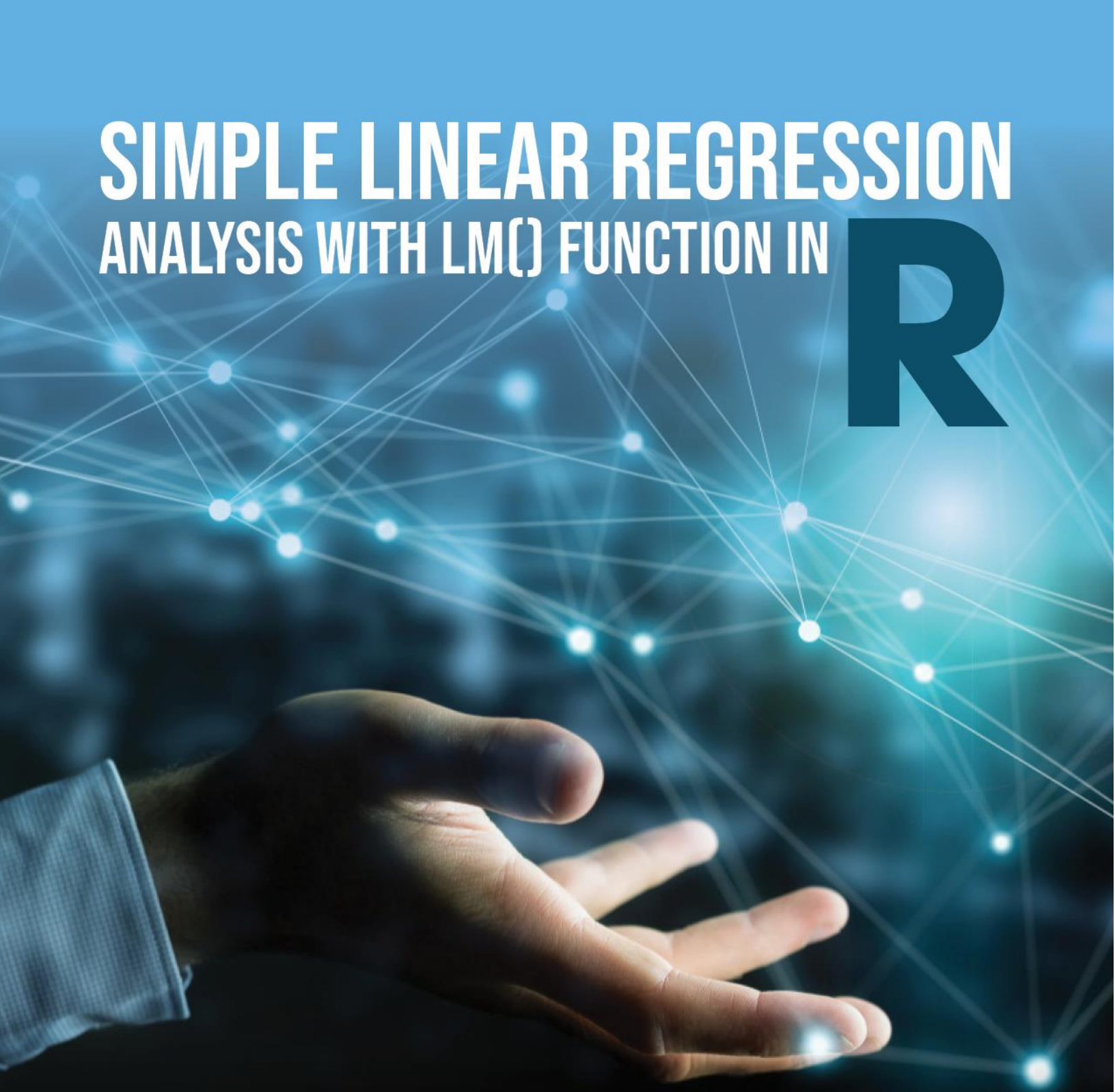
SIMPLE LINEAR REGRESSION ANALYSIS WITH LM() FUNCTION IN R



A ready-to-use handbook to apply simple linear regression analyses to time-series datasets, using lm() function.

Roberto Delgado Castro
San José, Costa Rica

519.53
D352s Delgado Castro, Roberto
Simple linear regression analysis with lm () function in R : how to apply regression analyses in time-series datasets [recurso electrónico] / Roberto Delgado Castro. – primera edición – San José, Costa Rica : D. Castro R., 2024.
E-book : pdf ; 800 Kb

ISBN 978-9968-03-939-0

1. ESTADÍSTICA – PREDICCIONES. 2. ANÁLISIS DE REGRESIÓN. 3. DATOS ESTADÍSTICOS – HISTORIA. 4. ESTADÍSTICA – MODELOS MATEMÁTICOS. 5. ESTADÍSTICA DESCRIPTIVA. I. Título.

Simple Linear Regression Analysis with lm() function in R

Roberto Delgado Castro

Foreword

Regardless of the methodology used, it is essential, in data science, to have the ability to analyze the temporal evolution of data series. The mechanism, by excellence, to carry out this type of analysis is through historical data series.

In this regard, Fernández-Morales and Bonilla-Carrión (2020, page 4) indicate the following:

From the historical behavior series, it allows us to model the basic components of the series, trend, cycle and seasonality, and thus be able to make predictions for the future, such as sales figures, consumption forecasts of a product or service, etc.

Therefore, the construction of time series allows not only to visualize the evolution over time of certain variables, but also allows estimations to be made based on this historical behavior. By visualizing historical trends, potential patterns can be observed and detected that explain different events that occurred in the past and that are reflected in disruptions.

Thus, one of the most widely used statistical techniques to analyze historical (time series) data series in order to define a mathematical model that allows estimations to be made, is the simple linear regression model.

The usage of regression models, in general terms, brings along the development of quantitative analysis of historic data series, in order to understand and explain, mathematically, its evolution in time, but to evaluate the possibility to perform estimations based on the equation of the best fir curve of such data set, or what is called like the regression equation.

What is the main purpose of this book? Put in your hands, respected reader, a ready-to-apply knowledge to put in practice simple linear regression models to time series data sets.

Acknowledgements

To God, who owns and governs everything. lucha

To mom… her example, sacrifice, guide and legacy will never be forgotten.

Contents

Introduction

Allen (2004), expresses that “statistical techniques are tools to enable us to answer questions about possible patters in empiral data. It is noty surprising then, to learn that many important techniques of statistical analysis were developed by scientists who were interested in answering very specific empirical questions. So it was with regression analysis. Simple regression and correlation form the basis of what is generally referred to today as regression analysis. In general, regression analysis is a statistical technique that attempts to predict the values of one variable using the values of one or more other variables. By convention, the variable that we are trying to predict is called the dependent variable, and the variables that we are using as predictors of the variable are called independent variables”.

As stated by Allen, regression analysis is a valuable tool to perform statistical analysis about specific datasets; what I call as historic data series or historic data sets. It is absolutely relevant, first, to be able to build such data sets. As soon as a data set has been built, regression analysis and many other statistical methods, could be performed.

