

# Experimental Data Analysis

*in ©MATLAB*

## **Lecture 8:** Introduction to models, regression analysis

Jan Rusz  
Czech Technical University in Prague



## Motivation

### Association

**Question:** Can be increased blood pressure associated with stress?

**Answer:** Correlation analysis.

### Connection

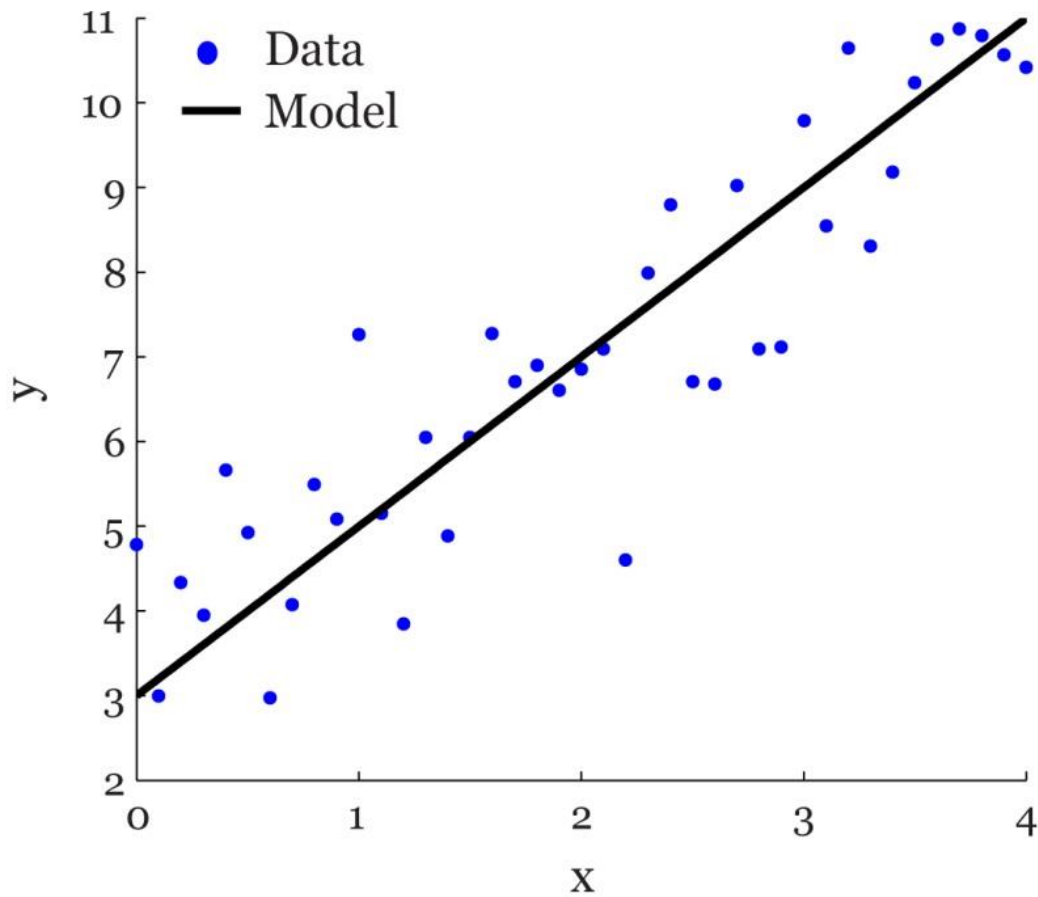
- Correlation indicates a relationship not causality.
- We need to find a connection to say that relationship is causal (i.e. examine that hormonal response to stress can elevate blood pressure).

### Prediction

**Question:** We interrogate the chief suspect (healthy) and measure his blood pressure. How much was the suspect stressed by our key question?

**Answer:** Regression analysis.

# Linear model



Model specification

$$y = ax + b$$



Fitted model

$$y = 2x + 3$$

# Types of models

Supervised learning

Unsupervised learning

Regression

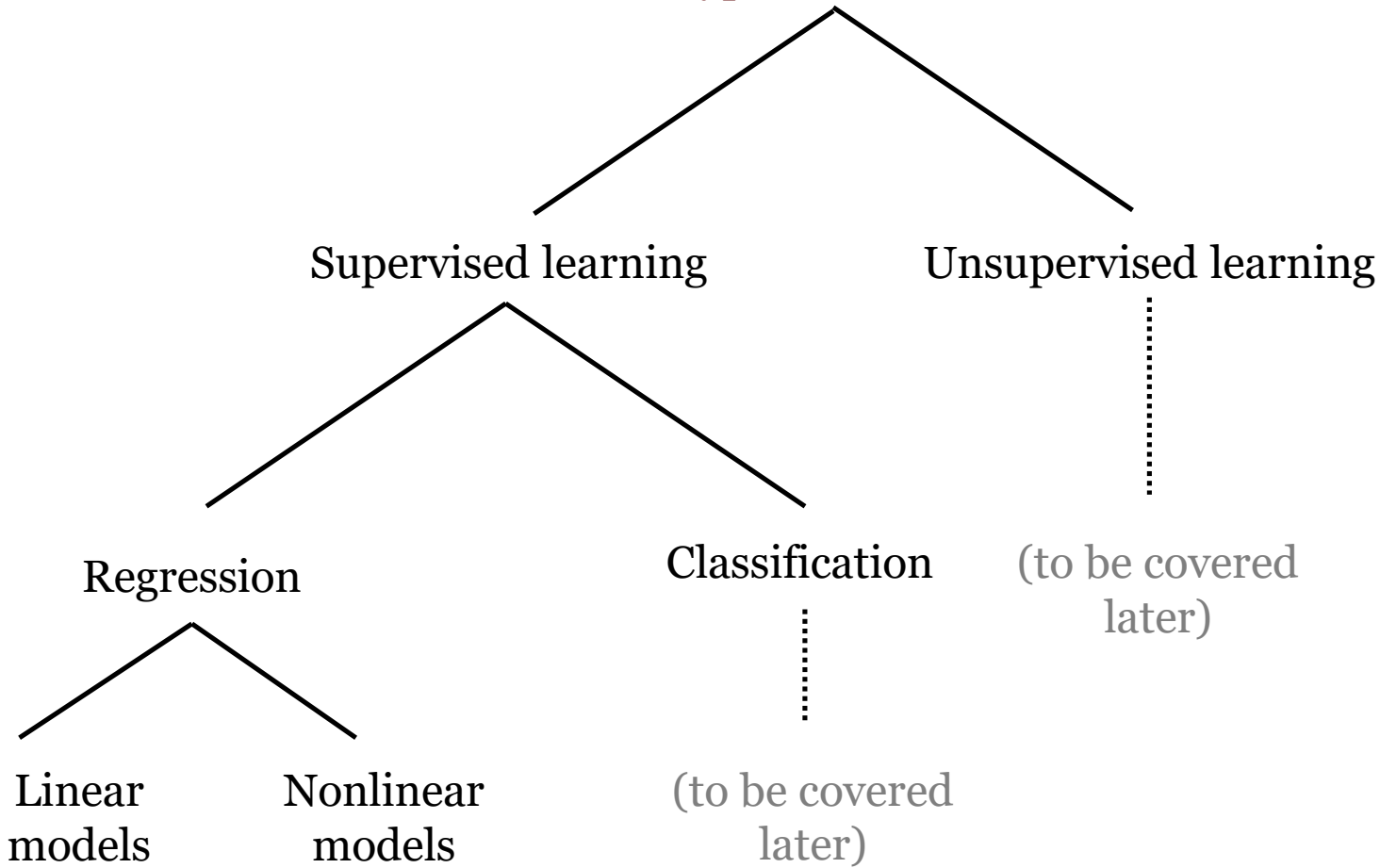
Classification

(to be covered  
later)

