Statistics and application of multi logistic model and negative binomial model in data

Lianjie Li[1, *, †], and Bolin Jing[2, †]

[1]Beijing National Day School, Beijing, China

[2]Changjun High School of Changsha City, Changsha, China

*Corresponding author: sniu5883@uni.sydney.edu.au

†These authors contributed equally.

Abstract. Contemporarily, statistics are widely adopted in various disciplines to verify the effectiveness of the analysis results scientifically. This article demonstrates two applications of the statistical. To be specifical, it mainly investigates negative binomial model and the way to judge the determinacy of factors with the help of Multi Logistic model. For the former, the relative superiority of random parameter binomial model is discussed. According to the data analysis of a traffic accident in a certain place, it shows a better model fitting effect. Another case of poultry waste recycling was used to introduce and compare the using conditions and advantages of Multi Logistic model. These results shed light on guiding further exploration of applications of statistics.

Keywords: Model Analysis; Logistic; Random; statistical applications.

1. Introduction

With the vigorous development of the automobile industry, the safety problems brought by road traffic come one after another. Contemporarily, the number of deaths caused by road traffic accidents has been rising all over the world, and traffic accidents have become a worldwide problem. According to data from the World Health Organization's worldwide status report on road safety 2018, 1.35 million people died in traffic accidents in 2016. It is the seventh most common cause of death across all age groups and the leading cause of death for children and young adults aged 5 to 29. Every year, a large number of people lose their lives or become disabled due to traffic accidents, and the direct economic losses arising therefrom account for 1%-3% of the gross national product of all countries. The purpose of this paper is to refer to the previous methods and improve them to study the traffic accident data of a certain area. In this case, it provide correspondings data analysis for the experiment (e.g., the pit on the road is an important factor leading to traffic accidents, so it is suggested that the traffic management department should pay special attention to filling the pit on the ground).

Further more, According to a 2018 survey by the Chinese government, with the growing number of rasorial poultry [1], many farmers have a higher demand for poultry waste recycling [2]. The government hopes to improve the waste disposal problem by improving the efficiency of poultry waste recycling. They sampled Shandong province because it is China's largest province for poultry production [3]. They want to find treatments that are currently most commonly used by farmers that can be optimized to make recycling more efficient [4]. Successful improvements in the largest farming areas could be extended to the rest of the country.

We hope to provide some model ideas and data processing schemes for environmental protection research through this article. In this paper, negative binomial model and Multi Logistic model will be introduced respectively, and the advantages and disadvantages of the model will be analyzed through a case study. At the end of the paper, we also introduce the available scope of these two models.

2. Introduction to negative binomial model

2.1 Statstic notes

The negative binomial distribution includes a series of independent experiments. In addition, each experiment has two results: success and failure. The experiment continues to R failures, and R can be any positive number. The probability density function is:

$$f(k;\ r,p) = \binom{k+r-1}{r-1} \cdot p^r \cdot (1-p)^k \tag{1}$$

Here, K is the number of parameters and l is the likelihood function. R is the number of trails. Akaike information criterion is another indicator, which can be calculated • :

$$AIC = 2p - 2\ln(L) \tag{2}$$

Here, P is the number of parameters in the model; L is the likelihood function. The smaller the AIC value, the better the model fitting effect. As for coeficient of ρ^2

$$\rho^2 = 1 - \frac{\ln(L)}{\ln(L_0)} \tag{3}$$

L_0 is the likelihood function value when only the constant term is included in the model. The larger the coefficient of ρ^2, the better the fitting effect of the model.

2.2 Data description:

Data is collected from 5 expressways in Heilongjiang Province and Liaoning Province (Suiman, hatong, Shenda, Shenshan and Shendan Expressway), with a total length of 2131km. Traffic accident data of five expressways from January 2014 to December 2018 were collected from the highway administration department. Since this study is aimed at the main line of the expressway, the accidents on the toll station and interchange ramp are excluded, and a total of 12097 accident data are obtained through screening. The annual traffic volume data of each section is obtained from the expressway flow observation station, including the annual average daily traffic (AADT) and truck traffic volume. The highway construction chart was obtained from the road design department, and the corresponding horizontal and vertical alignment and cross-sectional design data were obtained. The original data of Pavement Inspection over the years was collected from the expressway maintenance management center, including pavement damage rate, rut depth, international flatness index, lateral force coefficient and structural strength coefficient.

