İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 14, 2021, 2, 30-43

Geliş/Received:04.10.2021, Kabul/Accepted: 26.12.2021

Araştırma Makalesi / Research Article

Longevity Risk and Modelling in The Life and Pension Insurance Company: Mortality Forecasting with the Lee–Carter Method

İsmail YILDIRIM
Hitit University, Çorum, Türkiye
ismailyildirim@hitit.edu.tr
orcid.org/0000-0002-6408-0332

Abstract

Longevity risk is exactly the opposite of mortality risk and indicates that live longer than life expectancy has a cost for insurance companies. Longevity risk is one of the important topics which take part in actuary literature. One of the most widely used model is Lee-Carter (LC) model which allow to be expressed as a stochastic process of mortality models.

The study was carried out in order to model male and female mortality rates in Turkey by means of Lee-Carter (LC) method and in order to make predictions for the future. Thus, male and female death rates associated with age between 1950-2020 years of Turkish statistical institute in Turkey was used as data. At the end of study, it was found that death rate of men may be more than those females for the future.

Keywords: *Insurance, Longevity Risk, Mortality Risk, Lee Carter Method.*

Introduction

Life and pension insurance companies face with two important actuarial risks when conducting insurance operations. First one is the longevity risk life and pension companies are exposed to with regards to their insurance products based on the longevity of the insured. The second, on the other hand, is the mortality risk which may arise due to reduced mortality rate of the insured.

Longevity Risk may arise when individuals in a specific group live longer than they were expected to. It is the result of different outcomes obtained for longevity and mortality risk when compared to the assumptions insurance companies use when they assign prices for their products. In other words, longevity risk is the case when more than expected number of insured survives in terms of annuity products (Black and Skipper, 2000: 161). Private insurance companies, particularly life and pension

insurance companies along with social security institutions, take heed of longevity risk when developing their long-term plans and programs.

Longevity risk arises from the ambiguity of the survival of individuals which can be calculated using the mortality rate of the society in question reflecting a somewhat acceptable estimate for each individual (Hanewald; Piggott; Sherris, 2013: 88). On the other hand, more individuals may live longer than expected or the mortality rate may be lower than expected. Especially as the developments in health and medicine technologies may lead to short-term (i.e., one or two years) or long-term (10+ years) increase in the human life expectancy, this may lead to longevity risk (Coxa; Lin; Pedersen, 2010: 243). However, longevity risk is a long-term risk, it is mostly necessary to be evaluated for time periods as short as one year (Richards; Currie; Ritchie, 2012: 1).

Longevity risk is an important risk factor for life and pension insurance companies and it may have a significant impact on the risk status (Gatzert and Wesker, 2012: 1). It is known that mortality rate is related with socioeconomic and behavioral risk factors and this fact is taken into consideration in actuarial calculations for life insurance policies. It is of utmost importance to understand mortality fully for product development in life insurance, pricing, evaluation and profitability analysis (Kwon and Jones, 2008: 394). Life expectancy of insured is a source of costs for the insurance company in terms of health expenses, medical treatment expenses, etc. during the pension (Brouhns, Denuit and Vermunt, 2002:373). Increased life expectancy is a serious problem for life insurance companies and pension funds which are oblige to make monthly pension payments. In the light of these issues, it is important to take measures against this kind of a risk.

Lee and Carter (1992) offered a model which uses age and time factors in combination in order to model and estimate the age-specific mortality rates. Developed in 1992, this method is recognized as one of the most commonly used methods for the mortality rate estimations. Following the research conducted by Lee and Carter, Lee and Miller (2001), Booth et al. (2002), De Jong and Tickle (2006), and Hyndman and Ullah (2007) have further developed this method with several additions.

Among the research which estimates mortality rates in Turkey are Gençtürk and Genç (2012), and Demircioğlu and Büyükyazıcı (2013). In their study, Gençtürk and Genç (2012) estimated the mortality rates based on the mortality statistics available in Turkey using Trend method and Lee-Carter method and they have compared the results obtained from these methods. It was found that mortality rates obtained from Lee-Carter and Trend methods were similar while these two methods were more compatible for the data obtained from women. Demircioğlu and Büyükyazıcı (2013) estimated the age-specific mortality rates in Turkey using the Lee-Carter method and Poisson Log-Bilineer approach which was developed as

an alternative to Lee-Carter method. Researchers have compared the suggestions obtained from both methods and found that these two methods gave somewhat different results.

Wiśniowski et al., (2015) estimated the population changes in England until the year 2024 using the Lee-Carter method. This study modeled the population movements such as age-specific birth, mortality and immigration. Danesi, Haberman and Millossovich (2015) have estimated the mortality improvement rates using the Lee-Carter method based on the mortality data obtained from Italy between 1974 and 2008. In their study, Richards and Currie (2009) estimated the mortality rates for England and Whales using the Lee-Carter method with the data collected from a period between 1961 and 2006. The study showed that the Lee-Carter method should also take heed of liabilities such as pension payments. Antolin (2007) has conducted a mortality modeling study for some of the OECD countries (Canada, France, Germany, Italy, Mexico, United Kingdom, United States of America).

This study conducts a mortality modeling for Turkey using the Lee-Carter model, one of the most important mortality models for life insurance products and social security insurance policies priced with estimated longevity of individuals and masses. Data on the mortality rates of Turkey based on Turkish Statistical Institute (TUIK) for a period between 1950 and 2010 were taken as a reference point. Future estimations were then addressed calculating the longevity and mortality rates for the periods after 2010. This study includes longer-term estimations compared to other studies conducted in Turkey. By using historical data, life and death times in Turkey until 2060 were calculated. These results will guide insurance companies operating in Turkey in presenting their retirement plans.

Mortality Models and Lee-Carter (LS) Method

However, there is no universal method for mortality rate estimations, the method to be used is determined considering criteria such as accuracy, reliability, and simplicity. Mortality modeling most commonly uses variables such as age and gender while it is obvious that there are other factors affecting the mortality when health and mortality statistics are investigated. The effects of these descriptive variables on mortality model are disregarded as data is not available or insufficient for that variable. With this variables disregarded, an assumption of homogeneity of variance is not possible for mortality models commonly used in the literature. Some unexpected results may be obtained when the assumption of homogeneity of variance is not established (Gençtürk and Genç, 2012: 64)

Several methods are developed in order to estimate the mortality rates and pattern and also to anticipate the future mortality. The efforts to interpret mortality in the form of a curve go back to the 19th century, starting from the work of Gompertz, the "Law of Mortality". These first trials on mortality curve tried to divide mortality in early age, middle age and advanced age, and merely considered the age as a factor in

interpreting the mortality (Hári et al., 2007). In the recent years, accurate and reliable calculations in mortality models have gained importance for actuaries and policy-makers. Many models are developed for mortality calculations in time and it was found that stochastic models are most commonly preferred as they give more accurate and reliable outcomes as a result of the studies conducted in this respect (Koissi and Shapiro, 2008).

In 1992, Ronald Lee and Lawrance Carter have offered a model which includes age and time factors for age-specific mortality rate modeling and estimations (Lee and Carter, 1992). This model utilizes a time series model which is able to reflect the changes in the mortality rates of the past to a model. The estimations of mortality rate and life expectancy are in correlation with the estimation of the time-dependent mortality levels (Lee and Carter, 1992).

Lee-Carter model is a mortality projection approach which explains the multiplication of components namely a parameter which varies in time and reflects the general mortality rate as the log of age-specific death rate and a parameter which defines the speed of change in general mortality level for each age group and the addition of time-independent age-specific component values. This method suggests a linear approach with variables such as x (age) and t (time) and builds on the future estimations of mortality rates based on the recorded past mortality rates. (Haberman and Russolillo, 2005: 2-3).

Lee-Carter model has found a widespread use in this field in the last 18 years and it can be said that it is the golden standard of our time in anticipating the mortality using models (Li and Chan, 2007: 68).

Mortality Modeling

Let $\mu_x(t)$, represent the instantaneous death rate of an individual at the age of *x* and time of *t*. Then the probability of the deaths in that year is calculated (Kogure and Kurachi, 2010:162):

$$q_x(t) = 1 - \exp\left\{-\int_0^1 \mu_{x+s}(t+s)\mathrm{ds}\right\}.$$

$$\mu_{x+s}(t+u) = \mu_x(t) \qquad (1)$$

For x and t integers $0 \leq s, u < 1$. Exponential of death is constant in age and time.

$$q_x(t) = 1 - \exp\{-\mu_x(t)\}.$$

Lee-Carter method models the mortality rates.

$$q_x(t) = 1 - \exp\{a_x + \beta_x k_t\},$$

First, it is necessary to estimate the parameters.

$$q_x(t) = 1 - \exp\left\{-\exp\{a_x + \beta_x k_t\}\right\}$$

It is assumed that *x* number of individuals will be alive for *t* years.

$$ {}_tp_x(t_0) = \left(1 - q_x(t_0)\right) \mathrm{x} \left(1 - q_{x+1}(t_0 + 1)\right) \qquad (2)$$

$$\mathrm{x} \ldots \mathrm{x} \left(1 - q_{x+t-1}(t_0 + t - 1)\right)$$

$$= \exp\left\{-\sum_{j=0}^{t-1} \exp\left\{a_{x+j} + \beta_{x+j} k_{t_{0+j}}\right\}\right\}$$

Model Development

Lee-Carter Model is described as follows;

$$\mathrm{In}\left(m_{x,t}\right) = a_x + b_x k_t + \epsilon_{x,t},$$

$m_{x,t}$: represents the approximate mortality rate for the age of *x* at the time of *t* (Chan, 2013: 19-20).
a_x : Age-specific parameter; this set {a_x , *x* =0, 1 ,……..….} represents the general outlook of the death chart.

k_t : Time-dependent parameter; k_t represents the trend of the improvement in mortality in time.

b_x : Age-specific parameter; characterizes the k_t sensitivity at the age of *x*.

$\epsilon_{x,t}$: represents the error term.

