Blake Shurtz

6/7/18

STAT 6310-01

Multi-Level Regression Modeling with a Varying Intercept: A Walkthrough

Multi-level models (MLM) allow a statistician to compare multiple "levels" of data in a single regression framework. This paper introduces a multi-level regression model with a varying intercept and a constant slope.

There are two "levels" of data: the group level and the individual level. Presented in terms of the structure of the data, a MLM combines two data sets: an individual-level data set and a group-level data set.

Individual-Level Data

y	x	group ID
0	0	1
2	1	1
6	0	2
8	1	2
12	0	3
14	1	3

Group-Level Data

x	group ID
0	1
1	2
2	3

Data Structure

When a group-level predictor exists that a priori contains predictive power, a MLM should perform better than an unpooled regression. This paper puts this claim to the test.

Unlike a multi-level model, an individual-level regression can not factor in group-level predictors. Information is lost and the statistician is worse off for it. The best that an individual-level regression can do is an unpooled regression with a factor variable for the group ID.

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_j + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_y^2)$$

y_i is a continuous response variable for observation *i*. α is the intercept parameter. β_1 is the slope parameter which is assumed to be constant across all groups. x_i is the binary predictor.

x_j is the factor for group ID. The error term ε_i is assumed to be normally distributed around zero with a constant variance σ_y^2.

Although it converges to a single regression equation, a MLM is conceptually divided into two separate equations- one for each level of data.

At the individual-level, the MLM resembles the unpooled regression except that the intercept $\alpha_{j[i]}$ is a random variable.

$$y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_y^2), \quad i = 1, ..., n \ .$$

For each group *j*, there is a unique intercept that is a function of the group-level data. The error term ε_i is assumed to be normally distributed around zero with a constant individual-level variance σ_y^2.

$\alpha_{j[i]}$ is the response in the group-level regression equation, which is assumed to be normal with mean μ_α and group-level variance σ_α^2.

$$\alpha_j = \gamma_0 + \gamma_1 \mu_j + \eta_j, \eta_j \sim N(0, \sigma_\alpha^2)$$

γ_0 and γ_1 are unmodeled parameters with non-informative normal prior distributions. μ_j is the group level data. η_j is the error term with constant group-level variance σ_α^2.

Data was generated in R for the benefit of being able to specify the population regression lines. While model comparison can certainly be done with existing data, specifying target parameters beforehand ensures clarity in comparison between estimates and parameters.

```
set.seed(1234)
group_0A <- rnorm(15, 0, 2)
group_1A <- rnorm(15, 2, 2)
response<-c(group_0A, group_1A)
indv_pred <- c(rep(0, times = 15), rep(1, times = 15))
group_id <- c(rep(1, 30))
group_pred <- c(0, 30)
mydata <- cbind(response, indv_pred, group_id, group_pred)
```

Code for generating group 1 data

Efforts were made to keep the data values small and manageable. Model fit calculations were done by hand in an effort to get "down and dirty" with the results. j = 3 groups were chosen, which is the minimum number of groups required to prevent model convergence.

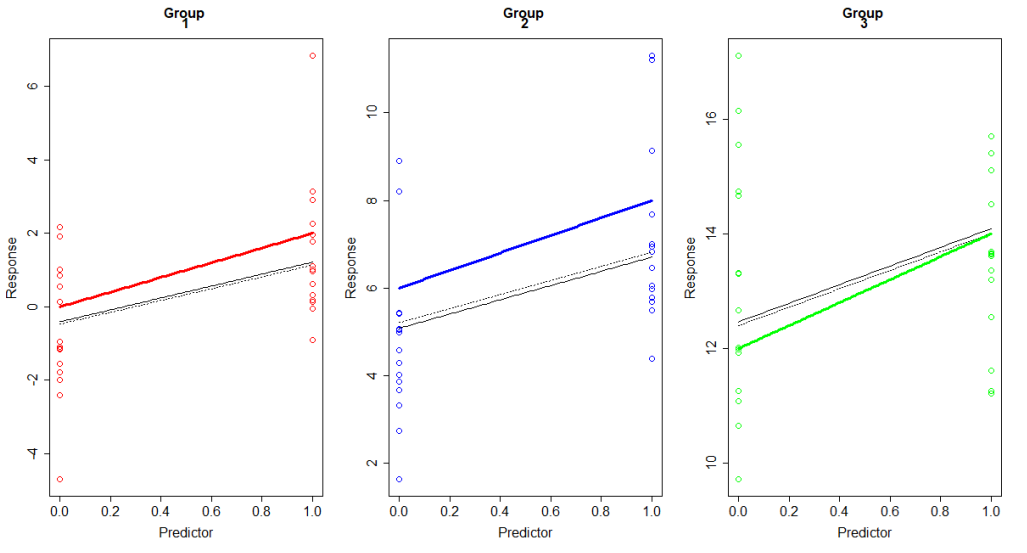
$$E(Y_1) = 2x_i + \varepsilon_i, \varepsilon \sim N(0, \sigma_y^2)$$
$$E(Y_2) = 6 + 2x_i + \varepsilon_i, \varepsilon \sim N(0, \sigma_y^2)$$
$$E(Y_3) = 12 + 2x_i + \varepsilon_i, \varepsilon \sim N(0, \sigma_y^2)$$

Population Regression Lines

The motivation for choosing these specific parameters is as follows: equal spacing on the intercepts with attention to standard deviation of the response; and a constant slope for all three lines.

