

# Workshop 6: Curve Fitting and $R^2$

We will start at 2pm. This is posted on Canvas under Files → Workshops → Workshop 6 if you'd like to download and follow along/annotate (highly recommended)

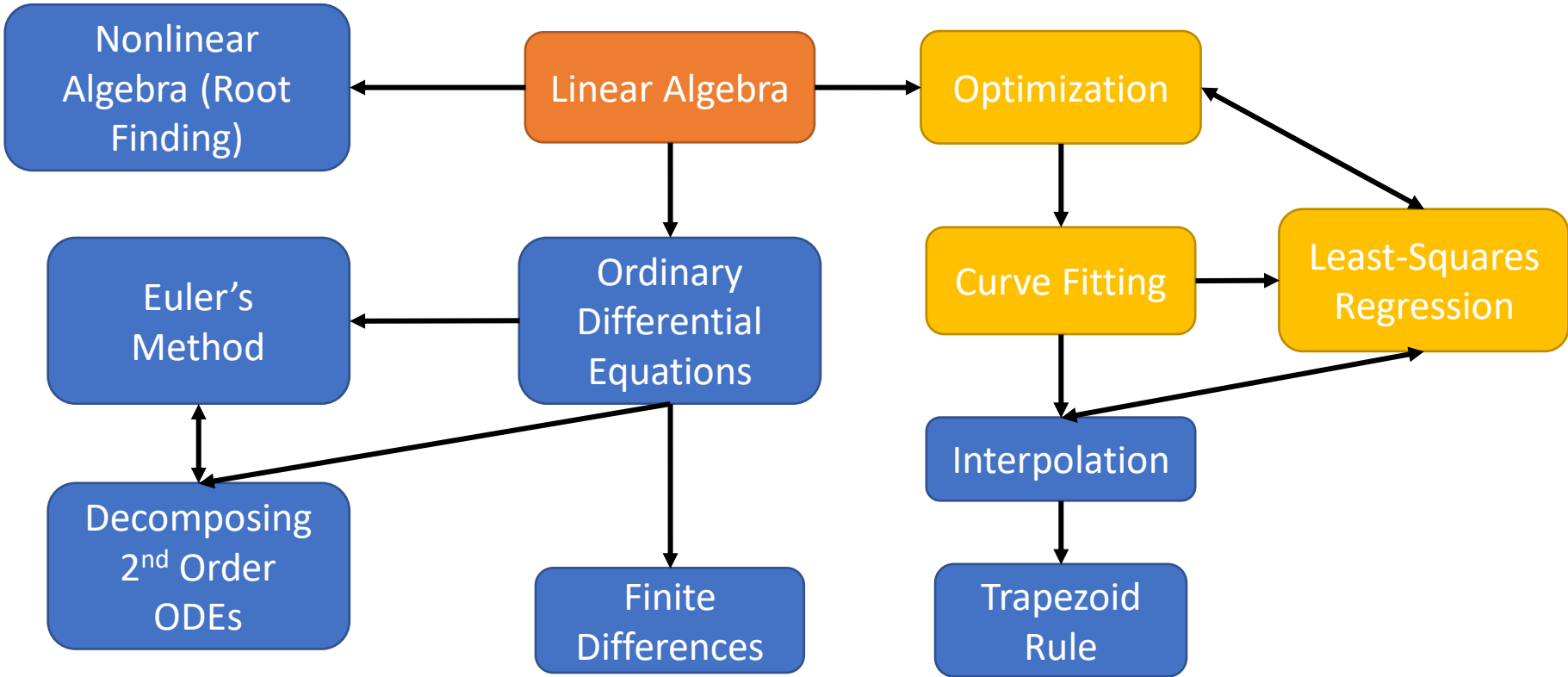
# Outline (Part 1, Part 2)

- 1.1: Curve Fitting Overview
  - Big Picture & Optimization Intro
  - Definitions
- 1.2: Best-Fit Line
  - Candidate Best-Fit Lines
  - Least-Squares Regression
- 2.1: Goodness of Fit
  - $R^2$  Definitions
- 2.2: Limitations and Warnings

# 1.1: Curve Fitting Overview



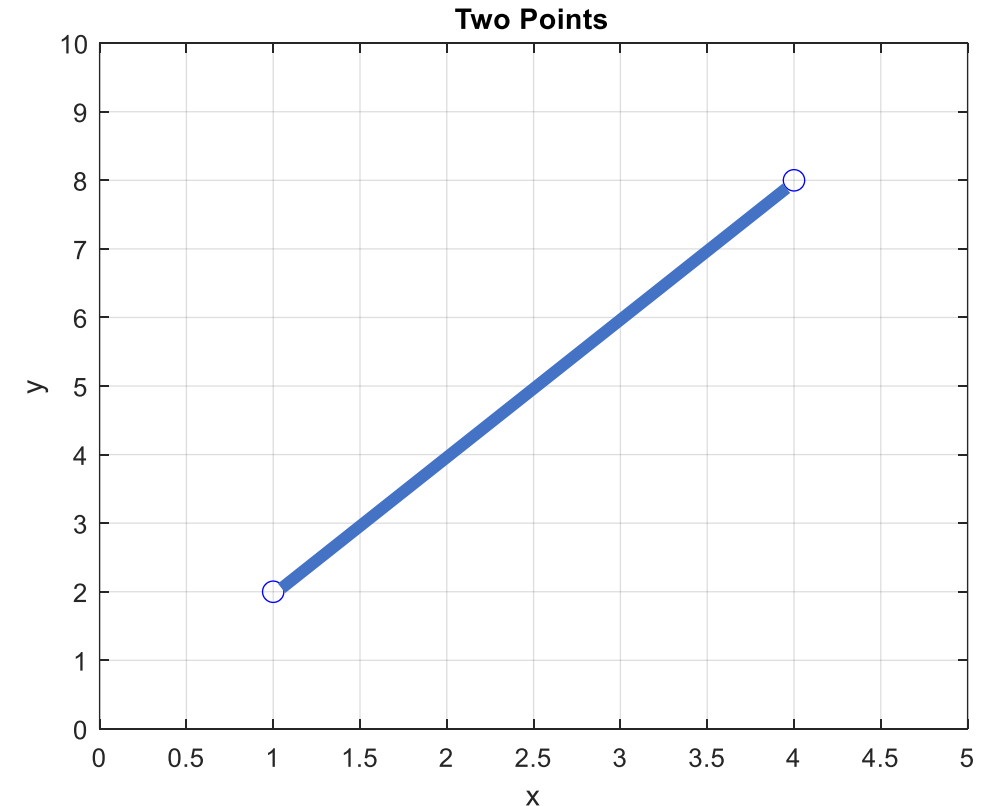
# Curve Fitting Overview



Calculus: Numerical differentiation
Calculus: Numerical differentiation
Calculus: Numerical integration workshop
Calculus: Numerical integration
Calculus: Numerical integration
Calculus workshop
Linear algebra
Linear algebra
Linear algebra workshop
Linear algebra
Curve fitting
Curve fitting workshop
Interpolation
Mid Term Exam
Fall Break
Non-linear algebra: root finding
Non-linear algebra: root finding
Non-linear algebra: root finding workshop
Non-linear algebra: root finding
Non-linear algebra: root finding
Non-linear algebra: root finding workshop
Ordinary differential equations
Ordinary differential equations
Ordinary differential equations workshop
Ordinary differential equations
Ordinary differential equations
Ordinary differential equations workshop
Ordinary differential equations
Ordinary differential equations
Ordinary differential equations workshop
Thanksgiving
Ordinary differential equations
Ordinary differential equations
Ordinary differential equations workshop
Ordinary differential equations

