The Statistical Analysis of HR Employee Retention, Salary Variation of Remote Work and Earthquake Occurrence Probability

Enxi Jiang[1, *, †], Jingyi Zheng[2, †] and Jingyuan Zhu[3, †]

[1]Department of Statistic, University of Toronto Mississauga campus, Toronto, Canada

[2]Yancheng No.1 High School CID, Yancheng, China

[3]Department of Statistic, Southern University of Science and Technology, Shenzhen, China

*Corresponding author: enxi.jiang@mail.utoronto.ca

†These authors contributed equally.

Abstract. As the fast development of statistical techniques, more and more statistical methods are widely used in different area. In this paper, we will conduct an analysis on employee retention problem, remote work salary problem, and the earthquake forecast problem based on statistical approaches. For the employee retention problem, we apply logistic regression to the HR employee retention data and successfully construct a model, which can predict each employees' probability of leaving the company. As for the remote work salary problem, we apply multivariable linear regression to find out the impact of remote work on salary. According to the analysis, the type of remote work may have an impact on the salary. Compared to participants working on site, participants with 50% remote ratio had the mean of 17298.72 lower salary ($P < 0.05$). For the earthquake forecast problem, we construct a Poisson model based on the previous seismic data from Chinese mainland. Then we calculate the probabilities of earthquake occurrence and predicts the amount of earthquake in certain time interval and then verifies. These results shed light on guiding further exploration of statistical techniques including logistic regression, multivariable linear regression and Poisson model.

Keywords: Logistic regression; Multivariable linear regression; Poisson model.

1. Introduction

Logistic regression model is widely used for classification problem. For instance, Greenhouse use logistic regression model to predict the results of language recovery after having a stroke [1]. Besides, Rose used logistic regression to find out the prospective predictors of smoking cessation [2]. A multivariate regression is a method of determining whether the dependent variables and the independent variables are linearly related. Poisson model is a discrete probability distribution which suitable to describe rare events. Because of this quality of Poisson model, Clarke used it to test whether he points of impact of the bombs tended to be grouped in clusters [3]. Employee retention is becoming increasingly challenging for employers these years. As Hale stated, attracting new employees is problematic for 86% of the employers and 58% of companies claim that they have difficulties keeping employees around [4]. As for the consequence of the employees' turnover, Fitz contended that a company would lose approximately $1 million on average if every 10 high-level employees had left the company. Considering both direct and indirect costs, the company's spending on each employee's turnover is at least equal to benefits and pay of one year [5]. Therefore, it is important for the employers to know why employees are leaving and how to retain them. There are many factors which might be correlated to the employees' turnover such as salary, working intensity, work accidents and so on. To figure out exactly how these factors affect the probability of employees' leaving, we need to construct logistic regression models to summarize the effect of these factors on the employees' turnover.

In addition to examining the impact of different factors on employee turnover, the statistical model can also assess the relationship between remote work types and wages. since the workforce was quarantined during the COVID-19 pandemic, a unique situation, remote work was universally applied in most companies [6]. However, there is debate regarding the type of remote work could have an influence on the salary [7].

In the conventional sense, geography has a significant role in defining compensation ranges since labor costs vary greatly from place to place. Nevertheless, with remote employment, workers are frequently based distant from a company's office, or, in the case of a totally remote business, there is no office location [8]. Susequently, there are the employees that commute farther than usual to work in areas with a high cost of living while residing in more affordable areas. As an alternative, some people could rent a tiny apartment close to their place of employment while keeping the family home in a less costly area. Companies typically base wage decisions on the compensation received by individuals with comparable qualifications and experience in the company's local region, as revealed through salary surveys [9]. The study hypothesis is that the wage range should not change depending on whether an employee works from home or in an office. Therefore, the study aims to evaluate the relationship between the type of remote work and salary.

Besides, statistical model can also be used to predict earthquakes. Owing to the large geographical area of China, earthquake events occur on a regular basis. Because of the long formation time and complex formation process, also, there are so many interference sources of earthquake observation and they are difficult to be identified. Moreover, the prediction efficiency of different earthquake prediction methods has large difference. As a result, it is more scientific and reasonable to express the uncertainty of earthquake prediction opinions in the form of probability. [10] To predict earthquake occurrence, we choose Poisson distribution, which defined as a discrete frequency distribution. The Poisson distribution is a mathematical relationship. One application of it is prediction of certain phenomena from basic data [11].

