Artificial Intelligence, Blockchain, e Criptovalute nello Sviluppo Software

Lezioni 10, 11 e 12: Fondamenti di Data Science per l'analisi dello sviluppo software

Giancarlo Succi
Dipartimento di Informatica – Scienza e Ingegneria
Università di Bologna
g.succi@unibo.it

Content

- Linear Regression
- Correlation and Covariance
- Toward Inference

Linear Regression

Linear Regression – Problem 1

- Suppose that:
  - I want to relate two random scalar phenomena, X and Y, to identify the relationships existing between them,
  - I can measure their values several times i, so I can have a set of pairs (x_i, y_i) with i spanning the interval of observation, say $i \in [0 \ldots n-1]$

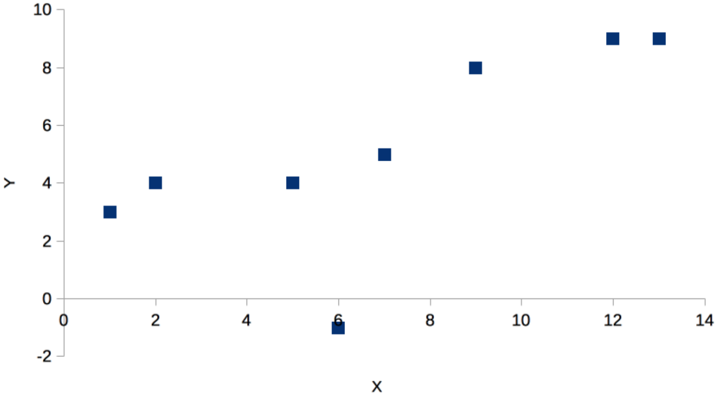
i	**X**	**Y**
0	1	3
1	2	4
2	5	4
3	6	-1
4	7	5
5	9	8

Linear Regression – Problem 2

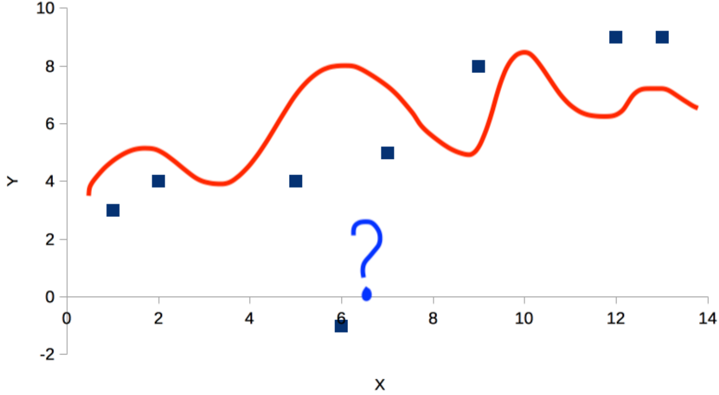
Using a simple and common approach, I may try to build a relationship between the two phenomena. However:

- What kind of relationships I am going to look for?
- How do I build it?

In other words, I have this set of points:



How can I build a line that represent the relationships between these two sets?



Linear Regression – Definition

We need to define:

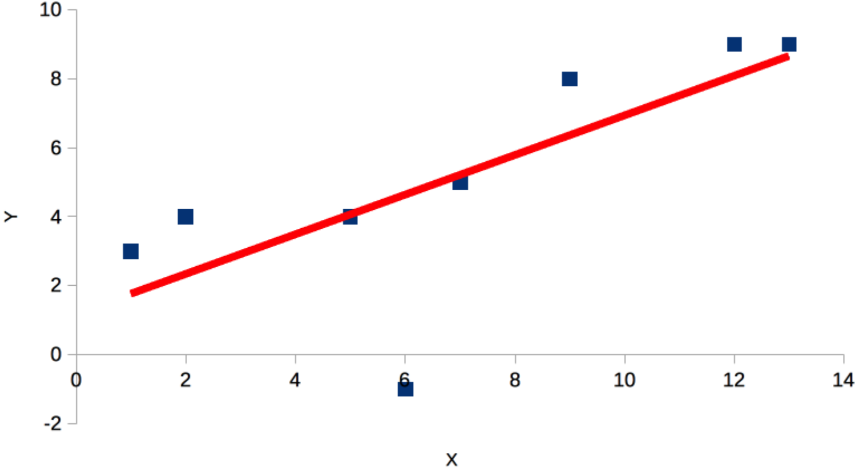
- A mean function that represents the relationship that I hypothesize between the phenomena X and Y
- A cost-minimization function to define the parameters of the mean function

We will use initially:

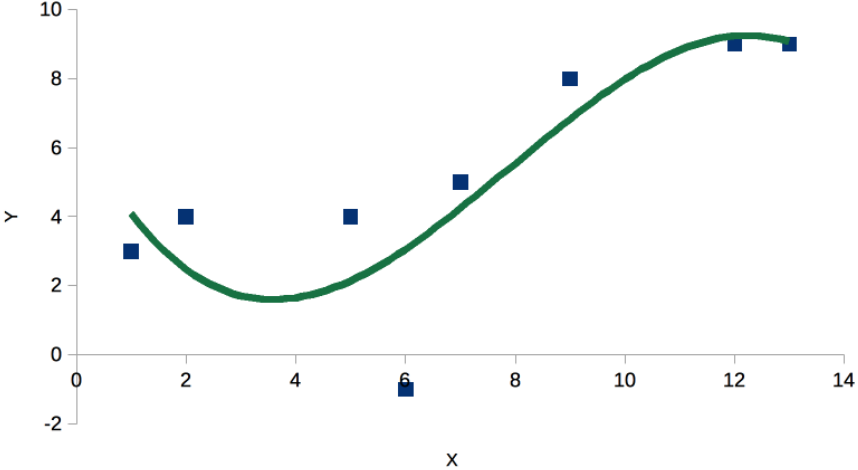
- As mean function the simple line
- As cost function the square of the errors between the modeled values and the real values

We define Ordinary Least Squares (OLS) Linear Regression as a simple line that minimizes a square error between modelled values and real values.

This is what we would like to build:

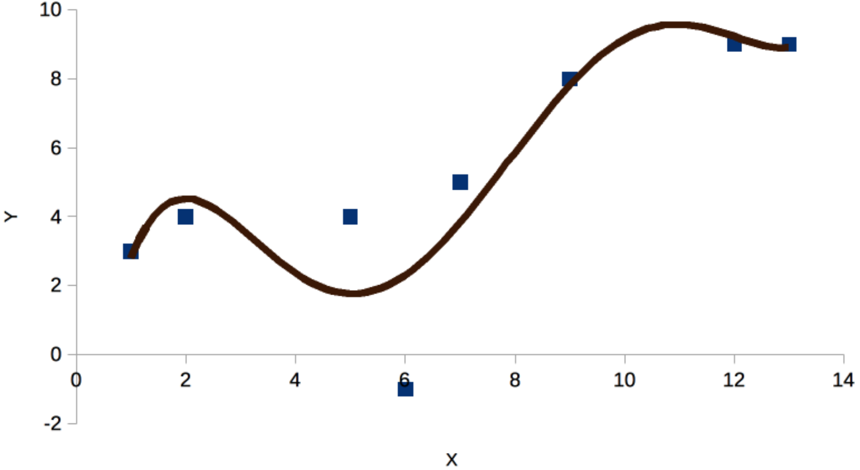


But we could have used as a mean function a cubic function:

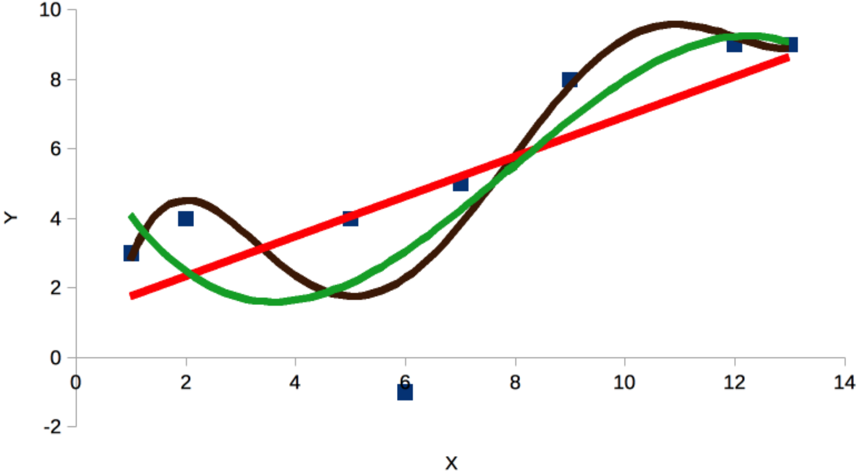


Linear Regression – Alternative Goal 2

But we could have used as a mean function a fifth order function:



What are the differences between all these 3?



