

MATH43515: Multilevel Modelling

Lecture 6: Longitudinal data

Module Convenor/Tutor:

Andy Golightly

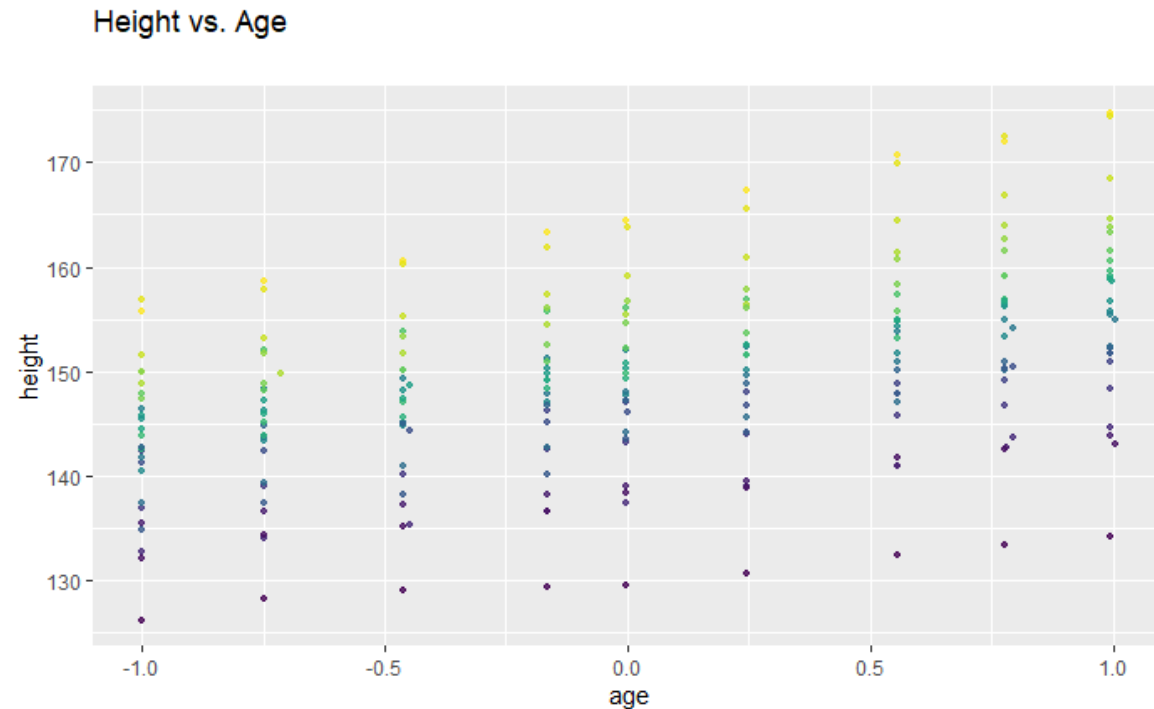
Outline (Lecture 6)

- Repeated measures data
- Example: Oxford boys data
- Wide and long data format
- Example: College student GPA data
- Modelling strategies including cross-level interactions and random slopes

Repeated measures data

- Consider repeated measurements on individuals taken over time
- It is intuitively clear that measurements pertaining to a certain individual will have larger correlation between themselves than with measurements from other individuals
- This “within-individual-correlation” can again be dealt with by two-level models; but now
 - The upper level corresponds to individuals
 - The lower level corresponds to repeated measurements on individuals

