A Simpler Model for the Impact of Atrial Fibrillation on Cognitive Trajectories in the Elderly

LaNae Schaal

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

2015

Committee:

Barbara McKnight

Mary Lou Thompson

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

A Simpler Model for the Impact of Atrial Fibrillation on Cognitive Trajectories in the Elderly

LaNae Schaal

Chair of the Supervisory Committee:
Professor Barbara McKnight
Biostatistics

In their 2013 paper "Atrial fibrillation and cognitive decline: A longitudinal cohort study", Dr. Thacker and colleagues investigated whether, in the absence of clinical stroke, incident atrial fibrillation was associated with faster cognitive decline in the elderly. They performed regression analyses on longitudinally collected data. When developing the regression model they relied on both prior scientific knowledge and the results of data driven model selection techniques. As a result, the regression model used thirty-nine terms to model the impact of thirteen covariates on cognition. In this paper we first explain the statistical and scientific rational for the terms in their regression model. Then through two simulation studies we investigate the consequences of having chosen a simpler but possibly misspecified model. Ultimately when deciding to utilize a simpler, possibly misspecified, model the researcher must

settle the question of whether a small bias is acceptable when the benefits will be greater precision and ease of communication with non-statistically minded readers.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

DEDICATION

To my supportive husband, Mark

Chapter 1. INTRODUCTION

In their 2013 paper "Atrial fibrillation and cognitive decline: A longitudinal cohort study", Dr. Thacker and colleagues investigated whether, in the absence of clinical stroke, incident atrial fibrillation was associated with faster cognitive decline in the elderly. They performed regression analyses on longitudinally collected data. To build their regression model they relied on decisions that were based on prior scientific knowledge and but also some decisions were data driven. As a result, the regression model used thirty-nine terms to model the impact of thirteen covariates on the outcome. Through two simulation studies this paper investigates the consequences of having chosen a simpler but possibly misspecified model to analyze the data. Ultimately when deciding to utilize a simpler possibly misspecified model the researcher must settle the question of whether a small bias is acceptable when the benefits will be greater precision and ease of communication with non-scientifically minded readers.

The paper is organized in two main parts: 1) rationale for the model that Dr. Thacker and colleagues developed, and 2) simulation studies comparing the performance of a simpler misspecified analysis model. There were many choices involved in the research used to address Dr. Thacker and colleagues' scientific question. Part 1 contains four chapters to provide the reader with background knowledge to understand Dr. Thacker's choice of analysis and specific regression model. Part 1 Chapter 2 focuses on how study design influenced his analysis. Part 1 Chapter 3 gives background about the broad concept of regression analysis and guides the reader to understand the choice of the specific regression analysis method "Mixed Effects Linear Regression". Part 1 Chapter 4 provides scientific rationale for the covariates and introduces notation relevant to the regression modeling. Part 1 Chapter 5 concludes with the explanation of

the modeling choice that was data driven: modeling age as a fifth degree polynomial rather than linearly.

The two chapters in Part 2 focus on the simulation studies. The aim of the simulation studies was to investigate how making a slight alteration to the analysis done by Dr. Thacker and colleagues' would alter the conclusions drawn from the research. Specifically the same analysis method (Mixed Effects Linear Regression) that was justified in Part 1 was retained. However, the regression model was simplified to model time as a quadratic rather than a quintic polynomial. In two separate simulation scenarios, subject observations were simulated to match characteristics of the actual sample that Dr. Thacker and colleagues analyzed. The main difference between the two simulation studies was the model used to originally generate the data. In Simulation 1, the more complicated modeling of time using up to a fifth degree polynomial was used to generate cognition scores. In Simulation 2, the simpler model using only linear and quadratic terms for time was used to generate cognition scores. The subsequent data analyses and contrasts of interest remained identical in the two simulation studies.

Part 2 Chapter 6 gives the reader the rationale and details of the data generation. With only minor alterations the data generation process was similar for the two simulation studies. Chapter 6 also provides information about the implementation of the data analysis. It contains sufficient information so that the reader could replicate the research. Part 2 Chapter 7 presents the results of the two simulation studies. Two different data generation models were used to allow for a variety of model misspecification scenarios. Would the benefits/drawbacks of the analysis using a simpler model of time be similar in the different data generation scenarios? The final chapter of the thesis discusses this question and presents the findings.

Chapter 2. CHARACTERISTICS OF STUDY DESIGN

2.1 CHARACTERISTICS OF AN OBSERVATIONAL, LONGITUDINAL STUDY DESIGN

The data set that Thacker and his colleagues analyzed was a subset of data from the Cardiovascular Health Study (CHS). The CHS was an observational, community-based, longitudinal study of risk factors for coronary heart disease and stroke in adults aged 65 and older (Fried, et al., 1991). The adjectives (observational and longitudinal) each identify details important for the choice of an analysis model. One main difficulty in doing observational research rather than experimental research is that the scientist is not able to manipulate subjects with respect to the levels of the variable of interest. Without the ability to assign subjects to have an incident atrial fibrillation (AF) or not, randomization could not be relied upon to make the comparison groups equal with respect to potential confounders. Therefore, to ensure that any estimate of an association between atrial fibrillation and cognitive decline would be unbiased the unequal distribution of any confounders across the groups defined by incident AF must be accounted for in the analysis. Chapter 4 will provide more detail on how Dr. Thacker and his colleagues' chosen analysis accommodated this difficulty of an observational study.

Dr. Thacker and his colleagues were researching a question that required contrasting the association between current age and cognition in groups of people without an AF with the association between current age and cognition in people who had experienced an AF. Utilizing data collected from a longitudinal study design was important to ensure that in each group defined by AF status the appropriate relationship between current age and cognition would be characterized. There were two potential choices of study design that could have been used. Researchers using a cross-sectional sample would randomly select one group of people and take cognitive measurements once for each member of the group. The ages of this group of people

would cover the entire age range which the researchers want to make inferences about. In contrast, researchers using a longitudinal study design would sample a group of people and take multiple cognition measurements on each individual in the group. The initial ages of the members in a group for a longitudinal study do not necessarily span the entire age range of interest. To obtain measurements for the entire age range, this group of people is followed forward in time until the age range of interest was captured. Repeated cognition measurements would be taken on each individual in the sample as they aged. Both study designs would allow the pattern of mean cognitive outcomes at different ages to be compared across the groups defined by AF event status. However, both study designs would not necessarily produce similar patterns for how mean cognition and current age were related. Work done by Schaie (2005) with The Seattle Longitudinal Study of Adult Development (SLS), illustrated that the study's design can lead to different conclusions about how cognition is related to current age.

2.2 CONTRASTING CROSS-SECTIONAL AND LONGITUDINAL STUDY DESIGNS: THE SEATTLE LONGITUDINAL STUDY OF ADULT DEVELOPMENT

The aging research of the SLS began in 1956 as a cross-sectional investigation about how cognitive abilities changed over the age range of 25 to 81 years old. Seven years later, in 1963, the study was converted to a multi-cohort longitudinal study. Follow-up of the original cohort continued in seven year intervals. At each observation six different attributes of normal cognition (identified in Figure 2-1 and Figure 2-2) were studied. In the cross-sectional data collected in 1956, four of the six cognitive attributes showed a consistent statistically significant negative linear trend for differences in ages between 25 and 81 years (Figure 2-1)[1]. While in the

[1] Figure 2-1 and Figure 2-2 are directly excerpted from Schaie (2005).

data from the longitudinal study all six cognitive factors are relatively stable from ages 25 until 60; at varying ages after 60, the six cognitive factors begin to demonstrate a consistent statistically significant negative linear trend (Figure 2-2). As Schaie (2005) noted, "The cross-sectional data clearly misinform if they are to be taken as estimates of decline in cognitive abilities". A person's cognitive abilities did not decline steadily from age 25 as the trend in the cross-sectional data indicated. By following people longitudinally, he saw that as people aged they tended to remain at their current level of cognition from age 25 until they reached 60 years old and then started to decline. The nature of a cross-sectional study design allowed this method to detect a trend in the relationship between current age and cognition that was confounded by the "cohort effect", the change in collective cognitive abilities of a generation due to forces other than natural aging, such as differences in access to education.

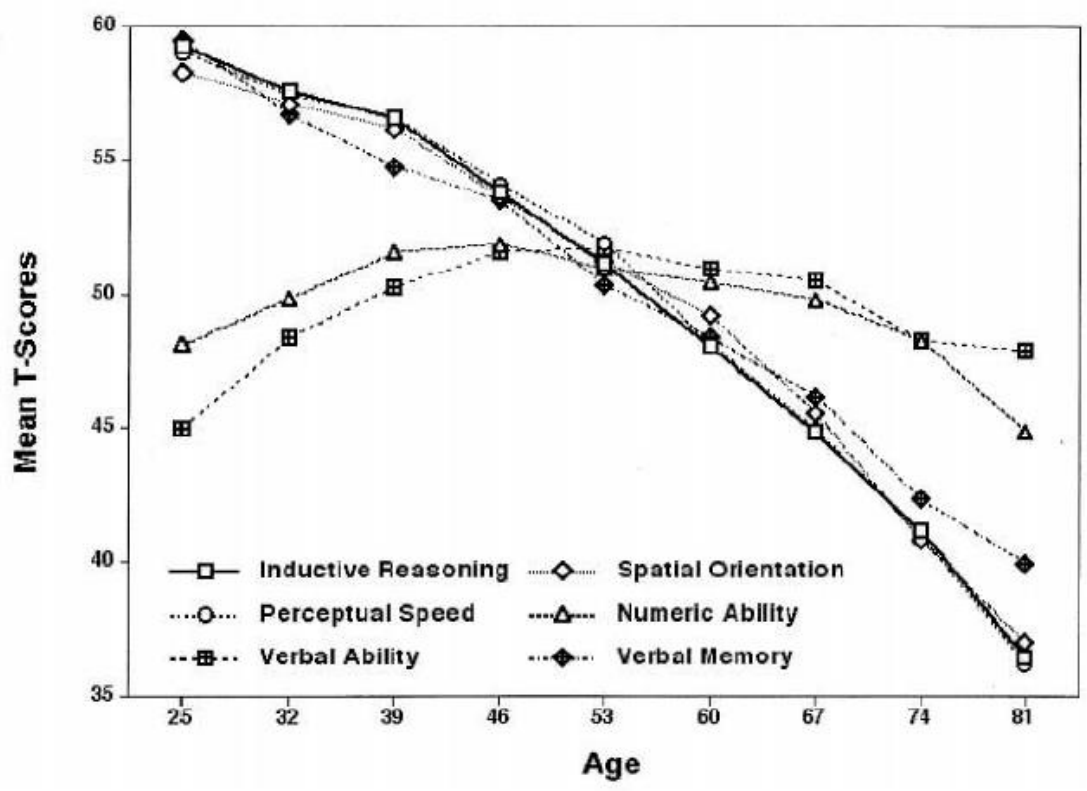

Figure 2-1 Age-difference changes in six cognitive characteristics from cross-sectional data as demonstrated in Schaie (2005).

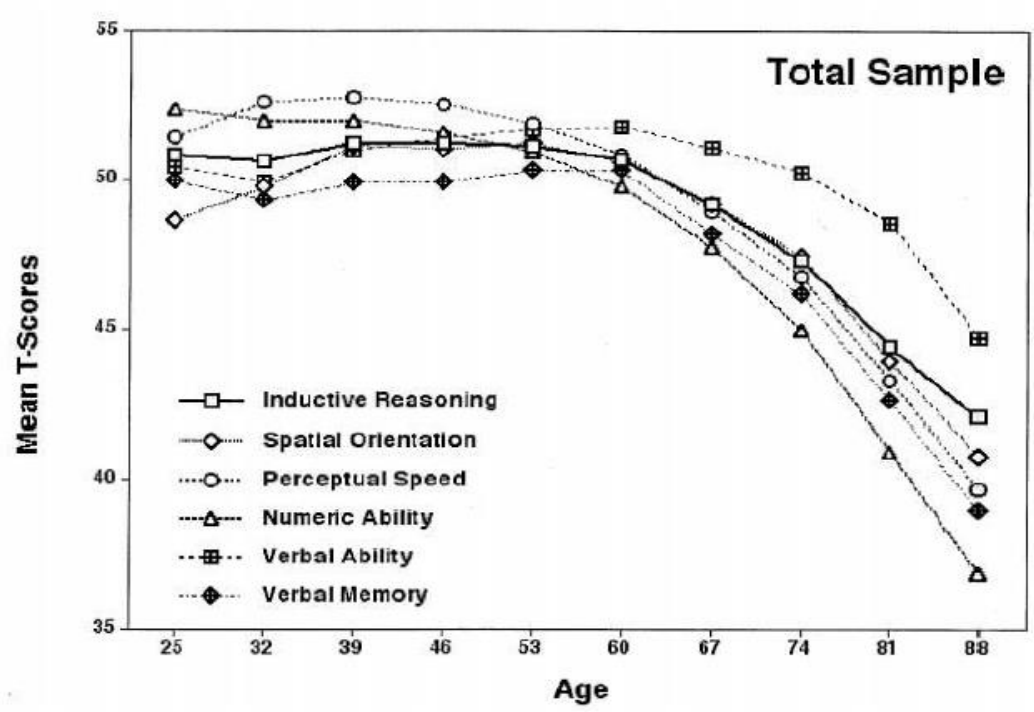

Figure 2-2 Estimated average intraindividual changes for six cognitive abilities from longitudinal data with 7-years of follow-up as reported in Schaie (2005).

For the analysis of the original cross-sectional data from 1956 people were independently sampled from different birth years. The group of thirty year olds were sampled from the population of individuals born in 1926, thirty-five year olds were sampled from the population born in 1921, and so on. The data in a sample using a cross-sectional design captured the cognitive abilities of each birth year at only one age level. Similar to the graphical representation of data used by Baltes (1968), I have created a hypothetical depiction to demonstrate that having only one measurement per birth cohort misrepresents the association between cognition and growing older by confusing the influences of non-age related generational factors and the effects of natural aging (Figure 2-3). The colored lines, cognitive trajectories, represent cognition measurements taken at different ages from people born in the same year.

When generating each trajectory, the intercept was shifted one cognition point for each cohort, but all other aspects used to generate the trajectories are the same. Therefore, the changes in cognition due to growing older remained the same for all birth cohort years. The different intercepts modeled the cohort effect. The black dots represent the cognitive scores that would be recorded by a cross-sectional sample of 25 to 81 year olds taken in 1956.

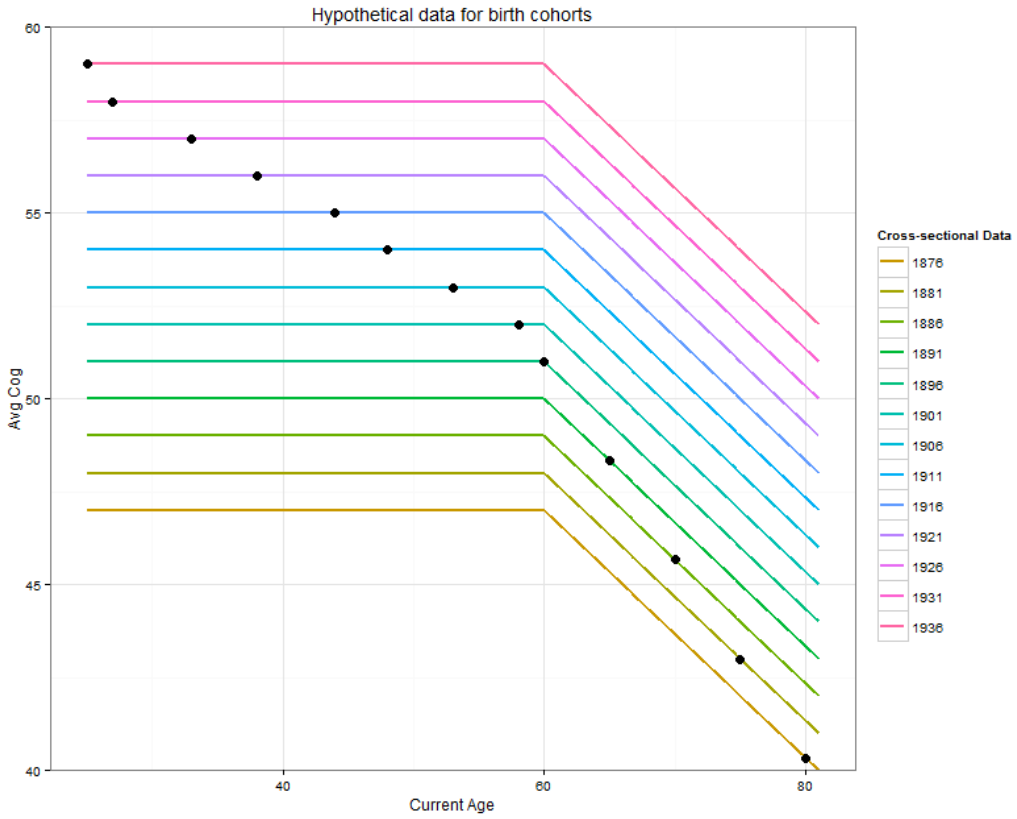

Figure 2-3 Hypothetical data from birth cohorts.

The pattern in the cross sectional dots is not parallel to the cohort trajectories; clearly the cross-sectional study is detecting a pattern in cognition that does not come from merely growing older. Prior to age 60, by design there was no change in cognition due purely to maturation. The cross-sectional association between age and cognition that was detected prior to age 60 was due solely to the differences in the baseline cognition for each birth year cohort, the cohort effect.

After age 60, the changes in cognition are due to a combination of natural aging and the cohort effect. By using a study design that can account for the cohort effect, the association that is detected between cognition and age would more closely tell the influence of natural aging. For a longitudinal study that consists of one group of people born in the same year and follows them forward in time, the cohort effect will not confound the association of current age and cognition. When the group of people entering the longitudinal study are not all of the same birth cohort, the cognitive trajectory is the average of all the trajectories from each cohort. An average of all the trajectories in Figure 2-3 would still tell the correct association between aging and changes in cognition: a flat line from ages 25 to 60 with a negative trend after age 60.

Being able to correctly characterize the changes in an outcome that are due to the passage of time is one of the main benefits of a longitudinal study design. An additional benefit of using a longitudinal study design is that we can directly study the factors that influence the changes in cognition over time. Dr. Thacker and colleagues' study was primarily interested in the factor (incident atrial fibrillation) and how it could alter the change in cognitive ability that occurs with normal aging. Thus the analysis of differences in cognitive trajectories between those who have had atrial fibrillation and those who have not had atrial fibrillation should be addressing the correct question. However, there are four other challenges encountered by using a longitudinal study design that must be recognized: selective survival, selective sampling, selective dropout, and the fact that the data consists of non-independent observations.

2.3 FOUR DESIGN RELATED CHALLENGES OF A LONGITUDINAL STUDY

The first challenge, selective survival, exists for both cross-sectional and longitudinal study designs. As defined by Baltes (1968) selective survival is the fact that "a given population at birth changes in its composition in conjunction with the aging process as a result of death or

incapacitation". Therefore, the population for a cohort born in 1925 will necessarily be different if we sample in 1956 or 1966. This type of selective bias is important if the inferential focus is on the original cohort of individuals born in 1925. Due to selective survival, the sampled individuals would no longer represent the birth cohort. However, for the CHS study the inference is not for the original birth cohort. Rather we are concerned with the effect of prior AF in live people at the time when the study started. So even though selective survival exists, it should have not caused biased results from this analysis.

The next two challenges, selective sampling and selective dropout, are specifically worrisome for longitudinal designs but can be somewhat mediated by intentional efforts in the design phase. Selective sampling creates a sample that does not represent the general population at the start of the study to whom the results should generalize. Selecting the initial sample utilizing random sampling techniques is a method to help align the sample with the general population. The CHS data were collected using random sampling methods that would cover 98% of the intended population (Fried, et al., 1991). The use of a national database and stratification by age brackets to determine who would initially be contacted was a good faith effort to obtain a representative sample. However, due to the long duration of longitudinal studies the subjects typically volunteering to participate do not represent the general population.

Statistical methods to address selective sampling during the analysis phase do exist; however, Thacker and colleagues did not report using such methods. For example, if there was demographic information about individuals who refused to participate statistical weighting can be applied to address possible sampling bias. Interested readers can find more information about these statistical methods in a paper by Höfler, Pfister, Lieb, & Wittchen (2005). Thacker and colleagues reported their results as if they were an unbiased estimate of the association in the

desired population. Reviewing the detailed demographic description of the sample of subjects who participated in the CHS study, researchers have the tools to determine for themselves whether the results would be generalizable to the correct population. These descriptions are available on the CHS website[2].

In a longitudinal study, it is likely that the sample completing the study will not be the same as those who initiated it. During the course of follow-up, subjects will cease to participate for reasons such as death, deteriorating health, relocation, financial burden, etc. While designing the study, procedures can be developed to reduce drop-outs due to scheduling difficulties, transportation issues, and other such logistics. A participant coordinator should be trained to provide reminder phone calls and to accurately track electronic data for non-clinic hospitalizations. These steps can reduce the number of people who drop-out of the study.

While minimizing all drop-outs was the goal, the most important drop-out to recognize is selective drop-out. In the CHS study selective dropout would have occurred if the characteristics of the group that dropped out were correlated with the outcome of interest: cognitive scores. In the CHS cohort many of the drop-outs occurred as a result of death. Prior to their death a large number of these subjects had severely declining cognition. Thus, it is likely that selective drop-out, which would have been informative about the distribution of the outcome, occurred in this study. Similar to selective sampling there are not consistently agreed upon analysis methods to adjust for selective drop-out. Reporting characteristics for the group of subjects who dropped out and those who completed the study provides the reader with information to determine whether they feel an analysis that did not accommodate the drop-out of subjects should still be trustworthy. In their research, Thacker and his colleagues did not report utilizing any specific

[2] https://chs-nhlbi.org/monograf/Monodocs

methods (i.e. multiple imputation of scores, or inverse probability/propensity weighting) to accommodate the missing values due to drop-outs in the study.

The final challenge introduced by using a longitudinal study design is the lack of independence in the observations. Independence of observations is a fundamental assumption to classical analysis methods. But it is expected that if we are following a group of subjects over time, the scores from one subject will be correlated because they are more similar to each other than to others' scores. The fact that the longitudinal study design necessitates some outcomes are dependent can be accommodated by choosing a correct analysis method. There are several analysis methods which have been developed to account for dependent measurements. Dr. Thacker and his colleagues chose to use a mixed effects linear regression to accommodate the dependent measurements. Chapter 3 provides more detail about this choice of analysis method.

Chapter 3. THE RATIONALE FOR USING "MIXED –EFFECTS LINEAR REGRESSION".

In their paper, Dr. Thacker and colleagues chose "mixed-effects linear regression" as the method of statistical analysis to answer the scientific question. Prior to addressing how "mixed-effects linear regression" accommodates the limitations of the observational and longitudinal study design, I will develop an understanding of what is meant by "regression". I begin by explaining the statistical concept of "regression" and clarifying some common estimation methods and regression terminology. Then in the setting of "classical linear regression" I describe how the assumptions made about the process of data collection influence whether the regression method produces "good" estimates of the association. Lastly, I identify what distinguishes "mixed-effects" linear regression from the other commonly used method, generalized estimating equations, for analysis of longitudinal data. In Chapter 4, I return to address how the "mixed effects linear regression" model chosen by Dr. Thacker and his colleagues addressed the limitations introduced by using an observational study design.

3.1 SUMMARY OF REGRESSION ANALYSIS: THE GENERAL CONCEPT AND VOCABULARY.

Regression analysis is one class of methods used to describe patterns in data. The following summary of regression analysis draws heavily on the concepts presented by Richard Berk in his 2004 book: Regression Analysis: A Constructive Critique. In the broadest sense, regression analysis describes how the conditional distribution of the response variable changes across groups defined by the values of a single predictor (or the values of a group of predictors like sex, smoking status, and years of education). For ease of exposition, I will focus on the case

where groups are defined by the levels of a single predictor. If the conditional distributions are the same in the groups, then we would say there is no association between the predictor and the response. Any difference in the conditional distributions indicates there is an association between the response and the predictor. Rather than compare every aspect of the conditional distributions of the outcome, some summary measure, θ, (e.g. the mean of the distribution) is used to describe an important characteristic of the distribution. Then a regression equation is constructed that describes how, for different values of the predictor variable, the summary measure of the conditional distribution changes. The left hand side of the equation is some function of the summary measure. The right hand (predictor) side of the equation is some function of the predictor, X, and a set of regression parameters, β. Possible regression equations[3] are: $\theta|X = \frac{\beta_1 x}{\beta_2 + x}$, $\theta|X = e^{\beta_0 + \beta_1 x}$, $\log(\theta|X) = \beta_0 + \beta_1 x$, $\log\left(\frac{\theta|X}{1-\theta|X}\right) = \beta_0 + \beta_1 x$, or $\theta|X = \beta_0 + \beta_1 x$. I have used $\theta|X$, rather than $\theta|X = x$, on the left hand side to simplify the notation. The use of lower case *x* on the right hand side indicates the specific value of X in which we are interested.

The single measure for a conditional distribution that is most commonly of scientific interest summarizes the central tendency of the distribution. The mean is often chosen as the summary measure for the central tendency. One benefit to using the mean is that it is sensitive to extreme values. If a distribution of outcomes at a particular value of a predictor contains unusually extreme values, the mean of the distribution will be changed whereas other measures of central tendency, such as the median, would not necessarily be altered. If at a given value for a predictor the conditional distribution has outliers not located at another value for the predictor,

[3]The first equation is the Michaelis–Menten model for enzyme kinetics.

the two conditional distributions are different. Comparing conditional summary measures that would be influenced by the outliers would give the correct conclusion that the predictor and the outcome are associated. Usually most regression equations use the mean as the conditional summary measure so we will set $\theta|X = E[Y|X]$.

