Biomath 204 2023-2024 Practice Exam

SIMON LEE

July 2024

1 Question A

A(i). To express SD_Y as a function of β_1, SD_X, and SD_e:
In the linear model $Y = \beta_0 + \beta_1 X + e$, we can use the property of variance for independent variables:

$$\begin{aligned} Var(Y) &= Var(\beta_0 + \beta_1 X + e) \\ &= Var(\beta_1 X) + Var(e) \quad \text{(since } \beta_0 \text{ is constant and } X \text{ and } e \text{ are independent)} \\ &= \beta_1^2 Var(X) + Var(e) \\ SD_Y^2 &= \beta_1^2 SD_X^2 + SD_e^2 \end{aligned}$$

Therefore,

$$SD_Y = \sqrt{\beta_1^2 SD_X^2 + SD_e^2}$$

A(ii). To express ρ as a function of β_1, SD_X, and SD_Y:
We're given that $\beta_1 = \rho \frac{SD_Y}{SD_X}$. To solve for ρ, we can rearrange this equation:

$$\begin{aligned} \beta_1 &= \rho \frac{SD_Y}{SD_X} \\ \rho &= \beta_1 \frac{SD_X}{SD_Y} \end{aligned}$$

This gives us the expression for ρ in terms of β_1, SD_X, and SD_Y.

2 Question B

(i) The median of $\log_e(Y)$ is μ since $\log_e(Y)$ has a normal distribution with mean μ.

(ii) Using the delta method, the approximate expected value of Y is:

$$\mathbb{E}[Y] \approx g(\mathbb{E}[X]) \qquad = g(\mu) \ = e^{\mu}$$

The approximate variance of Y is:

$$\text{Var}(Y) \approx [g'(\mathbb{E}[X])]^2 \text{Var}(X) \qquad = [e^{\mu}]^2 \sigma^2 \ = e^{2\mu}\sigma^2$$

(iii) If Y has a normal distribution, then $\log_e(Y)$ would have a normal distribution. However, the problem states that $\log_e(Y)$ has a normal distribution, so we cannot conclude that Y itself has a normal distribution. In fact, Y has a log-normal distribution, not a normal distribution.

3 Question C

1. What are the mean glucose levels at baseline and post drug for each group?

   For the standard drug group: Baseline: $\bar{x}_1 = 150$ mg/dL Post drug: $\bar{x}_2 = 150 - 10 = 140$ mg/dL For the new drug group: Baseline: $\bar{x}_3 = 150$ mg/dL Post drug: $\bar{x}_4 = 150 - 15 = 135$ mg/dL

2. Is this a factorial ANOVA or a repeated measure ANOVA?

   This is a repeated measure ANOVA, because the same subjects were measured at two time points

   (baseline and post drug) for each drug group.

   A factorial ANOVA would be if there were different subjects in each group and time point combination.

- Is this a balanced design (yes or no)?

  Yes, this is a balanced design because there are equal numbers of subjects in each group ($\sigma_1 = 3$, $\sigma_2 = 4$, total $\sigma = 5$).

- What kind of coding is being used, effect coding or dummy coding?

  The coding used here is effect coding, not dummy coding. We can tell because: Newdrug = -1 for baseline, 1 for post drug time = -1 for baseline, 1 for post drug In dummy coding, the baseline would typically be coded as 0 instead of -1.

- Are the factors orthogonal (yes or no)?

  Yes, the factors are orthogonal. Each factor has two levels, and all combinations of levels are present (standard drug at baseline and post drug, new drug at baseline and post drug).

- What is the mean change (post - baseline) for those taking the standard drug and its standard error? What is the mean change (post-baseline) for those taking the new drug and its standard error?

  For the standard drug: Mean change $= \bar{x}_2 - \bar{x}_1 = 140 - 150 = -10$ mg/dL

  For the new drug:

  Mean change $= \bar{x}_4 - \bar{x}_3 = 135 - 150 = -15$ mg/dL

  Standard errors are not provided in the question stem.

- What is the mean difference between these two changes and its standard error? Mean difference = New drug mean change - Standard drug mean change

  $$= (-15) - (-10) = -5 \text{ mg/dL}$$

  The standard error of this difference is not provided in the question stem.

4 Question D

To determine the best model choice and calculate the number of parameters, we need to analyze the given information:

- Dependent variable (Y): Number of colds elementary school students have in the winter (count data: 0, 1, 2, 3, 4, 5, ...)
- Independent variables: Gender, age, number of siblings, days in school

1. Best model choice

Given that Y is a count variable, the most appropriate model would be:
v. Factorial Poisson regression
This is because:

- Poisson regression is suitable for count data
- Factorial design allows us to examine the main effects and interactions of multiple independent variables

2. Number of parameters (βs) if all interactions are included

To calculate the number of parameters:

1. Count the number of main effects:
   - Gender (2 levels): 1 parameter
   - Age (assume continuous): 1 parameter
   - Number of siblings (assume continuous): 1 parameter
   - Days in school (assume continuous): 1 parameter
2. Count the number of interaction terms:
   - 2-way interactions: $\binom{4}{2} = 6$
   - 3-way interactions: $\binom{4}{3} = 4$
   - 4-way interaction: $\binom{4}{4} = 1$
3. Add the intercept: 1 parameter

Total number of parameters:

$$\begin{aligned} \text{Total} &= \text{Intercept} + \text{Main effects} + \text{2-way} + \text{3-way} + \text{4-way} \\ &= 1 + 4 + 6 + 4 + 1 \\ &= 16 \end{aligned}$$

Therefore, the total number of parameters (βs) in the model if all interactions are included is 16.

