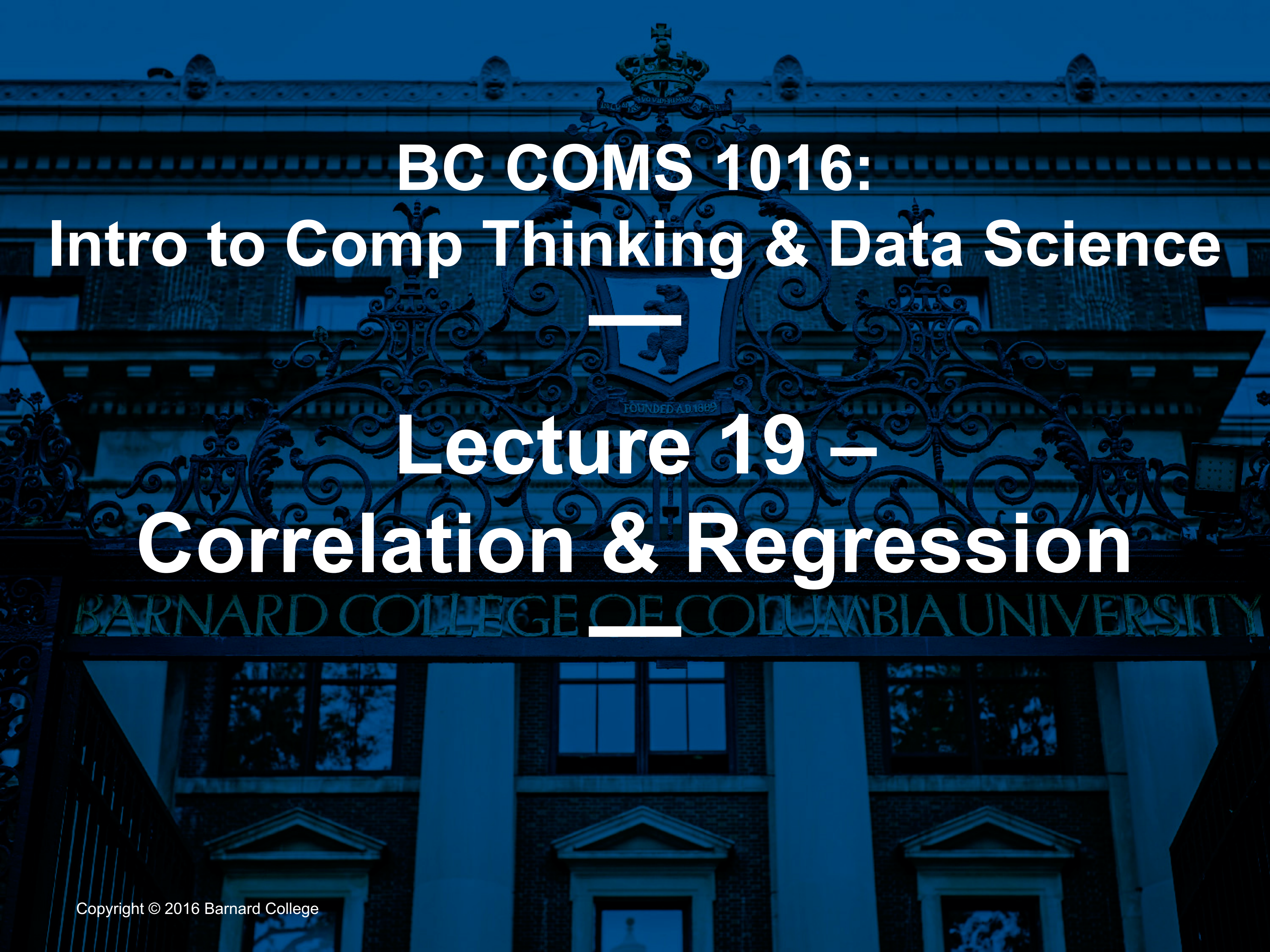


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 19 – Correlation & Regression

—

Announcements

- Lab07 – Normal Distribution and Variance of Sample Means (short)
  - Due Wednesday 11/23
- Homework 7 - Confidence Intervals, Resampling, the Bootstrap, and the Central Limit Theorem
  - Due Thursday 11/24
  - Not the shortest
- Homeworks:
  - Run all cells before submitting
- Dropping 2 homeworks and labs

