Residual Analysis and Outliers

Lecture 48

Sections 13.4 - 13.5

Robb T. Koether

Hampden-Sydney College

Wed, Apr 8, 2009

Outline

Homework Review

Introduction

Residual Analysis

Nonlinear Regression

Outliers and Influential Points

Assignment

1. Homework Review
2. Introduction
3. Residual Analysis
4. Nonlinear Regression
5. Outliers and Influential Points
6. Assignment

Homework Review

Exercise 13.4, page 821.

The following data represent trends in cigarette consumption per capita (in hundreds) and lung cancer mortality (per 100,000) for Canadian males:

Cigarette Consumption (x)	11.8	12.5	15.7	19.2	21.9	23.3
Mortality Rate (y)	10.4	16.5	22.9	26.6	33.8	42.8

(b) Give the equation of the least squares regression line of $y =$ mortality rate on $x =$ cigarette consumption.

(c) Interpret the slope of the regression line. (Be specific.)

(d) Use the least squares regression equation to predict the lung cancer mortality rate when the cigarette consumption per capita is 2000.

Homework Review

Solution

(b)

- Enter the x values into list $\texttt{L}_1$ and the y values into $\texttt{L}_2$.
- Then use $\texttt{LinReg(a+bx) L}_1\texttt{,L}_2\texttt{,Y}_1$ to get the regression line.
- The line is

$$\hat{y} = -15.474 + 2.35x.$$

Homework Review

Solution

(c) The slope, 2.35, means that if x increases by 1, then y increases by 2.35. That is, the mortality rate increases by 2.35 deaths per 100,000 for every additional 100 cigarettes consumed.

(d) If cigarette consumption were 2000, the model predicts that the mortality rate would be

$$\hat{y}(20) = -15.474 + 2.35(20) = 31.6$$

lung cancer deaths per 100,000.

Introduction

- How do we know that a linear regression model is the best choice?
- What other types of regression are there?
- There are many other types.
- How many would you like?
- The linear model is by far the simplest, but it is not the only choice.

Introduction

- How do we know that a linear regression model is the best choice?
- What other types of regression are there?
- There are many other types.
- How many would you like?
- The linear model is by far the simplest, but it is not the only choice.

Introduction

- How do we know that a linear regression model is the best choice?
- What other types of regression are there?
- There are many other types.
- How many would you like?
- The linear model is by far the simplest, but it is not the only choice.

Introduction

- How do we know that a linear regression model is the best choice?
- What other types of regression are there?
- There are many other types.
- How many would you like?
- The linear model is by far the simplest, but it is not the only choice.

Introduction

- How do we know that a linear regression model is the best choice?
- What other types of regression are there?
- There are many other types.
- How many would you like?
- The linear model is by far the simplest, but it is not the only choice.

TI-83 - Nonlinear Regression

TI-83 Nonlinear Regression

- 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:
$$\hat{y} = ax^3 + bx^2 + cx + d.$$

TI-83 - Nonlinear Regression

TI-83 Nonlinear Regression

- And...
  - `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:

$$\hat{y} = ab^x.$$

TI-83 - Nonlinear Regression

TI-83 Nonlinear Regression

- And...
  - `PwrReg` - Power regression:

$$\hat{y} = ax^b.$$

  - `Logistic` - Logistic regression:

$$\hat{y} = \frac{c}{1 + ae^{-bx}}.$$

  - `SinReg` - Sinusoidal regression:

$$\hat{y} = a \sin(bx + c) + d.$$

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

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

- Then we draw a scatterplot of x versus e and see whether there is a pattern.

The Appropriateness of the Linear Model

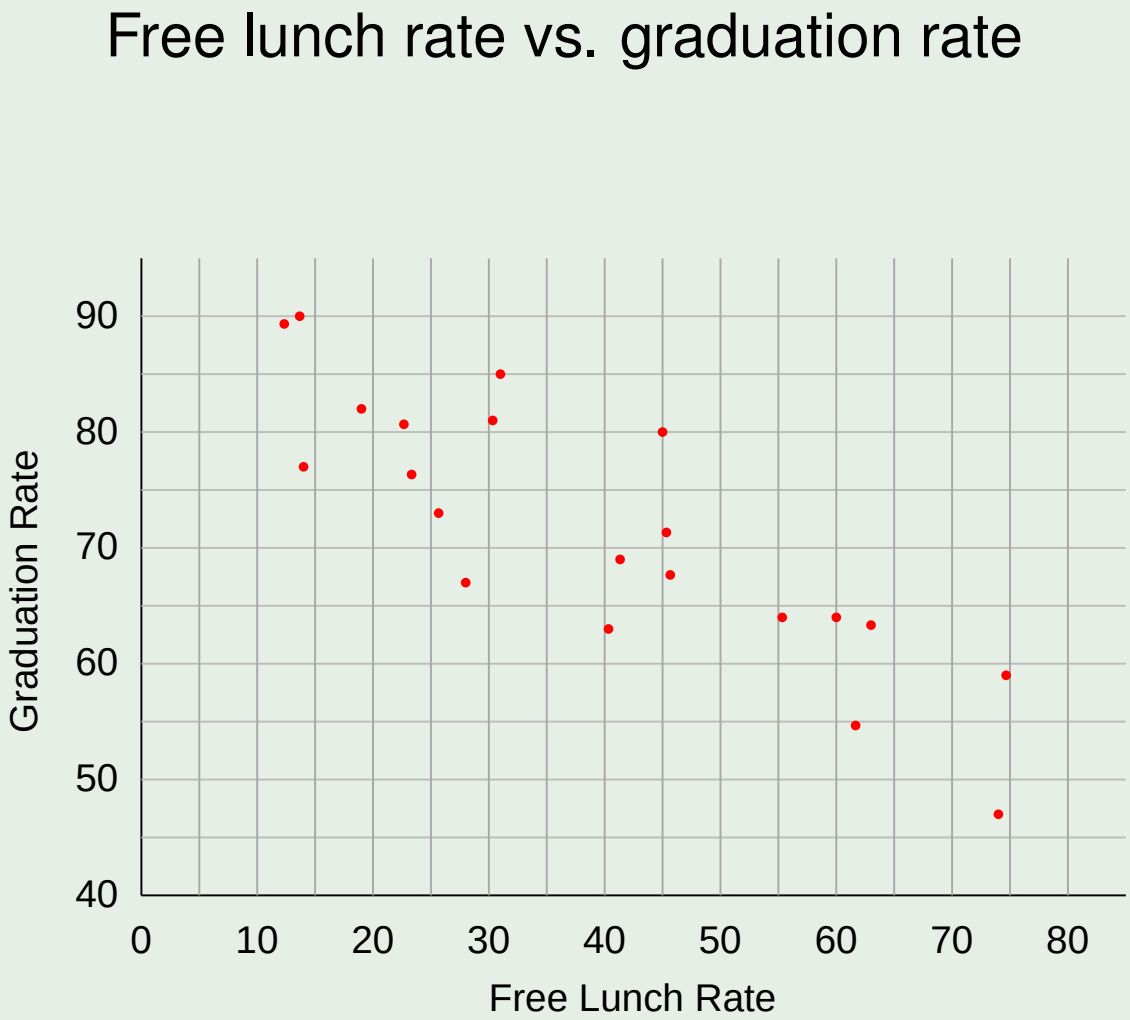
- To do this on the TI-83, after finding the equation of the regression line, enter

