Residual Analysis and Outliers

Lecture 47
Sections 13.4 - 13.5

Robb T. Koether

Hampden-Sydney College

Wed, Apr 16, 2008

Outline

Introduction

- We will look at a few issues related to linear regression.
- How can we tell whether the model should be linear rather than some other shape?
- What other types of regression are there?
- What are some of the pitfalls of linear regression?

The Appropriateness of the Linear Model

- We can learn a bit about the nature of the model by examining the residuals.
- This is called residual analysis.
- First, we need to find the residuals

$$e_i = y_i - \hat{y}_i.$$

The Appropriateness of the Linear Model

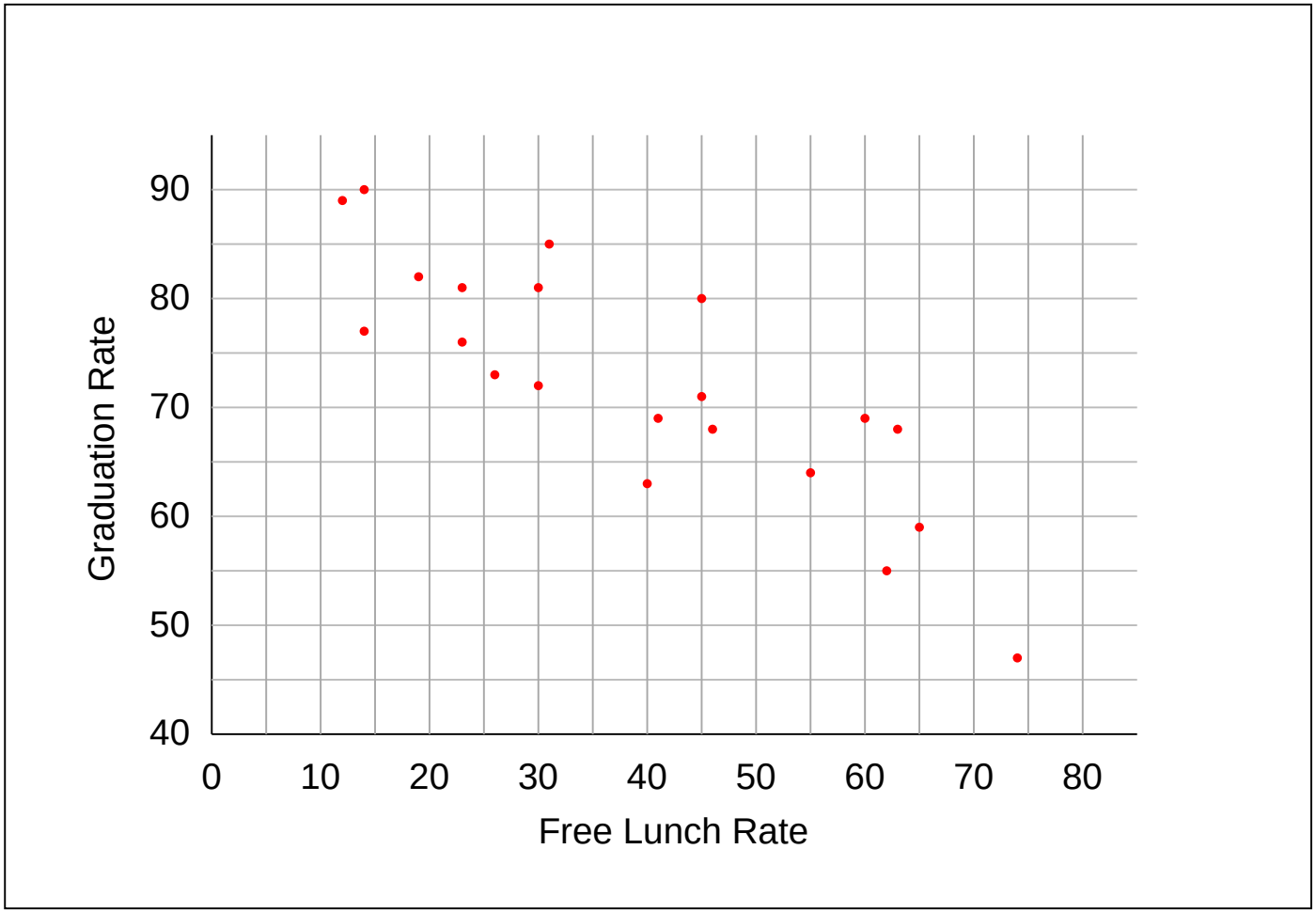
- To do this on the TI-83, after find the equation of the regression line, enter

  $$\texttt{L}_2 - \texttt{Y}_1\,(\texttt{L}_1) \longrightarrow \texttt{L}_3$$

  to store the residuals in $\texttt{L}_3$.
- Then draw a scatterplot of x versus e.

The Residual Plot

Free lunch rate vs. graduation rate



The Residual Plot

The residual plot

The Appropriateness of the Linear Model

- If the residual plot shows no clear pattern, then the linear model is appropriate.
- On the other hand, if the residual plot shows a distinct curvature, then the linear model may not be appropriate.

