

# Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 4

# Introduction to Regression

## How to fairly account for missing midterm score?

- ▶ In my first semester at Penn, several students missed Midterm 2 because of illness so I decided to up-weight their finals.
- ▶ Problem: Midterm 2 turned out easier than Midterm 1 and this put the students who had missed the second midterm at a disadvantage when I curved the class.
- ▶ In order to correct for this, I needed a way to *fill in* a score for the missing midterm.
- ▶ How could I do this fairly?

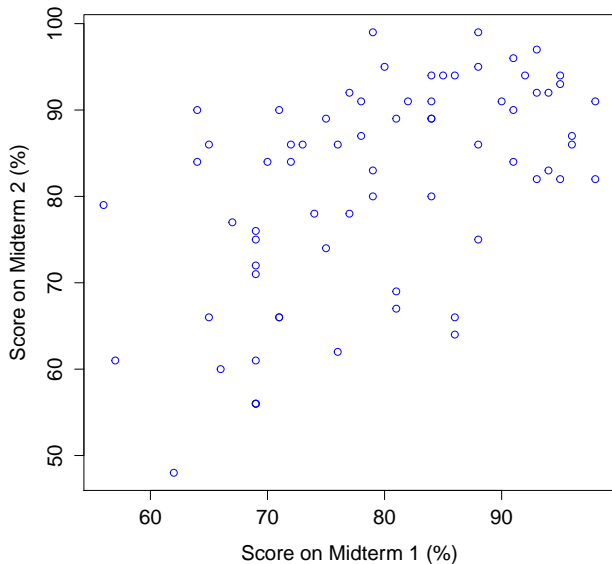
## How to fairly account for missing midterm score?

- ▶ In my first semester at Penn, several students missed Midterm 2 because of illness so I decided to up-weight their finals.
- ▶ Problem: Midterm 2 turned out easier than Midterm 1 and this put the students who had missed the second midterm at a disadvantage when I curved the class.
- ▶ In order to correct for this, I needed a way to *fill in* a score for the missing midterm.
- ▶ How could I do this fairly?
  - ▶ Just fill in mean score on second exam?

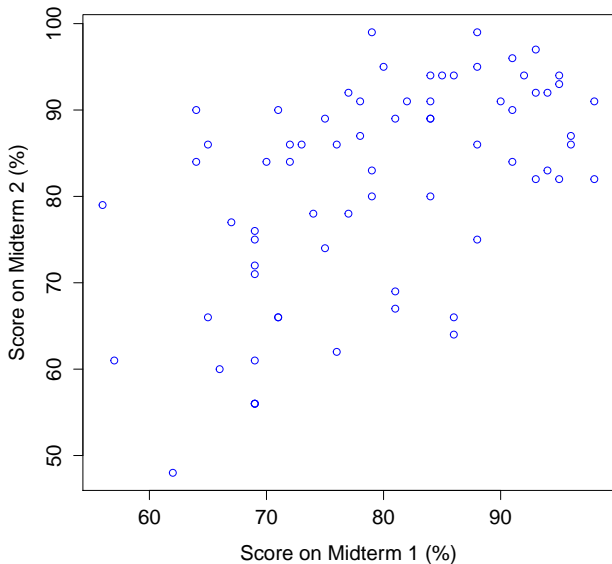
## How to fairly account for missing midterm score?

- ▶ In my first semester at Penn, several students missed Midterm 2 because of illness so I decided to up-weight their finals.
- ▶ Problem: Midterm 2 turned out easier than Midterm 1 and this put the students who had missed the second midterm at a disadvantage when I curved the class.
- ▶ In order to correct for this, I needed a way to *fill in* a score for the missing midterm.
- ▶ How could I do this fairly?
  - ▶ Just fill in mean score on second exam?
  - ▶ Use performance on first midterm to predict?

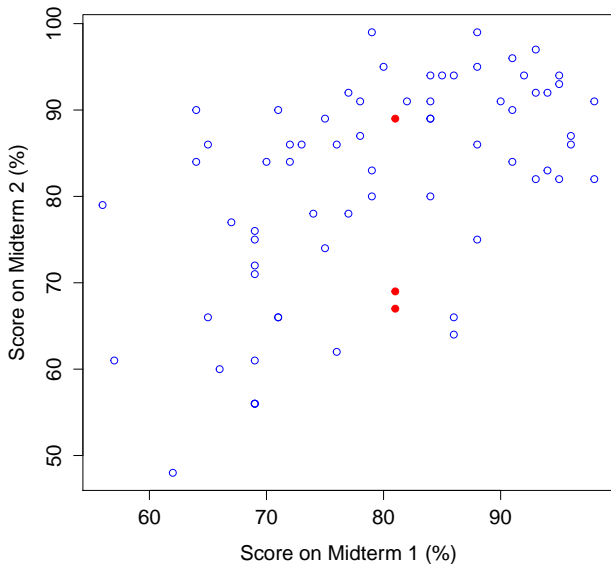
## Data for students who took both midterms:



