Modeling a Linear Relationship

Lecture 41
Sections 13.1 - 13.3.1

Robb T. Koether

Hampden-Sydney College

Mon, Apr 9, 2012

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.
- In Chapter 14, we will investigate the relationship between two or more *qualitative* variables.

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.
- In Chapter 14, we will investigate the relationship between two or more *qualitative* variables.
- In Chapter 13, the basic problems will be

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.
- In Chapter 14, we will investigate the relationship between two or more *qualitative* variables.
- In Chapter 13, the basic problems will be
  - Determine whether there is a relationship.

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.
- In Chapter 14, we will investigate the relationship between two or more *qualitative* variables.
- In Chapter 13, the basic problems will be
  - Determine whether there is a relationship.
  - If there is one, then describe it quantitatively.

Introduction

- In Chapter 13, we will investigate the relationship between two *quantitative* variables.
- In Chapter 14, we will investigate the relationship between two or more *qualitative* variables.
- In Chapter 13, the basic problems will be
  - Determine whether there is a relationship.
  - If there is one, then describe it quantitatively.
- Through this quantitative description, we will to be able to predict the value of one variable when we know the value of the other.

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Bivariate Data

Definition (Bivariate)

Data are called bivariate if two observations, which we will call x and y, are made for each member of the sample.

- x is the explanatory variable.
- y is the response variable.
- x is also called the independent variable.
- y is also called the dependent variable.

Free Lunches vs. Graduation Rates

Example (Free-lunch Rate vs. Graduation Rate)

- Is the free-lunch rate in a school district correlated with the graduation rate in that district?
- Recently the Richmond Times-Dispatch published data for school districts in the Richmond area.
- We will draw a scatterplot of the data and see what it looks like.

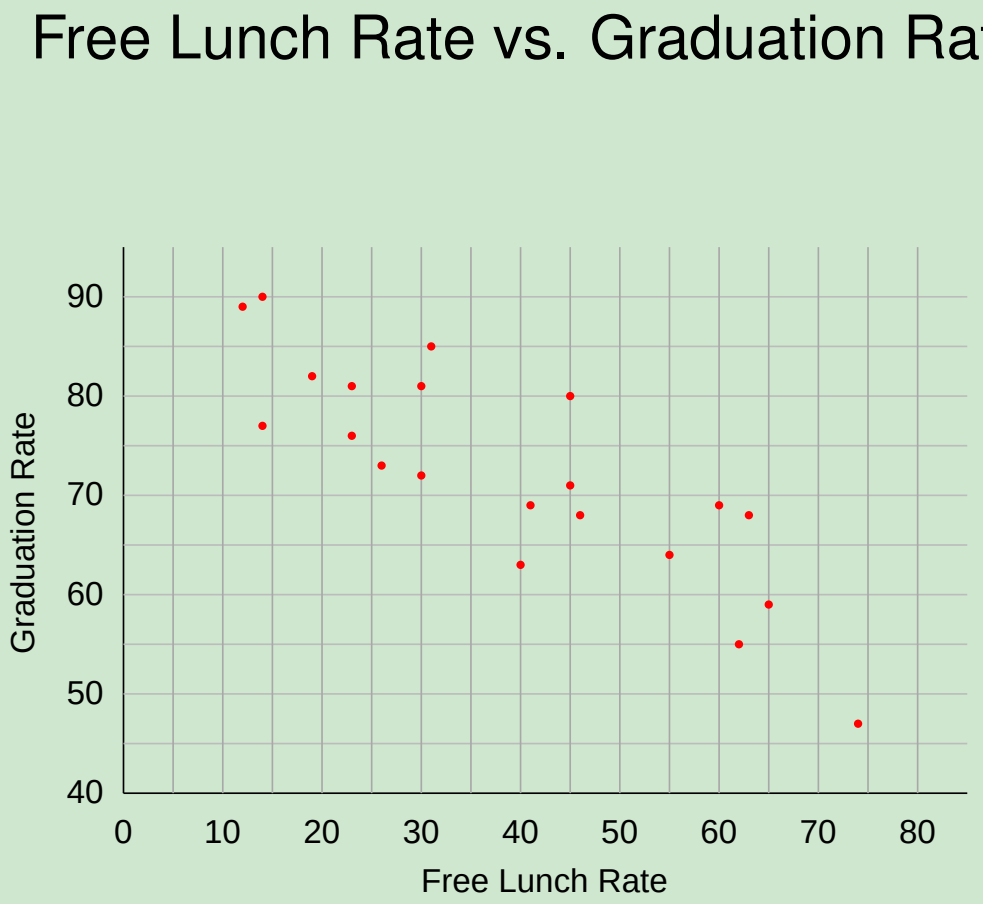
Free Lunches vs. Graduation Rates

Example (Free-lunch Rate vs. Graduation Rate)

District	Free Lunch	Grad. Rate	District	Free Lunch	Grad. Rate
Amelia	41.2	68.9	King and Queen	59.9	64.1
Caroline	40.2	62.9	King William	27.9	67.0
Charles City	45.8	67.7	Louisa	44.9	80.1
Chesterfield	22.5	80.5	New Kent	13.9	77.0
Colonial Hgts	25.7	73.0	Petersburg	61.6	54.6
Cumberland	55.3	63.9	Powhatan	12.2	89.3
Dinwiddie	45.2	71.4	Prince George	30.9	85.0
Goochland	23.3	76.3	Richmond	74.0	46.9
Hanover	13.7	90.1	Sussex	74.8	59.0
Henrico	30.2	81.1	West Point	19.1	82.0
Hopewell	63.1	63.4			

Scatter Plot

Example (Free-lunch Rate vs. Graduation Rate)



Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Describing a Relationship

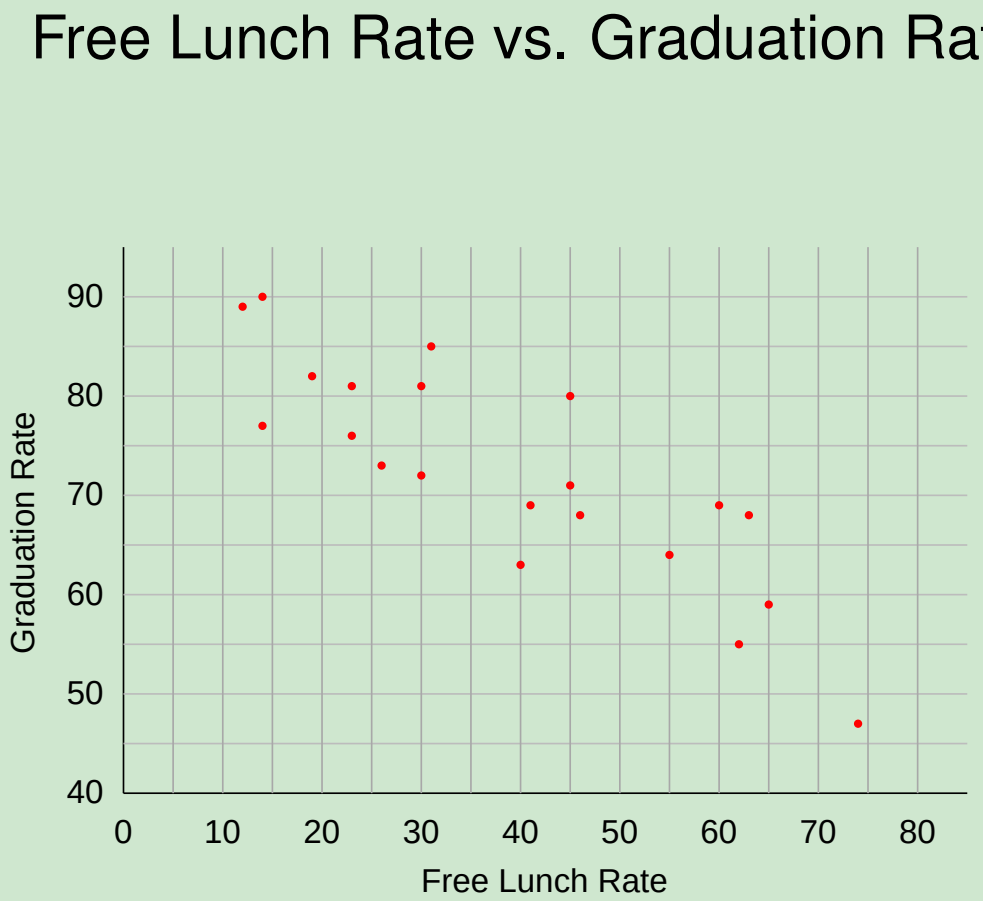
- Does there appear to be a relationship?
- How can we tell?
- How would we describe the relationship? (qualitatively and quantitatively)

Linear Association

- Draw (or imagine) an oval around the data set.
- If the oval is *tilted*, then there is some linear association.
- If the oval is tilted *upwards* from left to right, then there is positive association.
- If the oval is tilted *downwards* from left to right, then there is negative association.
- If the oval is not tilted at all, then there is no association.

