LECTURE 3

CEIC6789 NOTES

LINEAR REGRESSION

Linear approximation of a causal relationship between two or more variables

REGRESSION PROCESS

- Get sample data
- Design a linear model that fits the data
- Make predictions using the linear model

$$\hat{y}$$

Predicted

$$x_1, x_2, x_3 \dots$$

Predictors

$$\hat{y} = f(x_1, x_2, x_3 \dots)$$

Linear regression model

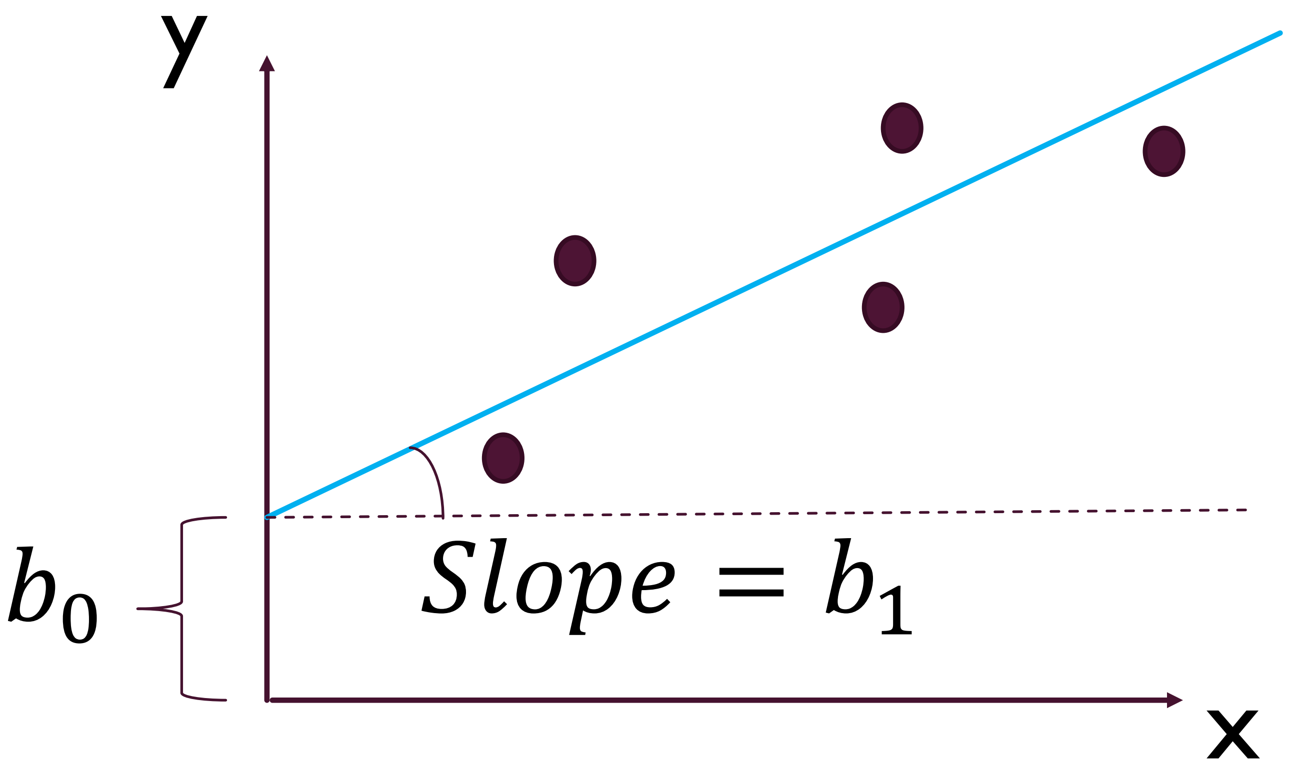
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$

