Biostatistics 2nd year Comprehensive Examination

Due: June 1st, 2018 by 5pm.

Instructions:

1. ***Complete this exam independently. Do not discuss this exam with anyone.***
2. The exam is divided into two sections. There are 6 theory questions in the first section and four data analysis questions in the second section.
3. Answer the questions to the best of your ability. Read the questions carefully.
4. Be as specific as possible and write as clearly as possible.
5. *This is a take-home examination. You may consult books, notes, and papers. You may use the Internet as a research resource. However, you may not consult, or discuss this exam, with another human being, directly or indirectly, nor may you seek help from another individual on the internet (e.g., no posting questions to chat rooms or message boards).*
6. If you have any questions, please contact Professor Blume by email, phone, or text (my cell is 615-545-2656). Texting is welcome. Do not worry about being polite. Contact Professor Blume as needed; call for emergencies.
7. Turn in your exam by emailing it to Professor Blume at j.blume@vanderbilt.edu **AND** Amanda Harding at amanda.harding@vanderbilt.edu. Your exam is not submitted until Professor Blume or Ms. Harding confirm that your exam was received. Alternatively, you may turn in a hard copy to either person by the deadline.
8. ***Vanderbilt's academic honor code applies; adhere to the spirit of this code.***

Question	Points	Score	Comments
1	50		
2	50		
3	50		
4	50		
5	50		
6	50		
Section II	300		50 pts per analysis question (4); 50 pts for overall report clarity and presentation; 50 points for overall thoroughness of approach.
Total	**600**		

Formatting of solutions: Answers to questions in Section I may be handwritten as long as they are neat and easily legible. Typsetting is welcome but not required, and it would be fine to typeset some solutions but not all. Note that Section II provides instructions for how to format the analysis report. Section II solutions should be typeset.

Section I

1. Let P and Q be two probability measures defined on the same sample space Ω and σ-algebra $\mathcal{F}$.

    a. Suppose that $P(A) = Q(A)$ for all $A \in \mathcal{F}$ with $P(A) \leq \frac{1}{2}$. Prove that $P = Q$, i.e. that $P(A) = Q(A)$ for all $A \in \mathcal{F}$.

    b. Give an example where $P(A) = Q(A)$ for all $A \in \mathcal{F}$ with $P(A) < \frac{1}{2}$, but such that $P(A) \neq Q(A)$ for some $A \in \mathcal{F}$.

2. Let $X_1, X_2, \cdots, X_n$ be independent with distribution $U[\theta - 1/2, \theta + 1/2]$, and define $M_n^* = \max\{X_1, X_2, \cdots, X_n\}$ and $M_n^{**} = \min\{X_1, X_2, \cdots, X_n\}$.

    a. Show that for each n, every point in $[M_n^* - 1/2, M_n^{**} + 1/2]$ is a maximum likelihood estimator of θ.

    b. Show that $M_n^{**} + 1/2$ and $M_n^* - 1/2$ are both consistent estimators of θ, in the sense of almost sure convergence.

3. Let h be an absolutely continuous function on $[0,1]$ with $0 \leq h(x) \leq 1$ for all x. Let $U_1, U_2, \cdots, U_n$ be an *iid* sample from a Uniform[0,1] distribution.

   a. Using this sample, argue that

$$\hat{I}_s(h) = \frac{1}{n} \sum_{i=1}^{n} h\left(U_i\right),$$

   is in L^1 and is an unbiased estimator of

$$I(h) \stackrel{\text{def}}{=} \int_0^1 h(x)\, dx.$$

   b. The method of *antithetic variables* leads to an alternative estimator

$$\hat{I}_{av}(h) = \frac{1}{2n} \sum_{i=1}^{n} \left[h\left(U_i\right) + h\left(1 - U_i\right)\right].$$

      i. Show that both $\hat{I}_s$ and $\hat{I}_{av}$ are *strongly consistent* estimators of $I(h)$: that is, as $n \to \infty$, $\hat{I}_s \stackrel{a.s.}{\to} I(h)$ and $\hat{I}_{av} \stackrel{a.s.}{\to} I(h)$.