2.3 Model construction

Since the regression analysis method is selected, the most used Poisson distribution is used for analysis. In the Poisson model, the probability p (n_{it}) of n_{it} accidents occurring within time t is

$$p(n_{it}) = \frac{\lambda_{it}^{n_{it}}(-\lambda_{it})}{n_{it}!} \tag{4}$$

where M is the number of road sections; T is the data analysis period; λ It is the expected number of accidents within time t, often expressed as:

$$\lambda_{it} = exp(\beta X_{it}) \tag{5}$$

Here, X_{it} is the accident influencing factor vector, β is its parameter vector. However, Poisson's model requires the variance to be equal to the mean, because $n \cdot p_n \to \lambda;\ n \cdot p_n \cdot (1-p_n) \to \lambda \cdot (1-\frac{\lambda}{n}) \xrightarrow{n\to+\infty} \lambda$. Whereas, the accident data are generally too discrete, i.e., the variance is greater than the mean value, resulting in the poor fitting effect of Poisson model. For this reason, the random error term is introduced ε It. Hence, the mean value λ It is expressed as :

$$\lambda_{it} = exp(\beta X_{it} + \varepsilon_{it}) \tag{6}$$

Here, exp(ε_{it}) is subject to a mean of 1 and a variance of α and also a Gamma distribution. At this time, Poisson model can be expanded to negative binomial model:

$$P(n_{it}) = \frac{\Gamma\left[\left(\frac{1}{\alpha}\right)+n_{it}\right]}{\Gamma\left(\frac{1}{\alpha}\right)n_{it}!} \times \left(\frac{\frac{1}{\alpha}}{\left(\frac{1}{\alpha}\right)+\lambda_{it}}\right)^{\frac{1}{\alpha}} \left(\frac{\lambda_{it}}{\left(\frac{1}{\alpha}\right)+\lambda_{it}}\right)^{n_{it}} \tag{7}$$

where Γ(•) represent gamma distribution.

The parameters of the variables in the negative binomial model represent the impact of this factor on the accident risk. Therefore, in order to characterize the heterogeneity of the impact of various factors on traffic accidents, the parameters of any variable x_{it} in the model can be β, the fixed value is set as a random variable subject to normal distribution, and then the traditional fixed parameter negative binomial model is improved to a random parameter negative binomial model. At this time, the variable parameters can be expressed as:

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta} + \sigma^2 \pi_{it} \tag{8}$$

Here, β_{it} is the parameter vector of accident influencing factor x_{it} in section i and time t, and it follows a normal distribution that mean value is β, the variance is σ^2; π_{it} is a random term following normal distribution. If σ equal to 0, indicating that the parameter of xit is a fixed parameter, i.e., there is no spatiotemporal heterogeneity in the impact of this factor on the accident. However, the mean value of the random parameter distribution shown in formula (8) is fixed on each sample because they are all β. Due to the potential interaction of various factors on traffic accidents, the mean value of random parameter distribution may be affected by other factors. Therefore, the mean value of the random parameter distribution can be further set as the function form of other factors. In this case, the parameter β It can be further rewritten as:

$$\boldsymbol{\beta}_{it} = \beta + \boldsymbol{\delta M}_{it} + \sigma^2 \pi_{it} \tag{9}$$

Here, β follows a normal distribution that it's mean is $\beta + \delta M_{it}$ and variance are σ^2; M_{it} is the accident influencing factor vector affecting the mean value of $\boldsymbol{\beta}_{it}$, δ Is its coefficient vector; If δ not equal to 0, it indicates that the magnitude of the factor M_{it} has a significant effect to β. Its mean value and likelihood function are expressed below:

$$\lambda_{it} = \exp(\boldsymbol{\beta}_{it} X_{it} + \varepsilon_{it}) \tag{10}$$

$$\ln(L) = \sum\nolimits_{\forall it} \ln \int_{\pi_{it}} \phi(\pi_{it}) P(n_{it} \mid \pi_{it}) d_{\pi_{it}} \tag{11}$$

