TOOLS FOR UNDERSTANDING MULTILEVEL REGRESSION MODELS

Iain Pardoe

Lundquist College of Business
University of Oregon

April 8, 2005

ipardoe@lcbmail.uoregon.edu

http://lcb1.uoregon.edu/ipardoe/

Joint work with Andrew Gelman, Columbia University
and S V Subramanian, Harvard School of Public Health

Presentation Outline

1. Application: income inequality and health in the U.S.
2. Multilevel regression modeling
3. Fitting a multilevel model
4. Displaying/understanding multilevel mean parameters
5. Displaying/understanding multilevel variance parameters
6. Diagnostics for model checking

1. Application: income inequality and health in the U.S.

- Poverty: risk factor for premature mortality/increased morbidity
- But, does unequal income distribution in a society pose an *additional* hazard to individual health within that society? (Subramanian and Kawachi, 2004)
- Individual-level model measures relation between income and health for individuals across all 50 states
- But, this ignores possibility that health outcomes within states are correlated (due to income inequality, say)
- State-level model measures relation between income inequality and aggregate (societal) health
- But, this fails to control for individual-level effects

Multilevel model: combines individual and state models

- 2002 Current Population Survey, 2000 Census
- $y_{ij} \sim$ Bernoulli(p_{ij}), i : individual, j : state
  $y = 1$ fair/poor health vs. $y = 0$ excellent/v.good/good health
- logit$(p_{ij}) = \alpha_j + \beta^T X_{ij}$ $(X_{ij}$ are individual-level predictors$)$
  - equivalized income categories
  - controls: age, gender, marital status, race, education, insurance
- $\alpha_j \sim$ N$(\gamma^T G_j, \sigma^2)$ $(G_j$ are state-level predictors$)$
  - median household income
  - Gini coefficient (income inequality)

2. Multilevel regression modeling

- Huge potential for multilevel models (MLMs) to improve undestanding of the world around us
- But, formulating, fitting, and understanding MLMs remains difficult
- Goal: emulate linear regression
  - idea (least squares)
  - algorithm (software)
  - choice of model specification
  - diagnostics
  - ways to understand results of fitted model

3. Fitting a multilevel model

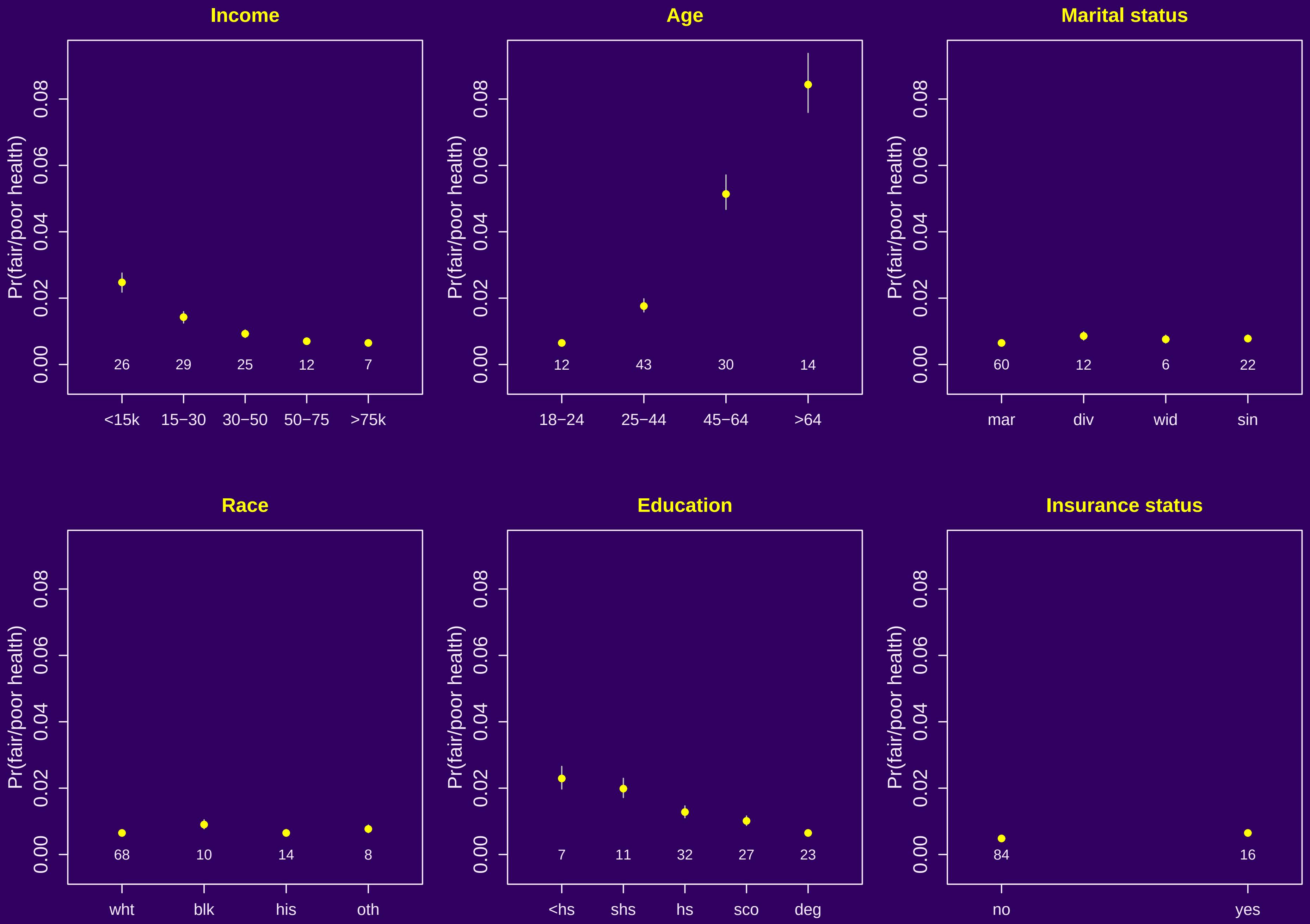
- Our approach: Bayesian inference using MCMC (Gibbs sampler and Metropolis algorithm)
- Software: R and Bugs
- Other software: MLwiN (IGLS/RIGLS or MCMC), HLM (ML)
- Issues:
  - redundant parameterization for Gibbs sampler
  - weakly-informative prior distributions
  - notation (depends on the model)
  - viewing categorical predictors as latent data

