Analyzing a Longitudinal Study of Cognitive Decline Using Mixed-Effect Models

by

Jeremy Faria

Senior Research Project
MA489
Supervised by Dr. Zilin Wang
Second Reader: Dr. Devan Becker

Department of Mathematics and Statistics
Wilfrid Laurier University
December 2025

Contents

1 Introduction

1.1 Cognition in the Elderly

Declining cognition is an unfortunate reality for many people as they age. It can lead to dementia, where patients with dementia struggle to lead independent lives and are a risk to their safety and others. There are many factors believed to contribute to dementia: this research aims to find which factors lead to the steepest decline in cognition. Some factors may be due to the lifestyle of people, such as smoking or lack of exercise, and other factors cannot be changed, such as ethnic background [3]. It is important for people to be aware that changing their lifestyle reduces their cognitive decline and to have their cognition tested sooner if they feel at risk, to lessen the effects of cognitive decline [2].

The Health and Retirement study provides data from a 30 year longitudinal study on US adults over 50. From the data, the factors contributing to the steepest decline in cognition as the respondents age can be investigated. This research uses weighted linear mixed-effects models to measure cognitive decline, where fixed effects coefficients are used to derive population level inferences. Previous cognition studies from the Health and Retirement study typically perform a survival analysis. One study on the HRS data using mixed-effect models measures the association of the big 5 personality traits with cognition, particularly that neuroticism has a high association with a lower cognition [7]. A potential flaw in their study is that neuroticism is highly associated with depression and depression is often identified as a factor contributing to cognitive decline in studies.

Section 1 will discuss general methodologies for longitudinal data and survey sampling. Section 2 will expand further on these methodologies and how they are applied to the HRS data. Section 3 will discuss the data source and how the weighting was handled. Section 4 will investigate whether a mixed-effect model is necessary for the data and explore potential covariates. Section 5 will discuss the next steps to take in January through April.

1.2 Longitudinal Data

Longitudinal data occur when repeated measures are collected on the same subjects over time, resulting in multiple observations per subject at various times.

The benefit of a longitudinal analysis over a cross-sectional analysis is the addition of the cohort and aging effects. Analyzing the cohort effect requires observing the differences in individuals from different cohort groups (generation, for example), whereas the aging effect requires observing the changes in the response variable over time within individuals [9]. Both effects cannot be analyzed in cross-sectional studies. Longitudinal studies can also identify causality as associations of what causes a certain change in units. The causal inferences found in uncontrolled studies will not hold as much weight as the inferences found in controlled studies, but the associations can still lead to meaningful insights.

Longitudinal studies can be carried out prospectively, where a subject is followed in time, or retrospectively, where data is collected on subjects from historical records. The quality of data collected in prospective studies is generally considered superior to that in retrospective studies, particularly when the quality of historical records is low or uncertain. Prospective longitudinal studies also have disadvantages, one of them being the costs of conducting the study. There is a lot of planning and logistics to keep track of in a longitudinal study because individuals are sampled repeatedly. Generally, the cost of a longitudinal study that goes through n waves is more expensive than n individual cross sectional studies, if the sample sizes are the same. An additional disadvantage in data collection in prospective longitudinal studies is panel conditioning. This occurs when the response from an individual on a later wave would differ from if it was their first wave. This can include a respondent learning new information through questions in a previous wave or a respondent feeling more comfortable being surveyed and revealing more sensitive information than they would during their first interview. Sample attrition is another problem in prospective longitudinal studies in which the respondent is not responsive [4].

Another factor that differentiates longitudinal studies from cross-sectional studies is that repeated measures on subjects result in correlation between observations for the same subject. Models used in cross-sectional studies such as the ordinary least squares regression require the assumption that all observations are independent, which is violated in longitudinal data because of the said within-person correlation. However, the observations between two different subjects are independent, resulting in a correlation of zero. The covariance matrix in longitudinal studies is block diagonal, elements where the row and column are of the same subject are non-zero, and the rest are zero. If all observations are kept in the vector $\vec{y} = [y_{11}, y_{1,2}, \ldots, y_{1,n}, y_{2,1}, \ldots, y_{m,n}]^T$ where y_{ij} is observation j for subject i, then the covariance matrix for $\vec{y}$ is written as:

$$var(\vec{y}) = \begin{bmatrix} B_1 & 0_{nxn} & 0_{nxn} & \cdots & 0_{nxn} \\ 0_{nxn} & B_2 & 0_{nxn} & \cdots & 0_{nxn} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0_{nxn} & 0_{nxn} & \cdots & \cdots & B_m \end{bmatrix}_{(m*n)x(m*n)}$$

$$\text{where } B_i = \begin{bmatrix} \sigma_{i1}^2 & \rho_{12}\sigma_{i1}\sigma_{i2} & \cdots & \rho_{1n}\sigma_{i1}\sigma_{in} \\ \rho_{12}\sigma_{i2}\sigma_{i1} & \sigma_{i2}^2 & \cdots & \rho_{2n}\sigma_{i2}\sigma_{in} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1n}\sigma_{in}\sigma_{i1} & \rho_{2n}\sigma_{in}\sigma_{i2} & \cdots & \sigma_{in}^2 \end{bmatrix}_{nxn} \quad \text{for subject i}$$

When modeling longitudinal data, the main focus is how the correlation between observations is handled. When correlation is ignored in modeling longitudinal data, the coefficient estimates are incorrect. There are three main approaches to handling such correlations: marginal models, mixed-effect models, and transition models [9]. Similarly to cross-sectional studies, marginal models focus on modeling the marginal mean. In marginal models the correlation is modeled explicitly, where typically it assumed that correlation with-in all units have the same structure, such as AR(1). Mixed-effect models model correlation through the use of individual random effects, such as random intercepts and potentially random slopes, and inferences can still be made on a population level through the use of fixed effects. The nuances of mixed-effect models will be discussed in section 2.1. The final approach is the transition model, which models the correlation structure through a conditional expectation of the outcome based on previous outcomes. It is important to note that all 3 approaches still use covariates in modeling.

1.3 Survey Sampling

Due to cost restrictions and feasibility, many surveys can only survey a sample of the target population. Through survey sampling methodologies, it can be ensured that the results of surveys are unbiased so that the insights extracted from them are meaningful. When a survey is performed well, the results of the sample may be considered to be more accurate than those of the entire population because a measurement error occurring is less likely.

When analyzing the population for a survey, 3 different subsets of the population can be considered: the target population, the frame population, and the sampled population. The target population is the set of all units that are part of the population of interest for a study. The frame population is a subset of the target population that is all units in the target population available to sample from. When the target and frame populations are not perfectly aligned, there are units that are not covered in the sampling framed, therefore, it is an incomplete sampling frame. The Sampled population is a subset of the frame population, which are units that have a non-zero probability of being selected. When the sampled population and the frame population are not perfectly aligned, there is a set of units that would not like to participate in the survey. One common mitigation to non-response is through reward incentives for completing a survey, helping to align the frame population and sampled population [8] .

