Multilevel/Hierarchical Models

Lim Wai Yee, *The Hanging Gardens of Babylon*

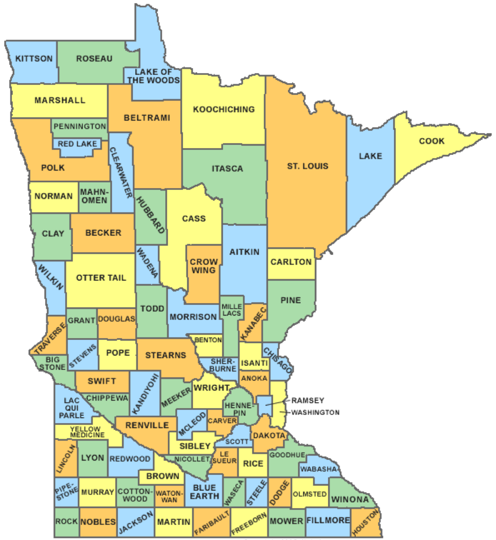
SML480: Pedagogy of Data Science, Spring 2020

Michael Guerzhoy

Case Study: Radon Levels in Minnesota

- Radon is a radioactive gas that is known to cause lung cancer, and is responsible for several thousands of lung cancer deaths per year in the US
- Radon levels vary in different homes, and also vary in different counties

Minnesota



Copyright 2005 digital-topo-maps.com

Minnesota counties

Goal

- Based on a limited set of measurements, want to know the average *log(radon level)* in each county

Complete Pooling

- Combine all the information from all the counties into a single “pool” of data
- Problem with complete pooling: the levels might differ for the different counties

No-Pooling Estimate

- Compute the average radon level for measurements in each county
- Compare pairs of counties using t-tests
- Equivalent to

```
    lm(log_radon~county, data=mn)
```

and looking at the coefficients for each county

No-Pooling Estimate: Problem

- We have just two data points for Lac Qui Parle, so we shouldn’t necessarily trust the data from there as much
- If we want to get at an estimate of the average log-radon level in Lac Qui Parle County, we probably want some kind of weighted average between what we observe in Lac Qui Parle County and the overall average

Multilevel Model

- Consider how the data is generated
- $y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$
- y_i is the i-th measurement
- j[i] is the county in which the i-th measurement was taken
- $\alpha_{j[i]}$ is the true log-radon level in county j[i]
- NEW:

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Estimate the best $\mu_\alpha, \sigma_\alpha^2$ from the data

Multilevel Model: Summary

$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$

Partial Pooling

$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Let $f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$
- (Approximate) Likelihood used by lme in R:

$$P\left(y_1, y_2, \dots, y_n \middle| \mu_\alpha, \sigma_y^2, \sigma_\alpha^2\right)$$
$$= (\Pi_j f\left(\alpha_j \middle| \mu_\alpha, \sigma_\alpha^2\right)) \left(\Pi_i f\left(y_i \middle| \alpha_{j[i]}, \sigma_y^2\right)\right)$$

- lme finds the $\alpha_j, \sigma_y^2, \mu_\alpha, \sigma_\alpha^2$ which maximize the likelihood
- Can now look at the different α_j

- (Look at R output)

