Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 1s (2025)

Modeling GHQ Total and SDQ Difficulty Score to Extract Incomplete Information on One Using the Other

Alka Sabharwal[1], Babita Goyal[2*], Lalit Mohan Joshi[3]
[1]Professor, Department of Statistics, Kirori Mal College, University of Delhi, Delhi-110007
Email: alkasabh@gmail.com
[2] Professor, Department of Statistics, Ramjas College, University of Delhi, Delhi-110007
Email: goyalbabita@gmail.com
[3]Research scholar, Department of Statistics, University of Delhi, Delhi-110007
Email: lalitjstats12@gmail.com *- Corresponding author

Article History:

Received: *05-08-2024*

Revised: *25-09-2024*

Accepted: *07-10-2024*

Abstract:

Background: A natural approach to analyze multidimensional data is use of multivariate statistical analysis. If the number of dimensions/variables is small and variables are correlated, Multivariate Normal Distribution (MVN) is applied frequently. If the distribution of underlying variables is not normal, transformations are applied to convert them to normal variables.

Objective: Many a times, information may be missing on one or more dimensions of the data. The study intended to estimate the missing information through the available information.

Method: In a series of three independent surveys to examine the psychological health of young adults during COVID-19 period, enrolled in higher educational institutions in India, Strength and Difficulty Questionnaire (SDQ) 17+ extended version was used. In addition, General Health Questionnaire (GHQ) was used in third survey. The data was divided into two datasets; based on third survey and based on first two surveys. MVN was used to estimate GHQ scores through difficulty score dimension of SDQ and vice versa. The model was applied to data of first two surveys to estimate GHQ scores at the time of these surveys. The model was applied on 162 respondents who were common in all the three surveys.

Result: The estimated values for the third survey data were consistent with the observed and simulated values. Further it was found that out of 64 respondents with high GHQ scores in third survey, 55 had it during first two surveys also.

Conclusion: The results can be extended to estimate any missing information whenever variables are correlated.

Keywords: Bivariate normal distributions, General Health Questionnaire, Psychological health, Strength and Difficulty Questionnaire, Weibull distribution.

1 Introduction

Any medical data, whether on an individual or on a cohort is multidimensional in nature, where the number of dimensions can be large. To analyze such data, a natural approach is to make use of multivariate statistical analysis. A dimension refers to specific type information collected on a subject, through variables; such as subject dimension includes information about the patient (name, age, gender etc.) Similarly disease dimension included duration of the disease; severity, comorbidities etc. If the number of variables used to collect the information is large, then the commonly used statistical techniques are General Linear Models, Discriminant Analysis, Factor Analysis, both for dimension reduction and analysis of the data (Johnson and Kotz, 1972; Khattree and Naik 2000; Lindsey, 2000;

Chi, 2012). If the number of variables is small (less than four), parametric multivariate methods are frequently applied. One of such methods is use of Multivariate Normal Distribution (MVN).

Multivariate normal distribution is used in case when the different variables of the data are correlated. A MVN involves the correlation matrix of the variables along with their marginal distributions. Although the MVN requires the marginal distributions to be normal, this condition is not met in general. Transformations are then applied to convert the non- normal variable(s) to normal variable(s). The most commonly used transformations are the Box-Cox transformation and the power transformations. If the number of variables is two, the MVN distribution reduces to a bivariate normal distribution (BVN). Lipow and Eidemiller applied BVN distribution to study the relationship between stress and strength in the reliability study problem (Lipow and Eidemiller, 1964). Yue applied BVN distributions to solve problems of hydrological engineering design and management (Yue, 1999). Grover et al. applied MVN and BVN distributions to estimate the duration of diabetes on the basis of Low-density Lipoprotein (LDL), Fasting Blood sugar (FBG) and systolic blood pressure (SBP) (Grover et al., 2014). In another study, Grover et al. applied MVN for estimating the length of stay (LOS) in the hospital, duration of disease and severity of the disease on a group of 146 inpatients diagnosed with mental and behavioural problems (Grover et al., 2015).

Psychological health of an individual is essentially assessed with the help of questionnaires which are multidimensional in nature. Sometimes more than one questionnaire is needed to have a comprehensive view of the psychological health of the individual. Purpose of using more than one questionnaire is to cross-validate the data collected through the different questionnaires as well as to complement the missing information, if any. In fact, a judicious choice of questionnaires not only ensures the consistency in the responses but also can be used to overcome the shortcoming of any questionnaire. Questionnaires are designed keeping in mind the target group under observation as well as the purpose of study/observation. In this study, the authors used the General Health Questionnaire (GHQ), and the **Strength and Difficulties Questionnaire** (SDQ) 17^{+} extended version.

The GHQ, developed by Dr. David Goldberg in 1970, is a commonly used, self-administered screening tool designed to detect current state mental disturbances and the health of the last 4-6 weeks (Goldberg and Hillier, 1979). It consists of 12 items that assess various aspects of an individual's mental health, including mood, anxiety, and social functioning. The SDQ, developed by Robert Goodman in 1997, is used to assess strength and difficulties among children and young adults (Goodman, 1997). It has two versions: the basic version and the extended version.

A series of three independent surveys was conducted during COVID-19 pandemic, from May–June 2020 to January–February 2022. Each of the survey was conducted after occurrence of some game-changing event. The first survey was held during May–June 2020, when the lockdown was freshly imposed. The second survey was held in October 2020–February 2021, when the first COVID-19 wave had almost subsided. However, soon after the deadly ‘delta’ wave had struck, causing a huge havoc on the society. The large number of morbidities and mortality severely affected the mental health of people at large. In this light, we conducted the third survey, during the months of January–February 2022. While all the three surveys were conducted using the SDQ in order to measure the psychological health of the young adults, in the third survey we used the GHQ-12 additionally, a questionnaire which is used to measure the recent (up to 4 weeks) psychological health. Whereas SDQ is students’ based questionnaire, GHQ-12 is widely used for every age group.

