Biomath 204 2020 Exam

SIMON LEE

July 2024

1 Question 1

1a. Log-Likelihood Function and MLE of λ

Given that y can take values $0, 1, 2, \ldots, n$, and the probability mass function is given by:

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

This is the probability mass function of the Poisson distribution.

For a sample of size n, $y_1, y_2, \ldots, y_n$, the likelihood function $L(\lambda)$ is:

$$L(\lambda) = \prod_{i=1}^{n} f(y_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$$

The log-likelihood function, $\ell(\lambda)$, is:

$$\ell(\lambda) = \log L(\lambda) = \sum_{i=1}^{n} \log\left(\frac{\lambda^{y_i} e^{-\lambda}}{y_i!}\right) = \sum_{i=1}^{n} (y_i \log\lambda - \lambda - \log(y_i!))$$

$$\ell(\lambda) = \left(\sum_{i=1}^{n} y_i\right) \log\lambda - n\lambda - \sum_{i=1}^{n} \log(y_i!)$$

To find the MLE of λ, we differentiate $\ell(\lambda)$ with respect to λ and set the derivative to zero:

$$\frac{d\ell}{d\lambda} = \frac{\sum_{i=1}^{n} y_i}{\lambda} - n = 0$$

$$\sum_{i=1}^{n} y_i = n\lambda$$

$$\hat{\lambda}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

This is the sample mean of y.

1b. Variance of the MLE

The variance of the MLE of a Poisson distribution is equal to:

$$\text{Var}(\hat{\lambda}_{MLE}) = \frac{\lambda}{n}$$

Since $\hat{\lambda}_{MLE}$ is the sample mean of y and for Poisson distribution $\text{Var}(Y) = \lambda$, the sample mean variance becomes $\frac{\lambda}{n}$.

1c. 95% Confidence Interval for λ

Using the standard error of λ, which is $\sqrt{\frac{\hat{\lambda}_{MLE}}{n}}$, the approximate 95% confidence interval is given by:

$$\hat{\lambda}_{MLE} \pm 1.96\sqrt{\frac{\hat{\lambda}_{MLE}}{n}}$$

This uses the normal approximation, which is reasonable for large n.

Thus, these are the comprehensive solutions to each part of the question.

2 Question 2

Let's address the questions from the image step by step.

2a. Expressions for Mean Y, Its Variance, and Standard Error for $G = 0$

Given:

$$Y = \beta_0 + \beta_1 G + \beta_2 X_2 + \epsilon$$

For $G = 0$, the equation simplifies to:

$$Y = \beta_0 + \beta_2 X_2 + \epsilon$$

Mean of Y when $G = 0$:

$$E[Y|G = 0] = \beta_0 + \beta_2 \bar{X}_2$$

Variance and Standard Error of Y when $G = 0$: Since ϵ is normally distributed with mean 0 and standard deviation σ_ϵ, the variance of Y when $G = 0$ is:

$$\text{Var}(Y|G = 0) = \sigma_\epsilon^2$$

The standard error of Y when $G = 0$ is:

$$\text{SE}(Y|G = 0) = \sigma_\epsilon$$

2b. Expressions for Mean Y and Its Variance and Standard Error for $G = 0$ with Unknown Parameters

Mean of Y when $G = 0$ with estimates:

$$E[Y|G = 0] = b_0 + b_2 \bar{X}_2$$

Variance of Y when $G = 0$ with estimates: The variance remains as:

$$\text{Var}(Y|G = 0) = \sigma_\epsilon^2$$

The variance of the estimate of Y also includes the variance due to the estimation error in b_0 and b_2:

$$\text{Var}(\hat{Y}|G = 0) = \text{Var}(b_0) + \bar{X}_2^2 \text{Var}(b_2) + 2\bar{X}_2 \text{Cov}(b_0, b_2) + \sigma_\epsilon^2$$

