Biomath 204 2021 Exam

SIMON LEE

July 2024

1 Question 1

1a. The Sample Mean Log Titer in Females After Vaccination and Its Variance and Standard Error
Given the regression model:

$$Y_{ijk} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + p_i + e_{ijk}$$

For females after vaccination: - $X_1 = 1$ (after vaccination) - $X_2 = 1$ (female)
The model simplifies to:

$$Y_{ijk} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + p_i + e_{ijk}$$

Mean

$$\text{Mean} = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

Variance and Standard Error Since p_i has variance σ_p^2 and e_{ijk} has variance σ_e^2, the total variance of Y_{ijk} for a single observation is:

$$\text{Var}(Y_{ijk}) = \sigma_p^2 + \sigma_e^2$$

The standard error for the mean (for n females) is:

$$\text{SE(Mean)} = \sqrt{\frac{\sigma_p^2 + \sigma_e^2}{n}}$$

1b. The Sample Difference in Mean Log Titer in Males After Vaccination Minus Before Vaccination and Its Variance and Standard Error
For males after vaccination ($X_1 = 1, X_2 = 0$): - $Y_{ijk}^{(\text{after})} = \beta_0 + \beta_1 + p_i + e_{ijk}$
For males before vaccination ($X_1 = 0, X_2 = 0$): - $Y_{ijk}^{(\text{before})} = \beta_0 + p_i + e_{ijk}$
Difference

$$\text{Difference} = \beta_1$$

Variance and Standard Error The variance of the difference for a single individual (considering the independence of errors):

$$\text{Var(Difference)} = \sigma_e^2 + \sigma_e^2 = 2\sigma_e^2$$

Since this is for individual males and not an average, the standard error is:

$$\text{SE(Difference)} = \sqrt{2\sigma_e^2}$$

1c. The Sample Covariance and Correlation Between the Observations in Males Before Vaccination and the Same Males After Vaccination
Given $Y_{ijk}^{(\text{after})}$ and $Y_{ijk}^{(\text{before})}$ have a shared random effect p_i:
Covariance

$$\text{Cov(before, after)} = \text{Var}(p_i) = \sigma_p^2$$

Correlation

$$\text{Correlation} = \frac{\sigma_p^2}{\sqrt{(\sigma_p^2 + \sigma_e^2)(\sigma_p^2 + \sigma_e^2)}} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_e^2}$$

1d. The 95% Prediction Interval for Log Titer in Females After Vaccination
For the prediction of an individual observation:

$$Y_{ijk} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + p_i + e_{ijk}$$

The mean prediction is:

$$\text{Mean prediction} = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

The variance of the prediction:

$$\text{Var(prediction)} = \sigma_p^2 + \sigma_e^2$$

The 95

$$\text{Mean prediction} \pm 1.96 \times \sqrt{\sigma_p^2 + \sigma_e^2}$$

These calculations provide the estimates required based on the model provided in the query.

2 Question 2

To determine the standard deviation of the errors (SDe) needed for the model to predict weight loss within ±5 lbs for 95% of subjects and to find the corresponding R-square value, follow these steps:
Determining the Required Standard Error of the Estimate (SDe)
The investigators want the prediction interval (PI) to cover ±5 lbs around the predicted weight loss for 95% of the subjects. A 95% prediction interval for a normally distributed variable corresponds approximately to ±1.96 standard deviations from the mean.
For the interval to be ±5 lbs, we use the formula for the width of the prediction interval, which is centered around the predicted value $\hat{y}$:

$$\hat{y} \pm z \times \text{SDe}$$

Given $z = 1.96$ (for 95% confidence) and the desired half-width of the interval 5 lbs, we set up the equation:

$$1.96 \times \text{SDe} = 5$$

Solving for SDe:

$$\text{SDe} = \frac{5}{1.96} \approx 2.55 \text{ lbs}$$

This means the standard error of the estimate needed to achieve the desired precision is approximately 2.55 lbs.
Calculating the Corresponding R-square Value
The R-square value in a regression context is a measure of how well the variability in the dependent variable (weight loss) is explained by the model. It is calculated as:

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

The variance of the residuals (Var(residuals)) is SDe^2 and the variance of Y (Var(Y)) is the variance in weight loss, which is given as 40 $\text{lbs}^2 = 1600$ lbs^2.

$$R^2 = 1 - \frac{\text{SDe}^2}{1600}$$
$$R^2 = 1 - \frac{(2.55)^2}{1600}$$
$$R^2 = 1 - \frac{6.5025}{1600}$$
$$R^2 = 1 - 0.004064$$
$$R^2 \approx 0.9959$$

This R-square value suggests that the model explains approximately 99.59
These calculations provide the required standard error of the estimate and the corresponding R-square value to achieve the desired prediction accuracy for the weight loss model.

3 Question 3

3a. The -log Likelihood Function and MLE of λ
The density function of an exponential distribution is given by:

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

For n independent observations $y_1, y_2, \ldots, y_n$, the likelihood function $L(\lambda)$ is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$

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

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

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

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

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

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

This expression $\frac{n}{\sum_{i=1}^{n} y_i}$ is the reciprocal of the sample mean $\bar{y}$, thus $\hat{\lambda}_{\text{MLE}} = \frac{1}{\bar{y}}$.

3b. Variance of the MLE
For exponential distribution, the variance of the MLE can be derived from the information in the log-likelihood function:

$$I(\lambda) = -E\left[\frac{d^2\ell}{d\lambda^2}\right]$$

Where:

$$\frac{d^2\ell}{d\lambda^2} = -\frac{n}{\lambda^2}$$

$$I(\lambda) = -E\left[-\frac{n}{\lambda^2}\right] = \frac{n}{\lambda^2}$$

The variance of the MLE is then:

$$\text{Var}(\hat{\lambda}) = \frac{1}{I(\hat{\lambda})} = \frac{\lambda^2}{n} = \frac{1}{n}\left(\frac{1}{\bar{y}}\right)^2$$

3c. 95% Confidence Bounds for λ
The standard error of $\hat{\lambda}$ is:

$$\text{SE}(\hat{\lambda}) = \sqrt{\text{Var}(\hat{\lambda})} = \sqrt{\frac{1}{n}\left(\frac{1}{\bar{y}}\right)^2} = \frac{1}{\bar{y}\sqrt{n}}$$

For 95% confidence bounds:

$$\hat{\lambda} \pm 1.96 \times \text{SE}(\hat{\lambda}) = \frac{1}{\bar{y}} \pm 1.96 \times \frac{1}{\bar{y}\sqrt{n}}$$

$$= \frac{1}{\bar{y}} \pm \frac{1.96}{\bar{y}\sqrt{n}}$$

These calculations give the -log likelihood function, the MLE of λ, its variance, and the confidence bounds for λ based on the data from n independent observations of an exponential distribution.

4 Question 4

Let's address each of the questions based on the details provided about the two models for predicting the success of removing a respirator from hospitalized COVID-19 patients:

4a. Are These Factorial or Repeated Measure (Mixed) Models?

Neither Model A nor Model B described in the scenario are explicitly stated to be mixed models. They appear to be standard logistic regression models used to predict an outcome based on several predictors (gender, age, hemodialysis status, and number of comorbidities). There is no mention of random effects or subject-specific variations that would characterize repeated measure or mixed models.

4b. Are the Regression Coefficients for Female in the Two Models Necessarily the Same?

No, the regression coefficients for "female" in the two models (a1 and b1) are not necessarily the same. This is because the structure of the models differs:

- - Model A includes age and number of comorbidities as linear predictors without any interaction with age.
- - Model B includes age squared (b4 $\times$ age^2) and interacts number of comorbidities directly with female (b6 $\times$ female $\times$ numcor), which can change the relationship of gender with the outcome, potentially modifying the effect of the "female" variable.

