Title
Mentally Unhealthy Days Among Los Angeles Immigrants: A Finite Mixture Modeling Approach

Author
Lu, Yi-Li

Publication Date
2012

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Mentally Unhealthy Days

Among Los Angeles Immigrants: A Finite Mixture Modeling Approach

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Yi-Li Lu

2012

Abstract of the Thesis

Mentally Unhealthy Days

Among Los Angeles Immigrants: A Finite Mixture Modeling Approach

by

Yi-Li Lu

Master of Science in Statistics

University of California, Los Angeles, 2012

Professor Susan Cochran, Chair

This thesis examines the impact of immigration on reports of mentally unhealthy days among respondents in the 2007 Los Angeles Health Survey. I argue that the findings in the extant literature are unstable due to unobserved heterogeneity in response. I fit 3 different regression models: the Poisson, the Negative Binomial, and the mixture of Negative Binomial models. From the test of goodness fits, the Mixture of the Negative Binomial models has a better fit than the other two traditional statistical models. A significant mixing proportion of my mixture model indicates that mixture of the Negative Binomial models is necessary. Two distinct distributions indicate that the model fits and identifies two kinds of people: distressed and non-distressed individuals. I use the finite mixture parameter estimates to calculate the posterior probability of being in the non-distressed group; meanwhile, I find evidence that race and economic status play important roles in classification but not migration-related factors, including years in US, citizenship and language ability.

The thesis of Yi-Li Lu is approved.

Jan de Leeuw

Rick Paik Schoenberg

Susan Cochran, Committee Chair

University of California, Los Angeles

2012

To my husband . . .

who—among so many other things—

always stands behind me anywhere and anytime

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I deeply appreciate my advisor, Professor Susan Cochran, for giving me the opportunity to work on this project. This thesis would not have been possible without her patience, motivation, enthusiasm and support. Her guidance helped me in all the time of research and writing of this thesis.

I would like to thank the members of my committee, Professor Rick P. Schoenberg and Professor Jan de Leeuw, for their helpful suggestions. I also thank Professor Vickie M. Mays, Department of Psychology, UCLA, for her instructions and assistance in various aspects.

I thank my friends and classmates at UCLA for their help and company. Without them, my graduate life would be boring. Thanks for all those who always support and encourage me during the completion of the thesis.

Finally, My deep gratitude goes to my parents and my husband, Sy-Chyuan Hwu, for their unconditional love and support throughout my entire graduate study. I also thank my lovely boy, Chi-Shiang Hwu, because his smile encourage me to move forward.

CHAPTER 1

Introduction

Los Angeles is a city historically rich in ethnic diversity. It is estimated that 35.5% of the population in Los Angeles County was not born in United States and the immigrants generally came from countries in Asia, the Pacific Islands, and Central and South America.[1] The process of migration and resettlement can create a risk for mental health distress.[1] Immigrants face many difficulties including coping with the past experience in their native countries, overcoming cultural and language barriers and encountering discrimination. All of these can provoke or aggravate depression and other mental problems.[2] Therefore, the task of preventing, recognizing and appropriately treating common mental health problems becomes quite critical in urban settings such as Los Angeles.

Pre-migration, migration and post-migration resettlement are the phases usually regarded as the three stages of immigration trajectory.[3] In the pre-migration stage, individuals might suffer from the disruptions of social support or the changes of economic, educational and occupational status in their original county. Once the migration has happened, individual's personality, age and gender, language ability, education level and family support may well play a role in managing the transition from one culture to another. In the last stage, post-migration, individuals start to learn new roles and become member of the recipient society.

[1] Source: U.S. Census Bureau, 2010 American Community Survey

The process of migration is extremely heterogeneous and not all migrants are likely to face similar experience before or after migration.[4] The process may involve one individual who moves to study, seek better employment, attempt to achieve a better future or to avoid political and religious persecution, or to marry. The duration of the process also differs from individual to individual. It may be temporary, permanent, seasonal or may occur once. The impact of migration is not only on the immigrants themselves but also on the second or third generations. Immigrants often represent healthier individuals such that the health of immigrants to the USA tends to be better than that of the US-born population of the same national origin. [5, 6]

The effects of immigration on psychological functioning is complicated. I shall review some of the key studies and discuss the risk factors of mental health problems among immigrants. Demographic characteristics of individuals are significant in understanding the migration experience. In a classic study, Odegaardd[1] stated that rates of schizophrenia in his Norwegians in the US sample were higher among those who had been in the US for 10-12 years than others and explained this increase as a consequence of migration. However, Sashidharan[7] challenged this interpretation and argued that experiences of migration by black and ethnic minority groups are not the same as those of white Norwegians. He also warned that researchers must take into account possible differences in the migration experience between ethnic groups.

Both age and gender are critical modifiers of psychiatric risk from the migration experience. Young adults are more likely to migrate and young immigrants are also more likely than older ones to be at the risk of developing mental disorders. Bhugra et al[8, 9] found that Asian women aged 18-24 were 2.5 times more likely than White women and seven times more likely than Asian men to attempt suicide. Moreover, Odegaard[1] reported that incidence of mental disorders was higher in females then males among immigrants. Additionally, Murphy also drew the same conclusion with Odegaard and explained that the reason may be the males who decide to migrate and females simply follow them. For men, it is may be easier to accept the stress when they chose to migrate. For females, they may perceive less control over the process and this may produce higher levels of stress.[10]

Another important factor is education level. Tseng [11] noted that when people decide to settle in a new country, educational attainment plays an important role during the process of adjustment. He suggested that those who are highly educated or poorly educated have more difficulties than those who are possess more normative educational level in obtaining appropriate and satisfactory occupations. Sometimes individuals with higher level qualifications may end up doing menial jobs and this situation can be distressing.

In addition to demographic factors, migration-related factors can not be ignored. Several community studies have pointed out that the rate of psychiatric disorder increases by length of stay in United States among immigrants of Hispanic ethnic origin.[12] In addition, researchers have established that fluency in the language of recipient society may accelerate the process of culture adjustment. In particular, evidence from previous studies in Japan[13] and in UK[14] show that not knowing the local language makes life difficult in the new environment. Other factors can also have an impact on the mental health of immigrants including worries about legal status.[15]

The process of immigration influences the mental conditions among immigrants heavily. Hypothetically, it is entirely possible that migrants are more depressed than natives because of the higher frequency of loss events and the stress of living in a new society, for example. However, this is not necessarily so; Regier et al.[16, 17] found a opposite conclusion. Several studies have also shown that the rates of common mental disorders are higher among groups without migration. I propose that this inconsistency comes from the ignorance of heterogeneity of effects for different latent classes of observations. For example, non-distressed and distressed individuals definitely perform differently when they face the stress of migration. However, these two latent types of respondents are hard to detect in traditional surveys. This phenomenon is frequently called unobserved heterogeneity. Thus, there are hypothetically latent subgroups in any data set, and in my data set in particular and I want to explain this heterogeneity by using covariates such as age, gender, races, and the migration-related factors to identify patterns of mentally unhealthy days among immigrants.

