Generalized Linear Model for predicting the life expectancy of countries as a function of air pollution

Santiago Pérez Lete

Abstract– This project deals with the development of different generalized linear models for the prediction of life expectancy of countries, based on different metrics of air pollution in these countries. Different families of the Generalized Linear Model (GLM) have been used to perform this work which is able to approximate the value of the dependent variable (Life Expectancy) using three independent variables. The implementation of these models is not a very complicated task, however, they can play an important role in the health of the world population. In this work, we have first chosen a dataset and the variables to work with by carrying out an exploratory analysis and then we have adapted these data for the different prediction models. Then, the implementation of the different prediction models using the Binomial, Poisson, Gaussian, Inverse Gaussian and Gamma family of the Generalized Linear Model (GLM). Finally, the results obtained from these models have been evaluated based on their Akaike Information Criterion (AIC), Residual Standard Error (RSE) and Residual Deviance (RD).

Keywords– Statistical analysis, multiple regression, logistic regression, generalized linear model

1 INTRODUCTION

In recent years, air pollution has become a global problem affecting the health of millions of people around the world. Both developed and developing countries face significant challenges in relation to air quality and its effects on the population. Exposure to harmful air pollutants, such as greenhouse gases and other particulate pollutants, has adverse consequences for human health, ranging from acute respiratory problems to chronic diseases and reduced life expectancy.

Air pollution can be caused by a variety of sources, including industrial emissions, the burning of fossil fuels, unsustainable agricultural practices, and improper waste management. These pollutants can be released both locally through the practices mentioned above and through transportation processes.

The detrimental health effects of air pollution are varied and range from acute respiratory problems, such as asthma and respiratory tract infections, to chronic diseases, such as cardiovascular disease, cerebrovascular disease, chronic obstructive pulmonary disease (COPD), and lung cancer. In addition, long-term exposure to high levels of air pollutants can reduce life expectancy. Importantly, air pollution does not affect everyone equally, as certain population groups, such as children, the elderly, and people with pre-existing health conditions, are more vulnerable to adverse effects. In addition, socioeconomic differences between countries and the lack of access to health care services in some countries may increase the health impacts of pollution on their populations.

Addressing the problem of air pollution and its effects on health requires concerted efforts at the national and international levels. Adopting sound environmental policies, promoting cleaner and renewable energy sources, implementing sustainable agricultural practices, and improving air quality in urban areas are crucial measures to reduce exposure to air pollutants and protect the health of the population.

The aim of the work is to predict the life expectancy of different countries in the world based on the air pollution in each country.

In order to make this prediction, the data will be subjected to different models that can be classified according to the type of the dependent variable or output (life expectancy):

- Binary: binomial model
- Grouped in sets: poisson model

• E-mail de contacte: santipl2001@hotmail.com
• Treball tutoritzat per: Walter Andrés Ortíz Vargas (Facultad de Matemáticas UAB)
• Curs 2022/23

- Real value:
  - Gaussian or Multiple model
  - Inverse Gaussian model
  - Gamma model

To make this prediction, we will start from a dataset containing 3 measures of pollution for each country, which are: Particulate Matter emissions, Carbon Dioxide emissions and Methane emissions, and the life expectancy of each country in the world. This dataset will be randomly divided, 80% will be used to train the model and the remaining 20% will be used for testing.
Finally, once the new results have been predicted, the prediction quality of the model will be evaluated through various evaluation metrics. Also, the conclusions obtained from the whole development of the work will be presented, as well as issues to be improved in a future development and a cost assessment.

2 OBJECTIVES

The overall aim of the work is to predict as accurately as possible life expectancy as a function of air pollution levels in countries around the world.
However, there are other, more specific objectives:

- Generate statistical prediction models using different GLM families (Binomial, Poisson, Gaussian, Inverse Gaussian and Gamma).
- Determine which of these models best fits the data and which one gives the best result.

