University of North Texas Health Science Center
UNTHSC Scholarly Repository

Theses and Dissertations

5-1-2010

Using Generalized Estimating Equations to Analyze Alcohol Consumption and Job Displacement among Older Workers

Nita Patel
University of North Texas Health Science Center at Fort Worth, patelnita@hotmail.com

Follow this and additional works at: http://digitalcommons.hsc.unt.edu/theses

Recommended Citation
Patel, N. , "Using Generalized Estimating Equations to Analyze Alcohol Consumption and Job Displacement among Older Workers" Fort Worth, Tx: University of North Texas Health Science Center; (2010).
http://digitalcommons.hsc.unt.edu/theses/91

This Dissertation is brought to you for free and open access by UNTHSC Scholarly Repository. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of UNTHSC Scholarly Repository. For more information, please contact Danelle.Orange@unthsc.edu.

Patel, N., Using Generalized Estimating Equations to Analyze Alcohol Consumption and Job Displacement among Older Workers. Doctor of Public Health (Biostatistics), May 2010, 94 pp, 9 tables, 2 illustrations, bibliography, 42 titles.

The objectives of this dissertation were to compare differences in alcohol consumption among the older workers (aged 51 to 61 years) who have experienced job displacement compared to those who remain continuously employed. Generalized estimating equations were used to model this relationship using longitudinal data from the Health and Retirement Study from 1992 to 2006. Approximately 39% of respondents had died during the study period. We analyzed four models. One model excluded data for deceased respondents. Another model retained data for deceased respondents. For the remaining two models, data was imputed using multiple imputation by chained equations. Data was imputed for only the predictors in one imputation, imputed for both the dependent variable and the predictors in the second imputation. All models were weighted and adjusted for key sociodemographic variables.

The results of this study show that being continuously employed, compared to experiencing job displacement, has a protective effect on the onset of alcohol consumption. Older workers who were not displaced were less likely to report consuming alcohol compared to those who had been displaced. This finding remained statistical significant even after adjusting for key sociodemographic variables. Complete case analysis and observed sample models provided biased estimates (i.e. wider confidence intervals, smaller *p* values) compared to the two multiple imputation models.

Our findings have important public health implications. Older workers are likely to have varied participation in the labor market. They are likely to be more experienced and hold senior or management positions, thereby earning higher wages. They may be at a higher risk of layoff during uncertain economic times, such as a recession. The effects of alcohol consumption among older individuals have been shown to be negative and particularly harmful, especially in terms of ethanol toxicity. Additional studies are needed to examine the health effects of late onset of drinking among older Americans.

USING GENERALIZED ESTIMATING EQUATIONS TO ANALYZE ALCOHOL CONSUMPTION AND JOB DISPLACEMENT AMONG OLDER WORKERS

Nita Patel, M.P.H.

APPROVED:

_____

Major Professor

_____

Committee Member

_____

Committee Member

_____

Department Chair

_____

Dean, School of Public Health

USING GENERALIZED ESTIMATING EQUATIONS TO ANALYZE ALCOHOL CONSUMPTION AND JOB DISPLACEMENT AMONG OLDER WORKERS

DISSERTATION

Presented to the School of Public Health

University of North Texas

Health Science Center at Fort Worth

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Public Health

By

Nita Patel, M.P.H.

Fort Worth, Texas

May 2010

ACKNOWLEDGMENTS

Thank you, Dr. Bae, for your support not only during my time as a student at UNTHSC, but even before that, when you provided guidance to me as a PHPS fellow. You used few words, but your words were full of life experience. I really appreciated having you as my adviser and mentor. Dr. Singh, thank you for bringing humor to biostatistics, a concentration that sometimes felt really daunting to me. You helped me see the value of balance – balancing life and work. Dr. Chen, you always had an open door policy. I always felt comfortable in coming to you when I had questions, even for classes you did not teach. Thank you. Finally, last but not least, Dr. Lykens, I will always be grateful to you for your HRSII class. Your knowledge and comfort in all things health policy and management really impressed me! I will always attribute my skills and knowledge of STATA to you. Thank you for showing me that whether it's learning new software or learning a new discipline (i.e. HMAP), it always gets easier with time and practice. I owe much gratitude to Dr. Biswas, Dr. Suzuki, and Alice for your help and support to me as a student.

TABLE OF CONTENTS

Page

LIST OF TABLES.....V

LIST OF FIGURES.....VI

LIST OF ABBREVIATIONS.....VII

Chapter

1. INTRODUCTION.....1

    Rationale
    Study Objectives

2. LITERATURE REVIEW.....5

    Job Instability, Health and Alcohol Consumption
    General Reviews of the Literature on Job Instability and Health
    Longitudinal Studies of Job Instability and Health among Older Workers
    Alcohol Consumption among the Elderly: Mixed Findings
    Alcohol Consumption among Older Workers
    Statistical Models for Longitudinal Data Analysis
    Traditional Models: Continuous Dependent Variable
    Current Approaches: Continuous and Discrete Dependent Variables
    Missing Data

3. METHODS.....29

    Background of the Health and Retirement Study
    Sampling Design
    Binary Dependent Variable
    Independent Variables
    Data Management
    Data Analysis

4. RESULTS.....43

Descriptive Analysis of Observed Data
Results of Missing Data Analysis
Descriptive Analysis of Key Variables among Complete Cases
Multiple Imputation Results
Determining the Working Correlation Structure
GEE Analysis of Complete Cases
GEE Analysis of Multiply Imputed Data

5. DISCUSSION.....59

6. BIBLIOGRAPHY.....63

APPENDICES.....77

A. Code for UVIS Implementation of ICE in STATA
B. STATA Output of Missing Data Analysis using MISSCHK Option
C. STATA Code for Imputation by Chained Equations (ICE)
D. Missing Value Imputation Diagnostics
E. STATA Output of GEE analysis of Complete Cases, Unadjusted and Adjusted
F. STATA Output of GEE analysis of Multiple Imputation, Unadjusted and Adjusted

LIST OF TABLES

Table *Page*

Table 1: Example of Patterns of Data Missingness in HRS.....24

Table 2: Missing Data Analysis of Unweighted Observed Sample:

HRS, 1992.....45

Table 3: Key Variables with Missing Data, Unweighted Observed Sample:

HRS, 1992 (n=4,334).....46

Table 4: Frequency of Alcohol Consumption and Displacement among

Complete Cases, 1992 - 2006 (n=2,633).....48

Table 5: Differences in Regression Coefficients and Standard Errors in Working

Correlation Structure by Type of Analysis.....54

Table 6: Results of GEE Analysis of Alcohol Consumption among

Displaced Workers, Complete Cases.....55

Table 7: Results of GEE Analysis of Alcohol Consumption among

Displaced Workers, Observed Sample.....56

Table 8: Results of GEE Analysis of Alcohol Consumption among

Displaced Workers, Multiple Imputation of Predictors.....57

Table 9: Results of GEE Analysis of Alcohol Consumption among

Displaced Workers, Multiple Imputation of all Variables.....58

LIST OF FIGURES

Figure *Page*

Figure 1 Health and Retirement Study. Overview of the Health and Retirement Study Surveys 29

Figure 2 Overview of Sample Size Selection ….....32

LIST OF ABBREVIATIONS

AHEAD	Assets and Health Dynamics of the Oldest Old
ANOVA	Analysis of variance
CDC	Centers for Disease Control and Prevention
CES-D	Center for Epidemiological Studies-Depression
GEE	Generalized estimating equations
GLIM	Generalized linear models
HRS	Health and Retirement Study
HU	Housing unit
ICE	Imputation using chained equations
MANOVA	Multivariate analysis of variance
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputation through chained equations
MLE	Maximum likelihood estimators
MNAR	Missing not at random
MSA	Metropolitan statistical area
NIA	National Institute on Aging
PPS	Probability proportionate to size

PSU	Primary sampling unit
SSA	Social Security Administration
SSU	Sampling of area segments

CHAPTER 1

INTRODUCTION

Deaths related to alcohol are increasing in the United States. According to the Centers of Disease Control and Prevention (CDC), there were approximately 75,766 deaths attributable to alcohol or 2.3 million years of potential life lost (about 30 years of life lost on average per alcohol attributed death) in 2001 (CDC, 2004). The effects of alcohol consumption on health are well documented. Alcohol is a major contributor in deaths due to unintentional injuries, including burns, drowning, falls firearm, and poisoning (Smith, Branas, & Miller, 1999). Alcohol abuse has been linked to intimate partner violence (Foran & O'Leary, 2008). Alcohol consumption has been linked to increased risk of cancers (including oral, esophagus and larynx) (Corrao, Bagnardi, Zambon, & La Vecchia, 2004). Alcohol disorders can induce anxiety disorders, and vice versa (Kushner, Abrams, & Borchardt, 2000).

Moderate alcohol use (≥ 1 drinks in past month to ≤ 2 drinks per day), however, has been found to have a protective effect on coronary heart disease (Elkind et al., 2006). Light (1 – 15 drinks per week) to moderate (15 – 42 drinks per week) alcohol use has a protective effect on type II diabetes (Wannamethee, Shaper, Perry, & Alberti, 2002). Changes in the way ethanol metabolizes, combined with other factors such as comorbidites and existing medications, make older adults more vulnerable to the effects of alcohol. The relationship between health, including behaviors such as alcohol consumption, and job instability is not clearly elucidated in published literature. The dynamics between health and employment are a complex and emerging research area that

are being examined by a variety of disciplines, including public health, sociology and economics. In the context of public health, the study of the relationship between health and employment falls under the purview of social epidemiology. Social epidemiology is the study of the social constructs of health. Current frameworks in social epidemiology that explain the distribution of diseases are psychological, social production of disease/political economy of health, and ecosocial and other emerging multilevel frameworks (Krieger, 2001).

Rationale

Does alcohol consumption lead to decreased participation in the job market, or are individuals experiencing job loss or facing career uncertainty at increased risk of or more susceptible to unhealthy behaviors, such as alcohol consumption? Two frameworks provide a contextual background in the study of the relationship between health and employment. These two frameworks are that of social causation and health selection. Both of these frameworks will be briefly discussed.

The systematic literature review conducted in the following section is inclusive of the effects of employment to broader aspects of health such as physical health, including chronic diseases, myocardial infarction and mental health. Few studies have been examined the relationship between employment and behaviors such as alcohol consumption and smoking. Findings from these studies will also be discussed. Research design is an important aspect of overall study design and hypothesis testing. A range of various research designs have to used to examine the relationship between health and employment. These include and are not limited to cross-sectional

and longitudinal studies. As its name implies, cross-sectional analysis examines a cross section of a population of interest. The longitudinal design, on the other hand, is the analysis of repeated measurements of individuals to asses change over time. Longitudinal data has advantages over the cross-sectional study design. In the cross-sectional design, causation can be erroneous because of issues of confounding (Mann, 2003). Longitudinal data allows the researcher to observe an individual from start to a designated endpoint. Also, each subject serves as his or her own control. Observations in longitudinal data are correlated and statistical techniques have to account for this correction (Diggle, Heagerty, Liang, & Zeger, 2002).

The literature review in this dissertation will focus primarily on longitudinal studies for reasons discussed above. Also, longitudinal data will be analyzed to test the objectives of this dissertation. Thus, a literature review of statistical techniques to analyze longitudinal data is also presented. Statistical methods for analyzing longitudinal data have evolved over time. Traditional methods, such as univariate analysis of variance (ANOVA) and multivariate ANOVA or MANOVA, are included in the literature review. Linear mixed effect

models have also been used to analyze longitudinal data. More recent work includes the broadening of classical regression modeling using different estimation methods for regression parameters. These approaches will also be reviewed. Techniques for determining the amount of missing data, as well as methods for handling missing data will be reviewed.

Study Objectives

The objectives of this dissertation are to compare differences in alcohol consumption among the older workers who have experienced job displacement compared to those who remain continuously employed. Generalized estimating equations will be used to model this relationship. The following literature review discusses issues of reverse causation, and potential confounders for the relationship between alcohol consumption and employment. Analyses will adjust for potential confounders.

Public access data from the Health and Retirement Study (HRS) will be used. HRS is a longitudinal study conducted by the University of Michigan specifically targeting Americans 50 years and older. It is an ongoing study, initiated in 1992, and administered every 2 years to approximately 22,000 adults. Types of information collected from HRS include physical and mental health, insurance coverage, financial status, family support systems, labor market status, and retirement planning. HRS is sponsored by the National Institute of Aging
(grant number NIA U01AG009740) and conducted by the University of Michigan.

CHAPTER 2

LITERATURE REVIEW

Job Instability, Health and Alcohol Consumption

There are at least two differing and conflicting hypotheses on the influence of health on labor force participation and vice versa. One such hypothesis is that of health selection, which proposes that health status determines labor force participation (Dooley & Prause, 2004). Those that are in poor health have limited participation in the labor force. The social causation hypothesis suggests that unemployment or underemployment attributes to poor health (Dooley & Prause, 2004). The majority of studies published in literature review presented in this section test the social causation hypothesis, or what is otherwise referred to as reverse causation. Many of the studies reviewed focus on individuals in their prime working age, and a few focus specifically on older workers.

General Reviews of the Literature on Job Instability and Health

The relationship between health and employment is a complex one. Several authors have attempted to review published literature to better understand this relationship. Two such reviews are presented here. These reviews differ from meta analysis in that meta analysis involves a statistical analysis of the studies reviewed, including determining the effect size and weight of each study in the meta analysis (Berman & Parker, 2002). Dooley, Fielding and Levi (1996) reviewed numerous studies to examine the relationship between health and unemployment. The studies in their review focused on many different aspects of health, including physical health, mental health, well-being or role functioning. One of finding

of their review was that statistical inference on the health effects of unemployment was inconsistent and depended on the research design of study reviewed. Dooley et al. (1996) distinguished between three study designs: aggregate, individual and cross-level. Aggregate level studies focused on the effect of the aggregate-level (i.e. community-level) unemployment on health outcomes. The authors discussed some inherent problems with this method, including ecological fallacy (making causal references from group data to the individual level) (Schwartz, 1994). Individual (cross-sectional) methods ran into the problem of reverse causation; however, longitudinal, or panel approaches, controlled for health status before individuals became unemployed, thus did not have the problem of reverse causation.

Dooley et al. (1996) found that unemployment was shown to increase risk factors for diseases such as alcohol and tobacco consumption, and unhealthy behaviors (i.e. diet, exercise), but further studies are needed. The authors, in their review, found that most of this research focused on the employed and unemployed. Few studies had been conducted on the underemployed, and how a change in employment status from employed to underemployed affected health (Dooley et al., 1996).

Another review of studies to examine the relationship between health and employment was conducted by Chirikos (1993). While the review conducted by Dooley et al. (1996) examined the relationship between health and employment from the standpoint of social causation (i.e. labor market affecting health outcomes), Chirikos' review provided insight from the health selection perspective (1993). Chirikos reviewed studies spanning two decades, considering U.S. studies published from the early 1970's,

and those that treated health as a determinant of labor market outcomes. Studies were distinguished based on methodological treatment of the variable, health. For example, studies in which health was an independent predictor in single-equation model of labor supply were classified as first-generation results. Those studies which modeled health (or some health attribute) as a dependent variable were classified as second-generation studies, and tended to have greater statistical complexity.

Chirikos' overall findings were that impaired health consistently changed labor market behavior (i.e. reduced wages, less time spent) of the individual, and/or other individuals in the household of the individual with impaired health. While the effect of health conditions on work effort had been well established, the magnitude of this health effect varied across studies. Some second-generation studies were now beginning to treat economic and health decision-making as interrelated or jointly determined. Chirikos' review highlighted the need for more research examining health differences across sociodemographic groups, stratified analysis of subgroups of growing workers (such as adult working women), for example.

