Stat637: Multilevel (Hierarchical) Modeling: What It Can and Cannot Do [1]

Arthur Lui

3 April 2015

1 Summary of the Paper

Gelman reviews the multilevel (hierarchical) model using a radon-measurement dataset from the Environmental Protection Agency (EPA). He demonstrates that using the hierarchical model is almost always an improvement from classical regression, and shows when it is essential, useful, or only helpful.

Of interest to the EPA is the distribution of radon levels across homes in the US. Radon (measured in picoCuries per Liter or pCi/L) is a carcinogenic gas that causes several thousand lung cancer deaths each year. It is known that radon comes from underground and enters more easily into homes that are built into the ground, or that have basements. Consequently, the presence of basements in homes an important predictor for radon levels. In addition, uranium (measured in parts per million or ppm), a solid that exists in soil, is a parent element that eventually decays to form radon gas. So, soil uranium measurements are also an important predictor of radon levels. The dataset from the EPA contains information on more than 80,000 houses throughout the country. But Gelman analyzes only data in Minnesota, and groups observations (to form a hierarchy) by county, within the state of Minnesota.

The structure for the hierarchical model that Gelman fits is similar to the following:

$$\begin{aligned} y_{ij}|\alpha_j, \beta, \sigma_y^2, x_{ij} &\sim \mathrm{N}(\alpha_j + \beta x_{ij}, \sigma_y^2), \text{ for } i = 1, ..., n_j, \ j = 1, ...J. \\ \alpha_j|\gamma_0, \gamma_1, \sigma_a^2, u_j &\sim \mathrm{N}(\gamma_0 + \gamma_1 u_j, \sigma_a^2), \text{ for } j = 1, ..., J, \\ \beta &\sim \mathrm{N}(0, 100) \\ \gamma_0 &\sim \mathrm{N}(0, 100) \\ \gamma_1 &\sim \mathrm{N}(0, 100) \\ \sigma_y^2 &\sim \mathrm{N}(2, 1) \\ \sigma_a^2 &\sim \mathrm{N}(2, 1) \end{aligned} \tag{1}$$

where y_{ij} is the *log* radon measurement in house i within county j, x_{ij} is an indicator (0 for "No", 1 for "Yes") for whether house i in county j has a basement, u_j is the *log* soil uranium measurement in county j. Consequently, the interpretation of α_j would be the expected *log* radon measurement in a houses without basements in county j (within Minnesota). Instead of interpreting β, it may be more informing to interpret $\alpha_j + \beta$, which is the the expected *log* radon measurement in a houses with basements in county j. γ_0 is the expected *log* radon measurement in houses without basements in Minnesota when county uranium level is 1. γ_1 is the expected increase in *log* radon measurement in houses in county j when county j's uranium measurement increase by $e \approx 2.718$. σ_y^2 is the variation that exists between houses within county j. Finally, σ_a^2 is the variation that exists between baseline *log* radon measurements between different counties in Minnesota.

2 Multilevel Hierarchical Generalized Linear Models

The ideas of multilevel modeling can be extended to generalized linear models (GLM's) where the likelihood is not necessarily normal. Specifically, if we were to set a new indicator variable $z_{ij} = \mathrm{I}(y_{ij} > \log(4))$, and

[1] https://github.com/luiarthur/Fall2014/blob/master/Stat637/mp

interpret it to be house i's log radon danger level (1 if dangerous, and 0 if safe), then we could model z_{ij} with a Bernoulli likelihood, and link the log odds to covariates of interest. (The EPA recommends homes to take corrective measures to reduce radon levels when they are measured to be 4 and above.) We could express this model as

$$z_{ij}|p_{ij} \sim \text{Bernoulli}(p_{ij}), \text{ for } i = 1, ..., n_j, \ j = 1, ...J.$$
$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + \beta x_{ij}$$
$$\alpha_j|\gamma_0, \gamma_1, \sigma_a^2, u_j \sim \text{N}(\gamma_0 + \gamma_1 u_j, \sigma_a^2), \text{ for } j = 1, ..., J$$

$$\implies z_{ij}|\alpha_j, \beta, x_{ij} \sim \text{Bernoulli}\left(\frac{exp(\alpha_j + \beta x_{ij})}{1 + exp(\alpha_j + \beta x_{ij})}\right), \text{ for } i = 1, ..., n_j, \ j = 1, ...J.$$
$$\alpha_j|\gamma_0, \gamma_1, \sigma_a^2, u_j \sim \text{N}(\gamma_0 + \gamma_1 u_j, \sigma_a^2), \text{ for } j = 1, ..., J, \tag{2}$$

with the same prior specifications as before.

