ST544, Fall 2020

Midterm Exam 2, due by 5:00pm on 4/4/2020

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 4/4. Please make your exam one single file (with the format lastname-st544-exam2.pdf) and submit it on moodle or email it to me.

1. (50 pts) A small clinical trial was conducted in 4 randomly selected clinics to compare a new treatment to an old treatment for patients with some disease. The data was presented in the following table

	Center 1		Center 2		Center 3		Center 4	
Treatment	S	F	S	F	S	F	S	F
New	4	6	6	6	8	6	8	4
Old	2	8	4	12	5	5	4	8

where "S" means the result is a succeess and "F" means the result is a failure.

Denote by $X = 1/0$ for the new/old treatment, $Z = 1, 2, 2, 4$ for 4 centers, and $Y = 1/0$ for S/F. Let $\pi(x, z) = P(Y = 1|x, z)$. Consider the following model for $\pi(x, z)$:

$$\text{logit}\{\pi(x, z = k)\} = \beta x + \beta_k^Z, \quad k = 1, 2, 3, 4.$$

Do the following:

(a) (10 pts) Show that the above model implies common odds-ratio ($\theta_{XY|Z}$) between X and Y across $Z = 1, 2, 3, 4$ centers.

Solution: From the model, we have for any $k = 1, 2, 3, 4$

$$\text{logit}\{\pi(x = 1, z = k)\} = \beta + \beta_k^Z, \quad \text{logit}\{\pi(x = 0, z = k)\} = \beta_k^Z.$$

Therefore,

$$\theta_{XY|Z} = \frac{\pi(x = 1, z = k)/\{1 - \pi(x = 1, z = k)\}}{\pi(x = 0, z = k)/\{1 - \pi(x = 0, z = k)\}} = e^{\beta},$$

free of $k = 1, 2, 3, 4$.

(b) (10 pts) Fit the above model to the data using ML approach, report the estimates of β, β_k^Z, interprete $e^{\widehat{\beta}}$, and find a 95% LR CI for e^{β}.

Solution: SAS program fitting the above model using ML method:

```
data prob1;
input center x y count @@;
datalines;
1 1 1 4 1 1 0 6
1 0 1 2 1 0 0 8
2 1 1 6 2 1 0 6
2 0 1 4 2 0 0 12
3 1 1 8 3 1 0 6
3 0 1 5 3 0 0 5
4 1 1 8 4 1 0 4
4 0 1 4 4 0 0 8
;

title "Problem 1(b)";
proc genmod descending;
freq count;
class center;
model y = center x / noint dist=bin link=logit aggregate lrci;
run;
```

MLE of β, β_k^Z: $\widehat{\beta} = 0.9340$, $\widehat{\beta_1^Z} = -1.3582$, $\widehat{\beta_2^Z} = -1.0172$, $\widehat{\beta_3^Z} = -0.3715$, $\widehat{\beta_4^Z} = -0.4670$.

$e^{\widehat{\beta}} = e^{0.9340} = 2.545$, meaning that in any clinic, the odds of success of the new treatment is 2.545 times the odds of success of the old treatment.

The 95% LR CI for β: [0.0935, 1.8039]. So the 95% CI for e^{β}: [1.1, 6.07]. (**Note**: you are aked to find a LR CI)

(c) (10 pts) Conduct the Cochran-Mantel-Haenszel test (by hand) for $H_0 : X \perp Y|Z$ at the significance level 0.05 (No need to do correction).

Solution: The CMH χ^2 is:

$$\chi^2_{CMH} = \frac{5.130952^2}{5.814305} = 4.5279 > 3.841 = \chi^2_{1,0.05},$$

so we reject $H_0 : X \perp Y|Z$ at level 0.05.

(d) (5 pts) Does the above model fit the data adequately?

Solution: The GOF test statistic: Deviance = 0.9201 with $df = 3$, $p - value = P[\chi^2_3 > 0.9201] = 0.82$, indicating that the model fits the data adequately.

(e) (5 pts) Conduct the exact Cochran-Mantel-Haenszel test for $H_0 : X \perp Y|Z$ at the significance level 0.05.

