**Notes on exams for TA session**

* I graded all the exams, everyone did really well, I think. I tried to create a standard answer for every question and grade consistently, and I left kind of standard comments explaining any points taken off. So, let me know if you have any questions about that. For what it’s worth, I thought this exam was quite hard and this is toughest assignment in the class, but now we don’t have anymore exams. Everyone is doing great on the homeworks which count for a ton of your grade, and then the final project is pretty open-ended and I’m sure that will be fine.
* It is worth remembering that the reason assumptions are made is so that statistical tests may be carried out and that the estimators have certain desirable properties (like unbiasedness and consistency). Many "assumptions" are better considered as "conditions" that are needed in order to make certain inferences.
  + Models don’t make assumptions. Models provide a way to make inference about the data generating process that produced our data and the parameters that govern this process.
  + These inferences are only valid (unbiased, consistent, etc.) under certain assumptions, depending on the modeling framework. For example, if we interpret the beta coefficient in an OLS model as the average treatment effect, we have to make tons of assumptions such that the conditions are met for this to be an unbiased estimate.
  + Do mixed effects models make more or less assumptions than OLS?
    - Less. We are parameterizing more things like the error covariance structure, but we are sort of implicitly doing that with an OLS model too. In an OLS model our beta is unbiased assuming the error term is iid, whereas we relax that assumption in mixed effects models because we are able to account for situations where the error is in fact not iid. Of course we have to get the covariance pattern correct, but it is more flexible than OLS.
* Review APC: often in panel data, age and time are perfectly colinear because of the structure of the survey.
  + For example, if we begin with 1000 infants at birth (age 0) and follow them every year, then “year of interview” and “age” will contain identical information. One example of a survey like this is the Fragile Families cohort study. This of course presents the APC problem, where it is impossible to infer whether variation at age/time=3 is truly a period effect or an age effect.
  + But often in surveys, time of interview and age of respondent capture two different things: time is when the interviews happen, age is a time-varying predictor. Think about period- or wave-effects and correlations here. There are lots of examples of this type of panel data: the HRS, the PSID, etc.

NAME

* II.C. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1). Interpretation for the intercept should be person’s baseline CD4PCT (-1).
* II.H. The matrix would be 7x7 because there are seven observations for each child (-2).
* II.J. Correct definition but didn’t include explanation of why you’d prefer one over the other. If we are only interested in the fixed effects results, we can choose either model. If we are interested in both the fixed- and random-effects parameters, we use the multilevel model. If we are uncertain about the misspecification of the variance-covariance of the longitudinal data, we use the GEE method. (-1).

NAME

* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.J. Correct definitions and interpretation of differences. The model specification is just an extension of GLM where you have a flexible link function, and the variance function is specified as a working correlation matrix for each subject. See attached (-2).

NAME

* II.C. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then differentiate fixed/random effects and specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.F. Here you want to interact the predictors (age at visit and treatment) with the “time” variable, visit – so these would both be included on Level 2. See attached (-3).

NAME

* II.C. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then differentiate fixed/random effects and specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.F. Your equations are correctly structured, but the “TIME” term in Level 1 should be Visit and (visage-baseage) and treatment should be interacted with this (so in Level 2). This is a tricky question because “time” is not synonymous with age at a given time. We’ll talk about this in the TA session. See attached equations (-2).
* II.H. You’re correct in the description of the Toeplitz structure, but the variance-covariance is for every pair of “time” steps. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-3).
* II.J. Correct definition of GEE, but your interpretation is a bit off. We don’t know how similar these would be because it would depend how poorly we choose the variance-covariance matrix of the random effects in the MLM. GEE is also more flexible in handling the errors, not less. If you only care about the fixed effects and do not care about inferring anything about the structure of the random effects, you could just use GEE (-2).

NAME

* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.H. The Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). In this case, we would have a 7x7 matrix because each individual is observed 7 times (-4).
* II.I. There are two issues here: balance and missingness. Having an unbalanced dataset is not an issue for MLM. However, missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data. Including observed predictors does not address this issue (-2).
* II.J. Your rationale for using the GEE model is correct because you care only about the fixed effects, but your description of the GEE specification is not correct. See attached (-3).

