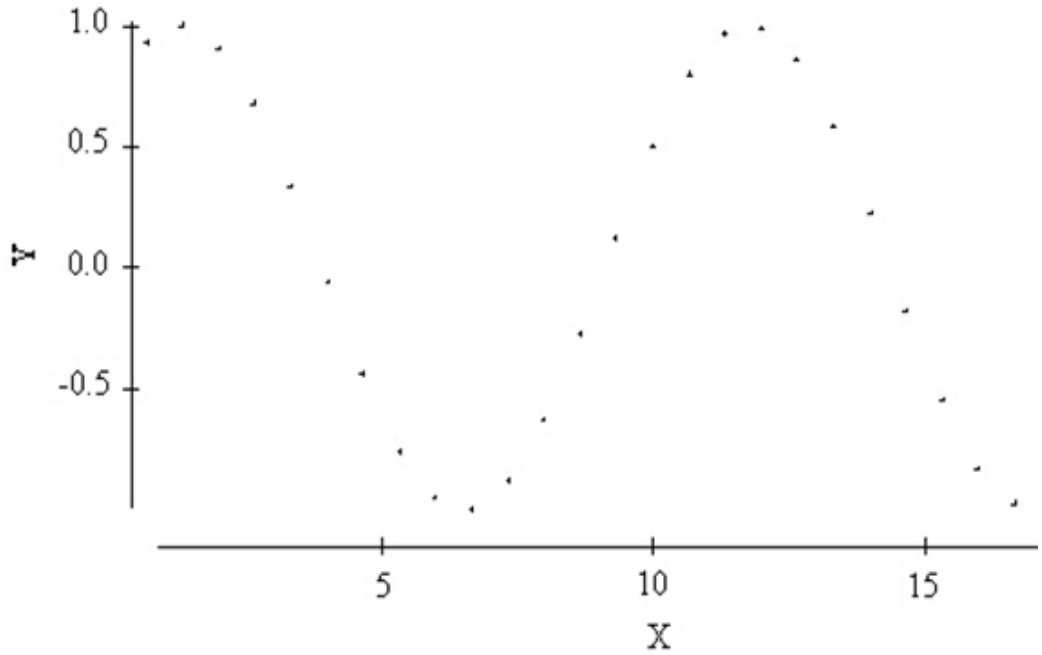
Simple Linear Regression

Lecture 15

02/15/2013

Turn on your clickers!

The scatterplot of a dataset is shown below. We notice a distinct curved pattern in the plot. It would be appropriate to conclude:



(a) r is small

(b) r is approximately -2/3 because Y decreases as X increases in approximately 2/3 of the plot.

(c) r is meaningless here

Example: *Midterm vs. final exam scores*

The *average* score in an exam was 80 (out of 100) and the distribution of the scores was quite symmetric.

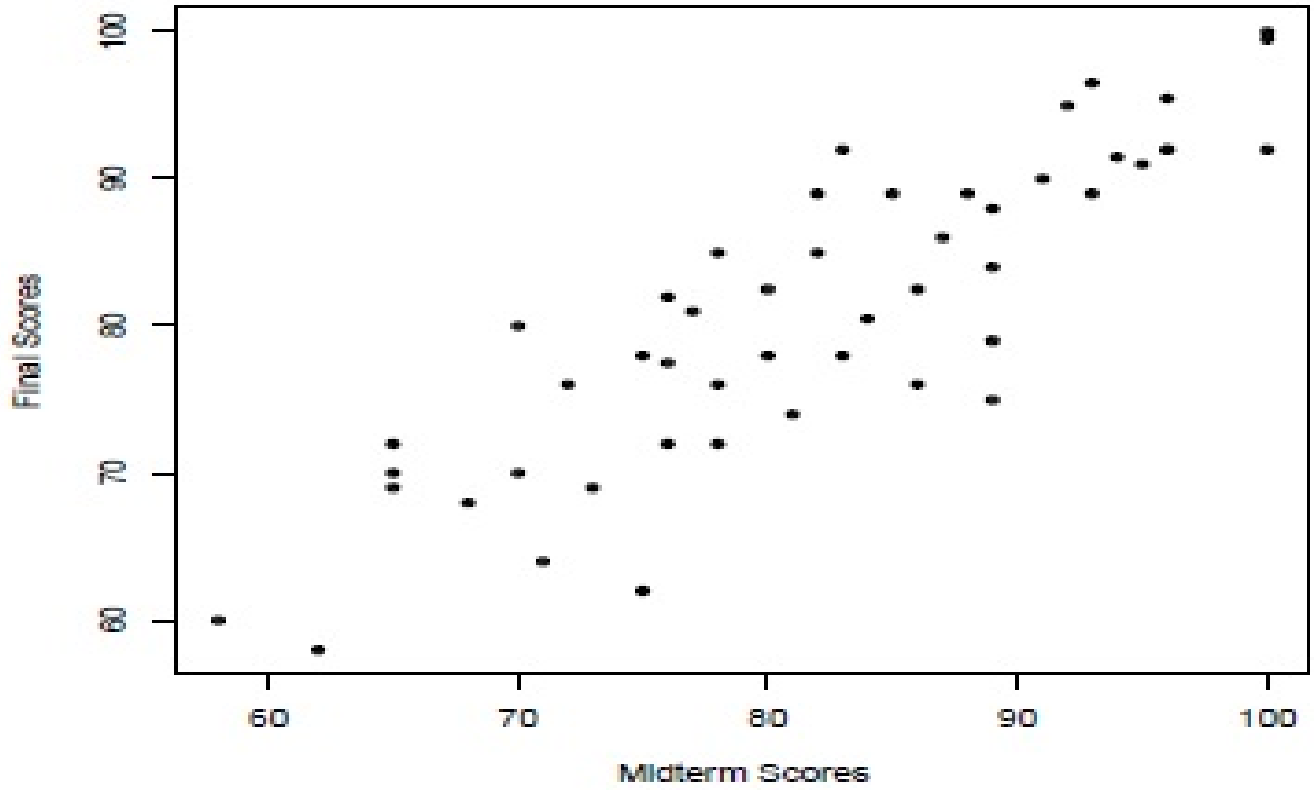
- If you meet a student from this class and you do not have any other information about him/her, what would be your guess for his/her final score?
  - 80/100: the average of the class.
- However, if this student was telling you that his/her midterm score was 90/100, would you be able to make a better guess?
  - Yes! We can do so using the linear regression.

Example: *Midterm vs. final exam scores*

- **Y** the variable that we want to predict/explain (i.e. the final score)
  - Y is called *response*
- **X** the variable that we use to make the prediction (i.e. the midterm score).
  - X is called the *predictor*.

