Unit 3, Lecture 4
Piecewise Linear Growth Models

Calling out some additional reading

Faraway, Chapter 9: Some case studies of longitudinal data, plus some model checking (see later slides)

Rabe-Hesketh & Skrondal text:

- ★ pp227-264 : discussion of the structure of longitudinal data, some common vocabulary for it.
- ★ pp278-282 : why MLM handles missing data nicely (i.e., if we are missing some observations for some people)
- ★ pp293-311 (chap 6): Marginal models, and error structure (with some technical detail)

Today's Goals

Review the idea of a *parametric growth curve*

- ★ Each student gets their own curve, but these curves are all defined by a small set of *parameters*.

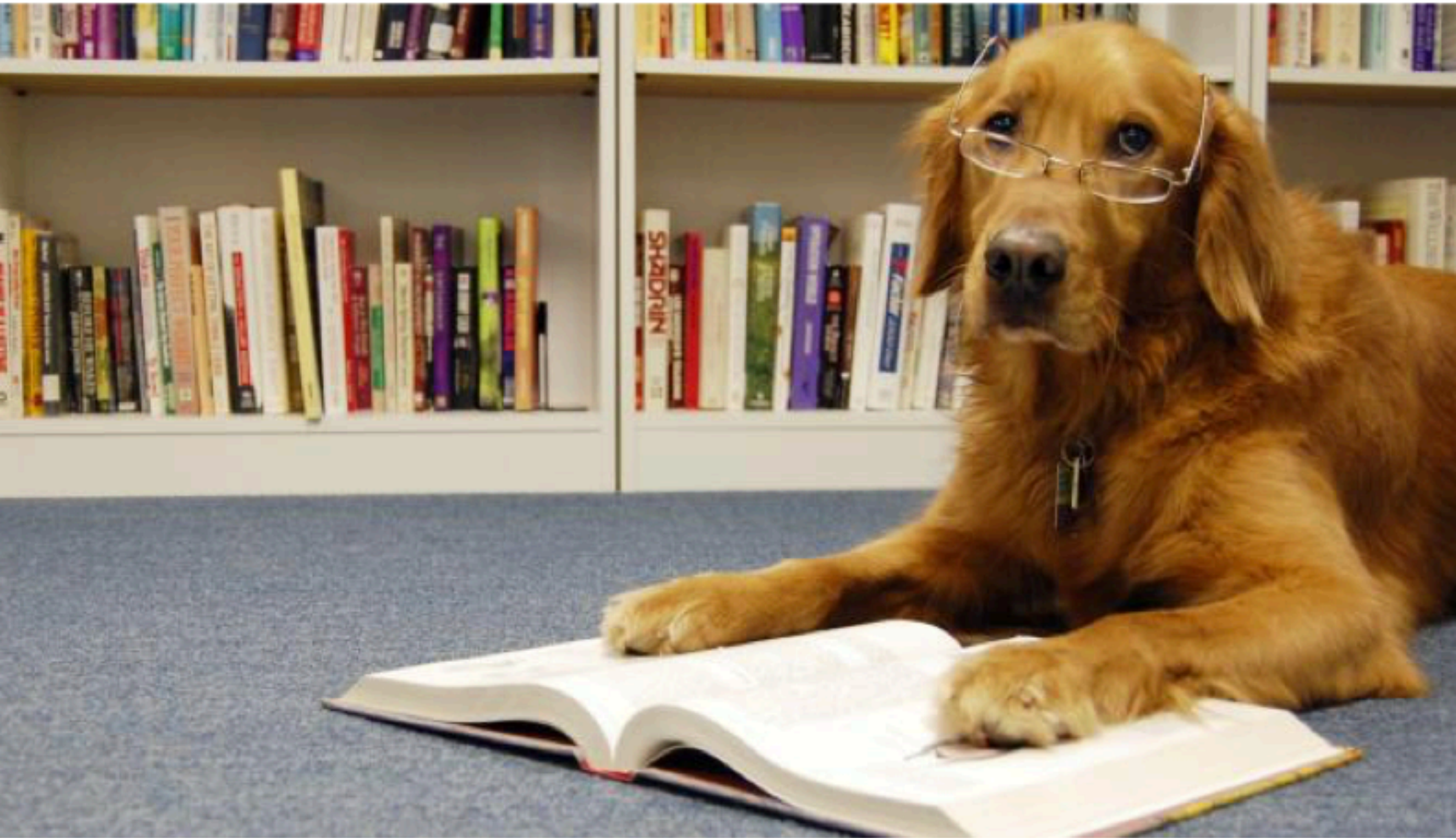
Introduce the idea of a *piecewise linear growth model*

- ★ *People will grow at different rates in different circumstances. We can use a linear model for each type of circumstance.*

Model comparison with AIC

- ★ A very simple way to statistically select from some competing models.

Our dataset and question today



Reading ability (scale score, ITBS)

- ★ Over 3,000 students from schools in North Carolina; we're only looking at students who were in 3rd grade at the start of the study.
- ★ Outcome is a scale score computed from responses to items on the ITBS test
- ★ Each child tested 4 times: spring of 2014, fall of 2014, spring of 2015, and fall of 2015.
- ★ Time defined as months from first data-collection point (roughly)

Motivating Research Question

How quickly does student reading comprehension grow over time?

- ★ time invariant stuff could predict reading ability and how reading changes over time (maybe girls enter school more advanced than boys; maybe LEP students grow more quickly during the school year than their peers)
- ★ time varying stuff could explain reading in a given period (maybe kids who have a high sense of belonging at school in a given period have higher scores in that period)

Goal for right now

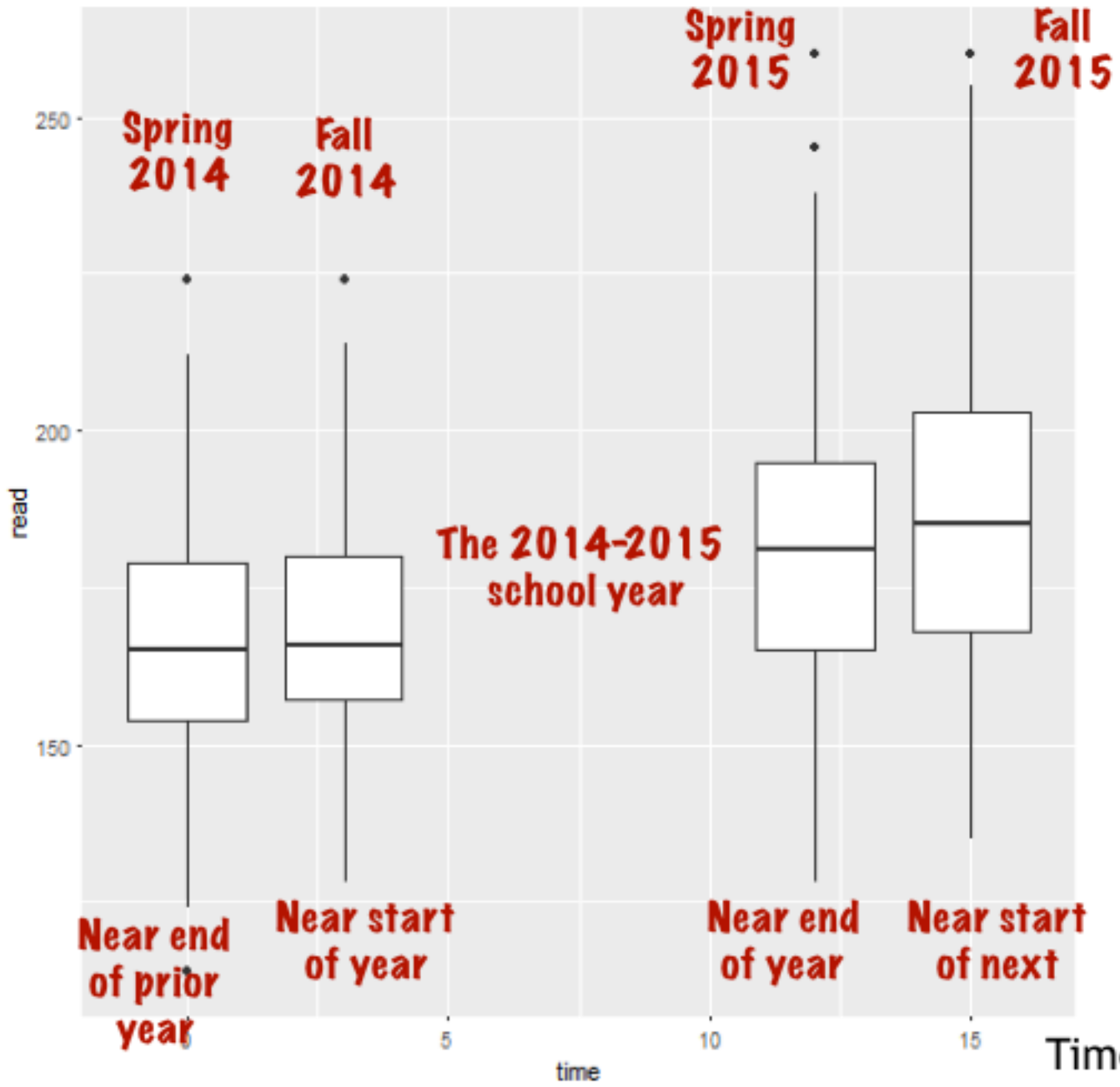
Want to decide on the general shape of the growth curve.

Each student's growth curve describes the general trajectory of her or his growth over this period of time.

We need to specify a model which can represent the shape of each individual student's growth curve.

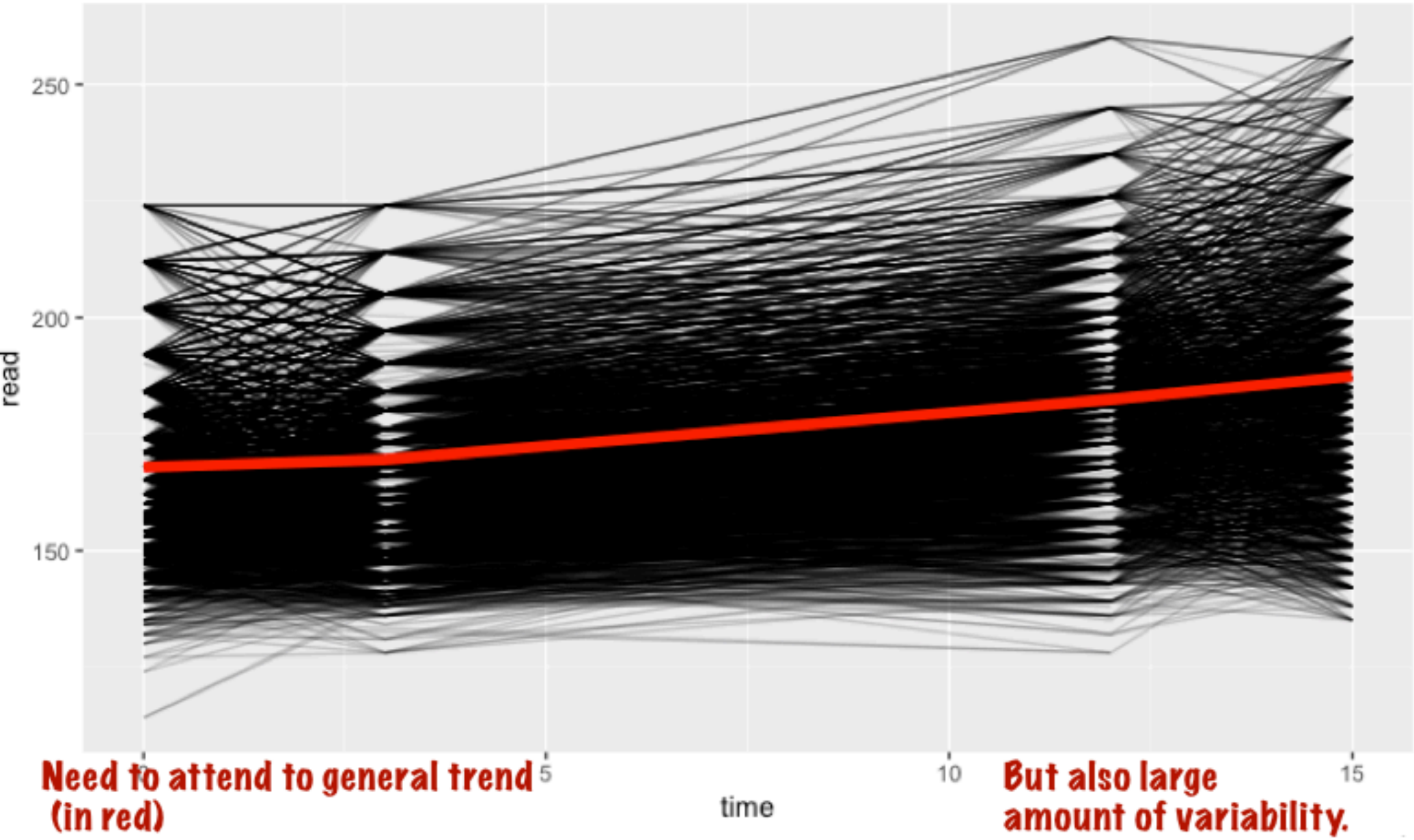
This requires us to look at both the average growth over the whole sample, but also individual curves.