In this paper, we have three primary purposes. First, we aim to help the employers predict which employees are more likely to leave so that the employers can take measures in advance. Second, the study aims to evaluate the relationship between the type of remote work and salary. Third, we predict the possible number of earthquakes occurring at a specific time. The rest of this paper organizes as follows. In Section 2, we introduce a theoretical analysis of the logistic regression model and use the logistic regression model to predict the employees' probability of turnover. Then we evaluate the effectiveness of the model. Section 3 used a multivariate linear model to predict the employees' salaries. In Section 4, we use the formula based on Poisson Distribution to calculate the earthquake occurrence probability and analyze if it is an effective method in seismic prediction. In Section 5, we make a summary of these applications.

2. Analysis of HR Employee Retention Data

2.1 Logistic Regresion Model

Logistic regression is a machine learning classification algorithm which is suitable for solving problems with two possible outcomes. It is a kind of extension of linear regression model. The main character of logistic regression is that it squeezes the output of a linear equation into a probability whose value is between 0 and 1 using logistic function. While the logistic function is defined as:

$$logit = \frac{1}{1+e^{x}} \tag{1}$$

Logistic function can force the output to be values between 0 and 1, which represents the probabilities. If we combine the logistic function with the linear regression model, one derives the formula:

$$P\left(y^{(i)} = 1\right) = \frac{1}{1+\exp\left(-\left(\beta_0+\beta_1 x_1^{(i)}+\cdots+\beta_p x_p^{(i)}\right)\right)} \tag{2}$$

Then we can construct logistic regression model based on this formula.

2.2 Data Processing

We collect the HR Employee Retention Data on Kaggle [12]. It contains 14999 employers' information including their salary levels, satisfaction levels, scores of last evaluations, average working hours each month, number of years spent in the company, number of work accidents in the company, whether they got promotion last 5 years and whether they have left the company. In order to predict which employee are more likely to leave, we take "whether they have left the company" as the response variable y and other information about the employees as independent variable $x^{(i)}$ to construct a logistic regression model. The results are listed in Table. 1.

Table 1. Results of logistic regression

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-1.4576948	0.1767825	-8.246	< 2e-16
Satisfaction level	-4.1336685	0.1011619	-40.862	< 2e-16
Last evaluation	0.6993192	0.1537692	40.862	5.42e-06
Number of project	-0.3119283	0.0219397	-14.218	< 2e-16
Average monthly hours	0.0043613	0.0005323	8.193	2.54e-16
Time spent company	0.2520152	0.0158690	15.881	< 2e-16
Work accident	-1.5020033	0.0917261	-16.375	< 2e-16
Promotion last 5 years	-1.4613510	0.2635496	-5.545	2.94e-08
Salary low	1.9636434	0.1308313	15.009	< 2e-16
Salary medium	1.4183997	0.1317296	10.768	< 2e-16

2.3 Evaluation of the model

In order to evaluate the effectiveness of the model, we conduct a test to see how well the model can predict whether the employees are going to leave. We randomly choose 13999 samples from the 14999 employees to construct the logistic regression model and use the model to predict the other 1000 samples on their probabilities of leaving. As shown in Fig. 1, one obtains the result of 1000 predicted probabilities.

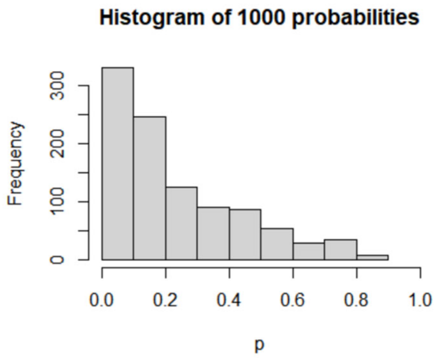

Fig. 1 The histogram of predicted probabilities

Then, we define 0.5 as the threshold of prediction. In this case, if the probability of turnover is larger than 0.5, we consider the employee is going to leave; if the probability of turnover is smaller than 0.5, we consider the employee is not going to leave. Finally, we check one by one that whether the predictions are correct. The result shows: 785 out of 1000 predictions got the correct answers. Therefore, we can conclude that the accuracy rate of the model is about 78.5%.