A Nonlinear Example

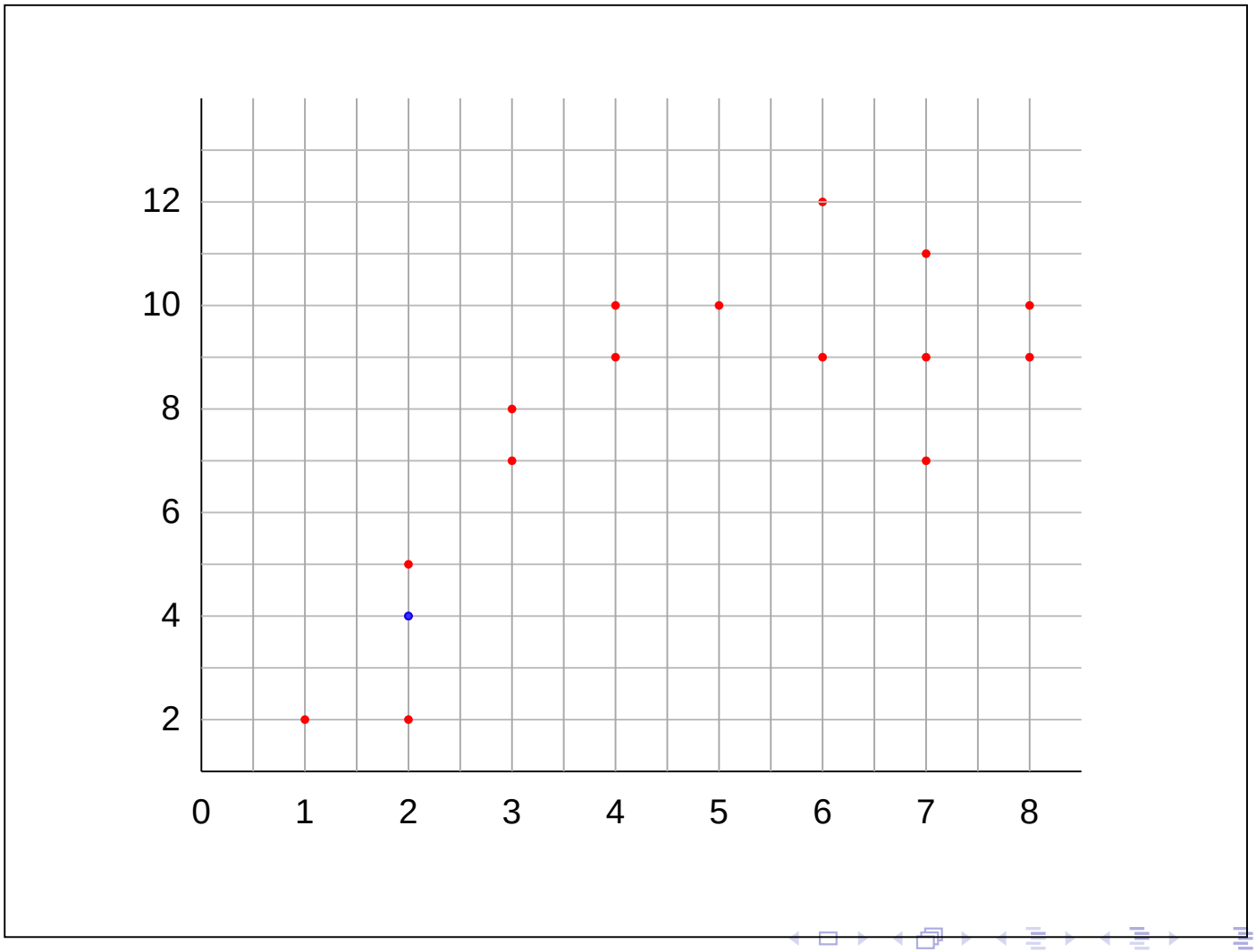
- Consider the following data.