3 PREPARATION

3.1 Literature Review

The World Factbook is an almanac published by the Central Intelligence Agency (CIA). It is produced for US policymakers and coordinated throughout the US Intelligence Community, presents the basic realities about the world in which we live. They share these facts with the people of all nations in the belief that knowledge of the truth underpins the functioning of free societies. And this is the site from which I have extracted the data for the paper.
Within this large document is included data from many different fields such as: Economic, Social, Political, etc. In my case, the options I will work with will be on the one hand climatic and on the other hand health (Life expectancy).
Prolonged exposure to air pollution is a major health risk, according to a study published in the journal Cardiovascular Research. The authors conclude that pollution shortens life expectancy by an average of 2.9 years.
Miguel Sánchez, a nurse at the General Council of Nursing Research Institute, explains: "This is because pollution causes, above all, damage to blood vessels due to increased oxidative stress, which in turn leads to increases in blood pressure, strokes, heart attacks, heart failure and diabetes. This health problem particularly affects older people, according to Sanchez: "It is estimated that 75 percent of deaths attributed to air pollution are in people over 60 years of age." (See [2]).
These pollutants include carbon dioxide (CO2), a colorless and odorless gas, which is extracted from the combustion of fossil materials (coal, oil derivatives, biomass, etc.) and the aerobic respiration of animals. Regarding its effect on health, carbon dioxide is not toxic or even harmful to human health, nor is it useful for breathing, so that high concentrations of this gas in the air produce an uncomfortable sensation because it displaces oxygen from the air and makes breathing more tiring (See [6]).
Another of these pollutants is methane, this gas is not toxic and is not dangerous if inhaled in small amounts; however, if a large amount of natural gas or methane displaces the air, the lack of oxygen could cause asphyxiation (See [5]).
Through different multivariate statistical prediction models, we are going to try to predict the life expectancy of countries as a function of the countries' pollution.
The Generalized Linear Model (GLM) has been used to develop these models using different families. The first works where the Generalized Linear Model is introduced and developed are, respectively, Nelder and Wedderburgn (1972) and McCullagh and Nelder (1989). The Generalized Linear Model (GLM) as well as statistical modeling are methodological tools that allow codifying all analysis situations within the same general scheme. Obviously, this facilitates the learning of new analysis models because it is simply a matter of contemplating them as particular cases of a more general model already known, the Linear Model (LM) (See [10]).
While in the ML there is an identity relationship between the fitted values and the linear predictor, $\eta_i = \mu_i$, in the GLM the linearity is established on the scale of the linear predictor but not on the scale of the fitted values. There is, therefore, no identity between the fitted values and the predicted values but between them there is a function that relates them, the linking function: $\eta_i = g(\mu_i)$. The random component of the Linear Model uses a normal distribution, this fact has a considerable importance: depending on the distribution of the errors will be the conditional distributions of the predicted values of the criterion. In the GLM it happens that the random component does not necessarily follow a normal distribution but uses any distribution of the exponential family and, consequently, the distributions of the predicted values of the criterion will not necessarily be normal. The distributions that will be used in the models used in this work are the binomial, Poisson, Gaussian, inverse Gaussian and gamma distributions.

3.2 Methodology

The study investigates how independent variables in each country affect the life expectancy of each country. With the aim of creating a model in which life expectancy can be predicted as a function of these variables, different methods have been applied. Figure 1 shows a summary of all the steps that were necessary to reach the objective of the study.

With regard to the creation of the model and its interpretation or evaluation, the steps to be followed are as follows:

1. Specification of the theoretical model. Determining which variables are of interest, as well as the relationships between them. Basically, it consists of determining the dependent and independent variables and defining the model to which they will be subjected.

2. Parameter estimation. Calculate the value of the coefficients of the model created from the chosen data set.

3. Model selection. Assess whether the level of discrepancy between the actual data and the predictions made by the model is low enough to consider the model or high enough to reject the model.

4. Model evaluation. Examine individual observations, influential data and outliers. As well as checking the assumptions of normality, linearity and independence.

5. Interpretation of the model. Understanding the implication of the model with respect to the response variable. This part needs a detailed explanation of the model parameters to check that they meet the statistical, logical and substantive criteria.

Finally, the model is accepted or not depending on all the evaluations mentioned above.

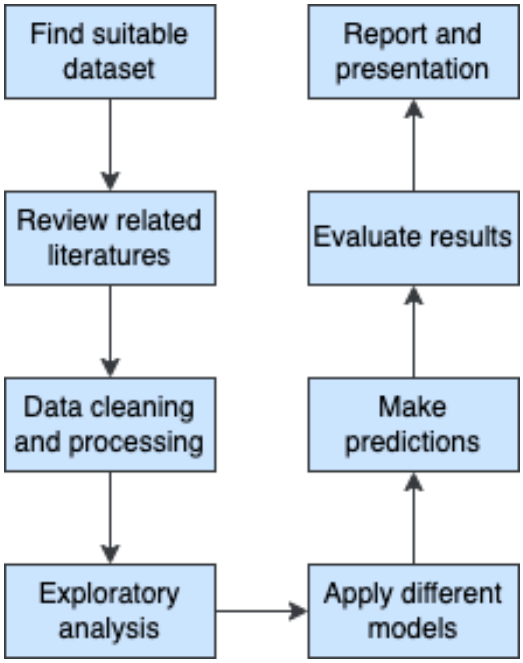

Fig. 1: Flowchart of the study's work development.

4 Data description and variables

This part of the paper describes some characteristics of the countries investigated. The following subsections show the different variables chosen and an exploratory analysis of them and the relationship between them.