Distribution by wave of data collection



Time in months

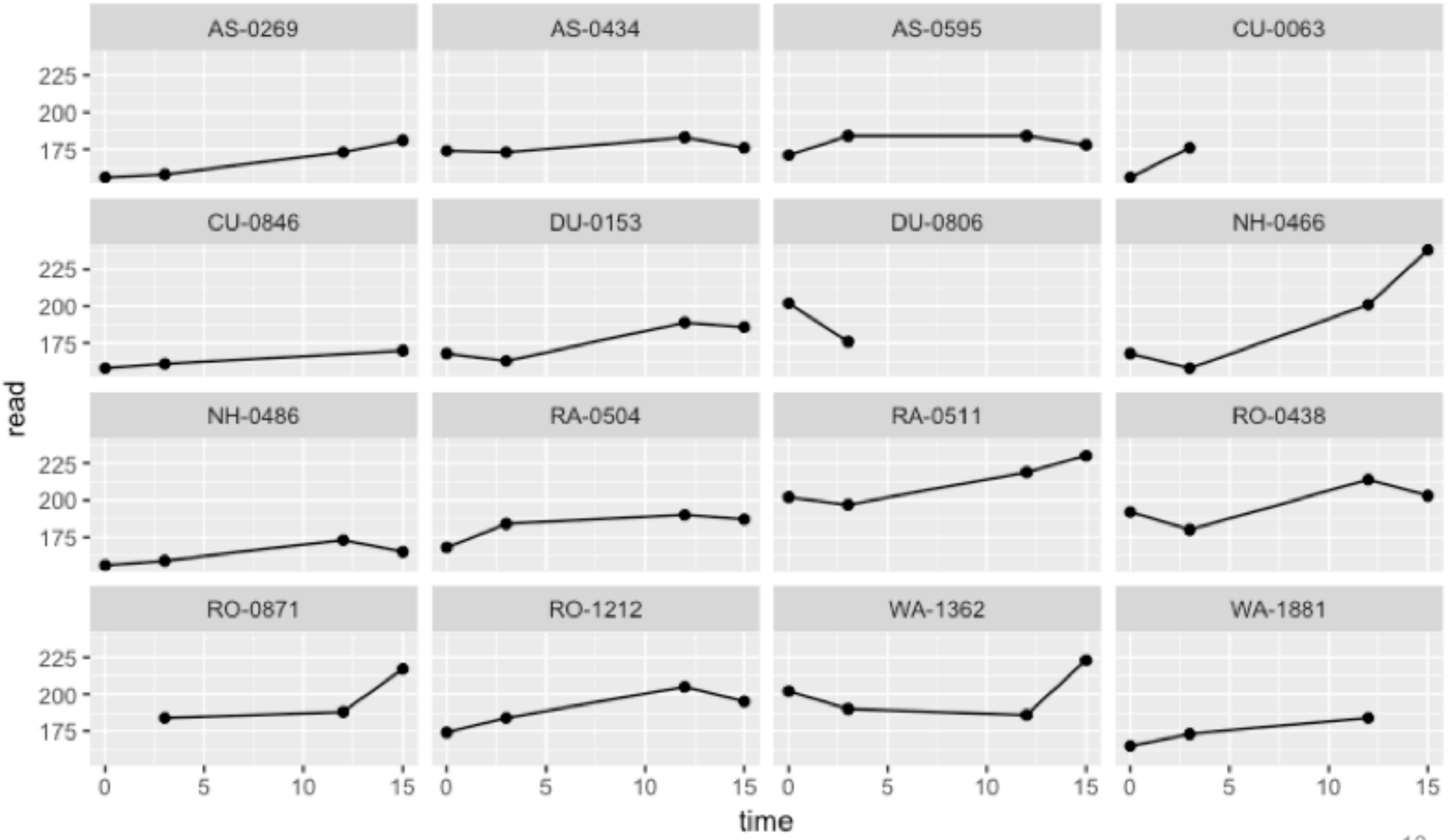
Growth over time (the raw data)



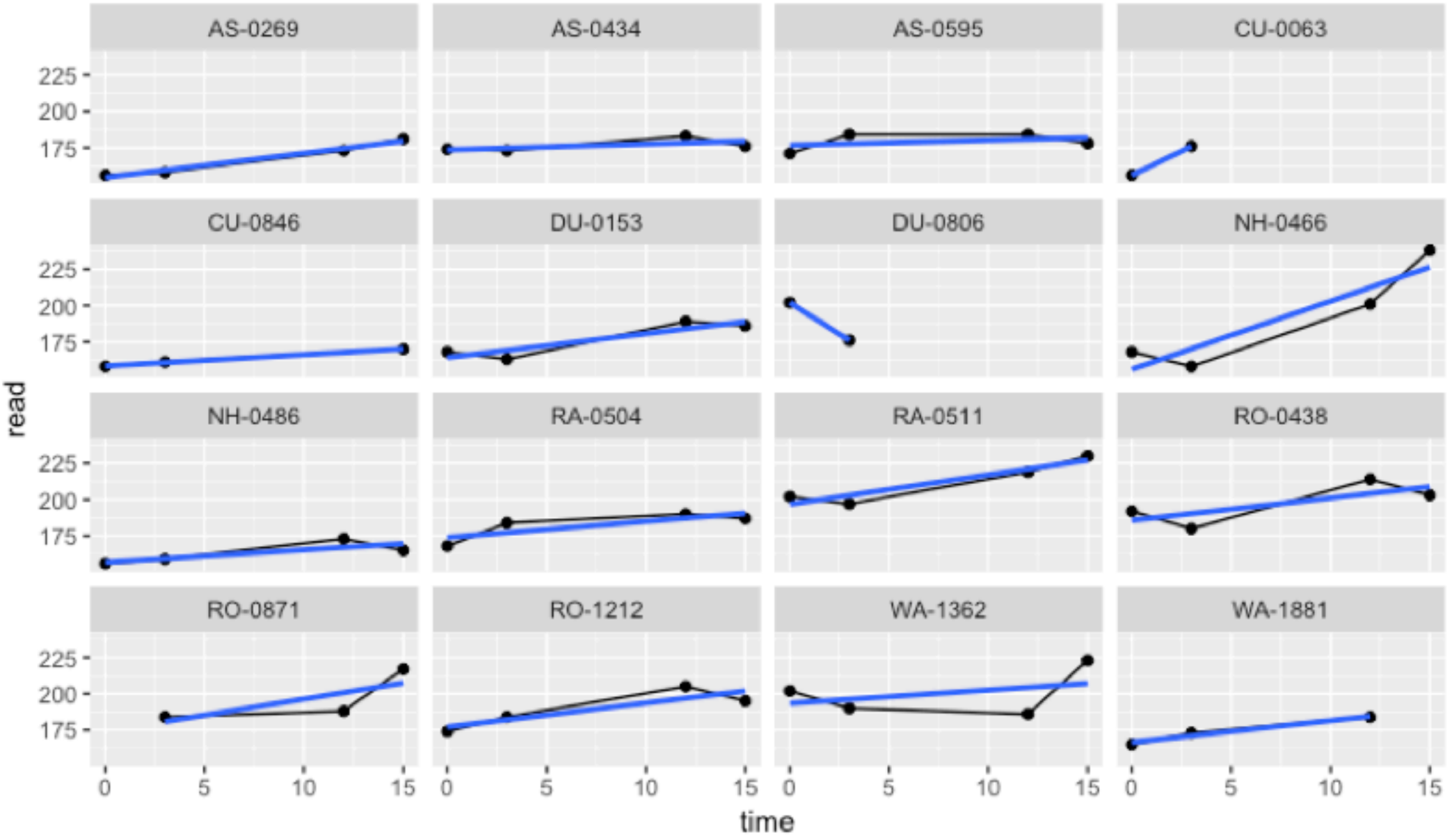
Need to attend to general trend (in red)

But also large amount of variability.

16 random children



Linear trajectories for these children



Unconditional Model Results

```
> mod <- lmer(read ~ time + (time|id),
data = dat)
> screenreg(mod)

========================================
                          Model 1
----------------------------------------
(Intercept)                166.61 ***
                            (0.32)
time                          1.32 ***
                            (0.02)
----------------------------------------
AIC                       88440.52
BIC                       88484.21
Log Likelihood           -44214.26
Num. obs.                 10730
Num. groups: id            3061
Var: id (Intercept)         251.65
Var: id time                  0.49
Cov: id (Intercept) time      6.09
Var: Residual                98.87
========================================
*** p < 0.001, ** p < 0.01, * p < 0.05
```

Let's look at a simple linear growth model to get a baseline of variability, etc.

screenreg() from texregp package gives this printout.

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

Students who start higher grow faster

Interpreting our linear growth results

★Mean Intercept is 166 – this is the average starting position

★Mean Slope is 1.32: We expect an average kid to grow about 1.32 points per month.

★sd of intercept around 15 points, so wide spread in initial ability.

★sd of slope around 0.7, so fairly substantial spread in growth rates; some students are predicted to lose ability over time

Note that $1.32 \pm 2 \cdot 0.7 = 1.32 \pm 1.4 = (-0.08, 2.72)$

A model incorporating predictors

	Model 1
(Intercept)	168.35 ***
	(0.45)
time	1.37 ***
	(0.03)
gendermale	-3.56 ***
	(0.64)
time:gendermale	-0.09 *
	(0.04)
AIC	88413.35
BIC	88471.60
Log Likelihood	-44198.68
Num. obs.	10730
Num. groups: id	3061
Var: id (Intercept)	248.57
Var: id time	0.49
Cov: id (Intercept) time	6.00
Var: Residual	98.86

*** p < 0.001, ** p < 0.01, * p < 0.05

These are now specific to girls

Boys start behind girls and grow more slowly

lmer code?

Comparing Models

Model	Initial status	Slope
Unconditional	251.65	0.49
Conditional on gender	248.57	0.49
Proportion of variance explained	.01	.00

- ★ 1 - Var(Cond)/Var(Uncond) = Proportion explained
- ★ We see that our two variables explain almost nothing about differences in initial status or rates of growth; there's a lot still on the table!

