Name ________________________

Psychology 5068
Hierarchical Linear Models
Homework 5
Due March 5, 2018

In this homework you will use the student popularity data (popularity.csv) to explore methods for examining the adequacy of hierarchical linear models.

Begin with the unconditional model:

Level 1

$$popular_{ij} = \beta_{0j} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

1. Test the homogeneity of Level 1 residual variances (σ^2) assumption by comparing a model that estimates a single Level 1 variance (the default, call it *Pop_Fit_1*) to a model that estimates a separate variance for each classroom (call it *Pop_Fit_2*). Reminder: This needs to be done with the *nlme* package, which allows specifying separate variances for each Level 2 unit.
2. Add the Level 1 residuals (and fitted values) from *Pop_Fit_1* to the popularity data file (name them *L1_residuals* and *L1_fitted*, respectively) and produce a boxplot figure showing the residual distributions by classroom (class).
3. Examine the Level 1 residuals:
   (a) Construct a Q-Q plot of the Level 1 residuals.
   (b) Construct a histogram of the Level 1 residuals with a normal distribution overlay.
   (c) Are the Level 1 residuals normally distributed?
   (d) Construct a scatterplot of the Level 1 residuals against the Level 1 fitted values. Comment on the assumption of homoscedasticity and what that means for this unconditional model.
4. Now determine if Level 1 predictors should be added to the model.
   (a) Correlate the Level 1 residuals with extraversion. Is there evidence that this predictor should be included?
   (b) Create a scatterplot showing the relationship between the Level 1 residuals and extraversion. Does there appear to be any need to model nonlinearity?
   (c) Correlate the Level 1 residuals with student sex. Is there evidence that this predictor should be included?
   (d) Should both predictors be included in the model? That is, do they appear to be unique predictors?

5. Examine the Level 2 residuals:

(a) Create a classroom-level data frame (call it *Class_Data*) that contains the Level 2 residuals (only intercept residuals are available so far; name it *R_Intercept*) and the grand-mean centered classroom means for extraversion (name it *Mean_E_GMC*).
(b) Construct a Q-Q plot of the Level 2 residuals
(c) Construct a histogram of the Level 2 residuals with a normal distribution overlay.
(d) Are the Level 2 residuals normally distributed?
(e) Correlate the Level 2 residuals with classroom mean extraversion. Is there evidence that this predictor should be included in the Level 2 model?
(f) Create a scatterplot showing the relationship between the Level 2 residuals and classroom mean extraversion. Does there appear to be any need to model nonlinearity at Level 2?

6. Fit a new model based on what you have discovered so far:

Level 1

$$popular_{ij} = \beta_{0j} + \beta_{1j}extrav_{ij} + \beta_{2j}sex_{ij} + r_{ij}$$

Level 2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Mean\_E\_GMC_j + \gamma_{02}Mean\_E\_GMC\_SQ_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Mean\_E\_GMC_j + \gamma_{12}Mean\_E\_GMC\_SQ_j + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}Mean\_E\_GMC_j + \gamma_{22}Mean\_E\_GMC\_SQ_j + u_{2j}$$

   To fit this model, you will need to create a new variable, *Mean_E_GMC_SQ*, that is the square of *Mean_E_GMC.* Merge both variables into the original data frame, and fit the new model (*call it Pop_Fit_3*). Use the lme4 package so that you don't encounter convergence problems. Is there any evidence of curvilinearity at Level 2?

7. Fit a model that eliminates the squared terms in Level 2 (call it *Pop_Fit_4*) and compare it to the full model. Are you justified in eliminating the squared terms?
8. Using the simpler *Pop_Fit_4* model, determine if either or both of the slope variances at Level 2 can be set to 0.
   (a) Eliminate the random effect for extrav (call this model *Pop_Fit_5*). Is this model indistinguishable from *Pop_Fit_4*?
   (b) Eliminate the random effect for sex (call this model *Pop_Fit_6*). Is this model indistinguishable from *Pop_Fit_4*?
   (c) Finally, eliminate them both (call this Pop_Fit_7) and compare it to Pop_Fit_4. Which simpler model is justified?
9. Use the simplest model justified from the previous question.
   (a) Retest the Level 1 homogeneity assumption. Is there any improvement compared to what was found for Question 1?
   (b) Check the multivariate normality assumption for the residuals at Level 2 using Mahalanobis distance. Are the residuals multivariate normal?
10. Use Cook's distance to examine how influential the classrooms are in the model fit for Question 9.
   (a) How many classrooms stand out as distinctly more influential than the others?
   (b) If those classrooms are excluded from the analysis and the model is refit, do any conclusions change?