Just to introduce its importance and application, linear regression models could be applied to quiet simple activities, such as studying how influences father´s height over his son´s height, estimate the price of a house depending on its area, predict interest rates of financial loans, aproximate the score obtained in a determined subject based on the quantiy of study hours and estimate processing time of a computer based on the speed or agility of its processor.

Based on such empirical data sets of historic data series, mathematical models could be performed to describe them. Precisely, Allen (2004) states that “it is important to bear in mind that regression analysis is nothing more than a mathematical model for describing and analyzing particular types of patterns in empirical data”.

Therefore, in the actual era in which large data sets are everywhere and the necessity to gather, order, classify and analyze them is rampant, regression analysis is a fantastic tool to fill such need, in order to develop mathematical models to generate an unvaluable added value to describe and analyze such datasets.

Simple regression models

Rodiguez-Jaume and Mora Catalá (2001), expresses that “linear regression analysis, in general, allows us to obtain a linear function of one or more independent or predictor variables (X_1, X_2, ... X_K) from which to explain or predict the value of a dependent variable or criterion (Y). In linear regression analysis we can differentiate between simple linear regression analysis and multiple linear regression analysis. In the first, an attempt is made to explain or predict the dependent variable (Y) from a single independent variable, (X_1); while, in the second, we have a set of independent variables, (X_1, X_2, ... X_K), to estimate the dependent variable (Y). In both cases, both the dependent variable and the independent variable(s) are measured on an interval or ratio scale”.

In addition, the aforementioned author established that "the simple linear regression analysis aims to predict and/or estimate the values of the dependent variable by obtaining the linear function of the independent variable”.

Therefore, Simple Linear Regression is a mathematical model that describes the relationship between variables. Linear regression models are statistical methods that help to predict the future. They are used in scientific and business fields.

Complementating the mentioned arguments, Talavera, Ocampo, Castellanos and Wacher-Rodarte (1995), mentioned that “simple linear regression involves the establishment of the relationship between two continuous quantitative variables, one called "X" and the other "Y"; the variable "X" within the graphical representation corresponds to the axis of the abscissae, while the variable "Y" to that of the ordinates; likewise, the variable "X" represents the independent variable, while the variable "Y" represents the dependent variable. The first step of the linear regression process is to determine the regression line, whose algebraic representation is shown in the following equation:

$$y = a + bx$$

or

$$y = mx + b$$

where,

y is the estimator given a certain value of x;

m is the slope of the curve.

x is the value on the axis of the abscissae (x axis).

b is the point of intersection of the curve with the axis of the ordinates (y axis).

Likewise, Laguna (2014) establishes the following:

"Regression is aimed at describing the relationship between two variables X and Y, in such a way that predictions can even be made about the values of the variable Y, based on those of X. When the association between both variables is strong, regression offers us a statistical model that can achieve predictive purposes. Regression assumes that there is a fixed variable, controlled by the researcher (it is the independent or predictor variable), and another that is not controlled (response or dependent variable). The correlation assumes that neither is fixed: the two variables are beyond the control of the researcher. Regression is its simplest form and is called simple linear regression. It is a statistical technique that analyzes the relationship between two quantitative variables, trying to verify if this relationship is linear. If we have two variables we speak of simple regression, if there are more than two variables, multiple regression. Its objective is to explain the behavior of a variable Y, which we will call an explained variable (or dependent or endogenous), from another variable X, which we will call an explanatory variable (or independent or exogenous). Once we have made the scatter plot and after observing a possible linear relationship between the two variables, we propose to find the equation of the line that best fits the point cloud. This line is called the regression line."

In a complementary way, Dagnino (2014, page 142), expresses:

"Linear regression makes it possible to predict the behavior of one variable (dependent or predicted) from another (independent or predictor). The procedure to be followed can be divided into four stages:

1) The first approach is through drawing the points on a Cartesian graph that shows the relationship between the two variables.

2) Then determine the equation of the line that best describes these points.

3) Next, the variability of the sample around the calculated regression line is calculated.

4) Inferences can finally be made."

However, so far the aforementioned authors have indicated that simple linear regression analysis consists of determining the degree of relationship between two variables and, from there, constructing an estimate of future values. In addition, the added value generated by this type of approximation consists of the development of the equation of the best fit line or regression line, which explains, mathematically, the relationship between both variables. Precisely, Laguna (2014) points out that "linear regression consists of finding (approximating) the values of one variable from those of another, using a linear functional relationship, that is, we look for quantities a (ordered at the origin) and b (slope of the linear line) such that we can write $Y = a + bx$, with the least possible error between $\hat{Y}$ and Y".

Typically, to determine the values of the regression line, the Least Squares method is used, which, according to Laguna (2014) "consists of looking for the values of parameters a and b so that the sum of the squares of the residuals is minimal. This line is the regression line by least squares."

Once the equation of the best-fit line or regression line is developed, it is possible to estimate future values. This estimate is executed by substituting estimated X values in the equation; the result, for each estimated X-value, is a consequent estimated Y-value.

Another very important variable in the simple linear regression analysis is the Coefficient of Determination R^2. Laguna (2014) expresses the following:

"The most important measure of goodness of fit is the coefficient of determination R^2. This coefficient indicates the degree of adjustment of the regression line to the values of the sample, and is defined as the percentage of the total variability of the dependent variable Y that is explained by the regression line. The less dispersed the residuals are (remember that residuals or errors are the difference between the observed values and the values estimated by the regression line), the better the goodness of the fit."

Simple regression analysis: the best fit curve

A linear line of best fit can be defined as a straight line provinding the best aproximation of a defined dataset. Also, it could be defined as the line that better describes the relation between an dependant and an independent variable. The result consists in a line that better adjusts all the points in a scatter plot.

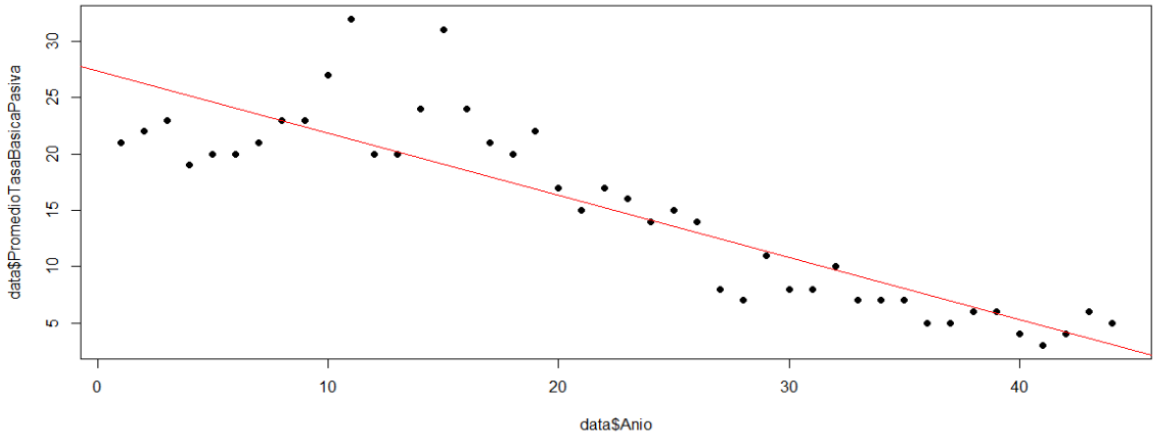
The distances between the points and the regression line are called residuals. They represent the portion of the response that is not explained by the regression equation; that is, the difference between the observed value and the approximate value is the residual.

