Analysing Comparative Longitudinal Survey Data Using Multilevel Models

(Day 2)

Dr Malcolm Fairbrother

School of Geographical Sciences · Centre for Multilevel Modelling
University of Bristol

Barcelona Summer School in Survey Methodology · July 2016

University of BRISTOL

Centre for Multilevel Modelling

To Download These Slides

http://bit.ly/29k1xtk

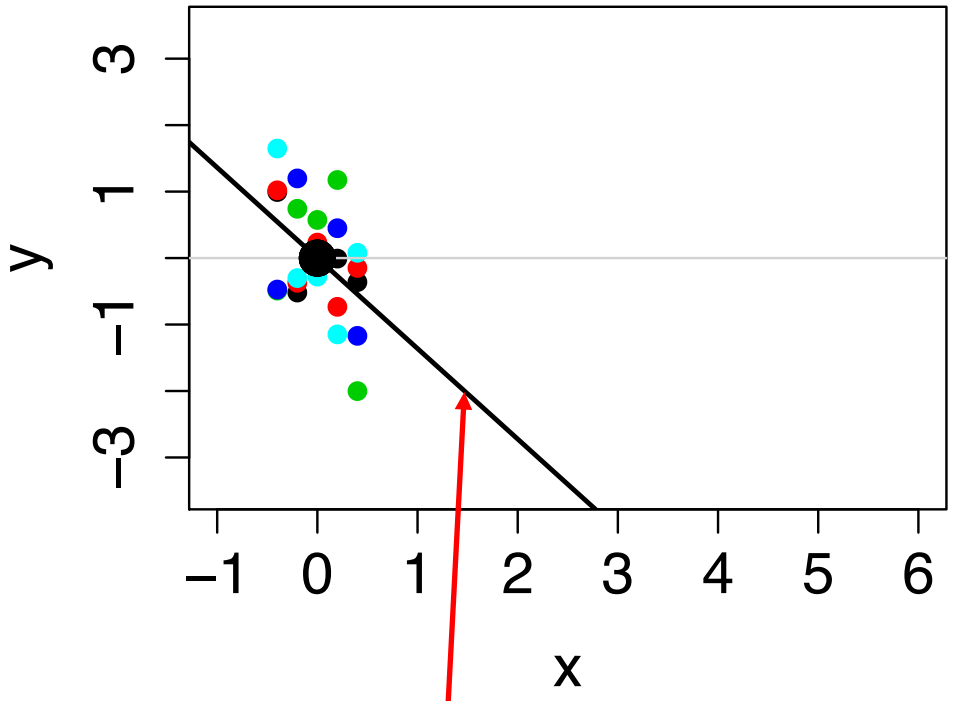
Agenda

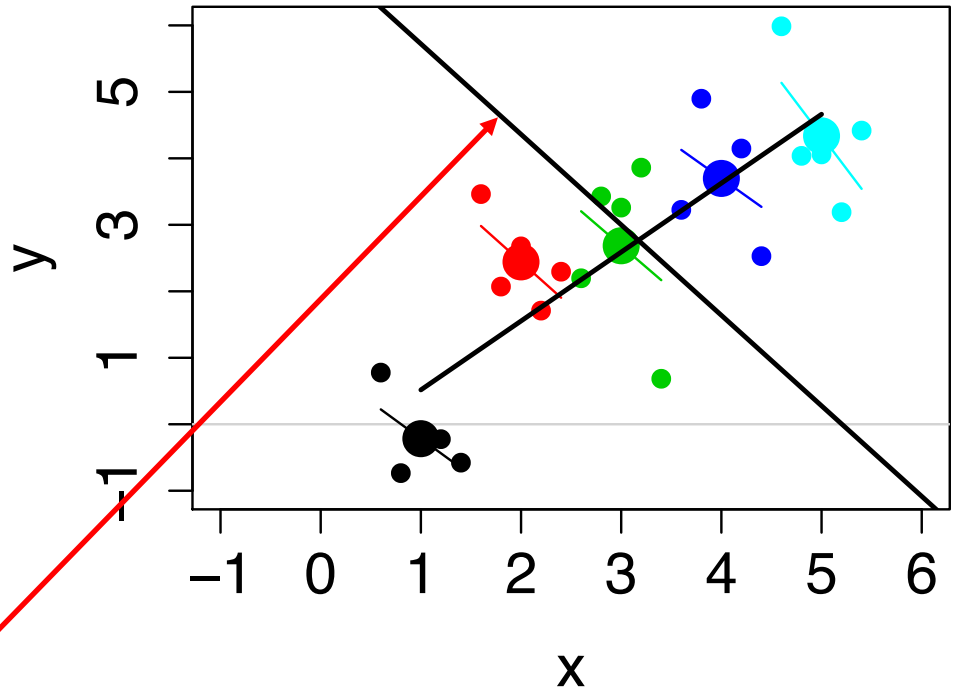
1. recap of yesterday
2. aggregating by country-wave
3. societal growth curves
4. random (country-specific) slopes
5. Bayesian/MCMC estimation
6. issues of GLMMs (versus just LMMs)
7. last miscellaneous issues and cautions

Recap

Fixed Effects

Random Effects





average “within” effect

An Application with Comparative Longitudinal Survey Data: European Values in the EVS

An Application with Comparative Longitudinal Survey Data: European Values in the EVS

"The worldviews of the peoples of rich societies differ systematically from those of low-income societies across a wide range of political, social, and religious norms and beliefs. These two dimensions reflect cross-national polarization between *traditional* versus *secular-rational* orientations toward authority; and *survival* versus *self-expression* values."

Inglehart and Baker (2000, in the *American Sociological Review*)

An Application with Comparative Longitudinal Survey Data: European Values in the EVS

CHILDFAITH

"Here is a list of qualities that children can be encouraged to learn at home. Which, if any, do you consider to be especially important? Please choose up to five. …

Religious faith?"

(two-point scale: mentioned or not mentioned)

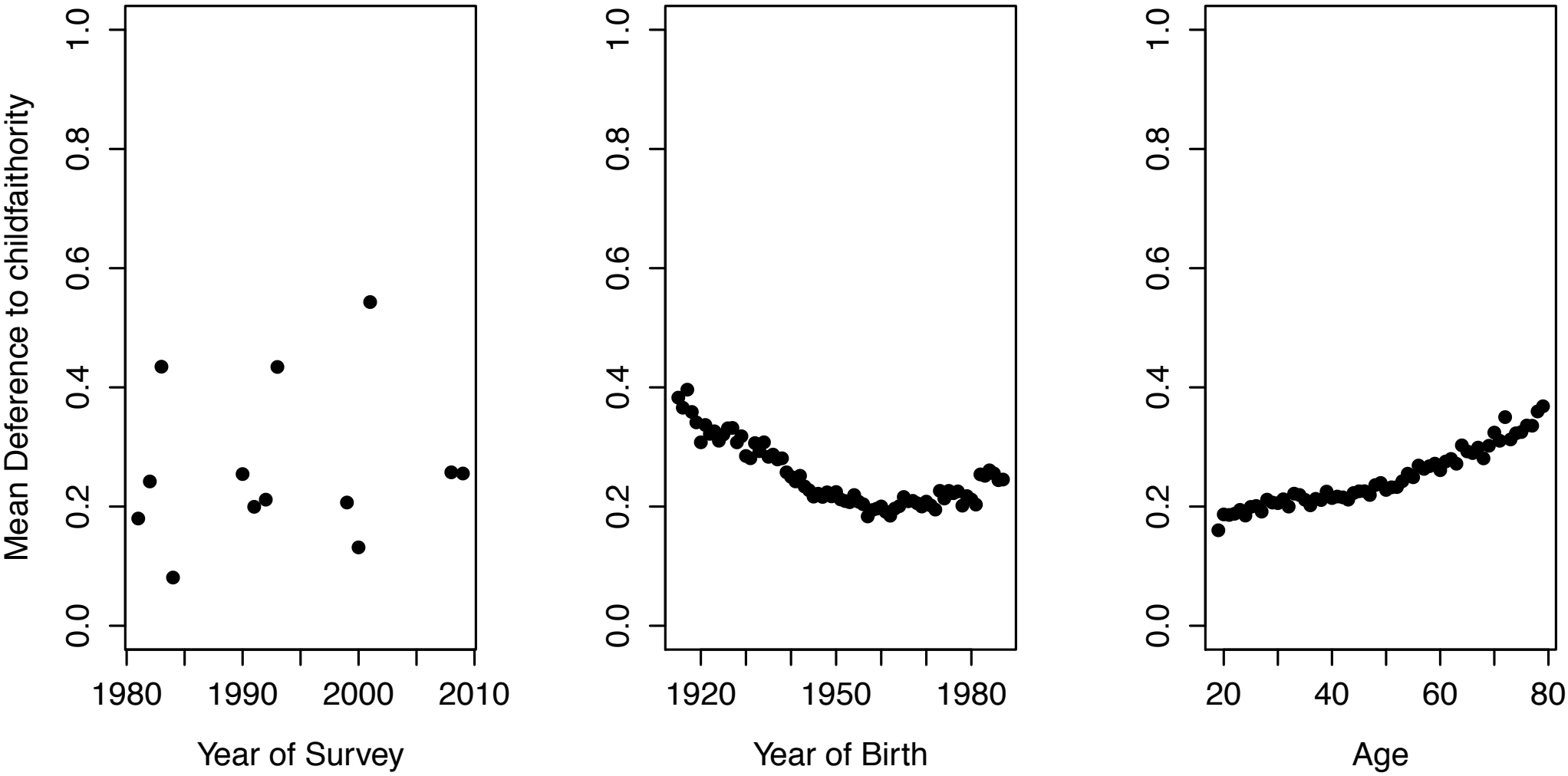
AUTH

"I'm going to read out a list of various changes in our way of life that might take place in the near future. Please tell me for each one, if it were to happen, whether you think it would be a good thing, a bad thing, or don't you mind? …

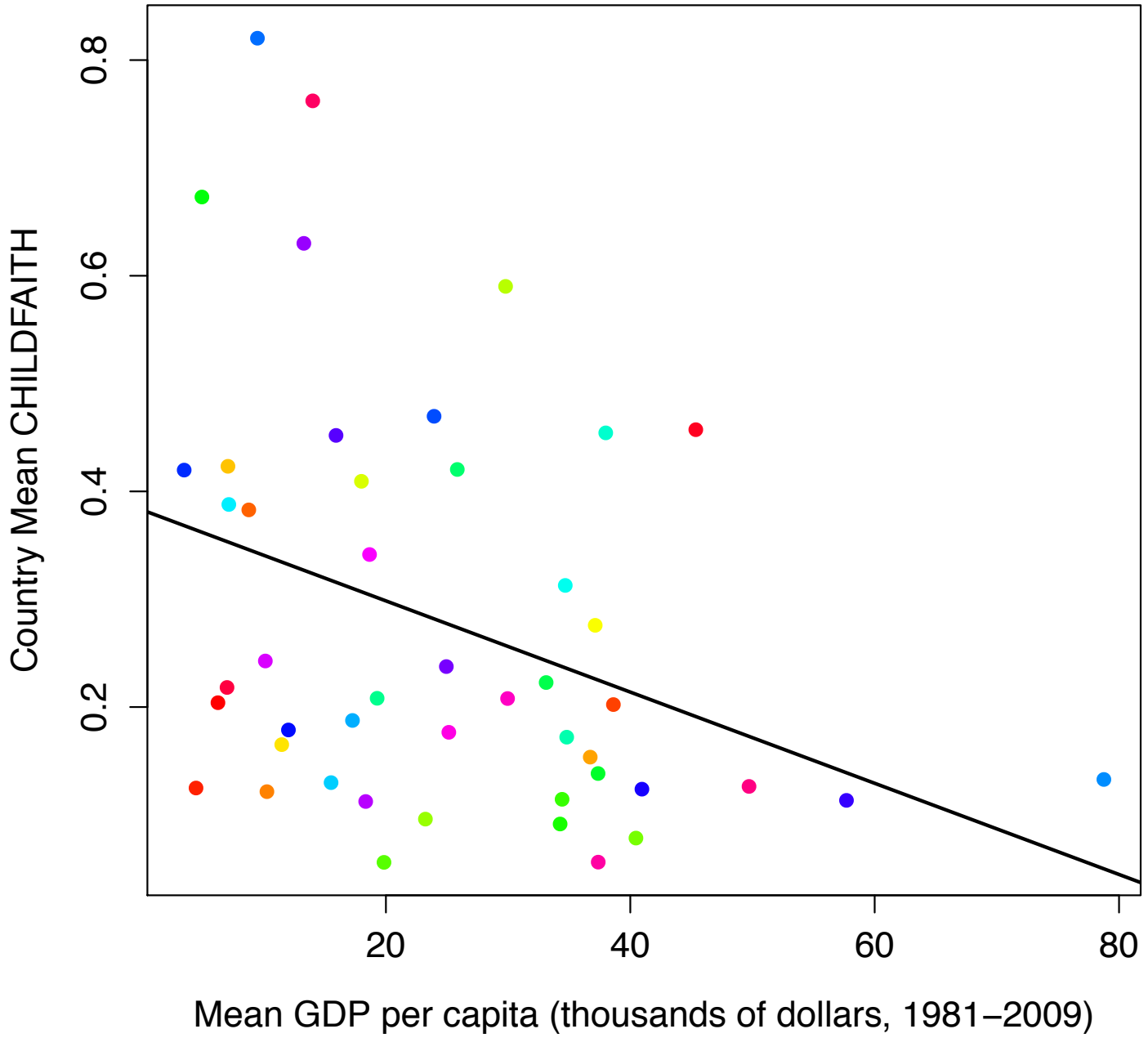
Greater respect for authority"

(three-point scale: bad thing, don't mind, good thing)

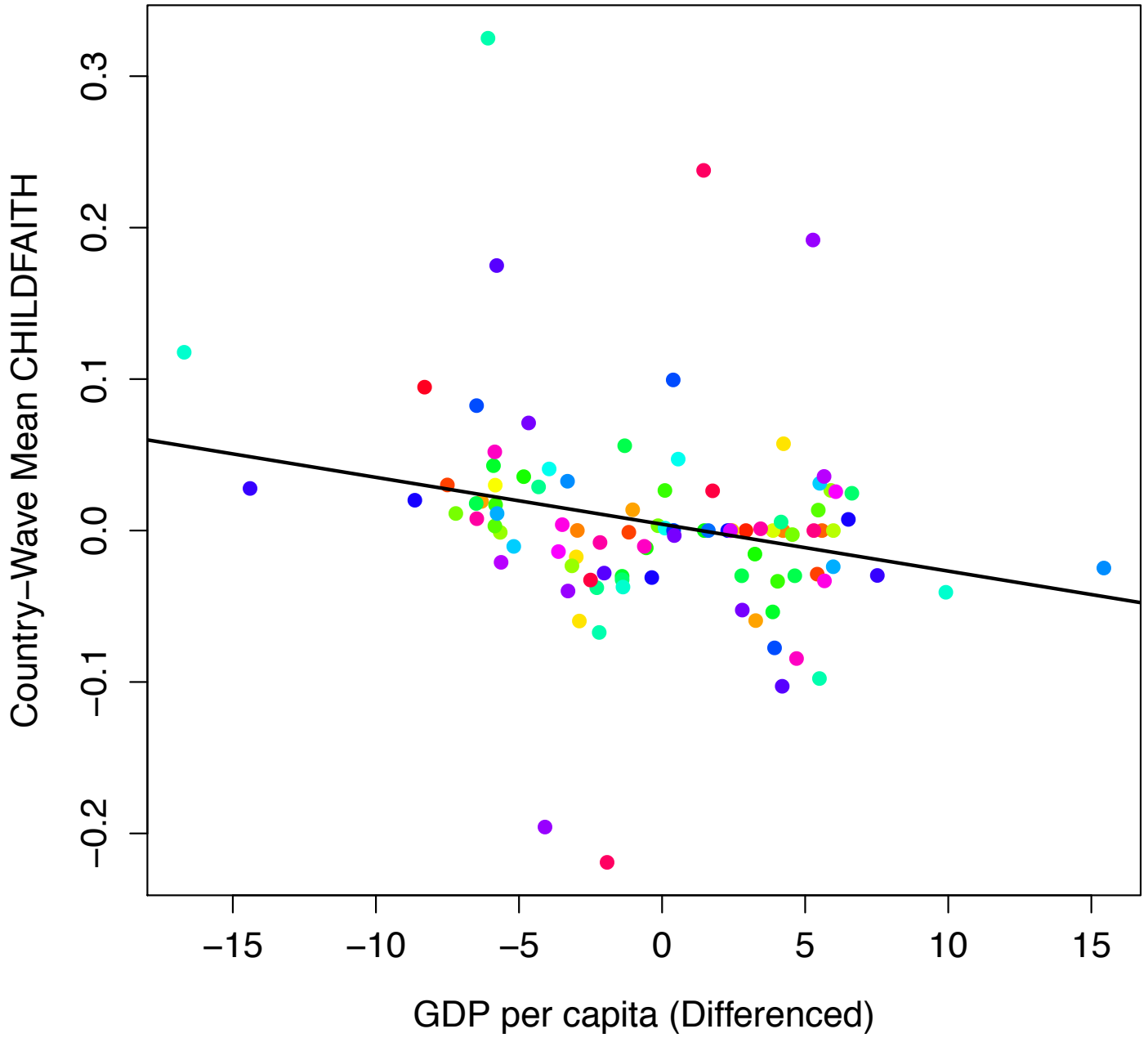
CHILDFAITH Over Time



CHILDFAITH versus GDP per capita (countries)



CHILDFAITH versus GDP per capita (country-waves)



Growth Curves

Growth Curves

- thus far, we've looked at how to model change over time in Y as a function of change over time in X

- different Q: what are the covariates (even time-invariant) covariates of faster (or slower) change in Y over time?

- standard technique in analyses of panel data on individuals: growth curves
  - multilevel: observations are nested within individuals
  - typical applications: growth/development of physical, psychological, behavioral characteristics of people

Growth Curves

- insofar as CLSD are panel data (on countries), we can estimate "growth curves" at the country level
  - assess the correlates of different rates of change
  - helps clarify how units arrive at different endpoints

- technically simply: just an interaction between a time-invariant (country-level) covariate and time

- typically includes a random slope for time (stay tuned)

Panel Data and Growth Curves

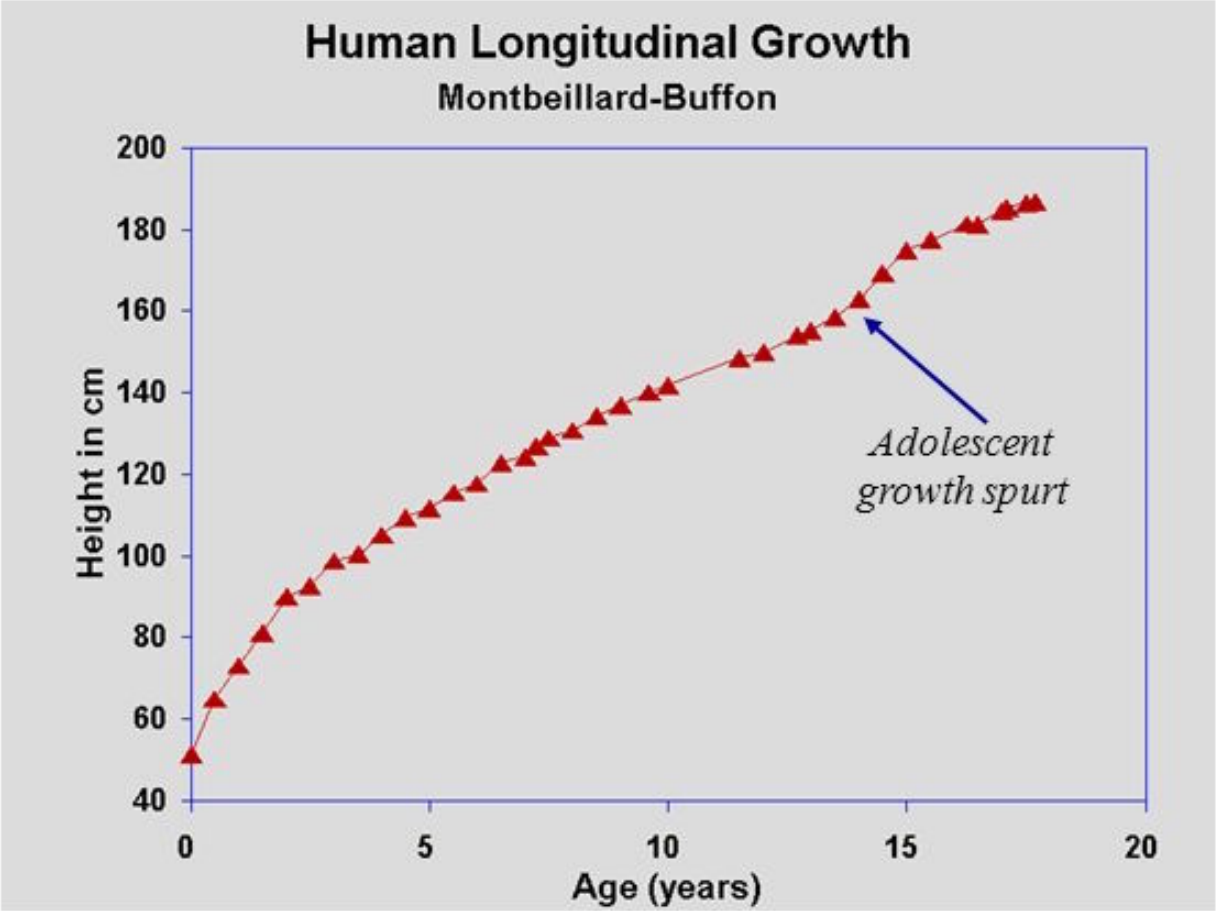
- growth curves = interactions between some time-invariant *x* and time
- the "curves" may in fact be straight lines, discontinuous, whatever

- examples of time-invariant properties of individuals:
  - mean income of city of birth
  - mother's level of education
  - test score at age 11
  - etc.

