COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 19: Least Squares and Residuals

Nov 24, 2025

The rest of the semester

- **Today**: Least Squares and Residuals ← HW 7 due
- ***Wednesday, Nov 26****: Holiday!*

No office hours this week

- **Monday, Dec 1**: Regression Inference ← HW 8 due
- **Wednesday, Dec 3**: Computing Fellows Workshop
  - Final Project Consultations during Lab

Progress Report
due Tuesday, Dec 2

- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due
- **Friday, Dec 12**: Final Projects Due

Last day of class :(

Progress Report

- Deadline has been moved to **Tuesday, Dec 2** (rather than Monday)

- Progress report we are checking both the written report AND your notebooks

  - You do *not* need to consolidate into a single notebook

  - *We will be expecting your code to be readable!*

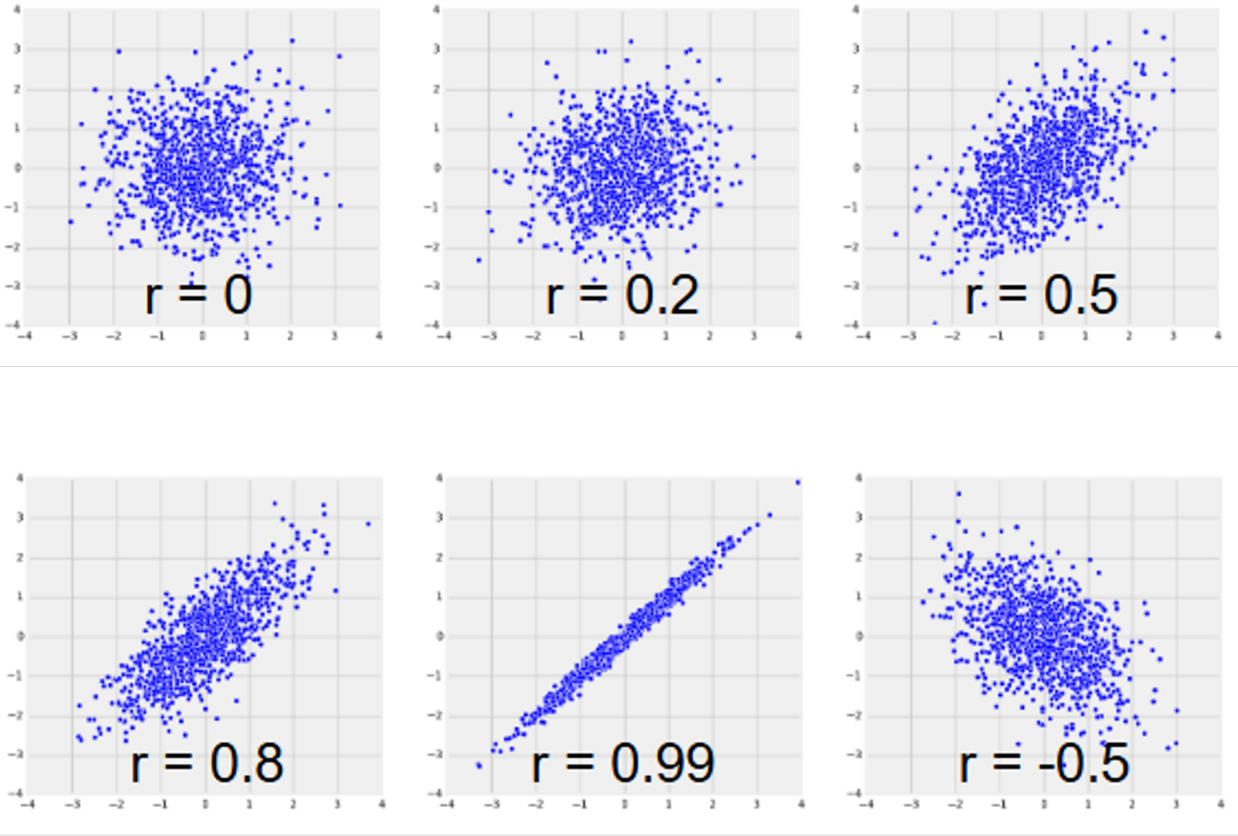
Progress Report

- You should be roughly 60% complete
    - Exploratory data analysis should be completed and should be started on other parts of the report
- List what remains and how you plan to approach them
    - Include anything you're stuck on or need assistance with
    - We will not be deducting points for conceptual errors
- **Please bold your hypothesis statements (null and alternative)**

Correlation & Regression Line Recap

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*)



Computing r

The **correlation coefficient** r is the **average product** of x in standard units and y in standard units. To compute:

- First convert our values in x & y to standard units

$$\vec{x}_{\text{su}} = \frac{\vec{x} - x_{\text{avg}}}{\text{SD}_x} \qquad \vec{y}_{\text{su}} = \frac{\vec{y} - y_{\text{avg}}}{\text{SD}_y}$$