Scatter Plot: *Final vs. Midterm*

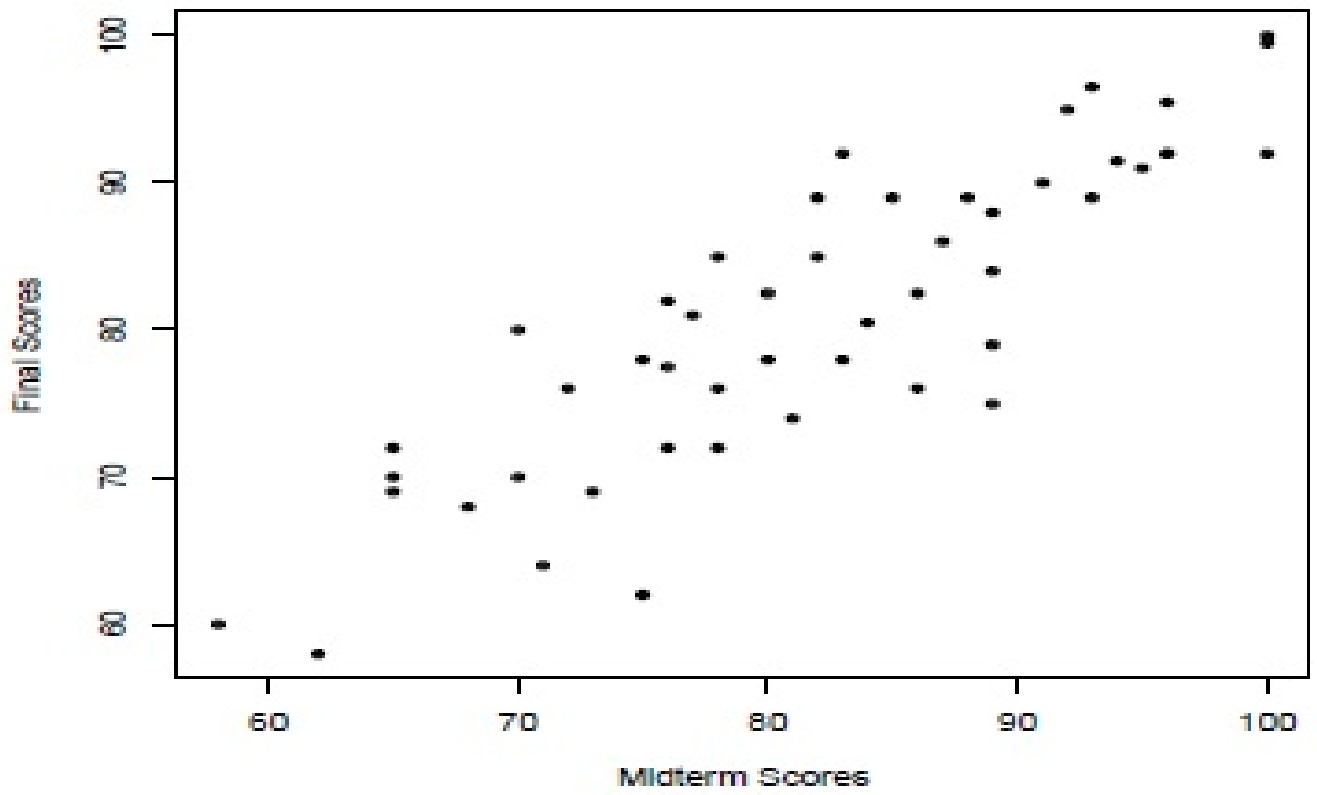
- Y: response
- X: predictor



The Regression line

- When the *scatterplot* looks like a **football-shaped cloud** of points, a straight line seems to be a good summary/fit.

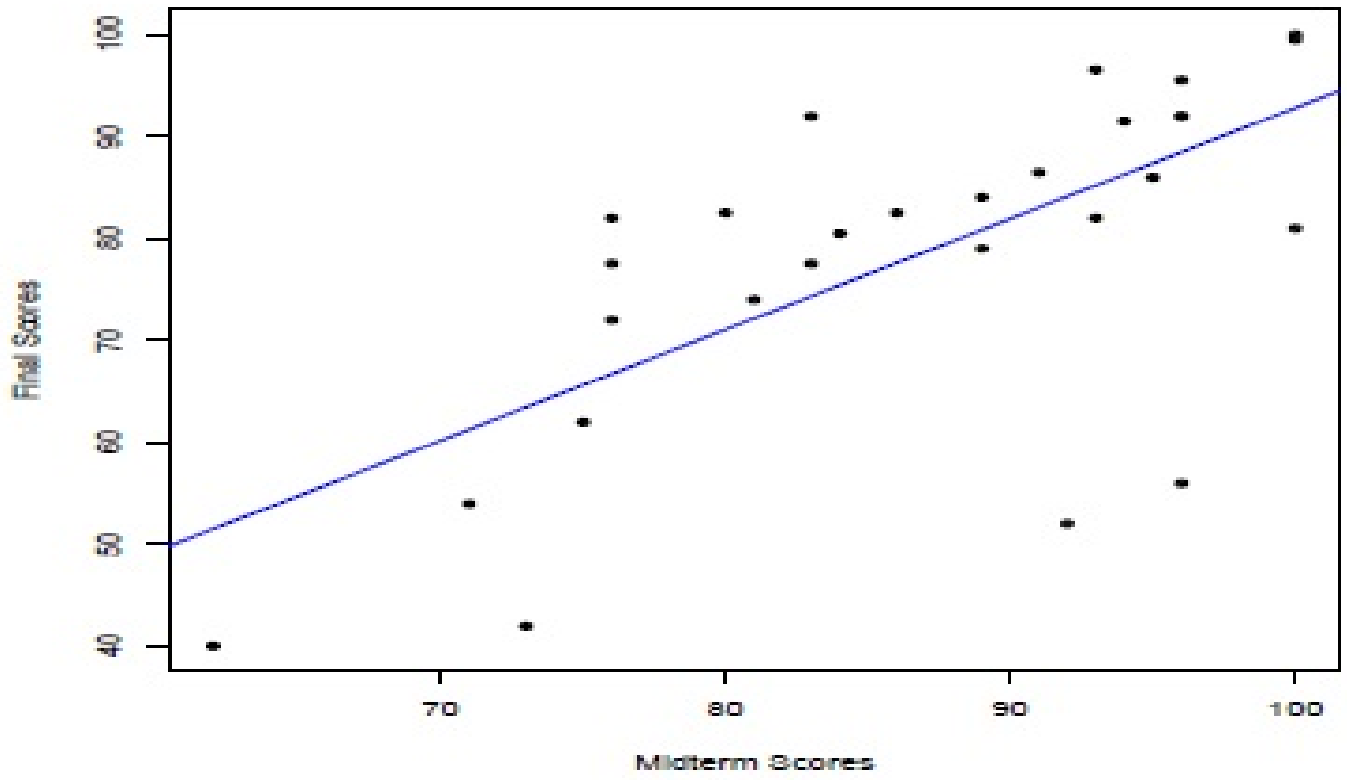
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

