**Homework 3 Due Monday by 1:10 pm, 3/04/2019**

**Miao Cai**

For the data used in homework 2, use the GEE model to assess the association between bmipct and all level-1 and level-2 independent variables. Answer the following questions.

1. Use the exchangeable and toep working correlation structures. Compare the resulting QIC values. Which structure seems better?
2. Compare the GEE model using an exchangeable working correlation with the single-level model. In general, which model produces greater standard errors of parameters?
3. Use the **backward** selection procedure to decide on the best model for these data by using the QIC and QICu as the selection criteria. Follow the instructions below:
4. Start with the model including all level-1 and level-2 independent variables. Use the exchangeable working correlation structure.
5. From the model in 1), remove the most insignificant variable (with the largest p value larger than 0.05). Refit the model and obtain QIC and QICu. If the QIC and QICu values are smaller than those in 1), then the model at this step is a better one. Otherwise, if the QIC and QICu values are substantially greater than those in 1), then keep this most insignificant variable regardless of its significance and return to model 1).
6. From the model selected at 2), remove the next most insignificant variable and refit the model and obtain QIC and QICu. If the QIC and QICu values are smaller than those in 2), then the model at this step is a better one. Otherwise, if the QIC and QICu values are substantially greater than those in 2), then keep this most insignificant variable regardless of its significance and return to model 2).
7. Repeat in this way until no more variables can be removed.
8. Report the parameter estimates from the final model.

**1. Step 1: Using the exchangeable correlation structure**:

/\*\*\*\*\*\*\*\*\*\* Q1 \*\*\*\*\*\*\*\*\*\*/

/\* Q1.1 Exchangeable correlation matrix \*/

**proc** **genmod** data = HW3.DATHW2clean;

class school\_ID meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime

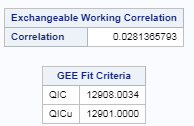
Region Urban Schooltype PctMinority;

model bmipct = meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV

Region Urban Schooltype PctMinority /type3 dist=normal;

repeated subject = school\_ID/type=exch corrw ;

**run**;



**Step 2: Using the toep correlation structure**:

/\* Q1.2 Exchangeable correlation matrix \*/

**proc** **genmod** data = HW3.DATHW2clean;

class school\_ID meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime

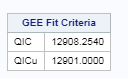
Region Urban Schooltype PctMinority;

model bmipct = meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV

Region Urban Schooltype PctMinority /type3 dist=normal;

repeated subject = school\_ID/type=toep corrw ;

**run**;



Using the two different correlation structures, it seems that using **the exchangeable correlation structure is slightly better than using the toep correlation structure** since the QIC is slightly smaller.

**2. Parameter estimates using GEE model:**

/\* Q2.1 Exchangeable correlation matrix \*/

**proc** **genmod** data = HW3.DATHW2clean;

class school\_ID meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime

Region Urban Schooltype PctMinority;

model bmipct = meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV

Region Urban Schooltype PctMinority /type3 dist=normal;

repeated subject = school\_ID/type=exch corrw;

**run**;



**Parameter estimates using the single-level model**:

/\* Q2.2 single-level analysis \*/

**proc** **glm** data=HW3.DATHW2clean;

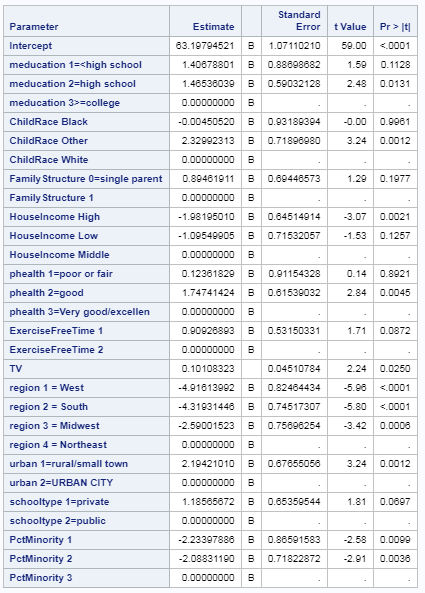
class school\_ID meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime

Region Urban Schooltype PctMinority;

model bmipct = meducation ChildRace FamilyStructure HouseIncome pHealth ExerciseFreeTime TV

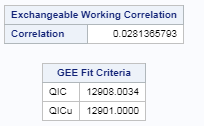
Region Urban Schooltype PctMinority/solution;

**run**;

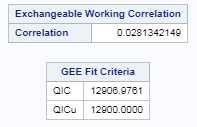


After comparing the standard errors of the two models, it seems that the **GEE models produce estimates with greater standard errors**.

3.

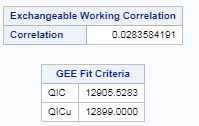
1) Step 1: the most insignificant variable is **FamilyStructure**, with the p-value = 0.2016. 

2) Step 2: as shown below, the QIC and QICu is much smaller after deleting FamilyStructure:



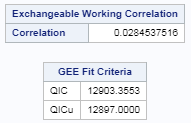
The most insignificant variable is this step is **schooltype**, with p-value = 0.1120.

3) Step 3: After removing schooltype, we found that the QIC and QICu are smaller than the previous model:



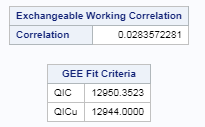
The most insignificant variable in this model is **meducation**, with p-value = 0.0702

4) Step 4: the QIC and QICu are much smaller after deleting meducation, compared with the previous model:



The most insignificant variable in this model is **ExerciseFreeTime**, with p-value = 0.093.

5) Step 5: the QIC and QICu are significantly greater than the previous model:



Therefore, the model is step 4 is the optimal one with regard to this data:

**proc** **genmod** data = HW3.DATHW2clean;

class school\_ID ChildRace HouseIncome pHealth ExerciseFreeTime

Region Urban PctMinority;

model bmipct = ChildRace HouseIncome pHealth ExerciseFreeTime TV

Region Urban PctMinority /type3 dist=normal;

repeated subject = school\_ID/type=exch corrw;

**run**;

The parameter estimates are:

