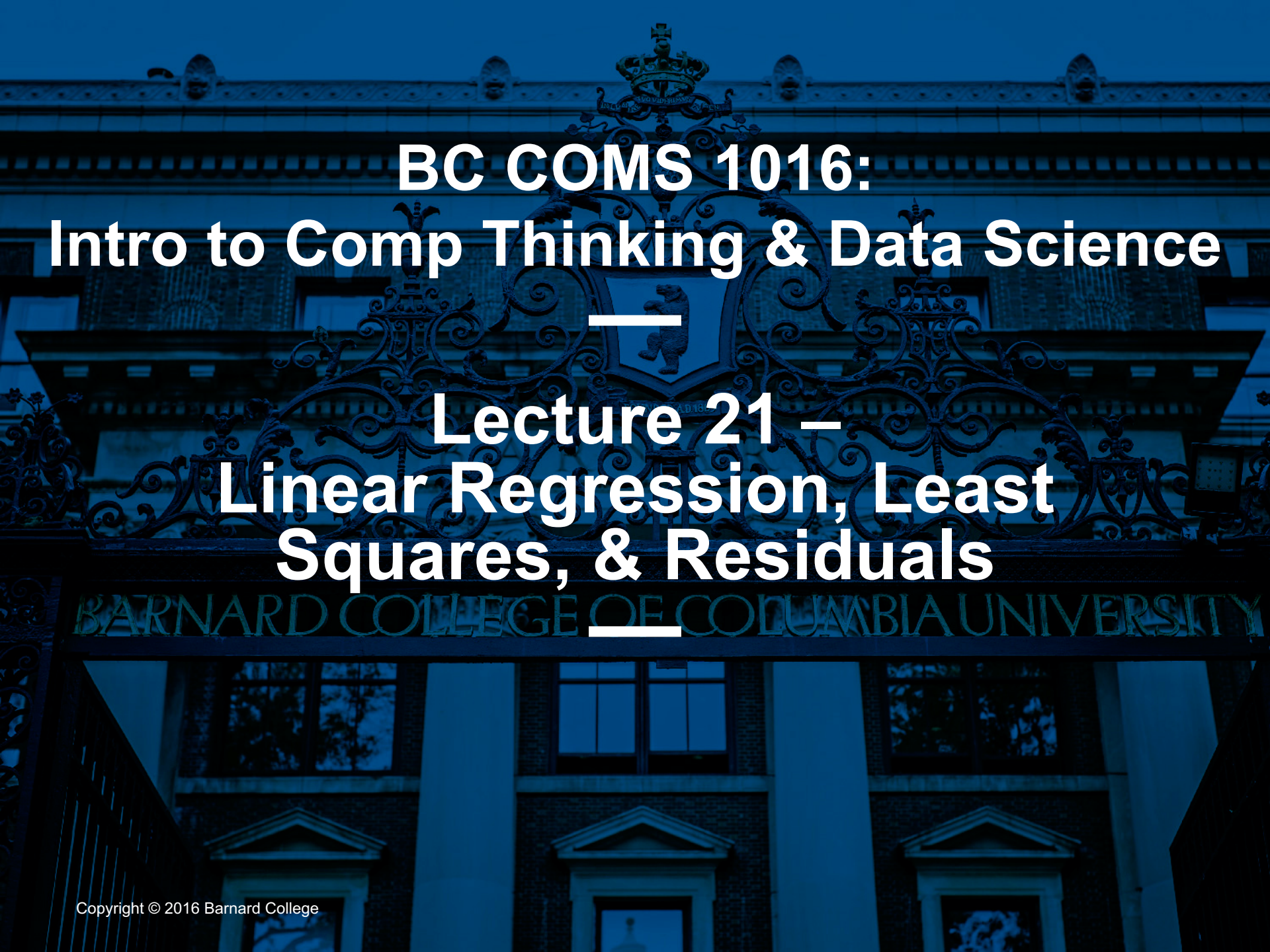


BC COMS 1016: Intro to Comp Thinking & Data Science

—

Lecture 21 – Linear Regression, Least Squares, & Residuals

Announcements

- No Monday/Thursday Lab
  - Last lab will be Wednesday/Thurday
- Homework 8 - Linear Regression
  - Due Monday 12/07
- Homework 9 - Regression Inference
  - Due Thursday 12/10
- Project 3:
  - Due Monday 12/14