Correlation

Prediction

- To predict the value of a variable:
  - Identify (measurable) attributes that are associated with that variable
  - Describe the relation between the attributes and the variable you want to predict
  - Then, use the relation to predict the value of a variable

Visualizing Two Numerical Variables

- Trend
  - Positive association
  - Negative association

- Pattern
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

Visualize, then quantify

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 standard(x) and standard(y)

Steps: 4 3 2 1

Measures how clustered the scattered data are around a straight line

Operations that leave *r* unchanged

R is not affected by:

- Changing the units of the measurement of the data
  - Because *r* is based on standard units
- Which variable is plotted on the x- and y-axes
  - Because the product of standard units is the same

Interpreting *r*

Causal Conclusion

Be careful …

- Correlation measures linear association
- Association doesn’t imply causation
- Two variables might be correlated, but that doesn’t mean one causes the other

Nonlinearity and Outliers

Both can affect correlation

- Draw a scatter plot before computing *r*

Ecological Correlation

- Correlations based on groups or aggregated data

- Can be misleading:
  - For example, they can be artificially high

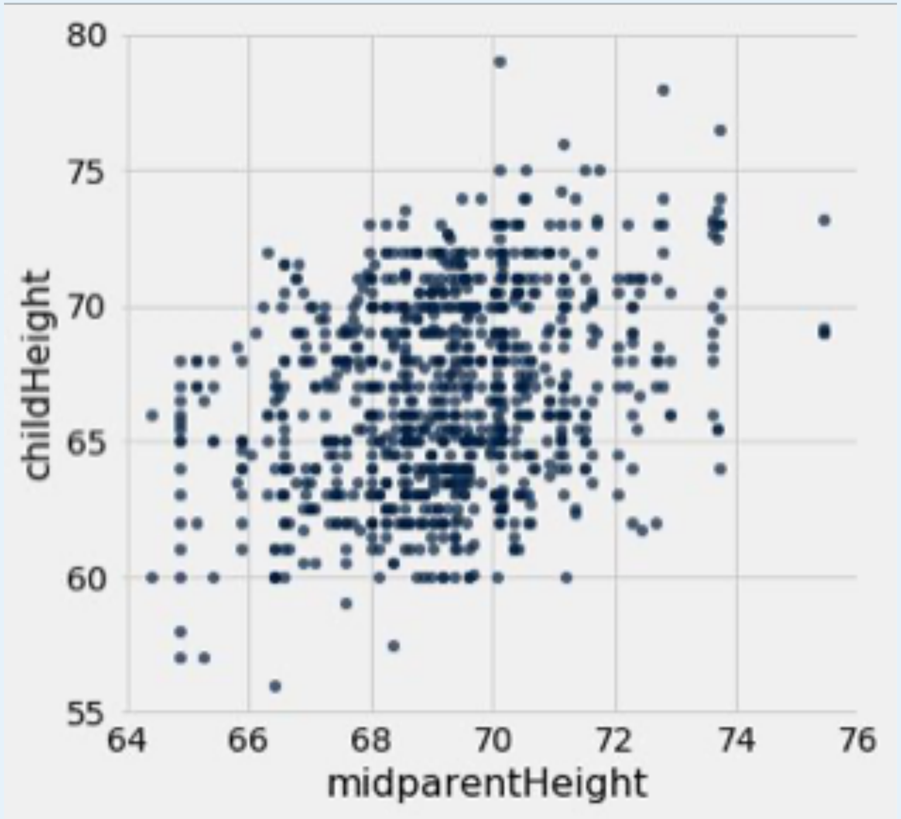
Prediction

Guess the future

- Based on incomplete information

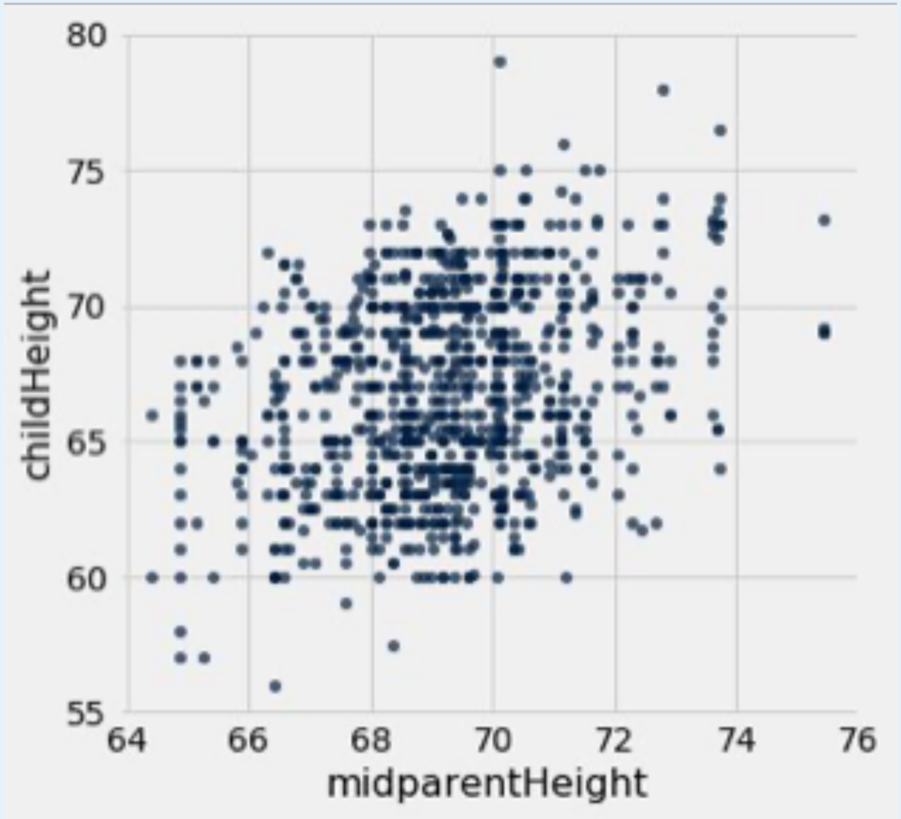
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

Galton's Heights



Goal: Predict the height of a new child, based on that child's midparent height

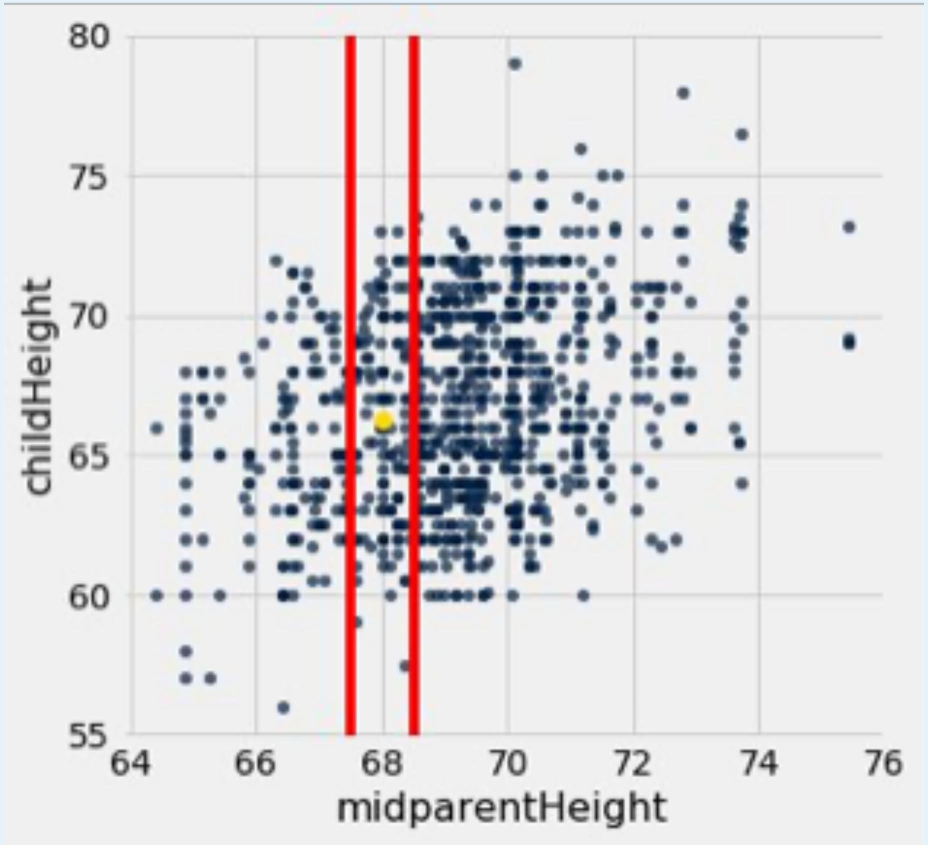
Galton’s Heights



How can we predict a child’s height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

