Abstract

In this paper we will discuss the topic: Analysis of categorized data using logistic regression. Starting with some basic knowledge about variables, random variables and discrete variables we get to the variables of the categorized type. The categorized data that will be analyzed in this paper will be of the type that will use the logistic regression method. The rest of the paper will focus on logistic regression and then the combination of these in the title of the topic mentioned above. Logistic regression is a technique widely used for categorical data analysis, offering increased flexibility compared to traditional intersection analysis. A binary result can be predicted using one or more categorical variables, continuous variables or combinations.

Keywords: Categorical, analysis, data, logistic.

1. Introduction

Random variables The random variable X is called any real function X conveys in space \( \Omega \) so \( X: \Omega \rightarrow R \) while, we mark the random variables with the letters X, Y, Z,… etc (Kastrati, 2017). When studying a random change the biggest problem is finding the probability of its different values, where this problem is smart because the values of the random changes are determined by the test results, where we are dealing with a probability \( P \). We note \( H = \{1x, 2x, 3x…… nx\} \) community and possible value of number X (Kastrati, 2017).
According to Pierce (2020), the random variations and possible values are specifically explained. This is illustrated by such an example: We get an important coin, and it starts from the fact that the coin has two sides. One side contains the "head" and another number. If I count 0 my head, while I count 1 we mark the number, we can get the changes as follows: With X we mark the random changes, 0 and 1 represent the possible probabilistic changes, while the coin represents the event and the possibility that may occur. In our reality only two events can happen, the head or the coin toss number can fall. \( X = \{0, 1\} \). \( X \) e random variable, \( \{0, 1 \text{ possible values}\} \). In the second example according to (Pierce, 2020), three coins are taken to see what are the possible cases that can occur when three coins are tossed at once. If we denote the random change by \( X \), we denote the sample space by \( P \) (this represents all possible cases that may occur after the coins are tossed). Coins can fall in eight different ways. If we want the head to fall there in every possible case let’s see how many times it can happen. Imagine if the coins are tossed and with \( K \) we mark the head while with \( N \) we mark the number:¹ Example below best expresses the values mentioned above.

\[
X = \{0, 1\}.
\]

K K K 3 times the head has dropped so the value of a combination is 3.

K K N 2

K N K 2

K N N 1

---

¹ Retrieved from: [https://www.mathsisfun.com/data/random-variables.html](https://www.mathsisfun.com/data/random-variables.html)
Looking at the cases above we see that we have 1 case that dropped 3 heads, 3 cases with two heads, 3 directly with 1 head and 1 case with 0 heads. Therefore we have:

\[ P(x = 3) = \frac{1}{8} \]

\[ P(x = 2) = \frac{3}{8} \]

\[ P(x = 1) = \frac{3}{8} \]

\[ P(x = 0) = \frac{1}{8} \]

1.2 Discrete and continuous random variables

Kastrati (2017), mentioned that the random variable X is called discrete if the set H is at most countable. Examples are: number of children in the family, number of students in a class, number of times we go to the beach, etc. Examples of discrete random variables are given above.

Definition: The random variable X is said to be continuous if a function \( f(x) \geq 0 \) (for every \( x \in \mathbb{R} \)) is found so that for every set \( B \subset \mathbb{R} \) we have:

\[ P(X \in B) = \int \! f(x) \, dx \]

The random variables X and Y are called independent if for \( \forall \, x, y \in \mathbb{R} \)

\[ P(X < x; Y < y) = P(X < x) \cdot P(Y < y) \]
Regarding the continuous variables we will take the following examples:

Example 2: It is known that the blood glucose level of diabetic persons follows a normal distribution pattern with an average of 106 mg / 100 ml and a standard deviation of 8 mg / 100 ml.

We calculate the probability of an ordinary person with diabetes having a glucose level of less than 120 mg / 100 ml.

Based on the above formulas we have: $P(X \leq 120) = 0.9599$.

What percentage of people have a glucose level between 90 and 120 mg / 100 mg?

$P(90 \leq X \leq 120) = 0.9372 \rightarrow 93.72\%$.

