An Investigation of the Statistical Modelling Approaches for MelC

Literature review and recommendations

By Jessica Thomas, 30 May 2011

Contents

1. Overview

In our quest to build a robust micro-simulation model we investigated various statistical models and methods. This document is a review of that investigation.

It is important in the simulation that an individual, whose life course will be simulated, will have a coherent trajectory or story. For variables that can change but that do not change often we needed to think about how this would be accomplished in the statistical modelling and simulation. The initial idea was to include, as a predictor/explanatory variable, the outcome variable's value at the previous year. We refer to this variable as the lagged dependent variable or LDV.

Following this came other ideas that do not require an LDV, such as, for dichotomous or categorical variables, we could build separate models for each prior state (providing that there are enough observations in each category). Those models are generally straight-forward but we may still have to consider models where we do include an LDV for continuous or count variables for which buildings models for each previous state will not be a practical solution.

The models considered in this document are pooled regression models, fixed effects models, random effects models and hybrid models.

Section 2 specifically considers the LDV and whether it should be included in any of the above types of models. The conclusion is that including an LDV in any of these models introduces the potential for bias; and it is very difficult to know how large the bias might be. To include an LDV in a model and get around the problem of bias one would need to consider an instrumental variable approach (or some other econometrics approach). Our best approach seems to be to:

- Exclude the LDV if it is not necessary
- If the previous value of the outcome does need to be taken into consideration, use separate models for each previous state value if possible
- If this is not possible, include the LDV (in a pooled, random effects or hybrid model) and accept the possibility of bias but check that our model is acceptable and performing well by comparing the predictions from the model to the real data.

Section 3 talks about how to choose between pooled regression models, random (mixed) effects models and fixed effects models. This section is focused around a series of four papers (different authors with different views) but presents a simple flowchart in Section 3.1.2 which can be used to easily choose between the three models.

Section 4 talks about the difference between ‘within’ and ‘between’ effects, which are an integral part of a second flow diagram in Section 7 which attempts to provide an overall procedure to choose between various modelling approaches. It is by no means fool-proof and won’t be able to be used in every situation. It is hoped, however, that it will be helpful in a number of situations.

In Section 5 I consider a major factor in our choice of modelling approach: the implementation of the model in the micro-simulation.

Section 6 is a bit of a mish-mash and simply lists a few other factors and facts that I came across in my reading that may be relevant to us.

Section 7 was mentioned earlier and is the culmination of the investigation into a single framework.

2. The LDV

By definition there is correlation between the LDV and the current disturbance/error. And hence, if the LDV is included as an explanatory variable in a classical linear regression model, one of the assumptions will be violated and estimators will be inconsistent (i.e. biased).

Is this a problem if we only had one row of data per child? (Barry’s idea). I.e. if it were effectively a cross-sectional analysis e.g. $y_5 = y_4 + x$, then could think of y_4 as any other predictor? But maybe it is still not exogenous (which we require it to be to meet even assumptions of this model).

Although we did find examples in the literature of people simply including an LDV in their statistical model, further reading led us to find that not everyone would agree with what they did. The literature around including LDV in statistical models was not straight-forward: one author would say one thing and another author would say another.

Green, Kim and Yoon write an article where they explain in which situations one should use a pooled model, a random (mixed) effects model or a fixed effects model (Green, Kim, & Yoon, 2001). In their fixed effects models they include an LDV but state in a footnote that they are aware that the LDV is an endogenous regressor and so they tried the analysis also using the Anderson-Hsiao methodology (instrumental variables) but did not present them because results were similar.

Oneal and Russett (Oneal & Russett, 2001) follow-up on Green, Kim and Yoon (whose article re-analysed a previous analysis Oneal and Russet had published (Oneal & Russet, 1997)) and present an analysis where they do include LDV in their final models without mentioning the issue of endogenetiy.

So these articles by Green, Kim and Yoon, Oneal and Russett and others that follow on in response to these include LDVs in their models but not explicitly talk about the issues it may cause.

Keele and Kelly do write an article that is expressly about the LDV (Keele & Kelly, 2005); however, it is about using an LDV in an OLS regression. Their conclusion is that one should first ask a theoretical question: Does the past matter for the current values of the process being studied? If the answer is yes, OLS with an LDV is appropriate so long as the stationarity condition holds and the model residuals are not highly autocorrelated. They do not mention the case when OLS is not used however (and we have not been using OLS yet as we have no continuous outcomes).

Beck and Katz state in a working document (Beck & J.N. Katz, 2004) that by switching estimation techniques from OLS to maximum likelihood, non-linear least squares or Cochrane-Orcutt we get around the problems of inconsistency when using OLS.

This may sound like the answer of our dreams (and maybe it is) as for all our count and dichotomous outcomes OLS is not used, but other authors (Goodrich (Goodrich, 2005) and Zorn (Zorn, 2001)) would say that what we have to think about is the within and between effects and whether these are the same for each variable or not.

Goodrich provides a very detailed and helpful working document that explains how to go about modelling TSCS data with correct estimates and standard errors for between and within effects. However the problem of the LDV causing bias still remains and he doesn't mention that it can be fixed simply by switching estimation techniques and not using least squares. In fact he says "I focus on least squares estimators, but all the conclusions hold if maximum likelihood, Markov Chain Monte Carlo, or the method of moments is used to estimate the linear model". Goodrich's approach, however is detailed in this document as overall it may help us build a better model.

2.1 LDV Specifically in a Random Effects Model

There were multiple books, articles, online power points and blogs that clearly stated that an LDV should not be used in a random effects model. These include:

- Goodrich (Goodrich, 2004): If an LDV is included in the REE, there will be correlation between the random effects and the LDV
- Microeconometrics: Methods and applications by Cameron (Cameron, 2005): "The estimators from the previous chapter are all inconsistent if the regressors include LDVs, even in the case of the random effects model" (p.764).
- An online blog[1]: If you have random effects in panel data then the lagged dependent variable will be correlated with the random effect in your error term. Random effects are just as troubling as fixed effects in the lagged dependent variable case. Better to just add more lags (*of the X variables*).
- Online power point presentation[2]

Ashley (Ashley, 2010) states, however, that "It is widely believed that the inclusion of lagged dependent variables in a panel data model necessarily renders the Random Effects estimators, based on OLS applied to the quasi-differenced variables, inconsistent. It is shown here [they give a proof] that this belief is incorrect under the usual assumption made in this context, that the other regressors are strictly exogenous" (abstract). I think that the quasi-differencing he is talking about may be the differencing/weighting (perhaps between the overall mean and a specific unit) that happens automatically when you run a random effects model. This may sound promising to us but

[1] http://www.mostlyharmlesseconometrics.com/2011/03/lagged-dependent-variables-with-random-effects/

[2] www.econ.pdx.edu/faculty/KPL/ec510/PD8.ppt

he also states that "If, however, there is feedback between, say, $X_{j,t}$ and$Y_{i,t}$, then $X_{j,t}$ is only weakly exogenous. The quasi-differenced lagged values of $X_{j,t}$ will in that case be correlated with the quasi-differenced current error term, leading to… inconsistency in the RE estimator. This inconsistency could be significant if the feedback is strong and the panels are short, in which case one would be better off using the Arellano and Bond or Keane and Runkle estimators". We have feedback by our theory as we run loops in our simulation so one Y eventually ends up affecting the X that was a regressor variable in its regression. Also, we have short panels (only 5 observations per child). This, however may change when we get more CHDS data.

2.2 The Instrumental Variable Approach

An instrumental variable (IV) approach would theoretically be able to correct the bias from the endogenous LDV. It does this at the expense of losing some information about the relationship with the outcome contained in these variables. The idea behind instrumental variables is that there is a variable that is correlated with the endogenous exposure variable (X, our LDV), but has no direct effect of its own on the outcome. The IV analysis works by running a regression of the endogenous variable (X) on the instrumental variable(s) and substituting the predicted value of X from this regression into another regression on the outcome variable, Y, that includes other covariates of importance. The coefficient from this predicted X variable (which is in the model in place of the endogenous variable itself) in the second equation can be interpreted as the 'IV estimate of the effect of X on Y'.

