IPS Instituto Politécnico de Setúbal

Session 4

Regression Analysis

Course Name:

Supervised Machine Learning

By Vala A. Rohani
vala.ali.rohani@estsetubal.ips.pt

Learning goals of this session:

By the end of this session, you will be able to do the following things:

- Being familiar with machine learning and its applications
- Regression algorithms
- Know how to build different regression models in R

What is **Machine Learning?**

- In a simple form, ML is getting computers to learn from past to predict the future.
- ML systems aim to perform human-like cognitive functions where the outputs are optimised by learning from historical datasets.

Some Machine Learning Applications

Retail: Market basket analysis, Customer relationship management (CRM), Recommender Systems

Finance: Credit scoring, fraud detection

Manufacturing: Optimization, troubleshooting

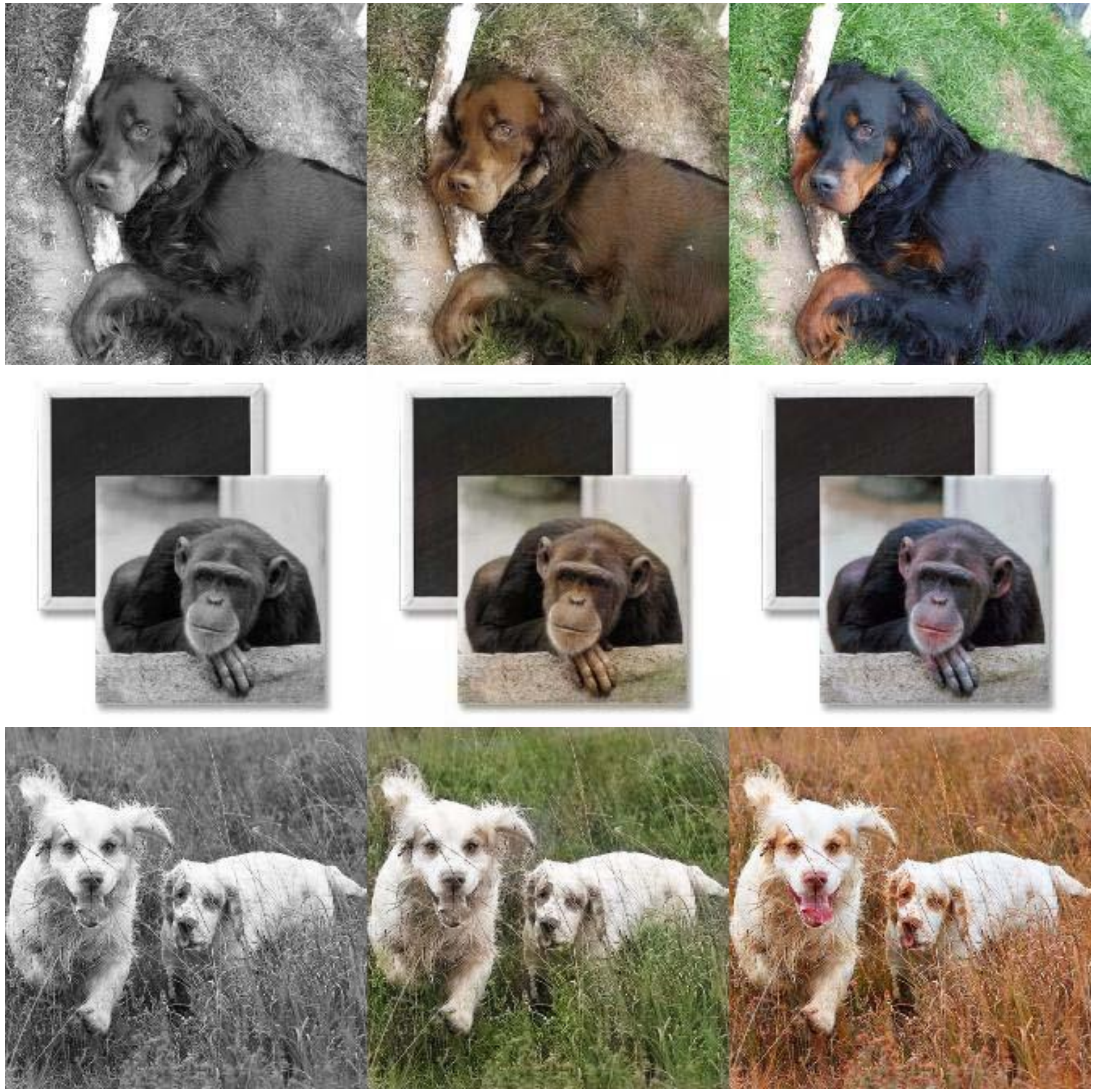
Medicine: Medical diagnosis

Telecommunications: Quality of service optimization

Bioinformatics: Motifs, alignment

Web mining: Search engines

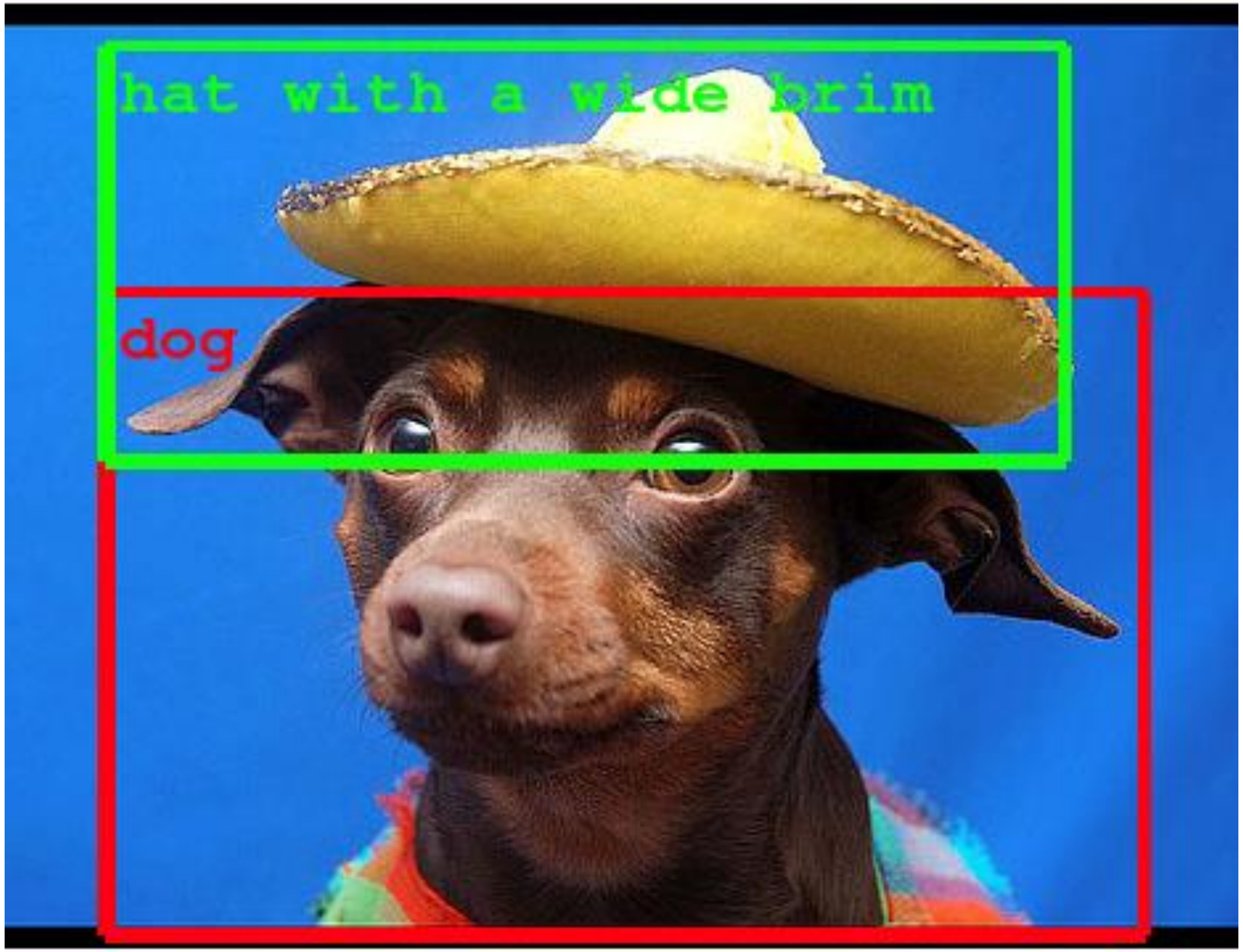
Let´s see some examples:



Colorize B&W images automatically

https://tinyclouds.org/colorize/

Let´s see some examples:



Classification

Object Recognition

https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html

Let´s see some examples:

? Where should detergents be placed in the Store to maximize their sales?

? Are window cleaning products purchased when detergents and orange juice are bought together?

? Is soda typically purchased with bananas? Does the brand of soda make a difference?

? How are the demographics of the neighborhood affecting what customers are buying?

Image source: deepclimate.org

Market Basket Analysis

Let´s see some examples:

Social Media Analytics

(Sentiment Analysis)

https://medium.com/@vini.bandeira.vb/creating-your-own-sentiment-analysis-service-to-classify-portuguese-phrases-eb2fb6613eb1

Let´s see some examples:

Stock Prediction

(Regression & Time Series Analysis)

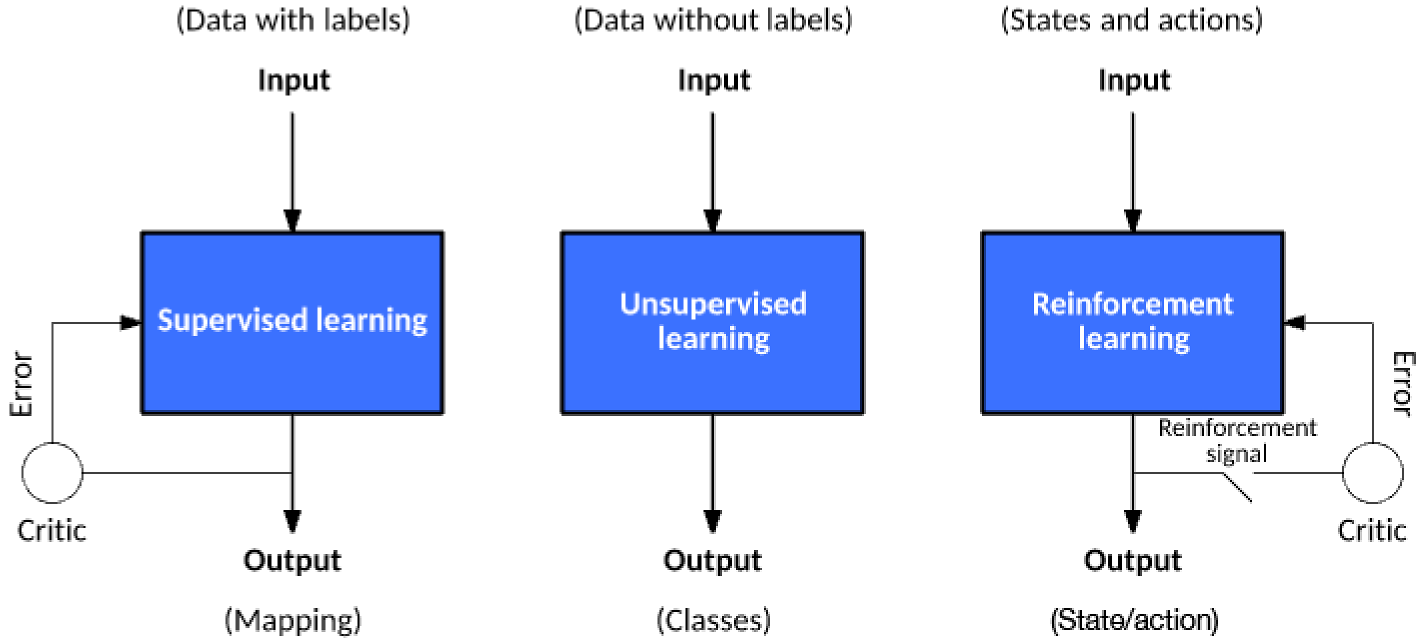
https://medium.com/daily-python/simple-stock-prediction-using-linear-regression-in-python-daily-python-18-ecfe23b76ce9

Let´s see some examples:

Recommender Systems

https://medium.com/@fenjiro/recommender-systems-d0e597424a98

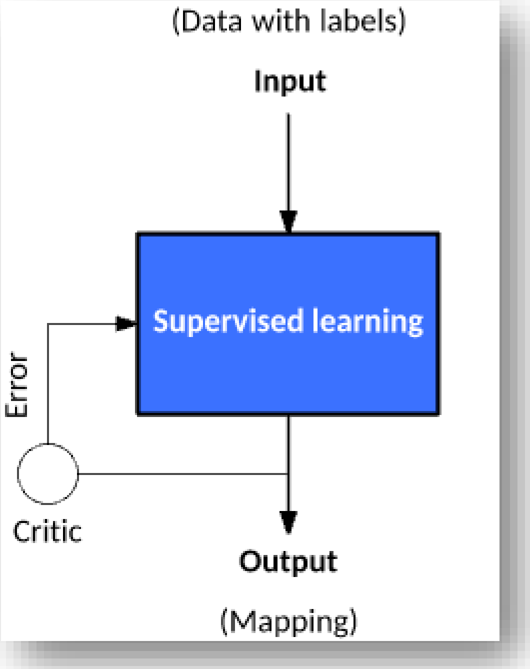
Machine Learning Models



Supervised Learning

In supervised learning, a data set includes its desired outputs (or *labels*) such that a function can calculate an error for a given prediction.