Considering that the sensitivity of the probability of traffic accidents is different from each factor, we need to do a sensitivity test[5]. By calculating the sensitivity of the number of accidents to each influencing factor. After analyzing the relative influence degree of each factor on the accident, the sensitivity of the number of accidents to continuous variables can be measured by the elastic coefficient E_k. The calculation method and fomulae are given as follows [6, 7]:

$$E_k = \frac{1}{T \times m} \sum\nolimits_{t=1}^{T} \sum\nolimits_{i=1}^{m} E_{X_{it}}^{\lambda_{it}} \tag{12}$$

$$E_{X_{it}}^{\lambda_{it}} = \frac{\partial \lambda_{it}}{\lambda_{it}} \times \frac{X_{it}}{\partial X_{it}} = \beta_{it} X_{it} \tag{13}$$

2.4 Results

From the result of AIC value [8] (the smaller the better) and ρ^2 coefficient [9] (the larger the better) indicates that the goodness of fit of the random parameter negative binomial model is higher than that of the traditional fixed parameter negative binomial model. Both the exposure variable AADT and the length of the road section are positively related to the number of accidents, and their parameters are 0.299 and 1.029, respectively (it can be inferred from equations (12) ~ (13) that the parameter of the exposure variable is its elastic coefficient), indicating that the number of accidents increases with the increase of traffic volume. The number of accidents is approximately linear with the length of the road section. The number of accidents will increase by 0.299% and 1.029% on average when the traffic volume and the length of the road section increase by 1%. The proportion of trucks is positively related to the number of accidents, and the number of accidents will increase by 0.093% for every 1% increase in the proportion of trucks. The results are summarized in Table. 1.

The potential reason is that with the increase in the proportion of trucks, the speed difference between vehicles will increase, and the behaviors endangering traffic safety (e.g., lane changing and overtaking) will increase, which will lead to more traffic accidents [10]. The number of accidents in one-way 3-lane and 4-lane road sections is 0.041 and 0.142 more than that in one-way 2-lane road sections, respectively. The potential reason is that the more lanes, the higher the frequency of vehicle lane change, which will lead to more traffic accidents. Curb width_ The parameter of 0.75 m follows a normal distribution with a mean value of -0.270 and a standard deviation of 0.057, that is, in most cases (> 99.99%), the road section with a curb of 0.75 m is safer than the road section with a curb of 0.5 m. The potential reason is that increasing the width of the curb can reduce the tension and anxiety of drivers in the inner lane to a certain extent, and also provide more lateral clearance for vehicles to deviate from the center line of the inner lane, which is conducive to improving the level of traffic safety.

Table 1. Estimation results for models （excluded non-significant variables）

variable name	Fixed parameter negative binomial model		Negative binomial model with random parameters	
	Parameter estimate		Standard error Parameter estimate	Standard error
Standard deviation of parameter distribution	0.618	0.19	0.683	0.162
logarithm of annual averaged aiy traffic	0.309	0.02	0.299	0.016
Logarithm of section length	0.992	0.021	1.029	0.019
Truck proportion	0.21	0.077	0.184	0.08
Number of lanes three	0.087	0.033	0.069	0.03
Number of lanes four	0.284	0.031	0.245	0.026
Curb width_0.75m	-0.299	0.04	-0.27	0.044
Curvature of horizontal curve	0.363	0.029	0.192	0.035
Longitudinal slope	0.078	0.011	0.068	0.01
Longitudinal slope direction_down hil slope	0.066	0.02	0.046	0.017
rut depth	0.021	0.004	0.019	0.004
Structural strength coefficient	-0.016	0.005	-0.02	0.005
number of samples	19730		19730	
AlC value	53422		53220	
coeficient of p^2	0.147		0.151	

The number of accidents is directly proportional to the gradient of the longitudinal slope, and the number of accidents in the downhill section is generally higher than that in the uphill section. The number of accidents will increase by 0.068% and 0.094% respectively for each 1% increase in the longitudinal gradient of the uphill and downhill sections. In addition, the longitudinal slope direction_ The downhill parameters follow normal distribution, and their mean value is positively correlated with the proportion of freight cars, that is, the downhill section with high proportion of freight cars has a high accident risk. The pavement damage rate is inversely proportional to the number of accidents. The elastic coefficient indicates that the number of accidents will decrease by 0.011% for every 1% increase. The potential reason is that the average pavement damage rate in the sample is only 0.06%, and the maximum damage rate is only 4.57%, and 99.8% of the sample pavement damage rate is less than 1%. The sensitive results are given in Table. 2.