Stratification is a technique for a more efficient and less biased sampling design. It involves dividing the population into non-overlapping strata, where each unit is in exactly one stratum. Typically, strata are divided using a demographic, such as region or age group, to ensure that all units fall into a single stratum. Given population U, the said population can be divided into H strata as follows:

$$U = U_1 \cup U_2 \cup \ldots \cup U_H$$

Clustering is the grouping of units that may be associated with each other. Similarly to stratification, units are only in one cluster. Sometimes, clustering can be done by dividing each strata into multiple clusters. Stratification reduces sampling variance, whereas clustering increases sampling variance [8].

In most modern surveys, a sample is designed by means of a probability measure. The probability measure combined with randomization eliminates potential biases that can potentially come from a non-probabilistic sampling design. Ω can be defined as the set of all possible subsets of the survey population U such that:

$$\Omega = \{S | S \subset U\}$$

$\mathscr{P}$ can be defined as a probability measure over Ω such that:

$$\mathscr{P}(S) \geq 0 \text{ for all } S \in \Omega$$

$$\text{and} \sum_{S:S\in\Omega} \mathscr{P}(S) = 1$$

Survey weights are an effective methodology that ensures that inferences from a model are relevant to the target population. For a unit, survey weights are determined by how their demographic is represented in the target population. Units in demographics that are undersampled from the target population will have higher weights, meaning observations from the said units will have more of a higher influence when using survey weights. Units in demographics that are oversampled from the target population will have lower weights, lessening their influence when using survey weights. The process of weighting units can vary from survey to survey, in section 3.3 the weighting of units in the Health and Retirement study will be elaborated on,

2 Methodology

2.1 Mixed-Effects Model

2.1.1 Linear Mixed-Effects Model

As mentioned in section 1.2 the mixed effects model contains two fundamental parts: fixed effects and random effects. The fixed effects coefficients are the same for the entire population, capturing population trends. The random effects coefficients are different for each individual, allowing for variability. For observation j of individual i the linear mixed-effects model can be written as:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + \gamma_{0i} + \gamma_1 z_{1ij} + \ldots + \gamma_k z_{qij} + \epsilon_{ij}$$

Or in matrix form:

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i$$

Where β is the vector of p fixed effects, X is the design matrix for fixed effects,

γ_i is the vector of q random effects, Z is the design matrix for random effects and ϵ_{ij} is the random error that follows the distribution:

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

The observation vector $\vec{y}$ comes from a multivariate normal distribution as follows:

$$\vec{y} \sim MVN(X\beta, \sigma^2 V)$$

Where $\sigma^2 V$ is the block diagonal matrix defined in section 1.2. The random effects are only the deviation from the population mean; therefore, the mean of the distribution is only the fixed effects.

Within the random effects of the linear mixed-effects model, there are two parts: random intercepts (γ_{0i}) and random slopes ($\gamma_{1i}, \ldots, \gamma_{qi}$). Random intercepts indicate how far an individual deviates from the sample average at the intercept. This allows different individuals to have a unique intercept, which is important in longitudinal settings with an uncontrolled environment. The random intercept typically follows a normal distribution centered at zero because the fixed random intercept is the average random intercept, so the random intercept will be the deviation away from the fixed intercept.

$$\gamma_{0i} \sim N(0, \sigma_b^2)$$

The resulting intercept for individual i is: $\beta_{0i} + \gamma_{0i}$. In addition, the random intercepts also account for correlation within units observation. The correlation function for two observations of individual i at different times j and k is:

$$Corr(y_{ij}, y_{ik}) = \frac{Cov(y_{ij}, y_{ik})}{\sqrt{Var(y_{ij)} Var(y_{ik})}}$$

$Cov(y_{ij}, y_{ik})$ can be expanded to:

$$Cov(y_{ij}, y_{ik}) = Cov[(X_{ij}\beta + \gamma_{0i} + \epsilon_{ij}), (X_{ik}\beta + \gamma_{0i} + \epsilon_{ik})]$$

Because the fixed effects are constant as they do not vary across samples, their covariance with anything is zero. This simplifies to covariance equation to:

$$Cov(y_{ij}, y_{ik}) = Cov[(\gamma_{0i} + \epsilon_{ij}), (\gamma_{0i} + \epsilon_{ik})]$$

The covariance term can be expanded to the following:

$$Cov(y_{ij}, y_{ik}) = Cov(\gamma_{0i}, \gamma_{0i}) + Cov(\gamma_{0i}, \epsilon_{ij}) + Cov(\epsilon_{ij}, \gamma_{0i}) + Cov(\epsilon_{ij}, \epsilon_{ij})$$

Because the random intercept is independent of random errors and the random errors at different times are independent of each other, the covariance term can be reduced to the following:

$$Cov(y_{ij}, y_{ik}) = Cov(\gamma_{0i}, \gamma_{0i}) + 0 + 0 + 0$$

If the response values are from 2 different observations, the covariance between them would be 0, because the random effects of two different individuals are independent.

Since the covariance of the same terms is equivalent to variance of the said term, the covariance term is:

$$Cov(y_{ij}, y_{ik}) = Var(\gamma_{0i}) = \sigma_b^2$$

The variance of y_{ij} is as follows

$$Var(y_{ij}) = Var(X_{ij}\beta + \gamma_{0i} + \epsilon_{ij})$$

As previously stated, $X_{ij}\beta$ do not vary between samples, simplifying the variance equation to the following:

$$Var(y_{ij}) = Var(\gamma_{0i} + \epsilon_{ij}) = Var(\gamma_{0i}) + Var(\epsilon_{ij}) + 2\ Cov(\gamma_{0i}, \epsilon_{ij})$$

As seen before, the random intercept and random errors are independent, resulting in the covariance term being equal to 0. so the variance is as follows:

$$Var(y_{ij}) = Var(\gamma_{0i}) + Var(\epsilon_{ij}) = \sigma_b^2 + \sigma^2)$$

Since the variances of two observations from the same individual are equal, the resulting correlation between said two observations is as follows:

$$Corr(y_{ij}, y_{ik}) = \frac{\sigma_b^2}{\sqrt{(\sigma_b^2 + \sigma^2)(\sigma_b^2 + \sigma^2)}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma^2}$$

The resulting correlation between two observations of the same individual is the proportion of total variance explained by the variance within a subject.

Some mixed-effect models only use random intercepts, but others can include random slopes. Random slopes are used for a covariate when its effects vary between individuals. The most commonly used random slope is for time, since time can cause the slopes of individuals to be at different rates. Other random slopes could be for time-varying covariates if it is expected that each persons effect of said covariate differs. Typically, time-invariate covariates, such as sex or ethnicity, do not require random slopes. Any covariate used for random effects is typically used for fixed effects as well. Since the fixed effect slope is the average slope for all observations, the distribution of the random slope is centered at 0, since it is the deviation from the fixed slope for that covariate. The distribution for the vector of random effects including the slope in individual i is as follows:

$$\gamma_i \sim MVN\left(\vec{0}_{qx1}, \begin{bmatrix} \sigma_{b0}^2 & \sigma_{b0}\sigma_{b1} & \cdots & \cdots \\ \sigma_{b0}\sigma_{b1} & \sigma_{b1}^2 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \sigma_{bq}^2 \end{bmatrix}\right)$$

2.1.2 Restricted Maximum Likelihood Estimation

The weighted least squares estimate for the vector of fixed effects coefficients β is the following:

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} \vec{y}$$

Given that V is the complete correlation structure of the data, it is hard to identify in practice, and usually an estimation is made with a loss of efficiency. In a longitudinal setting, there is a significant loss of efficiency when the number of measurements between units is varied. Units with more measurements have more information in total, but less information per measurement than units with fewer measurements, resulting in the fixed effect coefficient estimates having a higher variance and being less precise. Even when the efficiency is high, there are still issues with the variance of the fixed-effects coefficient estimates. Ordinary least squares assume the correlation structure is equal to the identity matrix and the variance of the fixed-effects coefficients is estimated with the residual mean square.