- Then compute r as the average product

$$r = \text{avg}\left(\vec{y}_{\text{su}} \times \vec{x}_{\text{su}}\right)$$

Linear Regression Line

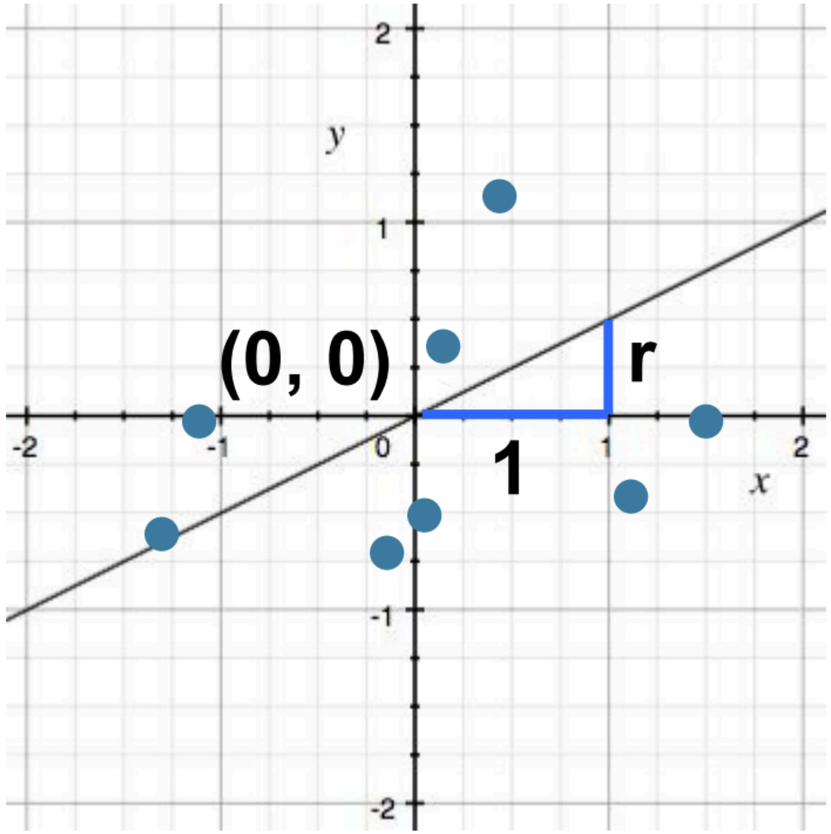
The correlation coefficient r can be used to plot the straight line that the points are clustered around:

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

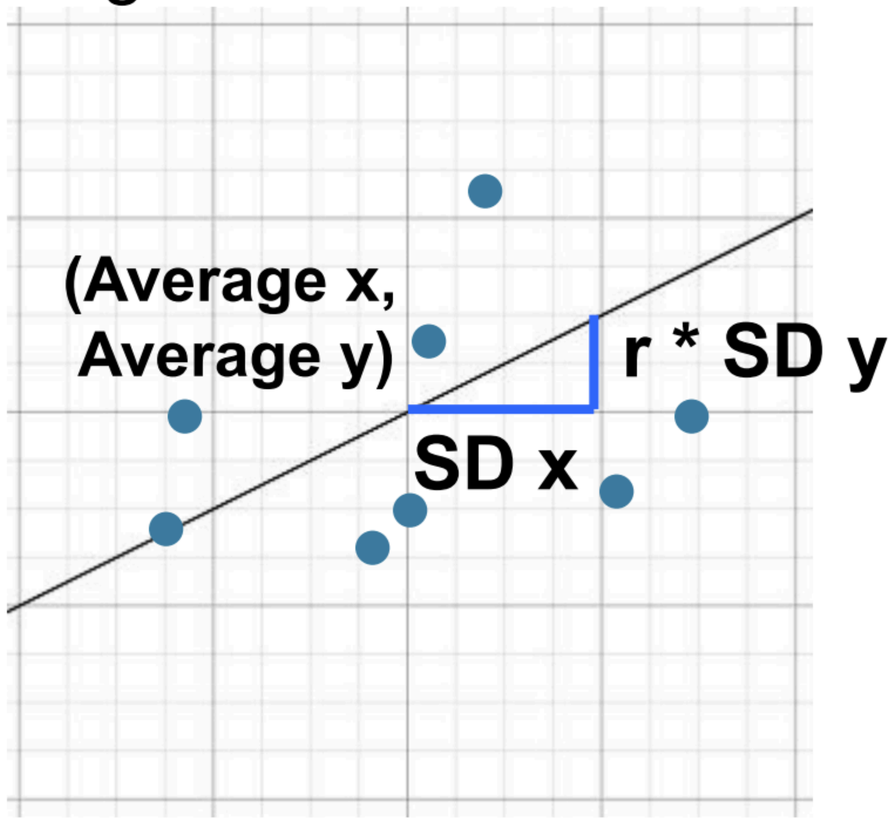
This is the **linear regression line**

Regression Line: Original Units

Standard Units



Original Units



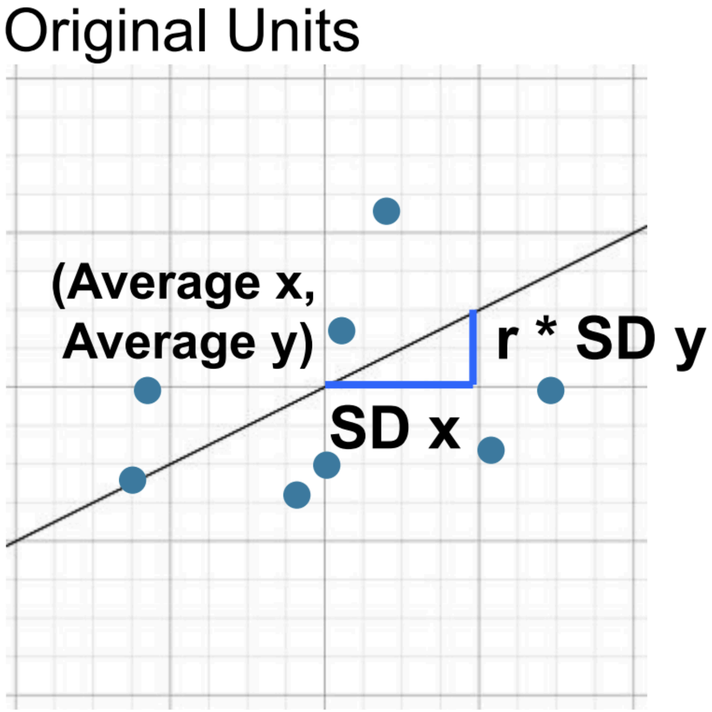
Regression Line: Converting to Original Units