Free-Lunch Participation vs. Graduation Rate

Example (Free-lunch Rate vs. Graduation Rate)



Free-Lunch Participation vs. Graduation Rate

Example (Free-lunch Rate vs. Graduation Rate)

Free Lunch Rate vs. Graduation Rate

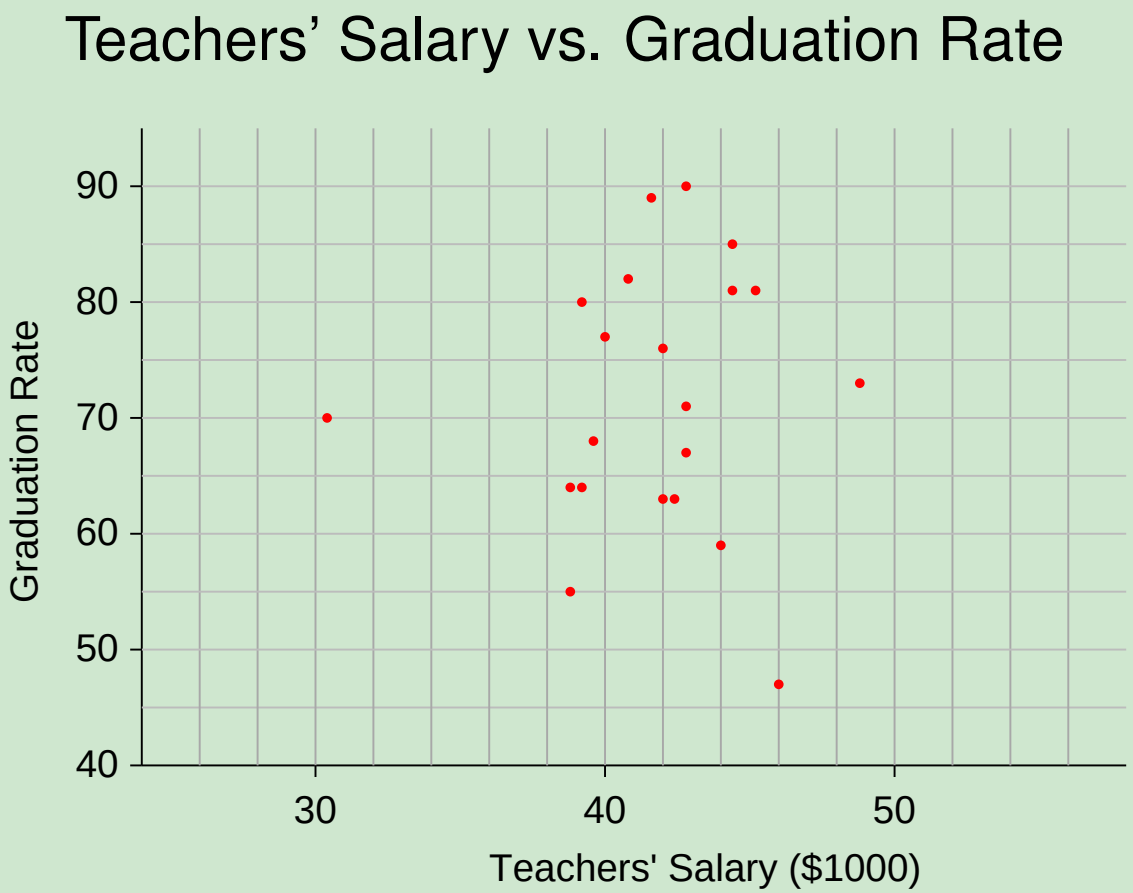
Teachers' Salary vs. Graduation Rate

Example (Teachers' Salary vs. Graduation Rate)

District	Avg. Salary	Grad. Rate	District	Avg. Salary	Grad. Rate
Amelia	30446	68.9	King and Queen	38803	64.1
Caroline	41935	62.9	King William	42750	67.0
Charles City	39530	67.7	Louisa	39010	80.1
Chesterfield	44417	80.5	New Kent	39891	77.0
Colonial Hgts	48999	73.0	Petersburg	38252	54.6
Cumberland	39380	63.9	Powhatan	41523	89.3
Dinwiddie	42866	71.4	Prince George	44529	85.0
Goochland	41893	76.3	Richmond	45875	46.9
Hanover	42715	90.1	Sussex	44142	59.0
Henrico	45021	81.1	West Point	40797	82.0
Hopewell	42351	63.4			

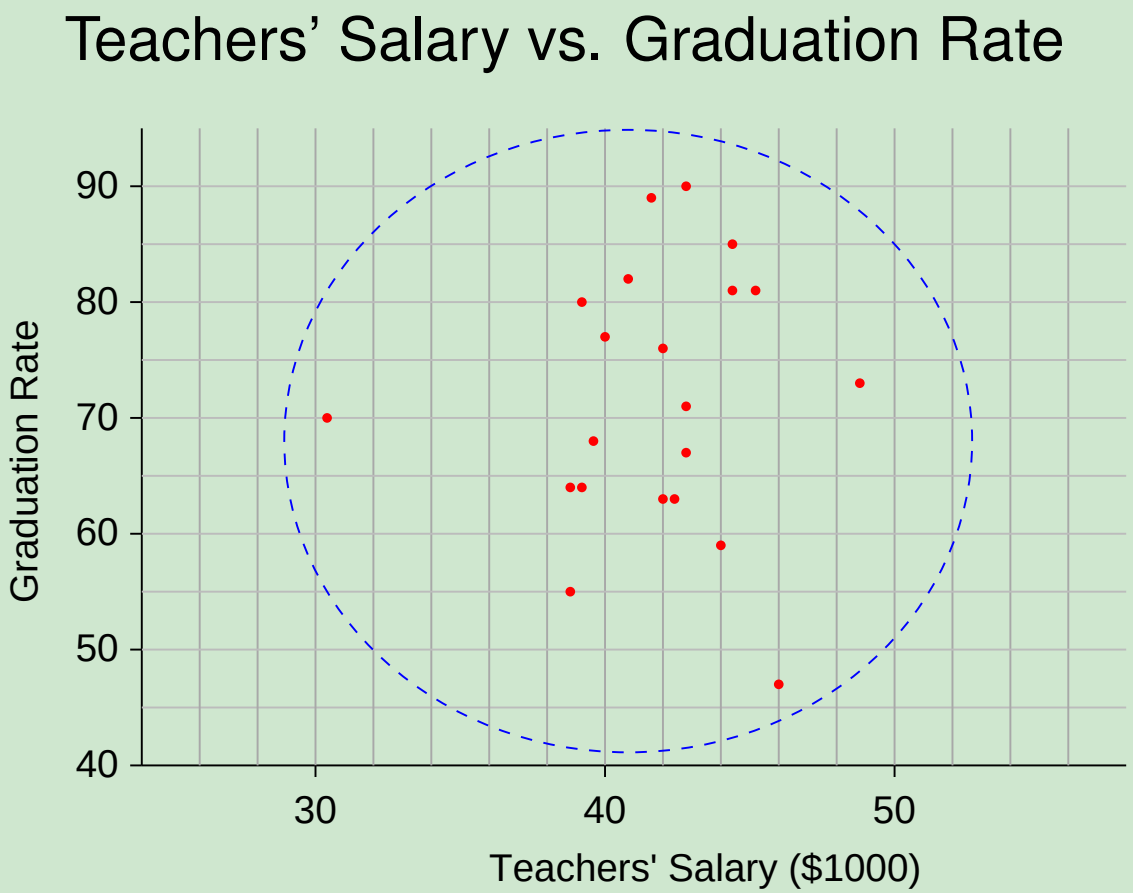
Teachers' Salary vs. Graduation Rate

Example (Teachers' Salary vs. Graduation Rate)



Teachers' Salary vs. Graduation Rate

Example (Teachers' Salary vs. Graduation Rate)



