Department of Statistics
University of Wisconsin-Madison
PhD Qualifying Exam Part II
January 15, 2009
1:00–4:00pm, Room 133 SMI

- There are a total of FOUR (4) problems in this exam. Please do a total of TWO (2) problems.
- Each problem must be done in a separate exam book.
- Please turn in TWO (2) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. **Definitions:** Let $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$. The sequence $(M_n)_{n\geq 1}$ is a martingale adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$ if M_n is $\mathcal{F}_n$-measurable, integrable and if

$$E(M_{n+1}|\mathcal{F}_n) = M_n \quad \text{for all } n \geq 1.$$

A stopping time is a map $\tau : \Omega \to \{1, 2, 3, \ldots, +\infty\}$ such that $\{\tau = n\} \in \mathcal{F}_n$ for all $n \geq 1$.

(a) Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. integrable random variables. Let $S_n = \sum_{i=1}^n X_i$ and $\mathcal{A}_n = \sigma(S_n, S_{n+1}, \ldots)$. Determine, for all integers i and n,

$$E(X_i|S_n), \quad E(X_i|\mathcal{A}_n) \text{ and } E(S_n|S_i)\,.$$

(b) Let $(X_n)_{n\geq 1}$ be a sequence of independent coin tosses with $P(X_n = 1) = p$ and $P(X_n = -1) = 1 - p$. Let $S_n = \sum_{i=1}^n X_i$. For any a and b positive integers, we define

$$\tau_a = \inf\{n \geq 1 : S_n \geq a\} \quad \text{and} \quad \tau_{a,-b} = \inf\{n \geq 1 : S_n \geq a \text{ or } S_n \leq -b\}$$

where the infimum of an empty set is $+\infty$.

i. Show that τ_a and $\tau_{a,-b}$ are stopping times.

ii. Show that $\tau_{a,-b}$ is almost surely finite.

iii. Determine the values of p for which τ_a is almost surely finite.

iv. Consider $p = 1/2$ for this question. Show that $M_n = S_n^2 - n$ is a martingale.

v. Find constants λ_n and η_n such that $M_n = (S_n - \lambda_n)^2 - \eta_n$ is a martingale.

2. The following problem consists of parts (a) and (b), where all dimensions involved are finite.

   (a) For a positive definite matrix M and vectors $\mathbf{d}$, $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$, define

   $$L(\boldsymbol{\theta}) = \|M(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \mathbf{d}\|_2^2.$$

   i. Find the minimizer $\widetilde{\boldsymbol{\theta}}$ of $L(\boldsymbol{\theta})$.

   ii. For $\widetilde{\boldsymbol{\theta}}$ in part (a)i, find a nontrivial lower bound of $L(\boldsymbol{\theta}) - L(\widetilde{\boldsymbol{\theta}})$ in terms of $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2^2$ for fixed M, $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\theta}}$, and show that this bound is independent of $\mathbf{d}$.

   (b) Suppose that for $n = 1, 2, \ldots$, $A_n(\boldsymbol{s})$ is a convex function of the form, $A_n(\boldsymbol{s}) = 2^{-1}\boldsymbol{s}^T\mathbf{V}\boldsymbol{s} + \mathbf{u}_n^T\boldsymbol{s} + C_n + r_n(\boldsymbol{s})$, where $\boldsymbol{s}$ is a vector,

   - $\mathbf{V}$ is a non-random positive definite matrix (not depending on $\boldsymbol{s}$),
   - $\mathbf{u}_n$ is a random variable (not depending on $\boldsymbol{s}$) and $\mathbf{u}_n = O_P(1)$,
   - C_n is an arbitrary random or non-random quantity that does not depend on $\boldsymbol{s}$, and
   - $r_n(\boldsymbol{s}) \xrightarrow{P} 0$ for each $\boldsymbol{s}$.