Grading – Rubric 1

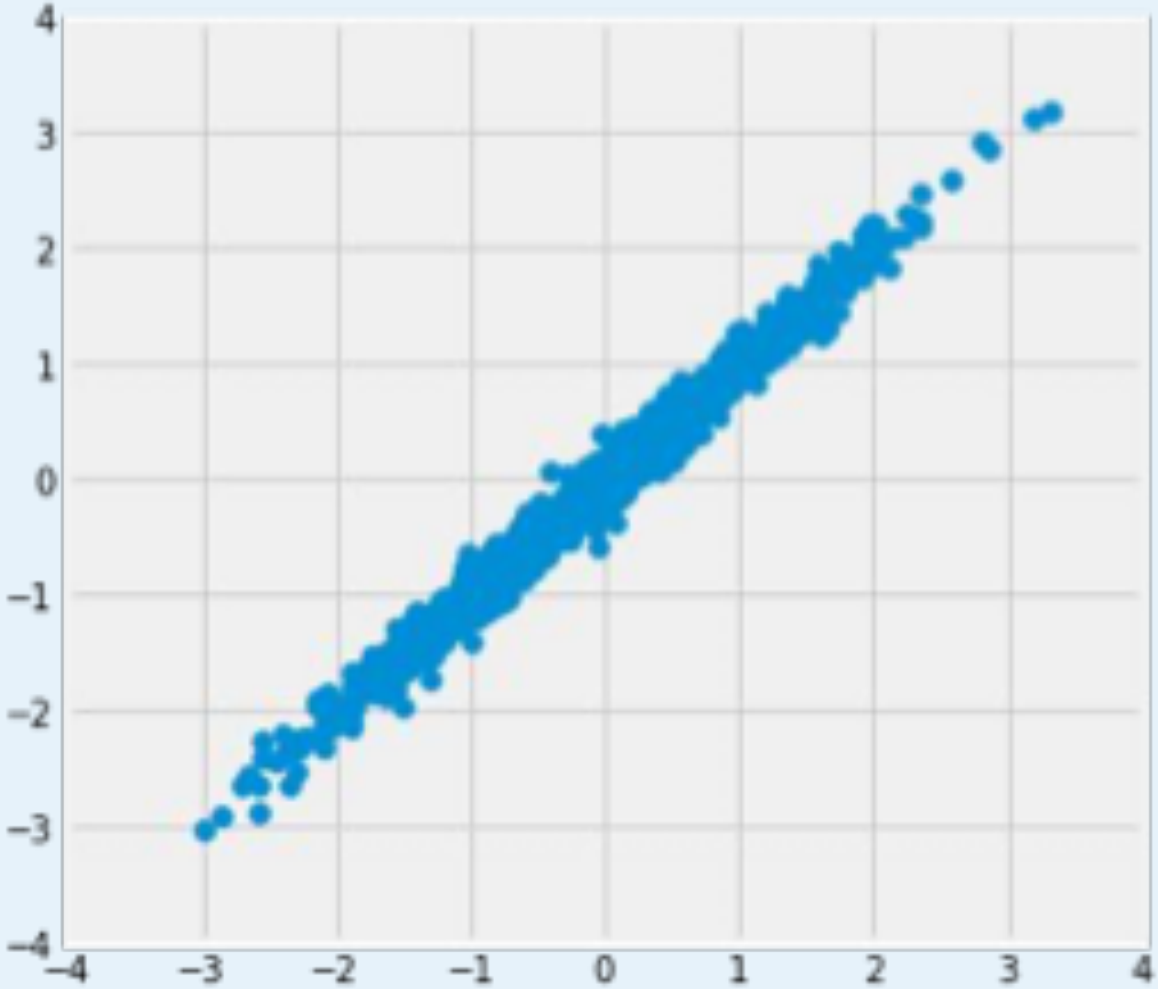
Participation	5%
Weekly HW	25%
Projects	20%
Midterm + daily quizzes	25%
Final Project	25%

