Biomath 204 2019 exam

SIMON LEE

July 2024

1 Question 1

1a. The log-likelihood function is obtained by taking the natural logarithm of the likelihood function L:

$\log L = \log(n!/(s!(n-s)!)) + s\log(\pi) + (n-s)\log(1-\pi)$

To find the maximum likelihood estimate (MLE) of π, we take the derivative of $\log L$ with respect to π, set it equal to 0, and solve for π:

$\frac{d}{d\pi}\log L = \frac{s}{\pi} - \frac{n-s}{1-\pi} = 0$

$\frac{s}{\pi} = \frac{n-s}{1-\pi}$

$s(1-\pi) = \pi(n-s)$

$s - s\pi = \pi n - \pi s$

$s = \pi n$

$\hat{\pi}_{MLE} = \frac{s}{n}$

1b. The large sample variance of the MLE is given by the inverse of the negative expected value of the second derivative of the log-likelihood function, evaluated at the MLE:

$Var(\hat{\pi}MLE) = -\left(\mathbb{E}\left[\frac{d^2\log L}{d\pi^2}\right]\pi = \hat{\pi}_{MLE}\right)^{-1}$

$\frac{d^2\log L}{d\pi^2} = -\frac{s}{\pi^2} - \frac{n-s}{(1-\pi)^2}$

$\mathbb{E}\left[\frac{d^2\log L}{d\pi^2}\right]\pi = \hat{\pi}MLE = -\frac{n}{\hat{\pi}MLE} - \frac{n}{1-\hat{\pi}MLE}$

$= -\frac{n^2}{s} - \frac{n^2}{n-s}$

$= -\frac{n^2(n-s+s)}{s(n-s)} = -\frac{n^3}{s(n-s)}$

$Var(\hat{\pi}MLE) = \left(\frac{n^3}{s(n-s)}\right)^{-1} = \frac{s(n-s)}{n^3} = \frac{\hat{\pi}MLE(1-\hat{\pi}_{MLE})}{n}$

So in summary:

The log-likelihood function is the logarithm of the likelihood function L

The MLE of π is $\hat{\pi}_{MLE} = \frac{s}{n}$

The large sample variance of the MLE is $Var(\hat{\pi}MLE) = \frac{\hat{\pi}MLE(1-\hat{\pi}_{MLE})}{n}$

2 Question 2

Question 2a states that the best model choice is to use repeated measures. From the options, this would be:

ii. Repeated measure linear regression

iv. Repeated measure analysis of variance

vi. Repeated measure Poisson regression

viii. Repeated measure logistic regression

x. Repeated measure ordinal logistic regression

xii. Repeated measure robust regression

Question 2b asks for the best model choice if Y is the number of flu episodes $(0, 1, 2, 3, \ldots)$ in a given season. This is count data, so a Poisson regression model would be most appropriate:

vi. Repeated measure Poisson regression

Question 2c mentions that log flu antibody titer has a normal distribution. Since the log of the dependent variable is normally distributed, this suggests a log-linear model:

ii. Repeated measure linear regression

where $Y = \log(\text{flu antibody titer})$
In summary:

- 2b: **Repeated measure Poisson regression** is best for modeling a count dependent variable Y.
- 2c: **Repeated measure linear regression** is best for modeling a log-transformed normally distributed dependent variable $\log(Y)$.

3 Question 3

The question is asking whether it's possible for the 95% confidence intervals for μ_a and μ_b to overlap even if the two-sided p-value for comparing the two means is less than $\alpha = 0.05$, i.e. if the mean difference is significantly different.
To answer this, we need to compute the Z statistic and corresponding confidence interval for testing $H_0 : \mu_a = \mu_b$.
The Z statistic is given by:

$$\begin{aligned} Z &= \frac{\bar{Y}_a - \bar{Y}_b}{\sqrt{\frac{\sum_1^n Y_{ai}^2}{n} + \frac{\sum_1^n Y_{bi}^2}{n}}} \\ &= \frac{\bar{Y}_a - \bar{Y}_b}{\sqrt{\frac{\bar{Y}_a}{n} + \frac{\bar{Y}_b}{n}}} \end{aligned}$$

The Gaussian 97.5th percentile is $Z = 1.96$. So the 95% confidence interval for $\mu_a - \mu_b$ is:

$$\bar{Y}_a - \bar{Y}_b \pm 1.96\sqrt{\frac{\bar{Y}_a}{n} + \frac{\bar{Y}_b}{n}}$$

If this confidence interval includes 0, then the null hypothesis $\mu_a = \mu_b$ cannot be rejected at the $\alpha = 0.05$ level.
The individual 95

$$\mu_a \in \bar{Y}_a \pm 1.96\sqrt{\frac{\bar{Y}_a}{n}}$$
$$\mu_b \in \bar{Y}_b \pm 1.96\sqrt{\frac{\bar{Y}_b}{n}}$$

It's possible for these individual intervals to overlap even if the confidence interval for $\mu_a - \mu_b$ does not include 0. This is because the variance of $\bar{Y}_a - \bar{Y}_b$ is less than the sum of the individual variances.
Therefore, the answer is yes, it is possible for the 95% confidence intervals for μ_a and μ_b to overlap even if the mean difference is significantly different at the $\alpha = 0.05$ level using a two-sided test.

4 Question 4

Here are the responses to the statistical questions presented in your image:
4a. How many regression (β) parameters are needed for this model? What is the corresponding degrees of freedom (df) for ethnic group?

