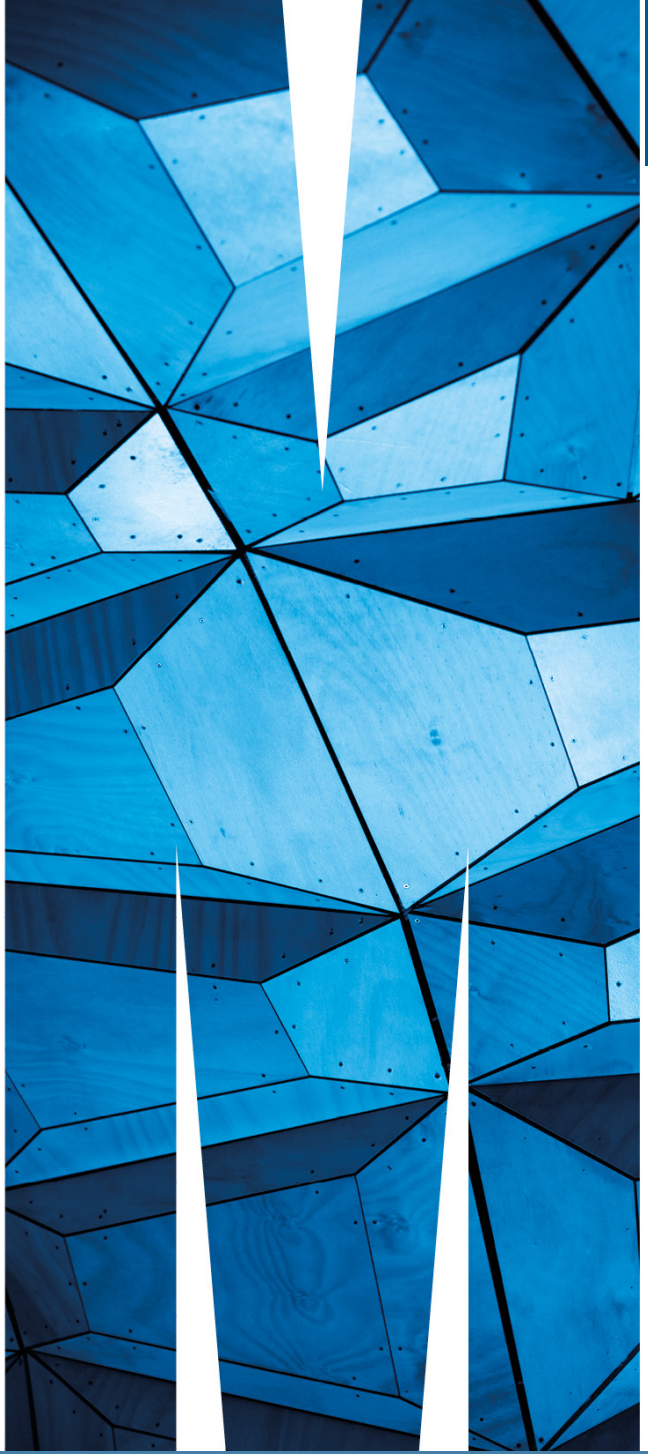
MONASH University

MONASH BUSINESS SCHOOL



Introduction to Multi-level Models using R

Professor Di Cook, Econometrics and Business Statistics

Workshop for the Institute for Safety, Compensation and Recovery Research

GROUP OF EIGHT AUSTRALIA

AACSB ACCREDITED EFMD EQUIS ACCREDITED ASSOCIATION AMBA ACCREDITED

Outline

- Session 1: Basic models, fitting multiple separate models
- Session 2: Putting it together, using mixed effects models
- Session 3: Summarising and visualising models
- Session 4: Advanced modeling

Session 2

- **Putting it together, using mixed effects models**

What is a multilevel model?

- Observations are not independent, but belong to a hierarchy
- Example: individual level demographics (age, gender), and state level information (health provider choice policy, waiting time, reimbursement policy)
- Multilevel model enables fitting accommodating different types of dependencies

Mixed effects models

For data organized in g groups, consider a continuous response linear mixed-effects model (LME model) for each group i, $i = 1, \ldots, g$:

$$\underset{(n_i \times 1)}{\mathbf{y}_i} = \underset{(n_i \times p)(p \times 1)}{\mathbf{X}_i \quad \beta} + \underset{(n_i \times q)(q \times 1)}{\mathbf{Z}_i \quad \mathbf{b}_i} + \underset{(n_i \times 1)}{\varepsilon_i}$$

- $\mathbf{y}_i$ is the vector of outcomes for the n_i level-1 units in group i
- $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices for the fixed and random effects
- β is a vector of p fixed effects governing the global mean structure
- $\mathbf{b}_i$ is a vector of q random effects for between-group covariance
- ε_i is a vector of level-1 error terms for within-group covariance

Fixed vs random

- *Fixed effects* can be used when you know all the categories, e.g. age, gender, smoking status
- *Random effects* are used when not all groups are captured, and we have a random selection of the groups, e.g. individuals (if you have multiple measurements), schools, hospitals

Data

- *radon*: 919 owner-occupied homes in 85 counties of Minnesota.
- *autism*: prospective longitudinal study following 214 children between the ages of 2 and 13 who were diagnosed with either autism spectrum disorder or non-spectrum developmental delays at age 2.
- *wages*: 6402 observations on labor-market experience of 888 male high school dropouts
- *Exam*: Exam scores of 4059 students from 65 schools in Inner London.

Your turn

- In each of the data examples, load into R, examine the help information
- Identify the response variable, hierarchical elements, individuals and groups, and the fixed vs random effects