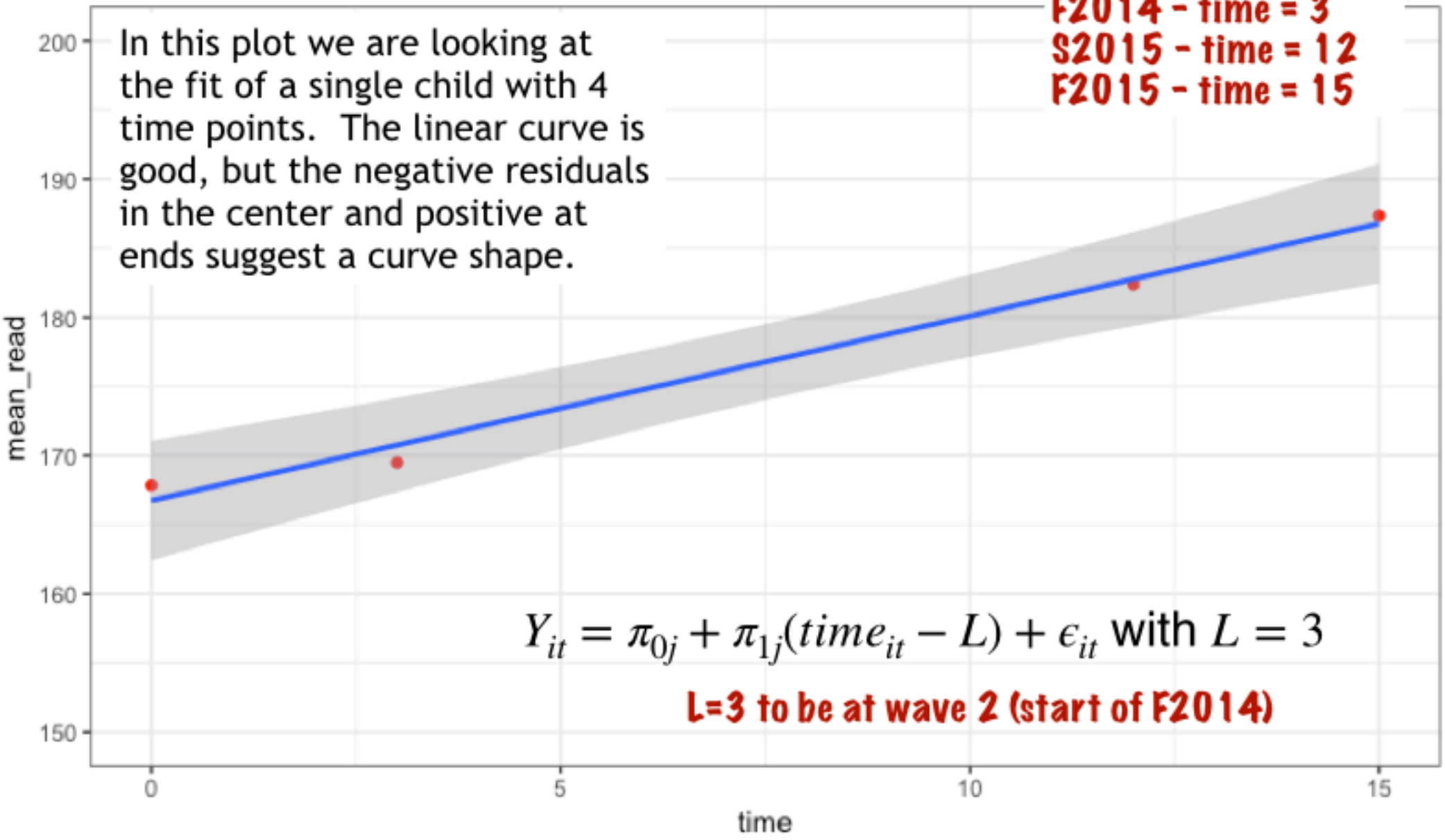
Recap:
Parametric Growth Curves

General Growth Curve Models

- ★ Growth Curve models means you specify a general form for each individual
  - E.g., a line
- ★ These forms are *parameterized* or *indexed* by a set of parameters
  - E.g., slope and intercept
- ★ We assume each individual has a particular member of this general form
  - E.g., a specific line with a specific slope and intercept
- ★ We finally put a distribution on the parameters to describe the population of individuals
- ★ We can fit any growth curve model that we can parameterize.

A linear curve

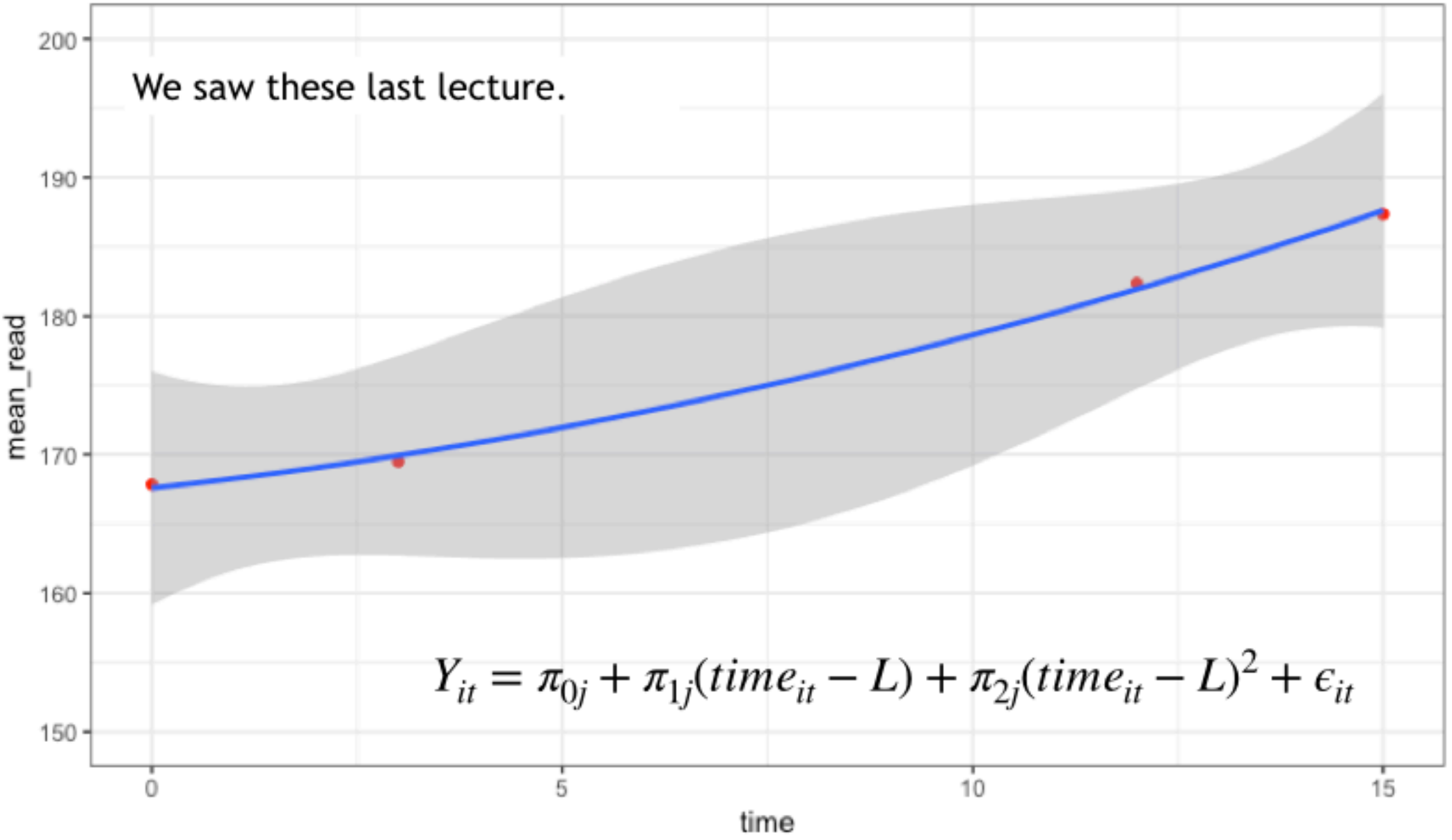
Recall we measure at
S2014 - time = 0
F2014 - time = 3
S2015 - time = 12
F2015 - time = 15



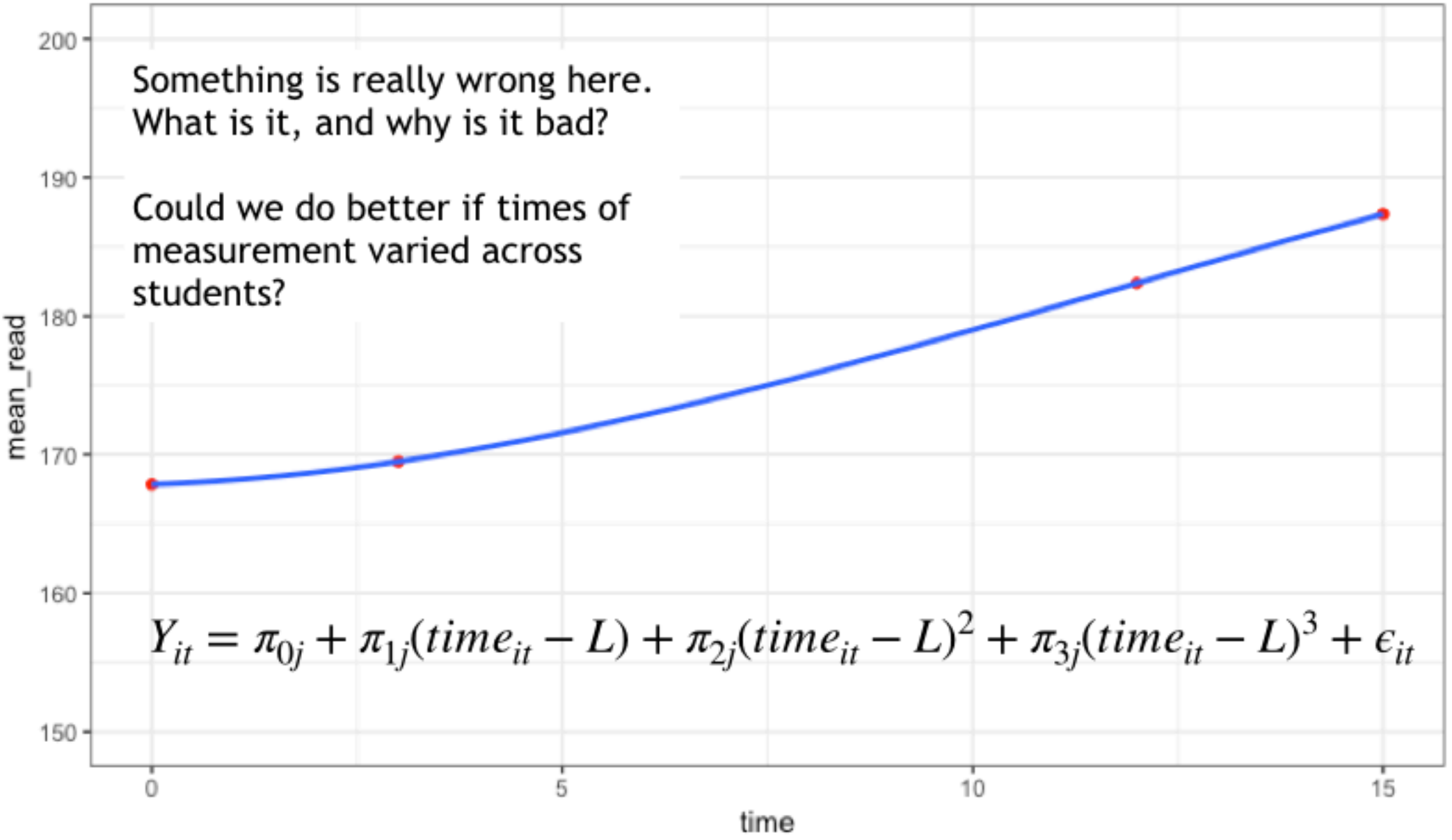
$$Y_{it} = \pi_{0j} + \pi_{1j}(time_{it} - L) + \epsilon_{it} \text{ with } L = 3$$

