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

- **Summarising and visualising models**

Assumptions

Recall:

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

- $\mathbf{b}_i$ is a random sample from $\mathcal{N}(\mathbf{0}, \mathbf{D})$ and independent from the level-1 error terms,
- ε_i follow a $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$ distribution
- $\mathbf{D}$ is a positive-definite $q \times q$ covariance matrix and $\mathbf{R}_i$ is a positive-definite $n_i \times n_i$ covariance matrix

Extract and examine level-1 residuals

```
radon_lmer_fit <- radon_sub
radon_lmer_fit$fit <- fitted(radon_lmer)
radon_lmer_fit$resid1 <-  HLMresid(radon_lmer,
            level=1)
ggplot(radon_lmer_fit, aes(x=resid1)) +
  geom_histogram(binwidth=0.5)
```

$\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$

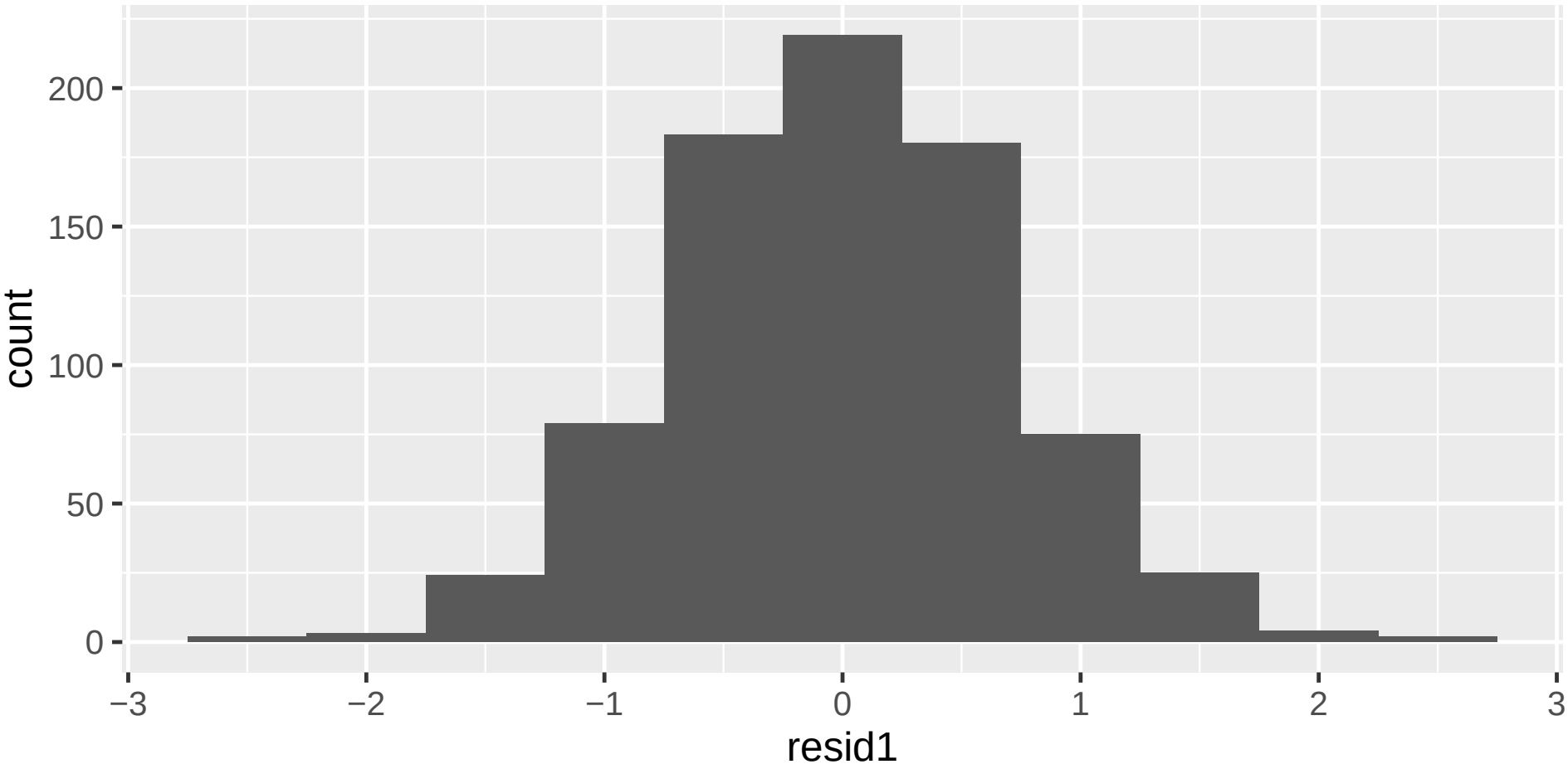
Your turn

For the radon data:

- What is p, q, g?
- And hence $n_i, i = 1, \ldots, g$?