Time-dependent parameter, k_t, represents the change in the logs of death rates in time. However, one should not expect to have similar changes in mortality values for every age group. It is not possible for any effect which may occur at the general level of mortality to be observed in the same pattern at any age. Therefore, any decrease in the general level of mortality will affect any age in a different manner. It is the b_x parameter which makes it possible to separately define the effects of these changes for each age group. a_x parameter, on the other hand, is the age-specific mortality level which is found using the averages over

the years for each age group and therefore is separated from the *t* index (Brouhns, Denuit, and Vermunt (2002).

For the estimation of the parameters b_x, which is the age-specific pattern of mortality that changes according to years, and k_t, which shows the variation of mortality by years, the $\epsilon_{x,t}$ matrix is obtained by subtracting the a_x vector from the logarithmic mortality matrix. In the application, the TDA method used by Lee and Carter in their original work and the two-stage estimation method were used to find the bx and k_t vectors. As a first step, b_x and kt values were obtained by applying the TDA method to the $\epsilon_{x,t}$ matrix. In this decomposition process, the Biplot software, which was prepared as an add-on to the Microsoft Excel program, and the Matlab program were used.

Application

Data Collection

Population related mortality rates are regularly announced by the Turkish Statistics Institution (TUIK). Data collection involved the censuses held by TUIK in Turkey in 1960, 1965, 1970, 1975, 1980, 1985, 1990, 1995, 2000, 2005, 2010 and 2020, then the mortality rates were calculated and the results were modeled using the Lee-Carter model. Data on the age-specific mortality rates of Turkey was not made available by Turkish Statistical Institute (TUIK) between 1960 and 1965. Data for this period was obtained from the research conducted by Yıldırım (2010). Data for the remaining years were obtained from the official website of TUIK.

Mortality Modeling Using the Lee-Carter Method

Mortality rate in Turkey was adapted to the TUIK data using the Lee-Carter method. Gender-specific mortality data of Turkey for the period between 1950 and 2020 was used. A data matrix was established for year and age values following the calculation of the conversions of the mortality rates in Turkey for ages between 1 and 85, both Women and Men.

The results of the model parameters are shown below in Figures based on the data collected between 1950 and 2020.

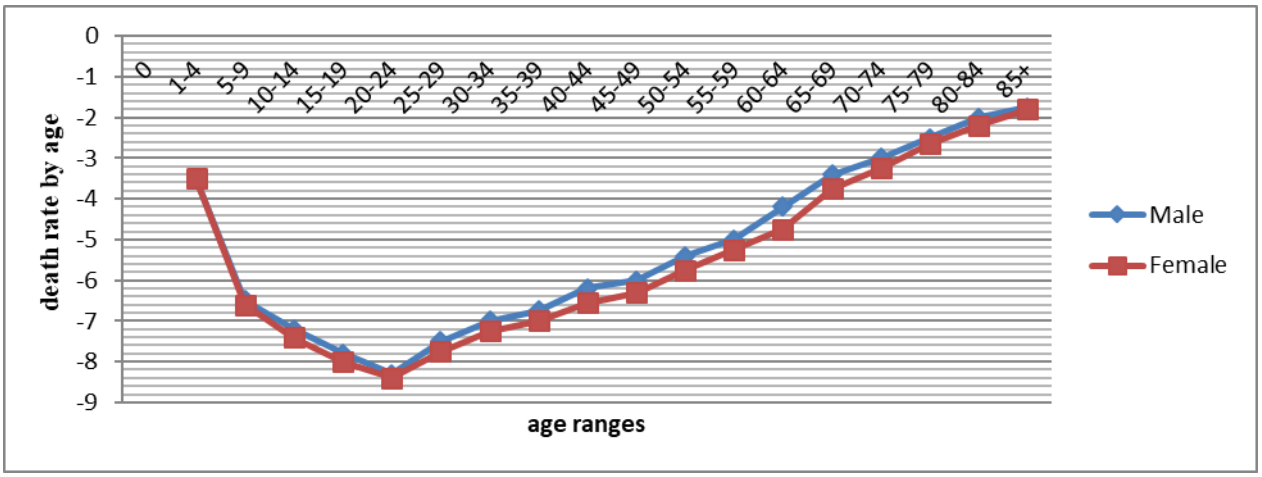

Figure 1: a_x parameter value

When the age-specific parameter of mortality, ax, is investigated (Figure 1), it was found that the mortality rate decreases between the age groups of 0 and 20-24 for both Women and Men. It is observed that mortality rates of early ages are lower than advanced ages for both genders and it turns into an increasing trend especially after the age group of 20-24. Nevertheless, it was found that mortality of men is higher than mortality of women for all age groups (0 to 85) according to the time-independent and age-specific parameters.

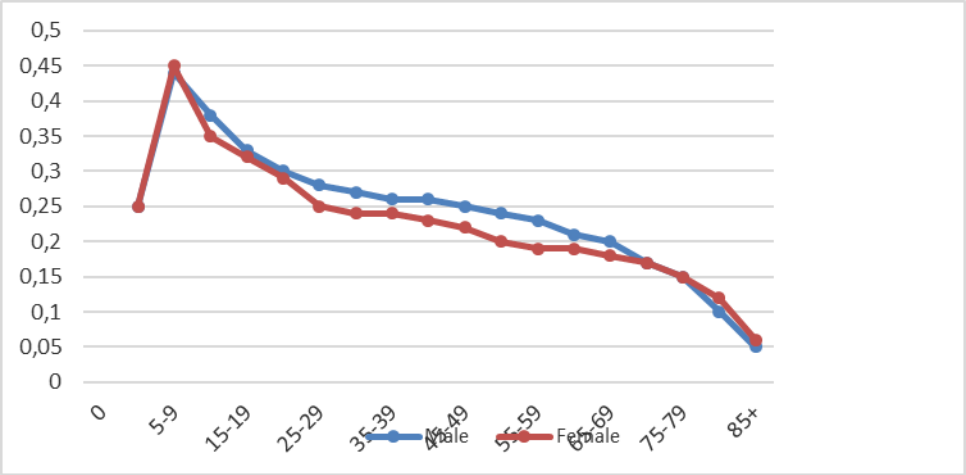

Figure 2: b_x parameter value

Figure 2 shows the results obtained for the bx parameter which gives the mortality rates for each age group. A review of the parameters shows that parameters have positive values for each age group. These results reveal that mortality rates in Turkey are decreasing for all age groups. They especially show that mortality rate will not increase for any age group based on a change in the mortality structure. According to these results, mortality rate is decreasing particularly for the age group of 1-4 and the mortality rate for this age group will be reduced in the future. This result confirms the reduced number of infant deaths in

Turkey reported in the recent years. bx values follow a decreasing pattern for the advanced age groups between 1-4 and 85. These results indicate that the effect of the change in the mortality rate will be reduced between age groups. Mortality rate increases in a dramatic manner especially with the age group of 50.

kt represents the general trend of mortality improvement in time. Calculated for mortality rates of both men and women, kt values are given in Figures 3 and 4.

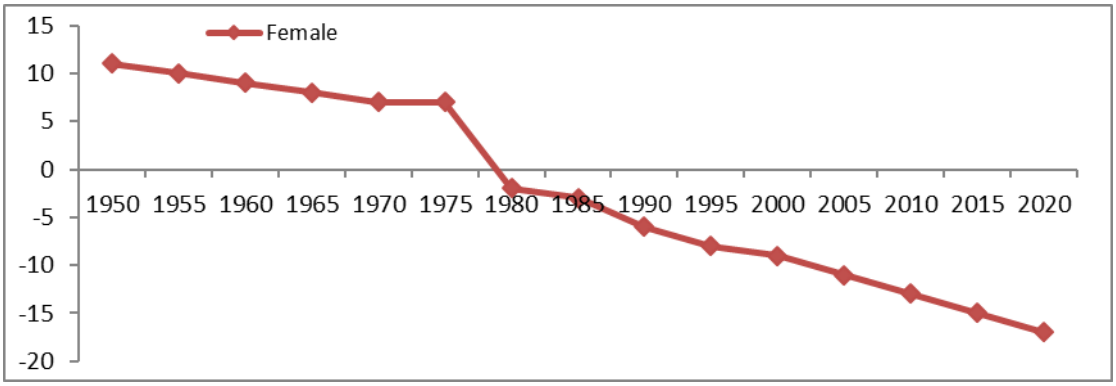

Figure 3: Change of k_t parameter according to years (female)

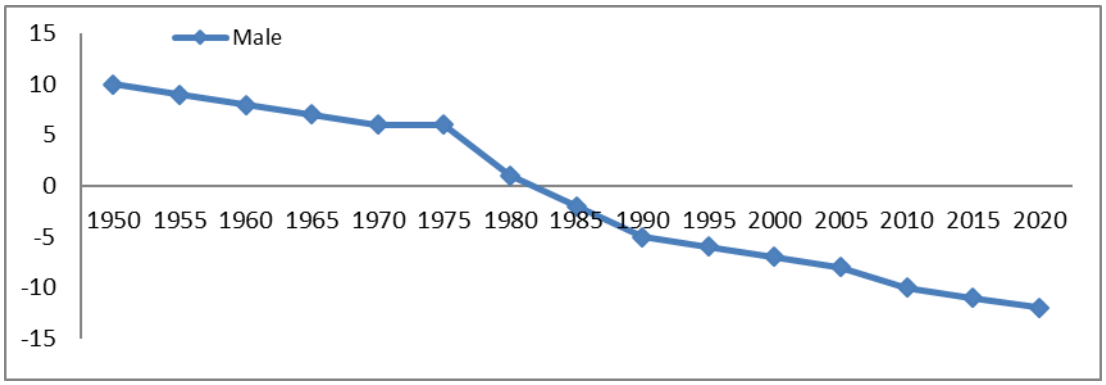

Figure 4: Change of k_t parameter according to years (male)

(kt) values show a similar pattern for both men and women. Ignoring the small changes, kt values obtained from men and women does not show a significant deviation. While having a similar pattern for both men and women, (kt) values are also irregular. In general, kt values have a decreasing trend for both men and women. It can be observed that these values have been decreasing for women as of 1975 and for men as of 1980.