The supervision comes when a prediction is made and an error produced (actual vs. desired) to alter the function and learn the mapping.



Supervised learning problems can be further grouped into two main categories:

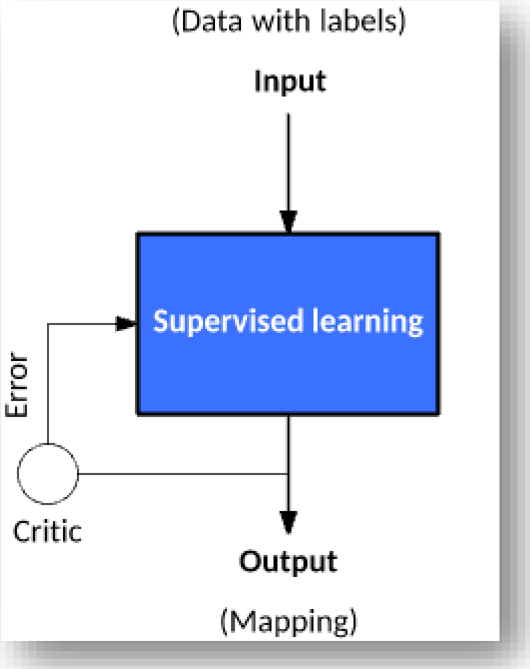
Classification: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised Learning

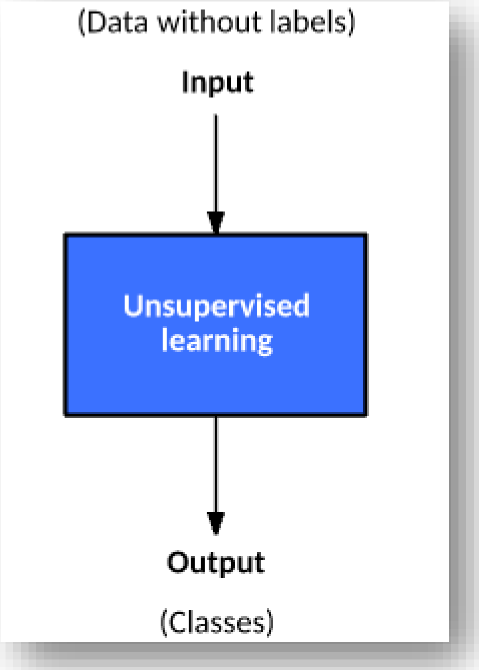
Supervised learning is effective for a variety of business purposes, including sales forecasting, inventory optimization, and fraud detection. Some examples of use cases include:

- •Predicting real estate prices
- •Classifying whether bank transactions are fraudulent or not
- •Finding disease risk factors
- •Determining whether loan applicants are low-risk or high-risk
- •Predicting the failure of industrial equipment's mechanical parts



Unsupervised Learning

In unsupervised learning, **a data set doesn't include a desired output**; No Label!
Therefore, there's no way to supervise the function.
Instead, the function attempts to segment the data set into **"classes"** so that each class contains a portion of the data set with common features.



Unsupervised learning problems can be further grouped into two main categories:

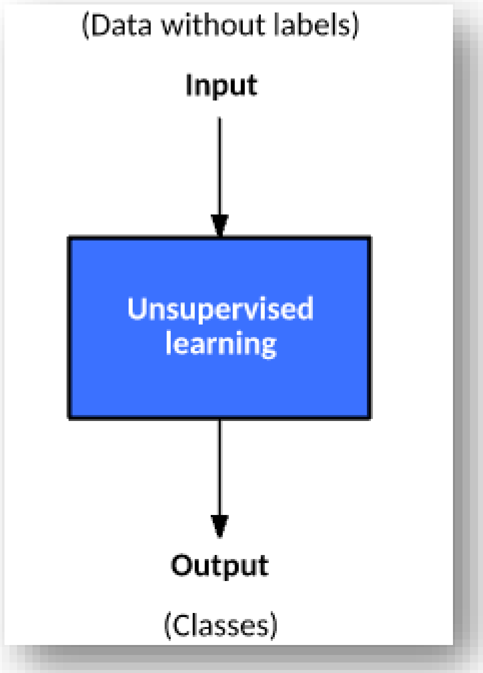
Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Unsupervised Learning

A few example use cases of unsupervised methods include:

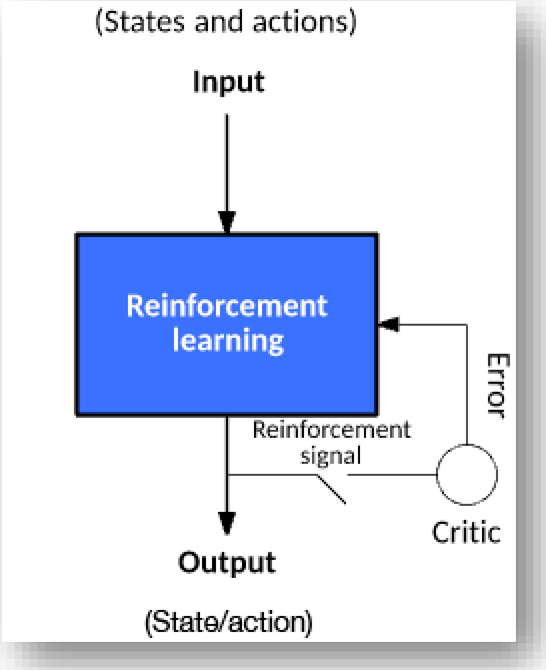
- Creating customer groups based on purchase behavior
- Grouping inventory according to sales and/or manufacturing metrics
- Pinpointing associations in customer data (for example, customers who buy a specific style of handbag might be interested in a specific style of shoe)



Reinforcement Learning

Finally, in reinforcement learning, **the algorithm attempts to learn actions for a given set of states that lead to a goal state**. An error is provided not after each example (as is the case for supervised learning) but instead on receipt of a reinforcement signal (such as reaching the goal state).

This behavior is similar to human learning, where feedback isn't necessarily provided for all actions but **when a reward is warranted**.

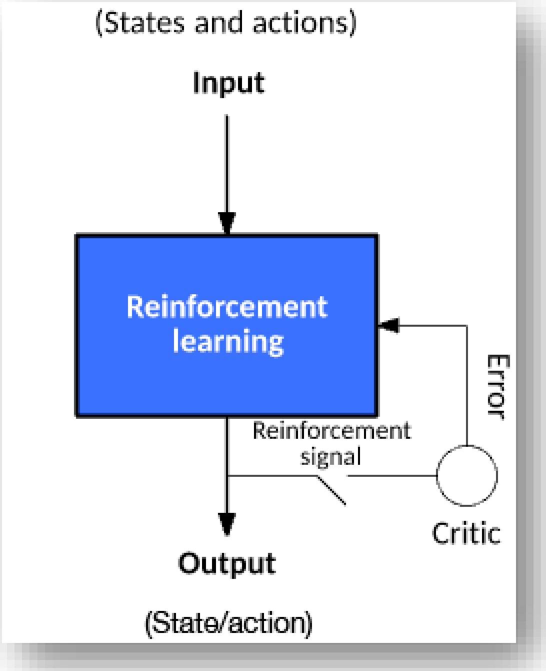


Reinforcement Learning

This type of machine learning requires less management than supervised learning, it's viewed as easier to work with dealing with unlabeled data sets. Practical applications for this type of machine learning are still emerging.

Some examples of uses include:

- Teaching cars to park themselves and drive autonomously
- Dynamically controlling traffic lights to reduce traffic jams
- Training robots to learn policies using raw video images as input that they can use to replicate the actions they see

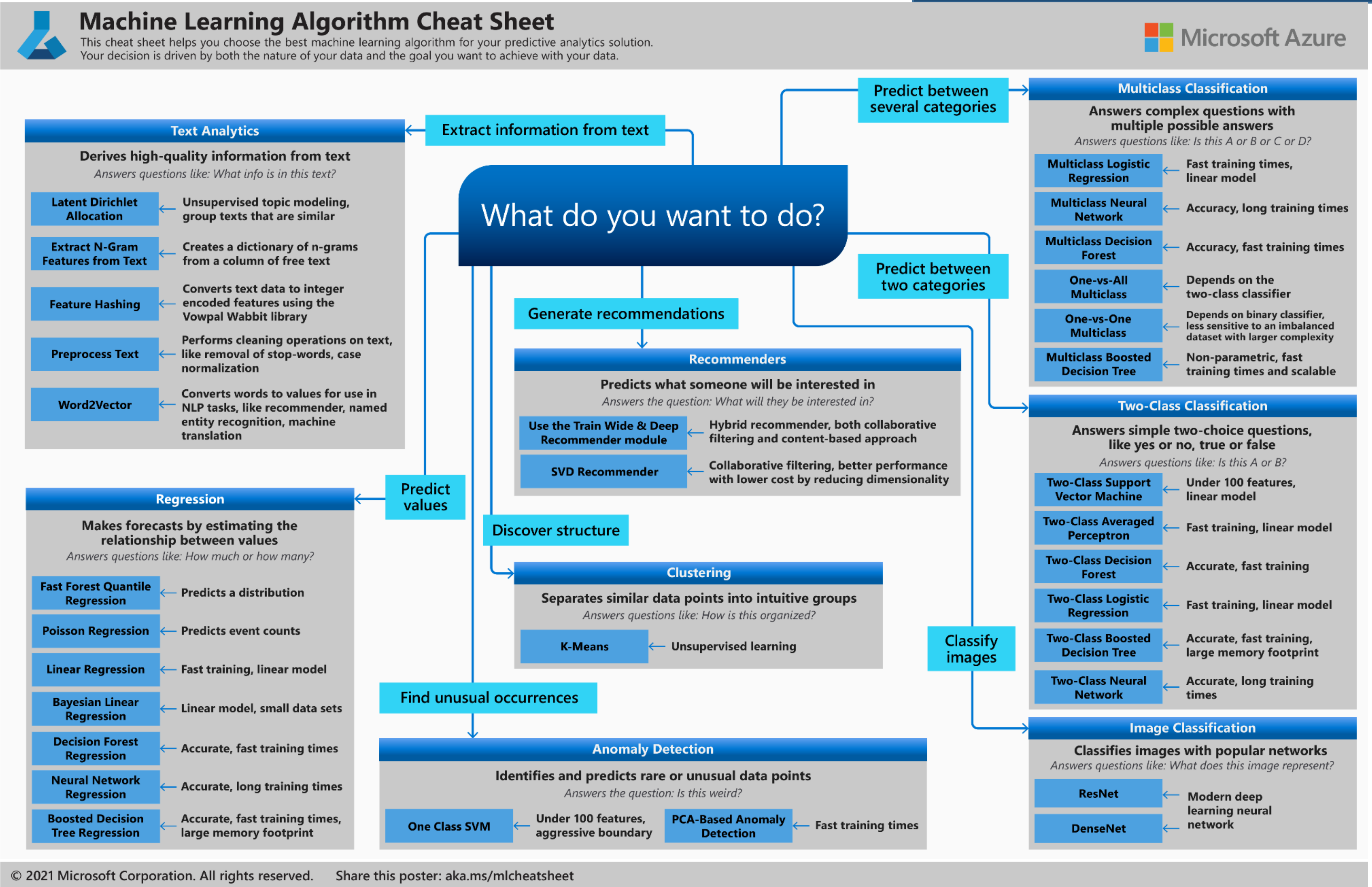


Reinforcement Learning

Let´s see

a good example of Reinforcement Learning!

https://www.youtube.com/watch?v=spfpBrBjntg



Machine Learning Algorithm Cheat Sheet

This cheat sheet helps you choose the best machine learning algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the goal you want to achieve with your data.