This study was initiated on the basis of the data collected during the third survey as this data contained information collected through SDQ as well as GHQ-12. The objective of this study was to examine if it is possible to estimate one score from the other through appropriate modeling. For this objective, we intended to obtain a relationship (if any) between the information obtained through the two

questionnaires (Difficulty score of SDQ and total score of GHQ-12, based on Likert scale 0-1-2-3). In order to achieve this objective, we selected appropriate distributions (on the basis of AIC and BIC criteria) to the two components of the data. The parameters of the selected distributions were estimated through the method of Maximum Likelihood Estimation (MLE). Through power transformations, both the components were transformed to follow normal distributions. Mardia test was applied to test the multivariate normality. On the basis of the outcome of the test, a bivariate normal distribution was fitted using the two correlated normal variables. Using the parameters of the transformed distributions (fitted to the Difficulty score of SDQ and total score of GHQ-12) and the correlation between them, we simulated 10000 values each of

(i) Marginal distribution of the Difficulty score

(ii) Marginal distribution of GHQ-12 total

(iii) Bivariate normal distribution

We used the data from survey 3 (Dataset_1) in order to estimate Difficulty score of SDQ through total score of GHQ and vice versa. The results were validated through simulation. The model was then used on the data of the first two surveys (Dataset_2, containing information on Difficulty score only) to estimate the GHQ-12 scores. The model was also validated by applying it on 162 respondents who had participated in all the three surveys.

Novelty of the study is establishment of the link between two different questionnaires, used under different circumstances through modeling. The results can be used to complete any missing piece of information. To the best of our knowledge, this is the first study to make use of real life data to link two different questionnaires.

The rest of this paper along with the introduction is as follows. In section 2, the materials and developed model are discussed. In section 3, the results are described. The paper concludes with discussion in section 4, discussion and conclusion in section 5.

2 Materials and Methods

2.1 Materials

During COVID-19 pandemic times, in a period of almost two years (From May-June 2020 to January-February 2022), three independent surveys were conducted on the young adults studying in higher educational institutions across India, using SDQ 17+ extended version. The number of responses obtained were respectively 1020, 743 and 934 respectively. Although the surveys were held independent of each other, 162 respondents were found to have participated in all the three surveys. In survey 3, GHQ-12 questionnaire was also used along with SDQ 17+ extended version. The scoring methods of SDQ 17+ extended version have been discussed in the earlier literature (Goodman, 1999; Goyal et al., 2023; Sabharwal et al., 2023).

The GHQ-12 questionnaire contains 12 items which are evaluated using the Likert (0-1-2-3/ 0-0-1-1) scale. The scores are then added to obtain the total GHQ scores. The scores lie in the range 0-36/0-12. Higher scores indicate a higher level of psychological distress or impaired mental well-being, although there are no clear-cut categories describing the severity of problems as is in case of SDQ scores (Goldberg, 1979; Goldberg et al., 1997; Anjara S. et al., 2020). The inclusion criteria for the study were the respondents enrolled in higher educational institutions and were participating willingly.

2.2 Methods

2.2.1 Bivariate Normal Distribution (BVN)

If two random variables X and Y are following normal distributions $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ respectively, with correlation coefficient ρ, the joint probability density function (pdf) of x and y is given by (Johnson and Kotz, 1972),

$$f(x,y) = \frac{\exp\left[\frac{-1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right\}\right]}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \tag{1}$$

$$-\infty < (x,y) < \infty, -1 < \rho < 1, (\sigma_x, \sigma_y) > 0$$

The conditional expectation for BVN distribution is

$$\mathrm{E}(x \mid y) = \mu_x - \frac{\sigma_{xy}}{\sigma_x}(y - \mu_y) \tag{2}$$

where, σ_{xy} is covariance between x; $-\infty < \sigma_{xy} < \infty$.

2.2.2 Distribution of Variables

The following distributions namely, Normal, Gamma, Weibull, Log-normal, and Exponential, were fitted to the data and the best fitted distribution was selected on the basis of minimum AIC and BIC values. Table **1** presents the pdf of all the fitted distributions in this study.

Table 1 The fitted distributions for present study

Name of distributions	Probability density function (pdf)	Name of parameters	Ranges
Normal	$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ = Mean and σ = Standard deviation	$-\infty < (x,\mu) < \infty, \sigma > 0$
Gamma	$f(x,\lambda,\gamma) = \frac{e^{-(x/\lambda)}x^{(\gamma-1)}}{\Gamma\gamma\ \lambda^{\gamma}}$	γ = Shape parameter and λ = Scale parameter	$(x,\lambda,\gamma) > 0$
Weibull	$f(x,\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp(-(x/\lambda)^k)$	k = Scale parameter and λ = Shape parameter	$x > 0$
Log-normal	$f(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\ln x-\mu)^2}{2\sigma^2}\right)$	μ = Mean log and	$x > 0$

		σ = Standard deviation log	
Exponential	$f(x,\theta)=\theta e^{\theta x}$	θ = Rate parameter.	$x \geq 0$

2.2.3 Transformations

Transformations are a group of statistical techniques, used to convert the data to a form which can be dealt with available standard methods. The most commonly used transformations are the square root, the logarithms, and the reciprocal (http://www-users.york.ac.uk/~mb55/msc/clinbio/week5/transfm_gif.pdf). For the bivariate data used in this study, the assumed distribution is a bivariate normal distribution (BVN) for which the marginal distributions should also be normal. Power transformations have been applied to the selected distributions to convert them to a normal form. The form of power transformation is (http://www.statsref.com/HTML/index.html?freeman-tukey.html).

$$z=(x+a)^{\lambda},\ \lambda>0, a>0, x>0 \tag{3}$$

where a is an optional constant.