Linear Regression – Formula (1/3)

I want to build a model of the kind:

$$Y = \theta_0 + \theta_1 X$$

Where X and Y are the phenomena that we are measuring.

Note:

- we know that there is no line passing for n arbitrary points with $3 \leq n$
- we need to introduce an approximation

$$\hat{Y} = \theta_0 + \theta_1 \hat{X} + \epsilon$$

- in our case ϵ is the error introduced by the approximation
- as we said, our cost function, our distance from the model, will be the square of the error ϵ^2
- θ_0 and θ_1 are called the regression coefficients

Linear Regression – Formula (2/3)

Altogether:

- we have a set of pairs (x_i, y_i) with $i \in [0 \ldots n-1]$
- we want to build n linear equations of the kind (the mean function):

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

- and we start with an approximation of the kind:

$$\hat{y}_i = \theta_0 + \theta_1 x_i$$

Linear Regression – Formula (3/3)

Altogether:

- our goal is to compute θ_0 and θ_1 that minimize the quadratic error (the cost function)

$$\sum_{i=0}^{n-1} \epsilon_i^2$$

- notice that:
  - we will denote as (x_i,y_i) the original data
  - we will denote as $(\hat{x}_i,\hat{y}_i)$ the approximation that we obtain in the linear regression
  - x_i and $\hat{x}_i$ are the same
  - there could be errors in the slides and you get extra credits by finding them

Linear Regression – Computation

- Since

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i$$

- therefore

$$\epsilon_i = y_i - \theta_0 - \theta_1 x_i$$

- we need to minimize:

$$\sum_{i=0}^{n-1} \epsilon_i^2 = \sum_{i=0}^{n-1} (y_i - \theta_0 - \theta_1 x_i)^2$$

- we need to zero the two partial derivatives:

$$\frac{\partial \sum_{i=0}^{n-1} (y_i - \theta_0 - \theta_1 x_i)^2}{\partial \theta_i}$$

- so we have to solve two simple equations and then to check the Hessian

Linear Regression – Computation for θ_0

$$\frac{\partial \sum_{i=0}^{n-1} (y_i - \theta_0 - \theta_1 x_i)^2}{\partial \theta_0} = 0 \Rightarrow$$

$$2 \sum_{i=0}^{n-1} (y_i - \theta_0 - \theta_1 x_i) = 0 \Rightarrow$$

$$\sum_{i=0}^{n-1} (y_i - \theta_0 - \theta_1 x_i) = 0$$

Linear Regression – Computation for θ_1

$$\frac{\partial \sum_{i=0}^{n-1}(y_i - \theta_0 - \theta_1 x_i)^2}{\partial \theta_1} = 0 \Rightarrow$$

$$2\sum_{i=0}^{n-1} x_i(y_i - \theta_0 - \theta_1 x_i) = 0 \Rightarrow$$

$$\sum_{i=0}^{n-1} x_i(y_i - \theta_0 - \theta_1 x_i) = 0$$

From the first equation:

$$\sum_{i=0}^{n-1}(\theta_0) = \sum_{i=0}^{n-1}(y_i - \theta_1 x_i) \Rightarrow$$

$$\sum_{i=0}^{n-1}(\theta_0) = \sum_{i=0}^{n-1}(y_i) - \theta_1 \sum_{i=0}^{n-1}(x_i) \Rightarrow$$

$$n\theta_0 = n\bar{y} - n\theta_1\bar{x} \Rightarrow$$

$$\theta_0 = \bar{y} - \theta_1\bar{x}$$

Linear Regression – In the second equation

$$\sum_{i=0}^{n-1} x_i(y_i - \theta_0 - \theta_1 x_i) = 0 \Rightarrow$$

$$\sum_{i=0}^{n-1} x_i y_i - \theta_0 \sum_{i=0}^{n-1} x_i - \theta_1 \sum_{i=0}^{n-1} x_i^2 = 0 \Rightarrow$$

$$\sum_{i=0}^{n-1} x_i y_i - n\theta_0 \bar{x} - n\theta_1 \bar{x^2} = 0 \Rightarrow$$

Linear Regression – Combining the result

Substituting $\theta_0 = \bar{y} - \theta_1 \bar{x}$:

$$\sum_{i=0}^{n-1} x_i y_i - n(\bar{y} - \theta_1 \bar{x})\bar{x} - n\theta_1 \overline{x^2} = 0 \Rightarrow$$

$$\sum_{i=0}^{n-1} x_i y_i - n\bar{y}\bar{x} + n\theta_1 \bar{x}^2 - n\theta_1 \overline{x^2} = 0$$

$$\sum_{i=0}^{n-1} x_i y_i - n\bar{y}\bar{x} + n\theta_1 \overline{x^2} - n\theta_1 \bar{x}^2 = 0$$

$$n\theta_1 (\overline{x^2} - \bar{x}^2) = \sum_{i=0}^{n-1} x_i y_i - n\bar{y}\bar{x}$$

Linear Regression – Final step

$$\theta_1 = \frac{\sum_{i=0}^{n-1} x_i y_i - n\bar{y}\bar{x}}{n(\overline{x^2} - \bar{x}^2)} = \frac{\frac{\sum_{i=0}^{n-1} x_i y_i}{n} - \cancel{n}\bar{y}\bar{x}}{\cancel{n}(\overline{x^2} - \bar{x}^2)} = \frac{\frac{\sum_{i=0}^{n-1} x_i y_i}{n} - \bar{y}\bar{x}}{(\overline{x^2} - \bar{x}^2)}$$

$$\theta_1 = \frac{Cov(x, y)}{Var(x)}$$

Which we can also write as:

$$\theta_1 = \frac{\sum_{i=0}^{n-1}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n-1}(x_i - \bar{x})^2}$$

Going back to our exercise...

Using the formula above we obtain that for the following dataset:

i	**X**	**Y**
0	1	3
1	2	4
2	5	4
3	6	-1
4	7	5
5	9	8
6	12	9
7	13	9

We have an equation:

$$\hat{Y} = \theta_0 + \theta_1 \hat{X}$$

with:

- $\theta_0 = 1.179$ and $\theta_1 = 0.574$

Our model

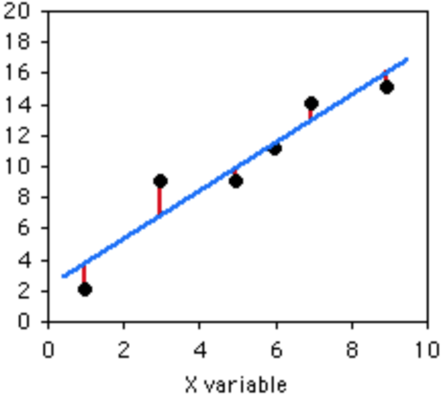
i	**X**	**Y**	$\hat{Y}$	ϵ
0	1	3	1.753	1.247
1	2	4	2.327	1.673
2	5	4	4.049	-0.049
3	6	-1	4.623	-5.623
4	7	5	5.197	-0.197
5	9	8	6.345	1.655
6	12	9	8.067	0.933
7	13	9	8.641	0.359

Linear Regression – Exercise

Build a linear regression for the following dataset:

X	Y
1	2
3	9
5	9
6	11
7	14
9	15

Linear Regression – Exercise



Linear Regression – Exercise

The regression equation for these numbers is $\hat{y} = 2.0286 + 1.5429x$. Now, fill the blanks using such equation and calculate the sum of squared deviations (last column).

x	y	Predicted y ($\hat{y}$)	Deviate from predicted (abs.)	Squared deviate
1	2			
3	9			
5	9			
6	11			
7	14			
9	15			

Results. The sum of squared deviations: 10.8

x	y	Predicted y ($\hat{y}$)	Deviate from predicted (abs.)	Squared deviate
1	2	3.57	1.57	2.46
3	9	6.66	2.34	5.48
5	9	9.74	0.74	0.55
6	11	11.29	0.29	0.08
7	14	12.83	1.17	1.37
9	15	15.91	0.91	0.83

Linear Regression – Modeling

In fact, we might think to use linear regression to model phenomena, assuming a linear dependence between input (the collected parameters) and output.

