Models for clustered data

Jon Zelner

February 17, 2020

EPID 601
University of Michigan School of Public Health

jzelner@umich.edu
www.jonzelner.net

Agenda

- Re-intro to GLMs

Agenda

- Re-intro to GLMs
- Why?

Agenda

- Re-intro to GLMs
- Why?
- Radon!

Agenda

- Re-intro to GLMs
- Why?
- Radon!
- Self-assessments.

What is a GLM?

- GLMs are essentially a framework for mapping linear predictions to outcomes with any kind of distributions.

What is a GLM?

- GLMs are essentially a framework for mapping linear predictions to outcomes with any kind of distributions.
- Gives flexibility in the types of outcomes and inputs we can model.

What is a GLM?

- GLMs are essentially a framework for mapping linear predictions to outcomes with any kind of distributions.
- Gives flexibility in the types of outcomes and inputs we can model.
- Model non-linear response using linear predictors

Advanced GLM tricks

- Represent hierarchical/multi-level structure

Advanced GLM tricks

- Represent hierarchical/multi-level structure
- Model spatial autocorrelation

Advanced GLM tricks

- Represent hierarchical/multi-level structure
- Model spatial autocorrelation
- Allow variance parameters to change with individual-level covariates to acommodate heteroskedasticity

Working definition of a generalized linear model?

- Model variation in outcome, Y, as a function of covariates X

Working definition of a generalized linear model?

- Model variation in outcome, Y, as a function of covariates X
- Linear relationship between X and Y quantified by regression coefficients β

Working definition of a generalized linear model?

- Model variation in outcome, Y, as a function of covariates X
- Linear relationship between X and Y quantified by regression coefficients β
- Outcome assumed to have distribution with mean determined by regression coefficients.

Working definition of a generalized linear model?

- Model variation in outcome, Y, as a function of covariates X
- Linear relationship between X and Y quantified by regression coefficients β
- Outcome assumed to have distribution with mean determined by regression coefficients.
- Link function, $g()$, translates between linear predictor and the mean of the distribution function.

Some notation

- Individuals are indexed by i.

Some notation

- Individuals are indexed by i.
- Y is a vector of measured outcomes, composed of individual outcomes, so, $Y = \{y_1, y_2, y_3, \dots, y_N\}$, where N is the total number of observations

Some notation

- Individuals are indexed by i.
- Y is a vector of measured outcomes, composed of individual outcomes, so, $Y = \{y_1, y_2, y_3, \dots, y_N\}$, where N is the total number of observations
- Linear predictor $(\alpha + \beta x_i)$ for individual i denoted by y_i^*

Some notation

- Individuals are indexed by i.
- Y is a vector of measured outcomes, composed of individual outcomes, so, $Y = \{y_1, y_2, y_3, \dots, y_N\}$, where N is the total number of observations
- Linear predictor $(\alpha + \beta x_i)$ for individual i denoted by y_i^*
- Conditional mean for individual i denoted by $\widehat{y}_i = g(y_i^*)$

If you can map from predictors to parameters of outcome distribution, you can use a GLM with it...

- Gaussian for real-valued outcomes, $\mathbb{R} = \{-\infty, \infty\}$

If you can map from predictors to parameters of outcome distribution, you can use a GLM with it...

- Gaussian for real-valued outcomes, $\mathbb{R} = \{-\infty, \infty\}$
- Bernoulli for binary outcomes (0,1)

If you can map from predictors to parameters of outcome distribution, you can use a GLM with it...

- Gaussian for real-valued outcomes, $\mathbb{R} = \{-\infty, \infty\}$
- Bernoulli for binary outcomes (0,1)
- Binomial for repeated bernoulli trials ($\{0, 1, \dots, n\}$, where n is number of trials)

If you can map from predictors to parameters of outcome distribution, you can use a GLM with it...

- Gaussian for real-valued outcomes, $\mathbb{R} = \{-\infty, \infty\}$
- Bernoulli for binary outcomes (0,1)
- Binomial for repeated bernoulli trials ($\{0, 1, \dots, n\}$, where n is number of trials)
- Poisson for count data (The Natural numbers, $\mathbb{N} = \{0, 1, 2, ...\}$)

If you can map from predictors to parameters of outcome distribution, you can use a GLM with it...

- Gaussian for real-valued outcomes, $\mathbb{R} = \{-\infty, \infty\}$
- Bernoulli for binary outcomes (0,1)
- Binomial for repeated bernoulli trials ($\{0, 1, \dots, n\}$, where n is number of trials)
- Poisson for count data (The Natural numbers, $\mathbb{N} = \{0, 1, 2, ...\}$)

And others:

- Gamma, Exponential, Negative Binomial...

GLM with Normally-distributed errors

GLM with normally distributed errors $\approx$ OLS regression

What are assumptions of ordinary least squares regression?

GLM translates OLS into an explicitly probabalistic framework

Linear predictor for individual i is a function of her covariates:

- $y_i^* = \alpha + \beta x_i$

GLM translates OLS into an explicitly probabalistic framework

Linear predictor for individual i is a function of her covariates:

- $y_i^* = \alpha + \beta x_i$

Outcomes assumed to be on the real line from $-\infty$ to ∞, so:

- $g()$ is the identity link function

GLM translates OLS into an explicitly probabalistic framework

Linear predictor for individual i is a function of her covariates:

- $y_i^* = \alpha + \beta x_i$

Outcomes assumed to be on the real line from $-\infty$ to ∞, so:

- $g()$ is the identity link function
- $\widehat{y}_i = y_i^*$

GLM translates OLS into an explicitly probabalistic framework

Linear predictor for individual i is a function of her covariates:

- $y_i^* = \alpha + \beta x_i$

Outcomes assumed to be on the real line from $-\infty$ to ∞, so:

- $g()$ is the identity link function
- $\hat{y}_i = y_i^*$
- $y_i = \hat{y}_i + \epsilon_i$