## Predict Second Midterm given 81 on First

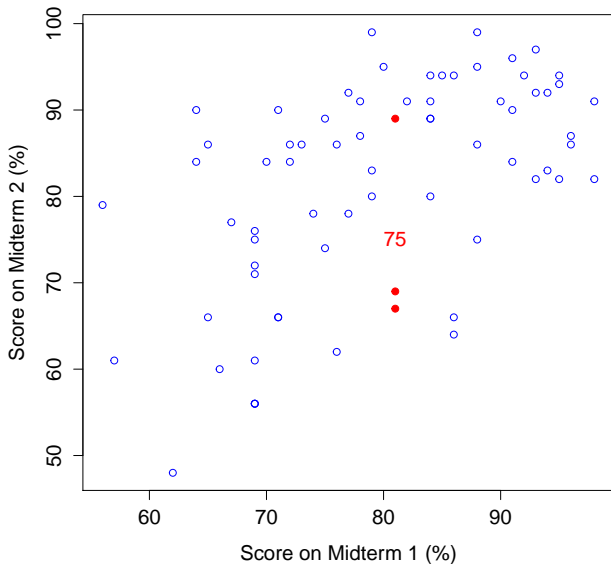


## Predict Second Midterm given 81 on First

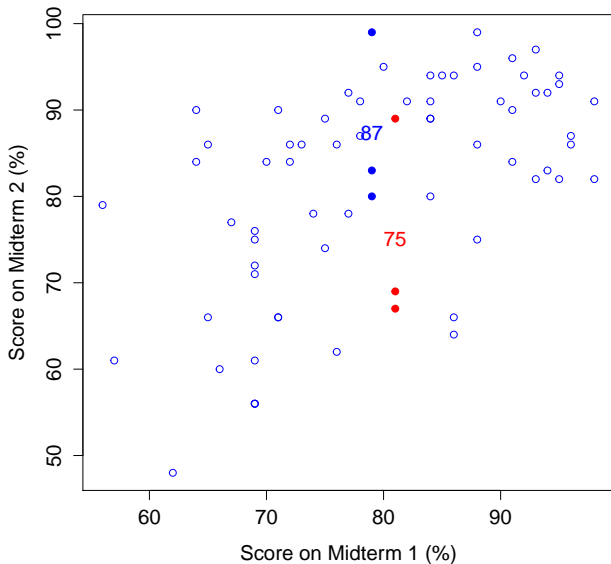




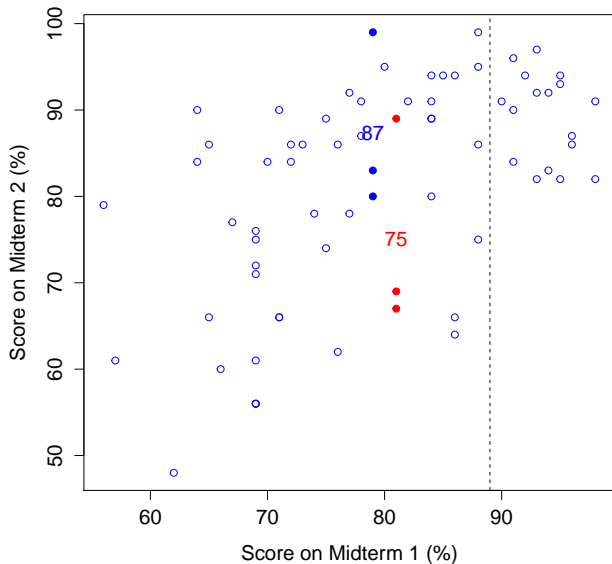
## Predict Second Midterm given 81 on First



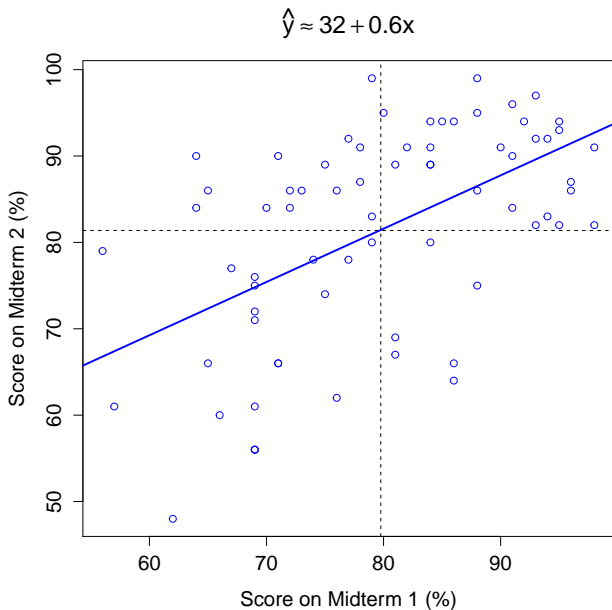
But if they'd only gotten 79 we'd predict higher?!



