

Application of Logit-Loglinear Model for Tuberculosis Disease

**Mohd Saiful bin Husain, Norlida Mohd Noor,
Fauhatuz Zahroh Shaik Abdullah**

School of Mathematical Sciences,
Universiti Sains Malaysia,
11800 Minden, Pulau Pinang, Malaysia

email: fauhatuz@usm.my, norlida@usm.my, md.epol@yahoo.com

(Received February 11 2021, Revised March 13, 2021,
Accepted June 6, 2021)

Abstract

Tuberculosis disease (TB) is known as one of the most deadly diseases in the world. Studies on TB are conducted to find out the main causes that lead to it and to promote a healthy lifestyle within the community. Our study is related to TB modeling at a location in a country in Southeast Asia. The secondary data consist of patients' information from demographic aspect, symptoms, factors, illness history and the results of chest X-rays. This study aims to indicate the primary factor that causes the individual to get a positive culture which result in TB illness and to indicate the relationship between the symptoms with the factors that cause TB. The logit-loglinear model selection in analyzing categorical data is used in this study. Three models were developed and combined which are demographic with TB symptoms, demographic with the factors that cause TB, and symptom with the factors that cause TB. From this logit-loglinear analysis, we found that the majority of the patients whose results were positive are from a particular race (Race B). In terms of gender, females were the majority group who tested positive. Hoarseness in voice was also discovered as the main symptom of TB and smoking is the main factor that initiated TB. On the other hand, factors such as alcohol, drug misuse, and HIV did not indicate that they cause TB. Meanwhile, there is an interaction between the hoarseness in voice and smoking that led to TB.

Key words and phrases: logit model, log-linear model.

AMS (MOS) Subject Classifications: 62H17, 62H20, 62J12.

ISSN 1814-0432, 2021, <http://ijmcs.future-in-tech.net>

1 Introduction

In analyzing categorical data, methods that can provide useful information to analysts such as correlation and regression analysis for continuous data are very important. Loglinear analysis for categorical data variables is one of the statistical techniques that has been continuously developed and has been widely used by many researchers including Anderson, Agresti, Bishop et al., and Fienberg [1][2][3][4].

Categorical data analysis such as loglinear model helps us to describe the associations among a collection of categorical variables with the response variables. When analyzing a disease that has only 2 categories of response variable such as 'positive' and 'negative', the logit-loglinear model could be applied [5].

Tuberculosis (TB) is a disease that attacks the human lungs and can lead to death if not treated immediately. The outbreak of TB is caused by the spread of bacteria with the scientific name 'Mycobacterium tuberculosis'. TB epidemic is not only just contagious through contact but is also transmitted through respiration, when a person inhales the air that has been contaminated by TB bacteria. These bacteria also spread into the air when a patient coughs, spits, sneezes or shouts.

The objectives of this study are:

- 1) to form a logit-loglinear model based on the selected variables to find out the main causes of TB disease,
- 2) to study the relationship between demographic data variables with symptom variables of TB disease,
- 3) to study whether smoking, alcohol intake, drug use and HIV also contribute to TB,
- 4) to study whether symptoms and factors are interrelated causing TB disease.

2 Literature Review

Loglinear model can be used to analyze infectious disease data. A study investigating the effects of temperature and relative humidity on daily new cases and daily new deaths of COVID-19 for 166 countries (excluding China) as of March 27, 2020 was done [6]. The effects of temperature and relative humidity on daily new cases and daily new deaths of COVID-19 were analyzed using a log-linear generalized additive model. The control variables include wind speed, median age of the national population, Global Health Security Index, Human Development Index and population density. The findings provide preliminary evidence

that the COVID-19 pandemic can be somewhat partially suppressed with the increase of temperature and humidity. In this regard, measures must be taken to control the source of infection to prevent further spread [6].

For any infectious disease, age and gender can be of importance because age can be a determinant of the rate at which the disease will spread. A study has been conducted to model the significance of age and gender on the spread of hepatitis B virus using data collected from Lagos State, Nigeria [7]. A 10-year data from years 2006 until 2015 was collected from the Nigeria Institute of Medical Research (NIMR). The loglinear model approach was used for the analysis and the Akaike Information Criterion (AIC) was used for the best model selection. Age and gender proved to have a significant effect on the transmission and propagation of hepatitis B virus infections in Lagos state [7].