Grading – Rubric 2

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

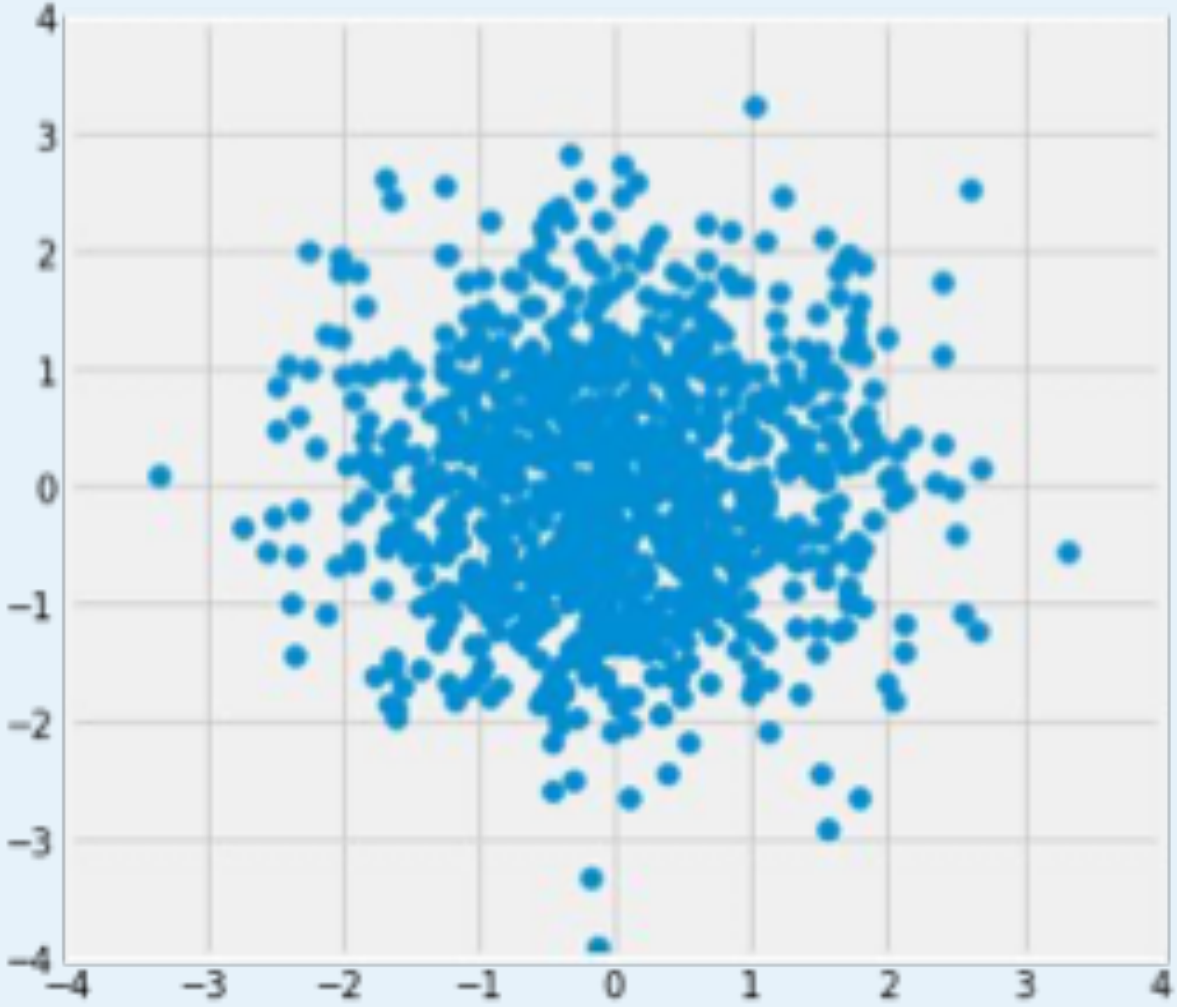
Linear Regression

Where is the prediction line?



