JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

Key Performance Indicators- A Bayesian Modelling for Enhanced Data Analysis

S. Mythreyi Koppur[1*], Dr. B. Senthilkumar[2]
[1*]Research Scholar, Department of Statistics, Periyar EVR College (Autonomous), Trichy, India.

[2]Assistant Professor, Department of Statistics, Periyar EVR College (Autonomous), Trichy, India.

Corresponding Author Email:mythreyikoppurs@gmail.com

Received: 2022 March 15; **Revised:** 2022 April 20; **Accepted:** 2022 May 10

Abstract

The aim of this article is to study about the key performance indicators as a "measure of performance" and can evaluate the accomplishment in any organization or other projects. The innate benefits in handling Bayesian hierarchical modelling has been exploited with the underlying models using appropriate transformation of underlying parameters. This study has considered illustrative datasets from open repositories to analyse the variability in the association between different categorical variables. Odds ratios are used to compare the measure of association between the variables of interest. Both individual and overall Odds ratios together with measure of heterogeneity estimates are used to obtain appropriate inferences and the effect of variables of interest.

Keywords:Bayesian, Heterogeneity, Key Performance Indicators, Odds Ratio

1. Introduction

In recent times,Meta analytic approach has witnessed extensive applications in various fields. This includes medicine , epidemiology , sports , social sciences etc.Engels et al. (2000), Hagger (2006), Viechtbauer (2007), Bowden et al. (2011), Riley et al. (2011), Davis et al. (2014) and Langan et al. (2015, 2016) are few but highly informative studies.This approach provides ample scope to obtain better / more insights from data when appropriate variables are considered for their association. In many cases one variable may act as an important variable of interest and others may impact on it. The major objective of data analysis could be to bring out and quantify these associations using appropriate statistical model. Ideally some practical situations or applications have natural key point indicators as one or more

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

variables. In such situations suitable application of statistical models would draw necessary insights from data. On the other hand, it may be required to treat the data suitably so that relevant statistical procedure can be applied to discover from the data. Such an approach might be based on context based inputs and / or intuitional hypothesis about the problem in hand. The latter approach is an integral part of many data analysis task in framing appropriate research questions and draft analysis plan and carry out the analysis with appropriate procedures, and computing tools. This paper has madea similar attempt with one of the most prominent data structures. Based on the understanding of context derived from illustrative data sets, analysis plans have been drafted and carried out in Bayesian statistical approach. This includes identifying suitable variables to study the association between them, selecting appropriate response variables as a major KPI and a measure/ metric to quantify the association between variables. Also necessary inputs have been studied to carry out typical Bayesian analysis; prior specifications to computing platform to realize the output of the chosen model more specifically. This study has focussed on multiple 2 x 2 contigencytables; odds ratio has been considered as an appropriate association measure along with measure of heterogeneity is identified as a significant measure for the analysis. STAN language(Gelman, A., Lee, D., &Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioural Statistics*, *40*(5), 530-543.)has been used to carry out the Bayesian analysis in R. The required inputs to run MCMC sampler (**Markov Chain Monte Carlo**) have been chosen and necessary convergence diagnostics are adopted in summarizing the outputs.(Grzegorczyk, M., &Husmeier, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning*, *71*(2-3), 265.) **.** A typical data set has been presented in section 2 as a motivational case; section 3 lists the methods and models applied in the paper; details of other data sets and analysis have been presented in section 4; section 5 provides the discussions and conclusions derived from the analysis.