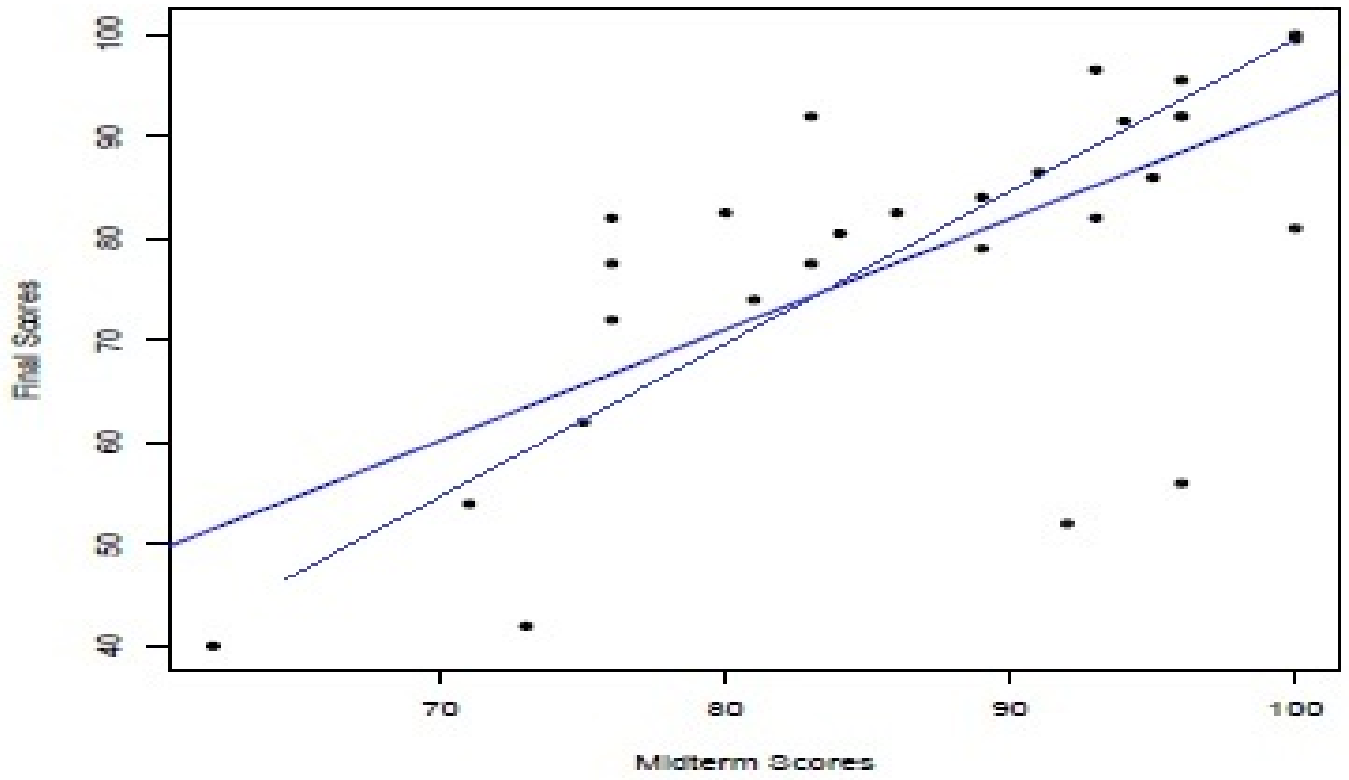
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

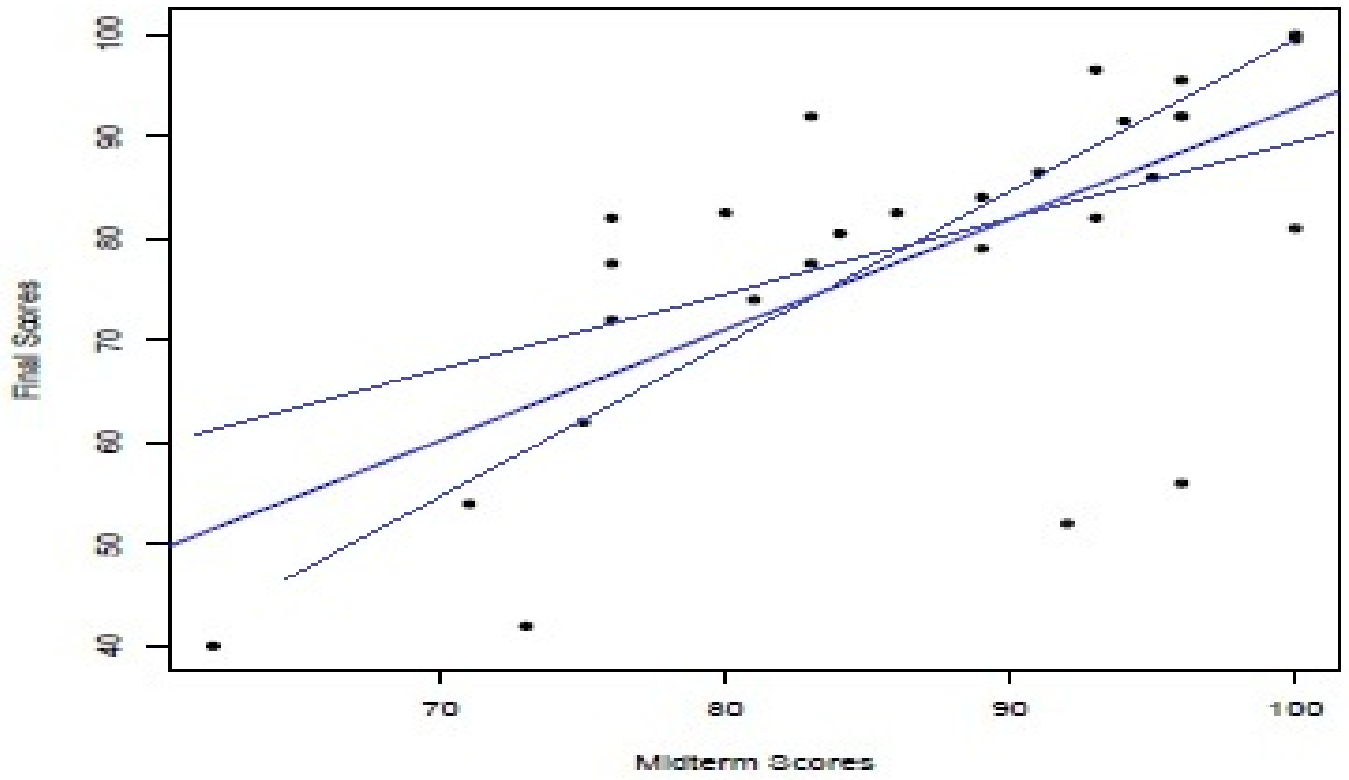
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

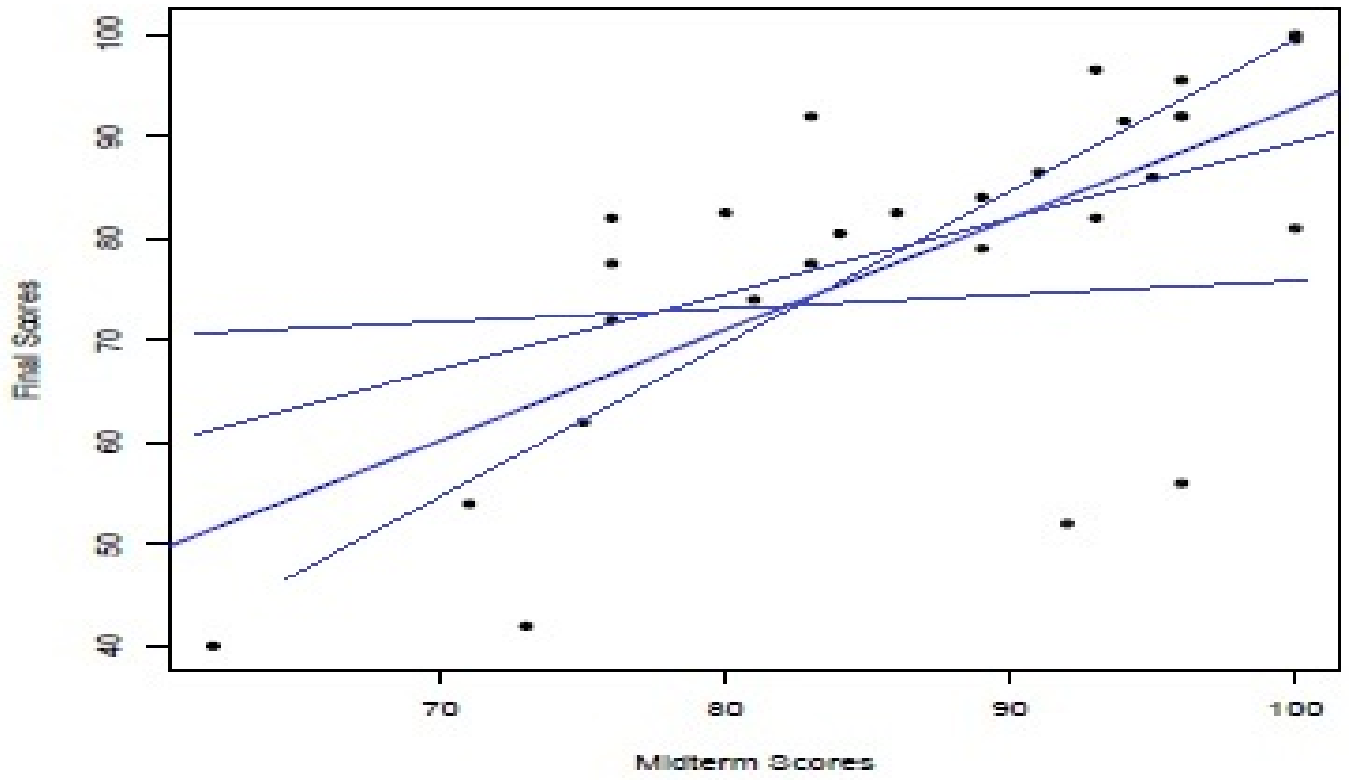
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

