Effort estimation with statistical and ML techniques

Project Management

Sergio de la Mata Moratilla

3º GII E

Index:

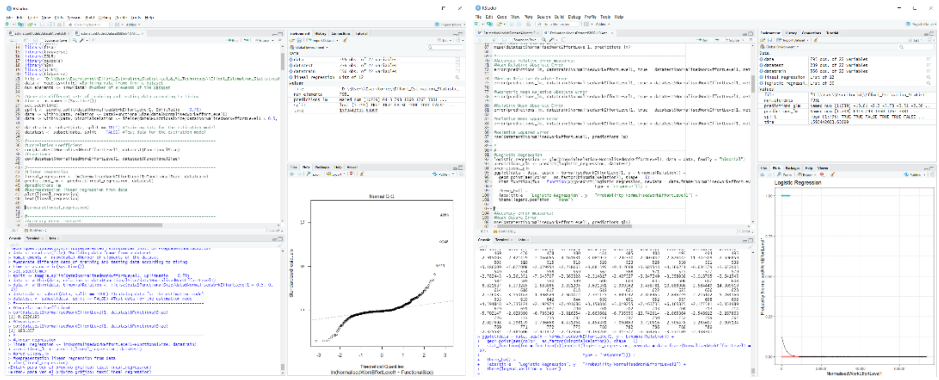
Introduction:

This document contains all the information regarding the performance of the practice. It will explain what it was made, the estimation models used, the evaluation measures used on them and some techniques used for the performance of the practice.

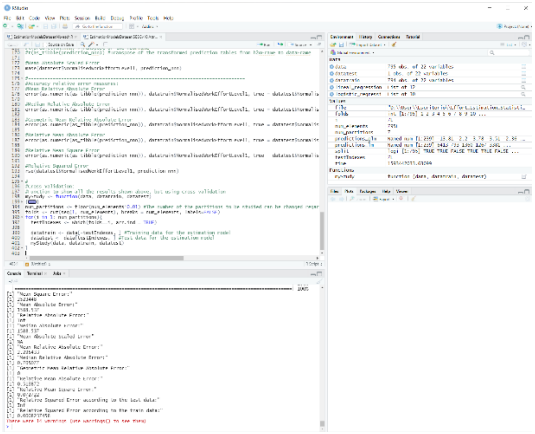
Summary Practice:

This practice asked to build different statistical or machine learning models in order to make estimations from a database of historical projects.

To do it, it was used two datasets given (one from the Github repository given and an another provided directly by teacher) where it was used linear regression, logistic regression and neural network models and multiple evaluation measures in order to have different estimation from the datasets.



It has been also used another dataset in order to represent the regression tree model and the technique of cross validation to have a different point of view of estimating dividing data provided in different subsets.

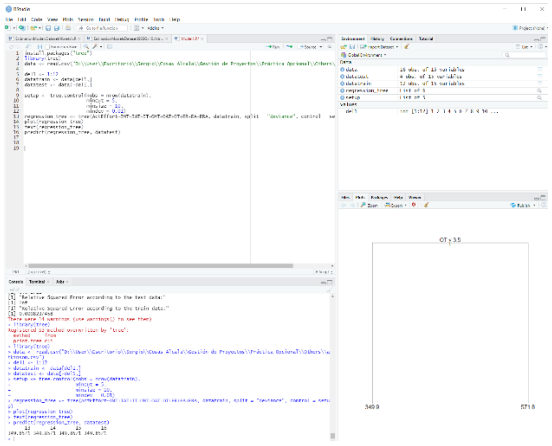


Estimation Models:

In order to make estimations and/or predictions from a set of historical data from a project, there are many estimation models which can be used considering the kind of data which is been used.

For the study of the datasets selected, were used the lineal regression, logistic regression and neural network models. Even though the main study makes use of

these estimation models, it was also seen with a different dataset the regression tree, but due to the restriction it had, it was decided not to be used with the rest of models.



