# BC COMS 1016: Intro to Comp Thinking & Data Science

## Lecture 22 – Linear Regression, Least Squares, & Residuals

# Announcements

- Lab 8 – Regression
  - Due Monday 04/18
- Homework 8 - Linear Regression
  - Due Monday 04/18
- Project 2
  - Due Monday 04/18

# Grading – Rubric 1

| | |
|---|---|
| Participation | 5% |
| Weekly HW | 25% |
| Projects | 20% |
| Midterm + daily quizzes | 25% |
| Final Project | 25% |

# Grading – Rubric 2

| | |
|---|---|
| Participation | 10% |
| Weekly HW | 35% |
| Projects | 30% |
| Midterm + daily quizzes | 0% |
| Final Project | 25% |

# Announcements

- Project 3
  - Optional, if electing to rubric 2

# Prediction

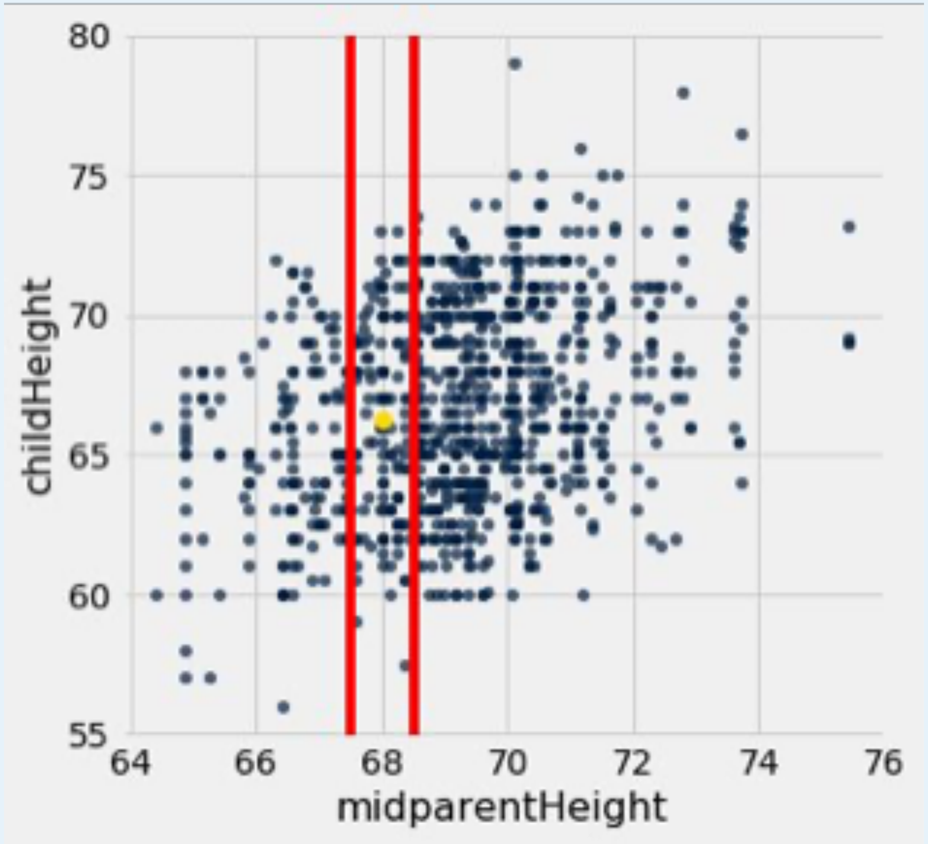