L=3 to be at wave 2 (start of F2014)

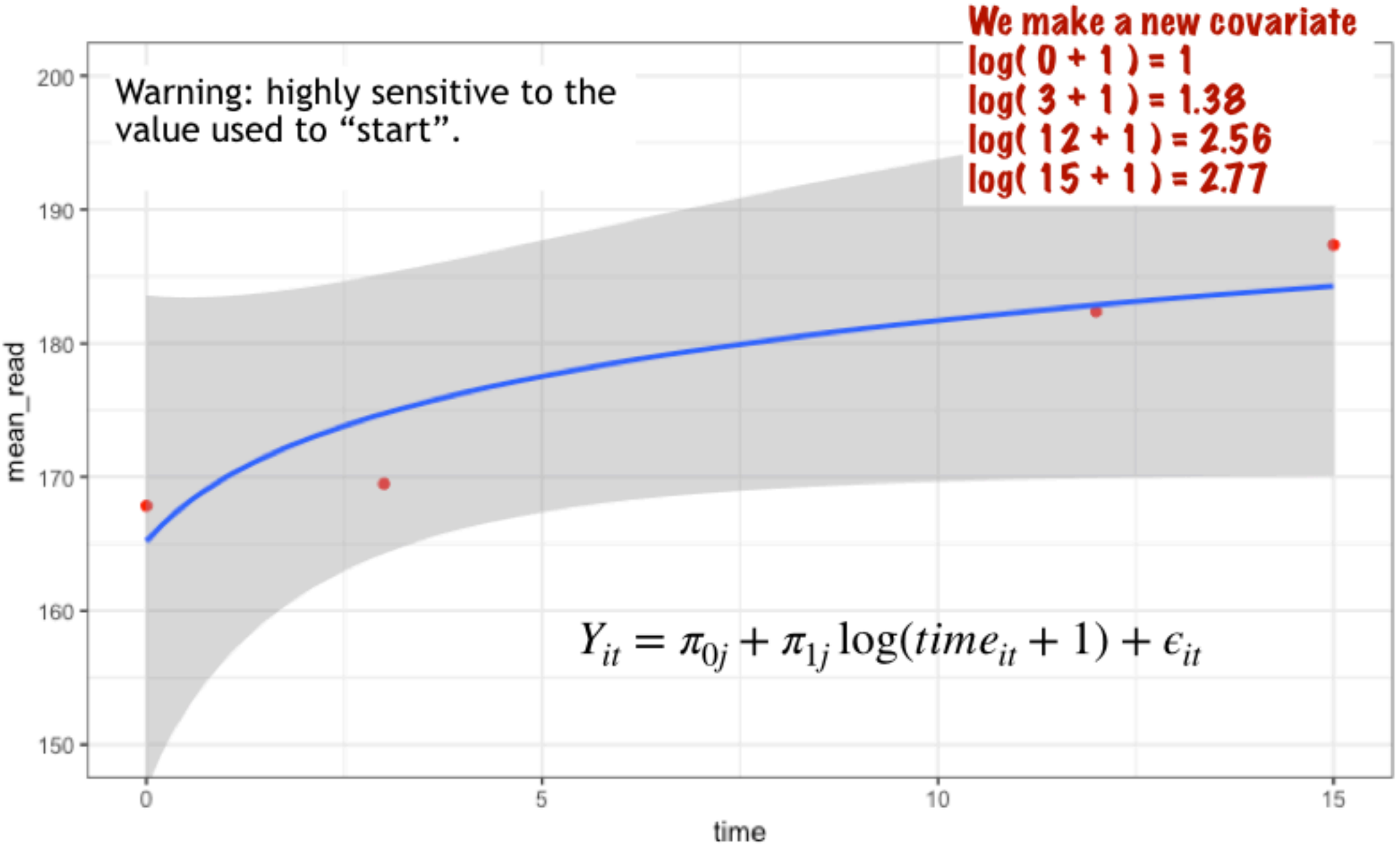
A quadratic curve



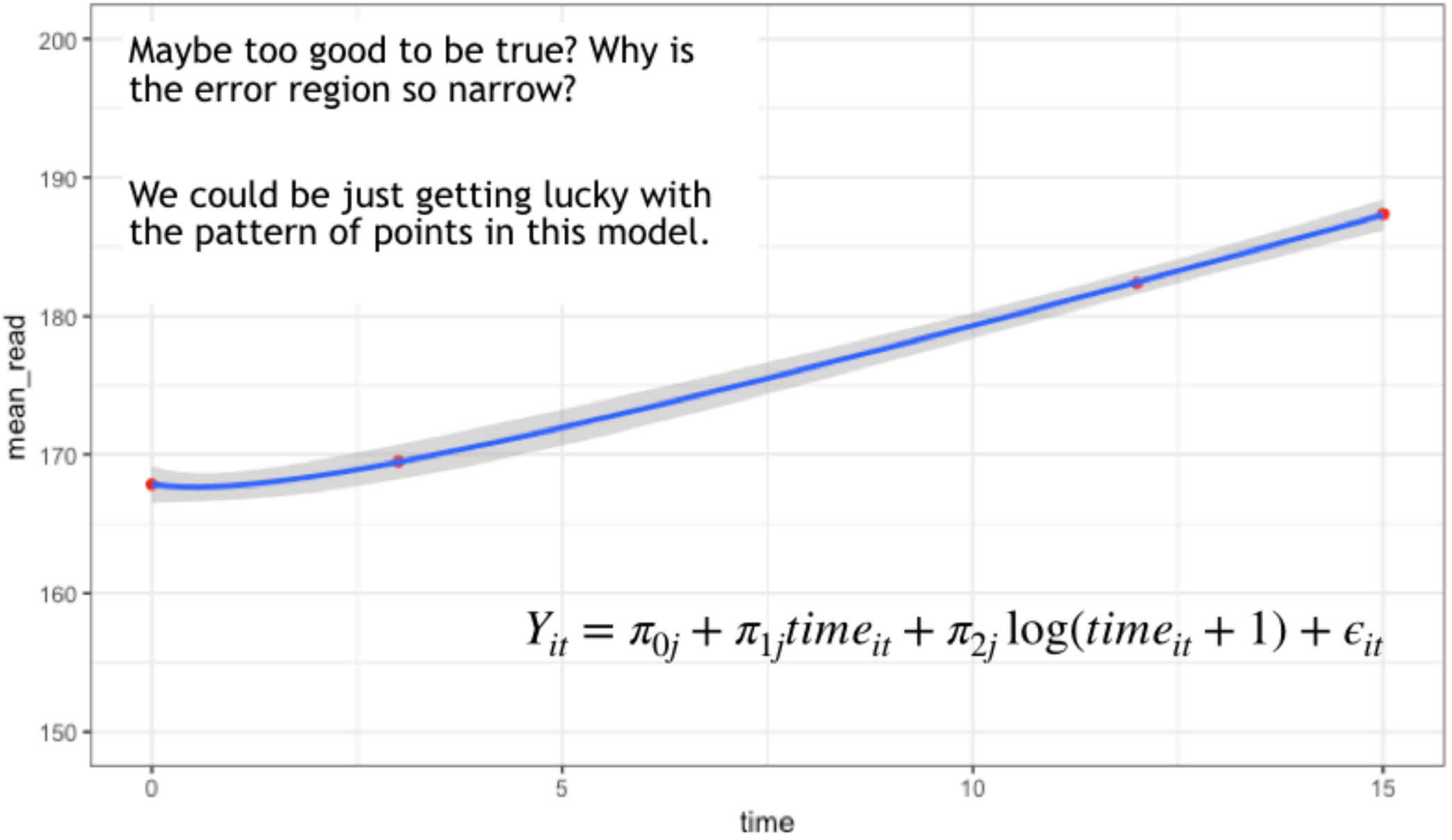
A cubic?


Something is really wrong here.
What is it, and why is it bad?
Could we do better if times of measurement varied across students?
$Y_{it} = \pi_{0j} + \pi_{1j}(time_{it} - L) + \pi_{2j}(time_{it} - L)^2 + \pi_{3j}(time_{it} - L)^3 + \epsilon_{it}$
mean_read
200
190
180
170
160
150
0
5
10
15
time

A logarithm


We make a new covariate
log(0 + 1) = 1
log(3 + 1) = 1.38
log(12 + 1) = 2.56
log(15 + 1) = 2.77
Warning: highly sensitive to the value used to "start".
$Y_{it} = \pi_{0j} + \pi_{1j} \log(time_{it} + 1) + \epsilon_{it}$
mean_read
200
190
180
170
160
150
0
5
10
15
time

Something weird: x + log(x)



Selecting a growth curve

- If possible, use theory/the research question
- Look for a curve that works overall, but also within each respondent (AIC can help here)
- Use a representation that can be interpreted in a meaningful way
- Need fewer coefficients (including the intercept) than observations
- In general, simpler is better!
- Remember that a model is *always* a simplification, but it will be useful if you simplify in the right way
- Don’t use complicated models to make predictions about what’s happening between waves

The fitted models for our different curve ideas

	Model 1	Model 2	Model 3	Model 4
(Intercept)	170.57 ***	169.65 ***	165.22 ***	173.20 ***
	(0.33)	(0.34)	(0.33)	(0.44)
time.c	1.32 ***	0.88 ***		1.83 ***
	(0.02)	(0.05)		(0.06)
time_sq		0.05 ***		
		(0.01)		
log_time			6.64 ***	-2.91 ***
			(0.11)	(0.29)
AIC	88434.22	88256.47	89600.59	88304.01
BIC	88477.90	88329.28	89644.27	88376.82
Log Likelihood	-44211.11	-44118.23	-44794.29	-44142.00
Num. obs.	10730	10730	10730	10730

Line

Quadradic (had convergence issues)

log(time+1)

linear + log(time+1)

Comparisons (via AIC scores)

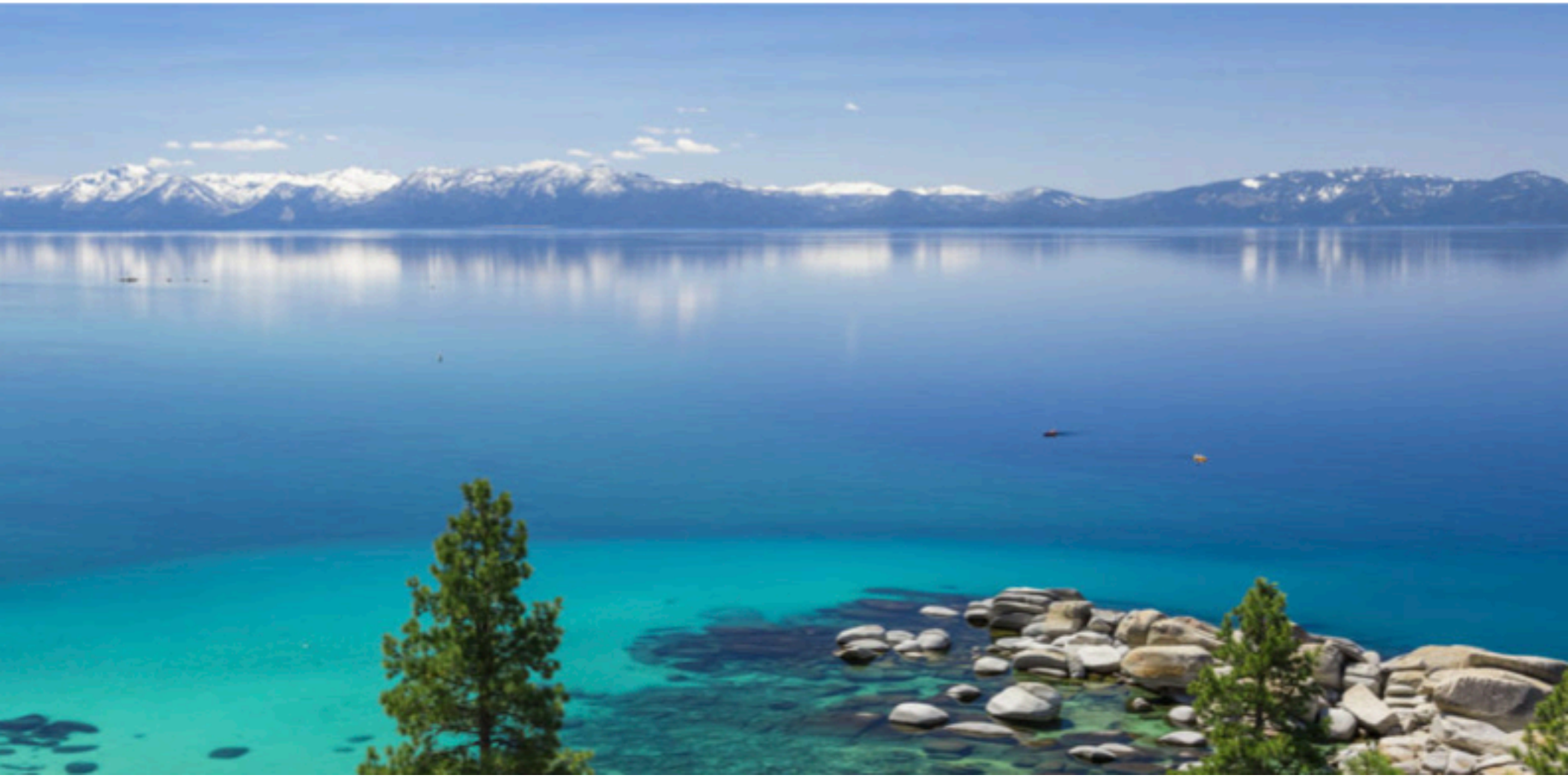
```
> m1 <- lmer(read ~ time + (time|id),
                    data = dat, REML = FALSE)
> m2 <- lmer(read ~ time + time_sq + (time + time_sq|id),
                    data = dat, REML = FALSE)
> m3 <- lmer(read ~ log(time + 1) + (log(time + 1)|id),
                    data = dat, REML = FALSE)
> m4 <- lmer(read ~ time + log(time + 1) + (time + log(time + 1)|id),
                    data = dat, REML = FALSE)
```

```
> AIC(m1, m2, m3, m4)
    df       AIC
m1   6 88434.22
m2  10 88256.47
m3   6 88770.36 winner
m4  10 88280.01
```

We will more formally do AIC and model selection in a later lecture.

The take-away here is that different models are possible, all are approximate, and you want something that (1) fits reasonably and (2) is interpretable.

Piecewise Linear Growth



Looking at growth in summer vs school year

- ★ Want to know if student growth is faster over the summer or over school year
- ★ Research suggests reading growth is slower over the summer
- ★ A Piecewise Linear Growth Curve Model has a general form of:

$$reading_{it} = \pi_{0i} + \pi_{1i} TimeA_{it} + \pi_{2i} TimeB_{it} + \epsilon_{it}$$

- ★ We track time with two different variables, $time_A$ and $time_B$, but how should we measure them?

In class activity:
Write out the lmer code to fit this

Defining time: First we can examine cumulative school and cumulative summer

The original measurements were taken such that

- ★ at wave 1 (time = 0), no months of school or summer had elapsed (by construction)
- ★ at wave 2 (time = 4), two months of school and two months of summer had elapsed
- ★ at wave 3 (time = 12), ten months of school and two months of summer;
- ★ at wave 4 (time = 16), twelve months of school and four months of summer.

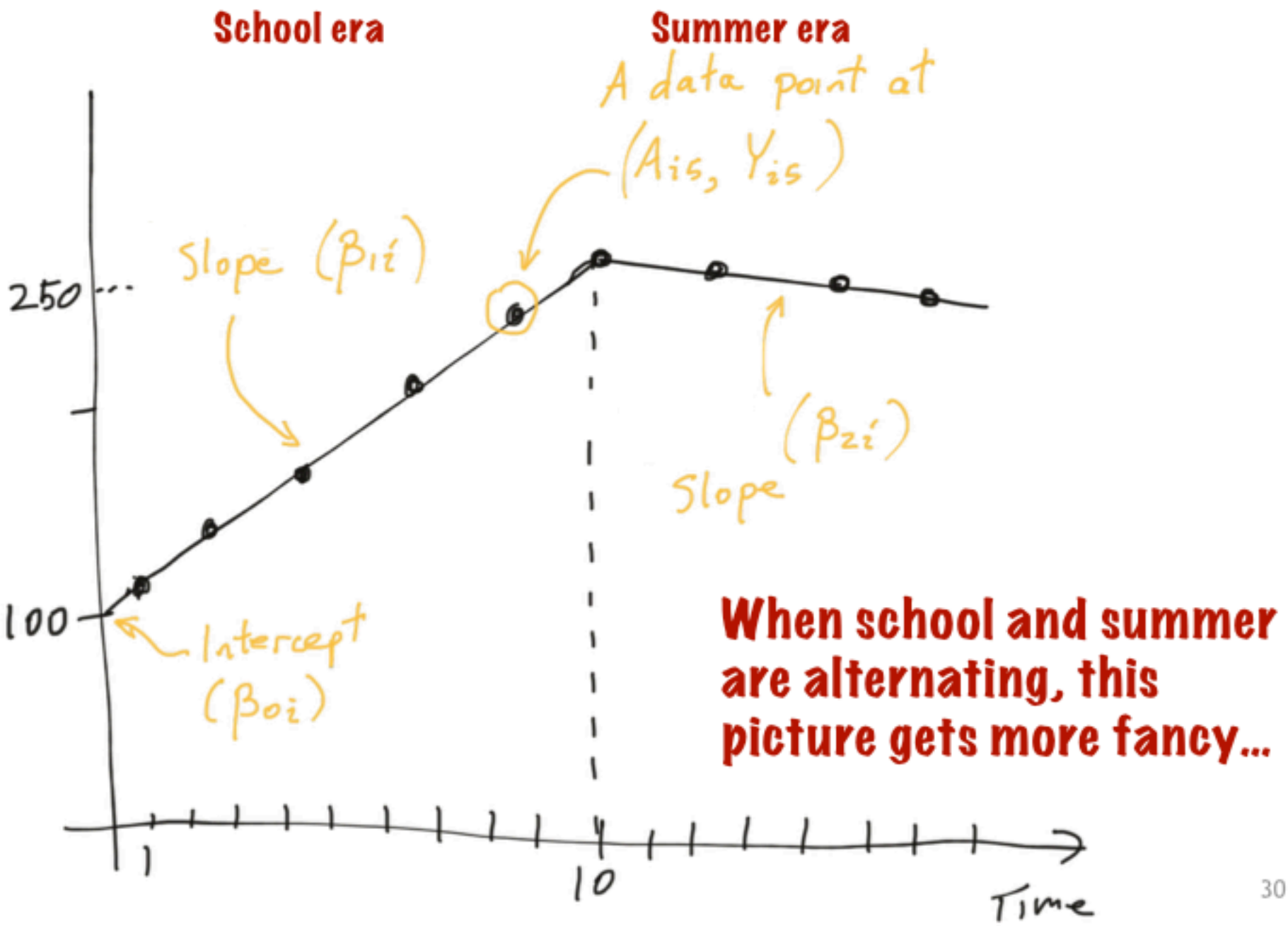
Option 1: Two different rates

★A_1 : number of **months of school** that have elapsed since the start of the study
(β_{1i} = school rate of growth for student i)

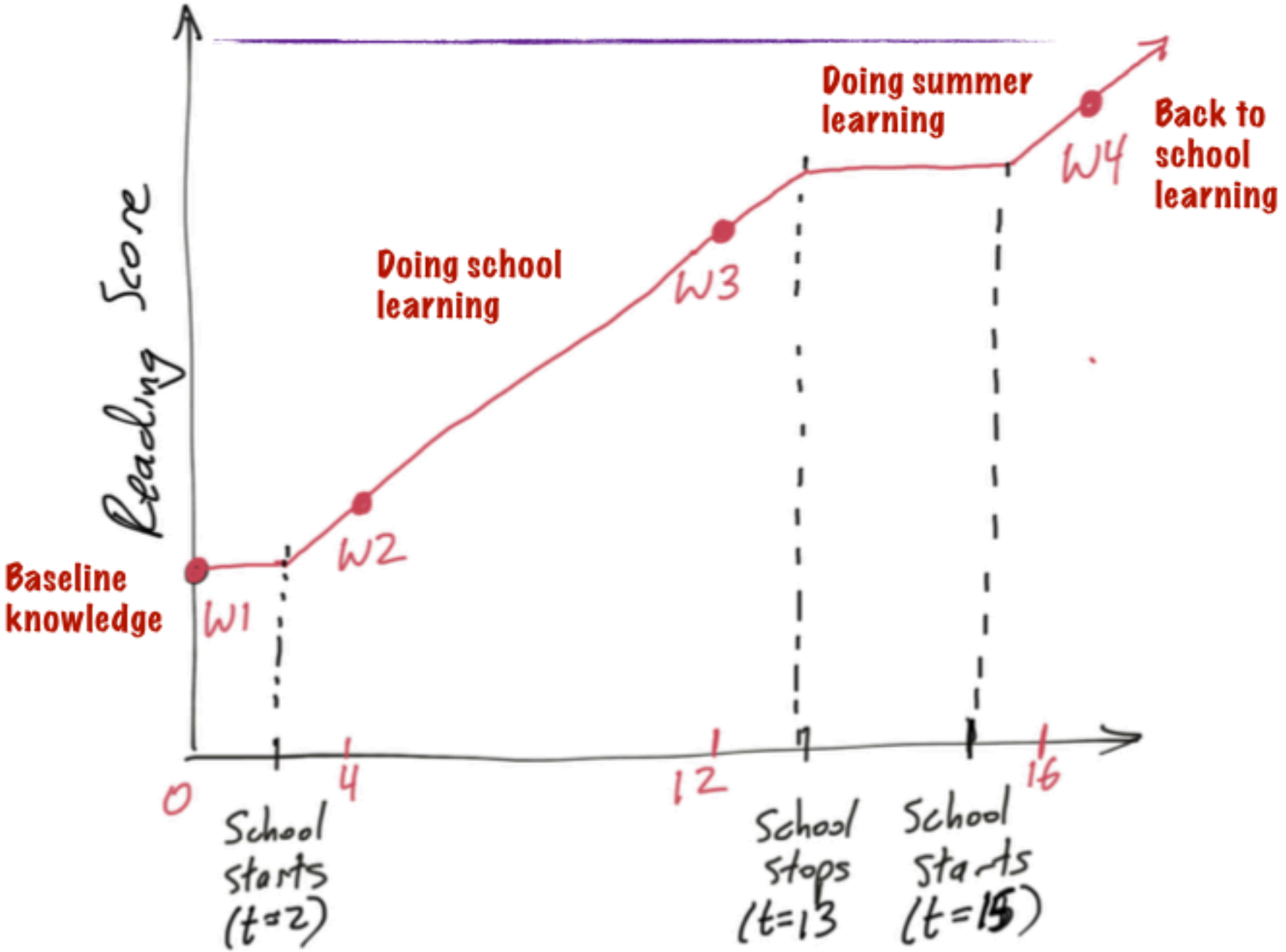
★B_2 : number of **months of summer** that have elapsed since the start of the study
(β_{2i} = summer rate of growth for student i)

