Biostatistics 2nd year Comprehensive Examination

Due: June 8th, 2017 by 5pm.

Instructions:

1. ***This is exam is to be completed independently. Do not discuss your work with anyone else.***
2. The exam is divided into two sections. There are 6 theory questions in the first section and a data analysis question in the second section.
3. Answer each question to the best of your ability. Read the exam 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 or by phone (cell: 615-545-2656). Texting is fine as well. 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 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**		

Section I

1. Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with distribution function

$$F(x) = \begin{cases} \frac{1}{2}e^{-x^2} & \text{for } x < 0 \\ 1 - \frac{1}{2}e^{-x^2} & \text{for } x \geq 0 \end{cases}$$

   a. Prove that $P\left\{\frac{X_n}{\sqrt{\log n}} > 1, i.o.\right\} = 1$.

   b. Prove that $P\left\{\frac{X_n}{\sqrt{\log n}} > 1 + \epsilon, i.o.\right\} = 0$ for every $\epsilon > 0$.

   c. Use (a) and (b), and possibly additional arguments, to show that

$$\limsup_n \frac{X_n}{\sqrt{\log n}} \overset{a.s.}{=} 1.$$

   d. Find the density function of X_1.

   e. Show that the characteristic function of X_1 is given by

$$\varphi(t) = 2\int_0^\infty x\, \cos(tx)\, e^{-x^2}\, dx.$$

   f. Prove that $\frac{S_n}{\sqrt{n}} \xrightarrow{d} Z$, where $S_n = \sum_{i=1}^n X_i$ and $Z \overset{d}{=} N(0,1)$.

2. A real number m is a *median* of the distribution of the random variable X if

$$P\{X \leq m\} \geq 1/2 \text{ and } P\{X \geq m\} \geq 1/2.$$

a. Prove that for any random variable a median exists, though it need not be unique.

b. Let φ_X denote the characteristic function of X. Show that if $e^{-imt}\varphi_X(t)$ is real-valued for all t, then m is a median for X.

c. Let $X_1, X_2, \ldots$ be a sequence of random variables such that for each n, m_n is a median of X_n. Prove that if $X_n \xrightarrow{d} X_\infty$ and the median of X_∞, m_∞, is unique, then $m_n \to m_\infty$.

d. Given an example of a distribution – other than a point mass – for which every value in its support is a median.

e. Give an example of a sequence of random variables, $X_1, X_2, \ldots$, with m_n denoting the unique median of X_n, such that $X_n \xrightarrow{d} X_\infty$ yet m_∞, a median of X_∞, is not unique.

3. Let $(\Omega, \mathcal{F}, P)$ be $([0,1], \mathcal{B}_{[0,1]}, P)$, where P is the (continuous) uniform distribution on $[0,1]$.

   a. Let $X_n(\omega) \equiv 1$ and $Y_n(\omega) = I(\omega > 1/n)$, where I denotes an indicator function. Does X_n/Y_n converge for every $\omega \in \Omega$? Does it converge almost surely to a random variable? If so, what random variable?

   b. Let

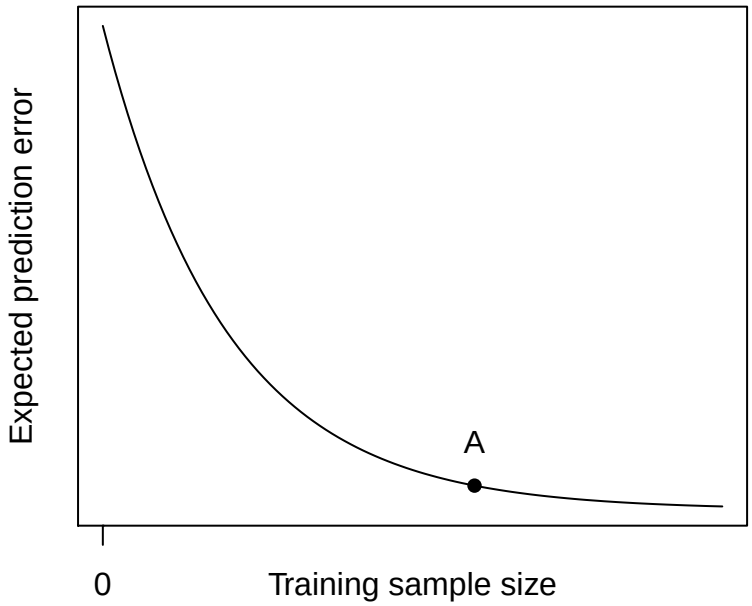
$$Y_n(\omega) = \begin{cases} (-1)^n & \text{if } \omega \text{ is rational} \\ \omega & \text{if } \omega \text{ is irrational} \end{cases}$$

   Does Y_n converge almost surely to a random variable? If so, specify a random variable on $(\Omega, \mathcal{F}, P)$ that Y_n converges to.

   c. Reverse the words "rational" and "irrational" in part (b). Does Y_n converge almost surely to a random variable? If so, specify a random variable on $(\Omega, \mathcal{F}, P)$ that Y_n converges to.

   d. For each n, divide $[0,1]$ into $[0,1/n), [1/n, 2/n), \ldots, [(n-1)/n, 1]$, and let $X_n(\omega)$ be the left endpoint of the interval containing ω. Prove that X_n converges almost surely to a random variable, and identify the random variable.

