YData: Introduction to Data Science

Lecture 30: regression

Overview

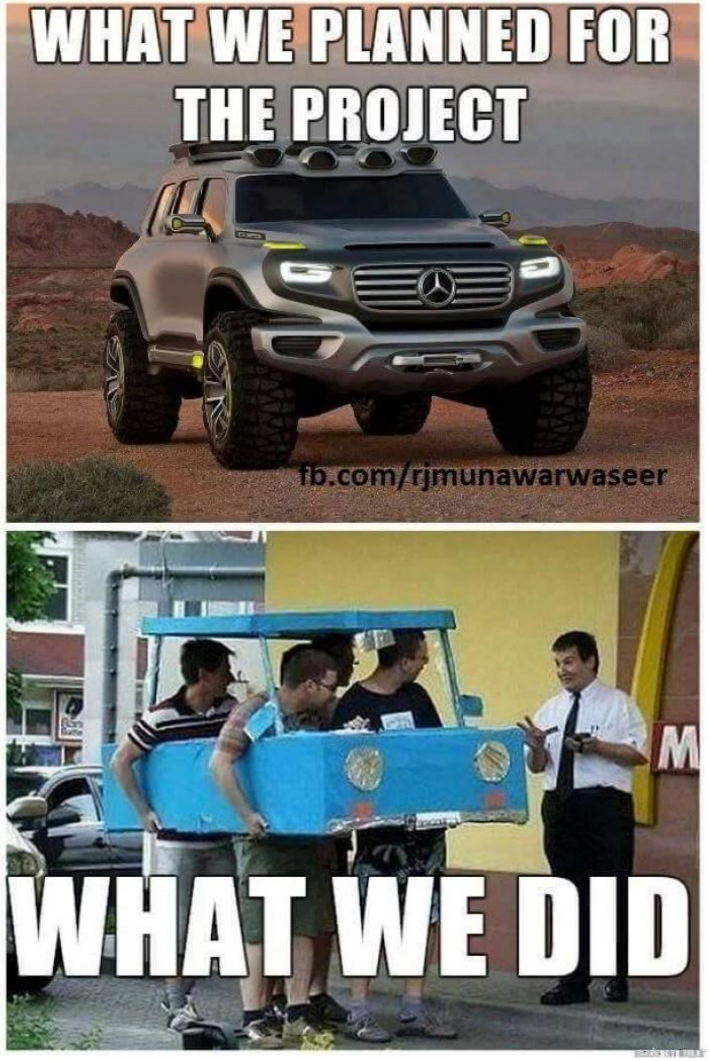
Correlation

- Predictions
- Associations
- The correlation coefficient
- Correlation cautions

Linear regression

- Linear predictions
- Relationship to the correlation coefficient

Announcements



Homework 9 has been posted

- It is due on Sunday the 17th

Project 3 dates have been slightly delayed

- It will be posted on Wednesday
- It is due Wednesday the 27th
  - Rather than on Friday the 22nd

Prediction

Guess the future

Predictions are based on incomplete information

One way 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

What examples of predictions have we seen in this class already?

- Class 9…
- Galton, predicting children's heights based on their parents' heights

Let's explore this in Jupyter!

Association

Two numerical variables

When we have two quantitative variables, we can explore trends in our data that are useful for making predictions

- Usual to visualize trends, and then to quantify them

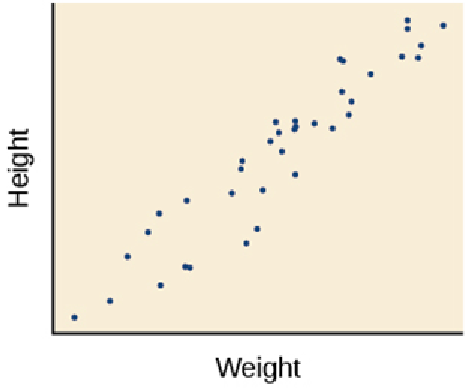
Trend

- Positive association
- Negative association

Pattern

- Any discernible "shape" in the scatter
- Linear
- Non-linear

Let's explore this in Jupyter!