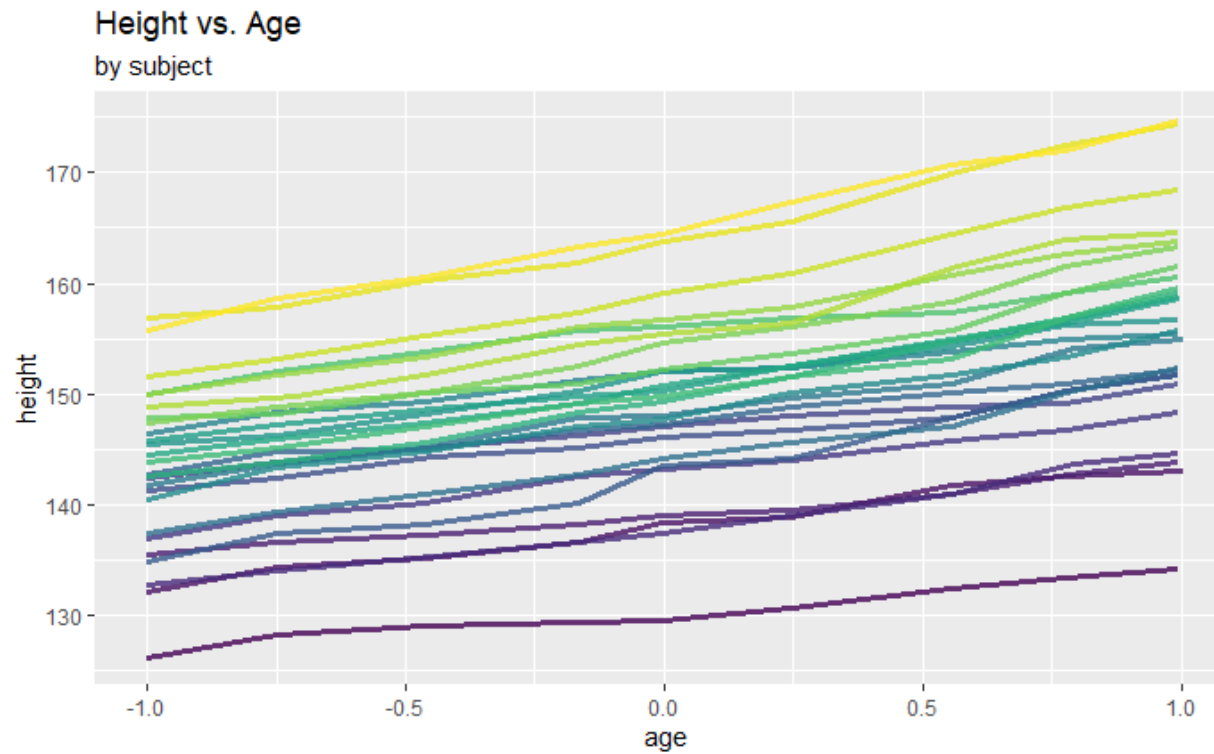
Oxford boys data



- Data on the height (in cm) of 26 boys from Oxford, England versus age
- Age is measured on a standardized and dimensionless scale with nine possible values, yielding a total of 234 observations

Oxford boys data (cont'd)

Same plot, just connecting the dots:



Repeated measures data formats

For traditional analysis, for software such as SPSS or Excel, repeated measures data are often presented in “wide format”:

age	-1	-0.7479	-0.463	-0.1643	-0.0027	0.2466	0.5562	0.7781	0.9945
Subject									
1	140.5	143.4	144.8	147.1	147.7	150.2	151.7	153.3	155.8
2	136.9	139.1	140.1	142.6	143.2	144	145.8	146.8	148.3
3	150	152.1	153.9	155.8	156	156.9	157.4	159.1	160.6
4	155.7	158.7	160.6	163.3	164.4	167.3	170.7	172	174.8
5	145.8	147.3	148.7	149.78	150.22	152.5	154.8	156.4	158.7
...									
26	132.2	134.3	135.1	136.7	138.4	138.9	141.8	142.6	143.1

The wide format is impractical for “modern” multilevel modelling, where we need a data frame with one row per measurement.

Repeated measures data formats (cont'd)

```
> head(Oxboys)
```

Grouped Data: height ~ age | Subject

	Subject	age	height	Occasion
1	1	-1.0000	140.5	1
2	1	-0.7479	143.4	2
3	1	-0.4630	144.8	3
4	1	-0.1643	147.1	4
5	1	-0.0027	147.7	5
6	1	0.2466	150.2	6

```
> dim(Oxboys)
```

```
[1] 234 5
```