Microsoft Azure

What do you want to do?

Extract information from text → **Text Analytics**

Derives high-quality information from text

Answers questions like: What info is in this text?

- Latent Dirichlet Allocation ← Unsupervised topic modeling, group texts that are similar
- Extract N-Gram Features from Text ← Creates a dictionary of n-grams from a column of free text
- Feature Hashing ← Converts text data to integer encoded features using the Vowpal Wabbit library
- Preprocess Text ← Performs cleaning operations on text, like removal of stop-words, case normalization
- Word2Vector ← Converts words to values for use in NLP tasks, like recommender, named entity recognition, machine translation

Predict values → **Regression**

Makes forecasts by estimating the relationship between values

Answers questions like: How much or how many?

- Fast Forest Quantile Regression ← Predicts a distribution
- Poisson Regression ← Predicts event counts
- Linear Regression ← Fast training, linear model
- Bayesian Linear Regression ← Linear model, small data sets
- Decision Forest Regression ← Accurate, fast training times
- Neural Network Regression ← Accurate, long training times
- Boosted Decision Tree Regression ← Accurate, fast training times, large memory footprint

Generate recommendations → **Recommenders**

Predicts what someone will be interested in

Answers the question: What will they be interested in?

- Use the Train Wide & Deep Recommender module ← Hybrid recommender, both collaborative filtering and content-based approach
- SVD Recommender ← Collaborative filtering, better performance with lower cost by reducing dimensionality

Discover structure → **Clustering**

Separates similar data points into intuitive groups

Answers questions like: How is this organized?

- K-Means ← Unsupervised learning

Find unusual occurrences → **Anomaly Detection**

Identifies and predicts rare or unusual data points

Answers the question: Is this weird?

- One Class SVM ← Under 100 features, aggressive boundary
- PCA-Based Anomaly Detection ← Fast training times

Predict between several categories → **Multiclass Classification**

Answers complex questions with multiple possible answers

Answers questions like: Is this A or B or C or D?

- Multiclass Logistic Regression ← Fast training times, linear model
- Multiclass Neural Network ← Accuracy, long training times
- Multiclass Decision Forest ← Accuracy, fast training times
- One-vs-All Multiclass ← Depends on the two-class classifier
- One-vs-One Multiclass ← Depends on binary classifier, less sensitive to an imbalanced dataset with larger complexity
- Multiclass Boosted Decision Tree ← Non-parametric, fast training times and scalable

Predict between two categories → **Two-Class Classification**

Answers simple two-choice questions, like yes or no, true or false

Answers questions like: Is this A or B?

- Two-Class Support Vector Machine ← Under 100 features, linear model
- Two-Class Averaged Perceptron ← Fast training, linear model
- Two-Class Decision Forest ← Accurate, fast training
- Two-Class Logistic Regression ← Fast training, linear model
- Two-Class Boosted Decision Tree ← Accurate, fast training, large memory footprint
- Two-Class Neural Network ← Accurate, long training times

Classify images → **Image Classification**

Classifies images with popular networks

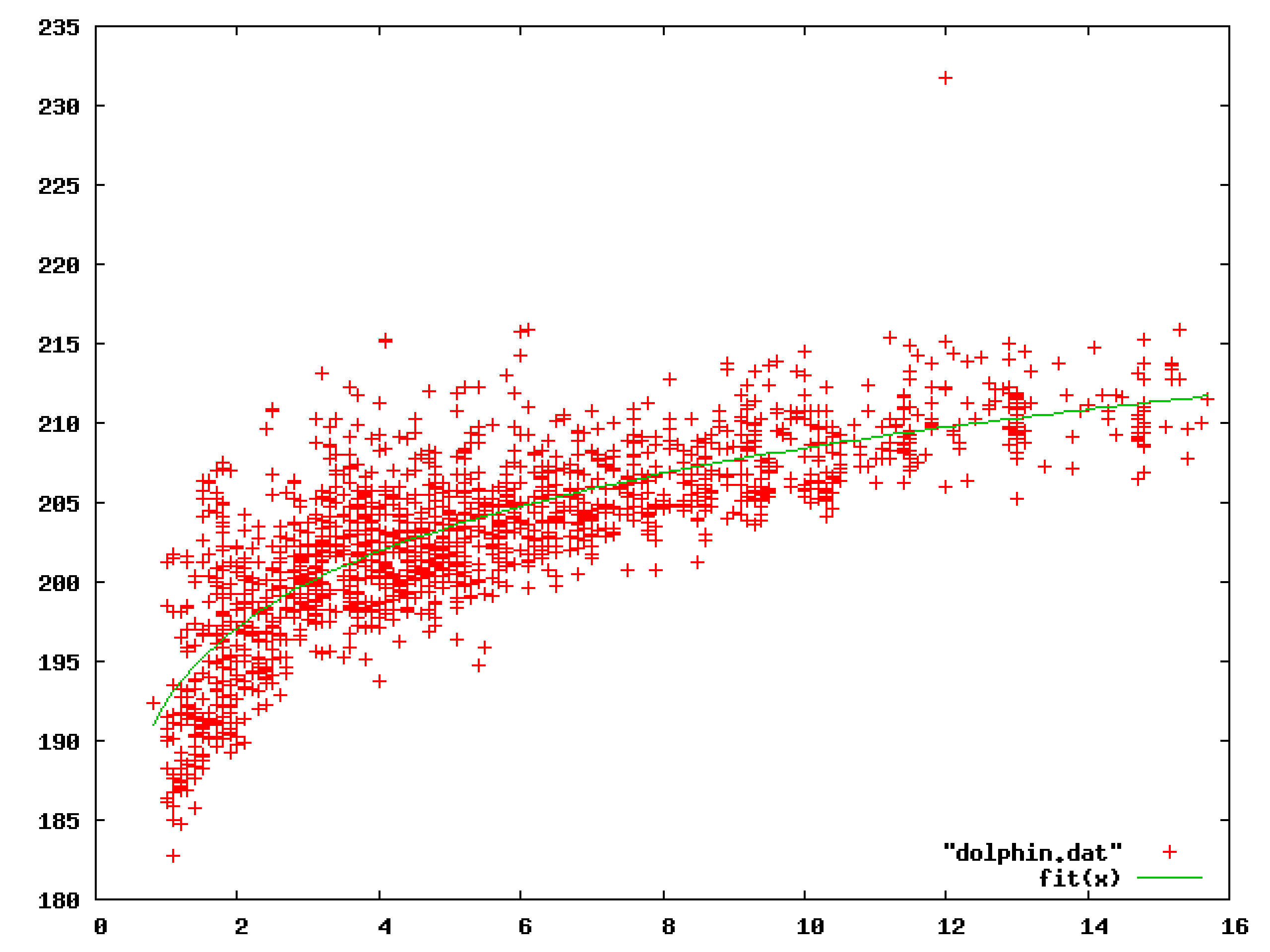
Answers questions like: What does this image represent?

- ResNet ← Modern deep learning neural network
- DenseNet ← Modern deep learning neural network

Share this poster: aka.ms/mlcheatsheet

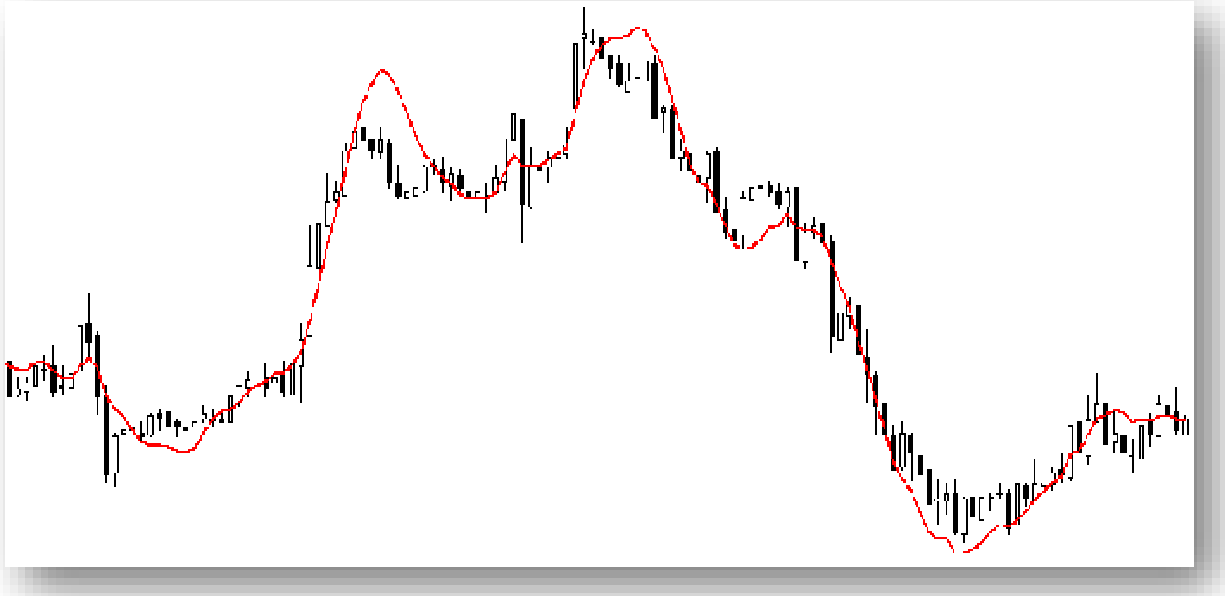
Source: https://docs.microsoft.com/pt-pt/azure/machine-learning/algorithm-cheat-sheet

Regression Models



What is Regression Analysis?

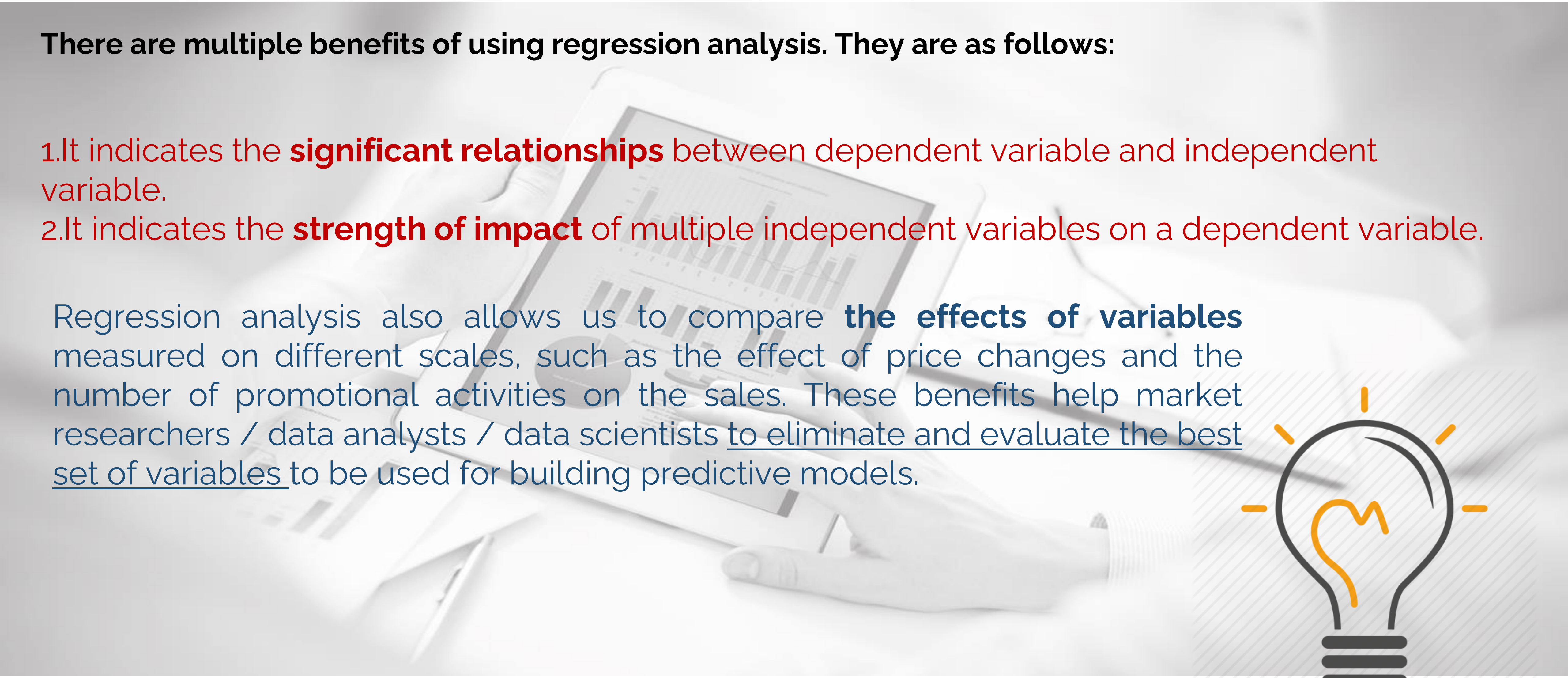
In this approach, we **fit a curve / line** to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.



