

Estimating Markov Transition Probabilities Between Health States in the Social Security Malaysia (SOCSCO) Dataset

Shamshimah Samsuddin, Noriszura Ismail

Abstract— Occupational injury represents a considerable part of injury burden to the society as it may affect workers in their most productive years. The objective of this paper is to estimate the Markov transition probabilities of a worker's health states over time using the Counting Method (CM) and the Proportional Odds Model (POM), focusing on disability among the Social Security Organization (SOCSCO) contributors in Malaysia. Four health states namely active/work (A), temporary disability (T), permanent disability (P) and death (D) are considered, where the transition probabilities are estimated at yearly intervals based on age, gender, year and disability category.

Keywords: Occupational injury, Markov transition probabilities, Counting Method, Proportional Odds Model

I. INTRODUCTION

Disability associated with work-related injuries has become an increasing burden to the workers' compensation in Malaysia. Therefore, understanding and awareness on disability among employees is important in assisting employees and employers for future retirement preparation such as retirement benefits and healthcare expenses. A study on the Social Security Organization (SOCSCO) in Malaysia conducted by [13] has reported increasing number of accidents in the workplace in the recent years, which resulted in increasing compensation payments by SOCSCO.

Estimation of transition probabilities amongst disabilities involves a challenging method, as a disable person could transfer from one level of disability to another, for instance, from temporary disability (T) to permanent disability (P), or vice versa, it also could transfer from one event to another, for instance, from T/P to active/work (A), or death (D). In other words, recurrent responses may be obtained at different time points in the related longitudinal studies. The study aims to estimate the Markov transition probabilities between health states using the Employment Injury Scheme (EIS) dataset which gathered from SOCSCO. The EIS is the most suitable dataset to construct the disability probabilities because this data provides four possible outcomes at yearly intervals for its health status variable namely Active/work (A), Temporary disability (T), Permanent disability (P) and Death, (D). We estimate the Markov transition probabilities using two separate methods, concentrating on the working age of 15 to 65 year olds.

The first method is the counting method (CM) which contains a simple counting procedure, and the second method is the Proportional Odds Model (POM). The estimates obtained from this study will be used in our future study to project the healthcare costs of disability and death of workers in the future in terms of longer time horizon. The results that obtained from this study can also be used to determine the gender and age that has a more positive (or negative) connection with the probabilities of transition, which allow us to categorize the group of workers with a risky health status.

This paper is organized in the following way. The next section which is Section 2 will describe the data. Sections 3 discusses the two methods, the CM and the POM, for estimating the transition probabilities, together with the related literatures. Section 4 contains the results from the CM and the POM. We conclude our analysis in Section 5

II. DATA

The files are combined together to form a single file with a complete event history of workers (counting process). Table 1 shows the benefits and active (working) states considered in this study, where a larger number refers to a higher rank of state

Table 1 Numbering and Definition of State

State Number	State Name
4	Death Benefit (Death), D
3	Permanent Disability, P
2	Temporary Disability, T
1	Active (Working), A

After merging the files, a total of 263,971 EIS recipients are identified in years 2009 to 2013. It should be noted that in this study only deaths and occupational accidents and injury claims lodged during years 2009 to 2013 are used, thereby excluding any deaths and earlier claims carried over from the previous years. The study is constrained to SOCSCO contributors aging between 15 to 65 years old (working age) and the yearly counts are made at the end of each year. The fundamental assumption is that an individual from the EIS recipients can enter the study at any transient state (except absorbing state which is death state), without any exit/withdrawal or new entry during the study period. If overlapped states from the same individual are registered during the one-year period, we consider only a single state, which is the state that belongs to a higher rank. For example,

Revised Manuscript Received on September 10, 2019.

Shamshimah Samsuddin, School of Mathematical Sciences, Faculty of Sciences and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
(E-mail: shamshimah@tmsk.uitm.edu.my)

Noriszura Ismail, Centre for Actuarial Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.

an individual who is registered at State 2 (T) in year 2010, and is also registered at State 3 (P) in the same year, will be considered as a State 3 (P) according to higher rank. Another assumption is the health state of an individual for a specific year depend only on the health state observed in the previous year.

III. METHODOLOGY

Markov Chain (MC) model can be used to estimate the probabilities of transition between states as age or time progresses. The model has been applied in several areas, such as in mobility study between labour markets [17], multi-state analysis of marital status [19], multi-state analysis of biological production [11], and multi-state study of chronic disease [18].