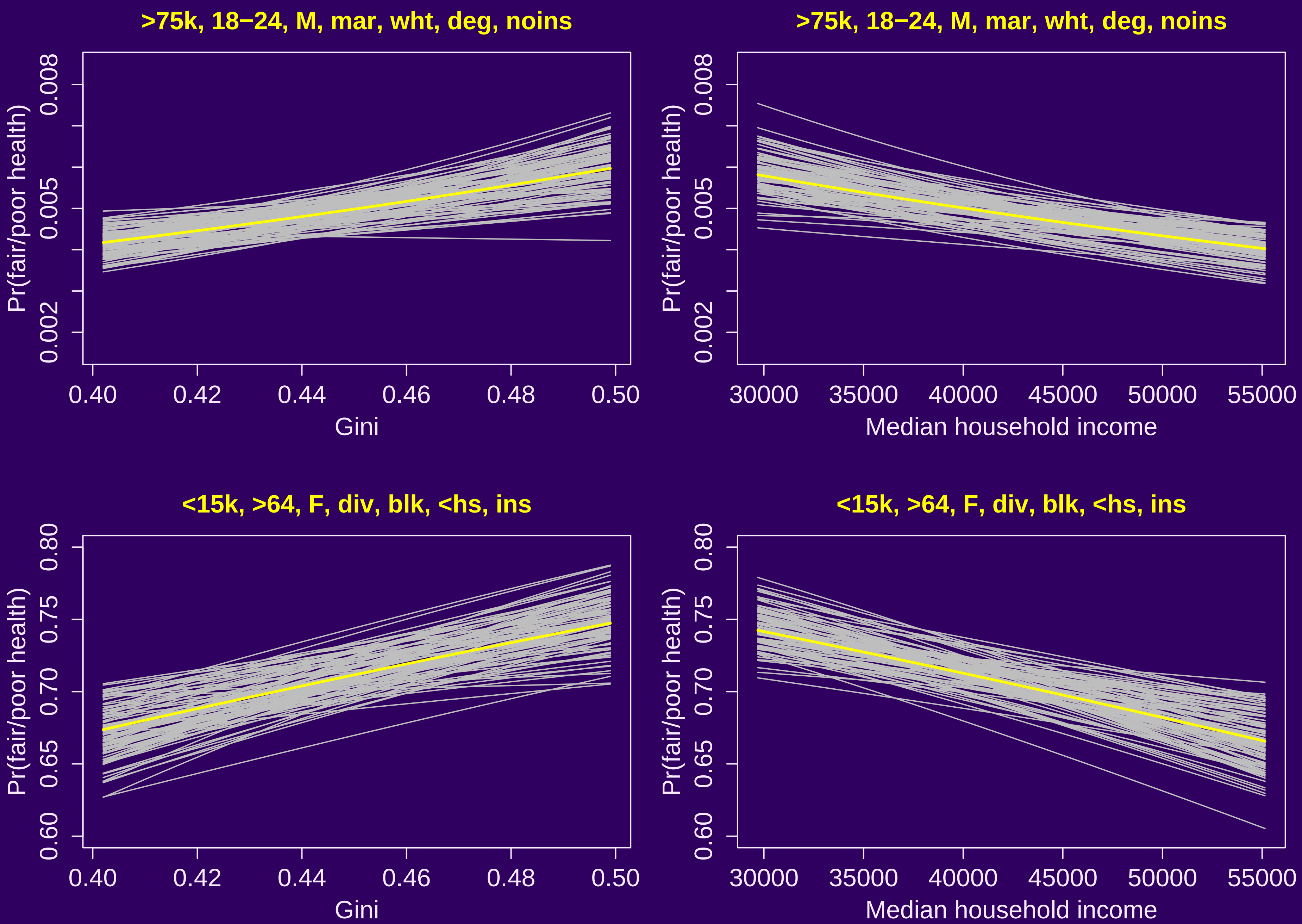
Predictor	Estimate	Std. error	95% interval
Intercept	−5.0	0.1	−5.1, −4.9
Gini	0.1	0.0	0.0, 0.1
Median income	−0.1	0.0	−0.1, 0.0
Age1 (25-44)	1.0	0.0	0.9, 1.1
Age2 (45-64)	2.1	0.0	2.0, 2.2
Age3 ($>$64)	2.6	0.1	2.6, 2.8
Divorced	0.3	0.0	0.2, 0.3
Widowed	0.2	0.0	0.1, 0.2
Single	0.2	0.0	0.1, 0.2
Black	0.3	0.0	0.3, 0.4
Hispanic	0.0	0.0	−0.1, 0.1
Other	0.2	0.0	0.1, 0.2

