Multilevel Models for Change:
With a focus on the between and within variation

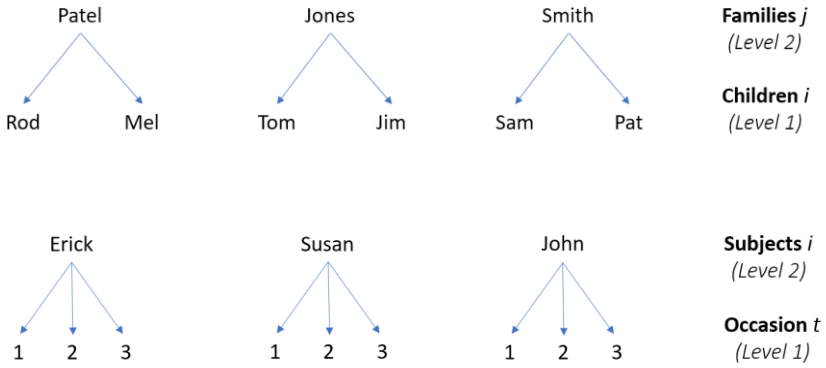
Naveed Sharif, M.S.[2]
November 25, 2018

Abstract

In this paper, I will be introducing the statistical multilevel model for change, demonstrating how it allows us to address within-individual and between-individual questions about change simultaneously. I begin by specifying the multilevel model for change while postulating a pair of subsidiary models—a level-1 submodel that describes how each individual change over time, and a level-2 submodel that describes how the changes differ across individuals. I then follow up with a composite model that collapses the level-1 and level-2 submodels together algebraically. A distinctive feature of the composite multilevel model is its composite residual. The mathematical form of the composite residual reveals two import properties about the residuals not readily apparent in the level-1 and level-2 specification: they can be both autocorrelated and heteroskedastic within individual.

Units of observations often fall into groups or clusters. For example, individuals could be nested in families, hospitals, schools, neighborhoods, or firms. Longitudinal data also consists of clusters of observations made at different occasions for the same individual or cluster. For two examples of clustered data, the nesting structure is depicted in figure [0].

Figure 0: Examples of clustered data



On the top panel of figure [0], children are nested within families and is cross-sectional data. Whereas, on the bottom panel of figure [0], occasions are nested in subjects and is longitudinal data. In clustered data, it is usually important to allow for dependence or correlations among the responses observed for units belonging to the same cluster. For example, the adult weights of siblings are likely to be correlated because siblings are genetically related to each other and have usually been raised with the same family. Multilevel models are designed to model and estimate such within-cluster correlations. Therefore, we would use multilevel modeling when there is reason to believe that observations are not necessarily independent of one another.

Before presenting the multilevel model for change itself, I feel it is wise to first briefly review the purpose of a statistical model. Statistical models are mathematical representations of population behavior. Statistical models attempt to describe features of the hypothesized process of interest among individuals in the target population. When you use a statistical model to analyze a set of data, you implicitly assume that the dataset you are using came from a target population. Therefore, statistical models are not statements about sample behavior; they are statements about the population process that generated the data. [2]

To be explicit here about the population process, statistical models are expressed using parameters—intercepts, slopes, variances, and so on—that represent specific population quantities of interest. For example, suppose you were to use the following simple linear regression model to represent the relationship between the *Education* and *Wage* of an individual, on a single cross-sectional data set;

[1.0] $$Wage_i = \beta_0 + \beta_1(Education_i - \overline{Education}) + \varepsilon_i$$

You would be declaring implicitly that in the population from which your sample was drawn: (1) β_0 is an unknown intercept parameter that represents the expected level of *Wage* for an individual with an average level of education, and (2) β_1 is an unknown slope parameter that represents the expected difference in functioning between individuals whose education level differs by one unit. You then fit the model and examine the statistical significance of its parameters, such as a *t-statistic* on *Education* and/or

a measure of "goodness-of-fit" such as an R^2 that quantifies the correspondence between the fitted model and the sample data. If the model fits well, you can use the estimated parameter values to draw conclusions about the direction and magnitude of the hypothesized effects in the population. For example, suppose you fit equation [1.0] and find that *Education* is statistically significant;

$$Wage_i = 20 + 5(Education_i - \overline{Education}) + \varepsilon_i$$

You will be able to predict that an average educated individual makes $20 an hour and that each additional year of education increases wage by $5. Hypothesized test (e.g. *t-statistic*) and confidence intervals could then be used to make inferences from the sample back to the population.

Equation [1.0] is designed for cross-sectional data (1-year data). The multilevel model for change requires an alternative approach to modeling the data and answers different research questions: level-1 questions about within-individual change and level-2 questions about between-individual differences in change. If the questions you were seeking on the relationship between the education level of an individual on wage was based on longitudinal data (repeated measures of each individual over time), we instead might ask: (1) How does each individuals wage change over time? and (2) do individual's trajectories of change vary by education level. This distinction among within-individual and between-individual provides the core rationale for specifying a statistical model for change. Taken together, these two components, level-1 and level-2 submodel's, form the multilevel statistical model.

For the remainder of this paper, I will be developing and explaining the multilevel model for change using an example of 7 waves of data collected by the Predictive Modeling and Analytics team in Kaiser Permanente under Raymond Fong's leadership. Figure [1], below, is a sample of the dataset where *region_account_number* is a repeated measure of each individual (in this example, individual equals group) over time. *effective_date_year* is the time interval used to measure the value of a variable at a corresponding period in time (e.g. CO_1000 in 2012 had penetration rate of 58%, in 2013 a 54%, in 2014 a 50%, and in 2015 the penetration rate was 48%).

Figure 1: Excerpts from the group-period dataset

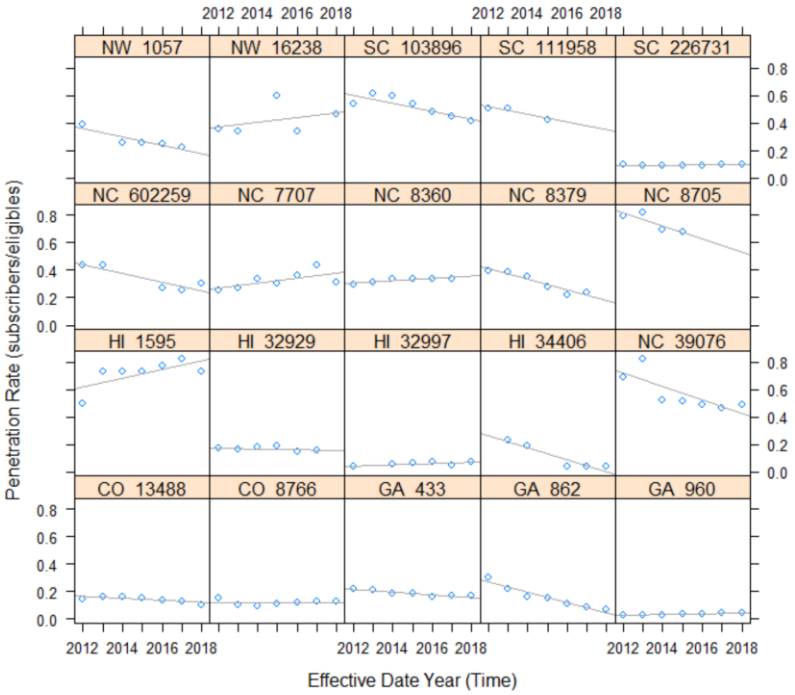
	region_account_number	effective_date_year	pen_rate	region	ri_mean	firm_size
1	CO_1000	2012	0.577	CO	NA	GT500_LTE1000
2	CO_1000	2013	0.538	CO	0.13	GT500_LTE1000
3	CO_1000	2014	0.500	CO	0.13	GT500_LTE1000
4	CO_1000	2015	0.483	CO	0.03	GT500_LTE1000
5	CO_1000	2016	0.374	CO	0.07	GT500_LTE1000
6	CO_1000	2017	0.345	CO	0.00	GT500_LTE1000
7	CO_1000	2018	0.301	CO	0.10	GT500_LTE1000
8	CO_10650	2012	0.174	CO	NA	GT1000_LTE3000
9	CO_10650	2013	0.177	CO	0.07	GT1000_LTE3000
10	CO_10650	2014	0.179	CO	0.15	GT1000_LTE3000
11	CO_10650	2016	0.039	CO	0.06	GT1000_LTE3000
12	CO_10650	2017	0.041	CO	0.01	GT1000_LTE3000
13	CO_10650	2018	0.048	CO	0.06	GT1000_LTE3000

Each group can have up to 7 waves of data (2012 through 2018). The dataset is unbalanced (some groups will have 4 or 5 waves of data versus all groups having a constant 7 waves of data). Additionally, the spread between each wave of data can vary. Most are 1 year, some are 2 years or more (e.g. CO_10650). The objective of the multilevel model will be to (1) evaluate the estimated change trajectory for each group and (2) why the change trajectories varies between groups.

The Level-1 Submodel for Group Change

The level-1 component of the multilevel model, also known as the individual growth model, represents the change we expect each individual of the population to experience during the time period under study. We must believe that the observed data could reasonably have come from a population in which the model is functioning. To align and reinforce expectations, I will precede level-1 submodel specification with visual inspection of the empirical growth plots. Figure [2], below, presents empirical growth plots of *Penetration_Rate* vs *Time* (*effective_date_year*). Since there are approximately 787 groups in the study, I took a random sample to avoid the clutter of attempting to plot all 787 groups.

