MODULE 9

Prediction

PREDICTION

GUESSING 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.

(Demo – Notebook 9.1, Prediction)

ASSOCIATION

TWO NUMERICAL VARIABLES

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible "shape" in the scatter?
    - Linear
    - Non-linear

Visualize, then quantify

(Demo – Notebook 9.1, Association)

CORRELATION COEFFICIENT

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*

(Demo – Notebook 9.1, Correlation)

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***

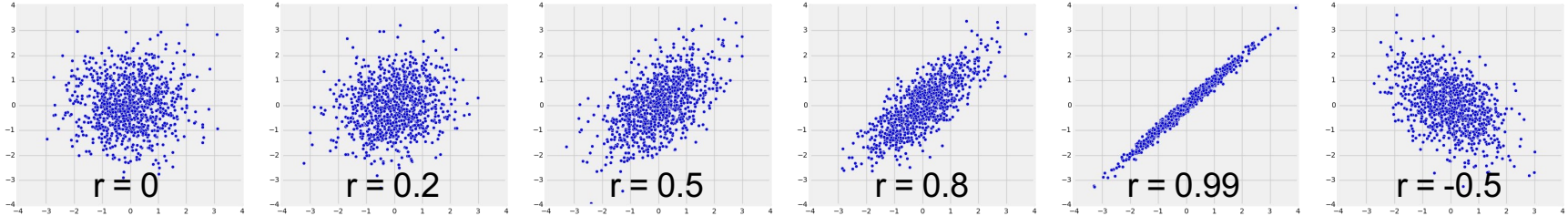
Demo()

PROPERTIES OF *r*

- *r* is a pure number. It has no units.
  - This is because r is based on standard units.
- *r* is unaffected by changing the units on either axis.
  - This too is because r is based on standard units.
- *r* is unaffected by switching the axes. (Demo – Notebook 9.1, Switching Axes)
  - Algebraically, this is because the product of standard units does not depend on which variable is called x and which y.
  - Geometrically, switching axes reflects the scatter plot about the line y=x, but does not change the amount of clustering nor the sign of the association.

RECAP - 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*



CARE IN INTERPRETATION

WATCH OUT FOR …

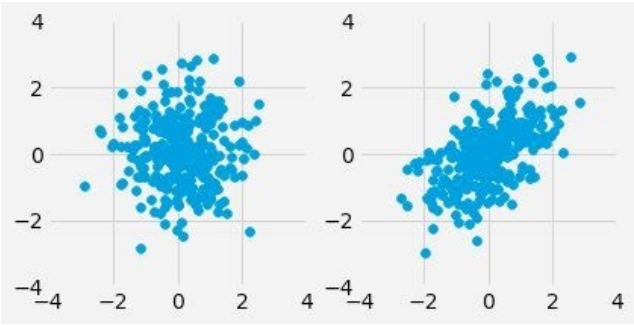
- False conclusions of causation
  - Association is NOT causation
  - Correlation is NOT causation
- Nonlinearity
- Outliers
- Ecological Correlations

(Demo – Notebook 9.1, Nonlinearity, Outliers, and Ecological Correlations)

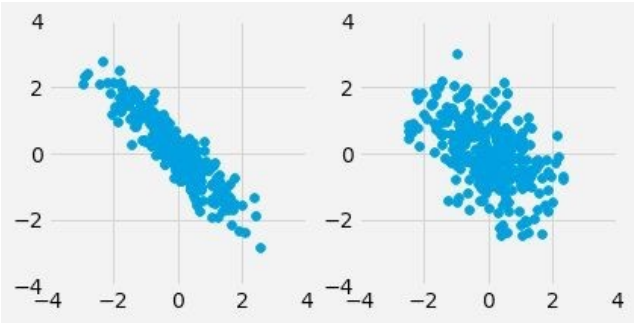
DISCUSSION QUESTION

For each pair, which one will have a higher value of r?

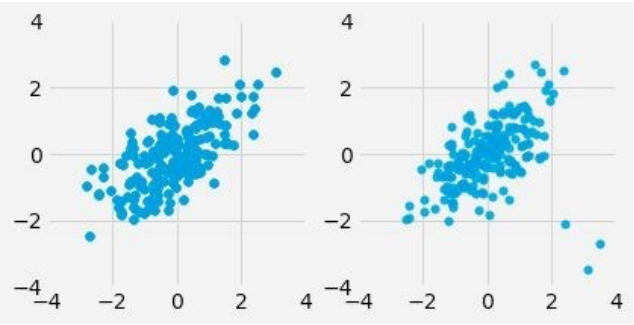
a)



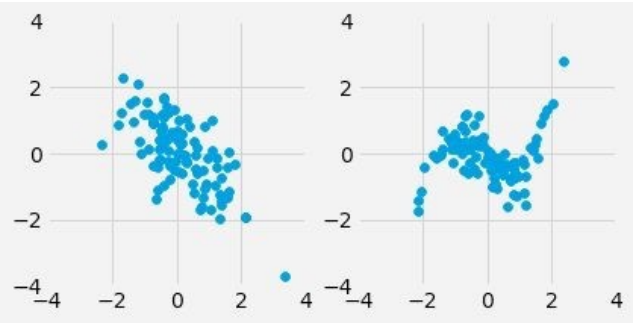
b)



c)



d)



