DSFA

Spring 2019

Lecture 25

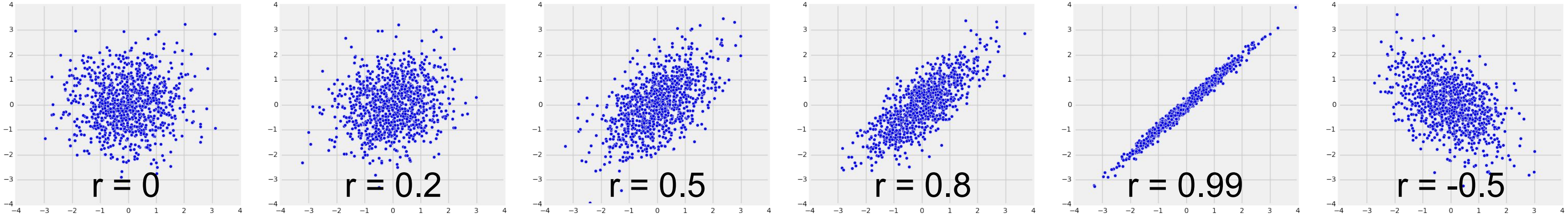
Linear Regression

Announcements

- **Final Exam**
  2pm Monday, May 13
  B14 Hollister Hall

The Correlation Coefficient *r*

- Measures linear association
- Based on standard 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*



Definition of *r*

Correlation Coefficient (*r*) =

average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

(Demo)

Properties of *r*

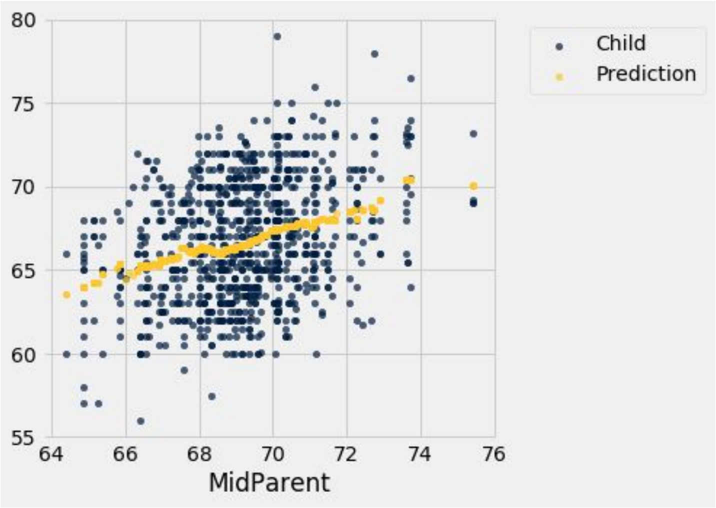
- r is a pure number, with no units
- r is not affected by changing units of measurement
- r is not affected by switching the horizontal and vertical axes (symmetric in x and y)

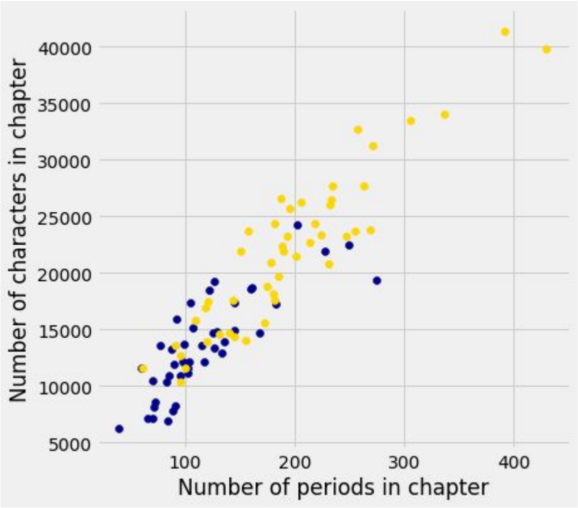
(Demo)

Prediction

Prediction

If we have a line describing the relation between two variables, we can make predictions



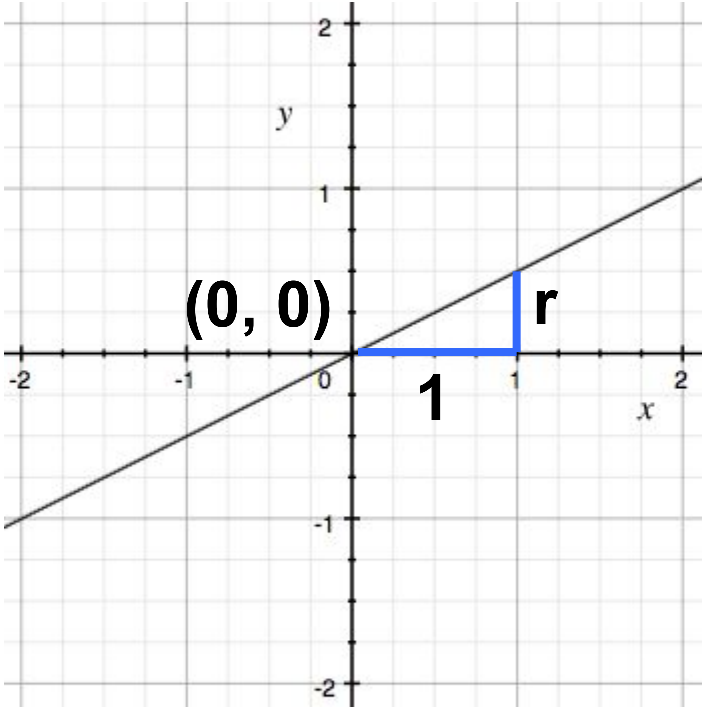


Prediction

- **Problem:** given a known *x* value, predict *y*, where both are in standard units
- **Solution:**
  - Compute *r*
  - Predict that $y = r * x$
- Why is that a line?

