HOW DOES INCOME INFLUENCE THE EFFECTS OF LIFESTYLE AND PHYSICAL CONDITION ON DIABETES?

JI QI
STAT5020 FINAL PROJECT

Abstract. This study investigates how income level impacts the effects of lifestyle and physical condition on diabetes risk. Using data from the Centers for Disease Control and Prevention submitted to the UCI Machine Learning Repository, structural equation models with multisample data were utilized, treating income level as a grouping indicator. The analysis divided contributing factors into lifestyle and physical condition categories, aiming to discern the influence of income levels on diabetes risk. Initial findings suggest significant differences between lower and higher income groups regarding lifestyle's impact on diabetes risk. For lower income individuals, unhealthy lifestyle factors mitigate diabetes risk, whereas for higher income individuals, these factors amplify the risk. Additionally, the influence of physical conditions on diabetes risk is slightly more significant for lower income groups, although the effects are relatively close between groups. Model comparison using Bayes Factor supports the hypothesis that coefficients in the structural equation vary between groups, indicating that income level indeed affects how lifestyle and physical condition factors influence diabetes risk.

Contents

1. Introduction

Diabetes is among the most prevalent chronic diseases in China, impacting millions of people each year and exerting a significant financial burden on the economy. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood, and can lead to reduced quality of life and life expectancy. Diabetes is generally characterized by either the body not making enough insulin or being unable to use the insulin that is made as effectively as needed. Complications like heart disease, vision loss, lower-limb amputation, and kidney disease are associated with chronically high levels of sugar remaining in the bloodstream for those with diabetes.

There are many factors will increase the probability of having diabetes, and we are interesting how the income level influences the effects of these factors on diabetes. Here, We will divide these factors affecting diabetes into three categories: lifestyle factors and physical condition factors. Lifestyle factors include veggie, fruit, and alcohol consumption, mental health level, physical activity, smoking habit, etc. Physical condition factors include BMI level, difficulty of walking or climbing stairs, history of stroke and heart attack, etc. High BP level and cholesterol level indicates high diabetes risk.

In the previous analysis of diabetes health indicators data[1], the influence of the income level on the effects of lifestyle and physical condition on diabetes is not studied. Hence, we plan to conduct another analysis to understand the interplay between income levels and the aforementioned variables on the incidence of diabetes. The aim is to discern whether higher or lower income levels amplify or mitigate the effects of lifestyle and physical condition factors on diabetes risk.

This project will consider structural equation models with multisample data, treating the income level as the group indicator. Besides, we will use the Bayes Factor to realize the model comparison.

2. Data

The dataset is submitted by Centers for Disease Control and Prevention to UCI Machine Learning Repository (which is the first recommended data source of our final project) [2]. The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. The raw data contains 253,680 samples. The 35 features consist of some demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy. For simplicity, we select lifestyle and physical conditions features, and then randomly sample 1000 samples from the raw data.

We construct one outcome latent variable, representing the level of diabetes risk, and two explanatory latent variables, representing the lifestyle condition factors and the physical condition factors. The list of symbols and corresponding meanings follows:

- individual indicator: i, representing each individual.
- group indicator: g, representing the income level group index.
  - 0 for less than \$35000, 1 for greater or equal to \$35000.
- latent variables: $\Omega = (\omega_i)$.
  - outcome latent variable: η_i, representing the risk factor of diabetes.
  - explanatory latent variables: ξ_i.
    * ξ_{1i}: the risk related to lifestyle factors.
    * ξ_{2i}: the risk related to physical condition factors.
- observed variables: $Y = (y_i)$.