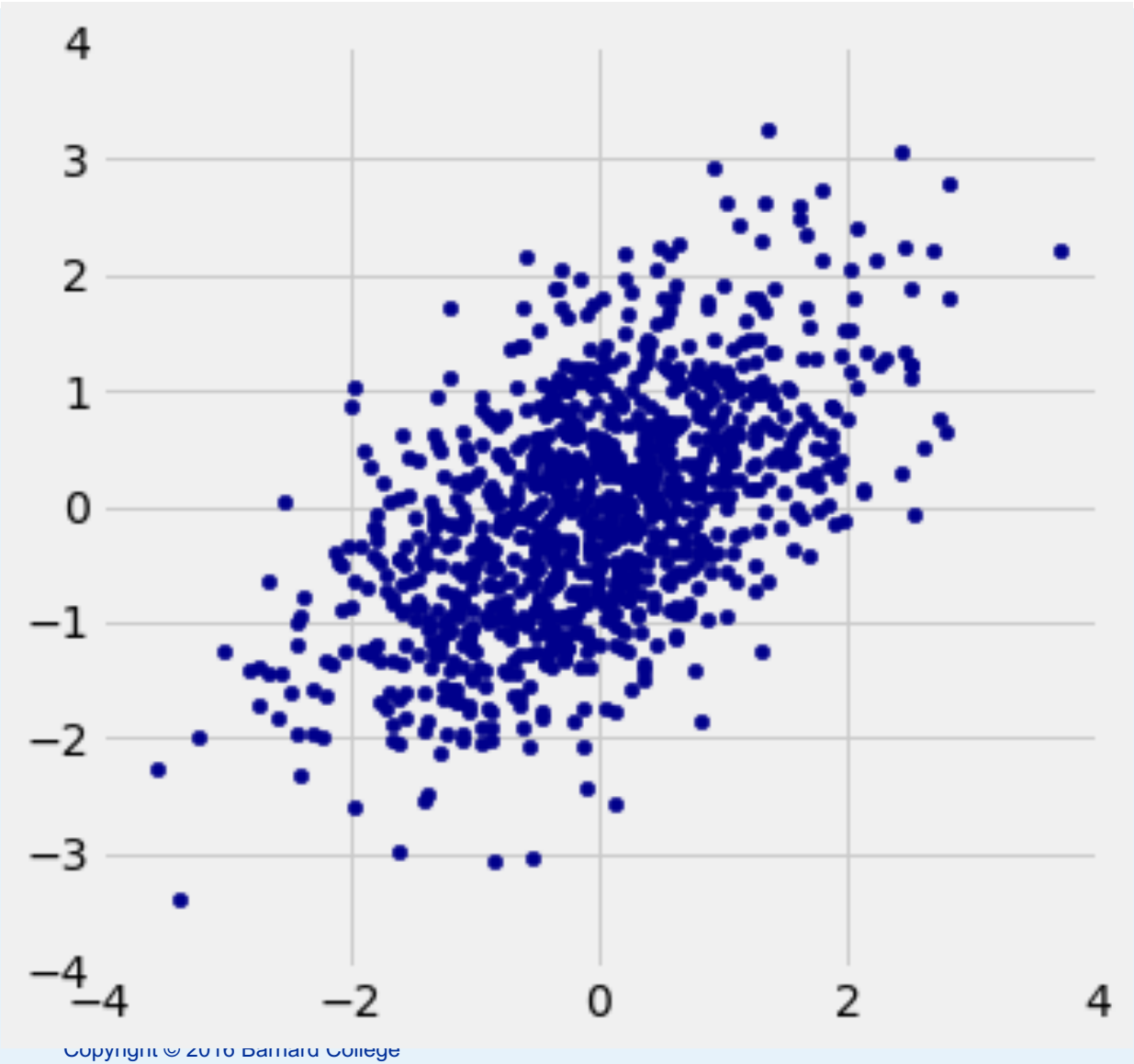
$r = 0.99$

Where is the prediction line?



$r = 0.0$

Where is the prediction line?



