Tiny Data and linear regression methods: a review.
Assignment 1.

Orlando Gonzalez Ortiz[A01327941], Edgar Ivan Rodríguez Medel[A00839205], and Benjamín Gutiérrez Padilla[A01732079]

Tecnológico de Monterrey
A01327941@tec.mx/A0839205@tec.mx/A01732079@tec.mx
https://tec.mx/es

Abstract. This work focuses on reproducing the case study of the article: "Wickham, H. (2014). Tidy Data." With the objective to understand and realize the importance of the process to transform a dataset from messy data to tidy data. The case of the study on the article was programmed in R language, and our main goal was to translate to Python language while we were analyzing each step involved in tidying data. In the end, we also compare three different methods for linear regression with the tidy data. In this way, we not only see which are the main steps in tidying data, but also, we could compare how tidy data is important when we try to make the analysis by linear regressions, and how it could be more difficult with messy data. In this work, we realize that one of the main steps in analyzing data is cleaning, which involves filtering and transforming the dataset into a new way easier to analyze the data. The steps we follow are described in this document and also the analysis of the article mentioned before.

Keywords: Data · linear · regression.

1 Introduction

1.1 Tidy and Messy Data [1]

The information in a database sometimes is saved as the best option to optimize space or facilitate the recovery of the data, but when you try to analyze and get specific data from the dataset the format could not be the best format to apply other methods as linear regressions.

Tidying data is part of a process of cleaning data that involves to transform the format or the visualization of a data set to make it easy to analyze. The format of the data before become tidy is called messy data.

The process to organize the data is one of the most important parts of the whole process of data analysis, tidying data provide us with a framework in which we can work easily and with better results. Obtaining the data is important, but making a good representation is the goal.

To get Tidy data there are some rules, and some of them depend on the rows, columns, and information of each cell. In tidy data every column is variable, every row is an observation and every cell is a single value. Following this pattern, we can transform from messy to tidy data.

Some of the most common problems with messy data are: column headers are values, multiple variables are stored in one column, variables are stored in both rows and columns, multiple types of observational units are stored in the same table, and a single observational unit is stored in multiple tables.

1.2 Objective

In this work, we reproduce the case study from the article: "Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1 - 23. doi: http://dx.doi.org/10.18637/jss.v059.i10Links to an external site. recovered from https://www.jstatsoft.org/article/view/v059i10Links to an external site." but in Python language, with the objective of understanding the importance of manipulating and cleaning data before making the analysis. After that, we compare three methods of linear models representing the data that is already tidy.

2 Case Study

2.1 Objectives

The case study illustrates the individual-level mortality data from Mexico in 2008, the goal is to find a relationship between causes of death with their deviation to find unusual temporal patterns and which diseases differ the most. In our case, the goal is to find the linear regression of this relationship.

2.2 Data Manipulation

There are two datasets, the code variable dataset which contains all the names of the diseases and their respective codes, and the full dataset which has information on 539,530 deaths in Mexico in 2008, and 55 variables, including location, time of death, cause of death, and graphics of the deceased.

The first thing that we did was to count the number of deaths in each hour for each cause and eliminate any null value presented in the dataset.

The table obtained before is joined with the code variable dataset, obtaining a new row that contains the name of the diseases.

Instead of comparing the proportion of the total deaths, it is more practical to group it by each hour and divided it by the total number of deaths from that cause (prop).

Also, is computed the overall average death rate for each hour (freq_all). Finally, is computed the overall proportion of people dying each hour (pro_all).

A mean squared deviation is used to find the deviation between the temporal pattern of each cause of death and the overall temporal pattern. Only diseases with more than 50 total deaths are considered to ensure that the disease is representative.

2.3 Fitting the relationship

It was needed to plot the graph to find the characteristics, using the formula of gplot.

$$R > ggplot(data = devi, aes(x = n, y = dist) + geom_point() \quad (1)$$

We found that the result was fig 1:

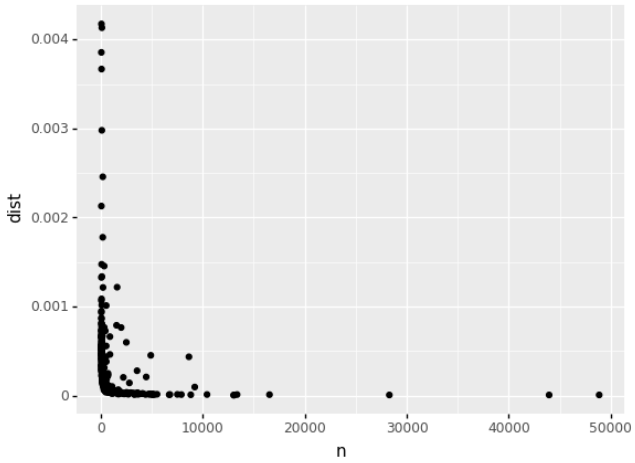

Fig. 1. Linear scale frequency vs deviation of frequency.