3. the Impact of Remote Work on Salary Data

As for the second case, the participants were evaluated based on the mean ± SD for continuous variables and the percentage for categorical variables. For descriptive analysis, the remote ratio was classified as 0, 50, and 100. Differences among three groups were analyzed using *t* tests for continuous variables and χ^2 tests for categorical variables. Multivariable linear regression was applied to investigate remote work and its association with salary. Three models were applied to reduce the confounding relationships. Model 1 was remote ratio. Model 2 was further adjusted for experience level, job title, employment type, work year, and employee residence. Model 3 was additionally adjusted for company size and company location. The statistical analyses were performed using R version 4.0.3. All *P* values were two-sided, and statistical significance was defined as $p < 0.05$. Table 2 presented the basic characteristics of participants by quantile of salary in USD. Among 607 participants included in the study, subjects with 100% remote work tended to have more salaries, senior levels, and full-time jobs ($P<0.001$). Table 3 shows the associations between remote work and salary. After adjusting multiple variables in model 3, compared to participants working on site, participants with a 50% remote ratio had a mean of 17298.72 lower salaries ($P < 0.05$). However, no statistical significance was found among participants with a 100% remote ratio.

Table 2. Baseline characteristics of 607 participants by remote ratio

Characteristics	Remote ratio			*P*
	0 (n=127)	**50 (n=99)**	**100 (n=381)**	
salary_in_usd (mean (SD))	106354.62 (67724.54)	80823.03 (57353.09)	122457.45 (72710.84)	<0.001
work_year (%)				<0.001
2020	15 (11.8)	21 (21.2)	36 (9.4)	
2021	34 (26.8)	66 (66.7)	117 (30.7)	
2022	78 (61.4)	12 (12.1)	228 (59.8)	
experience_level (%)				<0.001
EN	14 (11.0)	25 (25.3)	49 (12.9)	
EX	3 (2.4)	5 (5.1)	18 (4.7)	
MI	56 (44.1)	42 (42.4)	115 (30.2)	
SE	54 (42.5)	27 (27.3)	199 (52.2)	
employment_type (%)				0.006
CT	0 (0.0)	0 (0.0)	5 (1.3)	
FL	1 (0.8)	1 (1.0)	2 (0.5)	
FT	126 (99.2)	92 (92.9)	370 (97.1)	
PT	0 (0.0)	6 (6.1)	4 (1.0)	
company_size (%)				<0.001
L	32 (25.2)	60 (60.6)	106 (27.8)	
M	79 (62.2)	21 (21.2)	226 (59.3)	
S	16 (12.6)	18 (18.2)	49 (12.9)	

Table 3. Associations between remote work and salary

Remote ratio	β (SD)		
	Model 1	Model 2	Model 3
0	1.00 (ref.)	1.00 (ref.)	1.00 (ref.)
50	-25532 (9305) **	-12587.8 (7588.0)	-17298.72 (7848.62) *
100	16103 (7111) *	210.0 (5141.8)	2.43 (5216.60)

Model 1: adjusted for remote ratio.

Model 2: model 1+ experience level, job title, and employment type, work year, and employee residence.

Model 3: model 2+ company size and company location

*P < 0.05, **P < 0.01

4. Earthquake Occurrence Estimation & Precision

4.1 Calculating Previous Probability Based on Poisson Distribution

The Poisson distribution is a limiting case of a Binomial distribution. when the number of trials, n, gets very large, and p, the probability of success, is small, the Poisson distribution can provide a very good approximation to the binomial distribution, and at that time $\lambda = np$. That's why it is often used to describe the distribution of the amount of events with low probability. The occurrence of seismic events is rare enough, it satisfies the precondition for this model [13]. Therefore, the Poisson distribution is chosen to predict earthquake occurrence. It is assumed that the rate of earthquake occurrence in time is governed by the Poisson Distribution. Thus, the probability of some earthquakes occurring for k times in the time interval since the last earthquake have occurred is provided that [14]. The formula is given as:

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad k = 0,1,\dots \tag{3}$$

Here, k is the number of earthquakes, which has happened; λ is the average incidence of random events per unit time (or unit area). One sets the time interval since the last time an earthquake occurring at this region as T_e, and $\overline{T}$ is median of time interval of earthquake recurrence at the region. For seismic events, λ is $\frac{T_e}{\overline{T}}$. The probability of no earthquake occurring in the time interval (k = 0) is given by

$$P(k = 0) = \frac{\lambda^0}{0!}e^{-\lambda} = e^{-\lambda} = e^{-\frac{T_e}{\overline{T}}} \tag{4}$$

The event that there are some earthquakes occurring in the time intervalT_e(k≥1) is the complementary event of the event that no earthquake occurs (k = 0). Therefore, the probability of some earthquakes occurring in the time interval T_e(k≥1) is

$$P(k \geq 1) = 1 - P(k = 0) = 1 - e^{-\frac{T_e}{\overline{T}}} \tag{5}$$

Eq. (4) is a cumulative distribution function given by

$$\int_0^{T_e} f(T)dT = 1 - e^{-\frac{T_e}{\overline{T}}} \tag{6}$$

If there is no earthquake which is greater than certain magnitude from time of the last earthquake happening to T_e, the probability of an earthquake occurring between T_e and $T_e+\Delta T$ is

$$P(T_e \leq T \leq T_e + \Delta T |\, T > T_e) = \frac{\int_0^{T_e+\Delta T} f(T)dT - \int_0^{T_e} f(T)dT}{1 - \int_0^{T_e} f(T)dT} \tag{7}$$

Substituting Eq. (5) into Eq. (7), one derives

$$P(T_e \leq T \leq T_e + \Delta T |\, T > T_e) = 1 - e^{-\frac{\Delta T}{\overline{T}}} \tag{8}$$

On this basis, it is able to calculated the background earthquake probability of the region within a certain time range.

4.2 Data Selecting, Analyzing & Calculating

The data used are from the observation datum of China Earthquake Networks Center. All the seismic data of all the earthquakes whose magnitude is above 5.5 in Chinese mainland since 1970 are chosen, and the aftershocks and earthquake groups were excluded from the selected earthquake list. Based on the selected earthquake catalogue, the background earthquake probability of the region can be calculated using the method given above. According to the calculated background earthquake probability and 2018 earthquake data, the amount of earthquake occurrence for the next three months has been predicted monthly since February 2018 based on the calculation method above. Then, we compare the data predicted and the actual data obtained at that time to verify if the prediction is enough accurate.

4.3 Results

The prediction results of the Chinese mainland earthquake occurrence probability in the next 3 months produced from February to September 2018 was selected for verify and analyze. In 8 times of predicting earthquake occurrence probability, the largest hit rate was 75% in August while the least one was 0% in February since no earthquake occurred between February and April in 2018. Generally, as shown in Fig. 2 and Table. 4, average precision of prediction of earthquakes whose magnitude are greater than 5.5 is 72.5%, it indicates this method is effective in earthquake prediction practice.

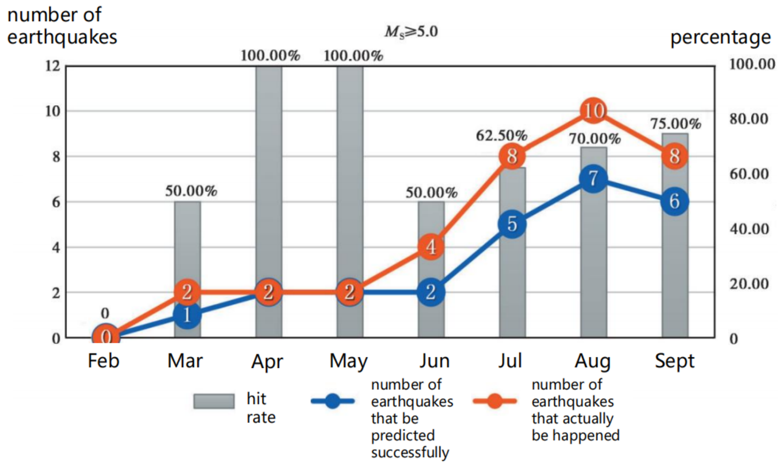

Fig. 2 Test and statistics of seismic hazard probability prediction results in Chinese mainland in every three months of 2018.