Complete/Partial/No-Pooling

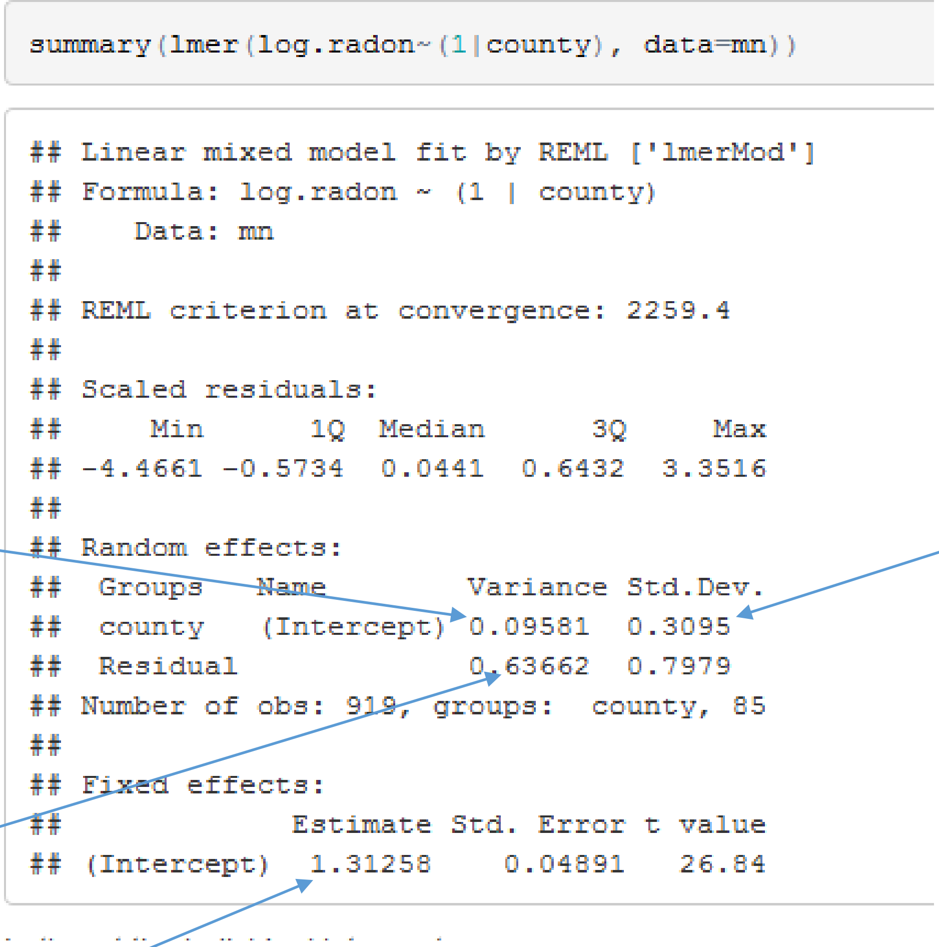
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$y_i \sim N\left(\alpha_{j[i]}, \sigma_y^2\right)$$

- No-Pooling: $\sigma_\alpha^2 = \infty$. That is, we assume that there is no connection at all between the log-radon levels in the different counties
  - `lm(log.radon~county, data=mn)`
- Complete pooling: $\sigma_\alpha^2 = 0$. Assume the true mean log-radon levels in all counties are the same
  - `lm(log.radon~1, data=mn)`
- Partial pooling: assume the mean log-radon levels are different in different counties, but their SD is σ_α (so they don’t differ by that much

R output

Random effects: coefficients that are *modelled* (i.e., generated by a distribution)
Fixed effects: coefficients that are note modelled

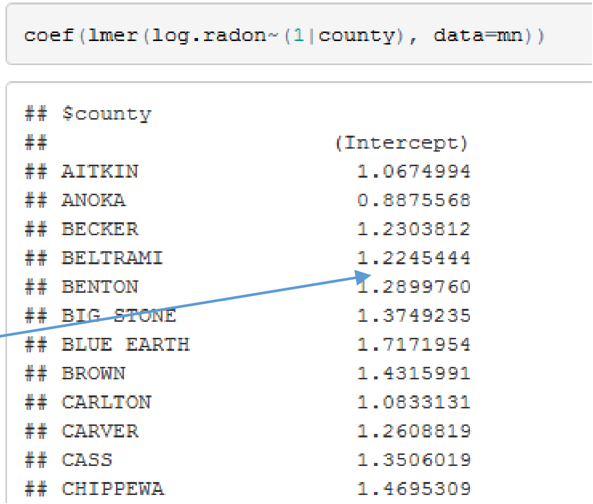
Note: the terminology is inconsistent in different places



```
summary(lmer(log.radon~(1|county), data=mn))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.radon ~ (1 | county)
##    Data: mn
##
## REML criterion at convergence: 2259.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.4661 -0.5734  0.0441  0.6432  3.3516
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  county   (Intercept) 0.09581  0.3095
##  Residual             0.63662  0.7979
## Number of obs: 919, groups:  county, 85
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.31258    0.04891   26.84
```

R output



```
coef(lmer(log.radon~(1|county), data=mn))
```

```
## $county
##                     (Intercept)
## AITKIN                1.0674994
## ANOKA                 0.8875568
## BECKER                1.2303812
## BELTRAMI              1.2245444
## BENTON                1.2899760
## BIG STONE             1.3749235
## BLUE EARTH            1.7171954
## BROWN                 1.4315991
## CARLTON               1.0833131
## CARVER                1.2608819
## CASS                  1.3506019
## CHIPPEWA              1.4695309
```