In any regression analysis it will be observed that some points are closer to the line and others much farther from it. The closer the points are to the line, the better the fit between the regression line and the data. Residuals allow you to check the equation in order to check how well the line fits the data.

Rudziewicz, Bossé, Marland and Rhoads (2017), stated that "through a scatter plot of data, a line of best fit can be used to represent that data. This is often accomplished by considering residuals. Also, they added that "many decisions must be made by participants as they attempt to visualise a line of best fit through data. They must decide how appropriate a linear approximation may be to represent the trend of the data, and must also decide whether outliers should be considered in the development of the line of fit or should be omitted from consideration".

In other words, the best fit curve, or the line of best fit, is a straight line that represents the trend of a scattered data plot on a graph. The following image shows an example of the best fit curve included into a scatter plot:

Image 1
Example. Best fit curve in a scatter plot



Importance of regression analysis

Sykes (1993) mentioned that "regression techniques have long been central to the weld of economic statistics ("econometrics"). Increasingly, they have become important to lawyers and legal policy makers as well".

In addtion, Marill (2004) states that "linear regression is a popular technique because many phenomena of interest have a linear relationship, and the technique is able to demonstrate mathematically and visually the relationships between clinically important variables".

Simple Linear Regression models are very popular among a wide variety of research fields, due to their agility and ease of interpretation. As a consequence of their capacity to transform data, they can be used to analyze a variety of relationships between variables. Linear Regression not only are used to predict scenarios; also has demonstrated its efectiveness describing systems.

Gogtay, Deshpande, and Thatte (2017) underlines that the utility of regression analysis lies in the fact that such models are able to explain the evolution of data sets (dependent and independent variables) due to the construction of scatter plots. Also, such analysis is capable to generate the regression line and the regression equation. With such equation, a prediction could be performed to determine estimated values based on such mathematical expression. Such authors also expresses that "there are three major

uses of regression analysis – attributing causality [cause and effect relationship], forecasting and prediction".

Berk (2004) points that "regression is used for description. It is used to describe the distribution of a variable under a number of different conditions". He also expresses that "regression is used for prediction". Therefore, he mentioned that such analysis have two pillars: description and prediction.

Indeed, based on my personal and professional experience, two of the most popular tasks that I have done within my career in data science is, precisely, describe and predict variables.

One popular usage and advantage of Simple Linear Regression Models is their application in market research. First, is a powerful statistical method that could help researchers to answer questions about market behavior. Second, it is a flexible method that could be applied to a wide variety of problems that usually appear in different markets. Third, this method is easy to undersand and utilized.

Thus, Simple Linear Regression Models could help researchers to analyze the effect of the changes in one determined variable over other independent variable. This is particularly useful to study the relationship between variables and market behavior, such as prices, income, expenses, etc. Finally, this statistical method could help researchers to predict future market behavior based on historical data sets.

It is frequent that the mentioned method could be used in some business fields, as follows:

1. Sales prediction.
2. Inventory management.
3. Prices analysis.
4. Cost analysys.
5. Cash flow analysis.
6. Employees performance analysis.
7. Credit risk avaluation.

Almost all data scientists and data professionals, in the majority of cases, regardless the nature and characteristics of their work, have to gather, combine, clean and organize data sets, describe them using some statistical method or mathematical model and, finally, based on that, perfom predictions. Thus, the work of describe and predict, as Berk mentioned, is very common within the data science ecosystem nowadays.

Therefore, the execution of regression models is, as well, very common actually. In my personal case, due to the fact that I have specialized in studying historic dataseries and in simple regression models, this kind of mathematical model is very useful to, precisely, describe and predict.

Now, it what fields should be applied the regression models? My answer must focus on the fact that it could be applied in almost every field of human activity: social science, medical industry, nutrition, sports, retail, biology, genetics, physics, supply chain, construction, chemical, purchases, imports, exports, energy, transportation, aeronautical, government, etc. As soon as a data set could be built, and a necessity to describe it and analyze through a mathematical model appears, a regression model would take part of stage for sure.

Even if you think in the most simple and trivial activity, like cutting the lawn in your house, a regression model could be developed. For instance, you hve collected data referred to the time spent cutting it within the last 25 chances in the past. Thus, you have a dataset composed of 25 observations (quantity of minutes). With it, you can use a simple regression model to explain and describe, mathematically, such data by developing the regression equation (the equation of the best fit curve of the dataset). With such analysis, you would have a general landscape in which you would know how you have done such a work. Therefore, based on that, you can predict the aproximate time you would spend cutting your lawn for the next five or then chances in the near future, in order to prepare your daily schedule in a more efficient way.

As seen, such analysis is very important in business. Just imagine the retail industry, exemplified by a supermarket: the king of data. A supermarket is a data factory. Based on customers purchases daily, they generate tons of data every minute. If the manager of a local store wants to know the record of how many cereal boxes of a specific brand have been sold within the last 12 months, in order to decide how many units he would have to purchase from the supplier in the near future, within lots of models, he would have to perform a regression analysis to find out what to do. Doing so, he might be able to describe the evolution of cereal´s purchases within the last year and, additionally, he would be able to predict how many units to buy for suppliers to fill the inventory. Thus, just imagine a worldwide giant composed of thousands of supermarkets in dozens of countries…it seems that there is a lot of work to do out there!

As well, Murray and Wilson (2021) points out that these kind of models are used to explore data and draw preliminary conclusions. In business, regardless of the industry, exploring data is a crucial task in

the decision-making process. Managers are always asking for solutions, specially, those based on data. Thus, regression models are wonderful tools to accomplish that need.

Now, one of the most common "headaches" of managers, not only in the retail industry of course, is the need to perform high quality forecasts of certain variables: sales, purchases, taxes, payroll, customers, expenses, income, etc. They need to predict what the organization would need or execute in the future in the most accurate way. Indeed, Wang and Jain (2003), points that "in our econocmic lives, our expectation of the future is more important than the events of the past. However, we are still interested in it because what has happened in the past would have a bearing on the future. Statistical data are a series of snap shots of the past on which we perform our analyses in order to draw inferences about the future. While we cannot change the past, with the lessons learned form it, we can change the future. Forecasting helps us to accomplish this task". In addition, they added that regression models are a method used to perform such forecasts. Precisely, they classified it as a cause-effect statistical procedure.

In addition, and according the last arguments, such authors added that "in recent years more and more companies are using regression models in their forecasting efforts. There are a number of factors, which have contributed to this growth. The most important one is the development of technologies in computing, accessing, processing and storing data".