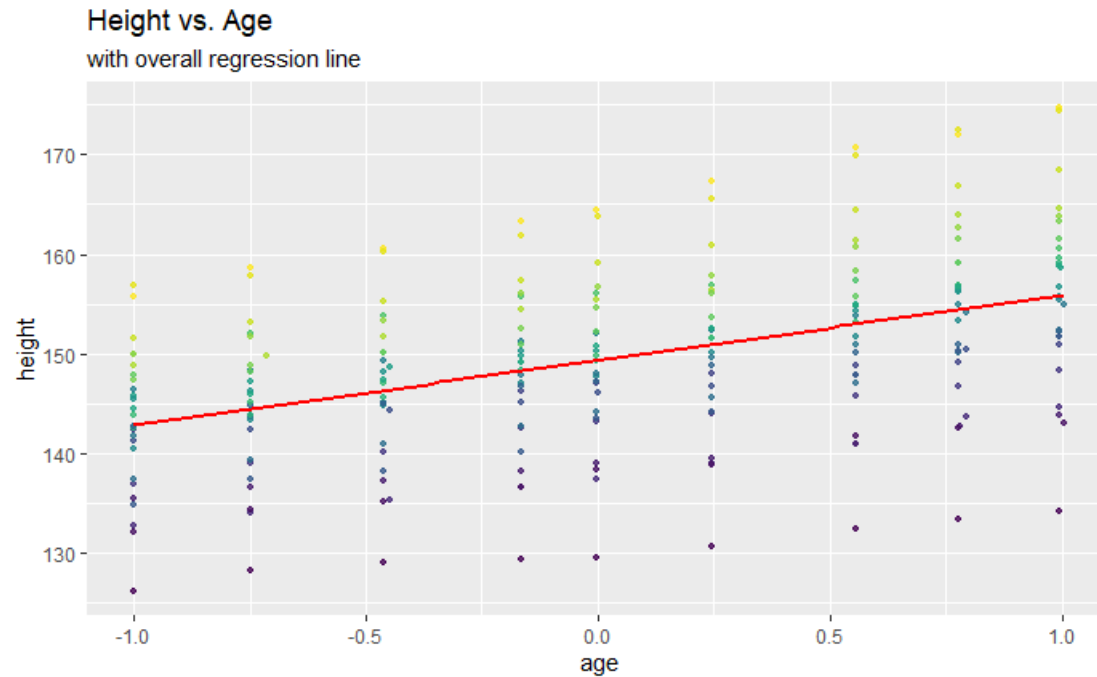
In R, repeated measures data will usually be stored in “**long format**”.

Directly usable in multilevel R functions such as lmer.

Very straightforward handling of missing values!

Modelling the Oxford boys data: linear model

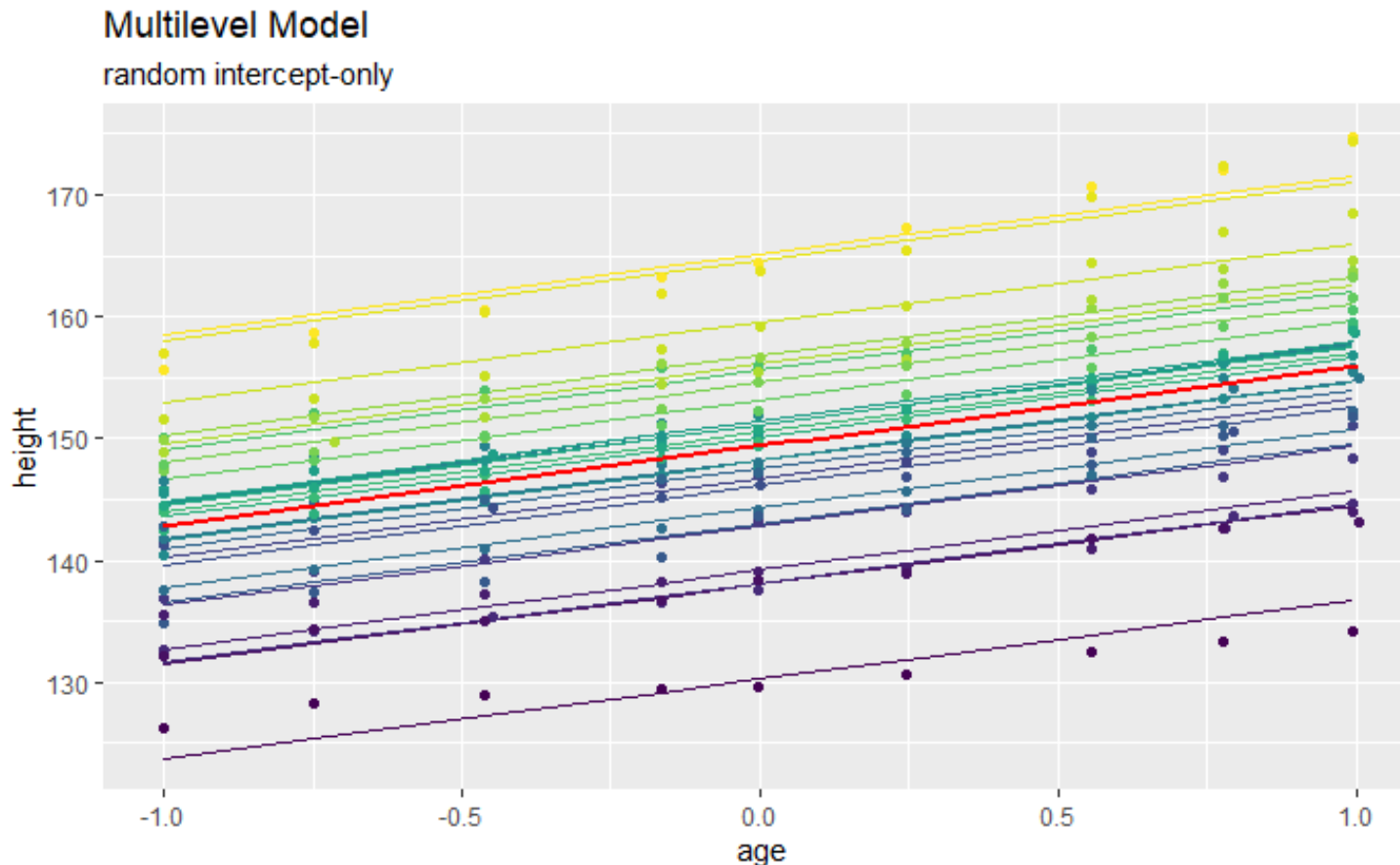
Overall least squares fit without consideration of group structure:



Captures the linear growth, but is otherwise a quite poor description of the data

Modelling the Oxford boys data: random effect model

Now with subject-level random effect:



We see that the vertical offset caused by the random intercept “sticks” with the individuals as they grow over time.

Comparing fixed and random effect models

```
> fit1 <- lm(height ~ age, data=Oxboys)
> round(summary(fit1)$coef, digits=3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	149.372	0.529	282.599	0
age	6.521	0.817	7.982	0

```
> fit2 <- lmer(height ~ age + (1 | Subject),
data=Oxboys)
> round(summary(fit2)$coef, digits=3)
```

	Estimate	Std. Error	df	t value	Pr(t)
(Intercept)	149.372	1.590	25	93.933	0
age	6.524	0.133	207	49.228	0

We see that a correct incorporation of the grouping structure changes the parameter standard errors considerably (in different directions for the two parameters!!).

The slope parameter itself does not change a lot, for this particular data set. But this does not remove the need to account for the grouping structure.

Repeated measures and longitudinal data

Repeated measures are often used synonymous with longitudinal data.

(To be more precise, one could say that repeated measures *lead to* longitudinal data).

The repeated measures can be produced at fixed or varying occasions. Within a multilevel framework, this is an irrelevant distinction. More traditional methods (“repeated measures ANOVA” etc) can deal only with fixed occasions.

The simple 2-level model for longitudinal data

Denote by y_{ti} the t -th measurement for individual i , and T_{ti} the time stamp (which can be measured in continuous or discrete time, or in form of some type of age measurement).

This gives rise to the model

$$y_{ti} = a_i + b_i T_{ti} + \epsilon_{ti}$$

where for individual i ,

$$a_i = a + u_i \text{ with } u_i \sim N(0, \sigma_u^2),$$

