DSC 10, Spring 2018
Lecture 22

Linear Regression

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Last Time

Review discussion question

Given a table with 3 columns:

Week

Beer: number of bottles of beer consumed in San Diego that week

Weddings: the number of weddings in San Diego that week

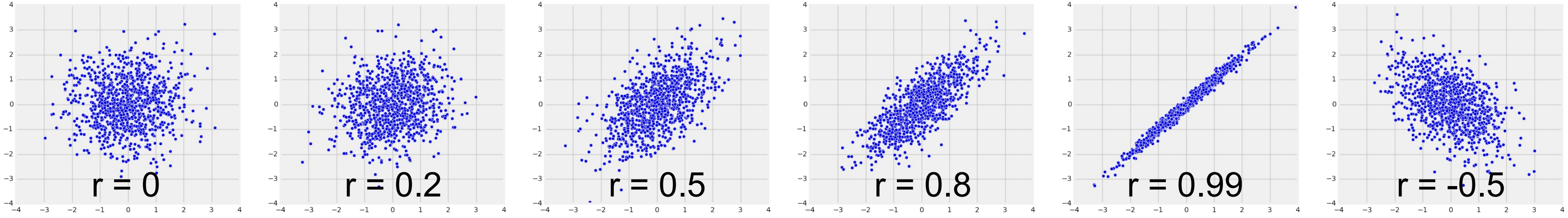
Let r be the *correlation* between beer and weddings.

Which statement is True?

A. If r = 1.5, this means that people consume an average of one and a half beers per wedding they attend.
B. If r is between -0.05 and 0.05, there is little association between beer consumption and weddings.
C. If r = 1, then an increase in weddings causes an increase in beer consumption.
D. More than one of the above.

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*

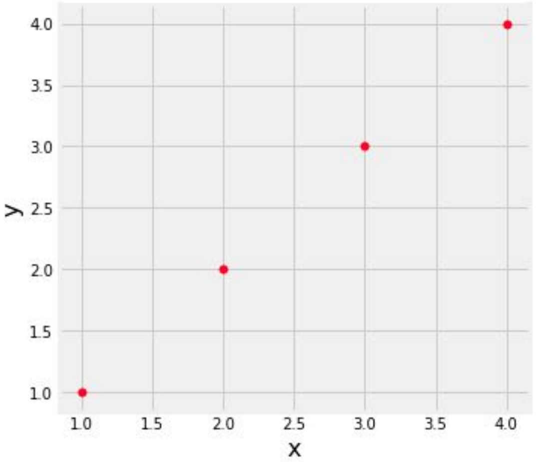


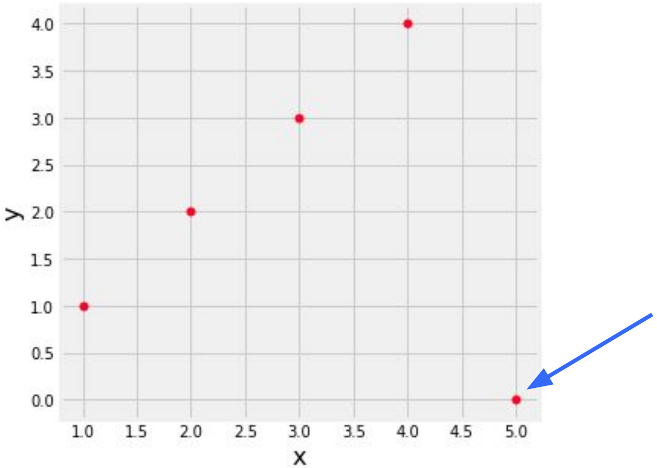
Interpreting *r*

Watch out for:

- Jumping to conclusions about causality
- Non-linearity
- Outliers
  - r is affected by outliers
  - r is based on the mean and sd
  - outliers change the mean and sd

Discussion Question





What are the correlations for these scatter plots? Note one outlier on the right plot.

A. $r = 1$, $r \sim 0.9$
B. $r = 1$, $r \sim 0.5$
C. $r = 1$, $r = 0$
D. $r = 0$, $r \sim 0.5$
E. None of the above

(Demo)

Linear Regression

Graph of Averages

A visualization of x and y pairs

- Group each x value with other nearby x values
- Average the corresponding y values for each group
- For each x value, produce one predicted y value

If the association between x and y is linear, then points on the graph of averages tend to fall on the regression line

Regression to the Mean

Regression Line

$$\boxed{y_{(\text{su})} = \underbrace{r}_{\text{Correlation}} \times x_{(\text{su})}}$$

- 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

Discussion Question

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

If the scatter diagram of midterm & final scores for students has a typical oval shape with **correlation 0.75**, then what is the average final exam score for students who scored **90 on the midterm**?

A. 76

B. 90

C. 68

D. 82

E. 67.5

Discussion Question: Solution

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

If the scatter diagram of midterm & final scores for students has a typical oval shape with **correlation 0.75**, then what is the average final exam score for students who scored **90 on the midterm**?