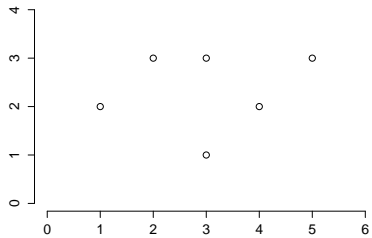
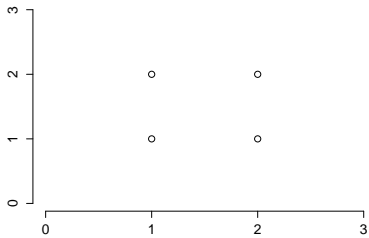
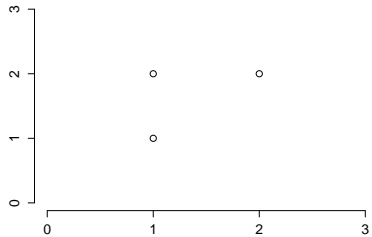
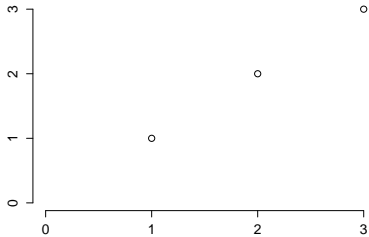
No one who took both exams got 89 on the first!

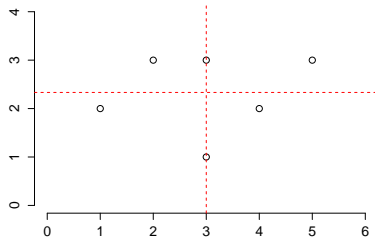
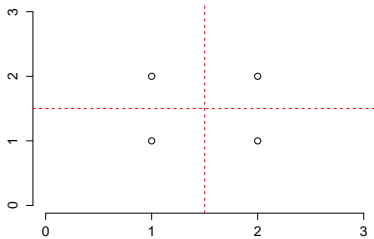
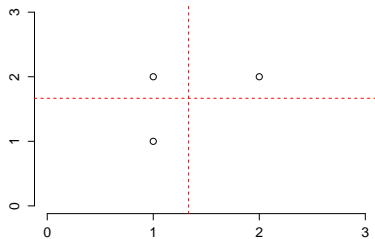
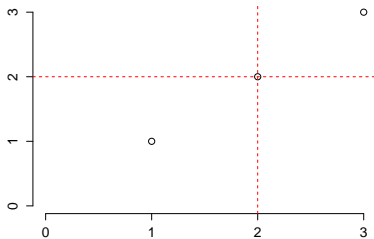


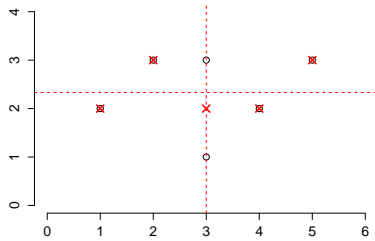
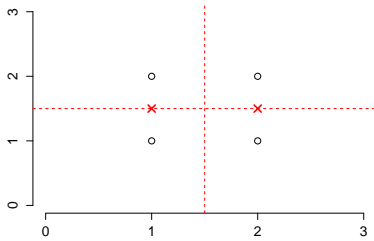
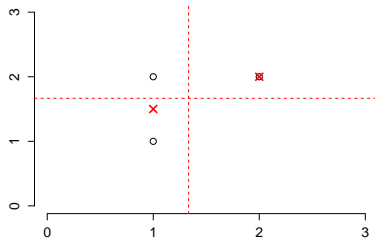
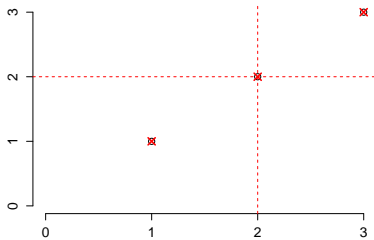
## Regression: “Best Fitting” Line Through Cloud of Points



