UW CSSS/POLS 512:
Time Series and Panel Data for the Social Sciences

Basic Concepts for Panel Data

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Panel Data Structure

Suppose we observe our response over both time and place:

$$y_{it} = \mathrm{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

We have units $i = 1, \ldots, N$, each observed over periods $t = 1, \ldots, T$, for a total of $N \times T$ observations

Panel Data Structure

Suppose we observe our response over both time and place:

$$y_{it} = \mathrm{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

We have units $i = 1, \ldots, N$, each observed over periods $t = 1, \ldots, T$, for a total of $N \times T$ observations

Balanced data: all units i have the same number of observations T.

Panel Data Structure

Suppose we observe our response over both time and place:

$$y_{it} = \mathrm{x}_{it}\boldsymbol{\beta} + \varepsilon_{it}$$

We have units $i = 1, \ldots, N$, each observed over periods $t = 1, \ldots, T$, for a total of $N \times T$ observations

Balanced data: all units i have the same number of observations T.

Unbalanced data: some units are shorter in T, perhaps due to missing data, perhaps due to sample selection

All of our discussion in class will assume balanced panels.

Small adjustments may be needed for unbalanced panels, unless the imbalance is due to sample selection, which could lead to significant bias.

You say Panel, I say TSCS. . .

Usages of the term *panel data* vary by field and sub-field

1. Data with large $N \approx 1000$ and small $T \approx 5$ (esp. in economics)

You say Panel, I say TSCS. . .

Usages of the term *panel data* vary by field and sub-field

1. Data with large $N \approx 1000$ and small $T \approx 5$ (esp. in economics)

2. Data with any N, T, and repeated observations on units $i = 1, \ldots, N$ (esp. in opinion research)

You say Panel, I say TSCS. . .

Usages of the term *panel data* vary by field and sub-field

1. Data with large $N \approx 1000$ and small $T \approx 5$ (esp. in economics)

2. Data with any N, T, and repeated observations on units $i = 1, \ldots, N$ (esp. in opinion research)

3. Any data with both $N > 1$ and $T > 1$ (sometimes in political science)

You say Panel, I say TSCS. . .

Usages of the term TSCS data vary by field and sub-field

1. Data with small $N \approx 20$ and medium to large $T > 15$ (esp. in political science)

You say Panel, I say TSCS. . .

Usages of the term TSCS data vary by field and sub-field

1. Data with small $N \approx 20$ and medium to large $T > 15$ (esp. in political science)

2. Data with any N, T, but each cross section has new units;
   so i in period t is a different person from i in period $t + 1$ (esp. opinion research)

You say Panel, I say TSCS. . .

Usages of the term TSCS data vary by field and sub-field

1. Data with small $N \approx 20$ and medium to large $T > 15$ (esp. in political science)

2. Data with any N, T, but each cross section has new units;
   so i in period t is a different person from i in period $t + 1$ (esp. opinion research)

3. Any data with both $N > 1$ and $T > 1$

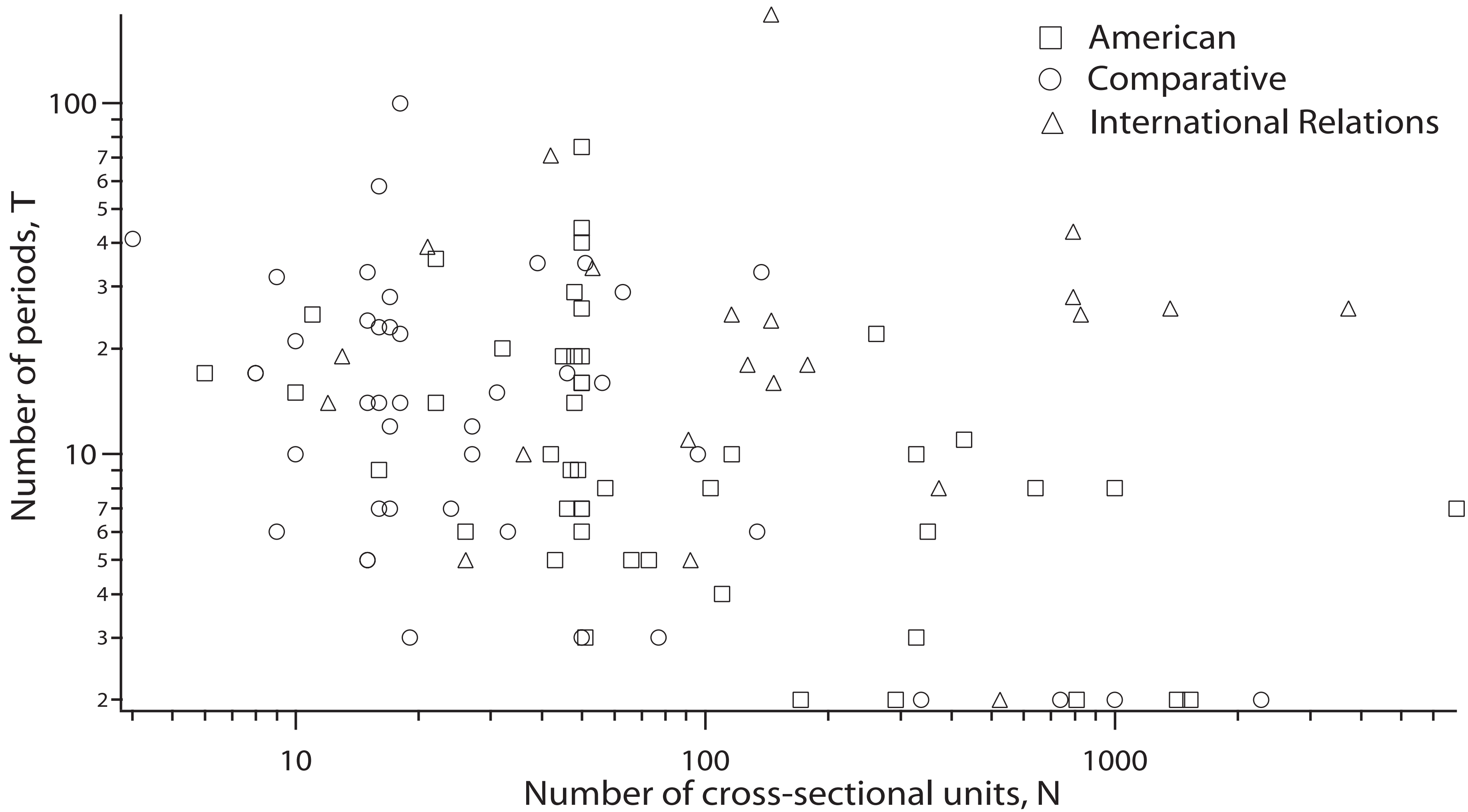
You say Panel, I say TSCS. . .

Data with large N and small T offer different problems and opportunities compared to data with small N and medium T

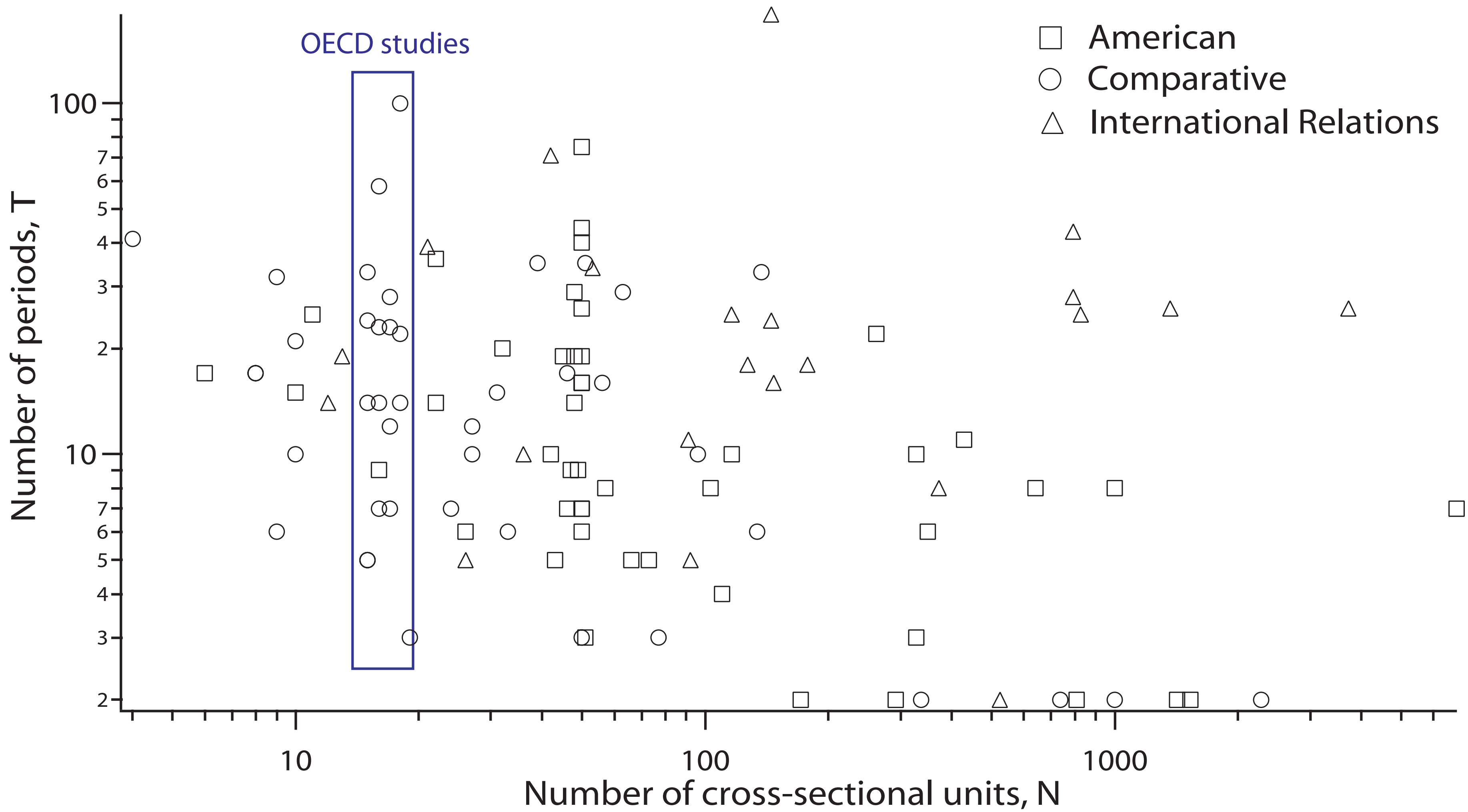
Beware blanket statements about *panel estimators* or *panel data*.

The author—even in a textbook—may be assuming an N and T ubiquitous in his field, but uncommon in yours!

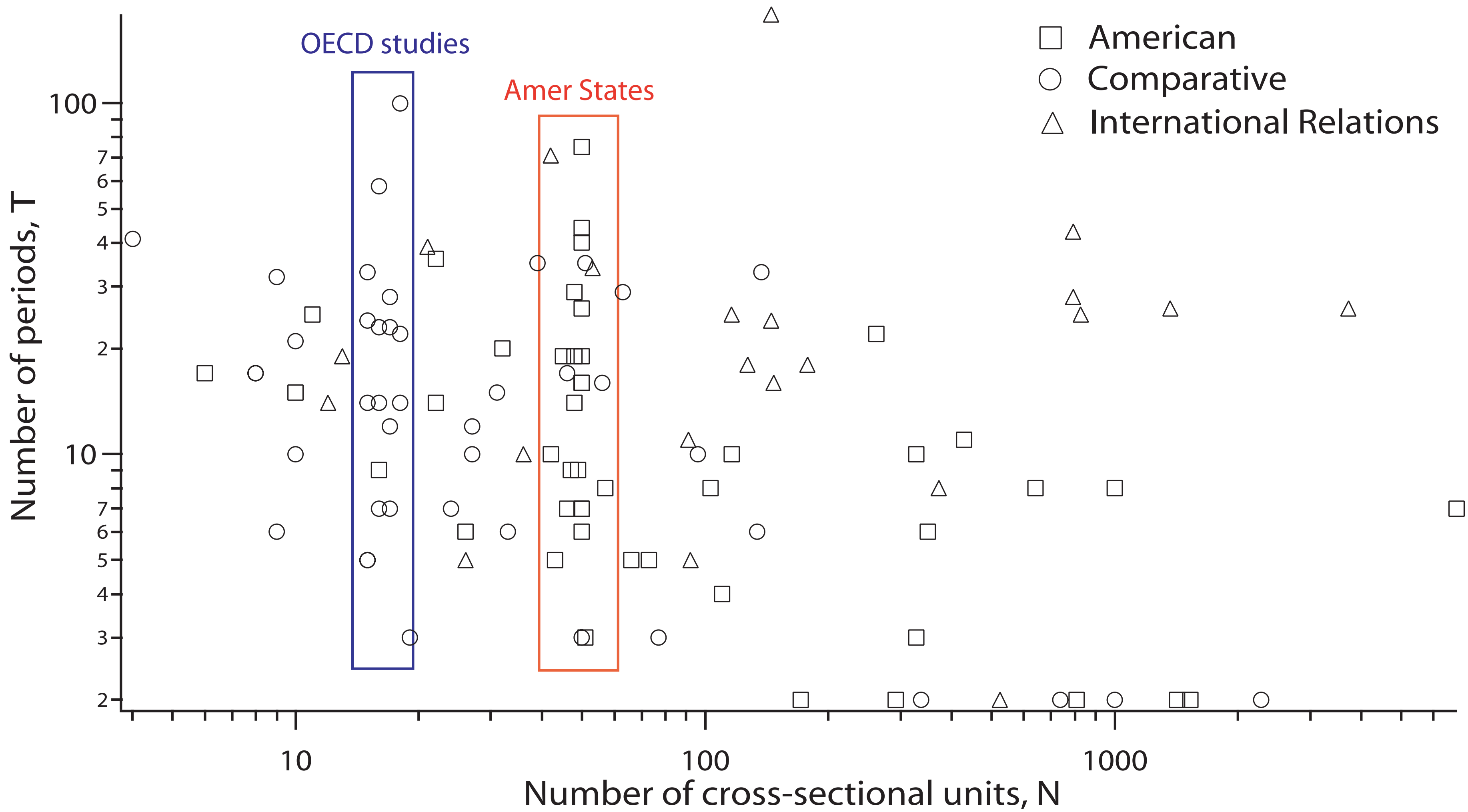
Especially a problem for comparativists learning from econometrics texts

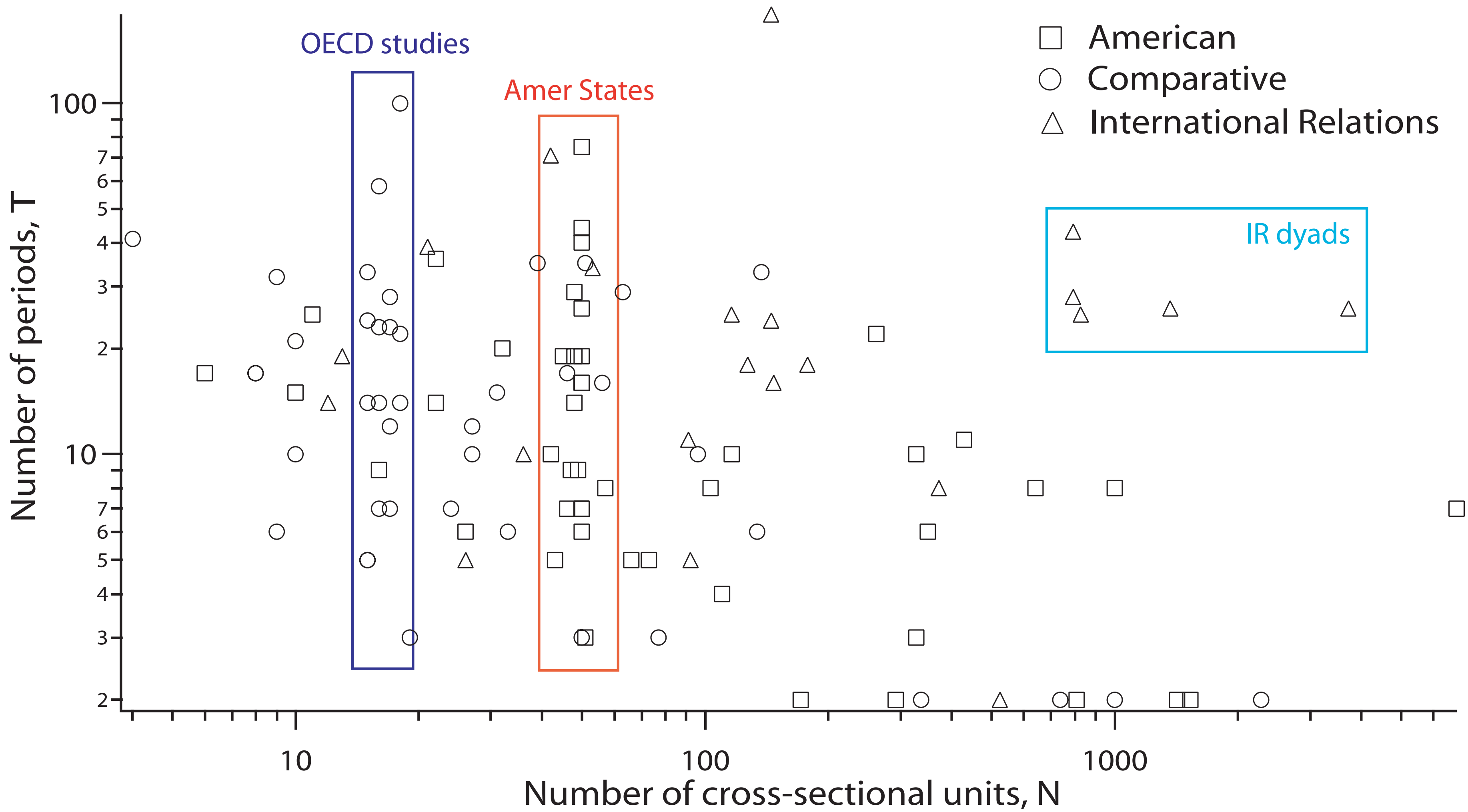


American
Comparative
International Relations
Number of periods, T
Number of cross-sectional units, N
100
10
7
6
5
4
3
2
10
100
1000

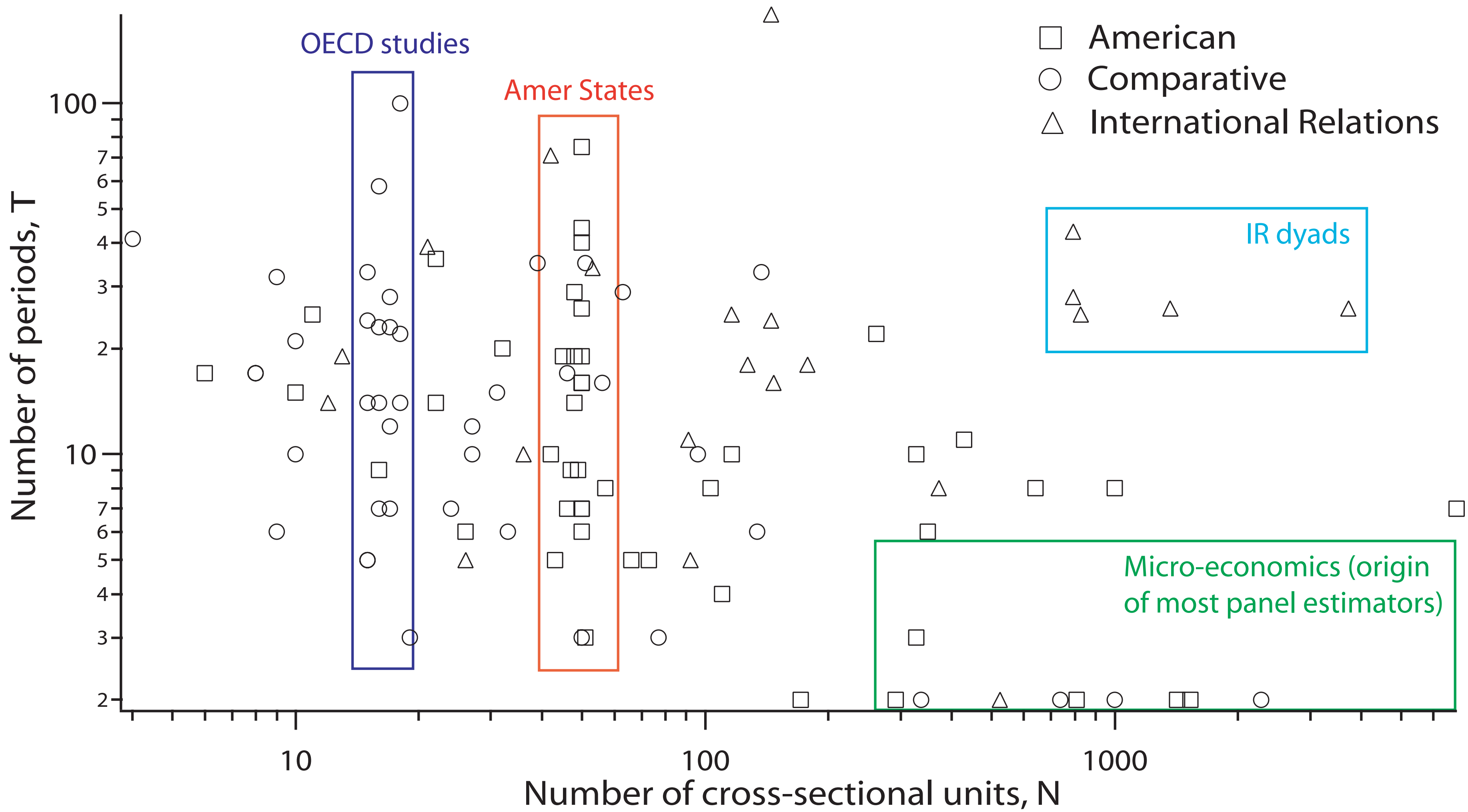


OECD studies
American
Comparative
International Relations
Number of periods, T
100
7
6
5
4
3
2
10
7
6
5
4
3
2
10
100
1000
Number of cross-sectional units, N


OECD studies
Amer States
American
Comparative
International Relations
Number of periods, T
100
7
6
5
4
3
2
10
7
6
5
4
3
2
10
100
1000
Number of cross-sectional units, N



OECD studies
Amer States
American
Comparative
International Relations
IR dyads
100
7
6
5
4
3
2
10
7
6
5
4
3
2
10
100
1000
Number of periods, T
Number of cross-sectional units, N


OECD studies
Amer States
American
Comparative
International Relations
IR dyads
Micro-economics (origin of most panel estimators)
Number of periods, T
Number of cross-sectional units, N
100
10
7
6
5
4
3
2
1000

A pooled TSCS model

$$\text{GDP}_{it} = \phi_1 \text{GDP}_{i,t-1} + \beta_0 + \beta_1 \text{Democracy}_{it} + \varepsilon_{it}$$

This model assumes the same effect of Democracy on GDP for all countries i (β_1)

And influence of past GDP on current GDP is the same for all countries i (ϕ_1)

The shared parameters make this a *Pooled* Time Series Cross Section model

Data storage issues

To get panel data ready for analysis, we need it *stacked* by unit and time period, with a time variable and a grouping variable included:

Cty	Year	GDP	lagGDP	Democracy
1	1962	5012	NA	0
1	1963	6083	5012	0
1	1964	6502	6083	0
...				
1	1989	12530	12266	0
1	1990	12176	12530	0
2	1975	1613	NA	NA
2	1976	1438	1613	0
...				
135	1989	6575	6595	0
135	1990	6450	6575	0

Data storage issues

To get panel data ready for analysis, we need it *stacked* by unit and time period, with a time variable and a grouping variable included:

Cty	Year	GDP	lagGDP	Democracy
1	1962	5012	NA	0
1	1963	6083	5012	0
1	1964	6502	6083	0
...				
1	1989	12530	12266	0
1	1990	12176	12530	0
2	1975	1613	NA	NA
2	1976	1438	1613	0
...				
135	1989	6575	6595	0
135	1990	6450	6575	0

Don't use `lag()` to create lags in panel data!
(*Exception:* okay inside `plm()` formulas)

You need a panel lag command that accounts for the breaks where the unit changes, such as `lagpanel()` in the `simcf` package.