Using the provided covariance matrix elements:

$$\text{Var}(\hat{Y}|G = 0) = c_{00} + \bar{X}_2^2 c_{22} + 2\bar{X}_2 c_{02} + \sigma_\epsilon^2$$

Standard Error of Y when $G = 0$ with estimates:

$$\text{SE}(\hat{Y}|G = 0) = \sqrt{\text{Var}(\hat{Y}|G = 0)}$$

2c. Mean Difference in Y for $G = 1$ minus $G = 0$

Mean Difference:

$$E[Y|G = 1] - E[Y|G = 0] = (\beta_0 + \beta_1 + \beta_2 \bar{X}_2) - (\beta_0 + \beta_2 \bar{X}_2) = \beta_1$$

Variance of the Mean Difference:

$$\text{Var}(\beta_1) = c_{11}$$

Standard error:

$$\text{SE}(\beta_1) = \sqrt{c_{11}}$$

2d(i). Orthogonality of G and X_2

G and X_2 are orthogonal if $\text{Cov}(G, X_2) = 0$. This can be inferred from c_{12} and c_{21} in the covariance matrix. If these values are zero, then G and X_2 are orthogonal.

2d(ii). Estimate of β_2 if G is Removed

Removing G from the model makes X_2 potentially a stronger predictor if it was previously confounded by G. The estimate of β_2 might change depending on the correlation between G and X_2. If they are uncorrelated (orthogonal), removing G does not affect the estimate of β_2. If they are correlated, β_2 could change.

These responses provide detailed answers for each part, considering both the case where the regression coefficients are known and when they are estimated.

Let's address the questions from the provided image:

2d(iii). Will the Estimated Standard Error of b_2 Necessarily be the Same if G is Removed from the Model?

No, the estimated standard error of b_2 will not necessarily be the same if G is removed from the model. The presence of G can control for the variance due to group differences that might also affect X_2. Removing G could increase the variability of Y that b_2 tries to explain, potentially increasing the standard error of b_2, especially if G and X_2 are correlated.

3 Question 3

3a. Is This a Factorial or Repeated Measure (Mixed) Model?

The model is a mixed model (or a random effects model). It incorporates fixed effects (smoking, male gender, interaction between smoking and male, and days) and a random effect (a_i), where a_i is a normally distributed random variable associated with each subject. This random effect accounts for variability between subjects that is not explained by the observed variables.

3b. Expression for the Odds Ratio (OR) for Testing Positive in Smokers vs. Non-Smokers

The odds ratio (OR) for testing positive in smokers versus non-smokers is calculated from the logistic regression coefficient for smoking (β_1):

$$\text{OR} = e^{\beta_1}$$

This expression assumes that other variables (male, days, a_i) are held constant, so it does not depend on male gender alone but on the interaction term's coefficient (β_3) as well.

3c. Do Two Subjects with the Same Attributes Have the Same Risk of Being Positive?

No, two subjects with exactly the same smoking status, gender, and test day may not have exactly the same risk of being positive due to the random effect a_i. Each subject i has a unique a_i that affects their individual susceptibility or response, leading to different risks even among subjects with identical observed characteristics.

3d. Interpretations if $\beta_3 = 0$

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

- True. If $\beta_3 = 0$, the interaction term in the logit model is non-contributory, implying that the effects of smoking and being male simply add up, with no multiplicative interaction effect on the logit scale.

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

- True. In logistic regression, the absence of an interaction term ($\beta_3 = 0$) means that the individual effects of smoking and being male on the odds are multiplicative (independent effects).

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

- True. Similar to the odds, the risk (probability) changes implied by independent logistic effects are multiplicative in the absence of an interaction.

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

- False. The risk (probability) is not additive; the logistic model inherently deals with odds and risks in a multiplicative fashion when the effects are independent (no interaction).

