Association of body mass index with risk of dementia using longitudinal and survival analysis

Nick Wibert

Department of Statistics

Florida State University

nlw22@fsu.edu

December 15, 2023

1 Introduction

Alzheimer's disease is a neurodegenerative disorder characterized by cognitive dysfunction, behavioral disturbances, and difficulty with general day-to-day activity. Existing literature suggests an association between one's body mass index (BMI) and the risk of dementia. BMI is calculated as weight (kg) divided by squared height (m^2) is widely regarded as one of the best measurements of fatness in adults [1]. We aim to study the relationship between BMI and dementia using a combined approach of longitudinal data analysis and survival analysis methods.

1.1 Data Description

The dataset used in this analysis comes from a longitudinal household survey conducted by the Institute for Social Research at the University of Michigan, and was downloaded from the RAND corporation's website. There are 38,558 subjects included in this dataset, and observations were collected in two-year intervals for up to 12 observations per subject (depending on censoring). Various demographic data were collected for each participant, including race, gender, age, and years of education. The pivotal measurements for each time point are the participant's height and weight (the quantities needed to compute BMI) as well as their dementia diagnosis (binary indicator) which was determined by researchers based on cognitive tests given at the relevant time point.

2 Methodology

2.1 Data Preparation

2.1.1 Data Cleaning

Before proceeding with any analysis, it was necessary to perform some data validation and cleaning. Of the 38,558 unique subjects in this dataset (identified by the field `HHIDPN`), there exist 180 subjects who are missing a dementia diagnosis for all available time points. Furthermore, there are an additional 211 patients who have no BMI data available whatsoever. Since this makes up a relatively small proportion of our dataset, rather than attempting to impute these values, we drop the aforementioned patients entirely, bringing our sample size down to $N = 38167$. In preparation for adjusting models by various demographic factors, we also make a second copy of this dataset where we drop an additional 224 patients who are missing years of education, race, gender, or age for all time points ($N = 37943$).

2.1.2 Feature Creation

While BMI is a continuous variable, it will be helpful to treat it as a categorical variable to determine differences between various subgroups, especially in the context of survival analysis. To prepare for categorical data analysis, we create a new column which categorizes BMI into one of four weight statuses defined by the Centers for Disease Control and Prevention (CDC) (Table 1). The BMI observations in our dataset have a precision of two decimal places, so to account for the gap between categories as defined below, we set the upper bound for each category as the (exclusive) lower-bound from the subsequent category (i.e. $BMI \in [18.5, 25)$ is considered healthy). Dummy variables were also created for convenience and ease of interpretation. Each patient also has a "baseline" BMI (`bmi_base`, the patient's BMI at first time point), so we perform the same feature creation for baseline BMI as well.

BMI	Weight Status
< 18.5	Underweight
$18.5 - 24.9$	Healthy
$25.0 - 29.9$	Overweight
≥ 30	Obesity

Table 1: BMI categories for adults as defined by the CDC

2.1.3 Survival Analysis Preparation

The longitudinal structure of the data is not directly compatible with survival analysis methods. Thus we need to prepare a new dataset which is set-up for survival analysis, where each patient has one observation for two variables: `time`, the time at which they were diagnosed with dementia or exited the study (censored); and `dementia`, a binary indicator with 1 indicating dementia diagnosis and 0 indicating censorship.

For uncensored patients, the column in our dataset represented as `time` indicates the point (between 1 and 12) where the first instance of dementia diagnosis occurred (some patients may be diagnosed with cognitive impairment at one time point, and then diagnosed as cognitively normal in a subsequent time point). So, these patients will simply be represented by a single row in the survival dataset with `time=time` and `dementia=1`. For patients who were censored at some point, the `time` column is blank; so, the time of censoring must be identified by observing the point at which the last observation was made. For example, if patient A has `time=NA`, cognitively normal outcomes (`dementia=00` for time points 1-10, and missing data for time points 11-12, this indicates that patient A was censored at the 10th time point. So, patient A will be represented by a single row in the survival dataset with `time=10` and `dementia=0`, to indicate that they were censored at time 10 (10^+). This process was performed for all patients to produce a new dataset ready for survival analysis (see Appendix A for data prep `R` code).

2.2 Statistical analysis

All data analysis was performed using the `R` programming language (see Appendices B and C for longitudinal data analysis and survival analysis codes, respectively).

2.2.1 Longitudinal data analysis with GEEs

To incorporate the longitudinal structure of the data in our exploration of the relationship between BMI and dementia, we use generalized estimating equations (GEE); specifically, the marginal logistic model for binary outcomes. This approach assumes that (1) dementia diagnosis depends only on the data collected at time t, and no other time before/after; and (2) the relationship between dementia diagnosis and the set of predictors which we regress on is *time-independent*; in context, we assume the relationship between BMI and dementia does not change over time. These are strong assumptions, and literature exists which demonstrates this relationship to be time-varying [3] but for our purposes, the GEE serves as a good starting point for investigating the association between BMI and dementia.

When fitting a GEE, one must choose the correlation structure; that is, the way that observations within a cluster (in our case, a single patient) are correlated with one another. A correctly specified correlation structure can lead to improvements in efficiency, but this is often an unrealistic goal with little benefit [2]. The simplest choice is an "independence" correlation structure, which assumes all observations for a single patient to be independent from one another. While this is not typically a reasonable assumption, its simplicity and consistency make it a popular choice amongst researchers, and the efficiency loss is usually insignificant. For comparison, we fit GEEs with various correlation structures: `independence`, `exchangeable` (assuming all pairs of observations for a single patient share the same correlation), and `unstructured` (which allows all correlations to vary freely).