For a model comparing mean Y by three ethnic groups (Asian, Hispanic, African American) using dummy coding: - You will need one parameter for the intercept (β_0) and two additional parameters for two of the ethnic groups (β_1 for one group and β_2 for another), while the third group's mean is captured by the intercept. - The degrees of freedom for the ethnic group is equal to the number of categories minus one: $3 - 1 = 2$.

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

For dummy coding with three groups, if we take the first group (e.g., Asian) as the reference category:

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

where the first column of ones represents the intercept, the second column represents Hispanic, and the third column represents African American.

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

For effect coding:

$$X = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

Here, negative ones for the first group (Asian) indicate it is the reference group, and the other columns represent the deviations of Hispanic and African American from the reference.

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

No, the estimated values of β coefficients will not be the same. Dummy coding estimates the mean of each group relative to the reference group, while effect coding estimates the deviation of each group's mean from the overall mean.

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

Yes, the estimated means for Y calculated from the model will be the same regardless of whether dummy coding or effect coding is used, although the interpretation and values of the β coefficients differ.

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?

A high p-value (e.g., $p = 0.7315$) indicates that there is not enough evidence to reject the null hypothesis that all ethnicity coefficients are zero simultaneously. This suggests that the ethnicity variables (i.e., the coefficients for Hispanic and African American in the model) do not have a statistically significant effect on Y. The intercept may still be significant if it represents the mean of the reference group (which is tested separately), but the effects of ethnicity are non-significant.

5 Question 5

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

To calculate the standard error and 95% confidence interval for Bilirubin in males with no transplant, we need to use the pooled standard deviation (SD) of 0.40 mg/dl and the sample size (n) of 10.

The standard error is calculated as:

$$
\begin{aligned}
SE &= \frac{SD}{\sqrt{n}} \\
&= \frac{0.40}{\sqrt{10}} \\
&= \frac{0.40}{\sqrt{10}} \\
&= 0.1265 \text{ mg/dl}
\end{aligned}
$$

The 95% confidence interval is calculated as:

$$
\begin{aligned}
95\% \text{ CI} &= \text{mean} \pm 1.96 \times SE \\
&= 9.0 \pm 1.96 \times 0.1265 \\
&= 9.0 \pm 0.2479 \\
&= (8.7521, 9.2479) \text{ mg/dl}
\end{aligned}
$$

Therefore, the standard error for Bilirubin in males with no transplant is 0.1265 mg/dl, and the 95% confidence interval is (8.7521, 9.2479) mg/dl.

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

The 95% prediction interval for Bilirubin in males with no transplant is (8.7521, 9.2479) mg/dl, as calculated in the previous question.

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

Based on the information provided, the gender x transplant term is statistically insignificant, as the p-value is greater than the typical significance level of 0.05. Therefore, we cannot determine if the interaction term is significant.

5d. In the equation:

$$\textbf{Bilirubin} = \beta_0 + \beta_1 \textbf{gender}' + \beta_2 \textbf{trt}' + \beta_3 (\textbf{gender}' \times \textbf{trt}') + \textbf{error}$$

where gender is coded 1=male, -1=female and trt is coded 1=transplant, -1= no are X1=gender, X2=trt and X3=gender x trt' mutually orthogonal? (yes, no, can't determine)?

The equation cannot be determined as the coefficients for X1, X2, and X3 are not provided in the given information.

6 Question 6

6a. The odds ratio of depression for a female compared to a male is:

$$
\text{Odds ratio} = \frac{\text{Odds of depression for female}}{\text{Odds of depression for male}} = \frac{\frac{0.1076}{1-0.1076}}{\frac{0.0476}{1-0.0476}} = \frac{0.1076/0.8924}{0.0476/0.9524} = \frac{0.1205}{0.0500} = 2.41
$$

The 95% confidence interval is given, so this odds ratio is the same for those with an income of 60 (thousand) compared to an income of 20 (thousand).

6b. Let p_1 be the probability of depression for income of 20 (thousand) and p_2 be the probability for income of 40

(thousand). From the model:

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= -1.8650 - 0.0164 \cdot \text{age} + 0.8755 \cdot \text{income} \\
\log\left(\frac{p_2}{1-p_2}\right) - \log\left(\frac{p_1}{1-p_1}\right) &= 0.8755 \cdot (40 - 20) \\
\log\left(\frac{p_2/(1-p_2)}{p_1/(1-p_1)}\right) &= 0.8755 \cdot 20 \\
\log\left(\frac{\text{Odds}_2}{\text{Odds}_1}\right) &= 17.51 \\
\frac{\text{Odds}_2}{\text{Odds}_1} &= e^{17.51} \\
\text{Odds ratio for 20 (thousand) increase} &= e^{17.51} \approx 4.0 \times 10^7
\end{aligned}$$

6c. As age increases, the coefficient for age (-0.0164) is negative, meaning the odds of depression decrease. The odds ratio for a 1 year increase in age is $e^{-0.0164} \approx 0.984$, so the odds of depression decrease by about 1.6% per year of age.

6d. If this data was obtained from a case control study, the absolute risk of depression cannot be computed. A case control study selects based on the outcome (depression), so it does not give information about the prevalence of depression in the general population. The model could still be used to compare the relative odds of depression between groups.

6e.

$$\begin{aligned}
\text{Logit score} &= -1.8650 - 0.0164 \cdot \text{age} + 0.8755 \cdot \text{income} \\
&= -1.8650 - 0.0164 \cdot 30 + 0.8755 \cdot 20 \\
&= -1.8650 - 0.492 + 17.51 \\
&= 15.1530
\end{aligned}$$

6f. The Hosmer and Lemeshow goodness of fit statistic tests if the model fits the data adequately. Since the p-value is not provided, there is not enough information to determine if the model has a good fit or not.