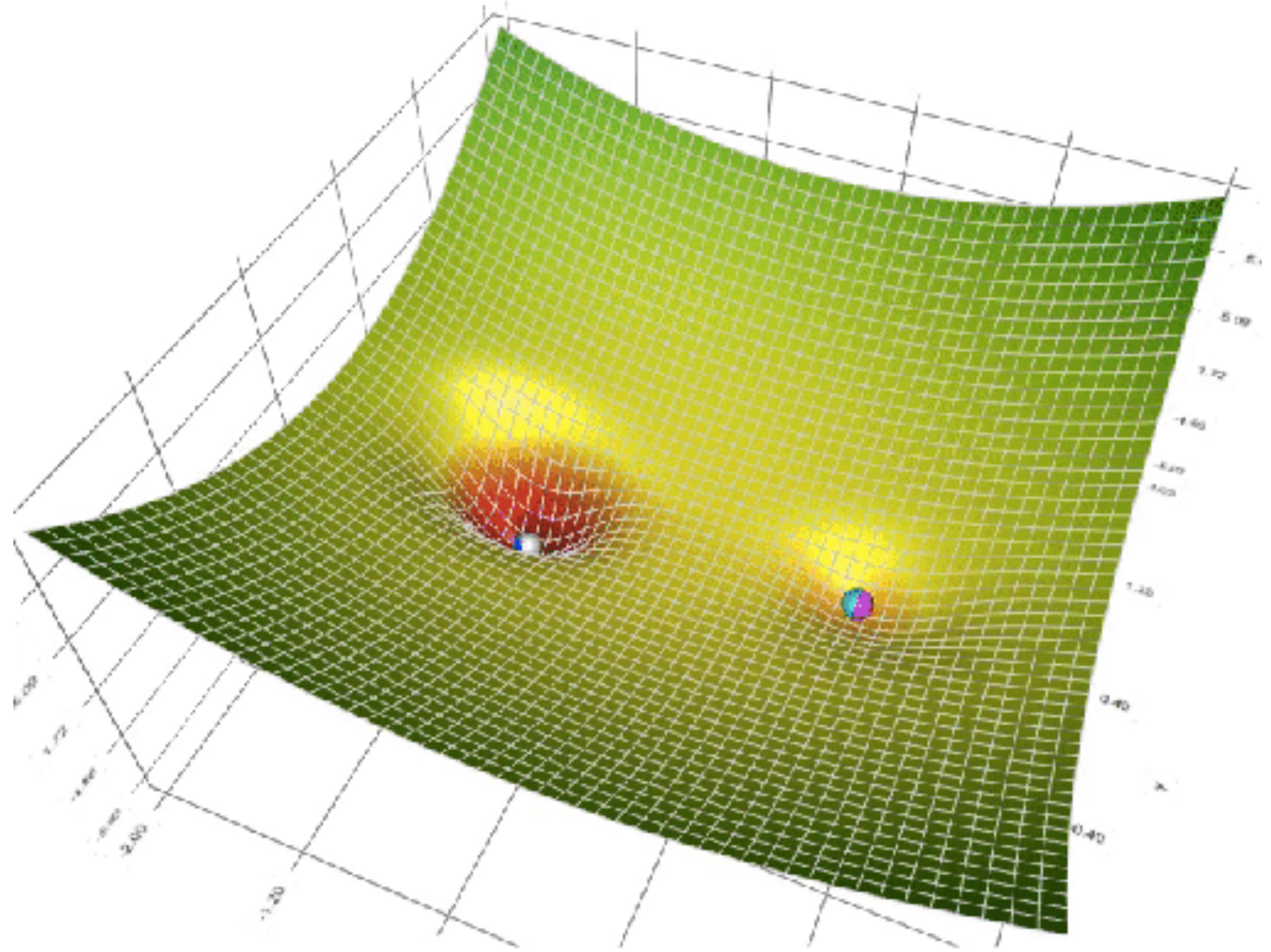
Alternatively (for differentiable function):

$$L(w_1) \geq L(w_0) + < \nabla L(w_0), w_1 - w_0 >$$

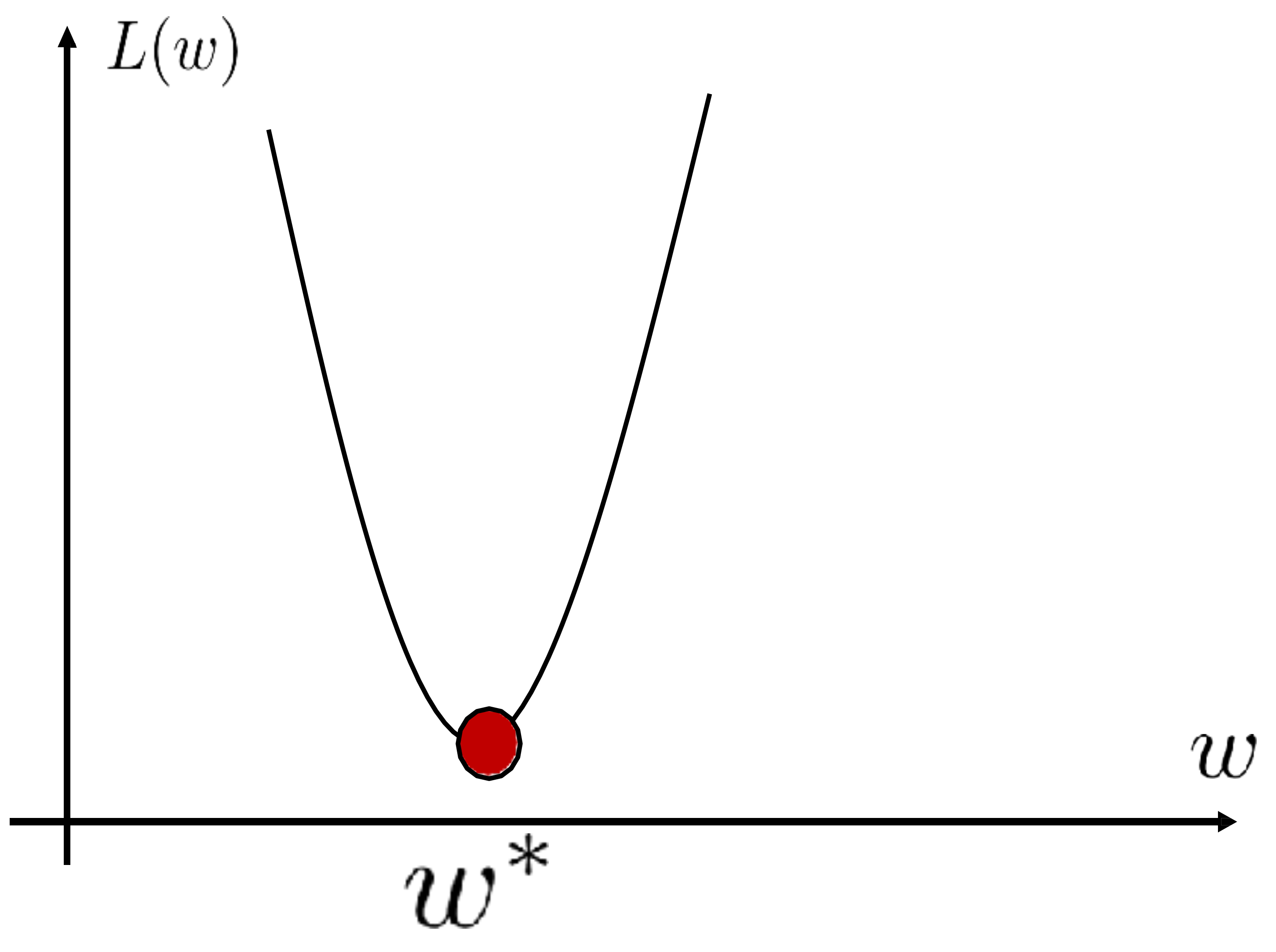
Non-convex functions that are differentiable



