Analysis of Relationship Between Personality and Favorite Places with Poisson Regression, ZINB Regression, and Quantile Regression

HA YOON SONG, HWA BAEK KANG
Department of Computer Engineering,
Hongik University
Sangsu, Mapo, Seoul
REPUBLIC OF KOREA
hayoon@hongik.ac.kr, khb12300000@gmail.com

Abstract Relationships between human personality and preferred locations have been a long conjecture for human mobility research. In this paper, we analyzed the relationship between personality and visiting place with Poisson Regression, Zero Inflated Negative Binomial regression, and Quantile regression. Poisson Regression can analyze correlation between countable dependent variable and independent variable. For this analysis, 33 volunteers provided their personality data and 49 location categories data are used. Raw location data is preprocessed to normalize into rates of visit, and outlier data is calibrated. For the regression analysis, independent variables are personality data and dependent variables are preprocessed location data. Several meaningful results are found. For example, persons with high tendency of frequent visiting to university laboratory has personality with high conscientiousness and low openness. As well, other meaningful location categories are presented in this paper. Zero Inflated Negative Binomial Regression is a usually good method for data with many zero values. As well, data are divided into quantiles and Quantile regression is applied. These three results are compared in order to verify the result of Poisson regression.

Key-Words: - Poisson, Personality, Location, ZINB, Quantile, Dependent, Independent

1 Introduction

Location Based Service (LBS) is one of the emerging topic with wide possibility of future service. Especially, understanding human mobility pattern is one of the core part of LBS.

In addition, it is widely recognized that human personality may affect personal favorite location. The relationship between human personality and human preferred location is valuable to understand.

In this paper, we are going to analyze the relationship between human personality and human favorite location by regression analysis. The personality data will be independent variable and the location data will be treated as dependent variable.

Linear regression is a general tool to analyze the correlation between independent variable and dependent variable. However, normal linear regression model such as stepwise regression cannot be a general tool in our case.

In our paper, the independent variable is from Big Five Factor (BFF) of human personality and the dependent variable is from the count of visit to favorable location categories. Since linear regression model is not an adequate to analyze such count data, we will use Poisson regression [1] to analyze data set containing count data.

Poisson regression model is based on Poisson distribution. Poisson distribution is a probability distribution regarding the count of occurrences in a given unit of time or area which probability mass function is shown as equation (1).

$$p(y) = \lim_{n \to \infty} \begin{bmatrix} n \\ y \end{bmatrix} p^y (1-p)^{n-y} = \frac{\lambda^y}{y!} e^{-\lambda} \quad (1)$$

Where y is a variable standing the number of occurrences, n is total number of events, and λ is expectation of probability variable y. Poisson distribution is a convergence of binomial distribution with n goes to infinity.

One of the properties in our data is zero inflated. For the verification of Poisson regression, we introduced Zero Inflated Negative Binomial (ZINB) Regression Analysis [2]. In addition, data are classified into quantiles so that Quantile regression analysis [3] also applied to verify the results from Poisson Regression Analysis.

The structure of this paper is like the followings. Section 2 discuss about the related researches. In section 3, we will discuss about the details of data used in this paper. In section 4, the pretreatment of data for Poisson regression will be discussed. Section 5 will present the result of Poisson regression and analyze the result. In section 6, Zero Inflated negative binomial (ZINB) Regression and Quantile Regression are applied and compared with the results of Poisson Regression. In section 7, we will conclude this paper and discuss about possible future research topics.

2 Previous Studies

Recently, positioning data can be collected by many of commercial devices including smartphones. Thus, the mobility data can be easily collected and these data can be used for various research purpose. For example, user can check in their favorite places voluntarily.

The question arise that human personality and human favorite locations are related. Since these data are personal data and privacy protection required, rare previous works and be found.

I. Burbey, [1] tried to predict next location based on past movement pattern. S. Y. Kim and H. Y. Song [2] used Back Propagation Network (BPN) to analyze mobility data. P. T.

Costa and R. R. McCrae [3] categorize personality into five element: openness, conscientiousness, extraversion, agreeableness, neuroticism. It is possible to enumerate each personality elements so that the corresponding numbers can be directly utilized in our research.

The most related work to this study can be found in H. Y. Song and E. Y. Lee [4], which used Stepwise regression to analyze the relationship between personality and visiting place. Stepwise regression is repetitive iteration of regression analysis, and is known to find the relationship between dependent variable and independent variable. However, stepwise

regression cannot analyze relationship between countable data while the frequency to visit a location is clearly a count data. Therefore, we are going to use Poisson regression to analyze the relationship in this paper.

