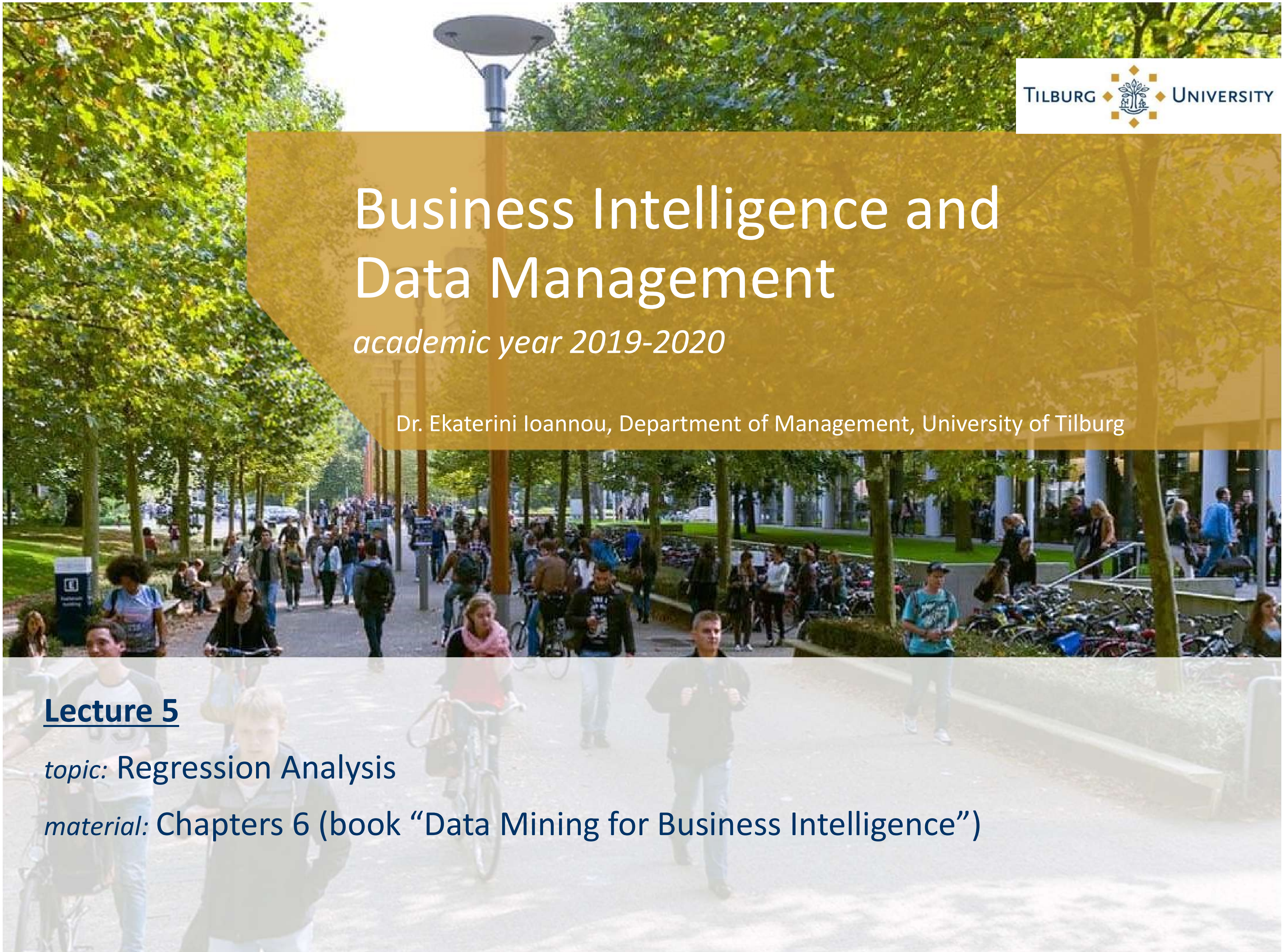


# Business Intelligence and Data Management

*academic year 2019-2020*

Dr. Ekaterini Ioannou, Department of Management, University of Tilburg

**Lecture 5**

*topic:* Regression Analysis

*material:* Chapters 6 (book “Data Mining for Business Intelligence”)

# Summary from previous lectures

**Data Mining:**

- Provides results useful for decision making
- Process involves multiple steps:

  Business understanding → Data preparation → Model building → Testing & Evaluation → Deployment
- Various options for the model:
  - Regression
  - Clustering
  - Decision trees
  - Association rules
  - etc.