A regression tree is a type of decision in which the leaves represent a numerical value and their idea is to predict the answer variable Y in function of covariables. The problem from this model is its disadvantages. These have made to not be used with the rest models as it could only be made in one of the data sets will in the order had many disadvantages its use. These were some of the disadvantages found:

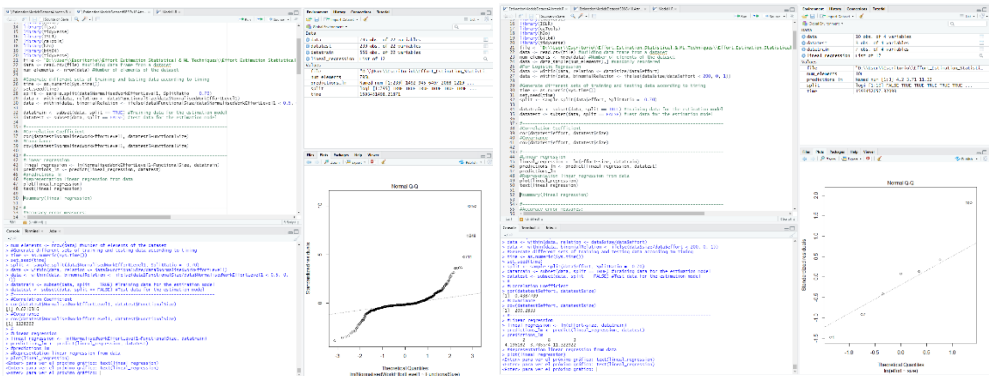
- If it is used the package “tree” as in the picture above, if the dataset provided is small, for example using 10 elements, it might not be able to be made in RStudio as it can distribute the element in a more accurate way.
- The regression tree may have also a lower number of nodes, if the series of terms provided for the prediction is low, being the lowest case only one term.
- When it is obtained a prediction for a series of terms given at first, the results obtained will only have the values which are at the nodes making very restrictive the variety of possibilities obtained and, in the cases in which there is a great variety of data, the error estimations obtained, can be high.

One of the datasets have only 10 elements with two fields each while the other have 795 elements and a huge number of fields. Therefore, the dataset with the lower number of elements is more restricted and it is not useful to be used it in this case.

The linear regression is an estimation model in which it is followed a linear approach between a dependent variable and an independent or a set of independent variables.

The problem with this estimation comes when the structure of the dataset according to its fields used for the estimation builds a different structure or the greater part of the elements are away from the estimation value which would correspond to that element in the estimation function straight line.

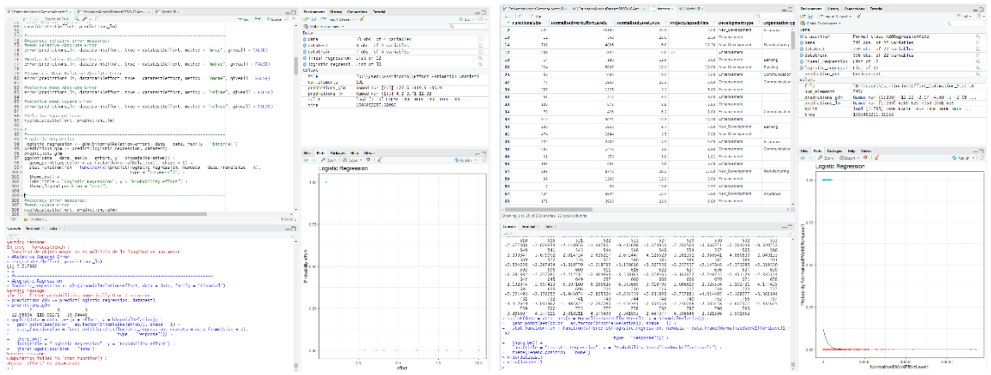
As it can be seen in the next pictures, the estimation obtained for the different values of the elements is really accurate to the different elements of the dataset.



The logistic regression is a statistical model which is used to predict the result of a variable in function from a set of independent variables. This estimation model is used when the dependent variable is binary in nature.

The main problem from this estimation is the binary nature from the dependent variable making impossible to consider intermediate variable unless it is obtained previously a mechanism to classify them using this criterion.