GLM translates OLS into an explicitly probabalistic framework

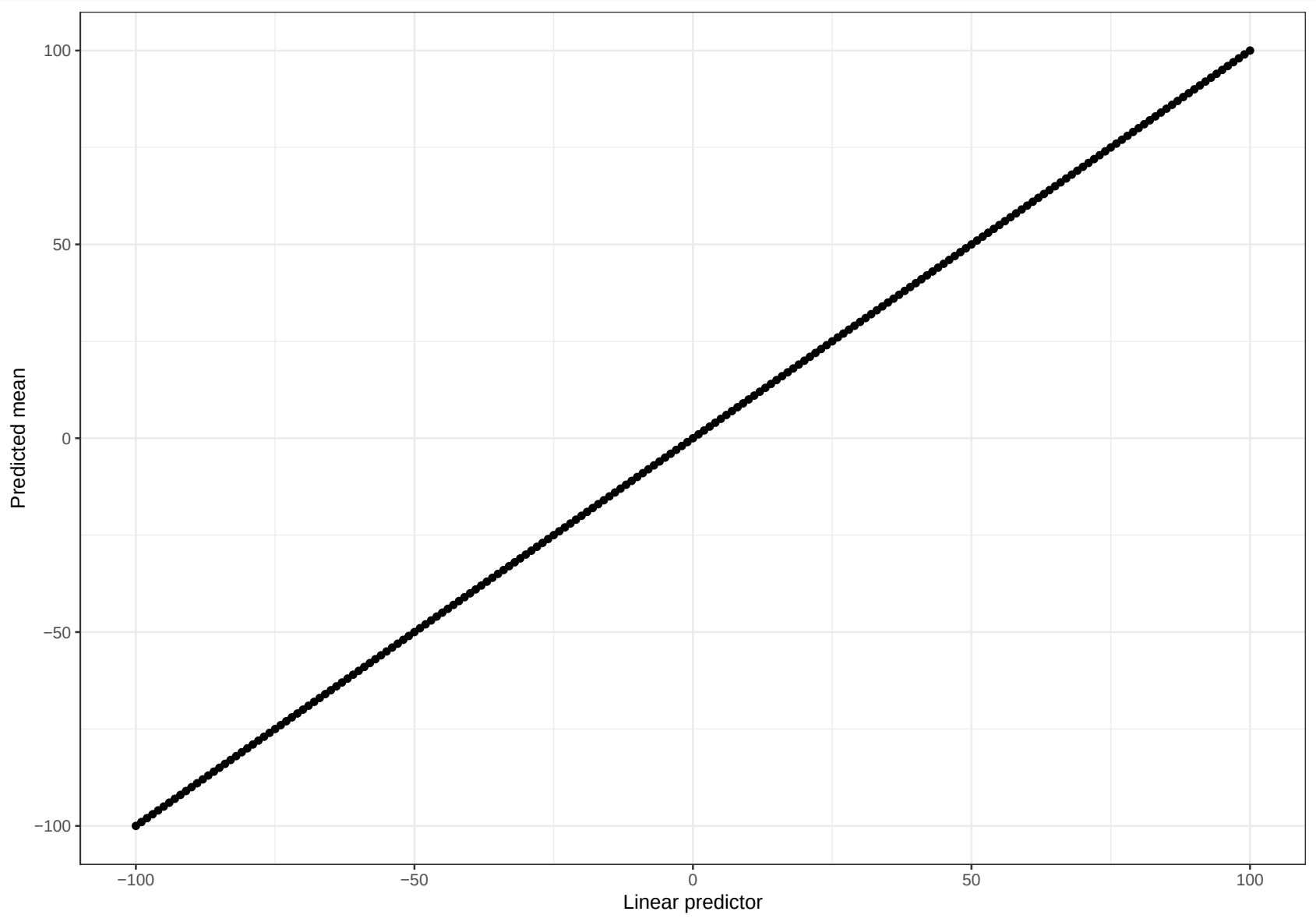
Linear predictor for individual i is a function of her covariates:

- $y_i^* = \alpha + \beta x_i$

Outcomes assumed to be on the real line from $-\infty$ to ∞, so:

- $g()$ is the identity link function
- $\hat{y}_i = y_i^*$
- $y_i = \hat{y}_i + \epsilon_i$
- $\epsilon_i \sim Normal(0, \sigma^2)$

Identity link function maps real numbers to real numbers



Interpreting results of Gaussian GLM

Interpreting results of Gaussian GLM

- α mean of y_i when $x_i = 0$

Interpreting results of Gaussian GLM

- α mean of y_i when $x_i = 0$
- β is change in mean of y_i for each one-unit change in x_i

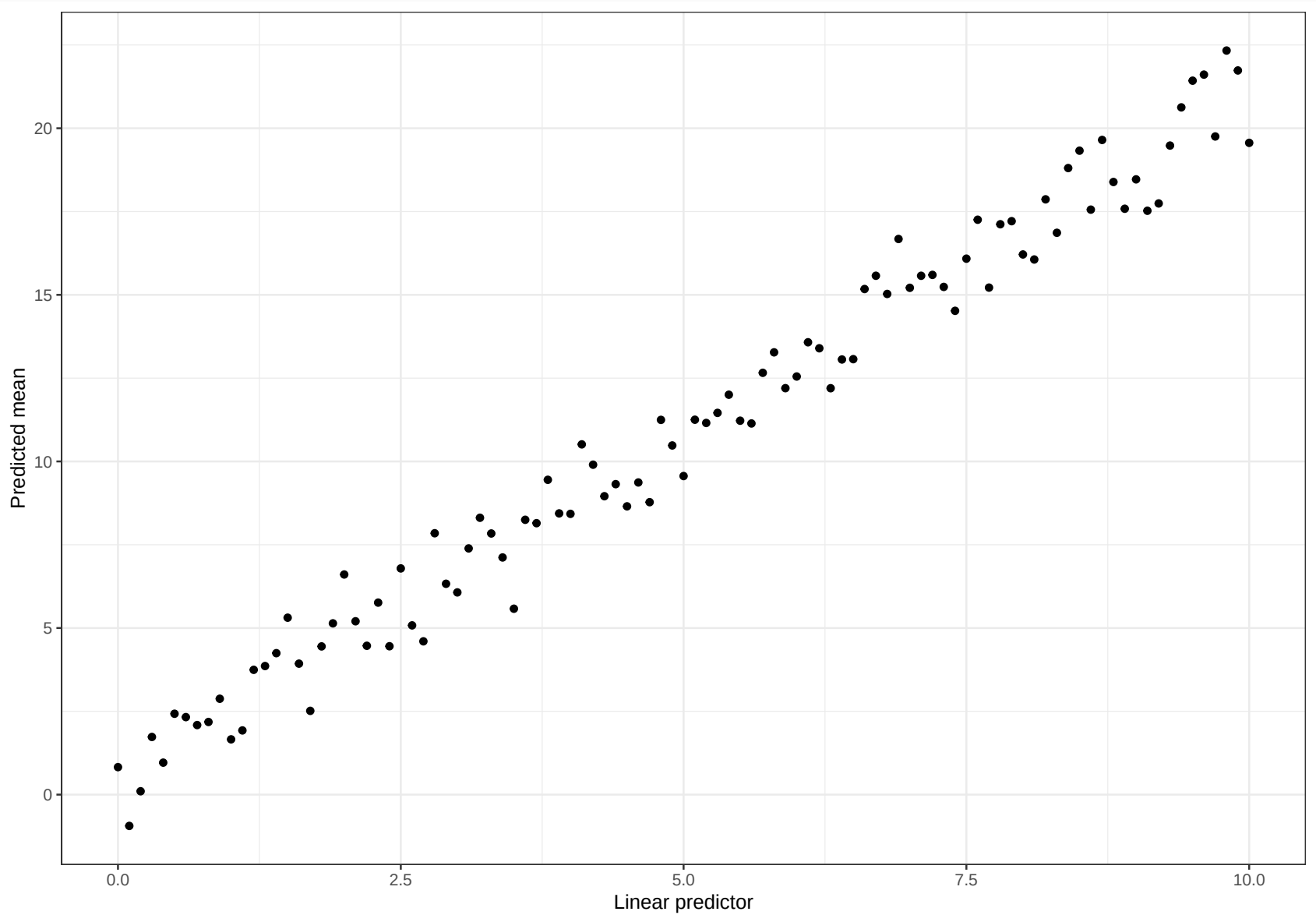
Interpreting results of Gaussian GLM

- α mean of y_i when $x_i = 0$
- β is change in mean of y_i for each one-unit change in x_i
- Just like in Merlo example, we also estimate the value of σ^2 that measures the amount of individual-level variability in outcomes