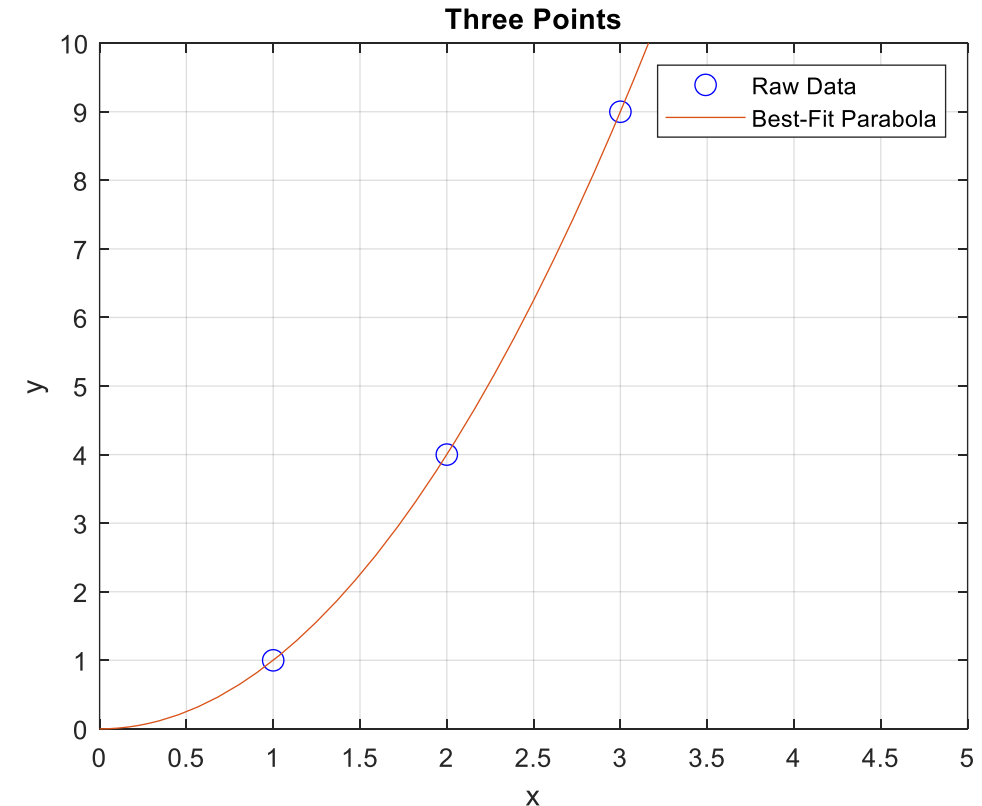
# Curve Fitting Overview

- Given experimental data, can we predict a trend between the *independent variable* and *dependent variable*?
- For this simple system: easy!
- Trend is governed by the *best-fit line*
  - 2 data points
  - 1<sup>st</sup> order polynomial



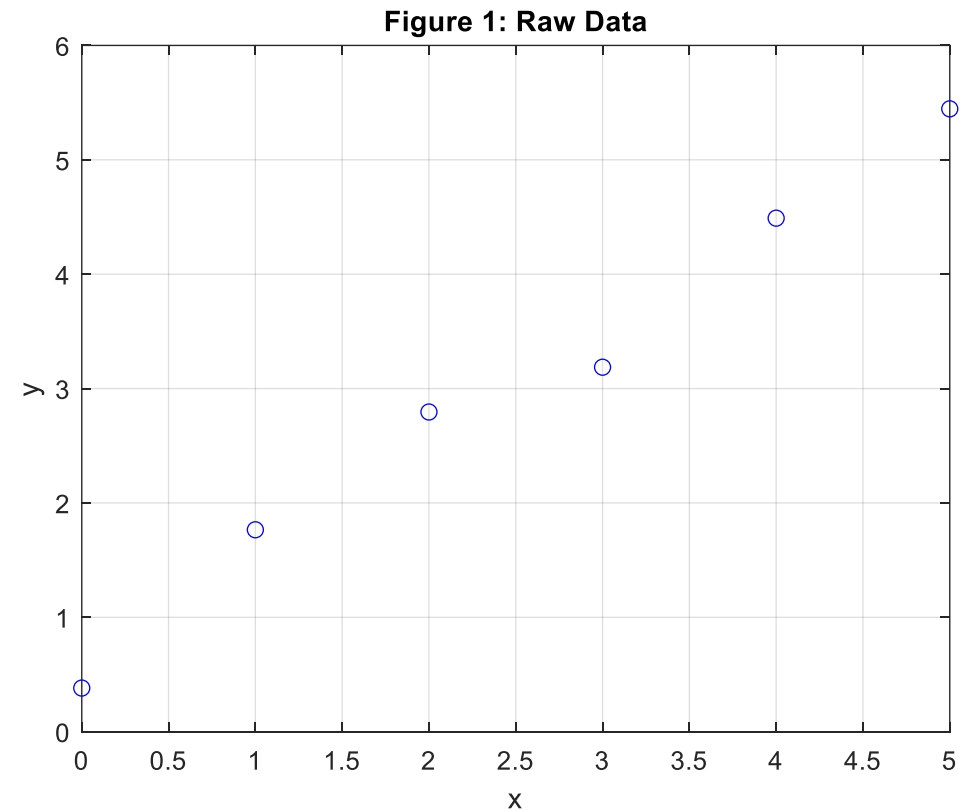
# Curve Fitting Overview

- Given experimental data, can we predict a trend between the *independent variable* and *dependent variable*?
- For this system: a little more work, but still easy
- Trend is governed by the *best-fit line*
  - 3 data points
  - 2<sup>nd</sup> order polynomial



# Curve Fitting Overview

- You can perfectly fit an  $(n - 1)th$  order polynomial through  $n$  data points
- This is the *unique* best-fit line (or curve)
- But in practice:
  - Lots of data
  - 99.999999999999999999999999999999% of experimental data contain errors



# Curve Fitting Overview

## Optimization problem

From Wikipedia, the free encyclopedia

*For broader coverage of this topic, see [Mathematical optimization](#).*

In [mathematics](#), [computer science](#) and [economics](#), an **optimization problem** is the [problem](#) of finding the best solution from all [feasible solutions](#).

The [standard form](#) of a [continuous](#) optimization problem is<sup>[1]</sup>

$$\begin{aligned} & \underset{x}{\text{minimize}} && f(x) \\ & \text{subject to} && g_i(x) \leq 0, \quad i = 1, \dots, m \\ & && h_j(x) = 0, \quad j = 1, \dots, p \end{aligned}$$

where

- $f: \mathbb{R}^n \rightarrow \mathbb{R}$  is the [objective function](#) to be minimized over  $\mathbb{R}^n$
- $g_i(x) \leq 0$  are called **inequality constraints**
- $h_j(x) = 0$  are called **equality constraints**, and
- $m \geq 0$  and  $p \geq 0$ .