# Summary of previous lectures / Agenda

| Date | | Lecture contents | Lecturer | Lab topics | Test |
|---|---|---|---|---|---|
| **Jan-27** | 1 | Intro. to BI+ Data Management | Caron | | |
| Jan-30 | | | | SQL-1 | 1 |
| **Jan-28** | 2 | Data warehousing | Caron | | |
| Feb-06 | | | | SQL-2 | 2 |
| **Feb-03** | 3 | OLAP business databases & dashboard | Caron | | |
| **Feb-13** | | | | SQL-3 & OLAP | 3a & 3b |
| **Feb-10** | 4 | Data mining introduction | Ioannou | | |
| | 5 | Regression models | Ioannou | | |
| **Feb-17** | 6 | Naïve Bayes | Ioannou | | |
| | 7 | k nearest neighbors | Ioannou | | |
| Feb-20 | | | | Bayes & neighbors | 4 |
| **Feb-27** | 8 | Performance measures | Ioannou | | |
| **Mar-02** | 9 | Decision trees | Ioannou | | |
| Mar-05 | | | | Dec. trees | 5 |
| **Mar-09** | 10 | Association rules | Ioannou | | |
| Mar-11,12&13 | | | | Ass. Rules | 6 |
| **Mar-16** | 11 | Clustering (+20 mins exam preparation) | Ioannou | | |
| Mar-19 | | | | Clustering | 7 |

Data Mining

# Motivation

**Data Mining:**
**Provides results useful for decision making**

- Examples:
  - Store X wants to know the amount of water bottles to place in the fridge during June
  - The tax office wants to detect if the family income relates to the overall monthly spending of families
  - The university wants to find out if reducing the lab sessions will result in lower course grades
  - The university wants to find out if the color of the student hair affects the final exam score

# Variables

- Store X wants to know the amount of water bottles to place in the fridge each week
- Variables are any measurement on the records
- Dependent variables (denoted as Y):
  - The ones we want to explain / predict
  - Their value depend on something else
  - Example? amount of water bottles to place for June
- Independent variables (denoted as X):
  - The variables that explain / justify the dependent ones
  - Example? mean temperature from past years with the expected temperature, i.e., forecast; stock control information for the stock of a new store; etc.

# Relationships

- Store X wants to know the amount of water bottles to place in the fridge each week
- Variables are any measurement on the records
- A relationship shows the variable association
- Examples?
  - Place more water bottles in the fridge during Summer
  - Special events near the store don't affect something
  - Etc.

# Definition

Regression Analysis: fit a relationship between a numerical outcome variable and a set of predictors

- Numerical outcome variable Y

  also called response, target, or dependent variable

- Set of predictors $X_1, X_2, \ldots, X_n$

  also referred to as independent variables, input variables, regressors, or covariates

# Definition

Regression Analysis: fit a relationship between a numerical outcome variable and a set of predictors

- Numerical outcome variable Y

  also called response, target, or dependent variable

- Set of predictors $X_1, X_2, \ldots, X_n$

  also referred to as independent variables, input variables, regressors, or covariates

Linear Regression Model

↳ arranged in, or extending, along a straight or nearly straight line

# Definition

Regression Analysis: fit a relationship between a numerical outcome variable and a set of predictors

- Numerical outcome variable Y

  also called response, target, or dependent variable

- Set of predictors $X_1, X_2, \ldots, X_n$

  also referred to as independent variables, input variables, regressors, or covariates

Single Linear Regression Model

↳ one independent predictor, i.e., single variable X

# Definition

Regression Analysis: fit a relationship between a numerical outcome variable and a set of predictors

- Numerical outcome variable Y

  also called response, target, or dependent variable
- Set of numeric predictors $X_1, X_2, \ldots, X_n$

  also referred to as independent variables, input variables, regressors, or covariates

