Unit 3, Lecture 1
Longitudinal Data and Growth Curves

Pop Quiz on Resources

You are planning on writing some math equations.

What do you do to get help?

Pop Quiz on Resources

You want to do a likelihood ratio test.

What do you do to get help?

Pop Quiz on Resources

You want to group mean center your data, but forget how.

What do you do to get help?

Pop Quiz on Resources

You want to know what class is going to cover in three weeks.

How do you find out?

Bonus: How do you find out how to prepare?

Group work means collaboration, not divide-and-conquer

Get-to-know-you

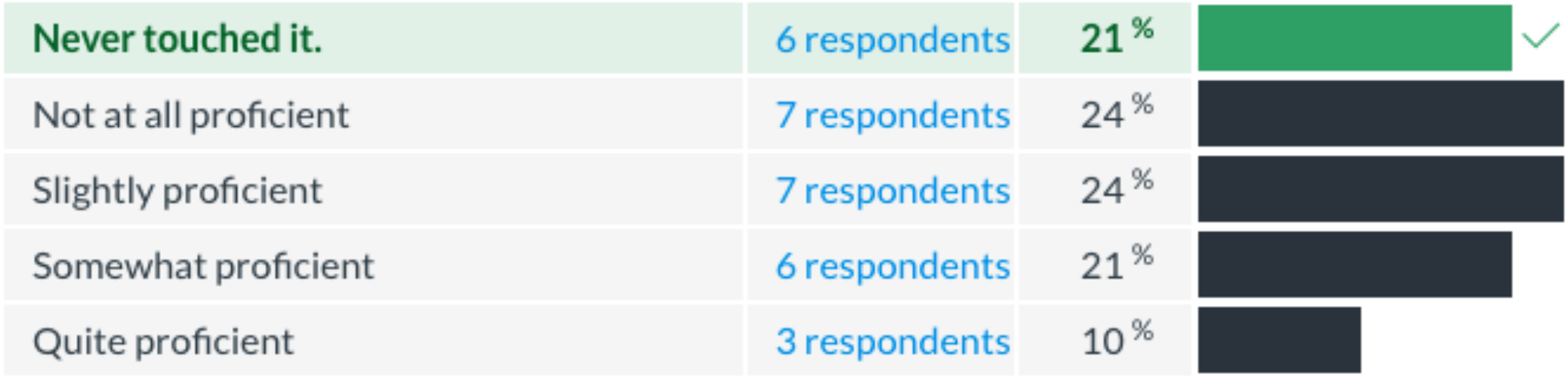
Thank you for the responses.

Lots of interesting results.

They mainly that show two things:

- ★There is a lot of variation.
- ★No one is alone.

Let's get more specific about R and your past exposure: how proficient were you at using R when you joined this course?



Looking back & moving forward

Review of Unit 2 and Connections to the Future (i.e. Unit 3)

★ What is a random intercept model?
- Foundational, not used in longitudinal

★ Level 2 predictors
- Super important - Level 2 will be "child" and we use their covariates everywhere.

★ Random slope models
- Super super important - a "growth model" is a random slope model, with the slope being outcome vs. time

★ Centering / Between vs. Within
- Centering time comes up a lot. Between vs. Within is secondary.

★ Standard Errors & Confidence Intervals
- All as before. We will demonstrate how to use these things moving forward.

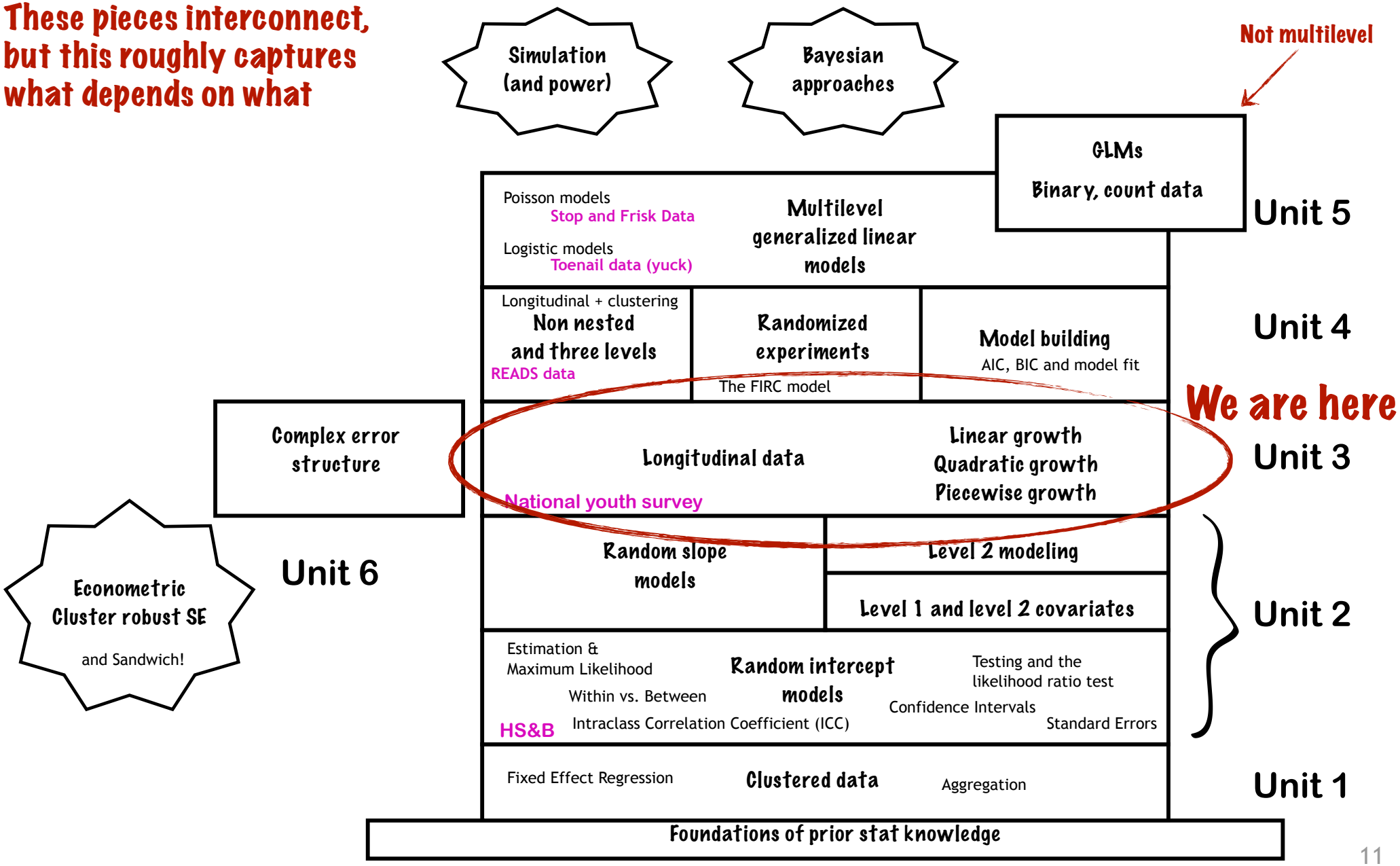
★ Inference with MLMs / Likelihood Ratio Test
- As above. Inference is the same, and we will now use it in our discussions of results.