Table 2. Sensitivities test result of crash for significant variables

variable name	elastic coefficient Ek	Marginal effect coefficient DL
AADT	0.299	
road length	1.029	
Truck proportion	0.093	
Number of lanes_ three		0.041
Number of lanes_ four		0.142
Curb width_ 0.75 m		-0.159
Curvature of horizontal curve	0.079	
Longitudinal slope	0.068	
Longitudinal slope direction_ downhill slope		0.026

Pavement damage rate	-0.011
rut depth	0.054
Structural strength coefficient	-0.064

If this small pavement damage rate is far from enough to cause the vehicle to lose control, on the contrary, the slight damage of the pavement on individual road sections will enhance the driver's vigilance, so that the accident risk is reduced to a certain extent. The parameters of rutting depth follow the normal distribution with the mean value of 0.019 and the standard deviation of 0.010, that is, in 97% of the cases, the increase of rutting depth will lead to more traffic accidents. The elastic coefficient indicates that every 1% increase in rutting depth will lead to an average increase of 0.054% in the number of accidents. The potential reason is that the deeper the rutting, the more difficult it is for the vehicle to maintain the predetermined track, especially when driving at high speed or changing lanes, Improper operation of the driver may cause large lateral deviation of the vehicle, which is not conducive to driving safety. However, in a few cases (3%) [11]. The rutting depth reduces the accident risk. The potential reason is that some drivers (especially those who are familiar with the road conditions) are vigilant, so they slow down in advance and drive carefully) [12], thus reducing the accident risk to a certain extent. The parameters of the structural strength coefficient follow the normal distribution with the mean value of -0.020 and the standard deviation of 0.044. In most cases, the higher the structural strength of the pavement is, the better the driving safety is, and the number of accidents will be reduced by 0.064% on average for every 1% increase. The potential reason is that the higher the structural strength, the stronger the bearing capacity of vehicles, especially heavy vehicles, so as to ensure the vehicles Stability.

In general, this paper has successfully studied) the relationship between factors, but at the same time, it is obvious that there are many areas that need to be improved. In order to analyze the data, the code efficiency is not very high, and the regression analysis method is relatively more suitable for experiments with large sample size [13]. As an improvement, more accurate research methods can be found next time. However, the research results are effective and can be used to improve the transportation sector [14].

3. Multi Logistic model

3.1 Descirption of the model

This model usually be used for estimating the correlation between different factors and the event. It can be used for determining if the factor will significantly influence or not to the event and if it is a positive effect or negative one. The model needs to collect the deterministic effect of each factor on multiple possibilities of an event. Therefore, in the process of collecting, researchers will encounter the problem of scoring subjective and objective factors. For subjective factors, the experimenters often let the participants evaluate the influence of the factors themselves. All objective factors will be assessed by the researcher's own custom rules to assess the impact of each factor on the outcome. If researchers want to use this model, the sample size of each factor must be larger than 30, and for each possible response factor, the sample must be larger than 10. Those data must be collected as random and the judgement for each factor’s value must be as same as possible. The researcher should collect a group which not use a specific response output, instead they use it as randomly. The model can be described as:

$$\ln\left(\frac{p_{t_i}}{p_0}\right) = \alpha_t + \sum_{k=1}^{22} \beta_{tk} x_{ki} \tag{14}$$

Here, the left-hand side is the dependent variable, which is the log of the probability of choosing a particular treatment, given the random treatment. P number i are respectively the probability of the i-th sample to give the response output by use t different categories. “t” is a number represent as randomly and other outputs, which their possibility sum is 1. X in the right side is called as independent variable, which is the factor. X_{ki} is the k-th factor which affects the i-th sample’s output of responsible. α are constant terms and β are the corresponding regression coefficients. It is adopted to determine whether this factor will influence this categories farmer's positively or negatively.