4. Consider the theoretical association between training sample size and expected prediction error, as illustrated below.



   a. Assuming that point "A" is the total training sample size, mark "CV2" at the point along the curve that corresponds (approximately) to the training sample size of each fold when using 2-fold cross-validation. Mark "CV5" at the point along the curve that corresponds to the training sample size of each fold when using 5-fold cross-validation.

   b. Using the result of part (a), describe the effect of fold count on the bias in estimating predition error using k-fold cross validation. Is the bias conservative or anticonservative?

   c. Mark "LOO" at the point along the curve that corresponds to leave-one-out validation, and "LOOB" for leave-one-out bootstrap validation.

   d. Which validation method has the smallest bias in estimating expected prediction error?

   Note: For answering parts (a) & (c), the graph pdf is availible at
   https://dl.dropboxusercontent.com/u/25204698/comps/tss-vs-epe.pdf .
   However, it is perfectly acceptable to simply mark up a printed page and scan it in.

5. Consider a dichotomous outcome classification problem with Bernoulli target variable $y_i \in \{0, 1\}$ and input x_i, for a sample $i = 1 \dots n$. Denote the probability $P(y_i = 1|x_i) = p_i$, and estimate p_i using the k-nearest neighbor (k-NN) method. Let the classifier $\hat{y}_i$ take the value that minimizes the expected loss, for the following loss function:

$$L(\hat{y}_i, y_i) = \begin{cases} 0 & \text{if } \hat{y}_i = y_i \\ L_{01} & \text{if } \hat{y}_i = 0 \; y_i = 1 \\ L_{10} & \text{if } \hat{y}_i = 1 \; y_i = 0 \end{cases}$$

where L_{01} and L_{10} are positive and finite.

   a. Using the loss function $L(\hat{y}_i, y_i)$, give an expression for the classification rule in terms of p_i.

   b. Argue that the effective degrees-of-freedom ($d.f.$), as given by the formula below, depends on the loss function.

$$d.f. = \sum_{i=1}^{n} \frac{\text{cov}(\hat{y}_i, y_i)}{\text{var}(y_i)}$$

   c. Compute $d.f.$ for the 1-NN classifier.

   d. Expand $\text{cov}(\hat{y}_i, y_i)$ and argue that $d.f.$ becomes smaller as both k and n become large, holding all else constant.

   e. Describe how the relative loss $L_{01}/(L_{01} + L_{10})$ affects $d.f.$, holding all else constant.

   f. Specify an alternative, nonsensical loss function that ensures $d.f.$ is always zero.

6. Let $\eta(x, \theta)$ model a vector of responses under experimental conditions $x = [x_1, \ldots, x_n]$, given p-variate parameter θ. The model parameter θ is said to be *estimable* if and only if $\eta(x, \theta) = \eta(x, \theta')$ implies that $\theta = \theta'$, for all θ and θ'. In words, the model parameters are estimable if and only if the response vector is sensitive to changes in the parameter value.

   a. Write an expression to approximate $\eta(x, \theta)$ in a neighborhood about θ' using a first order Taylor expansion. Argue that $J(x, \theta')(\theta' - \theta) \neq 0$ is a condition for estimability of θ in a neighborhood about θ', where

$$J(x, \theta')_{i,j} = \left[\frac{\partial \eta(x_i, \theta)}{\partial \theta_j}\right]_{\theta_j = \theta'_j}$$

   .

   b. What does the estimability condition derived in part a imply about the matrix $J(x, \theta')$? What does this imply about the *information matrix* $M(x, \theta') = J(x, \theta')^T J(x, \theta')$?

   c. Let $\eta(x, \theta) = x\theta$, where x is an $n \times p$ design matrix. What does the estimability condition imply about the p design variables?

End Section I

Section II

Background

Cardiovascular disease (CVD) affects approximately 40 million individuals over the age of 65 and is the leading cause of mortality. Hospitalization for an acute cardiovascular event is a significant stressor and can lead to functional decline both during the admission and at 12 months follow-up. Older adults who experience a decline in functional status are vulnerable to adverse health outcomes including an increased risk of hospitalization, institutionalization, and mortality. The extent of vulnerable functional status in hospitalized cardiovascular (CV) patients, however, is poorly characterized. Currently, there is no widespread standard for assessment of vulnerable functional status in the hospital setting. Although clinicians are able to recognize severe geriatric impairments, their sensitivity to detect moderate impairments or change in functional impairment is imperfect. Unrecognized impairments may result in new functional needs that are unmet after discharge and lead to increased risk of rehospitalization.

