Comprehensive exam – Biomathematics 204 – 2020

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 97.5th percentile of a normal distribution is Z=1.96.

1. The random variable y takes on the values 0, 1, 2, 3 … n

The density function of y and a parameter λ is given by

$$f(y|\lambda) = (\lambda^{y} e^{-\lambda}) / y!$$

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

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

1c. Using the standard error of the estimated λ, what would be the approximate 95% confidence bounds for λ?

2. In the linear regression, $Y = \beta_0 + \beta_1 G + \beta_2 X_2 + \varepsilon$

the estimated equation is $Y = b_0 + b_1 G + b_2 X_2 + e$

$\bar{X}_2$ is the mean of X_2. σ_e is the standard deviation of the errors (ε) and $\hat{\sigma}_e$ is the estimate of σ_e. The sample size is "n".

The variable G (group) is coded 0 or 1 (ie two groups). The variables Y and X_2 are continuous.

The estimated **<u>regression coefficien</u>**<u>t</u> covariance matrix is given by:

$$\text{Cov}(b) = \begin{array}{ccc} c_{00} & c_{01} & c_{02} \\ c_{10} & c_{11} & c_{12} \\ c_{20} & c_{21} & c_{22} \end{array}$$

2a. Write an expression for mean Y and its variance and standard error when G=0 and $X_2 = \bar{X}_2$ if β_0, β_1 and β_2 are <u>known</u>.

2b. Write an expression for mean Y and its variance and standard error when G=0 and $X_2 = \bar{X}_2$ as above when β_0, β_1 and β_2 are <u>unknown</u> and are estimated by b_0, b_1 and b_2 respectively.

2c. Write an expression for the mean difference in Y for G=1 minus G=0 when X_2 is the same in both groups. Write down its variance and standard error.

2d. The distribution of X_2, including the mean X_2 and standard deviation of X_2 is the same in group G=0 and G=1.

2d (i) Are G and X_2 orthogonal?

2d (ii) Will the estimate of β_2 (b_2) necessarily be the same if G is removed from the model?

2d (iii) Will the estimated **standard error** of b_2 necessarily be the same if G is removed from the model? Briefly explain. (Derivation NOT required).

3. Subjects were tested for absorption of a toxin (positive or negative) from an initial day 0. The effects of smoking (yes or no), male vs female gender and follow up time in days on a positive or negative test was modelled by a logistic regression of the form:

$$\text{logit}(P) = \beta_0 + \beta_1 \text{ smoking } + \beta_2 \text{ male} + \beta_3 \text{ smoking x male } + \beta_4 \text{ days } + a_i$$

where P is the proportion who test positive. The variable “smoking” is coded 1 for smoking and 0 for non smoking. The variable “male” is coded 1 for male and 0 for female. The variable “days” is the day of the test. The same subject can have a test on more than one day. For example, subject i=7 was tested on day 3 and day 12. A subject that is positive on one day does not necessarily have to be positive on another day.

The variable a_i in the model above is a normally distributed random variable with mean 0 and a constant standard deviation σ. All the observations from subject “i” have the same value of a_i. Observations from different subjects have different values of a_i.

3a. Is this a factorial or repeated measure (mixed) model?

3b. Give an expression for the odds ratio (OR) for testing positive in smokers versus non smokers. Does this depend on male gender?

3c. Under this model, will two subjects with exactly the same smoking status, gender, and test day have exactly the same risk of being positive? Briefly explain.

3d. If β_3=0 indicate which are true

i The effect of smoking and male on the **logit** is additive

ii The effect of smoking and male on the **odds** of testing positive is multiplicative

iii. The effect of smoking and male on the **risk** of testing positive is multiplicative

iv. The effect of smoking and male on the **risk** of testing positive is additive.

The output below shows the results of evaluating this model with data.

Variable	b	standard error of b	p value
Intercept	-2.00	0.010	< 0.0001
Smoking	0.15	0.065	0.0210
Male	0.20	0.060	0.0009
Smoking x male	-0.03	0.015	0.0455
Days	0.06	0.025	

Estimate of $\sigma = 0.25$

Sensitivity = 97%, Specificity = 98%

Hosmer-Lemeshow chi square = 18.168, df=8, p value = 0.020

n=10,000

3e. Compute the odds ratio of being positive on day 7 compared to day 1 and its 95% confidence bounds.

3f. Is the risk of testing positive higher in male non smokers or female non smokers?

3g. Is the risk of testing positive higher in male smokers or female smokers?

3h. Interpret exp(-0.03) = 0.970, the "interaction" of gender and smoking.

For example, is 0.970 the odds ratio of testing positive in male smokers compared to female smokers? Briefly explain.

3i. For female non smokers on day 5, the logit is -2 + 5 (0.06) = -1.7, **ignoring the a_i.**
Ignoring the a_i, the corresponding standard error of the -1.7 is 0.027.
(ie, sqrt(var(intercept) + var(days) + 2 cov(intercept, days)) = 0.027

What is the risk of testing positive for female non smokers on day 5 and its 95% prediction bounds, taking into account the a_i.

3j. 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, is this a valid concern? Briefly explain.

4. Indicate which of the following are methods for variable selection / elimination-retention.

4a. Hazard ratios

4b. Likelihood ratio test(s)

4c. Forward / backward stepping

4d. Loess

4e. Restricted cubic splines

4f. LASSO

4g. Leverage

4h. Outlier identification

4i. Density estimation

4j. Missing data imputation

5. Researchers wish to know how Y=log antibody titer (a continuous outcome) is affected by gender (male=0, female=1), age in years and body mass index (BMI) in kg/m^2.

They fit a model of the form

$$Y = b_0 + b_1 \text{ gender } + b_2 \text{ age} + b_3 \text{ BMI} + b_4 \text{ gender x age x BMI } + e$$

(The last variable is gender multiplied by age multiplied by BMI)

Results from the model fit are given below

Variable	**b-estimate**	**standard error of b**	**p value**
Intercept (b_0)	-2.10	0.20	< 0.0001
gender (b_1)	0.20	0.50	0.6892
age (b_2)	-0.40	0.09	< 0.0001
BMI (b_3)	-0.10	0.35	0.7751
gender-age-BMI (b_4)	0.08	0.06	0.1825

SD of Y = 4.0 (log titer units)   error SD = SD_e = 2.0 (log titer units)

It is known that age and BMI are not the same in both genders and that BMI tends to increase as age increases.

5a. Compute the R square for this model.

5b. Since b_1, b_3 and b_4 are all not statistically significant, can all of these terms be removed from the model? Briefly explain.

5c. Does this analysis show that log antibody titer decreases with increasing age in both genders and for all levels of BMI? Briefly explain.