$r = 0.5$

Regression Estimate

Goal: Predict *y* using *x*

To find the regression estimate oy *y*:

- Convert the given *x* to standard units
- Multiply by *r*

$$estimate\ of\ y_{su} = r \times given\ x_{su}$$

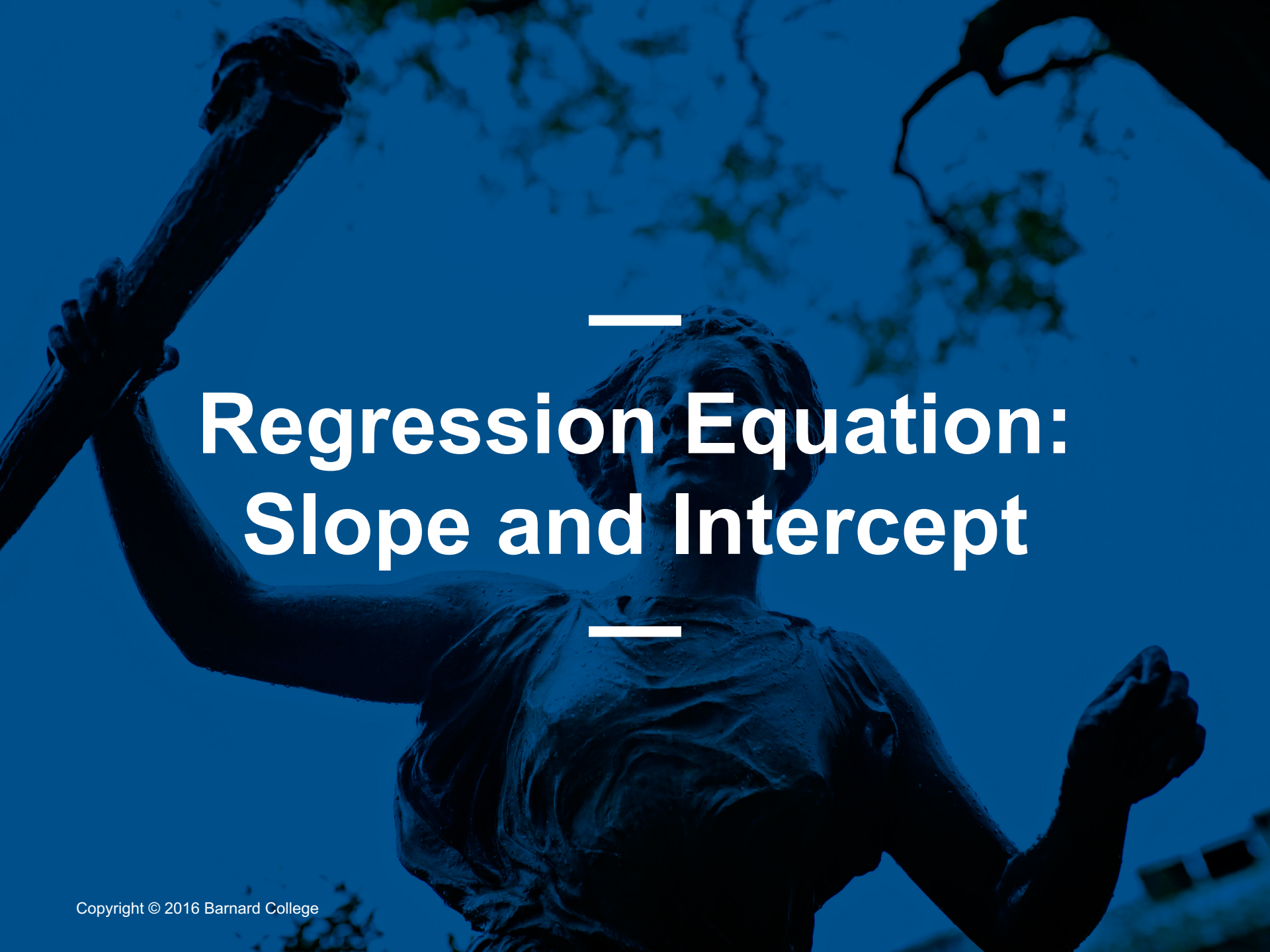
- Convert estimate back to the original units of *y*

Regression to the Mean

Regression to the Mean

$$estimate\ of\ y_{su} = r \times given\ x_{su}$$

- The regression estimate of y is closer to the mean than the given value of x is
- The regression estimate is an average. *On* average, the values of y at a fixed *x* are closer to the mean than *x* is.
- "Regression to the mean" is a statement about averages. It is not true for all individuals