Generate some fake data in R

```r
## Predictors
x <- seq(from = 0.0, to = 10.0, by = 0.1)
## Regression coefficient
b <- 2.1
## Intercept
a <- 0.4
## Variance
sd <- 1.0
## The Data
y <- a + b * x + rnorm(length(x), 0, sd)
```

The Data



Recover the parameters

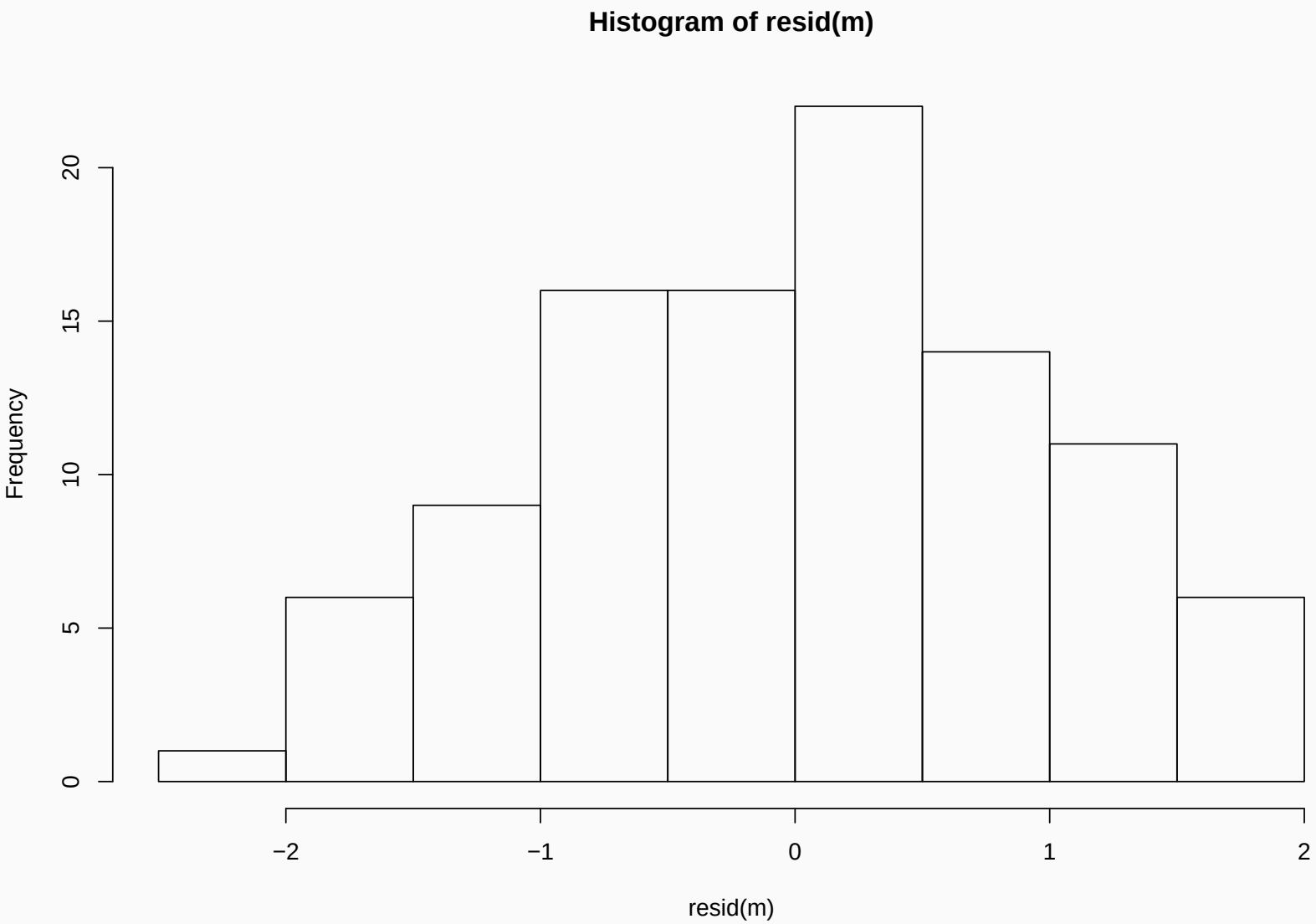
```
m <- glm(y ~ x, data = df)
```

Parameter estimates

Table 1: Fitting generalized (gaussian/identity) linear model: y ~ x

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.59	0.19	3.2	0.0019
x	2	0.032	64	2.5e-82

Model Residuals



Residual variance

Table 2: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	3604	3604	4091	2.5e-82
Residuals	99	87	0.88	NA	NA

How do we estimate GLM parameters and confidence bounds?

How do we estimate GLM parameters and confidence bounds?

- Maximum likelihood estimation

How do we estimate GLM parameters and confidence bounds?

- Maximum likelihood estimation
- Bayesian estimation, e.g. MCMC methods

How do we estimate GLM parameters and confidence bounds?

- Maximum likelihood estimation
- Bayesian estimation, e.g. MCMC methods
- We'll talk about these soon enough, but first...

How do we estimate GLM parameters and confidence bounds?

- Maximum likelihood estimation
- Bayesian estimation, e.g. MCMC methods
- We'll talk about these soon enough, but first...
- Both of these necessitate calculating the likelihood of the data.

What is likelihood?

What is likelihood?

Probability of observing:

What is likelihood?

Probability of observing:

- data y,

What is likelihood?

Probability of observing:

- data y,
- given model $f()$

What is likelihood?

Probability of observing:

- data y,
- given model $f()$
- and parameters, e.g. $\theta = \{\alpha, \beta\}$

What is likelihood?

Probability of observing:

- data y,
- given model $f()$
- and parameters, e.g. $\theta = \{\alpha, \beta\}$

For Gaussian GLM, f = Normal distribution, $\theta = \{\alpha, \beta, \sigma^2\}$

Motivations for Multilevel Modeling

Why?

- What is the distinction Gelman and Hill make between *hierarchical* and *multilevel* models?

Why?

- What is the distinction Gelman and Hill make between *hierarchical* and *multilevel* models?
- Did any of their motivations for multilevel modeling surprise you? What questions did it bring up?

Why?