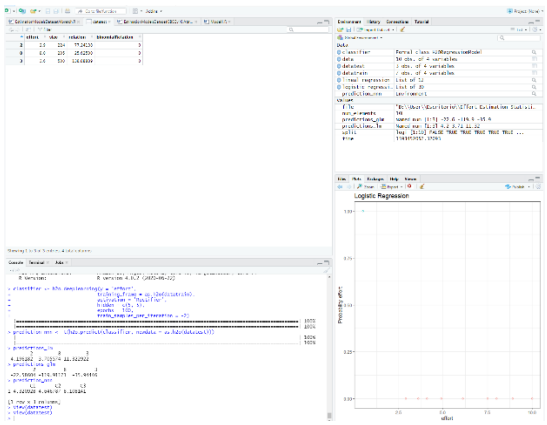
In the case of the datasets selected, it couldn't be done directly with the data given as none of the fields from their different elements had only the values of 0 and 1. To make the study, it was build a small function at the beginning from each of the R files in order to distribute according to a condition the different elements according to the value they had from the relation between the effort and the size of the project. The pictures below show the results from the predictions obtained.



Neural networks are computational model based on the biological neural connections in which a set of units are connected between them to transmit signals. The input information goes through the network obtaining different outputs passed to the next units until it is obtained a set of output data. Some of the disadvantages from this model are its hardware dependence, not quiet explained behaviour from the network, unknowledge from the proper structure of the network, difficulty of representation, … For this last reason, it has not been represented for the different datasets.

These are the results of the predictions obtained for each of the datasets according to the different test data provided for each of the datasets in each moment using the different estimation models and the values which in really would have. It is needed to

consider that the field value predicted is the one related with effort for all the models and both datasets.