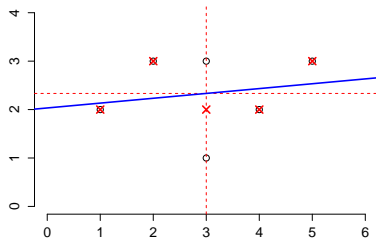
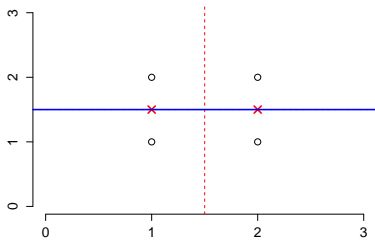
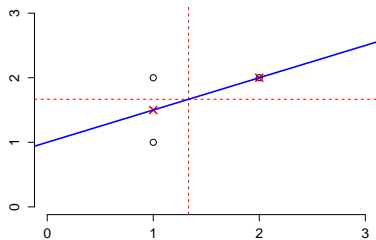
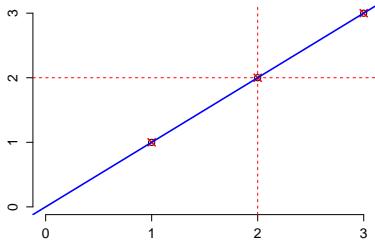
# Fitting a Line by Eye











But How to Do this Formally?

# Least Squares Regression – Predict Using a Line

Linear Model

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

# Least Squares Regression – Predict Using a Line

## Linear Model

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

Choose  $a, b$  to Minimize Sum of Squared Vertical Deviations

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

# Least Squares Regression – Predict Using a Line

## Linear Model

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

Choose  $a, b$  to Minimize Sum of Squared Vertical Deviations

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

## The Prediction

Predict score  $\hat{y} = a + bx$  on second midterm for someone with score  $x$  on first.

# Least Squares Regression – Predict Using a Line

## Linear Model

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

Choose  $a, b$  to Minimize Sum of Squared Vertical Deviations

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

## The Prediction

Predict score  $\hat{y} = a + bx$  on second midterm for someone with score  $x$  on first.

Why Vertical Deviations? Why Squared Deviations?

## Important Point About Notation

$$\text{minimize } \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2$$

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

- ▶  $(x_i, y_i)_{i=1}^n$  are the **observed data**
- ▶  $\hat{y}$  is our **prediction** for a given value of  $x$
- ▶ Neither  $x$  nor  $\hat{y}$  needs to be in our dataset!

## Key Point

- ▶ Each choice of  $a, b$  defines a line



# Key Point

- ▶ Each choice of  $a, b$  defines a line
- ▶ Given the data, each line defines collection of vertical devs.  
 $d_i = y_i - a - bx_i$  for  $i = 1, \dots, n$

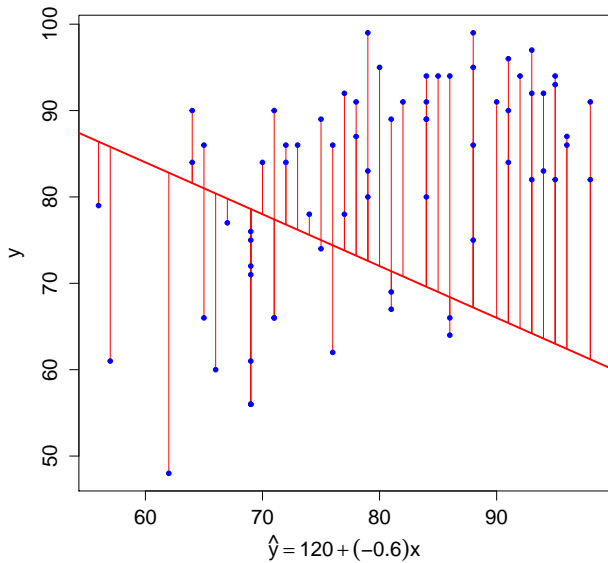
## Key Point

- ▶ Each choice of  $a, b$  defines a line
- ▶ Given the data, each line defines collection of vertical devs.  
 $d_i = y_i - a - bx_i$  for  $i = 1, \dots, n$
- ▶ Each collection of vertical devs. gives sum of squares  $\sum_{i=1}^n d_i^2$

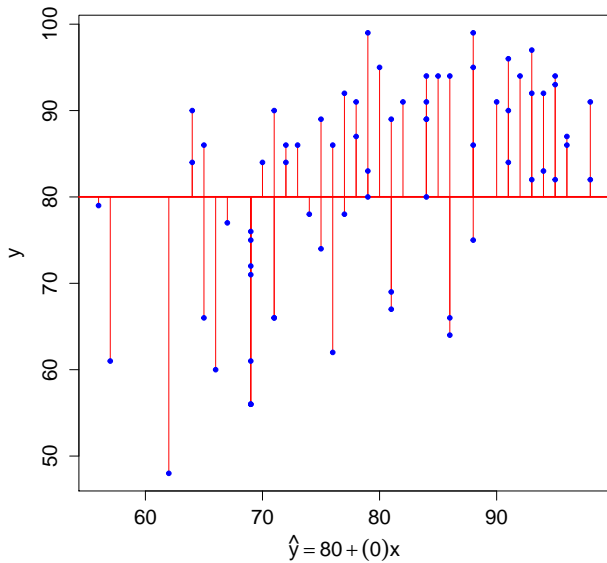
## Key Point

- ▶ Each choice of  $a, b$  defines a line
- ▶ Given the data, each line defines collection of vertical devs.  
 $d_i = y_i - a - bx_i$  for  $i = 1, \dots, n$
- ▶ Each collection of vertical devs. gives sum of squares  $\sum_{i=1}^n d_i^2$
- ▶ We choose  $a, b$  to minimize  $\sum_{i=1}^n d_i^2$

