Biostatistics 2nd year Comprehensive Examination

Due: June 9th, 2016 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 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. 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	20		
2	20		
3	40		
4	40		
5	40		
6	40		
7	40		
Section II	240		Two analyses with 40 pts each for methods, interpretation, presentation
Total	**480**		

Section I

1. Let $Y = \{Y_n\}$ be a sequence for which there exists the constant α with

$$E[Y_{n+1}|Y_0, \dots, Y_n] = \alpha Y_n + (1-\alpha)Y_{n-1}$$

for each n.

(a) For what values of α is the sequence Y a martingale?

(b) Show that there exists a real number c such that $X_n = cY_n + Y_{n-1}$ is a martingale with respect to Y.

2. Show that an event whose probability is either zero or one is independent of every event, and that an event that is independent of itself has probability zero or one.

3. Let $X = \{X_k\}$ be a sequence of independent random variables such that X_k is Poisson distributed with parameter $1/k$.

(a) Show that the sequence X converges in $\mathcal{L}^1$ to 0.

(b) Define the partial sum $S_n = \sum_{k=1}^{n} X_k$ for $n = 1,2, \dots$. Is there a finite constant c such that $S_n \overset{\mathcal{L}^1}{\to} c$? Explain and justify your answer.

(c) Prove or disprove that $(S_n - \mu_n)/\sigma_n \overset{d}{\to} Z$ where $\mu_n = E[S_n]$, $\sigma_n^2 = Var[S_n]$, and $Z \overset{d}{=} N(0,1)$.

4. You must choose to play one of two games. Each game consists of an infinite sequence of bets where the losses and winnings will add up. The two games are **Ploutos** and **Aequum**:

   Game **Ploutos**: On the n^{th} try, you win one dollar with probability $\frac{2^n}{2^n+1}$ and loose 2^n dollars with probability $\frac{1}{2^n+1}$.

   Game **Aequum**: On the n^{th} try, you win one dollar with probability $\frac{n}{n+1}$ and loose n dollars with probability $\frac{1}{n+1}$.

   To help decide which game to play, answer the following questions:

   (a) Show that both games are fair on the n^{th} bet (i.e., show that their expectation is zero on the n^{th} try) even though Ploutos' loss is always larger.

   (b) How many bets are lost, on average, in each game? (Remember a game consists of an infinite number of bets, $n = 1, 2, \ldots$)

   (c) Show that with probability one, you will lose only a finite number of bets in the Ploutos game.

   (d) If you were looking to make money, is there a reason to prefer one game over the other? Justify your answer.

   (e) Would your answer to part (d) change if you had only \$2 in your pocket, which you needed for a bus ride home, and the house only had \$1,000 left to play with? Explain your answer.

5. Consider a dichotomous outcome (y) classification problem in which the logit of the outcome probability is linearly related to a predictor(x), but that this relationship differs by some number (K) of latent subgroups. Model the phenomenon using a Bernoulli mixture density as follows:

$$p(y|x) = \sum_{k=1}^{K} \pi_k b\big(\mu(x^T \beta_k)\big)$$

where π_k is a mixing probability, $b(\cdot)$ is the Bernoulli mass function, $\mu(\cdot)$ is the inverse logit function, and β_k is the latent group linear coefficient vector.

(a) For a sample $\{y_i, x_i\}_{i=1}^N$, write down the marginal log-likelihood function.

(b) Introduce a latent variable z to represent the subgroup associated with each sample observation, then write down the conditional (on $\{z_i\}_{i=1}^N$) or "complete-data" log-likelihood function.

(c) Consider maximizing the marginal log-likelihood function using the EM algorithm. As far as possible, write explicit solutions for the E and M steps of the algorithm.

(d) Briefly describe the steps to implement the above EM algorithm in R using the **`glm`** or **`lrm`** function.

6. Consider a dichotomous classification problem in which the outcome (y) can take the value A or B. Denote the predictor function as $f(x)$; a function of input x. Let the loss function for prediction take the value L_{AB} when $y = A$ and $f(x) = B$ and the value L_{BA} when $y = B$ and $f(x) = A$, and zero otherwise. Thus, the loss associated with misclassification may be asymmetric.

(a) Write an expression for the expected prediction error (EPE) as a function of the predictor f.

(b) Argue that minimizing the conditional EPE $(y|x)$ for each value of x is sufficient to minimize the (marginal) EPE.

(c) Derive a classification rule for y given x by minimizing the EPE. Express the rule in terms of the odds associated with $y = A$ given x.

(d) Suppose that A is a disease state and $f(x)$ is a diagnostic test. Discuss the relationship between sensitivity/specificity and the values of L_{AB} and L_{BA}.

7. Using the model and EPE described above in problems (5) and (6):

   (a) Describe a method to select the number of subgroups/mixture components K by minimizing an estimate of EPE. Use expressions similar to those in Hastie, Tibshirani, and Friedman Section 7.4 to describe the EPE estimator.

   (b) Describe a method for approximating the effective number of parameters (degrees of freedom) associated with the model fitting process, including selection of K.