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$$\text{Where } \hat{\sigma}^2 = (nm - p)^{-1} (\vec{y} - X\hat{\beta})^T (\vec{y} - X\hat{\beta})$$

However, due to the positive correlation of observations from the same unit, $\hat{\sigma^2}$ is a biased estimator and results in a significant over or underestimation of the variance of $\hat{\beta}$. Therefore, the weighted least squares estimation cannot be used for mixed-effects models, so likelihood estimations should be considered instead.

Maximum likelihood estimation simultaneously estimates fixed effects coefficients (β), the correlation matrix of within unit observations (V), and the residual variance of observations (σ^2). The log-likelihood for the observed data $\vec{y}$ uses one block (V_0) of the correlation matrix and is as follows:

$$L(\beta, \sigma^2, V_0) = -0.5[nm\ ln(\sigma^2) + m\ ln(|V_0|) + \sigma^{-2} (\vec{y} - X\beta)^T V^{-1} (y - X\beta)]$$

β and σ^2 are estimated using the estimate of (V_0) where:

$$\hat{\beta}(V_0) = (X^T V^{-1} X)^{-1} X^T V^{-1} \vec{y}$$

$$\hat{\sigma^2}(V_0) = \frac{RSS(V_0)}{nm}$$

$$\text{Where } RSS(V_0) = (\vec{y} - X\hat{\beta}(V_0))^T V^{-1} (\vec{y} - X\hat{\beta}(V_0))$$

This gives the reduced log-likelihood for V_0 where:

$$L_r(V_0) = -0.5m(n\ ln(RSS(V_0)) + ln(|V_0|))$$

$\hat{V_0}$ is given by maximizing $L_r(V_0)$ and is then used to find $\hat{\beta}$ and $\hat{\sigma^2}$. In mixed-effect models, this severely underestimates σ^2 because it assumes β and V_0 is known and does not account for the loss of degrees of freedom because they are unknown. The unbiased estimator for σ^2 is $\frac{RSS}{nm-p}$ because it assumes that there is a loss of degrees of freedom. A consequence of this is that because the random effect variance is underestimated, the fixed effects absorb the variation that belongs to the random effects to optimize likelihood, resulting in the standard error getting too small and increasing the type 1 error rate.

The restricted maximum likelihood estimation (REML) transforms the data so that variance estimates do not contain fixed-effect information, resulting in the variance estimate being unbiased. The vector of transformed data Y^* does not depend on fixed effects β such that:

$$Y^* = AY \text{ where } A = (I - X(X^T X)^{-1} X^T) \text{ so that } AX = 0$$

$$\text{and } Y^* \sim MVN(0_{mxn}, \sigma^2) \text{ for all values of } \beta$$

Through this transformation, the data space is decomposed into 2 parts; the first part is explained by the fixed effects of dimension p, where $p = Rank(X)$, and the residual subspace of dimension $nm - p$ [6]. This ensures the variation is due to tje random effects and residuals by making the variance estimator unbiased as the following:

$$\hat{\sigma^2}(V_0) = \frac{RSS(V_0)}{nm - p}$$

2.1.3 Model Evaluation

There are two different approaches to comparing and evaluating mixed-effects models: statistical metrics and model visuals. Statistical metrics are very useful for model comparison due to quantifying the models performance into a value. Visuals can be useful for evaluating how a model fits the data by observing the residuals.

Because mixed-effects models use restricted maximum likelihood estimation, metrics such as AIC, BIC, and likelihood ratio tests cannot be used to compare models with different fixed effects because they use a different likelihood estimation technique than mixed-effect models. The restricted maximum likelihood estimation depends on design matrix X, and when different fixed effects are included, X changes, so models with different fixed effects cannot be changed. However, two models can be compared if they have the same fixed effects but different random effects. Therefore, the significance of including random random effects can be evaluated using metrics such as AIC, BIC, and likelihood ratio tests [5].

Comparing models with the same number of fixed effects can be used through marginal and conditional R^2. Marginal R^2 is the proportion of variance explained by the fixed effects, and conditional R^2 is the proportion of variance explained by the fixed effects and random effects. Models with different number of parameters cannot be compared using these metrics, as they will always increase when the number of parameters increases. There have been studies using an adjusted marginal and conditional R^2, however, there is no standard way to adjust the number of parameters when using restricted maximum likelihood estimation, so it will not be used in this study. To compare models with a different number of fixed effects, maximum likelihood estimation will need to be used. Therefore, metrics such as AIC, BIC and the likelihood ratio test can be used; however, they should be used with caution because maximum likelihood estimation underestimates the variance and can increase the type I error rate.

Visual metrics are an effective way of visualizing how the model fits the data. When plotting the residual versus fitted values plot for models, a good fit is when the residuals are scattered around the zero line with no visible pattern and do not fan out. If the points in the residual versus fitted values plot follow a trend, the model is likely not complex enough because of non-linearity or unexplained variance. If points fan out, it indicates heteroskedasticity and the data may need to be transformed or there may need to be a variance structure. heteroskedasticity can also be observed in mixed-effect models by using violin plots of residuals over time. If the variance of distributions changes as time increases, then there is heteroskedasticity. If the medians of the distributions change, then more random slopes or a fixed interaction effect will likely be needed. If the shape of the distributions changes over time, then the data will likely need to be transformed. In addition QQ-plots of residuals and random effects can be used to ensure they are normally distributed

2.2 Longitudinal Survey

2.2.1 Complex Survey Design

Compared to cross-sectional surveys, the survey design for longitudinal surveys is usually more complex and needs more considerations. If an in-person interview is used for a longitudinal survey, it can be very costly and logistically complicated to visit all units in a single wave. One mitigation is to use a complex survey design through multiple stages of stratification, clustering, and sampling. When the target population is very large and geographically dispersed, clustering and stratifying the population into regions and randomly sampling the said regions is an effective method of reducing costs. A drawback to this is that units from the same cluster could potentially be correlated with each other as they are more likely to have similar information.

In cross-sectional surveys, if subgroup analysis is of interest, then oversampling from the said subgroups is an effective approach if the selection probabilities are known. For time-invariate subgroups such as race, this method does not pose a problem; however, for time-variate subgroups such as location, this can be an ineffective approach, as people move in and out of the subgroup over time.

2.2.2 Weighting

When performing a longitudinal analysis on all waves, then the latest available weight for a respondent is used for them across all waves. This accounts for sample attrition and non-response bias that accumulates overtime, as well as population changes that occurred over the survey.