      ii. Show there exist positive constants $\sigma_s^2(h)$ and $\sigma_{av}^2(h)$ such that

$$\sqrt{n}\left[\hat{I}_s - I(h)\right] \stackrel{d}{\to} N(0, \sigma_s^2(h))$$
$$\sqrt{n}\left[\hat{I}_{av} - I(h)\right] \stackrel{d}{\to} N(0, \sigma_{av}^2(h)).$$

      iii. Derive the functional relation between $\sigma_{av}^2(h)$ and $\sigma_s^2(h)$ in terms of $\rho = Corr(h(U), h(1-U))$. Show that $\sigma_{av}^2(h) = \sigma_s^2(h)$ if and only if $h(U)$ is symmetric about $\frac{1}{2}$. Should one method be preferred over the other based on the smallest $\sigma^2(h)$?

   c. Using the estimators above, estimate the integral (point estimate and confidence interval)

$$\int_0^1 x^a (1-x)^b\, dx,$$

   where $a = -0.25$ and $b = 2$, using an *iid* uniform sample of size $n = 10,000$.

4. Consider a two-class classification problem. Classify target variable G given $X = x_0$ to class $\mathcal{G}_1$ only when the probability $P(G = \mathcal{G}_1|X = x_0) \geq 3P(G = \mathcal{G}_2|X = x_0)$. Let either probability be proportional to the following

$$\hat{\pi}_{\mathcal{G}_j} \hat{f}_{\mathcal{G}_j}(x_0, \lambda),$$

where $\hat{\pi}_{\mathcal{G}_j}$ is the estimated class prior and $\hat{f}_{\mathcal{G}_j}(x_0, \lambda)$ is a nonparametric density estimate for X in class $\mathcal{G}_j$, with smoothing parameter λ.

   a. Describe the loss function that gives rise to this classification rule.
   b. Describe the differences between this method and linear discriminant analysis.
   c. Describe how the smoothing parameter λ affects the bias-variance trade off in predictions.
   d. Describe a technique for selecting the value of the smoothing parameter.

5. Consider prediction of a quantitative response Y given predictor X using the following asymmetric loss function:

$$L_\tau(\hat{Y}, Y) = |\tau - I(Y \leq \hat{Y})||Y - \hat{Y}|$$

where $\tau \in (0, 1)$ and $I(\cdot)$ is the indicator function.

   a. Suppose that overpredictions (i.e., $\hat{Y} > Y$) incur three times the loss as underpredictions of the same magnitude. Find the value τ that encodes this asymmetric loss.
   b. Show that the estimator $\hat{Y}$ that minimizes the expected loss is the $\tau \times 100\%$ percentile of the distribution of Y given X.
   c. For a training sample of size n, consider the effective degrees-of-freedom ($d.f.$) defined as follows:

$$d.f. = \sum_{i=1}^{n} \frac{\text{cov}(\hat{Y}_i, Y_i|X = x_i)}{\text{var}(Y_i|X = x_i)}$$

   Formulate a hypothesis regarding the effect of τ on $d.f.$ Explain your rationale and describe how this hypothesis might be tested or proven. (Hint: think about the pairwise joint distributions of order statistics)

6. Let $l(\theta)$ represent a log likelihood function of p-variate parameter θ. The parameter θ is said to be *maximum likelihood (ML) estimable* if and only if there is a unique global maximum of the log-likelihood function.

   a. *The "label switching" problem:* Show that the parameters of a finite mixture model are not ML estimable. Propose a constraint on the parameter space that avoids the label switching problem.

   b. *The "ridge-in-the-likelihood" problem:* Let $\hat{\theta}$ be a global maximum (i.e., for all θ_0, $l(\hat{\theta}) \geq l(\theta_0)$) that satisfies the ML estimating equation $l'(\theta) = 0$, where $l'(\theta)$ is the score function. Write a first order Taylor approximation of $l'(\theta)$ about $\hat{\theta}$. Argue that, for θ in a neighborhood of $\hat{\theta}$, $l''(\hat{\theta})(\hat{\theta} - \theta) \neq 0$ is a condition for ML estimability. Note that $-l''(\hat{\theta})$ is the observed Fisher information matrix:

$$-l''(\hat{\theta}) = -\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right]_{\theta=\hat{\theta}}$$

