Comprehensive exam – Biomathematics 204 – 2019

Name (print) _____

If you use the back page or additional pages to write the answer to a question, be sure to identify each page with your name and the question number. Be sure to ask the exam proctor if any question is not clear.

1. Let x= 0 or 1 and s = $\sum_1^n x$ (s= the sum of the x's over n observations).

Let π = (population) probability that x=1.

The binomial distribution likelihood function (L) is given by

$$L = n!/(s!(n-s)!)\ \pi^s\ (1-\pi)^{(n-s)}$$

1a. What is the -log likelihood function and the maximum likelihood estimate (MLE) of π? Be sure to show your work.

1b. What is the (large sample) variance of the MLE above? (do not have to simplify).

2. An investigator wishes to assess the effects of age, gender, and season (winter or summer) on Y= having the flu (yes or no). Flu status was assessed in the winter and then in the summer on the **same** subject.

2a. The best model choice for this analysis is (indicate ONE)

i. Factorial linear regression ii. Repeated measure linear regression

iii. Factorial analysis of variance iv. Repeated measure analysis of variance

v. Factorial Poisson regression vi. Repeated measure Poisson regression

vii Factorial logistic regression viii Repeated measure logistic regression

ix. Factorial ordinal logistic regression x. Repeated measure ordinal logistic regression

xi Factorial robust regression xii Repeated measure robust regression

2b. What is the best model choice if Y is the number of flu episodes (0,1,2,3, …) in a given season?

2c. Log flu antibody titer is known to have a normal distribution although laboratory errors in measuring titer occur occasionally (i.e. outliers). What is the best model choice for Y = log flu antibody titer?

3. Y follows the normal (Gaussian) distribution. In group A, Y_{ai} has unknown mean μ_a and known standard deviation σ. In group B, Y_{bi} has unknown mean μ_b and known standard deviation σ. (The standard deviation is the same in both groups). Means based on a sample of size n are computed on each group. That is, the sample size is the same in each group.

$$\overline{Y_a} = \sum_1^n Y_{ai}/\text{n} \quad , \quad \overline{Y_b} = \sum_1^n Y_{bi}/\text{n}$$

3a. Give the Z statistic for testing the null hypothesis that $\mu_a=\mu_b$ or, equivalently, $\mu_a-\mu_b = 0$ and give the corresponding 95% confidence interval for the true mean difference, $\mu_a-\mu_b$. The Gaussian 97.5^{th} percentile is Z = 1.96.

Will the null mean difference value of zero be included in or excluded from the confidence interval above if the corresponding two sided p value is less than $\alpha=0.05$?

3b. Is it possible that the 95% confidence bounds for μ_a **overlaps** the 95% confidence bounds for μ_b even if the two sided p value for comparing the two means is less than $\alpha=0.05$? That is, can the two confidence intervals overlap even if the mean difference is significantly different?

4. Investigators wish to compare mean Y by ethnic group (Asian, Hispanic, African American). Y is known to have a normal distribution.

4a. How many regression (β) parameters are needed for this model? What is the corresponding degrees of freedom (df) for ethnic group?

4b. What is the design (X) matrix for a balanced design (equal n in all three groups) using dummy coding?

4c. What is the design (X) matrix for a balanced design using effect coding?

4d. Will the estimated values of the βs be the same using dummy versus effect coding?

4e. Will the estimated means for Y be the same using dummy versus effect coding?

4f. The F test for ethnic group gives a p value of $p = 0.7315$. Which regression coefficients, those for the intercept and/or those for ethnicity, are non significant?

5. Serum Bilirubin levels measure liver function. High levels imply poor liver function. Below are mean Bilirubin levels in mg/dl in male and female liver disease patients who either did or did not have a liver transplant. Bilirubin follows the normal distribution.

	n	Mean (mg/dl)
Males, no transplant	10	9.0
Males, transplant	10	3.0
Females, no transplant	10	7.0
Females, transplant	10	1.0
Overall (grand mean)	40	5.0

An analysis of variance gives a pooled SD = 0.40 mg/dl. The assumptions for the analysis of variance were met.

5a. What is the standard error and 95% **confidence** interval for Bilirubin in males with no transplant?

5b. What is the (at least approximate) 95% **prediction** interval for Bilirubin in males with no transplant?

5c. Based on the information given, is the gender x transplant term statistically significant ?

i) yes ii) no iii) unable to determine

5d. In the equation: **Bilirubin = β_0 + β_1 gender + β_2 trt + β_3 gender x trt + error** where gender is coded 1=male, -1=female and trt is coded 1=transplant, -1= no are X_1=gender, X_2=trt and X_3=gender x trt mutually orthogonal?

(yes, no, can't determine)?

6. The attached edited output shows a model for depression (depr:1=yes, 0=no) using data from 294 adults. The predictors are female gender ("female", 1=female, 0=male), age in years ("age") and "income" in 1000s of dollars (so "10" is 10,000). There is also an age x income interaction formed by multiplying the two variables together.

6a. According to this model, provide the estimated odds ratio of depression for a female compared to a male and its 95% confidence interval. Is this gender odds ratio the same for those with an income of 60 (thousand) compared to an income of 20 (thousand)?

6b. Give a formula for the depression odds ratio for a 20 (thousand) dollar increase in income or briefly explain if this cannot be done.

6c. According to this model, as age increases, does the odds of depression increase or decrease? Briefly explain.

6d. If this data was obtained from a case control (retrospective) study, could this model be used to compute the absolute risk of depression? (yes or no, briefly explain).

6e. The logit score for a 30 year old male with an income of 20 (thousand) is __. What is his model based risk of depression?

6f. Does the Hosmer and Lemeshow goodness of fit statistic indicate that the model fits the data (yes or no)?

Output for question 6

Variable	N	Min	Mean	Max	SD
age	294	18.00	44.41	89.00	18.09
income	294	2.000	20.57	65.00	15.29 (in 1000s)

female	Frequency	Percent
male=0	111	37.76
female=1	183	62.24

Logistic regression
Response Variable depression
Number of Observations 294

Value	depr	Frequency
1	yes	50
0	no	244

Probability modeled is depr='yes'.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	270.125	254.594
SC	273.808	273.011
-2 Log L	268.125	244.594

R-Square 0.0769 Max-rescaled R-Square 0.1286

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	p value
Intercept	1	0.1691	0.7736	0.0478	0.8270
female	1	0.9503	0.3882	5.9918	0.0144
age	1	-0.0412	0.0153	7.2819	0.0070
income	1	-0.1023	0.0433	5.5847	0.0181
age*income	1	0.00159	0.000928	2.9367	0.0466

Association of Predicted Probabilities and Observed Responses

Percent Concordant	70.4	Somers' D	0.413
Percent Discordant	29.1	Gamma	0.415
Percent Tied	0.5	Tau-a	0.117
Pairs	12200	c	0.707

Partition for the Hosmer and Lemeshow Test

		depr = yes		depr = no	
Group	Total	Observed	Expected	Observed	Expected
1	29	1	1.01	28	27.99
2	29	3	2.01	26	26.99
3	29	1	2.41	28	26.59
4	30	3	3.18	27	26.82
5	29	4	4.02	25	24.98
6	29	5	4.82	24	24.18
7	29	6	5.55	23	23.45
8	29	7	6.31	22	22.69
9	29	8	7.92	21	21.08
10	32	12	12.76	20	19.24

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	p value
1.6627	8	0.9897