UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

Estimating the Joint and Independent Effects of Body Mass Index and Age on Diabetic Mellitus Using Log-Linear Model

P Arumugam[1], Sornalatha M E[1*] and Jose P[2]

1 *Department of statistics, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, 627012, Tamil Nadu, India.*

2 *Department of CSE, Vel Tech Rangarajan Dr.SagunthalaR&D Institute of Science and Technology, Avadi,Chennai, Tamil Nadu, India.*

---

*Corresponding author. E-mail:swarnalathapramod@gmail.com;

Contributing authors: sixfacemsu@gmail.com; drjosep@veltech.edu.in;

***Abstract*—Aims: This study aims to investigate the relationship between body mass index, age, and the prevalence of diabetes mellitus using a log-linear model, exploring potential interaction effects between these variables.**

***Methods*: This study used a log-linear model to analyze the relationship between BMI, age, and diabetes in 768 females from the Akimel O'odham Indians Diabetes Dataset. The analysis considered BMI-age interactions and used goodness-of-fit statistics to select the most insightful model, focusing on identifying significant interactions and estimating the effects of BMI and age on diabetes prevalence.**

***Results*: Both BMI and age are significantly associated with diabetes prevalence. Higher BMI and older age correlate with increased diabetes likelihood. The interaction between age and BMI reveals that BMI's impact on diabetes risk is more pronounced in older individuals. These findings highlight the importance of considering both factors when assessing diabetes risk.**

***Conclusions*: BMI and age significantly influence diabetes prevalence in females, with their interaction adding complexity to risk assessment. This effect is more pronounced in older individuals. Further research is needed to understand the underlying mechanisms and develop targeted interventions. This study contributes to informing public health strategies for diabetes prevention and management.**

***Index Terms*—Health care, Diabetic mellitus, Body Mass Index, estimation, Loglinear model**

Introduction

Diabetes mellitus (DM) is a frequent and increasing public health problem. This condition arises either due to insufficient production of insulin by the pancreas or the inability of the body to effectively use the insulin it produces. The most influencing hormone that plays a key role in varying blood sugar levels is Insulin, which orchestrates the ingestion of glucose into cells for energy. Statistics reveal that every 1/10 persons in the world suffers from diabetes which comes to 537 million adults in the range of 20-79 years. It is predicted that these numbers can rise to around 643 million by 2030 and 783 million by 2045. Of these, it is projected that 3 out of 4 adults having diabetes come from underdeveloped and developing countries. This research is an attempt to analyse and ascertain the relationship between variables which are key factors in DM using the unique capabilities of the log-linear model. This includes analysing prominent variables like the pregnancies, BMI, diastolic BP, Triceps, skin fold thickness, serum, insulin level after two hour of the glucose, tolerance test and the diabetic pedigree function. In this research, both log linear and logistic regression methods are possible to be used for arriving the conclusion. In this research, both log linear and logistic regression methods are possible to be used for arriving the conclusion. However, (14) log-linear models are proven to be more effective than logistic regression in analysing most of the scenarios. A saturated model was constructed based on the recommendations (1; 10) on the multidimensional model for contingency tables (1; 3). Multidimensional variables can be analysed effectively in a log-linear model to get more precise results. The use of multiple methods for model framework (13) during the selection process, which includes the introduction of the Poisson model for the observed data helps to arrive at more accurate results. The article (3) provided insight on how the categorical variables can be effectively converted from the co-variants used in the data sets, how loglinear model (14; 2) applied in covid 19 dynamics, estimate unknown from known samples (7) and demographic variable effect on patients. (10) Compare the algorithms using real-world datasets presented in the research papers (5; 11; 12; 4) that are relevant to the research question, choose the method that best suits the specific needs. The methods discussed in the research papers (6; 7; 8; 9) aided in practically estimating the demographic variable effect on common patients. The methods discussed in the research papers aided in practically estimating the

demographic variable effect on common patients. From the figure1, it is very evident that the key highlight is the increase in the number of deaths due to diabetic mellitus. Also, if we look at the number of cases reported on diabetic mellitus, we see a sharp increase. The data collected from the Ministry of Health, India Government, indicates that the IGT (Impaired Glucose Tolerance) has also increased considerably. One of the key inferences from this attempt is the need for an analysis of the data sets from the affected patients to deduce the optimal control of various factors influencing diabetes mellitus. Currently, there are various studies available to identify the factors that impact the spread of diabetes mellitus. However, not many studies are focused on analysing the combination of these factors.

