Comparing Multilevel Models - ISyE 6420 Project

Gregory Schreiter, April 23, 2020

1 Introduction

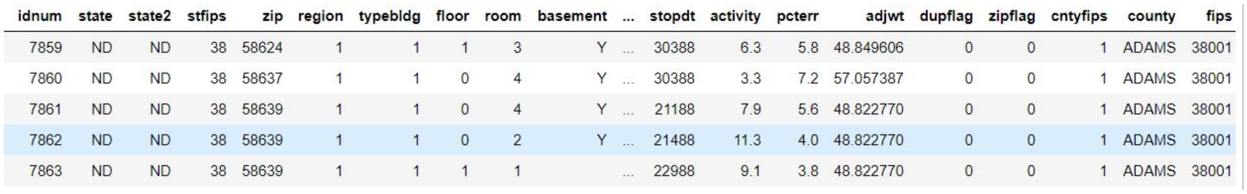
With this project we aim to compare and contrast several different approaches to Bayesian modeling using a dataset that measures Radon levels in houses along with various house features (geographical, spatial). The data may be accessed at the following URL:

`https://raw.githubusercontent.com/pymc-devs/pymc3/master/pymc3/examples/data/srrs2.dat`

Radon is a radioactive gas that enters homes through contact points with the ground. It is a carcinogen that is the primary cause of lung cancer in non-smokers. Radon levels vary greatly from household to household. The EPA did a study of radon levels in $80,000$ houses. Two important predictors:

- Measurement in basement or first floor (radon higher in basements)
- County uranium level (positive correlation with radon levels)

The hierarchy in this example is households within county (there are 53 counties in North Dakota). We studied 9 different Bayesian models that have varying levels of complexity. For full details about the data preprocessing and models chosen, please visit the included ipynb file and associated html rendering (open this in your favorite browser). Since we are limited by space in this report, we will focus on a comparison of the Complete Pooling, Partial Pooling, and the Contextual Effects models. Here is a quick snapshot of what the dataframe header looks like:

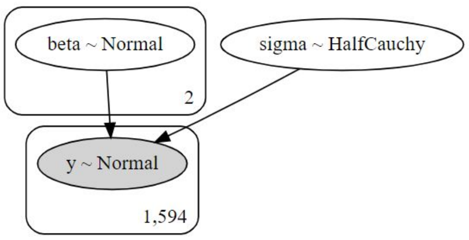


idnum	state	state2	stfips	zip	region	typebldg	floor	room	basement	...	stopdt	activity	pcterr	adjwt	dupflag	zipflag	cntyfips	county	fips
7859	ND	ND	38	58624	1	1	1	3	Y	...	30388	6.3	5.8	48.849606	0	0	1	ADAMS	38001
7860	ND	ND	38	58637	1	1	0	4	Y	...	30388	3.3	7.2	57.057387	0	0	1	ADAMS	38001
7861	ND	ND	38	58639	1	1	0	4	Y	...	21188	7.9	5.6	48.822770	0	0	1	ADAMS	38001
7862	ND	ND	38	58639	1	1	0	2	Y	...	21488	11.3	4.0	48.822770	0	0	1	ADAMS	38001
7863	ND	ND	38	58639	1	1	1	1		...	22988	9.1	3.8	48.822770	0	0	1	ADAMS	38001

2 Complete Pooling Model

This is the simplest model, in which we assume that all counties are the same and we estimate a single Radon level:

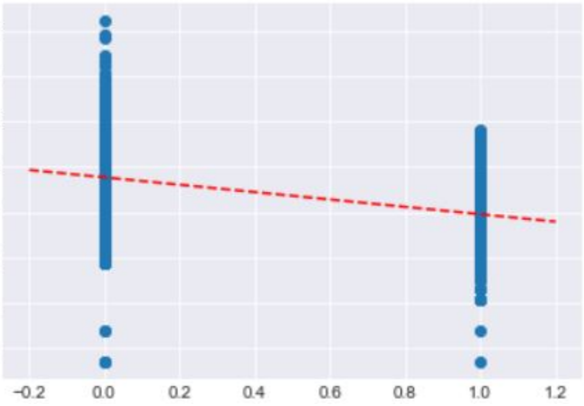
$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, 1594$$