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

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

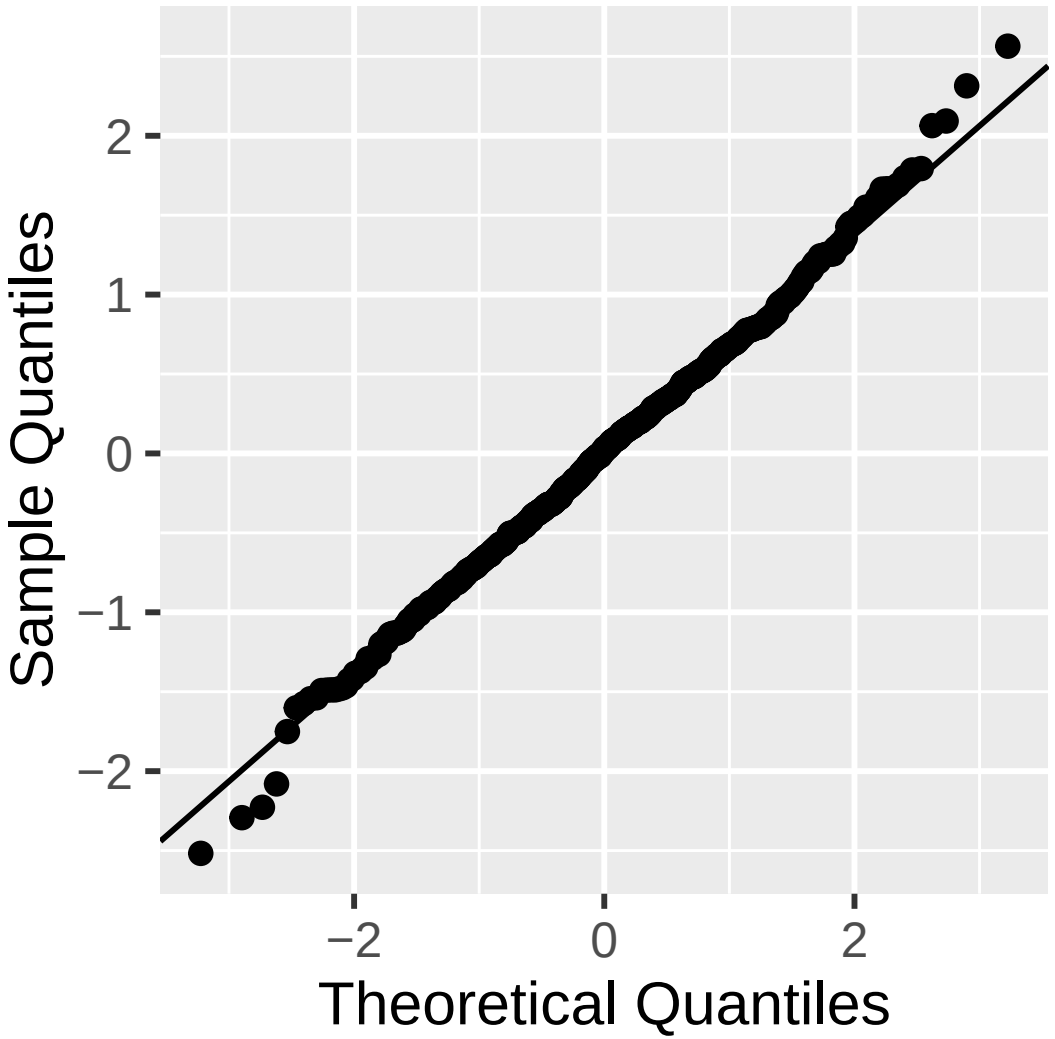


Level-1 (observation level) look normal.

Normal probability plot

```
ggplot_qqnorm(radon_lmer_fit$resid1, line="rlm") +
  theme(aspect.ratio=1)
```

```
ggplot_qqnorm(radon_lmer_fit$resid1, line="rlm") +
  theme(aspect.ratio=1)
```



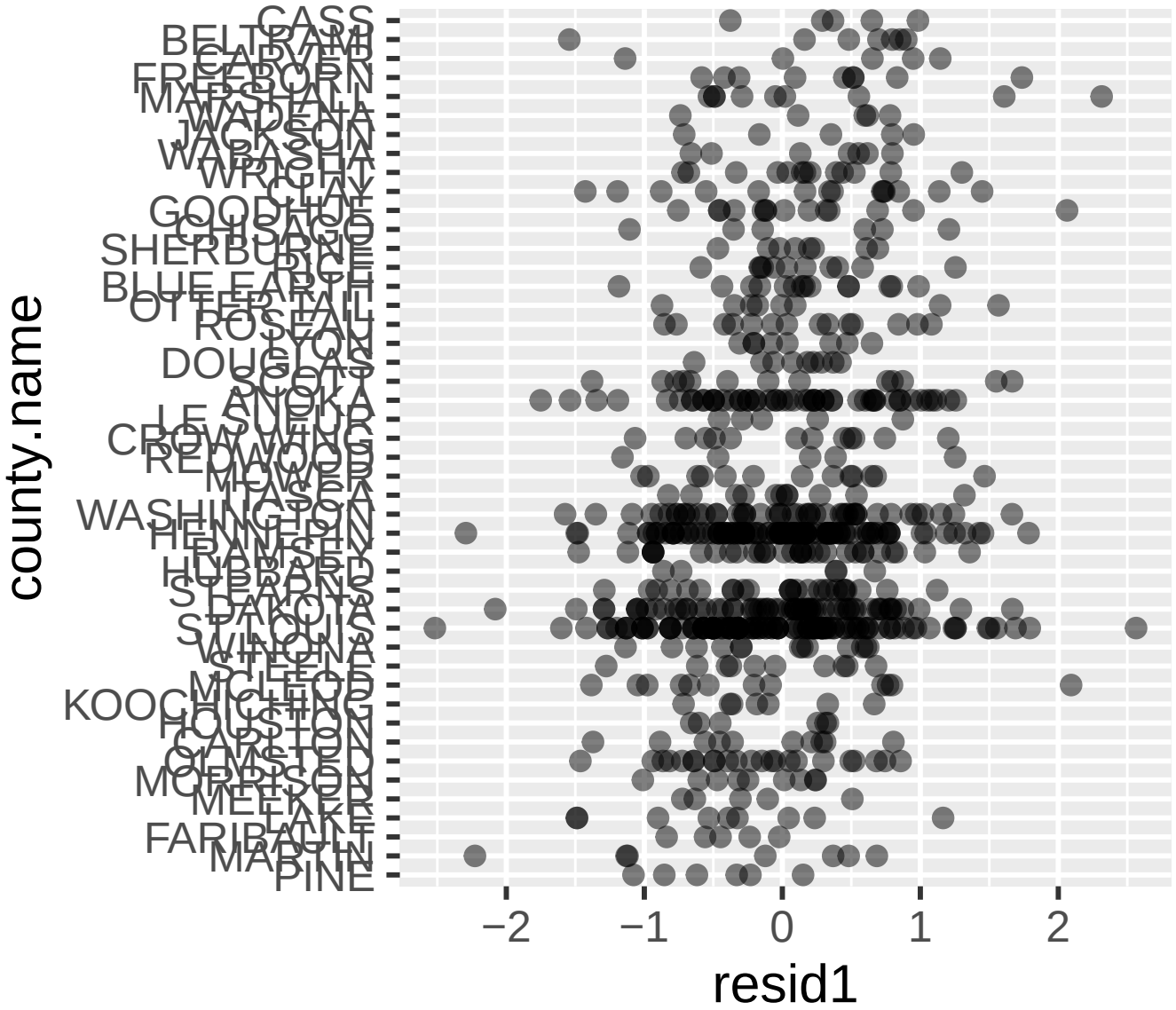
Level-1 (observation level) do nearly normal.

Examine within group

Summary statistics

```
radon_lmer_fit %>% group_by(county.name) %>%
  summarise(m = mean(resid1), s = sd(resid1), n = length(resid
  head()
# Source: local data frame [6 x 4]
#
#   county.name      m     s     n
#         (fctr)  (dbl) (dbl) (int)
# 1       ANOKA  0.051  0.72    52
# 2    BELTRAMI  0.335  0.87     7
# 3  BLUE EARTH  0.152  0.56    14
# 4     CARLTON -0.194  0.65    10
# 5      CARVER  0.322  0.92     5
# 6        CASS  0.383  0.50     5
```

```
ggplot(radon_lmer_fit, aes(x=county.name, y=resid1)) +
  geom_point(alpha=0.5) + coord_flip()
```



Normality tests

Anderson-Darling, Cramer-von Mises, Lilliefors (Kolmogorov-Smirnov)

```
library("nortest")
ad.test(radon_lmer_fit$resid1)
cvm.test(radon_lmer_fit$resid1)
lillie.test(radon_lmer_fit$resid1)
```

```
#
#   Anderson-Darling normality test
#
# data:  radon_lmer_fit$resid1
# A = 0.4, p-value = 0.4
```

all believe that the residuals are consistent with normality.

