STA/BST 224 Longitudinal Data Analysis

Problem Set 3
DUE: May. 19, 2020 (Tue)

Instruction:

- Please submit it through Canvas. You can scan or take picture of hand-written solution as long as it is clear.
- You can use either R/SAS/Stata or other software. Please include program and important results.
- The grade will be average of all problems. Due to policy of Graduate Studies, all students need to work on the same problems. You can work together on the homework and ask TA for help. However, each of you is responsible for your own statistical programming and for writing-up your solutions in your own words.

1. For the model

$$Y_{ij} = \beta_0 + \beta_1 \texttt{endog}_i + \beta_2 \texttt{week}_{ij} + \beta_3 \texttt{endog}_i * \texttt{week}_{ij} + U_{i1} + U_{i2}\texttt{week}_{ij} + Z_{ij}$$

   Use the results on Page 18-21 of Note 8 (note that, $\widehat{\text{var}}(Z_{ij}) = 12.21$), estimate:

   (a) $\text{var}(Y_{ij})$ and $\text{var}(Y_{ik})$ at two different time points, t_j and t_k
   (Note: $\text{var}(Y_{ij})$ is actually $\text{var}(Y_{ij}|X_i)$ because X is considered as fixed, while U and Z are random)

   (b) $\text{cov}(Y_{ij}, Y_{ik})$

   (c) $\text{corr}(Y_{ij}, Y_{ik})$

   (d) $\text{var}(Y_{ij}|U_i)$ (ie, $\text{var}(Y_{ij}|X_i, U_{i1}, U_{i2})$). Please explain the its difference with $\text{var}(Y_{ij})$.

2. The Study of Assets and Health Dynamics Among the Oldest Old (AHEAD) is a national panel study (longitudinal) with initial sample of 7444 respondents aged 70 years and older, and their spouses (if married). Objectives of the study include: (i) to monitor transitions in physical, functional, and cognitive health; (ii) to examine the relationship of late-life changes in physical and cognitive health to patterns of dissaving and income flows; (iii) to relate changes in health to economic resources and intergenerational transfers; (iv) to examine how the mix and distribution of economic, family, and program resources affect key outcomes, including institutionalization, dissaving, and health declines.

   In this problem, we will look at data from the first four waves of the study, collected in 1993, 1995, 1998 and 2000. A reduced data set is on the course web page (`aheadld.csv`). Some of the variables are described here.

id	subject ID
year	years since 1993
sex	1=Male, 2=Female
age	baseline Age (in years)
immword	immediate word recall score
delword	delayed word recall score
blks	difficulty walking several blocks without help
strs	difficulty climbing flight of stairs without help
push	difficulty pulling or pushing a living room chair
bag	difficulty lifting a bag of groceries
dime	difficulty picking up a dime

In this problem, we will explore the relationship of a test of cognitive function to some physical functioning indicators. The hypothesis is that, as physical function declines, movement becomes more difficult, and hence there are fewer stimuli, leading to cognitive decline.

(a) Download the data and do following data manipulation.

i. Notice that the age variable is really baseline age, so you should create a time-varying age variable (`realage`) by adding together age and year. For the remainder of the problem, this time-varying age will be our time scale of interest.

ii. Create a cognitive function score (`totword`) by summing up the immediate word recall (`immword`) and the delayed word recall (`delword`) scores. Each of these scores is the number (out of 10) of words that the subject can recall after hearing them. The first is the number recalled immediately after hearing them. The second is after a delay of a few minutes.

(b) Using an exchangeable correlation model, with total word recall as your response, fit a regression model using ReML that includes age (time-varying, centered at 80 before fitting), sex (you can create a dummy variable for indicator of female), their interaction, and main effects for the five "difficulty" indicators. Use model-based variance-covariance estimator. Provide confidence intervals and interpret the estimated coefficients for age, sex, their interaction and `blks` variable (ie, marginal interpretation as for cross-sectional data).

(c) From your model fit, estimate the sex effect for 70 year olds, provide a confidence interval for this estimate and interpret this estimated effect. (Hint: if age is centered at 80, this effect should be 1×coefficient of female −10×coefficient of age and sex interaction.)

(d) Using a $\mathcal{F}$-test, test the hypothesis that the five physical function variables are jointly associated with total word recall. State null and alternative hypotheses, provide a test statistic, degrees of freedom and P-value. Draw conclusions.
Note:
1. Software may use different calculation methods for denominator DF.
2. If software is not able to obtain denominator DF to do $\mathcal{F}$-test, you may approximately do a χ^2 test: χ^2 = (`numerator DF`) * $\mathcal{F}$, which has an approximate χ^2 distribution with DF=numerator DF, (ie, normalized χ^2 is the limiting distribution of $\mathcal{F}$ when denominator DF goes to infinity).

(e) Treating the exchangeable correlation structure as a *working correlation model*, refit your model from Part (b) using a robust (empirical) variance-covariance estimator. Provide confidence intervals and interpret the estimated coefficients for age, sex, their interaction and `blks` variable. Explain why you do not have to have the correlation structure exactly correct in order to obtain valid inferences.

3. The data for this problem are a subset of a larger data set. This subset consists "of the birth weights of babies who were born to 878 mothers from the state of Georgia, USA, all of whom has five babies. The analyses here focus on the effect of mother's age at birth." (Neuhaus and McCulloch, 2006).

Background: "Until recently, most population-based studies of women's reproductive experiences over pregnancies were based on cross-sectional, rather than longitudinal data. Adams et al. (1997) showed that such cross-sectional data analyses can lead to false conclusions that the analysis of longitudinal data would have avoided. For example, cross-sectional analyses indicated that perinatal mortality increases with each subsequent birth whereas the analysis of longitudinal data indicates the opposite: for a given women, the risk of perinatal mortality decreases with each of her subsequent pregnancies." (Neuhaus and McCulloch, 2006).

The raw data file `birthwt.raw` on the class website contains 5 observations on each of 878 mothers. The columns are: subject id (`id`), birth order (`birthorder`), birth weight (in gm) (`birthwt`), mother's age (`momage`), mother's average age (`momage_avg`), deviation from mother's age (`momage_dev`).

(a) Read the data into software and do following data manipulation: confirm the birth order is from 1 to 5 for mothers (can check frequency of `birthorder`). Divide the three mother's age variables by 10 so that they represent age in decades. As birth weight sometimes varies by birth order, create an indicator variable for "first birth" for each mother.

(b) Use maximum likelihood to fit a linear model with a random mother-specific intercept to these data. The response variable is birth weight. Your model should include mother's age and the first birth indicator.

i. Write down the model and indicate the random effect and assumptions about that random effect.

ii. Present the fitted model.

iii. Interpret the two regression coefficients in the model (ie, marginal interpretation as for cross-sectional data). From this model, what does one conclude about the association of birthweight to maternal age? How about first-born effect?

(c) Fit a linear fixed effects model with a mother-specific intercept to these data. Again, your model should include mother's age and the first birth indicator.

i. Explain any differences between this model and the model you wrote down in part (b)(i).

ii. Present the fitted model.

iii. Interpret the two regression coefficients in the model and explain how their interpretation differs from those in part (b)(iii). How do your conclude about the association of birthweight to maternal age change based on this new model? How about first-born effect?

iv. If the results from parts (b) and (c) diverge, provide some explanation as to why they are different.