Although one may get an unbiased estimate of X on Y using this approach, it is not ideal for the MelC project for a number of reasons:

- The IV approach would use the predicted values of the LDV, we may not be able to implement this approach in the micro-simulation or, if we were able, we may not get a very predictive model that will replicate the real data well.
- it may be difficult to find a good IV. The third lag has been suggested by others but this may still be correlated with the outcome in our case. Also, with only 5 time points using the third lag as an IV only allows us to make predictions for the 4th and 5th years.
- A separate IV is needed for each X variable. *Is this only X variables that we know are endogenous (i.e the LDV) or would we need an IV for each X in the model? (This second point is what I was thinking during Dalton's workshop but now am back to thinking maybe the first statement is correct)*

2.3 Other Approaches for Dealing with the LDV Problem

Dynamic or transition (Markov) models are useful when the past values of the outcome variable are influenced by the current values (state dependence) or when introducing time lags.
Marginal structural models are relatively new, building upon random effects and structural equation models.

Zucchini, W. and I.L. MacDonald, *Hidden Markov models for time series: an introduction using R*. Monographs on Statistics and Applied Probabilitt 110. 2009, Boca Raton: Chapman and Hall/CRC.
The above book can be accessed as a pdf e-book through the library – I have downloaded Chapter 1.

2.4 Panel Length

Another thing I read (e.g. in (Beck & J.N. Katz, 2004)) was that the length of the panel, i.e. the number of observations per unit, affects the degree of bias with the bias decreasing for longer panels. This means with our short panel that bias has the potential to be more an issue. If we get data for a longer time period we can at least be comforted by the fact that we are reducing any bias present due to endogeneity.

2.5 Conclusion on the LDV Problem

So the overall conclusion for including an LDV in any of the models we have been considering (pooled, random/mixed effects or hybrid) is that the coefficient estimates will be inconsistent. This means they may be biased. We do not know however, how biased they may be. The fixes to the LDV problem are econometric techniques such as instrumental variables or other fancy estimators.

Our best options would be to exclude the LDV when it is not strictly necessary and to estimate multiple models for each prior state (each value of the LDV) where we can. This will be possible for dichotomous variables and for some categorical variables depending on how many children are in each category and whether we can combine groups in a sensible way that provides enough observations to do multiple models. For other variables, if we do not want to group them, we will either have to look into complicated econometric techniques or something like hidden Markov models (need to think about time constraints and if we have any expert around to help us) or accept that there may be bias in our estimates.

If we can run our simulation and compare the simulated dataset to the real dataset and they are quite close, we can decide not to be worried about potential bias in the estimates – we are doing as good as we are going to do by matching the real data well.

3. Choosing Between Pooled Regression, Random Effects and Fixed Effects Models

3.1 A Series of Four Papers

A **series of articles** debating the use of fixed effects methods was published in International Organisation volume 55 issue 2. **Oneal and Russet** (Oneal & Russet, 1997) had published an analysis in which they simply pooled over years and dyads (pairs of countries- we can think of dyads as children for our purposes). This inspired **Green, Kim and Yoon** to write an article where they explained in which situations one should use a pooled model, a random (mixed) effects model or a fixed effects model (Green et al., 2001). The details of their paper are given in the next section along with a procedure (and flow-chart) for choosing between pooled, random and fixed effects based that was inspired from their paper.

3.1.1 An Explanation of When and When Not to Pool by Green, Kim and Yoon.

The following is lifted from Green, Kim and Yoon (the word dyad or dyads has been replaced with child or children except in the plots):

> Data is said to be pooled (in an analysis) if not distinction is made between observations in time and space (*space being individuals in our case*). This means that strong assumptions are imposed when one performs a pooled regression on panel data.
>
> The pooled cross-sectional model takes the form
>
> $$Y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_K X_{Kit} + u_{it} \quad (1)$$
>
> In this expression, the outcome Y_{it} is a function of K right-hand side variables that vary across both time and space. The subscript i refers to one of the N cross-sectional units (children in our case), and the subscript t refers to one of the T time points. The hallmark of this model is the inclusion of a *single intercept* (α) that reflects the expected value of the dependent variable when all of the independent variables are zero. In effect, this model makes the claim: “It doesn’t matter which child one picks; the intercepts are all the same.”
>
> The pooled cross-sectional model in Equation (1) differs from a fixed-effects panel model in which each child is assigned its own intercept. The regression model now includes $N - 1$ dummy variables, for each child (less one) in the data set:

$$Y_{it} = \alpha + \delta_1 Z_{1it} + \delta_2 Z_{2it} + \cdots + \delta_{N-1} Z_{N-1,it} + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_K X_{Kit} + u_{it} \qquad (2)$$

Here, the Z_{git} represent dummy variables marking each child, and the coefficients associated with each child are denoted δ_g. Thus the intercept for the first child is simple α + δ_1. Equation (1) is a subset of Equation (2), where all of the δ_g are constrained to be zero. Pooled cross-sectional regression, in other words, is a special form of a more general regression model.

Since pooled cross-sectional models omit variable that are included in the fixed effects panel model, it should come as no surprise that pooled cross-sectional model may generate biased estimates of the β_k. When the δ_g are not zero (that is, when the children really do have different intercepts) and when the Z_{git} are correlated with the X_{kit}, regression estimates will be biased. Note that these biases may be positive or negative, depending on how the intercepts covary with the regressors. In that sense, the problem of ignoring fixed effects is a special case of a more general problem, that of omitting variables in multivariate regression.

Scatter plots of hypothetical data illustrate how pooling may introduce bias. The plot in Figure 1 depicts a positive relationship between *X* and *Y* where *N* = 2 and *T* = 50. Because both children share a common intercept, pooling creates no estimation problems. One obtains similar regression estimates regardless of whether one controls for fixed effects by introducing a dummy variable for each child. A pooled regression is preferable in this instance because it saves a degree of freedom.

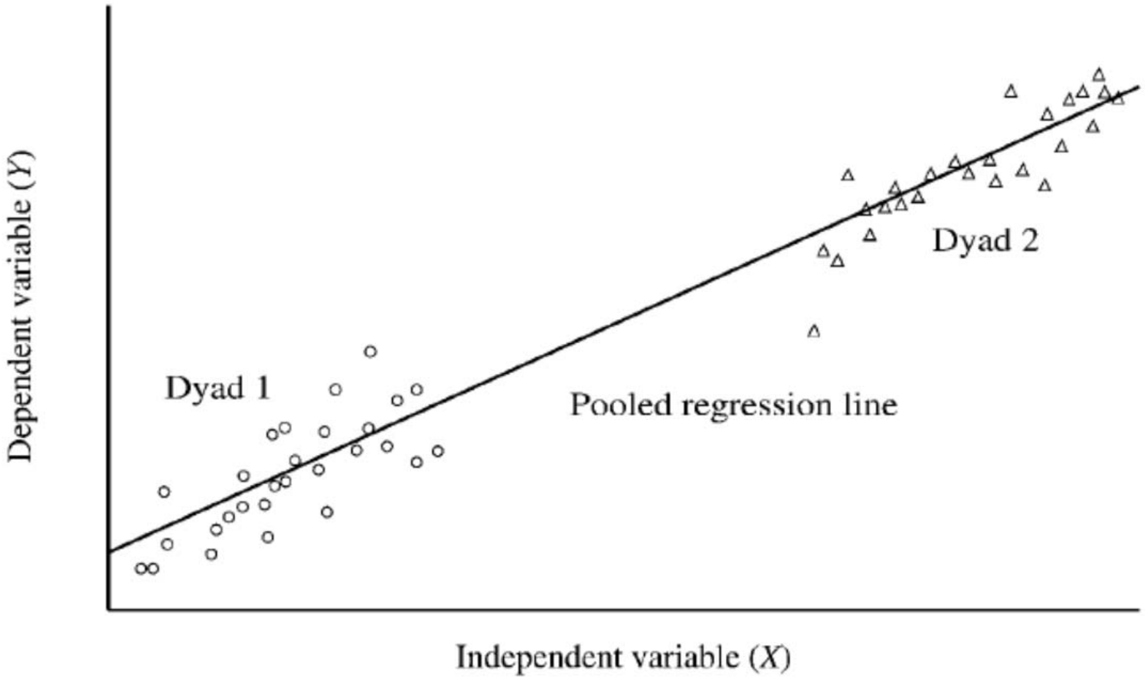


FIGURE 1. *Pooling homogenous observations*

Figure 1: An example situation of where pooled regression can be used. Dyads can be thought of as children.