Though we are interested specifically in the relationship between BMI and dementia, it may be helpful to adjust by various demographic factors. We started by fitting a simple GEE with BMI as the only covariate, and then a fully adjusted GEE containing all demographic features which could potentially be relevant. Insignificant features were then removed resulting, in a GEE adjusted by race, age, and years of education.

2.2.2 Survival analysis

All survival analysis was performed using the dataset discussed in section 2.1.3. Broadly speaking, survival analysis deals with modeling "time-to-event" outcomes; in our case, the event in question is either a dementia diagnosis, loss-to-followup (censoring), or the end of the study (treated as censorship at time point 12). We used three common survival analysis techniques to evaluate the effect of BMI in dementia: the Kaplan-Meier estimator for functions, log-rank tests, and the Cox proportional-hazards model. In all cases, we group patients according to their BMI category to determine what differences exist amongst the survival functions for each group, and investigate the relative risk of dementia between different BMI categories.

It is important to note that the BMI for a given patient in the survival analysis dataset is chosen as the BMI at the time of event (dementia diagnosis or censoring). So, the relationship that is being discussed in the previous paragraph is that between BMI and dementia at the instant of dementia diagnosis. To extend our analysis, we perform the same steps using baseline BMI (BMI at the first time point), and compare the results with those obtained using BMI at the time-of-event.

Similar to our process with GEEs, we want to consider potential confounding variables in our survival analysis. Upon exploring the distributions of the various continuous/categorical predictors across BMI levels, the only feature which seemed to be a potential confounder was gender (Figure 1), as the ratio of men to women seemed much different in the "overweight" BMI level relative to the other three levels. We performed log-rank tests and fit the Cox model while stratifying by gender, but found that this stratification had little impact on the results.

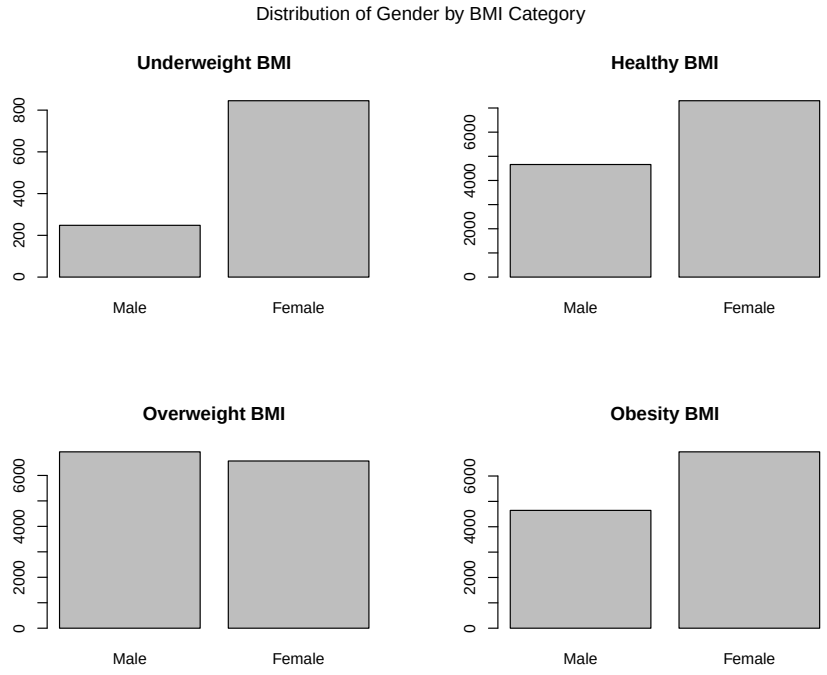

Figure 1: Distribution of gender within each BMI level

3 Results

3.1 Longitudinal data analysis results

The first and simplest model was the marginal logistic GEE with BMI as the only covariate. This simple model was fit using the three correlation structures discussed in section 2.2.1, and the resulting coefficient estimates for BMI are displayed in Table 2 along with the resulting odds ratio (OR) estimate and 95% confidence interval. We see that there is a slight efficiency gain with each subsequent choice of correlation structure moving down the table, but this gain is very minimal. There are also variations in the coefficient estimates, but again, very minimal differences. Observing the odds ratios (obtained by exponentiating the coefficient estimate for BMI, which represents log-odds) under the `independence` correlation structure, for each unit-increase in BMI, we estimate about 4.38% reduction in the odds of developing dementia. Similarly, we estimate about 5.78% and 4.89% reductions using the `exchangeable` and `unstructured` structures, respectively. In all cases, higher BMI appears to be associated with reduced odds of developing dementia, albeit a very slight reduction. We are confident in the direction of this relationship as the OR confidence intervals do not include 1.

As far as our choice for the correlation structure goes, the `exchangeable` correlation matrix estimates a correlation of about 0.236 between pairs of observations for a single observation. This isn't particularly strong, so we can reasonably assume a minimal loss of efficiency and feel justified in using the `independence` correlation structure for simplicity sake.

Correlation Structure	BMI Coef.	Robust S.E.	p-value	Odds Ratio (OR)	OR 95% lower	OR 95% upper
`independence`	-0.04482	0.003381	0	0.9562	0.9499	0.9625
`exchangeable`	-0.05957	0.003354	0	0.9422	0.9360	0.9484
`unstructured`	-0.05018	0.003229	0	0.9511	0.9451	0.9571

Table 2: Results from GEE (BMI only) using different correlation structures

Next we fit a "fully-adjusted" GEE, adjusting by all covariates that could potentially influence dementia diagnoses. The initial set of predictors included race (White, African-American, or Other), Hispanic (yes/no), gender (male/female), age, and years of education. The log-odds estimates for these predictors as well as their p-values are listed in Table 3. The indicators for Hispanic and gender are insignificant, so we drop these to obtain our final "adjusted" GEE (Table 4).