2.2.4 Mardia Test (Mardia K., 1970; Von Eye and Bogat, 2004)

Mardia test is used to examine the multivariate normality of a data. It makes use of skewness and kurtosis measurements, given by

$$\hat{\alpha}_{1,p}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}t_{ij}^3 \quad \text{and} \quad \hat{\alpha}_{2,p}=\frac{1}{n^2}\sum_{i=1}^{n}t_{ii}^2$$

where, $X_1, X_2, \ldots, X_n$ are a vector of size $1 \times p$, $t_{ij}=(x_i-\bar{x})' S^{-1}(x_j-\bar{x})$

$$S=\frac{1}{n}\left[\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})'\right] \quad \text{and} \quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

and, p is the number of variables.

The test statistics for skewness and kurtosis are, $\kappa_1=\frac{n\hat{\alpha}_{1,p}}{6}\underset{approx}{\sim}\chi^2_{p(p+1)(p+2)/6}$ and $\kappa_2=\frac{\left(\hat{\alpha}_{2,p}-p(p+2)\right)}{\sqrt{\left(8p(p+2)/n\right)}}\underset{asymp}{\sim}N(0,1)$ respectively.

2.2.5 Criteria for Model Selection

The most popular information criteria for choosing a model are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), given by:

$$AIC=-2\log(L)+2p \tag{4}$$

$$BIC=-2\log(L)+\log(n).p \tag{5}$$

where, L is the likelihood under the fitted model, p is the number of parameters, and n is number of total observations. The smallest value of AIC or BIC give best fit model for data (Kuha, 2004; Bradman M. et al., 2003).

2.2.6 Algorithm used in the Study

The following flow chart describes the algorithm used in this study.

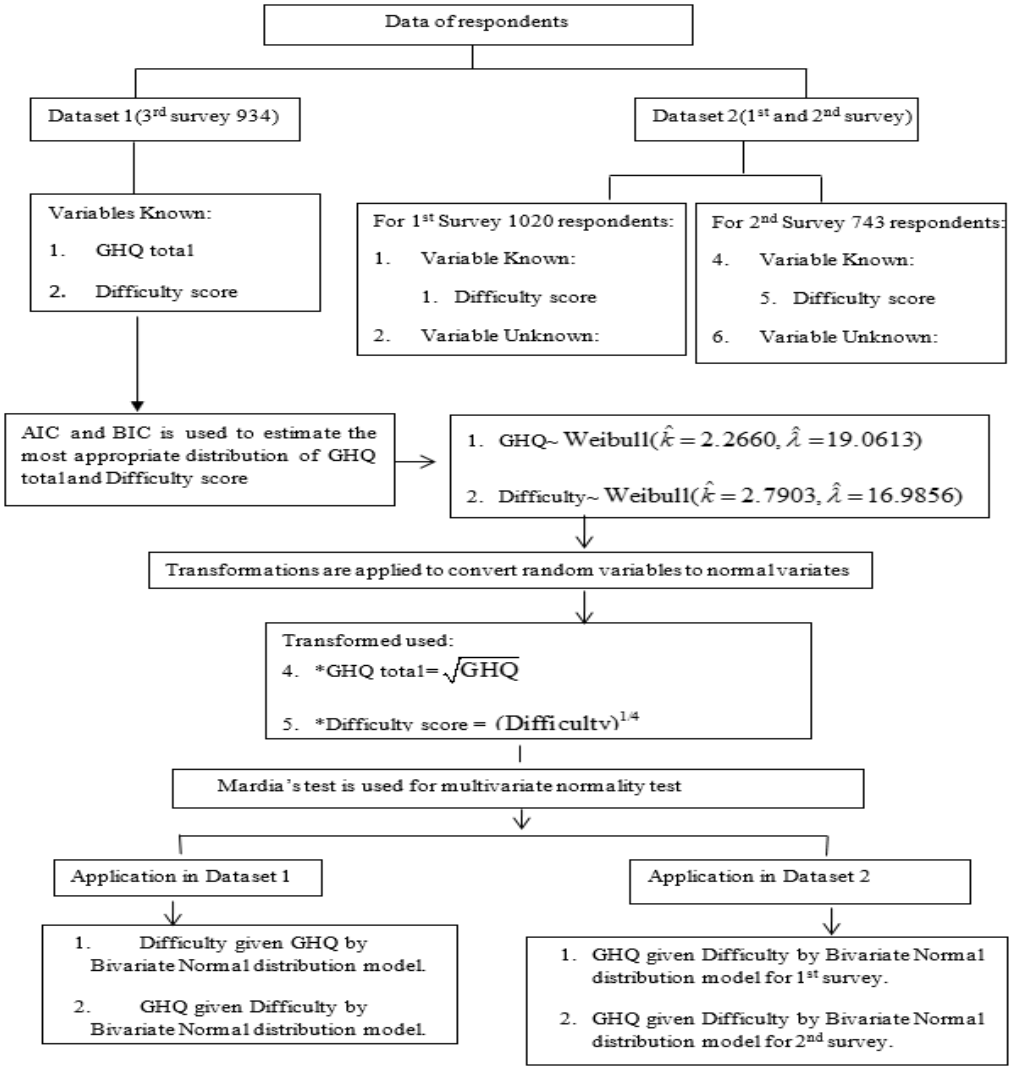

Fig. 1 Algorithm of the modeling and its application

3 Results

3.1 Data description

Table **2** presents the descriptive statistics of 934 respondents for Dataset_1(Survey 3 data). The mean and standard deviation value of GHQ total is 16.82 and 7.98 respectively. The mean and standard deviation value Difficulty score is 15.11 and 5.90 respectively. The maximum value of Difficulty score is 31 and that of GHQ total is 36.