Passing Rate on English SOL vs. Graduation Rate

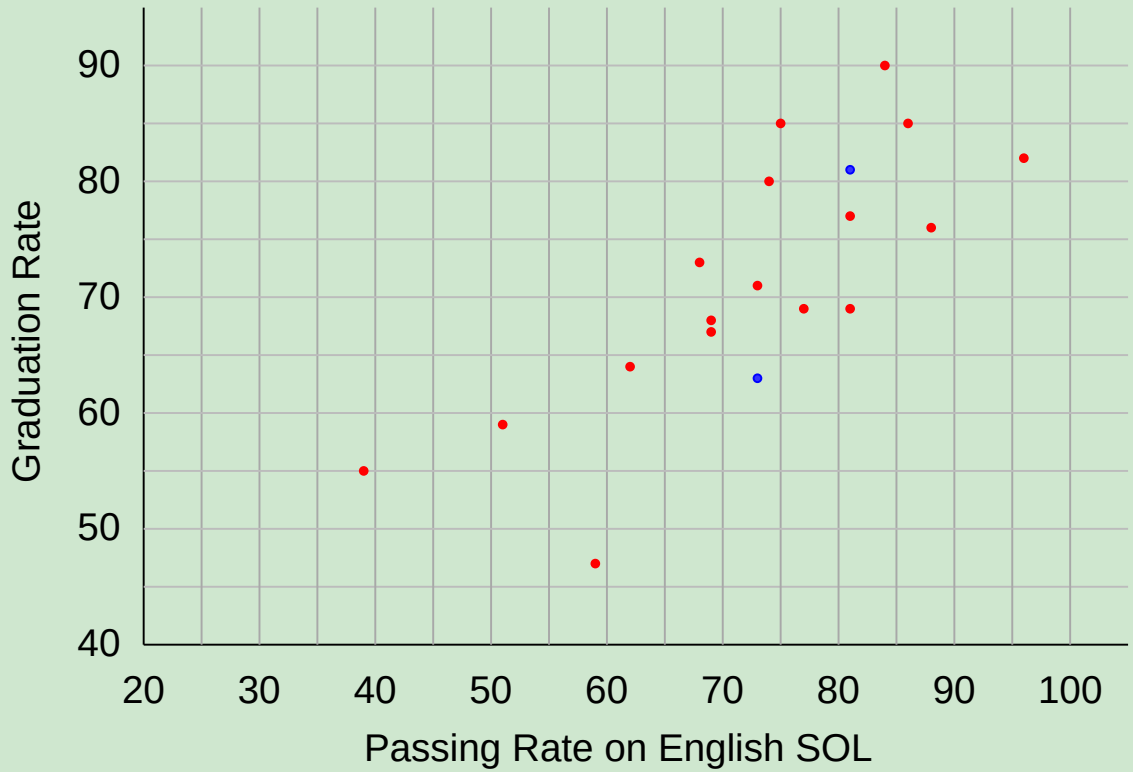
Example (Eng. SOL Passing Rate vs. Graduation Rate)

District	SOL Rate	Grad. Rate	District	SOL Rate	Grad. Rate
Amelia	77	68.9	King and Queen	62	64.1
Caroline	73	62.9	King William	69	67.0
Charles City	69	67.7	Louisa	74	80.1
Chesterfield	81	80.5	New Kent	81	77.0
Colonial Hgts	68	73.0	Petersburg	39	54.6
Cumberland	81	63.9	Powhatan	86	89.3
Dinwiddie	73	71.4	Prince George	75	85.0
Goochland	88	76.3	Richmond	59	46.9
Hanover	84	90.1	Sussex	51	59.0
Henrico	81	81.1	West Point	96	82.0
Hopewell	73	63.4			

Passing Rate on English SOL vs. Graduation Rate

Example (Eng. SOL Passing Rate vs. Graduation Rate)

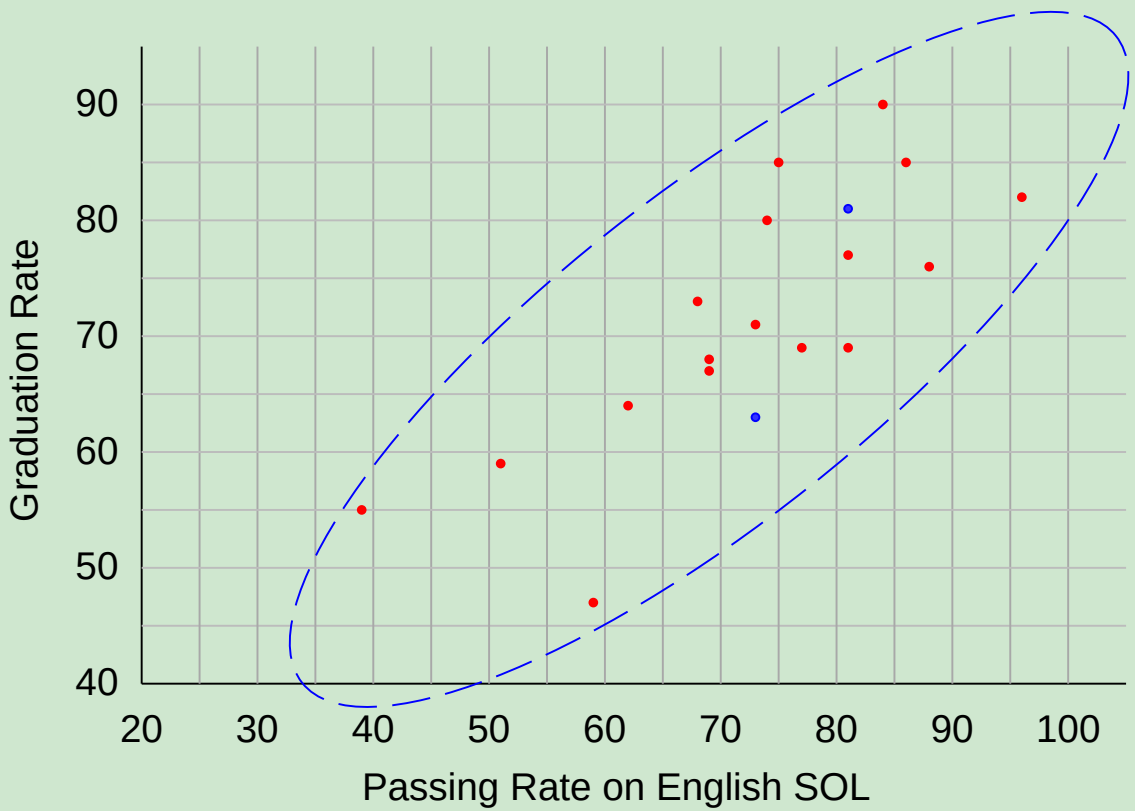
Passing Rate on English SOL vs. Graduation Rate



Passing Rate on English SOL vs. Graduation Rate

Example (Eng. SOL Passing Rate vs. Graduation Rate)

Passing Rate on English SOL vs. Graduation Rate

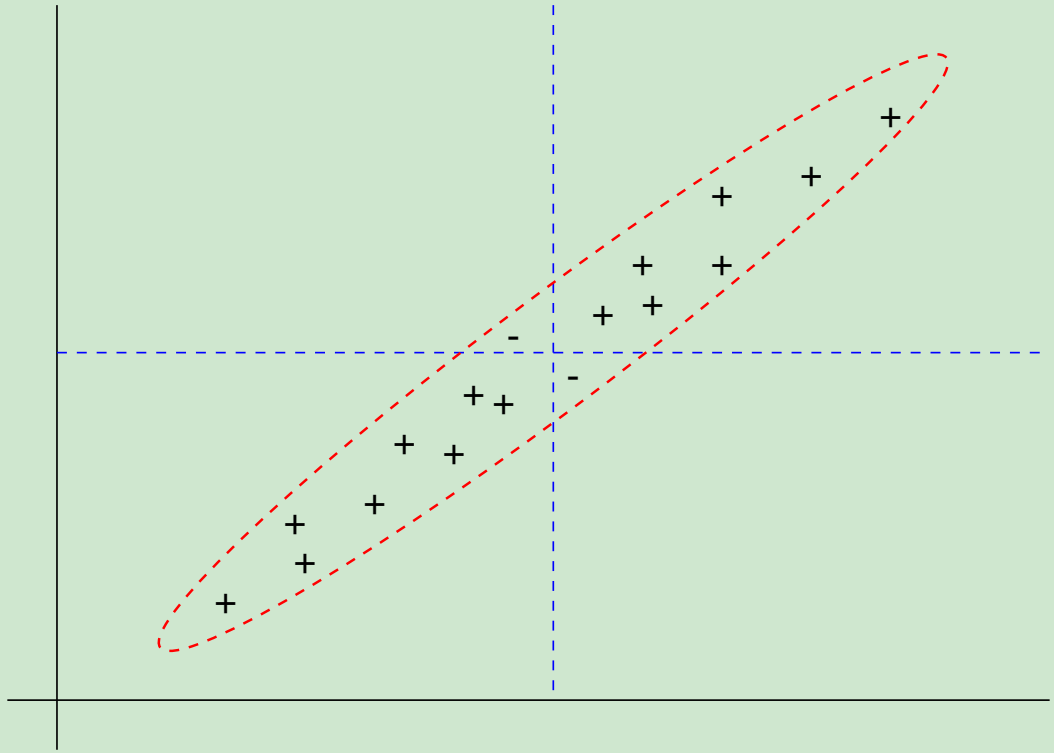


Strong vs. Weak Association

- The association is strong if the oval is narrow.
- The association is weak if the oval is wide.

