Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(Take-home portion)

May 25–26, 2023

Instructions: Please adhere to the following guidelines:

- This exam is scheduled to be administered on Thursday, May 25 at 9:00am, and will be due on Friday, May 26 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 and Dr. Simon Vandekar 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 our inboxes.
- 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 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.

---

6. [30 pts] A group of investigators sought to compare assays to quantify SARS-CoV-2 antibody levels in the blood. They compared a titer-based measure to an immunofluorescence measure in $N = 1024$ independently sampled individuals with recent SARS-CoV-2 infection. The titer is obtained similarly to the hemagglutination inhibition titer for influenza and takes on values of 10, 20, 40, ..., 2560; adjacent titer categories differ by a single dilution. The typical range of values for the immunofluorescence assay is 5-300 immunofluorescence units (IU). Both measures are defined on a multiplicative scale, with higher values signifying higher antibody levels. The data are provided to you in the file `assay.csv`, which contains the following variables:

`id`	unique study participant identifer
`titer`	value for the titer assay (10, 20, 40, ..., 2560)
`immunofl`	value for the immunofluorescence assay (5-300 IU, continuous)

Please approach the below questions carefully. You may (should) use statistical software to address them; your code should be attached as a clearly labeled appendix.

(a) Characterize the distribution of each assay individually using appropriate univariate descriptive statistics.

(b) Generate two scatter plots: (1) log-transformed immunofluorescence on the y-axis and untransformed titer on the x-axis; (2): log-transformed immunofluorescence on the y-axis and log-transformed titer on the x-axis. Briefly discuss and comment on patterns of note.

(c) Consider the following three linear regression models as methods to model the relationship between the two assays, each with log-transformed immunofluorescence as the outcome:

(I) Untransformed titer treated as a nine-category factor/nominal/categorical predictor.

(II) A natural cubic spline model on untransformed titer with knots at $k_1 = 30$, $k_2 = 160$, and $k_3 = 960$.

(III) A simple linear model with log-transformed titer as the predictor.

Use ordinary least squares with robust standard errors to construct point estimates and 95% confidence intervals for the mean log-transformed immunofluorescence at each titer value. Present your results in the format below, which includes a column for the sample size, n, in each titer category:

Titer	n	Model (I)	Model (II)	Model (III)
10				
20				
40				
80				
160				
320				
640				
1280				
2560				

(d) Comment on key patterns in interval width (e.g., across models and across titer groups) you observed in part (c); use your knowledge of regression modeling to account for these patterns.

(e) Perform appropriate diagnostics to investigate and comment on the degree to which you trust the validity of the confidence intervals you formed in part (c).

(f) On the titer scale, a difference of two dilutions between groups is considered the smallest that is clinically meaningful. Your collaborator is interested in estimating a single real-valued parameter that elucidates the degree to which subgroups differing in their titer by this amount compare on the immunofluorescence scale. Which of the three models is best equipped to accomplish this goal? Identify and interpret the corresponding quantity of interest that is provided by this model; then, provide a point estimate and 95% confidence interval for this quantity based on the model you've chosen.

---

7. [20 pts] Consider the setting of a three-group randomized trial involving a total of $N = 60$ independently sampled individuals. The predictor is given by treatment dose ($X = 0, 1, 2$) and the outcome is given by $Y = \alpha_0 + \alpha_1 1(X = 1) + \alpha_2 1(X = 2) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2 = 1)$. Consider four overall data generating mechanisms based on two study design scenarios and two outcome generation scenarios.

Study design:

i. Pure randomization: $(X_1, \ldots, X_{60}) \sim \text{Multinomial}(N = 60, k = 3, \mathbf{p} = (1/3, 1/3, 1/3))$; you can think of this as tossing a fair "three-sided die" to determine the randomization group of each patient.

ii. Fixed randomization: $X_1, \ldots, X_{20} = 0$, $X_{21}, \ldots, X_{40} = 1$, and $X_{41}, \ldots, X_{60} = 2$; you can think of this as a permuted block design so the number in each group is known in advance to be fixed at twenty.

Outcome generation mechanisms:

i. Linear dose-response relationship: $\alpha_0 = 100$, $\alpha_1 = 1$, and $\alpha_2 = 2$.

ii. Non-linear dose-response relationship: $\alpha_0 = 100$, $\alpha_1 = 1$, and $\alpha_2 = 5$.

Your pre-specified analysis plan was to treat X continuously by fitting the simple linear regression model $\text{E}[Y|X = x] = \beta_0 + \beta_1 x$ by ordinary least squares. Your task is to conduct a simulation to compare the following three methods to estimate $\text{Var}(\widehat{\beta}_1)$ from this model under each of the four data generation mechanisms:

(I) The sandwich variance (using `vcovHC(..., type = "HC0")` in R from the `sandwich` package).

(II) The non-parametric bootstrap, re-sampling observations with replacement ($B = 500$ replicates).

(III) The conditional non-parametric bootstrap, which stratifies the re-sampling specifically by treatment group in a way that treats the number of people in each group as fixed ($B = 500$ replicates).

Write R code to assess performance of these three methods for each of the four data generating mechanisms. For each data generating mechanism, I recommend the following organization for your code:

- Set the parameters of your simulation and create a place to store your results.
- **for** m = 1 **to** 1000
  - Generate the data under the parameters of the data generating mechanism.
  - Fit the simple linear regression model and extract $\widehat{\beta}_1$, call it $\widehat{\beta}_1^{(m)}$.
  - Estimate the variance using each method described above, call them $V_{(\text{I})}^{(m)}$, $V_{(\text{II})}^{(m)}$, and $V_{(\text{III})}^{(m)}$.

  end
- Determine the variance of $\widehat{\beta}_1^{(1)}, \ldots, \widehat{\beta}_1^{(1000)}$ across simulation replicates; this is the empirical variance to which you seek to compare the performance of each variance estimator.
- For each variance estimator, determine the average estimated value across the simulation replicates. For instance, $1000^{-1} \sum_{m=1}^{1000} V_{(\text{I})}^{(m)}$ is the average estimated variance based on the sandwich.

Your code must be turned in as clearly labeled supplementary .R file. Further, your code should be fully reproducible; it is best practice to include comments that make clear the goal of key steps in your code.

(a) Use the results of your simulations to fill in the table below.

Outcome generation: Randomization:	Linear Pure	Linear Fixed	Non-linear Pure	Non-linear Fixed
Emperical variance across simulations				
Average variance estimate based on the sandwich				
Average variance estimate based on the full bootstrap				
Average variance estimate based on the conditional bootstrap				

(b) Summarize what this study illustrates about the performance of each method across settings.
