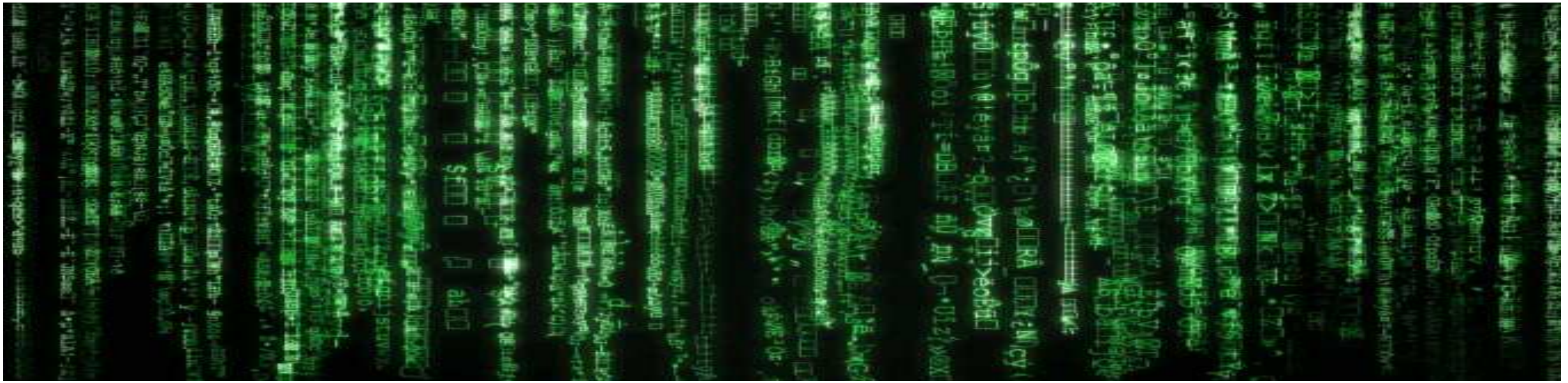


INF 110 Discovering Informatics

Prediction

NAU NORTHERN ARIZONA UNIVERSITY

Prediction and Informatics

- What can data tell us about the future?

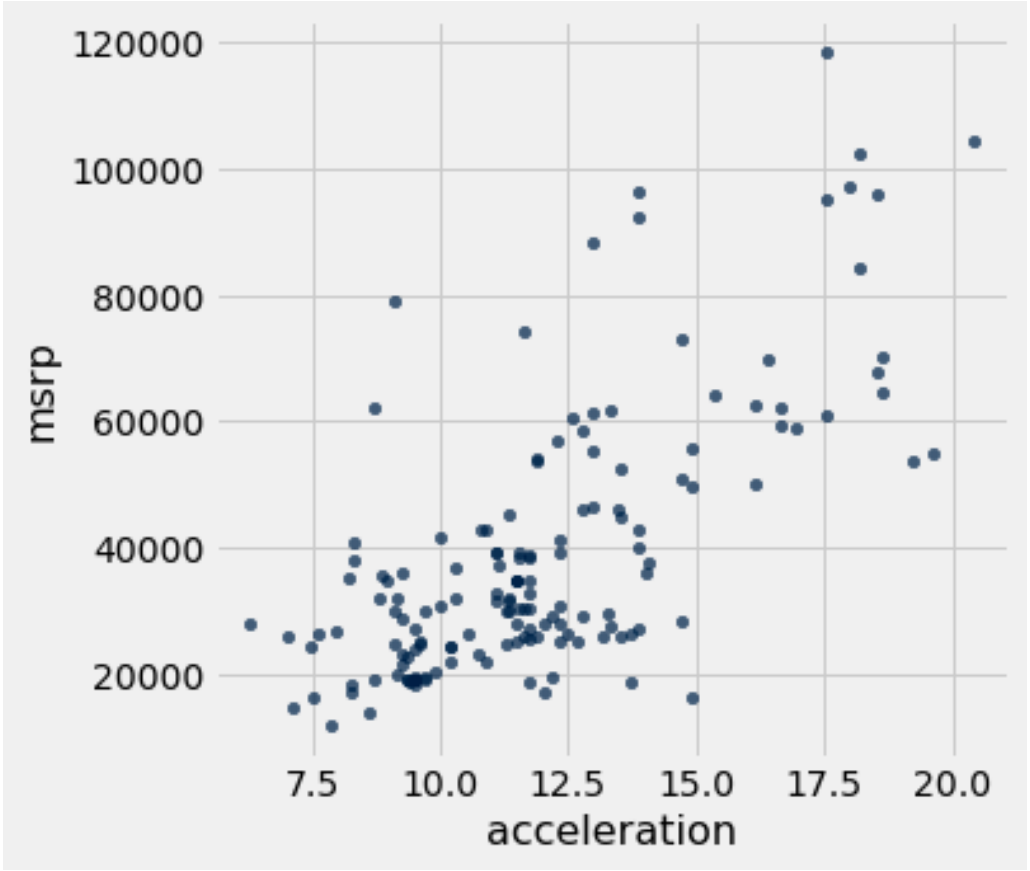
climate.nasa.gov

What can past climate data tell us about future temperatures?

Based on a person's social media usage, what websites will interest them?

