# Calculating the Regression Line

Lecture 42
Section 13.3.2

Robb T. Koether

Hampden-Sydney College

Tue, Apr 10, 2012

# Outline


1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# Outline

1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

- 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

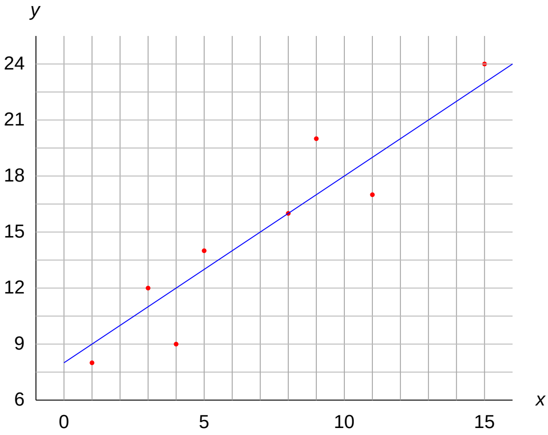
$$\hat{y} = a + bx.$$

- 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

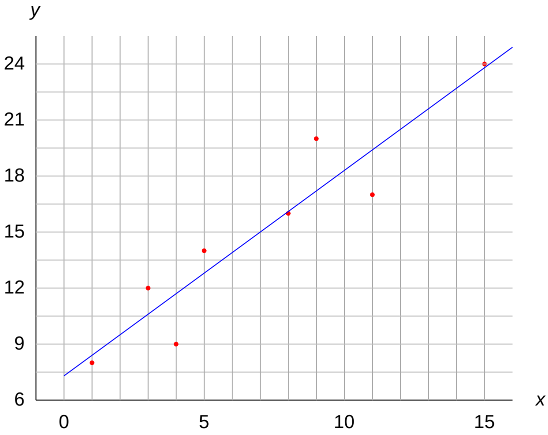
- Typically, there will be many lines that all look pretty good.
- To choose the best one, we need to measure how well a line fits the data.
- How do we measure how well a line fits the data?

# Measuring the Goodness of Fit




Which line better fits the data?

# Measuring the Goodness of Fit

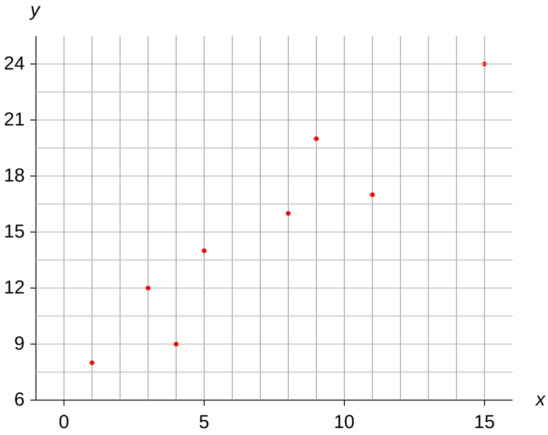



Which line better fits the data?

# Outline

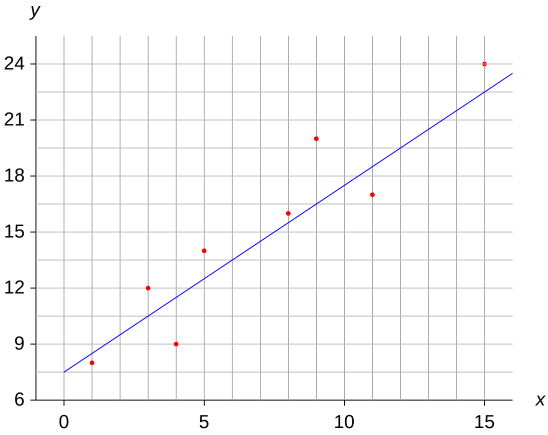

1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# Measuring the Goodness of Fit



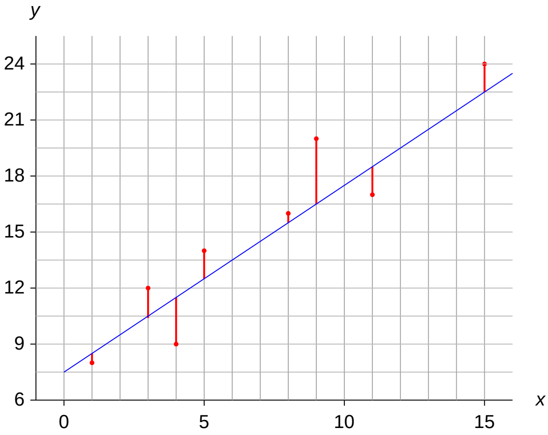

Start with the scatterplot.

