Machine Learning with Python

MTH786U/P 2020/21

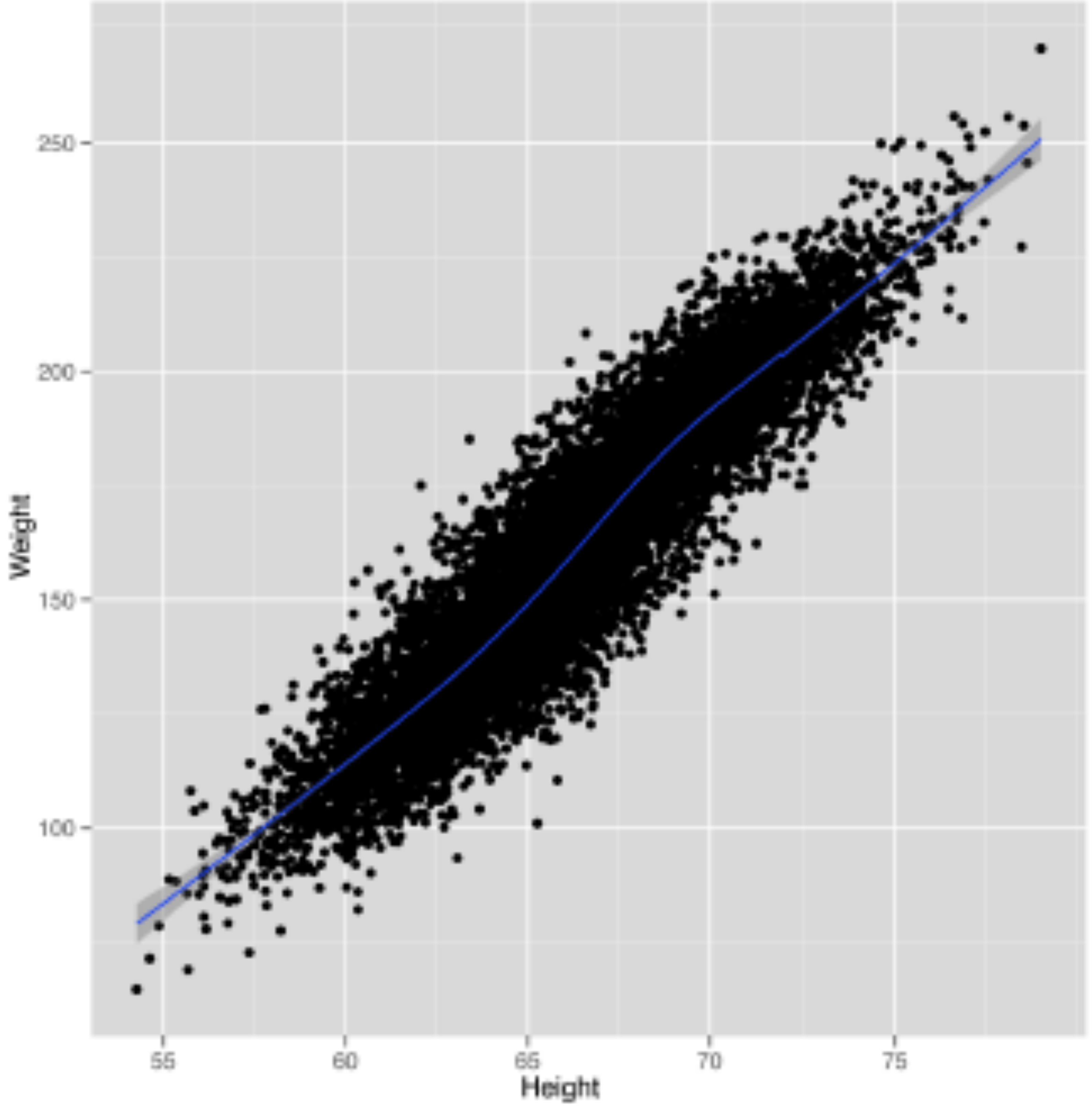
Estimating the height of a person

Mihail Poplavskyi, Queen Mary University of London (QMUL)

m.poplavskyi@qmul.ac.uk

What is regression?