Table 2 Descriptive statistics of variables: GHQ total and Difficulty score for Dataset_1(Survey 3 data)

Statistic	GHQ total	Difficulty score
Total	934	934
Minimum	0	1
Maximum	36	31
Range	36	30
Mean	16.82	15.11
Standard deviation	7.98	5.90

Table **3** presents the descriptive statistics of 1020 and 743 respondents giving minimum; maximum, range, mean, and standard deviation for Dataset_2 (survey 1 and survey 2 data). The mean value of Difficulty score is 13.64 and 12.73 for 1st and 2nd survey respectively. The maximum value of Difficulty score for 1st and 2nd survey is 31 and 29 respectively.

Table 3 Descriptive statistics variable; Difficulty score for Dataset_2 (survey 1 and survey 2 data)

Statistic	Difficulty score for 1st survey	Difficulty score for 2nd survey
Total	1020	743
Minimum	2	2
Maximum	31	29
Range	29	27
Mean	13.64	12.73
Standard deviation	5.35	5.17

3.2 Distribution Selection for GHQ total and Difficulty scores

The Table **4** below presents the AIC and BIC values of the fitted distributions on Dataset_1. On the basis of the least AIC and BIC criteria, Weibull distribution (with computed parameters) has been found to be the most appropriate fir for both the variables.

For GHQ total, Weibull distribution is obtained with parameters, the shape parameter ($\hat{k}$) is 2.266015 and the scale parameter ($\hat{\lambda}$) is 19.06131. For the Difficulty score, the Weibull distribution has the parameter shape ($\hat{k}$) 2.790381 and the scale ($\hat{\lambda}$) 16.98565.

Table 4 AIC and BIC values of different distributions for GHQ total and Difficulty score

Variable	Distribution	AIC values	BIC values	Selected distribution	MLE of the parameters
GHQ Total	Normal	6514.19	6523.87	Weibull	$\hat{k}$=2.266015
	Gamma	6488.11	6497.79		
	Weibull	**6447.84**	**6457.51**		

	Lognormal	6592.33	6602.005		$\hat{\lambda}$ = 19.06131
	Exponential	7131.75	7136.59		
Difficulty score	Normal	5970.19	5979.87	Weibull	$\hat{k}$ =2.790381 $\hat{\lambda}$ = 16.98565
	Gamma	6006.99	6016.67		
	Weibull	**5945.73**	**5955.41**		
	Lognormal	6117.06	6126.74		
	Exponential	6942.85	6947.69		

3.3 Transformations Applied on GHQ total and Difficulty score

The square root transformation has been applied on GHQ total to obtain an approximately normal *GHQ total and the fourth root transformation has been applied on Difficulty score to obtain an approximately normal *Diff score. The normality of the transformed variables *GHQ total and *Diff score has been justified by plotting the pdfs and the Q-Q plots of the transformed variables in Fig. **2(a)**, **2(b)**, **3(a)** and **3(b).**

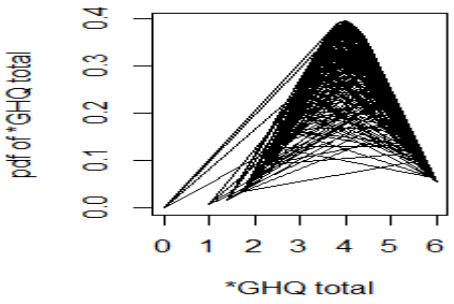

Fig. 2(a) Normal approximation of *GHQ total

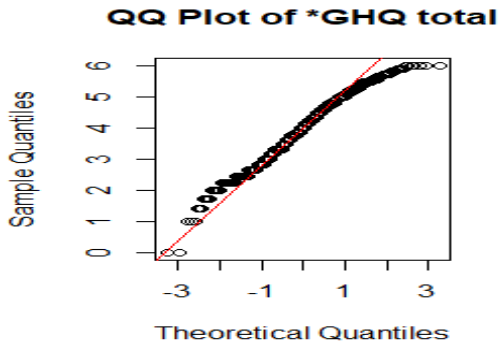

Fig. 2(b) QQ-plot for *GHQ total

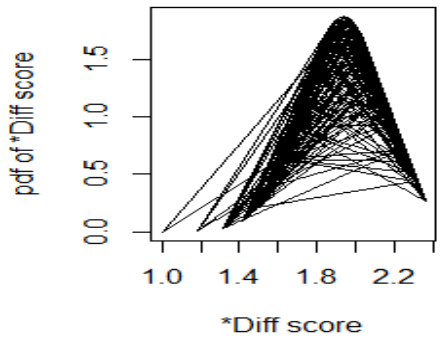

Fig. 3(a) Normal approximation of *Diff score

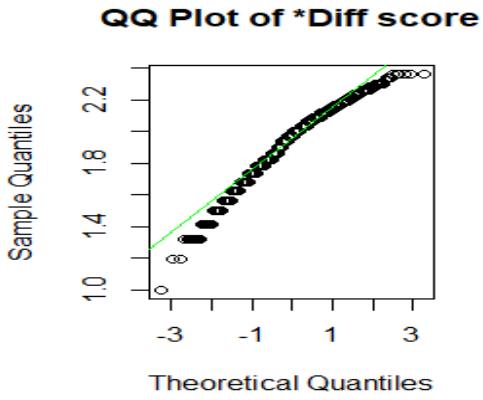

Fig. 3(b) QQ-plot for *Diff score

3.4 Checking Multivariate Normality of *GHQ total and *Diff score

Mardia test has been applied on *GHQ total and *Diff score to test the multivariate normality. Skewness and kurtosis of the bivariate data are κ_1 =2.4134 (at 4 degrees of freedom with *p-value* = 0.6601 > 0.05), and κ_2 = -0.3640 (with *p-value* =0.7158 > 0.05) respectively. Hence, the *GHQ total and *Diff score jointly follow BVN (3.971, 1.939, 1.05598, 0.04516, 0.49997).

