Linear Regression Analysis of Age and Dementia Onset*

A deep dive on the relationship between patient age and length before dementia onset

Christie Ngo

September 23, 2025

Dementia is a multifacted set of conditions whose prevalence is strongly associated with age. A simple linear regression model on Korea Health Panel survey data using biological age is constructed to examine the isolated influence on dementia onset duration. Results show a lack of linear relationship, with age alone only explaining a small proportion of the variation in time until onset. Future studies of age's impact on dementia require more complex modeling techniques and data transformations.

1 Introduction

Dementia refers to a set of conditions related to cognitive decline with various causes and risk factors ("Dementia - Symptoms and Causes" (n.d.)). Symptoms may include memory loss, disorientation, poor coordination, depression, and anxiety. As populations all around the world continue to age, the increasing physical and financial impact of dementia is a public health issue that should be thoroughly investigated. In the past, the highest risk factor has been recognized as chronological age as dementia prevalence increases with it (Daviglus et al. (2010)). However, subsequent research suggests that subjective age, the difference between one's biological age and how old they feel, may hold greater influence on risk of dementia (Kotter-Grühn (2016)).

Because dementia is a complicated condition, a simple linear regression model with the predictor variable of age will be insufficient in capturing nuanced associations but serves as a valuable starting point. Previous research tackled dementia modeling more holistically, but it remains important to fully understand the relationship with biological age. Subjective age is a qualitative assessment and is often not reported. In China, a team of researchers studied dementia risk

*Project repository available at: https://github.com/christiecngo/math261a_project1.

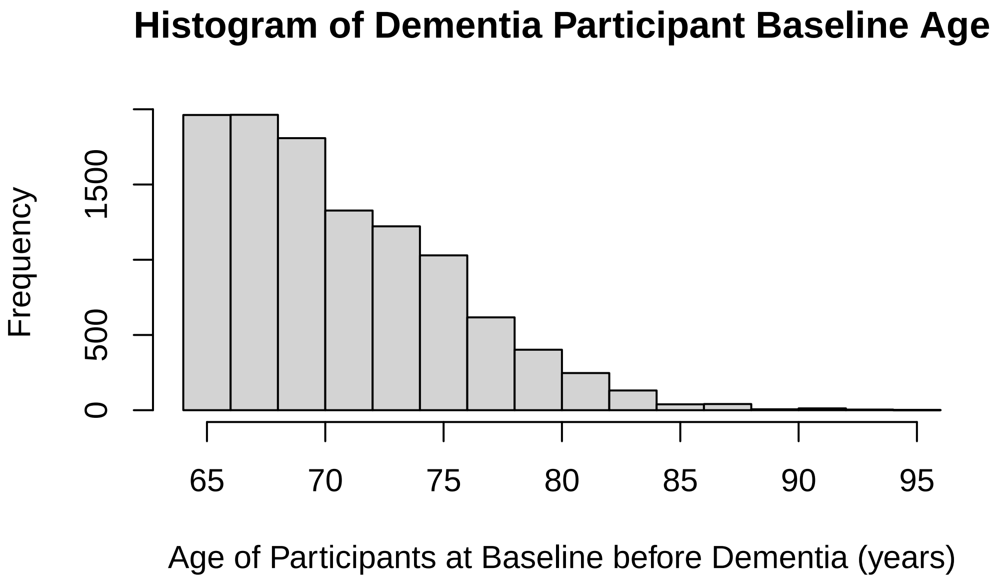
among rural seniors using Cox models but kept education, marital status, self-rated AD8 score, and stroke history in addition to age (Liu et al. (2025)).