SIMPLE LINEAR REGRESSION

$$\hat{y} = b_0 + b_1 x_1$$

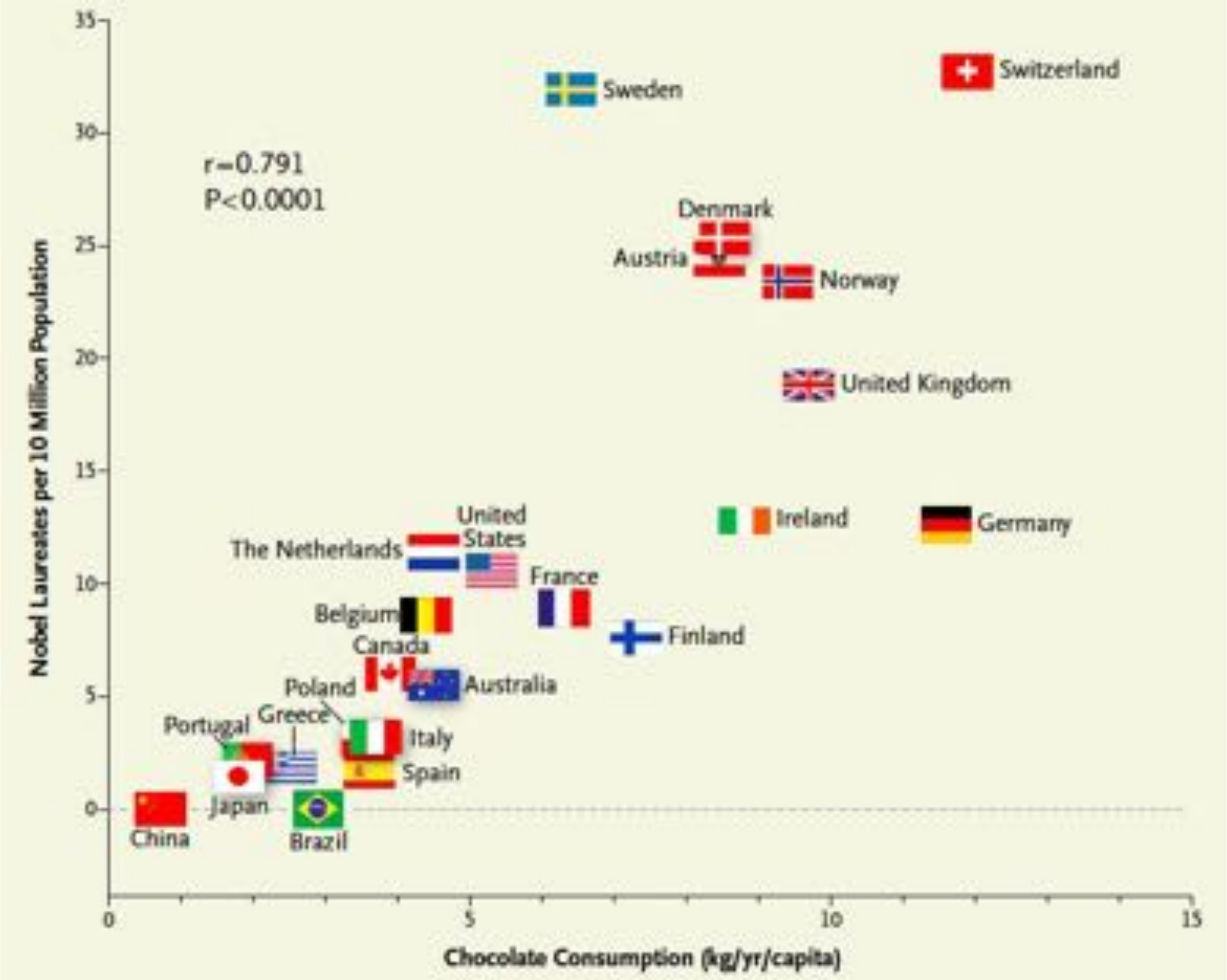
Profit (M$) = 1.3 M$ + 10 * advertising time (in months)

GEOMETRIC REPRESENTATION OF REGRESSION



CORRELATION VS. REGRESSION

- Correlation does not imply causation
- Correlation: No cause and effect
- Regression: Cause and effect



