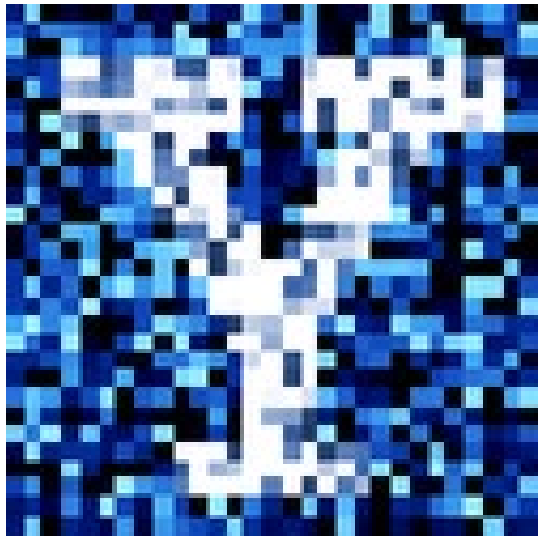
YData: Introduction to Data Science



Lecture 33: regression inference

Overview

Review: minimizing the RMSE

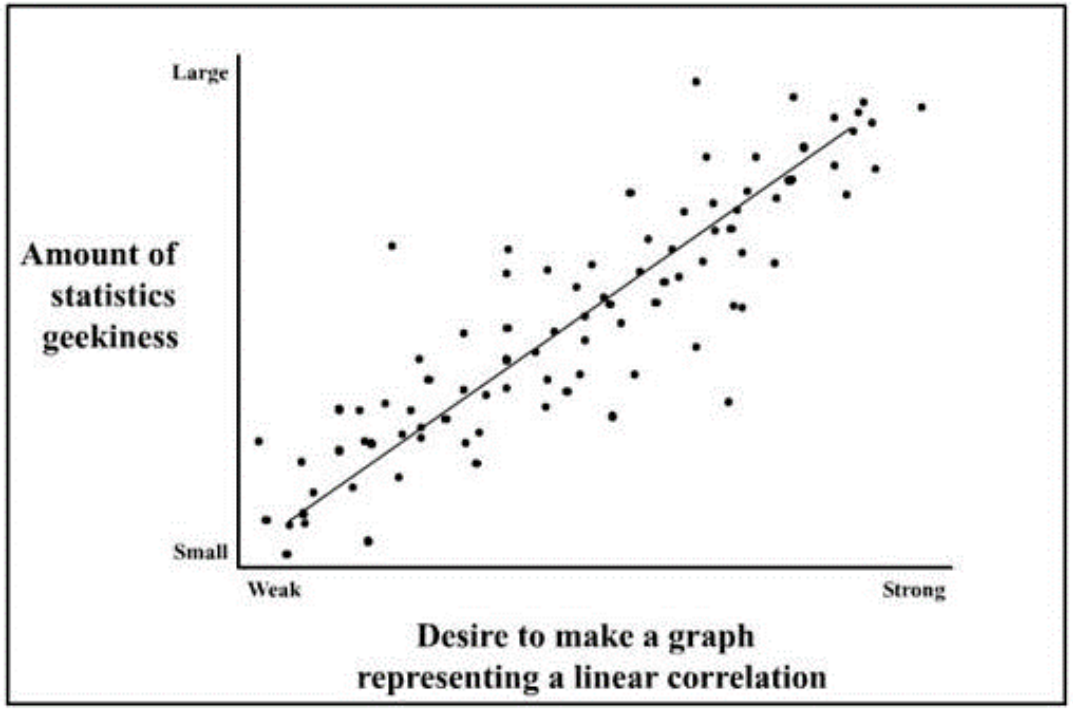
Regression diagnostic plots

Polynomial regression

The correlation coefficient as the proportion of variability explained

If there is time: inference for regression