x	y
1	2
2	2
2	4
2	4
2	5
3	7
3	8
4	9
4	10

x	y
5	12
6	9
6	12
7	7
7	9
7	11
8	9
8	10

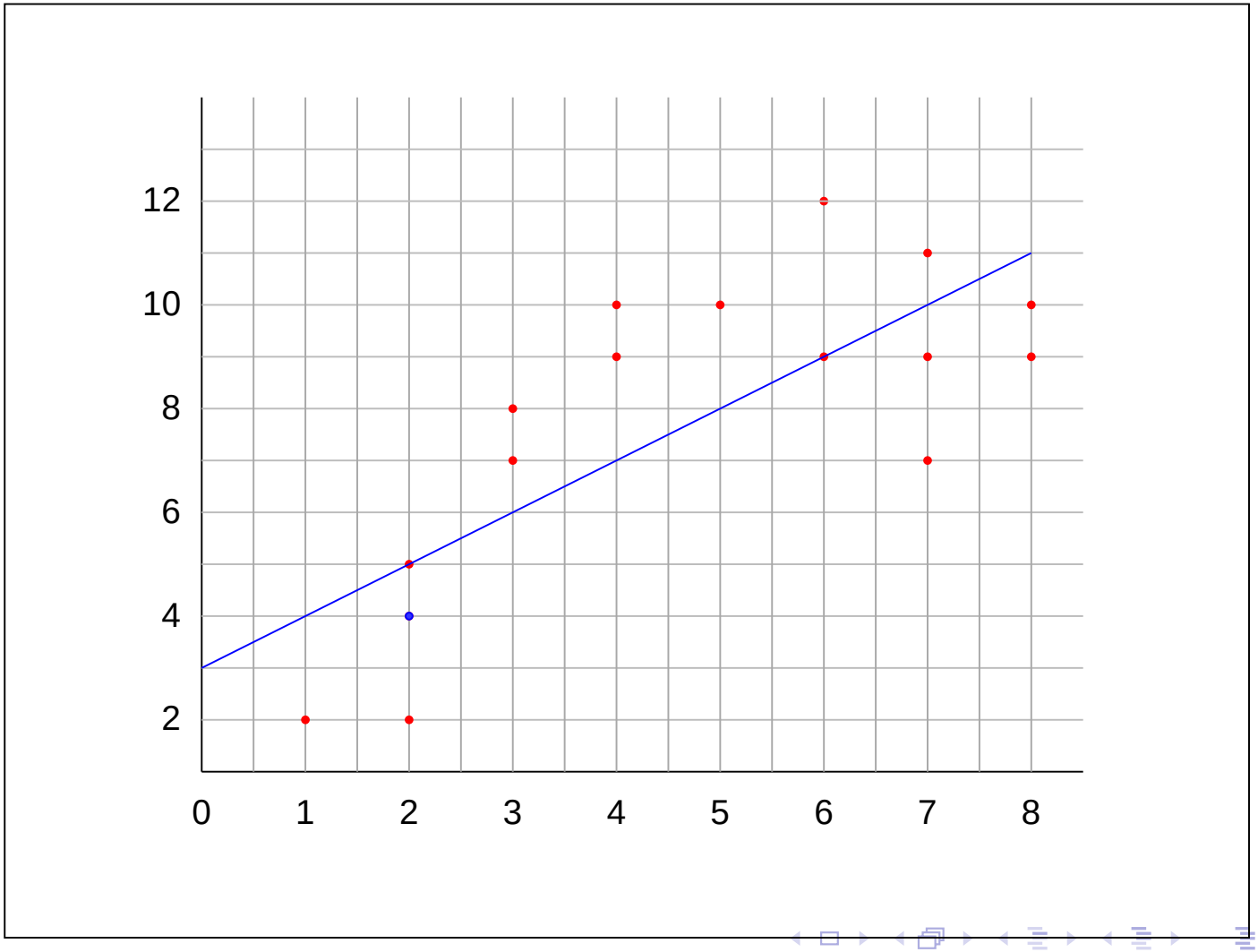
The Scatterplot

The scatterplot



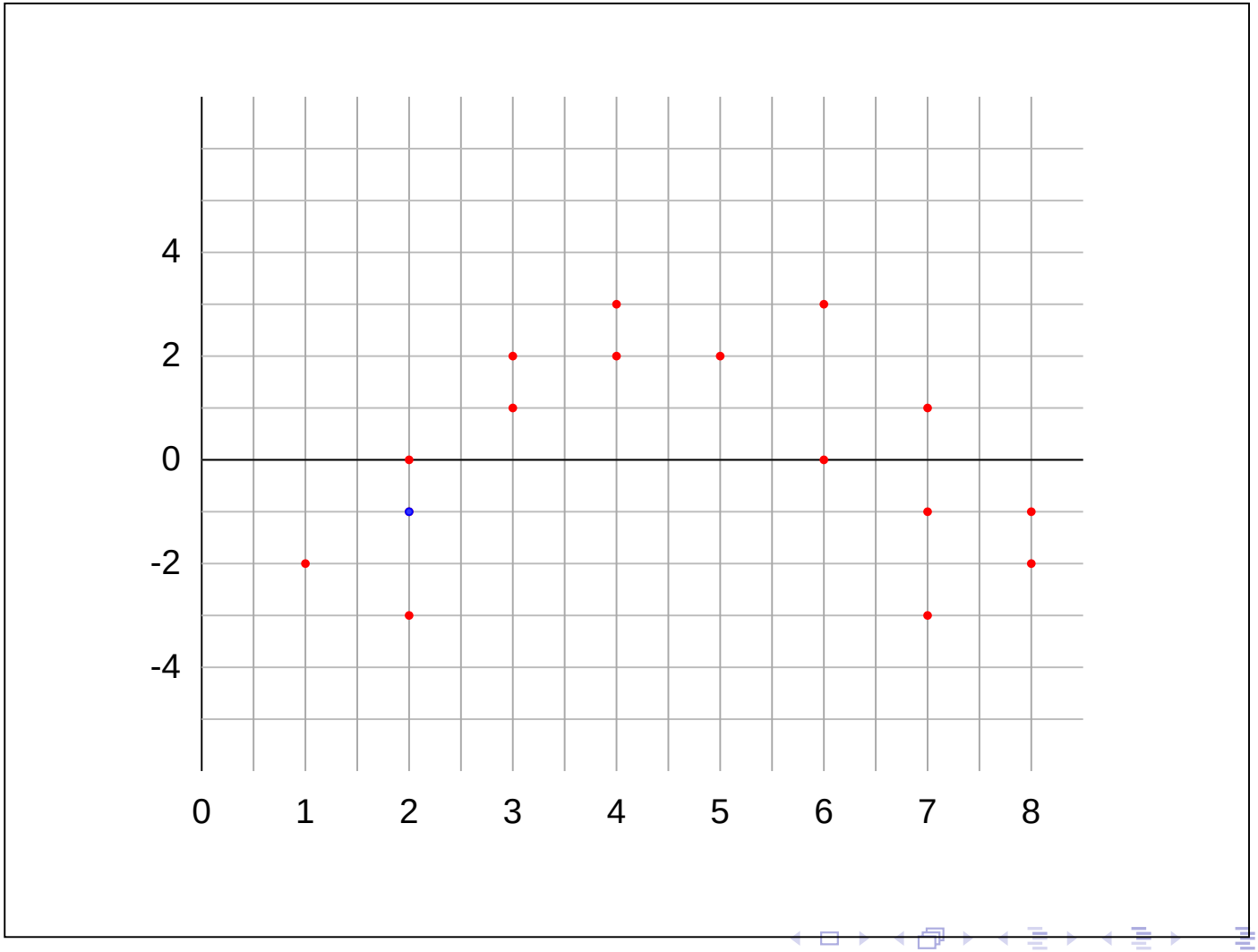
The Residual Plot

The regression line



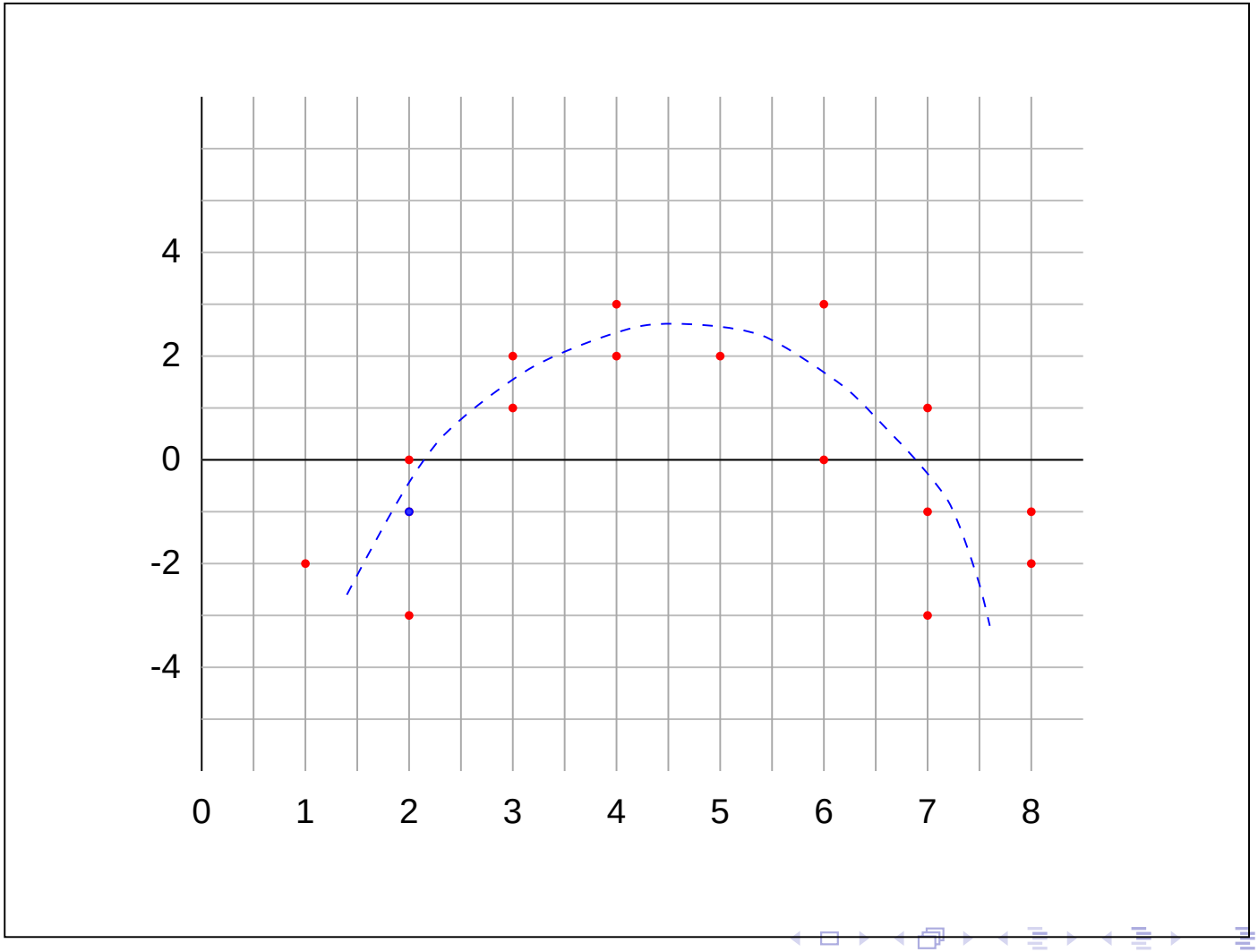
The Residual Plot

The residual plot

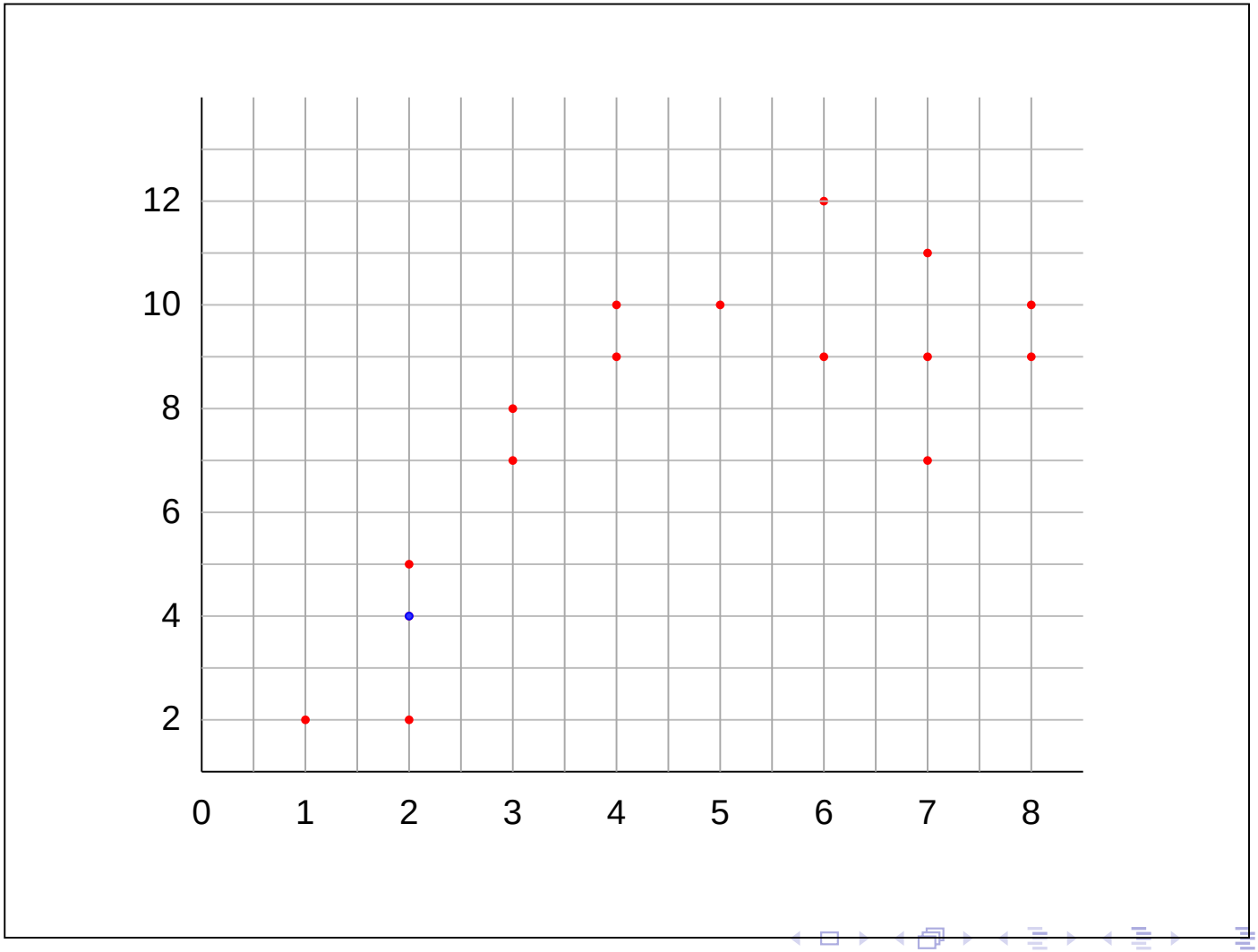


The Residual Plot

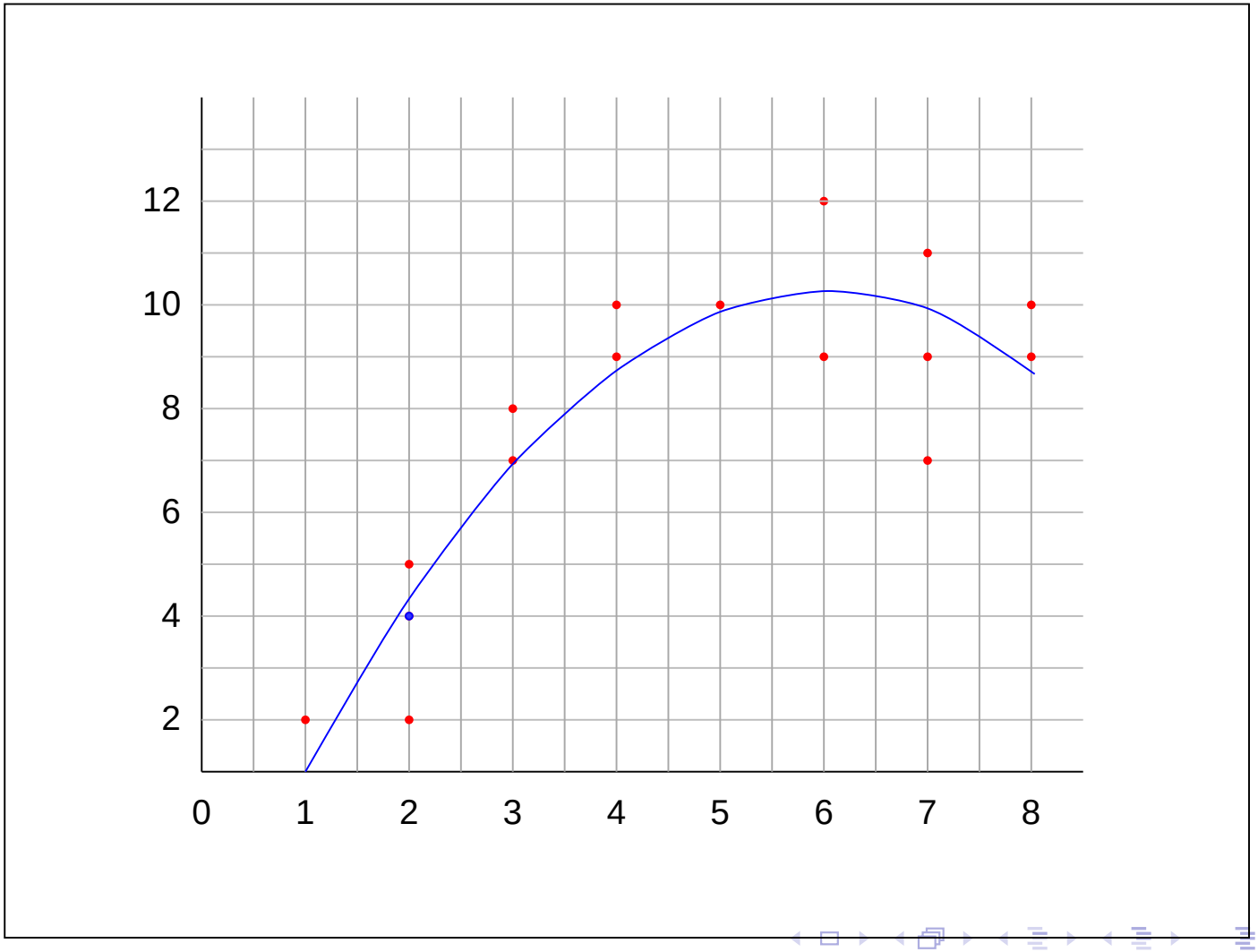
The residual plot



The Residual Plot



The Residual Plot



Nonlinear Regression on the TI-83

- The TI-83 will do a variety of nonlinear regressions.
- Press `STAT > CALC`.
- The list includes
  - LinReg - linear regression: $\hat{y} = a + bx$
  - QuadReg - quadratic regression: $\hat{y} = ax^2 + bx + c$
  - CubicReg - cubic regression: $\hat{y} = ax^3 + bx^2 + cx + d$
  - QuartReg - quartic regression: $\hat{y} = ax^4 + bx^3 + cx^2 + dx + e$
  - LnReg - logarithmic regression: $\hat{y} = a + b \ln x$
  - ExpReg - exponential regression: $\hat{y} = ab^x$
  - PwrReg - power regression: $\hat{y} = ax^b$
  - Logistic - logistic regression: $\hat{y} = \frac{c}{1+ae^{-bx}}$
  - SinReg - sinusoidal regression: $\hat{y} = a \sin(bx + c) + d$

Outliers

Definition (Outlier)

An outlier is a point with an unusually large residual (e.g., more than 2 standard deviations from 0).

Definition (Influential Point)

An influential point is a point that exerts a inordinate influence on the regression line.

- An outlier may or may not exert an inordinate influence on the regression line.
- An influential point may or may not be an outlier.