Correlation coefficient

The correlation coefficient

The **correlation** is measure of the strength and direction of a linear association between two variables

- The statistic is denoted with the symbol r
- The parameter is denoted with the symbol ρ (rho)

Based on standard units

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

It is always between -1 and 1:

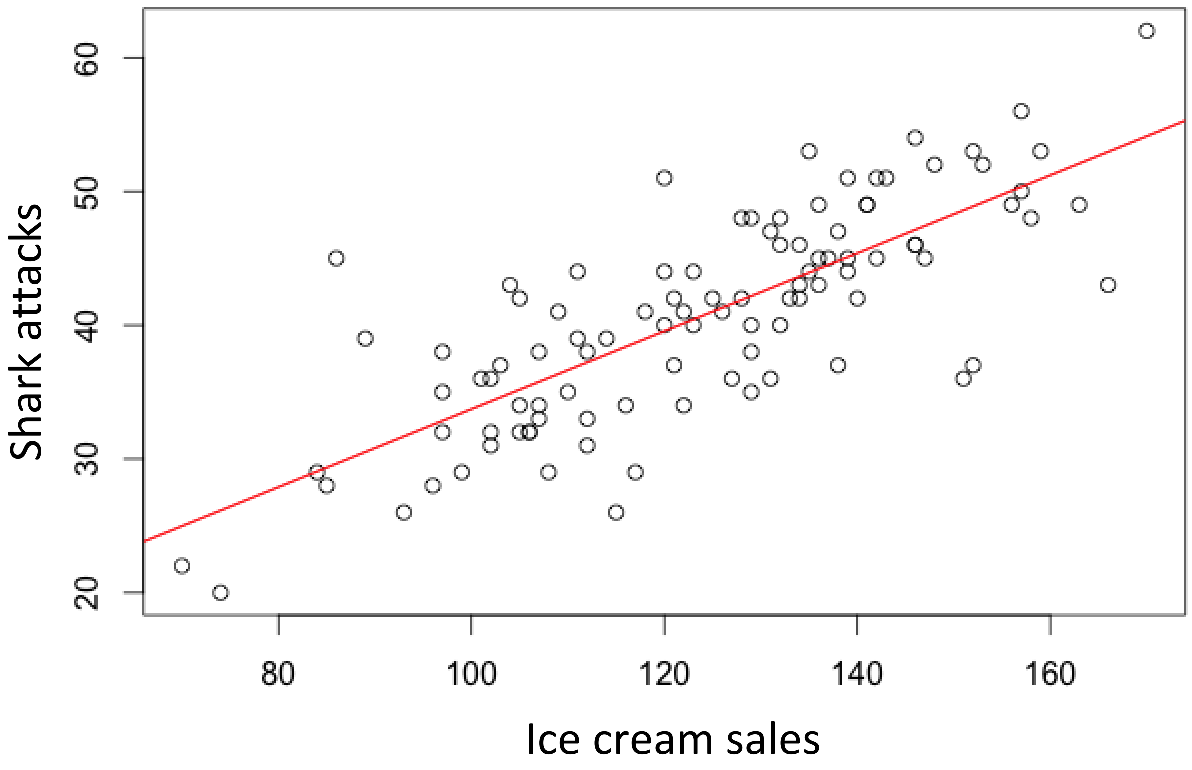
- r = 1: scatter is perfect straight line sloping up
- r = -1: scatter is perfect straight line sloping down
- r = 0: No linear association; uncorrelated

Let's explore this in Jupyter!