4.1 Data acquisition

The data used in the project comes from The World Factbook, produced for US policymakers and coordinated throughout the US Intelligence Community, presents the basic realities about the world in which we live. These facts are shared with the people of all nations in the belief that knowledge of the truth underpins the functioning of free societies. Within these records there is data on every country in the world, geographically, economically, environmentally, etc. All this is published on the website of the Central Intelligence Agency (CIA) and has been put in a dataset and published in kaggle, from where I have been able to download the csv. This dataset initially has 258 rows and 1054 columns. Given the diversity of characteristics of each country in the dataset, it was a complicated but important task to choose the variables. After trying several options such as: studying the Unemployment rate as a function of investment in education, inflation and GDP of each country. The variables chosen were:

- Dependent variable of the model:
  - Life Expectancy
- Independent variables of the model:
  - Particulate Matter emissions
  - Carbon Dioxide emissions
  - Methane emissions

Variable	Description	Type
Life Expectancy	Shows life expectancy from each country.	Numeric
Particulate Matter Emissions	This variable indicate, in micrograms per cubic meter, the particulate matter emissions.	Numeric
Carbon Dioxide emissions	Shows the carbon dioxide emissions in megatons.	Numeric
Methane emissions	Shows the methane emissions in megatones.	Numeric

Table 1: Variables features.

4.2 Data cleaning

The dataset, having been extracted and created manually by the CIA, as I mentioned earlier, is not very buggy and is quite easy to process and use. Despite all this, the dataset has been subjected to a filtering and cleaning process to adapt the data to our needs and objectives. The steps that have been taken for this cleaning have been as follows:

1. The first step has been to filter from the dataset the variables we want to use for our model using Python. Therefore, a sub-dataset has been created with the name, life expectancy, particulate matter emissions, carbon dioxide emissions and methane emissions of each country.

2. Once you have filtered the columns you want to use, they are renamed to make it easier to work with them for further statistical analysis..

3. Subsequently, it is searched for null values in any column of any row of the dataset and in case there is a null value in a row, that row is deleted in order to avoid inconsistencies in the models.

4. The next problem we have encountered is that the columns have a numeric part, which is what we need, and a text part. For example, the Life Expectancy: 76 years. Therefore, we proceed to delete all the text of

the variables that should be numeric. After all these steps the dataset has become 187 rows and 6 columns.

5. In the project different models will be realised and not all of them require the same type of dependent variable. Therefore, 3 different variants of this dataset have been created from the previously exported dataset.

    - One of the variants, identical to the exported one, requires the dependent variable (Life expectancy) in its real values.
    - Another of the variables requires Life expectancy to be a binary variable, so the original dataset has been modified and a 1 has been added to Life expectancy in countries where it is equal to or greater than 75 years and a 0 where it is less than 75 years.
    - For the last variant, the dependent variable needs to be divided into groups, so 4 groups have been created for Life expectancy, which are distributed as follows:
        - 0 if Life expectancy is smaller than 65.
        - 1 if Life expectancy is between 65 and 72,5.
        - 2 if Life expectancy is between 72,5 and 80.
        - 3 if Life expectancy is larger than 80.

6. Finally, for each variant, the dataset was randomly divided into train and test. With a distribution of 80% for the train and 20% for the test.

4.3 Exploratory analysis

After performing the data cleaning and having the data as we are interested in for our objective, we are going to analyse the variables we will use and how they are related to each other in order to see if they would work well in a statistical prediction model. Unlike data cleaning, the exploratory analysis will be done using the R language in RStudio.

To do this, first of all a correlation matrix has been made with the variables mentioned above. This is done by calculating the correlation matrix, of the 3 types of dataset that have been created for the project (explained in the previous section) one of these matrices is calculated for each one. Here we can see that the variable ”methane emissions” is highly correlated with the variable ”carbon dioxide emissions” and that the dependent variable of the model, Life Expectancy, is highly correlated with the variable ”particulate matter emissions”. The following image shows the correlation matrix of the dataset for the multi-models (without grouping Life expectancy). Although the correlation matrix of the three types of datasets are similar.

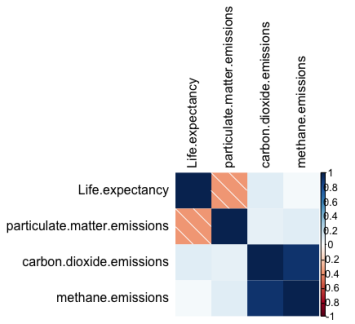

Fig. 2: Correlation matrix dataset ”multiple”.

Also as an exploratory analysis and to help in the division of the binary and grouped datasets, a histogram of the Life Expectancy variable has been made. With this we can see the frequencies of Life Expectancy and help us to divide the dataset in half or in equal groups.

5 Models implementation

In the project, in order to predict the life expectancy of countries based on their pollution levels as accurately as possible, different prediction models have been made in order to see which one best fits the data and gives the best results.