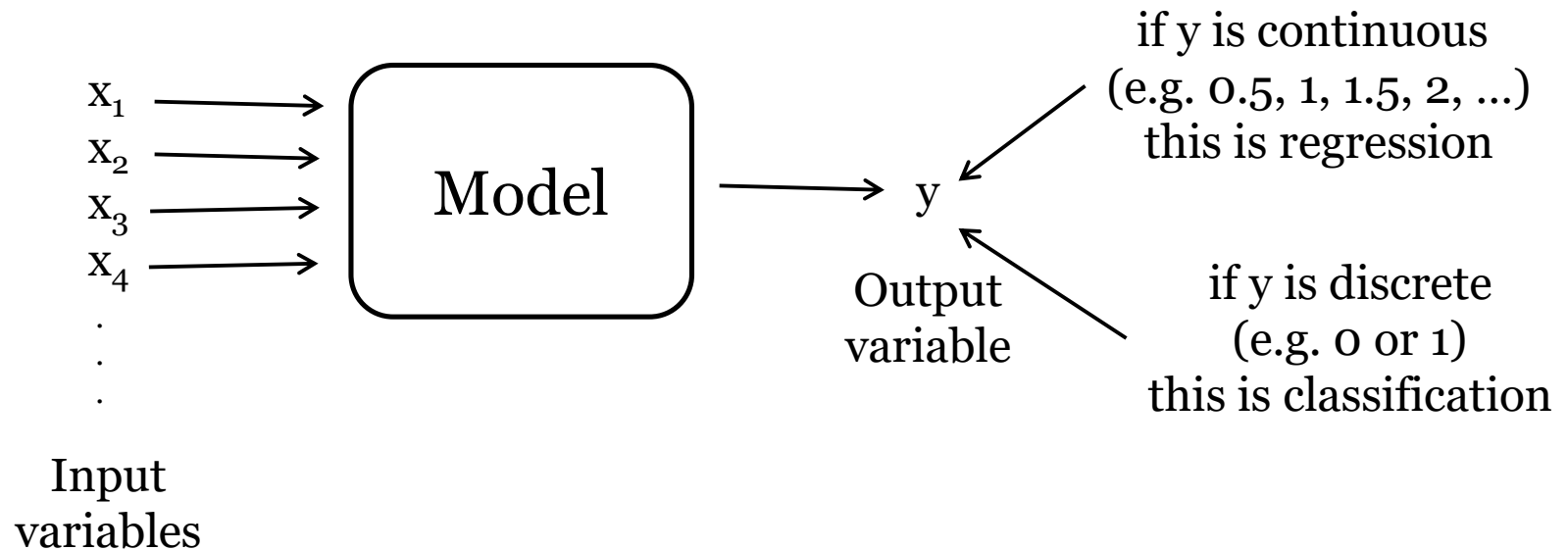
Linear  
models

Nonlinear  
models

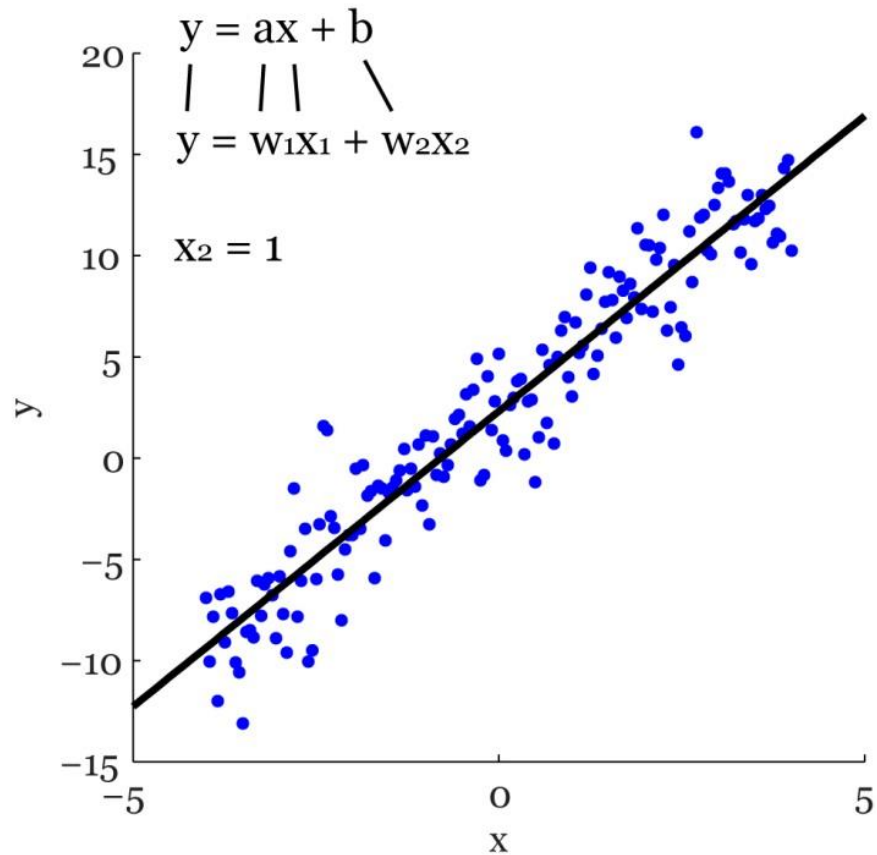
(to be covered  
later)