After running 1000 samples with a burn-in of 1000 (on 4 markov chains), our fitted model is

$$y = 1.7636648892888287 - 0.8129923389686631x$$

The following image shows this model (x-axis represents either basement or ground-level):



3 Partial Pooling Model

When we pool our data, we imply that they are sampled from the same model. This ignores any variation among sampling units (other than sampling variance).

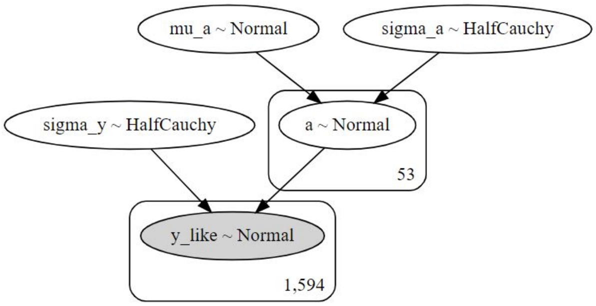
When we analyze data unpooled, we imply that they are sampled independently from separate models. At the opposite extreme from the pooled case, this approach claims that differences between sampling units are too large to combine them.

In a hierarchical model, parameters are viewed as a sample from a population distribution of parameters. Thus, we view them as being neither entirely different or exactly the same. We can use PyMC to easily specify multilevel models, and fit them using Markov chain Monte Carlo.

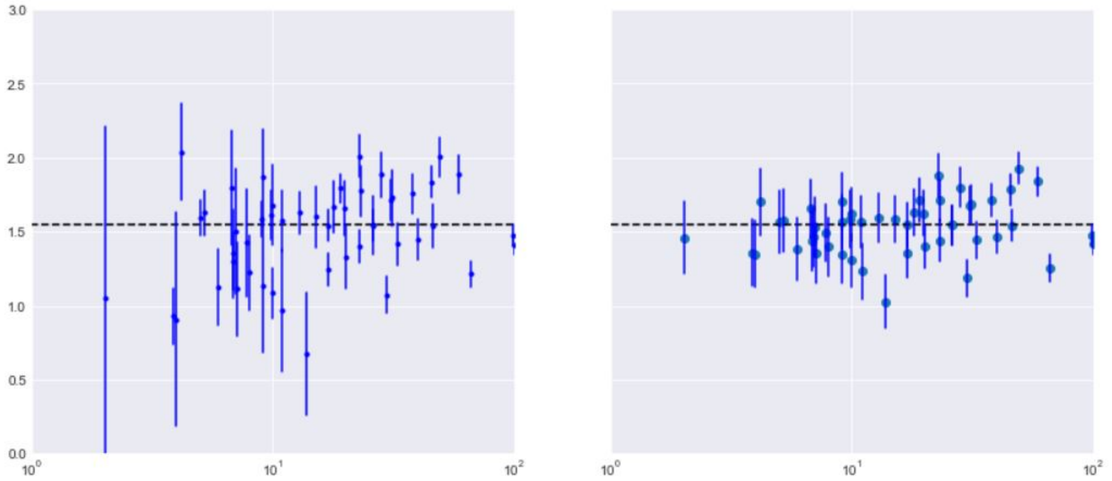
The simplest partial pooling model for the household radon dataset is one which simply estimates radon levels, without any predictors at any level. A partial pooling model represents a compromise between the pooled and unpooled extremes, approximately a weighted average (based on sample size) of the unpooled county estimates and the pooled estimates:

$$\hat{\alpha} \approx \frac{(n_j/\sigma_y^2)\bar{y}_j + (1/\sigma_\alpha^2)\bar{y}}{(n_j/\sigma_y^2) + (1/\sigma_\alpha^2)}$$