Compute $\frac{\partial L(w)}{\partial w}$ to solve the problem iteratively through gradient descent



Convex functions that are differentiable



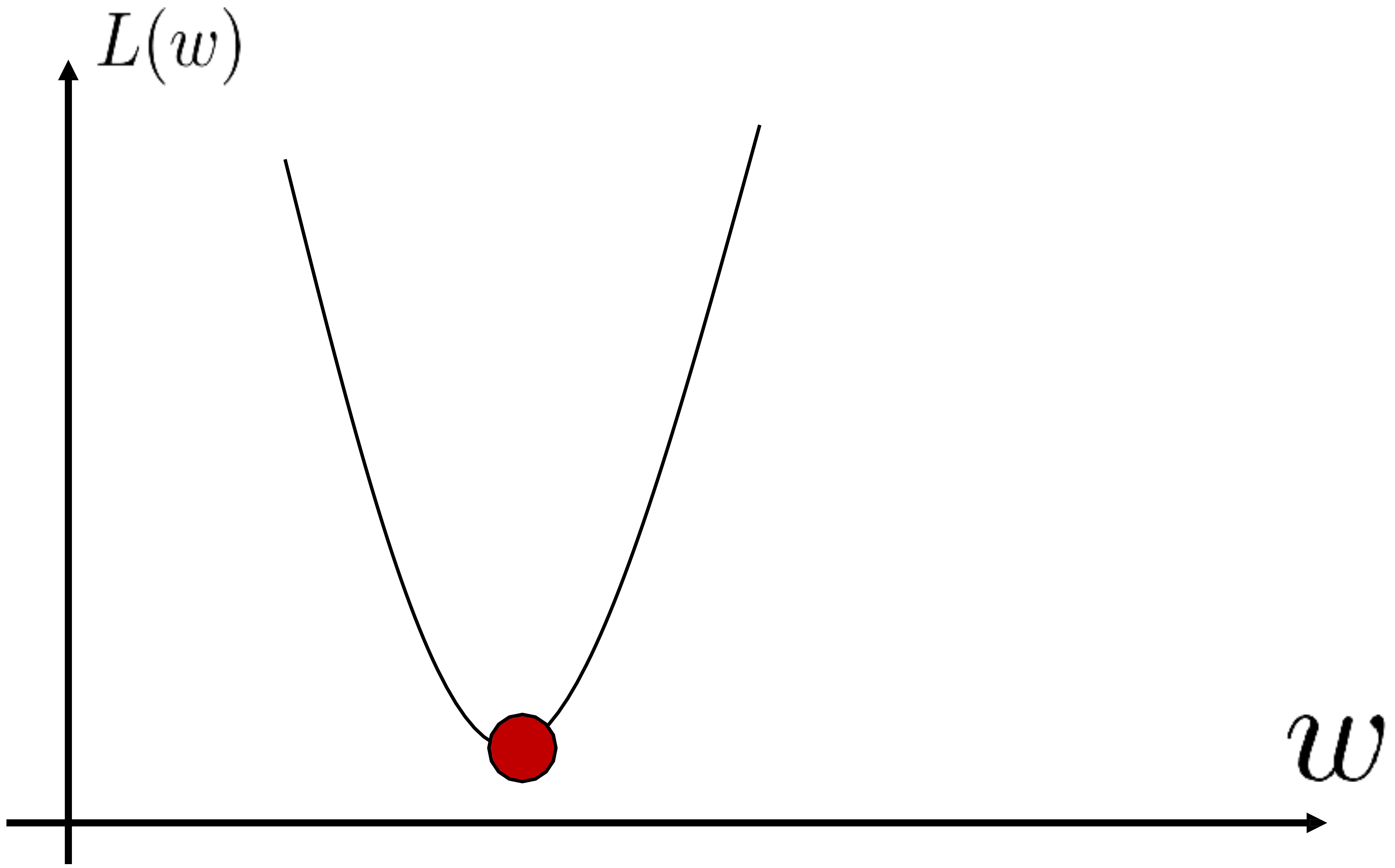
$$\frac{\partial L(w)}{\partial w}\Big|_{w^*} = 0$$

An analytical (closed form) solution exists!

$w^* = q(X, Y)$ where X and Y consists of your training data with the corresponding ground-truth labels.

You get your model in 1 shot, no iteration!