The objective of a regression equation is to detect patterns for how differences in a predictor are related to differences in the conditional mean. For heuristic reasons, the right hand side of the regression equation is modeled using a "linear" predictor, $\beta_0 + \beta_1 x$. In this context "linear" is used only to describe the form of the right hand side of the regression equation, i.e. the regression parameters are linear. This linear predictor form is the same form as the equation of a line. The regression parameter assigned to a covariate, β_1, easily provides an estimate of the association of interest. This linear predictor form represents that the measured association is constant when comparing any two groups that have the same difference in covariate values. Comparing Group A who have $x = 4$ and Group B who have $x = 1$ we have an association that is $3\beta_1$. The same magnitude of association would result when comparing Group C who have value $x = 7$ with Group A. In a linear predictor model the parameter β_1 has the same interpretation as a "slope" and will be constant across all groups defined by covariate values that are 1 unit apart

The left hand side of the regression equation is usually some function of the conditional mean: i.e.$E[Y|X]$, $\log(E[Y|X])$, $\log\left(\frac{E[Y|X]}{1-E[Y|X]}\right)$. When it is anticipated that the conditional mean changes linearly as a function of covariates, the left-hand side of the model is untransformed: $E[Y|X]$. When the conditional mean follows a non-linear path with respect to the changes in the covariates, one way to capture this is by modeling the left hand side with a link function (e.g. $\log(E[Y|X])$, $\log\left(\frac{E[Y|X]}{1-E[Y|X]}\right)$) that allows the right hand side of the equation to model the linear path of the transformed conditional mean. Further explanation regarding the choice of link

function for the left hand side of the regression model is beyond the scope of this paper. Interested parties can research "generalized" linear models (which is distinct from "general" linear models) in the regression textbooks (Fitzmaurice, Laird, & Ware, 2011; Vittinghoff, Glidden, Shiboski, & McCulloch, 2012). For ease of exposition, the remainder of this summary of regression analysis will represent only the untransformed mean as the left hand side of the regression equation.

The linear predictor form on the right hand side provides an easy method to detect (and describe) a simple increasing or decreasing pattern for how the mean of an outcome and the levels of a predictor are associated. If the predictor variable is continuous, when comparing groups who differ in the value of X there will be a β_1 change in the value of the conditional mean for every one unit higher value of X. Therefore, a slope regression parameter that is non-zero signifies that the data indicated there was either an increasing or decreasing trend in the conditional mean.

When the predictor variable is categorical, it can be represented in the regression as either one linear variable or a set of dummy variables defined by categories of the predictor variable. Treating a categorical variable as linear is appropriate when the levels of the variable are ordered and the magnitude of change from one level to the next is expected to be the same. For this representation of a categorical variable, β_1 will again describe the trend in the outcome. Specifically it quantifies the difference in the mean outcome as you move from the lower category group to the next higher category. Using a dummy variable representation of a categorical variable requires more than one regression parameter. If there are *k* levels of a variable, there will be *k* -1 regression parameters to represent this variable. One level of the variable is omitted from the regression model because it serves as a reference group. Each of the

regression parameters $\beta_1, \beta_2, \ldots, \beta_{k-1}$ describes the difference in the conditional mean when comparing the group defined by the value of X to that preselected reference group. The dummy variable parameterization is not necessarily describing a trend in the outcome across the levels of the categorical variable.

Thus far we have explained the choices for the right and left hand side of the regression equation to mathematically summarize how the mean outcome depends on the value of the predictor. For ease of exposition I have chosen to describe the choices that lead to the regression equation $E[Y|X] = \beta_0 + \beta_1 x$. It is important to note that if a regression analysis is used to predict the mean of a group with a certain value for the covariate X, it will usually provide a modeled mean rather than the true empirical mean in the group. That is to say, regression analysis is not using the layman's definition of conditional mean, $\bar{Y}|X = \frac{\sum_{i=1}^{n} Y_i|X}{n}$, to determine the mean of the outcomes at a given value of X. The layman's definition of conditional mean restricts the conditional mean to be evaluated only using the observed outcomes for people in the group with the specific value of the predictor. A conditional mean that results from a regression model is determined using the value of the predictor, X= *x*, and the estimate of the regression parameters $\hat{\beta}$. To determine the regression parameter estimates all available outcomes in the sample, regardless of their predictor group, are utilized. When a conditional mean is determined by this process it is referred to as a "fitted" mean.

There are two methods commonly used to estimate the regression parameters: least squares (LS) and maximum likelihood (ML). The least squares method attempts to find the set of regression parameters that minimize the sum of all squared residuals; a residual is the difference between the observation and the expected (or fitted) mean: $e_i|X = y_i|X - \hat{E}[Y|X]$. The LS method of estimation assumes no specific distribution from which the observed outcomes were

drawn. It merely finds the line that best fits the observed scatterplot of data according to the LS criterion. The maximum likelihood method attempts to find the set of regression parameters that maximizes the joint likelihood of all the observed outcomes; a joint likelihood is the function that describes the parameters of a distribution as a function of observed outcomes. A likelihood is mathematically identical to a density function; the designation for the fixed and known inputs (X_i,Y_i) and the unknown distributional parameters is merely reversed. Since the distributional parameters are utilized as fixed and known values, ML estimation relies on knowing the distribution from which the outcome variables were selected. Depending on the study design and assumptions that the researcher makes, one method of parameter estimation (LS or ML) may be more appropriate. This will be discussed in further detail at the conclusion of this chapter.

Regardless of whether LS or ML is chosen as the computational method to estimate the parameters, a single estimate $\hat{\beta}$ reflects only the association between the predictor and outcome that exists in one sample of N subjects from the population. If we had taken a different sample of N people from the population, we are very likely to realize a different numerical value for $\hat{\beta}$. Some numerical values have a higher chance of being seen than others. Thus like an outcome variable Y, the parameter estimate $\hat{\beta}$ can be thought of as coming from a distribution. In large samples it is the asymptotic distribution of $\hat{\beta}$ (not the outcome) that is important for making valid statistical inference (Lumley, Diehr, Emerson, & Chen, 2002). Whether we use LS or ML estimation methods, as long as the outcomes are independent, we can invoke the Central Limit Theorem to determine that the distribution of $\hat{\beta}$ is asymptotically Normal. Knowing the asymptotic distribution of the estimate is not enough. We would also like the data analysis method that we choose to provide "good" estimates.

There are two components to describe a "good" estimator: small bias (ideally completely unbiased) and high precision. To understand each quality we consider the situation when the method of analysis has been performed on repeated samples of size N from the population. Unbiasedness is when the long-run average of the estimators falls on the true value of the association in the population; $E(\hat{\beta}) = \beta$. Precision describes the variability of these estimators; high precision indicates that the estimators cluster tightly around one value, low precision describes estimators that are spread out. The four kinds of estimators are depicted in Figure 3-1 (Krishnan, 2014). Having an analysis method that returns an unbiased, highly precise estimator is equivalent to hitting the bull's-eye. However, it is often the case that one must sacrifice unbiasedness to achieve greater precision. The act of balancing the bias and variance of an estimator is called the "bias-variance trade-off".

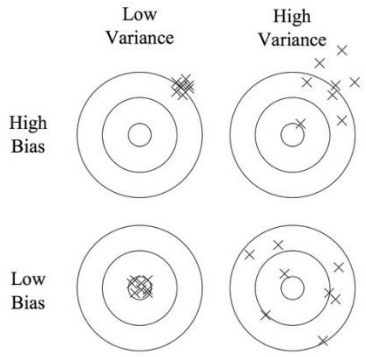

Figure 3-1 Illustrating the relationship for the concepts of bias and variance.

3.2 Assumptions of "classical linear regression" and "good" estimators

Thacker and colleagues were choosing an analysis method that would be appropriate for data with a continuous outcome. "Classical linear regression" is a very appealing model of analysis for this type of data because we can come close to hitting the bull's-eye. As traditionally taught in introductory classes, there are four unnecessarily restrictive assumptions for classical linear regression: 1) the association is correctly modeled by the linear predictor, 2) the

observations are independent, 3) errors are normally distributed about the regression line and the conditional mean of the errors is zero, and 4) the conditional variance of Y is equal across all values of the predictor. However, we can have less restrictive assumptions in a semi-parametric model that uses Least Squares estimation and still obtain a "good" estimator. In fact, using the assumptions of the Gauss-Markov theorem we find that the classical linear regression obtained by the ordinary least squares (OLS) method will be the "Best Linear Unbiased Estimator". We will obtain an estimator that is unbiased and has the greatest precision of all linear unbiased estimators.

I explain further the less restrictive assumptions for the Gauss-Markov theorem. First, it is still assumed that the linear predictor in the regression model is correct. This refers to the choice of the linear predictor which is a simple weighted sum of linear regression parameters (i.e. $\beta_0 + \beta_1 x$ rather than $\beta_0 + \beta_1^2 x \; or \; \beta_0 + \beta_0\beta_1 x$). It also refers to the fact that none of the necessary covariates have been omitted from the regression model. (It is possible that the regression model includes unnecessary covariates). Second, we still assume any two observed outcomes should be independent; knowing the value of one observation, Y_i , should not provide information about the value of another observation, Y_j. When outcomes are independent, it is a logical conclusion that they are also uncorrelated. The third assumption however is less restrictive; no specific distribution (i.e. Normal) is imposed on the conditional errors. Rather we assume that the conditional errors come from the same distribution with a mean of zero. Fourth, the variance of the residuals at each level of the predictor is constant, i.e. $Var(e_i|X\,) = \sigma^2$.

Under these assumptions, we can use the LS method to obtain an estimator for the regression parameters, $\hat{\beta}_{OLS}$. Let N represent the total number of observations (which for now is

equal to the number of subjects). The least squares method is finding the values for β_0 and β_1 that minimize:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - \hat{E}(Y|X)]^2 = \sum_{i=1}^{N} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Using the assumed linear predictor, we plug in the linear combination of the covariates to replace the conditional expectation. Then we can perform calculus and solve a system of equations which will minimize the above equation to obtain these values for β_0 and β_1:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Utilizing matrix notation to express the regression parameters makes it easier to see the two important characteristics of these regression parameter estimates: unbiasedness and the best efficiency of the estimates.

$$\text{Let } \boldsymbol{\beta} \equiv \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{X} \equiv \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{bmatrix} \text{ and } \boldsymbol{Y} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}.$$

Each row in the **X** design matrix and **Y** response vector represent the measurements from one individual in the sample. Written in condensed matrix notation the estimates for the regression parameters is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$.

The unbiasedness of these regression parameter estimates results from the facts that we have conditioned on the values of X, the conditional errors have mean 0, and the assumption that the linear model correctly modeled and included all necessary covariates (i.e. the covariate values are uncorrelated with the errors). When evaluating the expectation, the design matrix values are treated as fixed constants but the outcome is allowed to be random. The randomness in the outcome is reflected through the error term being the random variable. Thus we are taking the expectation with respect to the conditional error.

$$E_{e|X}[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}] = E_{e|X}[(\boldsymbol{X'X})^{-1}(\boldsymbol{X'Y})] = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}E_{e|X}[\boldsymbol{Y}|\boldsymbol{X}] = (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta} = \boldsymbol{\beta}.$$

$$\text{where } E_{e|X}[\boldsymbol{Y}|\boldsymbol{X}] = E_{e|X}[\boldsymbol{X\beta} + \boldsymbol{e}|\boldsymbol{X}] = \boldsymbol{X\beta} + \boldsymbol{E}_{e|X}[\boldsymbol{e}|\boldsymbol{X}] = \boldsymbol{X\beta}$$

From the above expectation, it is easy to see the necessity to assume that the conditional errors have a mean of zero. Without this assumption being true, we have an estimator that will give biased estimates.

To emphasize the importance of having the correct covariates in the model, we consider the case of a model that has omitted important variables. If we had incorrectly omitted a necessary variable from the regression model then we could summarize the situation with two regression models:

True Population model: $Y_i = \beta_0^* + \beta_1^* x + \delta z + \varepsilon_i$ where $\varepsilon_i|X, Z \sim (0, \sigma_p^2)$

Model resulting from fitting the sample: $Y_i = \beta_0 + \beta_1 x + e_i$

$$\text{where } e_i = \delta z + \varepsilon_i; \;\; e_i|X, Z \sim (\delta z, \sigma_s^2)$$

In the population there is an association between Z and Y, so $\delta \neq 0$. When we fit a model that does not allow for a parameter to estimate the association between Z and Y, we force that true non-zero relationship to be absorbed into the error component of the model. The new conditional error term may no longer have a mean of zero. We have violated an assumption of the "classical linear regression" model.

Omitting a variable from the model has an additional effect of altering the interpretation of the parameters in the two models. In the smaller model we are comparing groups of people defined only by their values of X. In the truth from the population, β_1^*, is comparing groups of people who differ by 1 in their value of X among those who are similar with respect to Z. When we have a smaller model that has omitted predictive variables the regression estimates will be answering a different scientific question than if we had the larger model.

As stated earlier if we are using a misspecified model to estimate the association between X and Y the conditional errors may no longer have a mean of zero. Our parameter estimator based on the variables observed in the sample remains the same $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$. However, this estimator is now biased when attempting to answer the question about the association between X and Y after controlling for the effects of Z. When taking the expectation we must determine the expected value **in the population** and use the true population model when substituting for $\boldsymbol{Y}|\boldsymbol{X}$**.**

$$E_{e|X,Z}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}\right] = E_{e|X,Z}[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E_{e|X,Z}[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}]$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}^{*} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\delta} = \boldsymbol{\beta}^{*} + bias$$

where $E_{e|X,Z}[\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}] = E_{e|X,Z}[\boldsymbol{X}\boldsymbol{\beta}^{*} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{e}|\boldsymbol{X},\boldsymbol{Z}] = \boldsymbol{X}\boldsymbol{\beta}^{*} + \boldsymbol{Z}\boldsymbol{\delta} + \boldsymbol{E}_{e|X,Z}[\boldsymbol{e}|\boldsymbol{X},\boldsymbol{Z}\,] = \boldsymbol{X}\boldsymbol{\beta}^{*} + \boldsymbol{Z}\boldsymbol{\delta}$

Returning to the situation where we have an unbiased estimator (i.e. assumptions 1 and 3 have been met), we can see how the additional assumptions lead to the OLS estimator being the most precise estimator of all unbiased estimators. To determine the variability of the regression parameter estimates we again treat the design matrix values as fixed and rely on the assumed covariance matrix of the outcome. From the assumptions that the conditional variance was constant and the observations were independent $Var[\boldsymbol{Y}] = \boldsymbol{\sigma^2 I} \equiv \boldsymbol{\Sigma}$**.** Therefore,

$$Var\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}\right] = Var[(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})] = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']Var[\boldsymbol{Y}][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\boldsymbol{\sigma^2 I}]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{\sigma^2}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Professor Flinn (2004) from New York University provides a proof of the Gauss- Markov theorem that the OLS estimator will have the smallest variance of all linear unbiased estimators. The details of this proof can be found in Appendix A. The steps, in brief, are: 1) Assume the existence of another unbiased estimator that is a linear function of **Y** (i.e. $\tilde{\beta} = m + My$) , 2) Use

the fact that the estimator must be unbiased to make conclusions about the values for m and M, 3) Compute the covariance matrix for the estimator $\tilde{\beta}$, and 4) Show that Var($\tilde{\beta}$) is the sum of Var$[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}]$ and a matrix that must consist of all positive terms. Therefore when the less restrictive assumptions of the Gauss-Markov theorem are met the estimator from LS method will be "best". It is important to note that we have obtained the "best" estimator without making any assumption that the outcomes (or the residuals) were normally distributed.

For classical linear regression, when we do add in the assumption of a normal distribution both LS and ML methods produce identical estimates for the regression parameters. Remember that if we want to utilize the ML method to estimate the regression parameters, we must assume a distribution for the outcomes. Assuming the outcomes follow a MVN ($\beta_0 + \beta_1 x$, $\sigma^2 I$), the ML method is finding the values for β_0 and β_1 that maximize the likelihood:

$$L(\beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{-N/2} \exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{N}[y_i - (\beta_0 + \beta_1 x_i)]^2\right).$$

This is equivalent to minimizing the sum in the exponent which is the identical problem that the LS method is solving. When the assumptions are true, we see that for linear regression either the LS or the ML method will produce regression parameter estimates that are unbiased, have the smallest variance of all unbiased estimators, and have valid statistical inference. Despite all of these qualities, Thacker and his colleagues chose a different analysis method than "classical linear regression".

3.3 WHY THE ASSUMPTIONS OF CLASSICAL LINEAR REGRESSION ARE NOT APPROPRIATE FOR THIS STUDY.

Regression estimates based on "classical linear regression" assumptions would not be appropriate due to the design of the study. The decision to use a longitudinal study design

allowed the data which was collected to appropriately address the scientific question but introduced a violation of the "classical linear regression" assumptions. Only from a longitudinal study design could repeated measures be obtained from subjects that would allow the construction of subject specific cognitive trajectories. Then the average trajectories in groups defined by AF status could be compared to address the scientific question of Dr. Thacker and his colleagues. However, using repeated measures over time on the same subjects, restricted the type of analysis that could be done on the data. By following the same people forward in time, the complete set of data at the end of the study would no longer be a set of independent measurements. For a single subject we would expect measurements taken at different ages to be somewhat similar. While measurements taken on different subjects would still be expected to be independent. Since the assumption of uncorrelated observations was violated, they needed an analysis method that appropriately accounted for the non-independence of the data.

The previously identified vector notation for the entire set of data will be redefined to correctly identify which measurements are from the same subject and thus possibly correlated. In addition, a new notation will be introduced that isolates the data specific to one subject.

$$\text{Let } \mathbf{X} \equiv \begin{bmatrix} 1 & x_{11} \\ \vdots & \vdots \\ 1 & x_{1n_1} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{2n_2} \\ \vdots & \vdots \\ 1 & x_{i1} \\ \vdots & \vdots \\ 1 & x_{in_i} \\ \vdots & \vdots \\ 1 & x_{Nn_N} \end{bmatrix}, \boldsymbol{Y} \equiv \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{i1} \\ \vdots \\ y_{in_i} \\ \vdots \\ y_{Nn_N} \end{bmatrix}, \boldsymbol{X_i} \equiv \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{ij} \\ \vdots & \vdots \\ 1 & x_{in_i} \end{bmatrix} \text{ and } \boldsymbol{Y_i} \equiv \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{in_i} \end{bmatrix}.$$

The new notation is explained as follows. A subscript *i* indicates the subject for whom the associated covariate or outcome was measured. The subscript *j* represents the ordinal position of

the measurement (i.e. first, second, etc.). The subscript n_i represents the total number of observations for the i[th] subject. The use of a subscript *i* for the total number of observations indicates that each subject may have a different amount of total observations. The design matrix and outcome vector without a subscript *i* contain a total of T measurements from the entire study. Some of the T outcome measurements remain independent (e.g. (y_{11},y_{21})) while others will be dependent (e.g. (y_{31},y_{32})). Specifically for each subject, his or her n_i measurements are not independent and any analysis needs to account for this correlation to have the correct variance of the regression parameters.

Both the LS and the ML method can be altered to accommodate the correlation between observations. The estimate for the regression parameter is modified by including a non-identity T x T covariance matrix, Ω. The least squares method that utilizes this new covariance matrix is called "generalized least squares" (GLS). [In practice, we do not usually make assumptions about the entire covariance matrix. Rather this covariance matrix is separated into a product of a variance and a correlation matrix $\Omega = \sigma^2\Sigma$ where Σ is the correlation matrix. Then assumptions are made about the correlation matrix. This is explained further in Section 3.4]. In theory, when the correlation matrix is assumed to be known, the regression parameter estimates from GLS becomes

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = (\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\Omega^{-1}\boldsymbol{Y}) = (\boldsymbol{X}'(\sigma^2\Sigma)^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'(\sigma^2\Sigma)^{-1}\boldsymbol{Y})$$
$$= (\boldsymbol{X}'(\Sigma)^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'(\Sigma)^{-1}\boldsymbol{Y}).$$

Note because of the dependence on the correlation this estimate will be different from the one derived previously: $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})$. The ML estimate will again be the same as the GLS estimate when we assume a multivariate normal distribution for the outcome that is parameterized using the known correlation matrix . A sketch of the derivation of the estimate

using both the GLS and ML methods is provided in Appendix B. (In the case when the data are not complete it is not necessarily true that GLS and ML methods provide the same parameter estimates. How the missingness in the data influences the estimates that result from the two estimation methods will be discussed in Chapter 3 Section 5).

The elements in the covariance matrix Ω, which describes the covariance of the data from all subjects at all observation time points and is known up to a constant scale factor, follow a pattern that is easily summarized after first introducing a covariance matrix that describes only the data for each subject Ω_i. The diagonal entries of Ω_i are the variances of the populations from which the *jth* outcome of subject *i* is drawn. Based on their empirical observations from many longitudinal studies, Fitzmaurice et al. (2011) recommend representing non-constant variances for outcomes that are drawn at different observation times. During data analysis we would allow for this heteroscedasticity by using the Huber White Sandwich estimator to generate robust standard errors. For the purposes of this discussion we will continue to assume homoscedasticity and use constant variance on the diagonals. The off-diagonal entries describe the covariance for any pair of outcomes from different observation times $(Y_{i,p}, Y_{i,q}): p \neq q$ and $p, q \in \{1,2, \dots n_i\}$.

$$\Omega_i = \begin{bmatrix} \sigma_i^2 & \sigma_{i,1,2} & \dots & \sigma_{i,1,n_i} \\ \sigma_{i,2,1} & \sigma_i^2 & \dots & \sigma_{i,2,n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{i,n_i,1} & \sigma_{1,n_i,2} & \dots & \sigma_i^2 \end{bmatrix} = \sigma_i^2 \begin{bmatrix} 1 & \rho_{i,1,2} & \dots & \rho_{i,1,n_i} \\ \rho_{i,2,1} & 1 & \dots & \rho_{i,2,n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{i,n_i,1} & \rho_{1,n_i,2} & \dots & 1 \end{bmatrix}$$

The entries of the covariance matrix for the set of all observations in the study, Ω, are easiest to understand when we consider it as a partitioned matrix. We subdivide the entire matrix into cells that describe the covariance of groups of data defined by subject. Along the diagonal of the subdivided covariance matrix we are describing how measurements vary within an individual across observations. Thus the diagonal entries are all the subject specific covariance

matrices. The off-diagonal partitions are describing how observations from different subjects vary together. Since each subject has data independent of other subjects' data, all of these entries will be zero matrices.

$$\Omega = \begin{bmatrix} \Omega_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Omega_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \Omega_N \end{bmatrix}$$

To see that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}$ is unbiased we recognize that the correlation among the outcomes has not affected the mean of the conditional errors, i.e.$\boldsymbol{E}[\boldsymbol{e}|\boldsymbol{X}] = 0$**;** so the conditional mean remains: $E[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X\beta}$. Treating the design and covariance matrixes as fixed, the regression parameter estimates are unbiased:

$$\boldsymbol{E}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}\right] = E[(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\Omega^{-1}\boldsymbol{Y})] = (\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}\boldsymbol{E}[\boldsymbol{Y}|\boldsymbol{X}] = (\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}\boldsymbol{X\beta}$$

$$= \boldsymbol{\beta}$$

However, the variance of the regression parameter estimates now differs from $\boldsymbol{Var}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}\right]$.

$$\begin{aligned} \boldsymbol{Var}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}\right] &= Var[(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\Omega^{-1}\boldsymbol{Y})] \\ &= [(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}]Var[\boldsymbol{Y}|\boldsymbol{X}][(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}]' \\ &= [(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}]Var[\boldsymbol{Y}|\boldsymbol{X}][\Omega^{-1}\boldsymbol{X}(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}] \\ &= [(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Omega^{-1}]\Omega[\Omega^{-1}\boldsymbol{X}(\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1}] = (\boldsymbol{X}'\Omega^{-1}\boldsymbol{X})^{-1} \end{aligned}$$

When the outcomes in the data set are assumed to be correlated, the distribution from which $\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}$ is drawn will have a different spread than the distribution from which $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$ is drawn. When the measurements are positively correlated, as is the typical situation for longitudinal data, $\boldsymbol{Var}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}\right] > \boldsymbol{Var}\left[\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}\right]$. Therefore, not accounting for the correlation in the data leads to inferential statements that are constructed from the wrong distribution. The given confidence intervals for estimates made under the assumption of independent data will be too narrow. Accounting for the correlation, will no longer give the estimator with the greatest estimated

precision. However, it will allow for correct inferential statements because the spread in the distribution of the parameters will be correctly evaluated.

3.4 TWO COMMONLY USED METHODS TO ANALYZE CORRELATED DATA.

So far I have focused mainly on how the computations used to generate parameter estimates could be adjusted to accommodate a longitudinal study design. Now I focus on the analysis methods that can be used to achieve estimates of the association that have been adjusted for the correlation in the data. There are two commonly used methods of regression analysis, Generalized Estimating Equations (GEE) and Linear Mixed Modeling (LMM), for longitudinal study designs. The GEE approach calculates the parameter estimates using assumptions similar to those of GLS, while the methods for parameter estimation in LMM are based on maximum likelihood. The GEE and LMM have differences other than in the methods used for parameter estimation. Specifically they model the association between the conditional mean and the covariates differently. As a result, the interpretation of the regression parameters usually have different meanings. GEE gives parameters that tell about how the average of a population changes over time. The focus is not for descriptions about how a subject may change over time. However, the LMM method is directed at conclusions about how the typical subject's outcome may change over time. This distinction in interpretation results from how the correlation within a subject is accounted for when developing the components of the GEE or LMM analysis.

The GEE analysis is a marginal model that separately models the mean response and the within subject correlation of the repeated measures. [We continue to use only the simple linear predictor in the modeling equations. For longitudinal data, the single covariate included in the model will represent time. Measures to extend the simple linear model to accommodate other covariates such as gender, AF status etc. will be discussed in a future chapter about multiple

linear regression.] The outcome for a subject at a specific time is determined by summing a systematic component and the random error.

Subject response: $Y_{ij} = \beta_o + \beta_1 X_{ij} + e_{ij}; \quad e_{ij} \sim N(0, \sigma_e^2)$

The systematic component, $\beta_o + \beta_1 X_{ij}$, describes how the mean in the population changes for different values of time. The population mean response is modeled using a regression equation that conditions only on the observed covariates.

Population mean model $E_e[\, Y_{ij} | X_{ij}] = \beta_0 + \beta_1 X_{ij}$

Note that this regression equation does not contain any information about the correlation of observations. The correlation is represented in a separate variance-covariance matrix that is the product of a diagonal matrix, A, for the assumed population variance of the outcome and an assumed "working correlation" matrix, V. The "working correlation" contains the structure that the researcher wishes to impose on the correlation of the observations. It is not necessarily representing what is true in the population or in the data. The following is an example of GEE variance and auto-regressive working correlation matrix.