$$\texttt{L}_2 - \texttt{Y}_1\,(\texttt{L}_1) \longrightarrow \texttt{L}_4$$

  to store the residuals in $\texttt{L}_4$.
- Then draw a scatterplot of $\texttt{L}_1$ versus $\texttt{L}_4$.

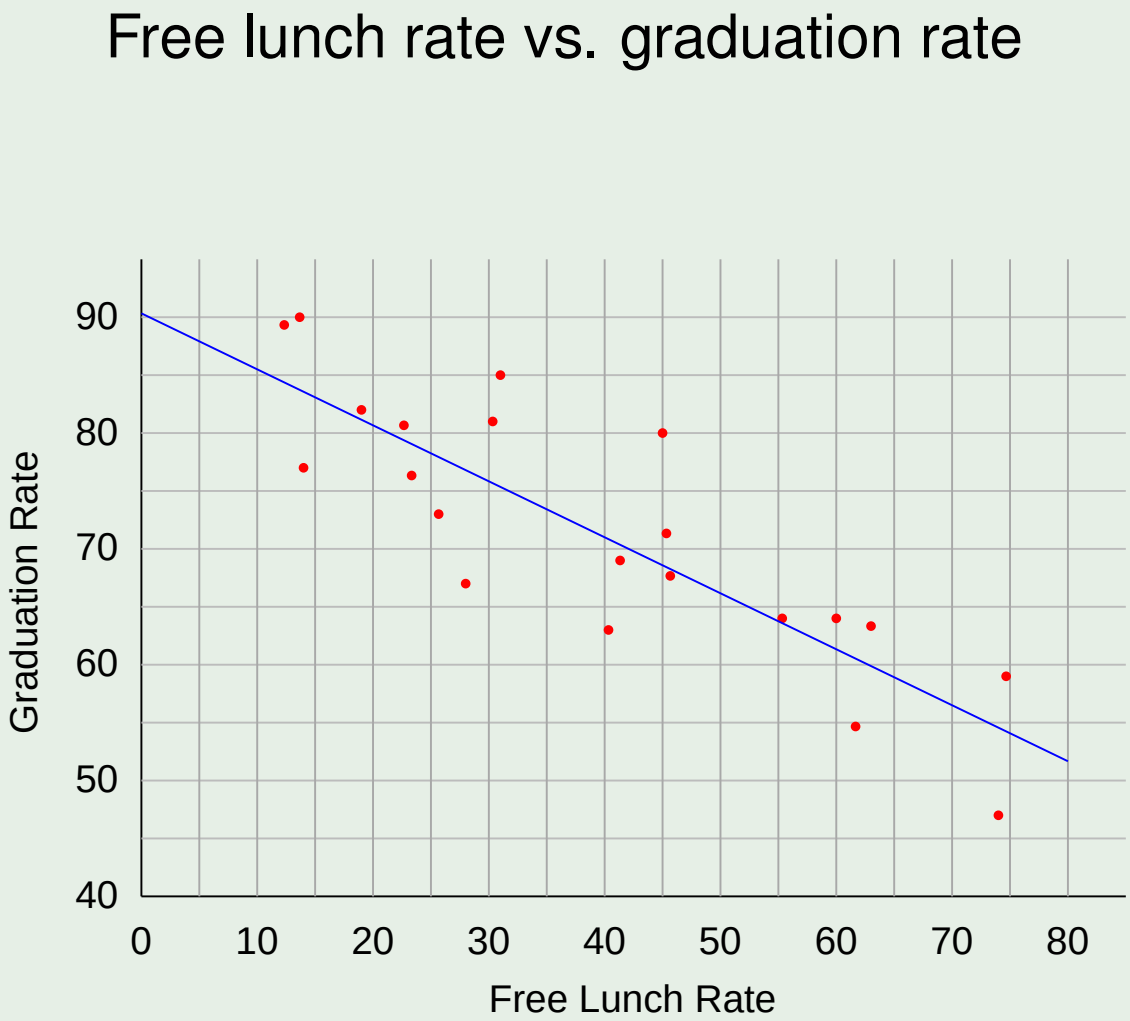
The Residual Plot

Example (Residual Plots)