2. Motivational Examples

Considering a data set ***Credit*** from CRAN repository . This is a typical rectangular data that has 400 observations (rows) and 12 variables (columns) including ID of an observation. The major aim is to know or predict defaulting customers as the aim lies in identifying a right KPI as a response variable. In this data set ,*Balance* (Average credit card

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

balance in $) could be a variable of interest. As an initial attempt, a few measures are required to know about *Balance.*

Table 1 Dataset

Measure	**Balance**
Minimum	0.00
1st Quarter	68.75
Median	459.50
Mean	520.01
3rd Quarter	863.00
Maximum	1999.00

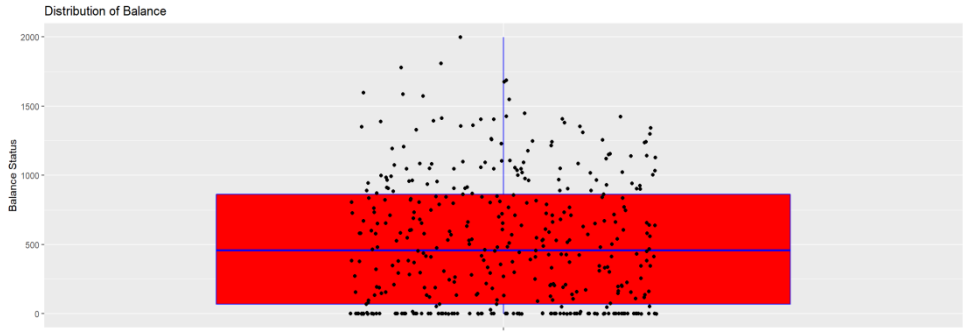

Figure 1 Distribution of Balance

This provides the way in which ***Balance*** is *'distributed'*. The next step is to focus on other variables to relate with the variable of interest. One such variable, ***Age*** could be the choice. To understand the relation between these two variables,

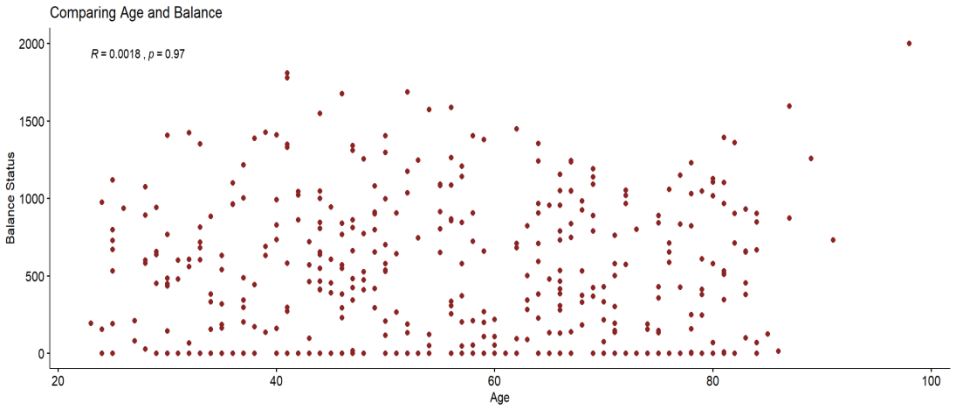

Figure 2 Comparing Age and Balance

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

To some extent it is possible to connect Age and Balance. In one side, low Balance prevails for all age group (in the bottom); on the other hand slightly different pattern is visible when the Age increases. However, numerical correlation indicates a poor correlation (0.0018) and statistically insignificant too; yet, it is not sure whether this is ruling out practical significance of relating Age and Balance.

A similar attempt can be tried with one more variable, student and Balance.

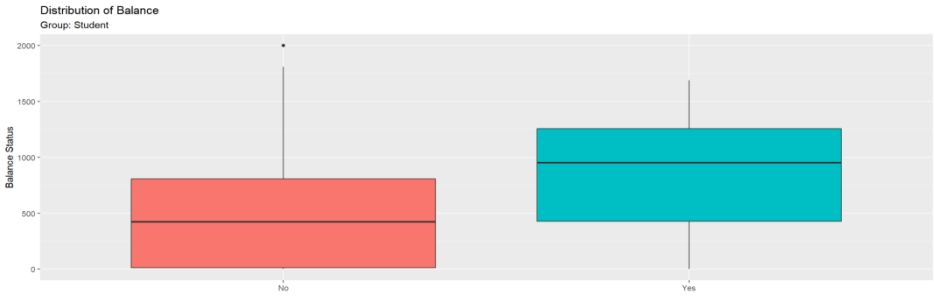

Figure 3 Distribution of Balance between students

Some visible difference in the 'distribution' of Balance could be noted among two groups of Students. But, how much this variability is accounted for and whether this difference is statistically significant beyond the notion of practical significance of connecting Student and Balance. It indicates that difference in the means of Balance in two groups of Student is statistically significant.