Regression Equation: 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*

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

Slope and Intercept

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

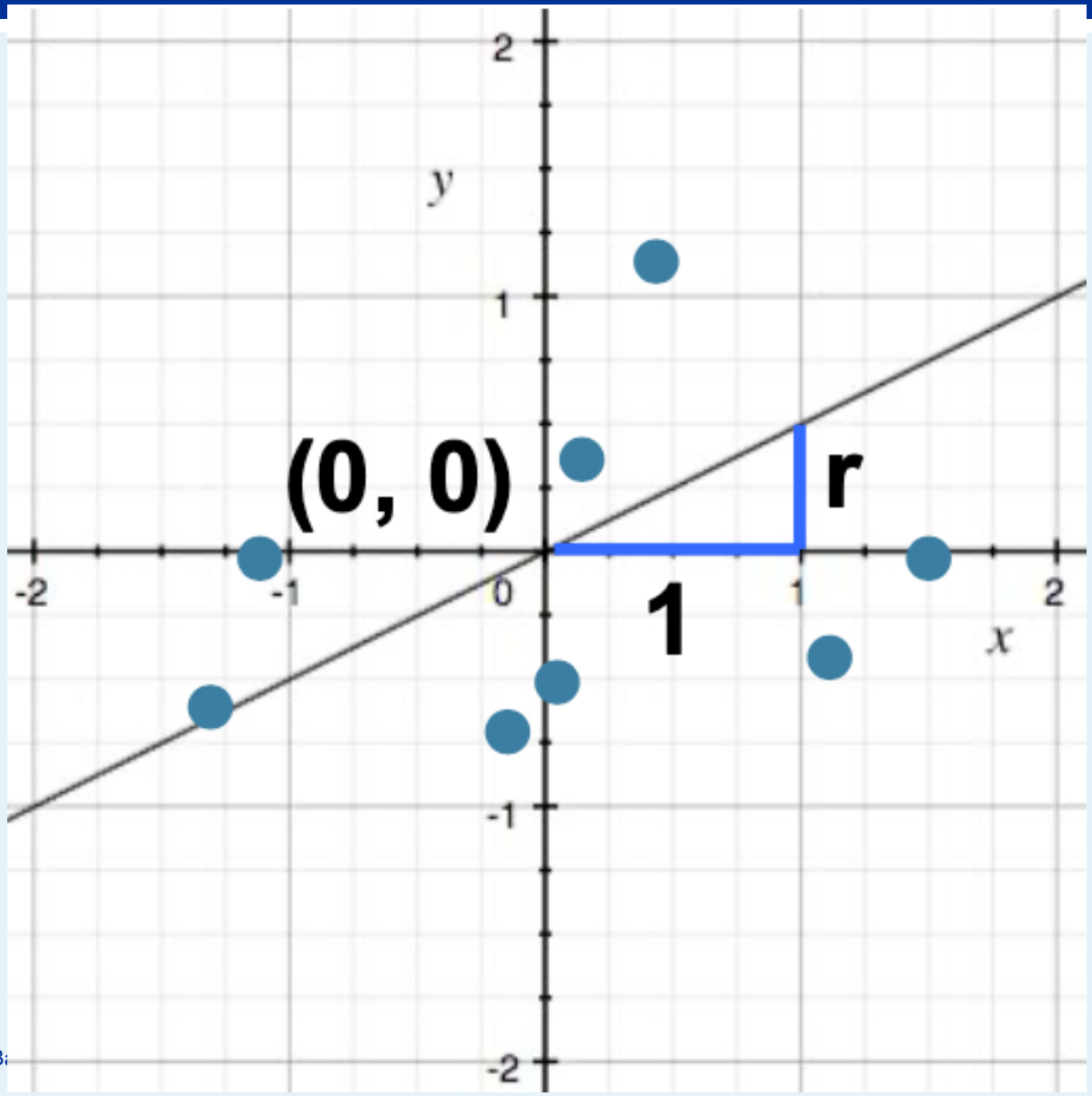
$$\boldsymbol{slope\ of\ the\ regression\ line}$$

