The three paths of hierarchical modeling

Jon Zelner

February 20, 2020

EPID 684

University of Michigan School of Public Health

jzelner@umich.edu
www.jonzelner.net

Agenda

- Hierarchical GLM Notation

Agenda

- Hierarchical GLM Notation
- Likelihood and model fit activity

Agenda

- Hierarchical GLM Notation
- Likelihood and model fit activity
- Radon!

A re-introduction to Generalized Linear Models (GLMs) for *clustered* data

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Where:

- y_i is continuous outcome measure: height, BMI, etc.

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Where:

- y_i is continuous outcome measure: height, BMI, etc.
- β is risk associated with some kind of exposure

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Where:

- y_i is continuous outcome measure: height, BMI, etc.
- β is risk associated with some kind of exposure
- $x_i \in [0, 1]$ is an indicator of exposure.

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Where:

- y_i is continuous outcome measure: height, BMI, etc.
- β is risk associated with some kind of exposure
- $x_i \in [0, 1]$ is an indicator of exposure.
- α is expected outcome when x_i = 0

Notation for a classic un-clustered GLM

Going to be seeing a lot of this:

- $y_i = \alpha + \beta x_i + \epsilon_i$

Where:

- y_i is continuous outcome measure: height, BMI, etc.
- β is risk associated with some kind of exposure
- $x_i \in [0, 1]$ is an indicator of exposure.
- α is expected outcome when x_i = 0
- ϵ_i are independently and identically distributed (i.i.d.) errors

Independent errors

Classic assumption is that:

- $\epsilon_i \sim N(0, \sigma^2)$

Independent errors

Classic assumption is that:

- $\epsilon_i \sim N(0, \sigma^2)$

In plain-ish English:

- Observation y_{ij} of individual i is a function of $\alpha + \beta x_i$ and normally distributed errors (ϵ_i) with mean zero and variance σ^2.

Independent errors

Classic assumption is that:

- $\epsilon_i \sim N(0, \sigma^2)$

In plain-ish English:

- Observation y_{ij} of individual i is a function of $\alpha + \beta x_i$ and normally distributed errors (ϵ_i) with mean zero and variance σ^2.

Another way of writing it:

- $y_i \sim N(\alpha + \beta x_i, \sigma^2)$

Which door will you choose?

Door #1: Ignore clustering and fit a normal GLM

- *Pool* data across all units, i.e. ignore clustering.
- i.e. fit model $y_{ij} = \alpha + \beta x_i + \epsilon_i$

Is this a good idea? Why or why not?

NO!

Complete pooling ignores potential sources of *observed* and *unobserved*. unit-level confounding.

Pooling clustered data violates assumption of independent errors

A pooled model:

$$y_i = \alpha + \beta x + \epsilon_i \tag{1}$$

Pooling clustered data violates assumption of independent errors

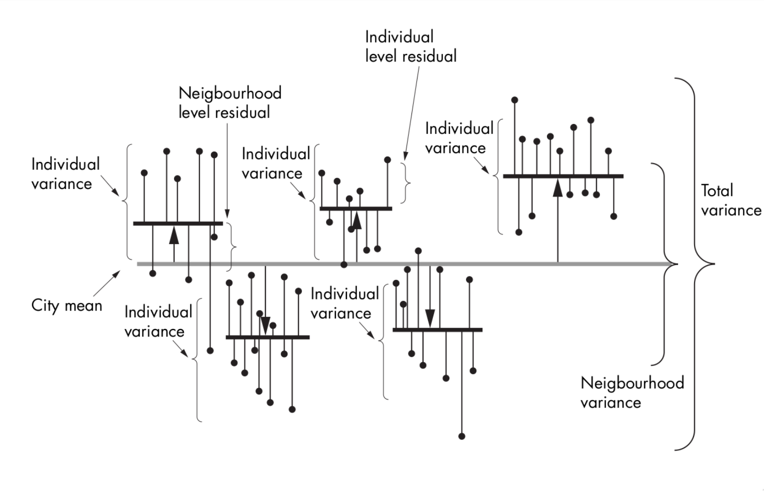
A pooled model:

$$y_i = \alpha + \beta x + \epsilon_i \tag{1}$$

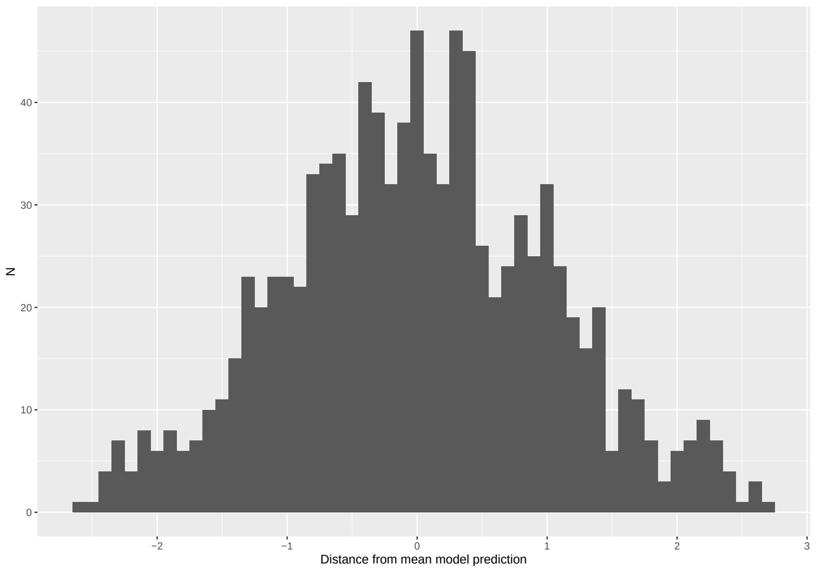
- y_i is a combination of systematic variation ($\alpha + \beta x$) and *uncorrelated* random noise (ϵ_i) where:

$$i.i.d.\ \epsilon \sim Normal(0, \sigma^2) \tag{2}$$

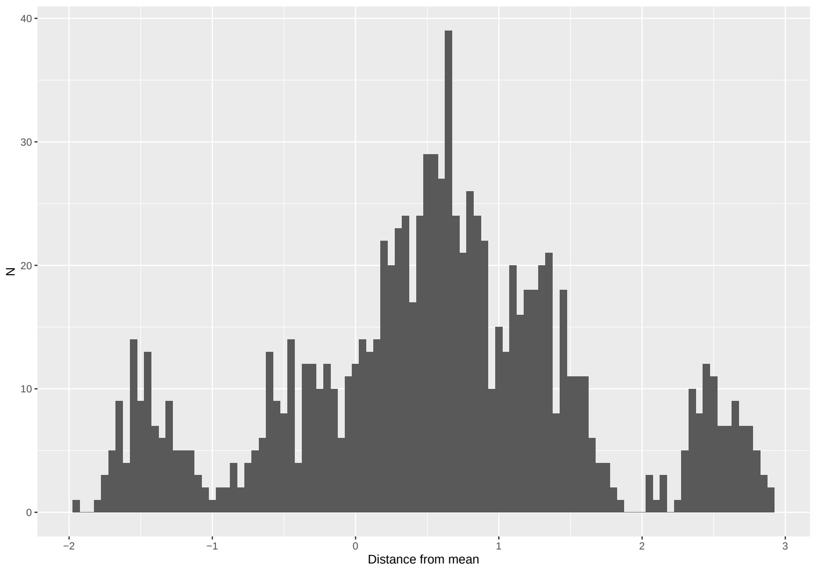
Clustering may result in correlation between average differences from mean



Your residuals should look like this



When you ignore clustering you may see something like:



Door #2: Fit a different model to each cluster

Fit *unpooled* model to each unit (j), assuming outcomes in each unit are independent:

- $y_{ij} = \alpha_j + \beta_j x_i + \epsilon_{ij}$
- $\epsilon_{ij} \sim N(0, \sigma_j^2)$

More danger!

Totally unpooled models run the risk of overfitting the data, particularly in small samples.

Specific dangers of unpooled models

What else could go wrong here?

Specific dangers of unpooled models

What else could go wrong here?

- Some units (e.g. counties) may have few observations, making *unpooled* models impractical

Specific dangers of unpooled models

What else could go wrong here?