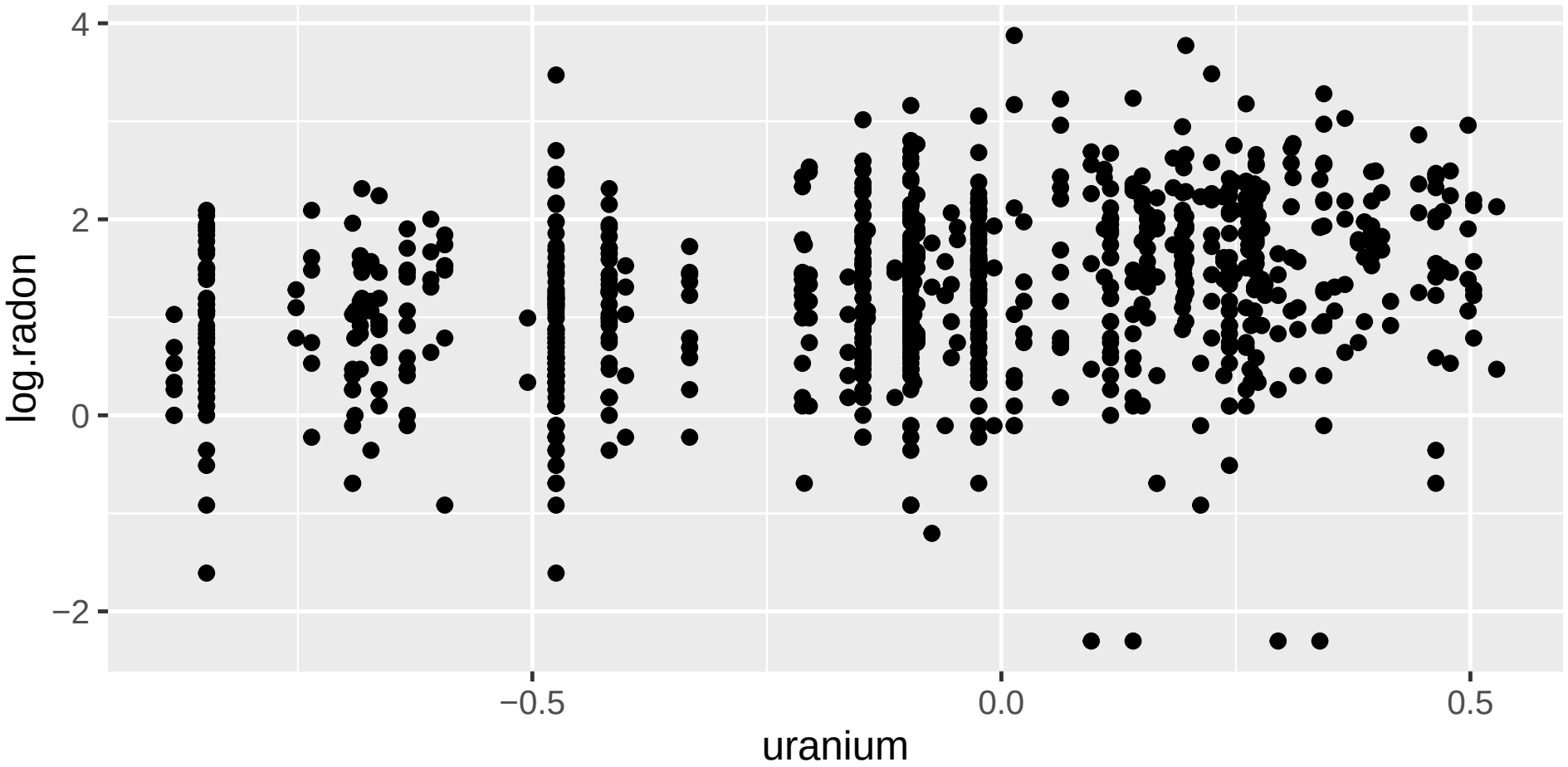
```
library(HLMdiag)
?radon
library(mlmRev)
?Gcsemv
```

Fitting a model to radon

- Response: *log.radon*
- Fixed: *basement* (categorical), *uranium* (quantitative)
- Random: *county* (house is a member of county)

Examining the data

```
ggplot(radon, aes(x=uranium, y=log.radon)) + geom_point()
```



Plot of response vs covariate. What do you see?

- Vertical stripes: each county is represented by an average uranium value
- Weak linear association, lots of variation for houses within county
- Four points inline horizontally at the base (be suspicious)
- Some counties only have 2, 3 points
- Scales?

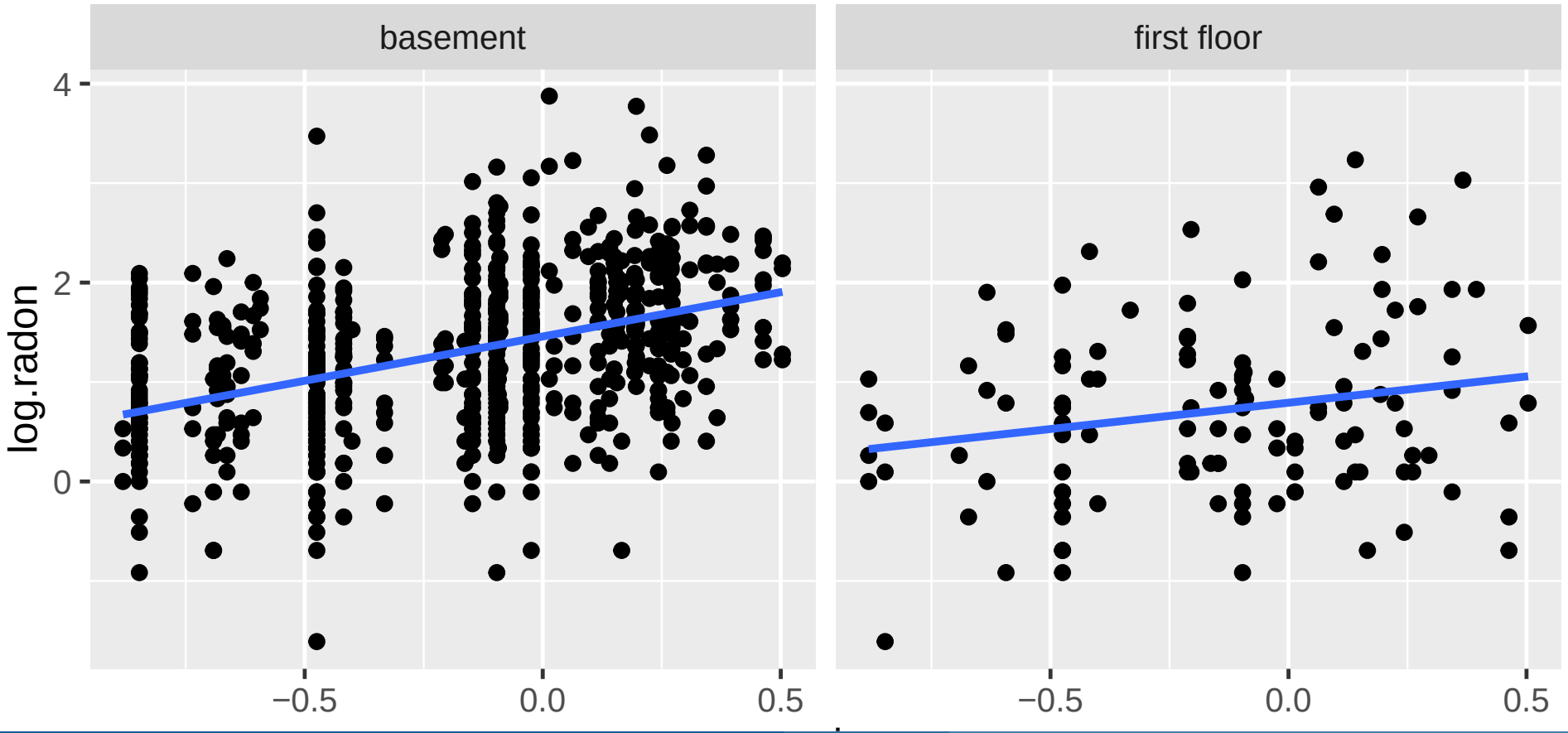
Pre-processing

- Counties with less than 4 observations removed
- Four flat-line observations should be removed, really suspect these were erroneously coded missing values

```
radon_keep <- radon %>% group_by(county) %>%
  tally() %>% filter(n > 4)
radon_sub <- radon %>%
  filter(county %in% radon_keep$county & log.radon > -2)
radon_sub$basement <-
  factor(radon_sub$basement, levels=c(0,1),
         labels=c("basement", "first floor"))
```

Look again

```
ggplot(radon_sub, aes(x=uranium, y=log.radon)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  facet_wrap(~basement)
```



Fit a simple model

$$log.radon = \beta_0 + \beta_1 basement + \beta_2 uranium + \varepsilon$$

```
radon_lm <- glm(log.radon ~ basement + uranium,
                data = radon_sub)
summary(radon_lm)
```

Your turn

What does this fitted model look like? Make a sketch.