Single Linear Regression Model

Multiple Linear Regression Model

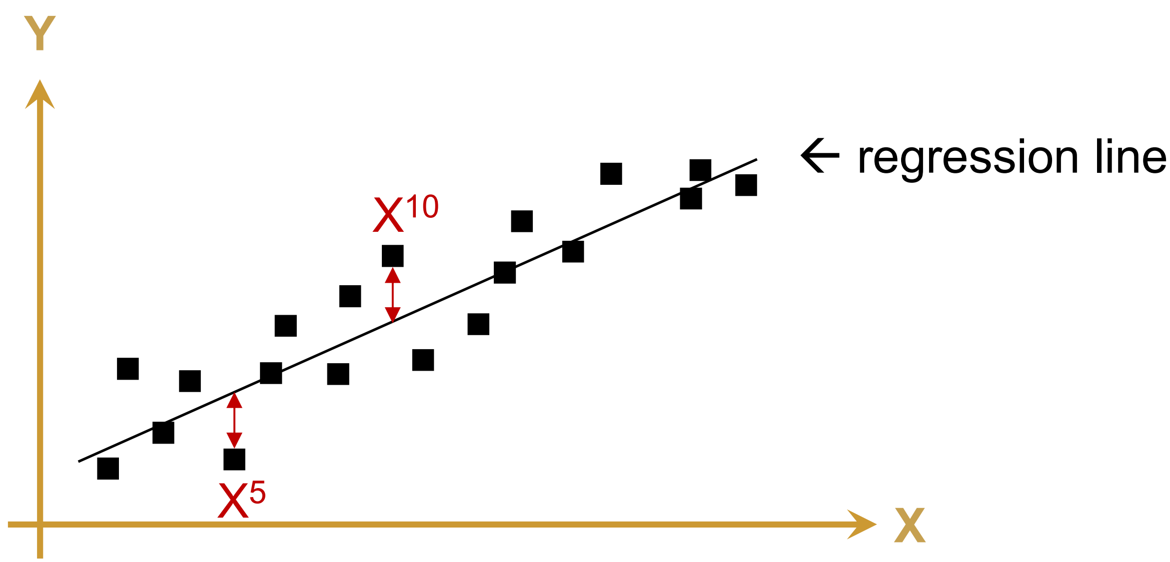
↳ two or more predictors, i.e., X1, X2, …

# Single / Multiple Linear Regression

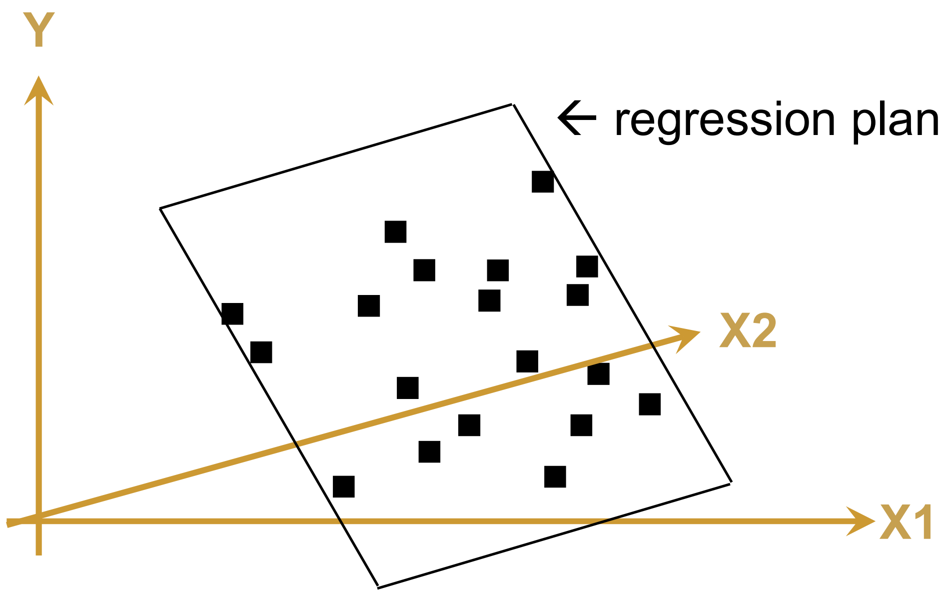
# Intuition for Single Linear Regression

- Fitting a linear relationship (i.e., equation) between:
  - a numerical outcome (i.e., variable Y) and
  - one predictor (i.e., variable X)
- Not all points go through the line

  → Error, i.e., deviation of data from line



# Intuition for Multiple Linear Regression




- Fitting a linear relationship (i.e., equation) between:
  - a numerical outcome (i.e., variable Y) and
  - two or more predictors (i.e., variables X1, X2, …)

