Summer School Overview

- Day 0: R bootcamp
- Day 1: Workflow, Google App Engine
- Day 2: Online Experiments
- Day 3: Data wrangling, visualization
- Day 4: Statistics, Probabilistic models
- Day 5: Experience sampling

Packages and programs

Please install the lme4, brms, tidybayes and BayesFactor packages in R, along with JAGS (see link on resources page of website)

Announcements

Day 4 materials

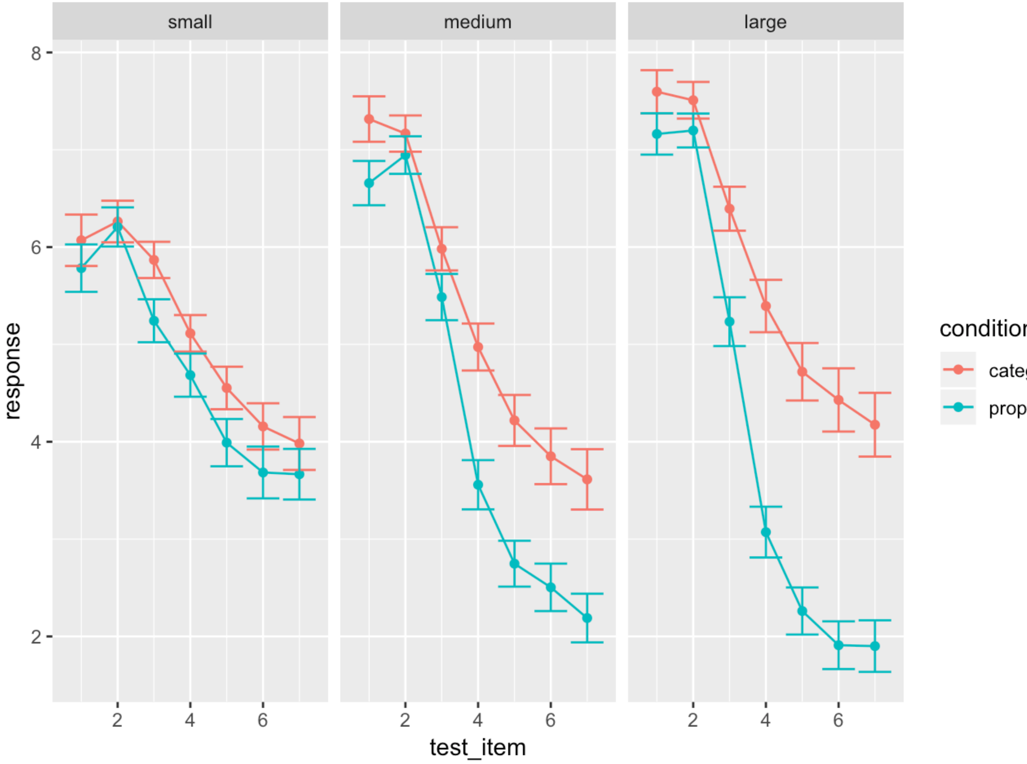
- Update your copy of the `chdss2019_content` repository

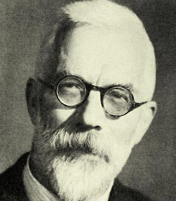
(type git pull at the terminal when working directory is Desktop/chdss2019_content)

Open `chdss2019_content.Rproj`

Goals

1. Introduce some statistical concepts, including Bayesian approaches and mixed effects models
2. Work towards a statistical analysis of the sampling frames data





Classical tests

- The `t.test()` function handles one-sample, independent samples and paired samples t-tests
- The `chisq.test()` function handles chi-square tests of independence and Pearson goodness of fit tests
- The `prop.test()` function tests for the equality of two proportions.
- The `binom.test()` function allows you to do a binomial test of choice proportion against a known rate
- The `wilcox.test()` function handles one- and two-sample nonparametric tests of equality of means
- The `cor.test()` function tests the significance of a correlation