3 Personality and Location Data

3.1 Personality Data

For personality data in our research, FFM (Five Factor Model) is utilized. Big Five Factor (BFF) of a person can be obtained by survey using Big Five Inventory (BFI) and 33 volunteers provided their BFF. BFF is composed of five personality factor openness, conscientiousness, extraversion, agreeableness and neuroticism L. R. Goldberg, [5].

The advantage of BFF is that the personality factors are presented in numerical form between 0 to 5 and these values can be independent variables of regression analysis.

Table 1. Personality Data

BFF / Person	O	C	E	A	N
person1	3.3	3.9	3.3	3.7	2.6
person2	2.7	3.2	3.3	2.7	2.8
person3	4.3	3.1	2.3	3.2	2.9
person4	3.6	3.3	2.8	3.2	2.8
person5	4.2	4.3	3.5	3.6	2.6
person6	4.0	3.7	4.0	3.9	2.8
person7	3.5	3.8	3.4	3.2	3.0
person8	2.2	3.4	3.0	3.1	2.6
person9	2.6	2.8	3.4	3.1	2.6
person10	3.3	2.9	3.1	3.1	3.3
person11	3.4	3.2	3.4	3.3	3.1
person12	3.1	3.7	3.4	3.2	3.5
person13	3.4	3.6	3.6	2.9	2.5

·O: Openness

·C: Conscientiousness

·E: Extraversion

·A: Agreeableness

·N: Neuroticism

Table 1 shows part of BFF of total 33 volunteers BFI survey. BFFs of thirteen volunteers are presented. Therefore, direct comparison of personality can be enabled in quantitative manner. For example, person1 has higher Openness than person2 and has lower Neuroticism than person2.

3.2 Location Data

Location data can be collected by checking-in user favorite places. For such purpose, smartphone app called SWARM [6] is used by volunteers. Table 2 shows parts of location data collected. It contains place names, place categories, and the count of visit to the categories for person2.

The topmost count shown in Table 2 is 69 at university laboratory by person2. The university laboratory is categorized as university building. The count to a category is a sum of count to places belonging to the category.

Table 2. Location Data

Count	Category	Place name
69	University building	University laboratory
36	Korean restaurant	Seorae Babsang
16	Pastfood restaurant	Burger king
11	Book store	Hongik book store
4	Theater	CGV hongik

For example, student restaurant, university library, and so on belong to category of university building and the counts of student restaurant, university library, university laboratory is summarized into the count of university building category.

Once the category count of a specific place is less than or equal to five, the place is categorized into etc. category. It is because to avoid the effect of minute count to regression analysis. In sun, we have 49 categories used in this paper.

4 Data Preprocessing

The counts of location categories are utilized as dependent variable of regression analysis. However, several preprocessing stages required in advance to the regression analysis. The counts need to be pruned.   The raw count may effect to regression analysis negatively with biases. Therefore, too high count and too low count must be excluded before regression analysis. Out of 36 data set of volunteers, two set of data are excluded because of their too high count and one set of data is excluded because of its too low count.

In addition, the counts need to be normalized. Therefore, we normalized the raw counts in t rates by equation (2).

$$Rate = \frac{visiting\ count\ by\ category}{total\ visiting\ count} \quad (2)$$

The normalized data can be used as dependent variable.

5 Poisson Regression Analysis

Poisson regression is the major tool we used for the analysis of relation between location data and personality data.
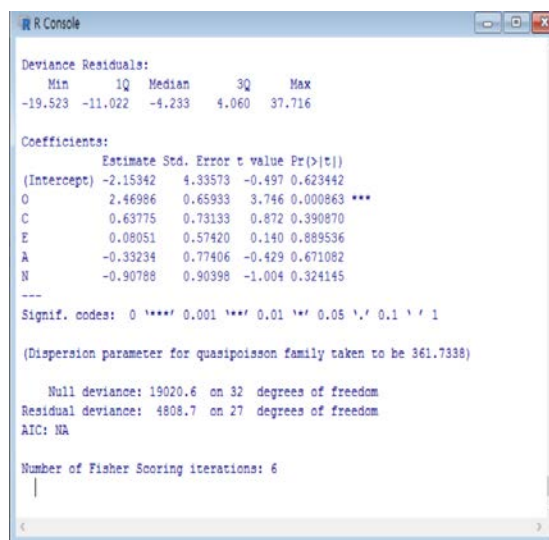
Fig 1. Poisson Regression Example

Calibrated count of location categories in Table 4 is used as dependent variable, and personality data in Table 1 is used as independent variable. The Poisson regression library incorporated in R studio is used. Figure 1 shows a sample analysis results executed in R studio environment. Poisson regression analysis is done for the location category 'theater' and personality data.