CHOCOLATE AND NOBEL PRIZES

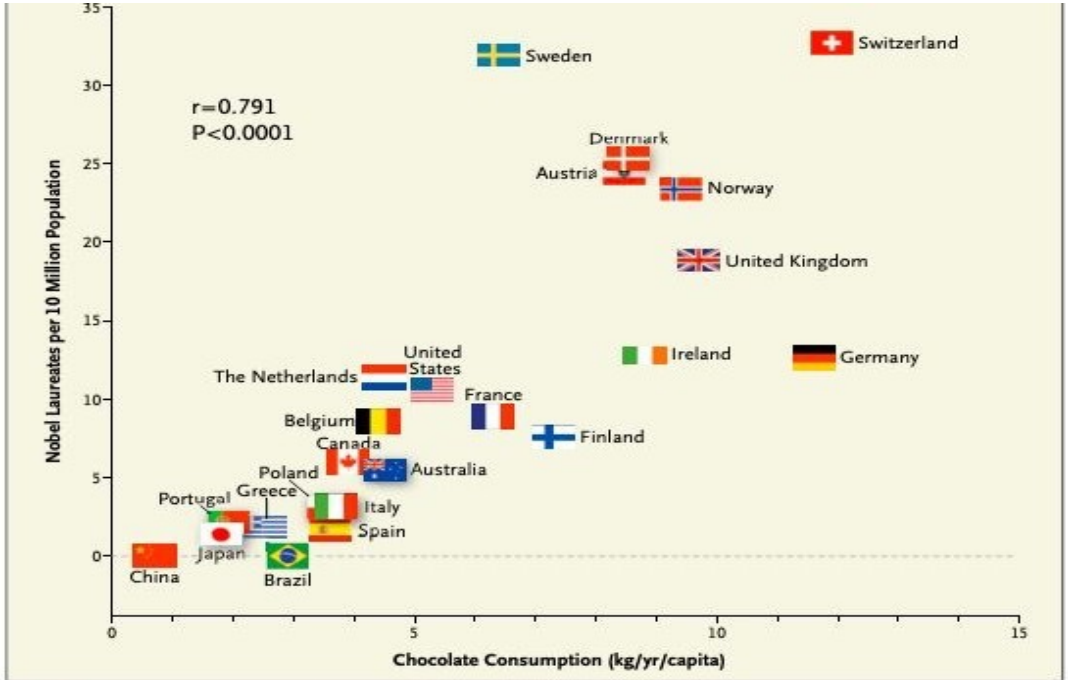


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Reference in course text

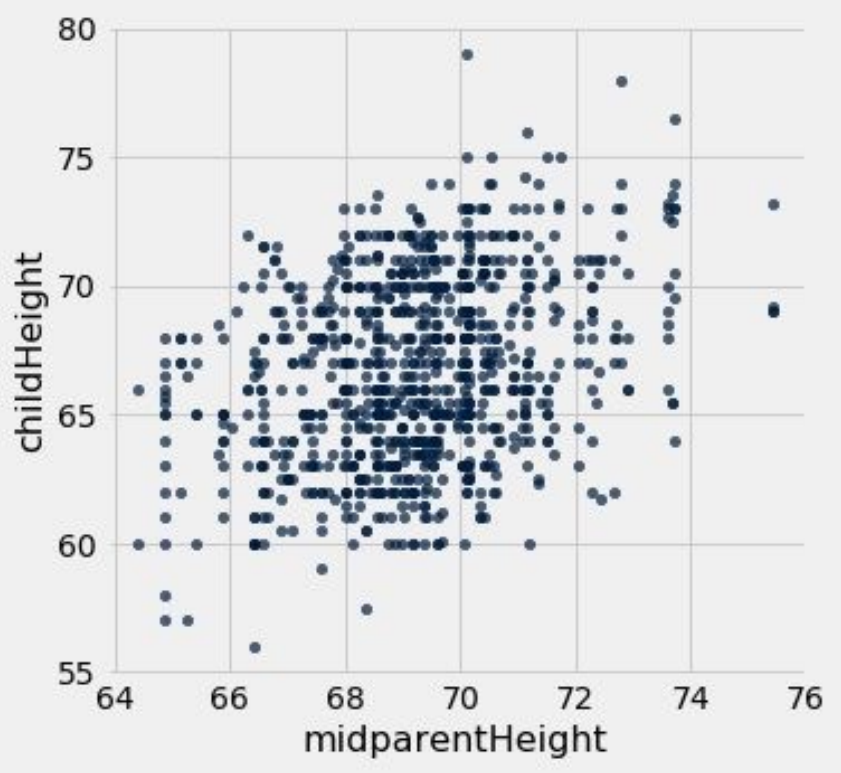
DISCUSSION QUESTION

True or False?

1. If x and y have a correlation of 1, then one must cause the other.
2. If the correlation of x and y is close to 0, then knowing one will never help us predict the other.
3. If x and y have a correlation of -0.8, then they have a negative association.