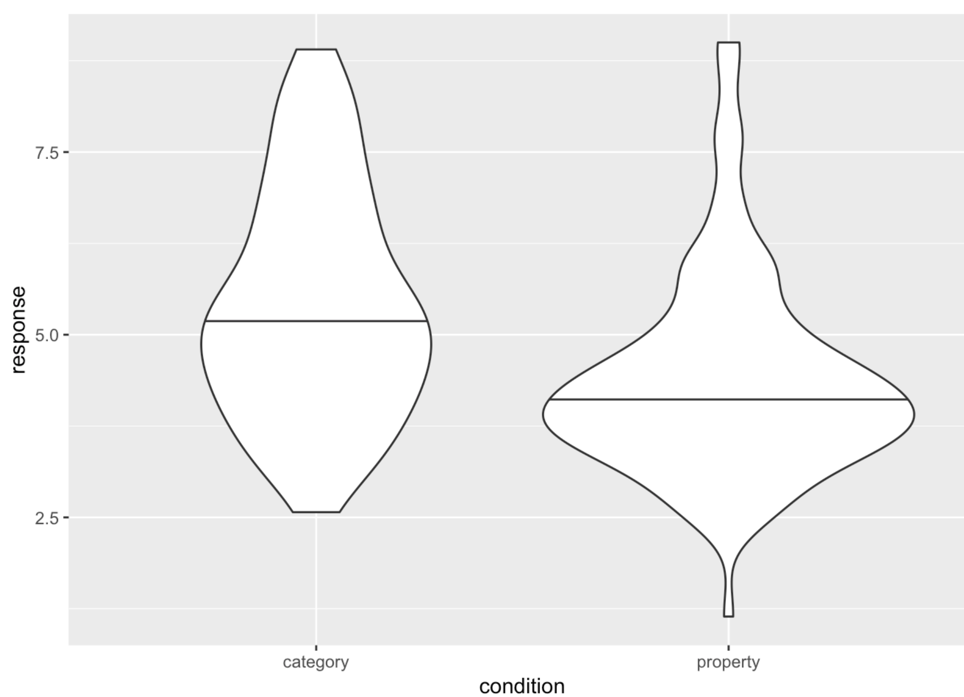
`tinyframes` data

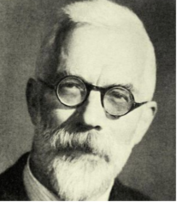
```
tinyframes <- frames %>%
  group_by(id, age, condition) %>%
  summarise(
    response = mean(response)
    ) %>%
  ungroup()
```

id	condition	response
1	category	5.333333
2	category	7.047619
3	property	4.857143
4	property	3.857143
5	property	9.000000
6	category	7.904762

`tinyframes` data

```
tinyframes <- frames %>%
  group_by(id, age, condition) %>%
  summarise(
    response = mean(response)
    ) %>%
  ungroup()
```

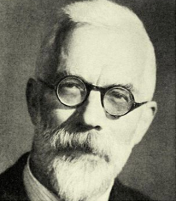




t-test

```
t.test(
  formula = response ~ condition,
  data = tinyframes,
  var.equal = TRUE
)
```


```
##  Two Sample t-test
##
## data:  response by condition
## t = 5.1625, df = 223, p-value = 5.388e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6259535 1.3988834
## sample estimates:
## mean in group category mean in group property
##               5.397661               4.385242
```

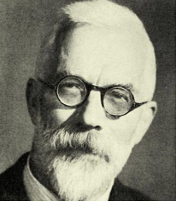


From a t-test to linear models

`mod1`: $\text{response}_i = \beta_0 + \epsilon_i$

`mod2`: $\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$

```
mod1 <- lm(formula = response ~ 1, data = tinyframes)
mod2 <- lm(formula = response ~ condition, data = tinyframes)
```



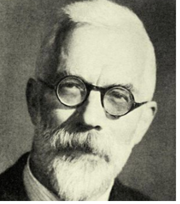
From a t-test to linear models

`mod1`: $\text{response}_i = \beta_0 + \epsilon_i$

`mod2`: $\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$

```
mod1 <- lm(formula = response ~ 1, data = tinyframes)
mod2 <- lm(formula = response ~ condition, data = tinyframes)
```

```
## 
## Call:
## lm(formula = response ~ condition, data = tinyframes)
## 
## Coefficients:
##       (Intercept)  conditionproperty  
##             5.398             -1.012
```