Why use Panel Data?

- More data, which might make inference more precise
(at least if we believe β is the same or similar across units)

Why use Panel Data?

- More data, which might make inference more precise
(at least if we believe β is the same or similar across units)

- Can help with omitted variables, especially if they are time invariant

Why use Panel Data?

- More data, which might make inference more precise
(at least if we believe β is the same or similar across units)

- Can help with omitted variables, especially if they are time invariant

- Some analysis only possible with panel data;
e.g., if variables don't change much over time, like institutions

Why use Panel Data?

- More data, which might make inference more precise
(at least if we believe β is the same or similar across units)

- Can help with omitted variables, especially if they are time invariant

- Some analysis only possible with panel data;
e.g., if variables don't change much over time, like institutions

- Heterogeneity is interesting! As long as we can specify a general DGP for whole panel, can parameterize and estimate more substantively interesting relationships

Why use Panel Data?

If modeled correctly, costs of panel data are born by researcher, not by model or data:

- Differences across the panel would appear the biggest problem,

but we can relax any homogeneity assumption to get a more flexible panel model

Why use Panel Data?

If modeled correctly, costs of panel data are born by researcher, not by model or data:

- Differences across the panel would appear the biggest problem,
but we can relax any homogeneity assumption to get a more flexible panel model

- The price of panel data is a more complex structure to conceptualize and model

Why use Panel Data?

If modeled correctly, costs of panel data are born by researcher, not by model or data:

- Differences across the panel would appear the biggest problem,
but we can relax any homogeneity assumption to get a more flexible panel model

- The price of panel data is a more complex structure to conceptualize and model

- Often need more powerful or flexible estimation tools

Building Time Series into Panel

Consider the ARIMA(p,d,q) model:

$$\Delta^d y_t = \alpha + \mathbf{x}_t \boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{t-p} \phi_p + \sum_{q=1}^{Q} \varepsilon_{t-q} \rho_q + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ is white noise

An encompassing specification for many time series processes

Includes as special cases:

Building Time Series into Panel

Consider the ARIMA(p,d,q) model:

$$\Delta^d y_t = \alpha + \mathbf{x}_t \boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{t-p} \phi_p + \sum_{q=1}^{Q} \varepsilon_{t-q} \rho_q + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ is white noise

An encompassing specification for many time series processes

Includes as special cases:

ARMA(p,q) models: Set $d = 0$

Building Time Series into Panel

Consider the ARIMA(p,d,q) model:

$$\Delta^d y_t = \alpha + \mathbf{x}_t \boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{t-p} \phi_p + \sum_{q=1}^{Q} \varepsilon_{t-q} \rho_q + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ is white noise

An encompassing specification for many time series processes

Includes as special cases:

ARMA(p,q) models: Set $d = 0$

AR(p) models: Set $d = Q = 0$

Building Time Series into Panel

Consider the ARIMA(p,d,q) model:

$$\Delta^d y_t = \alpha + \mathbf{x}_t \boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{t-p} \phi_p + \sum_{q=1}^{Q} \varepsilon_{t-q} \rho_q + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ is white noise

An encompassing specification for many time series processes

Includes as special cases:

ARMA(p,q) models: Set $d = 0$

AR(p) models: Set $d = Q = 0$

MA(q) models: Set $d = P = 0$

Building Time Series into Panel

Consider the ARIMA(p,d,q) model:

$$\Delta^d y_t = \alpha + \mathbf{x}_t \boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{t-p} \phi_p + \sum_{q=1}^{Q} \varepsilon_{t-q} \rho_q + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ is white noise

An encompassing specification for many time series processes

Includes as special cases:

ARMA(p,q) models: Set $d = 0$

AR(p) models: Set $d = Q = 0$

MA(q) models: Set $d = P = 0$

Linear regression: Set $d = P = Q = 0$

Could even be re-written as an error correction model

Multiple Time Series

Now notice that if we had several parallel time series $y_{1t}, y_{2t}, \ldots y_{Nt}$, as for N countries, we could estimate a series of regression models:

$$\Delta^{d_1} y_{1t} = \alpha_1 + \mathbf{x}_{1t}\boldsymbol{\beta}_1 + \sum_{p=1}^{P_1} \Delta^{d_1} y_{1,t-p}\phi_{1p} + \sum_{q=1}^{Q_1} \varepsilon_{1,t-q}\rho_{1q} + \varepsilon_{1t}$$

Multiple Time Series

Now notice that if we had several parallel time series $y_{1t}, y_{2t}, \ldots y_{Nt}$, as for N countries, we could estimate a series of regression models:

$$\Delta^{d_1} y_{1t} = \alpha_1 + \mathbf{x}_{1t}\boldsymbol{\beta}_1 + \sum_{p=1}^{P_1} \Delta^{d_1} y_{1,t-p}\phi_{1p} + \sum_{q=1}^{Q_1} \varepsilon_{1,t-q}\rho_{1q} + \varepsilon_{1t}$$

$$\Delta^{d_2} y_{2t} = \alpha_2 + \mathbf{x}_{2t}\boldsymbol{\beta}_2 + \sum_{p=2}^{P_2} \Delta^{d_2} y_{2,t-p}\phi_{2p} + \sum_{q=2}^{Q_2} \varepsilon_{2,t-q}\rho_{2q} + \varepsilon_{2t}$$

Multiple Time Series

Now notice that if we had several parallel time series $y_{1t}, y_{2t}, \ldots y_{Nt}$, as for N countries, we could estimate a series of regression models:

$$
\begin{aligned}
\Delta^{d_1} y_{1t} &= \alpha_1 + \mathbf{x}_{1t}\boldsymbol{\beta}_1 + \sum_{p=1}^{P_1} \Delta^{d_1} y_{1,t-p}\phi_{1p} + \sum_{q=1}^{Q_1} \varepsilon_{1,t-q}\rho_{1q} + \varepsilon_{1t} \\
\Delta^{d_2} y_{2t} &= \alpha_2 + \mathbf{x}_{2t}\boldsymbol{\beta}_2 + \sum_{p=2}^{P_2} \Delta^{d_2} y_{2,t-p}\phi_{2p} + \sum_{q=2}^{Q_2} \varepsilon_{2,t-q}\rho_{2q} + \varepsilon_{2t} \\
&\ldots \\
\Delta^{d_N} y_{Nt} &= \alpha_N + \mathbf{x}_{Nt}\boldsymbol{\beta}_N + \sum_{p=N}^{P_N} \Delta^{d_N} y_{N,t-p}\phi_{Np} + \sum_{q=N}^{Q_N} \varepsilon_{N,t-q}\rho_{Nq} + \varepsilon_{Nt}
\end{aligned}
$$

Each of these models could be estimated separately

Multiple Time Series

The results would be a panel analysis of a particular kind:

- one with maximum flexibility for heterogeneous data generating processes across units i,
- and no borrowing of strength across units i

Generally, we can write this series of regression models as:

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

We've just written all our time series equations in a single matrix

But estimation is still *separate* for each equation

Be clear what the subscripts and variables are

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it} \boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p} \phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q} \rho_{iq} + \varepsilon_{it}$$

Be clear what the subscripts and variables are

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

- $\mathbf{x}_{it}$ is the *vector* of covariates for unit i, time t. Not just a scalar.

Be clear what the subscripts and variables are

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

- $\mathbf{x}_{it}$ is the *vector* of covariates for unit i, time t. Not just a scalar.

- $\boldsymbol{\beta}_i$ is the *vector* of parameters applied to $\mathbf{x}_{it}$ *just* for a particular unit i, for all periods

Be clear what the subscripts and variables are

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

- $\mathbf{x}_{it}$ is the *vector* of covariates for unit i, time t. Not just a scalar.

- $\boldsymbol{\beta}_i$ is the *vector* of parameters applied to $\mathbf{x}_{it}$ *just* for a particular unit i, for all periods

- P_i is the number of lags of the response used for unit i. Could vary by unit.

Be clear what the subscripts and variables are

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

- $\mathbf{x}_{it}$ is the *vector* of covariates for unit i, time t. Not just a scalar.

- $\boldsymbol{\beta}_i$ is the *vector* of parameters applied to $\mathbf{x}_{it}$ *just* for a particular unit i, for all periods

- P_i is the number of lags of the response used for unit i. Could vary by unit.

- ϕ_{ip} is the AR parameter applied to the pth lag, $\Delta^{d_i} y_{i,t-p}$, for unit i.

Pooling and Partial Pooling

Alternative: we could "borrow strength" across units in estimating parameters

This involves imposing restrictions on (at least some of) the parameters
to assume they are either related or identical across units

Trade-off between flexibility to measure heterogenity,
and pooling data to estimate shared parameters more precisely

Same kind of trade-off is at work in *all* modeling decisions,
and all modeling involves weighing these trade-offs

All models are oversimplifications

Same trade-off is at work in *all* modeling decisions

For example, why can't we estimate, for a standard cross-sectional dataset with a Normally distributed y_i, this inarguably "correct" linear model:

$$y_i = \alpha_i + \mathbf{x}_i \beta_i + \varepsilon_i$$

All models are oversimplifications

Same trade-off is at work in *all* modeling decisions

For example, why can't we estimate, for a standard cross-sectional dataset with a Normally distributed y_i, this inarguably "correct" linear model:

$$y_i = \alpha_i + \mathbf{x}_i \beta_i + \varepsilon_i$$

To do any inference,

to learn anything non-obvious from data,

to reduce any data to a simpler model,

we must impose restrictions on parameters which are arguably false

Panel data simply offers a wider range of choices on which parameters to "pool" and which to separate out

The range of models available for panel data

Full flexibility:

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_i^2)$$

For each i, we need to choose p_i, d_i, q_i and estimate $\alpha_i, \boldsymbol{\beta_i}, \boldsymbol{\phi_i}, \boldsymbol{\rho_i}, \sigma_i^2$

The range of models available for panel data

Full flexibility:

$$\Delta^{d_i} y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_i^2)$$

For each i, we need to choose p_i, d_i, q_i and estimate $\alpha_i, \boldsymbol{\beta_i}, \boldsymbol{\phi_i}, \boldsymbol{\rho_i}, \sigma_i^2$

Full pooling:

$$\Delta^{d} y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^{d} y_{i,t-p}\phi_{p} + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_{q} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

We choose common p, d, q across all i, and estimate common $\alpha, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\phi}, \sigma^2$

Popular panel specifications

Variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Popular panel specifications

Variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Variable slopes and intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Popular panel specifications

Variable lag structures

$$\Delta^{d_i} y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Popular panel specifications

Variable lag structures

$$\Delta^{d_i} y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P_i} \Delta^{d_i} y_{i,t-p}\phi_{ip} + \sum_{q=1}^{Q_i} \varepsilon_{i,t-q}\rho_{iq} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Panel heteroskedasticity

$$\Delta^{d} y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^{d} y_{i,t-p}\phi_{p} + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_{q} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, {\sigma_i}^2)$$

Still more time series options

Variable intercepts with a unit-specific trend

$$\Delta^d y_{it} = \alpha_i + t\theta_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Still more time series options

Variable intercepts with a unit-specific trend

$$\Delta^d y_{it} = \alpha_i + t\theta_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Variable intercepts with unit-specific additive seasonality

$$\Delta^d y_{it} = \alpha_i + \kappa_{k,i} + +\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

Note that you could also assume the same trend and/or seasonality applies to the whole panel by dropping the relevant i subscripts

Models of variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$$

How do we model α_i?

Let the mean of α_i be α_i^*.

Models of variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

Then there are a range of possibilities:

Let α_i be a random variable with no systemic component
(this type of α_i known as a *random effect*)

$$\alpha_i \sim \mathrm{N}\left(0, \sigma_\alpha^2\right)$$

Models of variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

Then there are a range of possibilities:

Let α_i be a random variable with no systemic component
(this type of α_i known as a *random effect*)

$$\alpha_i \sim \mathrm{N}\left(0, \sigma^2_\alpha\right)$$

Let α_i be a systematic component with no stochastic component
(this type of α_i is known as a *fixed effect*)

$$\alpha_i = \alpha_i^*$$

Models of variable intercepts

$$\Delta^d y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

Then there are a range of possibilities:

Let α_i be a random variable with no systemic component
(this type of α_i known as a *random effect*)

$$\alpha_i \sim \mathrm{N}\left(0, \sigma_\alpha^2\right)$$

Let α_i be a systematic component with no stochastic component
(this type of α_i is known as a *fixed effect*)

$$\alpha_i = \alpha_i^*$$

Let α_i be a random variable with a unit-specific systematic component
(this type of α_i known as a *mixed effect)*

$$\alpha_i \sim \mathrm{N}\left(\alpha_i^*, \sigma_\alpha^2\right)$$

Random effects

$$\alpha_i \sim \mathrm{N}\left(0, \sigma_\alpha^2\right)$$

Intuitive from a maximum likelihood modeling perspective

A unit specific error term

Assumes the units come from a common population,
with an unknown (estimated) variance, σ_α^2

In likelihood inference, estimation focuses on this variance, not on particular α_i's

Uncorrelated with $\mathbf{x_{it}}$ by design

Need MLE to estimate

Random effects example

A (contrived) example may help clarify what random effects are.

Suppose that we have data following this true model:

$$
\begin{aligned}
y_{it} &= \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it} \\
\alpha_i &\sim \mathcal{N}(0, \sigma_\alpha^2) \\
\varepsilon_{it} &\sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

with $i \in \{1, \ldots N\}$ and $t \in \{1, \ldots T\}$

Note that we are ignoring time series dynamics for now

It may help to pretend that these data have a real world meaning
though remember throughout we have created them out of thin air and `rnorm()`

So let's pretend these data reflect undergraduate student assignment scores over a term for $N = 100$ students and $T = 5$ assignments

Random effects example: Student aptitude & effort

Let's pretend these data reflect undergraduate student assignment scores

over a term for $N = 100$ students and $T = 5$ assignments:

$$\begin{aligned} \text{score}_{it} &= \beta_0 + \beta_1 \text{hours}_{it} + \alpha_i + \varepsilon_{it} \\ \alpha_i &\sim \mathcal{N}(0, \sigma_\alpha^2) \\ \varepsilon_{it} &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

with $i \in \{1, \ldots N\}$ and $t \in \{1, \ldots T\}$

The response is the assignment score, score_{it}

and the covariate is the hours studied, hours_{it}

and each student has an unobservable aptitude α_i which is Normally distributed

Aptitude has the same (random) effect on each assignment by a given student

Random effects example: Student aptitude & effort

Let's pretend these data reflect undergraduate student assignment scores

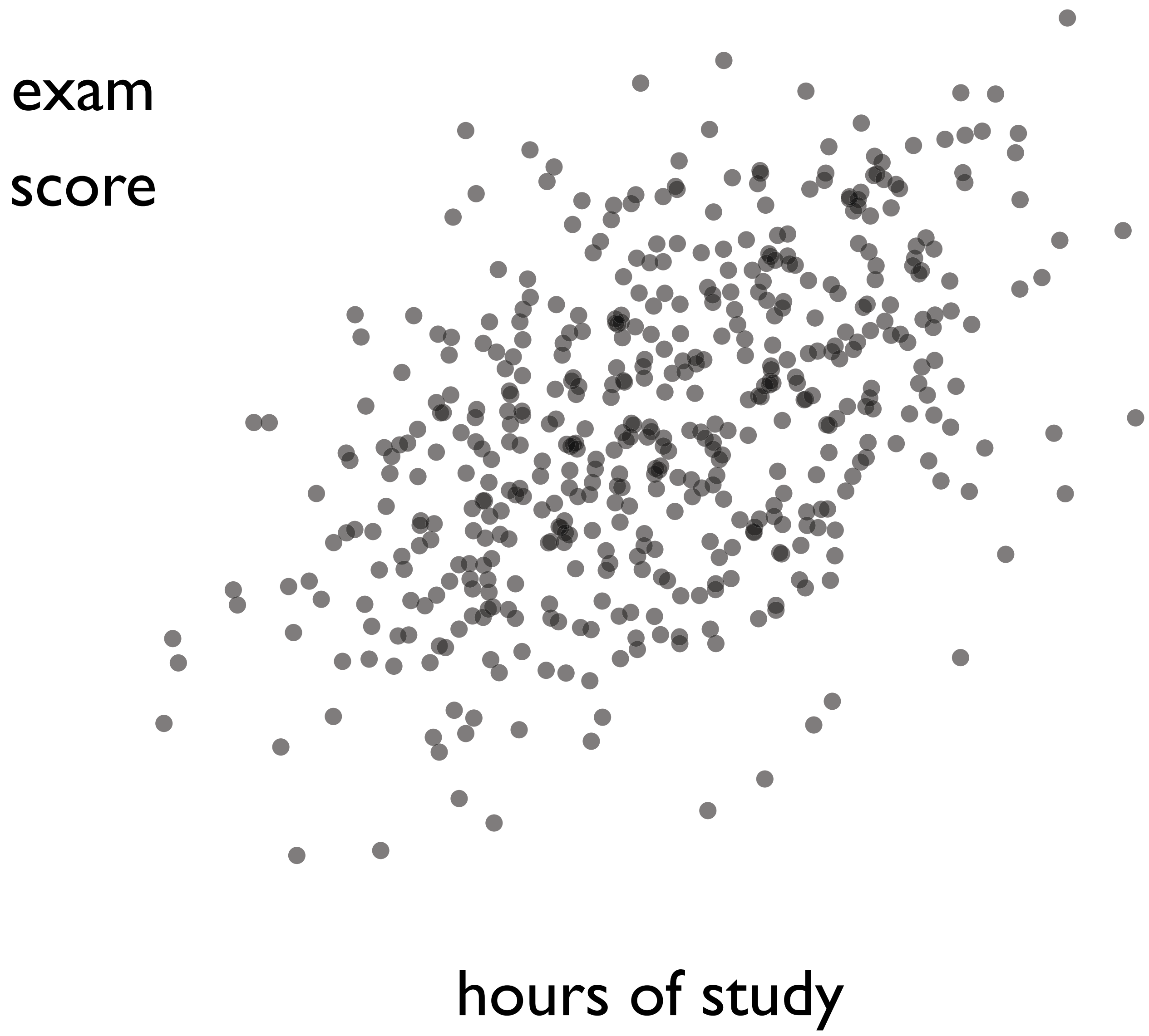
over a term for $N = 100$ students and $T = 5$ assignments:

$$
\begin{aligned}
\text{score}_{it} &= 0 + 0.75 \times \text{hours}_{it} + \alpha_i + \varepsilon_{it} \\
\alpha_i &\sim \mathcal{N}(0, 0.7^2) \\
\varepsilon_{it} &\sim \mathcal{N}(0, 0.2^2)
\end{aligned}
$$