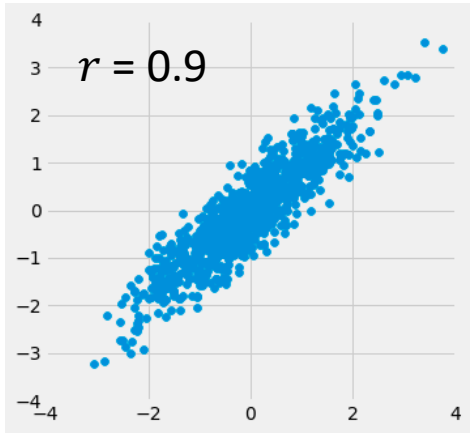
Correlation

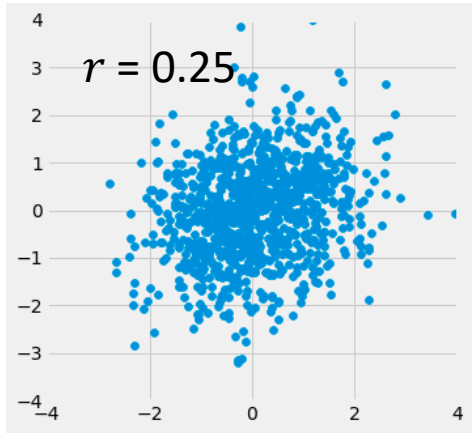
- Measuring linear association
- For instance, see how a hybrid car model's price is related to its acceleration rate.
- The scatter of points slopes upwards, indicating an ***association:***
  - cars with greater acceleration also tend to cost more.

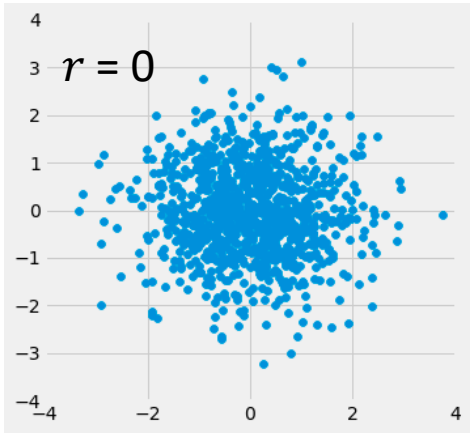


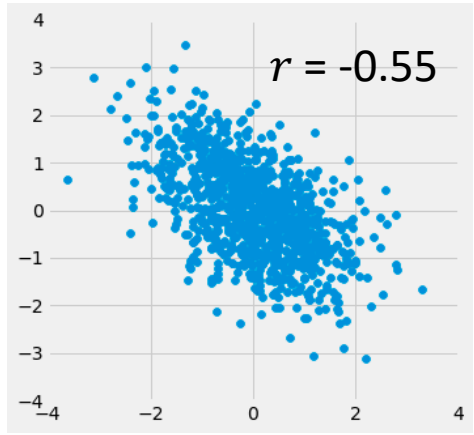
Correlation Coefficient (r)

- Measures the strength of the linear relationship
- How clustered is the scatter plot around a straight line?
- r is between -1 and 1.
- r = 1, the scatter is a straight line sloping upwards
- r = -1, the scatter is a straight line sloping downwards.