- Confidence intervals
- Hypothesis tests



Review: minimizing the RMSE

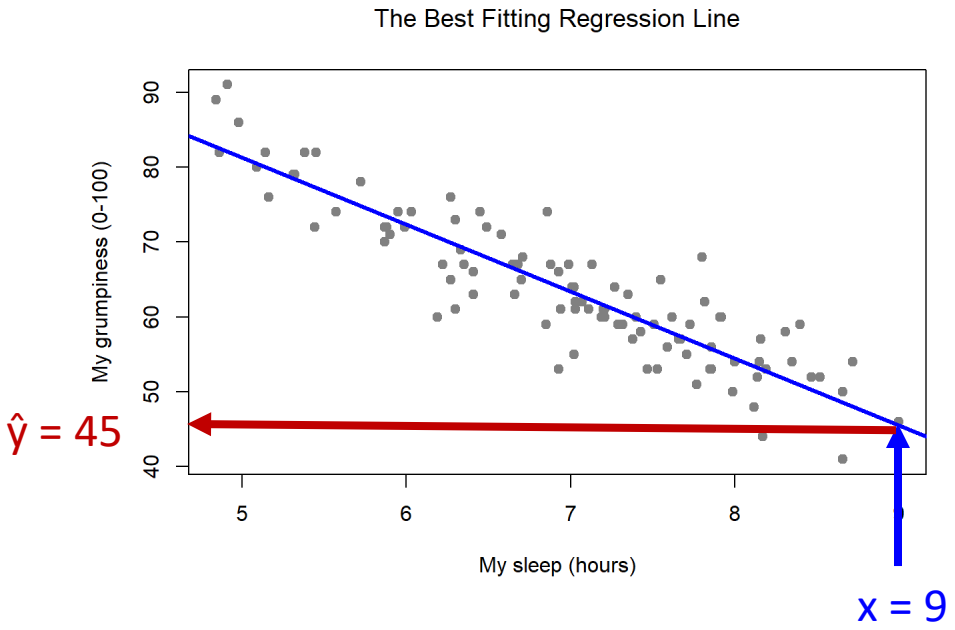
Regression

Regression is method of using one variable ***x*** *to predict* the value of a second variable ***y***

- i.e., $\hat{y} = f(x)$

In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = \text{slope} \cdot x + \text{intercept}$$