- What is the distinction Gelman and Hill make between *hierarchical* and *multilevel* models?
- Did any of their motivations for multilevel modeling surprise you? What questions did it bring up?
- When might you want model *slopes* to vary and *intercepts* to be fixed?

Why?

- What is the distinction Gelman and Hill make between *hierarchical* and *multilevel* models?
- Did any of their motivations for multilevel modeling surprise you? What questions did it bring up?
- When might you want model *slopes* to vary and *intercepts* to be fixed?
- When might you want them to both vary?

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:

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

Three Approaches to Modeling Clustered Data

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

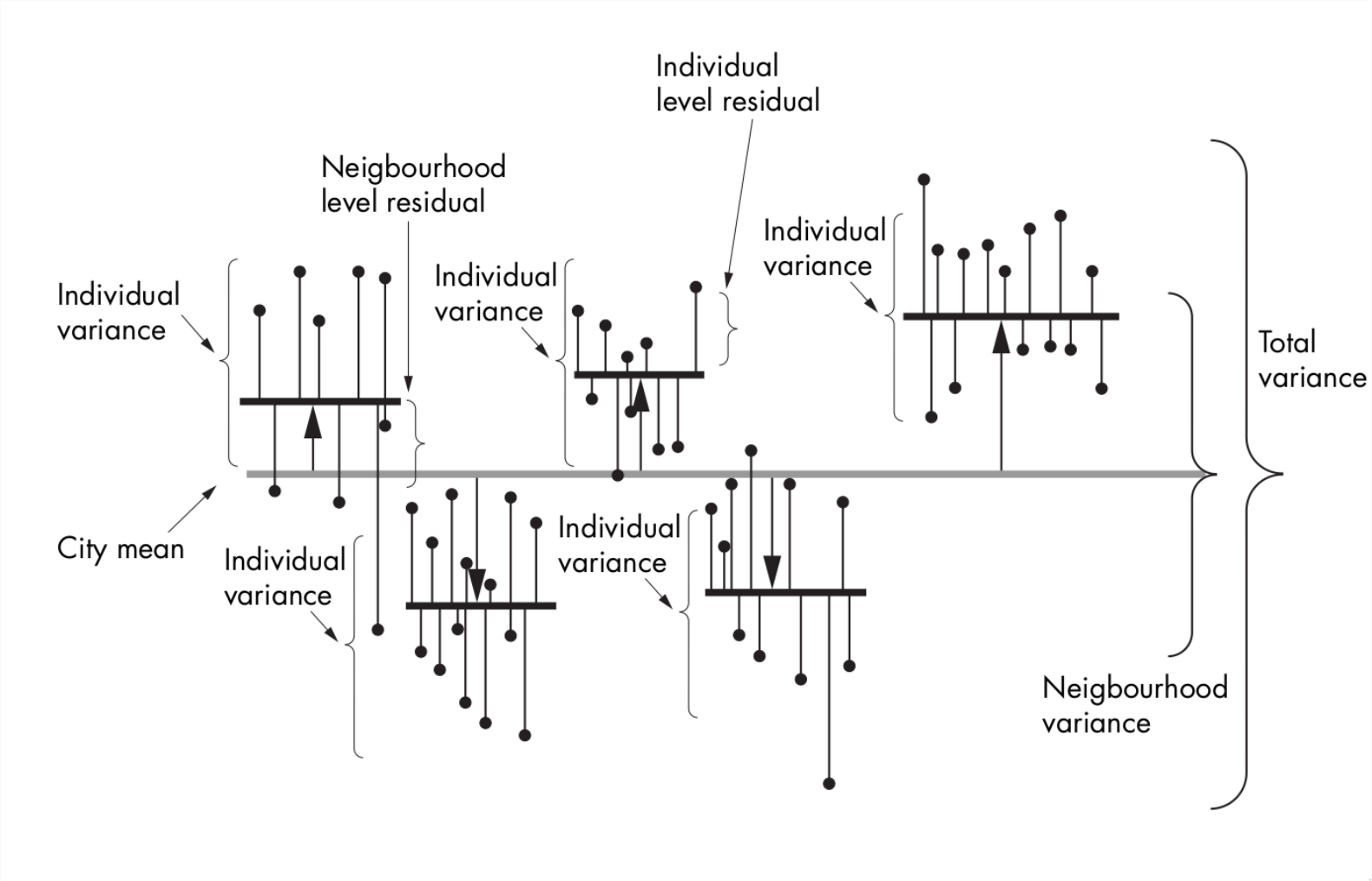
A pooled model:

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

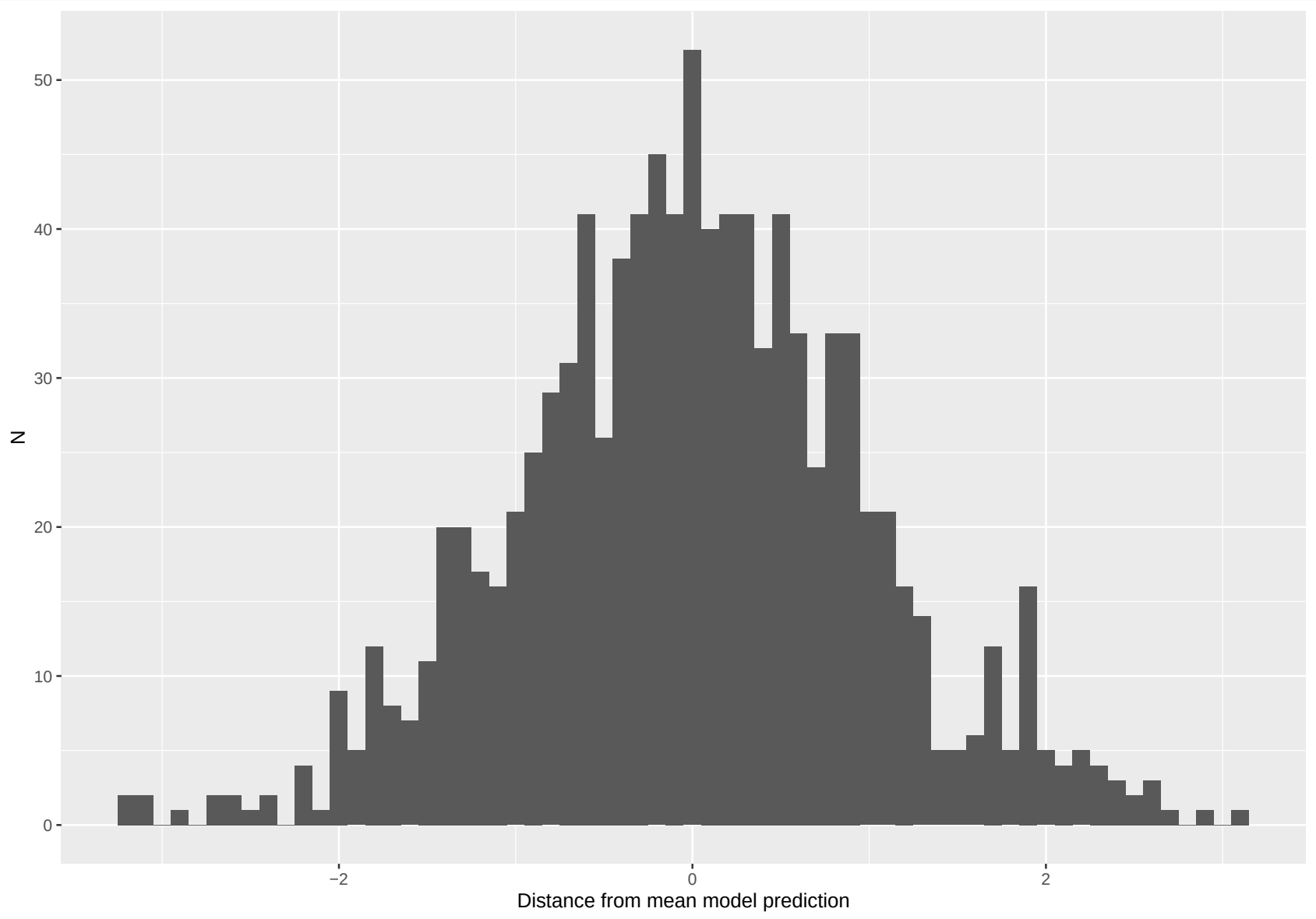
- y_i is a combination of systematic variation ($\alpha + \beta x$) and *uncorrelated* random noise (ϵ_i) where:

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

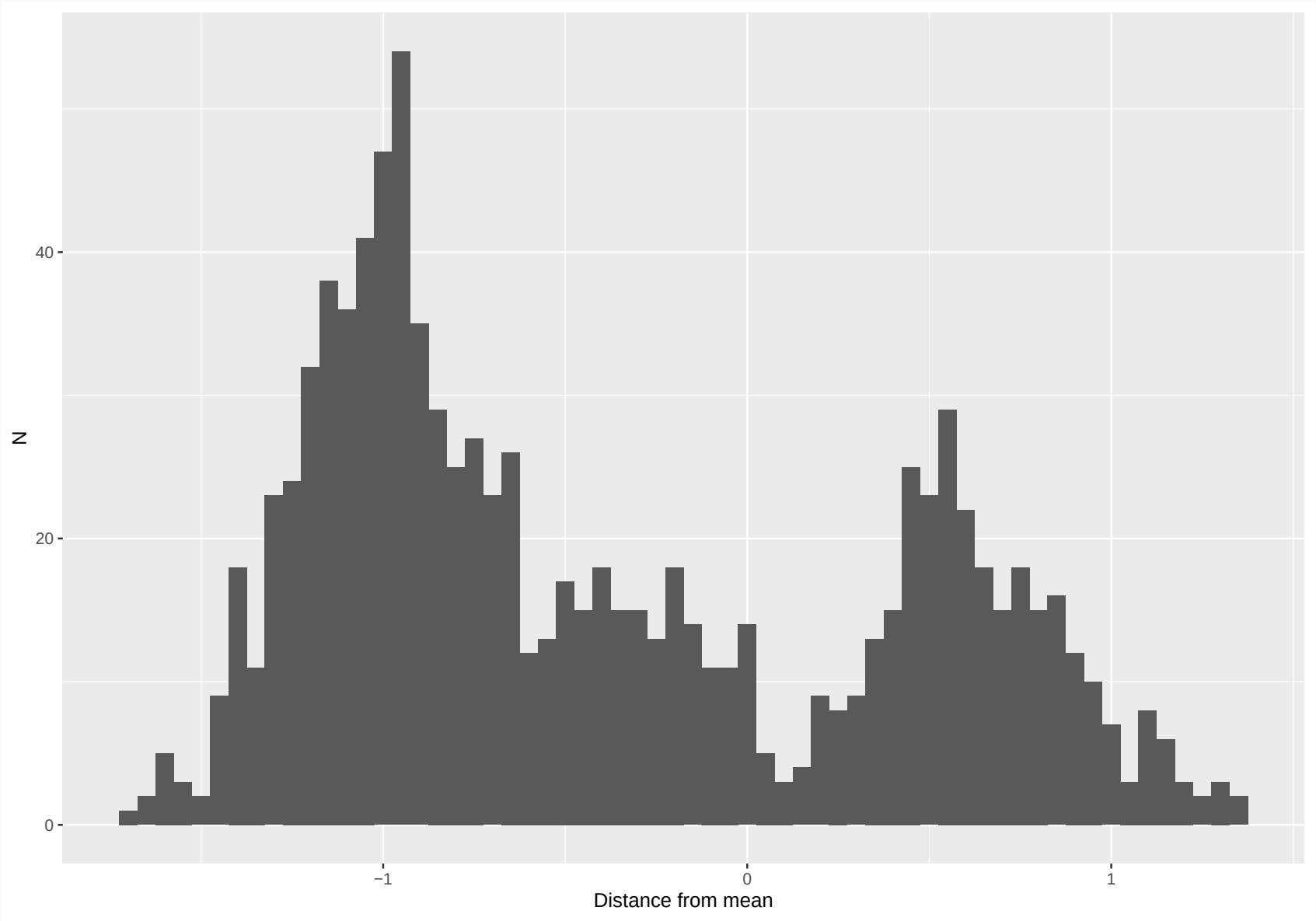
Clustering may result in correlation between average differences from mean



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.