	Estimate	Robust S.E.	Robust z	p-value
(Intercept)	-5.61364	0.199580	-28.1273	0.0000
bmi	-0.02964	0.003365	-8.8074	0.0000
race_AA	1.24958	0.042050	29.7169	0.0000
race_other	0.60572	0.083445	7.2590	3.9e-13
RAHISPAN	-0.05112	0.070654	-0.7236	**0.4693**
RAGENDER	0.02774	0.036418	0.7618	**0.4462**
age	0.07897	0.001681	46.9706	0.0000
RAEDYRS	-0.24112	0.005474	-44.0477	0.0000

Table 3: Coefficient estimates from fully-adjusted GEE

	Estimate	Robust S.E.	Robust z
(Intercept)	-5.61232	0.181585	-30.907
bmi	-0.02944	0.003363	-8.754
race_AA	1.25883	0.040796	30.857
race_other	0.58452	0.078349	7.460
age	0.07909	0.001660	47.658
RAEDYRS	-0.23903	0.004607	-51.890

Table 4: Coefficient estimates from adjusted GEE, after dropping RAHISPAN and RAGENDER

After adjusting for these other covariates, our new odds ratio estimate for BMI is $\exp(-0.02944) = 0.971$, indicating about a 2.9% reduction in odds of dementia diagnosis per-unit increase of BMI (holding all other covariates constant). So covariate adjustment improved precision/efficiency (lower robust S.E.) and led to a weaker estimate of the association between BMI and dementia.

Finally, we try a BMI-only GEE once again, but this time treating BMI as a categorical variable with "Healthy" BMI as the reference level (Table 5). The categorical BMI is still significant, and we can get an idea of how the odds of dementia vary between these broad groups as a result. Individuals with underweight BMI have 2.19 times the odds of developing dementia relative to those with healthy BMI (overweight: 0.79 OR, obese: 0.64 OR). This echoes our earlier conclusion that higher BMI is associated with lower odds of dementia.

	Estimate	Robust S.E.	Robust z	p-value	Odds Ratio	95% lower	95% upper
(Intercept)	-2.8556	0.02446	-116.728	0	0.05752	0.05483	0.06035
bmi_under	0.7827	0.06641	11.786	0	2.18741	1.92044	2.49151
bmi_over	-0.3043	0.03236	-9.403	0	0.73766	0.69233	0.78596
bmi_obese	-0.5276	0.04040	-13.060	0	0.59001	0.54509	0.63863

Table 5: Coefficient estimates for GEE (categorical BMI only)

3.2 Survival analysis results

Using the dataset prepared for survival, we can easily compute the Kaplan-Meier estimators in R. The resulting estimates for the survival function within each BMI level are plotted together in Figure 2 along with confidence bands.

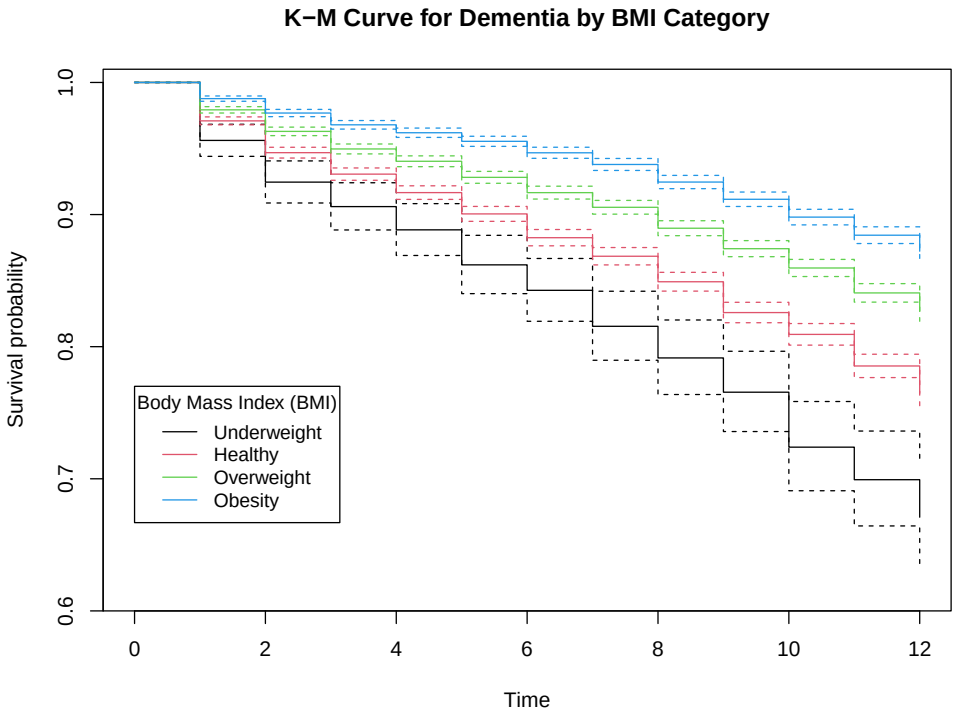

Figure 2: Kaplan-Meier estimators for each BMI level

Based on our insights from longitudinal data analysis, this plot is not surprising. We see clear differences in the survival amongst the BMI levels, and there is a clear ordering to them. The lower the BMI, the higher the risk of dementia. To obtain a more concrete result, we perform a log-rank test for the four levels of BMI. The test produced a test statistic $\chi^2_3 = 493$ with a corresponding p-value less than 2×10^{-16}, giving us strong evidence that the survival functions for these four groups do not overlap.