The Lancet Commission aims to reduce tuberculosis mortality by 90% from 2015. This is as proposed in the WHO's End TB strategy as one of the Sustainable Development Goal in order to achieve a tuberculosis-free world within a generation. Although there are many challenges to end tuberculosis, rapid growth in technology such as more sensitive diagnostic tools, and advanced tuberculosis treatment strategies may help achieve the goal [8].

Other studies utilizing loglinear models for infectious disease data include a study on the size of measles infection. One study aimed to estimate the number of measles cases that infect children under 10 years old around the Blacktown area [9]. In addition, loglinear regression models with generalized estimating equation were used in another study to identify the various factors which are associated with mother-to-child-transmission; i.e., vertical transmission, as well as late access to prevention of maternal to child transmission (PMTCT) services among HIV-infected women and risk factors for infant mortality among HIV-exposed infants. The study [10] was conducted using data from 26 counties of the Yunnan Province, China health care centers. Risk factors for late initiation of maternal antiretroviral (ARV) were found to be age, ethnicity, education, and a husband not tested for HIV. The mortality rate among HIV-exposed infants was 2.9/100 person-years. Ethnicity, low birth weight and pre-term birth were found to be associated with infant mortality in addition to late initiation of maternal ARV. Therefore, the PMTCT program in Yunnan proved to have achieved low rates of MTCT [10].

3 Methodology

The loglinear model of a contingency table can be explained as the relationship of interactions between various factors in a multi-dimensional table and usually has

high dimensions. This loglinear model can occur for dimensions higher than 3. The higher the dimensions of a loglinear model, the more complex the model is because there are many relationships and interactions that make the model selection more difficult.

A larger number of hierarchical models is appropriate to see a systematic approach to model selection. Therefore, the use of step by step model selection is very suitable for four-dimensional tables or more that is a high-dimensional table.

One particular way of choosing an initial model is to choose one that consists of all effects of a particular level. The initial model can be the model of all main effects, all two-factor effects or all three-factor effects, etc. According to [11], the initial model can be chosen as either the smallest of these models that fits the data or the largest of these models that do not fit the data. Elimination and selection can be carried out by using rules for backward elimination and forward selection [11]. It should be noted that different initial models may give different results [11]. Types of arrangements of model are arrangement 1 = [A][B][C][D], Arrangement 2 = [AB][AC][AD][BC][BD][CD], Arrangement 3 = [ABC][ABD][ACD][BCD] and Arrangement 4 = [ABCD].

Estimation of maximum probability for loglinear model indicators is by estimating the parameters β_j for a set. The goodness of fit test for this model is evaluated by comparison with one of the two statistical methods below [2]:

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \quad (3.1)$$

$$G^2 = 2 \sum (\text{observed}) \ln \left(\frac{\text{observed}}{\text{expected}} \right) \quad (3.2)$$

If the response variable is binary and the statistical model for the contingency table is generated from the response variable and the explanatory variable is log-linear then this model is considered a logit model.

By considering the case where only 3 explanatory variables are involved, the logit model is arranged as below [1].

$$\begin{aligned} \text{logit}(\Pi_{1|jkl}) &= 2\Pi_1^A + 2\Pi_{1j}^{AB} + 2\Pi_{1k}^{AC} + 2\Pi_{1l}^{AD} + 2\Pi_{1jk}^{ABC} \\ &+ 2\Pi_{1jl}^{ABD} + 2\Pi_{1kl}^{ACD} + 2\Pi_{1jkl}^{ABCD} \end{aligned} \quad (3.3)$$

The above model is suitable for the hypothesis:

$$H_{01} = \tau_{1jl}^{ABD} = \tau_{1kl}^{ACD} = \tau_{1jkl}^{ABCD}, \text{ for all } j, k \text{ and } l \quad (3.4)$$

In this equation, the logit model depends on the term τ_{1j}^{AB} , τ_{1k}^{AC} and τ_{1l}^{AD} , which states the direct influence of the three explanatory variables, and three interaction

factors τ_{1jk}^{ABC} . The process of rejection of the null hypothesis H_0 depends on the statistical value of the probability ratio of G^2 falling in the critical region at the confidence level of 95% where the p-value < 0.05 . The main purpose of the model selection analysis is to obtain the simplest and the most accurate logit-loglinear model. The logit-loglinear model developed from the analysis shows the main cause of TB, the relationship between demographic with TB symptoms, demographic with the TB factors and symptom with the TB factors.