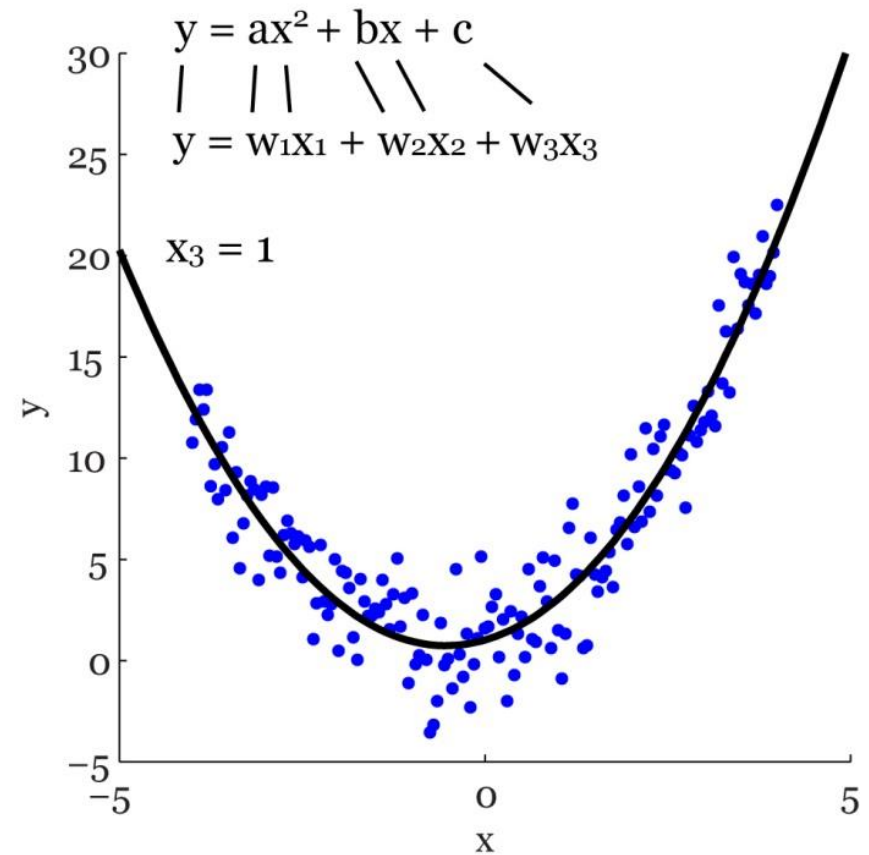
# Supervised learning



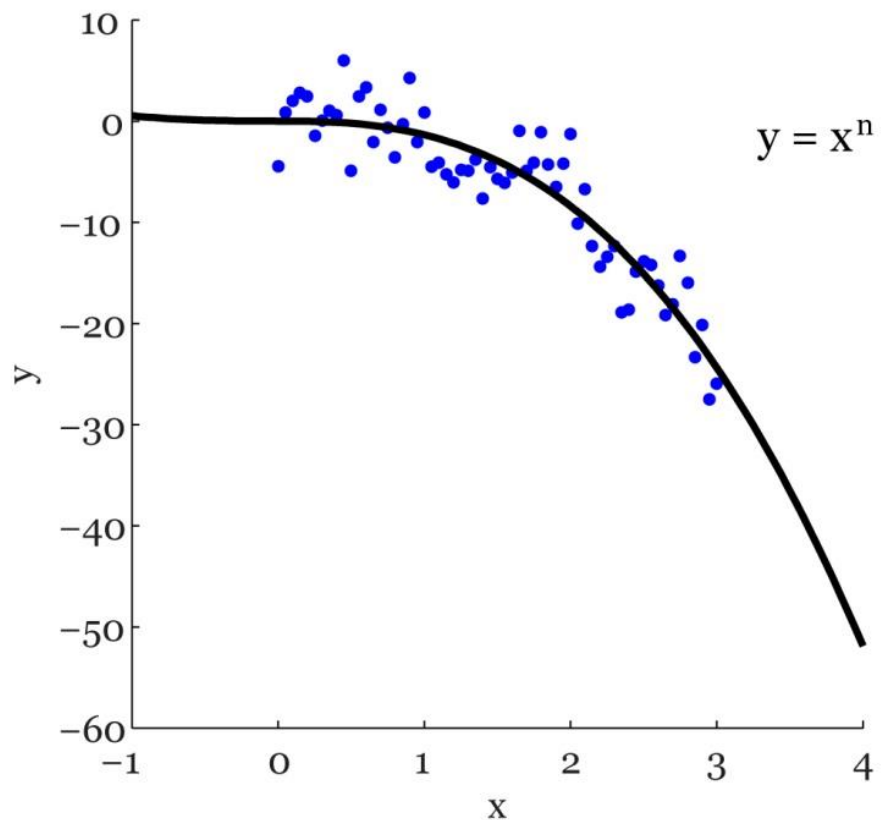
## Linear model



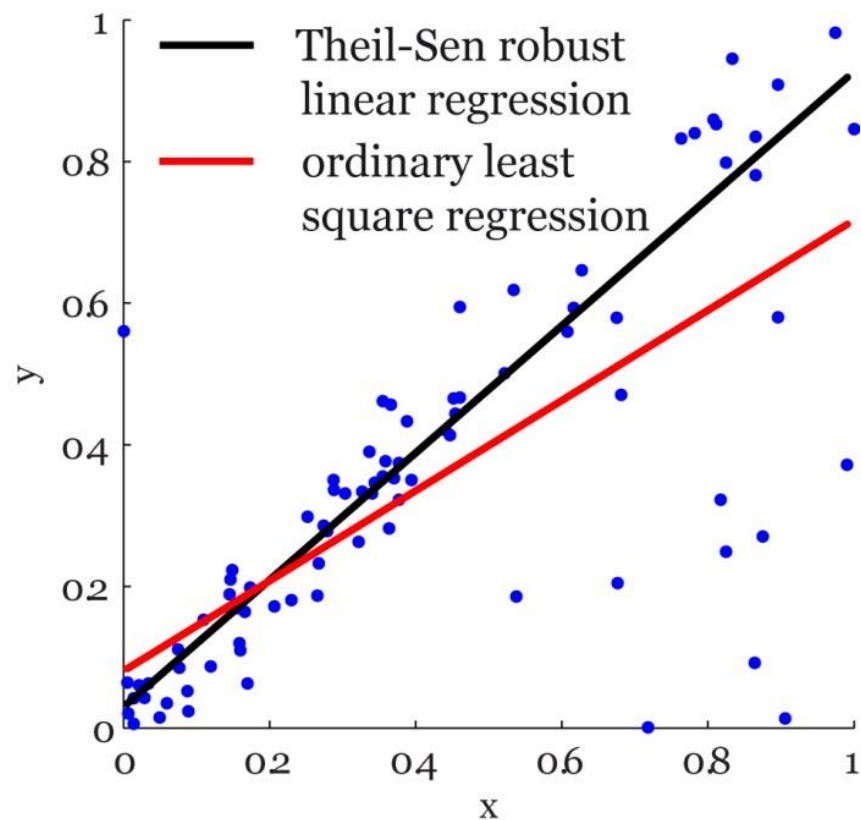
## Linearized model



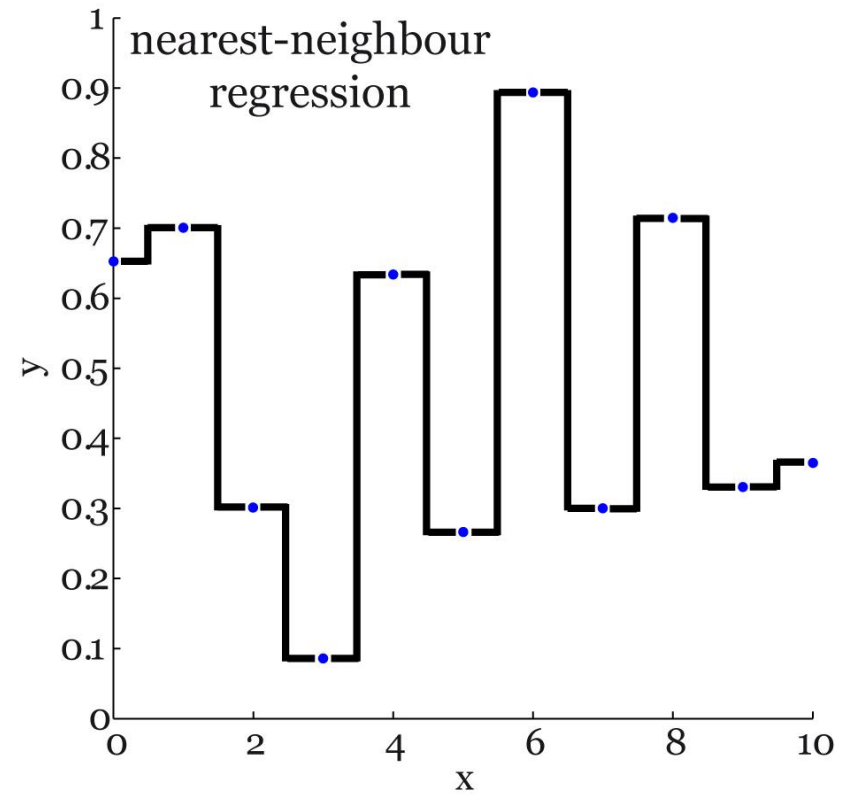
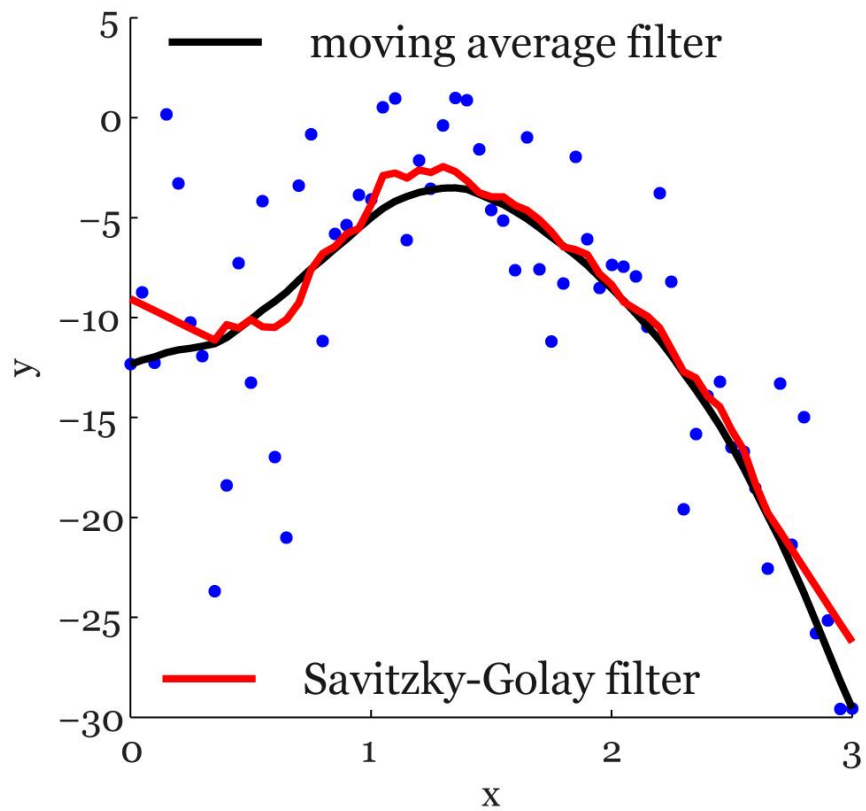
## Parametric nonlinear model



## Nonparametric linear model


















## Nonparametric nonlinear model

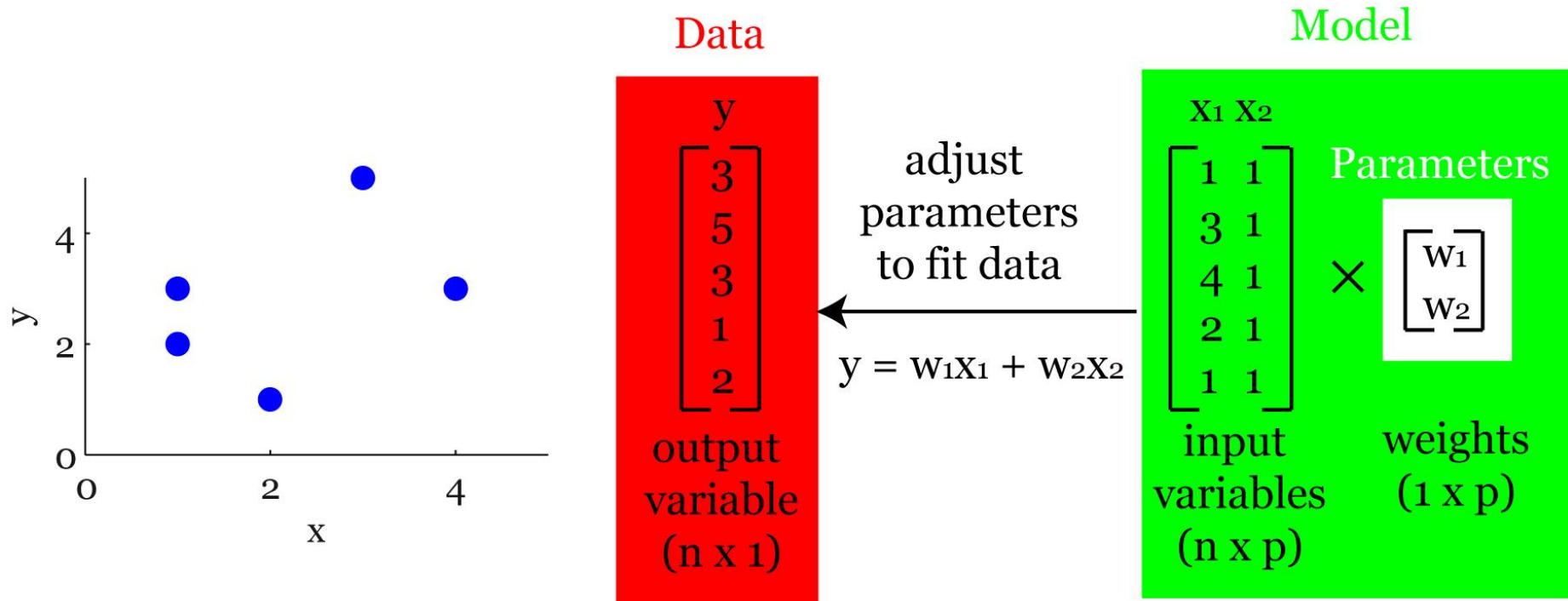




# Model characteristics

|                                | Linear  | Parametric   | Linear in parameters  |
|--------------------------------|---|--|---|
| Linear models                  |    |    |    |
| Linearized models              |    |    |    |
| Parametric nonlinear models    |    |    |    |
| Nonparametric linear models    |   |  |   |
| Nonparametric nonlinear models |  |  |  |