Figure 2: Using empirical growth plots to identify a functional form of the level-1 submodel



When examining empirical growth plots like these, it's important to have an eye toward the ultimate model specification. The majority of the groups from the sample seem to have a downward linear change trajectory. Therefore, an individual growth model in which change is a linear function of *Time*, can be written as;

[1.1: Level-1 Submodel] $$Penetration\_Rate_{it} = [\pi_{0i} + \pi_{1i}(Time_{it} - 2012)] + [\varepsilon_{it}]$$

Equation [1.1] uses two subscripts, *i* and *t*, to identify groups and occasions, respectively. For this dataset *i* runs from 1 through 787 (for the 787 groups in this analysis) and *t* runs from 2012 through 2018 (for the seven waves of data). I used brackets in equation [1.1] to distinguish two parts of the submodel: the *structural part* (in the first set of brackets) and the *stochastic part* (in the second bracket). The *structural* part of the level-1 submodel embodies our hypotheses about the shape of each groups true trajectory of change over time. Equation [1.1] stipulates that this trajectory is linear with *Time* and has individual growth parameters, π_{0i} and π_{1i}, that characterizes its shape for the *i*th group in the population. These individual growth parameters are the population parameters that lie beneath the individual intercepts and slopes obtained when we fit OLS-estimated individual change trajectories in my

analysis. To clarify what the individual growth model says about the population, examine figure [3], which maps the model into imaginary data for an arbitrary selected member of the population, group *i*.

Notice that equation [1.1] uses a special representation for the predictor *Time*. This practice is known as centering, facilitates interpretation. By using ($Time_{it} - 2012$) as a level-1 predictor, instead of *Time*, the intercept π_{0i} represents group *i*'s true value of *Pen Rate* (*Penetration_Rate*) at year 2012. If I had simply used *Time* as a level-1 predictor, with no centering, π_{0i} would represent *i*'s true value of *Pen Rate* at year 0; a year that doesn't have an intuitive interpretation in this analysis. There are other temporal representations, including those in which we center time on its middle and final values. The approach I used here aligns π_{0i} with the first wave of data collection. This allows me to interpret its value using simple jargon: it is group *i*'s true *initial status.* If π_{0i} is large, group *i* has a high true *initial status*; If π_{0i} is small, group *i* has a low true *initial status*.

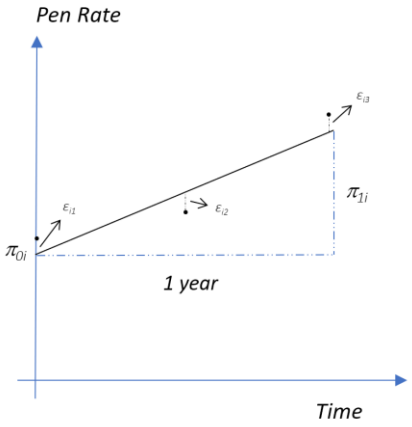
The second parameter in equation [1.1], π_{1i}, represents the slope of the postulated individual change trajectory. The slope is the most important parameter in the level-1 linear change submodel because it represents the rate at which individual *i* changes over time. Since *Time* units are in years, π_{1i} represents group *i*'s true annual *rate of change*. Because I hypothesize that each group in the population has their own slope and intercept, both parameters, π_{0i} and π_{1i}, are subscripted by *i*. Therefore, specifying a level-1 submodel attempts to describe everyone (all the *i*'s) in the population, I implicitly assume that all the true individual change trajectories have a common algebraic form. But I do not assume that everyone has the same exact trajectory. Each group will have their own individual growth parameters (π_{0i} and π_{1i}), so that different groups can have their own distinct change trajectories (more on this later).

Posting a level-1 submodel allows me to distinguish growth parameters. This leap is the cornerstone of individual growth modeling because it means that I can study interindividual differences in change by studying interindividual variation in the growth parameters. Imagine a population in which each member dips into a well of possible individual growth parameter values and selects a pair—a personal intercept and slope. These values determine their true change trajectory. Because each member draws their coefficients from an unknown random distribution of parameters, statisticians often call the multilevel model for change a *random effects* model. [7]

The Stochastic Part of the Level-1 Submodel

The stochastic part of the level-1 submodel appears in the second set of brackets on the right-hand side of equation [1.1]. Composed of just one term, the stochastic part represents the effect of random error ε_{it}, associated with the measurement of group *i* on occasion *t*. The level-1 errors appear in figure [3] as ε_{1t}, ε_{2t}, and ε_{3t}. Each group's true change trajectory is determined by the structural component of the submodel, π_{0i} and π_{1i}, but each group's observed change trajectory also reflects the measurement errors. The level-1 submodel accounts for these uncertainties—the differences between the true and observed trajectories—by including random errors: ε_{1t} for group *i*'s first measurement occasion, ε_{2t} for group *i*'s second measurement occasion, and so on.

Figure 3: Understanding the structural and stochastic features of the level-1 individual growth model



For these data points, each residual represents that part of groups *i*'s value of *Pen Rate* at *Time t* not predicted by its Year.

We can reduce the magnitude of the level-1 residuals by introducing selected predictors we believe are correlated with *Time*—this would reduce error in the form of bias in the variable *Time*. Additionally, we can introduce predictors that are correlated with *Pen Rate*, regardless if they are linked with *Time*. This additional method increases the goodness of fit in our model, typically measured by R^2.

To illustrate how confounding predictors correlated with *Time* reduces the magnitude of the level-1 residuals, ε_{it}, suppose I have the following regression estimated by OLS;

[2.1]
$$\hat{Y}_{it} = \hat{\beta}_{0i} + \hat{\beta}_1 X_{it} + \hat{\varepsilon}_{it}$$

where *X* is the predictor variable *Time* and *Y* is the *Pen Rate*. If we manipulate equation [2.1] by using covariance algebra (take covariance *X* throughout the equation), we get the following;

[2.2]
$$\begin{aligned} cov(\hat{Y}_{it}, X_{it}) &= cov(\hat{\beta}_{0i} + \hat{\beta}_1 X_{it} + \hat{\varepsilon}_{it}, X_{it}) \\ . &= \hat{\beta}_1 cov(X_{it}, X_{it}) + cov(\hat{\varepsilon}_{it}, X_{it}) \\ . &= \hat{\beta}_1 var(X_{it}) + cov(\hat{\varepsilon}_{it}, X_{it}) \\ \sigma_{yx} &= \hat{\beta}_1 \sigma_X^2 + \sigma_{\varepsilon x} \end{aligned}$$

Now divide the population variance of the question predictor *X* throughout equation [2.2];

[2.3]
$$\frac{\sigma_{yx}}{\sigma_x^2} = \hat{\beta}_1 + \frac{\sigma_{\varepsilon x}}{\sigma_x^2}$$

The $\frac{\sigma_{\varepsilon x}}{\sigma_x^2}$ ratio in equation [2.3] will give us the size of the residuals that leads to bias induced upon the $\hat{\beta}_1$ coefficient. The larger the ratio, the larger the bias. [5] The reduction in the residual term related to bias is only possible when the second term on the right-hand side of the equation is zero;

[2.4]
$$\frac{\sigma_{yx}}{\sigma_x^2} = \beta_1, \qquad when\ \sigma_{\varepsilon x} = 0$$

Now that we have a good understanding of where bias enters in our setup and increases the residual variance, it is also wise to spend time discussing how the introduction of predictors that are correlated with *Pen Rate*, regardless if they are linked with *Time*, reduces the residual variance.

The residual can be defined as;

[3.1]

$$\hat{\varepsilon}_{it} = Y_{it} - \hat{Y}_{it}$$

where,

$$\hat{Y}_{it} = \hat{\beta}_{0i} + \hat{\beta}_1 X_{it} + \hat{\beta}_2 X_{it} + \ldots + \sum_{j=1}^{k} \beta_k X_{it} + \hat{\varepsilon}_{it}$$

The total residual variance of your fitted model is calculated using the *SSR* (*Sum Squared Residual*);

[3.2]

$$SSR = \sum_{i=1}^{n} \hat{\varepsilon}_{it}$$

The total variation in *Y* can always be explained by the sum of the explained variation and the unexplained variation. We can use the R^2 to measure the explained variation compared to the total variation;

[3.3]

$$R^2 = 1 - \frac{SSR}{SSE}$$

where,

$$SSE = \sum_{i=1}^{n} (\hat{Y}_{it} - \bar{Y}_{it})^2$$

Now, to understand why the R^2 increases when you add additional *k* variables you have to recall that a regression minimizes the *SSE* by solving the *first order condition*; [10]

[3.4]

$$\min_{\beta 0, \beta 1 \ldots, \beta k} \sum_{i=1}^{n} (Y_{it} - \hat{\beta}_{0i} - \hat{\beta}_1 X_{it} - \hat{\beta}_2 X_{it} - \ldots - \hat{\beta}_k X_{it})^2$$

Equation [3.4] solves for the values of the coefficients such that the squared errors are minimized, or equivalently, for the values of the coefficients such that what you were able to explain, i.e. the R^2, is maximized. Therefore, whenever you add a variable to your model, the value of its estimated coefficient can either be zero, in which case the proportion of explained variance (R^2) stays unchanged or takes a nonzero value because it improves the quality of fit.

For example, suppose we have two regression models;

[Model 1]

$$\hat{Y}_{it} = \hat{\beta}_{0i} + \hat{\beta}_1 X_{it} + \hat{\varepsilon}_{it}$$

[Model 2]

$$\hat{Y}_{it} = \hat{\beta}_{0i} + \hat{\beta}_1 X_{it} + \hat{\beta}_2 X_{it} + \hat{\varepsilon}_{it}$$

The estimates are solved by minimizing *SSR*.

For model 1, we want to minimize *SSR*;

$$\min_{\beta 0,\beta 1} \sum_{i=1}^{n} (Y_{it} - \hat{\beta}_{0i} - \hat{\beta}_1 X_{it})^2$$

And for model 2, we want to minimize *SSR*;

$$\min_{\beta 0,\beta 1,\beta 2} \sum_{i=1}^{n} (Y_{it} - \hat{\beta}_{0i} - \hat{\beta}_1 X_{it} - \hat{\beta}_2 X_{it})^2$$

Let's say we have found the correct estimators for model 1, then you can obtain that exact same *SSR* in model 2 by choosing the same values for $\hat{\beta}_{0i}$, $\hat{\beta}_1$, and letting $\hat{\beta}_2 = 0$. Now you can find, possibly, a lower *SSR* by searching for a better value of $\hat{\beta}_2$.