$$\sum d^2 = 25596.88$$



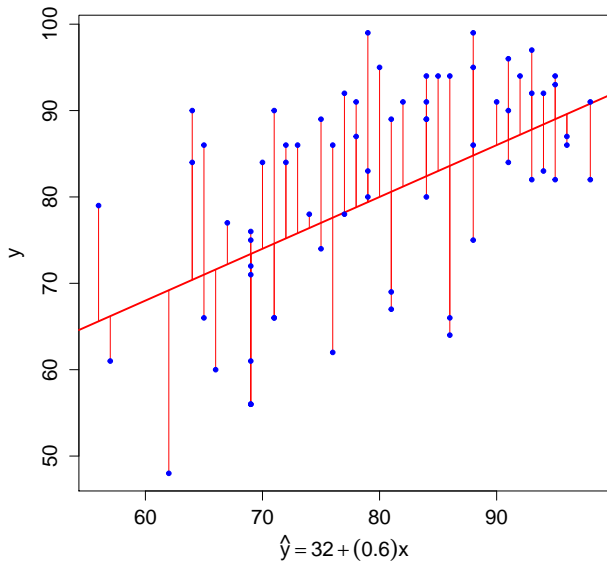
$$\sum d^2 = 10728$$



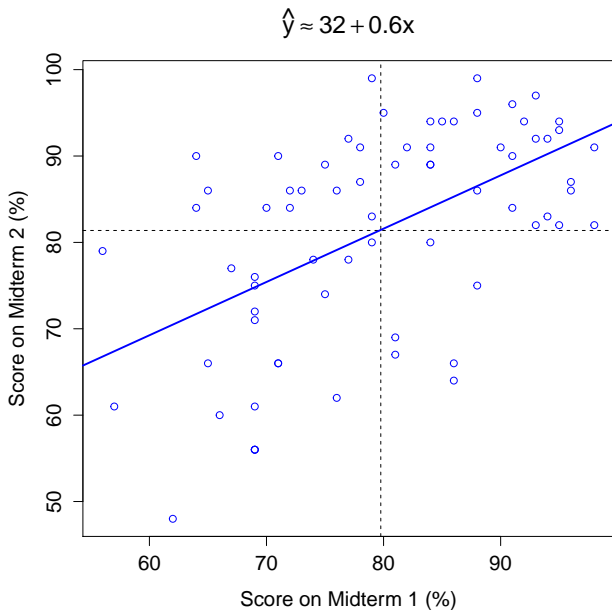
$$\sum d^2 = 8313.72$$



$$\sum d^2 = 7650.48$$

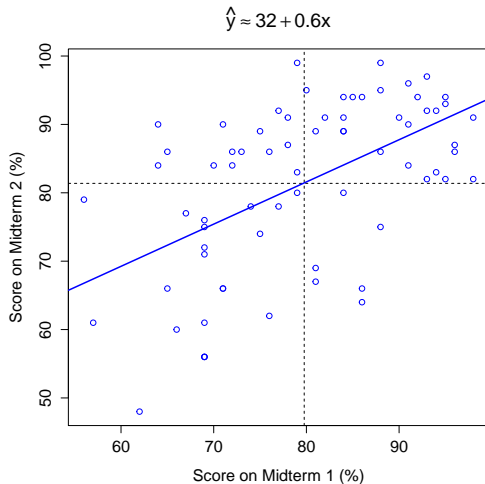


## Prediction given 89 on Midterm 1?





## Prediction given 89 on Midterm 1?



$$32 + 0.6 \times 89 = 32 + 53.4 = 85.4$$

# You Need to Know How To Derive This



Minimize the sum of squared vertical deviations from the line:

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

How should we proceed?

- (a) Differentiate with respect to  $x$
- (b) Differentiate with respect to  $y$
- (c) Differentiate with respect to  $x, y$
- (d) Differentiate with respect to  $a, b$
- (e) Can't solve this with calculus.