# Matrix representation of linear model



# Matrix representation of linear model

**Data**                      **Model**                      **Residuals**

The diagram illustrates the matrix representation of a linear model. It consists of three main components arranged horizontally, separated by an equals sign and a plus sign. On the left, under the heading 'Data' in red, is a red rectangular box containing a black vector  $y$  enclosed in square brackets. In the middle, under the heading 'Model' in green, is a green rectangular box. Inside this box, on the left, is a black matrix  $X$  enclosed in square brackets. To its right is a multiplication symbol  $\times$ , followed by a white rectangular box containing a black vector  $w$  enclosed in square brackets. Above this white box is the word 'Parameters' in white text. On the right, under the heading 'Residuals' in blue, is a blue rectangular box containing a black vector  $e$  enclosed in square brackets. The overall equation represented is  $y = Xw + e$ .

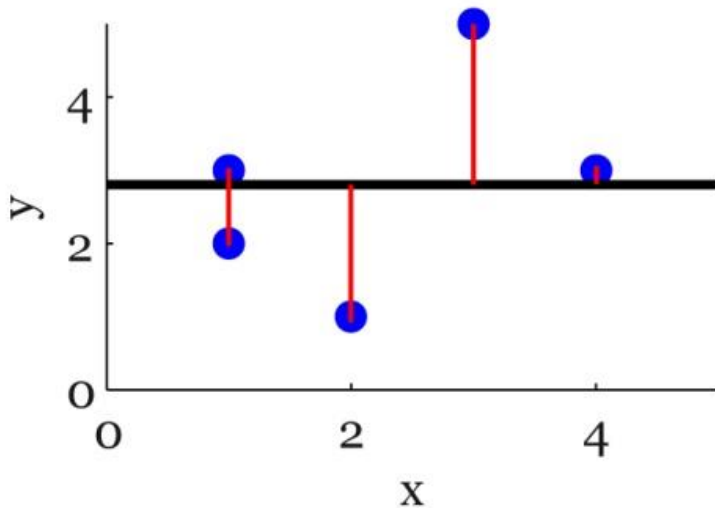
$$\begin{bmatrix} y \end{bmatrix} = \begin{bmatrix} X \end{bmatrix} \times \begin{bmatrix} w \end{bmatrix} + \begin{bmatrix} e \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}$$