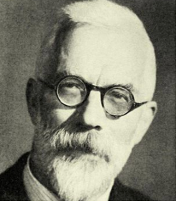
ANOVA for model comparison

`mod1`: $\text{response}_i = \beta_0 + \epsilon_i$

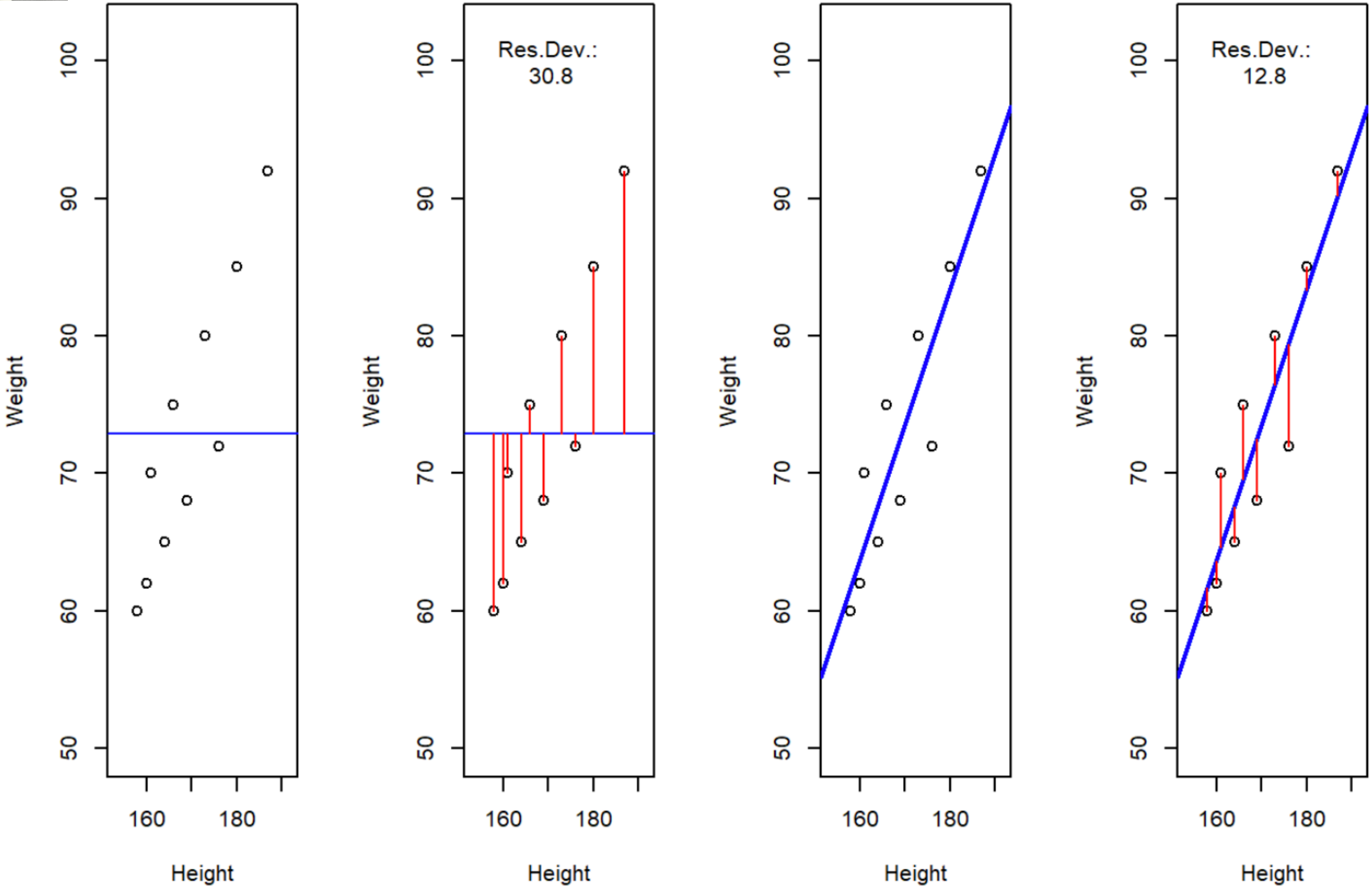
`mod2`: $\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$

```
anova(mod1, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: response ~ 1
## Model 2: response ~ condition
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    224 539.98
## 2    223 482.33  1    57.645 26.652 5.388e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Least squares regression



`y~1`