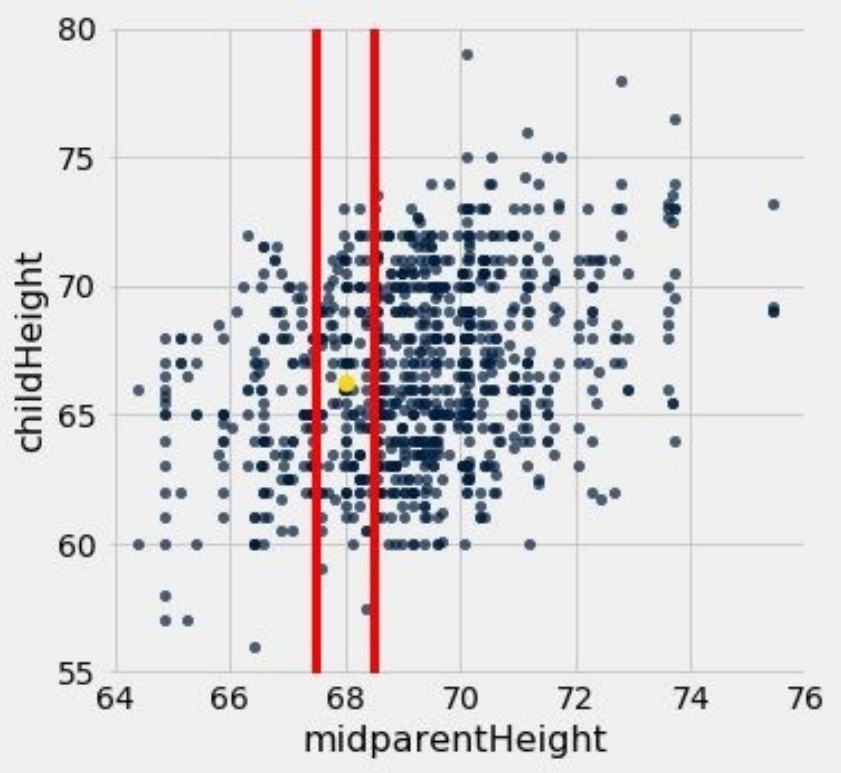
PREDICTION

GALTON'S HEIGHTS

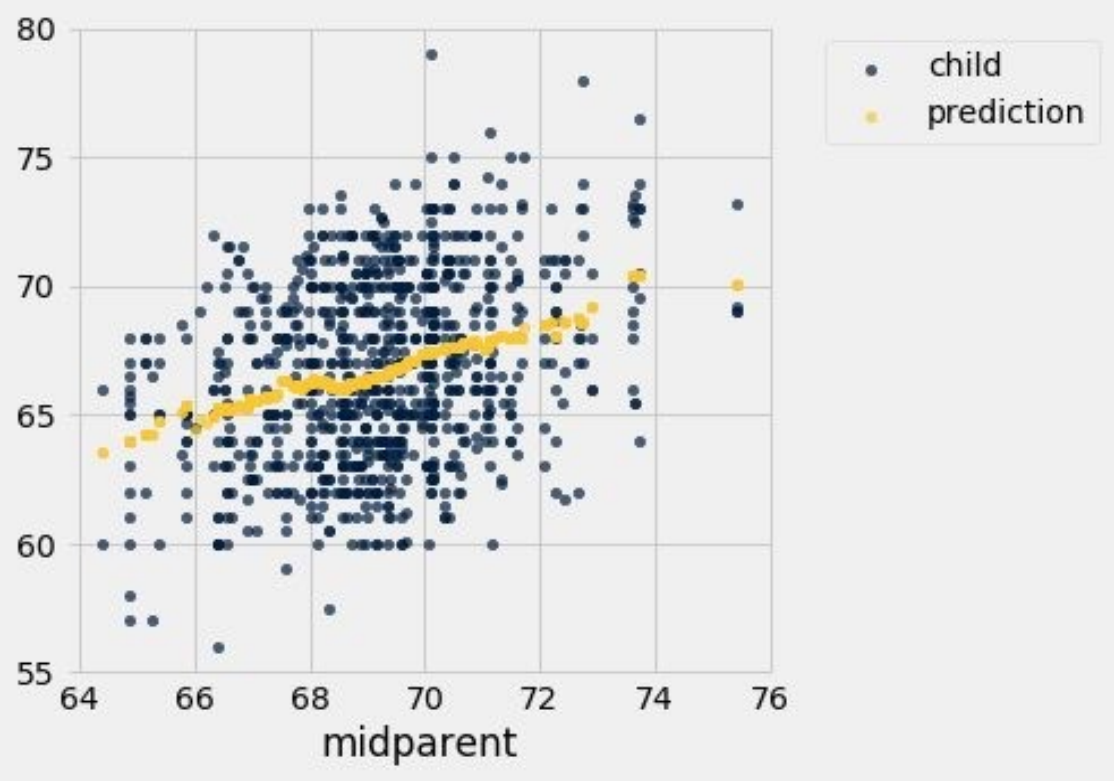


- Oval shaped
- Moderate positive correlation
- How can we predict child height from mid-parent height?