As we can see in the graph Fig. 2, it is better to convert it into a logarithmic scale, so it is needed to transform the total frequency (n) and the deviation of the frequency (dist) into logarithmic. In addition, they used a linear regression method of Robust fitting of Linear Modes (RLM).

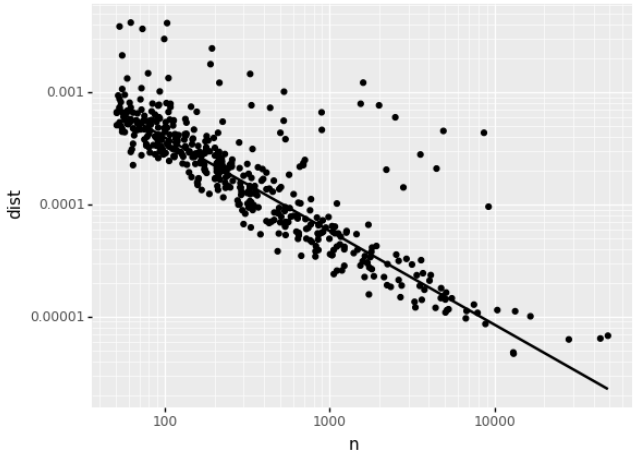

Fig. 2. Logarithmic scale frequency vs deviation of frequency.

2.4 Outliers analysis

Finally, the authors are interested in points that have high y-values relative to the x-neighbors. To find these outliers, it was needed to fit the values using robust linear models and plot the residuals. Fig. 3

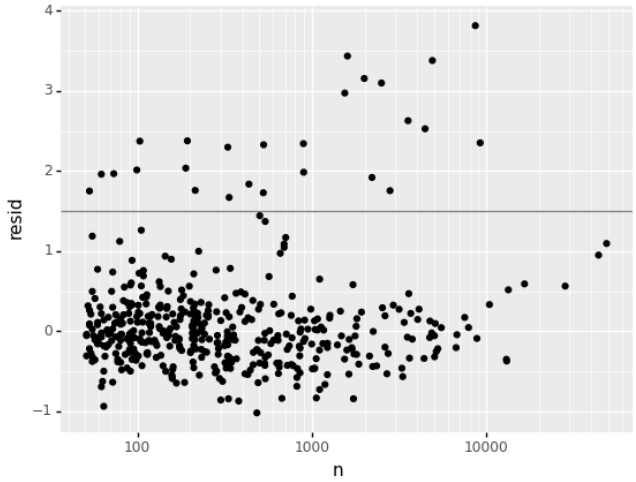

Fig. 3. Residuals from a robust linear model, predicting log of the deviation of the frequency (log(dist)) and log of the total frequency (log(n)).

As we can see in the graph, Fig. 3 it is selected a threshold of 1.5, so it is needed to select a value greater than this threshold (1.5). Finally, it is plotted the temporal course for each unusual cause (threshold>1.5). Cause of deaths, fewer than 350 deaths as it is shown in the graph Fig. 4 and greater than 350 deaths Fig. 5.