More Commentary on Connections

We learned:

- ★ What a multilevel model is
  - Many small worlds
  - Variance components
  - NOW: our worlds are individuals, with time inside them
- ★ How to fit a multilevel model
  - This will be the same
- ★ Inference with multilevel models
  - Longitudinal models are MLMs - the inference is identical (and we will see examples of it moving forward)
- ★ Interpreting multilevel model fits
  - How much variation is there?
  - How much variation is explained?

A Course Map: Building up layers



Todays Goals

★ Introduce idea of longitudinal data

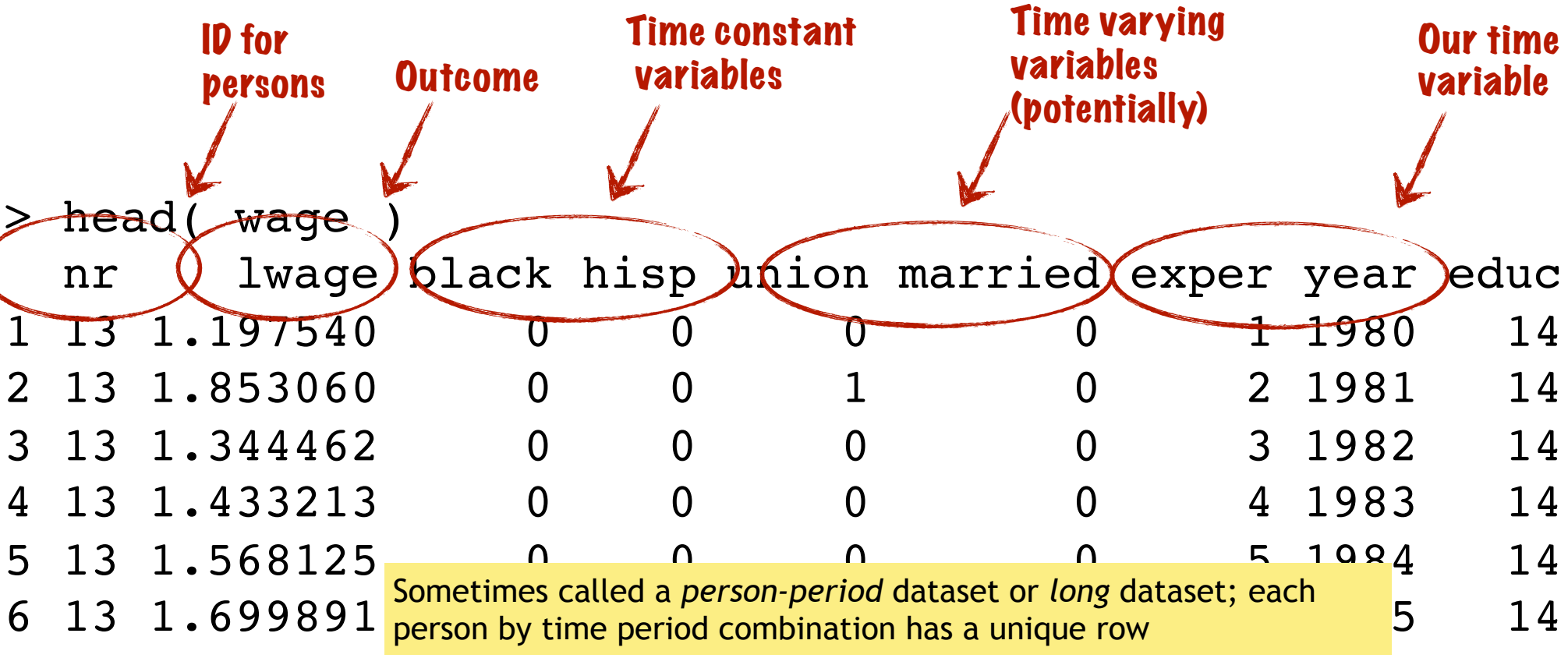
★ Provide a simple model for fitting such data: the linear growth model

★ Walk through a case-study taken from the textbook illustrating the interpretation of such models.

Overview of Longitudinal Data

Longitudinal Data Example: Wages

We watch individuals across time, collecting information from them at each of a series of time points:




```
> head( wage )
   nr    lwage black hisp union married exper year educ
1  13 1.197540     0    0     0       0     1 1980   14
2  13 1.853060     0    0     1       0     2 1981   14
3  13 1.344462     0    0     0       0     3 1982   14
4  13 1.433213     0    0     0       0     4 1983   14
5  13 1.568125     0    0     0       0     5 1984   14
6  13 1.699891     0    0     0       0     6 1985   14
```

Sometimes called a *person-period* dataset or *long* dataset; each person by time period combination has a unique row

Large scale goals

Motivating Research Question:
How and why do wages change across time?

- ★ time invariant stuff could predict initial wage and how wage changes over time
- ★ time varying stuff could explain wage in a given year

We might model in one fell swoop OR think of each year depending on past information (e.g., current year depends on *lagged* (prior) year).

This boils down to different concerns with the error structure for individual observations.

How are these data multilevel?

1. Individuals contribute multiple observations to the dataset.
2. Essentially, individuals are sampled, and then specific observations are "sampled" from within individuals.
3. Individuals are now **level 2.** *Time* is **level 1.**
4. Observations about individuals which do not change are level 2 variables.
5. Observations about individuals which *do* change are level 1 variables.
6. We often have additional levels (e.g., students in school or schools in district), but will ignore this for now.

Clustered data vs. Longitudinal data (Or, how are these multilevel data strange?)

Clustered Data:

★ We observe a bunch of students in a school. The students have no order. They are independent within the school.

Longitudinal Data:

★ We observe a bunch of times. The times are *ordered*. They are not independent within the student.

