Midterm Exam Questions

1. (50%) Find a "multilevel relevant" publication for which a replication data set is available. It is not necessary that the authors use multilevel concepts; you might propose to introduce multilevel concepts. It is OK if this publication, or replication of it, might become a term paper project. If you are unable to find any publications with replication data, it is OK to choose a data set in RHS and go find the original publications about it.

   a) Provide the citation in the style that is most suitable to your field, whether that is APA6, APSR, etc. I'll need to get that data from you, somehow. Maybe we can upload it in Blackboard (?). Don't worry, I'll work that out.

   b) Provide a succinct summary of the research question (e.g., protestantism is related to wealth accumulation, or the effect of scissor usage on industrial accidents depends on the presence of left handed laborers).

   c) What multilevel relationships might exist? Write out the level 1 and level 2 models that you think might be correct (in the style we've practiced a little bit so far). If this is in the published record already, just type it in. If it is not, make something up. Don't forget to label your subscripts i and j (or whatever).

   d) Combine the level 2 model into the level 1 model to obtain the one that can be estimated.

   e) Write a paragraph about the theories that fuel the formalization in c and d. Pretend you are writing this paragraph for your dissertation advisor who, bless his soul, knows not very much about multilevel modeling, but understands ordinary regression pretty well.

   f) If you had the data right now, and it was recoded and ready to go. Lets suppose I think the slopes are common among groups, but the intercepts may differ. If you gave that data to me, what would you say if I propose to do these things. Write a sentence or two to tell me what benefits I might harvest if I did these things, or what dangers and disappointments might confront me?

      i. Pooled regression

      ii. Regression on group averages (the between model)

      iii. Group dummy variables (fixed effects) regression analysis

      iv. Group-mean centered (fixed effects) regression analysis

      v. Random effects (mixed effects) regression analysis

2. (50%) Lets join the Army. In the RHS book, problem 4.5 (p. 219) concerns well-being in the Army. This is one of the problems for which web solutions are available! I chose this one because I don't want this test to be about Stata recoding drama, but rather about explaining the differences between different kinds of regression models. I suggest you work through the exercise as RHS have it. Lets assume that if you ask me to, I will pull out the worked example from the teacher cheat sheet and show it to

you. However, I truly believe in my heart that this is one of the more practical, least full-of-tedium exercises we've had. It is not as simple as the one about wheat varieties, but not too bad either.

wbeing the DV, well being

hrs hours worked per week

cohes unit horizontal social cohesion

lead vertical cohesion

grp the group identifier, army company

a) Describe what you have here. How many groups? How many observations? The handy Stata function and xtsummary might help here. If you do this in R, use split followed by summaries within, and then summaries between. That's a little bit fun. I notice xtdescribe does not work because there's no time variable.

b) Make a trellis graph to explore the relationship of wbeing and hours worked per week. I've worked out several examples like that during the past month, so I expect it will not be too challenging. It appears to me you might reduce the number of groups to 12 or 15 in order to make a manageable plot. (In my opinion, there's no need to fuss too much about choosing groups randomly, you can keep it simple by selecting them in order from the sample. But if you want to do the RHS thing of randomly choosing some groups, they write out the Stata code in tedious detail.) If you do this in R, both the lattice package and ggplot2 have more-or-less fool proof trellis graph tools.

c) Write a paragraph about your trellis graph.

d) Generate some new variables. Create the group-means of the numeric predictors and create group-deviation scores for each of them as well. I showed a Stata for loop that did this about 3 weeks ago.

e) Fit the following regressions and make 1 table to summarize them.

   i. Pooled regression of wbeing on hrs

   ii. Pooled regression of wbeing on hrs, cohes, and lead

   iii. Between regression (regression on group means) of wbeing on hrs

   iv. Between regression of wbeing on hrs, cohes, and lead.

f) Suppose you stop analysis at this point. Write a paragraph about the relationship between hrs and wbeing.

g) Fit the following regressions and make 1 table to summarize them.

   i. Dummy variable model: wbeing on hrs, cohes, lead and dummy variables for group membership. In your table, it is NOT necessary to include all of the estimated dummy variables. It would be nice to include the standard deviation among those estimates, but, honestly, I've not overcome the Great Wall of Stata for doing that. If you find a convenient way, show me! (In R, I can do that in 8 seconds). Lets say this that bit is Extra Credit 5%.

ii. Run the same thing, but omit the constant (if your software will allow it; if it does not, then skip this part). In your table, it is NOT necessary to include all of the estimated dummy variables.

iii. Fixed effects OLS regression estimated by replacing all variables with the group-deviation scores you calculated in part d). If you use software that has a built-in "fixed effects" estimator, you might run that. If the estimates differ from the fixed effects OLS you run, you should probably report that and be ready to explain.

iv. Random (mixed) effect of group membership.

h) Write a paragraph to address this question: If I run the fixed effects OLS estimator, what do I sacrifice? What shortcomings does it have in this particular example?

i) Fit the following random effects (mixed effect) regressions and make 1 table to summarize them. In question g) part iv, I did not specify how you ought to code the predictors. So I need to be explicit here.

i. Predictors cohes, lead and hrs are entered without recoding, centering.

ii. Predictors cohes, lead and hrs are included, but also are the group-means of cohes, lead, and hrs.

iii. Predictors cohes, lead, and hrs are replaced by group mean-deviation scores for cohes, lead, and hrs, but *also please* remember to include the group-means of cohes, lead, and hrs.

j) How would you describe/compare/contrast the estimates for the 3 models in the previous table?

k) Last question. The comparison between question (i) part i. and question (i) part iii is often thought to be an important specification test in multi-level regression. Please write a paragraph to briefly describe the idea we are testing and, if you are able to with the software you have, conduct the estimate and discuss it. (About software, we've found 2 ways in Stata to do this, but if you used R, I suspect we'd have to construct a Wald test (the fancy t-test) to get this done. If you took POLS706 with me, you may recall that.)

3. 0%. Please start thinking about what kind of term paper project you might do. If you know what it might be, hurray. If you don't, easier avenues might be 1) replication and embellishment of publication discussed in question 1, or 2) simulation of estimator performance under various conditions (as in my mlm_sim-6 or such).