Statistical prediction models are important tools for understanding and predicting the behaviour of data. One of the most useful models in statistics is the generalised linear model (GLM), which is used to model response variables that do not follow a normal distribution, such as binary variables or event counts. The GLM is an extension of the classical linear model and provides a unified theoretical framework for modelling a wide variety of responses.

Lineal Model (LM)	Generalized Lineal Model (GLM)
$y_i = \sum_j \beta_j X_{ij} + \varepsilon_i$	$y_i = \sum_j \beta_j X_{ij} + \varepsilon_i$
$\mu_i = E(Y_i)$	$\mu_i = E(Y_i)$
$\eta_i = \sum_j \beta_j X_{ij}$	$\eta_i = \sum_j \beta_j X_{ij}$
$\eta_i = \mu_i$	$\eta_i = g(\mu_i)$
y_i: response variable vector	
X_{ij}: predictor variables and covariates matrix	
β_j: parameters vector	
η_i: lineal predictor vector	

Table 2: ML and GLM comparison.

The GLM can be applied to many types of data and can be customised to fit specific data needs. It is a flexible model that can handle data that do not meet the assumptions of the classical linear model, such as normality and homoscedasticity.

Within the generalised linear model (GLM), there are different families of distributions that can be used to model the response variable, where each distribution family has different properties and is used depending on the nature of the data and the research question.

The families used in this project are the following: Binomial distribution, Poisson distribution, Gaussian distri-

bution (Normal or Multiple), Inverse Gaussian distribution and Gamma distribution. These families will be explained in more detail below.

5.1 Binomial family

The binomial distribution is one of the most common distributions in statistics and is used to model binary response variables, i.e. variables that can only take two possible values: success or failure, presence or absence, etc.

This distribution is an important tool in the construction of binomial prediction models, as it allows the binary response variable to be modelled using an appropriate probability function.

In a binomial prediction model, one or more predictor variables are used to predict the probability that the response variable will take a value of success or failure. For this, a link function is used that transforms the probability of success into a linear scale so that a linear regression model can be fitted. In fact, is used as the probability function in this type of model to model the probability of success or failure of the response variable at each observation. The most commonly used link function in binomial prediction models is the logit function, which transforms the probability of success into a linear scale.

The model then fits a straight line to the transformed data, allowing it to predict the probability of success or failure for new observations.

In the project, in order to analyse life expectancy with respect to pollution in countries with a binomial model, the dependent variable, i.e. life expectancy, has been modified so that when it is less than a certain value it is 0 and when it is greater than a certain value it is 1.

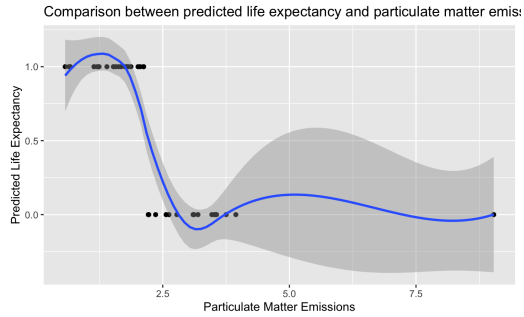

Fig. 3: Comparison between predicted values and particulate matter emissions.

The implementation that has been followed in the work to carry out the binomial model has been to create a model with the glm function of R-studio using the binomial family, all this applying it to the set of data that we have determined to train the binomial model. Once we had the model, the next step was to calculate the anova, which is used to analyse the variance in statistical models, and the confidence interval of the variables.

5.2 Poisson family

A Poisson statistical prediction model is a type of model used to predict the occurrence of events or counts of events in a given period of time or space. It is based on the Poisson distribution, which is a discrete probability distribution used to model the number of events occurring in a fixed interval of time or space, when the average rate of occurrence is constant.

The prediction model, a response variable representing event counts is used and related to one or more predictor variables that may influence the rate of occurrence of events. The objective is to find a statistical relationship between the predictor variables and the event occurrence rate, and to use this relationship to make predictions about future events.

On the other hand, the model is fitted to the data using estimation techniques, such as the maximum likelihood method, to find the parameters that best fit the observed data. Once the model has been fitted, predictions can be made about the rate of occurrence of events for new observations or situations.

In the model created in this project to carry out a study on the influence of certain pollution metrics on the life expectancy of countries, these events or counts required by the dependent variable have been determined by dividing this (Life Expectancy) into 4 options, as explained above on Data Cleaning. In this way a probability distribution is created between the number of options and with this a Poisson prediction model with our data.

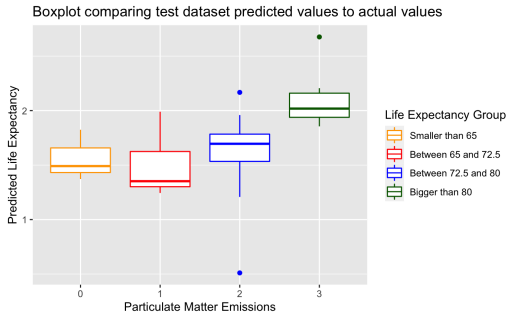

Fig. 4: Boxplot comparing test predicted values to actual values.