Your turn

Adapt the model to include a different slope for basement level. Write the model equation also.

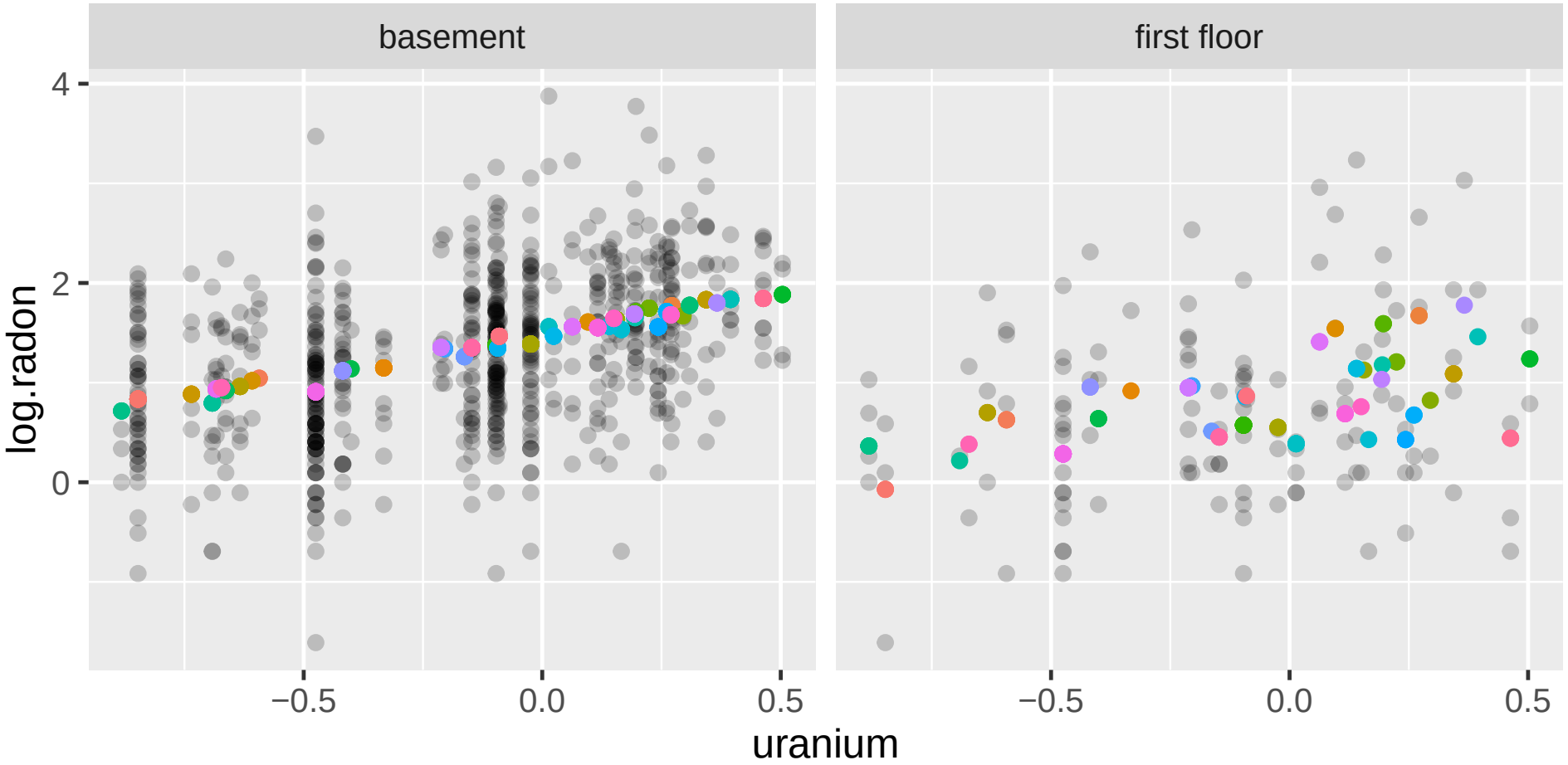
Mixed effects

$$log.radon_{ij} = \beta_0 + \beta_1 basement_{ij} + \beta_2 uranium_i + b_{0i} + b_{1i} basement_{ij} + \varepsilon_{ij}$$

$$i = 1, ..., \#counties; j = 1, ..., n_i$$

```
radon_lmer <- lmer(log.radon ~ basement + uranium +
  (basement | county.name), data = radon_sub)
summary(radon_lmer)
```

```
radon_lmer_fit <- radon_sub
radon_lmer_fit$fit <- fitted(radon_lmer)
ggplot(radon_lmer_fit, aes(x=uranium, y=log.radon)) +
  geom_point(alpha=0.2) +
  geom_point(aes(y=fit, colour=county.name)) +
  facet_wrap(~basement) + theme(legend.position="none")
```



Your turn

What does the syntax $(basement|county.name)$ provide in the model? How would the model fit if we had used $(1|county.name)$ instead?

Examining the model output: fixed effects

```
Fixed effects:
                     Estimate Std. Error t value
(Intercept)           1.48066    0.03856   38.40
basementfirst floor -0.59011    0.11246   -5.25
uranium               0.84600    0.09532    8.88
```

How do these compare with the simple linear model estimates?

Examining the model output: random effects

```
Random effects:
 Groups      Name                    Variance Std.Dev. Corr
 county.name (Intercept)             0.01388  0.1178
             basementfirst floor     0.22941  0.4790   0.02
 Residual                            0.50694  0.7120
Number of obs: 796, groups:  county.name, 46
```

This is saying that the variance of the estimates for first floor observations is larger than the basement.

Your turn

How does the mixed effects model differ from the simple linear model? (Hint: Think about the variance.)

