Multilevel Linear Models in Python Statsmodels

February 24, 2020

1 Introduction

Multilevel regression is a form of regression analysis, meaning that the goal is to relate one *dependent variable* (also known as the outcome or response) to one or more *independent variables* (known as predictors, covariates, or regressors). Multilevel models are used when there may be statistical dependence among the observations. More basic regression procedures such as least squares regression and generalized linear models (GLM) take the observations to be independent (or at least uncorrelated) with each other. Although it is sometimes possible to use OLS or GLM with dependent data, usually an alternative approach that explicitly accounts for any statistical dependence in the data is a better choice

One important thing to keep in mind when working with multilevel models is that you cannot take an arbitrary dataset and learn its dependence structure completely from the data alone. The structure of the dependence will usually be a function of the way in which the data were collected. We will discuss this further below.

Terminology

The following terms are mostly equivalent: mixed model, mixed effects model, multilevel model, hierarchical model, random effects model, variance components model.

Alternatives and related approaches

Here we focus on using mixed linear models to capture conditional mean relationships and statistical dependence among observed data values. Other analytic approaches with related goals include generalized least squares (GLS), generalized estimating equations (GEE), fixed effects regression, and various forms of marginal or panel regression.

Nonlinear mixed models

Here we only consider linear mixed models. Generalized linear mixed models ("GLIMMIX") and non-linear mixed effects models also exist, but are not covered here.

1.1 Mean and variance structure

Many regression approaches can be interpreted in terms of the way that they specify the *mean structure* and the *variance/covariance structure* of the population being modeled. The mean structure can be written as $E[Y|X = x]$, read as "the mean of Y given that X is equal to x". For example, if your dependent variable is a person's income, and the predictors are their age, number of years of schooling, and gender, you might model the mean structure as

$$E[\text{income} \mid \text{age}, \text{school}, \text{female}] = b_0 + b_1 \cdot \text{age} + b_2 \cdot \text{school} + b_3 \cdot \text{female}.$$

The *mean structure parameters* b_0, b_1, b_2, and b_3 are unknown constants to be estimated using the data, while income, age, education, and gender are observed data values for each case or observation. This is a *linear mean structure*, which is the mean structure used in linear regression (e.g. OLS), and in linear multilevel models. The term "linear" here refers to the fact that the mean structure is linear in the parameters (b_0, b_1, b_2, b_3). Note that it is not necessary for the mean structure to be linear in the covariates. For example, we would still have a linear model if we had specified the mean structure as

$$E[\text{income} \mid \text{age}, \text{school}, \text{female}] = b_0 + b_1 \cdot \text{age} + b_2 \cdot \text{age}^2 + b_3 \cdot \text{school} + b_4 \cdot \text{female}.$$

The variance structure can be written as $Var[Y|X = x]$, and is read as "the variance of Y given that X is equal to x". A very basic variance structure is a constant or *homoscedastic* variance structure. For the income analysis discussed above, this would mean that

$$\text{Var}[\text{income} \mid \text{age}, \text{school}, \text{female}] = \sigma^2,$$

where σ^2 is an unkown constant to be estimated from the data. We will see some non-constant variance structures below.

In the context of multilevel models, the mean and variance structures are often referred to as the *marginal mean structure* and *marginal variance structure*, for reasons that will be explained further below.

1.2 Dependent data

A common situation in applied research is that several observations are obtained for each person in a sample. These might be replicates of the same measurement taken at alsmost the same point in time (e.g. triplicate blood pressure measurements), longitudinal measurements of the same trait taken over time (e.g. annual BMI measurements taken over several years), or related traits measured at the same or different times (e.g. hearing levels in the left and right ear). When data are collected this way, it is likely that the measures taken within a single person are correlated with each other.

Dependent data also arise if we have multiple levels of sampling in our data collection process, even if there are no longitudinal repeated measures. For example, we may have test scores on students in a classroom, with the classroom nested in a school, which in turn is nested in a school district, and so on. In this case, the students in one classroom (or school, etc.) may tend to score higher, or lower than students in other classrooms or schools. This constitutes a form of statistical dependence. The generic terms *cluster variable* or *grouping variable* are often used to refer to such groups.