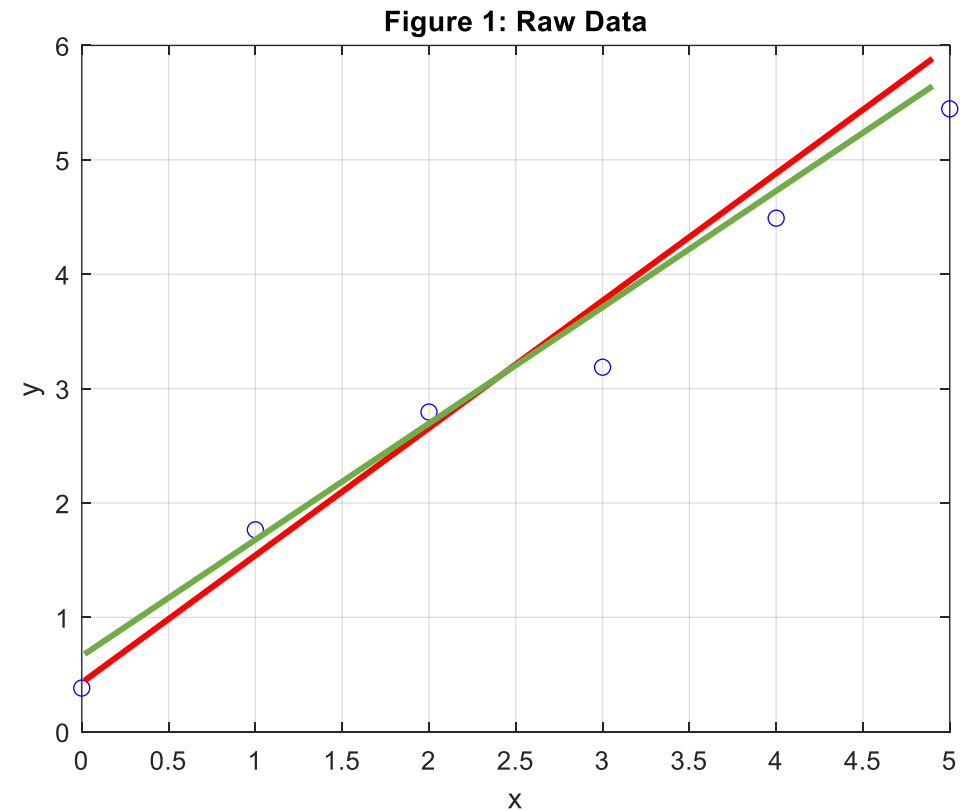
or *cost function*



# Curve Fitting Overview

- Objective: fit a 1<sup>st</sup> order polynomial to the data
  - 6 data points ( $n = 6$ )
  - $(n - 1) = 5 \neq 1??????$
- Can't perfectly fit a 1<sup>st</sup> order polynomial to the data, so you will incur a *penalty* (cost)
- Therefore, the best-fit line is the line which *minimizes* (optimizes) the penalty (error)

How do we *objectively* define THE best-fit line?



# Curve Fitting Overview

- *Best-fit line*: line (or curve) which minimizes the discrepancy between the data points (minimizes the penalty)
- Mathematical expression of a straight line:

$$y = a_0 + a_1x + e \quad (1)$$

$a_0$ : y-intercept;  $a_1$ : slope;  $e$ : residual (error) b/w data and curve

- Rearranging (1):

$$e = y - (a_0 + a_1x) \quad (2)$$

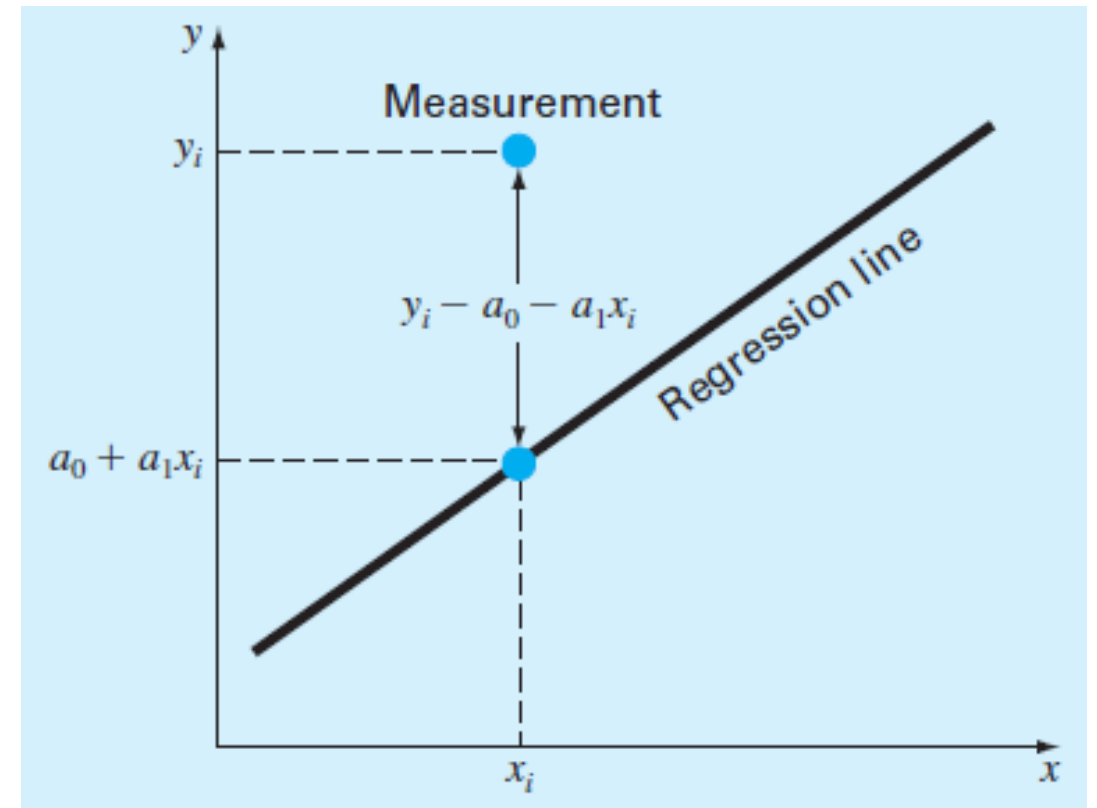
Residual (error)      true value      approximate value (best-fit line)

# Curve Fitting Overview

- Residuals (errors):

$$e_i = y_i - (a_0 + a_1 x_i) \quad (2)$$

- Visual representation of residuals
  - $y_i$  above best-fit line:  $e_i > 0$
  - $y_i$  below best-fit line:  $e_i < 0$
  - $y_i$  on best-fit line:  $e_i = 0$  (ideal)



## 1.2: Best-Fit Line



# Best-Fit Line

- How do we *objectively* define THE best-fit line?
- Five candidate methods:
  - A: Mean of the data
  - B: Minimize  $e$
  - C: Minimize  $|e|$
  - D: Minimax method
  - E: Least-squares regression
- Of the five, which is *objectively* the best?

