

Economics 103 – Statistics for Economists

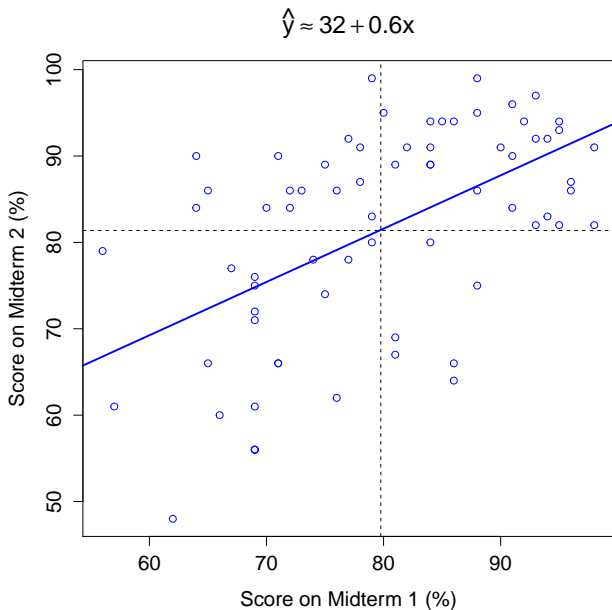
Minsu Chang

University of Pennsylvania

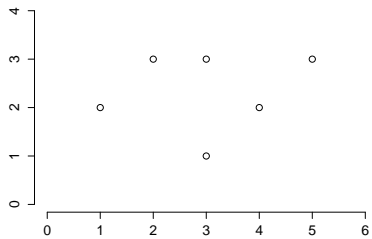
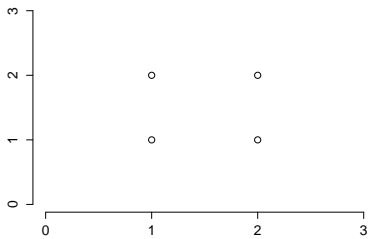
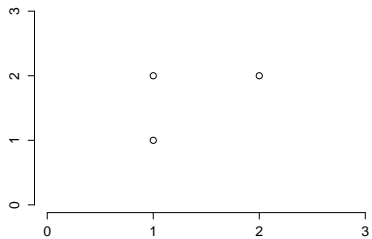
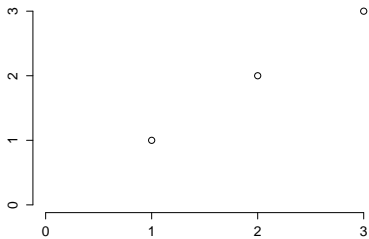
Lecture # 4

Introduction to Regression

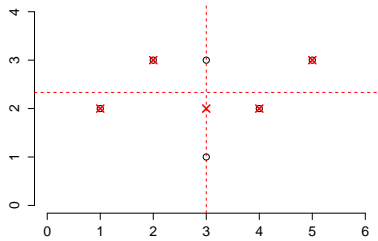
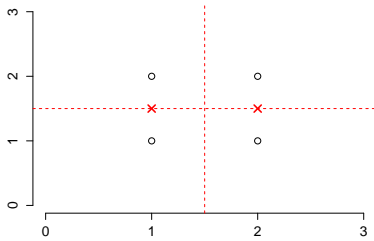
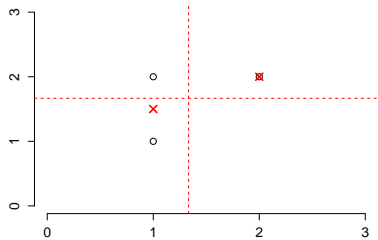
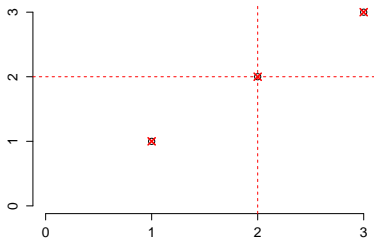
Regression: “Best Fitting” Line Through Cloud of Points

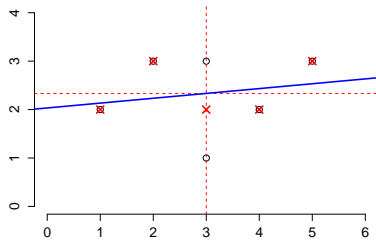
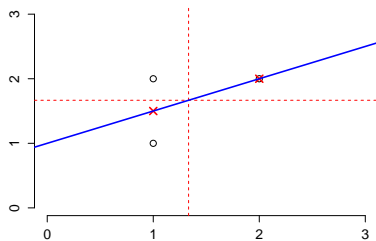


Fitting a Line by Eye









But How to Do this Formally?