3 Fitting The Normal Model

The posterior means for the α_j's range from 2.35 to 1.31. It appears from the trace plots that the chain has converged the correct distribution. Figure 1 shows the posterior densities and trace plots for α_1, α_{50}, and α_{85}, which are representative of the posteriors of the other α_j's.

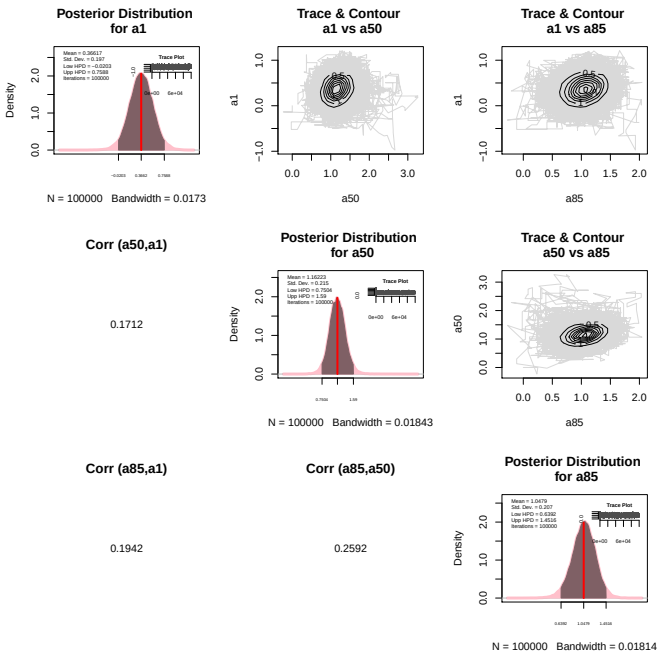

Figure 1: Posterior Distributions for α_1, α_{50}, and α_{85}

The chain also appears to have reached the correct distributions for the posterior of the other parameters and hyper parameters. Here we will not discuss the plots in detail, but simply display them in Figure 2.