To summarize, the models are nested, in the sense that everything we can model with model 1 can be matched by model 2, model 2 is more general than model 1. So, in the optimization, we have larger freedom with model 2 (there are two independent variables versus one), therefore we can always find a better solution compared to model 1.

Serial Correlation and Heteroskedasticity

Returning to equation [1.1] and fitting the level-1 submodel to data, I must invoke assumptions about the distribution of the level-1 residuals, from occasion to occasion and from group to group. OLS regression invokes "classical" assumptions: that residuals are independently and identically distributed, with homoscedastic variance across occasions and groups. This implies that regardless of group and occasion, each error is drawn independently from an underlying distribution with zero mean and an unknown residual variance. Most often, we also stipulate the form of the underlying distribution typically claiming normality. When we do, we can embody our assumptions about level-1 residuals;

[4.1] $$\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$$

However, assumptions like equation [4.1] are unlikely to hold in multilevel models for change because it is based on longitudinal data. [2] Each group's level-1 residual may be autocorrelated and heteroskedastic over time, not independent as equation [4.1] stipulates. Because the same group is measured on several occasions, any unexplained group-specific time-invariant effect in the residuals will create a correlation across occasions.

The no serial correlation assumption states that conditional on *X*, the errors in two different time periods are uncorrelated;

[4.2] $$Corr(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{is} \mid X) = 0, \text{ for all } t \neq s$$

The easiest way to think of it is to ignore the conditioning on *X*. Then, equation [4.2] is simply;

[4.3] $$Corr(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{is}) = 0, \text{ for all } t \neq s$$

When equation [4.3] is false, we say that the errors suffer from serial correlation, because they are correlated across time. Suppose that when $\hat{\varepsilon}_{t-1} > 0$ then, on average, the error in the next adjacent time period, $\hat{\varepsilon}_t$, is also positive. Resulting the $Corr(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{t-1}) > 0$, where the errors suffer from serial correlation.

So, too, the outcome may have different precision for groups at different times, perhaps being more suitable at some occasions than at others. The homoskedasticity assumption states that conditional on X, the variance of $\hat{\varepsilon}_{it}$ is the same for all t;

[4.4] $$Var\,(\hat{\varepsilon}_{it} \mid X) = Var\,(\hat{\varepsilon}_{it}) = \sigma^2,\ t = 1, 2, \ldots, n$$

Equation [4.4] means that $Var\,(\hat{\varepsilon}_{it} \mid X)$ cannot depend on X—that $\hat{\varepsilon}_{it}$ and X are independent while $\hat{\varepsilon}_{it}$ must be constant over time. When equation [4.4] does not hold, the error variance may differ over time and the level-1 residuals will be heteroskedastic over occasion within group.

Because the Gauss-Markov theorem requires both homoskedasticity and serially uncorrelated errors, and we know that longitudinal data analysis violates this assumption, then we can conclude OLS is no longer *BLUE* (*Best Linear Unbiased Estimator*) in the presence of serial correlation. [9] Even more importantly, the usual OLS standard errors and test statistics are not valid. I can see this by computing the variance of the OLS estimator under the first four Gauss-Markov assumptions and the *AR(1)* serial correlation model for the error term. More precisely I assume;

[4.5] $$\varepsilon_{it} = \rho\varepsilon_{it-1} + \varepsilon_{it}$$

For $t = 1, 2, \ldots, n$, where $|\rho| < 1$ and the ε_{it} are uncorrelated random variables with zero mean and variance σ^2_{ε}. Consider the variance of the OLS slope estimator from a simple linear regression and just to simplify the formula, I assume that the sample average of the X_{it} is zero. Then, the OLS estimator $\hat{\beta}_1$ of β_1 can be written as;

[4.6] $$\hat{\beta}_1 = \beta_1 + SST_x^{-1} \sum_{t=1}^{n} X_{it}\hat{\varepsilon}_{it}$$

Where $SST_x = \sum_{t=1}^{n} X_t^2$. Now in computing the variance of $\hat{\beta}_1$ (conditional on X), I must account for serial correlation in the ε_{it};

[4.7] $$\begin{aligned} Var(\hat{\beta}_1) &= SST_x^{-2}\, Var(\textstyle\sum_{t=1}^{n} X_{it}\varepsilon_{it}) \\ &= SST_x^{-2}\,(\textstyle\sum_{t=1}^{n} X_t^2\, Var(\varepsilon_{it}) + 2\sum_{t=1}^{n-1}\sum_{j=1}^{n-t} X_{it}X_{it+j}\,E(\varepsilon_{it}\,\varepsilon_{it+j})) \\ &= \underbrace{\frac{\sigma^2}{SST_x}}_{\text{Variance of } \hat{\beta}_1} + \underbrace{2\,(\frac{\sigma^2}{SST_x^2})\sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\rho^j X_{it}X_{it+j}}_{\text{Bias}} \end{aligned}$$

Where $\sigma^2 = Var(\varepsilon_{it})$ and $E(\varepsilon_{it}\varepsilon_{it+j}) = Cov(\hat{\varepsilon}_{it}, \hat{\varepsilon}_{it+j}) = \rho^j\sigma^2$. If I ignore the serial correlation and estimate the variance in the usual way, the variance estimator will be biased when $\rho \neq 0$ because it ignores the bias term on the right-hand side of equation [4.7]. The consequences of the bias, where $\rho > 0$, means that the usual OLS variance underestimates the true variance of the OLS. [10] Further, the independent variables in regression models are often positively correlated over time, so that $X_{it}\,X_{it+j}$ is positive for most pairs i,t and $i,t+j$. Therefore, in most economic applications, the term $\sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\rho^j X_{it}\,X_{it+j}$ is positive, and so the usual OLS variance formula $\frac{\sigma^2}{SST_x}$ understates the true variance of the OLS estimator.

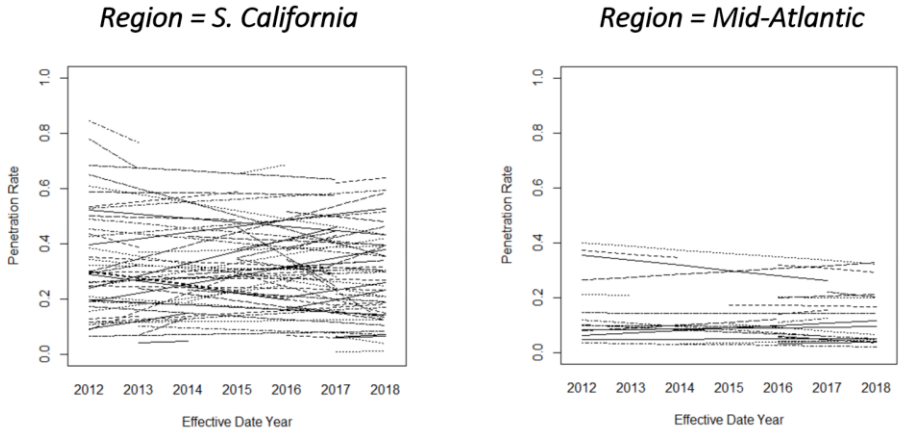
Because the standard error of $\hat{\beta}_1$ is an estimate of the standard deviation of $\hat{\beta}_1$, using the usual OLS standard error in the presences of serial correlation is invalid. Therefore, *t-statistics* are no longer valid for testing singly hypotheses. Since a smaller standard error means a larger *t-statistic*, the usual *t-statistic* will often be too large when $\rho > 0$.

The Level-2 Submodel for Systematic Interindividual Differences in Change

The level-2 submodel organizes the relationship between interindividual difference in the change trajectories and time-invariant characteristics of the individual. The ability to formulate this relationship using a level-2 submodel stems from the realization that adoption of common level-1 submodel forces people to differ only in the values of their individual growth parameters. When I use a level-1 linear change model, people can differ only in their intercepts and slopes.

Examine figure [4] below, which separately plots fitted OLS trajectories according to the groups region (groups within the Southern California region are plotted in the left panel; groups within the Mid-Atlantic region are plotted in the right panel). Groups within the Mid-Atlantic region tend to have a lower *Pen Rate* during their *initial status* and a flatter slope. Whereas groups within the Southern California region have substantial interindividual heterogeneity across their *Pen Rate*, with both high and low intercepts and slopes. The level-2 submodel must simultaneously account for both the general patterns (here, the between-group differences in intercepts and slopes) and interindividual heterogeneity in patterns within groups.

Figure 4: Identifying potential predictors of change by examining OLS fitted trajectories separately by levels of selected regional predictors



There are four specific features for the level-2 submodel that gives rise to the general patterns (in between-group differences in intercepts and slopes), and interindividual heterogeneity patterns within groups. [8]

First, its outcomes must be the individual growth parameters (from equation [1.1], π_{0i} and π_{1i}). As in regular regression, where we model the population distribution of a random variable by making it an outcome, here, where we model the population distribution of the individual growth parameters, they, too, must be the outcomes.

Second, the level-2 submodel must be written in separate parts, one for each level-1 growth parameter. When I use a linear change individual growth model at level-1 (as in equation [1.1]), I need two level-2 submodel's: one for the intercept, π_{0i}, another for the slope, π_{1i}.

Third, each part must specify a relationship between an individual growth parameter and the predictor (e.g. *Region*). As you move across the panels in figure [4], the value of the predictor, *Region*, shifts from Southern California (S. California) to Mid-Atlantic. This suggest that each level-2 model should ascribe differences in either π_{0i} or π_{1i} to *Region* just as in a regular regression model.

Fourth, each model must allow individuals who share common predictor values to vary in their individual change trajectories. This means that each level-2 submodel must allow for stochastic variation in the individual growth parameters.