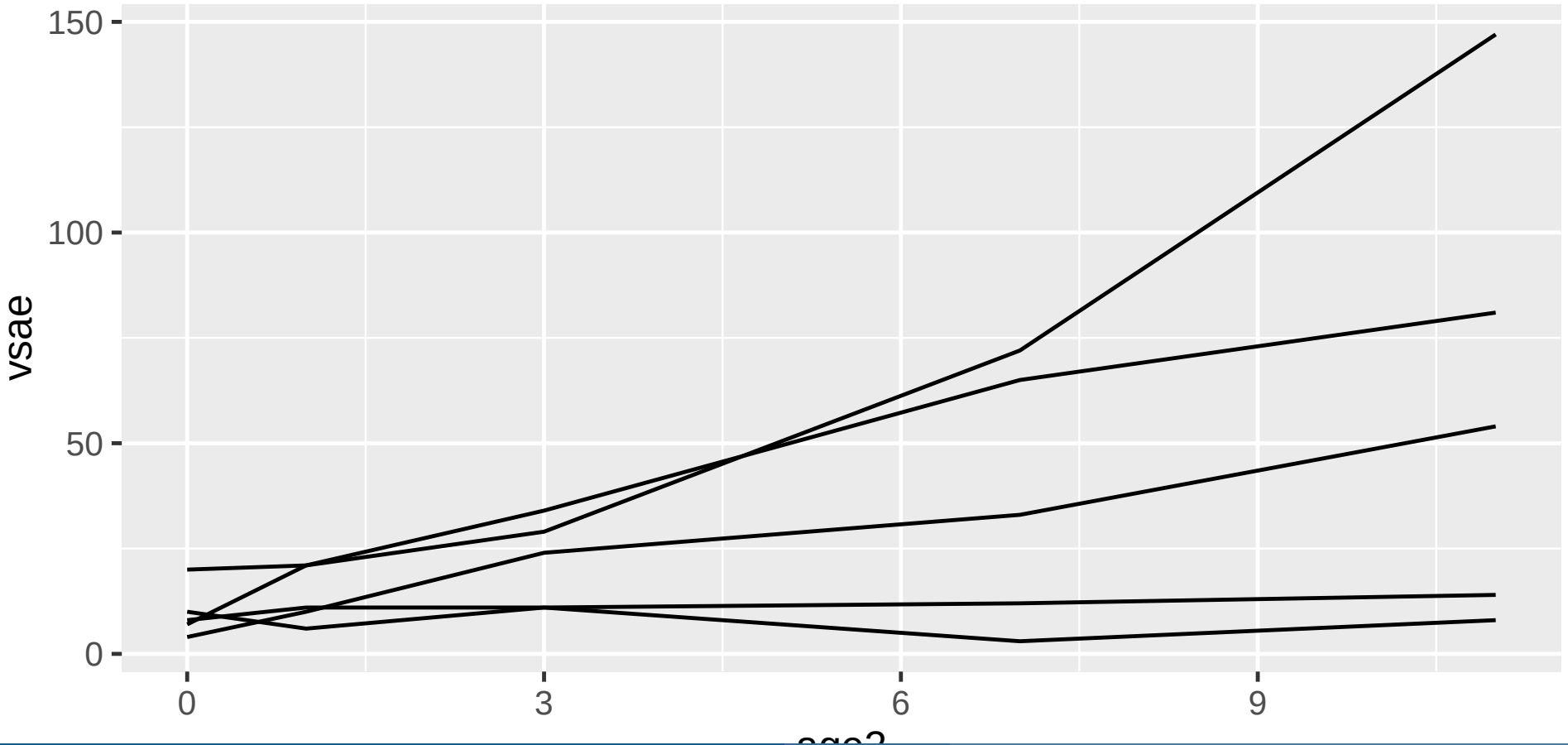
Autism data

- Response: *vsae* Vineland Socialization Age Equivalent
- Fixed: *gender*, *race*, *age2*, *bestest2*, *sicdegp*
- Random: *childid*

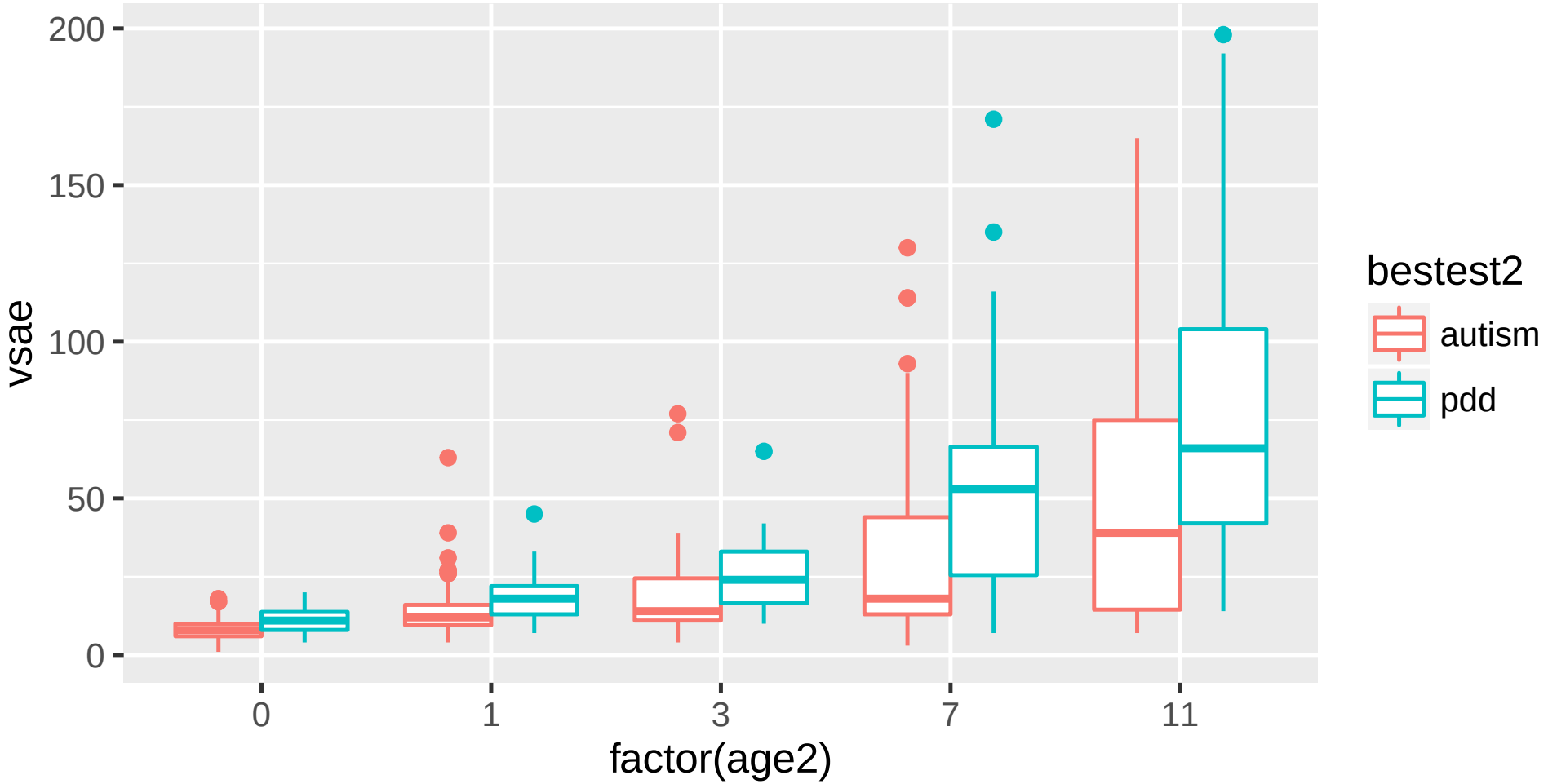
Take a look

Sample some profiles (age measured at 2, 3, 5, 9, 13)