Solution:
Solution: SAS program conducting the exact CMH test:

```
title "Problem 1(e)";
proc logistic descending;
class center / param=ref;
freq count;
model y=center x / aggregate scale=none;
exact x;
run;
```

The above program produced exact p-value = 0.0388 and exact mid p-value = 0.0301. So we reject $H_0 : X \perp Y|Z$ at level 0.05 using the exact CMH test.

(f) (5 pts) Fit a conditional logistic model to the data by removing the nuisance parameters β_k^Z's. Report the estimate of β from this conditional fit. Based on this model, test $H_0 : X \perp Y|Z$ at the significance level 0.05.

Solution: SAS program for the conditional approach for fitting the model:

```
title "Problem 1(f)";
proc logistic descending;
freq count;
model y = x;
strata center;
run;
```

The conditional MLE of β: $\widehat{\beta} = 0.8939$. The conditional approach gives 3 tests for $\beta = 0$ ($H_0 : X \perp Y|Z$): LRT $G^2 = 4.5515$, Score $\chi^2 = 4.5279$, Wald $\chi^2 = 4.4377$ with $df = 1$; all are significant at level 0.05.

(g) (5 pts) Suppose the total number of centers (K) is large and data in each center is sparse. What test/method would you use to test $H_0 : X \perp Y|Z$?

Solution: CMH test, exact CMH test (may be computationally very expensive when K is very large) and conditional logistic regression approach are all valid to test $H_0 : X \perp Y|Z$. Of course, the conditional logistic approach assumes the model given at the beginning has to be a good model.

2. (10 pts) After fitting a logistic regression model to a small data set, we got the following estimated success probabilities

Y	1	1	1	0	0	0	0
$\widehat{\pi}$	0.8	0.6	0.4	0.7	0.5	0.4	0.3

Construct the ROC curve for this logistic model. Find the area under the ROC curve and interpret the value.

Solution: The ROC curve looks like

The area under the ROC curve is $\frac{8.5}{12}$, which can be interpreted as the proportion of concordant pairs of $(Y, \widehat{\pi})$ among all pairs with different outcomes.

3. (40 pts) In a clinical trial to evaluate a treatment on curing a disease, we got following data:

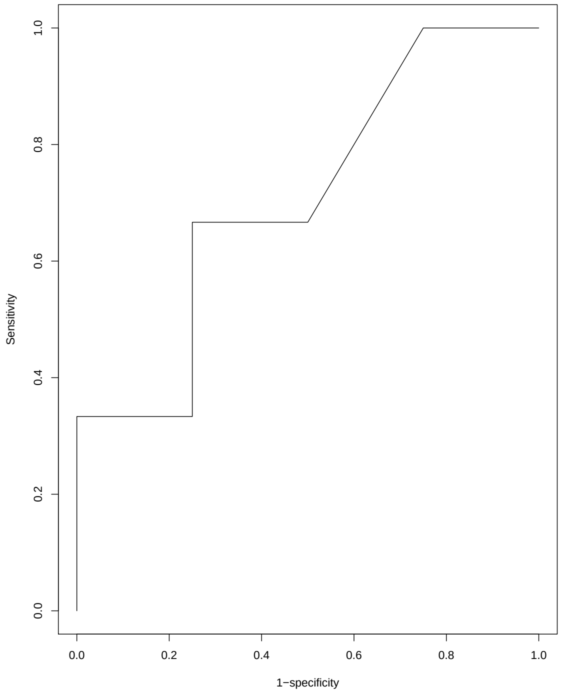
		Y		
Gender	X	Complete recovery (1)	Partial recovery (2)	No recovery (3)
Male	Treatment	22	10	8
Male	Placebo	10	8	12
Female	Treatment	24	8	6
Female	Placebo	12	10	8

Do the following:

(a) (10 pts) Fit a cumulative logit model with main effects of treatment and gender to the above data, write down the fitted model.