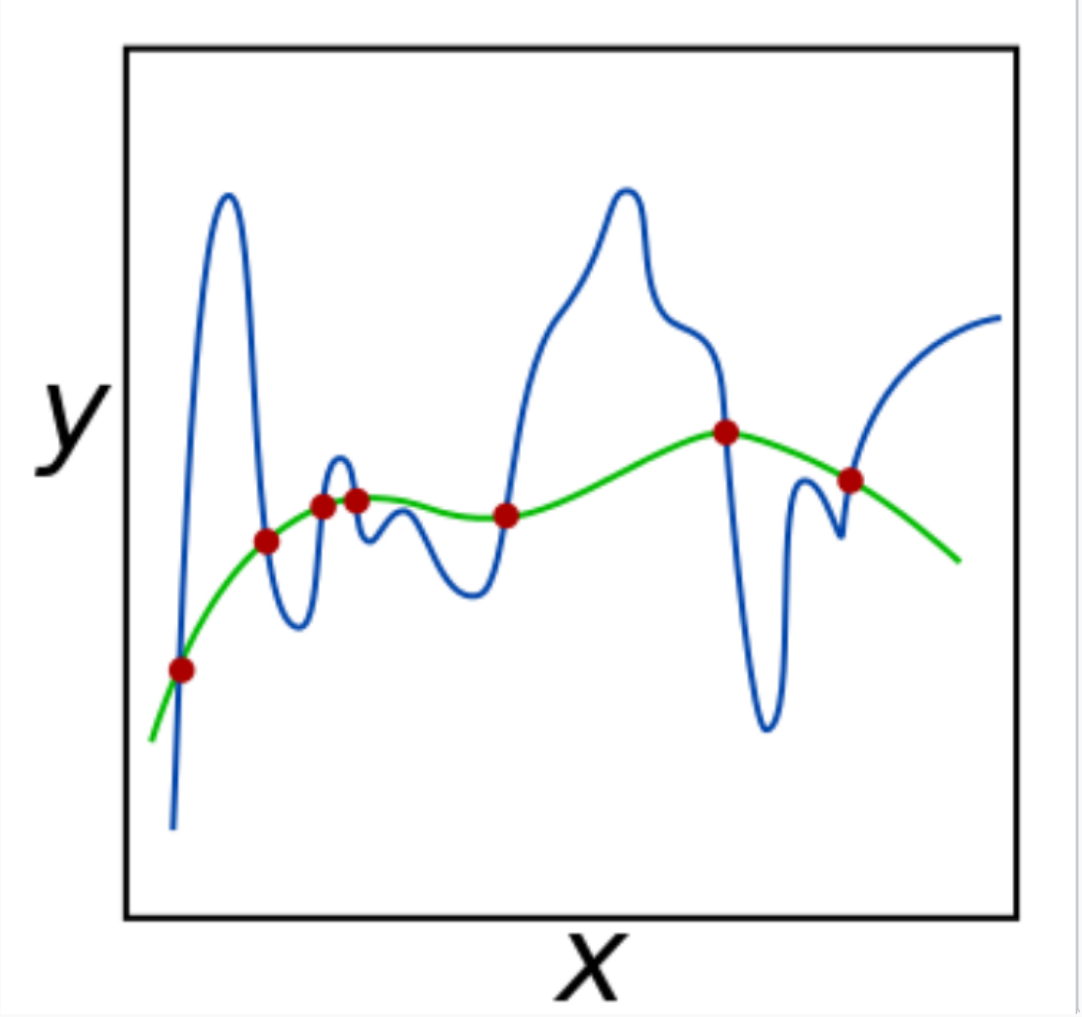
Door #3: Partial Pooling!

- 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.