Note that the statistical dependence discussed here can be seen as arising due to the way that the data were collected. Measurements of blood pressure, or any other trait, can be statistically independent or statistically dependent,

depending on the way the data were collected. In most cases, a cross sectional design, in which data are collected on one occasion for each subject, will be less likely to exhibit statistical dependencies among the data values, while longitudinal and cluster designs will be more likely to exhibit such dependencies. The way in which multilevel regression is used to analyze a data set is heavily dependent on understanding the way in which the data were collected.

1.3 Longitudinal data and random coefficients

There are various ways to accommodate statistical dependence in a regression analysis. One approach is to think in terms of "varying coefficients". In basic regression, the model coefficients, e.g. the coefficients b_j in the mean structure $E[Y|X = x] = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$ are constants. In multilevel modeling, these coefficients are taken to vary from one cluster (e.g. person) to the next. For example, if we have repeated measures of blood pressure and age over time within a person, we can regress blood pressure on age using a *conditionally linear mean structure* in which the intercept and slope are treated as subject-specific. That is, for subject i we have

$$E[\text{SBP}(\text{age}, i) \mid a_i, b_i] = a + b \cdot \text{age}_i + a_i + b_i \cdot \text{age}_i$$

This model states that for each subject, the conditional mean blood pressure (SBP) varies linearly with age, but each subject has their own intercept a_i and slope b_i. These subject-specific coefficients modify the population parameters a and b, i.e. the overall intercept for subject i is $a + a_i$, and the overall slope is $b + b_i$.

The model above is expressed in conditional mean form. Alternatively, it may be expressed as a fully generative model

$$\text{SBP}(\text{age}, i) = a + b \cdot \text{age}_i + a_i + b_i \cdot \text{age}_i + e(\text{age}, i),$$

where $e(\text{age}, i)$ are independent "errors", which represent unexplained or unstructured variation.

To simplify interpretation, suppose that age is coded as years beyond age 20. Then the varying intercepts a_i represent variation between people in their blood pressure at age 20, and the varying slopes b_i represent variation in the rate at which blood pressure changes with respect to age.

In a mixed model, the values of a_i and b_i are treated as random values. Each of these terms has a variance, e.g. $\mathrm{Var}[a_i] = \tau_a^2$, and $\mathrm{Var}[b_i] = \tau_b^2$. Also, there is a covariance $\tau_{ab} = \mathrm{Cov}[a_i, b_i]$ between the coefficients.

Here we provide a simple numerical example. Suppose the marginal slope is $b = 0.6$, meaning that an average person's blood pressure increases 6 units (mm Hg) per decade. If $\tau_b = 0.5$, this means that everyone has their own slope, with around 16% of the population having a slope greater than 0.6+0.5 = 1.1, and 16% of the population having a slope less than 0.6-0.5 = 0.1 (the random effects are taken to be Gaussian, and we use here the fact that 16% of a Gaussian population falls more than 1 standard deviation from the mean in a particular direction).

Note that if we did not have repeated measures on the same subjects, it would not be possible to separate the variation in the within-subject slopes and intercepts (variation in the a_i and variation in the b_i) from "between subject" variation, driven by differences between people.

1.4 Marginal covariance structure for longitudinal models

Above we expressed a linear mixed model using conditional equations that relate observed and unobserved values (latent variables or random effects). An alternative way to express a linear mixed model is in terms of its marginal mean and variance structure, $\mathrm{E}[Y|X]$ and $\mathrm{Var}[Y|X]$, as defined above. We can convert the conditional equations to marginal moments with some simple calculations.