- R syntax: use "*" between variables to generate an interaction effect, with each of the main effects included also by default:

  "x1*x2" -> $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

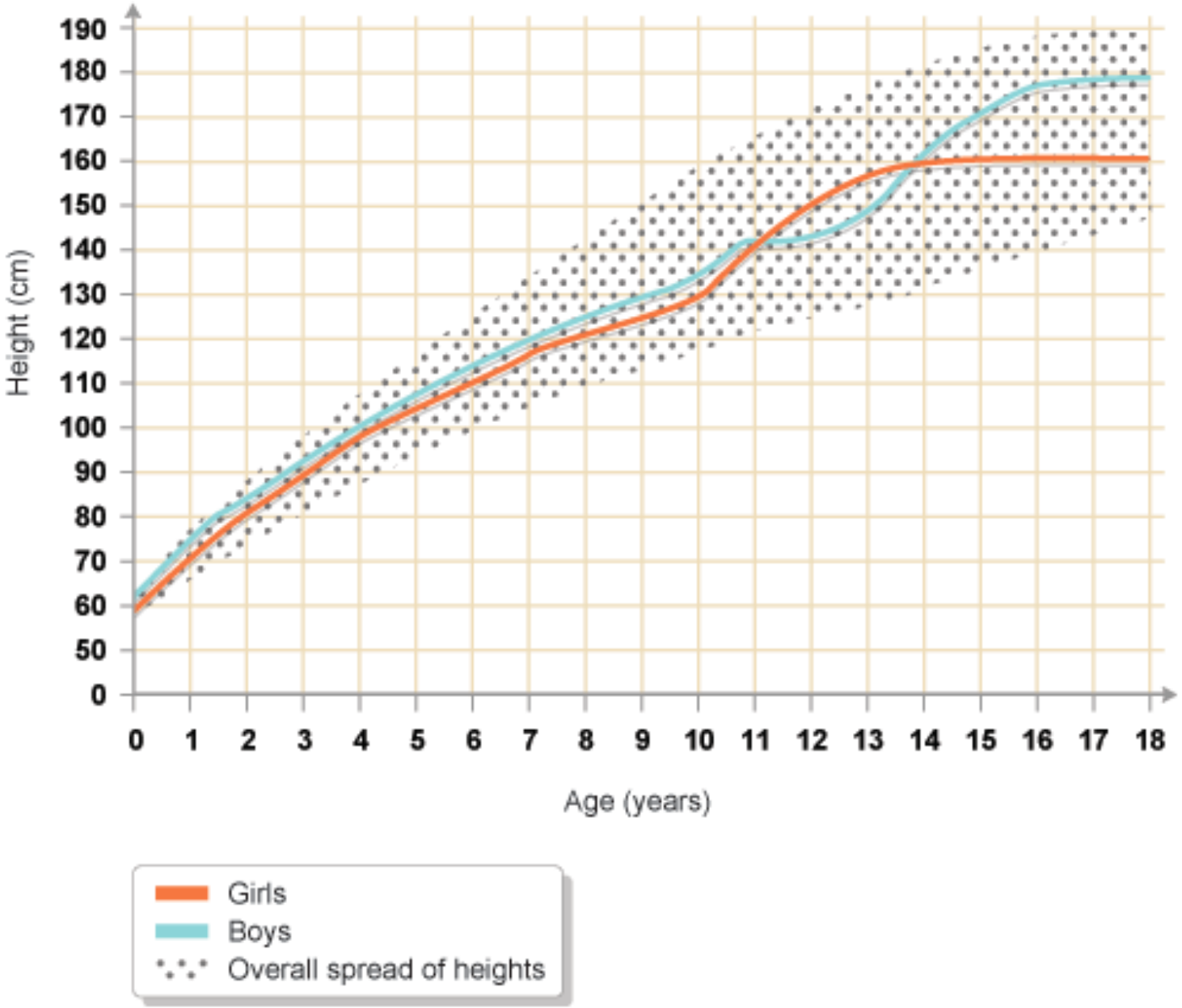
Why “Growth Curves”?



Source: Singer and Willett

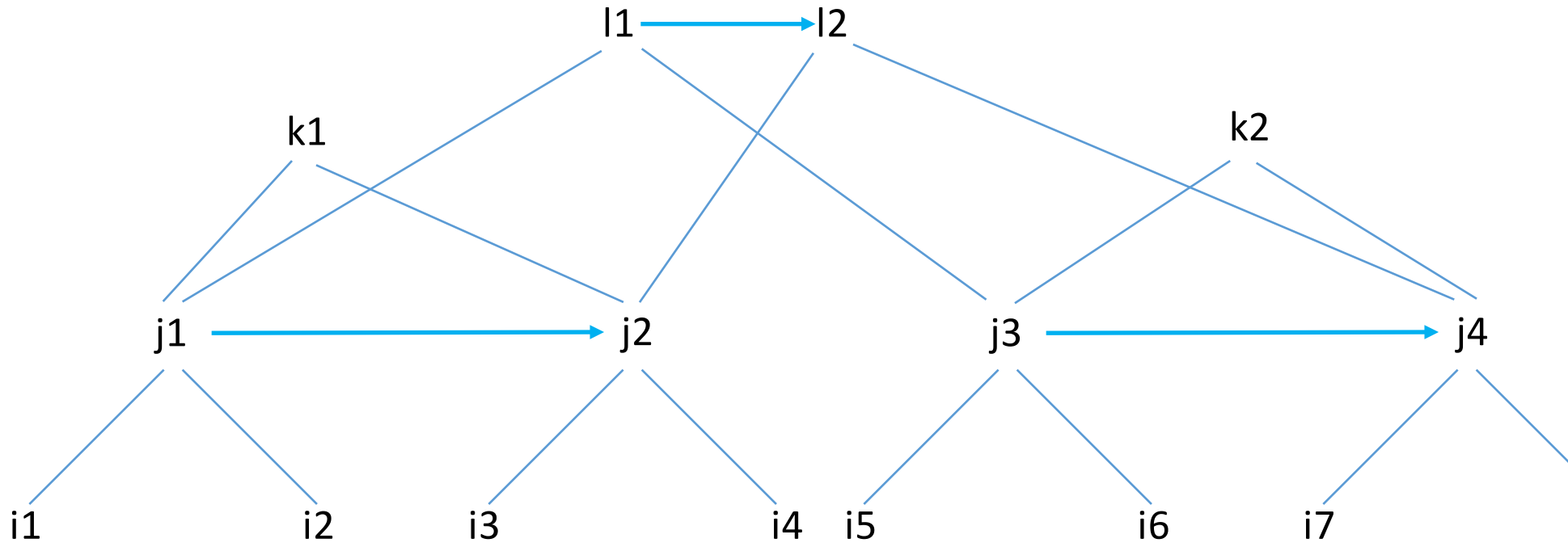
Buffon (1777) Histoire Naturelle & Scammon, RE (1927) The first seriation study of human growth, Am J of Physical Anthropology, 10, 329-336/

Why “Growth Curves”?

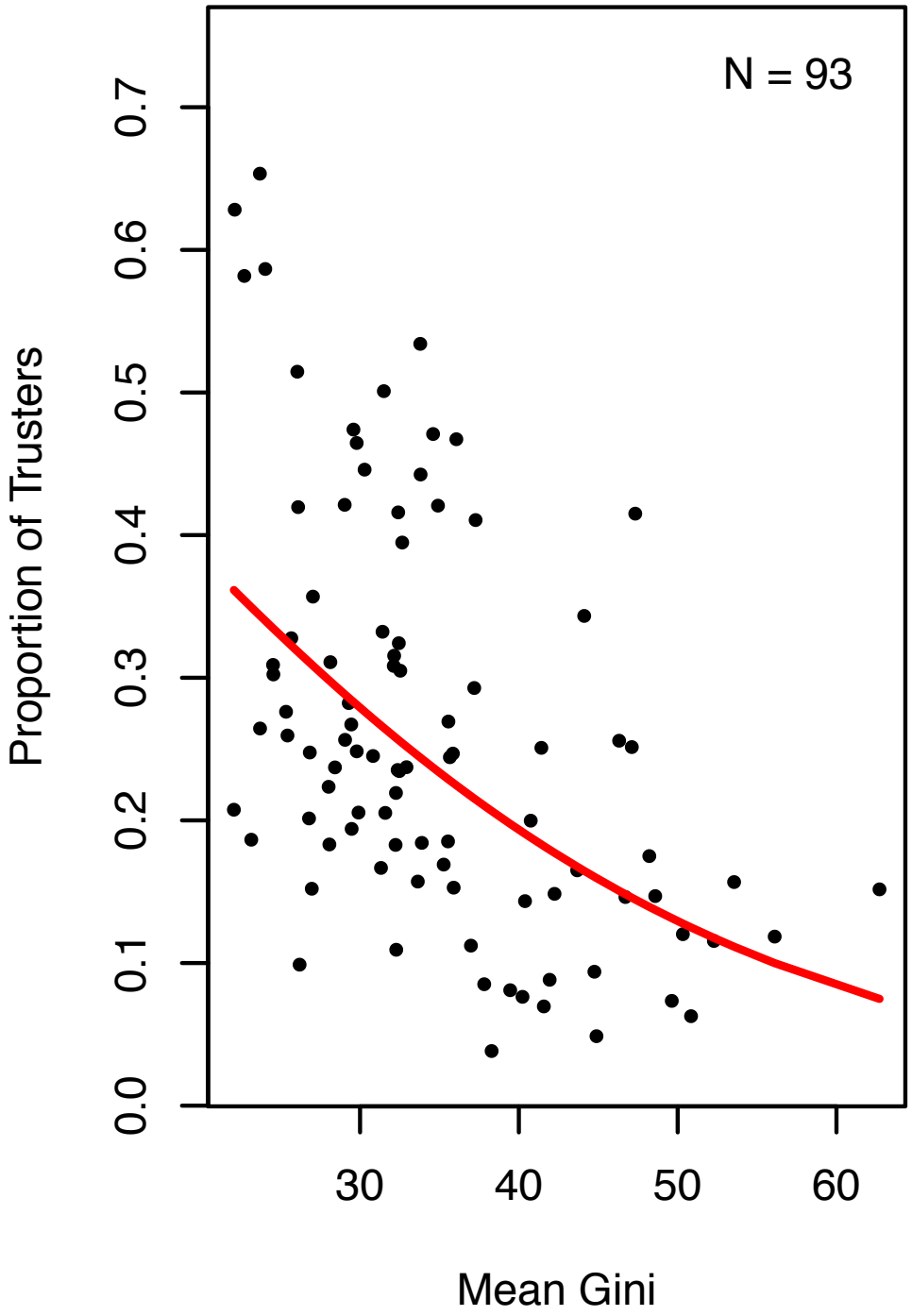


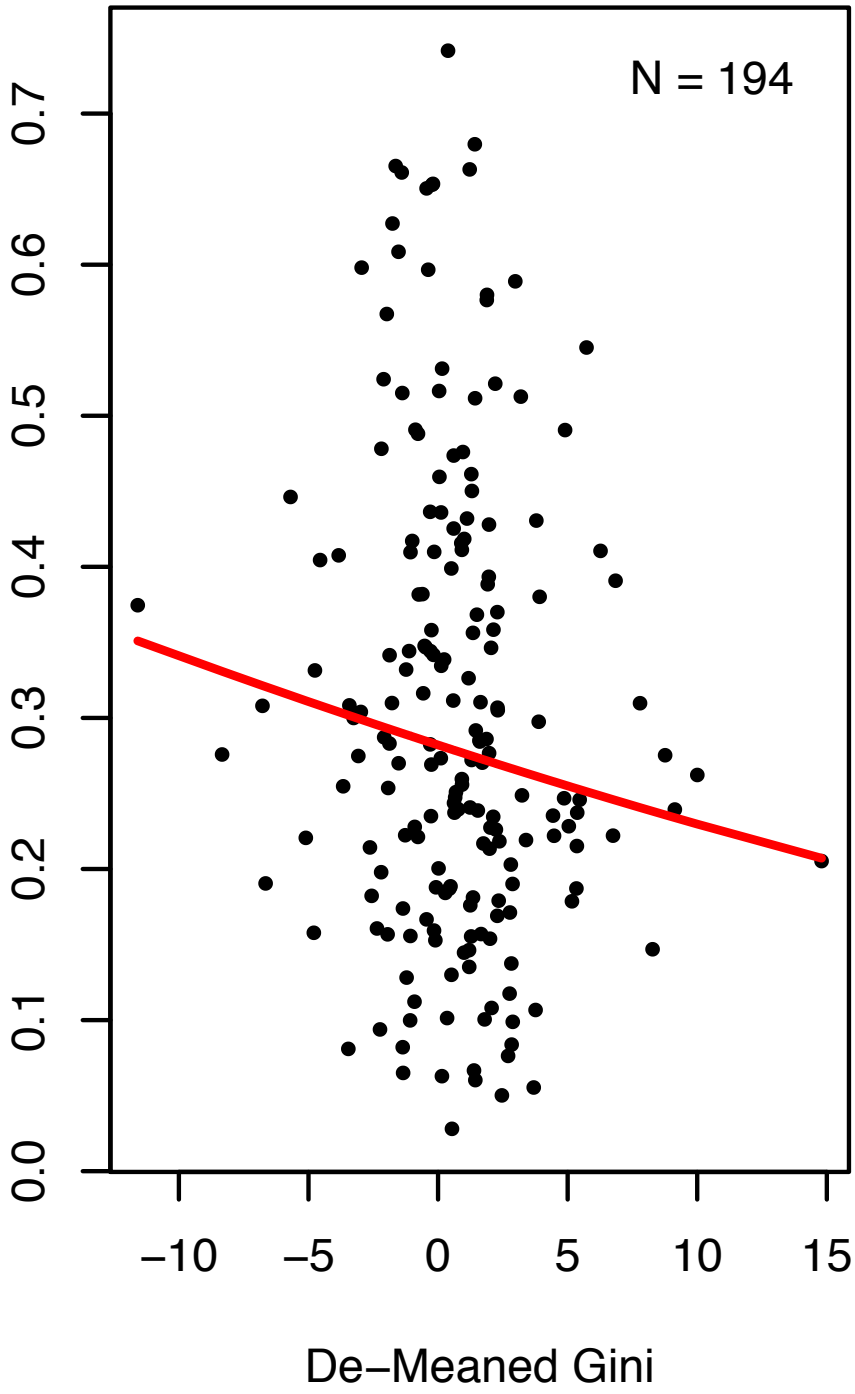
Societal Growth Curves

- an interaction between time and some enduring/time-invariant property of societies…
  - possible with time-series cross-sectional (TSCS) data, or comparative longitudinal survey data (CLSD)