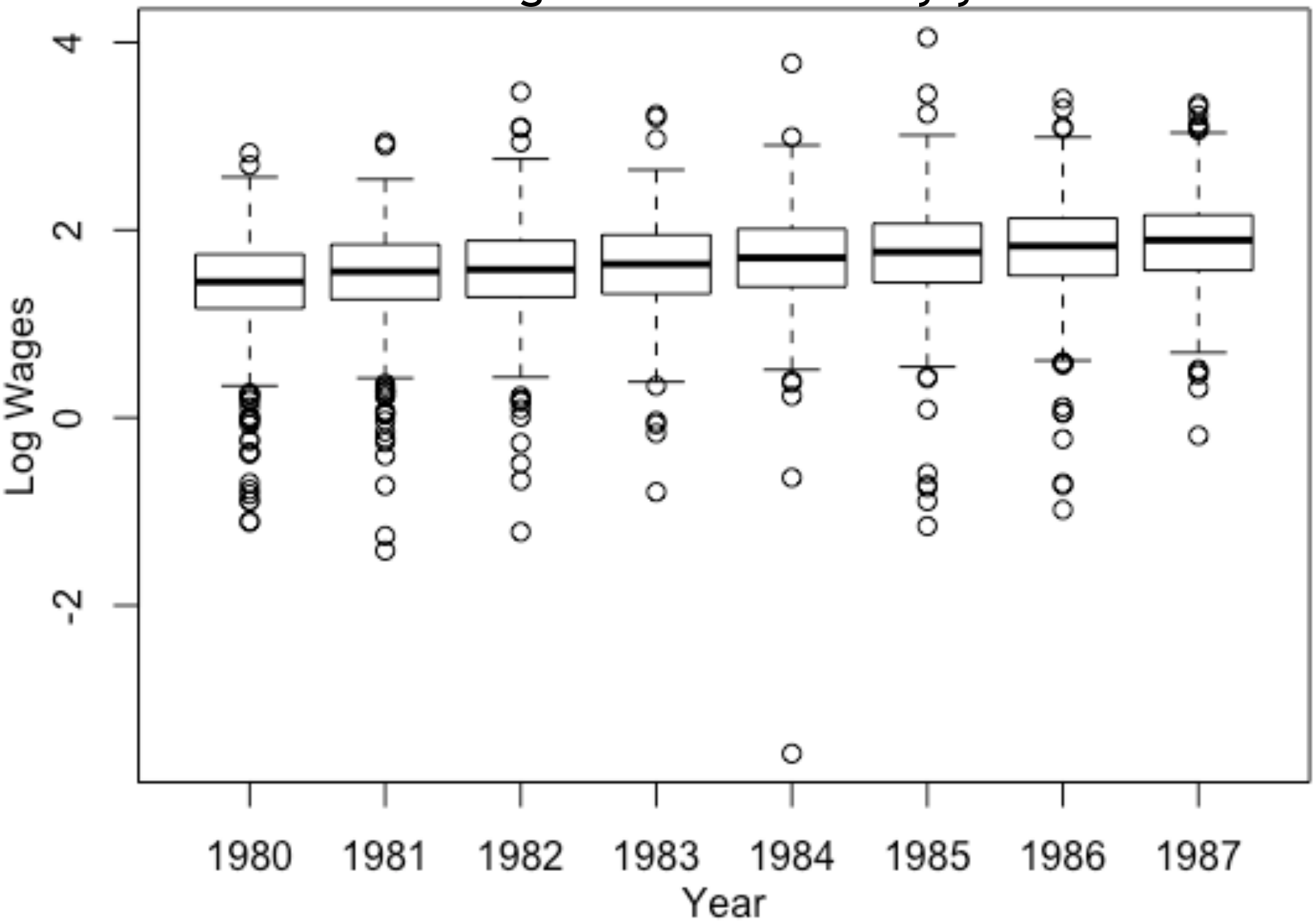
★ This ordering of points based on time of measurement is meaningful in a way that is not true of, e.g., students in a school

★ Consider: We expect the first and second observation to be closer than the first and last.

★ The time points aren't really sampled, so it looks less like classic multilevel structure.

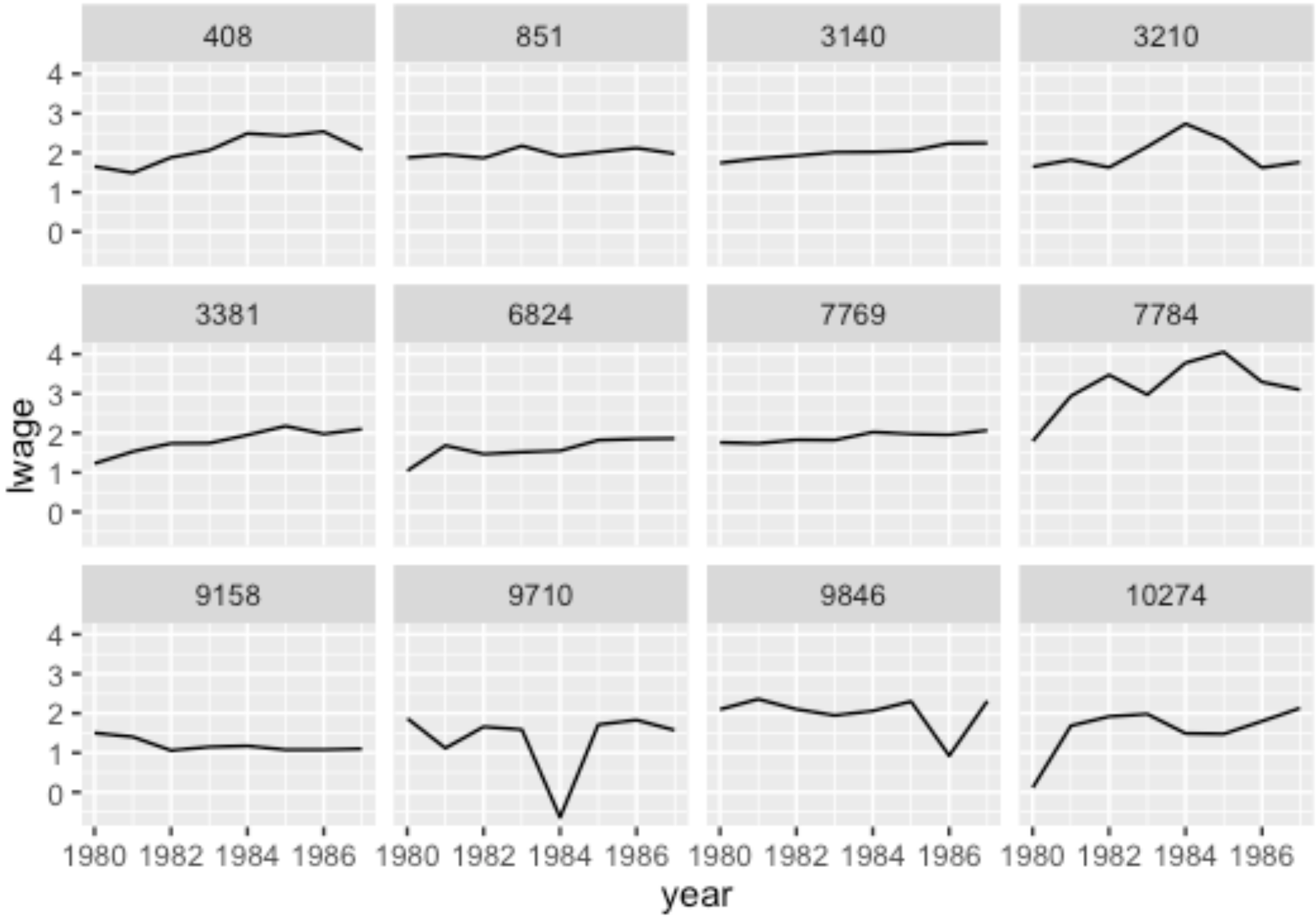
★ We typically have only a handful of level-1 units per level-2 unit (So lots of units with few observations instead of few units with many.)