Mahbobi and Tiemann (2015) unifies nicely the whole utility of regression models in business. They expresses that "regression analysis is one of the most used and most powerful multivariate statistical techniques for it infers the existence and form of a functional relationship in a population. Once you learn how to use regression, you will be able to estimate the parameters — the slope and intercept — of the function that links two or more variables. With that estimated function, you will be able to infer or forecast things like unit costs, interest rates, or sales over a wide range of conditions. Though the simplest regression techniques seem limited in their applications, statisticians have developed a number of variations on regression that greatly expand the usefulness of the technique. Being able to estimate the effect that one independent variable has on the value of the dependent variable in isolation from changes in other independent variables can be a powerful aid in decision-making and policy design. Being able to test the existence of individual effects of a number of independent variables helps decision-makers, researchers, and policy-makers identify what variables are most important. Regression is a very powerful statistical tool in many ways".

such authors pointed a very intertesting issue: “the mathematics of regression are not so simple, however. Instead of trying to learn the math, most researchers use computers to find regression equations”. In fact, R, as a statistical programming language widely used in data science, is a fantastic tool to perform such tasks. Hence, in the following chapters I will detail and discuss how to use R to execute these kind of mathematical procedure.

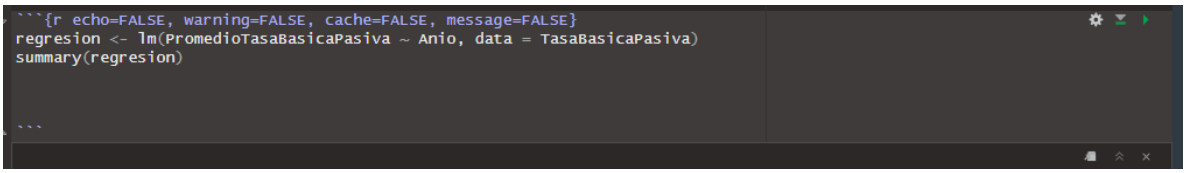
Lm() function in R: structure and results

According to the official site *rdocumentation.org*, **lm()** function (linear methods) "is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance". It´s most important arguments are as follows:

1. **Formula**: an object of class "formula": a symbolic description of the model to be fitted.
2. **Data**: an optional data frame, list or environment containing the variables in the model.
3. **Summary**: it prints a summary of the results obtained.

The following image shows an example of a code using the lm() function:

Image 2
Example. Lm() function



The image shows a typical code to execute the lm() function. Observations:

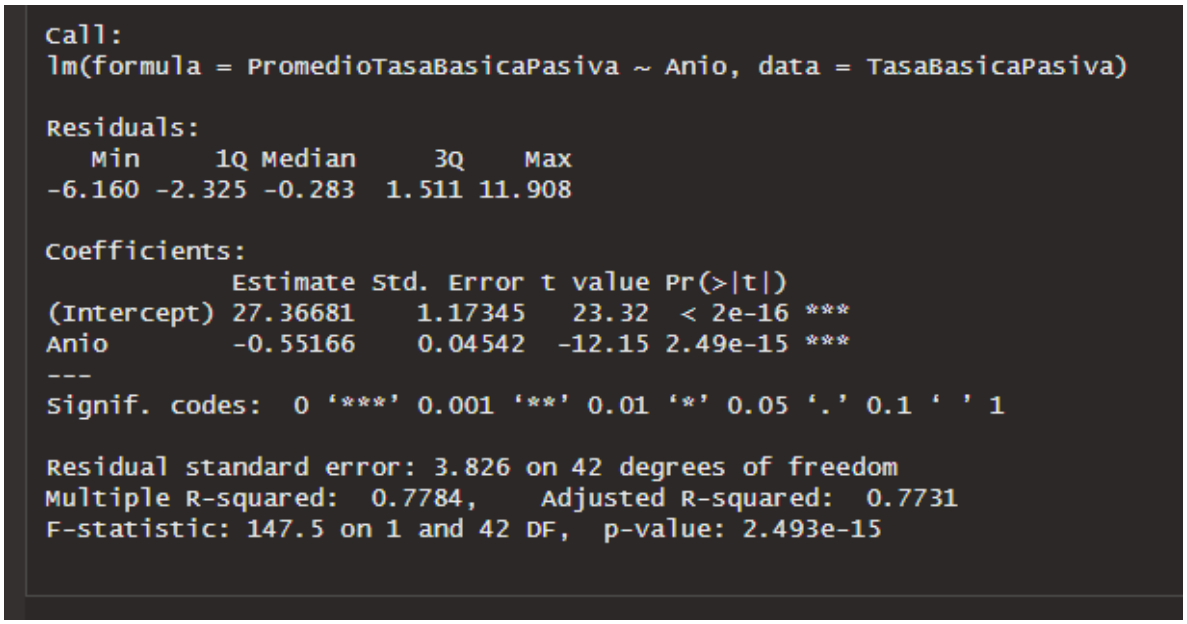
1. A new object must be defined in order to execute the function; in this case, the new object is "regresion".
2. Inside the parenthesis of the lm() function, it must be defined which of the two variables is the independent and which is the dependent; it brings along the need of defining which of them corresponds to the X axis, and which of them corresponds to the Y axis. In this specific example, both variables must be included inside the function.
3. Inside the parenthesis of the lm() function, it must be defined clearly the dataframe or database from which the original data will be taken to perform the calculations. In this case, the database-source (the dataframe that contains the data) must be defined and written after the "data" argument.
4. After defining the variables and the database-source, in order to let R print the general results of the components of the simple regression analysis, the name of the new variable definded in numeral 1 must be written inside the "summary" argument.

The argument "summary" contains the following results:

1. **Intercept**: corresponds to the estimation of the independent variable when all other variables are equal to zero, i.e., the value of y when $x = 0$ (ordered to the origin). In addition, it indicates the point of intersection of the best-fitting line with the Y-axis (ordinated).
2. **Y value (Anio)**: corresponds to the slope of the Best Fit Curve. In this case the value is negative, so the trend of the curve is towards decrease. In addition, this value corresponds to the variations in the value of the Y-axis (ordered) given changes in the values on the X-axis. The larger the slope value, the steeper of the curve will be. On the contrary, the lower the slope value, the lower the inclination of the curve will be.
3. **Multiple Correlation Coefficient**: it is a measure of how close together or apart the variables under study move or are shown (years and annual average of the Passive Base Rate). This coefficient shifts between -1 and +1.
4. **Adjusted Coefficient of Determination R^2**: this measure indicates how the developed model is adjusted taking into account the number of regression variables.
5. **Standard error or standard error**: this is a value that quantifies how far the values deviate from the mean of the data population under study.
6. **F-statistic**: it shows if the model is statistically significant or not. The F-test checks if at least one variable's weight is significantly different than zero.
7. **Residuals**: this section summarizes the residuals, the error between the prediction of the model and the actual results.
8. **Residual Standard Error**: This is the standard deviation of the residuals.