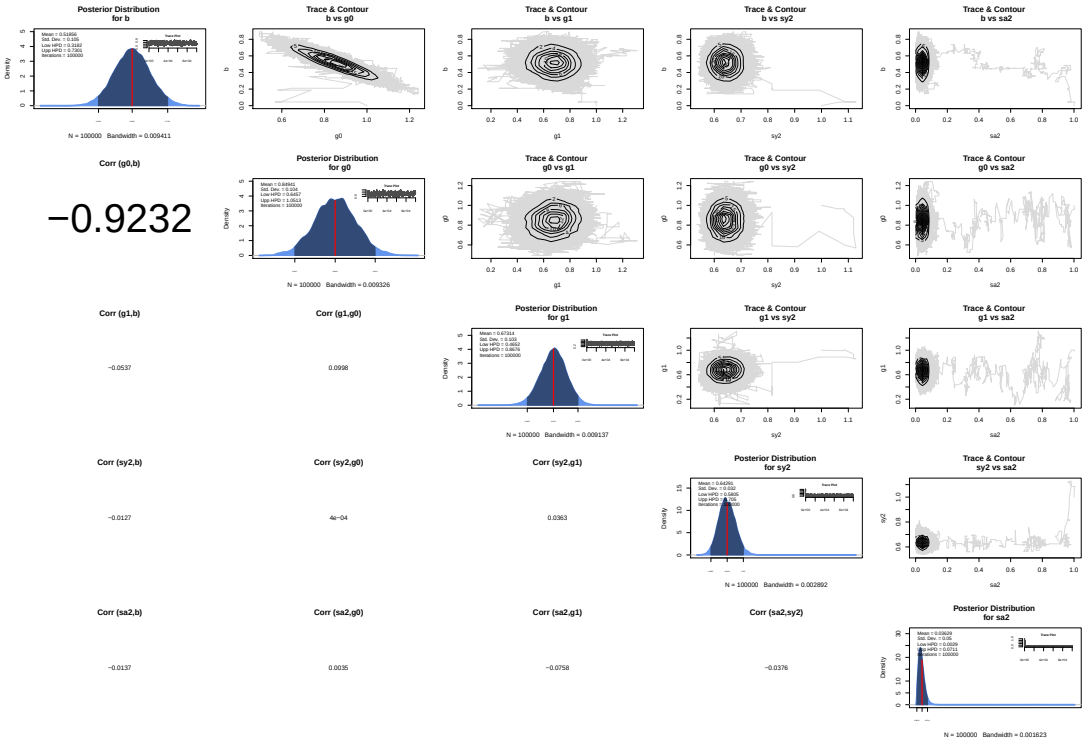

Figure 2: Posteriors and trace plots for parameters and hyper parameters

The EPA is interested in the relationship between county level radon measurements and county level uranium measurements. Figure 3 summarizes this relationship. On the y-axis is the predicted log radon levels computed using the estimated parameters. On the x-axis is the counties arranged in order of predicted log radon levels. The red line is the predicted log radon level for houses with basements, the blue line is the predicted log radon level for houses without basements. The grey line is the measured log uranium level. We see in general that as uranium levels increase, the radon levels increase. Moreover, we see that radon levels for houses with basements are higher than those without. It is interesting to note, however, that it is not always the case that counties with higher uranium levels have higher radon levels that that of counties with lower uranium levels. This provides useful information for the EPA as they can investigate why certain counties would have higher radon levels despite lower uranium levels. Also, note that the predicted radon levels for houses without basements are below 4 for all counties.

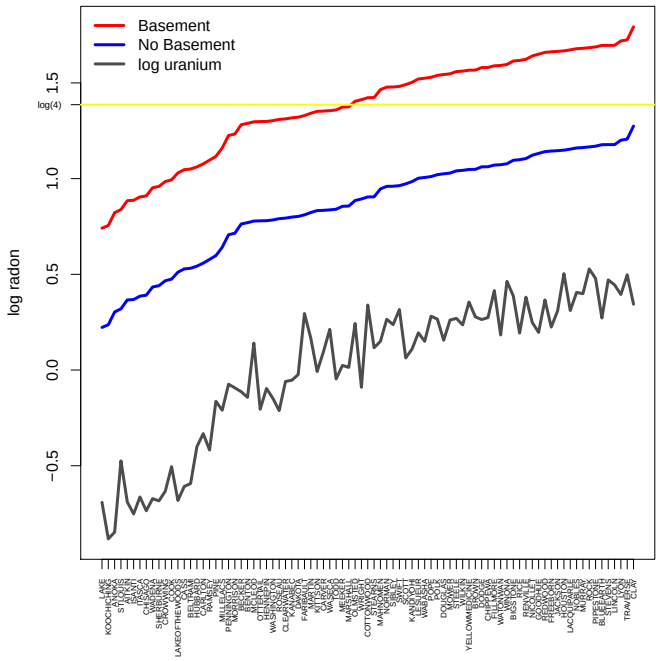

Figure 3: The red and blue lines are the predicted log radon levels (log pCi/L) for house with basements and houses without basements for each county respectively. The grey line is the measured log soil uranium levels (log ppm) for each county. The yellow line is level at log(4).

4 Fitting The Bernoulli Model

To extend these ideas to generalized linear models, model (2) was fit. The posterior means for the α_j's range from -2.55 to .274. It appears from the trace plots that the chain has converged the correct distribution. Figure 4 shows the posterior densities and trace plots for α_1, α_{50}, and α_{85}, which are representative of the posteriors of the other α_j's.
The chain also appears to have reached the correct distributions for the posterior of the other parameters and hyper parameters. Here we will not discuss the plots in detail, but simply display them in Figure 5.
Finally, Figure 6 summarizes the relationship between predicted probabilities of homes in different counties having radon levels above 4. For the plot above, on the y-axis is the predicted probabilities of houses having