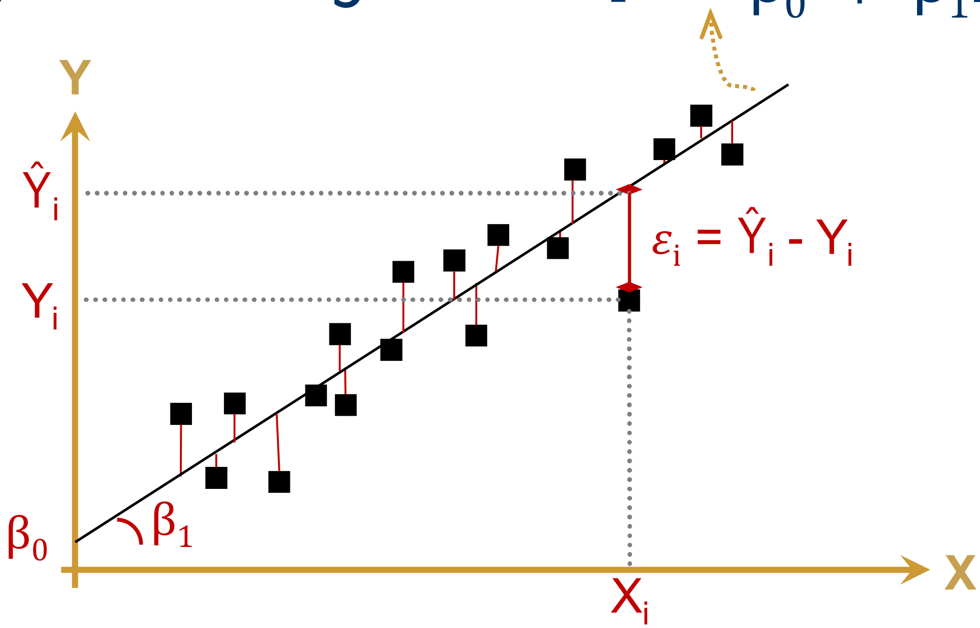
# Linear Regression Model

$$Y = \beta_0 + \beta_1 . X_1 + \beta_2 . X_2 + \ldots + \varepsilon$$

- **ε** is a term that represents the errors associated with the model
- $\boldsymbol{\beta_0, \beta_1, \ldots}$ are the regression coefficients (parameters)
- Values for attribute $X_1$ are $X_{11}, X_{12}, \ldots, X_{1i}, \ldots$ where $i$ is a "counter" representing the ith observation for the particular data collection