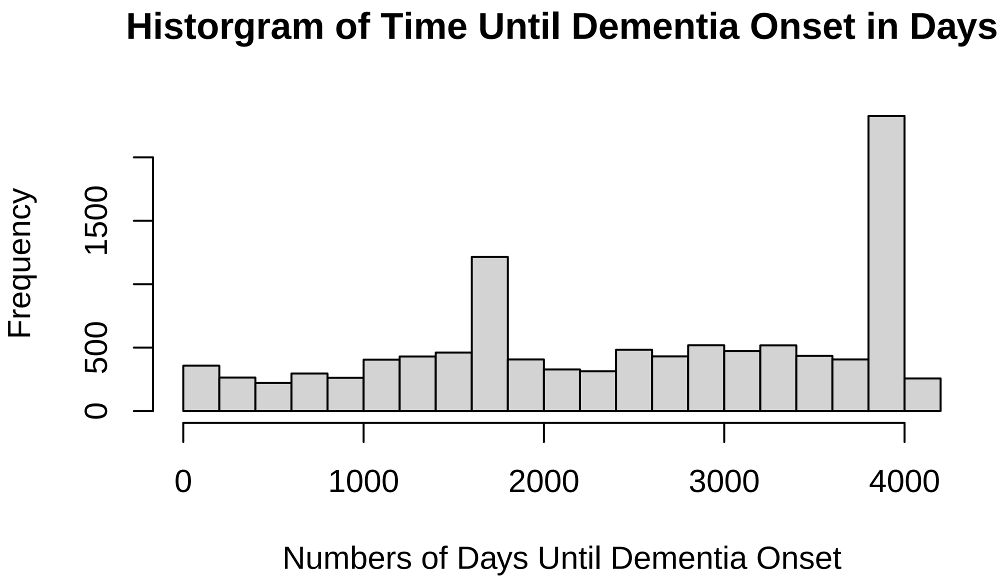
This paper aims to isolate the usefulness of considering the more widely available demographic, biological age, for determining time of dementia onset. We will first introduce the dementia study dataset of interest followed by construction of a simple linear regression model. After this, we assess the model's fit and discuss implications of these results.

2 Data

The data is a subset from the Korea Health Panel (KHP) surveys from 2006 to 2018 (Islam et al. (2025)). This computer-assisted data collection process is conducted by the Korea Institute for Health and Social Affairs and the National Health Insurance Service (NHIS); it occurs annually for households selected using clustered probability sampling on population census data (Chung 2022). More comprehensive alternative sources of this survey would include more participants of interest with other additional survey responses included. There are 10811 rows of observations, each unit representing participant responses during each survey wave since the study started. Participants, with unique identifiers, are all over 65 years old and may have multiple rows listed if they have been observed for more than one period. The cohort also includes many seniors who have not developed dementia.

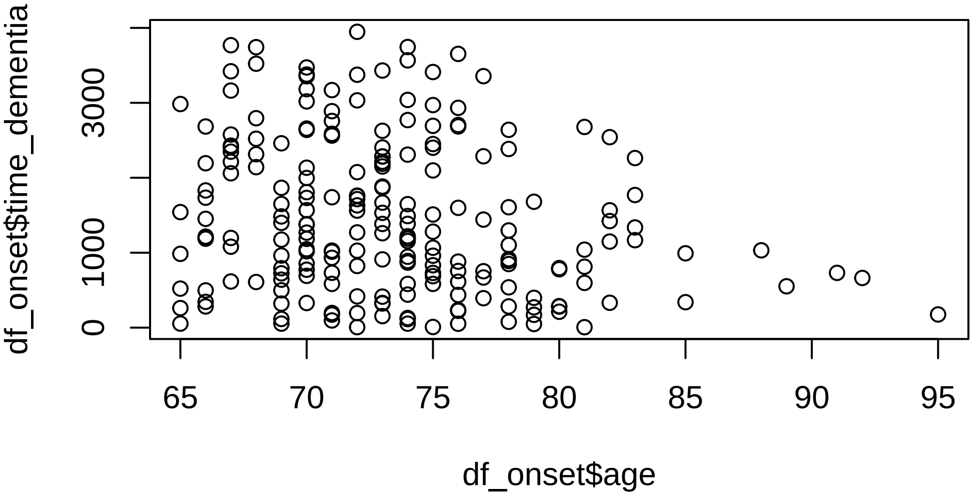
For the question of interest, filtering this dataset to only participants who ended up developing dementia led to only 2% of the cohort remaining. It should be noted that the interquartile range of biological age at baseline is only 7 years. The data is extremely right skewed. 65 year olds who are included in the study may already exhibit other risk factors that group them with older participants. The distribution of dementia onset days is almost uniform with a small peak in the middle of the range and a large one near the maximum.


