SPRING 2016

HUNTER COLLEGE

STAT 707

General Linear Model II

Final Exam

Last Name: ______________________________

First Name: ______________________________

May 23, 2017

Instructions

- There are 4 questions, each question is between 5-6 points. The maximal score is 22 points. A perfect score is **20** points.
- There are two versions of Question 2, 3 with different points. Please choice one version and only one to answer.
- Show all work. You will receive partial credit for partially completed problems.
- You may use any references.

Q1. Regression (5pt).
Consider a hierarchical linear model without random effects.

$$
\begin{aligned}
Y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + R_{ij},\\
\beta_{0j} &= \gamma_{00} + \gamma_{01}z_j,\\
\beta_{1j} &= \gamma_{10} + \gamma_{11}z_j,\\
R_{ij} &\sim N(0, \sigma^2),
\end{aligned}
$$

where $i = 0, 1, 2,\ j = 1, 2, 3$. Note here the index i starts from 0. The level-2 explanatory variable z_j is the observation of Y_{ij} when $i = 0$. Say, $z_j = y_{0j}$. Additionally, we have $x_{0j} = 0$. Denote $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$. Suppose that we have two maximum likelihood estimates of $\boldsymbol{\gamma}$, one is $\breve{\boldsymbol{\gamma}}$, where the estimation is based on the full data set $\{y_{ij}, z_j = y_{0j}, x_{ij}\}_{i=0,1,2,j=1,2,3}$; the other is $\tilde{\boldsymbol{\gamma}}$, where the estimation is based on the partial data set without the baseline data points $\{y_{ij}, z_j = y_{0j}, x_{ij}\}_{i=1,2,j=1,2,3}$.
(1) Show $\mathbb{E}[\breve{\boldsymbol{\gamma}}] = \mathbb{E}[\tilde{\boldsymbol{\gamma}}] = \boldsymbol{\gamma}$, say, the MLEs based on the full data set and based on the partial data set are both unbiased.
(2) Suppose that $y_{01} = 20$, $y_{02} = 24$, $y_{03} = 27$, and $x_{1j} = 1$, $x_{2j} = 2$. Numerically show that $\mathrm{var}[\breve{\boldsymbol{\gamma}}] < \mathrm{var}[\tilde{\boldsymbol{\gamma}}]$
Hint: `R` functions: transpose of A: `t(A)`; inverse of A: `solve(A)`; matrix multiplication: `A %*% B`.

Q2. Multilevel modeling: GLMM and GEE.

Please **choose only one** question from 2-A (5 pts) and 2-B (6 pts). Circle your option.

The table below shows a data set of time of half marathon of 6 amateur athletes. These athletes belong to one of two running clubs during one running event.

Athlete	Age (yr)	Club	Time (min)
1	38	0	95
1	40	0	94
1	43	1	93
2	53	1	96
2	55	1	98
2	56	1	91
2	58	1	93
3	37	1	83
3	40	1	82
3	42	0	82
4	41	0	91
4	45	1	94
4	46	0	99
5	54	1	105
5	58	1	111
6	57	1	90
6	60	1	89
6	62	1	95

Apply GLMM with Gaussian family and identity link to this data set to fit two models:

Model 1: `log(Time)` $\sim$ `Club + log(Age) + (1|Athlete)`

Model 2: `log(Time)` $\sim$ `log(Age) + (1|Athlete)`

Apply GEE with Gaussian family and identity link to this data set to fit two models:

Model 3: `log(Time)` $\sim$ `Club + log(Age)`

Model 4: `log(Time)` $\sim$ `log(Age)`

where the correlation matrix structure is exchangeable.

Report the estimates of these models. Briefly explain your estimations.

2-A. Additionally, test the significance of factor "club" by using either t-test or deviance test.

2-B. Additionally, test the significance of factor "club" by using both t-test and deviance test.

Q3. Spatial data analysis: Kriging.
Please **choose only one** question from 3-A (5 pts) and 3-B (6 pts). Circle your option.
Consider a 1-D Kriging regression. There are three observations along a line. The locations are $x1 = 0$, $x_2 = 3$, $x_3 = 10$. The coal qualities at these three locations are $z_1 = 7$, $z_2 = 6$, $z_3 = 10$. Find the Kriging prediction of coal quality $\widehat{z}_0$ at location $x_0 = 4$.

Hint: $\widehat{z}_0 = \sum_{i=1}^{3} \lambda_i z_i$. Denote the semivariogram between point i and point j as γ_{ij}. The semivariograms in the linear system for solving λ_i's are from fitted relation between semivariogram γ and distance d. Say, $\gamma_{ij} = \gamma(d_{ij})$, γ a function of d.

3-A. Use $\gamma = 1 + 2\sqrt{d}$ to find the Kriging prediction. Note, when $d = 0$, it follows that $\gamma = 1$. Find the Kriging estimation based on this $\gamma \sim d$ relation.

3-B. Suppose the relation between γ and d is $\gamma = 1 + k\sqrt{d}$, where k is a parameter. Determine k by using least-squares fitting. Find the Kriging estimation based on your fitted $\gamma \sim d$ relation.
Hint: Run a least-squares fitting without intercept between $\gamma - 1$ and $\sqrt{d}$.

Q4. Observational study: Propensity score matching (5pt).
There are twenty participants who are assigned to either control group ($z = 0$) or treatment group ($z = 1$), and the observed covariates include Age and Family history of disease.
Calculate the propensity scores, which are the conditional probability given vectors of observed covariates. Make a plot of Propensity score versus Group.

ID	Age	Family history of disease	Group	ID	Age	Family history of disease	Group
1	≤ 30	N	C	11	≤ 30	Y	T
2	≤ 30	N	C	12	$30 - 50$	N	C
3	≤ 30	N	C	13	$30 - 50$	N	C
4	≤ 30	N	C	14	$30 - 50$	Y	T
5	≤ 30	N	C	15	$30 - 50$	Y	T
6	≤ 30	N	Y	16	$30 - 50$	Y	T
7	≤ 30	N	Y	17	$30 - 50$	Y	T
8	≤ 30	Y	C	18	≥ 50	Y	C
9	≤ 30	Y	T	19	≥ 50	Y	T
10	≤ 30	Y	T	20	≥ 50	Y	T

Please list statistical topics which you are interested in or which are useful to you, especially which you have not learned from STAT 703, 706, & 707.

End of the final exam of Stat 707 (Instructor: Jiangtao Gou)