ST544, Spring 2020

Final, due by 5:00pm on 5/5/2020

Name: **Student ID:** **Signature:**

Honor Pledge: "By signing my name I swear that I have neither given nor received unauthorized aid in any form on this exam."

Some quantiles from the standard normal distribution:

	$\alpha = 0.01$	$\alpha = 0.0125$	$\alpha = 0.025$	$\alpha = 0.05$	$\alpha = 0.10$
z_α	2.326	2.241	1.960	1.645	1.282

Some critical values from χ^2 distributions

df	1	2	3	4	5	6	7	8
$\chi^2_{0.05,df}$	3.841	5.991	7.815	9.488	11.070	12.592	14.067	15.507
$\chi^2_{0.025,df}$	5.024	7.378	9.348	11.143	12.833	14.449	16.013	17.535
$\chi^2_{0.01,df}$	6.635	9.210	11.345	13.277	15.086	16.812	18.475	20.090

Instruction: This is a take-home exam. It is open-book and open-note. However, you are expected to work independently. When you are asked to fit a model to a data set, please provide your SAS code and relevant output to justify your answer. The exam is due by 5:00pm on 5/5. Please make your exam one single file (with the format lastname-st544-final.pdf) and submit it on moodle or email it to me.

1. (50 tps) Data set `findata1.txt` contains longitudinal data from 150 patients with certain disease randomized to a new and the standard treatments. It has four variables `id, y, trt, time`, where `id` is the patient ID, `y` = 1/0 is the binary outcome indicating whether or not the disease is under control at `time`, time in months (1,2,3) from treatment initiation, and `trt` is the treatment indicator (1=new treatment, 0=standard treatment). Do the following:

   (a) (10 pts) Let $\pi(trt, time) = P[Y = 1|trt, time]$. Fit the following model using the GEE method with the unstructured working correlation matrix for Y:

   $$\text{logit}\{\pi(trt, time)\} = \beta_0 + \beta_1 trt + \beta_2 time + \beta_3 trt \times time.$$

   Report the estimates of β's.

   (b) (10 pts) From the fitted model, estimate the following odds-ratio at time=1,2,3 months between the new and standard treatments:

   $$\theta(time) = \frac{\pi(trt = 1, time)/\{1 - \pi(trt = 1, time)\}}{\pi(trt = 0, time)/\{1 - \pi(trt = 0, time)\}}.$$

   Describe the treatment effects based on these odds-ratio estimates.

   (c) (5 pts) The estimated correlation matrix indicates that the underlying correlation matrix is not exchangeable. However, if you use an exchangeable working correlation matrix, will you get invalid inference? Why?

   (d) (15 pts) Define $\pi_i(trt, time) = P[Y_i = 1|trt, time, b_i]$ for subject i and fit the following generalized linear mixed model

   $$\text{logit}\{\pi_i(trt, time)\} = \beta_0 + b_i + \beta_1 trt + \beta_2 time + \beta_3 trt \times time$$

   where $b_i \sim N(0, \sigma^2)$. Interpret the time effect (odds-ratio of disease being under control with one month increase of time) for subject i if he/she was assigned to the standard treatment, or if he/she was assigned to the new treatment. Which treatment is better?

   (e) (10 pts) What is the estimate of σ^2 from the previous model? Interpret this estimate. Explain why the estimates of β's from the above two models are very different or close.

2. (50 pts) In a matched case-control study to investigate the association of smoking and lung cancer, each of 112 lung cancer patients was matched (with respect to age and gender) with a control (free of lung cancer) and their smoking status was ascertained. The data is presented in the following:

		Smoking status for cases	
		Yes (1)	No (0)
Smoking status for controls	Yes (1)	26	20
	No (0)	44	22

Denote $Y = 1/0$ for lung cancer/control, $X = 1/0$ for smoking/non-smoking, Z =matched pair. Do the following:

(a) (10 pts) Conduct the CMH test for $H_0 : X \perp Y | Z$ at significance level 0.05.

(b) (10 pts) For the above table, conduct the McNemar's test at the 0.05 level. What is the null hypothesis in this case?

(c) (10 pts) Ignore the matching, collapse the data across pairs and get the following table:

		Lung cancer	
		Yes (1)	No (0)
Smoking status	Yes (1)	70	46
	No (0)	42	66

Treat the data as if it is from a multinomial sampling, find an estimate of θ_{XY} and a 95% CI for θ_{XY}.

(d) (5 pts) Argue why the previous estimate and CI may not be valid.

(e) (15 pts) Let $\pi(x, k) = P(Y = 1|x, pair = k)$ for $k = 1, 2, \cdots, 112$ and assume the following model

$$\text{logit}\{\pi(x, k)\} = \beta_k + \beta x.$$

Estimate e^β and its 95% CI estimate using the conditional logistic regression approach. Interpret the estimate of e^β. Test $H_0 : X \perp Y | Z$ at significance level 0.05 using the conditional logistic regression approach. Is it consistent with the CMH test?

3. (50 pts) The following table presents results from four top tennis players in one year:

	Loser			
Winner	A	B	C	D
A	.	1	7	8
B	8	.	3	9
C	3	2	.	0
D	2	3	2	.

Do the following:

(a) (10 pts) Fit the Bradley-Terry model to the above data. Report the estimates of parameters in the model. Does the model fit the data well? Show the calculation of *df* for the deviance.

(b) (10 pts) Show that the Bradley-Terry model implies that if player i is better than player j, player j is better than player k, then player i is better than player k (here "player i is better than player j" means that player i wins player j with a probability > 0.5, etc.)

(c) (5 pts) Rank the players according to their winning probabilities.

(d) (10 pts) Find a 95% CI for the winning probability of player A against player D.

(e) (15 pts) Find a 95% CI for the winning probability of player A against player C.

4. (50 pts) Consider the data from 120 subjects in placebo group of the insomnia study (page 285 of the textbook):

		Y_2			
		< 20	$20-30$	$30-60$	> 60
	< 20	7	4	2	1
Y_1	$20-30$	14	5	1	0
	$30-60$	6	9	18	2
	> 60	4	11	14	22

Do the following:

(a) (10 pts) Ignore the ordinal scale of the table, test marginal homogeneity of the underlying probability table at the significance level 0.05.

(b) (10 pts) Repeat (a) by taking into account the ordinal scale.

(c) (10 pts) Using score 10, 25, 45, 75 (intended for the actual *time to falling asleep*) for the 4 four categories, test the null hypothesis that Y_1 and Y_2 have the same mean score at the significance level 0.05. What is the average increase/decrease in *time to falling asleep*?

(d) (10 pts) Conduct the Pearson χ^2 test for a symmetric underlying probability table.

(e) (10 pts) Suppose we are interested in the probability that the *time to falling asleep* two weeks later is shorter than that at the baseline, i.e., $\pi = P[Y_2 < Y_1]$. Estimate π and construct a 95% CI for π.