$$b_i = b + v_i \text{ with } v_i \sim N(0, \sigma_v^2),$$

and $\epsilon_{ti} \sim N(0, \sigma^2)$ as usual.

Of course,
this is just
exactly the same
as the two-level
model from
Lecture 4!

Additional covariates

There may be other covariates, x_{ti} , beyond “time”.
Specifically, including

- a lower-level covariate x_{ti}
- an upper (subject)-level covariate z_i

the model becomes

$$y_{ti} = a_i + b_i T_{ti} + c_i x_{ti} + \epsilon_{ti}$$

where $\epsilon_{ti} \sim N(0, \sigma^2)$ as usual, and for individual i

$$a_i = a + \alpha z_i + u_i \quad \text{with } u_i \sim N(0, \sigma_u^2),$$

$$b_i = b + \beta z_i + v_i \quad \text{with } v_i \sim N(0, \sigma_v^2),$$

$$c_i = c + \gamma z_i + w_i \quad \text{with } w_i \sim N(0, \sigma_w^2).$$

Repeated measures level
(lower)

Individual level (upper)

Additional covariates (cont'd)

Plugging the individual-level expressions in the lower-level ones, we obtain

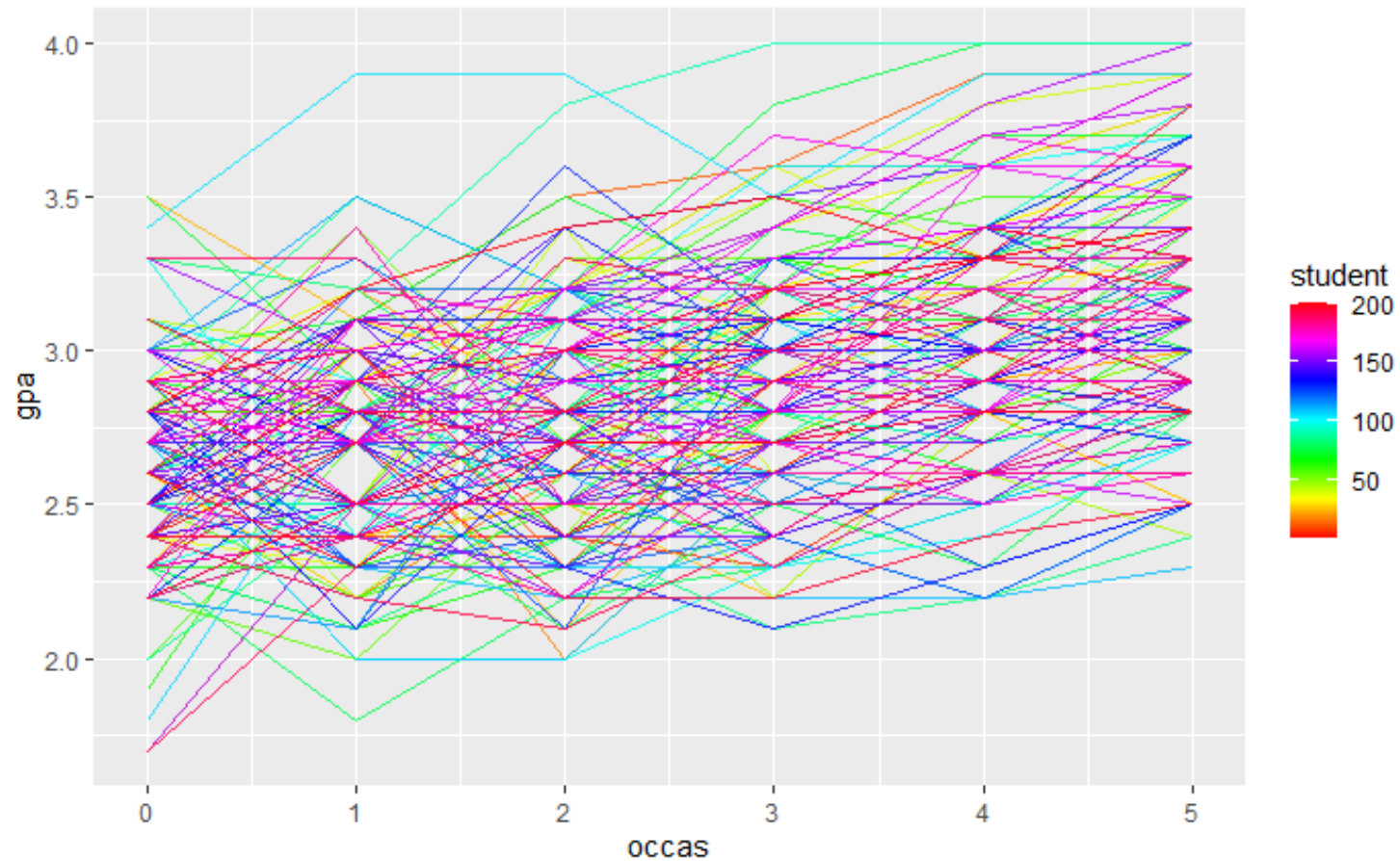
$$\begin{aligned} y_{ti} &= a + \alpha z_i + u_i + bT_{ti} + \beta z_i T_{ti} + v_i T_{ti} + cx_{ti} + \gamma z_i x_{ti} \\ &+ w_i x_{ti} + \epsilon_{ti} \end{aligned}$$

This is a complex model including **some cross-level interactions**. There may be good reasons to justify such a model, but in practice one will often seek to simplify the individual-level models.

Example: College student GPA data

- Longitudinal data from 200 college students ($i = 1, \dots, 200$) over six semesters ($t = 1, \dots, 6$).
- Response variable is Grade Point average $y_{ti} = gpa_{ti}$ of student i in semester t
- Predictor variables (covariates) are
 - Hours worked per day, $x_{ti} = job_{ti}$, by student i in semester t
 - gender $z_i = gend_i$ of student i (0=male, 1=female)
- Time variable (“measurement occasion”) given by $T_{ti} = occas_{ti} = t - 1$

College student GPA data



Intra-class correlation