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

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

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

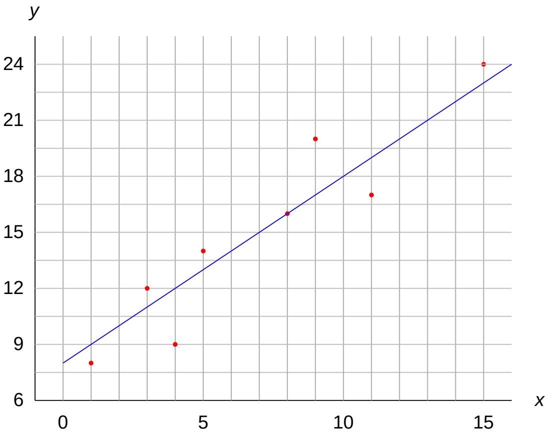
# Measuring the Goodness of Fit

## 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

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

Start with the data points.

# Example

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

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

# Example

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 | |
| | | | | |

Find the residues.

# Example

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 |
| | | | | |

Square the residues.

# Example

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 |

Add up the residues.

# Example

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

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

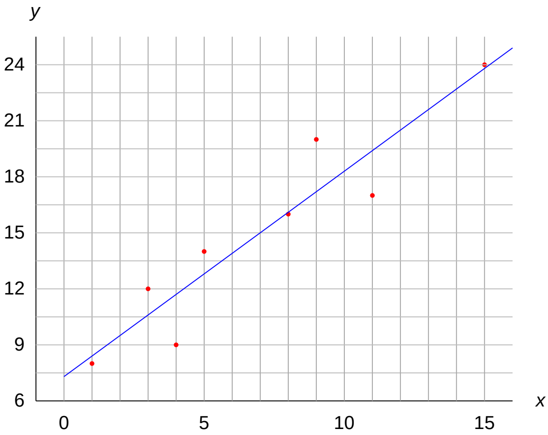
# TI-83 - Computing Residuals

## TI-83 Computing Residuals

- Enter the *x*-values in list $\mathtt{L}_1$ and the *y*-values in list $\mathtt{L}_2$.
- Compute $a + b{*}\mathtt{L}_1$ and store in list $\mathtt{L}_3$ ($\hat{y}$ values).
- Compute $\mathtt{(L_2-L_3)^2}$. This is a list of the squared residuals.
- Compute `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

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

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

# Least Squares Line

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 | | |
| | | | | |

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

# Least Squares Line

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 | |
| | | | | |

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

# Least Squares Line

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 |
| | | | | |

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

# Least Squares Line

| $x$ | $y$ | $\hat{y}$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
|---|---|---|---|---|
| 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 |

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

# 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.

# Outline


1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# The Least Squares Regression Line

- The equation of the regression line is of the form

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

- $b$ is the slope of the regression line.
- $a$ is the $y$-intercept.
- We need to find the coefficients $a$ and $b$ from the data.

# The Least Squares Regression Line

- The formula for $b$ is

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2}$$

- The formula for $a$ is

$$a = \overline{y} - b\overline{x}.$$

# The Least Squares Regression Line

- An alternate formula for $b$ is

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}.$$

- This is the one that the TI-83 uses.

## Example (Calculating the Regression Line)

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

Consider again the data set.

# Example

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ | $(x - \overline{x})(y - \overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | $-6$ | | | | |
| 3 | 12 | $-4$ | | | | |
| 4 | 9 | $-3$ | | | | |
| 5 | 14 | $-2$ | | | | |
| 8 | 16 | 1 | | | | |
| 9 | 20 | 2 | | | | |
| 11 | 17 | 4 | | | | |
| 15 | 24 | 8 | | | | |
| | | | | | | |

Compute the $x$ deviations.

# Example

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x-\overline{x}$ | $y-\overline{y}$ | $(x-\overline{x})^2$ | $(y-\overline{y})^2$ | $(x-\overline{x})(y-\overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | $-6$ | $-7$ | | | |
| 3 | 12 | $-4$ | $-3$ | | | |
| 4 | 9 | $-3$ | $-6$ | | | |
| 5 | 14 | $-2$ | $-1$ | | | |
| 8 | 16 | 1 | 1 | | | |
| 9 | 20 | 2 | 5 | | | |
| 11 | 17 | 4 | 2 | | | |
| 15 | 24 | 8 | 9 | | | |
| | | | | | | |

Compute the $y$ deviations.

# Example

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ | $(x - \overline{x})(y - \overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | $-6$ | $-7$ | 36 | | |
| 3 | 12 | $-4$ | $-3$ | 16 | | |
| 4 | 9 | $-3$ | $-6$ | 9 | | |
| 5 | 14 | $-2$ | $-1$ | 4 | | |
| 8 | 16 | 1 | 1 | 1 | | |
| 9 | 20 | 2 | 5 | 4 | | |
| 11 | 17 | 4 | 2 | 16 | | |
| 15 | 24 | 8 | 9 | 64 | | |
| | | | | | | |

Compute the squared $x$ deviations.