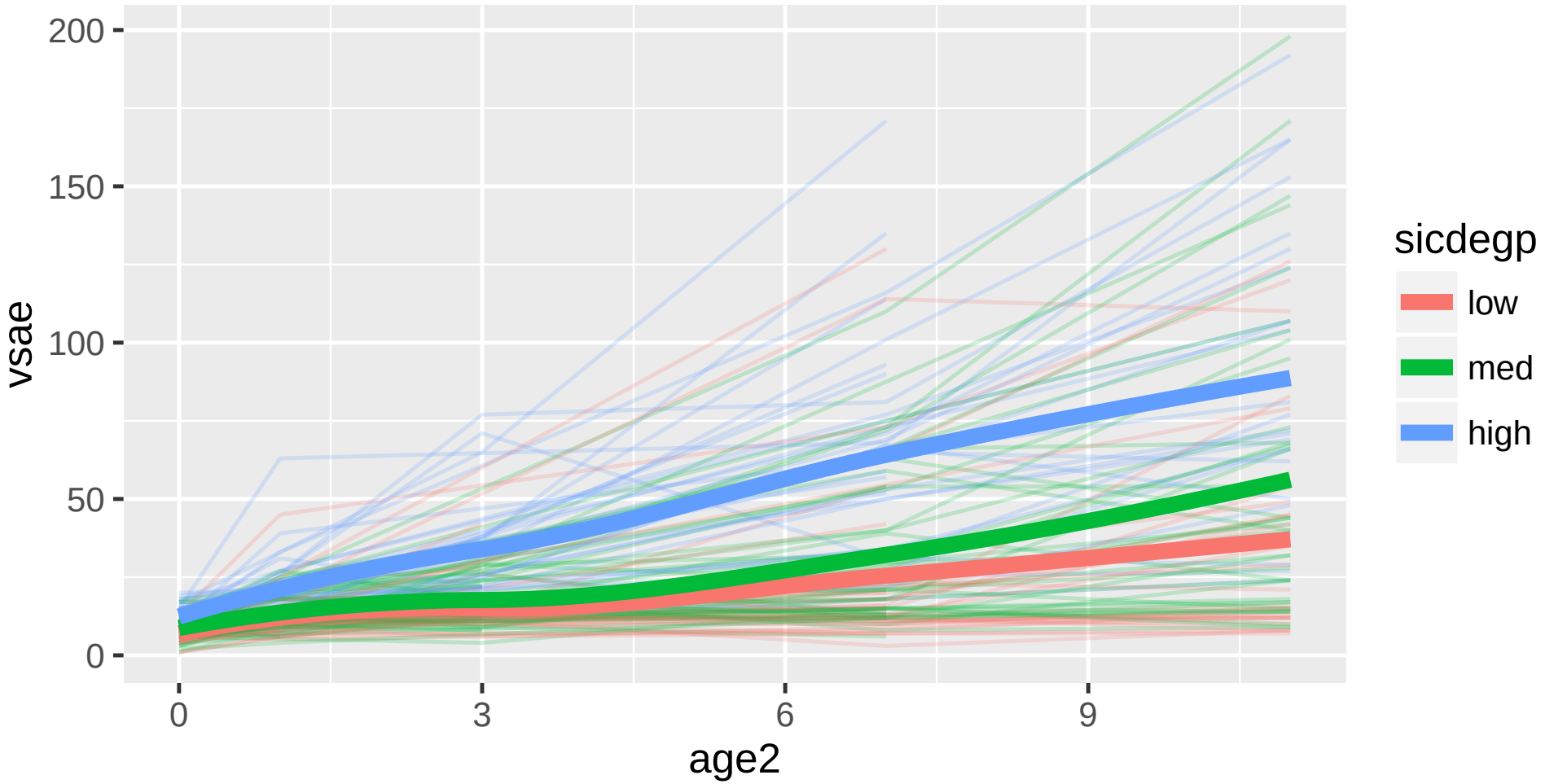
```
ggplot(filter(autism,
              childid %in% sample(unique(childid), 5)),
       aes(x=age2, y=vsae, group=childid)) + geom_line()
```



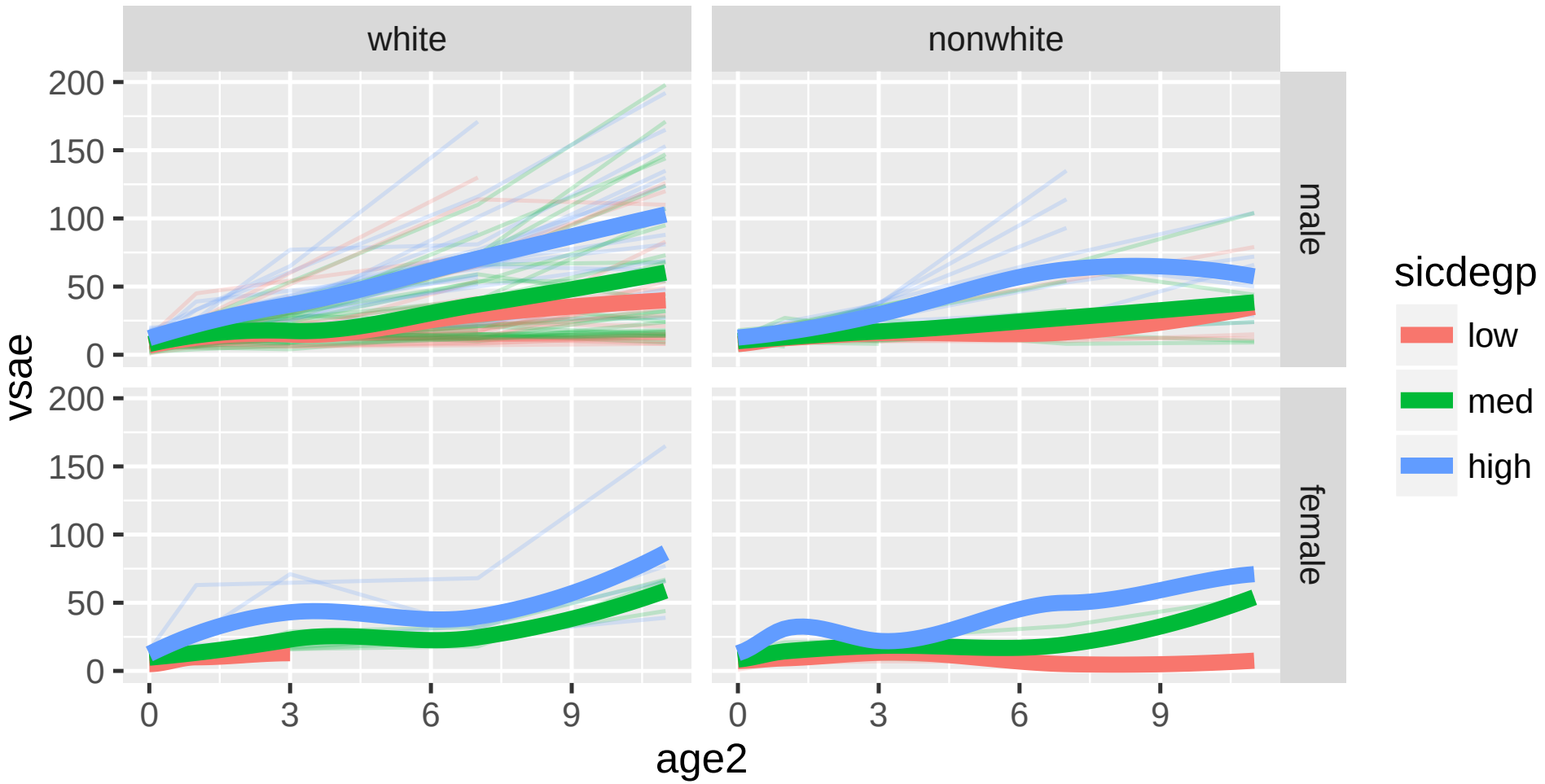

```
ggplot(autism, aes(x=factor(age2), y=vsae, colour=bestest2)) +
  geom_boxplot()
```



```
ggplot(autism, aes(x=age2, y=vsae, group=childid, colour=sicde
  geom_line(alpha=0.2) +
  geom_smooth(aes(group=sicdegp), se=F, size=2)
```




```
ggplot(autism, aes(x=age2, y=vsae, group=childid, colour=sicde
  geom_line(alpha=0.2) + facet_grid(gender~race) +
  geom_smooth(aes(group=sicdegp), se=F, size=2)
```