## Objective Function

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

FOC with respect to  $a$

## Objective Function

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

FOC with respect to  $a$

$$-2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

## Objective Function

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

FOC with respect to  $a$

$$-2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n a - b \sum_{i=1}^n x_i = 0$$

## Objective Function

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

FOC with respect to  $a$

$$-2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n a - b \sum_{i=1}^n x_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n y_i - \frac{na}{n} - \frac{b}{n} \sum_{i=1}^n x_i = 0$$

## Objective Function

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

FOC with respect to  $a$

$$-2 \sum_{i=1}^n (y_i - a - bx_i) = 0$$

$$\sum_{i=1}^n y_i - \sum_{i=1}^n a - b \sum_{i=1}^n x_i = 0$$

$$\frac{1}{n} \sum_{i=1}^n y_i - \frac{na}{n} - \frac{b}{n} \sum_{i=1}^n x_i = 0$$

$$\bar{y} - a - b\bar{x} = 0$$

## Regression Line Goes Through the Means!

$$\bar{y} = a + b\bar{x}$$



## Substitute: Eliminate $a$ from Objective Function

$$a = \bar{y} - b\bar{x}$$

$$\sum_{i=1}^n (y_i - a - bx_i)^2 =$$

## Substitute: Eliminate $a$ from Objective Function

$$a = \bar{y} - b\bar{x}$$

$$\begin{aligned}\sum_{i=1}^n (y_i - a - bx_i)^2 &= \sum_{i=1}^n (y_i - \bar{y} + b\bar{x} - bx_i)^2 \\ &= \end{aligned}$$

## Substitute: Eliminate $a$ from Objective Function

$$a = \bar{y} - b\bar{x}$$

$$\begin{aligned}\sum_{i=1}^n (y_i - a - bx_i)^2 &= \sum_{i=1}^n (y_i - \bar{y} + b\bar{x} - bx_i)^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2\end{aligned}$$

## Objective Function Without $a$

$$\sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

FOC with respect to  $b$

## Objective Function Without $a$

$$\sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

FOC with respect to  $b$

$$-2 \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})] (x_i - \bar{x}) = 0$$

## Objective Function Without $a$

$$\sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

FOC with respect to  $b$

$$-2 \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})] (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}) - b \sum_{i=1}^n (x_i - \bar{x})^2 = 0$$

## Objective Function Without $a$

$$\sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})]^2$$

FOC with respect to  $b$

$$-2 \sum_{i=1}^n [(y_i - \bar{y}) - b(x_i - \bar{x})] (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x}) - b \sum_{i=1}^n (x_i - \bar{x})^2 = 0$$

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y}) (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

# Simple Linear Regression

## Problem

$$\min_{a,b} \sum_{i=1}^n (y_i - a - bx_i)^2$$

## Solution

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$



## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} =$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} =$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} =$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} =$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} =$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b \frac{s_x}{s_y}$$

## Relating Regression to Covariance and Correlation

$$b = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b \frac{s_x}{s_y}$$

$$b = r \frac{s_y}{s_x}$$



# Comparing Regression, Correlation and Covariance

## Units

Correlation is unitless, covariance and regression coefficients ( $a$ ,  $b$ ) are not. (What are the units of these?)

## Symmetry

Correlation and covariance are symmetric, regression isn't. (Switching  $x$  and  $y$  axes changes the slope and intercept.)

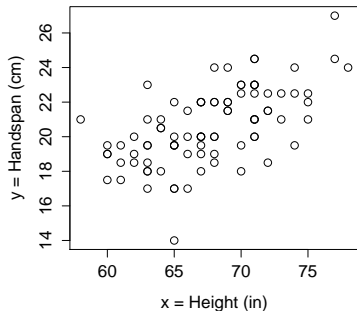
## On the Homework

Regression with z-scores rather than raw data gives  $a = 0$ ,  $b = r_{xy}$



$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

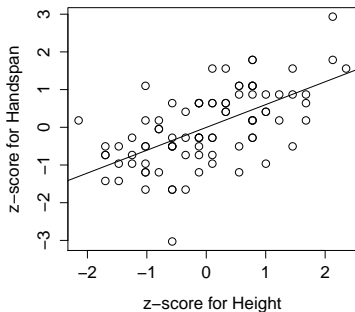
What is the sample correlation between height ( $x$ ) and handspan ( $y$ )?





$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the sample correlation between height ( $x$ ) and handspan ( $y$ )?



$$r = \frac{s_{xy}}{s_x s_y} = \frac{6}{5 \times 2} = 0.6$$



$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of  $b$  for the regression:

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

where  $x$  is height and  $y$  is handspan?