HYDROGEN FUEL FROM WATER

- EFFICIENCY INCREASES WITH CATALYST SIZE
- DATASET: EFFICIENCY AND CATALYST SIZE
- BUILD A LINEAR REGRESSION MODEL

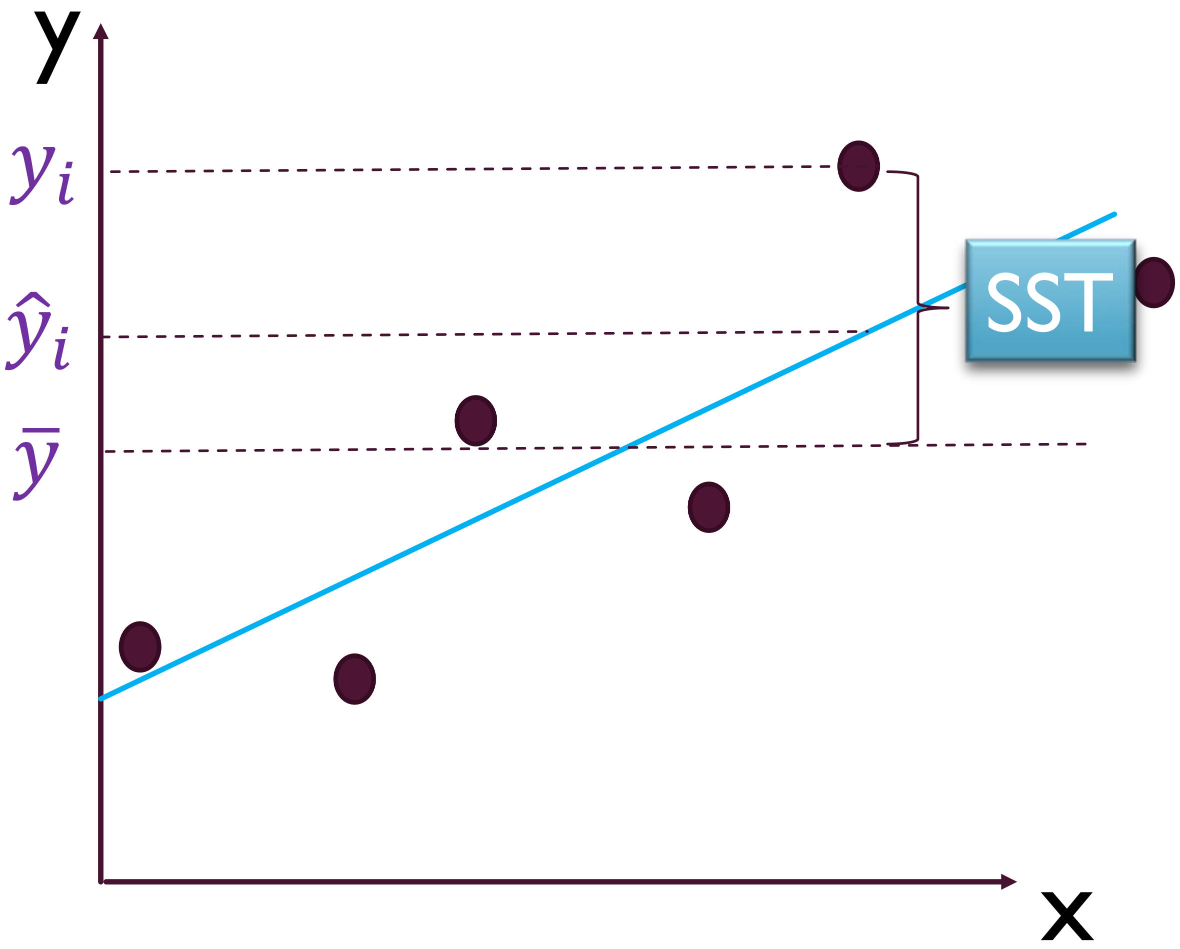
TERMS TO DECOMPOSE VARIABILITY IN REGRESSION

- Sum of Squares Total
- Sum of Squares Regression
- Sum of Squares Error

SUM OF SQUARES TOTAL (SST)

Measures the total variability of the dataset

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$



SUM OF SQUARES REGRESSION (SSR)

Measures how well the regression line fits the data (the explained variability of the dataset)