Longitudinal Studies of Job Instability and Health among Older Workers

While the above section focused on two studies that reviewed the literature to examine either the health selection or the social causation hypothesis regarding health and employment, this section includes individual, longitudinal studies on health and employment among older workers. The relationship between involuntary job loss, job displacement and unemployment on various health conditions among the elderly has been highlighted recently (Gallo, Bradley, Siegel, & Kasl, 2000). Gallo et al. (2000) used data

from 1992 and 1994 to examine health consequences of involuntary job loss, specifically among older workers. Involuntary job loss was defined by the authors as having lost a job due to plant or company closing or layoff. Physical functioning was determined by an index created using self-reported responses including activities of daily living and mobility tasks. Mental health was determined by using a portion of the Center for Epidemiological Studies-Depression (CES-D) scale.

Compared to workers who were continuously employed, those who involuntarily lost their jobs were more likely to have poorer physical functioning and mental health. However, Gallo et al. (2000) found evidence that the relationship between involuntary job loss and health was a combination of health selection and social causation. This conclusion was based on their finding of an association between baseline physical functioning and involuntary job loss (support for the health selection hypothesis). However, after controlling for health and sociodemographic and economic variables, the effect of involuntary job loss on physical functioning remained statistically significant (support for the social causation hypotheses).

Another study examined the effect of involuntary job loss on myocardial infarction and stroke among late career workers over a 10 year period (Gallo et al., 2004). After controlling for known confounders and predictors, late career workers who lost their jobs involuntarily were more than twice as likely at risk of developing myocardial infarction and stroke compared to individuals who continued to work.

The odds of cigarette smoking relapse was more than twice as high among older workers who experienced involuntary job loss (defined as job loss due to plant closing or layoff) compared to those who did not experience involuntary job loss (Falba, Teng, Sindelar, & Gallo, 2005). Smoking relapse was defined as smoking among individuals who were currently not smoking, but had smoked previously. Smoking increased among older smokers who remained unemployed.

Another study examined whether involuntary job loss among older men affected the mental health of their wives' (Siegel, Bradley, Gallo, & Kasl, 2003). Using data from 1992, 1994 and 1996, this study used a portion of the Center for Epidemiological Studies-Depression (CES-D) scale to measure depressive symptoms. Findings from this study indicate that wives' mental health was not statistically effected by their husbands involuntary job loss. However, data from 1992-1994 indicated that women who were more financially satisfied at baseline did experience an adverse effect from their husbands' job loss.

A longitudinal study specifically testing the health selection hypothesis that individuals in poor health self-select themselves out of employment was conducted by McDonough and Amick (2001). The study followed individuals aged 25 to 61 at baseline from 1984 to 1990. All individuals were employed at baseline and followed until they either left the labor force or the end of the study. This study found that the effect of poor health on labor force exit was mixed, and depended on gender, race and education. Younger men and white men were most likely to leave the labor force due to perceived poor health. McDonough and Amick also found that women continued to

work in the labor force despite poor health (2001). Age did not increase or decrease the hazard of workers exiting the labor force due to poor health. McDonough and Amick's study suggests that the relationship between health and employment is not a clear cut one, and that health is affected by social and economic resources, such as availability of pension and retirement benefits (2001). The above studies have provided mixed evidence on whether poor health causes individuals to exit the labor force, or the reverse of instability in the labor force causing poor health. The evidence on the effects of involuntary job loss and risk of myocardial infarction, stroke and smoking are much clearer. As with smoking, does the stress of involuntarily losing a job, or being unemployed cause an increase in alcohol consumption among the elderly?

Alcohol Consumption among the Elderly: Mixed Findings

Recent studies indicate that alcohol consumption among those 65 and older is increasing. The way in which alcohol, specifically ethanol, is metabolized changes with age (Meier & Seitz, 2008; Seitz & Stickel, 2007). Some of the enzymes involved with ethanol metabolism, alcohol and acetaldehyde dehydrogenase and cytochrome P-4502E1, have reduced activity in the elderly. Also, the volume of water distribution in the body changes (reduces) with age. These two factors have a direct and negative consequence on the elderly because they increase blood concentrations of ethanol in a population already vulnerable to the toxic effects of alcohol (Meier & Seitz, 2008). Seitz and Stickel discuss other factors that increase ethanol toxicity among the elderly, including the use of multiple drugs/medications and the presence of other types of liver disease (2007). The authors argue that alcohol consumption, even at lower levels, is more toxic to the aging

population. Other studies demonstrate that the effects of alcohol consumption on the elderly are mixed and based on the amount of consumption.

The following studies indicate that while heavy alcohol consumption may be harmful, light to moderate alcohol consumption may have some benefits to older Americans. Thun et al. (1997) examined death rates among adults 30 years and older who self-reported drinking alcohol. Their findings were consistent with previous studies on the harmful effects of heavy alcohol consumption. Mortality associated with alcohol consumption included cirrhosis, cancers of the mouth,
esophagus, pharynx, larynx, and liver combined, breast cancer, and injuries. Light alcohol consumption was found to have some beneficial health effects. Mortality rates from all cardiovascular diseases were 30 to 40 percent lower among men and women who reported having at least one drink daily (i.e. light alcohol consumption) than among nondrinkers. However, among heavier drinkers and those at lower risk of cardiovascular disease, death rates from all causes increased. Death rates among heavy consumers of alcohol followed a J-shape curve for individuals at low risk of cardiovascular disease, U-shape curve for those at intermediate risk, and L-shape curve for those at high risk of cardiovascular disease (Thun et al., 1997). Moderate alcohol consumption (one or two drinks of alcohol daily) was associated with lower overall mortality rates than no alcohol consumption, but cigarette smoking almost doubled the risk of death among those 35 to 69 years of age (Thun et al., 1997).

Another study also found that heavy alcohol consumption (either eight drinks per occasion or getting drunk at least monthly) increased mortality risk among light to

moderate drinkers (Rehm, Greenfield, & Rogers, 2001). Moore et al. (2006) examined the influence of alcohol use and comorbidity on 20-year mortality in adults aged 60 to 74 in a longitudinal study. At-risk drinking was defined as drinking 2 to 3 drinks per day and having gout or anxiety or taking pain medication. Those who drank at-risk had higher mortality rates compared to those who did not, but only among males. Abstinence was not associated with higher mortality rates among both males and females (Moore et al., 2006).

A longitudinal study examining alcohol consumption and the risk of atrial fibrillation (abnormal heart rhythm) among adults 65 and older found that current moderate alcohol consumption was not associated with risk of atrial fibrillation or with risk of death after diagnosis of atrial fibrillation (Mukamal et al., 2007). Low to moderate alcohol consumption may have a protective effect against incidence of dementia and Alzheimer's Disease, according to a meta analysis of longitudinal studies of adults 65 and older worldwide (Peters, Peters, Warner, Beckett, & Bulpitt, 2008). Light or moderate alcohol consumption could be protective against total and ischemic stroke (Reynolds, Lewis, Nolen, Kinney, Sathya, & He, 2003). Mild to moderate drinking among older adults has also been shown to slow cognitive decline. A longitudinal study examining the relationship between alcohol consumption and cognitive decline among older adults (65 and older) who did have have dementia at baseline. This study found that compared to nondrinkers, mild (once a month or less) to moderate (more than once a month, averaging between daily and weekly) drinkers had slower decline in cognitive domains (Ganguli, Bilt, Saxton, Shen, & Dodge, 2005).

Higher levels of alcohol consumption among older men have been shown to result in decreasing the rates of fatal and nonfatal coronary heart disease but increasing the risk of fatal and nonfatal neoplasms. Higher levels of alcohol consumption among light to moderate drinkers placed them at increased risk of fatal and nonfatal strokes (Goldberg, Burchfiel, Reed, Wergowske, & Chiu, 1994).

Alcohol Consumption among Older Workers

Very few studies have been conducted to examine the relationship between alcohol consumption and career instability among older workers in the United States. Though a number of longitudinal studies were found examining alcohol consumption among younger populations, we found only a small handful of longitudinal studies examining this relationship among older workers.

Mullahy and Sindelar (1991) studied the effects of alcoholism on employment to examine any differences by gender, using data form the Epidemiologic Catchment Area study among workers aged 18 to 64. They found that alcoholism (defined as whether an individual has met criteria for alcohol dependence or alcohol abuse) was associated with a negative influence in labor force participation and income for both men and women.

French and Zarkin (1995) wanted to better understand the relationship between alcohol use and wages. Age of workers in this study ranged from 21 to 68 years. They analyzed data for over 1,000 employees, randomly selected from four worksites and found that there was an inverse U-shaped relationship between alcohol consumption and wages. Moderate alcohol users (about 2 drinks per day on average) who were working had higher wages than abstainers and heavy drinkers. Their findings suggest that alcohol

use is related to wages through human capital variables, such as education, job tenure and health status.

Mullahy and Sindelar used data from the the Alcohol Supplement of the 1988 National Health Interview Survey (NHIS) to study problem drinking, employment, and unemployment among individuals 25 to 49 years old (1996). Problem drinking is defined by the authors to encompass heavy drinking, and the diagnosis of alcohol-related disorders, alcohol abuse and/or alcohol dependence. This cross-sectional analysis found that problem drinking reduced employment and increased unemployment (Mullahy & Sindelar, 1996).

Gallo, Bradley, Siegel, & Kasl (2001) examined the effects of involuntary job loss on alcohol consumption on the elderly (age range from 51 to 61 years). They used longitudinal data from 1992 and 1994 and defined involuntary job loss as job loss due to plant closing or layoff. They found that there was a significant association between job loss and alcohol use. Baseline nondrinkers were twice as likely to consume some amount of alcohol at follow up compared to those with continuous employment.

A longitudinal study examined the relationship between positive alcohol expectancies among older workers employed in either construction, manufacturing or transportation (Bacharach, Bamberger, Sonnenstuhl, & Vashdi, 2008). All individuals were retirement eligible as defined by their respective unions. Age ranged from 43 to 70 years. Positive alcohol expectancies were defined as positive beliefs about the effects of alcohol on behavior, moods and emotions. This study found that positive alcohol expectancies had a moderating effect on aging and drinking. High

expectancies were significantly associated with drinking problems, and vice versa. For individuals eligible for retirement but continued to work, the moderating effects of positive alcohol expectancies on aging and drinking were amplified. These studies highlight the fact that the effects alcohol vary by the amount of consumption among older workers, and those worker experiencing job instability may be at increased risk of consuming alcohol.

Statistical Models for Longitudinal Data Analysis

Methods for analyzing longitudinal data have expanded greatly over the past three decades. However, there still exists a gap in theoretical advances in statistical methods and their practical application. Historically, this had been in part due to a lack of statistical computing techniques that facilitated implementation of new statistical methods. Recent advances in computing and statistical programs has helped reduce this gap.

Longitudinal data analysis has many unique methodological aspects which require careful statistical consideration. Longitudinal data is a type of repeated measures data. Repeated measures data is data for a response variable on the same individual collected at different times. When data for a response variable on the same individual is collected multiple times, we have clustered data. When repeated measures data are collected on the same individual over a period of time,

we have longitudinal data (Klienbaum, Kupper, Muller, & Nizam, 1998). Longitudinal analysis can be conducted both prospectively or retrospectively (Diggle et al, 2002). Longitudinal studies have many advantages over cross-sectional studies (Kramer, 1983).

The longitudinal design allows both the exposure and outcome variables to be collected over a period of time. Since data are collected on individuals, rather than groups, it is ideal for studying individual change over time. Finally, the longitudinal design allows for more efficient estimation of within-subject covariates or time dependent covariates, such as age. However, this design is less efficient in estimating between-subject covariates (i.e. characteristics of a subject that do not change over time) (Ware, 1985).

Traditional Models: Continuous Dependent Variable

Many traditional methods for analyzing longitudinal data were based on ANOVA models (Fitzmaurice, Davidian, Verbeke, & Molenberghs, 2008). Repeated-measures ANOVA, like ANOVA, tests for equality of group means. In addition, repeated-measures ANOVA tests for equality of group means across repeated measurements of time. Time is the within-subject factor because different measurements are taken at different times on the same subject. Treatment is usually the between-subject factor because levels of treatment can change between subjects. Univariate repeated-measures ANOVA models for a single response variable, and allows for a positive correlation of the repeated measures on the same individual by including a random subject effect. Mathematically, the notation for repeated measures ANOVA is as follows:

$$Y_{ijk} = \mu + \tau_j + b_{jk} + \Upsilon_k + (\tau\Upsilon)_{jk} + e_{ijk}$$

where μ = grand mean
τ_j = effect of treatment on response variable
Υ_k = time effect
$\tau\Upsilon_{jk}$ = interaction effect of group *j* on time *k*
e_{ijk} = error term

MANOVA models for more than one response variable. MANOVA is expressed mathematically as follows:

$$\mathbf{Y}_{ij} = \boldsymbol{\mu} + \Upsilon_i + \mathbf{e}_{ij}$$

where $\boldsymbol{\mu} = n \times 1$ vector mean for time

$\Upsilon_i = n \times 1$ vector effect for population from with *i*th group of subjects was drawn

$\mathbf{e} = n \times 1$ vector of errors $N(\mathbf{0}, \boldsymbol{\Sigma})$ in each of the populations

Univariate ANOVA and MANOVA models have assumptions. Both the univariate and multivariate ANOVA models require that all units (i.e. individuals, observations) be observed at the same n "time" points, or an assumption of a balanced design. Not only must each data vector Y_i be of the same length, n, for all units, but each element Y_{ij} , j = 1,...,n must have been observed at the same number of times $t_1,. . . ,t_n$. Thus, the time intervals between repeated observations have to be equal (Fitzmaurice et al., 2008).

Univariate and multivariate ANOVA procedures assume that the variance covariance matrix ($\boldsymbol{\Sigma}$) of each data vector Y_i, i = 1, . . . ,m is the same for all i. If this assumption is believed to be reasonable, and $\boldsymbol{\Sigma}$ is a common (n x n) covariance matrix, the structure of $\boldsymbol{\Sigma}$ has to be examined. The univariate model assumes $\boldsymbol{\Sigma}$ to have compound symmetry. This means that there is a very specific pattern of correlation among observations taken on the same unit at different times. The correlation among all observations on a given unit is the same regardless of how near or far apart the observations are taken in time. Thus, the univariate methods are based on an assumption about the covariance structure that may be too restrictive if within-unit sources of correlation are not negligible (Kaplan, 2004). The multivariate model does not

have any assumptions about the structure of Σ. MANOVA does not take into account at all the way in which observations arise in

the longitudinal setting. It is likely that this assumption for the covariance structure is too general (Kaplan, 2004). Neither model explicitly incorporates time in the model (Diggle et al., 2002).

Current Approaches: Continuous and Discrete Dependent Variables

Among the restrictions of univariate ANOVA and MANOVA is independence. The inherent nature of longitudinal data (i.e. repeated observations on the *same* individual) violates the assumption of independence, because typically, observations closer together in time are more correlated than distant observations. Regression models have been expanded over the past 30 to 40 years to accommodate some of these restrictions, including accounting for correlated observations and longitudinal data. Hartley and Rao applied maximum likelihood methods to the mixed effects model (1967) to estimate variance components. They developed a set of equations from which specific estimates could be obtained by an

iterative (repeated) process. The method of maximum likelihood was developed by R. A. Fischer, and involves examining sample values and selecting those values that maximize the probability or probability density of obtaining sample values. Fischer was able to demonstrate that maximum likelihood estimators (MLE) were asymptotically minimum variance unbiased estimators (Miller & Miller, 2004). In other words, he was able to demonstrate that as the sample size increased, MLE were those with the lowest variance.