# Candidate Methods: (A) Mean of the Data

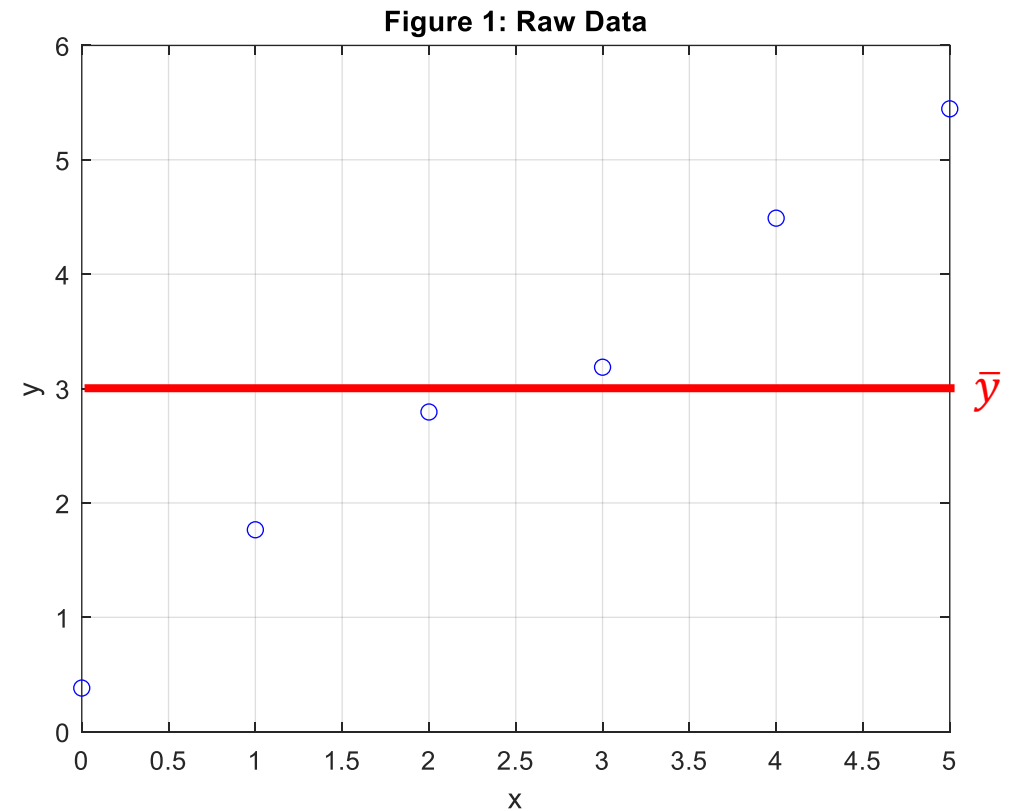
- (A): Mean of the data:

$$y = a_0 + e \rightarrow e = y - a_0 = y - \bar{y}$$

- For  $n$  data points:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \bar{y})$$

- Pros: unique line ✓
- Cons: bad ✗

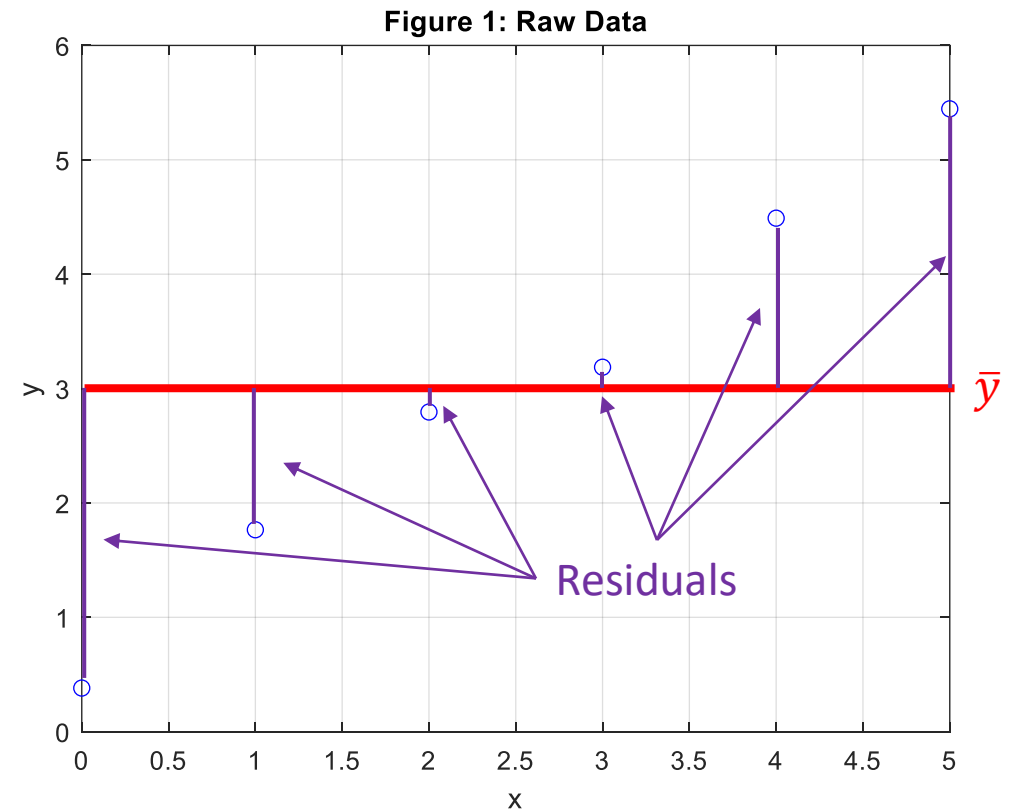


# Candidate Methods: (A) Mean of the Data

- Introduce  $S_t$ :

$$S_t = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

- Sum of the square of the residuals between the data points ( $y_i$ ) and the mean ( $\bar{y}$ )
- This simple dataset produces large residuals  $\rightarrow$  large  $S_t$ !



# Candidate Methods: (B) Minimize $e$

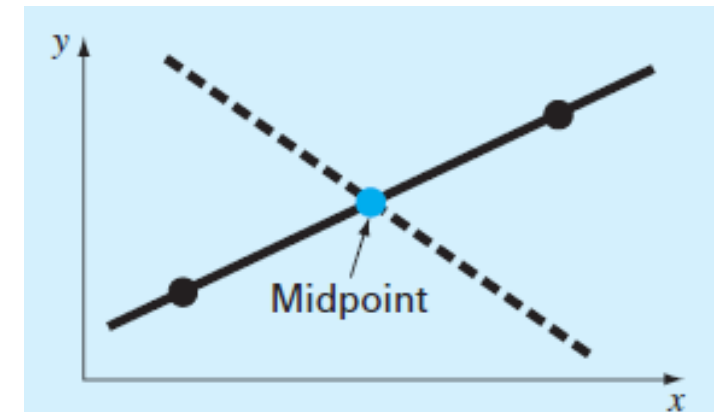
- (B): Minimize  $e$ :

$$e = y - (a_0 + a_1 x) \quad (2)$$

- For  $n$  data points:

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \quad (4)$$

- According to this definition, any line passing through the midpoint is a best-fit line (+ and – errors cancel) ❌





# Candidate Methods: (C) Minimize $|e|$

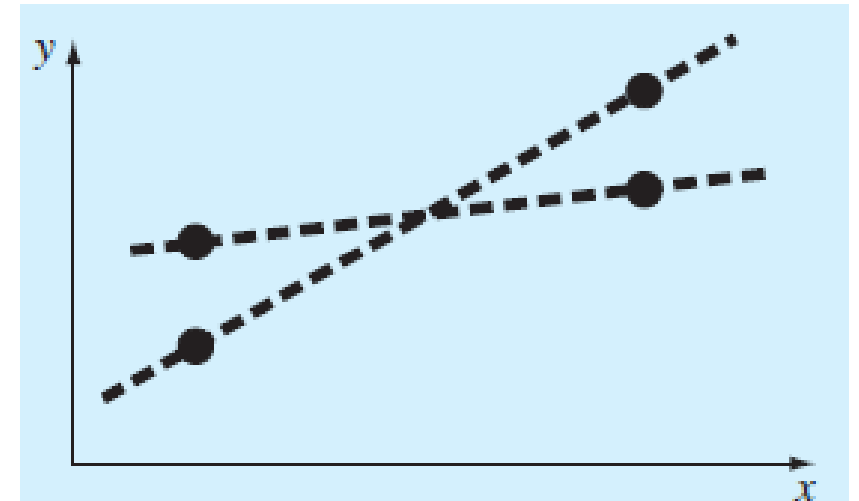
- (B): Minimize  $e$ :

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \quad (4)$$

- (C): Minimize  $|e|$ :