Convex function: differentiable



$$w^* = \arg min_\theta L(w)$$

1. (Convex) Function
2. Set Derivative to 0
3. Solve for w

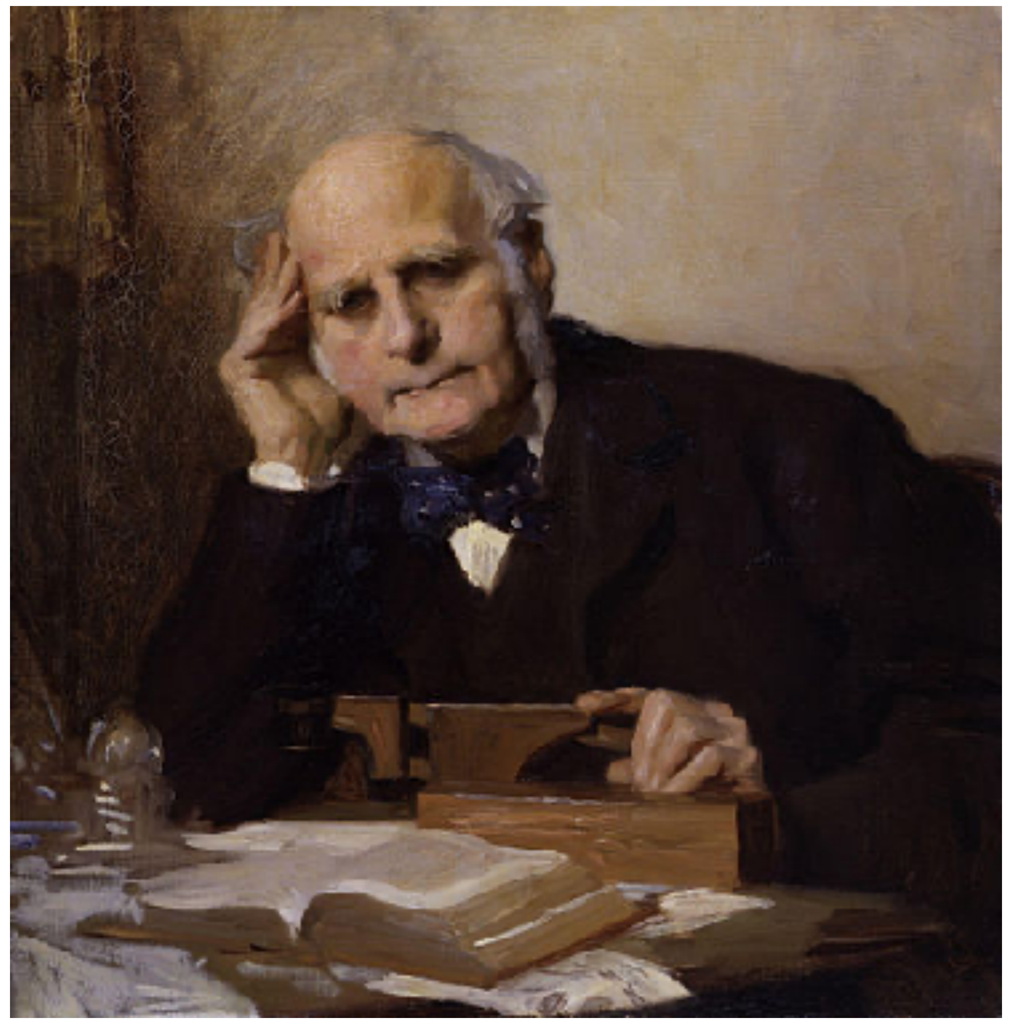
$$L(w) = (w-3)^2 + 4$$

$$\frac{dL(w)}{dw} = 2 \times (w-3) \qquad \frac{dL(w)}{dw} = 0$$

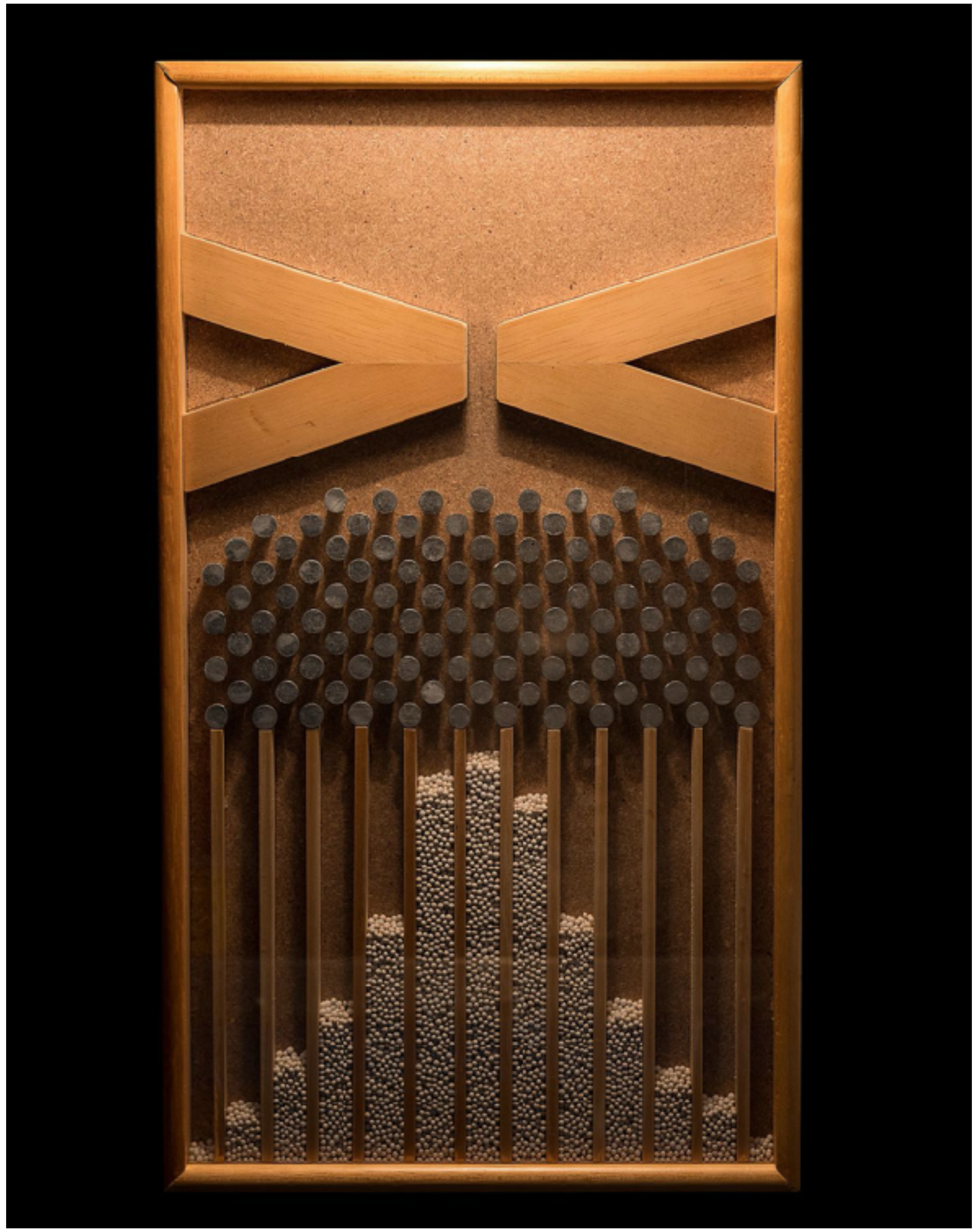
$$2 \times (w-3) = 0 \rightarrow w = 3$$

