PhD Take Home Qualifying Exam
June 1, 2022
Due June 8, 2022 at 5pm

The take home exam has 4 questions:

Q1: Requires data file: *MVPA_intervention.csv*

Q2. Requires data file: *data-lla.csv*

Q3. No file

Q4. Requires data file: *ATFIB.txt*

Instructions:

Prepare a separate report for each of the 4 questions.

Organize your reports so that your answers are labelled and are easy to follow.

Write or type your answers so that your work is legible.

Unless otherwise noted, limit each report to 12 pages; you can include SAS or R output as needed as an Appendix at the end of the report.

Write your exam number **not your name** on all pages of your reports.

If you have questions about the exam, please email Nichole Carlson (nichole.carlson@cuanschutz.edu)

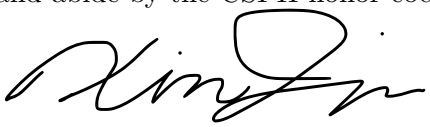
Submission Instructions: Email your reports to rocio.velezpesante@cuanschutz.edu before 5pm on Wednesday June 8, 2022.

- **Work individually on your exam. You may use outside resources, but you cannot discuss this exam with anyone.**

Please read the honor code and sign your agreement below:

> *I understand that my participation in this examination and in all academic and professional activities as a CSPH student is bound by the provisions of the CSPH Honor Code. I understand that work on this exam and other assignments are to be done independently unless specific instruction to the contrary is provided.*

I agree with and abide by the CSPH honor code:

Signed: 

Q1 A trial was performed to assess the impact of an intervention on the number of times a sedentary individual engages in 10 minutes of sustained moderate-to-vigorous physically active over a week. The study recruited participants in matched pairs (matching on age), assigning the intervention to one individual within each age pair (n total pairs). Denote each pair using the subscript i and let the number of active bouts for the person in the i^{th} pair be denoted as Y_{i1}, Y_{i2}, with Y_{i1} corresponding to individual who received the intervention. Let X_i be the age for individuals in the i^{th} pair. It is hypothesized that

$$Y_{i1}|X_i \sim \text{Poisson}(\gamma\beta(X_i))$$
$$Y_{i2}|X_i \sim \text{Poisson}(\beta(X_i))$$

where $\beta(\cdot)$ is an unspecified function of X_i (age). The investigators of the current study are uninterested in $\beta(\cdot)$. Assume that the distribution of age in the population follows some distribution $f_X(x)$.

If the number of pairs (n) observed is large and X is observed densely relative to the shape of $\beta(\cdot)$ then γ, β may be estimable using the likelihood of the data. However, if n is small or X is sparsely observed, estimating both γ and β may not be possible. This question involves deriving a conditional likelihood where you condition out the nuisance parameter(s) $\beta(\cdot)$ so that γ is estimable in such scenarios. You will implement a short simulation study and apply your method to data from the study referenced above.

(a) In the first part of this problem, we will condition out nuisance parameters (β) to obtain a likelihood for γ which depends only on the observed data.

i. What is the likelihood of $[X_i, Y_{i1}, Y_{i2}]^t|\gamma, \beta(\cdot)$? Your answer will involve the marginal distribution of X, $f_X(x)$.

ii. What is the likelihood of $Y_{i1} + Y_{i2}|X_i, \theta$? Identify the corresponding distribution.

iii. What is the likelihood of $Y_{i1}|Y_{i1} + Y_{i2}, X_i, \theta$? Identify the corresponding distribution.

iv. What is the maximum likelihood estimator of γ obtained from the conditional likelihood you derived in iii?

(b) Propose a method for constructing an asymptotic 95% confidence interval for γ based on your result from (a)iv using maximum likelihood theory (you may use the observed information here).

(c) A statistician colleague of yours suggests constructing of a confidence interval using simulation. Specifically, your colleague proposes the following procedure:

Fix α, the type-I error rate.

Step 1 Choose candidate interval endpoints $\boldsymbol{\gamma}^* = [\gamma_1^*, \gamma_2^*, \ldots, \gamma_S^*]^t$ which is dense over a range of plausible endpoints for the confidence interval.

Step 2 For each $\gamma_s^* \in \boldsymbol{\gamma}^*$

Step 2a Simulate $y_{i1}^s|y_{i1} + y_{i2}, \gamma_s^*$ using the distribution you derived in (a)iii.

Step 2b Calculate $\hat{\gamma}_s^*$ as the MLE you derived in (a)iv.

Step 2c Repeat (1)-(2) many times (e.g. 2000).

Step 2d Let $I_s = 1(\hat{\gamma} \in [\hat{F}^{-1}_{\gamma^*_s}(\alpha/2), \hat{F}^{-1}_{\gamma^*_s}(1-\alpha/2)])$, where $\hat{F}^{-1}_{\gamma^*_s}(\alpha/2)$ and $\hat{F}^{-1}_{\gamma^*_s}(1-\alpha/2)$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of $\hat{\gamma}^*_s$ obtained from Steps 2a-2c, respectively.