Least squares estimation

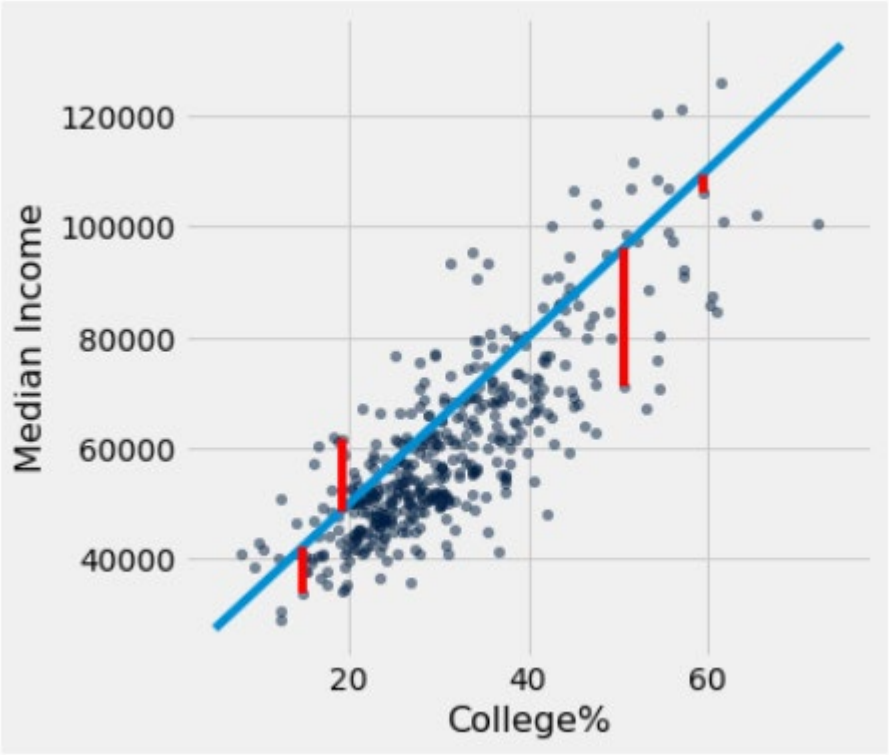
residual = actual value - estimate

$e_i = y_i - \hat{y}_i$

Typically, some errors are positive and some negative

To measure the rough size of the errors we calculate the root mean square error (RMSE):

- Square the errors to eliminate cancellation
- Take the mean of the squared errors
- Take the square root to fix the units



$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

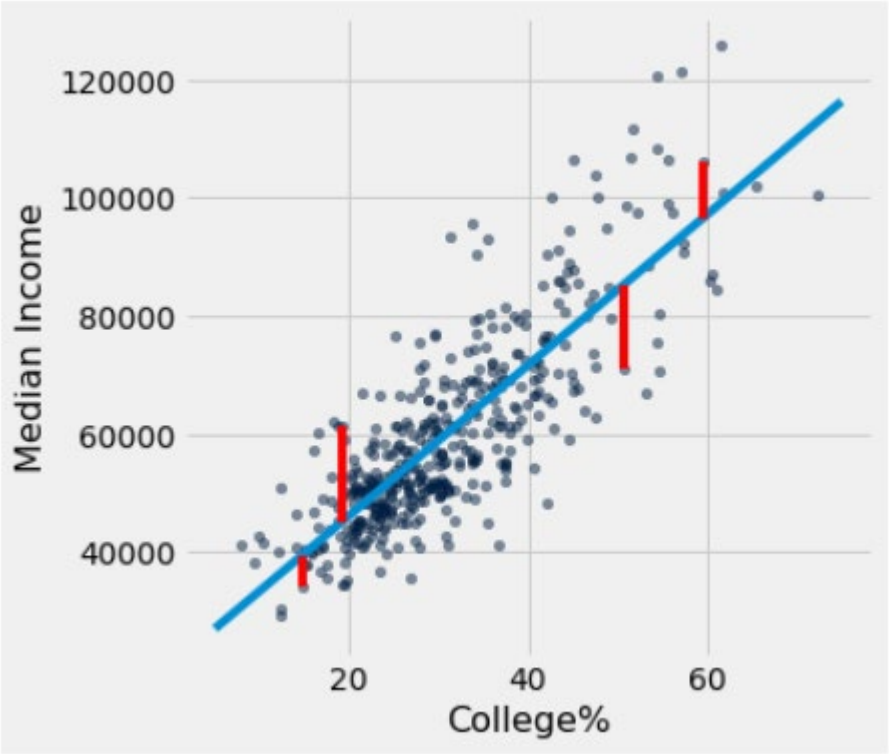
Least Squares Line

Minimizes the root mean squared error (RMSE) among all lines

- Equivalently, minimizes the mean squared error (MSE) among all lines

Names:

- “Best fit" line
- Least squares line
- Regression line



Numerical optimization

Numerical minimization is approximate but effective

Much of machine learning is based on numerical minimization

If the function mse(a, b) returns the MSE of estimation using the line "estimate = ax + b"

- then minimize(mse) returns array $[a_0, b_0]$
- a_0 is the slope and b_0 the intercept of the line that minimizes the MSE among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