The Least Squares Regression Line

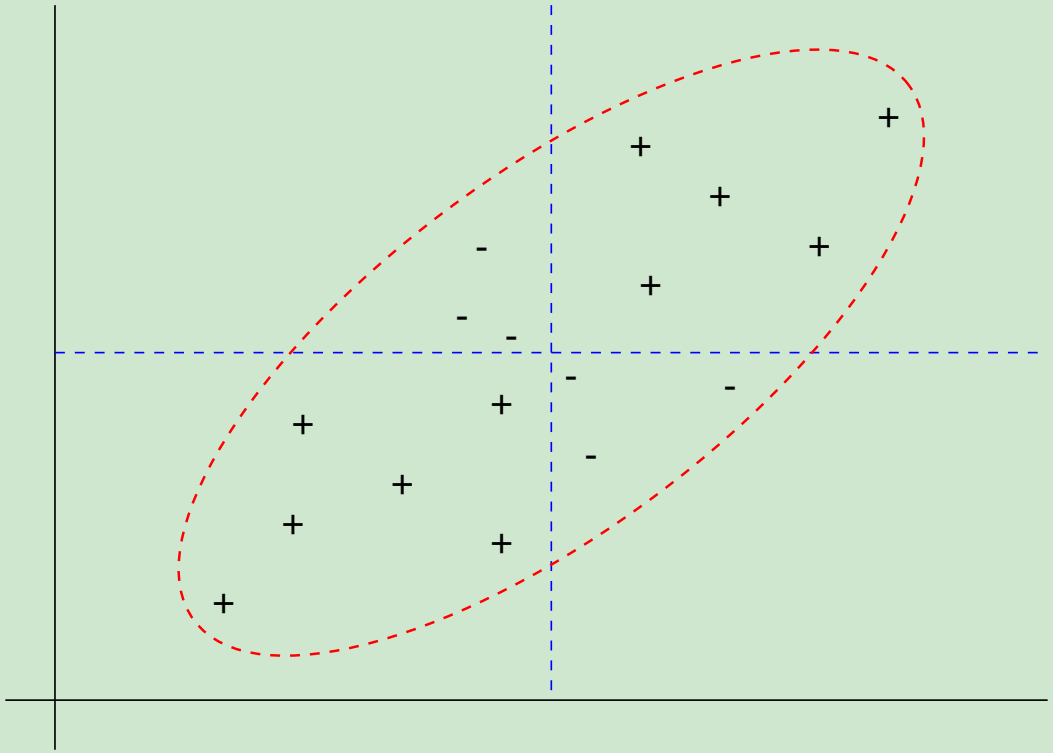
Example (The Least Squares Regression Line)



Strong positive association

The Least Squares Regression Line

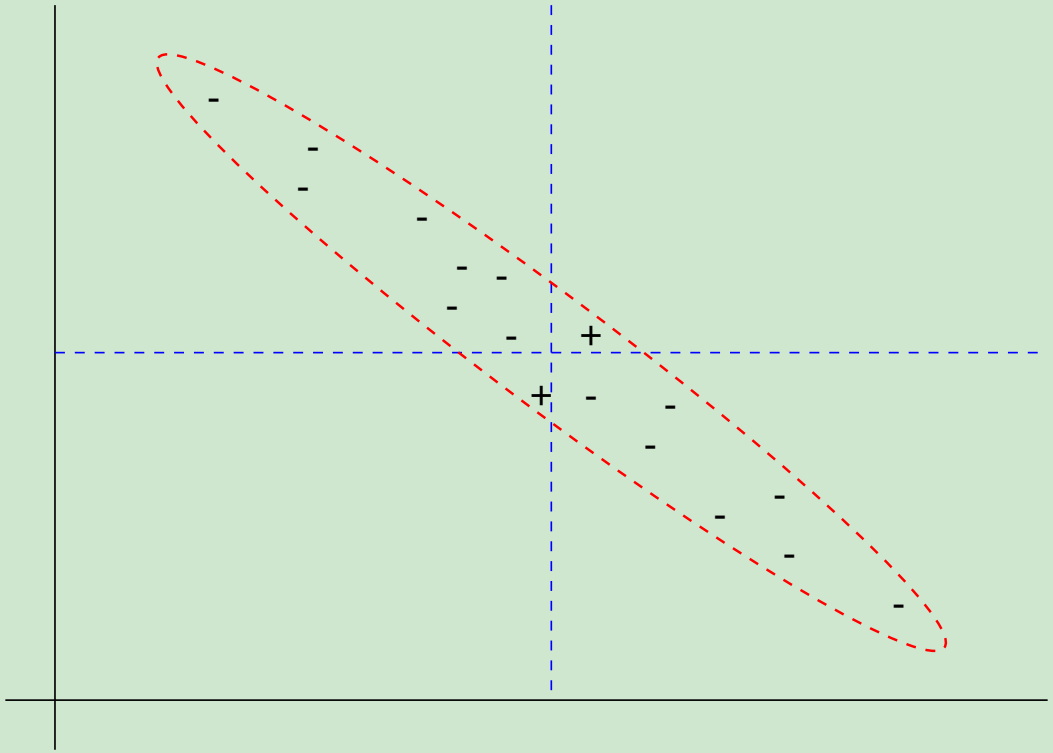
Example (The Least Squares Regression Line)



Weak positive association

The Least Squares Regression Line

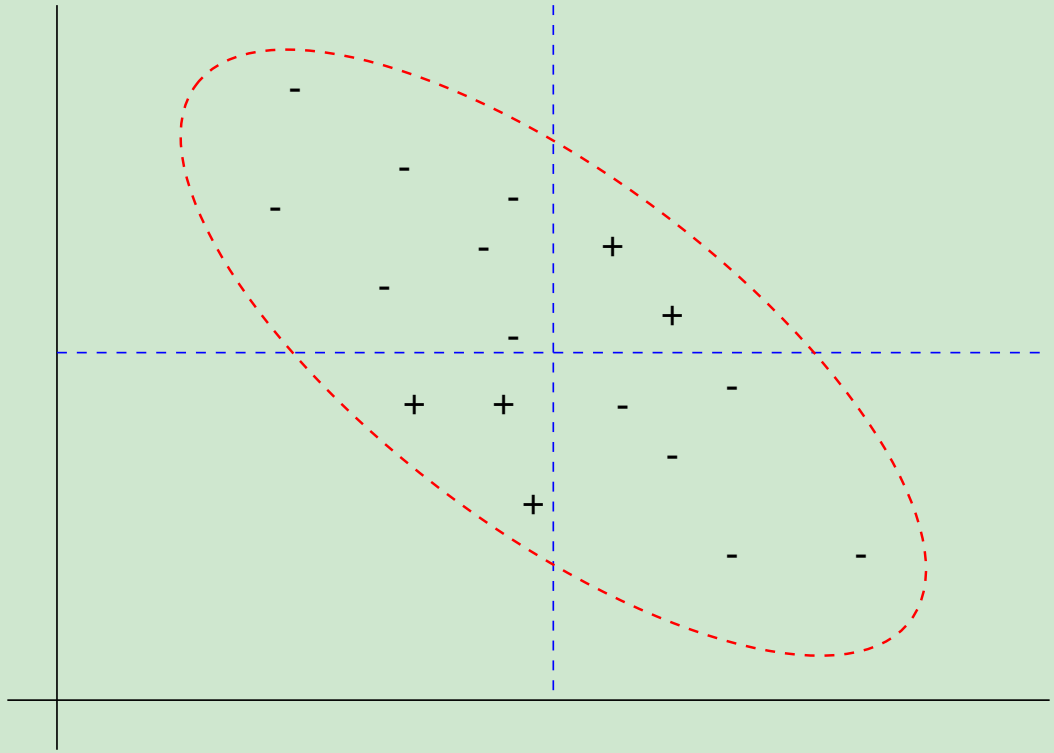
Example (The Least Squares Regression Line)



