DATA 8

Spring 2022

Lecture 33

Regression Inference

Announcements

- Lab 8 and HW 8 regrades due tonight
- Homework 10 due Thursday, 04/14
  - Turn in on **tonight** for a bonus point
- Project 2 due Friday, 04/15
  - OH party this Friday 1-5pm at SOCS 581

Weekly Goals

- Monday
  - Least squares: finding the "best" line for a dataset
  - Residuals: analyzing mistakes and errors
- **Today**
  - Regression inference
  - Quantifying uncertainty in the slope & intercept
- Friday
  - Data and privacy

Residuals

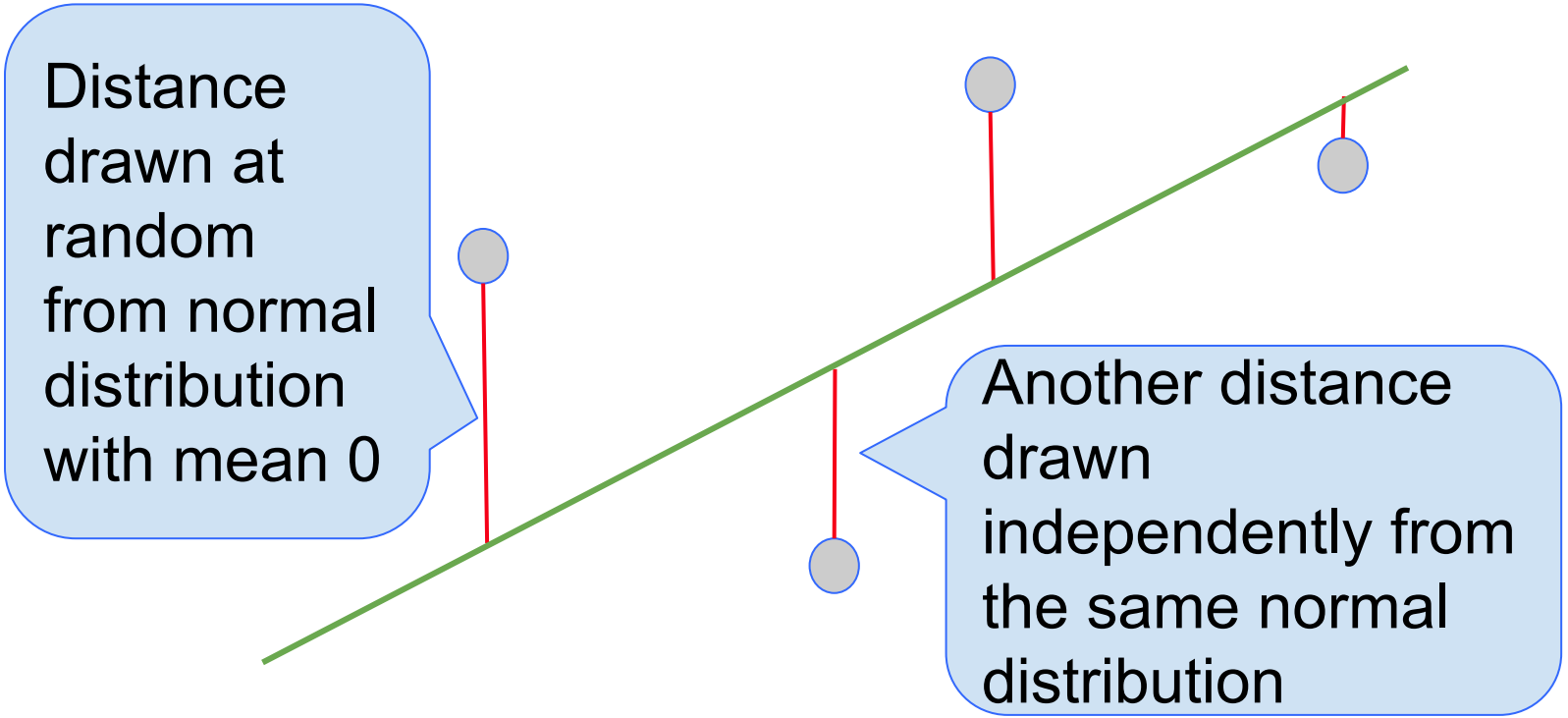
Discussion Questions

How would we adjust our line…

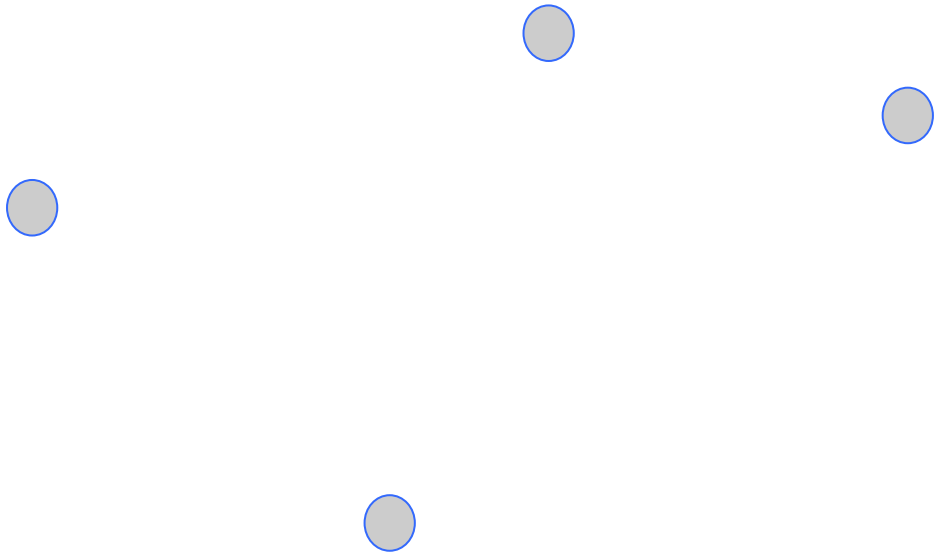
- if the average residual were 10?
- if the residuals were positively correlated with x?
- if the residuals were above 0 in the middle and below 0 on the left and right?

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 “true value” of the response *y* at a given value of *x* is the **height of the true line** at *x*
- We can’t see the true line, so we have to estimate this height
- The regression line is most likely close to true line
- Given a new value of *x*, predict y by finding the point on the regression line at that *x*

(Demo)

Prediction Interval

- **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. This is our range of predictions of *y*.
- It is an approximate **95% confidence interval for the height of the true line** at *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*.

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.

(Demo)