4 Results and Findings

The study is related to the TB disease of 141 patients at a location in a country in Southeast Asia which consists of 38 variables divided into four categories; namely, TB symptoms, chest X-ray results, TB factors, and patient lifestyle. It was conducted to find out the relationship between demographic variables and the symptoms of TB as well as to demographic with the factors and symptom with the factors that cause TB. This study also aims to provide awareness to the people in this country to practice healthy lifestyle by studying the effects of alcohol consumption, drug use, and HIV on the possibility of being infected with TB in the future. Before the analysis is carried out, the variables are coded as in Table 1 to facilitate the analysis.

Table 1: Code of categorical variables

Variable	Code	Category	Code
Age	A	$A \leq 21$	1
		$22 \leq A \leq 45$	2
		$46 \leq A \leq 65$	3
		$A \geq 66$	4
Race	B	Race A	1
		Race B	2
		Others	3
Gender	C	Male	1
		Female	2
Cough > 2 weeks	D	Yes	1
		No	2
Cough with phlegm	E	Yes	1
		No	2
Hemoptysis	F	Yes	1
		No	2

Variable	Code	Category	Code
Dyspnea	G	Yes	1
		No	2
Chest pain	H	Yes	1
		No	2
Loss of appetite	I	Yes	1
		No	2
Weight Loss	J	Yes	1
		No	2
Fever	K	Yes	1
		No	2
Sweating at night	L	Yes	1
		No	2
Hoarseness in voice	M	Yes	1
		No	2
Drug usage	N	Yes	1
		No	2
Alcohol	O	Yes	1
		No	2
Smoking	P	Yes	1
		No	2
IVDU	Q	Yes	1
		No	2
Culture result	R	Positive	1
		Negative	2

Table 2, selection of variables for models that include demographic data, symptoms and factors of TB. The backward selection method is done by removing any variables that do not lead to TB. This method of selection continues and stops at the 14th step. This means, step 14 contains the most suitable variables for the next analysis which includes the variables [B], [C], [M] and [P] that represent the patient's race, patient's gender, hoarseness in voice, and smoking. Next, the dependent variable which is the culture result is included. Suppose that F_{vwxyz} is a measurement of cells (v, w, x, y, z) , where v =race, w =gender, x = hoarseness in voice, y =smoking, and z =culture. An estimation where F_{vwxyz} is a multinomial and made for the purpose of hypothesis testing. The saturation model based on each study are shown in (4.5).

Table 2: Selection of variables

Step	Included Variables	Removed Variable	Method
1	[A][B][C][D][E][F][G][H][I][J][K][L][M][N][O][P][Q]		In
2	[A][B][C][D][E][F][G][H][I][J][L][M][N][O][P][Q]	K	Backward
3	[A][B][C][D][E][F][G][H][I][J][L][M][O][P][Q]	N	Backward
4	[A][B][C][D][E][F][G][H][I][J][L][M][P][Q]	O	Backward
5	[A][B][C][D][E][F][G][H][I][J][L][M][P]	Q	Backward
6	[A][B][C][D][E][F][G][H][J][L][M][P]	I	Backward
7	[A][B][C][E][F][G][H][J][L][M][P]	D	Backward
8	[A][B][C][E][F][G][H][L][M][P]	J	Backward
9	[A][B][C][E][F][G][L][M][P]	H	Backward
10	[A][B][C][F][G][L][M][P]	E	Backward
11	[A][B][C][F][G][M][P]	L	Backward
12	[A][B][C][G][M][P]	F	Backward
13	[B][C][G][M][P]	A	Backward
14	[B][C][M][P]	G	Backward

