CORNELL UNIVERSITY FOUNDED A.D. 1865

DSFA

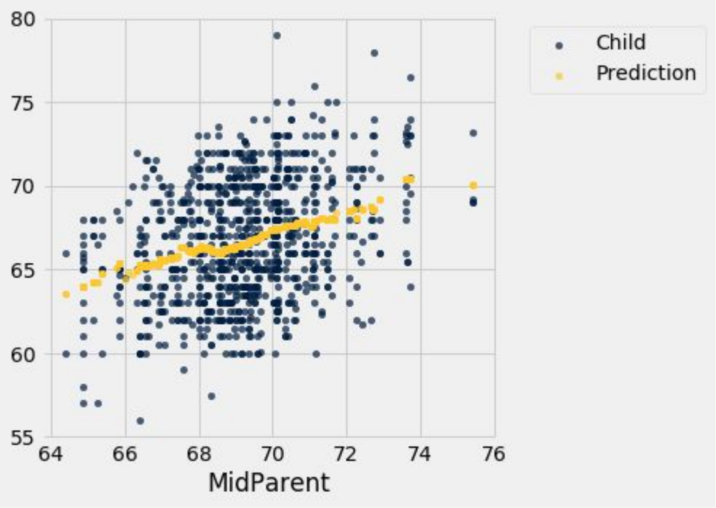
Spring 2020

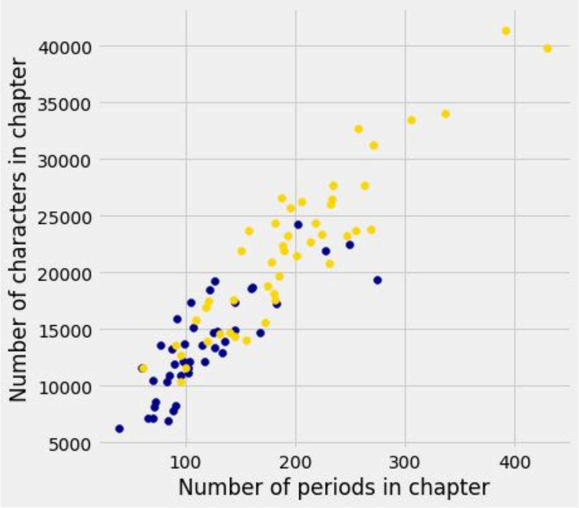
Lecture 24

Residuals

Prediction

If we have a line describing the relation between two variables, we can make predictions





Regression Line Equation

In original units, the regression line has this equation:

$$\underbrace{\frac{\text{estimate of } y \; - \; \text{average of } y}{\text{SD of } y}}_{\text{y in standard units}} = r \times \underbrace{\frac{\text{the given } x \; - \; \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

$$y = \text{slope} \times x + \text{intercept}$$

$$\textbf{slope of the regression line} \; = \; r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} \; = \; \text{average of } y \; - \; \text{slope} \cdot \text{average of } x$$

Errors and Predictions

- **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

(Demo)

Mean and Stdev of Residuals

No matter what the scatter looks like...

- mean(residuals) = 0
- 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

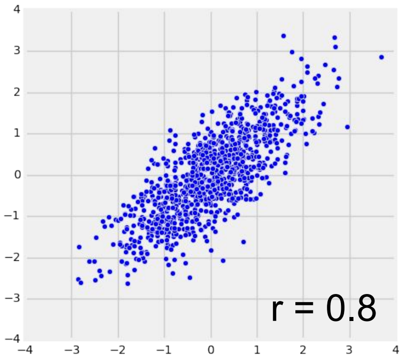
Rule of thumb:

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

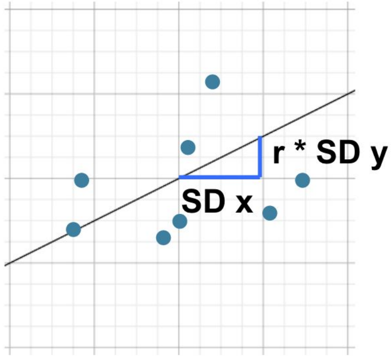
(Demo)

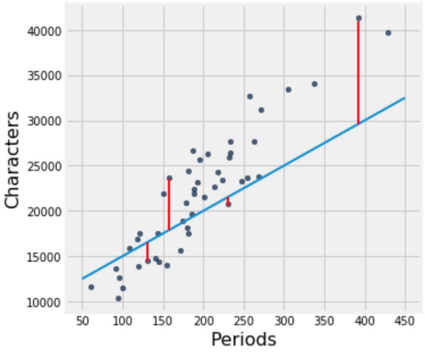
What we can learn from *r*

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



4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4
r = 0.8




Characters
Periods