Pre-processing

Looks like sample sizes in sub-groups is small

```
autism %>% filter(age2==0) %>% group_by(gender, race) %>%
  tally() %>% spread(race, n)
# Source: local data frame [2 x 3]
# Groups: gender [2]
#
#   gender white nonwhite
#   (fctr) (int)    (int)
# 1   male    88       46
# 2 female    12        8
```

May not be enough information to incorporate these factors.

Your turn

Tabulate the number of measurements for each child. How many children have less than 3 measurements?

```
autism_keep <- autism %>% group_by(childid) %>%
  tally(sort=TRUE) %>% filter(n>2)
autism_sub <- autism %>%
  filter(childid %in% autism_keep$childid)
```

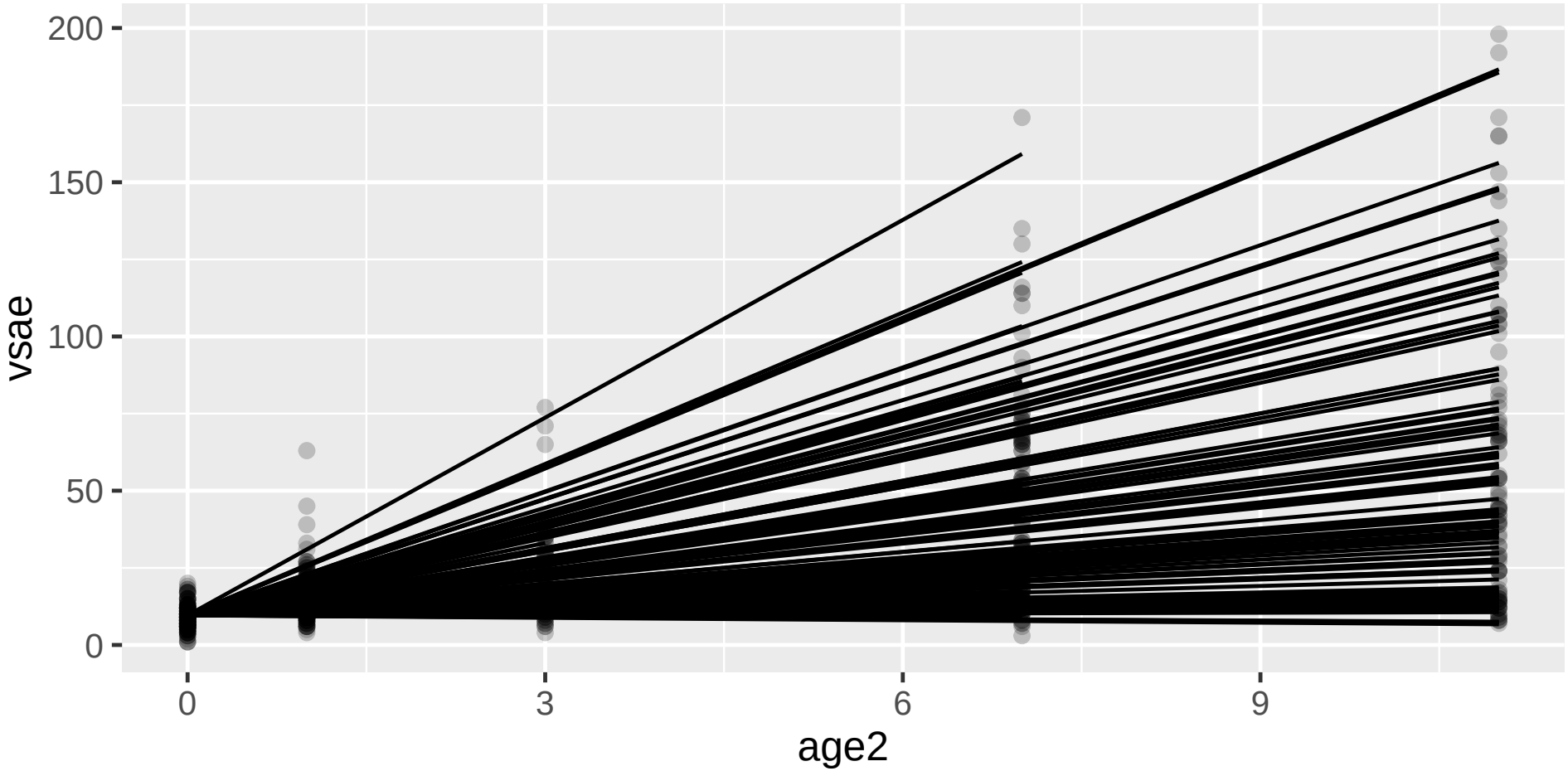
Model: random slopes

$$vsae = \beta_0 + \beta_1 age2 + b_1 childid + \varepsilon$$

```
autism_lmer <- lmer(vsae ~ age2 + ( age2 - 1 | childid ),
                    data = autism_sub)
summary(autism_lmer)
```

Take a look

```
autism_lmer_fit <- autism_sub
autism_lmer_fit$fit <- fitted(autism_lmer)
ggplot(autism_lmer_fit, aes(x=age2, y=vsae)) +
  geom_point(alpha=0.2) +
  geom_line(aes(y=fit, group=childid))
```

