Introduction to Statistical Learning and Application
CC1: Simple Linear Regression

Razan MHANNA
STATIFY team, Inria centre at the University Grenoble Alpes
LEMASSON/CHRISTEN team, Grenoble Institute of Neurosciences GIN

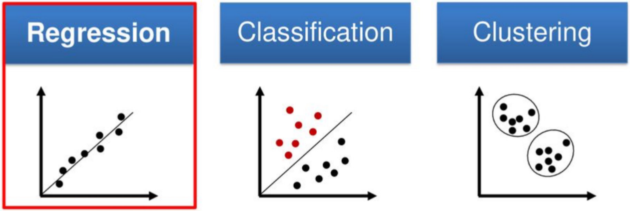
30 January 2024

Contents

Current Section

Course Information

- This is the complementary course of "Introduction to Statistical Learning and Applications" given to students from ENSIMAG and UGA by Professor Pedro Rodrigues.
- The classes will be given on Tuesdays from 11h30 to 13h at IM2AG, room F319.
- The materials used will be made available in this page https://github.com/ISLA-Grenoble/2024-complementary.



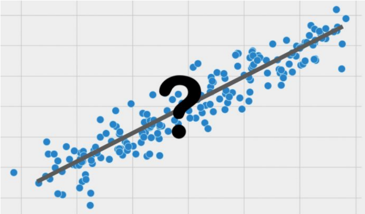
Current Section

What is Simple Linear Regression?

- Simple Linear Regression is a method used to fit the best straight line between a set of data points,
- After a graph is properly scaled, the data points must "look" like they would fit a straight line, not a parabola, or any other shape,
- The line is used as a model in order to predict a variable y from another variable x. A regression line must involve 2 variables,
- Finding the "best-fit" line is the goal of simple linear regression.

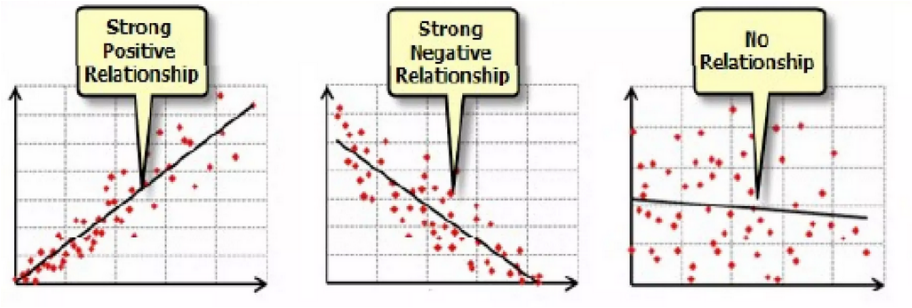
What is Simple Linear Regression?

Input, predictive or Independent varible x: this the variable whose value that is believed to influence the value of another variable,
Output, Response or Dependent variable y: this is the varaible whose value that is believed to be influenced by the value of another variable,
Best-fit Line: represents our model. It is the line that best-fits our data points. The line represents the best estimate of the y value for every given input of x,



Current Section

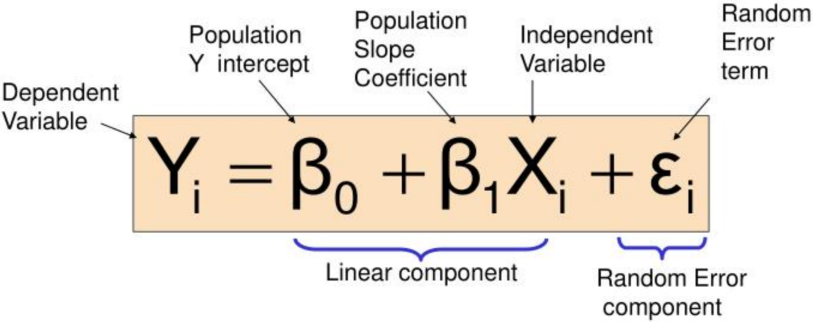
Types of Linear Regression Models



Current Section

Model equation

The simple linear regression equation provides an estimate of the population regression line.



The individual random error e_i have a mean of zero.

Assumptions of Regression

Use the acronym **LINE**

- Linearity: the underlying relationship between X and Y is linear,
- Independence of errors: error values are statistically independent,
- Normality of errors: error values are normally distributed for any given value of X,
- Equal Variance (homoscedasticity)! The probability distribution of the errors has constant variance.

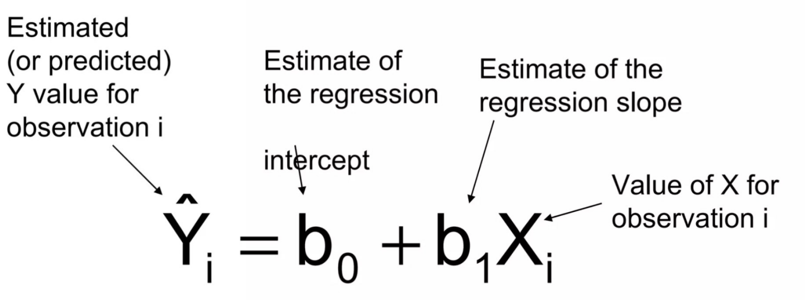
Current Section

Least Squarres Method

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squarred differences between y and $\hat{y}$

$\min \sum(y_i - \hat{y}_i)^2 = \min \sum(y_i - (b_0 + b_1 x_i))^2$

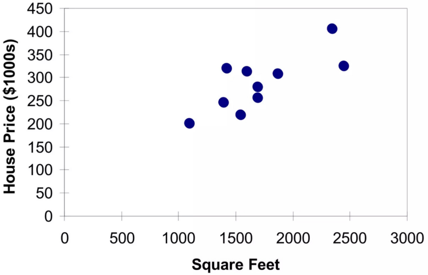
- b_0 is the estimated average value of Y when the value of X is zero,
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X.