From a statistical perspective, I desire to find a suitable model which identifies unobserved heterogeneity. Assuming that a latent structure underlies response patterns leads directly to considering a finite mixture model methodology to address the presumed complicated structure. I argue that the pattern of mental health distress among immigrants has a mixture structure and finite mixture models could help us to understand the heterogeneity.

CHAPTER 2

Data

2.1 Data Source

This thesis presents findings from the adult (18 years and older) respondents in the 2007 Los Angeles County Health Survey (LACHS), a population-based survey commissioned by the Los Angeles County Department of Public Health and conducted by the Field Research Corporation. The main purpose of LACHS is to provide updates on key health indicators and to identify emerging public health issues among the LA County population. The survey is periodic and the most recent version is 2007, following surveys conducted in 1997, 1999-2000, 2002-2003 and 2005.

The 2007 LACHS has a sample size of 7,200 adults, interviewed via structured telephone interviews. It was conducted in several languages, including English, Spanish, Chinese (Mandarin and Cantonese), Korean and Vietnamese. The design and weighting methodology are described in the methodology document on the website of Department of Public Health, LA County. [1] In this research, I focus on the respondents who were first generation immigrants only. Therefore, my analysis sample is restricted to the respondents who answered they were born outside the United States. The final analysis sample numbered 2635.

[1] Source:http://publichealth.lacounty.gov/ha/LACHSBackMeth2007.htm

2.2 Variable Selection

2.2.1 Mental health status

This thesis employs a measure of mental health status that was obtained by asking respondents "Thinking about your mental health, which includes stress, depression and problems with emotions, for how many days during the past 30 days was your mental health not good?". This question was developed by the Centers for Disease Control and Prevention (CDC)[18], which aimed to assess people's perceived mental health and mental distress. It has been a regular question in the CDC's Behavioral Risk Factor Surveillance System since 1993 and has been shown to have acceptable test-retest reliability[19], construct validity[20], "known-groups" validity[20] and predictive validity[21] across a number of studies. Conventionally, the poor mental health question is one of the common components of health-related quality of life(HRQOL) measure.[22, 23, 24]

2.2.2 Independent variables

- Years in US: In the thesis, there are three migration-related variables: years in US, citizenship and Language ability. The variable, years in US, is based on the question, "How many years have you lived in the United States?" I regarded years in US as a continuous variable and the range is from 0 to 86 years.

- Citizenship: Respondents who was not born in United States are also asked whether they are currently a U.S. citizen or not. Individuals who are citizens are coded as 0 while the individuals without citizenship are coded as 1.

- Language of interview: At the beginning of the interviews, respondents were informed that the survey could be conducted in the following languages: English, Spanish, Mandarin, Cantonese, Korean and Vietnamese. The language used in the survey was recorded as a binary variable: English and non-English(Spanish, Mandarin, Cantonese, Korean and Vietnamese). I use English speakers as the reference group.

- Age: I regarded age as a continuous variable and age ranged between 18 to 97 years.
- Gender: The gender of respondent is also included in the analysis. Male is coded as 0 (referent group) while female is coded as 1.
- Education: The question used to determine educational attainment was "What is the highest level of school you have completed or the highest degree you have received?" I grouped answers into one of four categories: less than high school, high school, some college or trade school and college or post graduate degree. Less than high school was designated as the referent group.
- Race/ethnicity: Respondents were asked to their ethnic and racial background in two questions. The first asked whether the respondent was Hispanic. The second asked respondents to identify the racial category or categories to which they belong from the following list: White, Black/African-American, Asian, Pacific Islander, American Indian/Alaskan native, Hispanic/Latino, and Others. I combined the answers of the two questions to categorize respondents into five groups: White, Black/African American, Latino, Asian/Pacific Islander and Others. (Respondents who indicated membership in more than one of the above groups were classified into Others.)
- Federal poverty level: Respondents reported their household income and number of family members. This was categorized into federal poverty level

(FPL)[2] thresholds (less than 100 %FPL, 100-199 %FPL, 200-299 %FPL, 300 %FPL or above). Less then 100%FPL is the reference group.

- Employment status: In the LACHS, there are several questions pertaining to respondents' current employment situation. The question, "Please tell me all that apply to you... are you self-employed or working for a family owned business, are you employed for pay by some other organization, are you looking for work, are you a homemaker or keeping house, are you retired from the labor force, are you unable to work because of a disability, are you not looking for work, or are you a student?", was used to divide respondents into three groups: employed, unemployed, not in workforce. People not in workforce include students, retired persons, homemakers and those unable to work.

All the missing values in the independent variables are imputed by the Hot Deck method. Hot-Deck imputation is one of the popular and widely used imputation methods.[25]

[2]Poverty status is based on U.S. Census 2006 FPL thresholds which for a family of four (2 adults, 2 dependents) correspond to annual incomes of $20,444 (100% FPL), $40,888 (200% FPL), and $61,332 (300% FPL).

CHAPTER 3

Models

3.1 Poisson and Negative Binomial model

Poisson regression is used to model numeric variables, but in the form of counts. Counts are all positive integers and follow a Poisson distribution rather than a Normal distribution. The density function for the Poission model is given by

$$Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

where $y! = y(y-1)(y-2)\ldots(2)1$, and $y \geq 0$. A Poisson regression model is sometimes known as a log-linear model because it is in the family of generalized linear models with the logarithm as the link function[26]. Poisson regression assumes the logarithm of the expected value of the response variable Y change linearly with equal increment increases in the covariates. In other words, the typical Poisson regression models expresses the log outcome as a linear function of a set of predictors.