Predictor	Estimate	Std. error	95% interval
Female	0.0	0.0	0.0, 0.1
Some college	0.4	0.0	0.4, 0.5
High school	0.7	0.0	0.6, 0.7
Some hi-sch	1.1	0.0	1.1, 1.2
< hi-sch	1.3	0.0	1.2, 1.4
Inc1 (50-75k)	0.1	0.1	0.0, 0.2
Inc2 (30-50k)	0.4	0.0	0.3, 0.5
Inc3 (15-30k)	0.8	0.0	0.7, 0.9
Inc4 (<15k)	1.4	0.0	1.3, 1.4
Uninsured	−0.3	0.0	−0.4, −0.2

4. Displaying/understanding multilevel mean parameters

- Extend ideas in Gelman et al. (2002):
  - summarize inferences by simulations
  - graph regression lines, uncertainties, and data
  - graph group-level regression lines, uncertainties, and group-level estimates
- Average predictive effects (Gelman and Pardoe, 2005a): change in response change as predictor changes, averaged over predictor distribution (incorporating parameter uncertainty)
  - works for nonlinear mean functions, interactions, and variance components (see also Gelman, 2005)
  - graphs of average predictive effects (Pardoe and Shor, 2005)


Income
Pr(fair/poor health)
0.00
0.02
0.04
0.06
0.08
26 29 25 12 7
<15k 15−30 30−50 50−75 >75k
Age
12 43 30 14
18−24 25−44 45−64 >64
Marital status
60 12 6 22
mar div wid sin
Race
68 10 14 8
wht blk his oth
Education
7 11 32 27 23
<hs shs hs sco deg
Insurance status
84 16
no yes


>75k, 18−24, M, mar, wht, deg, noins
Pr(fair/poor health)
0.002
0.005
0.008
0.40
0.42
0.44
0.46
0.48
0.50
Gini
>75k, 18−24, M, mar, wht, deg, noins
Pr(fair/poor health)
0.002
0.005
0.008
30000
35000
40000
45000
50000
55000
Median household income
<15k, >64, F, div, blk, <hs, ins
Pr(fair/poor health)
0.60
0.65
0.70
0.75
0.80
0.40
0.42
0.44
0.46
0.48
0.50
Gini
<15k, >64, F, div, blk, <hs, ins
Pr(fair/poor health)
0.60
0.65
0.70
0.75
0.80
30000
35000
40000
45000
50000
55000
Median household income