The following image shows an example of the results of the lm() function:

Image 3
Example. Lm() function results



```
Call:
lm(formula = PromedioTasaBasicaPasiva ~ Anio, data = TasaBasicaPasiva)

Residuals:
   Min     1Q Median     3Q    Max
-6.160 -2.325 -0.283  1.511 11.908

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.36681    1.17345   23.32  < 2e-16 ***
Anio        -0.55166    0.04542  -12.15 2.49e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.826 on 42 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7731
F-statistic: 147.5 on 1 and 42 DF,  p-value: 2.493e-15
```

Lm() function in R: scatter plot

Besides the main results of the lm() function, building a visualization in the form of a scatter plot is always important. Due to the usage of the argument "plot()", such scatter plot could be developed. The following images show and example of a code to develop a scatter plot in lm() function:

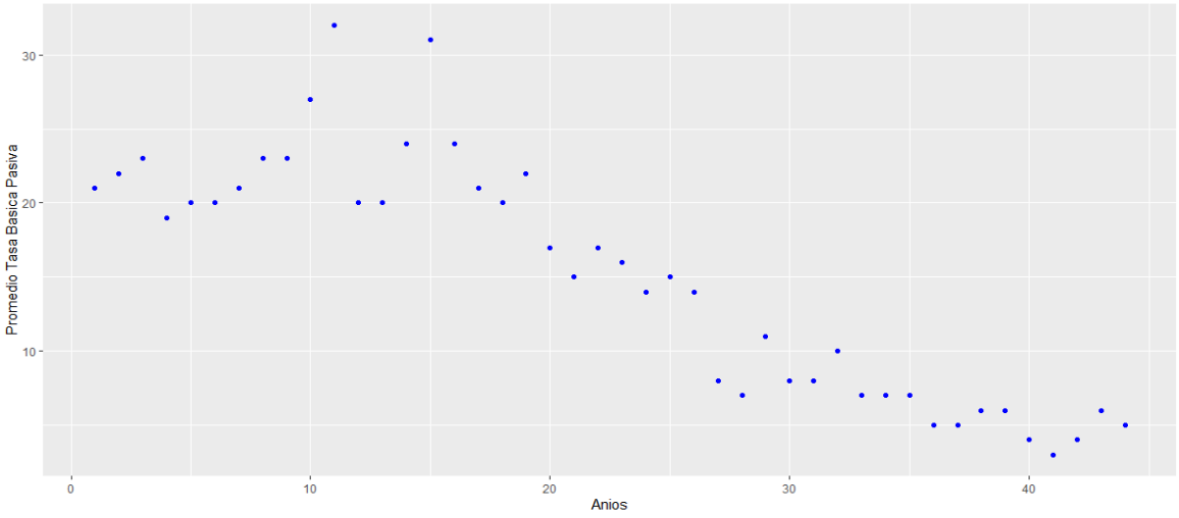
Image 4
Example. Lm() function scatter plot

```
#Regresion lineal con paquete LM

```{r echo=FALSE, warning=FALSE, cache=FALSE, message=FALSE}
data <- TasaBasicaPasiva
plot(data$Anio, data$PromedioTasaBasicaPasiva, pch = 16)
reg_model <- lm(PromedioTasaBasicaPasiva ~ Anio, data = data)
abline(reg_model, col = "red")
```

The following image shows the scatter plot as the result of the execution of the correspondent code:

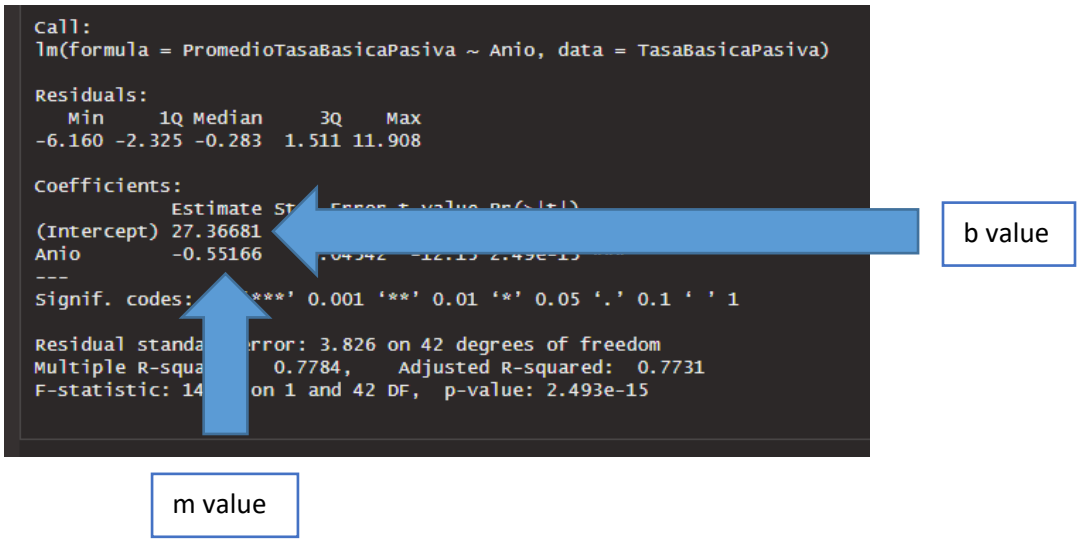
Image 5
Example. Lm() function scatter plot


```

Lm() function in R: simple regression equation

Once the lm() function is executed, as seen before, the "summary" argument displays an inventory of results of the regression model. In general basis, it can be established that the slope of a regression equation indicates the effect of the predictor variable on the response variable.

Under the main conception that the simple regression equation is composed by the structure of $y = mx + b$, the simple regression equation could be defined extracting the following arguments from the results of the "summary" argument of the lm() function:



- The "intercept" value corresponds the point of intersection of the best-fitting line with the *Y*-axis (ordinated).
- The "Anio" value corresponds to the slope of the curve, or the *m* value of the equation.

Predict() function: estimations from simple regression models

According to the official site **rdocumentation.org**, **predict()** function "is a generic function for predictions from the results of various model fitting functions". It´s most important arguments are as follows:

1. **Object**: a model object for which prediction is desired data: an optional data frame, list or environment containing the variables in the model.
2. **Newdata**: the predicted desired values.

The following image shows an example of a code using the predict() function:

Image 6
Example. Predict() function

```
```{r echo=FALSE, warning=FALSE, cache=FALSE, message=FALSE}
regresion <- lm(PromedioTasaBasicaPasiva ~ Anio, data = TasaBasicaPasiva)
NuevasTasas <- data.frame(Anio = c(45, 46, 47, 48, 49, 50, 51, 52, 53, 54))
predict(regresion, newdata = NuevasTasas, interval = "prediction")


```
```

In this specific example, the database-source was composed of 44 obervations; the objective consists in estimate values for a 10-year period of time in the future.

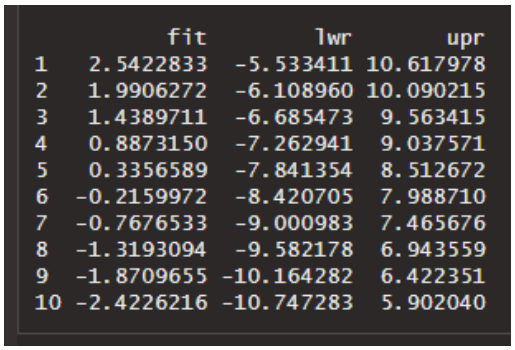
Under the object "regresion", the syntax of the lm() function must be included.

Under the object "NuevasTasas", inside the object of "data.frame", in the form of a vector, the ten estimated values must be included, in order let R calculate the projected values for the ten variables written.

Finally, inside the predict() function, the objects defined before of "regresion" and "NuevasTasas" must be included.