4c. Are the Regression Coefficients for the Female * Numcor Interaction in the Two Models (as a4 and b6) Necessarily the Same?

The coefficients a4 in Model A and b6 in Model B for the interaction between female and number of comorbidities (numcor) are also not necessarily the same. In Model A, the interaction term a4 captures the combined effect of being female and the number of comorbidities linearly, whereas in Model B, b6 captures this interaction under a model structure that also includes age squared. The presence of additional terms and their interactions in Model B can change the estimated effects compared to Model A.

4d. Likelihood Ratio Test for Age's Effect Being Linear vs Quadratic

To test whether the effect of age on the outcome is linear or quadratic, a likelihood ratio test (LRT) can be used, comparing Model A (linear age effect) and Model B (including a quadratic age term).

Calculation: The likelihood ratio test statistic is given by:

$$\chi^2 = -2(\text{log likelihood of simpler model} - \text{log likelihood of complex model})$$

$$\chi^2 = -2((-1061.4) - (-1038.9)) = -2(-22.5) = 45$$

Degrees of freedom for the LRT are calculated as the difference in the number of parameters between the models. Model B has one additional parameter (age squared), so:

$$df = 1$$

This statistic follows a chi-square distribution. A value of 45 is highly significant, suggesting that adding the quadratic term for age significantly improves the model fit.

Consistency with AIC and BIC:

- - AIC for Model A: 1073.4
- - AIC for Model B: 1052.9
- - BIC for Model A: 1107.0
- - BIC for Model B: 1092.1

Both AIC and BIC are lower for Model B, indicating that it provides a better fit to the data, consistent with the results of the LRT, supporting a non-linear relationship between age and the logit of success.

4e. In Model B, if $b6 = 0$ (Statements about numcor and gender):

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

True. If $b6 = 0$, it means there is no interaction term between numcor and female, suggesting that their effects on the logit (log odds) are additive.

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

True. In logistic regression, additive effects on the logit scale translate into multiplicative effects on the odds scale.
iii. The effect of numcor and gender on the model-based probability of success is multiplicative:
True. Since the odds effects are multiplicative, and since the probability is a function of odds ($P = \frac{odds}{1+odds}$), the effect on the model-based probability remains multiplicative.
iv. The effect of numcor and gender on the model-based probability is additive:
False. The probabilities do not add up but rather are transformed through the logistic function, leading to a multiplicative interaction of their odds.
4f. Expression for the Odds Ratio of Success in Females Compared to Males (Model B)
The odds of success for females compared to males, when no interaction terms involving gender are significant ($b6 = 0$), can be directly derived from the coefficient of the female variable ($b1$):
Odds Ratio (OR) for females vs. males:

$$OR = e^{b1}$$

In logistic regression, the coefficient of a binary variable (such as gender) represents the log of the odds ratio for the categories defined by the variable (here, female vs. male).
4g. Odds and Probability of Success for 40-year-old Males with Hemodialysis and No Comorbid Conditions (Model A)
Given:

- - Model A: logit $= a0 + a1 \cdot \text{female} + a2 \cdot \text{hemo} + a3 \cdot \text{age} + a4 \cdot \text{numcor} + a5 \cdot \text{female} \cdot \text{numcor}$
- - Variables: female $= 0$, hemo $= 1$, age $= 40$, numcor $= 0$

Logit for the specified male:

$$\text{logit} = a0 + a2 \cdot 1 + a3 \cdot 40$$
$$\text{logit} = -14.099 + (-0.097) + (0.333 \times 40)$$
$$\text{logit} = -14.099 - 0.097 + 13.32$$
$$\text{logit} = -0.876$$

Odds of Success:

$$\text{Odds} = e^{-0.876} \approx 0.416$$

Probability of Success:

$$P = \frac{\text{Odds}}{1 + \text{Odds}} = \frac{0.416}{1 + 0.416} \approx 0.294$$

