YData: An Introduction to Data Science

Lecture 33: Regression Inference

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- Homework assignment 10 is due tomorrow, April 22
- This Friday (April 23) is a break day – no class
- Project 3 checkpoint is due Monday, April 26
- Project 3 itself is due Friday, April 30
- The lowest project grade will be dropped (project checkpoint grade is included in project grade)
- Homework assignment 11 will be published on Monday (April 26) and will be due Thursday, May 6

Review. SD of Fitted Values

$$\frac{\text{SD of fitted values}}{\text{SD of y}} = |r|$$

$$\text{SD of fitted values} = |r| * (\text{SD of y})$$

Review. Residual Average and SD

- The average of residuals is always 0
- $\frac{\text{Variance of residuals}}{\text{Variance of y}} = 1 - r^2$
- SD of residuals $= \sqrt{1 - r^2}$ SD of y

(DEMO)

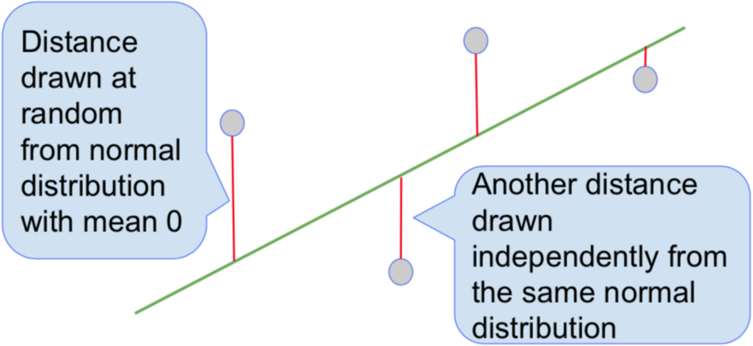
Discussion Question

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.

Regression Model

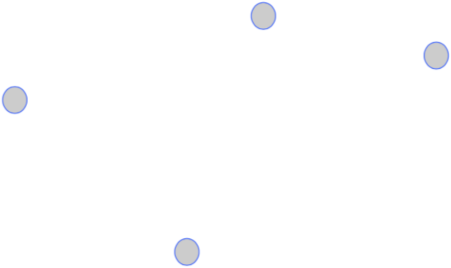
A “Model”: Signal + Noise



Distance drawn at random from normal distribution with mean 0

Another distance drawn independently from the same normal distribution

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.

(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.

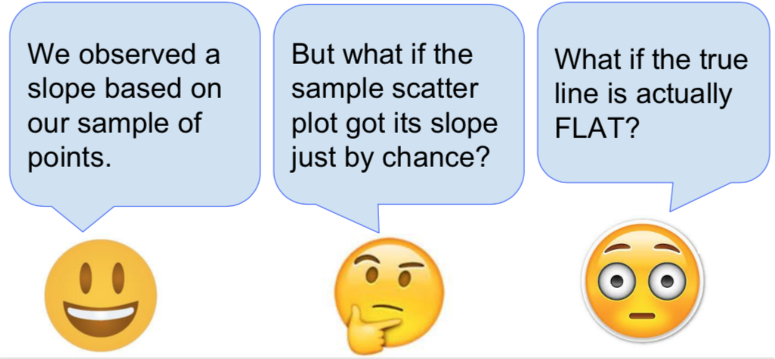
The True Slope

Confidence Interval for True Slope

- **Bootstrap the scatter plot**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat 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 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?

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)