Conclusion about level-1 residuals

The assumption:

$$\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}_i)$$

is probably ok, at the worst it is not badly violated.

Random effects

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad i = 1, \ldots g$$

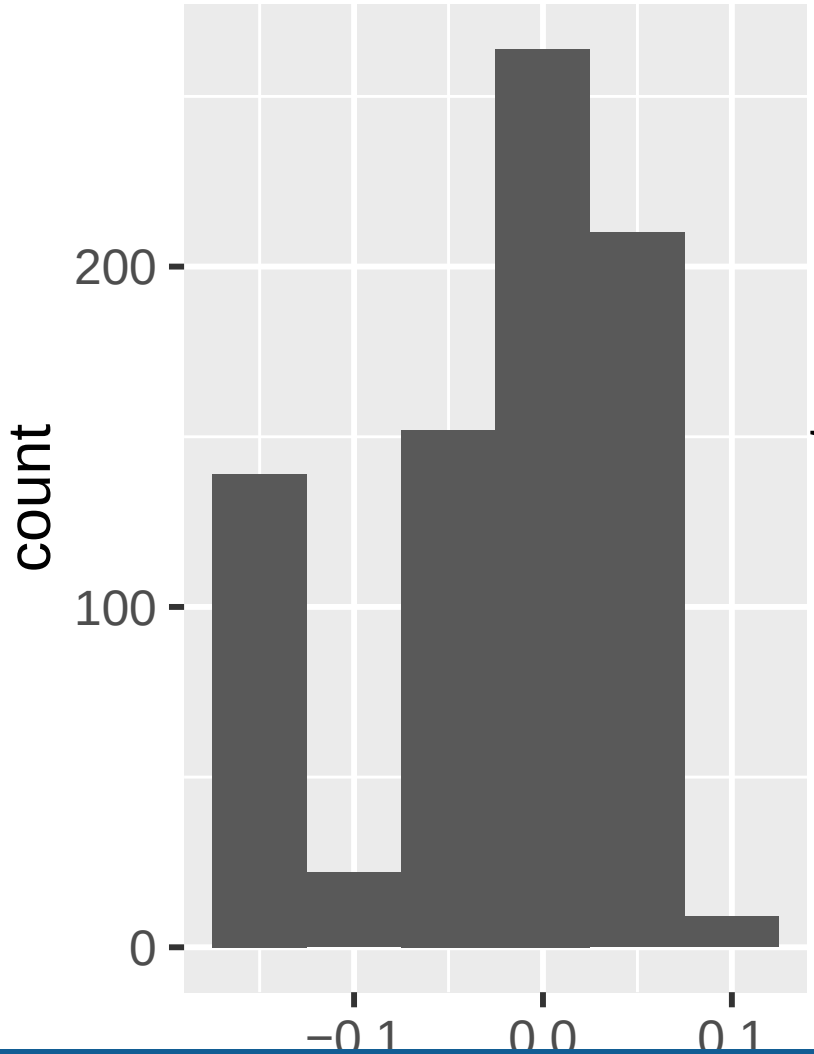
where $\mathbf{D}$ allows for correlation between random effects within group, and these should be independent from the level-1 error

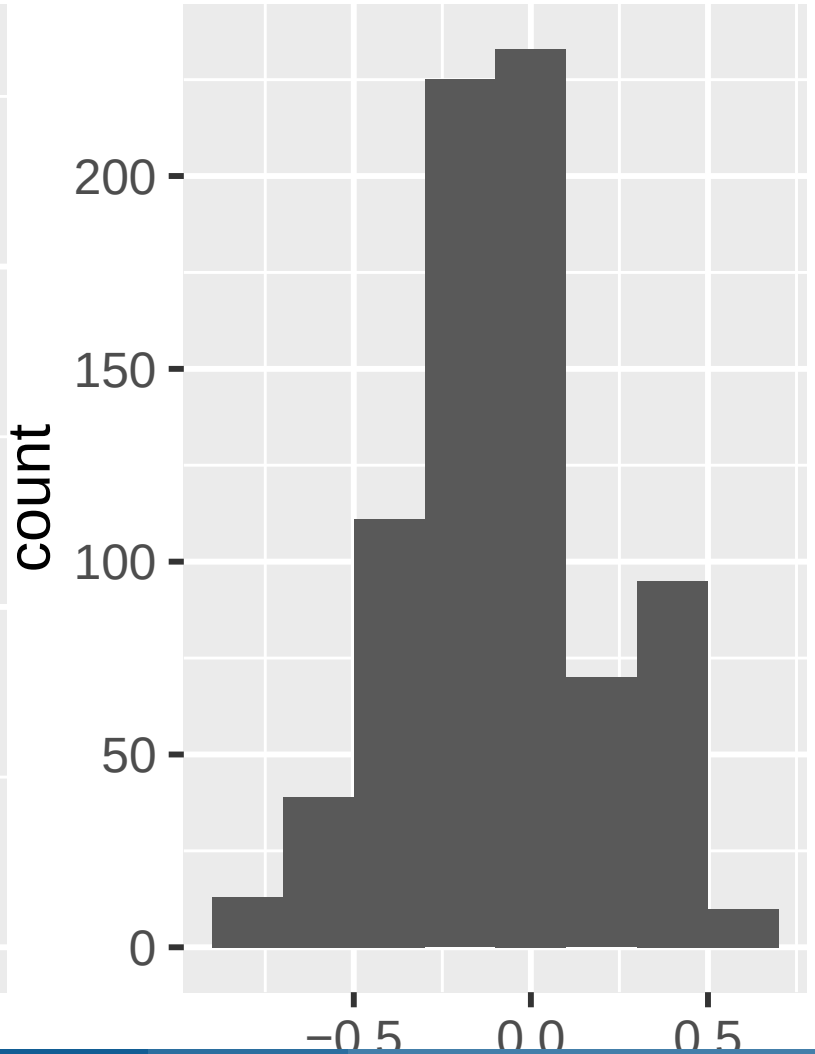
Level 2 (random effects)

```
rf <- HLMresid(radon_lmer, level="county.name")
# same as ranef(radon_lmer)
rf$county.name <- rownames(rf)
rf <- rf %>% rename(resid.basement=`(Intercept)`,
                    resid.ff=`basementfirst floor`)
radon_lmer_fit <- merge(radon_lmer_fit, rf,
                        by="county.name")
```

