Modeling Obesity Prevalence

A Stochastic Approach for the U.S. Population

A Thesis Submitted to the Faculty of Miami University in partial fulfillment of the requirements for the Master of Science degree

Department of Statistics

By

Palma Daawin

Miami University
Oxford Ohio

2017

Advisors: Tatjana Miljkovic, Seonjim Kim

Abstract

In recent years, obesity has become an inevitable health issue of interest in the United States due to its consequent effect of being a catalyst for many diseases such as arthritis and cancer that cause death. The topic of obesity has gained recognition as a premier public health issue in the U.S. The Center for Disease Control (CDC) reported that more than one-third of U.S. adults are obese. We use the cross-sectional obesity prevalence data and account for a source variation for the combined time and age effects, usually known as the cohort effect. The measurement and inclusion of cohort effect in various models have gained much interest recently in academic literature. In this study, we take advantage of the knowledge and academic tools in stochastic modeling to fit the curvilinear relationship of obesity prevalence and age for the cross-sectional observed obesity prevalence. We analyze the obesity trend based on 25 years of data for ages 23 to 90 provided by CDC's Behavioral Risk Factor Surveillance System (BRFSS) survey and fit proposed models. We assume a polynomial regression and make initial estimates using Ordinary Least Squares (OLS) regression; then we adjust for the cohort effect using an iterative algorithm to fit the quadratic structure of obesity prevalence. Finally, we also make use of time series and forecasting techniques to forecast obesity prevalence.

ACKNOWLEDGMENTS

I am forever grateful to the following people as this research could not have been completed without their immense support and contributions.

Committee members and my advisors: Dr. Tatjana Miljkovic, Dr Seonjin Kim and Dr Helaine Alessio, for their valuable guidance and feedback.

Faculty and members of the Department of Statistics, Miami University for their ideas, contributions, and thoughts which I found very useful.

Most of all my parents, family, friends, close associates and well-wishers for their prayers, encouragement, and support while I was pursuing the MS degree.

DEDICATION

Dedicated to my parents, family, friends and well-wishers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. Introduction

Researchers in medical sociology and epidemiology are often interested in studying the distribution and etiology associated with a large variety of health issues. Estimating and predicting obesity prevalence for the U.S. population has been an important concern for researchers working in this area. Alison at al. (1999) estimated that there are about 300,000 deaths in the U.S. attributed to obesity each year. Olshansky et al. (2005) forecasted a decline in life expectancy in the U.S. in the 21st century due to the effect of obesity on longevity. Many follow-up studies in medical field have focused on the same issue and concluded with similar findings. According to medical studies from the Center for Disease Control, obesity is a leading catalyst of inherent health conditions. Obese individuals are vulnerable to death –causing diseases such as coronary heart disease, high blood pressure, stroke, type 2 diabetes cancer, sleep apnea, gallstones and osteoarthritis. (Weight, 2015). Olshansky et al. (2005) argued that the steady rise in longevity observed in modern era may soon come to an end because the younger population today on an average live less healthy and possibly shorter lives than their parents. As a result, the U.S. is expected to face a decline in life expectancy and changes in the mortality during the 21st century.

Significant financial and pension planning consequences could arise due to the changes in mortality rates. This is because mortality and life expectancy affects the financial projections and sustainability of the U.S social security system. Insurance and investment firms use mortality rates to forecast liabilities and amounts required to be booked in periodic financial reporting. Recently, many life and health insurance firms have begun to include obesity as a major risk factor in their ratemaking and life insurance product pricing. It has become very common for the prospective insured to provide measurements of their height and weight during the underwriting

phase of the life insurance sales cycle. This information about an individual's obesity is included in the pricing of premiums payable for various life insurance and annuity products.

In the past two decades, various efforts have been made to estimate the prevalence of obesity for the U.S. population. Past methods have mainly revolved around the use of explanatory Analysis of Variance (ANOVA) models. (Keyes et al.,2010). The apparent focus of these methods is to explain the variation in obesity rates across different ages and years as well as different cohorts, no matter how many interest groups are interested in predicting obese prevalence. In the field of mortality estimation, several efficient models have been developed using stochastic methods. (Villegas et al., 2015). Since mortality and obesity prevalence have similar cross-sectional characteristics, which is vary across age and time period, we would take advantage of our knowledge in mortality estimation and apply them to approximate the prevalence of obesity. We can further make adjustments to cater for the cohort effect which we identify as a source of variation.

The intent of our study is to propose a model that not only explains the variation in obesity rates across different ages, years and cohorts, but also has predictive capability of obesity prevalence. In this study we would outline the various existing methods in academic literature and propose a model that fits the obesity prevalence data including the cohort effect.

Chapter 2. Literature Review

2.1 Obesity

The prevalence of obesity in the U.S in the past decades has become one of the top public health issues. Body-mass-index (BMI) is calculated as weight in kilograms divided by height in meters squared, and it serves as a measure to categorize people into the following categories: underweight (BMI < 18:5), normal weight (18:5 < BMI < 24:9), overweight (25 < BMI< 29:9), or obese (BMI > 30). A great number of medical studies have examined the issue of obesity and have recognized it as a risk factor to adult mortality. Based on the recent trends in BMI, with a larger proportion of people becoming overweight and obese, the distribution of BMI has shifted in means toward higher BMI levels with an increase in skewness and multimodality suggesting that mixture modeling of BMI may be more appropriate (Miljkovic et al., 2016). The diseases related to obesity include heart diseases, type-2 diabetes, and some types of cancers as discussed by Must et al. (1999) and Ebbeling et al. (2002). Fontaine et al. (2003) estimated that the expected number of years of life lost due to overweight and obesity for age 20-30 years is 13 for white men and 8 for white women based on the data from the National Health and Nutrition Examination Survey (NHANES; 1976-1994). In this study obesity rates are computed using data from the Behavioral Risk Factor Surveillance System (BRFSS) survey. In this study, observed obesity rates $p(t{:}\, x)$, were computed as

$$p(t{:}\, x) = \frac{Number\ of\ individuals\ aged\ x\ in\ year\ t\ \ with\ BMI > 30}{Total\ number\ of\ individuals\ aged\ x\ in\ year\ t}$$

2.2 The Cohort Effect

Ryder (1965) concluded that we impede on temporal analysis when the cohort effect is ignored and omitted as a source of variation in any form of Age-Period scalar analysis. The cohort effect is usually composed of the Age and Period effects. In order to appropriately define cohort effect, one needs to define the "age effects" and "period effects" (Keyes et al., 2010)

According to Keyes et al. (2010), age effects are the peculiar characteristics that are common and widely associated with individuals of a specific age group in their journey of life. Thus, the aging process comes with certain physiological alterations and cumulative vulnerabilities that have been medically determined to happen with certain age groups. On the contrary, period effects are usually concerned with the environment. Period effects are the peculiar widely known environmental experiences that exist at a certain period in time. Various fields have their way of defining the "cohort effect".

In epidemiology, cohort effect is often associated with events that have a general environmental occurrence or health epidemic. Thus cohort effect in epidemiology is defined as the obvious experience of an age specific exposure that is caused by a general occurrence.

In sociology, Ryder (1965) proposed a view about cohort effect that looks at this effect from the dimension of an exposure that can be used to explain the lives of members of the cohort over their course of life. It is an occurrence that is peculiar and associated with a specific age group that affects their generation into the future and is generally evident in the members of that cohort. Thus, it is a unique occurrence that achieves a structural effect in the way and life expectancy or characteristic of members of that group. Sociologists usually associate cohort effect with the full environmental influence of a specific group of individuals born in the same period.

Demography measures follow events such as birth, death, marriage, and migration and are all observed in the dimensions of age, period, and cohort. Palmore (1978) exemplifies the demographic view by illustrating cross-sectional, longitudinal, and time-lag dimensions of demographic events. In his paper, Palmore (1978) defines age and period as a longitudinal difference, age and cohort as a cross-sectional difference, and finally period and cohort as the time-lag difference. He further argues that age effects are as a result of biological, psychological and social changes in role. He describes period effects them as changes in the environment, measurement and practice and cohort effects as genetic shifts and interaction of historical situation with age of cohort.

A cohort is defined as "A group of people with a shared characteristic" (Oxford Advanced Learner's Dictionary, 2016). The cohort effect is also known as "generation effects" (Last, 2001). Cohort analysis refers to the methodology used to capture and quantify the specific at-risk birth cohorts.

In cohort analysis, the same individuals are studied over time. However, it is difficult to measure cohort effect not necessarily for the same individuals but for individuals who are the same age. Samples taken each year and grouped by ages and studied over time. Ideally, if the same individuals could be followed over time, the obesity effect would be measured more accurately and inadvertently lead to a measuring of the cohort effect since the same individuals would be followed over time. However, this is usually not the case. Obesity prevalence is estimated through random periodic surveys of samples of individuals from the population who are usually different from one period to another.

In the concept of mortality, inclusion of cohort and cohort estimation using stochastic methods has developed tremendously with many great ideas to estimating the cohort effect. Using a stochastic approach in estimate the cohort effect in obesity prevalence is a fairly new concept.

2.3 Background to the Development of Cohort Effect

Cohort analysis was first introduced as a concept in mortality in order to adjust mortality rates with the obvious diagonal effect of the cross-sectional age –period structure. This adjustment was done mainly for the purpose of predicting life expectancy (Tutt, 1953). Indeed, it is rare for academic researchers to discuss cohort effect without appreciating and dichotomizing the mortality origin of the cohort effect.