- y_1: Diabetes, 0 = no diabetes, 1 = prediabetes, 2 = diabetes.
- y_2: HighBP, 0 = no high BP, 1 = high BP.
- y_3: HighChol, 0 = no high cholesterol, 1 = high cholesterol.
- y_4: HvyAlcoholConsump, heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week) 0 = no 1 = yes.
- y_5: Fruits, consume Fruit 1 or more times per day, 0 = no 1 = yes.
- y_6: Veggies, consume Vegetables 1 or more times per day, 0 = no 1 = yes.
- y_7: PhysActivity, physical activity in past 30 days - not including job, 0 = no 1 = yes.
- y_8: Smoker, have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes.
- y_9: DiffWalk, do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes.
- y_{10}: Body Mass Index.
- y_{11}: Stroke, (ever told) you had a stroke, 0 = no 1 = yes.
- y_{12}: HeartDiseaseorAttack, coronary heart disease (CHD) or myocardial infarction (MI), 0 = no 1 = yes.

Remark 1. Since the synthetically assigned values for missing data are continuous, then all these selected observed variables are treated as continuous.

3. Model

3.1. M_0: vanilla one-sample model.

M_0 is the one-sample model which does not uses the income level information. The motivation of developing M_0 is to compare multi-sample model M_1 with vanilla model M_0 and then select the better model for the diabetes outcomes. Another motivation is to get the informative prior for our main model M_1.

$$
\begin{pmatrix} y_{1i} \\ y_{2i} \\ y_{3i} \\ y_{4i} \\ y_{5i} \\ y_{6i} \\ y_{7i} \\ y_{8i} \\ y_{9i} \\ y_{10i} \\ y_{11i} \\ y_{12i} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \\ \mu_7 \\ \mu_8 \\ \mu_9 \\ \mu_{10} \\ \mu_{11} \\ \mu_{12} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{31} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{52} & 0 \\ 0 & \lambda_{62} & 0 \\ 0 & \lambda_{72} & 0 \\ 0 & \lambda_{82} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{10,3} \\ 0 & 0 & \lambda_{11,3} \\ 0 & 0 & \lambda_{12,3} \end{pmatrix} \begin{pmatrix} \eta_i \\ \xi_{1i} \\ \xi_{2i} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \\ \varepsilon_{4i} \\ \varepsilon_{5i} \\ \varepsilon_{6i} \\ \varepsilon_{7i} \\ \varepsilon_{8i} \\ \varepsilon_{9i} \\ \varepsilon_{10i} \\ \varepsilon_{11i} \\ \varepsilon_{12i} \end{pmatrix}
$$

$$
\begin{aligned}
\eta_i &= \gamma_1 \xi_{1i} + \gamma_2 \xi_{2i} + \delta_i, \\
\varepsilon_{ik} &\sim N\left(0, \psi_{\varepsilon k}\right) \\
\delta_i &\sim N\left(0, \psi_\delta\right) \\
\xi_i &\sim N\left(0, \Phi\right)
\end{aligned}
$$

The prior distributions of parameters are

$$
\begin{cases}
\mu_k \sim \mathcal{N}\left(\bar{y}_k, 0.25\right) \\
\lambda \sim \mathcal{N}(0, 25) \\
\gamma_1, \gamma_2 \sim \mathcal{N}(0, 25) \\
\Phi \sim \mathrm{IW}_2\left(\begin{bmatrix} 0.04 & 0.05 \\ 0.05 & 0.7 \end{bmatrix}, 5\right) \\
\psi_{\epsilon k} \sim \mathrm{IG}(9, 4) \\
\psi_\delta \sim \mathrm{IG}(9, 4)
\end{cases}
$$

The sample traces of the chains are showed in 1. The posterior mean and standard deviation of estimated parameters are in 1. From the sample traces of the chains, we can see that the estimation converges after 2000 iterations.

3.2. M_1: **multi-sample model.** M_1 is the multi-sample model which uses the income level information. By observing the parameters of different group, we can better understand the influence of income levels on the effects of lifestyle and physical condition factors on diabetes risk.