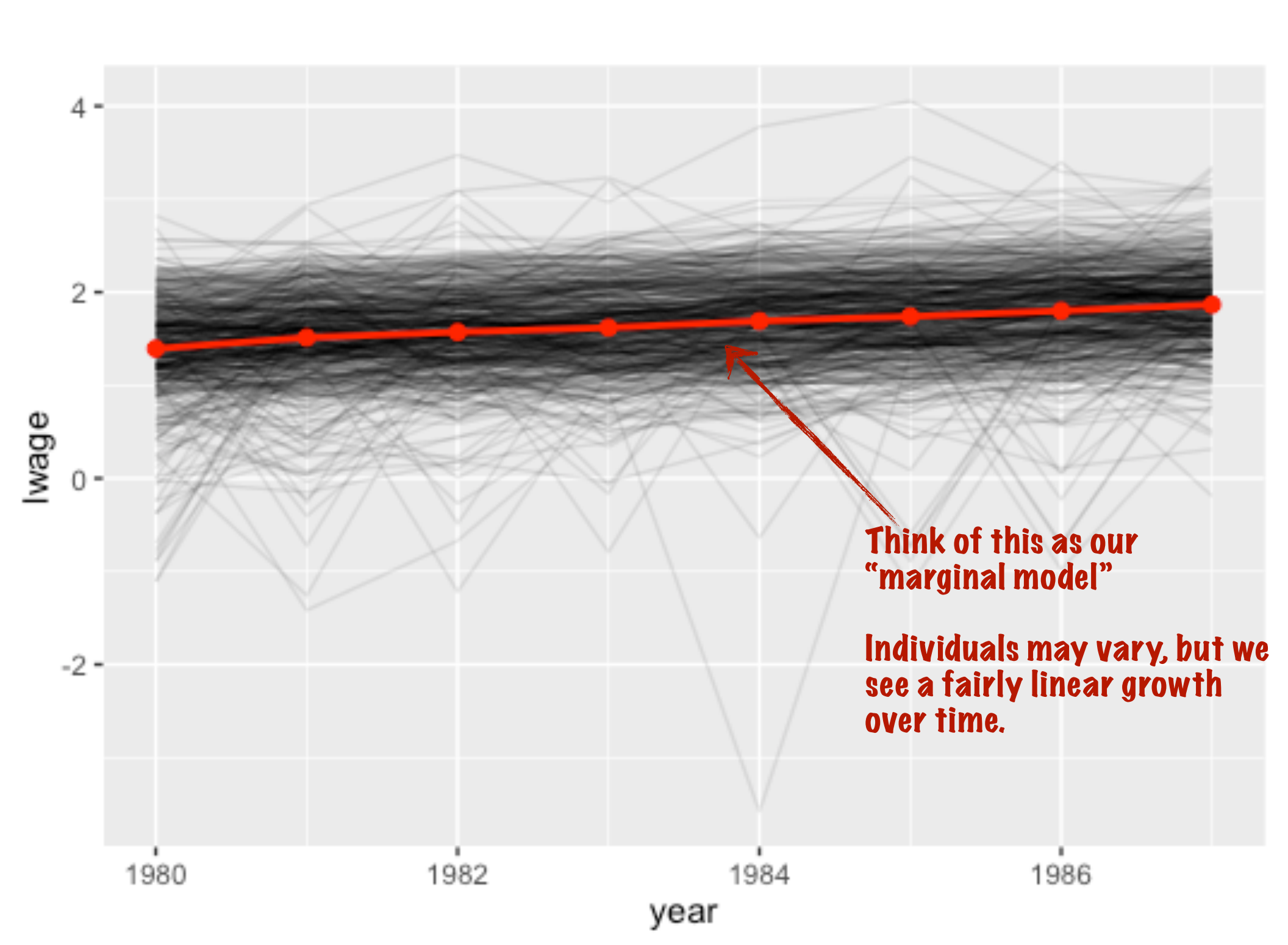
Looking at distribution by year



ALWAYS LOOK AT YOUR INDIVIDUALS

```
# plot 12 random people

# sample of individuals
samp.id = sample( unique( wage$nr ), 12 )

wage.samp = filter( wage, nr %in% samp.id )

ggplot( data=wage.samp, aes( x=year, y=lwage ) ) +
    facet_wrap( ~ nr ) +
    geom_line()


# plot everybody with overall trend line

# Calculate means for each time point
mean.wage = wage %>% group_by( year ) %>%
          summarize( lwage = mean( lwage ) )

# Plot all individuals with our means in red
ggplot( data=wage, aes( x=year, y=lwage )  )+
    geom_line( aes( group=nr ), alpha=.10 ) +
    geom_line( data=mean.wage, lwd=1, col="red" ) +
    geom_point( data=mean.wage, col="red", cex=2 )
```


408
851
3140
3210
3381
6824
7769
7784
9158
9710
9846
10274
lwage
year
0
1
2
3
4
1980
1982
1984
1986


4
2
0
-2
lwage
1980
1982
1984
1986
year
Think of this as our "marginal model"
Individuals may vary, but we see a fairly linear growth over time.

Panel Studies vs. Cohort Studies

Panel Studies

- ★ Track everyone at the same sequence of time points
- ★ These timepoints are called “waves”

Cohort Studies

- ★ Follow a group
- ★ Variable check-in times

Balance (panel data): When we have an observation for each person at each time point.

