Economics 103 – Statistics for Economists

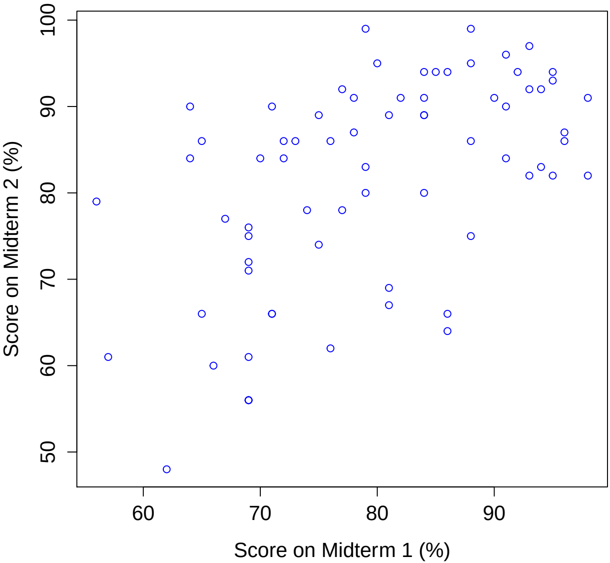
Mallick Hossain

University of Pennsylvania

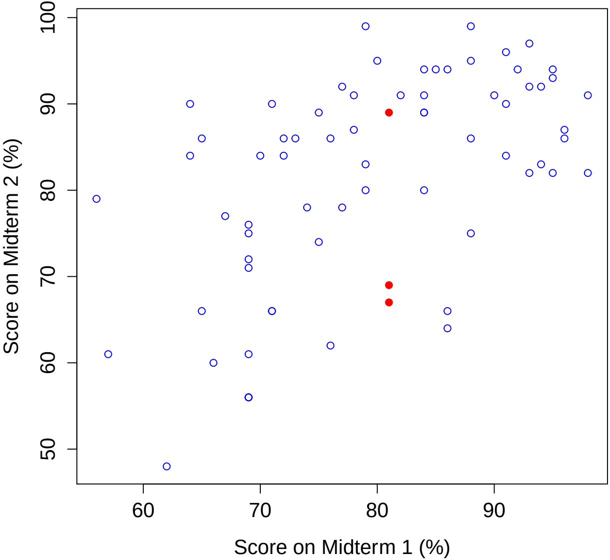
Lecture # 4

Introduction to Regression

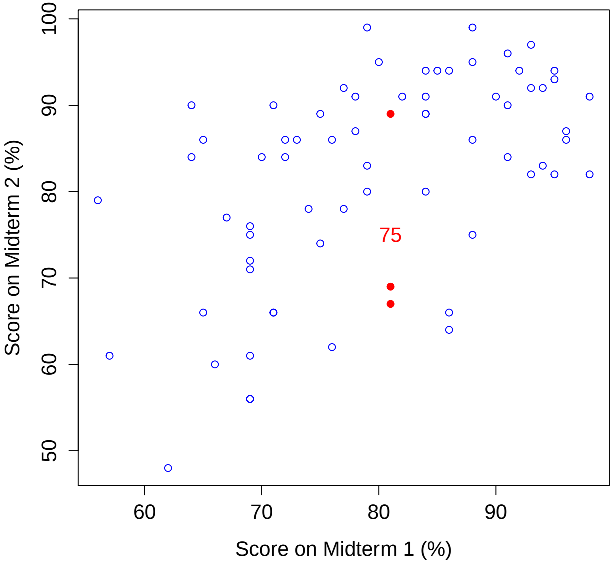
Predict Second Midterm given 81 on First



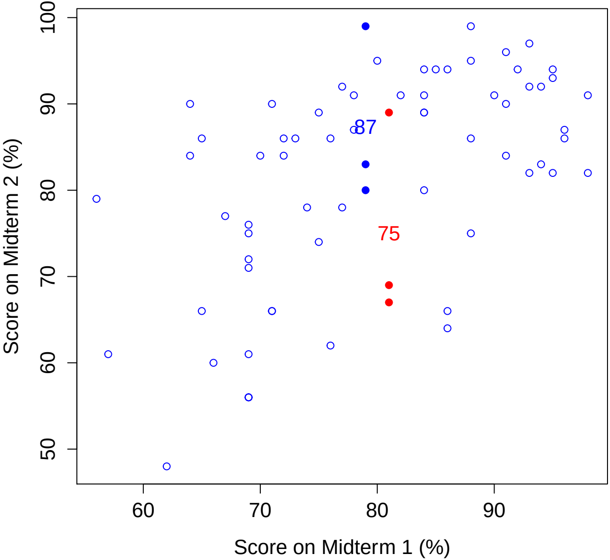
Predict Second Midterm given 81 on First