Regression

Regression (to the mean) - racists give us an important tool

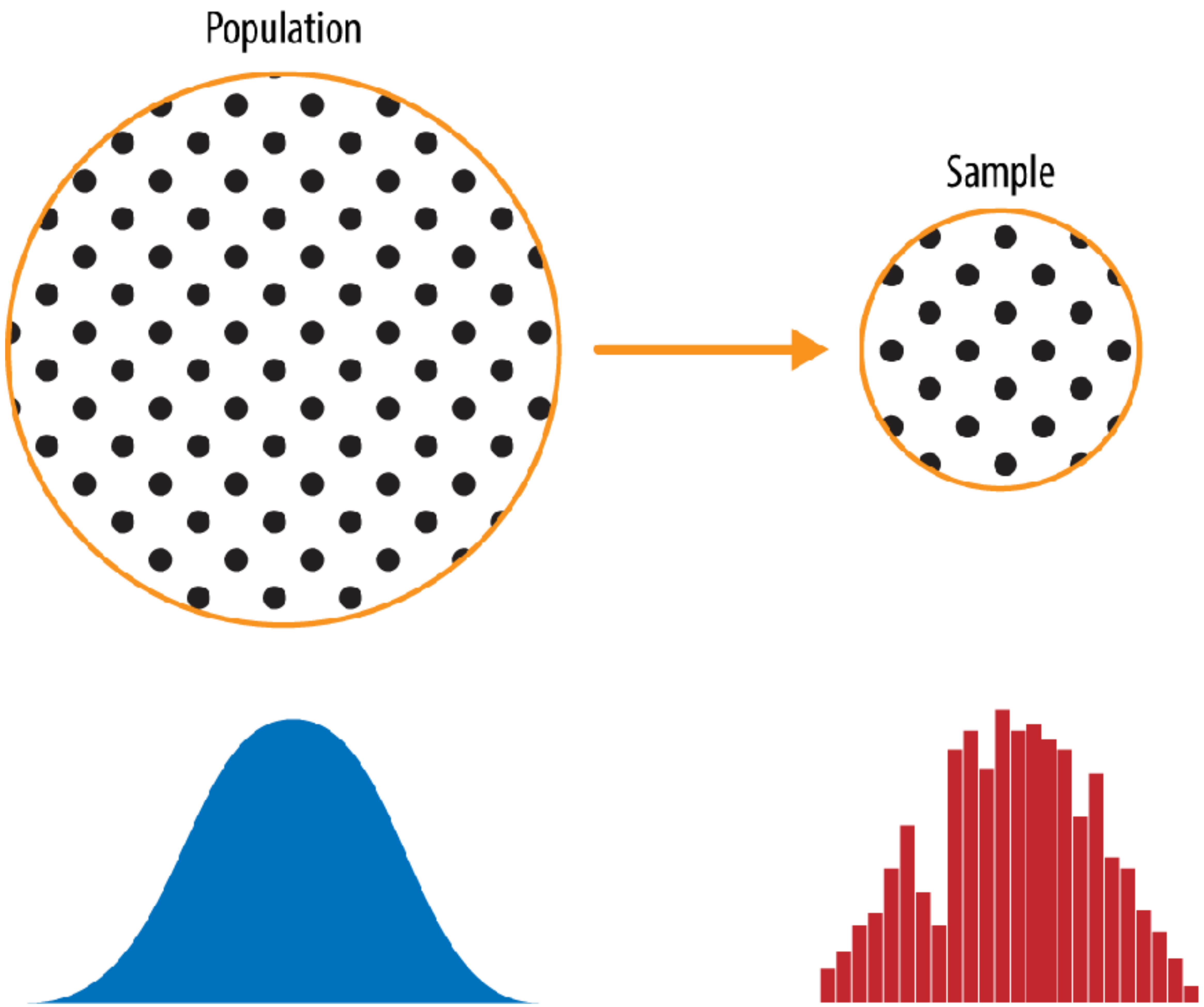


Sir Francis Galton,
polymath & eugenicist



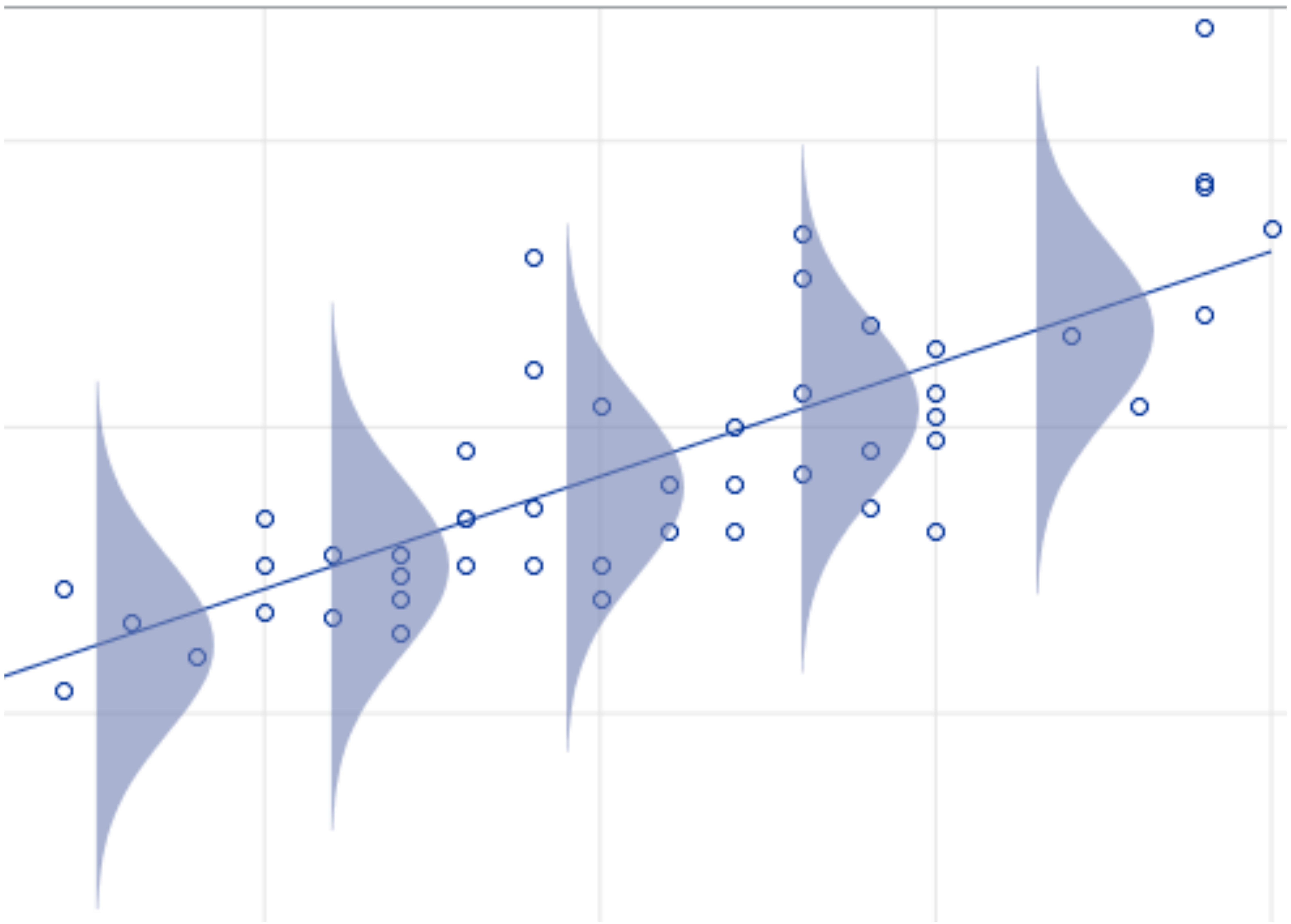
Regression

The intuition



Regression

The intuition



Regression

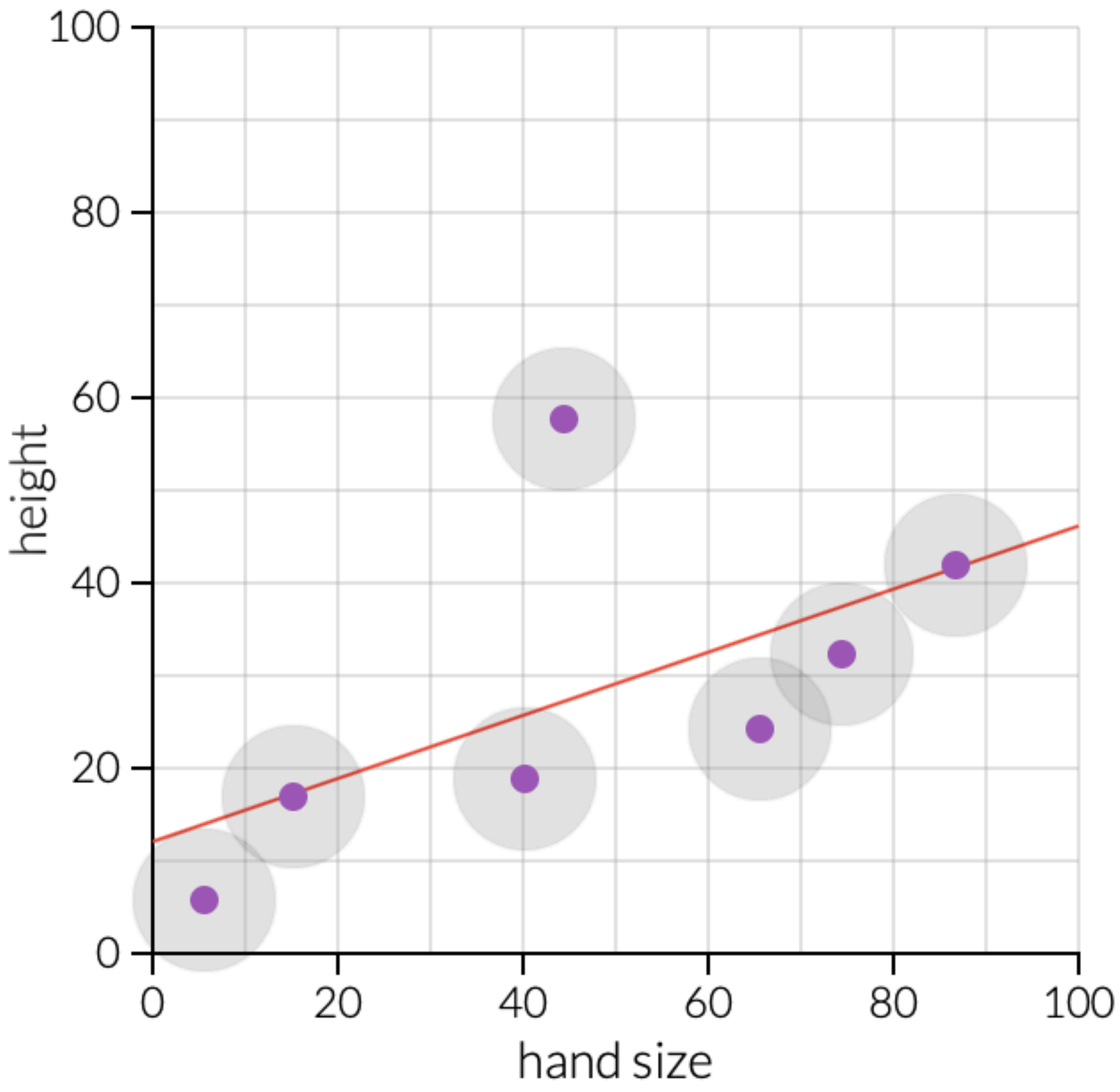
The intuition

$$y_i = f(x_i; W) + \epsilon$$

And if your sample is good enough the noise will cancel out

“regression to the mean”

https://setosa.io/ev/ordinary-least-squares-regression/



Beta 1 - The y-intercept of the regression line.

$$\underbrace{12.07}_{\text{Beta 1}} + \underbrace{0.34}_{\text{Beta 2}} * \text{hand size} = \text{height}$$