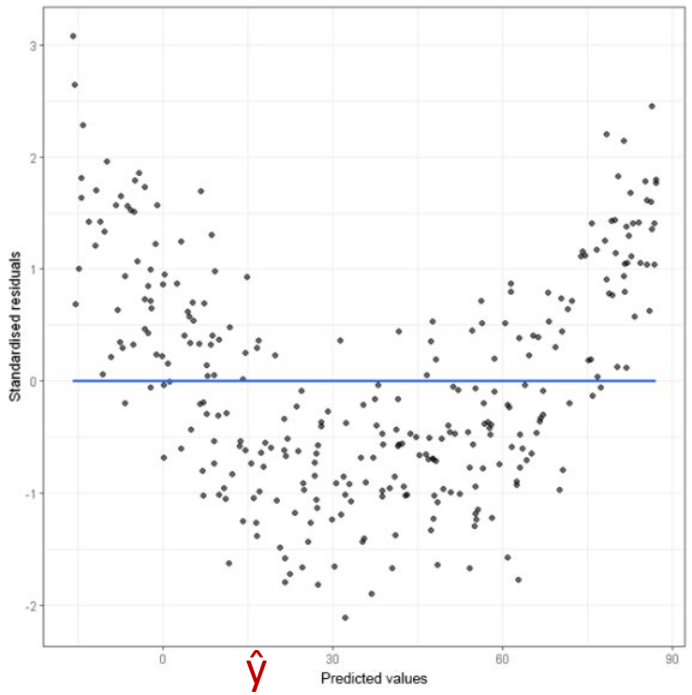
Let's explore this in Jupyter!

Regression diagnostics

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



Let's explore this in Jupyter!

Polynomial regression

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$\hat{y} = a + b \cdot x + c \cdot (x)^2 + d \cdot (x)^3$$

Need to find the coefficients: a, b, c, d

Still a linear equation but non-linear in original predictors

Polynomial regression

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$\text{child} = a + b \cdot \text{MidParent} + c \cdot (\text{MidParent})^2 + d \cdot (\text{MidParent})^3$$

Need to find the coefficients: a, b, c, d

Still a linear equation but non-linear in original predictors

Let's explore this in Jupyter!

The correlation coefficient as a measure of clustering around the regression line

Correlation revisited

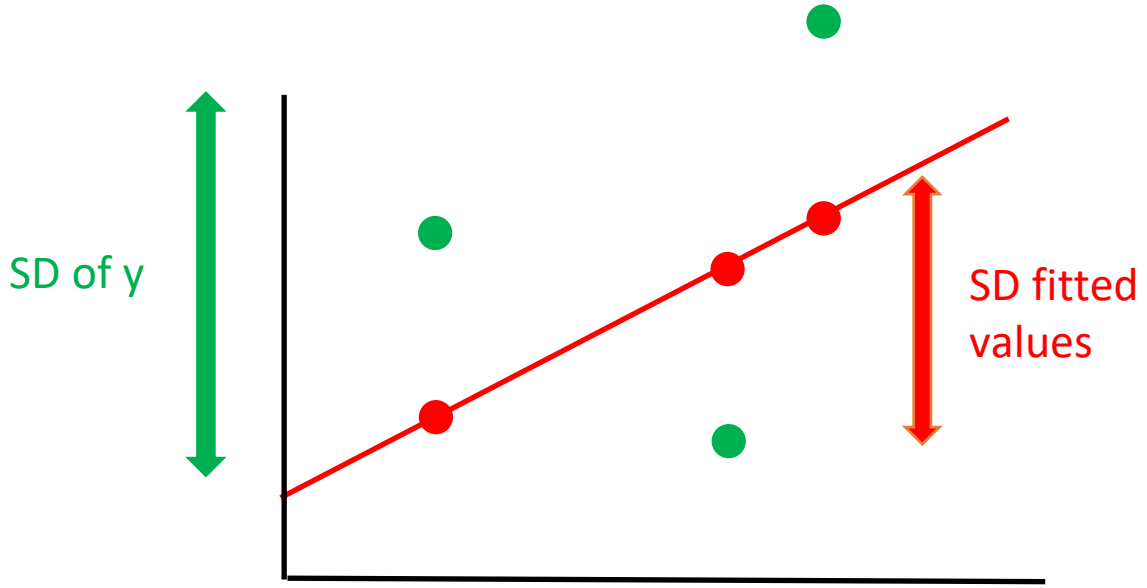
“The correlation measures how clustered the points are about a straight line."