The following image shows the results of the execution of the predict() function:

Image 7
Example. Predict() function results



	fit	lwr	upr
1	2.5422833	-5.533411	10.617978
2	1.9906272	-6.108960	10.090215
3	1.4389711	-6.685473	9.563415
4	0.8873150	-7.262941	9.037571
5	0.3356589	-7.841354	8.512672
6	-0.2159972	-8.420705	7.988710
7	-0.7676533	-9.000983	7.465676
8	-1.3193094	-9.582178	6.943559
9	-1.8709655	-10.164282	6.422351
10	-2.4226216	-10.747283	5.902040

How to develop simple regression analysis: where to begin

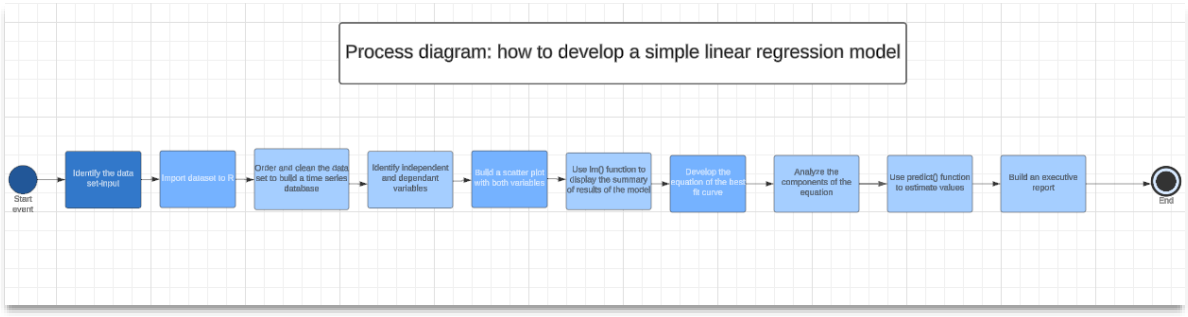
The most important ingredient of Regression Models is data. Hence, before performing any Simple Linear Regression Model, the best suggestion that could be applied is, obviously, obtain data.

The task of getting data depends on the researcher and its environment. Sometimes, due to its difficulty, such work could be considered as an art. Why? People often defines brilliant strategies to extract data from certain sources, in order to build logical and reasonable datasets. Such strategies are flexible and depend on researcher´s needs and particular features of their work or projects.

Therefore, getting data before performing a linear model is important not only to obtain the sufficient inputs to work with, but to identify and define, clearly, the variables that will be used in the model. This task is crucial: quality, efectiveness and reliability of a regression model depends, in a high percentage, in the definition of the variables. This fact is particularly relevant in simple linear regression models, in which independent and dependant variables are defined in order to find potential relations and to develop the equation of the Best Fit Curve, with the purpose to perform predictions based on specific input-datasets.

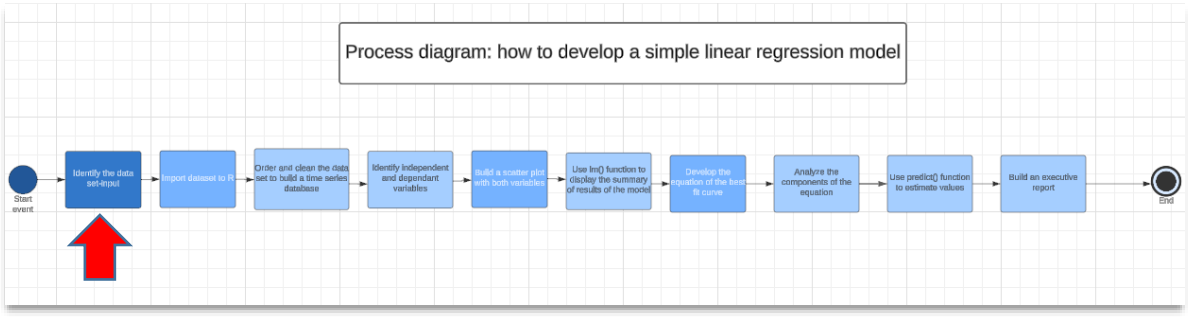
How to develop simple regression model: check list

Based on my own experience working with time series data and historic datasets, the following steps should be accomplished in order to develop a simple linear regression model in R:



1. Identify the dataset-input.
2. Execute proper tasks to import the dataset to R.
3. Order and clean the dataset to build a time series dataset.
4. Analyze the data: identify independent and dependant variables.
5. Build a scatter plot with both variables.
6. Use lm() function to display the summary results of the model.
7. Develop the equation of the best fit curve.
8. Analyze the components of the equation depending on the scatter plot.
9. Use predict() function to calculate estimated values in a future composed of a certain quantity of periods (years, months, weeks, days, etc).
10. Build an execuive report with the most relevant results.

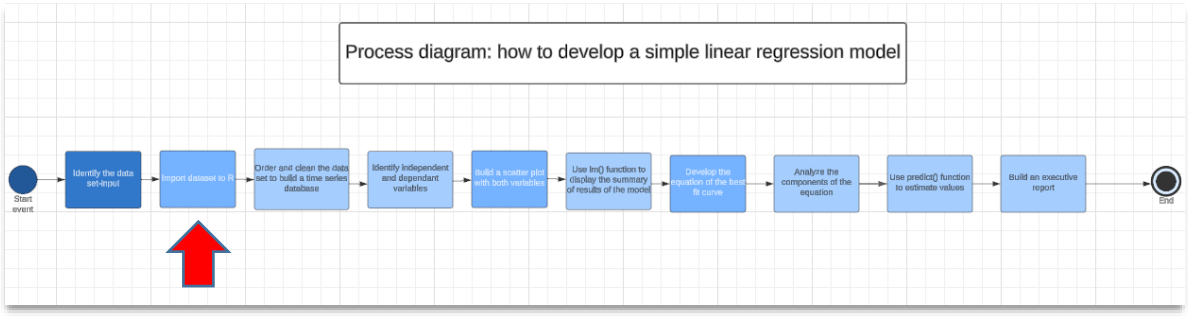
Identify the dataset-input



It is crucial to identify clearly the dataset or database that will contain all the information needed to include into the model.

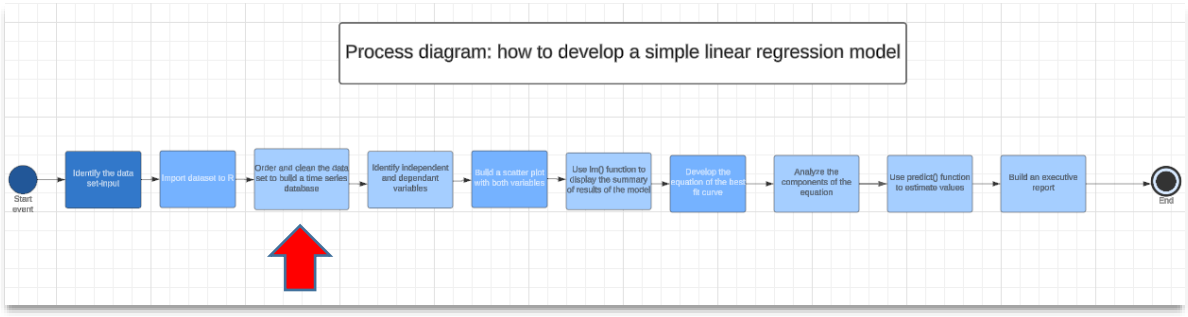
This task is alligned with the direct purpose or principal mission of the research. It is very important that you, as researcher, define what do you want to find, in order to identity the data you might need to fill such need.