4h. Expression for Odds Ratio for Hemo and 95% Confidence Bounds (Model A)
Odds Ratio for Hemo:

$$OR = e^{a2} = e^{-0.097} \approx 0.907$$

Variance and Confidence Bounds for Log(Odds): Since $a2$ has a standard error $SE = 0.028$, the variance of $a2$ is $SE^2 = (0.028)^2 = 0.000784$.
95% Confidence Interval for the log(odds):

$$\text{CI} = a2 \pm 1.96 \times SE = -0.097 \pm 1.96 \times 0.028$$
$$\text{CI} = [-0.097 - 0.05488, -0.097 + 0.05488]$$
$$\text{CI} = [-0.15188, -0.04212]$$

Converting back to Odds Ratio:

$$\text{CI for OR} = [e^{-0.15188}, e^{-0.04212}] \approx [0.858, 0.958]$$

This provides a detailed step-by-step solution to each part of the query based on logistic regression modeling for the scenario described.
Let's address the questions based on the logistic regression results provided:
4i. Odds Ratio for Females with One Comorbidity
Using Model A, where the interaction between the female gender and the number of comorbidities is represented by the coefficient $a5$:
Given:

- - **female = 1**
- - **numcor = 1**

The logit model would be:

$$\text{logit} = a0 + a1 \cdot \text{female} + a4 \cdot \text{numcor} + a5 \cdot \text{female} \cdot \text{numcor}$$

$$\text{logit} = a0 + a1 + a4 + a5$$

The odds ratio for a female with one comorbidity compared to a male with no comorbidities is then calculated by exponentiating the sum of the coefficients for female and the interaction term:

$$\text{OR} = \exp(a1 + a5)$$

Calculating this with provided coefficients:

- - $a1 = -0.253$ (female)
- - $a5 = 0.204$ (female*numcor interaction)

$$\text{OR} = \exp(-0.253 + 0.204) = \exp(-0.049) \approx 0.952$$

95% Confidence Bounds If the standard errors (SEs) of these coefficients were provided, you would typically compute the variance of this log(odds ratio) sum as:

$$\text{Var}(a1 + a5) = \text{SE}(a1)^2 + \text{SE}(a5)^2 + 2 \times \text{Cov}(a1, a5)$$

Here, we assume the covariance is negligible (or zero if $a1$ and $a5$ are assumed independent). Without specific standard errors or covariance data, precise confidence bounds cannot be computed directly.

4j. Comment on Model Fit with Hosmer-Lemeshow Test

The Hosmer-Lemeshow test is a goodness-of-fit test used in logistic regression to determine if the observed event rates match expected rates across different subgroups. A significant Hosmer-Lemeshow test (p-value = 0.041) suggests a poor fit, which might imply that the model does not adequately describe the observed data. This is a valid concern as it suggests that the model might not account well for the data complexities or subgroup variations.

4k. Interpretation of Female Gender p-value

A p-value of 0.2629 for the "female" term indicates that the effect of being female on the outcome (success of respirator removal) is not statistically significant at the 0.05 level. However, this does not mean that the gender has no effect whatsoever; it only means that the effect is not statistically detectable with this particular model and sample. It is essential to consider other potential factors or to gather more data. Also, the non-significant effect in the presence of significant interactions (like female*numcor) suggests that the effect of gender might be conditional or complex, requiring a nuanced interpretation rather than outright dismissal based on the p-value alone.

This interpretation aligns with statistical best practices that caution against equating "non-significant" with "no effect," especially in complex models or in the presence of interaction effects.

5 Question 5

5a. Linear Regression - Linear regression models the relationship between a continuous dependent variable and one or more independent variables using a linear function. However, when the dependent variable is time-to-event, which can include censored data (e.g., patients still on respirators at study end), linear regression is not suitable as it cannot handle censoring or the typically skewed distribution of time-to-event data.

5b. Analysis of Variance (ANOVA) - ANOVA is used to compare the means of three or more samples (using F-tests). It is not appropriate for time-to-event data and does not handle censoring, making it unsuitable for modeling the time until an event occurs.