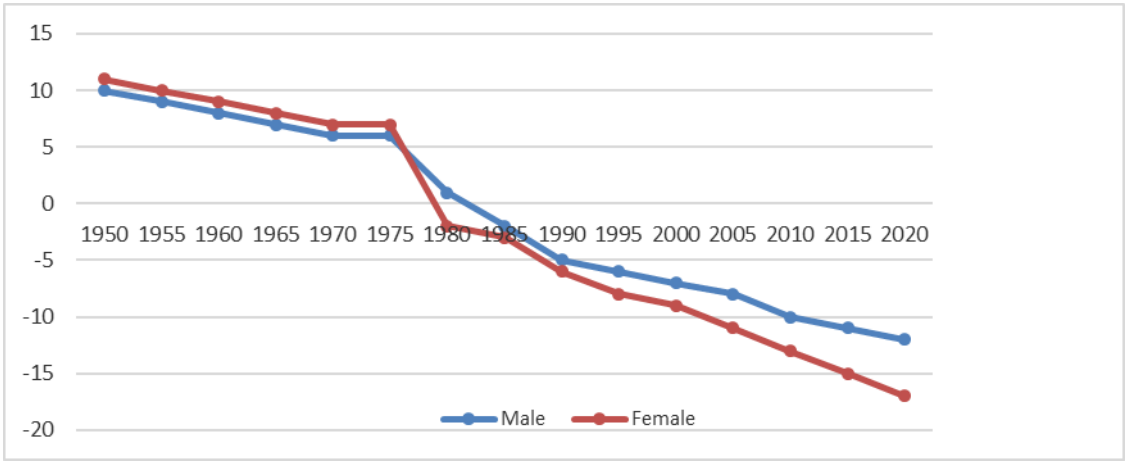

Figure 5: Comparison of $\boldsymbol{k_t}$ parameters by gender

kt, which is the death rate indicator in the Lee-Carter Model, shows the change in mortality rates over the years for all ages. The first stage kt estimation values obtained as a result of TKA are shown in Figure 5. The first-stage cut-off values of the kt parameter, which expresses the course of mortality over the years, show that the mortality of the Turkish population has decreased over the years for both men and women.

Mortality Estimation Modeling for the Future

Projections were created for the period between 2010 and 2060 according to the life expectancy, and age groups obtained from Lee-Carter method using the data recorded between 1950 and 2020. Figures 5 and 6 show the men and women mortality rates until 2060 based on the approximate death rates recorded between 1960 and 2020 in Turkey.

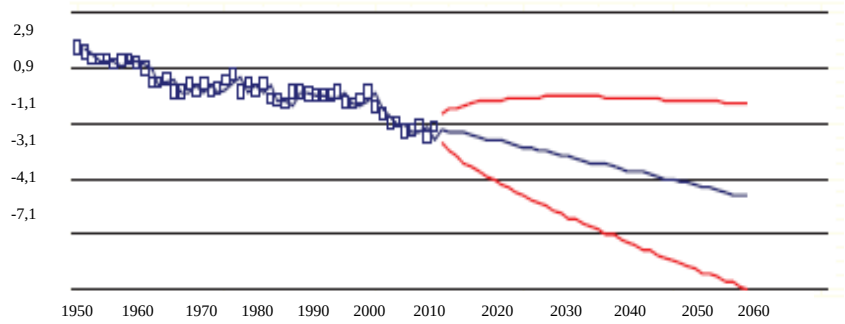

Figure 6: Projection Results for Women Until 2060

The projection results for women for the period between 2010 and 2060 show that these results comply with the data obtained from the period between 1950 and 2010 with a decreasing trend and the results are in a 95% confidence interval. Women mortality rates have been decreasing starting from 1950 and the

trend is continued until 2060. Mortality rates have been rapidly decreasing especially after 2000s with negative values.

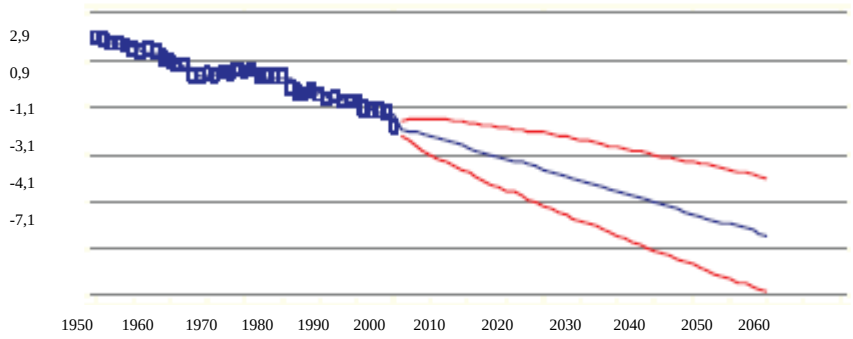

Figure 7: Projection Results for Men Until 2060

A review of the projection results for men until 2060 showed that these results are similar with the ones obtained for women. The confidence level of the projection results for men between 2010 and 2060 was 95% and the mortality rate follows a decreasing pattern.

Figure 8 shows the life expectancy estimations for both men and women in a time-depenndent manner using the Lee-Carter method.

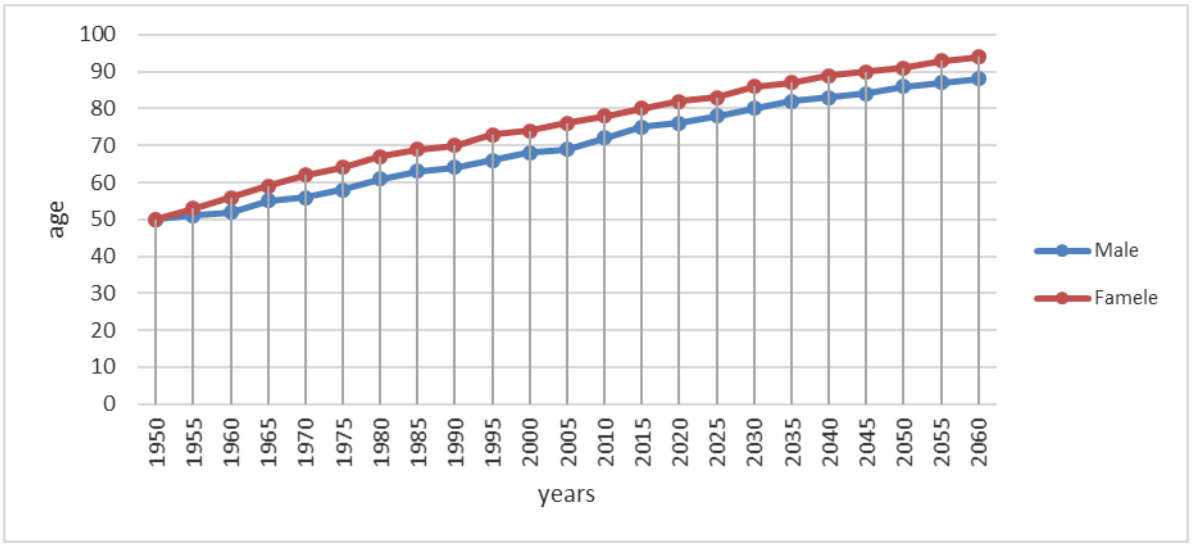

Figure 8: Life Expectancy Using Lee-Carter method.

Scientific research showed that life expectancy of women is longer than the life expectancy of men. Especially in Turkey, the life expectancy of women is longer than the life expectancy of man according to the data collected from TUIK. Based on the results of the estimations using Lee-Carter method, it was

found that the life expectancy of women is consistently longer then the life expectancy of men between the years 1950 and 2060.

As of 2010, life expectancy of women is 70+ and it increases consistently. It is expected that the average life expectancy of women in Turkey to become 80+. Life expectancy of men, on the other hand, has also been increasing yet it was shorter than the life expectancy of women.

Results and Discussion

Life and pension insurance companies face a number of insurance risks due to their operations. The subject of the insurance policies life and pension companies create is humans. Based on the human life, these kinds of insurances are exposed to several risks. The most important of these risks are longevity risk and mortality risk. Insurance companies must estimate mortality rates and life expectancies in order to be able to fulfill their obligations towards the insured. The mortality rate and life expectancy play an important role in the determination of the amount of insurance premiums to be paid for a life insurance. It is vital that life insurance companies and social security institutions make successful predictions about future death rates so that they can meet their financial obligations to their commitments. Because wrong estimations may cause the insured to pay more premiums unnecessarily by increasing premiums, or may cause institutions or companies to be insufficient in meeting their financial obligations with less premiums than they should be.

The estimation of mortality rates is an important subject of the actuarial sciences. Estimations of mortality rate are commonly used in many fields. The most important feature of mortality rates and life tables is that they are developed as a result of reliable calculations. It is of utmost importance to estimate mortality rates of a country based on its specific population and mortality statistics.

Lee-Carter method is one of the most commonly used methods in the actuary calculations related with the mortality rates. Lee-Carter method has been reported to be a successful resource since 1992 and has been used by several countries in order to identify their mortality patterns.

This study estimates the mortality rates based on the population and death statistics of Turkey using the Lee-Carter method, a stochastic method used to model mortality rates.

Mortality rates exhibit a decreasing trend for the age groups between 0 and 20-24 for both men and women between the ages of 0 to 85 and living in Turkey. It is observed that mortality rates of early ages are lower than advanced ages for both men and women and it turns into an increasing trend especially after the age group of 20-24. However, mortality rates of men and women are similar in their trends, it was found that mortality rate of men is higher than the mortality rate of women for every age group.

The results obtained for all parameters were positive for each age group when mortality rates are investigated per age group. It can be observed that mortality rates in Turkey are decreasing for all age groups. They especially show that mortality rate will not increase for any age group based on a change in the mortality structure. It was shown that the mortality rate is decreasing in Turkey for the age group of 1-4 while the mortality rate increases for the age group of 50+.