The measurement and structural equation of M_1 is:

$$\begin{pmatrix} y_{1i}^{(g)} \\ y_{2i}^{(g)} \\ y_{3i}^{(g)} \\ y_{4i}^{(g)} \\ y_{5i}^{(g)} \\ y_{6i}^{(g)} \\ y_{7i}^{(g)} \\ y_{8i}^{(g)} \\ y_{9i}^{(g)} \\ y_{10i}^{(g)} \\ y_{11i}^{(g)} \\ y_{12i}^{(g)} \end{pmatrix} = \begin{pmatrix} \mu_1^{(g)} \\ \mu_2^{(g)} \\ \mu_3^{(g)} \\ \mu_4^{(g)} \\ \mu_5^{(g)} \\ \mu_6^{(g)} \\ \mu_7^{(g)} \\ \mu_8^{(g)} \\ \mu_9^{(g)} \\ \mu_{10}^{(g)} \\ \mu_{11}^{(g)} \\ \mu_{12}^{(g)} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21}^{(g)} & 0 & 0 \\ \lambda_{31}^{(g)} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{52}^{(g)} & 0 \\ 0 & \lambda_{62}^{(g)} & 0 \\ 0 & \lambda_{72}^{(g)} & 0 \\ 0 & \lambda_{82}^{(g)} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{10,3}^{(g)} \\ 0 & 0 & \lambda_{11,3}^{(g)} \\ 0 & 0 & \lambda_{12,3}^{(g)} \end{pmatrix} \begin{pmatrix} \eta_i^{(g)} \\ \xi_{1i}^{(g)} \\ \xi_{2i}^{(g)} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1i}^{(g)} \\ \varepsilon_{2i}^{(g)} \\ \varepsilon_{3i}^{(g)} \\ \varepsilon_{4i}^{(g)} \\ \varepsilon_{5i}^{(g)} \\ \varepsilon_{6i}^{(g)} \\ \varepsilon_{7i}^{(g)} \\ \varepsilon_{8i}^{(g)} \\ \varepsilon_{9i}^{(g)} \\ \varepsilon_{10i}^{(g)} \\ \varepsilon_{11i}^{(g)} \\ \varepsilon_{12i}^{(g)} \end{pmatrix}$$

$$\eta_i^{(g)} = \gamma_1^{(g)}\xi_{1i}^{(g)} + \gamma_2^{(g)}\xi_{2i}^{(g)} + \delta_i^{(g)},$$
$$\varepsilon_{ik}^{(g)} \sim N\left(0, \psi_{\varepsilon k}^{(g)}\right)$$
$$\delta_i^{(g)} \sim N\left(0, \psi_{\delta}^{(g)}\right)$$
$$\xi_i^{(g)} \sim N\left(0, \Phi^{(g)}\right)$$

The prior distributions of parameters are

$$\begin{cases} \mu_k^{(g)} \sim \mathcal{N}\left(\mu_k^{M_0}, 0.25\right) \\ \lambda^{(g)} \sim \mathcal{N}(\lambda_k^{M_0}, 25) \\ \gamma_k^{(g)} \sim \mathcal{N}(\gamma_k^{M_0}, 25) \\ \Phi^{(g)} \sim \mathrm{IW}_2\left(\Phi^{M_0}, 5\right) \\ \psi_{\epsilon k}^{(g)} \sim \mathrm{IG}(\psi_{\epsilon k}^{M_0}, 4) \\ \psi_{\delta}^{(g)} \sim \mathrm{IG}(\psi_{\delta k}^{M_0}, 4) \end{cases}$$

Here, $\bullet^{M_0}$ represents the posterior mean of parameter $\bullet$ in M_0 estimation. The prior sets are identical for two groups since we do not know the difference between the low income group and the high income group.

The sample traces of the chains are showed in 2. The posterior mean and standard deviation of estimated parameters are in 2. From the sample traces of the chains, we can see that the estimation converges after 2000 iterations.