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SUM OF SQUARES ERROR (SSE)

Measures the unexplained variability in the dataset

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

SST = SSR + SSE

Total variability = Explained variability by regression + Unexplained variability (error)

$$\text{R-squared or } R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

$$0 <= R^2 <= 1$$

- R^2
  - Engg. and physical sciences: 0.7 – 0.99
  - Economics, finance, psychology: ~ 0.2

MATHEMATICS

ORDINARY LEAST SQUARES (OLS)

Least squares $\rightarrow$ min. SSE

Lower SSE results in a better explanatory power of the model

$$S(b) = \text{min. SSE}$$

$$= \text{min.} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \text{min.} \sum_{i=1}^{n} (y_i - x_i^T b)^2$$

$$= \text{min.} \ (y_i - x_i^T b)^T \ (y_i - x_i^T b)$$

Please go through the following derivation in Wikipedia: https://en.wikipedia.org/wiki/Ordinary_least_squares#Matrix/vector_formulation.

You can then write a few lines of code using numpy arrays to obtain the coefficients as per the derivation. This activity will help you understand how numpy works and how to deal with arrays in python. So, do give it a try!

ACTIVITY

ORDINARY LEAST SQUARES

GENERALIZED LEAST SQUARES

MAXIMUM LIKELIHOOD ESTIMATION

BAYESIAN REGRESSION

GAUSSIAN PROCESS REGRESSION

AND SO ON…

REGRESSION METHODS

MULTIPLE LINEAR REGRESSION

WHY DO WE NEED MULTIPLE LINEAR REGRESSION?

$R^2 = 74.5\%$

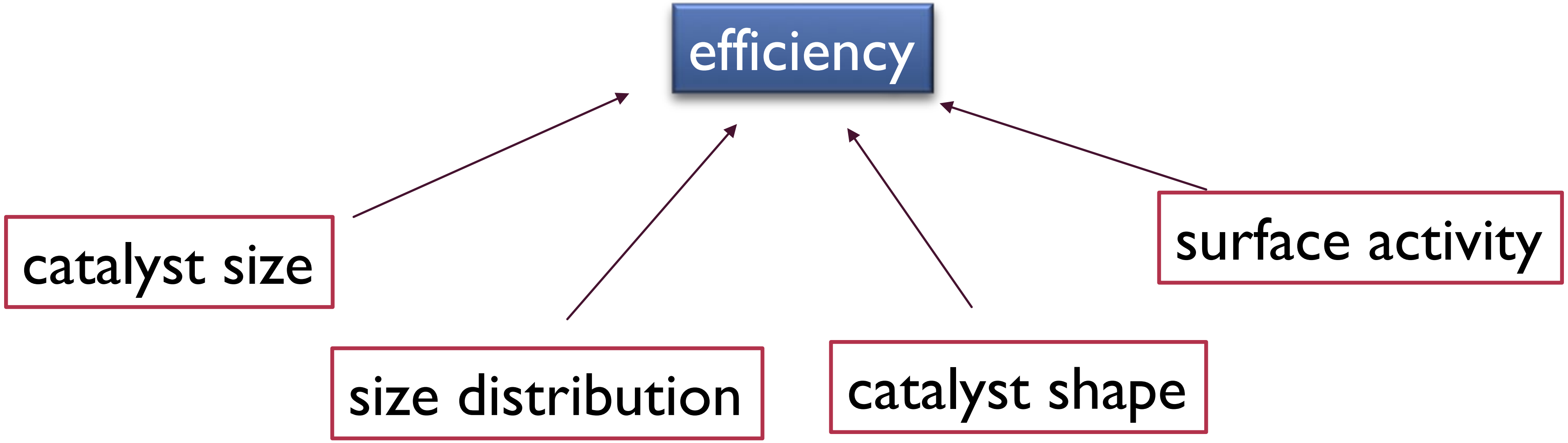
efficiency

catalyst size

- Catayst size explains 74.5% of the efficiency data
- Efficiency should be further explained by other factors

WHY DO WE NEED MULTIPLE LINEAR REGRESSION?