Predictive effect (PE) for an input variable:

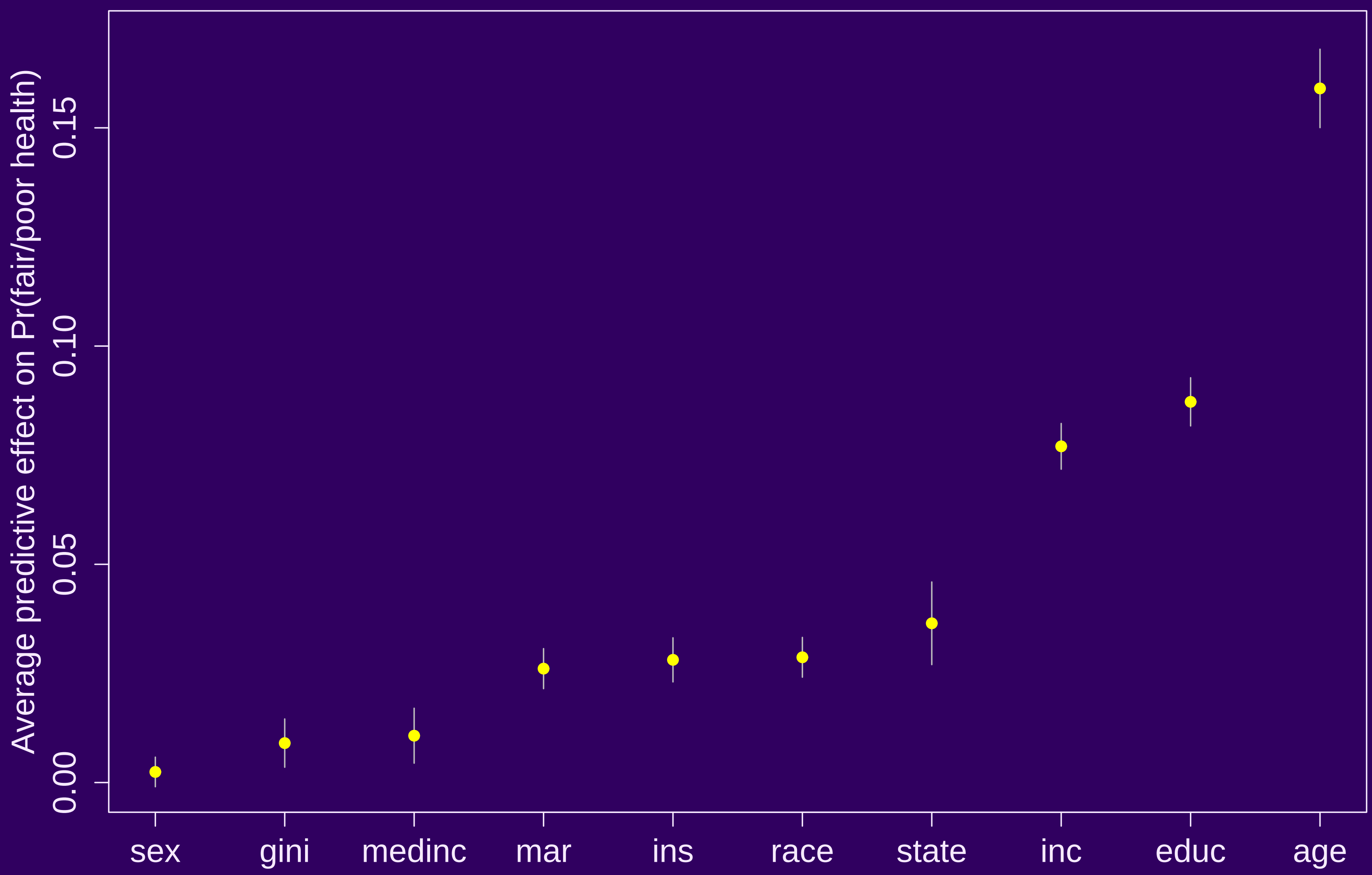
"The expected change in the response per unit change in the input, with all other inputs held constant"

$$\mathsf{PE}(u^{(1)} \to u^{(2)}, v, \theta) = \frac{\mathsf{E}(y|u^{(2)}, v, \theta) - \mathsf{E}(y|u^{(1)}, v, \theta)}{u^{(2)} - u^{(1)}}$$

where u is the input of interest, v represents the other inputs, $u^{(1)}$ is the initial value of u, and $u^{(2)}$ is the final value of u.