Finally, to quantify the relative risk for each BMI level using "Healthy" as a reference, we fit a Cox proportional-hazards model (Table 6). In this context, the exponentiated estimates give us the hazard ratio (HR) estimates, which is the multiplicative effect of the given variable on the hazard function (which can be thought of as the probability of an individual being diagnosed with dementia at a time t).. The results here align again with our conclusions from the longitudinal data analysis, in that higher BMI seems to be associated with lower risk of dementia, relative to healthy BMI. Specifically, having an underweight BMI is associated with an increase in the hazard by a factor of 1.46, or 46%. Having overweight BMI is associated with a 29.1% reduction in the hazard, and having obese BMI is associated with a 50.1% reduction in the hazard (all relative to healthy BMI).

	coef	robust se	z	Pr(>|z|)	Hazard Ratio	lower .95	upper .95
bmi_under	0.3806	0.06940	5.485	4.142e-08	1.4632	1.2771	1.6764
bmi_over	-0.3441	0.03230	-10.651	1.722e-26	0.7089	0.6654	0.7552
bmi_obese	-0.6957	0.03609	-19.278	8.195e-83	0.4987	0.4646	0.5353

Table 6: Coefficient estimates from Cox proportional-hazards model

We noted earlier that we have been using BMI at the time-of-event. For further investigation, we wish to perform survival analysis techniques using the baseline BMI. The Kaplan-Meier curve for baseline BMI is shown in Figure 3.

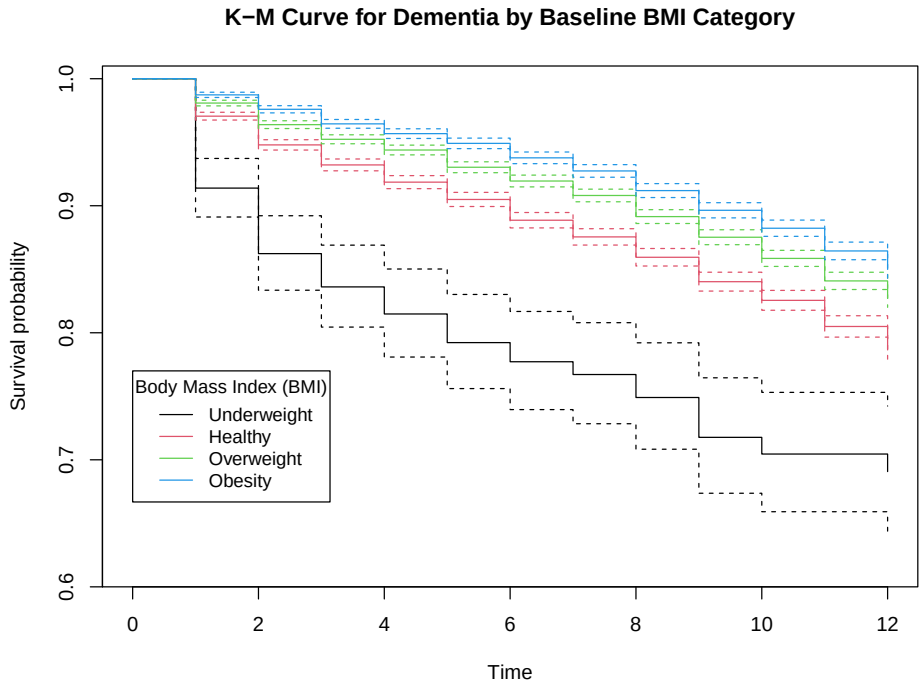

Figure 3: Kaplan-Meier estimators for each baseline BMI level

The resulting log-rank test statistic in this case was $\chi^2_3 = 259$ with a corresponding p-value less than 2×10^{-16}, once again giving us strong evidence that the survival (and hazards) for the four baseline BMI groups are different. Finally, we fit the Cox proportional hazards model using baseline BMI levels (Table 7). Baseline BMI is still significant in the Cox model, and the relative risk for each baseline BMI level is slightly different than the time-of-event BMI; namely, the risk-increase in underweight baseline BMI relative to healthy baseline BMI is even stronger than with time-of-event BMI, and the risk-reduction in overweight/obese baseline BMI relative to healthy BMI is weaker.

	coef	robust se	z	Pr(>|z|)	Hazard Ratio	lower .95	upper .95
bmi_base_under	0.6638	0.09675	6.861	6.833e-12	1.9421	1.6067	2.3477
bmi_base_over	-0.2581	0.03248	-7.947	1.910e-15	0.7725	0.7249	0.8233
bmi_base_obese	-0.4349	0.03533	-12.308	8.159e-35	0.6473	0.6040	0.6937

Table 7: Coefficient estimates from Cox proportional-hazards model using baseline BMI levels

Furthermore, to get a preliminary idea of the relationship of BMI and dementia over time, it may be insightful to perform a log-rank test to compare 2 groups of patients within each BMI level: (1) patients who remained at this BMI level in both baseline and time-to-event; and (2) patients who started at a different baseline BMI level, and changed over to this BMI level by time-of-event. Based on our time-independent assumption, the relationship between BMI and dementia risk should not change over time (that is, the hazard should only depend on BMI at the time of the event itself, and baseline BMI should matter).

The results for these log-rank tests are displayed in Table 8. Clearly, this time-independence assumption is violated; there is strong evidence to suggest that within a certain time-of-event BMI level, a patient's baseline BMI level *does* affect their risk of dementia.

