

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

## Lecture 22 – Residuals & Regression Inference

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

# Announcements



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



# 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
- Numerical Optimization:
  - Find slope and intercept to minimize
$$y = \text{slope} * x + \text{intercept}$$



# 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



## 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
- $(\text{Variance of residuals}) / (\text{Variance of } y) = 1 - r^2$
- $(\text{Variance of fitted values}) / (\text{Variance of } y) = r^2$
- Variance of  $y = (\text{Variance of fitted values}) + (\text{Variance of residuals})$



# SD of Fitted (Predicted) Values

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



# Variance of Fitted (Predicted) Values

- Variance = Square of the SD  
= Mean Square of the Deviations
- Variance has weird units, but good math properties
- $\frac{\text{Variance of fitted values}}{\text{Variance of } y} = r^2$

# A Variance Decomposition



By definition,

$$y = \text{fitted values} + \text{residuals}$$

$$\text{Var}(y) = \text{Var}(\text{fitted values}) + \text{Var}(\text{residuals})$$

# A Variance Decomposition



$$\text{Var}(y) = \text{Var}(\text{fitted values}) + \text{Var}(\text{residuals})$$

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



# A Variance Decomposition

$$\text{Var}(y) = \text{Var}(\text{fitted values}) + \text{Var}(\text{residuals})$$

- $\frac{\text{SD of fitted values}}{\text{Variace of } y} = |r|$
- $\frac{\text{SD of residuals}}{\text{Variace of } y} = \sqrt{(1 - r^2)}$



# Residual Average and SD

- The average of residuals is always 0
- $\frac{\text{Variance of residuals}}{\text{Variace of } y} = 1 - r^2$
- SD of residuals      = SD of y, not  $\sqrt{1 - r^2}$

# Question 1



# Midterm: Average 70, SD 10

# Final: Average 60, SD 15

$$r = 0.6$$

The SD of the residuals is .

