Comprehensive exam – Biomathematics 204 – 2021

Be sure to identify each answer with your name and the question number. **Be sure to ask the exam proctor if any question is not clear**. Written notes prepared by you are allowed. Any form of communication with anyone but the proctor or the proctor's designate is forbidden. Note that the 97.5^{th} percentile of a normal distribution is Z=1.96.

1. Y is continuous $\log_{10}$ antibody titer. The titer level can be affected by the administration of a vaccine (yes or no) and by gender (male or female). Antibody titer was measured before vaccine administration (X_1=0), and after vaccine administration (X_1=1) in both males (X_2=0) and females (X_2=1). That is, each subject was measured before and after vaccination.

A model for the data is $Y_{ijk} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1X_2 + p_i + e_{ijk}$

where Y_{ijk} is the $(j, k)^{th}$ observation on subject i. j=0,1 corresponding to X_1, k=0,1 corresponding to X_2.

The p_i have a normal distribution with variance σ_p^2. The e_{ijk} have a normal distribution with variance σ_e^2. The p_i and e_{ijk} are independent. There are "n" males and "n" females for a total of 2n subjects.

The ML estimates of the regression coefficients are b_0, b_1, b_2 and b_3 respectively.

Give expressions for the following estimates using the information above:

1a. The sample mean log titer in females after vaccination and its variance and standard error.

1b. The sample <u>difference</u> in mean log titer in males after vaccination minus before vaccination and its variance and standard error.

1c. The sample covariance and correlation between the observations in males before vaccination and the same males after vaccination.

1d. The 95% <u>prediction</u> interval (<u>not</u> the confidence interval) for log titer in females after vaccination.

2. Investigators are using linear regression to explore factors related to Y=weight loss in a group of severely overweight subjects on a diet. After 3 months on the diet, the mean weight loss in this group is 50 lbs with a standard deviation of 40 lbs. Weight loss has as normal distribution. The investigators would like to have a model that predicts weight loss to within ± 5 lbs in 95% of subjects. What value of SD_e is needed to achieve this much precision? What is the corresponding R square value?

3. The random variable $y \geq 0$ is continuous. The density function of y and a parameter λ is given by

$$f(y|\lambda) = \lambda \exp(-\lambda y)$$

3a. What is the -log likelihood function and the maximum likelihood estimate (MLE) of λ for a sample size of n where $y_1, y_2, \ldots y_n$ are independent and identically distributed. Be sure to show your work.

3b. What is the (large sample) variance of the MLE above? (do not have to simplify).

3c. Using the standard error of the estimated λ, what would be the approximate 95% confidence bounds for λ (sample size is n) ?

4. A group of hospitalized patients with severe COVID 19 were placed on respirators. A study looked into factors that might be associated with being able to successfully remove the respirator within 14 days in a living patient ("success"). Failure was the inability to remove the respirator and/or death in 14 days. Gender (male or female), age in years, hemodialysis treatment (yes or no) and number of comorbidities (0, 1,2,3,…) were potentially associated with success.

Below: logit = log odds of success.
female=0 for males, 1 for females.
hemo=0 if no hemodialysis, 1 if patient had hemodialysis
numcor = number of comorbidities (0,1,2,…)

Two models were fit to the data

Model A: $\text{logit} = a_0 + a_1\,\text{female} + a_2\,\text{hemo} + a_3\,\text{age} + a_4\,\text{numcor} + a_5\,\text{female} * \text{numcor}$

Model B: $\text{logit} = b_0 + b_1\,\text{female} + b_2\,\text{hemo} + b_3\,\text{age} + b_4\,\text{age}^2 + b_5\,\text{numcor} + b_6\,\text{female} * \text{numcor}$

Model	-2log L	AIC	BIC
A	1061.4	1073.4	1107.0
B	1038.9	1052.9	1092.1

4a. Are these factorial or repeated measure (mixed) models?

4b. Are the regression coefficients for female in the two models (a_1 and b_1) necessarily the same?

4c. Are the regression coefficients for the female * numcor interaction in the two models (a_5 and b_6) necessarily the same?

4d. Give a likelihood ratio test statistic and its degrees of freedom (df) for testing that the effect of age is linear vs quadratic. Are the results for this test consistent with the AIC and BIC results? Is there evidence that the relation of age with the logit is non linear?

4e. In model **B**, if $b_6=0$ indicate which are true

i The effect of numcor and gender on the **logit** is additive

ii The effect of numcor and gender on the **odds** of success is multiplicative

iii. The effect of numcor and gender on the **model based probability** of success is multiplicative

iv. The effect of numcor and gender on the **model based probability** is additive.

4f. Give an expression for the odds ratio of success in females compared to males based **on model B** above.

4g. Give an expression for the odds and model based probability of success in 40 year old males who had hemodialysis and no comorbid conditions based on **model A above**.

Model **A** results after data fitting are given below where n=10,000

	Estimate	SE	p value
intercept	-14.099	0.695	0.0000
female	-0.253	0.226	0.2629
hemo	-0.097	0.028	0.0004
age	0.333	0.017	0.0000
numcor	0.153	0.113	0.1730
female*numcor	0.204	0.068	0.0027

Sensitivity = 97%, Specificity = 96% ROC area = 0.99

Hosmer-Lemeshow chi square =16.09, df=8, p value = 0.041

4h. Using model A results above, give an expression for the odds ratio for hemo and its 95% confidence bounds. If the confidence bounds cannot be computed, briefly explain why.

4i. Using model A results, give an expression for the odds ratio for females with one comorbidity and its 95% confidence bounds. If the confidence bounds cannot be computed, briefly explain why.

4j. A critic says that the model does not fit the data, citing the Hosmer-Lemeshow statistic above, and therefore says the model is not very **accurate**. Given the information above, briefly comment on whether this is a valid concern regarding accuracy.

4k. The critic also says that, since the p value of 0.2629 for the "female" term is not significant at $p < 0.05$, that female gender has no significant effect on respirator success or failure. Is this interpretation correct?

5. A related outcome in hospitalized COVID patients who were put on respirators is time to respirator removal ("weaning"). Indicate the **best** model below for determining if factors such as gender, age, hemodialysis and number of comorbidities are related to time to respirator removal from initial start on the respirator.

5a. Linear regression

5b. Analysis of variance

5c. Logistic regression

5d. Cox (proportional hazard) regression

5e. Poisson regression

6. Indicate which of the following are methods directly useful for variable selection / elimination-retention.

6a. ROC area

6b. Likelihood ratio test(s)

6c. Forward / backward stepping

6d. LOESS

6e. Restricted cubic splines

6f. LASSO

6g. Leverage

6h. Outlier identification with Mahalanobis distance

6i. Density estimation

6j. Missing data imputation

6k. Random forest - variable importance