Let's address each of the questions based on the model results provided:

3e. Odds Ratio of Being Positive on Day 7 Compared to Day 1

The coefficient for "Days" ($\beta_{\text{days}} = 0.06$) represents the log of the odds ratio for each additional day. The odds ratio for day 7 compared to day 1, which is 6 days apart, is calculated as:

$$\text{OR}_{7 \text{ days}} = \exp(0.06 \times 6) = \exp(0.36)$$

Calculating exp(0.36):

$$\text{OR}_{7 \text{ days}} \approx 1.433$$

95% Confidence Bounds: The standard error for a 6-day increase is calculated as:

$$\text{SE}_{7 \text{ days}} = \sqrt{6^2 \times 0.025^2} = \sqrt{0.09} = 0.3$$

The 95

$$\text{CI}_{\text{log odds}} = 0.36 \pm 1.96 \times 0.3 = [0.36 - 0.588, 0.36 + 0.588] = [-0.228, 0.948]$$

Converting this back to the odds ratio:

$$\text{CI}_{\text{OR}} = [\exp(-0.228), \exp(0.948)] \approx [0.796, 2.58]$$

3f. Risk of Testing Positive: Male Non-Smokers vs. Female Non-Smokers
The logistic regression does not directly indicate higher risk between male non-smokers and female non-smokers unless an interaction term specific to non-smoking and gender was significant. Here, the coefficients suggest: - Smoking ($\beta_{\text{smoking}} = 0.15$) and - Male ($\beta_{\text{male}} = 0.20$),
But their interaction ($\beta_{\text{smoking x male}} = -0.03$) slightly decreases the effect when both are present. Thus, male non-smokers would have a higher base risk of testing positive compared to female non-smokers because β_{male} directly adds to the logit.

3g. Risk of Testing Positive: Male Smokers vs. Female Smokers
For smokers: - The effect for males is $\beta_{\text{smoking}} + \beta_{\text{male}} + \beta_{\text{smoking x male}} = 0.15 + 0.20 - 0.03 = 0.32$, - For females, it's just $\beta_{\text{smoking}} = 0.15$.
Thus, the effect is stronger in male smokers due to the additional positive coefficient for being male, despite the slightly negative interaction term.

3h. Interpretation of $\exp(-0.03) = 0.970$
This represents the interaction effect of gender and smoking. An odds ratio of 0.970 suggests that the combined effect of being a male smoker is slightly less than the product of their individual effects. This indicates a 3% decrease in the odds of testing positive for male smokers compared to what would be expected if their risks were purely multiplicative (without interaction).

3i. Risk of Testing Positive for Female Non-Smokers on Day 5
For female non-smokers on day 5, the logistic regression prediction is:

$$\text{logit} = -2 + 0.06 \times 5 = -1.7$$

This corresponds to a probability:

$$P = \frac{1}{1 + \exp(1.7)} \approx \frac{1}{1 + 5.474} \approx 0.154$$

95% Prediction Bounds: Standard error calculation is previously provided as 0.027. For a 95% CI:

$$\text{CI}_{\text{logit}} = -1.7 \pm 1.96 \times 0.027 = [-1.75292, -1.64708]$$

Converting this to probability:

$$\text{CI}_P = \left[\frac{1}{1 + \exp(-1.64708)}, \frac{1}{1 + \exp(-1.75292)}\right] \approx [0.161, 0.148]$$

Let's address the questions in the image:

3j. Valid Concern about the Model Fit Citing the Hosmer-Lemeshow Test
The Hosmer-Lemeshow test is a goodness-of-fit test for logistic regression models, which checks whether the observed event rates match expected event rates in subgroups of the model population. The critic's concern about the model not fitting well because of a significant Hosmer-Lemeshow test statistic (p-value = 0.020, chi-square = 18.168, df=8) is valid. A significant result (p-value ¡ 0.05) suggests that there are differences between the observed and predicted values, indicating a potential lack of fit. This could mean that the model may not accurately represent the data or that certain important predictors or interactions are missing or mis-specified.