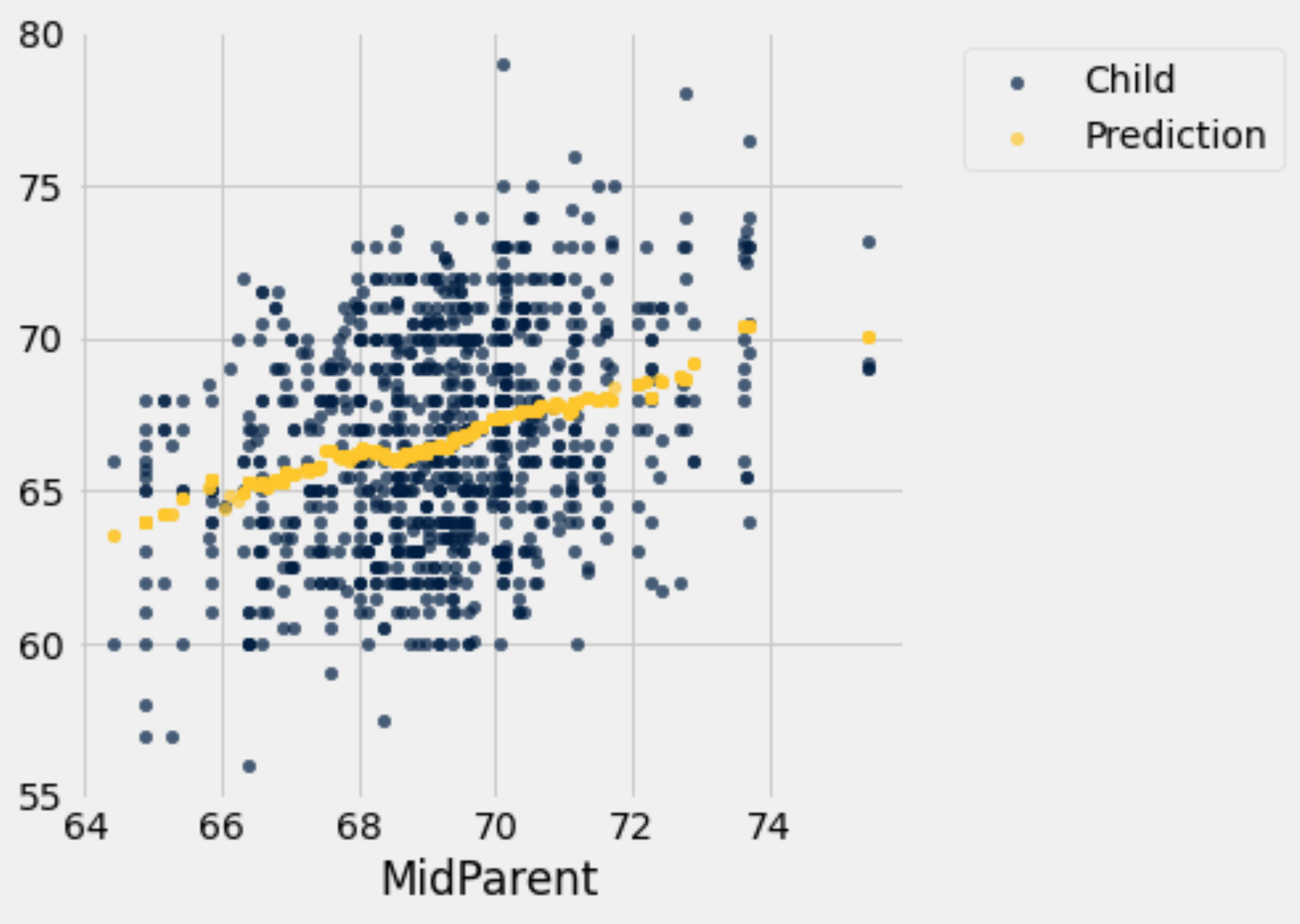
Galton’s Heights



How can we predict a child’s height given a midparent height of 68 inches?