Applications in the area of disability can also be found, such as in [8] who used the MC model to estimate the probabilities of transition for individual health status in the population of 40–90 year olds, and in [12] who applied the model for the United Kingdom to project the number of elderly with disabilities.

In this study, we assume that the probabilities of transition depend only on year, gender, age, as well as disability category, and thus, not taking into account the past experience and duration of disability of how or when an individual enter a category. According to [16], a Markov chain is a discrete-time stochastic process $\{X_t, t = 0, 1, 2, \dots\}$ that takes on a finite or countable number of possible values. According to [9], the advantage of discrete time model is to simplify the inferential procedures of processes with time-dependent transition probabilities, as it is easier to deal with matrix multiplication compared to the continuous time model. An approach for estimating the transition matrix of a discrete time Markov chain can be found in [7] and [3].

The set of possible values of the process is denoted by the set of non-negative integers $\{0, 1, 2, \dots\}$. The process is said to be in state i at time t when the health state is $X_t = i$. Whenever the process is in a certain state i , there is a fixed probability that it will be in state j one unit time later, P_{ij} , and that no more than one transition can occur during one year, apart from the possible death of the contributor.

Assuming a homogeneous Markov process, the matrix of transition rates, denoted as \mathbf{P} , is expressed as follows:

$$\mathbf{P} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} P_x^{1,1} & P_x^{1,2} & P_x^{1,3} & P_x^{1,4} \\ P_x^{2,1} & P_x^{2,2} & P_x^{2,3} & P_x^{2,4} \\ P_x^{3,1} & P_x^{3,2} & P_x^{3,3} & P_x^{3,4} \\ P_x^{4,1} & P_x^{4,2} & P_x^{4,3} & P_x^{4,4} \end{pmatrix} \end{matrix} \quad (1)$$

where the probability of X_{n+1} being in state j , given that X_n is in state i , is called the one-step transition probability and is denoted by ${}_{n,n+1}P_x^{i,j}$, that is, ${}_{n,n+1}P_x^{i,j} = \Pr\{X_{n+1} = j | X_n = i\}$. Since the transition probabilities are annual probabilities, the quantities P_{ij}

satisfy the conditions $P_{ij} \geq 0$ and $\sum_{j=0}^{\infty} P_{ij} = 1$ for $i, j \in \Omega$. The probabilities of transition in the MC model are easily obtained because the complete health state sequence (which is from years 2009 to 2013) is known for the entire population.

The first method in this study is the CM, which is used in [6] and [5]. Under this method, the numbers of transition from initial states to another states are counted, and then grouped according to gender and age (from 15 to 65 years old). In this study, the employees can contribute up to four states. The probabilities of transition are then obtained by observing the changes in the individual status in the 5-year

period as simple counts, $\hat{P}_{ij} = N_{ij} / \sum_{j=0}^k N_{ij}$, where $N_x^{i,j}$ is the number of transitions from state i to j of an observed aged x . A more detailed discussion on the CM can be found in [14] and [15].

IV. COUNTING METHOD (CM)

The first method is the CM, which is used in [6] and [5]. Under this method, the numbers of transition from initial states to another states are counted, and then grouped according to gender and age (from 15 to 65 years old). In this study, the employees can contribute up to four states. The transition probabilities are then obtained by observing the changes in the individual status in the 5-year period as

simple counts, $\hat{P}_{ij} = N_{ij} / \sum_{j=0}^k N_{ij}$, where $N_x^{i,j}$ is the number of transitions from state i to j of an observed aged x . A more detailed discussion on the CM can be found in [14] and [15].

V. PROPORTIONAL ODDS MODEL (POM)

The second method is based on [2] and [4]. The ordinal logistic regression model has been applied over the last few years for analyzing data where the response or outcome is presented in ordered categories [1]. There are several ordinal logistic regression models, and the most frequently used in practice is the constrained cumulative logit model called the proportional odds model (POM) [10]. The POM is appropriate for ordinal outcomes. Defining the cumulative probabilities as $w_j = p_1 + \dots + p_j$, the cumulative logit can be parameterized as

$$\log \left\{ \frac{w_j}{1-w_j} \right\} = \alpha_j + x\beta, \quad j = 1, \dots, J-1 \quad (2)$$

where each α_j is an intercept parameter for state j , and β is the vector of regression parameters. α_j must be non-decreasing in j to ensure monotonicity among the cumulative probabilities w_j . Therefore, the transition



probabilities are

$$p_j = \frac{1}{1 + e^{-(\alpha_j + x'\beta)}} - \frac{1}{1 + e^{-(\alpha_{j-1} + x'\beta)}} \quad (3)$$

In this study, the POM predicts the probabilities of transition for four health states respectively.