# Squared error

- Data
- Model
- Residuals

$$\text{squared error} = \sum_{i=1}^n (d_i - m_i)^2$$



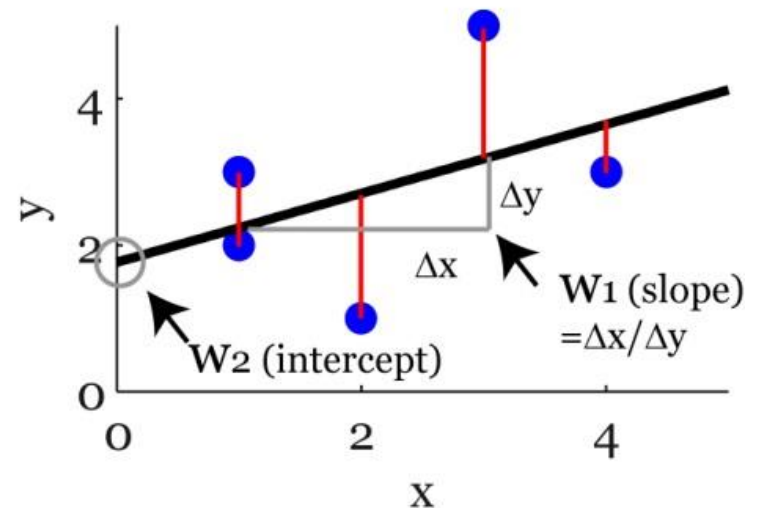
$$y = w_1x + w_2$$

$$w_1 = 0$$

$$w_2 = \text{mean}(x) = 2.8$$

$$y = 2.8$$

$$\text{squared error} = 8.8$$



$$y = w_1x + w_2$$

$$w_1 = 0.47$$

$$w_2 = 1.76$$

$$y = 0.47x + 1.76$$

$$\text{squared error} = 7.29$$

# Ordinary least squares solution

$$\text{regressors} \cdot \text{residuals} = 0$$

$$X^T e = 0$$

$$\text{regressors} \cdot (\text{data} - \text{modelfit}) = 0$$

$$X^T (y - Xw) = 0$$

$$X^T y - X^T X w = 0$$

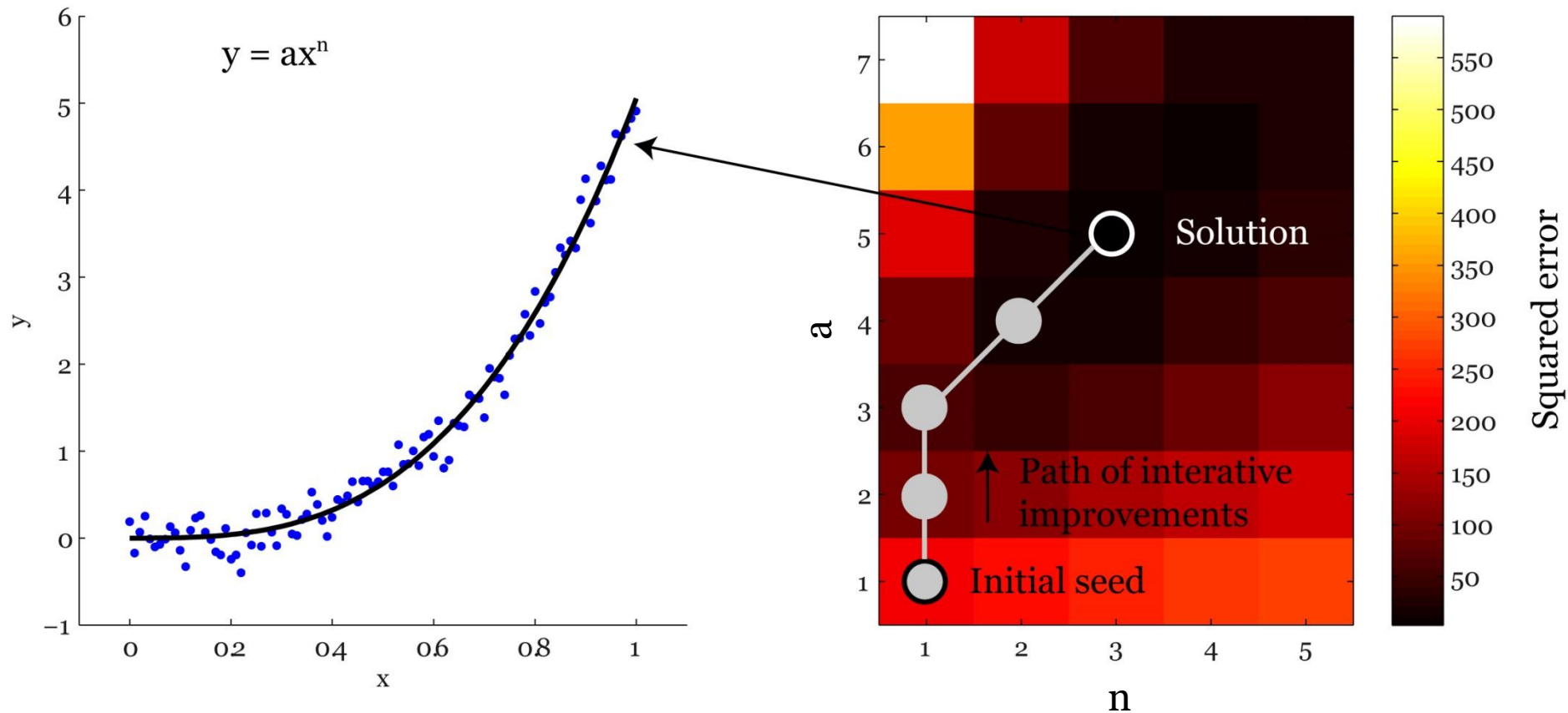
$$w = (X^T X)^{-1} X^T y$$

Diagram illustrating the Ordinary Least Squares (OLS) equation:  $y = Xw + e$ .

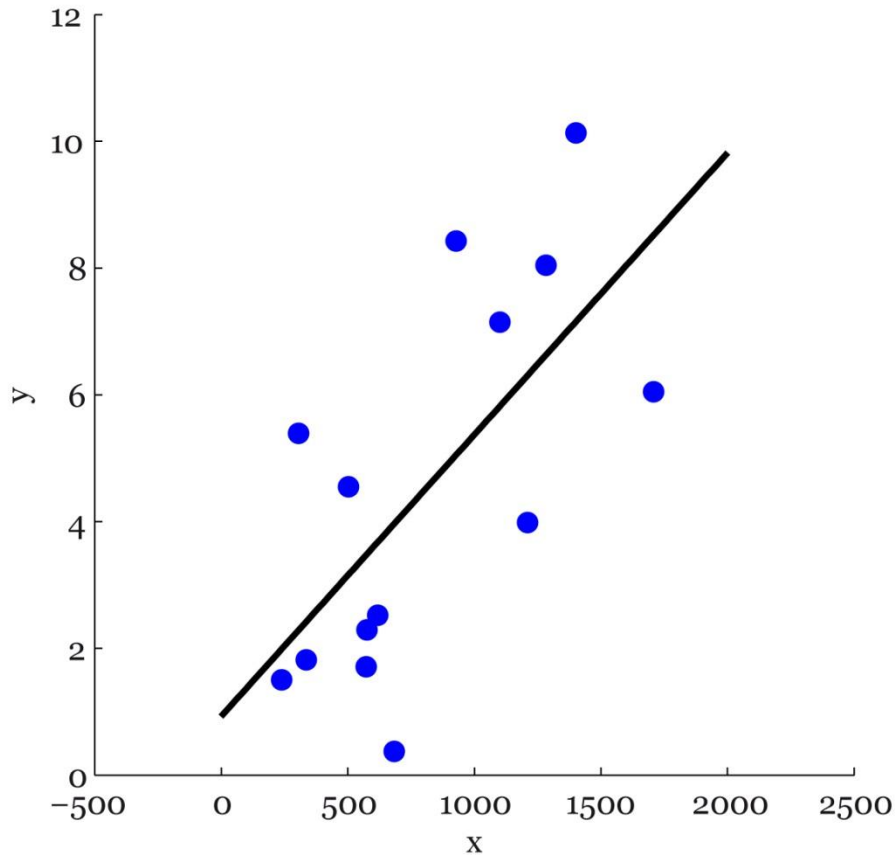
- Data** (Red box):  $y$
- Model** (Green box):  $X$
- Parameters** (White box):  $w$
- Residuals** (Blue box):  $e$

The equation is represented as:  $y = Xw + e$ .

# Fitting nonlinear model based on local, iterative optimization



## Quantifying model accuracy

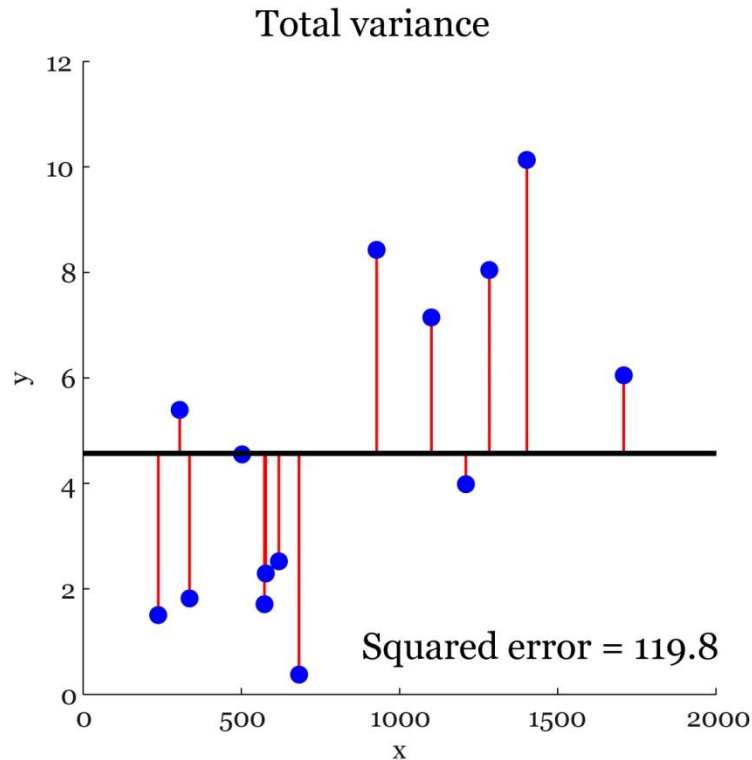


Squared error = 66.4  
(dependent on units, hard to interpret)

$R^2 = 44.6\%$   
(independent on units, easy to interpret)