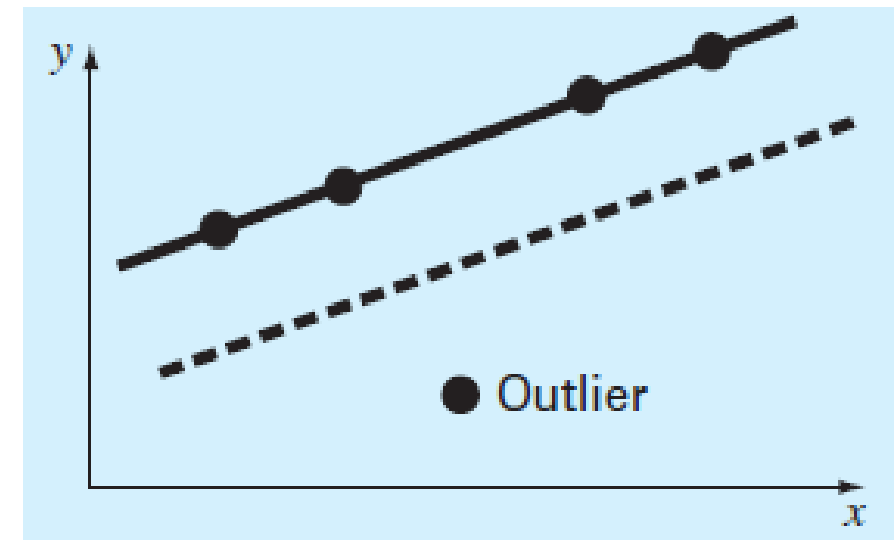
$$\sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - a_0 - a_1 x_i| \quad (5)$$

- According to this definition, any line in between the dashed lines is a best-fit line (+ and – errors cancel) ❌



# Candidate Methods: (D) Minimax Method

- (D): Minimax method
  - Best-fit line: line which minimizes max distance a single point lies from the line
- Results in a *single* line ✓
- Problem: gives undue influence to outliers ✗
- Used in game theory, AI, etc. to minimize possible loss for a worst-case scenario



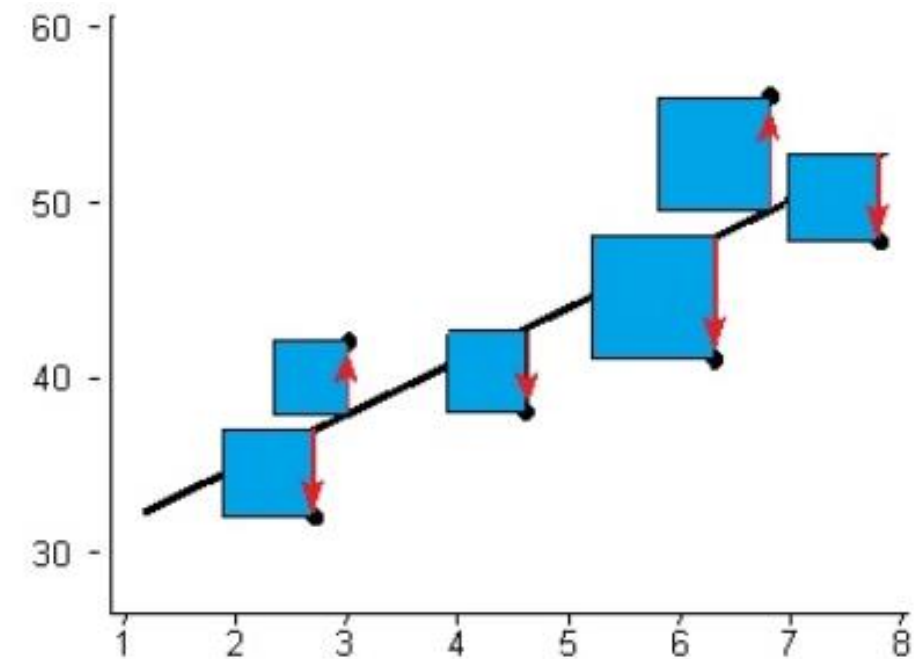
# Candidate Methods: (E) Least-Squares Regression

- (E): Least-squares regression

$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \quad (6)$$

- Minimizing sum of the squares of the residuals
- Ideal for numerous reasons
  - Most notably: produces a *unique line* for the given data

How do we find  $a_0$  and  $a_1$ ?



# Least-Squares: Computing $a_0$ and $a_1$

- Sum of the squares of the residuals:

$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \quad (6)$$

- Take partial derivative of  $S_r$  w.r.t  $a_0$  and  $a_1$  and set to 0:

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i) = 0 \quad (7)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum [(y_i - a_0 - a_1 x_i) x_i] = 0 \quad (8)$$

# Least-Squares: Computing $a_0$ and $a_1$

- Simplifying (7) and (8):

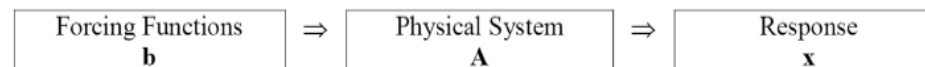
$$0 = \sum y_i - \sum a_0 - \sum a_1 x_i \quad (9)$$

$$0 = \sum x_i y_i - \sum a_0 x_i - \sum a_1 x_i^2 \quad (10)$$

- But...  $\sum a_0 = na_0$  ( $a_0$  is a constant). Rearranging:

$$\begin{aligned} (n)a_0 + (\sum x_i)a_1 &= \sum y_i \\ (\sum x_i)a_0 + (\sum x_i^2)a_1 &= \sum x_i y_i \end{aligned}$$

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (11)$$



# Least-Squares: Computing $a_0$ and $a_1$

- Solving (11) via  $Ax = b$ :

$$a_1 = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{n(\sum x_i^2) - (\sum x_i)^2} \quad (12)$$

$$a_0 = \bar{y} - a_1 \bar{x} \quad (13)$$

where  $\bar{y} = \text{mean}(y)$  and  $\bar{x} = \text{mean}(x)$

- Best-fit line:  $y = a_1 x + a_0$ 
  - Observation: best-fit line always passes through  $(\bar{x}, \bar{y})$ !

# Least-Squares: Computing $a_0$ and $a_1$

- This process extends to higher-order polynomials:

$$y = a_0 + a_1x + a_2x^2 + e$$

$$\rightarrow S_r = \sum_{i=1}^n (y_i - a_0 - a_1x_i - a_2x_i^2)$$

$$(n)a_0 + (\sum x_i) a_1 + (\sum x_i^2) a_2 = \sum y_i$$

$$(\sum x_i) a_0 + (\sum x_i^2) a_1 + (\sum x_i^3) a_2 = \sum x_i y_i$$

$$(\sum x_i^2) a_0 + (\sum x_i^3) a_1 + (\sum x_i^4) a_2 = \sum x_i^2 y_i$$

## 2.1: Goodness of Fit





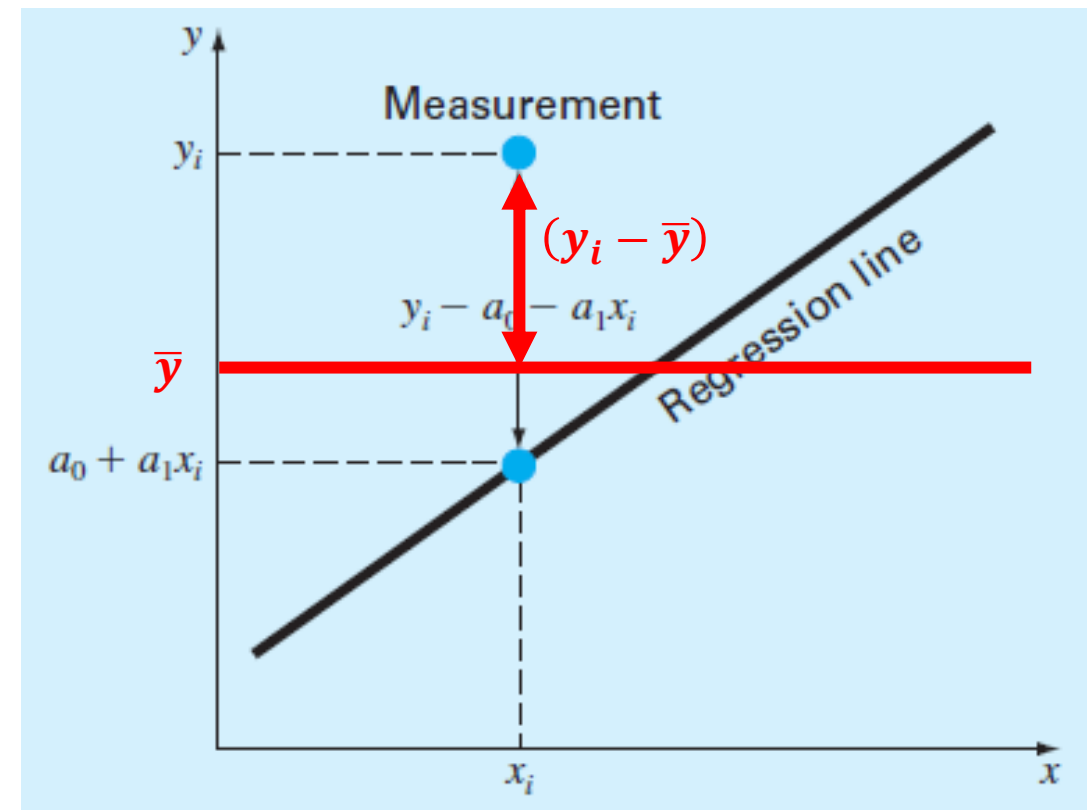
# Goodness of Fit

- Eq. (12) and (13) produce the single line which minimizes the sum of the square of the residuals  $S_r$ 
  - Any other line will have a nonminimal  $S_r$
- So...how good is it?

- Recall  $S_t$ :

$$S_t = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

- Sum of the square of the residuals between the data points ( $y_i$ ) and the mean ( $\bar{y}$ )



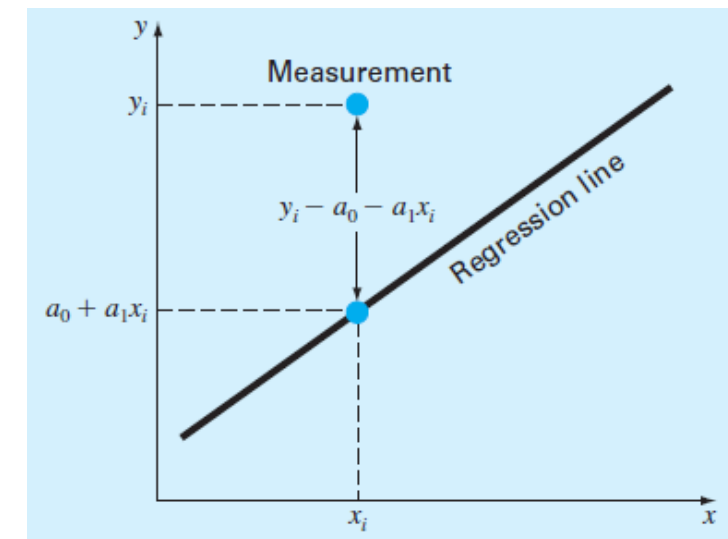
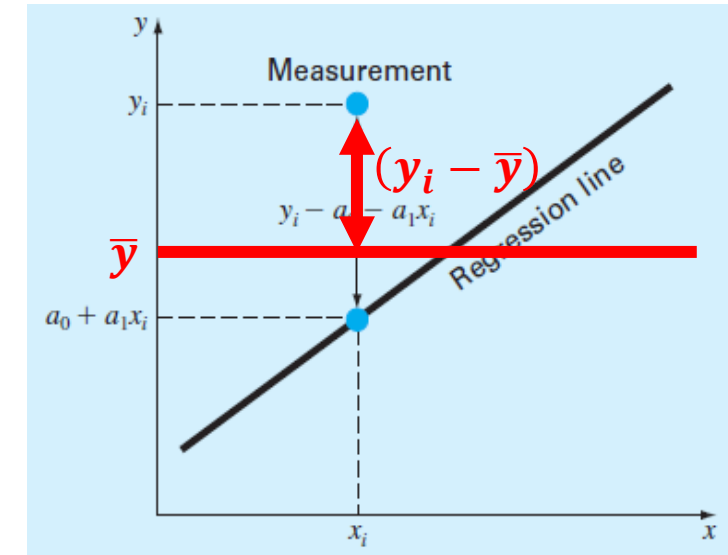
# Goodness of Fit

•  $S_t$ :

$$S_t = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (3)$$

•  $S_r$ :

$$S_r = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \quad (6)$$



# Goodness of Fit

- How can we compare the residuals before and after regression?

$$S_t - S_r$$

- This needs to be normalized w.r.t  $S_t$  because it's scale-dependent
- Coefficient of determination,  $R^2$ :

$$R^2 = \frac{S_t - S_r}{S_t} = 1 - \frac{S_r}{S_t} \quad (14)$$

# Goodness of Fit

- Coefficient of determination,  $R^2$ :

$$R^2 = \frac{S_t - S_r}{S_t} = 1 - \frac{S_r}{S_t} \quad (14)$$

- $R^2 = 0$ :  $S_t = S_r \rightarrow$  no improvement over mean
- $R^2 = 1$ :  $S_r = 0 \rightarrow$  ideal. Rarely happens in practice
- $R^2 < 0$ :  $S_r > S_t$ . Your model is worse than if you used the mean!

# Goodness of Fit

## THE correct $R^2$ interpretation

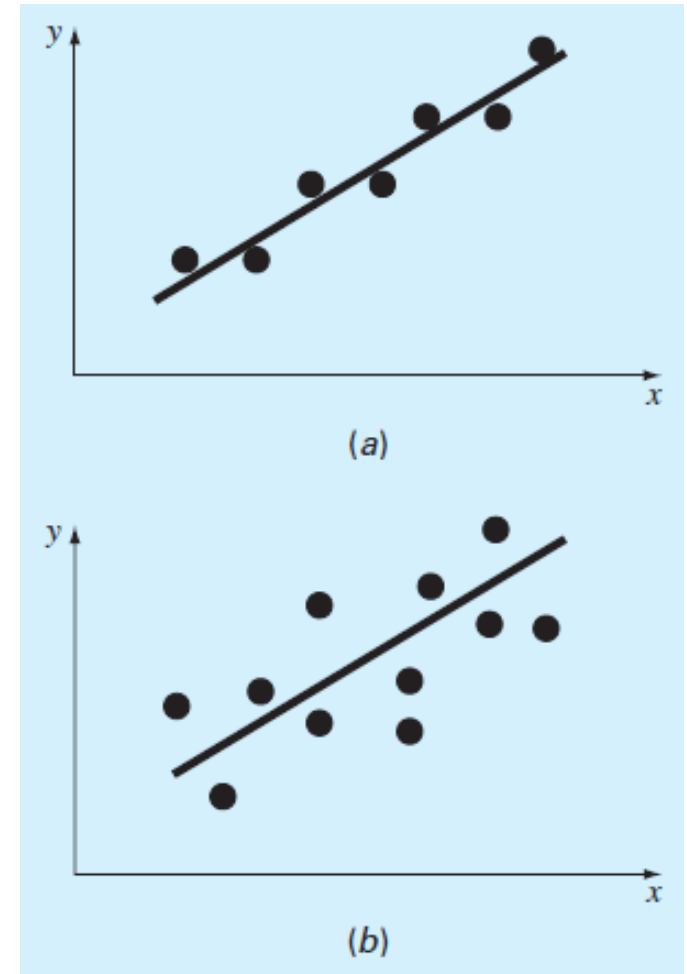
$$R^2 = \frac{\text{variance explained by the model}}{\text{total variance}}$$