The approximate MLE's for the $\alpha_j's$ for the different counties

Complete/Partial/No-Pooling

- No-Pooling
  - Doesn’t share information between data points
  - Estimates for different counties will be completely different from each other
- Complete pooling
  - Fully shares information between data points
  - Estimates for the different counties are all the same
- Partial pooling
  - Tries to share information between data points in an optimal way
  - Estimates for different counties are generally closer together than for the no-pooling estimate

Partial pooling with Predictors

- Let’s use the floor predictor (x) as well
  - The floor on which the measurement was taken
- Simplest variant:

$$y_i \sim N\left(\alpha_{j[i]} + \beta x_i, \sigma_y^2\right)$$
$$\alpha_{j[i]} \sim N(\mu_\alpha, \sigma_\alpha^2)$$

- Advantage: better estimates for the levels for the various counties would lead to better estimates for the β
- Interpretation of β: keeping everything else constant, the increase in radon levels going up one floor
- Better estimate of β is obtained by partially pooling information when estimating $\alpha_{j[i]}$

Random Slopes

$$y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right)$$

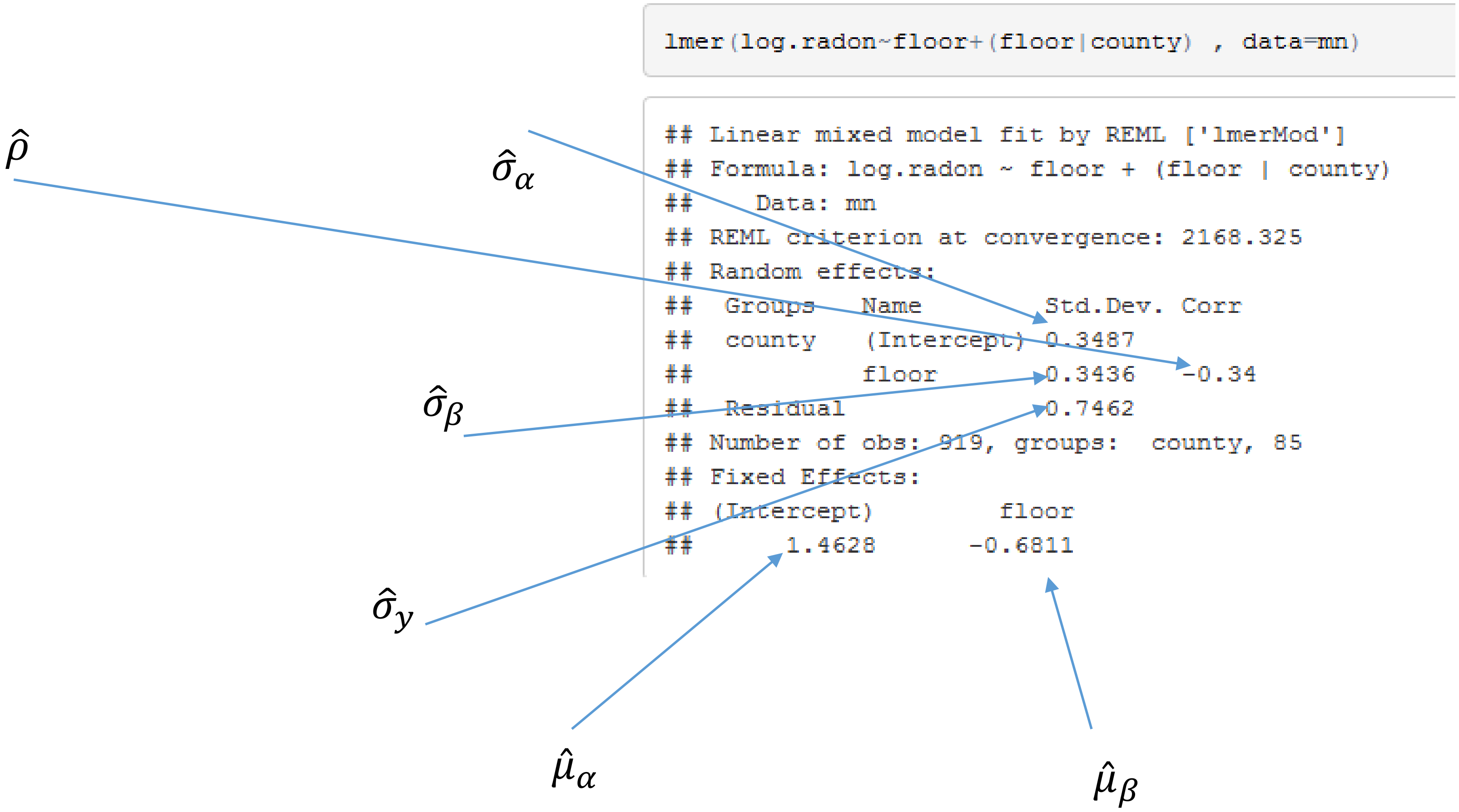
$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\alpha \\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right)$$