Cross-sectional surveys typically have some level of non-response from sampled units, occurring when sampled units decline to be surveyed. This can impose issues for the quality of the survey, but can usually be mitigated through survey weights. Non-response also occurs in longitudinal studies, but not just at the first wave, but in all subsequent waves as well. If a respondent drops out it can be for a multitude of reasons including, but not limited to: difficulty tracking a respondent, respondent fatigue from the interview, or the death of a respondent. Because non-response in later waves is never random, it is critical that as much information of why a respondent dropped-out is provided to make accurate inferences. Factors in survey design will likely play a role in drop-out, which should be considered. There is a tradeoff of the time between waves: shorter time periods between waves means that less respondents move, making it easier to track, but there is more likely to be respondent fatigue, and longer time periods between waves mean that it might be harder to track respondents at subsequent waves, but the respondents will be less likely to be fatigued. The time between waves should be dictated by this tradeoff and by the objectives of the study. Studies have found that survey attrition in longitudinal studies is more common in men, single people, and ethnic minorities [4].

3 Data Description

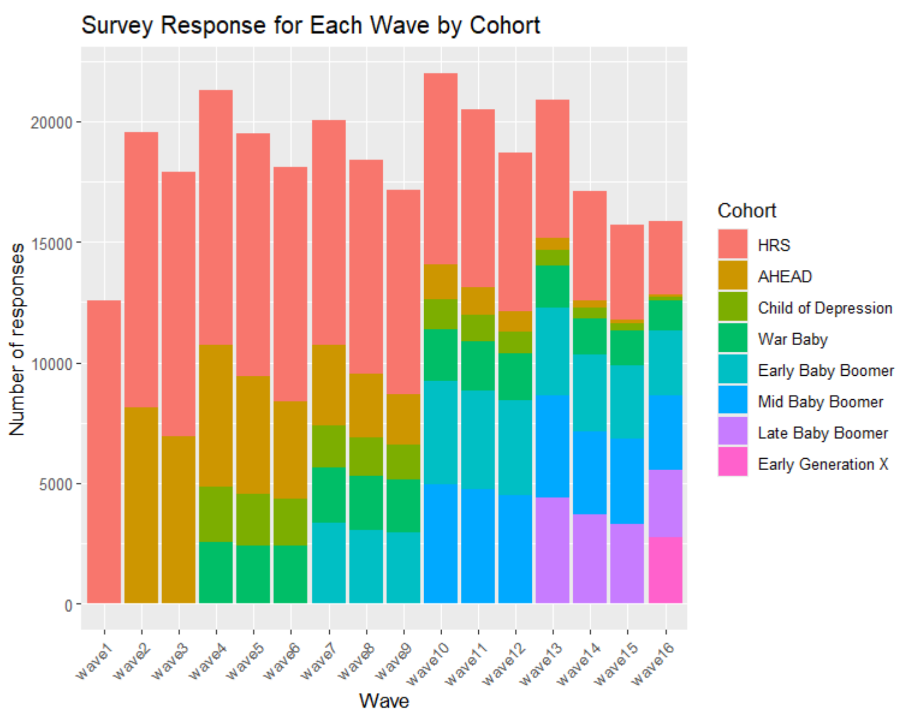
3.1 Health and Retirement Study

The Health and Retirement Study (HRS) is a national panel survey conducted from 1992-2022 of US citizens over 50 who do not live in institutions. The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan [1]. Respondents were surveyed every 2 years and new cohorts were added to the survey every 6 years. Below is a table identifying cohorts and when they were added to the survey.

Cohort Title	Birth Years	Year Added into Survey
HRS Cohort	1931-1941	1992
AHEAD Cohort	1924 or earlier	1993
Child of Depression	1924-1930	1998
War Baby	1942-1947	1998
Early Baby Boomer	1948-1953	2004
Mid Baby Boomer	1954-1959	2010
Late Baby Boomer	1960-1965	2016
Early Generation X	1966-1971	2022

Table 1: Cohorts of HRS Survey

The following plot is the frequency of each cohort in each wave:



From the plot it is evident that there is non-response throughout the survey as the frequency of people surveyed in a cohort decreases over time. It is recorded if the individuals that did not participate in a wave were alive or passed away, and if they did pass away, an exit interview was conducted from a family member of caregiver. The information provided on why a respondent dropped out will be critical to how inferences are made.

When the survey began in 1992, the interviews were conducted only in-person. From 2002, follow-up interviews were conducted by telephone, unless the respondents age is older than 80. Then starting in 2006, the respondents in follow up interviews were divided into 2 groups where every wave they alternated between face-to-face and telephone interviews. In addition, starting in 2018, respondents had the option to do an online survey instead. Of the eligible respondents who were able to complete the online survey, 40% of them continued to do it using the original methods to compare the results with the respondents who chose to complete the online survey. If a respondent cannot or is unwilling to do an interview themselves, they were given the option to use a proxy for the interview, which was typically a family member.

The sampling design was composed into 4 steps. The first step was to select the primary sampling units (PSU), which are counties in the United States. The counties were divided into metropolitan statistical areas (MSA) and non-MSA counties and then sampled separately from a proportionate probability to the population of those counties. In the second step smaller areas in the PSU's like cities or towns were sampled. Clusters of homes are selected in the third step and age-eligible individuals were selected for interviews in the fourth step.

The HRS surveys includes many different questions falling under one of these topics: demographics, health, functional limitations and helpers, financial and housing wealth, income, social security, pensions, health insurance, family structure, retirement plans, employment history and pyscho-social questions.

3.2 Cognition on the 27 Point Scale

HRS uses several different cognitive tests and questions in their surveys. Due to HRS being a multi-purpose study and using multiple interviewers due to its large scale, they needed to come up with standardized methods to obtain valuable information on cognition in a time efficient manner. At the time the survey began, there was a lack of literature and studies on cognition on a scale as large as HRS, so the tests needed to be validated through a team of clinicians.

The one used in many literature using the HRS data is the 27 point scale. The score a respondent received is made up of 3 tests, where the maximum score received is 27 points and the minimum is 0. The first test, which is worth 20 points, is immediate and delayed word recall, where the interviewer would read out 10 nouns and the respondent would have to name as many as they can immediately and after a set amount of time. Because there are multiple interviewers and the testing is carried out in the home of the respondents, impacts such as the interviewer reading speed and noise in the testing environment can affect the respondents score. The second test, which is worth 5 points, is counting down from 100 by 7's for 5 iterations. Respondents receive one less point for every error they make, but a mistake made at an earlier iteration will not affect later iterations. For example, if a respondent says 94 on the first iteration, they would lose a point, and then they say 87 on the second iteration, they will not lose a point despite not being on the correct track because 94-7 is 87. This introduces another issue in the methodology because it is the same test every wave, respondents can practice ahead of time, and therefore may not be an accurate assessment on cognition. The third test, which is worth 2 points is counting backwards by 1's from 20 and getting deducted a point for every mistake, where up to 2 points can be deducted. This test is purposely designed to make it easier than the other tests to identify respondents who are cognitively impaired.

This 27-point cognitive scale is based on the Langa-Weir classification of cognition defined by researchers at the University of Michigan. Scores between 12 and 27 points are considered normal cognition, 7 to 11 points are considered cognitive impairment without dementia, and 0 to 6 points are considered dementia. These thresholds are not clinical diagnoses, but have been validated to meet the criteria of dementia and cognitive impairment in the DSM-III-R and reviewed by a group of clinicians [7].

