Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(Take-home portion)

May 23–24, 2024

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Thursday, May 23 at 9:00am, and will be due on Friday, May 24 at 5:00pm. This deadline is strict: late submissions will not be accepted.
- To turn in your exam, please use your assigned Box folder and e-mail your word-processed exam to Dr. Andrew Spieker by the deadline. This level of redundancy is designed to ensure that your exam is received by the deadline. If you would like to e-mail exam drafts along the way, that is perfectly acceptable—do not be concerned about spamming my inbox.
- There are two problems (Problems 6 and 7). Note that not all questions and sub-questions are weighted equally. You are advised to pace yourself and to not spend too much time on any one problem.
- Answer each question clearly and to the best of your ability. Partial credit will be awarded for partially correct answers.
- Be as specific as possible in your responses.
- You may consult reference material (e.g., course notes, textbooks), though the work you turn in must be your own (this means no generative AI). This is an *individual effort.* Do not communicate about the exam with anyone. Vanderbilt University's academic honor code applies.
- Please direct clarifying questions by e-mail to Dr. Andrew Spieker and Dr. James Slaughter.

---

6. $\boxed{\text{35 pts}}$ Prevention of obesity during childhood is critical for children in underserved populations, for whom obesity prevalence and risk of chronic disease are highest. The objective of this analysis is to estimate associations between risk factors that contribute to the emergence of childhood obesity among low-income minority children. You were supplied with the file `pediatricbmi.csv`, which includes data from a prospective cohort study in which $n = 610$ parent-child pairs participated in a three-year study of a healthy lifestyle behavioral intervention. In the present analysis, we will not be considering the intervention, but will instead focus on the association of baseline risk factors with BMI measured 36 months later. Note that there is disagreement regarding how well BMI measures obesity in young children. For the purposes of this question, however, you are permitted to focus on BMI, as reported, as a measure of obesity. The following is a set of general instructions for the questions that follow:

(I) You are expected to use statistical software (R, in particular) to answer the set of questions below. You must include the syntax used to generate the results as a clearly labeled and organized `.R` file. A knitted .html file with a .Rmd or .Qmd file is acceptable, provided your code is made available and we do not have to run it in order to see your submission.

(II) In each response, summarize your approach in a way that your findings could be closely reproduced by an independent but statistically savvy investigator.

(III) When presenting regression results, scale covariates to have a meaningful interpretation as needed.

(IV) If your analysis relies on specific assumptions, state the assumptions, briefly defend them, and (where possible) assess possible violations to those assumptions.

(V) Specifically for parts (e) and (f), your models do not need to address missing data using the most principled/statistically efficient approaches. However, you should characterize extent of missingness in other parts of the problem as appropriate.

(VI) Do not present the results of every single figure you generate and every single analysis you try. Be judicious in your presentation.

(a) Perform univariate descriptive analysis of the variables measured at enrollment (demographic, BMI, and others). Use this descriptive analysis to summarize the sample.

(b) Using the descriptive analysis, identify any aspects of maternal BMI (`mother.bmi`) that may present statistical issues. Describe any issue you identify.

(c) Suppose a colleague was interested in estimating the association between age at enrollment (`age.enroll`) and BMI at 36 months (`bmi.36`). Based on the descriptive statistics, what statistical issues do you have with this proposed analysis?

(d) Create and interpret a descriptive plot of the bivariate association between `mother.bmi` and `bmi.36`. Identify any unusual points, if they exist.

(e) Fit an (unadjusted) regression model to estimate the association between `mother.bmi` and `bmi.36`. Interpret the results of the model you fit.

(f) Fit an (adjusted) regression model that additionally controls for baseline BMI of the child (`bmi.0`). Interpret the results of the model you fit.

(g) Compare the results of the unadjusted and adjusted model for the `mother.bmi` coefficient. Comment on whether you believe your analyses are consistent with the idea that baseline BMI confounds the association between maternal BMI and BMI of the child at 36 months.