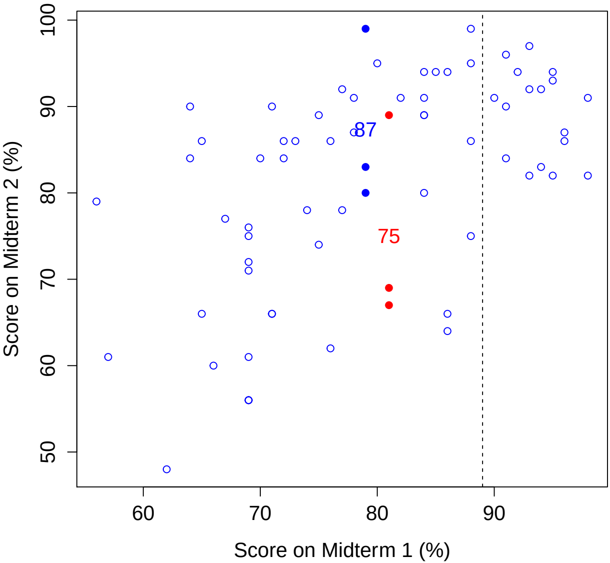
Predict Second Midterm given 81 on First



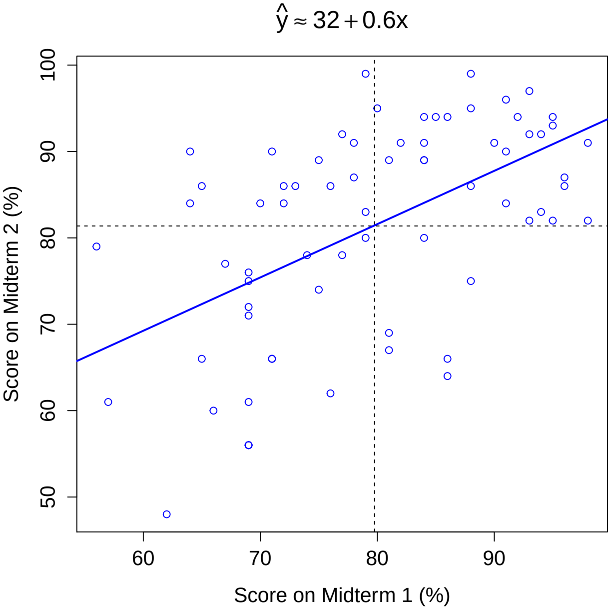
But if they'd only gotten 79 we'd predict higher?!



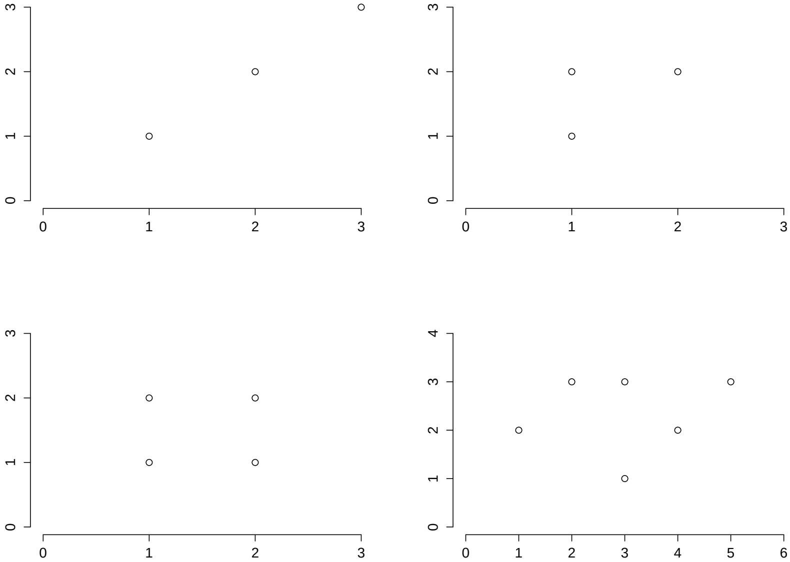
No one who took both exams got 89 on the first!