Beta 2 - The slope of the regression line.

Derive OLS

$$\mathbf{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$$

$$\hat{y}_n = b + m x_n = \mathbf{x}_n^T \mathbf{w}$$

$$\mathbf{w} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$e_n = y_n - \hat{y}_n = y_n - \mathbf{x}_n^T \mathbf{w}$$

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \end{bmatrix}$$

$$\hat{\mathbf{y}} = X\mathbf{w} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \end{bmatrix}$$

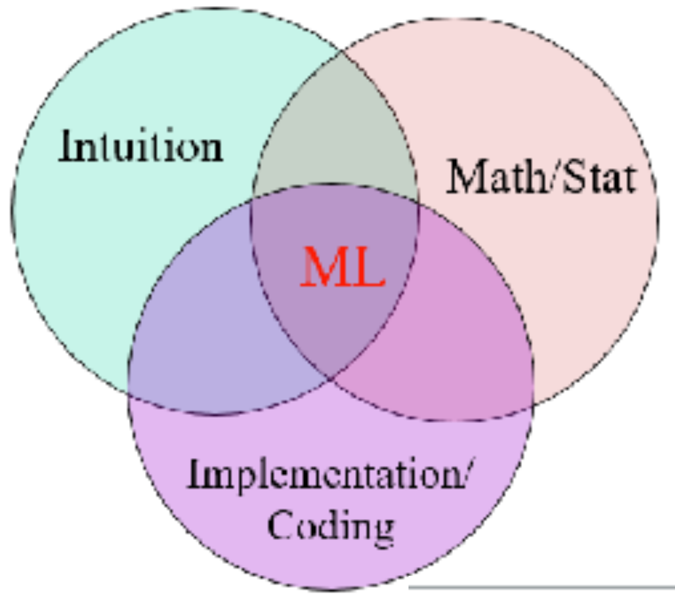
$$\mathbf{e} = \mathbf{y} - X\mathbf{w} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \end{bmatrix}$$

