DATA 8

Summer 2018

Lecture 30

Regression Inference

Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Contributions by Fahad Kamran (fhdkmrn@berkeley.edu) and Vinitra Swamy (vinitra@berkeley.edu)

Announcements

Review

Residual Plot

A scatter diagram of residuals

- Should look like an unassociated blob for linear relations
- But will show patterns for non-linear relations
- Used to check whether linear regression is appropriate

Fitted Values and Residuals

- SD of fitted values / SD of y = $|r|$
- SD of fitted values = $|r|$ * (SD of y)
- The average of residuals is always 0
- SD of residuals = $\sqrt{(1 - r^2)}$ SD of y

A Variance Decomposition

- Variance of fitted values

  ---------- = r^2

  Variance of y

- Variance of residuals

  ---------- = 1 - r^2

  Variance of y

Discussion Question

Midterm: Average 70, SD 10

Final: Average 60, SD 15

$r = 0.6$

Fill in the blank:

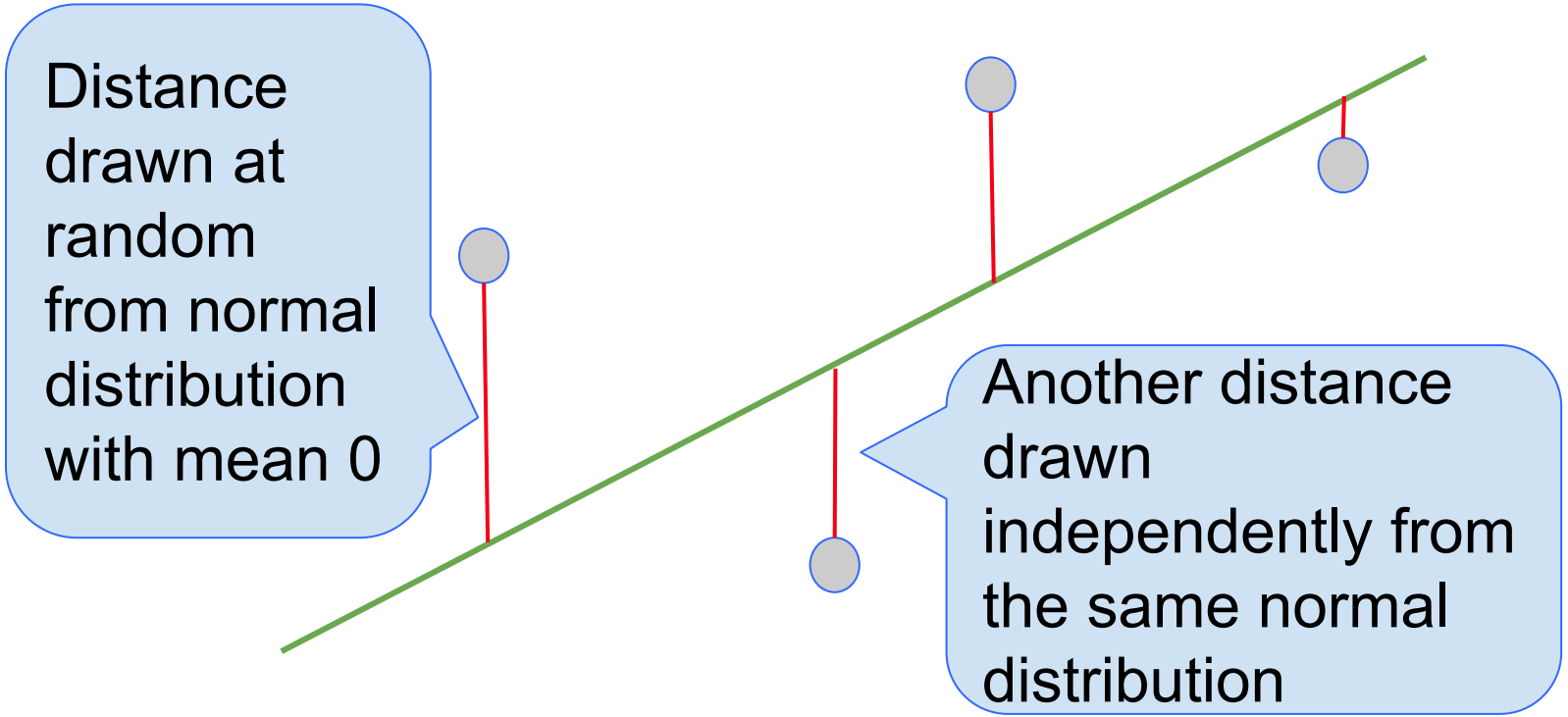
For at least 75% of the students, the regression estimate of final score based on midterm score will be correct to within _____ points.

Discussion Question

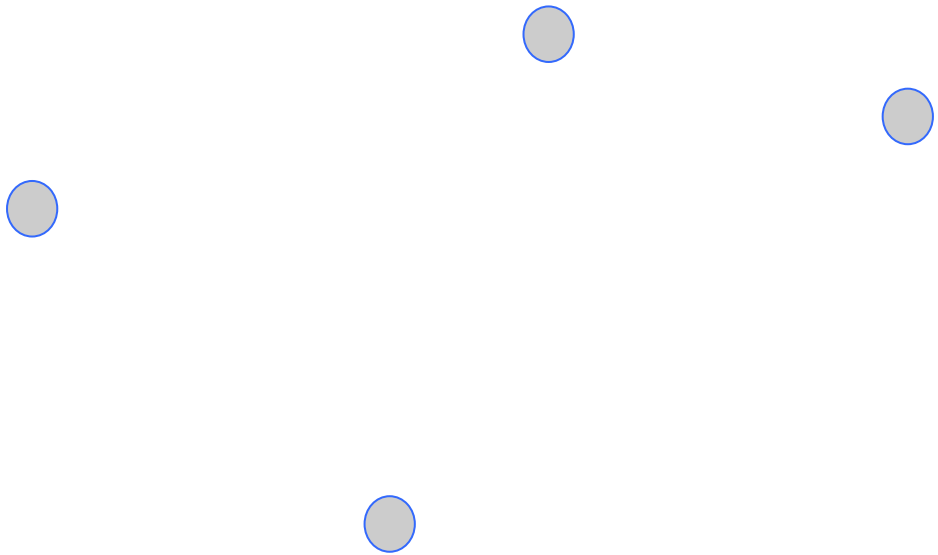
- On average, residuals are 0 so the regression is correct (on average)
- How far are we off?
  - Distribution of residuals
  - At-least 75% of data is within 2 SDs of the mean each direction -- Chebyshev's
- SD of residuals = $15 * \sqrt{(1 - .6^2)} = 12$
- On average correct, at least 75% of data is within 24 points

Regression Model

A “Model”: Signal + Noise



What We Get to See



(Demo)

Prediction Variability

Regression Prediction

- **If the data come from the regression model,**
- **and if the sample is large, then:**

- The regression line is close to the true line
- Given a new value of *x*, predict *y* by finding the point on the regression line at that *x*

(Demo)

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 height of the true line at *y*.

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*.

Discussion Question

True or **False**

Our goal of this method is to estimate what our regression line predicts, on average, what our y-value should be given a specific x-value.

Discussion Question

False

Our goal of this method is to estimate what **the true line** predicts, on average, what our y-value should be given a specific x-value.

We have the value for our regression line -- our regression line was based off of our sample.

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)

Discussion Question

True or **False**

Our goal of this method is to estimate what the slope of the regression line is.

Discussion Question

False

Our goal of this method is to estimate what the slope of **the true line** is.

We have the slope of our regression line -- it is calculated based on our sample.

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?

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)