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

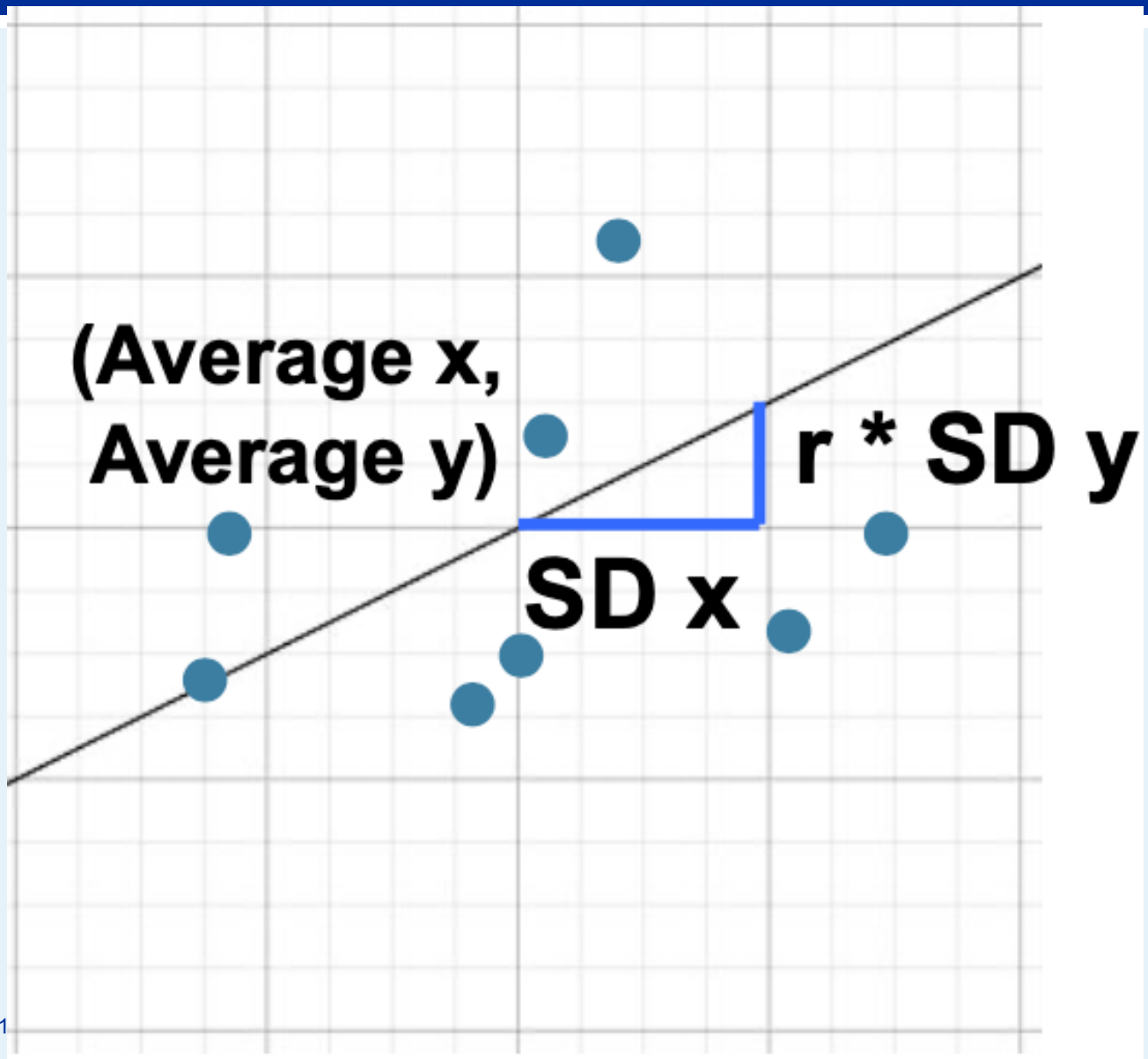
$$\boldsymbol{intercept\ of\ the\ regression\ line}$$