```
> lfit0 <- lmer(formula = gpa ~ 1 + (1|student), data = gpa.data)
> summary(lfit0)
....
Random effects:
Groups   Name      Variance Std.Dev.
student (Intercept) 0.05714  0.2390
Residual              0.09759  0.3124
Number of obs: 1200, groups: student, 200
...
> rho0 = 0.05714 / (0.05714 + 0.09759)
> rho0
[1] 0.3692884
```

About one third of the variance of the GPA measures is variance **between individuals**, and about two thirds is variance **within individuals across time**.

Modelling the college student GPA data

Let's begin with the full model including the cross-level interactions:

```
> lfit1 <- lmer(gpa ~ occas + job + gend + job:gend + occas:gend +  
(1 + occas + job | student), data=gpa.data)
```

The notation job:gend represents the term of $\gamma z_i x_{ti}$, and occas:sex represents $\beta z_i T_{ti}$.

Do we need the cross-level interactions?

```
> round(summary(lfit1)$coef, digits=4)
```

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	2.9110	0.0593	376.8731	49.0640	0.0000
occas	0.0852	0.0079	195.2881	10.8429	0.0000
job	-0.1604	0.0238	610.9215	-6.7341	0.0000
gend	-0.0318	0.0853	446.8028	-0.3727	0.7095
job:gend	0.0457	0.0351	650.8235	1.3023	0.1933
occas:gend	0.0309	0.0108	195.1070	2.8506	0.0048

Interpretation:

- The impact of the number of hours worked on the GPA does not depend on gender.
- The impact of the semester on the GPA does depend on gender.

Do we need the random slopes?

```
> summary(lfit1)
```

```
...
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
student	(Intercept)	0.067626	0.26005	
	occas	0.003477	0.05897	0.23
	job	0.002767	0.05261	-0.66 -0.88
Residual		0.041339	0.20332	

Number of obs: 1200, groups: student, 200

At first glance,
it looks like
rather
not...

Do we need the random slopes? (cont'd)

```
> ranova(lfit1)
```

ANOVA-like table for random-effects: Single term deletions

Model:

```
gpa ~ occas + job + sex + (1 + occas + job | student) + job:sex + occas:sex
```

	npars	logLik	AIC	LRT	Df
<none>	13	-101.65	229.30		
occas in (1 + occas + job student)	10	-151.84	323.68	100.386	3
job in (1 + occas + job student)	10	-104.80	229.60	6.302	3

	Pr(>Chisq)
<none>	
occas in (1 + occas + job student)	< 2e-16 ***
job in (1 + occas + job student)	0.09782 .

Interpretation: The random effect term for job is not needed ($p > 0.05$)

Let's review...

In two-level formulation:

$$y_{ti} = a_i + b_i T_{ti} + c_i x_{ti} + \epsilon_{ti}$$

where for individual i

$$a_i = a + \alpha z_i + u_i \quad \text{with } u_i \sim N(a, \sigma_u^2),$$

$$b_i = b + \beta z_i + v_i \quad \text{with } v_i \sim N(b, \sigma_v^2),$$

$$c_i = c + \cancel{\gamma z_i} + \cancel{w_i} \quad \text{with } w_i \sim N(b, \sigma_w^2).$$

Or in plug-in formulation:

$$\begin{aligned} y_{ti} \\ = a + \alpha z_i + u_i + b T_{ti} + \beta z_i T_{ti} + v_i T_{ti} + c x_{ti} + \cancel{\gamma z_i} x_{ti} + \cancel{w_i} x_{ti} + \epsilon_{ti} \end{aligned}$$