Least Squares Regression – Predict Using a Line

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?

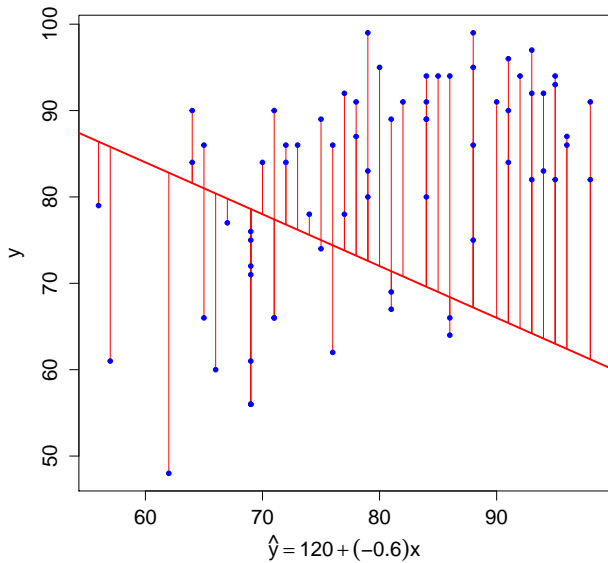
Important Point About Notation

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

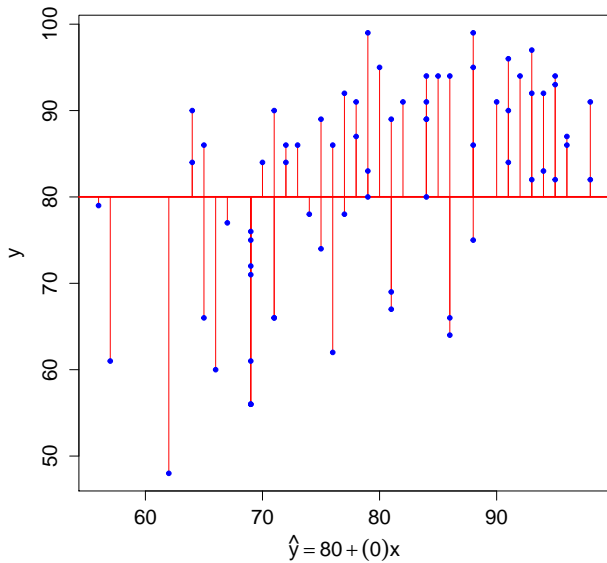
$$\hat{y} = a + bx$$

- ▶ $(x_i, y_i)_{i=1}^n$ 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!