Regarding the implementation of the Poisson prediction model, the first thing we have done is to create a model with the glm function of R-studio using the Poisson family, applying it to the data set that we have determined to train the Poisson model. Once we had the model, the next step was to calculate the ANOVA and the confidence interval of the variables.

5.3 Gaussian family

A Gaussian prediction model, also known as a multiple linear regression model, is an extension of the Gaussian prediction model that allows multiple predictor variables to

predict a continuous response variable.

In a multiple Gaussian prediction model, one seeks to establish a linear relationship between the response variable and several predictor variables. The idea is to find regression coefficients that weight the contribution of each predictor variable in predicting the response variable.

The multiple Gaussian prediction model is fitted using estimation methods, such as the least squares method, which finds the optimal values for the regression coefficients. Once the model has been fitted, predictions can be made for new observations using the values of the predictor variables.

This type of model allows multiple predictor variables and their simultaneous effects on the response variable to be taken into account. It is especially useful when it is suspected that several variables may influence the response variable and it is desired to know their relative impact.

When making this model in this work, the dependent variable has not had to be modified as in previous cases and the primary dataset has been used, since in this type of model the dependent variable must be a continuous numerical variable.

As for the implementation, the first thing we did was to create a model with the glm function of R-studio using the Gaussian family, applying it to the dataset that we determined to train the multiple model. Once we had the model, the next step was to calculate the ANOVA and the confidence interval of the variables.

5.4 Inverse Gaussian family

A Gaussian inverse distribution statistical prediction model is a model used to predict continuous positive values that follow a Gaussian inverse distribution.

This distribution is also known as the Wald distribution, is a continuous probability distribution used in a variety of statistical contexts. Unlike the ordinary (or normal) Gaussian distribution, which describes continuous values, describes positive values that follow an asymmetric distribution and is characterised by a long rightward tail.

The statistical prediction model of the inverse Gaussian distribution is based on the assumption that the response variable follows an inverse Gaussian distribution. This distribution is defined by two parameters: mu (μ) and lambda (λ). The parameter mu represents the mean of the distribution and lambda represents the scale or dispersion parameter.

In this type of model, regression coefficients representing the effect of the predictor variables on the mean of the inverse Gaussian distribution are fitted. The objective is to find the optimal coefficients that minimise the discrepancy between the observed values and the values predicted by the model.

As in the case of the previous Gaussian model, no changes have to be made in this model in order to be able to run the model, since the model needs its real values to predict these, so we will use the primary dataset.

The implementation that has been followed in the Inverse Gaussian prediction model has been very similar to the previous one, the first thing that has been done has been to create a model with the glm function of R-studio using the Inverse Gaussian family, all this applying it to the data set that we have determined to train the multiple model. Once we had the model, the next step was to calculate the ANOVA and the confidence interval of the variables.

5.5 Gamma family

A gamma statistical prediction model is a type of model used to predict continuous positive values that follow a gamma distribution. The gamma distribution is a continuous probability distribution that is widely used in statistics to model positive-valued and asymmetric variables.

The gamma statistical prediction model is based on the assumption that the response variable follows a gamma distribution. This distribution is characterised by two parameters: alpha (α) and beta (β). The α parameter represents the shape or skewness of the distribution, while the β parameter represents the scale parameter.

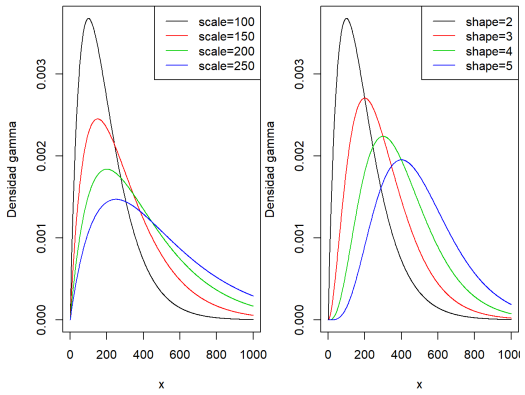

Fig. 5: Gamma Distribution.

In such a model, regression coefficients representing the effect of the predictor variables on the alpha and beta parameters of the gamma distribution are fitted. The aim is to find the optimal coefficients that minimise the discrepancy between the observed values and the values predicted by the model. Once the model has been fitted, predictions can be made for new observations using the values of the predictor variables.

When implementing this model, the primary dataset had to be used, as in the two previous cases. This is because the model, by using a gamma distribution, allows predicting the actual values of the dependent variable.

