Biostatistics 2\textsuperscript{nd} year Comprehensive Examination

Due: May 30\textsuperscript{th}, 2014 by 5pm.

Instructions:
1. The exam is divided into two parts. There are 6 questions in section I and a data analysis in section II.
2. Answer each question to the best of your ability.
3. Be as specific as possible and write as clearly as possible.
4. 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 statistical oracle, nor may you seek help from another individual on the internet (e.g., no posting questions to chat rooms or message boards).
5. 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.
6. Turn in your exam by emailing it to Professor Blume at j.blume@vanderbilt.edu \textbf{AND} Linda Wilson at linda.l.wilson@vanderbilt.edu. You must get confirmation from either Professor Blume or Ms. Wilson that your exam was received before 5pm on Friday 5/30. Alternatively, you may turn in a hard copy to either person by the deadline.
7. 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>Section II</td>
<td>180</td>
<td>60 pts per scientific question</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>360</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Section I

1. Suppose that

$$\varphi_X(t) = \frac{3\sin t}{t^3} - \frac{3\cos t}{t^2}, \quad t \neq 0$$

   a. Show that $X$ is symmetric.
   b. Show that for $n \geq 1$, $E[X^{2n}] = 3/(2n + 1)(2n + 3)$.
   c. Prove that $P(|X| > 1) = 0$.
   d. Show that $X$ is absolutely continuous.

2. Let $X_1, X_2, \ldots$ be i.i.d. $N(0,1)$ random variables.
   a. Prove Mills Ratio inequality: show that for any $x > 0$,

   $$\frac{x}{1 + x^2 e^{-x^2/2}} \leq \int_x^{\infty} e^{-y^2/2} dy \leq \frac{1}{x} e^{-x^2/2}$$

   b. Use the right-hand inequality in (a) and one of the Borel-Cantelli lemmas to prove that for every $\epsilon > 0$,

   $$P\left( \frac{X_n}{\sqrt{2\log n}} > 1 + \epsilon, \ i.o. \right) = 0$$

   c. Use the left-hand inequality in (a) and one of the Borel-Cantelli lemmas to prove that for every $\epsilon > 0$,

   $$P\left( \frac{X_n}{\sqrt{2\log n}} > 1 - \epsilon, \ i.o. \right) = 1$$

   d. Explain why (b) and (c) together imply that

   $$P\left( \limsup_{n \to \infty} \frac{X_n}{\sqrt{2\log n}} = 1 \right) = 1$$
3. There are \( t \) different types of coupons available in equal amounts and a collector is seeking to collect one of each. Show that if \( n \geq 1 \) coupons have been collected, then the probability \( p_n \) of having at least one type of each is

\[
p_n = 1 - \sum_{k=1}^{t} (-1)^{k-1} \binom{t}{k} \left( 1 - \frac{k}{t} \right)^n
\]

4. Consider a general linear model in which

\[
E(Y) = X\beta \quad \text{and} \quad Var(Y) = \sigma^2 I
\]

where \( Y \) is \( n \times 1 \), \( X \) is \( n \times p \), \( \beta \) is \( p \times 1 \) and \( I \) is the \( n \times n \) identity matrix. Let \( x_n^T \) denote the last row of \( X \) and let \( X(n) \) denote the \((n-1) \times p\) matrix with the last row of \( X \) excluded, i.e.

\[
X = \begin{bmatrix} X(n) \\ x_n^T \end{bmatrix}
\]

a. Show that

\[
X^TX = X^T(n)X(n) + x_n x_n^T
\]

b. For any \( p \times p \) matrix \( A \) of rank \( p \) and any \( p \times 1 \) vectors \( c \) and \( d \), show that

\[
[A + dc^T]^{-1} = A^{-1} - aA^{-1}dc^TA^{-1}
\]

where

\[
a = \frac{1}{1 + d^TA^{-1}c}
\]

Use this result and the result from (a) to show that the inverse of \( X^TX \) is

\[
[X^T(n)X(n)]^{-1} - [X^T(n)X(n)]^{-1}x_n x_n^T [X^T(n)X(n)]^{-1} \left[ \frac{1}{1 + h(n)} \right]
\]

where

\[
h(n) = x_n^T [X^T(n)X(n)]^{-1} x_n
\]
c. If we write \( Y \) as
\[
Y = \begin{bmatrix} Y(n) \\ y_n \end{bmatrix}
\]
show that
\[
X^TY = X^T(n)Y(n) + x_ny_n
\]

d. Show that the least squares estimate of \( \beta \), call it \( b \), is given by
\[
b(n) + [X^T(n)X(n)]^{-1}x_n[y_n - x_n^Tb(n)] \left[ \frac{1}{1 + h(n)} \right]
\]
where
\[
b(n) = [X^T(n)X(n)]^{-1}X^T(n)Y(n)
\]

e. Use the result of (d) to discuss the impact on the least squares estimates of the addition of an observation.