5c. Logistic Regression - Logistic regression is used for binary outcomes (e.g., success vs. failure) and does not accommodate time-to-event data directly. It's unsuitable for analyzing time until an event unless the time is discretized into a binary outcome, which could lead to loss of information.

5d. Cox (Proportional Hazards) Regression - Cox proportional hazards regression is specifically designed for time-to-event data analysis. It models the hazard rate as a function of several covariates and allows for censoring, making it ideal for studies where you want to assess the impact of various factors on the timing of an event (like respirator removal). This model does not assume a particular statistical distribution for survival times, which adds flexibility.
5e. Poisson Regression - Poisson regression is generally used to model count data or the number of events happening in a fixed period of time, under the assumption that these events happen with a known constant mean rate and independently of the time since the last event. It is less suited for time-to-event data, particularly if the event is not recurring on an individual basis.
Recommended Model: Cox Proportional Hazards Regression (5d) - Given the nature of the outcome (time to respirator removal), the Cox model is the most appropriate. It can effectively handle the types of explanatory variables mentioned (gender, age, hemodialysis, number of comorbidities) and accounts for right-censoring, which is likely present if some patients are not weaned off the respirator by the study's end or if the study period ends before weaning.
The Cox model's capacity to provide hazard ratios for each covariate makes it particularly useful for interpreting how each factor influences the time to respirator removal, allowing healthcare professionals to understand better which characteristics are associated with faster or slower weaning times.

6 Question 6

6a. ROC Area - ROC (Receiver Operating Characteristic) area measures the ability of a model to distinguish between classes. It is commonly used in binary classification to evaluate model performance but is not a method for variable selection.
6b. Likelihood Ratio Tests - Likelihood ratio tests compare the goodness-of-fit between two models—one with and one without a specific variable—to determine if the variable contributes significantly to the model. This method is directly used for variable selection by testing the significance of variables.
6c. Forward / Backward Stepping - Forward and backward stepping are forms of stepwise regression where variables are added or removed from the model based on certain criteria such as AIC, BIC, or p-values. These are classic methods for variable selection and retention.
6d. LOESS - LOESS (Locally Estimated Scatterplot Smoothing) is a non-parametric method that fits multiple regressions in local neighborhoods. Although it's useful for data smoothing and exploring non-linear relationships, it is not a variable selection method.
6e. Restricted Cubic Splines - Restricted cubic splines are used to model non-linear relationships in the data. They are more about modeling the form of the relationship rather than selecting or eliminating variables.
6f. LASSO - LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model. LASSO can shrink coefficients to zero, effectively performing variable selection.
6g. Leverage - Leverage identifies points that have a significant effect on the estimation of coefficients. While it is important for diagnosing model influence, it is not used directly for variable selection.
6h. Outlier Identification with Mahalanobis Distance - Mahalanobis distance is a measure used to identify outliers in a dataset. Identifying outliers is crucial for model diagnostics and can influence decisions on variable transformations or removal, but it's not a direct variable selection method.
6i. Density Estimation - Density estimation involves constructing an estimate of the distribution of a random variable. It is not a method for variable selection.
6j. Missing Data Imputation - Missing data imputation is a process to handle missing values in a dataset and does not contribute directly to variable selection. It prepares data for analysis but doesn't select variables.
6k. Random Forest - Variable Importance - Random forest variable importance measures derived from random forest models can indicate the strength and significance of predictors. This approach is commonly used for variable selection, particularly in complex datasets where relationships between variables may be non-linear or involve interactions.
Summary: Directly Useful for Variable Selection or Elimination-Retention

- - Likelihood Ratio Tests (6b)

- - Forward / Backward Stepping (6c)
- - LASSO (6f)
- - Random Forest - Variable Importance (6k)

These methods specifically target the inclusion or exclusion of variables in models based on their contribution to the model's predictive power or statistical significance.