- Interpretation: in each county, the effect of moving one floor up on the radon levels is different
    - Perhaps in one county, the ceilings are 2.5m high, and in another county, the ceilings are 2.2m high
        - What is the effect of that on the βs?
- Rewrite:

$$y_i \sim N\left((\mu_\alpha + \alpha_{j[i]}) + (\mu_\beta + \beta_{j[i]}) x_i, \sigma_y^2\right)$$

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right)$$

R Output



```
lmer(log.radon~floor+(floor|county) , data=mn)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log.radon ~ floor + (floor | county)
##    Data: mn
## REML criterion at convergence: 2168.325
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  county   (Intercept) 0.3487
##           floor       0.3436   -0.34
##  Residual             0.7462
## Number of obs: 919, groups:  county, 85
## Fixed Effects:
## (Intercept)        floor
##      1.4628      -0.6811
```

Prediction for a new observation in an existing group

$$y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2\right)$$

- Know α, β, and x, want to predict new y
- Simulate multiple y’s from the distribution
- (in R)

Prediction for a new observation in a new group

- For each simulation,
- First, generate

$$\begin{pmatrix}\alpha_j \\ \beta_j\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\alpha \\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right)$$

- Next, generate the new data

$$y_i \sim N\left(\alpha_{j[i]} + \beta_{j[i]}x_i, \sigma_y^2\right)$$

Voting Patterns Across States in 2004 and Bayesian Statistics

- Red (Republican)/Blue (Democratic) state terminology only stabilized in 2000
  - Was reversed before that
- Many claims about cultural and economic differences between “blue states” and “red states”
  - Richer states voted Democratic, but rich voters voted Republican
  - In 2016: https://www.washingtonpost.com/news/politics/wp/2017/12/29/places-that-backed-trump-skewed-poor-voters-who-backed-trump-skewed-wealthier/?noredirect=on&utm_term=.23b4af301f4f

Bayesian Inference and Multilevel Models

- Same as before (e.g.)

$$\begin{pmatrix}\alpha_j\\ \beta_j\end{pmatrix}\sim N\left(\begin{pmatrix}\mu_\alpha\\ \mu_\beta\end{pmatrix},\begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta\\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right)$$

$$y_i\sim N\left(\alpha_{j[i]}+\beta_{j[i]}x_i,\sigma_y^2\right)$$

- Prior distributions on $\mu_\alpha,\mu_\beta,\sigma_\alpha$ …
- Obtain posterior distributions for $\mu_\alpha,\mu_\beta,\sigma_\alpha$, …

Voting patterns in different states

- $\text{Person } i \text{ in the sample}$
- x_i: income, on a -2…2 scale
- $s[i]$: the state where person i lives
- $y_i = 1$ if i voted Republican
- Baseline model: $P(y_i = 1) = logit^{-1}(a_{s[i]} + bx_i)$

Model Checking

- Fit the baseline model

$$P(y_i = 1) = logit^{-1}(a_{s[i]} + bx_i)$$

- Assess model fit by simulating new data from the model, and comparing the generated data to the actual data
  - If the model is a poor fit, the generated data will look different from the actual data
  - In this case, the constant b was a problem

Better model

- $P(y_i = 1) = logit^{-1}(a_{s[i]} + b_{s[i]}x_i)$
- Income influences voting patterns differently in different states
- Observation: in poorer states, the income influences voting patterns more
    - Gelman’s interpretation: the cultural contrasts that correlate with different voting patterns in different states are mostly differences between *rich people’s* cultural consumption patterns
    - See Andrew Gelman, *Red State, Blue State, Rich State, Poor State Why Americans Vote the Way They Do* (PU Press, 2009)