4 Discussion

Through a comprehensive approach utilizing longitudinal analysis and survival analysis, we have established an inverse association between BMI and dementia; the higher one's BMI, the lower their associated risk of dementia. Since high BMI is typically associated with negative health outcomes, this association may seem a bit counterintuitive at first, however we must reiterate that the association explored in this paper is purely time-independent. Due to the debilitating nature of dementia, it seems likely that a patient who is approaching a dementia diagnosis will begin to drop in weight and thus have a lower BMI, and we might expect the time-of-event BMI of patients diagnosed with dementia to be lower on average than non-dementia patients. Furthermore, existing literature demonstrates a direct relationship between midlife BMI and latelife dementia, which we were unable to replicate with our data due to most of our subjects being in later-life [5]

While the association in our results was similar whether we used time-of-event BMI or baseline BMI, there was a weakening in the association for overweight/obese BMIs that may suggest baseline BMI has a different relationship with dementia when compared to time-of-event BMI. Furthermore, our log-rank tests within individual time-of-event BMI levels provide strong evidence that a patient's baseline BMI does in fact affect their hazard of dementia at the time of diagnosis, violating our previous time-independence assumptions. Our results here provide a motivation for further study into the time-varying relationship between BMI and dementia, a relationship which has been explored in existing literature [3][4].

Groups	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
`bmi_base_under=0, bmi_under=1`	700	137	171.4	6.92	28.2
`bmi_base_under=1, bmi_under=1`	394	94	59.6	19.92	28.2
$\chi^2_1 = 28.2, \quad p = 1 \times 10^{-7}$					

(a) Log-rank test for patients who changed to underweight BMI by time-of-event vs. patients who were underweight at baseline and time-of-event

Groups	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
`bmi_base_healthy=0, bmi_healthy=1`	3101	514	572	5.81	8.29
`bmi_base_healthy=1, bmi_healthy=1`	8867	1512	1454	2.29	8.29
$\chi^2_1 = 8.3, \quad p = 0.004$					

(b) Log-rank test for patients who changed to healthy BMI by time-of-event vs. patients who were healthy at baseline and time-of-event

Groups	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
`bmi_base_over=0, bmi_over=1`	4008	516	592	9.73	14.6
`bmi_base_over=1, bmi_over=1`	9497	1319	1243	4.63	14.6
$\chi^2_1 = 14.6, \quad p = 1 \times 10^{-4}$					

(c) Log-rank test for patients who changed to overweight BMI by time-of-event vs. patients who were overweight at baseline and time-of-event

Groups	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
`bmi_base_obese=0, bmi_obese=1`	2549	206	284	21.21	27.9
`bmi_base_obese=1, bmi_obese=1`	9051	1024	946	6.35	27.9
$\chi^2_1 = 27.9, \quad p = 1 \times 10^{-7}$					

(d) Log-rank test for patients who changed to obese BMI by time-of-event vs. patients who were obese at baseline and time-of-event

Table 8: Results for log-rank tests within time-of-event BMI levels

References

[1] José I. Baile. "IS IT USEFUL TO USE THE BODY MASS INDEX TO ASSESS OBESITY IN MUSCULAR PEOPLE?" In: *Nutr Hosp* 32.5 (2015), p. 2353. DOI: https://doi.org/10.3305/nh.2015.32.5.9598.

[2] Peng Ding. *Linear Model and Extensions*. Unpublished textbook. Chap. 25: Generalized Estimating Equation for Correlated Multivariate Data.

[3] Jie Shen et al. "Long-term Weight Change and its Temporal Relation to Later-life Dementia in the Health and Retirement Study". In: *The Journal of Clinical Endocrinology Metabolism* 107.7 (Apr. 2022), e2710–e2716. ISSN: 0021-972X. DOI: 10.1210/clinem/dgac229.

[4] Zhuowei Sun and Hongyuan Cao. "Regression analysis of multiplicative hazards model with time-dependent coefficient for sparse longitudinal covariates". In: (2023). URL: https://arxiv.org/pdf/2310.15877.pdf.

[5] Xu W.L. et al. "Midlife overweight and obesity increase late-life dementia risk: a population-based twin study". In: *Neurology* 76.18 (2011), pp. 1568–1574. DOI: 10.1212/WNL.0b013e3182190d09.