NAME

* I.3. These are the correct specifications, but you could describe the differences and when different specification will provide identical answers (DiD vs. FEs, FEs vs. REs) (-1).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.C. Incomplete (-7)
* II.F. You should include the interactions of visit and base age, see attached (-2).
* II.H. The Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). You do not include the dimensions of this matrix. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-4).
* II.I. Describe how we are handling missing data in these models (removing it from analysis) and describe how this might bias our estimates (-1).
* II.J. Incomplete (-7)

NAME

* I.2. OLS only produces unbiased estimates under various assumptions, including that the errors are uncorrelated with regressors. MLM does not violate this assumption but modifies it. In MLM we assume ahead of time that the errors are correlated, and we decide on how to model their correlation (the variance-covariance matrix of the random effects) (-1).
* I.3. These are the correct definitions, but you could describe the differences and when each specification will provide identical answers (DiD vs. FEs, FEs vs. REs) (-1).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1). To allow the intercept and slope on time (visit) to vary by predictors “age\_recentered” and “treatment” you need to interact both these variables with visit (-2). You also need to interpret the intercept (average CD4PCT at baseline) (-1). See attached.
* II.H. Incomplete (-7).
* II.J. Your rationale for using the GEE model is correct because you care only about the fixed effects, but you’re missing the GEE specification. See attached (-3).

NAME

* I.2. OLS only produces unbiased estimates under various assumptions, including that the errors are uncorrelated with regressors. MLM does not violate this assumption but modifies it. In MLM we assume ahead of time that the errors are correlated, and we decide on how to model their correlation (the variance-covariance matrix of the random effects) (-1).
* I.3. These are the correct definitions, but you could describe the differences and when each specification will provide identical answers (DiD vs. FEs, FEs vs. REs) (-1).
* II.C. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then differentiate fixed/random effects and specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.H. You’re correct in the description of the Toeplitz structure, but the variance-covariance is for every pair of “time” steps. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-3).
* II.J. This is correct regarding estimation but need more interpretation. We don’t know how similar these would be because it would depend how poorly we choose the variance-covariance matrix of the random effects in the MLM. If you only care about the fixed effects and do not care about inferring anything about the structure of the random effects, you could just use GEE. Also missing the GEE specification, see attached (-4).

NAME

* II.B. Here you would do a two-step estimation, first fitting an OLS model for every individual predicting their CD4PCT using VISIT. In the second step, you would predict the points estimates from this model (all child-specific intercepts, all child-specific slopes) using treatment and age at baseline. This is basically Level 1 and Level 2 from the MLM framework but doing them with two separate modelling steps rather than one (-2).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1). You should include the interactions of visit and base age to allow for “varying intercepts and slopes with child-level predictors for treatment and age at baseline” – see attached (-2).
* II.H. The Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). You do not include the dimensions of this matrix. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-4).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).
* II.J. This is the general link function in the GEE, but you needed to describe the whole specification and why you’d prefer one over the other. If we are only interested in the fixed effects results, we can choose either model. If we are interested in both the fixed- and random-effects parameters, we use the multilevel model. If we are uncertain about the misspecification of the variance-covariance of the longitudinal data, we use the GEE method. See attached (-5).

NAME

* II.C. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then differentiate fixed/random effects and specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).
* II.J. Correct definitions and interpretation of differences. The model specification is just an extension of GLM where you have a flexible link function as you note, but you needed to also note that the variance function is specified as a working correlation matrix for each subject (-2).

NAME

* II.C. This is correct about the problem with two-stage OLS but could describe how MLM handles this differently. In the multilevel framework, we use a single estimation step (one MLE function) rather than multiple estimation steps. We can then differentiate fixed/random effects and specify explicitly our assumptions about the correlated error terms expressed in the variance-covariance matrix of the joint distribution of the random terms (-1).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data. The issue of non-random missing data is not addressed by covariance pattern models (-3).
* II.J. MLM is more informative and preferred if you care about making inference on both fixed and random effects. If you only care about making inference on the fixed effects then GEE is more efficient, and you don’t have to worry about specifying the variance-covariance matrix incorrectly (as in MLM) (-2).