4. Model Comparison and Testing

In this subsection, we consider an hypothesis testing problem:

$$H_0 : \mathbf{\Gamma}^{(1)} = \mathbf{\Gamma}^{(2)} = \mathbf{\Gamma} \quad \text{v.s.} \quad H_1 : \mathbf{\Gamma}^{(1)} \neq \mathbf{\Gamma}^{(2)}$$

We will realize model comparison via Bayes factor. Rewrite M_1 in matrix form:

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \mathbf{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} = \mathbf{\Gamma}^{(g)} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)} \end{cases}$$

and consider M_2:

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \mathbf{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)} \\ \eta_i^{(g)} = \mathbf{\Gamma} \boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}. \end{cases}$$

Since the distributions of error measurements and explanatory output latent variables are all the same in this subsection,

$$\begin{cases} \boldsymbol{\epsilon}_i^{(g)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Psi}_{\boldsymbol{\epsilon}}{}^{(g)}\right), \\ \delta_i^{(g)} \sim \mathcal{N}\left(0, \psi_\delta^{(g)}\right), \\ \boldsymbol{\xi}_i^{(g)} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Phi}^{(g)}\right) \end{cases}$$

then we will ignore these distributions when describing the model in this subsection. We introduce an auxiliary Model M_a to link M_1 and M_2:

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \mathbf{\Lambda}^{(g)} \boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} = \delta_i^{(g)}. \end{cases}$$

We first estimate the Bayes factor between Model 1 (M_1) and Model a (M_a) via the path sampling method.

Construct a link model, Model t1a ($\mathrm{M}_{\mathrm{t1a}}$)

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} = t \cdot \boldsymbol{\Gamma}^{(g)}\boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}, \end{cases}$$

and notice that $\mathrm{M}_{\mathrm{t1a}}$ reduces to M_1 when $t = 1$, while $\mathrm{M}_{\mathrm{t1a}}$ reduces to M_a when $t = 0$. Consider the grids from 0 to 1 with step 0.05 , and on each fixed grid $t_{(s)}$, generate observations

$$\left(\boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}\right) \sim p\left(\boldsymbol{\Omega}, \boldsymbol{\theta} \mid \boldsymbol{Y}, t_{(s)}\right), \quad j = 1, \cdots, J = 1000,$$

after 500 warm-up iterations using MCMC methods.

Then calculate $U\left(\boldsymbol{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}\right)$ by

$$\begin{aligned} U(\boldsymbol{Y}, \boldsymbol{\Omega}, \boldsymbol{\theta}, t) &= \frac{d}{dt} \log p(\boldsymbol{\Omega}, \boldsymbol{\theta} \mid \boldsymbol{Y}, t)\bigg|_{t=t_{(s)}} \\ &= \sum_{g=1}^{2}\sum_{i=1}^{n_g} \frac{1}{\psi_\delta^{(g)}} \left(\eta_i^{(g)} - t_{(s)} \cdot \gamma_1^{(g)}\xi_{i1}^{(g)} - t_{(s)} \cdot \gamma_2^{(g)}\xi_{i2}^{(g)}\right)\left(\gamma_1^{(g)}\xi_{i1}^{(g)} + \gamma_2^{(g)} \right. \end{aligned}$$

then

$$\bar{U}_{(s)} = \frac{1}{J}\sum_{j=1}^{J} U\left(\boldsymbol{Y}, \boldsymbol{\Omega}^{(j)}, \boldsymbol{\theta}^{(j)}, t_{(s)}\right).$$

and then the Bayes factor between M_1 and M_a is estimated by

Similarly, we construct link model, Model t2a ($\mathrm{M}_{\mathrm{t2a}}$),

$$\begin{cases} \boldsymbol{y}_i^{(g)} = \boldsymbol{\mu}^{(g)} + \boldsymbol{\Lambda}^{(g)}\boldsymbol{\omega}_i^{(g)} + \boldsymbol{\epsilon}_i^{(g)}, \\ \eta_i^{(g)} = t \cdot \boldsymbol{\Gamma}\boldsymbol{\xi}_i^{(g)} + \delta_i^{(g)}, \end{cases}$$

The Bayes factor between M_1 and M_2 is estimated by

$$\widehat{\log}_{12} = \log \widehat{\mathrm{BF}}_{1a} - \log \widehat{\mathrm{BF}}_{2a}.$$

The estimated $\widehat{\log}_{12}$ is equal to 3.25.