   Let

   $$\begin{aligned} \widehat{\boldsymbol{\alpha}}_n &= \arg\min_{\boldsymbol{s}} A_n(\boldsymbol{s}), \\ \widehat{\boldsymbol{\beta}}_n &= \arg\min_{\boldsymbol{s}} \left\{ \frac{1}{2}\boldsymbol{s}^T\mathbf{V}\boldsymbol{s} + \mathbf{u}_n^T\boldsymbol{s} + C_n \right\}. \end{aligned}$$

   i. Show that

   $$\widehat{\boldsymbol{\alpha}}_n = \widehat{\boldsymbol{\beta}}_n + o_P(1).$$

   (Hint: you may use the result of part (a)ii. If you cannot solve part(a)ii, you may assume that $\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\|_2^2 \leq 1000\{L(\boldsymbol{\theta}) - L(\widetilde{\boldsymbol{\theta}})\}$.)

   ii. Assume $\mathbf{u}_n \xrightarrow{\mathcal{L}} \mathbf{u}$. Derive the asymptotic distribution of $\widehat{\boldsymbol{\alpha}}_n$ as $n \to \infty$.

3. The following results were obtained from a completely randomized experiment on the yield of three hybrids of corn subjected to four different fertilizer treatments.

		Hybrid type		
		1	2	3
Fertilizer type	a	71	57	54
	b	74	68	75
	c	57	57	68
	d	93	84	90

(a) Complete the following ANOVA table:

Source	Df	SumSq	MeanSq	F-value
Fertilizer	3			
Hybrid	2			
Residual	6			

(b) The fertilizer treatments were composed of nitrogen (N) and potassium (K) in the following factorial arrangement:

		Levels of K	
		0	60
Levels of N	60	a	b
	200	c	d

This suggests the ANOVA decomposition:

Source	Df	SumSq
Fertilizer	3	
N	1	
K	1	
N × K	1	
Hybrid	2	
Residual	6	

Fill in the sums of squares for the table.

(c) The experimenter is particularly interested in the nitrogen response and wants the ANOVA table:

Source	Df	SumSq
Fertilizer	3	
Between K	1	
Within K	2	
Between N in K=0	1	
Between N in K=60	1	
Hybrid	2	
Residual	6	

Fill in the sums of squares for this table and explain how they are related to the ones in the previous table.

(d) The experimenter decides to fit a regression model to the data using the variables $x_1 \in \{60, 200\}$ for nitrogen level and $x_2 \in \{0, 60\}$ for potassium level:

$$E(y) = \eta + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \gamma_1 I(\texttt{Hybrid} = 2) + \gamma_2 I(\texttt{Hybrid} = 3).$$

Find the least-squares estimates of β_1, β_2, and β_{12}.

4. An experiment was conducted to study the relationship between the growth of gerbil embryos in the laboratory to various formulations of the growth medium. In particular, embryos were grown in media prepared with different levels of several amino acids: taurine (`tau`), glycine (`gly`), glutamine (`glu`), histidine (`his`), and hypotaurine (`hyp`). Growth was measured by counting nuclei (`nucl`) in the developing embryo.