Some questions that Regression Analysis can answer?

- What is the effect of one more year of education on the income of an individual?
- What are the factors that indicate whether a particular individual will get a job?
- Can we predict how the price of a particular stock will change in the next few weeks?
- How will a particular policy such as change in taxes on cigarettes affect the incidence of smoking in a state?
- How long will a patient survive after being given a particular treatment as compared to not being given that treatment?

Why do we use Regression Analysis?



There are multiple benefits of using regression analysis. They are as follows:

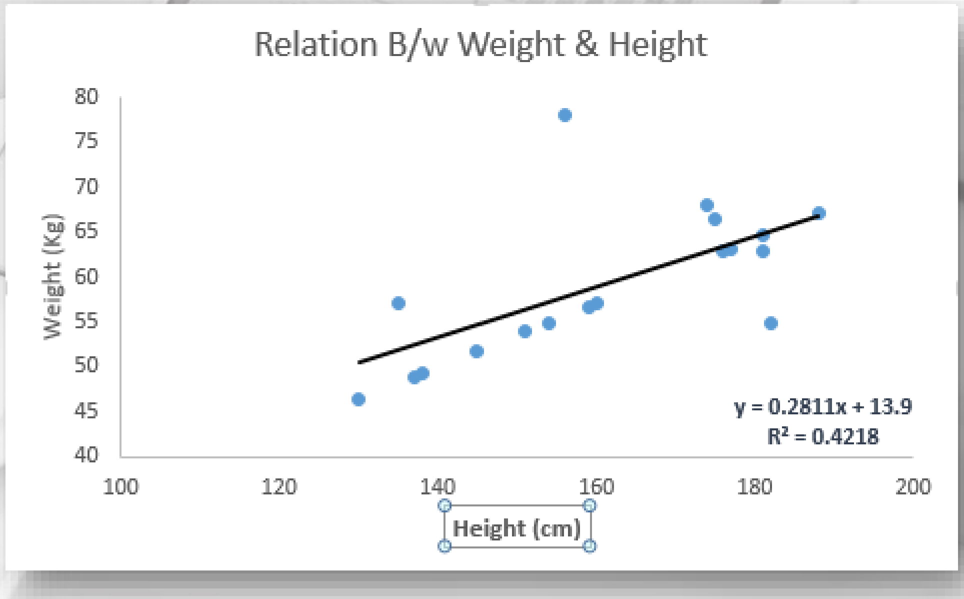
1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare **the effects of variables** measured on different scales, such as the effect of price changes and the number of promotional activities on the sales. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

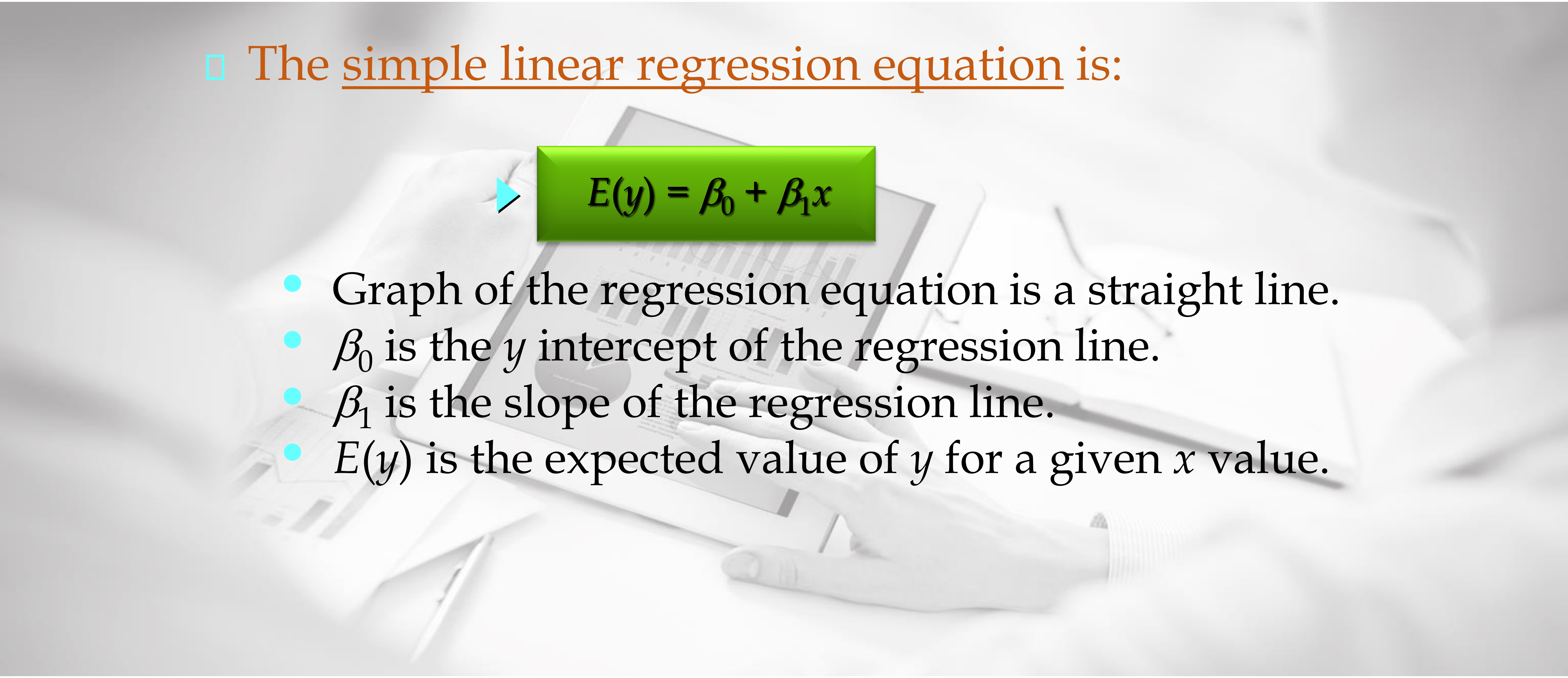
Linear Regression

It is one of the most widely known modeling technique. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).



Linear Regression



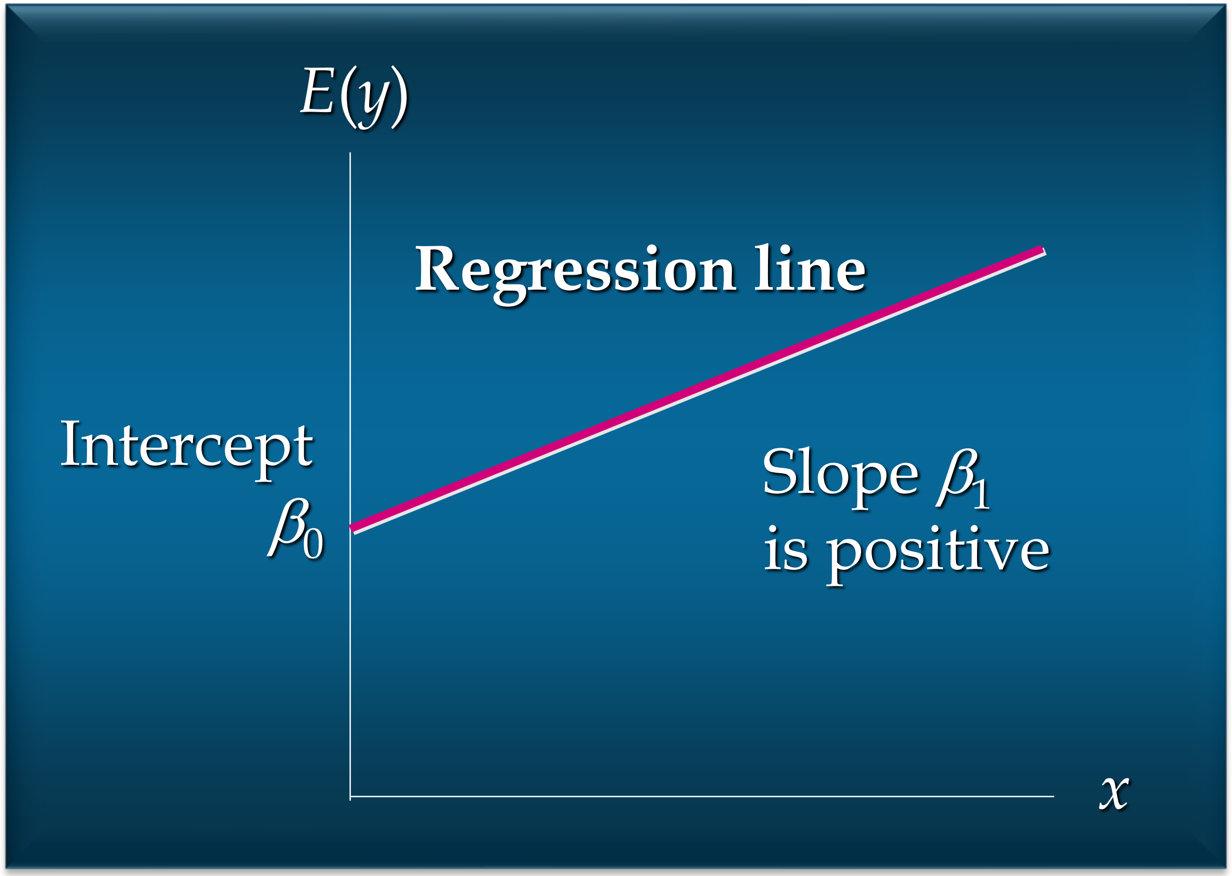
- The simple linear regression equation is:

$$E(y) = \beta_0 + \beta_1 x$$

- Graph of the regression equation is a straight line.
- β_0 is the y intercept of the regression line.
- β_1 is the slope of the regression line.
- $E(y)$ is the expected value of y for a given x value.