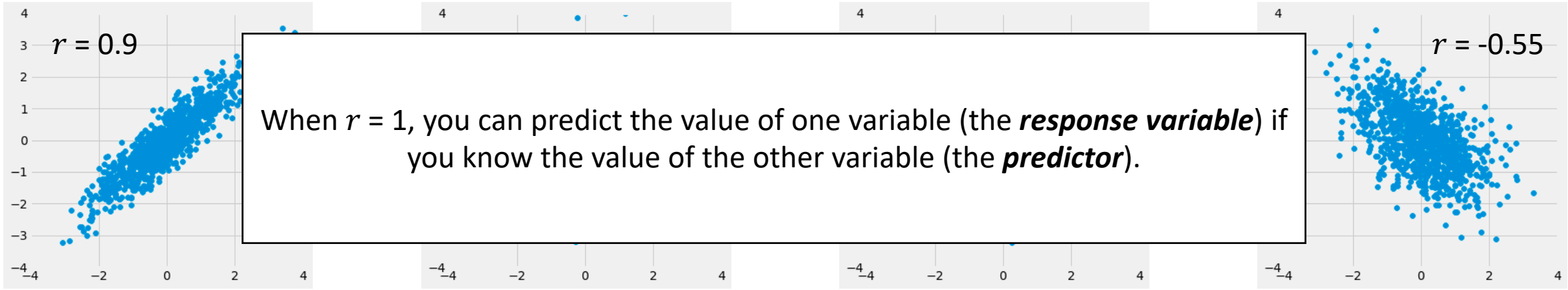






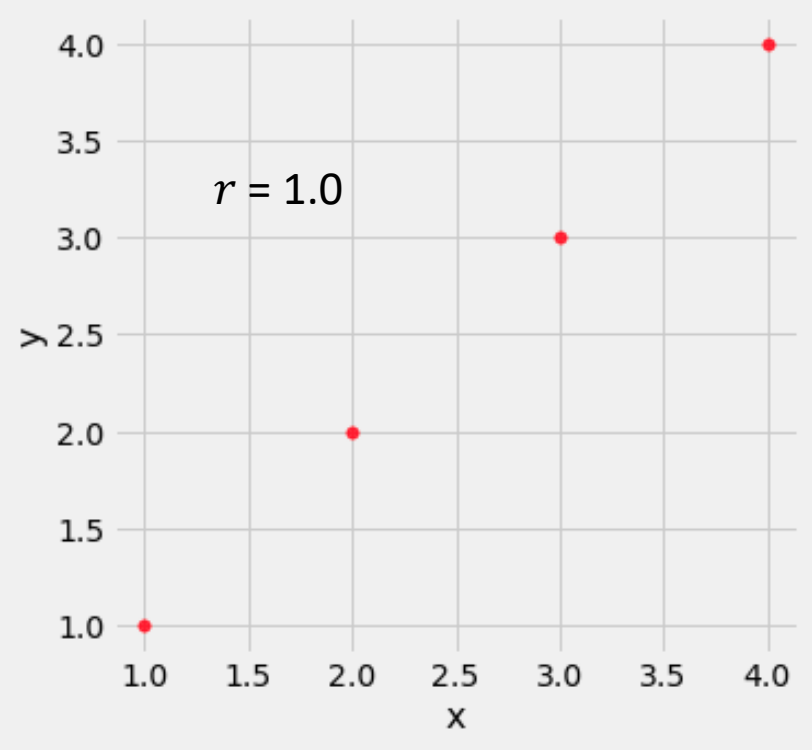
Correlation Coefficient (r)

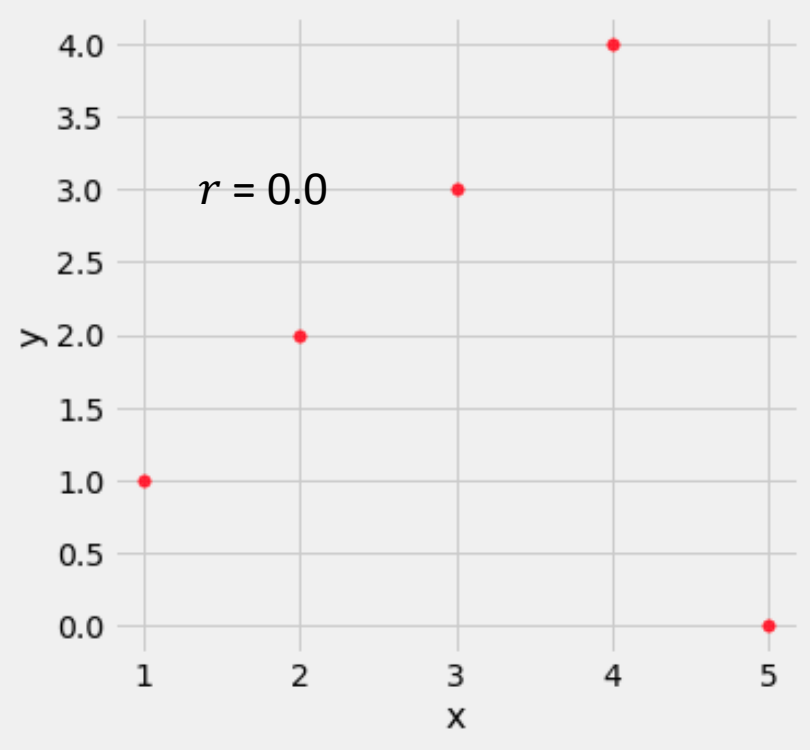
- Measures the strength of the linear relationship
- How clustered is the scatter plot around a straight line?
- r is between -1 and 1.
- r = 1, the scatter is a straight line sloping upwards
- r = -1, the scatter is a straight line sloping downwards.



When r = 1, you can predict the value of one variable (the ***response variable***) if you know the value of the other variable (the ***predictor***).