Histogram of Dementia Participant Baseline Age


Historgram of Time Until Dementia Onset in Days

3 Methods

Model fitting was completed using R's "lm()" function and data was preprocessed with help of the package "dplyr". The final cohort is obtained by filtering for the earliest instance of a dementia recording for patients with dementia. This led to a data size of 216 distinct participants.



The motivation for using a simple linear regression model is to utilize a formal mean to express a statistical relationship between two variables, where there is a probability distribution for the response Y for each level of predictor X; the means for the probability distributions vary with X in a systematic way (Kutner et al., 2005). A linear regression model is considered "simple" when there is only one predictor variable. The chosen model is linear in both parameters and predictor variables because all parameters appear without a mathematical operation and the predictor exists only in the first power. In our case, the sole predictor is the age of the participant at the baseline before the onset of dementia. Our measured response is duration in days before the onset of dementia from the initial baseline date. This overall model takes the form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

i ranges from 1 to the number of observations in the dataset. The term ϵ_i captures the noise or randomness within the real-world observations that is not included in the regression model. β_0 is the intercept that corresponds to the value if the participant's age were 0. β_1 is the slope of the model, representing the change in response Y when there is a unit increase in predictor X.

In order to consider how well our linear regression model fit the data, we will consider residuals. Residuals are the difference between the actual observed value and what the linear regression predicted. They are given by this formula:

$$e_i = Y_i - \hat{Y}_i$$

Summing up all residuals will give us the SSE, or error sum of squares. SSR is the regression sum of squares and accounts for the variation of the fitted values around the mean of observed values. SSTO is the total sum of squares, a combination of the two previous metrics.

$$SSTO = SSR + SSE$$

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{X})^2 + \sum(Y_i - \hat{Y}_i)$$

The amount of deviation explained by SSR, or regression line, out of the total deviation gives rise to the R^2 metric. We will be using this to determine how well the model fits the data.

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

4 Results

A simple linear regression model revealed that using age alone only accounts for a trivial proportion of relationship between dementia onset duration and biological age. Only approximately 5% of the variation in onset length was explained after factoring in age (R^2= 0.0499767). Another meaningful discovery includes the lack of linear association between these variables based on visual assessment of residuals in Figure 1. The QQ-plot is likewise showing a lack of linear relationship (Figure 2). However, the summary report of the linear model does reveal a notable negative association between participant's biological age and the time it takes for them to develop dementia. β_1 is `rlm`_fit$coefficients[2]‘, meaning that there is a 43 day decrease in dementia onset time for a 1 year increase in age. This comes a significant p-value of 9.3805132×10^{-4}.

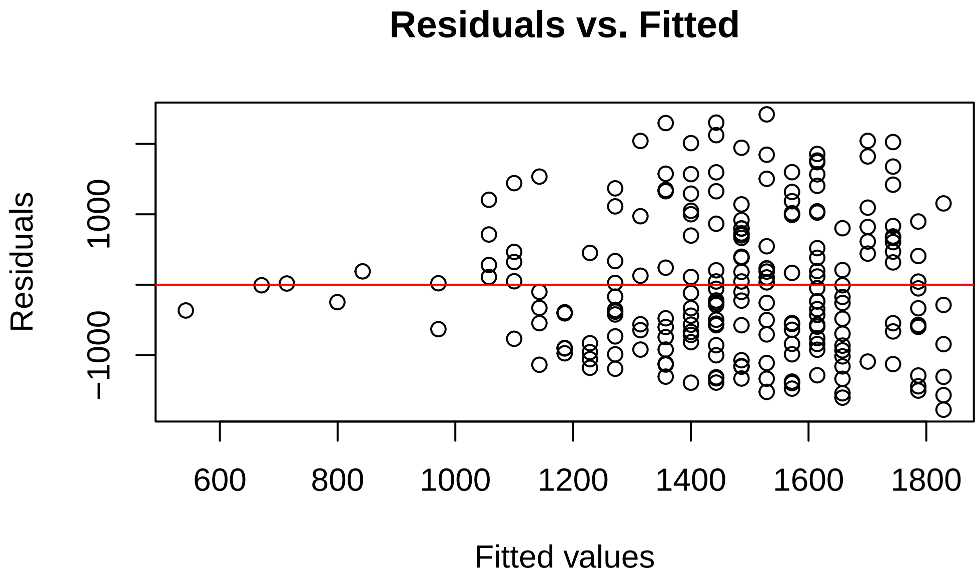


Figure 1: The residuals of the linear regression model are not randomly scattered