Prior to 1992, several methods were developed by various academic researchers for forecasting mortality. These approaches mainly focused on the use of stochastic methods. Stochastic models comprise random variables and probabilistic assumptions for terms of response of interest such as death and survival as well as assumptions of independence. (Olivieri and Pitacco, 2006). In 1992, Lee and Carter developed and published a new method for forecasting of trends and age patterns in mortality. Many adjustments have been made to the Lee and Carter (1992) model. One of the extensions, the Renshaw and Haberman (2006) model. This model generalized the Lee-Carter model to include a cohort effect. The model is characterized and formulated as

$$log\ (m(t:x)) = \beta_x^{(1)} + \beta_x^{(2)} k_t^{(2)} + \beta_x^{(3)} \gamma_{t-x}^{(3)} \qquad (1)$$

Here $m(t:x)$ is the death rate, $\beta_x^{(1)}\ and\ \beta_x^{(2)}$ are age predictors with a coefficient $k_t^{(2)}$ and $\gamma_{t-x}^{(3)}$ is the cohort predictor with its associated coefficient $\beta_x^{(3)}$.

Cairns, Blake & Dowd (2006b) successfully fit another simple model to mortality rates,

$$logit\big(q(t{:}\,x)\big) = k_t^{(1)} + k_t^{(2)}(x - \overline{\mathrm{x}}) \qquad (2)$$

Here $q(t{:}\,x)$ is the mortality rate, $k_t^{(1)}$ and $k_t^{(2)}$ are age coefficients associated with each period t. This model is popularly known as the CBD model, which is an abbreviation corresponding to the names of the authors. A generalization of model CBD model that adds a cohort effect and quadratic term into the age effect is popularly known as the M7 model. According to Carins et al. (2007), the inclusion of the quadratic term with the coefficient $k_t^{(3)}$ is inspired by some curvature identified in the logit $q(t;\, x)$ plots in the US data. The M7 model is characterized by the following link function;

$$logit(q(t{:}\,x)) = k_t^{(1)} + k_t^{(2)}(x - \overline{x}) + k_t^{(3)}((x - \overline{x})^2 - \sigma_x^2) + \gamma_{t-x}^{(4)} \qquad (3)$$

It is worthy to note that several extensions have been provided to their basic mortality model proposed by Carins et al., 2009. These models are known in academic literature as the Generalized Age-Period-Cohort (GAPC) stochastic mortality models. According to McCullagh and Nedler (1989), GAPC's have the following components;

i) Random component

Since death is count data, either a Poisson Distribution or Binomial Distribution is assumed such that the number of deaths $D(t;x) \sim Poisson(E(t;x)u(t;x))$ or $D(t;x) \sim Binomial(E(t;x)q(t;x))$, here $E(t;x)$ is number of life exposures, $u(t;x)$ is the force of mortality and $q(t;x)$ is the mortality rate for age x in year t .

ii) Systematic component

Hunt and Blake (2015) concluded that the effects of age α_x, period ,k_t^i and cohort γ_{t-x} could be fully expressed as a link predictor η_{xt} defined as

$$\eta_{xt} = \alpha_x + \sum_{i=1}^{N} \beta_x^i k_t^i + \beta_x^0 \gamma_{t-x} \qquad (4)$$

where β_x^i is usually a predictor defined as a function of age x.

iii) Link function

These associates the random and systematic components to each other. Like many GLMs, several possibilities exist with respect to the canonical link function that can be assumed. Usually the link function is chosen as either a "log" for a poison distribution or "logit" link for a binomial distribution.

iv) Parameter constraints

Due to the possible dependent structure and relationship between the various age, period and cohort variables, many of these mortality models have an identifiability problem hence making them inestimable via usual OLS regression. A set of parameter constraints have to be imposed to correct for the behavior of dependent relationship structure .An example of these parameter constraints is provided in our model in chapter 4 .

As can be seen from the above literature, there are multiple ways to estimate obesity prevalence using several different approaches. Firstly we can proceed to model the actual count of obese individuals across various age groups and periods. Here, a poisson or binomial can be assumed and a generalized linear models approach is pursued. Alternatively, we can model the actual obesity rates attributable to specific ages and periods. This can be done with OLS regression with appropriate constraints applied.

2.4 Statistical Methods for Modeling Cohort Effects in Obesity Prevalence

According to Keyes et al. (2010), over the years various efforts to model cohort effect could be categorized into three main approaches, the Age-Period-Cohort approach, the First order effects approach and finally the Second order effects approach. The Age-Period-Cohort analysis usually separates the variance structure into constituents which are attributable to age, period and cohort effects. The three components are usually defined as having some form of linear relation with obesity rates. Keyes et al. (2010) described these linear relationships as the "first-order effects".

Because of the apparent relationship of Cohort (Cohort = Period-Age) to Age and Period, a dependence structure is formed leading to collinearity and its associated problems. Applying usual regression or statistical methods without appropriate restrictions leads to a rank deficient design matrix which cannot be evaluated for unique coefficient. The identifiability problem is created when there are many solutions to the same equation, thus no unique solution exists.

A large volume of academic literature exists to address this identifiability problem. Recent literature in this area include Glenn, 2005 and Yang, Schlulhofer-Wohl,Fu and Land ,2008.

Many models have attempted to quantify the Age-Period-Cohort effect in obesity prevalence estimation. One of such models is the Constraint-Based model. The Constraint-Based Approach, suggested by Keyes et al. (2010), is a part of models that focuses on the first order effects. This model is basically an ANOVA with three-factors. It is characterized by the following equation

$$In(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + \varepsilon_{ijk} \qquad (5)$$

$Y_{ijk}, \mu, \alpha_i, \beta_j, \gamma_k \; and \; \varepsilon_{ijk}$ are the obesity rate , constant term , m-1 different age effects , n-1 different period effects , m+n-2 cohort effects and error term respectively . Here m is the number of different age groups and n is the number of different years in the data .The cohort predictor is obtained as (Cohort =Period-Age). This model has an identifiability problem because there exists some multicolinearity between age, period and cohort. In order to solve this, a number of constraints are applied so that the ends age groups have 0 effects. Thus, the constraint chosen could affect values of output from the model.

An alternative approach is the Holford approach, which is basically built on an underlying first order effects but has an outcome which is likened to a constraint invariant second order effects results. Holford modeled second order effects or curvature with a model using a 3 factor age-period –cohort linear contrasts. Holford (1983, 1991, and 1992) attempted to solve this identifiability issue with constraint Age-Period Cohort (APC) models by suggesting a new approach that dwells on second order effects with a focus on contrasts that are linear. Second order effects refer to relationships that are non-linear with the dependent variable of interest. The linear contrast attempt to estimate the change in direction of an underlying linear gradient.

The third approach is to fully focus on second order effects only. This approach basically does not model any first order effects. A good example of this is the Median Polish method which was first suggested by Tukey, 1977. This method is a two-factor model with no constraints and characterized by the following expression.

$$In(Y_{ij}) = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \qquad (6)$$

$Y_{ij}, \mu, \alpha_i, \beta_j and\ \varepsilon_{ij}$ are the outcome, constant term, m-1 age effects, n-1 period effects and error term respectively . No cohort term is represented since this approach does not assume an additive effect of the cohort term. This approach assumes that by combining age and period in the model, an inadvertent cohort effect is created.

Selvin (1996) suggested the use of this method for Age-Period-Cohort analysis. The Median Polish method is heavily reliant on a contingency table (see Table 1) having the various n age groups and m periods in rows and columns. This method basically deducts the median of each row and column iteratively. The residuals of the median are then regressed on dichotomous indicator variables in order to establish their classification to a cohort period using ordinary least squares regression. The cohort effect is then estimated as the degree of accuracy to which the computed variable correctly classifies the residual. (Keyes et al., 2010)

In mortality analysis, other researchers have suggested estimating the cohort effect through a multiplicative interaction of period and age. (see Keyes and Li, 2010). Our proposed model uses ages and the squared of ages shifted by the mean of all ages considered, as a predictor of the probability of obesity at a given age. We then use an iterative algorithm to continuously adjust parameters and estimates until convergence is attained. Details of our proposed model are described in chapter 4.

Chapter 3. Data Description

In order to explore the impact of obesity we analyze the longitudinal time series BMI data of the U.S. for the period 1988- 2012. The data is obtained from the Behavioral Risk Factor Surveillance System (BRFSS) survey, sponsored by Centers for Disease Control and Prevention (CDC) (BRFSS,1988-2012) .The BRFSS survey includes variables for more than 400,000 adults

interviewed each year. "BRFSS is the nation's premier system of health-related telephone surveys that collect state data about US residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. The survey collects data in all 50 states as well as the District of Columbia and three US territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world." (BRFSS, 1988-2012). We processed about 6 million individual BRFSS records over a period of 25 years. In addition, observations with height or weight as 77 ("do not know") and 99 ("refused") were deleted. Finally, observations from the top race groups - White, Black or African American, Asian, Native Hawaiian or Other Pacific Islander and American Indians are used in the analysis. This included interracial observations within the top five race groups. Only persons aged 23 to 90 were included. The individual level data over 50 states, for the period 1988-2012, are aggregated by age and year. We develop a longitudinal time series of BMI data by state with a primary focus on the proportion of population with minimum BMI > 30. Cross sectional obesity rates are then estimated using the formula described in section 2.3.