1. (90 - 70)/10 = 2 standard units on midterm,
2. estimate 0.75 * 2 = 1.5 standard units on final
3. estimated final score = 1.5 * 12 + 50 = 68 points

(Demo)

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

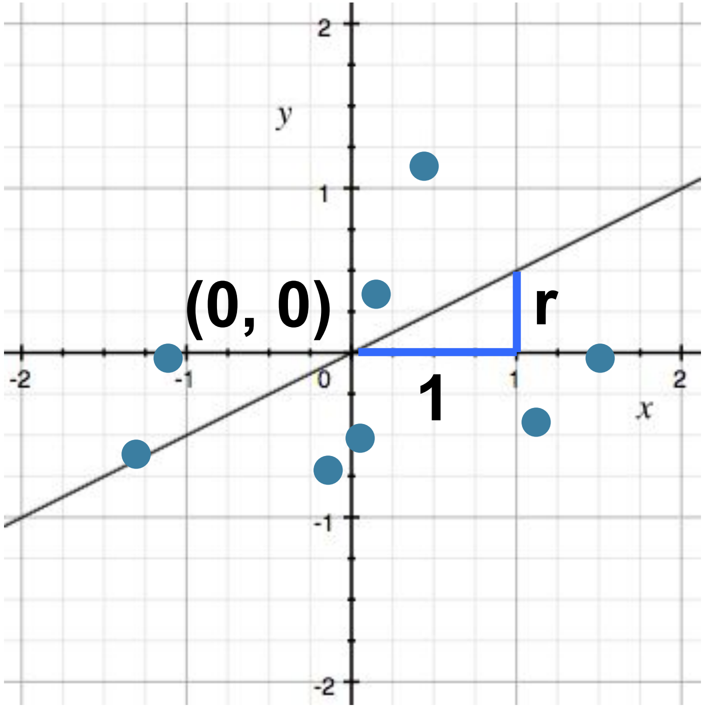
$$\underbrace{\boxed{\frac{\text{estimate of } y \; - \; \text{average of } y}{\text{SD of } y}}}_{\text{y in standard units}} = r \times \underbrace{\boxed{\frac{\text{the given } x \; - \; \text{average of } x}{\text{SD of } x}}}_{\text{x in standard units}}$$

Lines can be expressed by *slope* & *intercept*

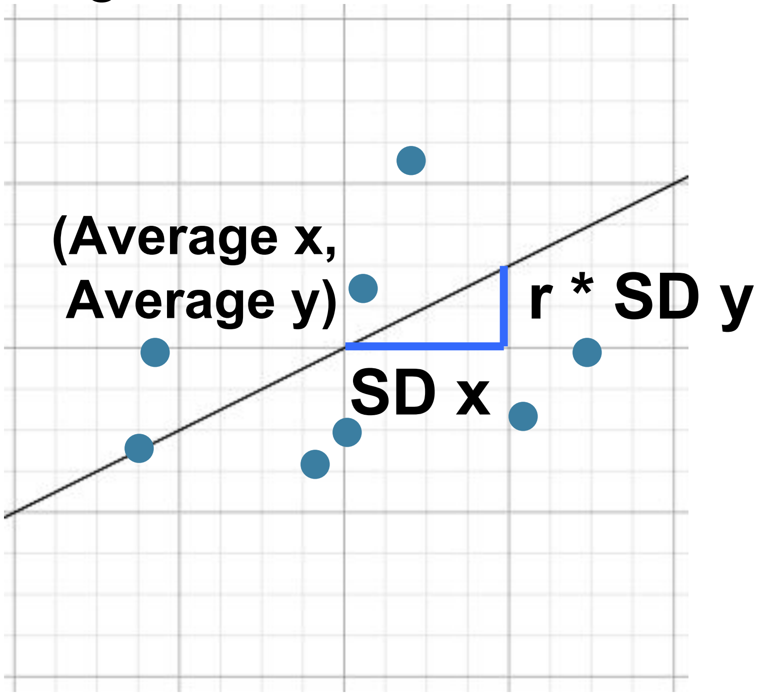
$$y = \text{slope} \times x + \text{intercept}$$

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)

Least Squares

Error in Estimation

- **error = actual value - estimate**

  = actual value - predicted value
- Typically, some errors are positive and some negative
  - What does a positive error mean? negative?
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square error** (rmse)

(Demo)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

Numerical Optimization

- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- Idea: Given a function that returns a real number,
  - Search among all possible inputs to the function
  - Find the input to the function that results in the function returning the smallest possible real number
  - Approximate because we cannot search *all* possible inputs

Numerical Optimization of MSE

If the function `mse(a, b)` returns the mse of estimation using the line "estimate = ax + b",

- then `minimize(mse)` returns array [a_0, b_0]
- a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

(Demo)

Discussion question

```python
def my_func(c):
    if c < -2:
        return 4
    elif c > 2:
        return 4
    else:
        return abs(c)+2
```

Pick the option that best completes the sentence:

“The expression **minimize(my_func)** evaluates to…”

A: -3
B: 0
C: 1
D: 2
E: 4