3.5 Estimating *GHQ total through *Diff score and vice versa using Dataset_1 through BVN

The joint distribution of variables *GHQ total and *Diff score is BVN, given by

$$\begin{pmatrix} *GHQ\ \text{total} \\ *Diff\ \ score \end{pmatrix} \sim BVN\left(\mu = \begin{pmatrix} \mu_{*GHQ\ \text{total}} \\ \mu_{*Diff\ \ score} \end{pmatrix}, \Sigma = \begin{bmatrix} \sigma_{*GHQ\ \text{total}} & \sigma_{*GHQ\ \text{total},*Diff\ \ score} \\ \sigma_{*GHQ\ \text{total},*Diff\ \ score} & \sigma_{*Diff\ \ score} \end{bmatrix} \right) \quad (6)$$

3.5.1 Generating the BVN population corresponding to transformed variables *GHQ total and *Diff score through simulation

Equation (6) is valid for population parameters. In order to generate the BVN population parameters, we conducted a simulation study as follows:

1. Distributions of GHQ total and Difficulty score were selected on the basis of minimum AIC and BIC value.

2. 10000 values of GHQ total were generated using the selected Weibull (2.2660,19.0613) distribution.

3. 10000 values of Difficulty score were generated using the selected Weibull (2.7903, 16.9856) distribution.

4. We applied the same power transformation as were applied on the real data to obtain the simulated transformed normal variables (^{sim} GHQ total and ^{sim} Diff score).

5. The parameters of the transformed variables along with their correlation coefficient were used to generate the bivariate normal population of size 10,000.

6. The values were put in equation (6).

3.5.2 Comparison of data based Mean Difficulty score with estimated Difficulty score as obtained through the BVN model

The following procedure is adopted to obtain the mean Difficulty score for various categories of GHQ total:

1. The intervals of *GHQ total (after transformation) range is defined as $1 < GHQ \geq 2$, $2 < GHQ \geq 3$, $3 < GHQ \geq 4$, $4 < GHQ \geq 5$, and $5 < GHQ \geq 6$; identified as GHQ_i; i = 1,2,3,4,5 respectively.

2. Take the first range of GHQ. Corresponding to this range, compute the mean GHQ from Dataset_1. Fixing this range of GHQ, compute the mean Difficulty score i.e., by taking only those Difficulty scores, corresponding to which GHQ lies in this range.

3. Repeat (2.) for all the ranges of GHQ, thus yielding all the mean GHQ totals and the corresponding mean Difficulty scores.

4. The same procedure (for the same GHQ total ranges) is applied for the simulated data.

5. Next, taking GHQ total from the sample data, the conditional expectation of Difficulty score for each GHQ category (GHQ_i, i = 1,2,3,4,5) has been computed using equation (7) where all the other parameters are population based (simulated data)

$$E_i(Diff.|GHQ) = \mu_{i,Diff.} - \frac{1}{\sigma_{i,Diff.}}\sigma_{i.GHQ,Diff.}\left(GHQ_i - \mu_{i,GHQ}\right) \quad (7)$$

6. The observed mean values of Diff score corresponding to each range of GHQ total, is compared with the estimated values obtained through equation (7).

7. Step (6) is used again after retransforming the transformed variables to the original variables.

The results are presented in Table **5** below:

Table 5 Estimated mean *Diff score given *GHQ total for 934 respondents for different interval of *GHQ total using a generated random sample of size 10000 for BVN distribution

Interval	Real Data		Simulated Data		E[*Diff. | *GHQ]	Observed difficulty score	Estimated difficulty score
	Mean *GHQ	Mean *Diff.	Mean *GHQ	Mean *Diff			
1<GHQ≥2	1.8279	1.3930	1.6794	1.4580	1.3313	3.7653	3.1413
2<GHQ≥3	2.6513	1.6924	2.6159	1.6552	1.6329	8.2037	7.1095
3<GHQ≥4	3.6000	1.8959	3.5307	1.8467	1.8506	13.0208	11.2787
4<GHQ≥5	4.5768	2.0663	4.4482	2.0363	2.0076	18.2308	16.2446
5<GHQ≥6	5.4316	2.2708	5.3821	2.3097	2.2431	26.5898	25.3159

The observed and estimated Difficulty scores are compared in Fig. **4** below:

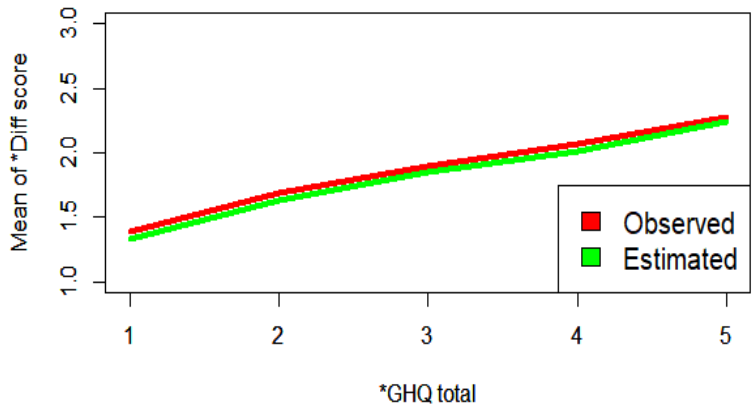

Fig. 4 Comparison of observed and estimated Difficulty score for Dataset_1

3.5.3 Comparison of data based Mean GHQ total with estimated Mean GHQ total as obtained through the BVN model

Categorizing the *Diff score (after transformation) range as 1< Diff ≥ 1.25, 1.25 < Diff ≥ 1.5, 1.5 < Diff ≥ 1.75, 1.75 < Diff ≥2, 2 < Diff ≥ 2.25, and 2.25 < Diff ≥ 2.5, identified as $Diff_i$: i = 1,2,3,4,5,6, the steps (2)-(6) in 3.5.2 are repeated to obtain the corresponding GHQ scores for the real and simulated

data. For estimated values for the respective ranges, equation (8) has been used where except for $Diff_i$, which is based on the real data, rest all have been taken from the population (simulated data):

$$\mathrm{E}_i(GHQ|\,Diff.) = \mu_{i,GHQ} - \frac{1}{\sigma_{i,GHQ}}\sigma_{i,Diff.,\mathrm{GHQ}}\left(Diff_i - \mu_{i,Diff.}\right) \tag{8}$$