Example

A real estate agent wishes to examine the relationship between the selling price of a home and its size measured in square feet.
A random sample of 10 houses is selected. The dependent variable Y is the house price in \$1000s. The independent variable X is the surface in squarre feet.

House Price	Square feet
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



Example

$\widehat{houseprice}$=98.24833 + 0.10977 square feet
Predict the price for a house with 2000 square feet: 98.25 + 0.1098(2000)=317.85.
The predicted price for a house with 2000 square feet is 317,850 dollars.

Current Section

Measures of variation

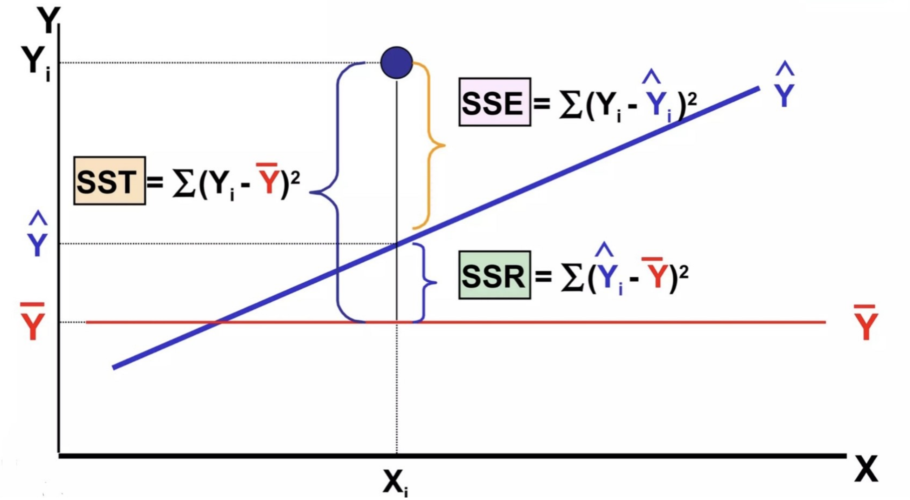
- SST= total sum of squarres
  Measures the variation of the Y_i values around their mean.
- SSR= regression sum of squares
  Explained variation attributable to the relationshiop between X and Y.
- SSE= error sum of squarres
  Variation attributable to factors other than the relationship between X and Y.

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

Measures of variation

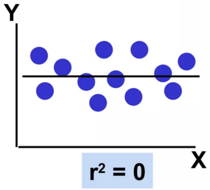


Current Section

Coefficient of Determination r^2

The coefficient of determination is the portion of total variation in the dependent variable that is explained by variation in the independent variable. It is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{regressionsumofsquarres}{totalsumofsquares}$$



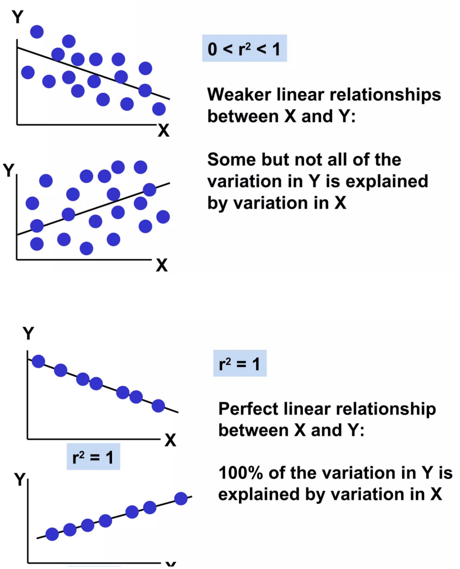
r² = 0

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

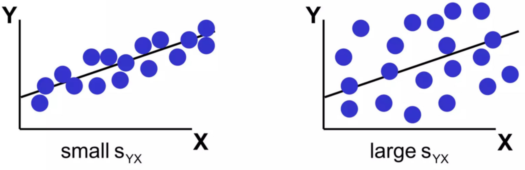
Note that r^2 values range b/w 0 and 1.

Examples of Approximate values



Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by
$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$
where SSE= error sum of squares and n=sample size.
- The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data, i.e. S_{YX}= \$41.33k is moderately small relative to house prices in the \$200 - \$300k range.



Current Section

Residual Analysis

The residual for observation i e_i is the difference between the observed and predicted data.
Check the assumptions of regression by examining the residuals

- Examine the linearity assumption,
- Evaluate the independence assumption,
- Evaluate normal distribution assumption assumption,
- Examine for constant variance for all levels of X.

Current Section

Questions

1. Define the following: regression, regression model, dependent variable.
2. Having that SST=SSR+SSE, what is the best case scenario and the worst? Explain.
3. What does the coefficient of determination r^2 mean.
4. What does residual analysis mean?
5. Under what conditions is the linear regression model reliable?