However, Poisson regression has another strong assumption, the conditional means equal the conditional variances.[26] It could be stated as $E(Y) = var(Y) = \lambda$. This means that once the mean is estimated, the variance is estimated. In applied situations, this assumption is quite limiting and data appropriate for Poisson regression do not happen very often.[27] One of the most common problems is over-dispersion. The variance is greater than expected from a simple Poisson distribution. Nevertheless, Poisson regression is often used as a starting point for modeling count data.[27]

A negative binomial regression model can be used in most situations where one would apply a Poisson model if there is concern about over-dispersion.[28] The density function for the Negative Binomial model is given by

$$Pr(Y = y|\mu, \alpha) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)y!}\left(\frac{\mu}{\alpha + \mu}\right)^y\left(\frac{\alpha}{\alpha + \mu}\right)^\alpha$$

where μ is the mean of y which is $E(y|\alpha, \mu) = \mu$; α is called over-dispersion and $var(y|\alpha, \mu) = \mu(1 + \frac{\mu}{\alpha})$. If the Poisson parameter λ is not considered fixed but assumed to follow a gamma distribution, the Negative Binomial distribution is obtained. This means, the Poisson distribution is a special case of Negative Binomial distribution when the parameter λ goes to infinity. This could be stated as $Poisson(\mu) = \lim_{\alpha\to\infty} NB(\mu, \alpha)$. Peter Schlattmann has proved that the Negative binomial distribution can be thought of as a Poisson distribution with unobserved heterogeneity, which can be conceptualized as a mixture of two probability distributions, namely, Poisson and gamma.[29] However, the choice of the gamma distribution as the mixing distribution is somewhat arbitrary and sometimes the data will not fit well under simple Negative Binomial regression model.

3.2 Finite Mixture Model

3.2.1 Introduction of Finite Mixture Model

Finite mixture models already have a long standing history in Statistics since Pearson's (1894) classic mixtures paper on a truly Bayesian approach to the mixture problem.[30] Because of the flexibility, finite mixture models have become popular and have been used in a wide range of applications. The books by Everitt and Hand[31], Mclachlan and Basford[32], Bohning[33] and more recently Fruhwirth-Schnatter[34] describes the theory of finite mixture models and its application thoroughly.[29]

A mixture model is a probability model for representing sub-populations each with an individual distribution for the overall population. In the finite mixture model, the random variable Y is independent, identically distributed p-dimensional observations drawn from one additive mixture of K distinct subgroups in proportions π_k. The general expression of the probability density function for the finite mixture model is as follows:

$$f(y; \pi_1, \pi_2, \dots, \pi_k) = \sum_{k=1}^{K} \pi_k f_k(y)$$

where,$0 < \pi_k < 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Here, K represents the total number of components with $\pi = (\pi_1, \pi_2, \dots, \pi_k)'$. Usually, $f_k(y)$ are assumed to be of parametric *i.e.*$f_k(y) \equiv f_k(y; \vartheta_k)$ and the functional form of $f_k(.;.)$ is completely known, but the parameterizing vector ϑ_k is unknown. If the component distributions are of the same distributional form, the mixture is called homogeneous. In most applications of homogeneous mixtures, the mixing probabilities do not depend on regression parameters. Thus, the general model could be simplified to

$$f(y; \vartheta) = \sum_{k=1}^{K} \pi_k f_k(y; \vartheta_k)$$

where

$$\vartheta = (\pi', \vartheta_1', \vartheta_2', \dots, \vartheta_k')$$

When the number of mixture components, K, is also unknown, K and the vector ϑ are both estimated. However, K is estimated by theoretical suggestion in most of the cases, and only ϑ has to be estimated.

In current case of count data, the finite mixture model of the Poisson distribution and the finite mixture model of the Negative Binomial distributions are considered instead of the simple Poisson model or the parametric mixture model, the Negative Binomial model. For the Poisson mixture, the mixture density for observation y is given by

$$f_k(y;\vartheta_k) = \frac{exp(-\lambda_k)\lambda_k^y}{y!}$$

Thus, each subpopulation is desired by a Poisson distribution with parameter λ_k. As to the mixture model of the Poisson distribution, the mixture density of the Negative Binomial for observation y is given by

$$f_k(y;\vartheta_k) = \frac{\Gamma(\alpha_k + y)}{\Gamma(\alpha_k)y!}\left(\frac{\mu_k}{\alpha+\mu_k}\right)^y\left(\frac{\alpha_k}{\alpha+\mu_k}\right)^{\alpha_k}$$

Again, each subpopulation is modeled with its own parameters, μ_k and α_k. The finite mixture models are estimated using maximum likelihood. Cluster-corrected robust standard errors are used throughout for inference purposes. These methods are implemented using the STATA package `fmm`.

Under this approach, clusters are represented as probability models in a model space; in other words, each model represents one particular cluster. This is the reason why finite mixture modeling is has been called model-based clustering as well.[35] However, in the vast clustering literature, the discriminative (or distance/similarity-based) approach is another way to combine cases into groups, which is fundamentally different from the model-based clustering approach.[36] In the discriminative approach, clustering relies on a measure of closeness or similarity of observations, and then groups similar objects together. The converse of similarity is distance, and many different similarity/distance measures are discussed in the literature.[36] Although discriminative methods usually produce desirable clustering results, model-based clustering methods provide better interpretation since the resulting model for each cluster is directly characterized observations within that cluster. Although clustering has evolved from finite mixture modeling, the two approaches have distinct goals: finite mixture modeling is typically associated with inference on the model and its parameters while the goal of model-based clustering is to provide a partition of the data into groups of homogeneous observations.

3.2.2 Posterior Probability

Although the main aim of mixture models is to understand the relationship between the dependent variable and the independent variables, the goal is also to estimate the probability of being in the latent groups for each observation. To achieve classification, model-based clustering requires an additional step after model-fitting that assigns each observation to different groups according to some pre-specified rule. Mixing proportions can be thought of as the prior probability that an observation originated from a specific mixing distribution. In this thesis, I use a Bayes rule at this step which allocates observations to clusters in accordance with their posterior probabilities. Thus, every observation will be assigned to the group having the highest posterior probability that the observation originated from this group.

As mentioned in the previous section, finite mixture modeling could help here to calculate the posterior probability of membership in each latent class, conditional on all observed covariates and outcomes. The classification is done by the Bayes Theorem, as

$$Pr(y_i \in k|y_i;\vartheta) = \frac{\pi_j f_j(y_i|\vartheta_j)}{\Sigma_{k=1}^{K}\pi_k f_k(y_i|\vartheta_k)}, j = 1, 2, \ldots, K$$

Thus, I will fit the data to a finite mixture model first and obtain parameter estimates which will be used to calculate the posterior probability of being in each of the latent classes by using the formula above. As a consequence, the estimated posterior probability will vary across observations. In the next step, this set of estimated posterior probability becomes the dependent variable, and efforts are now directed at determining which individuals have higher probability to be in component 2. That means what are the key factors associated with mental distress among immigrants classified to non-distressed individuals.

CHAPTER 4

Results

4.1 Descriptive Statistics of Mentally Unhealthy Days

The frequency distribution of mentally unhealthy days in Los Angeles County adults is shown in Figure 4.1. The distribution is strongly skewed with most people reporting no or few unhealthy days. Clearly a normal approximation model or a square-root transformation is not appropriate to use as a theoretical distribution for this process. At the same time, the box plot in Figure 4.1 shows there are many outliers in the dataset even with the assumption that the underlying distribution reflects a Poisson process. Although the Poisson model is one of the standard methods in analyzing count data, if over-dispersion occurs, the Negative Binomial model is commonly considered as an alternative.[29] From Table 4.1, the mean of mentally unhealthy days is 2.75 while the variance is 50.98. The variance is nearly 25 times larger than the mean and they are definitely not identical. The distribution of mentally unhealthy days is displaying signs of over-dispersion.