GALTON'S HEIGHTS



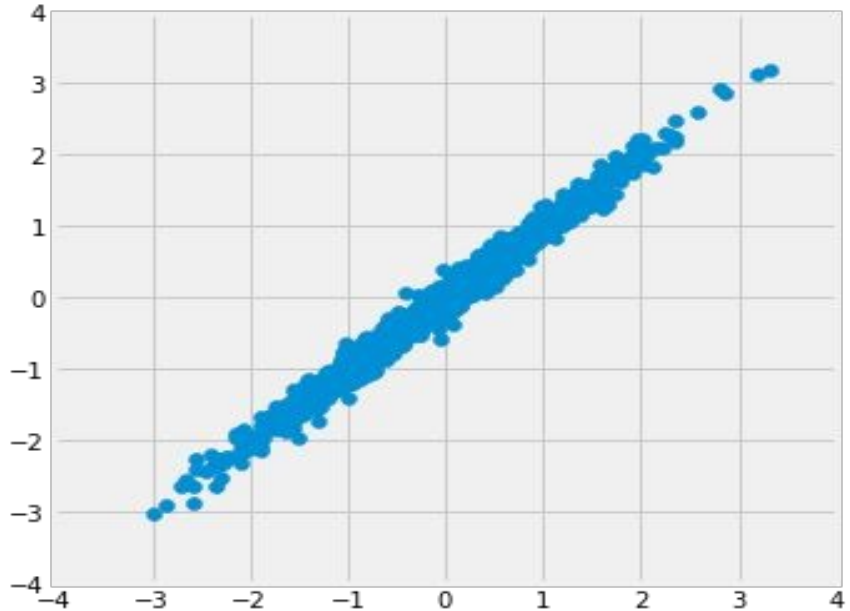
GALTON'S HEIGHTS



NEAREST NEIGHBOR REGRESSION

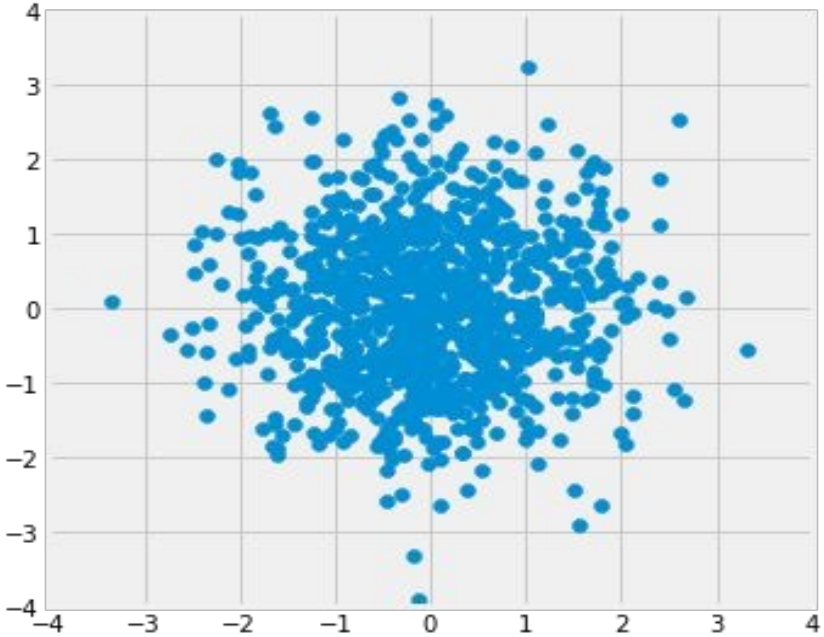
- A method for prediction:
  - Group each x with similar (nearby) x values
  - Average the corresponding y values for each group
- For each x value, the prediction is the average of the y values in its nearby group.
- The graph of these predictions is the "graph of averages".
- If the association between x and y is linear, then points in the graph of averages tend to fall on a line.

WHERE IS THE PREDICTION LINE?



r = 0.99

WHERE IS THE PREDICTION LINE?



r=0.0

(Demo – Notebook 9.2,
Prediction lines)

LINEAR REGRESSION

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

Regression Line

$$y_{(\text{su})} = r \times x_{(\text{su})}$$

Correlation

Not true for all points — a statement about averages

SLOPE & INTERCEPT

REGRESSION LINE EQUATION

In original units, the regression line has this equation:

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

estimated y in standard units

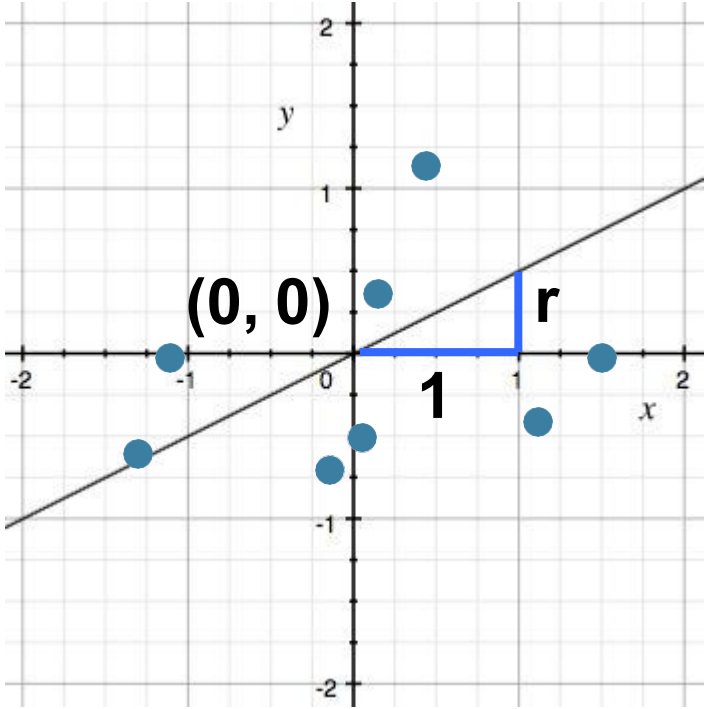
x in standard units

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

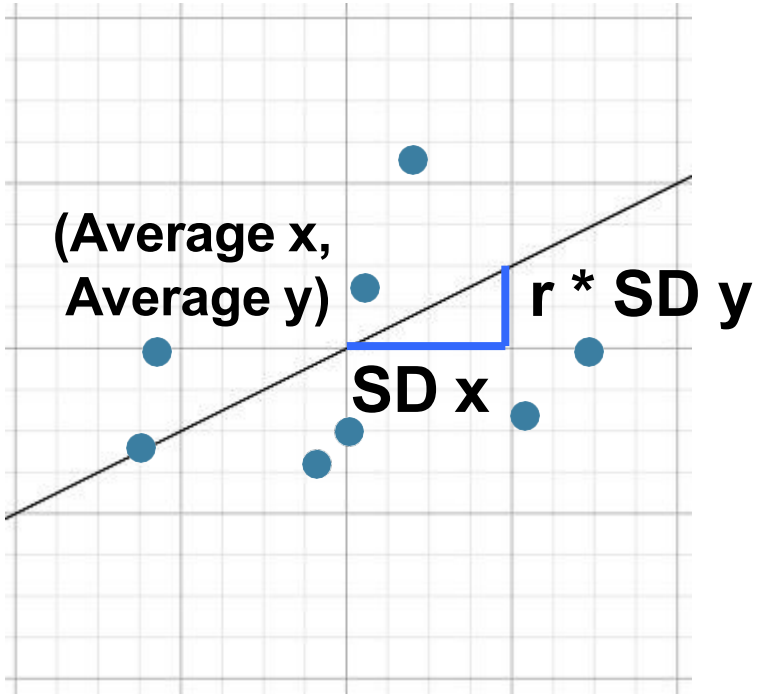
$$y = \text{slope} \times x + \text{intercept}$$

REGRESSION LINE

Standard Units



Original Units



SLOPE AND INTERCEPT

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

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

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

(Demo)

DISCUSSION QUESTION

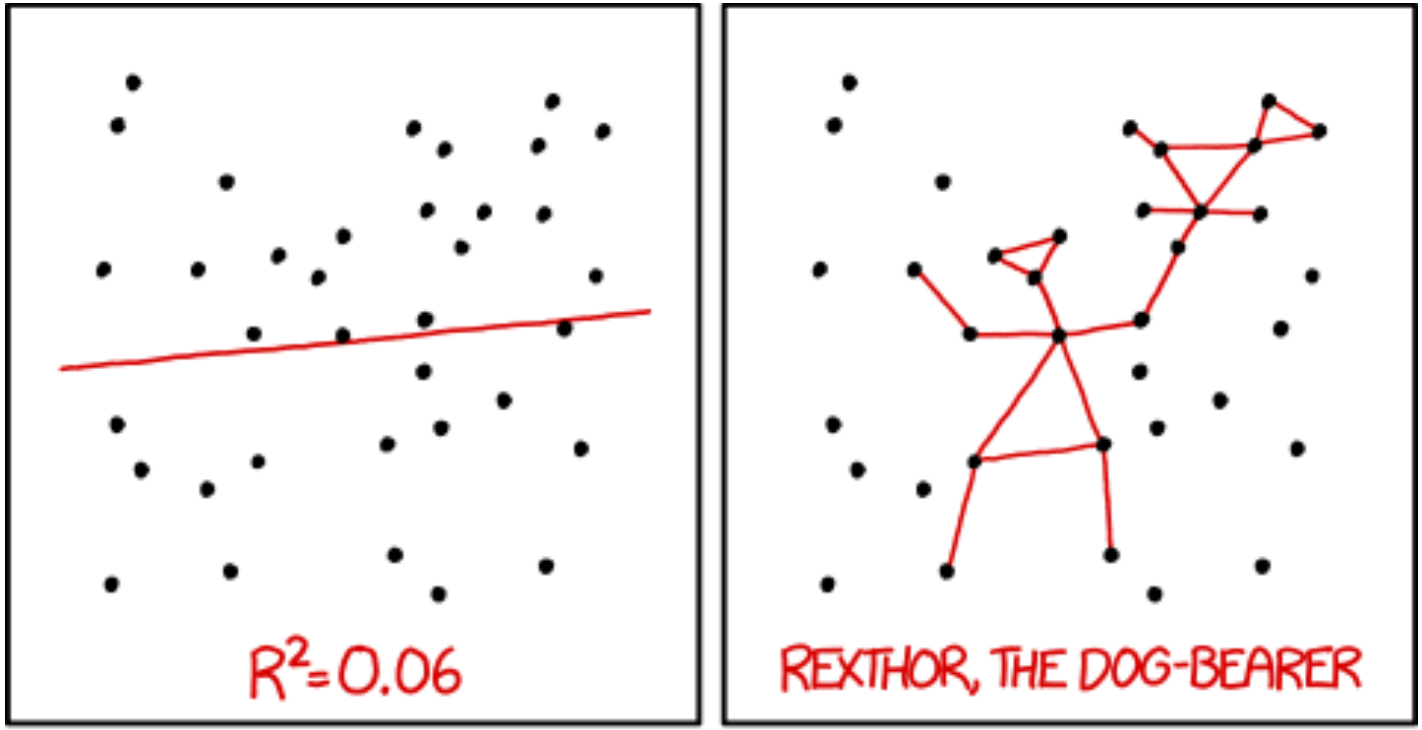
Suppose we use linear regression to predict candy prices (in dollars) from sugar content (in grams). What are the units of each of the following?

- r
- The slope
- The intercept

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 comparing midterm & final scores for students has an oval shape with correlation 0.75, then…
- What do you expect the average final score would be for students who scored 90 on the midterm?
- How about 60 on the midterm?

