Vanderbilt University Biostatistics Comprehensive Examination

MS Applied Exam/
PhD Applied Exam Series 1

(In-class portion)

May 22, 2024

Instructions: Please adhere to the following guidelines:

- This exam begins on Wednesday, May 22 at 9:00am. You will have until 1:00pm to complete it.
- There are five problems of varying length and difficulty. Note that not all 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, show your work when necessary, and please write legibly.
- This exam is closed-everything and is an *individual effort*. You are, however, permitted the use of a scientific calculator. Vanderbilt University's academic honor code applies.
- Please direct clarifying questions to the exam proctor.

---

1. [20 pts] **Background**: Circulating antibodies to influenza can be characterized via the *antibody titer*, obtained by sequentially diluting a serum sample and testing each dilution for the antibody of interest. The initial dilution of the original serum sample occurs at a ratio of 1:5, and the sample is diluted by a factor of two until the serum no longer responds. A participant's titer is defined as the relative concentration of the *final* dilution that responds to an antibody test (higher titer values are indicative of a greater concentration of antibodies in the blood). It will be helpful to keep in mind that circulating antibodies can signify ongoing or prior infection (symptomatic or asymptomatic), or even response to prior vaccination.

   A new assay for the antibody titer was developed. As a precursor to clinical use, it is of interest to evaluate the assay in a small "proof-of-concept" pilot study by characterizing how well it can distinguish between those who might and might not be expected to have higher levels of circulating antibodies. To that end, blood was collected from twelve participants: six adults sampled from an urgent care clinic with ongoing infection as confirmed by a gold standard (Group **A**), and six adults responding to an advertisement with no complaints of influenza symptoms (Group **B**). The data are shown in the table below:

Group A:	1280	2560	160	40	2560	1280
Group B:	5	160	320	20	320	160

   As a concrete example of how titers are derived, consider the first patient in Group **A**, whose sample achieved a positive response on the initial dilution and the next eight subsequent dilutions, but a negative response on the ninth dilution, resulting in a titer of $5 \times 2^8 = 1280$. Owing to the multiplicative nature of the antibody titer assay, it is typically considered appropriate to consider the geometric mean as a measure of central tendency rather than the arithmetic mean. As a hint for parts of this problem, you may consider applying the transformation $f(x) = \log_2(x/5)$ to the data to simplify certain calculations (but don't forget to back-transform as appropriate). Below are approximate 95^{th} and 97.5^{th} percentiles of the t-distribution having between four and twelve degrees of freedom (you may not need all of this information).

df	4	5	6	7	8	9	10	11	12
$\boldsymbol{t}_{0.95,\mathbf{df}}$	2.13	2.02	1.94	1.89	1.86	1.83	1.81	1.80	1.78
$\boldsymbol{t}_{0.975,\mathbf{df}}$	2.78	2.57	2.45	2.36	2.31	2.26	2.23	2.20	2.18

   (a) Compute a point estimate and 95% confidence interval for the geometric mean titer in each group.

   (b) Compute a point estimate and 95% confidence interval for the geometric mean titer ratio (i.e., the ratio of the Group **A** geometric mean titer to the Group **B** geometric mean titer). Particularly since you have not been supplied with enough information to do otherwise, please make use of the *pooled* variance—but do also supply a brief statement regarding why you might be comfortable with this choice.

   (c) Let θ denote the geometric mean ratio described in part (b), and consider the goal of testing the hypothesis $H_0 : \theta = 1$ vs. $H_1 : \theta \neq 1$. Make an educated guess at the approximate value of a t-statistic associated with this test based on your calculation in part (b). Your guess does not have to be exact.

   (d) Confirm your suspicions of part (c) by computing a t-statistic for this test.

   (e) Suppose it is of interest to evaluate this test as a diagnostic tool (i.e., as a test for influenza). Using a cut-off of ≥320 to define a positive test, compute point estimates for the test's sensitivity and specificity (i.e., P(Test +|Disease) and P(Test −|Healthy), respectively).

   (f) Is it possible to obtain estimate the positive and negative predictive values (i.e., P(Disease|Test +) and P(Healthy|Test −), respectively) of the test described in part (e)? If so, do so; if not, briefly explain.

   (g) Major limitations of this pilot study, apart from its small sample, lie in its sampling scheme. Comment (in about 4-6 sentences) on some of the limitations of the study design and what barriers they might pose to your ability to answer the key study questions. You are not expected to cover every limitation in your response, nor are you expected to propose a study design that would address these limitations.