Furthermore, Figure 4.2 provides another strong suggestion of over-dispersion. In the boxplot of mentally unhealthy days by age groups, each subgroup has a long tail and the mean is much smaller than the variance as well. Those outliers show that the variation within each subgroups cannot be ignored. On the other hand, the variation between each subgroup is also apparent. From Figure 4.3, the range of the average mentally unhealthy days between subgroups is from 1.55 to

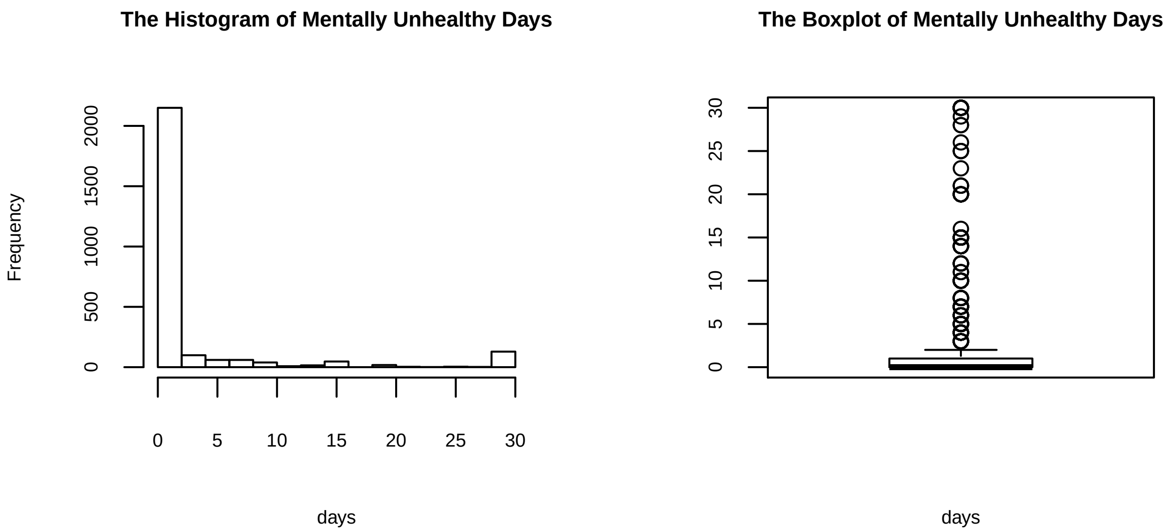

Figure 4.1: The Histogram and Box Plot of Mentally Unhealthy Days foradults in the 2007 Los Angeles County Survey

Min	1st Qu	Median	Mean	3rd Qu	Max	Variance
0	0	0	2.75	1	30	50.98

Table 4.1: The Summary of Mentally Unhealthy Days

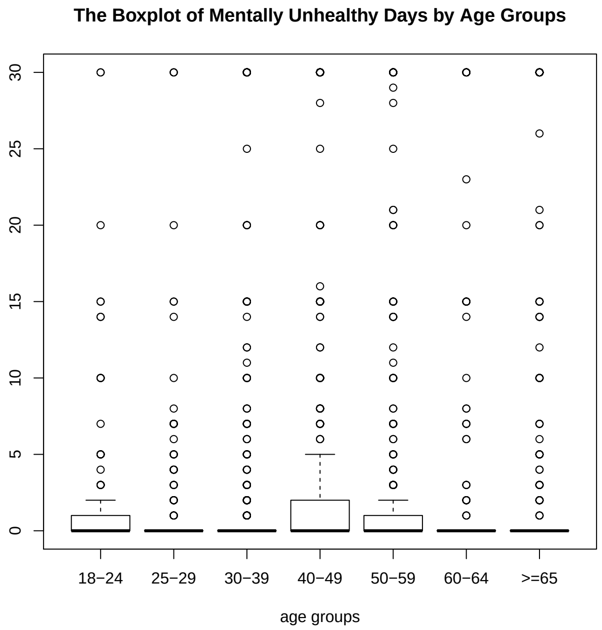

Figure 4.2: Boxplot of Mentally Unhealthy Days for Los Angeles County Adults by Age Groups

3.49. Although the median by subgroups are all zeros, the variability within and between subgroups are large. This leads to the rejection that the simple Poisson model is going to fit the data very well.

In the following sections, I try to fit different models and find an appropriate statistical model that provides reasonable interpretations in order to determine the important risk factors associated with reports of mentally unhealthy days among Los Angeles County adults. Poisson regression is still shown first, even though I believe that the simple Poisson model is not, on the face of it, a best choice. My first attempt to deal with the over-dispersion is to fit a Negative Binomial regression, which allows for the variance to be larger than the mean. As I mentioned before, heterogeneity may exist in the data set so a mixture of Negative Binomial models is used. Finally, I use the finite mixture parameter estimates to calculate the posterior probability of being in each of the latent

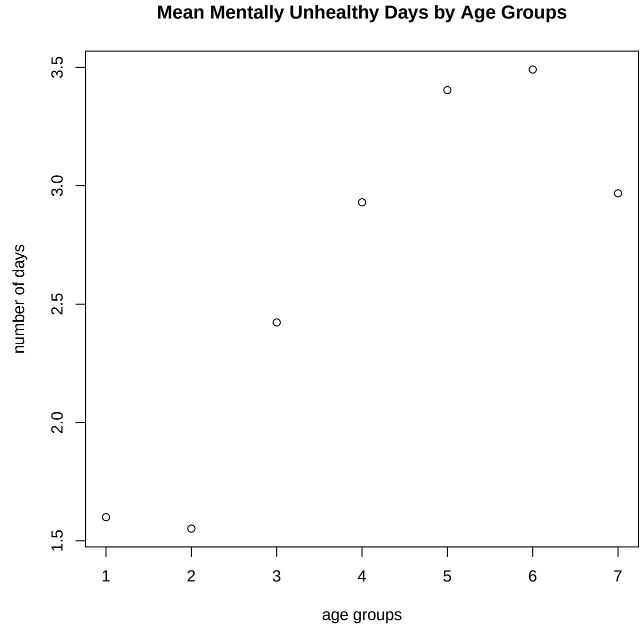

Figure 4.3: Scatter plot of the Mean of Mentally Unhealthy Days by Age Groups classes.