Solution: The SAS program:

ROC curve for the fitted model



```
data prob3;
input gender trt y count @@;
datalines;
1 1 1 22
1 1 2 10
1 1 3 8
1 0 1 10
1 0 2 8
1 0 3 12
0 1 1 24
0 1 2 8
0 1 3 6
0 0 1 12
0 0 2 10
0 0 3 8
;

title "cumulative logit model";
proc logistic;
freq count;
model y =  gender trt / link=clogit aggregate scale=none;
run;
```

The fitted model is

$$\begin{aligned}\text{logit}\{\widehat{\tau}_1(gender, trt)\} &= -0.3471 - 0.3781 gender + 0.8978 trt,\\ \text{logit}\{\widehat{\tau}_2(gender, trt)\} &= 0.8607 - 0.3781 gender + 0.8978 trt.\end{aligned}$$

(b) (5 pts) From the fitted model, find the estimate and a 95% CI of the odds-ratio of complete recovery between the treatment and placebo for patients with the same gender.

Solution: The estimate of the required odds-ratio is $e^{0.8978} = 2.45$. A 95% CI is $[e^{0.8978-1.96\times 0.3296}, e^{0.8978+1.96\times 0.3296}] = [1.29, 4.68]$.

(c) (5 pts) Find the deviance of this model and show the calculation of the degrees of freedom. Does the model fit the data well?

Solution: The deviance is 0.4599, with $df = 4$. The df is calculated as $(I-1)(J-1) -$ # of x's $= (4-1)(3-1) - 2 = 4$. The p-value $= P[\chi^2_4 > 0.4599] = 0.9773$, indicating a very good fit.

(d) (5 pts) What is the score statistic for testing goodness of fit of this model? Show the calculation of the degrees of freedom. What is the alternative model in this score test? Does this test show adequate fit of the model to the data?

Solution: The score $\chi^2 = 0.1941$ with $df = 2$. The df is calculated as $(J-2)dim(x) = (3-2)2 = 2$. The p-value$=P[\chi^2_2 > 0.1941] = 0.9075$, indicating a very good fit of the model to the data.

The alternative model in this score test is

$$\text{logit}\{\tau_j(gender, trt)\} = \alpha_j + \beta_{1j} gender + \beta_{2j} trt, \quad j = 1, 2.$$

(e) (5 pts) Estimate the 3 cell probabilities for male patients receiving the treatment.

Solution: The estimated cumulative logits are:

$$\begin{aligned}\text{logit}(\widehat{\tau}_1) &= -0.3471 - 0.3781 + 0.8978 = 0.1726,\\ \text{logit}(\widehat{\tau}_2) &= 0.8607 - 0.3781 + 0.8978 = 1.3804,\end{aligned}$$

implying $\widehat{\tau}_1 = e^{0.1726}/(1+e^{0.1726}) = 0.5430$, $\widehat{\tau}_2 = e^{1.3804}/(1+e^{1.3804}) = 0.7991$. Therefore, the 3 cell probability estimates are $\widehat{\pi}_1 = 0.5430, \widehat{\pi}_2 = 0.7991 - 0.5430 = 0.2561, \widehat{\pi}_3 = 1 - 0.7991 = 0.2009$.

(f) (10 pts) Fit a baseline category model to the above data with main effects of treatment and gender. What is the deviance and its df. Show the calculation of its df. Does this model fit the data well? Find the 3 cell probabilities for male patients receiving the treatment and compare them to the above 3 probabilities.

Solution: SAS program for fitting a baseline category logit model to the above data:

```
title "baseline category logit model";
proc logistic;
freq count;
model y =  gender trt / link=glogit aggregate scale=none;
run;
```

The deviance of this model is 0.3495, with $df = 2$. The calculation of df is $df = (J-1)(I - 1 - \#\text{ of x's}) = (3-1)(4-1-2) = 2$. The p-value $= P[\chi^2_2 > 0.3495] = 0.8397$, indicating a very good fit.

From the fitted model, we have for male receiving the treatment:

$$\begin{aligned}\log(\widehat{\pi}_1/\widehat{\pi}_3) &= 0.3569 - 0.5123 + 1.1121 = 0.9567\\ \log(\widehat{\pi}_2/\widehat{\pi}_3) &= 0.0899 - 0.3695 + 0.3695 = 0.0899.\end{aligned}$$

Solving this system, we got $\widehat{\pi}_1 = 0.5542, \widehat{\pi}_2 = 0.2329, \widehat{\pi}_3 = 0.2129$, very close to those found in (e).