The next step takes in connecting Age, student and Balance.

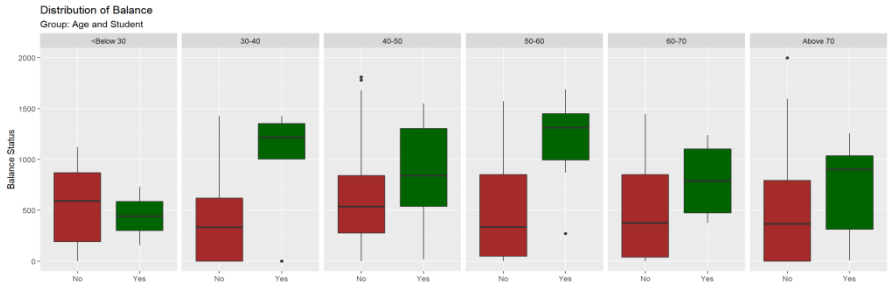

Figure 4 Distribution of Balance between Age and students

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

This output may indicate a possible different 'distribution' of Balance in the Age group further grouped by Student. The notion of variability is quite apparent in two or more Age groups. Then the question of interest may lead to quantify this variability.

Before trying to answer this, one more attempt of connecting three variables could be done by creating two groups for Balance.

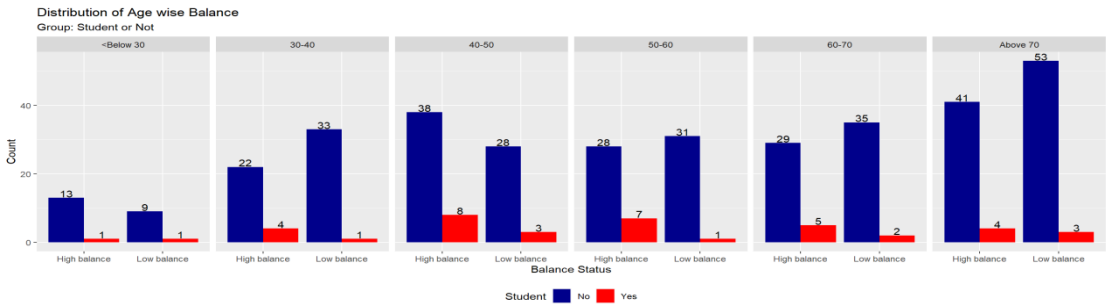

Figure 5 Distribution of Age wise Balance

In most of the cases variability is the key to know from data and the variables associated with the process. Modifying two variables to understand their relations more better, assuming there is a practical significance in investigating the relations. Such modifications might change the *Nature* of the variables and subsequent visual and numerical treatments. There is a scientific way to conclude and / or support the process of investigation. This may help in quantifying the variability so as to understand the heterogeneous behaviour of a key variable. Five more datasets have been considered in this study and the next chapter lists the methods and models considered in this study.