When approximate mortality rate of both men and women is considered, it can be observed that these values have been decreasing for women as of 1975 and for men as of 1980. The projection results for men and women between 2010 and 2060 shows that the mortality rate will be decreasing as it was the case in the previous years. An increase in the life expectancy of both men and women is estimated while the life expectancy of women is anticipated to be higher than the men.

Several issues may arise in the estimation of mortality rate and life expectancy in Turkey, as it was the case also in this study. Without a regulated recording system, numerically insufficient mortality data and unreliable results are inevitable. Application results and the deviations in the findings of further research will show the significant data insufficiency for the mortality statistics available in Turkey. Consistency for the future in modeling and foresight studies Finding predictions depends on the goodness of fit of the model obtained with the method used to the data, as well as how accurate and explanatory the data used in the study is. Therefore, for a more accurate estimation of the general population mortality in Turkey, obtaining data on mortality indicators is of great importance.

REFERENCES

Antolin, P. (2007). Longevity Risk and Private Pensions, OECD Working Papers on Insurance and Private Pensions, No. 3, OECD Publishing.

Arık, A. ve Sucu, M. (2011). Uzun Ömürlülük Bonolarını Fiyatlandırma: Uç Değer Kuramı ve Kübik Risk Fiyatlandırma Modeli, İstatistikçiler Dergisi, 4 (2011) 69-85.

Black, K. and H.D. Skipper.(2000). Life & Health Insurance, Prentice Hall, NJ, USA.

Booth, H., Maindonald, J. and Smith, L. (2002). Age-Time Interactions in Mortality Projection: Applying Lee-Carter to Australia, ANU, Demography and Sociology Program, Research School of Social Sciences, Working Papers No. 85.

Brouhns, N., Denuit, M., Vermunt, J. (2002) A Poisson Log-Linear regression approach to the construction of projected life tables, Insurance: Mathematics and Finance, 31, 373-393.

Butt Z. Haberman S. (2004). Application of Frailty-Based Mortality Models Using Generalized Linear Models, ASTIN Bulletin, Vol 34, Issue 1, 175-197.

Chan, W.S. (2013) Longevity Risk: Recent Developments and Actuarial Implications, The Chinese University of Hong Kong, Master Thesis, A SOA Center of Actuarial Excellence.

Coxa, S. H., Y. Lin and H. Pedersen. (2010). Mortality Risk Modeling: Applications to Insurance Securitization, Insurance: Mathematics and Economics, 46, 242-253, journal homepage: www.elsevier.com/locate/ime

Currie, L.D. (2013). Smoothing Constrained Generalized Linear Models With an Application to the Lee-Carter Model. Statistical Modelling 2013; 13(1): 69–93.

Danesi, L.I., Haberman, S., Millossovich, P. (2015) Forecasting mortality in subpopulations using Lee–Carter type models: A comparison, Insurance: Mathematics and Economics 62 (2015) 151–161.

De Jong, P. and Tickle, L. (2006). Extending Lee-Carter Mortality Forecasting, Mathematical Population Studies 13(1), 1–18.

Demircioğlu, S., Büyükyazıcı, M. (2013). Poisson log-bilineer yaklaşımıyla Lee-Carter modellemesi ve Türkiye Uygulaması . İstatistikçiler Dergisi:İstatistik ve Aktüerya , 6 (1) , 14-40 . Retrieved from https://dergipark.org.tr/tr/pub/jssa/issue/10048/123893

Gatzert, N. and Wesker, H. (2012). Mortality Risk and its Effect on Shortfall and Risk Management in Life Insurance, The Journal of Risk and Insurance, Vol. 00, No. 0, 1-34.

Gençtürk, Y., Genç, T. (2012). Türkiye İl-İlçe Merkezlerindeki Ölüm Oranlarının Trend ve Lee-Carter Yöntemleri ile Tahmini, Anadolu Unıversıty Journal Of Scıence And Technology. Cilt/Vol.: 2-Sayı/No: 1: 63-74.

Haberman, S. Russolillo, M. (2005). Lee-Carter Mortality Forecasting: Application to the Italian population, City University – Actuarial Research Paper No. 167.

Hanewald, K., J. Piggott and M. Sherris. (2013). "Individual Post-retirement Longevity Risk Management under Systematic Mortality Risk". Insurance: Mathematics and Economics 52, 87–97.

Hàri, N. Waegenaere, A.D. Melenberg, B. and Nijman, T.E. (2007). Estimating the Term Structure of Mortality. Insurance: Mathematics and Economics. 42, 492-504.

Heale, B. (2013). Stress and Scenario Testing: How Insurers Compare With Banks, Stress Testing: European Edition. Moody's Analytics. Vol 1, September, p.80-86.

Hyndman, R.J. and Ullah, M.S. (2007). Robust Forecasting of Mortality and Fertility Rates: A Functional Data Approach, Computational Statistics and Data Analysis 51(10), 4942–4956.

Ilic, M.M., V.Avdalovic and M. D. Obadovic. (2011). Development of Model for Insurance Risk Management and its Application to Insurance Companies Operating in the Serbian Market, Journal of Business Management and Economics, 2(6), 223-228

Kogure, A. and Kurachi, Y. (2010). A Bayesian Approach to Pricing Longevity Risk Based on Risk-Neutral Predictive Distributions. Insurance: Mathematics and Economics. 46 (2010) 162-172.

Koissi, M.C. and Shapiro, A.F. (2008). The Lee-Carter Model Under The Condition of VariablesAge-Specific Parameters, Actuarial Research Conference, Regina, Canada, 2.

Kwon, H.S. and Jones, B. L. (2008). Applications of a Multi-State Risk Factor/mortality Model in Life Insurance, Insurance: Mathematics and Economics, 43, 394-402.

Lee, R. and Miller, T. (2001). Evaluating the Performance of Lee-Carter Mortality Forecasts, Demography 38(4), 537-549.

Lee, R.D. and Carter, L. (1992). Modeling and Forecasting U.S. Mortality, Journal of the American Statistical Association. 87, 419.

Richards, S.J., Currie, I.D. (2009) Longevity Risk and Annuity Pricing With the Lee-Carter Model, British Actuarial Journal, 15, II, 317-365.

Richards, S.J., I. D. Currie and G. P. Ritchie. (2012). A Value-at-Risk Framework for Longevity Trend Risk, A Discussion Paper, Presented to the Institute and Faculty of Actuaries, London.

Wiśniowski, A., et al., (2015) Bayesian Population Forecasting: Extending the Lee-Carter Method, Demography (2015) 52:1035–1059.

Yıldırım, F. (2010) Türkiye Ölümlülük Yapısının Lee-Carter Ve Bulanık Lee-Carter İle Modellenmesi, Yayınlanmamış Yüksek Lisans Tezi, Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü.

İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 14, 2021, 2, 44-57

Geliş/Received:08.10.2021, Kabul/Accepted: 07.12.2021

Araştırma Makalesi / Research Article

Bayesian modelling of statistical region- and family-level clustered ordinal self-rated health outcome data from Turkey

Özgür ASAR

Department of Biostatistics and Medical Informatics,
Acıbadem Mehmet Ali Aydınlar University, 34752
İstanbul, Turkey
ozgurasarstat@gmail.com
orcid.org/0000-0003-0603-1409

Abstract

This study is concerned with the analysis of three-level ordinal outcome data with polytomous logistic regression in the presence of random-effects. It is assumed that the random-effects follow a Bridge distribution for the logit link, which allows one to obtain marginal interpretations of the regression coefficients. The data are obtained from the Turkish Income and Living Conditions Study, where the outcome variable is self-rated health (SRH), which is ordinal in nature. The analysis of these data is to compare covariate sub-groups and draw region- and family-level inferences in terms of SRH. Parameters and random-effects are sampled from the joint posterior densities following a Bayesian paradigm. Three criteria are used for model selection: Watenable information criterion, log pseudo marginal likelihood, and deviance information criterion. All three suggest that we need to account for both region- and family-level variabilities in order to model SRH. The extent to which the models replicate the observed data is examined by posterior predictive checks. Differences in SRH are found between levels of economic and demographic variables, regions of Turkey, and families who participated in the survey. Some of the interesting findings are that unemployed people are 19% more likely to report poorer health than employed people, and rural Aegean is the region that has the least probability of reporting poorer health.

Keywords: Bayesian statistics, categorical data analysis, income and living conditions, latent-variable models, multi-level analysis, self-rated health.

Öz

Türkiye istatistiki bölge ve aile düzeyinde kümelenmiş sıralı algılanan sağlık düzeyi sonuç verisinin Bayesçi modellemesi

Bu çalışma, üç seviyeli sıralı sonuç verisinin, rastgele etkili terimler içeren polytomous lojistik regresyon ile analizi üzerinedir. Rastgele etkili terimlerin, regresyon katsayıları için marjinal yorumlar elde edilebilmesini mümkün kılan logit linki için Bridge dağılımını takip ettikleri varsayılmıştır. Veri Türkiye Gelir ve Yaşam Koşulları Çalışması'ndan elde edilmiştir. Sonuç değişkeni sıralı bir yapıya sahip olan algılanan sağlık düzeyidir (ASD). Bu verinin analizi ile, bağımsız değişenlerin alt grupları, bölge ve aile düzeyinde ASD hakkında çıkarımlar yapılması amaçlanmaktadır. Bayesçi paradigma takip edilerek parametre ve rastgele etkilerin bileşik sonsal dağılımından örnekler elde edilmiştir. Model seçimi için üç kriter kullanılmıştır: Watanebe bilgi kriteri, log yalancı marjinal olabilirlik ve sapma bilgi kriteri. Üç kriter de, bölge ve aile düzeyindeki varyasyonların, algılanan sağlık düzeyinin modellenmesi için göz önünde bulundurulması gerektiğine işaret etmektedir. Modellerin, gözlenen veriye benzer verileri üretme yeterliliğini anlamak için sonsal kestirim kontrolleri yapılmıştır. Ekonomik ve demografik değişkenlerin seviyeleri, Türkiye'nin bölgeleri ve çalışmaya dahil edilen aileler arasında ASD açısından farklılıklar bulunmuştur. Örneğin, işsiz insanlar çalışan insanlara kıyasla %19 daha yüksek ihtimalle kötü sağlık durumu raporlarken, kırsal Ege kötü sağlık durumu raporlama konusunda en düşük olasılığa sahip bölgedir.