In Figure 2 we encounter another instance in which pooling is benign. The two children have different intercepts, but there is no correlation between the intercepts and *X*. The average value of the independent variable is the same in each dyad. Again, pooled regression and regression with fixed effects give estimates with the same expected value.

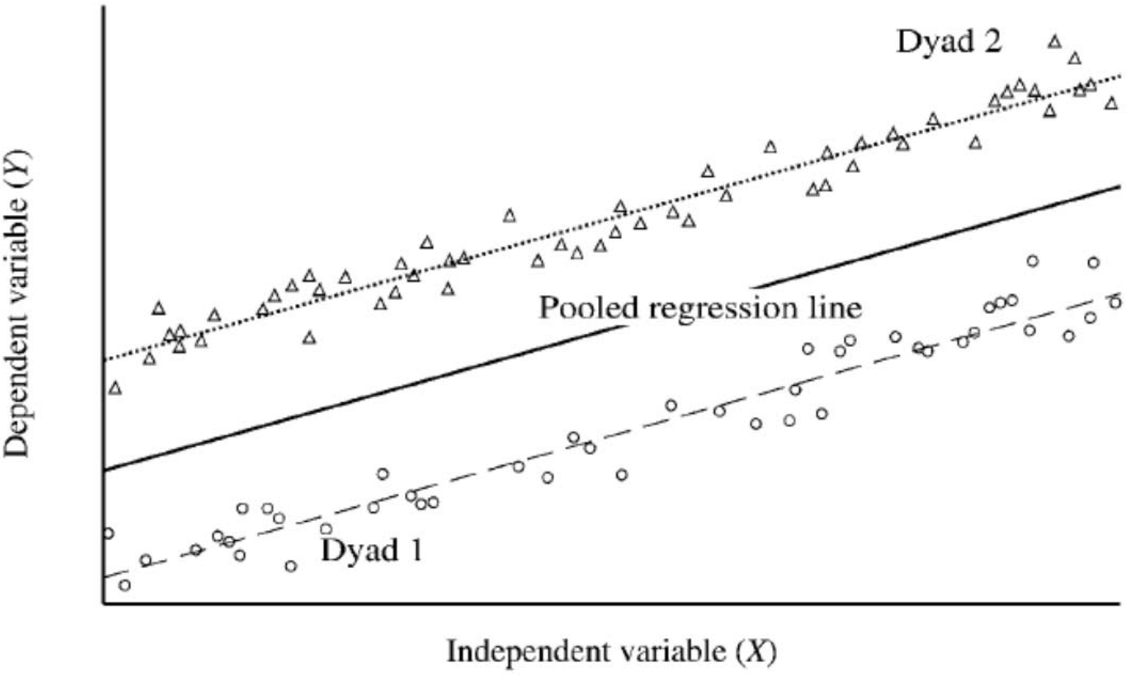


FIGURE 2. *Pooling dyads with randomly varying intercepts*

Figure 2: An example situation of where random effects should be used. Dyads can be thought of as children.

Figure 3 illustrates a situation in which pooled regression goes awry. Here, the causal relationship between *X* and *Y* is negative; in each dyad higher values of *X* coincide with lower values of *Y*. Yet when the dyads are pooled together, we obtain a spurious positive relationship. Because the dyad with higher values of *X* also has a higher intercept, ignoring fixed effects biases the regression line in the positive direction. Controlling for fixed effects, we correctly ascertain the true negative relationship between *X* and *Y*.

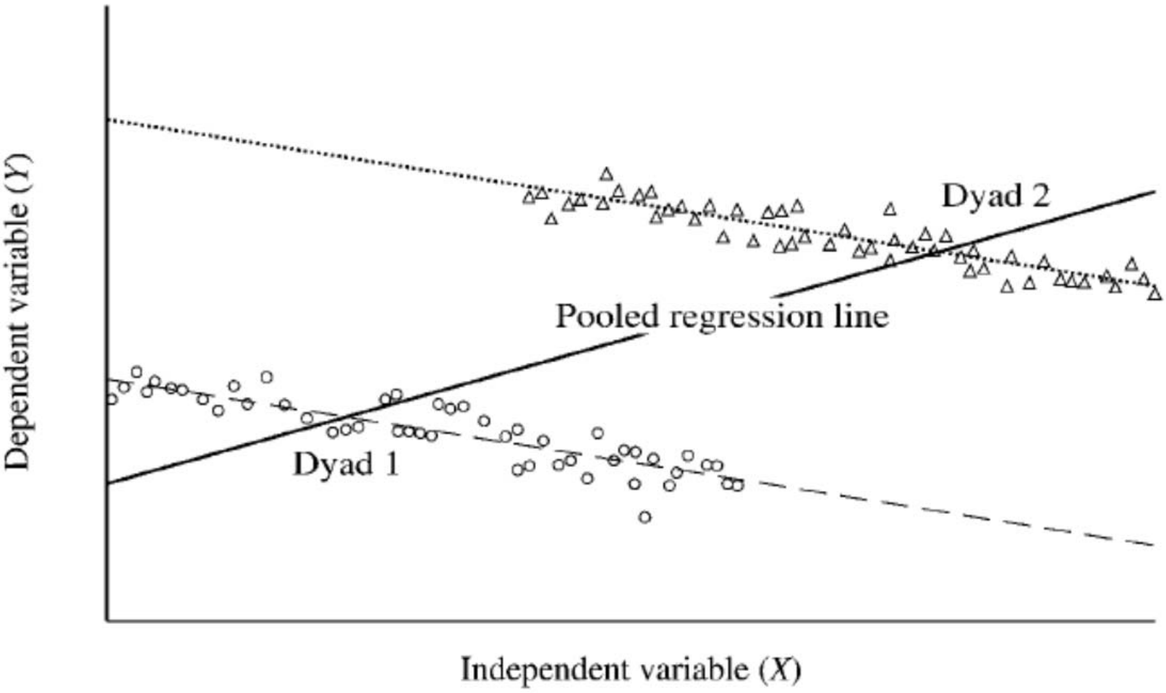


FIGURE 3. *Pooling observations ignoring fixed effects*

Figure 3: An example situation of where fixed effects estimates should be used. Dyads can be thought of as children

3.1.2 A Procedure to Choose Between Pooled Regression, Random Effects and Fixed Effects

Green, Kim and Yoon do not explicitly mention random effects models (strictly we should refer to these as mixed models as a random effects model is a model with only random effects we will use the term random effects to mean mixed effects to keep consistent with the econometric literature) in the above excerpt but one can see that their Figure 2 is an example of a situation where one would choose to use a random effects model; it meets the assumptions of a random effects model where the child intercepts are not correlated with the regressors. A random effects model would take the form

$$Y_{it} = \alpha + \delta_i + \beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_K X_{Kit} + u_{it} \qquad (3)$$

where the δ_i represent the child intercepts (or effects). These are assumed to be iid Normal(0, σ^2). This is nested within the fixed-effects model and the pooled model is nested within this random effects model.

So we can form a procedure for choosing between these models. This procedure is provide in a flow-chart form in Figure 4. First we need to ask if the children have different intercepts or whether they could be considered to have a common intercept. This is the question the "Child Effect?" box asks. We can test this with a statistical test for whether the variation of the child intercepts is significantly different from zero. We can use the Covtest statement in Proc Glimmix in SAS to do this.

If the child intercepts have very little variation (that is, they are very close) then we can use a pooled model (and include time as an independent variable). If the child intercepts do have significant variation then we need to ask whether these intercepts are correlated with the regressors. A statistical test for this question is the Hausman test which can also be performed in SAS[3]. If the child effects are not correlated with the regressors then a random effects model will give us unbiased estimates but if the child effects are correlated with the regressors then we should consider a model with fixed effects estimates. To get fixed effects estimates there are other models to choose in addition to the fixed effects model in Equation (2). These include hybrid models such as those described in Section 4.

[3] Allison (Allison, 2005) gives details on how to perform the Hausman test for a continuous outcome model on pages 29 – 30. He uses the TSCSREG procedure.