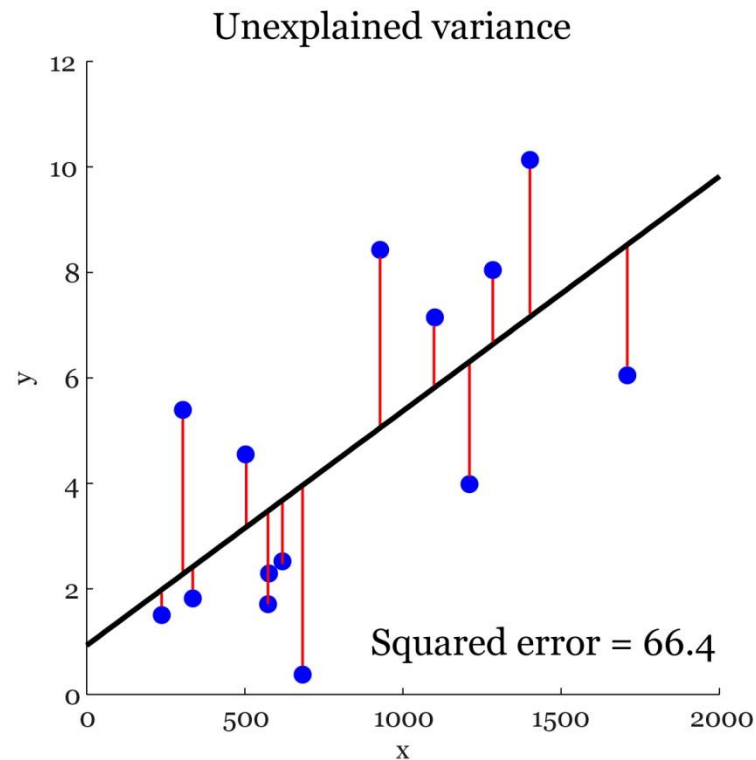
$$variance = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n - 1}$$

## Coefficient of determination ( $R^2$ )



$$R^2 = 100 \times \left( 1 - \frac{\text{unexplained variance}}{\text{total variance}} \right)$$

$$R^2 = 100 \times \left( 1 - \frac{\frac{\sum_{i=1}^n (d_i - m_i)^2}{n-1}}{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}} \right)$$

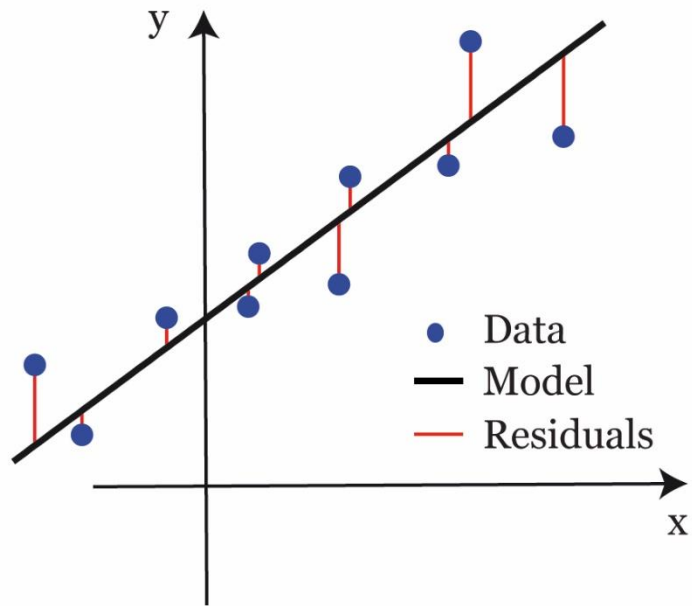


$$R^2 = 100 \times \left( 1 - \frac{SE \text{ model fit}}{SE \text{ model mean}} \right)$$

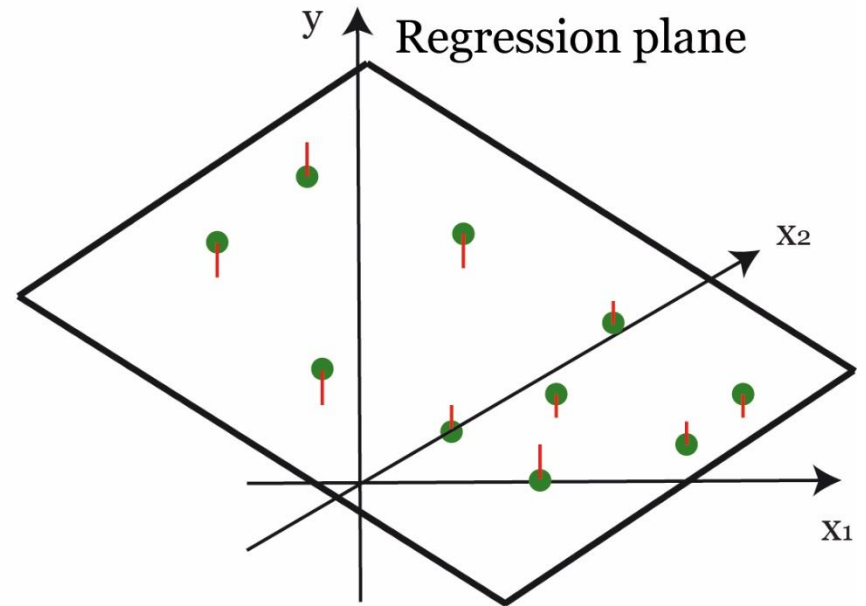
$$R^2 = 100 \times \left( 1 - \frac{\sum_{i=1}^n (d_i - m_i)^2}{\sum_{i=1}^n (d_i - \bar{d})^2} \right)$$