As for the implementation of the Gamma prediction model, it has been done in a very similar way to the previous one. The first thing we did was to create a model with the R-

studio glm function using the Gamma family, applying it to the dataset we had determined to train the multiple model. Once we had the model, the next step was to calculate the ANOVA and the confidence interval of the variables.

5.6 Functions

These are the betas of the different forecasting models:

Model	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$
Binomial	1.08313	-0.50065	0.03664	-0.11596
Poisson	0.887694	-0.174744	0.002992	-0.014372
Multiple	77.44023	-1.67877	0.03210	-0.13549
Inverse Gaussian	1.641e-04	9.291e-06	-1.569e-07	6.859e-07
Gamma	1.285e-02	3.314e-04	-5.824e-06	2.519e-05

Table 3: Prediction functions betas.

An example of what these functions would look like, where x_1 is the particulate matter emissions, x_2 is carbon dioxide emissions and x_3 is methane emissions:

- Binomial:
  $E(Y)=\frac{\exp(1.08313-0.50065x_1+0.03664x_2-0.11596x_3)}{1+\exp(1.08313-0.50065x_1+0.03664x_2-0.11596x_3)}$
  Where:
  - $E(Y)$ is the probability of success in a binary event.
- Poisson:
  $E(Y)=exp(0.887694-0.174744x_1+0.002992x_2-0.014372x_3)$
  Where:
  - $E(Y)$ is the expected value, which will be a number close to the sets of dependent variables created.
- Gaussian or Multiple:
  $E(Y)=77.44023-1.67877x_1+0.03210x_2-0.13549x_3$
  Where:
  - $E(Y)$ is the expected value, in this case it will be a value between the real ranges of the dependent variable (Life Expectancy).
- Inverse Gaussian:
  $E(Y)=1/(1.641e-04+9.291e-06x_1-1.569e-07x_2+6.859e-07x_3)$
  Where:
  - $E(Y)$ is the expected value, in this case it will be a value between the real ranges of the dependent variable (Life Expectancy).
- Gamma:
  $E(Y)=exp(1.285e-02+3.314e-04x_1-5.824e-06x_2+2.519e-05x_3)$
  Where:
  - $E(Y)$ is the expected value, in this case it will be a value between the real ranges of the dependent variable (Life Expectancy).

6 Results

To analyze the results and see which could be the best option, we have taken different metrics that all types of prediction models have in common and compared them with each other. The different metrics are:

6.1 Akaike Information Criterion (AIC)

This criterion measures the explanatory power of the model and at the same time penalises its complexity.

It measures the goodness of fit based on the maximum likelihood. It measures complexity based on the number of parameters. It is not very restrictive in terms of complexity, because if the model has a lot of data it will not penalise us. The formula is as follows:

$$AIC = 2K - 2ln(L),$$

where K is the number of independent parameters in the statistical model and L is the maximum likelihood function. The term -2ln(L) is known as the variance.

AIC must be calculated for each model and the model with the lowest AIC value must be selected as the best.

With the first term the number of parameters is penalised with the Parsimony Principle (See [7]), the more the number of parameters increases the higher the AIC value. With the second term the goodness of fit is measured. Penalising 2K is similar to doing cross-validation by leaving one data out. AIC is a fairly easy criterion to implement.

Calculating the difference in the AIC of the models can help us to see whether the other models are competitive or not. This difference is calculated as follows:

$$\triangle_i = AIC_i - AIC_{min}; \; i = 1...r$$

Where AIC_{min} is the lowest value of AIC for the compared models. That model will be the best of the set of models. If $\triangle_i \leq 2$ has substantial empirical support. If $\triangle_i$ is between 4-7 have considerably less empirical support. If $\triangle_i \geq 10$ the model does not compete to be the best.

6.2 Residual Standard Error (RSE)

The residual standard error (RSE) or residual standard error is a measure of the unexplained variability in a regression model. It is defined as the standard deviation of the residuals, which are the differences between the observed values and the values predicted by the model.

The formula for the RSE is as follows:

$$RSE = \sqrt{\frac{SSE}{(n-p-1)}},$$

where:

- SSE: sum of the Squares of Errors, which is the sum of the squares of the differences between the observed values and the values predicted by the model.
- n: number of observations in the data set.
- p: number of predictive variables in the model.

The RSE formula involves dividing the Sum of Squares of Errors (SSE) by the number of degrees of freedom remaining after fitting the model. The denominator n-p-1 is used to correct for the overestimation of the standard error due to the fit of the model to the training data.

In other words, the RSE measures how much the true values vary around the fitted regression line on average. A smaller RSE indicates that the model fits the data better and that most of the variation is explained by the model. A larger RSE indicates that the model does not fit the data well and that there is a large amount of variation that is not explained by the model.

The RSE is often used as a measure of the goodness-of-fit of a regression model. The lower the RSE, the better the fit of the model to the data. However, it is important to remember that the RSE only provides information on the quality of the model fit and not on the validity of the model itself.