Variance of the i^{th} subject's responses: $V_i = A_i^{\frac{1}{2}} Corr(Y_i) A_i^{\frac{1}{2}}$

$$A_i \equiv \begin{bmatrix} \sigma_i^2 & 0 & \dots & 0 \\ 0 & \sigma_i^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_i^2 \end{bmatrix} \quad Corr(Y_i) = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \dots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \dots & 1 \end{bmatrix}.$$

The LMM is a subject-specific model that does not have separate components for the mean response and correlation within subjects. The LMM analysis includes terms in the regression model to account for the correlation of outcomes with-in a subject. Like GEE the

subject's response at a specific time is considered to result from summing a systematic component and a random error.

$$Y_{ij}|b_{oi}, b_{1i} = \beta_o + b_{oi} + (\beta_1 + b_{1i})X_{ij} + e_{ij}\ ; \quad e_{ij} \sim N(0, \sigma_e^2)$$

However, the systematic component, $\beta_o + b_{oi} + (\beta_1 + b_{1i})X_{ij}$, consists of more than just terms that describe the population mean. Each subject is considered to also have some unmeasured attributes, b_{oi} and b_{1i}, which cause them to be systematically different from the population mean (i.e. always a bit higher or lower).

There are two types of random effects that can be included in a LMM regression model to account for the correlation of a subject's responses. A random intercept, b_{oi}, is used to indicate how the individual typically responds in comparison to the average of the population. Thus an individual who is above average at the first measurement would have the same offset, b_{oi}, applied to his or her outcome at all successive measurements to allow the outcome to tend to be above average. A random slope, b_{1i}, is used to capture how the individual's rate of change in the outcome compares to the average rate of change. If the individual has a slower decline in outcome over time, then the population regression parameter for rate of change over time β_1 is adjusted by b_{1i}.

The naming, random effects, can be misleading. Like the population level regression parameters, β_o and β_1, the subject specific random effects b_{oi} and b_{1i} are constant across time for a subject. However, for different subjects these values are determined by chance from a distribution and hence called random effects. The random effects are described as coming from a joint distribution because the subject specific intercept and slope are allowed to have some correlation. For example, above average people may also tend to have a slower decline than the general population.

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim MVN\left(\mathbf{0}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b0,b1} \\ \sigma_{b0,b1} & \sigma_{b1}^2 \end{bmatrix}\right)$$

Having two sources of randomness, random effects and individual observation variation, in the LMM equation for the individual's response allows two different mean models to be developed. When considering the expected outcome for an individual, the within subject random effects are considered fixed and we condition on them. The individual observation variation would still be random so we would average over the random error. This relationship is referred to as the subject-specific mean model.

Subject-specific mean model: $E_e[\boldsymbol{Y}_i|\boldsymbol{X_i}, b_i] = \beta_o + b_{oi} + (\beta_1 + b_{1i})\boldsymbol{X_i}$

However, this mean model is not scientifically useful because it describes a relationship between the outcome and characteristics which cannot be observed about an individual. Just like with individual observation variation, which is unobservable, we want to average over the random effects to obtain a summary of the relationship between the measured outcomes and the measured covariates. This model is referred to as the population mean model:

Population mean model $E_b\big[\, E_e[\boldsymbol{Y}_i|\boldsymbol{X}_i, b_i]\big] = E_b\, \big[E_e\big[\beta_o + b_{oi} + (\beta_1 + b_{1i})\boldsymbol{X}_i + e_{ij}\big]\big]$

$$E_b\big[\, E_e[\boldsymbol{Y_i}|\boldsymbol{X}_i, b_i]\big] = E_b\, [\beta_o + b_{oi} + (\beta_1 + b_{1i})\boldsymbol{X_i}]$$

$$E_b[\boldsymbol{Y}_i|\boldsymbol{X}_i] = \beta_0 + \beta_1\boldsymbol{X}_i$$

This second expectation is conceptually equivalent to a two-stage process: after finding out how all possible specific individuals change over time, we averaged the trajectories. Thus the LMM population level regression parameters, β_0, β_1, can be interpreted as the trajectory for a "typical" individual. These LMM population level parameters tell the story of how change occurs within a typical individual over time.

The derivation of the population mean models for LMM and GEE shows that they are remarkably similar in structure. Some important comments need to be made about this misleadingly apparent equivalence. First, the numerical estimates for the parameters that result from LMM and GEE will not necessarily be identical. As is discussed later, missing data leads to differences in the estimates for the parameters under the two methods. Also the estimates of the parameters will be different if the two methods assume different correlation structures. Second, the population mean model has a dual interpretation of the regression parameter in only a limited circumstance. It appears as if the regression parameter can be interpreted both as the "trajectory of the means" (how the population mean changes over time) and the "mean of the trajectories" (how the typical individual in the population changes over time). The possibility of this dual interpretation is only appropriate for the linear regression model (i.e. the untransformed mean is modeled by the linear predictor) and only when the population over which we average is always the same, no matter what time (i.e. there is no missing data).

As previously mentioned, different regression models can be generated that use a non-identity link function on the left hand side of the regression equation, (e.g. Logistic regression $\log\left(\frac{E[Y|X]}{1-E[Y|X]}\right) = \beta_0 + \beta_1 x$, Poisson regression $\log(E[Y|X]) = \beta_0 + \beta_1 x$). For those regression equations a Generalized Linear Mixed Modeling method is appropriate to account for correlation in the data. The mean model for regression situations that require a non-identity link function is $E_e[\,Y|X, b_{0i}, b_{1i}] = g^{-1}(\beta_0 + \ b_{0i} + (\beta_1 + \ b_{1i})x)$. However, in the GLMM the β_1 parameter does not have an equivalent interpretation as "subject specific change over time for the average individual" and "population average change over time". This is because

$$E_b[g^{-1}(\beta_0 + \ b_{0i} + (\beta_1 + \ b_{1i})x)] \neq g^{-1}(E_b[\beta_0 + \ b_{0i} + (\beta_1 + \ b_{1i})x]).$$

3.5 WHY CHOOSE LMM TO ANALYZE DR. THACKER AND COLLEAGUES' DATA?

I have introduced two methods to analyze correlated data, GEE and LMM. We see that when there is missing data, the numeric estimates and the interpretation may be different from the two methods. For a longitudinal study design there will often be missing data so a method that can produce unbiased estimates in the presence of missing data and answer the scientific question is needed. The decision to use LMM results from its ability to achieve both of these objectives in the setting of longitudinally collected data.

As explained in section 3.4, when we are using a linear regression model (compared to a generalized linear regression model) the parameter estimates from either GEE or LMM can be interpreted with the subject-specific meaning needed to answer the Dr. Thacker and colleagues' scientific question "In the absence of clinical stroke, do people with atrial fibrillation (AF) experience faster cognitive decline than people without AF?". In the presence of complete data, mixed models with random subject-specific intercepts and GEE with exchangeable working correlation matrix would provide identical estimates and could provide estimates of the typical subject specific decline in those without an AF and a separate estimate of the typical subject specific decline in people with an AF. The difference in these estimates (and the appropriate confidence intervals for the difference) could be calculated to produce an estimate to the scientific question of interest. However, the existence of a data set that is complete will be unlikely in the longitudinal data setting. Therefore even though both methods could be used to answer the scientific question only one would provide an unbiased estimate.

When the LS and ML methods are modified to accommodate the covariance in the data, Ω, the fact that the GLS methods do not require an assumption about the distribution of responses while the ML method does, becomes a very important distinction[4]. In the unrealistic

situation that information is collected for all subjects at every observation time (i.e. complete data set) the results from GLS and ML will continue to be identical. Likewise the results will be the same when data observations are missing completely at random (MCAR). MCAR missingness means that that the distribution of observed response values is not different from the distribution of response values from the counterfactual complete case study. Thus as Fitzmaurice et al. (2011) describe the "observed data can be thought of as a random sample of the complete data." In each one of these cases, the distribution of the outcome in the population is the same as the distribution of outcome in the observed sample. So the GLS analysis that is based solely on the information in the sample, will coincide with the ML analysis that is based on the assumed distribution in the population. However, with longitudinal data the MCAR pattern for missing data does not seem likely.

It is reasonable to assume that missing data from loss to follow-up would exhibit a pattern that is related to the observed responses. For example, we will tend to have more missing data when the cognition scores are low than when the cognition scores are high. This type of missingness is labeled missing at random (MAR). Fitzmaurice et al. (2011) acknowledge that "because the missingness mechanism now depends on the observed responses the distribution of the Y_i in each distinct stratum defined by the patterns of missingness are not the same as the distribution of Y_i in the target population". When missingness results from MAR, the estimates of the mean and variance that result from the sample data no longer correspond to the target population. Rather they are biased estimates. Since GLS does not make use of the assumption about the distribution of the outcome in the population, its reliance solely on sample data will produce biased estimates when data observations are MAR. Under the more realistic assumption that missing data observations are MAR the ML estimation method remains valid

provided that the assumed multivariate normal distribution is correct[4]. In an analysis where the regression equation models the mean of continuous outcomes using an identity link, the estimating equation for GEE is identical to GLS estimation. Thus the appropriate method to obtain an unbiased estimate from an analysis of the data from Dr. Thacker and colleagues' study would be the maximum likelihood based LMM "linear mixed effects" analysis. In addition to providing unbiased estimates under the Normal distribution assumption in this setting of missing data the LMM method provides the subject specific interpretation of the "mean trend" rather than the population average interpretation of the "trend in the means".

Chapter 4. A MULTIPLE REGRESSION MODEL: NOTATION AND SCIENTIFIC RATIONALE FOR PREDICTORS

For ease of exposition, the regression equations in Chapter 3 modeled the effect of one predictor, X, on the univariate outcome, Y. That simple mixed effects linear regression model would not be appropriate for this study's design. For different reasons, the longitudinal and observational aspect of the study design made inclusion of additional predictors necessary. In this chapter, I will describe the scientific rationale for including additional predictors in the model and demonstrate how the regression equation changes as these predictors are accounted for. Using a multi-predictor model does not change how the random effects, b_{0i} and b_{1i} are represented; for ease of exposition I will omit the random effects as I demonstrate how the model expands. Using the simple linear model with only one predictor variable (time), I begin with a brief explanation of the notation used to identify the observed repeated measures data with the appropriate aspect of the model: subject, observation, and variable. Then I expand the model assuming only linear relationships between each additional predictor and response. Lastly, I rationalize modeling certain predictors more flexibly than a mere linear modeling of the predictor would.

4.1 NOTATION

Rather than use the most general form of notation for longitudinal data, I will develop a system specific to the regression model used in the research of Dr. Thacker and colleagues. The outcome was cognitive score, and is referred to as Cog. Twelve scientifically relevant predictors were chosen: time, AF status, birth year, gender, race, history of diabetes, education, smoking history, alcohol use, blood pressure, history of heart disease, and history of heart failure. As

introduced in Chapter 3, the subscript, *i*, identifies the subject and the subscript, *j*, refers to the time-ordered position of the associated variable: 1st observation, 2nd observation, etc. Rather than using subscripts, the predictors are identified by an appropriate naming of the variable. When the subject's entire vector of observed predictors from a specific observation is needed for matrix operations the representation is $\boldsymbol{X_{ij}}$. The regression parameters are represented by the vector $\boldsymbol{\alpha}$. The error that is added to the systematic portion of the model to recover the specific individual's score is represented by e.

The following notation is for the j[th] observation for the i[th] subject:

Simple linear model notation

$$Cog_{ij} = \alpha_0 + \alpha_1(time_{ij}) + e_{ij}$$

Expected response of subjects with same covariate values.

$$E[Cog_{ij}|time_{ij}] = \alpha_0 + \alpha_1(time_{ij}) = \boldsymbol{X'_{ij}\alpha}$$

For all n_i observations from one subject the observation subscript, *j*, is dropped:

the response is identified by a vector $\boldsymbol{Cog}_i = \begin{bmatrix} cog_{i1} \\ cog_{i2} \\ \dots \\ cog_{in_i} \end{bmatrix}$,

the predictors are a matrix $\boldsymbol{X}_i = \begin{bmatrix} 1 & time_{i1} \\ 1 & time_{i2} \\ \dots & \dots \\ 1 & time_{in_i} \end{bmatrix}$ and the error terms are a vector $\boldsymbol{e}_i = \begin{bmatrix} error_{i1} \\ error_{i2} \\ \dots \\ error_{in_i} \end{bmatrix}$.

One individual's response vector.

$$\begin{bmatrix} Cog_{i1} \\ Cog_{i2} \\ \dots \\ Cog_{in_i} \end{bmatrix} = \begin{bmatrix} \alpha_0 + \alpha_1(time_{i1}) \\ \alpha_0 + \alpha_1(time_{i2}) \\ \dots \\ \alpha_0 + \alpha_1(time_{in_i}) \end{bmatrix} + \begin{bmatrix} error_{i1} \\ error_{i2} \\ \dots \\ error_{in_i} \end{bmatrix}$$

$$\boldsymbol{Cog}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{e}_i$$

Expected response of subjects with same covariate values.

$$\begin{bmatrix} E[Cog_{i1}|time_{i1}] \\ E[Cog_{i2}|time_{i2}] \\ \dots \\ E[Cog_{in_i}|time_{in_i}] \end{bmatrix} = \begin{bmatrix} \alpha_0 + \alpha_1(time_{i1}) \\ \alpha_0 + \alpha_1(time_{i2}) \\ \dots \\ \alpha_0 + \alpha_1(time_{in_i}) \end{bmatrix}$$

$$\boldsymbol{E}[\boldsymbol{Cog}_i|\boldsymbol{X}_i] = \boldsymbol{X}_i\boldsymbol{\alpha}$$

As discussed previously, the interest is in modeling the trajectory of mean responses at different observation times. Going forward, the notation will emphasize how the model for the conditional expectation changes as more predictors are included.

4.2 EXPANSION OF SIMPLE LINEAR REGRESSION TO MULTIPLE PREDICTOR LINEAR REGRESSION MODEL.

Thacker and colleagues' scientific question "In the absence of clinical stroke, do people with atrial fibrillation (AF) experience faster cognitive decline than people without AF?" necessitated a longitudinal study design. In a longitudinal study design, determining the association between time and the outcome variable is the first objective. A simple linear model where the only predictor, time (defined as Age to represent the continuous current age predictor), modeled the association between age and cognition. However, Dr. Thacker and colleagues were not primarily interested in the effect of current age on cognition; rather they were interested in how another factor, AF status, would influence the trajectory of cognition as people aged. Thus, in addition to the predictor current age, an indicator variable for whether the individual has had an AF event was needed. The regression parameter, β_2, that describes the difference in cognition for people who have the same age at observation *j* but have different AF event statuses is also added to the model. At a minimum the multiple regression model for subject, i, at observation *j*, which begins to address the scientific question is:

$$E\left[Cog_{ij} \middle| Age_{ij}, AF_{ij}\right] = \alpha_0 + \alpha_1\left(Age_{ij}\right) + \alpha_2(AF_{ij})$$

Since the study's specific interest was on how an incident AF event affected cognition, AF status was not constant for the duration of the study but varied with time. A criterion to enter the study was that at baseline, all subjects had not experienced an AF event. Thus, for all subjects the AF indicator was 0 at baseline. As the study progressed AF status was stochastic,

i.e. governed by a random mechanism so it could not be precisely predicted at any one observation. The use of the double subscript, *ij*, on the AF status (and any other variables) identified the time-varying nature of the variable. If by observation *j*, the subject had experienced an AF event, for that observation and all subsequent observations the indicator was changed to have a value of 1.

To obtain the vector of average cognition for all observation time points, the covariate design matrix must be updated too. An additional column that contains the AF status at each observation is appended to the design matrix. Thus the model for the vector that contains the trajectory of mean cognition remains $\boldsymbol{E}[\boldsymbol{Cog}_i|\boldsymbol{X}_i] = \boldsymbol{X}_i\boldsymbol{\alpha}$ but uses the updated design matrix and regression parameters. For all subsequent changes to the modeled equation, the design matrix and regression parameter vector will be updated in a similar manner. Implementation of these changes will be left to the reader.

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & time_{i1} & AF_{i1} \\ 1 & time_{i2} & AF_{i2} \\ \dots & \dots & \dots \\ 1 & time_{in_i} & AF_{in_i} \end{bmatrix} \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{bmatrix}$$

In a longitudinal study it is usually sufficient to model time with only one dimension. Dr. Thacker and colleagues had scientific reason to include an additional dimension to capture the passage of time. Using only the single dimension, current age, to measure time implicitly assumes that the time since an AF event has occurred is not important. At observation *j*, the modeled average cognition for 75 year olds would be the same under the following three scenarios: a) they just had an AF at age 75, b) they experienced the AF last year at 74 or, c) they had an AF t years ago at age 75 - t.

$$E[Cog_i|age = 75, AF = 1] = \alpha_0 + \alpha_1(75) + \alpha_2(1)$$

Research by Stefansdottir et al. (2013) supported the theory that AF has a cumulative negative effect on the brain and cognitive functioning in the elderly. It is not merely the type of AF that is important but also the duration of the disease. To capture the time since AF event a second continuous predictor, PostAF, was added to the model.

$$E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij}\right] = \alpha_0 + \alpha_1\left(Age_{ij}\right) + \alpha_2\left(AF_{ij}\right) + \alpha_3\left(PostAF_{ij}\right).$$

Through the added dimension of time, the cumulative effect of an AF is included in the modeled average cognition. Even though they are the same age, and are observed at the same time period, the individuals from scenario a) and b) below will have average cognitions that differ by α_3. Comparing subjects a) and c) the difference in average cognition is the product of the number of years since AF, *t*, and α_3. Thus the difference for each additional year beyond the AF event is constant. The choice of 75 years for the current age is not relevant; for the multiple regression model we must merely hold all other variables constant at some value. Evaluating the above model for all three scenarios, we can see that the per year effect of duration since an AF is constant for all choices of current age.

a) $E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij} = 0\right] = \alpha_0 + \alpha_1(75) + \alpha_2(1) + \alpha_3(0) = \alpha_0 + 75\alpha_1 + \alpha_2$

b) $E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij} = 1\right] = \alpha_0 + \alpha_1(75) + \alpha_2(1) + \alpha_3(1) = \alpha_0 + 75\alpha_1 + \alpha_2 + \alpha_3$

c) $E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij} = t\right] = \alpha_0 + \alpha_1(75) + \alpha_2(1) + \alpha_3(t) = \alpha_0 + 75\alpha_1 + \alpha_2 + t(\alpha_3)$

Additionally it is believed that the current age when cognition is assessed would modify the effect of duration since AF on cognition. The observations from Chen et al. (2014) suggest that "the association between cognitive decline and incident AF is mediated by the presence or development of subclinical cerebral infarcts (SCI)". Fanning et al. (2014) performed a systematic review to understand the incidence, prevalence and risk factors involved in silent brain infarction, an alternate name for SCI. They reported that age has been one of the most

clearly identified risk factors for SCI. Thus association between years post AF and cognition can be modified by the current age. Younger individuals will have fewer SCI that result from the AF so the effect on the cognition is not as strong. To model this, an interaction term between PostAF and Age was included in the model.

$$E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij}\right] = \alpha_0 + \alpha_1\left(Age_{ij}\right) + \alpha_2\left(AF_{ij}\right) + \alpha_3(PostAF_{ij}) + \alpha_4(Age_{ij})(PostAF_{ij})$$

As discussed previously in Chapter 2, the observational nature of the data means that the effect measured can be biased if potential confounding variables are not accounted for in the analysis. To control for the potential confounding of other factors, additional terms are included in the regression model for these variables. When discussing the results that come from cross-sectional versus longitudinal data, it was noted that a cohort effect can alter the association that is detected. Since the subjects entered the Cardiovascular Health Study at ages ranging from 65 to 85, a term to control for the potential confounding of the birth cohort effect was included. The BirthYear predictor is a continuous time-invariant variable with only integer values. It was recorded once at baseline for each subject.

$$E\left[Cog_{ij}\middle|Age_{ij}, AF_{ij}, PostAF_{ij}, BirthYear_{\boldsymbol{i}}\right] = \alpha_0 + \alpha_1\left(Age_{ij}\right) + \alpha_2\left(AF_{ij}\right) + \alpha_3\left(PostAF_{ij}\right) + \alpha_4\left(Age_{ij}\right)\left(PostAF_{ij}\right) + \alpha_5(BirthYear_i)$$

Other variables are known to be associated with age and cognition. But rather than introducing bias into a single estimate these variables have been studied enough to believe that between different levels of these predictors the association between age and cognition is different. Therefore, Dr. Thacker and colleagues included terms that modeled the effect modification for the remaining eight predictors: gender (Male), black race (BRace), history of diabetes (Diab), education (Educ), smoking history (Smk) , alcohol use (Alc), blood pressure

(BP), history of heart disease(HrtDis) and history of heart failure (HrtFail) (Bond, et al., 2005; Kilander, Nyman, Boberg, Hansson, & Lithell, 1998; Stott, et al., 2008; Waldstein, 2003; Wright, Elkind, Luo, Paik, & Sacco, 2006). Four of the variables were indicator variables, Male, BRace, Diab, HrtDis, and HrtFail; a value of 1 indicates that characteristic was assigned to the subject, while a value of 0 indicated the absence of that characteristic. One predictor, Smk, was originally measured as a categorical variable with three levels. Another predictor, Educ, was ordered categorical and could be treated as linear. The remaining two predictors Alc and BP were continuous. Further modifications of some of these predictors were performed to allow for more flexible representation of the association between that predictor and cognition. These modifications will be discussed at the end of this chapter. The resulting model thus far is:

$$\begin{aligned} E\left[Cog_{ij}|X_{ij}\right] = {} & \alpha_0 + \alpha_1\left(Age_{ij}\right) + \alpha_2\left(AF_{ij}\right) + \alpha_3\left(PostAF_{ij}\right) + \alpha_4\left(Age_{ij}\right)\left(PostAF_{ij}\right) \\ & + \alpha_5(BirthYear_i) + \alpha_6(Male_i) + \alpha_7(BRace_i) + \alpha_8(Diab_i) + \alpha_9(Educ_i) \\ & + \alpha_{10}(Smk_i) + \alpha_{11}(Alc_i) + \alpha_{12}(BP_i) + \alpha_{13}(HrtDis_i) + \alpha_{14}(HrtFl_i) \\ & + \alpha_{15}\left(Age_{ij}\right)(Male_i) + \alpha_{16}\left(Age_{ij}\right)(BRace_i) + \alpha_{17}\left(Age_{ij}\right)(Diab_i) \\ & + \alpha_{18}\left(Age_{ij}\right)(Educ_i) + \alpha_{19}\left(Age_{ij}\right)(Smk_i) + \alpha_{20}\left(Age_{ij}\right)(Alc_i) \\ & + \alpha_{21}\left(Age_{ij}\right)(BP_i) + \alpha_{22}\left(Age_{ij}\right)(HrtDis_i) + \alpha_{23}\left(Age_{ij}\right)(HrtFail_i) \end{aligned}$$

Modeling the effect modification with these eight new predictors is similar but not identical to the effect modification between the two dimensions of time. The effect modification is captured by terms that are the product of current age and the predictor. The main-effect terms for each component in the product are included; different levels of a predictor (or different values of current age) are not forced to have the same intercept. Unlike the representations of time, the values of these eight effect modifying predictors were all time-invariant. They were measured on a subject at baseline and did not change throughout the study. To indicate the constancy across time, the subscript is single, *i*, rather than double, *ij*. The potential confounder, birth year, also has this change in subscript notation.

At this stage, the model has expanded from the simple regression that included only two terms: an intercept and a predictor. The current model for average cognition includes twenty-four terms. The additional terms add detail to the model that explain more variation in cognition to improve the estimation of the association. However, more flexible modeling of the continuous predictors (Age, Educ, Alc, and BP) and the categorical predictor (Smk) can improve the fit of the model even more. With over 5000 subjects, the danger of having too many terms so that we are overfitting the model is quite small. The benefit of improved model fit is worth the additional parameter estimation. For maximum likelihood estimation of the model parameters the correct specification of the distribution is a necessary assumption. The correctly fitting regression model defines the mean of the distribution and thus is necessary for parameter estimation.

4.3 MODELING NONLINEAR RELATIONSHIPS BETWEEN A PREDICTOR AND COGNITION.

Using a linear representation of a predictor in the regression model might not adequately fit the relationship between the predictor and cognition. First the rate of change in Cog may not be constant over the entire domain of the predictor. Second, changes in Cog throughout the entire domain of the covariate data may not progress from one original intercept. For scientific reasons, four of the predictors (Age, Educ, Alc, and BP) were considered to have a non-linear relationship with Cog. As age increases, the rate of change in cognition increases (in an absolute sense) rather than staying constant. For education, there is a ceiling effect. Beyond high school accruing more years of education will not improve cognition indefinitely. It is known that, when compared to abstainers, people who are drinkers have slower cognitive decline (Bond, et al., 2005; Stott, et al., 2008; Wright, Elkind, Luo, Paik, & Sacco, 2006). Likewise for blood

pressure, hypertensive people perform more poorly on cognitive tasks than normotensives (Waldstein, 2003). Figures 4-1 to 4-4 provide comparisons of the hypothetical linear and non-linear associations with cognition for the four predictors (Age, Educ, Alc, and BP). Obtaining a better fit between the linear predictor and the observed data would require modifications to the regression model. The association between each of the four predictors and cognition should model the non-linear fit identified by the solid black lines rather than the dashed linear fit.

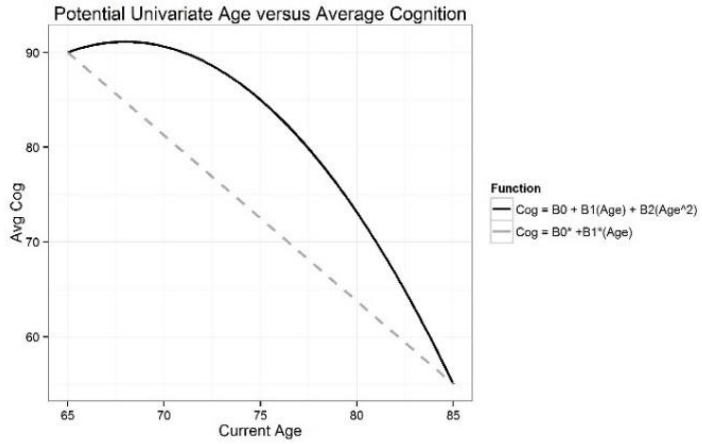

Figure 4-1 Comparing linear and nonlinear modeling of age.

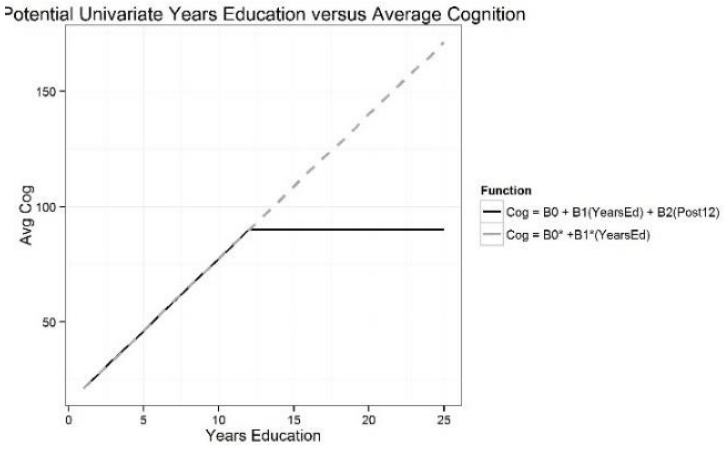

Figure 4-2 Comparing linear and nonlinear modeling of years of education.