## Question 2



**Midterm:** Average 70, SD 10

**Final:** Average 60, SD 15

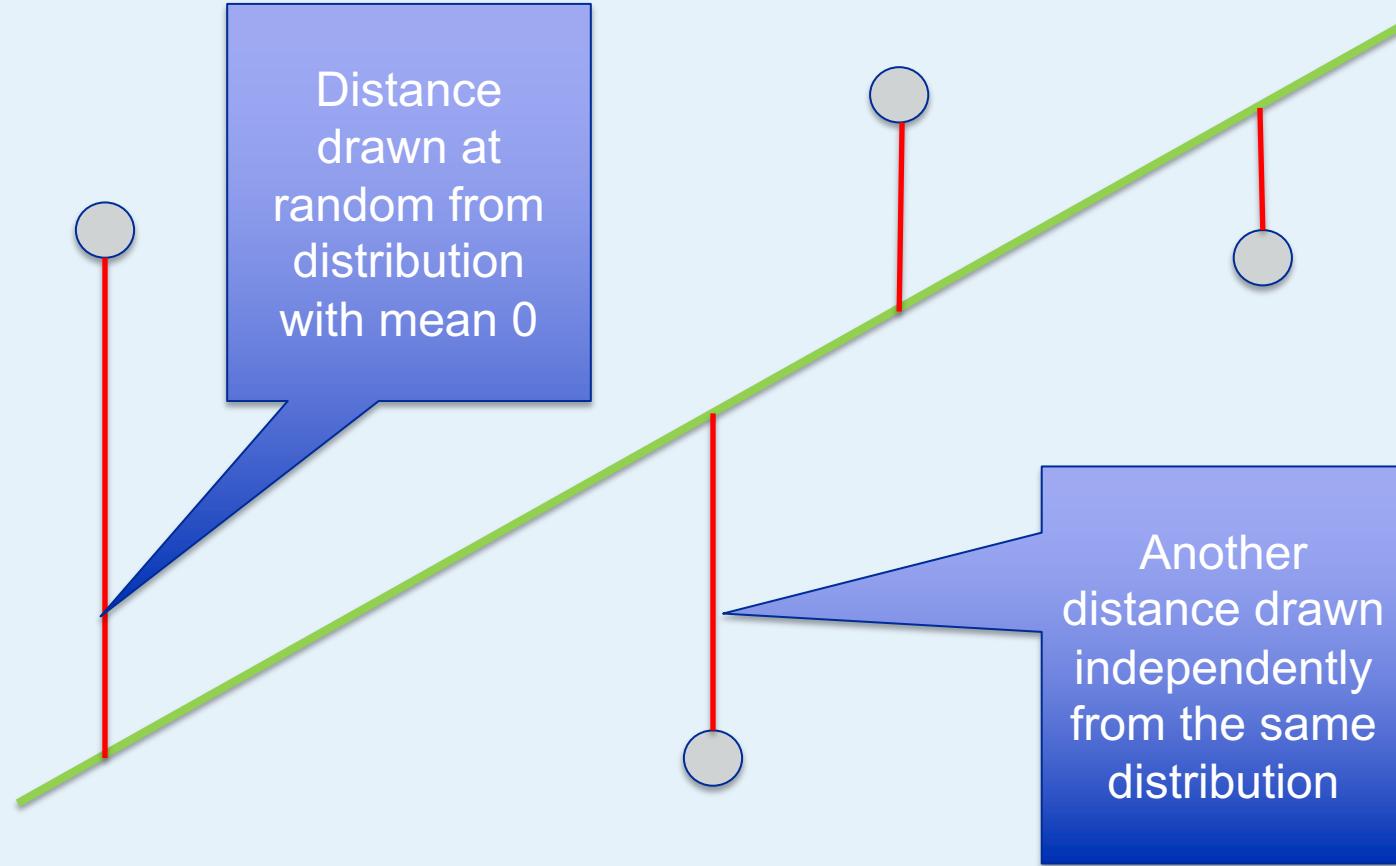
$$r = 0.6$$

For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within \_\_\_\_\_ points.



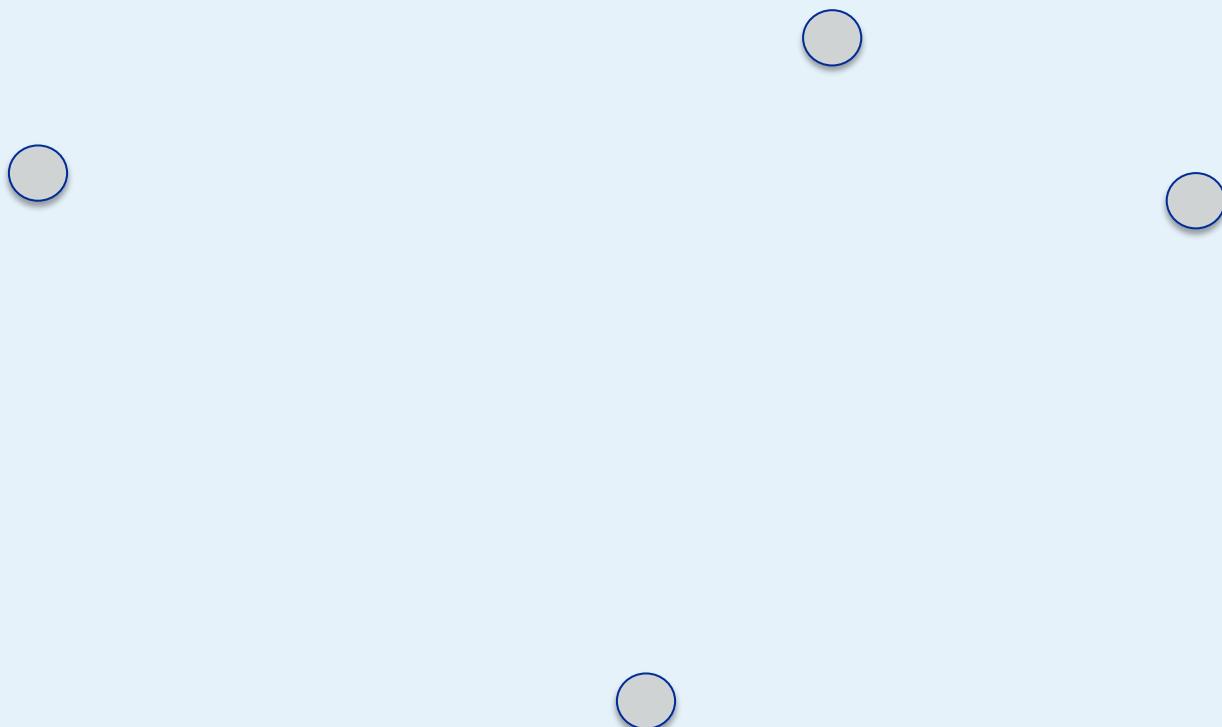
# Regression Model

# 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  $x$ .

# Confidence Interval for True Slope



- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- 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.

# Rain on the Regression Parade



# 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



- `minimize()` works no matter what\*!
- Define a function that computes the prediction you want, then the error you want, for example:
  - Nonlinear functions of  $x$
  - Multiple columns of the table for  $x$
  - Other kinds of error instead of RMSE
- Nonlinear functions can get complicated, fast!