Idea: Use the average height of the children of all families where the midparent Height is close to to 68 inches

Predicted Heights



Graph of Average

For each x value, the prediction is the average of the y values in its nearby group.

The graph of these predictions is the

graph of averages

If the association between x and y is linear, then points in the graph of averages tend to fall on a line. The line is called the **regression line**

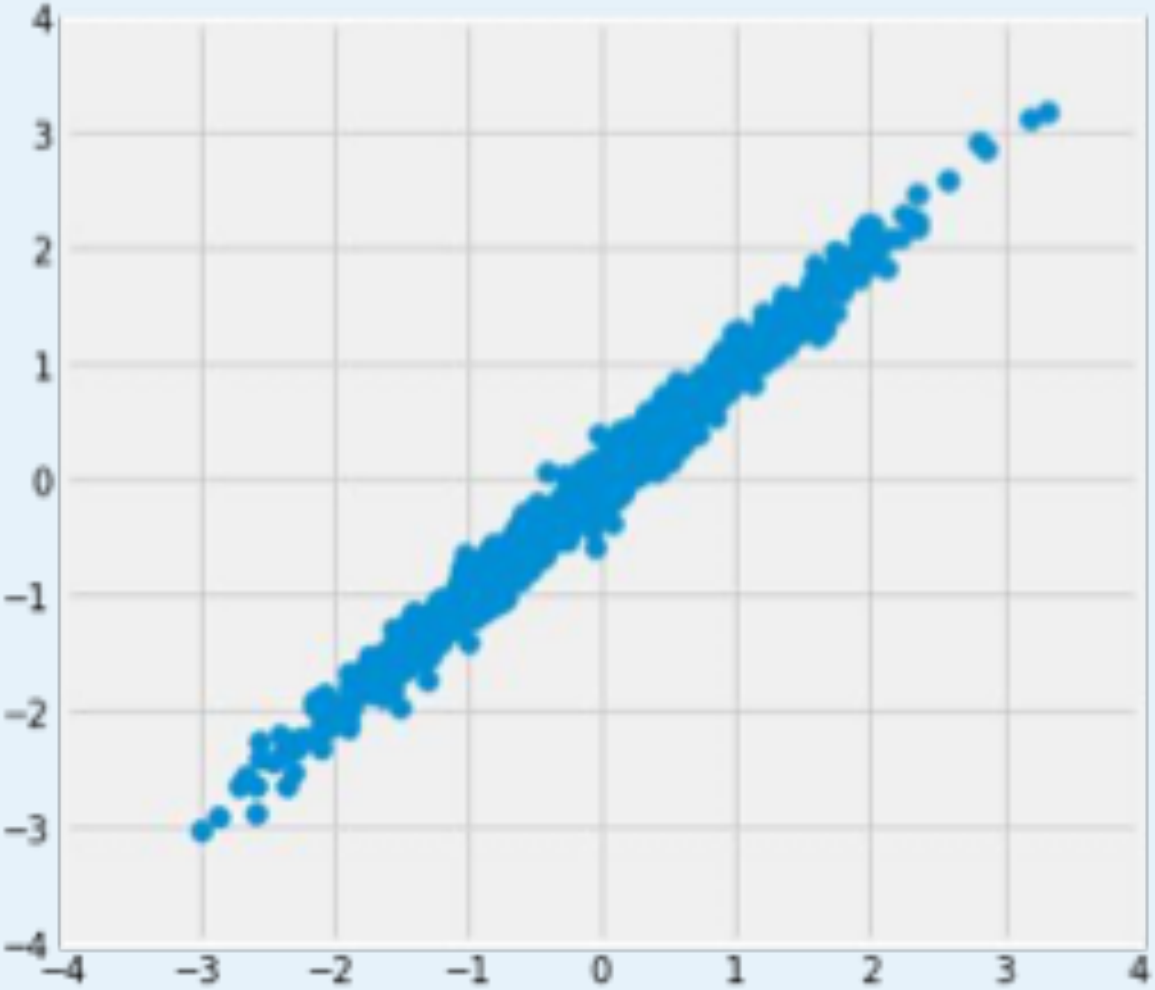
Nearest Neighbor Regression

A method for predicting a numerical y, given a value of x:

- Identify the group of points where the values of x are close to the given value
- The prediction is the average of the y values for the group