Here are some "real world" examples (w.r.t. certain assumptions):

- Impact of SAT Score (or GPA) on College Admissions;
- Impact of product price on number of sales;
- Impact of rainfall amount on the number of fruits yielded;
- Impact of blood alcohol content on coordination.

Linear Regression – Evaluation

We can evaluate the quality of linear regression, i.e. assess how good the model for the data that we have:

- by the sum of squares of residuals;
- by the coefficient of determination.

The sum of squared errors

The sum of squares of residuals, also called the residual sum of squares:

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2$$

In the case above SS_{res} is equal to 39.751672.

The coefficient of determination (R^2)

The coefficient of determination describes the proportion of variance of the dependent variable explained by the regression model. If the regression model is "perfect," SS_{res} is zero, and R^2 is 1.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

The total sum of squares:

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

In the example above

$$SS_{tot} = \sum_i (y_i - \bar{y})^2 = 82.875$$

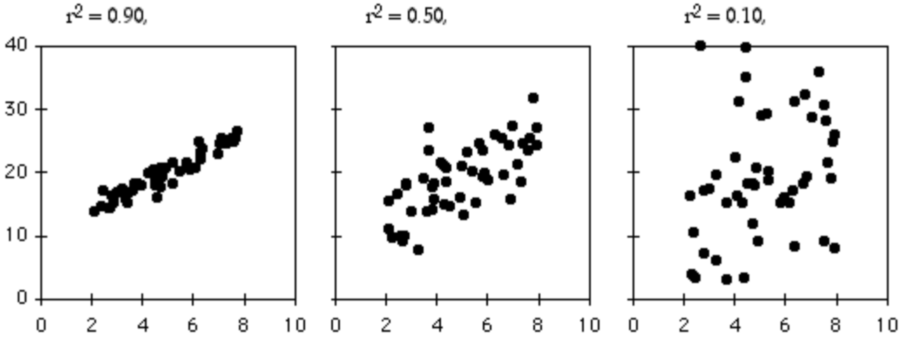
Remember that:

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 = 39.751672$$

Therefore:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{39.751672}{82.875} = 0.5203$$

Coefficient of determination (R^2)



Multivariate Linear Regression

- The "X" variable is often called "feature" in machine learning.
- Indeed, we could have multiple features, say, n.
- If we also have m observations, we could build a system of m equations of the kind:

$$y_i = \boldsymbol{\theta}^T \cdot \boldsymbol{x}_i + \epsilon_i, i = 1 \dots m$$

- and then we will build our linear regression (approximation) as:

$$\hat{y}_i = \boldsymbol{\theta}^T \cdot \hat{\boldsymbol{x}}_{\boldsymbol{i}}, i = 1 \dots m$$

- where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_{\boldsymbol{i}}$– are vectors of $n + 1$ features for the i-th observation

Question: Why here we use $n + 1$?

A closed-form solution of Linear Regression

To find the value of $\boldsymbol{\theta}$, there is a closed-form solution, a mathematical equation that gives the result directly.
This is called the **Normal Equation**:

$$\theta = (\boldsymbol{X} \cdot \boldsymbol{X}^T)^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{y}$$

Derivation of the closed-form solution (1/4)

- We start considering a set of m equations of the form:

$$\hat{y}_i = \boldsymbol{\theta}^T \boldsymbol{x}_i, i = 1 \dots m$$

  where $\boldsymbol{x}_i$ has dimension $n+1$
- We move all the model in matrix format:

$$\hat{\boldsymbol{y}} = \boldsymbol{X} \cdot \boldsymbol{\theta}$$

- Notice that $\hat{\boldsymbol{y}}$ and $\boldsymbol{y}$ have dimension (m,1), $\boldsymbol{X}$ (m,n+1), and θ (n+1,1). $\boldsymbol{X} \cdot \boldsymbol{\theta}$ has therefore dimension (m,1) as it should be.
- The error vector $\boldsymbol{\epsilon}$ is defined for each pair as:

$$\boldsymbol{\epsilon} = \hat{\boldsymbol{y}} - \boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y}$$

- And the square of the error is:

$$(\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})^T (\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})$$

Derivation of the closed-form solution (2/4)

- To determine the values of the parameters we take the partial derivatives and we null them:

$$\frac{\partial (\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})^T (\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})}{\partial \boldsymbol{\theta}} = 0$$

- Now we evaluate:

$$\frac{\partial (\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})^T (\boldsymbol{X} \cdot \boldsymbol{\theta} - \boldsymbol{y})}{\partial \boldsymbol{\theta}} =$$

$$= \frac{\partial ((\boldsymbol{X} \cdot \boldsymbol{\theta})^T (\boldsymbol{X} \cdot \boldsymbol{\theta}) - (\boldsymbol{X} \cdot \boldsymbol{\theta})^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{X} \cdot \boldsymbol{\theta} + \boldsymbol{y}^T \boldsymbol{y})}{\partial \boldsymbol{\theta}} =$$

$$= \frac{\partial ((\boldsymbol{X} \cdot \boldsymbol{\theta})^T (\boldsymbol{X} \cdot \boldsymbol{\theta}) - 2(\boldsymbol{X} \cdot \boldsymbol{\theta})^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y})}{\partial \boldsymbol{\theta}}$$

Derivation of the closed-form solution (3/4)

- Now we can consider that:

$$\frac{\partial(\boldsymbol{y}^T\boldsymbol{y})}{\partial\boldsymbol{\theta}} = 0 \quad \text{and} \quad \frac{\partial((\boldsymbol{X}\cdot\boldsymbol{\theta})^T\boldsymbol{y})}{\partial\boldsymbol{\theta}} = \boldsymbol{X}^T\boldsymbol{y}$$

- Notice that $\boldsymbol{X}^T\boldsymbol{y}$ has dimension (n+1,m) · (m,1), that is, (n+1,1).
- We can finally conclude that:

$$\frac{\partial((\boldsymbol{X}\cdot\boldsymbol{\theta})^T(\boldsymbol{X}\cdot\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$$

- Notice that $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}$ has dimension (n+1,m) · (m,n+1) · (n+1,1), that is, (n+1,1) as it should be.

Derivation of the closed-form solution (4/4)

- Substituting the results in the original formula:

$$2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{X}^T\boldsymbol{y} = 0 \Rightarrow$$

$$\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{y} \Rightarrow$$

- Notice that $\boldsymbol{X}^T\boldsymbol{X}$ has dimension (n+1,m) $\cdot$ (m,n+1), that is, (n+1,n+1). Notice that m $\gg$ n, so we *hope* that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible.

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

- QED.

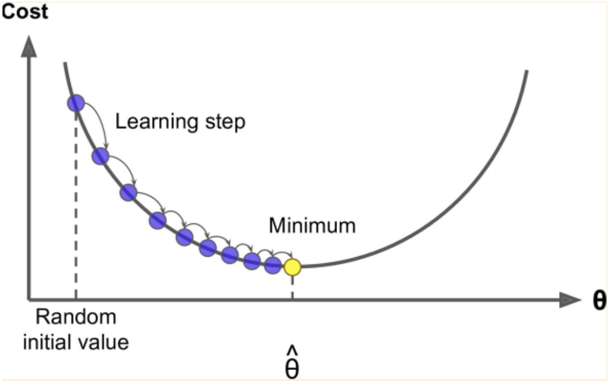
Computational complexity

The Normal Equation computes the inverse of $X^T \cdot X$, which is an n x n matrix (where n is the number of features).

The computational complexity of inverting such a matrix is typically about $O(n^{2.4})$ to $O(n^3)$ (depending on the implementation).