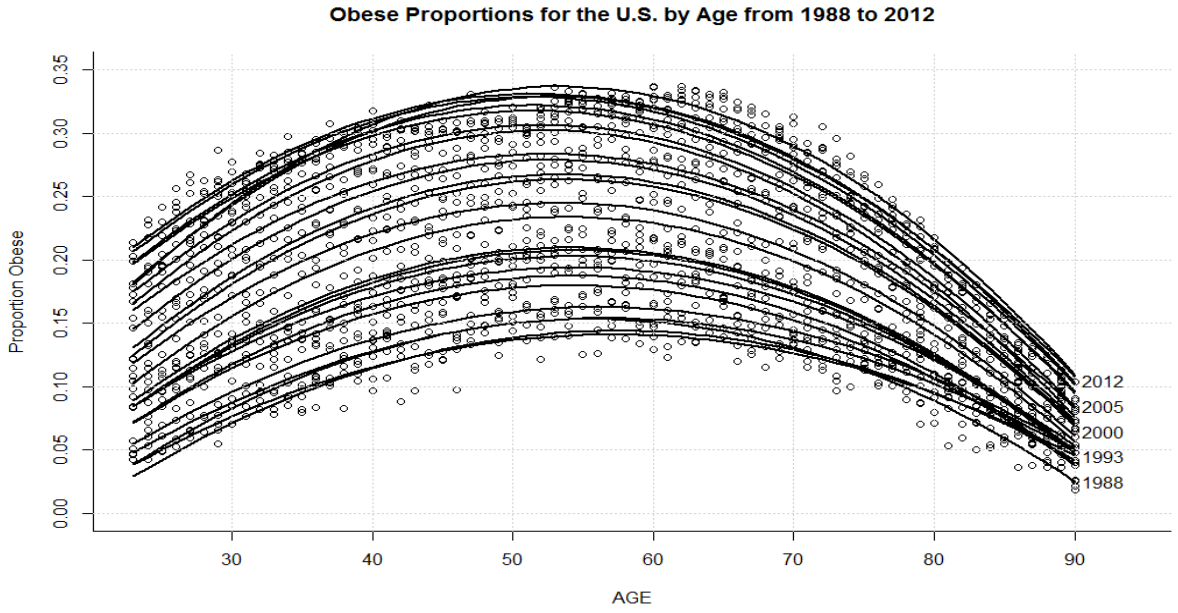
Table 1 Extract Age-Period Contingency table for obesity. Age (rows) and period (columns) Obesity (BMI>30) rates

	BRFSS [2006]	BRFSS [2007]	BRFSS [2008]	BRFSS [2009]	BRFSS [2010]	BRFSS [2011]	BRFSS [2012]
[23]	0.1729702	0.2093023	0.2025316	0.2135981	0.1990262	0.1811757	0.1783061
[24]	0.1938776	0.2167617	0.2273622	0.2178649	0.2319187	0.1905896	0.1812325
[25]	0.2041925	0.2194369	0.2298695	0.2415934	0.2412607	0.1952014	0.1943987
[26]	0.2120335	0.2378609	0.2452050	0.2416165	0.2560570	0.2228346	0.2246661
[27]	0.2172800	0.2307461	0.2526278	0.2674881	0.2612077	0.2286801	0.2130841
[28]	0.2316865	0.2541317	0.2523608	0.2543676	0.2626923	0.2200051	0.2308872
[29]	0.2346003	0.2580487	0.2502765	0.2637028	0.2866988	0.2616693	0.2526010
[30]	0.2290323	0.2602978	0.2386831	0.2773723	0.2697704	0.2531703	0.2520710
[31]	0.2502254	0.2601358	0.2537764	0.2682927	0.2618683	0.2443492	0.2550172
[32]	0.2602324	0.2727905	0.2743136	0.2747570	0.2843354	0.2755319	0.2742607

In Table 1 above, extracts of the full cross sectional data is presented. These rates were obtained by counting observations with BMI >30 divided by the number of samples individuals in the age group in a particular year.

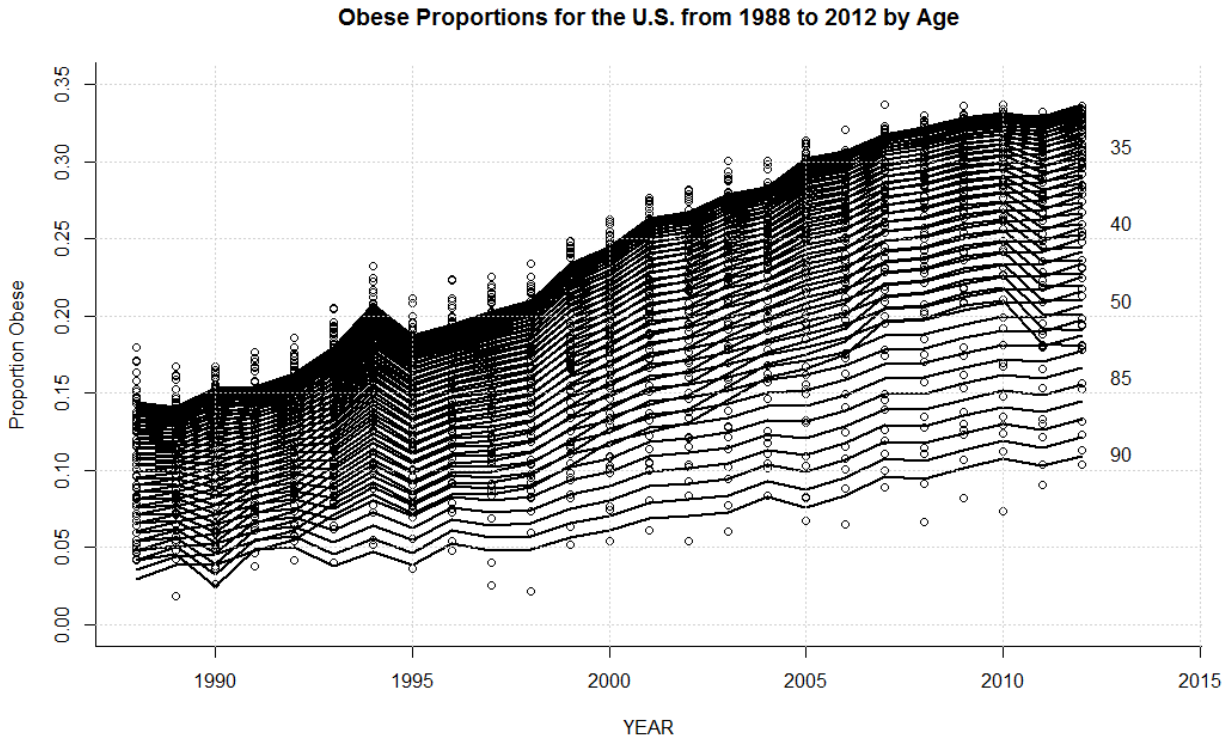
Past studies have used the National Health and Nutrition Examination Survey (NHANES) data to obtain information on the BMI and obesity prevalence .It is worthy to note that the two main surveys for important national health data, the Center for Disease Control (CDC)'s, NHANES and Behavioral Risk Factor Surveillance System (BRFSS) may vary in their sampling and mode of presentation of the data from the survey. The CDC has over the years maintained a sample that is consistent with the population proportion by race of the U.S in the BRFSS survey. In terms of our proposed modeling, we find the BRFSS survey appropriate for the form and structure of the proposed model.

Figure 1 Obesity prevalence in the U.S. by age and period



In Figure 1, we see the cross-sectional obesity prevalence graphed over age and year. The graphical display confirms our assertion that obesity has become more prevalent and a public health concern in recent years. We can see a general increasing prevalence in more recent years than in previous years. The fitted lines represent geometric smooth lines.

Figure 2 Obesity prevalence by period for various age groups



In Figure 2, we can obviously tell a good story and make a conclusion from the graph. Basically we can see that obesity is more prevalent for the middle-aged than for the aged or individuals of higher age groups.

Chapter 4. Methodology

4.1 Statement of Model

To follow the protocols of the stochastic model the random component is stated as

$$\boldsymbol{p(t:x)} \sim Normal(\boldsymbol{u(t;x)}, \boldsymbol{\sigma^2_{x\,t}})$$

The model which is the systematic component is stated as follows

$$p(t:x) = k_t^{(1)} + k_t^{(2)}(x - \overline{x}) + k_t^{(3)}((x - \overline{x})^2 - \sigma_x^2) + \gamma_{t-x}^{(4)} + \varepsilon_t \qquad (7)$$

σ_x^2 is the estimated variance for all ages under consideration in the model. This is included to center the quadratic $\left(x - \overline{x}\right)^2$ predictor.

$\overline{x}$ is the mean of the ages. This is included to center the age predictor.

$p(t:x)$: Probability of a person aged x being obese in year t. Here obese is defined as a person with BMI>30 .

$k_t^{(1)}, k_t^{(2)}, k_t^{(3)}$ are the age effect coefficients associated with various time periods to be estimated by fitting the model to the observed obesity data

$k_t^{(1)}$ is the intercept as usual .

$k_t^{(2)}$ is the coefficient associated with the age variable.

$k_t^{(3)}$ is the coefficient associated with the quadratic effect .

$\gamma_{t-x}^{(4)}$ is the random cohort effect obtained by regressing and applying the constraints .

ε_t is the random error term.

We assume an identity link function, in this case $\eta_{xt} = p(t:x)$.

A polynomial/quadratic regression is assumed, hence allowing us to use Ordinary Least Squares regression to obtain estimates. The advantage of this model is that we are able to predict obesity rates for various ages and particular periods of interest.

4.2 Parameter Estimation

The coefficients $k_t^{(1)}, k_t^{(2)}, k_t^{(3)}$ are obtained using estimation methods in multiple linear regressions. After this, estimates from the fitted obesity rates are then used to compute residuals.

Here we assume "n" is the total number of ages and "m" is length of years/period under consideration. Next, the algorithm uses an m x n matrix based on a unit matrix to set weights to zero (0) for cohorts with fewer than 5 observations and one (1) for cohort years with more than 5 observations. This is because we anticipate that the cohort effect is effective for years with at least more than 5 observations.