Appendices: R Code

Appendix A: Data Preparation

```
# Libraries ------------------------------------------------------------
library(gee)
library(survival)
library(dplyr)

# Load Data ------------------------------------------------------------
# set working directory
setwd("/Users/nicholasw30/Desktop/School/FSU/Fall 2023/STA5168 - Statistics in Applications III/Final Project/")
# Load dataset, drop index column
dementia <- subset(read.csv("20221013formatted_data-1.csv"), select=-c(X))

# Data Cleaning --------------------------------------------------------
# Drop patients which are missing outcome (dementia) for ALL time points
dementia <-
  dementia %>%
    group_by(HHIDPN) %>%
    filter(!all(is.na(dementia)))

# For patients missing BMI for some time points, impute with the patient's average
dementia <-
  dementia %>%
    group_by(HHIDPN) %>%
    mutate(bmi=ifelse(is.na(bmi), mean(bmi, na.rm=TRUE), bmi)) %>%
    ungroup()

# Check how many patients missing BMI for all time points
na.bmi.patients <-
  dementia %>%
    group_by(HHIDPN) %>%
    filter(any(is.na(bmi))) %>%
    ungroup() %>%
    select(HHIDPN) %>%
    distinct() %>%
    .$HHIDPN
length(na.bmi.patients) # 211 patients
length(na.bmi.patients) / length(unique(dementia$HHIDPN)) # < 1% of dataset

# Since this is only 1% of our patients, we drop these patients rather than trying to impute BMI
dementia <- dementia %>% filter(!(HHIDPN %in% na.bmi.patients))

# Check how many patients missing age for all time points
all.na.age.patients <-
  dementia %>%
  group_by(HHIDPN) %>%
  filter(all(is.na(age))) %>%
  ungroup() %>%
  select(HHIDPN) %>%
  distinct() %>%
  .$HHIDPN
length(all.na.age.patients)
length(all.na.age.patients) / length(unique(dementia$HHIDPN)) # <1% of dataset

# Create new dataset for adjusted GEE, which drops the patients missing age for all time points
# (for unadjusted GEE, we only care about BMI and dementia)
dementia.adj <- dementia %>% filter(!(HHIDPN %in% all.na.age.patients))

# Check how many patients missing gender for all time points
all.na.gender.patients <-
  dementia.adj %>%
  group_by(HHIDPN) %>%
  filter(all(is.na(RAGENDER))) %>%
  ungroup() %>%
  select(HHIDPN) %>%
  distinct() %>%
  .$HHIDPN
```

```
length(all.na.gender.patients)
length(all.na.gender.patients) / length(unique(dementia.adj$HHIDPN)) # <1% of dataset

# drop missing gender observations for adjusted GEE
dementia.adj <- dementia.adj %>% filter(!(HHIDPN %in% all.na.gender.patients))

# Check how many patients missing race for all time points
all.na.race.patients <-
  dementia.adj %>%
  group_by(HHIDPN) %>%
  filter(all(is.na(RARACEM))) %>%
  ungroup() %>%
  select(HHIDPN) %>%
  distinct() %>%
  .$HHIDPN
length(all.na.race.patients)
length(all.na.race.patients) / length(unique(dementia.adj$HHIDPN)) # <1% of dataset

# drop missing race observations for adjusted GEE
dementia.adj <- dementia.adj %>% filter(!(HHIDPN %in% all.na.race.patients))

# Check how many patients missing education for all time points
all.na.edu.patients <-
  dementia.adj %>%
  group_by(HHIDPN) %>%
  filter(all(is.na(RAEDYRS))) %>%
  ungroup() %>%
  select(HHIDPN) %>%
  distinct() %>%
  .$HHIDPN
length(all.na.edu.patients)
length(all.na.edu.patients) / length(unique(dementia.adj$HHIDPN)) # <1% of dataset

# drop missing education observations for adjusted GEE
dementia.adj <- dementia.adj %>% filter(!(HHIDPN %in% all.na.edu.patients))


# create ordered categorical BMI column based on the levels given at cdc.gov
dementia$bmi_cat <- factor(ifelse(dementia$bmi < 18.5, "Underweight", ifelse(
                                  dementia$bmi < 25, "Healthy", ifelse(
                                  dementia$bmi < 30, "Overweight", "Obesity"))),
                           levels=c("Underweight","Healthy","Overweight","Obesity"))
# dummy columns
dementia$bmi_under <- ifelse(dementia$bmi_cat == "Underweight", 1, 0)
dementia$bmi_healthy <- ifelse(dementia$bmi_cat == "Healthy", 1, 0)
dementia$bmi_over <- ifelse(dementia$bmi_cat == "Overweight", 1, 0)
dementia$bmi_obese <- ifelse(dementia$bmi_cat == "Obesity", 1, 0)

# Do the same for bmi_base
dementia$bmi_base_cat <- factor(ifelse(dementia$bmi_base < 18.5, "Underweight", ifelse(
  dementia$bmi_base < 25, "Healthy", ifelse(
    dementia$bmi_base < 30, "Overweight", "Obesity"))),
  levels=c("Underweight","Healthy","Overweight","Obesity"))
# dummy columns
dementia$bmi_base_under <- ifelse(dementia$bmi_base_cat == "Underweight", 1, 0)
dementia$bmi_base_healthy <- ifelse(dementia$bmi_base_cat == "Healthy", 1, 0)
dementia$bmi_base_over <- ifelse(dementia$bmi_base_cat == "Overweight", 1, 0)
dementia$bmi_base_obese <- ifelse(dementia$bmi_base_cat == "Obesity", 1, 0)

# Age categories
dementia$age_cat <- factor(ifelse(dementia$age < 30, "Early", ifelse(
  dementia$age < 60, "Midlife", "Late")),
  levels=c("Early","Midlife","Late"), ordered=TRUE)
# dummy columns
dementia$age_early <- ifelse(dementia$age_cat == "Early", 1, 0)
dementia$age_mid <- ifelse(dementia$age_cat == "Midlife", 1, 0)
dementia$age_late <- ifelse(dementia$age_cat == "Late", 1, 0)
```

```r
# Save out cleaned data, and cleaned data for adjusted GEE
write.csv(dementia, "dementia")
write.csv(dementia.adj, "dementia_adj.csv")


# Prepare data for survival analysis (one row per patient)

# For uncensored observations, `time` is the first instance of dementia.
# These are the rows where `time` equals `duration`
# (where duration is essentially the index for observations within a cluster, i.e. HHIDPN)
uncensored <-
  dementia %>%
  filter(time == duration)

# For censored observations, dementia and time are NA. So, we need to find
# the last non-NA row of dementia, and choose the corresponding value of `duration`
# to be our event of interest (i.e. censoring time)
censored <-
  dementia %>%
  group_by(HHIDPN) %>%
  filter(is.na(time), !is.na(dementia)) %>%
  slice(n()) %>%
  mutate(time=duration)

# Combine censored and uncensored patients
dementia.surv <- union(censored,uncensored)

# Make sure we have the correct number of patients; this should be TRUE
nrow(dementia.surv) == length(unique(dementia$HHIDPN))

# Save out survival data
write.csv(dementia.surv, "dementia_surv.csv")
```