The results are presented in Table **6** below:

Table 6 Estimated mean *GHQ total given *Diff score for 934 respondents for different intervals of *Diff score using a generated random sample of size 10000 for BVN distribution

Interval	Real Data		Simulated Data		E[*GHQ total |*Diff score]	Observed mean GHQ total	Estimated Mean GHQ total
	Mean *GHQ	Mean *Diff.	Mean *Diff.	Mean *GHQ			
1<Diff≥1.25	2.1019	1.1261	1.2885	1.8800	2.1034	4.4180	4.4243
1.25< Diff ≥150	2.6678	1.4284	1.4276	2.5538	2.5538	7.1171	6.5219
1.50<Diff ≥1.75	3.2810	1.6667	1.6588	2.6579	2.9532	10.7649	8.7214
1.75< Diff ≥2	3.8577	1.8947	1.8826	3.7115	3.7055	14.8818	13.7307
2< Diff ≥2.25	4.4266	2.1122	2.1038	4.7563	4.4027	19.5947	19.3838
2.25<Diff ≥2.50	5.8756	2.2923	2.3317	5.8411	5.8568	34.5227	34.3021

The observed and estimated GHQ total scores are compared in Fig. **5** below:

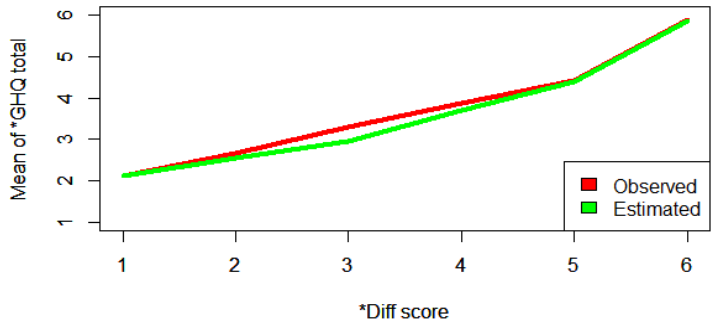

Fig. 5 Comparison of observed and estimated GHQ total for Dataset_1

3.5.4 Estimation of Mean GHQ total for Dataset_2

The results obtained through the simulation and Dataset_1 are applied to the Dataset_2 (survey 1 and survey 2 data) where we just had the Difficulty scores and not the GHQ scores. Similar to what we did for Dataset_1, appropriate distributions for Difficulty scores of survey 1 and survey 2 data were obtained. The distributions were Weibull (2.74574, 15.2954) for survey 1; and Weibull (2.64852, 14.3518) for survey 2 Difficulty scores (Table **7**).

Table 7 AIC and BIC values of different distributions for Difficulty score, for survey 1 and survey2

Variable	Distribution	AIC values	BIC values	Selected distribution	MLE of the parameters
Difficulty score (Survey 1)	Normal	6319.12	6328.98	Weibull	$\hat{k}$=2.74574 $\hat{\lambda}$ = 15.2954
	Gamma	6258.82	6268.68		
	Weibull	**6254.20**	**6264.06**		
	Lognormal	6314.92	6324.78		
	Exponential	7365.21	7370.14		
Difficulty score (Survey 2)	Normal	4554.51	4563.73	Weibull	$\hat{k}$=2.64852 $\hat{\lambda}$ = 14.3518
	Gamma	4521.42	4530.65		
	Weibull	**4518.56**	**4527.78**		
	Lognormal	4571.42	4580.64		
	Exponential	5269.36	5273.97		

We generated 10000 values from each of the two Weibull distributions and applied the (fourth root) power transformation to the generated and the real data. Assuming the same covariance structure of the transformed variables as was in case of survey 3 data (the reason being the scale and shape parameters of the two Weibull distributions were close to that in case of survey 3 data), equation (8), i.e.

$$\mathrm{E}_{i,j}(GHQ|\,Diff.) = \mu_{i,GHQ} - \frac{1}{\sigma_{i,GHQ}}\sigma_{i,Diff.,\mathrm{GHQ}}\left(Diff_{i,j} - \mu_{i,Diff_j.}\right);\ j = 1,2$$

; j denotes the values of survey 1 and survey 2.

was used to calculate conditional expectation of GHQ total given Difficulty scores.

The results are presented in Table **8** below. Columns 6 and 7 of the Table **8** present the estimated GHQ scores for survey 1 and survey 2 respectively as obtained through the BVN model.