$$\begin{aligned}
 \ln m_{vwxyz} = & \lambda + \lambda_{B(v)} + \lambda_{C(w)} + \lambda_{M(x)} + \lambda_{P(y)} \\
 & + \lambda_{R(z)} + \lambda_{BC(vw)} + \lambda_{BM(vx)} + \lambda_{BP(vy)} + \lambda_{BR(vz)} \\
 & + \lambda_{CM(wx)} + \lambda_{CP(wy)} + \lambda_{CR(wz)} + \lambda_{MP(xy)} \\
 & + \lambda_{MR(xz)} + \lambda_{PR(yz)} + \lambda_{BCM(vwx)} + \lambda_{BCP(vwy)} \\
 & + \lambda_{BCR(vwz)} + \lambda_{BMP(vxy)} + \lambda_{BMR(vxz)} \\
 & + \lambda_{BPR(vyz)} + \lambda_{CMP(wxy)} + \lambda_{CMR(wxz)} \\
 & + \lambda_{CPR(wyz)} + \lambda_{MPR(xyz)} + \lambda_{BCMP(vwxy)} \\
 & + \lambda_{BCMR(vwxz)} + \lambda_{BCPR(vwyz)} + \lambda_{BMPR(vxyz)} \\
 & + \lambda_{CMPR(wxyz)} + \lambda_{BCMPR(vwxyz)}
 \end{aligned} \tag{4.5}$$

Hypothesis

- H_0 :the parsimonious model is not different from the saturated model.
- H_1 :the parsimonious model is different from the saturated model.

Table 3: Goodness of fit test

	chi-square	df	p-value
Likelihood ratio, G^2	0.000	0	-
Pearson chi-square, X^2	0.000	0	-

Table 3, the value of chi-squared probability and Pearson chi-squared for saturated model is zero. Therefore the saturated model is not significant. The non-given p-value indicates that this model has a perfect fit and always has one hundred percent probability of adapting the data.

Table 4: The effect on K-way and high order effects

K	df	Likelihood ratio chi-square	p-value	Pearson chi-square	p-value	Number of iteration
1	47	477.781	0.000	1020.191	0.000	0
2	41	135.232	0.000	155.230	0.000	2
3	27	5.341	1.000	4.242	1.000	12
4	11	0.229	1.000	0.123	1.000	3
5	2	0.000	1.000	0.000	1.000	2

Table 4, the p-value is significant when the value of $K = 2$. The maximum interaction of the model is $2^k - 1 = 2^2 - 1 = 3$. Therefore the most suitable hypothesis is as follows:

H_0 :K-way and higher order effects are zero.

H_1 :K-way and higher order effects are not zero.

Table 5: K-way effects

K	df	Likelihood ratio chi-square	p-value	Pearson chi-square	p-value	Number of iteration
1	6	342.549	0.000	864.962	0.000	0
2	14	129.891	0.000	150.988	0.000	0
3	16	5.112	0.995	4.120	0.999	0
4	9	0.229	1.000	0.123	1.000	0
5	2	0.000	1.000	0.000	1.000	0

Table 5, the process of analysing the data starts from the bottom, where the fifth order of interactions is not significant based on the probability value of the likelihood ratio chi-square, G^2 where the p-value = 1.000 which is greater than 0.05. Furthermore, the third and fourth interaction are also not significant whereas the likelihood ratio are 1.000 and 0.995 respectively. On the other hand, the first and the second interaction are significant within the 5% significant level which is 0.000 for both interactions. Thus, the 2-way interaction model is based on hypotheses as follows:

H_0 :K-way are zero.

H_1 :K-way are not zero.

Table 6: Backward selection

Step	Model	Removed Variable	df	Likelihood ratio, G^2	p-value
1	[BCMP][BCMR][BCPR] [BMPR][CMPR]		2	0.000	1.000
2	[BCMR][BCPR] [BMPR][CMPR]	[BCMP]	4	0.000	1.000
3	[BCMR][BCPR][BMPR] [CMP]	[CMPR]	5	0.000	1.000
4	[BCMR][BCPR][BMPR]	[CMP]	6	0.000	1.000
5	[BCMR][BCPR][BMP] [MPR]	[BMPR]	8	0.000	1.000
6	[BCMR][BCPR][MPR]	[BMP]	10	0.000	1.000
7	[BCMR][MPR][BCP] [BPR][CPR]	[BCPR]	12	0.001	1.000
8	[BCMR][MPR][BCP] [BPR]	[CPR]	13	0.001	1.000
9	[BCMR][MPR][BCP]	[BPR]	15	0.187	1.000
10	[MPR][BCP][BCM] [BCR][BMR][CMR]		17	0.322	1.000