3. Methods And Materials

The major association metric used in this study is "ODDS RATIO", a descriptive measure for comparing groups on binary responses i.e., between the two groups 1 and 2 of the variable X. Agresti (2013) and Card (2012) can be referred for more details on odds ratio and its advantages so as to appreciate it as a desired summary measure in REM with binary data. One of the notable merits is its invariable nature over case control, follow-up and cross-sectional studies and thus it can be applied to differentiate findings of various study designs. OR has wide applications in statistical analysis as in case-control studies multicentre study, meta-analysis and diagnostic accuracy (Normand, 1999 and Suzuki, 2006). For a probability ω

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

of success, the odds are defined to be $\tau=\frac{\omega}{1-\omega}$. The ratio of the odds τ_1 and τ_2 in the two rows,

$\theta=\frac{\tau_1}{\tau_2}=\frac{\omega_1/1-\omega_1}{\omega_2/1-\omega_2}$ is called the Odds Ratio. An alternative name for θ is the cross-product ratio and a sample version of OR is ($\frac{ad}{bc}$) and can equal any non-negative number. When OR > 1 subjects in row 1 are more likely to make the first response than subjects in row 2 and if OR lies between 0 and 1 then the first response is less likely in row 1 than in row 2. Also the natural logarithm of Odds Ratio is widely used for convenience. Prior distributions play significant role in Bayesian modelling especially if the inferential problem is focussed on boundaries of the parameter space. However Bayesian procedures largely rest on the choice of priors. In recent days , Bayesian perspective to statistical inference has received more consideration in research in applied and conceptual statistics. It is mostpredominent to record that most of the advantages claimed for Bayesian approach follow from the ability to handle complex models and the three main aspects that reflects Bayesian modelling include,

- Computation.
- Incorporation of historical information.
- Inference on complex functions of parameters.

This study basically aims to exploit the inbuilt advantages of Bayesian approach in statistical inference on categorical data analysis with binary outcomes. This study also focus on Bayesian study designs, the choice of appropriate priors and the computational strategies involved in the analysis of a problem in practice. Bayesian analyses for complex models can be applied on a data set andmake it simple using Monte Carlo methods to generate posterior distributions. MCMC is originally Monte Carlo integration using Markov Chains which provides enormous scope for realistic statistical modelling through aintegratingstructure within which many complex problems can be explored using generic software. R is a language for analysis of data and graphics ,anextended source unrestricted access statistical software package . The entire exercise has been carried out using the computational tool R studio (version 1.1.463) especially with “R Stan: the R interface to Stan” (Stan Development Team, 2020)). Numerical and graphical summaries are quite straight forward with the tool. Stan Development Team (2020). “RStan:

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

the R interface to Stan." R package version 2.21.2.

3.1. Random Effect Model

Random Effect Model (REM) is a statistical method to merge the outcome of individual studies so as to enhance the accuracy of the estimates of study effect and assess whether study effects are similar enough to be combined. It can be shown that simple average of study effects may not be a proper method to summarize the results. It is important to understand the sources of variability, within-study and between study when making conclusions about the population. Such models are of considerable scientific interest and closely resemble the statistical principles of meta- analysis. Extensive studies are available in detailing the conceptual, statistical, computational and interpretative aspects of REM and / or meta-analysis. Though medical, epidemiological and health related studies dominate this field, many other faculties exploit the advantages of REM in terms of prospective, retrospective or cross sectional studies. Card (2012) provides a better overview of methods involved in meta-analysis for social science with binary or metric data.

In Random Effect Model, if $\boldsymbol{\delta_i}$ is an effect size estimate of a corresponding true effect size $\boldsymbol{\theta_i}$ with the Within-study variance σ_i^2 , then we could **estimate**$\boldsymbol{\theta_i}$, i=1,2,⋯k from the sample data; let us denote these estimated values as $\boldsymbol{\delta_i}$ i=1,2,⋯k; that is $\boldsymbol{\widehat{\theta}_i} = \delta_i$ i=1,2,⋯k. Here we estimate Odds Ratio based on the independent binomial distribution for the two rows. Precisely for the two rows in each table,

$\rho_1 \sim$ Bin (n_1, θ_1)

$\rho_2 \sim$ Bin (n_2, θ_2)

Then we define $\varphi = logit\,(\theta_1) - logit\,(\theta_2) = log\,\left(\frac{\theta_1/(1-\theta_1}{\theta_2/(1-\theta_2}\right)$ which is log odds ratio, the quantity of interest in logscale for each of the k tables. Also for k tables we define,

$\mu = \frac{logit\,(\theta_1)+logit\,(\theta_2)}{2}$.