2. Categorized data

A verifiable number of authors have explained in two different forms the data categorized in areas such as: medicine, economics, sports and other fields. Change of content categories of a category according to a measurement scale (Agresti, 2003). The diagnosis of a disease which has been made with a sophisticated modern apparatus can give a clue as follows: normal, harmless, dangerous, probable, suspicious, critical, etc. (James, 2007). Developing methods for category data and promoting research study in the social and medical sciences, moreover scales of category variables used today in social science to measure opinions and attitudes, so the categorized scales used in today’s medicine to measure different outcomes and to indicate whether medical treatment is successful (Akturk, 2007; Lipsitz, 1994).

Although these provide a great opportunity for social and medical science, it does not mean that they are limited to these scientific areas, but they often do not even reach the behavioral
science (for example: types of mental illness, with this category as: schizophrenia, depression, nervousness) (Feldman, 2009), in epidemiology and public health (for example: contraceptive methods, pills, IUD, etc.) (Preisser, 1997), in zoology (for example: alligators as a preference for primary food, fish category, invertebrates, reptiles) (Maddison, 2014), in education (for example: student answers to exam questions, data category and errors) (Chen, 2003), in marketing (Magidson, 1982), (for example: consumer preferences for the main brand of a product, category A, B and C). In addition to the types of science these also contribute to numerous quantum fields are engineering science and industry control. Examples to be able to take some items and classify them in groups with reference to how much I meet certain standards, also could not be subjective assessment of some characteristics of rubber, for example: as much as possible and may have a costume life, how good and tasty a food product can be (Ai, 2002).

2.1 Nominal, sorted and interval variables

Variables that have taken values not according to an order that have become nominal (Agresti, 2003) and changes vary are: names, tags, categories, gender, etc. Examples are: religion (Orthodox, Catholic, Protestant, Jewish, Muslim, etc.), types of public transport (cars, bicycles, buses, subways, etc.), type of music (classical, country, folk, jazz, rock, etc.) and residence (apartment, condominium, private house, etc.). For nominal variables, statistical analysis does not depend on ranking (Djamel, 2005). Arithmetic operations cannot be performed on them. For example gender: male, female; there is no sequencing nor arithmetic operations. Variables that take values according to categories sorted into ordered changes (Freese, 2000).

They can give according to a criterion. Even to me these give one cannot perform arithmetic operations. Examples are school course (year I, II, III,..), Social class (upper, middle,
lower), political philosophy (liberal, moderate, conservative) and mood (good, serious, dangerous etc.). Relating to the authors above we can conclude some of the definitions by some self-examples. The sorted variables have sorted categories. For example an individual classified as moderate who is more liberal than an individual and classified as conservative has no numerical value does not determine how liberal an individual is. Methods for sorting variables use category sorting. For example, blood pressure level, life of TVs, length of time in prison, annual income, etc. The way a change is conveyed relates to its classification. For example: “education” is nominally changed when it is counted as a public or private school; it has been changed to rank when enabling the categories: nine-year, high school, bachelor, master and doctorate; it is a value variable in the interval when measured by the random school number, as a number 0, 1, 2. Statistical methods for nominal variables can help for listed variables while those for listed variables cannot help for nominal variables. It is better to use the right methods in real schools. The methods of use for interval variables have a small number of possible values (such as, measured schooling < 10 years, 10 - 12 years, > 12 years).

3. Evaluation of parameters in the data dissemination function

For generating random numbers of the rbinom () function we use the example retrieved by “Felix” (2016),\(^2\) the uses command generates 30 random numbers from a binominal distribution with p = 0.4 and N = 15. The simulated data set will change to require a simulation due to stochasticity. The simulation can be represented for example 30 students who have issued 15

\(^2\) Retrieved from: https://rpubs.com/felixmay/MLE

By: Felix May13, January (2016)
coins \((N = 15)\) and counts the number of coins that have a head \((p = 0.4)\) with the function table () we can calculate how often each number of successes \((p \text{ sh. coin showing head})\) has occurred.