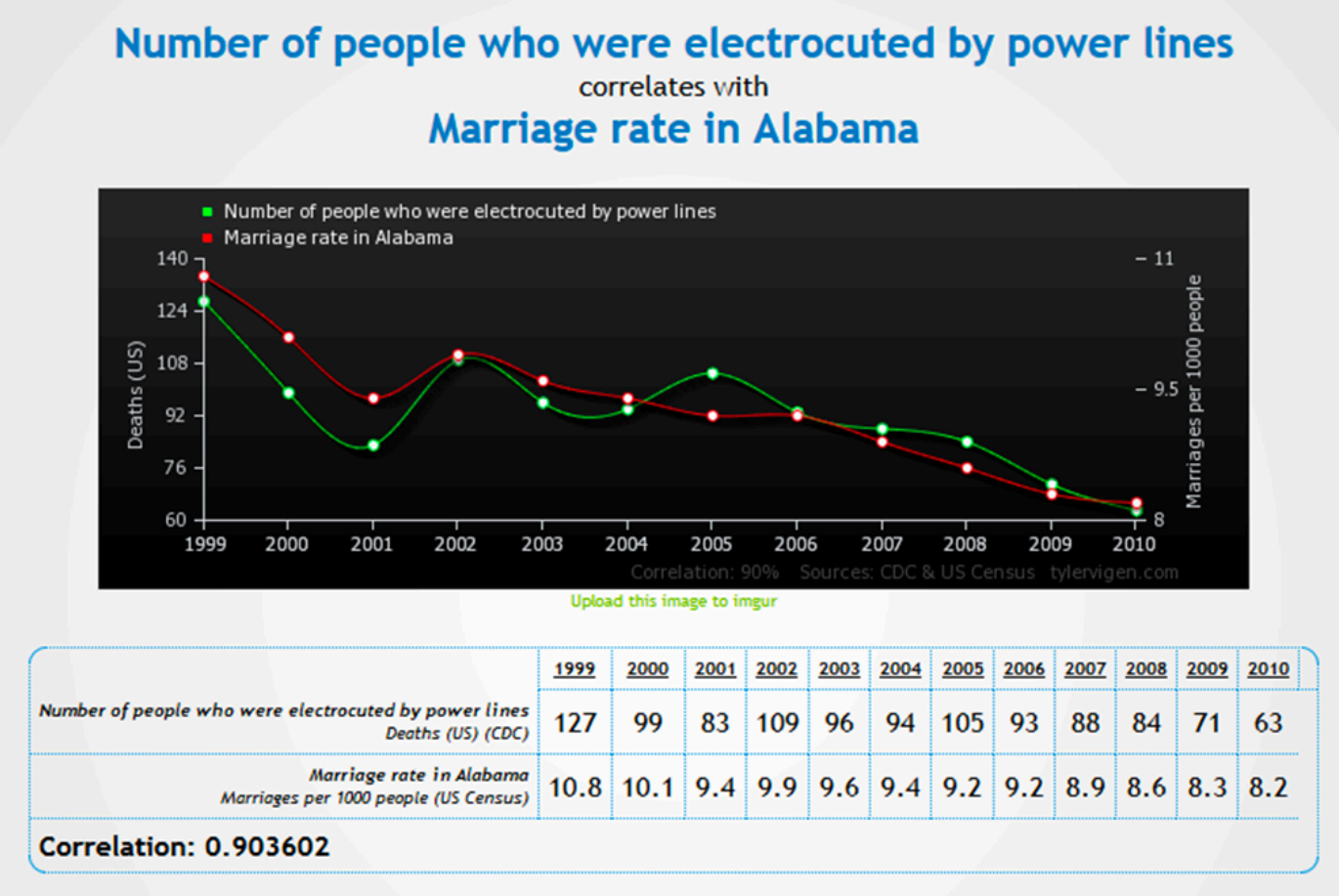
Correlation is affected by outliers





Correlation

- Only measures association
- ***Does not imply causation!***



	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Number of people who were electrocuted by power lines *Deaths (US) (CDC)*	127	99	83	109	96	94	105	93	88	84	71	63
Marriage rate in Alabama *Marriages per 1000 people (US Census)*	10.8	10.1	9.4	9.9	9.6	9.4	9.2	9.2	8.9	8.6	8.3	8.2