2. [20 pts] A study was conducted of $N = 672$ independently sampled children between 5 and 10 years old with asthma. The study objective was to characterize the association between asthma severity as measured by forced expiratory volume (`fev`, which is the volume of air, in liters, that can be blown out of one's lungs in one second) and physical activity (`phys.act`, measured as the average number of hours spent involved in physical activities per week over a six-week period). Consider the following regression model:

$$\mathrm{E}[\texttt{phys.act}|\texttt{fev}] \;=\; \beta_0 + \beta_1 \log(\texttt{fev}).$$

The following page presents important supplementary material for this problem. At the top are two figures: (A) a histogram of FEV, and (B) a scatter plot of log-transformed FEV and physical activity (with the fitted regression line included and a marker for the average log-transformed FEV). Two points, labeled **1** and **2**, are also circled in this plot. At the bottom of the page is the regression output from this model.

(a) State a literal interpretation for β_0 in plain but scientifically precise language. Briefly comment on the degree to which this interpretation is or is not meaningful in the real world.

(b) State an interpretation for $\beta_1 \log(1.5)$ in plain but scientifically precise language. Then,

(i) Recall the conditions often considered in a linear regression model: (1) linearity, (2) constant variance/homoscedasticity, and (3) error normality. Briefly discuss the extent to which each of these must hold for you to trust the point estimate and 95% confidence interval for β_1 presented in the supplementary material.

(ii) Briefly discuss (maximum of three sentences) the extent to which the information provided in the supplementary material demonstrates (or fails to demonstrate) evidence of violations to the assumptions you deemed critical in (b)(i).

(c) Briefly comment on the extent to which you agree or disagree with the following conclusion: "*Greater asthma severity causes children to be less physically active.*"

(d) Briefly comment on the extent to which you agree or disagree with the following statement: "*The regression model called for FEV to undergo a log-transformation due to its right-skewness.*"

(e) Which circled point, **1** or **2**, has higher leverage?

(f) Which circled point, **1** or **2**, is likely more highly influential (with respect to β_1)?

(g) Consider the following hypothesis test:

$$H_0 : \frac{\partial}{\partial x}\mathrm{E}[\texttt{phys.act}|\texttt{fev} = x]\Big|_{x=4} = \frac{1}{4} \quad \text{vs.} \quad H_1 : \frac{\partial}{\partial x}\mathrm{E}[\texttt{phys.act}|\texttt{fev} = x]\Big|_{x=4} \neq \frac{1}{4}.$$

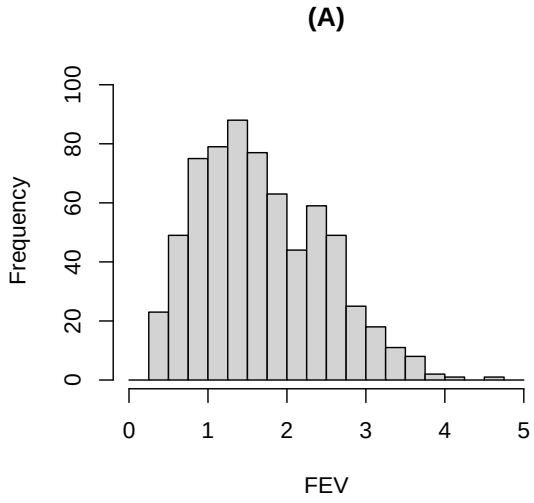
First, provide an interpretation and motivation for this hypothesis test. Although you cannot compute the p-value exactly, use information from the model output on the following page to determine whether a test of this hypothesis would achieve statistical significance under a nominal level of $\alpha = 0.05$ (two-sided).

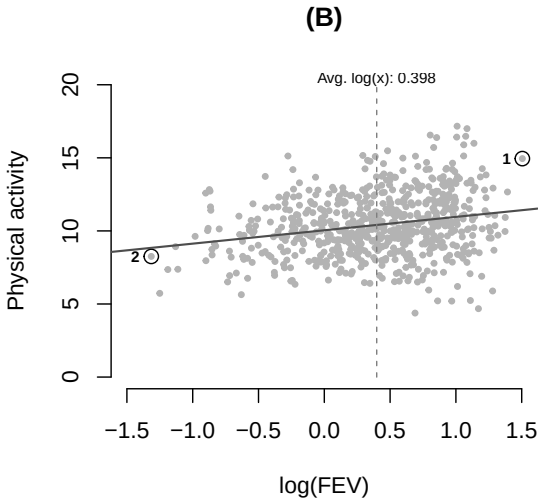
(h) Noting that the root mean squared error is given by $\widehat{\sigma} = 2.043$, consider formulating a 95% prediction interval (i.e., a reference range) for physical activity among individuals with an FEV of 3.00 based on the formula:

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 \log(3.00)\right) \pm 1.96 \times \widehat{\sigma}.$$

Comment on at least two major limitations of this prediction interval that would give you reason to be concerned about its validity.

