DATA 8

Fall 2016

Lecture 29, November 2

The Regression Line

Slides created by Ani Adhikari and John DeNero

Announcements

- Project 2 deadline is Tuesday 11/8 at 7 p.m.
- Homework due this week as usual.
- There will be a small lab in lab, in addition to project time

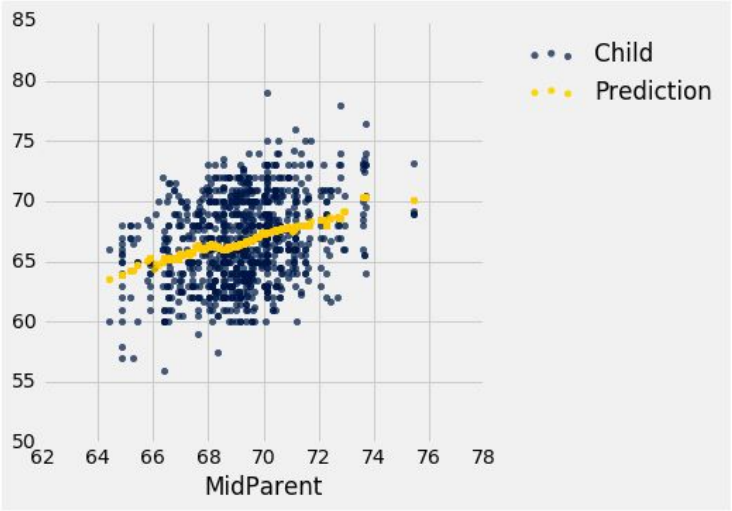
- Prob 140 (Statistics 140) is now open for enrollment

The Correlation Coefficient *r*

- Measures linear association
- Based on standard units; pure number, not affected by changing units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No *linear* association; *uncorrelated*
- Not affected by switching axes

Prediction

- Guess outcomes in the future, based on available data
- One simple goal:
  - Predict the value of one variable based on another



(Demo)

Regression to the Mean

- **estimate of $y = r \cdot x$**, when both variables are measured in standard units
- If $r = 0.6$, and the given x is 2 standard units, then:
  - The given x is 2 SDs above average
  - The prediction for y is 1.2 SDs above average
- On average (though not for each individual), regression predicts y to be closer to the mean than x is

Regression Estimate, Method I

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

If the scatter of midterm & final scores for students looks like a typical oval with correlation 0.75, then...

What do you expect the average final score would be for students who scored 90 on the midterm?

2 standard units on midterm,
so estimate 0.75 * 2 = 1.5 standard units on final.
So estimated final score = 1.5 * 12 + 50 = 68 points

Regression Equation

In original units, the regression line has this equation:

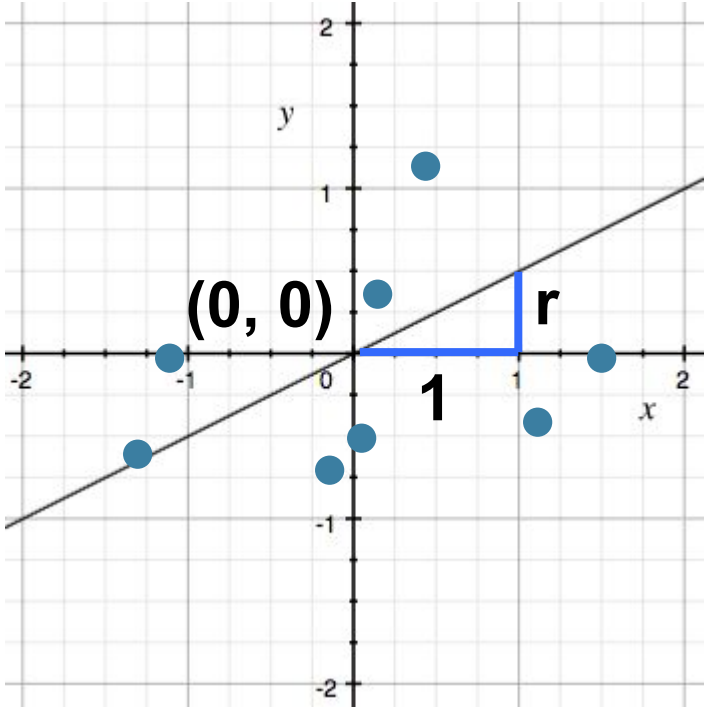
$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimate of *y* in standard units

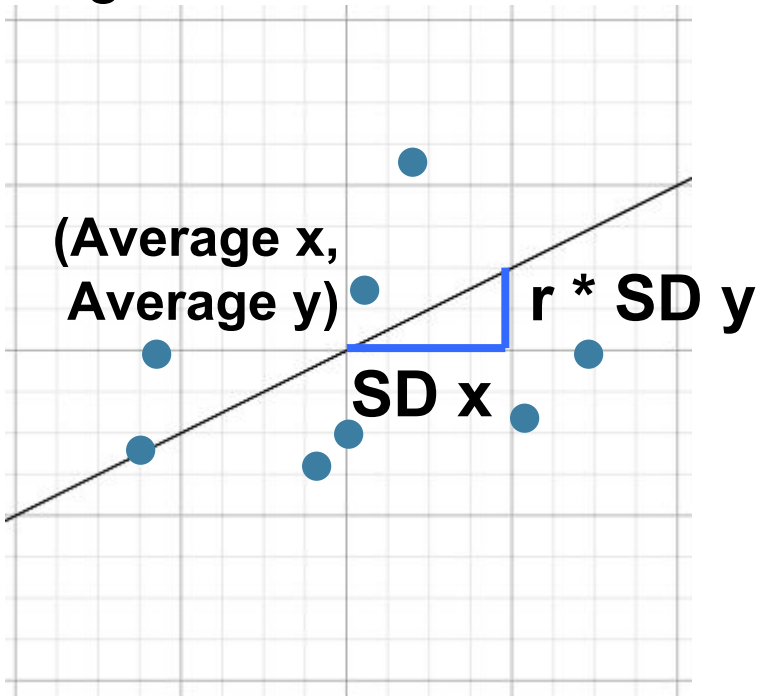
x in standard units

Regression Line

Standard Units



Original Units



Slope and Intercept

estimate of y = slope * x + intercept

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

(Demo)

Regression Estimate, Method II

The equation of a regression line for estimating child's height based on midparent height is

estimated child's height = 0.64·midparent height + 22.64

Estimate the height of someone whose midparent height is 69 inches.

0.64*69 + 22.64 = 66.8 inches