5. Conclusion

Based on M_1, we can conclude that

- From λ_{52} to λ_{82}, fruit and veggies consumption, and physical activity decrease the risk related to lifestyle factors for both low income group and high income group. Smoking increase the risk related to lifestyle factors for both groups.

- From $\lambda_{10,3}$ to $\lambda_{12,3}$, higher BMI, stroke history, and heart disease or attck history increase the risk related to physical condition factors for both two groups. However, for low income group, this influence is amplified compared to the high income group.

- $\gamma_1^{(1)} << \gamma_1^{(2)}$ indicates the influences of the risk related to lifestyle factors on the diabetes risk are very different between the two groups. For low income group, the risk related to lifestyle factors mitigates the diabetes risk. For high income group, the risk related to lifestyle factors amplifies the diabetes risk.

- $\gamma_1^{(1)} > \gamma_1^{(2)}$ indicates that the effect of the risk related to physical condition factors on the diabetes risk in low income group is greater than this effect in high income group. However, the effect power in the two group are close.

- Hence, our final conclusion is that (i) High income people are more likely to have diabetes if they have the same unhealthy lifestyles than the low income people. (ii) High income people are slightly less likely to have diabetes if they have the same physical conditions and disease histories than the low income people.

- Based on the estimate of Bayes factor between M_1 and M_2, since $2\log\widehat{\mathrm{BF}}_{12} = 3.26$, then M_1 is better than M_2, the data provide positive evidence to support the alternative hypothesis $H_1 : \Gamma^{(1)} \neq \Gamma^{(2)}$, i.e. the data

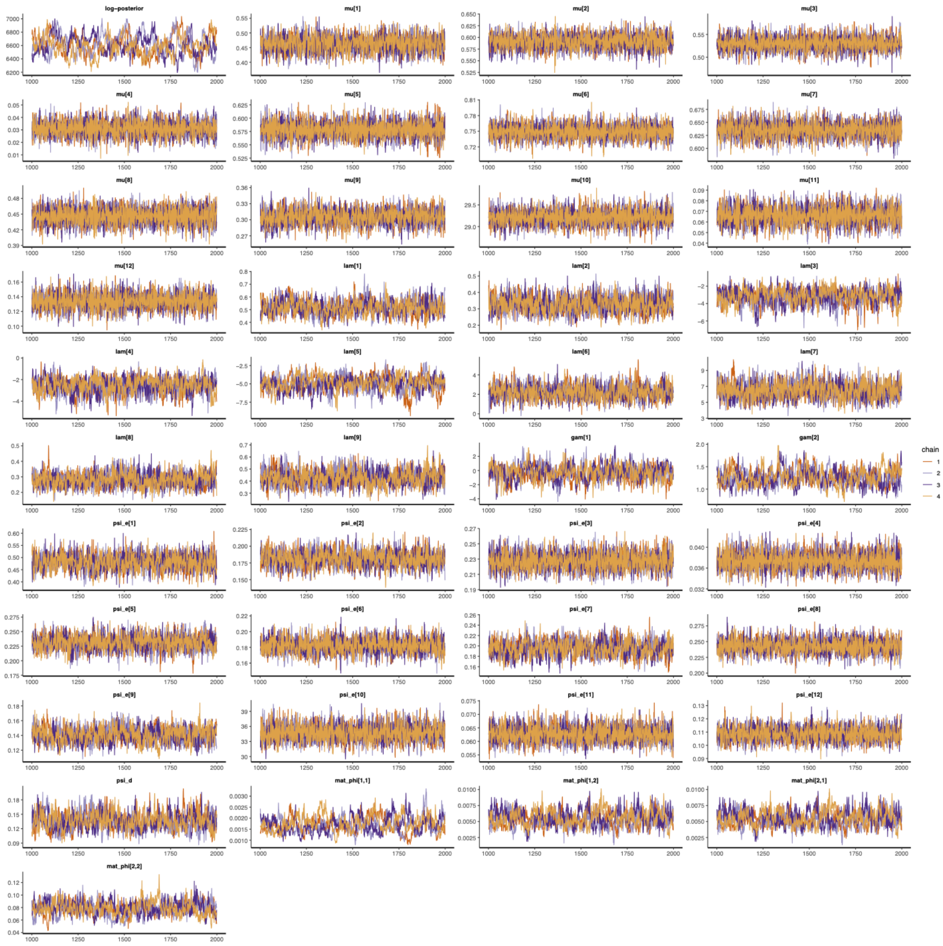

FIGURE 1. Sample Traces Plot of M_0

supports our main model M_1. Hence, we can conclude that the coefficients in the structural equation in different group are different.