The main result in Fig 1 shows the effect of five independent variables, O, C, E, A, and N

Table 4. Raw Location Data

	airport	church	Train station	Outdoor Activity	library	University laboratory	Bus station	Shopping mall	Subway station	home
person1	23	5	15	72	422	107	6	11	7	375
person2	17	3	1	23	306	23	0	201	1	341
person3	7	4	2	2	84	11	0	5	2	79
person4	11	0	2	20	515	13	0	1	0	516
person5	5	2	0	9	270	10	1	4	0	169
person6	5	0	1	3	257	33	4	6	4	121
person7	5	38	0	4	199	46	11	2	39	145

Table 5. Pruned and Normalized Location Data

	airport	church	Train station	Outdoor Activity	Library	University laboratory	Bus station	Shopping mall	Subway station	home
person1	0.014715	0.003199	0.009597	72	0.269994	0.068458	0.003839	0.007038	0.004479	0.239923
person2	0.013127	0.002317	0.000772	23	0.236293	0.017761	0	0.155212	0.000772	0.26332
person3	0.014737	0.008421	0.004211	2	0.176842	0.023158	0	0.010526	0.004211	0.166316
person4	0.005624	0	0.001022	20	0.263292	0.006646	0	0.000511	0	0.263804
person5	0.007163	0.002865	0	9	0.386819	0.014327	0.001433	0.005731	0	0.24212
person6	0.006297	0	0.001259	3	0.323678	0.041562	0.005038	0.007557	0.005038	0.152393
person7	0.007418	0.05638	0	4	0.295252	0.068249	0.01632	0.002967	0.057864	0.215134
person8	0.005814	0	0.004651	10	0.196512	0.001163	0.002326	0.016279	0	0.276744

toward the location category theater. First, results with p-value less than 0.05 must be observed. This is because that, statistically, result of regression analysis with p-value judged to be meaningful. In this example, the visit to theater and factor O (Openness) is related. The values in Estimate column is quantitative effect of independent variable to dependent variable.

The estimate of O in Figure 1 is 2.46986, which leads to analysis result that a person with high Openness is tent to visit theater frequently. On the contrary, once we have negative values on Estimate for the effect of one of BFF to a location category, it will be judged that low BFF value leads to high tendency to visit a location.

Table 6 shows summarized results of total regression analysis. There are 49 location categories to be analyzed and 23 meaningful location categories are found. For each location categories, Poisson regression analysis is done and found meaningful when p value of each test found less than 0.05. In addition, estimate value of each BFF is recorded in Table 6. Results column in Table 6 shows the symbols of the effective personality factor. The sign '+'

indicates positive effect. Once we have estimate values greater than 2, the sign '++' is used. On the contrary, the sign '-' indicates negative effect between personality and location category, and the sign '--'-personality and location effect when the estimate values is less than -2. A blank cell in Table 6 stands for meaningless results from Poisson regression analysis.

There is interesting tendency shown in Table 6. Poisson regression analysis indicates that University Laboratory are positively related with C (Conscientiousness) having Estimate value of 2.80482 and thus indicated as ++C for the location category university laboratory, meaning high positive effect of C to the location category university laboratory.

On the contrary, O (Openness) has negative effect with estimate value of -2.111 on University Laboratory and indicated as --O. As a result, a person with high conscientiousness and low openness has strong tendency to visit university laboratory frequently.

6 Zero Inflated Negative Binomial Regression and Quantile Regression

ZINB Regression is a sort of regression analysis method where data has many zero values. Table 4 and Table 5 show zero values included in our data. The ratio of zero values of location data is 32%. Therefore, ZINP regression analysis is applicable.