With longitudinal data, we must face some data wrangling issues that bring great pain

This is

LONG vs. WIDE

Concept: tidy data

One critical coding philosophy is to think of your work as a series of steps that you do IN ORDER:

1. Get your data ready
2. Visualize your data
3. Analyze your data

"Getting data ready" is when you tidy your data. In short, you want to turn your data into a nice rectangle! *Which rectangle you use depends on your purposes*.

A worked, toy example: wide data

This data is in so-called *wide* format. The waves correspond to different time points

```
> dta
  id   X    wave1    wave2    wave3
   A 1.3 3.398506 2.854276 3.659517
   B 2.2 3.121525       NA 4.281374
   C 1.5 3.379317 4.905583 4.595625
```

This is a good rectangle for many purposes

But a bad rectangle for other purposes!

Say we know wave1 is at start, wave2 was 1 year out, wave3 is 4 years out.

Generally, data is generally easiest to use in *long* format where no variable is repeated multiple times in a row.

We want our data to be 1 outcome per row

"Gathering" a variable to go to long format

```
library( tidyverse )
dat <- gather( dta, wave1, wave2, wave3,
                    key='wave', value='score' )
```

```
> dat
  id   X  wave       score time
1  A 1.3 wave1 0.01055856    0
2  B 2.2 wave1 3.54820520    0
3  C 1.5 wave1 2.09044686    0
4  A 1.3 wave2 3.49256775    1
5  B 2.2 wave2          NA    1
6  C 1.5 wave2 3.37300287    1
7  A 1.3 wave3 3.02156096    4
8  B 2.2 wave3 4.15792712    4
9  C 1.5 wave3 6.23645127    4
```

"Gathering" (in the tidyverse) is the cleanest way to reshape data. (But this can only reshape a single variable at a time.)

Read R for Data Science, section 12

For more powerful reshape commands, use reshape()- see appendix and handout for more

Pro Tip: Save your cleaned data

Once your data are clean, save them to a file.

```
write.csv( dta, file="clean_data.csv",
                        row.names=FALSE )
```

or

```
saveRDS( dta, file="clean_data.rds" )
```

Then load the data later on.

This decouples your data cleaning from your main work.

Making your own time

```
> dat$time = 0
> dat$time[ dat$wave=="wave2" ] = 1
> dat$time[ dat$wave=="wave3" ] = 4

> dat
  id   X  wave       score time
1  A 1.3 wave1 0.01055856    0
2  B 2.2 wave1 3.54820520    0
3  C 1.5 wave1 2.09044686    0
4  A 1.3 wave2 3.49256775    1
5  B 2.2 wave2          NA    1
6  C 1.5 wave2 3.37300287    1
7  A 1.3 wave3 3.02156096    4
8  B 2.2 wave3 4.15792712    4
9  C 1.5 wave3 6.23645127    4
```

Turn our wave (a factor) into time (a number) This is especially important if different waves are separated by different lengths of time

Missing data bad

```
> dat
   id   X  wave       score time
1   A 1.3 wave1 0.01055856    0
2   B 2.2 wave1 3.54820520    0
3   C 1.5 wave1 2.09044686    0
4   A 1.3 wave2 3.49256775    1
5   B 2.2 wave2          NA   1
6   C 1.5 wave2 3.37300287    1
7   A 1.3 wave3 3.02156096    4
8   B 2.2 wave3 4.15792712    4
9   C 1.5 wave3 6.23645127    4
```

Missing data in your dataset can cause trouble. Means will return NA, plots will drop stuff. =(

```
# How to drop missing data
> dat = na.omit( dat )
```

This drops all rows with an NA

The resulting data are "complete cases" leading to a "complete case analysis"
IMPORTANT: drop unused variables so you don't drop cases missing irrelevant variables

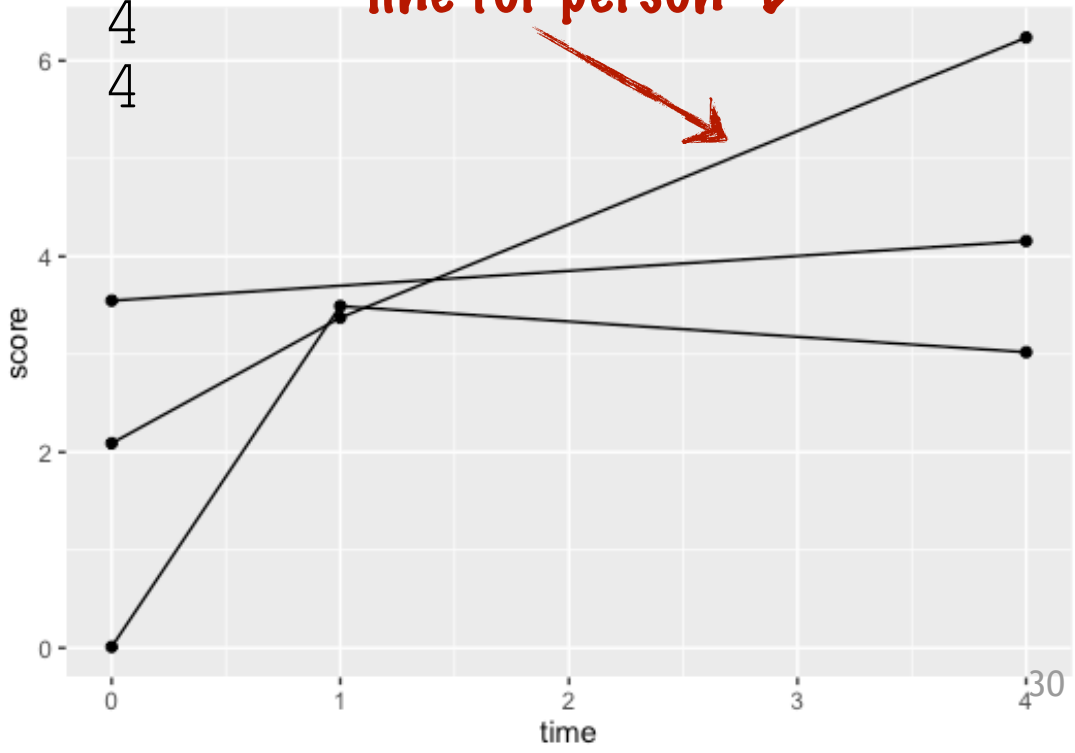
Plotting and missing data

```
> dat = na.omit( dat )
> dat
   id   X  wave       score time
1   A 1.3 wave1 0.01055856    0
2   B 2.2 wave1 3.54820520    0
3   C 1.5 wave1 2.09044686    0
4   A 1.3 wave2 3.49256775    1
6   C 1.5 wave2 3.37300287    1
7   A 1.3 wave3 3.02156096    4
8   B 2.2 wave3 4.15792712    4
9   C 1.5 wave3 6.23645127    4
```

```
> ggplot( data=dat,
    aes(x=time, y=score,
          group=id) ) +
        geom_line() +
        geom_point()
```

This is a nice rectangle for our purposes

We now get a (two-point) line for person "B"



Multilevel models help with missing data

One powerful thing about multilevel models is that they allow you to work with respondents for whom some data are missing.