Algebra review:

Equation of a Line



$$y = r * x$$

In general:

$$y = a * x + b$$

(a is slope, b is intercept)

Prediction

- **Problem:** given a known *x* value, predict *y*, where both are in standard units
- **Solution:**
  - Compute *r*
  - Predict that $y = r * x$
- Why is that a line?
- Why use *that* line?

(Demo)

Prediction

- **Problem:** given a known *x* value, predict *y*, where both are in standard units
- **Solution:**
  - Compute r
  - Predict that $y = r * x$
- Why is that a line?
- Why use *that* line?
  - It is a version of the graph of averages, smoothed to a line

(Demo)

Prediction

- **Predict** $y = r * x$ (in standard units)

- Example:
  - x = 2 (in standard units)
  - r = .75
  - What is the prediction for y (in standard units)?
    - A. 0.0
    - B. 0.75
    - C. 1.5
    - D. 2.0

Prediction

- **Predict** $y = r * x$ (in standard units)

- Example:
  - A course has a typical prelim (mean=70, std=10), and a hard final (mean=50, std=12)
  - The scores on the exams look linearly related when visualized, with $r = .75$
  - **Predict** a student's final exam score, given that their prelim score was 90 *(go ahead and work on that)*

Prediction

- Prelim: mean=70, std=10
  - x = 90 = 70 + 2*10 in original units = 2 standard units
- Prediction:
  - y = r * x = .75 * 2 = 1.5 standard units
- Final: mean=50, std=12
  - y = 50 + 1.5 * 12 = **68** in original units

Prediction

- **Predict** $y = r * x$ (in standard units)
- If r = .75 and x is 2 std above mean,
  then prediction for y is 1.5 std above mean
- So y predicted to be closer to its mean than x is

- "Regression to the mean"
  - Children with exceptionally tall parents tend not to be as tall
  - Galton called it "regression to mediocrity"

(Demo)

Linear Regression

(Demo)

Equation for regression line

(y in su) = r * (x in su)

Equation for regression line

$$(y \text{ in su}) = r * \frac{x - \text{mean}(\text{all } x)}{\text{std}(\text{all } x)}$$

Equation for regression line

$$\frac{y - \text{mean(all y)}}{\text{std(all y)}} = r * \frac{x - \text{mean(all x)}}{\text{std(all x)}}$$

Equation for regression line

$$\frac{y - \text{mean(all } y)}{\text{std(all } y)} = r * \frac{x - \text{mean(all } x)}{\text{std(all } x)}$$

Do some algebra to put that in the form y = slope * x + intercept...

Slope and Intercept

$$y = \text{slope} * x + \text{intercept}$$

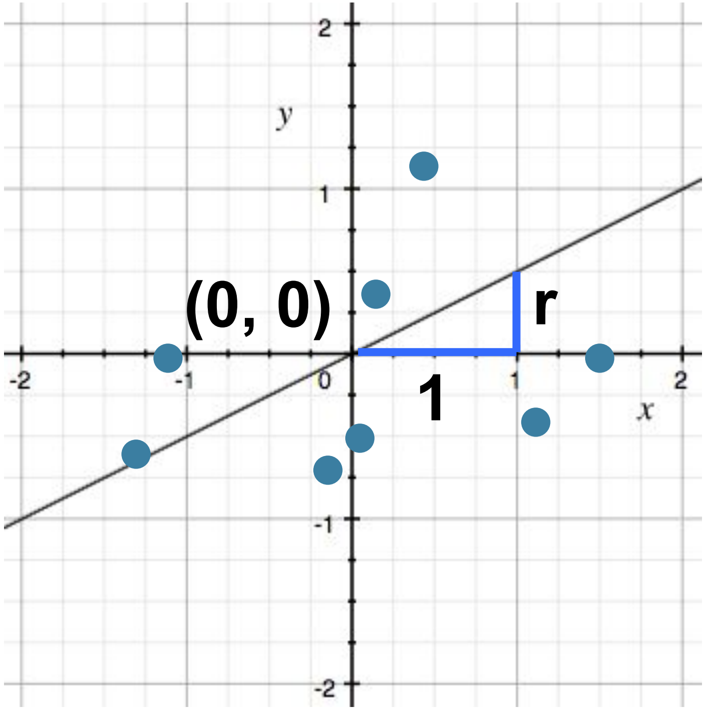
$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