Correlation cautions

Correlation caution #1

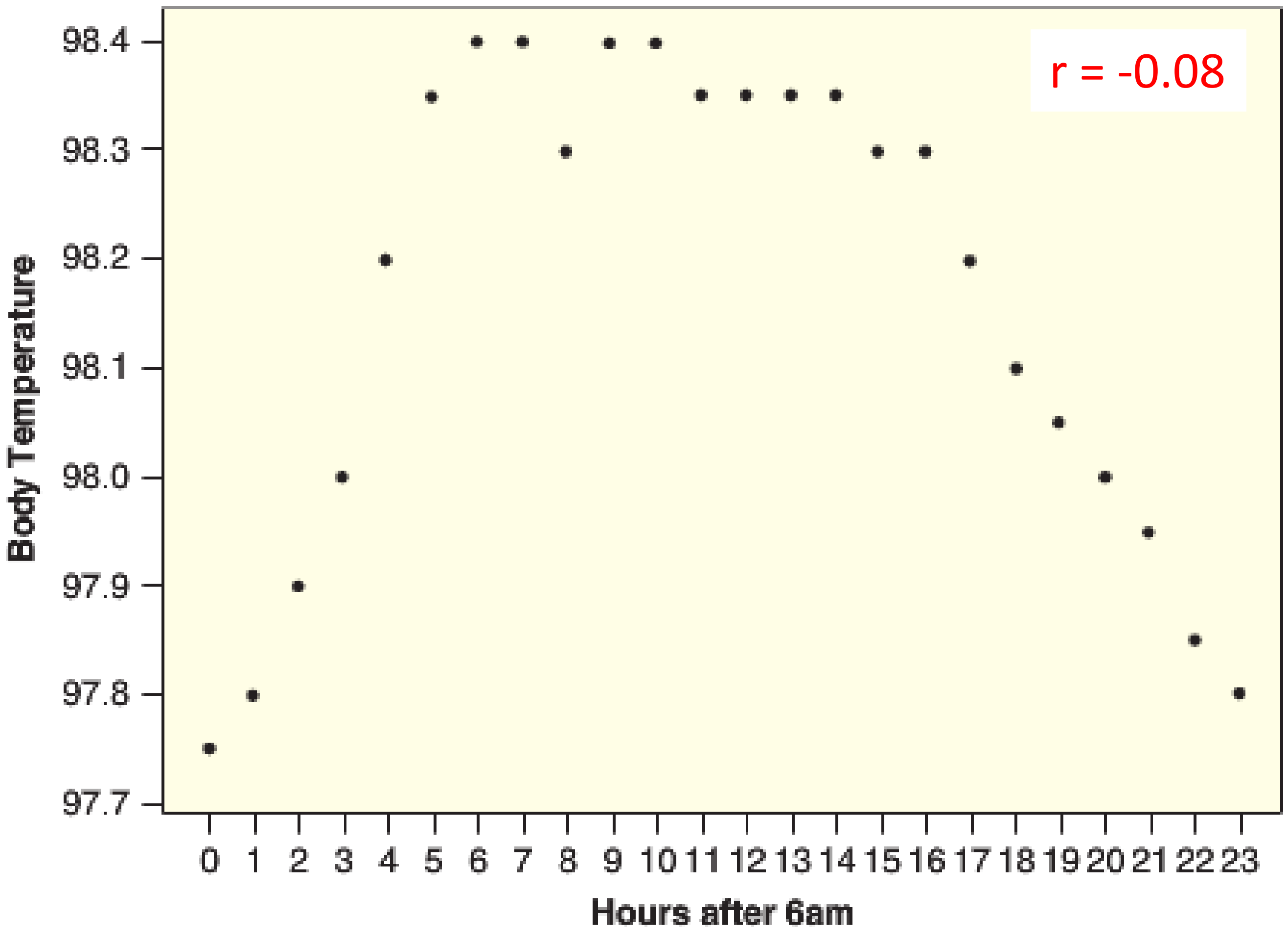
A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between two variables