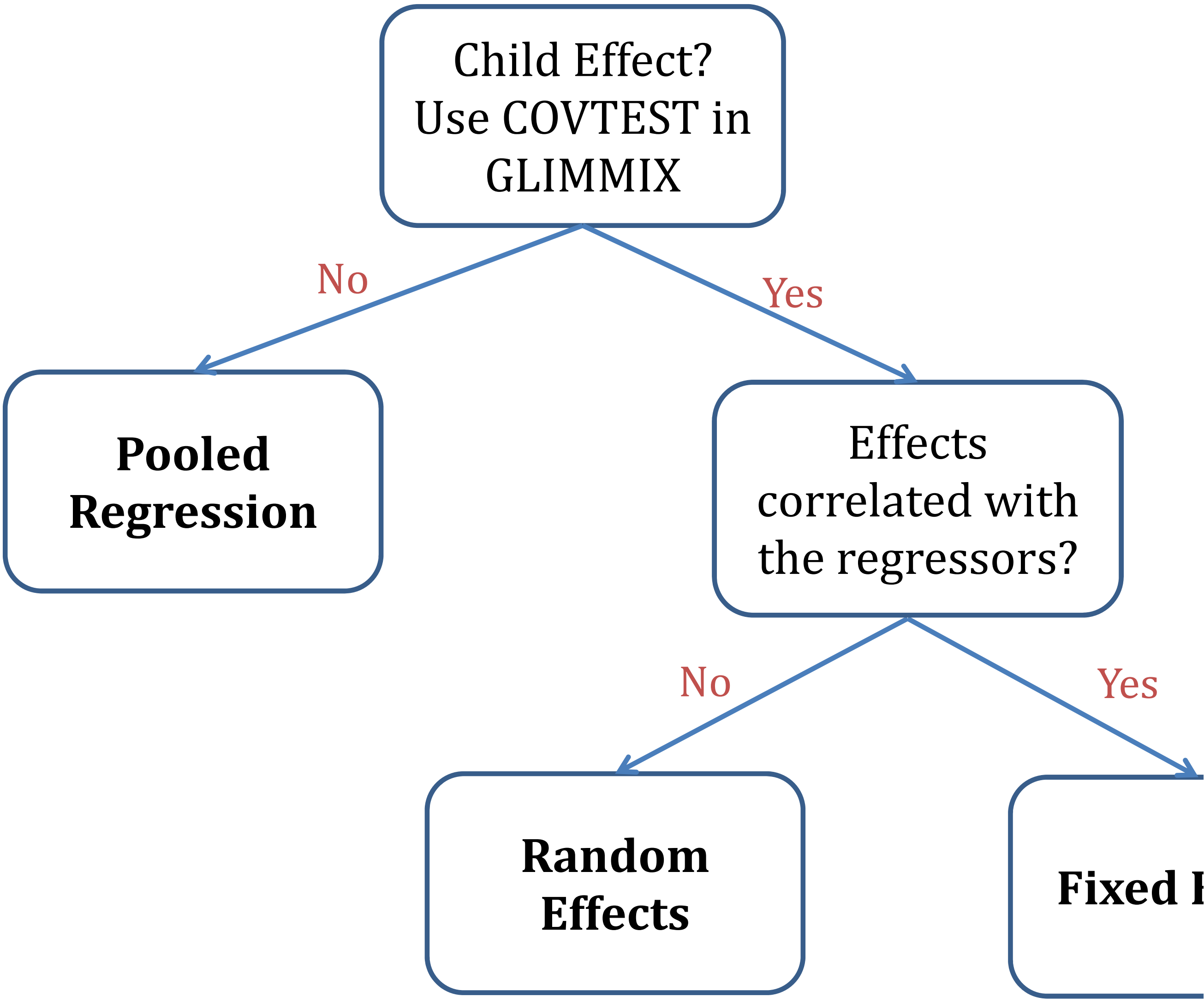

Child Effect? Use COVTEST in GLIMMIX
No
Yes
Pooled Regression
Effects correlated with the regressors?
No
Yes
Random Effects
Fixed E

Figure 4: Flow diagram of how to choose between pooled regression, random effects and fixed effects models. Note that this diagram does not take into account issues with the LDV.

3.1.3 The Two Papers Following Green, Kim and Yoon

Oneal and Russett strongly disagree with Green, Kim and Yoon in a response article (Oneal & Russett, 2001). They ask "what kind of theory would purport to explain variation through time but not accrued differences across groups?" (p.481) They do not like fixed effects because they not do allow cross-sectional information to inform their understanding. They state "the best summary of the consequences of our theoretical variables is the average of the within- and across-groups effects" (p.482). They state that fixed effects relies on variation over time in the variables and when there is limited variation over time (and this is more likely to be the case with binary data) that fixed effects are not ideal.

They also talk of the balance between type 1 and type 2 errors. They state that fixed effects place too much emphasis on avoiding type 1 error. For us, working on MelC, we also need to consider the type 2 error. In fact, in our case making a type 2 (falsely excluding a relevant variable) error may be worse than making a type 1 (falsely including an irrelevant variable). However, the main point of Green, Kim and Yoon was that of bias in the estimates not types of errors.

Oneal and Russett make a number of other statements in their article but those I have presented here are ones that I feel have relevance to MelC and that I agree with (at least to some extent).

The next in this sequence of articles is one by **Beck and Katz** (Beck & Jonathan N. Katz, 2001) who agree with Oneal and Russett that fixed effects models are not appropriate for international relations (IR) time-series cross-section (TSCS) models with binary dependent variables. They have stated IR and not all time-series cross-section models with binary dependent variables. I am also not sure that we could really consider the CHDS data to be TSCS data so I would not take this statement (or others in the TSCS literature) as is and apply to MelC but I do feel that we can gain relevant knowledge from looking at the TSCS literature.

The last article in this sequence of articles is one by **Gary King** (King, 2001). He takes an approach which is more balanced between Green, Kim and Yoon and Oneal and Russett and Beck and Katz. He focuses on the amount of information per unit, i.e. the number of observations per unit and also the number of events (or in our case, the amount of change) for each unit (in our case, each child). Simply put: you do not get good estimates if you do not have much information. I make a note myself that if there is no within child variation in the outcome to explain why would we try and explain it with the independent variables? A sensible first (or early-on step) would be to look into the amount of within-child variation in the outcome variable.

King states as his solution that we use a model that borrows strength from similar units to help estimate quantities of interest in each one but that if we use these approaches (e.g. Bayesian hierarchical, random effects or spilt population models) that we must change the assumption that the unobserved, but estimated heterogeneity is independent of the X variables. By just assuming this independence the omitted variable problem (bias) is still present but he states that this assumption is not difficult to change. However, I know of no way to change this assumption.

4. The Between and Within Approach of Goodrich

Two versions of Goodrich's working paper "Problems with and Solutions for Two-Dimensional Models of Continuous Dependent Variables" can be found online: a 2004 version (Goodrich, 2004) and a 2005 version (Goodrich, 2005). Although the 2005 version may be more up-to-date the maths is in a more digestible format in the 2004 version.

First I need to define within and between effects. A *within* effect is defined as the expected change in a child's dependent variable when its independent variable increases by one, holding everything else constant. A *between* effect is defined as the expected difference in the dependent variable between two children (who are the same except for the difference in the independent variable). An example Goodrich gives is from Gould (Gould, 2001): Suppose the dependent variable is a sample of Americans' wages over time, and the independent variable is a dummy variable that codes 1 for living in a southern state. Wages in southern states are lower than in other parts of the country, on average, and the so the cross-sectional or between effect of the "South" dummy is expected to be negative. However, if people *move* to the south, they are likely taking better-paying jobs. Thus, the temporal or within effect of the "South" dummy variable is expected to be positive.

Goodrich names a model presented by Zorn (Zorn, 2001) the simultaneous parsed model (SPM) – *parsed* because the effects of the covariates are split into their between and within components and *simultaneous* because the between and within estimates are obtained at the same time (rather than consecutively with Goodrich then proceeds to do).

To implement the SPM is easy: one just decomposes each variable into what are called *meaned* and *deviation* variables. To construct the meaned variables, for each variable, calculate the mean (over time) for each child and to construct the deviation variables, for each child, for each independent variable, calculate the difference (or deviation) from their mean. The meaned variables contain the between-child variation and the deviation variables contain the within-child information. One then simply includes the meaned and deviation variables in the model rather than the variables in their original format. The dependent variable in included in its original format for the SPM.

This SPM appears to be the same to me as Allison's hybrid model (found in (Allison, 2005)) and this is the term I will generally use to refer to a model with meaned and deviation variables included instead of the variables in their original formats.