with $i \in \{1, \ldots 100\}$ and $t \in \{1, \ldots 5\}$

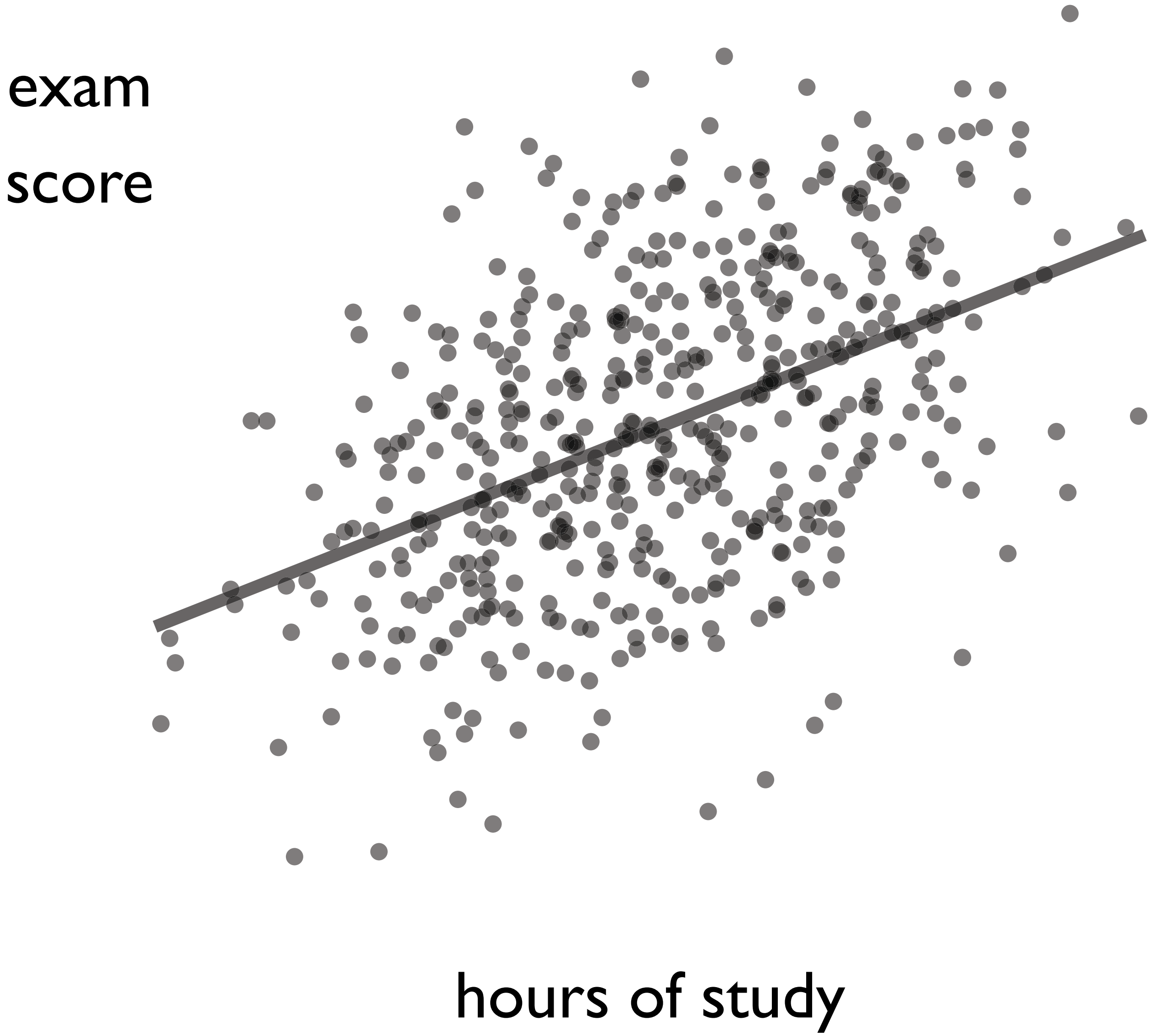
the above are the true values of the parameters I used to generate the data

let's see what role the random effect α_i plays here



The 500 obervations

A relationship between effort & scores seems evident

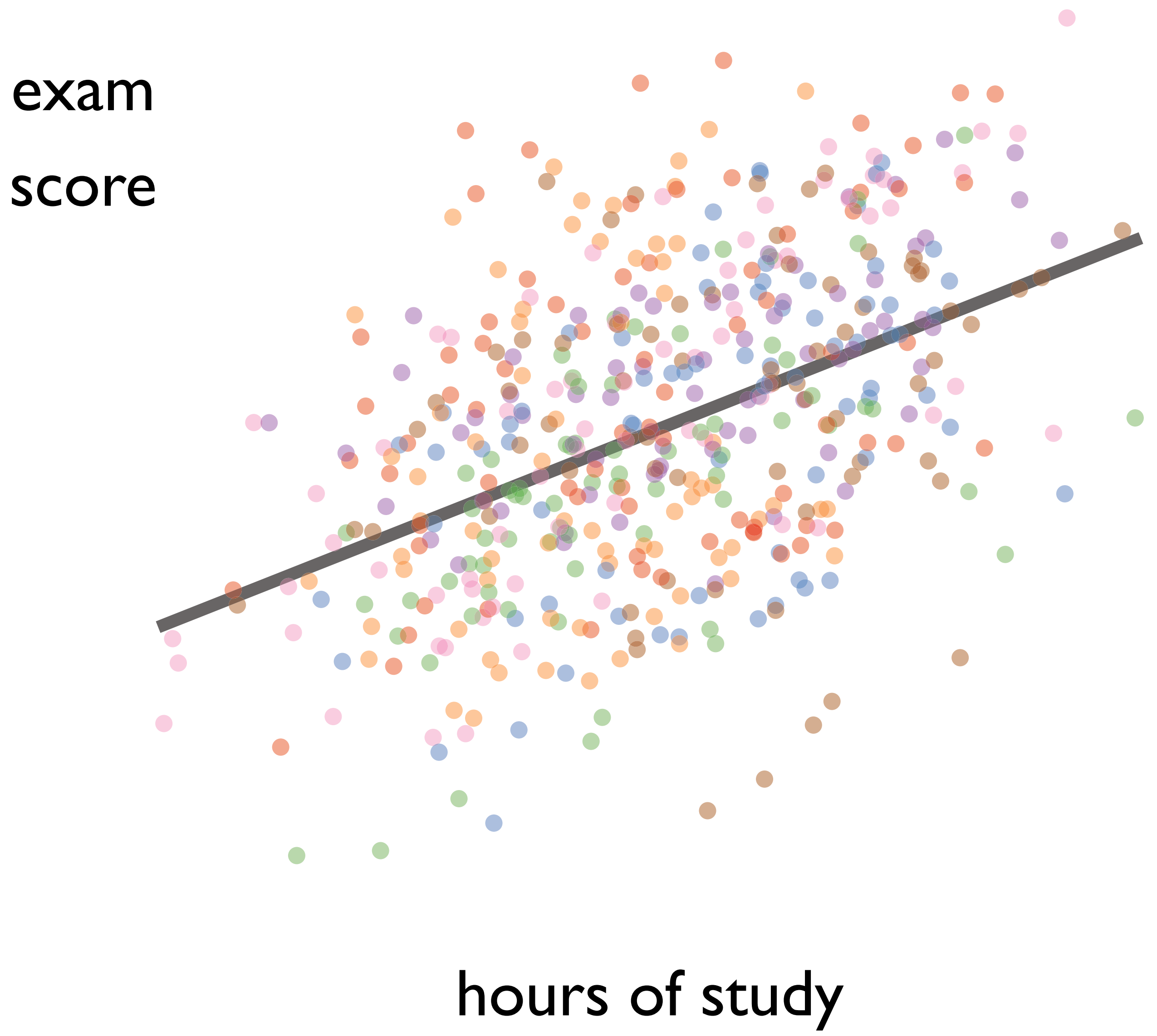


Let's summarize the relationship using the least squares $\hat{\beta}_1$

Approximately equal to the true $\beta_1 = 0.75$

Haven't discussed, used, or estimated the random effects yet

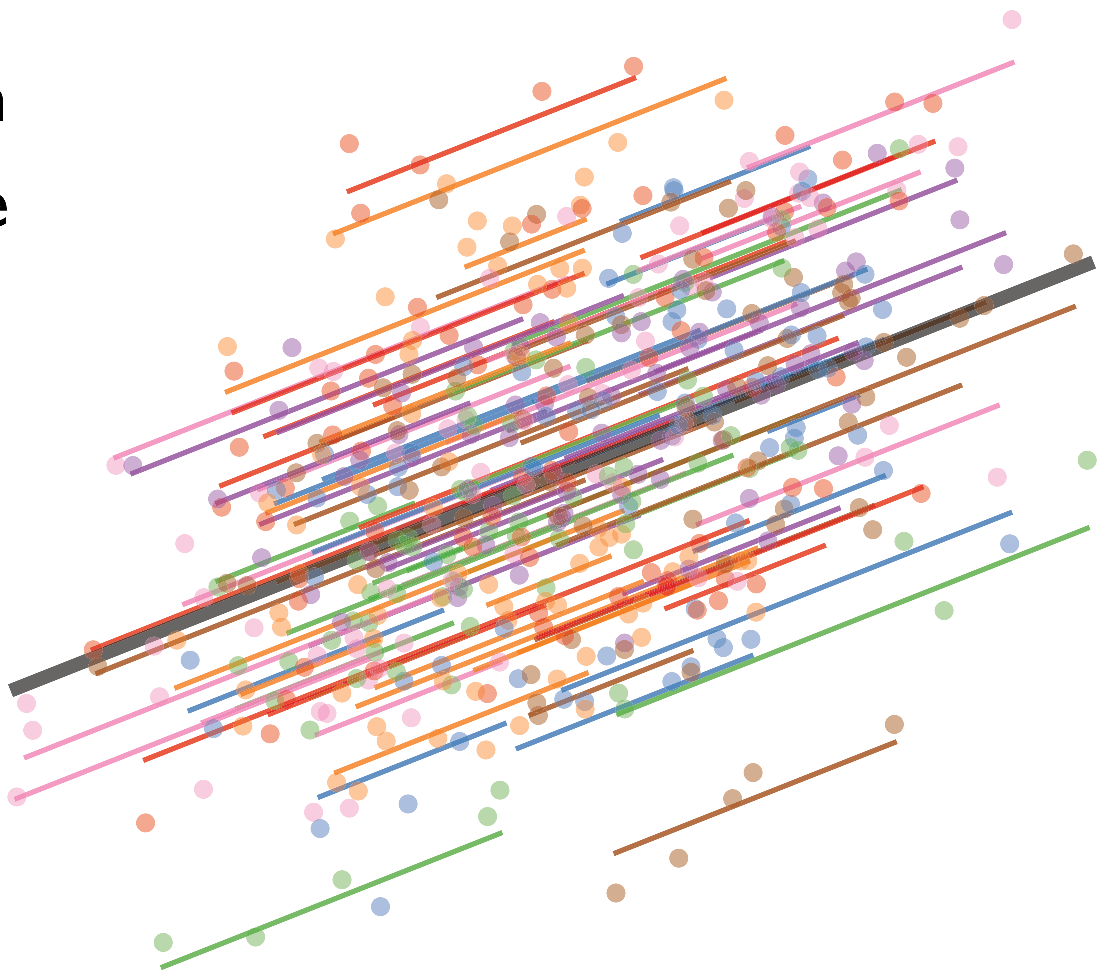
Do we need them?



Identify each of the 100 students using colored dots (we have 8 colors; they repeat)

Clear that each student's scores are tightly clustered

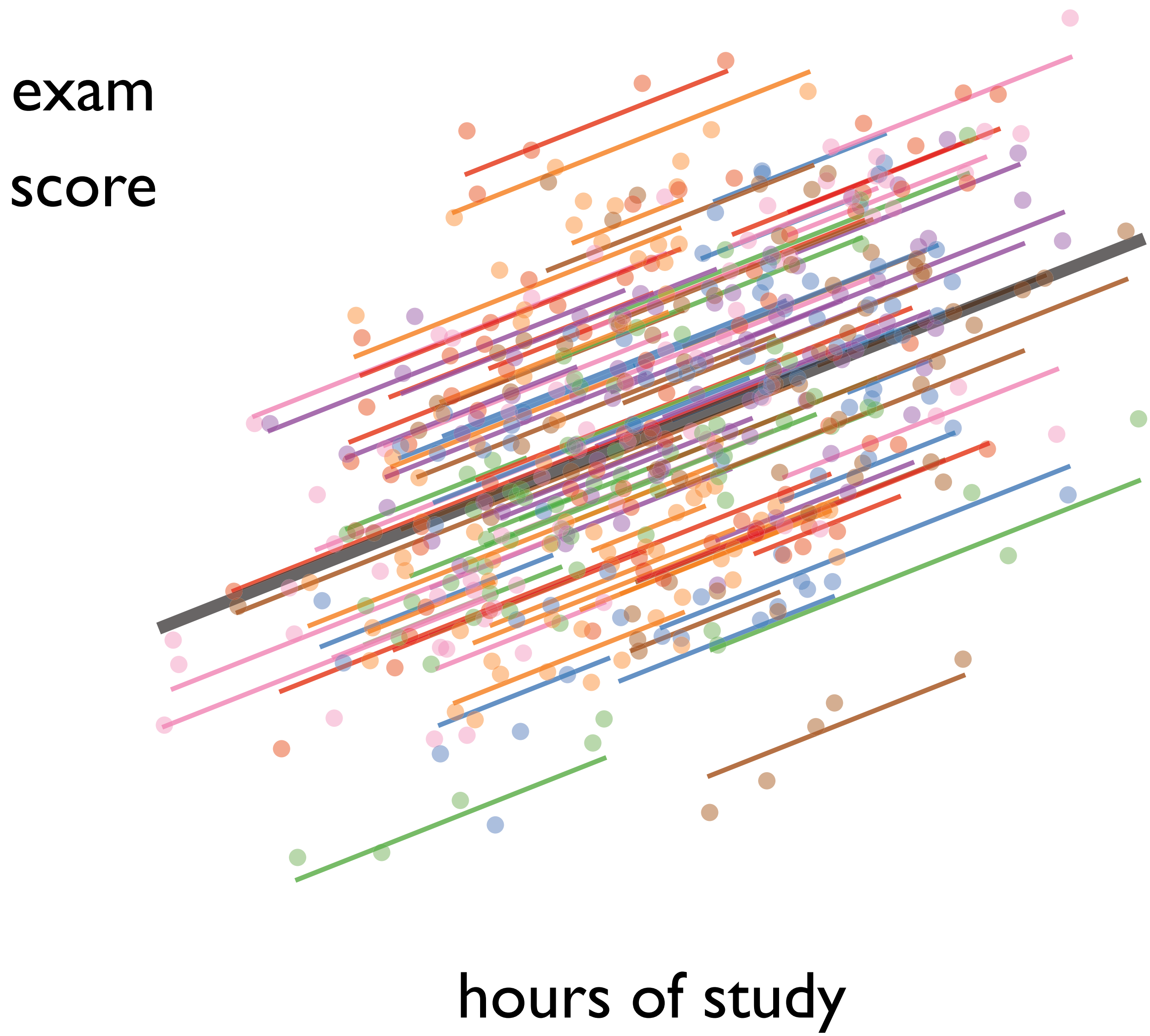
Note the student-level slopes



Each student follows the same regression line as the whole class, but with a unique intercept

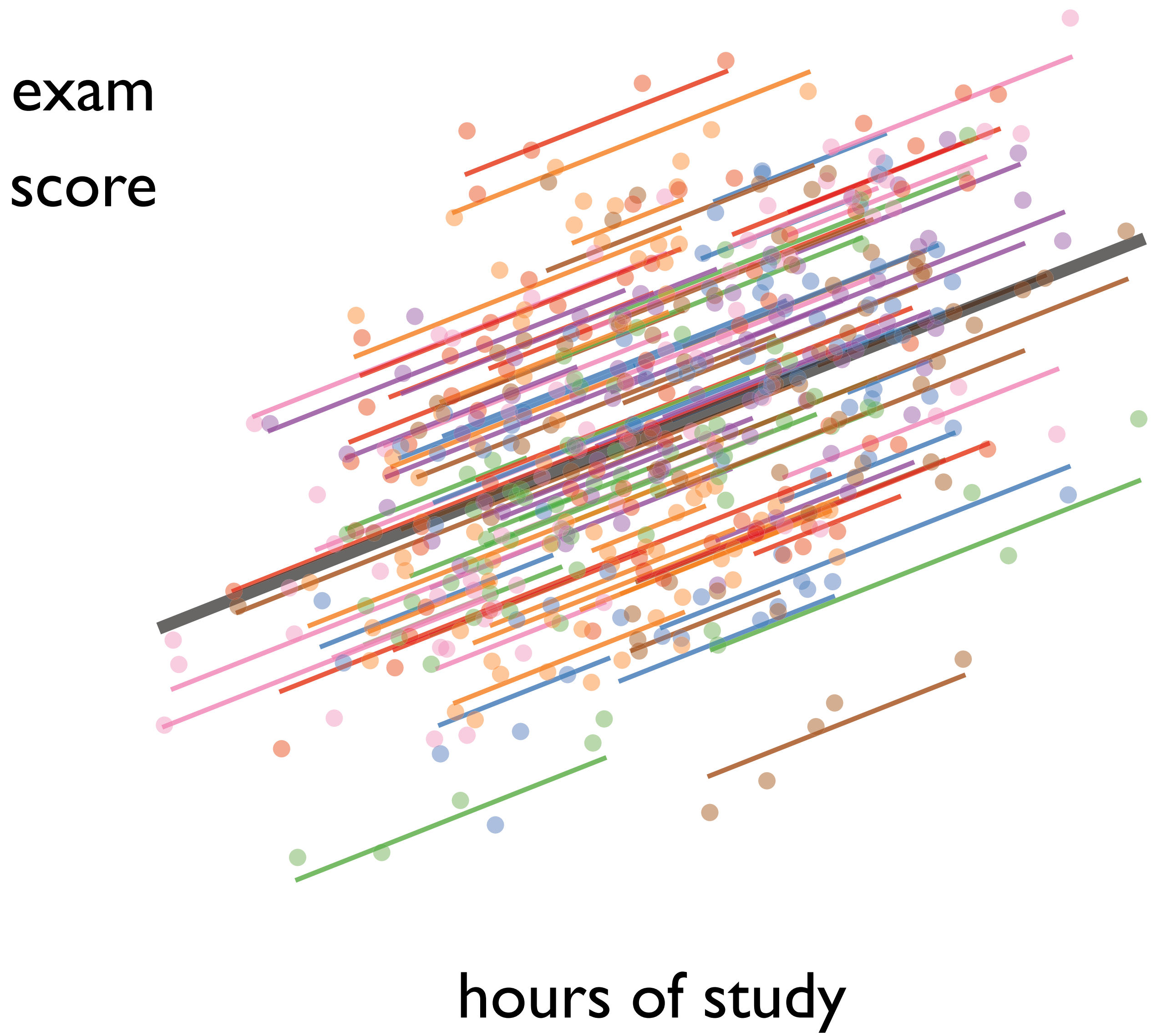
That intercept is the random effect α_i

It's also the average difference between student i's scores and the class-level regression line



The student random effect is the student-specific component of the error term

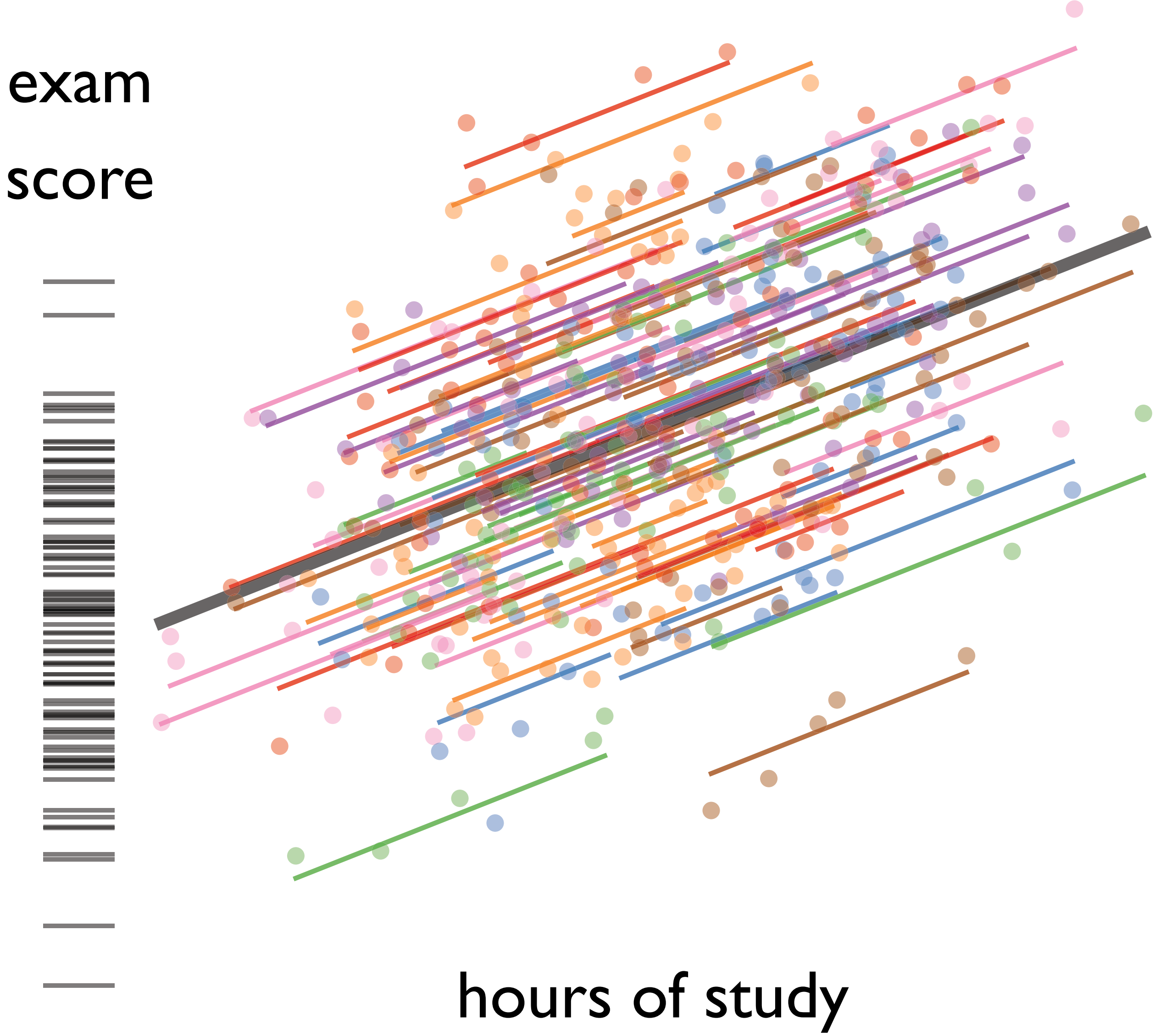
After we remove it, a student's scores across exams exhibit white noise variation around a student-specific version of the overall regression line