(Demo)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Source: xkcd.com/1725

LINEAR REGRESSION RECAP

PREDICTION TASK

Goal: Predict *y* using *x*

Examples:

- Predict # *hospital beds available* using *air pollution*
- Predict *house prices* using *house size*
- Predict # *app users* using # *app downloads*

REGRESSION ESTIMATE

Goal: Predict y using x

To find the regression estimate of y:

- Convert the given x to standard units
- Multiply by r
- That's the regression estimate of y, but:
  - It's in standard units
  - So convert it back to the original units of y

REGRESSION LINE EQUATION

In original units, the regression line has this equation:

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

estimated y in standard units

x in standard units

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

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

REGRESSION LINE EQUATION

In original units, the regression line has this equation:

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

what we want

what we observe

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

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

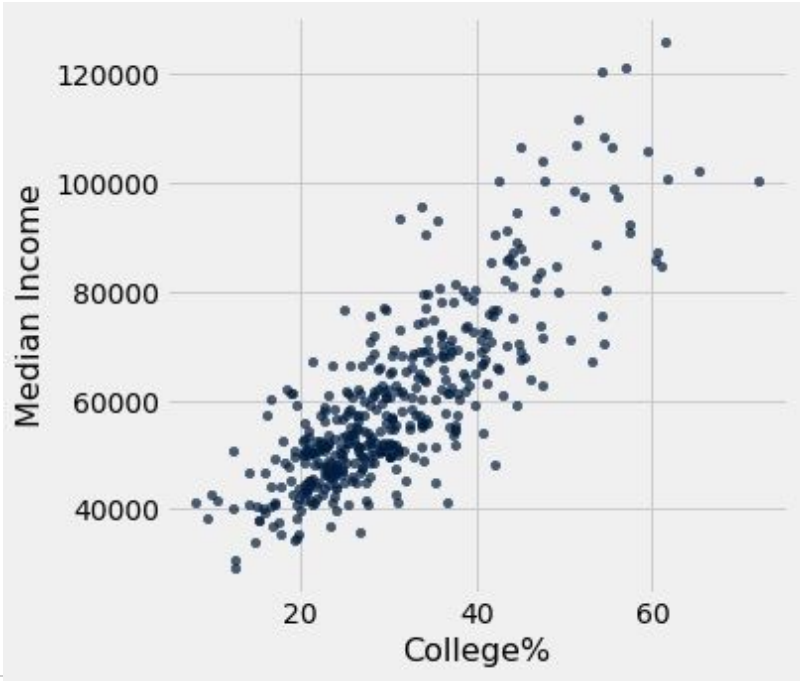
LEAST SQUARES

DISCUSSION QUESTION

Based only on the graph, which must be true? Explain.

1. Going to college causes people to get higher incomes.
2. For any district, having more college-educated people live there causes median incomes to rise.
3. For any district, having a higher median income causes more college-educated people to move there.

USA Congressional Districts, 2016



ERROR IN ESTIMATION

- **error = actual value − estimate**
- Typically, some errors are positive and some 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 (Demo)
  - Regression line

NUMERICAL OPTIMIZATION

- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- If the function **mse(a, b)** returns the mse of estimation using the line "estimate = a*x* + 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)

ERRORS AND RESIDUALS

RESIDUALS

- Error in regression estimate
- One residual corresponding to each point (x, y)
- **residual**
  = observed *y* - regression estimate of *y*
  = observed y - height of regression line at x
  = vertical distance between the point and the best line
- In other words:
  - **observed y = regression estimate + residual**

(Demo)

REGRESSION DIAGNOSTICS

EXAMPLE: DUGONGS

Image Source: National Geographic

(Demo)

RESIDUAL PLOT

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate
- Look for curves, trends, changes in spread, outliers, or any other patterns

PROPERTIES OF RESIDUALS

- Residuals from a linear regression **always** have
  - **Zero** mean
    - (so **rmse = SD of residuals**)
  - **Zero** correlation with x
  - **Zero** correlation with the fitted values

- These are all true **no matter what the data look like**
  - Just like deviations from mean are zero on average

(Demo)

DISCUSSION QUESTIONS

How would we adjust our regression line…

- if the average residual were 10?
- if the residuals were positively correlated with x?
- if the residuals were above 0 in the middle and below 0 on the left and right?

A MEASURE OF CLUSTERING

CORRELATION, REVISITED

- "The correlation coefficient measures how clustered the points are about a straight line."

- We can now quantify this statement.

(Demo)

SD OF FITTED VALUES

- SD of fitted values

$$\frac{\text{SD of fitted values}}{\text{SD of } y} = |r|$$

- SD of fitted values $=$ $|r|$ * (SD of y)

VARIANCE OF FITTED VALUES

- Variance = Square of the SD
  = Mean Square of the Deviations
- Variance has weird units, but good math properties

- $\dfrac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2$

A VARIANCE DECOMPOSITION

By definition,

$$y = \text{fitted values} + \text{residuals}$$

Tempting (but wrong) to think that:

~~$$SD(y) = SD(\text{fitted values}) + SD(\text{residuals})$$~~

But it **is** true that:

$$Var(y) = Var(\text{fitted values}) + Var(\text{residuals})$$

(a result of the **Pythagorean theorem!**)

A VARIANCE DECOMPOSITION

Var(y) = Var(fitted values) + Var(residuals)

- $\dfrac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2$

- $\dfrac{\text{Variance of residuals}}{\text{Variance of } y} = 1 - r^2$

RESIDUAL AVERAGE AND SD

- The average of residuals is always 0

- $$\frac{\text{Variance of residuals}}{\text{Variance of } y} = 1 - r^2$$

- SD of residuals $= \sqrt{(1-r^2)}$ SD of y

(Demo)

RESIDUAL AVERAGE AND SD

- The average of residuals is always 0

- SD of residuals $= \sqrt{(1 - r^2)} * \text{SD of } y$

- SD of predictions $= |r| * \text{SD of } y$

(Demo)

DISCUSSION QUESTION 1

Midterm: Average 70, SD 10

Final: Average 60, SD 15

$r = 0.6$

Fill in the blank:

The SD of the residuals is _____.