Table 4. Test of seismic hazard probability in Chinese mainland in every three months of 2018.

S/N	time	hit rate of prediction
1	Jan-Apr	0/0
2	Mar-May	1/2
3	Apr-Jun	2/2
4	May-Jul	2/2
5	Jun-Aug	2/4
6	Jul-Sept	5/8
7	Aug-Oct	7/10
8	Sept-Dec	6/8
	Average precision	72.5%

5. Conclusion

In summary, this paper discusses applications of statistical models. To be specific, HR employee retention, salary variation of remote work and earthquake occurrence probability are discussed. For the first case, we make an analysis of HR employee retention based on logistic regression model. Aiming at helping the employers to predict which employees are more likely to leave, we fit a logistic regression model to the data. Then, we evaluate the model by a random test. The result shows that the model has an accuracy rate of about 78.5%. Based on the analysis, logistic regression model is a great model to predict binary variable problems. In addition, as for the second topic, for participants with 50% remote ratio, it has lower salaries compared with worker on site. However, there is no statistical significance among participants with 100% remote ratio. Future studies are warranted to elucidate the reasons behind the findings. Regarding to the last topic, the result shows that, it is an effective method for seismic hazard prediction. In building statistical models of seismic activity over time, because of its simplicity of model formulation, and the small number of parameters to be estimated, the Poisson model can be widely used in seismic risk assessment and long-term prediction of seismic activity. Nevertheless, this model is only an ordinary statistical model, without any physical meaning. Therefore, with Poisson distribution, there must be some deficiencies to describe the seismicity in some areas. Overall, these results offer a guideline for the application of logistic regression, multivariable linear regression and Poisson model.

References

[1] Greenhouse JB, Bromberg JA, Fromm D. An introduction to logistic regression with an application to the analysis of language recovery following a stroke. J Commun Disord. 1995 Sep, 28(3): 229-46.

[2] Rose J. S., Chassin L., Presson C. C., Sherman S. J. rospective predictors of smoking cessation: A logistic regression application. Multivariate applications in substance use research: New methods for new questions, 2000, pp. 289–319.

[3] Clarke R D. An application of the Poisson distribution. Journal of the Institute of Actuaries, 1946, 72(3): 481-481.

[4] Hale J. Strategic Rewards: Keeping your best talent from walking out the door. Compensation & Benefits Management, 1998, 14(3), 39-50.

[5] Ma Kunlong. Short term distributed load forecasting method based on big data. Changsha: Hunan University, 2014.

[6] Daniels R A, et al. One Size Does NOT Fit All: Understanding Differences in Perceived Organizational Support during the COVID-19 Pandemic. Business and Society Review, 2022, vol. 127, no. S1, pp. 193–222. Wiley Online Library.

[7] Menger Pierre-Michel. Artistic Labor Markets and Careers. Annual Review of Sociology, 1999, vol. 25, pp. 541–74.

[8] Olson G M, Judith S. O. ‘Distance Matters’. Human–Computer Interaction, Sept. 2000, vol. 15, no. 2–3, pp. 139–78.

[9] Carpenter M A, et al. Bundling Human Capital with Organizational Context: The Impact of International Assignment Experience on Multinational Firm Performance and CEO Pay’. Academy of Management Journal, June 2001, vol. 44, no. 3, pp. 493–511.

[10] Deng S G, Zhou L Q, Ma Y W, et al. Research on probability prediction of earthquake risk based on Bayesian theorem. Earthquake Research in China, 2019, 35(1): 1- 13

[11] Gerlough D L, Schuhl A. Use of Poisson distribution in highway traffic. Saugatuck, CT: Eno Foundation for Highway Traffic Control, 1955.

[12] GUMMULA SRIKANTH, 2017, Retrieved from: https://www.kaggle.com/datasets/gummulasrikanth/hr-employee-retention

[13] Wang J P, Huang D, Chang S C, et al. New evidence and perspective to the Poisson process and earthquake temporal distribution from 55,000 events around Taiwan since 1900. Natural Hazards Review, 2014, 15(1): 38-47.

[14] Wen X Z. Segmented seismic hazard assessment of time-dependent active faults and its problems. Chinese Science Bulletin, 1998, 43(14): 1457- 1466.