Correlation: 0.903602

http://www.setheliot.com/blog/2015/05/10/misleading-correlations/

Correlations

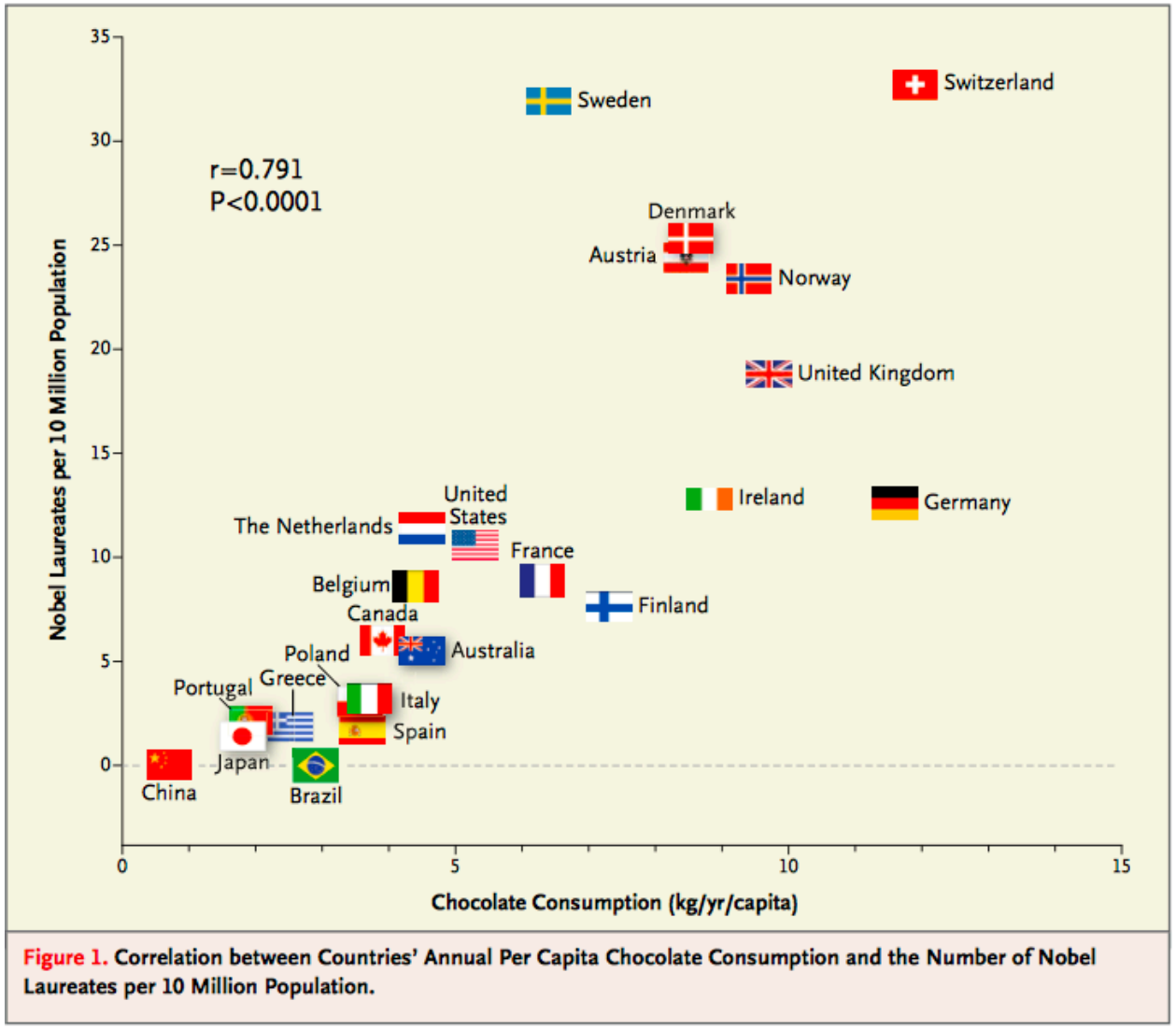
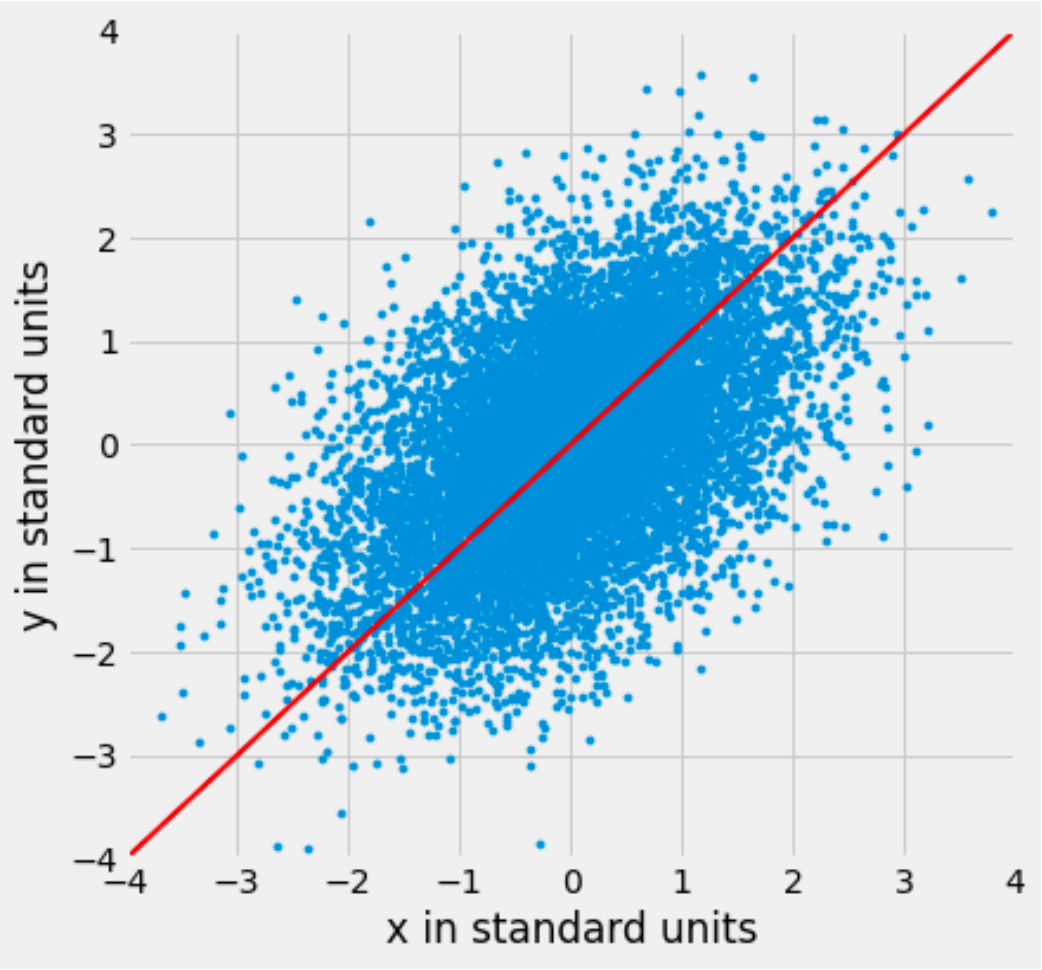


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Messerli FH. 2012. Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine.*

