

HOMEWORK 4
BIOSTATISTICS 755
DUE APRIL 3RD, 2025

1. **(50 points)** The Study of Assets and Health Dynamics among the Oldest Old (AHEAD) is a national longitudinal study with initial sample of 7444 respondents aged 70 years and older, and their spouses (if married). Objectives of the study include: (i) to monitor transitions in physical, functional, and cognitive health; (ii) to examine the relationship of late-life changes in physical and cognitive health to patterns of saving and income flows; (iii) to relate changes in health to economic resources and intergenerational transfers; (iv) to examine how the mix and distribution of economic, family, and program resources affect key outcomes, including institutionalization, saving, and health declines. In this problem, we look at data from the first four waves of the study, collected in 1993, 1995, 1998 and 2000.

A reduced data set `ahead.xlsx` is on the course website. We explore the relationship of a test of cognitive function to some physical functioning indicators. The hypothesis is that, as physical function declines, movement becomes more difficult, and hence there are fewer stimuli, leading to cognitive decline. Here is a description of the variables contained:

```
id- subject identification number
year- study wave
sex- 1=male, 2=female
age- age in years
immword- immediate word recall
delword- delayed word recall
blks- difficulty walking several blocks without help
strs- difficulty climbing flight of stairs without help
push- difficulty pulling or pushing a living room chair
bag- difficulty lifting a bag of groceries
dime- difficulty picking up a dime
iadlany- perform any instrumental activities of daily living
```

The focus of the study is to assess the association of difficulty performing instrumental activities of daily living (IADLs) to memory and cognition problems. IADLs include activities such as shopping, managing money, taking medications, using the telephone, preparing meals. They are supposed to represent the types of integrative activities that one needs to be able to perform in order to live independently in the community.

Use total word recall (`totword = immword + delword`) as the main variable of interest and `iadlany`, an indicator for whether or not the subject reports difficulty with any (at least one) IADL, as the outcome variable. The goal of this analysis is to determine if there is a relationship between total word recall and the IADL variable using GEE. Consider the variables `age` and `sex` as possible confounders and/or effect modifiers for this analysis.

- (a) **(5 points)** Discuss whether you feel that GEE is the appropriate analysis choice for this project.
- (b) **(45 points)** Complete an analysis of this data using a GEE that results in the best possible answer to the the scientific aims of the study. This should start with exploratory analysis of the data, explore different statistical models based on model fit and scientific hypotheses, and finish with a clear summarization of your findings including the interpretation of the coefficients of interest. Please include only relevant output to how you found the best model. It is encouraged to give the code to replicate your findings.
2. **(50 points)** The Skin Cancer Prevention Study was a randomized, double-blind, placebo-controlled clinical trial of beta carotene to prevent non-melanoma skin cancer in high-risk subjects. A total of 1805 subjects were randomized to either placebo or 50mg of beta-carotene per day for 5 years. Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The main objective of the analyses is to compare the effect of beta carotene on skin cancer rates.
- The data “skin.xlsx” is available on github. This file contains a description of the data. Briefly, the outcome variable (Y) is a count of the number of new skin cancers per year. The categorical variable “Treatment” is coded 1 = beta-carotene, 0 = placebo. The variable “Year” denotes the year of follow-up. The categorical variable “Gender” is coded 1 = male, 0 = female. The categorical variable “Skin” denotes skin type and is coded 1 = burns, 0=otherwise. The variable “Exposure” is a count of the number of previous skin cancers. The variable “Age” is the age (in years) of each subject at randomization. Complete data are available on 1683 subjects comprising a total of 7081 measurements.
- (a) **(10 points)** Consider a Poisson-generalized linear mixed model with random intercepts for the subject-specific log rate of skin cancers ($\log E(Y_{ij}|b_i)$) with time, treatment, and a time by treatment interaction. Find the best way to include time and the time by treatment interaction.
- (b) **(3 points)** Regardless of significance, give an interpretation for time by treatment interaction.
- (c) **(4 points)** From these results, what **conclusions** do you draw about the effect of beta carotene on skin cancers and why?
- (d) **(10 points)** Repeat the random effect model analysis in (a) while appropriately adjusting for skin type and age. What conclusions do you draw about effect of beta carotene on skin cancers and why?
- (e) **(3 points)** Does adjusting for skin type and age fit the data better? Give some quantification as to how you made your choice.

For the rest of the questions we will slightly change the outcome variable. A colleague recommends that the Y variable be modeled as a 0/1 outcome

instead of a count. That is, create a new Y variable, say Y^* , where $Y^* = \min(Y, 1)$.

- (f) **(10 points)** Rerun the model in (a) using Y^* with and a logistic instead of a Poisson model. Find the best way to include time and the time by treatment interaction.
- (g) **(3 points)** Regardless of significance, give an interpretation for time by treatment interaction.
- (h) **(3 points)** From these results, what **conclusions** do you draw about the effect of beta carotene on skin cancers and why?
- (i) **(4 points)** Comment on the differences between the logistic and Poisson GLMM approaches. Which approach would you recommend? Don't use quantitative measures, discuss this qualitatively.