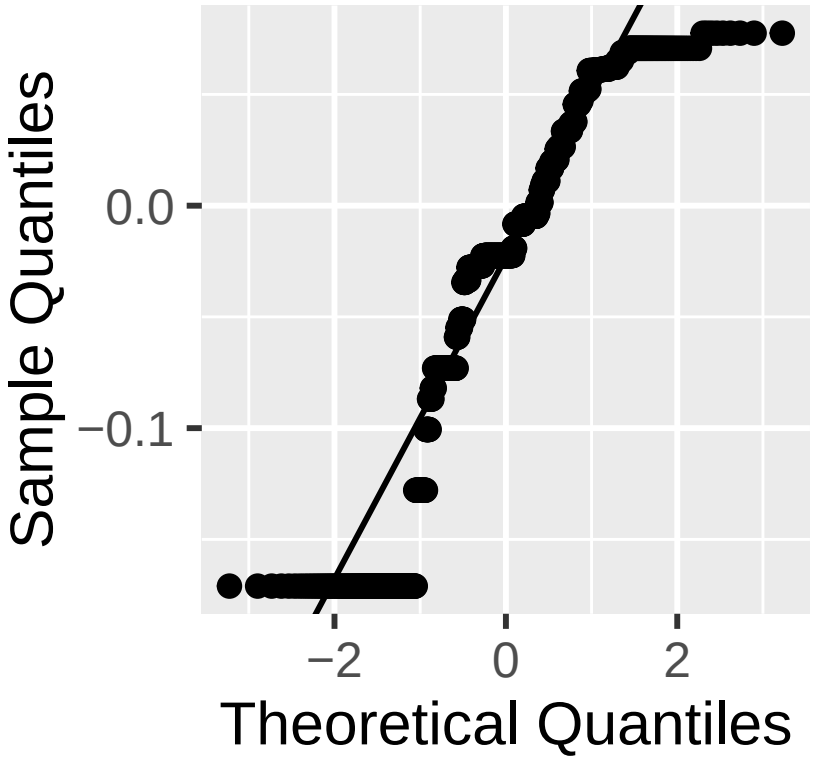
We have both intercepts (basement) and slopes (first floor)

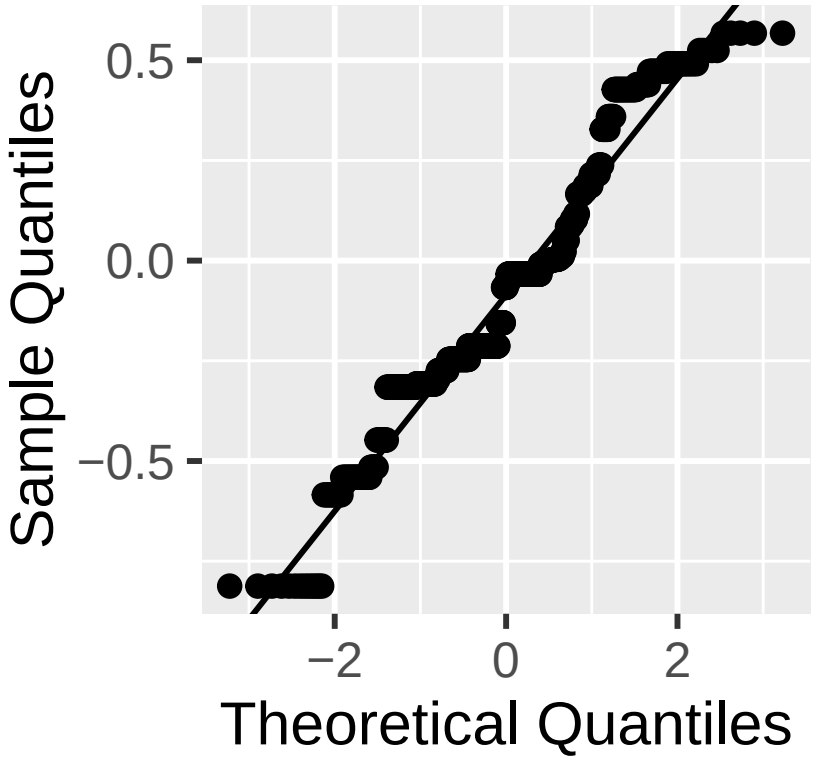
```
ggplot(radon_lmer_fit, aes(x=resid.basement)) +
  geom_histogram(binwidth=0.05)
ggplot(radon_lmer_fit, aes(x=resid.ff)) +
  geom_histogram(binwidth=0.2)
```





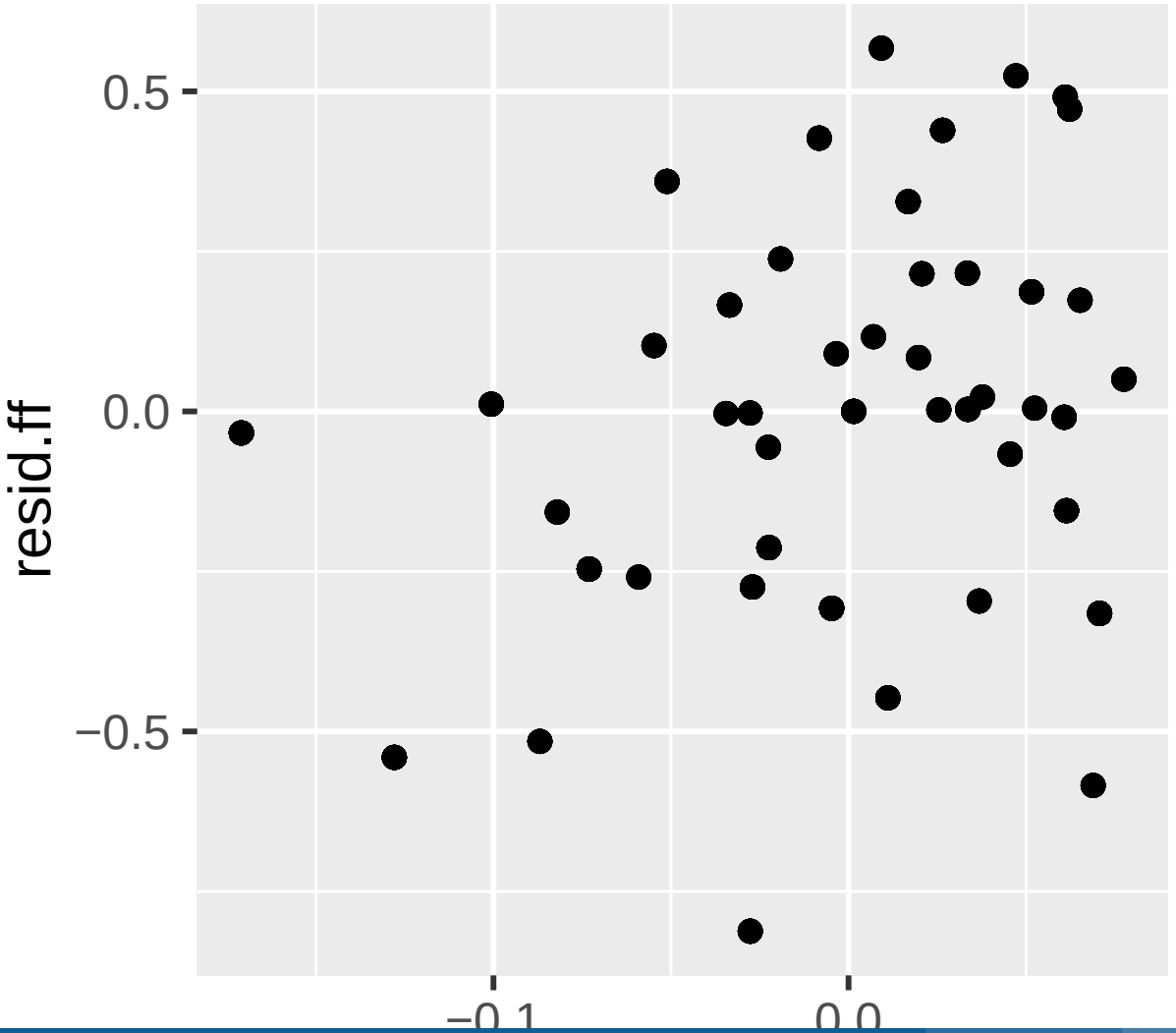
```
ggplot_qqnorm(radon_lmer_fit$resid.basement, line="rlm") +
  theme(aspect.ratio=1)
ggplot_qqnorm(radon_lmer_fit$resid.ff, line="rlm") +
  theme(aspect.ratio=1)
```