**Programming on R:**

```r
sample_size <- 30
> p <- 0.4
> N <- 15
> dat1 <- rbinom(sample_size, prob = p, size = N)
> table(dat1)
```

**Results:**

```
dat1
4 5 6 7 8 9 10
4 5 7 3 6 2 3
```

Example discussion: In contrast to the `rbinom()` function, which generates random numbers, the `dbinom()` function provides security for \(x\) success having parameters \(p\) and \(N\). D on behalf of the function is derived from density, since for redundant distribution functions, example; normal distributions, this function is called the probability density function. What else for discrete distribution included binomial distribution this function is called the probability measure function. This means even if this terminology is not confusing the `dbinom()` function ensures the operation of the binomial distribution probability measure. First we do not visualize this function for a parameter specification set. Now we see that options should be needed up to 1 by definition. When we toss 15 coins we will see that I will live between zero and fifty heads, i.e. the probability of a result between zero and fifty is equal to one.

```r
> p <- 0.4
> N <- 15
> binom1 <- dbinom(x = 0:15, prob = p, size = N)
> sum(binom1)
```
Results:

The likelihood function is defined as the probability of the data given in the model. With `dbinom()` we get the probability for every single set, assuming a value of this parameter of p. In this case we do not know the true value of p because we have simulated "information". Finally, when we have to give real we do not have we do not know the true value of p. The N parameter is determined by the design of the experiment or sampling. The following function represents the value of the similarity function with values that we have represented:

```
dbinom(x = dat1, prob = 0.4, size = 15)
```

Results:

```
[1] 0.20659761 0.11805577 0.06121411 0.06121411 0.18593784 0.18593784 0.20659761
0.20659761 0.20659761 0.12677580 0.02448564 0.20659761 0.18593784 0.12677580
0.11805577 0.11805577
[17] 0.20659761 0.17708366 0.12677580 0.17708366 0.12677580 0.02448564 0.17708366
0.11805577 0.18593784 0.18593784 0.20659761 0.11805577 0.02448564 0.11805577
```

Above it is seen that this is a very small number. To avoid numerical problems the probabilities of the logarithm can be calculated, and which provides values that are numerically more controllable (i.e. negative values that are not so close to zero).

4. Conclusion

The paper has elaborated on the categories of information by making analyzes on the change of the case, respectively their categorization to come to the required result. The understanding of the random variations and the main distribution discussed above in Bionomial, Multinomial, cannot be reflected in the results we seek. In addition to the examples taken from everyday life and the choice of rubber in the classical way, in this topic the appearance and signs
of realization by means of the code in the language program R. The idea of working with this programming language is that it enables us to quickly find the truth of results we do not seek.

Examples are taken from everyday life that we have combined the ideas of different authorities to better understand what it does not require. To this analysis to categorize the data is added the logistic regression and what can be done to complete this topic, leaving us to understand even better the use of information that I need to analyze more delicately and have a whole presented with the results.

5. Bibliography

Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons,.

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. *Biometrics*, 117-128. doi:10.1111/j.0006-341x.1999.00117.x

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2013). *Statistika për biznes dhe ekonomik*. Tiranë.

Association, A. S. (2000). Generalized Linear Models. *Journal of the American Statistical Association*, 1320-1324.

Assumptions of Logistic Regression. (n.d.). *Statistics Solutions*.

Cameron, A. C. (1998). Regression Analysis of Count Data.

Cameron, A. C. (2009). Advances in Count Data Regression. *Talk for the Applied Statistics Éorkshop*. Retrieved from http://cameron.econ.ucdavis.edu/racd/count.html

Cameron, A. C. (2009). Microeconometrics Using Stata. *College Station*.

Chamilka. (2012). Maximum Likelihood Estimation Poisson distribution mle {stats4}. Retrieved from https://stat.ethz.ch/pipermail/r-help/2012-July/317433.html

Chen, G. &. (2003). Hoé to Deal éith Missing Categorical Data: Test of a Simple Bayesian Method. *Organizational Research Methods*, 309-327. doi:10.1177/1094428103254672

David W. Hosmer, J. S. (2013). *Applied Logistic Regression* (Vol. Third Edition). John Éiley & Sons,.

Davidson, A., & Hinkley, D. (1997). *Bootstrap Methods and their Application*. Neë York: Cambridge University Press.
Dupont, É. D. (2002). Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. Neē York: Cambridge Press.

Duygu Akturk, S. G. (2007). Multiple Correspondence Analysis Technique Used in Analyzing the Categorical Data in Social Sciences. Journal of Applied Sciences, 585-588. doi:10.1016/j.jas.2007.585.588

Feldman, B. J. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. Developmental Psychology, 652-676. doi:10.1037/a0014851

Freese, J. S. (2000). Listing and Interpreting Transformed Coefficients from Regression Models for Categorical and Limited Dependent Variables. Retrieved from https://jslsoc.sitehost.iu.edu/files_research/reprints/STBlistcoef.pdf

He, Z. X. (2002). An efficient algorithm for clustering categorical data. Comput. Sci. & Technol, 611-624. doi:10.1007/BF02948829

Heckman, J. &. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. Econometrica, 271. doi:10.2307/1911491

Hilbe, J. M. (2009). Models, Logistic Regression. Neē York.