Refit the simplified model

After removing the two unnecessary terms,

```
lfit2 <- lmer(gpa ~ occas + job+ gend + occas:gend + (occas | student), data=gpa.data)
```

Note this is the same as

```
lfit2 <- lmer(gpa ~ occas*gend + job+ (occas | student), data=gpa.data)
```

Refit the simplified model (cont'd)

```
> summary(lfit2)
```

Random effects:

Groups	Name	Variance	Std.Dev.	Corr
student	(Intercept)	0.041347	0.20334	
	occas	0.003676	0.06063	-0.19
Residual		0.041597	0.20395	

Number of obs: 1200, groups: student, 200

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(> t)
(Intercept)	2.849e+00	4.582e-02	9.193e+02	62.177	< 2e-16 ***
occas	8.783e-02	7.994e-03	1.967e+02	10.987	< 2e-16 ***
job	-1.321e-01	1.727e-02	1.035e+03	-7.648	4.65e-14 ***
gend	6.790e-02	3.559e-02	1.979e+02	1.908	0.05787 .
occas:gend	2.956e-02	1.102e-02	1.957e+02	2.683	0.00791 **

Note: The fixed effects slope for gender is not significant at the 5% level, but usually we would not remove a main effect (gend) when keeping an interaction effect involving it (occas:gend).

Interpreting the model

$$\begin{aligned}y_{ti} &= a + \alpha z_i + u_i + bT_{ti} + \beta z_i T_{ti} + cx_{ti} + \epsilon_{ti} \\&= 2.849 + 0.068 \times gend_i + 0.088 \times occas_{ti} - 0.132 \times job_{ti} \\&\quad + 0.029 \times gend_i \times occas_{ti} + \\&\quad + u_i + v_i \times occas_{ti} + \epsilon_{ti}\end{aligned}$$

with $u_i \sim N(0, 0.041)$, $v_i \sim N(0, 0.0037)$ and $\epsilon_{ti} \sim N(0, 0.042)$.

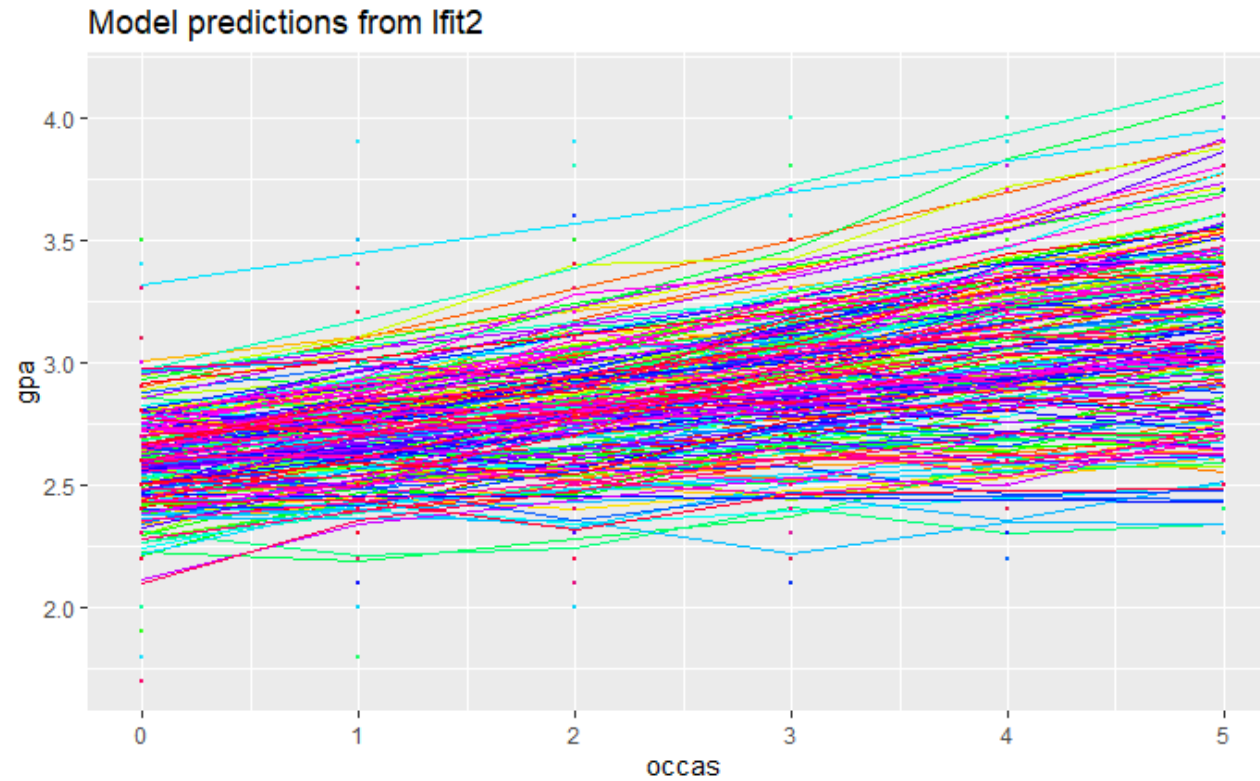
Recall:

$$\begin{aligned}z_i &= gend_i \\x_{ti} &= job_{ti} \\T_{ti} &= occas_{ti}\end{aligned}$$

So,

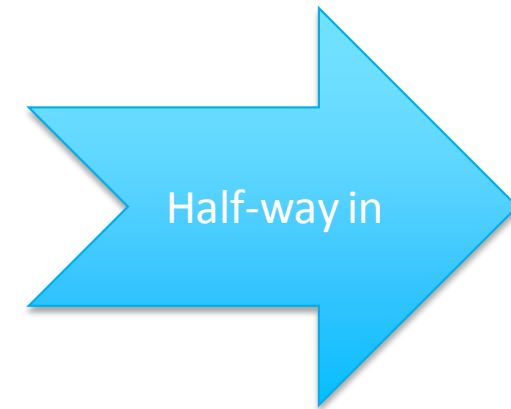
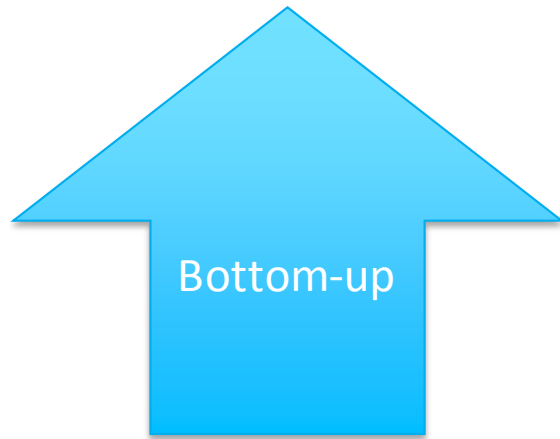
- The expected GPA is 0.068 higher for female than for male students
- The expected GPA decreases with 0.132 for each hour worked
- For male students, the expected GPA increases with 0.088 each term
- For female students, the expected GPA increases with $0.088 + 0.029 = 0.117$ each term.

Predictions from fitted model



A word on modelling strategies

In the light of the preceding analysis, there appear to be several potential strategies for the analysis of multilevel data (not only longitudinal data):



Bottom-up modelling strategy

1. Fit the empty model (without covariates, only group-level random intercept). Compute ICC.
 2. Include lower-level explanatory variables
 3. Include higher-level explanatory variables
 4. Include random coefficients
 5. Add cross-level interactions
- at each step checking appropriately for significance/relevance of terms.

Advantage: Systematic, principled approach.

Disadvantage: Many options to consider, feels like fishing in the dark.

This is the technique suggested by Hox et al (2018), and what we started doing for the 3-level data.

Top-down modelling strategy

1. Compute the full two-level model including all random effects and cross-level interactions
2. Remove irrelevant cross-level interactions
3. Remove irrelevant random slopes
4. Remove irrelevant higher level covariates
5. Remove irrelevant lower level covariates

This is what we did for the GPA data (taking the computation of ICC aside)


Advantage: With the first fit, one gets a good impression on relevant/irrelevant terms

Disadvantage: Initial model may be too complex, suffer from multicollinearities, or may not fit at all

Halfway-in strategy

1. Carry out an explanatory analysis of the data (may include ICC)
2. Based on this, fit a suitable base model, for instance including the identified fixed effects and a group level intercept.
3. Inspect fitted model, carry out some diagnostics. Simplify or expand model as appropriate.

This is what we did for the student extraversion data.



Thank
you!!!!!!