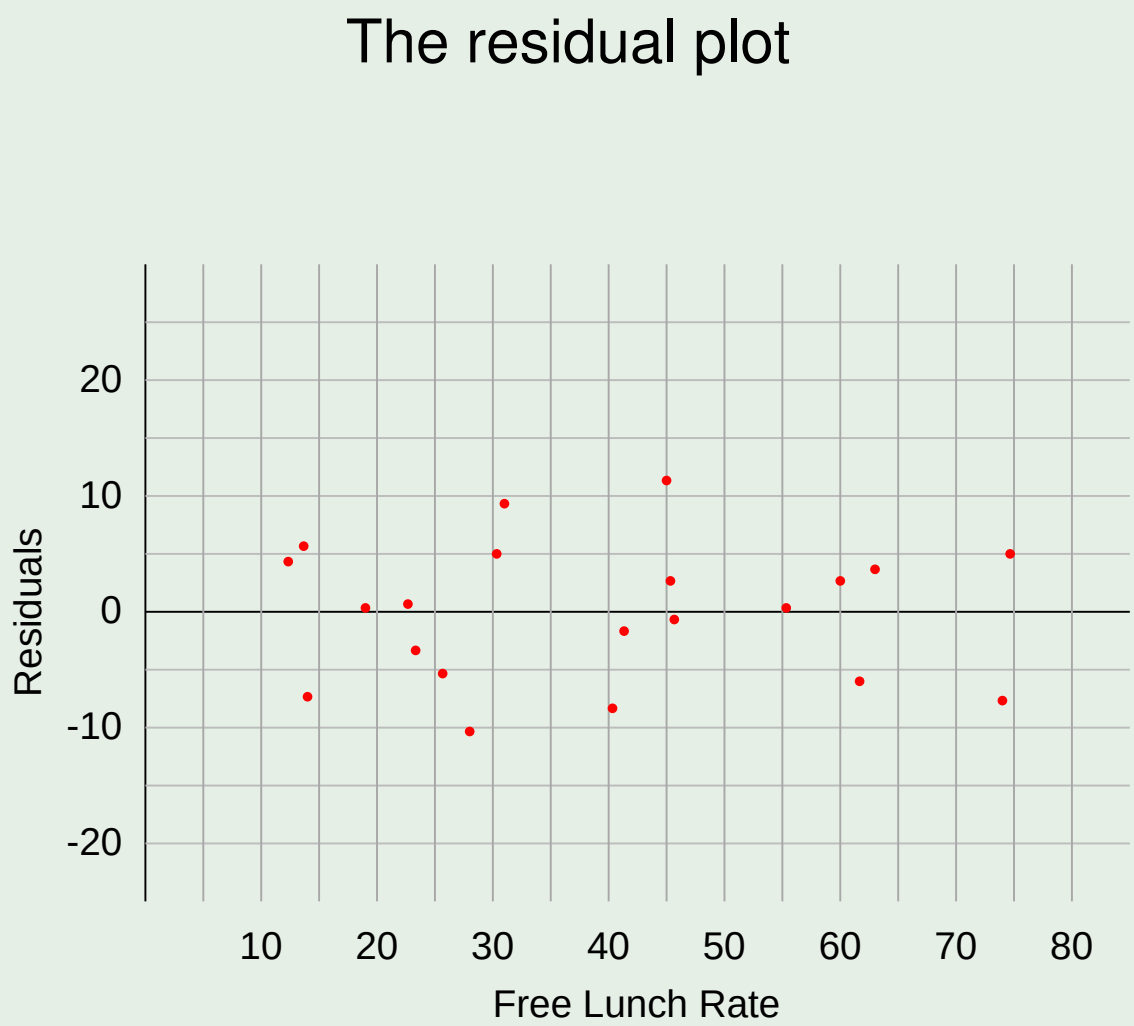
The Residual Plot

Example (Residual Plots)



The Residual Plot

Example (Residual Plots)



The Appropriateness of the Linear Model

- If the residual plot shows no clear pattern, but just a big blob of points, 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 Model

Example (A Nonlinear Model)

- Consider the following data.

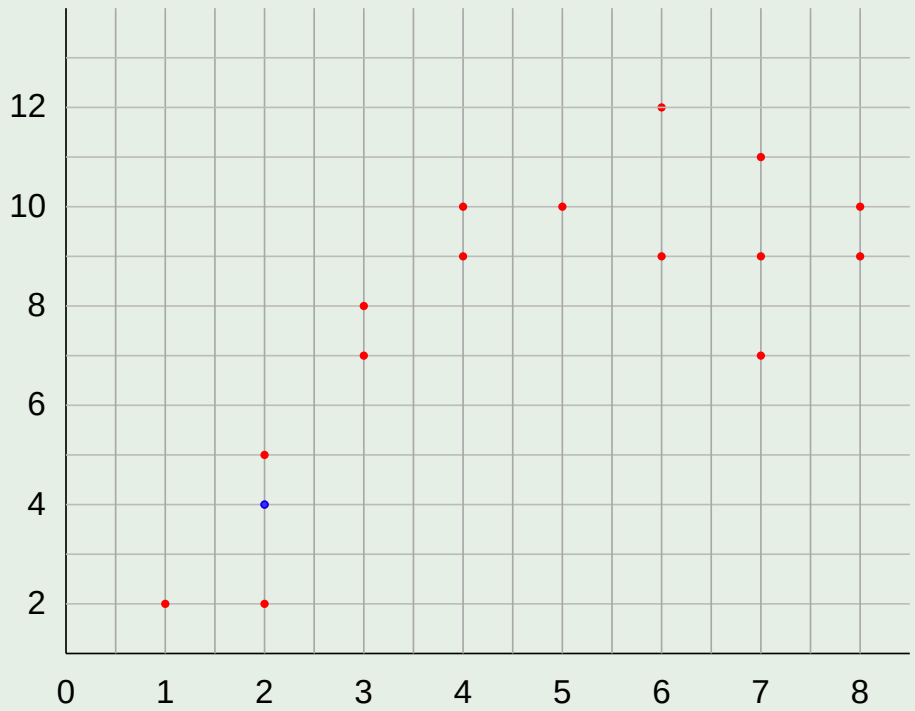
x	y
1	2
2	2
2	4
2	4
2	5
3	7
3	8
4	9
4	10