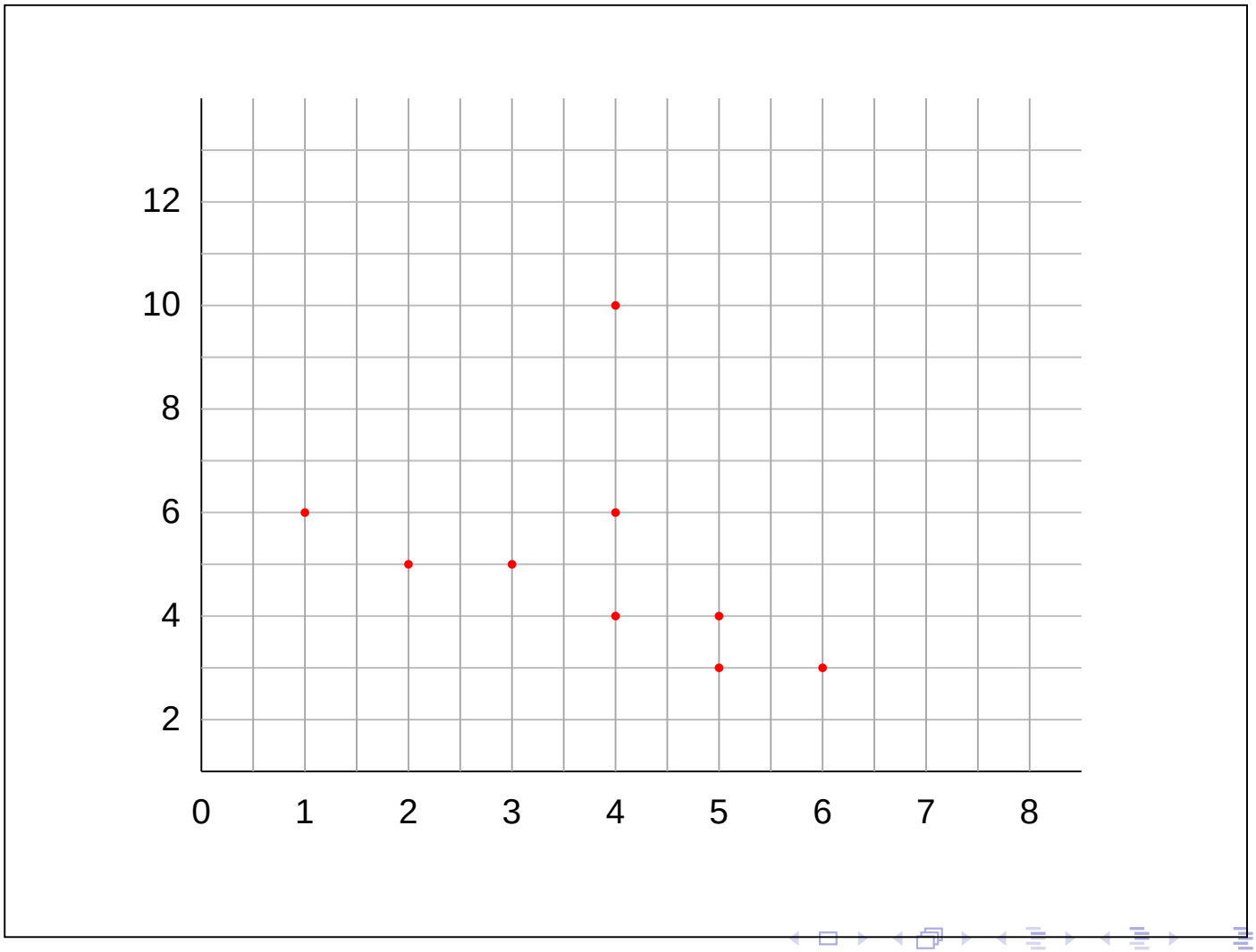
Outlier vs. Influential Point

- Consider the following data.

x	y
1	6
2	5
3	5
4	6
4	4
4	10
5	3
5	4
6	3

Outlier vs. Influential Point

The scatterplot



Outlier vs. Influential Point

- The regression line is $\hat{y} = 7.0 - 0.5x$.

x	y	$\hat{y}$	$y - \hat{y}$
1	6		
2	5		
3	5		
4	6		
4	4		
4	10		
5	3		
5	4		
6	3		

Outlier vs. Influential Point

- The regression line is $\hat{y} = 7.0 - 0.5x$.

x	y	$y - \hat{y}$	$y - \hat{y}$
1	6	6.5	-0.5
2	5	6.0	-1.0
3	5	5.5	-0.5
4	6	5.0	1.0
4	4	5.0	-1.0
4	10	5.0	5.0
5	3	4.5	-1.5
5	4	4.5	-0.5
6	3	4.0	-1.0

Outlier vs. Influential Point

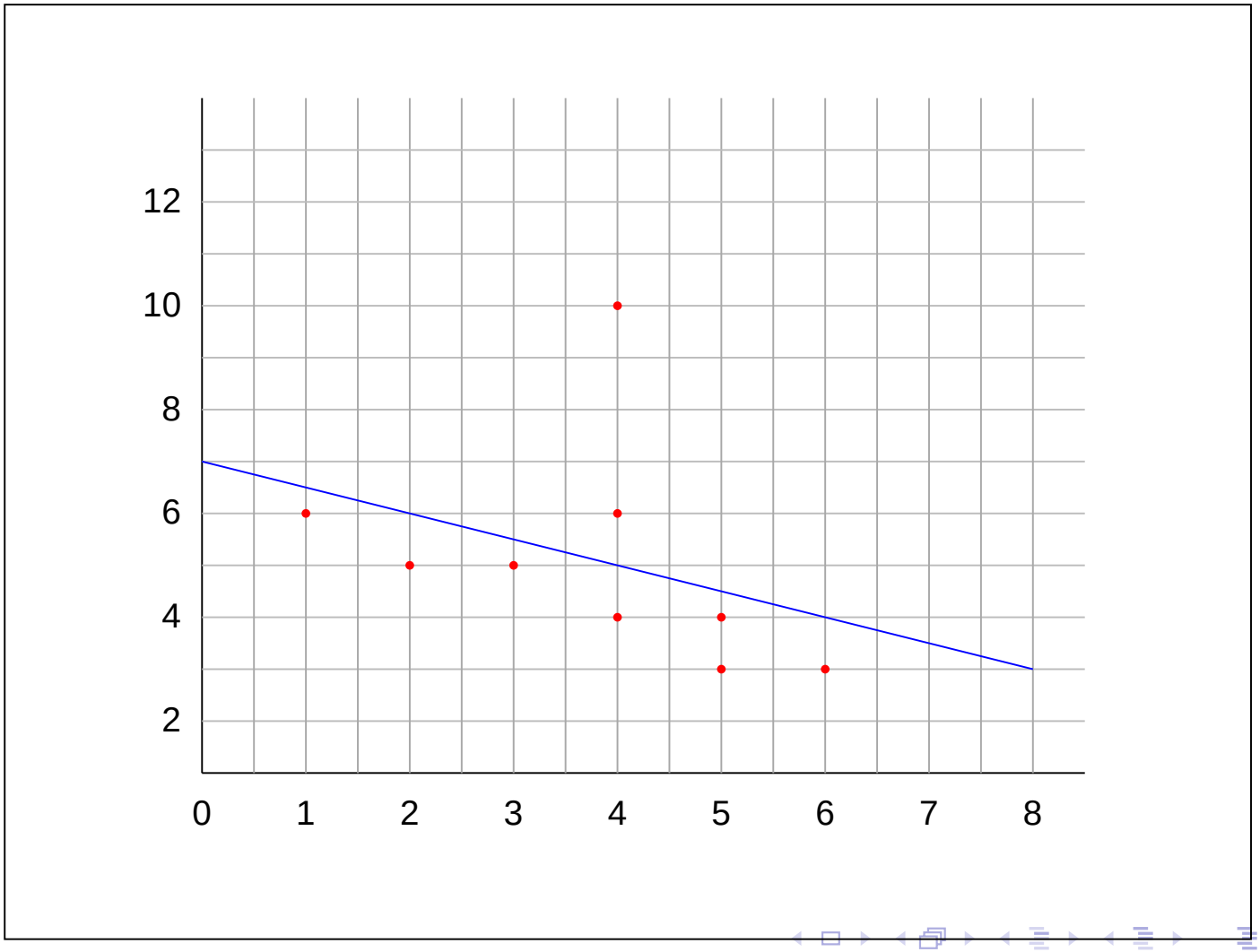
- The mean residual is 0.0 and the standard deviation of the residuals is 2.0.
- Thus, the residual 5.0 is 2.5 standard deviations above the mean, an outlier.
- But, it the point $(4, 10)$ influential?
- Remove it and see what the difference is.

