DATA 8

Fall 2016

Lecture 32, November 9

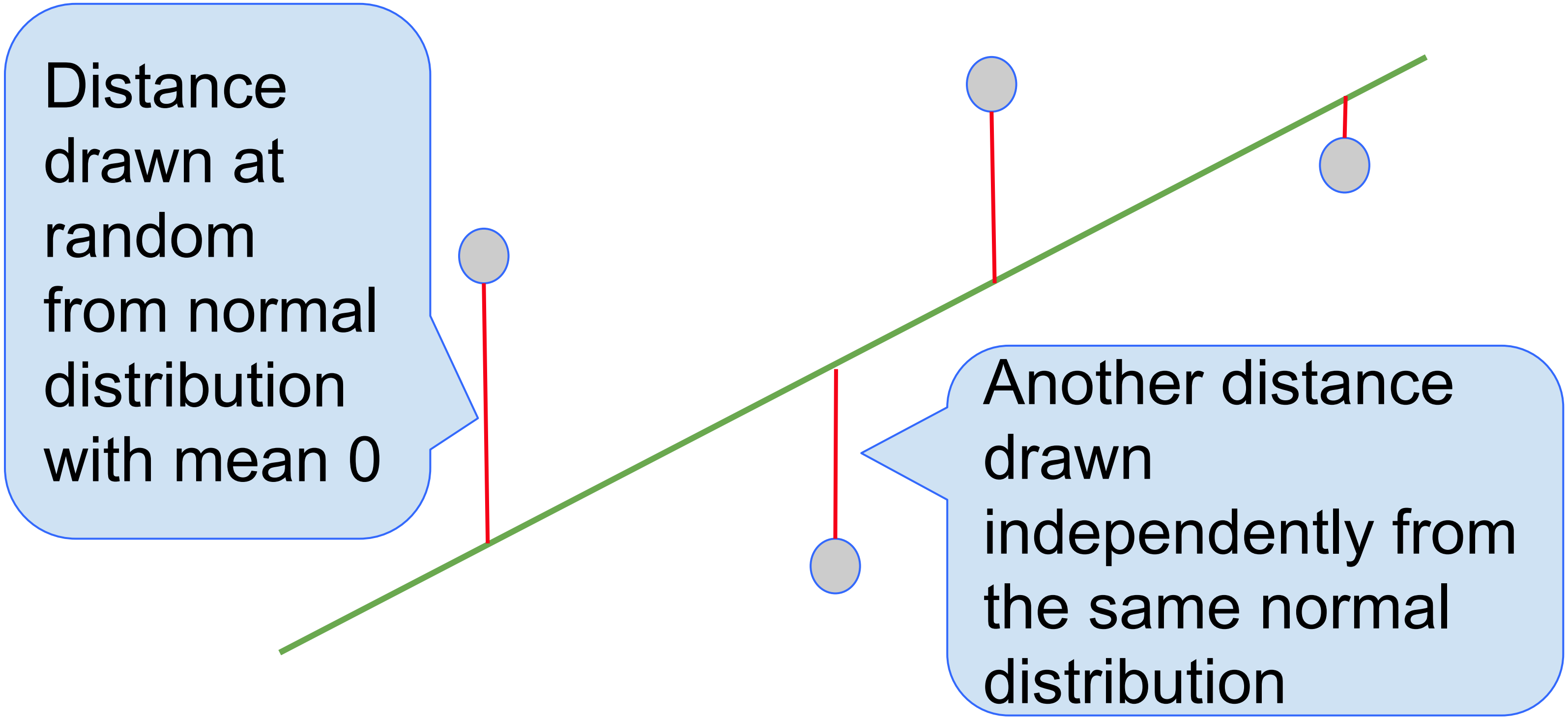
Inference for Regression

Slides created by Ani Adhikari and John DeNero

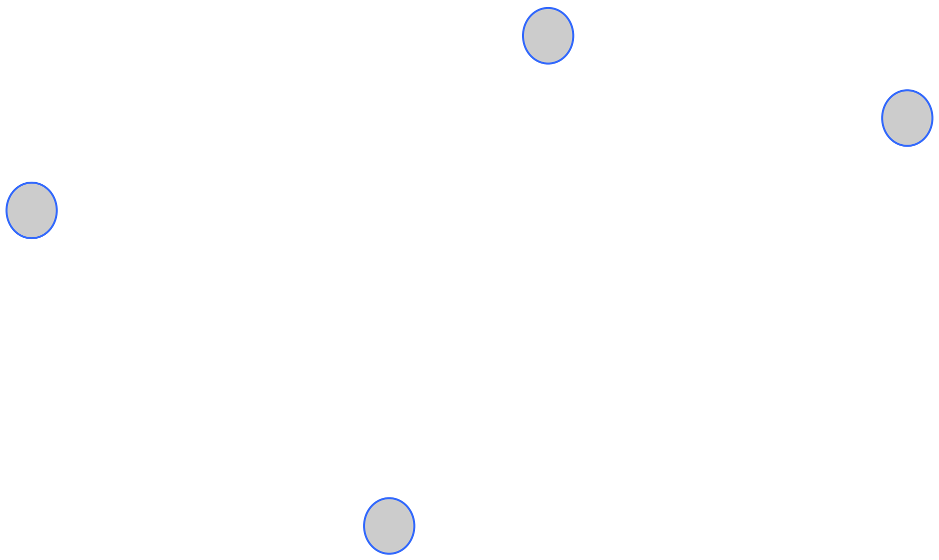
Announcements

- Congratulations on submitting the project!
- Homework due Wed/Thurs as usual.
- Friday is an Academic and Administrative Holiday. No lecture and no office hours.

A “Model”: Signal + Noise



What We Get to See



(Demo)

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.

(Demo)

Rain on the Regression Parade

We observed a slope based on our sample of points.

But what if the sample scatter plot got its slope just by chance?

What if the true line is actually FLAT?

(Demo)

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, reject the null hypothesis.
  - If the interval does contain 0, there isn’t enough evidence to reject the null hypothesis.

Regression Prediction

If the model is good,

- Regression line is close to true line
- Given a new value of *x*, predict y by finding the point on the regression line at that *x*
- **Bootstrap the scatter plot**
- **Get a new prediction using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Get an interval of predictions of *y* for the given *x*

(Demo)

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 interval also depends on *x*.
  - Typically, intervals are wider for values of *x* that are further away from the mean of *x*.