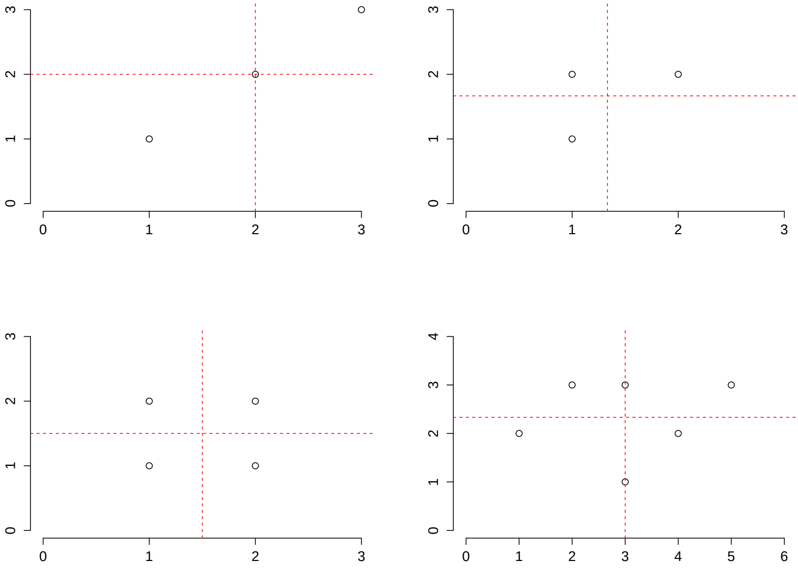


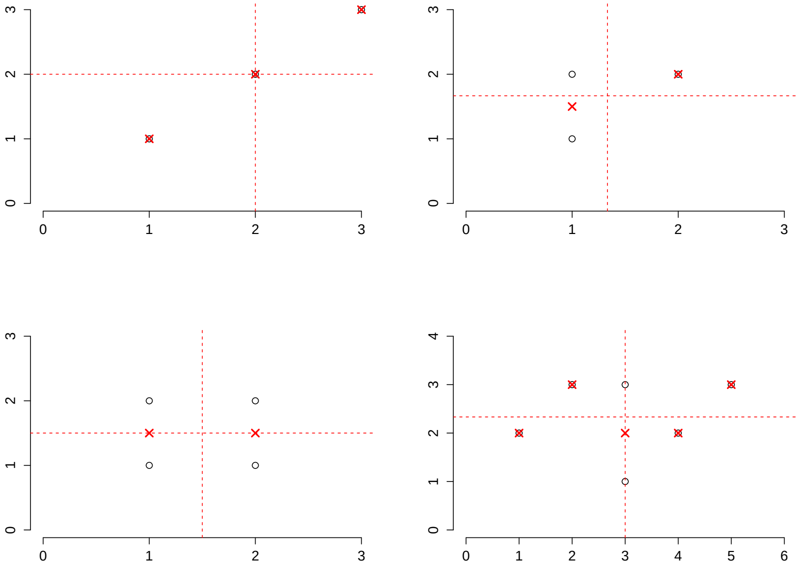
Regression: “Best Fitting” Line Through Cloud of Points