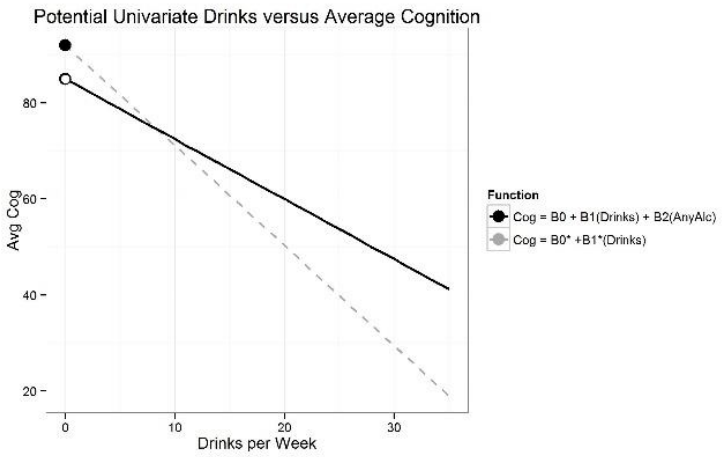

Figure 4-3 Comparing linear and nonlinear modeling of drinks per week

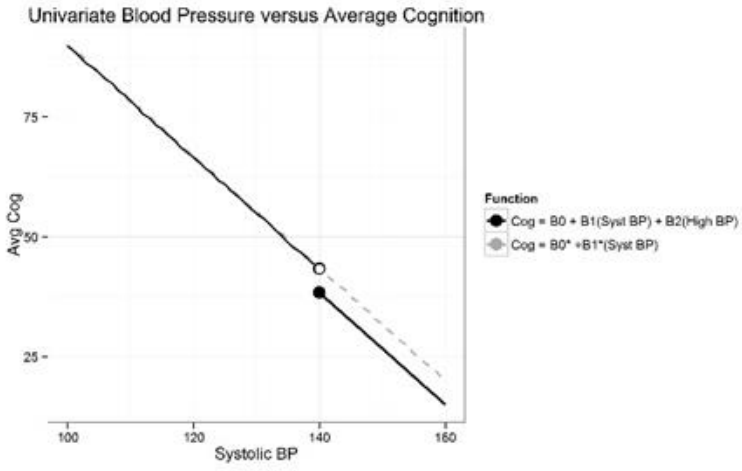

Figure 4-4 Comparing linear and nonlinear modeling of systolic blood pressure

When science indicates that we have a reason to believe a continuous predictor has a non-linear association with the outcome, we have many options to improve our model. Using a polynomial representation of the predictor will allow for a non-constant rate of change. Allowing the intercept to change based on the domain of the predictor is achieved by selecting a cut-point and modeling the predictor with two terms: an indicator for groups above/below the cut-point and continuous variable for the domain of the predictor. The variable age has a non-

linear association with cognition of the first type. To capture the non-constant rate of decline across current age, Dr. Thacker and colleagues modeled the predictor Age using a fifth degree polynomial. For Educ, Alc, BP the non-linear association was of the second type; so, these variables were split at a cut-point and modeled using indicators and linear terms (Educ: AnyEdPost12, YearEd12; Alc: AnyAlc, DrinksPerWeek; BP: Hypertension; SBP). The regression parameters in the model have been renamed from α to β to indicate the associations represented are not the same for the model using linear representation.

$$
\begin{aligned}
E[Cog_{ij}|X_{ij}] = {} & \beta_0 + \beta_1(Age_{ij}) + \beta_2(Age_{ij})^2 + b_{2i}(Age_{ij})^2 + \beta_3(Age_{ij})^3 + \beta_4(Age_{ij})^4 \\
& + \beta_5(Age_{ij})^5 + \beta_6(AFIndic_{ij}) + \beta_7(FUPostAF_{ij}) + \beta_8(Age_{ij})(FUPostAF_{ij}) \\
& + \beta_9(Age_{ij})^2(FUPostAF_{ij}) + \beta_{10}(BirthYear_i) + \beta_{11}(Male_i) + \beta_{12}(BRace_i) \\
& + \beta_{13}(Diab_i) + \beta_{14}(YrEd12_i) + \beta_{15}(AnyEdPost12_i) + \beta_{16}(PriorSmk_i) \\
& + \beta_{17}(CurrSmk_i) + \beta_{18}(AnyAlc_i) + \beta_{19}(DrinkPerWk_i) \\
& + \beta_{20}(Hypertension_i) + \beta_{21}(SBP_i) + \beta_{22}(HrtDis_i) + \beta_{23}(HrtFl_i) \\
& + \beta_{24}(Male_i)(Age_{ij}) + \beta_{25}(BRace_i)(Age_{ij}) + \beta_{26}(Diab_i)(Age_{ij}) \\
& + \beta_{27}(YrEd12_i)(Age_{ij}) + \beta_{28}(AnyEdPost12_i)(Age_{ij}) \\
& + \beta_{29}(PriorSmk_i)(Age_{ij}) + \beta_{30}(CurrSmk_i)(Age_{ij}) + \beta_{31}(AnyAlc_i)(Age_{ij}) \\
& + \beta_{32}(DrinksPerWk_i)(Age_{ij}) + \beta_{33}(Hypertension_i)(Age_{ij}) \\
& + \beta_{34}(SBP_i)(Age_{ij}) + \beta_{35}(HrtDis_i)(Age_{ij}) + \beta_{36}(HrtFl_i)(Age_{ij})
\end{aligned}
$$

Even though the association of age and cognition is represented using the polynomial terms, the interactions do not occur with all terms of the age polynomial. Using only the linear term to model the interaction with the appropriate other predictors (except PostAF) preserves the notion that the difference in the rate of change in cognition over time between the levels of that predictor will be constant. For example, comparing the rates of change in cognition over time between normotensive people whose SBP differs by 1 point will be constant. There is no dependence on the age at which we compare the rates of cognitive decline.

The interaction term between the two dimensions of time did incorporate the more complex modeling of age. By including the interaction of Age^2 and FUPostAF, yet another

nuance to the relationship between current age and years since AF was preserved. The difference in the rates of cognitive decline for people who have different durations since their AF is not constant. Comparing younger people who have an AF event, we see that an additional year since the AF event occurred will not create as large a difference in the rates of decline as if we compared older people.

To complete the scientific development of the model, I close with the explanation for the inclusion of the three random effects terms: $b_{i0}, b_{i1}(Age_{ij})$ and $b_{i2}(Age_{ij})^2$. In Chapter 3, an explanation was provided for using a random intercept and random slope for the linear term of age. The random slope for the quadratic term of age ensures that not only the rates of change but also the speed with which it changes can vary from person to person.

One further component of the model is necessary for computational reasons rather than scientific. The polynomial representation of Age introduced potential computational difficulties. Solving for the estimates of the regression parameters requires that the terms in the model are not collinear, that is none of the terms in the model should be able to be linearly predicted from other terms in the model. However the terms in the polynomial representation of age are highly correlated. Centering of the current age variable by subtracting off a constant prior to the polynomial transformation reduces the correlation between the polynomial terms. Thus Age_{ij} was replaced by CAC_{ij} to indicate that the current age was centered by the mean of the baseline ages.

Thus Dr. Thacker and colleagues used to the following model to analyze the data:

$$
\begin{aligned}
E\left[Cog_{ij}\middle|X_{ij}\right] = {} & \beta_0 + b_{i0} + (\beta_1 + b_{i1})\left(CAC_{ij}\right) + (\beta_2 + b_{i2})\left(CAC_{ij}\right)^2 + \beta_3\left(CAC_{ij}\right)^3 \\
& + \beta_4\left(CAC_{ij}\right)^4 + \beta_5\left(CAC_{ij}\right)^5 + \beta_6\left(AFIndic_{ij}\right) + \beta_7\left(FUPostAF_{ij}\right) \\
& + \beta_8\left(CAC_{ij}\right)\left(FUPostAF_{ij}\right) + \beta_9\left(CAC_{ij}\right)^2\left(FUPostAF_{ij}\right) + \beta_{10}(BirthYear_i) \\
& + \beta_{11}(Male_i) + \beta_{12}(BRace_i) + \beta_{13}(Diab_i) + \beta_{14}(YrEd12_i) \\
& + \beta_{15}(AnyEdPost12_i) + \beta_{16}(PriorSmk_i) + \beta_{17}(CurrSmk_i) + \beta_{18}(AnyAlc_i) \\
& + \beta_{19}(DrinkPerWk_i) + \beta_{20}(Hypertension_i) + \beta_{21}(SBP_i) + \beta_{22}(HrtDis_i) \\
& + \beta_{23}(HrtFl_i) + \beta_{24}(Male_i)\left(CAC_{ij}\right) + \beta_{25}(BRace_i)\left(CAC_{ij}\right) \\
& + \beta_{26}(Diab_i)\left(CAC_{ij}\right) + \beta_{27}(YrEd12_i)\left(CAC_{ij}\right) + \beta_{28}(AnyEdPost12_i)\left(CAC_{ij}\right) \\
& + \beta_{29}(PriorSmk_i)\left(CAC_{ij}\right) + \beta_{30}(CurrSmk_i)\left(CAC_{ij}\right) + \beta_{31}(AnyAlc_i)\left(CAC_{ij}\right) \\
& + \beta_{32}(DrinksPerWk_i)\left(CAC_{ij}\right) + \beta_{33}(Hypertension_i)\left(CAC_{ij}\right) \\
& + \beta_{34}(SBP_i)\left(CAC_{ij}\right) + \beta_{35}(HrtDis_i)\left(CAC_{ij}\right) + \beta_{36}(HrtFl_i)\left(CAC_{ij}\right)
\end{aligned}
$$

Chapter 5. DETAILS ABOUT THE SELECTION OF THE POLYNOMIAL MODEL FOR AGE.

The aim of the research by Thacker and colleagues was to "examine the association of incident atrial fibrillation with cognitive trajectories using annual cognitive assessments in a large longitudinal study of older adults". As explained in Chapter 4, the variables selected for inclusion in the regression model characterizing the cognitive trajectories were chosen for scientific reasons. They were selected *a priori* to any examination of the data. However, the nonlinear modeling for current age was determined by applying the Bayesian Information Criteria (BIC) to the data set. In this chapter I briefly explain the rationale for using this model selection process.

In Chapter 3, the distinction was made between two commonly used methods of parameters estimation, least squares and maximum likelihood. For this analysis of longitudinal data that is investigating within subject changes in cognition over time, the maximum likelihood (ML) based estimation method was determined to be most appropriate. When using ML methods, a distribution with a likelihood function will be assumed. The mean of the distribution will be parameterized by the conditional mean that was modeled by the regression equation. Changes in how covariates are modeled in the regression equation will imply different parameterizations of the likelihood function. Determining which modeling of the covariates is best is done by comparing some form of the likelihood that results from different regression models. Three commonly used methods of model selection based on a likelihood function are: Likelihood Ratio Tests, Akaike Information Criteria (AIC), and Bayesian (Schwartz) Information Criteria (BIC). The formulas for the methods are given below.

$$LRT = -2 * \left[\log \frac{f\left(y|\hat{\theta}_{null}\right)}{f\left(y|\hat{\theta}_{alternative}\right)}\right]$$
$$AIC = -2 * \log(f\left(y|\hat{\theta}_k\right) + 2 * k$$
$$BIC = -2 * \log(f\left(y|\hat{\theta}_k\right) + k \log(n)$$

The simplest method to determine the goodness of fit for a regression model is to use the LRT. In a LRT two likelihoods are compared. For this test, the two models must be nested: one likelihood contains the additional parameters and the other is a simplified null version with the parameters in question set equal to zero. Using more parameters tends to give better model fit. However, the trade-off is that the regression equation can over-fit the data.

The two Information Criterion based methods attempt to account for the overfitting to the data by including a term to penalize for the number of parameters in the criterion. Both methods begin with the goodness of fit term: the product of -2 and the maximized log likelihood. Then the goodness of fit is penalized by a factor of the number of parameters. The AIC formula penalizes by a factor of two. Whenever the sample size n is larger than 8, the BIC formula utilizes a larger factor: log(n).

Both the AIC and BIC criteria can be used for selecting between models that are not necessarily nested. When two models are nested, one model consists of a subset of the terms from the first model. The two versions of the potential model are fit by the ML method. Then the information criterion is calculated using the estimated regression parameters. The model with the smallest information criterion is selected as the best fit for the data. When trying to select the appropriate degree for the polynomial representation of age and the modeling of the covariate interactions with the polynomial representation of age, either selection criterion would have been appropriate. Dr. Thacker and colleagues utilized the BIC selection criteria to give a larger penalty for each additional term in the model. The conclusion from their BIC model

selection process was that a fifth degree polynomial for current age and interaction terms between time since AF and at most the squared term of current age would fit the data best.

The scientific rationale for using a polynomial representation of age was that the decline in cognition was thought to be non-constant over time. As people get older, cognition is expected to decline at a faster rate. As described in Chapter 4, the polynomial representation of age was suggested by a non-linear pattern in the univariate scatterplot of age and cognition. The fifth degree polynomial captures a very complex pattern for how the rate of cognition changes as people age. From the graphical depiction given in Figure 1 in the paper by Dr. Thacker and colleagues, the average cognition for the NOAF group appears to always be concave down. The rate of change of cognition is non-constant but does not appear to have more than one local maximum. This led me to wonder whether representing age using a less flexible quadratic polynomial would capture the non-constant rate of decline in cognition adequately enough. If we ignored the results from the BIC analysis and chose a simpler model for cognition, would the results have been remarkably different? The research that I conducted investigated whether using a simpler polynomial model for age in the regression analysis model would alter the conclusions made about the change in cognition when comparing people with AF and without AF.

Chapter 6. SIMULATION DESIGN

6.1 SOURCES/REFERENCES FOR SIMULATED DATA

I utilized simulated datasets rather than real data to compare two analysis models for assessing the impact of atrial fibrillation (AF) on cognitive trajectories. The simulated datasets differed only in how the cognition scores for individuals were created. Simulation 1 utilized the more complex fifth degree modeling of age from Dr. Thacker and colleagues to generate cognition scores. Simulation 2 utilized the simple quadratic modeling of age which I proposed to create cognition scores. With respect to covariates the data generation in the two simulations was identical.

While generating the simulated datasets I relied on three sources to provide summary information about variables which I generated: Thacker and colleagues (2013), the associated online supplemental materials, and publically available information[1] pertaining to the entire Cardiovascular Health Study (CHS) sample. In the simulated dataset three different types of observations were created for each subject: baseline covariate values, time-varying covariate values, and the time-varying outcome 3MSE cognition score. For all the baseline covariates and the time-varying covariate current age, the summary statistics in Table 1 of Thacker and colleagues and descriptive information about the larger CHS sample served as the references. Information to construct the time-varying covariate AF status came specifically from Thacker and colleagues. The material to construct the time-varying 3MSE scores was found in three locations: Figure 1 of the original paper, Table 2 of the original paper, and Table e-1 of the supplemental materials. In none of the references did I locate information about the correlation of covariates within an individual. Therefore, I made the simplifying assumption that the covariates are independent within and across subjects.

6.2 OVERVIEW FOR THE PROCESS OF DATA GENERATION

For the simulation study I made the additional simplifying assumptions that all subjects would share identical visit dates and remain in the study until completion. During the first step of data generation, a wide format database that contained one row for each subject was built. It contained the dates for baseline visit and each of 9 follow-up visits that were one year apart. Since the covariate of interest for this study is AF status, I tried to develop a dataset that preserved differences across groups defined by AF status. From Thacker and colleagues' Table 1 descriptive statistics, it was evident that some baseline characteristics varied considerably between the two groups defined by AF status. (Table 1 from Thacker et al. has been reproduced as part of Table I in Chapter 6.) Prior to assigning baseline characteristics, I separated the subjects into two groups: those who would eventually have an AF during follow-up and those who would never have an AF during follow-up. Then baseline characteristics were assigned to members within each of the two AF status groups so that the baseline covariates were similar to the summaries given in Thacker and colleagues Table 1. An age when the AF event occurred was determined for subjects who were assigned to have an AF event. To conclude the first step, the two wide format datasets were merged into one main wide format dataset.

During the second phase of data generation, the wide format dataset was converted to a long format dataset so that every visit time for each subject was a separate row. The values for time varying variables (AF indicator, age, follow-up time post AF event, and 3MSE cognition score) were assigned according to visit status. The terms to model correlation of cognition scores within an individual across visit times were generated: a random intercept, and two random slopes. An error term modeling the variation in cognition due to sources not accounted for in the regression model was produced for each subject at each visit time. Lastly, according to

the appropriate data generating model, the values for covariates, random effects, and the error were summed to create a cognition score for each individual at each observation time. The details about the distributions from which the random values were chosen will be explained in section 6.3.

The final phase of data generation was confirmation that the simulated data set resembled the real world sample. For the group of individuals who never have an AF event, the group of people who eventually have an AF event, and the sample overall, the Table 1 summary statistics were computed for baseline covariates. As will later be presented, comparison of this data generation Table 1 with Thacker and colleagues Table 1 indicated that the parameters used to generate the baseline covariates in the data were adequate. Then the overall cognition scores at various ages were summarized. The means were compared to Thacker and colleagues' Figure 1 to determine whether the trajectory was approximately correct. The minimum and maximum cognition values at these ages were checked to see that they were reasonable. When any of these validations returned implausible values, the parameters used to generate the data were altered and a new data set was generated. With these preliminary data checks met, the process of data generation was deemed appropriate and progress was made to the analysis phase of the research.

6.3 SPECIFICS OF DATA GENERATION – WIDE FORMAT

The data generation methods used in this simulation study attempted to recreate a sample that, in general, would have resulted from Thacker and colleagues' sampling methods and data collection techniques. The number of subjects in the entire sample in the simulation was set to 5000. We were not attempting to recreate the different entry times or mistiming of follow-up visits. Thus the first step when building the wide format of the data set was to assign the values

that would be constant for each subject. The date for the baseline visit was predetermined as June 22, 1989. The nine follow-up visits all occurred on June 22 of the successive year.

To simulate that the sample has been drawn from a larger population, each covariate value (e.g. gender, smoking status, etc.) for a subject was assigned based on random draws from a distribution. The random sampling is first evident in the data generation process with the assignment of a person's eventual atrial fibrillation status. Thacker and colleagues reported that 10.7% of the participants experienced an incident AF by the conclusion of their study. This percentage served as the binomial distribution probability of success when randomly generating each individual's static AF status. While the dataset is in wide format each person's values for the baseline covariates are assigned. After the assignment of baseline covariates and the later conversion of the dataset to long format, this initially assigned static AF status was transformed to a time-varying covariate: AF indicator. The time-varying nature of the AF indicator will be described in the future section 6.4 which discusses the long format of the dataset.

The baseline covariates can be separated in three categories: categorical, those modeled only as continuous terms, and those modeled with a combination of binary and continuous terms. A binomial distribution was used to generate each individual's status with respect to the categorical baseline variables which had only two levels: male, black race, diabetes, history of coronary heart disease, and history of heart failure. Since each individual's status on every binary variable was independent of any other subject's status, independent draws from the binomial distribution with the correct parameterization were repeated for each subject. Table A lists the probability of success which parameterized the binomial distribution of each variable. Smoking status was a three level categorical variable (never, former, and current). For this multilevel variable, the probability of being in each category was related to the probability of being in each

of the other categories. Thus the smoking status was modeled using a multinomial distribution. Like draws from the binomial distribution, each person's assignment to a category was independent of other subjects' smoking status. A multinomial distribution with parameters listed in Table B was used to assign each individual's smoking level: 0 - never smoked, 1 - former smoker, and 2 - current smoker.

Table A Binary baseline covariates: Probability parameters for data generation distributions

	never AF	Eventual AF
Male	0.399	0.520
Black Race	0.160	0.092
Diabetes	0.147	0.163
History of Coronary Heart Disease	0.169	0.286
History of Heart Failure	0.031	0.053

Table B Multilevel categorical smoking variable: Sample size and probability parameters for data generation distributions.

	Never AF	Eventual AF
Sample Size	~ 4465	~ 535
Never Smoke	0.465	0.464
Prior Smoke	0.416	0.431
Current Smoke	0.119	0.105

Unlike the categorical variables, the creation of a subject's continuous baseline age required a two-step process of random draws from different types of distributions. The CHS cohort was conceptualized to have a specific baseline distribution of ages across four age brackets: 65-69 year olds, 70-74 year olds, 75-79 year olds, and older than 80 years. Initially subjects in the dataset were assigned a baseline age that matched the youngest age in a designated age bracket (i.e. 65, 70, 75, or 80) using the ratio 0.35: 0.25: 0.20: 0.20. These ages were assigned by making one draw from a multinomial distribution parameterized by the sample size and CHS theoretical age probability ratio. Then, a process that allowed for even distribution of individuals across the years in a given age bracket was utilized. For each individual a random

number from a Uniform (0,1) distribution was generated. For people with initial baseline ages of 65, 70, or 75 the age bracket spanned five years. The oldest age bracket extended from 80 to either 82 years or 84 years depending on which AF event status dataset was being generated. To determine an individual's specific age in the assigned age bracket, the random uniform number was multiplied times the width of the age bracket and the product was added to the initial baseline age.

Examination of the summary statistics of Thacker and colleagues' sample, guided the decision to vary the widths of the 80 and older age bracket based on whether the dataset was for people who would or would not eventually experience an AF event. The sample of people who have an AF were older and had slightly wider spread (mean 74.4 years sd(5.8)) than the people who did not have an AF (mean 72.9 years sd(5.3)). The upper limits were determined by trial and error until the mean baseline ages of each dataset defined by AF status were similar to the means reported by Thacker and colleagues. Table I comparing the descriptive results for age (and the other covariates) is described later.

As explained in Chapter 4, the remaining continuous baseline covariates (education, alcohol use, and blood pressure) were each represented in the regression analysis using a combination of an indicator variable and a continuous variable. Examining each variable's pair of descriptive statistics in Thacker and colleagues' Table 1, the proportion of participants above the variable specific threshold (12th grade, 0 drinks, 140 mm HG) exceeded what would be expected if the sample was selected from a normal distribution parameterized by the values in Table 1 of Thacker and colleagues. Table C offers a comparison of what proportion of a normal distribution should exceed each cut-point. Thus distributions for years of education, alcohol use, and blood pressure must all have been asymmetrical about the stated means. To select the

appropriate distribution from which to generate the data, more detailed information about how these variables were distributed in the original CHS sample was utilized. This process is described below.

Table C Mismatching percents above the threshold show the distributions of Education, Alcohol Use, and Blood Pressure are not Normal with the mean and variance in Thacker et al.'s Table 1.

	All participants	Never AF event	Eventual AF
Years Ed through 12th	11.0 (2.0)	11.0 (2.0)	11.0 (2.0)
Reported % beyond 12th	43.5%	43.3%	45.1%
True % of Normal distribution above 12th grade	30.9%	30.9%	30.9%
Drinks per week	5.0 (8.4)	5.0 (8.5)	5.1 (7.7)
Reported % exceeding 0 drinks per week	50.7%	50.4%	52.5%
True % of Normal distribution above 0 drinks/wk	72.4%	72.2%	74.6%
Systolic Blood Pressure, mm HG	136.0 (21.5)	135.7 (21.4)	138.1 (22.6)
Reported % that exceeds 140 mm HG	57.2%	56.7%	61.2%
True % of Normal distribution above 140 mm HG	42.6%	42.1%	46.7%

Using the histogram from the original CHS sample[4] for years of education,

Figure 6-1, I decided to generate a distribution of years of education that was skewed and bimodal. First, the individuals who would be assigned education post high school were chosen using a binomial distribution. Then based on their post high school education status, the years of education value was selected from a truncated normal distribution. A truncated normal was necessary so that the grade levels could be restricted to the feasible ranges appropriate to the pre/post high school categorization. For use in the regression analysis, the years of education variable was trimmed so that people who had education beyond high school were represented with a value of 12 rather than the number generated from the truncated normal. The indicator

[4] All histograms for the CHS sample can be found at http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/dataset.cgi?study_id=phs000287.v4.p1&phv=100502&phd=2774&pha=3548&pht=1452&phvf=3&phdf=1&phaf=&phtf=&dssp=11&consent=&temp=1

variable, any post high school education, represented the information that these individuals had exceeded twelfth grade education.

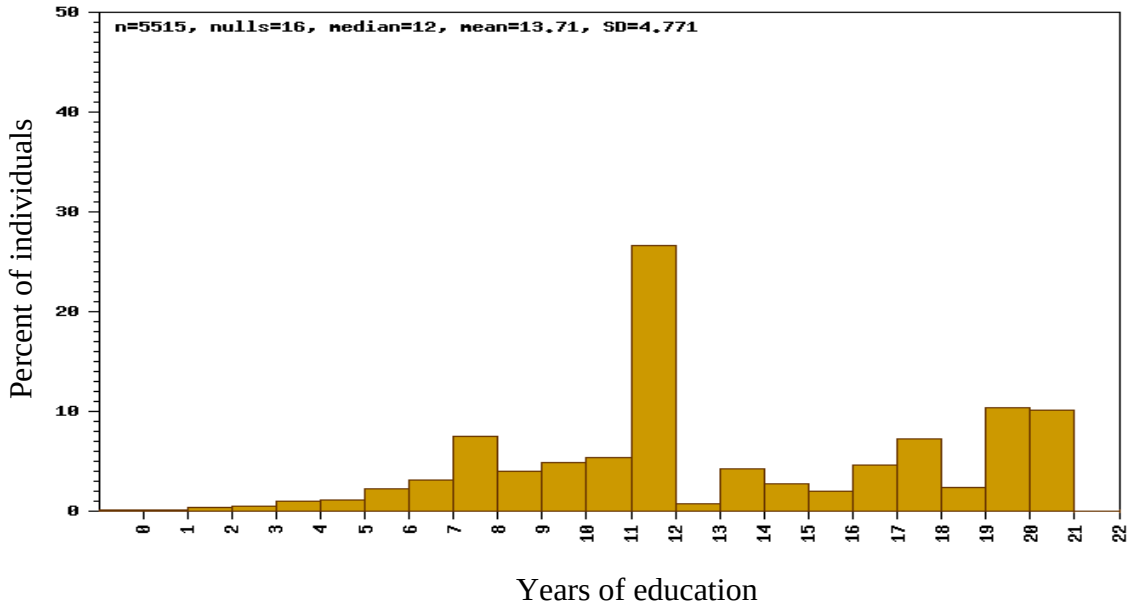

Figure 6-1 Distribution for years of education in the CHS sample. On the CHS website the variable name is Grade 01 and the variable accession is phv00100301.v1.p1

Even though the density function for a normally distributed random variable condensed within left and right endpoints is known, generating simulated draws from this distribution can be inefficient. So the creators of the R package *truncnorm* utilized the accept-reject method to generate random variables X from the truncated normal density f(x) (Trautmann, Steuer, Mersmann, & Bornkamp, 2015). This method relies on one assumption and two facts from probability theory to ensure that the random variable X will follow the truncated normal distribution. First, identify another random variable Y with a density function g(y) for which an algorithm exists that allows for easy simulation of random draws. Make the assumption that the ratio of the densities of the two random variables are close so that for all values of x in the domain of X, there exists a constant c such that $\frac{f(x)}{g(x)} \leq c$. Then using multivariable calculus, the

definition of a cumulative distribution, and the definition of joint probability it has been proved that the distribution of X generated from the accept-reject method will follow the correct distribution (Robert, 1995; Sigman, 2007). An annotated version of the proof has been included in Appendix C.