These considerations lead me to postulate the following level-2 submodel for this dataset:

[5.1: Level-2 Submodel, Intercept]

$$\pi_{0i} = [\gamma_{00} + \gamma_{01}Region_i + \gamma_{02}Firm\_Size_i + \gamma_{03}Group\_Tenure_{it} + \gamma_{04}Log(Member\_Risk\_Score_{it}) + \gamma_{05}Prop\_BlackHisp_{it} + \gamma_{06}Prop\_Dependents_{it}] + [\xi_{0i}]$$

[5.1: Level-2 Submodel, Slope]

$$\pi_{1i} = [\gamma_{10} + \gamma_{11}Region_i + \gamma_{12}Firm\_Size_i + \gamma_{13}Group\_Tenure_{it} + \gamma_{14}Log(Member\_Risk\_Score_{it}) + \gamma_{15}Prop\_BlackHisp_{it} + \gamma_{16}Prop\_Dependents_{it}] + [\xi_{1i}]$$

Where, in this dataset, *Region* is a categorical variable that can equal Northern California, Southern California, Colorado, Georgia, Hawaii, Mid-Atlantic, or Northwest. *Firm_Size* is a categorical variable that can equal less than or equal to (LTE) 50 eligibles, greater than (GT) 50 but LTE 200 eligibles, GT 200 but LTE 500, GT 500 but LTE 1000, GT 1000 but LTE 3000, or GT 3000. *Group_Tenure* is a continuous variable that quantifies the annual tenure of Kaiser Permanente healthcare coverage. *Member_Risk_Score* is a continuous variable that has been log transformed to normalize the skewness of the distribution. *Member_Risk_Score* quantifies the groups health risk of members. *Prop_BlackHisp* quantifies the proportion of Black and Hispanic members within a group. *Prop_Dependents* quantifies the proportion of members with dependent coverage within a group.

Like all level-2 submodels, equation [5.1] has more than one component, each resembling a regular regression model. Taken together, the two components treat the intercept (π_{0i}) and the slope (π_{1i}) of a groups growth trajectory at level-2 outcomes that may be associated with predictors such as *Region*, or *Firm_Size*. Each component also has its own residual—here, ξ_{0i} and ξ_{1i}—that permits the level-1 parameters (the π's) of one group to differ stochastically from those of others.

Structural Components of the Level-2 Submodel

The structural parts of the level-2 submodel, (the γ's) are known as the *fixed effects*. The *fixed effects* capture systematic interindividual differences in change trajectory according to values of the level-2 predictor(s). in equation [5.1], seven of the *fixed effects* are level-2 intercepts (γ_{00}, γ_{01}, γ_{02}, γ_{03}, γ_{04}, γ_{05}, and γ_{06}); seven are level-2 slopes (γ_{10}, γ_{11}, γ_{12}, γ_{13}, γ_{14}, γ_{15}, and γ_{16}). I can interpret the level-2 parameters much as I do regular regression coefficients, except that I must remember that they describe variation in outcomes that are themselves level-1 individual growth parameters (π_{0i} and π_{1i}).

The Stochastic Components of the Level-2 Submodel

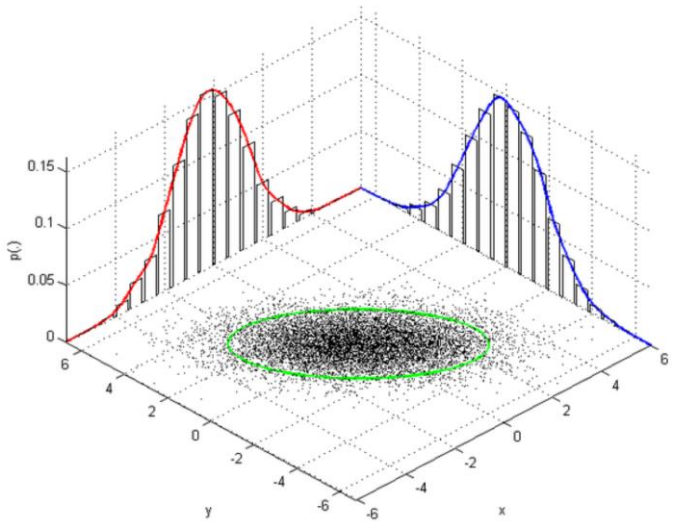
Each part of the level-2 submodel contains a residual that allows the value of each group's growth parameters to be scattered around the relevant population averages. These residuals, ξ_{0i} and ξ_{1i} in equation [5.1], represent those portions of the level-2 outcomes—the individual growth parameters—that remain unexplained by the level-2 predictor(s). As is true for most residuals, we are interested less in their specific values than in their population variance and covariance, σ_0^2, σ_1^2, and σ_{01}.

The level-2 residuals represent deviations between the individual growth parameters and their respective population averages. Their variances, σ_0^2, σ_1^2, summarize the population variation in true individual intercept and slope around these averages. Because they describe those portions of the intercept and slope left over after accounting for the effect(s) of the model's predictor(s), they are *conditional residual variances*. [1] Conditional on the presence of the model's predictors, σ_0^2 represents the population residual variance in true *initial status*, and σ_1^2 represents the population residual variance in true annual *rate of change*. These variance parameters allow me to address the question: How much heterogeneity in true change remains after accounting for the effects of the model's predictor(s).

When I posit a level-2 submodel, I also allow for a possible association between the individual *initial status* and individual *rate of change*. Groups who begin at a higher level may have higher (or lower) slopes. To account for this possibility, I permit the level-2 residuals to be correlated. Since ξ_{0i} and ξ_{1i} represent the deviations of the individual growth parameters from their population averages, their population covariance summaries the association between true individual intercepts and slopes. Again because of their conditional nature, the population covariance of the level-2 residuals, σ_{01}, summarizes the magnitude and direction of the association between true *initial status* and true annual *rate of change*, after controlling for the model's predictor(s).

To fit the multilevel model for change to data, I must make some assumptions about the level-2 residuals. But because I have two level-2 residual (ξ_{0i} and ξ_{1i}—which could also be more based on the additional level-2 outcomes, π_{0i} and π_{1i}), I describe their underlying behavior using a bivariate distribution.

Figure 5: A hypothetical example of level-2 residuals, ξ_{0i} and ξ_{1i}, that are bivariate normal with mean 0, with variances σ_0^2, and σ_1^2, and covariance σ_{01}



The standard assumption is that the two level-2 residuals, ξ_{0i} and ξ_{1i}, are bivariate normal with mean 0, with unknown variances, σ_0^2 and σ_1^2, and an unknown covariance, σ_{01}. I can express these assumptions compactly using matrix notation;

[5.2]

$$\begin{bmatrix} \xi_{0i} \\ \xi_{1i} \end{bmatrix} \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}$$

We interpret equation [5.2] in the same way we interpret the assumptions about the level-1 residuals in equation [4.1]. The first matrix on the right of the right of the tilde, specifies the bivariate distribution's mean vector; here, I assume it to be 0 for each residual. The second matrix specifies the bivariate distribution's variance-covariance matrix, also known as the level-2 error covariance matrix because it captures the covariation among the level-2 residuals. Two variances, σ_0^2 and σ_1^2, appear along the diagonal, the covariance, σ_{01}, appears on the off-diagonal. Because the covariance between ξ_{0i} and ξ_{1i}, is the same as the covariance between ξ_{1i} and ξ_{0i}, the off-diagonal elements are identical, $\sigma_{01} = \sigma_{10}$

The Composite Specification of the Multilevel Model for Change

Above, I showed that I used a pair of linked statistical models to establish the multilevel model for change (equations [1.1] and [5.1]). Within the representation, a level-1 submodel (equation [1.1]) describes how each group changes over time and a level-2 submodel (equation [5.1]) relates interindividual differences in change to predictors. The level-1/level-2 representation is not the only specification of the multilevel model for change. A more parsimonious representation arises if you collapse the level-1 and level-2 submodels together algebraically into a single *composite model*. [8]

To derive the composite specification, first notice that any pair of linked level-1 and level-2 submodels share some common terms. Specifically, the individual growth parameters of the level-1 submodel (π_{0i} and π_{1i}) are the outcomes of the level-2 submodel. I can therefore collapse the submodels together by substituting for π_{0i} and π_{1i} from the level-2 submodel into the level-1 submodel;

$$\begin{aligned} Y_{it} &= \pi_{0i} + \pi_{1i}Time_{it} + \varepsilon_{it} \\ &= [\gamma_{00} + \gamma_{01}Region_i + \gamma_{02}Firm\_Size_i + \gamma_{03}Group\_Tenure_{it} + \gamma_{04}Log(Member\_Risk\_Score_{it}) \\ &\quad + \gamma_{05}Prop\_BlackHisp_{it} + \gamma_{06}Prop\_Dependents_{it} + \xi_{0i}] + [\gamma_{10} + \gamma_{11}Region_i \\ &\quad + \gamma_{12}Firm\_Size_i + \gamma_{13}Group\_Tenure_{it} + \gamma_{14}Log(Member\_Risk\_Score_{it}) \\ &\quad + \gamma_{15}Prop\_BlackHisp_{it} + \gamma_{16}Prop\_Dependents_{it} + \xi_{1i}]Time_{it} + \varepsilon_{it} \end{aligned}$$

Multiplying out and rearranging the *structural* and *stochastic* terms then yield the composite multilevel model for change;

[6.1]

$$\begin{aligned} Y_{it} &= [\gamma_{00} + \gamma_{01}Region_i + \gamma_{02}Firm\_Size_i + \gamma_{03}Group\_Tenure_{it} + \gamma_{04}Log(Member\_Risk\_Score_{it}) \\ &\quad + \gamma_{05}Prop\_BlackHisp_{it} + \gamma_{06}Prop\_Dependents_{it} + \gamma_{10}Time_{it} + \gamma_{11}(Region_i \times Time_{it}) \\ &\quad + \gamma_{12}(Firm\_Size_i \times Time_{it}) + \gamma_{13}(Group\_Tenure_{it} \times Time_{it}) \\ &\quad + \gamma_{14}((Log(Member\_Risk\_Score_{it})) \times Time_{it}) + \gamma_{15}(Prop\_BlackHisp_{it} \times Time_{it})] \\ &\quad + \gamma_{16}(Prop\_Dependents_{it} \times Time_{it})] + [\xi_{0i} + \xi_{1i}Time_{it} + \varepsilon_{it}] \end{aligned}$$

Even though the composite specification in equation [6.1] appears more complex than the level-1 and level-2 specification, the two forms are logically and mathematically equivalent. Each posits an identical set of links between the outcome, Y_{it}, and predictors. The specifications differ only in how they organize the hypothesized relationships, each providing valuable insight into what the multilevel model represents. The advantage of the level-1 and level-2 specification is that it reflects our conceptual framework directly: we focus first on individual change and next on interindividual differences in change. It also provides an intuitive basis for interpretation because it directly identifies which parameters describe interindividual differences in *initial status* (γ_{00}, γ_{01}, γ_{02}, γ_{03}, γ_{04}, γ_{05}, γ_{06}) and which describe interindividual differences in *rate of change* (γ_{10}, γ_{11}, γ_{12}, γ_{13}, γ_{14}, γ_{15}, γ_{16}).

A distinctive feature of the composite multilevel model is its *composite residual*, the three terms in the second set of brackets on the right of equation [6.1] that combine together the level-1 residual and the two level-2 residuals;

[6.1: Composite Residual] $[\xi_{0i} + \xi_{1i}Time_{it} + \varepsilon_{it}]$

The *composite residual* is not a simple sum. Instead, the second level-2 residual, ξ_{1i}, is multiplied by the level-1 predictor, *Time*, before joining the other residual terms. Despite its unusual construction, the interpretation of the *composite residual* is straightforward: it describes the difference between the observed and the expected value of *Y* for individual *i* on occasion *t*. The mathematical form of the *composite residual* reveals two important properties about the occasion-specific residuals not readily apparent in the level-1 and level-2 specification: they can be both autocorrelated and heteroskedastic within person. [7]

When the residuals are heteroskedastic, the unexplained portions of each person's outcome have unequal variances across occasion of measurement. Although heteroscedasticity has many roots, one major cause is the effects of omitted predictors. Because their effects have nowhere else to go, they bundle together, by default, into the residuals. If their impact differs across occasions, the residual's magnitude may differ as well, creating heteroskedasticity. The *composite model* allows for heteroskedasticity via the level-2 residual ξ_{1i}. Because ξ_{1i} is multiplied by *Time* in the *composite residual*, its magnitude can differ across occasions. If there are systematic differences in the magnitudes of the *composite residual*'s across occasions, there will be accompanying differences in the residual variance.

When residuals are autocorrelated, the unexplained portions of each group's outcome are correlated with each other across repeated occasions. Once again, omitted predictors, whose effects are bundled into the residuals, are a common cause. Because their effects may be present identically in each residual over time, a group's residual my become linked across occasions. The presence of the time-invariant ξ_{0i}'s and ξ_{1i}'s in the *composite residual* of equation [6.1] allows the residuals to be autocorrelated. Because they have only an *i* subscript (and no *t*), they feature identically in each group's *composite residual* on every occasion, creating the potential for autocorrelation across time.

Estimation Procedure: First Step—Fitting the Unconditional Means Model

Now that I have went through the general assumptions and construction ideas of the multilevel model for change, I will estimate it to the Kaiser Permanente dataset and review the output of the model.

Before I am tempted to begin by fitting the full model (with all hypothesized predictors) of equation [6.1], it is recommended to begin by fitting the *Unconditional Means Model* first. Model 1 in table [1], as noted in page 26, represent the *Unconditional Means Model* fitted to the Kaiser Permanente dataset.

Instead of describing change in the outcome over time, it simply describes and partitions the outcome variation;

[7.1: Unconditional Means]

$$Penetration\_Rate_{it} = \pi_{0i} + \varepsilon_{it}$$
$$\pi_{0i} = \gamma_{00} + \xi_{0i},$$

Where I assume that;

$$\varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^2) \text{ and } \xi_{0i} \sim N(0, \sigma_0^2)$$

Notice that because there is only one level-2 residual, ξ_{0i}, I assume univariate normality at level-2 residuals (not bivariate normality, as we do when we have two level-2 residuals).

The *Unconditional Means Model* stipulates that, at level-1, the true individual change trajectory for group *i* is completely flat, sitting at elevation π_{0i}. Because the trajectory lacks a slope parameter, π_{1i}, associated with a temporal predictor, it cannot tilt. The single part of the level-2 submodel stipulates that while these flat trajectories may differ in elevation, their average elevation, across everyone in the population is γ_{00}. Any interindividual variation in elevation is not linked to predictors. It is recommended to begin with the *Unconditional Means Model* because it partitions the total variation in the outcome meaningfully.

To understand how this variance partition operates, notice that flat individual change trajectories are really just means. The true mean of *Y* for group *i* is π_{0i}; the true mean of *Y* across every group in the dataset is γ_{00}. Borrowing terminology from analysis of variance, π_{0i} is the *group-specific mean* and γ_{00} is the *grand mean*. The *Unconditional Means Model* postulates that the observed value of *Y* for group *i* on occasion *t* is composed of deviations about these means. On occasion *t*, Y_{it} deviates from group *i*'s true mean (π_{0i}) by ε_{it}. The level-1 residual is thus a "within-group" deviation that assesses the distance between Y_{it} and π_{0i}. Then, for group *i*, their true mean (π_{0i}) deviates from the population average true mean (γ_{00}) by ξ_{0i}. This level-2 residual is thus a "between-group" deviation that assesses the distance between π_{0i} and γ_{00}.

The variance components of equation [7.1] summarize the variability in these deviations across every group in the population σ_{ε}^2: is the within-group variance, the pooled scatter of each group's data around their own mean; σ_0^2 is the between-group variance, the pooled scatter of the *group-specific means* around the *grand mean*. The primary reason I fit the *Unconditional Means Model* is to estimate these variance components, which assess the amount of outcome variation that exists at each level.

Model 1 of table [1] presents the results of fitting the *Unconditional Means Model* to the Kaiser Permanente dataset. Its one *fixed effect*, $\hat{\gamma}_{00}$, estimates the outcome's grand mean across all occasions and groups within the dataset. Rejection of its associated null hypothesis ($p < .001$) confirms that the average *Pen Rate* for groups in 2012 is 27%.

Next, I examine the *random effects*, which is the major purpose for fitting the *Unconditional Means Model*. The estimated within-group variance $\hat{\sigma}_{\varepsilon}^2$, is 0.005; the estimated between-group variance $\hat{\sigma}_0^2$, is 0.033. The *Unconditional Means Model* also allows us to evaluate numerically the relative magnitude of the within-person and between-person variance components.

A useful statistic for quantifying the variance components relative magnitude is the *intraclass correlation coefficient (icc),* ρ_{icc}, which describes the proportion of the total outcome variation that lies between groups. Because the total variation in *Y* is just the sum of the of the within and between-group variance components, the population *intraclass correlation coefficient* is;

[7.2]

$$\rho_{icc} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_\varepsilon^2}$$

Therefore, substituting the two estimated variances from table [1] equals;

$$\hat{\rho}_{icc} = \frac{0.033}{0.033 + 0.005} = 0.79$$

Indicating that more than ¾ of the total variation in *Pen Rate* is driven by the differences among groups.

The *intraclass correlation coefficient* has another role as well: is summarizes the size of the residual autocorrelation in the composite *Unconditional Means Model*. [5] To understand how it does this, substitute the level-2 submodel in equation [7.1] into its level-1 submodel to yield the following composite unconditional means model;

[7.3]

$$Penetration\_Rate_{it} = \gamma_{00} + [\xi_{0i} + \varepsilon_{it}]$$

In this representation, Y_{it} is composed of one *fixed effect*, γ_{00}, and one *composite residual* $[\xi_{0i} + \varepsilon_{it}]$. Each group has a different *composite residual* on each occasion of measurement. But notice the differences in the subscripts of the pieces of the *composite residual*: while the level-1 residual ε_{it}, has two subscripts (*i* and *t*), the level-2 residual, ξ_{0i}, has only one (*i*). Each group can have a different ε_{it} on each occasion, but only has one ξ_{0i} across every occasion. The repeated presence of ξ_{0i} in group *i*'s *composite residual* links their *composite residual* across occasions. The error autocorrelation coefficient quantifies the magnitude of this linkage; in the *Unconditional Means Model*, the error autocorrelation coefficient is the *intraclass correlation coefficient*. Thus, I estimate that for each group, the average correlation between any pair of *composite residual*'s—between occasion 2012 and 2013, or 2013 and 2014, …, or 2017 and 2018—is 0.79. This is very large, and far from the zero residual autocorrelation that an OLS analysis of these data would require.

Estimation Procedure: Second Step—Fitting the Unconditional Growth Model

The next logical step is the introduction of predictor *Time* into the level-1 submodel;

[7.4: Unconditional Growth]

$$Penetration\_Rate_{it} = \pi_{0i} + \pi_{1i} Time_{it} + \varepsilon_{it}$$
$$\pi_{0i} = \gamma_{00} + \xi_{0i}$$
$$\pi_{1i} = \gamma_{10} + \xi_{1i}$$

Where I assume that;

$$\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2) \text{ and } \begin{bmatrix} \xi_{0i} \\ \xi_{1i} \end{bmatrix} \sim N \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}$$

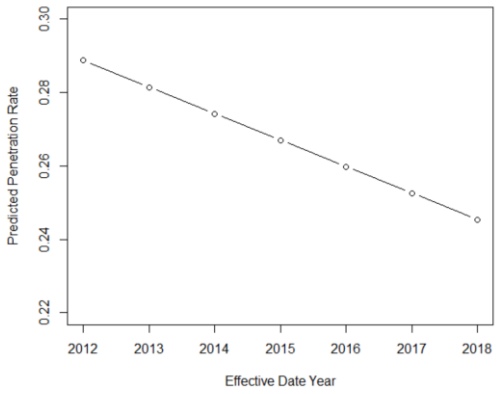
Because the only predictor in equation [7.4] is *Time* I call this the *Unconditional Growth Model*. [8]

From table [1], I can compare the *Unconditional Growth Model* (model 2) to the *Unconditional Means Model* (model 1). Instead of postulating that group *i*'s observed score on occasion *t*, Y_{it}, deviates by ε_{it} from their group-specific mean (π_{0i}), it specifies that Y_{it} deviates by ε_{it} from their true change trajectory. In other words, altering the level-1 specification alters what the level-1 residuals represent. In addition, I now have a second part to the level-2 submodel that depicts interindividual variation in the *rate of change* (π_{1i}). But, because the model includes no substantive predictors (such as equation [6.1]), each part of the level-2 submodel simply stipulates that an individual growth parameter (π_{0i} or π_{1i}) is the sum of an intercept (either γ_{00} or γ_{10}) and a level-2 residual (ξ_{0i} or ξ_{1i}).