APE: average over distributions for $x = (u, v)$ and θ.

Weights, $w_{ij} = \frac{|u_j - u_i|}{1 + (v_i - v_j)^T \Sigma_v^{-1} (v_i - v_j)}$.


Average predictive effect on Pr(fair/poor health)
0.00
0.05
0.10
0.15
sex
gini
medinc
mar
ins
race
state
inc
educ
age

5. Displaying/understanding multilevel variance parameters

- Gelman (2005) revisits ANOVA to motivate Bayesian ANOVA, and finite-population and super-population variances
- Gelman and Pardoe (2005b) generalize explained variance (R^2) at each level of an MLM. Equivalent to usual definition of R^2 in classical least-squares regression. Average over regression parameter uncertainty: "Bayesian adjusted R^2."
- Gelman and Pardoe (2005b) also propose a related variance comparison to summarize degree to which estimates at each level of MLM are pooled together based on level-specific regression relationship, rather than estimated separately. In simple random-intercepts MLM, related to "shrinkage."

General multilevel model:

$$\theta_k^{(m)} = \mu_k^{(m)} + \epsilon_k^{(m)}, \quad \text{for } k = 1, \ldots, K^{(m)}$$

Explained variance:

$$R^2 = 1 - \frac{\mathrm{E}\left(\overset{K}{\underset{k=1}{\mathrm{V}}}\epsilon_k\right)}{\mathrm{E}\left(\overset{K}{\underset{k=1}{\mathrm{V}}}\theta_k\right)} = 0.35$$

1. Compute the vectors of “responses” θ_k, “predicted values” μ_k, and “errors” $\epsilon_k = \theta_k - \mu_k$
2. Compute the sample variances, $\overset{K}{\underset{k=1}{\mathrm{V}}}\theta_k$ and $\overset{K}{\underset{k=1}{\mathrm{V}}}\epsilon_k$
3. Average over the simulation draws to estimate $\mathrm{E}\left(\overset{K}{\underset{k=1}{\mathrm{V}}}\theta_k\right)$ and $\mathrm{E}\left(\overset{K}{\underset{k=1}{\mathrm{V}}}\epsilon_k\right)$, and then use these to calculate R^2

General multilevel model:

$$\theta_k^{(m)} = \mu_k^{(m)} + \epsilon_k^{(m)}, \quad \text{for } k = 1, \ldots, K^{(m)}$$

Pooling factor:

$$\lambda = 1 - \frac{\overset{K}{\underset{k=1}{V}} \mathrm{E}(\epsilon_k)}{\mathrm{E}\left(\overset{K}{\underset{k=1}{V}} \epsilon_k\right)} = 0.19$$

1. For each k, estimate the posterior mean $\mathrm{E}(\epsilon_k)$ of each of the errors ϵ_k as defined previously
2. Compute $\overset{K}{\underset{k=1}{V}} \mathrm{E}(\epsilon_k)$—that is, the variance of the K values of $\mathrm{E}(\epsilon_k)$—and then use this, along with $\mathrm{E}(\overset{K}{\underset{k=1}{V}} \epsilon_k)$ from the R^2 calculation to calculate λ

6. Diagnostics for model checking

- Questions naturally arise as to whether an MLM provides an adequate fit to the data
- Is the computational burden of an MLM over a non-multilevel model justified?
- Bayes marginal model plots Pardoe (2004) can be used to visualize goodness of fit in multilevel settings
  - can clearly demonstrate the need to consider MLMs when analyzing such data

References

Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discussion). *The Annals of Statistics 33*.

Gelman, A. and I. Pardoe (2005a). Average predictive effects for models with nonlinearity, interactions, and variance components. Technical report, Dept. Statistics, Columbia U.

Gelman, A. and I. Pardoe (2005b). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technical report, Dept. Statistics, Columbia University.

Gelman, A., C. Pasarica, and R. Dodhia (2002). Let's practice what we preach: Turning tables into graphs. *The American Statistician 56*, 121–130.

Pardoe, I. (2004). Model assessment plots for multilevel logistic regression. *Computational Statistics and Data Analysis 46*, 295–307.

Pardoe, I. and B. Shor (2005). Graphical displays of predictive effects. Technical report, Lundquist College of Business, University of Oregon.

Subramanian, S. V. and I. Kawachi (2004). Income inequality and health: What have we learned so far? *Epidemiological Reviews 26*, 78–91.