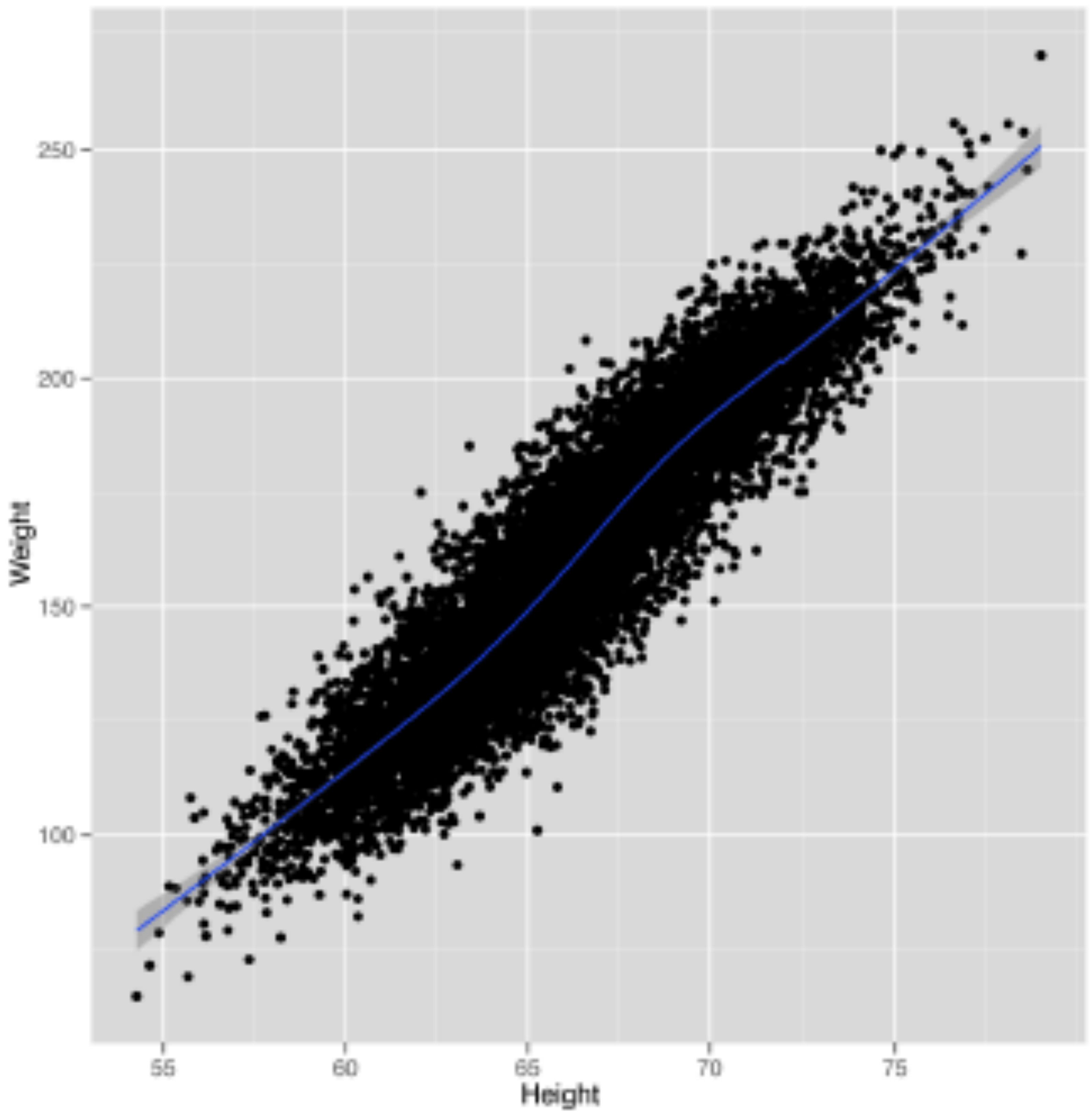
Examples:



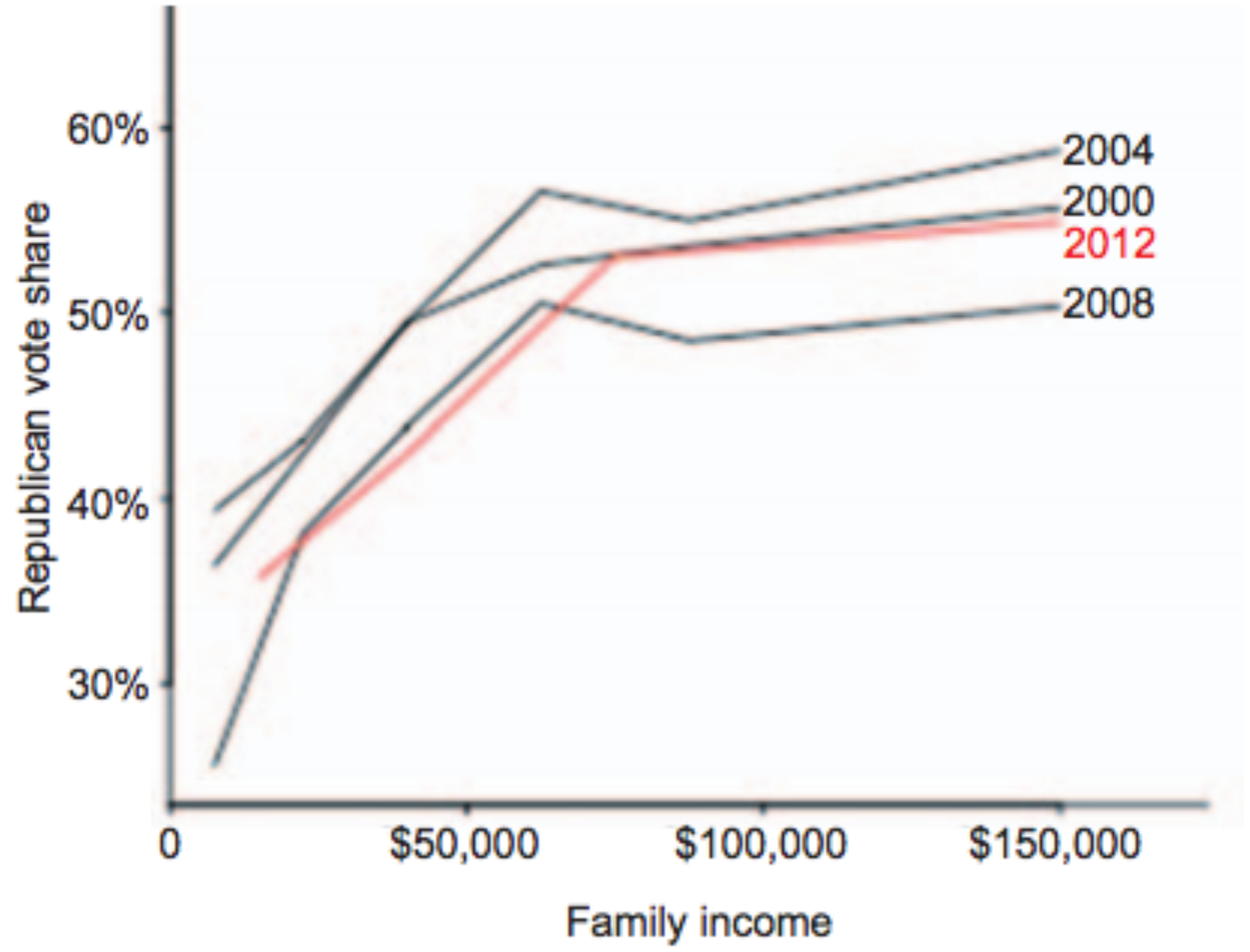
From “Machine Learning for Hackers” by Conway & White