Efficiency = *f* (catayst size, size distribution, catalyst shape, surface activity)



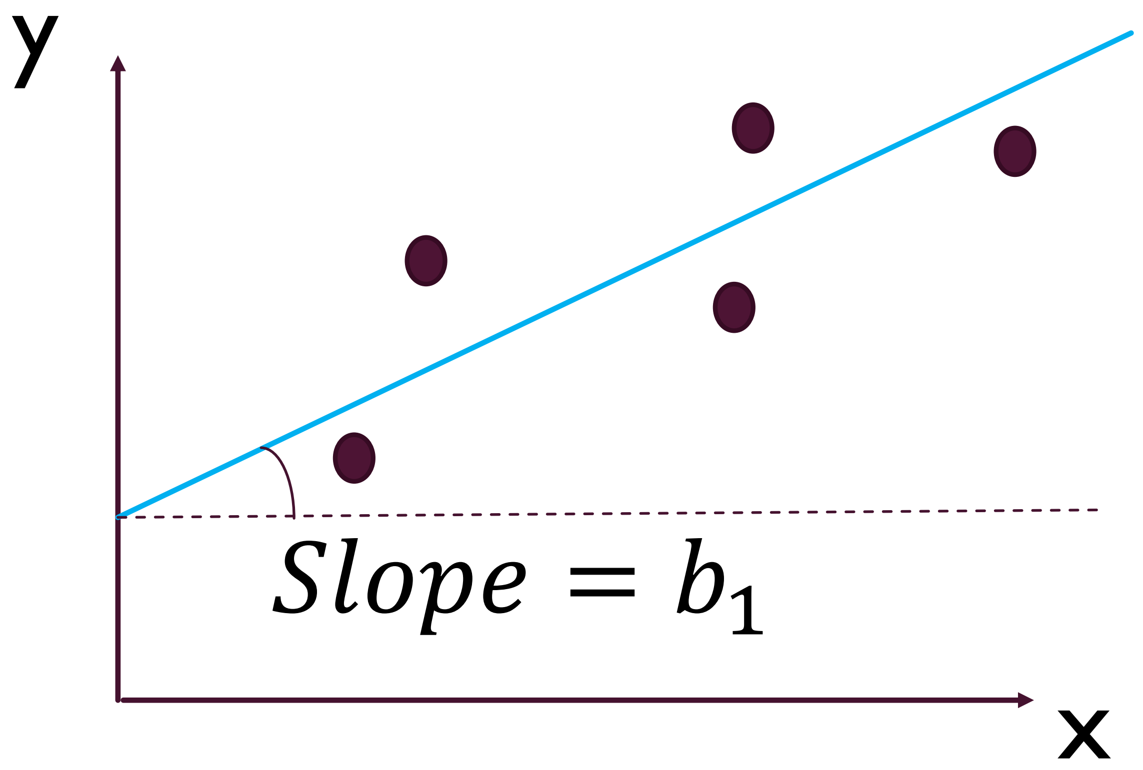
Simple Linear regression

$$\hat{y} = b_0 + b_1 x_1$$

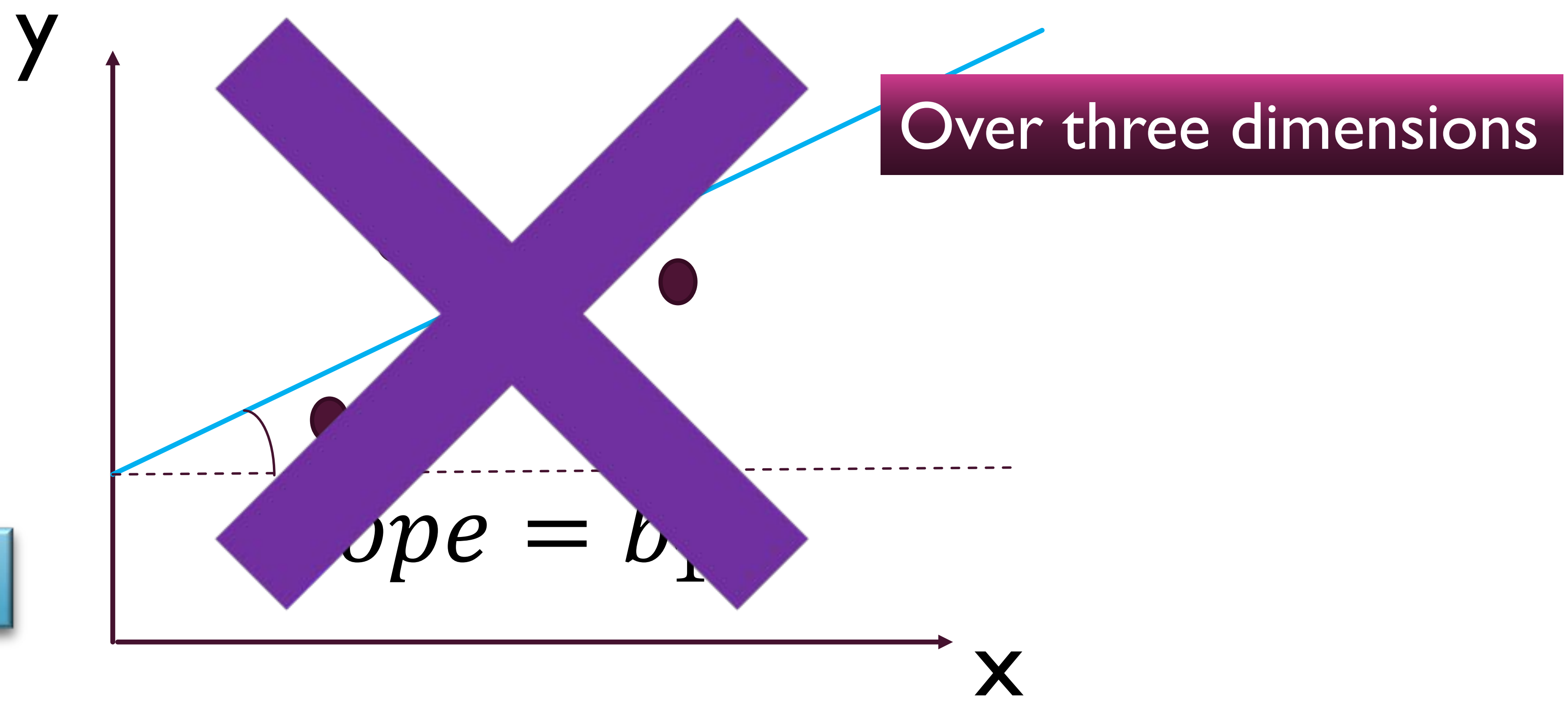
Multiple Linear regression