Appendix B: Longitudinal data analysis

```
# unadjusted GEE
bmi.only.gee <- gee(dementia ~ bmi, id=HHIDPN,
                    family = binomial(link="logit"),
                    corstr = "independence",
                    data = dementia)
summary(bmi.only.gee)

# Multiplicative effect of bmi on odds of developing dementia
exp(summary(bmi.only.gee)$coefficients["bmi","Estimate"]) # 0.956167
# Therefore we associate a 1-unit increase in BMI with a ~4.38% reduction
# in the odds of developing dementia in the unadjusted GEE.


# Compare BMI estimate results for different covariance matrices
corstr_list <- c("independence", "exchangeable", "unstructured")#, "fixed")
bmi.effects <- data.frame("corstr" = corstr_list,
                          "bmi Estimate" = rep(NA, length(corstr_list)),
                          "Robust S.E." = rep(NA, length(corstr_list)),
                          "p-value" = rep(NA, length(corstr_list)),
                          "exp(bmi Estimate)" = rep(NA, length(corstr_list)),
                          "95% lower" = rep(NA, length(corstr_list)),
                          "95% upper" = rep(NA, length(corstr_list)),
                          check.names = FALSE)

for (corstr in corstr_list)
{
  # fit logistic GEE using given correlation structure
  bmi.only.gee <- gee(dementia ~ bmi, id=HHIDPN,
                      family = binomial(link="logit"),
                      corstr = corstr,
                      data = dementia)
  # p-value of `bmi`
  bmi.pvalue <- 2 * (1 - pnorm(abs(summary(bmi.only.gee)$coefficients["bmi","Robust z"])))
  bmi.est <- summary(bmi.only.gee)$coefficients["bmi","Estimate"]
  bmi.se <- summary(bmi.only.gee)$coefficients["bmi","Robust S.E."]
  # Store estimate, mult. effect, p-value
  bmi.effects[bmi.effects$corstr == corstr,2:5] = c(bmi.est, bmi.se, bmi.pvalue, exp(bmi.est))
  bmi.effects[bmi.effects$corstr == corstr,
              c("95% lower","95% upper")] = exp(summary(bmi.only.gee)$coefficients["bmi","Estimate"]
                                          + c(-1,1) * 1.96 * summary(bmi.only.gee)$coefficients["bmi","Robust S.E."])
}
bmi.effects

# Fully-adjusted GEE
# See which covariates are significant in determining dementia outcome
# (i.e. which variables may be confounders with BMI)
f.reg <- dementia ~ bmi + race_AA + race_other + RAHISPAN + RAGENDER + age + RAEDYRS

dementia.adj.gee <- gee(f.reg,
                        id=HHIDPN,
                        family = binomial(link="logit"),
                        corstr = "independence",
                        data = dementia.adj)
summary(dementia.adj.gee)

# p-values
gee.coef <- summary(dementia.adj.gee)$coefficients
gee.coef <- cbind(gee.coef, "p-value" = rep(0,nrow(gee.coef)))
for (coef in rownames(gee.coef))
{
  gee.coef[coef,"p-value"] = 2 * (1 - pnorm(abs(summary(dementia.adj.gee)$coefficients[coef,"Robust z"])))
}
gee.coef
```

```
# gender and Hispanic-indicator appear insignificant; drop them from model
f.reg <- dementia ~ bmi + race_AA + race_other + age + RAEDYRS

dementia.adj.gee <- gee(f.reg,
                        id=HHIDPN,
                        family = binomial(link="logit"),
                        corstr = "independence",
                        data = dementia.adj)
summary(dementia.adj.gee)
# Multiplicative effect of bmi on odds of developing dementia
exp(summary(dementia.adj.gee)$coefficients["bmi","Estimate"]) # 0.9695784

# p-values
gee.coef <- summary(dementia.adj.gee)$coefficients
gee.coef <- cbind(gee.coef, "p-value" = rep(0,nrow(gee.coef)))
for (coef in rownames(gee.coef))
{
  gee.coef[coef,"p-value"] = 2 * (1 - pnorm(abs(summary(dementia.adj.gee)$coefficients[coef,"Robust z"])))
}
gee.coef
# everything significant

# Compare BMI estimate results for different covariance matrices
corstr_list <- c("independence", "unstructured", "exchangeable")#, "fixed")
bmi.effects <- data.frame("corstr" = corstr_list,
                          "bmi Estimate" = rep(NA, length(corstr_list)),
                          "Robust S.E." = rep(NA, length(corstr_list)),
                          "exp(bmi Estimate)" = rep(NA, length(corstr_list)),
                          "p-value" = rep(NA, length(corstr_list)),
                          "95% lower" = rep(NA, length(corstr_list)),
                          "95% upper" = rep(NA, length(corstr_list)),
                          check.names = FALSE)

for (corstr in corstr_list)
{
  # fit logistic GEE using given correlation structure
  dementia.adj.gee <- gee(f.reg,
                          id=HHIDPN,
                          family = binomial(link="logit"),
                          corstr = corstr,
                          data = dementia.adj)
  # p-value of `bmi`
  bmi.pvalue <- 2 * (1 - pnorm(abs(summary(dementia.adj.gee)$coefficients["bmi","Robust z"])))
  bmi.est <- summary(dementia.adj.gee)$coefficients["bmi","Estimate"]
  bmi.se <- summary(dementia.adj.gee)$coefficients["bmi","Robust S.E."]
  # Multiplicative effect of bmi on odds of developing dementia
  bmi.effect <- exp(bmi.est)
  # Store estimate, mult. effect, p-value
  bmi.effects[bmi.effects$corstr == corstr,2:5] = c(bmi.est, bmi.se, bmi.effect, bmi.pvalue)
  bmi.effects[bmi.effects$corstr == corstr,
              c("95% lower","95% upper")] = exp(summary(dementia.adj.gee)$coefficients["bmi","Estimate"]
                                                + c(-1,1) * 1.96 * summary(dementia.adj.gee)$coefficients["bmi","Robust
S.E."])
}
bmi.effects

# Treating BMI as categorical
# unadjusted GEE w categorical BMI
bmi_cat.only.gee <- gee(dementia ~ bmi_under + bmi_over + bmi_obese,
                        id=HHIDPN,
                        family = binomial(link="logit"),
                        corstr = "independence",
                        data = dementia)
bmi_cat.coefficients <- summary(bmi_cat.only.gee)$coefficients
bmi_cat.coefficients <- cbind(bmi_cat.coefficients,
                              "p-value" = rep(0,nrow(bmi_cat.coefficients)),
                              "exp(Estimate)" = rep(0,nrow(bmi_cat.coefficients)),
                              "95% lower" = rep(NA, nrow(bmi_cat.coefficients)),
                              "95% upper" = rep(NA, nrow(bmi_cat.coefficients)))
```

```
for (coef in rownames(bmi_cat.coefficients))
{
  bmi_cat.coefficients[coef,"p-value"] = 2 * (1 - pnorm(abs(bmi_cat.coefficients[coef,"Robust z"])))
  bmi_cat.coefficients[coef,"exp(Estimate)"] = exp(bmi_cat.coefficients[coef,"Estimate"])
  bmi_cat.coefficients[coef,c("95% lower","95% upper")] = exp(bmi_cat.coefficients[coef,"Estimate"]
                                                              + c(-1,1) * 1.96 * bmi_cat.coefficients[coef,"Robust S.
E."])
}
bmi_cat.coefficients
# Categorical BMI still significant
```