A commonly used mixed effects linear model for continuous response

variables was proposed by Laird and Ware (1982). They proposed a family of two-stage models for repeated measurements when the outcome variable is approximately normal. Their work, based on Harville (1977), included a unified approach using growth models and repeated-measures models. Both MLE and empirical Bayes estimation were discussed. Repeated measurements on each individual are assumed to follow a regression model with distinct regression parameters for each individual. The distribution of these individual subject parameters is modeled in the second stage as random effects to account for between-subject variation. Correlated continuous outcomes are treated as fixed effects. Additional models for longitudinal data analysis have been proposed, including one by Zhang and Davidian for a semi-nonparametric linear mixed model (2001).

Mixed effects linear models have less restrictive assumptions than repeated-measures ANOVA. The variance covariance structure does not have to be independent. Data can be unbalanced and repeated measurements do not have to have equal time intervals. The model can accommodate missing data if the missing data is missing at random (MAR). Some of increased popularity of mixed effect linear models is due to improvements in statistical computing. An example is the availability of programs that can fit multilevel models (Singer, 1998).

Generalized linear models (GLIM) are likelihood based models, proposed by by Nelder and Wedderburn that expanded upon classical linear models to include the normal, Bernoulli (discrete), Poisson (discrete) and gamma distributions (1972). Classical linear models had many restrictive assumptions that could not be

applied to longitudinal data. For example, classical linear models assume that the error term is normally distributed. However, there are many instances in which the error term in continuous data is not normally distributed. Linear model assumptions were also not met when data was count or categorical data. GLIM expanded upon the classical linear model to be more inclusive of probabilistic distributions, including continuous and discrete response variables. GLIM specifies a more indirect relationship between the predicted mean and predictor variables. A link function relating the linear predictor to the predicted mean of the response variable is chosen. Also, a function defining the error terms probability distribution is chosen.

Liang and Zeger expanded GLIM for longitudinal data analysis (1986). GLIM assumptions require a multivariate normal distribution for continuous outcome variables. GLIM can handle data missing completely at random (MCAR) or missing at random (MAR). There is a distinction between data that is MAR or MCAR. When the missing data is independent of both the observed data and the missing data, it is considered MCAR. When the data that is missing is only independent of the missing data, it is MAR (Rubin, 1976). Approaches to missing data in longitudinal data analysis are discussed in the following section.

GLIM have been extended into different classes of regression models for longitudinal data analysis. Some of these classes include: (i) marginal or population averaged models, (ii) random effects or subject-specific models, and (iii) transition or response conditional models. Some of the key differences in these models is how the

model accounts for the correlation among repeated measurements, and how the regression parameters are interpreted.

Marginal models or population-average models, are an extension of the generalized linear models. They are relevant when the main focus of a study is investigating the effect of covariates on the population mean and not necessarily at individual level (Zeger, Liang, & Albert, 1988). Marginal models are considered more flexible than classic generalized linear models since they can handle unbalanced longitudinal data with repeated measurements and therefore they can handle as well some patterns of missing data. Marginal models do not require precise specification of the outcome distribution.

In the marginal models, the estimation of the parameters is performed using generalized estimating equations (GEE). GEE is an extension of simple linear regression. It accounts for repeated measures and correlated responses resulting in longitudinal data. It is an equation obtained by generalizing another estimating equation. Liang and Zeger introduced it as a method for calculating consistent estimates of regression parameters and their variances under weak assumptions about the joint distribution (1986). Their method for parameter estimation was based on the quasi-likelihood method (Wedderburn, 1974; McCullagh, 1983) in which only the mean and covariance structure is specified (Zeger & Liang, 1986). Mathematically, the marginal response of y is noted as:

$\mu_{ij} = E(y_{ij})$ has a link function to linear covariates $g(\mu_{ij}) = x_{ij}\beta$

where y_{ij} = response of subject *i* at time *j*
$x_{ij} = (x_{ij1}, \ldots x_{ijp})$ corresponding 1x *p* vector of covariates
$\beta = (\beta_1, \ldots \beta_p)'$ *p* x 1 vector of unknown parameters
g(.) = known link function

The variance of y is given by $Var(y_{ij}) = \Phi V(\mu_{ij})$ where Φ is a dispersion parameter that is known or can be estimated. For binary response variables, the variance function is as follows:

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \text{ and } V(\mu_{ij}) = 1 - \mu_{ij},\ \Phi = 1$$

Correlation among repeated measurements are considered nuisance parameters (Liang & Zeger, 1986). This correlation must be taken into consideration to obtain correct parameter estimators. An important step in choosing a specific correlation structure is to find the simplest structure which fits the observed data well. Correlation structures differ in their assumptions. The following are common correlation structures:

- Autoregressive: Observations taken closer in time are more correlated than the observations taken further apart for the same individual. Shults et al. (2009) showed that this correlation structure is appropriate for binary longitudinal data.
- Exchangeable: Every observation for an individual is equally correlated with every other observation for the same individual. The intraclass correlation coefficient provides a measure of this degree of correlation (Johnston & Strokes, 1997).
- Independent: Observations for a given individual are uncorrelated with every other observation for the same individual.
- Unstructured: No assumptions are made about the correlation coefficients between any two pairs of observations.

- User fixed: Correlation coefficients are fixed by the user rather than being estimated from the data. These values are fixed before the analysis.

GEE estimators are robust to departures from the true correlation patterns. However, selecting an appropriate working correlation structure can aid in making correct statistical inferences (Shults et al., 2009). Barnhart and Williamson have proposed goodness-of-fit test for modeling binary response variables using GEE (1998). Statistical software such as SAS, STATA, SUDAAN, and S-Plus, are well suited to fitting GEE regression models (Horton & Lipsitz, 1999).

Random effects models can also be used to analyze longitudinal data with binary outcomes. Random effects models can provide estimates of regression coefficients that are specific to a subject. They can also be used when the response variable is categorical (Stiratelli, Laird, & Ware, 1984). Although both GEE and random effects models can be used to analyze longitudinal data with a binary response variable, each method provides different standard errors of covariates (Kuchibhatla & Fillenbaum, 2003). The GEE methodology has been applied to analyzing repeated nominal and ordinal data by using the correlation coefficient as the measure of association (Miller, Davis, & Landis, 1993; Lipsitz, Kim, & Zhao, 1994). The odds ratio can also be used as a measure of association for analyzing ordinal data using GEE (Williamson, Kim, & Lipsitz, 1995).

Missing Data

Data can be missing in longitudinal studies for a number of reasons. These include participants who are lost to follow-up due to withdrawal, drop out, death or

adverse events. When data is collected over multiple points in times or waves, such as in the HRS, there may be complete data for a respondent for one Wave, and incomplete data for subsequent waves. For example, Table 1 illustrates that individual 1 (denoted by HRS unique identifier, HHIDPN) has complete information for the alcohol consumption variable for all 6 waves. Individual 2, however, has intermittent information (i.e. information for Waves 1, 3,4 and 6). Individual 3 has information for Waves 1 through 3, and appears to be lost to follow-up for the remaining Waves. Individual 4 has information for Wave 1, appears to be lost to follow-up for Waves 2 and 3, and has returned for the remainder of the study. Thus, patterns of missing data vary.

Table 1: Example of Patterns of Data Missingness in HRS

	Alcohol Consumption (Wave 1)	Alcohol Consumption (Wave 2)	Alcohol Consumption (Wave 3)	Alcohol Consumption (Wave 4)	Alcohol Consumption (Wave 5)	Alcohol Consumption (Wave 6)
HHIDPN 1	x	x	x	x	x	x
HHIDPN 2	x		x	x		x
HHIDPN 3	x	x	x			
HHIDPN 4	x			x	x	x

Mechanisms of missing data have been described by Rubin in terms of probability as indicated below (1976).

$\mathbf{y}$ = complete data matrix
$\mathbf{y}^{observed}$ = observed part of $\mathbf{y}$
$\mathbf{y}^{missing}$ = missing part of $\mathbf{y}$
$\mathbf{R}$ = missing data matrix

- MAR: $P(\mathbf{R}|\mathbf{y}) = P(\mathbf{R}|\mathbf{y}^{observed})$
- MCAR: $P(\mathbf{R}|\mathbf{y}) = P(\mathbf{R})$ for all $\mathbf{y}$

- MNAR: P(**R**|**y**) depends on $\mathbf{y}^{\mathbf{missing}}$

Data that is MAR is only dependent on the observed values. Data MNAR *is* dependent upon unobserved values, and therefore, is typically the worst kind of missing data. Because MCAR data is not dependent on either observed or unobserved values, data can be estimated for missing values for a dataset with missing values that are MCAR. Data can be estimated regardless of the patterns of missingness because of the assumption with MCAR that cases with missing data are a random sample of all the cases (Graham, 2009).

There are several techniques to address missing data analysis. A traditional technique, but one that is typically not recommended is complete case analysis or listwise deletion. In type of analysis, only observations with complete data (i.e. comple cases) are analyzed. Incomplete observations are excluded from analysis. Excluding observations with missing data can significantly impact a study. This type of analysis results in biased estimators, decreased power in the study and a loss of precision because incomplete observations are simply excluded from analysis (van der Heijden, Donders, Stijnen, & Moons, 2006). Another bias in
complete case analysis is introduced for data missing because of death. In this instance, complete case analysis is essentially analysis of the healthiest individuals, if the cause of death among the deceased is ill health (Diehr, Johnson, Patrick, & Psaty, 2005.

Another traditional technique for missing data analysis include single imputation by taking the average of observed values (i.e. mean imputation), or last observation carried forth, but again, these techniques are generally not recommended (Baraldi &

Enders, 2010). Last observation carried forth, especially in application to binary longitudinal data analyzed by GEE is not recommended (Cook, Zeng, & Yi, 1999).

Multiple imputation and MLE based techniques are current popular techniques for missing data analysis. Both require missing data to be MAR and multivariate normal (Baraldi & Enders, 2010). Unlike single imputation techniques, multiple imputation methods generate several imputed data sets using observed data that is assumed to be MAR. The number of imputations is specified by the analyst. Analysis is conducted on each imputed data set to produce parameter estimates and standard errors. Each of these separate data sets can be combined to produce one set of parameter estimates and standard errors (Donders et al., 2006). Single imputation tend to underestimate standard errors. This is not the case in multiple imputations, which correctly estimate standard errors and confidence intervals (Donders et al., 2006).

Maximum likelihood based methods to analyze missing data differ from multiple imputation in that this method estimates parameters by the value that maximizes the likelihood of the sample. Other imputation techniques not discussed at length here include Bayesian simulation method (Little, 1995), and inverse probability weighting methods (Philipson et al., 2008). Techniques such as inverse probability weighting has been applied to marginal models estimated through GEE (Hogan et al., 2004). GEE analysis assumes that missing data is MCAR. Paik used both mean imputation and multiple imputation to simulate bivariate and multivariate missing data not MCAR in GEE analysis (1997).

Another technique for imputing missing values is multivariate imputation by chained equations or MICE (van Buuren, Boshuizen, & Knook, 1999). The MICE method is based on the conditional distribution of the missing data given the rest of the data. MICE imputes a single predictor and cycles it over all predictors with missing values. A regression model is specified regarding conditional distribution of the missing data. Incomplete data for binary variables can be imputed by specifying logistic regression, polytomous regression can be specified for categorical data and linear regression for continuous data (van Buuren & Oudshoorn, 2000). Much of the statistical software today is capable of implementing statistical techniques for imputing missing data (Horton & Kleinman, 2007). To name a few, SAS, STATA, R, S-PLUS are all capable of implementing all of the missing data methods discussed here.

Imputation through chained equations or **ICE** is a user written multiple imputation program for STATA by Royston (Royston, 2005). **ICE** works in conjunction with UVIS (univariate imputation sampling) to impute missing values that are MAR (see Appendix A for UVIS code written by Royston and provided by UCLA, Academic Technical Services). A concern regarding using chained equations in imputing missing data is that compared to other methods, there may be need for additional theoretical justification of this technique (Stuart, Azur, Frangakis, & Leaf, 2009). A multivariate distribution is assume, but there is no assumed multivariate joint distribution. However, the ability of MICE to be model each variable based on its distribution adds to its appeal. In practice, a few studies have shown that MICE can produce unbiased estimators, including a study by Ambler and colleagues (Ambler, Omar, & Royston, 2007).

Missing data can be imputed using any of the techniques discussed above, but can data missing due to a respondent dying be imputed? There are different perspectives in the literature regarding imputing data for deceased individuals. Revicki and colleagues imputed missing data for physical health status scores among deceased individuals in a simulation study to examine bias (2001). When the amount of data missing was small (i.e. <15%) and mortality rates comparable among groups, all imputation techniques provided consistent and unbiased estimates. However, their findings are only generalizable to physical health and functioning (Revicki et al., 2001).

CHAPTER 3

METHODS

Background of the Health and Retirement Study

The Health and Retirement Study (HRS), initiated in 1992, is an ongoing, prospective, longitudinal study of a cohort of males and females aged 51 – 61 years and their spouses, focusing on economic and demographic issues (Burkhauser & Gertler, 1995; Anderson, 2008). HRS surveys labor force participation and pensions; health conditions and health status; family structure and transfers; and economic status. The survey originated as a social science survey to study economic and health factors among the elderly as they move into retirement (Juster & Suzman, 1995). A variety of cohorts have been sampled since the surveys' inception (see Figure 1).

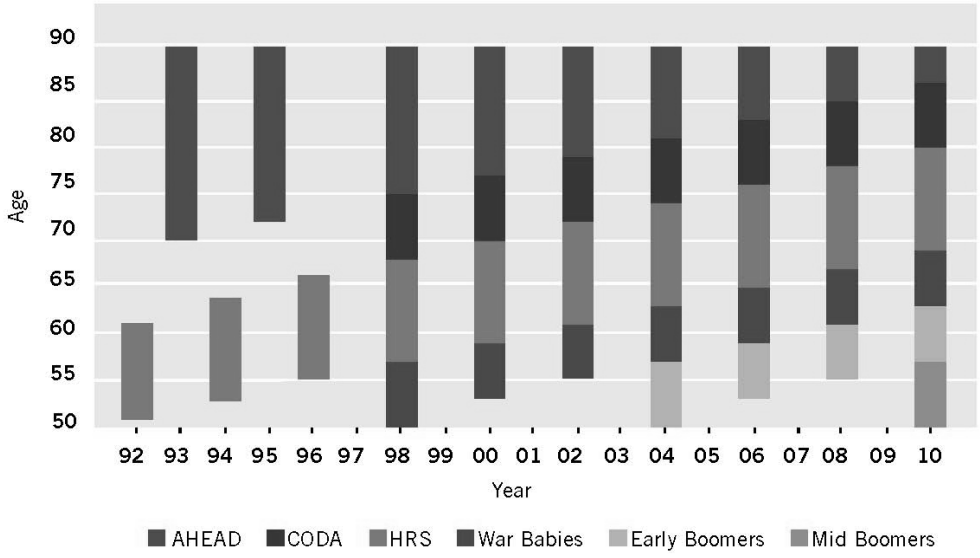

Figure 1. Health and Retirement Study. (n.d.). Overview of the Health and Retirement Study Surveys. Retrieved May, 2009 from http://hrsonline.isr.umich.edu/sitedocs/databook/HRS_Text_WEB_intro.pdf.