An m+n-1 vector of weights is then created with a 0 if the cohort is completely excluded and 1 assigned if otherwise. The permutation order is also obtained for the vector of cohort years. For years with weights assigned 1, the in the vector of weights. The model further imposes the constraint that the fitted quadratic function from the least squares fit of a quadratic function of (t-x) to $\gamma_{t-x}^{(4)}$ be identically zero. The constraints are imposed in order to resolve the identifiability problem. To resolve the identifiability problem this generalization switches to

$$\bar{\Upsilon}_{t-x}^{(4)} = \gamma_{t-x}^{(4)} + \emptyset_1 + \emptyset_2(t - x - \overline{x}) + \emptyset_3(t - x - \overline{x})^2.$$

At every stage of the iteration initial values of $\gamma_{t-x}^{(4)}$ are obtained as the mean of the residuals for cohort years which have weights of 1 after fitting the model below

$$p(t:x) - \gamma_{t-x}^{(4)} = k_t^{(1)} + k_t^{(2)}(x - \overline{x}) + k_t^{(3)}((x - \overline{x})^2 - \sigma_x^2) \quad (8)$$

The parameters $\emptyset_1$,$\emptyset_2$, and $\emptyset_3$, are coefficients estimated by fitting the model using linear regression. The algorithm iteratively regresses the residuals and applies constraints continuously to improve the estimates until convergence is achieved.

Programming to fit the models was performed in the free-to-use R Studio software. Ideas for fitting the cohort effect were adapted from M7 model in the "Lifemetrics" package.

4.3 Identifiability Constraints

Like with many models in the stochastic models literature, our proposed model has an identifiability problem. This is caused by the fact that parameters can be rescaled and shifted but yet have no impact on the proportion of obese individuals $p(t:x)$. That is, many different parameters can be used to obtain the same or similar values of obese proportion $p(t:x)$. As described by Carins et al. (2009) we impose the following constraints to resolve this problem.

$$\sum_{c=c_0}^{c_1} \gamma_c^{(4)} = 0 \quad (9)$$

$$\sum_{c=c_0}^{c_1} c\gamma_c^{(4)} = 0 \quad (10)$$

$$\sum_{c=c_0}^{c_1} c^2\gamma_c^{(4)} = 0 \quad (11)$$

Here c_0 and c_1 are defined as the first and last periods of birth to which the cohort effect are fitted to. The constraint in the expression (9) and (10) are imposed to ensure that the mean-

reverting model is appropriately applied to $\gamma_{c.}^{(4)}$ This is so because the estimates do not carry a linear drift on the over the time period under consideration.

4.4 Forecasting Obesity Using Proposed Model

We can make use of Time Series and forecasting techniques to appropriately forecast obesity prevalence. After observing the stationary process of the fitted values and time series of obesity prevalence an appropriate ARIMA model can be fitted. We investigate the use of appropriate ARIMA models to forecast the period effects coefficients $k_t^{(1)}, k_t^{(2)}, k_t^{(3)}$ as well as the cohort effect $\gamma_{t-x}^{(4)}$.

4.5 Goodness of Fit and Model Evaluation

Generally speaking from a statistical point of view, the best estimates or projections are those from models that minimize the error or residuals. We can simply evaluate models using the Least Residual Errors usually measured by the Mean Squared Error.

Alternatively, we can use some more powerful model selection techniques like the Akaike Information criterion (AIC), the Bayesian Information Criterion (BIC) and the Mean Absolute Percentage Error (MAPE) to evaluate the models and check for goodness of fit of the model.

4.5.1 Bayesian Information Criterion (BIC)

Bayesian Information Criterion (BIC) is usually used for selecting the best model among a set of models. It is also known as Schwarz criterion (SBC, SBIC). Schwarz (1978) proposed

$$BIC = -2\,\boldsymbol{L}(\boldsymbol{\theta}) + kIn(n) \qquad (12)$$

Where k is the number of free parameters estimated and for the Gaussian case the Log likelihood $\boldsymbol{L}(\boldsymbol{\theta})$ can be replaced by$In\left(\widehat{\sigma_e^2}\right)$.

Here $\widehat{\sigma_e^2}$ is defined as the estimated variance associated with the error estimated as

$$\widehat{\sigma_e^2} = \frac{1}{NM}\sum_{x,t}^{NN}(\widehat{p_{x,t}} - p_{x,t})^2 \ ,$$

Models with smaller BIC are preferred.

4.5.2 Akaike Information Criterion (AIC)

AIC was proposed by Akaike (1974) .The Akaike Information criterion (AIC) measures the quality of a statistical model relative to others for a given set of data. It is defined as

$$AIC = -2\boldsymbol{L}(\boldsymbol{\theta}) + 2k \qquad (13)$$

Where k is the number of free parameters estimated and for the Gaussian case the Log likelihood $\boldsymbol{L}(\boldsymbol{\theta})$ can be replaced by $In\left(\widehat{\sigma_e^2}\right)$. Models with smaller AIC are preferred.

4.5.3 Mean Absolute Percentage Error (MAPE)

The Mean Absolute Percentage Error (MAPE) in utility for our purpose is defined below:

$$MAPE = \frac{1}{NM}\sum_{x,t}^{NN}\frac{|\widehat{p_{x,t}} - p_{x,t}|}{p_{x,t}} \qquad (14)$$

Where N (N=25) is the period dimension and M (M=68) is the age dimension. Models with smaller MAPE are preferred.

Chapter 5. Results and Analysis

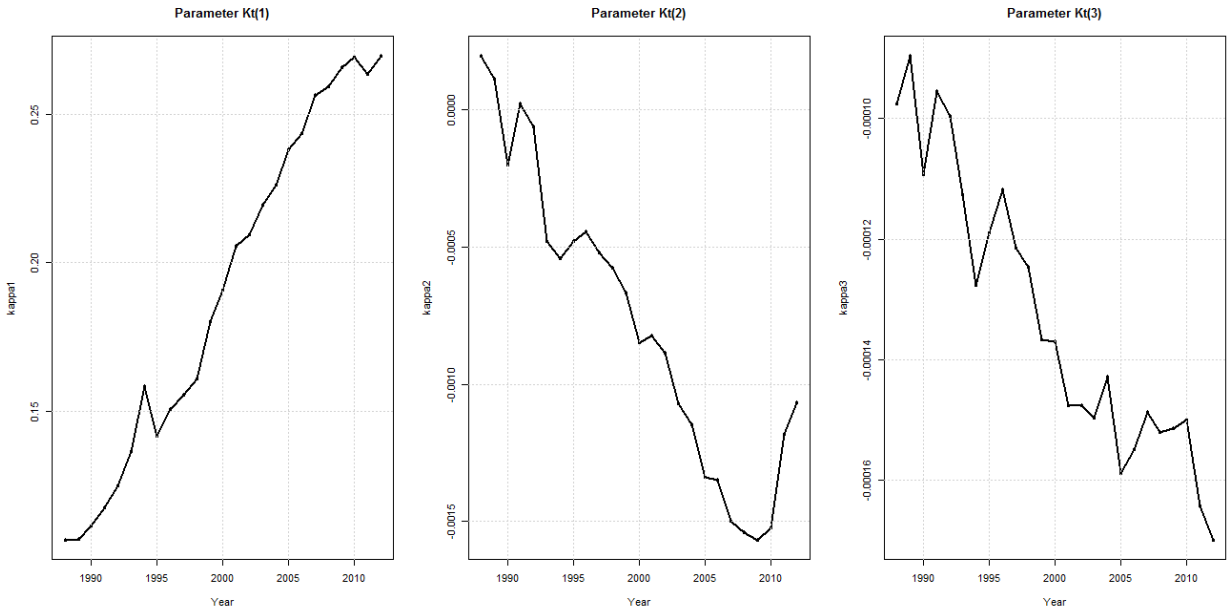
In this section, we explore the results of the various models and make appropriate discussions.

5.1 Results of Parameter Estimates

In our proposed model, we had the task of estimating $k_t^{(1)}, k_t^{(2)}$ and $k_t^{(3)}$. Basically these are period coefficients of the age variable. Figure 3 presents the results of the estimation from fitting the model. There are 25 sets of each of the estimates of the predictors $k_t^{(1)}, k_t^{(2)}$ and $k_t^{(3)}$. Each triplet of $k_t^{(1)}, k_t^{(2)}$ and $k_t^{(3)}$ estimated for each of the 25 years.

We can see that $k_t^{(1)}$ has a positive relationship with time period and a clearly rising trend. This parameter in our model is similar to the intercept of a regression model. The second parameter $k_t^{(2)}$, has a decreasing trend when observed over time. The third parameter $k_t^{(3)}$, which models the quadratic effect also has a decreasing trend when observed over time.