★Corresponding model estimates each rate as distinct parameter

Anatomy of a piecewise linear model



Our school & summer growth model



Option 2: Increment/Decrement

★ A_1 : **total number of months** that have elapsed since the start of the study (so this is just our original time variable)

(β_{1i} = school rate of growth for student i)

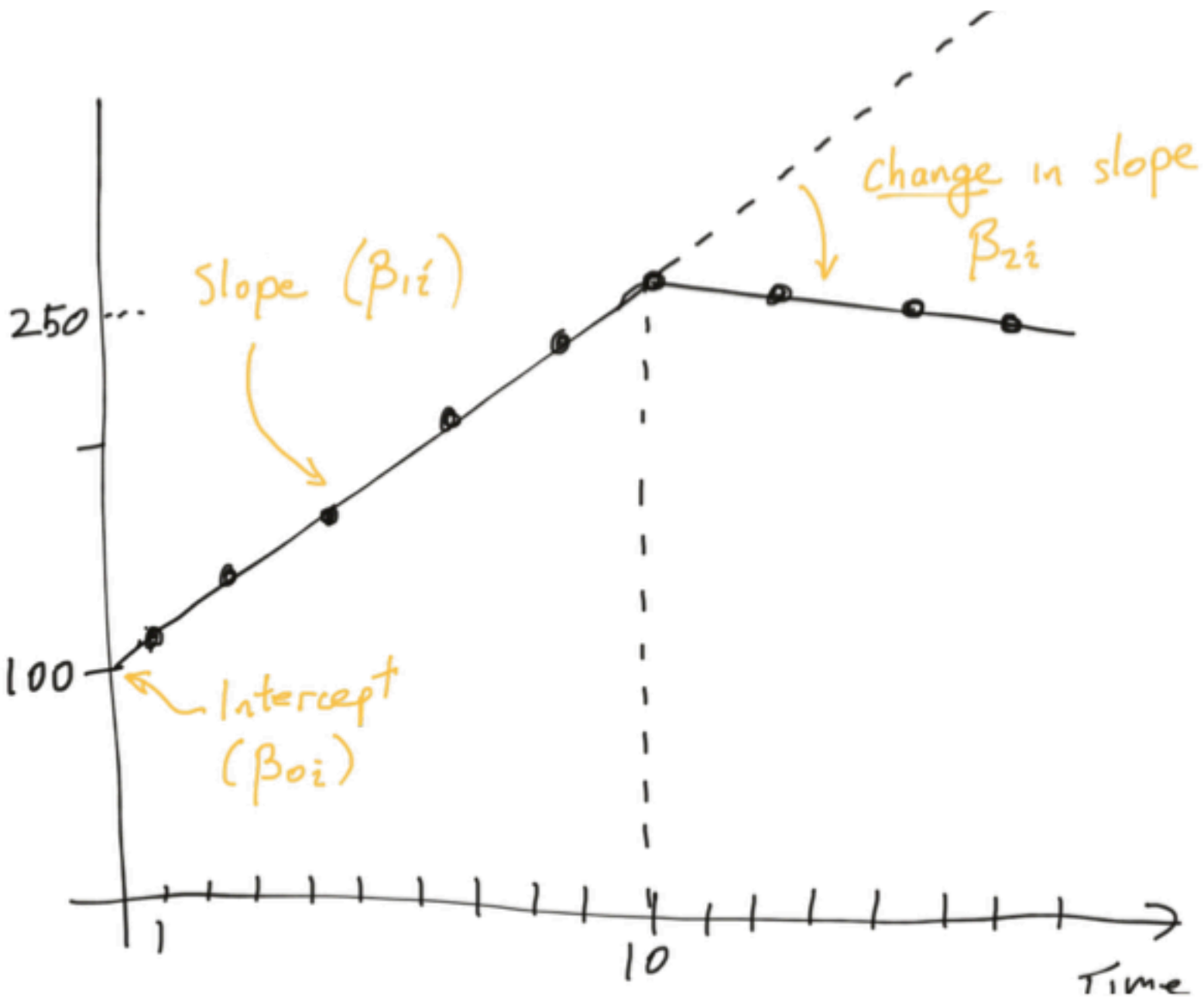
★ B_2 : number of **months of summer** that have elapsed since the start of the study

(β_{2i} = ***difference*** between summer and school rates of growth for student i)

★ Corresponding model has summer rate as a comparison to school rate (this is basically an interaction)

- ★ This model lets us ask "How is summer learning different from school learning?"

The increment/decrement view



The two models

$$reading_{it} = \pi_{0i} + \pi_{1i} School_{it} + \pi_{2i} Summer_{it} + \varepsilon_{it}$$

$$\pi_{0i} = \gamma_{00} + u_{0i}$$

$$\pi_{1i} = \gamma_{20} + u_{1i}$$

$$\pi_{2i} = \gamma_{20} + u_{2i}$$

We will directly estimate growth during the school and during the summer.

$$reading_{it} = \pi_{0i} + \pi_{1i} Time_{it} + \pi_{2i} Summer_{it} + \varepsilon_{it}$$

$$\pi_{0i} = \gamma_{00} + u_{0i}$$

$$\pi_{1i} = \gamma_{20} + u_{1i}$$

$$\pi_{2i} = \gamma_{20} + u_{2i}$$

We will directly estimate overall average growth, plus the additional growth during summer that is added to baseline growth.

Comparing those models

	Model 1	Model 2
(Intercept)	166.96 ***	166.96 ***
	(0.33)	(0.33)
school	1.45 ***	
	(0.03)	
summer	0.78 ***	-0.67 ***
	(0.11)	(0.13)
time		1.45 ***
		(0.03)
AIC	88390.91	88390.91
Var: id (Intercept)	249.11	249.11
Var: id school/time	0.25	0.25
Var: id summer	2.77	1.46
Cov: id (Intercept) school/time	4.84	4.84
Cov: id (Intercept) summer	8.33	3.49
Cov: id school/time summer	0.78	0.53
Var: Residual	96.70	96.70

These are the three gamma terms (note that the summer's gamma (γ_{20}) changes depending on model).

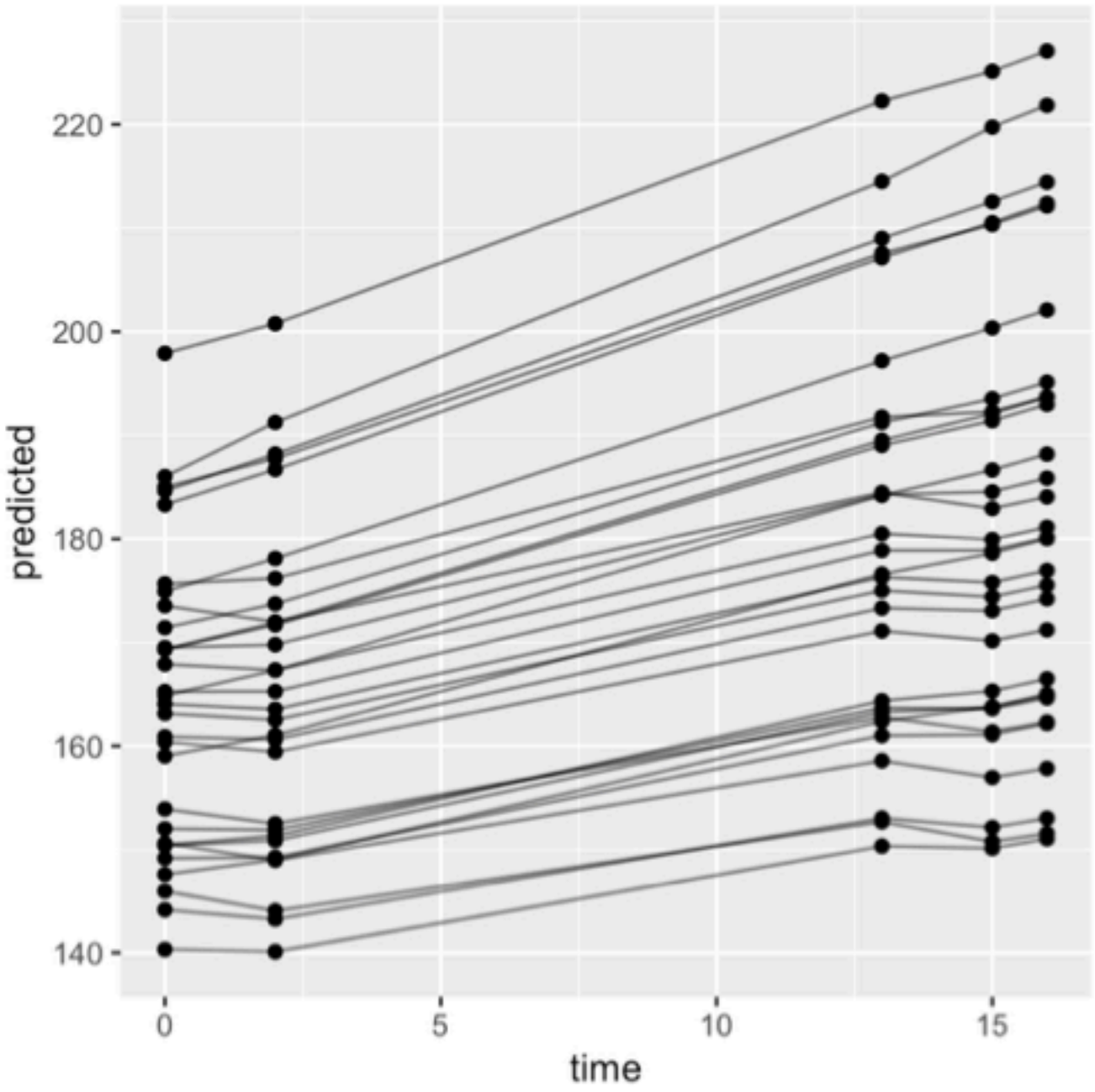
Residual variation exactly the same. These are the same model, with different parameters.

The Growth Curves (for random sample of 30 kids)

Each child has two slopes, one for summer, and one for school.

We make both summers the same.

These curves will be the same regardless of choice of coding of *a* (i.e., summer + school vs. summer + total time).



Substantive conclusions

	Model 1	Model 2
(Intercept)	166.96 ***	166.96 ***
	(0.33)	(0.33)
school	1.45 ***	
	(0.03)	
summer	0.78 ***	-0.67 ***
	(0.11)	(0.13)
time		1.45 ***
		(0.03)

Students grow at about 1.45/month when in school.

Reading growth in summer is less than in school (by a lot, on average).

Substantive conclusions

	Model 1	Model 2
(Intercept)	166.96 ***	166.96 ***
	(0.33)	(0.33)
school	1.45 ***	
	(0.03)	
summer	0.78 ***	-0.67 ***
	(0.11)	(0.13)
time		1.45 ***
		(0.03)

Students grow at about 1.45/month when in school.

Reading growth in summer is less than in school (by a lot, on average).

More interpretation

```
> display( mod_school_summer )
lmer(formula = read ~ school + summer + (school + summer |
id),
    data = dat)
             coef.est coef.se
(Intercept) 166.96     0.33
school        1.45     0.03
summer        0.78     0.11

Error terms:
 Groups   Name          Std.Dev. Corr
 id       (Intercept)   15.78
          school         0.50    0.62
          summer         1.66    0.32 0.94
 Residual                9.83
---
number of obs: 10730, groups: id, 3061
AIC = 88390.9, DIC = 88353.9
deviance = 88362.4
```

There is considerably more variation in learning rate over the summer than in school.