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

A Nonlinear Model

Example (A Nonlinear Model)

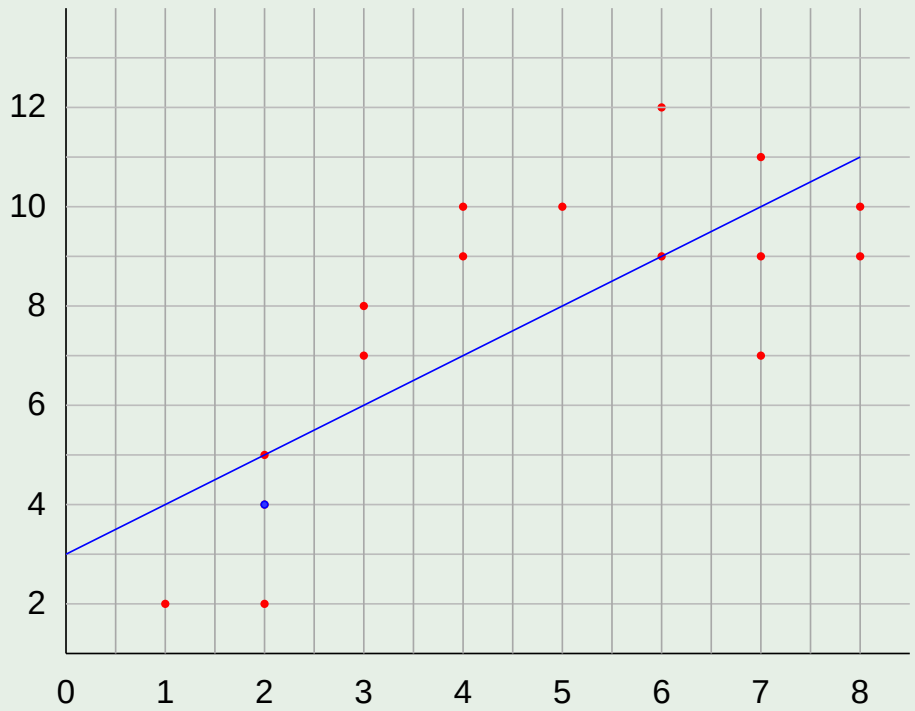
The scatterplot



A Nonlinear Model

Example (A Nonlinear Model)

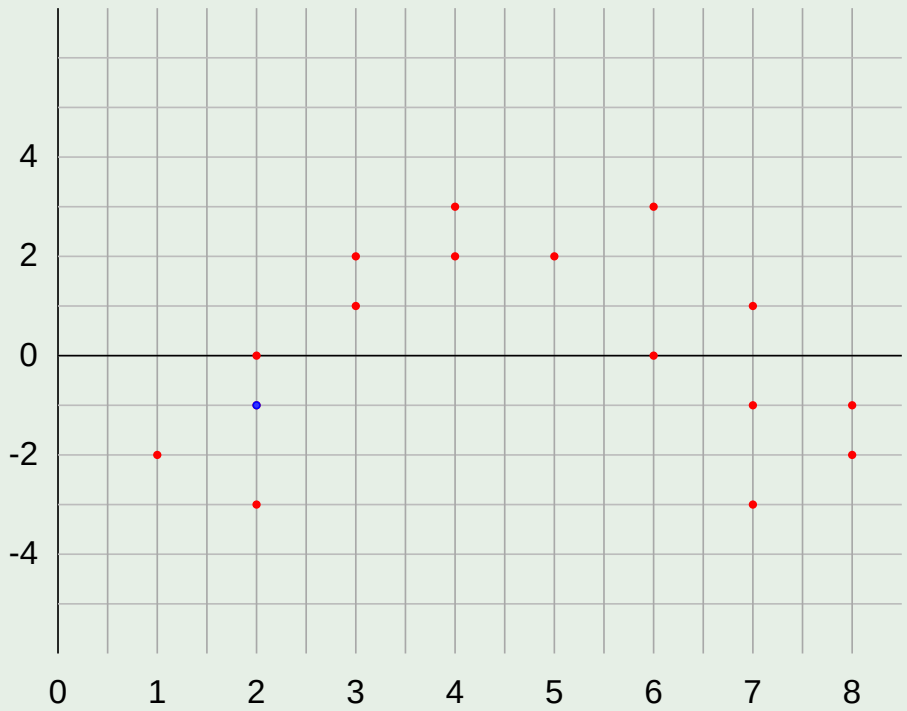
The regression line



A Nonlinear Model

Example (A Nonlinear Model)