Table 6, the backward selection method was applied for best model selection. The best model are listed in (4.6):

$$\begin{aligned}
 \ln m_{vwxyz} = & \lambda + \lambda_{B(v)} + \lambda_{C(w)} + \lambda_{M(x)} + \lambda_{P(y)} \\
 & + \lambda_{R(z)} + \lambda_{BC(vw)} + \lambda_{BM(vx)} + \lambda_{BP(vy)} + \lambda_{BR(vz)} \\
 & + \lambda_{CM(wx)} + \lambda_{CP(wy)} + \lambda_{CR(wz)} + \lambda_{MP(xy)} \\
 & + \lambda_{MR(xz)} + \lambda_{PR(yz)} + \lambda_{BCM(vwx)} + \lambda_{BCP(vwy)} \\
 & + \lambda_{BCR(vwz)} + \lambda_{BMR(vxz)} + \lambda_{CMR(wxz)} + \lambda_{MPR(xyz)}
 \end{aligned} \tag{4.6}$$

Table 7: Goodness of fit test

	chi-square	df	p-value
Likelihood ratio, G^2	0.322	17	1.000
Pearson chi-square, X^2	0.250	17	1.000

Table 7, the goodness of fit test. The value of the likelihood ratio and Pearson chi-square is 0.322 and 0.250 respectively. The p-value for both test is 1.000 which is more than 0.05. Thus, these two tests are not significant. Therefore hypothesis H_0 is retained where the parsimonious model is no different from the saturated model. From the best model selection, the best parsimonious model design are as shown in (4.7), where B = Race, C = Gender, M = hoarseness in voice, P = Smoking and R = Culture:

$$\begin{aligned}
 & \text{Constant} + B + C + M + P + R + BC + BM + CM + BP + CP \\
 & + MP + BR + CR + MR + PR + BCM + BCP \\
 & + BCR + BMR + CMR + MPR
 \end{aligned} \tag{4.7}$$

For logit-loglinear analysis, culture result variables were used as dependent variables and other variables namely race, gender, hoarseness in voice and smoking were used as independent variables. Based on the parsimonious model for model selection only the variable that are related to the culture result are taken into account. Therefore the parsimonious model after modification are shown in (4.8):

$$\begin{aligned}
 & \text{Constant} + R + BR + CR + MR + PR \\
 & + BCR + BMR + CMR + MPR
 \end{aligned} \tag{4.8}$$

Table 8: Parameter estimate for logit model

Parameter	Estimate	Standard Error	Z	p-value	95% C.I	
					Lower Bound	Upper Bound
[R=1]	-0.087	0.168	-0.519	0.604	-0.416	0.242
[R=1][B=1]	-0.186	0.227	-0.820	0.412	-0.631	0.259
[R=1][B=2]	0.154	0.244	0.624	0.533	-0.326	0.630
[R=1][C=1]	-0.048	0.168	-0.284	0.776	-0.377	0.281
[R=1][M=1]	0.133	0.168	0.791	0.429	-0.196	0.462
[R=1][P=1]	-0.042	0.168	-0.247	0.805	-0.371	0.288
[R=1][B=1][C=1]	-0.107	0.227	-0.473	0.636	-0.552	0.338
[R=1][B=2][C=1]	0.049	0.244	0.200	0.841	-0.429	0.527
[R=1][B=1][M=1]	0.140	0.227	0.619	0.536	-0.305	0.585
[R=1][B=2][M=1]	-0.060	0.244	-0.248	0.804	-0.538	0.417
[R=1][C=1][M=1]	0.002	0.168	0.011	0.991	-0.327	0.331
[R=1][M=1][P=1]	-0.004	0.168	-0.025	0.980	-0.333	0.325

Table 8, all p-value are not significant. Therefore the null hypothesis is retained where the parameter effect is zero. Therefore, the logit model are as shown in (4.9):

$$\begin{aligned}
 \text{logit} &= \ln \left(\frac{m_{BCMPR1}}{m_{BCMPR2}} \right) \\
 &= -0.087 - 0.186V_1 + 0.154V_2 - 0.048W_1 + 0.133X_1 \\
 &\quad - 0.042Y_1 - 0.10V_1W_1 + 0.049V_2W_1 + 0.140V_1X_1 \\
 &\quad - 0.060V_2X_1 + 0.002W_1X_1 - 0.004X_1Y_1
 \end{aligned} \tag{4.9}$$

where $V_1=1$ if B=1 and $V_1=0$ if B = 2, 3;
 $V_2=1$ if B=2 and $V_2=0$ if B = 1, 3;
 $W_1=1$ if C=1 and $W_1=0$ if C = 2;
 $X_1=1$ if M=1 and $X_1=0$ if M = 2;
 $Y_1=1$ if P=1 and $Y_1=0$ if P = 2;