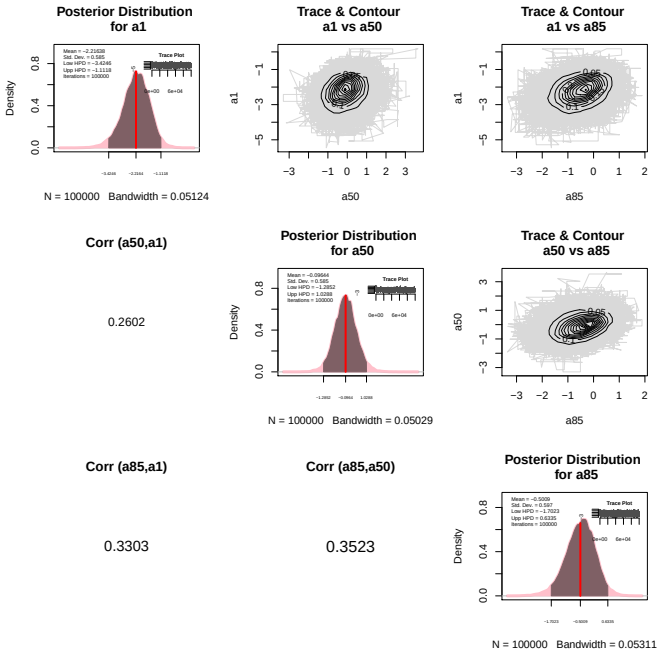

Figure 4: Posterior Distributions for α_1, α_{50}, and α_{85}

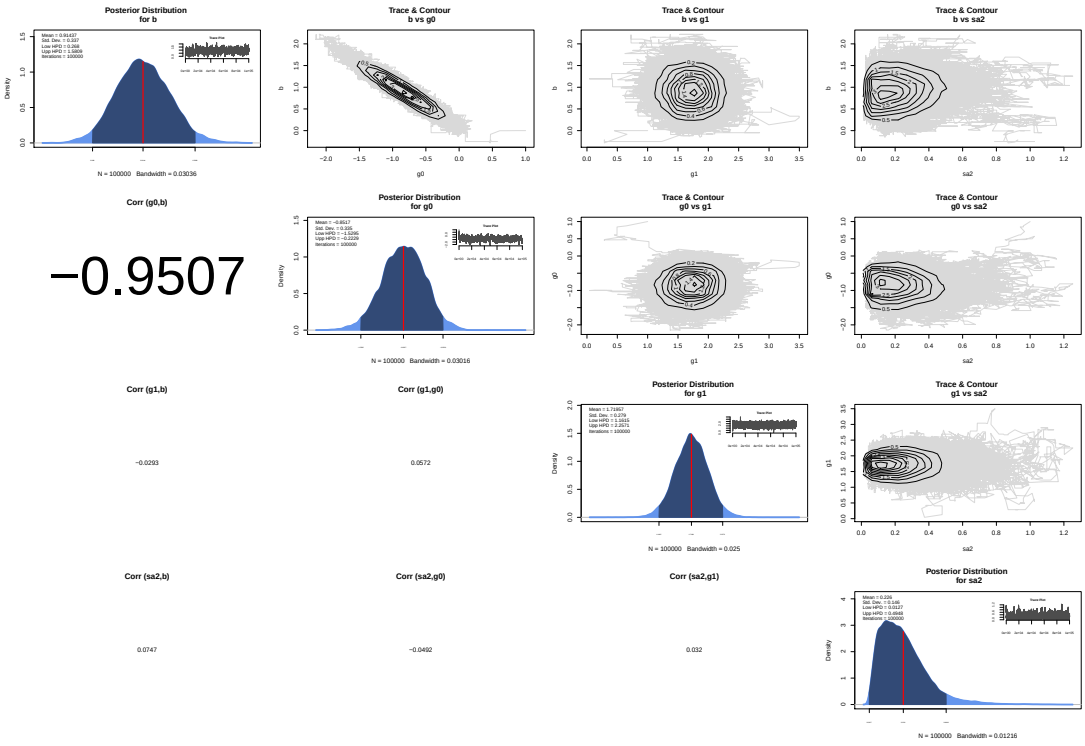

Figure 5: Posteriors and trace plots for parameters and hyper parameters

radon levels above 4, computed using the estimated parameters. On the x-axis is the counties arranged in order of predicted probabilities. The red line is the predicted probabilities for houses with basements, the blue line is the predicted probabilities for houses without basements. The grey line in the plot below is the measured log uranium level. Again, we see in general that as uranium levels increase, the probabilities increase. Moreover, we see that houses with basements are more likely to have radon levels above 4 than those without. Again, it is not always the case that counties with higher uranium levels are more likely to have radon levels above 4.