In other words, if you double the number of features, you multiply the computation time by roughly $2^{2.4} = 5.3$ to $2^3 = 8$.

Gradient Descent is a very generic optimization algorithm capable of finding optimal solutions to a wide range of problems. The general idea of Gradient Descent is to tweak parameters iteratively in order to minimize a cost function.



Gradient Descent - Computation

To implement Gradient Descent, you need to compute the gradient of the MSE cost function with regards to each model parameter θ_j.
Mean squared error (MSE) cost function for a Linear Regression model:

$$MSE(\theta) = \frac{1}{m}\sum_{k=1}^{m}(\boldsymbol{\theta}^T \cdot \boldsymbol{x}^{(k)} - \boldsymbol{y}^{(k)})^2$$

$x^{(k)}$ - k-th observation vector ($x^{(k)}$ is an n-dimensional vector)

Gradient Descent - Computation

To implement Gradient Descent, you need to compute the gradient of the MSE cost function with regards to each model parameter θ_j.

$$\frac{\partial}{\partial \theta_j} MSE(\theta) = \frac{2}{m} \sum_{i=1}^{m} (\theta^T \cdot \boldsymbol{x}^{(i)} - \boldsymbol{y}^{(i)}) x_j^{(i)}$$

Gradient Descent - Computation

In vector form:

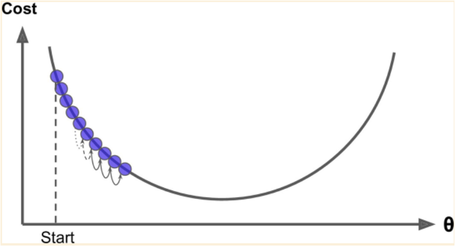
$$\nabla_\theta MSE(\theta) = \frac{2}{m}\boldsymbol{X}^T(\boldsymbol{X}\cdot\boldsymbol{\theta} - \boldsymbol{y})$$

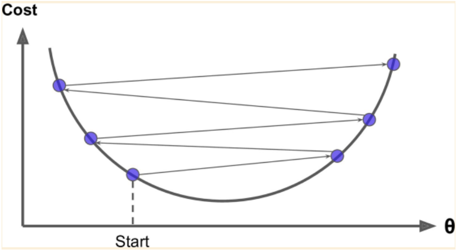
We update vector $\boldsymbol{\theta}$ step by step:

$$\boldsymbol{\theta}^{next} = \boldsymbol{\theta} - \eta\nabla_\theta MSE(\theta)$$

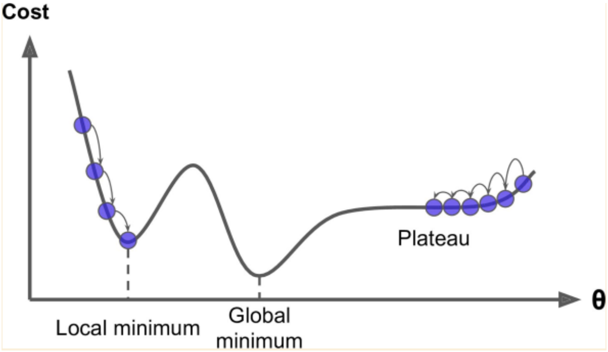
η – learning rate

Learning rate





Pitfalls of Gradient Descent



Linear Regression and Machine Learning

Linear Regression is a statistical model developed in the field of Regression Analysis.
Later it was borrowed for the use of Machine Learning field.

Terminology difference

Regression analysis	**Machine Learning**
estimation, fitting	training, learning
regressors	features
response	target

References

1) http://www.cs.umd.edu/~djacobs/CMSC426/Convolution.pdf

2) https://www.researchgate.net/post/Difference_between_convolution_and_correlation

3) https://www.tutorialspoint.com/signals_and_systems/convolution_and_correlation.htm

Correlation and Covariance

Content

- Covariance
- Correlation (aka Pearson product-moment correlation coefficient)
- Relationship between Pearson correlation and linear regression

Covariance

- To proceed further with our analysis we will use the concept of **covariance**, which we have already seen
- It expresses the degree in which the variation of a random variable is connected to the variation of another random variable
- It is defined as follows:
  - Given two random variables X and Y

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Covariance – graphically

Source : https://www.geeksforgeeks.org/mathematics-covariance-and-correlation

About the covariance - 1

- We notice that:
  - The covariance of a random variable with itself is the variance:
    $$Cov(X, X) = Var(X)$$
  - There is a similar property as for the variance
    $$Cov(X, Y) = E(XY) - E(X)E(Y)$$
    since:
    $$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] =$$
    $$= E(XY) - E(XE(Y)) - E(E(X)Y) + E(E(X)E(Y)) =$$
    $$= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) =$$
    $$= E(XY) - E(X)E(Y)$$
    QED.

About the covariance - 2

- The covariance is symmetric:

$$Cov(X, Y) = Cov(Y, X)$$

- The covariance is linear with respect to multiplications by constants:

$$(\forall a, b \in \mathbb{R}) \;\; Cov(aX, bY) = abCov(X, Y)$$

- If $e \sim N(0, \sigma)$, $Cov(X, e) = 0$

$$Cov(X, e) = E(Xe) - E(X)E(e)$$

Moreover, X and e are independent and $E(e) = 0$
QED.

Pearson Correlation Coefficient

- AKA Pearson product-moment correlation coefficient or just correlation coefficient
- It expresses the linear correlation between two random variables
- It is defined as follows:
  - Given two random variables X and Y

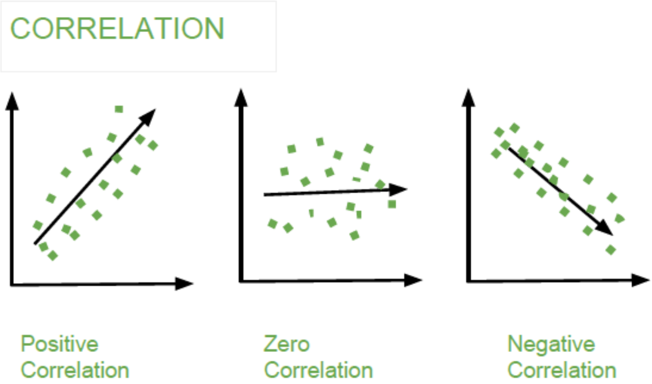
$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

  - Where:

$$\sigma_Z = \sqrt{Var(Z)}$$

For the time being we intentionally ignore the difference between sample and population.

Pearson Correlation Coefficient – graphically



Source : `https://www.geeksforgeeks.org/mathematics-covariance-and-correlation`

About the Pearson Correlation Coefficient (1/2)

- The Pearson correlation coefficient is also often expressed as:

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

- It is symmetric: $r_{X,Y} = r_{Y,X}$
- It is invariant with respect to multiplications by, and additions of constants:
  $(\forall a, b, c, d \in \mathbb{R}, b \neq 0, d \neq 0) \;\; r_{X,Y} = r_{(a+bX),(c+dY)}$

About the Pearson Correlation Coefficient (2/2)

- The Pearson correlation coefficient ranges from -1 to 1: $-1 \leq r_{X,Y} \leq 1$
  - $r_{X,Y} = 1$ means perfect linear relationship
    - all points lie on a monotonically increasing line
  - $r_{X,Y} = -1$ means perfect opposite linear relationship
    - all points lie on a monotonically decreasing line
  - $r_{X,Y} = 0$ means no linear relationship between X and Y

Back to Linear Regression (1/2)

- We now focus our attention to the case of the case of the linear regression
- Suppose we have two phenomena that we want to measure, X and Y
- Let us assume
  - that there is a linear relationship between them
  - that I can express the data I collect as:

$$\boldsymbol{y} = \boldsymbol{\theta}_0 + X\boldsymbol{\theta}_1 + \boldsymbol{\epsilon}$$

  - where ϵ is a stationary gaussian process $N(0, \sigma^2)$
- We know the solution that minimizes the square error

Back to Linear Regression (2/2)