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

$$y_{su} = \frac{y - \text{avg}(y)}{\text{SD of } y} \qquad x_{su} = \frac{x - \text{avg}(x)}{\text{SD of } x}$$

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

Regression Line: Converting to Original Units



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

$$\text{slope} = r \times \frac{\text{SD of } y}{\text{SD of } x}$$

$$\text{intercept} = \text{avg}(y) - \text{slope} \times \text{avg}(x)$$

Prediction

Goal: Predict y given x

$$z = \frac{v - \mu}{\text{SD}}$$

To find the regression estimate of y:

1. Convert the given x to standard units
2. Multiply by r to get y in standard units
3. Convert y in standard units back to original units of y

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

Grade Example

A course has a **midterm** (average: 70, standard deviation: 10) and a hard **final exam** (average: 50, standard deviation: 12).

We create a linear regression line to predict what a final exam score would be for a given midterm score.

In this case:

1. What is our y (i.e., what do we want to predict)?

2. What is our x (i.e., we want to predict y given x)?

Grade Example

A course has a **midterm** (average: 70, standard deviation: 10) and a hard **final exam** (average: 50, standard deviation: 12).

We create a linear regression line to predict what a final exam score would be for a given midterm score and compute the correlation coefficient r=0.75.

3. What do you expect the average final exam score to be for students who scored a 90 on the midterm?

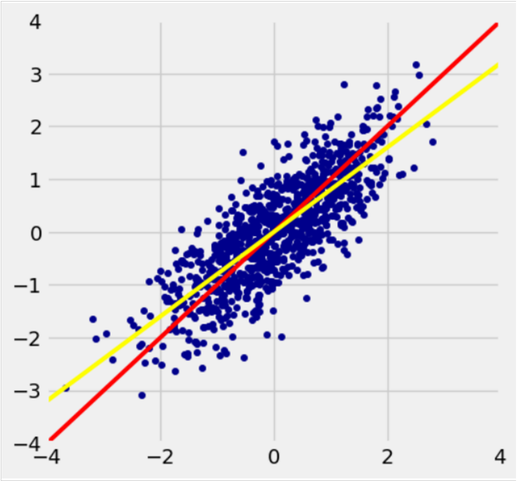
$$z = \frac{v - \mu}{\text{SD}}$$

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

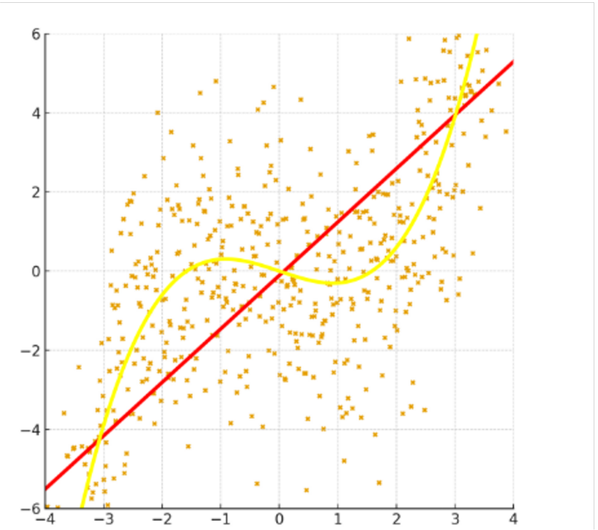
Least Squares and Residuals

Least Squares and Residuals

How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?