Because the 27-point cognitive scale requires a telephone or in-person interview with a respondent, the tests were not run for web-based and proxy interviews. In addition, the 27-point cognition scale was implemented the third wave, however, there is other cognition data available prior to the third wave.

3.3 Survey Weights

The survey weights in the HRS are composed of 3 parts: the base weight, non-response adjustment, and post-stratifcation. The base weight is from the sample design, accounting for the inverse of the probability of selection. The

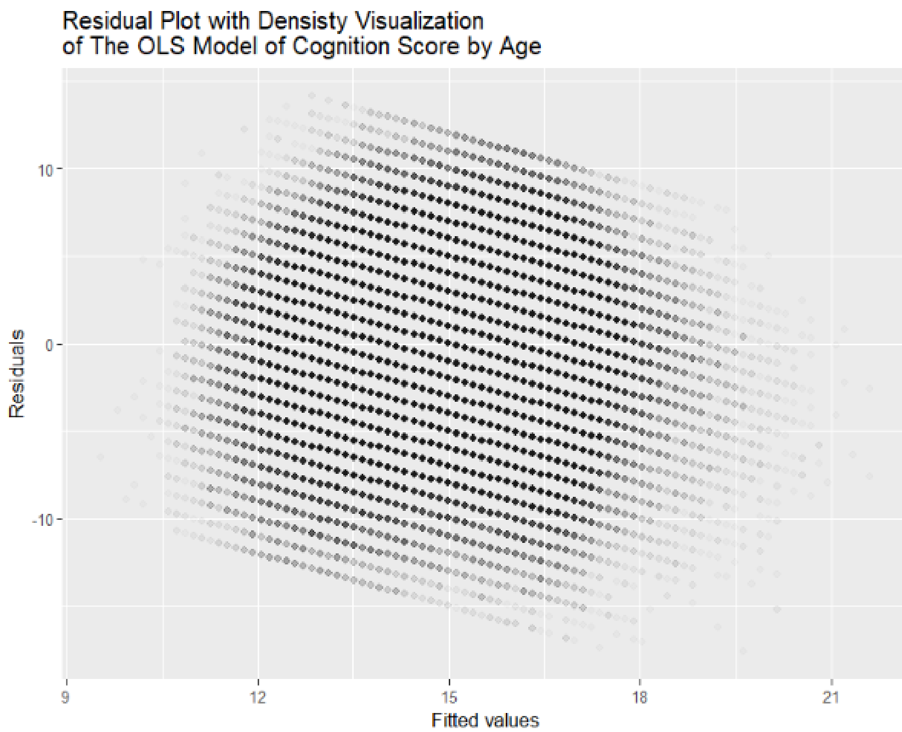
HRS survey conducted oversampling on black, Hispanic, and Floridian residents: therefore, their survey weights will be lower as each individual represents less of the population. Because of survey attrition the HRS accounts for the non-response at every wave. This weighting accounts for attrition from death and attrition from voluntarily dropping out of the survey. Post-stratification was performed when the sample was projected to the Current Population Survey and weights in the HRS were corrected for the CPS data.

4 Exploratory Data Analysis

As stated in section 3.1, in the HRS dataset new cohorts are introduced at various waves in the survey. Because the new cohorts are often younger than the previous cohorts, their cognition at time points is likely to be higher than that of older cohorts at the same time points, even if the cognition score of a member of the new cohort is worse than the cognition score of a member of the old cohort at that age. Therefore, the time point can not be a fixed effect or random effect in the model, unless the analyzes is only run on one cohort; however, in doing that we lose lots of valuable information from the previous cohorts. In addition, the age range within a cohort could also distort cognition scores within a cohort. To mitigate these effects, age can be included as a fixed effect and potentially a random effect instead of time, since it puts all individuals in the sample on the same level. Because the minimum age for the survey is 50, the x axis of the model will be age minus 50, resulting in the intercept being at age 50.

4.1 Justification of a Mixed-Effects Model

One of the main purposes for using a mixed-effects model is to effectively model correlated data. If the data are not correlated, then using a regression model estimated through ordinary least squares would be a valid approach to modeling. An indicator that a linear regression model is not suitable for analyzes is analyzing the residual plot of a regression model. If the residuals of a regression model are randomly scattered with no discernible pattern and do not fan out, then the model captures the full pattern of the data. A linear regression model of the cognition score of individuals based on age has the following residual plot:



The residuals are not randomly distributed in the linear regression model, indicating that it does not capture the full pattern of the data. The residuals tend to decrease linearly as the fitted value increases. The likely cause of this is that there is a linear residual pattern for each individual because the residuals are correlated with each other. Because ordinary least squares assumes observations are independent, the within-person correlation is ignored, causing the mean residual to be linear. Individuals consistently being above or below the mean residual at the same level indicates the need for random intercepts, since that will account for individuals starting at different baselines. Random intercepts will also account for the correlation within individuals, as discussed in section 3.1.1.

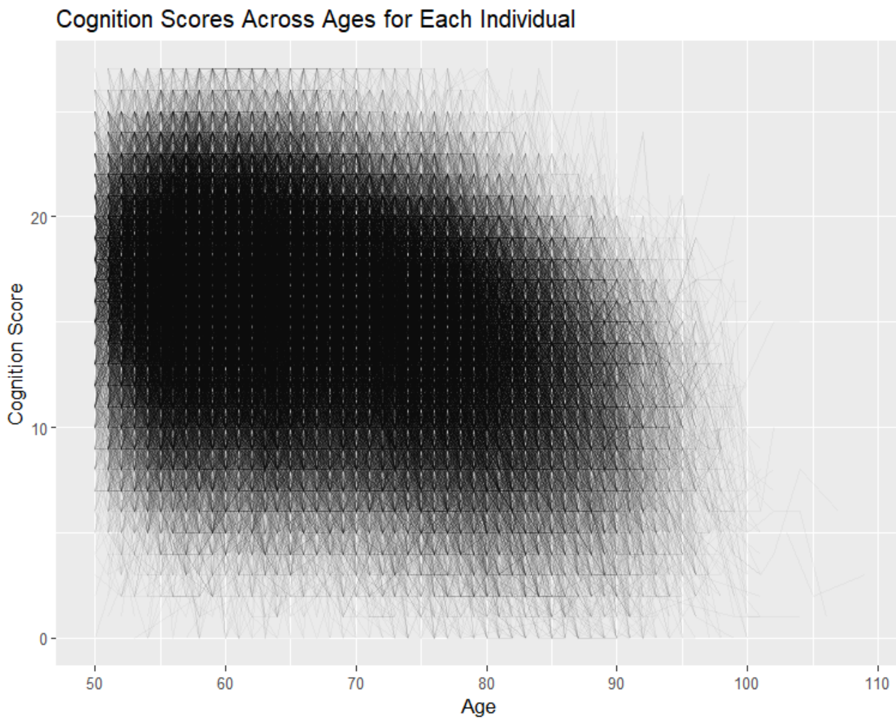
To explore the correlation structure of the data, the correlation can be explicitly modeled. To model the correlation structure, we can find the empirical correlation for a lag, which is the age difference. To find the empirical correlation at a given lag, all pairs of cognition scores are found where the age difference is the lag, and the correlation is found between all pairs of the said lag. Pairs of cognition scores are only used in calculating the correlation if they are from the same individual. Given a vector of scores at age t as x and vector y, which are scores at age t+u, where x_i and y_i correspond to the same individual, the empirical correlation at lag u is calculated as follows:

$$corr(u) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

The correlation between observations in individuals is expected to decrease as the lag increases. To interpret the correlation structure in a correlation plot, a locally estimated scatterplot smoothing (LOESS) line can be fitted to the points in the said correlation plot. LOESS can be used to derive a smooth pattern from data points by fitting a local quadratic regression line for each data point using their neighbours, and then the local quadratic regression lines are put together into a smooth curve. Due to survey attrition and cohorts being added throughout the survey, there are more pairs at smaller lags than at larger lags. To adjust for this discrepancy, a weight based on the number of pairs at a lag can be used to have the LOESS line be closer to points at smaller lags than to larger lags, because there is more confidence in the precision of the correlation at the smaller lags. The correlation plot for age lag using weighted LOESS smoothing is as follows:

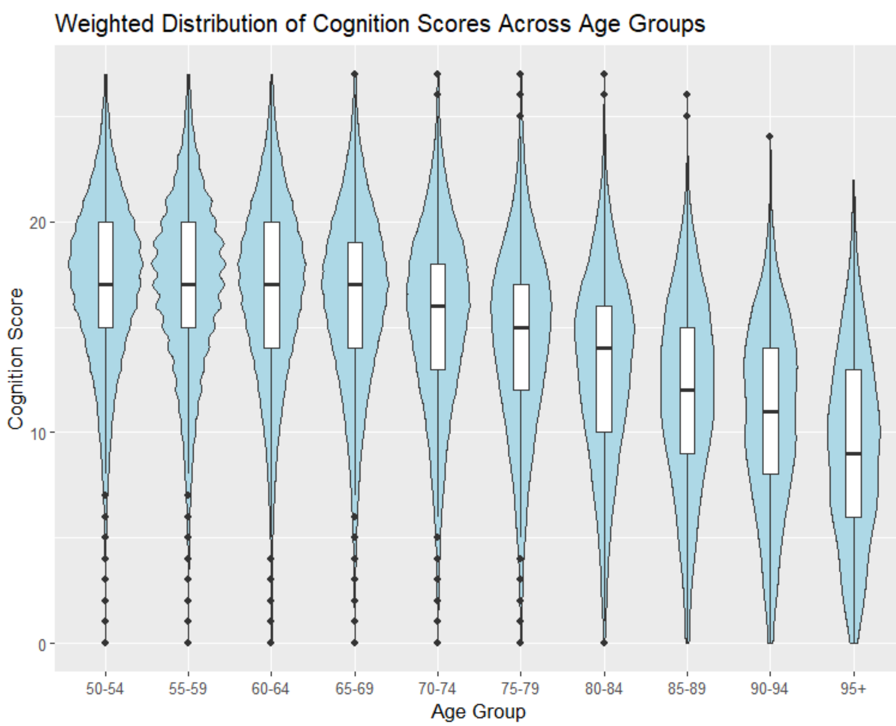
In the correlation plot, there appears to be a slow decay in correlation as the lag increases, indicating that there is strong within-person stability. The correlations at lags 2 through 4 are very high at around 0.7. Individuals cognition scores do not change dramatically on a year-to-year basis, and the measurement is reliable. The correlation is decreasing at a moderate rating, indicating that individuals cognition scores diverge more as aging progresses and the potential need for random slopes in mixed-effects models. Because the confidence ribbon of LOESS smoothing line gets significantly wider at around lag 20 large age gaps are not as common and correlation estimates at this lag are more sensitive to noise and less precise. The correlation pattern indicates an auto-regressive(1) correlation structure which can be modeled explicitly in mixed-effects models.

In longitudinal studies, a common plot that is used in exploratory data analysis is the spaghetti plot. A spaghetti plot is where the x axis is time, the y axis is the observation, and there is a line for every individual. The distribution at the baseline in this plot can indicate if random intercepts will be required and the requirement of random slopes can be investigated if the rate of cognitive decline is different between individuals.



From the spaghetti plot, it appears that random intercepts are needed for the model, as individuals are different baselines. The opacity in the spaghetti plot is turned down, so darker regions have more individuals at that point than lighter regions. It appears at the intercept (50 years old) that the individuals are normally distributed where the points are very dark at a cognition score of 16 and then the further away a score is from 16, the lighter the points are. This would result in a fixed intercept being at 16 and the random intercept is normally distributed at 0 with variance σ_b^2. The spread of lines changes as the ages increase. First, the spread of line gets narrower from age 50 to about age 75, meaning that most people with a high cognition score at 50 decline at a faster rate than people with a low cognition score at 50. From age 75 and on, the spread of lines is wider, meaning that individuals are declining at different rates at this age. This could mean that we could need a random effect for age, because time could have different effects on different people because the rate of decline is different for different individuals.

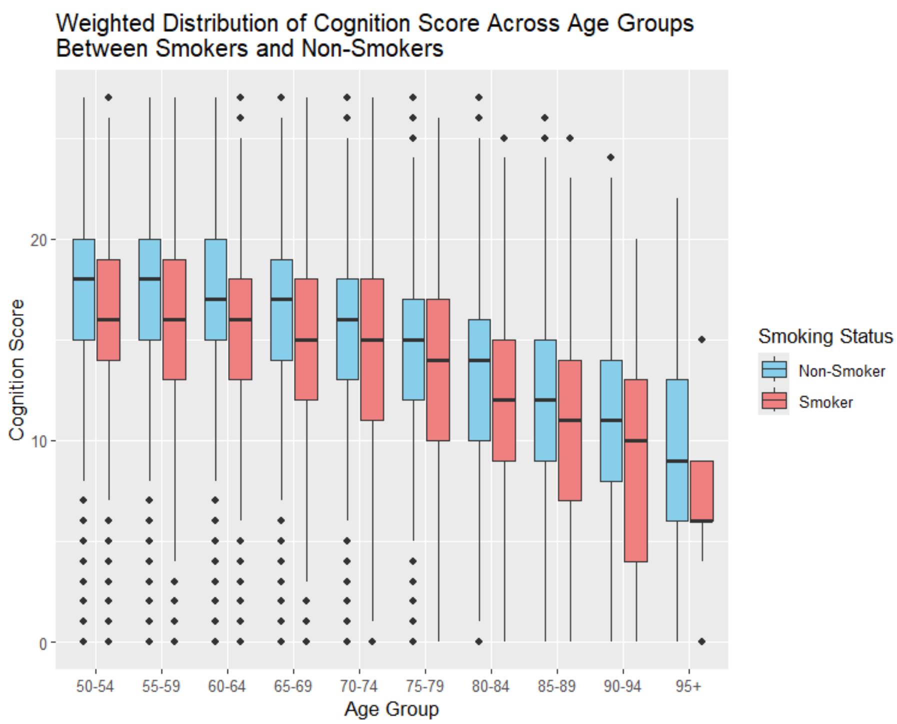
Random effects can also be investigated through distribution plots, such as box plots and violin plots. Box plots identify the quartiles and the range of a continuous variable, and violin plots visualize the distribution across a continuous variable. If the distribution across ages changes, it means that random slopes might be needed because cognition is changing at different rates for individuals. Due to ages in the HRS survey ranging from 50 to 109, ages were grouped into bins for easier analysis of the distribution plots. The bins only covered an age range of 5 years, with the exception of all observations of individuals being 95 in one bin, so that important patterns are not missed out. Survey weights are also applied as they are used to ensure the insights gained are relevant to the mixed effects models that will be used, as they will be using survey weights.