Suppose we observe longitudinal data with three time points per person, taken at the same three time points. If we model these values as above, we have

$$Y_{it} = a + b \cdot t + a_i + b_i \cdot t + e_{it}$$

where $t = 1, 2, 3$. We can directly calculate the mean struture as

$$Y_{it} = a + b \cdot t$$

The variance structure is

$$\text{Var}[Y_{it}] = \tau_a^2 + \tau_b^2 \cdot t^2 + 2\tau_{ab} \cdot t + \sigma^2,$$

where $\sigma^2 = Var[e(i,t)]$, and the covariance between two different time points is

$$\text{Cov}[Y_{is}, Y_{it}] = \tau_a^2 + \tau_{ab}(s+t) + \tau_b^2 \cdot s \cdot t.$$

If the three time points are coded $t = -1, 0, 1$, we can more explicitly write the marginal mean as

$$\begin{pmatrix} a-b \\ a \\ a+b \end{pmatrix}$$

and the marginal covariance as

$$\begin{pmatrix} \tau_a^2 + \tau_b^2 - 2\tau_{ab} + \sigma^2 & \tau_a^2 - \tau_{ab} & \tau_a^2 - \tau_b^2 \\ \tau_a^2 - \tau_{ab} & \tau_a^2 + \sigma^2 & \tau_a^2 + \tau_{ab} \\ \tau_a^2 - \tau_b^2 & \tau_a^2 + \tau_{ab} & \tau_a^2 + \tau_b^2 + 2\tau_{ab} + \sigma^2 \end{pmatrix}.$$

1.5 Nested variance components

Above we focused on a longitudinal setting in which the repeated measures are predicted by a quantitative variable (age) and are taken over time within each subject. A different setting is when the repeated measures are taken for various grouping variables that may be nested or crossed. These are often described as *variance components*, but can also be called *random intercepts* or simply *random effects*.

A simple example would be if we had data on body mass index (BMI) for subjects where we also know their residential location in terms of neighborhood, city, and state. The neighborhoods are nested in the cities, and the cities are nested in the states. It would be possible to approach this analysis using *fixed effects regression*, in which we allocate a parameter to each clustering unit (e.g. to each neighborhood). This can be useful, but is subject to the "Neyman-Scott" problem, in which allocating too many parameters leads to inconsistency.

Rather than estimating a large number of fixed effects parameters, we can focus instead on estimating the variance contributed by each level of the nesting. A model for these data could be

$$Y_i = \mu + N[n_i] + C[c_i] + S[s_i] + e_i$$

where Y_i is the BMI for subject i, μ is the population mean, n_i, c_i, and s_i are nested geographical regions, for example, the neighborhood, city, and state where subject i lives, and $N[\cdot]$, $C[\cdot]$, and $S[\cdot]$ are the random effects for each of these levels. For example, suppose that we are studying people in Mexico, and subject i lives in the San Pedro district of Monterrey city, which is in Nueva Leon state. Then n_i = San Pedro, c_i = Monterrey, and s_i = Nueva Leon.

There are unkown random terms associated with each of these levels of clustering. For example perhaps $N[\text{San Pedro}] = 1$, $C[\text{Monterrey}] = 2$, and $S[\text{Nueva Leon}] = -1$. This means that people in San Pedro tend to have 1 unit higher BMI than people in other districts of Monterrey city, people in Monterrey city tend to have 2 units higher BMI than people in other cities of Nueva Leon state, and people in Nueva Leon state have on average one unit lower BMI than people in other states of Mexico. These terms are statistically independent and combine additively, so that subject i has conditional mean value $m + 1 + 2 - 1 = m + 2$.

In variance components modeling, we imagine that all the $N[\cdot]$ terms come from a common distribution, say with mean 0 and variance τ_N^2, all the $C[\cdot]$ terms come from a distribution with mean 0 and variance τ_C^2, and all the $S[\cdot]$ terms come from a distribution with mean 0 and variance τ_S^2. Our goal here is to estimate τ_N^2, τ_C^2, and τ_S^2, to better understand how the different levels of geography explain variation in people's BMI values.