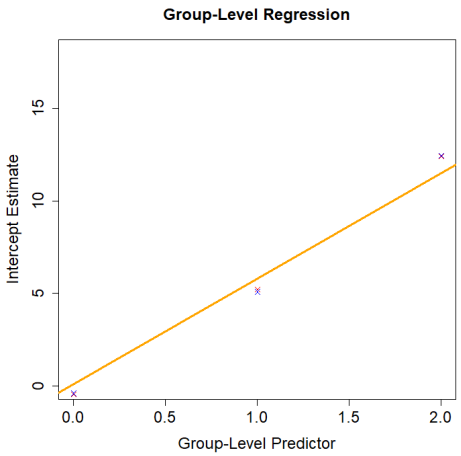
A normal distribution was chosen to ensure that the residuals would be normally distributed and the model would meet the OLS assumption of homoscedastic residuals. A group size of n=30 (n=15 for each response) led to some unexpected distributions. For example, almost all of the response values for group 2 when x=0 is below the specified parameter.

A constant variance of $\sigma_y = 2$ allows for some overlap among the groups as about 2/3 of observations fall within y=±2. Several outliers were found- as expected but not predicted- the most extreme outlier being about 2.45 standard deviations from the mean.



Population Regression Line (colored), individual-level regression (solid) and MLM regression line (dashed)

Group-level regression data was chosen as $x = (0,1,2)$ for group $i = (1,2,3)$ respectively. These values were chosen in a somewhat arbitrary manner, noting the correlation between the group-predictor and the responses of 16%, where the lowest value for the group level predictor is paired with the lowest intercept-value for the population regression line.



Group	Group $\hat{\alpha}_j$	MLM $\hat{\alpha}_j$
1 (y=0)	0.1	-.47
2 (y=6)	5.8	5.2
3 (y=12)	11.5	12.41

Group level regression line with intercept values for group-level and MLM regression.

The metric used to compare the models is the sum of squares of the difference between the estimated regression coefficient and the population regression coefficient for both the intercept and the slope parameter for each group. Let's call it the **sum of squares of the coefficients (SSC)**.

$$SSC_j = \sum_{j=1}^{3}(\hat{\alpha}_j - \alpha_j)^2 + \sum_{j=1}^{3}(\hat{\beta}_j - \beta_j)^2$$

The reason that this metric is used is that multi-level models and individual-level models traditionally use different measurements of model fit: R^2 for an individual-level model and the deviance information criterion (DIC) for a multi-level model.

The results of the model comparison are that the multi-level model has a SSC that is 12% smaller than the individual-level model (0.4/3.5).

Group 1	Y=0	Y=2	Group 2	Y=6	Y=8	Group 3	Y=12	Y=14
Classic	-0.4	1.22	Classic	5.09	6.08	Classic	12.5	14.1
MLM	-0.47	1.16	MLM	5.21	6.84	MLM	12.4	14

Table 1: Predicted y-values versus modeled parameters for each group.

Group 1	SSC	Group 2	SSC	Group 3	SSC	Total
Classic	0.77	Classic	2.51	Classic	0.23	3.5
MLM	0.93	MLM	1.98	MLM	0.16	3.1

Table 2: Sum of squared coefficients for each group, along with cumulative SSC.

Using the lme4 package, the point estimate for the residual standard error for the group-level regression is $\sigma_\alpha = .69$ and the individual-level residual standard error is $\sigma_y = 1.9$. The intraclass correlation is $\rho = 12\%$, meaning that 12% of the variation can be explained by the group-level data.

Surprisingly, ρ is equivalent to the ratio of coefficients of SSC (to the 3rd decimal place). A future research topic would be to investigate the relationship between these metrics.

During this project, a number of "forking paths" presented themselves. The biggest challenge was sticking to a single topic! Future research would involve exploring MLMs in more detail.

First, I would choose a different metric for model fit. It is possible to calculate a deviance information criterion for an individual-level regression if it is done in a Bayesian framework with uninformative priors.

Varying group-size when simulating data is another option. One benefit to using a multi-level model is that groups with smaller sample sizes have their multi-level estimates "pulled" towards the group-level regression. While this effect could be demonstrated using simulated data, it did not fit the goal of the research project. I consider this feature of MLM a major benefit in a real-world, messy data analysis.

There are a number of software options for Bayesian analysis. Using the lme4 package in R is the easiest option, but it only gives point estimates the parameters. WinBUGS uses MCMC to create posterior distributions. Another option is to use Stan. This option is recommended by Professor Gelman, whose work (and blog) I followed closely during this project.

Finally, I am most excited to expand my research from a Bayesian *regression* framework to a more general Bayesian framework. Bayesian analysis isn't only for hierarchical data, but has the potential for modeling more general frameworks. Indeed, it seems that as long as a data set provides *some* predictive power in another data set, they can be integrated into a Bayesian model.

Works Cited

Gelman and Hill, Data Analysis Using Regression and Multilevel/Hierarchical Models. (2007)

Home Page for Gelman and Hill. http://www.stat.columbia.edu/~gelman/arm/

Radon Data Set. http://www.stat.columbia.edu/~gelman/arm/examples/radon/

LEMMA (Learning Environment for Multilevel Methodology and Applications). https://www.cmm.bris.ac.uk/lemma/

National Centre for Research Methods. University of Bristol.

Cross Validated. https://stats.stackexchange.com/