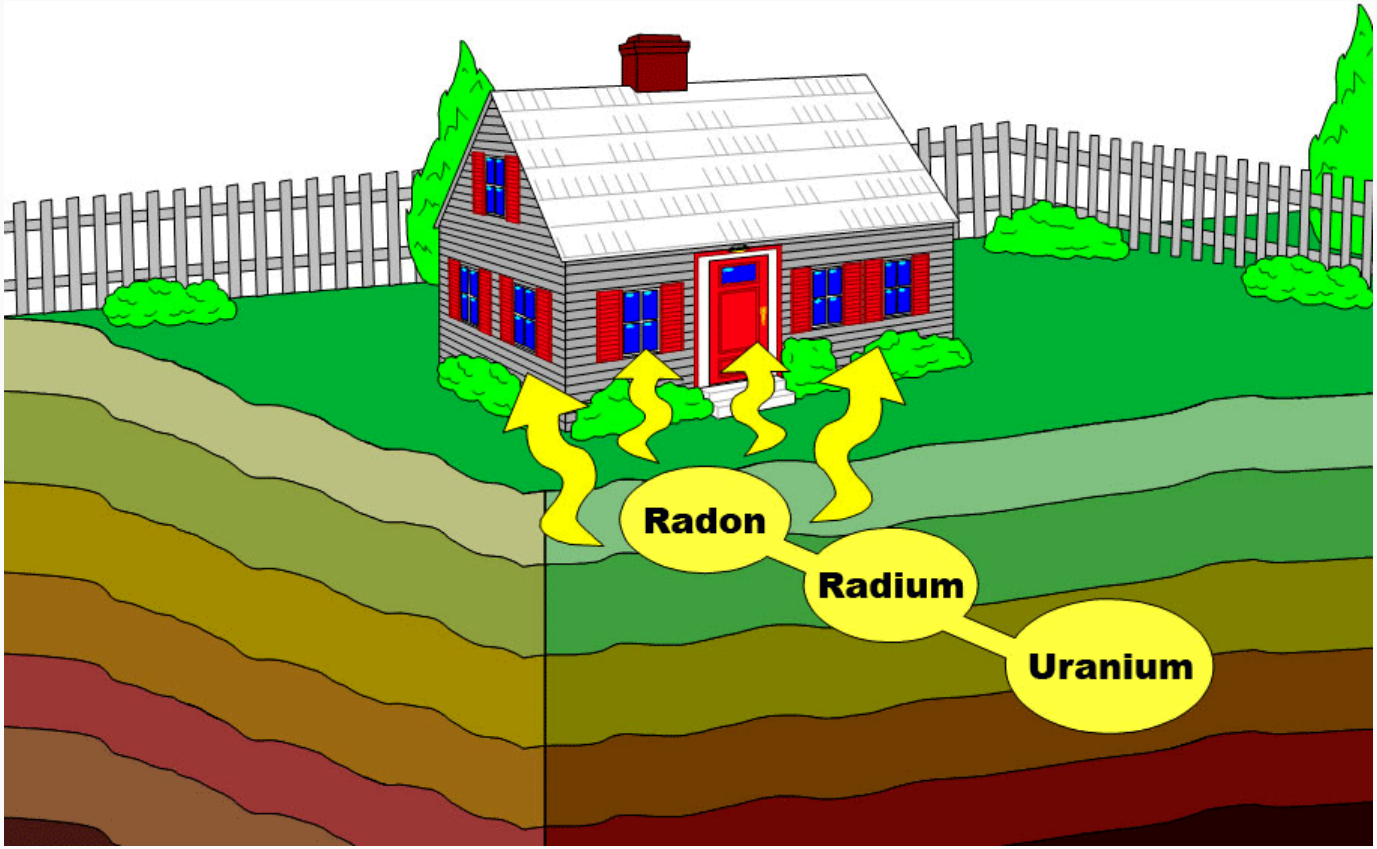
Partial pooling = Regularization



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

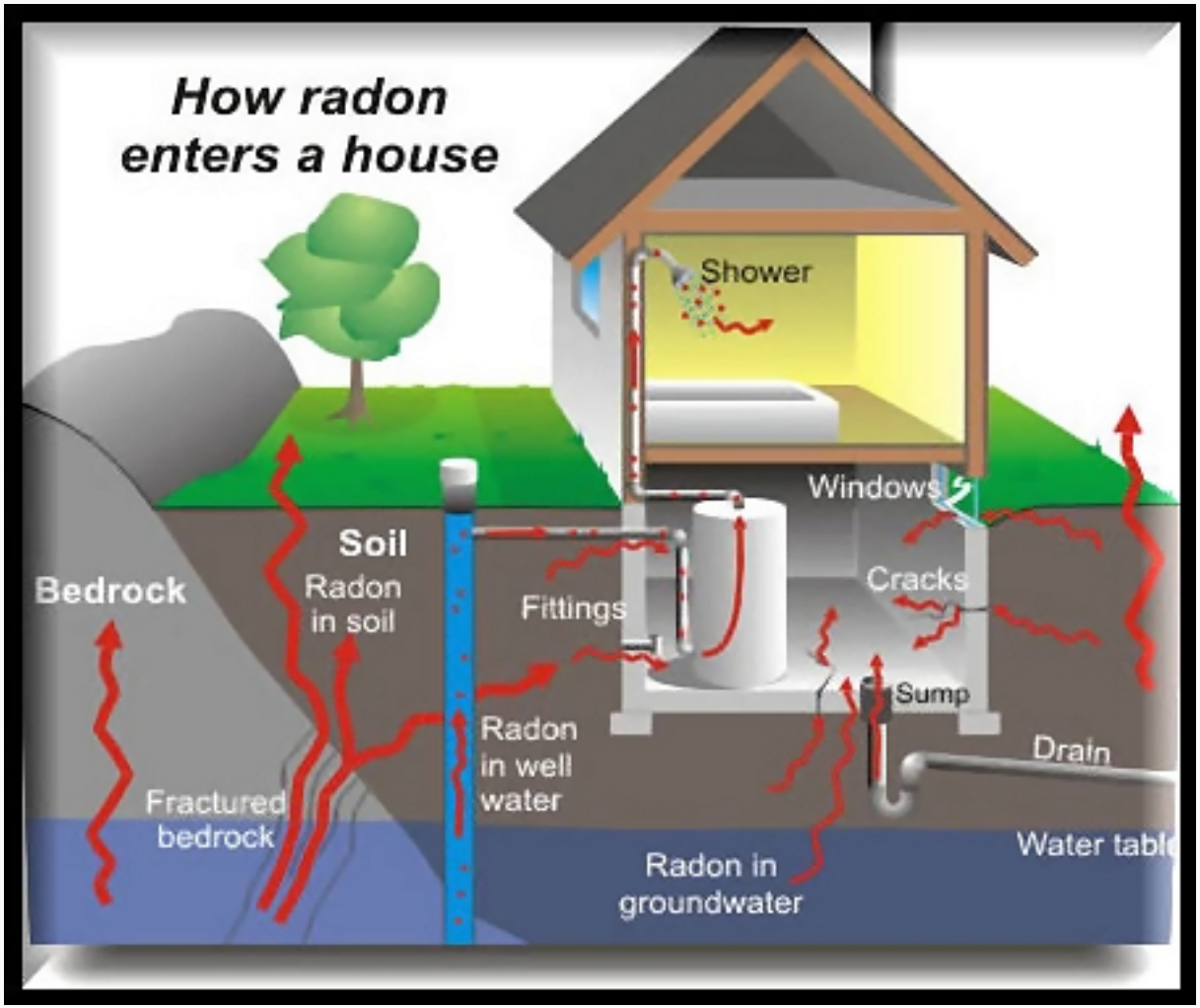
Radon Example

Radon is a carcinogenic gas



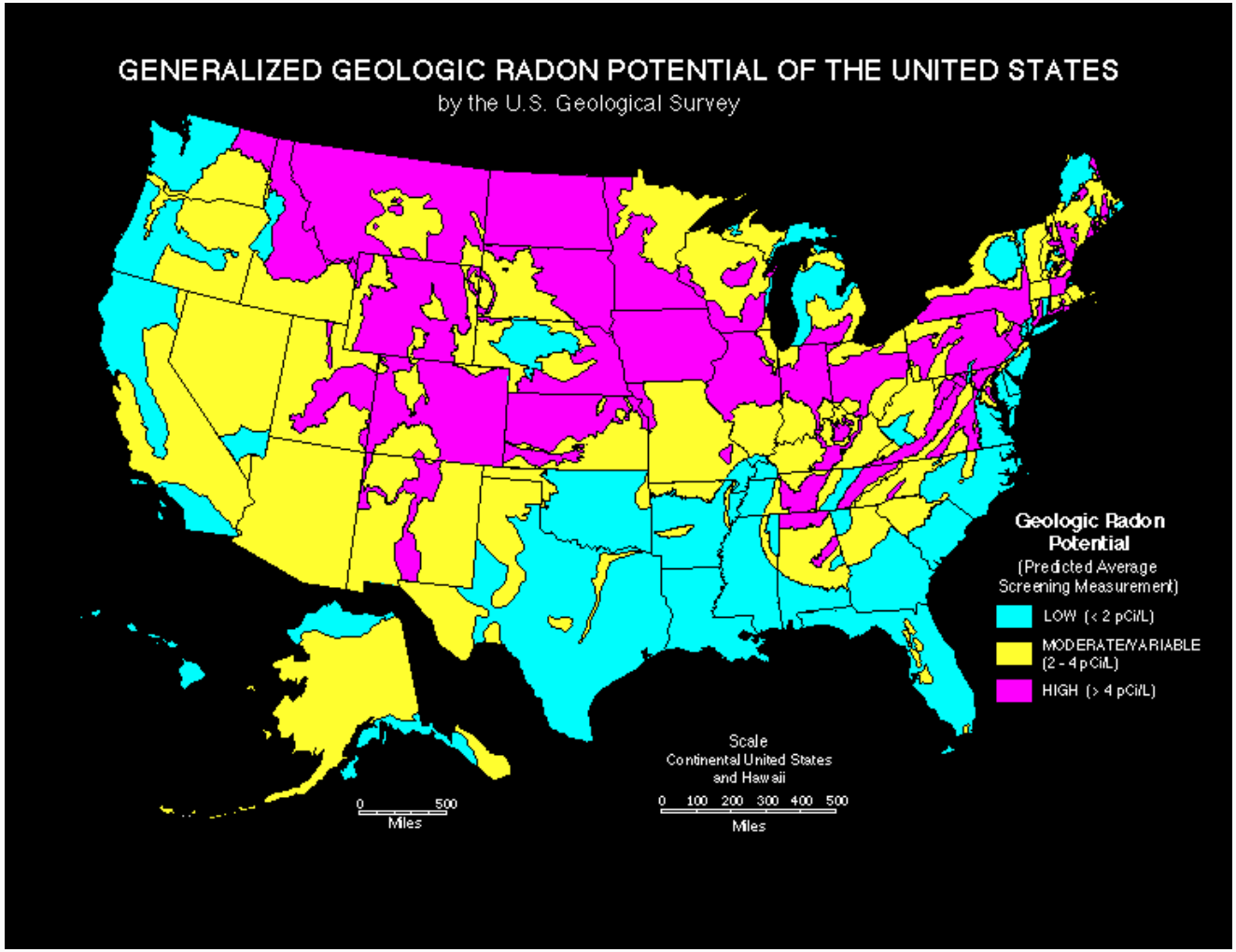
Radon is a byproduct of decaying soil uranium.