- Some units (e.g. counties) may have few observations, making *unpooled* models impractical
- We may want to allow some effect of exposure (e.g. having a basement) to be consistent across counties.

Door #3: Partial Pooling!

- Allow effects to vary across clusters, but constrain them with a prior distribution.

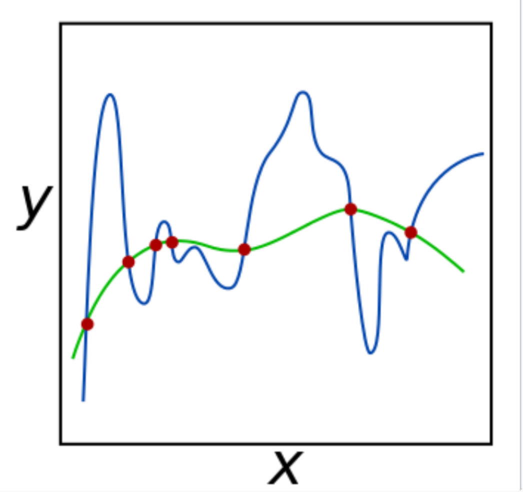
Door #3: Partial Pooling!

- Allow effects to vary across clusters, but constrain them with a prior distribution.
- This approach accommodates variation across units without assuming they have no similarity.

Door #3: Partial Pooling!

- Allow effects to vary across clusters, but constrain them with a prior distribution.
- This approach accommodates variation across units without assuming they have no similarity.
- More likely to make accurate out-of-sample predictions than the fully-pooled or unpooled examples.

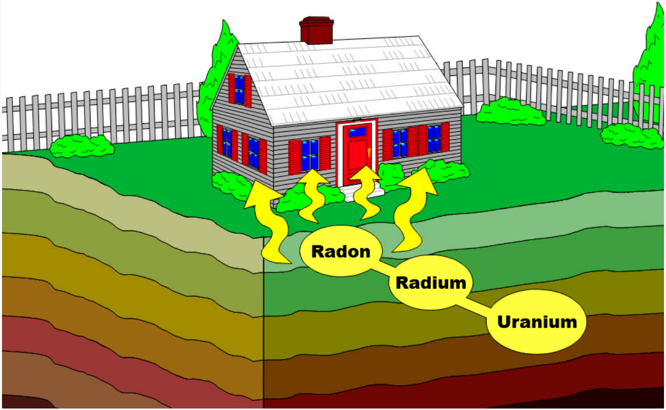
Partial pooling = Regularization



Both functions fit the data perfectly...which one should you prefer?

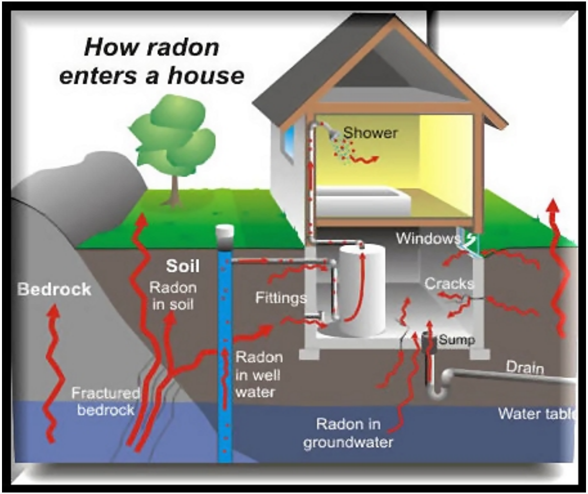
Radon Example

Radon is a carcinogenic gas



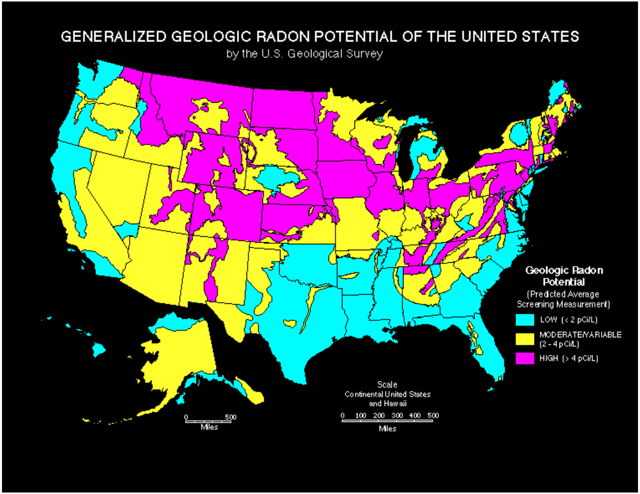
Radon is a byproduct of decaying soil uranium.

