Jonathon Hirschi

## Project Overview

Productive work is hypothesized to reduce the rates of incident dementia. In this project, we analyze the association between having a paid job and the rates of dementia for older adults. Additionally, the difference of this effect by race is examined.

## Data Description

The data consists of longitudinal observations from 3,075 highly functioning black and white adults over 11 years. The response variable is a binary diagnosis of dementia, and the associated year of onset. This response variable had to be constructed using hospital visits, medical records, and a cognitive test called the 3M score.

The main predictors of interest are whether the individual has a paid job and the race (black vs white) of the individual. Additional variables include gender, family income, education level, age at year 1 of the study, the site the study took place (one of 2 locations), and a genetic marker of a protein “apoehap” that is meant to be an indicator of genetic propensity for dementia. Family income, education level, and the apoehap protein were only measured in year 1 of the study, so they are considered fixed in time for this analysis. Additionally, race and gender are also constant in time. Due to the complexity of the data and the project, competing risk of other diseases and attrition due to death are not considered at this stage.

## Methods

The variables used are summarized in the following table:

| **Variable Name** | **Variable Type** |
| --- | --- |
| Dementia | (Response) Binary |
| Race | Binary |
| Gender | Binary |
| Site | Binary |
| Apoehap | Categorical |
| Family Income | Continuous |
| Education Level | Categorical |
| Year 1 Age | Continuous |

Gender was coded as binary in the data. Education level was coded as 3 categories, so it will be treated as a categorical variable. Missing data will be excluded from the analysis, then imputation methods will be examined if time permits. A critical challenge of this study was that different variables were encoded with different names and stored in different datasets. Pattern matching on the variable suffixes was necessary to construct the dataset.

The model used is a Cox Proportional Hazards model, a type of statistical survival analysis. The goal of the model is to estimate how the predictor variables affect the rate of onset of dementia. The model estimates a *hazard ratio* for each predictor. There are also significance tests which examine the proportional hazards assumption, which will also be applied. If this ratio is greater than 1, then this predictor is associated with an increased rate in the onset of dementia. If the ratio is less than 1, there is an associated decrease in the rate of onset of dementia.

The main assumption of the model is that hazards are proportional. This means that the ratio of the hazards is constant in time for any two individuals. This assumption can be assessed with “Kaplan-Meier” curves. The curves should have consistent shapes in time, rather than crossing or leveling out in inconsistent ways. Additionally, there is an assumed linear relationship between the log hazard and the predictor variables. The linearity assumption will be tested on the continuous variables using “Martingale residuals”. The plots of the residuals versus the continuous variables should be roughly linear. Finally, there should be no outliers present in the data. The outliers assumption can be examined using the deviance of the residuals. The necessary code and methods will be used from two R packages: “survival” and “survminer”.

Modeling Sources:

<http://www.sthda.com/english/wiki/cox-proportional-hazards-model#the-need-for-multivariate-statistical-modeling>

<https://www.theanalysisfactor.com/assumptions-cox-regression/>