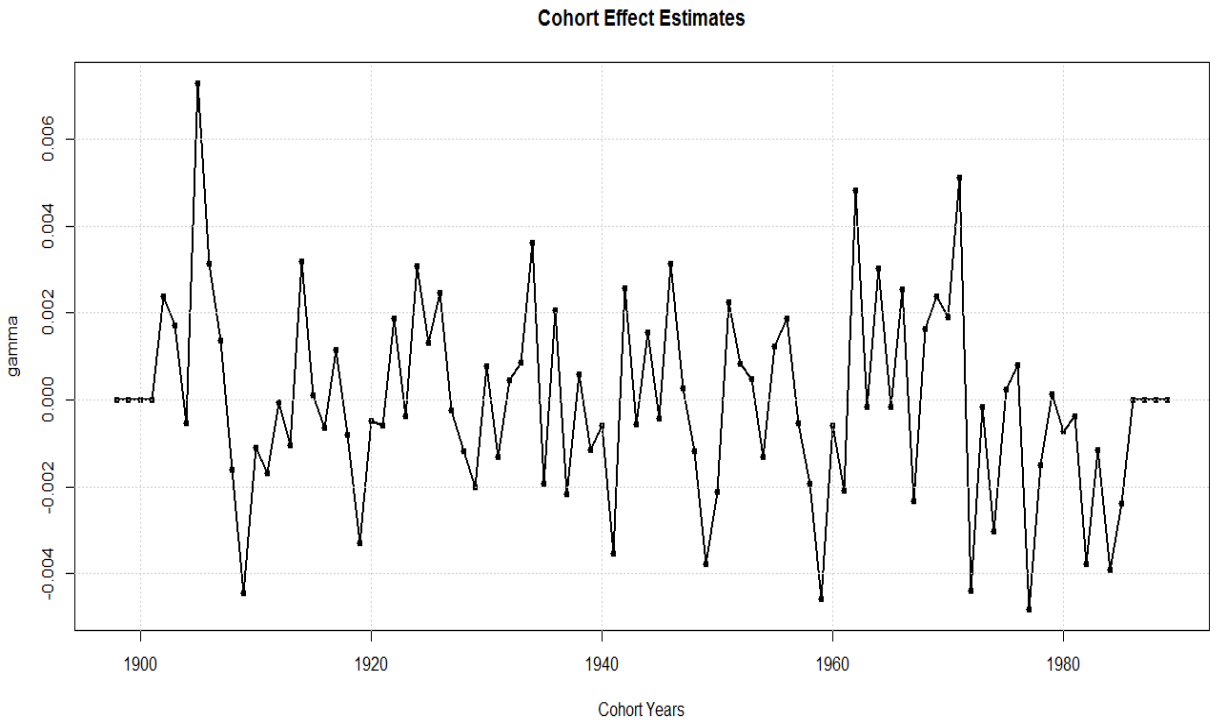
Figure 3 Graph of Parameter estimates from proposed model



5.2 Results of Cohort Effect Estimates

In Figure 4 below, the cohort effect estimates are plotted. In our data the birth cohorts considered are from the year 1898 to the year 1989. With the lowest period under consideration being 1988 and the highest age under consideration being 90, the lowest birth cohort estimated as (Cohort = Priod- Age) was 1898 while the highest was 1989. In Figure 4, we plot the cohort effect estimates for the various periods from 1898 to 1989.

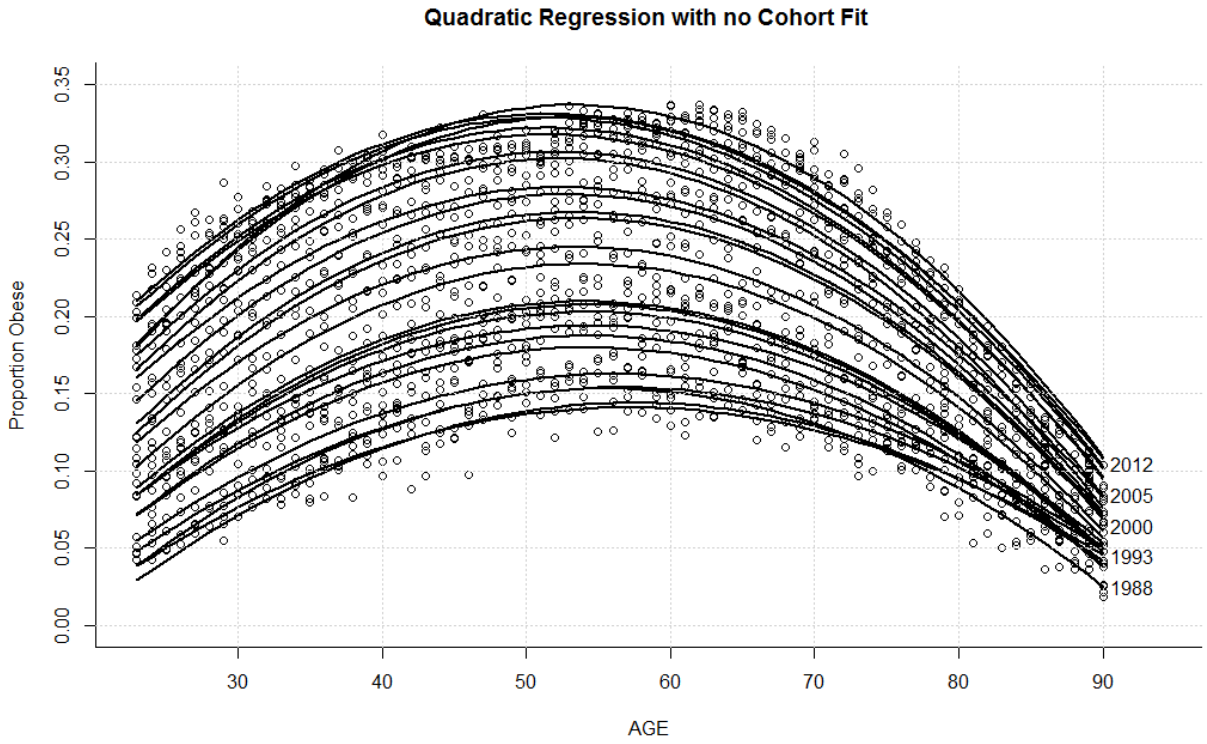
Figure 4 Graph of Cohort effect estimates



As described earlier, we impose identifiability constraints and make a switch of the cohort term $\gamma_{t-x}^{(4)}$ such that the cohort estimates revolve around 0. It is clear from the graph in Figure 4 that the cohort estimates fluctuate either on the negative or positive side of 0.