An import consequence of altering the level-1 specification is that the meaning of the variance components changes as well. The level-1 residual variance (within-group variation), σ_ε^2, now summarizes the scatter of each group's data to their own linear change trajectory (not their own group-specific mean). The level-2 residual variances σ_0^2 and σ_1^2, now summarize between-group variability in *initial status* and *rate of change*. Estimating these variance components allows me to distinguish level-1 variation from the two different kinds of level-2 variation and to determine interindividual differences in change are due to interindividual differences in true *initial status* or *rate of change*.

Model 2 in table [1] presents the results of fitting the *Unconditional Growth Model* to the Kaiser Permanente Dataset. The *fixed effects*, $\hat{\gamma}_{00}$ and $\hat{\gamma}_{10}$, estimate the starting point and slope of the population average change trajectory. I can reject the null hypothesis for each ($p < .001$), estimating that the average true change trajectory for *Pen Rate* has a non-zero intercept of 0.29 and a non-zero slope of -0.01. Because there are no level-2 predictors, it is simple to plot this trajectory (figure [6]).

Figure 6: Displaying the results of the fitted unconditional growth multilevel model for change



If the true change trajectory is linear with *Time*, the *Unconditional Growth Model* will do a better job of predicting the observed outcome data than the *Unconditional Means Model*, resulting in a smaller level-1 residuals and a smaller level-1 residual variance. Comparing $\hat{\sigma}_\varepsilon^2$ in model 2 to that of model 1, we find a decline of 0.80 (from 0.005 to 0.001). Therefore, I can conclude that 80% of the within-group variation in *Pen Rate* is systematically associated with linear *Time*.

The level-2 variance components (σ_0^2 and σ_1^2) quantify the amount of unpredicted variation in the individual growth parameters (π_{0i} and π_{1i}). σ_0^2 assesses the unpredicted variability in the true *initial status* (the scatter of the π_{0i} around γ_{00}); σ_1^2 assesses the unpredicted variability in true *rates of change* (the scatter of the π_{1i} around γ_{10}). The variance components—0.054 and 0.001—will provide

benchmarks for quantifying the predictors' effects. Note, I do not compare these variance components with estimates from the *Unconditional Means Model* because introduction of *Time* into the model changes their interpretation. [8] The correlation coefficient, ρ_{01}, in table [1] model 2, assesses the relationship between the level-2 residuals—it quantifies the relations between the *initial status* and *rate of change*. This means that I can assess whether groups who have a large *Pen Rate* during their *initial status* increases (or decreases over time).

We can learn more about the residuals in the *Unconditional Growth Model* by examining the composite specification of the multilevel model;

[7.5] $$Penetration\_Rate_{it} = \gamma_{00} + \gamma_{10}Time_{it} + [\xi_{0i} + \xi_{1i}Time_{it} + \varepsilon_{it}]$$

Each group has *t composite residual*, one per occasion of measurement. The structure of the *composite residual*, which combines the original level-1 residual and level-2 residual (with ξ_{1i} multiplied by *Time* before being bundled into the sum), provides the anticipated heteroskedasticity and autocorrelation that longitudinal data analysis may demand.

First, I examine the variances of the *composite residual*. I can write the population variance of the *composite residual* on the *t'*th occasion of measurement as;

[7.6: Heteroskedasticity] $$\sigma^2_{Residual_t} = \sigma_0^2 + \sigma_1^2 Time^2{}_t + 2\sigma_{01}Time_t + \sigma_\varepsilon^2$$

Substituting the estimated variance components from model 2 in table [1];

$$\sigma^2_{Residual_t} = (0.054 + 0.001Time^2{}_t + 2(-0.005)Time_t + 0.001)$$

Substituting values for *Time* at effective date year 2012 (*Time = 0*), effective date year 2013 (*Time = 1*), effective date year 2014 (*Time = 2*), effective date year 2015 (*Time = 3*), effective date year 2016 (*Time = 4*), effective date year 2017 (*Time = 5*), and effective date year 2018 (*Time = 6*), I find estimated *composite residual* variances of 0.055, 0.046, 0.039, 0.034, 0.031, 0.030, 0.031, respectively. These residual variances do not seem heteroskedastic. Effective date years 2012 through 2014 could however be argued that it is beyond the bland homoskedasticity we assume of residuals in cross-sectional data.

Estimation Procedure: Fitting the Final Model

In this final model, I am going to see how individual trajectories of *Pen Rate* differ according to the groups, *Region*, *Firm_Size*, *Group_Tenure*, *Prop_BlackHisp*, *Prop_Dependents*, and *Member_Risk_Score*. The multilevel model I will be estimating is based on equation [6.1] or equivalently equation [1.1] and [5.1]. Both are the same but specified differently. As mentioned earlier, equation [6.1] is a more parsimonious representation. It collapses the level-1 and level-2 submodels together algebraically into a single *composite model*. Whereas equation [1.1] is the level-1 submodel and equation [5.1] is the level-2 submodel. I will be referring to both specifications while examining the final model.

Figure [7], displays the *Pen Rate* trajectories by *Region*. The *fixed effects*, γ, describe the average change trajectories for groups distinguished by their level-2 predictor values. For example, in figure [7], Mid-Atlantic's (MA) change trajectory is determined by its intercept (γ_{00}, γ_{05}, and γ_{09}) and slope (γ_{10}, γ_{15}, and γ_{19}). The intercept and slope are displayed in table [1]. Thus, MA's intercept equals 0.134 (0.401 - 0.180 - 0.087) and its slope equals -0.008 (-0.009 - 0.004 + 0.005). Whereas HI's intercept (γ_{00}, γ_{04}, and γ_{09})

equals 0.264 (0.401 - 0.050 - 0.087) and its slope (γ_{10}, γ_{14}, and γ_{19}) equals +0.001 (-0.009 + 0.005 + 0.005). I can interpret MA's average *initial status* (for all group *i*'s in the MA region) lower than HI's *initial status*. Additionally, MA's *rate of change* declines year-over-year whereas HI's increases slightly year-over-year. Note, not all *fixed effects*, γ, used in this example are statistically significant. This example was used to illustrate calculating differences between regions *initial status* and *rate of change*.

Figure 7: Displaying the results of the change trajectories by region

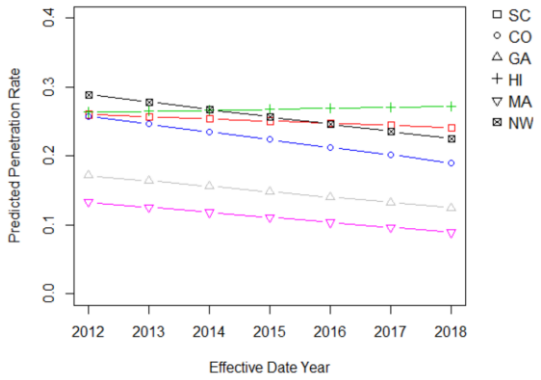
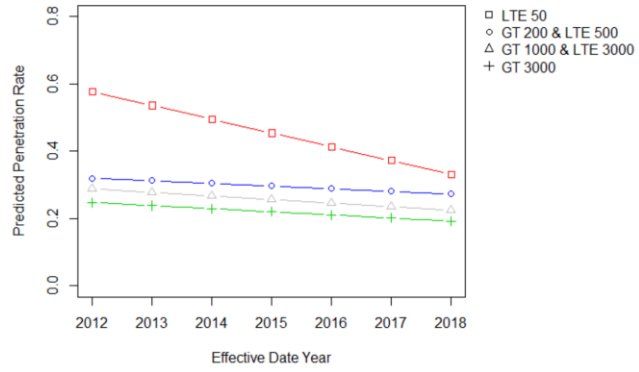


Figure [8] displays the *Pen Rate* trajectories by *Firm_Size*. The *fixed effects*, γ, describe the average change trajectories for groups distinguished by their level-2 predictor values. I can evaluate the change trajectories for each firm size similarly as I did in the regional example for figure [7] above.

Figure 8: Displaying the results of the change trajectories by firm size where region equals Northwest



Before proceeding to the evaluation of the individual growth parameters (π_{0i} and π_{1i}), I feel it is wise to pause and think of multilevel modeling as a regression that includes a categorical input variable representing an individual group. From this perspective, the group index is a factor with *J* levels, corresponding to *J* predictors in the regression model (or *2J* if they interacted with a predictor *X* in a varying intercept and varying slope model. Equation [5.1] clearly illustrates this—both π_{0i} and π_{1i} are outcome variables for each group *i*). Multilevel models typically have so many parameters (this multilevel model has 787+ parameters—one for each group) that it is not feasible to closely examine all their numerical estimates. Instead I will take a random sample and examine change trajectories within a group (estimating a π_{0i} and π_{1i} for each individual group *i*).