These random effects α_i reflect the portion of the error term that results from unmeasured student characteristics

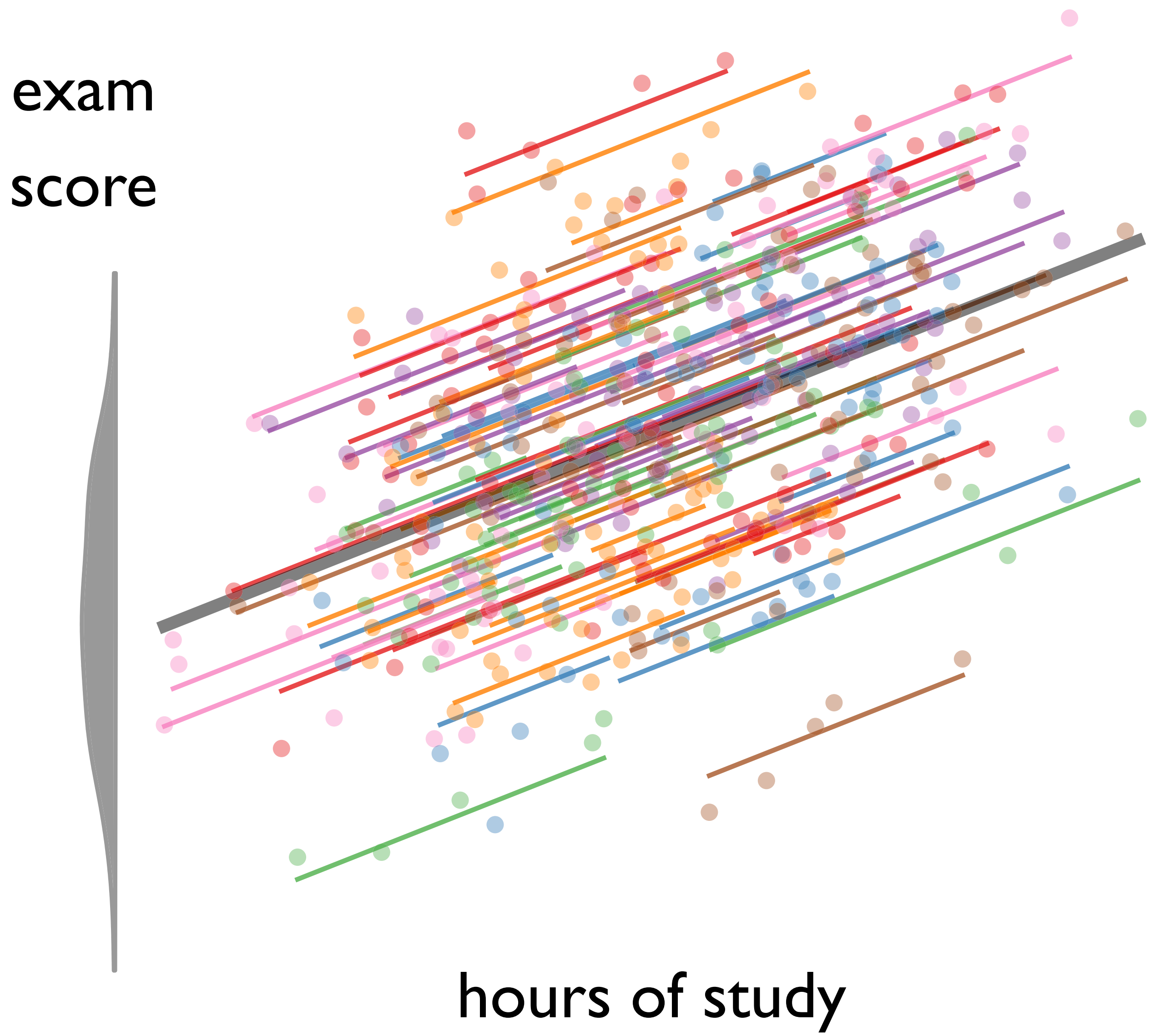
I've labelled this random component "aptitude"

But that's is just a word for everything related to a student's ability

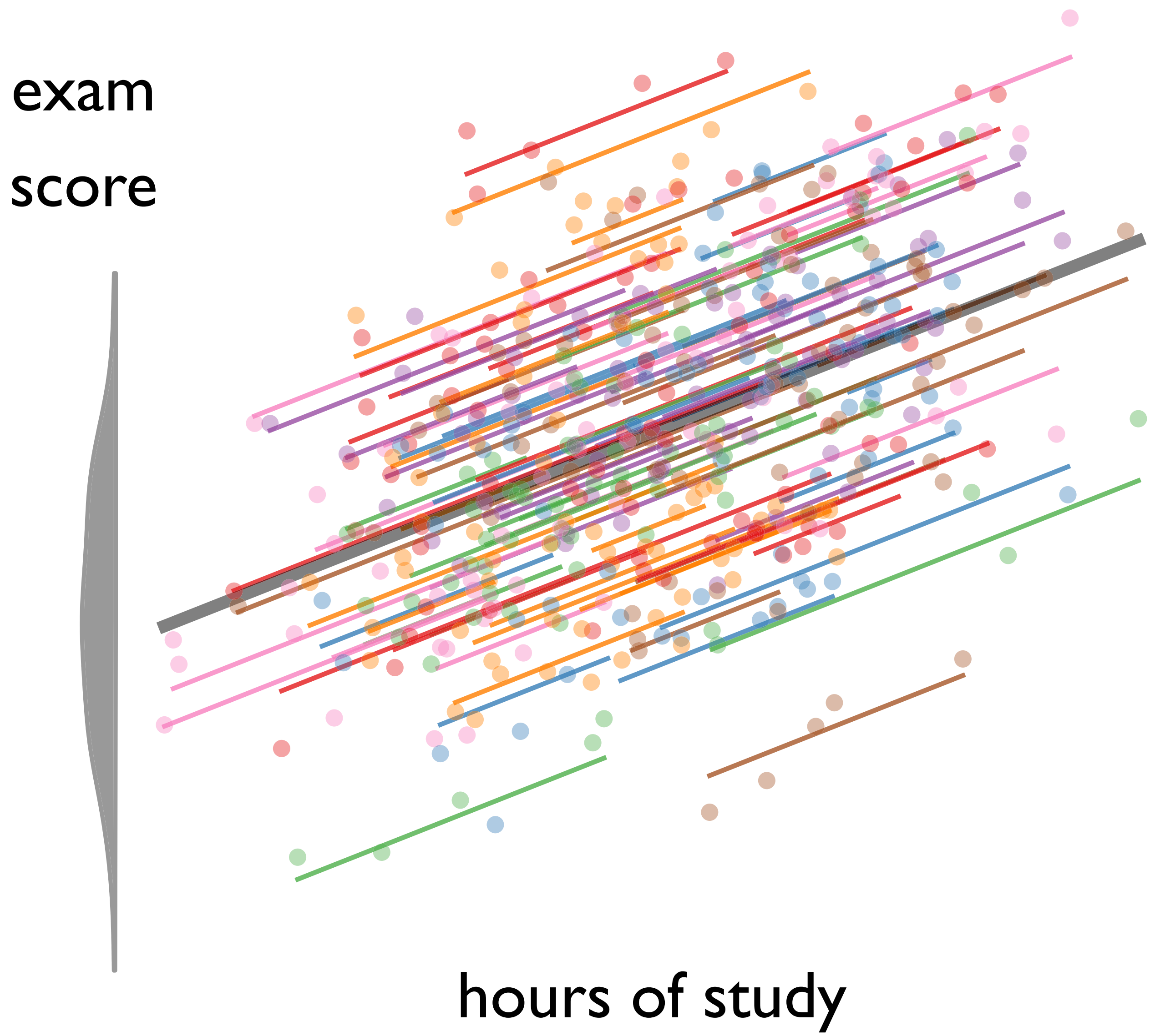


The distribution of the random effects is shown at the left

A plot of a marginal distribution on the side of a scatterplot is called a "rug"



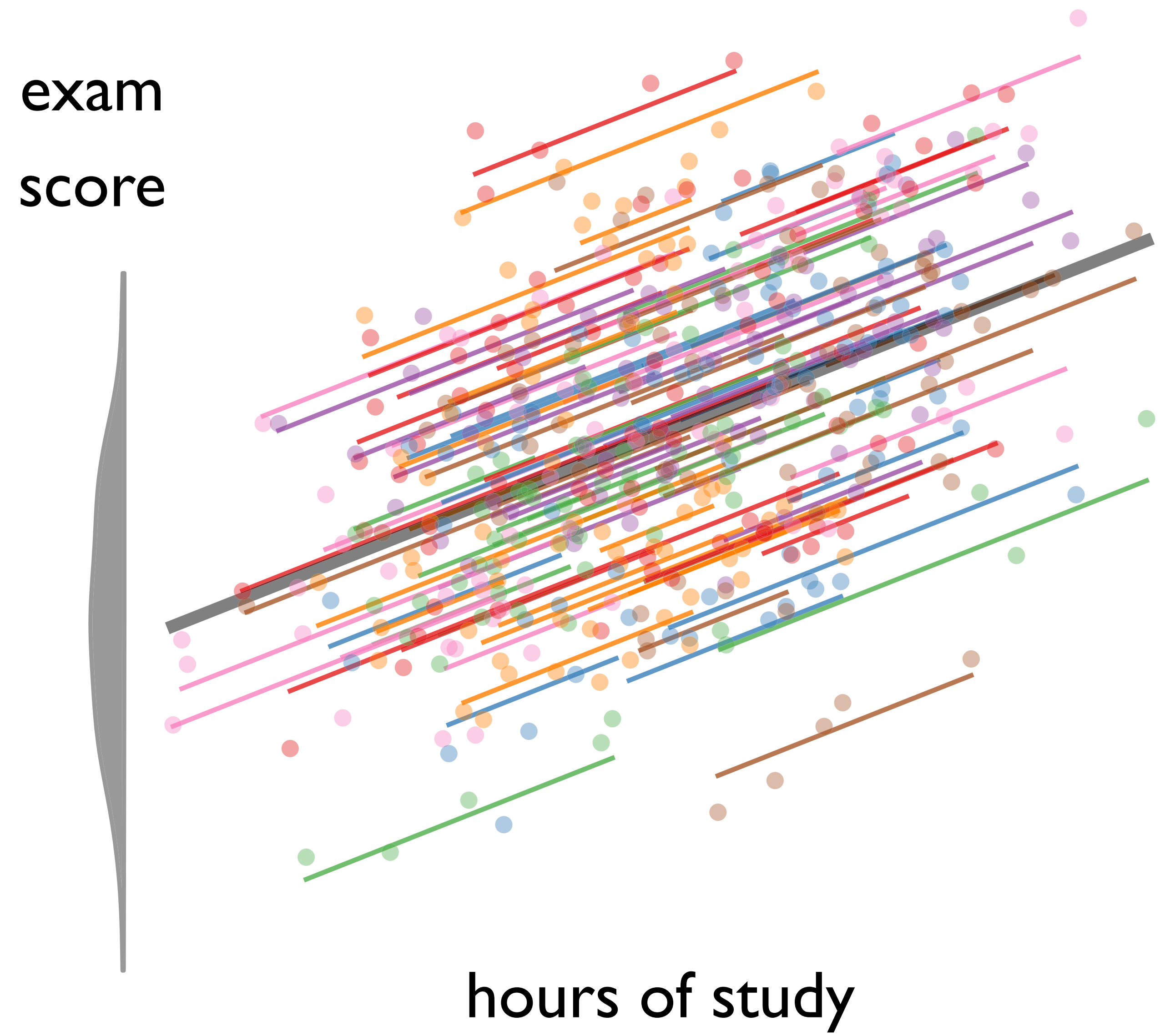
A density plot of the distribution of random effects suggests they are approximately Normal



Random effects are a decomposition of the error term into

1. a unit-specific part
2. an idiosyncratic part

Random effects are determined after we have the overall regression slope and cannot change that slope



The model is now hierarchical or multilevel

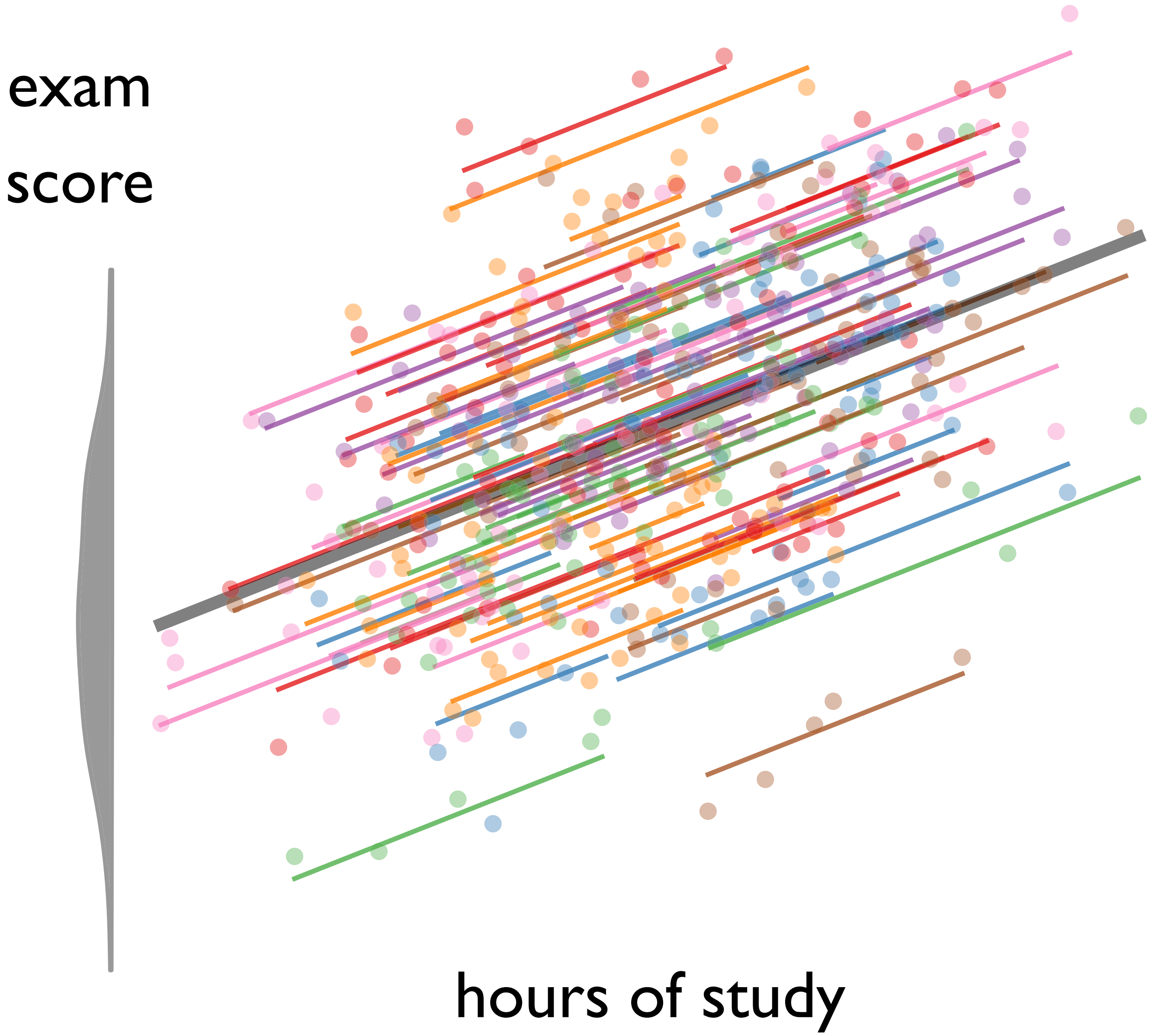
Level 1: Student level

sits above

Level 2: Student × exam level

There's random variation at both levels

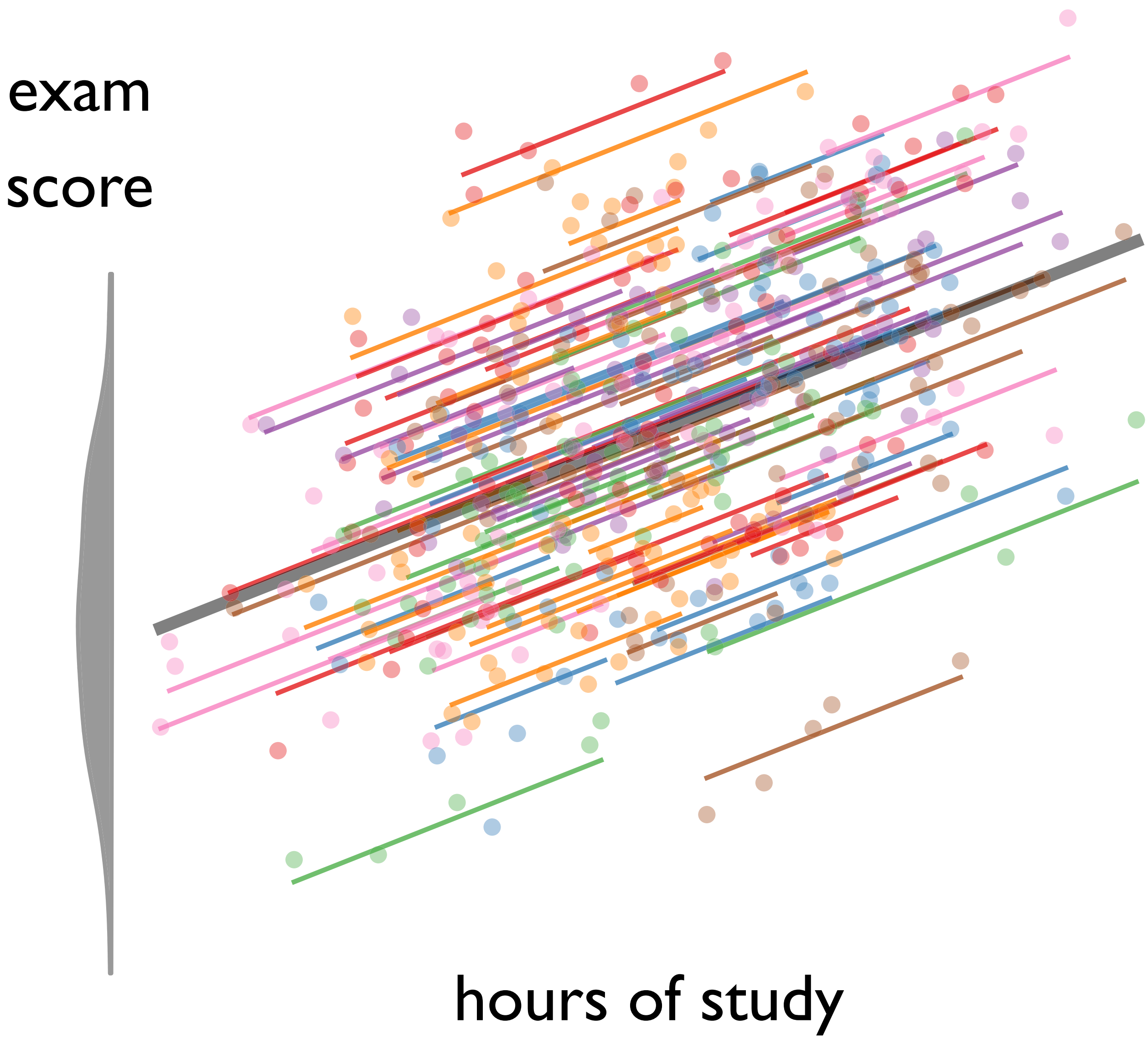
But mainly at the student level



Students randomly vary a lot: $\sigma_\alpha = 0.7$

Exams for a given student vary little: $\sigma_\varepsilon = 0.2$

Student level random effects comprise $100\% \times \sqrt{0.7^2/(0.7^2 + 0.2^2)} = 96\%$ of the total error variance



We haven't controlled for any omitted confounders

What if unmeasured ability were correlated with study effort?