RSS - residual sum of squares

$$L(\mathbf{w}) = \mathbf{e}^T \mathbf{e} = \sum_{n=1}^{N} e_n^2$$

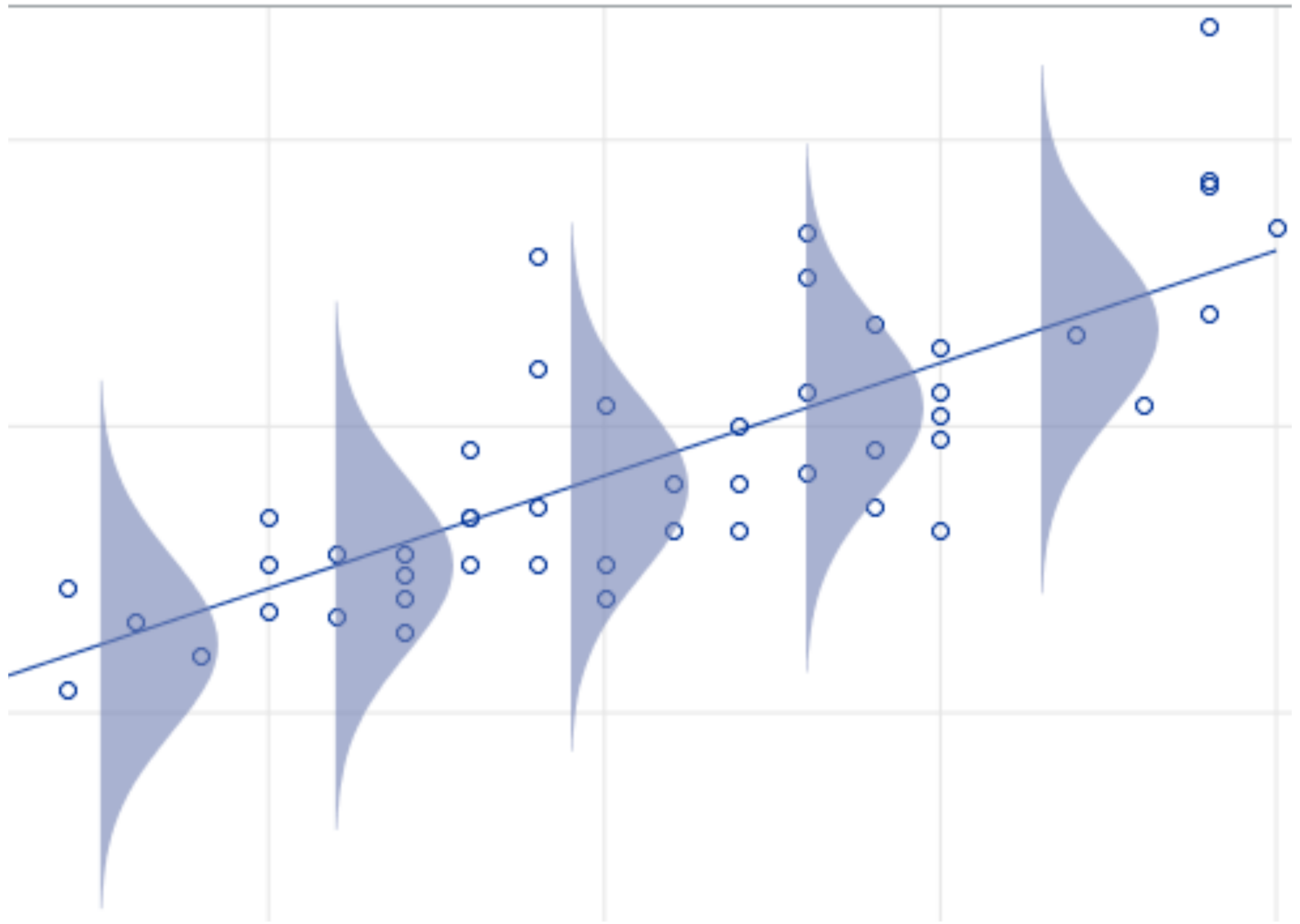
Derive OLS

Derive OLS

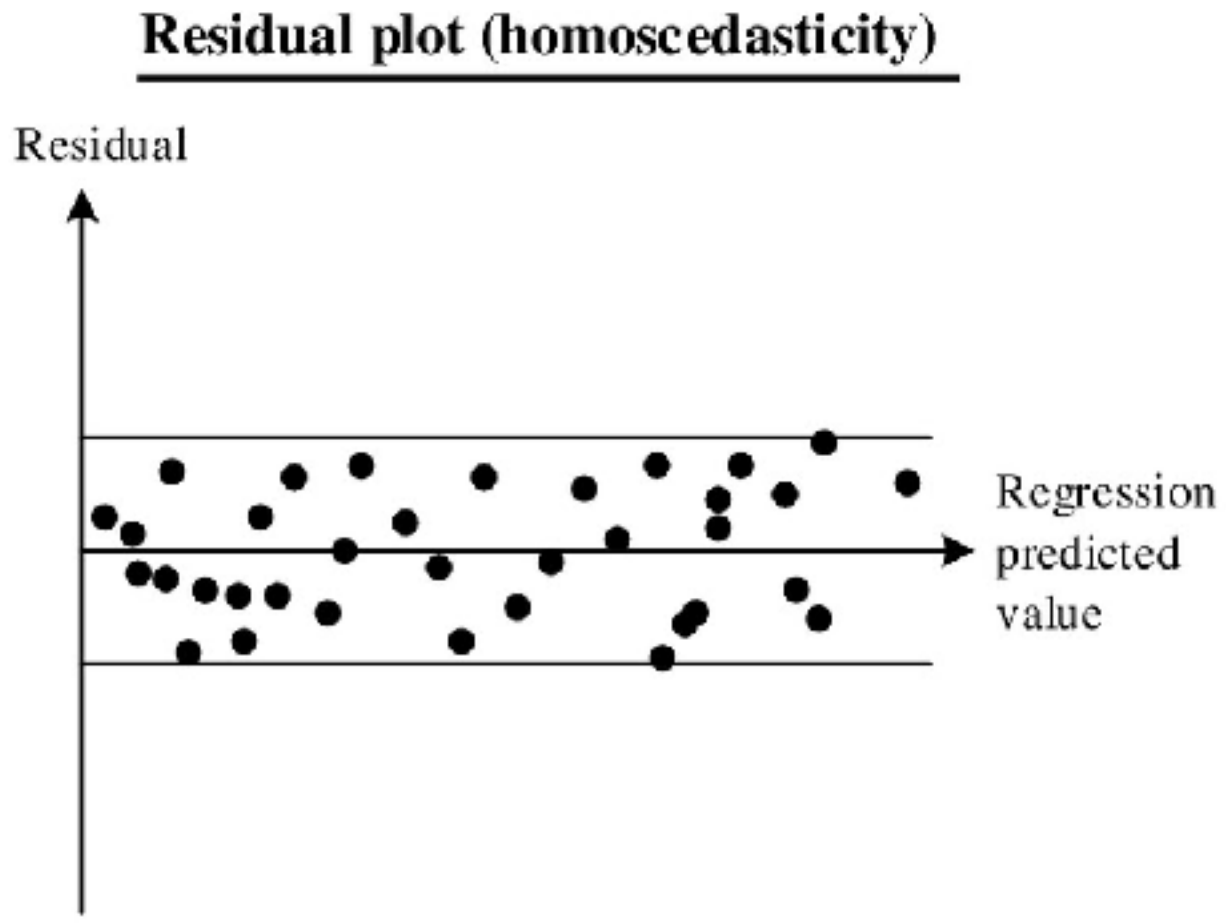


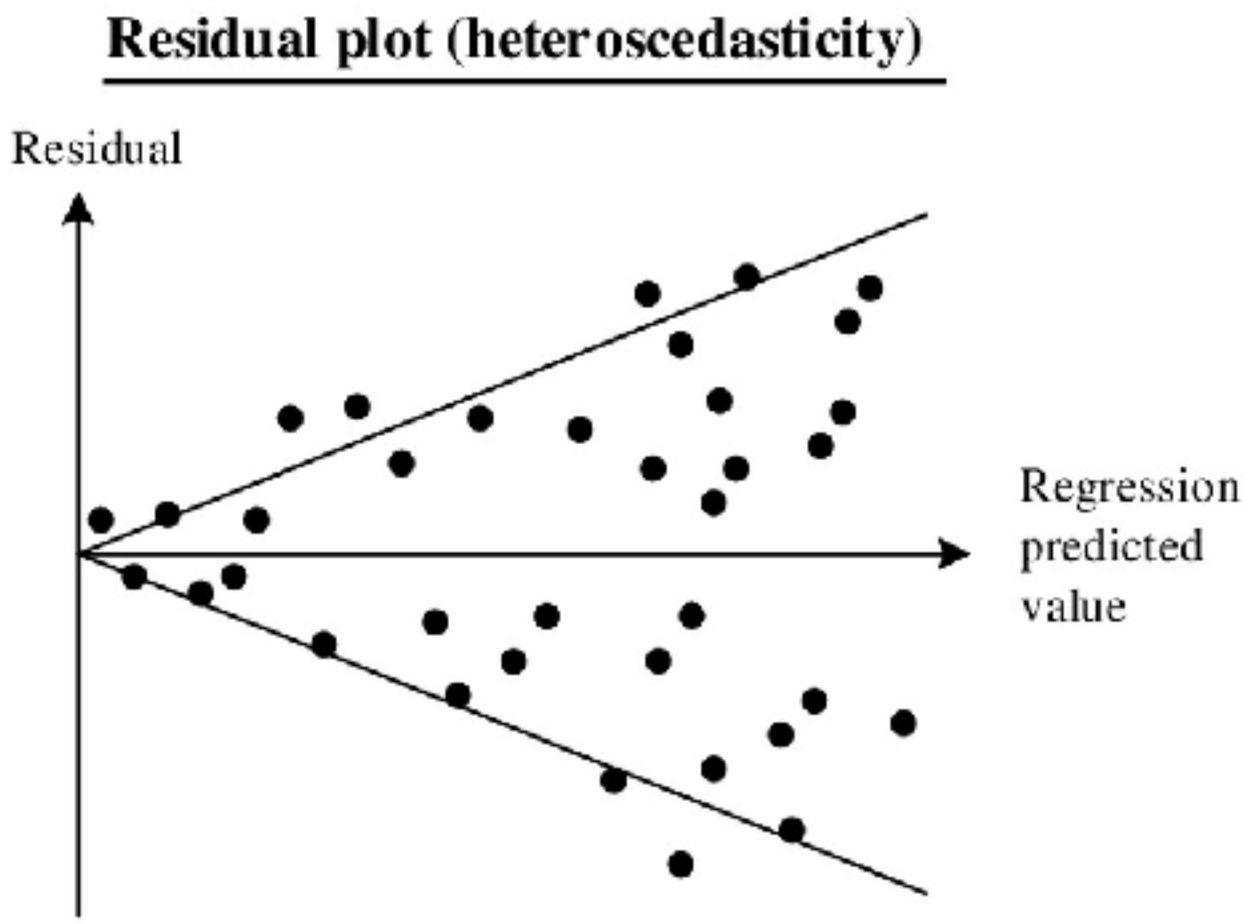
ASS-U-ME D properties

Linear Regression using Ordinary Least Squares



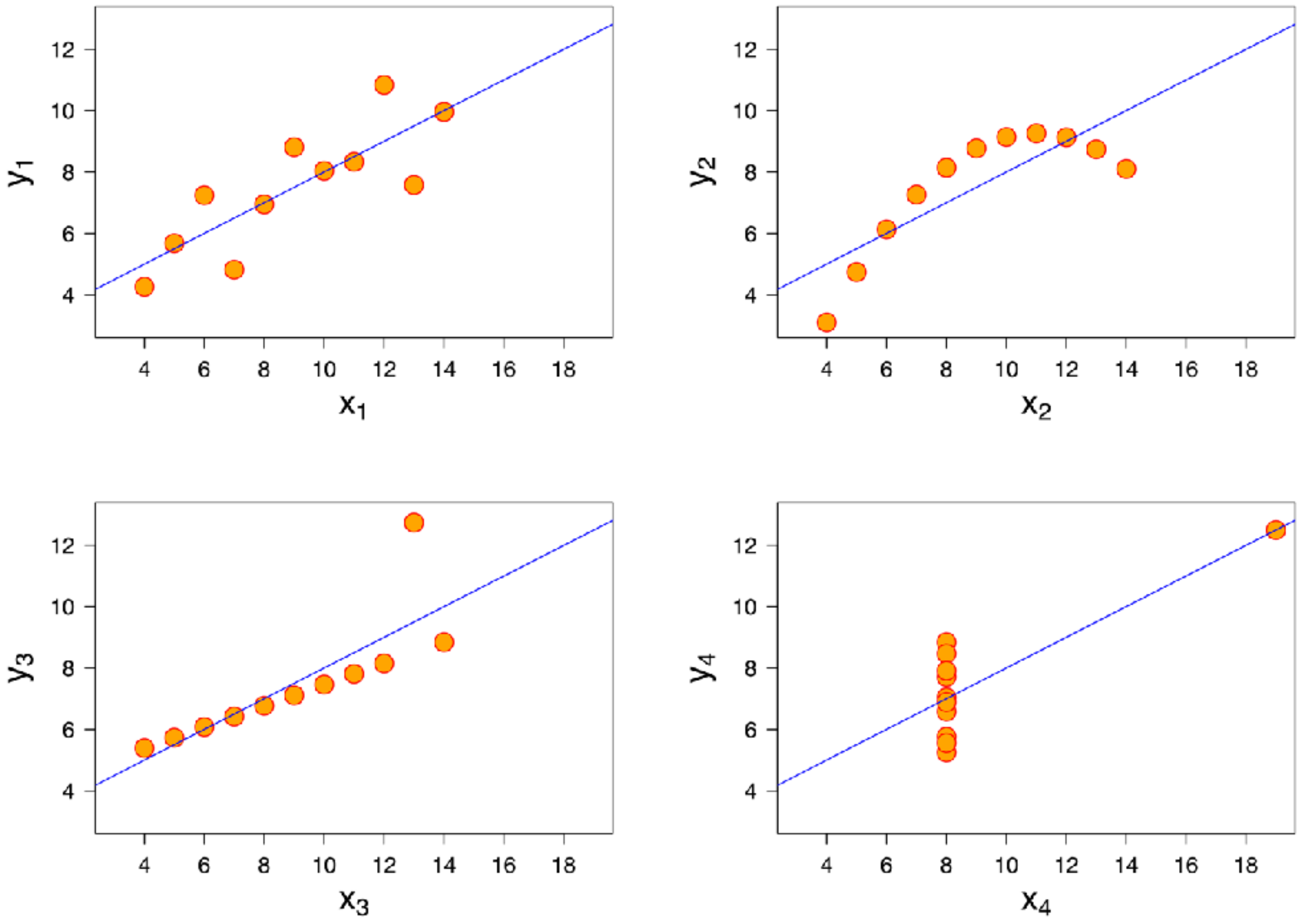
A1. The linear regression model is “linear in parameters.”