This parameterization is equivalent to $\theta_1 = logit^{-1}\left(\mu + \frac{\varphi}{2}\right) = \frac{\exp(\mu+\frac{\varphi}{2})}{1+\exp(\mu+\frac{\varphi}{2})}$ and $\theta_2 = logit^{-1}\left(\mu - \frac{\varphi}{2}\right) = \frac{\exp(\mu-\frac{\varphi}{2})}{1+\exp(\mu-\frac{\varphi}{2})}$.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

Hence the second stage of the model is specifying priors for $\mu \ and \ \varphi$.

Modelling assumptions are,

$$\mu \sim \mathrm{N}\,(\mu_0\,, \sigma_0^2) \tag{1}$$

$$\varphi \sim \mathrm{N}\,(d, \tau^2) \tag{2}$$

σ_0^2amounts sampling variability in the effect size θ_i estimate.τ^2 amounts variability between the effect size (among the grouping or stratifying variables).The statistical inference aims to provide following summaries to understand the association between the variables in the individual and overall levels together with the amount of heterogeneity.

- Point estimate and confidence interval for the true θ_i
- Point and interval estimates of μ to understand the presence or absence of an

  overall effect and its statistical significance.
- Estimates of variability measures indicating the variation between strata.

4. Data Analysis

Six data sets have been collected from the repository representing different areas. The motivation for collecting the data sets is to have as many variables of different type. We aim at identifying a right KPI (Key Performance Indicator) as a response variable. The focus lies in selecting a right procedure to treat the response metric variable in to a categorical variable with binary outcomes. The next step follows in pickingsuitable associated variables to get two more dimensions. The final data set will be a K x 2 x 2 data set. The details of the datasets considered in this study with the description are given in table 1.

Table 2Detailed explanation of datasets examined in this present study.

Data Set	Description	No. of observations	No. ofvariables	Factorvariables	No. of models

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

D1	CREDIT CARD BALANCE DATA	400	12	BALANCE	3
D2	ADULT CENSUS DATA	32561	15	INCOME	2
D3	HEALTH INSURANCE	8802	11	INSURANCE	2
D4	Ph.D., PUBLICATIONS	915	6	ARTICLES	2
D5	WEIGHT OF CHICKS ON DIFFERENT DIETS	578	4	WEIGHT	1
D6	END SEMESTER MARKS	281	10	MARKS	1

Various models can be achieved through possible combination of variables. The details of the variables for all the data sets, together with models is explained in detail in table 2.

Table3Details of variables for the data sets considered in this study.

Data set	**Model**	**Grouping Variable**	**Predictor Variable**	**Response Variable**
D1	M1	AGE	STUDENT	BALANCE
	M2	AGE	GENDER	BALANCE
	M3	AGE	MARRIED	BALANCE
D2	M1	AGE	GENDER	INCOME
	M2	EDUCATION	GENDER	INCOME
D3	M1	AGE	EMPLOYMENT	INSURANCE
	M2	AGE	GENDER	INSURANCE
D4	M1	KIDS	GENDER	ARTICLES

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

	M2	MENTOR	GENDER	ARTICLES
D5	M1	DIET	AGE	WEIGHT
D6	M1	DEPARTMENT	GENDER	MARKS

While the main interest in Bayesian inference is to get relevant insights from a posterior distribution π (θ|X), it is expected that the sample generated from a MCMC algorithm should adequately represent π (θ|X). Hence, convergence becomes an essential part of this computational procedure. Convergence can be achieved by running many chains of comparatively smaller in length. Graphical tools such as kernel density plots help in understanding the convergence of MCMC chains.In MCMC the most important part is to identify, the number of initial iterations M, that are to be removed, and then for a further N iterations every K^{th} value has to be stored.MCMC procedures for Prior-Data-Posterior modeling could be implemented in a suitable computing platform. This study follows one of the most widely used techniques Random Effect Model as parameter estimation procedure. This will yield Odds Ratio $\widehat{\boldsymbol{\theta}}_{\boldsymbol{i}} i$ = 1, 2,…k with point and interval estimates, secondly an estimate $\widehat{\boldsymbol{\theta}}$for the overall odds ratio and an estimate for between variability, major quantity of interest.