Keywords: *Bayesçi istatistik, kategorik veri analizi, gelir ve yaşam koşulları, gizli değişken modelleri, çok seviyeli analiz, algılanan sağlık düzeyi.*

1. Introduction

In this study, we consider the analysis of three-level ordinal outcome data. The data come from the Turkish Income and Living Conditions Surveys (TR-SILC) conducted by the Turkish Statistical Institute since 2006. In TR-SILC, the data are collected as panels of four years and cross-sectionally. Since regional information is only available in the cross-sectional data, in this study we consider the cross-section of one year; for three-level analysis of panel data, interested reader is referred to [1].

In the cross-sections of TR-SILC, data are collected on individuals that are nested within families. One would expect individuals from the same family to be more similar compared to individuals from other families, e.g. due to genetic factors, lifestyle, economic conditions, etc. The data is further nested within the statistical regions of Turkey. There are 12 statistical regions, defined according to the Nomenclature of Territorial Units for Statistics (NUTS) classification for Turkey, and in addition, we have the information about rural and urban areas. Thus, there are 24 regional units in total. It is expected that individuals from the same region are more similar than those from other regions.

The outcome of interest is self-rated health (SRH), which can take one of the following values: very poor, poor, fair, good, very good. A number of family and individual level explanatory variables are available. The main research interest of this study is to understand:

- the relationships between SRH and explanatory variables, and
- the region- and family-specific characteristics.

To address these, we consider a polytomous logistic regression model with random-effects. The presence of random-effects in a regression framework makes the interpretation of the regression coefficients, i.e. the first research interest, conditional on two persons from different covariate groups having the same random-effect. This is a restrictive assumption, as one would typically expect the random-effects associated with these two persons to be different. Following [1] and [2], and the references therein, we assume that the random-effects have a Bridge distribution for the logit link [3]. This assumption allows for an unconditional (or marginal) interpretation of the regression coefficients as in the classical regression setting (without random-effects). We take a Bayesian paradigm, and sample the parameters and random-effects from the joint posterior densities using Hamiltonian Monte Carlo (HMC, [4]).

We shall note that both the panel data analysis of [1] and the cross-sectional analysis of the current work are on three-level ordinal SRH outcome data and both works consider the same modelling strategy. The main differences between the two are as follows. In the panels, the repeats that are collected through time are nested within individuals, and the individuals are further nested within families. In the cross-sectional data, the repeats belong to different members of a single family, hence there is no time aspect, and the families are nested within regions. In [1], the main aim was to obtain interpretations of the regression coefficients, whereas in the current work we also consider interpreting the random-effects, as comparing the regions is one of the main research interests of this study.

The rest of the paper is organized as follows. In Section 2, we present the 2013 cross-section of TR-SILC. In Section 3, we present the modelling framework and the model selection criteria. Section 4 presents the results, while Section 5 the posterior predictive checks. Section 6 closes the paper with conclusion and discussion.

2. Data

The Turkish Income and Living Conditions Study (TR-SILC) surveys collect detailed information on income, poverty, social exclusion, living conditions, housing, labour, education and health. Turkey has been conducting the survey since 2006 as part of its integration into the EU, in the form of 4-year panels

and cross-sectional surveys. For the details of TR-SILC and SILC in general, the interested reader is referred to [1] and [5] and the references therein.

In this study, we consider a cross-section (specifically, the 2013 data) to examine, in particular, regional differences in health, as regional information is not available in the panels. The outcome variable is self-rated health (SRH) which is ordinal and can take one of the following values: very poor, poor, fair, good, very good. SRH represents the general health status of an individual and is considered as a predictor of morbidity and mortality [6]. Following [7] and [8], we consider a re-categorized version of the variable as good health (good/very good), fair health and poor health (poor/very poor). Mean household disposable income, defined as total annual family income in 2012 divided by family size (MHDI, in Turkish Lira), gender (male, female), marital status (married, never married, other), age (15 - 34, 35 - 64, 65+), education level (primary school or less, secondary or high school, higher education), working status (full/part time work, unemployed, student, housekeeping, other) are the explanatory variables. Note that, MHDI is a family-level variable, while the other variables are at individual-level.

The 2013 cross-section includes 53,496 individuals from 19,899 families. The SRH distribution with respect to regions is depicted in Figure 1. Urban Istanbul is the region with the lowest percentage of poor SRH, rural East Black Sea with the highest. Summary statistics for the explanatory variables can be found in Table 1, where we present the statistics both with respect to the levels of SRH and overall. In the analyses, the MDHI will be used in natural logarithm scale, because the variable is right-skewed. Since there are only 74 individuals from rural Istanbul, the data from rural and urban Istanbul are combined in the analyses (and simply referred to as Istanbul). There is no missing data in the variables considered.

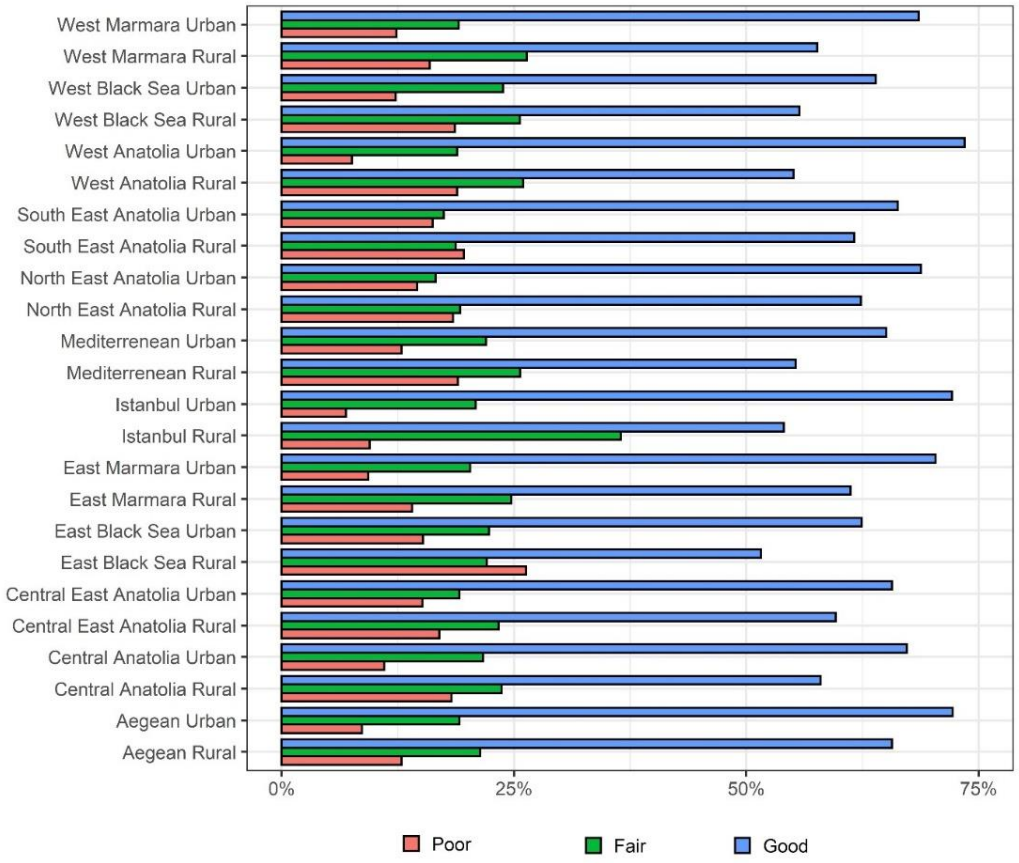

Figure 1. SRH distributions with respect to the statistical regions.

Table 1. Summary statistics for the 2013 cross-section of the TR-SILC data.

	Poor	Fair	Good	All
MHDI				
Minimum	375.7	44.2	6.3	6.3
25th percentile	4,125.0	4,991.0	5,186.4	4,969.2
Median	6,316.0	7,550.0	8,057.2	7,674.9
Mean	7,494.8	9,532.3	10,934.1	10,178.1
75th percentile	9,186.4	11,300.0	12,663.4	11,807.8
Maximum	178,842.3	210,667.3	373,924.6	373,924.6
Standard deviation	5,862.5	8,832.9	11,175.3	10,213.3
# of individuals	7,162	11,280	35,054	53,496
Family size				
Minimum	1.0	1.0	1.0	1.0
25th percentile	2.0	2.0	2.0	2.0
Median	3.0	3.0	3.0	3.0
Mean	3.2	3.1	3.4	3.3
75th percentile	4.0	4.0	4.0	4.0
Maximum	13.0	17.0	17.0	17.0
Standard deviation	1.6	1.5	1.6	1.6
# of individuals	7,162	11,280	35,054	53,496
Gender				
Female	4,377 (15.8%)	6,436 (23.3%)	16,820 (60.9%)	27,633 (51.7%)
Male	2,785 (10.8%)	4,844 (18.7%)	18,234 (70.5%)	25,863 (48.3%)
Marital Status				
Married	4,753 (13.2%)	8,721 (24.2%)	22,521 (62.6%)	35,995 (67.3%)
Never married	660 (5.2%)	883 (6.9%)	11,168 (87.9%)	12,711 (23.8%)
Other	1,749 (36.5%)	1,676 (35.0%)	1,365 (28.5%)	4,790 (9.0%)
Age				
15-34	844 (3.8%)	1,986 (9.0%)	19,342 (87.2%)	22,172 (41.4%)
35-64	3,602 (14.3%)	6,998 (27.7%)	14,641 (58.0%)	25,241 (47.2%)
65+	2,716 (44.6%)	2,296 (37.7%)	1,071 (17.6%)	6,083 (11.4%)
Education level				
Primary school or	6,235 (21.4%)	8,438 (29.0%)	14,406 (49.5%)	29,079 (54.4%)
Secondary or high	807 (4.3%)	2,207 (11.7%)	15,830 (84.0%)	18,844 (35.2%)
Higher education	120 (2.2%)	635 (11.4%)	4,818 (86.5%)	5,573 (10.4%)
Working status				
Full/part time	1,550 (6.3%)	4,558 (18.6%)	18,414 (75.1%)	24,522 (45.8%)
Unemployed	127 (6.2%)	305 (15.0%)	1,606 (78.8%)	2,038 (3.8%)
Housekeeper	1,945 (13.7%)	3,696 (26.1%)	8,534 (60.2%)	14,175 (26.5%)
Retired	961 (22.1%)	1,556 (35.8%)	1,835 (42.2%)	4,352 (8.1%)
Student	65 (1.5%)	183 (4.2%)	4,102 (94.3%)	4,350 (8.1%)
Other	2,514 (61.9%)	982 (24.2%)	563 (13.9%)	4,059 (7.6%)

3. Modelling framework

3.1. Notation and model

Let $Y_{ijk} \in \{1 = \text{good health}, 2 = \text{fair health}, 3 = \text{poor health}\}$ be the outcome belonging to individual $k\ (k = 1, \dots, n_{ij})$ from family $j\ (j = 1, \dots, m_i)$ and region $i\ (i = 1, \dots, s)$. Also let x_{ijk} a $p \times 1$ dimensional covariate matrix, where p is the number of coefficients.