1.6 Marginal covariance structure for nested model

As above, we can determine the marginal mean and covariance corresponding to the conditionally-specified model above. The marginal mean is μ. The variances simply add, so the marginal variance of any observation is

$$Var[Y] = \tau_N^2 + \tau_C^2 + \tau_S^2 + \sigma^2,$$

where $\tau_N^2 = \text{Var}[N]$, for the neighborhood random effects N, $\tau_C^2 = \text{Var}[C]$ for the city random effects C, and so on.

The covariance between two observations depends on how many levels of grouping the observations share in common. Since the grouping levels are nested, if two observations are in the same cluster at a given level, they are also in the same level for alll coarser clusters. For example, two people who live in the same city must also live in the same state.

To put this to use, suppose we have three people, who live in states S1, S1, S2 (i.e. the first two people live in the same state, and the third person lives in a different state), and they live in cities C1, C1, C2 and neighborhoods N1, N2, N3. Then the marginal covariance matrix for these three people is:

$$\begin{pmatrix} \tau_N^2 + \tau_C^2 + \tau_S^2 + \sigma^2 & \tau_C^2 + \tau_S^2 & 0 \\ \tau_C^2 + \tau_S^2 & \tau_N^2 + \tau_C^2 + \tau_S^2 + \sigma^2 & 0 \\ 0 & 0 & \tau_N^2 + \tau_C^2 + \tau_S^2 + \sigma^2 \end{pmatrix}$$

1.7 Crossed variance components

Variance components can also be "crossed", which basically means "not nested". In a crossed model, variance terms for different variables can occur in arbitrary combinations with each other.

Fitting a crossed model puts more stress on the software, but it can be done if there aren't too many levels of crossing. Suppose for example that we have data on the number of emails sent among people all pairs of people in a sample. For simplicity, we model these counts with a Gaussian distribution so we can use linear mixed models. We can imagine that person i has a propensity $A[i]$ to send emails, and also a propensity $B[i]$ to receive emails.

Higher values of A_i indicate that a person writes a lot of emails, while higher values of B_i indicate that a person recieves a lot of emails. In practice, these values may be correalated, but here we model them as if they were independent, i.e. someone who sends a lot of emails does not necessarily receive a lot of emails, and vice versa.

A simple additive variance component would be

$$Y_{ij} = \mu + A_i + B_j + e_{ij},$$

where Y_{ij} is the number of emails sent from subject i to subject j. The random effect A_i and B_j are crossed, meaning that any of the A_i terms can occur in combination with any of the B_j terms. We are mainly interested in the variance parameters τ_a^2 and τ_b^2, describing, respectively, the variation in the population of email sending and email receiving propensities.

1.8 Marginal covariance structure for crossed models

The crossed model specified above has a very simple mean structure, in which every observation has the same mean μ, which is an unknown parameter to be estimated from the data.

The variance is also fairly simple. Every observation has variance

$$\text{Var}[Y] = \tau_a^2 + tau_b^2 + s^2.$$

There are three possible covariances:

The covariance between Y_{ij} and Y_{ik} (i.e. between the counts for the same sender to different receivers) is τ_a^2.

The covariance between Y_{ik} and Y_{kj} (between the counts for the same receiver with two different senders) is tau_b^2.

The covariance between Y_{ij} and Y_{kl} (different receivers and different senders) is 0.

1.9 Parameters and random effects

Like the widely-used routines in R, Stata, and SAS, Python Statsmodels uses maximum likelihood to fit mixed models to data. (Technically, the default estimator is restricted maximum likelihood, but the difference is not important here). This means that we are optimizing the parameter values in a class of parametric models to best fit the data. The random effects, e.g. random intercepts a_i or $N[i]$ in the examples discussed above, are random variables, not parameters, but unlike the data (which are also treated as random variables), these random effects are not observed. We therefore marginalize the random effects out of the model's likelihood function before fitting to the data. This means that the fitting process does not directly involve these random effects, although it does involve the parameters defining their distribution.