Strong negative association

The Least Squares Regression Line

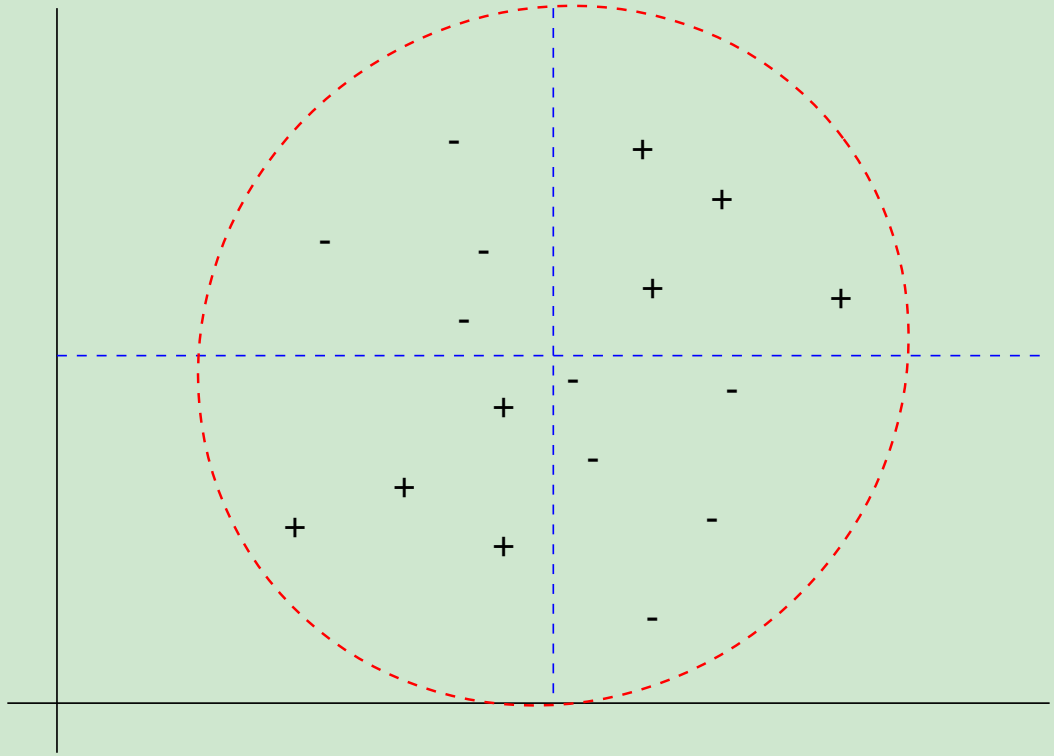
Example (The Least Squares Regression Line)



Weak negative association

The Least Squares Regression Line

Example (The Least Squares Regression Line)



No association

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Scatterplots on the TI-83

Example (Scatterplot)

- Draw a scatterplot of the following data.

x	y
1	8
3	12
4	9
5	14
8	16
9	20
11	17
15	24

TI-83 - Scatterplot

TI-83 Scatterplot

- Enter the x values in $\texttt{L}_1$.
- Enter the y values in $\texttt{L}_2$.
- Press `2nd STAT PLOT`.
- Select `Plot1` and press `ENTER`. The `Stat Plot` display appears.

TI-83 - Scatterplot

TI-83 Scatterplot

- Select `On` and press `ENTER`.
- Under `Type`, select the first icon (a small image of a scatterplot) and press `ENTER`.
- For `XList`, enter L_1.
- For `YList`, enter L_2.
- For `Mark`, select the one you want and press `ENTER`.

TI-83 - Scatterplots

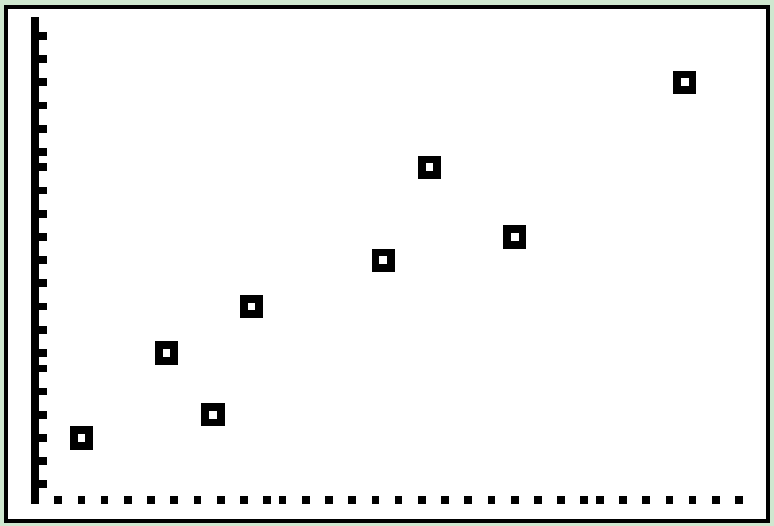
TI-83 Scatterplot

- Press `ZOOM`. The `Zoom` menu appears.
- Select `ZoomStat` (#9) and press `ENTER`. The scatterplot appears.
- Press `TRACE` and use the arrow keys to inspect the individual points.

TI-83 - Scatterplots

Example (TI-83 Scatterplot)

TI-83 Scatterplot



Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Simple Linear Regression

- We quantify the linear relationship between x and y by finding the equation of the line that "best" fits the data.
- That equation will be written in the form

$$\hat{y} = a + bx.$$

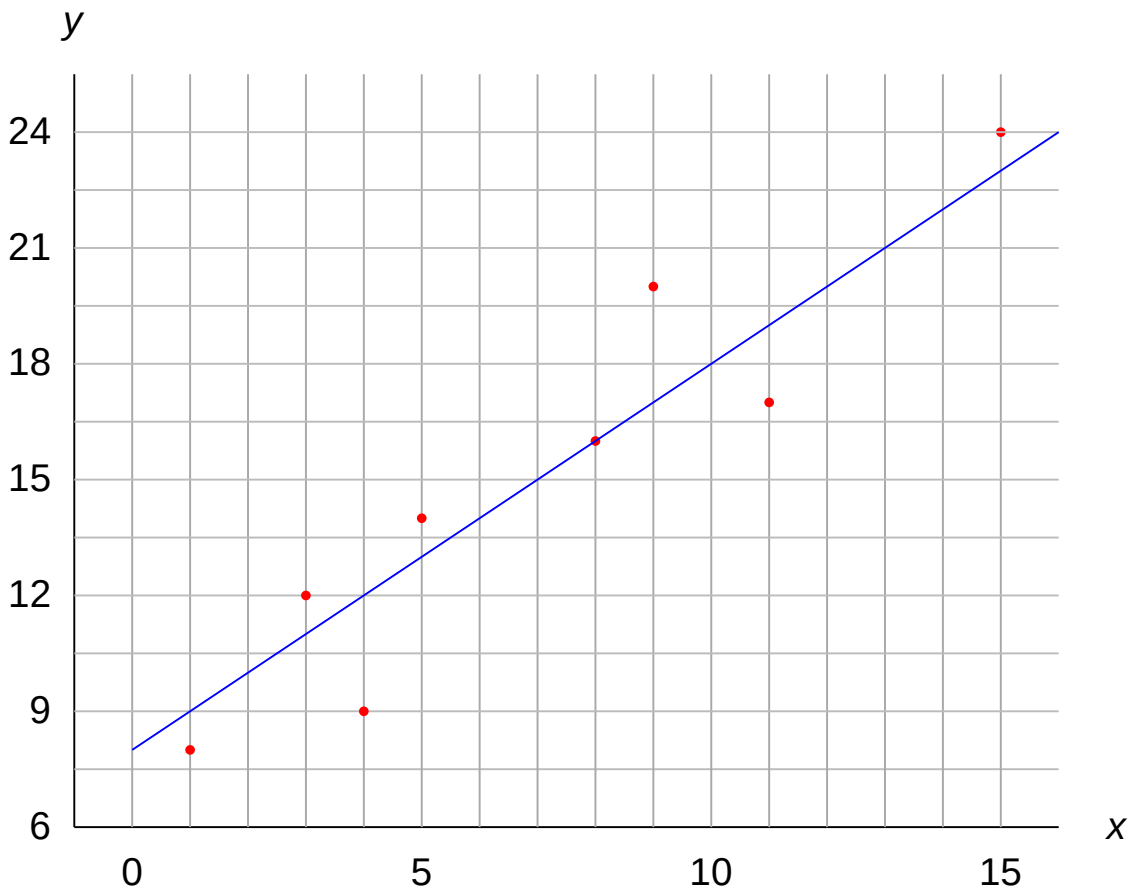
- The variable y represents the value that was observed.
- The variable $\hat{y}$ represents the value of y that is predicted by the model.