CODEBOOK FOR PROBLEM 6

`id`	Study ID
`bmi.36`	BMI of child at 36 months (kg/m^2)
`bmi.enroll`	BMI of child at enrollment (kg/m^2)
`mother.bmi`	BMI of mother at enrollment (kg/m^2)
`male`	Male sex (0=Female, 1=Male)
`birthweight`	Weight of child at birth (grams)
`gestage`	Gestational age of child at birth (weeks)
`hispanic`	Self-reported Hispanic ethnicity
`mother.depress`	Maternal depression score at enrollment (higher scores are indicative of more depression)
`food.security`	Measure of food security at enrollment (1=low, 2=moderate, 3=high)
`age.enroll`	Age of the child at enrollment (months)

7. $\boxed{\text{15 pts}}$ A study consists of an exposed group and an unexposed group, each of sample size n. For each observation (all independent), we measure a binary outcome. Let Y_0 and Y_1 denote the number of responses among the unexposed and exposed groups, respectively. The logistic regression model is natural for these data: $Y_0 \sim \text{Binomial}(n, p = \text{expit}(\beta_0))$ and $Y_1 \sim \text{Binomial}(n, p = \text{expit}(\beta_0 + \beta_1))$; you may assume $0 < Y_0, Y_1 < n$.

(a) Show that the maximum likelihood estimator (MLE) for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is given by:

$$\begin{aligned} \widehat{\beta_0} &= \text{logit}(Y_0/n), \text{ and} \\ \\ \widehat{\beta_1} &= \text{logit}(Y_1/n) - \text{logit}(Y_0/n). \end{aligned}$$

(b) The odds ratio places the measure of association between the exposure and outcome on a scientifically meaningful scale and can be expressed as $\varphi = \exp(\beta_1)$. Use the result of part (a) to show that the MLE for the odds ratio of interest is given by

$$\widehat{\varphi} = \frac{Y_1}{Y_0} \frac{(n - Y_0)}{(n - Y_1)}.$$

(c) Although you do not need to show it, standard likelihood theory asserts that:

$$\sqrt{n}(\widehat{\beta_1} - \beta_1) \xrightarrow{d} \mathcal{N}(0, v(\boldsymbol{\beta})),$$

where

$$v(\boldsymbol{\beta}) = \frac{(1 + \exp(\beta_0 + \beta_1))^2}{\exp(\beta_0 + \beta_1)} + \frac{(1 + \exp(\beta_0))^2}{\exp(\beta_0)}.$$

Use this knowledge together with the delta method to determine an expression for an asymptotically valid symmetric 95% confidence interval (CI) for φ.

(d) The above formulation of a CI for an odds ratio does not reflect usual practice. Typically, we first form a symmetric CI for β_1 based on standard likelihood theory and then exponentiate its endpoints. Your task is to conduct a simulation study in which you determine and compare the coverage properties for each method in a variety of scenarios pertaining to strength of association and group-specific sample size. Let $M = 10{,}000$ denote the total number of replicates per scenario. The table below characterizes the scenarios in terms of p_0 and p_1, the probability of a response for an individual in the unexposed and exposed groups, respectively; you should determine and fill in the corresponding values of β_0, β_1, and φ. Fill in the rightmost two columns with your simulation-based approximations of the coverage (to three decimal places). **You must turn in your code as a .R file with your submission**.

						Method to determine 95% CI	
p_0	p_1	β_0	β_1	φ	n	Delta method	Exponentiate endpoints
0.50	0.50				25		
0.50	0.50				100		
0.50	0.50				1000		
0.25	0.75				60		
0.25	0.75				100		
0.25	0.75				1000		

(e) Briefly describe your findings from the simulation study. Account for patterns of note; you may use graphical methods to display simulation results that bolster your arguments but you do not need to include mathematical justification in your response.

(f) You were not asked to conduct a simulation under the combination of parameters $p_0 = 0.25$, $p_1 = 0.75$, and $n = 25$. Speculate on why this is (*Hint*: try it and see what goes wrong). You should justify your answer (e.g., mathematically) as part of your response.