4.2 Poisson Regression and Negative Binomial Regression

Frequently for count data, a Poisson distribution is assumed. Because of the issue of over-dispersion, the Negative Binomial regression model is also considered. Table 4.2 shows the results of regressing number of days in poor mental health on our independent variables, including age, gender, non-English speaker, education level, citizenship, years in US, race, federal poverty level and employment status. The parameter, alpha, is highly significant meaning that the problem of over-dispersion exists and the Negative Binomial model is a significant improvement over the standard Poisson model. Based on the Wald test of all the predictors, whether the person has some college or trade school education, is African American or Asian/Pacific Islander, is unemployed and has a lower income appear to be significant predictors of the number of mentally unhealthy days.

Looking at the χ^2 goodness-of-fit test statistics, I find poor fits in both of the Poisson and the Negative Binomial regression models ($\chi^2 = 78.58$ in Poisson regression and $\chi^2 = 66.89$ in Negative Binomial regression, 17 degree of freedom). The tests support rejection of the hypothesis that the deviance follows a chi-square distribution with degrees of freedom equal to the model residual are significant, meaning that neither a standard Poisson nor a Negative Binomial regression is probably appropriate in our case. Because of unobserved heterogeneity in response, I have to extend my model to mixture models. By taking latent clusters into account, I hope a finite mixture model with Negative Binomial-distributed subpopulations may have better performance over simple Poisson and simple Negative Binomial models.

4.3 Mixture of Negative Binomial Models

For finite mixture models, estimating the number of components is always an important issue. In the LACHS case, respondents are randomly selected and can be easily divided into two groups: non-distressed and distressed people. Thus, I arbitrarily determined the number of components, K, and fit the mixture of Poisson models with 2 components, meaning that $K = 2$. In addition, the simple Negative Binomial model could be regarded as a special case of mixture of Negative Binomial models with $K = 1$. Hence, the simple Negative Binomial model is a nested model to achieve a mixture of Negative Binomial models. To compare fits between simple and mixture of Negative Binomial models, the Akaike information criterion (AIC) and the Bayesian information Criterion (BIC) are examined.

AIC was developed by Hirotsugu Akaike in 1974 [37] and he proposed AIC as

	Poisson	Negative Binomial
YEARS IN US	0.008 (0.006)	0.006 (0.008)
NON-CITIZEN	-0.056 (0.147)	0.051 (0.142)
NON-ENGLISH SPEAKER	0.036 (0.193)	-0.134 (0.211)
AGE	0.006 (0.004)	0.012 (0.006) **
GENDER	-0.014 (0.133)	0.079 (0.129)
EDUCATION LEVEL		
high school	0.009 (0.189)	0.085 (0.208)
some college or trade school	0.358 (0.184) *	0.419 (0.194) **
college or post graduate degree	0.052 (0.207)	0.161 (0.199)
RACE		
Latino	-0.026 (0.239)	0.123 (0.248)
African-American	1.161 (0.493) **	1.039 (0.623) *
Asian/Pacific Islander	0.549 (0.232) **	0.650 (0.249) ***
Others	-0.022 (0.255)	0.130 (0.255)
FEDERAL POVERTY LEVEL		
100% to <200%FPL	-0.402 (0.140) ***	-0.457 (0.154) ***
200% to <300%FPL	-0.476 (0.195) **	-0.460 (0.209) **
300%FPL or above	-0.547 (0.211) ***	-0.546 (0.211) ***
EMPLOYMENT STATUS		
Unemployed	0.756 (0.232) ***	0.829 (0.249) ***
Not in workforce	0.277 (0.143) *	0.213 (0.143)
Constant	0.544 (0.358)	0.173 (0.386)
α		10.16167 (.552)***

Table 4.2: Poisson Regression and Negative Binomial Regression for Mentally unhealthy days

	NegBin	Component1	Component2
YEARS IN US	1.01 (0.01)	1.01 (0.01) *	1.10 (0.04) **
NON-CITIZEN	1.05 (0.14)	0.93 (0.15)	6.81 (0.49) ***
NON-ENGLISH SPEAKER	0.87 (0.21)	1.04 (0.17)	0.09 (1.01) **
AGE	1.01 (0.01) **	1.01 (0.00) *	0.85 (0.10) *
GENDER	1.08 (0.13)	0.95 (0.13)	5.98 (0.56) ***
EDUCATIONAL LEVEL			
high school	1.09 (0.21)	1.06 (0.18)	0.00 (2.17) ***
some college or trade school	1.52 (0.19) **	1.38 (0.18) *	7.81 (0.98) **
college or post graduate degree	1.17 (0.20)	1.20 (0.20)	4.38 (0.66) **
RACE			
Latino	1.13 (0.25)	1.12 (0.22)	13.52 (1.53) *
African-American	2.83 (0.62) *	3.04 (0.35) ***	12.08 (1.55)
Asian/Pacific Islander	1.91 (0.25) ***	2.01 (0.22) ***	13.15 (1.54) *
Others	1.14 (0.25)	1.24 (0.25)	14.94 (1.40) *
FEDERAL POVERTY LEVEL			
100% to <200%FPL	0.63 (0.15) ***	0.69 (0.14) ***	2.07 (0.64)
200% to <300%FPL	0.63 (0.21) **	0.70 (0.21)	2.21 (0.84)
300%FPL or above	0.58 (0.21) ***	0.52 (0.24) ***	4.01 (0.54) ***
EMPLOYMENT STATUS			
Unemployed	2.29 (0.25) ***	2.09 (0.24) ***	6.12 (0.63) ***
Not in workforce	1.24 (0.14)	1.27 (0.14)	0.00 (1.17) ***
Constant	1.19 (0.39)	3.83 (0.37) ***	0.01 (2.57) **
π_1			0.67 (0.02) ***

Table 4.3: Incidence Rate Ratios in Negative Binomial and Finite Mixture of Negative Binomial Models for Mentally Unhealthy Days

a measure of the relative goodness of fit of a statistical model among candidate models. The definition of AIC score is as below.

$$AIC = -2 \times ln(L) + 2 \times p$$

where ln(L) is the log-likelihood of the model and p is the number of parameters. The AIC score takes into account both the statistical goodness of fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters. Hence, the model with the smallest AIC is deemed the "best" model.

Another frequently used numeric measure for goodness fit of the model is Bayesian information criterion(BIC), or alternately the Schwarz criterion, which was introduced by Gideon E. Schwarz[38]. It is also used as a tool of model selection and is defined as

$$BIC = -2 \times ln(L) + ln(N) \times p$$

where, again, ln(L) is the log-likelihood of the model, p is the number of parameters and N is the number of observations. Lower values of the BIC scores indicate the preferred model.