Numerous multidimensional assessment tools have been developed to measure vulnerable functional status, the majority of which have been developed for use in the ambulatory setting and range from a composite score of reported clinical deficits to physical performance based criteria. An optimal assessment in hospitalized older adults would include an objective physical performance test (e.g., gait speed and hand grip strength). However, at the time the patient is admitted these tests may not be feasible to implement. The Vulnerable Elders Survey (VES-13) is one tool utilized in the community to identify 'at-risk' older adults. This instrument may be useful because it is short, simple, and can be collected without the need for specialists. The instrument is attached, and it was captured at admission to the hospital in elderly patients (age 65 and above) participating in the Vanderbilt Inpatient Cohort Study (VICS). Ideally, vulnerable patients can be identified early so that services can be deployed to minimize risk for poor patient outcomes such as long hospital length of stay, readmission to the hospital, and death. While VICS collected the VES-13, the score was not used to render care. This analysis seeks to examine if collecting VES-13 at hospital admission can help to identify patients more likely to experience poor outcomes.

We have provided a number of variables for analyses that are being conducted. They include the VES-13 score (ves13score), demographics and baseline characteristics (age, bmi, gender, race, patient transfer), socioeconomic status variables (employ4cat, edu), overall patient health at the time of hospital admission (hosp12m, elix, diagnosis2), the length of the hospital stay in days (los), where the patient was discharged from the hospital (home3), readmission or death within 90 days from discharge (readm.death.90), and one year mortality (death1yr). Note that all patients lived at home prior to being admitted to the hospital and enrolling in VICS. The primary analyses for this study examine the extent to which knowledge of patient vulnerability as measured by the VES-13 could provide useful information for patient treatment. If patient vulnerability is associated with poor patient outcomes, we might begin to seek approaches to remediate unnecessary risk. As a secondary analysis, we seek to build a prognostic model to characterize risk of death in the year following discharge.

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

1. At the time of admission, does VES-13 provide useful information about patient length of stay at the hospital? Address this question by examining and describing the independent VES-13 association with length of hospital stay (los).

2. At the time of hospital discharge, does VES-13 provide useful infomormation about risk of readmission or death within 90 days? Address this question by examining and describing the independent VES-13 association with patient readmission or death within 90 days ($readm.death.90$).

3. For the above models, discuss the relationship between all covariates that are associated with the above outcomes.

4. At the time patients are discharged from the hospital, we would like predict risk of death within the next year. Construct and evaluate a model for predicting death within 1 year after discharge ($death1yr$).

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 4–7 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 please 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.

The next several pages contian the VES instrument as well as a description of the analysis dataset. Links to identical versions of these documents are included below.

Links to data and supporting files
Data: https://dl.dropboxusercontent.com/u/25204698/comps/VICSData.Rdata
VES Instrument: https://dl.dropboxusercontent.com/u/25204698/comps/ves13.pdf
Data Summary: https://dl.dropboxusercontent.com/u/25204698/comps/Datasummary2017.pdf

End Section II

VES-13

1. Age _____

SCORE: *1 POINT FOR AGE 75-84*
3 POINTS FOR AGE ≥ 85

2. In general, compared to other people your age, would you say that your health is:

☐ Poor,* *(1 POINT)*
☐ Fair,* *(1 POINT)*
☐ Good,
☐ Very good, or
☐ Excellent

SCORE: *1 POINT FOR FAIR or POOR*

3. How much difficulty, on average, do you have with the following physical activities:

	No Difficulty	A little Difficulty	Some Difficulty	A Lot of Difficulty	Unable to do
a. stooping, crouching or kneeling?	☐	☐	☐	☐ *	☐ *
b. lifting, or carrying objects as heavy as 10 pounds?	☐	☐	☐	☐ *	☐ *
c. reaching or extending arms above shoulder level?	☐	☐	☐	☐ *	☐ *
d. writing, or handling and grasping small objects?	☐	☐	☐	☐ *	☐ *
e. walking a quarter of a mile?	☐	☐	☐	☐ *	☐ *
f. heavy housework such as scrubbing floors or washing windows?	☐	☐	☐	☐ *	☐ *

SCORE: *1 POINT FOR EACH * RESPONSE IN Q3a THROUGH f . MAXIMUM OF 2 POINTS.*

4. Because of your health or a physical condition, do you have any difficulty:

a. shopping for personal items (like toilet items or medicines)?