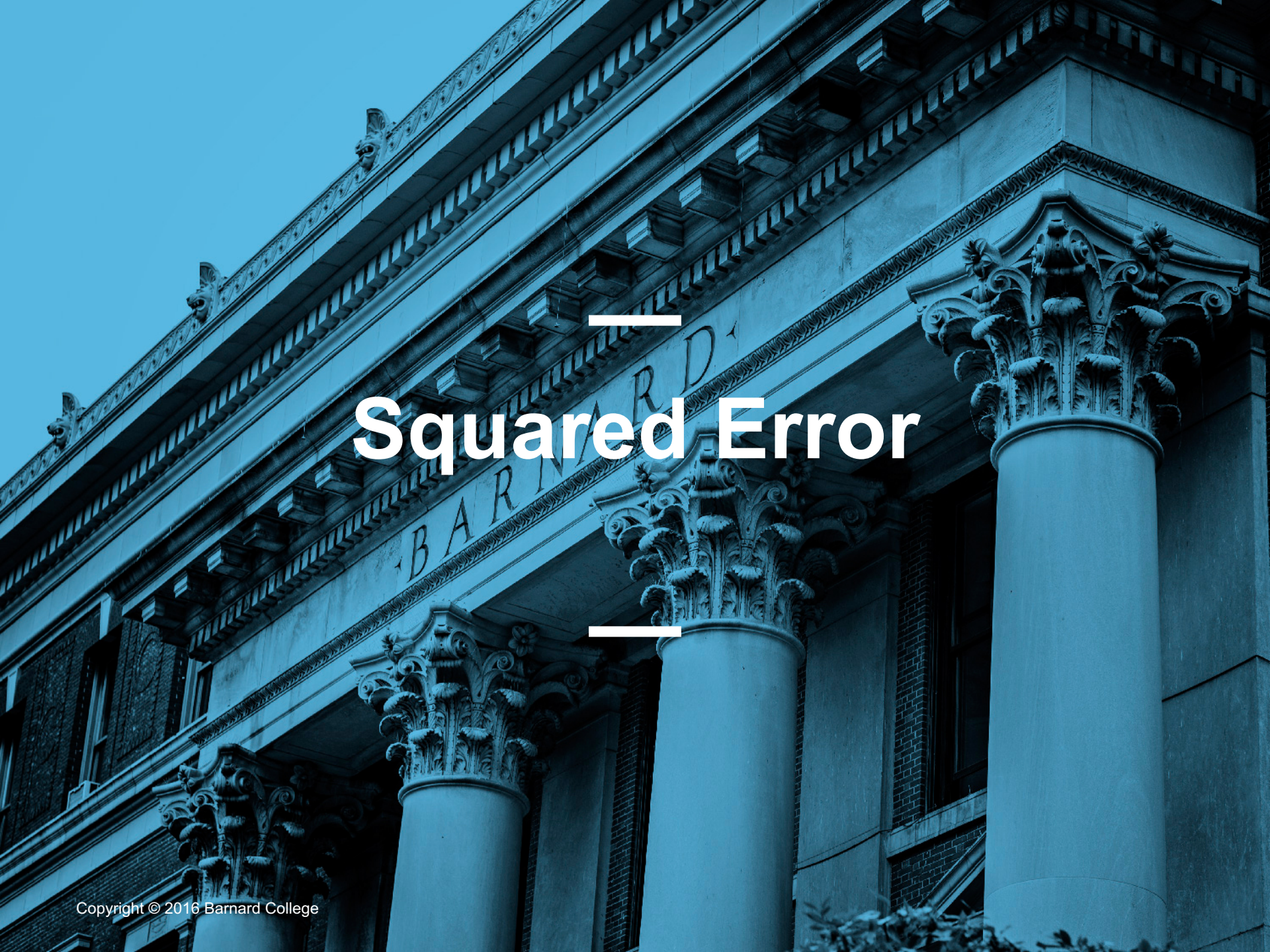
$$mean(y) - slope \times mean(x)$$

Regression Line – Standard Units



Regression Line – Original Units





Squared Error

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

Least Squares Line

- Minimizes 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
- If the function **mse(a, b)** returns the mse of estimation using the line "estimate = a*x* + 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 Dlagnostics

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)