DISCUSSION QUESTION 2

Midterm: Average 70, SD 10

Final: Average 60, SD 15

$r = 0.6$

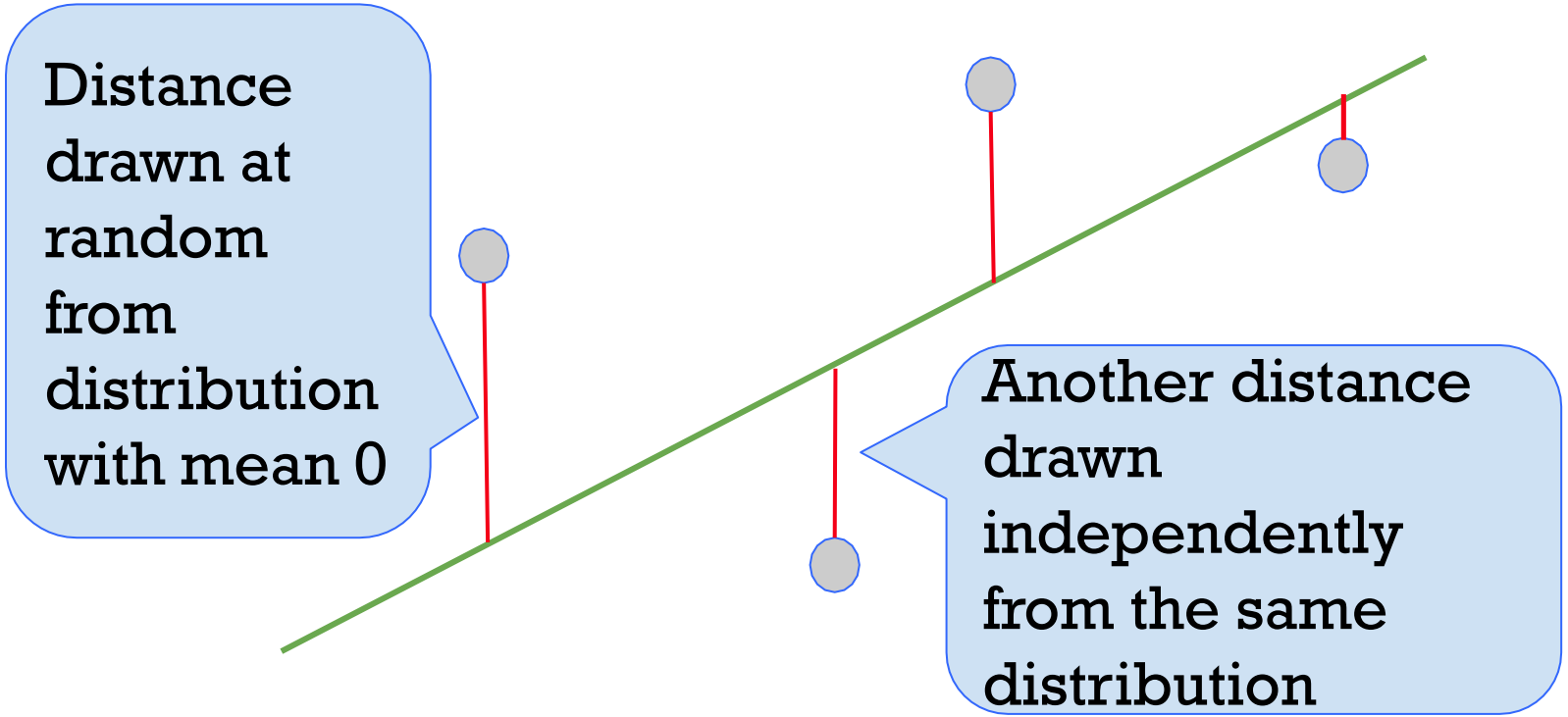
Fill in the blank:

For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within

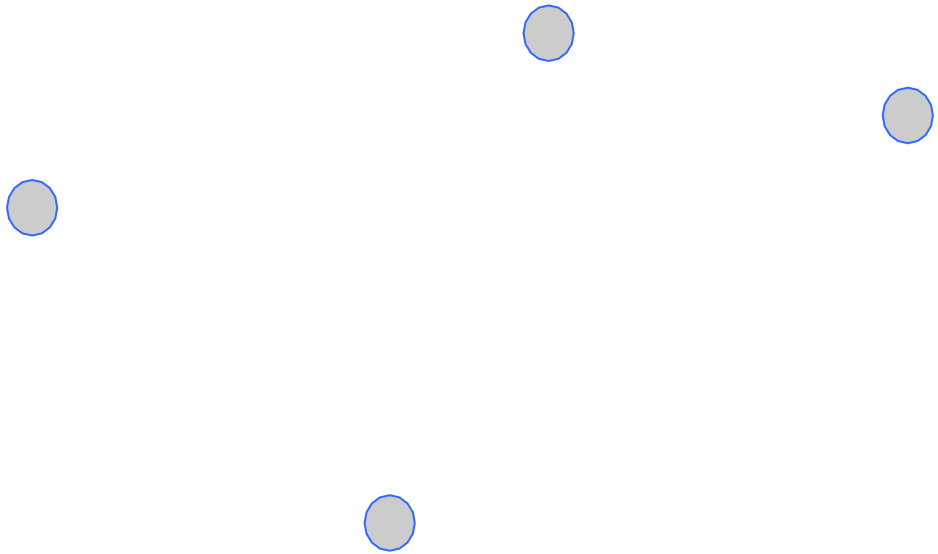
_____points.

