DSFA

Spring 2018

Lecture 33

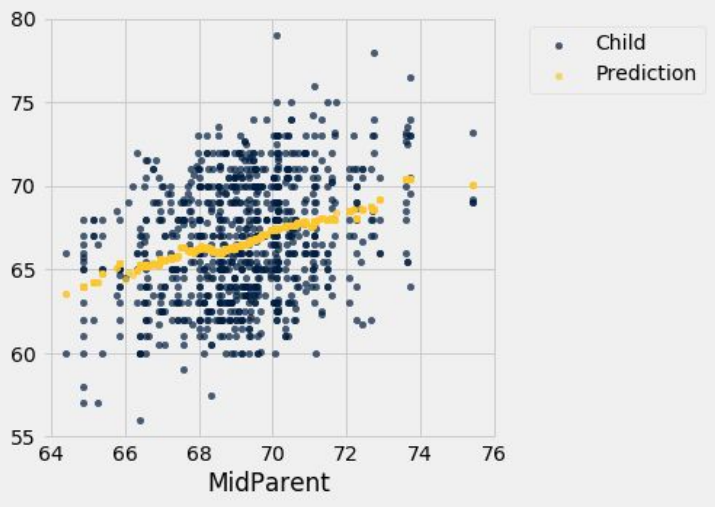
Regression Inference

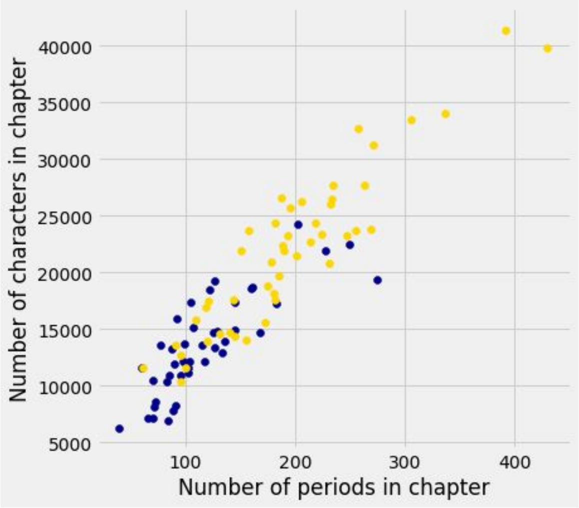
Announcements

- Hw07: make sure your kernel is “Python 3.6 (beta)” not “Python 3”. See Piazza post @133.

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

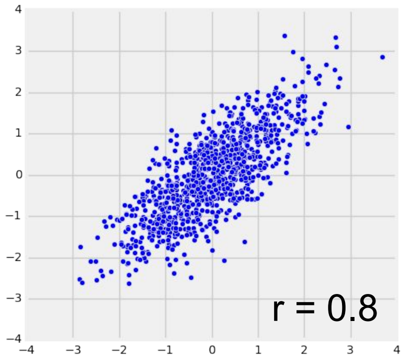
Bounds

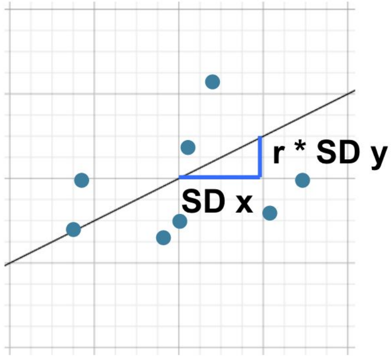
Rule of thumb:

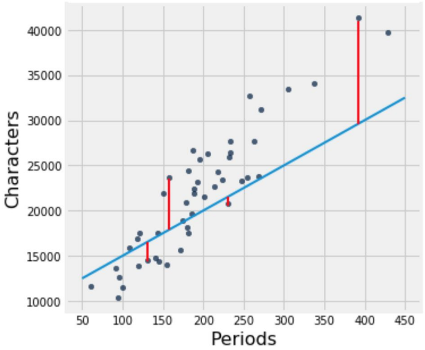
- About 68% of values within 1 RMSE of prediction
- About 95% of values within 2 RMSE of prediction
- etc.

Summary: 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







Prediction from a Sample

Prediction from a Sample

- We’ve been treating dataset as though it were population
- What if we had to make predictions from samples?

(Demo)

Confidence Interval for Prediction

- **Bootstrap:**
  - **Resample the data**
  - **Get a prediction for *y* using the regression line that goes through the resampled data**
  - **Repeat the above two steps, many times**
- Draw the empirical histogram of all the predictions
- Get the “middle 95%” interval
- That’s an approximate 95% confidence interval for the predicted value of *y*

(Demo x 2)

Regression Inference

Applying inference to regression

- Inference techniques: bootstrap, hypothesis testing, confidence intervals
- Regression: correlation, prediction, slope, intercept, RMSE, etc.

Test Whether Variables are Correlated

- **Null hypothesis:** The correlation is 0
- **Alternative hypothesis:** It’s not
- **Test statistic:** abs(sample correlation)
- **Method:**
    - Construct a bootstrap simulated distribution for the abs(correlation)
    - Compute a p-value for the observed abs(correlation)

(Demo)