Execute proper tasks to import the dataset to R



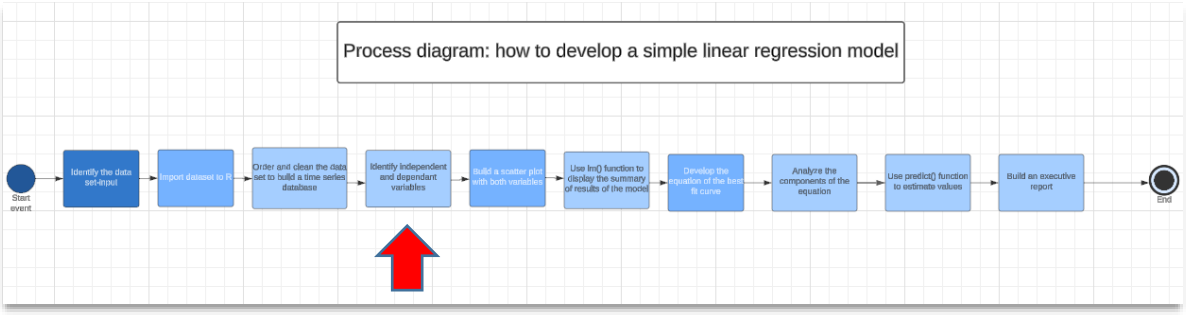
Once the dataset has been identified and built, it must be imported to R in order to begin with the project. The import process depends on the type and format of the dataset (file). It is crucial to check out that the import process completes totally and the total quantity of observations and variables are imported properly.

Order and clean the dataset to build a time series dataset



After the import process, it is relevant that the researcher cleans and orders the dataset based on its purposes and the nature of the research. If it is necessary, all unacceptable registers with incorrect formats must be removed, in order to guarantee the consistency of the dataset and the reliability of the results.

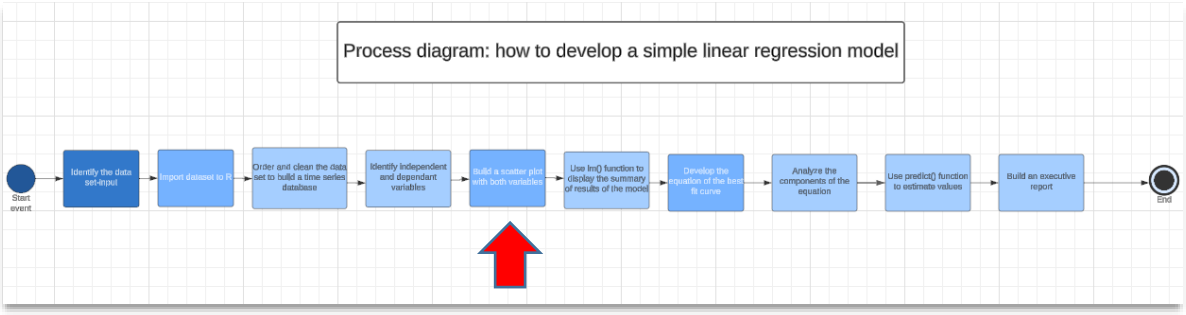
Analyze the data: identify independent and dependant variables



After the order and clean process, once the dataset is consistent, researcher must identity the independent and dependant variable. In time series data, usually the independent variable is composed of time units, like years, months, weeks or days; the dependant variables are composed of correspondent values related with such time units.

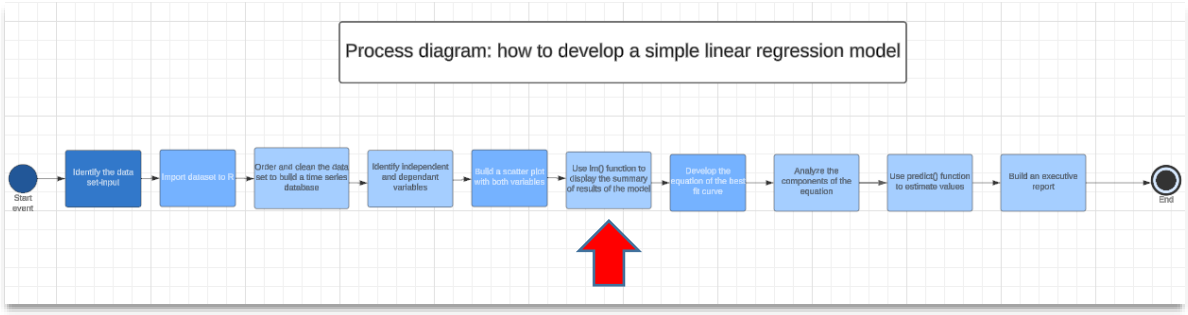
In these kind of processes (simple linear regression models), it is a good idea to build a database composed of two columns: one for the independent values and another one for the dependent values.

Build a scatter plot with both variables



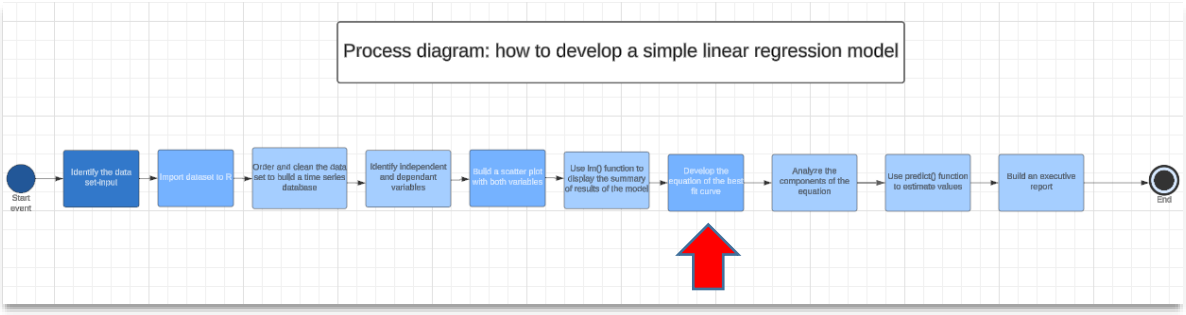
A scatter plot could be build using the database built in the former stage. It is importat to develop such visualization in order to take a look to the data displayed in a graph to visualize its variability.