4 Question 4

Here's what each listed method generally pertains to:

- - 4a. Hazard ratios: Used to describe the effect of variables in survival models, not typically a method of variable selection.
- - 4b. Likelihood ratio test(s): Used to compare the goodness-of-fit of two models - one with and without a potential predictor. It is a method for variable selection.
- - 4c. Forward / backward stepping: These are stepwise regression methods used to add or remove predictors based on certain criteria (like AIC, BIC, p-value thresholds). These are methods for variable selection.
- - 4d. Loess: This is a method for fitting local regressions, used in smoothing data, not for variable selection.
- - 4e. Restricted cubic splines: Used to model non-linear relationships in data. Not a method for variable selection but rather for modeling the form of relationships.
- - 4f. LASSO (Least Absolute Shrinkage and Selection Operator): A regularization technique that can both select variables and shrink coefficients to zero. Definitely a method for variable selection.
- - 4g. Leverage: Typically related to influence in a regression model but not a direct method for variable selection.
- - 4h. Outlier identification: While identifying outliers can inform the need to modify a model, it is not a variable selection method.
- - 4i. Density estimation: This is more about estimating the distribution of data, not selecting variables.
- - 4j. Missing data imputation: While necessary for preparing data, it does not select variables but rather fills in missing values.

Summary of Methods for Variable Selection/Elimination-Retention:

- - Likelihood ratio tests
- - Forward/backward stepping
- - LASSO

These methods specifically aim at enhancing the model by choosing the most significant variables.

5 Question 5

5a. Compute the R-square for this model
The R-square for a model in the context of linear regression is typically computed as the proportion of the variance in the dependent variable that is predictable from the independent variables. However, for models fit using methods other than ordinary least squares (like logistic regression or when dealing with transformed outcomes), this computation might involve different approaches such as pseudo R-squared values. Given the standard deviation (SD) of Y and the error SD (standard error of the estimate, SDe), you might approximate an R-square using:

$$R^2 = 1 - \frac{\text{Var}(\epsilon)}{\text{Var}(Y)}$$

$$R^2 = 1 - \frac{SDe^2}{SD^2} = 1 - \frac{4}{16} = 1 - 0.25 = 0.75$$

This method, however, assumes that the residual variance (SDe^2) and the total variance (SD^2) are based on the same scale, which might not always be accurate without specific model outputs.
5b. Significance of b_1, b_3, and b_4
The p-values for gender (b_1), BMI (b_3), and the interaction term (b_4) are above the conventional significance level (0.05), suggesting that these variables are not statistically significant at the 95% confidence level. However, deciding to remove these terms from the model should not be based solely on p-values. Consideration of clinical or research significance is essential, especially since the interaction of these variables could be meaningful, or there could be a power issue.
Removing b_3 (BMI) and b_4 (gender-age-BMI interaction) without further analysis could oversimplify the model, possibly omitting important effects or interactions that could be significant in different contexts or subgroups. Model reduction should follow a structured approach, considering domain knowledge and potentially using other criteria like AIC or BIC for a more holistic decision.
5c. Interpretation of Analysis Regarding Age
The coefficient for age (b_2) is statistically significant with a p-value ¡ 0.0001, and the estimate is -0.40, suggesting that, on average, an increase in age is associated with a decrease in the log antibody titer. The interpretation would be that each additional year of age is associated with a decrease in the log antibody titer by 0.40 units, all else being equal.
This effect holds regardless of gender and BMI since the interaction term (gender-age-BMI) and the main effect of BMI are not statistically significant. However, the absence of statistical significance in the interaction term does not necessarily mean there is no effect modification by gender or BMI; it could also be due to lack of power or other modeling issues. Therefore, the statement that the log antibody titer decreases with age across both genders and all levels of BMI is supported by b_2 but should be interpreted with caution given the broader model context.