Blacks and Hispanics were oversampled (2:1 compared to Whites) because factors effecting their health and retirement were thought to differ significantly than Whites. Similarly, residents of Florida were oversampled due to a higher density of the

aging population in the State. Women, typically underrepresented in retirement studies, were also oversampled. Individuals in prisons, jails, nursing homes, long-term or dependent care facilities are excluded from the sample. HRS questionnaires are available in Spanish, and administered by bilingual interviewers when required. Individuals who do not speak English or Spanish, and for whom proxy informants are not obtained, are recorded as non-respondents and dropped from HRS. Exit interviews are conducted for each deceased respondent. Data for HRS is collected in waves, beginning with Wave 1 (1992-94), with a new Wave

conducted every 2 years. Analysis for this dissertation includes eight waves of data (Wave 1 (1992-93), Wave 2 (1994-95), Wave 3 (1996-97), Wave 4 (1998-99), Wave 5 (2000-01), Wave 6 (2002-03), Wave 7 (2004-05), and Wave 8 (2006-07)).

Sampling Design

There were 12,652 respondents in Wave 1 of HRS (baseline) and 11,596 at 1994 followup. Of the 12,652 respondents, 4,334 were eligible for inclusion in the study sample (see Figure 2). These respondents were 51 to 61 years old in 1992 and at risk of job displacement. Workers at risk of job displacement were defined as those who were working in 1992, not self-employed, and had been working for the same employer for three or more years. The three-year job tenure is comparable to that used the Displaced Worker Survey. The U.S. Department of Labor conducts the Displaced Worker Survey, a supplement to the Current Population Survey and defines displaced workers as those workers who were laid off due to a business or plant closing in which they were worked for three or more years. (Hipple, 1999). The three-year tenure has also been used in

similar research studies on job displacement (Couch, 1998). The final sample was comprised of 4,334 respondents.

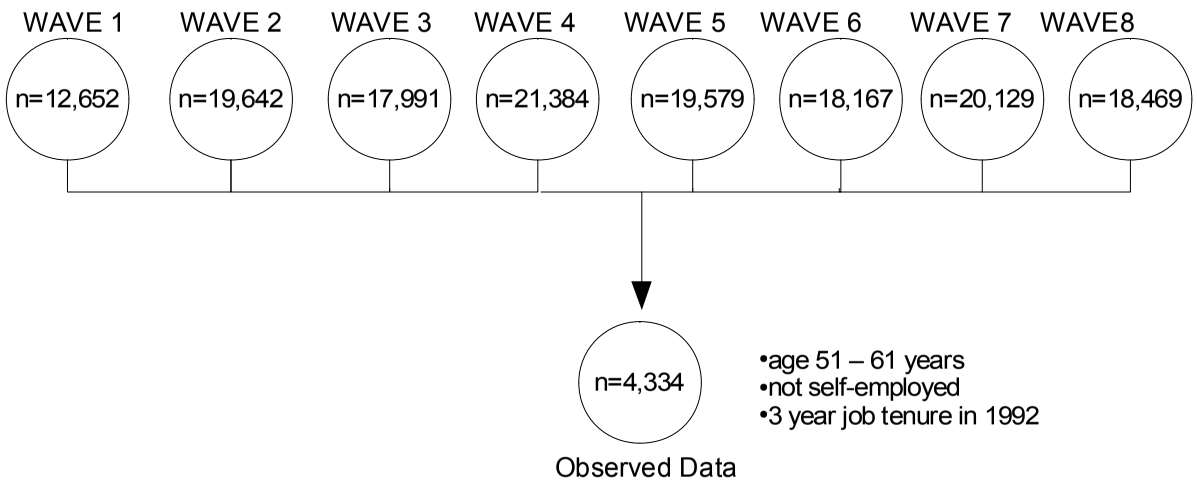

Figure 2. Overview of sample size selection.

HRS uses a multistage, stratified, probability sampling design. Heeringa and Connor (1995) detail the technical aspects of HRS survey sample design. The observational unit for HRS is the household financial unit which includes at least one eligible member born during 1931 to 1941. This age eligible individual can be single and unmarried, part of a married couple where either one or both individuals are age eligible. HRS is a probability sample involving four stages. The first stage is the probability proportionate to size (PPS) selection of U.S. Metropolitan Statistical Areas (MSAs) and non-MSA counties. In the second stage, sampling of area segments (SSUs) takes place within the sampled primary sampling units (PSUs). In the third stage, all housing units (HUs) within the selected SSU are enumerated and systematically selected. The fourth stage involves selecting the household financial unit within a sample HU.

HRS's design and coverage has changed since its inception in 1992 (Hauser & Willis, 2005). Assets and Health Dynamics of the Oldest Old (AHEAD) was added in 1993 to include cohorts born between 1890 to 1923, aged 70 and older. HRS and AHEAD questionnaires were combined in 1995. HRS data has to be weighted to account for its the sampling design, especially the oversampling of Blacks, Hispanics and households in Florida. The following section describes how HRS data is weighted (using weight variables developed for HRS, Wave 1) (Heeringa & Connor, 1995). Household and person level weights are provided. Household analysis weight is formulated using the HU selection weight, an adjustment factor for non-listed segments, an adjustment factor for sub-sampled areas, a household non-response adjustment factor, and a household post-stratification factor. Person-level analysis weight is composed of the

household analysis weight, the respondent selection weight and a person level post-stratification factor. HRS household selection weight is a relative weight value designed to be used with statistical software that support weighted estimation and data analysis.

The complex survey sampling design of HRS, particularly oversampling, stratification and clustering, must be taken into account in the analysis phase to provide correct standard errors and variance estimates. HRS provides two variables (STRATUM and SECU) to enable either Taylor Series or Replicated estimation of sampling errors (Heeringa & Connor, 1995). STRATUM, SECU and person-level analysis weights were used in this analysis.

Binary Dependent Variable

The dependent variable, alcohol consumption, was measured by the following question asked in Wave 1 of HRS, "Do you ever any alcoholic beverages, such as beer, wine, or liquor?" Respondents who answered "yes" were asked, "In general, do you have less than one drink a day, one to two drinks a day, three or four drinks a day, or five or more drinks a day?" Baseline drinking status (Wave 1) was coded as a binary variable. Subsequent alcohol consumption (Waves 2 – 6) was also assessed. The amount of alcohol consumed (i.e. number of drinks consumed among those who consumed alcohol) was coded as an ordinal variable using HRS created categories (stated above). There are cross wave differences in the question assessing current amount of alcohol consumption. Waves 1 and 2 of HRS recorded the amount of alcohol consumed in an ordinal variable as noted above. Subsequent waves (1996 to 2006) recorded the actual number of drinks consumed. Because alcohol consumption was initially an

ordinal variable and transitioned to a continuous variable from Wave 3 onwards, the amount of alcohol consumed could not be coded as a continuous variable without losing the ordinal data from Waves 1 and 2.

Among the variables adjusting for potential confounding in this study include individuals with a history of problem drinking. HRS asked several questions from the CAGE questionnaire which enabled discerning problem drinkers. The CAGE questionnaire is an easy to use, validated tool to assess alcohol dependence (Dhalla & Kopec, 2007; O'Brien, 2008). It was developed by Ewing in 1985 and is comprised of four questions: 1) Have you ever felt that you ought to cut down on your drinking?; 2) Have people annoyed you by criticizing your drinking?; 3) Have you ever felt bad or guilty about your drinking?, and 4) Have you ever had a drink first thing in the morning to steady your nerves or to get rid of a hangover (eye-opener)? (Ewing, 1985). Typically, a score of 1 is given for each positive response (for a total score of 4). Scores of two or more are recommended to detect alcohol abuse or dependence, and provide high sensitivity, specificity, and positive predictive value (Dhalla & Kopec, 2007). In addition to the two questions regarding alcohol, HRS also asks the four CAGE questions above. Respondents were assessed for a history of alcohol abuse or dependence using a cutoff score of 2 or more positive responses to any of the four CAGE questions. A variable comprised of respondents who were problem drinkers was created.

Independent Variables

There are a number of employment related questions available for analysis. Multiple variables will be used. Seasonal workers and those employed temporarily have

typically been excluded from employment studies. This analysis will also exclude these two groups of workers. Seasonal employment status will be assessed from a question (derived by RAND) as to whether the respondent works for someone else, or is self-employed in their current job. Labor force status of each respondent will be determined by a variable noting whether the respondent is working full-time, part-time, unemployed, partly retired, retired, disabled, or not in the labor force.

The definition of job instability in the literature varies. It ranges from job instability defined as job tenure (Jaeger & Stevens, 1999), job turnover (Gottschalk & Moffitt, 1999), and job change or separation from employer (both voluntary and involuntary) (Bernhardt, Morris, Handcock, & Scott, 1999). He, Colantonio and Marshall (2006) conducted a longitudinal study on career instability and long term health conditions among older Canadian workers and measured job instability by the number of jobless spells, number of weeks unemployed and the number of weeks not in the labor force. Job instability in this dissertation will be measured using the following questions:

- Is the respondent currently working for pay?
- Is the respondent working full-time, part-time, unemployed, partly retired, retired, disabled, or not in the labor force?
- Respondent's years of tenure in current job.
- Respondent's last month and year worked.

Unless otherwise indicated, HRS imputed variables will be used for sociodemographic variables such as age, gender, marital status, income and education. Age, provided as a continuous variable, will be categorized. Blue collar workers will be

classified as such if they worked in the following industries: farming, forestry, fishing, production and operations, and military. White collar workers were those who reported working in managerial or professional positions, sales, clerical, administration, or service industries. The marital status category was collapsed from 3 categories: married, single, and divorced/widowed/separated into 2 (i.e. marrried; single/divorced/widowed/separated). Only 10 respondents in the unweighted data (n=4,334) were single.

Baseline mental health will be determined by adapting the Center for Epidemiological Studies-Depression (CES-D) scale (Radloff, 1977). CES-D is a validated, 20 question scale, using a Likert scale. HRS differs from the CES-D scale in several key respects. For the baseline year, 1992, HRS only uses 10 of the 20 CES-D questions. These questions indicate whether the respondent, during the past week, felt depressed, everything he/she did was an effort, sleep was restless, was happy, felt lonely, enjoyed life, felt sad, felt that people disliked him/her, could not "get going," did not feel like eating (his/her appetite was poor). In 1994, only 8 of the 20 questions of the CES-D scale were included in the HRS questionnaire and responses were dichotomized to whether the respondent experienced an event "most" of the time. Due to this cross-wave difference, a 10 point scale will be determined using only 1992 data to create a baseline mental health score.

Data Management

RAND HRS data will be used (Source: http://hrsonline.isr.umich.edu/). RAND provides two types of data to researchers: RAND HRS Data file (v.1) and RAND-

Enhanced HRS Fat files. The former is available for download on RAND's website, while the latter is available by request on CD-ROM. The RAND HRS Data file (v.1) is a cleaned version of the HRS. It contains derived variables that cover a broad but not complete range of measures from the original HRS. Data includes imputations for income, assets, and medical expenditures developed at RAND. The development and continued maintenance of the RAND HRS Data are supported by the National Institute on Aging (NIA) and the Social Security Administration (SSA). RAND also provides RAND-Enhanced HRS Fat files. These provide raw variables in a single line per wave. RAND HRS Data file (v.1) does not include HRS drinking variables based on the CAGE questionnaire. RAND-Enhanced HRS Fat files do contain all of the variables regarding the CAGE questionnaire. Due to this reason, only RAND-Enhanced HRS Fat files will be used. Analysis was conducted using statistical software such as SAS or STATA (StataCorp LP., College Station, TX). Eight waves of data, or 12 years of longitudinal data were analyzed. The variable HHIDPN was used to track each individual through the eight waves. HHIDPN is a unique identifier developed by RAND, comprised of HRS identifiers.

Before GEE analysis was conducted, several data management steps took place in SAS. RAND-Enhanced HRS Fat Files are provided by wave; therefore, each wave will be merged by HHIDPN. Files were merged using a one-to-one merge option. Data were checked for errors using common PROC (procedure) statements in SAS: means, univariate, freq, print, contents, sort, and format. All variables that were either recoded or newly created were verified by creating a r x c table of old variables in SAS and using

the list missing option to compare values among the old and recoded variables. The number of missing values were assessed for key variables used in the analysis. Every major statistical programming step in SAS was documented. Data was converted from SAS format (.sas7bdat extension) to STATA format (.dta extension) using STAT/transfer.

Before conducting the analysis, data was reshaped from wide form (each HHIDPN has one line of data for the 6 waves) into long form using the **`reshape`** option in STATA. The long form had data for each HHIDPN in a separate line by year. After converting the data into long form, the data was weighted to account for oversampling, clustering and stratification of the survey sample. Specifically, HRS variables SECU and STRATUM were used as the clustering and stratification variables. Person level analysis weight were used as the weight variables.

Data Analysis

Descriptive analysis of the sample eligible to be included in the study (n=4,334) was conducted. This initial analysis was not weighted. Frequency distributions of categorical and binary variables were tabulated. Additionally, means and standard errors of continuous variables were be determined.

Approximately 39% of respondents had died during sometime during the followup period from 1992 to 2006. The mortality status of each respondent was provided by HRS in their tracker files. The STATA command **`misschk`** was used to determine missing data patterns among respondents and deceased respondents. Comparisons between these two groups for the dependent and key independent variables

was conducted using Chi-Square (χ^2) tests and t-tests. There were statistically significant differences between the two groups on several variables.

In our first model, we excluded data for deceased respondents and conducted a complete case analysis. GEE analysis requires that missing data be MCAR. In the second model, we assumed that although deceased and non-deceased respondents differed on several sociodemographic variables, data for deceased respondents was MCAR. Thus, the next model analyzed the entire observed sample (n=4,334).

However, missing data could also be assumed to be MAR. Thus, multiple imputation using **ICE** in STATA was performed prior to conducting GEE analysis. In the first multiply imputed data set, only missing data for the predictors was imputed. Data was imputed for all variables (dependent and predictors) in the second multiply imputed data set. **ICE** was selected as the multiple imputation technique because the dependent variable (alcohol consumption) was binary and did not meet the multivariate normal assumption of similar multiple imputation methods used in SAS. **ICE** assumes a multivariate distribution but not a multivariate normal distribution. **ICE** is relatively easy to download and implement. It can be downloaded directly from within the STATA software by typing findit **ICE**. Additionally, **ICE** models each missing variable based on the type of regression analysis (i.e. logistic for a binary variable, multinomial for a nominal variable). The following steps were taken to complete the imputation process:

1. Observed data (n = 4,334) reshaped into wide form
2. Imputation syntax created and verified in STATA using **dryrun** option
3. Variables recoded if necessary

4. Data ready to be imputed. Data imputed 10 times.
5. Imputed data reshaped into long form to prepare for pooling
6. Imputed data pooled using **MIM** prefix in STATA

There are no clear requirements in the literature regarding the number of imputations; generally, the greater the amount of missing data, the more imputations are recommended to achieve stable estimates. Several simulations using **ICE** have shown good results with just 5 impuations. After imputation, the STATA prefix, **MIM**, can be used to pool imputed datasets into one so analysis can be conducted on the one dataset. Imputed data will be distinguished from the eligible sample (n=4,334). The latter will be referred to as observed data throughout this dissertation.

After data was imputed, the **svyset** command was used in STATA to prepare the data for the complex survey sampling analysis. The STATA command **xtgee** was used to analyze a population average model using GEE. A comparison of naïve and robust standard errors were requested. Robust standard errors provide valid standard errors even if the working correlation structure is misspecified.