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

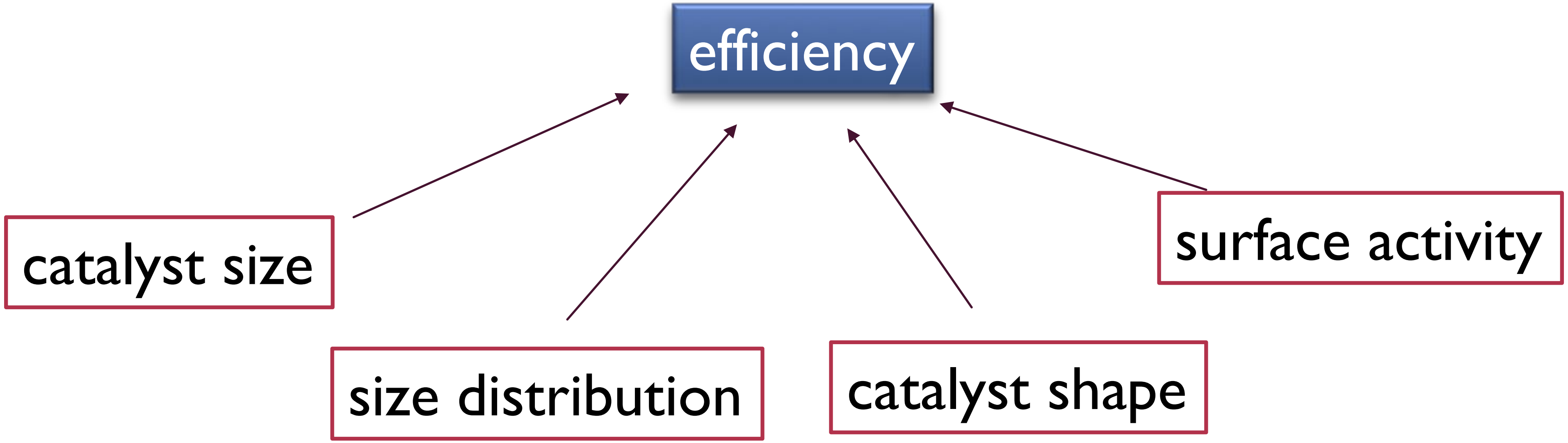
GEOMETRIC REPRESENTATION OF SIMPLE REGRESSION



GEOMETRIC REPRESENTATION OF MULTIPLE REGRESSION



Best fitting model

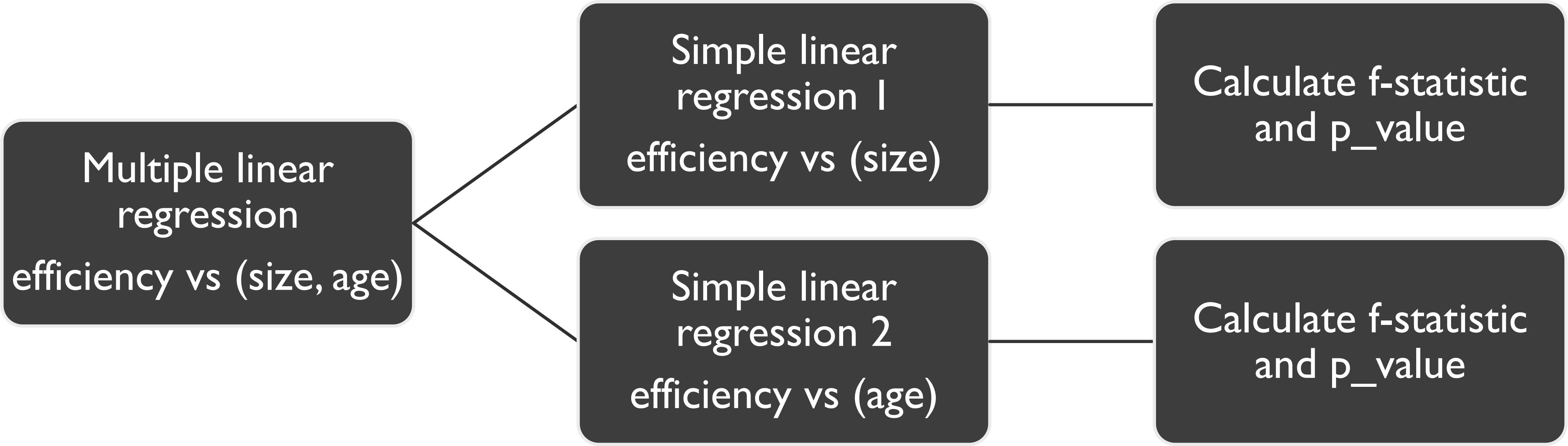


By adding more independent variables:

- Explanatory power increases by zero or more than zero
- R^2 remains the same or increases, we cannot lower it!

FEATURE SELECTION

f_regression from sklearn

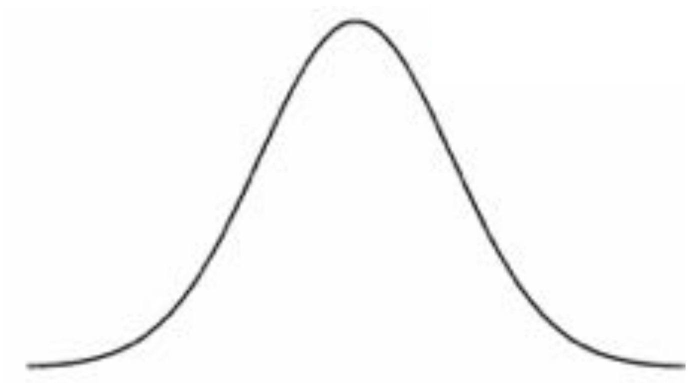


WHAT DOES F_REGRESSION DO?