   Use the following results from R to answer the questions that follow.

```
> print(dat)

   tau gly glu his hyp nucl
1    9   5  13  28  27   19
2    2  18  23   3  29   15
3   12  29  26  23  18   27
4   11  21  24  17  20   23
5    2  17  25  27  12   18
6   16   6  19  27   7   21
7    4  18  18   9  22   19
8   26  20  19  25   1   28
9    3   9  19  24  17   17
10  10   6  27  29  10   20
11   3   8   2  29   7   12
12  19  30   6  13  13   27
```

```
> cor(dat)
             tau         gly         glu         his         hyp         nucl
tau    1.0000000  0.32558432 -0.07383340  0.15473029 -0.53244181  0.85032214
gly    0.3255843  1.00000000  0.08570427 -0.53235326  0.08353151  0.64574335
glu   -0.0738334  0.08570427  1.00000000 -0.05179112  0.17573905  0.20294565
his    0.1547303 -0.53235326 -0.05179112  1.00000000 -0.58321040 -0.01474397
hyp   -0.5324418  0.08353151  0.17573905 -0.58321040  1.00000000 -0.26838364
nucl   0.8503221  0.64574335  0.20294565 -0.01474397 -0.26838364  1.00000000
```

```
Call:
lm(formula = nucl ~ tau + gly + glu + his + hyp)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  3.31154    2.92610   1.132 0.300944
tau          0.51109    0.06830   7.483 0.000294
gly          0.28607    0.06181   4.628 0.003585
glu          0.12067    0.05171   2.334 0.058354
his          0.16095    0.07086   2.271 0.063556
hyp          0.13890    0.07041   1.973 0.095999

Residual standard error: 1.313 on 6 degrees of freedom

Analysis of Variance Table

Response: nucl
          Df  Sum Sq Mean Sq  F value    Pr(>F)
tau        1 197.392 197.392 114.5284 3.929e-05
gly        1  41.555  41.555  24.1107  0.002683
glu        1  13.522  13.522   7.8459  0.031121
his        1   3.482   3.482   2.0205  0.205017
hyp        1   6.707   6.707   3.8913  0.095999
Residuals  6  10.341   1.724

Call:
lm(formula = nucl ~ tau + gly + glu + hyp)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.95270    1.95324   4.584 0.002533
tau          0.51900    0.08612   6.026 0.000528
gly          0.20510    0.06376   3.217 0.014714
glu          0.13640    0.06470   2.108 0.072974
hyp          0.05107    0.07429   0.687 0.513985

Residual standard error: 1.658 on 7 degrees of freedom

Call:
lm(formula = nucl ~ gly + his + hyp)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.70402    8.10281   1.198   0.2653
gly          0.48247    0.16326   2.955   0.0183
```

```
his          0.20890    0.20284   1.030   0.3332
hyp         -0.07503    0.17635  -0.425   0.6817

Residual standard error: 3.797 on 8 degrees of freedom

Analysis of Variance Table

Response: nucl
          Df  Sum Sq Mean Sq F value  Pr(>F)
gly        1 113.837 113.837  7.8976 0.02283
his        1  41.241  41.241  2.8612 0.12920
hyp        1   2.609   2.609  0.1810 0.68171
Residuals  8 115.313  14.414
```

(a) I fit the model $\texttt{nucl} = b_0 + b_2\texttt{gly} + b_1\texttt{tau} + b_3\texttt{glu} + e$. What is R^2 for this model?

(b) The adjusted R^2 for a given model be calculated as:

$$R^2_{\text{adj}} = 1 - (1 - R^2_u)\frac{n-1}{\text{dfError}}$$

where R^2_u is the usual R^2.

Consider the model:

$$\texttt{nucl} = b_0 + b_2\texttt{gly} + b_4\texttt{his} + b_5\texttt{hyp} + e.$$

i. Find the observed value of R^2_{adj} for this model.

ii. Denote the observed value found in the preceding part by $\widetilde{R}^2_{\text{adj}}$. Consider repetitions of this experiment (with the same sample size); and for each repetition, imagine fitting the model

$$\texttt{nucl} = b_0 + b_2\texttt{gly} + b_4\texttt{his} + b_5\texttt{hyp} + e$$

as before. In that case, R^2_{adj} can be considered a random variable. Assuming that the data from the experiments indeed arise from the model being fitted, find the probability that

$$R^2_{\text{adj}} > \widetilde{R}^2_{\text{adj}}$$

conditional on the value of $\widetilde{R}^2_{\text{adj}}$ observed in the actual experiment described above.

(c) Consider the model $\texttt{nucl} = b_0 + b_1\texttt{tau} + b_2\texttt{gly} + b_3\texttt{glu} + b_4\texttt{his} + e$. Test the hypothesis: $H_0 : b_2 = b_3 = b_4 = 0$.

(d) Consider the model $\texttt{nucl} = b_0 + b_1\texttt{tau} + b_2\texttt{gly} + b_3\texttt{glu} + b_5\texttt{hyp} + e$.
Test the hypothesis: $H_0 : b_1 = b_3 = 0$.

(e) Consider the model $\texttt{nucl} = b_0 + b_4\texttt{his} + e$. Peform a lack of fit test for this model.

(f) The experimenters admit that the choice of the levels of the amino acids was made quite haphazardly, with little intentional design. They ask you to comment on whether this is a problem or not, and whether you would advise them to choose the levels differently if they were to repeat the experiment.