6.3 Residual Deviance (RD)

Residual deviance is a measure of the goodness of fit of a logistic regression model in statistics. It is used to assess how much the observed data deviate from the values expected by the model.

Deviance is a measure of the discrepancy between the model-adjusted probability and the observed probability of an event. Residual deviance measures the residual deviance after fitting the model to the data.

RD is calculated as the difference between the null deviance and the residual deviance:
RD = Null Deviance - Deviance of the fitted model

The Null Deviance or total sum of squares (SST) is the deviance obtained by fitting a model with no predictor variables to the data. The other value, the residual sum of squares (SSR) is the deviance obtained by fitting the model with predictor variables to the data.

For practical purposes, this formula would be carried out as follows:

$$RD = -2log(LR),$$

where LR is the likelihood ratio statistic (also known as the chi-square test statistic) for the comparison of the fitted model with the null model.

The likelihood ratio (LR) statistic is used to compare the goodness-of-fit of the fitted model with the null model. The null model is a model that has no predictor variables and only one parameter to estimate the probability of success. The fitted model is the model that fits the data with the predictor variables included.

The LR is calculated as the difference between the deviances of the fitted model and the null model, multiplied by -2. The deviance is a measure of the lack of fit of the model to the data, where a lower deviance indicates a better fit.The Residual Deviance is the final result of the LR calculation and provides a measure of the amount of deviance remaining in the fitted model after fitting it to the data.

A low Residual Deviance value indicates that the model fits the data well and that there is little variation unexplained by the model. A high Residual Deviance value indicates that the model does not fit the data well and that there is a lot of variation unexplained by the model.

In general, one seeks to minimise the Residual Deviance to obtain a model with a good fit to the data.

6.4 Results analysis

The result of the different models in these three metrics can be found in the following table:

Model	AIC	RSE	RD
Binomial	186.23	0.4681377	178.23
Poisson	407.88	0.7703707	78.315
Multiple	969.69	6.141677	5469.4
Inverse Gaussian	981.45	6.141677	0.015375
Gamma	976.82	6.141677	1.0834

Table 4: Results analysis.

As for the AIC, as previously mentioned, the lower the AIC value, the better the model. Therefore, in the 5 different models I have run, the best in AIC is the Binomial, which beats the rest by quite a lot as can be seen. This is because the AIC measures the goodness of fit, but penalising the complexity of the model, as the Binomial model is a binary response, it is a simpler or less complex model than the rest, as the output of the model is between 0 and 1. With the Poisson model, whose output is between 4 options, we see that the AIC rises considerably and in the rest of the models, because the ouput seeks to resemble the real value, its values oscillate between many options and therefore, being more complex, its AIC is a much higher value even than the Poisson model and therefore higher than the Binomial model.

As far as the RSE is concerned, the lower its value the better the model fits the data. Knowing this, we see that the best model is still the Binomial model. The RSE measures how much the actual values vary around the fitted

regression line on average. Therefore, it is logical that the model with the binary dependent variable has a lower value since the variability of its output is much lower than in the other models.

Finally, we analyse the RD, which in this case also means that the smaller the value, the better the model fits the data. In the RD, contrary to the previous metrics, the binomial model is not the one that best fits the data. The best fitting model is the Inverse Gaussian model, followed not far behind by Gamma.

7 Conclusions

1. In conclusion, the main objective of the work was to be able to predict the value of life expectancy as accurately as possible using as dependent variables the emission of methane, carbon dioxide and particulate matter in the air in each country in the air in each country.This objective has been achieved with the implementation of the different statistical prediction models created. As for example in Albania x_1=1.787, x_2=0.454, x_3=0.255:

$$LifeExpectancy = 77.44023 - 1.67877x_1 + 0.03210x_2 - 0.13549x_3 = 77.44023 - 1.67877(1.787) + 0.03210(0.454) - 0.13549(0.255) = 74.42$$

   Here we can see that in this country the real value of life expectancy is: 79.47 and the one predicted by the GLM prediction model of the Gaussian family is: 74.42 . Therefore we can state that the value of the dependent variable can be predicted with good accuracy.

2. With respect to the more specific objectives, the first of them was to generate statistical prediction models using different GLM families, it can be said that this specific objective has been successfully completed since the families that were originally set out in the objectives have been used, namely: Binomial, Poisson, Gaussian, Inverse Gaussian and Gamma.

3. Regarding the second specific objective, determine which of these different models best fit the data, we can conclude by saying that the one that has obtained the best overall results, and therefore best fits the data, is the Binomial due to the simplicity of its dependent variable (binary). On the other hand, if we only take into account the models that predict a value between the real ranges of the dependent variable, i.e. the Gaussian, Inverse Gaussian and Gamma families, the best performing family that best fits the data depends on which metric we are looking at. For example, if we evaluate the models by AIC, the one that best fits the data is the Gaussian (or Multiple) model, on the contrary, if we evaluate the models by Residual Deviance (RD), the one that best fits the data is the Inverse Gaussian.