In the prior plot, we do have a sense of what the trajectory with the missing data point is like. When we model, we dynamically model each group with the data we have.

A Simple Model for Growth

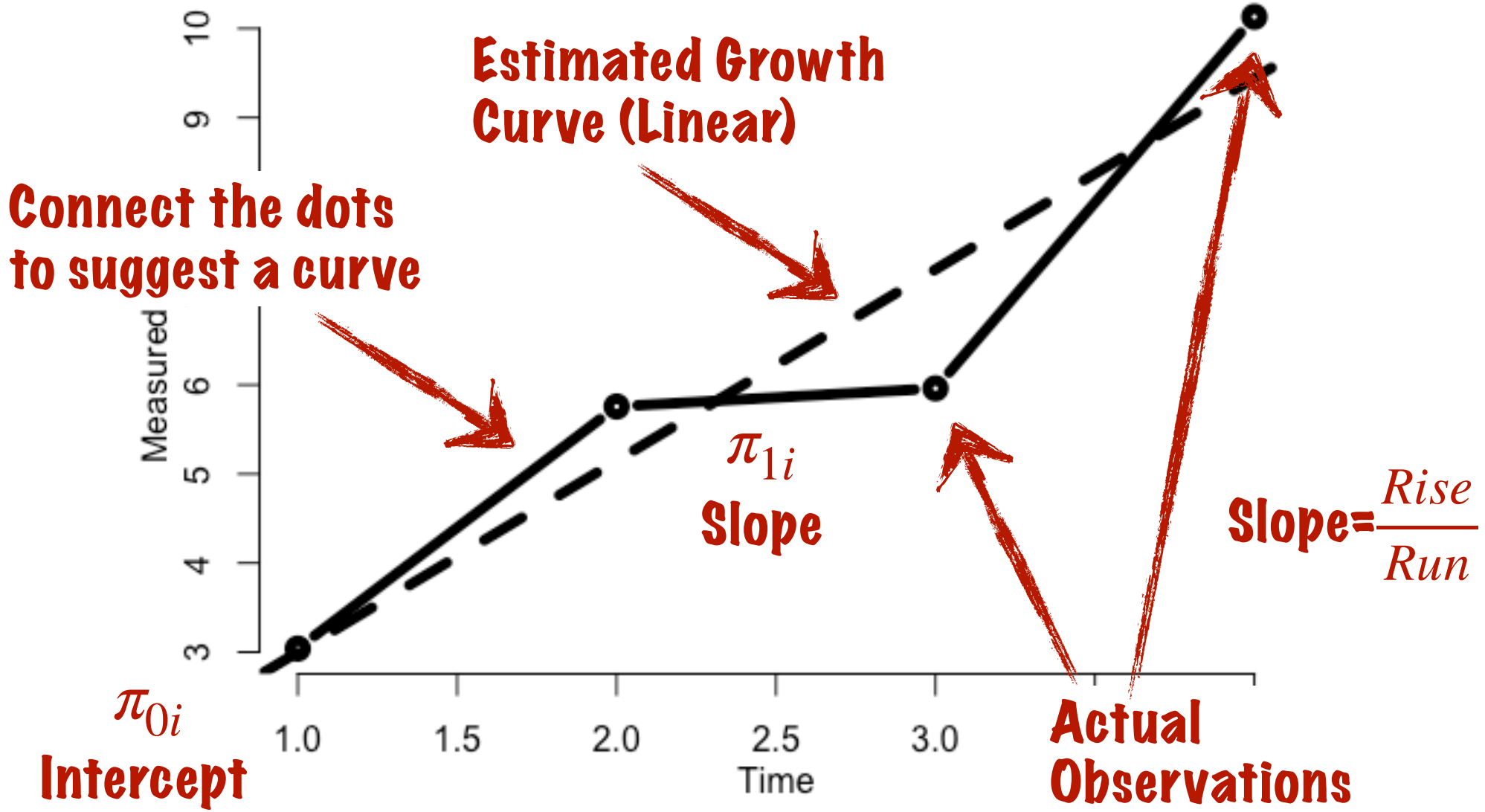
Example from Chapter 6 of R&B

Natural Sciences Knowledge (HeadStart)

Data described on R&B pg 164

- ★ 143 children enrolled in Head Start
- ★ Outcome is an IRT scaling of a set of test items, a "logit metric" (and so is a measure of student knowledge)
- ★ Data are lost; using tables and discussion from R&B chapter 6
- ★ Each child (supposed to be) tested 4 times along year
- ★ Age a_{ti} defined as *time from first data-collection point* (in months)
- ★ Also collected home language (Spanish or English) and number of hours of direct instruction

A Hypothetical Child



The meaning of our parameters

Intercept: this will be the predicted/expected outcome for everyone at a specific point in time.

- ★ Be sure to *choose* where that time is by centering your time variable appropriately.

Slope: this is the "rate of growth" or how much the outcome changes per unit in time.

- ★ Slope = Rise / Run (old Algebra!)
- ★ Slope = derivative of your line (Calculus!)
- ★ The linear model assumes that, while some people are faster growers than others, each person grows at a constant rate

An Unconditional Linear Growth Model

Notation from R&B pg 163-164

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + e_{ti}$$

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

$$\pi_{0i} = \beta_{00} + r_{0i}$$

$$\pi_{1i} = \beta_{10} + r_{1i}$$

$$(r_{0i}, r_{1i}) \sim N(\mathbf{0}, \Sigma)$$

Our Covariance Matrix

$$\Sigma = \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{10} & \tau_{11} \end{bmatrix}$$

We use Sigma (Σ) to denote the whole matrix.

★ a_{ti}, *age*, is the actual time of the t^{th} observation of child *i* (this allows for different measurement occasions).

★ The above is a classic random-slope model, nothing more, nothing less.

★ a_{ti} in this case is "time elapsed from first measure", which gives a specific interpretation to the intercept.

Notation Alert!

- ★ Raudenbush and Bryk *kept* the betas for individual students
- ★ So now the betas (ßs) are parameters at *level 2*.
- ★ We have “tucked in” a level underneath, for time.
- ★ These are the pi’s (πs)

- ★ In a later unit we will get our gammas back… on level 3!