   (c) Reconcile the maximum likelihood approach of problem (5) with the minimum EPE approach in problem (6). Should the model parameters π_k and β_k be estimated using an alternative criterion? Explain and justify your answer.

End Part I.

Section II

Heart failure (HF) and acute coronary syndrome (ACS) are significant and costly public health concerns. Over 5.1 million people in the United States are affected by HF, and each year, approximately 1.4 million patients are hospitalized for ACS, of which 810,000 are for myocardial infarctions. Hospitalizations account for a large fraction of the costs associated with HF and ACS, and costs of hospitalizations are driven very much by length of stay in the hospital (LOS) and also readmissions to the hospital after discharge (READM). Understanding factors that influence LOS and READM risk can potentially facilitate allocation of hospital resources to more accurately identify patients at increased risk. Though many predictors of LOS and READM may not be modifiable by the time of hospitalization, more complete understanding of risk factors can facilitate deployment of interventions geared toward reducing LOS and READM risk. These interventions include tailoring inpatient care and adjusting the timing of patient education and discharge planning.

We will consider a subset of N=1239 patients who participated in the Vanderbilt Inpatient Cohort Study (VICS). One purpose of VICS is to obtain a more complete understanding of the risk factors associated with READM in a population of patients enrolled with HF and ACS. If we are able to learn what predicts READM and long LOS, we can begin to build clinical decision support (CDS) tools that run in the background of the electronic health record. Such tools can be programmed to automatically alert hospital staff when high risk patients are admitted to the hospital. We can then dedicate resources and tailor therapy to ensure that their care is optimized in the hospital to reduce LOS if necessary, and to reduce READM risk.

For the purpose of this project, we are interested in two analyses:

1. Examine and describe the association between hospital LOS and variables contained in the following constructs: 1) health competence, 2) medical history, and 3) social support. We are interested in individual variable associations with LOS as well the strength of individual construct associations with LOS. The constructs are shown below. Demographic variables and other readily available data are also included in the dataset.

2. Construct a prediction model that can be used at the time a patient is admitted to the hospital to identify who is at increased risk for being readmitted to the hospital within 90 days after discharge. Summarize the results from this model. If VUMC implemented a decision support tool that notified hospital staff if a patient's 90 day READM risk was greater than or equal to 0.4, 0.5, or 0.6, how many people would we estimate being flagged?

Data: https://dl.dropboxusercontent.com/u/25204698/comps/VICSData.Rdata

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 5 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 for this), 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.

- When interpreting regression model parameter estimates, please do so carefully. You are being tested on your ability to understand precise interpretations of parameter estimates.

- Unedited statistical output is not acceptable.

- R commands should not be included in your write-up, but 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.

All Variables

- **id:** subject identifier

- **Demographics** and other easily ascertained variables
  - **age**: Continuous variable, age in years.
  - **gender2**: Two level categorical variable (i.e., Female/Male).
  - **racecat1**: Three level categorical variable (i.e., White/African American/Other).
  - **transfer**: Two level categorical variable (Was the patient a transfer patient from another hospital).
  - **diagnosis1**: Three level categorical variable (i.e ACS/ADHF/ACS and ADHF).

- **Health competence construct** variables that could potentially be stored in the HER.
  - **edu**: Continuous variable, number of years of school completed.
  - **stofhla**: Continuous variable, an objective measure of health literacy on a 0 to 36 point scale.
  - **snsmean**: Continuous variable, subjective measure of self-perceived health numeracy on a 1 to 6 point scale.

- **Medical history construct** variables mostly from the recent past.
  - **elix**: Continuous variable, modified elixhauser comorbidity score that captures a number of possible co-morbid conditions. Note that scores can be negative. Higher scores are associated with more severe comorbid conditions
  - **hosp12m**: Continuous variable, measures number of hospitalization in last 12 months prior to VICS enrollment.

- **Social Support construct** variables.
  - **maritalstatus1**: Three level categorical variable (i.e., Married/Separated-Divorce-Single/Widowed).
  - **nhome**: Number of people living in the household other than yourself; continuous variable.
  - **essi6sum**: Continuous variable, sum of 6 items on a questionnaire measuring self-perceived social support; 6 to 30 point scale.

- **Variables measured while hospitalized and after discharge.**
  - **icu**: Was the patient admitted to the intensive care unit while hospitalized (yes/no)?
  - **los**: length of hospital stay (i.e in days): a continuous variable.
  - **readm.days**: days from discharge to readmission up to 90 days (i.e in days): If NA, the participants was not readmitted.
  - **death.days**: days from discharge to death (i.e in days): If NA, the participants has not passed away.

The next two pages contain a description of the analysis dataset. Please look over these summaries carefully.

End Part II.