Let's begin first rearranging both level-1/level-2 submodels from equation [5.1];

[Level-2 Submodel, Intercept]

$$\pi_{0i} = [\gamma_{00} + \xi_{0i}] + \gamma_{01}Region_i + \gamma_{02}Firm\_Size_i + \gamma_{03}Group\_Tenure_{it} + \gamma_{04}Log(Member\_Risk\_Score_{it}) + \gamma_{05}Prop\_BlackHisp_{it} + \gamma_{06}Prop\_Dependents_{it}]$$

[Level-2 Submodel, Slope]

$$\pi_{1i} = [\gamma_{10} + \xi_{1i}] + \gamma_{11}Region_i + \gamma_{12}Firm\_Size_i + \gamma_{13}Group\_Tenure_{it} + \gamma_{14}Log(Member\_Risk\_Score_{it}) + \gamma_{15}Prop\_BlackHisp_{it} + \gamma_{16}Prop\_Dependents_{it}]$$

I have rearranged terms where the intercepts γ_{00} and γ_{10}, have been coupled with its own group *i* residual (ξ_{0i} and ξ_{1i}). Recall that each ξ_{0i} and ξ_{1i}, permits the level-1 parameters, π_{0i} and π_{1i}, of group *i* to differ stochastically from those of other *i*'s. Another way put is that each *fixed effect*, γ_{00} and γ_{10}, is coupled with an individual *random effect* (hence where the name *mixed effect* comes from) that allows each group *i* to deviate from the average intercept. The *random effect* tells us how much the intercept is shifted up or down in a particular group.

I can easily extract the estimated *random effect* intercept and slope terms for each group using most statistical software (I used R programming);

Figure 9: Displaying a sample of the estimated random effect for each group (where intercept equals ξ_{0i} and slope equals ξ_{1i})

	region_account_number	intercept	slope
1	CO_1000	0.27803600	-0.030714323
2	CO_10650	-0.03463230	-0.016241604
3	CO_1071	0.15011018	-0.004589831
4	CO_1095	0.25560047	-0.025977470
5	CO_11551	0.01081298	-0.004351499

Thus, for example, group (*region_account_number*) CO_1000, the estimated intercept is ~0.278 higher than the average, so that the regression line intercept is 0.679 (0.401 + 0.278);

[Level-2 Submodel, Intercept]

$$\pi_{0,CO\_1000} = [0.401 + 0.278] + \gamma_{01}Region_{CO\_1000} + \gamma_{02}Firm\_Size_{CO\_1000} + \gamma_{03}Group\_Tenure_{CO\_1000,t} + \gamma_{04}Log(Member\_Risk\_Score_{CO\_1000,t}) + \gamma_{05}Prop\_BlackHisp_{CO\_1000,t} + \gamma_{06}Prop\_Dependents_{CO\_1000,t}]$$

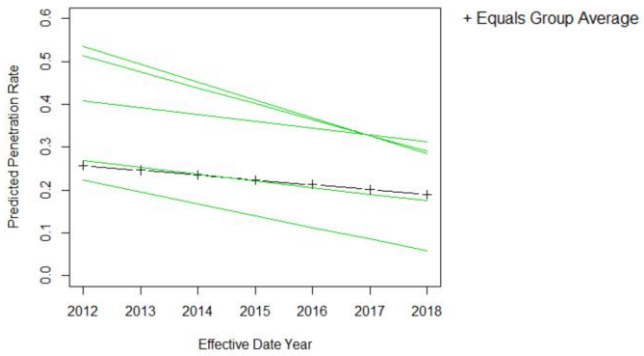
The estimated slope is ~0.031 lower than the average, therefore the slope is -0.040 (-0.009 - 0.031);

[Level-2 Submodel, Slope]

$$\pi_{1,CO\_1000} = [-0.009 - 0.031] + \gamma_{11}Region_{CO\_1000} + \gamma_{12}Firm\_Size_{CO\_1000} + \gamma_{13}Group\_Tenure_{CO\_1000,t} + \gamma_{14}Log(Member\_Risk\_Score_{CO\_1000,t}) + \gamma_{15}Prop\_BlackHisp_{CO\_1000} + \gamma_{16}Prop\_Dependents_{CO\_1000,t}]$$

I can repeat the above example to estimate all 787 group *i*'s individual intercept and slope. Below, in figure [10], I have plotted the intercept and slope of five groups (displayed in figure [9]) compared against its regional (Colorado) group average. The first thing that might seem obvious is the interindividual differences between the intercepts and slope.

Figure 10: Displaying Colorado's group average (+) change trajectory against individual group change trajectories from groups listed in figure [9]



Counterfactuals

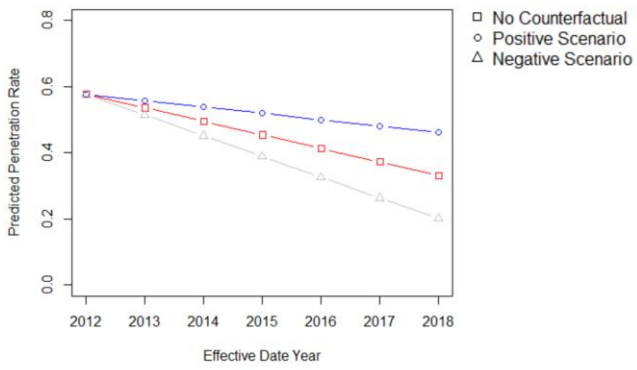
In conducting a multilevel analysis for change, my central objective is to demonstrate how multilevel models allow me to address within-individual and between-individual questions about change simultaneously. However, it is also of my interest to understand how individual change trajectories would differ if some groups were exposed to an intervention.

The condition to which the groups would have been exposed in the absence of the experimental treatment is called the *counterfactual*. From a theoretical standpoint, the way to obtain an ideal counterfactual would be to use the same participants under both a treatment and a control condition, resetting all internal and external conditions to their identical initial values before participants experienced either condition. [4]

So, you might have a representative sample of groups from the population, administer the treatment to them, and measure their outcome afterwards. Then, to learn what the outcomes would be under the *counterfactual* condition, you would need to transport these same groups back to a time before your research was conducted, erase all their experiences of the treatment and the outcome measurement from their memories, and measure their values of the outcome again, after they had transpired under the control condition. If this were possible, it could be argued convincingly that any difference in each group's outcome values between the two conditions must be only to their experiences of the treatment.

In figure [11], I have plotted *counterfactual* change trajectories where groups proportion of Black/Hispanic members increases/decreases, and groups proportion of Dependents members increases/decreases while *holding other factors fixed*. The change trajectories noticeably differ between the negative, positive, and no *counterfactual* change trajectory. As mentioned earlier, deriving a causal interpretation of the outcomes positive or negative *counterfactual* is challenging. The assumption, *holding other factors fixed* is also a difficult assumption to hold—hence my reasonings in mentioning it prior to jumping into evaluating the differences in negative, positive, and no *counterfactual* change trajectories below.

Figure 11: Displaying the results of the counterfactual change trajectories where region equals Northwest, firm size equals less than 50. The two scenario counterfactuals are (1) Positive Scenario: Both Prop Black/Hispanics and Prop of Dependents increases 1.5 times annually. (2) Negative Scenario: Both Prop Black/Hispanics and Prop of Dependents decreases by 1.5 times annually.



Evaluating the Tenability of the Model Assumptions

Whenever you fit a statistical model, you invoke assumptions. For example, I assume that the errors are independent and normally distributed with constant variance. Assumptions allows us to move forward, estimate parameters, interpret results, and test hypotheses. But the validity of our conclusions rests on our assumptions' tenability. Fitting a model with untenable assumptions is as senseless as fitting a model to data that are knowingly flawed. Violations lead to biased estimates, incorrect standard errors, and erroneous inferences. [1]

When I fit a multilevel model for change, I also invoke assumptions. And because the model is more complex, its assumptions are more complex as well, involving both *structural* and *stochastic* features at each level. The *structural* specification embodies assumptions about the true functional form of the relationship between outcome and predictors. At level-1, I specify the shape of the hypothesized individual change trajectory, declaring it to be linear (as I have assumed so far) or non-linear. At level-2, I specify the relationship between each individual growth parameter and time invariant predictors. The *stochastic* specification embodies assumptions about the level's outcome (either Y_{it} at level-1 or π_{0i} and π_{1i} at level-2) that remains unexplained by the model's predictors. Because I know neither their nature nor value, I make assumptions about these error distributions, typically assuming univariate normality at level-1 and bivariate normality at level-2.

The first assumption I will examine is the functional form assumption in the multilevel model for change. The ideal way of inspecting it is to examine the outcome versus the predictors. I have plotted this earlier in figure [2], page 4. Where I had used empirical plots of a random sample of groups. The linear change trajectory does seem reasonable, however there are some obvious groups that might have been better suited with a polynomial trajectory (e.g. NW_16238, SC_103896, and NC_39076).

The next assumption to examine is normality. Most multilevel modeling packages can output estimates of the level-1 and level-2 errors, ε_{it}, ξ_{0i} and ξ_{1i}. I can visually inspect the residual distribution to check for normality. For each estimated residual—the one at level-1 and the two at level-2—I will examine a normal probability plot, a plot of their value against their associated normal scores. If the distribution is normal, the points will form a line. Any departure from linearity indicates a departure from normality. Figures [12], [13], and [14] allows us to visually inspect the normality assumption.