REFERENCES

[1] N.R. Burrows, I. Hora, L.S. Geiss, E.W. Gregg, and A. Albright. Incidence of end-stage renal disease attributed to diabetes among persons with diagnosed diabetes — united states and puerto rico, 2000–2014. *MMWR Morb Mortal Wkly Rep*, 66:1165–1170, 2017. `http://dx.doi.org/10.15585/mmwr.mm6643a2`.

[2] Centers for Disease Control and Prevention. CDC Diabetes Health Indicators Dataset. `https://archive.ics.uci.`

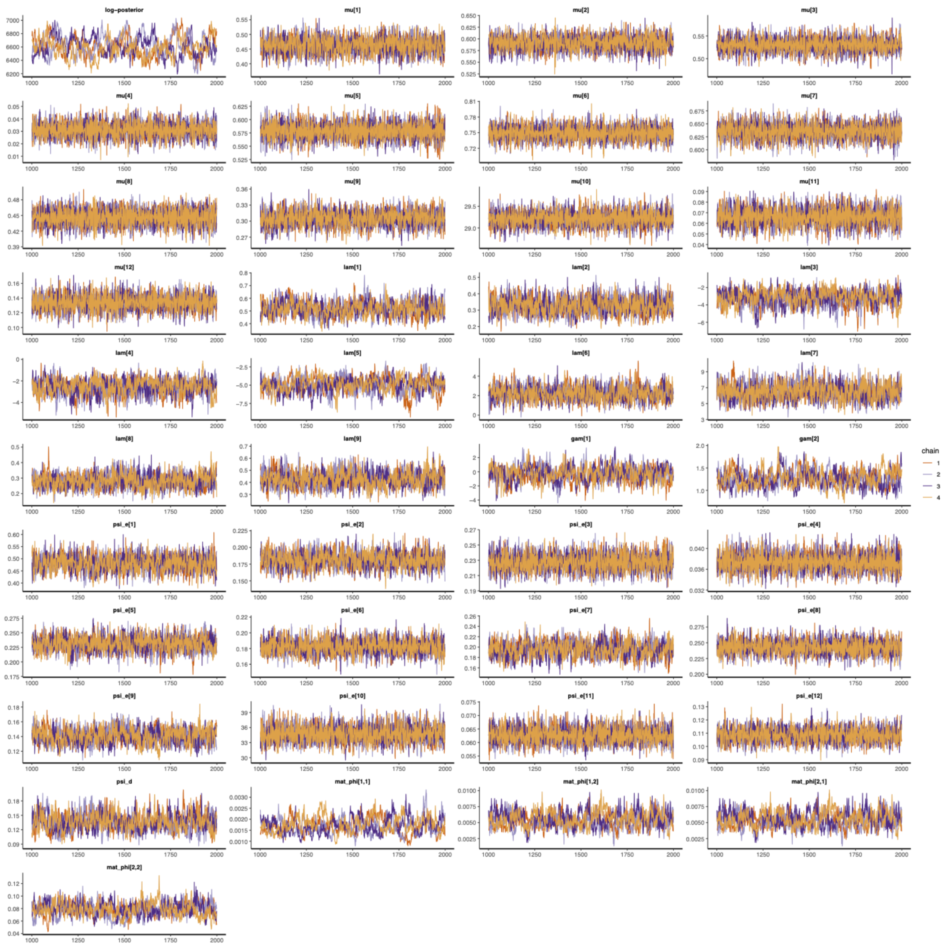

FIGURE 2. Sample Traces Plot of M_1

`edu/dataset/891/cdc+diabetes+health+indicators`, 2023. Accessed: 2023-05-DD.