The dataframe is called VICSData and it can be loaded with: load("VICSData.Rdata")

VICSData
18 Variables 1239 Observations

```
id
      n  missing  unique  Info   Mean    .05    .10    .25    .50     .75     .90     .95
   1239        0    1239     1  715.6   74.9  140.8  353.5  716.0  1074.5  1291.2  1371.1

lowest :    1    2    3    4    5, highest: 1451 1452 1453 1458 1459
```

```
age : Age
      n  missing  unique  Info  Mean  .05  .10  .25  .50  .75  .90  .95
   1239        0      70     1  60.3   39   44   52   61   69   76   81

lowest : 20 22 23 25 26, highest: 87 88 90 91 95
```

```
gender2 : Gender
      n  missing  unique
   1239        0       2

Male (684, 55%), Female (555, 45%)
```

```
racecat1 : Race
      n  missing  unique
   1236        3       3

White (1030, 83%), AA (186, 15%), Other (20, 2%)
```

```
transfer : Transferred from another hospital
      n  missing  unique
   1224       15       2

No (761, 62%), Yes (463, 38%)
```

```
diagnosis1 : Diagnosis
      n  missing  unique
   1239        0       3

ACS (797, 64%), HF (352, 28%), Both (90, 7%)
```

```
edu : Years of school completed
      n  missing  unique  Info   Mean  .05   .10   .25   .50   .75   .90   .95
   1239        0      23  0.96  13.52  9.0  11.0  12.0  13.0  15.0  17.2  18.0

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

```
stofhla : sTOFHLA
      n  missing  unique  Info   Mean    .05    .10    .25    .50    .75    .90    .95
   1198       41      36  0.99  29.41  11.85  16.00  26.00  33.00  35.00  36.00  36.00

lowest :  0  2  3  4  5, highest: 32 33 34 35 36
```

```
snsmean : Subjective Numeracy
      n  missing  unique  Info  Mean    .05    .10    .25    .50    .75    .90    .95
   1238        1      16  0.99  4.38  1.667  2.333  3.667  4.667  5.583  6.000  6.000

1 (28, 2%), 1.33333333333333 (14, 1%), 1.66666666666667 (26, 2%)
2 (36, 3%), 2.33333333333333 (34, 3%), 2.66666666666667 (54, 4%)
3 (49, 4%), 3.33333333333333 (67, 5%), 3.66666666666667 (67, 5%)
4 (111, 9%), 4.33333333333333 (93, 8%), 4.66666666666667 (102, 8%)
5 (121, 10%), 5.33333333333333 (126, 10%), 5.66666666666667 (118, 10%)
6 (192, 16%)
```

```
elix : Elixhauser score
      n  missing  unique  Info   Mean  .05  .10  .25  .50  .75  .90  .95
   1224       15      55     1  11.92   -1    0    3   11   19   25   31

lowest : -10  -8  -7  -5  -4, highest:  44  47  50  53  56
```

```
hosp12m : Number of times hospitalized in the 12 months preceding VICS
       n  missing   unique    Info    Mean   .05   .10   .25   .50   .75   .90   .95
    1236        3       14    0.88   1.386     0     0     0     1     2     4     5

            0   1   2  3  4  5  6 7 8 9 10 11 12 14
Frequency 586 243 160 94 58 38 25 8 6 6  4  2  4  2
%          47  20  13  8  5  3  2 1 0 0  0  0  0  0
```

```
maritalstatus1 : Marital Status
       n  missing   unique
    1239        0        3

Married/living with partner (759, 61%)
Separated/Divorce/Single (330, 27%), Widowed (150, 12%)
```

```
nhome : Number of people in the household other than you.
       n  missing   unique    Info    Mean   .05   .10   .25   .50   .75   .90   .95
    1239        0       11    0.87     1.4     0     0     1     1     2     3     4

            0   1   2   3  4  5 6 7 8 11 14
Frequency 243 600 205 109 50 19 3 4 4  1  1
%          20  48  17   9  4  2 0 0 0  0  0
```

```
essi6sum : ENRICHD Social Support Instrument (sum of 6 of the 7 items)
```

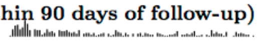
```
       n  missing   unique    Info    Mean   .05   .10   .25   .50   .75   .90   .95
    1239        0       25    0.98   25.79    16    19    24    27    29    30    30

lowest :  6  7  8  9 10, highest: 26 27 28 29 30
```

```
icu
       n  missing   unique
    1224       15        2

No (844, 69%), Yes (380, 31%)
```

```
los : length of hospital stay (in days)
```

```
       n  missing   unique    Info    Mean     .05     .10     .25     .50     .75     .90     .95
    1239        0      612       1   4.435   1.110   1.250   1.880   3.040   5.480   8.888  11.216

lowest :  0.03  0.17  0.29  0.48  0.54
highest: 31.46 34.94 36.35 54.96 91.07
```

```
readm.days : days to readmission within 90 days (if NA, then not readmitted within 90 days of follow-up)
```



```
       n  missing   unique    Info    Mean   .05   .10   .25   .50   .75   .90   .95
     368      871       88       1   33.23     2     4     9    25    55    76    83

lowest :  0  1  2  3  4, highest: 85 86 87 88 90
```

```
death.days : days to death after discharge (if NA, then still alive)
```

```
       n  missing   unique    Info    Mean    .05    .10    .25    .50    .75    .90    .95
     192     1047      177       1   396.8   23.0   45.0  138.0  308.5  603.2  866.8  984.4

lowest :    6    7   10   11   12, highest: 1072 1148 1177 1189 1260
```