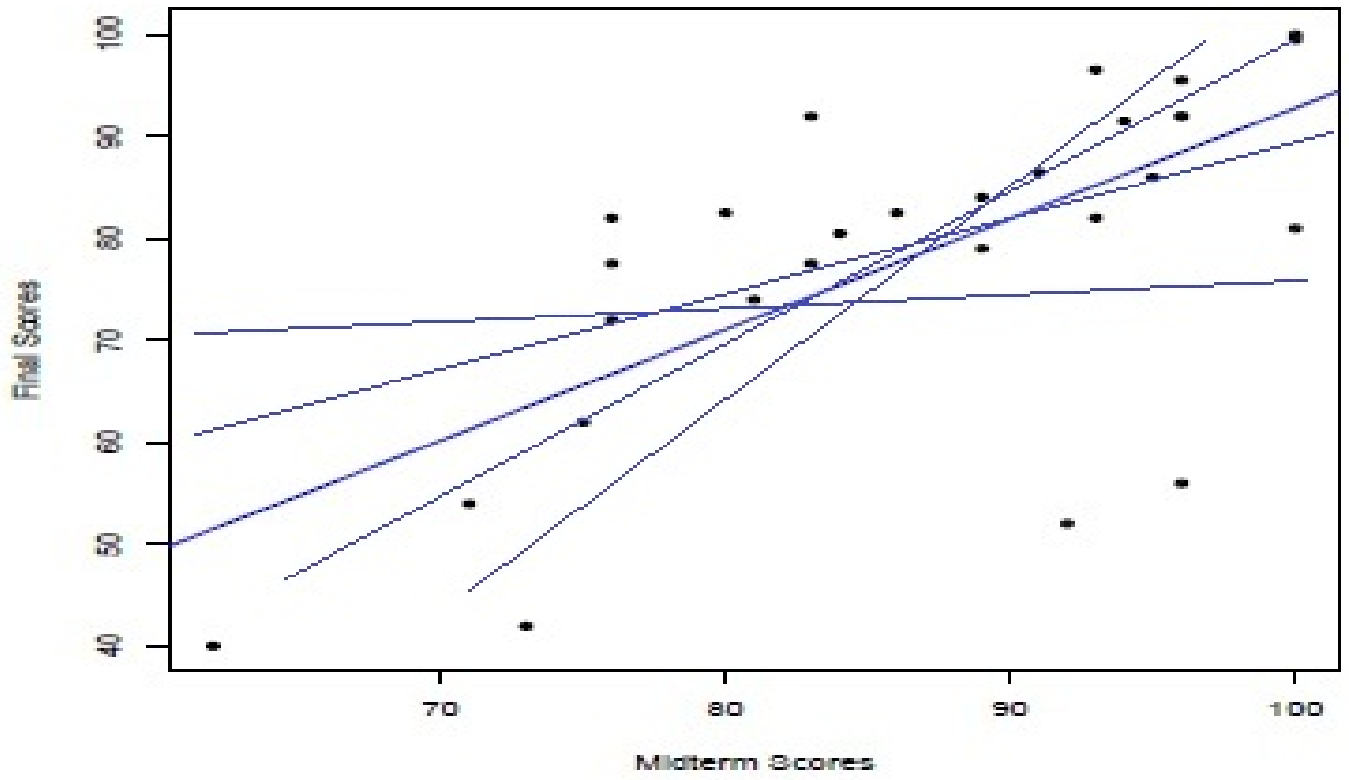
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