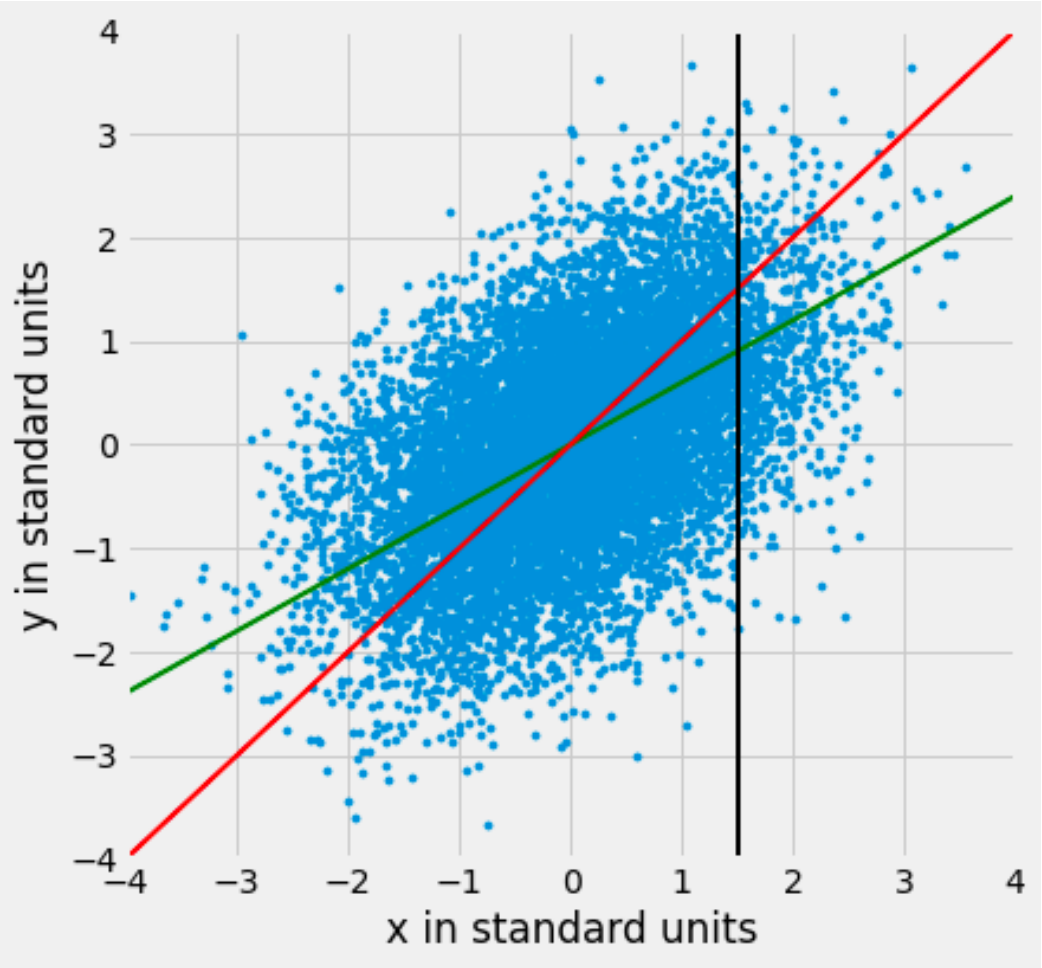
The Regression Line

- The straight line around which the points in a scatter plot are clustered
- Here is a football shaped plot
  - 45 degree line is in red.



The Regression Line

- The green line represents the "graph of averages"
  - It goes through the center of the vertical strips
  - ***It is flatter than the red line.***
- The slope of the 45 degree line is 1.
- The slope of the green line is less than 1.
- It's r!



Residuals

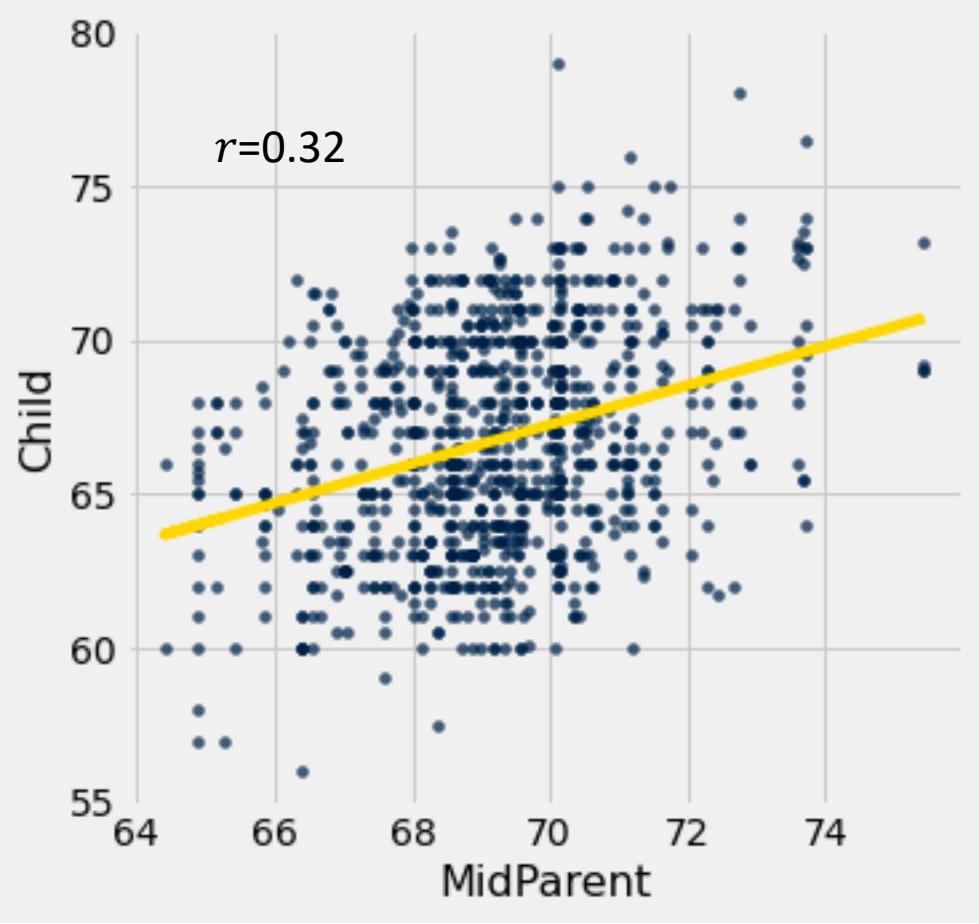
- Regression can be used to predict values of one variable based on the values of another variable.