# Example

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ | $(x - \overline{x})(y - \overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | $-6$ | $-7$ | 36 | 49 | |
| 3 | 12 | $-4$ | $-3$ | 16 | 9 | |
| 4 | 9 | $-3$ | $-6$ | 9 | 36 | |
| 5 | 14 | $-2$ | $-1$ | 4 | 1 | |
| 8 | 16 | 1 | 1 | 1 | 1 | |
| 9 | 20 | 2 | 5 | 4 | 25 | |
| 11 | 17 | 4 | 2 | 16 | 4 | |
| 15 | 24 | 8 | 9 | 64 | 81 | |
| | | | | | | |

Compute the squared $y$ deviations.

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ | $(x - \overline{x})(y - \overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | −6 | −7 | 36 | 49 | 42 |
| 3 | 12 | −4 | −3 | 16 | 9 | 12 |
| 4 | 9 | −3 | −6 | 9 | 36 | 18 |
| 5 | 14 | −2 | −1 | 4 | 1 | 2 |
| 8 | 16 | 1 | 1 | 1 | 1 | 1 |
| 9 | 20 | 2 | 5 | 4 | 25 | 10 |
| 11 | 17 | 4 | 2 | 16 | 4 | 8 |
| 15 | 24 | 8 | 9 | 64 | 81 | 72 |
| | | | | | | |

Compute the product of $x$ and $y$ deviations.

# Example

## Example (Calculating the Regression Line)

| $x$ | $y$ | $x - \overline{x}$ | $y - \overline{y}$ | $(x - \overline{x})^2$ | $(y - \overline{y})^2$ | $(x - \overline{x})(y - \overline{y})$ |
|---|---|---|---|---|---|---|
| 1 | 8 | −6 | −7 | 36 | 49 | 42 |
| 3 | 12 | −4 | −3 | 16 | 9 | 12 |
| 4 | 9 | −3 | −6 | 9 | 36 | 18 |
| 5 | 14 | −2 | −1 | 4 | 1 | 2 |
| 8 | 16 | 1 | 1 | 1 | 1 | 1 |
| 9 | 20 | 2 | 5 | 4 | 25 | 10 |
| 11 | 17 | 4 | 2 | 16 | 4 | 8 |
| 15 | 24 | 8 | 9 | 64 | 81 | 72 |
| | | 0 | 0 | 150 | 206 | 165 |

Find the sums.

# Example

## Example (Calculating the Regression Line)

- Compute the coefficients from the formula.

$$
\begin{aligned}
b &= \frac{165}{150} = 1.1. \\
a &= 15 - (1.1)(7) = 7.3.
\end{aligned}
$$

- The equation is

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

# Outline

1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# The Regression Line on the TI-83

## TI-83 The Regression Line (`2-Var Stats`)

- Enter `2-Var Stats` $L_1, L_2$.
- Press `ENTER`.
- The calculator reports that
  - $n = 8$
  - $\Sigma x = 56$
  - $\Sigma x^2 = 542$
  - $\Sigma y = 120$
  - $\Sigma y^2 = 2006$
  - $\Sigma xy = 1005$

- Use `2-Var Stats` to get the basic summations.
- Then use the formulas.

# The Regression Line on the TI-83

## TI-83 The Regression Line (LinReg(a+bx))

- Put the *x* values in $L_1$.
- Put the *y* values in $L_2$.
- Select STAT > CALC > LinReg(a+bx) (item #8).
- Press Enter. LinReg(a+bx) appears in the display.
- Enter $L_1$,$L_2$.
- Press ENTER.

- Or, use the LinReg(a+bx) function.

# The Regression Line on the TI-83

## TI-83 The Regression Line (`LinReg(a+bx)`)

- The following appear in the display.
  - The title `LinReg`.
  - The equation `y=a+bx`.
  - The value of $a$.
  - The value of $b$.
  - The value of $r^2$ (to be discussed later).
  - The value of $r$ (to be discussed later).

# The Regression Line on the TI-83

## TI-83 Graphing the Regression Line

- Follow procedure for using `LinReg(a+bx)`, except...
- Enter `LinReg(a+bx)` $L_1, L_2,$
- Press `VARS > Y-VARS > Function >` $Y_1$. Now the display shows `LinReg(a+bx)` $L_1, L_2, Y_1$
- Press `ENTER`.
- Press `ZOOM > ZoomStat` to draw the graph.
- Press `Y=` to see the regression equation.