The modelling framework that we consider to understand the relationships between SRH and the explanatory variables whilst taking into account the region- and family-level variabilities has the following form:

$$\text{logit}\{P(Y_{ijk} \leq a | x_{ijk}, U_i, V_{ij}, \theta)\} = \alpha_a^c - x_{ijk}^T \beta^c - U_i - V_{ij},\ a = 1,2, \quad (1)$$

where in addition to the notation introduced before, $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$, $P(\cdot)$ the probability operator, α_a^c category-specific threshold parameters, T transpose of a matrix, β^c regression coefficients, U_i and V_{ij} are random-effects, and θ a generic notation for parameters. In this setting, the interpretations of α_a^c and β^c are conditional on the U_i and V_{ij} terms being the same for two persons belonging to two different covariate groups; the super-script "c" stands for conditional interpretation. Assuming the random-effects having a Bridge distribution for the logit link allows us to directly obtain the unconditional/marginal interpretation, i.e. as in the usual regression setting. We call these parameters as the marginal parameters and denote by α_a^m and β^m.

3.2. Bridge distributed random-effects

One can obtain the relationships between α_a^m and α_a^c, and β^m and β^c by solving the following equation:

$$P(Y_{ijk} \leq a | x_{ijk}, \alpha_a^m, \beta^m) = E_{U,V}\left(P(Y_{ijk} \leq a | x_{ijk}, U_i, V_{ij}, \alpha_a^c, \beta^c, \theta_{U,V})\right), \quad (2)$$

where $E(\cdot)$ is the expectation operator, $\theta_{U,V}$ are the parameters of U_i, V_{ij}. The relationships would be available in closed-form, when one assumes Bridge distribution for the random-effects, as follows. Let $U_i = U_i^*/\emptyset_V$, where $[U_i^*]$ = Bridge($\emptyset_{U^*}$), and $[V_{ij}]$ = Bridge($\emptyset_V$), with $0 < \emptyset_{U^*}, \emptyset_V < 1$, and "$[\cdot]$" denotes "the distribution of". One can then obtain the marginal estimates as $\alpha_a^m = \emptyset_{U^*}\emptyset_V \alpha_a^c$ and $\beta^m = \emptyset_{U^*}\emptyset_V \beta^c$, see [1].

Under the above specification, note that U_i is no longer Bridge-distributed, but it has a Modified Bridge distribution. Properties of the Bridge and Modified Bridge distributions are presented below.

The probability density function of the Bridge distribution for logit link [3] is given by

$$f(x|\emptyset) = \frac{1}{2\pi} \frac{\sin(\emptyset\pi)}{\cosh(\emptyset x) + \cos(\emptyset\pi)}, -\infty < x < \infty, 0 < \emptyset < 1. \quad (3)$$

where $\cosh(\cdot)$ is the hyperbolic cosine, defined as $\cosh(x) = \frac{1}{2}(\exp(x) + \exp(-x))$. It is a symmetric distribution, has zero-mean and a variance of $\frac{\pi^2}{3}(\emptyset^{-2} - 1)$. The density function of the modified Bridge distribution, for generic X, Y and Z with $X = Y/\emptyset_Z$, $[Y|c] = \text{Bridge}(\emptyset_Y)$, $[Z|\emptyset_Z] = \text{Bridge}(\emptyset_Z)$, is given by

$$f(x|\emptyset_Y, \emptyset_Z) = \frac{\emptyset_Z}{2\pi} \frac{\sin(\emptyset_Y \pi)}{\cosh(\emptyset_Y \emptyset_Z x) + \cos(\emptyset_Y \pi)}, -\infty < x < \infty, 0 < \emptyset_Y, \emptyset_Z < 1, \quad (4)$$

Modified Bridge is also symmetric, zero-mean, and has a variance of $\frac{\pi^2}{3\emptyset_Y^2}\left({\emptyset_Y}^{-2} - 1\right)$.

3.3. Priors and inference

We select weakly informative prior distributions for the parameters following the literature. For $\alpha_a^c$and β^c, Cauchy distribution with location parameter 0 and scale parameter 5 is considered [9]. For Bridge distribution, the standard deviation, $\frac{\pi^2}{3}(\emptyset^{-2} - 1)$, is assumed to be half-Cauchy with location 0 and scale 5 [10, 11]. We sample the parameters and the random-effects from the joint posterior densities using the No-U-Turn Sampler [12], which is a modified version of Hamiltonian Monte Carlo [4]. Details of the posterior distributions are skipped here; for details one can consult the work of [1]. For computation, we use the R [13] package mixed3 (https://github.com/ozgurasarstat/mixed3).

3.4. Model selection

For model selection, we consider three widely used criteria that are used within the Bayesian framework. First of these is the Watanebe Information Criterion (WIC, [14]):

$$\text{WAIC} = -2(\text{lppd} - \rho), \quad (5)$$

where, "lppd" stands for log point-wise posterior density that is calculated as

$$\text{lppd} = \sum_{i=1}^{s}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} \log\left(\frac{1}{M}\sum_{l=1}^{M}[Y_{ijk}|U_i^{(l)}, V_{ij}^{(l)}, \theta^{(l)}]\right), \quad (6)$$

and ρ is the effective number of parameters and calculated as

$$\rho = \sum_{i=1}^{s}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} V_{l=1}^{M}\left(\log\left([Y_{ijk}|U_i^{(l)}, V_{ij}^{(l)}, \theta^{(l)}]\right)\right), \quad (7)$$

with

$$V_{l=1}^{M}(\text{a}) = \frac{1}{M}\sum_{l=1}^{M}\left(a^{(l)} - \bar{a}\right)^2. \quad (8)$$

In (6-8), the superscript (l) denotes the lth draw of the associated term from the joint posterior densities, M the size of the HMC sample. Note that lower values of WAIC indicate better model performance.

The second is the log pseudo marginal likelihood (LPML, [15, 16]). It is calculated as

$$\text{LPML} = \sum_{i=1}^{s}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}} \log\left(\widehat{CPO}_{ijk}\right), \quad (9)$$

where CPO stands for conditional predictive ordinate that is defined as leave-one-out cross-validated predictive density, $CPO_{ijk} = \left[Y_{ijk}|Y_{-(ijk)}\right]$, where $Y_{-(ijk)}$ denotes the full set of outcomes without the observation (ijk). The estimate of CPO that we use is the harmonic mean estimate [15],

$$\widehat{CPO}_{ijk} = \left(\frac{1}{M}\sum_{l=1}^{M}\frac{1}{\left[Y_{ijk}|U_i^{(l)},V_{ij}^{(l)},\theta^{(l)}\right]}\right)^{-1}. \quad (10)$$

Larger values of LPML indicate better model fit.

The third criterion is the deviance information criterion (DIC, [17]) for which the formula is given by

$$DIC = 2\bar{D} - D(\bar{\theta},\bar{U},\bar{V}), \quad (11)$$

where

$$\bar{D} = \frac{1}{M}\sum_{l=1}^{M} -2\log\left(\left[Y_{ijk}|U_i^{(l)},V_{ij}^{(l)},\theta^{(l)}\right]\right), \quad (13)$$

$$D(\bar{\theta},\bar{U},\bar{V}) = -2log\left(\sum_{i=1}^{s}\sum_{j=1}^{m_i}\sum_{k=1}^{n_{ij}}\left[Y_{ijk}|\bar{\theta},\bar{U}_i,\bar{V}_{ij}\right]\right), \quad (14)$$

and $\bar{\theta} = \frac{1}{M}\sum_{l=1}^{M}\theta^{(l)}$, $\bar{U}_i = \frac{1}{M}\sum_{l=1}^{M}U_i^{(l)}$, $\bar{V}_{ij} = \frac{1}{M}\sum_{l=1}^{M}V_{ij}^{(l)}$. Lower values of DIC indicate better fit.

4. Results

We fit the following three models to the 2013 cross-section of the TR-SILC:

- fixed-effects: no U_i and V_{ij} terms in model (1),
- two-level: no U_i term in model (1),
- three-level: the model described in (1).

For each model, we run 4 parallel HMC chains started from random initials. Each chain has the length of 2,000, first halves of which are discarded as the burn-in. In total, the HMC chains have the size of 4,000. To assess the convergence of the chains, we use trace-plots, density plots, and the R-hat statistic [18]. Trace-plots indicate that the 4 chains for each parameter converge to the same target and mix well, density plots indicate the chains have similar distributions, and all the R-hat statistics were close to 1. These collectively indicate convergence of the HMC chains. It took about 1.8, 8.6, and 8.8 hours to fit the fixed-effects, two-level and three-level models, respectively, on a 64-bit personal laptop with 12 GB RAM and Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz running Windows 10. Means, standard deviations (sd) and 2.5%th and 97.5%th percentiles of the HMC samples are presented in Table 2. For the two- and three-level models, we directly present the α_a^m and β^m, as α_a^c and β^c are not of primary interest. The model selection criteria are presented in Table 3. All of the LPML, WAIC and DIC indicate that the three-level model is the best fitting model, whereas the fixed-effects model is the worst. This indicates that both the regional- and family-level dependencies need to be taken into account in order to appropriately analyze the TR-SILC data.