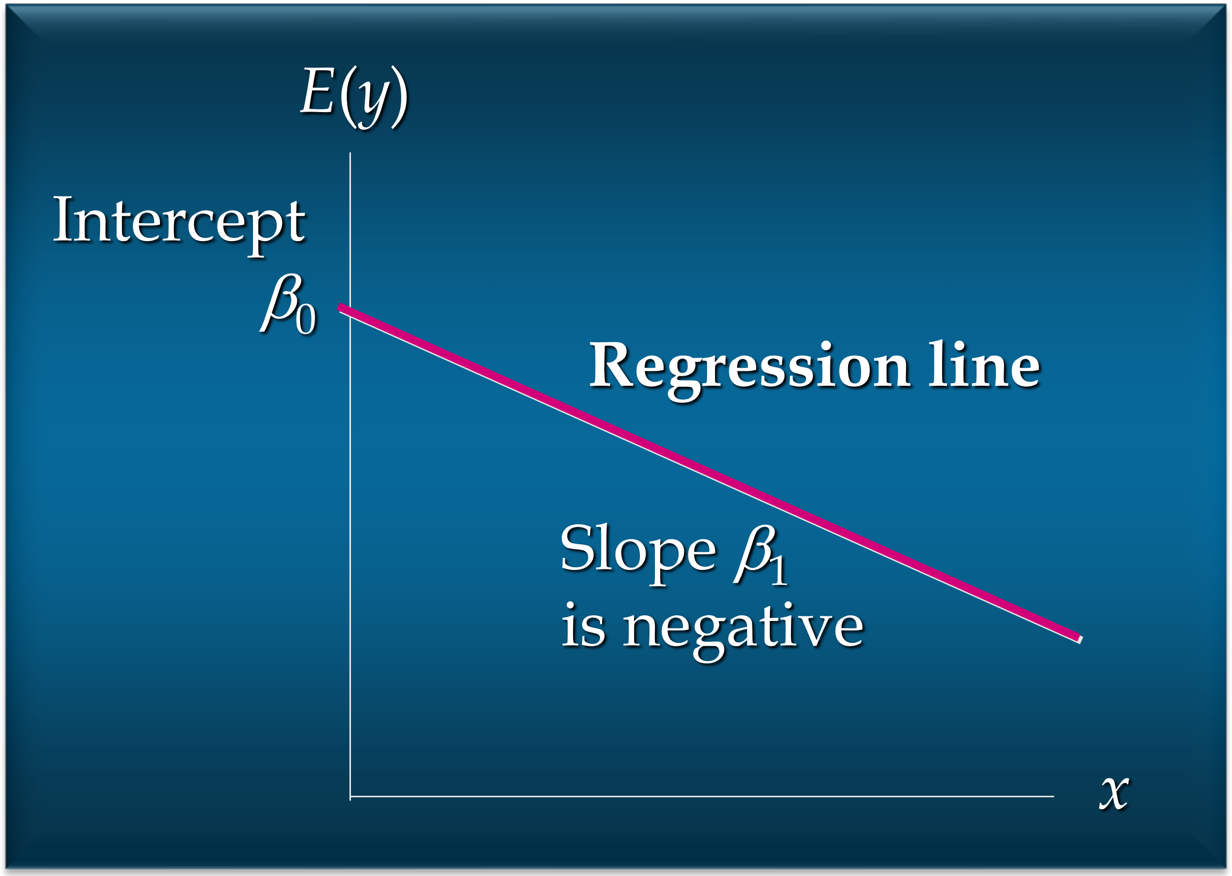
Linear Regression

- Positive Linear Relationship



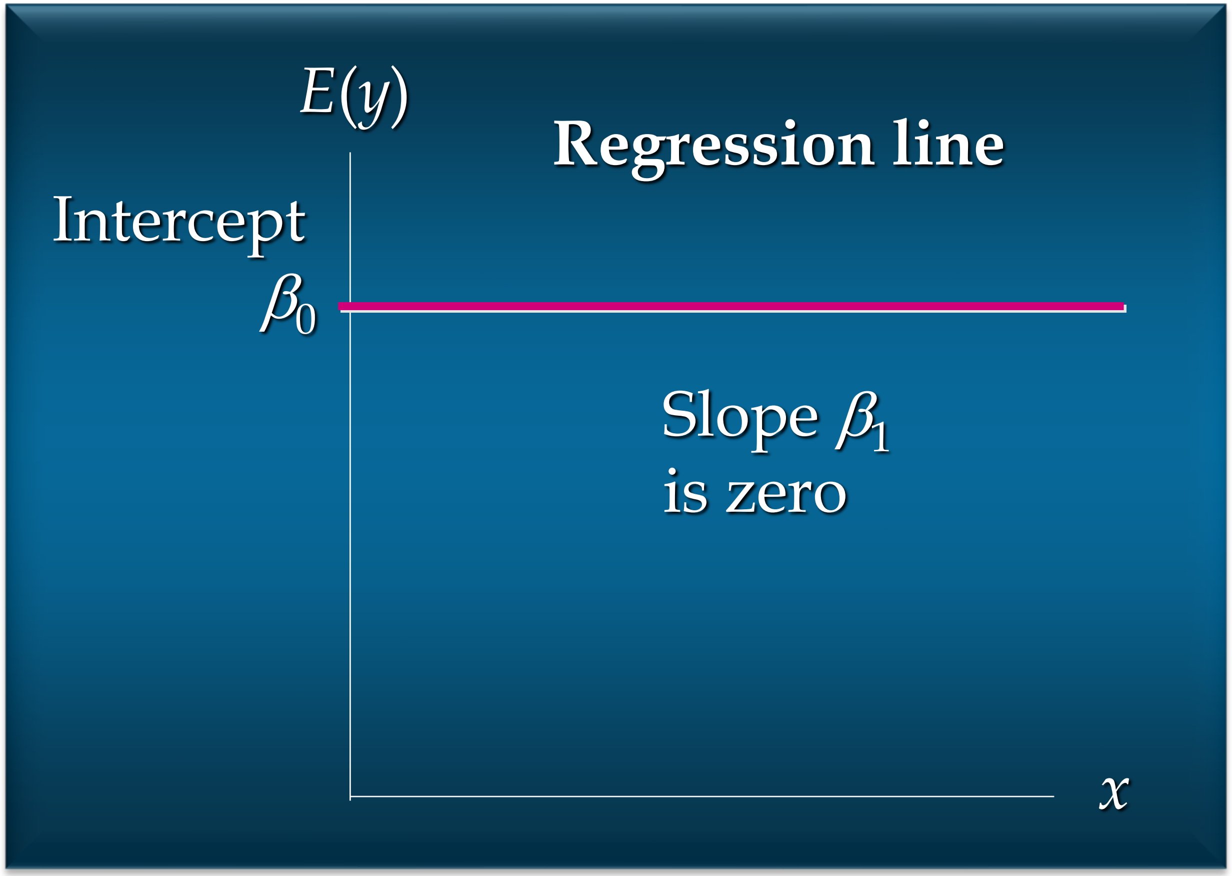
Linear Regression

- Negative Linear Relationship



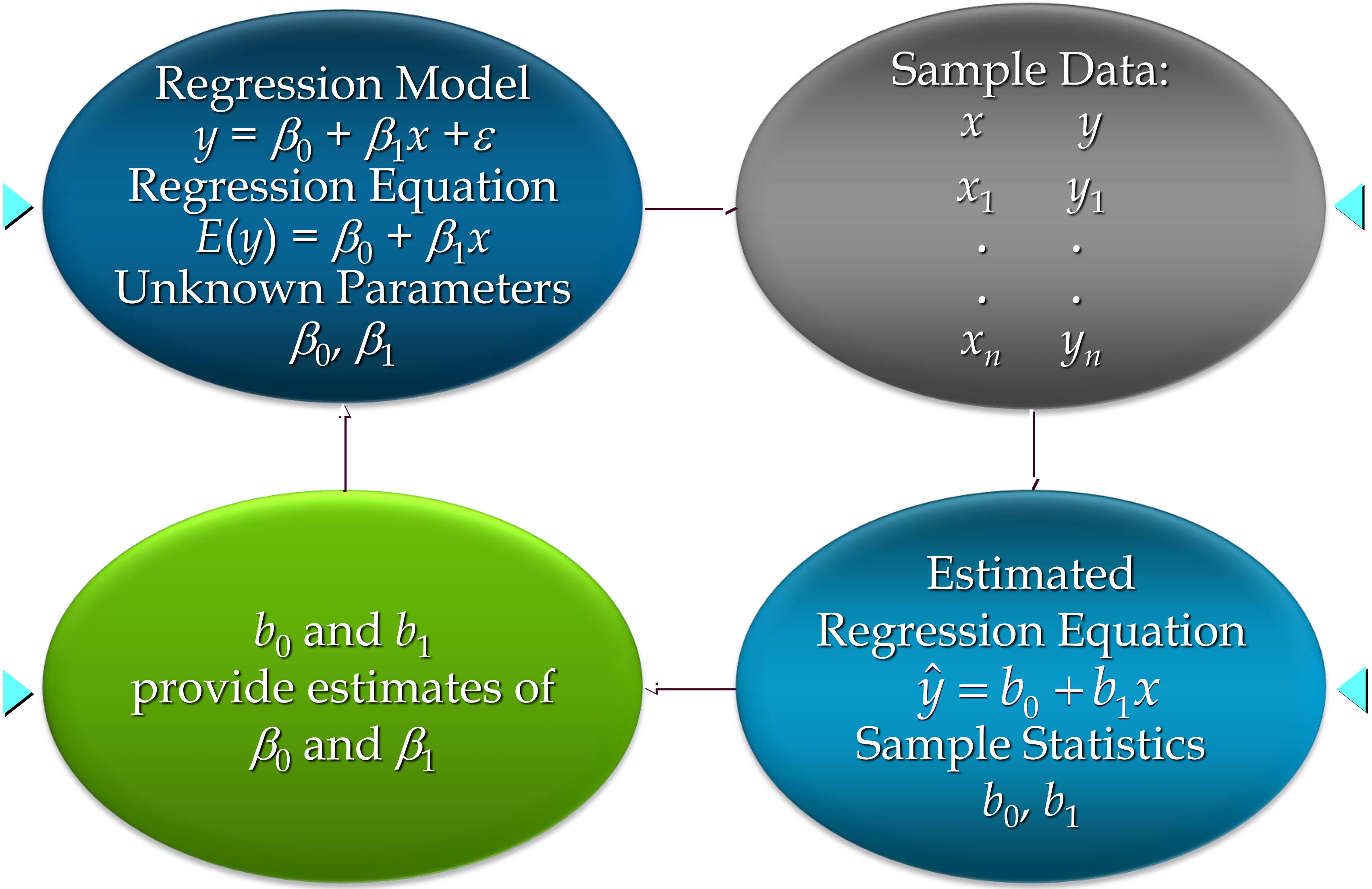
Linear Regression

- No Relationship



Linear Regression

Estimation Process



Regression Model
$y = \beta_0 + \beta_1 x + \varepsilon$
Regression Equation
$E(y) = \beta_0 + \beta_1 x$
Unknown Parameters
β_0, β_1

Sample Data:

x	y
x_1	y_1
.	.
.	.
x_n	y_n

Estimated
Regression Equation
$\hat{y} = b_0 + b_1 x$
Sample Statistics
b_0, b_1

b_0 and b_1
provide estimates of
β_0 and β_1

Linear Regression

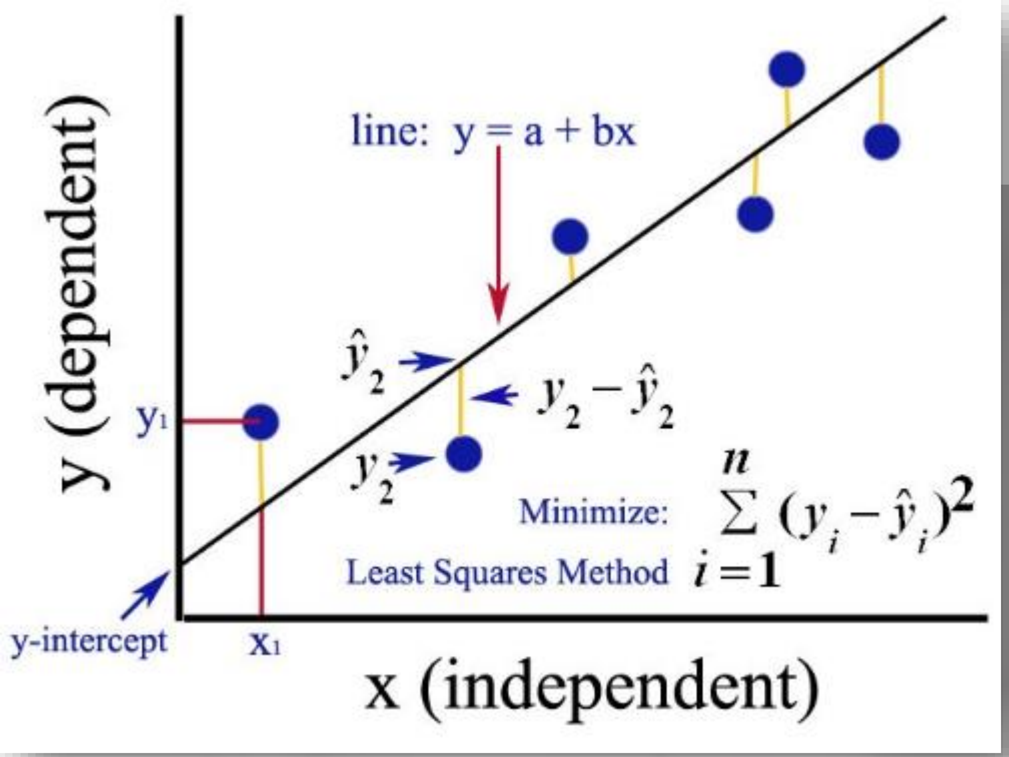
Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