A2. There is a random sampling of observations.
A3. The conditional mean should be zero.
A4. No multi-collinearity (or perfect collinearity).
A5. Homoscedasticity
A6: Error terms should be normally distributed





Anscombe's quartet

An example of what happens when data violates assumptions of OLS



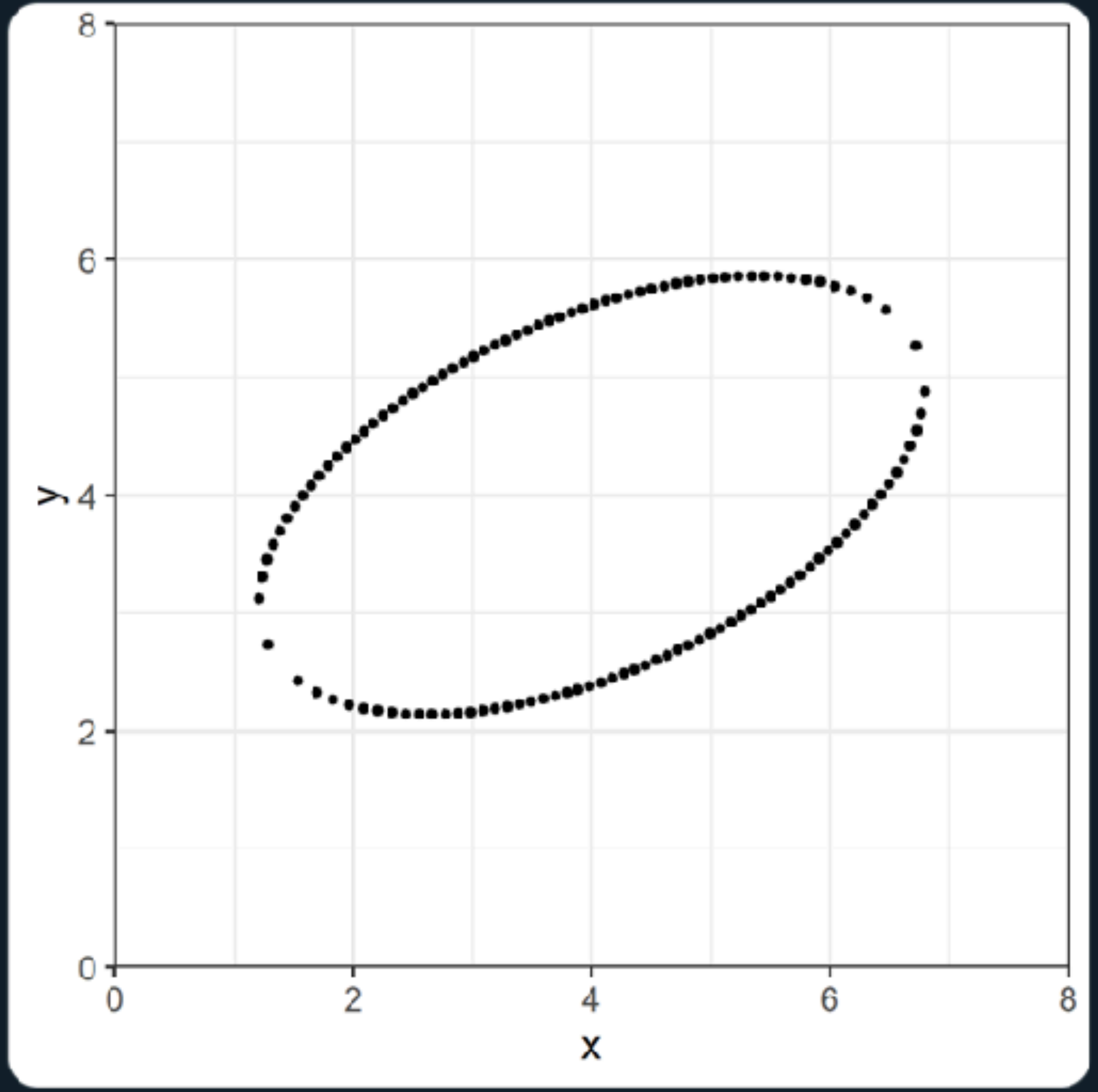
OLS isn't the only game in town

Noah Haber
@NoahHaber

STATS QUIZ!