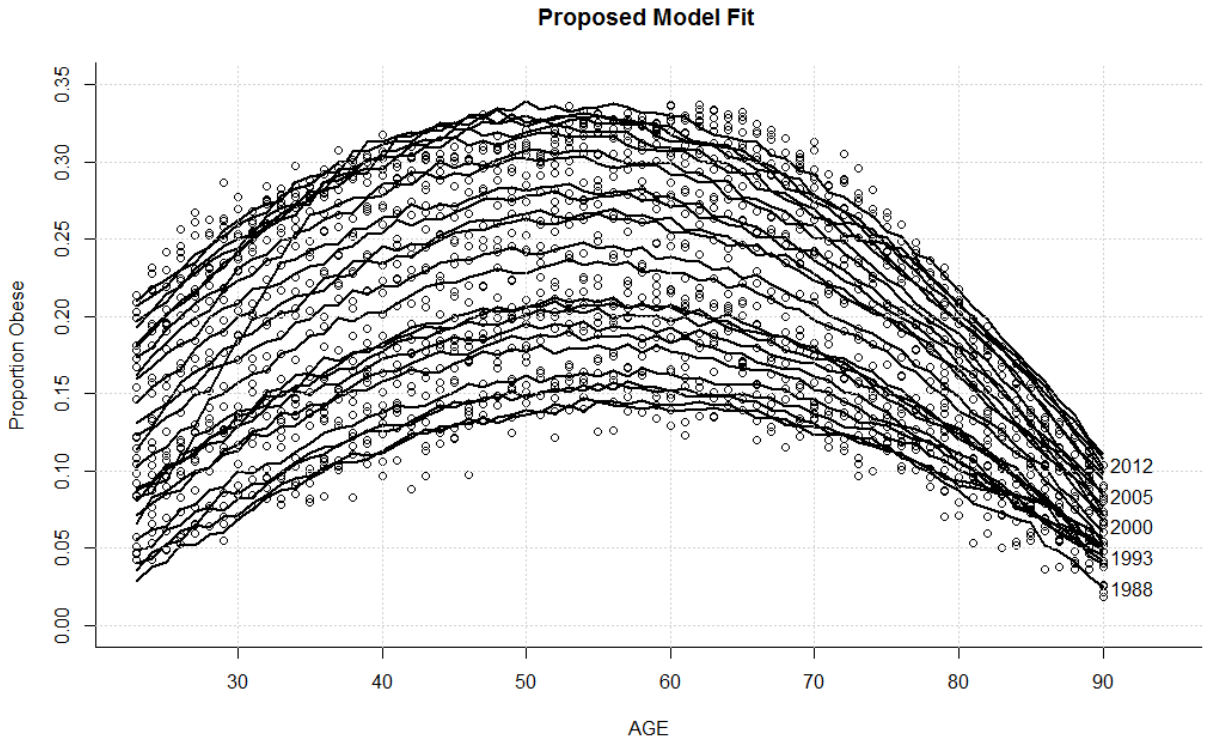
5.3 Results of Fitted Values

Figure 5 Graph of fitted values from Quadratic regression with no cohort



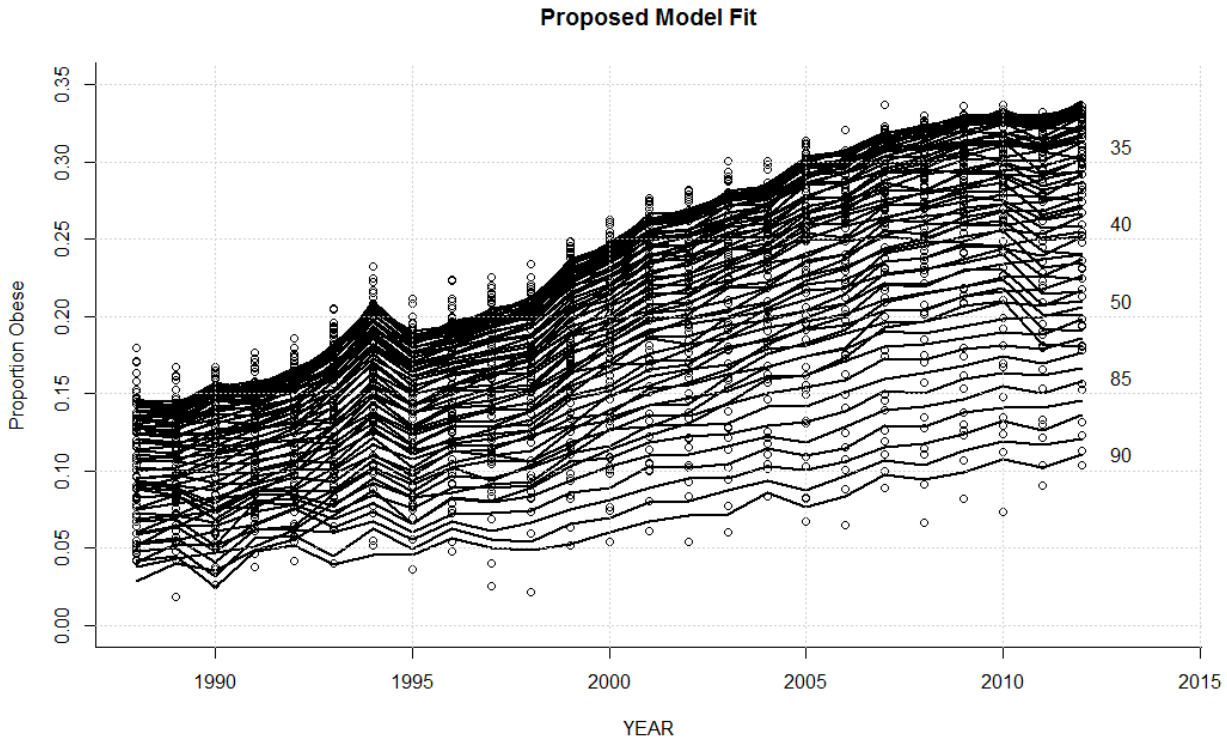
In Figure 5 we illustrate the plot of the quadratic regression portion of our model that is our proposed model without the cohort effect. We can see the fitted lines are smooth curves which basically estimate the mean of the proportion obese at each age value. Here we plot 25 fitted curves representing the model fit for the various periods from 1988 to 2012. It is evidently clear that obesity has become a major issue in recent years as the general average obese proportions have consistently kept on increasing over the period from 1988 to 2012. Recent years generally have had higher proportions of obese individuals at the various ages from 23 to 90.

Figure 6 Graph of fitted values from the proposed model by age



In Figure 6 we illustrate the plot of the proposed model , with the cohort effect . We can see the fitted lines have some adjustments as compared to the smooth curves from the quadratic regression model with no cohort effect. Again, here we plot 25 fitted curves representing the model fit for the various periods from 1988 to 2012. It is evidently clear that obesity has become a major issue in recent years as the general average obese proportions have consistently kept on increasing over the period from 1988 to 2012. Recent years generally have had higher proportions of obese individuals at the various ages from 23 to 90.

Figure 7 Graph of fitted values from the proposed model by year



For the purposes of performing time series analysis and procedures, it was beneficial to obtain the time series of obesity proportions for various age groups from 23 to 90 over the period under consideration, in this case 1988 to 2012.

From Figure 7, we observe the time series plot for various age groups. Usually the lower series represents older ages since in general older people are less likely to be obese. We observe that the series is non-stationary and has an increasing trend for most age groups. The variance of the series appears to be constant over time as there is low variability.

Figure 8 Graph of fitted values from Constraint Based Approach

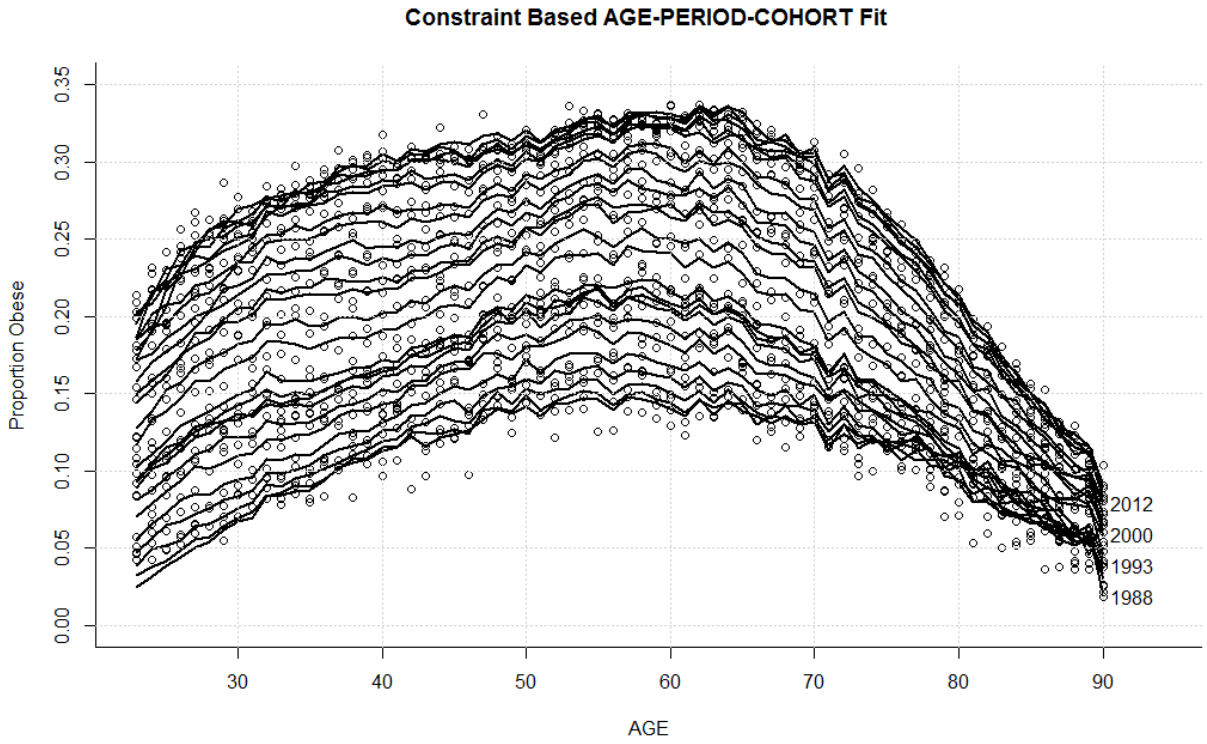
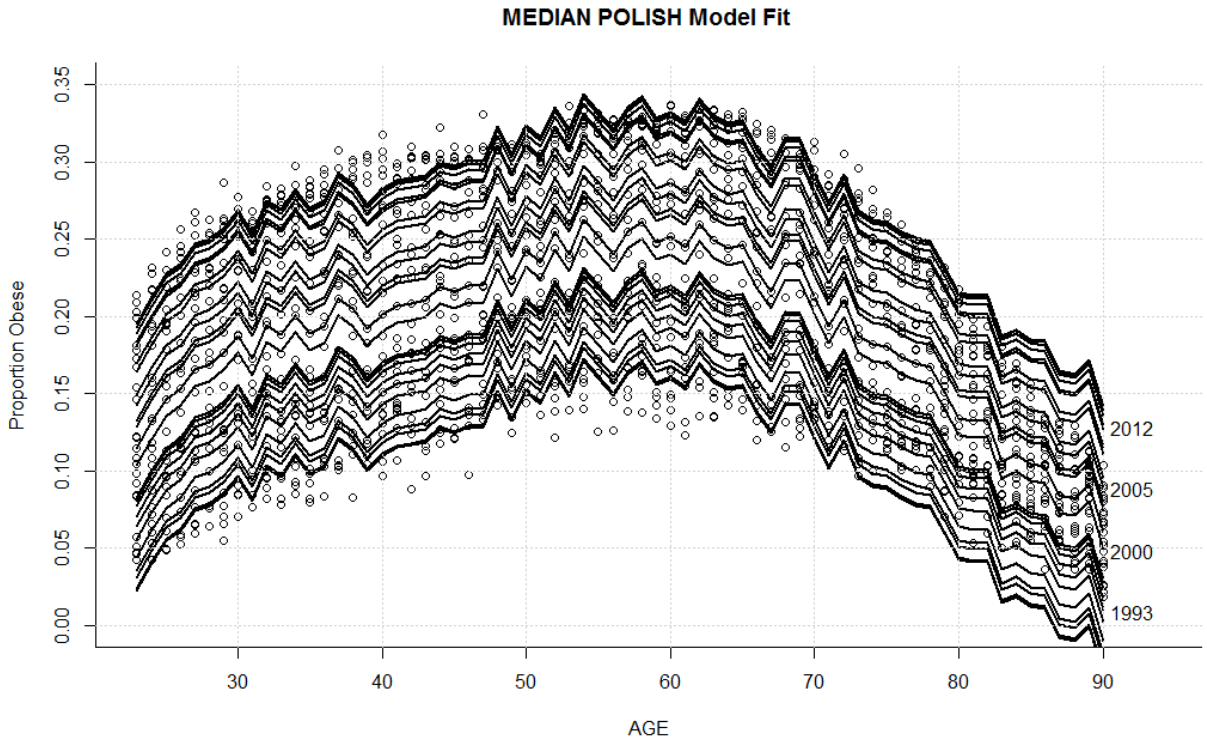


Figure 8 depicts the estimates of fits of obesity rates from the Constraint Based Approach. The model stipulation is equivalent to an ANOVA model with parameters to estimate the cohort effect. Thus in this model the cohort effect is viewed as a 3rd factor effect in the ANOVA model hence there are m+n-1 parameters to estimate. A marked difference between our proposed approach and the Constraint -Based approach is that here the cohort effect is treated as a parameter.