# Correlation

# 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.

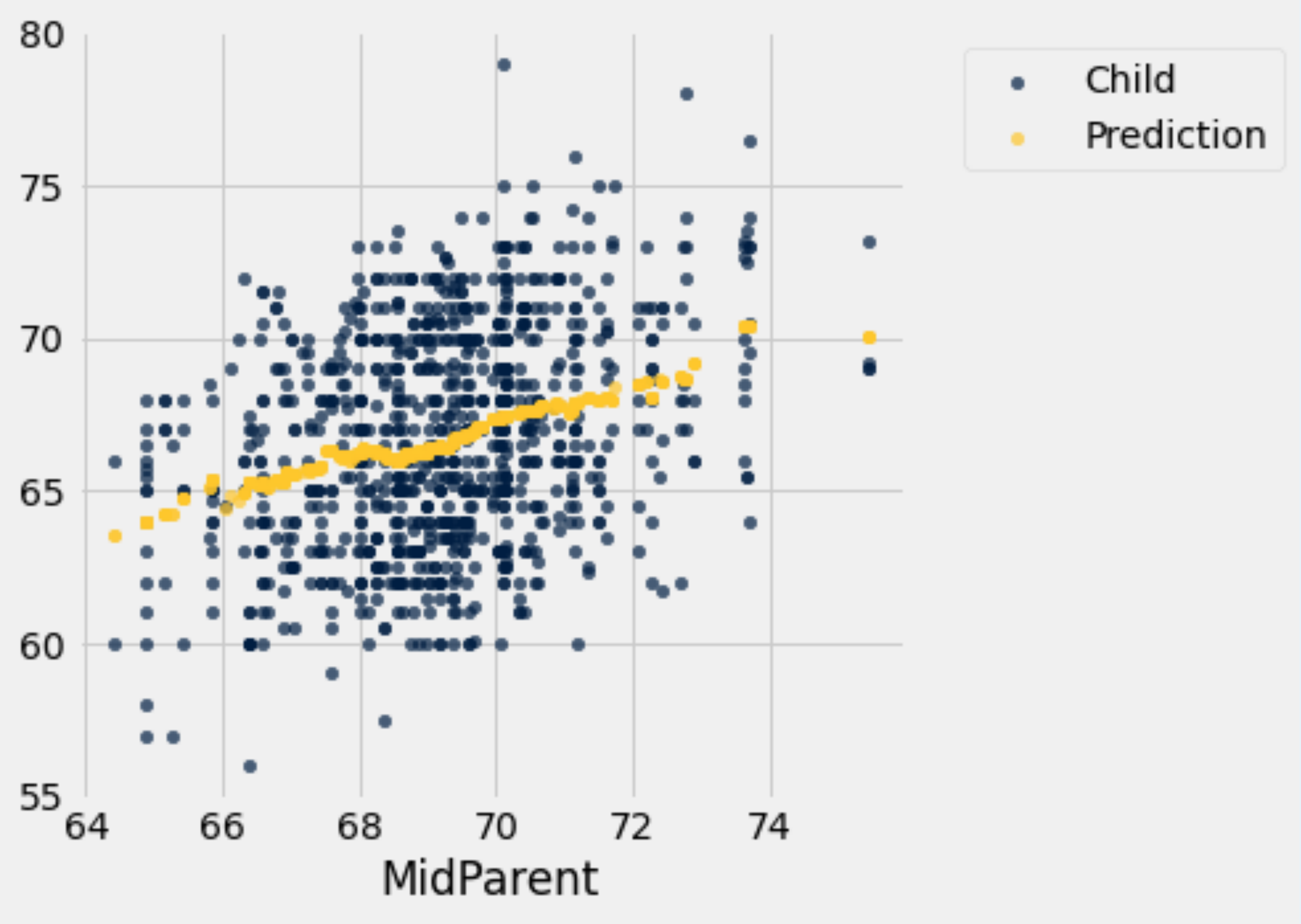
# Galton's Heights




**Goal:** Predict the height of a new child, based on that child's parents' heights. predict a child's height.

**Idea:** Use the average height of the children of all families where the midparent Height is close to the child's parents

# 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

# 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:

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

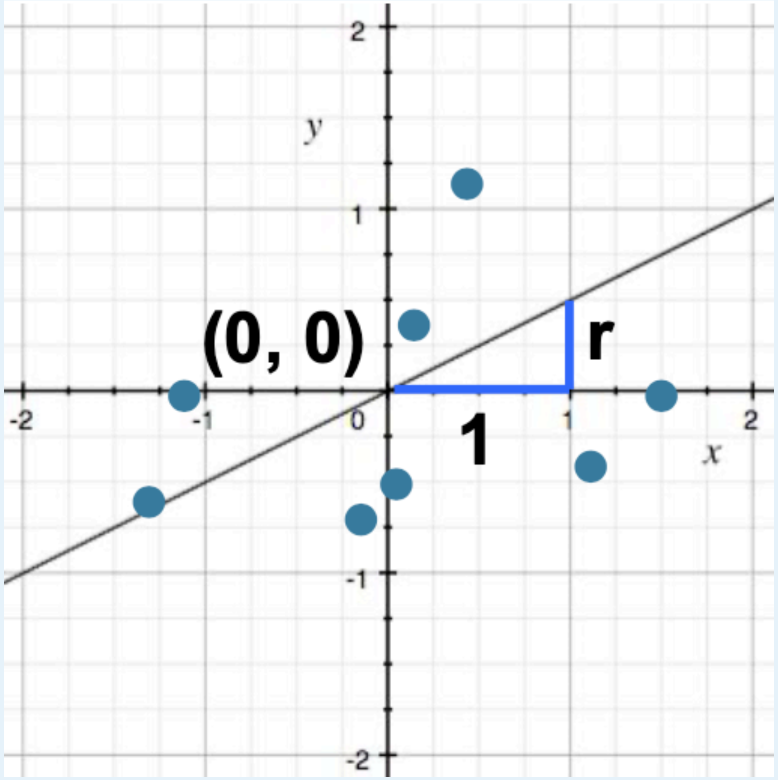
$$\frac{estimate\ of\ y\ - mean(y)}{SD\ of\ y} = r \times \frac{given\ x\ - mean(x)}{SD\ of\ x}$$