What is regression?

Examples:



From “Machine Learning for Hackers” by Conway & White



From Avi Feller et al. 2013

What is regression?

Mathematical formulation:

Given input/output pairs $\{(x_i, y_i)\}_{i=1}^{s}$ find function f with

$$y_i \approx f(x_i) \qquad \forall i \in \{1, \ldots, s\}$$

Example: linear regression

$$y_i \approx f(x_i) \qquad \forall i \in \{1, \ldots, s\}$$

Example: linear regression

$y_i \approx f(x_i) \qquad \forall i \in \{1,\ldots,s\}$

How do we parametrise f ?

Example: linear regression

$y_i \approx f(x_i)$ $\forall i \in \{1, \ldots, s\}$ How do we parametrise f ?

Example: $$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

Example: linear regression

$y_i \approx f(x_i) \qquad \forall i \in \{1,\ldots,s\}$

How do we parametrise f ?

Example:

$$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

Affine linear transformation of vector $x = (x_1, \ldots, x_d)$ with weights $w \in \mathbb{R}^{d+1}$

Example: linear regression

$y_i \approx f(x_i) \qquad \forall i \in \{1, \ldots, s\}$ How do we parametrise f ?

Example:

$$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j$$

Affine linear transformation of vector $x = (x_1, \ldots, x_d)$ with weights $w \in \mathbb{R}^{d+1}$

Note that $i \neq j$:

$$f(x_i) = w_0 + \sum_{j=1}^{d} w_j \, x_{ij}$$

Cost function

Notation: $f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle$ with $x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$

Cost function

Notation: $$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j \;=\; \langle w, x \rangle$$ with $$x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

(slight abuse of x-notation)

Cost function

Notation: $f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle = w^\top x = x^\top w$ with $x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$

(slight abuse of x-notation)

Cost function

Notation: $$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle = w^\top x = x^\top w \quad \text{with} \quad x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

How do we choose w such that $y_i \approx f(x_i)$?

(slight abuse of x-notation)

Cost function

Notation: $f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle = w^\top x = x^\top w$ with $x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$

How do we choose w such that $y_i \approx f(x_i)$?

(slight abuse of x-notation)

Imagine $s = 3$ and $d = 2$:

$$w_0 + x_{11}w_1 + x_{12}w_2 = y_1$$
$$w_0 + x_{21}w_1 + x_{22}w_2 = y_2$$
$$w_0 + x_{31}w_1 + x_{32}w_2 = y_3$$

Cost function

Notation: $f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle = w^\top x = x^\top w$ with $x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$

How do we choose w such that $y_i \approx f(x_i)$?

(slight abuse of x-notation)

Imagine $s = 3$ and $d = 2$:

$$\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

Cost function

Notation: $$f(x) = w_0 + \sum_{j=1}^{d} w_j x_j = \langle w, x \rangle = w^\top x = x^\top w \quad \text{with} \quad x := \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

How do we choose w such that $y_i \approx f(x_i)$?