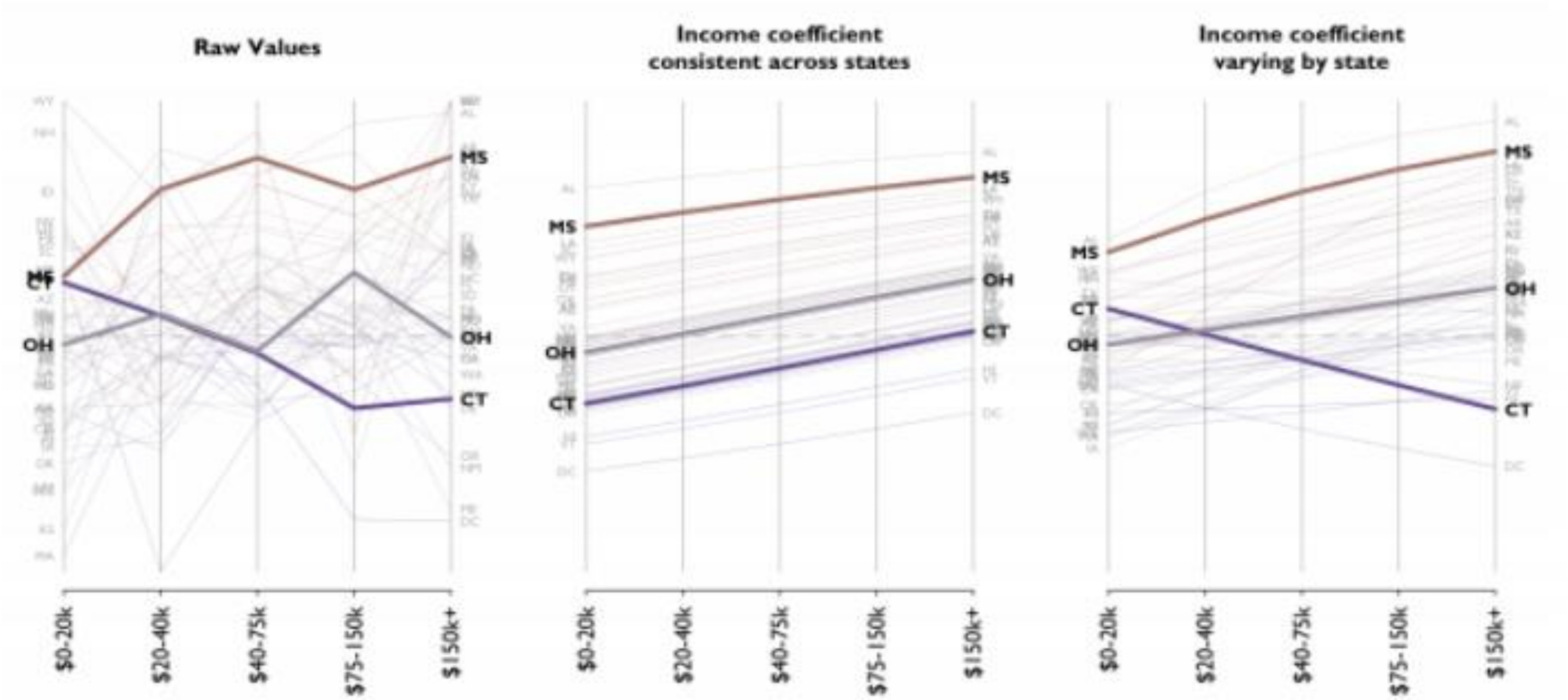


Figure 1: *The evolution of a simple model of vote choice in the 2008 election for* state × income *subgroups, non-Hispanic whites only. The colors come from the 2008 election, with darker shades of red and blue for states that had larger margins in favor of McCain or Obama, respectively. The first panel shows the raw data; the middle panel is a hierarchical model where state coefficients vary but the (linear) income coefficient is held constant across states; the right panel allows the income coefficient to vary by state. Adding complexity to the model reveals weaknesses in inferences drawn from simpler versions of the model. Three states—MS, OH, and CT—are highlighted to show important trends.*

(From Ghitza and Gelman, Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups (2012))

The "Usual" Story About Bayesian Inference

- Start with a prior distribution and a model, get some data, get a posterior distribution for the model parameters
- Everything you need to know is contained in the posterior distribution

A View of Bayesian Modelling

- Fit increasingly complex models to the data until the model fit the data
- Use priors for coefficients that help the posterior prediction be smooth if there is too little data
  - E.g., for the Radon levels example, make the prior for σ_α small enough
- Do model checking using predictive simulation: fake data generated by the model should look like real data