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

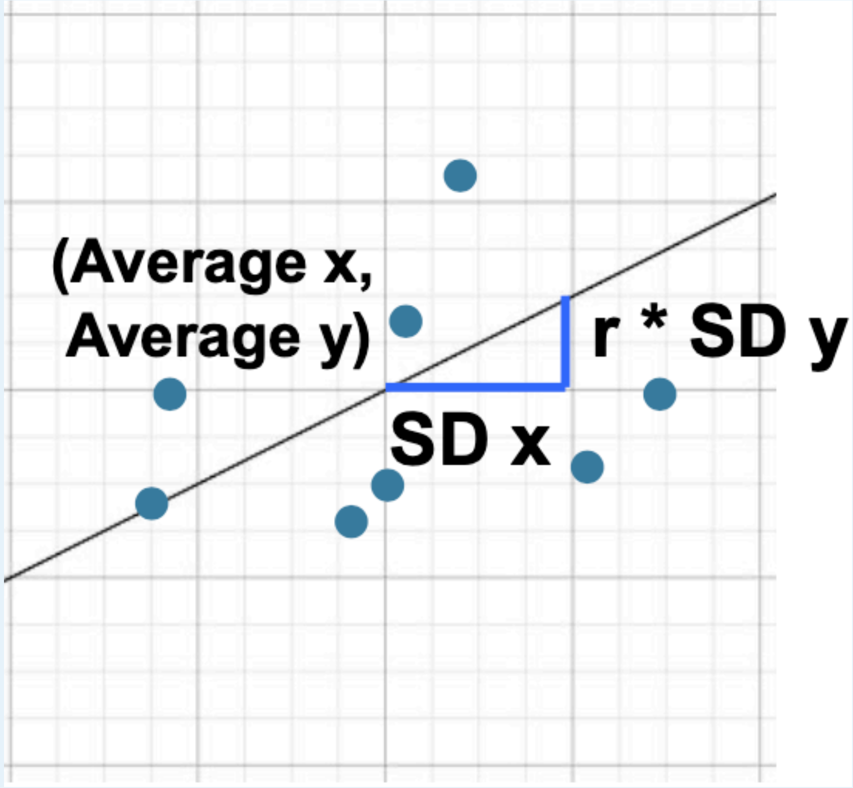
$$y = slope \times x + intercept$$

# Regression Line

## Standard Units




## Original Unites



# Slope and Intercept

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

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

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

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

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

# 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

# Prediction with Linear Regression

**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 oy *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:

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

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

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

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

What we want

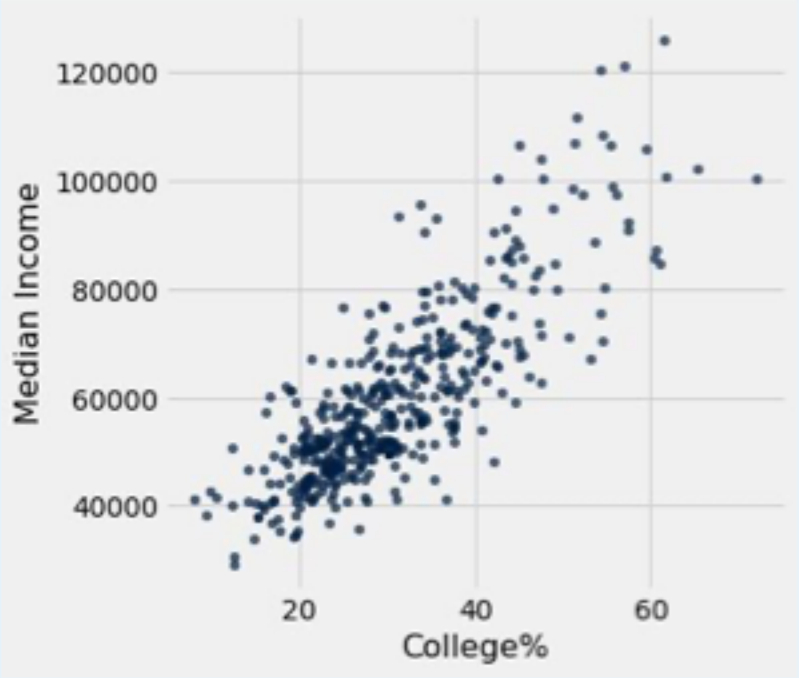
What we observe

# Discussion Question

Based only on the graph, which must be true?

1. Going to college causes people to earn more.
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



# Least Squares

# Error in Estimation

- **error = actual value – estimate**
- Typically, some errors are positive and some are 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)

# Least Squares Line

- Minimized the root mean squared error among all lines

- Equivalently, minimizes the mean squared error 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 (demo)
- If the function **mse(a, b)** returns the mse of estimation using the line “estimate = ax + b”,
  - then **minimize(mse)**returns array [a0, b0]
  - a0 is the slope and b0 the intercept of the line that *minimizes* the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

# Residuals

# Residiuals

- 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 line

# Regression DIagnostics

# Residual Plot

A scatter diagram of residuals

- For linear relations, plotted residuals should look like an unassociated blob
- For non-linear relations, the plot will show patterns
- Used to check whether linear regression is appropriate
- Look for curves, trends, changes in spread, outliers, or any other patterns

# Properties of residuals

- The mean of residuals is always 0
- Variance is standard deviation squared
- (Variance of residuals) / (Variance of y) = 1 – $r^2$
- (Variance of fitted values) / (Variance of y) = $r^2$
- Variance of y =
  (Variance of fitted values) + (Variance of residuals)