## Simple linear regression



## Multiple linear regression



## Research project: Parkinson's disease (PD), stuttering & L-dopa

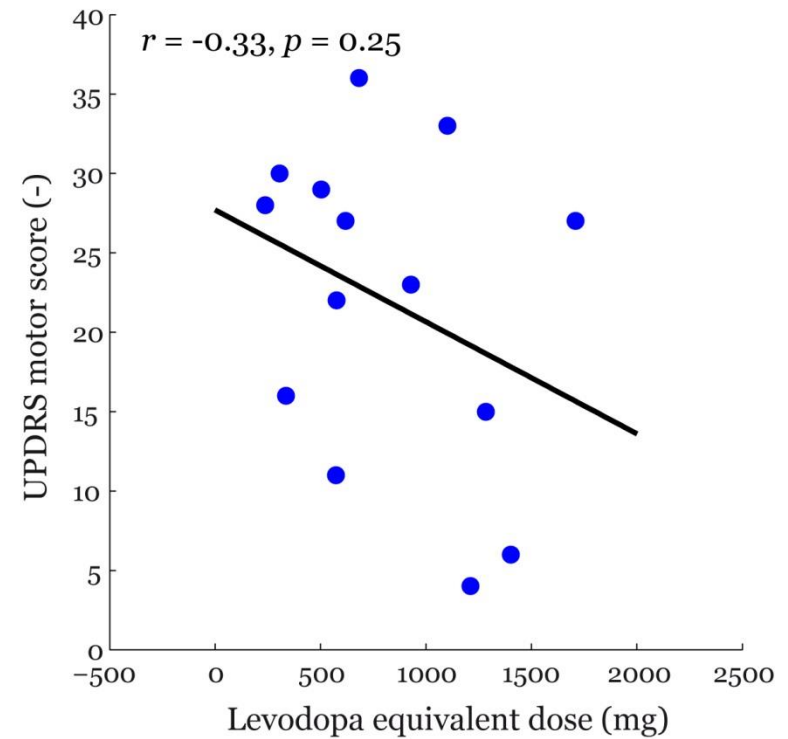
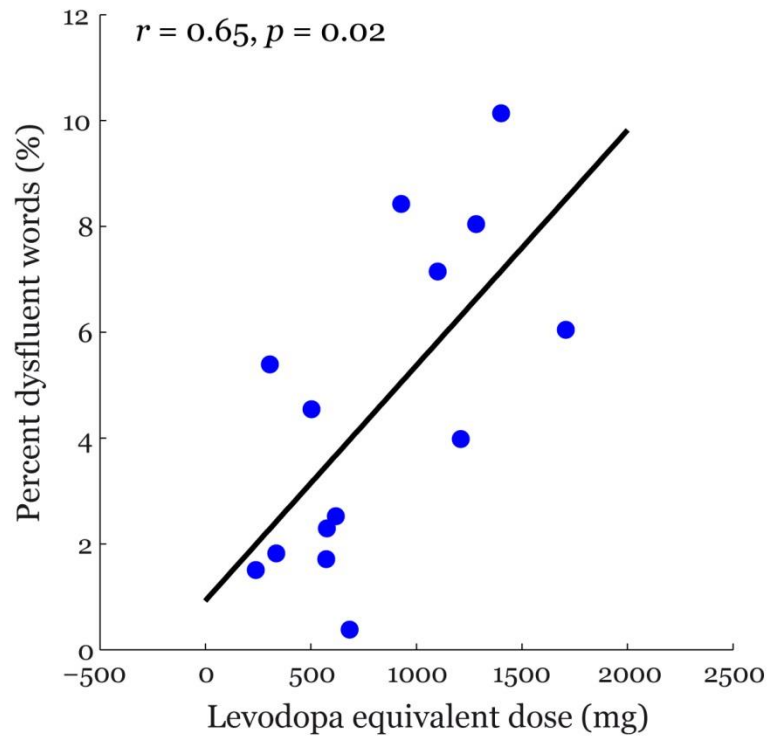
### Background:

- Excess dopamine theory of stuttering suggests that stuttering may be related to an excess amount of dopamine within the brain
- Some patients with PD develop stuttering in the course of their illness
- Levodopa is precursor of dopamine used to treat motor manifestations of patients with PD

### Hypothesis:

- Stuttering is related to extent of L-dopa doses
- Stuttering is not related to motor speech manifestations in PD

## Research project: Parkinson's disease (PD), stuttering & L-dopa



## Research project: Parkinson's disease (PD), stuttering & L-dopa

Linear regression model:

$$y \sim 1 + x1 + x2$$

Estimated Coefficients:

|                        | Estimate   | SE        | tStat    | pValue   |
|------------------------|------------|-----------|----------|----------|
| stuttering (Intercept) | 1.1987     | 2.4258    | 0.49413  | 0.63093  |
| L-dopa x1              | 0.0043749  | 0.0015763 | 2.7755   | 0.018049 |
| UPDRS x2               | -0.0097572 | 0.071821  | -0.13586 | 0.89439  |

Number of observations: 14, Error degrees of freedom: 11

Root Mean Squared Error: 2.46

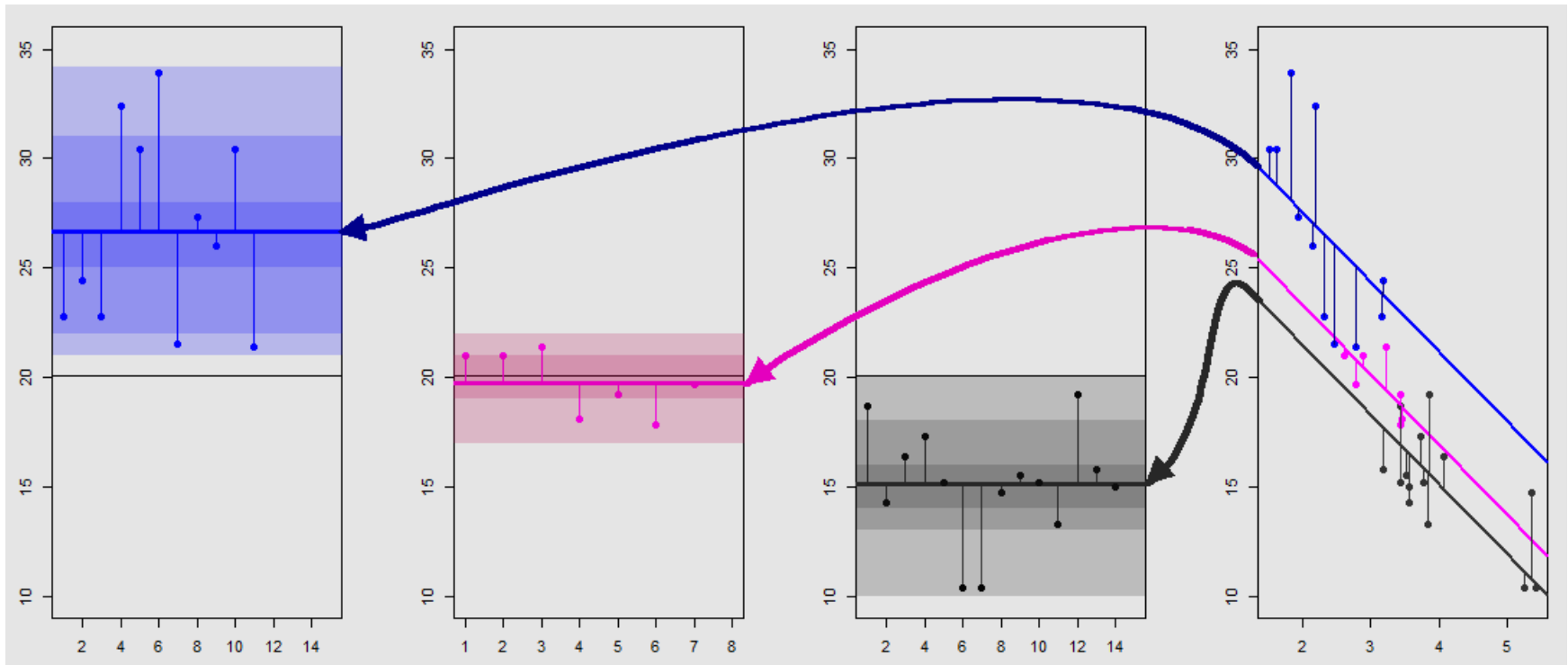
R-squared: 0.447, Adjusted R-Squared 0.346

F-statistic vs. constant model: 4.44, p-value = 0.0386

How to report results of regression?

Our case:  $[F(2,11) = 4.4, p = 0.04, R^2 = 0.45]$

- ANOVA and linear regression analysis are the “same thing”



- Intercept is the mean of the reference group*
- The coefficients for the other two groups are the differences in the mean between the reference group and the other groups*