WHAT IS F-STATISTIC?

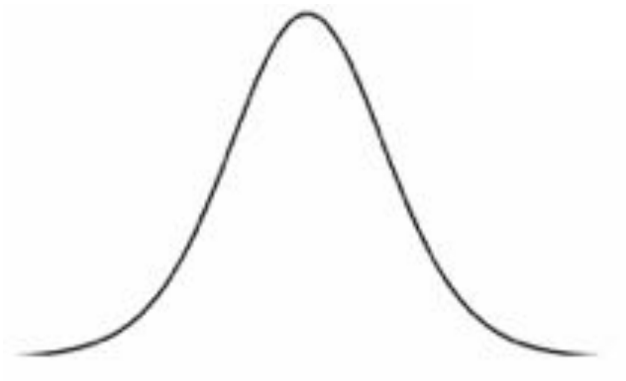
z-statistic

- Normal distribution



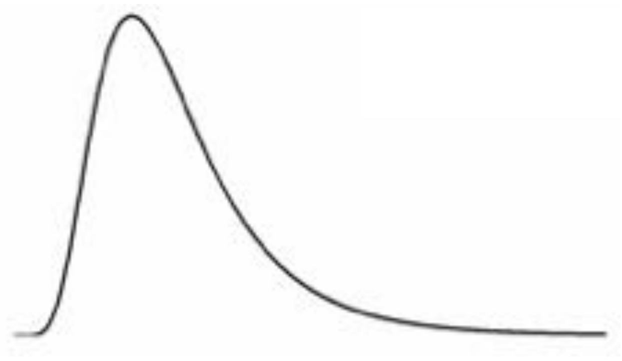
t-statistic

- Student's t-distribution



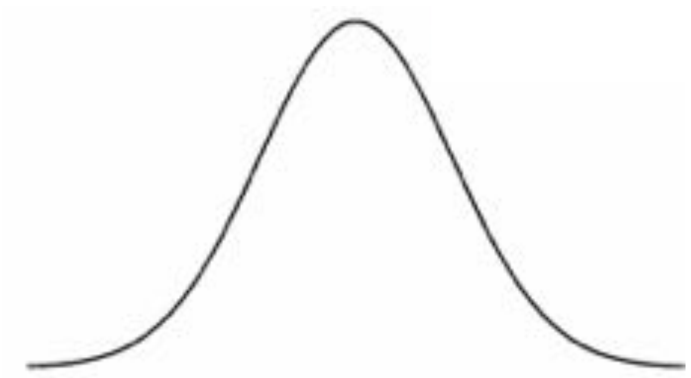
f-statistic

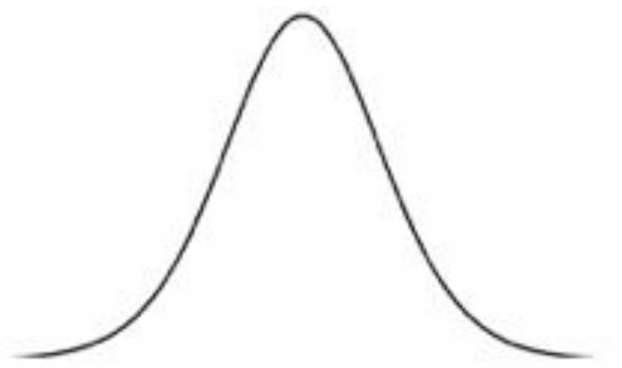
- f distribution



WHAT IS F-STATISTIC?

z-statistic	t-statistic	f-statistic
• Normal distribution	• Student's t-distribution	• f distribution





p value

p value < 0.05
feature useful

p value > 0.05
feature useless

ACTIVITY

Here's an activity: Familiarize yourself with the f-statistic and the concept of p-values.

You can read this informative webpage to understand the f-statistic: https://online.stat.psu.edu/stat501/lesson/6/6.2