Now I introduce some notation: $\bar{\boldsymbol{X}}_i$ represents a meaned independent variable, $\bar{\boldsymbol{y}}_{i[t-1]}$ represents a meaned LDV, $\bar{\boldsymbol{y}}_i$ represents a meaned dependent variable, $\tilde{\boldsymbol{X}}_{it}$ represents an independent deviation variable, $\tilde{\boldsymbol{y}}_{it-1}$ represents a lagged dependent deviation variable and $\tilde{\boldsymbol{y}}_{it}$ represents a dependent deviation variable. We have child *i* where *i* = 1, …, *N* and measurement at time *t* where *t* = 1, …, *T*.

The SPM can be seen as the sum of the between estimator (B-E) and the within estimator (W-E) where the B-E is

$$\bar{\boldsymbol{y}}_i = \alpha + \varphi_b \bar{\boldsymbol{y}}_{i[t-1]} + \boldsymbol{\beta}_b \bar{\mathbf{X}}_i + \bar{\boldsymbol{\epsilon}}_i$$

and the W-E is

$$\tilde{\boldsymbol{y}}_{it} = 0 + \varphi_w \tilde{\boldsymbol{y}}_{it-1} + \boldsymbol{\beta}_w \tilde{\boldsymbol{X}}_{it} + \tilde{\boldsymbol{\epsilon}}_{it}$$

as $\boldsymbol{y}_{it} = \bar{\boldsymbol{y}}_i + \tilde{\boldsymbol{y}}_{it}$. Also note that the W-E is the same as the standard fixed effects model.

The pooled model,

$$\boldsymbol{y}_{it} = \alpha + \varphi \boldsymbol{y}_{it-1} + \boldsymbol{\beta} \boldsymbol{X}_{it} + \boldsymbol{\epsilon}_{it},$$

can be thought of as an SPM with constraints $\varphi_b = \varphi_w$ and $\boldsymbol{\beta}_b = \boldsymbol{\beta}_w$, i.e. the between and within effects are assumed to be equal. It also assumes that the child effects ($\bar{\boldsymbol{\epsilon}}_i$) do not exist (or that they are all equal).

Goodrich states that “the constraints that $\boldsymbol{\beta}_b = \boldsymbol{\beta}_w$ never have a theoretical justification when an LDV is included in the SPM. I do not fully understand the reasons for this but it may be because it does not make sense for one person's lagged value to be able to affect a different person – the effect must be within person. But for our simulation purposes we need the LDV to keep the coherent trajectory of an individual. Goodrich also gives reasons why we should expect φ_w and φ_b to differ (on page 11 of the 2005 version) which again aren't very clear.

A random effects model produces estimates that are a weighted average of the between and within estimates also with weights dependent on the variances of the error terms. A fraction of the unit means is subtracted from the two-dimensional variables, where the fractional parameter, $\hat{\omega}$, equals $1 - \sqrt{\frac{\widehat{Var}(\tilde{\epsilon}_{it})}{\widehat{Var}(\bar{\epsilon}_i)}}$.

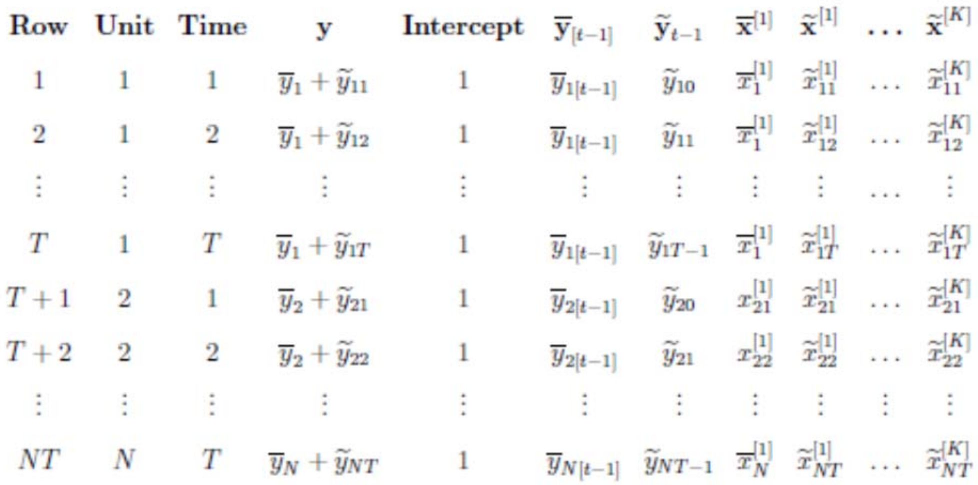
The pooled model also produces estimates that are a weighted average of the between and within estimates. The weights are determined differently from the random effects model[4]. For both models, more weight is given to the dimension that has more variance.[5] For the random effects model

Goodrich, although happy with Zorn's coefficient estimates is not happy with his standard errors. The standard errors are correct if the B-E and W-E models are estimated separately, but, when done all together in Zorn's model, standard errors for both the between and within estimates are wrong. It is intuitive to see that the standard errors for the between estimates are too small because the means are copied *T* times for each child. The standard errors for the within estimates are also too small; we need to subtract *N* from the degrees of freedom because *N* child-specific means were estimated for each covariate when calculating the deviation variables.

To go forward from this point Goodrich rearranges that dataset so it looks like this

[4] Not exactly sure how they are calculated – it may depend on the variance of the within and between components and it may also depend on the ratio of the number of units and the number of observations within units.

[5] I think! It is not 100% percent clear whether he is talking about the random effects model or the pooled model. See page 20 of the 2004 version.



Row	Unit	Time	y	Intercept	$\overline{\mathbf{y}}_{[t-1]}$	$\widetilde{\mathbf{y}}_{t-1}$	$\overline{\mathbf{x}}^{[1]}$	$\widetilde{\mathbf{x}}^{[1]}$	…	$\widetilde{\mathbf{x}}^{[K]}$
1	1	1	$\overline{y}_1 + \widetilde{y}_{11}$	1	$\overline{y}_{1[t-1]}$	$\widetilde{y}_{10}$	$\overline{x}^{[1]}_1$	$\widetilde{x}^{[1]}_{11}$	…	$\widetilde{x}^{[K]}_{11}$
2	1	2	$\overline{y}_1 + \widetilde{y}_{12}$	1	$\overline{y}_{1[t-1]}$	$\widetilde{y}_{11}$	$\overline{x}^{[1]}_1$	$\widetilde{x}^{[1]}_{12}$	…	$\widetilde{x}^{[K]}_{12}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…	⋮
T	1	T	$\overline{y}_1 + \widetilde{y}_{1T}$	1	$\overline{y}_{1[t-1]}$	$\widetilde{y}_{1T-1}$	$\overline{x}^{[1]}_1$	$\widetilde{x}^{[1]}_{1T}$	…	$\widetilde{x}^{[K]}_{1T}$
$T+1$	2	1	$\overline{y}_2 + \widetilde{y}_{21}$	1	$\overline{y}_{2[t-1]}$	$\widetilde{y}_{20}$	$x^{[1]}_{21}$	$\widetilde{x}^{[1]}_{21}$	…	$\widetilde{x}^{[K]}_{21}$
$T+2$	2	2	$\overline{y}_2 + \widetilde{y}_{22}$	1	$\overline{y}_{2[t-1]}$	$\widetilde{y}_{21}$	$x^{[1]}_{22}$	$\widetilde{x}^{[1]}_{22}$	…	$\widetilde{x}^{[K]}_{22}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
NT	N	T	$\overline{y}_N + \widetilde{y}_{NT}$	1	$\overline{y}_{N[t-1]}$	$\widetilde{y}_{NT-1}$	$\overline{x}^{[1]}_N$	$\widetilde{x}^{[1]}_{NT}$	…	$\widetilde{x}^{[K]}_{NT}$

If were we to run the SPM or pooled models on this dataset we would be estimating what he calls the SPM2 or the pooled2 model.

Goodrich is of the opinion that it is best to estimate the W-E and B-E separately (an approach he calls the consecutive parsed estimator (CPE)) using a dataset of the above form. He provides degrees of freedom adjustments necessary to obtain correct standard errors for models with a continuous outcome; no solution is provided for the standard errors for models with a dichotomous or count outcome[6].