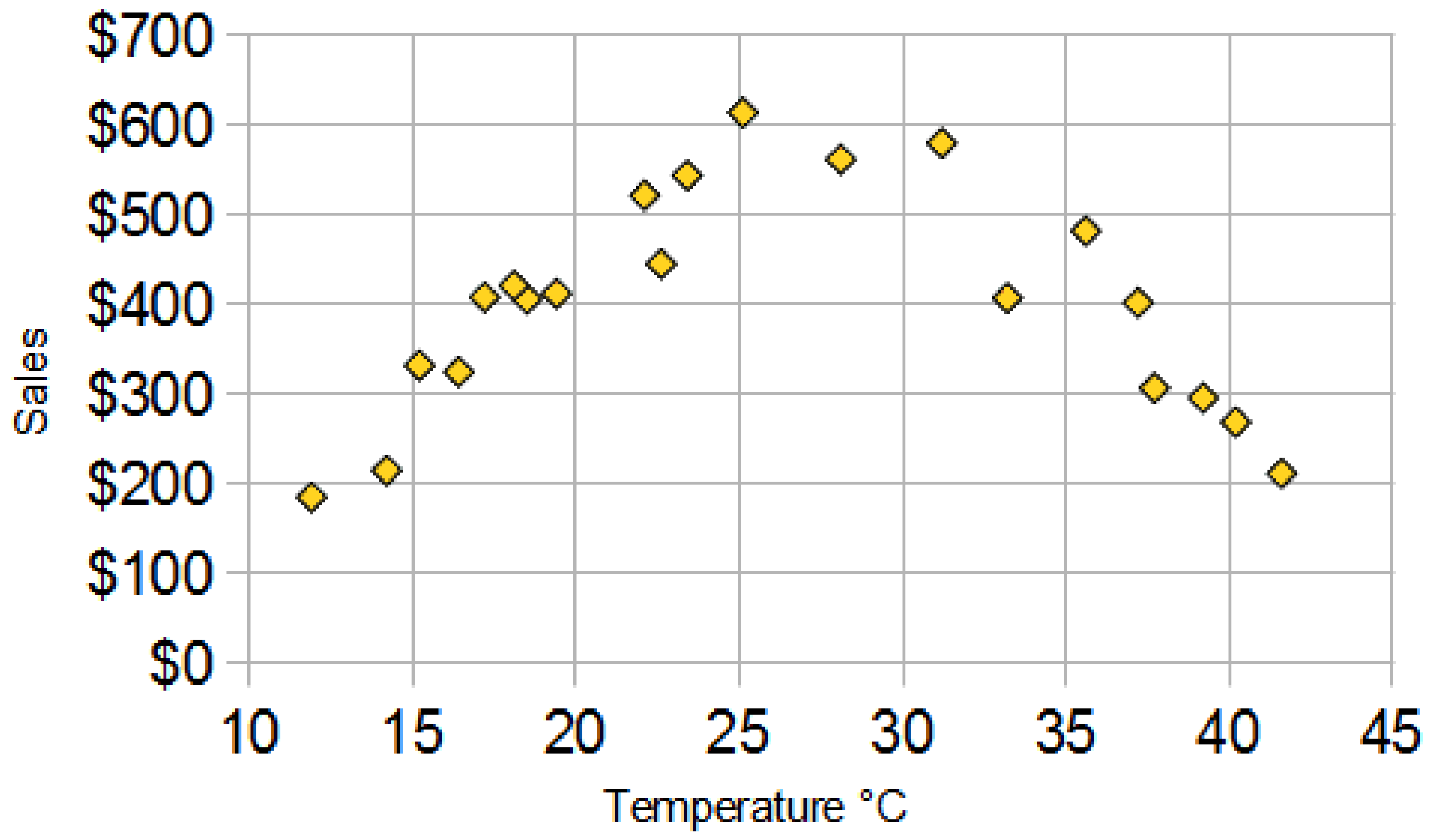
Correlation caution #2

A correlation near zero does not (necessarily) mean that two variables are not associated. Correlation only measures the strength of a <u>linear</u> relationship.