REGRESSION MODEL

A "MODEL": SIGNAL + NOISE



WHAT WE GET TO SEE



(Demo)

PREDICTION VARIABILITY

REGRESSION PREDICTION

- **If the data come from the regression model,**
- **and if the sample is large, then:**

- The regression line is close to the true line
- Given a new value of x, predict y by finding the point on the regression line at that x

(Demo)

CONFIDENCE INTERVAL FOR PREDICTION

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the "middle 95%" interval.
- That's an approximate 95% confidence interval for the height of the true line at *y*.

(Demo)

PREDICTIONS AT DIFFERENT VALUES OF *X*

- Since *y* is correlated with *x*, the predicted values of *y* depend on the value of *x*.

- The width of the prediction's CI also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.

(Demo)

THE TRUE SLOPE

CONFIDENCE INTERVAL FOR TRUE SLOPE

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- Draw the empirical histogram of all the generated slopes.
- Get the "middle 95%" interval.
- That's an approximate 95% confidence interval for the slope of the true line.

(Demo)

RAIN ON THE REGRESSION PARADE

We observed a slope based on our sample of points.

But what if the sample scatter plot got its slope just by chance?

What if the true line is actually FLAT?

(Demo)

TEST WHETHER THERE REALLY IS A SLOPE

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn't contain 0, the data are more consistent with the alternative
  - If the interval does contain 0, the data are more consistent with the null

(Demo)

ADVANCED REGRESSION

ADVANCED REGRESSION

- minimize() works no matter what*!

- Define a function that computes the prediction you want, then the error you want, for example:
  - Nonlinear functions of *x*
  - Multiple columns of the table for *x*
  - Other kinds of error instead of RMSE
- Nonlinear functions can get complicated, fast!

(Demo)

PREDICTION

GUESSING THE VALUE OF AN ATTRIBUTE

- 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.

- Two Types of Prediction
  - Classification = Categorical; Regression = Numeric

PREDICTION EXAMPLE: SPAM OR NOT?

You made a Wells Fargo payment - wellsfargo.com You recently submitted a payment The ...

BUSINESS TRUST - -- I have a legal business proposal for you worth $23,000,000. If you kn...

Hi - Today???!!!! What a wonderful day! Congrats again! I am definitely not doing s...

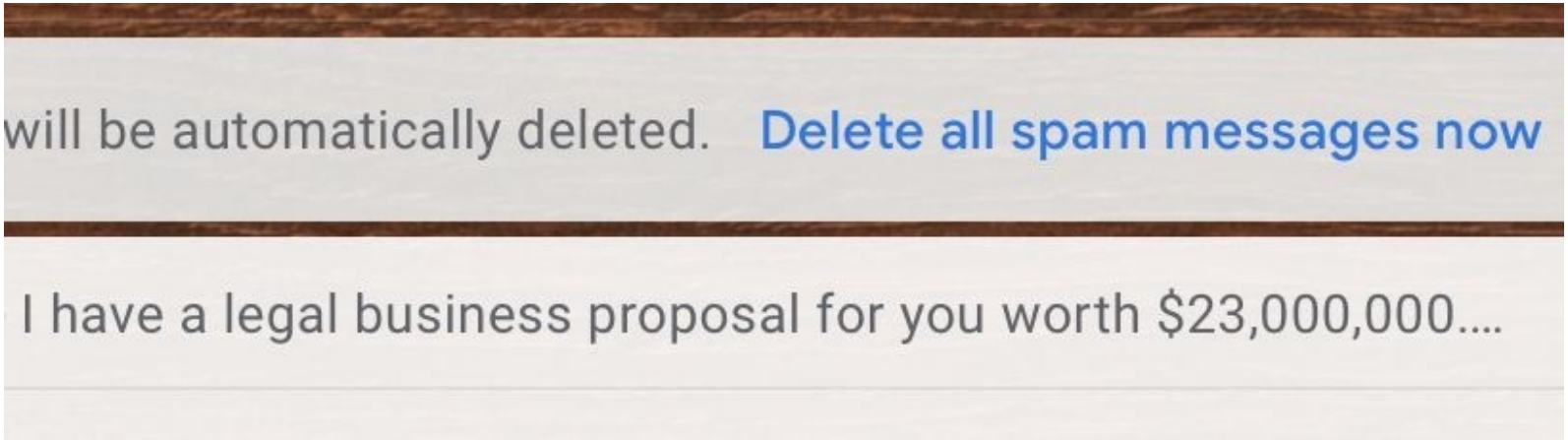
Michael Kors Handbags Up To 84% Plus Free Shipping! - Shop Handbags Online & In Store...

MACHINE LEARNING ALGORITHM

- A mathematical model
  - calculated based on sample data ("training data")
  - that makes predictions or decisions without being explicitly programmed to perform the task

CLASSIFICATION

CLASSIFICATION EXAMPLES



CLASSIFICATION EXAMPLES

QUESTIONS?