Table 4.4 displays the AIC and BIC scores of the simple and the Mixture of Negative Binomial models. The Mixture of Negative Binomial models is the better choice than the simple Negative Binomial model whether by AIC or by BIC scores. Next, I am going to look at the Mixture of Negative Binomial models closely and make statistical inference among those predictors.

Table 4.3 presents both the results of the simple and Mixture of Negative Binomial models. Unlike the simple Negative Binomial model, mixture models are

	NB	Mixture of NB
AIC	8,777,045	8,533,863
BIC	8,777,157	8,523,092

Table 4.4: AIC and BIC scores of the Negative Binomial and Mixture of Negative Binomial models

	Min	1st Qu	Median	Mean	3rd Qu	Max	Variance
Component 1	2.836	5.830	7.669	9.098	11.034	54.365	23.553
Component 2	0.000	0.000	0.000	0.056	0.002	17.331	0.299

Table 4.5: The predicted mentally unhealthy days in two Components.

capable of modeling the heterogeneity between respondents and allow for drawing better conclusions about the associations of predictors reports of mentally unhealthy days. While gender, being a non-English speaker, citizenship and years in US had no association with reports of mentally unhealthy days in the simple Negative Binomial regression, those factors are highly significant to Component 2.

The mixing weights of two components are 0.33 and 0.67, correspondingly. More individuals are classified as being in Component 2 than in Component 1. If we assume respondents are divided by their mental distress conditions, the Component 1 should be the people who are mentally distressed; on the other hand, the Component 2 should be the people who are mentally non-distressed. Following the theory of finite mixture models, I expect that two subgroups have their own distributions and parameter estimates.

According to the estimates of the constants in the model, if all of the predictors in the model are evaluated at zero, the predicted number of mentally unhealthy days would be 3.83 in Component 1 and 0.01 in Component 2. The result indi-

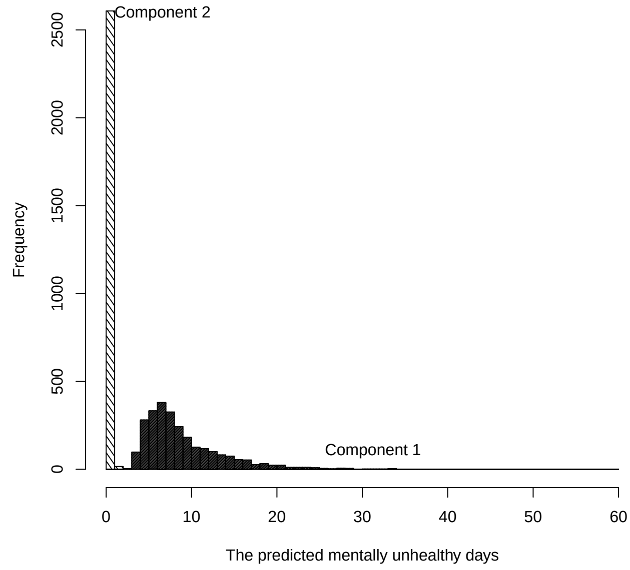

Figure 4.4: The Histograms of Predicted Mentally Unhealthy Days within Component 1 and Component 2

cates that one group of people report no mentally unhealthy days while the others do. The two distinct distributions are observed in the predicted unhealthy days of the observations as well. From Table 4.5, the means and variance of the predicted mentally unhealthy days in Component 1 are much larger than those in Component 2. Fig 4.4 shows the two distinct distributions divided by mental health distress condition clearly. Predicted mentally unhealthy days in Component 1 are spread out wider while most of the individuals are predicted around 0 mentally unhealthy day in Component 2.

The Negative Binomial model uses a log link and models the mean number of events, so the coefficients are easier to interpretation on the mean scale rather than on the log scale. The incident rate ratios (IRR) describe the change in days associated with a one-unit increment in an explanatory variable relative to the reference group. First of all, the migration-related factors, such as years in US, citizenship, and language ability, do not show a strong association with mentally unhealthy days in Component 1 but they are highly significantly correlated to mentally unhealthy days in Component 2.

Furthermore, although years in US is not statistically significant in simple Negative Binomial model, it is significant in both Component 1 and Component 2 when I use the Mixture of Negative Binomial models. Residing longer in US is associated with reporting more mentally unhealthy days when the individual is distressed. Next, there is no association between citizenship and mentally unhealthy days among individuals in Component 1; however, people who do not have US citizenship are expected to have 6.81 times more mentally unhealthy days in Component 2. It tells us that for non-distressed people, having US citizenship or not has a strong association with the mentally unhealthy days.

The issue of language fluency deserves to be carefully considered. Non-English speakers have statistically insignificant increases in mentally unhealthy days if they belong to the non-distressed group. There is a 91% reduction in the mean number of mentally unhealthy days for non-English speakers in Component 2. This result is at odds with my earlier hypothesis and suggests that English ability is not a relevant factor for the Component 1 and also, for non-distressed people, non-English speakers report less distress than English speakers. However, this conclusion is tentative due to possible measurement bias arising from misunderstanding during the phone interview.

In addition to migration-related factors, other demographic factors are in the mixture model. For ages, the model of Component 1 indicates that 1% increase in the predicted mentally unhealthy days are expected for every one year increase. For Component 2, age is a statistically significant variable as well but in a different direction. For people who are classified as non-distressed individuals, 15 % decrease in the predicted mentally unhealthy days are expected for every one year increase in age . As to gender, it is a highly significant factor in Component 2 but not in Component 1. Females are predicted to have 5.98 times more mentally unhealthy days than males if they are regarded as non-distressed.

Education level is another notable factor in the mixture models. Relative to people who are less than high school educated, people who are high school educated are much happier in Component 2, but people who are some college or trade school educated, people who have college or post graduate degree are predicted to have 7.81 and 4.38 times more mentally unhealthy days, relatively speaking. These results are consistent with the hypothesis that lower and higher educated people have more difficulties if they are the non-distressed respondents. However, the hypothesis do not hold in the distressed respondents group because the association between education level and mentally unhealthy days is not strong.

There were also racial/ethnic differences in the models. In Component 1, African -American and Asian/Pacific Islander, when compared to whites, are predicted to have 3.04 and 2.01 times more mentally unhealthy days, respectively. For people who are classified as non-distressed individuals, Latino, Asian/Pacific Islander, and other races as compared to whites were estimated to 13.52, 13.15 and 14.94 times more mentally unhealthy days.