Radon enters a house more easily when it is built into the ground



Ann Arbor is a radon hotspot!

Considerable geographic variation in radon potential



Ann Arbor is a radon hotspot!

Trust me on this one...

My very own radon mitigation system.

What should a model that accounts for important sources of variation in household radon potential include?

What should a model that accounts for important sources of variation in household radon potential include?

- County-level variation in soil uranium.
- Whether or not the radon measurement was taken in a basement.

Random intercepts account for county-level variation

Gelman [@Gelman2006] proposes a multi-level model to measure household radon in household i in county j, y_{ij}:

- $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$

Random intercepts account for county-level variation

Gelman [@Gelman2006] proposes a multi-level model to measure household radon in household i in county j, y_{ij}:

- $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$, for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$

Where:

- α_j is average, non-basement radon measure at county level
- β is fixed effect measuring average change in radon level in houses with a basement.
- σ_y^2 represents within-county variation in risk

Include predictors of county-level variation in second level

County-level random intercept is a function of county soil uranium measure, u_j:

- $\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$, for $j = 1, \ldots, J$

Include predictors of county-level variation in second level

County-level random intercept is a function of county soil uranium measure, u_j:

- $\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$, for $j = 1, \dots, J$

Where:

- γ_0 is expected household radon measure when $u_j = 0$
- γ_1 scales expected county-level uranium with u_j
- σ_α^2 is between-county variation in radon risk not measured by u_j.

Putting it all together

County-level intercept is a function of county soil uranium measure, u_j:

- $\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$

Putting it all together

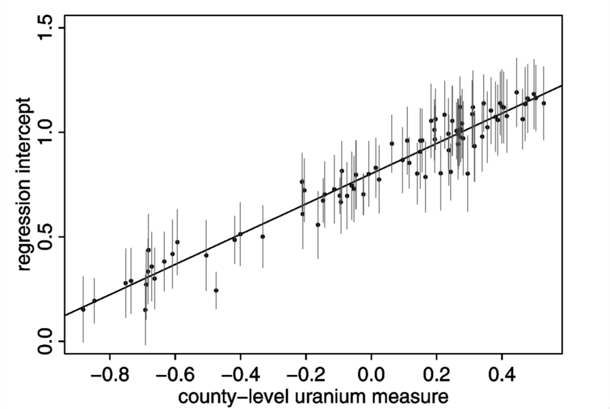
County-level intercept is a function of county soil uranium measure, u_j:

- $\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$

Household-level radon measure is a function of having a basement and county-level intercept:

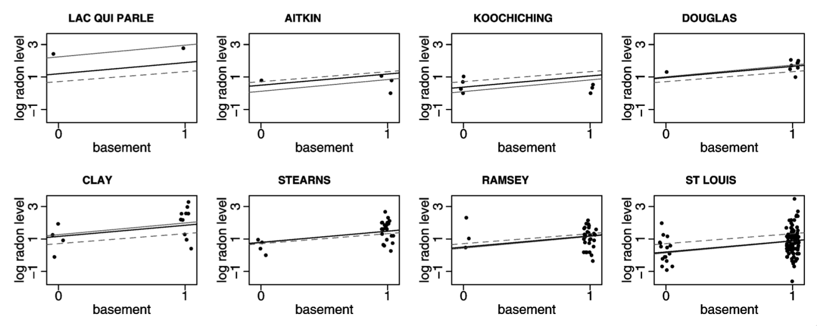
- $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$

County-level radon levels vary with soil uranium measures



County-level intercept, α_j, (± 1 standard error) as a function of county-level uranium.

Model predictions vs. radon measures by county



Multi-level regression line, $y = \alpha_j + \beta x$, from 8 Minnesota counties. Unpooled estimates = light grey line; Totally pooled estimates = dashed grey line.

Next Time

- Hands-on with the Radon example

References

References i