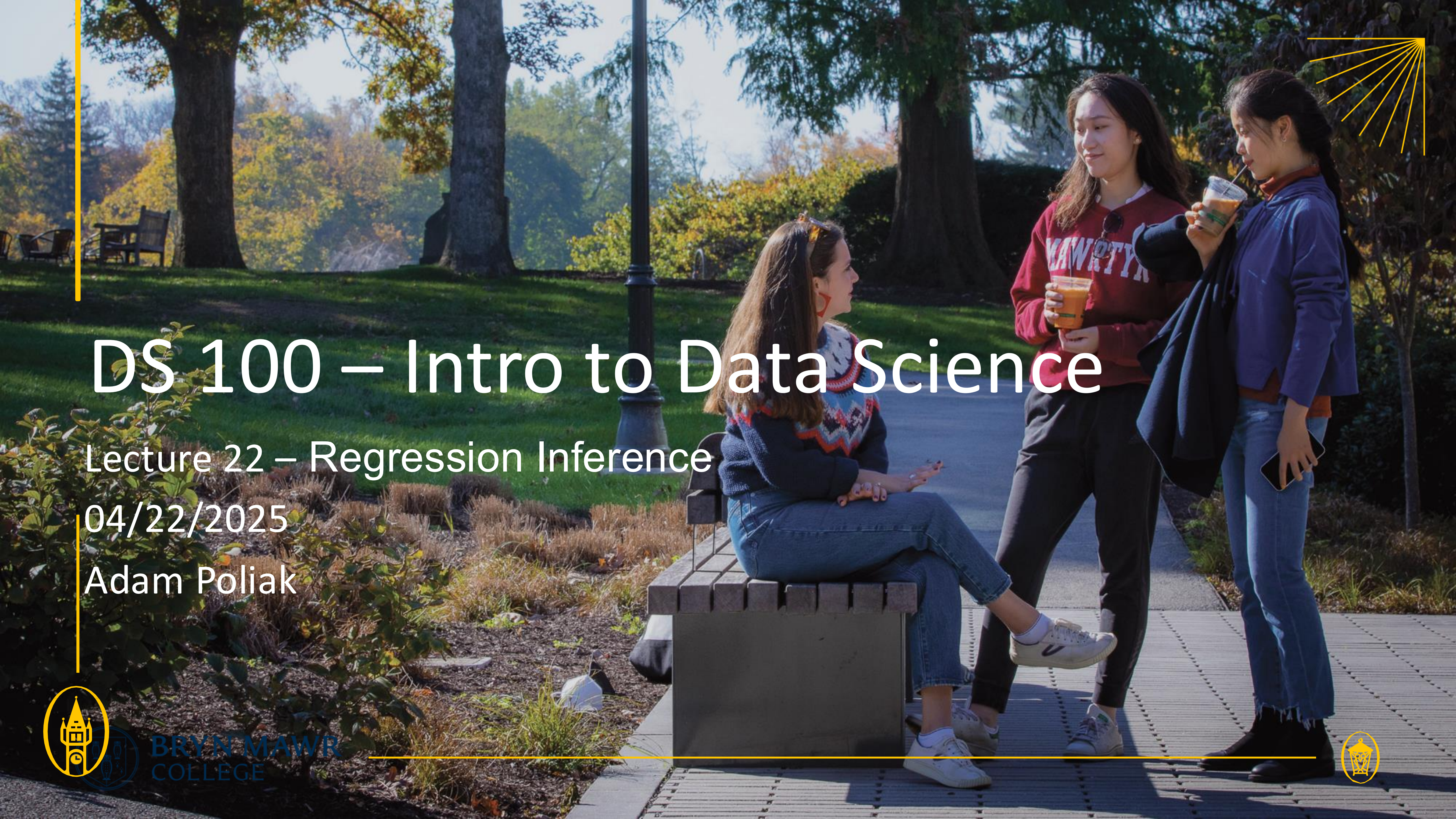


DS 100 – Intro to Data Science

Lecture 22 – Regression Inference

04/22/2025

Adam Poliak

BRYN MAWR COLLEGE

Announcements

Midsemester feedback form - https://forms.gle/M4jVdGTDgQknWAwo6

HW10 (due Friday May 2nd)