The parameters necessary to generate an individual's value for years of education varied depending on the AF status and whether the individual surpassed a high school education. Table D contains the parameters for the binomial distributions for whether post high school education was achieved and the truncated normal distributions to assign years of highest education. The parameters necessary to utilize the *rtruncnorm* package are the number of observations, the left and right endpoints for the truncation, and the mean and standard deviation of the non-truncated normal distribution. The number of observations varied depending on the results of the draws from the binomial distributions. The values for the mean and standard deviation were determined by trial and error to imitate the bimodal distribution referenced in

Figure 6-1 and summary values in Table 1 from Thacker and colleagues.

Table D Years of education: Parameters for data generation distributions by AF status.

	Binomial Parameter		Truncated Normal Parameters			
	Probability of post 12^{th} education		Min	Max	Mean	SD
Never AF	0.433	HS only	0	12.01	22	6
		More than HS	12.01	21	20	6
Eventual AF	0.451	HS only	0	12.01	13.75	4.69
		More than HS	12.01	21	13.75	4.69

The process to represent an individual's alcohol consumption was similar to that of generating years of education. A binomial distribution was used to assign whether an individual would use alcohol. Then the number of drinks per week was assigned from another distribution. The CHS data were consulted regarding the distribution for the drinks per week. The grouping of the categories in the histogram for

Figure 6-2 are wide-ranging and did not provide sufficient detail to replicate the distribution. The CHS data website provided additional information, a list of the 10 most frequently occurring valid values, which gave further insight into the distribution within the category 0 to 20 drinks per week. The detail for the drinks per week reported with the highest frequencies is presented in Figure 6-3 (the non-drinkers, who represented the highest category of responses, were omitted from this visualization).

Figure 6-2 and Figure 6-3 illustrate characteristics appropriate for a chi-square distribution: 1) the possible values for drinks per week was necessarily greater than zero, 2) the distribution was skewed to the right, and 3) the variance is larger than the mean. Based on these observations, this distribution for drinks per week was first modeled using a central chi-square distribution.

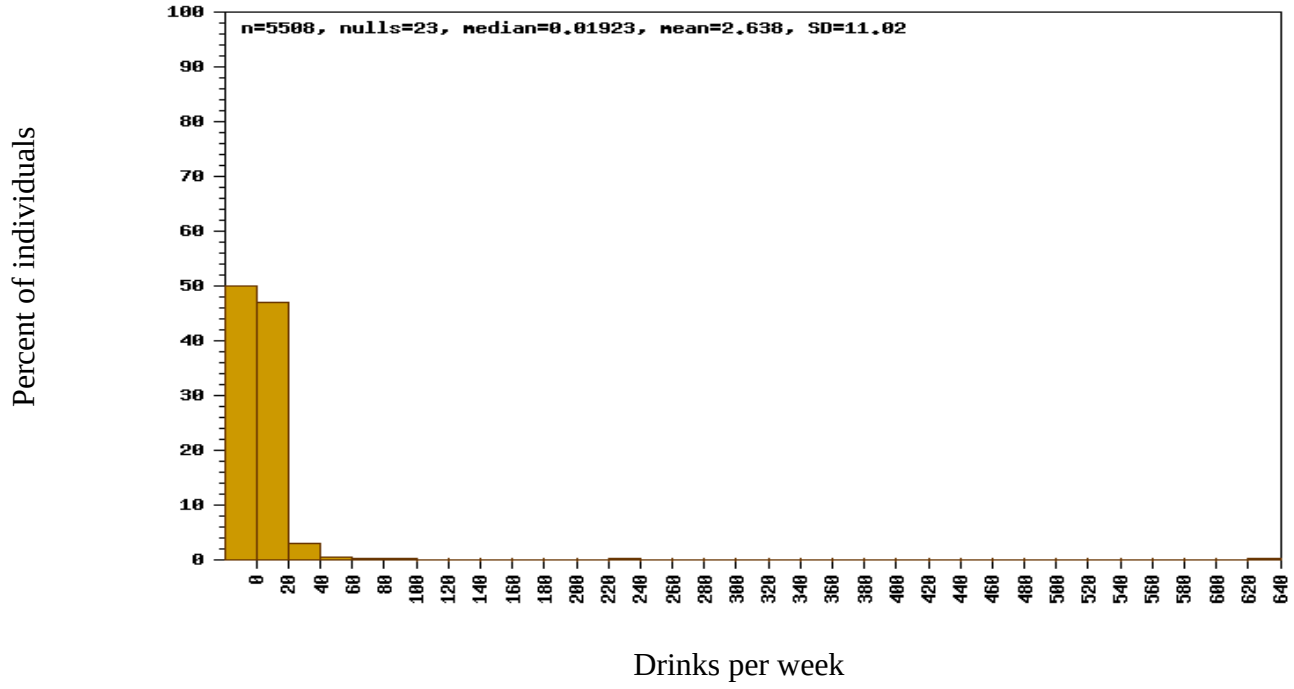

Figure 6-2 Distribution for drinks per week in CHS sample. On the CHS website the variable name is ALCOH and the variable accession is phv00100502.v1.p1.

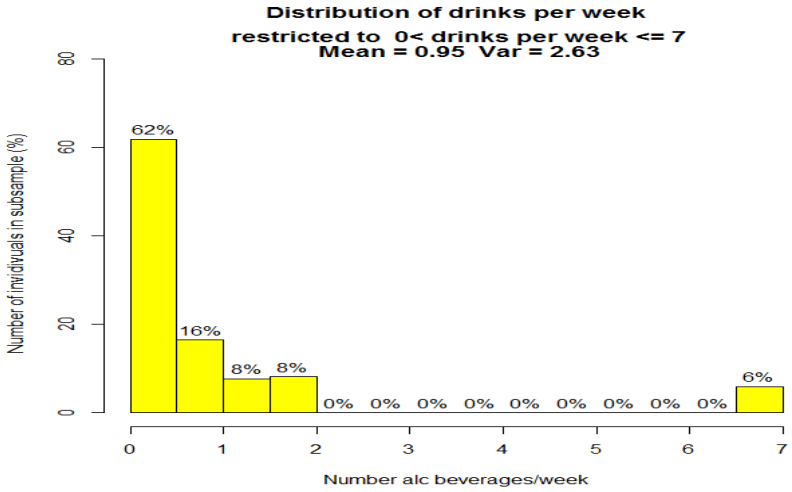

Figure 6-3 Drinks per week in the CHS subsample of the 9 most reported frequencies of drinking among those who drank.

Representing the drinks per week by a central chi-square distribution did not reproduce the targeted mean and distribution for the combined sample of drinkers and non-drinkers. Therefore, the drinks per week were generated using a non-central chi-square distribution. A non-central chi-square random variable, **Y**, is the sum of the squares of the elements of a random multivariate normal vector **X** that follows a normal distribution centered at $\boldsymbol{\mu}$ rather than **0** ($\boldsymbol{Y} = \sum_{\boldsymbol{k}} \boldsymbol{X}^2$; $\boldsymbol{Y} \sim \chi^2_k(\lambda)$; $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$). The shifting of the center of the normal distribution leads to the non-central chi-square distribution having a smaller proportion of the density located near 0 than in the density of the central chi-square distribution (Figure 6-4). Therefore the parameters necessary to generate the values for drinks per week were the degrees of freedom, k, and the non-centrality parameter, λ. Table E contains the parameters for the binomial distributions and the non-central chi-square distributions used to generate the data alcohol consumption data.

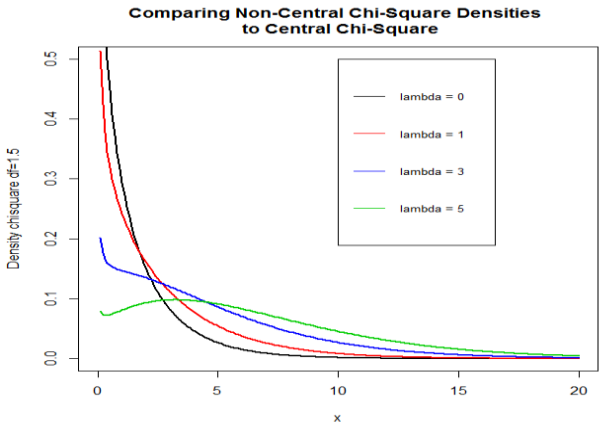

Figure 6-4 Comparing non-central Chi-Square densities to a central Chi-Square density.

Table E Alcohol consumption: Parameters for data generation distributions by AF status.

	Binomial Parameter	Noncentral Chi-Square Parameters	
	Probability of any alcohol consumption	Degrees freedom	Non-Centrality
Never AF	0.504	1.5	11
Eventual AF	0.525	1.6	11.3

The final randomly generated baseline variables represented information about the risk of cardiovascular disease as measured by hypertension. A binary variable reported whether the individual was hypertensive. The hypertension status of an individual was generated using a binomial distribution. Another continuous variable represented the individual's systolic blood pressure. From visual inspection of the CHS data set, displayed in Figure 6-5, a skewed normal distribution appeared appropriate. The skewed normal package by Azzalini (2015) used five parameters: sample size, location, scale, slant, and "hidden mean". The "hidden mean" is used when the mean of the normal distribution is itself a random variable with a normal distribution. Table F contains all the parameters utilized to generate the two variables for hypertension and systolic blood pressure.

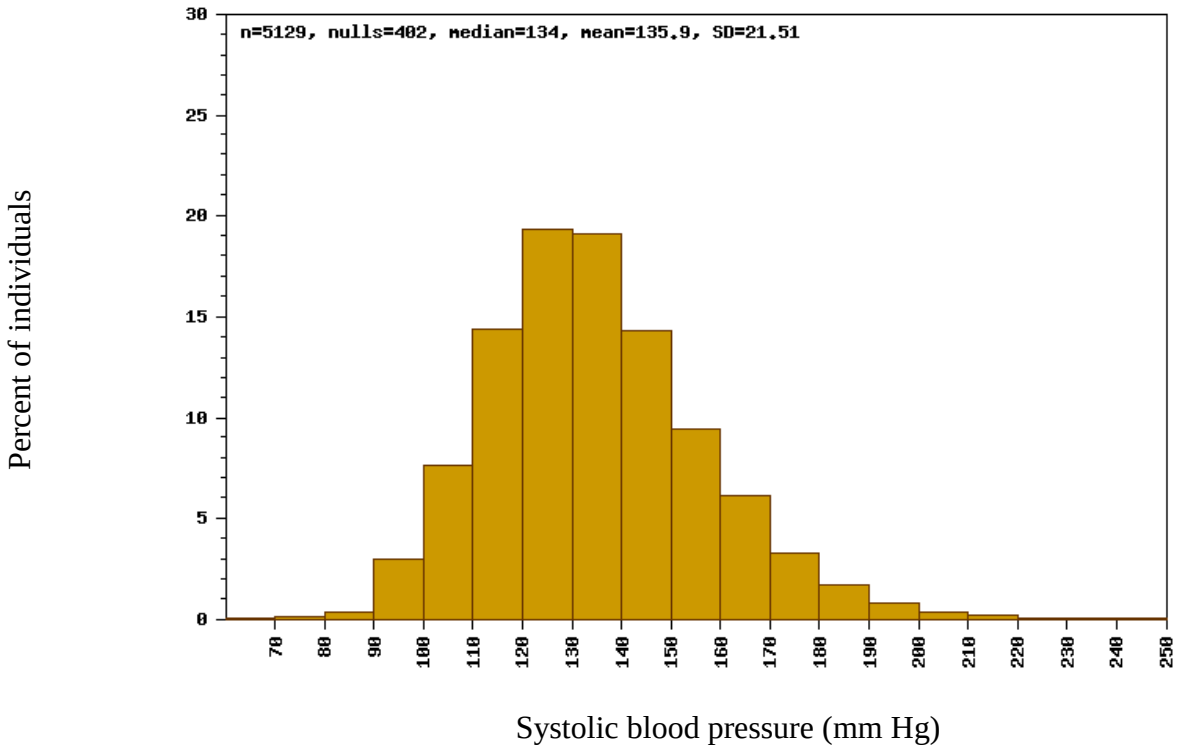

Figure 6-5 Distribution of systolic blood pressure in the CHS sample. On the CHS website the variable name is STDSYS16 and the variable accession is phv00100401.v1.p1.

Table F Hypertension and systolic blood pressure: Parameters for data generation distributions

	Binomial Parameter	Skewed Normal Parameters			
	Probability of being hypertensive	location	scale	slant	Hidden mean
Never AF	0.567	128	27	3	1
Eventual AF	0.612	130	35	3.5	1

Per the study's definition of hypertension, individuals with systolic blood pressure (SBP) at least 140 mm Hg met the criteria for having hypertension. Additional measures were utilized while generating the values for the binary and continuous hypertension related variables to ensure that the information from these two variables was consistent. After the SBP for all individuals had been generated, a count of the number of individuals with at least 140 mm Hg was taken. If this count exceeded the number of individuals assigned to have hypertension by the binomial distribution, then the systolic blood pressures measurements were redrawn. This

process was repeated until the number of people assigned as hypertensive was at least as large as the number of individuals with SBP that met the criteria for hypertension. Finally, when appending the columns to the data set the correct alignment of hypertension status with individuals having high SBP was verified.

While the data set remained in wide format thirteen additional variables (prior smoker, current smoker, birth year, ages at follow-up visit, and age at AF) were derived from the previously randomly generated data values. The two indicator variables, prior smoker and current smoker, would represent in the regression model the smoking behavior of the subjects rather than the three level variable smoking status. Those subjects previously randomly assigned smoking status 1 were assigned 1 in the indicator variable prior smoker. All other subjects were assigned 0 for prior smoker. Individuals whose smoking status had previously been randomly assigned 2 were assigned 1 to the current smoker indicator. Again, all other subjects were assigned 0 for this indicator variable current smoke.

Birth year and ages at follow-up were derived from the randomly generated baseline age. To determine these values, the age at baseline was subtracted from the baseline visit date to establish a birthdate for the subject. For the R program to properly perform the calculations on the date objects, age at baseline was first converted from years to days by multiplying by 365.25. For the baseline covariate birth year, the four digits of the birth year were extracted from the birthdate object. Since all subjects were known to have a birth year in the format 19XX, only the last two digits were recorded for the birth year. The ages at follow-up were calculated by taking the difference between the date of follow-up and the subject's birthdate. To convert the age from days to years, the values were divided by 365.25.

Lastly, an age at which an AF event would occur was generated. Even though this variable was not directly used in the regression analysis, it would be utilized at each visit time to calculate the duration of follow-up after an AF event. For subjects who would eventually have an AF event, generating the age at AF event was a three step process. A list of visit dates from the baseline visit to the penultimate visit was created. This list was sampled with replacement to select a follow-up visit after which subjects would have an AF event. Then a random number from the Uniform(0,1) distribution was generated to determine the fraction of a year that would elapse after the follow-up visit prior to the AF event occurring. Finally the sum of the age at baseline, the randomly selected years of follow-up, and the Uniform(0,1) value was recorded as the age at AF variable. Subjects that would never have an AF event, were merely assigned an age for an AF event that exceeded the reasonable range of age values: 200. This age of AF variable would be essential for all subjects once the data set was converted to long format.

The process outlined above was executed twice to generate separate wide format datasets based on the static status of the occurrence of an AF event. Each dataset contained each individual's visit dates, current age at the visit dates, static AF status, age at AF, and the randomly generated baseline covariate values for birth year, gender, race, diabetes history, years of education through 12^{th} grade, any post high school education, status as a never, prior, or current smoker, any history of alcohol consumption, number of alcoholic drinks per week, diagnosis of hypertension, systolic blood pressure measurement, history of heart disease, and history of heart failure. The two datasets were merged into one large dataset. In the next phase of data generation, this combined dataset was converted to long format so that variables that changed based on the visit date could be added at each subject's appropriate visit.

6.4 SPECIFICS OF DATA GENERATION – LONG FORMAT

From the second phase of data generation a preliminary data set ready for analysis was produced. Unlike the first phase of data generation, in this phase random selection from a distribution representing the population's dispersal of a variable is not the primary technique to assign values to the variables. Rather for all the time-varying covariates the values are mathematically derived from variables randomly generated in the first phase. For the time-varying outcome cognition, the Thacker and colleagues' regression model introduced at the end of Chapter 5 was used to derive the values. The values of the coefficients in the regression model to generate cognition scores were initially selected by educated guess. They were revised during the third step of data generation. The final values for these coefficients are reported later in the paper in Table J.

During the conversion to long format many of the columns from the wide format dataset were removed. In the long format each subject's data were recorded across 10 rows; one row representing the subject's covariates and outcome at one visit date. One column, Visit, was introduced to the long format dataset to identify which row coincided with which visit: baseline, follow-up 1, etc. So when converting to the long format, the ten columns of the wide format dataset containing the visit dates for each subject were collapsed into one column: Date. Similarly the 10 columns containing ages in the wide format were collapsed into one long format column: current age. In the wide format dataset the time varying nature of having an AF event was recorded across two columns: the static AF status, and age at AF. For the long format data set, the information in these two columns was combined into one time-varying variable: AF indicator. At each visit, if the current age was less than the age at AF, the AF indicator was set to 0. For visits with current ages meeting or exceeding the age at AF, the AF indicator value was

set to 1. Since the baseline covariates remained constant across time, the values in each column was replicated across the ten rows for each subject.

In order to calculate the individual's cognition scores from the regression model, several new columns containing derived variables were appended to the long format dataset. For visit times when the AF indicator was set to 1, the subject's amount of follow-up time since his or her AF event was recorded by subtracting the age at AF from the current age at the visit. When no AF event had occurred, the follow-up time since AF was designated as 0. Centered current age at each visit was calculated by subtracting the mean of all the baseline ages from the current age. Centered current age polynomial terms, through the fifth degree, were produced by raising the centered current age to the appropriate power.

Next the randomly generated intercept, centered age slope, squared centered age slope, and error term for determining the cognition scores at each visit were attached to the dataset. Similar to the first phase, each randomly generated variable was selected from a univariate distribution which was intended to represent variation of the values in the entire population. Using three independent univariate distributions to model the random effects differed from what Dr. Thacker et al. allowed for in their analysis. They assumed an unstructured covariance matrix for the random effects. In the simulations the random intercepts were generated from a standard normal distribution, while the random slopes and error terms came from truncated normal distributions. Table G contains the parameter values for each distribution. These parameters were chosen to generate cognition scores for all individuals that would remain within a plausible range. They were determined by trial and error.

Table G Parameters for random effects and error term distributions

	Sample size	Mean	SD	Left trunc value	Right trunc value
Random intercept	5,000	0	1	NA	NA
Random slopes	5,000	0	0.25	-1	1
Error term	50,000	0	0.75	-1.75	1.75

The random effects for intercept, centered age slope, and squared centered age slope were generated only once for each subject, then replicated across all ten rows for that subject. Maintaining identical random effects for a subject ensured that they were modeling how the cognition scores were correlated within a subject across visits. The error term was generated independently for each visit and subjectHaving generated the random effects and error terms, the dataset now contained in each row all covariate values, random effects, and error terms needed to determine the individual's cognition at each visit. The initial guess for true beta values were stored in a vector. The cognition scores were calculated using the appropriate multiplications and sums as noted in the data generation regression models below. The data dictionary for the coding of the variables is located in Table H.

Thacker and colleagues' data generation model:

$$
\begin{aligned}
Cog_{ij} = {} & \beta_0 + b_{0i} + \beta_1(CAC_{ij}) + b_{1i}(CAC_{ij}) + \beta_2(CAC_{ij})^2 + b_{2i}(CAC_{ij})^2 + \beta_3(CAC_{ij})^3 \\
& + \beta_4(CAC_{ij})^4 + \beta_5(CAC_{ij})^5 + \beta_6(AFIndic_{ij}) + \beta_7(FUPostAF_{ij}) \\
& + \beta_8(CAC_{ij})(FUPostAF_{ij}) + \beta_9(CAC_{ij})^2(FUPostAF_{ij}) + \beta_{10}(BirthYear_i) \\
& + \beta_{11}(Male_i) + \beta_{12}(BRace_i) + \beta_{13}(Diab_i) + \beta_{14}(YrEd12_i) \\
& + \beta_{15}(AnyEdPost12_i) + \beta_{16}(PriorSmk_i) + \beta_{17}(CurrSmk_i) + \beta_{18}(AnyAlc_i) \\
& + \beta_{19}(DrinkPerWk_i) + \beta_{20}(Hypertension_i) + \beta_{21}(SBP_i) + \beta_{22}(HrtDis_i) \\
& + \beta_{23}(HrtFl_i) + \beta_{24}(Male_i)(CAC_{ij}) + \beta_{25}(BRace_i)(CAC_{ij}) \\
& + \beta_{26}(Diab_i)(CAC_{ij}) + \beta_{27}(YrEd12_i)(CAC_{ij}) + \beta_{28}(AnyEdPost12_i)(CAC_{ij}) \\
& + \beta_{29}(PriorSmk_i)(CAC_{ij}) + \beta_{30}(CurrSmk_i)(CAC_{ij}) + \beta_{31}(AnyAlc_i)(CAC_{ij}) \\
& + \beta_{32}(DrinksPerWk_i)(CAC_{ij}) + \beta_{33}(Hypertension_i)(CAC_{ij}) \\
& + \beta_{34}(SBP_i)(CAC_{ij}) + \beta_{35}(HrtDis_i)(CAC_{ij}) + \beta_{36}(HrtFl_i)(CAC_{ij}) + \varepsilon_{ij}
\end{aligned}
$$

Schaal data generation model:

$$
\begin{aligned}
Cog_{ij} = {} & \alpha_0 + a_{0i} + \alpha_1(CAC_{ij}) + a_{1i}(CAC_{ij}) + \alpha_2(CAC_{ij})^2 + a_{2i}(CAC_{ij})^2 + \alpha_3(AFIndic_{ij}) \\
& + \alpha_4(FUPostAF_{ij}) + \alpha_5(CAC_{ij})(FUPostAF_{ij}) + \alpha_6(CAC_{ij})^2(FUPostAF_{ij}) \\
& + \alpha_7(BirthYear_i) + \alpha_8(Male_i) + \alpha_9(BRace_i) + \alpha_{10}(Diab_i) + \alpha_{11}(YrEd12_i) \\
& + \alpha_{12}(AnyEdPost12_i) + \alpha_{13}(PriorSmk_i) + \alpha_{14}(CurrSmk_i) + \alpha_{15}(AnyAlc_i) \\
& + \alpha_{16}(DrinkPerWk_i) + \alpha_{17}(Hypertension_i) + \alpha_{18}(SBP_i) + \alpha_{19}(HrtDis_i) \\
& + \alpha_{20}(HrtFl_i) + \alpha_{21}(Male_i)(CAC_{ij}) + \alpha_{22}(BRace_i)(CAC_{ij}) \\
& + \alpha_{23}(Diab_i)(CAC_{ij}) + \alpha_{24}(YrEd12_i)(CAC_{ij}) + \alpha_{25}(AnyEdPost12_i)(CAC_{ij}) \\
& + \alpha_{26}(PriorSmk_i)(CAC_{ij}) + \alpha_{27}(CurrSmk_i)(CAC_{ij}) + \alpha_{28}(AnyAlc_i)(CAC_{ij}) \\
& + \alpha_{29}(DrinksPerWk_i)(CAC_{ij}) + \alpha_{30}(Hypertension_i)(CAC_{ij}) \\
& + \alpha_{31}(SBP_i)(CAC_{ij}) + \alpha_{32}(HrtDis_i)(CAC_{ij}) + \alpha_{33}(HrtFl_i)(CAC_{ij}) + \varepsilon_{ij}
\end{aligned}
$$

Table H Data dictionary for the coding of regression covariates

Abbreviation	Meaning	Variable type; values of variable
CAC	Continuous age centered	Continuous variable; Years which can be represented as a decimal
AFindic	Atrial fibrillation indicator	Indicator; 1-had AF 0- has not had AF
FUPostAF	Follow-up post atrial fibrillation	Continuous variable; Years which can be represented as a decimal
BirthYear	Birth year	Years since 1900; only integer values allowed
Male	Gender	Indicator; 1-Male 0- Female
BRace	Black Race	Indicato;r 1-black race 0 - other
Diab	History of Diabetes	Indicator; 1-had diabetes 0 - did not have diabetes
YrEd12	Years of education	Continuous variable; Years which can be represented as a decimal
AnyEdPost12	Education post high school	Indicator; 1-some education post high school 0- no post high school education
PriorSmk	History of prior smoking	Indicator; 1 - Prior Smoker 0- No prior smoke
CurrSmk	Currently smoking	Indicator; 1 - Current smoker 0- Not current smoker
AnyAlc	History of alcohol use	Indicator; 1 - drinker at any time 0 - does not drink
DrinkPerWk	Drinks per week	Continuous variable; which can be represented as a decimal
Hypertension	History of hypertension	Indicator; 1-had hypertension 0- did not have hypertension
SBP	Systolic Blood Pressure	Continuous variable; mm HG
HrtDis	History of heart disease	Indicator; 1-had heart disease 0- has not had heart disease
HrtFail	History of heart failure	Indicator; 1-had heart failure 0- has not had heart failure

6.5 SPECIFICS OF DATA GENERATION – VERIFY THE PLAUSIBILITY OF GENERATED DATA OBSERVATIONS

As noted previously, trial and error was relied upon extensively when determining the values for parameters of a continuous covariate's distribution(s).When establishing the parameterization for the distribution of a covariate I could be changing numerous attributes: mean, variance, non-centrality parameter, etc. My aim was to generate a sample that had similar mean and variance as the Thacker and colleagues' sample for the components used to describe the characteristic (i.e. education needed to have both the YearsED and AnyPost12). If the results in one sample were reasonably close to the values from Thacker et al.'s sample, I accepted those values for the parameters of the covariate's distribution. I then took the long run average from a thousand samples. Table I contains the summary statistics for a data set generated with the seed 6468, the average of 1000 datasets, and Thacker and colleagues' sample. In the single sample, the summary statistics for all the covariates are close to those reported by Thacker and colleagues. Over 1000 simulated samples, the average of the results came even closer to the Thacker et al targets. However, the simulated samples tended to be about half a year older, half a grade level less educated, and consume 1.3 beverages per week more than the sample that Thacker and his colleagues collected. These differences between the simulated data set and the sample from Thacker and his colleagues were considered acceptable and no further adjustments were made to the parameters.