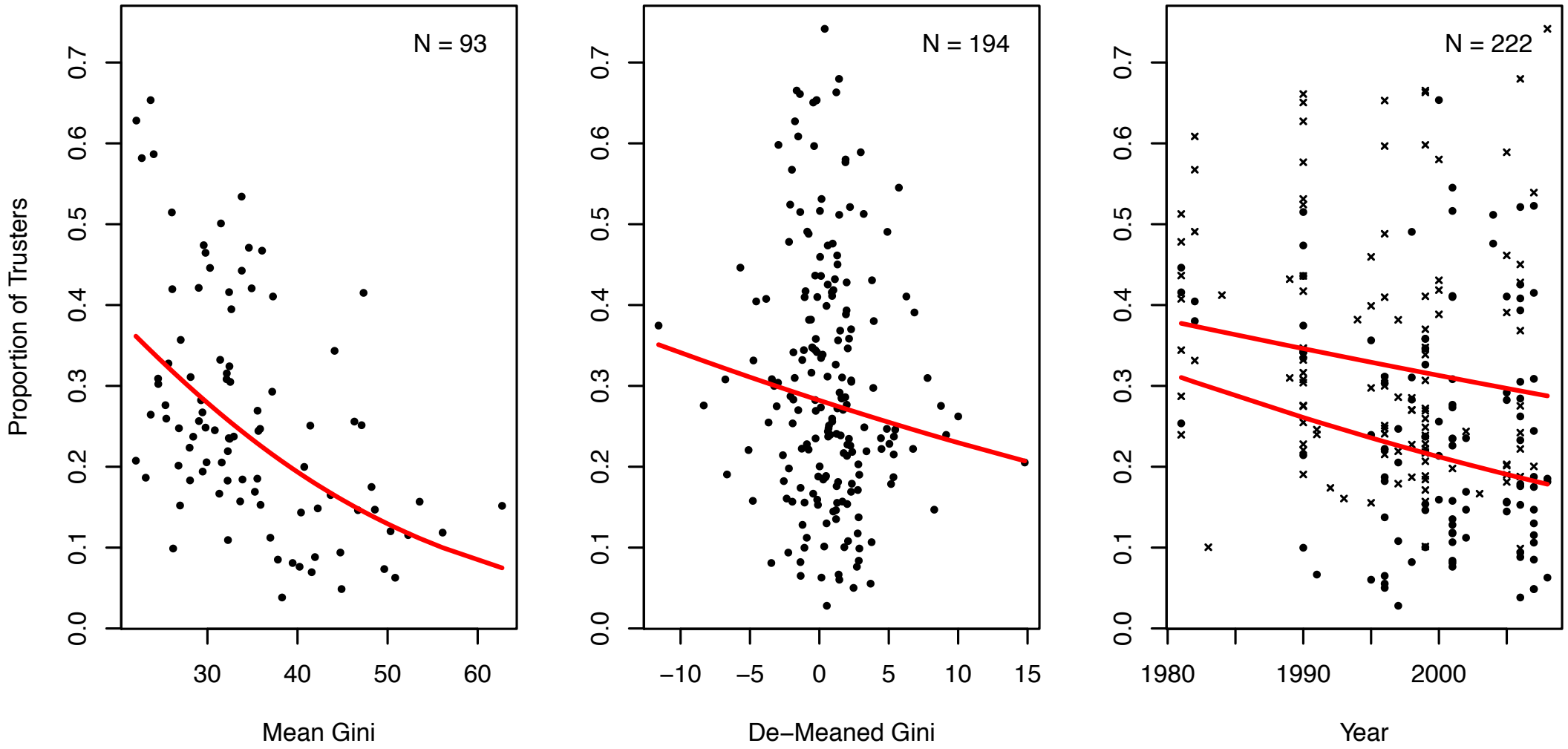


Income Inequality and Social Trust

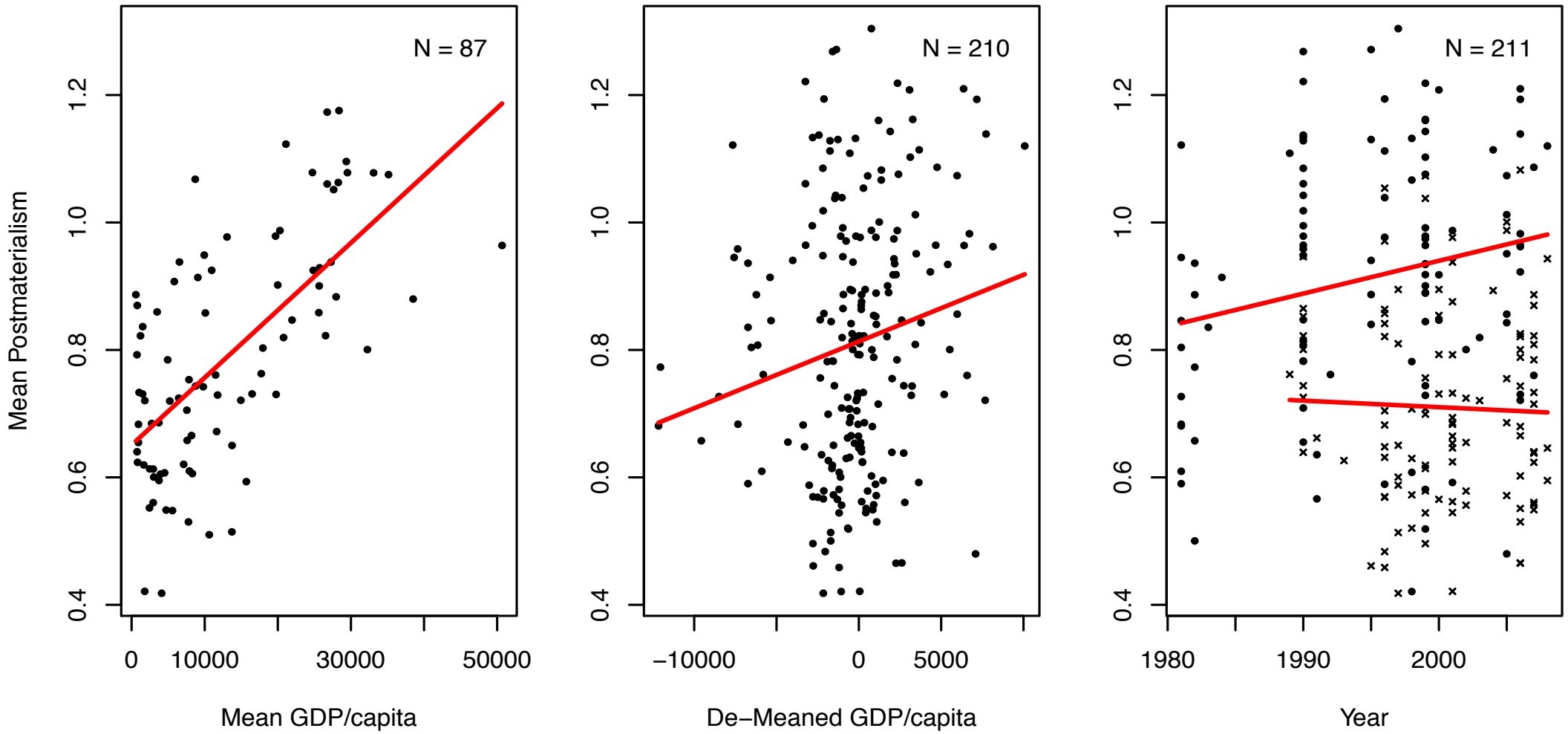




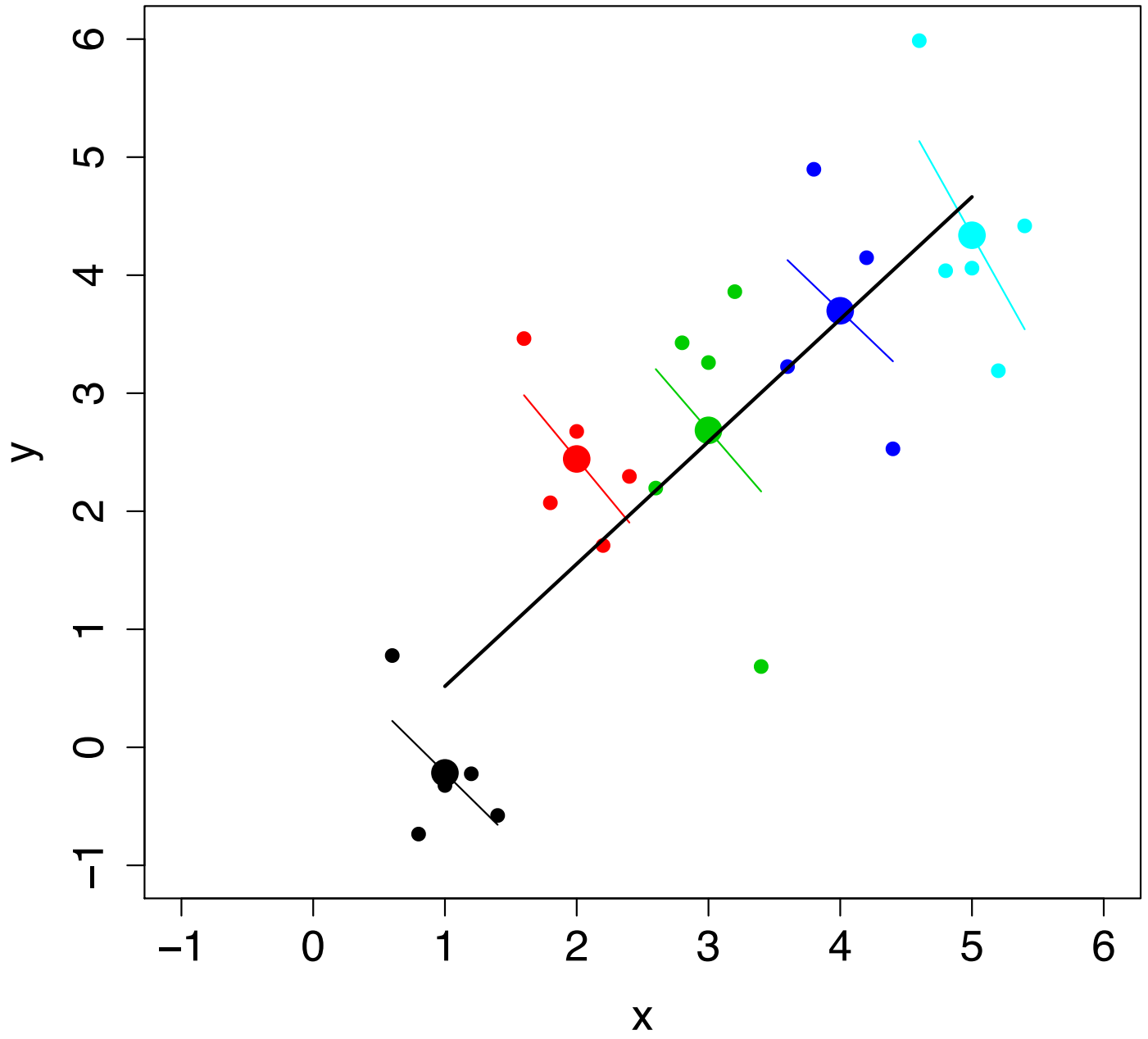
Income Inequality and Social Trust

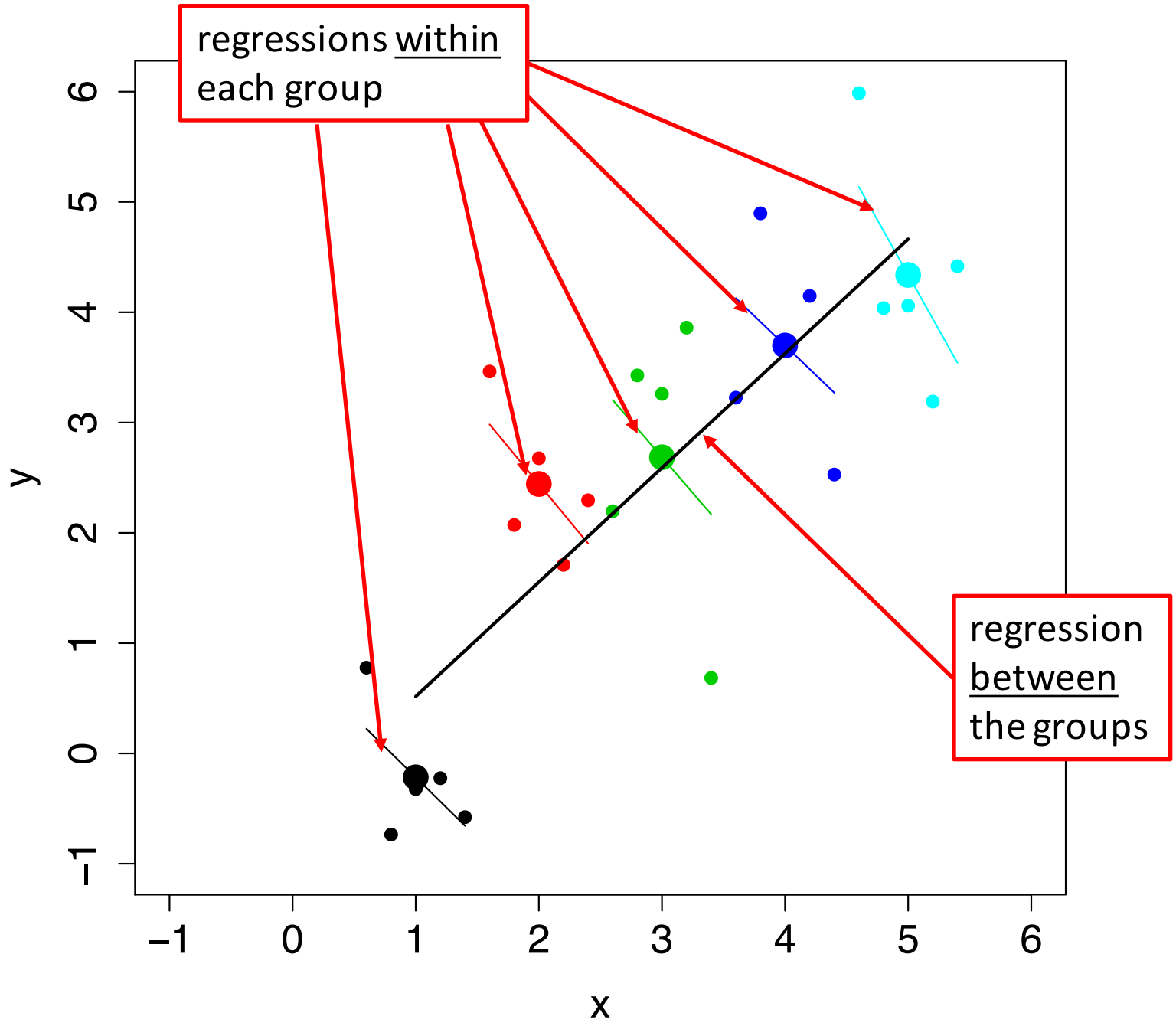


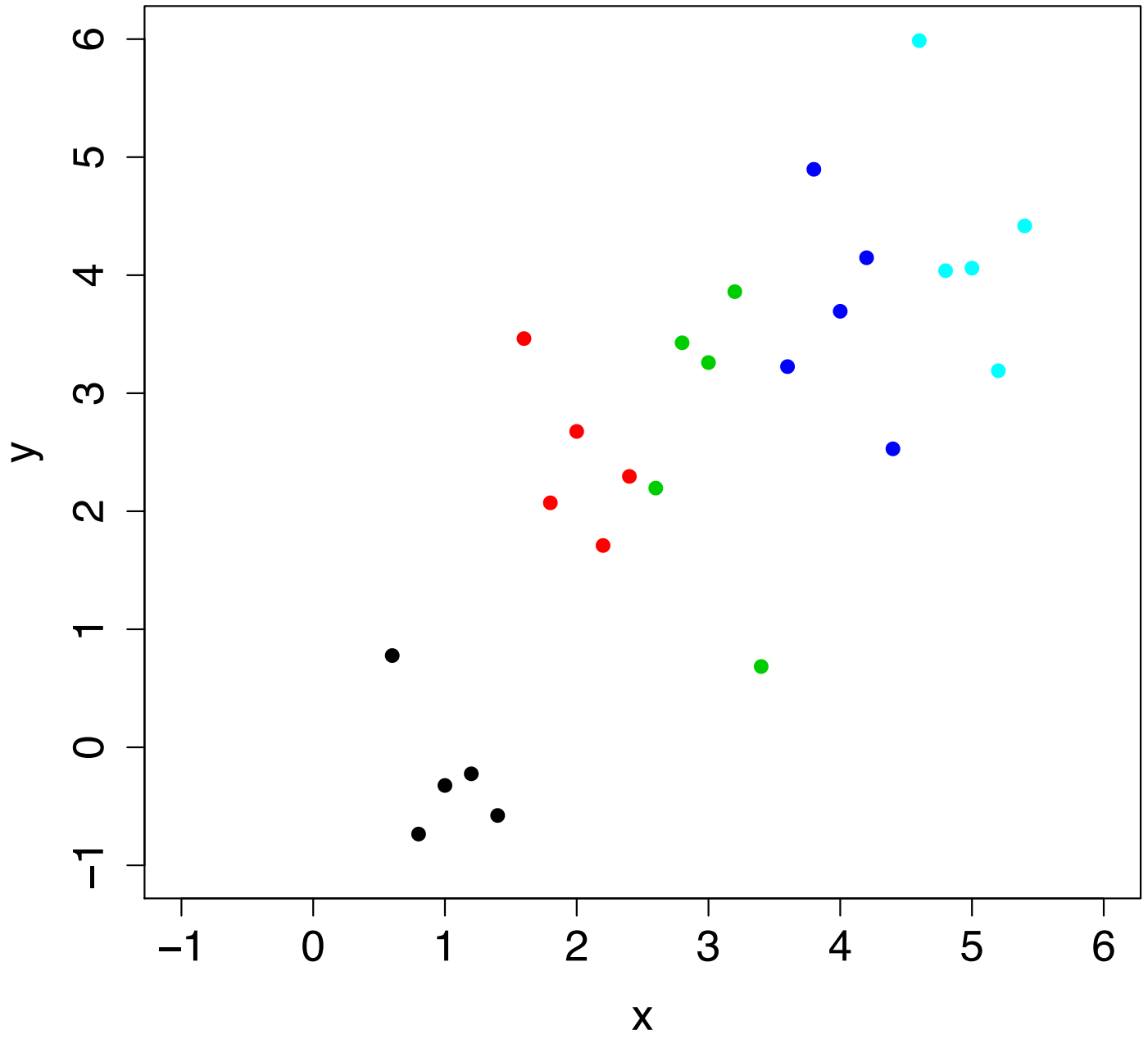
GDP/capita and Post-Materialist Values

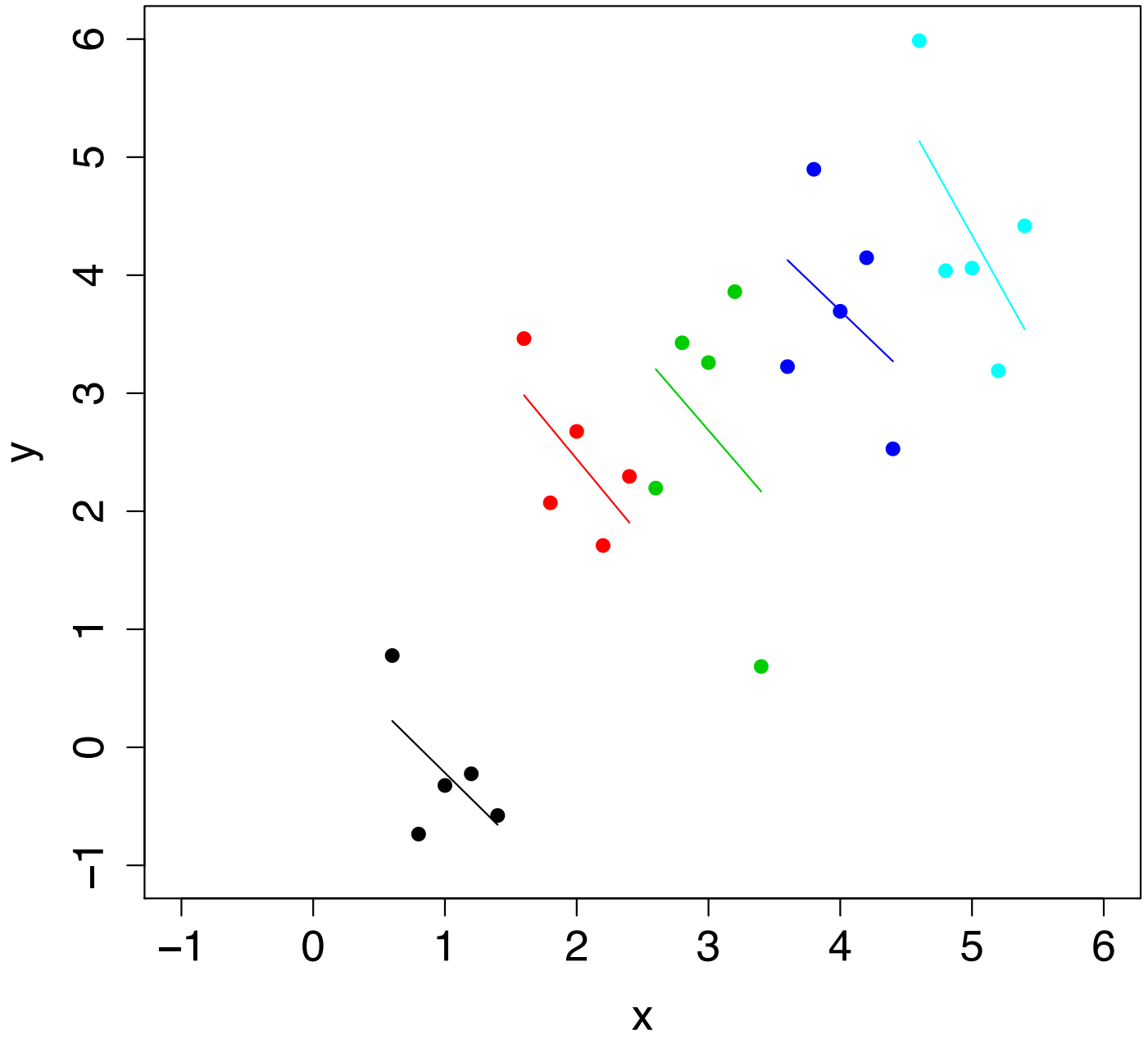


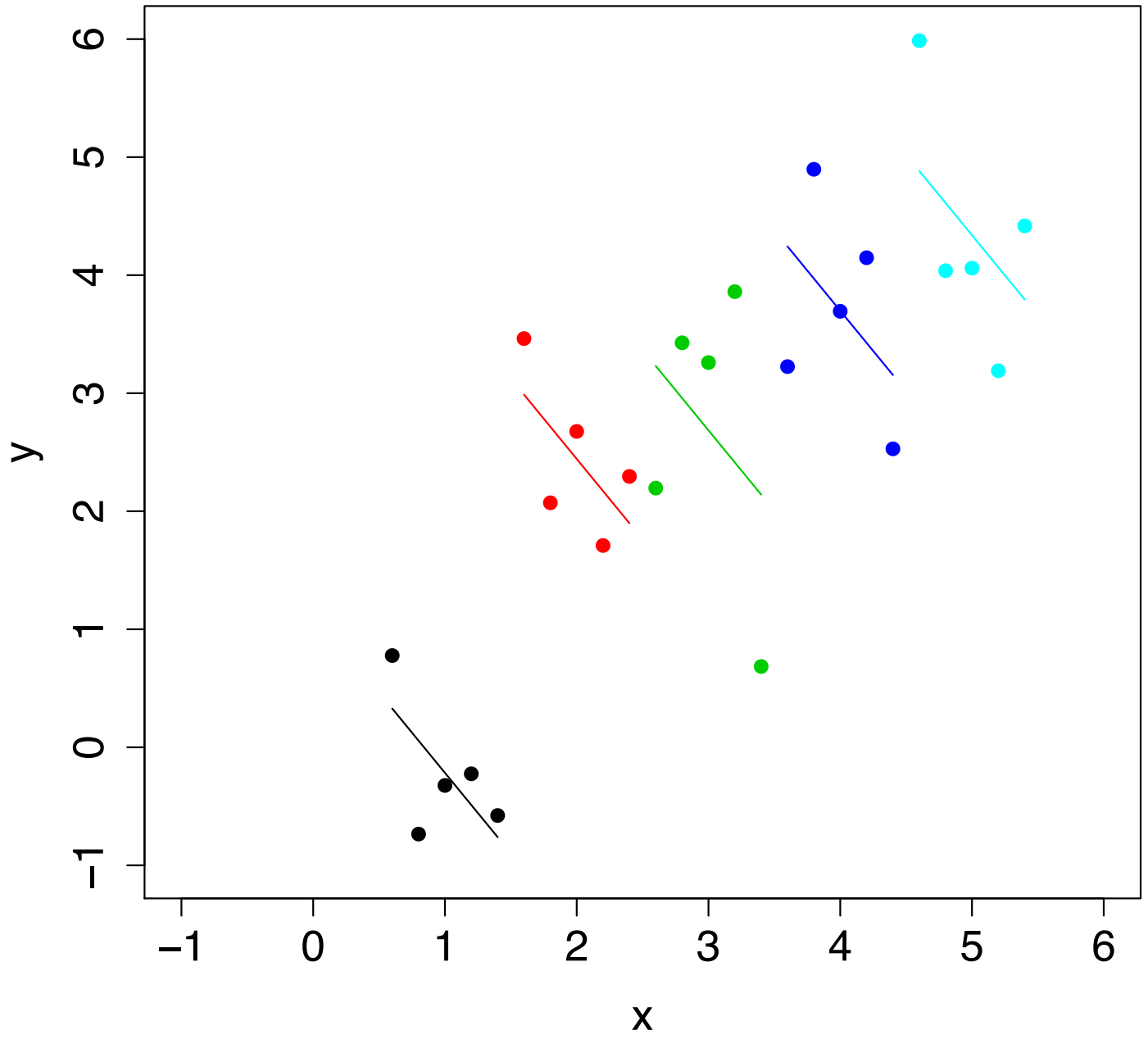
Random Slopes

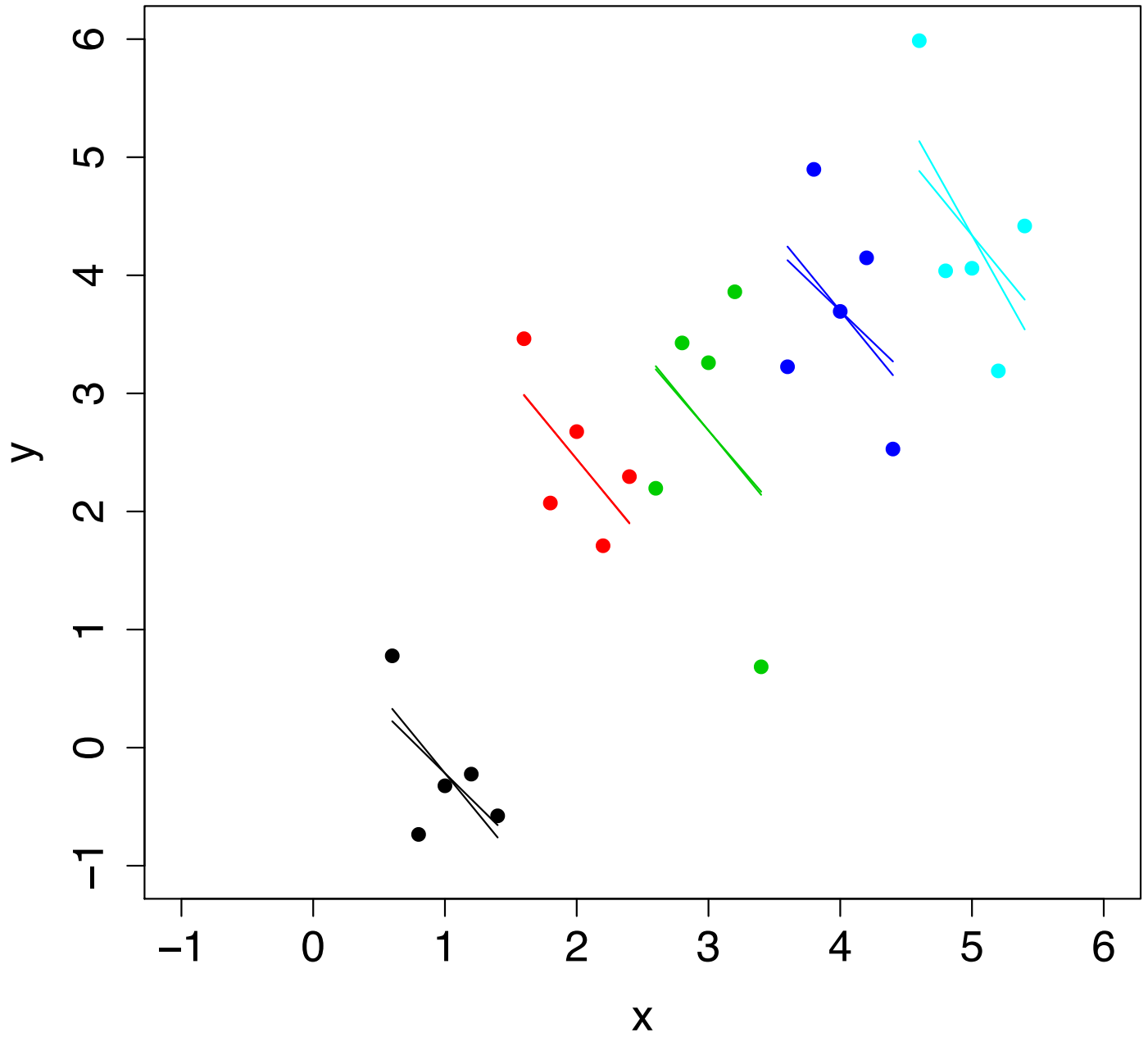

y
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
x

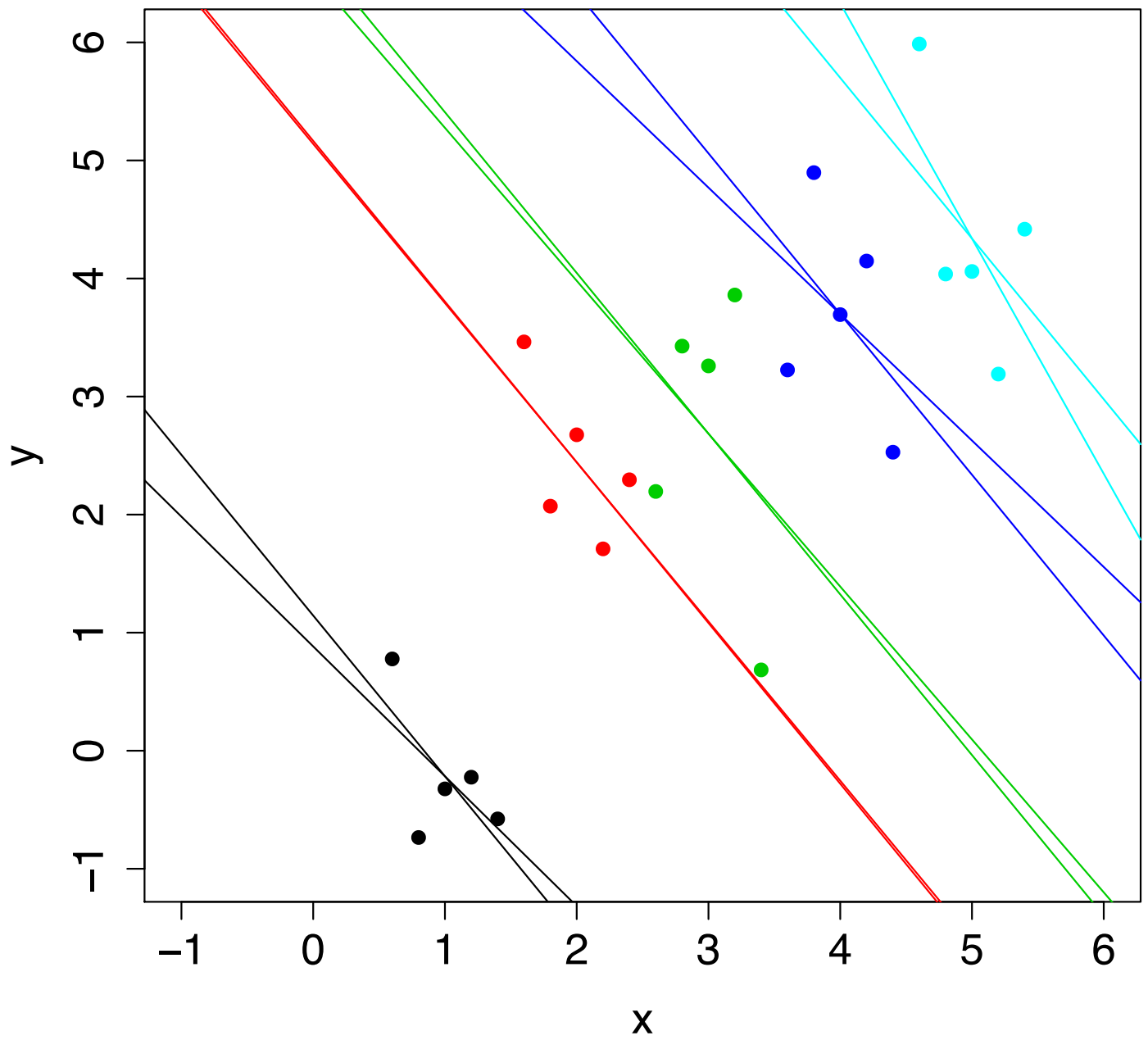

regressions within each group
regression between the groups