James, N. T. (2007). Reliability procedures for categorical data in Performance Analysis. International Journal of Performance Analysis in Sport. doi:10.1080/24748668.2007.11868382

Jorgensen, B. (1997). The theory of dispersion models. London, UK: Chapman&Hall.

Kastrati, E. (2017). MODELIMI MATEMATIK I TË DHËNAVE KATEGORIKE DHE ZBATIMEVE. UNIVERSITETI POLITEKNIK I TIRANËS, 99. Retrieved from https://www.upt.al/images/stories/phd/MSc.%20Erjola%20KASTRATI%20(CENAJ).pdf

Lillis, D. (2017). Generalized Linear Models in R, Part 1: Calculating Predicted Probability in Binary Logistic Regression. Retrieved from https://www.theanalysisfactor.com/r-tutorial-glm1/

Lipsitz, S. R. (1994). Analysis of repeated categorical data using generalized estimating equations. Statistics in Medicine, 1149-1163. doi:10.1002/sim.4780131106

Maddison, W. P. (2014). The Unsolved Challenge to Phylogenetic Correlation Tests for Categorical Characters. Systematic Biology, 127-136. doi:10.1093/sysbio/syu070

Magidson, J. (1982). Some Common Pitfalls in Causal Analysis of Categorical Data. American Marketing Association. doi:10.1177%2F002224378201900408

Menard, S. (2002). APPLIED LOGISTIC REGRESSION ANALYSIS (Vol. 106). SAGE UNIVERSITY Paper.

MILLER, M. E., HUI, S. L., & TIERNEY, W. M. (1991). VALIDATION TECHNIQUES FOR LOGISTIC REGRESSION. STATISTICS IN MEDICINE.

Nettleton, D. (2017). A Generalized Linear Model for Binomial Response Data. Retrieved from https://dnett.github.io/S510/27GLMbinomialAnnotated.PDF
Pierce, R. (2020). "Random Variables". Math Is Fun. Retrieved from http://www.mathsisfun.com/data/random-variables.html

Pregibon, D. (1981). Logistic Regression diagnostics. The annals of statistics, 9, 705-724.

Preisser, J. S. (1997). CATEGORICAL DATA ANALYSIS IN PUBLIC HEALTH. Annual Review of Public Health, 51-82. doi:doi.org/10.1146/annurev.publhealth.18.1.51

R.Nuhiu, A. (1995). Bazat e Statistikës. Prishtinë.

Rodríguez, G. (2007). Logit Models for Binary Data. Retrieved from https://data.princeton.edu/ãs509/notes/c3.pdf

Rodriguez, G. (2007). Poisson Models for Count Data. 2-14.

Selvaratnam, S. Y. (2018). Maximum likelihood estimation of generalized linear models for adaptive designs: Applications and asymptotics. Biometrical Journal. doi:doi:10.1002/bimj.201800181

Taboga, M. (2017). Normal distribution - Maximum Likelihood Estimation", Lectures on probability theory and mathematical statistics. Third edition. Retrieved from https://www.statlect.com/fundamentals-of-statistics/Poisson-distribution-maximum-likelihood

(2015). The Method of Maximum Likelihood for Simple Linear Regression.

Treiman, D. J., Dhembo, E., Bejko, E., & Dizari, V. (n.d.). Analiza sasiore e te dhenve; Te kryesh kerkim social per te testuar ide.

Cameron, A. C. and Trivedi, P. K. 2009. Microeconometrics Using Stata. College Station, TX: Stata Press.

Cameron, A. C. and Trivedi, P. K. 1998. Regression Analysis of Count Data. New York: Cambridge Press.

Cameron, A. C. Advances in Count Data Regression Talk for the Applied Statistics Workshop, March 28, 2009. http://cameron.econ.ucdavis.edu/racd/count.html.

Dupont, W. D. 2002. Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. Neê York: Cambridge Press.

Long, J. S. 1997. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Long, J. S. and Freese, J. 2006. Regression Models for Categorical Dependent Variables Using Stata, Second Edition. College Station, TX: Stata Press.