When establishing the regression coefficients which would be used to generate the cognition scores, the overall objective was to choose values that would generate an average cognition trajectory similar to Figure 1 from Thacker and colleagues. It was also necessary to know that the individual cognitive scores at a specific age would be on the correct order of

magnitude. With a seed set to 6468, cognition scores for each individual at each observation were generated using the person's covariates, random effects and the guess for regression coefficients. It was deemed that the regression coefficients identified as "True Beta Values" in Table J produced a mean cognitive trajectory with a plausible spread for cognition scores. Table K itemizes some age specific summary statistics for the cognition generated using the final set of regression coefficients. Figure 6-6 shows the scatterplot of cognition scores from seed 6468. A loess smoother overlaid in blue indicates the trajectory of the mean cognition without accounting for AF status. It is noted that, compared to Figure 1 from Thacker et al., the cognitive trajectory from my simulated data is not as flat between the ages of 65 to 80 as the predicted trajectory for individual's without an AF. (For reference Figure 1 from Thacker et al. is reproduced as Figure 6-7). In addition the simulated cognition scores start at a slightly higher value at age 65 than the observed scores. These differences were deemed acceptable because this research would not be comparing simulated results with the observed values in Thacker and colleagues. Rather, the comparisons of the contrasts from the two analysis models would occur from analyses that were both conducted on identical simulated data sets.

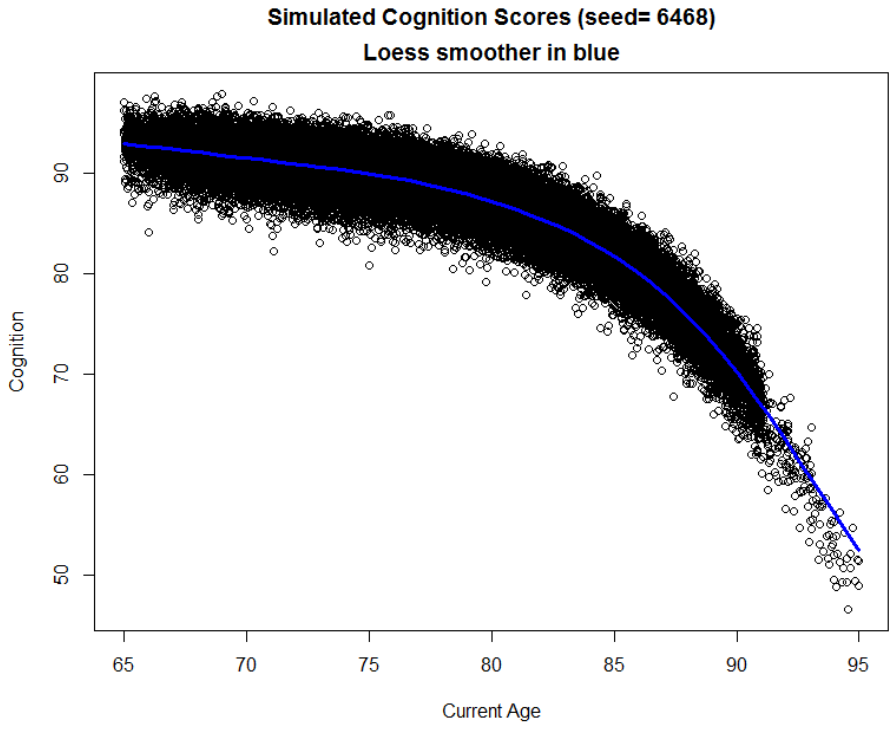

Figure 6-6 Simulated cognition scores with seed = 6468. (Blue line is the LOESS smoother.)

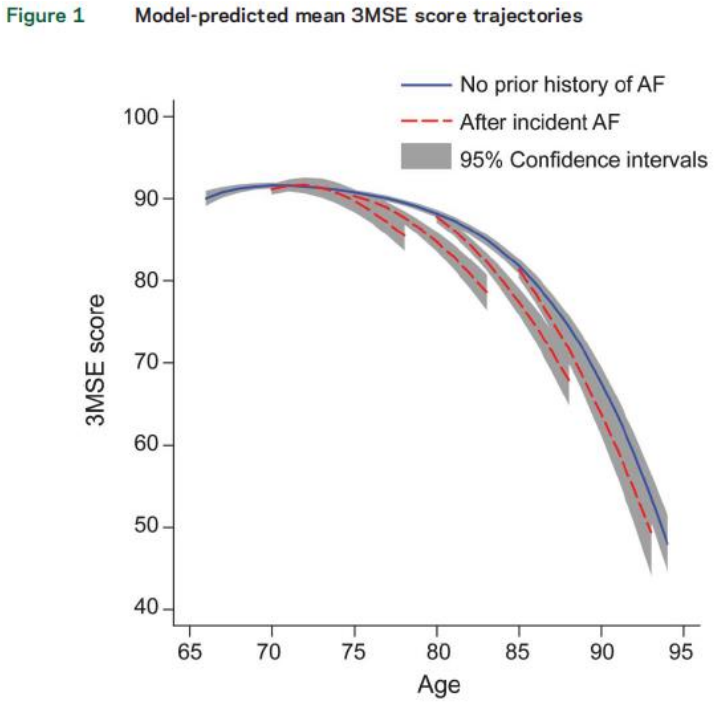

Figure 6-7 Figure 1 from Thacker et al. 2015

Table I Summary of baseline covariates for three situations: One sample generated when the seed was 6468, the average of 1000 simulations, and Thacker and colleagues' results.

	Entire Dataset			NOAF Dataset			AF Dataset		
	Seed 6468	1000 Sims	Thacker	Seed 6468	1000 Sims	Thacker	Seed 6468	1000 Sims	Thacker
Age 65 - 69: %	36.1	35.0	NA	36.2	35.0	NA	35.4	35.2	NA
Age 70 - 74: %	23.9	25.0	NA	24.3	25.0	NA	21.2	25.0	NA
Age 75 - 79: %	19.8	20.0	NA	19.5	20.0	NA	21.8	19.9	NA
Age 80 or older: %	20.2	20.0	NA	20.0	20.0	NA	21.6	19.9	NA
Had atrial fibrillation: %	11.3	10.7	10.7	0	0	0	100	100	100
Baseline age: Mean (SD)	73.5 (5.5)	73.5 (5.5)	73.0 (5.4)	73.4 (5.4)	73.4 (5.4)	72.9 (5.3)	74.0 (6.1)	73.8 (6.0)	74.4 (5.8)
Birth year: Mean (SD)	13.5 (5.5)	13.5 (5.5)	17 (5.5)	13.6 (5.4)	13.5 (5.4)	17 (5.4)	12.9 (6.0)	13.1 (6.0)	15 (5.9)
Male: %	39.4	41.2	41.2	37.7	39.9	39.9	52.9	51.9	52.0
Black race: %	15.7	15.2	15.2	16.6	16.0	16.0	8.8	9.1	9.2
Education through 12^{th}: Mean (SD)	10.6 (2.1)	10.6 (2.1)	11.0 (2.0)	10.6 (2.1)	10.6 (2.0)	11.0 (2.0)	10.2 (2.5)	10.3 (2.4)	11.0 (2.0)
Any education post 12^{th}: %	43.0	43.4	43.5	42.9	43.3	43.3	43.4	45.1	45.1
History of prior smoking: %	41.9	41.7	41.8	41.4	41.6	41.6	46.0	43.1	43.1
Current smoking: %	11.4	11.7	11.7	11.5	11.9	11.9	11.0	10.5	10.5
Any alcohol use: %	50.9	50.6	50.7	50.2	50.4	50.4	56.8	52.5	52.5
Drinks per week: Mean (SD)	6.3 (7.9)	6.3 (7.9)	5.0 (8.4)	6.2 (7.9)	6.3 (7.9)	5.0 (8.5)	7.2 (8.4)	6.8 (8.2)	5.1 (7.7)
History of diabetes: %	15.2	14.8	14.9	14.8	14.7	14.7	18.2	16.3	16.3
Hypertension: %	56.6	57.1	57.2	56.0	56.7	56.7	61.2	61.0	61.2
Systolic blood pressure : Mean (SD)	136.0 (22.8)	135.9 (22.8)	136.0 (21.5)	135.5 (22.1)	135.4 (22.0)	135.7 (21.4)	140.4 (27.7)	139.7 (28.3)	138.1 (22.6)
History of heart disease: %	18.2	18.2	18.2	16.8	16.9	16.9	29.0	28.6	28.6
History of heart failure: %	3.0	3.3	3.3	2.7	3.1	3.1	5.1	5.2	5.3

Table J a) Validation that the simulation process will correctly estimate the True Beta Values: Sim 1

Sim 1 Data Gen: Thacker	True β	Mean of $\hat{\beta}$	SE ($\hat{\beta}$)	Mean of SE($\hat{\beta}$)
Intercept	87.6	87.60141	3.78E-03	0.0643
Current age	-0.10043	-0.10044	3.67E-04	0.0467
(Current age)^2	-0.00243	-0.0024	2.21E-05	0.0483
(Current age)^3	-0.00115	-0.00115	3.16E-06	0.0824
(Current age)^4	-0.00016	-0.00016	0.00E+00	0.1933
(Current age)^5	3.00E-06	0.000003	0.00E+00	0.1218
Follow-up post AF	-0.28	-0.27995	5.41E-04	0.0169
AF indicator	-0.4	-0.40087	1.49E-03	0.0453
Birth year	0.15	0.15001	1.68E-04	0.0024
Male	-0.3	-0.30023	4.43E-04	0.0147
Black race	-1.4	-1.39931	6.01E-04	0.0202
History of diabetes	-1.7	-1.69996	5.95E-04	0.0203
Year of education through 12th	0.19	0.190091	1.30E-04	0.0044
Any education post 12th	0.35	0.349395	5.34E-04	0.0184
History of prior smoking	-0.25	-0.2499	4.49E-04	0.0154
Current smoking	-1.5	-1.49891	6.99E-04	0.0237
Any alcohol use	-0.21	-0.20882	6.74E-04	0.0236
Drinks per week	-0.04	-0.04007	4.43E-05	0.0015
Hypertension	0.145	0.145847	6.26E-04	0.0235
Systolic blood pressure	-0.00443	-0.00445	1.26E-05	0.0116
History of heart disease	-1.1	-1.10014	5.66E-04	0.0188
History of heart failure	-1.7	-1.69924	1.22E-03	0.0406
(Current age)(Follow-up Post AF)	0.014545	0.01456	1.20E-04	0.0221
(Current age)^2(Follow-up Post AF)	-0.00041	-0.00041	6.32E-06	0.0153
(Current age)(Male)	-0.00182	-0.00184	6.64E-05	0.0128
(Current age)(Black race)	-0.01	-0.01001	8.85E-05	0.0176
(Current age)(Diabetes)	-0.00109	-0.00114	8.54E-05	0.0177
(Current age)(Years education up to 12th)	-0.01364	-0.01364	1.90E-05	0.0038
(Current age)(Any education post 12th)	0.014545	0.014516	8.22E-05	0.016
(Current age)(History of prior smoking)	-0.01091	-0.01091	6.64E-05	0.0134
(Current age)(Current smoking)	-0.01546	-0.01555	1.04E-04	0.0206
(Current age)(Any alcohol use)	-0.01636	-0.01644	9.80E-05	0.0205
(Current age)(Drinks per week)	-0.00055	-0.00055	6.32E-06	0.0013
(Current age)(Hypertension)	0.004	0.00389	9.80E-05	0.0204
(Current age)(Scaled systolic blood pressure)	0.000137	0.000138	3.16E-06	0.0101
(Current age)(History of heart disease)	-0.01364	-0.01368	8.22E-05	0.0164
(Current age)(History of heart failure)	-0.01455	-0.01487	1.74E-04	0.0353

Table J b) Validation that the simulation process will correctly estimate the True Beta Values: Sim 2

Sim 2 Data Gen: Schaal	True β	Mean of $\hat{\beta}$	SE ($\hat{\beta}$)	Mean of SE($\hat{\beta}$)
Intercept	87.6	87.6008	3.53E-03	0.0611
Current age	-0.10043	-0.10022	3.45E-04	0.0432
(Current age)^2	-0.043	-0.04299	6.32E-06	0.0115
Follow-up post AF	-0.28	-0.28003	5.22E-04	0.0169
AF indicator	-0.4	-0.40006	1.48E-03	0.0452
Birth year	0.15	0.149903	1.64E-04	0.0023
Male	-0.3	-0.30064	4.46E-04	0.0147
Black race	-1.4	-1.40014	6.10E-04	0.0202
History of diabetes	-1.7	-1.69901	6.17E-04	0.0203
Year of education through 12th	0.19	0.189983	1.36E-04	0.0044
Any education post 12th	0.35	0.350482	5.38E-04	0.0184
History of prior smoking	-0.25	-0.25012	4.40E-04	0.0154
Current smoking	-1.5	-1.5004	7.12E-04	0.0237
Any alcohol use	-0.21	-0.20917	6.99E-04	0.0236
Drinks per week	-0.04	-0.04001	4.43E-05	0.0015
Hypertension	0.145	0.144931	6.29E-04	0.0235
Systolic blood pressure	-0.00443	-0.00443	1.26E-05	0.0116
History of heart disease	-1.1	-1.09997	5.53E-04	0.0188
History of heart failure	-1.7	-1.70115	1.27E-03	0.0406
(Current age)(Follow-up Post AF)	0.014545	0.014611	1.14E-04	0.021
(Current age)^2(Follow-up Post AF)	-0.00041	-0.00041	6.32E-06	0.0137
(Current age)(Male)	-0.00182	-0.00182	6.64E-05	0.0128
(Current age)(Black race)	-0.01	-0.00999	8.85E-05	0.0176
(Current age)(Diabetes)	-0.00109	-0.00132	8.85E-05	0.0177
(Current age)(Years education up to 12th)	-0.01364	-0.01368	1.90E-05	0.0038
(Current age)(Any education post 12th)	0.014545	0.014633	7.91E-05	0.016
(Current age)(History of prior smoking)	-0.01091	-0.0108	6.64E-05	0.0134
(Current age)(Current smoking)	-0.01546	-0.01538	1.04E-04	0.0206
(Current age)(Any alcohol use)	-0.01636	-0.01633	1.01E-04	0.0205
(Current age)(Drinks per week)	-0.00055	-0.00056	6.32E-06	0.0013
(Current age)(Hypertension)	0.004	0.003954	9.80E-05	0.0204
(Current age)(Scaled systolic blood pressure)	0.000137	0.000138	3.16E-06	0.0101
(Current age)(History of heart disease)	-0.01364	-0.0137	7.91E-05	0.0164
(Current age)(History of heart failure)	-0.01455	-0.01476	1.77E-04	0.0353

Table K Summary Statistics for Cognition Scores predicted at various ages for the entire data set when the seed was set to 6468.

	Mean	Min	Max
Age 70	91.3	84.44	97.18
Age 75	89.62	80.78	95.78
Age80	86.77	80.11	92.72
Age 85	80.89	71.92	86.9
Age 90	68.64	61.54	75.45
Age 94	51.31	46.6	56.3

6.6 OVERVIEW OF DATA ANALYSIS

Using the data generated in the previous steps to understand how an AF event would affect the cognitive score of elderly subjects required multiple steps of data analysis. First, we needed to analyze the entire data set to obtain estimates of the regression coefficients, $\vec{\boldsymbol{\beta}}$. We used the observed values $(\vec{\boldsymbol{X}}, Y)$ from all subjects in the sample to estimate how cognition changed with alterations in a covariate's value. Next, we created predictions for the average cognitive trajectories of various subgroups of the population. Using the set of estimates of the regression coefficients and values of the covariates for a specific subgroup $\{\vec{\boldsymbol{\beta}}, \mathbb{X}\}$ we generated Y values at successive ages for an average person in that group. Last, we obtained estimates of the contrasts in the predicted average cognition for the various groups we investigated.

We investigated the same two types of contrasts which Thacker et al reported in their paper: effectOfAF and AFdrop. In the effectOfAF contrast we compared the predicted cognitive values at the same ages for the counterfactual situation: the average individual who never had an AF event and the average individual after having experienced an AF event five years earlier. There were four ages at which cognition was compared: 75, 80, 85, and 90. The next type of contrast, AFdrop, recorded the rate of cognitive decline in cognition for people who had initially experienced an AF event five years earlier. This contrast of AFdrop was analyzed at various

ages: 70, 75, 80, and 85 years. The data analysis concluded with a set of eight contrasts: effectOfAF at 4 ages (75, 80, 85, 90) and AFdrop from 4 ages (70, 75, 80, 85).

We applied all of these steps separately to each of the proposed regression models allowing us to have three different stories about how an AF event affected cognition trajectories. One story was known to be the truth because the regression coefficient values used to simulate the data were substituted when making the predictions. The other two stories came from the fitted regression model used by Thacker and colleagues and the fitted simpler regression model proposed by myself.

6.7 SPECIFICS OF DATA ANALYSIS

Determining the input, $\vec{\boldsymbol{\beta}}$, to be used to create predictions for the cognition in the five groups was the first step in the data analysis. The observed values $(\vec{\boldsymbol{X}}, Y)$ from all subjects in the sample created during the data generation phase were analyzed using random effects regression models from the LME4 package in R. Several judgment decisions were required by the researchers to perform the data analysis. By default, the *lmer* function in LME4 uses the restricted maximum likelihood (REML) method to estimate the regression coefficients (Bates, et al., 2015). Oehlert (2012) provides a summary of this method and the benefits and drawbacks of its use. REML is a method for estimating the variances of the random effects in a mixed effects model, based on a model for the distribution of a restricted subset of the residuals chosen so that their distribution is independent of the regression coefficients. Both REML and MLE estimates are asymptotically unbiased, but in finite samples, particularly when the number of parameters p is large compared to the sample size n, maximum likelihood estimates of variance covariance parameters are biased and REML estimates of them are not biased. As a simple example, in ordinary linear regression, standard maximum likelihood methods divide the estimate of the

variance by n which gives biased estimates of the variance when sample sizes are finite, and the REML estimate divides the variance estimate by $n - p$ (where p is the number of parameters estimated) rather than n, thus giving an unbiased estimate of the variance.

In mixed models, this correction for the number of fixed effects also occurs when calculating the restricted likelihood value and maximizing it to obtain regression coefficients. As a result of correcting for the number of fixed effects parameters in the model, the likelihood functions that are maximized during the REML method are no longer equivalent in two models having different numbers of fixed effects. Thus the REML method is not appropriate for likelihood based methods of comparing models that have different numbers of fixed effects.

During their model selection process Thacker and colleagues decided between models with differing numbers of fixed effects. When deciding how to model the polynomial representation of age and its interactions, they used BIC, a likelihood based criterion, to compare their possible models. Likewise, our general interest was with comparing the performance of nested models. The models we compared differed only in how the main effect terms for centered age were represented by the polynomial. The first model utilized by Thacker and colleagues included polynomial terms through the fifth power (see section 4.3). While our proposed model contained only the linear and quadratic terms (see section 4.3). We ultimately did not use a likelihood based method, such as BIC, to compare the performance of the models. But the initial interest in using these methods guided the decision to override the default of the *lmer* function and use standard maximum likelihood methods rather than REML methods to analyze the data.

Even though the two models had different numbers of fixed effects, the analysis models both employed three random effects: a random intercept, a random slope for the linear centered age term, and a random slope for the quadratic centered age term. The *lmer* package allowed for

several different methods of inputting the three random effects. I opted for the notation that allowed for some correlation between the random effects:

(1 + Scaled_Center_Age + Scaled_Squared_Center_Age)|Subject.

This notation allows for the subject specific random intercept, scaled centered age slope, and scaled squared centered age slope to be correlated within that subject. For example, subjects who tended to have higher than average intercepts could also have rates of change for the scaled center age that were higher than average but rates of change for the scaled squared centered age that could have been lower than average. This decision followed Dr. Thacker and colleagues' analysis that assumed an unstructured correlation matrix for the random effects. (However, in both simulations the random effects for a subject were generated independently of each other.)

When trying to fit the data set with the observed values, using either the Thacker and colleagues' model or my model, the *lmer* function generated warnings regarding convergence issues. The continuous variables for the polynomial terms of centered age and systolic blood pressure tended to have values that were orders of magnitude different from the other covariate values. Because of this, we scaled the polynomial age terms by their respective standard deviations from the observed data. The systolic blood pressure was centered by its observed mean and then scaled by its standard deviation. Interaction terms that involved a continuous variable were scaled by the corresponding standard deviation. Then the analysis was rerun on the data set with the scaled continuous variables and unscaled binary covariates. Appendix D has the derivation for the back-transformation of the regression coefficients from the scaled data set. The back-transformed regression estimates were saved as the $\vec{\boldsymbol{\beta}}$ input used to predict the cognition for the various subgroups.

Prior to moving to the second step of data analysis, we verified that the analysis was producing unbiased results. Using the *lmer* package in R, each dataset was analyzed using the regression model that generated the cognition scores. Then the estimates of the coefficients ($\hat{\beta}$) and the estimates of the standard deviations of the coefficients ($\widehat{sd}(\hat{\beta})$) were averaged. Table J contains a comparison of the true coefficient values used to generate the data and the mean of the 1000 $\hat{\beta}$'s. (Since we actually did two simulation studies, there is one table when Thacker and colleagues's model was used to generate the data, and a second table when Schaal's simpler model was used to generate the data.) The virtually identical results for each coefficient indicated that the data simulation and analysis methods were working correctly.

The second step in the data analysis was to generate predictions for the cognitive trajectory of an average individual in each of the five AF scenarios (the scenarios will be explained later). But first I identify the five different subgroups for whom a design matrix was generated. It was anticipated that the age at which the AF event occurred would impact cognition differently. Therefore the prediction of cognitive trajectories was computed for five different groups: individuals who through the end of the study would never have an AF event and individuals who progressed to experiencing the AF at either 70, 75, 80, or 85 years of age. To obtain predictions for each subgroup, indexed by *i*, we needed a design matrix that contained the values of its covariates.

As explained in Chapter 4, the design matrix $\mathbb{X}_i$ contains the values of all covariates for people of group *i,* at each observation. Previously, for a specific individual, the AF status and time since AF entries in the design matrix were allowed to change from one observation to the next. However, the interest in comparing cognitive predictions across groups that are defined by AF status and time since AF, necessitated that these values remained constant within a design

matrix that represented each of the five groups. Thus, for the design matrix in the prediction model, the only remaining time varying covariate was centered age.

$$\mathbb{X}_i = \begin{bmatrix} 1 & centeredAge_{i,j=1} & AF_i & PostAF_i & \overline{male} & \dots & \left(centeredAge_{i,j=1}\right)\left(\overline{HrtFl}\right) \\ 1 & centeredAge_{i,j=2} & AF_i & PostAF_i & \overline{male} & \dots & \left(centeredAge_{i,j=2}\right)\left(\overline{HrtFl}\right) \\ 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & centeredAge_{i,j=n} & AF_i & PostAF_i & \overline{male} & \dots & \left(centeredAge_{i,j=n}\right)\left(\overline{HrtFl}\right) \end{bmatrix}$$

The time varying nature of a covariate is indicated by the subscript, *j*, which is only found on centered age. To obtain a smooth trajectory the ages that were modeled for a given subgroup were 0.05 years apart. Thus if at observation , j=1, the equation was modeling 70 year olds, then at observation, j=2, we were modeling the average cognition of the group of 70.05 year olds with the same values for other covariates. The design matrix for the group of subjects who never had an AF event had current age entries from 65 to 95 years old, while the design matrix for subjects who had an AF event were followed from the age of their AF event until five years post AF.

Depending on which of the five subgroups the design matrix represented, the time-invariant variables represented by the single subscript *i*, took on different values. The group of individuals who would never have an AF event were assigned an AF indicator value and time since AF value both equal to zero. The group of individuals who would have an AF event had the AF indicator set to one, while follow-up time after an AF event was predetermined to be five years.

We decided to use the population average for each of the baseline covariates. Since we intended to utilize the population average, the specific subgroup that we were making predictions for would not influence the value for the covariate. The lack of a subscript, *i*, for all of the baseline covariates indicates that these values were constant across each of the five subgroups. As explained previously, the process of selecting distributional parameters for data generation of

some of the covariates was complex. Obtaining the true mean in the population from which the data were generated for all of the covariates was beyond the scope of this project. Therefore, the long run average of 1000 simulated data sets was used as an estimate of the population mean for each covariate (Table I). These values estimated the expected covariate measurement for a typical individual from the population.

After the design matrixes for all subgroups were completed, both elements from the set of inputs necessary to create the cognitive predictions for a subgroup existed. Using the set of input $\{\vec{\boldsymbol{\beta}}, \mathbb{X}_i\}$ relevant for each subgroup, the vector of predictions $\vec{\boldsymbol{Y}}$ for the cognitive values across the appropriate age range were computed and stored. Once the cognitive predictions for all subgroups were made, contrasting the cognitive trajectories in the subgroups, could be completed.

As introduced previously there were two types of contrasts, effectOfAF and AFdrop, that were extracted from these sets of predictions. For example, the effectOfAF contrast was the difference in the predicted cognitive value for the average 75 year old who never had an AF event versus the average 75 year old who previously had an AF event at age 70. In this contrast the current age was held constant in the two groups being compared. The rate of cognitive decline in people who had an AF event, AFdrop, was calculated by subtracting the predicted cognition for an average individual at two different time points. For example, the difference in predicted cognition at 85 and 80 years, for people who had an AF at 80 years old. When all of the relevant contrasts had been computed, a data set was summarized with eight contrasts: effectOfAF at 4 ages (75, 80, 85, 90) and AFdrop from 4 ages (70, 75, 80, 85). The entire process from data generation to calculation of contrasts was executed 1000 times. Then we computed the mean and standard deviations for each of the eight contrasts in the 1000 samples.