Figure 9 Graph of fitted values from Median Polish Approach



We can see that the estimates of the fitted line from the median polish approach resemble the use of some form of central tendency for the fitting of the lines to the observed obesity rates. We can see the almost strait lines within certain age groups. The nature of the graph is expected, the median polish approach makes use of row and column medians of the two-way contingency table of observed obesity rates across several ages and years

5.4 Results of Evaluation of Goodness of Fit

In this section, we shall make use of commonly known and used statistical evaluation criteria to make a comparison of the models. In Table 2 below, we present results of some of the widely accepted model comparison and evaluation tools computed after fitting our model.

Table 2 Summary Results of Proposed Model and Constraint Based Model

Statistic	Proposed Model	Constraint-based (APC ANOVA model)	Quadratic Regression (No Cohort Model)
MAPE	7.20%	5.71%	7.04%
BIC	1237.00	1370.93	575.10
AIC	345.17	375.71	167.23
Number of Parameters	164	183	75
Mean Squared Error(MSE)	1.8716e-04	1.186409e-04	1.812762e-04

In comparing models, we normally prefer models with smaller values of MAPE, BIC, AIC, the number of Parameters and MSE over their counterparts. From Table 2, we see that on most of the evaluation criteria, the proposed model will be preferred over the Constraint Based Age Period Cohort model. We see that on this basis, AIC, BIC, number of parameters of the proposed model are smaller relative to the Constraint –Based APC Approach. However the MAPE and MSE of the Constraint Based Approach are lower hence better than the proposed model. This is expected as the Constraint-Based APC model has more parameters hence better prediction accuracy. However, more parameters may lead to an inconvenience in prediction as the model will be too complicated. Removing the cohort parameters from our proposed model leads to even better results as can be seen from the table. The quadratic regression model with no cohort has an advantage of fewer parameters, lower AIC and BIC values.

5.6 Forecasting Obesity Prevalence

We undertook time series analysis with our estimated coefficients as well as the cohort effect. We observe that the ARIMA (0,1,0) fits the series of the coefficients $k_t^{(1)}, k_t^{(2)}$ and $k_t^{(3)}$. Since the proposed model without the cohort effect was better compared to the other models, there was no need to forecast and include the cohort effect in making predictions of obesity prevalence. The ARIMA (0,1,0) model is stated as eg. $k_{t+1}^{(1)} = u + k_t^{(1)}$ where u is the long-term drift in $k_t^{(1)}$.

Figure 10 6-yearTime Series Forecast of Model Coefficients and Cohort Effect

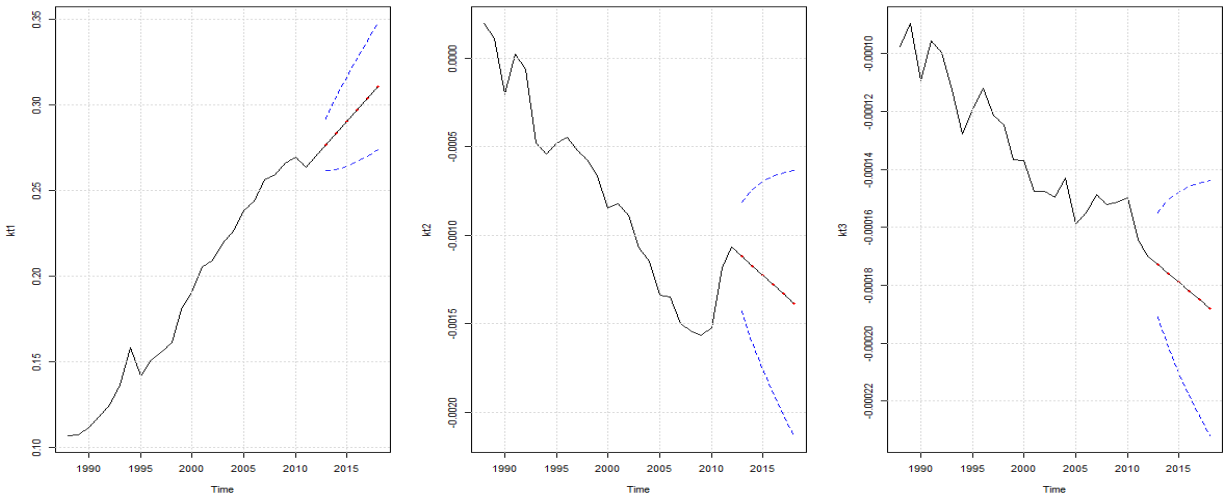


Figure 10 depicts a 6-yearTime Series Forecast of Model Coefficients and Cohort Effect. We observe a rising trend for the intercept $k_t^{(1)}$, a decreasing trend for the coefficient of the age predictor $k_t^{(2)}$, a decreasing trend for the coefficient of the quadratic term $k_t^{(3)}$.

After performing usual time series procedures on the parameters $k_t^{(1)}, k_t^{(2)}$ and $k_t^{(3)}$ we predict future obesity prevalence with the following forecasting model

$$\boldsymbol{p(t+1:x) = k_{t+1}^{(1)} + k_{t+1}^{(2)}\left(x - \overline{x}\right) + k_{t+1}^{(3)}\left(\left(x - \overline{x}\right)^2 - \sigma_x^2\right)} \qquad \textbf{(15)}$$

for selected ages between 23 to 90 years for 6 years .

5.6.1 Residual Bootstrapping of Confidence Intervals from Forecast

To obtain confidence intervals from the proposed forecasts we explore the use of the residual bootstrapping. This method involves simulating and randomizing the residuals from the fitted model and adding back the randomized residuals matrix to obtain bootstrap data. The following steps are used to obtain the bootstrap confidence intervals.

1) Fit the quadratic regression model with no cohort.
2) Obtain and randomize the residuals from the fitted model and obtain 1000 simulated residual error matrix.
3) Obtain 1000 bootstrap matrix of data by adding the randomized residuals to the model fit data.
4) Obtain 6- year forecasts of the parameters from the various simulated bootstrap data using appropriate time series models of the bootstrap fit data. This provides the bootstrap fitted component. To model the residual component we also obtain a 6-year forecast of the residual component of the bootstrap data including the 95% confidence upper and lower limits.
5) Apply forecast parameters to estimate the model forecast projections using the proposed forecasting model in formula (15) and obtain the 95% confidence limits.

Figure 11 6-year Time Series forecast of Obese Proportions for various ages confidence intervals

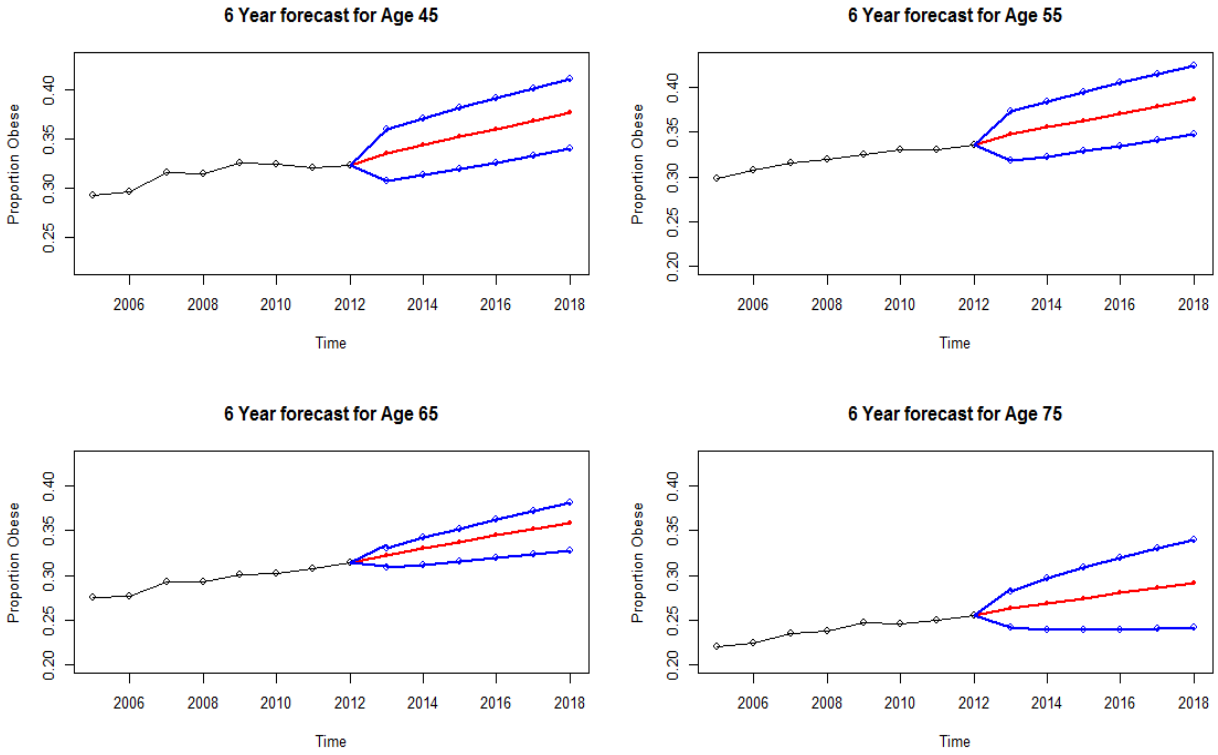
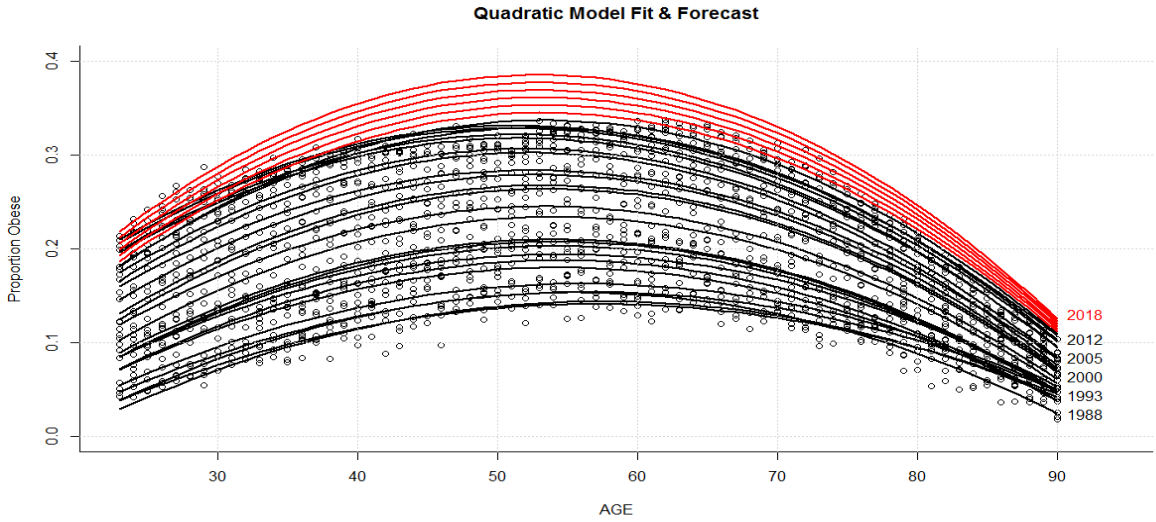