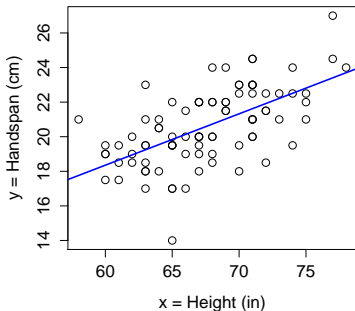


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of  $b$  for the regression:

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

where  $x$  is height and  $y$  is handspan?



$$b = \frac{s_{xy}}{s_x^2} = \frac{6}{5^2} = 6/25 = 0.24$$

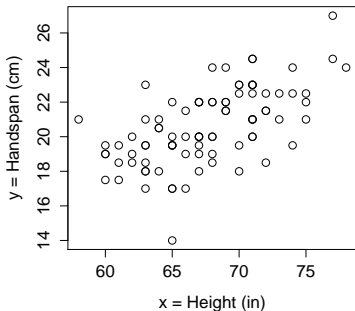


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of  $a$  for the regression:

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

where  $x$  is height and  $y$  is handspan?  
(prev. slide  $b = 0.24$ )



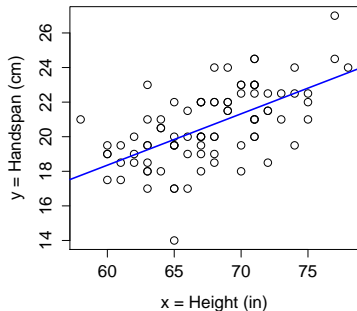


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of  $a$  for the regression:

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

where  $x$  is height and  $y$  is handspan?  
(prev. slide  $b = 0.24$ )



$$a = \bar{y} - b\bar{x} = 21 - 0.24 \times 68 = 4.68$$

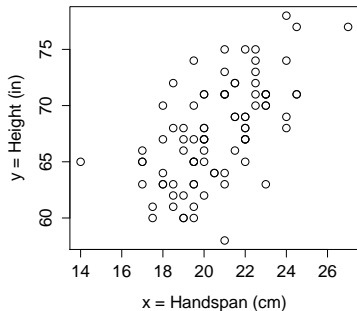


$$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$$

What is the value of  $b$  for the regression:

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

where  $x$  is handspan and  $y$  is height?





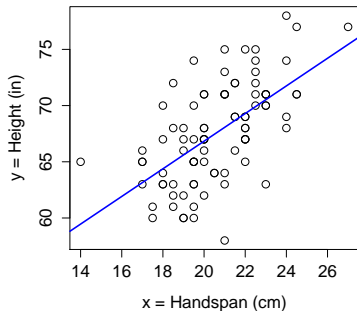


$$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$$

What is the value of  $b$  for the regression:

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

where  $x$  is handspan and  $y$  is height?

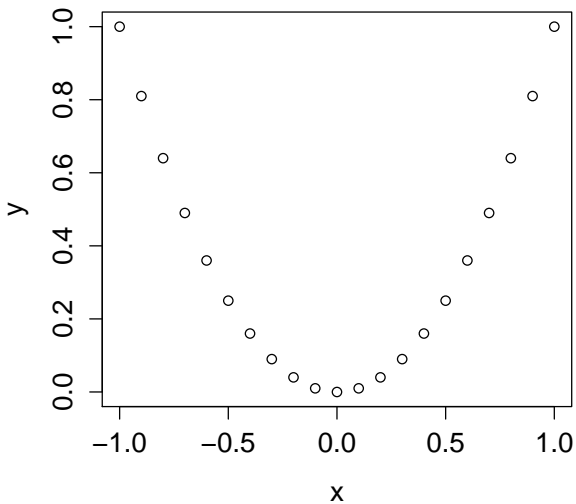


$$b = \frac{s_{xy}}{s_x^2} = 6/2^2 = 1.5$$

## EXTREMELY IMPORTANT

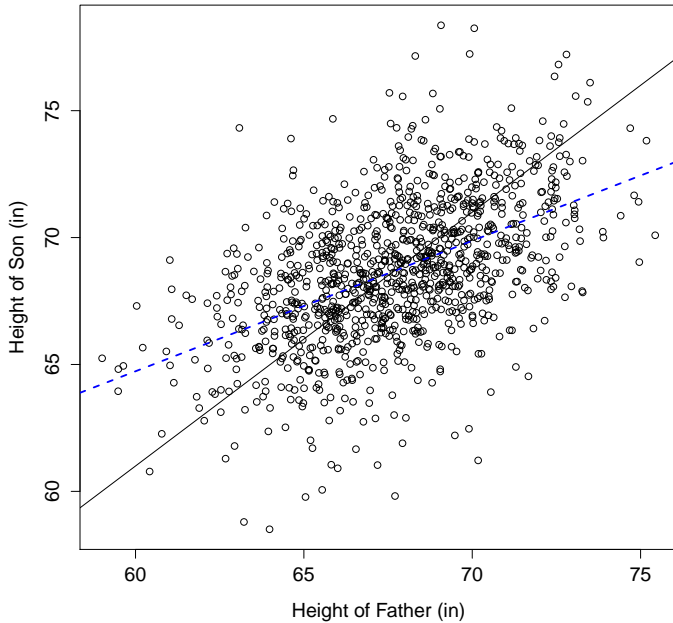
- ▶ Regression, Covariance and Correlation: linear association.
- ▶ Linear association  $\neq$  causation.
- ▶ Linear is not the only kind of association!