From the analysis, we found that race and gender play part in the causes of TB. Hoarseness in voice was a symptom of TB and smoking was a factor causing

TB. Race B patients were most likely to be infected with TB and female patients were more likely to be infected with TB. Race B males were the most vulnerable patients. As for the interaction between demographic and symptom, Race A patients with hoarseness in voice and male patients with hoarseness in voice were more likely to be infected with TB. As for symptom and factor, there was a slight interaction between smoking and hoarseness in voice.

5 Conclusion

Logit-loglinear analysis is one of the efficient techniques in analysing categorical data in high-dimensional contingency tables. Logit-loglinear can detect complex relationships among the categorical variables through the best logit-loglinear model with related analysis.

According to the analysis done at a location in a country in Southeast Asia, it turned out that Race B patients were most likely to be infected with TB and Race A patients and female patients are more likely to get TB compared to male patients. Hoarseness in voice, which may be due to high fever, very hot weather or smoking, was one of the symptoms of TB. As for the factor, smoking was the only factor that affects a person to be infected with TB.

Race B patients are the most vulnerable people to be infected with TB with most of them being male. On the other hand, Race A patients with hoarseness in voice and male patients with hoarseness in voice are more likely to get TB. This may be due to different races adopting different lifestyles and male patients having different habits compared to female patients.

From this study, we found that there is interaction effects between smoking and hoarseness in voice that affect a patient to get TB. However, the effect of this interaction does not have a profound effect in causing TB disease. Only a small number of patients with TB were those who smoked and had hoarseness in voice in this analysis. This may be due to the analysis done involving data from only 141 patients. However, TB disease is a high risk disease and should not be taken lightly.

The conclusion that can be drawn is that TB disease can infect anyone because the main medium of infection is through the air. Public awareness is therefore quite important to prevent the spread of the disease. Consequently, the health sectors and government should take wise measures, like organizing a campaign and/or distributing leaflets warning people about the dangers of TB. To further improve the accuracy of this study, we recommend the use of more data as well as conducting the study nationwide to obtain more accurate results.

References

- [1] E. B. Anderson, Introduction to the Statistical Analysis of Categorical Data, 1997.
- [2] A. Agresti, Categorical Data Analysis, 2002.
- [3] Y. M. Bishop, S. E. Fienberg, P. W. Holland, Discrete Multivariate Analysis: Theory and Practice, 1975.
- [4] S. E. Fienberg, The Analysis of Cross-Classified Categorical Data, 1980.
- [5] J. S. Preisser, G. G. Koch, Categorical data analysis in public health, Annual Review of Public Health, **18**, (1997), 51–82.
- [6] M. Liu, Y. Wu, W. Jing, J. Liu, Q. Ma, J. Yuan, Y. Wang, M. Du, Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries, Science of The Total Environment, **729**, 2020.
- [7] O. A. Odetunmibi, A. O. Adejumo, T. A. Anake, Log-Linear Modelling of Effect of Age and Gender on the Spread of Hepatitis B Virus Infection in Lagos State, Nigeria, Macedonian Journal of Medical Sciences, **7**, no. 13, (2019), 2204–2207.
- [8] The Lancet Commissions, **393**, (2019).
- [9] C. A. McGilchrist, L. F. McDonnell, L. R. Jorm M. S. Patel, Loglinear models using capture-recapture methods to estimate the size of a measles epidemic, Journal of Clinical Epidemiology , **49**, no. 3, (1996), 293–296.
- [10] K. Meyers, H. Qian, Y. Wu, Y. Lao, Q. Chen, X. Dong, H. Li, Y. Yang, C. Jiang, Z. Zhou, Early Initiation of ARV During Pregnancy to Move towards Virtual Elimination of Mother-to-Child-Transmission of HIV-1 in Yunnan, China, PlosOne, **10**, no. 9, (2015).
- [11] R. Christensen, LogLinear Models and Logistic Regression, 1997.