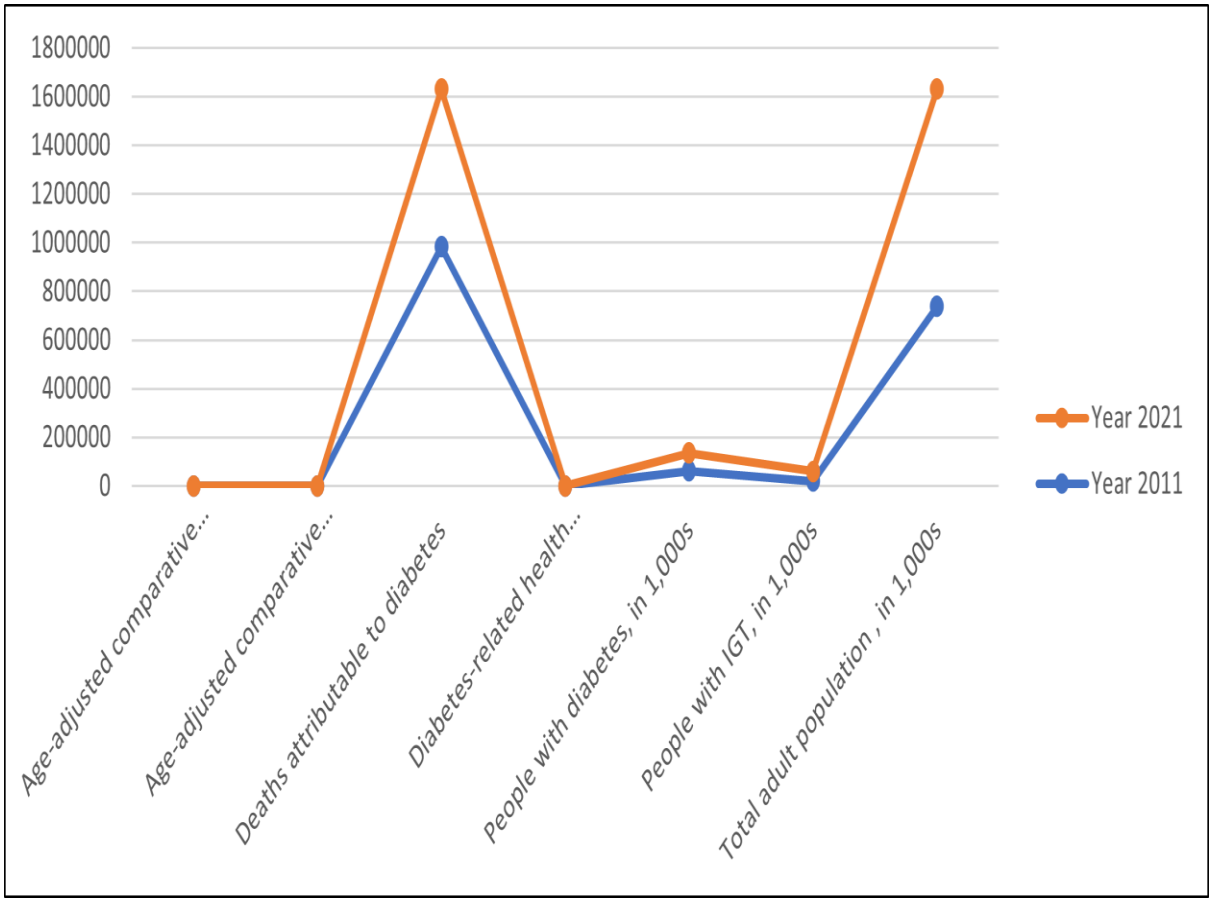

Fig 1: Comparison of Key factors

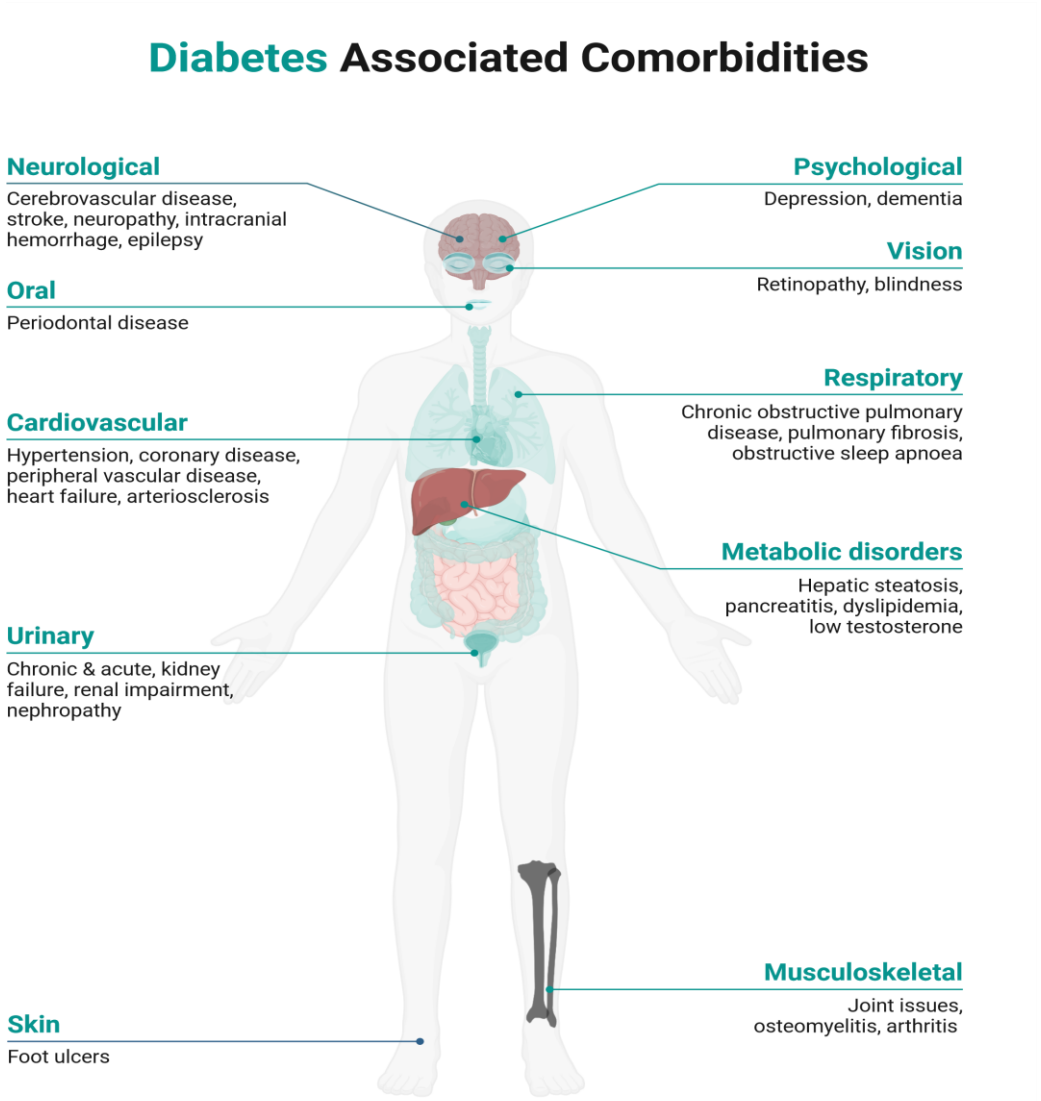

Fig 2: Diabetes Associated Comorbidities

Diabetes mellitus often doesn't occur in isolation. It frequently coexists with other chronic conditions, known as comorbidities, which can worsen its impact and complicate management. Fig.2 Brief overview of common diabetes-related comorbidities. Log-linear models have emerged as a powerful tool in the field of predictive analytics, offering a versatile and robust approach to modelling complex relationships between variables. This study explores the application of log-linear models in Diabetic Mellitus conditions, highlighting their advantages and potential limitations