VI. RESULTS

Table 2 shows the demographic characteristics of the population at the entering state in year 2009. It can be noted that the total number of contributors under the EIS submitted their claims is 263,971, and categorized according to gender and age. The majority of claimants are male employees, which is 212,416 out of 263,971. The average age of claimants for male and female are 34 ± 11.65 and 35 ± 11.59 , respectively.

Table 2 Summary of Descriptive Statistics at entering state, 2009

N:263,971			
Male		212,416	Female 51,555
Age	Min	34	Age Min 35
	Median	33	Median 34
	Mode	23	Mode 21
	Standard deviation	11.65	Standard deviation 11.59

The number of employees according to gender and health status from 2009 until 2013 are shown in Table 3. State 1 (A) consists of employees who do not lodge any reports of accidents at work, or death. It can be noted that State 2 (TD) shows increasing number of both male and female employees from year to year. The number of male employees increases from 31,329 in year 2009 to 33,764 in 2013, while the number of female employees increases from

7,193 in 2009 to 8,981 in 2013. State 3 (permanent disability) also shows increasing number of employees for both gender throughout the years, except for female employees where the number declines in 2013. State 4 (death) shows slightly increasing numbers for both genders in 2011, but the numbers decline over the next two years (2012 and 2013). Since 2009 is considered as the beginning year, information on the number of deaths is not recorded for that year

Table 3 Number of Employees (Age: 15-65)

Gender	Health State	Year				
		2009	2010	2011	2012	2013
Male	1	169,609	166,825	163,869	162,521	161,526
	2	31,329	31,881	33,014	33,338	33,764
	3	11,478	12,683	13,424	13,609	13,443
	4	-	1,027	082	839	735
Female	1	42,188	41,330	40,484	39,716	39,517
	2	7,193	7,674	8,224	8,730	8,981
	3	2,174	2,448	2,640	2,805	2,664
	4	-	103	104	97	89

Table 4 shows the results of transition probabilities from both methods, the CM and the POM. The results from both methods are similar (up to two decimal places). The transitions probabilities of state 1 (A) are larger than the transition probabilities of other states (states 2, 3 and 4), implying that a large proportion of disabled workers have improve their health conditions, and are able to work. However, several workers who are in state 2 (T) have deteriorate to permanently disabled (state 3), as shown from the probabilities from state 2 to state 3, which are 0.0107 for male and 0.0051 for female.

Table 4 Transition Probabilities from Counting Method, CM and POM (Age: 15-65)

Method	Male				Female					
	Final State									
		1	2	3	4	1	2	3	4	
CM	1	0.7267	0.1906	0.0773	0.0054	0.7325	0.2017	0.0634	0.0024	
	2	0.9451	0.0435	0.0107	0.0006	0.9734	0.0218	0.0048	0.0001	
	3	0.9548	0.0225	0.0223	0.0004	0.9692	0.0178	0.0128	0.0002	
	4	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	
POM	Initial State	1	0.7259	0.1941	0.0752	0.0048	0.7357	0.1879	0.0718	0.0046
		2	0.9451	0.0436	0.0107	0.0006	0.9734	0.0213	0.0051	0.0003
		3	0.9548	0.0229	0.0219	0.0004	0.9693	0.0157	0.0148	0.0003
		4	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000

VII. DISCUSSION

The use of MC model in population studies has been extensively studied over past years. The main advantage of using the Markov chain model is that it can applied for studying multiple populations. The observations of all state transitions must be included as it is the most important and informative data for Markov chain model. In this study, we used the EIS dataset from SOCSO to estimate the Markov transition probabilities between health states in Malaysia. This study implements two different methods, the CM and the POM, to form the Markov probabilities. The results show that both methods produced similar results, which are also consistent with other study carried out in [8]. The results also show that gender has greatly affect the probabilities of transition.