y
x
−1
0
1
2
3
4
5
6


y
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
x


y
x
−1
0
1
2
3
4
5
6


y
x
−1
0
1
2
3
4
5
6


y
x
−1
0
1
2
3
4
5
6


y
x
−1
0
1
2
3
4
5
6

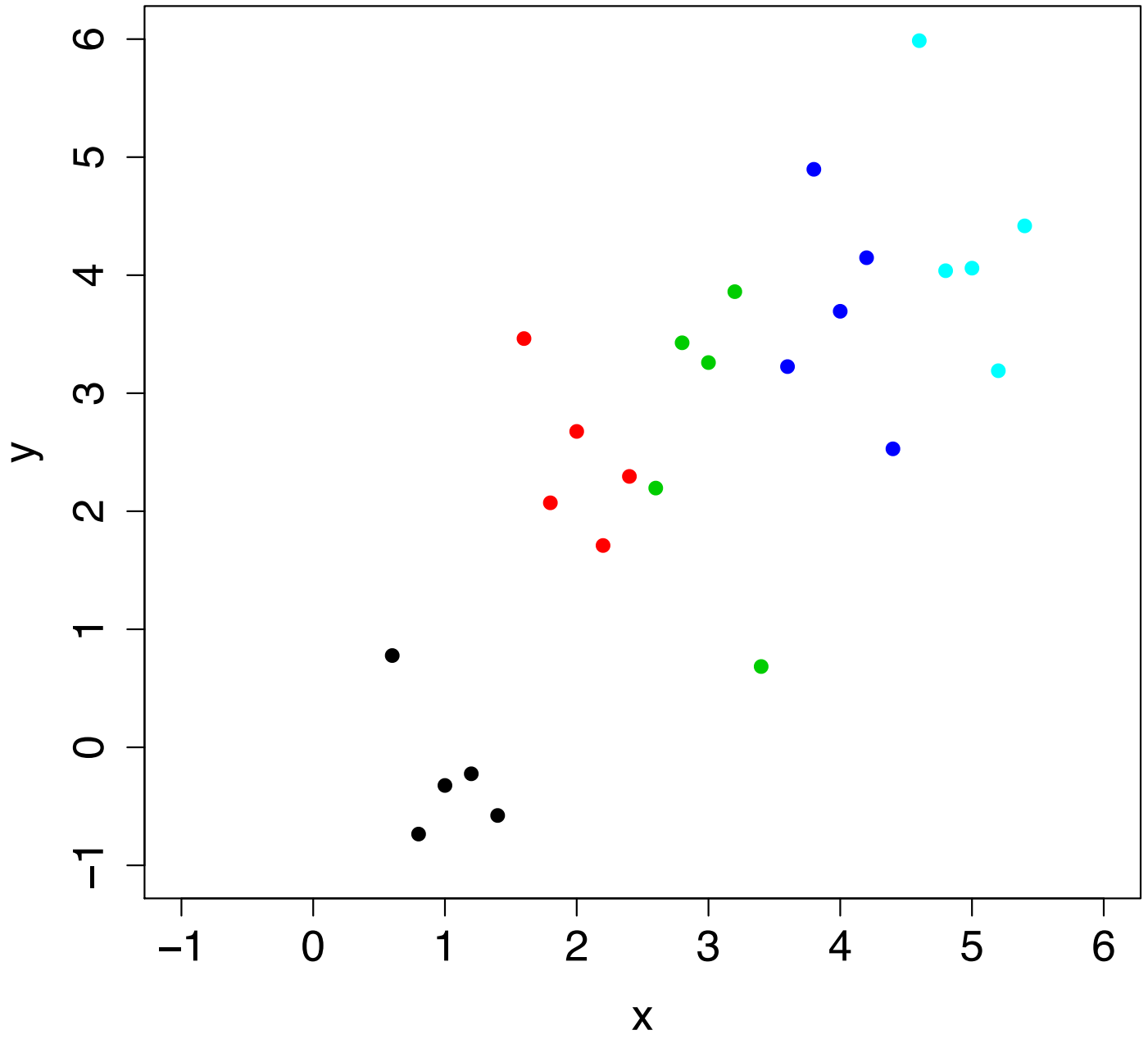
A “Random Intercept” Model

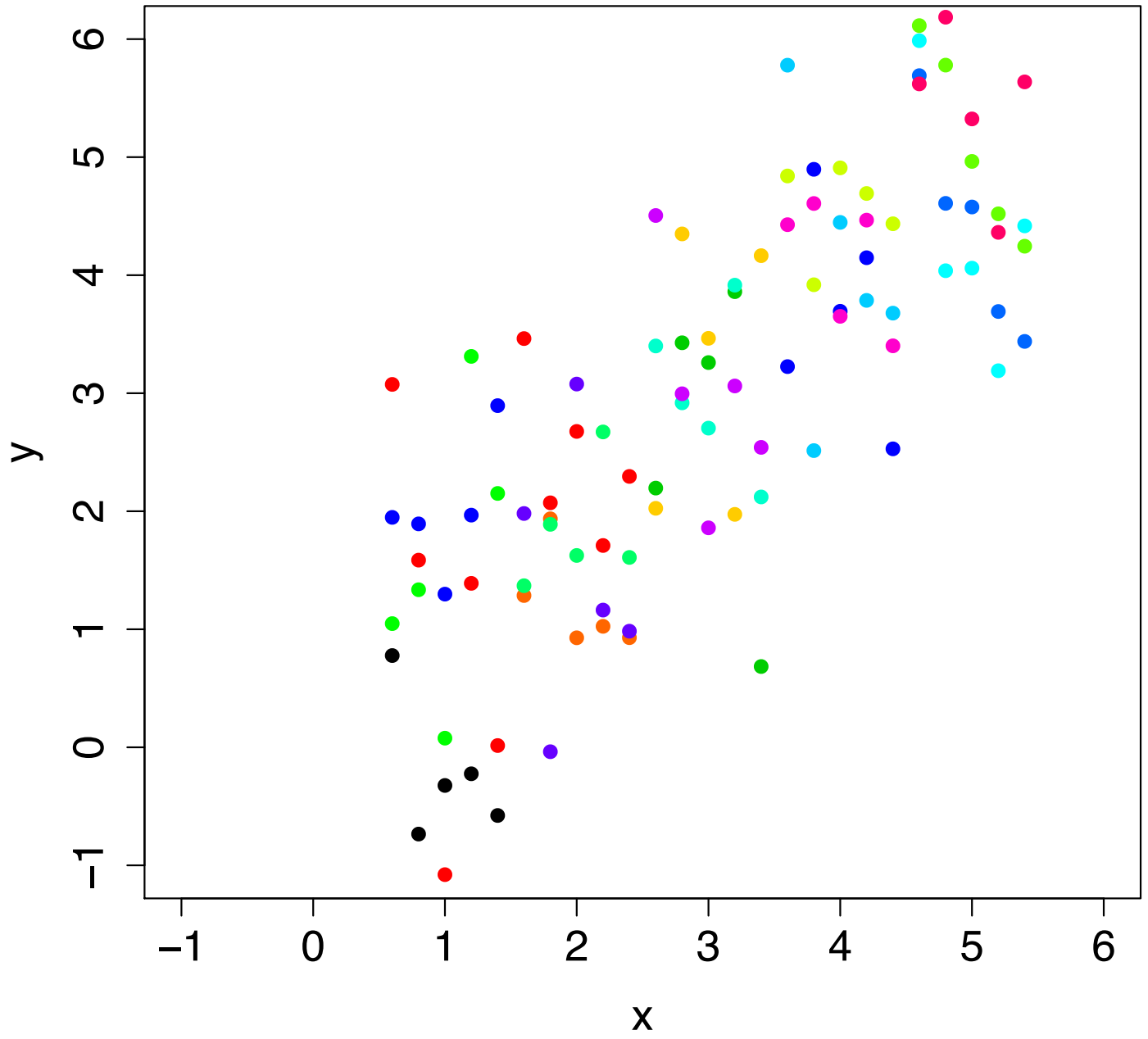
$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$

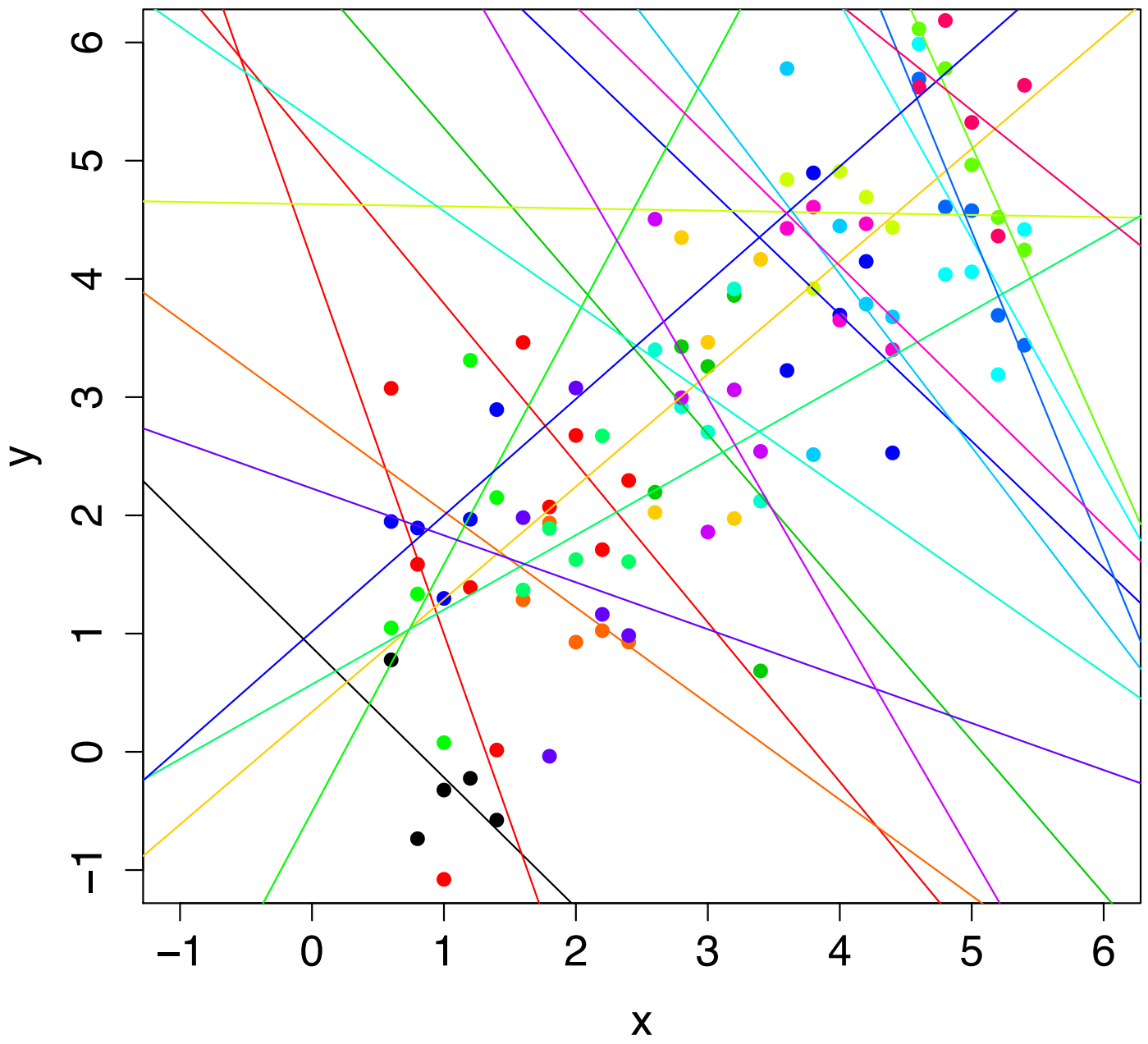
$$\left[u_{0j}\right] \sim N(0, \Omega_u) : \ \Omega_u = [\sigma^2_{u0}]$$

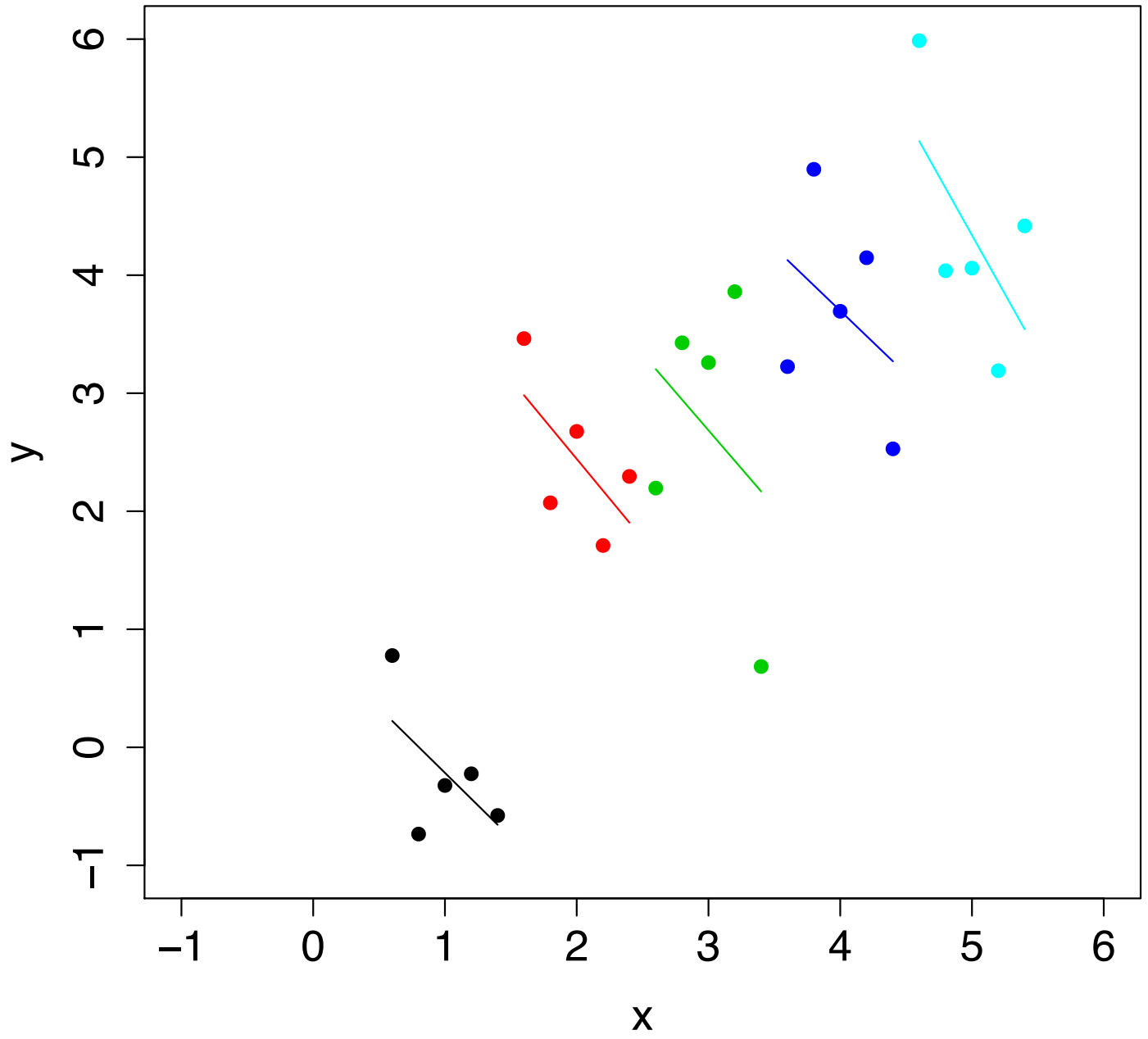
$$e_{ij} \sim N(0, \sigma^2_e)$$

(an example with one individual- and one-group level variable: x_{1ij} and x_{2j})