Body temperature as a function of time of the day

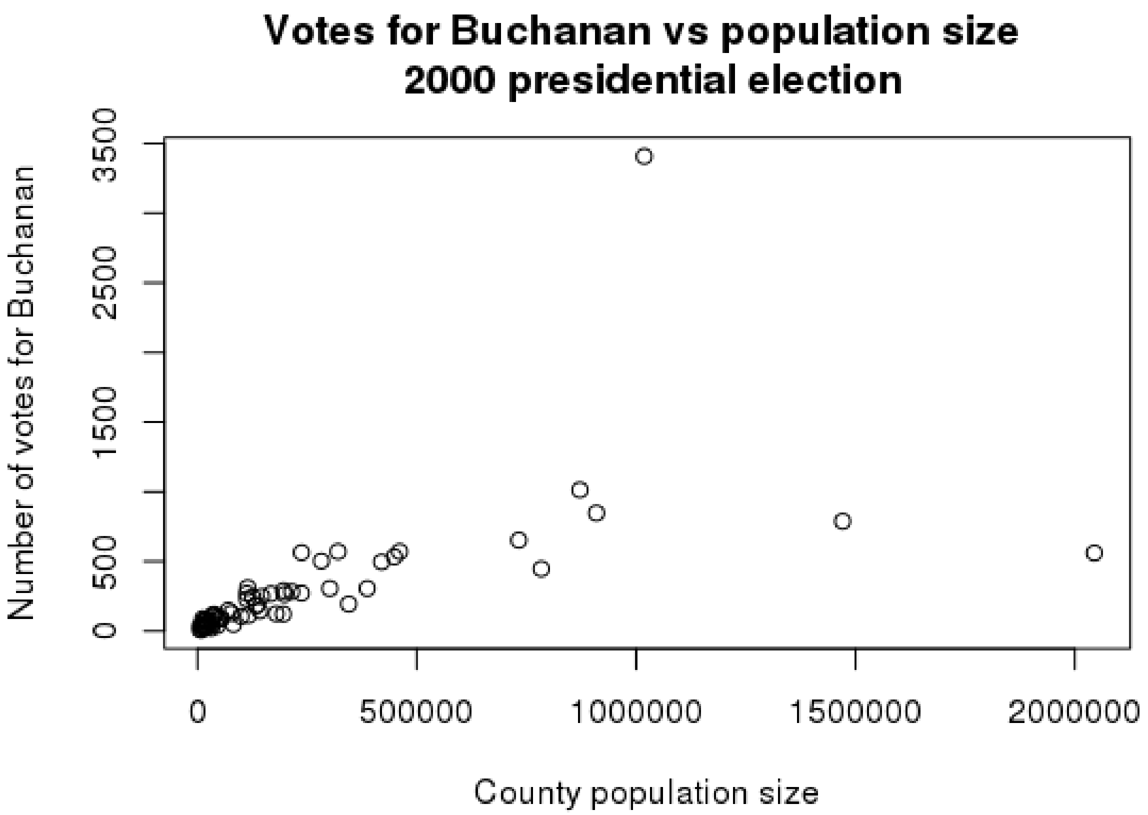


Ice cream sales and temperature



Correlation caution #3

Correlation can be heavily influenced by outliers. Always plot your data!



With Palm Beach
r = 0.61

Without Palm Beach
r = .78

Let's explore this in Jupyter!

Linear regression

Regression

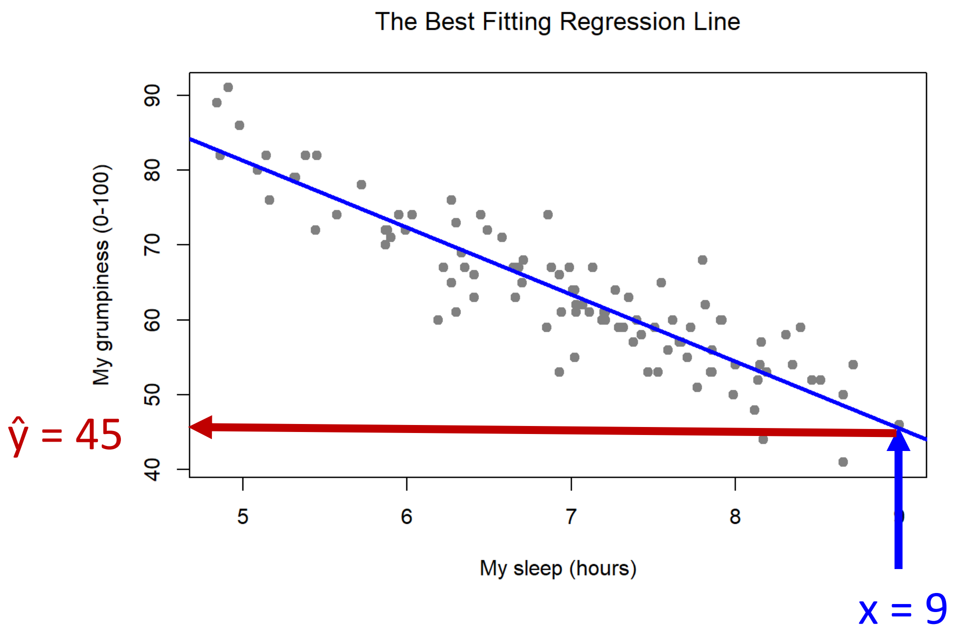
Regression is method of using one variable ***x*** ***<u>to predict</u>*** the value of a second variable ***y***

- i.e., ŷ = f(x)

In **linear regression** we fit a <u>line</u> to the data, called the **regression line**

Lines can be expressed by a slope and intercept:

$$\hat{y} = \text{slope} \cdot x + \text{intercept}$$

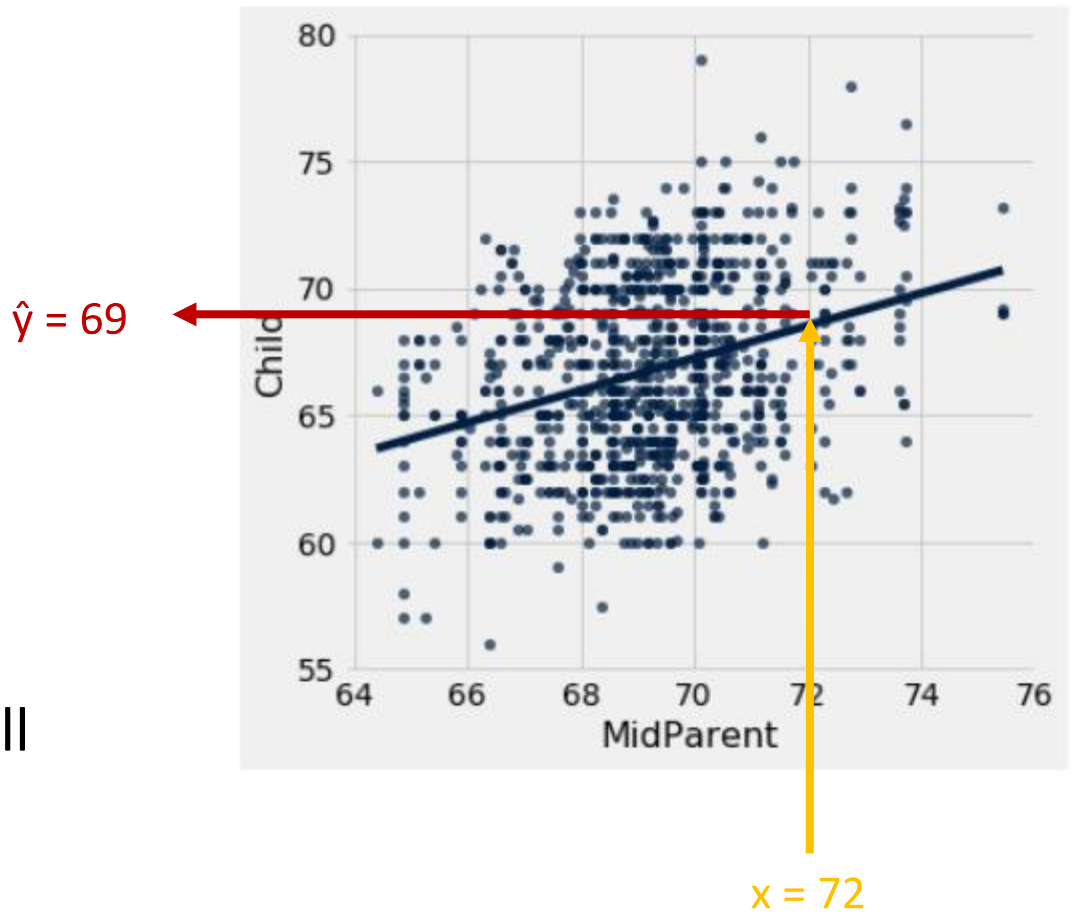


Regression predictions

The regression line predicts an "average" value:

- For a given x value, the average y could be considered the "best" prediction

Example: Take all children whose midparent height is 72 standard unit. The average height of these children is somewhat less than 70 inches



It doesn't say that all of these children will be somewhat less than 70 inches in height. Some will be taller, and some will be shorter.

Let's explore this in Jupyter!

Slope and intercept

Regression with standardized units

Suppose we standardize our x and y variables through a z-score transformation:

- $y_{(SU)} = (y - \bar{y})/SD_y$ where $\bar{y}$ and SD_y are the mean and SD of y
- $x_{(SU)} = (x - \bar{x})/SD_x$ where $\bar{x}$ and SD_x are the mean and SD of x

The we can relate our predictions of these standardized x and y variables to the correlation coefficient r:

Regression Line

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

Regression line

Our equation for the regression line in standardized units is:

Regression Line

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

Expanding the definition of standardized units we have:

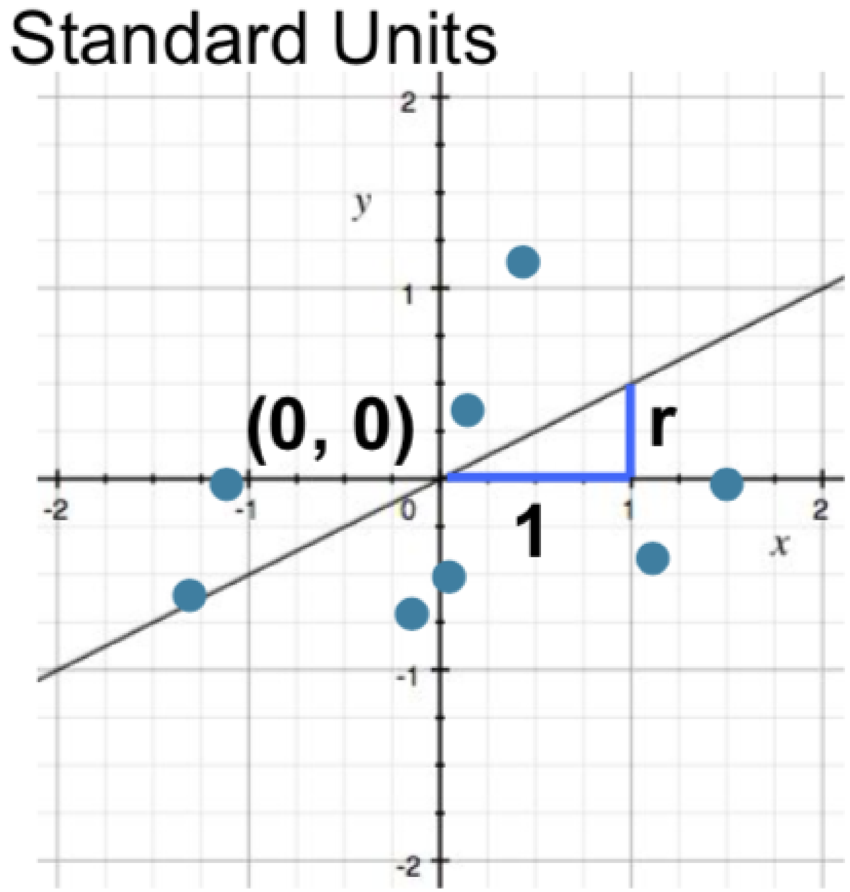
$$(\hat{y} - \bar{y}) / SD_y = r \cdot (x - \bar{x}) / SD_x$$

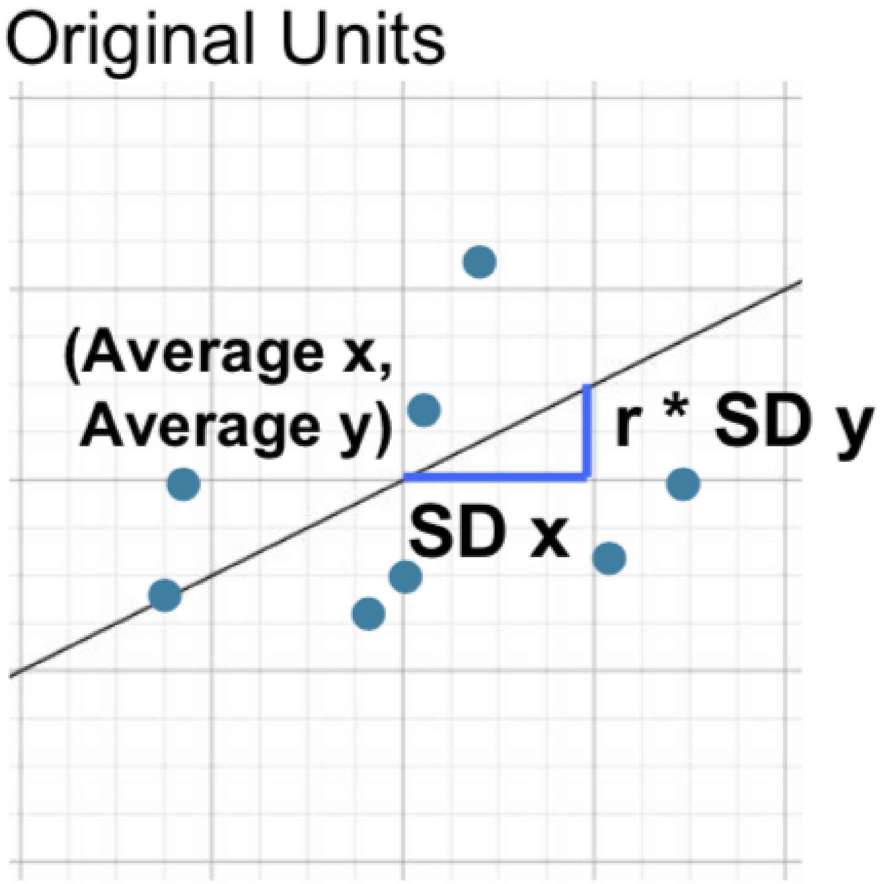
Solving in our original units: $\hat{y} = \text{slope} \cdot x + \text{intercept}$

$$\text{Slope} = r \cdot SD_y / SD_x$$

$$\text{Intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

Regression line





Regression to the mean

Our equation for the regression line in standardized units is:

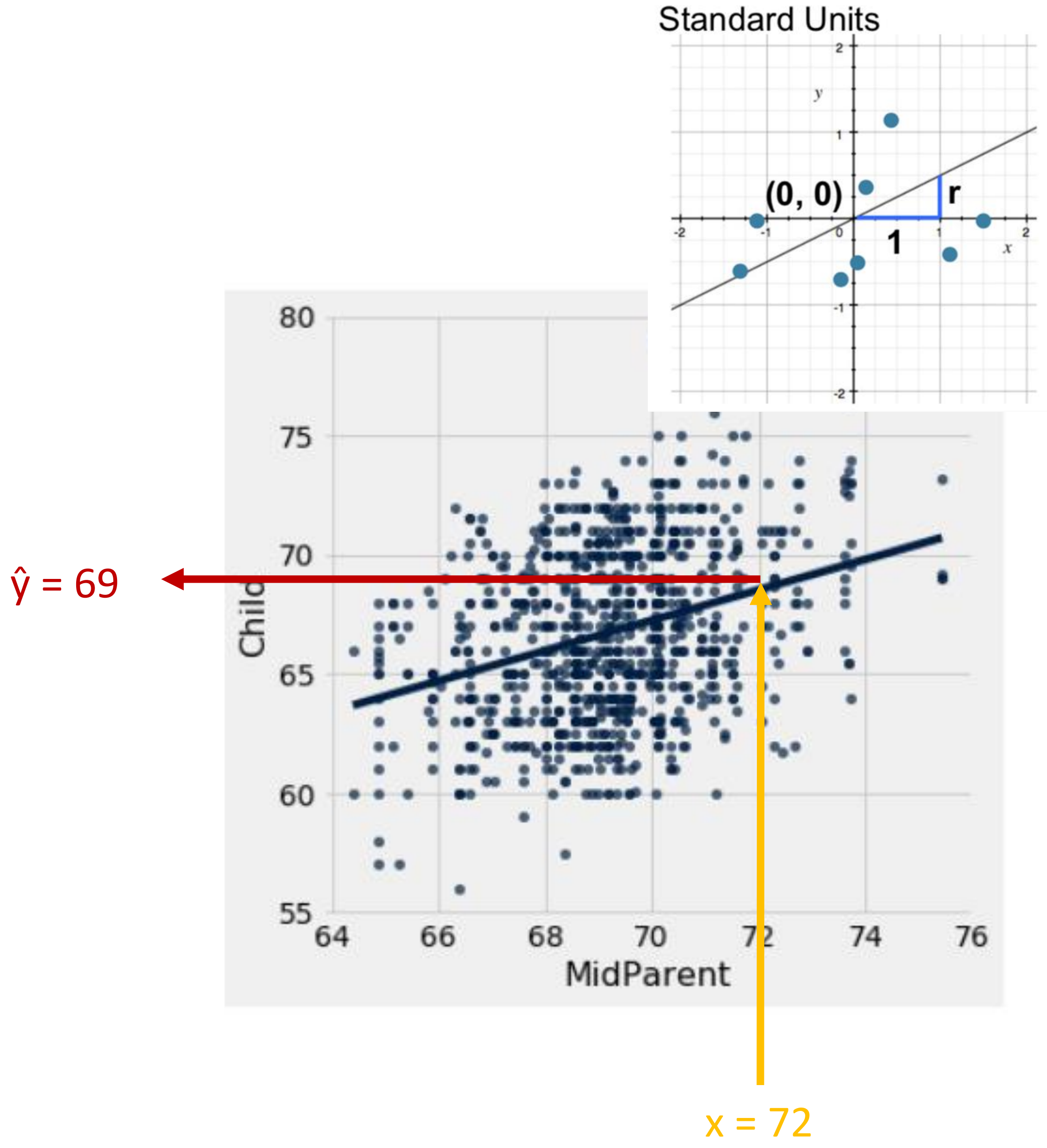
Regression Line

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

Correlation

Because -1 ≤ r ≤ 1 this means that standardized predicted y values will be closer to their mean the than standardized x values used for the prediction

This phenomenon is called "regression to the mean" or "regression to mediocrity"



Let's explore this in Jupyter!