Simple Linear Regression

- Typically, there will be many lines that all look pretty good.
- To choose the best one, we need to measure how well a line fits the data.
- How do we measure how well a line fits the data?

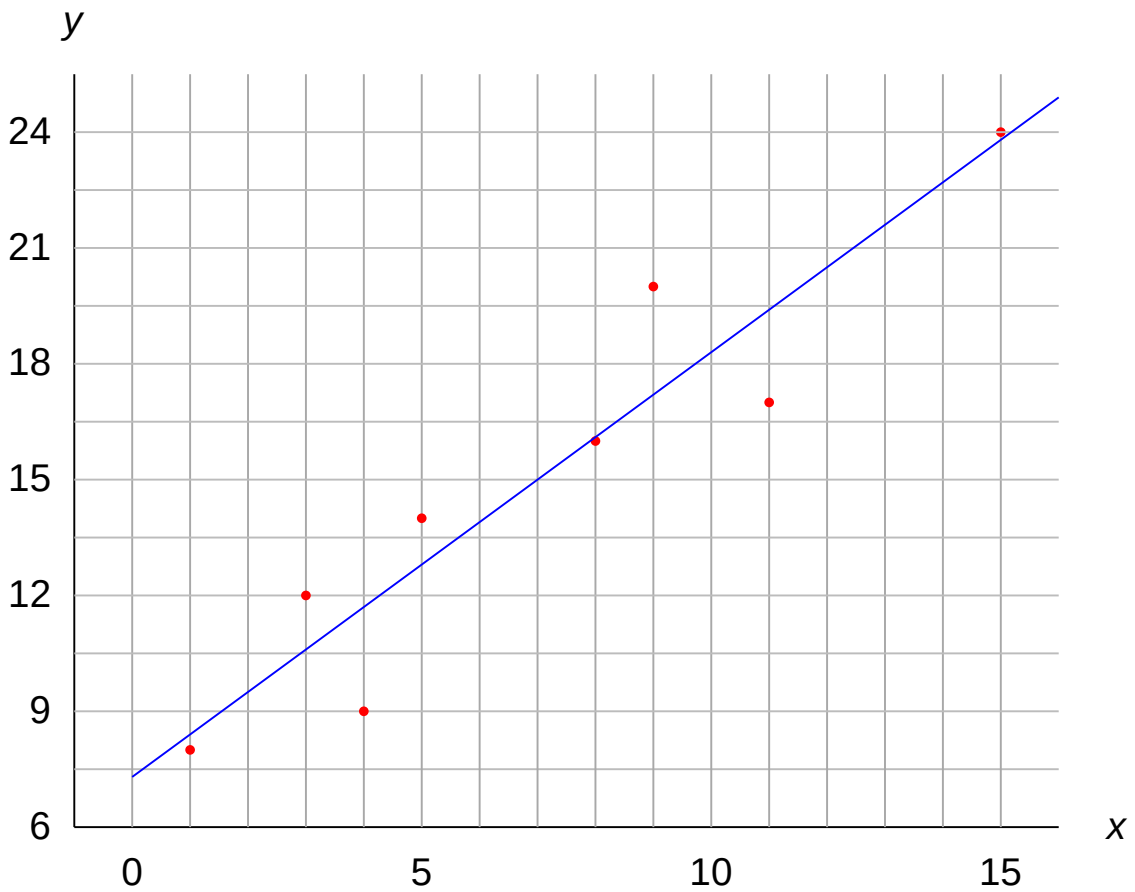
Measuring the Goodness of Fit

- Which line better fits the data?



Measuring the Goodness of Fit

- Which line better fits the data?