Supplementary Material for Problem 2






```
model <- ols(phys.act ~ log.fev, x = TRUE, y = TRUE)

> robcov(model)
Linear Regression Model

ols(formula = phys.act ~ log.fev, x = TRUE, y = TRUE)

                 Model Likelihood    Discrimination
                       Ratio Test           Indexes
Obs     672    LR chi2     35.73    R2       0.052
sigma 2.043    d.f.            1    R2 adj   0.050
d.f.    670    Pr(> chi2) 0.0000    g        0.537

Residuals

    Min      1Q  Median      3Q     Max
-6.4357 -1.3757 -0.1381  1.3100  6.1974


          Coef    S.E.   t      Pr(>|t|)
Intercept 10.0437 0.0880 114.14 <0.0001
log.fev    0.9150 0.1556   5.88 <0.0001

> confint(robcov(model))
              2.5 %    97.5 %
Intercept 9.8709360 10.216495
log.fev   0.6094497  1.220564
```

3. [20 pts] An investigative team sought to study the immunogenicity associated with different doses of a vaccine for respiratory syncytial virus (RSV) in adults 50 years of age or older. They conducted a randomized study in which patients were randomized on a 1:1:1 basis to receive a placebo formulation ($X = 0$), a standard-dose formulation ($X = 1$), or a high-dose formulation ($X = 2$). Of further interest was to compare vaccine effects across continuous age, Z, in years. The outcome, Y, was a cellular response to RSV antigen, measured continuously. To that end, consider the following model:

$$\mathrm{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 I(x = 1) + \beta_2 I(x = 2) + \beta_3(z - 50) + \beta_4 I(x = 1)(z - 50) + \beta_5 I(x = 2)(z - 50),$$

where $I(\cdot)$ denotes the indicator function. Note that in this model, age has undergone a transformation (namely, a shift by fifty years).

(a) Translate the null hypothesis, $H_0 : \beta_1 = 0$, into plain scientific language. Show the work that leads to your conclusion.

(b) Translate the null hypothesis, $H_0 : \beta_1 = \beta_4 = 0$, into plain scientific language. Show the work that leads to your conclusion.

(c) Translate the null hypothesis, $H_0 : \beta_1 = \beta_2$, into plain scientific language. Show the work that leads to your conclusion.

(d) Showing your work, determine the parameter or combination of parameters that marks the mean cellular response among 60 year-olds assigned to the high-dose formulation of the RSV vaccine.

(e) Showing your work, determine the parameter or combination of parameters that compares the mean immune response between 75 year-olds assigned to the high-dose formulation of the RSV vaccine to 50 year-olds assigned to the standard-dose RSV vaccine.

(f) The proposed model embeds a key linearity assumption. Articulate this assumption in plain, scientific language and describe how you might assess violations to it given a data set. Then, explain an alternative modeling strategy you could employ to avoid having to make this assumption.

(g) Describe a circumstance under which you might benefit from including baseline (i.e., pre-vaccine) presence of an autoimmune disorder as a covariate. Make clear as part of your response what that advantage would be.

---

4. [20 pts] Below is a sample of R code used to compare the finite-sample behavior of two competing methods via simulation. Investigate the code carefully before responding to the questions that follow.

```
1       simulation <- function(n, p0, p1, alpha, M=50000, seed=2024) {
2           set.seed(seed)
3           reject <- matrix(0, nrow = M, ncol = 2)
4           for (m in 1:M) {
5             x0 <- rbinom(1, size = n, prob = p0)
6             x1 <- rbinom(1, size = n, prob = p1)
7             zz <- chisq.test(matrix(c(sum(x0), n - sum(x0), sum(x1), n - sum(x1)), nrow = 2))
8             p.chisq <- as.numeric(zz$p.value)
9             p.hat.null <- (x0 + x1)/(2*n)
10            p.nrml <- 2*pnorm(q = -abs((x1 - x0)/n),
11                              sd = sqrt(p.hat.null*(1 - p.hat.null)*(2/n)))
12            reject[m,1] <- as.numeric(p.chisq < alpha)
13            reject[m,2] <- as.numeric(p.nrml < alpha)
14           }
15          out <- colMeans(reject)
16          names(out) <- c("Chisq", "Nrml")
17          return(out)
18      }
```

Results of this simulation are shown below for select combinations of values for `n`, `p0`, `p1`, and `alpha` (with some results not reported, labeled `D1` through `D12`).

Simulation parameters				Results	
n	p0	p1	alpha	Chisq	Nrml
20	0.50	0.50	0.10	0.04194	0.08324
200	0.50	0.50	0.10	D1	0.09824
500	0.50	0.50	0.10	0.08854	0.10046
1000	0.50	0.50	0.10	0.09510	D2
200	0.45	0.55	0.10	D3	D4
200	0.45	0.55	0.05	0.47748	0.51896
200	0.55	0.45	0.05	D5	D6
200	0.44	0.56	0.05	D7	D8
200	0.10	0.20	0.05	D9	D10
50000	0.67	0.67	0.05	D11	D12

(a) Briefly explain the major purpose of the `seed` argument (referred to on Line 2).

