HOMEWORK 5
BIOSTATISTICS 755
DUE MARCH 31ST, 2023

1. 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) 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) Regardless of significance, give an interpretation for the impact of time on the outcome.

   (c) Regardless of significance, give an interpretation for time by treatment interaction.

   (d) From these results, what **conclusions** do you draw about the effect of beta caroteneon skin cancers and why?

   (e) Carry out a marginal model analysis using GEE, using the same covariates as in (a). Fit this model with exchangeable and unstructured covariance matrices, then choose the best according to QIC.

   (f) Regardless of significance, give an interpretation for the impact of time on the outcome.

   (g) Regardless of significance, give an interpretation for time by treatment interaction.

   (h) For the GEE analysis, what do you **conclude** about the effect of beta carotene on skin cancers and why? Comment on the difference (or lack of) with the random effect model.

   (i) Which approach do you think is more appropriate, the GEE or GLMM?

(j) 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?

(k) Does the model in (a) or (j) 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$(Y, 1)$.

(l) 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.

(m) Regardless of significance, give an interpretation for the impact of time on the outcome.

(n) Regardless of significance, give an interpretation for time by treatment interaction.

(o) From these results, what **conclusions** do you draw about the effect of beta caroteneon skin cancers and why?

(p) Comment on the differences between the logistic and Poisson GLMM approaches. Which approach would you recommend?