DATA 8

Fall 2016

Lecture 31, November 7

Residuals

Slides created by Ani Adhikari and John DeNero

Announcements

- Project 2 deadline is Tuesday (tomorrow) 11/8 at 7 p.m.
- Homework due Wed/Thurs as usual.
- Friday is an Academic and Administrative Holiday. No lecture and no office hours.

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 erro**r (rmse)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

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 difference between point and line

(Demo)

Residual Plots

For good regressions, the regression plot

- Should look like a blob
- About half above and half below the horizontal line at 0
- Similar vertical spread throughout
- No pattern

Spotting a Problem

Residual plots can be used to detect:

- Non-linearity
  - Shape of scatter plot is curved, not a straight line

Dugong

(Demo)

Spotting another Problem

Residual plots can be also be used to detect:

- Heteroscedasticity
  - Uneven spread

(Demo)

Residual Plots are Flat Overall

- No matter what the shape of the original scatter:
- The residual plot cannot have any overall trend, neither upwards nor downwards
- The correlation between the residuals and the predictor variable is 0.

(Demo)

The Average of the Residuals

- It’s 0.
- Always.
- No matter how nasty the scatter diagram is.

(Demo)

Rough Size of Error in Regression

$$\text{SD of residuals} = \sqrt{1 - r^2} \times \text{SD of } y$$

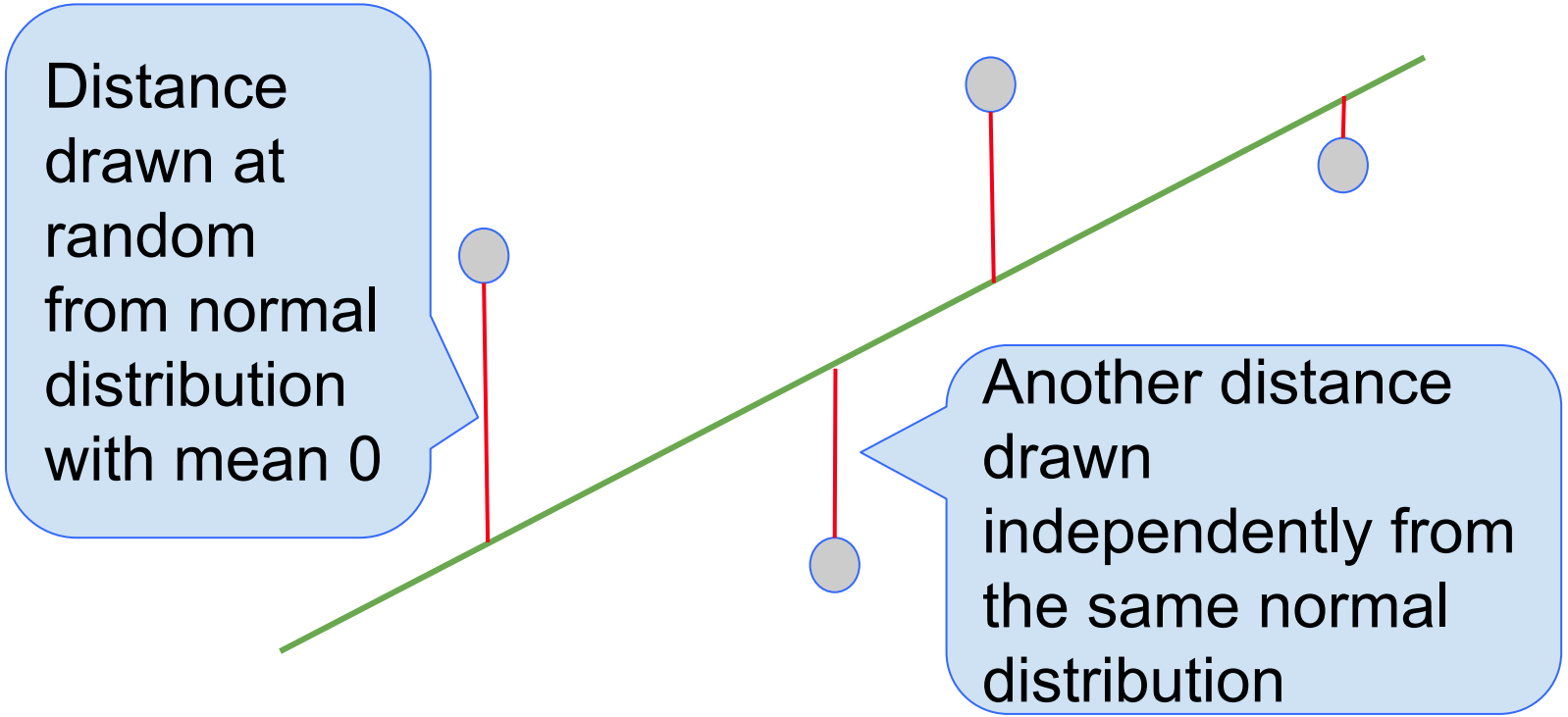
(Demo)

Another Way to Think About *r*

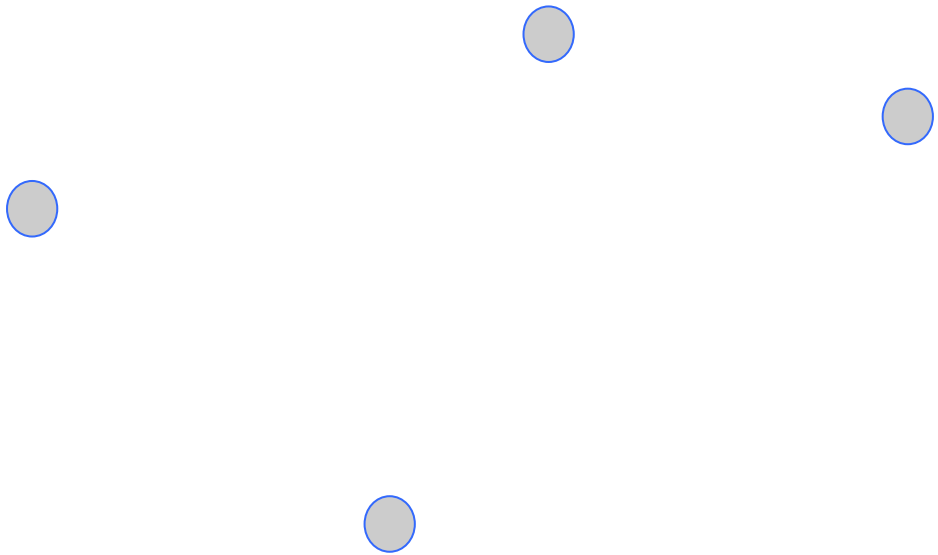
$$|r| = \frac{\text{SD of fitted values of } y}{\text{SD of observed values of } y}$$

Tyche, the Goddess of Chance

A “Model”: Signal + Noise



What We Get to See



(Demo)