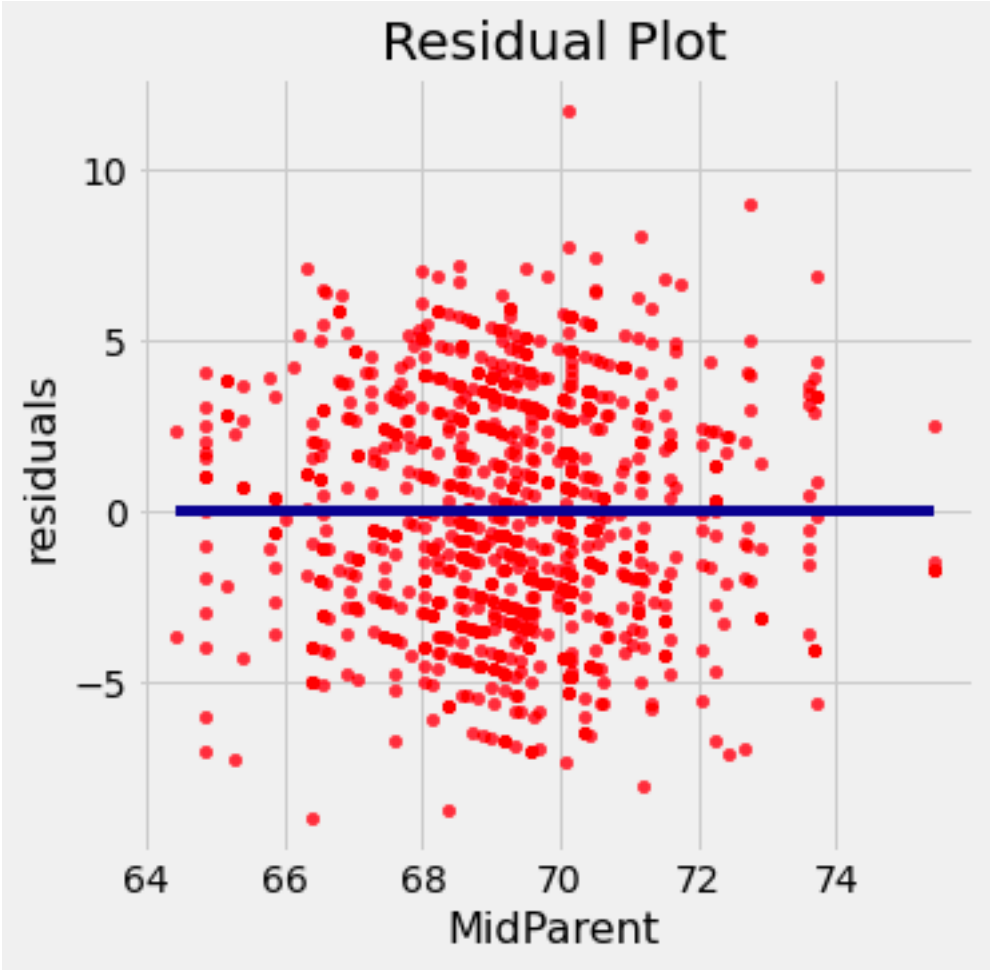
When r = 1, you can predict the value of one variable (the ***response variable***) if you know the value of the other variable (the ***predictor***).

- But how far off are the estimates?
  - How good is the model?