Appendix C: Survival Analysis

```
# Kaplan-meier plots
par(mfrow=c(1,1))
sfit <- survfit(Surv(time=time, event=dementia) ~ bmi_cat, data=dementia.surv)
summary(sfit)
plot(sfit, conf.int=TRUE, col=1:4, ylim=c(0.6,1.01),
     main="K-M Curve for Dementia by BMI Category", xlab='Time', ylab='Survival probability')
legend(0, 0.77, c("Underweight","Healthy","Overweight","Obesity"), title="Body Mass Index (BMI)", lty=1, col=1:4, cex=.
9)

# create more clean categorical variables
dementia.surv$Gender <- factor(ifelse(dementia.surv$RAGENDER == 1, "Male", ifelse(
                                      dementia.surv$RAGENDER == 2, "Female", NA)),
                               levels=c("Male","Female"))
dementia.surv$Race <- factor(ifelse(dementia.surv$RARACEM == 1, "White", ifelse(
  dementia.surv$RARACEM == 2, "Black", ifelse(
    dementia.surv$RARACEM == 3, "Other", NA))))
for (cat_col in c("Gender","Race"))
{
  par(mfrow=c(2,2), oma=c(0,0,2,0))
  for (level in levels(dementia.surv$bmi_cat))
  {
    barplot(table(dementia.surv[(dementia.surv$bmi_cat == level), cat_col]),
            main=paste(level,"BMI"))
  }
  mtext(paste("Distribution of", cat_col, "by BMI Category"), side=3, line=0, outer=TRUE)
}

for (num_col in c("RAEDYRS", "age"))
{
  par(mfrow=c(2,2), oma=c(0,0,2,0))
  for (level in levels(dementia.surv$bmi_cat))
  {
    hist(dementia.surv[(dementia.surv$bmi_cat == level),] %>% pull(num_col),
         xlab=num_col, nclass=25, main=paste(level,"BMI"))
  }
  mtext(paste("Distribution of", num_col, "by BMI Category"), side=3, line=0, outer=TRUE)
}


# Log-rank test
# 4 groups, based on the BMI categories
survdiff(Surv(time=time, event=dementia) ~ bmi_cat, data=dementia.surv)

# Cox proportional hazards model
dementia.cox <- coxph(Surv(time=time, event=dementia) ~ bmi_under + bmi_over + bmi_obese,
                      robust=TRUE, data=dementia.surv)
summary(dementia.cox)


# Repeat survival analysis using bmi_base_cat
par(mfrow=c(1,1))
sfit <- survfit(Surv(time=time, event=dementia) ~ bmi_base_cat, data=dementia.surv)
summary(sfit)
plot(sfit, conf.int=TRUE, col=1:4, ylim=c(0.6,1.01),
     main="K-M Curve for Dementia by Baseline BMI Category", xlab='Time', ylab='Survival probability')
legend(0, 0.77, c("Underweight","Healthy","Overweight","Obesity"), title="Body Mass Index (BMI)", lty=1, col=1:4, cex=.
9)

survdiff(Surv(time=time, event=dementia) ~ bmi_base_cat, data=dementia.surv)

# Use baseline bmi instead
dementia.cox.base <- coxph(Surv(time=time, event=dementia) ~ bmi_base_under + bmi_base_over + bmi_base_obese,
                           robust=TRUE, data=dementia.surv)
summary(dementia.cox.base)
```

```
# Use log-rank tests to compare patients whose BMI does not change between baseline/event
survdiff(Surv(time=time, event=dementia) ~ bmi_base_under + bmi_under,
         data=dementia.surv[dementia.surv$bmi_under==1,])

survdiff(Surv(time=time, event=dementia) ~ bmi_base_healthy + bmi_healthy,
         data=dementia.surv[dementia.surv$bmi_healthy == 1,])

survdiff(Surv(time=time, event=dementia) ~ bmi_base_over + bmi_over,
         data=dementia.surv[dementia.surv$bmi_over == 1,])

survdiff(Surv(time=time, event=dementia) ~ bmi_base_obese + bmi_obese,
         data=dementia.surv[dementia.surv$bmi_obese == 1,])
```