`y~1+x`

Coughing patient

- *d*: Jen is coughing

- h_1: Jen has a cold
  h_2: Jen has emphysema
  h_3: Jen has a stomach upset

Posterior probability

Evidence (Likelihood)

Prior knowledge

$$P(h|d) = \frac{\mathbf{P(d|h)}\ \mathbf{P(h)}}{P(d)}$$

Coughing patient

- *d*: Jen is coughing

- h_1: Jen has a cold
  h_2: Jen has emphysema
  h_3: Jen has a stomach upset

Posterior probability

Evidence (Likelihood)

Prior knowledge

$$P(h|d) \propto \mathbf{P(d|h)\ P(h)}$$

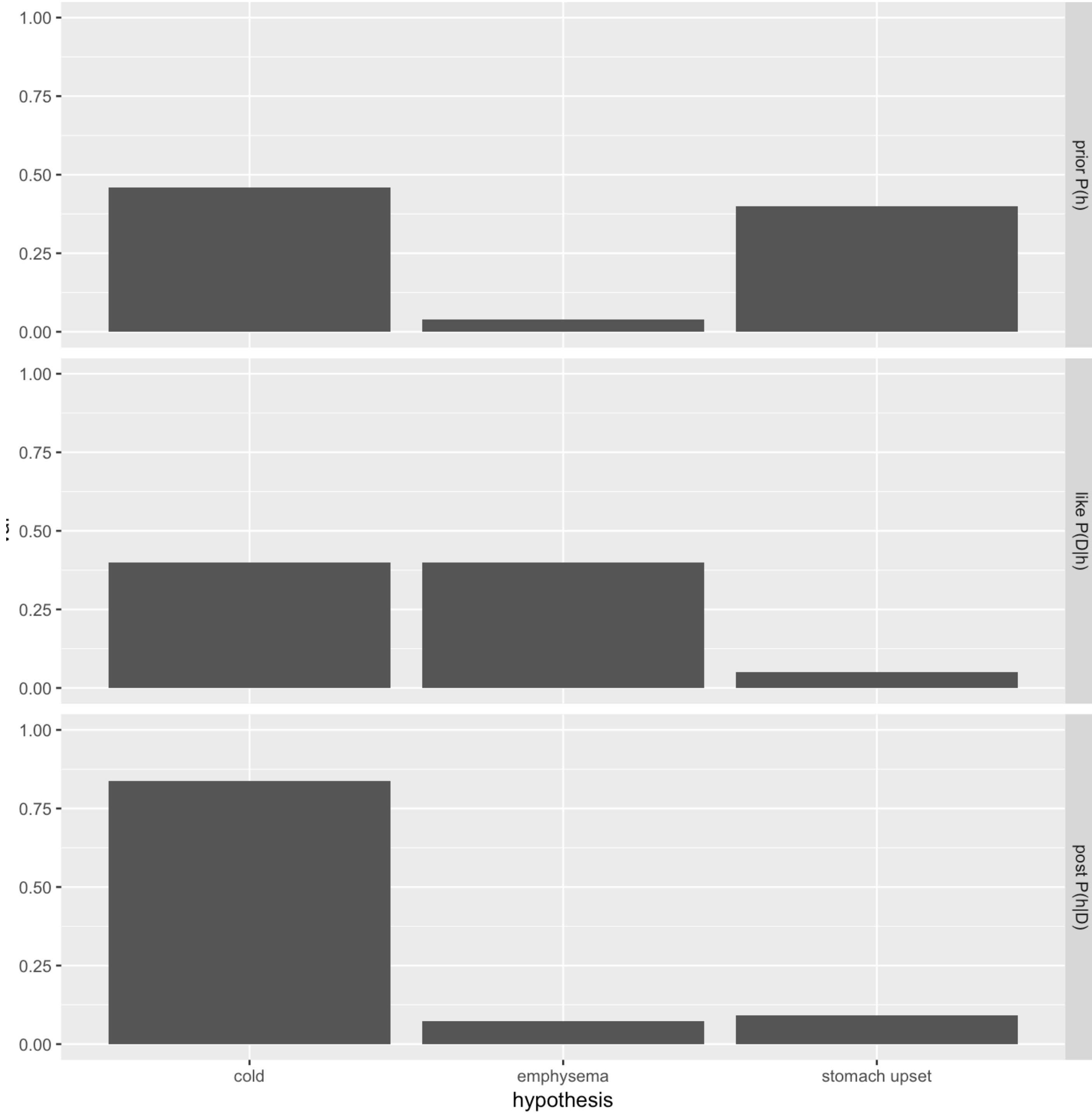
Specifying prior and likelihood