(b) Briefly explain the most important reason to specify a sufficiently high value for `M`.

(c) Using statistical vocabulary, describe what the `simulation()` function does in a sentence.

(d) Describe your conclusions from this simulation based on the limited information supplied in the table.

For parts (e)-(k) you are asked to make a reasonable hypothesis about the values of `D1` through `D12`. Do not attempt to supply mathematical justification; instead, explain your answers by appealing to your broader knowledge of statistics and relying on the information already supplied in the table. The phrase "make a reasonable hypothesis" can mean supplying an actual number or characterizing a reasonable range.

(e) Make a reasonable hypothesis about the value of `D1`.

(f) Make a reasonable hypothesis about the value of `D2`.

(g) Make a reasonable hypothesis about the values of `D3` and `D4`.

(h) Make a reasonable hypothesis about the values of `D5` and `D6`.

(i) Make a reasonable hypothesis about the values of `D7` and `D8`.

(j) Make a reasonable hypothesis about the values of `D9` and `D10`.

(k) Make a reasonable hypothesis about the values of `D11` and `D12`.

5. [20 pts] Venovenous extracorporeal membrane oxygenation (ECMO) in patients with severe acute respiratory distress syndrome is controversial, as previous studies have shown a benefit while others have shown harm. A randomized controlled trial was conducted to evaluate the efficacy of ECMO in this population. The primary endpoint, Y, is 60-day mortality (you may assume no censoring: all subjects who survive are observed for at least 60 days). The following logistic regression model was considered for $\mathrm{P}(Y = 1)$, the probability of death within 60 days among subjects randomized to received ECMO ($X = 1$) or standard care ($X = 0$):

$$\text{logit}(\mathrm{P}(Y = 1|X = x)) = \beta_0 + \beta_1 x.$$

To augment the interpretation of the trial, a Bayesian analysis was performed under four different priors for β_1. The table below provides a description of the four prior distributions. Also presented are various summaries of the posterior distribution for the odds ratio, $\text{OR} = \exp(\beta_1)$.

$\pi(\beta_1)$	Characterization	Posterior OR (95% CrI)	P(OR < 1/1.25)	P(OR > 1.25)	P(1/1.1 < OR < 1.1)
$\mathcal{N}(0, 0.4^2)$	"Neutral"	1.24 (0.98-1.55)	<0.01	0.47	0.17
$\mathcal{N}(-0.4, 0.4^2)$	"Optimistic"	1.19 (0.95-1.51)	<0.01	0.35	0.26
$\mathcal{N}(0.4, 0.4^2)$	"Pessimistic"	1.28 (1.01-1.62)	<0.01	0.56	0.13
$\mathcal{N}(0, 10^2)$	"Flat"	1.24 (0.92-1.56)	<0.01	--	0.15

Although you have not been supplied information on the prior for β_0, you may assume that $\pi(\beta_0)$ was held constant across the four analyses and plays a negligible role in all questions that follow.

(a) Which of the four priors would you expect to be most consistent with a frequentist analysis of the same data? Briefly explain your reasoning.

(b) The value of $\mathrm{P}(\text{OR} > 1.25)$ under the "flat" prior has intentionally been excluded from the table. Explaining your reasoning, determine its approximate value.

(c) Sketch the prior and posterior distribution for β_1 under the "optimistic" prior (i.e., with β_1 on the x-axis and density on the y-axis).

(d) Suppose the investigators also conducted the a test of the hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ in a frequentist analysis, obtaining a p-value of $p = 0.12$. A clinical colleague used this result to declare that 60-day mortality was "the same" among those receiving and not receiving ECMO. Briefly explaining, summarize the extent to which you agree or disagree with your colleague's conclusion.

(e) Use the results of all of the Bayesian analyses to evaluate which of the following appears most likely, given the data:

- ECMO is associated with important benefit (OR < 1/1.25).
- There is approximate equivalence between ECMO and no ECMO (1/1.1 < OR < 1.1).
- ECMO is associated with significant harm (OR > 1.25).

(f) Using the results from the "neutral" prior, provide an interpretation for the posterior distribution of $\exp(\beta_1)$. Include in your summary an interpretation of the posterior median and 95% credible interval.

(g) Suppose it is known that 60-day mortality is strongly associated with sex in this population. Accordingly, randomization was stratified by sex (so that males and females were each randomized, 1:1, to ECMO or no ECMO). Consider the following logistic regression model that controls for sex, W:

$$\text{logit}(\mathrm{P}(Y = 1|X = x, W = w)) = \gamma_0 + \gamma_1 x + \gamma_2 w.$$

Compared to β_1 from the unadjusted model, do you expect the posterior median for γ_1 to be higher, lower, or the same? Explain your reasoning.