Working correlation structures were determined by running models with each of the following four structures: exchangeable, independent, autoregressive and unstructured on both the observed data and the imputed data. After the working correlation structure was determined, final analysis was conducted. Four models were analyzed. These included unadjusted models for the observed data, complete cases, and multiply imputed data. Models were adjusted using sociodemographic variables recommended in the literature. These included age, marital status, gender, race/ethnicity,

occupation, income, years of education completed, prevalence of heart disease, prevalence of hypertension, prevalence of diabetes, history of problem drinking and mental health score. The regression coefficients for GEE analysis for this study are in the log form since the link selected is binomial. Therefore, regression coefficients and accompanying confidence intervals are exponentiated to provide odds ratios. Odds ratios are used because of their ease in interpreting the results.

CHAPTER 4

RESULTS

Descriptive Analysis of Observed Data

Table 2 displays results of key demographic characteristics of the observed data (n=4,334). Of the 4,334 respondents, 60.7% (n=2,633) had complete data (or complete cases). The remainder of the respondents (39.3%) had died during the followup period (1992 to 2006) and subsequently had missing data due to non response. The majority of complete cases were in the age group of 55 – 59, female, White, married, and working in white collar jobs. Those with missing data were also predominantly 55 to 59 years old, White, married and working in white collar jobs. However, those with missing data were primarily male. Gender, race/ethnicity and occupational differences between these two groups was statistically significant. The difference in prevalence of hypertension, heart disease and diabetes among respondents who had complete data versus those with missing data was statistically significant ($p < 0.05$).

The groups differed statistically ($p = 0.002$) in self-reports of alcohol consumption at baseline (1992). There was no statistical difference in the groups in terms of history of problem drinking. However, there were differences in the number of drinks consumed in 1992 ($p = 0.014$). Those with missing data were more likely to report a greater consumption of five or more drinks daily. Finally, there were no differences among the two groups for job displacement in 1994. Overall, the two groups differed statistically for nine of the twelve variables for which comparisons were made in Table 2.

Results of Missing Data Analysis

Running a missing data description using the **`misschk`** option in STATA (see Appendix B) revealed that data was missing for key variables, including the dependent variable. Approximately 8.7 percent to 77.6 percent of data was missing for the variables assessing alcohol consumption (currdrink) and displacement (displaced) (see Table 3). For variables not listed in Table 3, data was complete. About 3.8% of data was missing for marital status of the respondent. The largest amount of missing data was for the variables measuring displacement and alcohol consumption. More than 50% of data for a respondents' displacement status from 2000 to 2006 was missing. Less data was missing for the binary variable assessing alcohol consumption. For this variable, approximately 25% percent of data for 2002, 2004 and 2006 was missing.

Table 2. Missing Data Analysis of Unweighted Observed Sample: HRS, 1992 (n=4,334)[a]

Variable	Complete Data[b] Frequency (Column Percentage)[c]	Missing Data[b] Frequency (Column Percentage)	P-value
Age group			
50 - 54	1,122 (60.78)	724 (39.22)	0.141
55 - 59	1,194 (61.77)	739 (38.23)	
60 - 64	317 (57.12)	238 (42.88)	
Sex			
Male	1,284 (57.60)	945 (42.40)	0.00***
Female	1,349 (64.09)	756 (35.91)	
Race			
White	2,140 (62.94)	1,270 (37.06)	0.00***
Black	413 (52.48)	374 (47.52)	
Hispanic	27 (55.10)	22 (44.90)	
Other	53 (54.08)	45 (45.92)	
Marital Status			
Married/Partner	2,041 (61.64)	1,270 (38.36)	0.089
Single/Sep/Div/Widowed	501 (58.46)	356 (41.54)	
Occupation			
Blue collar	742 (55.96)	584 (44.04)	0.000***
White collar	1,889 (62.95)	1,112 (37.05)	
History of problem drinking	314 (56.58)	241 (43.42)	0.031*
Prevalence of hypertension	921 (56.64)	705 (43.36)	0.000***
Prevalence of heart disease	239 (53.59)	207 (46.41)	0.001**
Prevalence of diabetes	198 (53.66)	171 (46.34)	0.004**
Consume alcohol, 1992	1,757 (62.46)	1,056 (37.54)	0.002**
Number of drinks consumed, 1992			
Less than 1 per day	1,358 (63.76)	772 (36.24)	0.014*
1 – 2 per day	277 (59.57)	188 (40.43)	
3 – 4 per day	96 (59.63)	65 (40.37)	
5 or more per day	26 (45.61)	31 (54.39)	
Displaced, 1994	81 (71.68)	32 (28.32)	0.220

[a]There may be some differences in sample size within categories due to missing values, blanks or refusals.
[b]For complete data, n=2,633. For missing data, n=1,701.
[c]Percent shown represent row percentages.

Table 3. Key Variables with Missing Data, Unweighted Observed Sample: HRS, 1992 (n=4,334)

Variable	**Percent of Missing Data**
Marital status	3.8
Displaced	
1992	0.0
1994	19.0
1996	33.7
1998	44.1
2000	56.8
2002	66.4
2004	71.5
2006	77.6
Currently drink	
1992	0.0
1994	8.7
1996	13.0
1998	16.9
2000	21.1
2002	23.7
2004	26.7
2006	29.8

Descriptive Analysis of Key Variables among Complete Cases

Prior to the multiple imputation, detailed descriptive analysis was conducted for the dependent and independent variables measuring alcohol consumption and job displacement in this study. As discussed in the previous section, there was a considerable amount of missing data for these two variables. However, it is important to get a sense of any patterns/trends in these two variables since they are the dependent and predictor variables, respectively. A detailed tabulation of frequency of these variables for the study period, 1992 to 2006, is shown in Table 4. This table is restricted to those individuals with complete data (i.e. complete cases).

More than 50% of respondents reported drinking alcohol from 1992 to 2006. There was a decreasing trend in alcohol consumption. Almost 68% of respondents reported drinking in 1992. This percent slowly decreased to 54% reporting drinking in 2006. Since all individuals at baseline (year 1992) were currently employed, the cell representing displacement in 1992 in Table 4 is correctly reflecting 0%. Less than 10% of 2,633 complete cases were displaced anytime during the study period. There was an overall decreasing trend in job displacement among respondents, with slight increases in 1998, 2002 and 2006. The variable measuring the number of drinks consumed had more than 50% missing values. Thus, frequency distribution of this variable is not provided.

Table 4. Frequency of Alcohol Consumption and Displacement among Complete Cases, 1992 - 2006 (n=2,633)

Variable	**Frequency (Row Percentage)[a]**	**Frequency (Row Percentage)[a]**
Currently Drink	Yes	No
1992	1,757 (66.73)	876 (33.27)
1994	1,580 (61.26)	999 (38.74)
1996	1,536 (58.36)	1,096 (41.64)
1998	1,476 (56.06)	1,157 (43.94)
2000	1,410 (53.55)	1,223 (46.45)
2002	1,413 (53.67)	1,220 (46.33)
2004	1,393 (52.93)	1,239 (47.07)
2006	1,429 (54.29)	1,203 (45.71)
Displaced	Yes	No
1992	0 (0.00)	2,633 (100.00)
1994	81 (3.48)	2,248 (96.52)
1996	60 (2.95)	1,976 (97.05)
1998	73 (4.04)	1,736 (95.96)
2000	31 (2.10)	1,443 (97.90)
2002	61 (5.09)	1,137 (94.91)
2004	46 (4.42)	995 (95.58)
2006	48 (5.56)	815 (94.44)

[a]Unweighted. There may be some differences in sample size within categories due to missing values, blanks or refusals.

Multiple Imputation Results

Data was imputed 10 times. The code for imputation using **ICE** in STATA is provided in Appendix C. Additionally, results of the **dryrun** are also provided because they illustrate the process of **ICE** visually. After imputation was completed, missing data diagnostics were conducted. Scatterplots of missing data versus imputed data are a useful diagnostic tool. Additionally, quantile-quantile plots of observed and imputed values can also be useful. However, all the variables with missing data in this study were binary in nature. Thus, scatterplots quantile-quantile plots were not created. Instead, each variable with missing data was compared with the imputed variable using **tabmiss**. These results for multiple imputation of all variables (dependent variable and predictors) are provided in Appendix D. The variable measuring current alcohol consumption is assessed before and after imputation as an example. Missing data for this variable ranged from 8 to 29%. However, after imputation, this variable had no missing data.

Determining the Working Correlation Structure

Data, once imputed, was now ready for GEE analysis. One of the prerequisites before GEE analysis, is specifying a working correlation structure (i.e. unstructured, independent, autoregressive or exchangeable). Misspecification of the working correlation structure can result in incorrect standard errors, especially when a naïve analysis is conducted. However, when robust standard errors are requested in STATA, misspecification of the working correlation matrix is not as much of a concern. Table 5

displays differences in regression coefficients and semi-robust standard errors for complete cases, the observed sample and multiple imputation analysis.

Regardless of the type of model, semi-robust estimates yielded higher *p* values and slightly higher standard errors than naïve estimates. The model analyzing the observed sample had the lowest *p* values and three types of correlation structures significant (unstructured, exchangeable and autoregressive). The multiple imputation model of all variables had the highest *p* values and the least number of significant correlation structures. Only the exchangeable correlation structure was significant.

The complete case model and the multiple imputation of predictors only model were similar in that both indicated that the unstructured and exchangeable correlation structures were statistically significant. However, of all models, the complete case model had the highest semi-robust standard errors, while multiple imputation of predictors only yielded the smallest semi-robust standard errors. One reason for the higher semi-robust standard errors in the complete case analysis is that complete cases had the smallest sample size of all models (i.e. n=2,633) and thus, the smallest denominator in the calculation of standard error.

GEE Analysis of Complete Cases

GEE analysis results, both unadjusted and adjusted, for complete cases (n=2,633) is provided in Table 6. Results show that those respondents who were not displaced had lower odds of alcohol consumption in both the unadjusted model (odds ratio (OR) = 0.80; standard error (SE) = 0.076; 95% confidence interval (CI) = 0.69 to 0.93; $p < 0.01$) and adjusted model. In the adjusted model, job displaced remained a

statistically significant predictor of alcohol consumption (OR = 0.79; SE = 0.083; 95% CI = 0.67 to 0.93; $p < 0.01$). Additionally, sex, race/ethnicity, income, years of education, and prevalence of diabetes were significant predictors of alcohol consumption among those who were displaced.

GEE Analysis of Observed Sample

GEE analysis of the observed sample (see Table 7) also showed job displaced to be a statistically significant predictor of the onset of alcohol consumption in the unadjusted model (OR = 0.76; SE = 0.051; 95% CI = 0.66 to 0.87; $p < 0.01$). The observed sample model had slightly smaller standard error, a smaller confidence interval and smaller p value than the complete case model. After controlling for age, marital status, sex, race/ethnicity, occupation, income, years of education, prevalence of heart disease, prevalence of diabetes, prevalence of diabetes, history of problem drinking and mental health score, the probability of the onset of alcohol consumption was lower among those not displaced versus those who were displaced (OR = 0.75; SE = 0.056; 95% CI = 0.65 to 0.87; $p < 0.01$). Two additional predictors were significant in this model: occupation and baseline mental health status.

GEE Analysis of Multiply Imputed Data – Predictors Only

Job displacement was a significant predictor of the onset of alcohol consumption in the GEE analysis of the multiple imputation of predictors model (see Table 8). The unadjusted model was statistically significant (OR = 0.85; SE = 0.053; 95% CI = 0.76 to 0.94; $p < 0.01$). After controlling for age, marital status, sex, race/ethnicity, occupation,

income, years of education, prevalence of heart disease, prevalence of diabetes, prevalence of diabetes, history of problem drinking and mental health score, the probability of the onset of alcohol consumption was lower among those not displaced versus those who were displaced (OR = 0.85; SE = 0.057; 95% CI = 0.75 to 0.95; $p < 0.01$). This model differed slightly from the adjusted, observed sample model. Prevalence of hypertension and heart disease among the displaced were significant in the onset of alcohol consumption in this model, while baseline mental health status was not significant.

GEE Analysis of Multiple Imputation of all Variables

GEE analysis of multiple imputation of all variables (see Table 9) also showed job displaced to be a statistically significant predictor of alcohol consumption in the unadjusted model (OR = 0.84; SE = 0.069; 95% CI = 0.72 to 0.97; $p < 0.05$). After controlling for age, marital status, sex, race/ethnicity, occupation, income, years of education, prevalence of heart disease, prevalence of diabetes, prevalence of diabetes, history of problem drinking and mental health score, the probability of alcohol consumption was lower among those not displaced versus those who were displaced (OR = 0.84; SE = 0.078; 95% CI = 0.71 to 0.99; $p < 0.05$).

The unadjusted, complete case analysis produced slightly smaller confidence intervals and higher p values than adjusted, multiple imputation model. Even the adjusted, complete case model resulted in smaller confidence intervals and higher statistical significance than the adjusted, multiple imputation model. Variables statistically significant in the adjusted, complete case model included gender,

race/ethnicity, income, years of education, and prevalence of diabetes. These variables remained significant in the adjusted, multiple imputation model. Additional variables that were statistically significant in the latter model were occupation, and prevalence of hypertension and heart disease.

Overall comparison of the predictor job displacement in the four adjusted models shows that the observed sample model had the lowest p value ($p < 0.01$), and the lowest standard errors. The odds ratio was also the lowest in this model (OR = 0.75). Both the observed sample analysis and the complete case analysis produced wider confidence intervals than the multiple imputation analyses.

Table 5. Differences in Regression Coefficients and Standard Errors in Working Correlation Structure by Type of Analysis

Variable	**Working Correlation Structure**	**Complete Case Analysis[a]**				**Observed Sample Analysis[b]**			
		Naïve Estimates		**Semi-Robust Estimates**		**Naïve Estimates**		**Semi-Robust Estimates**	
		Est[a] (SE)[b]	P-value	Est (SE)	P-value	Est[a] (SE)[b]	P-value	Est (SE)	P-value
Displaced	Unstructured	-0.170 (0.065)**	0.009	-0.170 (0.075)*	0.023	-0.234 (0.056)***	0.000	-0.234 (0.066)***	0.000
	Exchangeable	-0.223 (0.067)***	0.001	-0.223 (0.076)**	0.003	-0.276 (0.058)***	0.000	-0.276 (0.068)***	0.000
	Independent	-0.049 (0.103)	0.636	-0.049 (0.109)	0.654	-0.114 (0.087)	0.188	-0.114 (0.092)	0.216
	Auto regressive	-0.174 (0.079)*	0.028	-0.174 (0.104)	0.095	-0.250 (0.070)***	0.000	-0.250 (0.091)**	0.006
		Multiple Imputation Analysis (predictors only)[c]				**Multiple Imputation Analysis (dependent variable and predictors)[d]**			
		Naïve Estimates		**Semi-Robust Estimates**		**Naïve Estimates**		**Semi-Robust Estimates**	
		Est[a] (SE)[b]	P-value	Est (SE)	P-value	Est (SE)	P-value	Est (SE)	P-value
Displaced	Unstructured	-0.111 (0.048)*	0.026	-0.111 (0.051)*	0.035	-0.107 (0.071)	0.157	-0.107 (0.073)	0.164
	Exchangeable	-0.166 (0.050)**	0.002	-0.166 (0.053)**	0.003	-0.177 (0.068)*	0.020	-0.177 (0.069)*	0.021
	Independent	-0.165 (0.089)	0.074	-0.165 (0.093)	0.087	-0.085 (0.084)	0.322	-0.085 (0.087)	0.339
	Auto regressive	-0.081 (0.053)	0.134	-0.081 (0.059)	0.172	-0.068 (0.079)	0.413	-0.068 (0.082)	0.425

[a]Complete case analysis consists of sample with no missing values.

[b]Observed sample consists of complete cases and respondents who died during 1992-2006.