NAME

* II.B. The model setup is correct, but the Step 1 model should be visit instead of age at baseline. And then the Step 2 models should include both treatment and age at baseline to predict the points estimates from Step 1 (-2).
* II.D. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-2).
* II.F. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-2).
* II.J. In your last statement, it would make sense to use GEE if you only care about making inference on your fixed effects. The MLM approach requires correctly specifying the variance-covariance, while the GEE does not (-1).

NAME

* II.F. You should include the interactions of visit and base age to allow for “varying intercepts and slopes with child-level predictors for treatment and age at baseline,” and also include an interpretation of gamma\_00 (-2).
* II.H. The Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). In this case, we would have a 7x7 matrix because each individual is observed a maximum of 7 times (-3).

NAME

* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).

NAME

* II.B. This question was getting at the idea of specifying a “multilevel model” using two-stage OLS: first estimating each child’s intercept and slope, and then predicting these points estimates using treatment and age at baseline. See attached (-4).
* II.C. Unlike the two-step approach, we estimate a multilevel model using a single step with one maximum likelihood function. We differentiate between fixed and random effects. We further specify assumptions about the correlation between u0i and u1i expressed in the variance-covariance matrix of the joint distribution of the two random terms (-3).
* II.D. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1). This is the correct model form, though “time” in this case is visit rather than age (-1).
* II.F. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-1). You should include the interactions of visit and base age to allow for “varying intercepts and slopes with child-level predictors for treatment and age at baseline” – see attached (-1).
* II.H. Your description of the structure is correct, but the Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). You do not include the dimensions of this matrix. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-2).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).
* II.J. Almost correct definition of GEE, but your interpretation is a bit off. It does consider correlated errors: the variance function is specified as the working correlation matrix for each subject, Ri. But this makes it difficult to infer anything about the general structure of the random effects compared to MLM. Overall, if we are only interested in the fixed effects, we can choose either model. If we are interested in both the fixed and random effects, we use the MLM. If we are uncertain about the misspecification of the variance-covariance of the longitudinal data, we use the GEE method (-2).

NAME

* II.C. Unlike the two-step approach, we estimate a multilevel model using a single step with one maximum likelihood function. We further specify assumptions about the correlation between u0i and u1i expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.D. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-2).
* II.F. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-2).
* II.H. The Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). In this case, we would have a 7x7 matrix because each individual is observed at most 7 times (-2).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).
* II.J. This is the general link function in the GEE, but you needed to describe the whole specification and why you’d prefer one over the other. If we are only interested in the fixed effects results, we can choose either model. If we are interested in both the fixed- and random-effects parameters, we use the multilevel model. If we are uncertain about the misspecification of the variance-covariance of the longitudinal data, we use the GEE method. See attached (-5).

NAME

* I.3. Need definition and description of FE and RE models and when they are identical (-2).
* II.C. Unlike the two-step approach, we estimate a multilevel model using a single step with one maximum likelihood function. We further specify assumptions about the correlation between u0i and u1i expressed in the variance-covariance matrix of the joint distribution of the two random terms (-1).
* II.D. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time (-1).
* II.F. The “time” variable should be visit, so your model here is correct if you just replace the recentered age with recentered time. When you have multiple random effects, you need to specify the variance-covariance of their joint distribution (-2).
* II.H. Your description of the structure is correct, but the Toeplitz matrix describes the variance-covariance for every pair of “time” steps and assumes these do not vary by individual (e.g. the covariance between time 2 and time 3 is equal across all children, but differs from the covariance of time 3 and time 4, etc.). You do not include the dimensions of this matrix. In this case, we would have a 7x7 matrix because each individual is observed 7 times (-2).
* II.I. You needed to describe what happens to missing data in these models (it is simply dropped from analysis). Missing data can result in biased estimates of parameters if the data are not missing at random. This can be a big problem if you have a lot of missing values and they are not missing randomly, but rather represent individuals that are systematically different than the ones for which you do observe complete data (-2).
* II.J. Incomplete (-7).