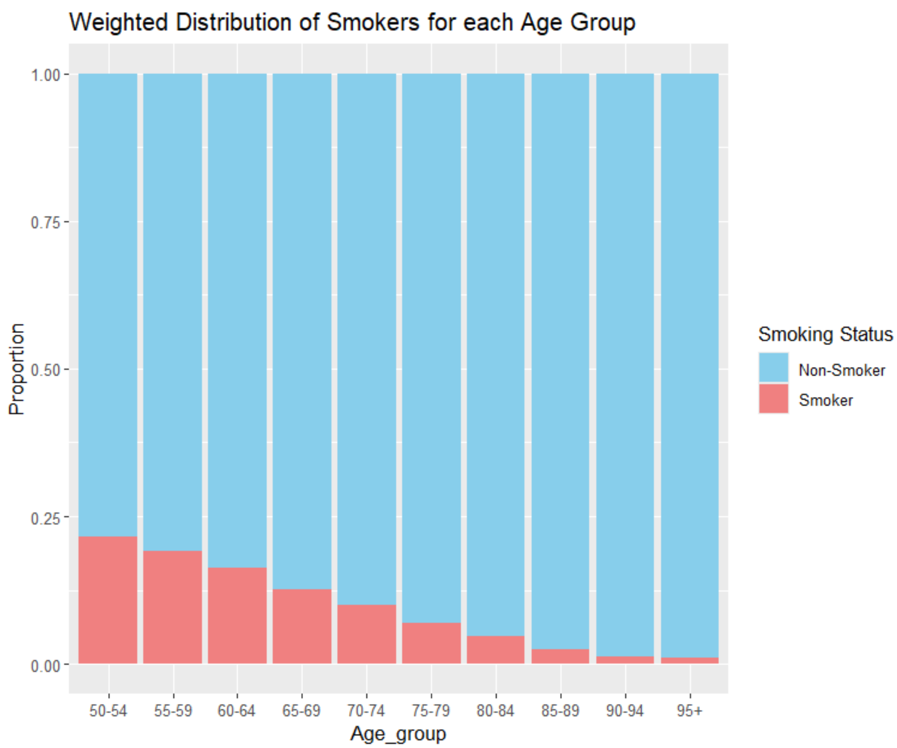
From the violin and box plots, the cognition scores are distributed differently in different age groups. In the three youngest age groups the scores are less spread out than in the older age groups. This means that random slopes are possibly needed because changing distributions indicate that individuals experience cognitive decline at different rates.

4.2 Covariates

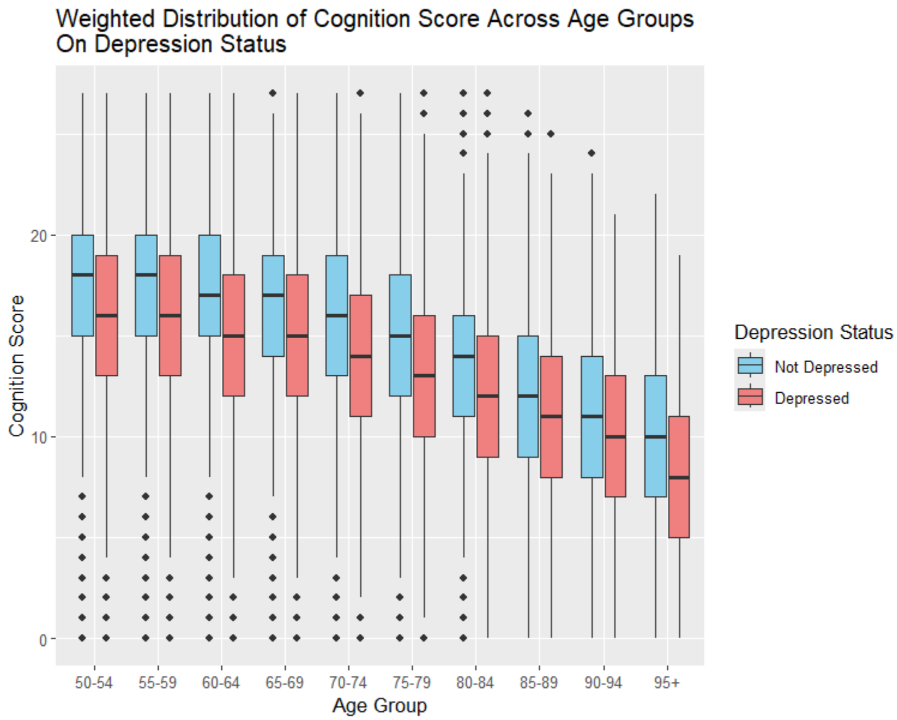
Studies have shown that there are non-modifiable and modifiable factors that contribute to cognition, such as race, physical inactivity, smoking, high systolic blood pressure, and depression [3]. The relationship between these factors and cognition in the HRS data will be explored. Survey weights will be used in the exploratory analysis as they will be used in the model.



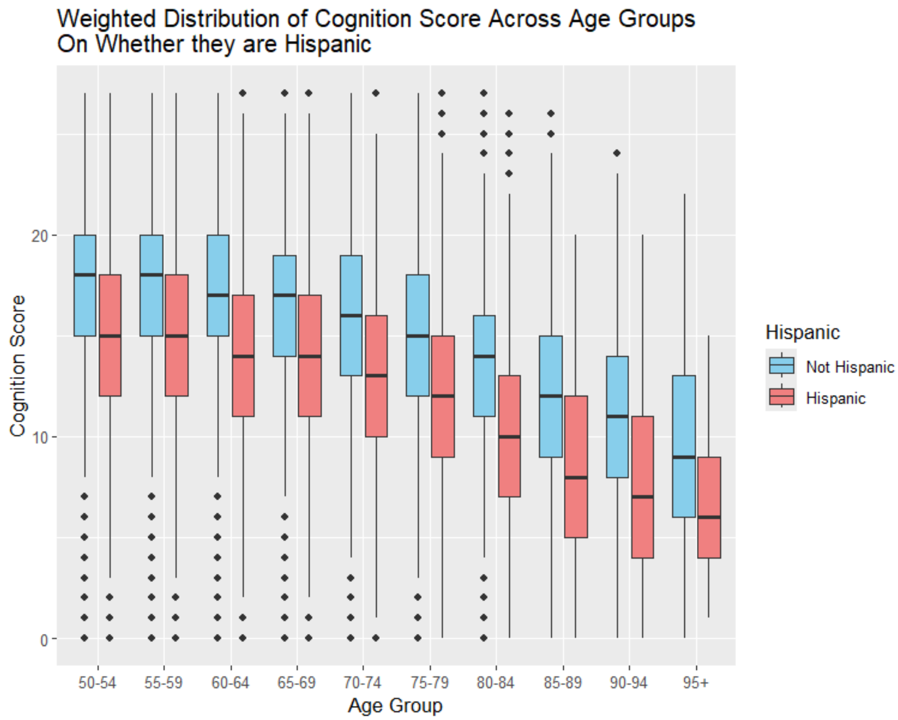
Based on the series of boxplots, it appears that smokers have a higher baseline cognition score than non-smokers but decline at a slower rate. In a linear mixed-effects model the differences in the declining rates could be solved with a fixed interaction affect of time and smoking status. In older age groups, the distribution of cognition scores is wider for smokers than for younger age groups, and then in the 95+ age group the median is about the same as the 25% quartile. Because the distribution changes dramatically for the non-smokers, a random effect may be needed, but the frequency of non-smokers should be found first.