(slight abuse of x-notation)

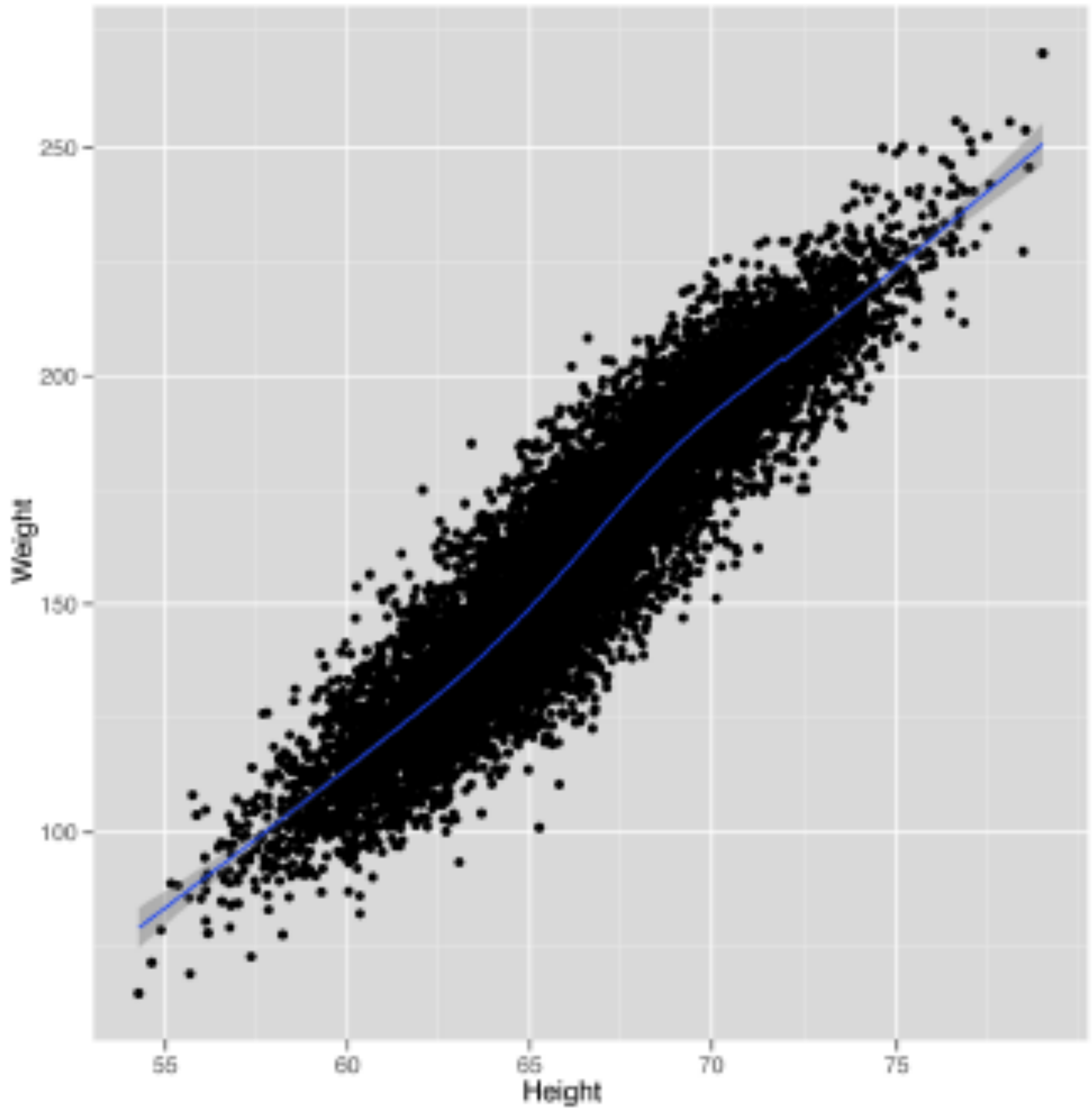
Imagine $s = 3$ and $d = 2$:

$$\begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$

This system of linear equations has a unique solution if…?