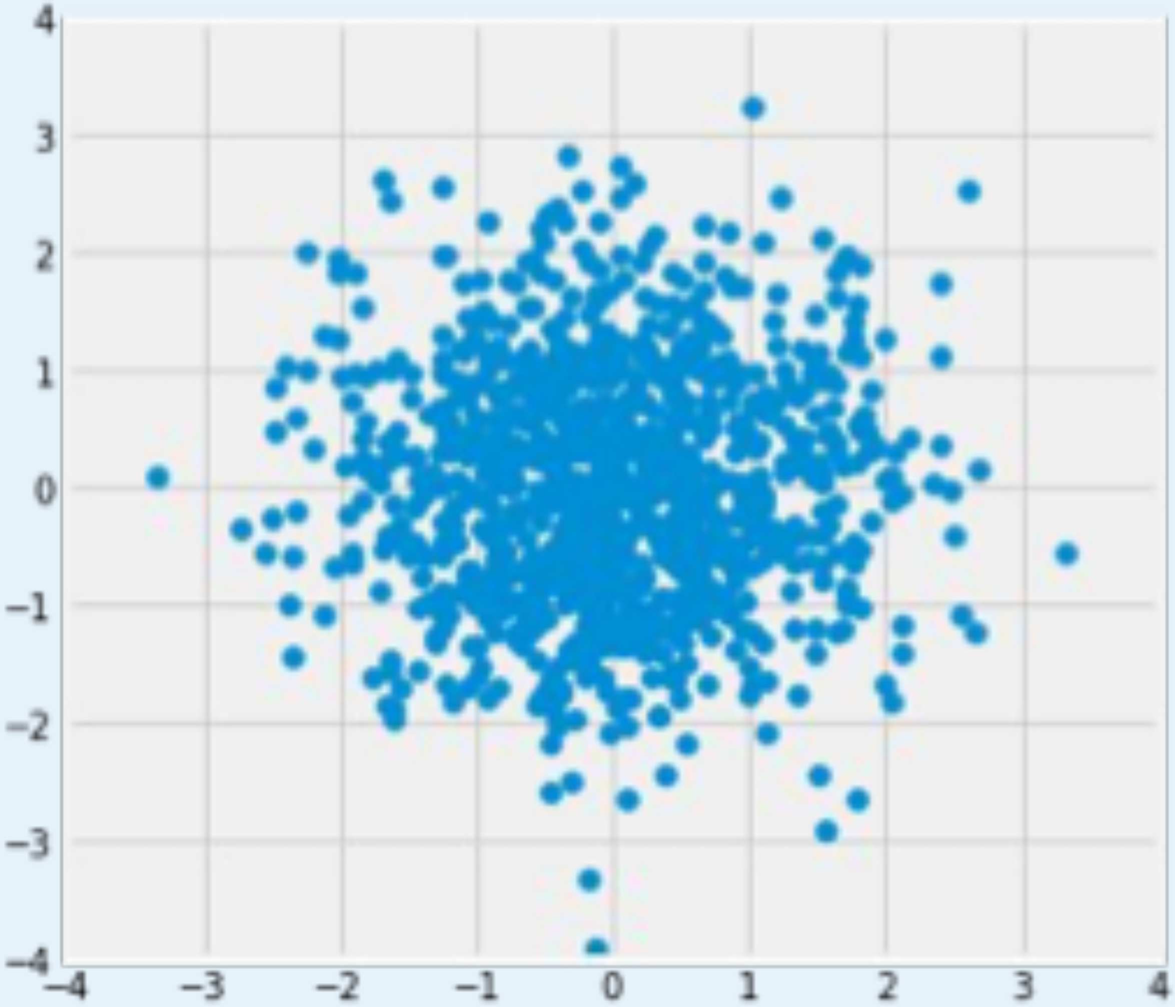
Linear Regression

Where is the prediction line?



