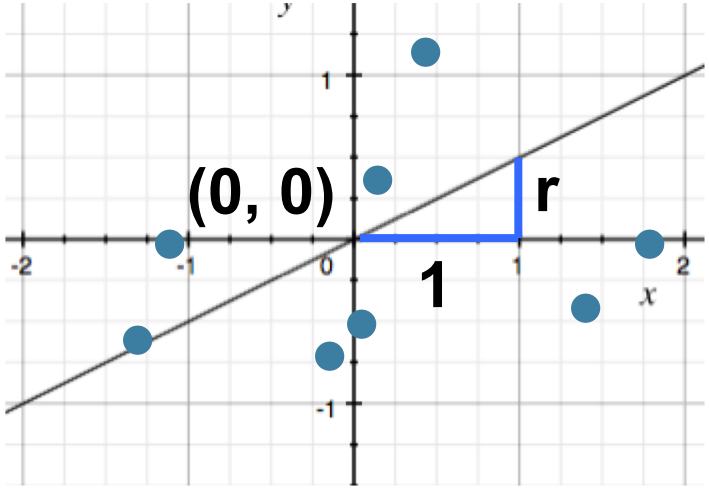
Least Squares

Regression Line Equation

In standard units, the equation of the regression line is:



Fitted value

Observed value

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

Correlation coefficient

Regression Line Equation

In original units, the regression line has this equation:

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

y in standard units

x in standard units

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

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

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

```
# Linear Regression Code

def standard_units(nums):
    return (nums - np.mean(nums)) / np.std(nums)

def correlation(t,x,y):
    return np.mean(standard_units(t.column(x)) * standard_units(t.column(y)))

def slope(t,x,y):
    r = correlation(t,x,y)
    return r * np.std(t.column(y)) / np.std(t.column(x))

def intercept(t,x,y):
    r = correlation(t,x,y)
    return np.mean(t.column(y)) - slope(t, x, y) * np.mean(t.column(x))
```

Demo

Least Squares

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
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

Numerical Optimization

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

Time to Work On Group Project