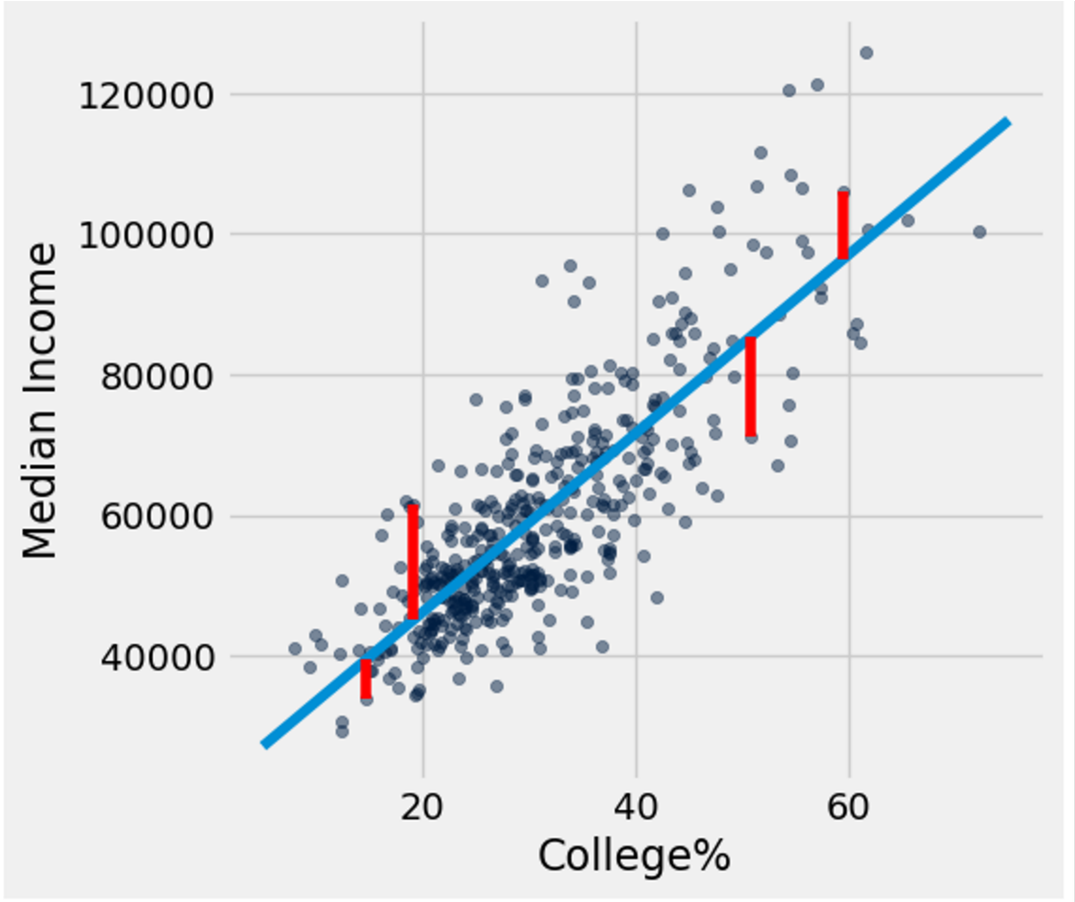


How can we check whether a line is appropriate (versus a non-linear model)?