Project 3 (due Friday May 2nd)

Linear Regression

Finding the best-fit line

Compute correlation coefficient *(r)*

- Prediction in standard units

Find slope and intercept of the data

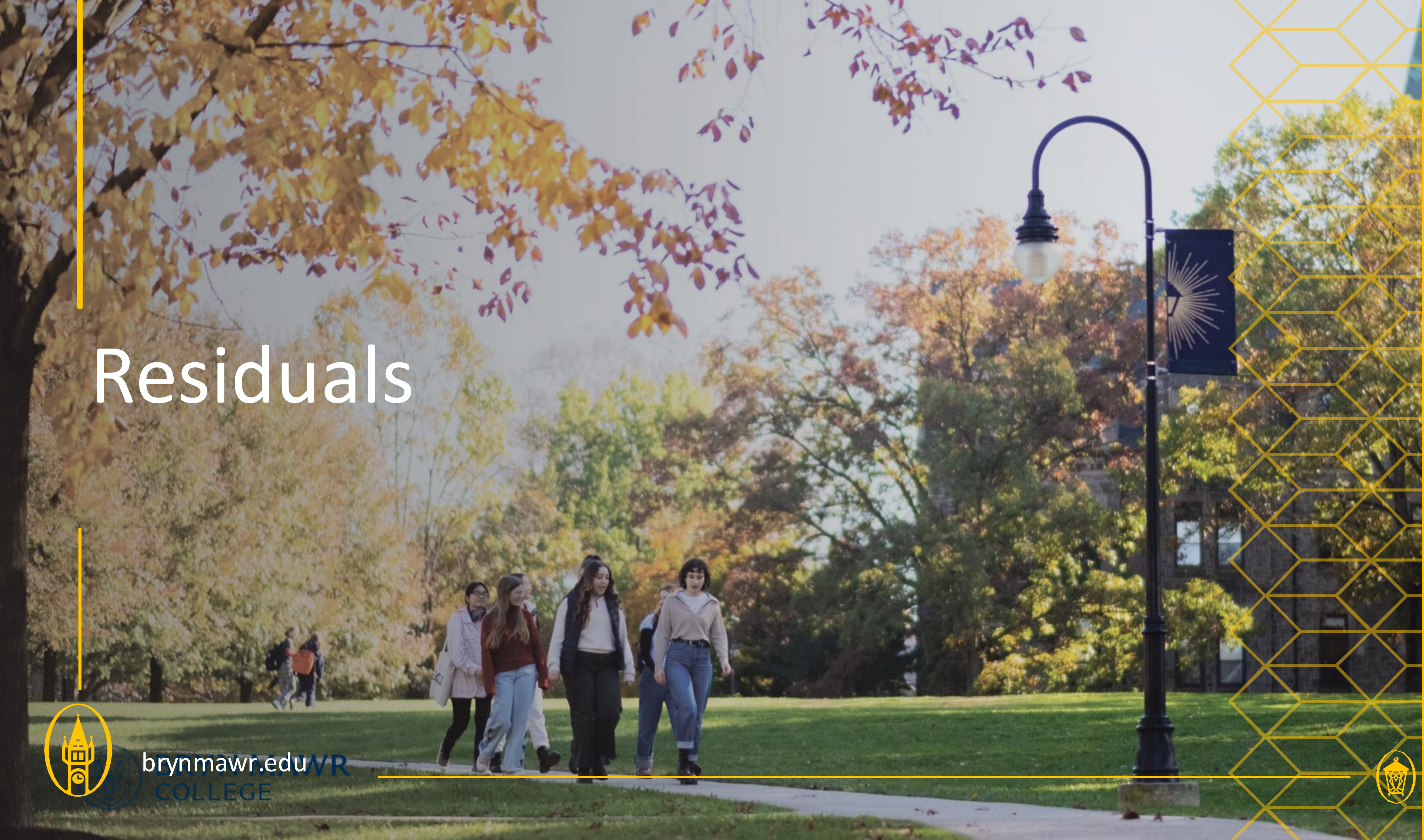
- Prediction in original units
- slope = r * sd(y) / sd(x)
- intercept = mean(y) – slope * mean(x)

Numerical Optimization:

- Use a compute to find slope and intercept to minimize y

y = slope * x + intercept

Residuals



Residuals

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)

Standard Deviation of Fitted (Predicted) Values

We just said

- (Variance of fitted values) / (Variance of y) = r^2
- variance is standard deviations squared,

So:

- $\frac{SD\ of\ fitted\ values}{SD\ of\ y} = |r|$
- $SD\ of\ fitted\ values\ = |r|\ * (SD\ of\ y)$

- $\frac{Variance\ of\ fitted\ values}{Variance\ of\ y} = r^2$

A Variance Decomposition

By definition,

y = fitted values + residuals

Var(y) = Var(fitted values) + Var(residuals)

A Variance Decomposition

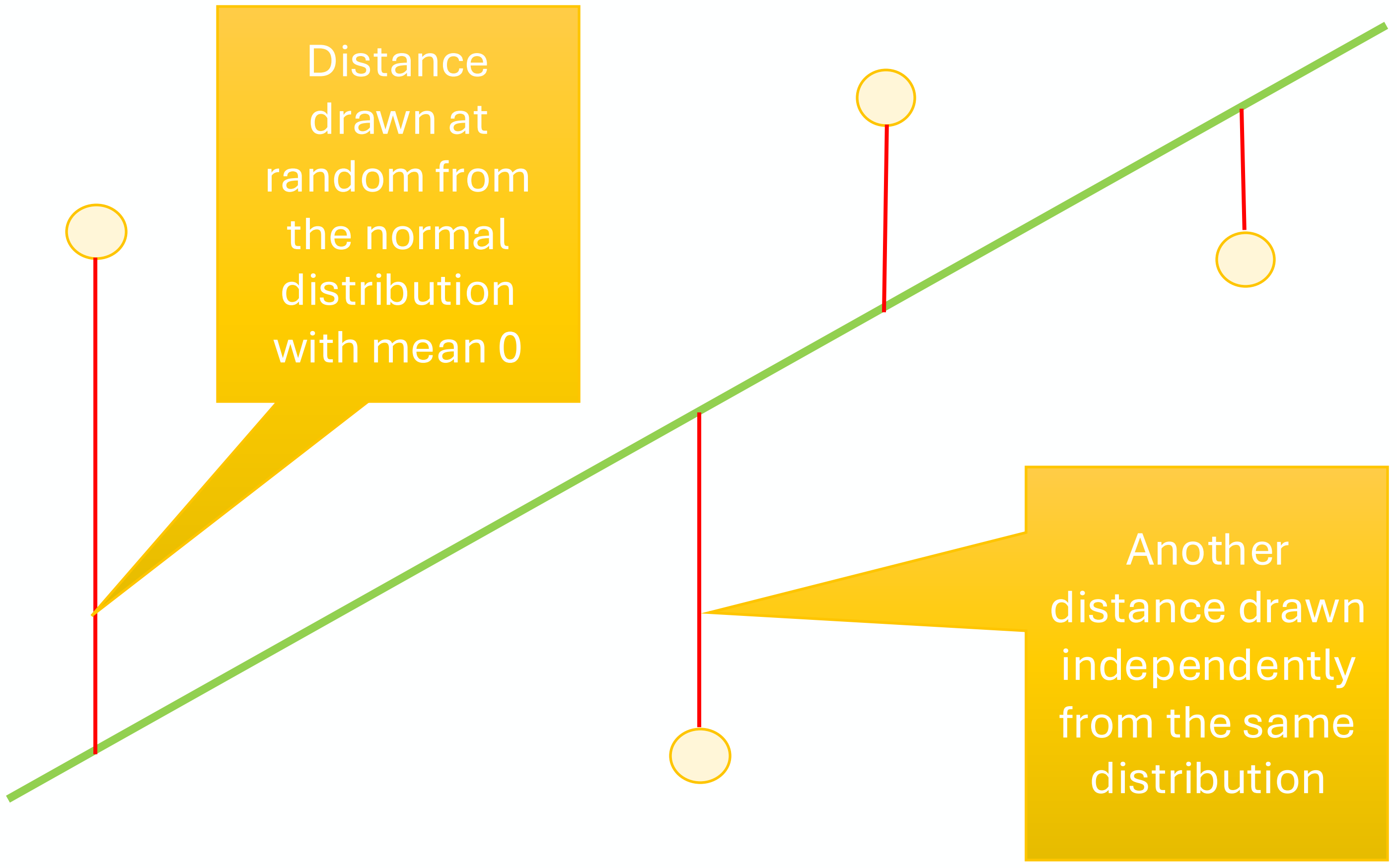
Var(y) = Var(fitted values) + Var(residuals)