As discussed above, the parameters in a mixed model can broadly be considered as being one of the following types:

Mean structure parameters: this includes regression intercepts and slopes. These parameters determine the marginal mean structure (defined above) but are not sufficient to describe the conditional mean structure, which also depends on a subject's random effects. These parameters are sometimes called "fixed effects" because they describe the marginal trends in the population, not the unique trends for individual subjects.

Variance/covariance structure parameters: this includes variances of random effects, and covariance parameters describing how various random effects are correlated. These are structural parameters describing how the random effects are distributed, not the random effects themselves.

Since the random effects are not parameters, they are not estimated (this is a good thing). However it is possible to predict the value of a random effect after fitting a model. There are various ways to do this, but the most common approach uses a "Best Linear Unbiased Predictor" (BLUP).

In the longitudinal mixed model

$$E[SBP(age, i) \mid a_i, b_i] = a + b \cdot age + a_i + b_i \cdot \text{age}_i$$

where a and b are mean structure parameters (fixed effects), and τ_a^2, τ_b^2,

τ_{ab}, and the "error variance" $V[SBP(age, i) \mid a_i, b_i]$ are variance/covariance structure parameters. The a_i and b_i are the actual "realized" random effects.

In the nested variance components model

$$Y_i = \mu + N[n_i] + C[c_i] + S[s_i] + e_i,$$

the only mean structure parameter is μ. The variance structure parameters are τ_N^2, τ_C^2, τ_S^2, and σ^2.

In the crossed variance components model

$$Y_{ij} = \mu + A_i + B_j + e_{ij},$$

the mean structure parameter is μ, and the variance structure parameters are τ_a^2, τ_b^2, and σ^2.

1.10 Software and algorithms

Estimation routines for linear mixed models are much more challenging to implement than routines for fitting more basic regression approaches such as OLS and GLM. However a series of developments in the past 20 years has led to algorithms that are reasonably fast and stable. Statsmodels utilizes many of these best practices, such as internally re-parameterizing the covariance parameters through their Cholesky factor, and profiling out certain parameters during the estimation process.

The earlier specifications of linear models (e.g. Laird and Ware 1982) were explicitly group-based. This means that there was a grouping variable such as a person, such that observations made on different groups are taken to be independent. Many applications of mixed modeling are compatible with this group-based approach. However heavily crossed models that are widely used in, for example, experimental psychology and linguistics are not.

Recent versions of R's LMER have taken a somewhat different algorithmic approach, utilizing the sparse Cholesky factorization (http://faculty.cse.tamu.edu/davis/welcom Statsmodels does not use this approach, partly because the sparse Cholesky code is not available with a Python-compatible license. The sparse Chleksy

approach may be somewhat more efficient for handling large crossed models as noted above. However Python Statsmodels does use sparse matrices and exploits some matrix factorizations to allow crossed models to be fit.

Another important disctinction between Python Statsmodels and LMER in R (which is the most mature open-source implementation of mixed models) is that the Statsmodels code is written in Python, whereas LMER is mostly written in C that is then linked to R. The Python MixedLM code makes use of advanced Numpy and Scipy techniques (which are written in C) and therefore the distinction is not as clear as it may at first seem. There are many trade-offs in this decision, but at present the Python code generally runs somewhat slower than LMER. There are many innovations underway to accelerate numerical Python code so it is likely that the Statsmodels code will become faster over time.

1.11 Other practicalities

To fit a mixed model to data using Python Statsmodels (or most other software tools), it should be in "long format". This means that there is one row of data for each observed outcome (not for each group). If the data are originally represented in wide format, like this

Subject	Time1Y	Time2Y	Time1X	Time2X
1	34	39	12	9
2	31	27	19	15
...	...	...	...	...

then it should be restructured to long form:

Subject	Time	Y	X
1	1	34	12
1	1	39	9
2	2	31	19
2	2	27	15
...	...	...	...

There are various tools for doing this in Python, including many powerful data manipulation routines in the Pandas library (http://pandas.pydata.org).