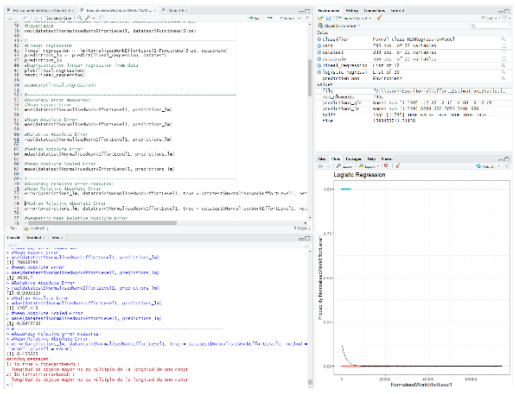
```
> predictions_lm
          2          13          14          17          18          22          25          27          30          32
6288.6391   620.3211  7053.1162  2400.9934   638.9669  3529.0633  1319.5380  1226.3091  4116.4054   983.9139
         33          34          39          40          41          45          48          49          53          54
1440.7356  1161.0488  3398.5428  2373.0247  4927.4970  5533.4849  4302.8633  1077.1428  1962.8175  3295.9910
         59          63          64          65          67          69          75          76          78          79
1431.4127   573.7067  4461.3524  2102.6609  2354.3789  4843.5910  1869.5886   610.9982  1543.2874  1254.2777
         82          83          84          85          90          96          98         104         107         109
3836.7187  2913.7524 12003.5716  3510.4175  3258.6994   638.9669  1832.2970  4172.3428  2568.8054 15481.0101
        113         116         118         123         124         125         126         131         135         136
1114.4344  1254.2777  1244.9549  2624.7428 31777.4244  3659.5837  2317.0874  1468.7042 27815.1956  1338.1838
        140         145         146         147         151         152         155         158         159         162
1095.7886  1869.5886   750.8416  1207.6633  1263.6006  1692.4536   638.9669  1254.2777  1757.7139   862.7163
        163         164         165         166         169         176         178         179         180         181
4787.6536  1431.4127  5123.2777  1627.1934  5533.4849 16851.4752  2289.1187  1422.0898   834.7476  6987.8560
        187         189         192         194         195         218         223         225         229         233
 937.2994  2298.4416  2261.1500  2596.7741  1226.3091 17541.3692   909.3308  1561.9332  1338.1838  2251.8271
        234         236         237         238         250         253         260         268         269         272
 638.9669  7472.6464  6969.2102  1916.2030   993.2368  2167.9211   732.1958  3668.9066  2764.5862   536.4151
        277         283         284         285         292         300         302         304         308         312
6400.5138  3510.4175 12423.1017  2400.9934  1850.9428  2951.0440  4079.1139  3491.7717  1114.4344  2373.0247
        313         318         326         327         329         330         331         332         340         344
1748.3910  1114.4344  5831.8175  3510.4175  3780.7813  1645.8392 12432.4246  1822.9741  1310.2151  1981.4633
        347         352         353         364         367         370         373         378         387         388

> predictions_glm
           2          13          14          17          18          22          25          27          30          32
-12.220148  -2.169847  -4.686794  -3.004129  -2.077901  -5.285852  -2.882286  -3.068434  -2.673011  -2.211589
         33          34          39          40          41          45          48          49          53          54
 -2.471632  -2.222307  -5.344517  -3.067870  -2.903722  -3.335246  -2.806699  -2.175488  -2.484606  -6.924519
         59          63          64          65          67          69          75          76          78          79
 -2.653268  -2.029954  -5.893372  -4.187579  -2.588398  -2.579937  -3.445243  -2.058158  -3.094382  -3.389963
         82          83          84          85          90          96          98         104         107         109
 -4.430135  -2.755367 -13.281192  -2.525221  -4.658026  -3.401809  -6.222233  -5.733172  -2.328355 -11.695549
        113         116         118         123         124         125         126         131         135         136
 -2.324970  -4.391213  -2.870440  -3.499395 -19.819501  -7.618907  -3.078024  -3.059409  -6.477764  -2.530861
        140         145         146         147         151         152         155         158         159         162
 -2.464299  -3.362887  -2.080157  -4.475262  -2.353175  -2.363892  -2.097080  -2.266306  -2.137694  -2.656652
        163         164         165         166         169         176         178         179         180         181
 -3.502216  -2.596295  -2.814596  -3.713747  -5.346773  -6.132544  -2.437787  -2.714189  -2.112310  -7.576601
        187         189         192         194         195         218         223         225         229         233
 -2.337380  -2.678652  -2.617166  -3.354989  -2.118515  -3.250070  -2.048569  -2.086927  -3.568778  -2.169847
        234         236         237         238         250         253         260         268         269         272
 -2.066619  -5.793529  -9.033068  -2.871005  -2.180001  -3.592469  -2.375738  -6.855701  -6.325461  -1.989340
        277         283         284         285         292         300         302         304         308         312
 -6.067110  -2.447377 -18.019506  -5.222111  -2.216666  -6.070494  -4.978426  -8.846920  -2.538759  -6.574222
        313         318         326         327         329         330         331         332         340         344
 -2.164770  -2.620551  -7.050874  -6.404433  -2.607013  -6.483969  -3.989585  -2.878902  -2.597987  -4.311678
        347         352         353         364         367         370         373         378         387         388
 -2.705728  -3.089870  -4.066301  -2.538194  -3.044743  -6.179363  -4.195476  -3.464986 -10.196776  -2.490811
        389         390         393         406         408         409         410         413         420         422

> prediction_nnn
          C1         C2         C3         C4         C5         C6         C7         C8         C9        C10        C11        C12
1 21735.67 -171.6798 3533.098 1725.635 -407.6558 5596.785 1341.137 1579.224 1153.103 357.6075 703.5852 1076.23
         C13        C14        C15        C16        C17        C18        C19        C20        C21        C22        C23        C24
1 7004.671 1668.03 1302.217 2053.204 1857.642 685.8113 808.1316 11514.2 1213.285 -496.33 7253.482 3657.13
         C25        C26        C27        C28        C29        C30        C31        C32        C33        C34        C35        C36
1 1101.932 836.5793 2307.763 1367.592 1757.435 1818.332 3433.321 1561.17 25767.6 1262.686 5591.238 2623.471
         C37        C38        C39        C40        C41        C42        C43        C44        C45        C46        C47        C48
1 8313.418 6326.697 778.2928 23559.55 410.0503 4772.255 1345.171 2165.298 35522.18 11862.07 1856.693 1680.368
         C49        C50        C51        C52        C53        C54        C55        C56        C57        C58        C59        C60
1 5481.09 928.5229 631.6041 1989.033 242.5084 6603.812 792.2433 1100.72 318.8876 364.8605 119.4041 958.3733
         C61        C62        C63        C64        C65        C66        C67        C68        C69        C70        C71        C72
1 2331.087 1299.274 1324.984 2483.364 5266.53 5135.841 828.8398 1264.423 -206.2018 12953.65 561.1115 2132.568
         C73        C74        C75        C76        C77        C78        C79        C80        C81        C82        C83
1 1057.654 2104.213 -67.29053 -72.78251 -151.4594 -166.0004 2551.156 -137.3358 -354.0649 6670.08 16136.18
         C84        C85        C86        C87        C88        C89        C90        C91        C92        C93        C94        C95
1 1487.317 71.45228 2088.393 351.366 17384.12 8412.67 -571.1261 8843.178 1258.074 28411.6 4852.894 483.5168
         C96        C97        C98        C99       C100       C101       C102       C103       C104       C105       C106       C107
1 7664.544 4965.98 17137.72 855.6864 9702.24 115.8439 1035.718 10174.58 11865.19 966.0574 9248.885 2132.884
        C108       C109       C110       C111       C112       C113       C114       C115       C116       C117       C118       C119
1 1735.186 1053.029 4075.024 1196.038 1774.065 3236.836 678.9105 1634.345 8508.895 4415.389 2261.76 22555.71
        C120       C121       C122       C123       C124       C125       C126       C127       C128       C129       C130       C131
1 1346.638 964.5892 1353.487 1445.867 368.5013 454.581 901.5855 3167.624 2049.877 1390.364 622.8749 1690.651
        C132       C133       C134       C135       C136       C137       C138       C139       C140       C141       C142
1 276.3033 5122.955 1110.222 -323.2631 1043.181 1417.884 17707.29 2448.478 1691.165 2386.968 802.0612
```