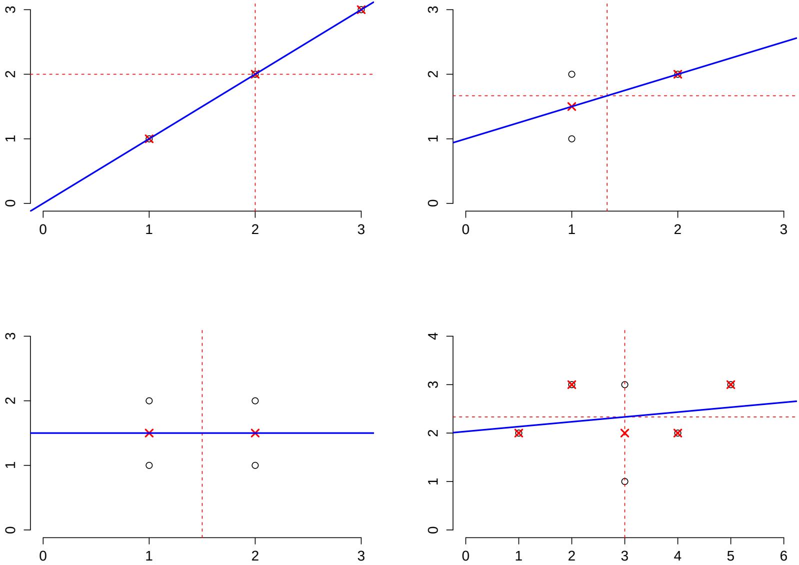
Fitting a Line by Eye


0
1
2
3


0
1
2
3
0
1
2
3
0
1
2
3
4
5
6



3
2
1
0
0
1
2
3


0
1
2
3
0
1
2
3
4
5
6

But How to Do this Formally?

Least Squares Regression – Predict Using a Line

The Prediction

Predict score $\hat{y} = a + bx$ on 2nd midterm if you scored x on 1st

How to choose (a, b)?

Linear regression chooses the slope (b) and intercept (a) that

minimize the sum of squared vertical deviations

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

Why Squared Deviations?

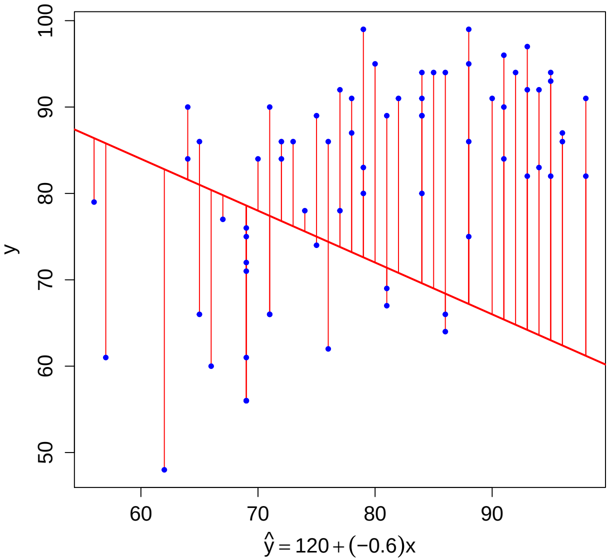
Important Point About Notation

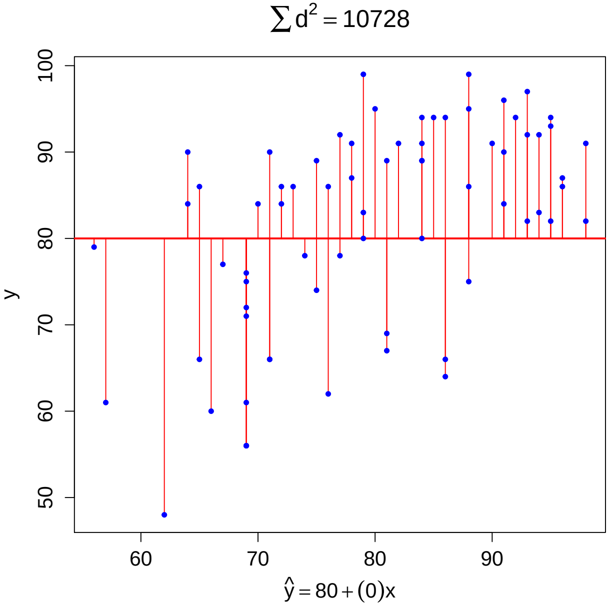
$$\boxed{\underset{a,b}{\text{minimize}} \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2}$$