The residual plot



A Nonlinear Model

Residual Analysis and Outliers

Robb T. Koether

Homework Review

Introduction

Residual Analysis

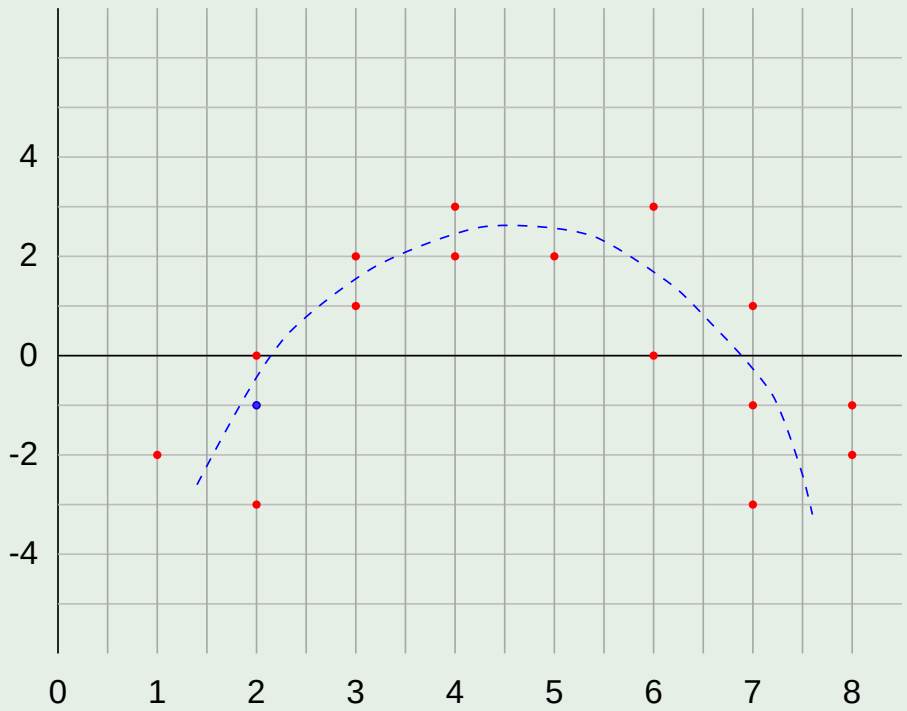
Nonlinear Regression

Outliers and Influential Points

Assignment

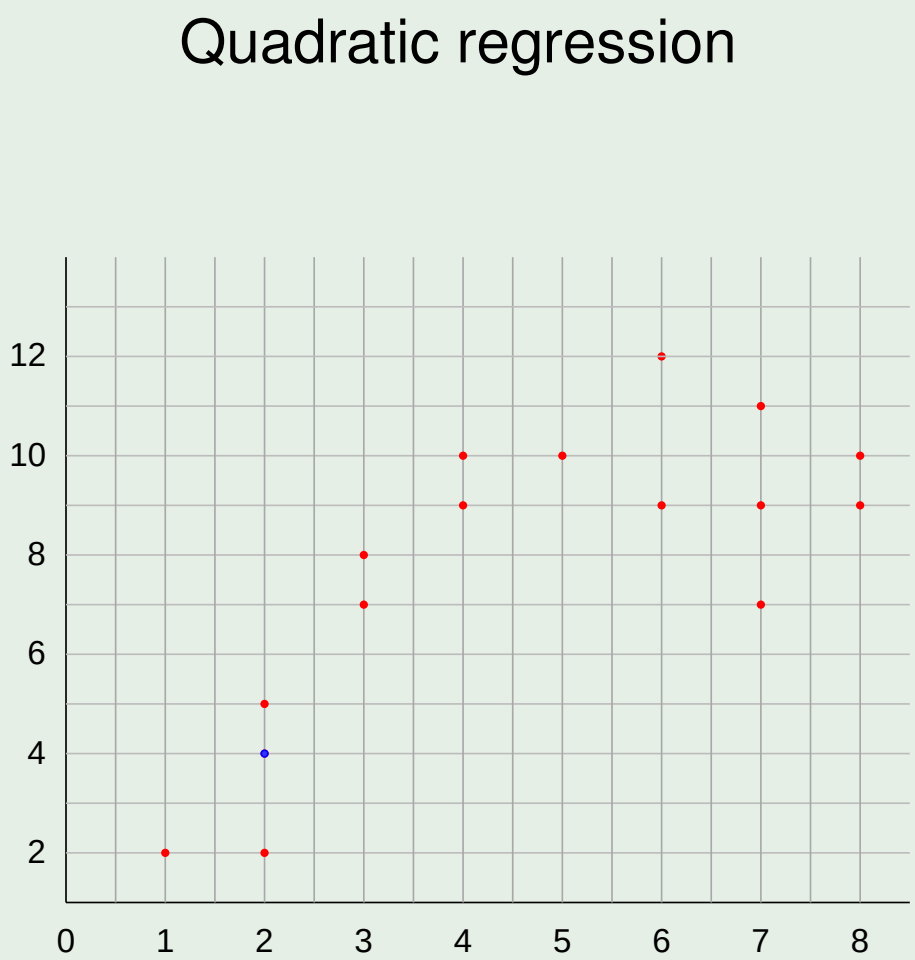
Example (A Nonlinear Model)