**Correlation = 0**



Why is it called “regression?”

## Pearson Dataset

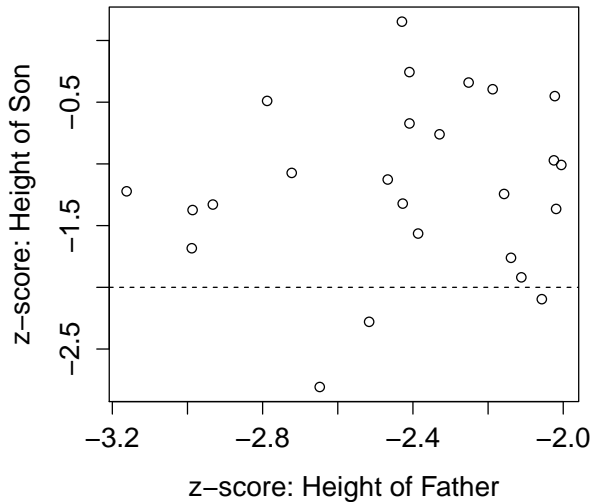




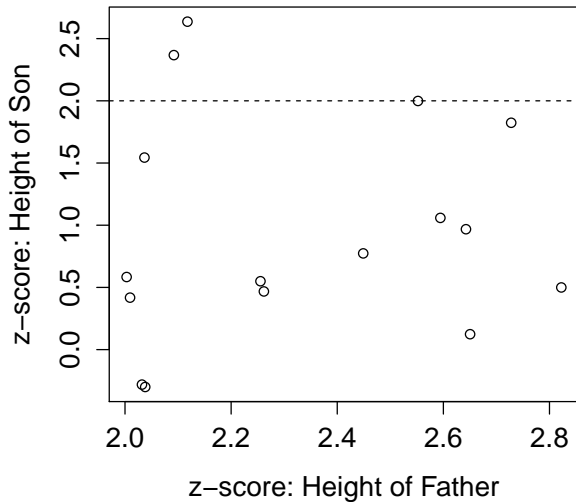
Suppose a father is very short compared to other fathers (very negative z-score). Would you expect his son to be:

- (a) Shorter
- (b) About as short
- (c) Taller

## Very Short Fathers and Their Sons

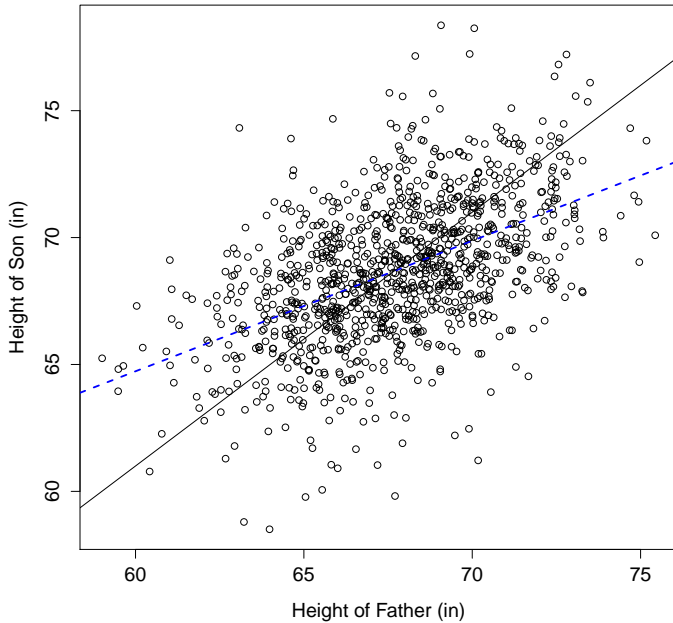


## Very Tall Fathers and Their Sons





## Pearson Dataset



# Regression to the Mean

Skill and Luck / Genes and Random Environmental Factors

Unless  $r_{xy} = 1$ , There Is Regression to the Mean

$$\frac{\hat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}$$

Least-squares Prediction  $\hat{y}$  closer to  $\bar{y}$  than  $x$  is to  $\bar{x}$

You will derive the above formula in this week's homework.

# Regression Fallacy

For More, See the Document Posted on Piazza

## Pre-test

Which students are strongest, which are weakest?

## Intervention

Put the best performing in an enrichment program and the worst performing in a remedial class

## Post-test

The weak students did better than on their first test, but the strong students did *worse*.

## Mistaken Conclusion

Remedial classes are beneficial, enrichment programs are harmful