Since the three-level model is found to be the best fitting model, here we only interpret the related coefficients. One percent decrease in MDHI was associated with approximately $0.3\%\ (= (\exp(0.293 \times \log(1.01)) - 1) \times 100)$ increase in the odds of reporting poorer health. Females were approximately $28\%\ (= (\exp(0.244) - 1) \times 100$ more likely to report poorer health compared to males. People who never married were approximately 33% less likely to report poorer health compared to people who were married, whereas people whose marital status was different than married/never married were approximately 30% more likely to report poorer health compared to those who were married. People whose age was in the $35-64$ and $65+$ categories were 2.3 and 5.7 times more likely to report poorer health compared to those who were in the $15-34$ category, respectively. As the education level increased the probability of reporting poorer health decreased. Students were less likely to report poorer

health compared to employed people. People in all the other working status categories were more likely to report poorer health. For example, unemployed people were 19% more likely to report poorer heath compared to those were employed. Means, and 2.5%th and 97.5%th percentiles of the HMC samples of the U_i terms are displayed in Figure 2. Rural and urban Aegean regions are the ones with the lowest chance of reporting poorer health. Urban and rural East and West Marmara regions, Istanbul and urban West Anatolia are also amongst the lowest risk regions. Rural and urban East Black Sea regions are the ones that had the highest chance of reporting poorer health. Both urban and rural Central East Anatolia are also amongst the regions that had the highest chance. Means, 2.5%th and 97.5%th percentiles of the HMC samples of the V_{ij} terms for randomly selected 50 families are displayed in Figure 3. Two- and three-level models largely agree on the mean estimates, whereas we see minor differences in the 95% credibility intervals.

Table 2. Estimation results. "sd" stands for standard deviation.

		Fixed-effects			Two-level			Three-level		
Variable	**Parameter**	**Mean**	**sd**	**2.5%, 97.5%**	**Mean**	**sd**	**2.5%, 97.5%**	**Mean**	**sd**	**2.5%, 97.5%**
Threshold	α_1^m	-0.557	0.169	-0.894, -0.232	-0.835	0.186	-1.198, -0.470	-0.229	0.194	-0.607, 0.162
Threshold	α_2^m	1.126	0.168	0.793, 1.444	0.821	0.186	0.460, 1.186	1.371	0.192	0.997, 1.757
log(MHDI)	β_1^m	-0.351	0.017	-0.385, -0.319	-0.377	0.019	-0.413, -0.340	-0.293	0.020	-0.333, -0.253
Male (Ref)	-	-	-	-	-	-	-	-	-	-
Female	β_2^m	0.250	0.027	0.196, 0.303	0.247	0.025	0.200, 0.297	0.244	0.024	0.196, 0.291
Married (Ref)	-	-	-	-	-	-	-	-	-	-
Never married	β_3^m	-0.288	0.039	-0.366, -0.212	-0.297	0.040	-0.374, -0.220	-0.282	0.039	-0.358, -0.205
Other	β_4^m	0.255	0.035	0.188, 0.326	0.265	0.034	0.198, 0.333	0.263	0.033	0.198, 0.329
15-34 (Ref)	-	-	-	-	-	-	-	-	-	-
35-64	β_5^m	1.194	0.030	1.137, 1.253	1.218	0.030	1.158, 1.277	1.186	0.032	1.122, 1.251
65+	β_6^m	1.963	0.042	1.881, 2.048	1.960	0.042	1.874, 2.041	1.906	0.049	1.810, 2.002
Higher education (Ref)	-	-	-	-	-	-	-	-	-	-
Primary or less	β_7^m	0.925	0.045	0.834, 1.015	0.861	0.046	0.772, 0.952	0.843	0.046	0.751, 0.931
Secondary or high school	β_8^m	0.292	0.046	0.202, 0.384	0.283	0.046	0.190, 0.373	0.286	0.045	0.198, 0.377
Full/part time (Ref)	-	-	-	-	-	-	-	-	-	-
Housekeeper	β_9^m	0.278	0.030	0.220, 0.336	0.280	0.029	0.224, 0.337	0.264	0.028	0.207, 0.321
Other	β_{10}^m	2.131	0.043	2.047, 2.212	2.108	0.042	2.027, 2.192	2.020	0.049	1.921, 2.114
Retired	β_{11}^m	0.785	0.035	0.716, 0.853	0.730	0.035	0.662, 0.799	0.724	0.034	0.658, 0.791
Student	β_{12}^m	-0.458	0.076	-0.608, -0.311	-0.294	0.070	-0.435, -0.160	-0.284	0.066	-0.411, -0.155
Unemployed	β_{13}^m	0.202	0.062	0.082, 0.320	0.177	0.057	0.062, 0.288	0.173	0.058	0.060, 0.284
U^*	$\emptyset_{U^*}$	-	-	-	0.816	0.006	0.805, 0.827	0.959	0.013	0.930, 0.979
V	$\emptyset_V$	-	-	-	-	-	-	0.821	0.006	0.810, 0.832

Table 3. Model selection results

Model	LPML ↑	WAIC ↓	DIC ↓
Fixed-effects	-37,267.8	74,535.5	74,535.
Two-level	-35,919.6	71,315.8	71,781.
Three-level	-35,830.2	71,158.0	71,603.

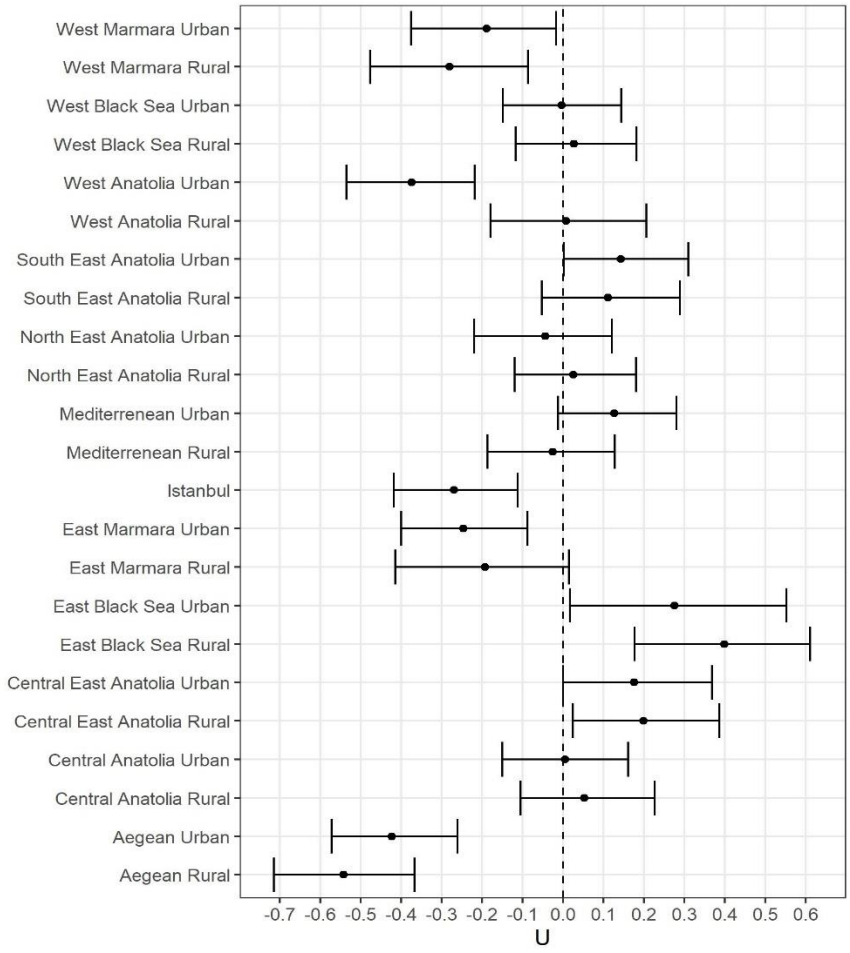

Figure 2. Means (in black dots) and 2.5%th and 97.5%th percetiles (as error bars) of the posterior distributions of the U terms based on the three-level model.