3.2 Data

The data example is given in Table. 3. On the left side of the table, there are 22 factors that we called it as X [15], and above the table are the possibility that a farmer uses one category out of randomly. For these four classifications of factors that totally these 12 factors are subjective factors. This part are 10 objective factors. The researchers put the X value and P value with their formula in the SPSS to get the coefficient β, the S.D. for each factor and the significance, symbolled by the stars here. So, this data table show us that which factor will influence farmer's specific choices more and is that a positive effect or negative one.

Table 3. SASS output for rasorial waste strategy

Empty Cell	**Direct Return/Discard Randomly**		**Compost Fermentation/Discard Randomly**		**Biogas Fermentation/Discard Randomly**		**Fresh-Packed Sale/Discard Randomly**	
Variable	**Coefficient**	**Std**	**Coefficient**	**Std**	**Coefficient**	**Std**	**Coefficient**	**Std**
rePE								
MPE	0.2	0.2	0.02	0.3	−0.3	0.4	0.1	0.3
SPE	0.4**	0.2	0.8***	0.2	1.2***	0.4	0.5*	0.25
EPE	−0.2	0.2	−0.24	0.25	0.4	0.4	−0.3	0.3
EE								
KEE	0.05	0.2	0.07	0.2	−0.15	0.3	0.2	0.2
LEE	0.1	0.2	0.15	0.2	−0.2	0.35	0.1	0.2
MEE	0.1	0.25	0.5*	0.3	0.4	0.3	0.3	0.3
SI								
SFI	0.7***	0.2	0.7***	0.2	1.0***	.03	0.65***	0.25
SNI	0.001	0.002	0.001**	0.001	0.001	0.001	−0.002	0.002
PRI	−0.05	0.2	−0.3	0.2	−0.04	0.296	−0.2	0.2
FC								
LFC	0.02	0.2	−0.1	0.2	−0.15	0.3	0.03	0.2
TFC	0.15	0.2	0.2	0.2	0.2	0.35	−0.15	0.2
EFC	−0.2	0.6	0.2	0.7	0.65*	1.1	0.4	0.85
IC								
SEX	-0.1	0.3	−0.8**	0.4	0.7	0.7	−0.4	0.4
AGE	0.01	0.02	0.03	0.02	−0.1*	0.04	−0.01	0.02
MRI	−0.6	0.7	−1.4*	0.9	−0.6	1.2	−0.4	0.9
EDU	0.02	0.1	−0.01	0.1	0.1	0.1	−0.02	0.07
ICM	0.4**	0.15	0.2	0.2	0.3	0.2	0.01	0.2
LC								

Empty Cell	**Direct Return/Discard Randomly**		**Compost Fermentation/Discard Randomly**		**Biogas Fermentation/Discard Randomly**		**Fresh-Packed Sale/Discard Randomly**	
Variable	**Coefficient**	**Std**	**Coefficient**	**Std**	**Coefficient**	**Std**	**Coefficient**	**Std**
LFI	0.5**	0.2	0.5*	0.3	0.7*	0.4	0.8***	0.3
LFP	0.1	0.15	0.15	0.2	0.6*	0.3	0.2	0.2
LFB	−0.3	0.6	−0.2	0.7	−0.5	0.8	−0.5	0.7
LFS	0.05	0.3	0.9**	0.4	0.4	0.5	0.3	0.4
LFQ	0.3*	0.2	0.3*	0.2	0.4*	0.2	0.5***	0.2
_cons	−3.0*	1.7	−6.4***	2.120	−10.0***	3.5	−3.9*	2.2
	−2 Log likelihood		1280.0 (Intercept Only)		Chi-Square		253.8 (p = 0.000)	
			1020.0 (Final)		pseudo R-Square		0.455	

Note: *, **, and *** refer to significance at the 10%, 5%, and 1% levels, respectively.

3.3 Explanition of the results

These data in total 462 samples [16] which are socred by farmer themselves [17] show that factors such as SFI, SPE [18] and LFI [19], as give in Table 3, will siginificantly influence farmer's decision on which strategy that they will use. Using SASS to find the β and α's value, the whole model could also successfully help the research learn weather the corelation between factors and each determinates is strong positive or weak negetive.