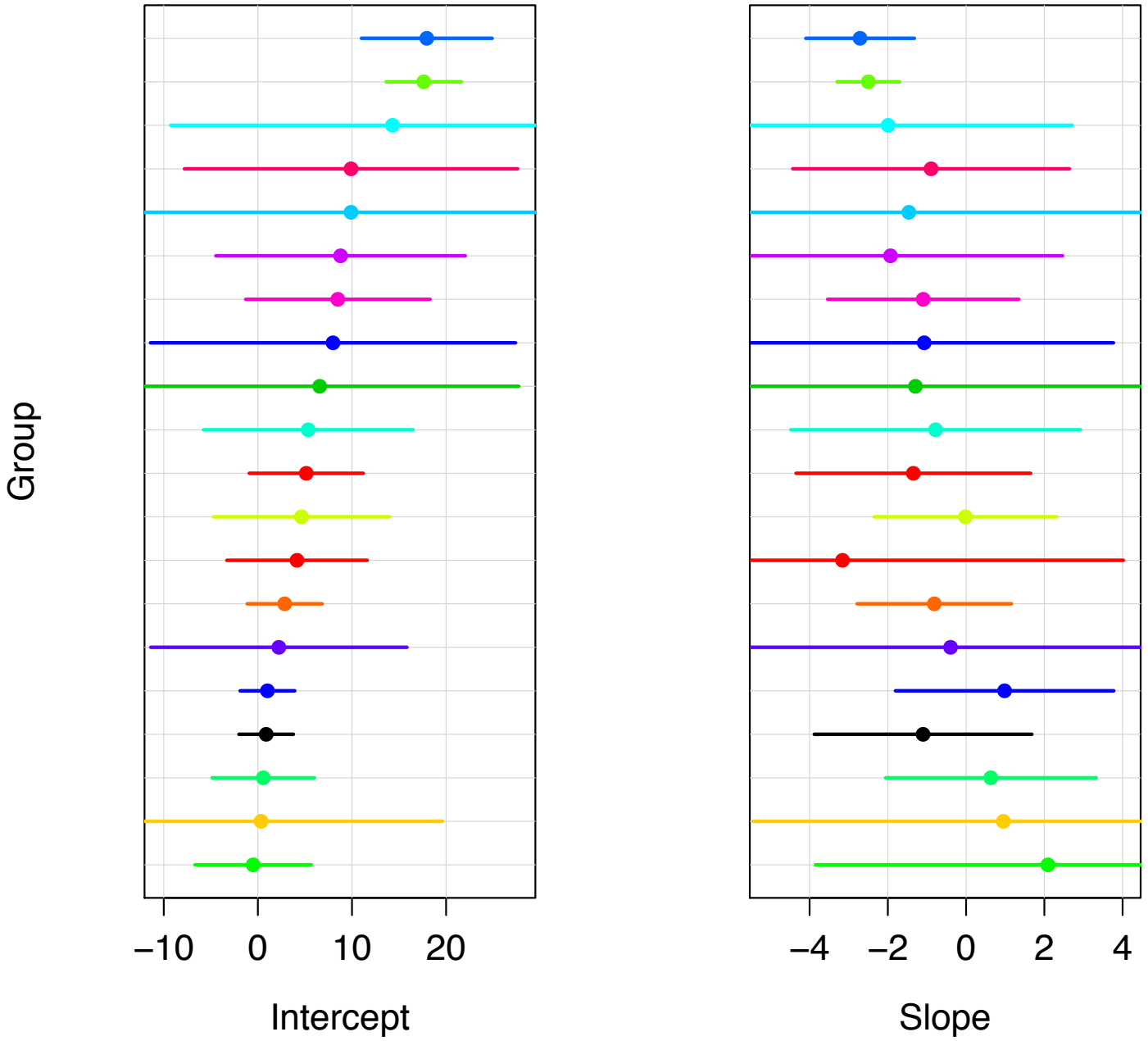

y
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
x


y
x
−1
0
1
2
3
4
5
6

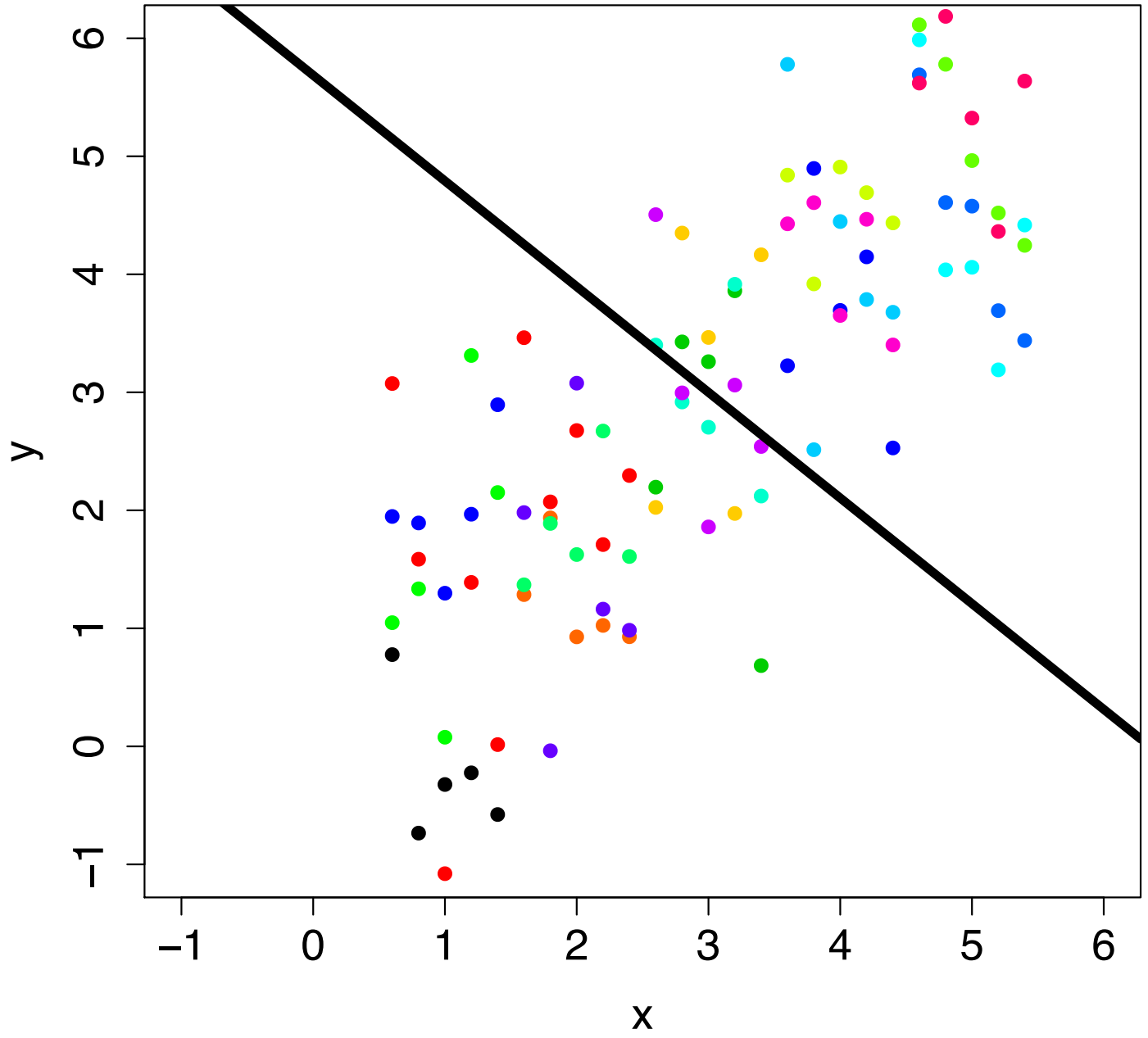

y
6
5
4
3
2
1
0
−1
−1
0
1
2
3
4
5
6
x


y
x
−1
0
1
2
3
4
5
6

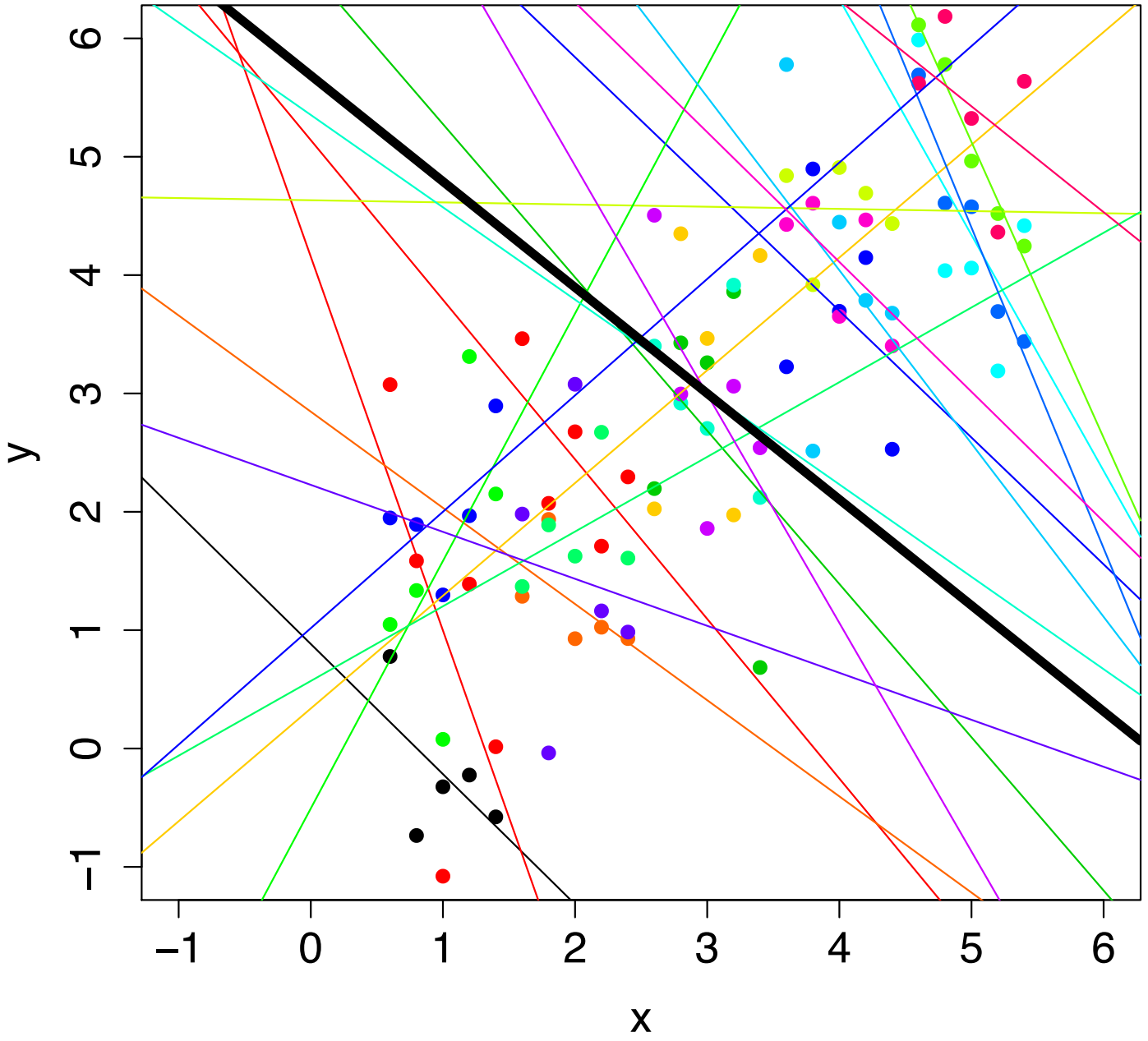
A Separate Regression for Each Group



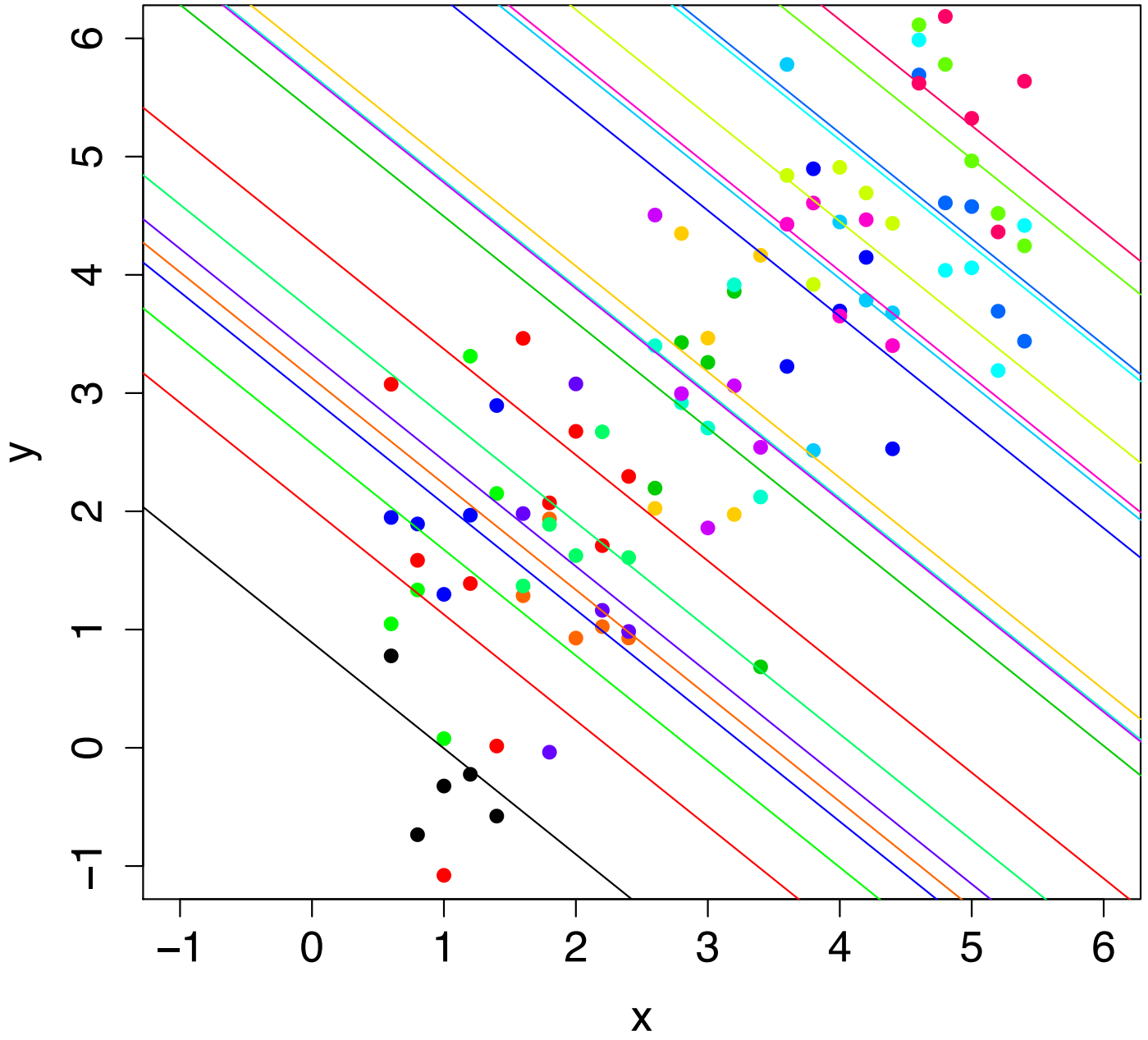
A Fixed Effects ("Within") Model



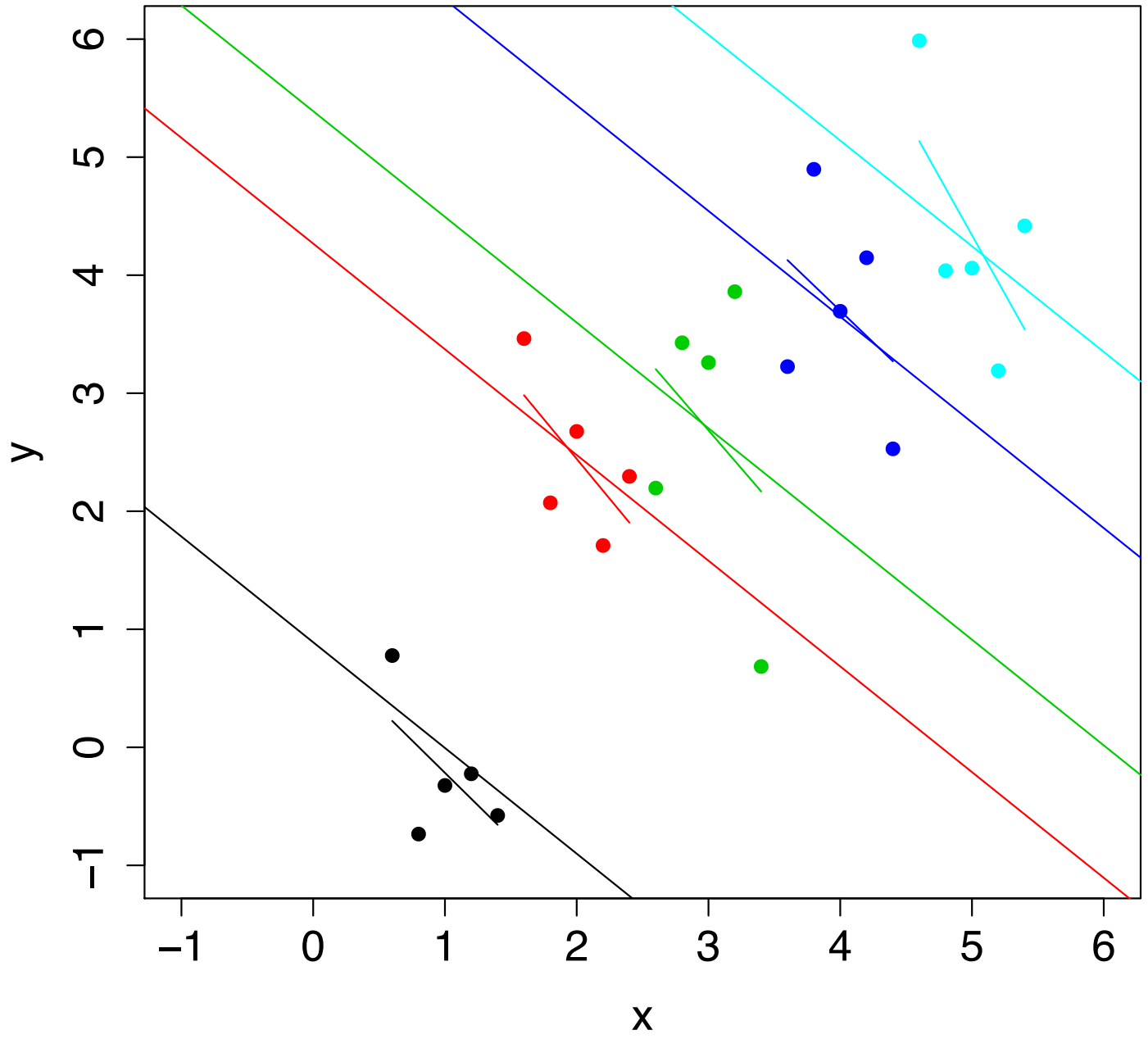
A Separate Regression for Each Group



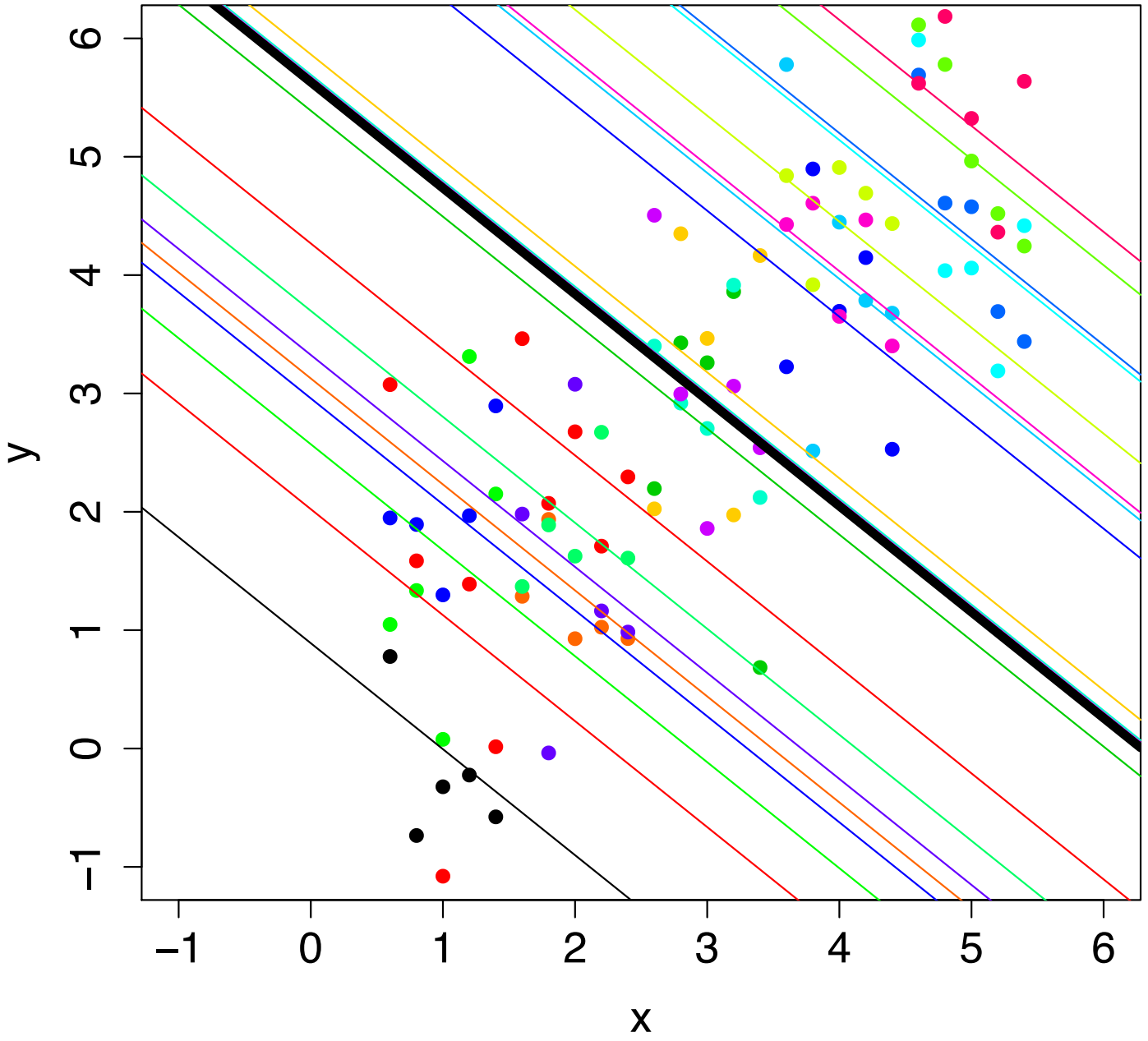
A Random Intercepts Model (with “Within” Lines)



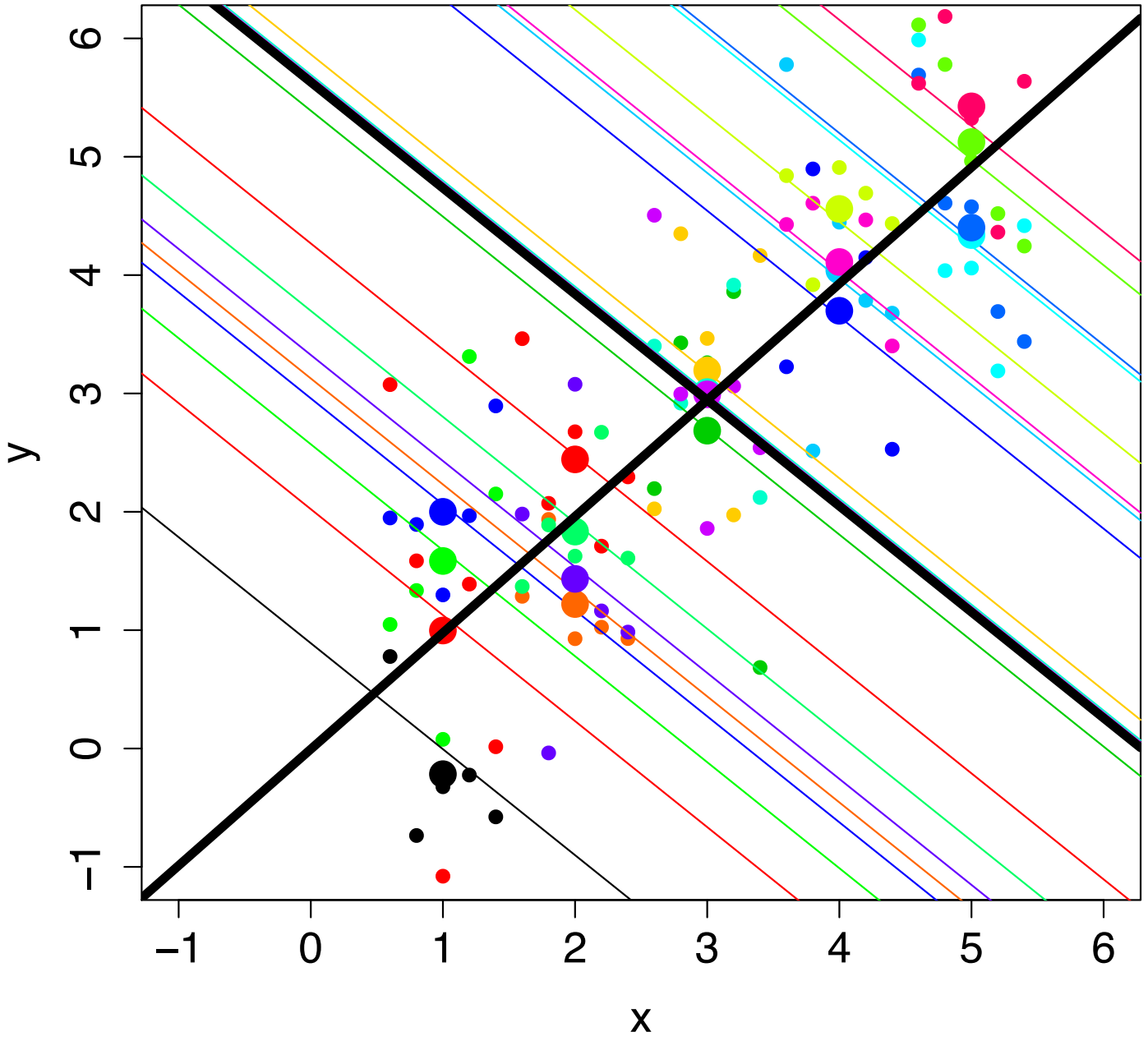
Random Intercepts versus Separate Regressions



Random Intercepts Model with Group-Specific "Within" Lines and an Overall "Within" Line



Group-Specific “Within” Lines, an Overall “Within” Line, and a “Between” Line



Adding a Random Slope…
A "Random Intercept, Random Slope" Model

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2_e)$$

(an example with one individual- and one-group level variable: x_{1ij} and x_{2j})

Adding a Random Slope…
A "Random Intercept, Random Slope" Model

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2j} + e_{ij}$$

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + u_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$

$$[u_{0j}] \sim N(0, \Omega_u) : \ \Omega_u = [\sigma^2_{u0}]$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2_e)$$

- "the term $u_{1j}x_{1ij}$ can be regarded as a random interaction between group and x_1" (Snijders and Bosker)
- it is possible to exclude the covariance, but rarely if ever makes sense