This research attempts to analyse and ascertain the relationship between variables which are key factors in DM using the unique capabilities of the log-linear model. This includes analysing prominent variables like number of pregnancies, BMI, diastolic BP,

Triceps, skin fold thickness, serum, insulin level after two hours of the glucose tolerance test and the diabetic pedigree function. The prime objective of this study is to identify and estimate the parameters factored in the model which are significant interactions in the DM. Based on the significance values from the Pearson chi-square and likelihood ratio statistics used in the log-linear model analysis, we identified the significant parameters that influence diabetes mellitus. To estimate this, we rely on the log-linear methodology.

Materials and Methods

This study considered 768 Females from the data sets available at Akimel O'odham Diabetes Database on Kaggle. Among the models generated in the log-linear approach, the most insightful model was decided by filtering the higher order model to lower. The significant variables are identified after parameter estimation of the most fitting model. We can relate the analysis of variance concept (ANOVA) for the continuously distributed independent variables against the log-linear analysis method for Poisson distribution. In case of ANOVA, we require to check the normality condition for all variables used and have the limitation that the categorical variables cannot be compared. In contrast, the log-linear model gives the flexibility to compare the categorical variables and arrive at better insights. Another pressing reason for involving log linear model is due to the limitations in the chi-square and the ANOVA test in determining the associations between categorical variables. While chi-square technique restricts the determination of the associations between the two variables only, ANOVA model allows to determine multiple variables, it never allows to determine associations for any categorical variables. When we consider the log-linear models, we have the flexibility to include multiple rows and columns involving both the two-way and three-way contingency tables. In addition to this, log linear model supports testing more than one hypothesis, when compared to other models in this category. Hence this is the preferred approach. In the case of multi-dimensional contingency tables for log-linear models, where a model is created to investigate the relationships between the variables, the parameters in the model are estimated and the significance of this model is also tested. The overall goodness-of-fit of a model is evaluated by comparing the expected frequencies against the observed cell frequencies for each model. The Pearson chi-square statistics or the likelihood ratio statistic is mostly used to test a model fit.

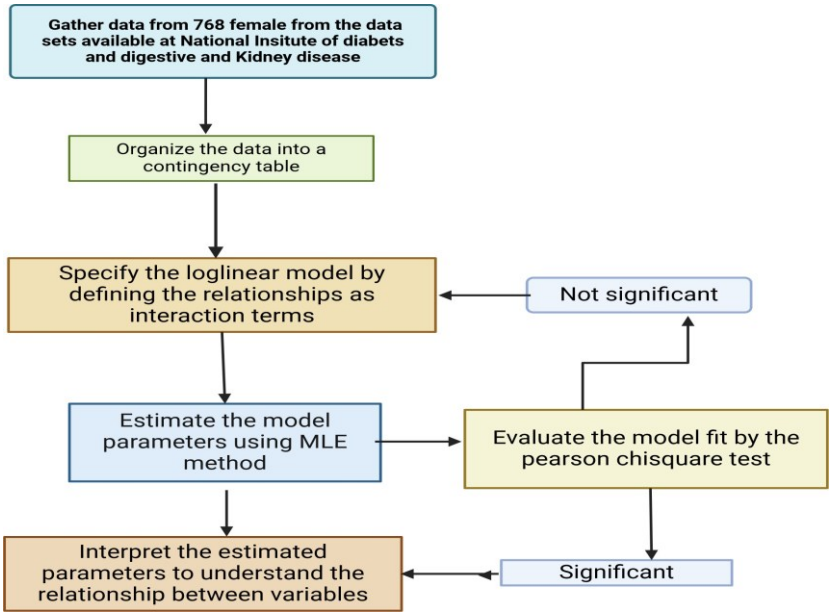

Fig 3: Flowchart of the Log-linear Modelling Process

Fitting procedure for log-linear models

Log-linear model methodology for independent two-way table is

$$\log(\mu ij) = X_O + X_i^1 + X_j^2 + X_{ij}^{12} \quad (1)$$

The constraints of X term for all "i" and "j" sum to zero. such as

$$\sum_{i=1}^{I} X_i^1 = \sum_{j=1}^{J} X_j^2 = \sum_{i=1}^{I} X_{ij}^{12} = \sum_{j=1}^{J} X_{ij}^{12} = 0$$