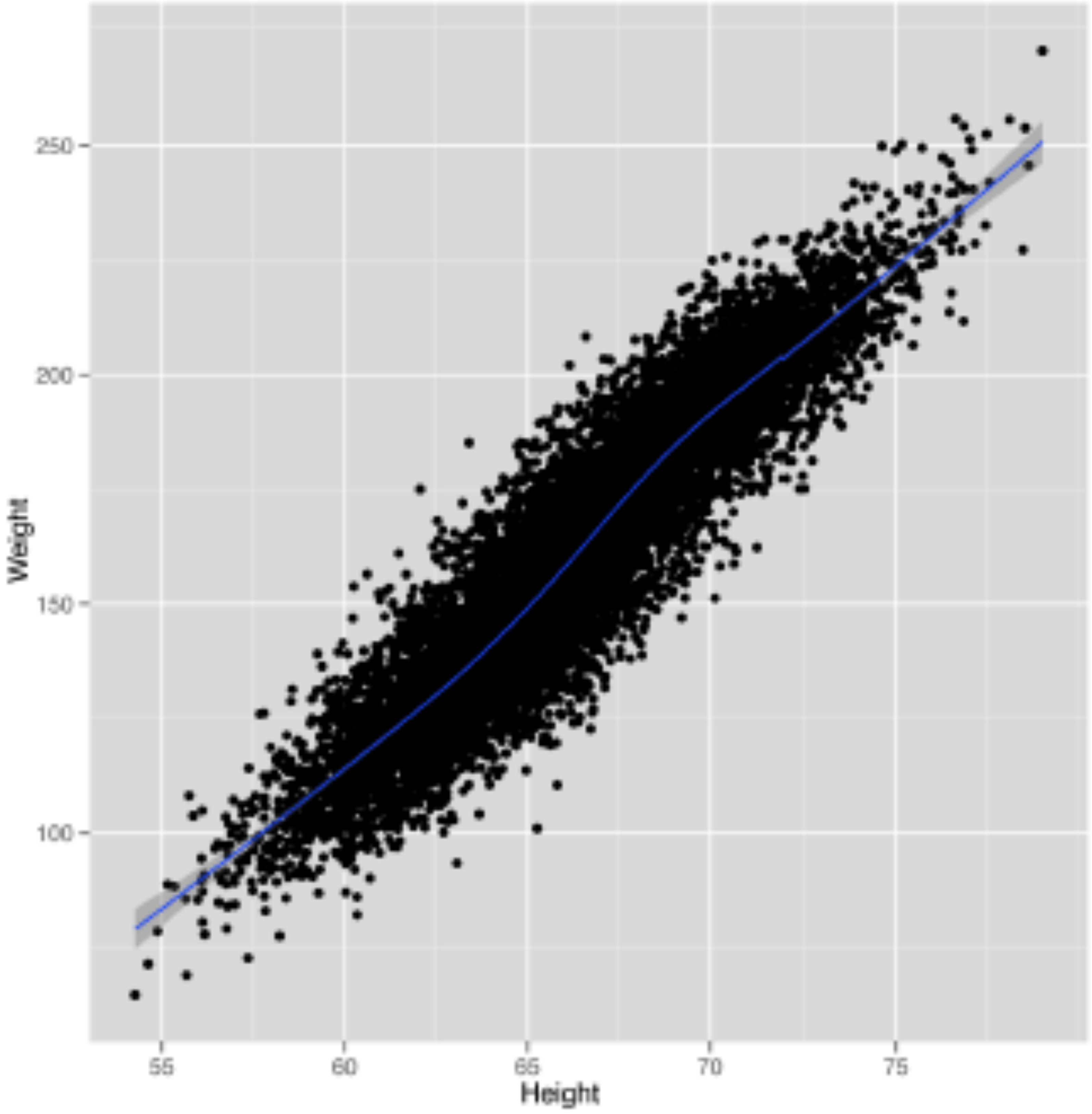
But is it realistic to assume $s = d + 1$?

But is it realistic to assume $s = d + 1$?



m.poplavskyi@qmul.ac.uk

But is it realistic to assume $s = d + 1$?



$$s \gg d + 1 = 2$$

m.poplavskyi@qmul.ac.uk

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathrm{MSE}(w) := \frac{1}{2s}\sum_{i=1}^{s} |f(x_i) - y_i|^2$$

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathsf{MSE}(w) := \frac{1}{2s}\sum_{i=1}^{s} |f(x_i) - y_i|^2$$

Obtain 'optimal' parameters $\hat{w}$ by minimising MSE:

$$\hat{w} = \arg\min_{w} \mathsf{MSE}(w)$$

Instead we need to find an approximation that is optimal in some sense

Example: Mean-Square Error (MSE)

$$\mathrm{MSE}(w) := \frac{1}{2s} \sum_{i=1}^{s} |f(x_i) - y_i|^2$$

Obtain 'optimal' parameters $\hat{w}$ by minimising MSE:

$$\hat{w} = \arg\min_{w} \mathrm{MSE}(w)$$

How can we do this?

How do we compute $\hat{w}$?

Example: $f(x_i) = w_0 \qquad \forall i \in \{1, \ldots, s\}$

How do we compute $\hat{w}$?

Example: $f(x_i) = w_0 \qquad \forall i \in \{1,\ldots,s\}$

MSE cost function: $\mathrm{MSE}(w_0) := \frac{1}{2s}\sum_{i=1}^{s} |w_0 - y_i|^2$

How do we compute $\hat{w}$?

Example: $f(x_i) = w_0 \qquad \forall i \in \{1,\ldots,s\}$

MSE cost function: $\mathrm{MSE}(w_0) := \frac{1}{2s}\sum_{i=1}^{s} |w_0 - y_i|^2$

We do what we did in school: we compute the derivative and set it to zero:

$$\nabla \mathrm{MSE}(\hat{w}_0) = \mathrm{MSE}'(\hat{w}_0) = \frac{1}{s}\sum_{i=1}^{s}(\hat{w}_0 - y_i) \overset{!}{=} 0$$

How do we compute $\hat{w}$?