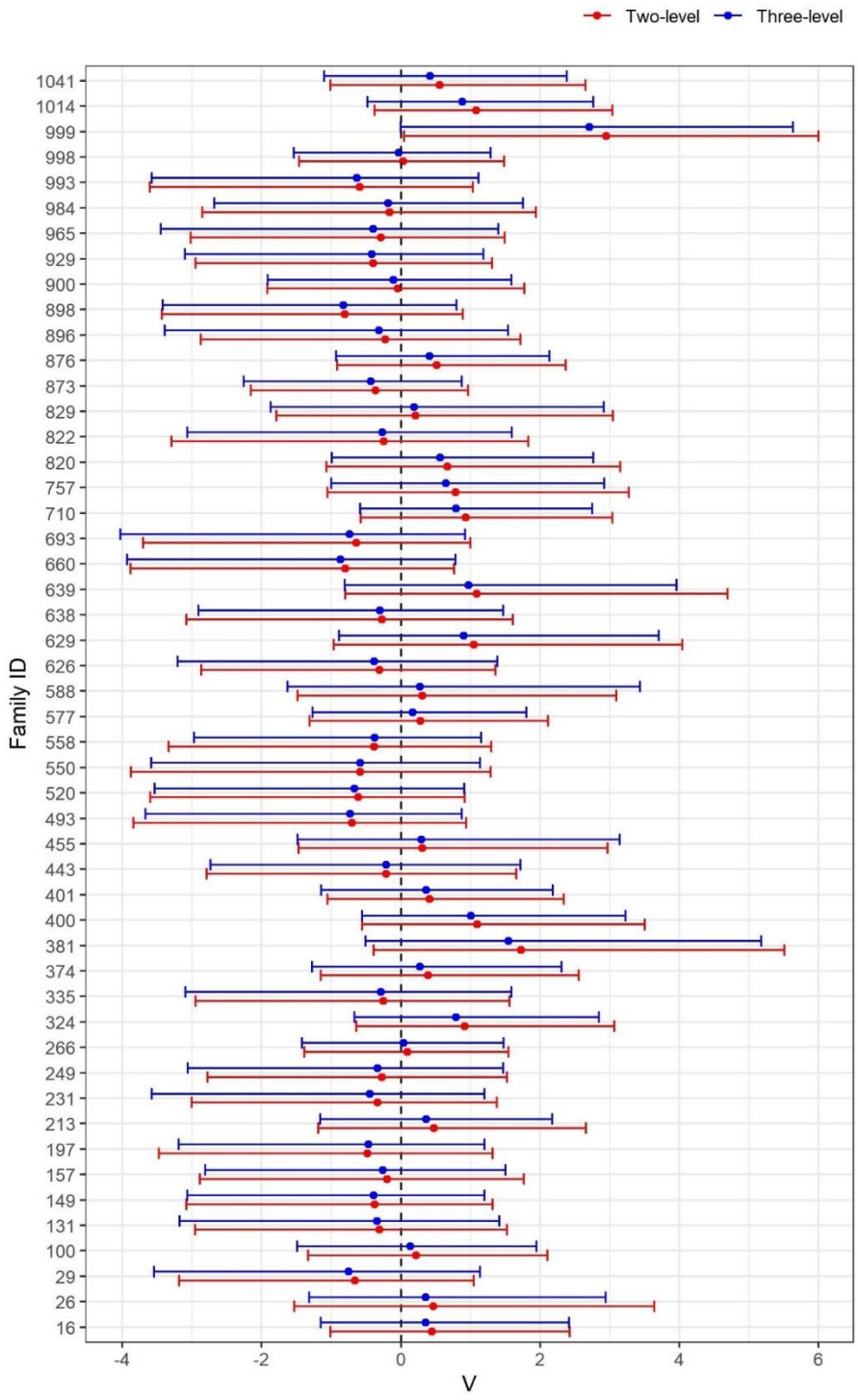

Figure 3. Means (in dots) and 2.5%th and 97.5%th percentiles of the HMC samples of the family-level random-effects (V_{ij}) for randomly selected 50 families, based on the two-level model (in red) and three level model (in blue).

5. Posterior predictive checks

In order to see how the fitted models replicate the SRH outcome data, we performed posterior predictive checks. We simulated data for each of the 4,000 elements of the HMC samples from

$$[Y_{ijk}^{sim}|Y] = \int\int\int [Y_{ijk}^{sim}|U_i, V_{ij}, \theta][U_i|Y,\theta][V_{ij}|Y,\theta]\,[\theta|Y] dU\, dV\, d\theta \quad (14)$$

for the three-level model, from

$$[Y_{ijk}^{sim}|Y] = \int\int [Y_{ijk}^{sim}|V_{ij}, \theta][V_{ij}|Y,\theta]\,[\theta|Y]\, dV\, d\theta \quad (15)$$

for the two-level model, and from

$$[Y_{ijk}^{sim}|Y] = \int [Y_{ijk}^{sim}|\theta]\,[\theta|Y]\, d\theta \quad (16)$$

for the fixed-effects model, Y indicates the set of observed SRH outcomes. We then compared the simulated data-sets with the observed SRH outcomes. We report means, standard deviations and 2.5%th and 97.5%th percentiles for the percentages of matches and mis-matches between the observed and simulated SRH outcomes, see Table 4. Here, matches and mis-matches are defined as

- ``-2": observed outcome being ``good health" and simulated being ``poor health";
- ``-1": observed being ``good health" and simulated being ``fair health", or observed being ``fair health" and replicated being ``poor health";
- ``0": observed and simulated being the same;
- ``1": observed being ``fair health" and simulated being ``good health", or observed being ``poor health" and simulated being ``fair health";
- ``2": observed being ``poor health" and replicated being ``good health".

Note that non-zero values mean mis-match, whereas ``-2" and ``2" would mean the most mis-match. Two- and three-level models seem to perform similarly in terms of replicating the observed data, whereas fixed-effects model seems to be the worst.

	Fixed-effects			**Two-level**			**Three-level**		
Diff	**Mean**	**sd**	**2.5%, 97.5%**	**Mean**	**sd**	**2.5%, 97.5%**	**Mean**	**sd**	**2.5%, 97.5%**
-2	8.63	0.13	8.37, 8.89	6.54	0.12	6.32, 6.78	7.10	0.26	6.59, 7.60
-1	16.69	0.18	16.34, 17.03	14.46	0.17	14.12, 14.81	15.44	0.37	14.69, 16.13
0	49.30	0.19	48.93, 49.66	51.95	0.17	51.61, 52.29	51.04	0.39	50.30, 51.82
1	16.58	0.10	16.38, 16.78	17.49	0.09	17.31, 17.66	17.14	0.14	16.86, 17.42
2	8.80	0.07	8.67, 8.93	9.56	0.06	9.43, 9.68	9.29	0.12	9.06, 9.51

Table 4. Posterior predictive check results. ``Diff" stands for difference, ``sd" for standard deviation.

6. Conclusion and discussion

In this study, we analyzed the 2013 cross-section of the TR-SILC study. The outcome variable is the SRH which has three categories: poor, fair and good health. A number of economic and demographic variables are considered to explain the variability in SRH. The data has two sources of dependency: statistical regions and families. We considered a polytomous logistic regression with Bridge distributed random-effects. The Bridge distribution specifically allows us to obtain marginal interpretations of the regression coefficients, while making inferences at the region- and family-level. Inferences for parameters and

random-effects are obtained under the Bayesian paradigm. The methods are implemented in the R package mixed3.

We found differences between covariate subgroups with respect to SRH. People with higher income and education were less likely to report poorer health overall. Gender, marital status, and age also appear to explain variability in SRH. People who have never been married appear less likely to have poorer health. Similarly, students seem to be less likely to report poorer health compared to those who are employed. We shall note that both of these results can be explained by the age factor.

It is interesting to observe differences between regions in terms of reporting poorer health. The Aegean and Marmara regions have the lowest probability of reporting poorer health, while East Black Sea and Central East Anatolia have the highest probability of reporting poorer health. It is also interesting to observe differences between the families through the random-effects, which can be considered as proxies for unmeasured characteristics, e.g. genetic factors. Besides these observations, the model selection criteria we considered suggest that both regional- and family-level dependencies need to be taken into account when analyzing the TR-SILC data.

This paper is the first to consider appropriate statistical modelling for the analysis of cross-sections of TR-SILC, where we analyzed data from the 2013 cross-section. Other cross-sections can also be analyzed and the results are compared. Causal inference can be considered to draw causal interpretations, as the TR-SILC data is observational. These are the beyond the scope of this work.

Acknowledgements

The author acknowledges the encouragements of and helpful discussions with Dr. Kutsev Bengisu Özyörük while writing this paper. The author also thanks to Dr. Mahmut Yardım for introducing the SILC studies and helpful discussions.

References

[1] Ö. Asar, 2021, Bayesian analysis of Turkish Income and Living Conditions data, using clustered longitudinal ordinal modelling with Bridge distributed random-effect, *Statistical Modelling*, 21(5), 405-427.
[2] L. Boehm, B. J. Reich, Bandyopadhyay, 2013, Bridging conditional and marginal inference for spatially referenced binary data, *Biometrics*, 69, 545-554.
[3] Z. Wang, T. A. Louis, 2003, Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function, *Biometrika*, 90(4), 765-775.
[4] R. Neal, 2011, MCMC using Hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo*, Brooks, S., Gelman, A., Jones, G. L., Meng, X. L. (eds.), Chapman & Hall/CRC Press, Boca Raton. s. 113-162.
[5] V. S. Arora, M. Karaniokolos, A. Clair, A. Reeves, D. Stuckler, M. McKee, 2015, Data resource profile: the European Union Statistics on Income and Living Conditions (EU-SILC), *International Journal of Epidemiology*, 44(2), 451-461.
[6] B. Burstörm, P. Fredlund, 2001, Self-rated health: is it as good a predictor of subsequent mortality among adults in lower as well as in higher social classes?, *Journal of Epidemiology and Community Health*, 55, 836-840.
[7] D. S. Abebe, A. G. Toge, E. Dahl, 2016, Individual-level changes in self-rated health before and during the economic crisis in Europe, *International Journal for Equity in Health*, 15(1), 1-8.
[8] M. S. Yardım, S. Üner, 2018, Equity in access to care in the era of health system reforms in Turkey, *Health Policy*, 122(6), 645-651.
[9] A. Gelman, A. Jakulin, M. G. Pittau, Y.-S. Su, 2008, A weakly informative default prior distribution for logistic and other regression models, *The Annals of Applied Statistics*, 2(4), 1360-1383.
[10] A. Gelman, 2006, Prior distributions for variance parameters in hierarchical models, *Bayesian Analysis*, 1(3), 515-534.
[11] N. G. Polson, J. G. Scott, 2012, On the half-Cauchy prior for a global scale parameter, *Bayesian Analysis*, 7(4), 887-902.

[12] M. D. Hoffman, A. Gelman, 2014, The No-U-Turn sample: adaptively setting path lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research*, 15, 1593-1623.
[13] R Core Team, 2021, R: a language for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.
[14] S. Watanebe, 2010, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *Journal of Machine Learning Research*, 11, 3571-3594.
[15] D. K. Dey, M. H. Chen, H. Chang, 1997, Bayesian approach for nonlinear random effects models, *Biometrics*, 53, 1239-1252.
[16] A. Gelman, J. Hwang, A. Vehtari, 2014, Understanding predictive information criteria for Bayesian models, *Statistics and Computing*, 24, 997-1016.
[17] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, A. van der Linde, 2002, Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society, Series B*, 64(4), 583-639.
[18] S. P. Brooks, A. Gelman, 1997, General methods for monitoring convergence of iterative simulations, *Journal of Computational and Graphical Statistics*, 7, 434-455.