Unconditional Model Results

Fixed Effect		*Coefficient*	*se*	*t Ratio*
Mean initial status, β_{00}	$\hat{\beta}_{00}$	−0.135	0.005	−27.00
Mean growth rate, β_{10}	$\hat{\beta}_{11}$	0.182	0.025	7.27

Wald tests, clearly significant

Random Effect		*Variance Compenent*	*df*	χ^2	*p Value*
Initial status, r_{0i}	$\sqrt{\hat{\tau}_{00}} = 1.30$	1.689	139	356.90	<0.001
Growth rate, r_{1i}	$\sqrt{\hat{\tau}_{11}} = 0.20$	0.041	139	724.91	<0.001
Level-1 error, e_{ti}	$\hat{\sigma} = 0.65$	0.419			

They use a chi-squared test. I've taught you LR tests. They are going to be about the same.

Reliability of OLS Regression Coefficient Estimate	
Initial status, π_{0i}	0.854
Growth rate, π_{1i}	0.799

We also have: $\rho = \dfrac{\tau_{01}}{\sqrt{\tau_{00}\tau_{11}}} = -0.278$

We see there is significant variation in initial knowledge and rate of growth.

R&B pg 165

Interpreting our results

★ Mean Intercept $\hat{\beta}_{00}$ is -0.135:

- At initial measure, average knowledge is -0.135.

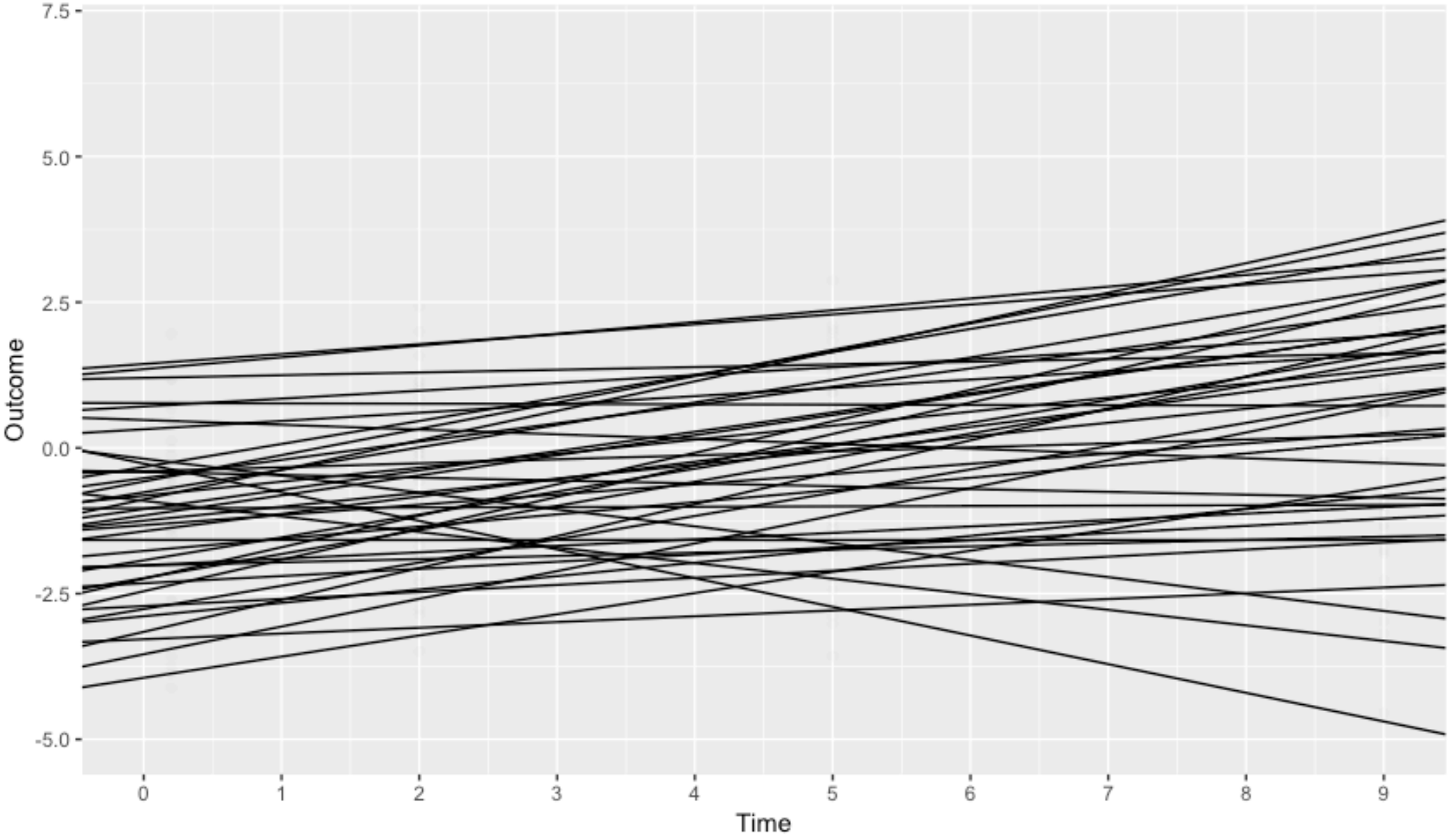
★ Mean Slope $\hat{\beta}_{10}$ is 0.182:

- We expect a median kid to grow about 0.182 units per month.