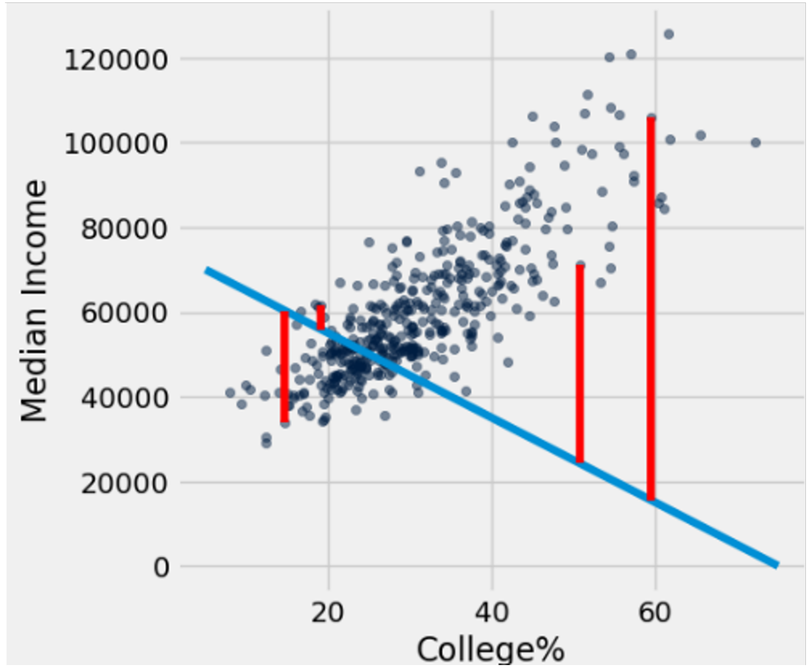


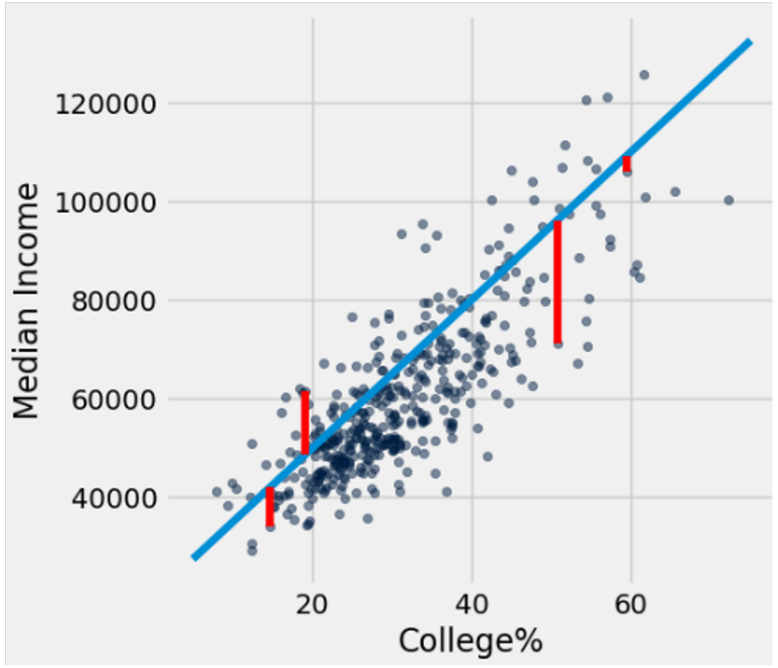
Error in Estimation

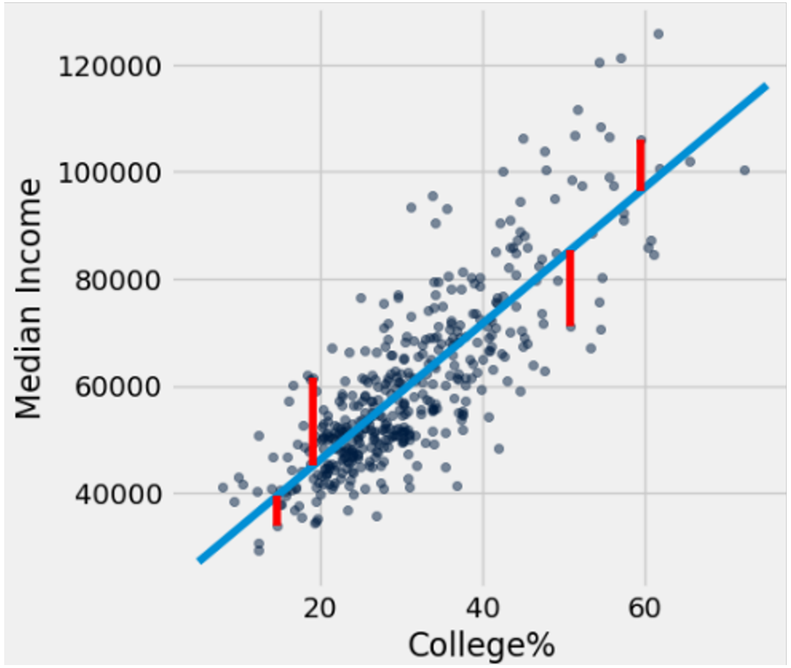
error = actual value - estimate



Different Prediction Lines







Which one is the best and why?

Root Mean Square Error (RMSE)

Process to calculate the size of the error:

1. Compute the errors between the regression line and actual value and square them

2. Compute the mean of the squared errors

3. Compute the square root

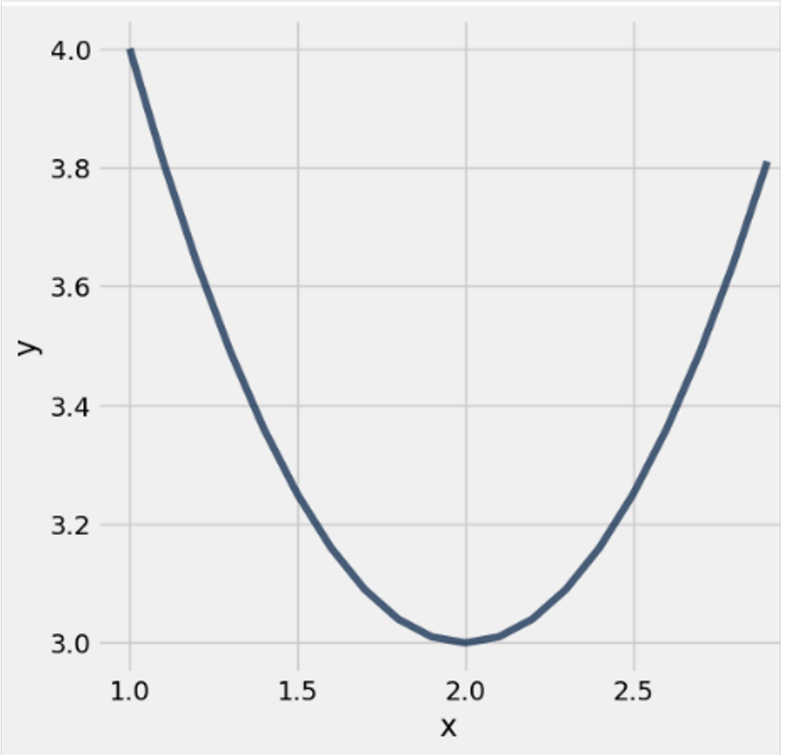
This gives us the **root mean square error** (**rmse**)

Least Squares Line

- **Minimizes** the **root mean squared error** (**rmse**) among all possible lines
  - Equivalently, minimizes the **mean square error** (**mse**) among all lines
- Other names for this line include:
  - “Best fit” line
  - Least squares line
  - Regression line

How to find the minimum?