The above log-linear model is the saturated model for the statistical dependency between any two variables. This is similar to multiple regression and ANOVA with factors. This implies that the overall mean is arrived by the below equation

$$X_O = \frac{1}{IJ} \sum_{i=1}^{I} X_i^1 \log \mu_{ij} \quad (2)$$

In the below equation, the key impact of the two variables is estimated by

$$X_J^1 = \frac{1}{J} \sum_{I=1}^{J} 1 \log \mu_{ij} - X_O \quad (3)$$

$$X_i^1 = \frac{1}{I} \sum_{i=1}^{I} 1 \log \mu_{ij} - X_O \quad (4)$$

For the same variables, the two-factor effect can be arrived by

$$X_{ij}^{12} = log\mu_{ij} = \left(X_i^1 + X_j^2\right) - X_0 \qquad (5)$$

Thus, the key and two-factor effects can be found out by the odds, odds ratio, which can be represented as

$$X_i^1 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} log\frac{\mu_{ij}}{\mu_{i'j}} \qquad (6)$$

$$X_i^2 = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} log\frac{\mu_{ij}}{\mu_{j'i}} \qquad (7)$$

$$X_{ij}^{12} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \log\left(\frac{\mu_{ij}\mu_{ij}}{\mu_{i'j}\,\mu_{j'i}}\right) \qquad (8)$$

The cell probability for the independent model, considering the statistically independent two variables, can be represented by marginal probabilities of μ_{i+} and μ_{+j} as follows

$$\mu_{ij} = \mu_{i+}\mu_{+j} where \mu_i = \sum_{j=1}^{J}\mu_{ij} and \mu_{+j} = \sum_{i=1}^{I}\mu_{ij}$$

Hence the two-factor outcome is

$$X_{ij}^{12} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{log(\mu_{i+}\mu_{+j}\,\mu_{i^1+}\,\mu_{+j^1})}{log(\mu_{i^1j}\mu_{+j}\,\mu_{i+}\,\mu_{+j^1})} = 0 \qquad (9)$$

and the log linear model for the independence model can be mentioned as

$$\log\left(\mu_{ij}\right) = X_o + X_i^1 + X_j^2 forall"i"and"\ j" \qquad (10)$$

In case of three-way table, the log linear model for independence will be

$$log_{(ijk)} = X_0 + X_I^1 + X_J^2 + X_K^3 + X_{IJ}^{12} + X_{JK}^{13} + X_{JK}^{23} + X_{IJK}^{123} for\ all\ "i","j","k"$$

For scenario where the X term satisfies the below constraints,

$$\sum_{i=1}^{I} X_i^1 + \sum_{j=2}^{J} X_j^2 + \sum_{k=3}^{K} X_k^3 = 0$$

$$\sum_{i=1}^{I} X_{ij}^{12} + \sum_{j=2}^{J} X_{ik}^{13} + \sum_{k=3}^{K} X_{jk}^{23} = 0$$

$$\sum_{i=1}^{I} X_{ijk}^{123} = sum_{j=1}^{J} X_{ijk}^{123} = \sum_{k=1}^{K} X_{ijk}^{123} = 0$$

Then we can define the X terms as below

The overall mean is given by

$$X_0 = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=2}^{J}\sum_{k=2}^{K} log\mu_{ijk} \quad (11)$$

The main effects of the three variables are estimated by

$$X_i^1 = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} log\mu_{ijk} - X_0 \quad (12)$$

$$X_i^2 = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} log\mu_{ijk} - X_0 \quad (13)$$

$$X_i^3 = \frac{1}{JI}\sum_{j=1}^{J}\sum_{I=1}^{I} log\mu_{ijk} - X_0 \quad (14)$$