Our $\hat{\beta}_1$ estimate would be biased

This bias persists even if we allow for random effects

Random effects example: Student aptitude & effort

Suppose that ability *is* correlated with effort

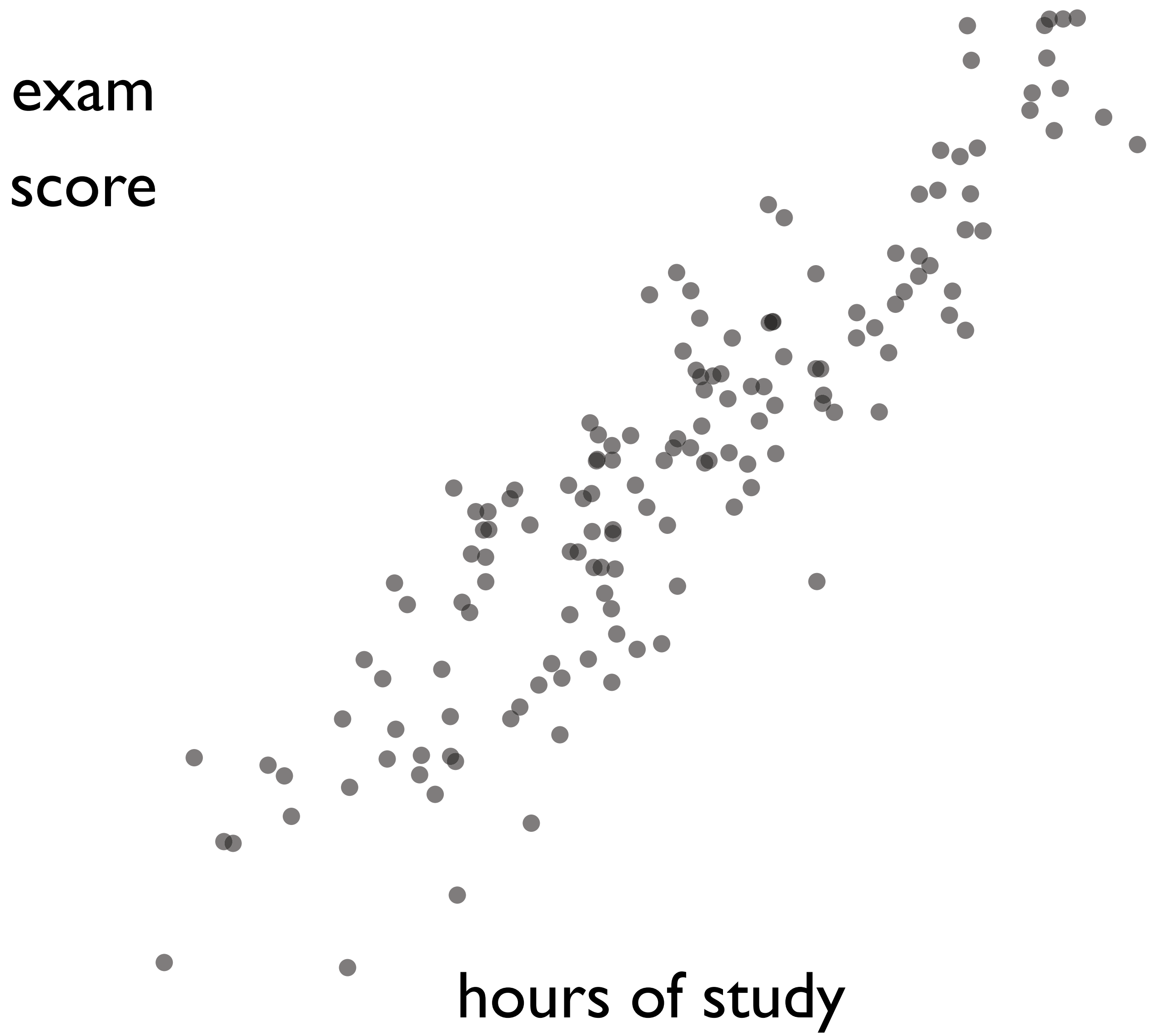
For example, perhaps high ability students rationally choose to study harder as their best available human capital investment opportunity

We have the same model, but now hours_{it} is a function of α_i:

$$\begin{aligned}
\text{score}_{it} &= 0 + 0.75 \times \text{hours}_{it} + \alpha_i + \varepsilon_{it} \\
\text{hours}_{it} &= 0 + 0.5 \times \alpha_i + \text{uniform}(-0.7, 0.7) \\
\alpha_i &\sim \mathcal{N}(0, 0.7^2) \\
\varepsilon_{it} &\sim \mathcal{N}(0, 0.2^2)
\end{aligned}$$

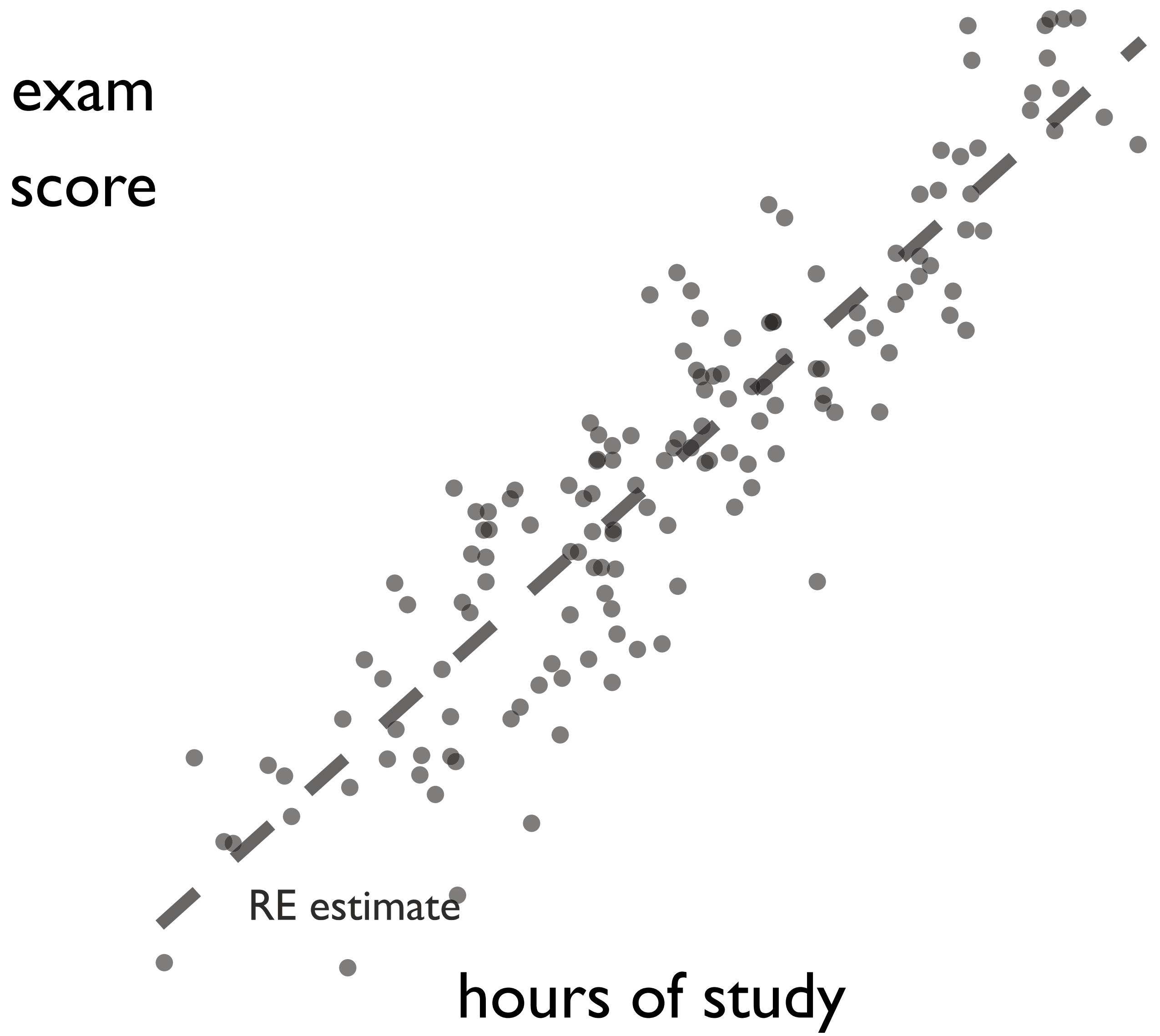
with $i \in \{1, \ldots 100\}$ and $t \in \{1, \ldots 5\}$

What happens when we estimate a treat α_i as a random effect and estimate $\hat{\beta}_1$?



I've shown only the first 30 students to make the graph easier to read

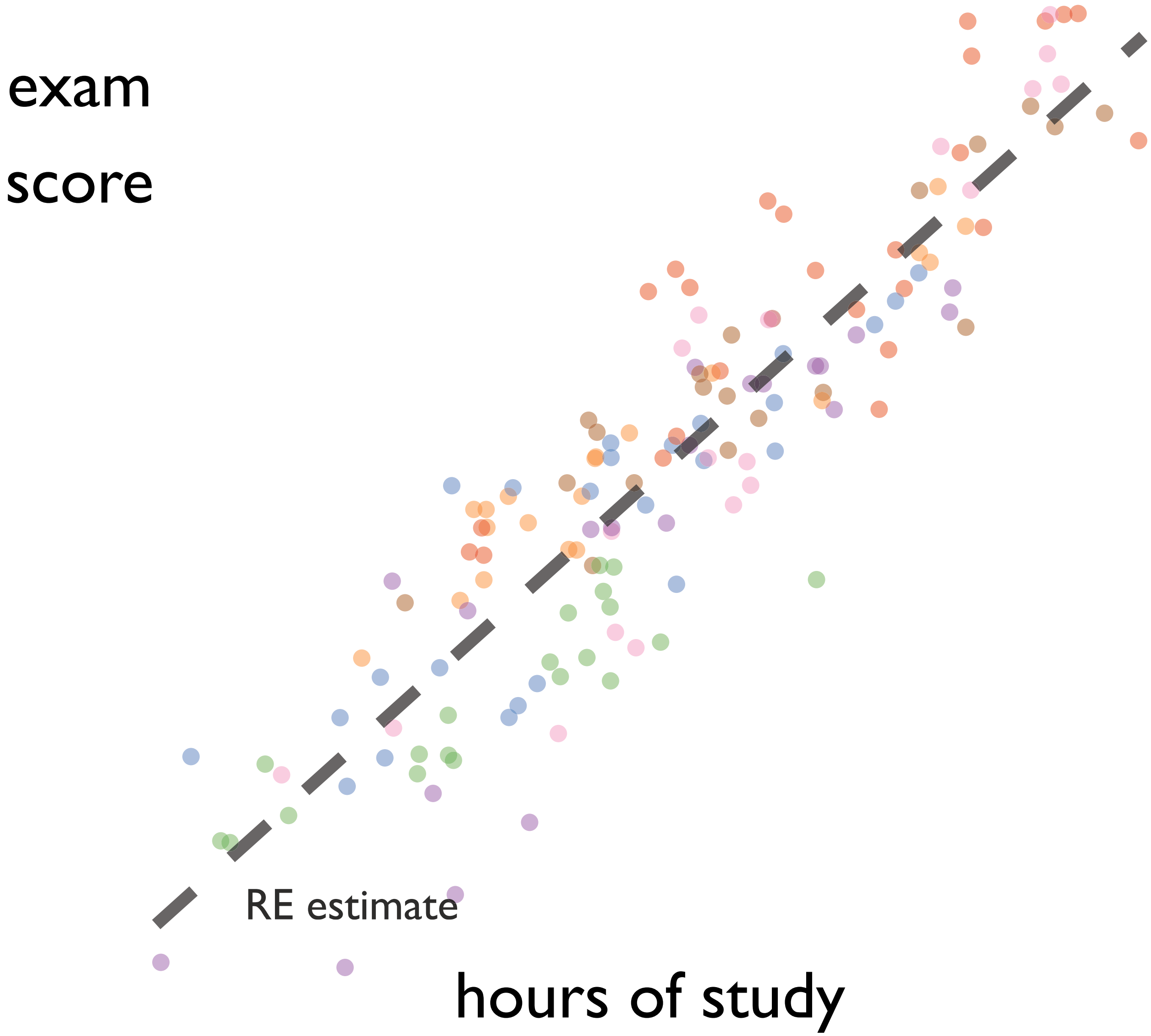
A *stronger* relationship between effort and grades seems evident



Least squares model finds $\hat{\beta}_1 \approx 1.6$

More than double the true value of 0.75!

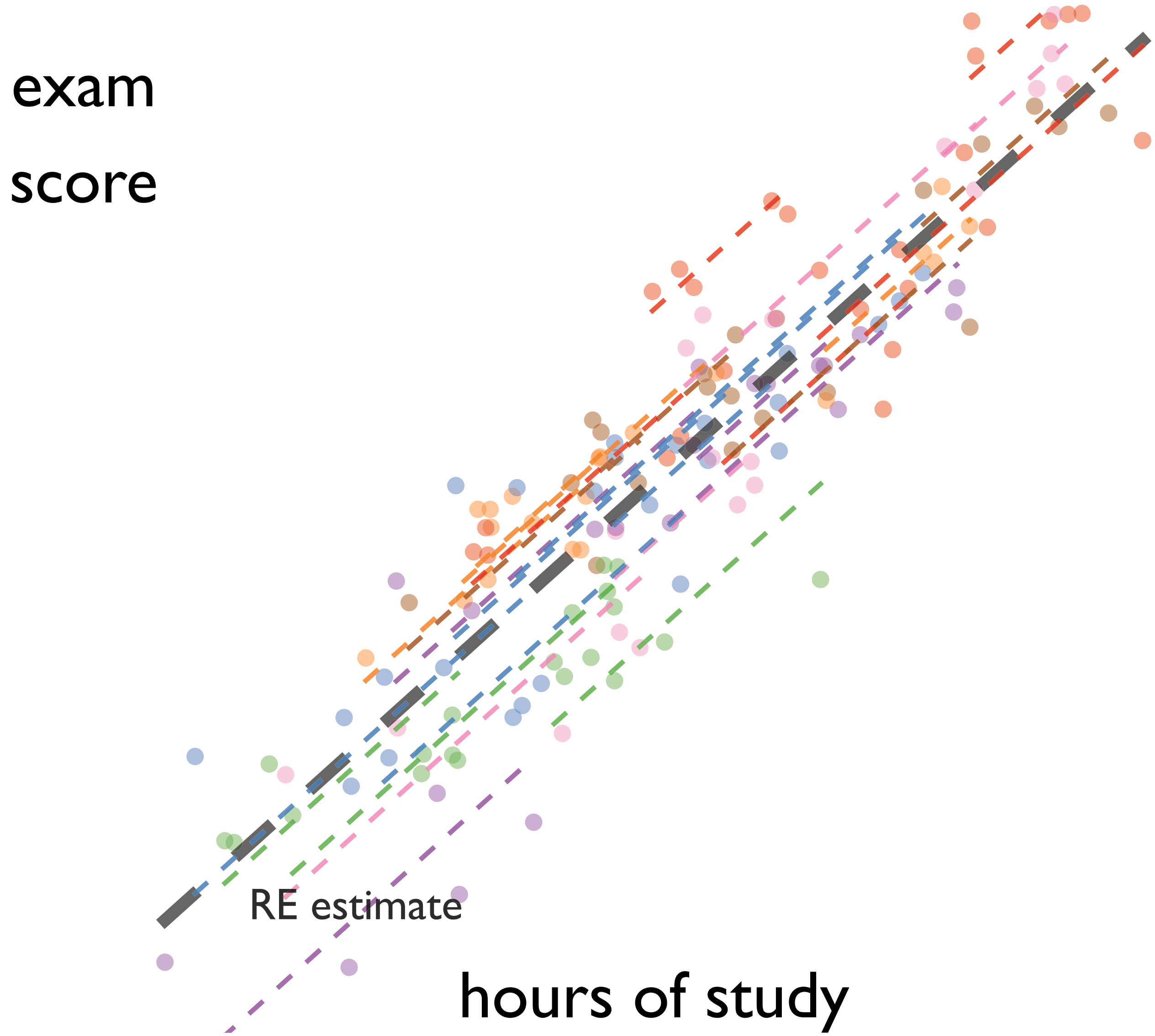
Where did the bias come from?



With multilevel data, it helps to start at the lowest level

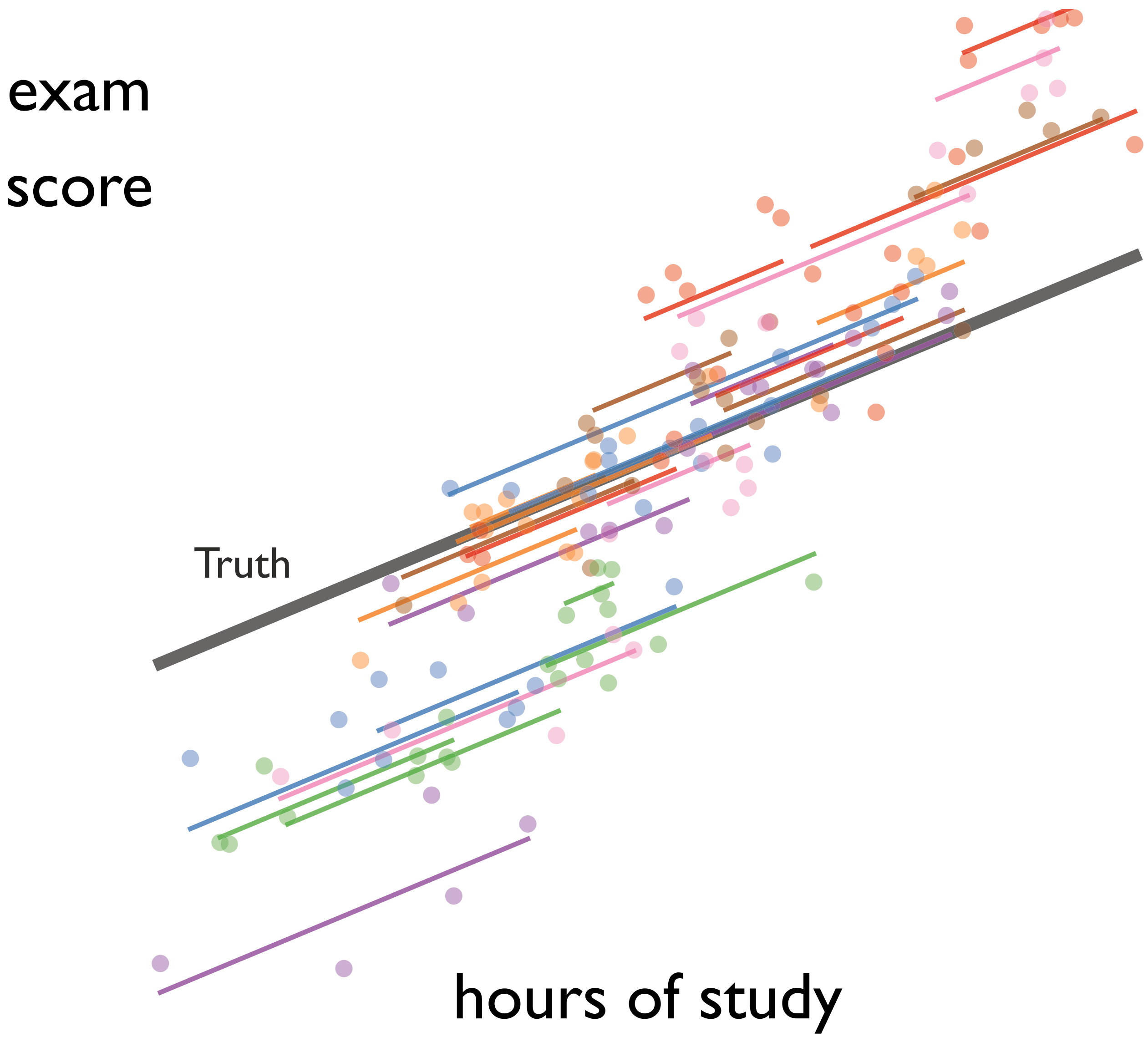
I've colored the points by student

A random effects model finds the student specific intercept *after* estimating the slope of the main regression line



The student specific relationships between effort and scores as estimated by a random effects model

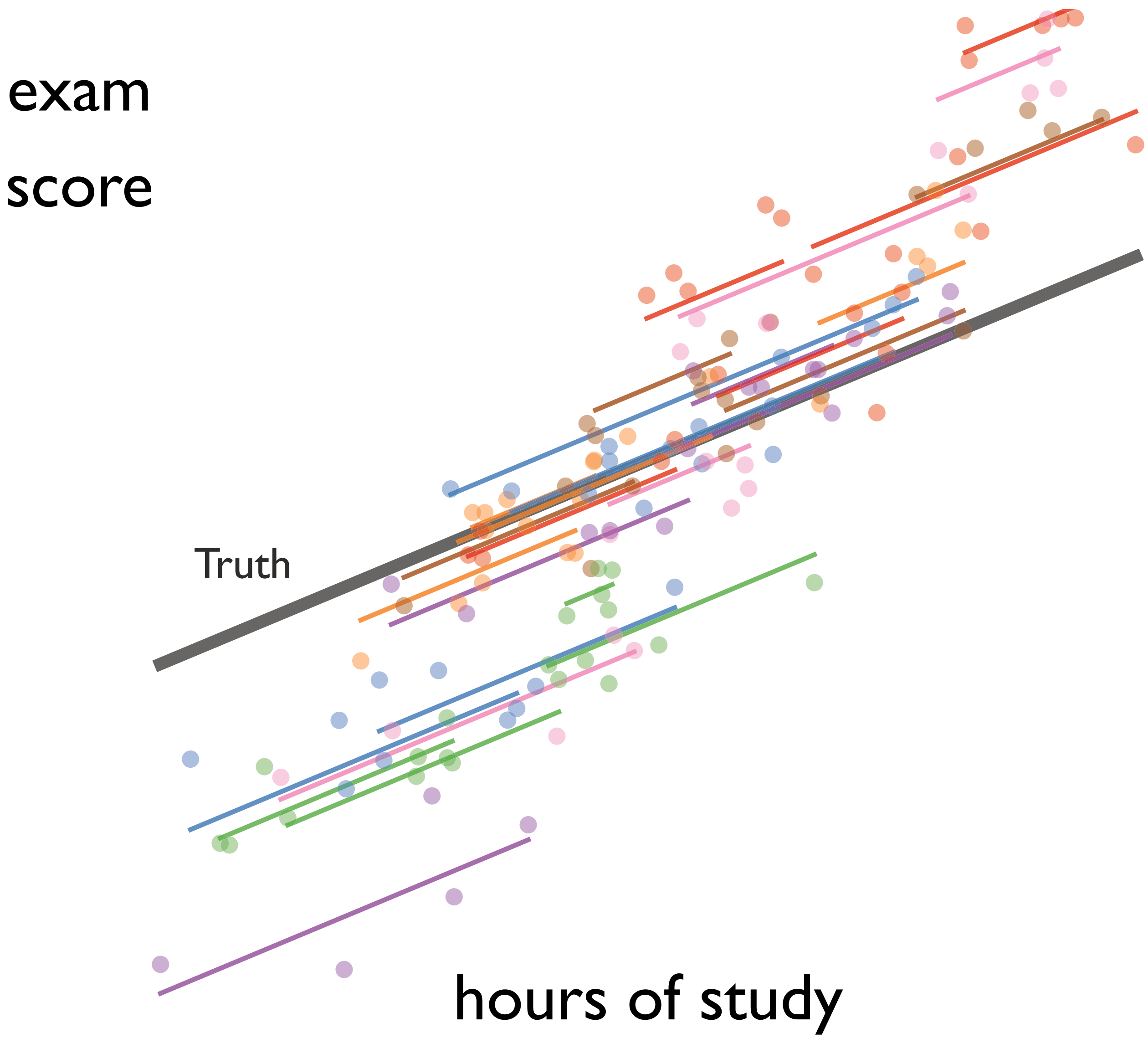
Are these estimates right?