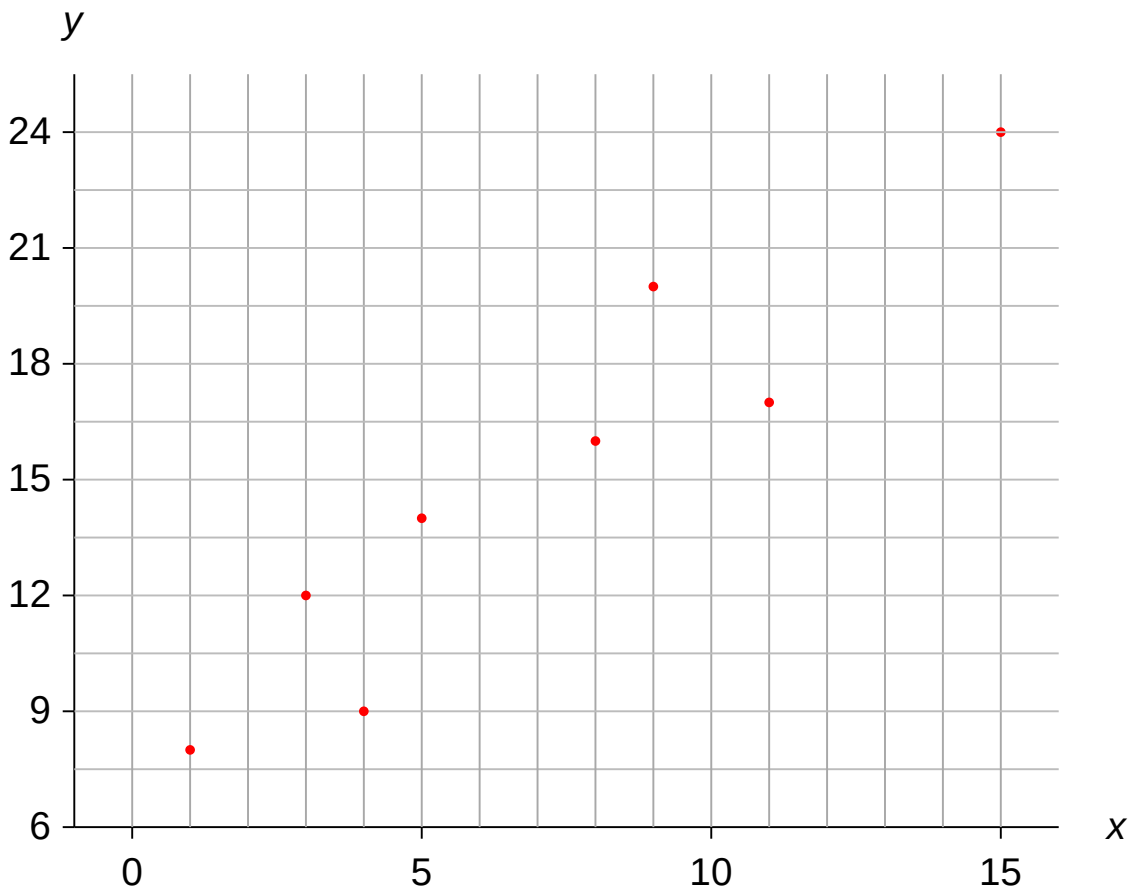


Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

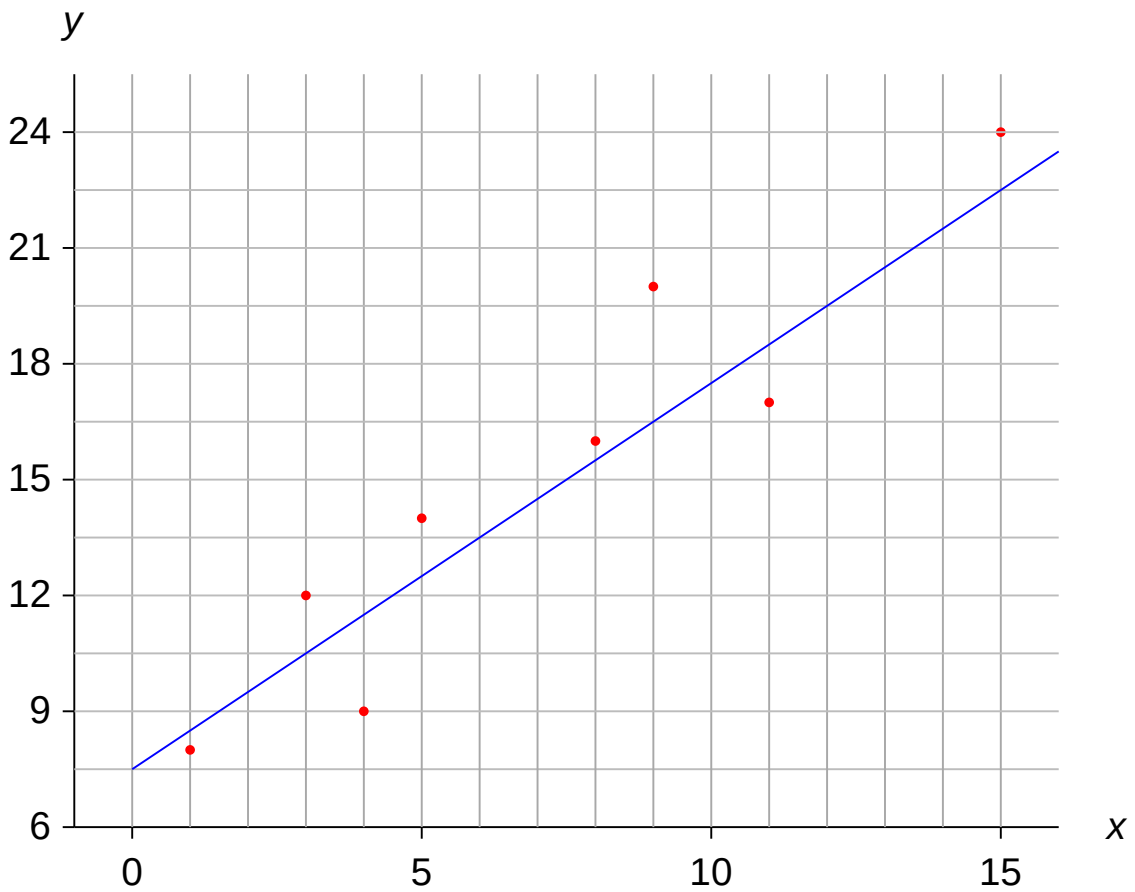
Measuring the Goodness of Fit

- Start with the scatterplot.



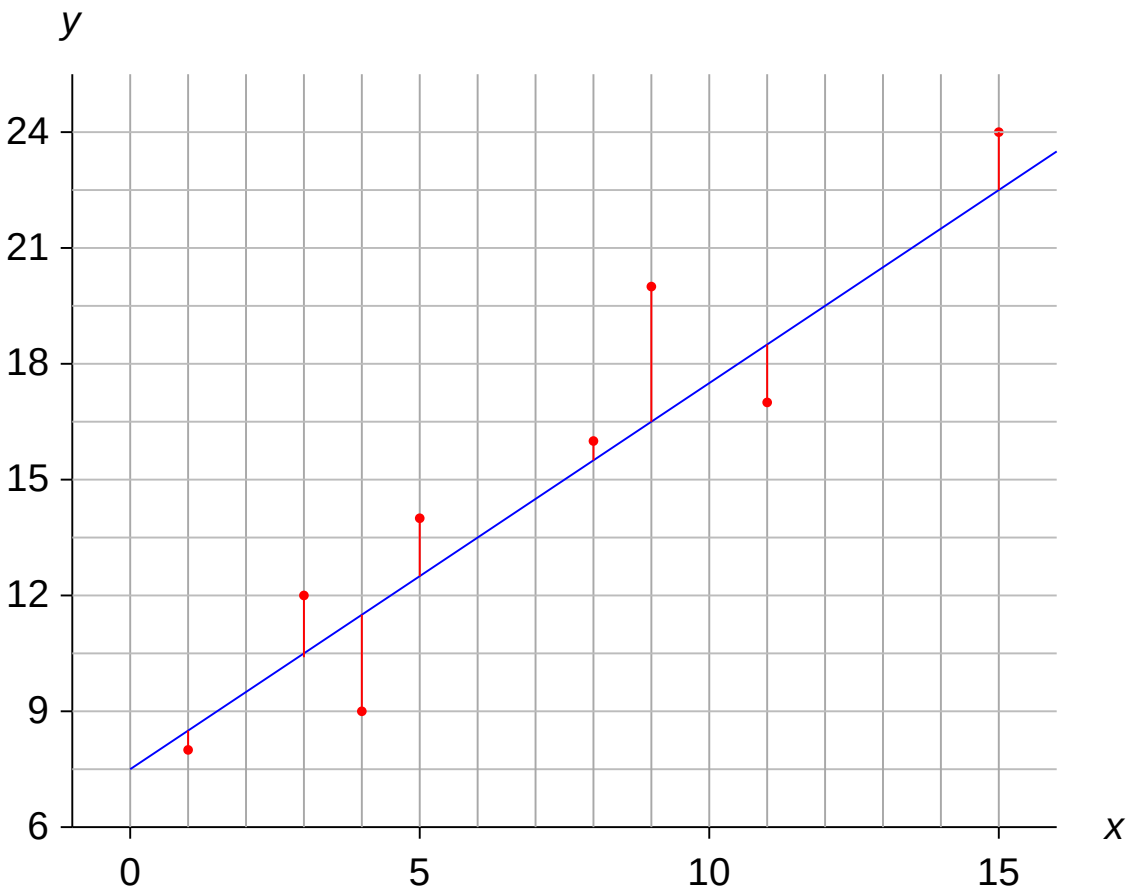
Measuring the Goodness of Fit

- Draw the line through the scatterplot.



Measuring the Goodness of Fit

- Measure the *vertical* distances to the line.



Residuals

Definition (Residual)

The i^{th} residual is the difference between y_i and $\hat{y}_i$.

- The vertical distances are called the residuals.
- The formula for the i^{th} residual is

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

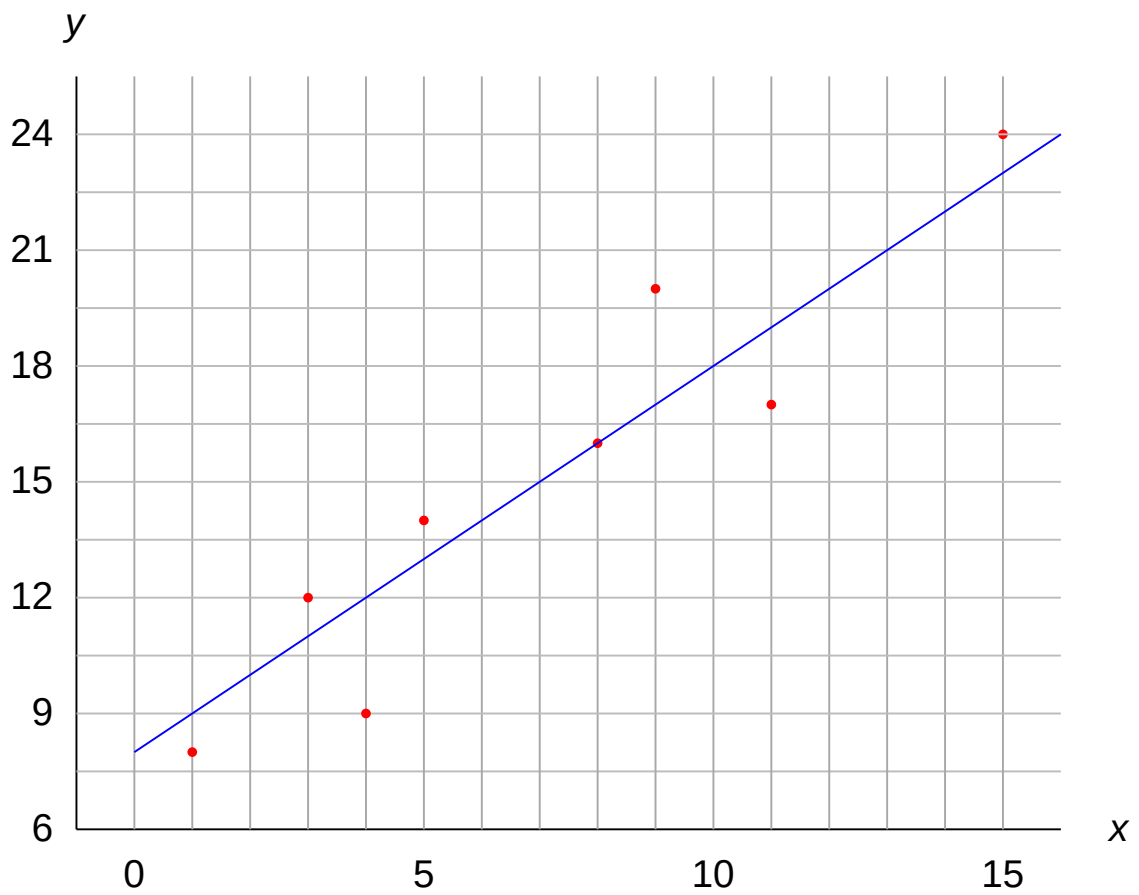
Measuring the Goodness of Fit

Definition (Line of Best Fit)

The line of best fit is the line with the smallest sum of squared residuals. This line of best fit is also called the least squares line and the regression line.