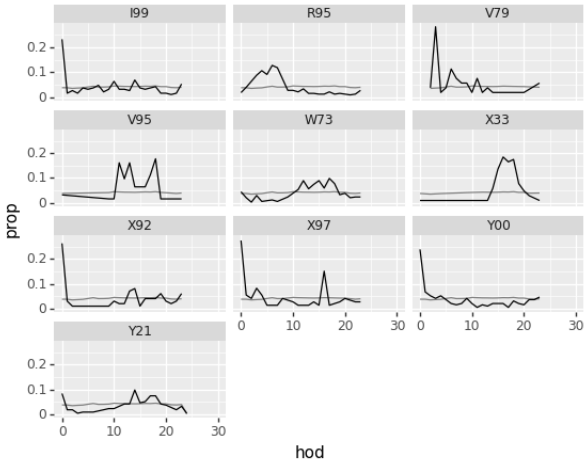

Fig. 4. Causes of death with unusual temporal courses with fewer than 350 deaths

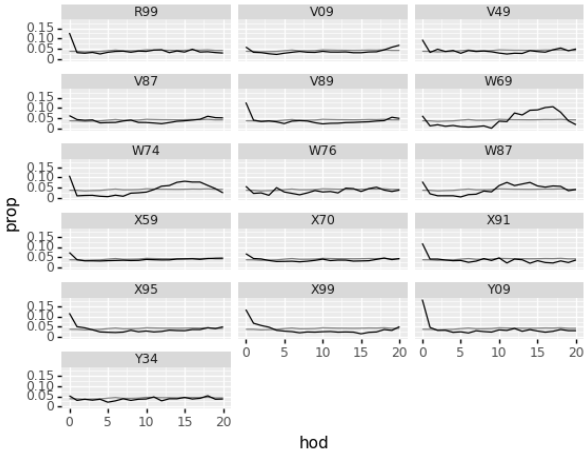

Fig. 5. Causes of death with unusual temporal courses with more than 350 deaths

3 Benchmark

3.1 Linear Regression

A Regression analysis is defined as a statistical technique for modeling the relationship between variables, and it is used extensively and widely in the science [2]. In case study, it has many observations that are plotted in the Figure 2, this type of graph is called **scatter diagram** that suggests a relationship between x and y variable* and the behavior of the data points suggest that it would be represented as a line relating that is:

$$y = A + Bx \tag{2}$$

where A is called the intercept and B is the slope[3]. As seen in the graph, the data points do not fall exactly on a straight line, and in fact, it would adjust any line with arbitrary numbers. But what of all possible lines should you adjust better? You need to use some specific methods and judge them to find the best adjustment.

3.2 Ordinary least squares Method

A particular line is a good fit if all data points are close to the line, therefore it means that the deviations from the line will be small [3]. A standard approach is to square the deviations and then sum it, this method is called ***Ordinary Least Squares Method***.
For a set of bivariate data $(x_1,y_1),(x_2,y_2),(x_3,y_3);...,(x_n,y_n)$[3]:

$$\sum [y-(a+bx)]^2 = [y_1-(a+bx_1)]^2+[y_2-(a+bx_2)]^2+...+[y_n-(a+bx_n)]^2 \tag{3}$$

the slope is:

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} \tag{4}$$

and the intercept is:

$$a = \bar{y} - b\bar{x} \tag{5}$$

where $\bar{y}$ and $\bar{x}$ are the mean of the values.

3.3 Ridge regression Method

Ridge regression is a type of *penalized regression method* that is a derivative of ordinary least squares regression, but they were designed to overcome the basic limitations of the past method. Especially, the method reduces or shrinks the values of the coefficients on size [5]. Our extra complexity parameter that controls the amount of shrinkage is denoted as $\alpha \geq 0$ where if $\alpha = 0$, the problem converts to ordinary least squares regression [4].

3.4 Lasso regression Method

Lasso is another type of *penalized regression method*. The advantage of this shrinkage methods is that the estimated models exhibit less variance than least squares estimates [5]. The difference between ridge regression and Lasso regression is the measure of length that each one uses for penalizing [4].

3.5 Comparison: similarities and differences

As it was explained in the section 4.1, we need to adjust our data points to a line to trying to describe the behavior in one linear equation (3). Table 1 contains the results obtained for each linear regression methods that were described in the last sections. Additionally, we add a third row that contains additional random values for A and B to see what happens if only approximate values are chosen randomly. As we can see in Table 1, the approximation values are very similar, it is not possible to get an obvious answer or use a qualitative method to choose an answer or make a simple decision, therefore we apply root-mean-square error (RMSE), that is calculated by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_e(i) - y(i))^2} \tag{6}$$

where N is the total number of data, y_e is the value estimated by the model of the proposed line regression method and y is the data point recorded by the data. The closer the metric value is to zero, the greater the similarity between the real data values and the estimated values.

Table 1. Results of three linear regression methods

Method	A (Intercept)	B (Slope)	y= A + Bx	RMSE
Ordinary Least Squares	-1.8360	-0.7733	y=-1.8360 -0.7733x	0.2962
Ridge regression	-1.8420	-0.7710	y=-1.8420 -0.7710x	0.2967
Lasso regression	-1.9061	-0.7455	y=-1.9061 -0.7455x	0.2962
Random values	-2.5	-0.8	y=-2.5-0.8x	0.3759

As it can see in the Table 1 and in the Fig. 6, the best methods are the Ordinary Least Squares and Lasso regression, however, all three methods adjust linear regression considerably since they obtained almost the same results and the RMSE values are close to zero, changing only in the order of hundredths or even thousandths.

On the other hand, when we compare the results using regression methods and the choice of random values, we can see the great importance of using linear

regression methods because the value of RMSE is considerably further from zero compared to those results in which that the regression methods were used, this because they use criteria to approximate the data as far as possible.