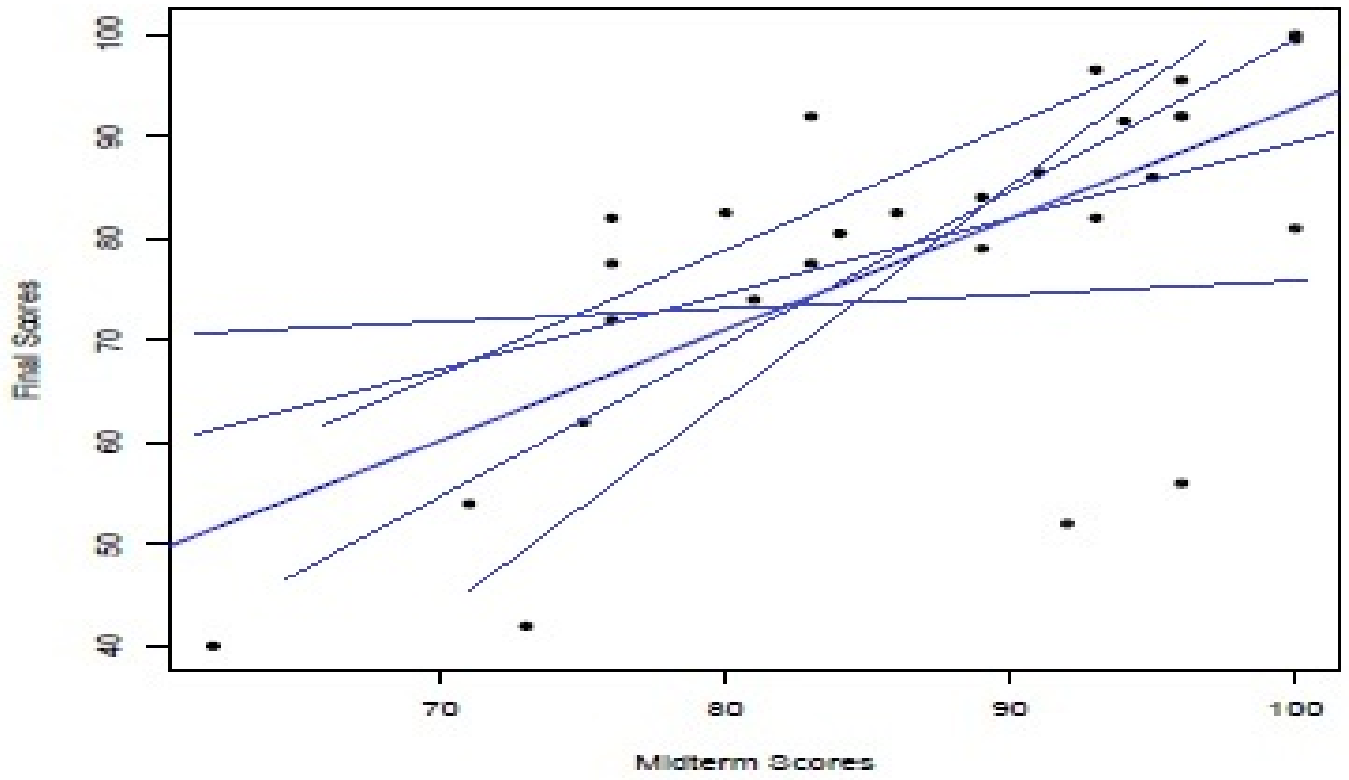
Final vs. Midterm



The Regression line

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

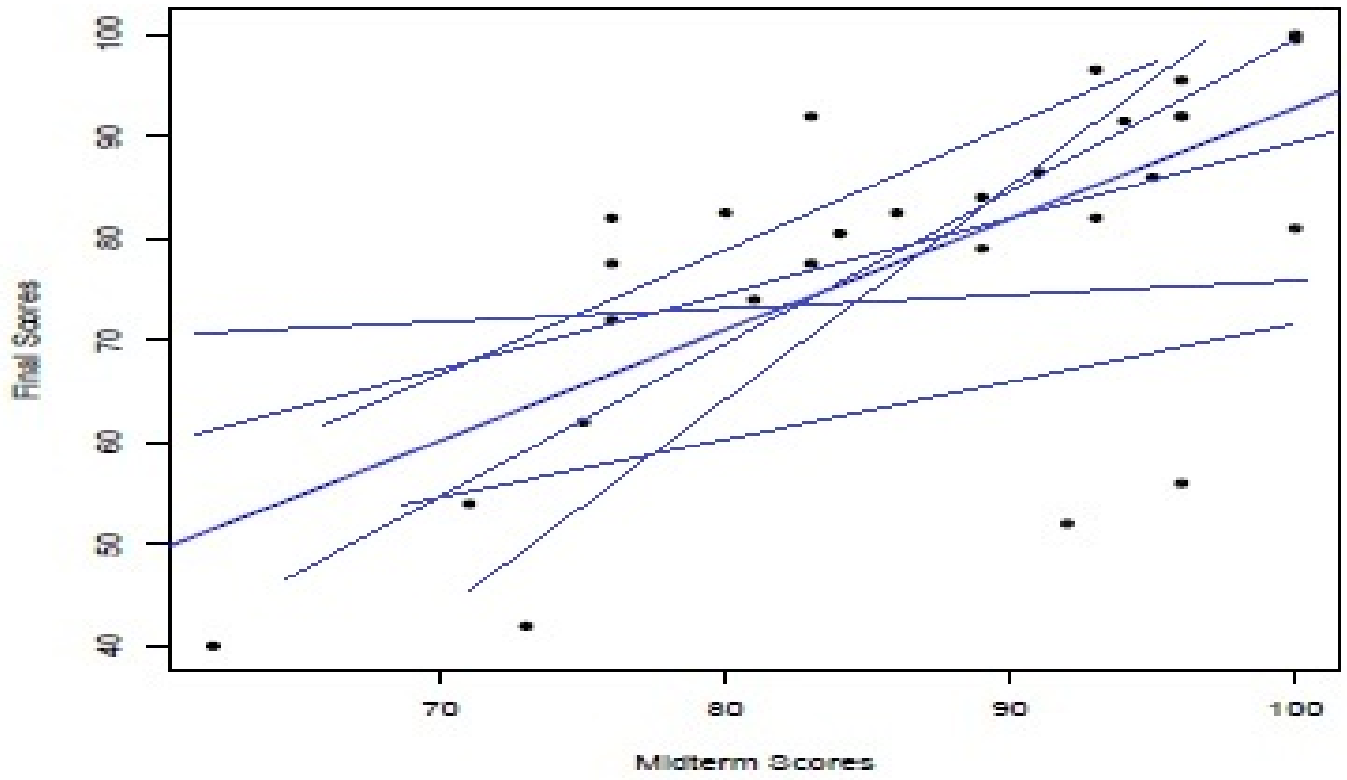
Final vs. Midterm



Lines to summarize the scatterplot

- There are many possible lines that could be used to summarize this scatterplot. Which one should we pick?