"Random Intercept" versus "Random Slope" versus "Fixed Effects" Models

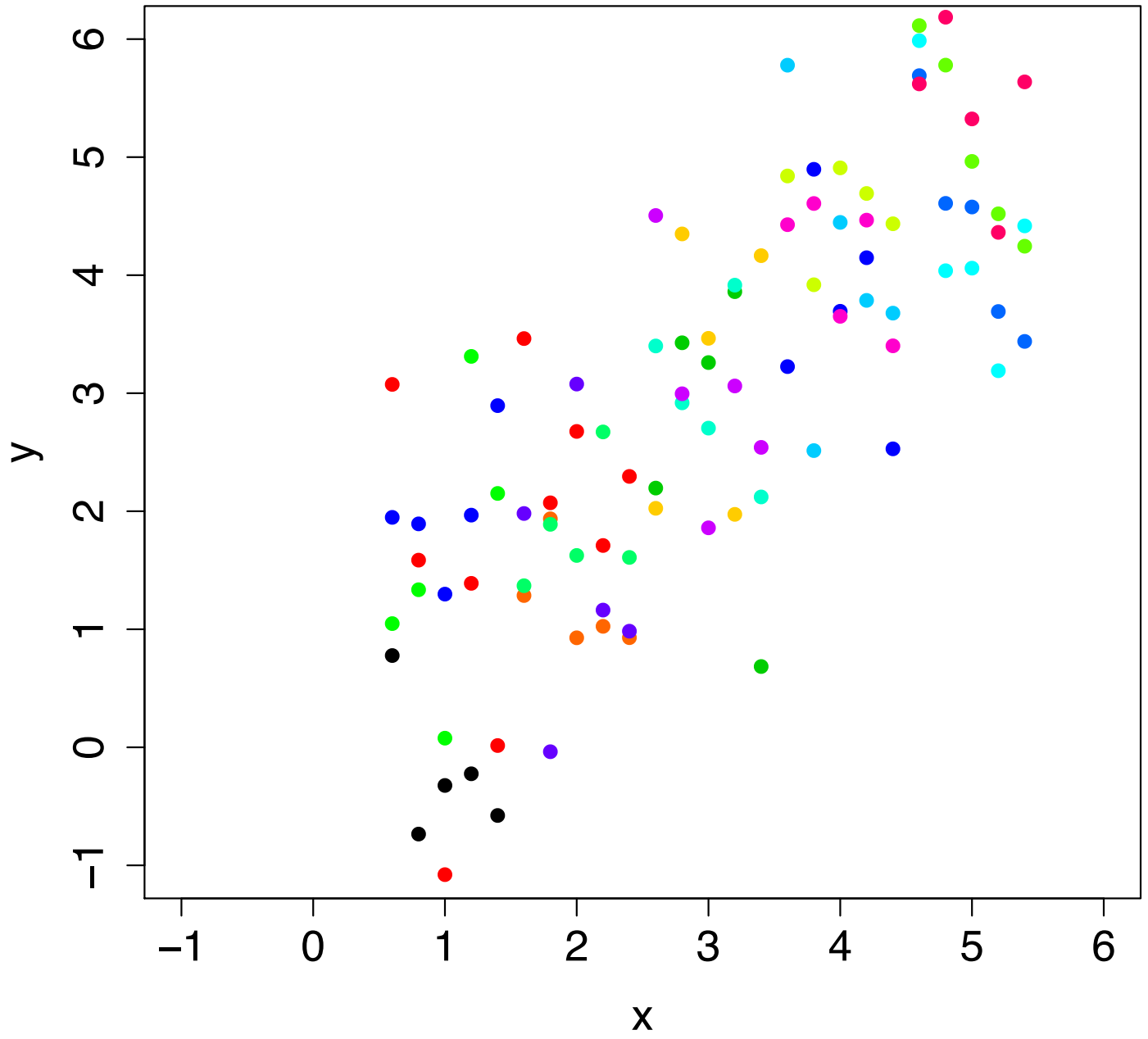
$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + u_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$

$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + u_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$

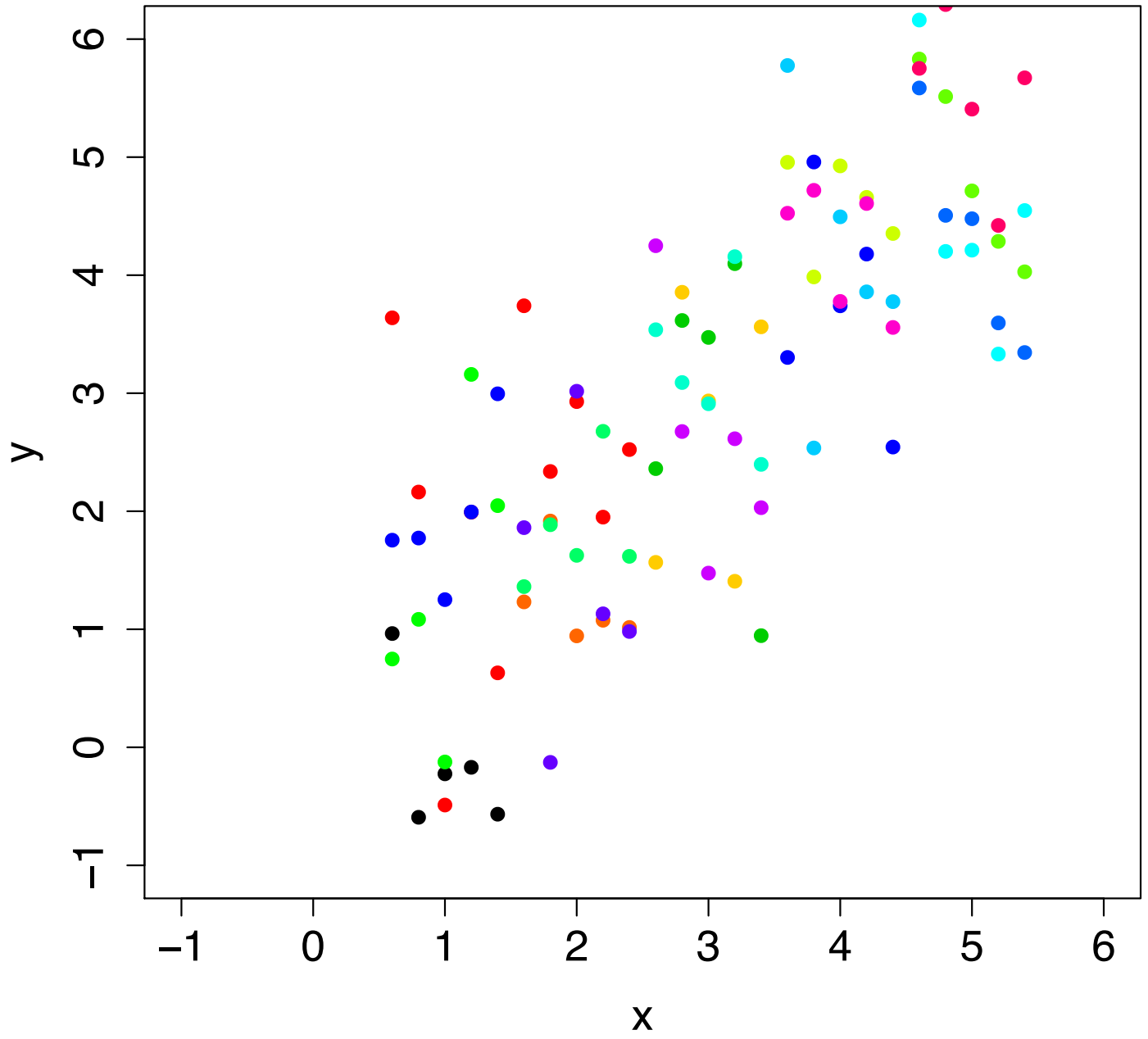
$$y_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1ij} + u_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$

- in fixed effects models, the u_j's (dummy variables for each higher-level unit, other than a reference unit) only apply to the intercepts… not the slopes on x_{ij}

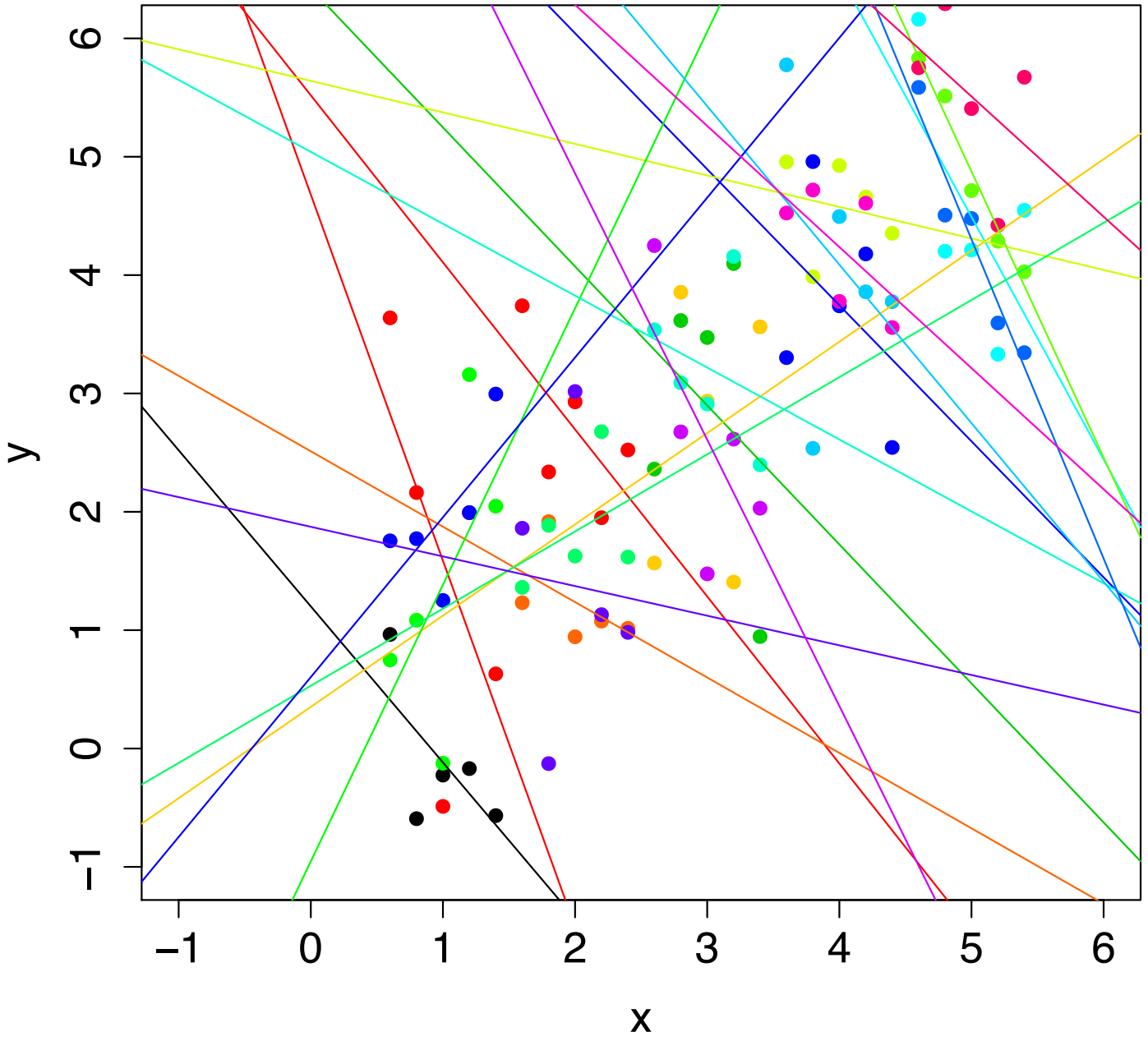
* in practice "random slope" models include random intercepts too


y
x
−1
0
1
2
3
4
5
6

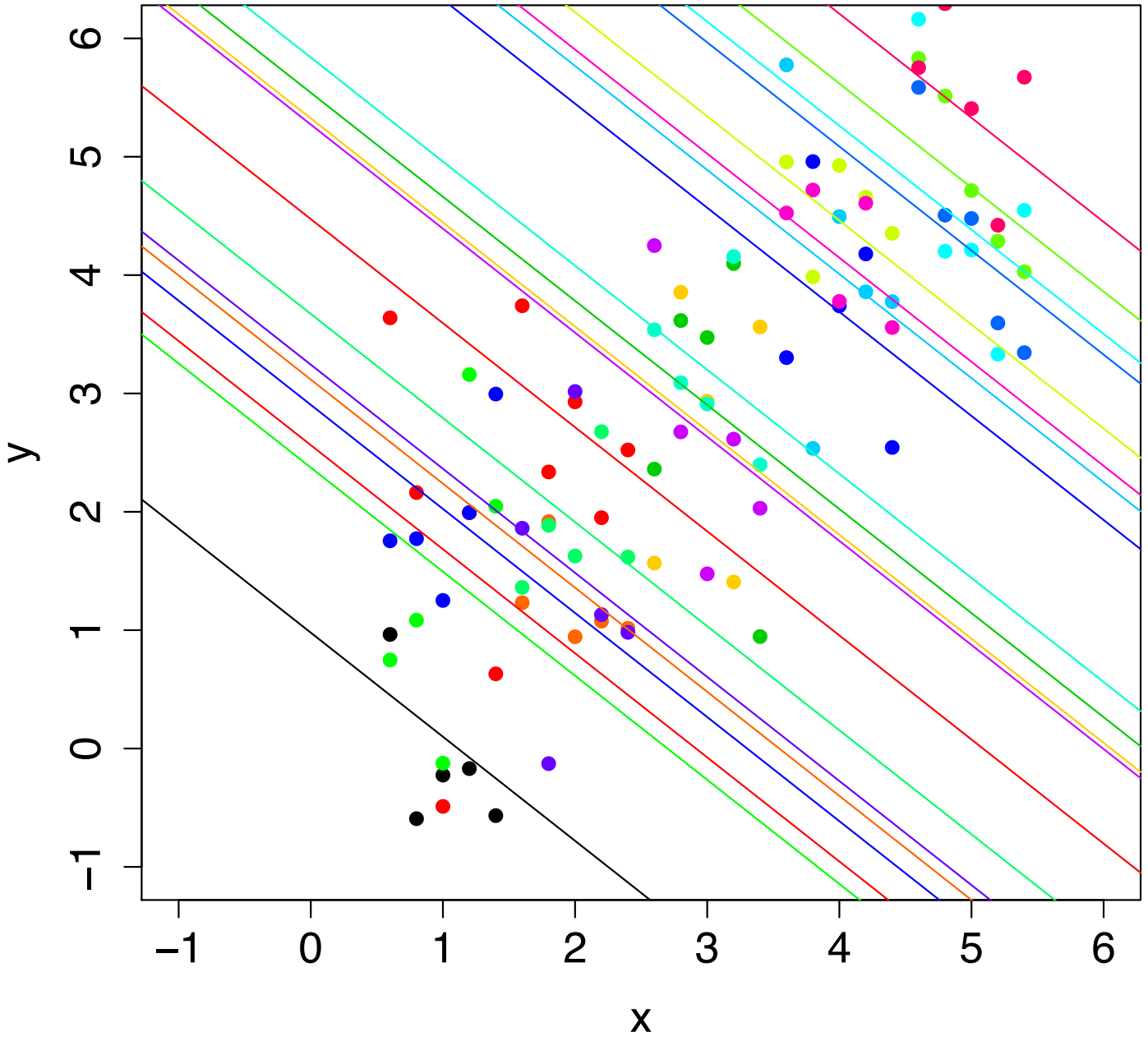
Some Slightly Different (Fake) Data



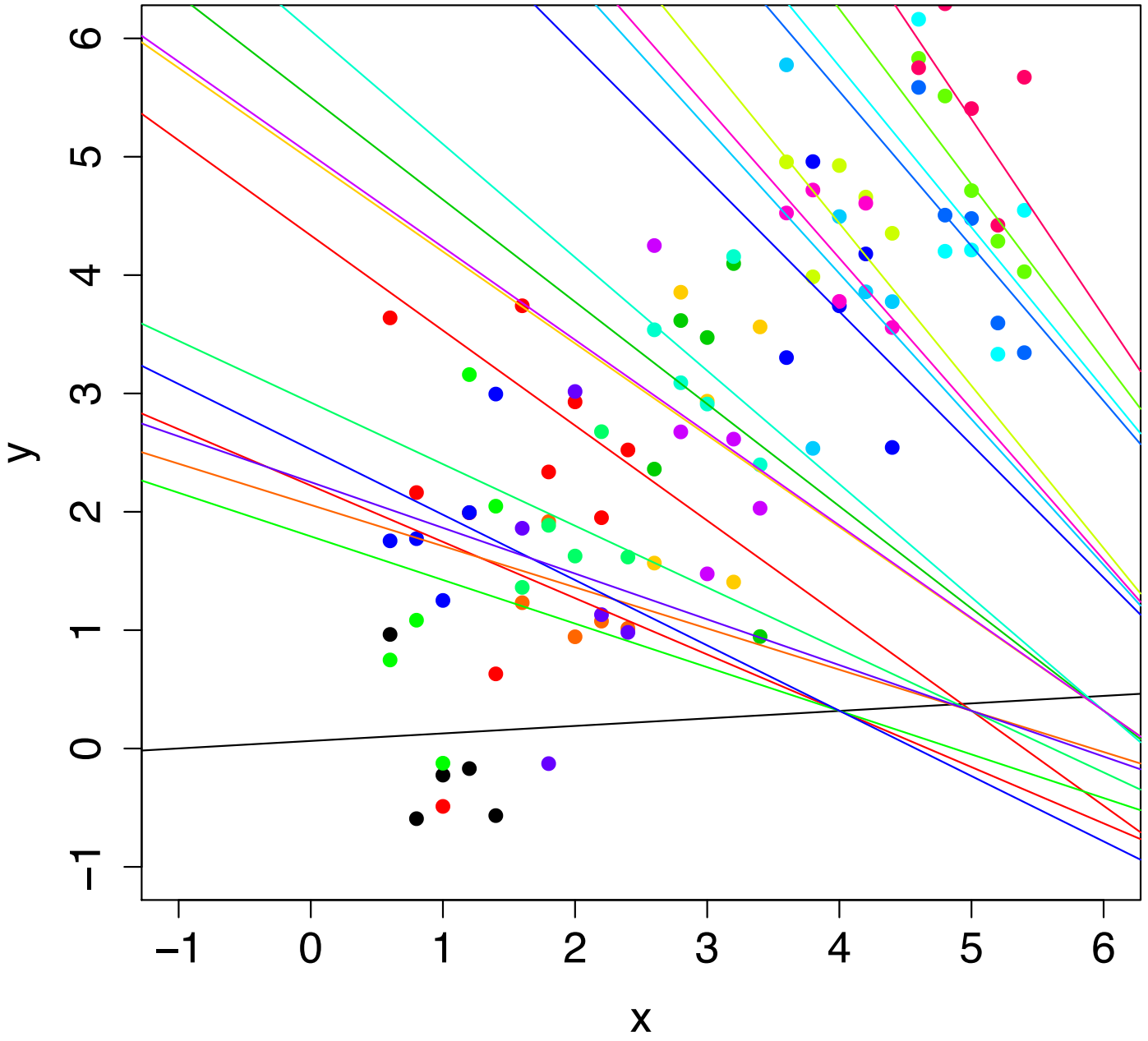
A Separate Regression for Each Group



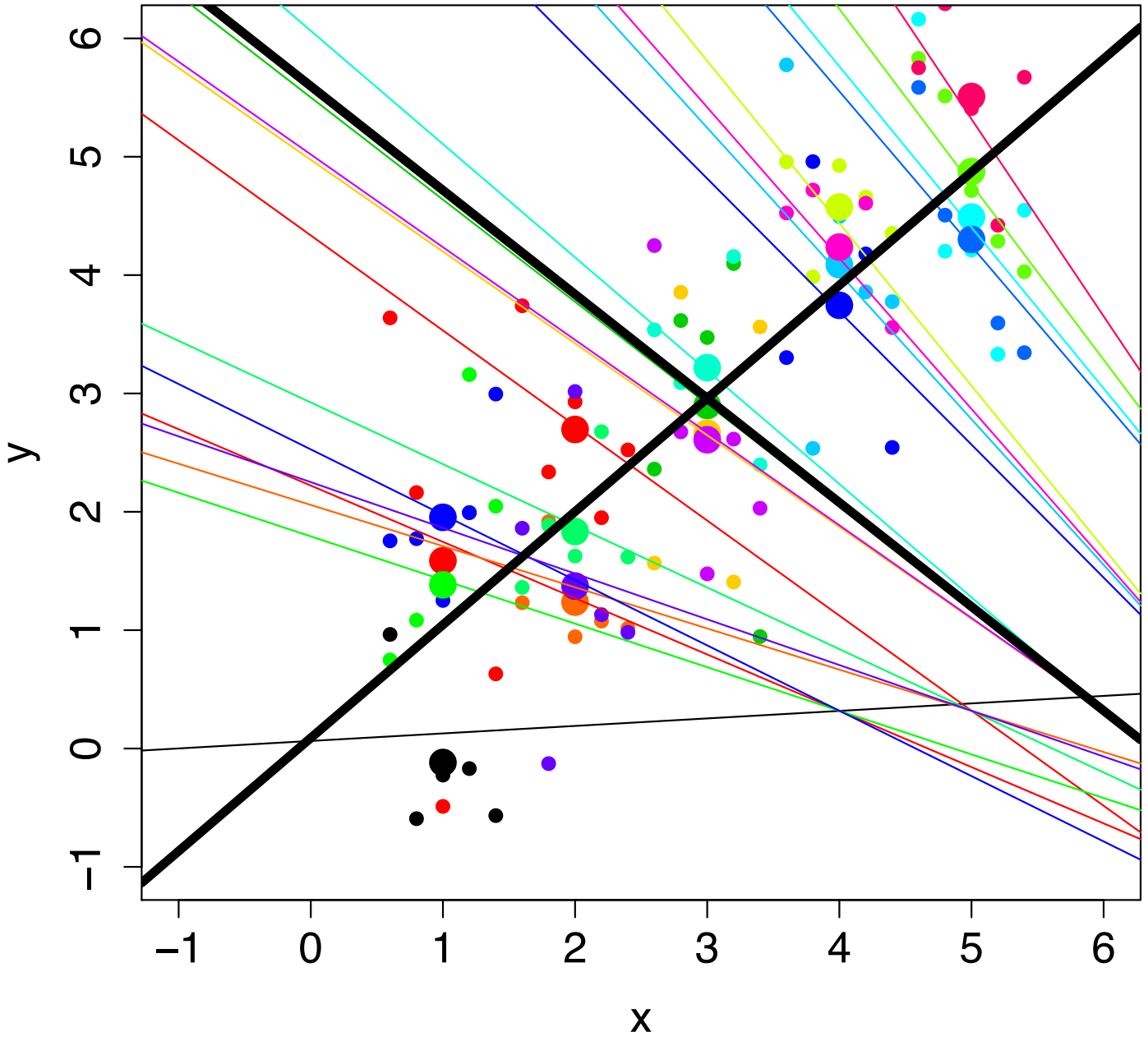
Random Intercepts ("Within" Relationships Constrained to a Common Slope)