Every interaction above is estimated by the equation

$$X_{ij}^{12} = \frac{1}{K}\sum_{k=1}^{K} log\mu_{ijk} - \left(X_i^1 + X_j^2\right) - X_0 \quad (15)$$

$$X_{ik}^{13} = \frac{1}{J}\sum_{j=1}^{J} log\mu_{ijk} - (X_i^1 + X_k^3) - X_0 \quad (16)$$

$$X_{jk}^{23} = \frac{1}{I}\sum_{i=1}^{I} log\mu_{ijk} - \left(X_j^2 + X_k^3\right) - X_0 \quad (17)$$

and

$$X_{ijk}^{123} = log\mu_{ijk} - \left(X_{ij}^{12} + X_{ik}^{13} + X_{jk}^{23}\right) - \left(X_i^1 + X_j^2 + X_k^3\right) - X_0 \quad (18)$$

The parameter estimates from the loglinear model represent the logarithmic transformation of the odds ratios for different combinations of variables. Converting these coefficients back to their original scale yields odds ratios, which are more readily interpretable. Significance testing allows to evaluate whether individual interaction terms in the model are statistically significant, which helps identify the key relationships.

Observations and Results

Table 1

Summary of significant two-way and one-way effects

Factor	Degrees of freedom	Significance
Outcome * BMI group	8	.005
Age group * TSFT	8	.049
BMI group * DBP	7	.013

Factor	Degrees of freedom	Significance
Outcome * GLC	6	.000
Age group * GLC	7	.006
outcome * PL	6	.006
Age group * PL	6	.000
BMI group * PL	8	.045
Outcome	2	.000
Age group	2	.000
BMI group	2	.000

Using the backward elimination process, we summarised the test results of the significant interrelations on the eight variables used in this study in Table 1. Among the several two-way and three-way models generated from the saturated model, 3 three-way models and 8 two-way models showed significant inter relations. We have focused on the best two-way model with variables representing 'the glucose level after 2 hours of glucose tolerance tests and age group. This model was used in log linear model under the null hypothesis to derive the estimates.

Table 2

Cell Counts of BMI and Diabetic condition from Log-linear two -way model

Outcome	BMI	Count	Percentage
Not having diabetic	Underweight	10	1.2
	Normal	450	58.5
	Overweight	33	4.2
	Obesity	11	1.5
Having diabetic	Underweight	3	0.3
	normal weight	213	27.7
	Overweight	51	6.6
	Obesity	5	0.6

There are other prominent models to understand the relationship between two variables. In this, the chi-square test of independence can show only whether any two variables are related. However, this cannot find the expected frequencies of each parameter. In the case of the log-linear model, this can be found in the estimates of parameters under the independence model (model under the null hypothesis) and the

maximum model (model under the alternative hypothesis). Table 2 exhibits the most significant interrelation between normal weight and not having a diabetic.

Table 3

Parameter estimation of the model glucose level and age group for DM patients

Parameter	Estimate
X_1	2.603
X_1^1	2.164
X_2^1	1.532
X_1^2	0.071
$X_1^1 X_1^2$	1.288
$X_2^1 X_1^2$	0.452

Table 3 manifests estimate of the cell counts of two-way model glucose level and age group.

Fig. 4 provides a visual representation of the probabilities across different age and BMI groups. The colour intensity represents the probability level. This heatmap would visually represent how the probability of diabetes progression varies across different combinations of age and BMI. While the log-linear model itself provides the statistical significance of these relationships, the heatmap makes them visually apparent and easier to grasp.