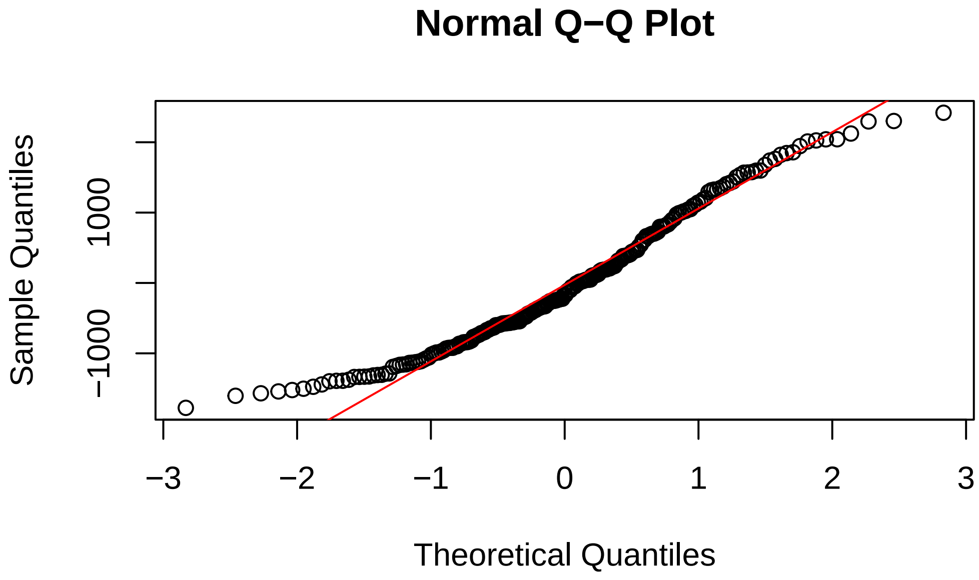


Figure 2: The QQ-plot shows curvature towards the tail ends

5 Discussion

Ultimately, because of the low explainability from this simple linear regression model, more advanced models that include interaction terms and data transformations may see greater success in inference of dementia onset using a participant's demographics. Though age has a significant positive association, there is still a lack of ability to predict dementia onset duration. Due to linear assumption violations, nonparametric techniques can yield greater results. Our limited dataset of dementia participants hinders the generalizability of these regression results. Additionally, since age is a factor that can contribute intricate relationships with other comorbidities, studying interaction terms can help improve the model.

References

Daviglus, Martha L, Carl C Bell, Wade Berrettini, Phyllis E Bowen, E. Sander Connolly, Nancy Jean Cox, Jacqueline M Dunbar-Jacob, et al. 2010. "National Institutes of Health State-of-the-science Conference Statement: Preventing Alzheimer Disease and Cognitive Decline." *Annals of Internal Medicine* 153 (3): 176–81. https://doi.org/10.7326/0003-4819-153-3-201008030-00260.

"Dementia - Symptoms and Causes." n.d. *Mayo Clinic*. Accessed September 24, 2025. https://www.mayoclinic.org/diseases-conditions/dementia/symptoms-causes/syc-20352013.

Islam, Md. Akhtarul, Prosanta Kumar Mondal, Hyun J. Lim, and Zahid A. Butt. 2025. "Dementia." Harvard Dataverse. https://doi.org/10.7910/DVN/ANLJSG.

Kotter-Grühn, Dana. 2016. "Looking Beyond Chronological Age: Current Knowledge and Future Directions in the Study of Subjective Age." *Gerontology* 62 (1): 86–93. https://doi.org/10.1159/000438671.

Liu, Keke, Tingting Hou, Yuqi Li, Na Tian, Yifei Ren, Cuicui Liu, Yi Dong, et al. 2025. "Development and Internal Validation of a Risk Prediction Model for Dementia in a Rural Older Population in China." *Alzheimer's & Dementia* 21 (2): e14617. https://doi.org/10.1002/alz.14617.