Evaluation Measures:

In order to evaluate how the different estimation projects were more or less accurate to the data given from each dataset, it was used a set of evaluation measures. These would be the ones used:

- Correlation Coefficient.
- Covariance.
- Mean Square Error.
- Mean Absolute Error.
- Relative Absolute Error.
- Median Absolute Error.
- Mean Absolute Scaled Error.
- Mean Relative Absolute Error.
- Median Relative Absolute Error.
- Geometric Mean Relative Absolute Error.
- Relative Mean Absolute Error.
- Relative Mean Square Error.

According to the kind of measure is provided from the ones above, it can be seen when they are executed different values like “Inf” (referring to infinite), “NA” (referring to not available, which happens when there are not enough data to obtain the result) and integer/float numbers and it can be joined with a set of warnings in relation with how the train and test datasets are built.



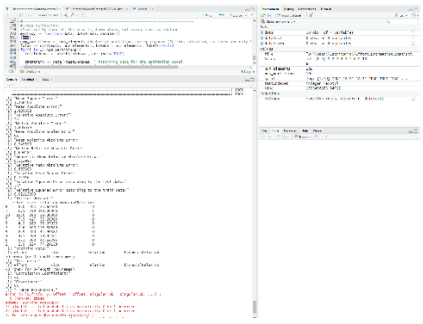
All the measures except the correlation coefficient and the covariance, can change according to the different kind of prediction used if the train and test data are maintained for each of them while correlation and covariance only depend on the dependent and independent variables established to be studied.

Other Techniques:

With all the estimation models and evaluation measures used, it was also used the cross validation which is used to evaluate the different results from different statistical analysis which guarantees the independency between the train and the test data.

The cross validation used for the two datasets is differentiated on how many partitions each of them are represented. In the case of the dataset with a higher number of elements, it was only shown a few of the possible analysis which could be obtained having only one element as train data. In the case of the other dataset, the number of analysis obtain are the ten possible resulting analysis if only it is selected one of its elements as part of the training data.

This technique uses all the operations used to build each of the different models and evaluation measures used and is repeated for each analysis made. In this case, it is necessary to print the different results obtained as the command is not directly executed.



Considerations:

Regarding this practice, it is needed to consider that:

- It was made using RStudio IDE in order to use R language.
- All the packages used for both R files it was needed to be the first time installed with the command 'install.packages("name_package")'
- To execute any of the operations, it is needed to do the next sequence of keys: Ctrl + Enter.
- The packages used for a practice need to be loaded each time, RStudio is initialized. The command needed is "library(name_package)".
- At the linear regression it is needed sometimes to execute again the command given for its representation to have a correct picture of data. This mainly happens with the linear regression.
- The practice is divided in different sections which can be done at the order the user wants, but it is compulsory to have executed the operations established before the section for correlation coefficient and covariance.
- It is needed to consider that the section related to the cross validation it is recommended to be executed after or without executing the rest of sections, as the variables "data_train" and "data_test" will be changed from the one given initially making to obtain the same results as when it is made alone.
- The operations at the function used for the cross validation are the same as the ones used for the rest of sections, but with the difference that the results obtained each operation are printed at screen and it is not needed to execute them.
- The idea of having both representations at the R document is to show a specific study having specific set of train and test data using the approximations used for this kind of studies (70-80% of data are for training and 20-30% of data are for testing) distributing them in a random way, and using a cross validation (in this case, the set of train data could only be made of one element).
- The number of validations which can be made with cross validation can be changed at any moment on the source code depending on the specifications of the user.
- The project contains three datasets and three R files. Two of each are used for the study of the linear regression, the logistic regression and the neural networks models while the other show an example of use of the regression tree.

