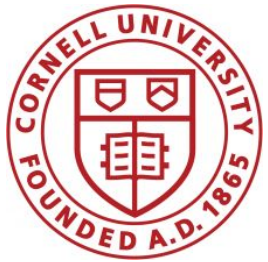

DSFA

Spring 2018

Lecture 32

Residuals

Announcements

Regression line

- **error = actual value − prediction**
- RMSE = root mean square error
- Regression line has the minimum RMSE of all lines

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

Non-linear regression

(Demo)

Residuals

Residuals

- Error in regression prediction
- **residual**
  = observed *y* - regression prediction of *y*
  = vertical distance between each point and the best line

(Demo)

Residual Plot

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But still contains patterns for non-linear relations
- Can reveal whether linear regression is appropriate

(Demo)

Dugong

Mean and Stdev of Residuals

- The mean of the residuals is always 0, no matter what the scatter looks like
- SD(residuals) = RMSE = SD(y) * sqrt(1 - r^2)

(Demo)

Clustering around line

- “The correlation measures how clustered the points are about a straight line.”

- SD(residuals) = RMSE = SD(y) * sqrt(1 - r^2)
- so, RMSE / SD(y) = sqrt(1 - r^2)

(Demo)

Bounds

Rule of thumb:

- About 68% of values within 1 RMSE of prediction
- About 95% of values within 2 RMSE of prediction
- etc.

(Demo)

What we can learn from *r*

- How clustered points are around a line
- How *y* depends on *x*
- How accurate linear regression predictions will be