```r
h  <- c('cold', 'emphysema', 'stomach upset')
p_h <- c(0.46, 0.04, 0.4)
p_d_given_h <- c(0.4, 0.4, 0.05)
```

prior

likelihood

posterior



Exercise: Coughing patient

```r
h  <- c('cold', 'emphysema', 'stomach upset')
p_h <- c(0.46, 0.04, 0.4)
p_d_given_h <- c(0.4, 0.4, 0.05)
```

Bayesian inference

Two distinct applications:

1. Bayesian Data analysis

2. Bayesian cognitive models

Bayesian regression

$\mathtt{M}_1$: $$\text{response}_i = \beta_0 + \epsilon_i$$

$\mathtt{M}_2$: $$\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$$

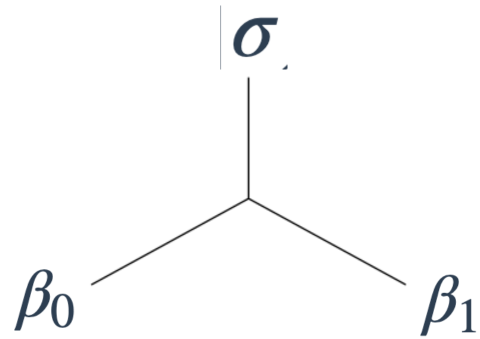
Both models assume $\epsilon_i \sim N(0, \sigma^2)$

Fitting $\mathtt{M}_2$: compute $P(\beta_0, \beta_1, \sigma | D)$

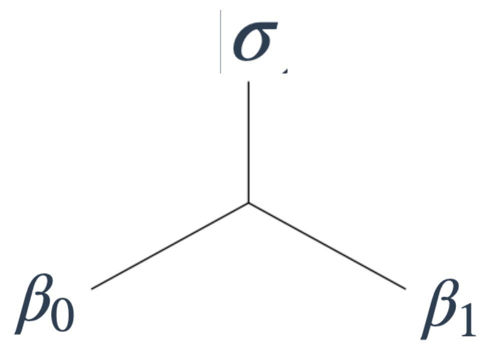
where *D* is the observed data

Bayesian regression

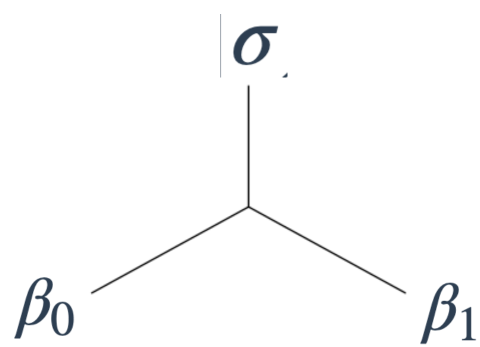
prior



likelihood



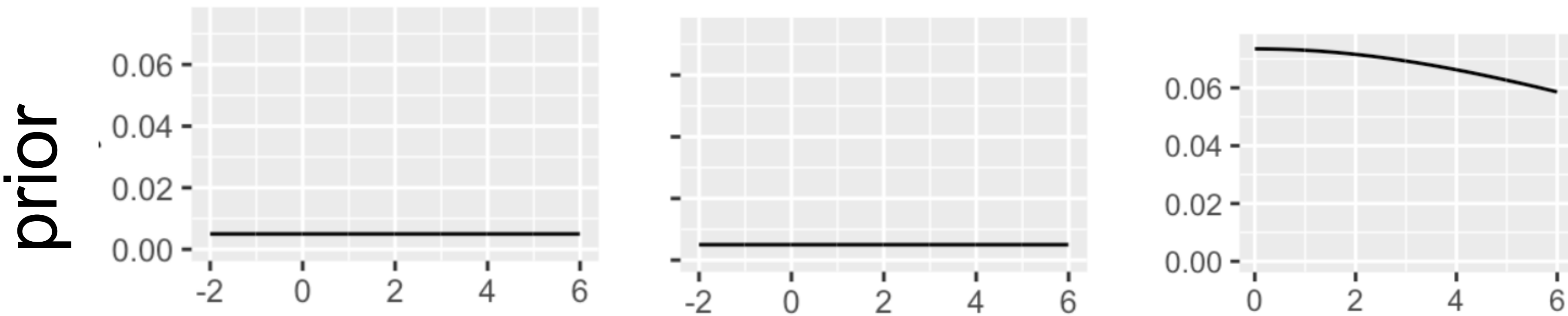
posterior



Bayesian inference

$$P(\beta_0, \beta_1, \sigma | D) \propto P(D | \beta_0, \beta_1, \sigma) P(\beta_0, \beta_1, \sigma)$$

β_0 β_1 σ



Assume $P(\beta_0, \beta_1, \sigma) = P(\beta_0) P(\beta_1) P(\sigma)$

Bayesian inference

$$P(\beta_0, \beta_1, \sigma | D) \propto \boxed{P(D | \beta_0, \beta_1, \sigma)} P(\beta_0, \beta_1, \sigma)$$

likelihood

id	condition	response
1	category	5.333333
2	category	7.047619