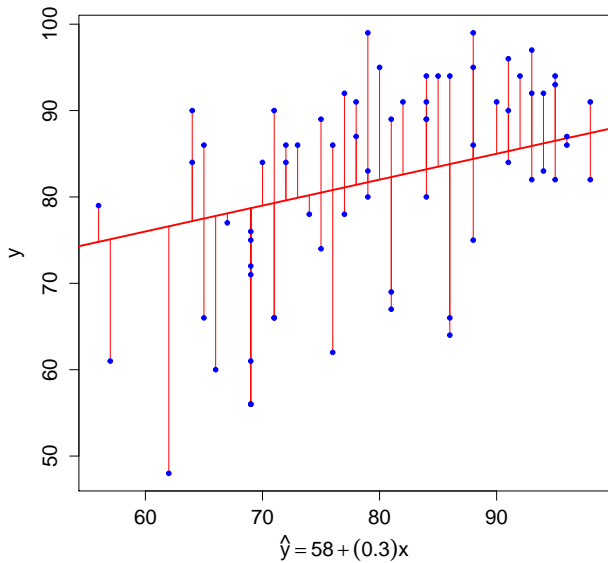
$$\sum d^2 = 25596.88$$



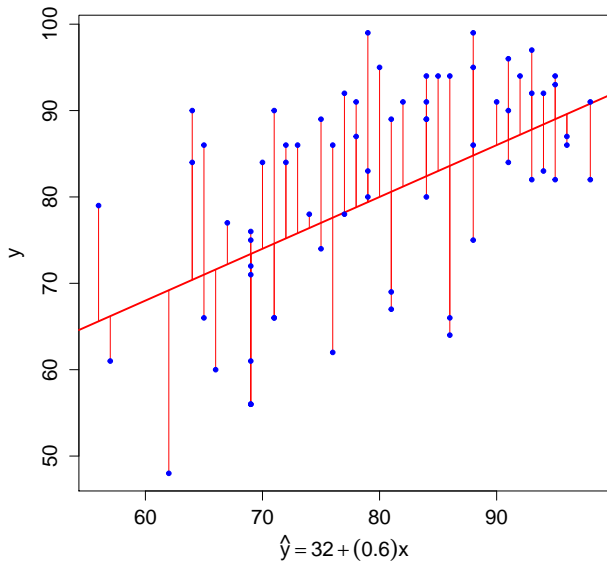
$$\sum d^2 = 10728$$



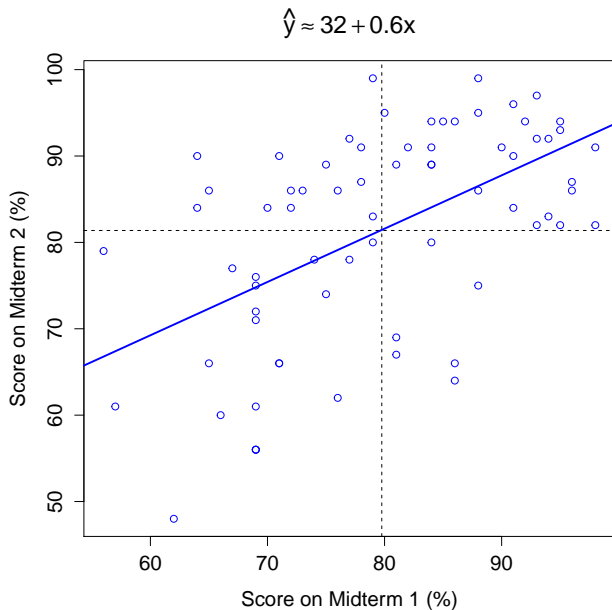
$$\sum d^2 = 8313.72$$



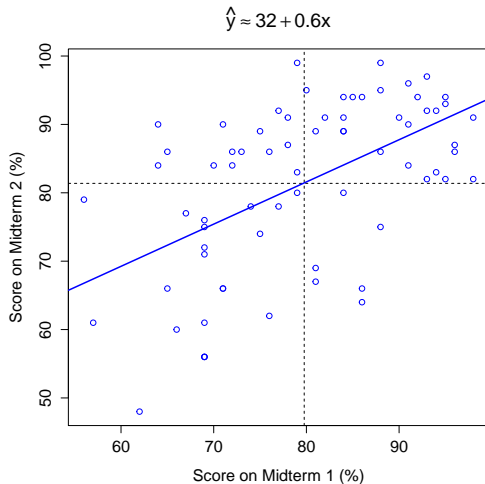
$$\sum d^2 = 7650.48$$



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

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

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

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

FOC wrt b

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

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

FOC wrt b

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

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

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

FOC wrt b

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

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

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

FOC wrt 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

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

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

Comparing Regression, Correlation and Covariance

Units

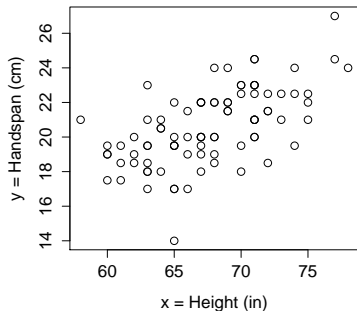
Correlation is unitless, covariance and regression coefficients (a, b) are not. (What are the units of these?)

Symmetry

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

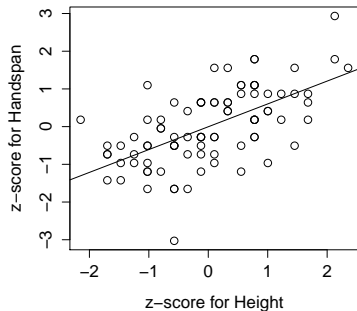
$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

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)?



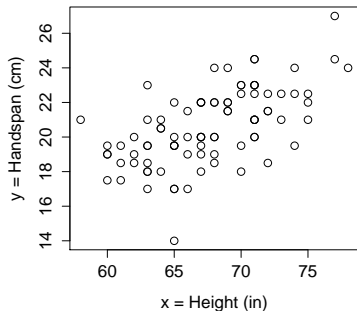
$$r = \frac{s_{xy}}{s_x s_y} = \frac{6}{5 \times 2} = 0.6$$

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

What is the value of b for the regression:

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

where x is height and y is handspan?

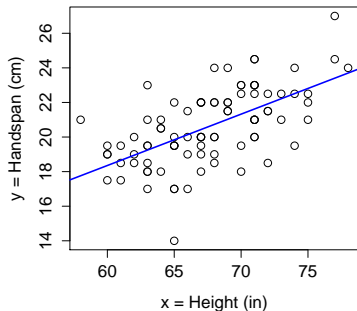


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of b for the regression:

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

where x is height and y is handspan?