   What does this ML estimability condition imply about $l''(\hat{\theta})$?

   c. *Bayesian estimability:* State a Bayesian maximum *a posteriori* (MAP) criterion similar to that for ML. Show that the Bayesian maximum *a posteriori* (MAP) estimator can be estimable even when the ML estimability condition is violated. Provide an example, and explain why this result might be useful in practice.

End Section I

Section II

Background

Alzheimers disease (AD) and related dementias are a major public health crisis and early detection is essential to mitigate the associated burden. Mild cognitive impairment (MCI) is widely regarded as a prodromal stage of dementia, as many individuals diagnosed with MCI convert to Alzheimers disease (AD). Cognitive complaint or a concern regarding changes in cognition is a diagnostic criterion for early MCI, because such complaints purportedly represent a clinically relevant perceived change in cognitive health. Despite evidence that cognitive complaint is an early manifestation of unhealthy brain aging, it remains unclear how cognitive complaint aligns with objective cognitive performance in non-demented adults. Some studies have shown individuals with cognitive complaint correlates with cognitive decline, whereas others suggest no relation between cognitive complaint and objective cognitive performance among non-demented older adults.

One complication in understanding how complaints relate to cognition is the source of complaint. Self-reported complaints may be less reliable than when they are confirmed by a friend, family member or close associate. Complaints from a friend, family member or close associate are called informant complaints. Self-reported cognitive complaint is highly prevalent among older adults but lacks specificity (i.e., cognitively normal elders frequently mention cognitive problems). Also, elders with an underlying neurodegenerative disease sometimes lack insight and may self-assess their cognitive ability incorrectly (e.g., optimistically). One potential solution is to confirm self-reports with an assessment from friend, family member or close associate who knows the patient well (i.e., an informant complaint). While this type of outside confirmation is not accepted as a gold-standard assessment, it can be helpful when evaluating the self-reported complaint.

To date, there have been limited empirical studies of how externally confirmed complaints, either by themselves or in combination with the original self-report, relates to AD diagnostic outcomes and cognitive decline. Your job is to use the National Alzheimers Coordinating Center (NACC) data to examine this issue and examine the effect of self- and confirmed-complaints on diagnostic outcomes and cognitive decline. NACC maintains a database of participant information collected from 30 national Alzheimers Disease Centers (ADCs). A subset of NACC participants 55 to 90 years of age, who had normal cognition at the initial visit between 2005-09-01 and 2014-12-01, were followed annually for three years and are included in this analysis.

Analysis Questions
The goals of the analysis are to address the following:

1. How does baseline cognitive complaint relate to diagnostic conversion by 3 years? Address this question by examining and describing the association between baseline cognitive complaint (complaint.factor: a 4 level factor) and diagnostic conversion at the last clinic visit (last.convert.factor: 2 levels, Stable or Convert, relative to the baseline diagnosis). Be sure to discuss the advantages and disadvantages of using only the last clinical diagnosis, as opposed to using repeated clinical diagnosis over time, in the assessment of this relationship.

2. The scientific team you are working with would like to see the analysis conducted under a Bayesian or Likelihood framework. Pick one and repeat the analysis and summarize the results. Be sure to fully describe the approach (assumptions, background computations, etc.), and quantify the strength of statistical evidnece in these data for a relationship between cognitive complaint and diagnostic conversion.

3. How does baseline cognitive complaint (complaint.factor: a 4 level factor) relate to longitudinal global cognitive decline (where global cognition is measured using Mini-Mental State Examination, denoted as MMSE)? Address this question by comparing trajectories of MMSE scores between different types of baseline cognitive complaint. Be sure to discuss and interpret the relationships between all covariates that are associated with the outcomes.