Step 3 Define a $1-\alpha\%$ CI for γ as $\left[\min_{\{s:I_s=1\}}(\gamma^*_s), \max_{\{s:I_s=1\}}(\gamma^*_s)\right]$

Argue that your colleague's algorithm is a valid approach for constructing a 95% confidence interval. Write a function which implements this algorithm for arbitrary:

- Type-I error α
- Input vectors $\boldsymbol{y}_1, \boldsymbol{y}_2$.
- Candidate interval endpoints $\boldsymbol{\gamma}^*$
- Number of simulated datasets created in step 2c

(d) Implement a fully reproducible simulation study which assesses the performance of the confidence intervals obtained in (b) and (c). Specifically, implement a simulation study as follows:

- $X \sim \text{Unif}(0, 1)$
- $\gamma = 2.2$
- $\beta(x) = 2e^{-2x}$
- $n = 15$ and $n = 100$
- $\boldsymbol{\gamma}^*$ is an equally spaced grid on $[0, 5.5]$ with interval length of 0.01
- Simulate 2000 datasets in Step 2c of the algorithm from (c)

Estimate 95% confidence intervals using the approaches you derived in both (b) and (c) in 5000 simulated datasets and evaluate power for testing the null hypothesis $H_0 : \gamma = 1$. Compare approaches on coverage probability and power. Summarize your results in 3-4 sentences as if for a statistical journal. Support your claims using 1-2 tables or figures. Provide well commented code which reproduces your simulation study and recreates your tables/figures.

(e) Apply your methods to the dataset "MVPA_intervention.csv", a data matrix which contains two columns, "Y1" and "Y2", corresponding to the number of 10 minute bouts of moderate-to-vigorous physical activity in matched pairs. Estimate α and construct 95% confidence intervals using both of the approaches you derived in (b) and (c). Write a brief, non-technical, summary of both methods as if communicating to a non-statistician, recommend and justify one approach, and conclude with an interpretation of the estimated coefficient and confidence interval.

Q2 **Carry out the steps of the statistical analysis outlined below. Turn in commented statistical code used to obtain the results reported.** Use no more than 3 decimal places.

Lower-limb amputation (LLA) results primarily from complications of severe peripheral artery disease, diabetes mellitus, or trauma. Sustained exercise has been suggested by previous studies as a promising rehabilitation target to improve the long-term health after LLA. Using telehealth, a clinical trial has been conducted at the Veterans Affairs Eastern Colorado Regional Amputation Center to test the potential of sustaining walking exercise using exercise self-management (EXP) versus an attention-control education (CTL) program. The primary outcome is step count per day and a few other secondary outcomes are also of interest including presence of any adverse events, as binary outcome. Outcomes were collected at baseline (time=0), 6 months (time=6) and 18 months (time=18). The primary hypothesis of the trial was that the intervention would improve daily steps in at least 1000 steps. Because randomization was stratified by age (<60 and ≥ 60) and level of amputation (below knee and above knee), these two variables should be adjusted for in the analyses.

The longitudinal dataset of the study is in the table data-lla.csv with data dictionary below.

Variable name	Description
`treat`	treat=EXP for individuals in EXP group, and treat=CTL for individuals in CTL group
stepc	Step count
amp.above	amp.above=1 for amputation above the knee, amp.above=0 for amputation below the knee
time	Month from baseline
older	older=1 for individuals 60 or older, and older=0 for individuals younger than 60 years of age

Table 1: Description of variables in data-lla.csv dataset.

Use R or SAS to answer the following questions regarding the hypothesis above.

(a) The primary hypothesis of the study was to test whether there was a difference of more than 1000 steps in the change of step count/day at 6 months between exposure groups. This particular type of trial is often referred to as a superiority trial by a margin. Carry out this hypothesis test and interpret your results for a clinical collaborator.

i. Write the model formally in mathematical form. Allow for a difference in daily step count at baseline.

ii. Write down the null and alternative hypothesis of interest in terms of the model parameters specified in (a); keep the terms consistent with (a).

iii. Calculate the change (from baseline) in expected means and standard error of the step counts at 6 months for each the intervention and the control group.

iv. Calculate the difference (EXP - CLT) in change of expected means in the two intervention groups. Report a 1-sided 95% confidence interval for this difference.

v. Make a graph, and include any appropriate descriptive statistics, that you would present as evidence related to the primary hypothesis of interest.

vi. Write your conclusion of the hypothesis test in (b).

vii. Comment on your assumptions of the correlation structure for the repeated measurements within subject.

viii. Write a summary paragraph interpreting the results for a clinical collaborator.

(b) A secondary hypothesis of interest was whether or not the change in step counts was sustained from 6 to 18 months, specifically, interest was in testing the difference between M18-M0 and M6-M0 in the intervention group. Carry out the appropriate hypothesis to test whether there was a sustained change and interpret the results for a clinical collaborator.

