Bios 765: Final take-home problem set, due Nov 18, 2020

Honor Pledge

I have not discussed the problems on this final problem set with anyone; I have not had any communications including email regarding the problems in this final set with anyone. I have not shared my work on this final problem set with anyone. I have not received help on this final problem set nor have I given help.

Name (please print):

Signature:

1. **(Applied, 15 pts) Weighted least squares and logistic regression: Table A** displays a 4-way cross-classification of data related to complaints of symptoms of a respiratory disease, byssinosis, which occurs among textile mill workers. With complaints as the response variable, assume a simple stratified random sampling framework with strata defined by workplace conditions, years employment and smoking.

Table A. Frequency table of byssinosis Complaints

WORKPLACE CONDITIONS	YEARS EMPLOYMENT	SMOKING	COMPLAINTS yes	no
Dusty	<10	yes	30	203
Dusty	<10	no	7	119
Dusty	>=10	yes	57	161
Dusty	>=10	no	11	81
Not Dusty	<10	yes	14	1340
Not Dusty	<10	no	12	1004
Not Dusty	>=10	yes	24	1360
Not Dusty	>=10	no	10	986

a. Write, using appropriate notation, the main effects linear model for the empirical logit functions of the presence of complaints. Clearly define your explanatory variables, using effect coding (1/-1).

b. Report the weighted least squares estimates of the regression parameters and their standard errors.

c. Provide a goodness-of-fit statistic for the model fit in part b). Based on $\alpha = 0.05$ significant level, is this model fit adequate?

d. Next, write, using appropriate notation, the main effects logistic regression model for the probability of the presence of complaints. Clearly define your explanatory variables, using effect coding (1/-1).

e. Report the maximum likelihood estimates of the regression parameters and their standard errors for the model in part d) and report the results side-by-side those from part b.

f. Provide a goodness-of-fit statistic for the model fit in part e). Based on $\alpha = 0.05$ significant level, is this model fit adequate?

g. Comment on the similarity or dissimilarity of the estimates from parts b) and e). Interpret the result from one of the models, declaring the model you are interpreting, in words that a non-statistician would understand. Specifically, report 95% confidence intervals for the effects of explanatory variables on the presence of complaints.

2. **Applied (10 pts.) Generalized Estimating Equations.**

Consider longitudinal data on the number of occurrences of green tobacco sickness or GTS for each of 182 migrant farmworkers who worked in tobacco. The data set tobacco2nomiss.dat can be downloaded from the course website. Each farm worker (identified by ID) was interviewed from 1 to 5 times (or periods) and asked about the number of days they got sick within the previous week (NUMSICK) and the number of days they actually worked in tobacco during that week (DAYSRISK).

(a) For each period, t=1,2,3,4,5, determine the incidence density of disease as ID_t = (total number of events)/(total number of days at risk), where the numerator and denominator are calculated across all individuals.

(b) Define a population-averaged longitudinal Poisson regression model for the log incidence density of GTS including the following covariates as main effects: time (use dummy variables for MID and LATE season with early season as the reference; early season refers to period=1 or period=2, mid season refers to period=3, and late season refers to period=4 or period=5); YRSWORK = number of years worked in tobacco; and MEDICINE = takes medicine to prevent getting sick (1=yes, 0=no). Clearly define the notation for all terms in your model.

(c) Fit the model in part b) under both working exchangeable and first order auto-regressive correlation matrices using the generalized estimating equations procedure (hint: don't forget the model offset). In each case, report the estimated regression coefficients along with their model-based (assuming ϕ=1) and empirical sandwich standard errors. Comment on any disparities between the model-based and empirical standard error estimates.

(d) Using the results from part c.), calculate the incidence density ratio and corresponding 95% confidence interval (using empirical standard errors and AR-1 correlation structure) for each independent variable in the model and summarize your conclusions.

3. **(Applied, 15 pts)** Apply **Mantel-Haenszel** hypothesis tests and **nonparametric analysis of covariance** to the outcome 'unus' from the sensory retraining clinical trial, where unus describes the extent of unusual feelings six months after jaw surgery. Download the data file h128un.dat from the class web site [Ignore the variable SITE in this analysis]. The ordinal response variable unus has seven categories, where unus=1 means the subject has no problem with unusual feelings in the face and unus=7 means the subject has the most severe problems with unusual feelings. Treat the outcome unus as an integer score ("table score" in PROC FREQ). Hint: For parts (d) and (e), **download** and use the SAS **macro NParCov4** for nonparametric analysis of covariance from the course website and refer to **Chapter 2 "Advanced Randomization Methods"** by Zink *et al.,* (2017) as needed.

a. Conduct statistical tests with minimal assumptions (randomization only) to assess the association of gender, genioplasty, and number of jaws with unusual feelings. Interpret the results of each test, examining each of the three factors separately.

b. Similarly, conduct a statistical test with minimal assumptions to assess the association of age at surgery with unusual feelings. Interpret the results of the test. Calculate a statistic that summarizes the strength of the association between age at surgery and the extent of problems with unusual feelings at six months follow-up.

c. Conduct a statistical test with minimal assumptions to assess the association of exercise group with unusual feelings. Interpret the results of the test.

d. Using the NParCov4 software, test the hypotheses that the mean score for unusual feelings at six months post-surgery does not differ between exercise groups. First provide the unadjusted test results (not adjusting for any covariates), then provide the test result adjusting for age, gender, genioplasty and number of jaws operated on. Report the relevant Chi-square statistics and pvalues. Interpret results, explaining the minimal assumptions required for this method. Comment on any similarities or dissimilarities with the test results in part (c).

e. Using the NParCov4 software, estimate the difference in mean scores for unusual feelings at six months follow-up post-surgery. First provide an estimate and 95% confidence interval for the mean difference not adjusting for covariables. Second, provide an estimate and 95% confidence interval of the mean difference adjusting for age, gender, genioplasty and number of jaws operated on. Was variance reduction achieved through covariate adjustment? Interpret your results, explaining assumptions regarding sampling underlying your approach.

4. **(Theory, 10 pts)** Consider the nonparametric comparison between two randomized groups for a univariate response with adjustment for a single covariable within one stratum. The set-up is:

Group	Sample size	Mean response	Mean covariables
1	n_1	$\bar{y}_1$	$\bar{x}_1$
2	n_2	$\bar{y}_2$	$\bar{x}_2$
Difference		$d = (\bar{y}_1 - \bar{y}_2)$	$u = (\bar{x}_1 - \bar{x}_2)$

where interest is in the linear model $\mathrm{E}\begin{bmatrix} d \\ u \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}\gamma = \mathbf{z}\gamma$, which would be fit by weighted least squares.

a. Assuming only randomization, derive an expression for $\mathrm{Var}\begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$ under the hypothesis $H_0 : \gamma = 0$. (Hint: apply results from Appendix 6 of Notes 13).

b. Based on the results in part (a), derive the formula for $\mathbf{V}_0 = \mathrm{Var}\begin{bmatrix} d \\ u \end{bmatrix}$ whose general form (for multiple covariables) is given on page 11 of notes 19.