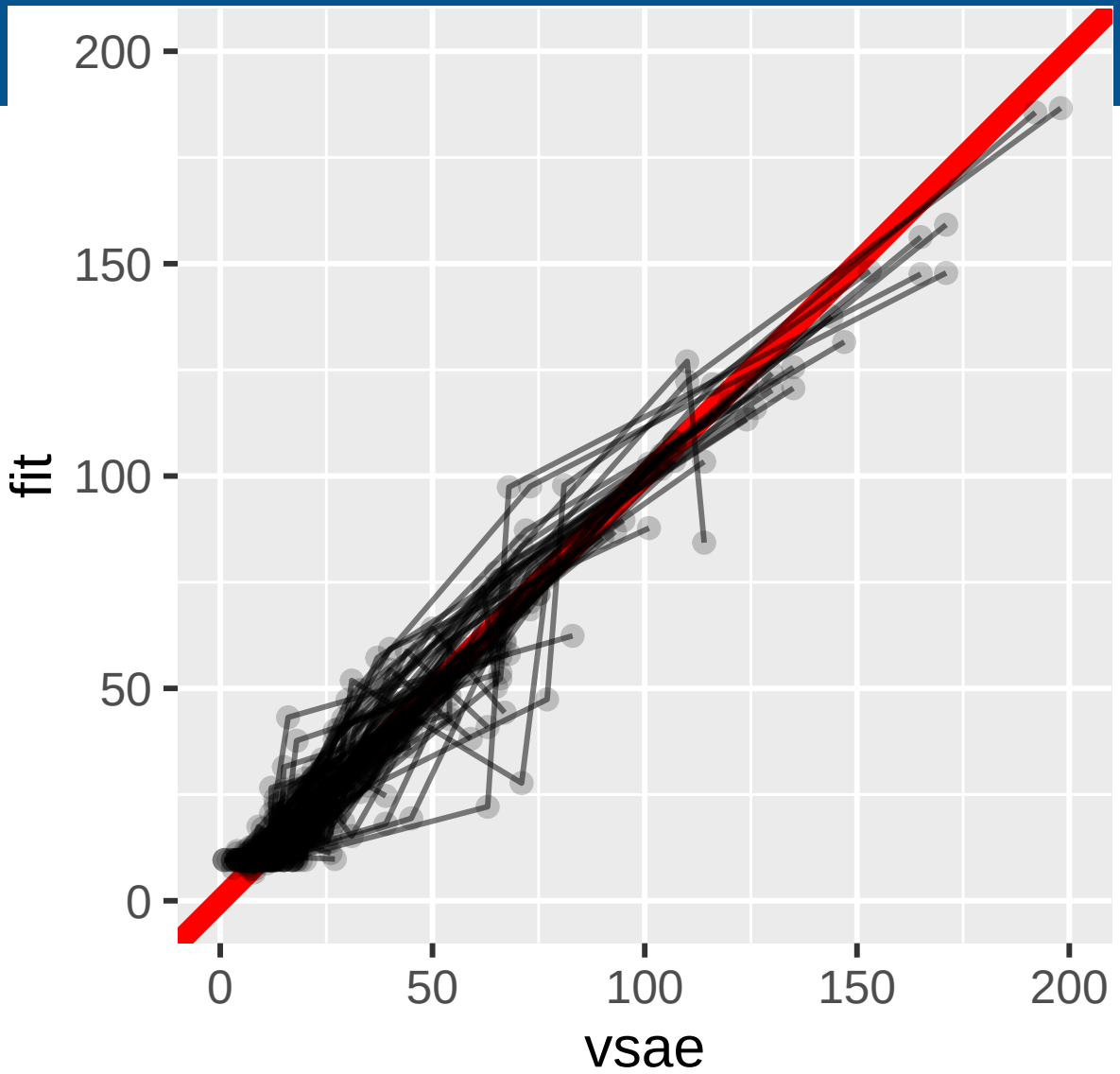


Your turn

What does `age2 - 1` do for the model fit?

Check: fitted vs observed

```
ggplot(autism_lmer_fit, aes(x=vsae, y=fit, group=childid)) +
  geom_abline(intercept=0, slope=1, color="red", size=2) +
  geom_point(alpha=0.2) + geom_line(alpha=0.5) +
  xlim(c(0, 200)) + ylim(c(0, 200)) +
  theme(aspect.ratio=1)
```



Fitted look like observed. No child stands out as being badly fit. Perhaps slightly nonlinear relationship.

Model building

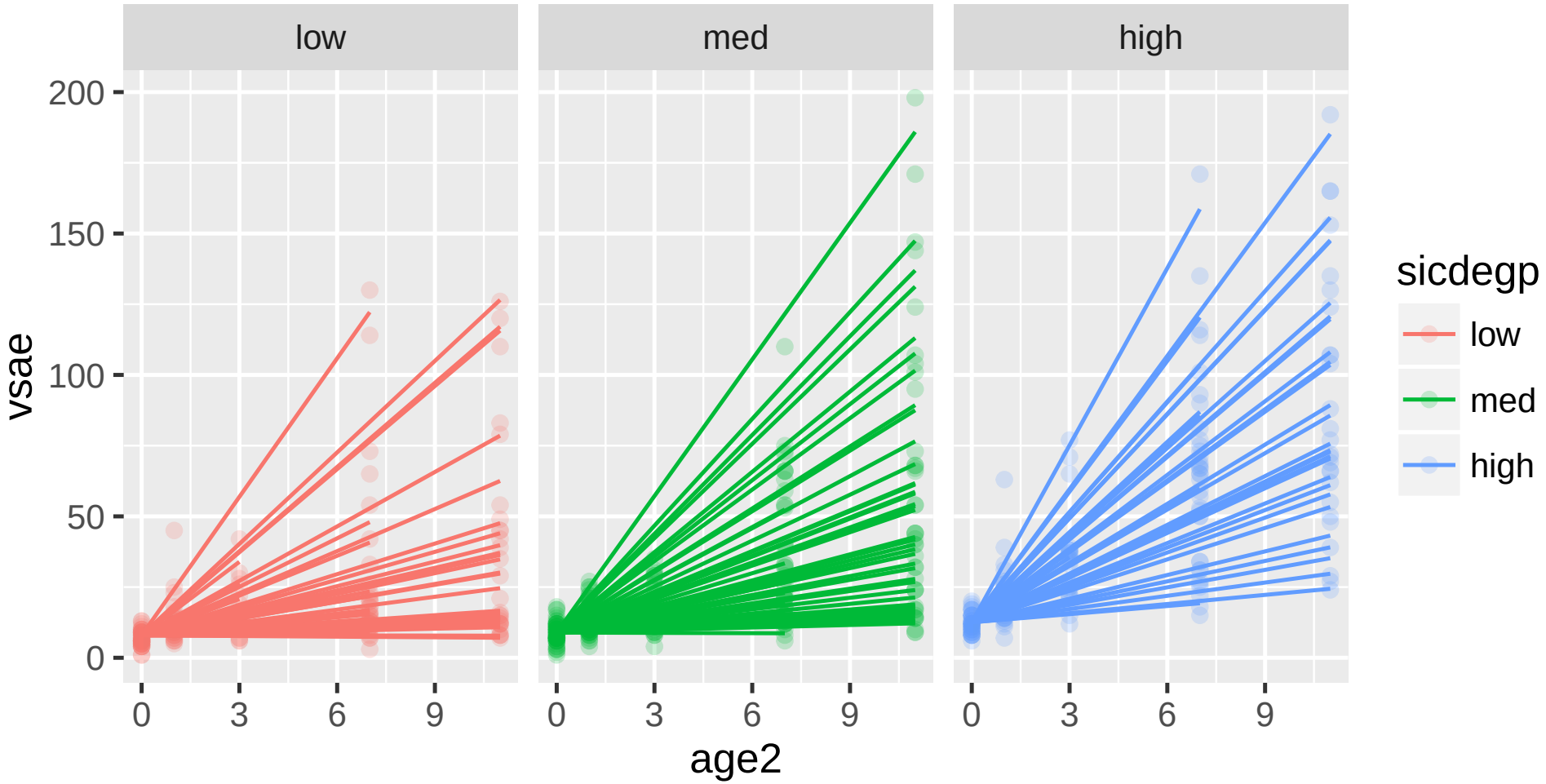
- Variables to include: sicdegp, bestest2, age, gender
- Compare models