- From this solution we have extracted the coefficient of determination
$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$
- Where:
  - $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$
  - SS_{res} is the distance between the reality and the 1-degree best approximation, that is, the OLS model
- and
  - $SS_{tot} = \sum_i (y_i - \overline{y})^2$
  - SS_{tot} is the distance between the reality and the 0-degree best approximation, that is, the mean
- I want to know the relationship between R^2 and the correlation coefficient between X and Y, $r_{X,Y}$

Our goal – understanding R and r

- We focus on 1D
- We are now going to prove a fundamental point.
- Under the assumption that the noise is gaussian and centered in 0, in a linear regression:

$$R^2 = r_{X,Y}^2$$

$R^2 = r^2_{X,Y}$ (1/4)

- Since

$$\hat{y} = \theta_0 + \theta_1 x$$

- we have from above (see page 56) that:

$$r_{X,Y} = r_{\hat{Y},Y}$$

- We define now the explained sum of squares (ESS)
  - $ESS = \sum_i (\hat{y}_i - \overline{y})^2$
  - ESS is the additional knowledge we get on the random variable using a polynomial of degree 1 vs. using a polynomial of degree 0
- We will now prove that **under our hypotheses**:

$$ESS + SS_{res} = SS_{tot}$$

$[R^2 = r^2_{X,Y}]$ – $ESS + SS_{res} = SS_{tot}$ (1/6)

- We start from:

$$(y_i - \overline{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

- which we square:

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

- and then we sum:

$$\sum_i (y_i - \overline{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \sum_i (\hat{y}_i - \overline{y})^2$$

Source with modifications: `https://en.wikipedia.org/wiki/Explained_sum_of_squares`

$[R^2 = r^2_{X,Y}]$ - $ESS + SS_{res} = SS_{tot}$ (2/6)

- Now we focus on:

$$\sum_i 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

- and we want to prove that it is 0, that is $\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = 0$; considering:

$$y_i = \hat{y}_i + \epsilon_i$$

$$E(y_i) = E(\hat{y}_i + \epsilon_i) = E(\hat{y}_i) + E(\epsilon_i) = E(\hat{y}_i)$$

because ϵ is a stationary gaussian process $N(0, \sigma^2)$

Source with modifications: https://en.wikipedia.org/wiki/Explained_sum_of_squares

- We can build a system:

$$\begin{cases} \hat{y}_i = \theta_0 + \theta_1 x_I \\ \overline{y} = \theta_0 + \theta_1 \overline{x} \end{cases}$$

- from which we deduce by subtraction:

$$\hat{y}_i - \overline{y} = \theta_1 (x_i - \overline{x})$$

- remembering that:

$$\theta_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Source with modifications: `https://en.wikipedia.org/wiki/Explained_sum_of_squares`

- So:

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_i (y_i - \hat{y}_i)(\theta_1(x_i - \overline{x})) =$$

$$= \theta_1 \sum_i (y_i - \hat{y}_i)(x_i - \overline{x})$$

- Now, let's consider that:

$$(y_i - \hat{y}_i) = y_i - \hat{y}_i + \overline{y} - \overline{y} = (y_i - \overline{y}) - (\hat{y}_i - \overline{y}) =$$

$$= (y_i - \overline{y}) - \theta_1(x_i - \overline{x})$$

- Substituting $(y_i - \hat{y}_i)$ above we get:

$$\theta_1 \sum_i (y_i - \hat{y}_i)(x_i - \overline{x}) = \theta_1 \sum_i [(y_i - \overline{y}) - \theta_1(x_i - \overline{x})](x_i - \overline{x})$$

Source with modifications: `https://en.wikipedia.org/wiki/Explained_sum_of_squares`

- We can conclude:

$$\theta_1 \sum_i [(y_i - \overline{y}) - \theta_1(x_i - \overline{x})](x_i - \overline{x}) =$$

$$= \theta_1 [\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \theta_1 (x_i - \overline{x})(x_i - \overline{x})] =$$

$$= \theta_1 [\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \frac{\sum_j (x_j - \overline{x})(y_j - \overline{y})}{\sum_j (x_j - \overline{x})^2} (x_i - \overline{x})^2] =$$

Source with modifications: `https://en.wikipedia.org/wiki/Explained_sum_of_squares`

- And simplifying what is in [•]:

$$\sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_i \frac{\sum_j (x_j - \overline{x})(y_j - \overline{y})}{\sum_j (x_j - \overline{x})^2}(x_i - \overline{x})^2 =$$

$$= \sum_i (y_i - \overline{y})(x_i - \overline{x}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) \sum_i \frac{(x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} =$$

$$= \sum_i (x_i - \overline{x})(y_i - \overline{y}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) \frac{\sum_i (x_i - \overline{x})^2}{\sum_j (x_j - \overline{x})^2} =$$

$$= \sum_i (x_i - \overline{x})(y_i - \overline{y}) - \sum_j (x_j - \overline{x})(y_j - \overline{y}) = 0$$

QED.

Source with modifications: `https://en.wikipedia.org/wiki/Explained_sum_of_squares`

$R^2 = r^2_{X,Y}$ (2/4)

- Now we know that, under the assumption to deal with a Gaussian noise centered in 0 we have:

$$ESS + SS_{res} = SS_{tot}$$

- Under this hypothesis we have:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = \frac{ESS}{SS_{tot}}$$

$R^2 = r^2_{X,Y}$ (3/4)

- We now consider the square of $r_{X,Y} = r_{\hat{Y},Y}$

$$r^2_{\hat{Y},Y} = \left(\frac{Cov(\hat{Y},Y)}{\sqrt{Var(Y)Var(\hat{Y})}} \right)^2 = \frac{Cov(\hat{Y},Y)Cov(\hat{Y},Y)}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{Cov(\hat{Y},\hat{Y}+\epsilon)Cov(\hat{Y},\hat{Y}+\epsilon)}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{(Cov(\hat{Y},\hat{Y}) + Cov(\hat{Y},\epsilon))(Cov(\hat{Y},\hat{Y}) + Cov(\hat{Y},\epsilon))}{Var(Y)Var(\hat{Y})} =$$

$$= \frac{Cov(\hat{Y},\hat{Y})Cov(\hat{Y},\hat{Y})}{Var(Y)Var(\hat{Y})}$$

Source with modifications: `https://economictheoryblog.com/2014/11/05/proof/`

$R^2 = r^2_{X,Y}$ (4/4)

- But we know that $Cov(\hat{Y}, \hat{Y}) = Var(\hat{Y})$, therefore we get that

$$r^2_{X,Y} = \frac{Var(\hat{Y})Var(\hat{Y})}{Var(Y)Var(\hat{Y})} = \frac{Var(\hat{Y})}{Var(Y)} =$$

$$= \frac{\frac{\sum_i (\hat{y}_i - \overline{\hat{y}})^2}{n}}{\frac{\sum_i (y_i - \overline{y})^2}{n}} = \frac{\sum_i (\hat{y}_i - \overline{\hat{y}})^2}{\sum_i (y_i - \overline{y})^2} = \frac{ESS}{SS_{tot}}$$

since we have already proven that $\overline{y} = \overline{\hat{y}}$

QED

Source with modifications: `https://economictheoryblog.com/2014/11/05/proof/`

Comment on $R^2 = r_{X,Y}^2$

- This is a major result
- It is the center of our subsequent investigation, in the case of normality of error we can model, interconnect, and understand relationships in an easy way
- The next question is on how the slope of the regression line (θ_1) relates to the correlation coefficient $r_{X,Y}$

$r_{X,Y}$ and θ_1

- We know that:

$$\theta_1 = \frac{Cov(X,Y)}{Var(X)}$$

- And that:

$$r_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- Therefore:

$$\theta_1 Var(X) = r_{X,Y} \sigma_X \sigma_Y$$

- We can then conclude that:

$$\theta_1 = \frac{\sigma_X \sigma_Y}{Var(X)} r_{X,Y}$$

$$r_{X,Y} = \frac{Var(X)}{\sigma_X \sigma_Y} \theta_1$$

Comment on $r_{X,Y} \sim \theta_1$

- $r_{X,Y}$ and θ_1 are therefore directly and monotonically proportional
- It means that a positive relationships implies a positive slope and viceversa

General remark

- Right now we work with samples of larger populations of data
- We measure properties of samples, like mean, standard deviation, covariance, correlation coefficient
- All these properties are also random variable and have a distribution
- Our question is therefore, what kind of distribution is the one of the correlation coefficient
- Knowing its distribution allows us to understand the relationships existing between the variables it connect

Toward Inference

Content

- Premises of the Law of Large Numbers
- Markov's inequality
- Chebyshev's inequality
- Proof of the Law of Large Numbers
- Central Limit Theorem in the Linderberg-Lévy formulation
- Moment
- Moment generating function
- Proof of the Central Limit Theorem in the Linderberg-Lévy formulation

Last words...