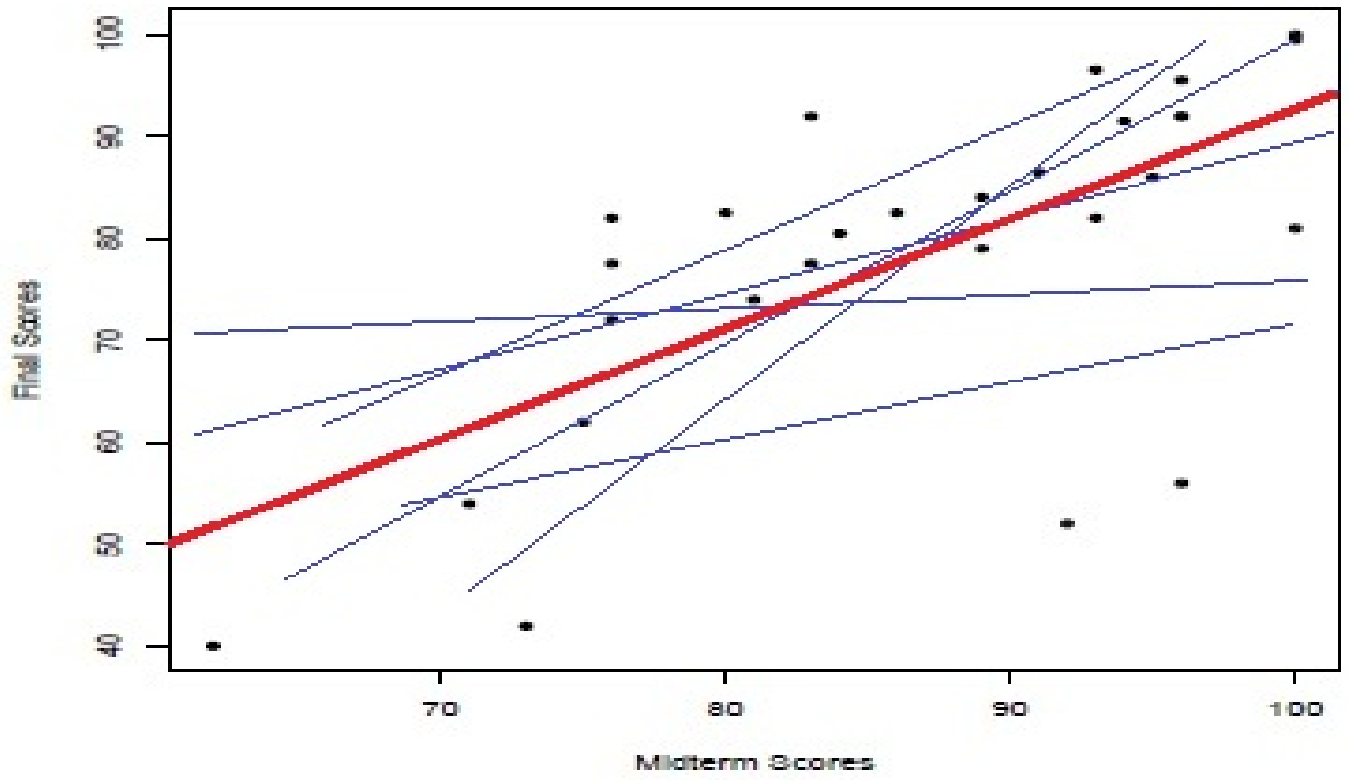
Final vs. Midterm



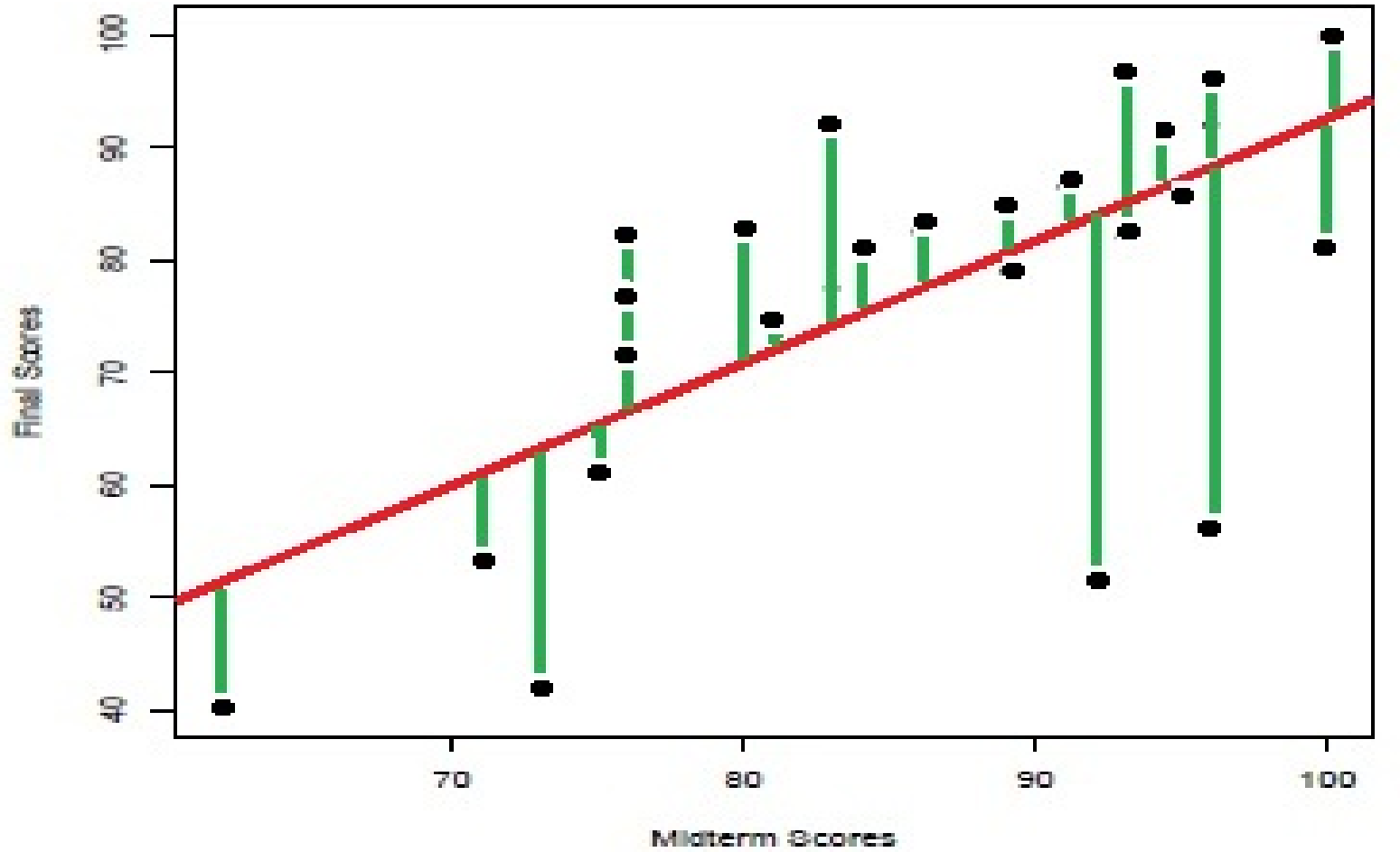
The Regression Line

- According to the method of least squares, the *best* line that summarizes the scatterplot is