TABLE4Combined Odds Ratio and heterogeneity measure (τ^2) for various data sets considered in this study, together with the lower and upper limits of 97.5% confidence interval.

DATA SET	MODEL	OVERALL OR POINT ESTIMATE	OVERALL OR INTERVAL ESTIMATE	HETEROGENEITY (τ^2) POINT STIMATE	HETEROGENEITY (τ^2) INTERVAL ESTIMATE
	M1	0.330	(0.129,0.823)	0.390	(0.130,1.088)
D1	M2	1.060	(0.561,2.031)	0.400	(0.135,0.975)
	M3	1.040	(0.549,1.960)	0.320	(0.120,0.801)
D2	M1	0.270	(0.171,0.419)	0.300	(0.120,0.685)

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

	M2	0.290	(0.154,0.473)	0.310	(0.110,0.808)
D3	M1	0.400	(0.241,0.652)	0.290	(0.110,0.698)
	M2	0.730	(0.453,1.190)	0.280	(0.109,0.703)
D4	M1	1.160	(0.640,2.083)	0.330	(0.123,0.853)
	M2	1.230	(0.570,2.705)	0.380	(0.127,1.067)
D5	M1	0.010	(0.002,0.020)	0.480	(0.138,1.450)
D6	M1	0.810	(0.335,1.999)	0.440	(0.137,1.279)

From Table 3, a clear specification of the numerical summaries of overall odds ratio with the interval estimates and the measure of heterogeneity ($\boldsymbol{\tau^2}$)with the interval estimates is provided. The overall odds ratio is more than 1 in datasets 1& 4, except for dataset1 model 1.Also to note that in dataset 1 (model 2&3) and dataset 4(model 1&2) the estimates are not statistically significant. In dataset 1(model1) , balance is compared with student across various age groups and the overall OR estimate is 0.330, which shows that the odds of being a student and having high balance is more than not being a student and having a high balance. In dataset 1 (model 2) , balance is compared with gender across various age groups and the overall OR estimate is 1.060, which shows that the odds of being female and a high balance is more than being a male and having high balance. In model 3 , balance is compared with married across age groups and the overall OR estimate is more than 1 , which shows the odds of not married and having a high balance is more than married and having a high balance. Further individual measure such as dataset specific odds ratio (θ_i) together with 97.5% confidence interval is presented in forest plot (Lewis and Clarke, 2001) for easier visual interpretations and understanding the variability. Figure 1 is the forest plot for the dataset1.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