f. Use the fact that
\[
[X^T(n)X(n)] = X^TX - x_nx_n^T
\]
and the results in (b) to show that the inverse of \([X^T(n)X(n)]\) is given by
\[
[X^TX]^{-1} + [X^TX]^{-1}x_nx_n^T[X^TX]^{-1}\left[ \frac{1}{1 - h_{nn}} \right]
\]
where
\[
h_{nn} = x_n^T[X^TX]^{-1}x_n
\]

\[g.\] Use the result in (f) to show that \( b(n) \) is given by
\[
b - [X^TX]^{-1}x_n(y_n - x_n^Tb) \left[ \frac{1}{1 - h_{nn}} \right]
\]

h. Use the result in (g) to discuss the impact on the least squares estimate of the deletion of an observation and relate the result to the interpretation of the hat matrix.
5. This problem examines Bayes classification and risk. For $Y \in \{0,1\}$ and predictor(s) $X$, let $f(x) = P(Y = 1|X = x)$ and let $\hat{f}(x)$ be an estimate of $f(x)$. Denote the Bayes classifier by $h(x) = I\left(f(x) \geq \frac{1}{2}\right)$ and denote its estimate by $\hat{h}(x) = I\left(\hat{f}(x) \geq \frac{1}{2}\right)$.

a. Show that the risk (i.e., expected loss) of a 0-1 loss function is the misclassification probability.

b. Show that the misclassification probability at $X = x_0$ is given by

$$P(Y \neq \hat{h}(X)|X = x_0) = P(Y \neq h(X)|X = x_0)$$

$$+ |2f(x_0) - 1|P(\hat{h}(X) \neq h(X)|X = x_0)$$

where $P(Y \neq h(X)|X = x_0)$ is the ‘irreducible’ error rate associated with the Bayes classifier.

c. Assuming that $\hat{f}(x_0) \sim N\left(E[\hat{f}(x_0)], \text{Var}[\hat{f}(x_0)]\right)$, show that

$$P(\hat{h}(X) \neq h(X)|X = x_0) = \Phi \left[ \text{sign}\left(\frac{1}{2} - f(x_0)\right) \frac{\left(E[\hat{f}(x_0)] - \frac{1}{2}\right)}{\sqrt{\text{Var}[\hat{f}(x_0)]}} \right]$$

d. Explain the implications of (b) and (c) for the bias-variance tradeoff in estimating $f(x)$ when that estimate, $\hat{f}(x)$, is used for classification purposes.
6. Let $Y_n = (y_1, ..., y_n)^T$ denote a random vector of responses, $X_n$ an $n \times p$ matrix of predictors ($p$ predictors per subject), and $\beta$ a $p \times 1$ parameter vector with $j^{th}$ component $\beta_j$. We are interested in linear models of the form

$$Y_n \sim N(X_n\beta, \sigma^2 I_n)$$

where $I_n$ is the identity matrix and $\sigma^2$ is an unknown constant. The table below lists different penalized least squares approaches for fitting the model. $\lambda$ is a tuning parameter.

<table>
<thead>
<tr>
<th>Method</th>
<th>Objective function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p}</td>
</tr>
<tr>
<td>Ridge</td>
<td>$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$</td>
</tr>
<tr>
<td>Inverse Lasso</td>
<td>$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \frac{1}{</td>
</tr>
<tr>
<td>Inverse Ridge</td>
<td>$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \frac{1}{\beta_j^2}$</td>
</tr>
</tbody>
</table>

a. Show that for each method, when $Y_n \sim N(X_n\beta, \sigma^2 I_n)$, there exists a penalized likelihood that, when maximized, will yield the same solution set for $\beta$. It may be helpful to assume the columns of $X_n$ have been standardized and/or that the data have been centered so that the intercept is identically zero. Be sure to state your assumptions.

b. Propose a method that might reasonably be called “Inverse elastic net”.

The next three questions ask you to speculate about the behavior of these methods and contrast them. Consider supporting your answers with some analytical observations or simple empirical examples (brief simulations or toy examples) when appropriate.

c. Speculate about and describe your anticipated effect of the penalty term in each method. What type of shrinkage might it encourage? Which penalty terms might encourage greater shrinkage and when?

d. Explain how the tuning parameter(s) should be selected. What algorithm would you use to do this? Should the tuning parameter be determined separately for each method? Would a common tuning parameter facilitate comparisons between the methods?

e. What criteria would you suggest for comparing fitted models obtained from these methods? Describe a simulation that would assess these properties.