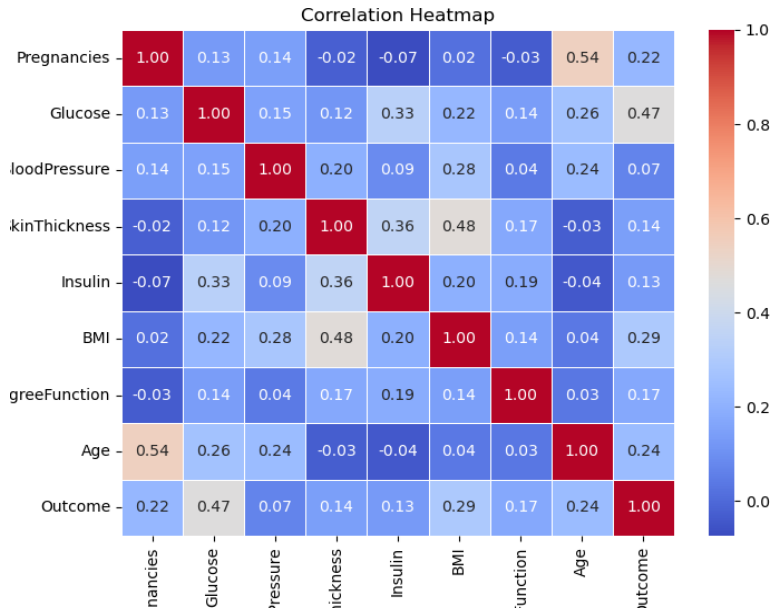

Fig 4: Correlation Heatmap

Table 4

Observed and Expected frequency of the model, age group, BMI and Glucose Level with their corresponding standardized residual for DM patients

cell			Observed	Expected	Standardized residual
i	j	k			
1	1	0	9	9	0.092
1	1	1	403	405	-0.121
1	1	2	36	36	0.050
1	1	3	9	8	0.696
1	2	0	1	2	-0.710
1	2	1	95	93	0.192
1	2	2	9	8	0.278
1	2	3	0	2	-1.281
1	3	0	1	0	1.417
1	3	1	13	12	0.165
1	3	2	0	1	-1.046
1	3	3	0	0	-0.468
2	1	1	87	91	-0.461
2	1	2	28	23	1.148
2	1	3	2	3	-0.600
2	2	1	50	48	0.225
2	2	2	9	12	-0.853
2	2	3	3	2	1.090
2	3	1	13	10	0.892
2	3	2	0	3	-1.583
2	3	3	0	0	-0.582

Discussion

Log-linear analysis revealed significant differences in residual errors between specific age, BMI, and glucose level combinations. From Table 4 individuals aged 20-40, both underweight and overweight groups with glucose levels below 140 showed small

residual errors (0.092 and 0.050, respectively), suggesting a strong fit between the model and the data. Conversely, individuals above 60 showed larger residual errors (1.417 for underweight and -1.583 for normal weight with glucose levels less than 140 and between 140-200, respectively), indicating weaker relationships. These findings suggest the log-linear model's predictions are more precise for certain subgroups, particularly younger individuals.

This study investigated the impact of age and BMI on Diabetic Mellitus using a log-linear model. Our analysis revealed significant interrelationships between these demographic factors and DM. Specifically, strong associations were observed between glucose levels and age group, and between BMI and diabetic condition. The risk of progressing to a more severe stage of diabetes increases more rapidly with age for individuals with higher BMI compared to those with lower BMI. The increasing prevalence and mortality rates of DM underscore the importance of understanding these contributing factors

Table 5

Goodness of fit and degrees of freedom for associated variables for 3-way model

Variables	Likelihood ratio	Pearson chi square statistic	degrees of freedom
Outcome * BMI * GL	19.046	21.36	16
Age group * GL * PL	2.815	2.917	6
Age group* GL * BMI	18.369	13.613	18

The log-linear model proved to be a valuable tool for analyzing these complex relationships, offering more specific predictions compared to other models. By employing backward elimination, we identified the most significant two-way and three-way interactions, refining the model and highlighting key contributing factors. While many three-dimensional models did not yield significant outcomes, the significant three-way interactions provided valuable insights into the combined effects of age, BMI, and other relevant variables. These findings are presented in the Table 5.