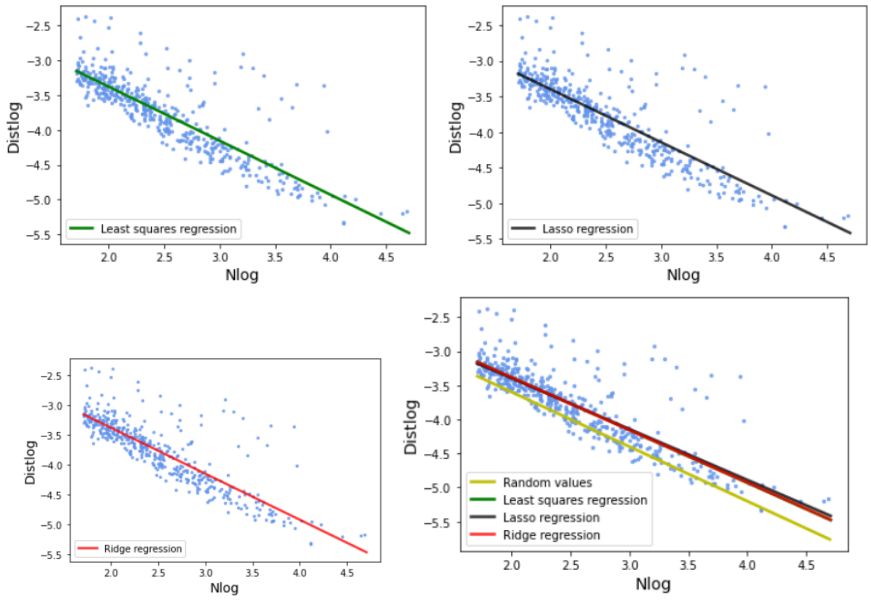

Fig. 6. Linear regression methods

As main steps, we consider that first a qualitative inspection must be done to see if the correlated data has a linear trend or not. Second, find the best regression method that makes a better approximation, taking into account the characteristics of the input data. And finally, make a comparison of the different models used, as well as evaluate each method with metrics such as RMSE to ensure the best interpretation of the data in a linear model.

4 Conclusion

4.1 Conclusion 1 (Edgar Ivan Rodríguez Medel/ A00839205)

As presented in the review and in the article, one of the characteristics of a database is the way in which they are ordered. This feature will always help us when we start to manipulate the databases. The data analysis process begins from the moment they are received until the results are obtained. As described in the article, the reality is that databases are almost never correctly tidy, which is why it is important to know the ways in which they can achieve an tidy database. In the second section of the document, it is concluded that there is no definitive method for linear regression, there may be thousands of methods, however, the main objective is to choose the appropriate one for the correlated data, this is achieved with evaluation metric (RMSE, for example). On the other hand, it is important to mention that we can apply the best linear regression method that exists, but if we do not have a correlation or a linear trend in our data, we will never be able to obtain real results. It is important to know and study the data before using any regression technique to obtain the correct methods and results.

4.2 Conclusion (Benjamín Gutiérrez Padilla/A01732079)

This work was very helpful in understanding the fundamental steps in the process of analyzing data. We mainly focused on the part of cleaning the data and converting it into Tidy Data, which is a very important process to filter and sort the dataset in a way that is easier to analyze. Reading the article I could realize how important this process is and why it should be done. With the case study we were able to test with an example how tidying data and finally when doing the linear regression models we could use the sorted data and make a better analysis than we would have done with the Messy Data. Finally we were able to use these data to compare linear regression models and conclude that there is no perfect method, and that the choice of any one method comes from experimentation with various methods.

4.3 Conclusion (Orlando Gonzalez Ortiz)

One of the main objectives of this work was to be able to process data before doing analysis, the core of this project was to clean the data frame and use some data analysis tools. The first part was to process the data and analyze them. As the article said and exemplifies the importance of processing the data before doing analysis, in that sense, it is easier for the program to be performed and also useful for the programmer to know what is he doing. Finally, the data obtained is analyzed via three linear regression models, to find the relationship between two sets of data. We use three different linear models: Ordinary least squares method, which is one of the most useful linear methods and the advantage of this method is the reduction of a quadratic error making easier the complexity of the algorithm. Secondly, it was used a lasso regression method, the main point to

focus on here is that there is less variance compared to least squares estimates. Finally, there was used the Ridge regression method, which adds a parameter α to overcome the limits that the square method has. As we can see in the results, the three methods overcome the difficulties and make a great linear relationship between the variables, the RMSE is almost the same for the three methods.

Acknowledgements Thanks to Wickham, H., author of the paper *Tidy Data. Journal of Statistical Software* [1].

References

1. Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23. https://doi.org/10.18637/jss.v059.i10
2. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to Linear Regression Analysis (6th ed.). Wiley Global Research (STMS). https://tec.vitalsource.com/books/9781119578758
3. Peck, R. (2014). Statistics: Learning from Data. Cengage Learning US. https://tec.vitalsource.com/books/9781285966083
4. Bowles, M. (2019). Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics. John Wiley & Sons.
5. Chan-Lau, J. A. (2017). Lasso Regressions and Forecasting Models in Applied Stress Testing. INTERNATIONAL MONETARY FUND.