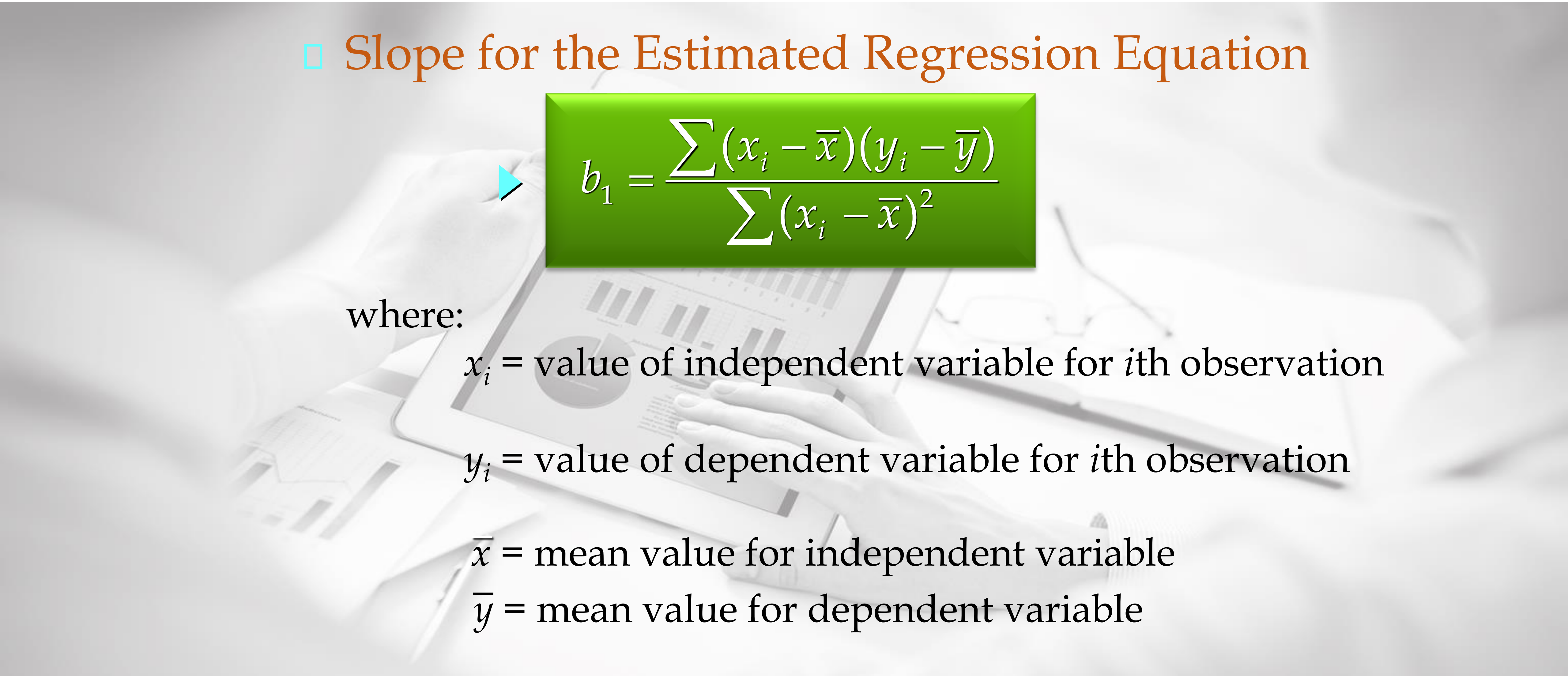
where:

y_i = <u>observed</u> value of the dependent variable for the ith observation

$\hat{y}_i$ = <u>estimated</u> value of the dependent variable for the ith observation



Linear Regression



- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

where:

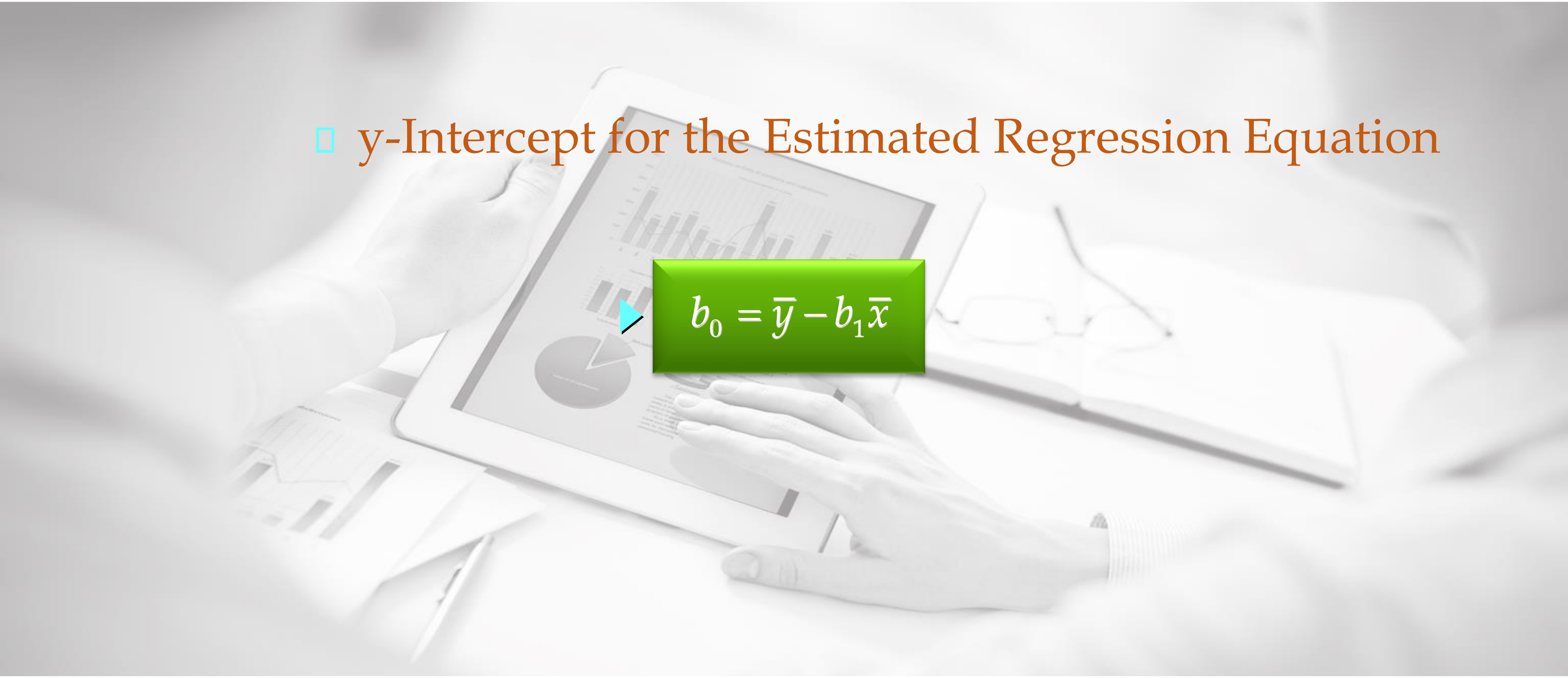
x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

Linear Regression



- y-Intercept for the Estimated Regression Equation

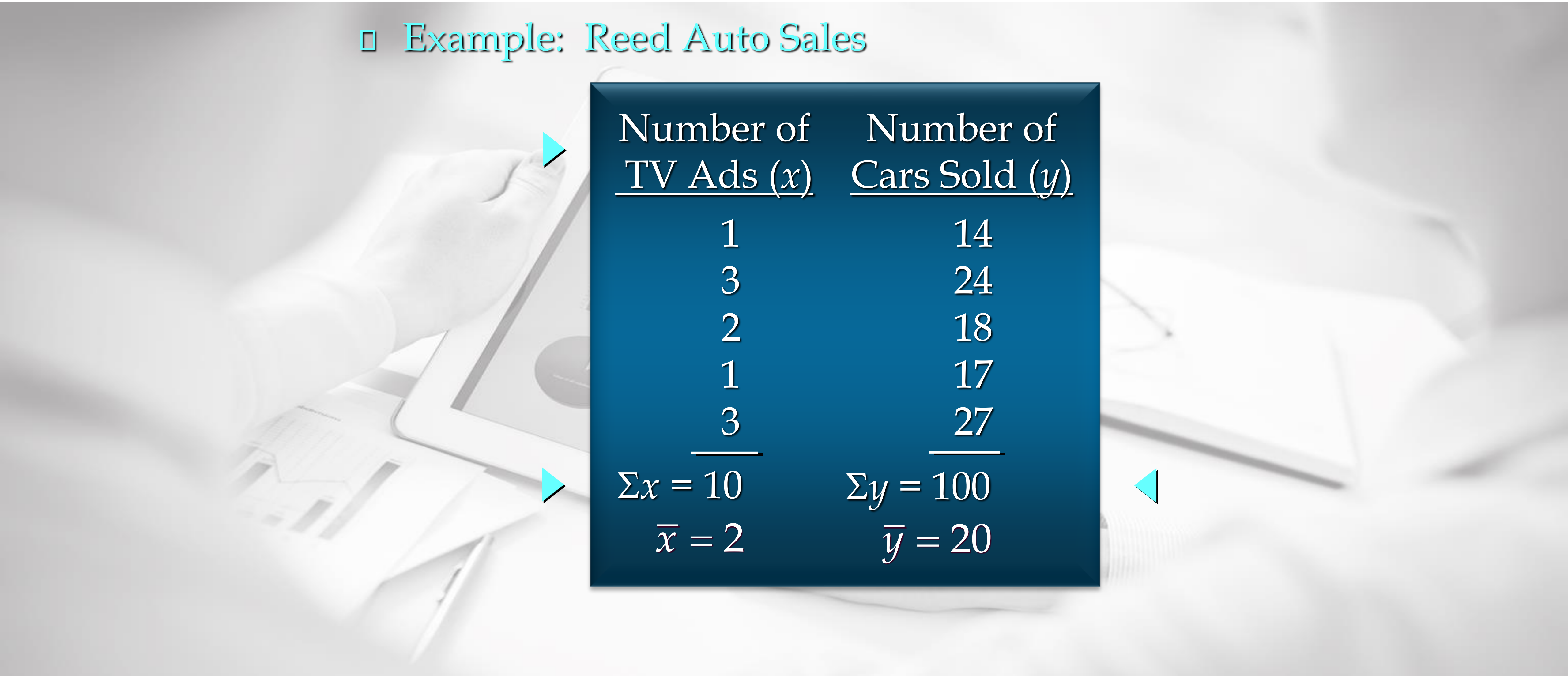
$$b_0 = \bar{y} - b_1\bar{x}$$

Example: Reed Auto Sales

Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

Linear Regression

- Example: Reed Auto Sales



Number of TV Ads (x)	Number of Cars Sold (y)
1	14
3	24
2	18
1	17
3	27
$\Sigma x = 10$	$\Sigma y = 100$
$\bar{x} = 2$	$\bar{y} = 20$

Linear Regression

- ▯ Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{20}{4} = 5$$

- ▯ *y*-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x} = 20 - 5(2) = 10$$

- ▯ Estimated Regression Equation

$$\hat{y} = 10 + 5x$$

Coefficient of Determination

- Relationship Among SST, SSR, SSE

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$$

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination

- The <u>coefficient of determination</u> is:

  ▶ $$r^2 = SSR/SST$$

  where:

  SSR = sum of squares due to regression

  SST = total sum of squares

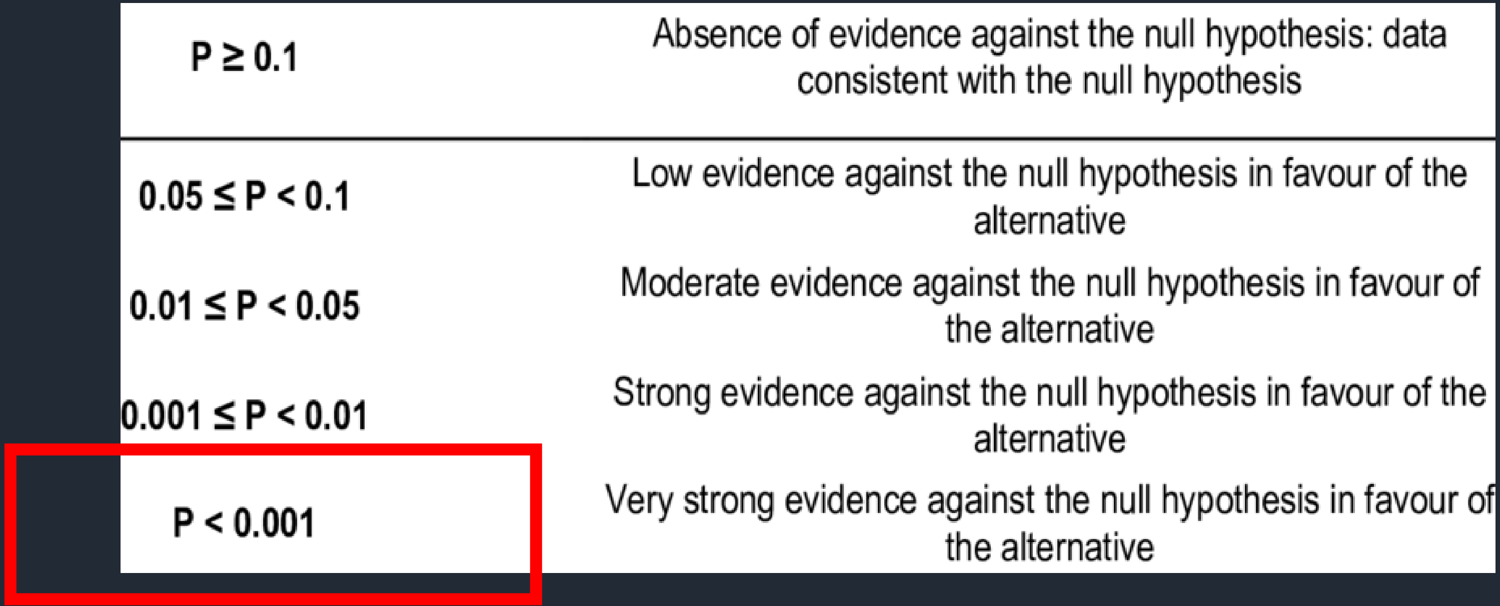
Linear Regression

Coefficient of Determination

- $r^2 = \text{SSR}/\text{SST} = 100/114 = \boxed{.8772}$
- The regression relationship is very strong; 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