## Some incorrect $R^2$ interpretations

- $R^2$  = number of original data points which pass through the regression line
- “I am [ $R^2 * 100$ ]% confident in my model”
- “My regression line is [ $R^2 * 100$ ]% accurate”

# Goodness of Fit

- (a) has a higher  $R^2$  than (b) because (a)'s residuals are smaller  $\rightarrow$  less variability in  $y$
- Ok...so we have:
  - Unique best-fit line equation
  - One quantitative way of assessing the best-fit line
- That's it! We're done with curve fitting!



## 2.2: Limitations and Warnings



# $R^2$ Limitations

$R^2 \approx 1$  DOES NOT  
NECESSARILY MEAN THE  
FIT IS “GOOD”

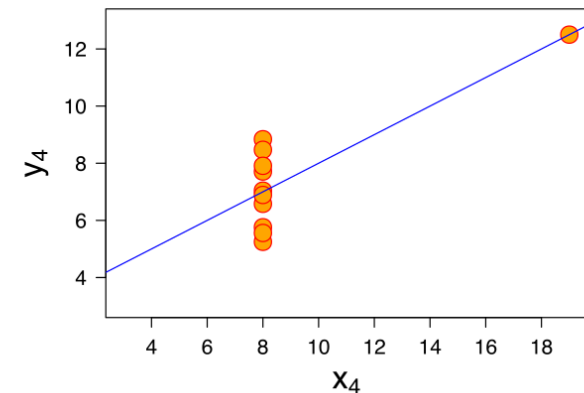
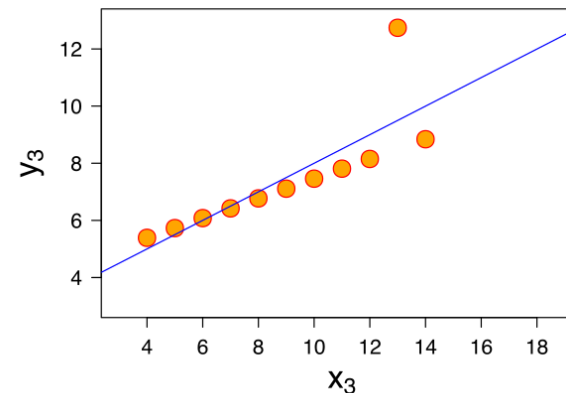
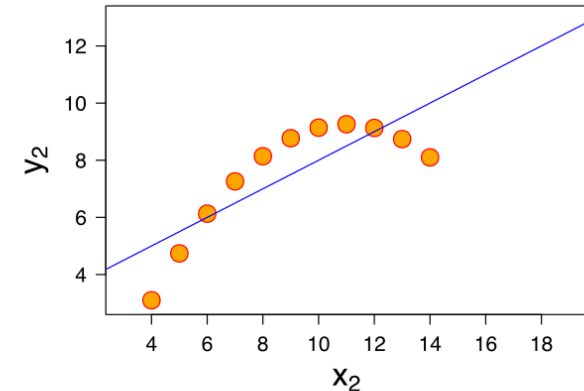
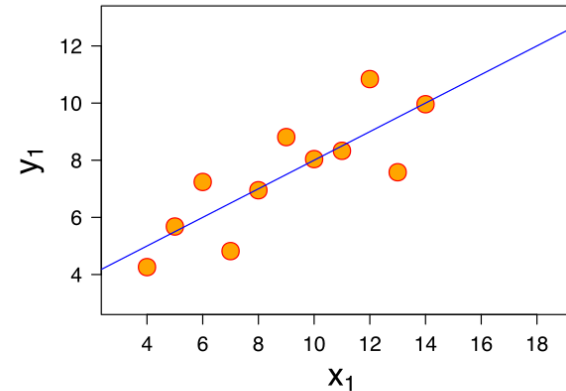


# $R^2$ Limitations

- *Anscombe's Quartet*: four datasets w/ nearly identical stats:

Property	Value	Accuracy
Mean of $x$	9	exact
Sample variance of $x : \sigma^2$	11	exact
Mean of $y$	7.50	to 2 decimal places
Sample variance of $y : \sigma^2$	4.125	$\pm 0.003$
Correlation between $x$ and $y$	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : $R^2$	0.67	to 2 decimal places

- Upper right isn't even linear!
- Bottom row's regressions are impacted by the outliers



# $R^2$ Limitations

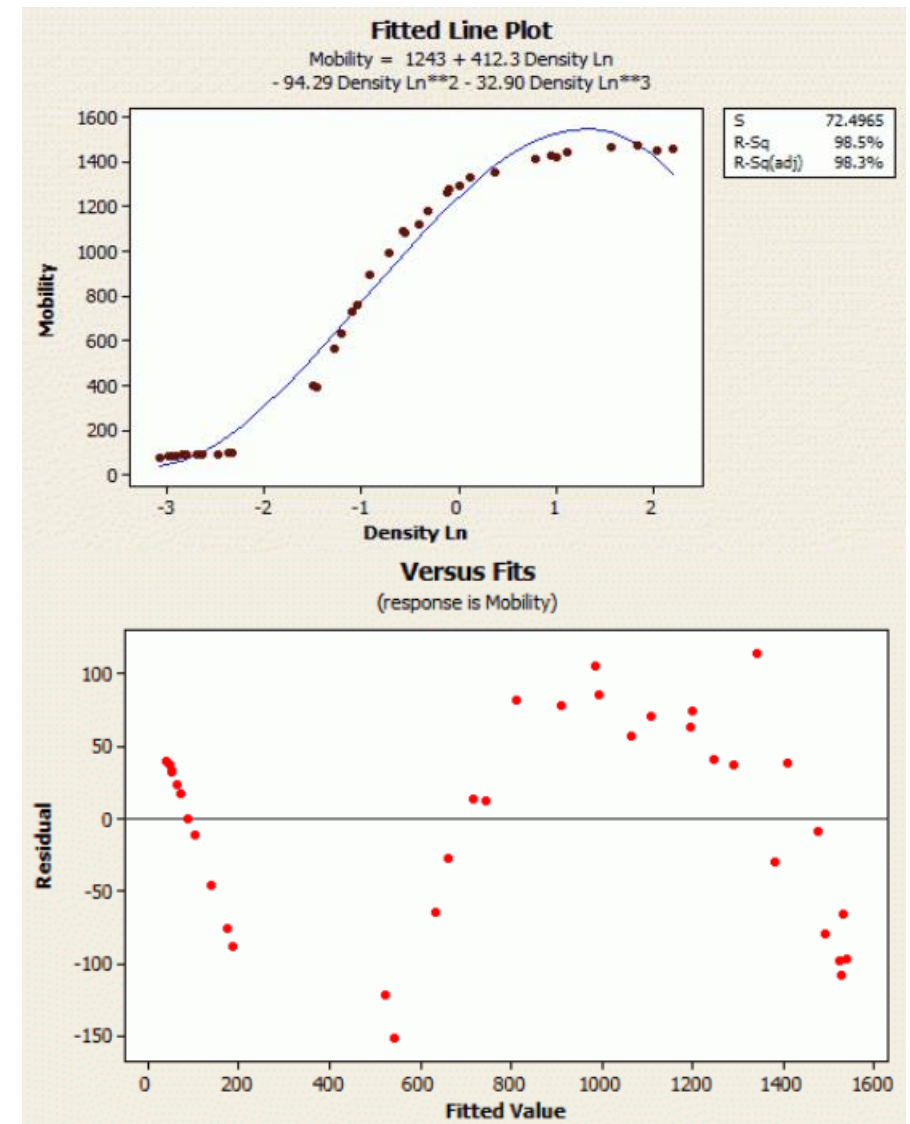
**Anscombe's quartet** comprises four [data sets](#) that have nearly identical simple [descriptive statistics](#), yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the [statistician Francis Anscombe](#) to demonstrate both the importance of graphing data before analyzing it and the effect of [outliers](#) and other [influential observations](#) on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."<sup>[1]</sup>