4. Construct a prediction model that can be used to identify cognitive normal participants who are at increased risk of progressing to MCI or AD within 3 years. Use the time-to-convert (t2convert) variable as the survival outcome and conversion indicator (convert) as the event indicator variable. Summarize the results from this model. If a clinical trial planned to use this risk prediction tool to identify cognitive normal participants with 3-year risk of progression greater than 0.3, 0.4 and 0.5 respectively, how many participants will be identified?

Notes

1. The file `nacc.Rdata` (link below) is an R workspace that contains two data sets: `nacc` and `nacc.long`. Simply type `load(nacc.Rdata)` to load both datasets into R. The dataset `nacc` should be used for questions 1, 2 & 4. Use dataset `nacc.long` for question 3.

2. The file `datasummary2018.pdf` provides a summary of the variables and data in these datasets. The data include different sources of complaint (complaint.factor), demographic status variables (age, race.factor, sex.factor), socioeconomic variable (edu), vital signs (bpsys, height, weight), medical history (htnrx.factor, diab.factor, smoke.factor, cvd.factor, afib.factor) and genetic risk factor (apoe4pos.factor).

3. The dataset `nacc` set includes convert indicator (convert), time to the last visit in years (last.timeinuds.years), the last clinical diagnosis (last.naccudsd.factor), conversion status at the last clinic visit (last.convert.factor) which is the outcome for question 1 (a collapsed last clinical diagnosis of MCI and AD into one level), time to the first clinical diagnosis of MCI or AD (t2convert) which is the outcome for question 4 and is subject to censoring at the last clinic visit.

4. The `nacc.long` data set to be used for question 3 includes follow up time (timeinuds.years) and MMSE (mmse). Please note that you can use baseline covariates and baseline independent variables for question 4, although the data includes time-dependent covariates.

5. For questions 1, 2 & 3 the investigators think that the minimum set of covariates to be *considered* for inclusion in the models are demographic status, socioeconomic variables, medical history of diabetes (diab.factor), current smoking status (smoke.factor) and genetic risk factor. For question 4, they are open to any approach. Be sure to describe your strategy for covariate selection in your methods section.

Report Format
Present your results in the form of an analysis report, consisting of four main sections:

1. **Introduction**: Provide (briefly) any relevant scientific background and state the scientific questions of interest.

2. **Methods**: Summarize and justify the statistical methods used in the analysis as relevant to the scientific questions of interest. It is important to explain how the statistical methods address the scientific questions.

3. **Results**: Present the analysis results regarding the scientific questions of interest, using language understandable to a non-statistician.

4. **Summary**: Provide a brief conclusion of the analysis.

Your report should be 4 to 7 single-spaced pages, excluding figures, tables, and R commands. You will be evaluated based on the appropriateness of the statistical analysis, the quality of the presentation, and the interpretation of the results.

General guidelines

- Be sure to justify the statistical procedures that you use. This includes discussion of any key model decisions and/or any appropriate model evaluation.

- Do not present the results of every analysis that you've done; rather, present the key results.

- Tables and figures should be informative and presented in a format appropriate for a journal article (properly labeled with figure legends and descriptive headings).

- Scale variables appropriately and use significant digits to report results.

- You may include an appendix, but it should contain supplemental information only.

- R commands should not be included in your write-up, but submit all R commands as a seperate appendix.

- Unedited statistical output is not acceptable, but may be included in an appendix for reference purposes.

- Be sure to address each of the analysis questions. If you think a question needs to be modified or expanded, explain your reasoning and describe how such a change impacts the answer.