Figure 10 depicts a 6-year bootstrap forecast of obesity rates for individuals of age 45, 55, 65 and 75. The forecasts were based on the proposed forecasting model in formula (15) and the simulations using the residual bootstrapping method .We can see a general increasing trend of obesity prevalence over the period under consideration for all of the ages. Also, we notice from the confidence limits that there is more prediction variability for higher ages than the mid and lower age groups.

Figure 12 6 year projection of obesity across age



In Figure 12, we show a 6-year projection of obesity across various ages. After the 6 year time series forecasts we fit the projected obese proportions across various ages for the years 2013 through to 2018. We generally observe an increasing obese prevalence across the various ages over time.

Table 3 Forecast obese prevalence from model

Age	Observed 2012	Projected 2018	Difference
25	19.44%	24.00%	+4.56%
30	25.21%	28.75%	+3.55%
35	28.41%	32.57%	+4.16%
40	31.73%	35.45%	+3.72%
45	31.03%	37.39%	+6.37%
50	32.04%	38.39%	+6.35%
55	33.16%	38.45%	+5.29%
60	32.00%	37.57%	+5.57%
65	32.96%	35.75%	+2.79%
70	30.72%	32.99%	+2.27%
75	26.71%	29.29%	+2.58%
80	21.75%	24.65%	+2.90%
85	15.66%	19.07%	+3.41%
90	10.37%	12.55%	+2.18%

In Table 3 we show the projected obesity prevalence for selected ages with 5 year age intervals. We observe that the mid age group (35 to 55) has the highest absolute increase over the 6 year period from 2012 to 2018.

Chapter 6. Conclusion

While stochastic methods in estimation and model fitting are not a new concept in the field of mortality and life expectancy studies, it is a relatively new concept in obesity prevalence analysis.

In this study, we have introduced and elaborated on the background of cohort effect estimation which arose from estimation of the log mortality rate at various ages for various years various years. Predicting obesity prevalence is useful as this enables various states, localities and counties to take appropriate action to educate citizens about the consequent effect of obesity on health and life expectancy.

Researchers in medical sociology and epidemiology interested in studying the distribution and etiology associated with obesity will benefit from the use of the proposed stochastic methods.

Life and health insurance firms who are interested in accurately classifying and quantifying risk will benefit from the proposed methods to account for the obesity risk factor in pricing of life insurance and annuity products. Actuaries and underwriters in these firms will also be able to appropriately classify risks associated with prospective life and health insurance customers.

Existing methods have usually used the age bin (21-25,26-30,31-35, etc.) and year bin (1971-1975,1976-1980,1981-1985 etc) approach ,however the proposed method makes use of

individual age and year groups and uses a convenient matrix iterative algorithm to estimate obesity rates for various ages and appropriate periods of interest.

While existing ANOVA/Effects models may have better fitting, our proposed model tend to have an advantage of using fewer predictors to achieve a fit almost as good as the existing methods. By including the cohort effect we are able to account for some of the variability in the data and hence reducing the amount of unexplained variability.

We have also performed time series analysis and suggested appropriate ARIMA and forecasting models that are best for forecasting the prevalence of obesity.

The ultimate goal of predicting these obesity rates is to enables us to appropriately adjust mortality risk and hence life expectancy as well as other important demographic indicators. We are hopeful that future studies in this area will aim at incorporating some form of obesity adjustment into fitting of mortality rates.

References

Allison D B, Fontaine KR, Manson JE, Stevens J, VanItallie TB. (1999). Annual deaths attributable to obesity in the United States.JAMA;282:1530-8.

Akaike, H. (1974). A new look at the statistical modeling identification. IEEE Transactions on Automatic Control, vol. 19, 716-723.

Behavioral Risk Factor Surveillance System (BRFSS) survey (1988-2012). National Center for Chronic Disease Prevention and Health Promotion. Available from. https://www.cdc.gov/brfss/annual_data/annual_data.htm Accessed 05.18.16

Cairns, A.J.G., Blake, D., and Dowd, K. (2006b) A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration," *Journal of Risk and Insurance,* 73: 687-718

Cairns, A.J.G., Blake, D., Dowd, K , Coughlan G.D, Epstein D., Ong A., Balevich I. (2007). A quantitative comparison of stochastic mortality models using data from England & Wales and the United States . DISCUSSION PAPER PI-0701

Cairns A.J.G , Blake D, Dowd K, Coughlan G. D , Epstein D , Ong A , Balevich I.(2009) A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States, North American Actuarial Journal, 13:1, 1-35, DOI: 10.1080/10920277.2009.10597538

Ebbeling, C., Pawlak, D., Ludwig, D.(2002) Childhood obesity: Public-health crisis, common sense cure**.** *The Lancet*. 360:473–482.

Fontaine R. K. , Redden .T.D, Wang C., Westfall O. A., Allison B. D.(2003). Years of Life Lost due to Obesity . JAMA, 289:187-193

Glenn, N. D. (2005). Cohort analysis (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. Biometrics, 39, 311–324.

Holford, T. R. (1991). Understanding the effects of age, period, and cohort on incidence and mortality rates. Annual Reviews in Public Health, 12, 425–457.

Holford, T. R. (1992). Analysing the temporal effects of age, period and cohort. Statistical Methods in Medical Research, 1(3), 317–337.

Hunt, A., Blake, D. (2015). On the Structure and Classification of Mortality Models. Pension Institute Working Paper. http://www.pensions-institute.org/workingpapers/wp1506.pdf Accessed 02.10.16

Keyes, K. M., Utz, R. L., Robinson, W., & Li, G. (2010). What is a cohort effect? Comparison of three statistical methods for modeling cohort effects in obesity prevalence in the United States, 1971–2006. *Social science & medicine*, *70*(7), 1100-1108.

Last, J. M. (2001). A dictionary of epidemiology (4th ed.). Oxford University Press.

McCullagh, P., Nelder, J., (1989). Generalized Linear Models, 2nd Edition. Chapman & Hal, London.

Miljkovic T, Shaik S, Miljkovic D.,(2016) .Redefining standards for body mass index of the US population based on BRFSS data using mixtures. Journal of Applied Statistics.

Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH.(1999). The Disease Burden Associated With Overweight and Obesity. *JAMA.*;282(16):1523-1529. doi:10.1001/jama.282.16.1523

National Center for Health Statistics. (2005). Analytic and reporting guidelines: the national health and nutrition examination survey (NHANES). Available from. http://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_ guidelines_dec_2005.pdf. Accessed 02.05.09.

Olivieri, A., Pitacco, E. (2009). Stochastic mortality: the impact on target capital. *Astin Bulletin*, *39*(02), 541-563.

Olshansky, S.J., Passaro, D.J., Hershow, R.C., Layden, J., Carnes, B.A., Brody, J., Hayflick, L., Butler, R.N., Allison, D.B. and Ludwig, D.S.(2005). A potential decline in life expectancy in the United States in the 21st century.*New England Journal of Medicine*, *352*(11), pp.1138-1145.

Oxford Advanced learner's Dictionary *OED Online*. (2016) Oxford University Press, 2016. Accessed 02.10.17

Palmore, E. (1978).When can age, period, and cohort be separated? *Social Forces*,*57*(1), 282295.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Renshaw, A.E., and Haberman, S. (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors," *Insurance: Mathematics and Economics*, 38: 556-570.

Ryder, N. (1965). The cohort as a concept in the study of social change. American Sociological Review, 30(6), 843–861.

Schwarz, G. (1978). Estimating the dimension of model. The Annals of Statistics, vol.6, 461-464.

Selvin, S. (1996). Statistical analysis of epidemiologic data. New York: Oxford University Press.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MS: Addison-Wesley Publishing Company.

Tutt, L. W. G. (1953). The mortality aspect of population projections. Transactions of the Faculty of Actuaries, 21, 3–50.

Villegas, A. M., Kaishev, V. K., & Millossovich, P. (2015). StMoMo: An R Package for Stochastic Mortality Modeling.

Weight, H. (2015) . The Health Effects of Overweight and Obesity. Centers for Disease Control and Prevention (CDC) .Available from. https://www.cdc.gov/healthyweight/effects/?iframe=true&width=95%&height=95% Accessed 01.02.17

Yang, Y., Schulhofer-Wohl, S., Fu, W. J., & Land, K. C. (2008). The intrinsic estimator for age–period–cohort analysis: what it is and how to use it? American Journal of Sociology, 113, 1697–1736

Appendix

For R code and data files used in this study contact daawinpa@miamioh.edu