The residual plot



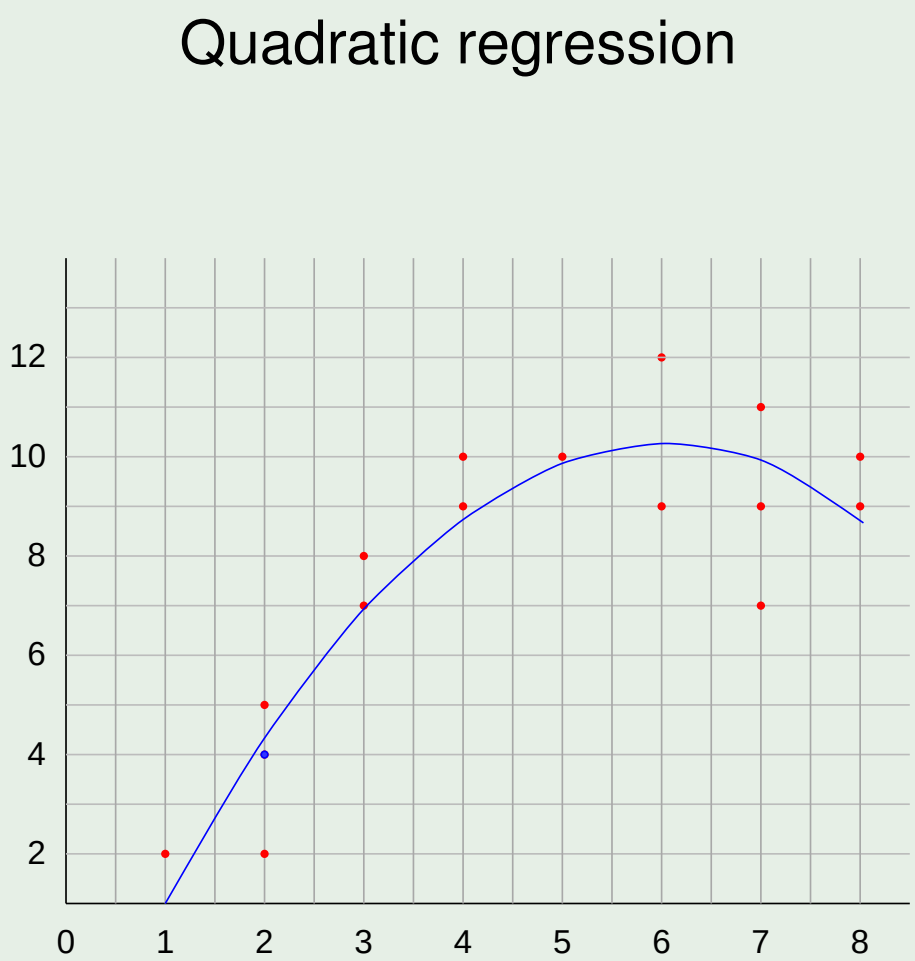
A Nonlinear Model

Example (A Nonlinear Model)



A Nonlinear Model

Example (A Nonlinear Model)



Outliers

Definition (Outlier)

An outlier is a point with an unusually large residual (e.g., at least 2.5 standard deviations from the mean).

Definition (Influential Point)

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

Outliers

- An outlier may or may not be influential.
- An influential point may or may not be an outlier.

Outliers and Influential Points

Example (Outliers and Influential Points)

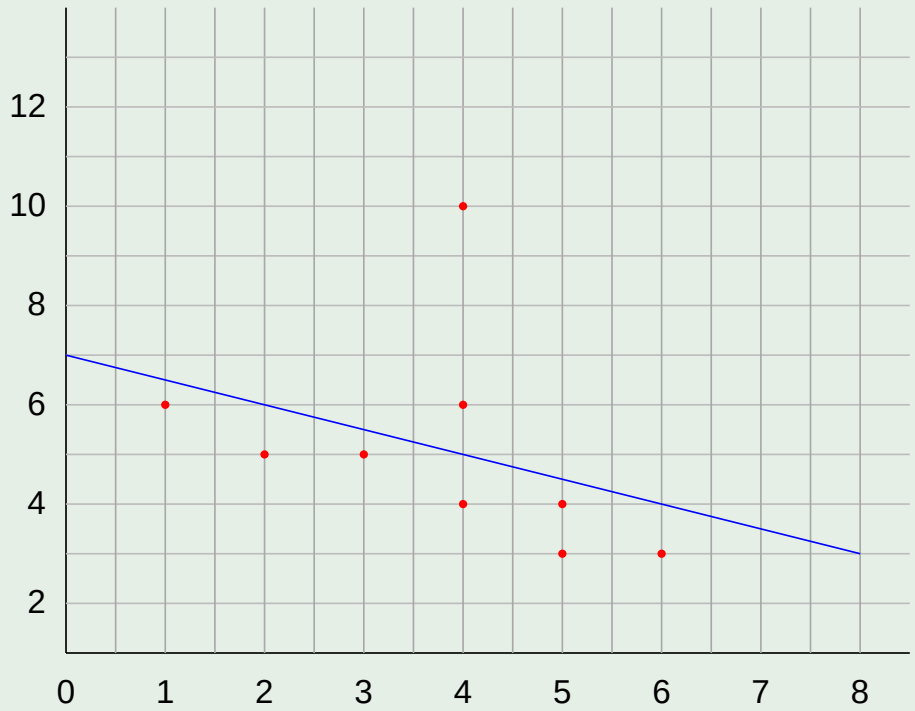
- Consider the following data.

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

Outliers and Influential Points

Example (Outliers and Influential Points)

The scatterplot



Outliers and Influential Points

Example (Outliers and Influential Points)

- 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		

Outliers and Influential Points

Example (Outliers and Influential Points)

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

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

Outliers and Influential Points

Example (Outliers and Influential Points)