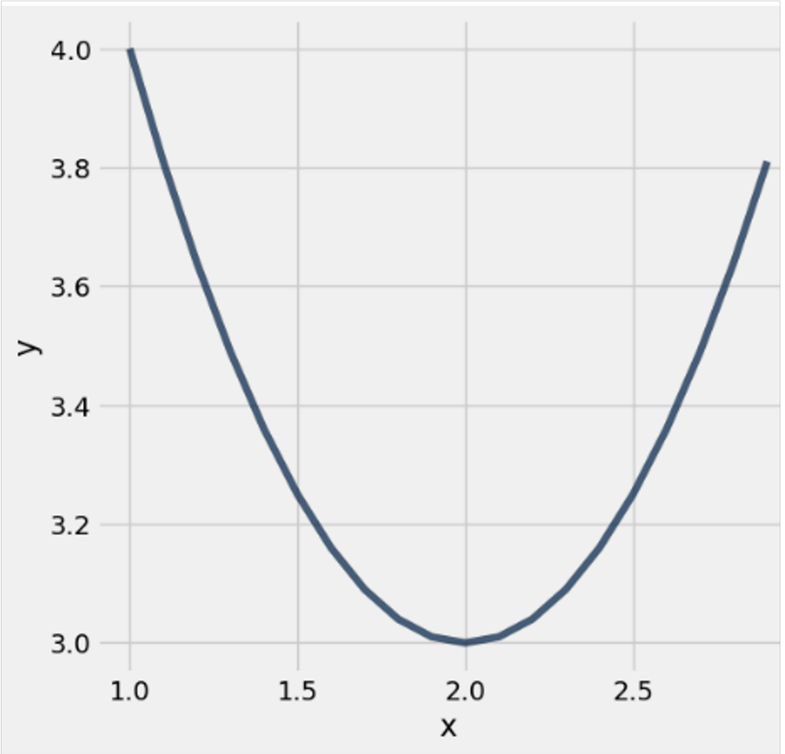
Goal: $\min_x f(x)$



$$f(x) = (x-2)^2 + 3$$

How to find the minimum?

- Derivatives?
  - Find the x where the derivative is closest to zero
  - Only works for simple functions
- Brute force?
  - Evaluate at 1, 2, 3, …
  - This might take a while… How many decimal places should we check?



$$f(x) = (x - 2)^2 + 3$$

Numerical Optimization

- This is a generally difficult task and outside the scope of this class!
- We will use the `datascience` function **`minimize`**
  - **`minimize(f)`** will output **arguments** to a function `f` that minimize the output of `f`

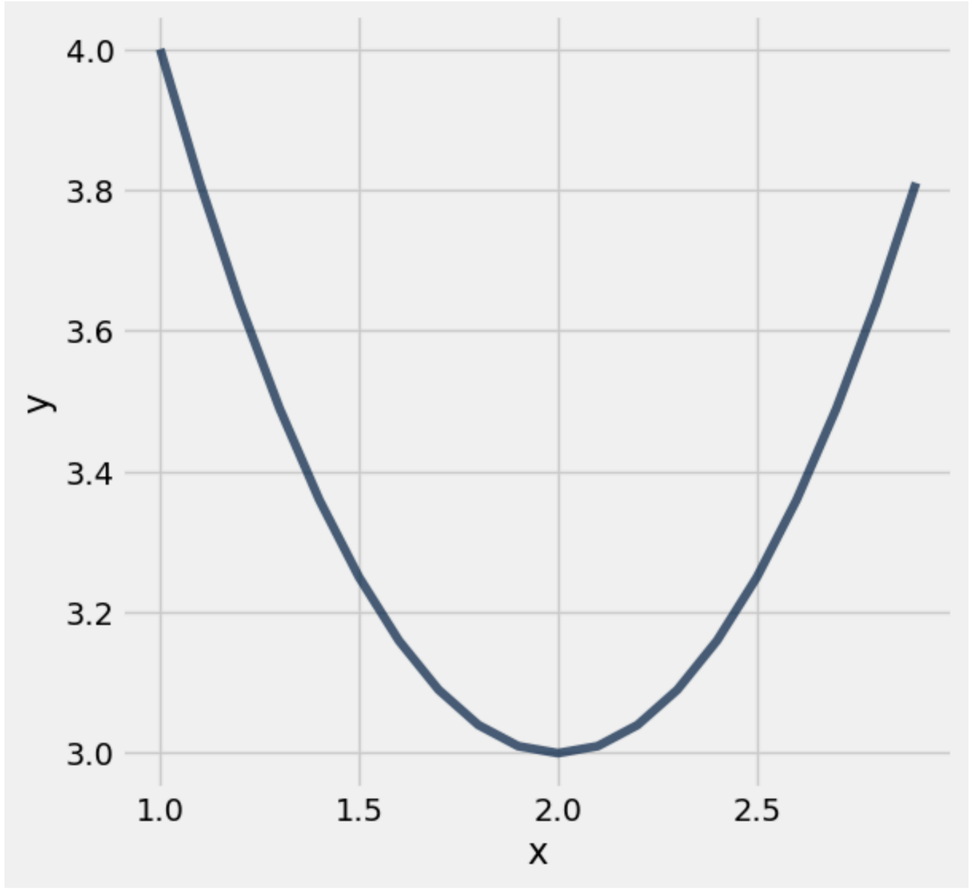
`minimize` Examples

`minimize(f)` will output **arguments** to `f` that minimize the output of `f`

- Suppose we have a function for computing $(x-2)^2+3$:

```
def f(x):
    return ((x-2)**2) + 3
```

- `minimize(f)` outputs 1.9999



`minimize` Examples

- We can also use it for functions with multiple inputs:

```
def f2(x1, x2):
    return 2 * np.sin(x2*np.pi) + x1 ** 3 + x1 ** 4
```

- `minimize(f2)` outputs an array corresponding to `x1` and `x2` that minimize the value of `f2`

```
minimize(f2)
```

```
array([-0.75000601, -0.5       ])
```

Using `minimize` with mse to minimize errors

Suppose we have a dataset with x, y pairs.