- Right now we work with samples of larger populations of data
- We measure properties of samples, like mean, standard deviation, covariance, correlation coefficient
- All these properties are also random variable and have a distribution
- Our question is therefore, what kind of distribution is the one of the correlation coefficient
- Knowing its distribution allows us to understand the relationships existing between the variables it connect

Knowing the sample ...

- What can we infer of populations now that I know the properties of the sample?
- Now we know the mean, the standard deviation, the distribution of the sample, what would be the mean, the standard deviation, and the distribution of the population?
- Moreover, from two samples we can build a regression, what would be the regression of the population?

We start from the mean

- We suppose that we have an unknown population $\mathfrak{P}$ of entities on a ratio scale from which we extract n samples $\mathfrak{S}_i$ with $i \in [1 \ldots n]$
- Each sample i is composed by $\mathfrak{n}_\mathfrak{i}$ elements $e_{i,j}$ with $j \in [1 \ldots \mathfrak{n}_\mathfrak{i}]$
- We can compute the set of the means of each sample $\mathfrak{S}_i$, $\mathfrak{m}_i$ with $i \in [1 \ldots n]$
- $\mathfrak{m}_i$ is a random variable, so we would like to know what is its structure
- There are two fundamental theorems about the distributions of such $mathfrakm_i$, the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT)
- Since we are not making **any** assumption on the population $\mathfrak{P}$, we ignore it and consider simply a sequence of random variables x_i.

LLN – Premises

- From now on, we will use the notation "iid" to denote the property of a set of random variables to be **i**ndependent and **i**dentically **d**istributed
- **Let $\{\mathfrak{Xn}_1, \mathfrak{Xn}_2, \ldots, \mathfrak{Xn}_n\}$ a set of n iid random variables drawn from a population with mean μ**
- Each $\mathfrak{Xn}_i$ could be considered the average of a sample $\mathfrak{S}_i$ of size 1, that is $\mathfrak{S}_i = \{\mathfrak{Xn}_i\}$
- **Let us consider $\overline{\mathfrak{Xn}}$, the average for this sample of size n**
- $\overline{\mathfrak{Xn}}$ is like the average of the n averages of each sample $\mathfrak{S}_i$

Source with modifications: https://en.wikipedia.org/wiki/Law_of_large_numbers

LLN – Weak formulation

- Let $\{\mathfrak{Xn}_1, \mathfrak{Xn}_2, \ldots, \mathfrak{Xn}_n\}$ a set of n iid random variables drawn from a population with mean μ
- Let us consider $\overline{\mathfrak{Xn}}$, the average for this sample of size n
- the Law of Large Number in its weak formulation states that:

$$(\forall \epsilon \in \mathbb{R}^+) \quad \lim_{n \to \infty} \mathbb{P}(|\overline{\mathfrak{Xn}} - \mu| > \epsilon) = 0$$

- This means that $\overline{\mathfrak{Xn}}$ tends to get the value of μ probabilistically

Source with modifications: `https://en.wikipedia.org/wiki/Law_of_large_numbers`

LLN – Proof (1/4)

- We are now going to prove LLN
- To do so, we need to prove two other interesting theorems:
  - The Markov's inequality
  - The Chebyshev's inequality

Source with modifications: `https://en.wikipedia.org/wiki/Law_of_large_numbers`

[LLN – Proof] Markov's inequality (1/3)

- The Markov's inequality put a first boundary on the distribution of a random variable
- Let $X \geq 0$ be a random variable with mean $\mu \in \mathbb{R}$
- Then:

$$(\forall k \in \mathbb{R}^+) \;\; \mathbb{P}(X \geq k) \leq \frac{\mu}{k}$$

- Proof:

$$\mu = \int_{-\infty}^{+\infty} x f_x(x) dx$$

Source with modifications: `https://en.wikipedia.org/wiki/Markov%27s_inequality`

[LLN – Proof] Markov's inequality (2/3)

- Since $X \geq 0$

$$\int_{-\infty}^{+\infty} x f_x(x) dx = \int_{0}^{+\infty} x f_x(x) dx =$$

And given $k \in \mathbb{R}^+$

$$= \int_{0}^{k} x f_x(x) dx + \int_{k}^{+\infty} x f_x(x) dx$$

Since $\int_0^k x f_x(x) dx \geq 0$

$$\mu \geq \int_{k}^{+\infty} x f_x(x) dx \geq k \int_{k}^{+\infty} f_x(x) dx = \mathbb{P}(X \geq k)$$

Source with modifications: `https://en.wikipedia.org/wiki/Markov%27s_inequality`

[LLN – Proof] Markov's inequality (3/3)

- Therefore we have

$$\mu \geq k\mathbb{P}(X \geq k)$$

- And from this we conclude:

$$\mathbb{P}(X \geq k) \leq \frac{\mu}{k}$$

Source with modifications: `https://en.wikipedia.org/wiki/Markov%27s_inequality`

[LLN – Proof] Chebyshev's inequality (1/3)

- The Chebyshev's inequality put a further limit on the distribution of a random variable
- Let X be a random variable with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$
- Then:

$$(\forall k \in \mathbb{R}^+) \;\; \mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- Proof:
  Let us define a new random variable

$$Y = (X - \mu)^2 \geq 0$$

Let us define

$$h = (k\sigma)^2$$

Source with modifications: `https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

[LLN – Proof] Chebyshev's inequality (2/3)

- By the Markov inequality we have for the nonnegative random variable Y and for the positive real h:

$$\mathbb{P}(Y \geq h) \leq \frac{\overline{Y}}{h}$$

- And this means:

$$\mathbb{P}((X-\mu)^2 \geq (k\sigma)^2) \leq \frac{\overline{(X-\mu)^2}}{(k\sigma)^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

Source with modifications: `https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

[LLN – Proof] Chebyshev's inequality (3/3)

- This can be rewritten into:

$$\mathbb{P}(|X - \mu| \geq |k\sigma|) \leq \frac{1}{k^2}$$

- Since we know that both k and σ are strictly positive:

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

QED

Source with modifications: `https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

LLN – Proof (2/4)

- We want to prove that:

$$(\forall \epsilon \in \mathbb{R}^+) \quad \lim_{n\to\infty} \mathbb{P}(|\overline{\mathfrak{Xn}} - \mu| > \epsilon) = 0$$

- we add the additional hypothesis that $\sigma_i > 0$
- Let us consider σ_i;
  - since the variables $\mathfrak{Xn}_i$ are iid

$$(\forall i, j)\ (\sigma_i = \sigma_j = \sigma)$$

  - we also assume that $\sigma > 0$
  - finally, since the variables $\mathfrak{Xn}_i$ are independent of one another:

$$Var(\overline{\mathfrak{Xn}}) = \frac{\sigma^2}{n} = \mathfrak{s}_{\mathfrak{n}}^2$$

Source with modifications: `https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

- Let us define:

$$k = \frac{\epsilon}{\mathfrak{s}_\mathfrak{n}}$$

k exists, since $\mathfrak{s}_\mathfrak{n}$ is strictly positive; therefore:

$$\epsilon = k\mathfrak{s}_\mathfrak{n}$$

- By Chebyshev's inequality we have:

$$\mathbb{P}(|\overline{\mathfrak{X}\mathfrak{n}} - \mu| \geq k\mathfrak{s}_\mathfrak{n}) \leq \frac{1}{k^2}$$

- That is:

$$\mathbb{P}(|\overline{\mathfrak{X}\mathfrak{n}} - \mu| \geq \epsilon) \leq \frac{{\mathfrak{s}_\mathfrak{n}}^2}{\epsilon^2}$$