Should be no correlation

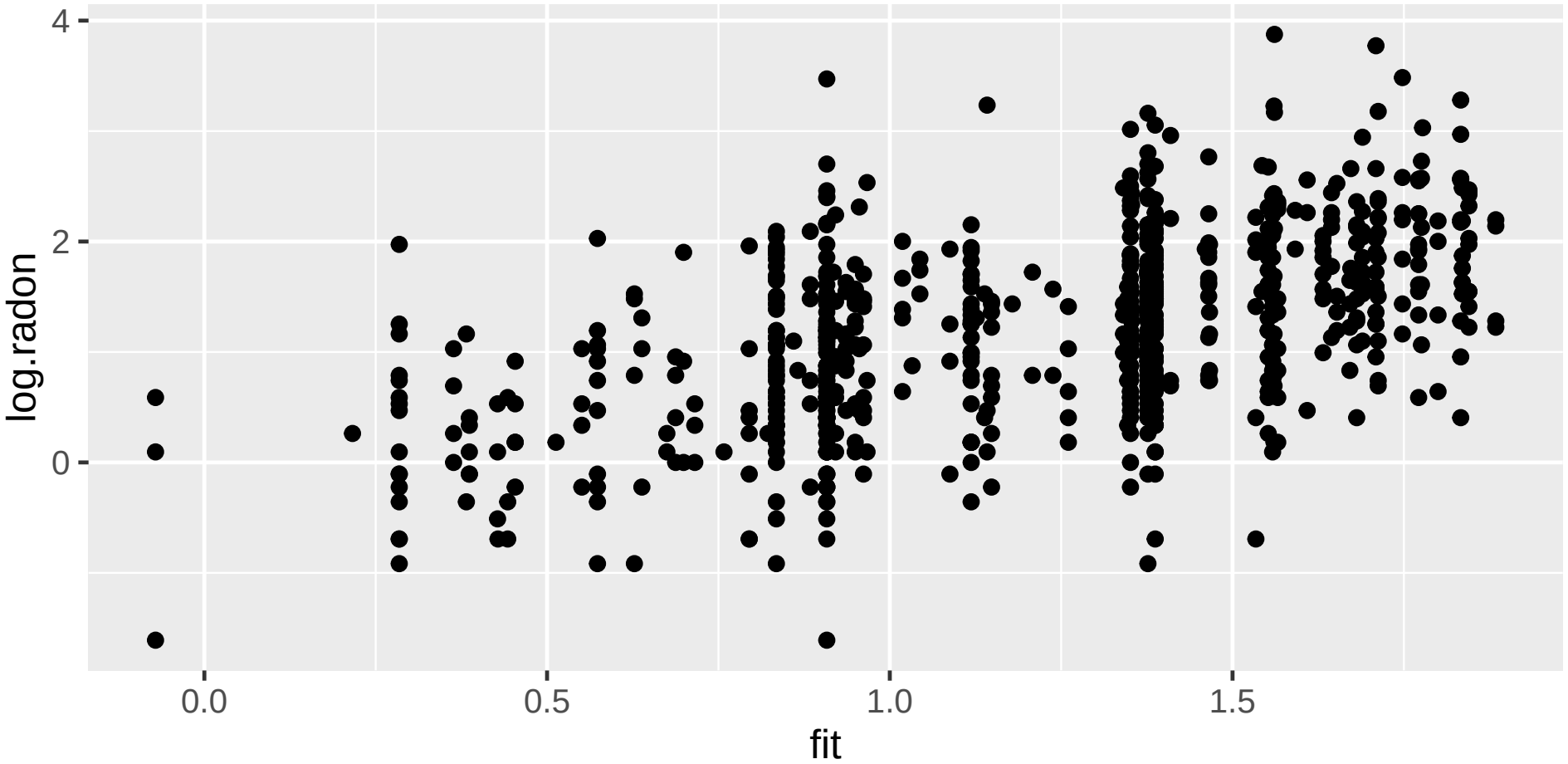
```
ggplot(radon_lmer_fit, aes(x=resid.basement, y=resid.ff)) +
  geom_point() + theme(aspect.ratio=1)
```



Fitted vs Observed

Plotting the observed vs fitted values, gives a sense for how much of the response is explained by the model. Here we can see that there is still a lot of unexplained variation.

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



Autism model

```
autism_lmer4 <- lmer(vsae ~ age2*sicdegp +
                        age2*bestest2 + age2*race +
                        ( age2 - 1 | childid ),
                     data = autism_sub)
autism_lmer_fit <- augment(autism_lmer4)
```

Your turn

- What is p, q, g?
- And hence $n_i, i = 1, \ldots, g$?