If we define a function **`mse(a,b)`** to compute the mean square error of

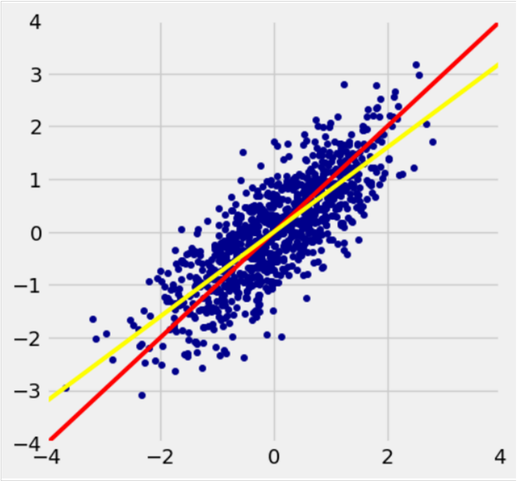
$$\text{estimate of } y = a \times x + b$$

then `minimize(mse)` returns an array $[a_0, b_0]$

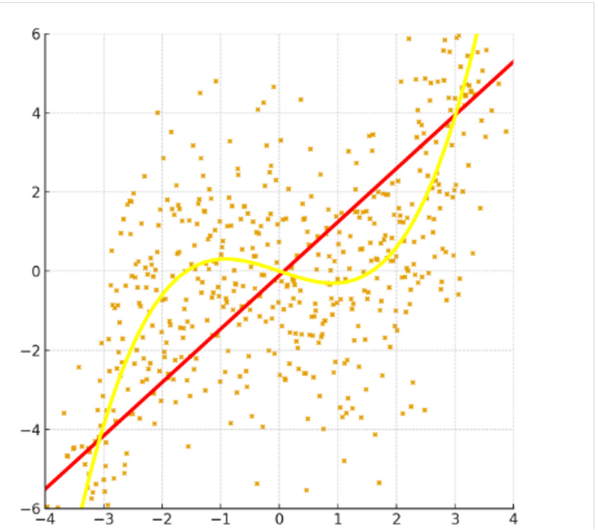
where a_0 is the **slope** and b_0 is the **intercept** that minimizes mse

Least Squares and Residuals

How can we know we've created the best line to fit through our data (i.e., that we've minimized error)?



How can we check whether a line is appropriate (versus a non-linear model)?



Residuals

Residuals

Residual: The error for *individual* regression estimates

$$\text{residual} = \text{observed } y - \text{regression estimate of } y$$

- Can calculate the residual for each individual (x, y) point
- It's the **vertical distance** between the point and the line of best fit