Use lm() function to display the summary results of the model



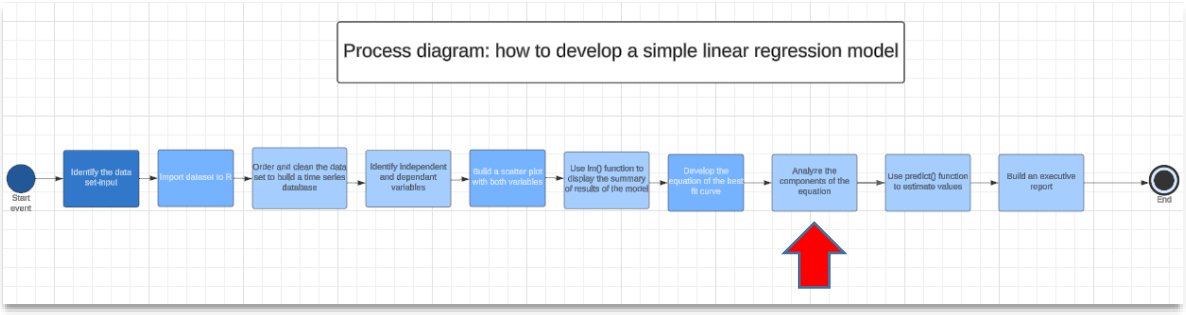
Using the correspondent code, the lm() function must be executed in R in order to develop the model properly.

Develop the equation of the best fit curve



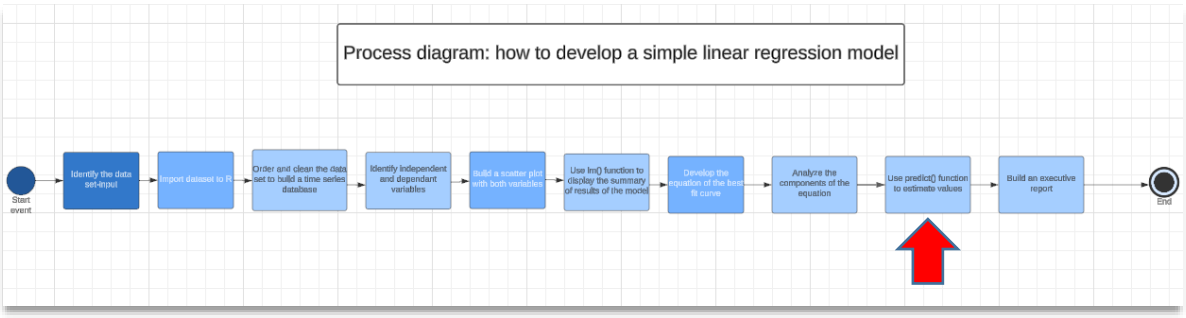
Based on the results displayed by the "summary" object of the lm() function executed, researcher must build the equation of the best fit curve.

Analyze the components of the equation depending on the scatter plot.



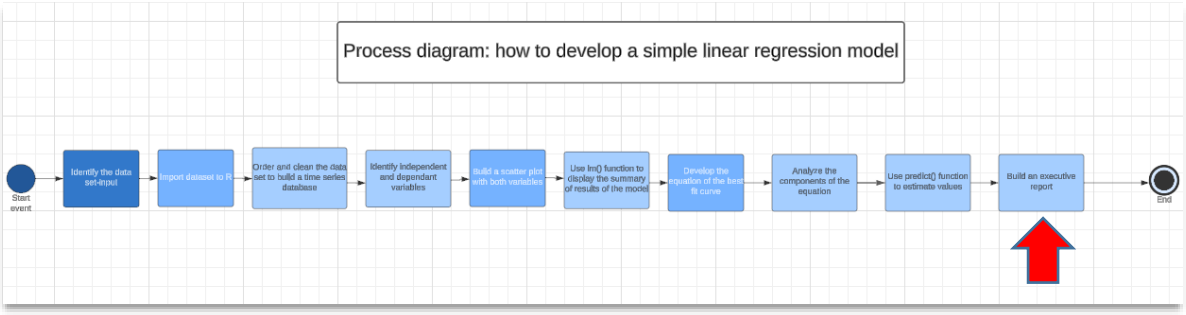
Based on the results displayed by the "summary" object of the lm() function executed, and once the researcher has built the equation of the best fit curve, it is recommended that he/she analyzes its components based on the tendency of the data displayed in the scatter plot: slope and point of intersection with the Y axis.

Use predict() function to calculate estimated values in a future composed of a certain quantity of periods



Once the researcher has built the equation of the best fit curve, he/she, according with the main purpose and nature of the research, should perfom estimations using the proper code of the predict() function. As mentioned, the quantity of estimated periods (years, months, weeks or days) will depend on researcher´s needs.

Build an execuive report with the most relevant results



The most important stage of the process consists in the proper documentation of the results obtained. It is crucial to remember that what has not been documented correctly, does not exist. Thus, researcher must make sure to document all results and relevant information in a executive and consolidated report, in order to share it with strategic people, depending of his/her professional environment: board members, managers, supervisors, consultants, colleagues, students, etc.

Closing

If you, my dear friend reader, has reached this chapter of the book, it is clear that you has shown interest in the application of simple linear regression analyses to time-series datasets.

As me, I am pretty sure that you are a data geek, and a super-fan of time series databases. Maybe that fact has pushed you to places and realities that you might never thought some time before.

Therefore, from the deepest part of my heart, I hope you have learned the most important components of these kind of models and how to solve problems regarding the application of this topic to time-series datasets.

The most important mission of this book consists in upgrading your knowledge in data science and helping you move forward some steps up in your desire to be a better professional. I really hope I have achieved such objective.

You can reach me and find complementary information in my official website:

https://roberto-delgado.com/

References

Allen, M. P. (2004). Understanding regression analysis. Springer Science & Business Media.

Berk, R. A. (2004). Regression analysis: A constructive critique (Vol. 11). Sage.

Dagnino, J. (2014). Regresión lineal. Rev. Chil. Anest, 43(2).

Fernandez-Morales, M., & Bonilla-Carrión, R. (2020). Bibliominería, datos y el proceso de toma de decisiones. Revista Interamericana de Bibliotecología, 43(2).

Gogtay, N. J., Deshpande, S. P., & Thatte, U. M. (2017). Principles of regression analysis. J. Assoc. Physicians India, 65(48), 48-52.

Laguna, C. (2014). Correlación y regresión lineal. Instituto Aragonés de Ciencias de la Salud, 4, 1-18.

Mahbobi, M., & Tiemann, T. K. (2015). Regression basics. Introductory Business Statistics with Interactive Spreadsheets-1st Canadian Edition.

Marill, K. A. (2004). Advanced statistics: linear regression, part I: simple linear regression. Academic emergency medicine, 11(1), 87-93.

Murray, L. L., & Wilson, J. G. (2021). Generating data sets for teaching the importance of regression analysis. Decision Sciences Journal of Innovative Education, 19(2), 157-166.

Rodríguez-Jaume, M. J., & Mora Catalá, R. (2001). Análisis de regresión simple.

Rudziewicz, M., Bossé, M. J., Marland, E. S., & Rhoads, G. S. (2017). Visualisation of lines of best fit. Int. J. Math. Teach. Learn, 18, 359-382.

Sykes, A. O. (1993). An introduction to regression analysis.

Talavera Piña, J. O., Ocampo, A. A., Castellanos Olivares, A., & Wacher Rodarte, N. H. (1995). Regresión lineal simple.

Wang, G. C., & Jain, C. L. (2003). Regression analysis: modeling & forecasting. Institute of Business Forec.

Simple Linear Regression Analysis with lm() function in R

Roberto Delgado Castro