Bibliography:

- Correlations in R: https://www.statmethods.net/stats/correlations.html
- Read table in R: https://www.rdocumentation.org/packages/utils/versions/3.6.2/topics/read.table
- Data structures in R: http://ocw.uc3m.es/estadistica/aprendizaje-del-software-estadistico-r-un-entorno-para-simulacion-y-computacion-estadistica/estructuras-de-datos-en-r
- Introduction to R: https://cran.r-project.org/doc/contrib/R-intro-1.1.0-espanol.1.pdf
- Package mlr: https://cran.r-project.org/web/packages/mlr/mlr.pdf
- Model Training and Turing: https://topepo.github.io/caret/model-training-and-tuning.html
- Confidence Interval: https://fhernanb.github.io/Manual-de-R/ic.html
- Measures of Evaluation in Software Engineering: http://danrodgar.github.io/DASE/evaluationSE.html
- Mean Squared Error in Linear Regression in R: https://stats.stackexchange.com/questions/107643/how-to-get-the-value-of-mean-squared-error-in-a-linear-regression-in-r
- Package ftsa, error function: https://www.rdocumentation.org/packages/ftsa/versions/3.4/topics/error
- Documentation package Metrics: https://cran.r-project.org/web/packages/Metrics/Metrics.pdf
- Forecast function in R: https://www.rdocumentation.org/packages/forecast/versions/8.12/topics/forecast
- Documentation package ftsa: https://cran.r-project.org/web/packages/ftsa/ftsa.pdf
- Forecast error measure in R: http://finzi.psych.upenn.edu/library/ftsa/html/error.html
- Forecast error measure in R: https://rdrr.io/cran/ftsa/man/error.html
- Simple and multiple logistic regression in R: https://rpubs.com/Joaquin_AR/229736
- Introduction Artificial Neural Networks in R: https://rpubs.com/rdelgado/402754
- Aggregation and Restructuring data in R: https://www.r-statistics.com/tag/transpose/
- Conversion H2OFrame to a data frame: https://stackoverflow.com/questions/43189340/how-to-convert-a-column-in-h2oframe-to-a-python-list
- Median in R need numeric data: https://stackoverflow.com/questions/13204008/median-in-r-need-numeric-data

- Cross validation in R: https://www.diegocalvo.es/validacion-cruzada-en-r/
- Conditionals in R: https://stackoverflow.com/questions/54825426/how-to-use-conditional-statement-and-return-value-for-a-function-in-r
- Add calculated fields to data in R: https://www.dummies.com/programming/r/how-to-add-calculated-fields-to-data-in-r/
- Logistic Regression in R: http://idaejin.github.io/courses/R/2019/euskaltel/regresion-logistica.html
- Simple and multiple logistic regression: https://rpubs.com/Cristina_Gil/Regresion_Logistica
- Example logistic regression with R: https://rpubs.com/emilopezcano/logit
- Simple and multiple logistic regression in R: https://rpubs.com/Joaquin_AR/229736
- Creation of random numbers from time stamp: https://decisionstats.com/2013/09/28/using-r-for-random-number-creation-from-time-stamps-rstats/
- Introduction to cross validation in R: https://rpubs.com/rdelgado/405322
- Minimum sample size to make a correlation test in R: https://www.researchgate.net/post/What_is_the_minimum_sample_size_to_run_Pearsons_R
- Neural Networks: https://es.wikipedia.org/wiki/Red_neuronal_artificial
- Regression Trees: https://www.sciencedirect.com/topics/computer-science/regression-tree#:~:text=Analogously%20to%20decision%20trees%2C%20regression,by%20discretizing%20the%20target%20variable.
- Linear Regression: https://en.wikipedia.org/wiki/Linear_regression#:~:text=In%20statistics%2C%20linear%20regression%20is,is%20called%20simple%20linear%20regression.