Biostatistics 2nd year Comprehensive Examination

Due: May 29th, 2015 by 5pm.

Instructions:
1. *This is exam is to be completed independently. Do not discuss your work with anyone else. Please read the exam carefully.*
2. The exam is divided into two sections. There are 7 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.
4. Be as specific as possible and write as clearly as possible.
5. *This is a take-home examination. You may consult books, notes, papers, and you may use the Internet. However, you may not consult or discuss this exam with another human being or oracle, 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). Do not worry about being polite; email Professor Blume as needed and call for emergencies.
7. Turn in your exam by emailing it to Professor Blume at j.blume@vanderbilt.edu AND Linda Wilson at linda.l.wilson@vanderbilt.edu. Your exam is not submitted until you receive confirmation from either Professor Blume or Ms. Wilson that your exam was received on time. Alternatively, you may turn in a hard copy to either person by the deadline.
8. *Vanderbilt's academic honor code applies; be sure to adhere to the spirit of this code.*

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Score</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section II</td>
<td>210</td>
<td></td>
<td>70 pts per scientific goal</td>
</tr>
<tr>
<td>Total</td>
<td>420</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section I

1. Suppose that a population consists of a fixed number, say $m$, of genes in any generation. Each gene is one of two possible genetic types. If any generation has exactly $i$ (of its $m$) genes being Type-1, then the next generation will have $j$ Type-1 (and $m-j$ Type-2) genes with probability

$$\binom{m}{j} \left( \frac{i}{m} \right)^i \left( \frac{m-i}{m} \right)^{m-j}, \quad j \neq 0, 1, ..., m$$

Let $X_n$ denote the number of Type-1 genes in the $n^{th}$ generation, and assume that $X_0 = i_0$.

a. Find $E[X_n]$.
b. What is the probability that eventually all genes will be Type-1?

2. Let $X_1, X_2, ...$ be independent random variables such that

$$\begin{cases} n^2 - 1 & \text{with probability } n^{-2} \\ -1 & \text{with probability } 1 - n^{-2} \end{cases}$$

a. Prove that $X_n \overset{L^1}{\to} 0$.
b. Let $S_n = \sum_{k=1}^{n} X_k$. Prove $S_n/n \overset{a.s.}{\to} -1$.
c. Given part (a), one might expect that $S_n/n$ would converge to 0, almost surely. Explain in words why this is not the case. [You may use an example backed up with figures, etc., if you wish.]

3. Let $X_1, X_2, ...$ be an i.i.d. sequence of random variables with sum $Z = \sum_{i=1}^{N} X_i$ where $N \overset{d}{=} P(\lambda)$ is a random variable that is independent of $X_i$. Denote the characteristics function for $X$ as $\phi_X$.

a. Prove that the characteristic function for $Z$ is $\phi_Z = e^{\lambda(\phi_X-1)}$.
b. Use part (a) to express the expectation and variance of $Z$ in terms of the expectation and variance of $X_i$ (denote them by $\mu$ and $\sigma^2$) and $\lambda$. 

Biostatistics 2nd year comprehensive exam
Spring 2015

1 of 10
4. Consider a prediction problem in which the set of allowable predictors is
\(V = \{aY + b: a, b \in \mathbb{R}\}\), where \(Y \in L^2\) and \(\sigma^2 = Var(Y) > 0\).

a. Show that \(Z_1 = (Y - E[Y])/\sigma\) and \(Z_2 \equiv 1\) constitute an orthonormal basis of \(V\).

b. Show that for \(X \in L^2\), the MMSE predictor of \(X\) within \(V\) is
\[
\hat{X} = \frac{\text{Cov}(X,Y)}{\sigma^2} (Y - E[Y]) + E[X]
\]

5. Consider the classification of target variable \(G\) given input features \(X\), such that
the probability \(P(G = j | X = x_0)\) is maximized with respect to \(j\). Let this
probability take the form
\[
\frac{\hat{\pi}_j \hat{f}_j(x_0, \lambda)}{\sum_{k=1}^{J} \hat{\pi}_k \hat{f}_k(x_0, \lambda)}
\]
where \(\hat{\pi}_j\) is the estimated class prior and \(\hat{f}_j(x_0, \lambda)\) is a nonparametric density
estimate for \(X\) in class \(j = 1, ..., J\) with smoothing parameter \(\lambda\).

a. Describe the loss function that gives rise to this classification rule.

b. Describe how the smoothing parameter \(\lambda\) affects the bias-variance trade off in prediction error.

c. Describe a technique for selecting the values of the smoothing parameters.

d. Describe a method for computing the effective number of model parameters (degrees of freedom).
6. Consider the dichotomous classification of phonemes using generalized linear regression and, as inputs, the intensity of 256 ordered and evenly spaced audio frequencies that are measured when the phoneme is spoken. Specifically, let \( y \) represent the binary target variable and \( x \) the vector of intensities.

a. Write an expression for \( g(\pi(x, \beta)) \), where \( g \) is a link function and \( \pi(x, \beta) \) is the probability that \( y = 1 \) given \( x \), and \( \beta \) represents a vector of linear coefficients.