- Plot everything. Everything. Always. Plot even if we don't require a plot. Plot even if the pcode produces a plot.

← From the "WS3 Pointers" document on Canvas

# $R^2$ Limitations

- What if you have a low  $R^2$ ?
  - This isn't always bad
  - Context!!!!
- If your independent variables are *statistically significant*, you can still draw important conclusions about  $y$  vs.  $x$ 
  - More on this soon
- High  $R^2$  is a **necessary but insufficient condition** to ensure accurate, precise predictions



# $R^2$ Limitations

- How high does  $R^2$  need to be?
- What's your goal?
  - Understand the  $x - y$  relationship?
  - Predict  $y$ ?

# $R^2$ Limitations

- If your goal is to understand the  $x - y$  relationship:  **$R^2 = \text{irrelevant}$**

$$y = a_0 + a_1x + e \quad (1)$$

- Contextually interpreting  $a_1$  and  $a_0$  doesn't depend on  $R^2$ !
  - Assuming statistical significance
- *Instead, you should be asking:*
  - Can I trust the data?
  - Do the results fit the theory?
  - How do I interpret  $a_1$  and  $a_0$ ?

# $R^2$ Limitations


- If your goal is to predict  $y$ :  **$R^2 = \text{a consideration}$**
- Prediction implies a margin of error

$$R^2 = \frac{S_t - S_r}{S_t} = 1 - \frac{S_r}{S_t} \quad (14)$$

- *Instead, you should be asking:*
  - What are my **prediction intervals**?
  - Are my prediction intervals precise enough for my application?

# $R^2$ Limitations

- Plenty of additional goodness of fit metrics to use in tandem with  $R^2$
- MATLAB `fit()` documentation:


**gof — Goodness-of-fit statistics**  
 gof structure

Goodness-of-fit statistics, returned as the gof structure including the fields in this table.

Field	Value
sse	Sum of squares due to error
rsquare	R-squared (coefficient of determination)
dfe	Degrees of freedom in the error
adjrsquare	Degree-of-freedom adjusted coefficient of determination
rmse	Root mean squared error (standard error)

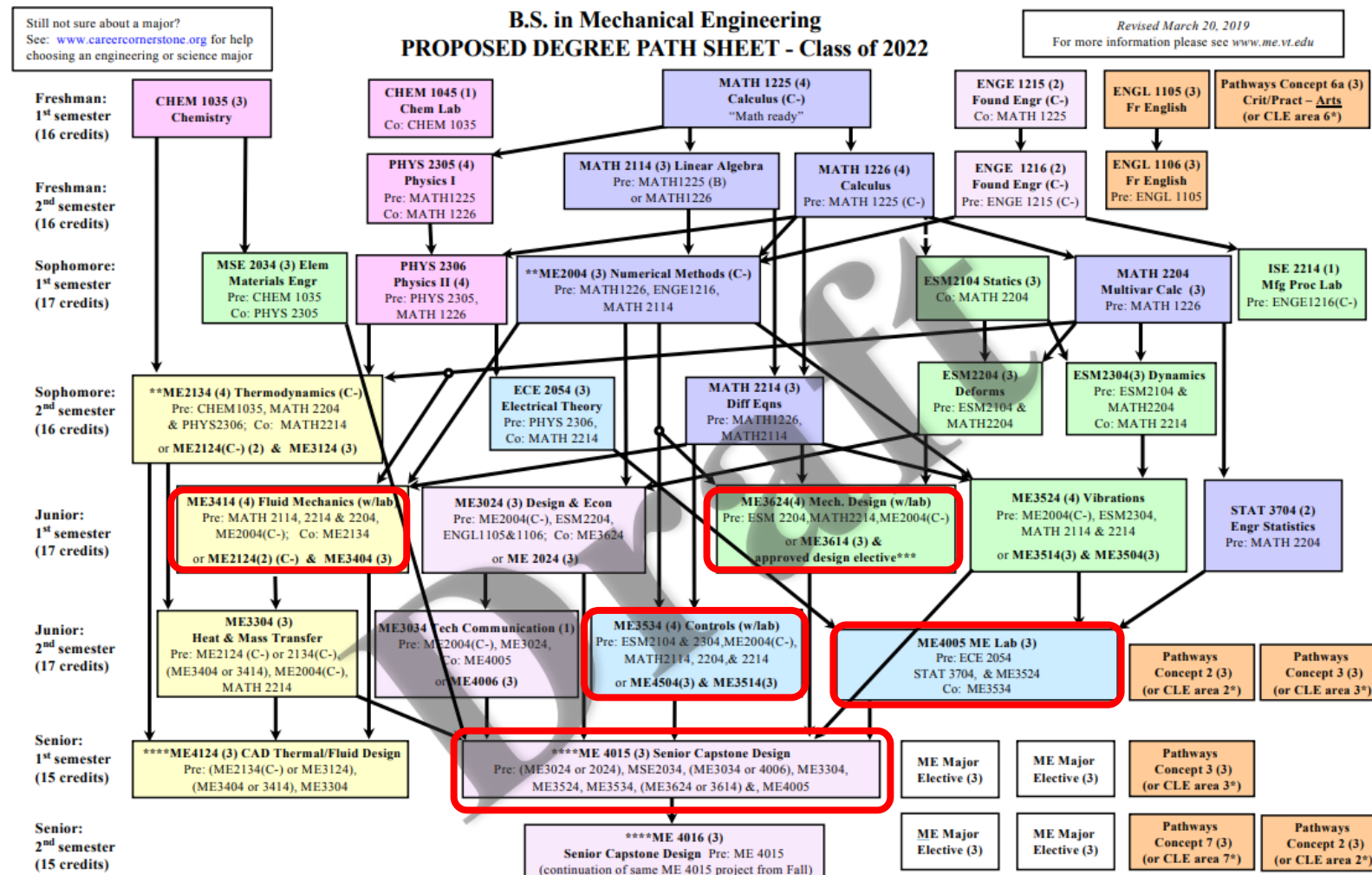
# Conclusion

ME 2004 Fall 2020 >

## Workshop 6

Visible: 08 Oct 2020 8:00 AM

## Midterm Exam





# Conclusion

- Least-squares regression produces a unique best-fit line
  - Solve for the coefficients  $(a_1, a_0 \dots)$  via linear algebra
- $R^2$  is one of many goodness of fit measures you can use to assess your model *in the context of your problem*
  - $R^2$  is not the be all, end all