`

Data Analysis 1

The data set “skindata” is on the Canvas site. 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 the skin susceptibility and is coded 1=burns easily, 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.

Variable List: ID, Center, Age, Skin, Gender, Exposure, Y, Treatment, Year.

Reference: Greenberg, E.R., Baron, et. Al. (1990). A clinical trial of beta carotene to prevent basal-cell and squamous-cell cancers of the skin. *New England Journal of Medicine*, 323, 789-795.

1. For these data, an “intention to treat” (ITT) analysis which only looks at Treatment as a factor, while adjusting for clustering and longitudinal structure is conventional
   - a. Fit a generalized linear mixed model (glmer), Poisson family, with Y as an outcome, a log link function, an effect for Treatment, Year as a continuous variable, and a random intercept for the individual. Write equations to specify this model and state the assumptions.
   - b. Give the estimated variance of the random effects and show a histogram of the estimated random effects empirical Bayes estimates. With glmer, EB estimation~~es~~ requires the package merTools. The function REextract() extracts the EB estimates. Does the model seem reasonable?
   - c. Using the same data, fit generalized estimating equation models with a Poisson family and log link and compound symmetry (“exchangeable”) working correlation matrices using the same fixed effects as in 1a.
   - d. Compare the treatment effect estimates with the generalized linear mixed model, and discuss any differences in interpretation.
2. Add the “Exposure” variable to the models above.
   - a. Evaluate the strength of the association of this variable with the outcome.
   - b. Generate missing data indicators for each observation where the missingness probabilities depends on the Exposure variable being above or below a threshold. Apply this to the outcome Y (eg. make it NA), and refit the model. At least 20% of the outcomes should be missing. Compare the missing data rates in each arm. Comment of the difference in treatment effect estimates and confidence intervals for the estimates when applying the ITT analyses to the new dataset with and without using the Exposure ~~analysis~~variable.
   - ~~c. Now repeat this, but have different Exposure thresholds for each treatment group. The difference in the rates between the arms should be at least 15%. Compare the ITT estimates to the analysis in b.~~
   - ~~d. Now add the Exposure variable as a covariate to both the b and c analyses.~~
   - ~~e.~~c. Comment on the need for adjustments for Exposure in this randomized study.

Data Analysis 2

The dataset reports survival (or censoring) time in years. Individuals with an event (death) are coded as 1, and censoring is coded as 0. The independent variable of primary interest is Treatment (1 = invasive surgery, 0 = less invasive procedure). Other covariates are Female (biological sex is 1 if female, o otherwise), Age, am ordinal Disease Score from 1 to 5 and a Biomarker for which higher is meant to meant worse prognosis.

1. a. Describe the distribution of the biomarker.
   b. Report the Pearson and Spearman correlations of the Biomarker age, disease score and biological sex (or the latter do a two-sample t-test).
   c. Develop a model for the association of the biomarker age, disease score and sex. Comment on the findings.
   d. Using the model above how does a unit increase in the disease score affect the biomarker.
   e. For the multivariable model in part c which yields better fit, the log-transformed biomarker or non-transformed.
2. Plot Kaplan-Meier survival curves stratified by Treatment group.
3. Is there a significant difference in survival between the two treatment groups.
4. Run a multivariable Cox P.H. model for how the variables Female, Age, Disease Score and Biomarker affect survival.
5. Add Treatment to this model and report the hazard ratio with 95%CI comparing the invasive to less invasive procedure adjusted for the variables in part 4. Is it a statistically significant effect.
6. Test the proportionality of hazards assumption for each variable in the multivariable model and comment.
7. Plot the Schoenfeld residuals corresponding to treatment and their smoother as a function of time.
8. Report the hazard ratio for treatment adjusted for sex, age, disease score and the biomarker for the following time windows, a. < 0.25 years, b. 0.25 to < 1 year and c. 1 year and above.
9. a. Derive a propensity score for Treatment based on sex, age, disease score and biomarker and calculate IWP (inverse weighted propensities).
   b. What covariates influence treatment selection ?
   c. Plot Kaplan-Meiers for the two treatment groups weighted by IWP.
   d. Calculate the hazard ratio for treatment weighted by IWP.
10. Derive the doubly robust estimator of the hazard ratio for treatment by combining a multivariable Cox model with weighting by IWP.