End Part I.
Section II

Instructions

Prepare a 4 to 6 page statistical analysis report on the following data analysis. A detailed description of the data and scientific aims are described below. Tables and figures should be included in an appendix; they are not counted against the page total. Be sure to include the following sections in your report: Introduction, Methods, Results, and Conclusions. You may cite this description as a source document for background information. Your Methods section should be comprised of your analytical approach and may be proportionately longer than a typical scientific article. Results and interpretations should be well written, formatted and presentable to a scientific investigator. However, they should also include enough statistical details for a statistical reviewer. Check model assumptions and/or other diagnostics as appropriate. Be sure to provide sufficient justification for all modeling decisions.

Introduction

Among patients undergoing cardiac surgery, 20-30% experience kidney injury that is defined by rises in serum creatinine concentrations. Since kidney injury promotes systemic infections, is thought to increase risk of myocardial injury, and is associated with an independent, eight-fold increase in the odds of death at 30 day, investigators wish to understand risk factors for kidney damage. Ultimately, investigators would like to identify effective mitigation strategies that lower the risk of kidney injury and therefore the associated morbidities.

We will consider data from a randomized clinical trial that tested the hypothesis that angiotensin-converting enzyme inhibition (ramipril) or aldosterone receptor antagonism (spironolactone) decreased the incidence of atrial fibrillation after elective cardiac surgery. We have attached a manuscript based on data that substantially overlap those considered here (same study). The manuscript should be useful for background knowledge.

Even though surgical, anesthesia, and intensive care unit care have improved dramatically over time, Acute Kidney Injury (AKI) related morbidity and mortality continues to be on the rise among cardiac surgery patients. A commonly held hypothesis is that the rise in obesity can explain the rise in post surgery AKI. Further, obesity is thought to cause inflammation and oxidative stress (radical oxygen species floating in the blood) that may lead to kidney injury.
Background
Billings et. al. (2012) analyzed a similar data set and reported their findings in the attached paper. Use this paper as background reading for the following scientific questions. The paper can be found at: https://dl.dropboxusercontent.com/u/25204698/comps/Billings2012.pdf

The Scientific Questions
This analysis seeks to address two questions and to build a model for predicting acute kidney injury risk. Specifically, we would like to address the following:

1) To what extent is body mass (as measured by body mass index) associated with creatinine concentrations (as measured by maximum change from baseline) during the four days following surgery?

2) It has been shown that increased amounts of perioperative blood transfusion products during cardiac surgery are associated with worse patient outcomes such as increased postoperative bleeding and increased hospital stay. To what extent are oxidative stress (as measured by f2-isoprostanes) and inflammation (as measured by IL-6) biomarkers associated with the number of intra-operative blood units transfused? To address this question, consider the number of units of red blood cells transfused as a 3-level categorical variable taking values 0, 1-2, and >=3, indicating none, moderate, and excessive blood transfusion, respectively.

3) A 0.3 mg/dl change in creatinine concentrations is regarded as a substantial change and is indicative that kidney injury has taken place. Construct and evaluate a risk model that seeks to identify those most likely to develop kidney injury at the time the surgery has ended. Based on this model, if an intervention were developed to reduce the risk of AKI for all those with BMI >25 kg/m² to that of a person with a BMI of 25 kg/m² (all else being equal), how many fewer patients are expected to have experienced AKI if applied to all patients in this sample? Provide a graphical comparison of the distribution of risk between the observed sample and the hypothetical (intervention-based) sample

Note: A key component of these questions is the choice of what variables should and should not be included in a given model. Be sure to provide sufficient justification for all modeling decisions.
The Data  (https://dl.dropboxusercontent.com/u/25204698/comps/AKI.Rdata)

bmi0: body mass index (kg/m2). Measured at randomization.

hx_dm: history of type II diabetes mellitus

il6.0: il6 concentrations measured at the time of randomization

il6.3: peri-operative il6 concentrations, measured at end of surgery.

f2i.0: f2-isoprostane concentrations measured at the time of randomization

f2i.3: peri-operative f2-isoprostane concentrations, measured at end of surgery.

Tx: treatment group: 1=placebo, 2=ramipril, 3=spironolactone

gender: 0=Female, 1=Male

hx_dm: History of type 2 diabetes; 0=No, 1=Yes (History of diabetes)

valve: Valve surgery; 0=No, 1=Yes

race: Patient race; 0=Caucasian, 1=African American

cpb: Cardiac bypass surgery as part of the surgery; 0=No, 1=Yes

cpb.tm: Time on bypass (mins)

prbc.xfus: Number of units of red blood cells transfused.

crt0: creatinine concentration measured just prior to surgery

crt.pod1: creatinine concentrations measured on post-op day 1

crt.pod2: creatinine concentrations measured on post-op day 2

crt.pod3: creatinine concentrations measured on post-op day 3

crt.pod4: creatinine concentrations measured on post-op day 4

Important note: If a patient was discharged from the hospital prior to POD 4, their creatinines will be missing. On can assume that creatinine was normal on such days (i.e., set it equal to the baseline concentration on such days).

End Section II.