I have the datapoints below. Nothing hidden, no tricks, just a bunch of data making roughly an ellipse.

In your head, draw what you think the ordinary least squares line (i.e. good ol' y= mx+b) line looks like for these data.



7:05 AM · Sep 12, 2020 · Twitter Web App

521 Retweets **80** Quote Tweets **1.8K** Likes

---

10 6 70

Noah Haber
@NoahHaber

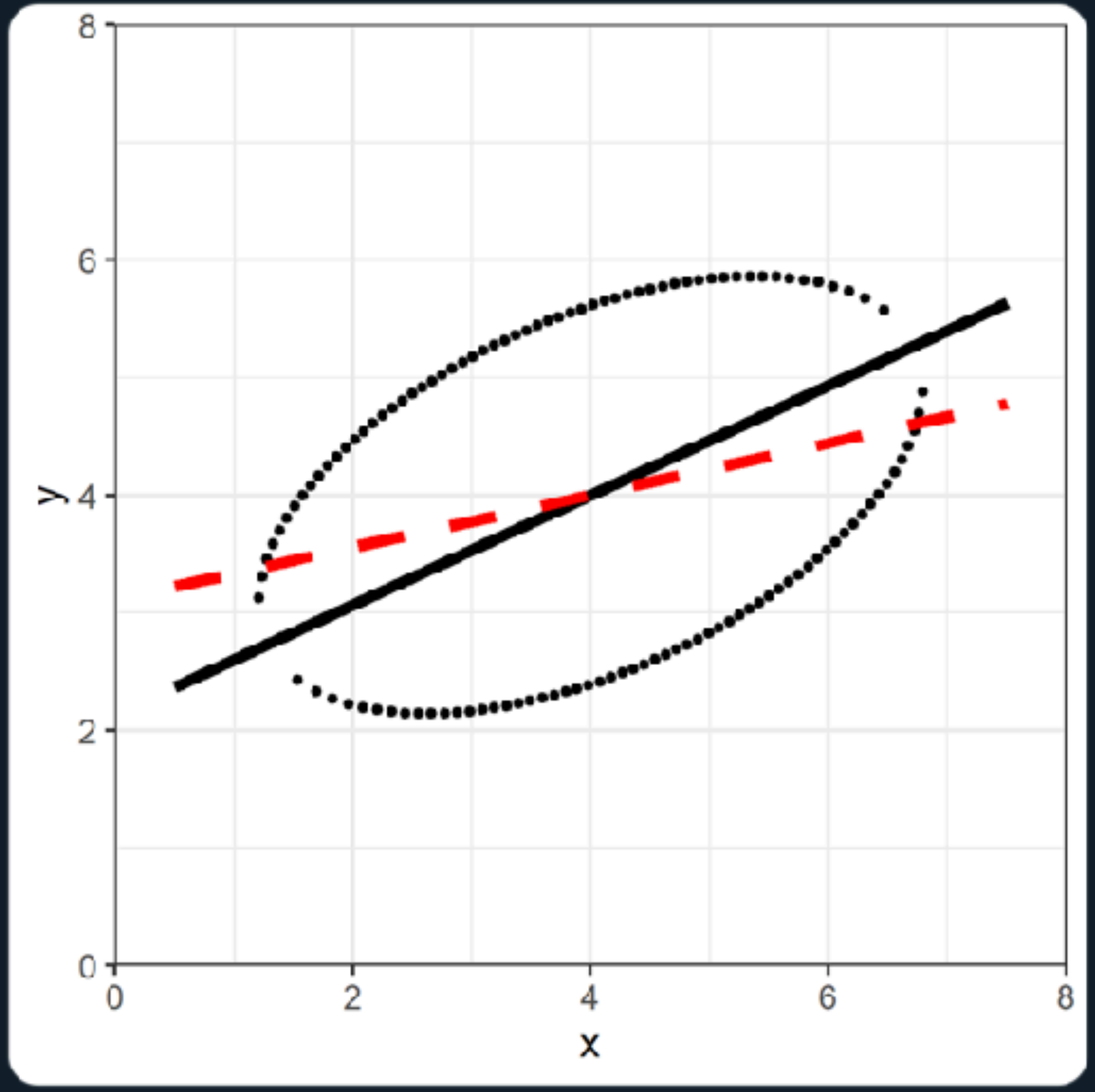
Replying to @NoahHaber

Seems "obvious" right?

Except that's not the OLS line.

The red dashed line is the OLS line.

What's going on here?



7:12 AM · Sep 12, 2020 · Twitter Web App

Noah Haber @NoahHaber · Sep 12, 2020
Replying to @NoahHaber
Ordinary least squares is the line which minimizes the VERTICAL squared distances.

Take a vertical line on the data, and draw a midpoint between the top and bottom of the ellipse. Do that for all points.

OLS is the line that minimizes the sum of those (squared) distances.

1 9 120

Noah Haber @NoahHaber · Sep 12, 2020
In your head, you were probably trying to find the line that minimizes the TOTAL distance, not just the distance on the vertical.

That's the ORTHOGONAL (or total) least squares line. Totally natural, and it isn't wrong. Makes you wonder why we care only about the vertical, eh?

2 4 117

Noah Haber @NoahHaber · Sep 12, 2020
RABBIT HOLE TIME!

What we are implying here is that Y is the DOMINANT axis.

That's great and fine if we are trying to predict Y from X, or estimate how much X causes Y.

But you've heard the phrase "it's just association" right?

6 4 88

Noah Haber @NoahHaber · Sep 12, 2020
If all we care about is that X and Y move together, we shouldn't have a dominant axis at all! We'd probably want to use something more like TLS.

In other words, the form of our estimates pushes us toward prediction/causation thinking, even if we really don't want to.

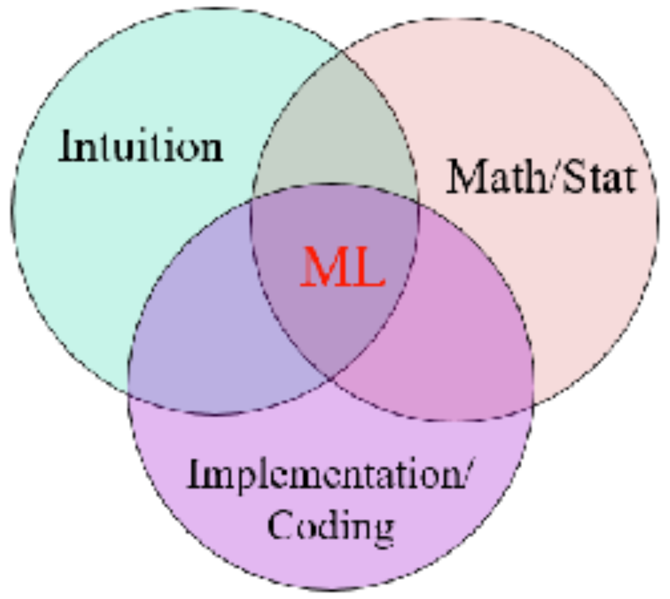
4 9 153

Noah Haber @NoahHaber · Sep 12, 2020
Small clarification OLS is (was a bit unclear):

The OLS line is the line in which the sum of the squared vertical distances from the line and the datapoints are minimized.

The midpoint bit is an illustration for intuition, not strictly how OLS actually "works"

2 2 43



Implementation

Linear Regression using Ordinary Least Squares

https://colab.research.google.com/github/COGS118A/demo_notebooks/blob/main/lecture_04_linear_regression.ipynb

https://github.com/COGS118A/demo_notebooks.git