If one wishes, they can then combine the estimates by taking a simple mean. He also gives a formula to calculate standard errors for the averages estimates but we would not need this in our application to MelC. However, if $\widetilde{\boldsymbol{y}}_{it-1}$ is included in the W-E then $\widehat{\boldsymbol{\beta}}_w$ represents short-term effects and there is no basis for averaging a short-term effect with a between effect. To obtain a plausible average one must first calculate long-run within estimates. This long-run effect can then be used in place of $\hat{\beta}_w$ in taking the average. The formula for the long-run effect is $\frac{\widehat{\beta}_w+\widehat{\gamma}_w}{1-\varphi_w}$ and its standard error (which can be used in place of the standard error for $\widehat{\boldsymbol{\beta}}_w$ in his formula) can be estimated by simulation or by the delta method.

Some notes about the CPE method:

- the coefficient estimates from the estimating the CPE are identical to those of the good SPM2 (as opposed to the bad SPM I assume).
- B-Es in textbooks never include the unit means of an LDV on the right hand side (for reasons I don't really understand given on p.11 of the 2005 version).

Goodrich's CPE approach is that it is developed for continuous data. Goodrich states in his 2005 version that the same problems, but not necessarily the same solutions, apply with generalised linear models. To take means and differences of a binary outcome does make sense and also one cannot go and estimate a logistic model that will give an equation for a probability if the response variable is not 0s and 1s.

[6] For the W-E the degrees of freedom correction is to multiply the standard errors by sqrt((N * J + N - W.E.$rank) / (N * J - N - W.E.$rank)) where N is the number of units, J is the number of observations taken on each unit (Goodrich assumes a balanced dataset but perhaps the mean could be used here if it were not?) and W.E.$rank is the rank of the W-E model which is equal to the number of parameters estimated in the W-E model.
For the B-E the degrees of freedom correction is to multiply the standard errors by sqrt((N * J + N - B.E.$rank) /(N - B.E.$rank)) where N, J and B.E.$rank are defined similar to those for the W-E.

Goodrich does seem to think that including an LDV in an RE model is worse than including it in a pooled model[7]

5. Implementing Models in the Simulation

A major factor in choosing a modelling approach for the MelC project is how the model will be implemented in the micro-simulation.

5.1 Thoughts about Hybrid Models in the Simulation

If we found that, for an outcome variable, there were child intercepts that were correlated with the regressors then a hybrid model may appear to be the best modelling option. We get the fixed effects estimates but also can include time-invariant variables. For continuous variables we could use Goodrich's CPE and for dichotomous variables we could use Zorn's logistic hybrid model. To recap, this model gives unbiased estimates but the standard errors are too small. The consequences of this seem small to me. We would end up including a variable that is not strictly necessary. It would seem much worse, in the simulation, to exclude a variable that is needed.

Although a hybrid model may be a good modelling approach, we do not only want an accurate model – more important is constructing the simulation. When we come to thinking about the simulation the hybrid model may pose problems. The estimates are split up into between and within components. This would mean in the simulation that we would have to work out means and deviations at each year. If these were used with estimates from a single hybrid model estimated using all years at once, the simulated data may not end up being very similar to the real data due to the means and deviations never being able to be very small in the second year. The deviation in the second year will change as we learn more about an individual and we will not have represented the individual well in the first few years.

To get around this problem we would need to build separate hybrid models for each year (prior to the simulation) and use at each year in the simulation.

Another option may be to average the between and within estimates. We could then apply these and not have to calculate means and deviations year by year in the simulation and so we would not need to build separate hybrid models for each year. If we were to average the between and within effects one may ask what the difference is between this and between the weighted averaged effects of the pooled or random effects models. The difference is the weights. When we average the effects ourselves from the hybrid model we get to choose the weights and we also get to choose which effects to average. Goodrich implies that we would want to choose equal weights but there may be an argument for letting the data determine the weights.

5.2 Pooled and Random Effects Models Implemented in the Simulation

In contrast to the hybrid model, the pooled and random/mixed effects models are straight-forward and intuitive to implement in the simulation. The random effects model has the extra step of drawing a random effect for each child (for each modelled outcome). The effect may be drawn from a normal distribution with parameters estimated from the real data or it may drawn from the empirical distribution of random effects. At each year in the simulation, the same individual intercept or effect is simply added in when using the formula (coefficients) from the estimated model. To clarify, each child will have a different intercept from other children but, for a specific child, their own intercept will remain the same over the five years.

5.3 Keeping Dynamism and Coherency in the Simulation

Including an LDV in the model was a way to keep coherency or dynamism in the simulation. Another way of keeping dynamism in the model without including an LDV is to estimate multiple models for each prior state (each value of the LDV) as mentioned in Section 2.5. This is also straight-forward to

[7] I think but not 100% sure – see comment in conclusion p.33-34, 2004 version.

implement in the simulation but can only be used if there are enough observations in each prior state group.

A random effect does not add dynamism into the model. A random effect keeps a child's probability of an event (or level for a continuous variable) fluctuating around a certain level. One would end up with a good predictive model but a change in one year will have no effect on the predicted values of the next year. If we looked at individual trajectories we would have sudden "blips" that were not taken into account when simulating the next year. A random effects model without an LDV will certainly keep a more consistent story for individuals than an ordinary regression without an LDV but, for any regression model without an LDV, a change in the outcome does not affect the probability of the outcome in the next simulated year.

If one were to build separate models for each previous state and use these depending on the previous state of the outcome variable, there would be dynamism in the model and a change in one year of the outcome would affect the next year. This is an excellent solution for variables that have enough observations in each group and have a limited number of groups.

6. Other Things to Consider when Choosing a Modelling Approach

From Beck and Katz (Beck & J.N. Katz, 2004)**:**

- Is the number of observations per child large enough to do serious averaging over time? (p. 3). Does small T matter for the averaging?
- Should we consider an instrumental variable (IV) approach?
  - These approaches were developed for the panel data case (where *T*, the number of observations per unit, is much smaller than in TSCS data) because the smaller *T*, the larger the bias due to including an LDV. (p. 9)
  - Proposed IVs: second lag of the dependent variable
  - But would we lose too much data doing this? But they said this was specifically developed for the case where T = 3 or 4 so maybe they have some way around it.
    - To get an estimate - in simulation probably could still apply this estimate to all years
- While consistency is a desirable property, what we really care about is the mean squared error properties for the sample sizes we have, not for infinite sample sizes. (p. 6)
- While an IV estimator may be unbiased, it may dramatically increase mean squared error if the instrument is not highly correlated with the problematic variable (p.9)
  - We may be trading a small reduction in bias for a large decrease in efficiency.
  - But I think the second lag of the dependent variable would be highly correlated with the first lag.
  - But is the IV approach only for fixed effects methods?
- Non-statistical costs to using fancy methods (p. 7)

7. A Procedure to Choose Between Modelling Options

7.1 A Flow Diagram

From all the various issues thought about in this document I attempted to put together a flowchart that provides a way to navigate through all the different aspects and questions that must be asked in choosing a modelling approach. This flowchart is presented in **Error! Reference source not found.**. We can first ask about the amount of within-child change. If there is no within-child change then we cannot use fixed effects estimates to explain it. Hence, if there is no within-child change the fixed effects and hybrid models are ruled out. If there is very little within-child change then the fixed effects or hybrid models may also not be the best option (see Section 3.1.3). How we decide whether there is enough within-child change to consider using a hybrid model is a question we have not yet given much thought.

If there is not much within-child change then the next question we ask is whether, theoretically, we need to consider the previous state in the model. If the answer is not we can use either a pooled regression or a random effects model. We can choose between these models by testing whether the variation in the individual child intercepts is significantly different from zero using a Covtest statement as stated earlier in Section 3.1.2.

If we do need to consider the previous state of the outcome variable then if the outcome is a count of dichotomous variable we can build multiple models – one for each prior state. This is dependent on there being enough observations in each group to build a model. For categorical variables with more than two categories we may be able to group similar categories in order to get enough observations in the group.

If the outcome in continuous, a count variable or a categorical variable with too many categories or too few observations in each group to build multiple models then we go back to using a pooled of random effects model. We can try models with and without the LDV and compare the simulated data to the real data to help us choose between them. It is possible as mentioned in Section 5.3 that the random effect in the simulation may negate the need for the LDV in the model.