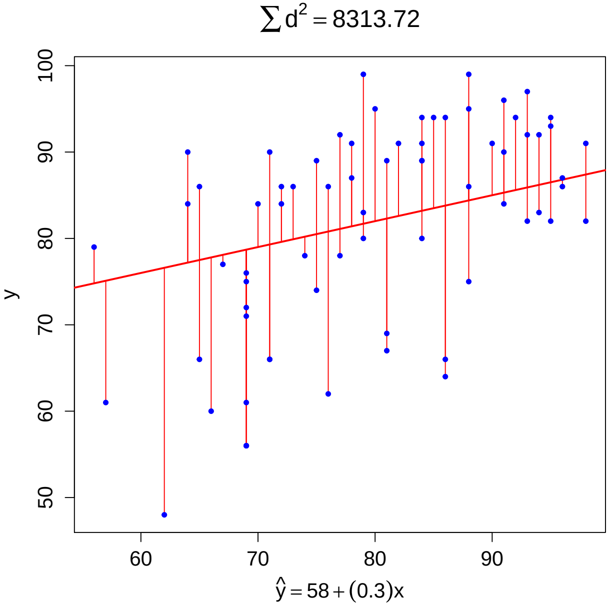
$$\boxed{\hat{y} = a + bx}$$

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

$\sum d^2 = 25596.88$

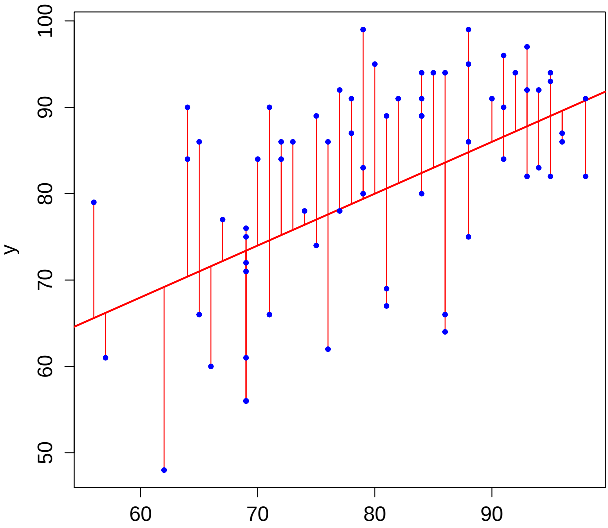



$\sum d^2 = 10728$
y
100
90
80
70
60
50
60
70
80
90
$\hat{y} = 80 + (0)x$



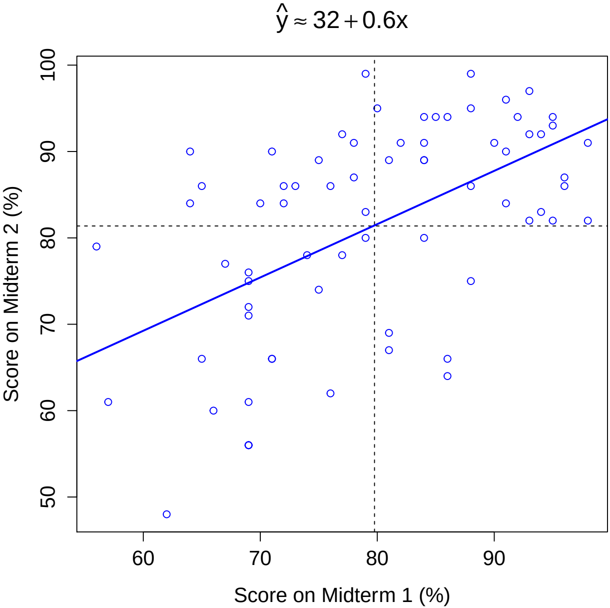
$\sum d^2 = 8313.72$
100
90
80
70
60
50
y
60
70
80
90
$\hat{y} = 58 + (0.3)x$

$$\sum d^2 = 7650.48$$

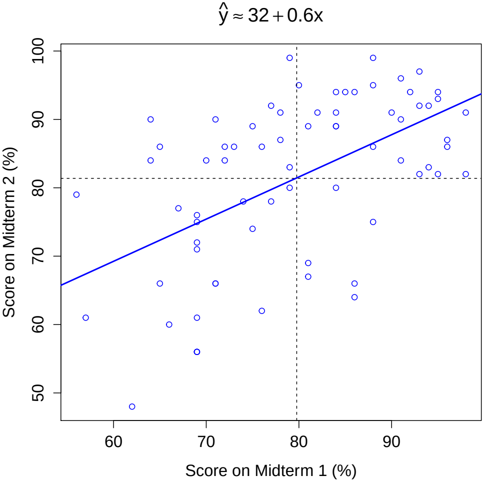


$$\hat{y} = 32 + (0.6)x$$

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*

$$\begin{aligned} -2\sum_{i=1}^{n} (y_i - a - bx_i) &= 0 \\ \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \end{aligned}$$