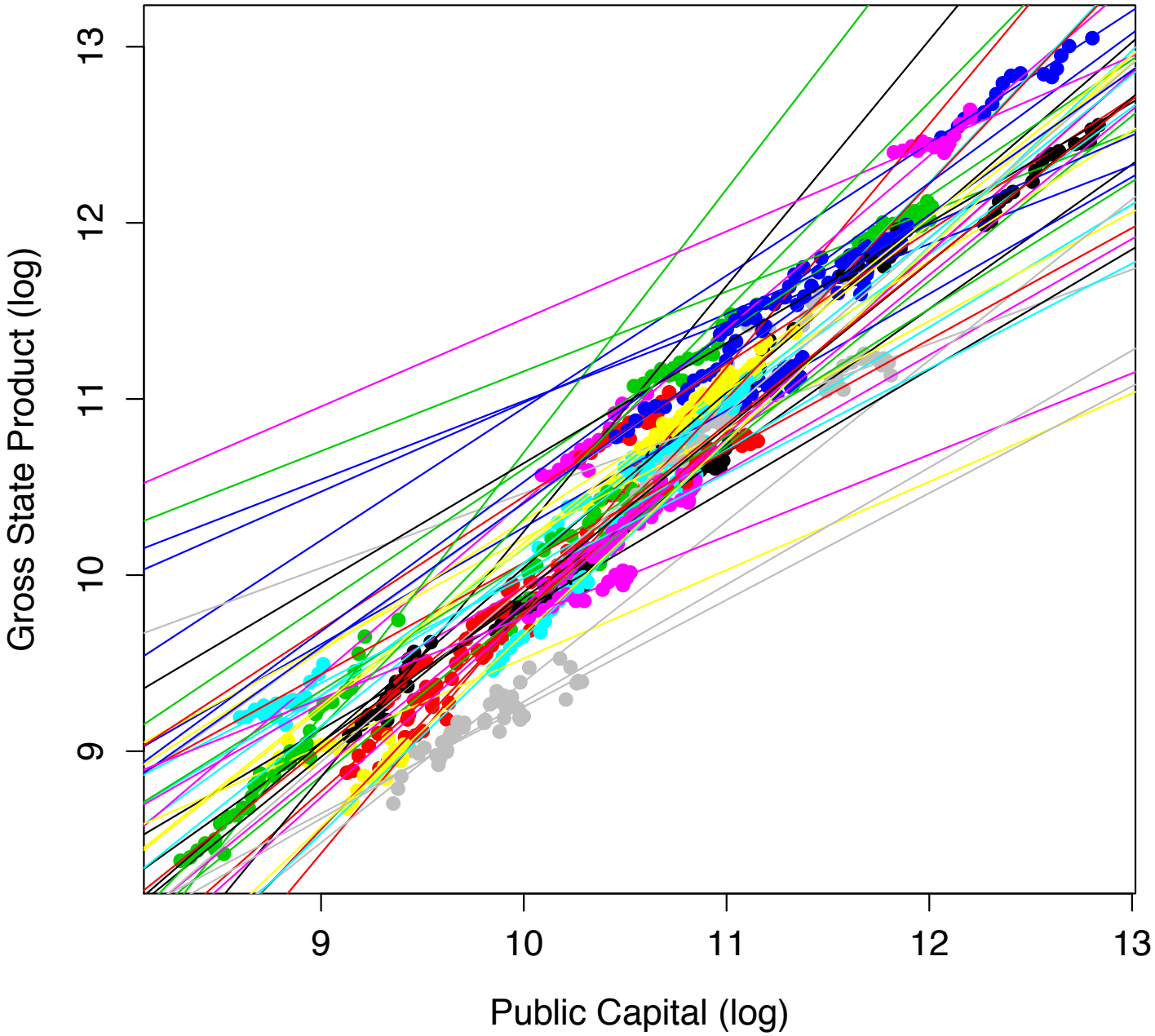
Random Slopes ("Within" Relationships Allowed to Vary, Correlated with Random Intercepts)



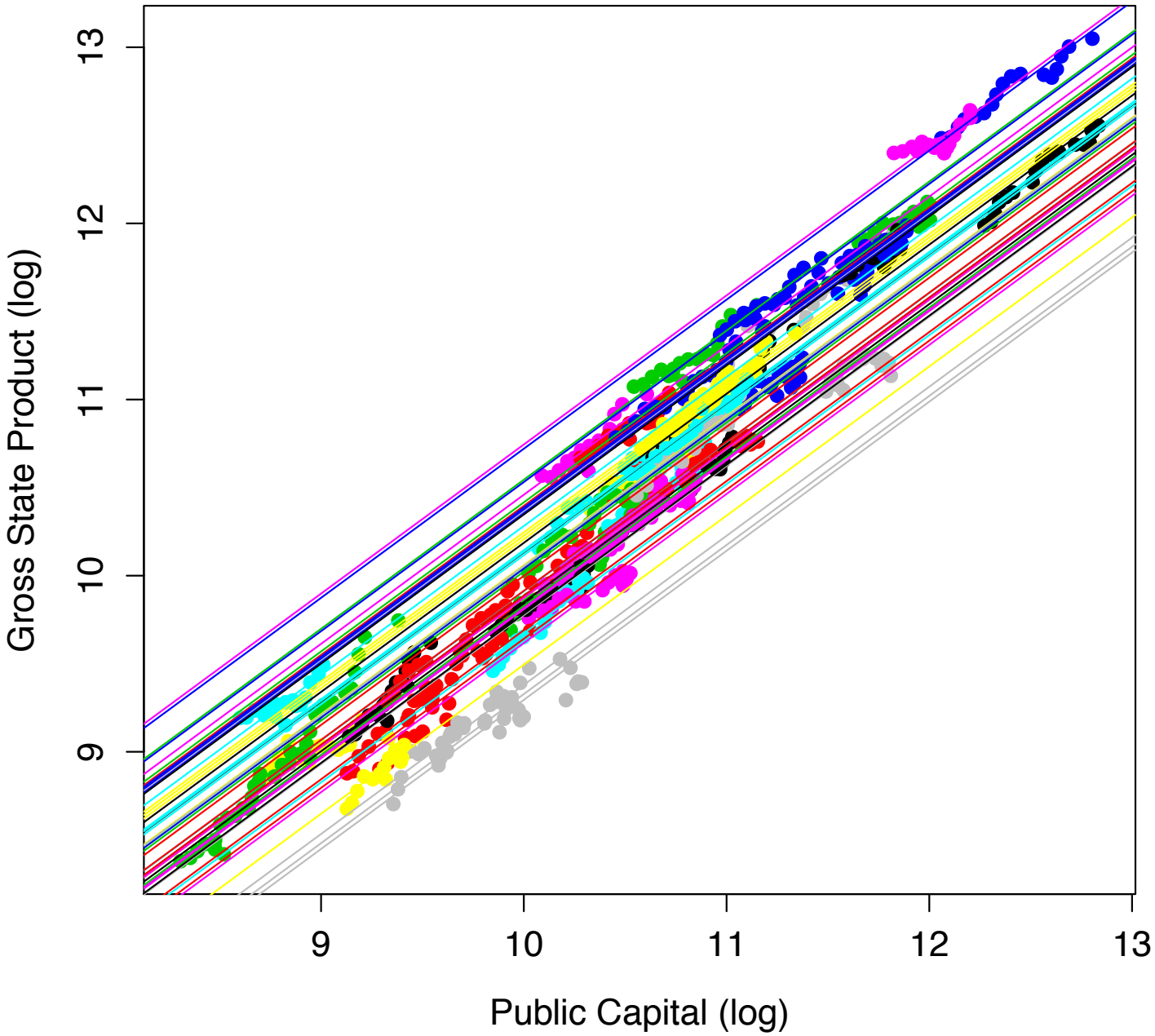
Random (Group-Specific) “Within” Lines, an Overall “Within” Line, and a “Between” Line



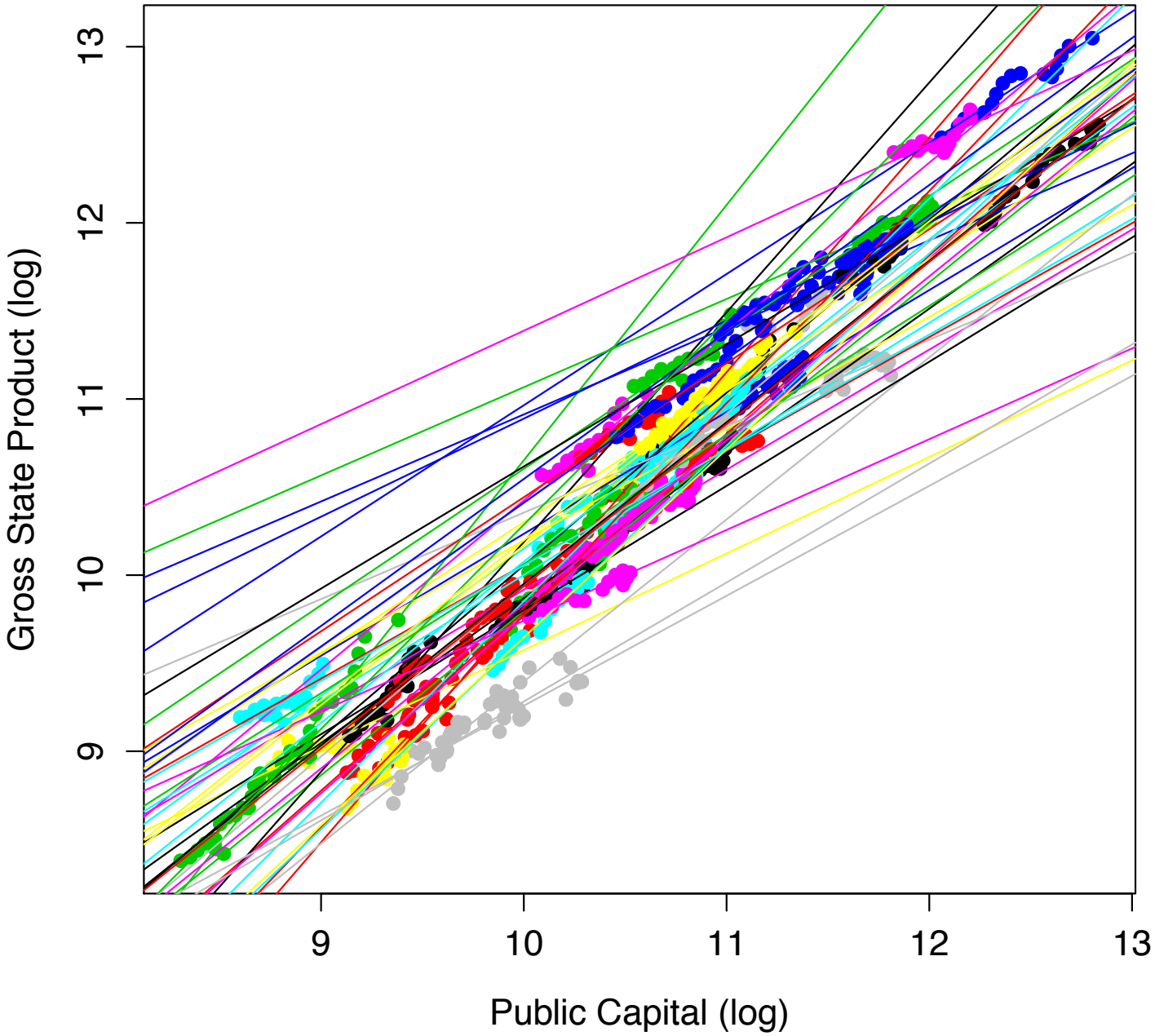
A Separate Regression by State (Produc data)



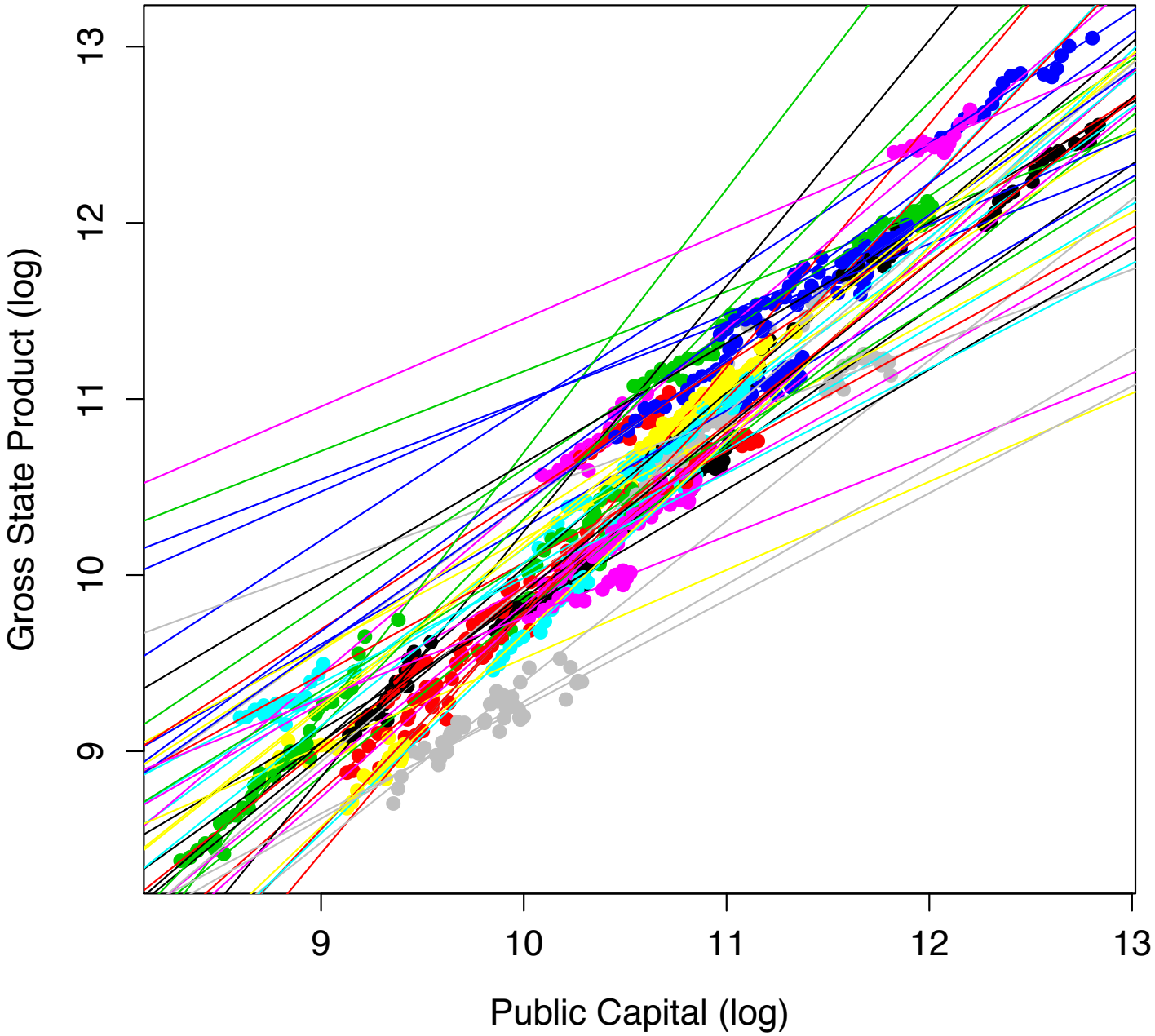
A Random Intercepts Model (Produc data)



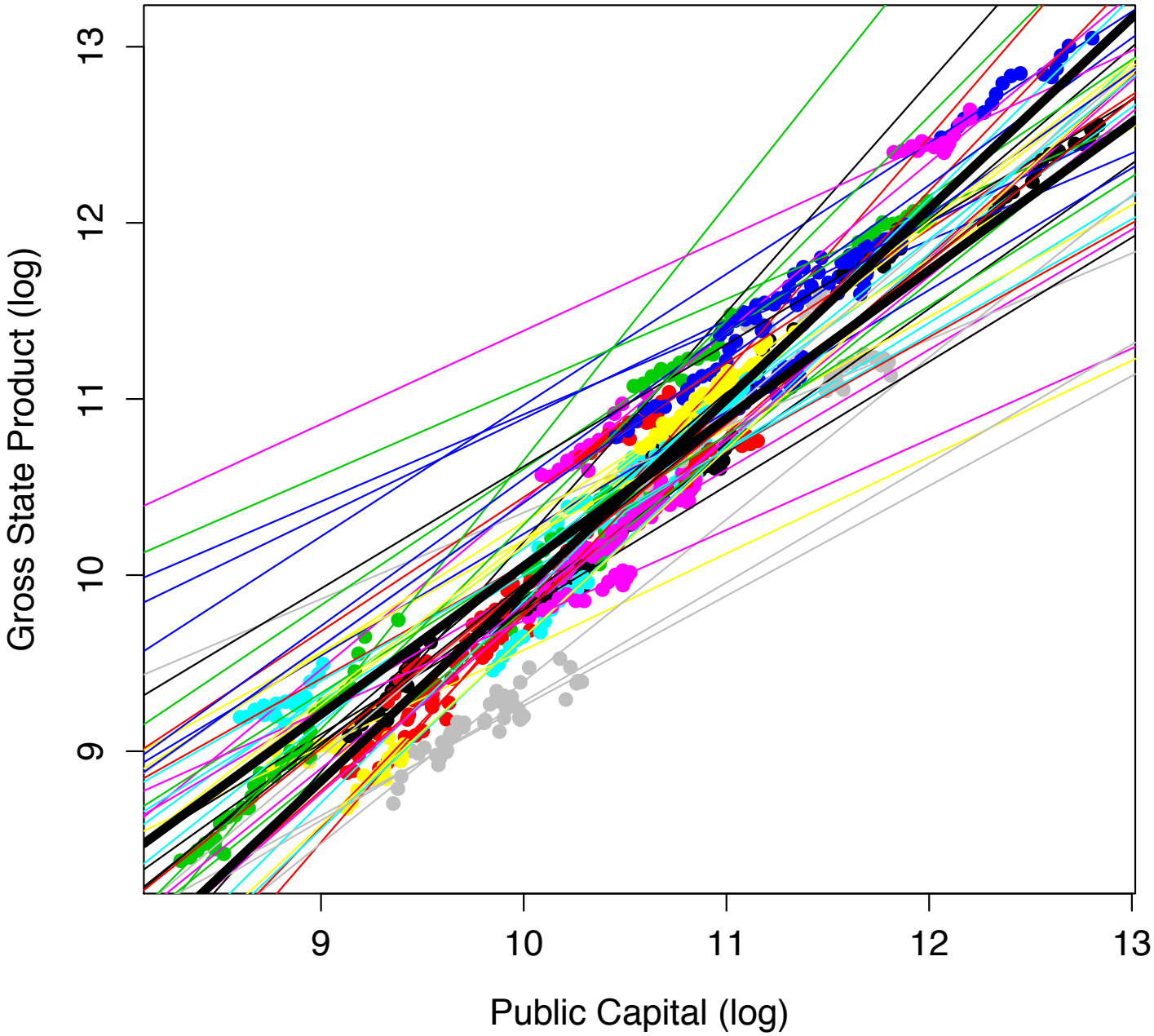
A Random Slopes Model (Produc data)



A Separate Regression by State (Produc data)



Random (State-Specific) "Within" Lines, an Overall "Within" Line, and a "Between" Line (Produc data)



Random Intercept and Random Slope Models

```
> summary(lmer(log(gsp) ~ lpcD + lpcM + (1 | state), prod))
Linear mixed model fit by REML ['lmerMod']
Formula: log(gsp) ~ lpcD + lpcM + (1 | state)
   Data: prod

REML criterion at convergence: -1935.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1488 -0.5983  0.0350  0.5721  4.2660

Random effects:
 Groups   Name        Variance Std.Dev.
 state    (Intercept) 0.07842  0.28003
 Residual             0.00381  0.06172
Number of obs: 816, groups:  state, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.76844    0.47260   -1.63
lpcD         0.84651    0.01305   64.85
lpcM         1.06770    0.04454   23.97
```

```
> summary(lmer(log(gsp) ~ lpcD + lpcM + (lpcD | state), prod))
Linear mixed model fit by REML ['lmerMod']
Formula: log(gsp) ~ lpcD + lpcM + (lpcD | state)
   Data: prod

REML criterion at convergence: -2173.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2236 -0.6011  0.0200  0.6044  3.5322

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 state    (Intercept) 0.079094 0.28124
          lpcD        0.057762 0.24034  0.14
 Residual             0.002418 0.04917
Number of obs: 816, groups:  state, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.93620    0.47016  -1.991
lpcD         0.84021    0.03671  22.888
lpcM         1.08356    0.04430  24.459
```

Random Intercept and Random Slope Models

```
> summary(lmer(log(gsp) ~ lpcD + lpcM + (1 | state), prod))
Linear mixed model fit by REML ['lmerMod']
Formula: log(gsp) ~ lpcD + lpcM + (1 | state)
   Data: prod

REML criterion at convergence: -1935.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1488 -0.5983  0.0350  0.5721  4.2660

Random effects:
 Groups   Name        Variance Std.Dev.
 state    (Intercept) 0.07842  0.28003
 Residual             0.00381  0.06172
Number of obs: 816, groups:  state, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.76844    0.47260   -1.63
lpcD         0.84651    0.01305   64.85
lpcM         1.06770    0.04454   23.97
```

```
> summary(lmer(log(gsp) ~ lpcD + lpcM + (lpcD | state), prod))
Linear mixed model fit by REML ['lmerMod']
Formula: log(gsp) ~ lpcD + lpcM + (lpcD | state)
   Data: prod

REML criterion at convergence: -2173.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2236 -0.6011  0.0200  0.6044  3.5322

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 state    (Intercept) 0.079094 0.28124
          lpcD        0.057762 0.24034  0.14
 Residual             0.002418 0.04917
Number of obs: 816, groups:  state, 48

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.93620    0.47016  -1.991
lpcD         0.84021    0.03671  22.888
lpcM         1.08356    0.04430  24.459
```

A Random Slope Model

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \beta_2 x_{2j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2_e)$$

(an example with one individual- and one-group level variable: x_{1ij} and x_{2j})

A Random Slope Model: With Panel Data

$$y_{ij} = \beta_{0j} + \beta_{1j} time_{ij} + \beta_2 x_{2j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{u00} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

$$e_{ij} \sim N(0, \sigma^2_e)$$

(an example with one individual- and one-group level variable: $time_{ij}$ and x_{2j})

A Random Slope Model: With Panel Data

$$y_{ij} = \beta_{0j} + \beta_{1j} time_{ij} + \beta_2 x_{2j} + \beta_3 time_{ij} x_{2j} + e_{ij}$$
$$\beta_{0j} = \beta_0 + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N(0, \Omega_u) : \ \Omega_u = \begin{bmatrix} \sigma^2_{u00} & \\ \sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$

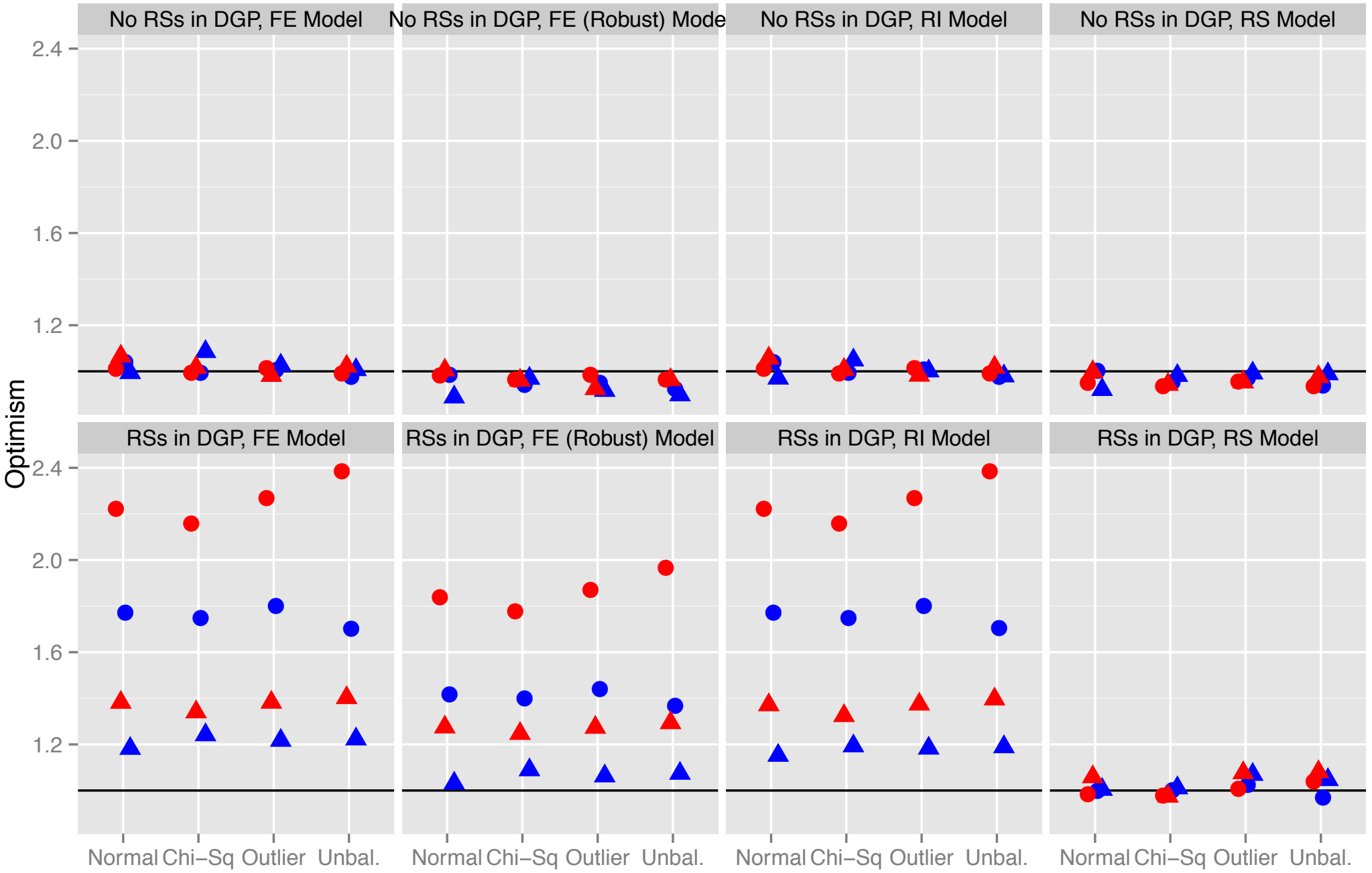
$$e_{ij} \sim N(0, \sigma^2_e)$$

(x_{2j} is some <u>time-invariant</u> property of the unit observed repeatedly)

So… (When) Do You Need Random Slopes?