If there is a notable amount of within-child change then we can next ask whether the between and within effects are the same for each variable. Following this, regardless of the answer is the theoretical question again of whether we need to consider the previous state. **If the between and within effects are the same for all the explanatory variables** (or are not too different) and we do not think that the previous state of the outcome influences the current state of the outcome then we will consider using a pooled regression or a random effects model. The choice between these will be based on whether there is a child effect (that is, whether the variation in the individual child intercepts is significantly greater than zero).

If the between and within effects are the same for all the explanatory variables (or are not too different) and we do think that the previous state of the outcome influences the current state of the outcome then, if we are able we will build separate pooled or RE models for each previous state. If we are not able to do this (continuous or count outcome or not enough observations in the categories) then we will try a single pooled or random effects model for the outcome. Again the presence or absence of a child effect will be the deciding factor in choosing between the pooled and RE models. In this situation we want to include the LDV but we are aware that there may be bias if we do. So we will try models with and without the LDV and, for each, compare the simulated data to the real data.

If the between and within effects differ for some of the explanatory variables (which seems a highly likely outcome) and if we do not think that the previous state of the outcome influences the current state of the outcome, we could build a single hybrid model for the outcome. To get around the problems of implementing the hybrid model in the simulation, we have to ask the question here whether we will build separate hybrid models for each year (only using the data for the current year and previous years) and whether, in using the model coefficients in the simulation we will average the between and within estimates.

If the between and within effects differ for some of the explanatory and if we do think that the previous state of the outcome influences the current state of the outcome, we could build separate hybrid models for each value of the previous state. Again we would have to consider building separate models for each year and think about whether we want to average between and within estimates.

If we cannot build separate models for each previous state then we could build a hybrid model (taking into consideration again the separate models for each year and the question of averaging) and again we would try models with and without the LDV and for each compare the simulated data to the real data.

A second version of the flow diagram is provided in . After team discussion we considered that the possible solutions to the problems of implementing a hybrid in the simulation would be a lot of extra work for a small amount of again. We were also not sure if the solutions we have thought of for the problems associated with the hybrid model will work well in practice and there may be other problems we have not yet thought of. For these reasons, in the second version of the flow diagram, if one finds oneself at one of the hybrid model boxes, the question of whether the hybrid model is problematic in the simulation is asked. The answer to this question is always 'yes' and the user of the flow diagram is directed back to an appropriate point in the diagram that will result in him or her ending up at one one of the boxes that declares that the model model to use is a pooled or random effects model for either all the data at once or for subsets of the data based on the value of the previous state. This is the flow diagram that was used in practice for choosing the modelling approach for each outcome.

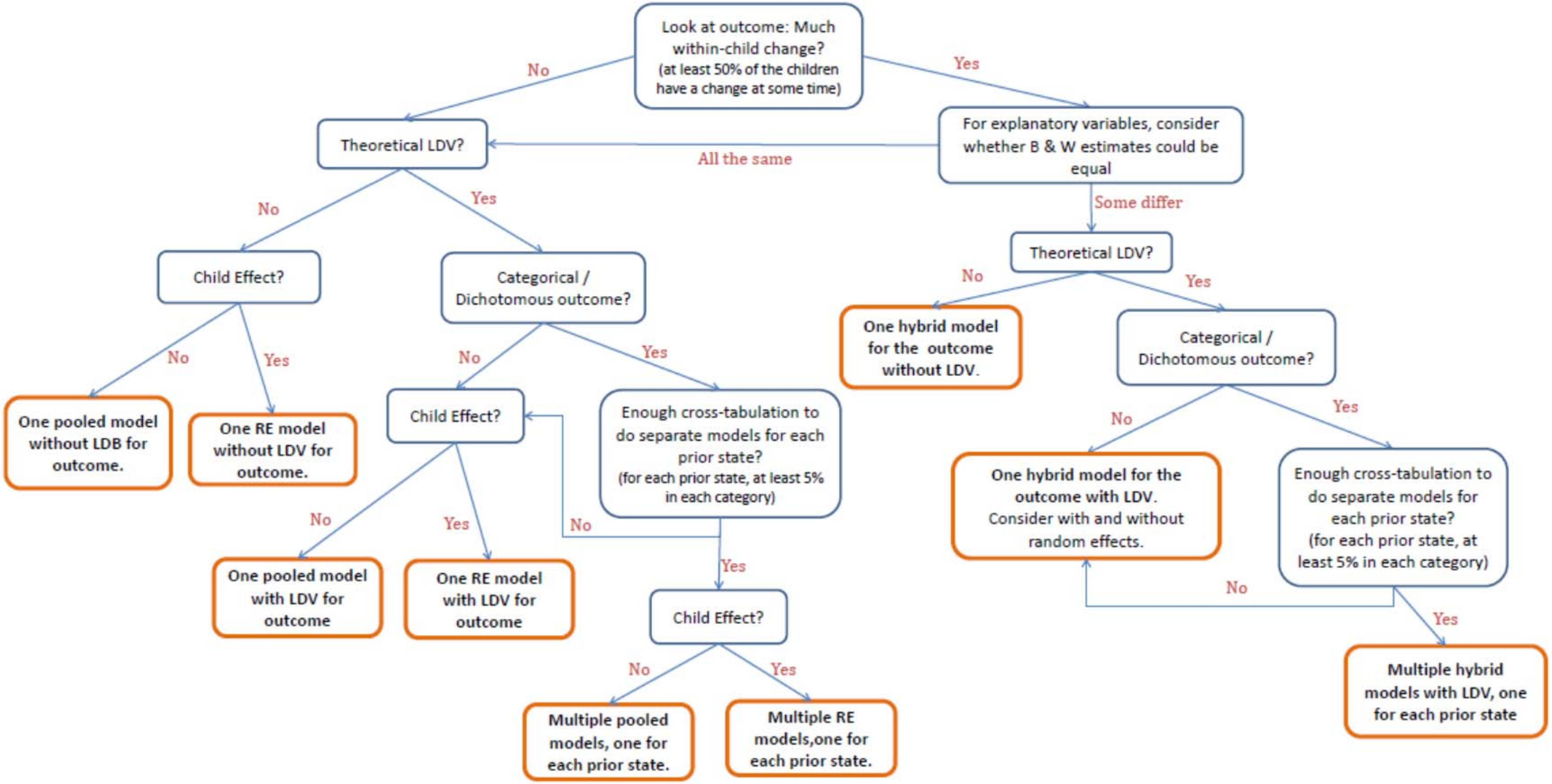

Look at outcome: Much within-child change? (at least 50% of the children have a change at some time)
No
Yes
Theoretical LDV?
For explanatory variables, consider whether B & W estimates could be equal
All the same
Some differ
No
Yes
Child Effect?
Categorical / Dichotomous outcome?
Theoretical LDV?
No
Yes
One hybrid model for the outcome without LDV.
Categorical / Dichotomous outcome?
No
Yes
No
Yes
One pooled model without LDB for outcome.
One RE model without LDV for outcome.
Child Effect?
Enough cross-tabulation to do separate models for each prior state? (for each prior state, at least 5% in each category)
No
Yes
One hybrid model for the outcome with LDV. Consider with and without random effects.
Enough cross-tabulation to do separate models for each prior state? (for each prior state, at least 5% in each category)
No
Yes
No
One pooled model with LDV for outcome
One RE model with LDV for outcome
Yes
No
Yes
Child Effect?
Multiple hybrid models with LDV, one for each prior state
No
Yes
Multiple pooled models, one for each prior state.
Multiple RE models,one for each prior state.