We can now quantify this statement?

Analysis of variance for regression

There are relationships between:

- The overall variability of a response variable y
- The variability of the fitted values
- The correlation coefficient



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

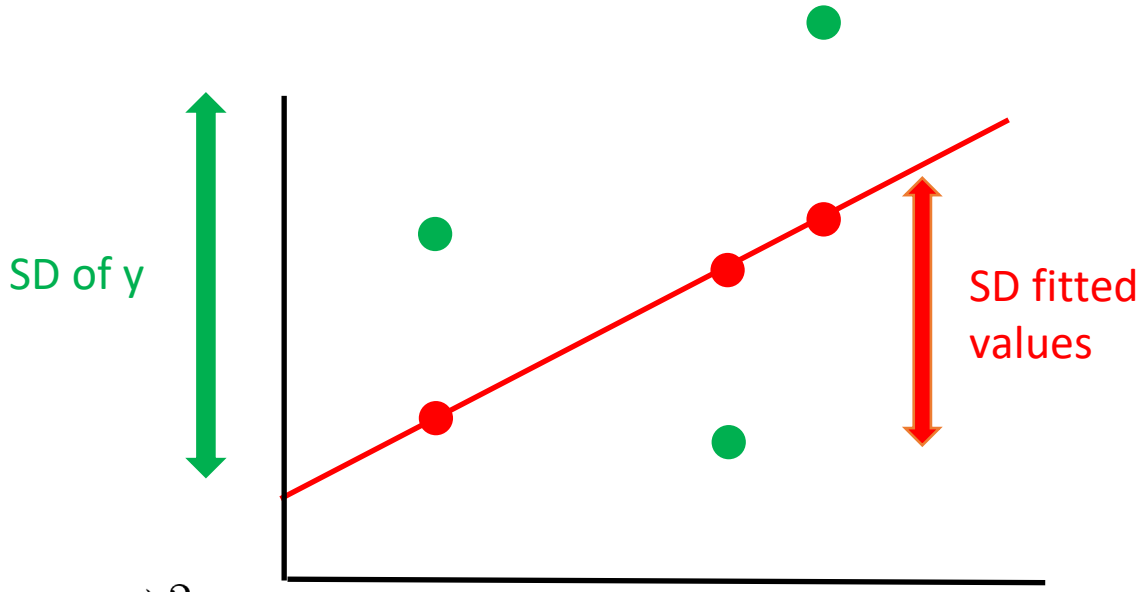
$$(\text{SD y})^2 = (\text{residuals})^2 + (\text{SD fitted values})^2$$

Just requires a lot of algebra to show this

Implications

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

The proportion of the total variability (SD y) accounted for by the regression line is |r|



$$(\text{SD y})^2 = (\text{SD residuals})^2 + (\text{SD fitted values})^2$$

The more variability accounted for by the regression line (larger slope) the less variability left in the residuals

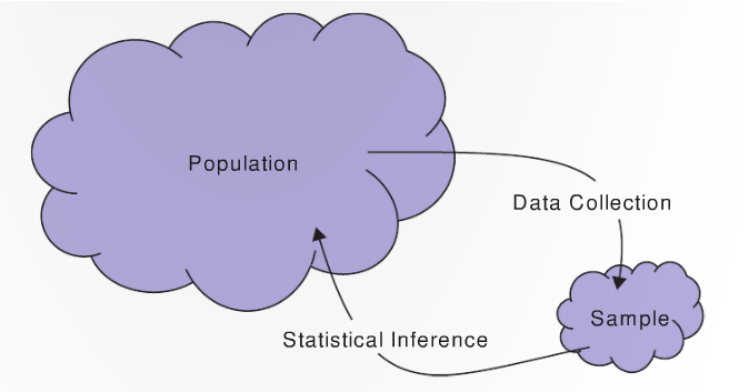
Let's explore this in Jupyter!