Estimates for counties with smaller sample sizes will shrink towards the state-wide average, whereas estimates for counties with larger sample sizes will be closer to the unpooled county estimates:



We run 1000 samples with a burn-in of 1000 (on 4 markov chains) of this Partial Pooling model. We compare the coefficient estimates for each county obtained via the unpooled and partial pooling methods:



We see that the difference between the unpooled and partially-pooled estimates are that the credible sets for the former are both more extreme and more imprecise.

4 Contextual Effects Model

We briefly describe the remaining models that we looked at, because these are key in motivating the contextual effects model:

1. The Varying Intercept model allows intercepts to vary across county, according to a random effect:

   $$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma_y^2) \text{ with intercept random effect } \alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

   As with the the *no-pooling* model, we set a separate intercept for each county, but rather than fitting separate least squares regression models for each county, multilevel modeling shares strength among counties, allowing for more reasonable inference in counties with little data.

2. Alternatively, we can posit a model that allows the counties to vary according to how the location of measurement (basement or floor) influences the radon reading:

   $$y_i = \alpha + \beta_{j[i]} x_i + \epsilon_i$$

3. The most general model allows both the intercept and slope to vary by county:

   $$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$$

4. A primary strength of multilevel models is the ability to handle predictors on multiple levels simultaneously. If we consider the varying-intercepts model above:

   $$y_i = \alpha_{j[i]} + \beta x_i + \epsilon_i$$

we may, instead of a simple random effect to describe variation in the expected radon value, specify another regression model with a county-level covariate. Here, we use the county uranium reading u_j, which is thought to be related to radon levels:

$$\alpha_j = \gamma_0 + \gamma_1 u_j + \zeta_j, \quad \zeta_j \sim N(0, \sigma_\alpha^2)$$

Thus, we are now incorporating a house-level predictor (floor or basement) as well as a county-level predictor (uranium).

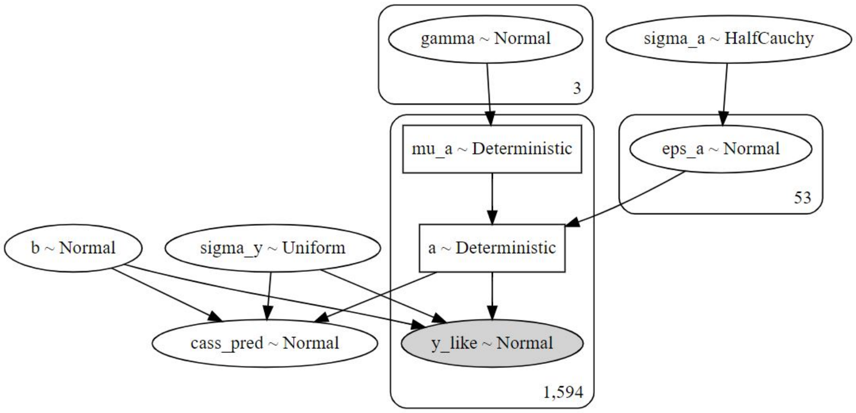
Note that the model has both indicator variables for each county, plus a county-level covariate. In classical regression, this would result in collinearity. In a multilevel model, the partial pooling of the intercepts towards the expected value of the group-level linear model avoids this.

Group-level predictors also serve to reduce group-level variation σ_α. An important implication of this is that the group-level estimate induces stronger pooling.