b. Rather than treating each element of \( \beta \) as a free parameter, consider applying a filter to the frequencies by representing \( \beta \) as a smooth function of frequency. Write an expression for \( \beta(f) \) as a function of frequency using a spline basis \( H(f) = [h_0(f), ..., h_M(f)] \) with coefficients \( \theta = [\theta_1, ..., \theta_M] \), and then rewrite the expression from part (a) in these terms by evaluating \( \beta(f) \) at each of the observed frequencies \( [f_1, ..., f_{256}] \).

c. Let \( \hat{\theta} \) be the maximum likelihood estimator (MLE) for \( \theta \). Provide approximate expressions for
   i. the sampling distribution of the MLE \( \hat{\beta}(f) \),
   ii. a pointwise confidence band for \( \beta(f) \),
   iii. a simultaneous confidence region for \( \beta \).

d. Explain why (with justification) the distribution of \( \pi(x, \hat{\beta}) \) would be approximately normal. Provide an approximate expression for pointwise confidence bands for \( \pi(x, \beta) \) and describe how they could be computed using typical computer output, e.g., that provided by the \( \text{glm} \) function in R.
7. Consider a log likelihood function \( l(\beta) = \sum_{i=1}^{n} l_i(\beta) \) where \( l_i(\beta) \) is the contribution of observation \( i \) and \( l_{(-i)}(\beta) \) represents the log likelihood associated with all observations except \( i \). Denote the first and second derivatives \( l'_{(-i)}(\beta) \) and \( l''_{(-i)}(\beta) \), respectively. Let \( \hat{\beta} \) and \( \hat{\beta}_{(-i)} \) represent the arguments that maximize \( l(\beta) \) and \( l_{(-i)}(\beta) \), respectively.

a. Show that the contribution of observation \( i \) can be written as

\[
l_i(\beta) = l(\beta) - l_{(-i)}(\beta)
\]

b. Write the first-order Taylor series expansion of \( l'_{(-i)}(\beta) \) about \( \hat{\beta} \).

c. Use part (b) to show that

\[
\hat{\beta}_{(-i)} \approx \hat{\beta} - \left[l''_{(-i)}(\hat{\beta})\right]^{-1} l'_{(-i)}(\hat{\beta})
\]

d. Describe how part (c) could be used to generalize the idea of leave-one-out cross validation. Use part (c) to write an approximate expression for the cross validated estimate of prediction error, where \( y_i \) is the target of interest, \( f_{\hat{\beta}}(\beta) \) is the associated predictor or classifier, and loss is quantified using loss function \( L(y_i, f_{\hat{\beta}}(\beta)) \).

End Part I.
Section II

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 that 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, in the inpatient setting these tests may not be feasible to implement due to monitoring devices, continuous IV therapies, acute pain or discomfort, or other physical limitations, particularly among CVD patients who may have recently undergone an invasive procedure such as radial or femoral access coronary angiography.

One method utilized in the community to identify older adults at risk is a self-report instrument called the Vulnerable Elders Survey (VES-13). This instrument is thought to be useful because it is simple and is not affected by aforementioned challenges. 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, patients who are identified as more vulnerable will be provided, at discharge, the healthcare related support that is needed to minimize risk for poor patient outcomes such as readmission to the hospital and death.

We have provided a number of variables for the analyses that are being conducted. They include a number of measures collected at admission (e.g., health literacy, health numeracy, diagnosis upon admission (i.e., acute coronary syndrome, acute decompensated heart failure, or both), social support, social isolation, depression, recent medical history) and two collected at discharge (preparedness at the time of hospital discharge and whether or not discharged with some degree of support). Some demographics information is included as well. The data were collected as part of VICS, which seeks to examine sociodemographic determinants of patient outcomes among those admitted to the hospital with acute coronary syndrome and/or acute decompensated heart failure.

Data: [https://dl.dropboxusercontent.com/u/25204698/Comps/VESdata.RData](https://dl.dropboxusercontent.com/u/25204698/Comps/VESdata.RData)
The goal of your analysis is to address the following:

1. Describe and identify patient characteristics (demographic, social, baseline characteristics, etc.) that are associated with vulnerability at the time the patient was admitted to the hospital and enrolled in the study.

2. One particular focus of VICS is to examine the impact of health literacy (as measured by the stOFHLA; stofhlasmus) on patient vulnerability. Identify, describe and carefully interpret the relationship between health literacy and vulnerability.

3. We hope that our MDs and RNs are discharging patients with the appropriate support needed to minimize the risk for poor outcomes. Since patient vulnerability is likely to be associated with poor outcomes, evaluate the extent to which vulnerability is used to determine whether healthcare support is provided after discharge (DischWithSupport).

Present and report your results and findings in the form of a statistical analysis report with 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 (use the Methods section to describe your planned analysis), the quality of the presentation, and the interpretation of the results.

**General guidelines: Important!**
- Tables and figures should be informative and presented in a format appropriate for a journal article (properly labeled with figure legends and descriptive headings).
- *Edit numerical results to a reasonable number of significant digits and scale variables appropriately.*
- Unedited statistical output is not acceptable.
- R commands should not be included in your write-up, but please submit all R commands as an appendix.
- 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.