Example: $f(x_i) = w_0 \qquad \forall i \in \{1, \dots, s\}$

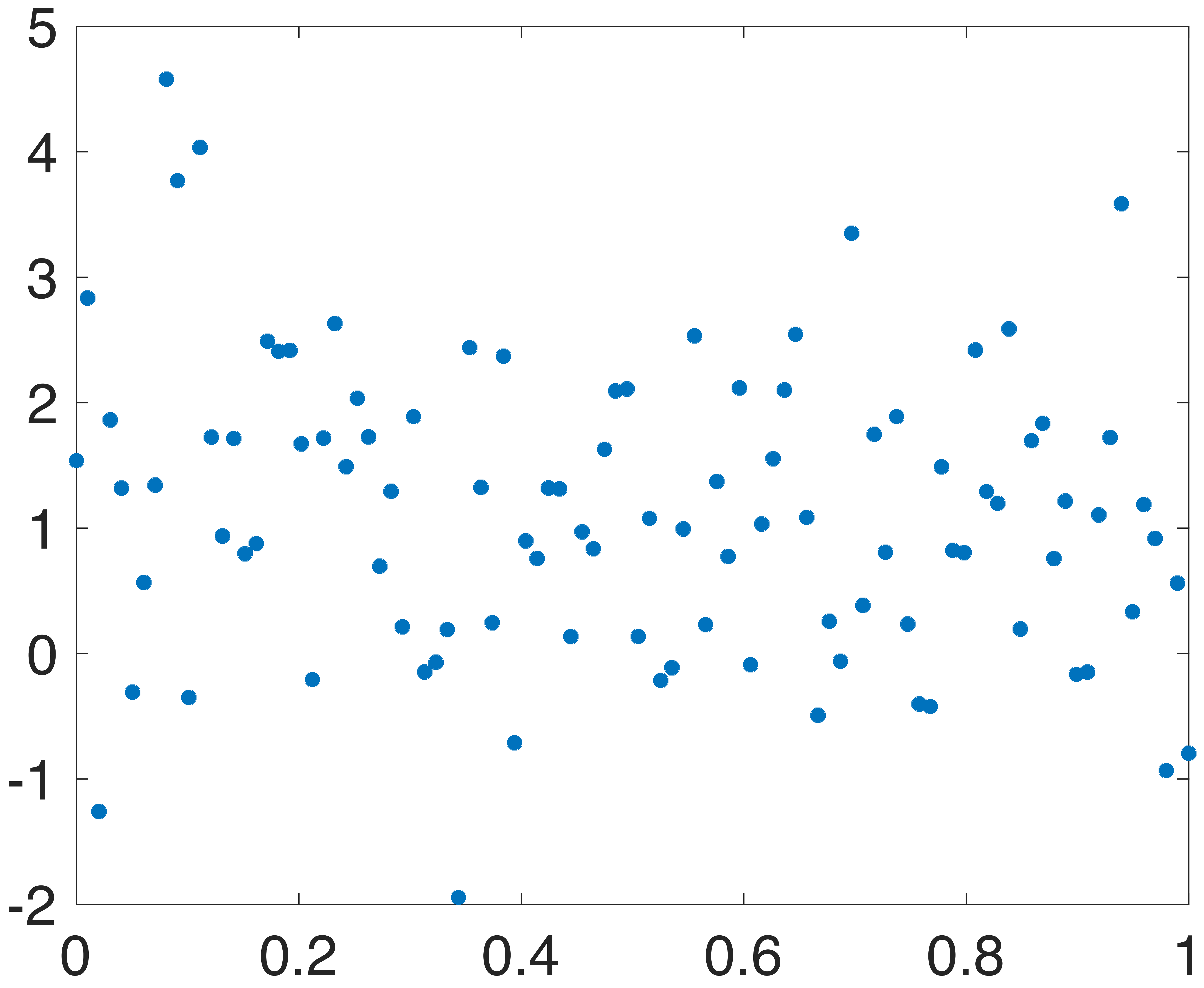
MSE cost function: $\mathrm{MSE}(w_0) := \frac{1}{2s} \sum_{i=1}^{s} |w_0 - y_i|^2$

We do what we did in school: we compute the derivative and set it to zero:

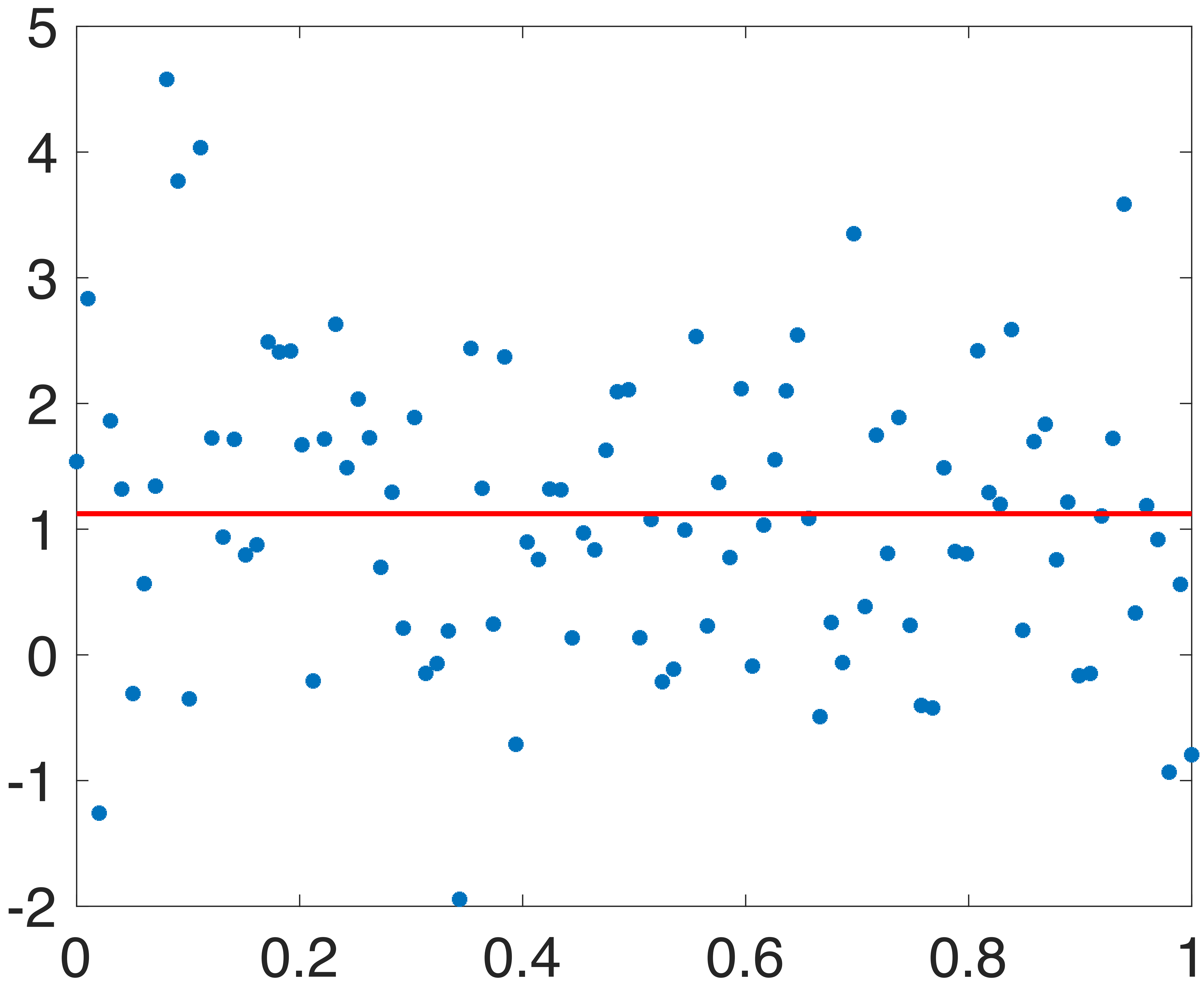
$$\nabla \mathrm{MSE}(\hat{w}_0) = \mathrm{MSE}'(\hat{w}_0) = \frac{1}{s} \sum_{i=1}^{s} (\hat{w}_0 - y_i) \overset{!}{=} 0$$

$$\Rightarrow \qquad \hat{w}_0 = \frac{1}{s} \sum_{i=1}^{s} y_i$$

Example:

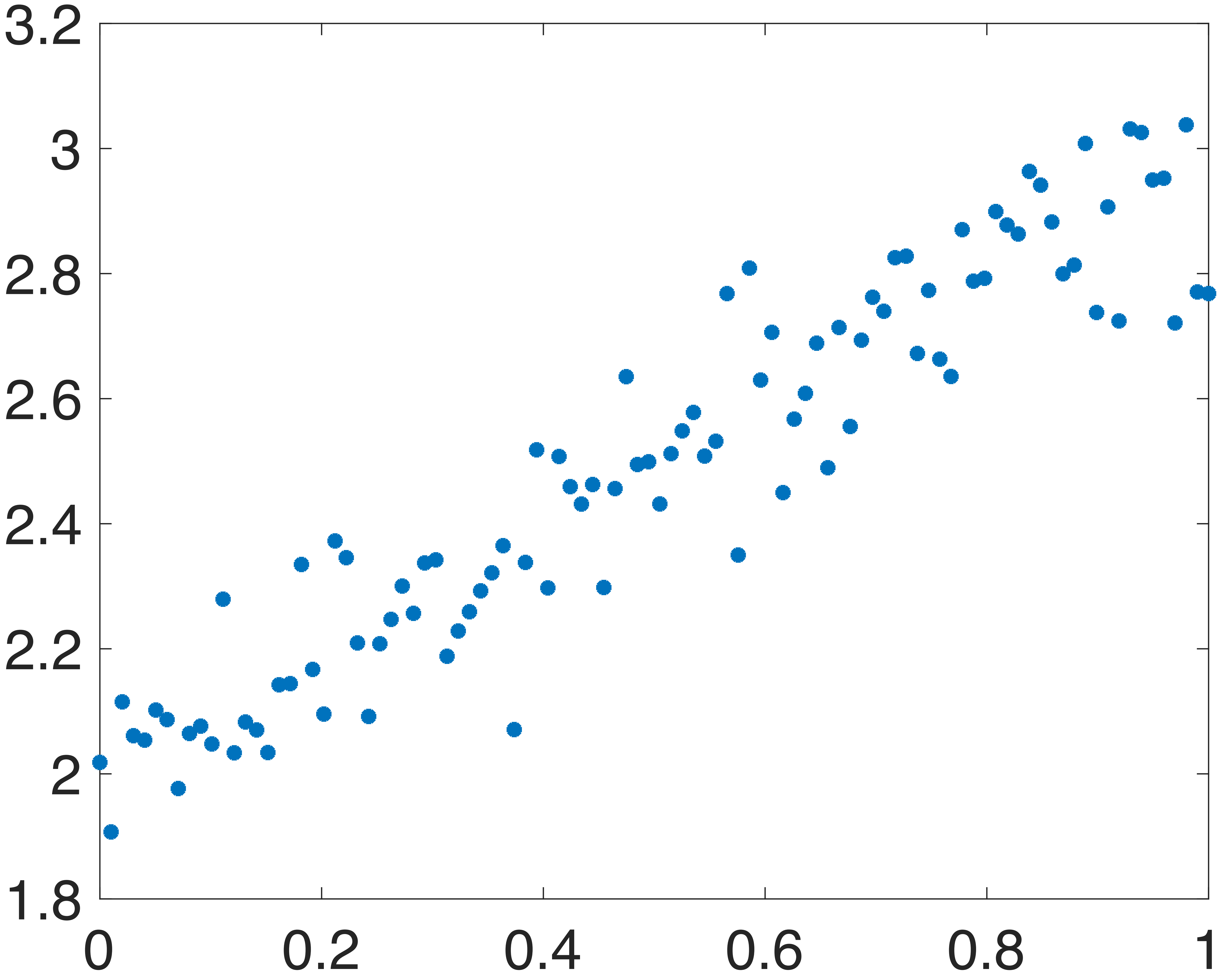


Example:

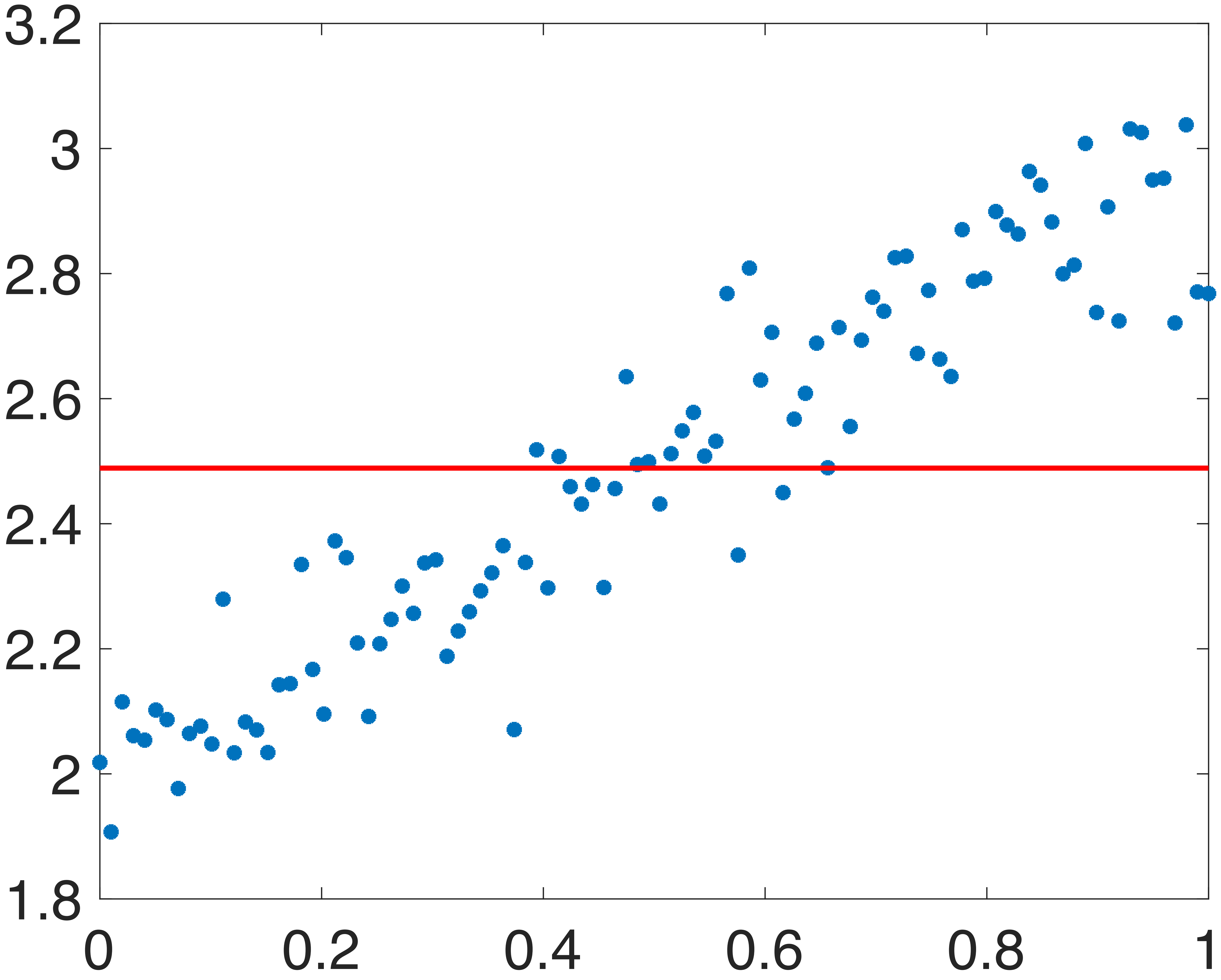


$\hat{w}_0 \approx 1.1231$

Example:



Example:



$\hat{w}_0 \approx 2.4889$

We will discuss how to compute a better approximation now