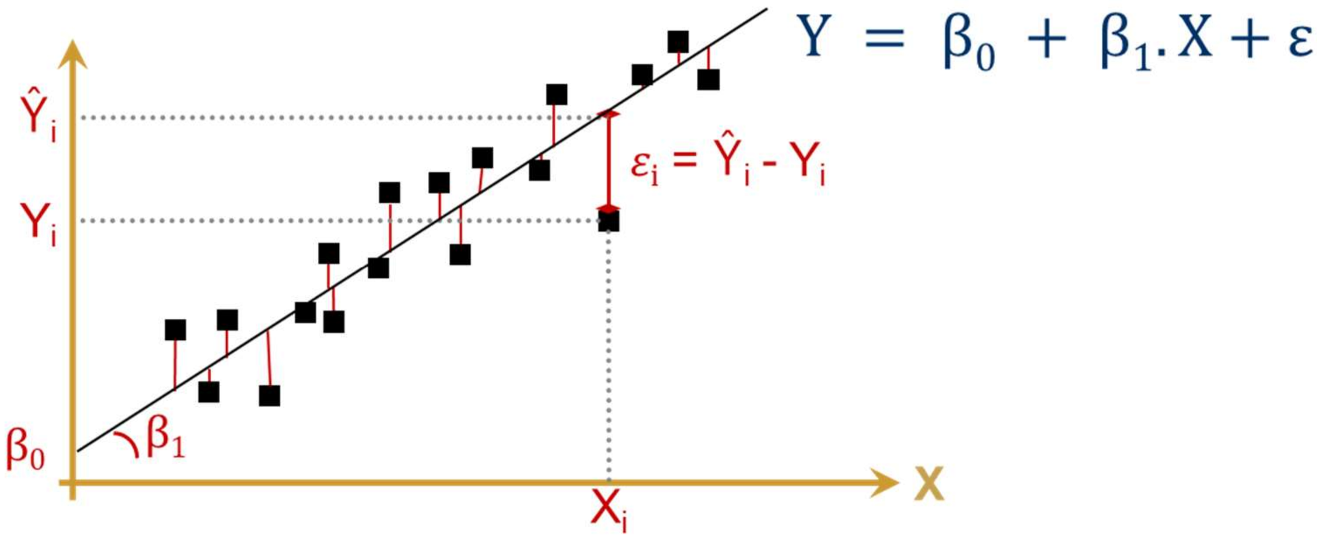
# Graphical Visualization

- Single Linear Regression: $Y = \beta_0 + \beta_1 . X + \varepsilon$



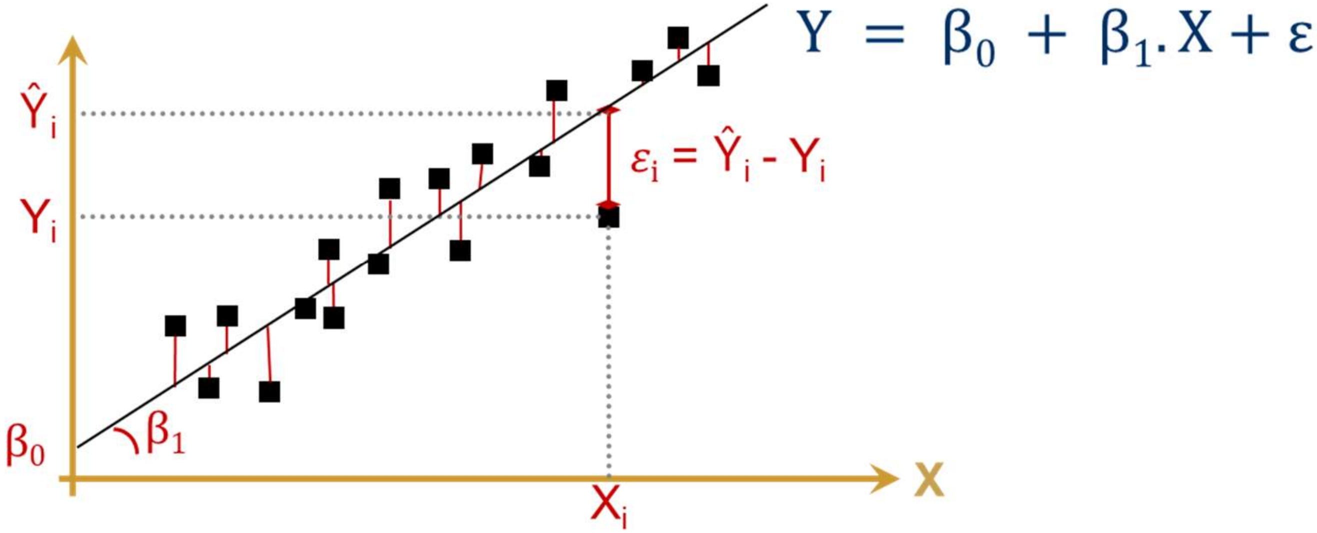

- $\beta_0$ the value of our dependent variable when the independent one is equal to zero, i.e., value of Y when X=0
- $\beta_1$ the slope of the regression line
- $\varepsilon_i$ corresponds the difference between the observations and the predictions of our model, i.e., distance from the line

# Ordinary Least Squares (OLS)




- Method for estimating the unknown parameters in a linear regression model
- It minimizes the errors associated with predicting values for the dependent variable Y
- It uses a least squares criterion because without square we would allow positive and negative deviations from the model to cancel each other out

# Ordinary Least Squares (OLS)




- Find $\beta_0$ and $\beta_1$ such that total error is minimal
- This criterion is very common and for example also used in Neural Network models
- Error for $Y_i$ is $\varepsilon_i$ and the value of $\varepsilon_i$ is equal to $\hat{Y}_i$ - $Y_i$