5 Question E

(i) β_1 will not necessarily be the same as γ_1. While both coefficients relate to age, β_1 represents a linear effect of age in Model A, whereas γ_1 represents the effect of the restricted cubic spline function of age in Model B. These can differ substantially in their values and interpretations.

(ii) Based on the model comparisons, there is some evidence that the relation between the logit and age is non-linear, but it's not particularly strong. The AIC for Model B (4012.0) is slightly lower than for Model A (4004.9), suggesting a marginally better fit for the non-linear model. However, the difference is small (7.1), which doesn't provide strong evidence for non-linearity. A formal test statistic is not directly available from the information given. To properly test this, we would need additional information such as the degrees of freedom for both models to conduct a likelihood ratio test.

(iii) Based on Model A, the expression for the BC odds ratio for those with a family history who smoke compared to those without a family history who do not smoke is:

$$\text{OR} = \exp(0.030 + 0.050 - 0.015) = \exp(0.065) \approx 1.067$$

This odds ratio will be affected by age and Hispanic status. The full expression including these factors is:

$$\text{OR} = \exp(0.030 + 0.050 - 0.015 + 0.020 \times \text{age} - 0.20 \times \text{Hispanic})$$

where "age" is the person's age in years and "Hispanic" is 1 for Hispanic individuals and 0 for non-Hispanic individuals.

(iv) The odds ratio for a 5-year increase in age is:

$$\text{OR} = \exp(5 \times 0.020) = \exp(0.100) \approx 1.105$$

The 95% confidence interval can be computed using:

$$\exp(5 \times 0.020 \pm 1.96 \times 5 \times 0.005)$$

$$\approx \exp(0.100 \pm 0.049) = (1.053, 1.161)$$

(v) For a 60-year-old non-Hispanic woman who smokes but has no family history:

$$\text{logit(BC)} = -3 + 0.020 \times 60 + 0.050 = -1.75$$

$$\text{odds} = \exp(-1.75) \approx 0.174$$

$$\text{risk} = \frac{0.174}{1 + 0.174} \approx 0.148 \text{ or } 14.8\%$$

(vi) Among those with a family history, the odds of BC is higher in smokers. The difference in log-odds is:

$$0.050 - 0.015 = 0.035$$

This positive value indicates higher odds for smokers.

(vii) Based on the given information:

b. The model does not account for all the variation in breast cancer

This is true because the Cox-Snell R-square is 0.12, indicating that the model explains only 12% of the variation.

(viii) For a non-Hispanic, non-smoker who is 60 years old and has a family history:

$$\text{logit} = -3 + 0.020 \times 60 + 0.030 = -1.77$$

$$\text{odds} = \exp(-1.77) \approx 0.170$$

$$\text{risk} = \frac{0.170}{1 + 0.170} \approx 0.145 \text{ or } 14.5\%$$

(ix) If this is a case-control study, it is not valid to compute absolute risk of BC from this model. Case-control studies can provide odds ratios but not absolute risks, as the sampling is based on the outcome rather than being representative of the population. The intercept in a case-control logistic regression does not reflect the true baseline risk in the population.

6 Question F

F1. The correct answer is b. it is more than 5 years.

Explanation: Given that S(5) = 0.75, which means 75% of patients are free of prostate cancer recurrence or death at 5 years, the median survival time (when 50% of patients have experienced the event) must be greater than 5 years.

F3. The hazard ratio (HR) for 70-year-old smokers compared to 50-year-old non-smokers is:

$$HR = \exp(0.04 \times (70 - 50) + 0.20 \times (1 - 0)) = \exp(0.8 + 0.2) = \exp(1) \approx 2.72$$

F4. The hazard ratio for a "k" year increase in age where smoking is constant is:

$$HR = \exp(0.04k)$$

This formula does not depend on smoking status, as the smoking term cancels out when comparing two individuals with the same smoking status.

F5. For those age 70 who smoke:

$$\text{2-year risk} = 1 - S_0(2)^{\exp(0.04\times 70+0.20)}$$
$$\text{5-year risk} = 1 - S_0(5)^{\exp(0.04\times 70+0.20)}$$

For those age 50 who do not smoke:

$$\text{2-year risk} = 1 - S_0(2)^{\exp(0.04\times 50)}$$
$$\text{5-year risk} = 1 - S_0(5)^{\exp(0.04\times 50)}$$

Where $S_0(t)$ is the baseline survival function at time t, which is not provided in the given information.