Figure 5: Model Choice Flow-diagram version 1

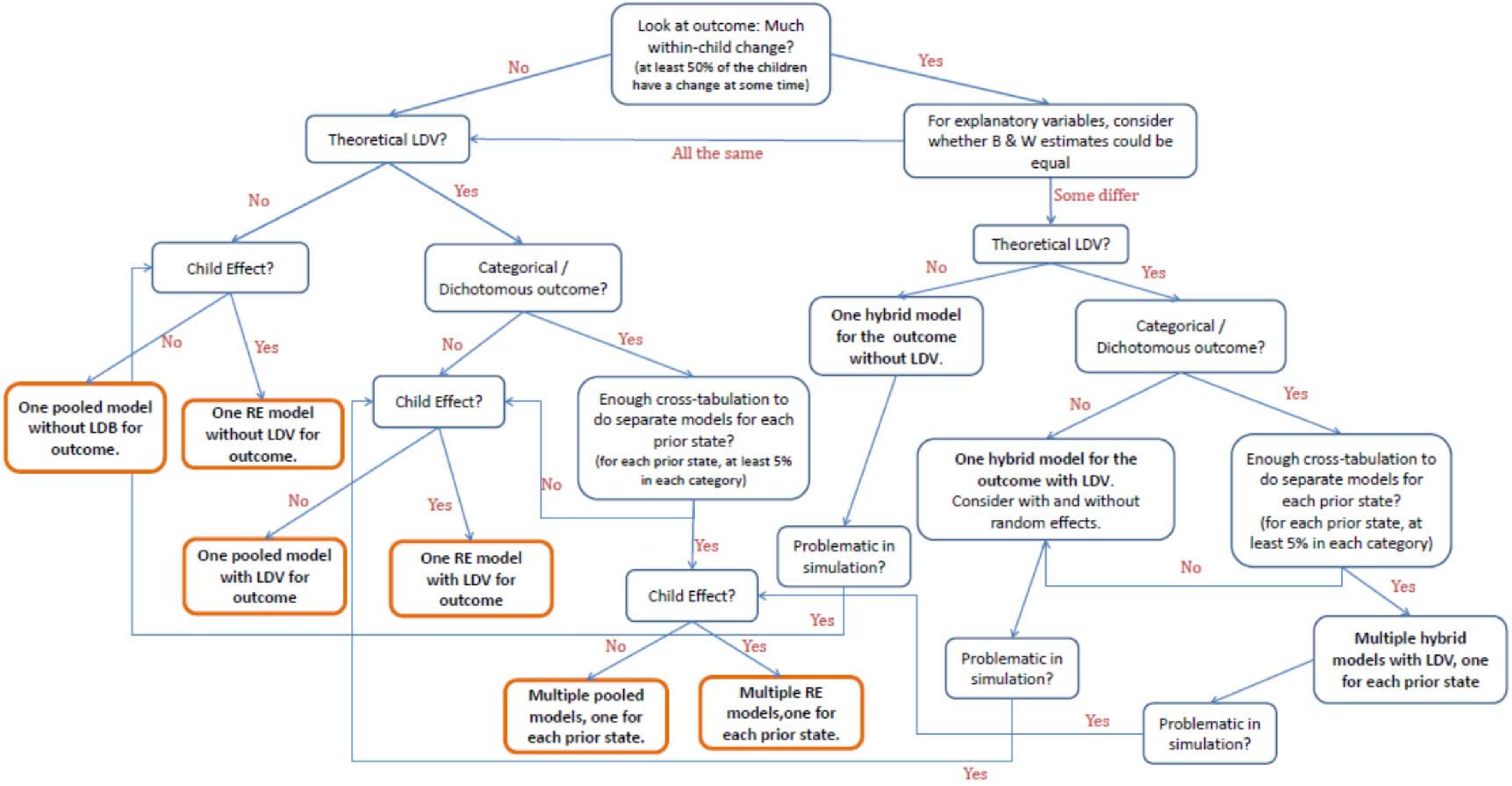

Look at outcome: Much within-child change?
(at least 50% of the children have a change at some time)
No
Yes
Theoretical LDV?
For explanatory variables, consider whether B & W estimates could be equal
All the same
Some differ
No
Yes
Child Effect?
Categorical / Dichotomous outcome?
Theoretical LDV?
No
Yes
One hybrid model for the outcome without LDV.
Categorical / Dichotomous outcome?
No
Yes
No
Yes
One pooled model without LDB for outcome.
One RE model without LDV for outcome.
Child Effect?
Enough cross-tabulation to do separate models for each prior state?
(for each prior state, at least 5% in each category)
No
Yes
One hybrid model for the outcome with LDV. Consider with and without random effects.
Enough cross-tabulation to do separate models for each prior state?
(for each prior state, at least 5% in each category)
No
Yes
No
One pooled model with LDV for outcome
One RE model with LDV for outcome
Yes
Problematic in simulation?
No
Yes
Child Effect?
Yes
Multiple hybrid models with LDV, one for each prior state
No
Yes
Problematic in simulation?
Multiple pooled models, one for each prior state.
Multiple RE models,one for each prior state.
Yes
Problematic in simulation?
Yes

Figure 6: Model choice Flow-diagram version2. This is the version that was used in practice

7.2 How Do We Assess the Amount of Within-Child Change and Whether the Between and Within Effects are the Same?

These questions have not yet been addressed in this document. There is no formal statistical test available to determine whether there is a lot of within-child change or only a little. This must be determined somewhat subjectively by the analyst/team.

In response to the question of whether the within and between effects could be the same, it is not accurate to use the SPM (or Zorn's) model to determine whether this because the standard errors are wrong for this model. Goodrich recommends using the SPM2, but the SPM2 is only available for continuous outcomes. So, for cases where we have a non-continuous outcome, we can either use a non-reliable test (from the SPM), simply eyeball the coefficients or use theory to assess whether we think the effects should be the same or not. I would recommend using all three of these in combination to make the decision.

References

The pdfs of these papers are available in: S:\Symonds Group\soc\Sociology Research Group\Projects\FRST - MEL-C\LITERATURE\JessicasInvestigationsDocumentPapers.

Allison, P. D. (2005). *Fixed effects regression methods for longitudinal data using SAS*. Cary, NC: SAS Institute Inc.

Ashley, R. A. (2010). A Reconsideration of Consistent Estimation of a Dynamic Panel Data Model in the Random Effects (Error Components) Framework. Retrieved from http://ashleymac.econ.vt.edu/ashleyprofile.ht.

Beck, N., & Katz, J.N. (2004). Time-series-cross-section issues: dynamics, 2004. *annual meeting of the Society for Political Methodology, Stanford University*. Retrieved May 11, 2011, from http://www.nyu.edu/gsas/dept/politics/faculty/beck/beckkatz.pdf.

Beck, N., & Katz, Jonathan N. (2001). Throwing Out the Baby with the Bath Water: A Comment on Green, Kim, and Yoon. *International Organization*, *55*(2), 487-495. doi: 10.1162/00208180151140658.

Cameron, A. C. (2005). *Microeconometrics: methods and applications*. New York: Cambridge University Press. Retrieved May 12, 2011, from http://books.google.com/books?hl=en&amp;lr=&amp;id=Zf0gCwxC9ocC&amp;oi=fnd&amp;pg=PR15&amp;dq=Microeconometrics:+methods+and+applications&amp;ots=CY29mN0IpY&amp;sig=HW_Qe85UliRMygb7BnrafVVFMP8.

Goodrich, B. (2004). Problems with and Solutions for Two-Dimensional Models of Continuous Dependent Variables. *Department of Government, Harvard University. Typescript*. Retrieved May 11, 2011, from http://www.people.fas.harvard.edu/~goodrich/Research/Methods/TSCS/ProblemsSolutions.pdf.

Goodrich, B. (2005). Problems with and Solutions for Two-Dimensional Models of Continuous Dependent Variables. *Department of Government, Harvard University*. Retrieved May 11, 2011, from http://www.people.fas.harvard.edu/~goodrich/Research/Methods/TSCS/ProblemsSolutions.pdf.

Gould, W. (2001). What is the between estimator? *Stata Frequently Asked Questions*. Retrieved May 12, 2011, from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:What+is+the+between+estimator#0.

Green, D. P., Kim, S. Y., & Yoon, D. H. (2001). Dirty Pool. *International Organization*, *55*(2), 441-468. doi: 10.1162/00208180151140630.

Keele, L., & Kelly, N. J. (2005). Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. *Political Analysis*, *14*(2), 186-205. doi: 10.1093/pan/mpj006.

King, G. (2001). Proper Nouns and Methodological Propriety: Pooling Dyads in International Relations Data. *International Organization*, *55*(2), 497-507. doi: 10.1162/00208180151140667.

Oneal, J. R., & Russet, B. M. (1997). The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950-1985. *International Studies Quarterly*, *41*(2), 267-294. doi: 10.1111/1468-2478.00042.

Oneal, J. R., & Russett, B. (2001). Clear and Clean: The Fixed Effects of the Liberal Peace. *International Organization*, *55*(2), 469-485. doi: 10.1162/00208180151140649.

Zorn, C. (2001). Estimating between- and within-cluster covariate effects, with an application to models of international disputes. *International Interactions*, *27*(4), 433-445. doi: 10.1080/03050620108434993.