Your turn

Write down the model statement that corresponds to the R code fit:

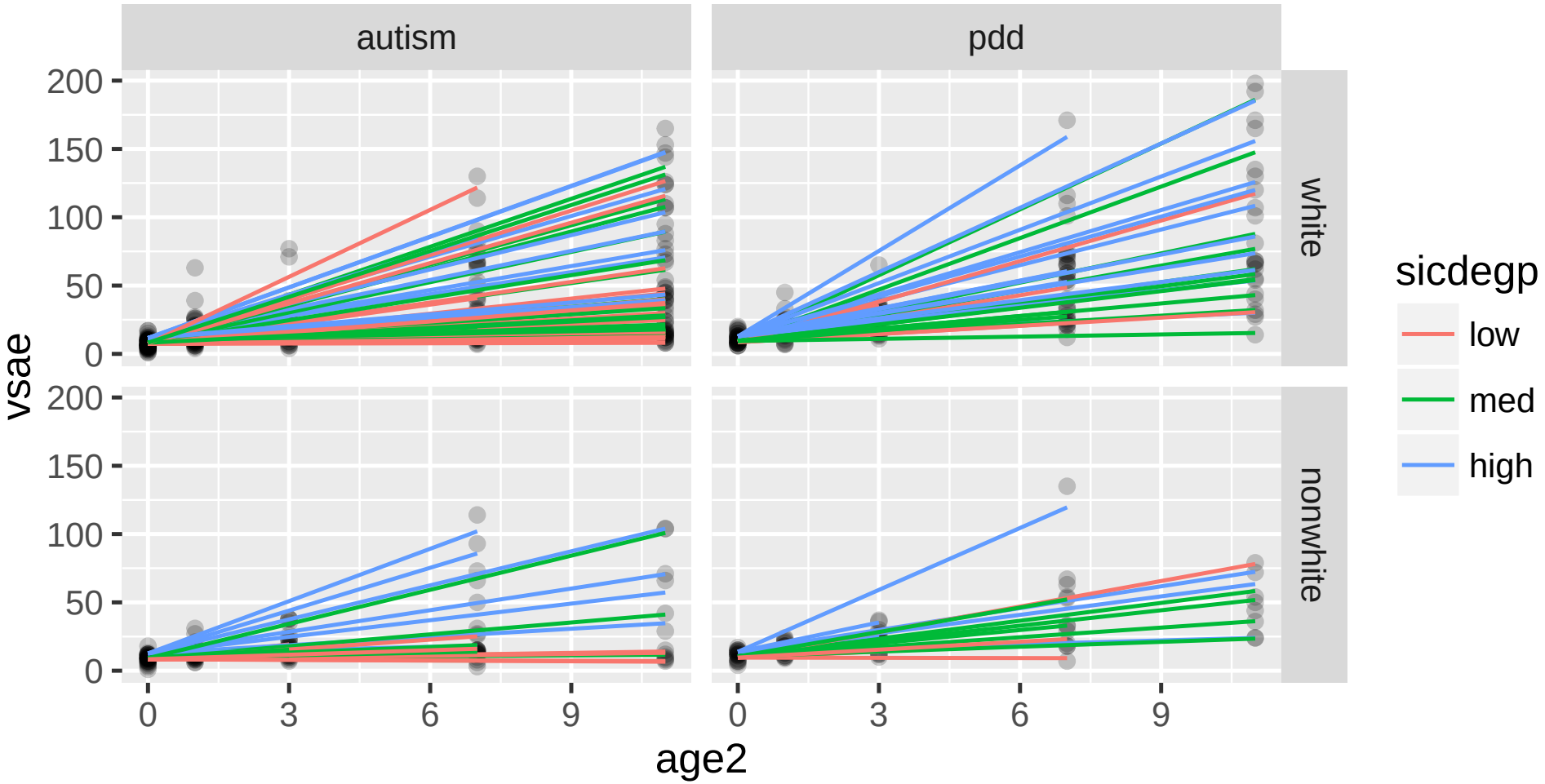
```
autism_lmer4 <- lmer(vsae ~ age2*sicdegp +
                        age2*bestest2 + age2*race +
                        ( age2 - 1 | childid ),
                     data = autism_sub)
```

Your turn

What does the function `augment` do? (Hint: it is in the `broom` package.)

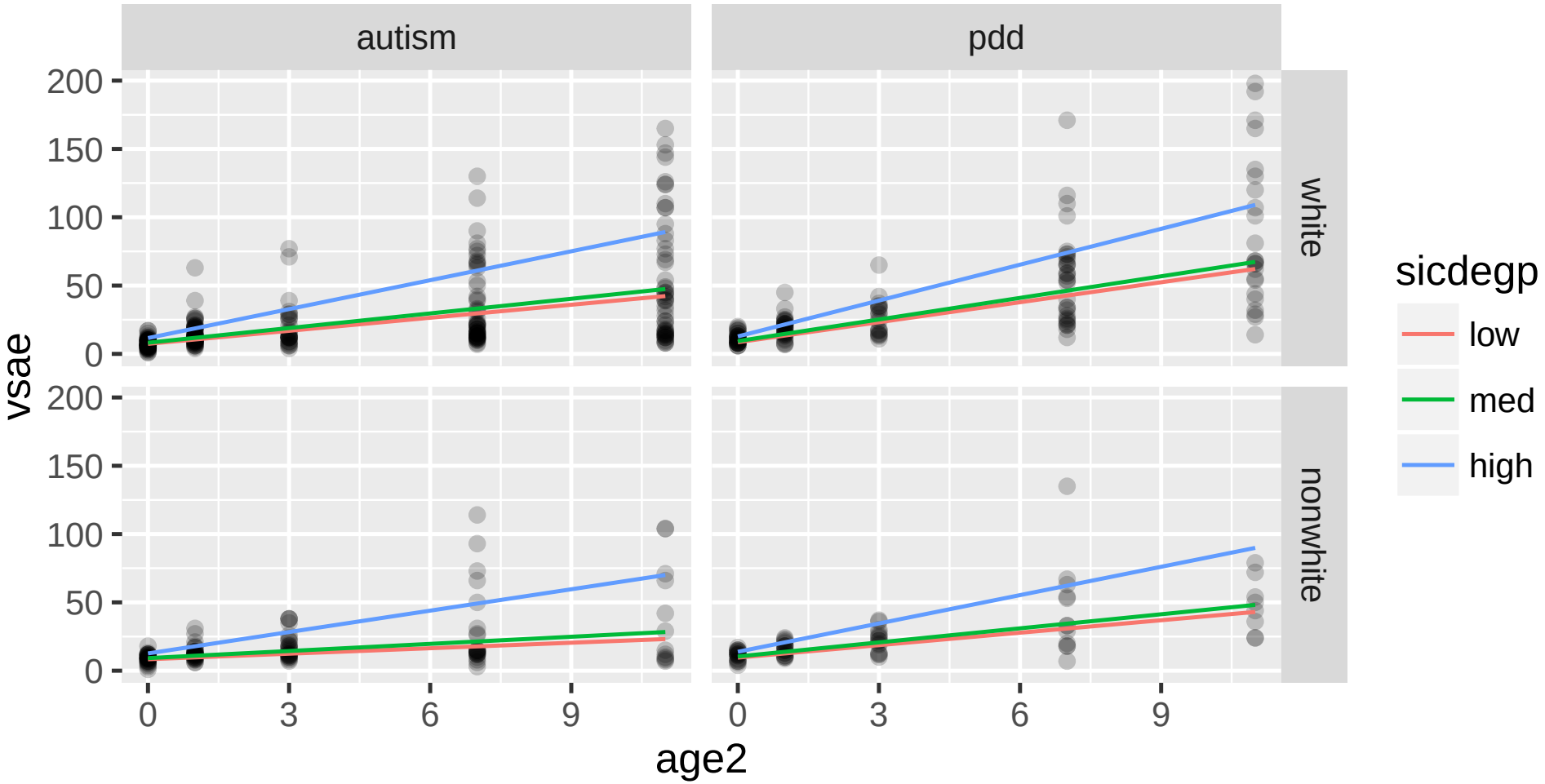
Plot the model: random effects

```
ggplot(autism_lmer_fit, aes(x=age2, y=vsae)) +
  geom_point(alpha=0.2) +
  geom_line(aes(y=.fitted, group=childid, color=sicdegp)) +
  facet_grid(race~bestest2)
```