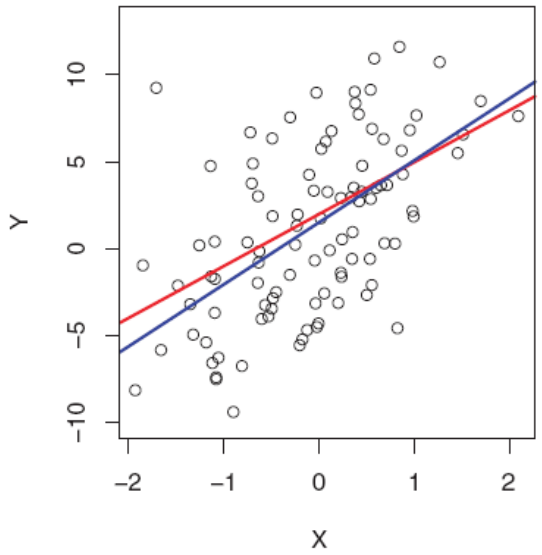
Inference for regression

Inference for regression

A regression line from a sample of data is only an estimate of the true population regression line

- i.e., if we had a different sample of data, we would get a different regression line





Population: regression lines

Sample estimates:
"lines of best fit" based on a sample of data

Q: How accurate is our "line of best fit" from a sample at capturing the true relationship?

Let's explore this in Jupyter!

Confidence interval for linear regression

Confidence interval for regression lines

We can use the bootstrap to create confidence intervals for:

- The regression slope
- The regression intercept
- The whole regression line

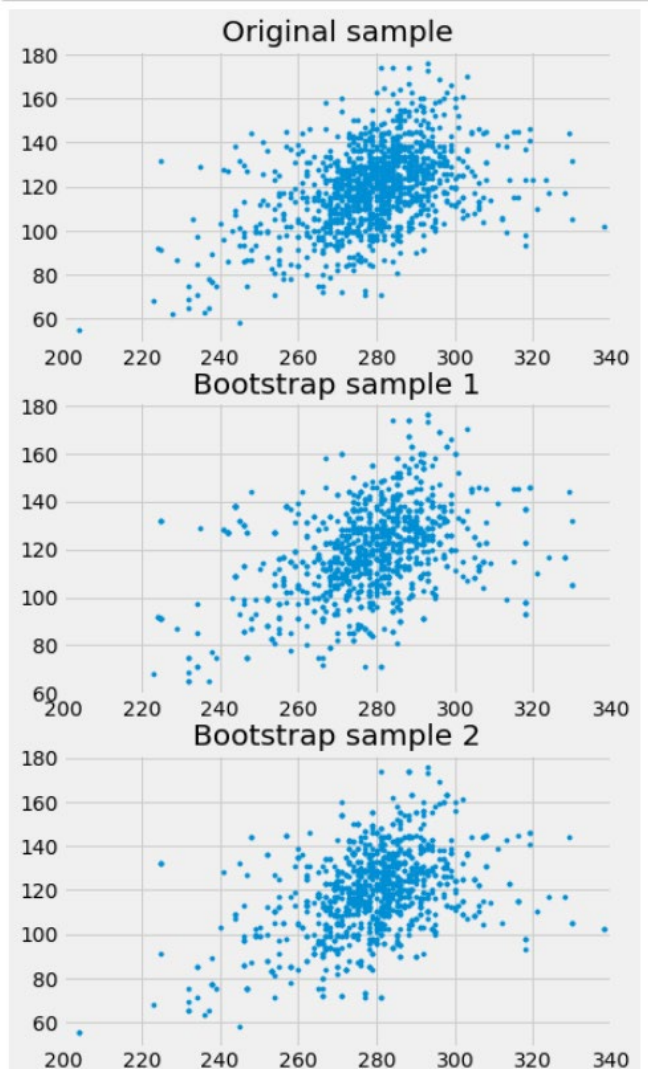
To run the bootstrap we need to:

- Resample our data with replacement
- Fit a regression line to the sample of data
- Save the regression slope and intercept
- Repeat many times