Source with modifications: `https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

LLN – Proof (4/4)

- Since:

$$\mathfrak{s}_{\mathfrak{n}}^2 = \frac{\sigma^2}{n}$$

- We have that

$$\lim_{n\to\infty} \frac{\mathfrak{s}_{\mathfrak{n}}^2}{\epsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = \frac{\sigma^2}{\epsilon^2} \lim_{n\to\infty} \frac{1}{n} = 0$$

- Therefore:

$$\lim_{n\to\infty} \left(\mathbb{P}(|\overline{\mathfrak{X}\mathfrak{n}} - \mu| \geq \epsilon)\right) \leq \lim_{n\to\infty} \frac{\mathfrak{s}_{\mathfrak{n}}^2}{\epsilon^2} = 0 \Rightarrow \lim_{n\to\infty} \left(\mathbb{P}(|\overline{\mathfrak{X}\mathfrak{n}} - \mu| \geq \epsilon)\right) =$$

QED

Source with modifications: https://en.wikipedia.org/wiki/Chebyshev%27s_inequality

CLT – Lindeberg–Lévy formulation

- Let $\{\mathfrak{Xn}_1, \mathfrak{Xn}_2, \ldots, \mathfrak{Xn}_n\}$ a set of n iid random variables drawn from a population with mean μ and standard deviation σ
- Let us consider for this sample of size n:
  – the mean, $\overline{\mathfrak{Xn}}$
  – the variance, σ^2
  – the modulated difference, $\mathfrak{Dn}$, defined as:

$$\mathfrak{Dn} = \sqrt{n}(\overline{\mathfrak{Xn}} - \mu)$$

- Central Limit Theorem (Lindeberg–Lévy formulation):

$$\mathfrak{Dn} \xrightarrow{d} N(0, \sigma^2)$$

- This means that $\mathfrak{Dn}$ tends to be normal.

Source with modifications: `https://en.wikipedia.org/wiki/Central_limit_theorem`

[CLT – LLf] Moment (1/2)

- To prove the CLT – LLf we need to introduce a few additional statistical concepts that could be useful also in the continuation of this course series
- We define the r^{th} **moment** of a random variable X as the expected value of the r^{th} power of X; formally:

$$\mu_X(r) = E(X^r)$$

clearly: $\mu_X(1) = \mu_X = E(X)$

- Example:
  - If $P(X = 0) = 0.25$ and $P(X = 4) = 0.75$: $\mu_X(1) = 3$, $\mu_X(2) = 12$, $\mu_X(3) = 48$, and $\mu_X(4) = 192$

Source with modifications: `https://www.statlect.com/fundamentals-of-probability/moments`

[CLT – LLf] Moment (2/2)

- We define the **central r^{th} moment** of a random variable X as the expected value of the r^{th} deviation of X; formally:

$$\overline{\mu_X(r)} = E((X - \mu_X)^r)$$

clearly: $\overline{\mu_X(2)} = \sigma_X^2 = E((X - \mu_X)^2)$

- Example:
  - If $P(X = 0) = 0.25$ and $P(X = 4) = 0.75$:
    $\overline{\mu_X(1)} = 0$
    $\overline{\mu_X(2)} = 3$
    $\overline{\mu_X(3)} = -6$
    $\overline{\mu_X(4)} = 21$

Source with modifications: `https://www.statlect.com/fundamentals-of-probability/moments`

- Let X be a random variable defined over a set S and let f_X be its probability density function
- We define the **moment generating function** (**mgf**) M_X over X as:

$$M_X(t) = E(e^{tX}) = \int_S e^{tx} f_X(x)dx$$

if there exists $h \in \mathbb{R}^+$ so that $E(e^{tX})$ is defined in $(-h, +h)$

- Note that:
  - The mgf may not exist
  - The mgf has interesting properties

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/72/`

- Mgf and first moment:

$$\left[\frac{dM_X(t)}{dt}\right](t=0) = \mu_X(1) = \mu_X = E(X)$$

Since:

$$\left[\frac{dM_X(t)}{dt}\right](t=0) = \left[\frac{d\int_S e^{tx} f_X(x)dx}{dt}\right](t=0) =$$

$$= \left[\int_S x e^{tx} f_X(x)dx\right](t=0) = \int_S x e^{0x} f_X(x)dx = \int_S x f_X(x)dx =$$

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/73/`

- In general:

$$\left[\frac{d^n M_X(t)}{dt^n}\right](t=0) = \mu_X(n) = E(X^n)$$

- This comes from:

$$\frac{d^n M_X(t)}{dt^n} = \int_S x^n e^{tx} f_X(x)dx$$

- Proof. By induction, n=1, see above
- Let us assume that the proposition holds for n-1:

$$\frac{d^{n-1} M_X(t)}{dt^{n-1}} = \int_S x^{n-1} e^{tx} f_X(x)dx$$

Source with modifications: https://onlinecourses.science.psu.edu/stat414/node/73/

- We check it holds for n:

$$\frac{d^n M_X(t)}{dt^n} = \frac{d\left[\frac{d^{n-1}M_X(t)}{dt^{n-1}}\right]}{dt} =$$

$$= \frac{d\left[\int_S x^{n-1}e^{tx}f_X(x)dx\right]}{dt} = \int_S x^n e^{tx} f_X(x)dx$$

QED

- This confirms:

$$\left[\frac{d^n M_X(t)}{dt^n}\right](t=0) = \mu_X(n) = E(X^n)$$

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/73/`

[CLT – LLf] Mgf (5/10)

- Mgf and second moment:

$$\sigma_X^2 = E(X^2) - (E(X))^2 = \left[\frac{d^2 M_X(t)}{dt^2}\right](t=0) - \left\{\left[\frac{dM_X(t)}{dt}\right](t = 0\right.$$

And if the mean is 0:

$$\sigma_X^2 = \left[\frac{d^2 M_X(t)}{dt^2}\right](t=0)$$

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/73/`

[CLT – LLf] Mgf (6/10)

- Fundamental fact:

  > If the mgf for a random variable exists, it characterizes fully such random variable.

  Proof: omitted.
- It means that mgf and pdf are interchangeable
- We need now to determine the mgf for a normally distributed random variable $N(0, \sigma^2)$
- We will then use this to prove the CLT – LLf
- Let Z be a random variable, $Z \sim N(0, 1)$ then, the mgf for Z is:

$$M_Z(t) = e^{\frac{1}{2}t^2}$$

- Proof

$$M_Z(t) = \int_{-\infty}^{+\infty} e^{zt}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz = \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{zt-\frac{1}{2}z^2}dz =$$

$$= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{\frac{1}{2}(2zt-z^2)}dz = \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z^2-2zt+t^2-t^2)}dz =$$

$$= \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z^2-2zt+t^2)}e^{\frac{1}{2}t^2}dz = \int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}e^{\frac{1}{2}t^2}dz$$

$$= e^{\frac{1}{2}t^2}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-t)^2}dz = e^{\frac{1}{2}t^2}$$

QED

Source with modifications: `https://www.le.ac.uk/users/dsgp1/COURSES/MATHSTAT/6normgf.pdf`

Extending to the case of general Gaussian variables:

- Let X be a random variable, $X \sim N(\mu, \sigma_X^2)$, then the mgf for X is:
$$M_X(t) = e^{t\mu + \frac{1}{2}t^2{\sigma_X}^2}$$
- We can first define $Z = \frac{X-\mu}{\sigma_X}$ and $Z \sim N(0,1)$

$$M_X(t) = E(e^{tX}) = E(e^{t(\mu+\sigma_X Z)}) = E(e^{t\mu}e^{t\sigma_X Z}) = e^{t\mu}E(e^{t\sigma_X Z}) =$$
$$= e^{t\mu}M_X(t\sigma_X) = e^{t\mu}e^{\frac{1}{2}t^2{\sigma_X}^2} = e^{t\mu+\frac{1}{2}t^2{\sigma_X}^2}$$

QED

Source with modifications: https://www.quora.com/What-is-the-MGF-of-normal-distribution

The last piece of information that we miss are the following two properties:

- **Property 1: Moment of the Sum** Let $Y = \sum_{i=1}^{i=n} X_i$ where X_i are iid random variables then:

$$M_Y(t) = \prod_{i=1}^{i=n} M_{X_i}(t)$$

Proof:

$$M_Y(t) = E(e^{tY}) = E(e^{t\sum_{i=1}^{i=n} X_i}) = E(\prod_{i=1}^{i=n} e^{tX_i}) = \prod_{i=1}^{i=n} M_{X_i}(t)$$

QED

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/170/`

- **Property 2: Moment of the LC** Let $Y = a + bX$ where X is a random variable and $a, b \in \mathbb{R}, b \neq 0$ then:

$$M_Y(t) = e^{at} M_X(bt)$$

Proof:

$$M_Y(t) = E(e^{(a+bX)t}) = E(e^{at+bXt}) = E(e^{at}e^{bXt}) = e^{at}E(e^{bXt})$$

$$= e^{at}E(e^{btX}) = e^{at}M_X(bt)$$

QED

- Corollary: the sum of randomly iid Gaussian r.v. is still Gaussian.