Not even close

The *true* regression lines by student and overall

Random effects estimates of effort is *biased* because the student-specific effect is correlated with effort



Random effects are an inadequate model when the grouping indicator is correlated with our covariates

In this case we have *omitted variable bias*

We need a different model of α:
fixed effects

Fixed effects

$$\alpha_i = \alpha_i^*$$

Easiest to conceptualize in a linear regression framework

Easiest to estimate: just add dummies for each unit, and drop the intercept

Can be correlated with $\mathbf{x}_{it}$: FEs control for *all* omitted time-invariant variables

Indeed, that's usually the point.
FEs usually included to capture unobserved variance potentially correlated with $\mathbf{x}_{it}$.

Comes at a large cost:
we're actually purging the cross-sectional variation from the analysis

Then assuming a change in $\mathbf{x}$ would yield the same response in each time series

Fixed effects models use over-time variation in covariates to estimate parameters;
Cannot be added to models with perfectly time invariant covariates

More on fixed effects

$$\alpha_i = \alpha_i^*$$

Fixed effects specifications incur an incidental parameters problem: MLE is consistent as $T \to \infty$, but *not* as $N \to \infty$.

More on fixed effects

$$\alpha_i = \alpha_i^*$$

Fixed effects specifications incur an incidental parameters problem:
MLE is consistent as $T \rightarrow \infty$, but *not* as $N \rightarrow \infty$.

Of concern in microeconomics, where panels are sampled on N with T fixed.
Not of concern in CPE/IPE, where N is fixed, and T could expand

Monte Carlo experiments indicate small sample properties of fixed effects pretty good if $t > 15$ or so; we'll see some of these results later

Fixed effects are common in studies where N is not a random sample, but a (small) universe (e.g., the industrialized countries).

More on fixed effects

$$\alpha_i = \alpha_i^*$$

Fixed effects specifications incur an incidental parameters problem:
MLE is consistent as $T \to \infty$, but *not* as $N \to \infty$.

Of concern in microeconomics, where panels are sampled on N with T fixed.
Not of concern in CPE/IPE, where N is fixed, and T could expand

Monte Carlo experiments indicate small sample properties of fixed effects pretty good if $t > 15$ or so; we'll see some of these results later

Fixed effects are common in studies where N is not a random sample, but a (small) universe (e.g., the industrialized countries).

Sui generis: Fixed effects basically say "France is different because it's France," "America is different because it's America," etc.

Fixed effects example

Another example may help clarify what fixed effects are.

Suppose that we have data following this true model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_i + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

with $i \in \{1, \ldots N\}$ and $j \in \{1, \ldots M_i\}$

j indexes a set of M_i counties drawn from state i

There are $N = 15$ states total, and we drew $M_j = M = 15$ counties from each state

Note that we are ignoring time series dynamics completely now

(We could add them back in if j were ordered in time)

Fixed effects example

Suppose the data represent county level voting patterns for the US

(I.e., let's illustrate Gelman *et al*, *Red State, Blue State, Rich State, Poor State* w/ contrived data)

$$\begin{aligned}\mathrm{RVS}_{ij} &= \beta_0 + \beta_1 \mathrm{Income}_{ij} + \beta_2 \mathrm{ConservativeCulture}_i + \varepsilon_{ij} \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2)\end{aligned}$$

with $i \in \{1, \ldots N\}$ and $j \in \{1, \ldots M_i\}$

j indexes a set of M_i counties drawn from state i

Remember: the data I'm using are fake, and contrived to illustrate a concept simply

Gelman et al investigate this in detail with real data and get similar but more nuanced findings

We will review the real data later in this lecture

Fixed effects example: What's the matter with Kansas?

Suppose the data represent county level voting patterns for the US

(I.e., let's illustrate Gelman *et al*, *Red State, Blue State, Rich State, Poor State* using similar but contrived data)

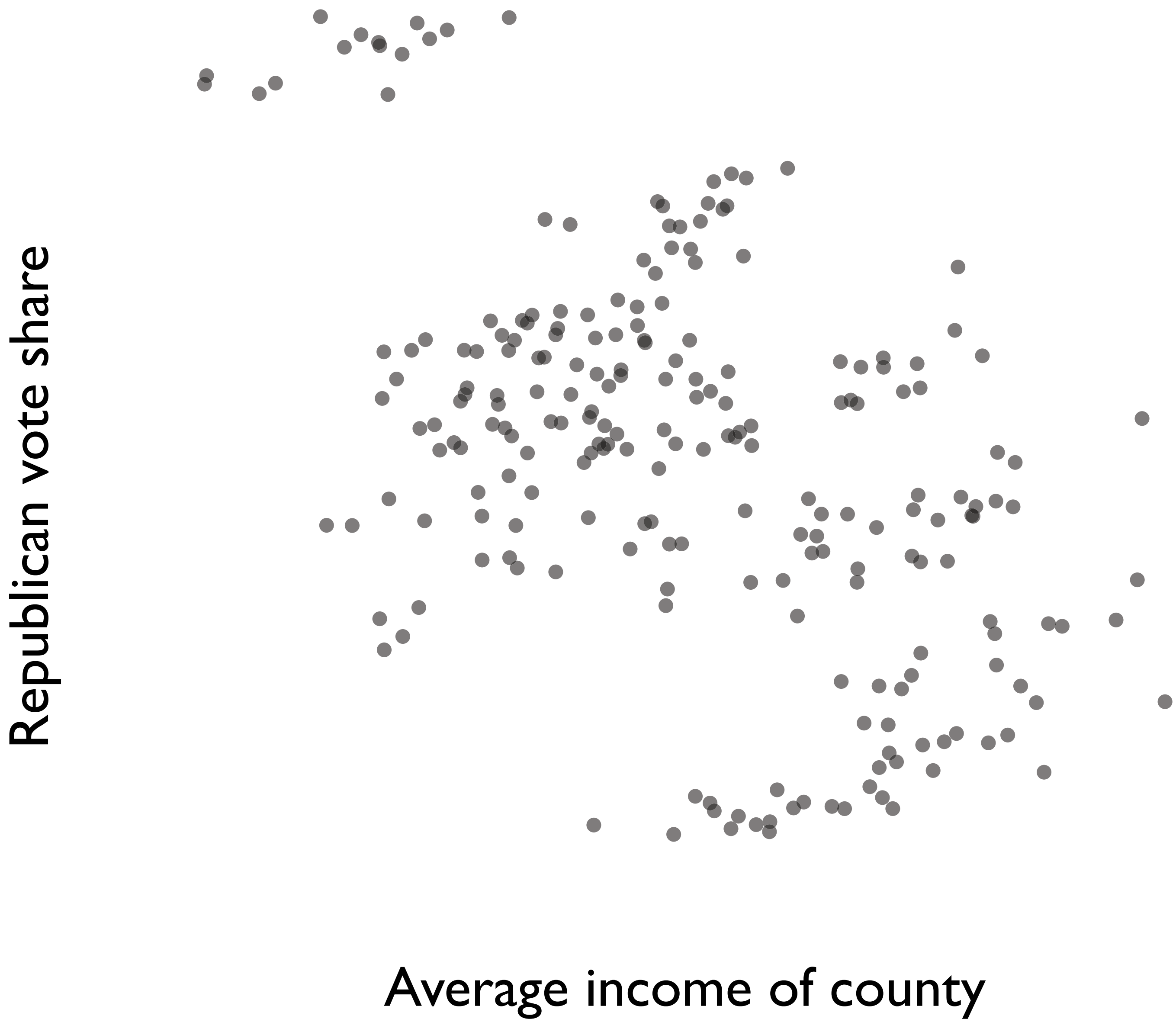
$$\text{RVS}_{ij} = \beta_0 + \beta_1 \text{Income}_{ij} + \beta_2 \text{Conservatism}_i + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

with $i \in \{1, \ldots N\}$ and $j \in \{1, \ldots M_i\}$

A problem:
suppose we don't have (or don't trust) a measure of state-level Conservatism

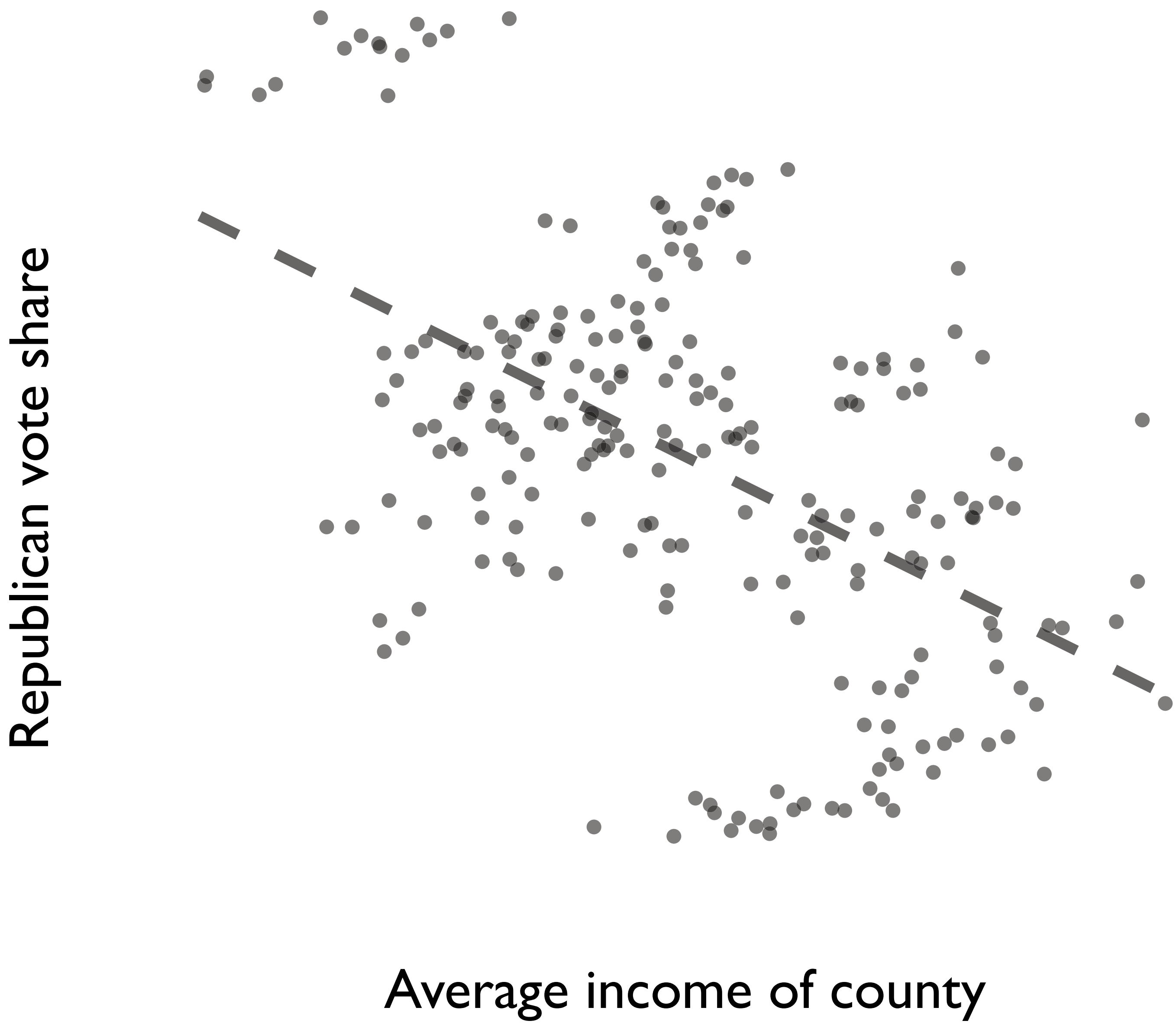
If we exclude it, or mismeasure it, we could get omitted variable bias in $\hat{\beta}_1$

This leads to potentially large misconceptions. . .



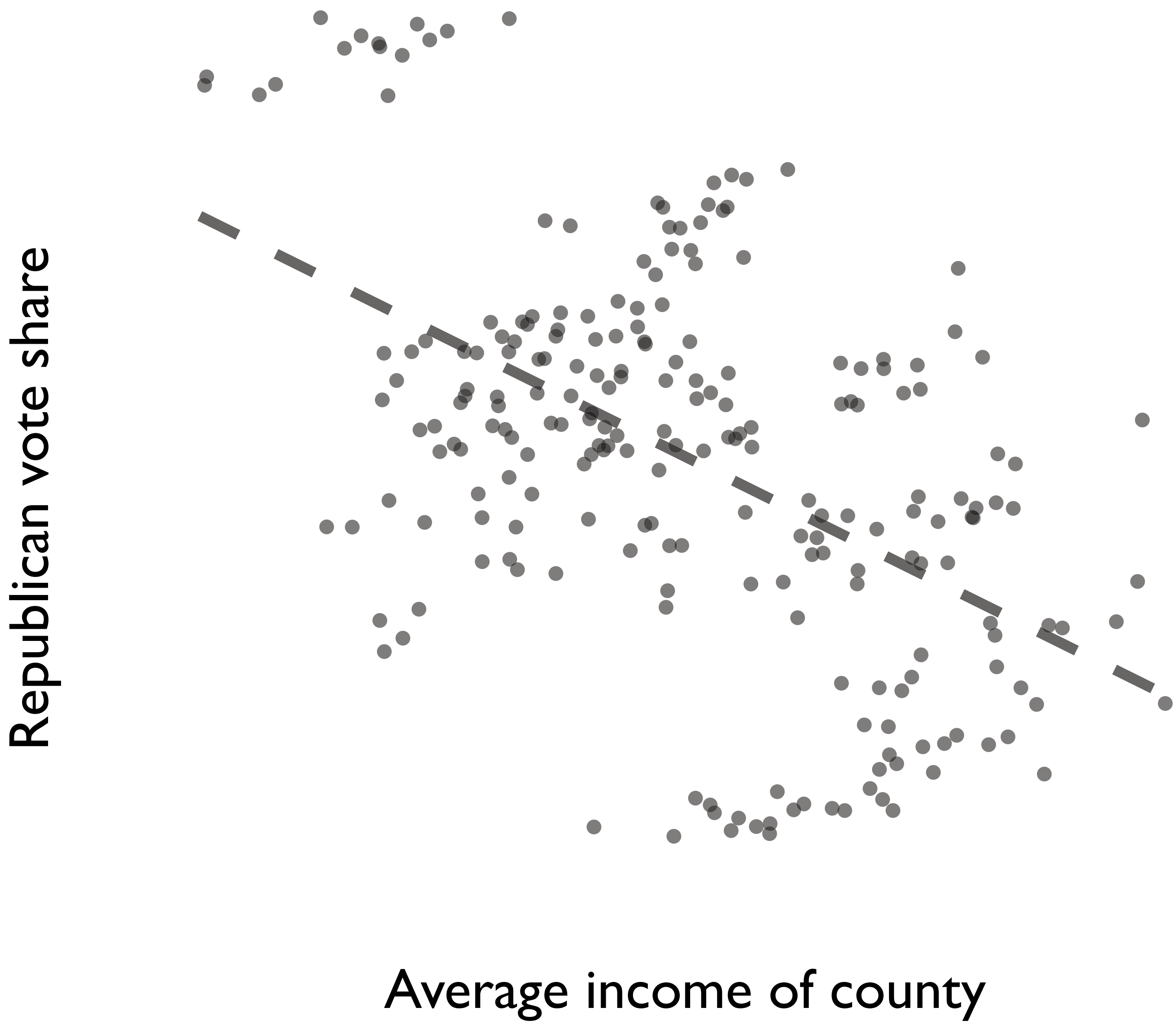
Suppose we observe 15 counties from each of 15 states (225 observations)

Our first cut is to estimate this simple linear regression: $y_{ij} = \beta_0 + \beta_1 \text{Income}_{ij} + \varepsilon_{ij}$



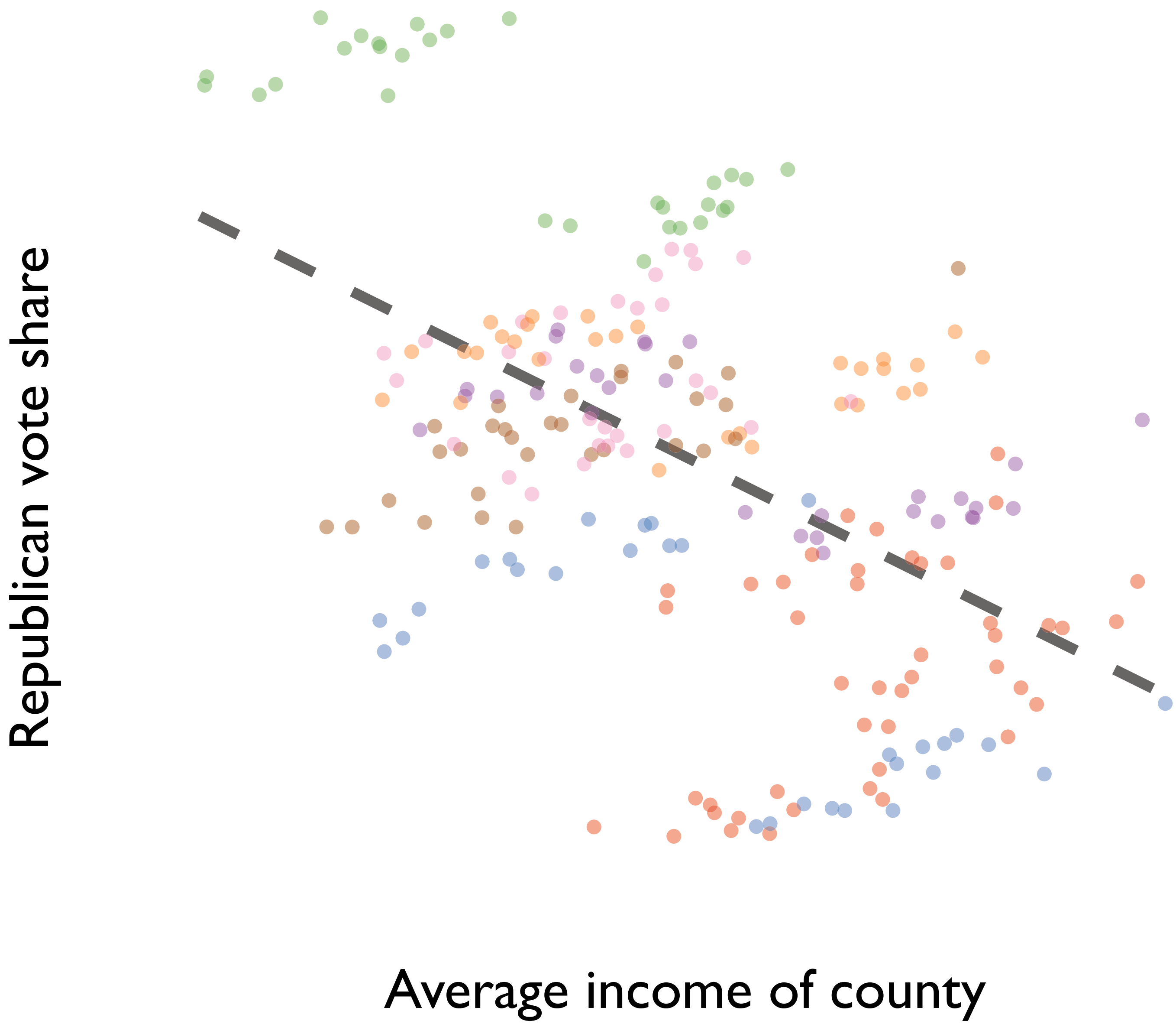
We find that $\hat{\beta}_1$ is negative:

Poor counties seem to vote more Republican than rich counties!



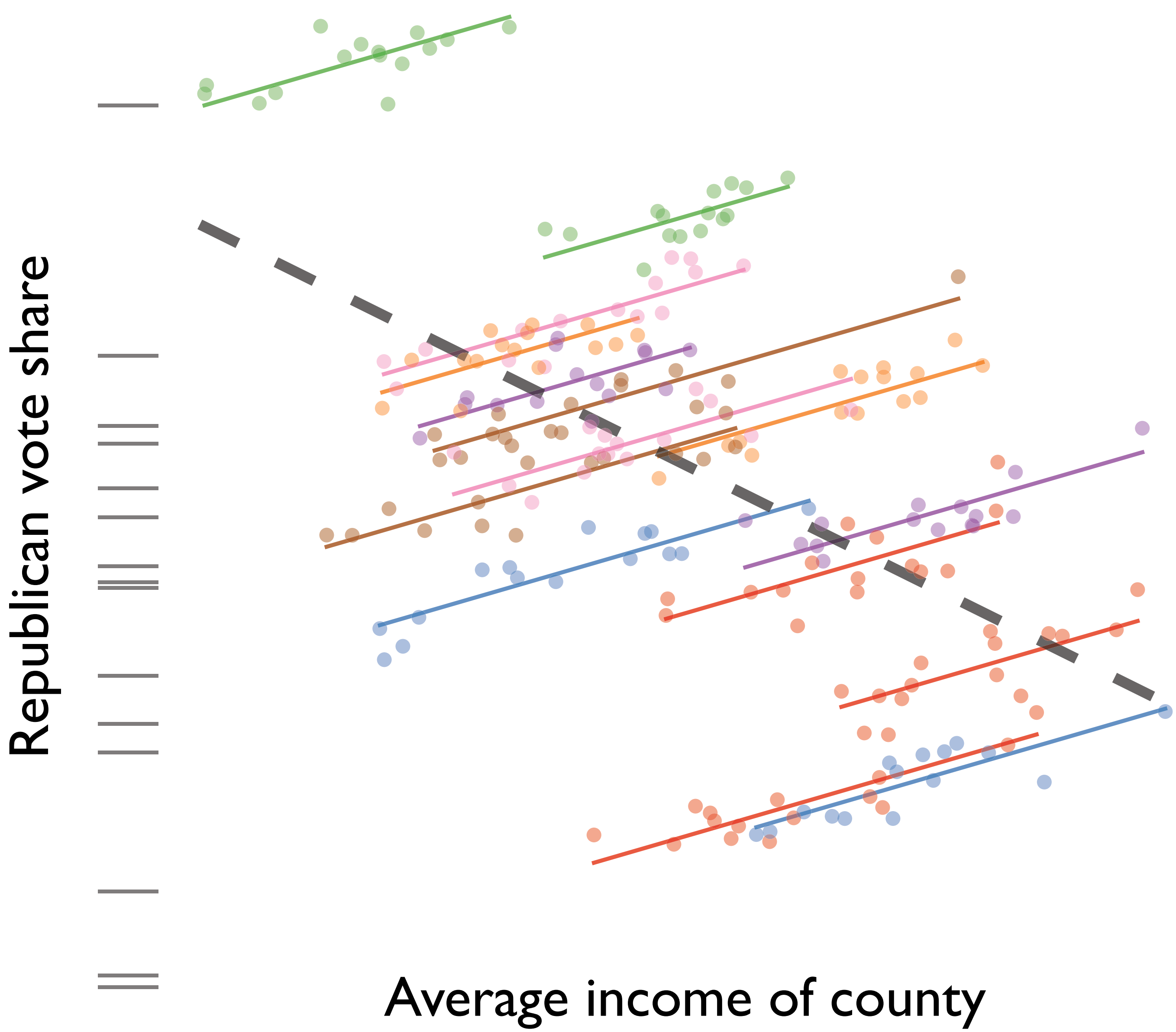
But Republican elected officials attempt to represent the affluent

What's the matter with (poor counties in) Kansas, as Thomas Frank asked?