8 Difficulties

During the course of the project I encountered different problems. At the beginning, I had a lot of trouble finding a dataset that I liked, that was useful for all the types of models I wanted to implement and that was not too clean. I also experienced difficulties in finding metrics that were common for all model types and in getting the values of these metrics. Finally, writing the report in latex was a challenge for me, as I had never used this language before. All the adversities that I have had throughout the work have helped me to gain experience and improve in different areas that at first I did not think I would improve.

9 Future improvements

In order to improve the work, future work could be improved, for example, by considering more measures of pollution than the three used so far, such as ozone or sulphur dioxide in the air, or even adding some other pollutant that is not in the air and that affects people's life expectancy. And also improve the models to increase the accuracy of the predictions.

Acknowledgements

First of all, I would like to express my sincere thanks to my tutor, Walter Andrés Ortiz, for his guidance, support and knowledge throughout my Final Degree Project in Data Engineering. His experience and dedication have been fundamental in the development of this project.

I would also like to thank my family, including my parents, partner, sister, grandparents, aunts, uncles and cousins, for their unconditional support. Their encouragement and affection have been a great boost to me throughout the course of the work.

In summary, the work and all my academic success would not have been possible without the support and love of these special people in my life. Thank you all so much for being there!

References

[1] *Causas de la contaminación del aire.*[Internet]; 2006 [Quoted 28 March 2023]. Url: https://www.ecologistasenaccion.org/5681/causas-de-la-contaminacion-del-aire/#: :text=Las%20principales%20causas%20de%20la,del%20transporte%20por%20carretera%2C%20principalmente

[2] *¿Cuáles son las consecuencias de la contaminación del aire ambiental exterior en la salud?*[Internet]; 2018 [Quoted 28 March 2023]. Url: https://www.paho.org/es/temas/calidad-aire-salud/contaminacion-aire-ambiental-exterior-vivienda-preguntas-frecuentes#: :text=La%20contaminaci%C3%B3n%20del%20aire%20puede,impactos%20adversos%20en%20la%20salud

[3] Dalgaard, P.: *Introductory Statistics with R*. 2008.

[4] *Dataquest* [Internet],[Quoted 24 Febreury2023].Url: https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/

[5] *El metano y la seguridad.*[Internet], [Quoted 27 March 2023]. Url: https://www.socalgas.com/es/stay-safe/methane-emissions/methane-and-health-and-safety

[6] *¿Es el Dióxido de carbono tóxico para la salud humana?*[Internet]; 2023 [Quoted 27 March 2023]. Url: https://www.solerpalau.com/es-es/blog/dioxido-de-carbono/

[7] Gori, M.: *Parsimony Principle.* [Internet]; [Quoted 14 May 2023]. Url: https://www.sciencedirect.com/topics/computer-science/parsimony-principle

[8] *Kaggle* [Internet],[Quoted 20 Febraury 2023].Url: https://www.kaggle.com/

[9] Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., Münzel, T.: *Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective.*[Internet];2020 [Quoted 27 March 2023]. Url: https://academic.oup.com/cardiovascres/article/116/11/1910/5770885?login=false#207231589

[10] López González, E., Ruiz Soler, M.: *Análisis de datos con el Modelo Lineal Generalizado. Una aplicación con R*. Universidad de Malaga. 2011.

[11] *Máxima Formación* [Internet],[Quoted 22 February 2023].Url: https://www.maximaformacion.es/blog-dat/el-papel-de-la-estadistica-en-el-machine-learning/#: :text=El%20Machine%20Learning%20es%20una%20rama%20de%20las%20estad%C3%ADsticas%2C%20por,el%20campo%20del%20Machine%20Learning.

[12] Pando Fernándezand, V., San Martín Fernández, R.: *Regresión Logística Multinomial.* 2004.

[13] Peña, M. J., Escobar, J. A.: *Por qué la contaminación acorta casi tres años de vida*[Internet]; 2022 [Quoted 27 March]. Url: https://dkv.es/corporativo/blog-360/medioambiente/contaminacion/esperanza-de-vida#:˜:text=en%20nuestra%20salud%3F-,La%20contaminaci%C3%B3n%20atmosf%C3%A9rica%20acorta%20la%20esperanza%20de%20vida,media%20de%202%2C9%20a%C3%B1os.

[14] *Towards Data Science* [Internet], [Quoted 24 February 2023].Url: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

ANNEX

The code created for the realisation, both the data cleaning part and the implementation of the models, is in a github repository which can be accessed through this link: https://github.com/santipe01/TFG_AnalisisEstadistico.git