$$\frac{Variance\ of\ fitted\ values}{Variace\ of\ y} = r^2$$

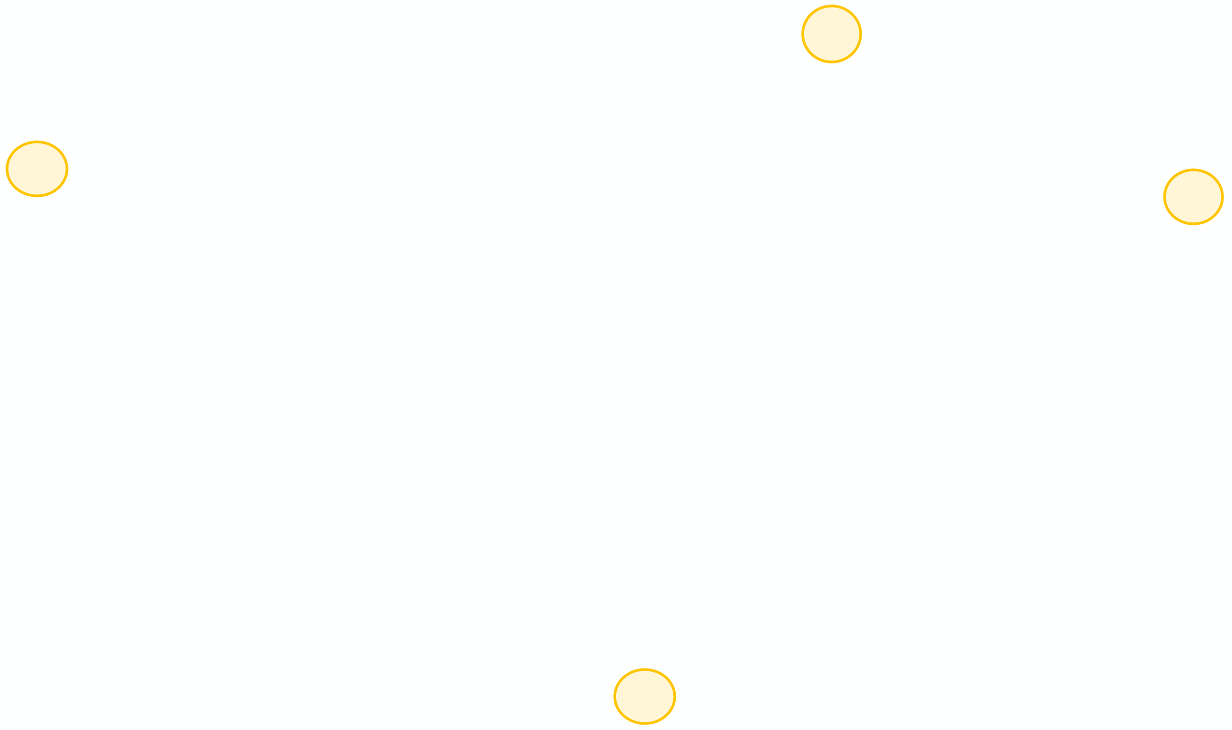
$$\frac{Variance\ of\ residuals}{Variace\ of\ y} = 1\ -\ r^2$$

Regression Model

A “Model”: Signal + Noise



What we get to see



Prediction Variability

Regression Prediction

- If the data come from the regression model,
- And if the sample is large, then:

- The regression line is close to the true line
- Given a new value of *x*, predict *y* by finding the point on the regression line at that *x*

Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at *y*.