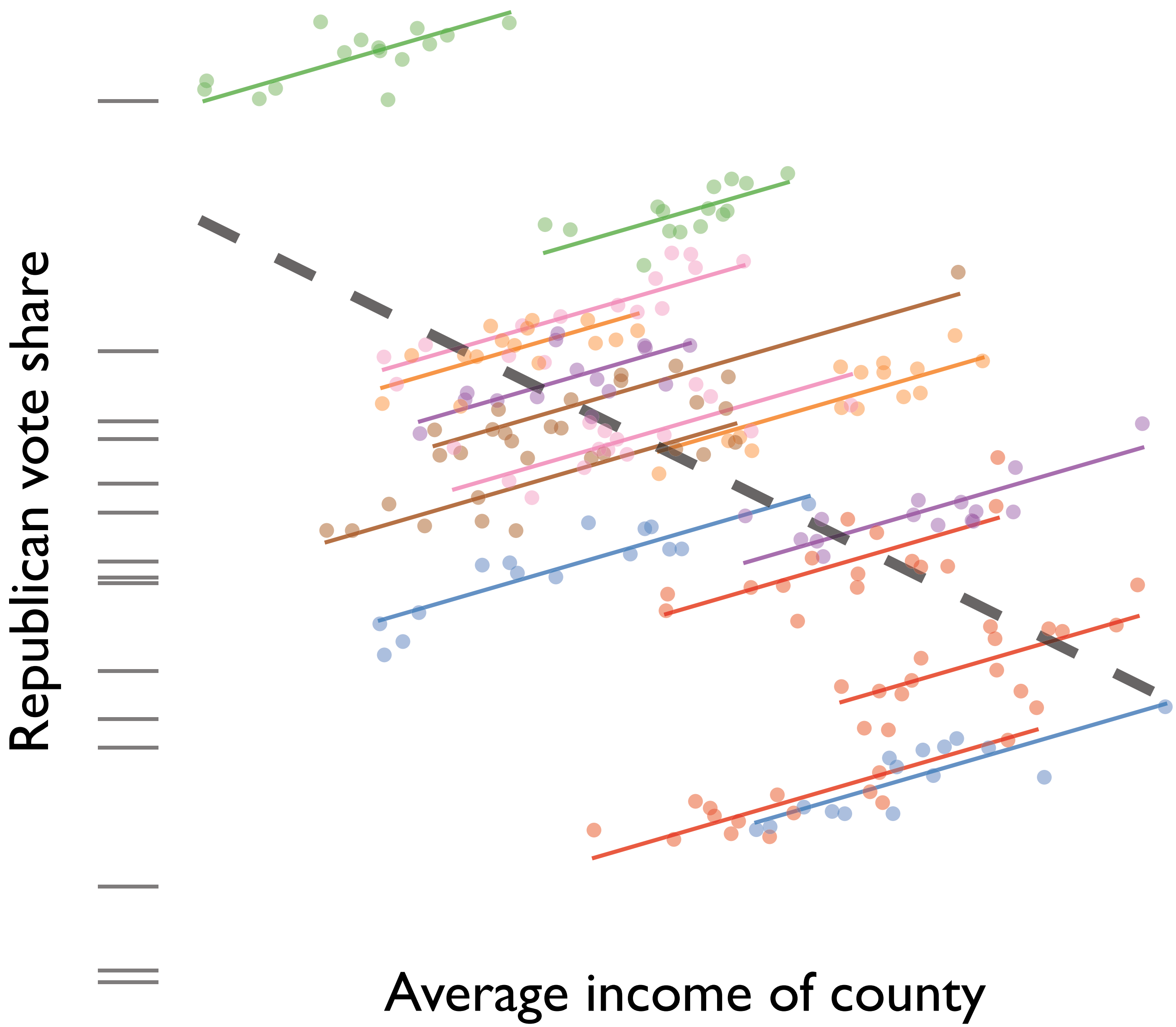


Let’s look at which observations come from which states

Clearly, counties from the same state are clustered

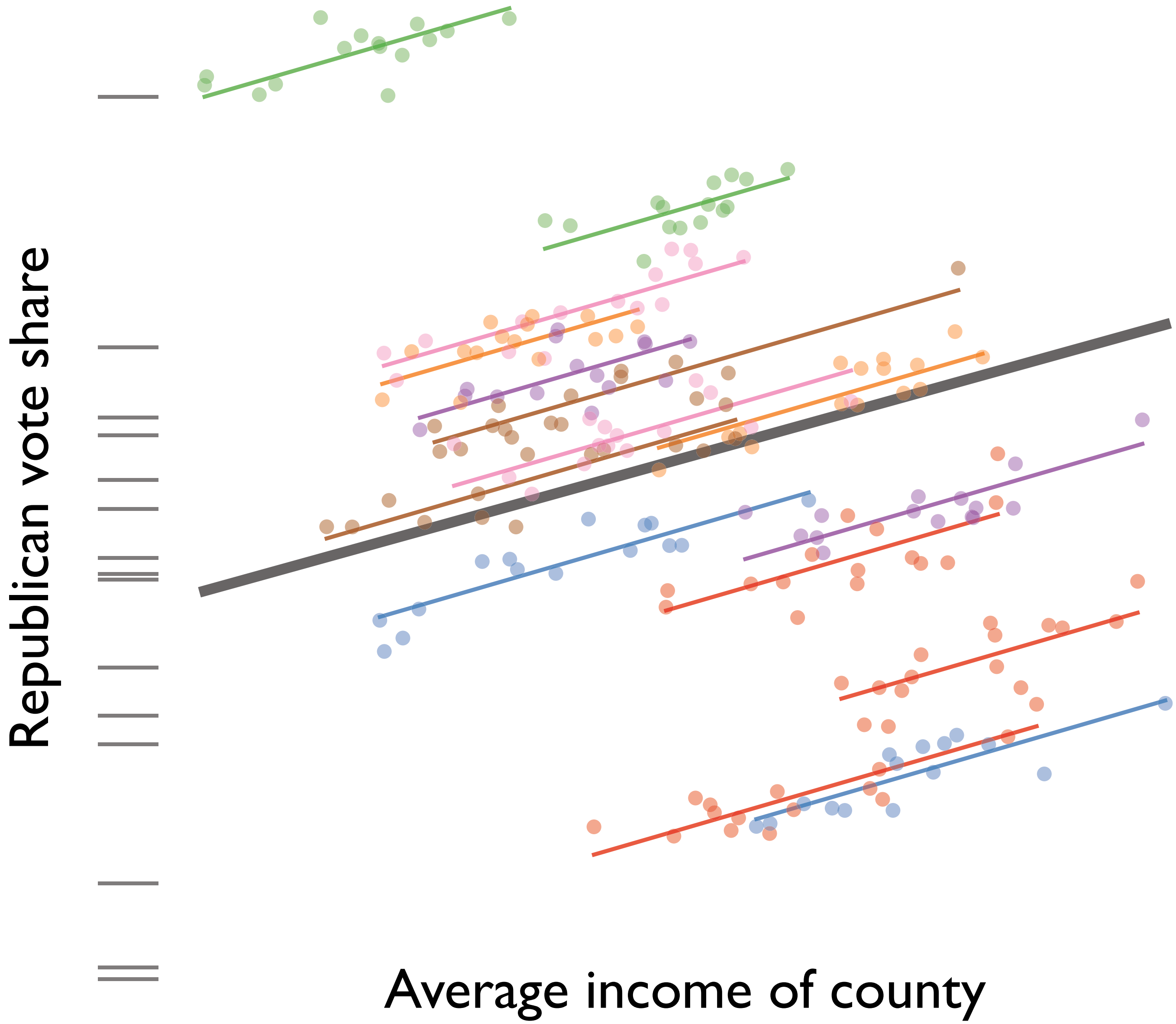


Within each state, there's a *positive* relationship
between income & voting Republican



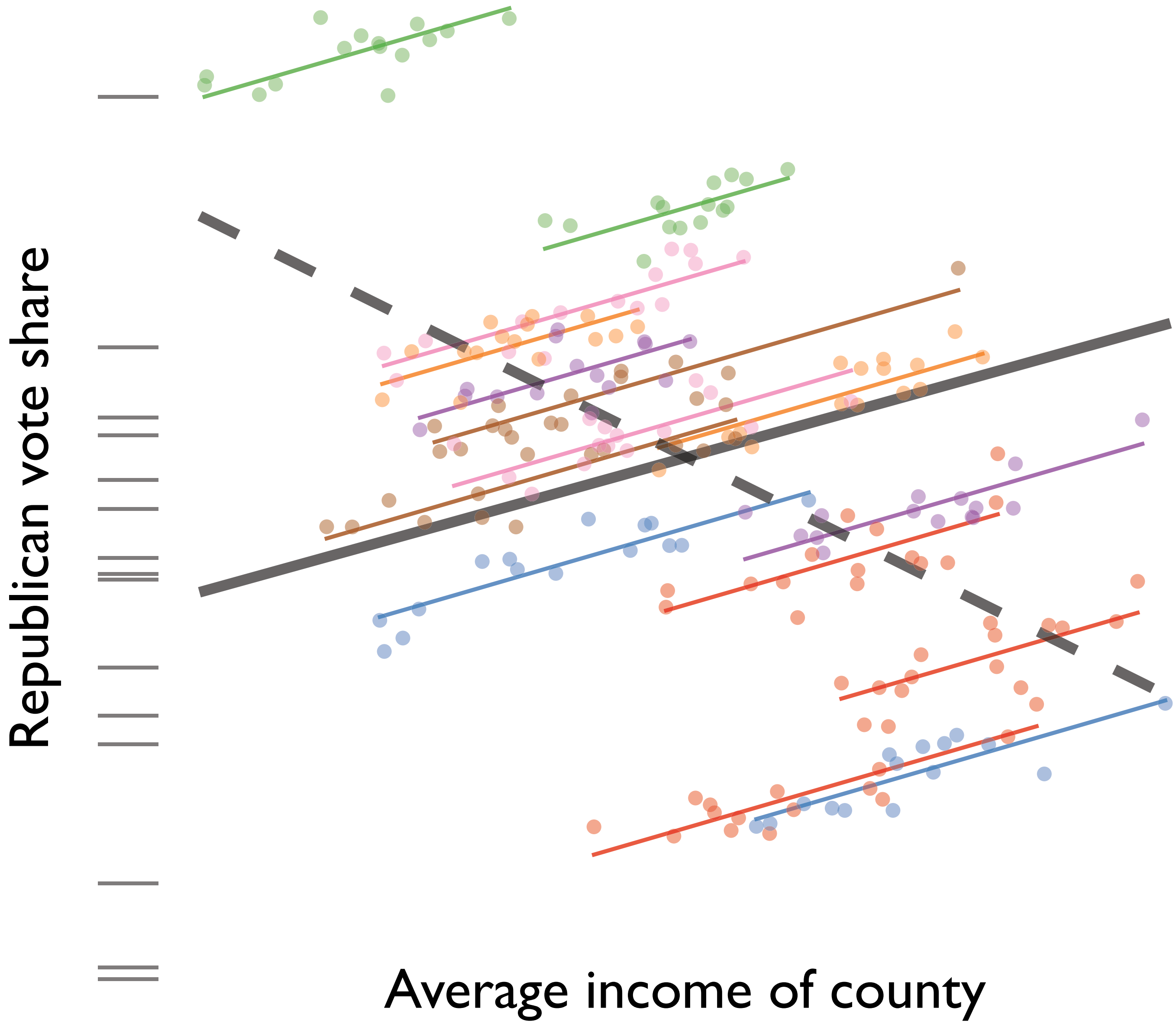
Suggests we need to control for variation at the state level,

either by collecting the state level variables causing the variation, or. . .



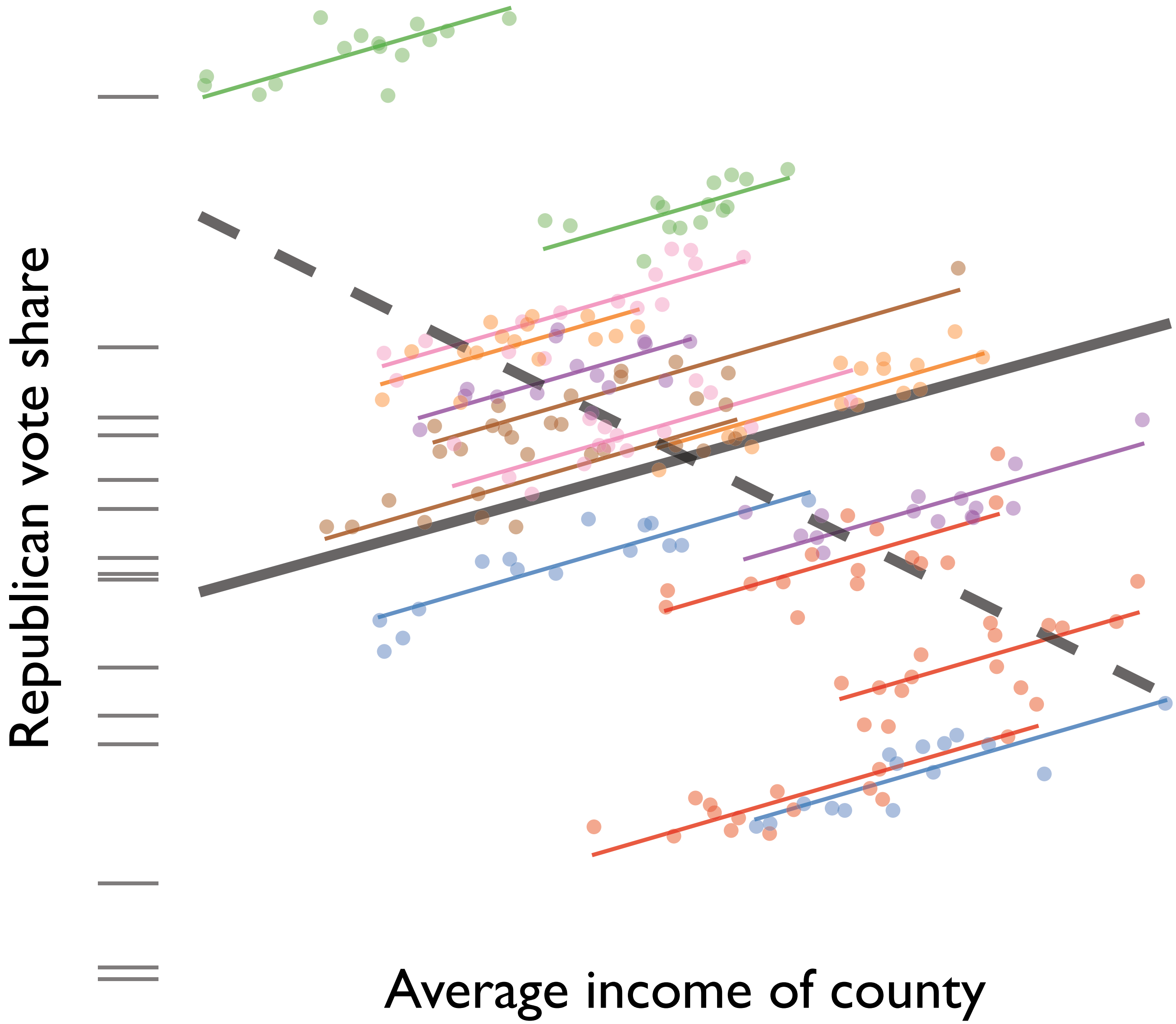
use brute force: add a dummy for each state to the matrix of covariates

to purge the omitted variable bias

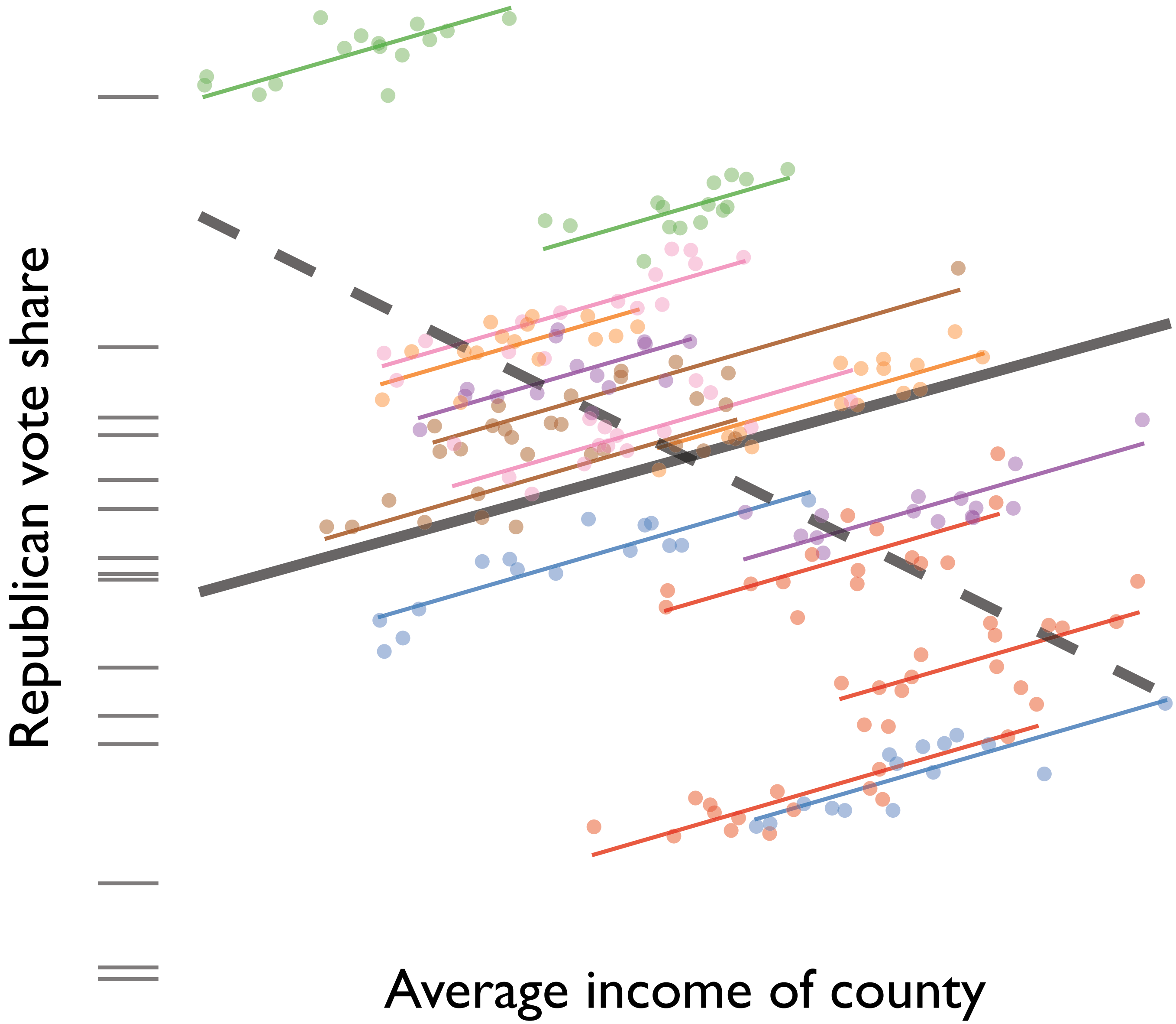


Controlling for state fixed effects, $\hat{\beta}_1$ flips signs!