Table 8 Estimated unknown mean *GHQ total given *Diff score for dataset 2 using a generated random sample of size 10000 for BVN distribution

1	2		3		4	5	6	7
Interval	**Real Data**		**Simulated Data**		**E[*GHQ| *Diff.] for survey 1**	**E[*GHQ| *Diff.] for survey 2**	**Estimated GHQ survey 1**	**Estimated GHQ survey 2**
	Mean Diff_survey 1	**Mean Diff_survey2**	**Mean Diff_survey1**	**Mean diff_survey2**				
1<Diff≥1.25	1.1892	1.1892	1.2336	1.1796	1.9411	2.2819	3.7677	5.2071

1.25<Diff ≥1.5	1.4493	1.4346	1.4308	1.4304	2.5374	2.5501	6.4384	6.5030
1.5<Diff ≥1.75	1.6704	1.6631	1.6575	1.6531	2.9500	2.9473	8.7025	8.6866
1.75<Diff ≥2	1.8857	1.8870	1.8554	1.8720	3.6963	3.7039	13.6626	13.7188
2<Diff ≥2.25	2.1044	2.0982	2.0936	2.0909	4.7516	4.7531	22.5777	22.5919
2.25<Diff ≥2.5	2.2868	2.2775	2.3182	2.3238	5.8536	5.8595	34.2646	34.3337

3.5.5 Validation of the model using information on respondents who participated in all the three surveys.

Although the three surveys were conducted independently on dynamic population of young adults, 162 respondents were found to have participated in all the three surveys. The observations on these respondents were not independent. The model was applied on survey 3 data of these 162 respondents to estimate GHQ total for known Difficulty scores. The results are presented in Table **9** below:

Table 9 Estimated mean *GHQ total given *Diff score for survey 3, of 162 respondents (who participated in all the three surveys) for different intervals of *Diff score

Interval (1)	**Real Data (2)**		**E[*GHQ |*Diff.] (3)**	**Observed mean scores of GHQ (4)**	**Estimated GHQ (after reverse transformation) (5)**
	Mean *GHQ	**Mean *Diff.**			
1<Diff≥1.25	2.0819	1.1621	1.9671	4.3343	3.8695
1.25< Diff ≥150	2.9142	1.4953	2.4935	8.4925	6.2175
1.50<Diff ≥1.75	3.2533	1.6764	2.8471	10.5840	8.1059
1.75< Diff ≥2	3.8880	1.8999	3.7028	15.1165	13.7107
2< Diff ≥2.25	4.5194	2.1286	4.7454	20.4250	22.5188
2.25<Diff ≥2.50	5.6587	2.3131	5.8468	32.0209	34.1851

The columns (4) and (5) of Table **9** are compared graphically in Fig. **6** below:

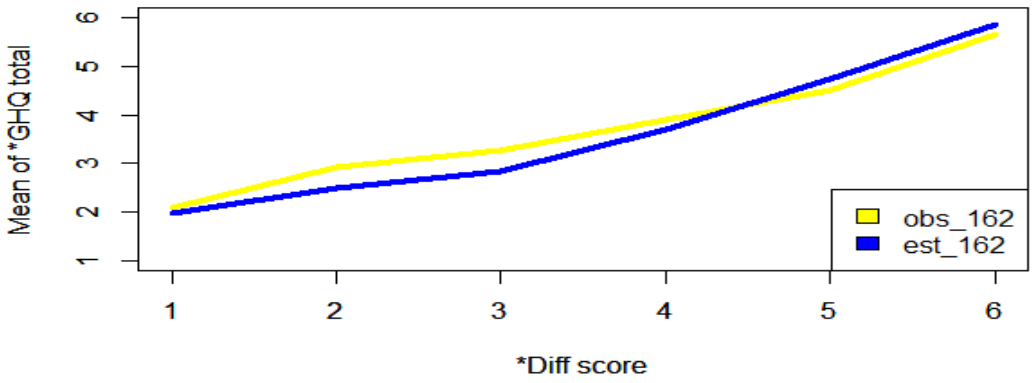

Fig. 6 Estimated mean *GHQ total given *Diff scores in survey 3, of 162 respondents common to all the three surveys

It is evident from Fig. 6 that for this data set also, the values estimated through the model are in close proximity to the observed mean scores.

Next, we applied the model on survey 1&2 data of these 162 respondents to complete the information about GHQ scores (the then recent health at the time of these surveys). The results are presented in Table **10** below:

Table 10 Estimated unknown mean *GHQ total given *Diff score for survey 1 and survey 2, of 162 respondents (who participated in all the three surveys) for different intervals of *Diff score

Interval	Known *Diff score mean	Estimated *GHQ total	Estimated GHQ (after reverse transformation)
1<Diff≥1.25	1.1892	1.9412	3.7682
1.25<Diff ≥1.5	1.4547	2.5326	6.4140
1.5<Diff ≥1.75	1.6875	2.6397	6.9680
1.75<Diff ≥2	1.8959	3.6913	13.6257
2<Diff ≥2.25	2.1464	4.7507	22.5692
2.25<Diff ≥2.5	2.2833	5.8550	34.2810

All the calculations have been done using R software, version 4.2.1.

4 Discussion

Psychiatric experts suggest to either use a single questionnaire multiple times or to use a number of questionnaires at a single point of time in order to confirm consistency in the observations. During three (independent) surveys conducted by the authors during COVID-19 pandemic times, using SDQ, the observations were found to be consistent (using Chronbach alpha and Guttman Lambda statistics) (Goyal et al., 2023; Sabharwal et al., 2023; Sabharwal et al., 2024). The psychological health of young adults was assessed through SDQ in the first two surveys. In third survey, due to deadly ‘delta’ wave, the authors felt the need to assess the recent general health of the respondents using GHQ-12, along with SDQ.

The GHQ-12 is a questionnaire which is used to assess the recent general health of the respondents (4-6 weeks). The main components of the questionnaire are Social dysfunction, Depression and Lack of confidence. On the other hand, SDQ extended version assesses strength (Prosocial behaviour), difficulties (Emotional symptoms, Conduct problems, Hyperactivity- inattention symptoms and Peer problems) and social dysfunction (Impact score). The correlation coefficient between the GHQ-12 score and the SDQ Difficulty score was 0.49997, which was found to be significant with p-value < 0.001. As the responses were found to be consistent, we applied a parametric model to extract information on one questionnaire using information from the other.