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

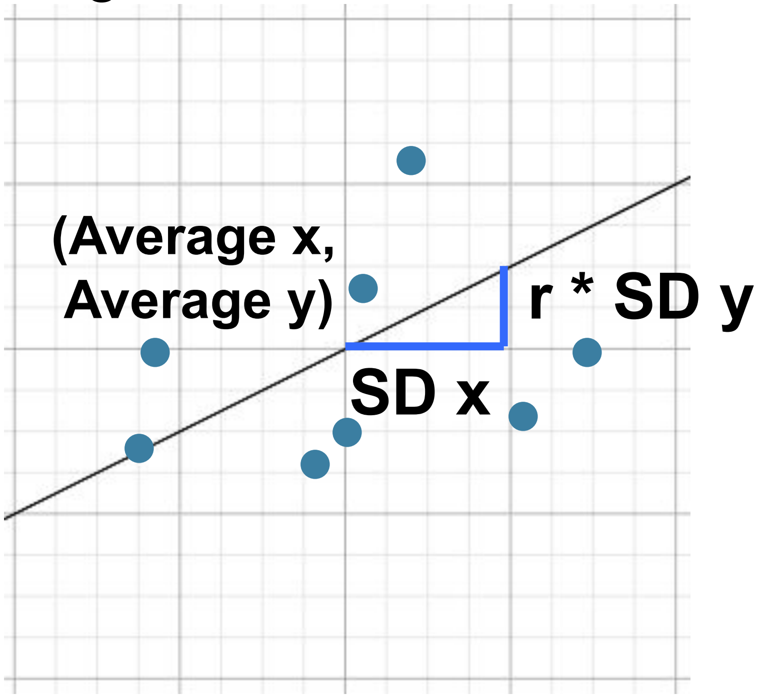
(Demo)

Regression Line

Standard Units



Original Units



Abuses of *r*

- Summarizing non-linear data with *r*
- Eliminating outliers to “improve” *r*
- Drawing conclusions about individuals based on data about groups (*ecological* correlations)
- Jumping to conclusions about causality

Correlation is not causation

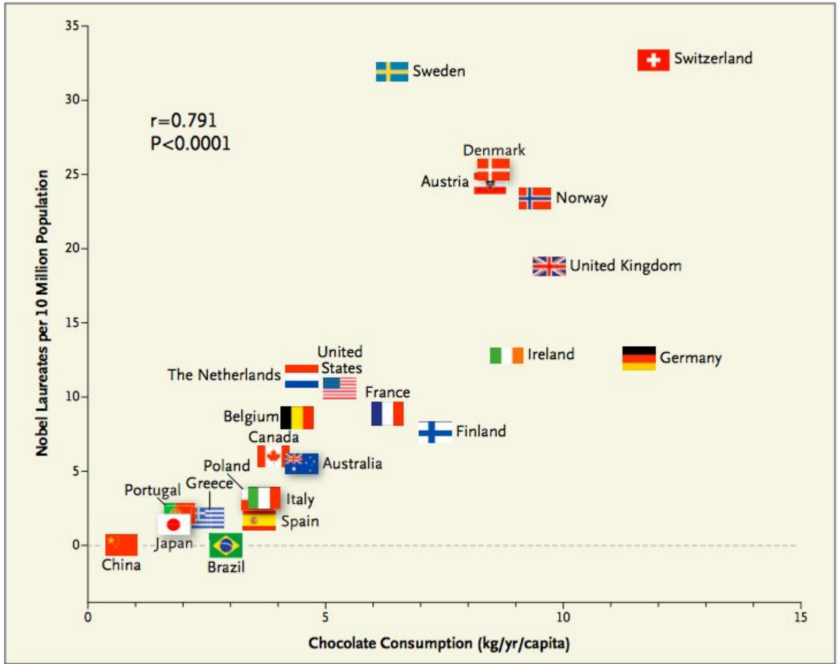


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Quantifying Error

Error in Prediction

- How good is the regression line at making predictions?
  - Hard to say for unknown data
  - But easy for data we already have

- **error = actual value − prediction**

(Demo)

Error in Prediction

- How good is the regression line at making predictions?
  - Hard to say for unknown data
  - But easy for data we already have

- **error = actual value − prediction**
- RMSE = root mean square error
  4 3 2 1
- RMSE = root mean square of deviation from prediction
  5 4 3 2 1

(Demo)

RMSE

RMSE = root mean square error

RMSE = std(y) * sqrt(1 - r^2)

- If r = 1, what is RMSE? 0
- If r = 0, what is RMSE? std(y)

Compare regression line to other lines using RMSE...

(Demo)

Line with smallest RMSE?

- SciPy function `minimize(f)` returns the value `x` that produces the minimum output `f(x)` from `f`
- Also works for functions that make multiple arguments
- How to use to find best line:
  - Write function `rmse(a, b)` that returns the RMSE for line with slope `a` and intercept `b`
  - Call `minimize(rmse)` and get output array [a_0, b_0]
  - a_0 is slope and b_0 intercept of line that minimizes RMSE

(Demo)

Regression line

- Regression line has the minimum RMSE of all lines
- Names:
  - Regression line
  - Least squares line
  - “Best fit” line