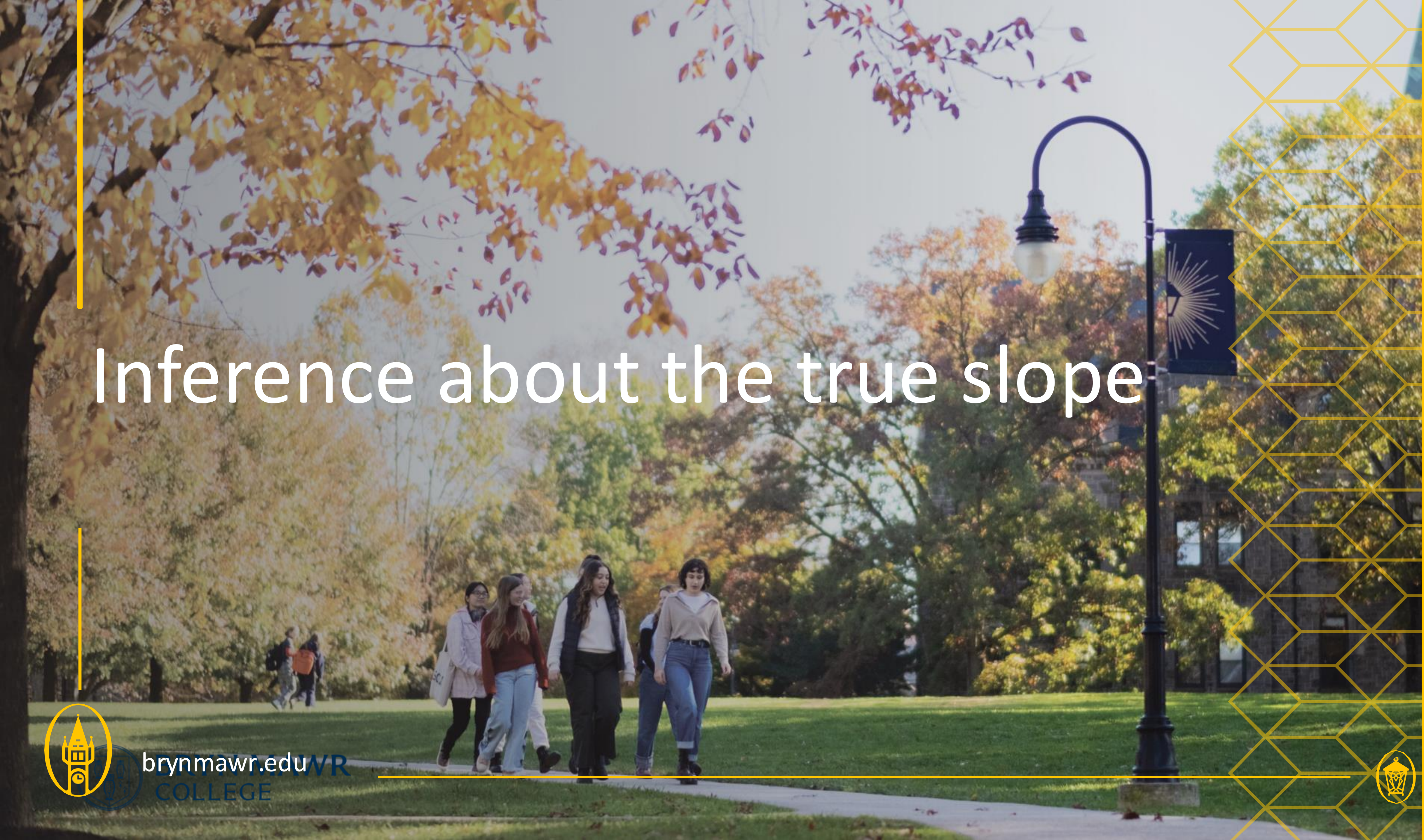
Predictions at Different Values of x

Since y is correlated with x, the predicted values of y depend on the value of x.

The width of the prediction's CI also depends on x.

- Typically, intervals are wider for values of x that are further away from the mean of x.

Inference about the true slope



Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat the first two steps.
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.

Test Whether There Really is a Slope

Null hypothesis: The slope of the true line is 0.
Alternative hypothesis: No, it's not.
Method:

- Construct a bootstrap confidence interval for the true slope.
- If the interval doesn't contain 0, the data are more consistent with the alternative
- If the interval does contain 0, the data are more consistent with the null