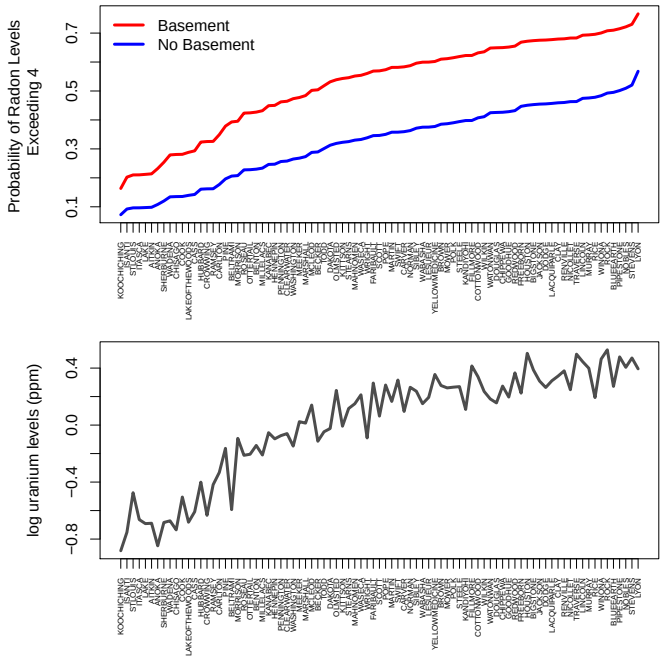

Figure 6: The figure above is the estimated probability of a house with a basement (red) and that of a house without a basement (blue) having radon levels above 4 for each county. The figure below is the measured log uranium levels (log ppm) for each county.

Comparisons

The choice of likelihoods obviously provides different interpretations for parameters. It is interesting to note that the counties with higher predicted probability of having radon levels above 4 are not necessarily the counties with higher predicted radon levels. For instance, Clay county has the highest predicted radon levels, but it 14 from the county (Lyon) with the highest probability of having radon levels above 4. This is due to Lyon having more houses slightly above 4 piC/L radon measurements and Clay having only a few houses with much higher radon measurements, pulling the average radon measurements above 4 (See Table 1). The EPA could potentially use this information to tailor further analyses to different counties. I suspect that houses in Lyon may have higher radon levels due to other unmodeled factors that are influencing the slightly higher-than-4 radon levels across Lyon. It may be the case that home owners cannot do much to reduce radon levels in their homes. I would also suppose that the variation in radon levels in Clay could be due to location. Certain regions in Clay could may have higher radon levels. In which case, homeowners may not be able to reduce the radon levels in their homes, but may have to evacuate their homes.

Observation	County	Activity	Basement	U(ppm)
1	LYON	6.50	Y	1.48
2	LYON	6.90	N	1.48
3	LYON	5.10	Y	1.48
4	LYON	12.00	Y	1.48
5	LYON	5.10	Y	1.48
6	LYON	8.90	Y	1.48
7	LYON	5.80	Y	1.48
8	LYON	4.60	Y	1.48
Observation	County	Activity	Basement	U(ppm)
1	CLAY	12.90	Y	1.41
2	CLAY	2.60	Y	1.41
3	CLAY	26.60	Y	1.41
4	CLAY	13.00	Y	1.41
5	CLAY	8.80	Y	1.41
6	CLAY	19.50	Y	1.41
7	CLAY	2.50	N	1.41
8	CLAY	9.00	Y	1.41
9	CLAY	13.10	Y	1.41
10	CLAY	3.60	Y	1.41

Table 1: Radon measurements (piC/L) for Lyon county and Clay county. The all the houses sampled in Lyon, with and without basements, had radon levels above 4, while 7 out of 9 houses with basements sampled in Clay had radon levels above 4. The one house without basement sampled in Clay had radon levels below 4. Notice, however, that the mean radon levels in Clay is 14.7 for houses with basements while the mean radon levels for houses with basements in Lyon is only 6.86 even though Lyon had a higher sampled proportion of houses with radon levels greater than 4.