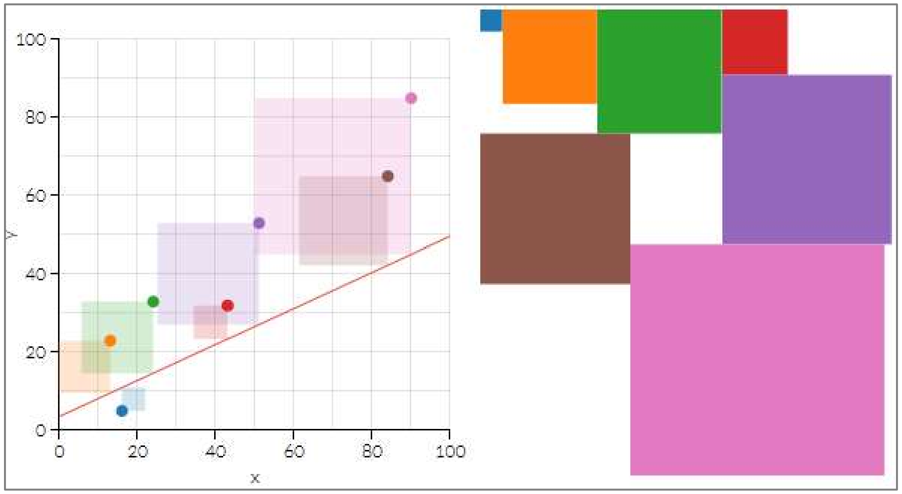
$$\Rightarrow \min \sum_{i=1}^{n} (\varepsilon_i)^2 = \min \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2$$

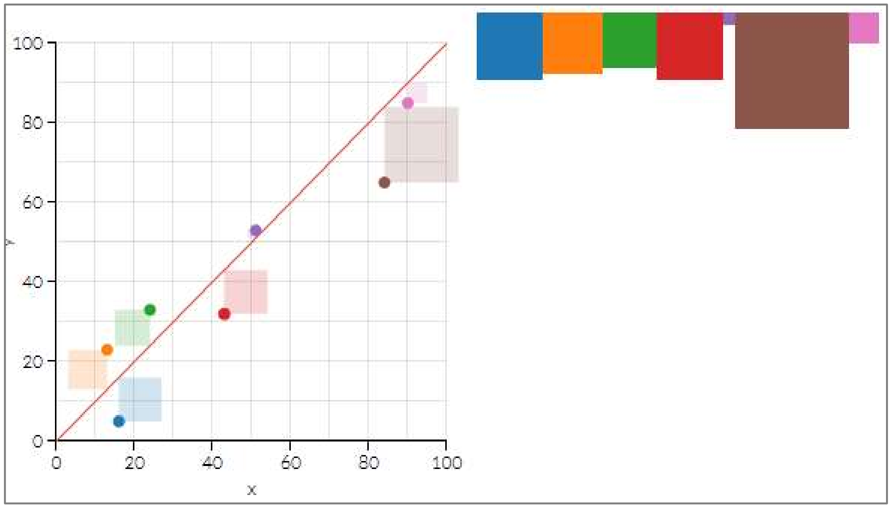
$$= \min \sum_{i=1}^{n} ( Y_i - \beta_0 - \beta_1 Xi )^2$$

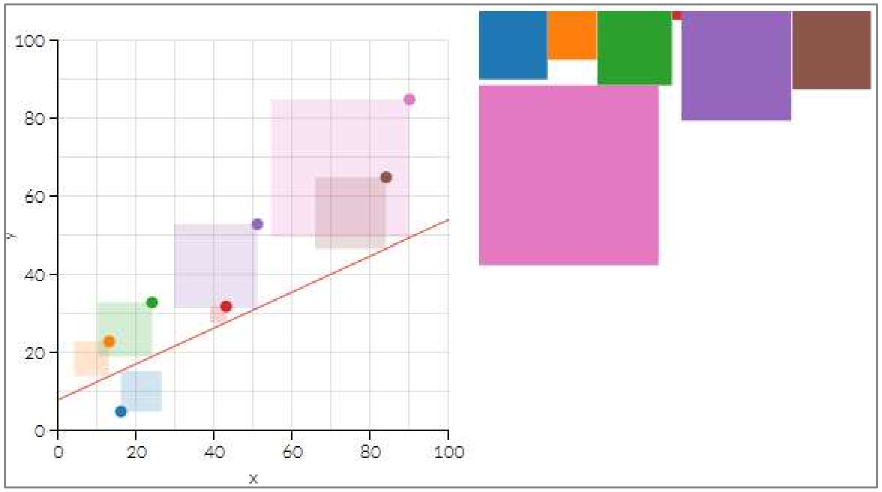
# Graphical Visualization

OLS: choose betas so that the total area of all the squares is as small as possible

$$\sum_{i=1}^{n}(\varepsilon_i)^2 = \min \sum_{i=1}^{n} ( Y_i - \beta_0 - \beta_1 Xi )^2$$




Plots taken from here:

http://setosa.io/ev/ordinary-least-squares-regression/

the site contains interactive plots to understand the concepts!