Source with modifications: `https://onlinecourses.science.psu.edu/stat414/node/170/` and `https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf`

- Remember that we want to prove that:

$$\mathfrak{Dn} \xrightarrow{d} N(0, \sigma^2)$$

- This is like proving that:

$$\frac{\mathfrak{Dn}}{\sigma} \xrightarrow{d} N(0, 1)$$

- We can rewrite $\mathfrak{Dn}/\sigma$:

$$\frac{\mathfrak{Dn}}{\sigma} = \frac{\sqrt{n}}{\sigma}(\overline{\mathfrak{Xn}} - \mu) = \frac{\sqrt{n}}{\sigma}\left[\frac{\sum_{i=1}^{i=n} \mathfrak{Xn}_i}{n} - \mu\right] = \frac{\sqrt{n}}{\sigma}\frac{\sum_{i=1}^{i=n} \mathfrak{Xn}_i - n\mu}{n}$$

$$= \frac{\sum_{i=1}^{i=n} \mathfrak{Xn}_i - n\mu}{\sigma\sqrt{n}}$$

Source with modifications: https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf and https://en.wikipedia.org/wiki/Central_limit_theorem

- Note: We can assume that $\mu = 0$. If it is not, we could define a new set of variables $\mathfrak{Y}_\mathfrak{i} = \mathfrak{X}_\mathfrak{i} - \mu$ and we would have that:

$$\sum_{i=1}^{i=n} \mathfrak{X}\mathfrak{n}_i - n\mu = \sum_{i=1}^{i=n} \mathfrak{Y}_\mathfrak{i}$$

Preserving the same proof.

- Let now define $\mathfrak{W}_\mathfrak{n} = \mathfrak{D}\mathfrak{n}/\sigma$

$$\mathfrak{W}_\mathfrak{n} = \frac{\sum_{i=1}^{i=n} \mathfrak{X}\mathfrak{n}_i}{\sigma\sqrt{n}}$$

- We want to prove that $\mathfrak{W}_\mathfrak{n} \sim N(0,1)$ demonstrating that its moment is the same as the one of $N(0,1)$

Source with modifications: https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf and https://en.wikipedia.org/wiki/Central_limit_theorem

- Note: We recall Property 1 (Slide 105) and 2 (Slide 106) about the momentum of combining random variables and we have:

$$M_{\mathfrak{D}\mathfrak{n}}(t) = \left[M_{\mathfrak{X}_\mathfrak{i}}(\frac{t}{\sqrt{n}}) \right]^n$$

and likewise:

$$M_{\mathfrak{W}\mathfrak{n}}(t) = M_{\mathfrak{D}\mathfrak{n}}\left(\frac{t}{\sigma}\right) = \left[M_{\mathfrak{X}_\mathfrak{i}}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

- In essence we need to evaluate the limit for n going to infinite of $\left[M_{\mathfrak{X}_\mathfrak{i}}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$
- We want to prove that such limit is equal to:

$$M_{N(0,1)}(t) = e^{\frac{1}{2}t^2} \quad \text{(the momentum of } N(0,1) \text{)}$$

Source with modifications: `https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf` and `https://en.wikipedia.org/wiki/Central_limit_theorem`

- For simplicity we take the natural logarithm:

$$\ln\left[M_{\mathfrak{x}_i}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n = n\ln\left[M_{\mathfrak{x}_i}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]$$

- Now we define

$$q = \frac{1}{\sqrt{n}}$$

Therefore n is $1/p^2$ and $n \to \infty \Rightarrow p \to 0$. This means that we want to compute:

$$\lim_{p\to 0}\frac{\ln M_{\mathfrak{x}_i}(\frac{tp}{\sigma})}{p^2} =$$

- This is an indeterminate form, so we can take the derivative of both side by the theorem of de l'Hôpital

Source with modifications: https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf and https://en.wikipedia.org/wiki/Central_limit_theorem

CLT – LLf – Proof (5/7)

- This results to:

$$= \lim_{p \to 0} \frac{\frac{1}{M_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})} \frac{dM_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp} \frac{t}{\sigma}}{2p} = \frac{t}{2\sigma} \lim_{p \to 0} \frac{\frac{dM_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp}}{pM_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})} =$$

- This is again an indeterminate form, so we can take the derivative of both side by the theorem of de l'Hôpital

$$= \frac{t}{2\sigma} \lim_{p \to 0} \frac{\frac{d^2 M_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp^2} \frac{t}{\sigma}}{M_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma}) + p\frac{dM_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp} \frac{t}{\sigma}} = \frac{t^2}{2\sigma^2} \lim_{p \to 0} \frac{\frac{d^2 M_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp^2}}{M_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma}) + p\frac{dM_{\mathfrak{x}_\mathfrak{i}}(\frac{tp}{\sigma})}{dp} \frac{t}{\sigma}}$$

- We now take the limits at numerator and denominator and we are done.

Source with modifications: `https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf` and `https://en.wikipedia.org/wiki/Central_limit_theorem`

CLT – LLf – Proof (6/7)

- Numerator:

$$\lim_{p\to 0} \frac{d^2 M_{\mathfrak{X}_i}(\frac{tp}{\sigma})}{dp^2} = \left[\frac{d^2 M_{\mathfrak{X}_i}(\frac{tp}{\sigma})}{dp^2}\right](0) = E({\mathfrak{X}_i}^2) =$$

$$= E(\mathfrak{X}_i)^2 + Var(\mathfrak{X}_i) = 0 + \sigma^2 = \sigma^2$$

- Denominator:

$$\lim_{p\to 0}\left[M_{\mathfrak{X}_i}(\frac{tp}{\sigma}) + p\frac{dM_{\mathfrak{X}_i}(\frac{tp}{\sigma})}{dp}\frac{t}{\sigma}\right] = M_{\mathfrak{X}_i}(0) + 0\left[\frac{dM_{\mathfrak{X}_i}(\frac{tp}{\sigma})}{dp}\frac{t}{\sigma}\right](0) =$$

$$= M_{\mathfrak{X}_i}(0) = 1$$

Source with modifications: https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf and https://en.wikipedia.org/wiki/Central_limit_theorem

- And now we pull everything together and we obtain:

$$\lim_{p \to 0} \frac{\ln M_{\mathfrak{X}_\mathfrak{i}}\left(\frac{tp}{\sigma}\right)}{p^2} = \frac{t^2}{2\sigma^2}\frac{\sigma^2}{1} = \frac{t^2}{2}$$

- And, therefore

$$\lim_{n \to +\infty} M_{\mathfrak{W}\mathfrak{n}}(t) = e^{\frac{1}{2}t^2}$$

QED

Source with modifications: `https://www.stat.berkeley.edu/~mlugo/stat134-f11/clt-proof.pdf` and `https://en.wikipedia.org/wiki/Central_limit_theorem`

Status

- Now we know that the means of samples of a population tend to be distributed normally.
- This is an essential assumption to perform several numeric operations, like Montecarlo simulations, Bootstrap, etc.
- We can now understand the distribution of the Pearson momentum correlation coefficient of the sample
- Moreover, we have an open infinite issue on what to do if the data is NOT on a ratio scale
- This is an open issue for followup courses