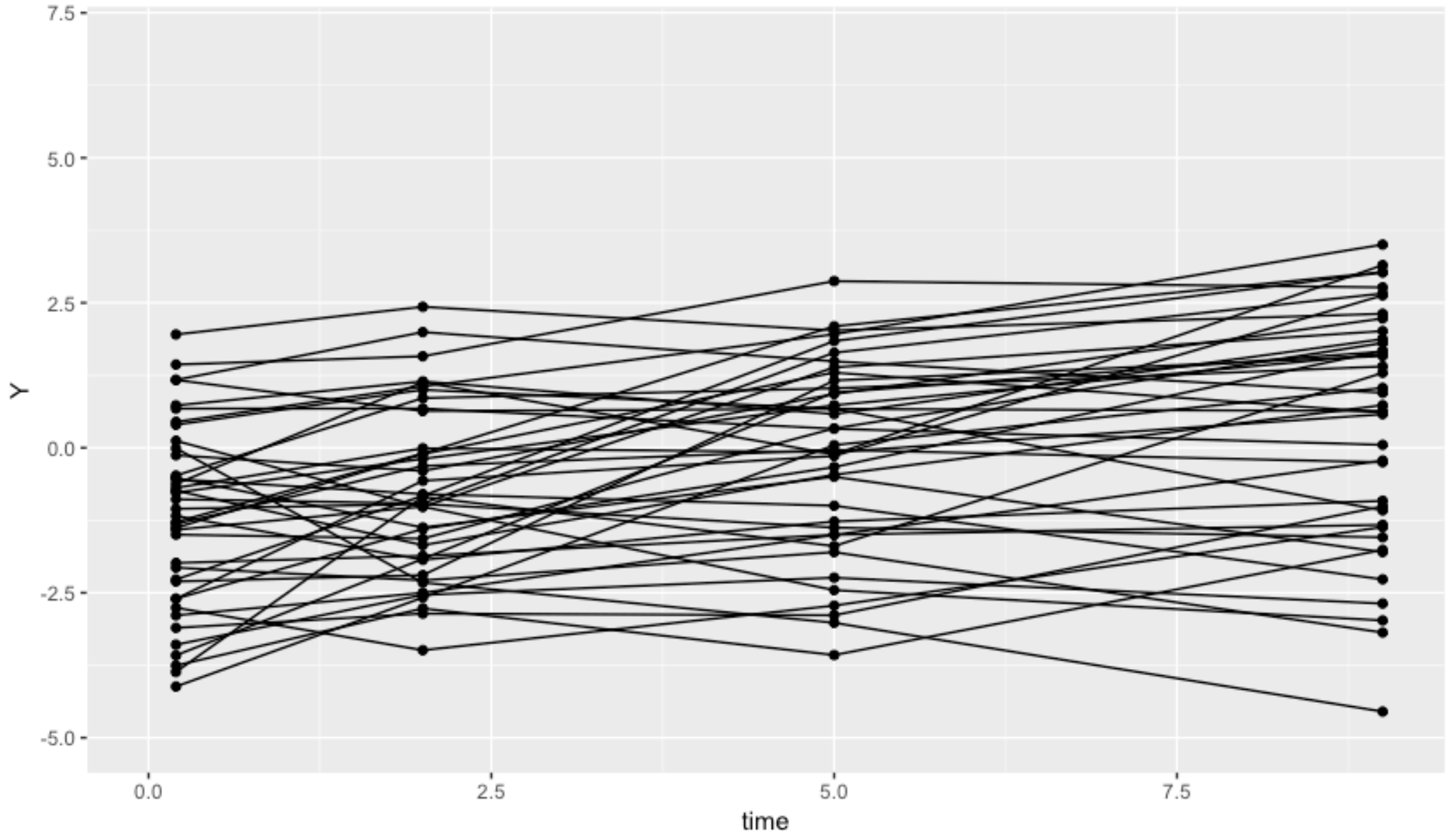
★ SD of random intercept is $\sqrt{\hat{\tau}_{00}} = 0.202.$

- A child one SD above average will grow at 0.182 + 0.202 = 0.384 logits/month
- A child one SD below average will decline (according to the model)...
- A 95% ***prediction interval*** of growth rates is -0.20 to 0.59 units per month. So 95% of kids will have growth rates in this interval.

Simulated data for 40 students to illustrate model (plotting the lines defined by π_{0j} and π_{1j})



Simulated data to illustrate model (with individual time points)



Reliability

We can break down the variance of the individual slope estimates into measurement error plus true variation:

$$Var(\hat{\pi}_{1i}) = Var(\hat{\pi}_{1i}|\pi_{1i}) + Var(\pi_{1i})$$

$$= v_{11i} + \tau_{11}$$

The v_{11i} are the Empirical Bayes SEs for the random slopes

The *reliability* is then how much noise there is in estimating our individual slope (compared to true variation)

$$\text{reliability}(\hat{\pi}_{1i}) = \frac{\tau_{pp}}{v_{ppi} + \tau_{pp}}$$

Estimate for all kids, and then average to get overall reliability.

The Correlation of Growth Rate with Initial Status

- ★ Frequently wish to know if initial status predicts/ is associated with subsequent rate of growth.
- ★ Simple pre-post scenarios FAIL at this (due to regression to the mean issues)
- ★ Multiple time points help with this.
- ★ With MLM, the correlation of random effects parameter specifically estimates this quantity of interest.
- ★ For our example we estimate -0.278 (see R&B pg 167.)

I believe there is a typo in the text
Equation 6.8 (pg 167) should not have "+" but instead "*"

Looking back

Lecture Recap

Check-In
http://cs179.org/lec31

★ We talked about what longitudinal data looks like, and how it often comes in "wide" form but we want it in "long" form

★ We talked about the simplest growth curve model, the linear growth model

★ We dove into a case study of children learning natural science

- We can talk about variation in child growth rates
- We can talk about how well we are estimating individual growth curves (reliability)
- Next class we will use covariates to *predict* different rates of growth

Appendix:
Model Convergence Issues

Uh-oh, my model didn't converge?

Sure, that happens

OLS doesn't have these issues (usually) because it only requires some matrix algebra

But even with OLS you can still run into issues (in theory) if variables are measured on very different scales

- These come down to how R represents matrices internally; the right level of precision for one column may not be right for other columns, especially after squaring.

I don't care about OLS. What about my model? It didn't converge!

Convergence failure happens quite often with MLE

Likelihood surfaces can be hard to explore

Lots of almost flat regions

Mathematical operations tend to be much more complex

Your algorithm can get tired

Especially an issue in longitudinal models where each subject only has one or two more measurements than coefficients

Things to try to get convergence

Rescale variables

Algorithms work better with variables measured to have mean 0, sd 1

Play around with the optimizer

One algorithm might succeed where another fails

Increase the number of iterations the algorithm will go through before it quits

Just fit your model overnight

Simplify your RQ/model

Fully pooling parameters and dropping correlations can help.
But before that, check your model to make sure it's doing what you want – be sure that the fixed and random parts are both correct.

Go fully Bayesian

We may talk about this later.

Don't feel bad!

It happens to everyone!

Appendix:
reshape() and debugging

More flexible wrangling: reshape()

"reshaping" data to long format
(reshaping is generally VERY FRUSTRATING)

```
dat <- reshape(dta, direction = 'long',
       varying = list(c('wave1', 'wave2', 'wave3') ),
       times = c('w1', 'w2', 'w3'),
       v.names = 'score',
       timevar = 'wave',
       idvar = 'id')
```

Variable names in wide format

The values of the time variable

Variable names in the long format

This way is more powerful, but more difficult to use. See reshape example handout for more.

You can avoid using this in most cases.

Smashing bugs in your R code

Can’t get the reshape function to work? You’re not alone!

Start by simplifying the problem; e.g., only convert one variable from wide to long to start.

Practice on a tiny dataset with only 3 people or so.

If some aspect of reshape is giving you trouble, consider trying to fix it after reshaping in an ad-hoc way, e.g., merge your student race variable in after reshaping.