+sicdegp

```
autism_lmer2 <- lmer(vsae ~ age2*sicdegp + ( age2 - 1 | childi
                     data = autism_sub)
summary(autism_lmer2)
```

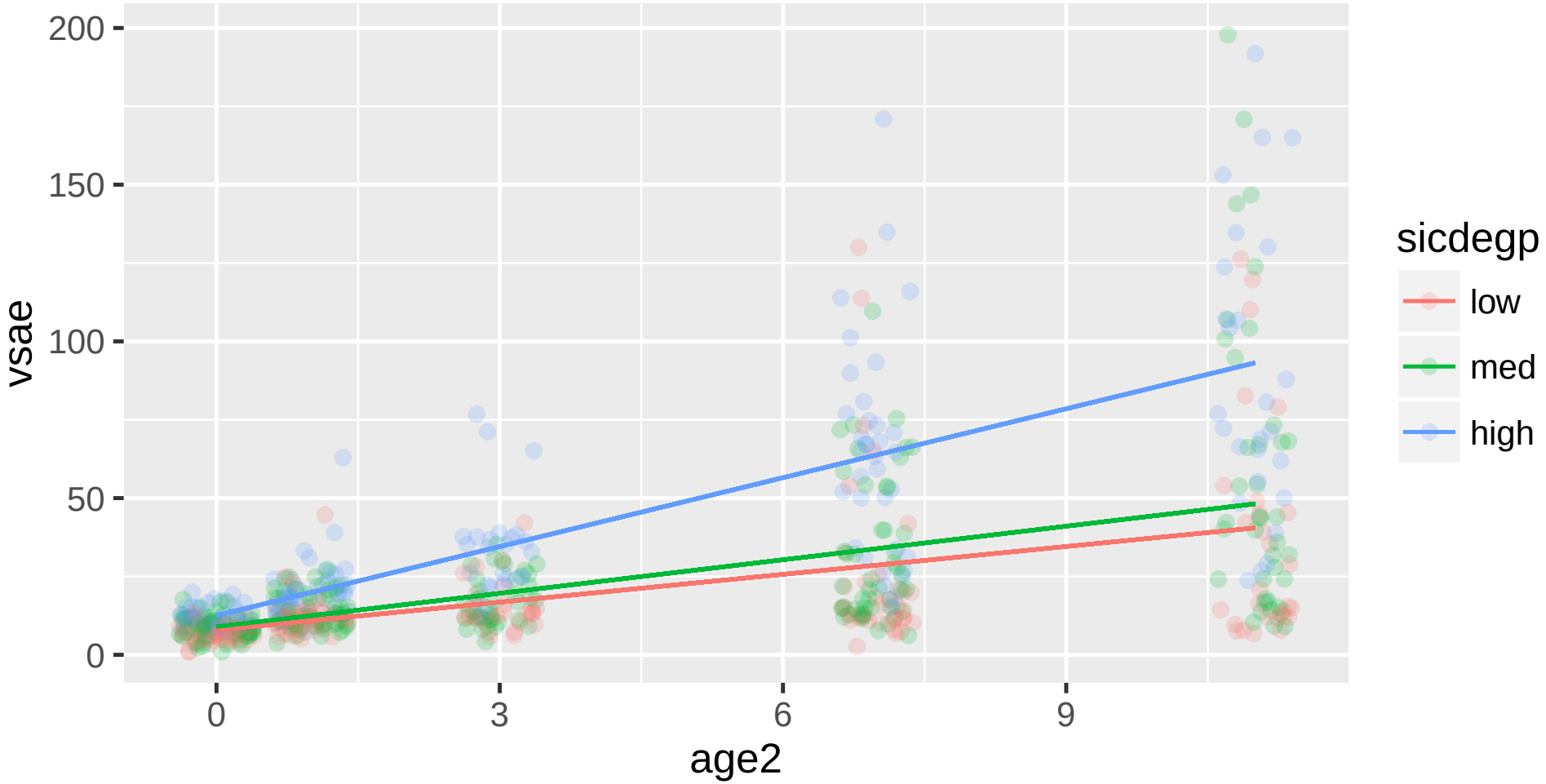
Take a look

```
autism_lmer2_fit <- autism_sub
autism_lmer2_fit$fit <- fitted(autism_lmer2)
ggplot(autism_lmer2_fit, aes(x=age2, y=vsae, colour=sicdegp))
  geom_point(alpha=0.2) + facet_wrap(~sicdegp) +
  geom_line(aes(y=fit, group=childid))
```



Fixed effects

```
autism_lmer2_fit <- augment(autism_lmer2)
ggplot(autism_lmer2_fit, aes(x=age2, y=vsae, colour=sicdegp))
  geom_jitter(alpha=0.2) +
  geom_line(aes(y=.fixed, group=childid))
```



Compare models

```
anova(autism_lmer, autism_lmer2)
# Data: autism_sub
# Models:
# autism_lmer: vsae ~ age2 + (age2 - 1 | childid)
# autism_lmer2: vsae ~ age2 * sicdegp + (age2 - 1 | childid)
#              Df  AIC  BIC logLik deviance Chisq Chi Df Pr(>C
# autism_lmer   4 4556 4574  -2274     4548
# autism_lmer2  8 4514 4549  -2249     4498  50.1      4    3.
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Your turn

Examine the remaining variables. Fit the best model that you can. Make a plot of the model.

Reference material

- R bloggers post: Getting Started with Mixed Effect Models in R
- Bates and Pinheiro "Mixed-Effects Models in S and S-PLUS"
- Gelman and Hill "Data Analysis Using Regression and Multilevel/Hierarchical Modeling"

Credits

Notes prepared by Di Cook, using material developed by Hadley Wickham, Heike Hofmann and Adam Loy.