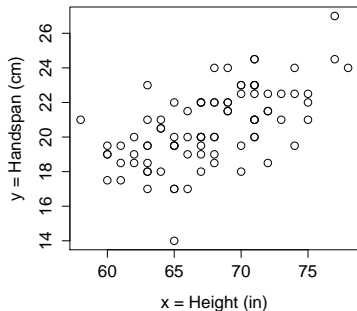
$$b = \frac{s_{xy}}{s_x^2} = \frac{6}{5^2} = 6/25 = 0.24$$

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

What is the value of a for the regression:

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

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

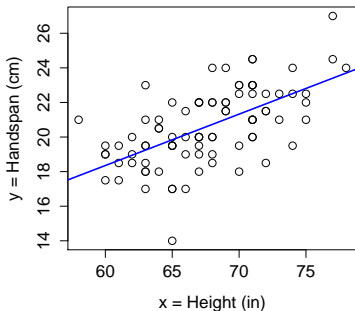


$$s_{xy} = 6, \quad s_x = 5, \quad s_y = 2, \quad \bar{x} = 68, \quad \bar{y} = 21$$

What is the value of a for the regression:

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

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



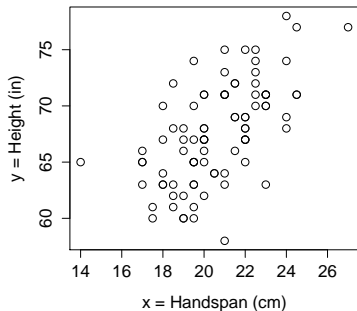
$$a = \bar{y} - b\bar{x} = 21 - 0.24 \times 68 = 4.68$$

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

What is the value of b for the regression:

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

where x is handspan and y is height?

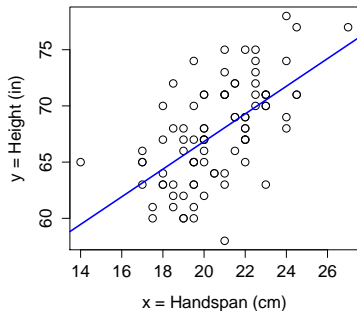


$$s_{xy} = 6, \quad s_y = 5, \quad s_x = 2, \quad \bar{y} = 68, \quad \bar{x} = 21$$

What is the value of b for the regression:

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

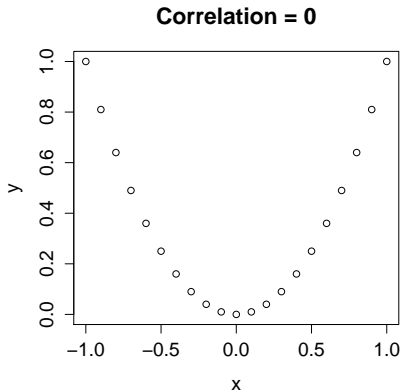
where x is handspan and y is height?



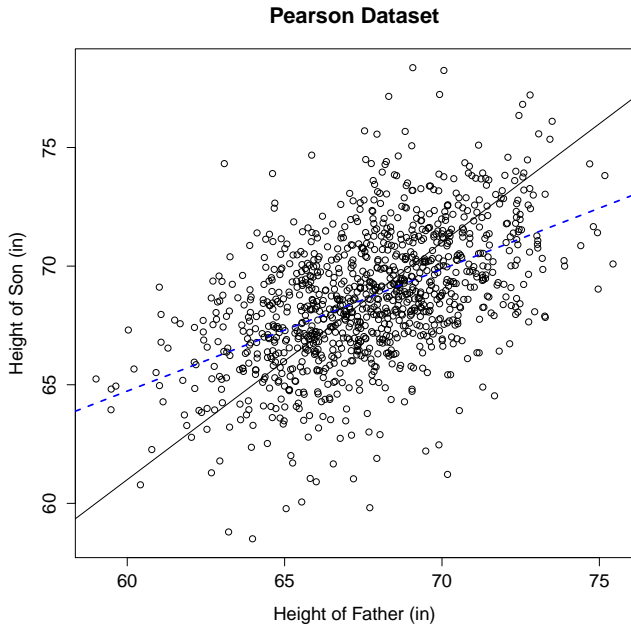
$$b = \frac{s_{xy}}{s_x^2} = 6/2^2 = 1.5$$

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



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

If you're rusty on permutations, combinations, etc. from High School math, read [ClassicalProbability.pdf](#) on the webpage.