- sometimes they're a key point of the analysis (why do specific countries differ?), so you want them
  - and if random slopes make no difference, then it's up to you

- what about when they're not the point, and they do make a difference?
  - it can happen that statistically significant results… disappear
  - so do you need to include them (when you do not want your result to evaporate)?
  - definitely a useful robustness check
  - a current debate in the methodological literature…



Source: Bell, Fairbrother, and Jones (under review),
"Fixed and Random effects: making an informed choice"

Bayesian/MCMC Estimation

Bayesian/MCMC Estimation

- philosophy
- practicalities
- software options

1701 –1761

Bayesian/MCMC Estimation: Philosophy

- what's "probability"?
- what's a "p value"?
  - and why aren't there any in the output from lmer?

- what we really want to know…

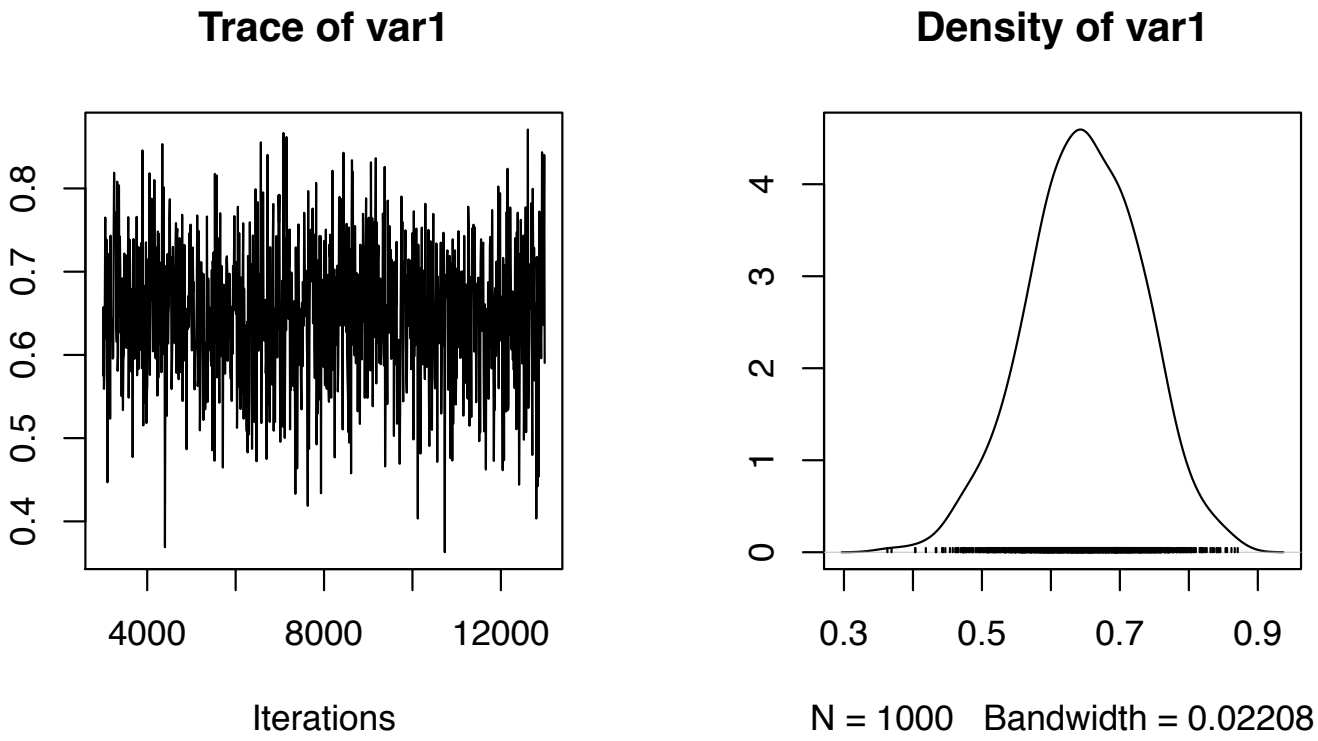
Bayesian/MCMC Estimation: Practicalities

- "Markov chain Monte Carlo"
  - Markov chain -> one thing leads to another…
  - Monte Carlo -> and there's luck involved

- difference from (versions of) maximum likelihood

Bayesian/MCMC Estimation: Practicalities

- "Markov chain Monte Carlo"
  - Markov chain -> one thing leads to another…
  - Monte Carlo -> and there's luck involved



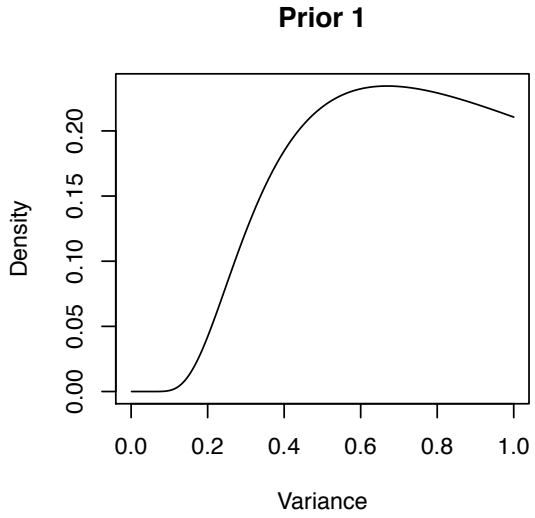
Bayesian/MCMC Estimation: Practicalities

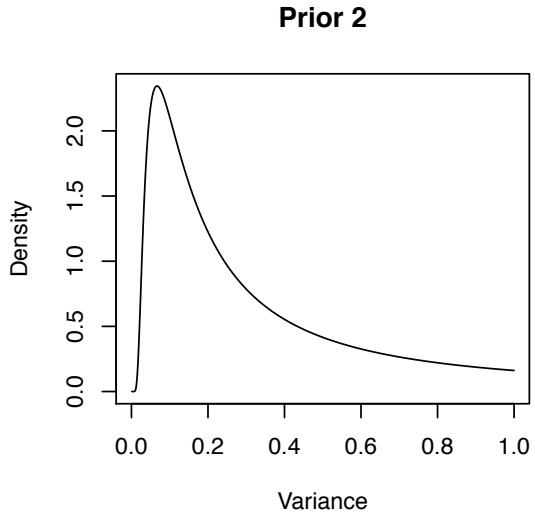
- advantages
  - philosophical
  - inference
  - flexibility

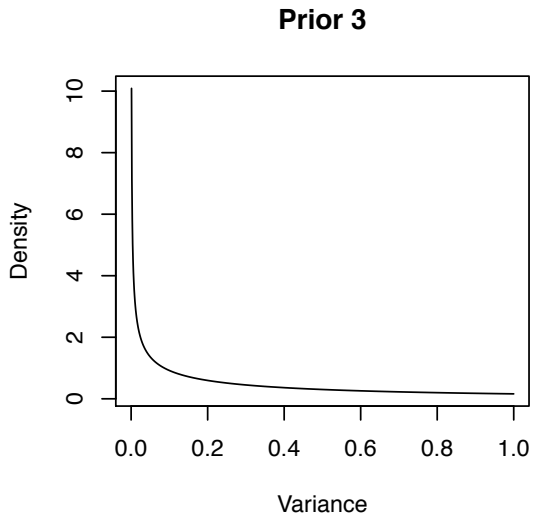
- disadvantages
  - time/computing power
  - a little more technically demanding

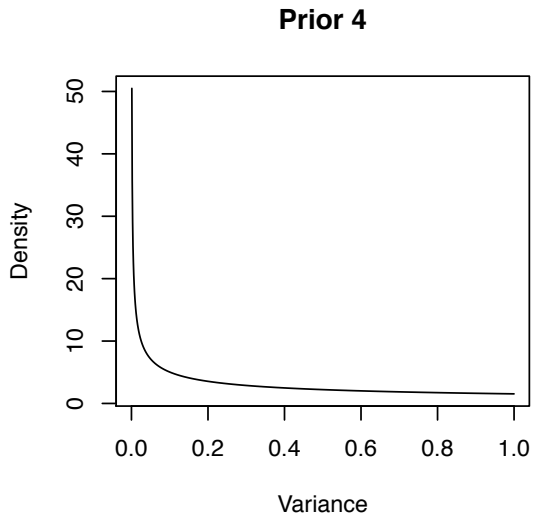
Bayesian/MCMC Estimation: Practicalities

- computer power
- priors
  - in theory
  - in practice
- convergence and mixing
  - things to watch…
  - thinning
  - burn-in



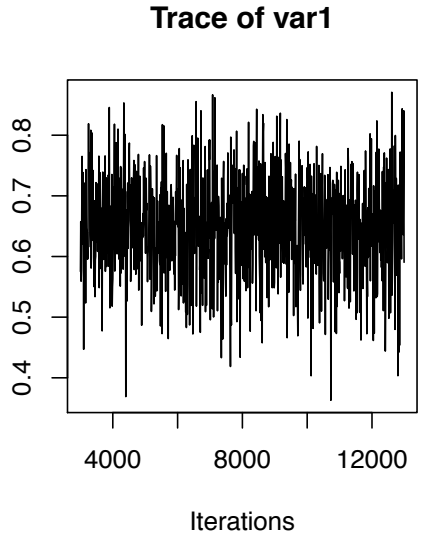


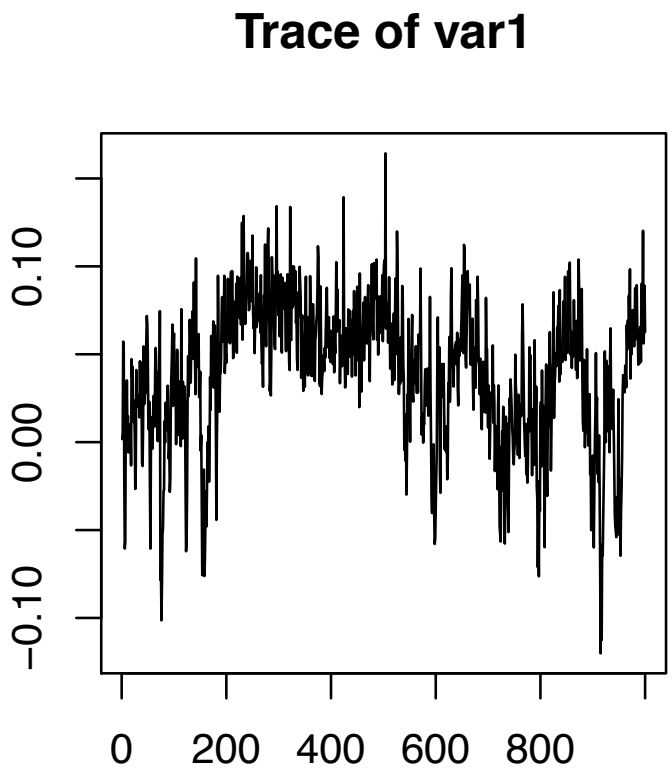


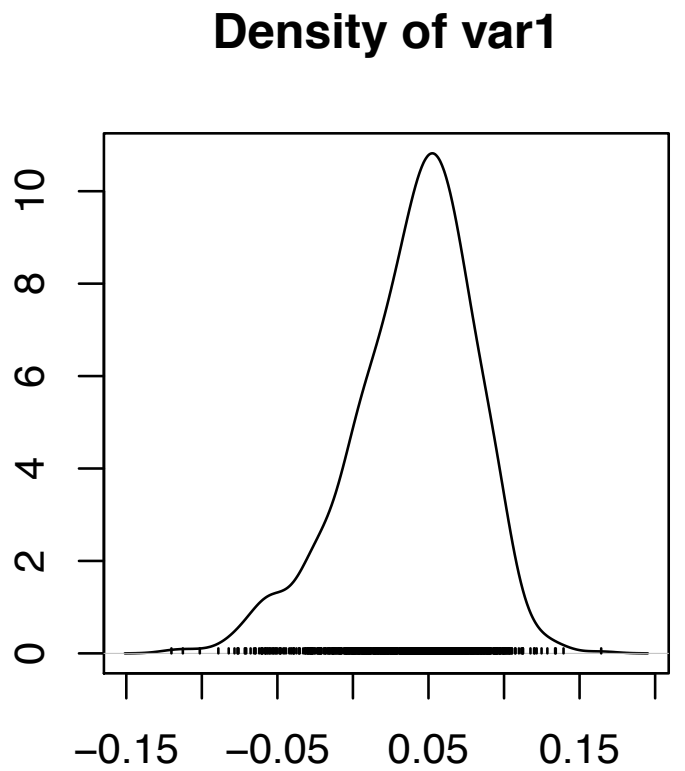


Bayesian/MCMC Estimation: Practicalities

- convergence and mixing
  - things to watch







```
> summary(model.g)
Linear mixed model fit by REML ['lmerMod']
Formula: alcuse ~ ccoa + cpeer * age_14 + (age_14 | id)
   Data: alcohol1

REML criterion at convergence: 603.7

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.59995 -0.40432 -0.07739  0.44372  2.27436

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.2595   0.5094
          age_14      0.1469   0.3832   -0.05
 Residual             0.3373   0.5808
Number of obs: 246, groups:  id, 82

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.65148    0.08119   8.024
ccoa         0.57120    0.14898   3.834
cpeer        0.69518    0.11322   6.140
age_14       0.27058    0.06203   4.362
cpeer:age_14 -0.15138    0.08556  -1.769
```

Maximum Likelihood (lmer)

MCMC (MCMCglmm)

```
> summary(gMC)

 Iterations = 3001:12991
 Thinning interval  = 10
 Sample size  = 1000

 DIC: 534.7168

 G-structure:  ~us(1 + age_14):id

                           post.mean l-95% CI u-95% CI eff.samp
(Intercept):(Intercept).id  0.253921  0.07672  0.44053    566.6
age_14:(Intercept).id      -0.003427 -0.10214  0.09309    392.5
(Intercept):age_14.id      -0.003427 -0.10214  0.09309    392.5
age_14:age_14.id            0.146984  0.04807  0.26052    530.0

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units    0.3575   0.2593   0.4778    678.4

 Location effects: alcuse ~ ccoa + cpeer * age_14

             post.mean l-95% CI u-95% CI eff.samp  pMCMC
(Intercept)    0.64850  0.47150  0.79690     1000 <0.001 ***
ccoa           0.56377  0.30248  0.87402     1115 <0.001 ***
cpeer          0.69780  0.49080  0.91836     1000 <0.001 ***
age_14         0.26943  0.14864  0.39828     1000 <0.001 ***
cpeer:age_14  -0.14898 -0.31903  0.01702     1000  0.082 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

pMCMC

- "pMCMC" is the two times the smaller of the two quantities: MCMC estimates of
  i) the probability that a<0, or
  ii) the probability that a>0,

where a is the parameter value.

- another possibility: calculate the proportion of samples on the other side of zero
  - what's the probability that the sign of the parameter is actually the opposite?

Bayesian/MCMC Estimation: Software

- R package: MCMCglmm (by Hadfield)
- MLwiN
- WinBUGS (http://www.mrc-bsu.cam.ac.uk/software/bugs/)
- JAGS (and rjags, http://mcmc-jags.sourceforge.net/)
- Stan (http://mc-stan.org/)

-> see Gelman and Hill's *Data Analysis Using Regression and Multilevel/ Hierarchical Models*

Generalised Linear Mixed Models (categorical/non-Normal data—binary, etc.)

- in general, GLMMs are to LMMs (linear mixed models) what GLMs are to LMs
- note that "within" results from a FE model for categorical data will not generally be exactly the same as those from a RE model
  - correlation between x_j and u_j will bias estimates of β capturing within relationships (unlike for linear/identity-link models, mean-centering is not a perfect solution)
  - but methodological research suggests the differences in practice never more than trivial

- GLMMs tend to take longer to converge… but there is a trick that can help

Compressing Information:
An Example with Indian Census Data

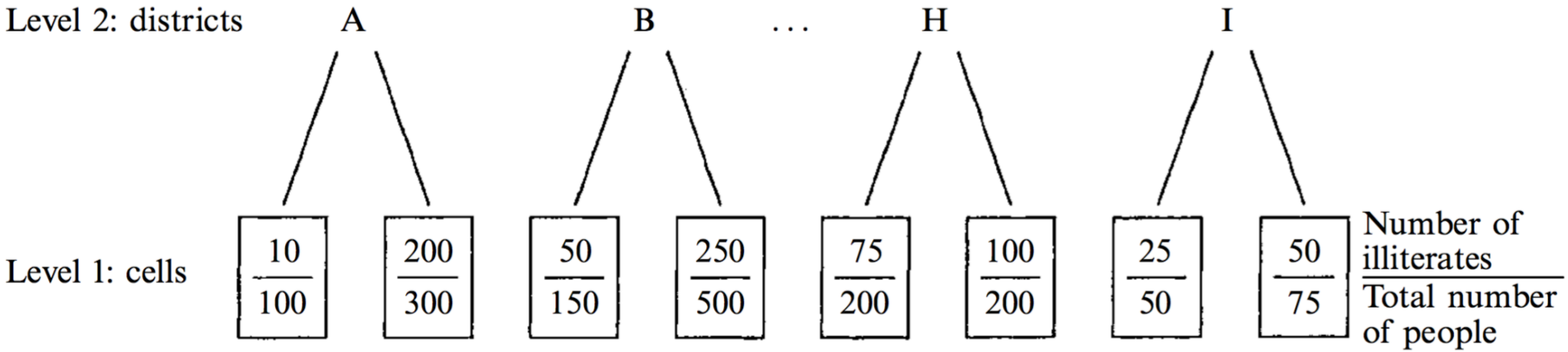


Figure 1. Two-level structure of cells within districts.

Compressing Information: An Example with Indian Census Data

Table 2. Illiteracy and social categories as individual responses.

Respondent	Illiterate	Social categories	District
1	yes	general	A
⋮	⋮	⋮	A
400	no	caste	A
1	no	caste	B
⋮	⋮	⋮	B
650	no	general	B
⋮	⋮	⋮	⋮
1	no	general	H
⋮	⋮	⋮	H
400	yes	caste	H
1	yes	caste	I
⋮	⋮	⋮	I
125	yes	caste	I

Compressing Information:
An Example with Indian Census Data

Table 1. Illiteracy counts by social categories.

Districts	Number of illiterates out of total	
	general	caste
A	10 out of 100	200 out of 300
B	50 out of 150	250 out of 500
⋮	⋮	⋮
H	75 out of 200	100 out of 200
I	25 out of 50	50 out of 75

Illiteracy as a Binomial Response (Series of Bernouilli Trials)

District	Illiterate	Literate	Caste
A	10	90	N
A	200	100	Y
B	50	100	N
B	250	250	Y
…			
H	75	125	N
H	100	100	Y
I	25	25	N
I	50	25	Y

Last Miscellaneous Issues and Cautions

Last Miscellaneous Issues and Cautions: Time

- controlling for time is essential
  - in order to rule out association because of common trending
  - the best functional form may be linear, quadratic, discontinuous, etc.

- with multilevel models, we recognise *clustering*, but (generally) ignore relationships and ordering *within* clusters
  - e.g., in panel data, t=1 is likely to be more similar to t=2 than t=5
  - so watch out for autocorrelation (standard errors will be anticonservative)

Last Miscellaneous Issues and Cautions: Space

- with multilevel models, we recognise *clustering*, but (generally) ignore relationships and ordering *across* clusters
  - but clusters are related in space
  - e.g., Italy shares a border with France… not with Peru
  - so watch out for spillovers (units are not independent)

Last Miscellaneous Issues and Cautions
Statistical Power

- with messy data, as always, it's harder to find relationships
- think carefully about what your N (at different levels) can and can't do for you

Last Miscellaneous Issues and Cautions Mean-Centering… How?

- FE models center by the mean *in the data*
- it may make sense to use all the (country-level) data, not just the years for which you have survey data

Lab 2

http://bit.ly/29qnonA