[c]Multiple imputation of predictors (independent variables) only.

[d]Multiple imputation of both dependent variable and predictors.

[a]Est = Regression coefficients
[b]SE = standard error
[§]Unweighted data
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 6. Results of GEE analysis of alcohol consumption among displaced workers, Complete Cases

Variable	**Complete Cases, Unadjusted[a]**			**Complete Cases, Adjusted[b]**		
	OR (SE)[§]	95% CI	P-value	OR (SE)	95% CI	P-value
Displaced	0.80 (0.076)**	(0.69,0.93)	0.003	0.79 (0.083)**	(0.67,0.93)	0.005
Age	-	-	-	1.05 (0.055)	(0.95,1.17)	0.341
Marital status	-	-	-	0.99 (0.091)	(0.83,1.18)	0.898
Sex	-	-	-	0.66 (0.098)***	(0.55,0.80)	0.000
Race/ethnicity	-	-	-	0.72 (0.063)***	(0.63,0.81)	0.000
Occupation	-	-	-	0.84 (0.094)	(0.70,1.01)	0.068
Income	-	-	-	1.17 (0.033)***	(1.09,1.24)	0.000
Years of education	-	-	-	1.08 (0.016)***	(1.05,1.12)	0.000
Hypertension[±]	-	-	-	0.92 (0.078)	(0.79,1.07)	0.275
Heart disease[±]	-	-	-	0.82 (0.130)	(0.63,1.05)	0.120
Diabetes[±]	-	-	-	0.53 (0.138)***	(0.40,0.70)	0.000
Problem drinking	-	-	-	1.14 (0.123)	(0.90,1.45)	0.279
Mental health score	-	-	-	1.07 (0.051)	(0.97,1.18)	0.188

[a]Weighted, unadjusted GEE analysis of complete cases (n=2,633). Complete cases are those respondents with no missing data.
[b]Weighted, adjusted GEE analysis of complete cases. Adjusting variables are those listed above.
[±]These variables indicate prevalence of hypertension, heart disease and diabetes.
[§]SE is standard error (semi-robust).
*$p < .05$; **$p < .01$; ***$p < .001$.

Table 7. Results of GEE analysis of alcohol consumption among displaced workers, Observed Sample

Variable	**Observed Sample, Unadjusted[a]**			**Observed Sample, Adjusted[b]**		
	OR (SE)[§]	95% CI	P-value	OR (SE)	95% CI	P-value
Displaced	0.76 (0.051)***	(0.66,0.87)	0.000	0.75 (0.056)***	(0.65,0.87)	0.000
Age	-	-	-	1.06 (0.045)	(0.98, 1.16)	0.141
Marital status	-	-	-	0.99 (0.071)	(0.86, 1.14)	0.918
Sex	-	-	-	0.63 (0.048)***	(0.55, 0.73)	0.000
Race/ethnicity	-	-	-	0.73 (0.035)***	(0.67, 0.81)	0.000
Occupation	-	-	-	0.83 (0.061)*	(0.72, 0.96)	0.012
Income	-	-	-	1.14 (0.027)***	(1.09, 1.20)	0.000
Years of education	-	-	-	1.08 (0.013)***	(1.06, 1.11)	0.000
Hypertension[±]	-	-	-	0.92 (0.056)	(0.82, 1.04)	0.175
Heart disease[±]	-	-	-	0.86 (0.083)	(0.71, 1.04)	0.122
Diabetes[±]	-	-	-	0.57 (0.058)***	(0.47, 0.70)	0.000
Problem drinking	-	-	-	1.10 (0.103	(0.91, 1.32)	0.316
Mental health score	-	-		1.10 (0.043)*	(1.01, 1.18)	0.032

[a]Weighted, unadjusted GEE analysis of observed sample (n=4,334). Observed sample consists of respondents with complete data and respondents who died between 1992-2006.
[b]Weighted, adjusted GEE analysis of observed sample. Adjusting variables are those listed above.
[±]These variables indicate prevalence of hypertension, heart disease and diabetes.
[§]SE is standard error (semi-robust).
*$p < .05$; **$p < .01$; ***$p < .001$.

Table 8. Results of GEE analysis of alcohol consumption among displaced workers, Multiple Imputation of Predictors

Variable	**Multiple Imputation (predictors only), Unadjusted[a]**			**Multiple Imputation (predictors only), Adjusted[b]**		
	OR (SE)[§]	95% CI	P-value	OR (SE)	95% CI	P-value
Displaced	0.85 (0.053)**	(0.76, 0.94)	0.003	0.85 (0.057)**	(0.75, 0.95)	0.005
Age	-	-	-	0.98 (0.038)	(0.91, 1.06)	0.605
Marital status	-	-	-	1.01 (0.066)	(0.89, 1.15)	0.830
Sex	-	-	-	0.61 (0.068)***	(0.75, 0.95)	0.000
Race/ethnicity	-	-	-	0.73 (0.044)***	(0.67, 0.79)	0.000
Occupation	-	-	-	0.82 (0.067)**	(0.72, 0.93)	0.003
Income	-	-	-	1.13 (0.022)***	(1.09, 1.18)	0.000
Years of education	-	-	-	1.09 (0.011)***	(1.07, 1.12)	0.000
Hypertension[±]	-	-	-	0.89 (0.055)*	(0.80, 0.99)	0.032
Heart disease[±]	-	-	-	0.84 (0.088)*	(0.70, 1.00)	0.044
Diabetes[±]	-	-	-	0.52 (0.090)***	(0.44, 0.62)	0.000
Problem drinking	-	-	-	1.07 (0.084)	(0.91, 1.26)	0.409
Mental health score	-	-	-	1.07 (0.036)	(1.00, 0.91)	0.055

[a]Weighted, unadjusted GEE analysis of data multiply imputed for all predictors listed above.
[b]Weighted, adjusted GEE analysis of data multiply imputed for all predictors listed above. Adjusting variables are those listed above.
[±]These variables indicate prevalence of hypertension, heart disease and diabetes.
[§]SE is standard error (semi-robust).
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table 9. Results of GEE analysis of alcohol consumption among displaced workers, Multiple Imputation of all Variables

Variable	**Multiple Imputation, Unadjusted[a]**			**Multiple Imputation, Adjusted[b]**		
	OR (SE)	95% CI	P-value	OR (SE)	95% CI	P-value
Displaced	0.84 (0.069)*	(0.72,0.97)	0.021	0.84 (0.078)*	(0.71,0.99)	0.038
Age	-	-		0.97 (0.039)	(0.90,1.05)	0.421
Marital status	-	-		1.05 (0.069)	(0.91,1.20)	0.524
Sex	-	-		0.62 (0.070)***	(0.54,0.71)	0.000
Race/ethnicity	-	-		0.72 (0.044)***	(0.66,0.78)	0.000
Occupation	-	-		0.81 (0.070)**	(0.71,0.93)	0.003
Income	-	-		1.14 (0.022)***	(1.09,1.19)	0.000
Years of education	-	-		1.10 (0.011)***	(1.08,1.13)	0.000
Hypertension[±]	-	-		0.88 (0.055)*	(0.79,0.98)	0.023
Heart disease[±]	-	-		0.83 (0.089)*	(0.70,0.99)	0.035
Diabetes[±]	-	-		0.51 (0.096)***	(0.42,0.62)	0.000
Problem drinking	-	-		1.09 (0.084)	(0.92,1.29)	0.314
Mental health score	-	-		1.07 (0.036)	(1.00,1.15)	0.051

[a]Weighted, unadjusted GEE analysis of multiply imputed data.
[b]Weighted, adjusted GEE analysis of multiply imputed data. Adjusting variables are those listed above.
[±]These variables indicate prevalence of hypertension, heart disease and diabetes.
[§]SE is standard error (robust).
*$p < .05$; **$p < .01$; ***$p < .001$.

CHAPTER 5

DISCUSSION

We examined longitudinal data from HRS from 1992 to 2006 and tested for differences in current alcohol consumption among those displaced versus those not displaced within the study period. This was substantiated by our data. All models showed that continuous employment had a protective effect on alcohol consumption. These findings are consistent with Gallo et al. (2001). This study confirmed that there are statistical differences in alcohol consumption among the older workers who have experienced job displacement compared to those who remain continuously employed. Older worker who were displaced during the study period were more likely to begin consuming alcohol than those who were not displaced. This finding has important public health implications. Older workers are likely to have varied participation in the labor market. They are likely to be more experienced and hold senior or management positions, thereby earning higher wages. Economic changes in the labor market, such as recession, lead companies to adjust by downsizing, which can include laying off working or closing businesses. In the case of layoffs, sometimes the first to go are those in management positions. This had a direct implication for older workers. The effects of alcohol consumption among older individuals has been shown to be negative and particularly harmful in terms of ethanol toxicity. Although we did not examine whether alcohol consumption measured by the number of drinks consumed increased among current drinkers experiencing job displacement, the finding of increased onset of drinking

among non-drinkers who were displaced is of particular concern. Further studies are needed to examine the health effects of late onset of drinking in this population.

Our study was longitudinal in nature. This is an important strength. The advantage of longitudinal studies over cross sectional studies is ability to examine and test for causality between dependent variables and predictors. Cross sectional studies may be able to show association and correlation, but do not have the design and statistical elements to attribute causality. The fact that this analysis included 8 waves of longitudinal data speaks to the methodological soundness of its study design.

Loss of participants or respondents over the length of a study is of particular concern in longitudinal studies. Longitudinal studies typically have a much longer study duration than other designs. Participants can be followed for years. This study was 14 years in duration. We limited analysis from 1992 to 2006 due to the available data. However, HRS is currently still being conducted biannually. As the length of a study increases, an expected outcome is to see participants either dropping out/withdrawing, being lost to followup, relocating, or even dying. The latter was especially of concern in our study population, since the study sample was older workers aged 51 to 61 years in 1992. By 2006, the last year of analysis, this sample would be in the 65 to 75 age range. We found that death was a primary reason for participants being lost to followup. There are likely additional reasons for loss of followup. We were unable to determine these reasons from the available data.

We found about 30% data missing for the dependent variable and almost 77% of data missing for the independent variable assessing job displacement. GEE analysis

assumes that missing data is MCAR. There were statistical differences between respondents who were lost to followup due to death versus those who responded to each Wave of data. However, we made no underlying assumptions about the pattern of missingness. Instead, several statistical options were carefully considered prior to data analysis.

The observed sample was comprised of 4,334 individuals. Complete data (or data for non-deceased respondents) was available for 2,633 individuals. Complete case analysis or case-wise deletion, although not recommended in the literature, could be conducted to meet GEE requirements of missing data MCAR. A better option is to impute missing data through multiple imputation. Multiple imputation has been shown to produce unbiased results (Greenland & Finkle, 1995; Little, 1992; Rubin, 1976). We found differences in the results, including statistical inference, based on the type of method used to handle missing data. The direction of the estimates was the same regardless of the type of analysis. We used real data with missing values (rather than simulated data). The differences between complete case analysis and multiple imputation were smaller than expected. A larger sample size could potentially show more pronounced differences in comparative analysis of complete cases versus multiple imputation.

Several limitations of this study warrant discussion. Alcohol consumption in this study was assessed with the HRS question, "Do you ever any alcoholic beverages, such as beer, wine, or liquor?" This question is essentially a snapshot of incidence of alcohol consumption. It does not assess trends in alcohol consumption, or whether a

respondent has never consumed alcohol. Nor does the question distinguish between the type of alcohol consumed among drinkers. Alcohol consumption is self-reported. There may be a tendency to under report socially undesirable behavior such as consuming alcohol. In our study, this can lead to bias in the frequency of both the dependent and independent variable measuring job displacement. Finally, the multiple imputation technique, **`ice`**, is a fairly new technique. It lacks the theoretical rigor of other techniques implemented via widely available software, such as SAS. Future comparative studies can be conducted to assess differences in statistical inferences using different multiple imputation software.

CHAPTER 6

BIBLIOGRAPHY

Ambler, G., Omar, R. Z., & Royston, P. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research, 16*(3), 277-298.

Andersen, R. M. (2008). National health surveys and the behavioral model of health services use. *Medical Care, 46*(7), 647-653.

Bacharach, S., Bamberger, P. A., Sonnenstuhl, W. J., & Vashdi, D. (2008). Aging and drinking problems among mature adults: The moderating effects of positive alcohol expectancies and workforce disengagement. *Journal of Studies on Alcohol and Drugs, 69*(1), 151-159.

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*(1), 5-37.

Barnhart, H. X., & Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics, 54*(2), 720-729.

Berman, N. G., & Parker, R. A. (2002). Meta-analysis: Neither quick nor easy. *BMC Medical Research Methodology, 2*, 10.

Bernhardt, A., Morris, M., Handcock, M., & Scott, M. (1999). Trends in job instability and wages for young adult men. *Journal of Labor Economics*, *17*(4), S65-S90.

Burkhauser, R. V., & Gertler, P. J. (1995). Introduction to special issues on the Health and Retirement Survey: Data quality and early results. *Journal of Human Resources, 30* (Suppl.), S1–S6.

Centers for Disease Control and Prevention (CDC). (2004). Alcohol-attributable deaths and years of potential life lost--United States, 2001. *MMWR. Morbidity and Mortality Weekly Report, 53*(37), 866-870.

Chirikos, T. N. (1993). The relationship between health and labor market status. *Annual Review of Public Health, 14*, 293-312.

Cook, R. J., Zeng, L., & Yi, G. Y. (2004). Marginal analysis of incomplete longitudinal binary data: A cautionary note on LOCF imputation. *Biometrics, 60*(3), 820-828.

Corrao, G., Bagnardi, V., Zambon, A., & La Vecchia, C. (2004). A meta-analysis of alcohol consumption and the risk of 15 diseases. *Preventive Medicine, 38*(5), 613-619.

Couch, K.A. (1998). Late life job displacement. *The Gerontologist, 38*(1), 7-17.

Crimmins, E. M., Reynolds, S. L., & Saito, Y. (1999). Trends in health and ability to work among the older working-age population. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 54B*(1), S31-40.

Dhalla, S., & Kopec, J. A. (2007). The CAGE questionnaire for alcohol misuse: A review of reliability and validity studies. *Clinical and Investigative Medicine, 30*(1), 33-41.

Diehr, P., Johnson, L. L., Patrick, D. L., & Psaty, B. (2005). Methods for incorporating death into health-related variables in longitudinal studies. *Journal of Clinical Epidemiology, 58*(11), 1115-1124.

Diggle, P.J., Heagerty, P., Liang, K.-Y., & Zeger, S.L. (2002). *Analysis of Longitudinal Data, 2nd edition*. New York: Oxford University Press.

Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*(10), 1087-1091.

Dooley, D., Fielding, J., & Levi, L. (1996). Health and unemployment. *Annual Review of Public Health, 17*, 449-465.

Dooley, D., & Prause, J. (2004). *The Social Costs of Underemployment: Inadequate Employment as Disguised Unemployment*. New York: Cambridge University Press.

Elkind, M. S., Sciacca, R., Boden-Albala, B., Rundek, T., Paik, M. C., & Sacco, R. L. (2006). Moderate alcohol consumption reduces risk of ischemic stroke: The northern Manhattan study. *Stroke; a Journal of Cerebral Circulation, 37*(1), 13-19.

Ewing, J.A. (1985). Detecting Alcoholism: The Cage Questionnaire. *JAMA, 252*(14), 1905-1907.

Falba, T., Teng, H. M., Sindelar, J. L., & Gallo, W. T. (2005). The effect of involuntary job loss on smoking intensity and relapse. *Addiction, 100*(9), 1330-1339.

Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds.). (2008). *Longitudinal Data Analysis: A handbook of modern statistical methods*. Florida: Chapman & Hill/CRC Press.

Foran, H. M., & O'Leary, K. D. (2008). Alcohol and intimate partner violence: A meta-analytic review. *Clinical Psychology Review, 28*(7), 1222-1234.

French, M. T., & Zarkin, G. A. (1995). Is moderate alcohol use related to wages? evidence from four worksites. *Journal of Health Economics, 14*(3), 319-344.

Gallo, W. T., Bradley, E. H., Falba, T. A., Dubin, J. A., Cramer, L. D., Bogardus, S.T.,Jr, et al. (2004). Involuntary job loss as a risk factor for subsequent myocardial infarction and stroke: Findings from the health and retirement survey. *American Journal of Industrial Medicine, 45*(5), 408-416.

Gallo, W. T., Bradley, E. H., Siegel, M., & Kasl, S. V. (2001). The impact of involuntary job loss on subsequent alcohol consumption by older workers: Findings from the health and retirement survey. *Journals of Gerontology Series B-Psychological Sciences & Social Sciences, 56*(1), S3-9.

Gallo, W. T., Bradley, E. H., Siegel, M., & Kasl, S. V. (2000). Health effects of involuntary job loss among older workers: Findings from the health and retirement survey. *Journals of Gerontology Series B-Psychological Sciences & Social Sciences, 55*(3), S131-40.

Ganguli, M., Bilt, J. V., Saxton, J. A., Shen, C., & Dodge, H. H. (2005). Alcohol consumption and cognitive function in late life: A longitudinal community study. *Neurology, 65*(8), 1210-1217.

Goldberg, R., Burchfiel, C., Reed, D., Wergowske, G., & Chiu, D. (1994). A prospective study of the health effects of alcohol consumption in middle-aged and elderly men. *Circulation, 89*(2), 651-659.

Gottschalk, P., & Moffitt, R. (1999). Changes in job instability and insecurity using monthly survey data. *Journal of Labor Economics, 17*(4), S91-126.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549-576.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology, 142*(12), 1255-1264.

Hartley, H. O., & Rao, J. N. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*(1), 93-108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association,* 72, 320-340.

Hauser, R.M., & Willis, R.J. (2005). Survey design and methodology in the Health and Retirement Study and the Wisconsin Longitudinal Study. In L.J. Waite (Ed.), *Aging, Health, and Public Policy: Demographic and Economic Perspectives.* New York: Population Council. Supplement to *Population and Development Review* Vol 30 (2004).

He, Y., Colantonio, A., & Marshall, V. (2006). The relationships between career instability and health condition in older workers: A longitudinal data analysis of

the Survey of Labour and Income Dynamics. In Leroy Stone (Ed.), *New Frontiers of Research on Retirement.* Ottawa: Statistics Canada, pp. 321-342.

Heeringa, S. G., & Connor, J. H. (1995). Technical description of the health and retirement survey sample design. HRS/AHEAD Documentation Report DR-002.

Hipple, Steven. (1999). Worker displacement in the mid-1990's. *Monthly Labor Review, 122*(7), 15-32.

Hogan, J. W., Roy, J., & Korkontzelou, C. (2004) Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine, 23,* 1455-1497.

Horton, N. J., & Lipsitz, S. R. (1999). Review of software to fit generalized estimating equation regression models. *The American Statistician, 53*(2), 160-169.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61*(1), 79-90.

Jaeger, D. A., & Stevens, A.H. (1999). Is job stability in the United States falling? Reconciling trends in the CPS and PSID. *Journal of Labor Economics, 17*, S1-S28.

Johnston, G., & Strokes, M. (1997). Application of GEE methodology using the SAS system. *North East SAS®Users Group, Inc.*

Juster, F. T., & Suzman, R. (1995). An overview of the health and retirement study. *The Journal of Human Resources, 30*(Special Issue on the Health and Retirement Study: Data Quality and Early Results), S7-S56.

Kaplan, D. (Ed.) (2004). *The Sage Handbook of Quantitative Methodology in the Social Sciences*. Newbury Park, CA: Sage Publications.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. (1998). *Applied Regression Analysis and other Multivariate Methods.* 3rd ed. Pacific Grove: Duxbury Press.

Kramer, G. H. (1983). The ecological fallacy revisited: Aggregate- versus individual-level findings on economics and elections, and sociotropic voting. *The American Political Science Review, 77*(1), 92-111.

Krieger, N. (2001). Theories for social epidemiology in the 21st century: An ecosocial perspective. *International Journal of Epidemiology, 30*(4), 668-677.

Kuchibhatla, M., & Fillenbaum, G. G. (2003). Comparison of methods for analyzing longitudinal binary outcomes: Cognitive status as an example. *Aging & Mental Health, 7*(6), 462-468.

Kushner, M. G., Abrams, K., & Borchardt, C. (2000). The relationship between anxiety disorders and alcohol use disorders: A review of major perspectives and findings. *Clinical Psychology Review, 20*(2), 149-171.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*(4), 963-974.

Liang, K., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

Lipsitz, S. R., Kim, K., & Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medicine, 13,* 1149-1163.

Little, R.A. (1992). Regression with missing X's; a review. *J Am Stat Assoc, 87,* 1227-1237.

Little, R.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90,* 1112–1121.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics, 11*(1), 59-67.

McDonough, P., & Amick, B. C.,3rd. (2001). The social context of health selection: A longitudinal study of health and employment. *Social Science & Medicine (1982), 53*(1), 135-145.

Mann, C. J. (2003). Observational research methods. research design II: Cohort, cross sectional, and case-control studies. *Emergency Medicine Journal, 20*(1), 54-60.

Martin, L. G., Freedman, V. A., Schoeni, R. F., & Andreski, P. M. (2009). Health and functioning among baby boomers approaching 60. *The Journals of Gerontology.Series B, Psychological Sciences and Social Sciences, 64*(3), 369-377.

Meier, P., & Seitz, H. K. (2008). Age, alcohol metabolism and liver disease. *Current Opinion in Clinical Nutrition & Metabolic Care, 11*(1), 21-26.

Miller, I. & Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications.* 7th ed. NJ: Prentice Hall.

Miller, M. E., Davis, C. S., & Landis, J. R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares. *Biometrics, 49,* 1033-1044.

Moore, A. A., Giuli, L., Gould, R., Hu, P., Zhou, K., Reuben, D., et al. (2006). Alcohol use, comorbidity, and mortality. *Journal of the American Geriatrics Society, 54*(5), 757-762.

Mukamal, K. J., Psaty, B. M., Rautaharju, P. M., Furberg, C. D., Kuller, L. H., Mittleman, M. A., et al. (2007). Alcohol consumption and risk and prognosis of atrial fibrillation among older adults: The cardiovascular health study. *American Heart Journal, 153*(2), 260-266.

Mullahy, J., & Sindelar, J. L. (1991). Gender differences in labor market effects of alcoholism. *The American Economic Review, 81*(2, Papers and Proceedings of the Hundred and Third Annual Meeting of the American Economic Association), 161-165.

Mullahy, J., & Sindelar, J. L. (1993). Alcoholism, work, and income. *Journal of Labor Economics, 11*(3), 494-520.

Mullahy, J., & Sindelar, J. (1996). Employment, unemployment, and problem drinking. *Journal of Health Economics, 15*(4), 409-434.

Mutchler, J. E., Burr, J. A., Massagli, M. P., & Pienta, A. (1999). Work transitions and health in later life. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 54B*(5), S252-261.

Mutchler, J. E., Burr, J. A., Pienta, A. M., & Massagli, M. P. (1997). Pathways to labor force exit: Work transitions and work instability. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 52B*(1), S4-12.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, Series A 135, 370–384.

O'Brien, C.P. (2008). The CAGE questionnaire for detection of alcoholism: A remarkably useful but simple tool. *JAMA, 300*(17), 2054-2056.

Paik, M.C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association, 92*(440), 1320- 1329.

Peters, R., Peters, J., Warner, J., Beckett, N., & Bulpitt, C. (2008). Alcohol, dementia and cognitive decline in the elderly: A systematic review. *Age and Ageing, 37*(5), 505-512.

Philipson, P. M., Ho, W. K., & Henderson, R. (2008). Comparative review of methods for handling drop-out in longitudinal studies. *Statistics in Medicine, 27*(30), 6276-6298.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385-401.

RAND HRS Data, Version H. Produced by the RAND Center for the Study of Aging, with funding from the National Institute on Aging and the Social Security Administration. Santa Monica, CA (February 2008).

Rehm, J., Greenfield, T. K., & Rogers, J. D. (2001). Average volume of alcohol consumption, patterns of drinking, and all-cause mortality: Results from the U.S. national alcohol survey. *American Journal of Epidemiology, 153*(1), 64-71.

Revicki, D. A., Gold, K., Buckman, D., Chan, K., Kallich, J. D., & Woolley, J. M. (2001). Imputing physical health status scores missing owing to mortality: Results of a simulation comparing multiple techniques. *Medical Care, 39*(1), 61-71.

Reynolds, K., Lewis, B., Nolen, J. D., Kinney, G. L., Sathya, B., & He, J. (2003). Alcohol consumption and risk of stroke: A meta-analysis. *JAMA, 289*(5), 579-588.

Royston, P. (2005). Multiple Imputation of Missing Values: Update. *The Stata Journal, 5:*188- 201.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

Schwartz, S. (1994). The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences. *American Journal of Public Health, 84*(5), 819-824.

Seitz, H. K., & Stickel, F. (2007). Alcoholic liver disease in the elderly. *Clinics in Geriatric Medicine, 23*(4), 905-921.

Shults, J., Sun, W., Tu, X., Kim, H., Amsterdam, J., Hilbe, J. M., et al. (2009). A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. *Statistics in Medicine.* doi:10.1002/sim.3622

Sickles, R. C., & Taubman, P. (1986). An analysis of the health and retirement status of the elderly. *Econometrica, 54*(6), 1339-1356.

Siegel, M., Bradley, E. H., Gallo, W. T., & Kasl, S. V. (2003). Impact of husbands' involuntary job loss on wives' mental health, among older adults. *Journals of Gerontology Series B-Psychological Sciences & Social Sciences, 58*(1), S30-7.

Singer, J. D. 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24,* 323-355.

Smith, G. S., Branas, C. C., & Miller, T. R. (1999). Fatal nontraffic injuries involving alcohol: A metaanalysis. *Annals of Emergency Medicine, 33*(6), 659-668.

Stata Corporation. (2009). Stata Statistical Software: Release 10.1. College Station, TX: StataCorp LP.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics, 40*(4), 961-971.

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology, 169*(9), 1133-1139.

Thun, M. J., Peto, R., Lopez, A. D., Monaco, J. H., Henley, S. J., Heath, C. W.,Jr, et al. (1997). Alcohol consumption and mortality among middle-aged and elderly U.S. adults. *The New England Journal of Medicine, 337*(24), 1705-1714.

UCLA: Academic Technology Services, Statistical Consulting Group. (n.d.) *Multiple imputation using ICE.* Retrieved from https://www.ats.ucla.edu/stat/Stata/library/ice.htm

van Buuren, S., Boshuizen, H.C., & Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18,* 681–694.

van Buuren, S., & Oudshoorn C.G.M. (2000). *Multivariate imputation by chained equations: MICE V1.0 User's Manual*. Report PG/VGZ/00.038. Leiden: TNO Preventie en Gezondheid.

van der Heijden, G. J. M. G., T. Donders, A. R., Stijnen, T., & Moons, K. G. M. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *Journal of Clinical Epidemiology, 59*(10), 1102-1109.

Wannamethee, S. G., Shaper, A. G., Perry, I. J., & Alberti, K. G. (2002). Alcohol consumption and the incidence of type II diabetes. *Journal of Epidemiology and Community Health, 56*(7), 542-548.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39*(2), 95-101.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika, 61*(3), 439-447.

Wheaton, B. (1990). Life transitions, role histories, and mental health. *American Sociological Review,* 55, 209–223.

Williamson, J.M., Kim, K., & Lipsitz, S.R. (1995). Analyzing bivariate ordinal data using a global odds ratio. J*ournal of the American Statistical Association*, *90*(432), 1432-1437.

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics, 42*(1), 121-130.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics, 44*(4), 1049-1060.

Zhang, D., & Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics, 57*(3), 795-802.

APPENDICES

Appendix A

Code for UVIS Implementation of `ICE` in STATA

```
uvis regress write math science, seed(12457) gen(w1)

     set seed 12457
     regress write math science
        tempname b e V chol bstar
        tempvar xb u
        matrix `b'=e(b)
        matrix `e'=e(b)
        matrix `V'=e(V)
        matrix `chol'=cholesky(`V')
        local colsofb=colsof(`b')

       local rmse=e(rmse)
       local df=e(df_r)
       local chi2=2*invgammap(`df'/2,uniform())

       local rmsestar=`rmse'*sqrt(`df'/`chi2')
        matrix `chol'=`chol'*sqrt(`df'/`chi2')

       forvalues i=1/`colsofb' {
         matrix `e'[1,`i']=invnorm(uniform())
          }
       matrix `bstar'=`b'+`e'*`chol'' /*disturbance here*/
       gen `u'=uniform()
       matrix score `xb'=`bstar' /*score the data with the new
     coefficient*/
       gen w2 = write
       replace w2=`xb'+`rmsestar'*invnorm(`u') if write==.
     /*disturbance here again*/
```

To create one imputed data set for multiple variables x_1, x_2, ..., x_k, with missing observations, ice does the following:

- Ignore observations for which every member of x_1, x_2, ..., x_k has a missing value. This step will eliminate the observations that are impossible to impute;
- For each variable with any missing data in x_1, x_2, ..., x_k, randomly order that variable and replicate its observed values across the missing cases. This step initializes the iterative procedure by filling in missing data at random;
- For each of x1, x2, ..., x_k, in turn, impute missing values by applying **uvis** with the remaining variables as covariates. Repeat the step above # times specified by the **cycles(#)** option. The default is 10.