The interest of this simulation research was to compare the stories that two methods of analysis tell about the effect of an AF event on cognitive trajectories in the elderly. Two simulation studies were performed that utilized different regression models to generate the cognition for each individual. The first simulation study, Sim1, generated each participant's cognition score using the original Thacker fifth degree polynomial model. In the second simulation, Sim2, cognitive scores were computed using only the quadratic model. For each simulation study, all three steps of data analysis (estimation of $\hat{\beta}$, prediction of average cognition, and calculation of the estimates for the contrasts effectOfAF and AFdrop) performed on the data set were identical. Each simulation study analyzed the generated data twice using each of the two proposed analysis models: 1)Thacker and colleagues and 2)Schaal. To determine a benchmark for the truth against which the results of these two analysis models could be compared a third analysis was performed. In this third analysis, called "Truth", the regression parameters known to have generated the cognition scores were used in place of regression estimates of $\hat{\beta}$. The latter two steps of the data analysis (prediction of average cognition, and calculation of the estimates for the contrasts effectOfAF and AFdrop) were performed identically as previously described. Repeating the simulation study using different underlying models of truth allowed us to investigate whether the strengths (or weaknesses) of the simpler regression model were broadly applicable or depended upon the structure of the underlying truth in the population.

Chapter 7. COMPARISON OF SIMULATION RESULTS

Through their research Thacker and his colleagues contributed information about the effect of an incident atrial fibrillation event on cognitive decline in the elderly. Two different simulation studies were performed to determine whether a simpler analysis model could perform better at estimating this effect than the original fifth degree polynomial model which Thacker et al used. Note the only difference between the true data generation models in Sim1 and Sim2 was the exclusion of fixed-effect terms for the third, fourth, and fifth degree of centered age.

A cursory inspection of the graphical representation of the simulation results provides no unexpected revelations. First, when we inspect the predicted cognitive trajectories in the group without an AF we see that the model misspecification performs as expected. In Sim 1, when there are omitted variables we see that the misspecified model is not capturing the truth (Figure 7-1). The Schaal model does not follow the red line that indicates the true data generation model while the Tacker et al model overlays the truth. In Sim 2, when the misspecified model contains additional terms that are not necessary we see that this misspecification does not add bias (Figure 7-2). All three models are tracing the identical trajectory and even when jittered the three models continue to trace virtually the same trajectories. The Schaal model displayed in green is beneath the other two models.

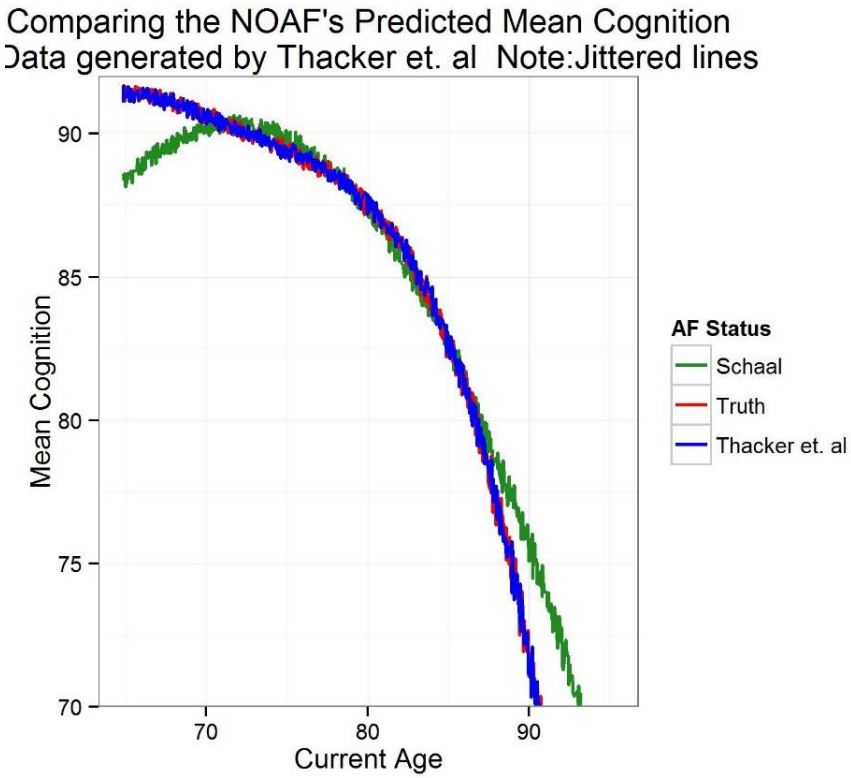

Figure 7-1 Simulation 1 predicted cognition for NOAF group from all three analysis models

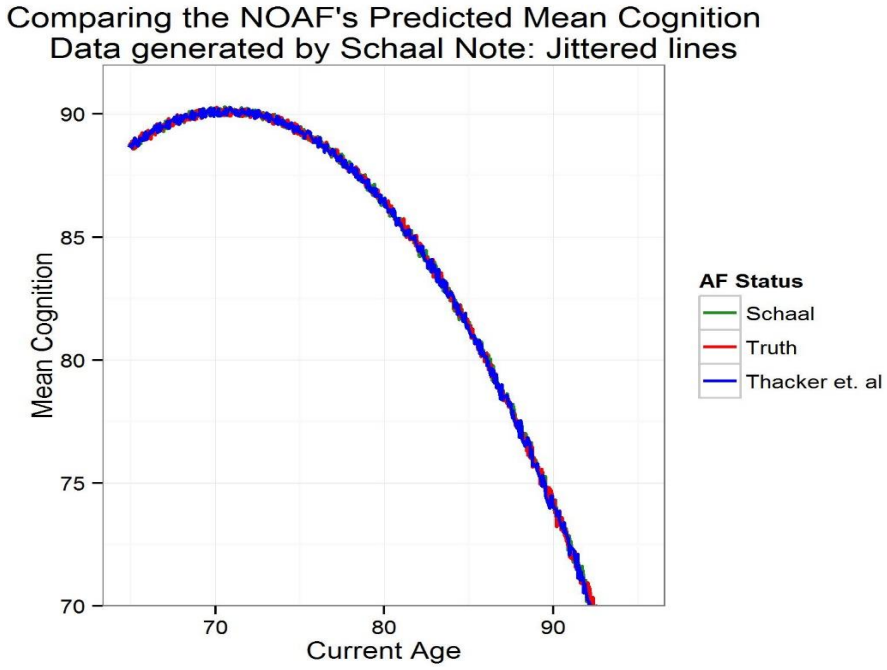

Figure 7-2 Simulation 2 predicted cognition for NOAF group from all three analysis models

When attempting to look at patterns in the predicted cognitive trajectories for all five age scenarios, the visualization becomes too cluttered when including the three models (Schaal, Truth, and Thacker et al.) in one graph. At this point we are still gleaning initial observations from the visualization of the results. Later we will examine the performance of the different models with tabular presentation of the numeric results. Figure 7-3, Figure 7-4, and Figure 7-5 each contain the cognitive trajectories from each of the three different scenarios in Sim 1. Figure 7-5 illustrates that the Schaal model has bias in predicting for all groups (those with AF at 70, 75, 80 and 85) not just for individuals without an AF. Figure 7-6, **Error! Reference source not found.**, and **Error! Reference source not found.** each contain the cognitive trajectories from the three different scenarios in Sim 2. No obvious differences between these three models are immediately apparent from these graphs.

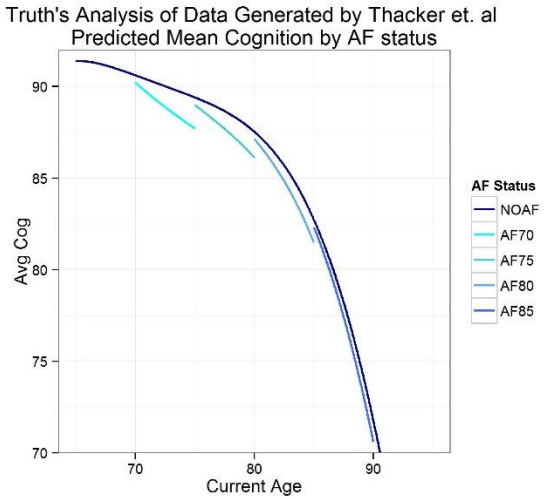

Figure 7-3 Simulation 1 predicted cognition using the true regression parameters.

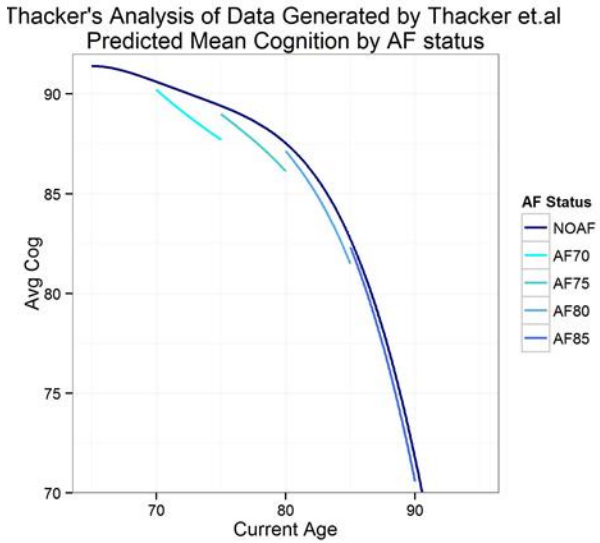

Figure 7-4 Simulation 1 predicted cognition using Thacker et al.'s regression parameters.

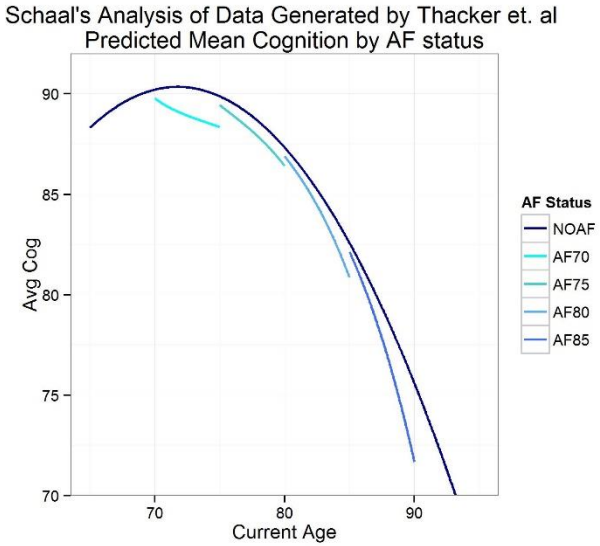

Figure 7-5 Simulation 1 predicted cognition using Schaal's regression parameters.

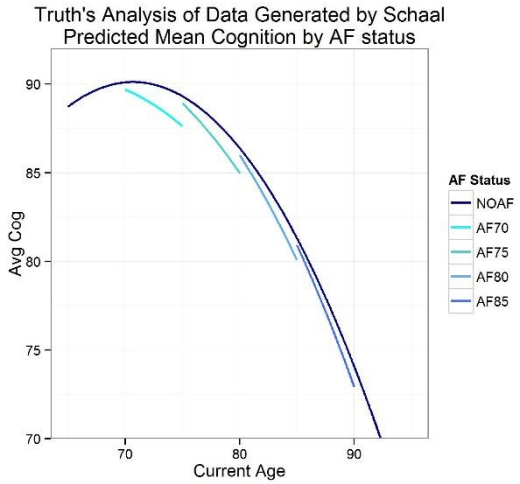

Figure 7-6 Simulation 2 predicted cognition using the true regression parameters.

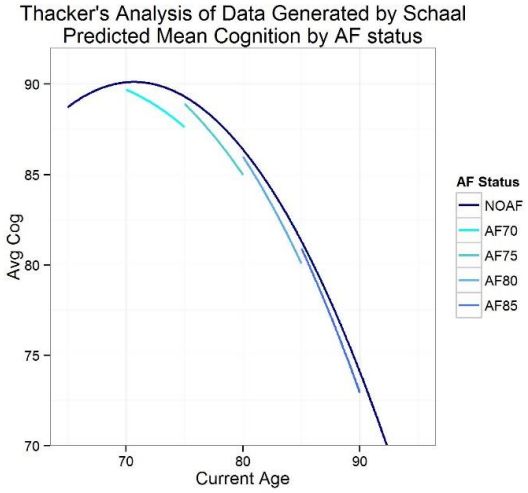

Figure 7-7 Simulation 2 predicted cognition using Thacker et al.'s regression parameters.

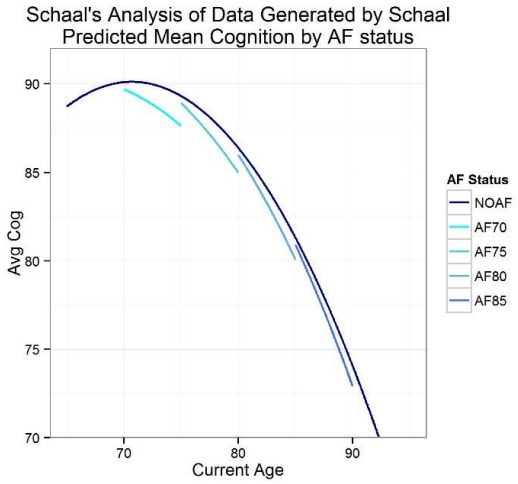

Figure 7-8 Simulation 2 predicted cognition using Schaal's regression parameters.

Previously, I introduced two difference contrasts that could be made to compare the trajectories of people who do not have an AF with people who do have an AF: effectOfAF and AFdrop. I first discuss the observations that can be made by examining numerically the effectOfAF contrast. Afterwards, I will discuss the AFdrop contrast. In Appendix E, the derivation of the effectOfAF contrast which compares people of the same age who have an AF to people without an AF (under the data generation model of Sim 1) demonstrates that the third, fourth, and fifth degree of centered age cancel out. The contrast effectOfAF depends only on the coefficients for PostAF, AFib, and the coefficients for the interaction Age(PostAF) and Age^2(PostAF). Since the contrast does not depend on any interaction terms that include higher order terms for Age, the true population value of the contrast effectOfAF should be exactly the same in Sim2 as in Sim 1. In Tables L and M we see that Sim1 and Sim2 had identical contrasts for the true model. Whether the data are generated using a complex model with fifth degree age or a simpler second degree polynomial, the true contrast in the data remains the same. Therefore, even though there are different data generation models, in both simulations we are attempting to measure the same underlying true contrast. For this contrast, differences between the data generation models were not as important as they were for the AFdrop contrast, but the differences in coefficient estimates could still influence the value of the estimated contrast.

Table L Simulation 1: The difference between average cognition for people without AF and people who have AF from three models of analysis; data were generated from the fifth degree polynomial model.

Data Generation:	Thacker											
effectOfAF	Truth's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Thacker's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Schaal's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev
DiffAtAge75	1.695	1.695	1.695	0.000	1.695	1.692	1.698	0.049	1.538	1.535	1.542	0.063
DiffAtAge80	1.413	1.413	1.413	0.000	1.413	1.410	1.417	0.051	0.923	0.919	0.927	0.060
DiffAtAge85	1.235	1.235	1.235	0.000	1.234	1.230	1.237	0.054	1.716	1.711	1.721	0.084
DiffAtAge90	1.158	1.158	1.158	0.000	1.156	1.150	1.163	0.101	3.917	3.906	3.929	0.186

Table M Simulation 2: The difference between average cognition for people without AF and people who have AF from three models of analysis; data were generated from the second degree polynomial model.

Data Generation:	Schaal											
effectOfAF	Truth's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Thacker's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Schaal's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev
DiffAtAge75	1.695	1.695	1.695	0.000	1.695	1.692	1.698	0.048	1.695	1.692	1.698	0.048
DiffAtAge80	1.413	1.413	1.413	0.000	1.412	1.409	1.415	0.049	1.412	1.409	1.415	0.048
DiffAtAge85	1.235	1.235	1.235	0.000	1.233	1.229	1.236	0.053	1.233	1.230	1.236	0.053
DiffAtAge90	1.158	1.158	1.158	0.000	1.155	1.150	1.161	0.094	1.157	1.151	1.162	0.087

The second and third columns in Tables L and M contain the estimates for the effectOfAf contrast that is generated under the two different methods of analysis. In Table M, when the data are generated using a simpler model, we see that both the Thacker and colleagues and Schaal analysis models provide unbiased estimates (out to two decimal places) at all four ages (75, 70, 85, 90). Both analysis methods also provide nearly equal precision at all four ages. It is noted that, in the oldest age bracket with the fewest data points, the Thacker and colleagues estimate has a slightly larger standard deviation than the Schaal model. The greater flexibility in the Thacker and colleagues model can give it a tendency to over-fit the data at the extremes.

In Table L, when the data are generated using a more complex relationship between age and cognition, the two methods of analysis are no longer providing similar stories. At all four ages, the Thacker and colleagues analysis provides unbiased estimates of the contrast. We see similar precision among the first three ages groups. In the oldest age group, with the fewest data points, we again see more variation in the estimate. In this age bracket, the standard deviation of the estimate is twice what it was at younger ages. Since the data were generated using Thacker and colleagues' model, the unbiased and precise results are to be expected.

The Schaal model applying a different analysis model returns biased estimates of the contrast at all four ages. For younger subjects (age 75 and 80) the estimates for the contrast in

cognition are more optimistic than they should be. The estimates are 9.2% and 34.7% smaller than the true contrast. For older subjects (age 85 and 90) the estimates of the contrast are more pessimistic than they should be. The estimates are 38.9% and 70.4% larger than they should be. However, the fact that the estimates are biased are not reason to conclude that the simpler Schaal model should be discarded.

When these contrasts are examined on an absolute scale, we see that the magnitude of the difference in the estimates is quite small. At age 75 we are comparing the truth of 1.695 to the biased estimate of 1.54 (95% CI 1.535, 1.542), a difference of only about 0.16 points in cognition. For 80 and 85 the absolute difference between the truth and the Schaal estimate is about 0.5 points. The truth at age 80 is 1.413 compared to the Schaal estimate 0.923 (95% CI 0.919, 0.927). At age 85 the truth is 1.235 with a Schaal estimate of 1.716 (95% CI 1.711, 1.721). Among the oldest individuals the bias becomes the largest, about 3 points (truth 1.158 Schaal: 3.917 (95% CI 3.906, 3.929)). If errors of 0.16 to 3 cognition points are acceptable, then using a simpler model to analyze the data may be justified. The benefit to the simpler model is that explaining it to a non-mathematically inclined individual would be easier.

For the other contrast, AFdrop, the scientific interest shifts slightly to focus on how cognition changes after an AF event rather than how having an AF affects cognitive decline. Even in this different scenario we see that the results from the two simulation studies give similar conclusions regarding the appropriateness of using a simpler analysis model. In Table O we see that when the data are generated using a simpler modeling of age, the estimates of the contrast from both the Thacker and colleagues analysis and the Schaal analysis are unbiased (out to 2 decimals) at all four ages. The precision of the estimates of the rate of decline for the three younger ages (70, 87, 80) is also quite similar for Thacker and Schaal. However, at the oldest

age bracket (85) we find that the standard deviation of Thacker's estimate is three times the estimate for Schaal. Again this can be thought of as an artifact of the more flexible model overfitting the extreme of the data. It also points to an advantage of using a simpler model to analyze the data, when in fact the data generation mechanism is simpler.

Table N Simulation 1: The change in cognition for people with an AF from the time when they have just experienced it to 5 years later; data were generated from the second degree polynomial model.

Data Generation:	Thacker											
AFdrop	Truth's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Thacker's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Schaal's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev
Drop: 70 to 75	2.504	2.504	2.504	0.000	2.504	2.500	2.509	0.078	1.432	1.426	1.438	0.095
Drop: 75 to 80	2.881	2.881	2.881	0.000	2.880	2.875	2.884	0.075	3.033	3.028	3.038	0.085
Drop: 80 to 85	5.645	5.645	5.645	0.000	5.648	5.642	5.654	0.093	6.042	6.036	6.048	0.094
Drop: 85 to 90	11.731	11.731	11.731	0.000	11.751	11.732	11.770	0.304	10.459	10.448	10.471	0.182

Table O Simulation 2: The change in cognition for people with an AF from the time when they have just experienced it to 5 years later; data were generated from the second degree polynomial model.

Data Generation:	Schaal											
AFdrop	Truth's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Thacker's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev	Schaal's Mean 1000 sims	Lower 95CI	Upper 95CI	Stand Dev
Drop: 70 to 75	2.074	2.074	2.074	0.000	2.076	2.071	2.080	0.075	2.075	2.070	2.079	0.074
Drop: 75 to 80	3.943	3.943	3.943	0.000	3.942	3.938	3.947	0.072	3.942	3.937	3.946	0.070
Drop: 80 to 85	5.914	5.914	5.914	0.000	5.912	5.906	5.917	0.091	5.912	5.908	5.917	0.074
Drop: 85 to 90	7.988	7.988	7.988	0.000	7.990	7.970	8.009	0.318	7.986	7.979	7.992	0.103

Table N displays the results for the decline in cognition when the data were generated using the more complex relationship between age and cognition. Thacker and colleagues' analysis is unbiased and has good precision for all four ages. This again is to be expected because the data generation and analysis models were the same. Schaal's analysis tends to underestimate the decline that each age group would experience. On an absolute scale the bias again appears small, only about 0.2 points to 1.3 points depending on the age. As with the simpler data

generation model, we see that in the older age bracket the variability of the estimate from the Schaal model (sd= 0.182) is much smaller than that from the Thacker et al. model (sd= 0.304).

In practice, we are unaware of the true data generating model and must make decisions on how to model the predictors in a regression model. In both scenarios when the model is mis-specified such that variables that describe the relationship are omitted the estimated contrast is biased. We also see that the simpler model (whether it was correct or misspecified) tended to have greater precision for the older age brackets. Therefore, when a researcher is planning a model that may have omitted variables, he or she must decide whether to accept a small amount of absolute bias while gaining greater precision and greater comprehension of the model with a wider audience.

Chapter 8. CONCLUSION

In Part 1, I provided the justification for Dr. Thacker and colleagues' choice of study design and analysis method. The use of a longitudinal study design ensured that the data could tell the story of how the average person's cognition changed over time. By following people forward in time, we would not confuse the "cohort effect" with the effect of aging on cognition. But using a longitudinal study design introduced correlated observations within the data collected from individuals. An analysis that failed to account for this correlation and merely treated the data as if it were all independent would have incorrectly estimated the variance of the regression parameter estimates. The result would be that the confidence intervals would be too narrow and could give false conclusions about the statistical significance of the results. From the two methods appropriate to analyze longitudinal data, GEE and mixed-effects regression models (LMM), only one would give unbiased results in the presence of data that is missing completely at random (MCAR). Since the parameter estimation in LMM is based on the maximum likelihood method, it would correctly identify the association that existed in the population even when subject observations were likely to be missing.

Having determined the appropriate method of analysis, I went on to build a multiple regression model. Dr. Thacker and colleagues were interested in determining "In the absence of clinical stroke, do people with atrial fibrillation (AF) experience faster cognitive decline than people without AF?". Before being able to determine the effect of an AF on cognitive decline an accurate association between current age and cognitive function needed to be modeled. Based on scientific reasons, the association between current age and cognitive function needed to be adjusted for several demographic covariates. Terms that allowed for expected effect modification were also included. Specific attention was given to modeling the nonlinear aspects

of the univariate relationships between a covariate and cognition. As all of these elements were considered and added to the model, the regression equation grew to contain a large number of terms (over thirty). I performed two simulation studies to compare performance of a simpler model, using a lower order polynomial representation of the age trajectory, with that of a more complex model, with a higher order polynomial.

From the simulation studies, I conclude that there is no one correct answer to the question that I propose. Depending on the unknown underlying truth, a simpler model may provide near identical results as the more complicated model. However, the simpler model with omitted terms may also provide biased results. Albeit, the bias tended to be quite small and possibly even at an acceptable level. The benefit of accepting the biased results was a decrease in the variance of the estimate of the contrast in the highest age brackets. A researcher considering using a simpler model with omitted variables would need to weigh the trade-off between the bias and the decrease variance of the estimate.

I close with a few limitations of these findings and possible extensions of the research. Most importantly, it is noted that the simulations did not investigate the effect of missing data due to drop-out on the relative performance of the simpler model. When conducting analysis on data collected in the real-world, it is extremely likely that drop-out will occur. Future research could simulate missing data and determine whether the bias that is observed remains at an acceptable level. It could also consider how the variation of the estimate is impacted. Another extension would be to generate the data under a model that is vastly different from either of the two proposed analysis models. Would both models have biased results with approximately the same magnitude? Would one model tend to over-estimate and the other tend to under-estimate the contrasts? A third alteration could be to generate the data assuming the random effects were

correlated rather than independent. The analysis methods used in this simulation study and the research of Dr. Thacker and colleagues allowed for the random effects to be correlated. These extensions would further address my research question about whether analyzing data using a model that is simpler than the data generation process vastly alters the conclusions that can be made from the research.

BIBLIOGRAPHY

Azzalini, A. (2015). 'sn' package [Documentation for R]. Retrieved November 10, 2015, from https://cranr-projectorg/web/packages/sn/snpdf

Baltes, P. B. (1968). Longitudinal and Cross-sectional Sequences in the Study of Age and Generational Effects. *Human Development, 11*, 145-171. Retrieved November 10, 2015, from http://library.mpib-berlin.mpg.de/ft/pb/PB_Longitudinal_1968.pdf

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H., Singmann, H., . . . Grothendieck, G. (2015). Package 'lme4' [Documentation for R]. Retrieved November 10, 2015, from https://cran.r-project.org/web/packages/lme4/lme4.pdf

Berk, R. (2004). *Regression Analysis: A Constructive Critique Advanced. Quantitative Techniques in the Social Sciences* (Vol. 11). Thousand Oaks, California: Sage Publications.

Bond, G. E., Burr, R. L., McCurry, S. M., Rice, M. M., Borenstein, A. R., & Larson, E. B. (2005). Alcohol and Cognitive Performance: A Longitudinal Study of Older Japanese Americans The Kame Project. *International Psychogeriatrics, 17*(4), 653-668. doi:http://dxdoiorg/101017/S1041610205001651

Cardiovascular Health Study data. (2015, November 10). Retrieved from https://chs-nhlbi.org/monograf/Monodocs

Chen, L. Y., Lopez, F. L., Gottesman, R. F., Huxley, R. R., Agarwa, S. K., Loehr, L., . . . Alonso, A. (2014). Atrial Fibrillation and Cognitive Decline: The Role of Subclinical Cerebral Infarcts. *Stroke, 45*, 2568-2574. doi:101161/STROKEAHA114005243

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Data Analysis* (2nd ed.). New York, New York: Springer.