VIII. ACKNOWLEDGEMENTS

The authors are indebted to the staff of the Social Security Organization (SOCSO) for providing the EIS data to make this paper possible. This research also would not have been possible without the sponsorship from Universiti Teknologi MARA (UiTM) and Universiti Kebangsaan Malaysia (UKM). This research was funded by Universiti Kebangsaan Malaysia (UKM-GUP-PI-08-34-318) and Ministry of Higher Education (FRGS/1/2015/SG04/UKM/02/2).

REFERENCES

1. Abreu, Mery Natali Silva, Siqueira, Arminda Lucia, & Caiaffa, Waleska Teixeira, 2009. Ordinal logistic regression in epidemiological studies. *Revista de Saude publica*, 43(1), 183-194.
2. Cole, Bernard F, Bonetti, Marco, Zaslavsky, Alan M, & Gelber, Richard D., 2005. A multistate Markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Statistics in Medicine*, 24(15), 2317-2334.
3. Craig, Bruce A, & Sendi, Peter P., 2002. Estimation of the transition matrix of a discrete-time Markov chain. *Health economics*, 11(1), 33-42.
4. De Jong, Piet, & Heller, Gillian Z., 2008. *Generalized linear models for insurance data* (Vol. 10): Cambridge University Press Cambridge.

5. Diehr, Paula, & Patrick, Donald L., 2001. Probabilities of transition among health states for older adults. *Quality of Life Research*, 10(5), 431-442.
6. Diehr, Paula, Patrick, Donald L, Bild, Diane E, Burke, Gregory L, & Williamson, Jeff D., 1998. Predicting future years of healthy life for older adults. *Journal of clinical epidemiology*, 51(4), 343-353.
7. Fleming, Thomas R, & Harrington, David P., 1978. Estimation for discrete time nonhomogeneous Markov chains. *Stochastic Processes and their Applications*, 7(2), 131-139.
8. Jung, Juergen., 2006. *Estimating Markov transition probabilities between health states in the HRS dataset*. Indiana University.
9. Lievre, Agnes, Brouard, Nicolas, & Heathcote, Christopher., 2003. The estimation of health expectancies from cross-longitudinal surveys. *Mathematical population studies*, 10(4), 211-248.
10. McCullagh, Peter., 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109-142.
11. Park, Wonjun, & Moon, Il., 2007. A discrete multi states model for the biological production of hydrogen by phototrophic microalga. *Biochemical engineering journal*, 36(1), 19-27.
12. Rickayzen, Ben D., and Duncan EP Walsh., 2002. A Multi-state Model of Disability for the United Kingdom: Implication for Future Need for Long-term Care for the Elderly *British Actuarial Journal*, 8(2), 341-393.
13. Samsuddin, Shamshimah, & Ismail, Noriszura., 2015. Isu Hilang Upaya Dikalangan Pencarum PERKESO di Malaysia. *Malaysia Labour Review*, 11(No. 2), 85-94.
14. Samsuddin, Shamshimah, & Ismail, Noriszura., 2016. Multi-state Markov model for disability: A case of Malaysia Social Security (SOCSO). Paper presented at the INNOVATIONS THROUGH MATHEMATICAL AND STATISTICAL RESEARCH: Proceedings of the 2nd International Conference on Mathematical Sciences and Statistics (ICMSS2016).
15. Samsuddin, Shamshimah, & Ismail, Noriszura., 2017. Transition probabilities of health states for workers in Malaysia using a Markov chain model. Paper presented at the AIP Conference Proceedings.
16. Taylor, Howard M, & Karlin, Samuel., 1998. *An introduction to stochastic modeling*: Academic press.
17. Uhlendorff, Arne., 2006. From no pay to low pay and



- back again? A multi-state model of low pay dynamics.
18. van de Kastelee, Jan, Hoogenveen, RT, Engelfriet, PM, Van Baal, PHM, & Boshuizen, HC., 2012. Estimating net transition probabilities from cross-sectional data with application to risk factors in chronic disease modeling. *Statistics in medicine*, 31(6), 533-543.
 19. Willekens, Frans J, Shah, I, Shah, JM, & Ramachandran, P., 1982. Multi-state analysis of marital status life tables: theory and application. *Population Studies*, 36(1), 129-144.