Appendix B

STATA Output of Missing Data Analysis using `MISSCHK` Option

```
. misschk
Variables examined for missing values

   #  Variable        # Missing   % Missing
--------------------------------------------
   1  groupage              0         0.0
   2  sex                   0         0.0
   3  race                  0         0.0
   4  married             166         3.8
   5  v207                  0         0.0
   6  occupation            7         0.2
   7  newincome            87         2.0
   8  pdrinker              0         0.0
   9  hyper                 0         0.0
  10  heart                 0         0.0
  11  diabetes              0         0.0
  12  mhscore               0         0.0
  13  age1992               0         0.0
  14  displaced1992         0         0.0
  15  currdrink1992         0         0.0
  16  displaced1994       822        19.0
  17  currdrink1994       378         8.7
  18  displaced1996      1461        33.7
  19  currdrink1996       565        13.0
  20  displaced1998      1913        44.1
  21  currdrink1998       733        16.9
  22  displaced2000      2463        56.8
  23  currdrink2000       916        21.1
  24  displaced2002      2876        66.4
  25  currdrink2002      1029        23.7
  26  displaced2004      3098        71.5
  27  currdrink2004      1158        26.7
  28  displaced2006      3363        77.6
  29  currdrink2006      1292        29.8
  30  displaced1992         0         0.0
  31  currdrink1992         0         0.0
```

Appendix C

STATA Code for Imputation by Chained Equations (`ICE`)

Dry Run

```
. ice currdrink1992 currdrink1994 currdrink1996 currdrink1998
currdrink2000 currdrink2002 currdrink2004 currdrink2006 displaced1992
displaced1994 displaced1996 displaced1998 displaced2000 displaced2002
displaced2004 displaced2006 v207 married m1-m3 groupage a1-a3 sex
occupation hyper heart diabetes pdrinker mhscore s1 s2 s3 s4 race r1-r4
newincome, substitute(married: m1 m2 m3) cmd (m1-m3: mlogit, a1-
a3:ologit) seed (1001) m(10) dryrun
```

Actual Run

```
ice currdrink1992 currdrink1994 currdrink1996 currdrink1998
currdrink2000 currdrink2002 currdrink2004 currdrink2006 displaced1992
displaced1994 displaced1996 displaced1998 displaced2000 displaced2002
displaced2004 displaced2006 v207  marital groupage y1-y3 sex occupation
hyper heart diabetes pdrinker mhscore s1 s2 s3 s4 race newincome, cmd
(s1-s4:ologit, y1-y3:ologit) seed(1001)saving(imputednew) m(10)
```

Results of Dry Run

```
 Variable | Command | Prediction equation
------------+-------------------------------------------------------------
currdr~1992 |         | [No missing data in estimation sample]
currdr~1994 | logit   | currdrink1992 currdrink1996 currdrink1998
            |         | currdrink2000 currdrink2002 currdrink2004
            |         | currdrink2006 displaced1992 displaced1994
            |         | displaced1996 displaced1998 displaced2000
            |         | displaced2002 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~1996 | logit   | currdrink1992 currdrink1994 currdrink1998
            |         | currdrink2000 currdrink2002 currdrink2004
            |         | currdrink2006 displaced1992 displaced1994
            |         | displaced1996 displaced1998 displaced2000
            |         | displaced2002 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~1998 | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink2000 currdrink2002 currdrink2004
            |         | currdrink2006 displaced1992 displaced1994
            |         | displaced1996 displaced1998 displaced2000
            |         | displaced2002 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~2000 | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2002 currdrink2004
            |         | currdrink2006 displaced1992 displaced1994
            |         | displaced1996 displaced1998 displaced2000
            |         | displaced2002 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
```

```
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~2002 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2004
             |          | currdrink2006 displaced1992 displaced1994
             |          | displaced1996 displaced1998 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~2004 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2006 displaced1992 displaced1994
             |          | displaced1996 displaced1998 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
currdr~2006 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2004 displaced1992 displaced1994
             |          | displaced1996 displaced1998 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~1992 |          | [No missing data in estimation sample]
displa~1994 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2004 currdrink2006 displaced1992
             |          | displaced1996 displaced1998 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~1996 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2004 currdrink2006 displaced1992
             |          | displaced1994 displaced1998 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~1998 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2004 currdrink2006 displaced1992
             |          | displaced1994 displaced1996 displaced2000
             |          | displaced2002 displaced2004 displaced2006 v207
             |          | groupage y1 y2 y3 sex occupation hyper heart
diabetes
             |          | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~2000 | logit    | currdrink1992 currdrink1994 currdrink1996
             |          | currdrink1998 currdrink2000 currdrink2002
             |          | currdrink2004 currdrink2006 displaced1992
```

```
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2002 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~2002 | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2000 currdrink2002
            |         | currdrink2004 currdrink2006 displaced1992
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2000 displaced2004 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~2004 | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2000 currdrink2002
            |         | currdrink2004 currdrink2006 displaced1992
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2000 displaced2002 displaced2006 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
displa~2006 | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2000 currdrink2002
            |         | currdrink2004 currdrink2006 displaced1992
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2000 displaced2002 displaced2004 v207
            |         | groupage y1 y2 y3 sex occupation hyper heart
diabetes
            |         | pdrinker mhscore s1 s2 s3 s4 race newincome
       v207 |         | [No missing data in estimation sample]
   groupage |         | [No missing data in estimation sample]
         y1 | ologit  | [No missing data in estimation sample]
         y2 | ologit  | [No missing data in estimation sample]
         y3 | ologit  | [No missing data in estimation sample]
        sex |         | [No missing data in estimation sample]
 occupation | logit   | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2000 currdrink2002
            |         | currdrink2004 currdrink2006 displaced1992
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2000 displaced2002 displaced2004
            |         | displaced2006 v207 groupage y1 y2 y3 sex hyper
heart
            |         | diabetes pdrinker mhscore s1 s2 s3 s4 race
newincome
      hyper |         | [No missing data in estimation sample]
      heart |         | [No missing data in estimation sample]
   diabetes |         | [No missing data in estimation sample]
   pdrinker |         | [No missing data in estimation sample]
    mhscore |         | [No missing data in estimation sample]
         s1 | ologit  | [No missing data in estimation sample]
         s2 | ologit  | [No missing data in estimation sample]
         s3 | ologit  | [No missing data in estimation sample]
         s4 | ologit  | [No missing data in estimation sample]
```

```
       race |         | [No missing data in estimation sample]
  newincome | regress | currdrink1992 currdrink1994 currdrink1996
            |         | currdrink1998 currdrink2000 currdrink2002
            |         | currdrink2004 currdrink2006 displaced1992
            |         | displaced1994 displaced1996 displaced1998
            |         | displaced2000 displaced2002 displaced2004
            |         | displaced2006 v207 groupage y1 y2 y3 sex
occupation
            |         | hyper heart diabetes pdrinker mhscore s1 s2 s3
s4 race
----------------------------------------------------------------------

End of dry run. No imputations were done, no files were created.
```

Appendix D

Missing Value Imputation Diagnostics

****Frequency of missing values in variable currdrink, observed data**

```
. tabmiss  currdrink1992 currdrink1994 currdrink1996 currdrink1998
currdrink2000 currdrink2002 currdrink2004 currdrink2006

    Variable |     Obs       Missings   Feq.Missings    NonMiss
Feq.NonMiss
-----------+-------------------------------------------------------------
currdri~1992 |    4334           0            0           4334          100
currdri~1994 |    4334         378        8.722           3956        91.28
currdri~1996 |    4334         565        13.04           3769        86.96
currdri~1998 |    4334         733        16.91           3601        83.09
currdri~2000 |    4334         916        21.14           3418        78.86
currdri~2002 |    4334        1029        23.74           3305        76.26
currdri~2004 |    4334        1158        26.72           3176        73.28
currdri~2006 |    4334        1292        29.81           3042        70.19
```

****Frequency of missing values in variable currdrink, imputed data**

```
. tabmiss  currdrink1992 currdrink1994 currdrink1996 currdrink1998
currdrink2000 currdrink2002 currdrink2004 currdrink2006

    Variable |     Obs       Missings   Feq.Missings    NonMiss
Feq.NonMiss
----------+-------------------------------------------------------------
currdri~1992 |   43340           0            0          43340          100
currdri~1994 |   43340           0            0          43340          100
currdri~1996 |   43340           0            0          43340          100
currdri~1998 |   43340           0            0          43340          100
currdri~2000 |   43340           0            0          43340          100
currdri~2002 |   43340           0            0          43340          100
currdri~2004 |   43340           0            0          43340          100
currdri~2006 |   43340           0            0          43340          100
```

Appendix E

STATA Output of GEE analysis of Complete Cases, Unadjusted and Adjusted

**GEE: COMPLETE CASE ANALYSIS, UNADJUSTED

```
. svyset v13 [pweight=v15], strata (v11)

      pweight: v15
          VCE: linearized
  Single unit: missing
     Strata 1: v11
         SU 1: v13
        FPC 1: <zero>


. xtset hhidpn year
       panel variable:  hhidpn (strongly balanced)
        time variable:  year, 1992 to 2006, but with gaps
                delta:  1 unit


. xtgee currdrink displaced, i(hhidpn) family(binomial) link(logit) cor(exc)
robust

Iteration 1: tolerance = .16555279
Iteration 2: tolerance = .00051904
Iteration 3: tolerance = 5.088e-06
Iteration 4: tolerance = 4.172e-08

GEE population-averaged model                   Number of obs      =     13381
Group variable:                      hhidpn     Number of groups   =      2633
Link:                                 logit     Obs per group: min =         1
Family:                            binomial                    avg =       5.1
Correlation:                   exchangeable                    max =         8
                                                Wald chi2(1)       =      8.73
Scale parameter:                          1     Prob > chi2        =    0.0031

                                 (Std. Err. adjusted for clustering on hhidpn)
------------------------------------------------------------------------------
             |             Semi-robust
   currdrink |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   displaced |  -.2230685   .0755124    -2.95   0.003      -.37107    -.075067
       _cons |    .421243   .0347892    12.11   0.000     .3530574    .4894286
------------------------------------------------------------------------------


. xtgee, eform

GEE population-averaged model                   Number of obs      =     13381
Group variable:                      hhidpn     Number of groups   =      2633
Link:                                 logit     Obs per group: min =         1
Family:                            binomial                    avg =       5.1
Correlation:                   exchangeable                    max =         8
                                                Wald chi2(1)       =      8.73
Scale parameter:                          1     Prob > chi2        =    0.0031

                                 (Std. Err. adjusted for clustering on hhidpn)
------------------------------------------------------------------------------
             |             Semi-robust
```

```
   currdrink | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   displaced |     .80006   .0604144    -2.95   0.003     .6899956    .9276813
-------------------------------------------------------------------------------
```

**GEE: COMPLETE CASE ANALYSIS, ADJUSTED

```
. xtgee currdrink displaced groupage marital sex race occupation newincome v207
hyper heart diabetes pdrinker mhscore, i(hhidpn) family(binomial) link(logit)
cor(exc) robust

Iteration 1: tolerance = .29534786
Iteration 2: tolerance = .00588614
Iteration 3: tolerance = .00015194
Iteration 4: tolerance = 3.619e-06
Iteration 5: tolerance = 8.518e-08

GEE population-averaged model                   Number of obs      =     12709
Group variable:                      hhidpn     Number of groups   =      2500
Link:                                 logit     Obs per group: min =         1
Family:                           binomial                     avg =       5.1
Correlation:                  exchangeable                     max =         8
                                                Wald chi2(13)      =    233.01
Scale parameter:                          1     Prob > chi2        =    0.0000

                                 (Std. Err. adjusted for clustering on hhidpn)
-------------------------------------------------------------------------------
             |             Semi-robust
   currdrink |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   displaced |  -.2340451   .0827871    -2.83   0.005    -.3963048   -.0717854
    groupage |   .0525301   .0552229     0.95   0.341    -.0557048    .1607649
     marital |  -.0115904   .0905517    -0.13   0.898    -.1890684    .1658877
         sex |  -.4091784   .0976237    -4.19   0.000    -.6005173   -.2178394

        race |   -.334787   .0632178    -5.30   0.000    -.4586916   -.2108825
  occupation |  -.1711959   .0937098    -1.83   0.068    -.3548638     .012472
   newincome |   .1530999   .0330051     4.64   0.000      .088411    .2177887
        v207 |   .0807268   .0158426     5.10   0.000     .0496759    .1117777
       hyper |  -.0854101    .078228    -1.09   0.275    -.2387342    .0679139
       heart |  -.2012713   .1296039    -1.55   0.120    -.4552903    .0527478
    diabetes |  -.6342311   .1380979    -4.59   0.000    -.9048981   -.3635641
    pdrinker |   .1330151   .1229643     1.08   0.279    -.1079905    .3740206
     mhscore |   .0669776   .0508918     1.32   0.188    -.0327685    .1667236
       _cons |  -.3992077   .2715868    -1.47   0.142    -.9315081    .1330928
-------------------------------------------------------------------------------


. xtgee, eform

GEE population-averaged model                   Number of obs      =     12709
Group variable:                      hhidpn     Number of groups   =      2500
Link:                                 logit     Obs per group: min =         1
Family:                           binomial                     avg =       5.1
Correlation:                  exchangeable                     max =         8
```

```
                                                  Wald chi2(13)      =    233.01
Scale parameter:                            1     Prob > chi2        =    0.0000

                                   (Std. Err. adjusted for clustering on hhidpn)
------------------------------------------------------------------------------
             |               Semi-robust
   currdrink | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   displaced |   .7913261   .0655116    -2.83   0.005     .6728016    .9307306
    groupage |   1.053934   .0582013     0.95   0.341     .9458183    1.174409
     marital |   .9884765   .0895082    -0.13   0.898     .8277299    1.180441
         sex |   .6641957   .0648412    -4.19   0.000     .5485278    .8042546
        race |   .7154905   .0452317    -5.30   0.000     .6321102    .8098692
  occupation |   .8426565   .0789652    -1.83   0.068      .701269     1.01255
   newincome |   1.165441   .0384655     4.64   0.000     1.092437    1.243324
        v207 |   1.084075   .0171745     5.10   0.000      1.05093    1.118264
       hyper |   .9181356   .0718239    -1.09   0.275     .7876242    1.070273
       heart |   .8176906   .1059759    -1.55   0.120     .6342638    1.054164
    diabetes |   .5303431   .0732393    -4.59   0.000     .4045831    .6951942
    pdrinker |   1.142267   .1404581     1.08   0.279     .8976362    1.453567
     mhscore |   1.069271   .0544171     1.32   0.188     .9677626    1.181428
------------------------------------------------------------------------------
```

Appendix F

STATA Output of GEE analysis of Multiple Imputation, Unadjusted and Adjusted

```
. mim:  xtgee currdrink displaced, i(hhidpn) family(binomial) link(logit)
cor(exc) robust
```

```
Multiple-imputation estimates (xtgee)                    Imputations =      10
                                                         Minimum obs =   34672
                                                         Minimum dof =    15.8

------------------------------------------------------------------------------
   currdrink |     Coef.  Std. Err.     t    P>|t|    [95% Conf. Int.]     FMI
-------------+----------------------------------------------------------------
   displaced |  -.176541   .069112   -2.55   0.021   -.323182   -.0299   0.759
       _cons |   .186686   .025417    7.34   0.000    .136802   .23657   0.027
------------------------------------------------------------------------------


. mim:  xtgee currdrink displaced groupage marital sex race occupation
newincome v207 hyper heart diabetes pdrinker mhscore, i(hhidpn)
family(binomial) link(logit) cor(exc) robust

Multiple-imputation estimates (xtgee)                    Imputations =      10
                                                         Minimum obs =   34672
                                                         Minimum dof =    15.4

------------------------------------------------------------------------------
   currdrink |     Coef.  Std. Err.     t    P>|t|    [95% Conf. Int.]     FMI
-------------+----------------------------------------------------------------
   displaced |  -.176242    .07774   -2.27   0.038    -.34158 -.010904   0.768
    groupage |  -.031059    .03857   -0.81   0.421   -.106776  .044658   0.050
     marital |   .043955    .06886    0.64   0.524   -.091627  .179537   0.157
         sex |  -.484291   .069907   -6.93   0.000   -.621675 -.346907   0.102
        race |  -.332403   .043629   -7.62   0.000   -.418029 -.246776   0.027
  occupation |  -.206352   .069466   -2.97   0.003   -.342877 -.069828   0.104
   newincome |   .126303   .022368    5.65   0.000    .082154  .170452   0.208
        v207 |   .096189     .0111    8.67   0.000    .074402  .117976   0.038
       hyper |  -.126634   .055386   -2.29   0.023   -.235396 -.017872   0.068
       heart |  -.188668   .089076   -2.12   0.035   -.363587 -.013749   0.068
    diabetes |  -.670168   .095907   -6.99   0.000   -.858975  -.48136   0.153
    pdrinker |   .084967   .084332    1.01   0.314   -.080635  .250569   0.068
     mhscore |   .070327   .036002    1.95   0.051   -.000354  .141009   0.054
       _cons |  -.563424   .200525   -2.81   0.005   -.957628 -.169219   0.113
------------------------------------------------------------------------------
```