Conclusion and Recommendations

Loglinear models can be valuable in data mining when your focus is on understanding relationships between categorical variables and interpretability is crucial. We have considered 768 samples from female patients for this analysis. Among various statistical models, the log-linear model provided more insights, considering the ability to consider the significant interrelations among various levels of factors that affect diabetic mellitus. In this context, we decided to further analyse the significant interrelationships among these factors to provide a more focused conclusion. Hence, we considered the log-linear model approach to analyse the data sets and come up with more mature suggestions, which in turn can help the patients to optimally control diabetes mellitus. From various inferences on the interrelations among the variables, it is recommended to involve a more focused mathematical model to measure the precise level of interrelations effect which in turn, can help to develop better drugs to control diabetic mellitus optimally. log-linear analysis highlights the complex interplay among age, BMI, and diabetes mellitus. Recognizing these interactions is crucial for effective prevention, diagnosis, and management of this chronic condition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work re- ported in this article.

Funding

This research was supported by the Manonmaniam Sundaranar University for the M.G. Ramachandran Centenary Fellowship. (MSU/R/RES/R2)

Acknowledgement

The Second author gratefully acknowledges Manonmaniam Sundaranar University's financial support for the M.G. Ramachandran Centenary Fellowship. (MSU/R/RES/R2)

References

[1] A. Agresti, Categorical data analysis, volume 792. John Wiley & Sons, 2012.

[2] A. Bhaskar, C. Ponnuraja, R. Srinivasan, and S. Padmanaban, Distri- bution and growth rate of covid-19

outbreak in Tamil nadu, A log-linear regressionapproach, Indian Journal of Public Health, 64(6):188, 2020.

[3] S. E. Fienberg, The analysis of cross-classified categorical data, Springer- Science & Business Media, 2007.

[4] M. S. b. Husain, N. M. Noor, and F. Z. S. Abdullah, Application of logit-loglinear model for tuberculosis disease, International Journal of Mathematics & Computer Science, 16(3), 2021.

[5] O. A. Odetunmibi, A. O. Adejumo, and T. A. Anake, Log-linear modellingof effect of age and gender on the spread of hepatitis b virus infection inlagos state nigeria, Open Access Macedonian Journal of Medical Sciences,7(13):2204, 2019.

[6] M. Tiensuwan, P. Yimprayoon, and Y. Lenbury, Application of log- linear models to cancer patients: A case study of data from the national cancer institute, Southeast Asian journal of tropical medicine and public health,36(5):1283, 2005.

[7] Fokianos, Konstantinos and Tjøstheim, Dag, Log-linear Poisson autoregression,Journal of multivariate analysis, Elsevier,102(3),563-578,2011.

[8] Jamaludin, Siti Zulaikha Mohd and Romli, Nurul Atiqah and Kasih- muddin, Mohd Shareduwan Mohd and Baharum, Aslina and Mansor, Mohd Asyraf and Marsani, Muhammad Fadhil,Novel logic mining incorporating log linear approach,Elsevier,34(10),9011-9027,2022.

[9] Gilbert, Nigel, Analyzing tabular data: Loglinear and logistic models for social researchers, Routledge,2022.

[10] Iwasaki, Yuichi and Fukaya, Keiichi and Fuchida, Shigeshi and Matsumoto, Shinji and Araoka, Daisuke and Tokoro, Chiharu and Ya- Yutaka, Tetsuo, Projecting future changes in element concentrations of approximately 100 untreated discharges from legacy mines in Japan by a hierarchical log-linear model, Science of the Total environment, Elsevier,786,147500,2021.

[11] Yuan, Rui and Du, Simon S and Gower, Robert M and Lazaric, Alessan- dro and Xiao, Lin, linear convergence of natural policy gradient methods with log-linear policies, International Conference on Learning Rep- resentations,2023.

[12] P.Arumugam, V.Ambikavathi, Feature Selection Using Bisection Branch and bound algorithm for diabetes,Studies In Indian Place Names,40,70,2020.

[13] Coons, Jane Ivy and Langer, Carlotta and Ruddy, Michael,Classical iterative proportional scaling of log-linear models with rational maximum likelihood estimator,International Journal of Approximate Reasoning,Elsevier,164,109043,2024.

[14] Alzahrani, Salem Mubarak, A log linear Poisson autoregressive model to understand COVID-19 dynamics in Saudi Arabia,Beni-Suef University Journal of Basic and Applied Sciences,Springer,11(1),118,2024