*The next four pages contain the VES instrument and a description of the analysis dataset. Please look over these documents carefully.*

End Part II.
1. Age ______________________________

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

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

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

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.
The dataframe is called VESData and it can be loaded with: `load(VESData.RData)`

**VESData**  
17 Variables 452 Observations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>n</th>
<th>missing</th>
<th>unique</th>
<th>Info</th>
<th>Mean</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>ves13score</td>
<td></td>
<td>445</td>
<td>7</td>
<td>11</td>
<td>0.98</td>
<td>3.315</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>age: Age</td>
<td></td>
<td>452</td>
<td>0</td>
<td>27</td>
<td>1</td>
<td>72.87</td>
<td>65</td>
<td>66</td>
<td>68</td>
<td>71</td>
<td>77</td>
<td>82</td>
<td>84</td>
</tr>
<tr>
<td>edu: Edu</td>
<td>Number of years of school completed</td>
<td>452</td>
<td>0</td>
<td>22</td>
<td>0.98</td>
<td>13.84</td>
<td>8.55</td>
<td>10.00</td>
<td>12.00</td>
<td>13.00</td>
<td>16.00</td>
<td>18.00</td>
<td>20.00</td>
</tr>
<tr>
<td>gender</td>
<td>Gender</td>
<td>452</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>racecat: Race</td>
<td></td>
<td>451</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>diagnosis: Diagnosis</td>
<td></td>
<td>452</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maritalcat1</td>
<td></td>
<td>452</td>
<td>0</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LiveAlone: Living Alone</td>
<td></td>
<td>452</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>phqsum: PHQ8 Depression Sum (higher implies more depressed)</td>
<td></td>
<td>447</td>
<td>5</td>
<td>24</td>
<td>0.99</td>
<td>7.324</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>stofhlasum: Objective measure of health literacy: higher implies higher health literacy</td>
<td></td>
<td>427</td>
<td>25</td>
<td>35</td>
<td>1</td>
<td>26.12</td>
<td>9</td>
<td>12</td>
<td>19</td>
<td>30</td>
<td>34</td>
<td>35</td>
<td>36</td>
</tr>
</tbody>
</table>

- **ves13score**: Frequency 75 66 66 86 19 8 19 71 27 4 4
- **age**: Age:
  - Frequency 17 15 15 19 4 2 4 16 6 1 1
- **edu**: Number of years of school completed:
  - Frequency 75 66 66 86 19 8 19 71 27 4 4
- **gender**: Male (240, 53%), Female (212, 47%)
- **racecat**: Race:
  - White (394, 87%), African American (45, 10%), Other (12, 3%)
- **diagnosis**: ACS (267, 59%), Heart Failure (145, 32%), Both (40, 9%)
- **maritalcat1**: Married/Living partner (281, 62%), Unmarried (64, 14%), Widowed (107, 24%)
- **LiveAlone**: Living Alone:
  - No (341, 75%), Yes (111, 25%)
- **phqsum**: PHQ8 Depression Sum (higher implies more depressed):
  - Frequency 75 66 66 86 19 8 19 71 27 4 4
- **stofhlasum**: Objective measure of health literacy: higher implies higher health literacy:
  - Frequency 75 66 66 86 19 8 19 71 27 4 4
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>n</th>
<th>missing</th>
<th>unique</th>
<th>Info</th>
<th>0.05</th>
<th>0.10</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.90</th>
<th>0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>snsmean</td>
<td>Subjective Numeracy</td>
<td>451</td>
<td>1</td>
<td>16</td>
<td>0.99</td>
<td>4.49</td>
<td>2.167</td>
<td>2.667</td>
<td>3.667</td>
<td>4.667</td>
<td>5.667</td>
<td>6.000</td>
</tr>
<tr>
<td>hosp12m</td>
<td>Number of hospitalizations within the 12 months prior to the present study</td>
<td>449</td>
<td>3</td>
<td>11</td>
<td>0.88</td>
<td>2.687</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>midussum</td>
<td>MIDUS cognition score (higher implies higher cognition)</td>
<td>452</td>
<td>0</td>
<td>28</td>
<td>1</td>
<td>15.2</td>
<td>8</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>essi6sum</td>
<td>ENRICHED social support score (higher implies more social support)</td>
<td>451</td>
<td>1</td>
<td>20</td>
<td>0.98</td>
<td>26.12</td>
<td>18</td>
<td>20</td>
<td>24</td>
<td>27</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>DischWithSupport</td>
<td>Was the patient discharged with at least some level of health care support</td>
<td>409</td>
<td>43</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(e.g., home nurse / nursing home)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bprep</td>
<td>B-prepared: patient's perceived preparedness at the time of discharge</td>
<td>452</td>
<td>0</td>
<td>17</td>
<td>0.94</td>
<td>19.1</td>
<td>11.0</td>
<td>13.1</td>
<td>18.0</td>
<td>20.0</td>
<td>22.0</td>
<td>22.0</td>
</tr>
<tr>
<td>hosp30</td>
<td>Hospitalization Last 30 days</td>
<td>448</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

No (261, 64%), Yes (148, 36%)