Objective Function

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

FOC with respect to a

$$\begin{aligned}
-2\sum_{i=1}^{n}(y_i - a - bx_i) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{na}{n} - \frac{b}{n}\sum_{i=1}^{n} x_i &= 0
\end{aligned}$$

Objective Function

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

FOC with respect to a

$$\begin{aligned}
-2\sum_{i=1}^{n}(y_i - a - bx_i) &= 0 \\
\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} x_i &= 0 \\
\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{na}{n} - \frac{b}{n}\sum_{i=1}^{n} x_i &= 0 \\
\bar{y} - a - b\bar{x} &= 0
\end{aligned}$$

Regression Line Goes Through the Means!

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

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

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

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

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

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

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

FOC wrt b

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

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

FOC wrt b

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

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

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

FOC wrt b

$$\begin{aligned}-2\sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) &= 0 \\ \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - b\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 &= 0\end{aligned}$$

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

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

FOC wrt b

$$\begin{aligned}-2\sum_{i=1}^{n}\left[(y_i - \bar{y}) - b\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) &= 0 \\ \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - b\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 &= 0\end{aligned}$$

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

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

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

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

Comparing Regression, Correlation and Covariance

Units

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

Symmetry

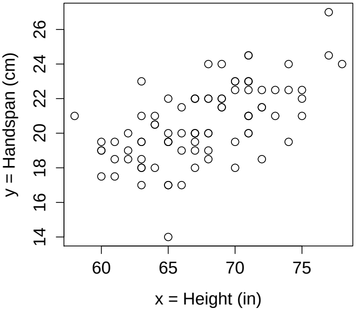
Correlation and covariance are symmetric, regression isn't.
(Switching x and y axes changes the slope and intercept.)

On the Homework

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

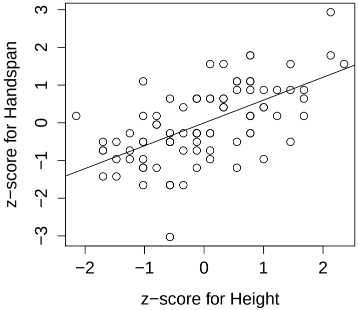
$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

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)?



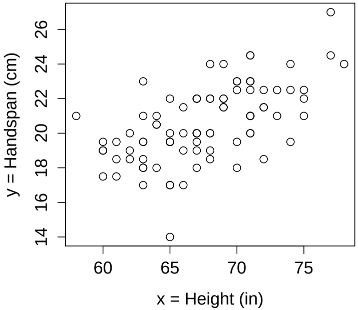
$$r = \frac{s_{xy}}{s_x s_y} = \frac{6}{5 \times 2} = 0.6$$

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

What is the value of b for the regression:

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

where x is height and y is handspan?

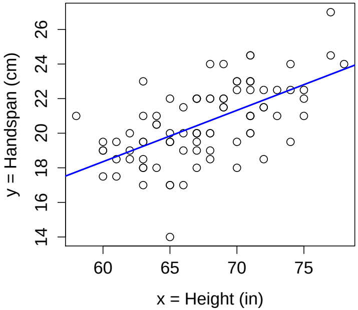


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of b for the regression:

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

where x is height and y is handspan?