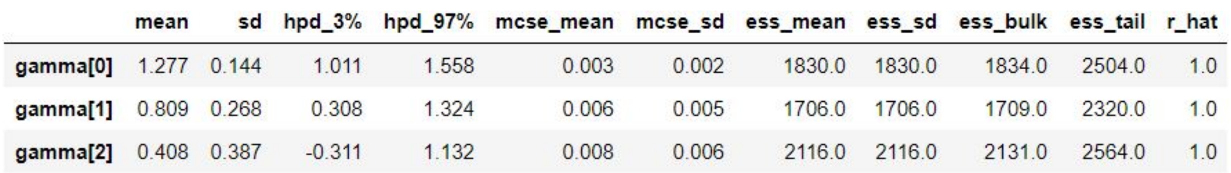
5. In some instances, having predictors at multiple levels can reveal correlation between individual-level variables and group residuals. We can account for this by including the average of the individual predictors as a covariate in the model for the group intercept.

$$\alpha_j = \gamma_0 + \gamma_1 u_j + \gamma_2 \bar{x} + \zeta_j$$

These are broadly referred to as **contextual effects**:



We run 1000 samples with a burn-in of 1000 (on 4 markov chains) of this Contextual Effects model, obtaining estimates for the three gamma coefficients:



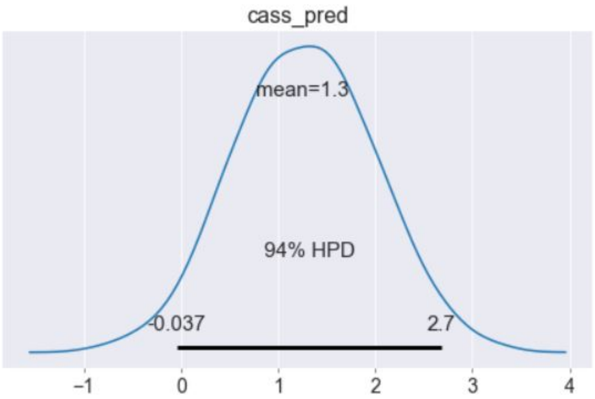
	mean	sd	hpd_3%	hpd_97%	mcse_mean	mcse_sd	ess_mean	ess_sd	ess_bulk	ess_tail	r_hat
gamma[0]	1.277	0.144	1.011	1.558	0.003	0.002	1830.0	1830.0	1834.0	2504.0	1.0
gamma[1]	0.809	0.268	0.308	1.324	0.006	0.005	1706.0	1706.0	1709.0	2320.0	1.0
gamma[2]	0.408	0.387	-0.311	1.132	0.008	0.006	2116.0	2116.0	2131.0	2564.0	1.0

We may infer from the above that counties with higher proportions of houses without basements tend to have higher baseline levels of radon. Perhaps this is related to the soil type, which in turn might influence what type of structures are built.

We may also use this model for prediction. There are two types of prediction that can be made in a multilevel model:

- a new individual within an existing group
- a new individual within a new group

For example, if we wanted to make a prediction for a new house with no basement in Cass county, we just need to sample from the radon model with the appropriate intercept:



We see that the predicted value is 1.3.

5 Summary and Conclusion

In general, the complexity of the model chosen will depend on many factors such as domain knowledge of the problem, interpretability of the model and what the analysis aims to achieve. While the Complete Pooling model is easy to interpret, the fact that it does not account for geographic variation among counties is a disadvantage. The Partial Pooling model yields stronger estimates compared to either the Complete-Pooling or No-Pooling models, and nicely balances interpretability with accuracy. The final Contextual Effects model is quite complex, but this yields the greatest flexibility as far as modeling specific interactions and making predictions.

There are many benefits of Multilevel Models despite their complexity. They help account for natural hierarchical structure of observational data, they allow for estimation of coefficients for (under-represented) groups, they incorporate individual-level and group-level information when estimating group-level coefficients, and they allow for variation among individual-level coefficients across groups.

6 References

Gelman, A., & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models (1st ed.). Cambridge University Press.

Gelman, A. (2006). Multilevel (Hierarchical) modeling: what it can and cannot do. Technometrics, 48(3), 432–435.

A primer on Bayesian Methods using Multilevel Modeling