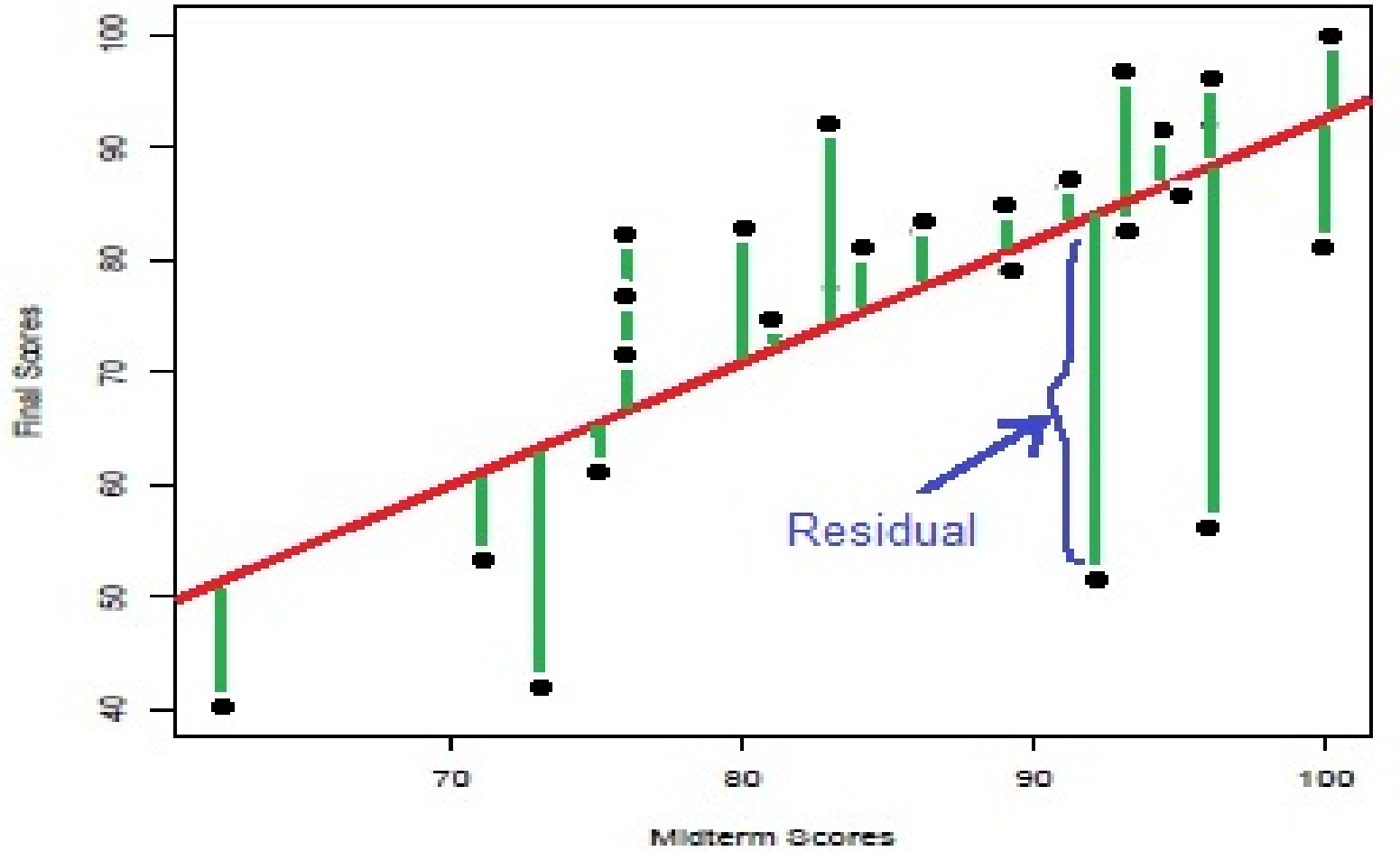
Final vs. Midterm



Error

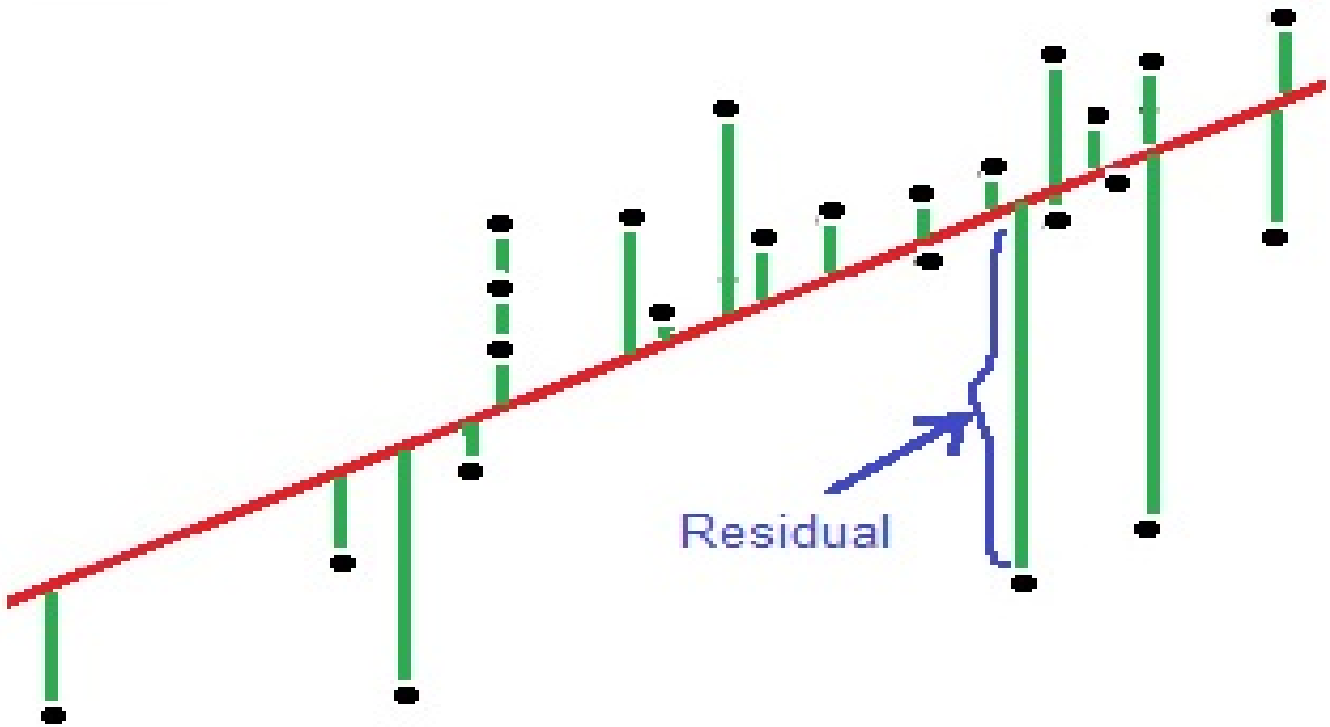


Residuals


Final Scores
40
50
60
70
80
90
100
70
80
90
100
Midterm Scores
Residual

Least Squares Line

- Error at each observation
- Residual = $y_i - \hat{y}_i = y_i - (b_0 + b_1\ x_1)$
- Minimize the sum of squared residuals
- Keep the total error as small as possible



How to compute the Regression line

Least-Squares Linear Regression

$$Y = b_0 + b_1 X,$$

where

$$b_1 = r\left(\frac{s_Y}{s_X}\right)$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

- b_0: *intercept*
- b_1: *slope*

How to compute the Regression line

- The slope and the intercept are summary statistics, since they are computed based on the data and they are used to summarize the relationship between the two variables.