Links to data and supporting files
Data: https://www.dropbox.com/s/eh7s229f29v9l1w/nacc.RData?dl=0
Data Summary: https://www.dropbox.com/s/z39ho04wiolvkb7/naccdata.pdf?dl=0

End Section II

NACC Cross-Sectional Dataset for Q1, Q2 and Q4
20 Variables 5319 Observations

---

naccid : Subject ID Number

n	missing	distinct
5319	0	5319

```
lowest : NACC000403 NACC000792 NACC000868 NACC000875 NACC000920
highest: NACC999529 NACC999663 NACC999675 NACC999729 NACC999854
```

---

naccvnum : Visit Number

n	missing	distinct	Info	Mean	Gmd
5319	0	1	0	1	0

```
Value         1
Frequency  5319
Proportion    1
```

---

age : Age in Years

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	36	0.999	71.15	9.31	58	60	65	71	77	83	85

```
lowest : 55 56 57 58 59, highest: 86 87 88 89 90
```

---

educ : Years of Education

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	30	0.971	16.23	4.086	12	12	14	16	18	20	20

```
lowest :  0  1  2  3  4, highest: 25 26 27 28 99
```

---

race.factor : Race

n	missing	distinct
5319	0	2

```
Value          White non-White
Frequency       4403       916
Proportion     0.828     0.172
```

---

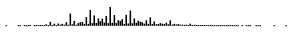
sex.factor : Sex

n	missing	distinct
5319	0	2

```
Value        Male Female
Frequency    1781   3538
Proportion  0.335  0.665
```

---

bpsys : Systolic blood pressure (mmHg)



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	119	0.999	132.8	19.99	106	110	120	131	144	157	164

```
lowest :  78  84  85  87  88, highest: 203 204 205 212 218
```

---

height : Height (inches)

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	192	0.998	66.06	5.28	60.0	61.0	63.0	65.0	68.5	71.5	73.1

```
lowest : 49.0 52.0 53.9 54.0 54.5, highest: 77.0 77.5 78.0 78.4 88.8
```

---

weight : Weight (lbs)

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	217	1	178	61.36	116	124	140	163	190	220	244

```
lowest :  83  85  89  90  91, highest: 320 323 328 332 888
```

---

htnrx.factor : Current use of any type of an antihypertensive or blood pressure medication

n	missing	distinct
5319	0	2

```
Value         No   Yes
Frequency   2635  2684
Proportion 0.495 0.505
```

---

diab.factor : History of Diabetes

n	missing	distinct
5319	0	2

```
Value         No   Yes
Frequency   4723   596
Proportion 0.888 0.112
```

---

smoke.factor : Current Smoker

n	missing	distinct
5319	0	2

```
Value         No   Yes
Frequency   5129   190
Proportion 0.964 0.036
```

---

cvd.factor : History of CVD

n	missing	distinct
5319	0	2

Value	No	Yes
Frequency	4822	497
Proportion	0.907	0.093

---

afib.factor : History of Atrial fibrillation

n	missing	distinct
5319	0	2

Value	No	Yes
Frequency	5029	290
Proportion	0.945	0.055

---

naccudsd : Clinical Diagnosis

n	missing	distinct	Info	Mean	Gmd
5319	0	1	0	1	0

Value	1
Frequency	5319
Proportion	1

---

apoe4pos.factor : APOE4 allel Carrier Status

n	missing	distinct
5319	0	2

Value	No	Yes
Frequency	3707	1612
Proportion	0.697	0.303

---

complaint.factor : Cognitive Complaint

n	missing	distinct
5319	0	4

Value	No Complaint	Self Complaint Only
Frequency	3982	780
Proportion	0.749	0.147

Value	Informant Complaint Only	Both Self and Informant Complaint
Frequency	176	381
Proportion	0.033	0.072

---

convert.factor : Conversion Status within 3 Years of Follow-up

n	missing	distinct
5319	0	2

Value	Stable	Convert
Frequency	5111	208
Proportion	0.961	0.039

---

t2convert : Time to Conversion to MCI or AD in Years

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
5319	0	805	0.96	1.2	1.285	0.000	0.000	0.000	1.207	2.127	2.689	2.913

lowest : -8.681725 -8.117728 -8.084873 -6.381930 -6.228611
highest: 2.986995 2.989733 2.992471 2.995209 2.997947

---

convert : Conversion Status within 3 Years of Follow-up: 0=No, 1=Yes

n	missing	distinct	Info	Sum	Mean	Gmd
5319	0	2	0.113	208	0.03911	0.07517