3.4 Limitations

The flaw in this model is that the accuracy of the results is not excellent. From the subjective factor score, everyone's standard is not the same, it is easy to appear answer error, i.e., participants can not accurately provide the real information, resulting in a large error in the experiment. Second, this approach cannot be applied to large scale factors and outcomes. The reason is that too many factors will lead to the score output value of each factor is very close, which makes it difficult to judge the real influencing factors.

4. Conclusion

In conclusion, this paper discusses two state-of-art applications of statics in other fields. According to the analysis, the Multi Logistic model can help to determine wheather each factor will siginificantly or slightly influence the output or not. It can also shown the postive or negetive relationship between factors and outputs. However, this model need huge sample size and still not accurate enough, but it can still be used in macroexamination for surveys or researches which only need a solving direction on the problem. Overall, these results offer a guideline for implemntation of statistical tools in cross-displinary.

References

[1] Zheng C. H., et ak. Managing manure from China's pigs and poultry: the influence of ecological rationality Ambio, 43 (5) (2014), pp. 661-672.

[2] Barr, S., 2003. Citizens and responsible environmental action are two sustainable development

[3] Gao Y., et al. A study about the factors that would influence the time when farmers adopted organic agriculture-- Take 325 organic vegetable farmers in Shandong Province as an example. J. Huazhong Agricculteral. University. (Soc. Sci. Ed.), 2016, 1, 56–63.

[4] Kassie M., et al. Adoption of interrelated and sustainable agricultural practices in the system of smallholder farmers planting people: evidence where comes from rural Tanzania Soc. Change, 2013, 80 (3), 525–540.

[5] Tang N S. Test for multiple upper and lower outliers in exponential samples. Athematical statistics and applied probability, 1998, 13(3): 28-38.

[6] Hong Zhu. Detection of multiple abnormal data in exponential distribution samples. Acta Electronica Sinica, 22(12), 1994, 95-98.

[7] Guo S, Fraser M W. Propensity Score Analysis: Statistical Methods and Applications, 2nd edition. Thousand Oaks: SAGE Publications, 2010, 186-195.

[8] Zhang Z. Propensity score method: a non-parametric technique to reduce model dependence. Annals of translational medicine, 2017, 5(1).

[9] Wood J, Donnell E T. Safety assessment of continuous green T-shaped intersection: propensity score genetic matching potential result method: Mathematical statistics and applied probability, 1998, 13(1): 63-70.

[10] Leite L, et al. A survey of DevOps concepts and challenges. ACM Computing Surveys (CSUR), 2019, 52(6): 1-35.

[11] Cafiso S. Et al. Collision correction function of pavement condition and geometric design index. Accident Analysis & Prevention，2021, 149: 105887.

[12] Dash A K. Analysis of accidents due to slope failure in Indian opencast coal mines. Current Science (00113891), 2019, 117(2).

[13] Meng Xianghai, Hou Qinzhong, Shi Yongyi, Jing Linpeng, Zheng Lai. IHSDM Expressway accident prediction model. Journal of Traffic and Transportation Engineering, 2016, 16(1): 123-132.

[14] Wang Lei, LV Pu, Lin Yongjie. Analysis of influencing factors and injury estimation of expressway traffic accidents. China safety science journal, 2016, 26(3): 5.

[15] M.H. Xing, J.B. Zhang, G.T. Huang., 2009. Environmental protection cognition of farmers' nonparticipation in circular agriculture and its influencing factors -- Based on the investigation of the rural economy in Shanxi and Hubei Provinces, 4 (2009), pp. 72-79.

[16] Considerations, farmers' method for using livestock and poultry wasteon a case study of 462 rural households in Hunan. J. Hunan Agric. Univ. (Soc. Sci. Ed.) 2009, 18 (4), 37–43 (in Chinese).

[17] Wen, K.L., F.M. Zhou, and M.R. Bin. Research on adoption's influence and openness, 2019, preprint.

[18] Mueller W. The effectiveness of recycling policy options: waste diversion or just diversions?. Waste management, 2013, 33(3): 508-518.,

[19] Pan D., Kong F.B. Analysis on the way farmers treat livestock's poultry wastes as resources and the influencing factors of the way Chinas Rural area Econ 2015, 9, 17–29 (in Chinese).