190F Foundations of Data Science

Spring 2020

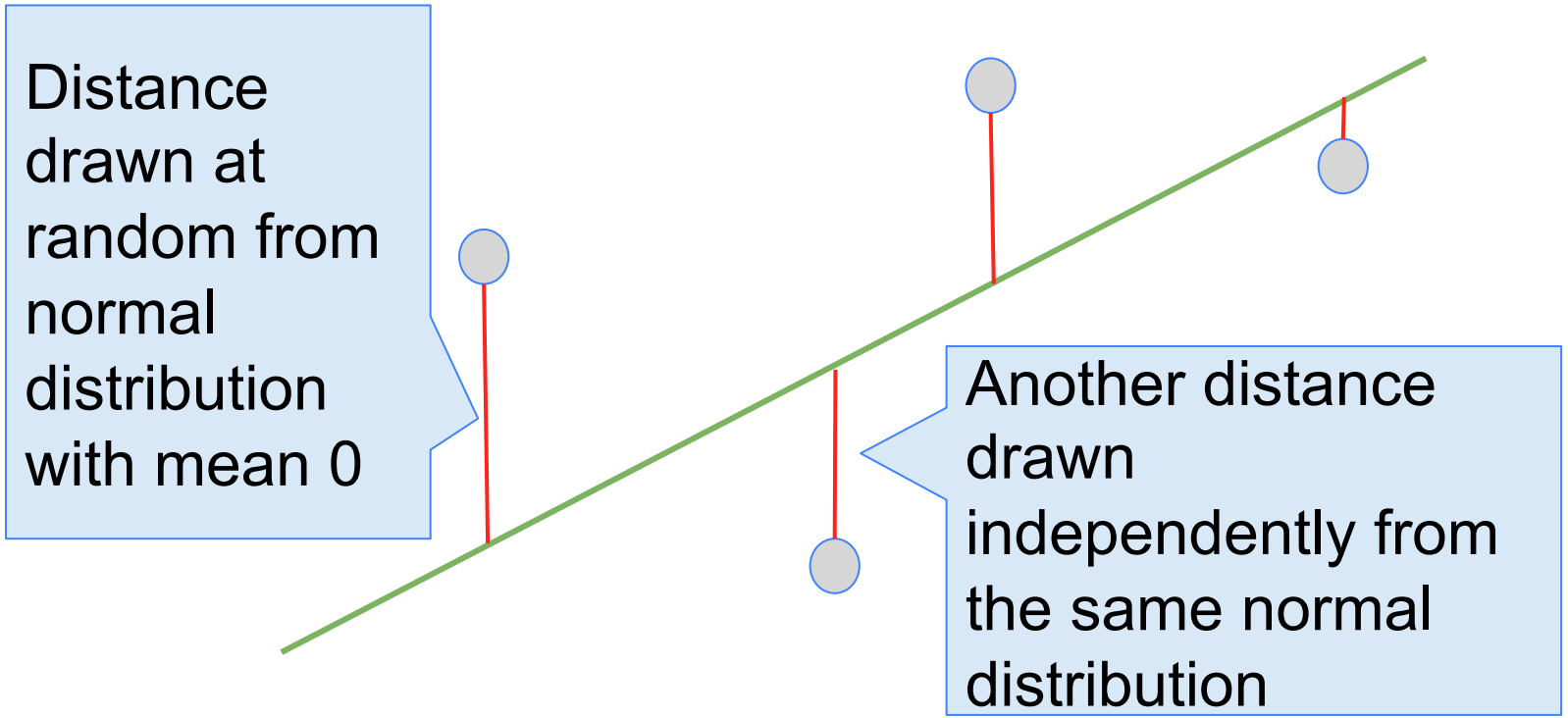
Lecture 20

Regression Inference

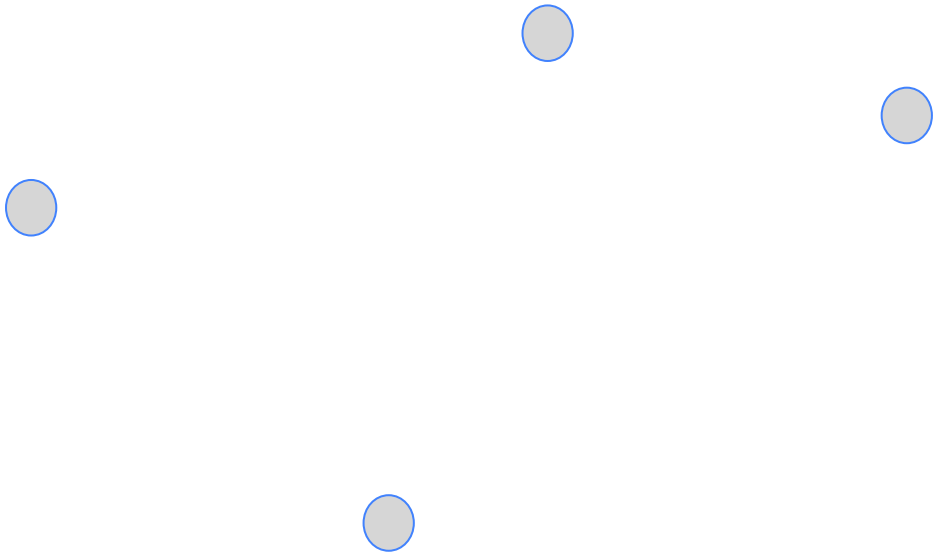
Announcements

Regression Model

A “Model”: Signal + Noise



What We Get to See



(Demo)

Prediction Variability

Regression Prediction

If the data come from the regression model,

- The 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*

Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above 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)

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*.

The True Slope

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.