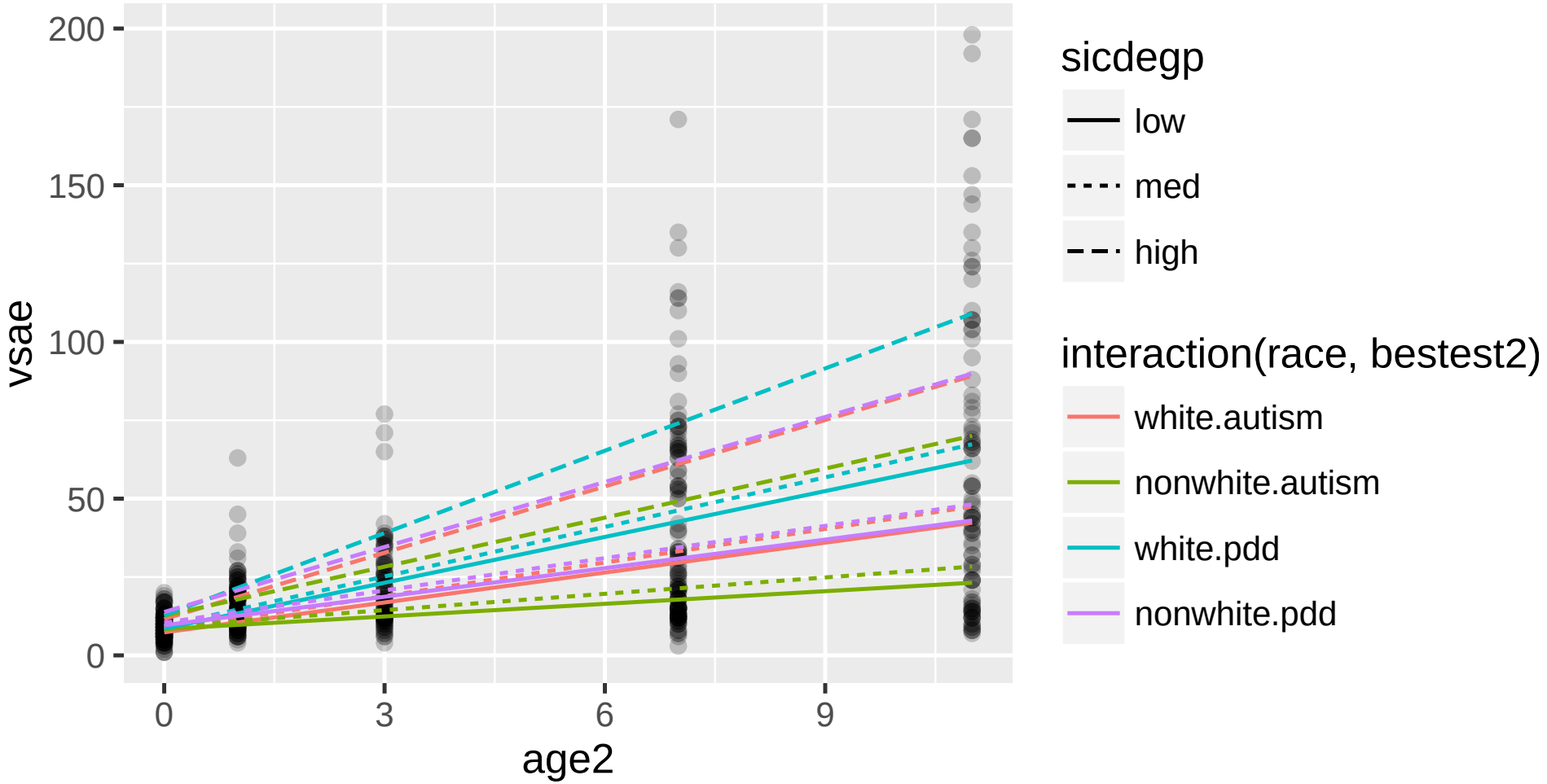
Fixed effects

```
ggplot(autism_lmer_fit, aes(x=age2, y=vsae)) +
  geom_point(alpha=0.2) +
  geom_line(aes(y=.fixed, color=sicdegp)) +
  facet_grid(race~bestest2)
```



Fixed effects

```
ggplot(autism_lmer_fit, aes(x=age2, y=vsae)) +
  geom_point(alpha=0.2) +
  geom_line(aes(y=.fixed, color=interaction(race,bestest2), li
```



Your turn

- Compute the level-1 residuals
- Conduct a Lilliefors test of normality
- Do the residuals look normal?

Your turn

- Plot the observed vs fitted data
- Does the model explain a substantial amount of the variation?

Diagnostics of influence, outlier detection

- Leave-one-out statistics form the basis of diagnostics. Examine the change in the model estimates with and without the case.
- For multilevel models, there are multiple levels of removal. Could be one of the group level structures, or individuals within groups
- The `HMLdiag` package makes it easy to compute and examine these, e.g. Cooks distance, mdffits, leverage

Exam data from mlmRev package

```
library("mlmRev")
glimpse(Exam)
# Observations: 4,059
# Variables: 10
# $ school   (fctr) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
# $ normexam (dbl) 0.26, 0.13, -1.72, 0.97, 0.54, 1.73, 1.04,
# $ schgend  (fctr) mixed, mixed, mixed, mixed, mixed, mixed,
# $ schavg   (dbl) 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0.17, 0
# $ vr       (fctr) mid 50%, mid 50%, mid 50%, mid 50%, mid 50
# $ intake   (fctr) bottom 25%, mid 50%, top 25%, mid 50%, mid
# $ standLRT (dbl) 0.619, 0.206, -1.365, 0.206, 0.371, 2.189,
# $ sex      (fctr) F, F, M, F, F, M, M, M, F, M, M, M, M, M,
# $ type     (fctr) Mxd, Mxd, Mxd, Mxd, Mxd, Mxd, Mxd, Mxd, Mx
# $ student  (fctr) 143, 145, 142, 141, 138, 155, 158, 115, 11
```

Fit model

```
fm4 <- lmer(normexam ~ standLRT + I(standLRT^2) +
              I(standLRT^3) + sex + schgend + schavg +
              (standLRT | school), data = Exam, REML = FALSE)
fm4
```