To create a 95% confidence interval:

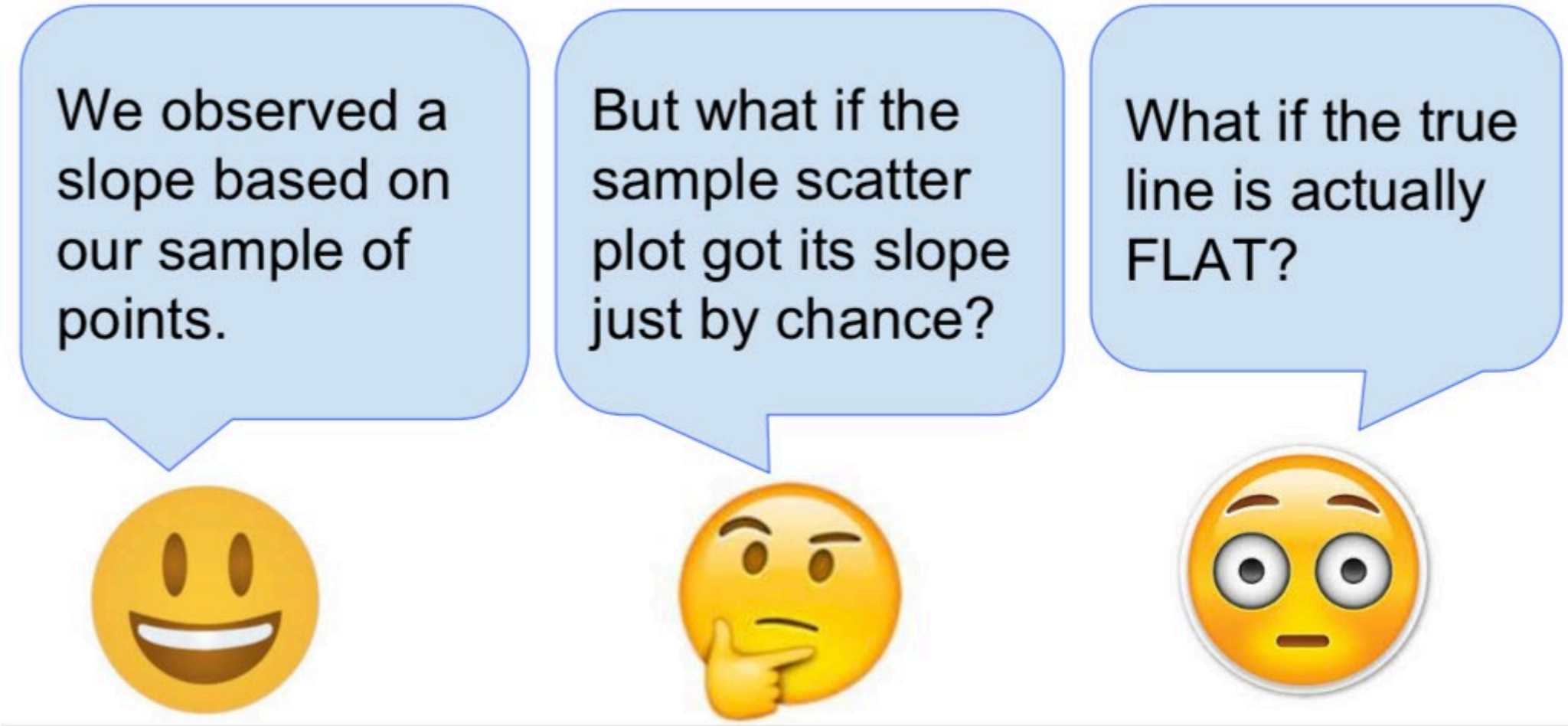
- Get the "middle 95%" of our regression slopes (or intercepts)

Let's explore this in Jupyter!



Hypothesis tests for linear regression

Rain on the regression parade



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

Let's explore this in Jupyter!