residuals = observed value – regression estimate

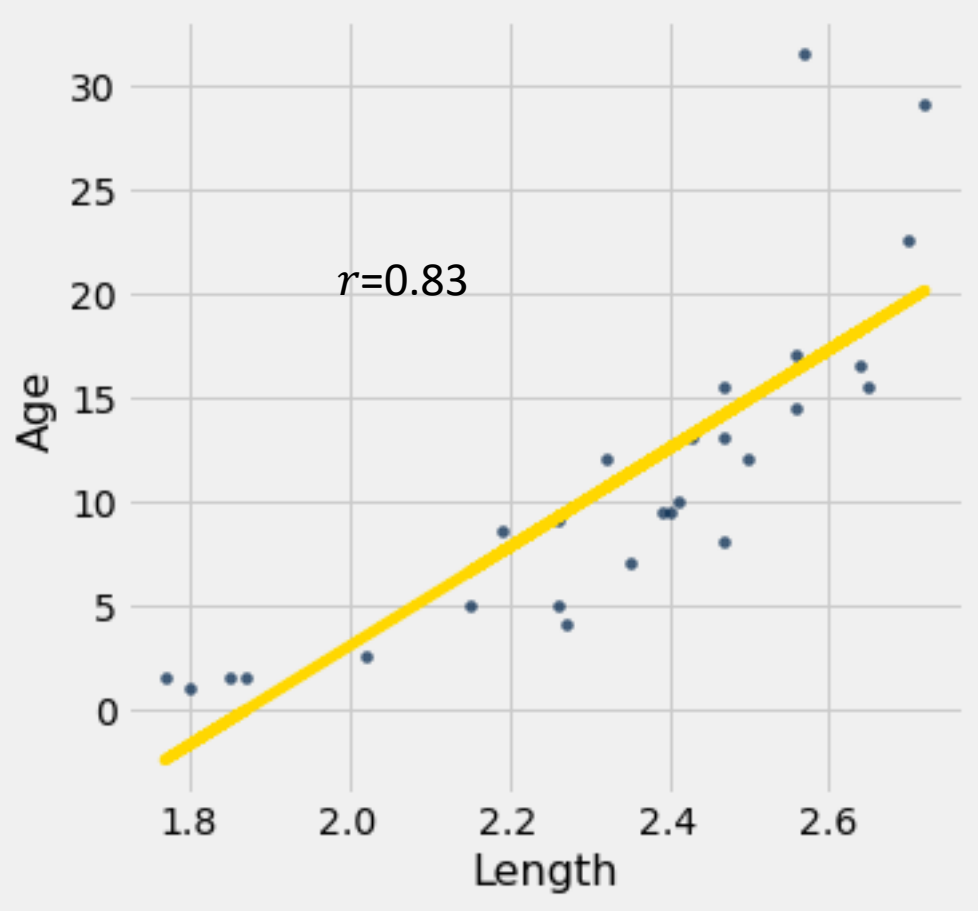
Visual Diagnostics: Linearity





residuals = observed value – regression estimate

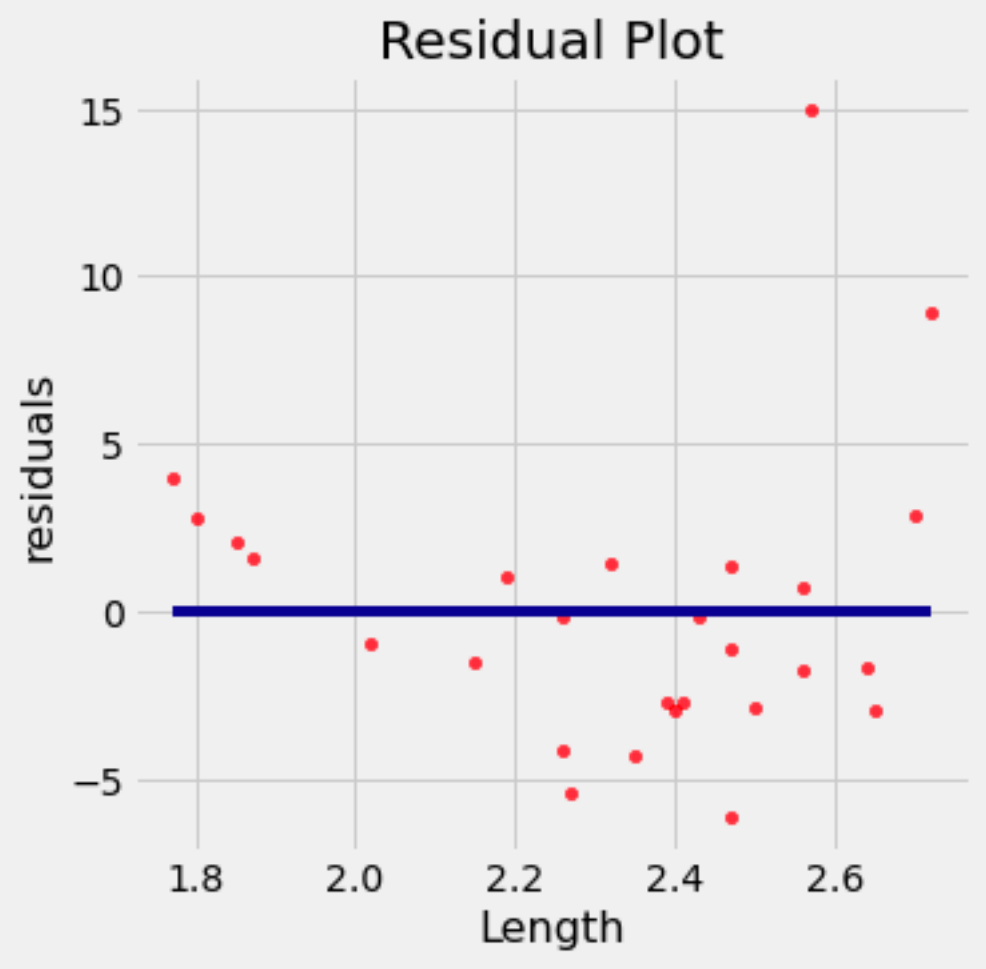
Visual Diagnostics



Dugong (*Dugong dugon*)

If we know the length of a dugong, can we estimate it's age?

Visual Diagnostics: Non-linearity



Dugong (*Dugong dugon*)

This residual plot shows a pattern.
First the predicted weights are too high.
Then they are too low.
Then they are too high again.


end