Q3 Assume that $(Y_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ are independent observations, where Y is binary and $\mathbf{X}$ is a $d-$dimensional vector of covariates . We wish to fit a logistic regression model of Y to $\mathbf{X}$, i.e.,

$$P(Y_i = 1) = \frac{\exp(\mathbf{X}_i'\beta)}{1 + \exp(\mathbf{X}_i'\beta)},$$

where β are the regression coefficients. Note that we are fitting a model without an intercept

(a). Determine a generative model based on the multivariate normal distribution for $\mathbf{X}|Y$ with $d = 3$ and $n = 102$ that leads to estimates of α and β that are either negative infinity or infinity. We will refer to this as a *nonconvergence situation.* List out the following values: the mean vector and variance-covariance matrix of $\mathbf{X}|Y$ and p, the probability of $Y = 1$.

(b). Give a 2-3 sentence description which describes when we can expect nonconvergence.

(c). Give a 2-3 sentence description which describes how nonconvergence relates to multicollinearity.

(d). Modify the model in (a) to achieve convergence. Based on the modified model, fit a logistic regression of y on $\mathbf{X}$ using 100000 observations from $y = 1$ and 100000 observations from $y = 0$. Report the estimated regression coefficients and standard errors.

(e). Perform a simulation study using the generative model in (a). comparing different penalization schemes. As in (a), we will assume a sample size of 103. Please use the following methods: (1) LASSO regression (using the **glmnet** package in R); (2) Firth correction (using the **logistf** package in R). For both methods, you will actually fit the model **with** the intercept. Thus, we are fitting a misspecified model, and the goal is to see how the corrections fare in terms of bias. For the simulation study, specify the following:

- Tuning parameters used for the two penalized regressions;
- The number of simulated datasets being used.

Report on the mean bias for the parameters using each of the methods.

(NB: For the LASSO, you will get a warning message of the form "In lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, ... : one multinomial or binomial class has fewer than 8 observations; dangerous ground"; you can ignore that. For the Firth correction, you will get warning messages of the form "fitted probabilities numerically 0 or 1 occurred for variable V2"; you can disregard those. In the event that any of your simulation values do not converge, you may remove those from the simulation study. Finally, for the Firth correction, you will want to set the maximum number of iterations to be at least 10000).

Q4 You have received data on patients with cardiac monitoring devices that detect atrial fibrillation events. In reporting the counts of atrial fibrillation events within one month, for some patients (Group A) the number of events is available while for the rest (Group B) you only have a binary indicator of whether the count was non-zero. There is no information in the data set that will let you know whether the value presented is a count or a binary indicator (i.e., data from groups A and B are mixed together). The data is available in *ATFIB.txt* and contains the following columns:

- `id:` unique patient identifier
- `outcome:` observed outcome (i.e., the direct actual count of atrial fibrillation events or the binary indicator of a non-zero count)

(a) Given what you know about the data collection, you realize that you could consider the data to arise from a Poisson random variable Y with mean λ for which you observe either Y directly (Group A), or the indicator of $Y > 0$ (Group B). Which group each individual belongs to can be considered missing data. Define a latent variable $Z \sim Bernoulli(p)$, independent of Y, as a binary indicator of whether Y is observed directly (i.e., $Z = 1$ indicates an observation is from Group A and $Z = 0$ indicates an observation is from Group B). The observed outcome X can then be defined in terms of $Y \sim Poi(\lambda)$, $I(Y > 0)$ and Z.

Write out the likelihood for:

i. The observed data (X). *Hint:* Consider the cases (i) $X = 0$, (ii) $X = 1$, (iii) $X = k, k > 1$.

ii. The complete data (Y, Z). *Hint:* Consider the cases for when (i) $Z = 1, Y = k$, (ii) $Z = 0, Y = 0$, (iii) $Z = 0, Y = k, k > 0$

(b) Numerically optimize the observed data log-likelihood using "optim" in R, to obtain the estimates, standard errors, and 95% confidence intervals for λ and p. Provide non-technical interpretations for these parameters. Comment on how you think these results would differ if you had data on Z, and the reason for those differences.

(c) You begin to suspect that maybe the data was actually collected consistently, and only represents counts that arise from a $Poi(\lambda)$ distribution. Perform an appropriate statistical test to formally test this assumption for your data. State the null and alternative hypotheses for your test, and the asymptotic distribution of the test statistic under the null. Present your conclusions based on this test.

(d) Design and conduct a simulation study that evaluates the efficiency of estimation using the incomplete data versus the complete data (i.e., the incomplete data augmented with a column for Z). Specifically, report bias and efficiency for estimating λ. Fix $\lambda = 1$. Vary sample size ($N = 100, 1000$) and p ($p = 0.2, 0.5, 0.8$) to generate data from 6 settings. Simulate a reasonable number of replicates.

i. Submit commented code that describes (i) data generation process, (ii) estimation, and (iii) computation of metrics.

ii. Summarize your results in a short report (2-3 paragraphs) and 1-2 figures or summary tables. Give hypotheses, conclusions, and recommendations based on what you have found.