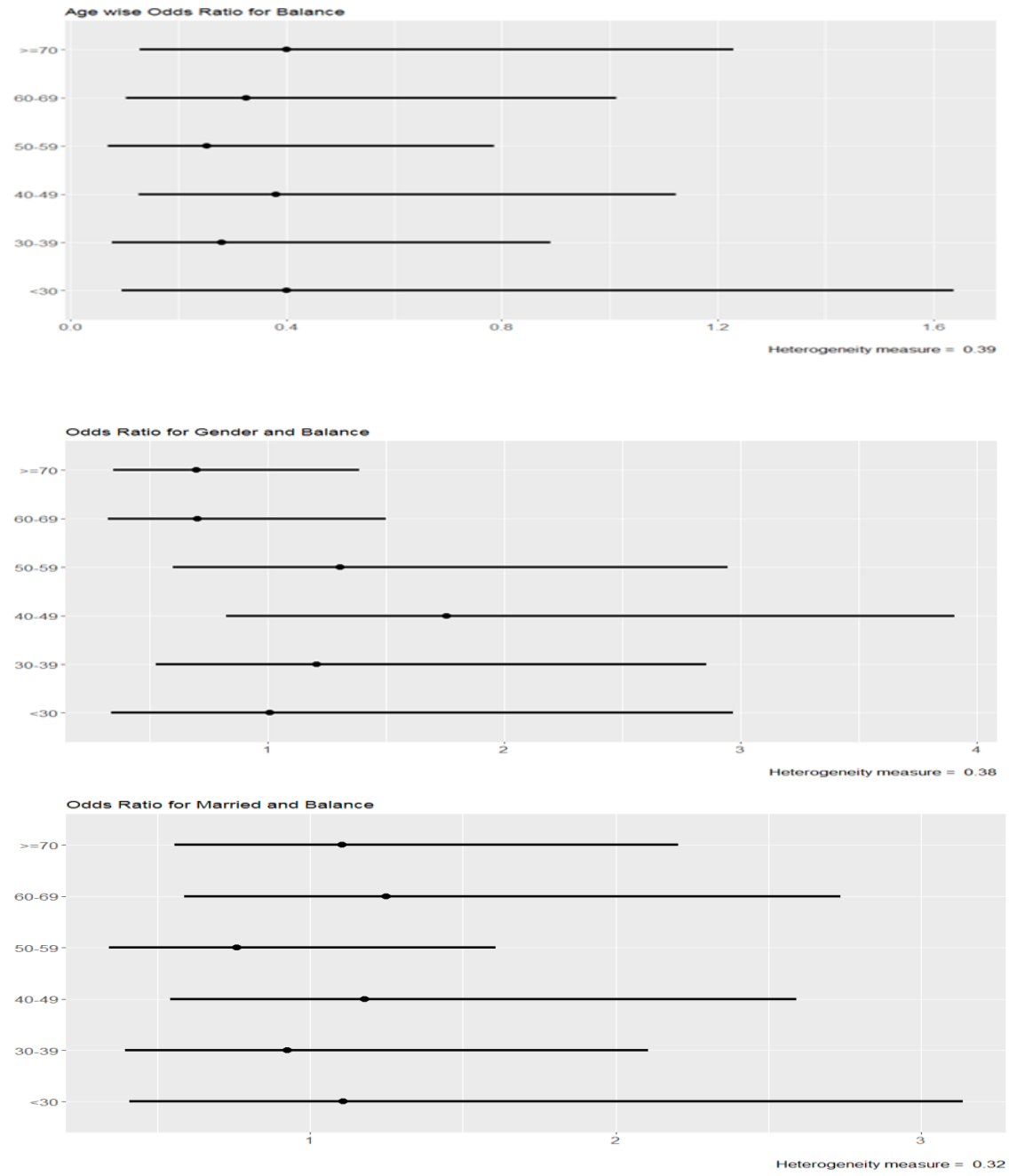

Figure 6 Forest plotof the point and interval estimates of the individual odds ratio for having a high balance with respect to student, gender and married across various age groups

Figure 1 shows the possibility of having a high balance corresponding to student, gender and married across various age groups. In model1, except for the age groups 30-39 and 50-59 the estimates are not statistically significant. The age group <=30 shows a different behaviour with a wider interval. In model 2 except for 60-69 &>=70, the results are not statistically significant. In model 3 except for the age groups of 30-39 and 50-59

Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

the results are not statistically significant. Further, forest plot helps to understand the variability between the models in the respective data sets. Higgins and Thompson (2002) provides the necessary interpretation of the metric (τ^2) for measuring such heterogeneity.