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

x	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

Outliers and Influential Points

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

Outliers and Influential Points

- The regression line of the remaining points is

$$\hat{y} = 6.615 - 0.564x.$$

- This is nearly the same as

$$\hat{y} = 7.0 - 0.5x.$$

Outliers and Influential Points

Residual Analysis and Outliers

Robb T. Koether

Homework Review

Introduction

Residual Analysis

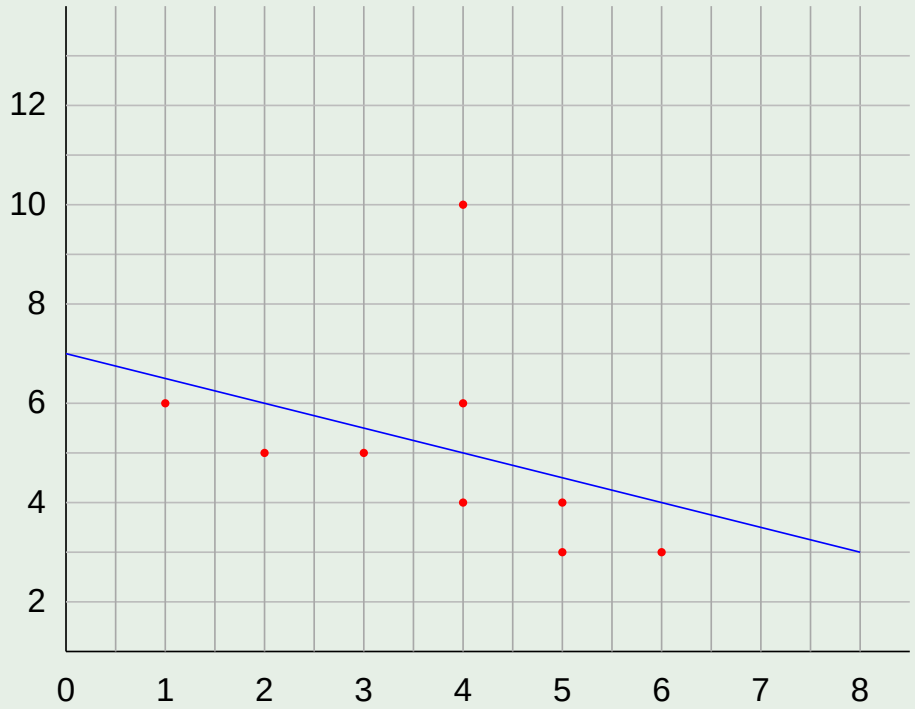
Nonlinear Regression

Outliers and Influential Points

Assignment

Example (Outliers and Influential Points)

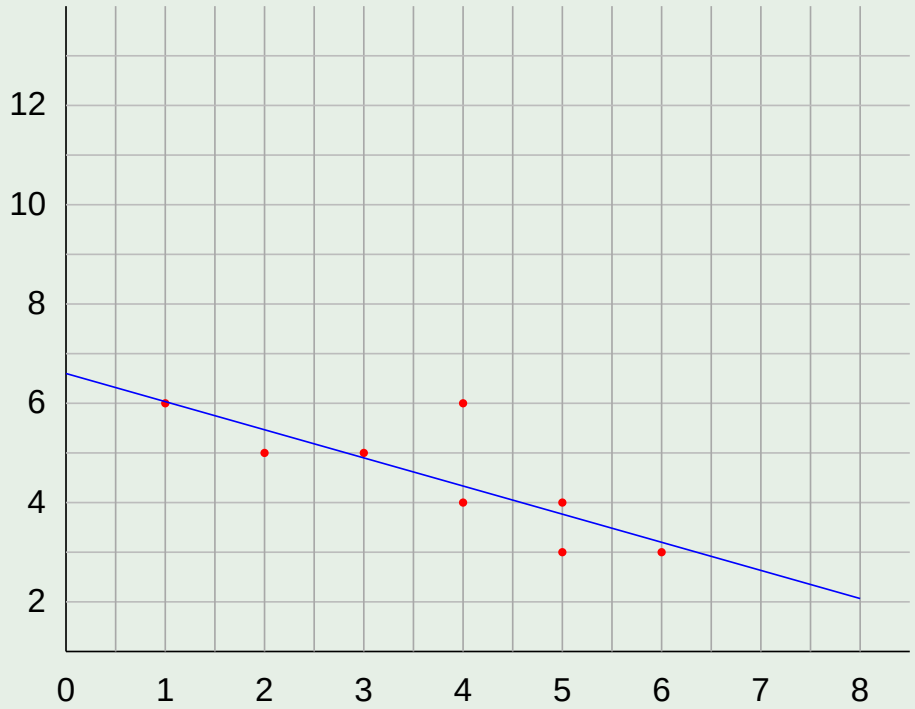
Including the point $(4, 10)$



Outliers and Influential Points

Example (Outliers and Influential Points)

Excluding the point $(4, 10)$



Outliers and Influential Points

- Now change the point $(4, 10)$ to the point $(12, 12)$.

x	y
1	6
2	5
3	5
4	6
4	4
5	3
5	4
6	3
12	12

Outliers and Influential Points

- The regression line is

$$\hat{y} = 6.615 - 0.564x$$

.