P values and statistical significance

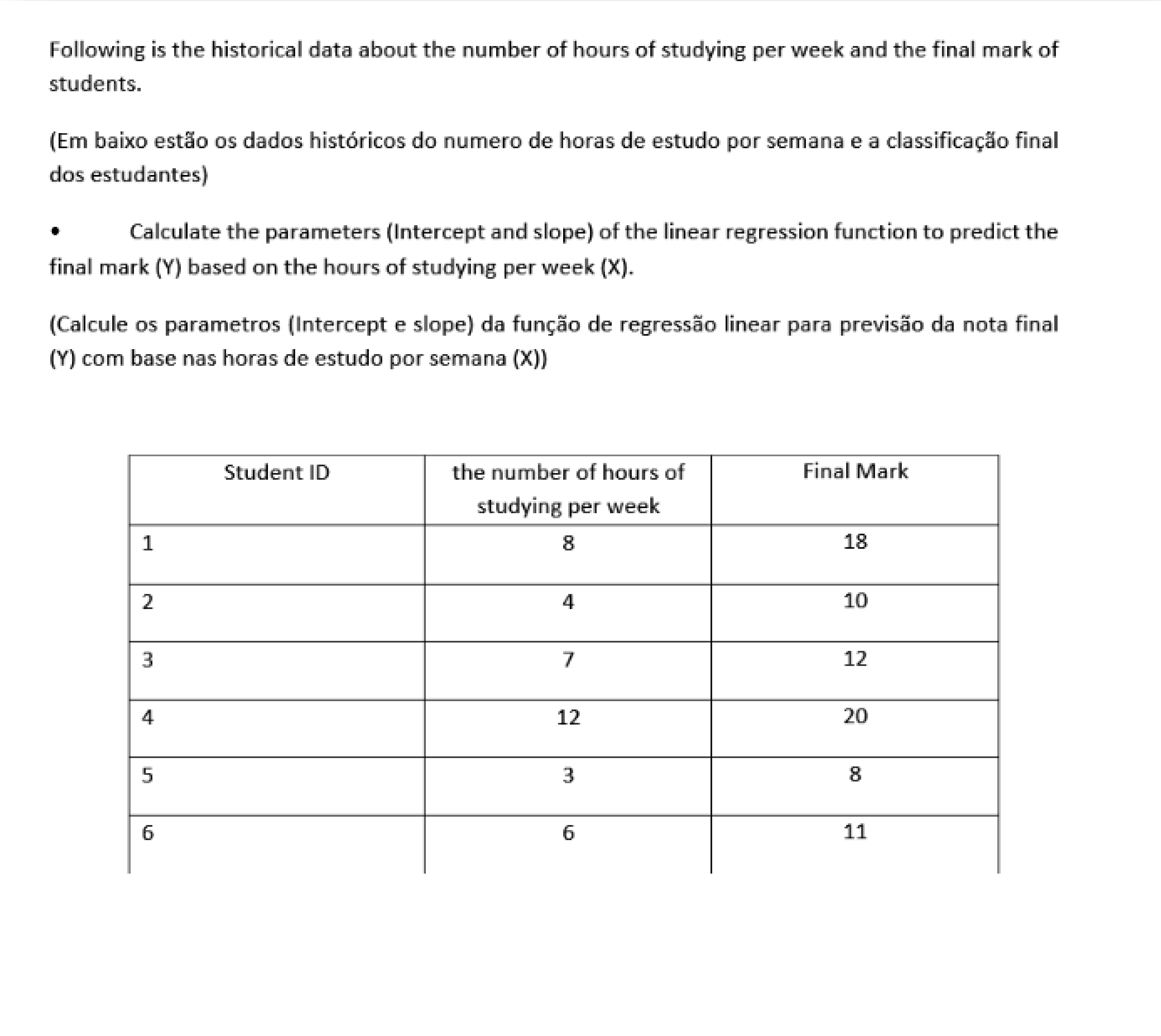
P values are most often used by researchers to say whether a certain pattern they have measured is statistically significant.



$P \geq 0.1$	Absence of evidence against the null hypothesis: data consistent with the null hypothesis
$0.05 \leq P < 0.1$	Low evidence against the null hypothesis in favour of the alternative
$0.01 \leq P < 0.05$	Moderate evidence against the null hypothesis in favour of the alternative
$0.001 \leq P < 0.01$	Strong evidence against the null hypothesis in favour of the alternative
$P < 0.001$	Very strong evidence against the null hypothesis in favour of the alternative

Linear Regression

Lab Activity:



Following is the historical data about the number of hours of studying per week and the final mark of students.

(Em baixo estão os dados históricos do numero de horas de estudo por semana e a classificação final dos estudantes)

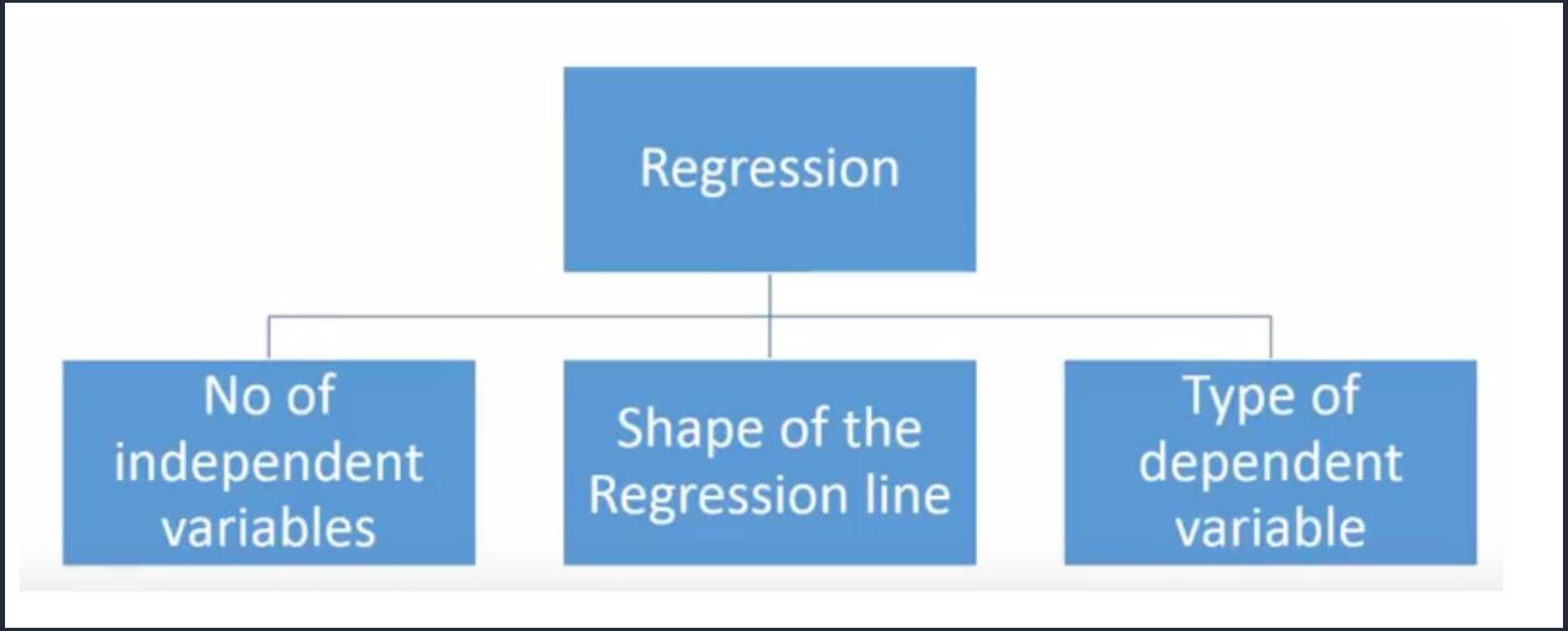
- Calculate the parameters (Intercept and slope) of the linear regression function to predict the final mark (Y) based on the hours of studying per week (X).

(Calcule os parametros (Intercept e slope) da função de regressão linear para previsão da nota final (Y) com base nas horas de estudo por semana (X))

Student ID	the number of hours of studying per week	Final Mark
1	8	18
2	4	10
3	7	12
4	12	20
5	3	8
6	6	11

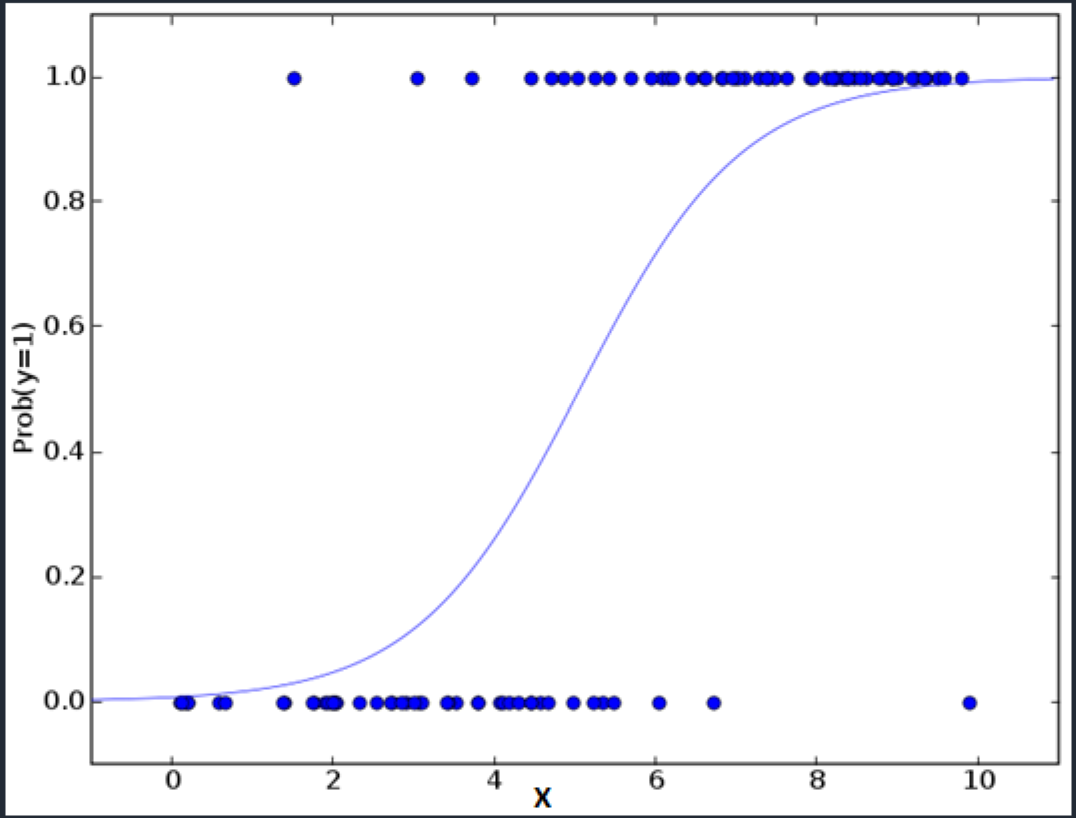
What are the main Regression Analysis techniques?

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line). We'll discuss them in detail in the following sections.


Regression
No of independent variables
Shape of the Regression line
Type of dependent variable

Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature.



Some questions that Logistic Regression can answer?