Outlier vs. Influential Point

- The regression line of the remaining points is $\hat{y} = 6.615 - 0.564x$.
- This is nearly the same as $\hat{y} = 7.0 - 0.5x$.

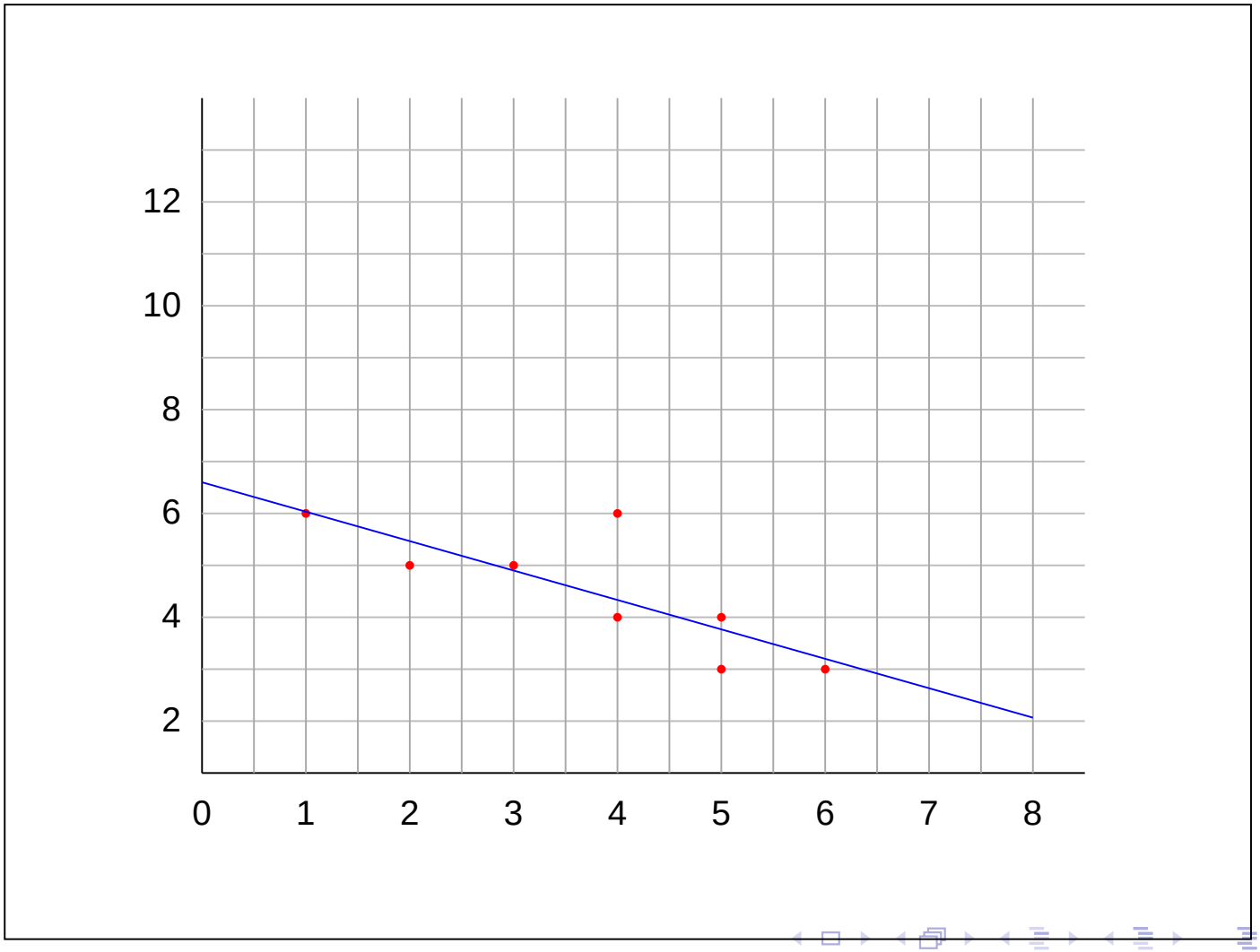
Outlier vs. Influential Point

Including the point $(4, 10)$



Outlier vs. Influential Point

Excluding the point $(4, 10)$

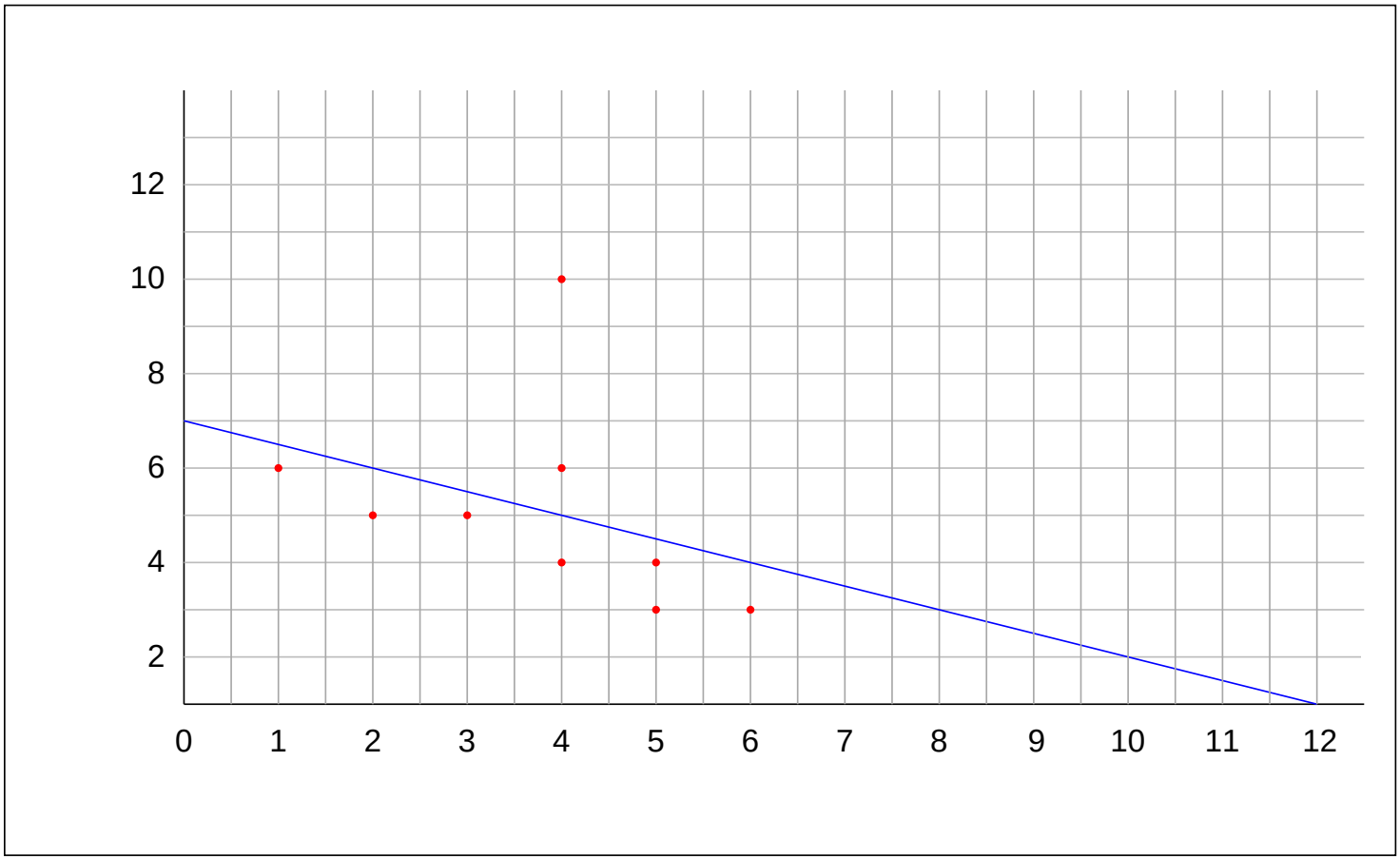