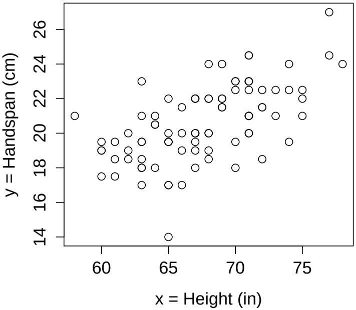
$$b = \frac{s_{xy}}{s_x^2} = \frac{6}{5^2} = 6/25 = 0.24$$

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

What is the value of a for the regression:

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

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

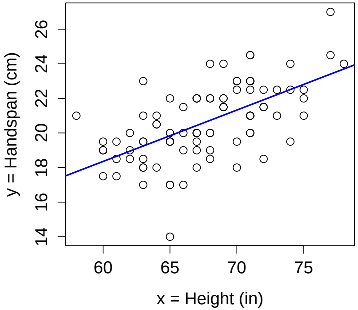


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of a for the regression:

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

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



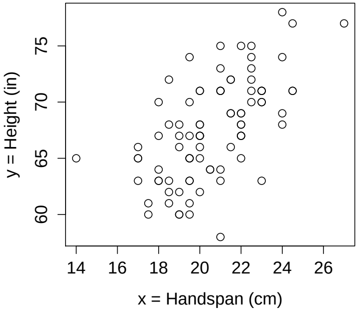
$$a = \bar{y} - b\bar{x} = 21 - 0.24 \times 68 = 4.68$$

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

What is the value of b for the regression:

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

where x is handspan and y is height?

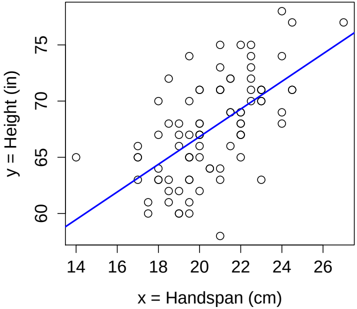


$$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$$

What is the value of b for the regression:

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

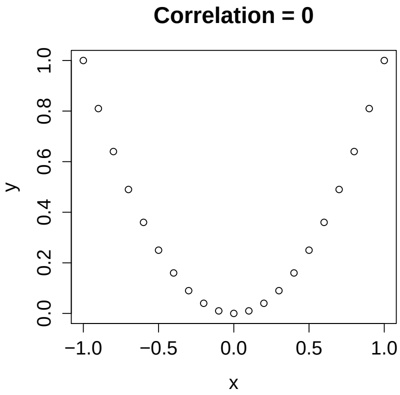
where x is handspan and y is height?



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

EXTREMELY IMPORTANT

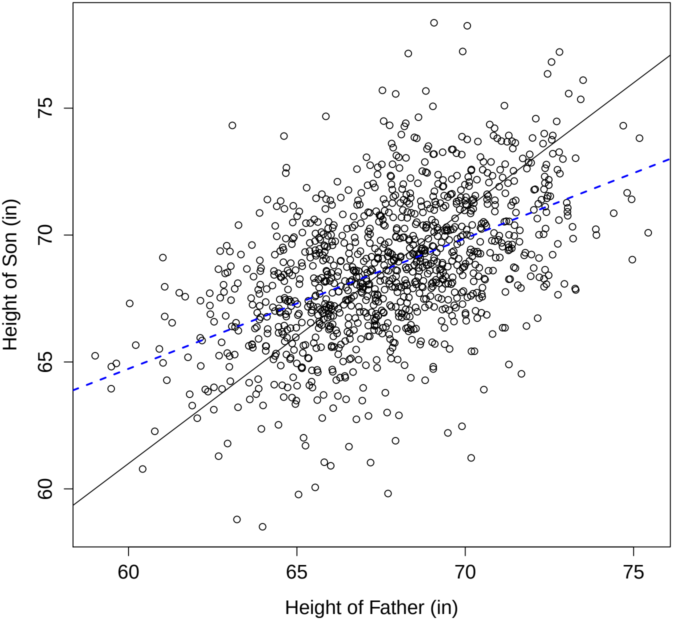
- Regression, Covariance and Correlation: linear association.
- Linear association $\neq$ causation.
- Linear is not the only kind of association!



Regression to the Mean and the Regression Fallacy

Please read Chapter 17 of "Thinking Fast and Slow" by Daniel Kahnemann which I have posted on Piazza. This reading is fair game on an exam or quiz.

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.

Next Up: Basic Probability

Please do the following before our next class:

1. Complete the “Odd Questions” quiz posted on Piazza `OddQuestions.pdf` – we’ll be discussing these in class.
2. If you’re rusty on permutations, combinations, etc. from High School math, read this review `http://ditraglia.com/Econ103Public/ClassicalProbability.pdf`