☐ YES → Do you get help with shopping? ☐ YES * ☐ NO
☐ NO
☐ DON'T DO → Is that because of your health? ☐ YES * ☐ NO

b. managing money (like keeping track of expenses or paying bills)?

☐ YES → Do you get help with managing money? ☐ YES * ☐ NO
☐ NO
☐ DON'T DO → Is that because of your health? ☐ YES * ☐ NO

Continued

c. walking across the room? USE OF CANE OR WALKER IS OK.

☐ YES → Do you get help with walking? ☐ YES * ☐ NO
☐ NO
☐ DON'T DO → Is that because of your health? ☐ YES * ☐ NO

d. doing light housework (like washing dishes, straightening up, or light cleaning)?

☐ YES → Do you get help with light housework? ☐ YES * ☐ NO
☐ NO
☐ DON'T DO → Is that because of your health? ☐ YES * ☐ NO

e. bathing or showering?

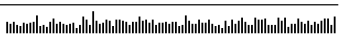
☐ YES → Do you get help with bathing or showering? ☐ YES * ☐ NO
☐ NO
☐ DON'T DO → Is that because of your health? ☐ YES * ☐ NO

SCORE: *4 POINTS FOR ONE OR MORE * RESPONSES IN Q4a THROUGH Q4e*

.

VICSData
16 Variables 1067 Observations

studyid



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	1067	1	1546	179.6	322.2	811.5	1535.0	2314.5	2748.8	2916.7

```
lowest :    3    4    5    6    7, highest: 3036 3037 3038 3039 3040
```

readm.death.90 : Readmitted or died within 90 days from discharge (1=yes, 0=no)

n	missing	unique	Info	Sum	Mean
1067	0	2	0.64	329	0.3083

death1yr : Died in the year following hospital discharge (1=yes, 0=no)

n	missing	unique	Info	Sum	Mean
1067	0	2	0.34	141	0.1321

los : Number of days hospitalized when discharged

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	30	0.98	4.898	1	1	2	3	6	10	13

```
lowest :  1  2  3  4  5, highest: 31 36 38 49 72
```

age : Age (in years)

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	30	1	73	65	66	68	71	77	83	85

```
lowest : 65 66 67 68 69, highest: 90 91 92 94 95
```

bmi: BMI (kg / m^2)

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1065	2	826	1	29.21	20.66	22.25	24.96	28.36	32.45	37.30	40.80

```
lowest : 14.11 14.71 17.02 17.12 17.25
highest: 51.59 52.60 52.80 53.49 56.15
```

gender : Gender

n	missing	unique
1067	0	2

```
Male (612, 57%), Female (455, 43%)
```

racecat : Race

n	missing	unique
1067	0	2

```
White (966, 91%), Non-White (101, 9%)
```

transfer : Was the patient transferred to VUMC from another hospital? (1=yes, 0=no)

n	missing	unique	Info	Sum	Mean
932	135	2	0.68	321	0.3444

employ4cat : Current employment status variable (4 levels)

n	missing	unique
1067	0	4

```
Employed (183, 17%), Not employed for wages (38, 4%), Retired (789, 74%)
Unable to work (disabled) (57, 5%)
```

edu : Number of years of education (e.g. 12=finished high school, 16=finished college)

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	23	0.97	13.97	9	11	12	14	16	18	20

```
lowest :  3  4  5  6  7, highest: 21 22 23 24 25
```

diagnosis2 : Patient diagnosis that led to the hospitalization and participation in VICS

n	missing	unique
1067	0	2

```
ACS (643, 60%), ADHF (424, 40%)
```

hosp12m : Number of times admitted to the hospital in the 12 months preceding VICS

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1066	1	13	0.87	1.214	0	0	0	1	2	3	5

	0	1	2	3	4	5	6	7	8	9	10	12	14
Frequency	515	243	133	71	41	28	13	6	9	2	2	2	1
%	48	23	12	7	4	3	1	1	1	0	0	0	0

---

elix : Elixhauser Comorbidity Index: A measure of comorbidity severity. Higher numbers imply worse patient health

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	60	1	16.9	0.0	2.0	8.0	16.0	24.0	32.4	37.0

lowest : -3 -2 -1 0 1, highest: 52 53 54 55 59

---

home3 : Where was the patient discharged from the hospital to? (3-levels)

n	missing	unique
1067	0	3

Home without Services (799, 75%), Home with Services (147, 14%)
Hospice / Rehab (121, 11%)

---

ves13score : VES-13

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1067	0	11	0.98	3.138	0	0	1	2	6	7	8

	0	1	2	3	4	5	6	7	8	9	10
Frequency	213	187	145	162	49	24	45	169	49	16	8
%	20	18	14	15	5	2	4	16	5	1	1