Outlier vs. Influential Point

- Now change the point $(4, 10)$ to the point $(12, 12)$.
- This changes the regression line to $\hat{y} = 2.767 + 0.55x$.

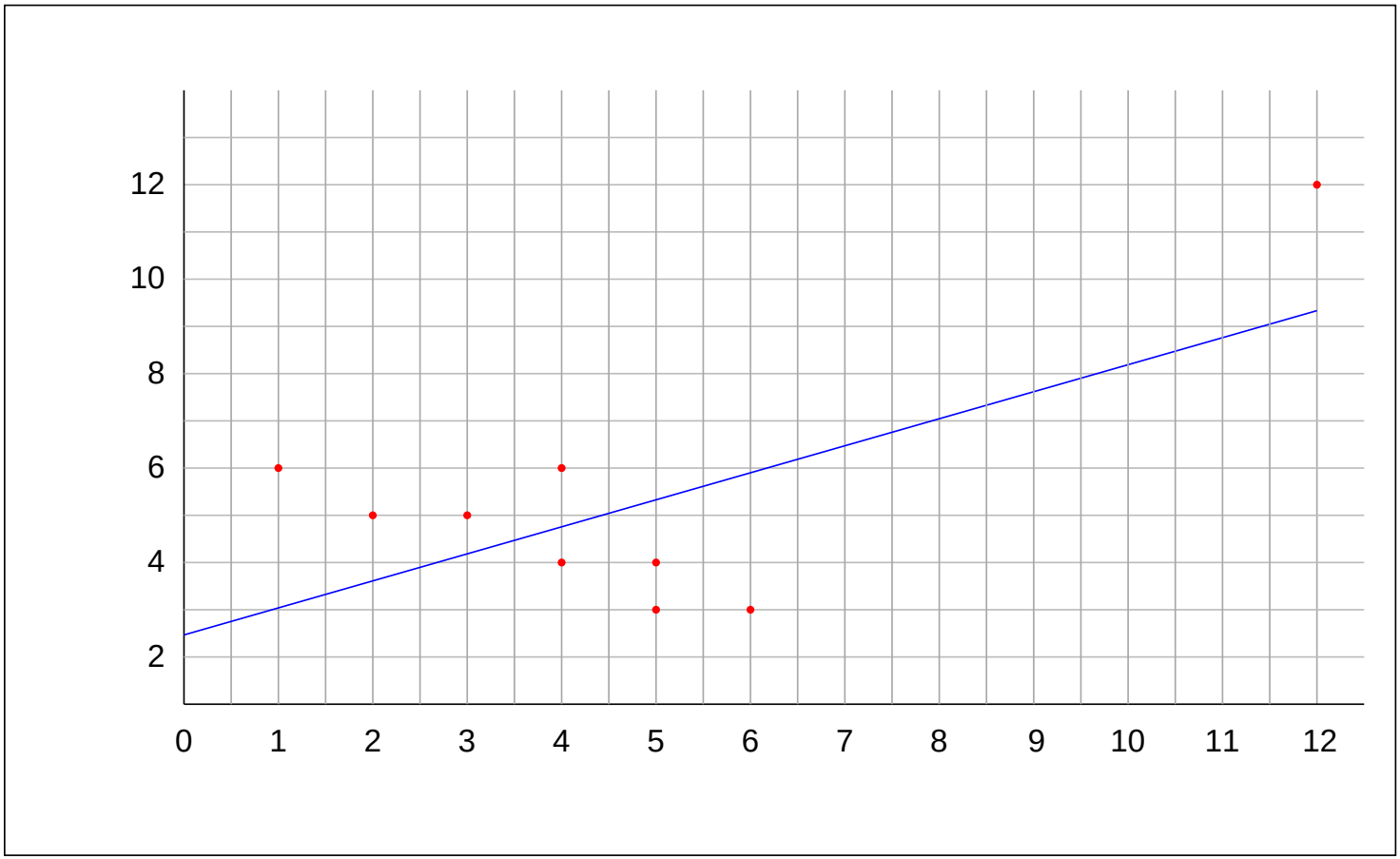
Outlier vs. Influential Point

Including the point $(4, 10)$



Outlier vs. Influential Point

Changing the point $(4, 10)$ to the point $(12, 12)$



Summary

Summary

-