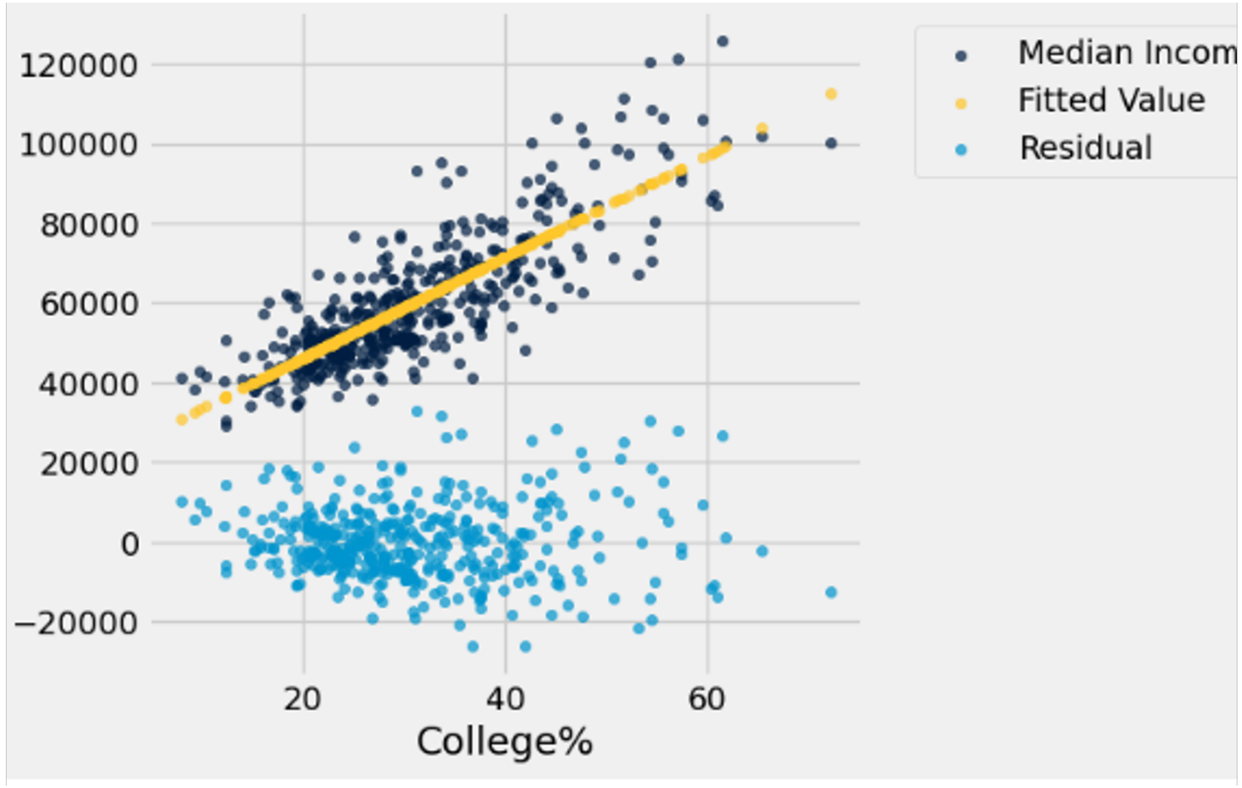
Residual Plots

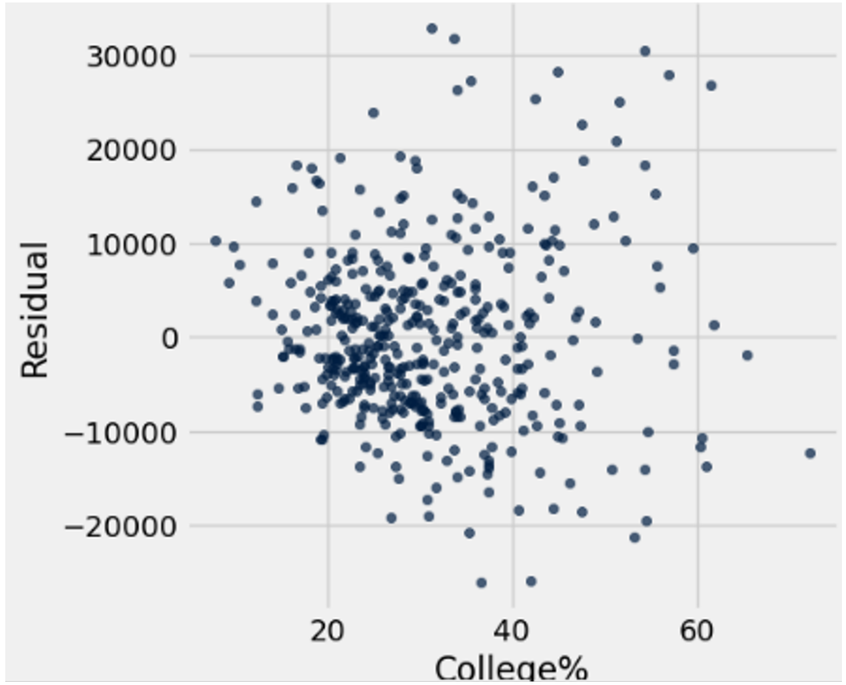
Plots of residuals can be a diagnostic for whether a linear model is appropriate (versus when a non-linear model might be better)

Scatter diagrams of residuals:

- Should look like unassociated blobs for linear relationships
- Will show patterns for non-linear relationships
- Look for curves, trends, changes in spread, outliers, etc. as examples of non-linear patterns

Residual Plots



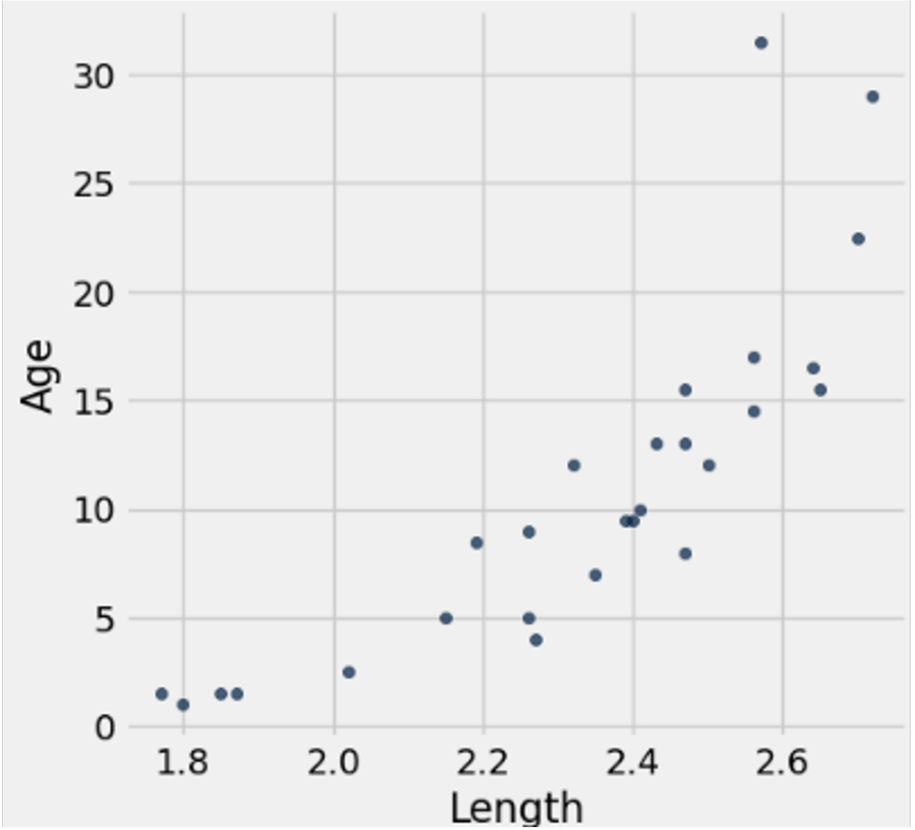


Dugong example

Suppose we have a dataset containing age and length of dugongs

- Is there an association between length and age?

- Is there a *linear* association between length and age?



Notebook Demo: Residuals & Regression Diagnostics

The next two weeks

- **Today**: Least Squares and Residuals ← HW 7 due
- ***Wednesday, Nov 26****: Holiday!*

No office hours this week

- **Monday, Dec 1**: Regression Inference ← HW 8 due
- **Wednesday, Dec 3**: Computing Fellows Workshop
  - Final Project Consultations during Lab

Progress Report
due Tuesday, Dec 2

- **Monday, Dec 8**: Special Topics (Data Privacy) ← HW 9 due
- **Friday, Dec 12**: Final Projects Due

Last day of class :(