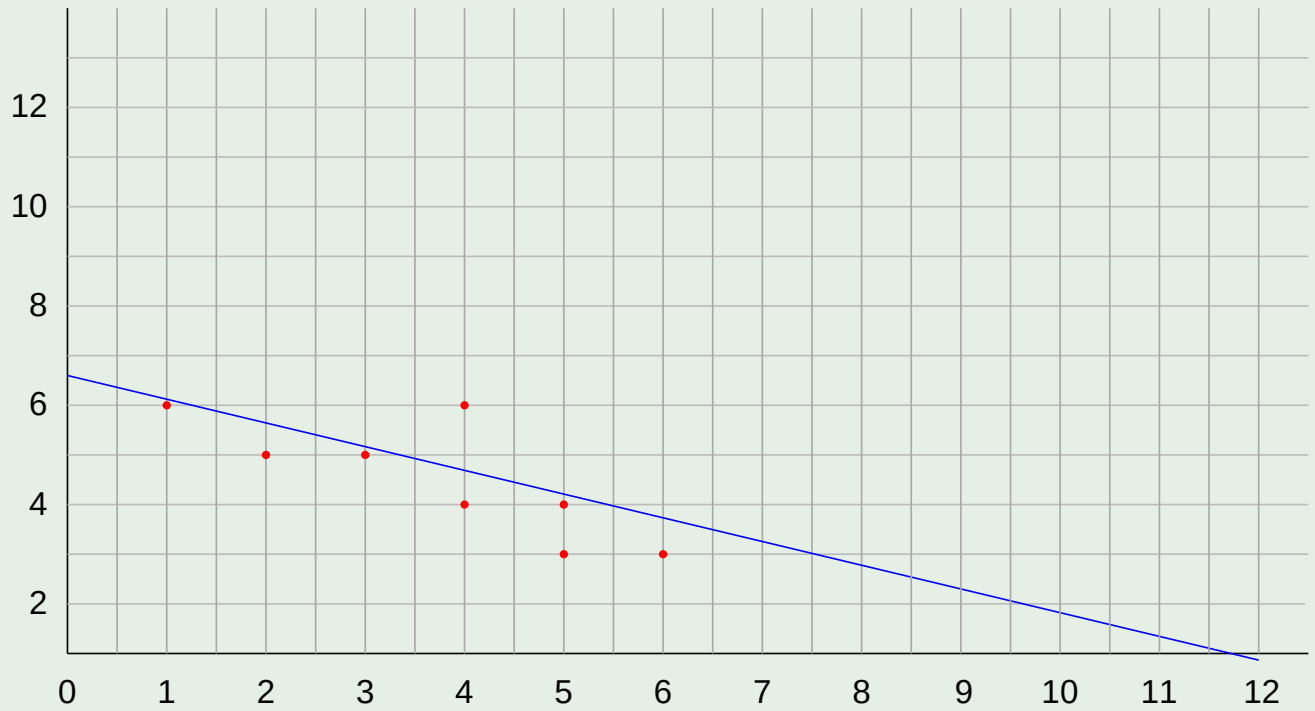
- Removing $(12, 12)$ changes it to

$$\hat{y} = 2.767 + 0.55x.$$

Outliers and Influential Points

Example (Outliers and Influential Points)

Excluding the point $(12, 12)$



Outliers and Influential Points

Residual Analysis and Outliers

Robb T. Koether

Homework Review

Introduction

Residual Analysis

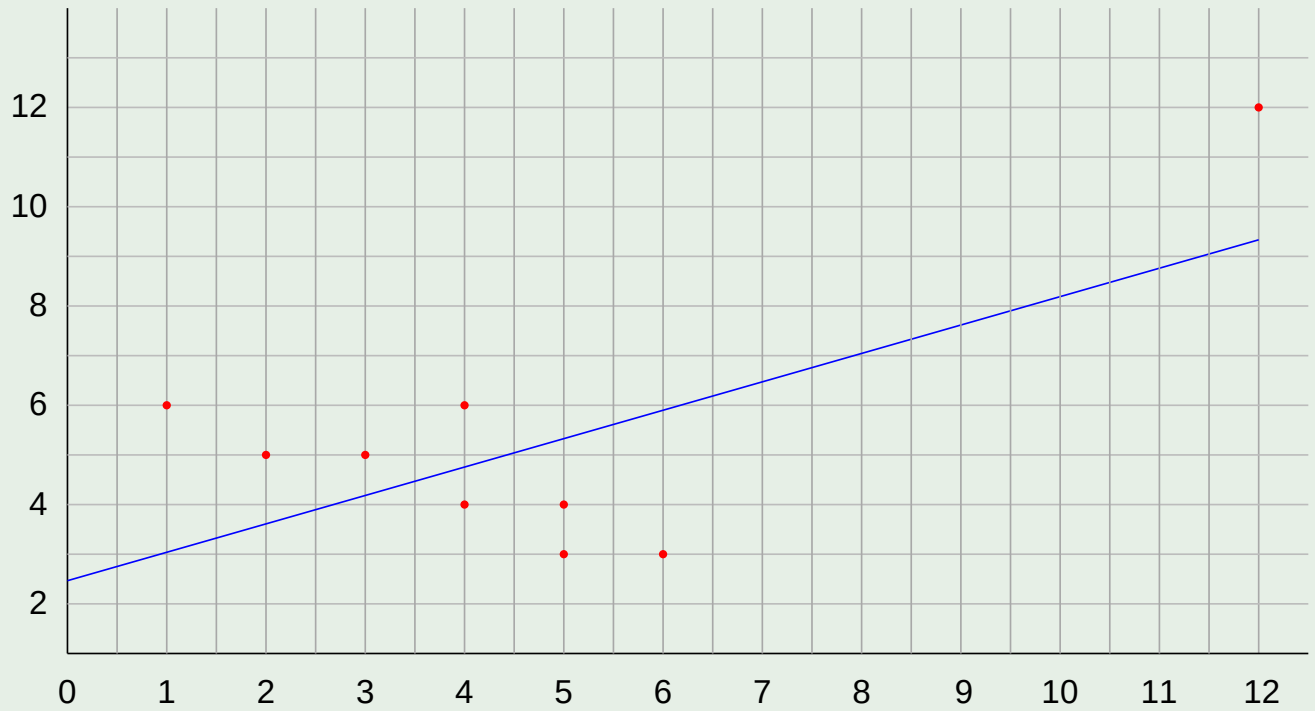
Nonlinear Regression

Outliers and Influential Points

Assignment

Example (Outliers and Influential Points)

Including the point $(12, 12)$



Outliers and Influential Points

- Yet the residual of $(12, 12)$ is only 2.63.
- The standard deviation of the set of residuals is 2.12.
- Therefore, $(12, 12)$ is not an outlier, but it is influential.

Assignment

Homework

- Read Sections 13.4, 13.5, pages 823 - 834.
- Let's Do It! 13.5, 13.6.
- Exercises 8, 9, 10, page 835.