Moreover, there is another difference between individuals across latent classes. The estimates are in opposite directions between Component 1 and Component 2. Compared to people who report incomes under 100% FPL, people in other FPL levels reported fewer mentally unhealthy days if they are grouped as distressed individuals. In contrast, if they are grouped as non-distressed individuals, people who are 300% FPL or above compared to people who are under 100% FPL were estimated to have a rate of 4.01 times greater mentally unhealthy days. Finally, employment status was also significantly predictive of mentally unhealthy days in both latent classes. Whether in Component 1 or 2, unemployed persons reported more mentally unhealthy day than people who were employed. However, students, retired persons, homemakers and those unable to work reported much fewer mentally unhealthy days than employed persons later when they are identified as non-distressed people.

To sum up, the mixture of Negative Binomial model fits and identifies two distinct groups of people. One group of people report nearly no mentally unhealthy days while the other do. The associations of predictors and reports of mentally unhealthy days are different between those two subgroups. I did not observe an association between language ability and mentally unhealthy days nor citizenship and mentally unhealthy days in the distressed group. However, for the non-distressed group, most of the variables are statistically associated with the dependent variable.

4.4 Model-based Clustering

In this section, I present estimates of latent class membership, or the posterior probability of belonging to one of the subgroups identified in the mixture model analysis. By investigating the determinants of the posterior probability of being assigned to Component 2, it provide clarity about the relationship between latent classes and predictors. I use model selection techniques to choose a subset of independent variables to best explain the dependent variable. Both the AIC and the Best Subset Selection method are used here to achieve this mission.

number of variables	Names of variables	logLikelihood	AIC
0	intercept	-1491.2	2982.35
1	intercept, igender	-1488.3	2978.58
2	intercept, igender, ifpl_300	-1487.0	2977.99
3	intercept, igender, irace_asian, ifpl_300	-1485.7	2977.49
4	intercept, igender, irace_latino, irace_others, ifpl_300	-1484.3	2976.68
5	intercept, igender, irace_latino, irace_others, ifpl_200, ifpl_300	-1483.2	2976.31
6*	intercept, igender, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300	-1482.1	2976.20
7	intercept, igender, iedu_tradsch, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300	-1481.7	2977.35
8	intercept, igender, iedu_highsch, iedu_tradsch, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300	-1481.1	2978.20
9	intercept, igender, iedu_highsch, iedu_tradsch, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300, ijob_un	-1480.9	2979.88
10	intercept, igender, eng, iedu_highsch, iedu_tradsch, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300, ijob_un	-1480.9	2981.86
11	intercept, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, irace_latino, irace_black, irace_others, ifpl_200, ifpl_300, ijob_un	-1481.0	2984.07
12	intercept, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, irace_latino, irace_black, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un	-1481.2	2986.34
13	intercept, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, ibplcitizn, irace_latino, irace_black, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un	-1481.7	2989.40
14	intercept, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, ibplcitizn, irace_latino, irace_black, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un, ijob_out	-1482.1	2992.24
15	intercept, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, ibplcitizn, irace_latino, irace_black, irace_asia, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un, ijob_out	-1482.9	2995.76
16	intercept, age_nomiss, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, ibplcitizn, bplyr_nomiss, irace_latino, irace_black, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un, ijob_out	-1483.5	2998.94
17	intercept, age_nomiss, igender, eng, iedu_highsch, iedu_tradsch, iedu_collegesch, ibplcitizn, bplyr_nomiss, irace_latino, irace_black, irace_asia, irace_others, ifpl_100, ifpl_200, ifpl_300, ijob_un, ijob_out	-1484.6	3003.25

Table 4.6: Best subsets for one, two, up to 17 variables for mentally unhealthy days data

	Coefficient	standard error
Gender	-0.074	(0.084)
Latino	0.197	(0.106) *
African-American	-0.036	(0.834)
Other races	0.313	(0.123) **
200% to < 300%FPL	0.117	(0.136)
300% FPL or above	0.183	(0.108) *
constant	0.559	(0.112) ***

Table 4.7: Determinants of the posterior probability of being in Component 2

The Best Subset Selection method uses the simple exhaustive search algorithm.[39] This approach carries out calculations for all models with or without each of the regression terms that are specified in the model. For example, if one has two models A and B, each having the same number of explanatory variables, model A is considered to be better than model B if the sum of squares for A is less than that for B. I list all the "best" models which have the lowest sum of squares compared with other models having the same number of independent variables. Then, AIC is used as the selection criterion to choose among the "best subsets" of various sizes.

Table 4.6 shows all the best subsets models for one, two, and, up to 17 variables for mentally unhealthy days among LA county immigrants. Based on the AIC scores, the model with intercept, gender, race and FPL is chosen to be the best model. Surprisingly, three migration-related factors, years in US, citizenship and language ability, do not play significant roles in allocating individuals to components. The hypothesis that living longer in the US would be associated with lower levels of distress is not supported by the statistical result. As shown in Table 4.7 , race and federal poverty level are the two significant factors in the classification between the distressed and non-distressed groups. Those who are Latino and other races and whose FPL is located above 300% are significantly more likely to be in Component 2. In other words, the characteristic for being assigned to the non-distressed group is being Latino or other vs. White and Asian and having high income.

CHAPTER 5

Conclusion

In this thesis, I used Los Angeles Health Survey data to examine the risk factors for reporting mentally unhealthy days among immigrants. The simple Poisson, simple Negative Binomial and mixture of Binomial models are applied. By checking the goodness of fit individually, I found the mixture of Negative Binomial models fits the data well. From the result, the mixture model approach is an appropriate way to understand reports of mentally unhealthy days among immigrants. Additionally, a substantial heterogeneity in mental distress is shown in the model. From the predicted dependent variable for each individual, two distinct distributions are clearly observed. One with less or no distress and one with distress present. This suggests that tradition statistical analysis obscures the underlying distribution. However, my hypothesis that migration-related factors would lie at the heart of these distributional differences was not supported.

The effectiveness of finite mixture model technique in excavating latent structure of the data has been demonstrated, but at the same time this analysis could be improved. First, although LA county health survey was conducted in 5 languages, I still believe that there is a large proportion of non-English speakers who were not reached by the survey. It is highly possible that the phone calls are always picked up by persons who are fluent in English in their family. Second, the number of components are not examined in the thesis. Model selection among mixture models involves choosing K number of components. I arbitrarily choose

2 as the number of components. However, it would be desirable to pursue an investigation of how many components should be specified in the model.

Last, other mixture models may fit the data better. From the histogram of mentally unhealthy days, I recognize that there are some respondents who feel sad in all the past 30 days and the censoring problem should be taken into account in future analyses. Censoring is a form of a missing data problem that occurs because of the limit of observation time. Special techniques could be used to handle censored data. Survival models examine the time it takes for events to occur and deal with the problems caused by censoring.[40] In the next step of this project, I intend to fit a survival model, such as Cox proportional hazard model or accelerated failure time model, in the context of mixture model approach.