Initial knowledge, summer learning, and school learning rates are all correlated. Summer and school are very correlated.

Learning rate in school:
1.45 +/- 1.0 = 0.45 - 2.45
Summer:
0.78 +/- 3.32 = -2.54 - 4.10

Lecture review

Recap of Piecewise Models

Check in: http://cs179.org/lec34

- ★ Piecewise growth is inherently **cumulative**: as time passes we grow at our various rates.
- ★ Our prediction for time *t* is our intercept plus our total accumulated growth for all the time from *t* to now.
- ★ If we are predicting for negative time, we do the same, just with subtracting the growth to go backwards.
- ★ Different parameterizations make our coefficients have different meanings, but the model shape is the same.

Another piecewise example: Chicago schools

Read chapter 6 of R&B for a nice case study of modeling growth across grades.

Chicago Public School Reading

- ★ Grades 1 - 6
- ★ End of year testing (with some missing testing)
- ★ Research suggests reading in grades 1-3 follows different pattern (faster) than grades 4-6
- ★ A *Piecewise Linear Growth Curve Model* has a general form of:

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{1ti} + \pi_{2i} a_{2ti} + \epsilon_{ti}$$

- ★ Question is: what should the time covariates a_1, a_2 look like?

Frank and Seltzer (1990)
from R&B pg 178

Options for the covariates a_1 and a_2

(a) Two-Rate Model

	Grades 1	2	3	4	5	6	*Interpretation of πs:*
a_{1t}	0	1	2	2	2	2	π_1 growth rate period 1
a_{2t}	0	0	0	1	2	3	π_2 growth rate period 2
							π_0 status Grade 1
a_{1t}	−2	−1	0	0	0	0	π_1 growth rate period 1
a_{2t}	0	0	0	1	2	3	π_2 growth rate period 2
							π_0 status Grade 3

(b) Increment (Decrement) Model

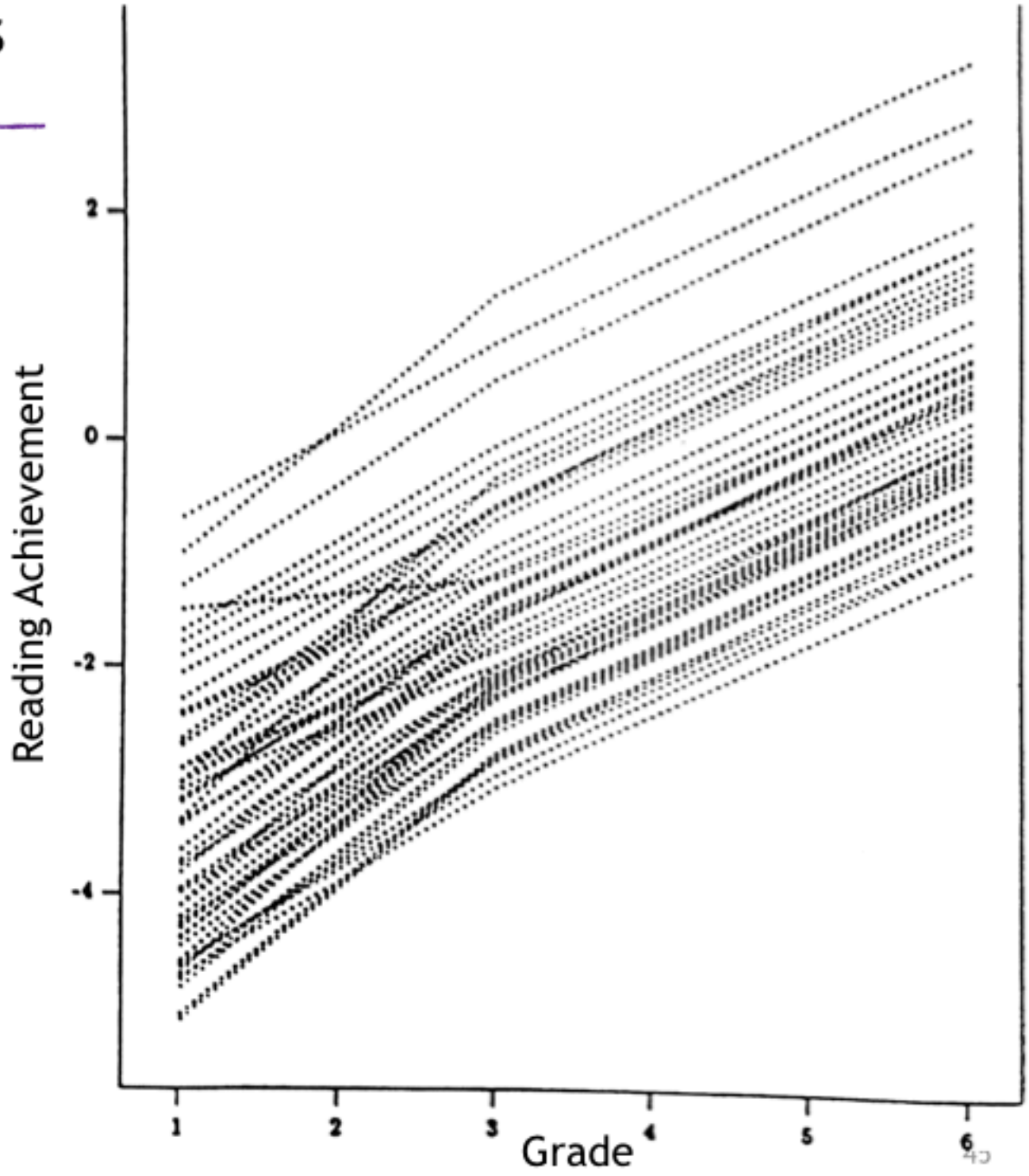
	Grades 1	2	3	4	5	6	*Interpretation of πs:*
a_{1t}	0	1	2	3	4	5	π_1 base growth rate
a_{2t}	0	0	0	1	2	3	π_2 increment (decrement) to growth in period 2
							π_0 status Grade 1
a_{1t}	−2	−1	0	1	2	3	π_1 base growth rate
a_{2t}	0	0	0	1	2	3	π_2 increment (decrement) to growth in period 2
							π_0 status Grade 3

The Growth Curves

Each child effectively has two slopes

Student level curves will be the same regardless of choice of coding of *a*

Think of these curves as showing us the students' different average rates of growth in the two periods.



Supplemental Material: A pointer to some more case studies and a taste of model diagnostics

See Faraway's "Extending the Linear Model with R", chapter 9

This chapter is a very recipe/nuts-and-bolts chapter on analyzing data

Panel Study of Income Dynamics (PSID)

Longitudinal study of a representative sample of US individuals in US.

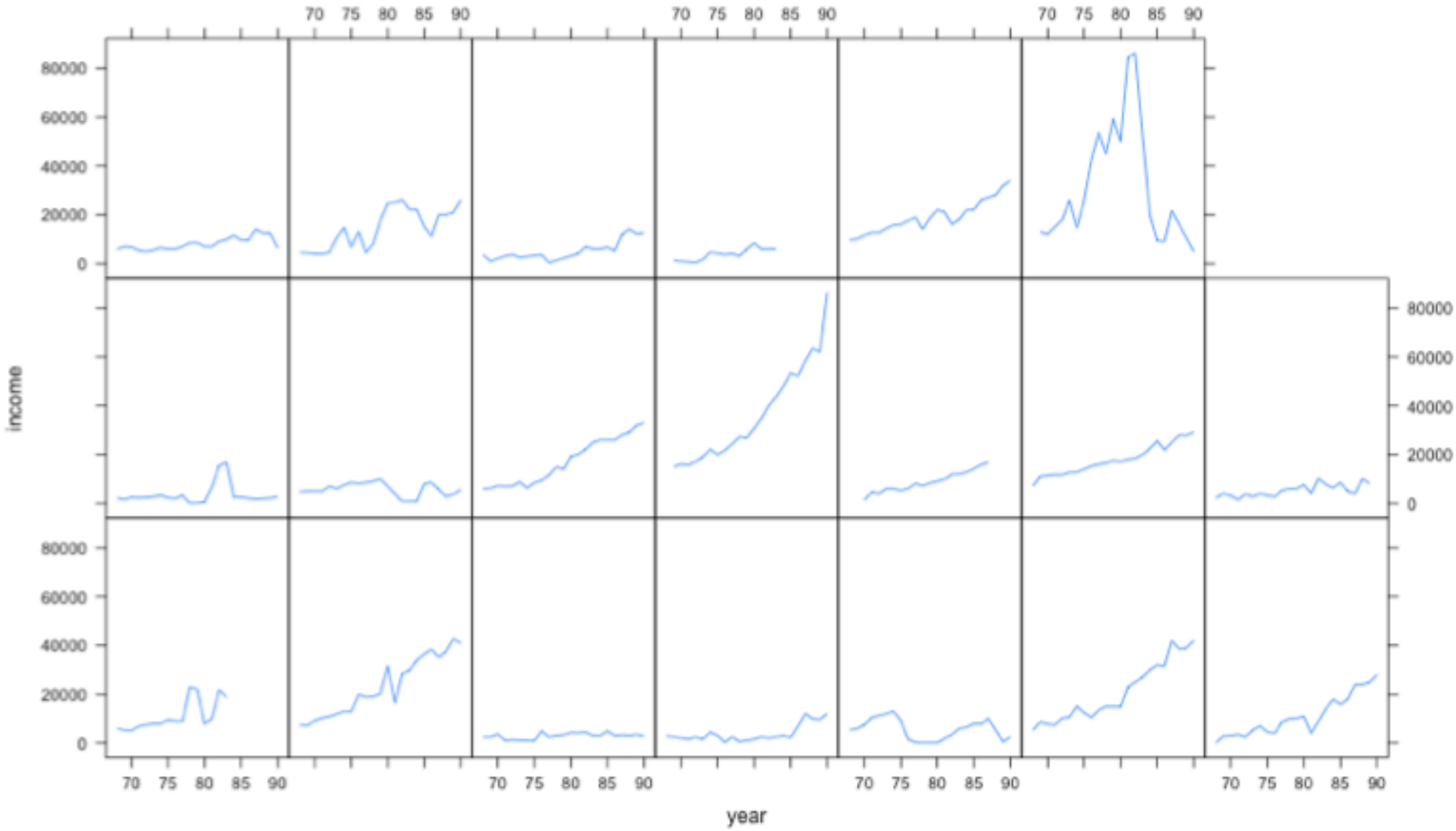
Age 25-39 in 1968

We look at subset of 85 of 8700 households with complete data for at least 11 years in 1968-1990