# Outline


1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# TI-83 - Prediction

## TI-83 Prediction

- Press `VARS > Y-VARS > Function > ` $Y_1$.
- Press `ENTER`.
- Press `(`.
- Enter the value of *x*.
- Press `)`.
- Press `ENTER`. The value predicted by the model appears.

- Once the function is entered as $Y_1$, it is easy to interpolate and extrapolate.

# Outline


1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# Example

## Example (Free Lunch Rate vs. Graduation Rate)

- Find the equation of the regression line for the school-district data on the free-lunch participation rate vs. the graduation rate.
- Let $x$ be the free-lunch participation.
- Let $y$ be the graduation rate.

# Example

## 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 | | | |

# Example

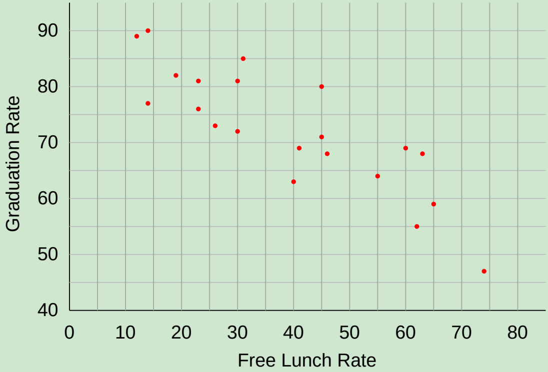
## Example (Free Lunch Rate vs. Graduation Rate)

- The regression equation is

$$\hat{y} = 91.047 - 0.494x.$$

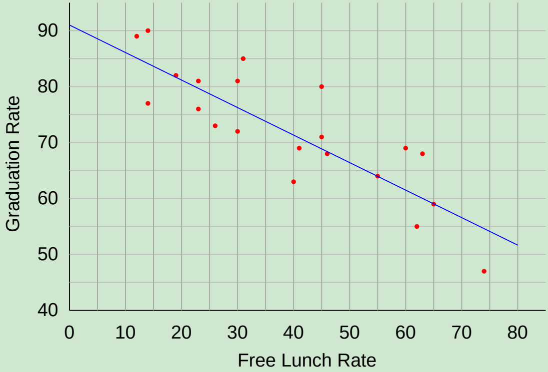
# Example

## Example (Free Lunch Rate vs. Graduation Rate)



# Example

## Example (Free Lunch Rate vs. Graduation Rate)



# Example

## Example (Predicting $\hat{y}$)

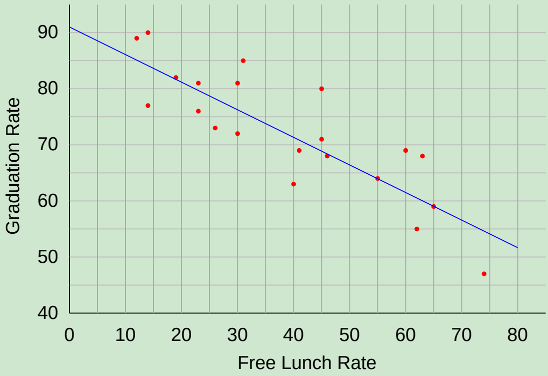
- What graduation rate would we predict in a district if we knew that the free-lunch participation rate was 50%?
- Calculate

$$\hat{y}(50) = 91.047 - 0.494(50) = 66.347.$$

- The model predicts a graduation rate of 66.3%.

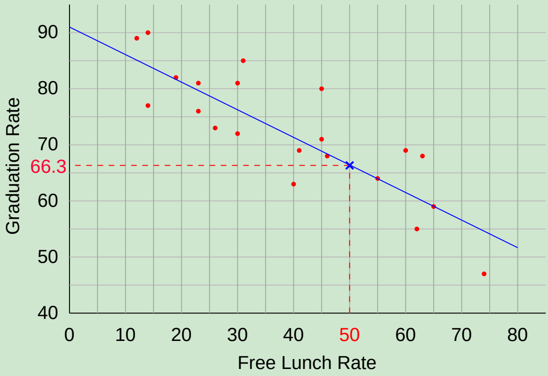
# Example

## Example (Predicting $\hat{y}$)



# Example

## Example (Predicting $\hat{y}$)



# Outline


1. Which Line is Better?
   - Measuring the Goodness of Fit
2. Formulas for the Regression Line
3. The Regression Line on the TI-83
4. Prediction
5. Free Lunch vs. Graduation Rate
6. Assignment

# Assignment

## Homework

- Read Section 13.3.2, page 815 - 820.
- Let's Do It! 13.3.
- Exercises 3(bc), 4(bcd), 5(b), 6, page 821.