References

[1] Odegaard O. Emigration and insanity. Acta Psychiatrica et Neurologica Supplementum 1932; 4:1-206.

[2] Berry JW. Acculturation as varieties of adaptations. In: Padilla AM, ed. Acculturation. Boulder, CO, USA: Westview Press, 1980: 9-26

[3] Kirmayer LJ, Narasiah L, Munoz M, Rashid M, Ryder AG, Guzder J, et al. Common mental health problems in immigrants and refugees: general approach in primary care. CMAJ. 2011 Sep 6;183(12):E959-67. Epub 2010 Jul 5.

[4] Bhugra D. Migration and mental health. Acta Psychiatr Scand. 2004 Apr;109(4):243-58.

[5] Burnam MA, Hough RL, Karno M, Escobar JI, Telles CA. Acculturation and lifetime prevalence of psychiatric disorders among Mexican Americans in Los Angeles. J Health Soc Behav. 1987 Mar;28(1):89-102.

[6] Alegra M, Canino G, Shrout PE, Woo M, Duan N, Vila D, Torres M, Chen CN, Meng XL. Prevalence of mental illness in immigrant and non-immigrant U.S. Latino groups. Am J Psychiatry. 2008 Mar;165(3):359-69. Epub 2008 Feb 1.

[7] Sashidharan SP. Afro-Caribbeans and schizophrenia. Int Rev of Psychiatry 1993;5:129-144.

[8] Bhugra, D. Attempted suicide in West London. I: Rates across ethnic communities. Psychological Medicine, 29, 1125-1130

[9] Bhugra, D. Attempted suicide in West London. II: Rates across ethnic communities. Psychological Medicine, 29, 1131-1139

[10] Murphy HBM. Socio-cultural factors in schizophrenia. In: Zubin A, Freyhan V, eds. Social Psychiatry. 1968. NY, USA: Grune and Stralton, 74-92.

[11] Tseng W.T. handbook of cultural psychiartry. San Diego: Academic Press, 2001.

[12] Odd SD, Suraj BT. Immigration, social integration and mental health in Norway, with focus on gender differences. Clinical Practice and Epidemiology in Mental Health 2007, 3:24

[13] Akiyama T. Onset study of English speaking temporary residents in japan. Soc Psych Psychiat Epid 1996; 31:194-198

[14] Hussain N, Greed F, Tamenson B. Adverse social circumstances and depression in people of Pakistani origin in the UK. Br J Psych 1977; 171: 434-438.

[15] Gee GC, Spencer M, Chen J, Yip T, Takeuchi DT. The association between self-reported racial discrimination and 12-month DSM-IV mental disorders among Asian Americans nationwide. Social Science & Medicine. 2007, 64:1984-1996.

[16] Gillam S, Jarman B, White P et al. Ethnic differences in consultation rates in urban general practice. BMJ 1989;299:953-958.

[17] Regier DA, Myers JK, Karmer M et al. The NIMH epidemiologic catchment area program. Arch Gen Psych 1984;41:934-941.

[18] Centers for Disease Control and Prevention. Behavioral risk factor surveillance system.2006: Behavioral risk factors surveillance system questionnaire. January 23, 2006: Question 2.2. 2006:6 Retrieved October 3, 2007, from http://www.cdc.gov/brfss/questionnaires/pdf-ques/2006brfss.pdf

[19] Andresen EM, Catlin TK, Wyrwich KW, Jackson-Thompson J. Retest reliability of surveillance questions on health related quality of life. J Epidemiol Community Health. 2003 May;57(5):339-43.

[20] Newschaffer CJ. Validation of behavioral risk factor surveillance system (BRFSS) HRQOL measures in a statewide sample. Atlanta, GA: Centers for Disease Control and Prevention; 1998.

[21] Dominick KL, Ahern FM, Gold CH, Heller DA. Relationship of health-related quality of life to health care utilization and mortality among older adults. Aging Clin Exp Res. 2002 Dec;14(6):499-508.

[22] Gandek B, Sinclair SJ, Kosinski M, Ware JE Jr. Psychometric evaluation of the SF-36 health survey in Medicare managed care. Health Care Financ Rev 2004;25(4):5-25.

[23] McHorney CA. Health status assessment methods for adults: past accomplishments and future directions. Annual Rev Public Health 1999; 20:309-35.

[24] Selim AJ, Rogers W, Fleishman JA, Qian SX, Fincke BG, Rothendler JA, Kazis LE. Updated U.S. population standard for the Veterans RAND 12-item Health Survey (VR-12). Qual Life Res. 2009;18(1):43-52.

[25] Rebecca R A, Roderick JAL. A review of Hot Deck imputation for survey non-response. International Statistical Review 2010; 78(1): 40-64.

[26] Cameron AC, Trivedi PK. Regression analysis of count data. Cambridge University Press. 1998.

[27] Stata Data Analysis Examples Poisson Regression. UCLA: Academic Technology Services, Statistical Consulting Group. from http://www.ats.ucla.edu/stat/stata/dae/poissonreg.htm (accessed May 16, 2012).

[28] Cook, John D. Notes on the Negative Binomial Distribution. Unknow, October 28, 2009 http://www.johndcook.com/negative_binomial.pdf

[29] Peter Schlattmann. Medical Applications of Finite Mixture Models. 2010; pp30-33 Springer, Berlin

[30] McLachlan, Georey J. and Peel, David. Finite mixture models. 2000. Wiley series

[31] Everitt, B. and D. Hand. Finite Mixture Distributions. Monographs on Applied Probability and Statistics. 1981. London, New York: Chapman and Hall.

[32] McLachlan, G. J. and K. E. Basford. Mixture models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs, 84. 1988. New York etc.: Marcel Dekker, Inc. xi, 253 p.

[33] Bohning, D. C.A.MAN-Computer Assisted Analysis of Mixtures and Applications. 1999. Chapman and Hall.

[34] Fruhwrith-Schnatter, S. Finite mixture and Markov switching models. Springer Series in Statistics. Berlin: Springer. xix, 492 p.

[35] Tatiana B, Didier C, David RH, Derek SY. Mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software. 2009; 32(6).

[36] Shi Zhong,Joydeep Ghosh. A unified framework for model-based clustering. The Journal of Machine Learning Research Volume 4. 2003 Mar: 1001-1037

[37] Akaike, Hirotugu. A new look at the statistical model identi

cation. IEEE Transactions on Automatic Control 1974; 19 (6): 716723.

[38] Schwarz, Gideon E. Estimating the dimension of a model", Annals of Statistics 1978; 6 (2): 461464.

[39] Morgan JA, Tatar JF .Calculation of the Residual Sum of Squares for all Possible Regressions. Technometrics 1972; 14, 317-325.

[40] Klein, JP, Moeschberger, ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer, 2003.