Fixed effects

```
Fixed effects:
                 Estimate Std. Error t value
(Intercept)    -0.019994   0.049379  -0.405
standLRT        0.594155   0.025732  23.090
I(standLRT^2)   0.013283   0.009004   1.475
I(standLRT^3) -0.014178   0.005503  -2.577
sexM          -0.170294   0.033830  -5.034
schgendboys     0.186847   0.094959   1.968
schgendgirls    0.164680   0.075212   2.190
schavg          0.287547   0.098420   2.922
```

Random effects

```
Random effects:
 Groups   Name         Variance Std.Dev. Corr
 school   (Intercept) 0.06507  0.2551
          standLRT    0.01419  0.1191   0.48
 Residual             0.54904  0.7410
Number of obs: 4059, groups:  school, 65
```

Model summary

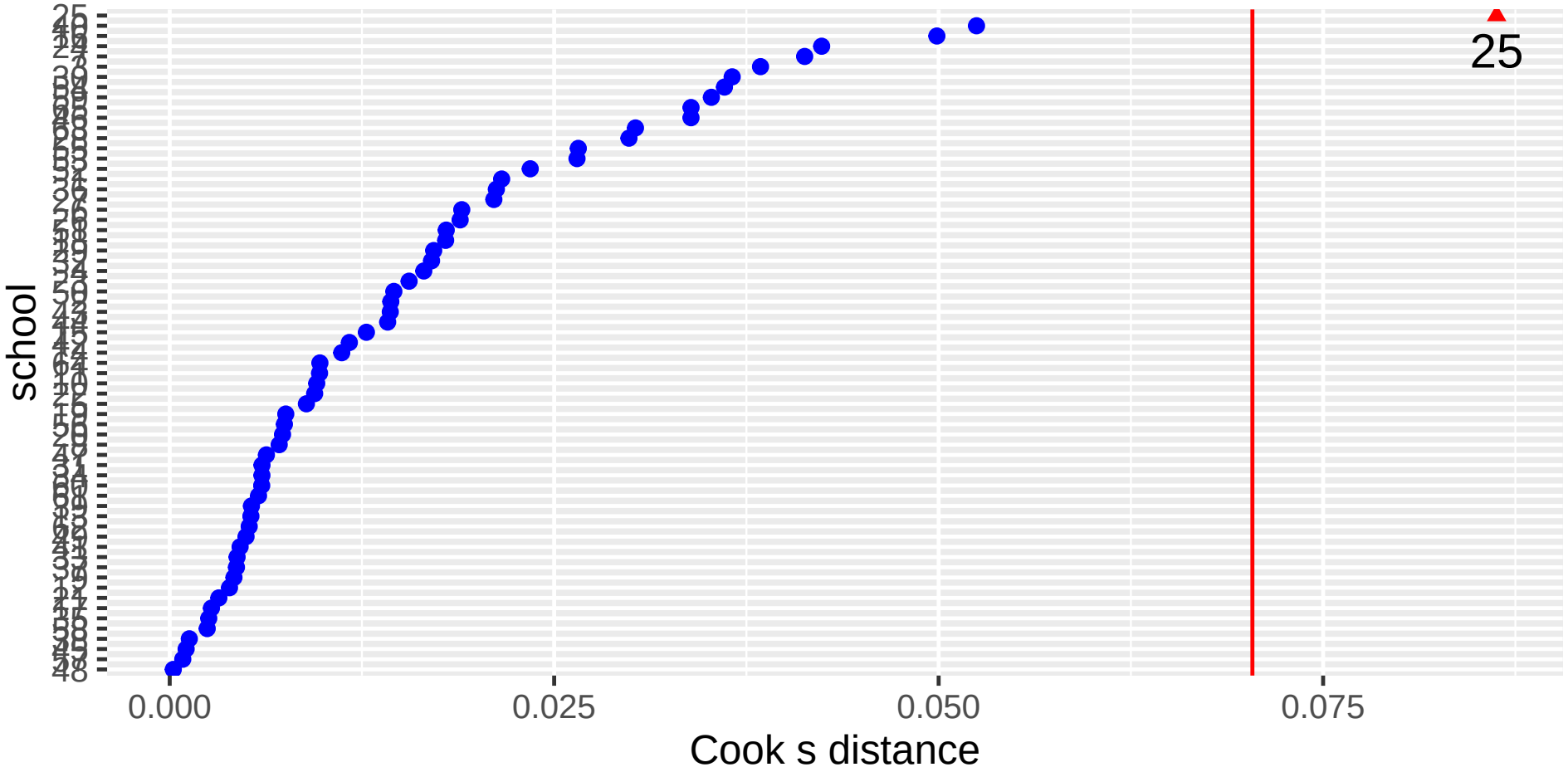
```
     AIC       BIC   logLik deviance df.resid
  9289.3    9365.0  -4632.6    9265.3     4047

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8524 -0.6301  0.0238  0.6853  3.4516
```

Cooks distance

Measures difference in standardised residual, based on prediction with model computed with and without the unit

```
library("HLMdiag")
cooksd_fm4  <- cooks.distance(fm4, group = "school")
dotplot_diag(x = cooksd_fm4, cutoff = "internal",
   name = "cooks.distance") + ylab("Cook s distance") + xlab('
```



This would indicate that school 25 is an anomaly.

mdffits

The statistic *dffits* measures the difference in the fitted value, with and without the unit.

```
mdffits_fm4 <- mdffits(fm4, group = "school")
sort(mdffits_fm4)
#   [1] 0.00023 0.00084 0.00105 0.00125 0.00240 0.00250 0.0025
#   [9] 0.00383 0.00408 0.00425 0.00429 0.00429 0.00488 0.0050
#  [17] 0.00514 0.00561 0.00588 0.00589 0.00592 0.00616 0.0066
#  [25] 0.00731 0.00734 0.00868 0.00871 0.00872 0.00910 0.0093
#  [33] 0.01140 0.01229 0.01334 0.01365 0.01410 0.01412 0.0151
#  [41] 0.01596 0.01609 0.01712 0.01725 0.01739 0.01762 0.0191
#  [49] 0.02100 0.02179 0.02560 0.02599 0.02892 0.02949 0.0324
#  [57] 0.03302 0.03467 0.03598 0.03613 0.03786 0.03940 0.0468
#  [65] 0.08106
```

Influence

```
leverage_fm4 <- leverage(fm4, level = "school")
head(leverage_fm4)
#   overall  fixef ranef ranef.uc
# 1   0.022 0.0019 0.020     0.16
# 2   0.027 0.0024 0.024     0.18
# 3   0.026 0.0026 0.023     0.17
# 4   0.020 0.0016 0.018     0.15
# 5   0.031 0.0019 0.029     0.14
# 6   0.018 0.0019 0.016     0.18
```

Reference material

- Loy and Hofmann (2015) "Are You Normal? The Problem of Confounded Residual Structures in Hierarchical Linear Models", Journal of Computational and Graphical Statistics
- Loy and Hofmann (2014) HLMdiag: A Suite of Diagnostics for Hierarchical Linear Models in R, Journal of Statistical Software

Credits

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