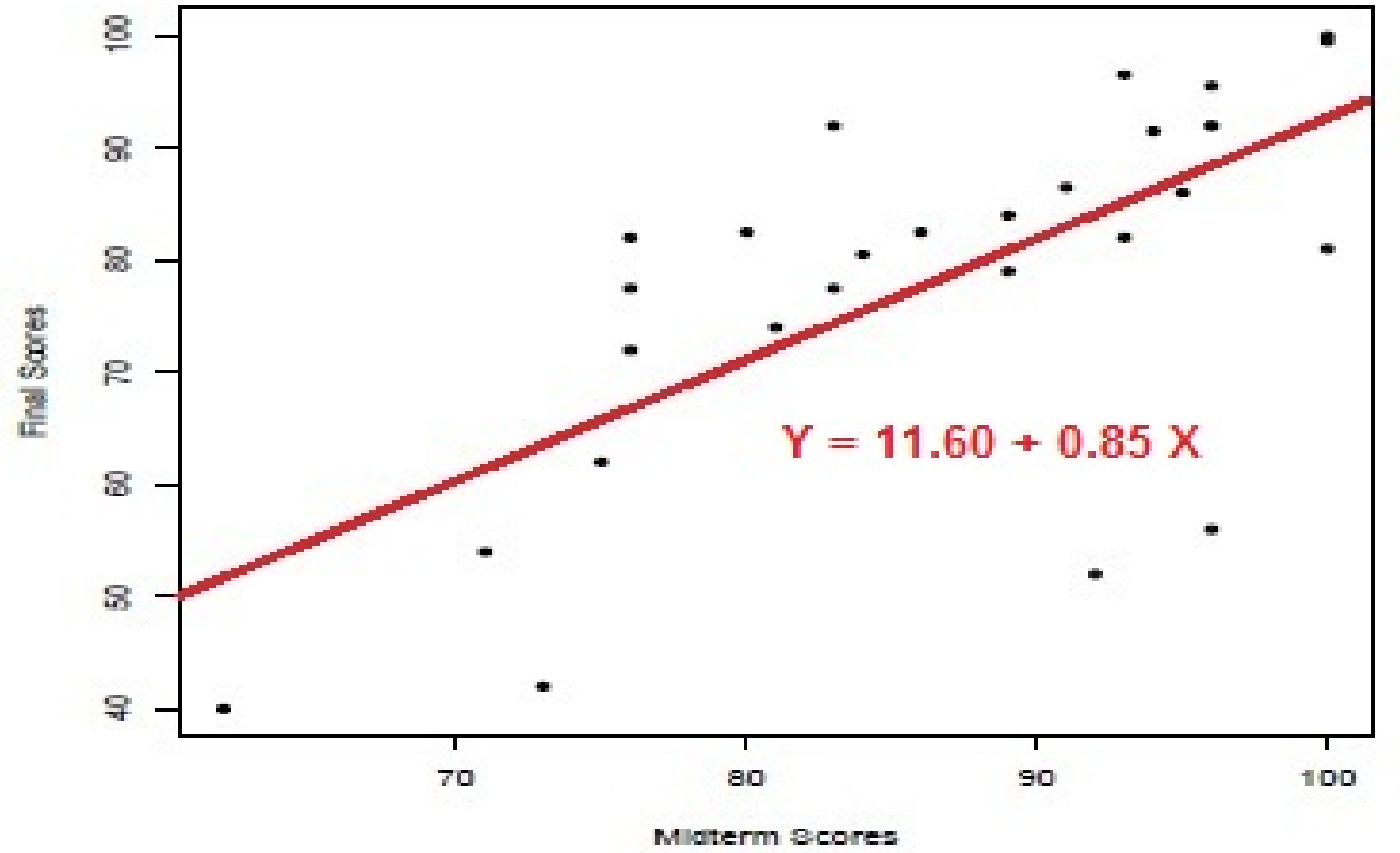
Midterm vs. final score Example

$\bar{y} = 79.6$	$s_y = 2.1$
$\bar{x} = 80$	$s_x = 1.86$
$r = 0.75$	

$$b_1 = 0.75 \cdot \frac{2.1}{1.86} = 0.85, \quad b_0 = 79.6 - 0.85\,(80) = 11.60$$

$$Y = 11.60 + 0.85 \cdot X$$

Add the regression line to the scatterplot



Turn on your clickers!

Which of the following is correct with repsect to the correlation coefficient r and the slope of the regression line?

(a) They will always have the same sign.

(b) They will have opposite signs.

(c) Nothing, because they are two different measures that are not related one to another.

Interpretation of the regression line

- A student that you met from last year's PSTAT 5A scored 70/100 on the midterm.
- What is the prediction for his/her final score?
  (a) 80
  (b) 71.10
  (c) 81.60
  (d) We do not know

Interpretation of the regression line

- A student that you met from last year's PSTAT 5A scored 70/100 on the midterm.
- What is the prediction for his/her final score?

$$\hat{Y} = 11.60 + 0.85 \cdot 70 = 71.10.$$

$$\text{(Final Score)} = 11.60 + 0.85 \cdot \text{(Midterm Score)}$$

Interpretation of the regression line

If his/her midterm score is X,
the **prediction** for his/her final score will be

$$\hat{Y} = b_0 + b_1 X$$

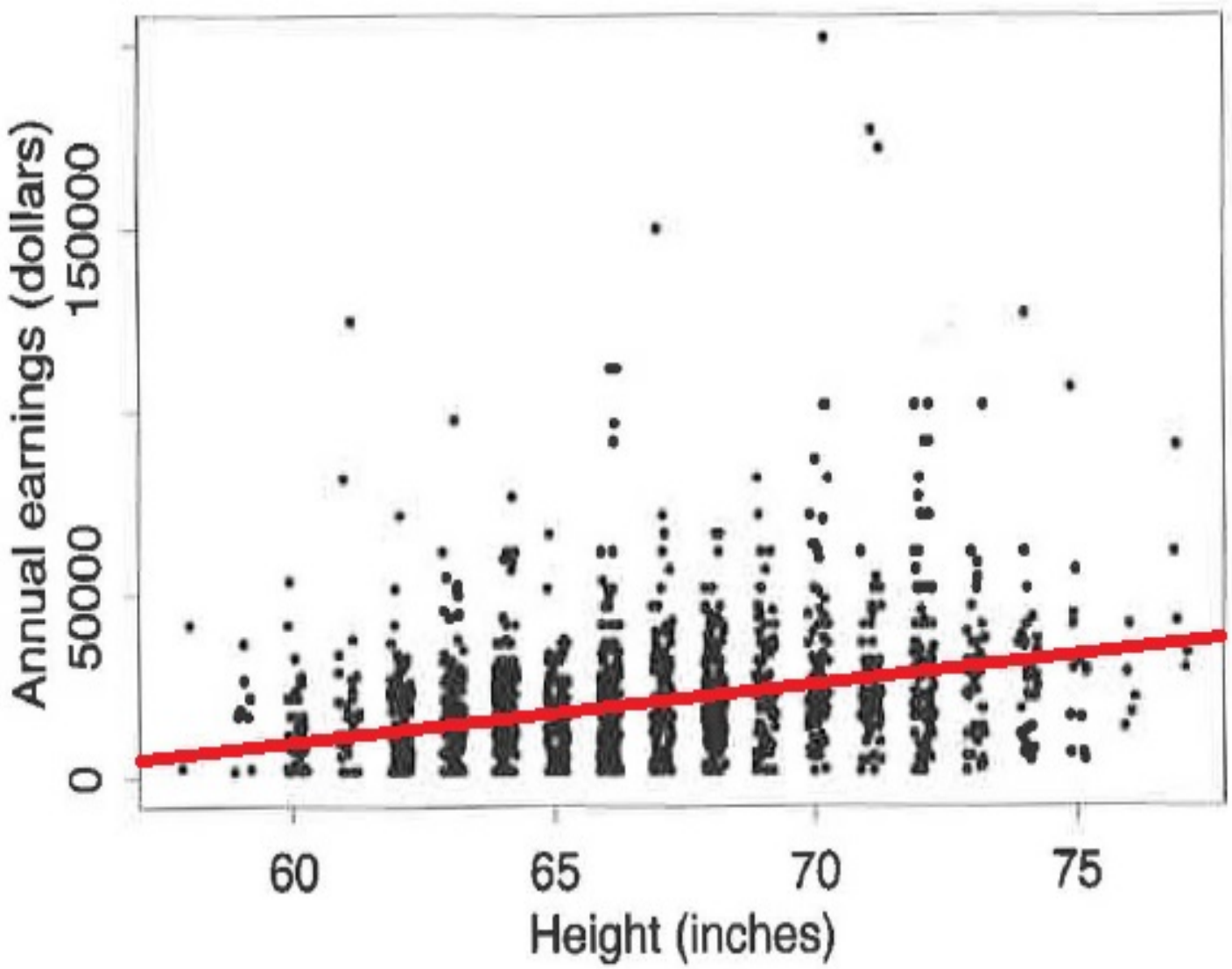
Interpretation of the slope

- Pick two students at random from the class whose midterms differ by one point.
- The *prediction* for the difference in their final scores is b_1. Indeed,

$$\begin{aligned} Y &= b_0 + b_1 X \\ Y &= b_0 + b_1 (X + 1) \end{aligned}$$

$$b_1 = \Big(b_0 + b_1 (X + 1)\Big) - \Big(b_0 + b_1 X\Big).$$

Example: *Taller people have higher incomes?*



Example: *Taller people have higher incomes?*

Regression Line

- Y response variable: **income**
- X predictor variable: **height**

$$Y = -84,000 + 1,560 \cdot X,$$

Example: *Taller people have higher incomes?*

Interpretation of the Intercept ($b_0 = -84,000$)

- -84,000 is the Y-value of the regression line when the X variable is equal to zero.
- The predicted value of income for an adult who is *zero inches tall* is -84,000.
- Such an *extrapolation* is meaningless!

Example: *Taller people have higher incomes?*

Interpretation of the Slope ($b_1 = 1,560$)

- The slope is positive.
- A positive slope indicates that *if one person is one inch higher than another one, then the prediction for the difference in their salaries is $1,560.*
- This implies that *taller people are more likely to have higher earnings*.