Data Set 5 has the largest τ^2 = 0.480 with 95% CI (0.138, 1.450) followed by Data Set 6 which has 44% variability. Further a positive measure of heterogeneity across all data sets indicating with a notable wider 97.5% credible interval indicating a heterogeneous effect size across all the data Sets considered in this study. Forest plot for all the dataset has been generated but due to paucity of space restrictions forest plot for the remaining datasets considered in this study could not be presented in the paper.

5. Conclusion

This paper has made an attempt in identifying Key Performance Indicators which evaluates the accomplishment in any organisation or any other projects. Six datasets taken from the repository have been analysed and the focus lied largely in analysing a right KPI. Converting a rectangular dataset in to 2 x 2 dataset with k levels after the metric variable is treated as a categorical variable; the results have become more visible in a numerical form.Random Effect model which has been considered as the underlying model and Bayesian methods as a statistical principle in order to provide a better insight. This study had provided a better understanding of relationship between the key variables, in understanding the uncertainty and predicting the future events associated with the process. The results obtained from the study using Random Effect Model (REM) approach provided a better insight in terms of Variability quantification.

References

1. Agresti, A. (2003). Categorical data analysis (Vol. 482). John Wiley & Sons.
2. Bowden, J., Tierney, J. F., Copas, A. J., & Burdett, S. (2011). Quantifying, displaying and accounting for heterogeneity in the meta-analysis of RCTs using standard and generalised Q statistics. BMC medical research methodology, 11(1), 1-12.
3. ISLR refer, https://cran.r-project.org/package=ISLR.
4. https://vincentarelbundock.github.io/Rdatasets/datasets.html
5. Card, N. A. (2015). Applied meta-analysis for social science research. Guilford Publications.
6. Davis, J., Mengersen, K., Bennett, S., &Mazerolle, L. (2014). Viewing

systematic reviews and meta-analysis in social research through different lenses. SpringerPlus, 3(1), 1-9.

7. Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Statistics in medicine, 19(13), 1707-1728.
8. Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in medicine, 21(11), 1539-1558.
9. Langan, D., Higgins, J. P., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. Research synthesis methods, 8(2), 181-198.
10. Langan, D., Higgins, J. P. T., & Simmonds, M. (2015). An empirical comparison of heterogeneity variance estimators in 12894 meta-analyses. Research Synthesis Methods, 6 (2), 195–205.
11. Lewis, S., & Clarke, M. (2001). Forest plots: trying to see the wood and the trees. Bmj, 322(7300), 1479-1480.
12. Team, R. C. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.
13. Riley, R. D., Higgins, J. P., &Deeks, J. J. (2011). Interpretation of random effects meta-analyses. Bmj, 342.
14. Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in medicine*, *26*(1), 37-52.
15. Chan, A. P., & Chan, A. P. (2004). Key performance indicators for measuring construction success. *Benchmarking: an international journal*.
16. Badawy, M., Abd El-Aziz, A. A., Idress, A. M., Hefny, H., &Hossam, S. (2016). A survey on exploring key performance indicators. *Future Computing and Informatics Journal*, *1*(1-2), 47-52.
17. Bhatti, M. I., Awan, H. M., &Razaq, Z. (2014). The key performance indicators (KPIs) and their impact on overall organizational performance. *Quality & Quantity*, *48*(6), 3127-3143.
18. Lindberg, C. F., Tan, S., Yan, J., &Starfelt, F. (2015). Key performance indicators improve industrial performance. *Energy procedia*, *75*, 1785-1790.
19. Hristov, I., & Chirico, A. (2019). The role of sustainability key performance indicators (KPIs) in implementing sustainable
strategies. *Sustainability*, *11*(20), 5742.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 4380-4394
https://publishoa.com
ISSN: 1309-3452

20. Parmenter, D. (2015). *Key performance indicators: developing, implementing, and using winning KPIs*. John Wiley & Sons.
21. del-Rey-Chamorro, F. M., Roy, R., Van Wegen, B., & Steele, A. (2003). A framework to create key performance indicators for knowledge management solutions. *Journal of Knowledge management*.