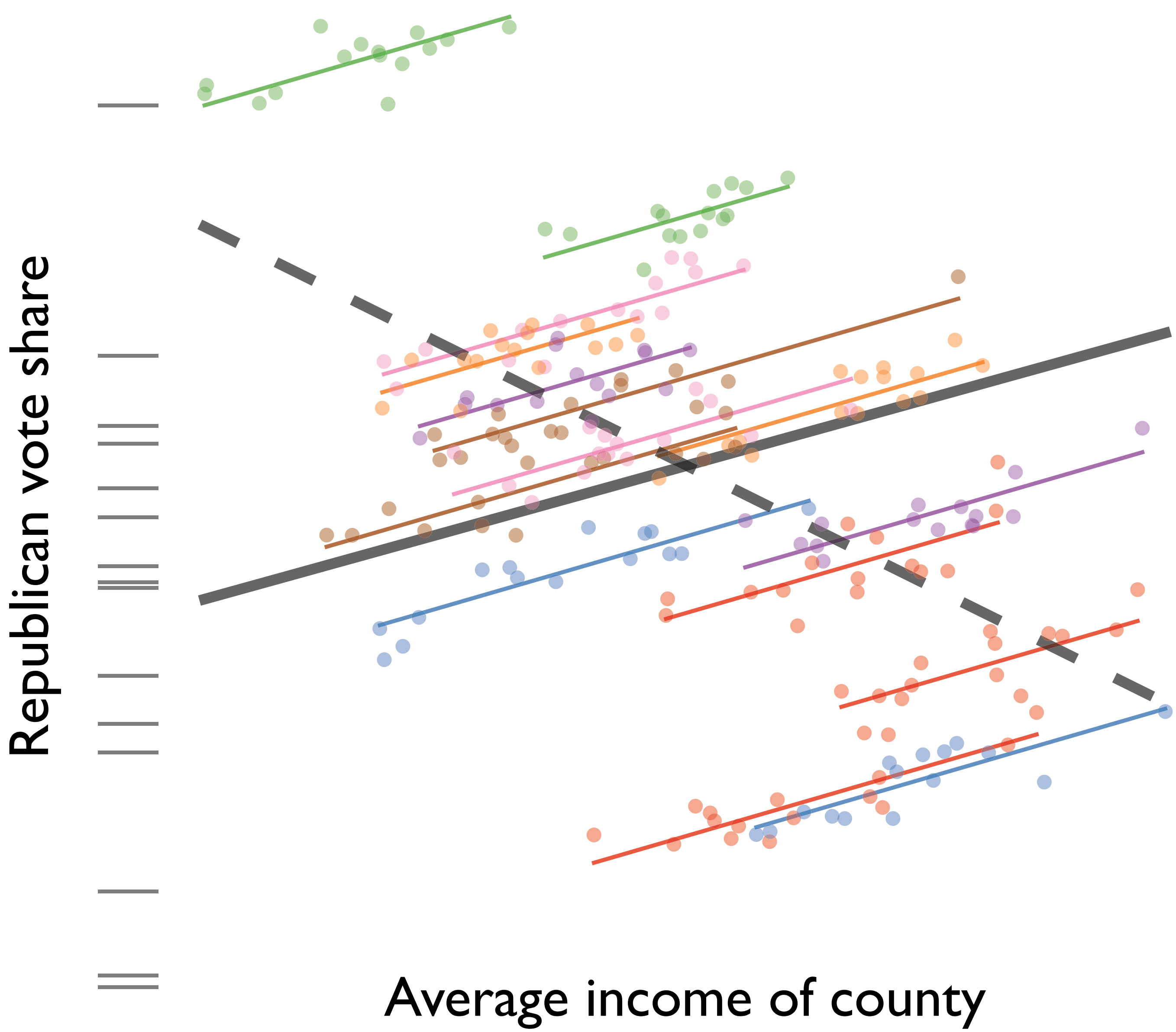
Including fixed effects for each state removes state-level omitted variable bias



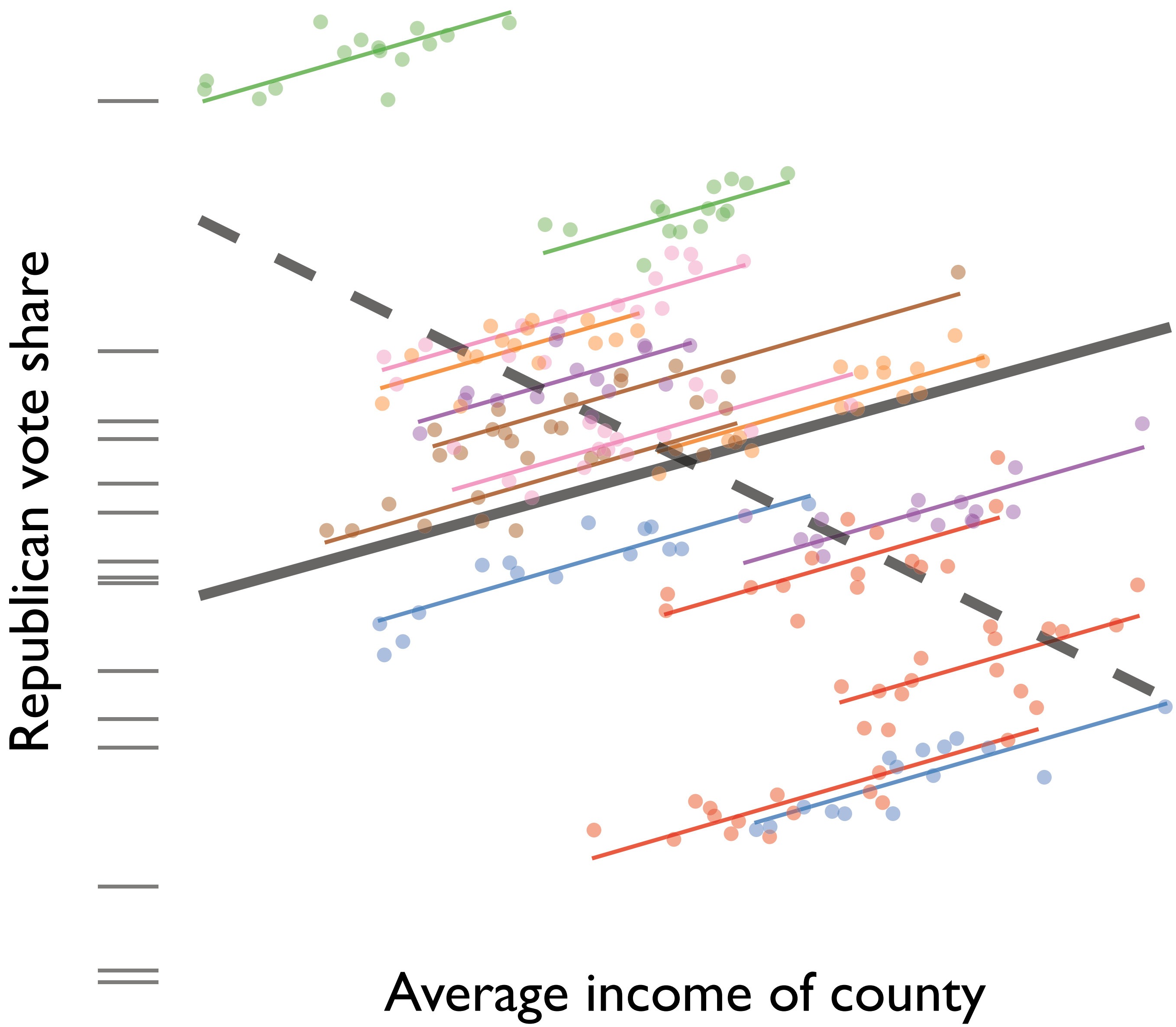
What’s the matter with Kansas? On average, Kansans are more conservative than other Americans, but within Kansas, the same divide between rich and poor holds



What's the matter with Kansas? On average, Kansans are more conservative than other Americans, but within Kansas, the same divide between rich and poor holds *. . . or at least it did until 2016*

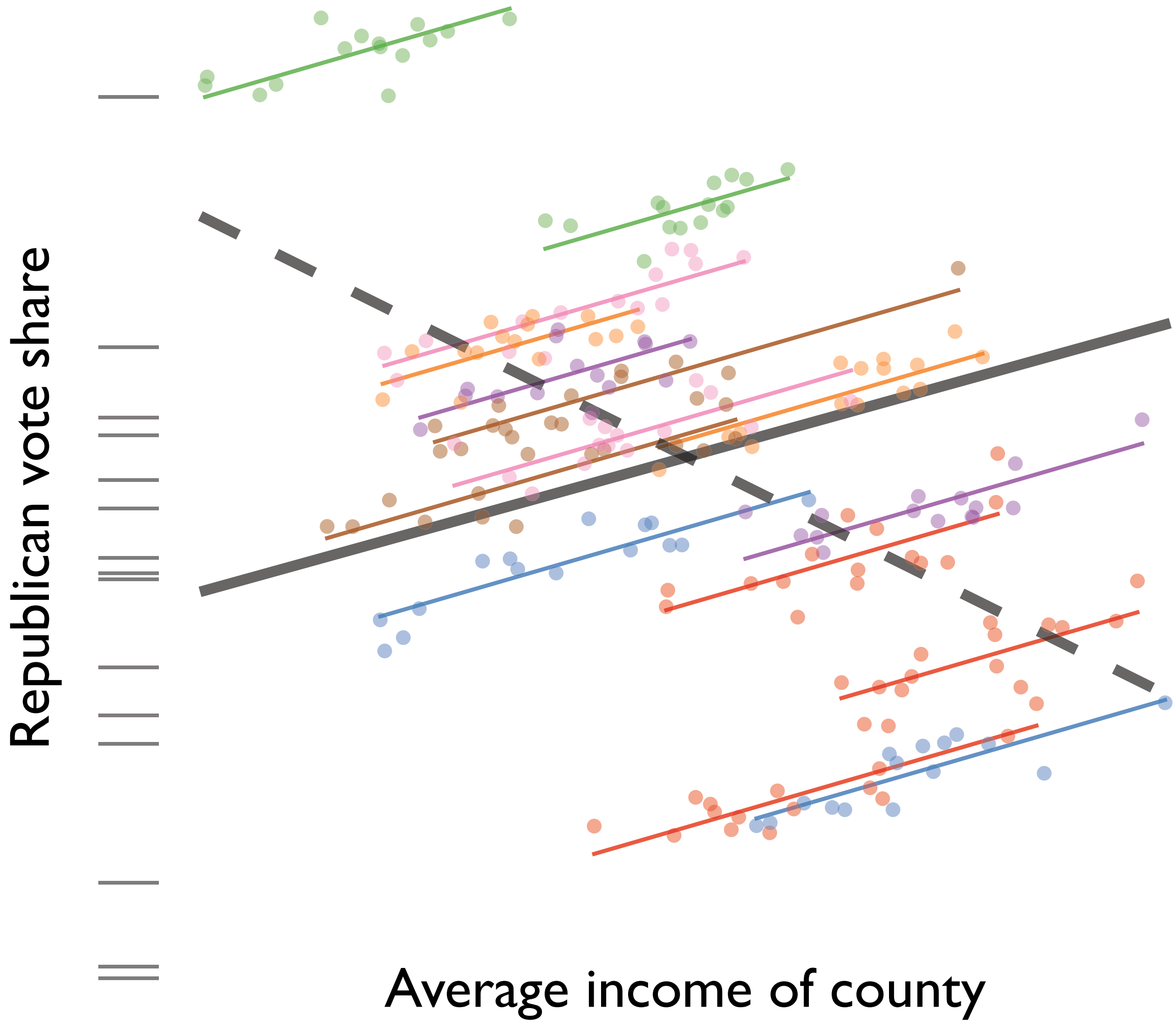


How are fixed effects different from random effects?



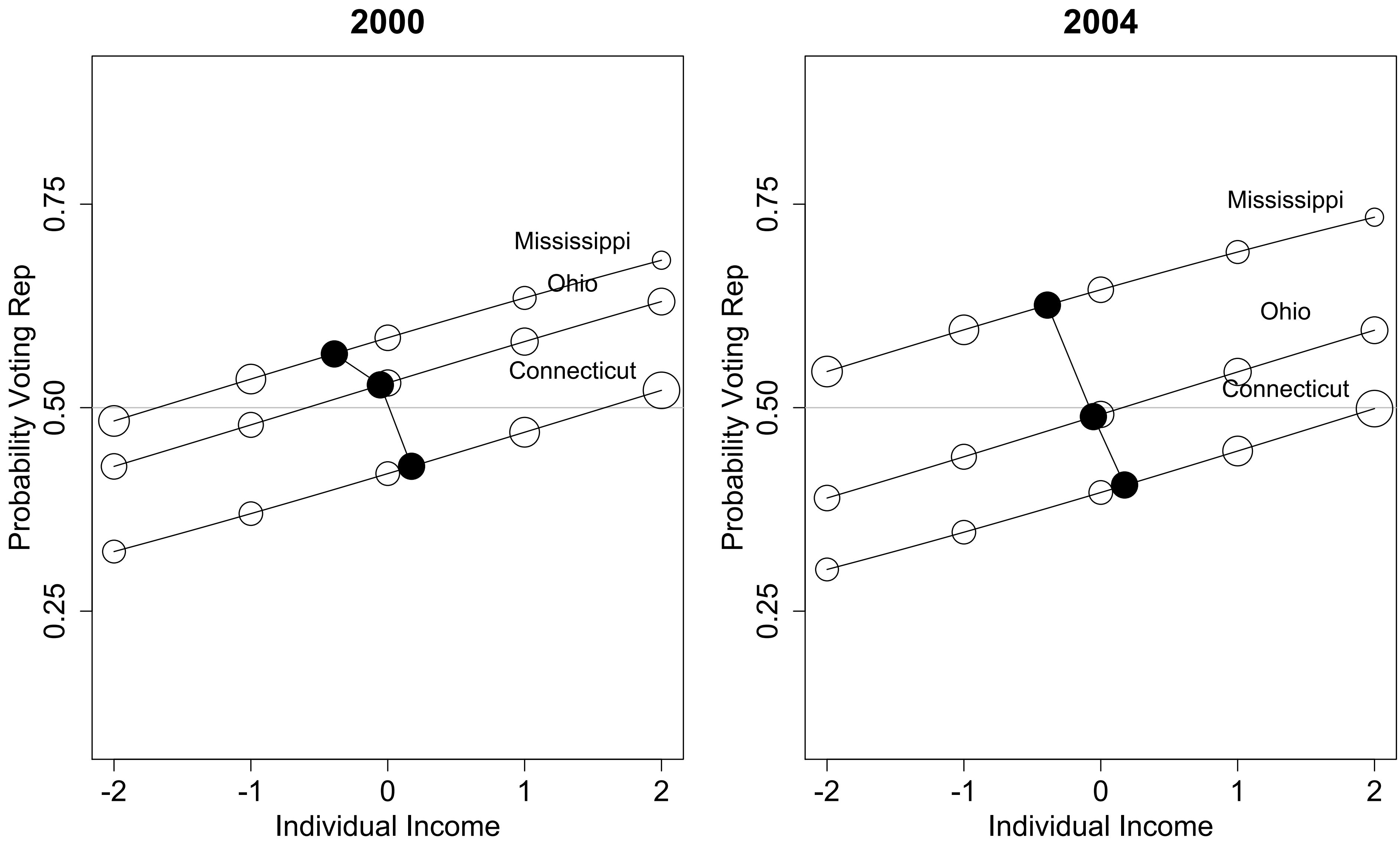
Fixed effects control for omitted variables *random effects don't*

Fixed effects don't follow any particular distribution *random effects do*



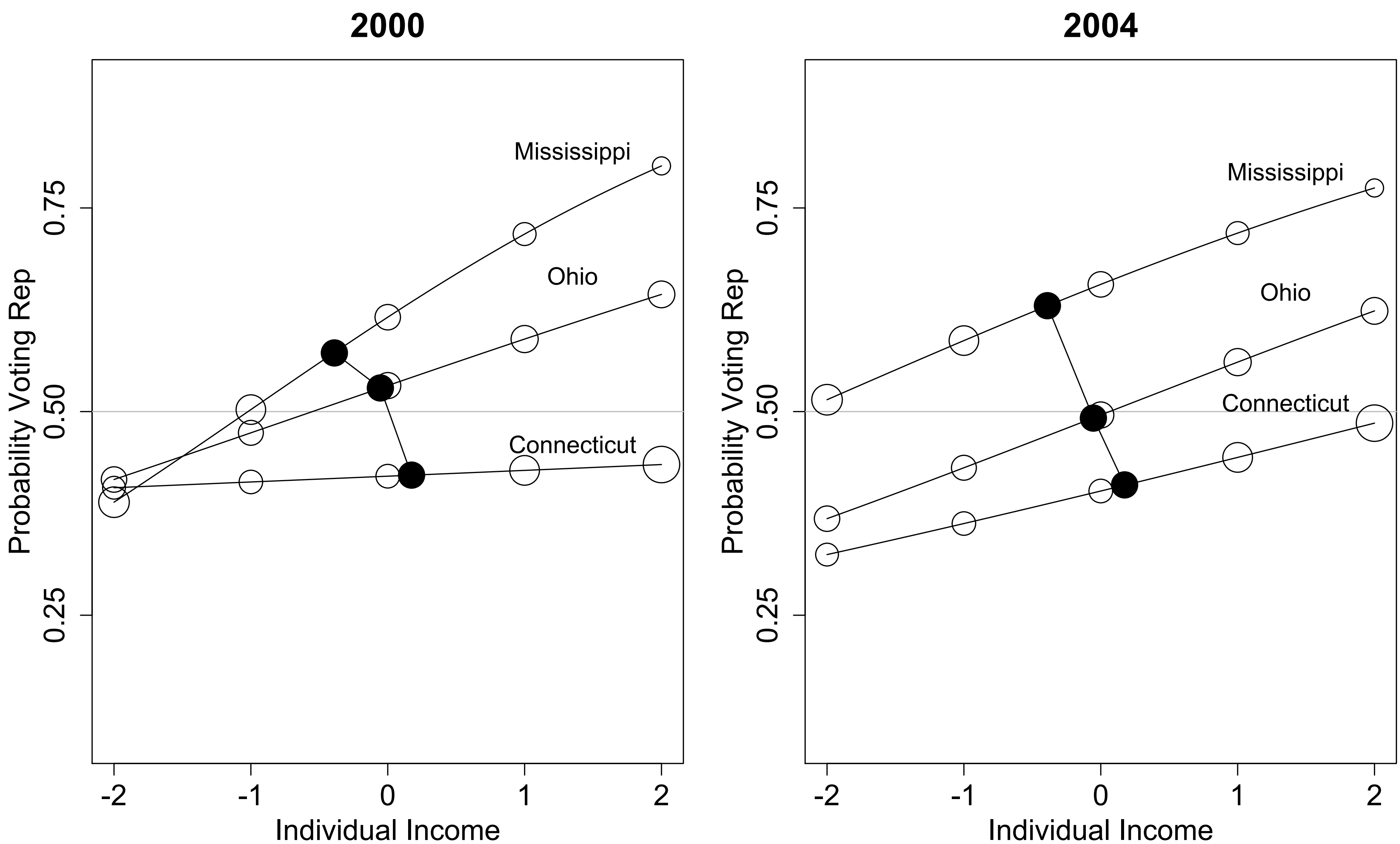
Aside 1 the above reversal is an example of the *ecological fallacy*, which says that aggregate data can mislead us about individual level relationships

Here, the pattern across states mislead us as to the pattern within states



Aside 2 Above are results on actual data from Gelman *et al*

This version of their model assume intercepts (but not slopes) vary by state



Aside 2 When Gelman *et al* allow slopes $\hat{\beta}_{1i}$ to vary across states, they find the rich-poor divide is actually *steeper* in poor states!

Variable slopes and intercepts

$$\Delta^d y_{it} = \alpha_i + \mathrm{x}_{it}\boldsymbol{\beta}_i + \sum_{p=1}^{P} \Delta^d y_{i,t-p}\phi_p + \sum_{q=1}^{Q} \varepsilon_{i,t-q}\rho_q + \varepsilon_{it}$$

How do we let $\boldsymbol{\beta}_i$ vary over the units?

For the kth covariate x_{kit}, let $\boldsymbol{\beta}_{ki}$ be random, with a multivariate Normal distribution

$$\boldsymbol{\beta}_{ki} \sim \mathrm{MVN}(\boldsymbol{\beta}^*_{ki}, \Sigma_{\boldsymbol{\beta}_{ki}})$$
$$\boldsymbol{\beta}^*_{ki} = \mathbf{w}_i\boldsymbol{\zeta}$$

That is, the β_{ki}'s are now a function of *unit-level covariates* $\mathbf{w}_i$
and their associated *hyperparameters* $\boldsymbol{\zeta}$

Variable slopes and intercepts

$$\text{GDP}_{it} = \phi_1 \text{GDP}_{i,t-1} + \alpha_i + \beta_1 \text{Democracy}_{it} + \varepsilon_{it}$$

Variable slopes and intercepts

$$\text{GDP}_{it} = \phi_1 \text{GDP}_{i,t-1} + \alpha_i + \beta_1 \text{Democracy}_{it} + \varepsilon_{it}$$

$$\alpha_i \sim \text{N}(0, \sigma^2_\alpha)$$

Variable slopes and intercepts

$$\begin{aligned}
\text{GDP}_{it} &= \phi_1 \text{GDP}_{i,t-1} + \alpha_i + \beta_1 \text{Democracy}_{it} + \varepsilon_{it} \\
\alpha_i &\sim \text{N}(0, \sigma^2_{\alpha}) \\
\beta_1 &\sim \text{N}(\boldsymbol{\beta}^*_{1i}, \sigma^2_{\boldsymbol{\beta}_{1i}})
\end{aligned}$$

Variable slopes and intercepts

$$
\begin{aligned}
\mathrm{GDP}_{it} &= \phi_1 \mathrm{GDP}_{i,t-1} + \alpha_i + \beta_1 \mathrm{Democracy}_{it} + \varepsilon_{it} \\
\alpha_i &\sim \mathrm{N}(0, \sigma^2_\alpha) \\
\beta_1 &\sim \mathrm{N}(\boldsymbol{\beta}^*_{1i}, \sigma^2_{\boldsymbol{\beta}_{1i}}) \\
\beta^*_{1i} &= \zeta_0 + \zeta_1 \mathrm{Education}_i
\end{aligned}
$$

Now the effect of Democracy on GDP varies across countries, as a function of their level of Education *and* a country random effect with variance $\sigma^2_{\boldsymbol{\beta}_{1i}}$

This is now a *multilevel* or *hierarchical* model

See Gelman & Hill for a nice textbook on these models

Easiest to accomplish using Bayesian inference
(place priors on each parameter and estimate by MCMC)

Variable slopes and intercepts: Poor man's version

$$\text{GDP}_{it} = \phi_1 \text{GDP}_{i,t-1} + \alpha_i + \beta_1 \text{Democracy}_{it}$$
$$+\beta_2 \text{Democracy} \times \text{Education} + \varepsilon_{it}$$
$$\alpha_i \text{ is a matrix of country dummies}$$

This version omits the random effects for α_i and β_i; instead, we have fixed country effects

and a fixed, interactive effect that makes the relation between Democracy and GDP conditional on Education

Note that we can't include an Education base term—it's part of the fixed effects already

But we can include the time invariant Education variable *within* a time-varying interaction

Should have approximately similar results to hierarchical