The distributions of the original (survey 3 data based) GHQ total and Difficulty score were found to be Weibull. After applying power transformations to these variables, the transformed variables were normal, which were correlated as well. Mardia test was applied to test the multivariate normality of the transformed variables, yielding a *p-value* =0.7158. Subsequently, a bivariate normal distribution was fitted on the observed data. Using the fitted model, we estimated GHQ scores given Difficulty scores and vice versa. The estimated values were compared with the observed values. The results obtained through the model were consistent with the observed values. Then the model was used to extract the unknown GHQ total scores of the respondents in survey 1 and survey 2 (Dataset_2) from the observed Difficulty scores.

When the results of the model were applied on 162 respondents who had participated in all the three surveys, it was found that 64 respondents were at advance level of both the Difficulty score and GHQ total at the time of survey 3. The model was then applied on surveys 1&2 data for these respondents to estimate their then recent health (based on estimated GHQ total). Out of these 64 respondents (having higher GHQ total in survey 3), the model estimated that 55 had higher GHQ total at the times of surveys 1&2 also.

5. Conclusion

The results suggest that whenever the assumptions of the multivariate normal model are met (irrespective of the distribution(s) of original variables), the model can be used to extract information on one variable from the other (Grover et al., 2014; Grover et al., 2015). In medical field, the missing or incomplete observations are very frequent. The above model can be applied in these circumstances to explore and extract the missing observations.

Reference

[1] Johnson, N. L., and Kotz, S., Distributions in statistics: Continuous multivariate distributions, John Wiley & Sons, New York, 1972.
[2] Khattree, R., & Naik, D. N. (2000). Multivariate Data Reduction and Discrimination with SAS Software. Wiley-SAS.
[3] Lindsey, J. K. (2000). Applying generalized linear models. Springer Science & Business Media.
[4] Chi, Y. Y. (2012). Multivariate methods. Wiley Interdisciplinary Reviews: Computational Statistics, 4(1), 35-47.
[5] Lipow, M., and Eidemiller, R.L., 1964. Application of the bivariate normal distribution to a stress vs strength problem in reliability analysis. Technometrics, 6(3), 325-328. http://www.jstor.org/stable/1266043.
[6] Yue, S., 1999. Applying bivariate normal distribution to flood frequency analysis. Water International, 24(3, 248-254. http://dx.doi.org/10.1080/02508069908692168.
[7] Grover, G., Sabharwal, A., & Mittal, J. (2014). Application of Multivariate and Bivariate Normal Distributions to Estimate Duration of Diabetes. International Journal of Statistics and Applications, 4(1), 46–57. http://www.sapub.org/global/showpaperpdf.aspx?doi=10.5923/j.statistics.20140401.05.
[8] Grover, G., Sabharwal, A., & Kaushik, S. (2015). Estimating Length of Stay and Duration of Illness for Psychiatric Inpatients using Multivariate Modelling. American Journal of Mathematics and Statistics, 5(6), 329–353. doi=10.5923/j.ajms.20150506.02
[9] Goldberg, D.P. & Hillier, V.F. (1979). A scaled version of the General Health Questionnaire. Psychological Medicine, 9, 139-45.
[10] Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. Journal of Child Psychology and Psychiatry, 38(5), 581–586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

[11] Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child Psychiatric caseness and consequent burden. Journal of Child Psychology and Psychiatry, 40(5), 791–799. https://doi.org/10.1111/1469-7610.00494
[12] Goyal B, Sabharwal A, Chauhan V, & Joshi, LM (2023). Psychological health of Indian youth during COVID-19: a study through three chronological surveys. International Journal of Public Health, 12(2), 752-763.
[13] Sabharwal, A., Goyal, B., Chauhan, V. S., Joshi, L. M., & Goyal, V. (2023). Evaluating the effect of COVID-19 pandemic on the psychological health of young adults in India. International Journal of Public Health Science (IJPHS), 12(1), 311. https://doi.org/10.11591/ijphs.v12i1.21527
[14] Goldberg, D. (1979). GHQ and Psychiatric Case. The British Journal of Psychiatry, 134(4), 446–447. https://doi.org/10.1192/bjp.134.4.446b
[15] Goldberg, D. P., Gater, R., Sartorius, N., Ustun, T. B., Piccinelli, M., Gureje, O., & Rutter, C. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychological Medicine, 27(1), 191–197. https://doi.org/10.1017/s0033291796004242
[16] Anjara, S., Bonetto, C., Van Bortel, T., & Brayne, C. (2020). Using the GHQ-12 to screen for mental health problems among primary care patients: psychometrics and practical considerations. International Journal of Mental Health Systems, 14(1). https://doi.org/10.1186/s13033-020-00397-0.
[17] Transformations, http://www-users.york.ac.uk/~mb55/msc/clinbio/week5/transfm_gif.pdf.
[18] Box-Cox and Power transforms,http://www.statsref.com/HTML/index.html?freeman-tukey.html.
[19] Mardia, K.V. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3):519–530, 1970.
[20] Von Eye, A., & Bogat, G. A. (2004). Testing the assumption of multivariate normality. Psychology Science. http://pabst-publishers.de/psychology-science/2-2004/ps_2_2004_243-258.pdf
[21] Kuha, J. (2004). AIC and BIC. Sociological Methods & Research, 33(2), 188–229. https://doi.org/10.1177/0049124103262065
[22] Bradman, M.J., Clark, T.G., Love, S.B., Altman, D.G., 2003, survival analysis part III: Multivariate data analysis-choosing a model and assessing its adequacy and fit, British Journal of Cancer, 89, 605-11, doi: 10.1038/sj.bjc.6601120.
[23] Sabharwal, A., Goyal, B., & Joshi, L. M. (2024). Psychological Health of Young Adults as Measured through Internalising and Externalising Scores of Strength and Difficulty Questionnaire. Migration Letters, 21(S4), 1511-1520.