Discussion of *Random-Effects Models for Longitudinal Data*

March 2024

1 A motivating example

Between 1980-95, scientists at Harvard conducted a study involving 8000 people across 6 cities in the US. The scientists selected 6 cities in the US that roughly represented all American cities at large and chose 8000 people across these cities. From time to time, certain demographic and health characteristics of these people were measured. The goal was to study the effects of air pollution on mortality rates.

What would be a reasonable statistical approach to take in this case? Since the outcome we want to look at is mortality, we could measure each person's lung function from time to time and study it's relationship with the air quality in their city. To make an effective case for better air quality, we need to adjust for people's personal characteristics. For instance, if they smoke or don't have good diet, that needs to be accounted for in order to accurately assess the effect of air pollution. This means some sort of a regression model.

2 Motivating the model

Suppose we want to set up a simple linear regression model. We would write the model as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon.$$

Here, $\mathbf{y}$ a vector consisting of all lung function measurements taken for all people and $\mathbf{X}$ is a matrix consisting of the covariates for all the people across the entire experiment. The vector ϵ is the vector of random noise that is part of linear regression. We could simply estimate β using the usual methods and arrive at an answer. However, there is an issue with this.

One common assumption in linear regression models is that the noise across measurements is independent. However, different people are being measured under different conditions. Thus the sources of noise are not the same for everyone. Furthermore, we have multiple measurements per person which means there is correlation. Due to these reasons, the simple assumption of i.i.d noise cannot be taken for granted.

There is yet another problem, this time with β. The model above is assuming that the effect of each variable on lung function is the same for each person. However, depending on each person's individual traits the effects of different variables could actually be different. There could be genetic, dietary, demographic factors that also have a say in how air quality affects a person's lung function. Thus, we are looking to fit a different β for every person.

Accounting for all these introduces more and more parameters into the model which causes the model to be too complex to be fit. What we need is a something that is flexible enough to account for the variation coming from personal factors and test conditions but that is also simple enough to be fit. One such middle ground is the mixed effects model which was studied extensively in [3]. This model is written as follows.

$$y_i = X_i\alpha + Z_i b_i + e_i,$$
$$b_i \overset{\text{iid}}{\sim} N(0, D), e_i \overset{\text{iid}}{\sim} N(0, R_i).$$

In this model, we divide the covariates into two groups. The X_i's denote those who's effect for all individuals is the same. There is a vector of coefficients α associated with these covariates. The Z_i's denote the covariates who's effect can change person to person. For the Z_i's, we need a different vector of coefficients b_i for person i. Instead of assuming that all these are different vectors that need to be fitted individually, we assume that they are IID draws from a normal distribution. This allows the model to have different effects for different people without introducing new parameters. The only parameter that is introduced is the covariance matrix D. This also frees up enough parameters to allow for a different covariance matrix R_i for each individual's noise vector.

3 Model Fitting

In statistical machine learning, one of the most common ways to fit probabilistic models is through maximum likelihood estimation. Roughly speaking, we would like to find the set of parameters under which the data is the most likely of being observed. These estimators are called maximum likelihood estimators (MLE). In this paper, the authors advocate for the use of the Expectation Maximization (EM) algorithm to numerically calculate maximum likelihood estimators. Intuitively, the EM algorithm "augments" the observable data to a set of comlete data, and treat the problem as an incomplete-data problem to facilitate computation. It employs an iterative two-step process: the E-step (or estimation step), where it estimates missing or latent variables, and the M-step (or maximization step), where it optimize the parameters of the model to best explain the data. This paper makes a significant contribution by offering a unified strategy for modeling random effects, presenting an approach that is both implementable and user-friendly compared to the more complex, hard to implement methods available at the time. However, it's worth noting that the original EM algorithm had its limitations, such as sensitivity to initial estimates and slow convergence. Over time, numerous enhancements have been proposed and successfully integrated, significantly improving its efficiency and accessibility.

4 Statistical Software

Following this seminal publication, the statistical and machine learning communities have developed mature, user-friendly software and packages for fitting and making inference on random-effects models. These tools have made sophisticated data analysis more accessible to researchers and practitioners across various fields. For instance, the lme4 package in R stands out as one of the most popular, with over 70,000 citations by 2024. This package, along with others like the Proc Mixed package in SAS and the HLM software, demonstrates the enduring legacy of the paper's contributions to statistical analysis.

5 Conclusion

In summary, The authors laid the foundation for setting up and fitting models with random-effects. Its significance is underscored by its extensive application across various fields to this day, with a remarkable citation count of 11,201 times as of 2024. The methodologies proposed have greatly enhanced data analysis and inference, contributing to advancements in fields as diverse as policy making[1], education[2], psychology[4], etc. This paper represents a milestone in the journey toward making data analysis more approachable and applicable to real-world problems. It showcases how an innovative approach can revolutionize methodologies across disciplines, making complex analyses not only possible but practical for everyday research and decision-making.

References

[1] Atif Ansar, Bent Flyvbjerg, Alexander Budzier, and Daniel Lunn. Should we build more large dams? the actual costs of hydropower megaproject development. *Energy Policy*, 69:43–56, 2014.

[2] Frances A Campbell, Elizabeth P Pungello, Shari Miller-Johnson, Margaret Burchinal, and Craig T Ramey. The development of cognitive and academic abilities: growth curves from an early childhood educational experiment. *Developmental psychology*, 37(2):231, 2001.

[3] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[4] Rohan Puthran, Melvyn WB Zhang, Wilson W Tam, and Roger C Ho. Prevalence of depression amongst medical students: A meta-analysis. *Medical education*, 50(4):456–468, 2016.