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 $i = 1, \dots, n_j$, $j = 1, \dots, J$

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 $i = 1, \dots, n_j$, $j = 1, \dots, J$

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:

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

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

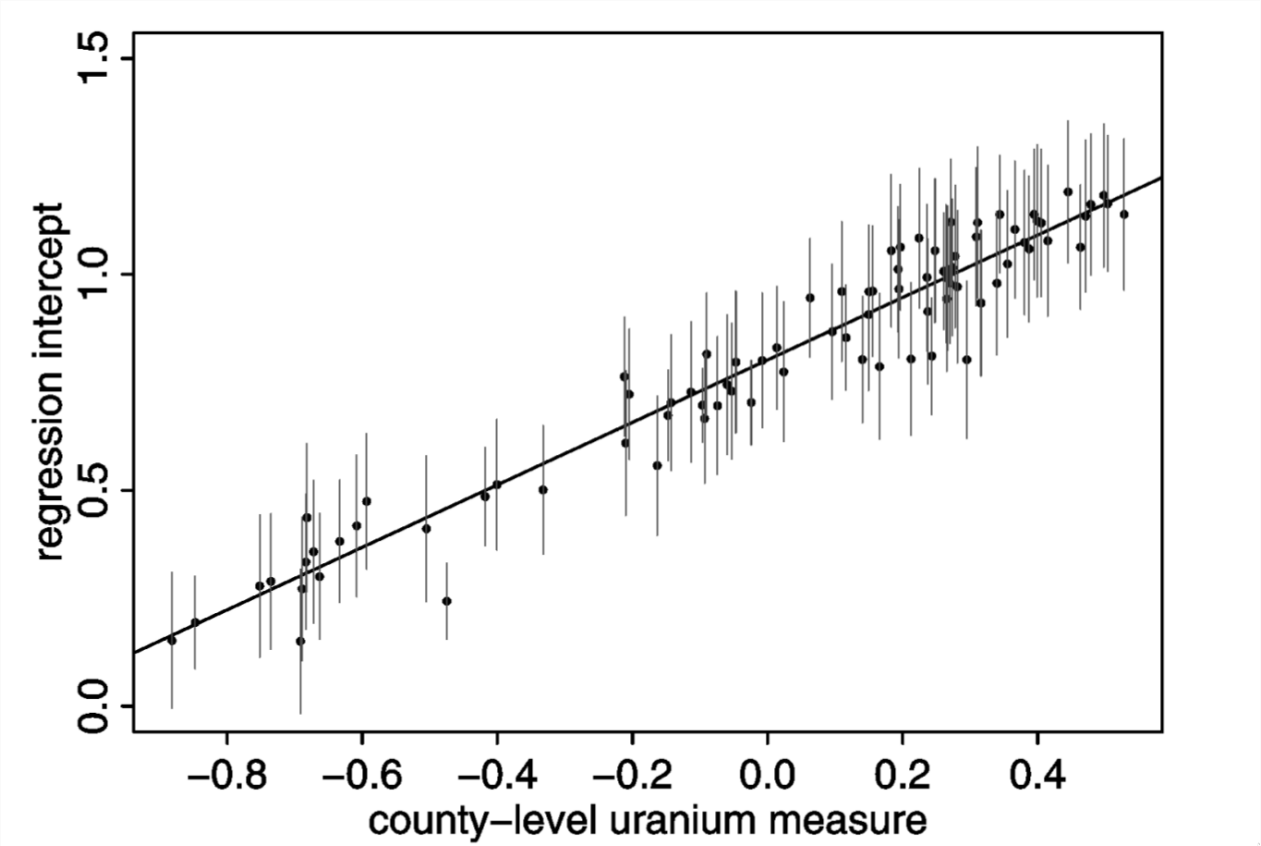
County-level intercept is a function of county soil uranium measure, u_j:

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

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

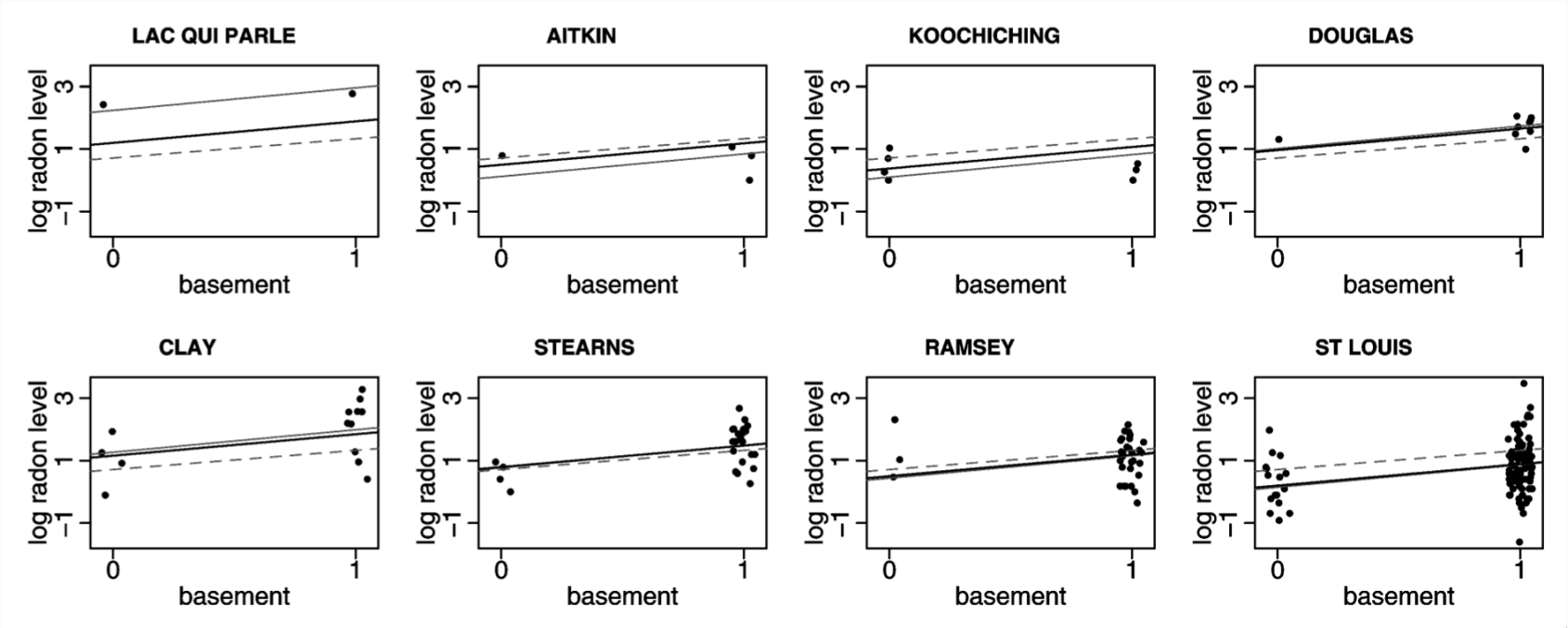
- $y_{ij} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2)$

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