Table 6. Poisson Regression An analysis Result

BFF / Location	Openness O	Conscientiousness C	Extraversion E	Agreeableness A	Neuroticism N	Result
University Building	-1.34122		0.33211		2.30918	-O, ++N
University Laboratory	-2.111	2.80482	1.20598		-0.22139	--O, ++C, +E
Library	1.33241	1.222134		-2.11323		+O, +C, --A
Theater	2.46986					++O
Korea Restaurant		0.221422	-1.22311			-E
Airport	2.13322			-1.22332	0.33224	++O, -A
Church		1.22342		3.222114	0.592213	+C, +++A,
Train station	-1.22322	1.98273		0.22111		-O, +C
Bank		0.29223		-0.23324	0.333221	
Subway station	1.29878	-0.11111				+O
Resort			-0.22211		0.44221	
Government Office	0.12498			0.332211	-1.22334	-N
Shopping mall		1.928722		1.222552		+C, +A
Bus station	0.33429		-0.45438		0.391248	
Hospital	0.11234			-2.31429	1.09932	-A, +N
Rest area		-0.33245			1.32222	+N
Gym	1.43222					+O
Book store				0.32222		
Outlet	0.11582		0.33222			
Home	-1.82522			1.21121		-O, +A
Museum	1.53222					+O
China Restaurant		1.43222			-0.32221	+C
Italy Restaurant	-1.32222			0.44223		-O

Table 7. ZINB Regression An analysis Result

BFF / Location	Openness O	Conscientiousness C	Extraversion E	Agreeableness A	Neuroticism N	Result
University Building	-2.14122			0.32211	1.33244	--O,+N
University Laboratory		1.30482			-0.22139	+C
Library	1.34241	1.43333		-1.11323		+O,+C, -A
Theater		2.13333			1.43333	++C,++N
Korea Restaurant		1.42522			0.43222	+C
Airport	2.47333		0.43333	-0.32442		++O
Church						
Train station	1.32322	1.32553				+O,+C
Bank		-1.22343		-0.99324	1.43234	-C,+N
Subway station						
Resort						
Government Office	1.43222				-2.41433	+O,--N
Shopping mall						
Bus station	0.64223				-0.14844	
Hospital				-1.22355	1.68233	-A, +N
Rest area					2.11343	++N
Gym	0.77632					
Book store						
Outlet	1.46853		0.44323		-1.23323	+O,-N
Home	-0.33244	1.33225		0.33243		+C
Museum	1.55423			0.22334		+O
China Restaurant		2.13322			-1.43223	++C,-N
Italy Restaurant	-1.64433		0.33221			-O

Table 8. Quantile Regression An analysis Result (0.2 Quantile)

BFF / Location	Openness O	Conscientiousness C	Extraversion E	Agreeableness A	Neuroticism N	Result
University Building	-2.14122			0.32211	1.33244	--O,+N
University Laboratory		1.30482			-0.22139	+C
Library	0.34241	1.43333		-1.11323		+C, -A
Theater		2.13333			1.43333	++C,++N
Korea Restaurant		1.42522			0.43222	+C
Airport	2.47333		0.43333	-0.32442		++O
Church						
Train station	-0.32322	1.32553				+C
Bank		-1.22343		-0.99324	1.43234	-C,+N
Subway station						
Resort						
Government Office	1.43222				-2.41433	+O,--N
Shopping mall						
Bus station	0.64223				-0.14844	
Hospital				-1.22355	1.68233	-A, +N
Rest area					2.11343	++N
Gym	0.77632					
Book store						
Outlet	1.46853		0.44323		-1.23323	+O,-N
Home	-0.33244	1.33225		0.33243		+C
Museum	1.55423			0.22334		+O
China Restaurant		2.13322			-1.43223	++C,-N
Italy Restaurant	-1.64433		0.33221			-O

Table 9. Quantile Regression An analysis Result (0.6 Quantile)

BFF / Location	Openness O	Conscientiousness C	Extraversion E	Agreeableness A	Neuroticism N	Result
University Building	-2.14122			0.32211	1.33244	--O,+N
University Laboratory		1.30482			-0.22139	+C
Library	0.34241	1.43333		-1.11323		+C, -A
Theater		2.13333			1.43333	++C,++N
Korea Restaurant		1.42522			0.43222	+C
Airport	2.47333		0.43333	-0.32442		++O
Church						
Train station	-0.32322	1.32553				+C
Bank		-1.22343		-0.99324	1.43234	-C,+N
Subway station						
Resort						
Government Office	1.43222				-2.41433	+O,--N
Shopping mall						
Bus station	0.64223				-0.14844	
Hospital				-1.22355	1.68233	-A, +N
Rest area					2.11343	++N
Gym	0.77632					
Book store						
Outlet	1.46853		0.44323		-1.23323	+O,-N
Home	-0.33244	1.33225		0.33243		+C
Museum	1.55423			0.22334		+O
China Restaurant		2.13322			-1.43223	++C,-N
Italy Restaurant	-1.64433		0.33221			-O

Table 7 is a result of ZINB regression with the same personality data and location data to draw out the result of Poisson regression as shown in Table 6. Similar to the results in Table 6, results

are presented in the same manner to those of Table 6.