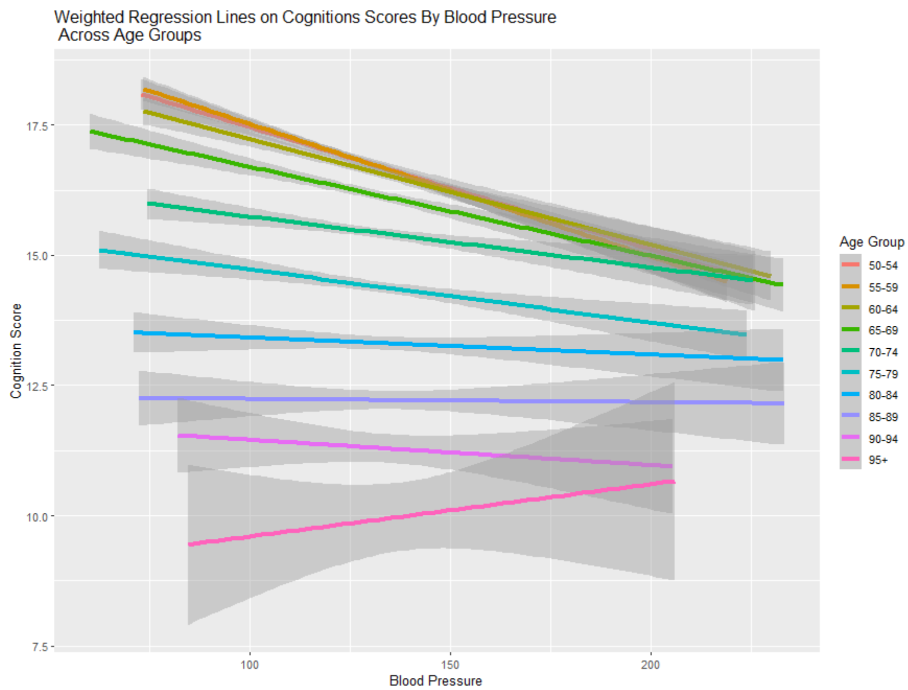
Based on the weighted frequency plot, the proportion of smokers decreases as the age group increases, becoming near 0 in age-groups above 90. The implication for the mixed-effect model is that random effects for smoking will likely not be necessary due to the little information on cognition among smokers older than 90 years. There is less confidence in the rate of cognitive decline of this group; therefore, the interaction effect between non-smokers and age proposed earlier may not be required, as the cognition in both groups could decline at the same rate. However, the impact of the interaction effect should be investigated in the modeling phase.



From this series of boxplots, people with depression have a lower cognition score in all age groups, indicating the need for a fixed effect. The skewness of the distribution for non-depressed cognition scores does seem to change at various different age groups, as the median relative to the quartiles seems to change from the 60-64 age group to the 65-69 age group. However, in other age groups it does seem consistent, so it is likely that no random effects would be needed for depression.



As depicted in the accompanying boxplots, people of Hispanic descent tend to have lower cognition scores than people of non-Hispanic descent at all age groups. In the HRS data set, they oversampled Hispanic individuals, so there is a greater confidence that there is a difference in cognition. If Hispanic was to be included as a covariate, it would only need to be a fixed effect with no interaction or random effect, as the distributions are consistent and decrease at the same rate.



From the weighted regression lines it can be found that in the younger age groups, systolic blood pressure has a negative relationship with cognition scores. In the older age groups, the slope flattens, or in the case of the 95+ age group the slope is actually positive. This could mean that a fixed effect with a negative coefficient and an interaction effect with a positive coefficient may be required for the model. The standard error for the 95+ age group is very wide, indicating that the slope may not be positive; however, the standard errors are small for the age groups with flatter lines, so an interaction effect should still be explored.

5 Next Steps

In the months of September through December, I studied how longitudinal studies differed from cross-sectional studies, how to use survey design and weighting, the inner-workings of mixed effects models, and performed some exploratory data analysis. Throughout the months of January through April, I will continue to perform exploratory data analysis and go through the modeling process. The modeling process will involve multiple cycles of selecting models and parameters, performing hypothesis tests on estimates, and reviewing model diagnostics.

6 Bibliography

References

[1] (RAND HRS Longitudinal File 2022) public use dataset. Produced HRS and (). distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor MI. In: ().

[2] J Rasmussen H Langerman. "Alzheimer's Disease - Why We Need Early Diagnosis". In: *Degenerative Neurological and Neuromuscular Disease* 9 (2019), pp. 123–130. DOI: `https://doi.org/10.2147/DNND.S228939`.

[3] Lisa Bransby E. Rosenich P. Maruff Y.Y. Lim. "How Modifiable Are Modifiable Dementia Risk Factors? A Framework for Considering the Modifiability of Dementia Risk Factors". In: *The Journal of Prevention of Alzheimer's Disease* 11 (2024), pp. 22–37. DOI: `https://doi.org/10.14283/jpad.2023.119`.

[4] Peter Lynn. *Methodology of Longitudinal Surveys*. Wiley Series in Survey Methodology. Wiley, 2009. ISBN: 9780470018712.

[5] Jinh Cheng Lloyd J. Edwards Mildred M. Maldonado-Molina Kelli A. Komro Keith E. Muller. "Real longitudinal data analysis for real people: Building a good enough mixed model". In: *Statistics in Medicine* 29 (2010), pp. 504–520. DOI: `https://doi.org/10.1002/sim.3775`.

[6] Patterson H.D Thompson R. "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". In: *Journal of the American Statistical Association* 72 (1977), pp. 320–338. DOI: `https://doi.org/10.2307%2F2286796`.

[7] Antonio Terracciano Yannick Stephan Martina Luchetti Emiliano Albanese Angelina R Sutin. "Personality traits and risk of cognitive impairment and dementia". In: *Journal of Psychiatric Research* 89 (2017), pp. 22–27. DOI: `https://doi.org/10.1016/j.jpsychires.2017.01.011`.

[8] Changbao Wu Mary E. Thompson. *Sampling Theory and Practice*. ICSA Book Series in Statistics. Springer Nature Switzerland, 2020. ISBN: 9783030442460.

[9] Peter J. Diggle Kung-Yee Liang Scott L. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. Claredon Press, 1994. ISBN: 9780198522843.