Table 1. Parameter Estimates with Mean and Standard Deviation

Parameter	Mean	SD	Parameter	Mean	SD
μ_1	0.4633	0.0268	γ_1	0.5896	0.6963
μ_2	0.591	0.0158	γ_2	0.2856	0.0774
μ_3	0.5299	0.0159	ψ_{e1}	0.4842	0.0334
μ_4	0.7491	0.0138	ψ_{e2}	0.1815	0.0115
μ_5	0.5778	0.0153	ψ_{e3}	0.2264	0.011
μ_6	0.0309	0.0063	ψ_{e4}	0.1854	0.0086
μ_7	0.6361	0.0152	ψ_{e5}	0.2385	0.0116
μ_8	0.4458	0.0161	ψ_{e6}	0.0375	0.0016
μ_9	29.2168	0.182	ψ_{e7}	0.1826	0.0137
μ_{10}	0.305	0.0148	ψ_{e8}	0.2465	0.011
μ_{11}	0.0659	0.0083	ψ_{e9}	35.589	1.6914
μ_{12}	0.1341	0.0109	ψ_{e10}	0.1364	0.01
λ_{21}	0.5186	0.059	ψ_{e11}	0.0627	0.003
λ_{31}	0.3348	0.0506	ψ_{e12}	0.1087	0.0053
λ_{52}	1.1698	0.3583	ψ_d	0.1379	0.0186
λ_{62}	0.0342	0.1032	$\Phi_{1,1}$	0.0075	0.0023
λ_{72}	2.7931	0.5927	$\Phi_{1,2}$	-0.0737	0.0209
λ_{82}	-0.811	0.3399	$\Phi_{2,1}$	-0.0737	0.0209
$\lambda_{10,3}$	0.2214	0.0515	$\Phi_{2,2}$	1.8944	0.7955
$\lambda_{11,3}$	0.0608	0.0164			
$\lambda_{12,3}$	0.0914	0.0244			

TABLE 2. Posterior Mean of Parameters for Group 1 and Group 2

Parameter	Group 1	Group 2	Parameter	Group 1	Group 2
μ_1	0.5799	0.3971	γ_1	-0.3572	0.495
μ_2	0.699	0.5312	γ_2	1.4313	1.3355
μ_3	0.5754	0.5049	ψ_d	0.2397	0.1661
μ_4	0.0084	0.0433	ψ_{e1}	0.4775	0.3852
μ_5	0.5948	0.5685	ψ_{e2}	0.1875	0.2042
μ_6	0.6527	0.8029	ψ_{e3}	0.2456	0.2317
μ_7	0.5789	0.6681	ψ_{e4}	0.0306	0.0524
μ_8	0.4269	0.4562	ψ_{e5}	0.2262	0.2359
μ_9	0.4758	0.2095	ψ_{e6}	0.221	0.1623
μ_{10}	29.3582	29.1233	ψ_{e7}	0.2395	0.186
μ_{11}	0.1096	0.042	ψ_{e8}	0.2474	0.245
μ_{12}	0.1682	0.1146	ψ_{e9}	0.1964	0.1183
λ_{21}	0.3219	0.4413	ψ_{e10}	37.215	31.1068
λ_{31}	0.1712	0.3071	ψ_{e11}	0.1079	0.0489
λ_{52}	-1.0168	-2.6068	ψ_{e12}	0.1423	0.1029
λ_{62}	-0.6502	-1.3539	$\Phi_{1,1}$	0.0019	0.0014
λ_{72}	-0.997	-4.3268	$\Phi_{1,2}$	0.0014	0.0747
λ_{82}	1.0257	1.9918	$\Phi_{2,1}$	0.0025	0.0061
$\lambda_{10,3}$	9.569	6.4879	$\Phi_{2,2}$	0.0061	0.0607
$\lambda_{11,3}$	0.3292	0.2082			
$\lambda_{12,3}$	0.4497	0.3884			