Flinn, C. (2004, October 1). Dicussion of the Gauss-Markov Theorem: Introduction to Econometrics. Retrieved October 20, 2015, from http://www.nyu.edu/econ/user/flinnc/courses/0266/gauss-markov%20theorem.pdf

Fried, L. P., Borhani, N. O., Enright, P., Furberg, C. D., Gardin, J. M., Kronmal, R. A., . . . Weiler, P. (1991). The Cardiovascular Health Study: Design and rationale. *Annals of Epidemiology, 1*(3), 263-76. doi:10.1016/1047-2797(91)90005-W

Kilander, L., Nyman, H., Boberg, M., Hansson, L., & Lithell, H. (1998). Hypertension is Related to Cognitive Impairment: A 20-year Follow-up of 999 Men. *Hypertension, 31*, 780-786.

Krishnan, A. (2014, September 29). How would you explain the bias-variance tradeoff to a five year old? Retrieved November 10, 2015, from http://www.quora.com/How-would-you-explain-the-bias-variance-tradeoff-to-a-five-year-old.

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The Importance of the Normality Assumption in Large Public Health Data Sets. *Annual Review of Public Health, 23*, 151-169.

Oehlert, G. W. (2012). A Few Words about REML [PDF document]. Retrieved November 10, 2015, from http://usersstatumnedu/~gary/classes/5303/handouts/REMLpdf

Robert, C. P. (1995). Simulation of Truncated Normal Variables. *Statistics and Computing, 5*, 121-125.

Schaie, K. W. (2005). What Can We Learn from Longitudinal Studies of Adult Development? *Research in human development, 2*(3), 133-158. doi:10.1207/s15427617rhd0203_4

Sigman, K. (2007). Acceptance-Rejection Method [PDF document]. Retrieved November 10, 2015, from http://wwwcolumbiaedu/~ks20/4703-Sigman/4703-07-Notes-ARMpdf

Stefansdottir, H., Arnar, D. O., Aspelund, T., Sigurdsson, S., Jonsdottir, M. K., Hjaltason, H., . . . Gudnason, V. (2013). Atrial Fibrillation is Associated with Reduced Brain Volume and Cognitive Function Independent of Cerebral Infarcts. *Stroke, 44*(4), 1020-1025. doi:101161/STROKEAHA12679381

Stott, D., Falconer, A., Kerr, G. D., Murray, H. M., Trompet, S., Westendorp, R. G., . . . Ford, I. (2008). Does Low to Moderate Alcohol Intake Protect Against Cognitive Decline in Oder People? *Journal of the American Geriatrics Society*, *56*(12), 2217 - 2224. doi:1. 101111/j1532-5415200802007x

Trautmann, H., Steuer, D., Mersmann, O., & Bornkamp, B. (2015). Package 'truncnorm'. Retrieved November 10, 2015, from https://cranr-projectorg/web/packages/truncnorm/truncnormpdf

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measure Models.* Hoboken, New Jersey: John Wiley & Sons.

Waldstein, S. R. (2003). The Relation of Hypertension to Cognitive Function. *Current Directions in Psychological Science, 12*(1), 9 - 12. doi:1. 101111/1467-872101212

Weiss, N. S., & Koepsell, T. D. (2014). *Epidemiologic Methods: Studying the Occurrence of Illness* (2nd ed.). New York, New York: Oxford University Press.

Wright, C. B., Elkind, M. S., Luo, X., Paik, M. C., & & Sacco, R. L. (2006). Reported Alcohol Consumption and Cognitive Decline: The Northern Manhattan Study. *Neuroepidemiology, 27*, 201-207. doi:101159/000096300

APPENDIX A - PROOF OF THE GAUSS-MARKOV THEOREM

Discussion of the Gauss-Markov Theorem
Introduction to Econometrics (C. Flinn) October 1, 2004

We start with estimation of the linear (in the parameters) model $y = X\beta + \varepsilon$, where we assume that:

1. $E(\varepsilon|X) = 0$ for all X (mean independence)
2. $Var(\varepsilon|X) = E(\varepsilon\varepsilon'|X) = \sigma^2_\varepsilon I_N$ (homoskedasticity)

The Gauss-Markov Theorem states that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = (\boldsymbol{X'X})^{-1}(\boldsymbol{X'Y})$ is the Best Linear Unbiased Estimator (BLUE) if ε satisfies (1) and (2).

Proof: An estimator is "best" in a class if it has smaller variance than others estimators in the same class. We are restricting our search for estimators to the class of linear, unbiased ones. Since the data are the y (not the X), we are looking at estimators that are linear functions of y, or

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{m} + \boldsymbol{M}y$$

where β is a $k \times 1$ parameter vector, m is a $k \times 1$ vector of constants, M is a $k \times n$ matrix of constants, and the data vector y is $n \times 1$.

Second, we are restricting attention to the class of unbiased estimators, that is we require that $E(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, for any "valid" possible value β could take, i.e., for all β in the parameter space Ω_β.

First note that if $\widetilde{\boldsymbol{\beta}}$ is to be unbiased, then

$$\begin{aligned} E(\widetilde{\boldsymbol{\beta}}|X) &= \boldsymbol{m} + \boldsymbol{M}\,E(\boldsymbol{y}|\boldsymbol{X}) \\ &= \boldsymbol{m} + \boldsymbol{M}\,E(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}|\boldsymbol{X}) \\ &= \boldsymbol{m} + \boldsymbol{MX\beta} + M\,E(\boldsymbol{\varepsilon}|\boldsymbol{X}) \\ &= \boldsymbol{m} + \boldsymbol{MX\beta}, \end{aligned}$$

where the last line follows from the mean independence assumption. To be unbiased for any possible value of $\boldsymbol{\beta}$ then requires

$$\boldsymbol{m} = 0$$

and

$$\boldsymbol{MX} = \boldsymbol{I}_k. \qquad (1)$$

We note that the least squares estimator satisfies this requirement for unbiasedness, since for $\widehat{\boldsymbol{\beta}}$,

$$\boldsymbol{M} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}$$

and

$$\boldsymbol{MX} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'X} = \boldsymbol{I}_k$$

Thus looking for linear unbiased estimators requires us to look for estimators of the form

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{My}$$

for an M that satisfies [1].

Without any loss of generality, we can redefine the *M* matrix to be of the form

$$\boldsymbol{M} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}$$

where *C* is some *k*×*n* matrix. Using the condition [1] again, we know that for our estimator to be unbiased requires that

$$\begin{aligned}
&\boldsymbol{MX} = \boldsymbol{I}_k \\
&\Rightarrow [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}]\boldsymbol{X} = \boldsymbol{I}_k \\
&\Rightarrow \boldsymbol{I}_k + \boldsymbol{CX} = \boldsymbol{I}_k \\
&\Rightarrow \boldsymbol{CX} = \boldsymbol{0}
\end{aligned}$$

that is, *CX* is a *k* × *k* matrix of 0's.

Now we can compute the covariance matrix of all alternative estimator $\widetilde{\boldsymbol{B}}$. We can write

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \boldsymbol{My} \\
&= \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\beta} + \boldsymbol{M\varepsilon})
\end{aligned}$$

Thus $$\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \boldsymbol{M\varepsilon}$$

Since it is unbiased by construction, $E\left[\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}|\boldsymbol{X}\right] = 0$, so the covariance matrix of the estimator is

$$\begin{aligned}
E\left[\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'|\boldsymbol{X}\right] &= \boldsymbol{E}[\boldsymbol{M\varepsilon}(\boldsymbol{M\varepsilon})'|\boldsymbol{X}] \\
&= \boldsymbol{E}[\boldsymbol{M\varepsilon\varepsilon}'\boldsymbol{M}'|\boldsymbol{X}] \\
&= \boldsymbol{ME}[\boldsymbol{\varepsilon\varepsilon}'|\boldsymbol{X}]\boldsymbol{M}' \\
&= \boldsymbol{M}\sigma_{\varepsilon}^2\boldsymbol{I_n}\boldsymbol{M}'
\end{aligned}$$

which is a *k* × *k* matrix. Now

$$\boldsymbol{MM}' = [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}]'$$

$$\boldsymbol{MM}' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}' + \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{CC}'$$

Since $\boldsymbol{CX}$ = 0 (and of course $\boldsymbol{X}'\boldsymbol{C}'$= 0 as well),

$$\boldsymbol{MM}' = (\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{CC}'$$

Now the matrix $\boldsymbol{CC}'$ is a *k*×*k* "cross products" matrix, which by construction cannot be negative definite. The best estimator in a class of estimators is the one with the "smallest" covariance matrix, where by small we mean that the covariance matrix associated with any other estimator in the class (that is, linear and unbiased in the current context) minus the covariance matrix of the best estimator is a positive definite matrix. Formally, the matrix difference

$$\boldsymbol{MM}' - COV \text{ (best estimator)}$$

is positive definite. Since $\boldsymbol{MM}'$ is minimized when we set the matrix *C* equal to 0 (that is, it contains *k* × *n* 0*s*), the best estimator in the class $\hat{\beta}$. Any other estimator *M* in this class (in which the *C* matrix does not contain 0*s* in every row and column) has a strictly "larger" covariance matrix. We conclude that the OLS estimator $\hat{\beta}$ is BLUE under the two conditions set forth (mean independence and homoskedastic).

APPENDIX B – SKETCH OF THE DERIVATIONS FOR $\hat{\beta}_{GLS}$ AND $\hat{\beta}_{ML}$ THAT INCLUDE A COVARIANCE MATRIX.

By least squares (GLS)

1) Assume a linear model $Y = X\beta + \varepsilon$
2) Transform the equation by a matrix C, where C is chosen such that the inverse of the covariance matrix for all the data is a function of C: $\Omega^{-1} = C'C$.
$$CY = CX\beta + C\varepsilon$$
3) Define new variables $Y^* = CY, X^* = XC, \varepsilon^* = C\varepsilon$
$$Y^* = X^*\beta + \varepsilon^*$$
4) The new equation meets the four assumptions for OLS
   a. Linear predictor in the model and each covariate contributes non-redundant information about the outcome (no multicollinearity)
   b. Strict exogeneity (i.e. conditional errors have mean 0)
   c. Observations are independent
   d. No heteroscedasticity (ie conditional variance $Var(\varepsilon|X) = \sigma^2$
5) From OLS on the new equation we have
$$\hat{\beta} = [(X^*)'X^*]^{-1}((X^*)'Y^*)$$
$$\hat{\beta} = [(CX)'CX]^{-1}((CX)'CY)$$
$$\hat{\beta} = [X'C'CX]^{-1}(X'C'CY)$$
$$\hat{\beta}_{GLS} = [X'\Omega^{-1}X]^{-1}(X'\Omega^{-1}Y)$$

By maximum likelihood (ML)

1) Assume outcomes for one individual are MVN $Y_i \sim MVN(X_i\beta, \Omega_i)$
$$f(y_{i1}, y_{i2} \ldots, y_{in_i}) = f(y_i) = (2\pi)^{-n_i/2}|\Omega_i|^{-1/2} exp\left\{\frac{-1}{2}(y_i - X_i\beta)'\Omega_i{}^{-1}(y_i - X_i\beta)\right\}$$
2) For all subjects *i* = 1, 2, … N we get a joint likelihood of all observed outcomes by multiplying the individual joint likelihoods
$$f(\vec{y}) = f(y_1)f(y_2) \ldots f(y_N)$$
$$= \prod_{i=1}^{N}(2\pi)^{-n_i/2}|\Omega_i|^{-1/2} exp\left\{\frac{-1}{2}(y_i - X_i\beta)'\Omega_i{}^{-1}(y_i - X_i\beta)\right\}$$
$$= (2\pi)^{-\sum n_i/2}\prod_{i=1}^{N}|\Omega_i|^{-1/2}\, exp\left\{\frac{-1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)'\Omega_i{}^{-1}(y_i - X_i\beta)\right\}$$
3) Maximize joint likelihood by minimizing log likelihood
$$L = \frac{-1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)'\Omega_i{}^{-1}(y_i - X_i\beta)$$
$$\frac{\partial L}{\partial \beta} = \frac{-1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)'\left(\Omega_i^{-1} + \left(\Omega_i^{-1}\right)'\right)(-X_i)$$
Because we have a symmetric matrix for the covariance matrix $\left(\Omega_i^{-1}\right)' = \Omega_i^{-1}$
$$\frac{\partial L}{\partial \beta} = \frac{-1}{2}\sum_{i=1}^{N}(y_i - X_i\beta)'\left(2\Omega_i^{-1}\right)(-X_i) = \frac{\partial L}{\partial \beta} = -1\sum_{i=1}^{N}(y_i - X_i\beta)'\left(\Omega_i^{-1}\right)(-X_i)$$

4) Just drop the subscripts and minimize

$$0 = (Y - X\beta)'(\Omega^{-1})(-X)$$

5) Transpose each side

$$0 = (-X)'(\Omega^{-1})(Y - X\beta)$$
$$X'(\Omega^{-1})Y = (X)'\Omega^{-1}X\beta$$
$$\hat{\beta}_{ML} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

APPENDIX C – ANNOTATED PROOF OF THE ACCEPT-REJECT METHOD

Annotations to the Sigman 2007 proof that the accept-reject method used to generate the truncated normal variables will provide the correct distribution for the random variable. The accept-reject method relies on having two distributions for random variables. X is the random variable of interest and it has a known density function f(x); however generating random variables from this distribution is not easily performed using a known algorithm. A second random variable Y has density function g(y) and generating random variables from this distribution is straightforwardly done with a known algorithm. The two densities are related such that there exists a c close to 1 where $supremum\left\{\frac{f(x)}{g(y)}\right\} \leq c$.

Define three random variables: X has density function f(x), Y has density function g(y), and U follows the Uniform(0,1) density so h(u) = u where $0 \leq u \leq 1$. The cumulative distribution functions for each of the random variables are: F(x), G(y), U(u). Define sets of the random variables as: $A = \{Y \leq y\}$, $B = \left\{U \leq \frac{f(Y)}{c*g(Y)}\right\}$, and $X = \{Y\text{: for } (u,y), u \leq \frac{f(y)}{c*g(y)}\}$.

The accept-reject method consists of three steps:

Generate Y=y from the density g(y).

Independently generate U=u from the Uniform (0,1).

If $u \leq \frac{f(y)}{c*g(y)}$, then set X= y. Otherwise reject that value and begin step 1 again.

To prove that the accept-reject method produces variables from the correct density it is necessary to show that P(A|B) = F(x). The definition of joint probability will be manipulated multiple times to determine P(A|B). The joint probability for A and B equals the product of the

conditional probability and the marginal probability. There are two equivalent expressions to write the joint probability.

$$P(AB) = P(A|B)P(B) \quad \text{and} \quad P(AB) = P(B|A)P(A)$$

Equating the two right hand sides of these equations and solving for the desired probability, we know:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

There are three parts to the right hand side of this equation that we must determine in order to know P(A|B).

1) From the definition of a cumulative distribution function we determine P(A):

$P(A) = P(Y \leq y) = \int_{-\infty}^{y} g(w)dw = G(Y)$. Therefore P(A) = G(y).

2) Since set B contains two random variables, U and Y, finding P(B) is more challenging. The upper limit of the random variable U is not fixed, but varies as Y changes. For a simpler case, we will find the conditional probability of B at a fixed value of Y : P(B|Y=y). At this fixed value of Y, the upper limit for U is determined because c and Y are both known values; $u \leq \frac{f(y)}{c*g(y)}$.

Since U ~ Uniform (0,1) the probability density is known: h(u) = u, $0 \leq u \leq 1$. Thus the lower limit of integration is known to be 0.

$$P(B|Y = y) = P\left(U \leq \frac{f(y)}{c * g(y)}\right) = \int_{-\infty}^{\frac{f(y)}{c*g(y)}} h(u)\, du = \int_{0}^{\frac{f(y)}{c*g(y)}} u\, du = \frac{f(y)}{c * g(y)}$$

To determine the marginal probability, P(B), we evaluate the integral of the joint density over all possible values of Y but retain the conditional restriction for the values of U. Since the variables U and Y are independent of each other, the joint density is the product of the two marginal densities. In the final step, we recognize that by definition the area under the density function is 1 when evaluated across the entire range of possible values.

$$P(B) = \int_{-\infty}^{\infty} \int_{0}^{\frac{f(y)}{c*g(y)}} h(u)g(y)du\ dy = \int_{-\infty}^{\infty} g(y) \left[\int_{0}^{\frac{f(y)}{c*g(y)}} h(u)\ du \right] dy$$

$$= \int_{-\infty}^{\infty} g(y) \frac{f(y)}{c * g(y)}\ dy = \frac{1}{c} \int_{-\infty}^{\infty} f(y)\ dy = \frac{1}{c}$$

3) Lastly we determine the conditional probability when we are conditioning on Y as a continuous random variable P(B|A). Using the relationships established above we make substitutions:

$$conditional\ probability = \frac{joint\ probability}{marginal\ probability}$$

$$P\left(U \leq \frac{f(Y)}{c * g(Y)} | Y \leq y\right) = \frac{P\left(U \leq \frac{f(Y)}{c * g(Y)}, Y \leq y\right)}{P(Y \leq y)} = \frac{1}{P(Y \leq y)} P\left(U \leq \frac{f(Y)}{c * g(Y)}, Y \leq y\right)$$

Rewrite the joint probability as conditional on a value for Y, w, but restrict this realization of Y to be not more than the previously established value, y.

$$= \frac{1}{G(y)}\ P\left(U = u \leq \frac{f(w)}{c * g(w)} | Y = w \leq y\right) P(Y = w \leq y)$$

$$= \frac{1}{G(y)} \int_{-\infty}^{y} \int_{0}^{\frac{f(w)}{c*g(w)}} h(u)g(w)du\, dw$$
$$= \frac{1}{G(y)} \int_{-\infty}^{y} g(w) \left[\int_{0}^{\frac{f(w)}{c*g(w)}} h(u)du \right] dw$$
$$= \frac{1}{G(y)} \int_{-\infty}^{y} g(w) \left[\frac{f(w)}{c * g(w)} \right] dw$$
$$= \frac{1}{c\ G(y)} \int_{-\infty}^{y} f(w)\ dw$$
$$= \frac{1}{c\ G(y)} F(y)$$

Combining all the results from 1), 2), and 3) we determine the value for P(A|B).

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{\frac{1}{c\,G(y)}F(y)G(y)}{\frac{1}{c}} = F(y)$$

The accept-reject process by which we select the values of X, states that the randomly chosen value, y, for Y will be set as the value for the random variable X only for the pair (u,y) such that $u \leq \frac{f(y)}{c*g(y)}$. Remembering that F(y) is the cumulative distribution function for the random variable X (we have set X=y because the stated condition has been met) we have proved that the accept-reject method will generate random variables that have the correct distribution for X.

APPENDIX D – BACK-TRANSFORMATION OF REGRESSION COEFFICIENTS

Due to the difference in scale of some of the covariates it was necessary to standardize the continuous covariates that would be entered into the regression model. The scientific interest focused on differences between groups that occurred on the scale when covariates were not standardized. Below is a summary to explain the process of obtaining the estimates of effect on the appropriate scale.

The desired regression model is:

$$E(Cog|CAC\ CAC^2, BP, Cov) = \beta_0 + \beta_1(CAC) + \beta_2(CAC^2) + \beta_3(BP) + \beta_4(Cov) + \beta_5(CAC)(Cov) + \beta_5(CAC)(BP)$$

The regression model that used standardized covariates:

$$E[Cog|CAC, CAC^2, BP, Cov] = \alpha_0 + \alpha_1(ScCAC) + \alpha_2 ScCAC^2 + \alpha_3(ScBP) + \alpha_4(Cov) + \alpha_5(ScCAC)(Cov) + \alpha_6(ScCAC)(ScBP)$$

The transformation of the covariates:

$$ScCAC = \frac{CAC}{sd(CAC)}; ScCAC^2 = \frac{CAC^2}{sd(CAC^2)}; ScBP = \frac{BP - mean(BP)}{sd(BP)}$$

The standard deviations were calculated individually for each variable from the set of values when the data were generated. For CAC^2 this value was determined for each individual in the data set,then the standard deviation of the new variable CAC^2 was calculated; we did not simply square the standard deviation from CAC.

By substituting equivalent values into the analysis regression model we can obtain the back transformations necessary to get the estimates on the correct scale.

$$E[Cog|CAC, CAC^2, BP, Cov]$$
$$= \alpha_0 + \alpha_1\left(\frac{CAC}{sd(CAC)}\right) + \alpha_2\left(\frac{CAC^2}{sd(CAC^2)}\right) + \alpha_3\left(\frac{BP - mean(BP)}{sd(BP)}\right) + \alpha_4(Cov)$$
$$+ \alpha_5\left(\frac{CAC}{sd(CAC)}\right)(Cov) + \alpha_6\left(\frac{CAC}{sd(CAC)}\right)\left(\frac{BP - mean(BP)}{sd(BP)}\right)$$

$$E[Cog|CAC, CAC^2, BP, Cov]$$
$$= \left[\alpha_0 - \frac{\alpha_3\, mean(BP)}{sd(BP)} - \frac{\alpha_6\, mean(BP)}{sd(CAC)sd(BP)}CAC\right] + \left(\frac{\alpha_1}{sd(CAC)}\right)CAC$$
$$+ \left(\frac{\alpha_2}{sd(CAC^2)}\right)CAC^2 + \left(\frac{\alpha_3}{sd(BP)}\right)BP + \alpha_4(Cov) + \left(\frac{\alpha_5}{sd(CAC)}\right)(CAC)(Cov)$$
$$+ \left(\frac{\alpha_6}{sd(CAC)sd(BP)}\right)(CAC)(BP)$$

By comparing the above regression equation to the regression on the desired scale, we see the regression coefficients on the desired scale can be obtained through the following transformations:

$$\beta_0 = \alpha_0 - \frac{\alpha_3\, mean(BP)}{sd(BP)} - \frac{\alpha_6\, mean(BP)}{sd(CAC)sd(BP)}CAC$$

$$\beta_1 = \frac{\alpha_1}{sd(CAC)};\ \beta_2 = \frac{\alpha_2}{sd(CAC^2)};\beta_3 = \frac{\alpha_3}{sd(BP)}$$

$$\beta_4 = \alpha_4;\ \beta_5 = \frac{\alpha_5}{sd(CAC)};\ \beta_6 = \frac{\alpha_6}{sd(CAC)sd(BP)}$$

APPENDIX E – DERIVATION OF THE CONTRAST EFFECTAF (I.E. NOAF-AF AT A FIXED AGE)

To simplify the calculations I am not representing each baseline covariation separately.

The notation $\overrightarrow{Cov}$ represents the vector of 12 population average covariates: Birth Year through Heart Failure. As before CAC represents Current Age Centered.

Simulation 1:

$$E[Cog] = \beta_0 + \beta_{Age}(CAC) + \beta_{Age2}(CAC^2) + \beta_{Age3}(CAC^3) + \beta_{Age4}(CAC^4) + \beta_{Age5}(CAC^5) + \beta_{AFib}(AFib) + \beta_{PostAF}(PostAF) + \beta_{AFib}(AFib) + \vec{\beta}_{cov}(\overrightarrow{Cov}) + \beta_{AgePostAF}(CAC)(PostAF) + \beta_{Age2PostAF}(CAC^2)(PostAF) + \vec{\beta}_{CovAge}(\overrightarrow{Cov})(CAC)$$

We are finding the difference in mean cognition for two groups with the following characteristics:

Group 1: CAC = A, $CAC^2 = A^2$, $CAC^3 = A^3$, $CAC^4 = A^4$, $CAC^5 = A^5$, AFib = 0, PostAF= 0, $\overrightarrow{Cov} = \vec{C}$

Group 2: CAC = A, $CAC^2 = A^2$, $CAC^3 = A^3$, $CAC^4 = A^4$, $CAC^5 = A^5$, AFib = 1, PostAF= 5, $\overrightarrow{Cov} = \vec{C}$

For each of the *i* datasets that is generated the contrast for the above two groups is:

$$\Delta Cog = -5\beta_{i,PostAF} - \beta_{i,AFib} - 5A\beta_{i,AgePostAF} 5A^2\beta_{i,Age2PostAF}$$

From 1000 simulations we obtain

$$\frac{\sum \Delta Cog_i}{1000} = \frac{\sum -5\beta_{i,PostAF} - \beta_{i,AFib} - 5A\beta_{i,AgePostAF} - 5A^2\beta_{i,Age2PostAF}}{1000} = -5\overline{\beta_{PostAF}} - \overline{\beta_{AFib}} - 5A\bar{\beta}_{AgePostAF} - 5A^2\bar{\beta}_{Age2PostAF}$$

Simulation 2:

$$E[Cog] = \beta_0 + \beta_{Age}(CAC) + \beta_{Age2}(CAC^2) + \beta_{AFib}(AFib) + \beta_{PostAF}(PostAF) + \beta_{AFib}(AFib) + \vec{\beta}_{cov}(\overrightarrow{Cov}) + \beta_{AgePostAF}(CAC)(PostAF) + \beta_{Age2PostAF}(CAC^2)(PostAF) + \vec{\beta}_{CovAge}(\overrightarrow{Cov})(CAC)$$

We are finding the difference in mean cognition for two groups with the following characteristics:

Group 1: CAC = A, $CAC^2 = A^2$, AFib = 0, PostAF= 0, $\overrightarrow{Cov} = \vec{C}$

Group 2: CAC = A, $CAC^2 = A^2$, AFib = 1, PostAF= 5, $\overrightarrow{Cov} = \vec{C}$

For each of the *i* datasets that is generated the contrast for the above two groups is:

$$\Delta Cog = -5\beta_{i,PostAF} - \beta_{i,AFib} - 5A\beta_{i,AgePostAF} 5A^2\beta_{i,Age2PostAF}$$

From 1000 simulations we obtain

$$\frac{\sum \Delta Cog_i}{1000} = \frac{\sum -5\beta_{i,PostAF} - \beta_{i,AFib} - 5A\beta_{i,AgePostAF} - 5A^2\beta_{i,Age2PostAF}}{1000}$$
$$= -5\overline{\beta_{PostAF}} - \overline{\beta_{AFib}} - 5A\bar{\beta}_{AgePostAF} - 5A^2\bar{\beta}_{Age2PostAF}$$