Figure 12: Examining normality assumptions in the multilevel model for change. The plot below, is for the level-1 residuals $\hat{\varepsilon}_{it}$, which indicates a departure from normality

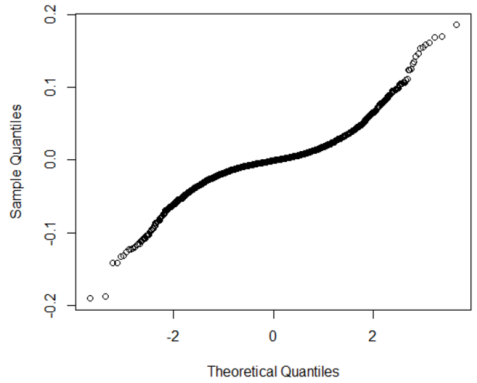


Figure 13: Examining normality assumptions in the multilevel model for change. The plot below, is for the first level-2 residuals, $\hat{\xi}_{0i}$ (intercept). This distribution does seem normal

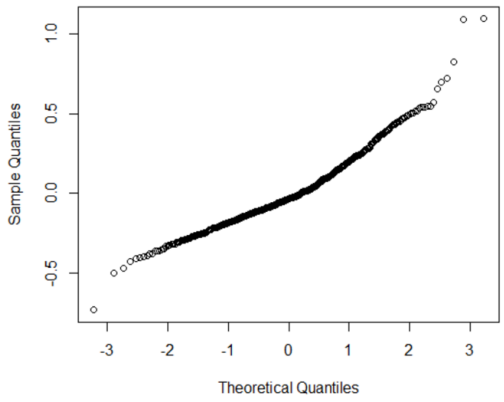
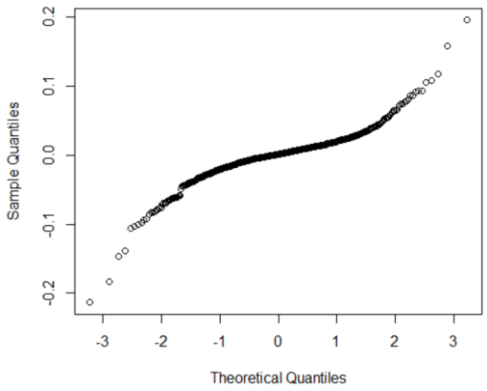
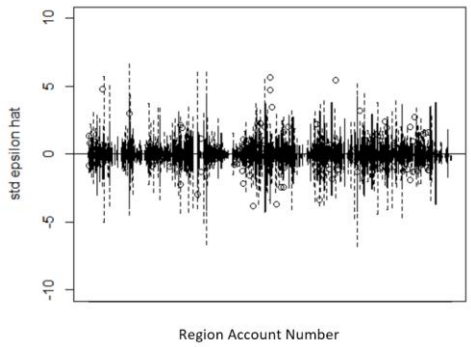


Figure 14: Examining normality assumptions in the multilevel model for change. The plot below, is for the first level-2 residuals, $\hat{\xi}_{1i}$ (slope), which indicates a departure from normality



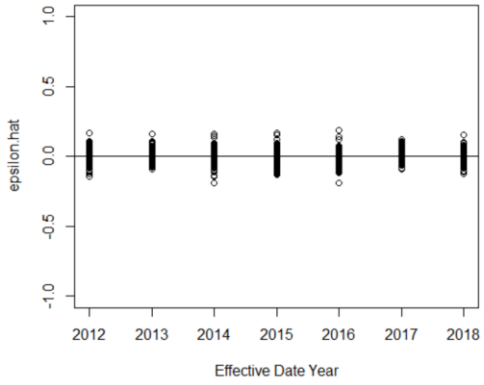
I can also plot the standardized residuals by *region_account_number* to identify extreme outlier groups. Although the majoriy of the groups fall within 2 standard deviations of center, there are several that fall between 2 and 5 standard deviations.

Figure 15: Identifing extreme groups by plotting the standardized residuals by *region_account_number*



The final assumption that I would like to examine is checking for homoskedasticity. I can examine the homoskedasticity assumption by plotting raw resdiuals against predictors. If the assumption holds, residaul variability will be approximately equal at every predictor value. Figure [16] reveals that the level-1 residauls, have approximately equal range and variability across the *effective_date_year* (*Time*).

Figure 16: Examining the homoskedasticiy assumption in the multilevel model for change



Evaluating the Accuracy of the Estimated Models

Forecasts are uncertain and since the multilevel model for change can be used as a forecast for the years outside of the span of *Time*, it is prudent to measure the forecast accuracy. I will be using the *Root Mean Squared Forecast Error* (*RMSFE*) statistic to measure the forecast accuracy of the *Unconditional Means Model*, *Unconditional Growth Model*, and the *Final Model*;

[8.1: RMSFE] $$RMSFE = ([E(Y_{Pen\ Rate} - \hat{Y}_{Pen\ Rate})^2]^{1/2})$$

The interpretation of the *RMSFE* is straightforward—it is a measure of the differences between values forecasted by a model and the values observed. *RMSFE* is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower *RMSFE* is better

than a higher one. In addition to using the *RMSFE* statistic, I will be plotting (figure [17]) the fitted values of *Pen Rate* against the actual values of *Pen Rate*. This will provide me with a visual inspection of the forecast accuracy.

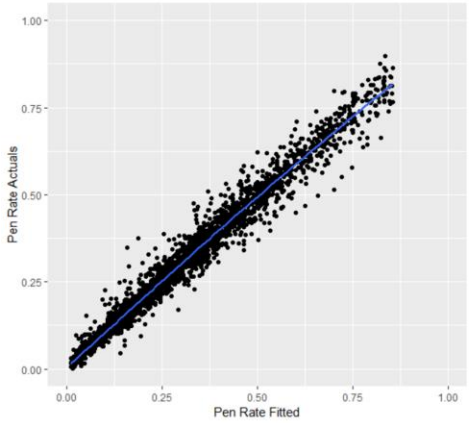
[RMSFE: Unconditional Means] $RMSFE = .039$

[RMSFE: Unconditional Growth] $RMSFE = .018$

[RMSFE: Final Model] $RMSFE = .018$

The *RMSFE* of the *Unconditional Means Model* improved over 50%, while the *Final Model RMSFE* stayed flat (note, I coupled these statistics with the *AIC* statistic before disregarding models).

Figure 17: A visual inspection of the forecast accuracy



Conclusion and Final Thoughts

In conclusion, multilevel models are extremely advantageous when modeling longitudinal or hierarchical data. They allow us to account for serial correlation and heteroskedasticity. Additionally, multilevel models can answer questions surrounding between-group and within-groups variations. While fitting the multilevel model for change, draw your attention on the random or variance components. These variance components will help us determine the performance of the fitted models. Finally, don't forget to use empirical plots to visually inspect the structural forms of the level-1 submodels.

It is worth noting that the final model could be improved if I added additional level-2 outcomes other than π_{0i} and π_{1i} (π_{2i}, π_{3i}, …, π_{ki}). This would be based on the notion that there could be strong covariance amongst the level-2 parameters. Additionally, treating time more flexibly and modeling discontinuous and nonlinear change on the slope parameter (π_{1i}) might have resolved the non-normality results in figures [12] and [14].

Table 1: Results of fitting a taxonomy of multilevel models for change to the Kaiser Permanente dataset (*n* = 3,854, *i* = 787. Coefficients in **Bold** are statistically significant at a *p-value* of 10% or less)

		Parameter	Model 1	Model 2	Model 3
Fixed Effects					
Initial Status, π_{0i}	Intercept	$\gamma_{0,0}$	**0.267**	**0.289**	**0.401**
			0.007	0.009	0.019
Region	Southern California	$\gamma_{0,1}$			**-0.053**
					0.026
	Colorado	$\gamma_{0,2}$			**-0.056**
					0.029
	Georgia	$\gamma_{0,3}$			**-0.142**
					0.030
	Hawaii	$\gamma_{0,4}$			-0.050
					0.038
	Mid-Atlantic	$\gamma_{0,5}$			**-0.180**
					0.030
	Northwest	$\gamma_{0,6}$			-0.025
					0.032
Firm Size	LTE 50	$\gamma_{0,7}$			**0.201**
					0.048
	GT 200 & LTE 500	$\gamma_{0,8}$			**-0.056**
					0.016
	GT 1000 & LTE 3000	$\gamma_{0,9}$			**-0.087**
					0.018
	GT 3000	$\gamma_{0,10}$			**-0.128**
					0.024
	Group Tenure	$\gamma_{0,11}$			**0.001**
					0.000
	Log (Risk Score)	$\gamma_{0,12}$			0.006
					0.007
	Prop Black/Hispanic	$\gamma_{0,13}$			**0.157**
					0.060
	Prop Dependent	$\gamma_{0,14}$			**-0.098**
					0.040
Rate of Change, π_{1i}	Intercept	$\gamma_{1,0}$		**-0.007**	**-0.009**
				0.001	0.003
Region	Southern California	$\gamma_{1,1}$			0.000
					0.005
	Colorado	$\gamma_{1,2}$			**-0.008**
					0.005
	Georgia	$\gamma_{1,3}$			-0.005
					0.005
	Hawaii	$\gamma_{1,4}$			0.005
					0.006
	Mid-Atlantic	$\gamma_{1,5}$			-0.004
					0.005
	Northwest	$\gamma_{1,6}$			-0.007
					0.006
Firm Size	LTE 50	$\gamma_{1,7}$			**-0.025**
					0.008
	GT 200 & LTE 500	$\gamma_{1,8}$			**0.008**
					0.003
	GT 1000 & LTE 3000	$\gamma_{1,9}$			**0.005**
					0.003
	GT 3000	$\gamma_{1,10}$			**0.007**
					0.004
	Group Tenure	$\gamma_{1,11}$			**0.000**
					0.000
	Log (Risk Score)	$\gamma_{1,12}$			**-0.004**
					0.002
	Prop Black/Hispanic	$\gamma_{1,13}$			0.004
					0.013
	Prop Dependent	$\gamma_{1,14}$			0.011
					0.010
Variance Components					
Level 1	Within-Group	σ^2_ε	0.005	0.001	0.001
Level 2	In Initial Status	σ^2_0	0.033	0.054	0.047
	In Rate of Change	σ^2_1		0.001	0.001
	Correlation	ρ_{01}		-0.62	-0.65
Goodness of Fit	Psuedo R^2			75%	1%
	Deviance		-6838.4	-9135.6	-9363.8
	AIC		-6832.4	-9123.6	-9295.8

References

[1] Fox, John. *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[2] Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[3] Luke, Douglas A. *Multilevel modeling*. Vol. 143. Sage, 2004.

[4] Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference*. Cambridge [2] University Press, 2015.

[5] Murnane, Richard J., and John B. Willett. *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press, 2010.

[6] Pickup, Mark. *Introduction to time series analysis*. Vol. 174. Sage Publications, 2014.

[7] Rabe-Hesketh, Sophia, and Anders Skrondal. *Multilevel and longitudinal modeling using Stata*. STATA press, 2008.

[8] Singer, Judith D., John B. Willett, and John B. Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.

[9] Shumway, Robert H., and David S. Stoffer. "Time series regression and exploratory data analysis." *Time series analysis and its applications*. Springer New York, 2011. 47-82.

[10] Wooldridge, Jeffrey M. *Econometric analysis of cross section and panel data*. MIT press, 2010.