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 8 + x$.



Example

- Start with the data points

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8			
3	12			
4	9			
5	14			
8	16			
9	20			
11	17			
15	24			

Example

- Compute the predicted y, using $\hat{y} = 8 + x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	9		
3	12	11		
4	9	12		
5	14	13		
8	16	16		
9	20	17		
11	17	19		
15	24	23		

Example

- Find the residues.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	9	-1	
3	12	11	1	
4	9	12	-3	
5	14	13	1	
8	16	16	0	
9	20	17	3	
11	17	19	-2	
15	24	23	1	

Example

- Square the residues.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	9	-1	1
3	12	11	1	1
4	9	12	-3	9
5	14	13	1	1
8	16	16	0	0
9	20	17	3	9
11	17	19	-2	4
15	24	23	1	1

Example

- Add up the residues.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	9	-1	1
3	12	11	1	1
4	9	12	-3	9
5	14	13	1	1
8	16	16	0	0
9	20	17	3	9
11	17	19	-2	4
15	24	23	1	1
				26

Example

- The sum of the squared residues is called the sum of squared errors (SSE).

$$\begin{aligned} \text{SSE} &= \sum (y - \hat{y})^2 \\ &= 1 + 1 + 9 + 1 + 0 + 9 + 4 + 1 \\ &= 26. \end{aligned}$$

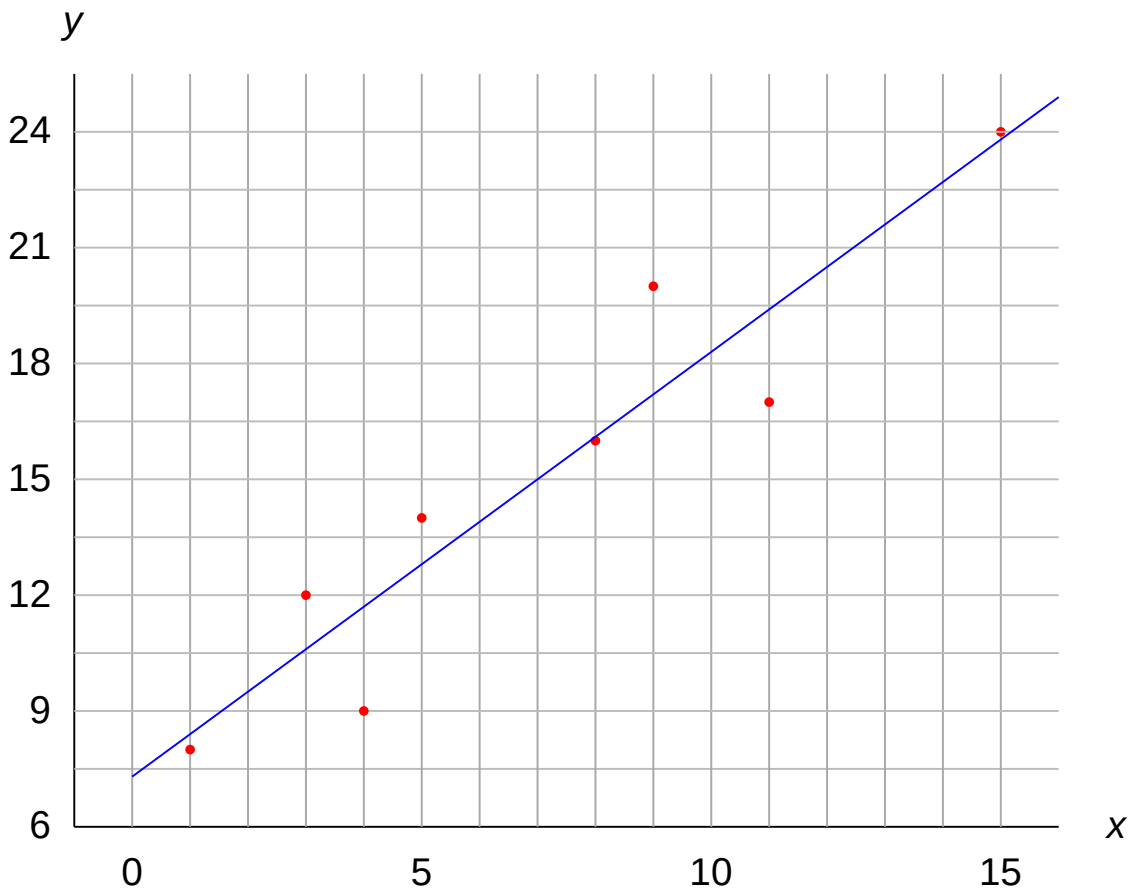
TI-83 - Computing Residuals

TI-83 Computing Residuals

- Enter the x-values in list $\texttt{L}_1$ and the y-values in list $\texttt{L}_2$.
- Compute $a + b*\texttt{L}_1$ and store in list $\texttt{L}_3$ ($\hat{y}$ values).
- Compute $(\texttt{L}_2-\texttt{L}_3)^2$. This is a list of the squared residuals.
- Compute $\texttt{sum(Ans)}$. This is the sum of the squared residuals.

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.



Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8			
3	12			
4	9			
5	14			
8	16			
9	20			
11	17			
15	24			

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	8.4		
3	12	10.6		
4	9	11.7		
5	14	12.8		
8	16	16.1		
9	20	17.2		
11	17	19.4		
15	24	23.8		

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	8.4	-0.4	
3	12	10.6	1.4	
4	9	11.7	-2.7	
5	14	12.8	1.2	
8	16	16.1	-0.1	
9	20	17.2	2.8	
11	17	19.4	-2.4	
15	24	23.8	0.2	

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	8.4	-0.4	0.16
3	12	10.6	1.4	1.96
4	9	11.7	-2.7	7.29
5	14	12.8	1.2	1.44
8	16	16.1	-0.1	0.01
9	20	17.2	2.8	7.84
11	17	19.4	-2.4	5.76
15	24	23.8	0.2	0.04

Least Squares Line

- Let's see how good the fit is for the line $\hat{y} = 7.3 + 1.1x$.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	8	8.4	-0.4	0.16
3	12	10.6	1.4	1.96
4	9	11.7	-2.7	7.29
5	14	12.8	1.2	1.44
8	16	16.1	-0.1	0.01
9	20	17.2	2.8	7.84
11	17	19.4	-2.4	5.76
15	24	23.8	0.2	0.04
				24.50

Sum of Squared Residuals

- We conclude that $\hat{y} = 7.3 + 1.1x$ is a better fit than $\hat{y} = 8 + x$.
- Is it the *best* fit?
- It turns out that it is the best possible fit.
- Therefore,

$$\hat{y} = 7.3 + 1.1x$$

is the regression line for this data set.

Prediction

Definition (Interpolation)

To interpolate is to use an x value that is within the observed extremes of x values to predict y.

Definition (Extrapolation)

To extrapolate is to use an x value that is beyond the observed extremes of x values to predict y.

- Interpolated values are more reliable then extrapolated values.
- The farther out the values are extrapolated, the less reliable they are.

Interpolation vs. Extrapolation

- Use the regression line to predict y when
- $x = 6$
- $x = 12$
- $x = 30$

Interpolation vs. Extrapolation

- Use the regression line to predict y when
- $x = 6$: $\hat{y}(6) = 7.3 + 1.1(6) = 13.9$.
- $x = 12$: $\hat{y}(12) = 7.3 + 1.1(12) = 20.5$.
- $x = 30$: $\hat{y}(30) = 7.3 + 1.1(30) = 40.3$.

Outline

1. Introduction
2. Scatterplots
3. Describing Relationships
4. Scatterplots on the TI-83
5. Linear Regression
   - Which Line is Better?
   - Measuring the Goodness of Fit
6. Assignment

Assignment

Homework

- Read Sections 13.1 - 13.3.1, pages 808 - 820.
- Let's Do It! 13.1, 13.2.
- Exercises 1, 2, 3(a), 4(a), 5(a), page 821.