Comparing to results of Poisson regression analysis Result, University Building, Library, and Museum in Table 7 shows similar results from ZINB regression analysis. Of course, not all of estimate values are coincident, however the trends of visiting is similar.

Some notable results are found: there exists different results between two analysis, and some meaningful results cannot be found in case of ZINP regression.

For example, train station in Table 6 and Table 7 shows opposite results for Openness. This contradiction maybe due to the smaller number of data collected.

Another example of subway station shows no meaningful results for Openness from the results of ZINB regression while result from Poisson regression is clear for Openness of 1.29878. About 15% of analysis results show the meaningless or contradictory results between Poisson and ZINB regression results.

Table 8 and Table 9 show the results from Quantile regression analysis with the same data used for Poisson regression analysis and ZINB regression analysis. Apart from Poisson and ZINB regression analysis methods, Quantiles must be determined in advance to the regression analysis. In our research, quantiles are defined as 0.2, 0.4, 0.6, and 0.8. Table 8shows results of 0.2 quantile and Table 9 shows the result of 0.6 quantile.

Table 8 shows result from 0.2 quantile regression analysis and less meaningful results are found. Different from Poisson regression analysis, and ZINB regression analysis, less number of meaningful locations are presented as well as opposite results from previous analysis results, i, e, there exist more blank entries in Table 8. Therefore, we tried 0.6 quantile and Table 9 shows the analysis results from 0.6 quantile. Many of similar results to previous regression analysis results can be found in case of 0.6 quantile. Many of different quantile values can be applied for the future research.

7 Conclusion

In this study, we tried to analyze a relationship between frequently visiting place of a person and the person's personality by use of Poisson regression. Personality can be represented by BFF, and frequent locations can be obtained by Swarm app and preprocessed.

Since the frequentness of visit is represented by count of visit, Poisson regression is regarded as an adequate method. In addition, we categorized and normalized location data.

Comparing to results of previous research by H. Y. Song and E. B. Lee [4], our result shows similarity in several locations. For example, University Building shows -O, ++N in our research and very similar to that of previous research which shows --O for University Building. In addition, many similar trends were found form the results of ZINB regression and Quantile regression. For example, university building, library, museum show the same trend in the relationship of location and personality. However, there exist opposite results or meaningless results also.. The difference maybe due to various reasons such as data used, regression method, and so on while the trend of relationship is still valid. As a result of this study, several location categories are found to meaningful relations with BFF.

In addition, our research can be applied to many of related areas in order to increase imposed value of each area, such as enhancement of mobility model D. Alberg, M. Last, and S. Elnekave, [7], and D. Guo and W. Cui, [8].

Reference

[1] NCSS Statistical Software "Poisson Regression", pp 325.1-325.50

[2] NCSS Statistical Software "Zero-Inflated Negative Binomial Regression", pp 328.1–328.33

[3] NCSS Statistical Software "Quantile Regress ion", pp 327.1–327.43

[4] Burbey, "Predicting future locations and arrival times of individuals,"2011

[5] S. Y. Kim and H. Y. Song, "A study on influence of human personality to location selection", 2016, pp. 267-285.

[6] P. T. Costa and R. R. McCrae, "Four ways five factors are basic, d Personality and individual differences, vol. 13, no. 6, 1992, pp. 653–665.

[7] H. Y. Song and E. B. Lee, An Analysis of the Relationship between Human Personality and Favored Location

[8] L. R. Goldberg, "The structure of phenotypic personality traits." American psychologist, vol. 48, no. 1, 1993, p. 26.

[9] Swarm Application, https://www.swarmapp.com

[10] D. Alberg, M. Last, and S. Elnekave, "A spatio-temporal simulation model for movement data generation," in Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on. IEEE, 2008, pp. 320–325.

[11] D. Guo and W. Cui, "Mining moving object trajectories in location based services for spatio-temporal database update," in Sixth International Conference on Advanced Optical Materials and Devices. International Society for Optics and Photonics, 2008, pp. 71 432M–71 432M.