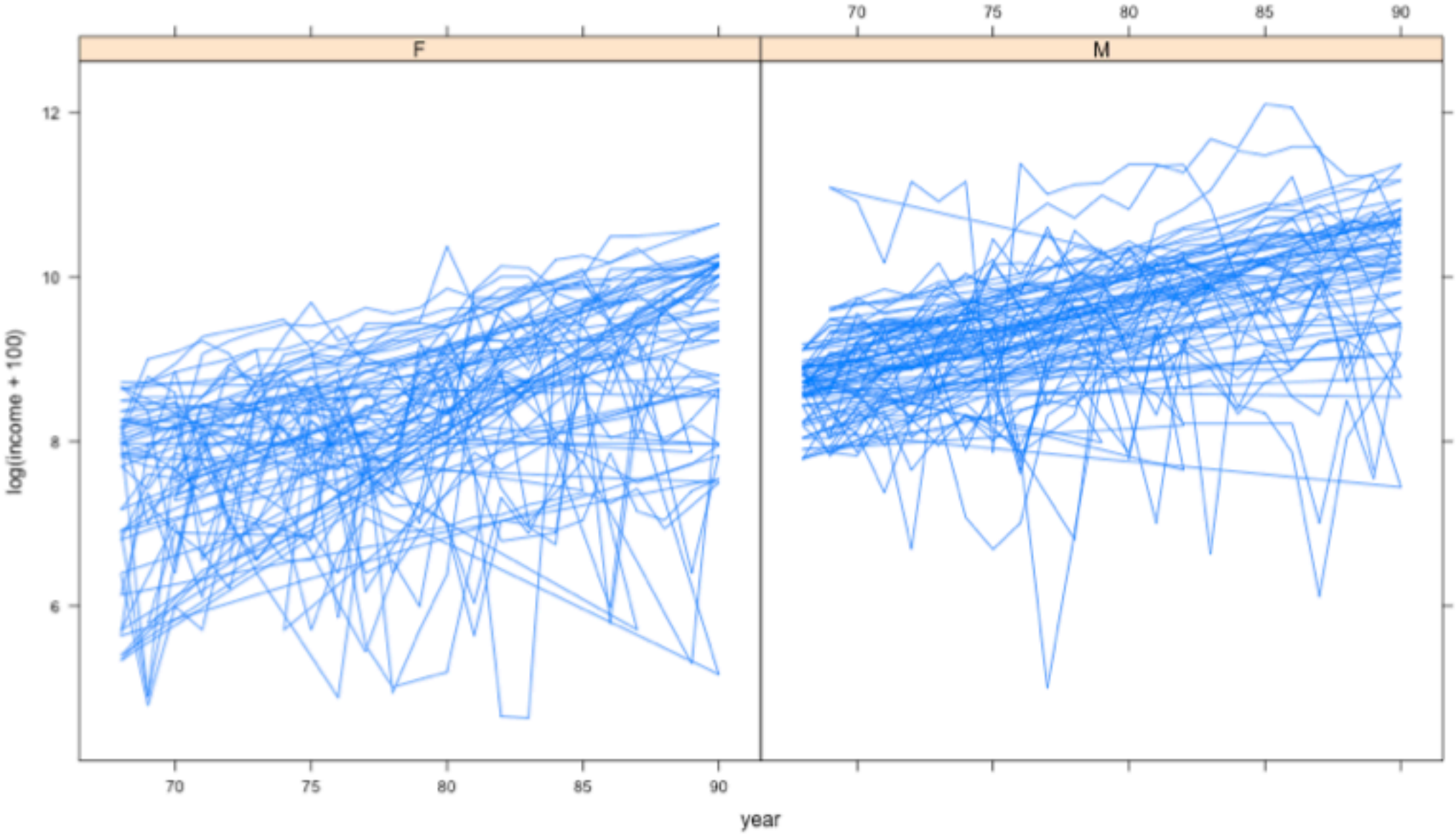
```
> data(psid)
> head(psid)
  age educ sex income year person
1  31   12   M   6000   68      1
2  31   12   M   5300   69
```

Faraway chap 9

Panel Study of Income Dynamics (PSID)

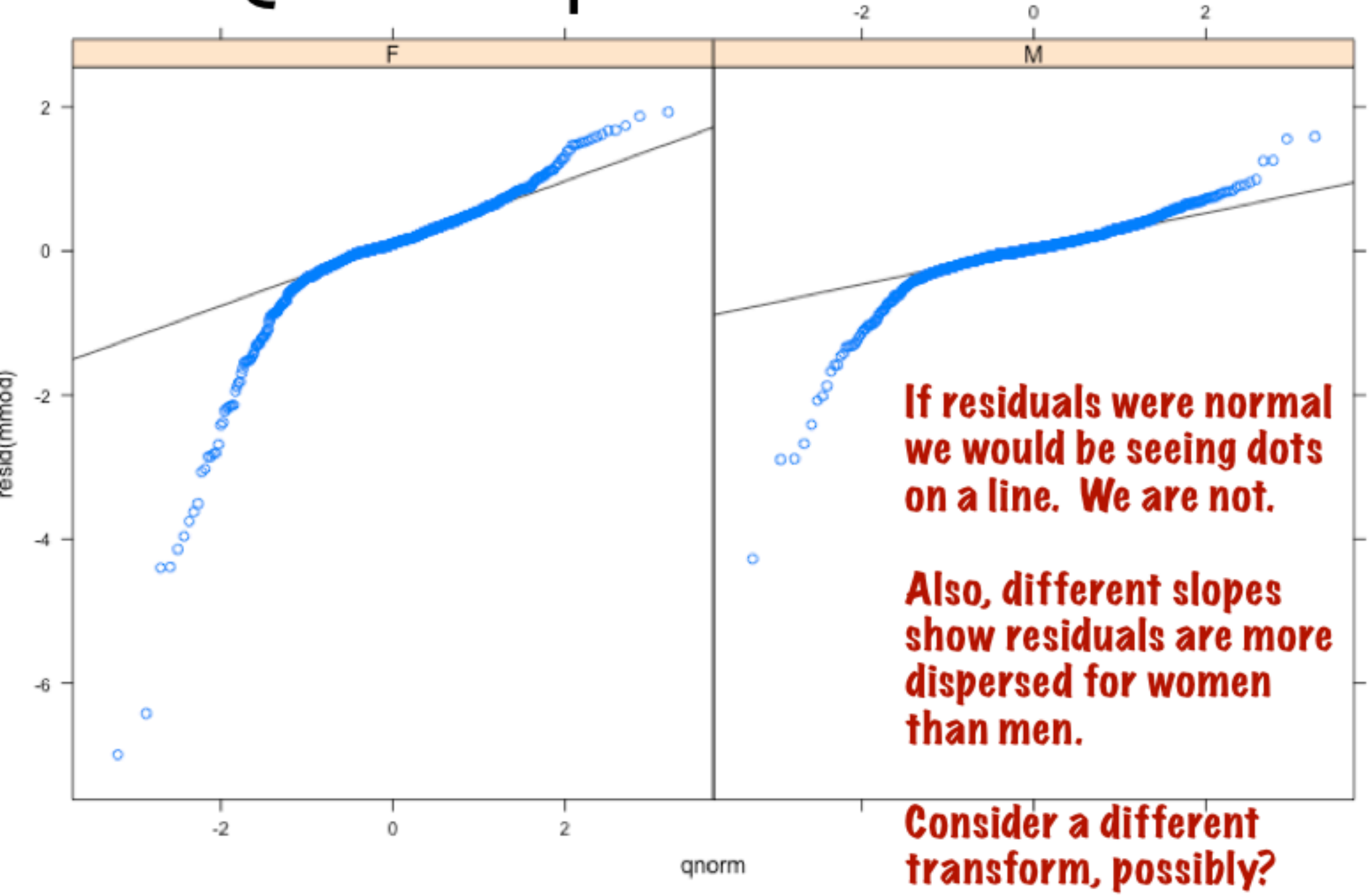


Panel Study of Income Dynamics (PSID)



Y variable is log transformed income (adding 100 to avoid 0s and undefined values)

Quantile plot of residuals

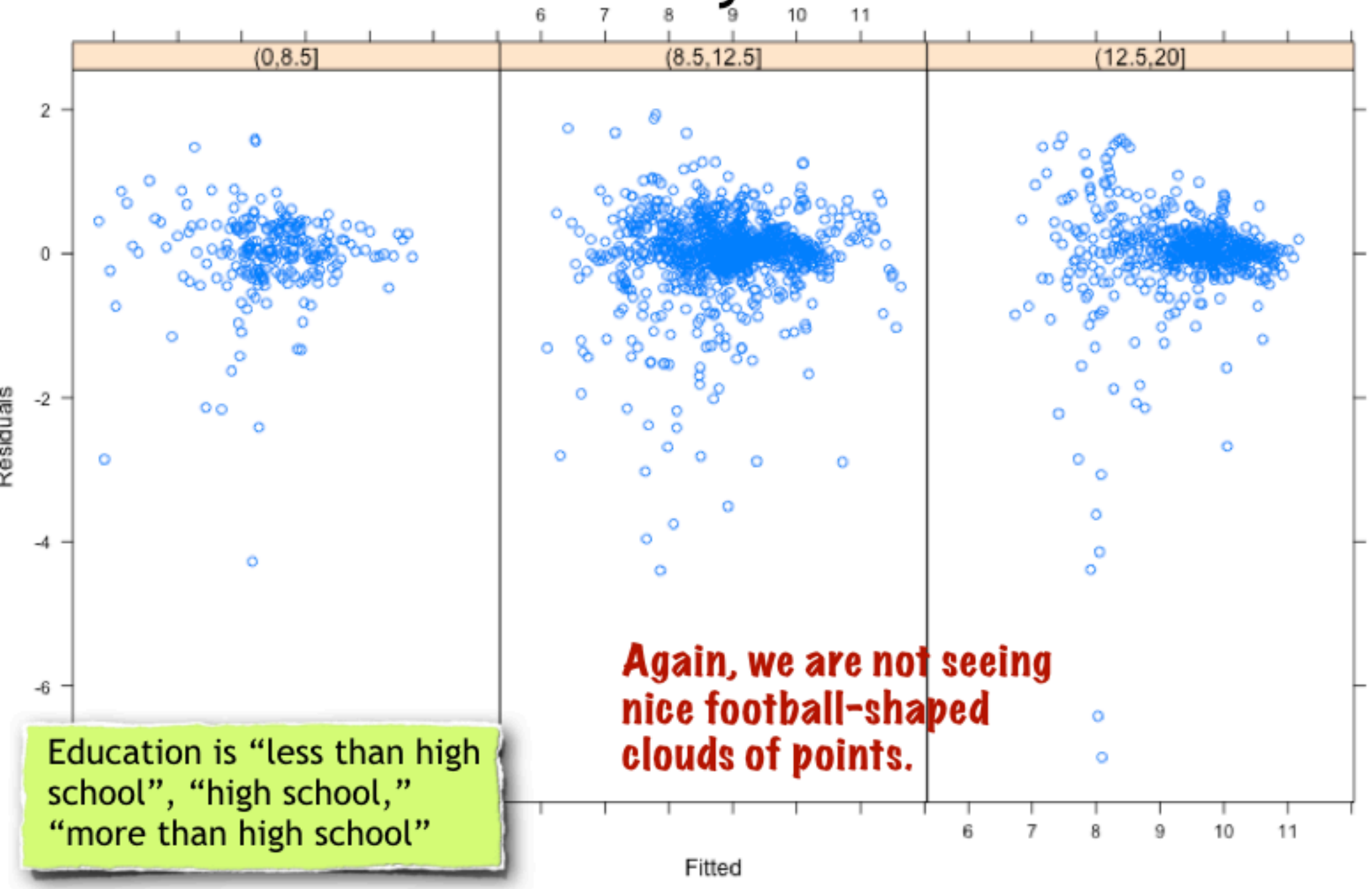


If residuals were normal we would be seeing dots on a line. We are not.

Also, different slopes show residuals are more dispersed for women than men.

Consider a different transform, possibly?

Fitted vs Residual by Education Levels



Education is “less than high school”, “high school,” “more than high school”

Again, we are not seeing nice football-shaped clouds of points.

Pointer to other examples in Faraway, Chapter 9

9.2 “Repeated Measures”

- ★Vision against four different powers of lens
- ★7 subjects, so small data.

9.3 “Multiple response multilevel models”

- ★Analysis of a bivariate outcome of math test and english test.
- ★Models this as an extra level of nesting (so two observations per student)
- ★This introduces many interactions to examine.