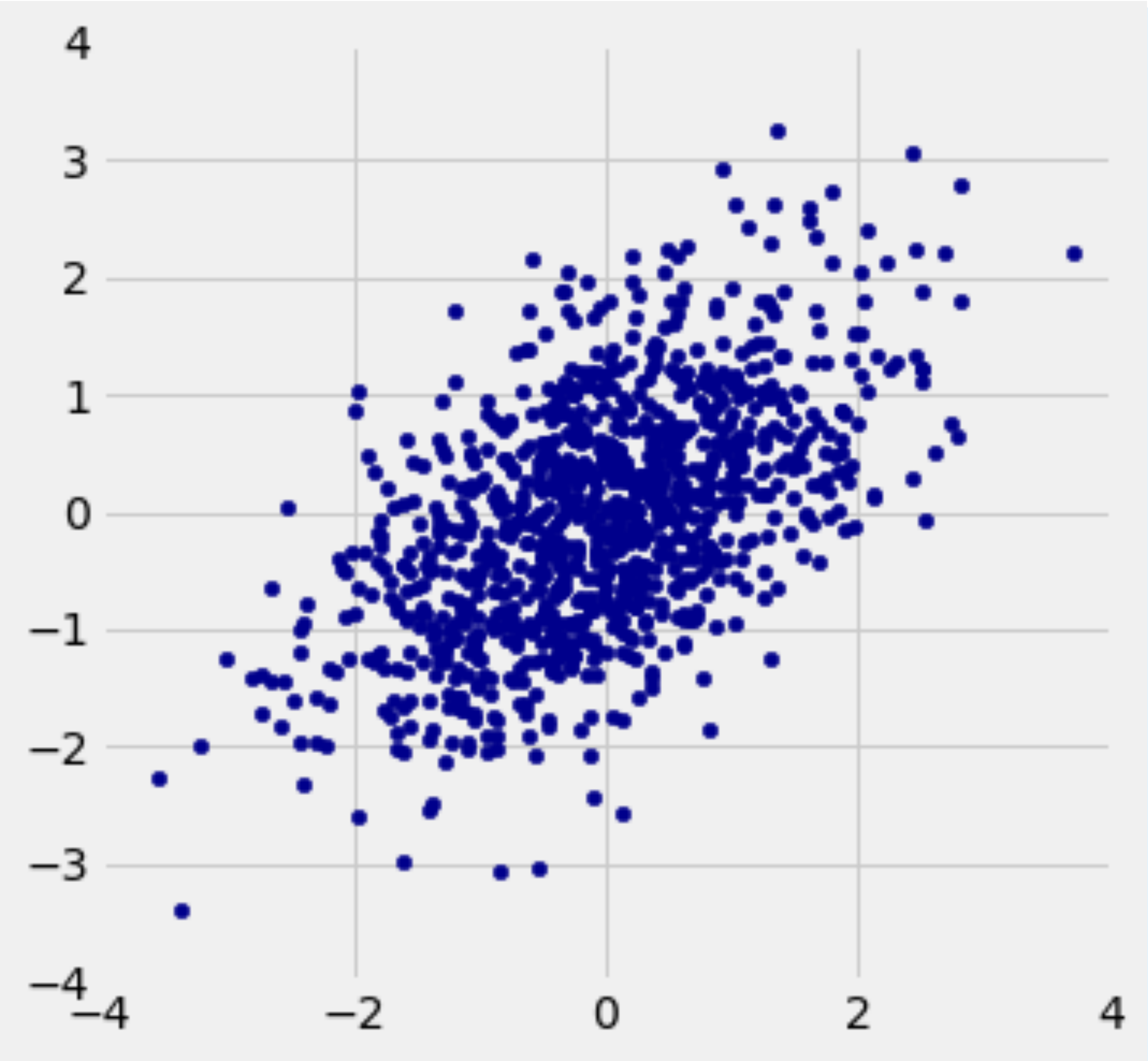
$r = 0.99$

Where is the prediction line?



$r = 0.0$

Where is the prediction line?



$r = 0.5$

Identifying the Line

- If the scatter plot is oval shaped, then we can spot an important feature of the regression line

Linear Regression

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y’s)

On average,

y deviates from 0 less than x deviates from 0

$$y_{su} = r \times x_{su}$$

Slope and Intercept

Regression Line Equation

In original units, the regression line has this equation:

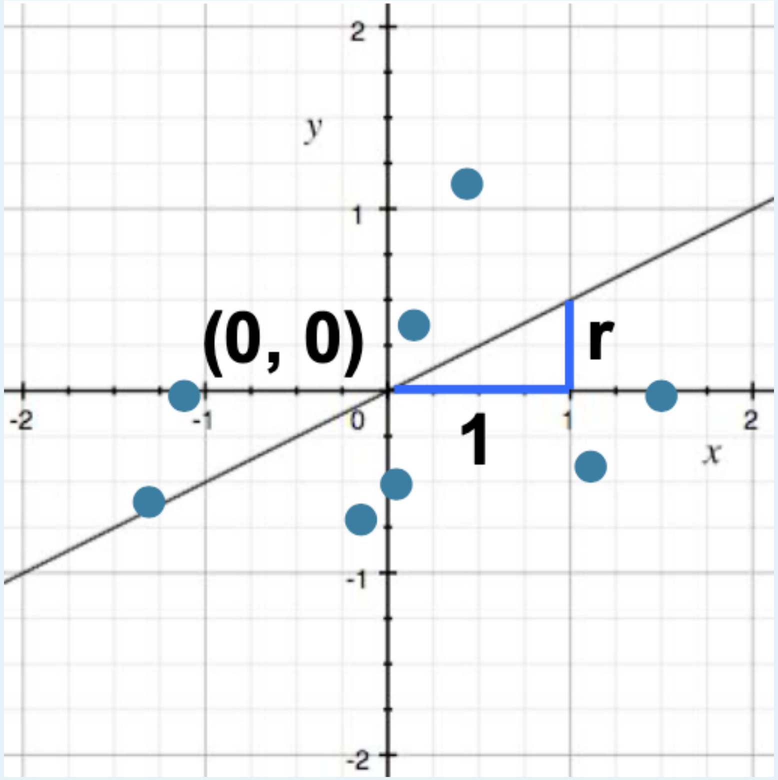
$$\frac{estimate\ of\ y\ - mean(y)}{SD\ of\ y} = r \times \frac{given\ x\ - mean(x)}{SD\ of\ x}$$

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

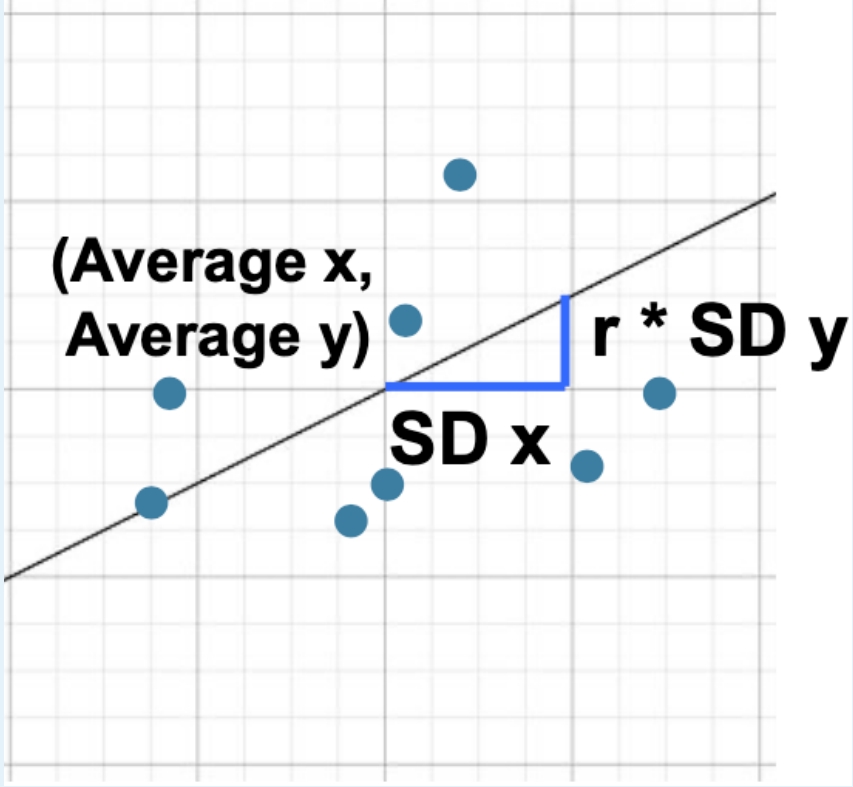
$$y = slope \times x + intercept$$

Regression Line

Standard Units



Original Unites



Slope and Intercept

$$estimate\ of\ y = slope\ * x + intercept$$

$$\boldsymbol{slope\ of\ the\ regression\ line}$$

$$r\ * \frac{SD\ of\ y}{SD\ of\ x}$$

$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

$$mean(y) - slope \times mean(x)$$