# Explanatory vs. Predictive

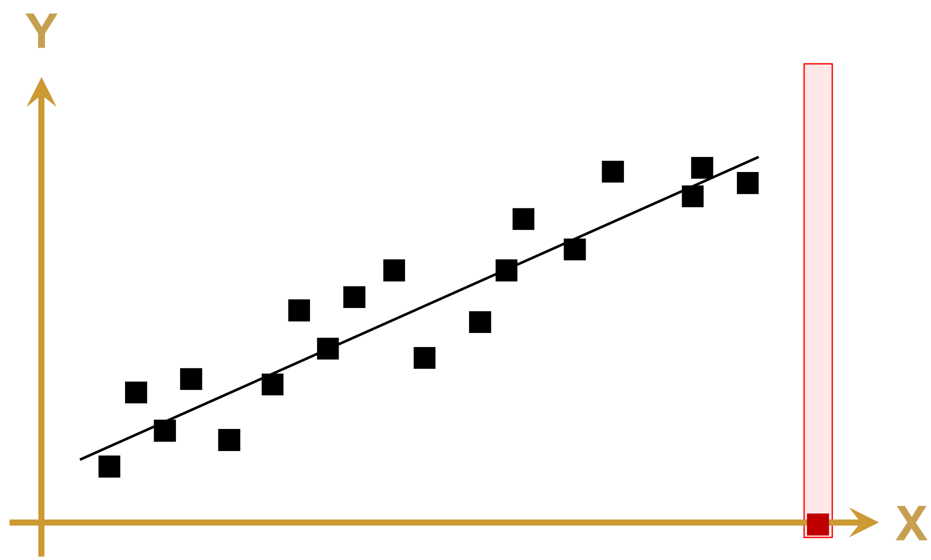
# Objectives for single/multiple regression

Two popular but different objectives behind fitting a regression model:

1. Predictive
   - Detect the outcome value for new records, given their input values

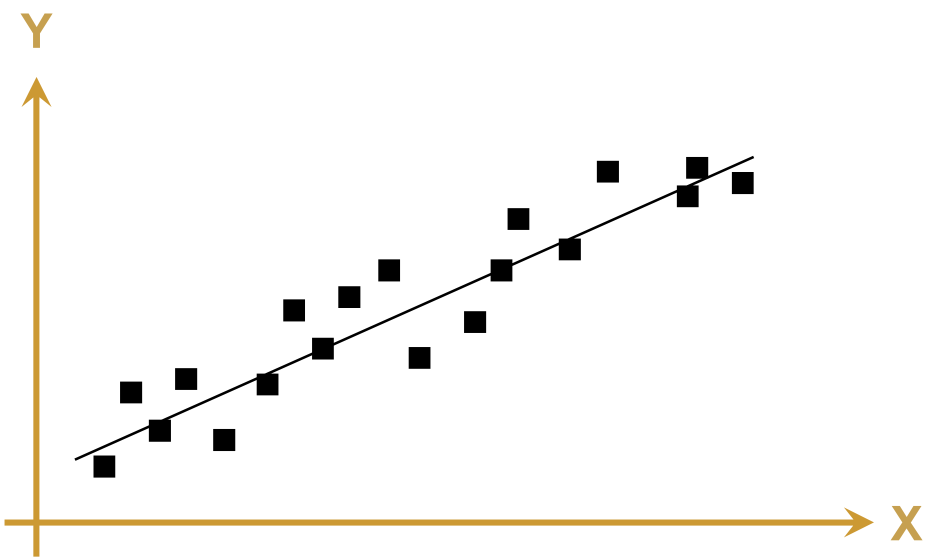
2. Explanatory or descriptive
   - Quantifying / explaining the average effect of inputs on an outcome
   - Data are treated as a random sample from a larger population of interest

# Predictive




- Given a new value for X, estimate the value for Y

# Explanatory




- Generate statements useful for decision making

  E.g., a unit increase in X is associated with

  an average increase of 2 points in Y