- **How does the probability of getting lung cancer (yes vs. no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day?**
- **Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?**

Logistic Regression

Important Points:

•It is widely used for **classification problems**

•Logistic regression doesn't require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio

•To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression

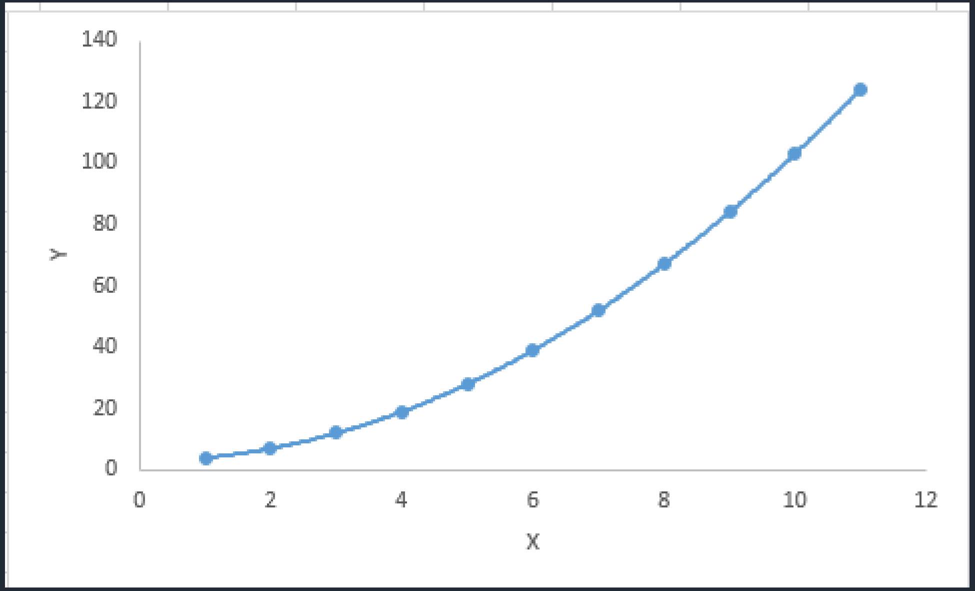
•It requires **large sample sizes**

•The independent variables should not be correlated with each other i.e. **no multi collinearity**.

Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

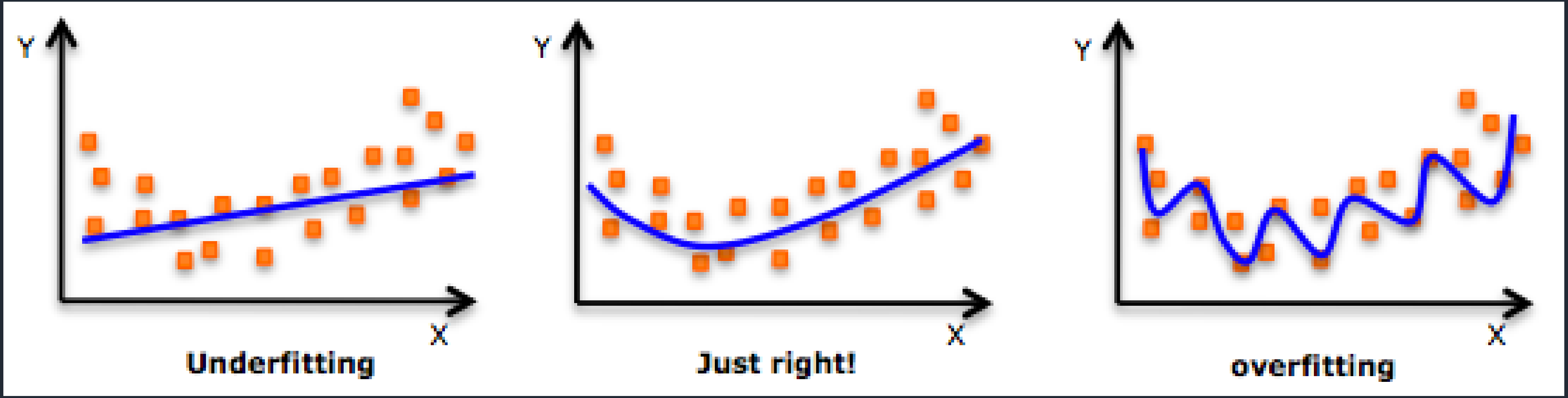
$$y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \ldots + \beta_p x_j^p + \varepsilon$$



Polynomial Regression

Important Points:

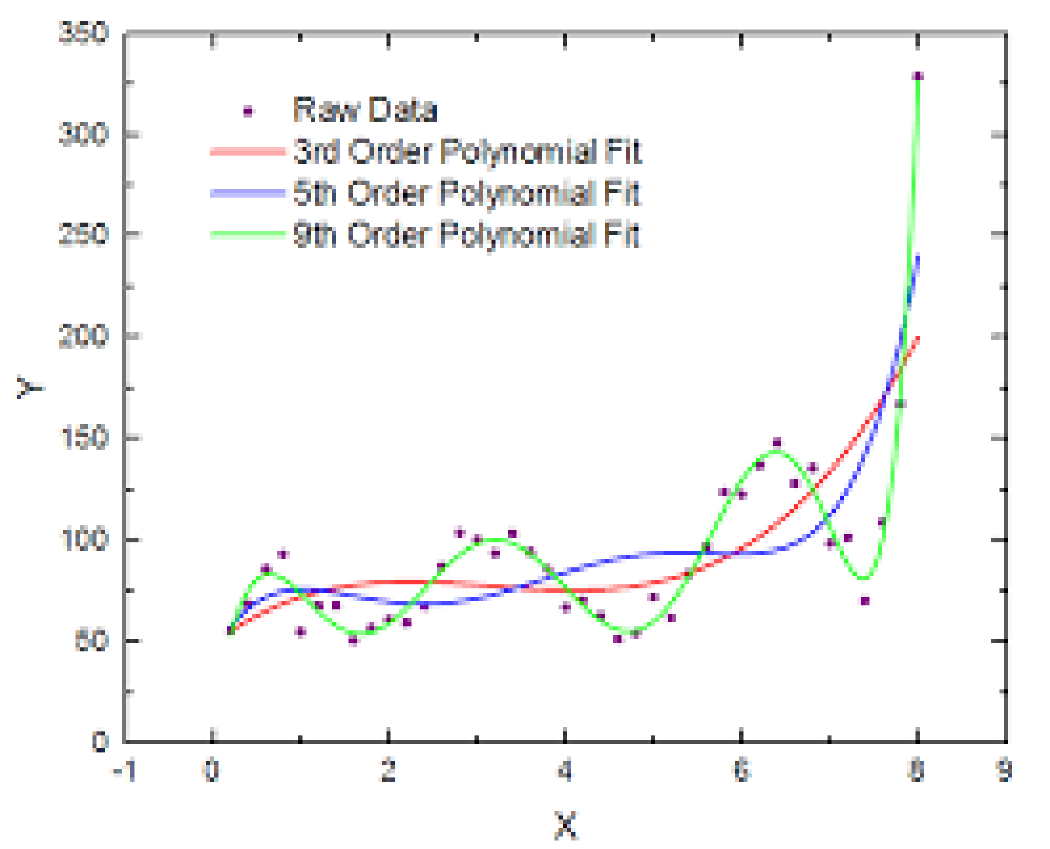
- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:



Polynomial Regression

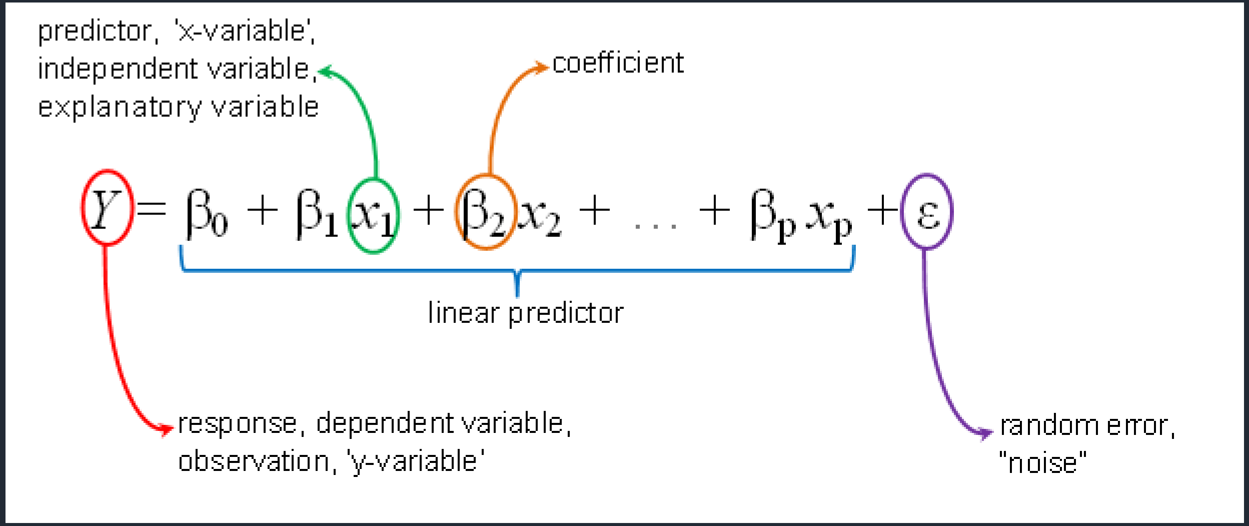
Important Points:

- Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing weird results on extrapolation.



Multiple Regression

Multiple regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummy coded as appropriate).



predictor, 'x-variable', independent variable, explanatory variable

coefficient

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}_{\text{linear predictor}} + \varepsilon$$

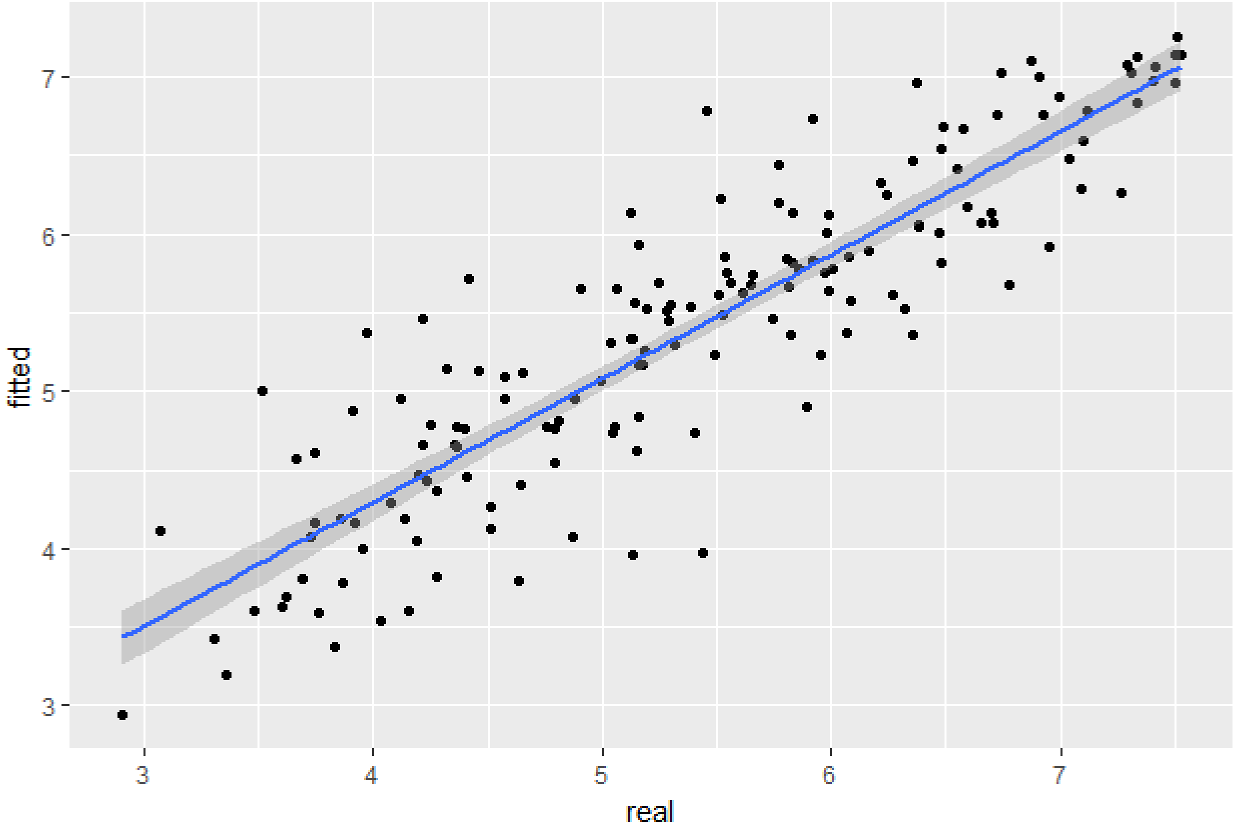
response, dependent variable, observation, 'y-variable'

random error, "noise"

Some questions that Multiple Regression can answer?

- Do age and IQ scores effectively predict GPA?
- Do weight, height, and age explain the variance in cholesterol levels?

Practical Session: Regression Models in R



Any questions?

Vala.ali.rohani@ips.pt