3	property	4.857143
4	property	3.857143
5	property	9.000000
6	category	7.904762

Bayesian inference

$$P(\beta_0, \beta_1, \sigma | D) \propto P(D | \beta_0, \beta_1, \sigma) P(\beta_0, \beta_1, \sigma)$$

β_0 β_1

likelihood

id	condition	response	modelfit
1	category	5.333333	5.397661
2	category	7.047619	5.397661
3	property	4.857143	4.385242
4	property	3.857143	4.385242
5	property	9.000000	4.385242
6	category	7.904762	5.397661

Bayesian inference

$$P(\beta_0, \beta_1, \sigma | D) \propto \boxed{P(D|\beta_0, \beta_1, \sigma)} P(\beta_0, \beta_1, \sigma)$$

$\beta_0 \ \beta_1$ $\quad\quad$ σ

likelihood

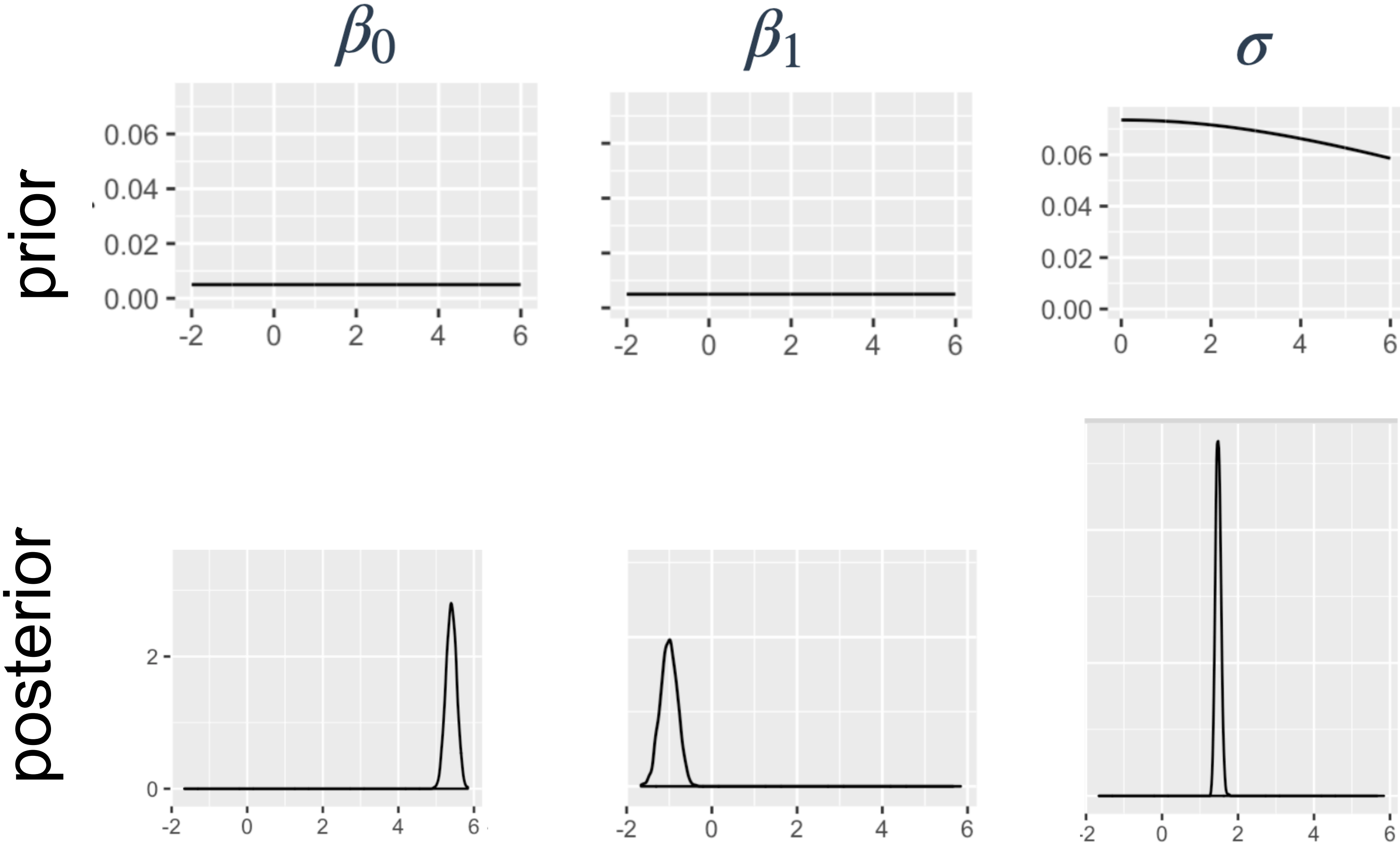
id	condition	response	modelfit	ϵ_i
1	category	5.333333	5.397661	-0.064327485
2	category	7.047619	5.397661	1.649958229
3	property	4.857143	4.385242	0.471900472
4	property	3.857143	4.385242	-0.528099528
5	property	9.000000	4.385242	4.614757615
6	category	7.904762	5.397661	2.507101086

$P(\epsilon_i | \sigma)$

$\vdots$

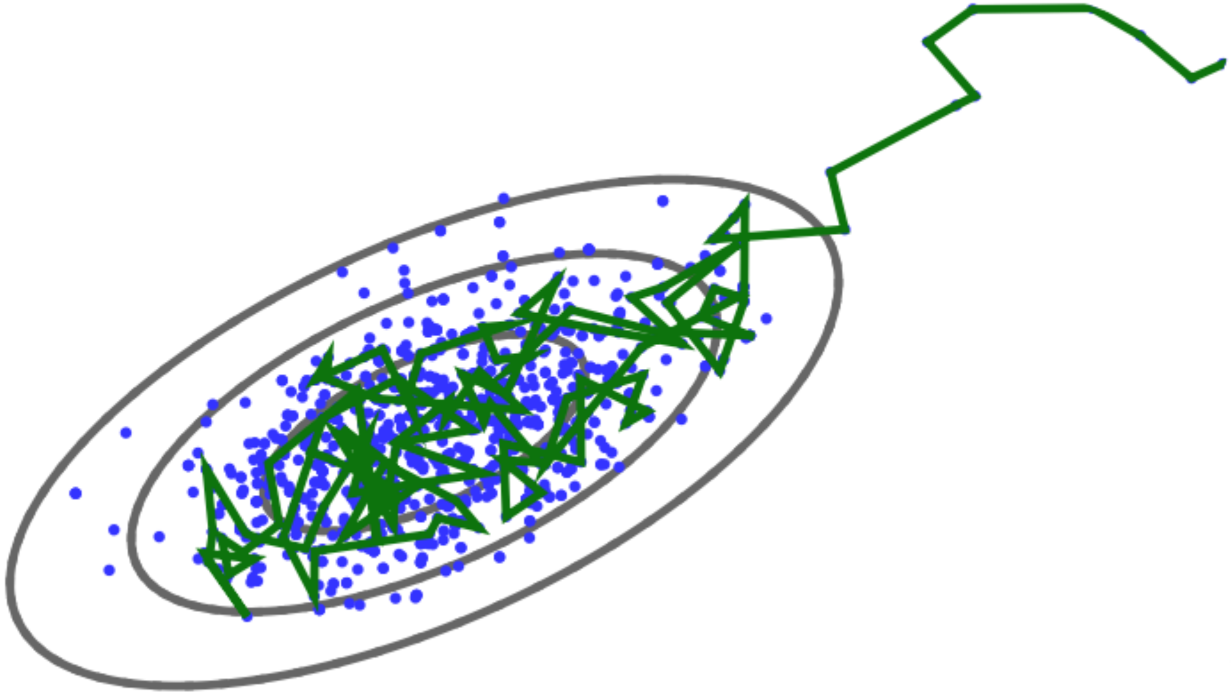
Bayesian inference

$$P(\beta_0, \beta_1, \sigma | D) \propto P(D|\beta_0, \beta_1, \sigma) P(\beta_0, \beta_1, \sigma)$$



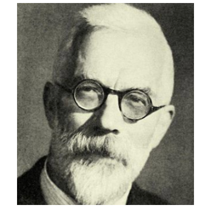
Markov-Chain Monte Carlo (MCMC) methods

$$P(\beta_0, \beta_1, \sigma | D) \propto P(D | \beta_0, \beta_1, \sigma) P(\beta_0, \beta_1, \sigma)$$



Regression

- Least-squares:



```
mod2 <- lm(
  formula = response ~ condition,
  data = tinyframes)
)
```

- Bayesian:

```
mod2_bayes <- brm(
  formula = response ~ condition,
  data = tinyframes,
)
```

Bayes factors for model comparison

M$_1$: $$\text{response}_i = \beta_0 + \epsilon_i$$

M$_2$: $$\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$$

$$BF_{21} = \frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)P(M_2)}{P(D|M_1)P(M_1)}$$

Bayes factors for model comparison

M$_1$: $\text{response}_i = \beta_0 + \epsilon_i$

M$_2$: $\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$

$$BF_{21} = \frac{P(M_2|D)}{P(M_1|D)} = \frac{P(D|M_2)}{P(D|M_1)}$$

$$P(D|M_2) = \int P(D|\beta_0, \beta_1, \sigma, M_2) P(\beta_0, \beta_1, \sigma|M_2) d\beta_0 d\beta_1 d\sigma$$

Bayes factors for model comparison

Value	Interpretation
$BF_{21} < 1$	Negative: supports M_1 rather than M_2
$1 < BF_{21} < 3$	Barely worth mentioning
$3 < BF_{21} < 10$	Substantial
$10 < BF_{21} < 100$	Strong
$100 < BF_{21}$	Decisive

Bayes factors for model comparison

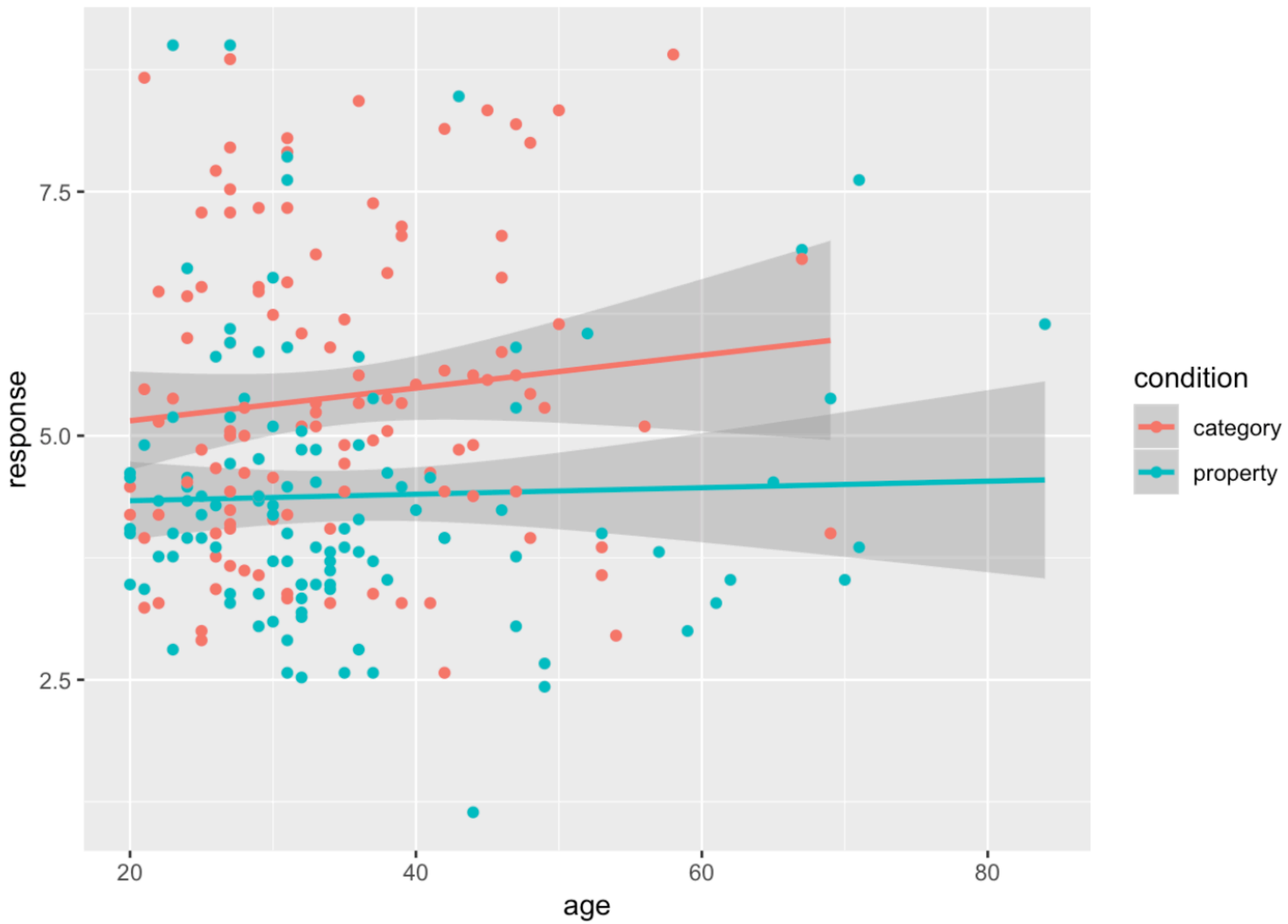
`mod1`: $\text{response}_i = \beta_0 + \epsilon_i$

`mod2`: $\text{response}_i = \beta_0 + \beta_1 * \text{condition}_i + \epsilon_i$

```
mod2_bayes <- brm(
  formula = response ~ condition,
  data = tinyframes,
  save_all_pars = TRUE
)
```

```
BF <- bayes_factor(mod2_bayes, mod1_bayes)
```

Multiple predictors



Multiple predictors

`mod3`:

```
## Call:
## lm(formula = response ~ condition + age, data = tinyframes)
##
## Coefficients:
##       (Intercept)  conditionproperty                age
##          5.102572          -1.018548           0.008536
```

Model selection:

```
anova(mod1, mod2, mod3)
```

Model comparison with AIC and BIC

For model with parameters θ

Find $\hat{\theta}$ that maximizes $P(D|\theta)$

AIC: $$2k - 2\ln(P(D|\hat{\theta}))$$

BIC: $$\ln(n)k - 2\ln(P(D|\hat{\theta}))$$

where k is number of parameters,
n is number of data points

Model comparison with AIC and BIC

Find $\hat{\theta}$ that maximizes $P(D|\theta)$

AIC: $$2k - 2\ln(P(D|\hat{\theta}))$$

BIC: $$\ln(n)k - 2\ln(P(D|\hat{\theta}))$$

where k is number of parameters,
n is number of data points

Important points:

- lower is better
- both penalize model complexity (BIC has heavier penalty)

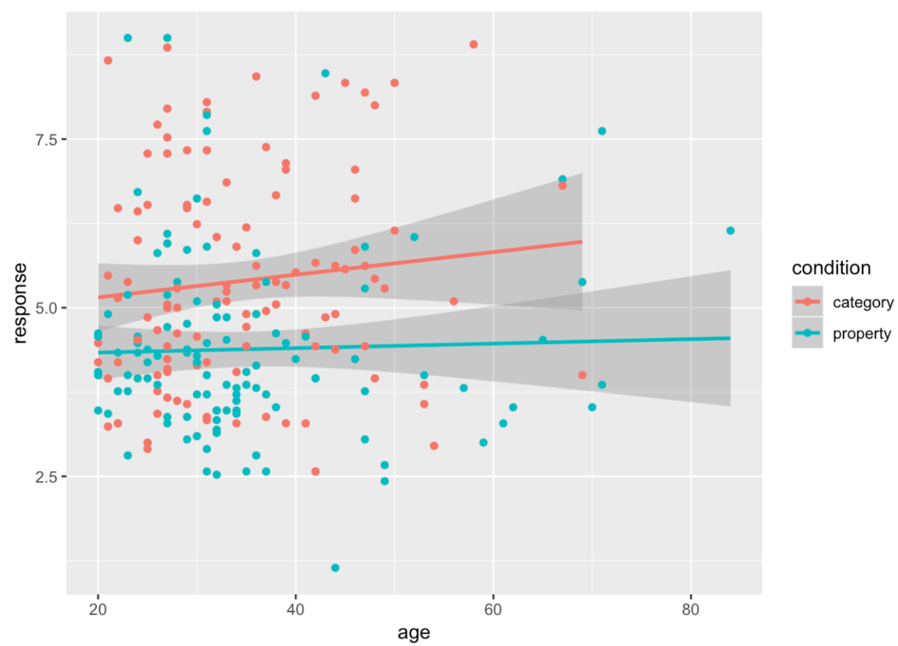
Model comparison with AIC and BIC

```
AIC(mod1, mod2, mod3)
```

```
##        df      AIC
## mod1    2 839.4940
## mod2    3 816.0928
## mod3    4 817.0575
```

```
BIC(mod1, mod2, mod3)
```

```
##        df      BIC
## mod1    2 846.3262
## mod2    3 826.3411
## mod3    4 830.7219
```

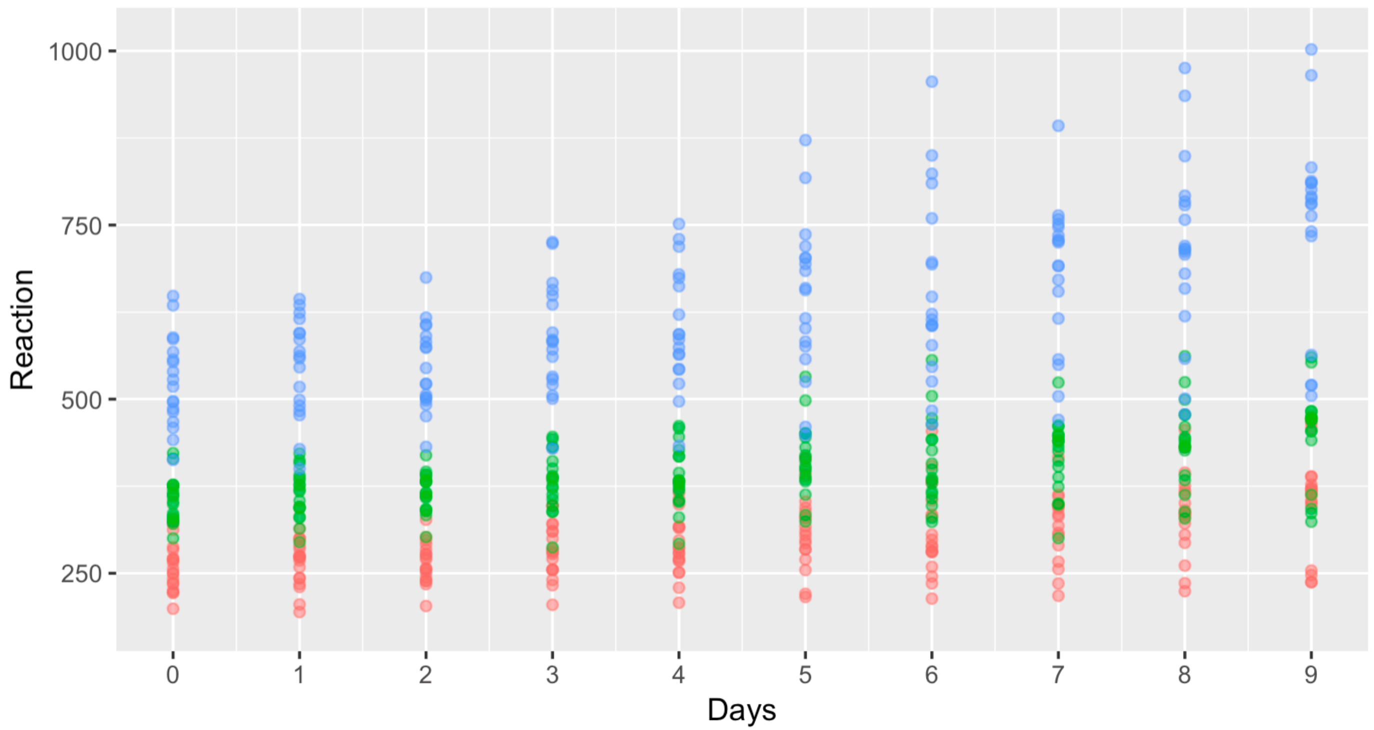


Mixed effects models

- ANOVA models used to be the go-to approach in psychology, but the field is shifting to mixed-effects models.

- Advantages of mixed-effects models:
  - extend naturally to complex situations (e.g. cases with nested structure, factors that overlap in complex ways)
  - deal well with missing data

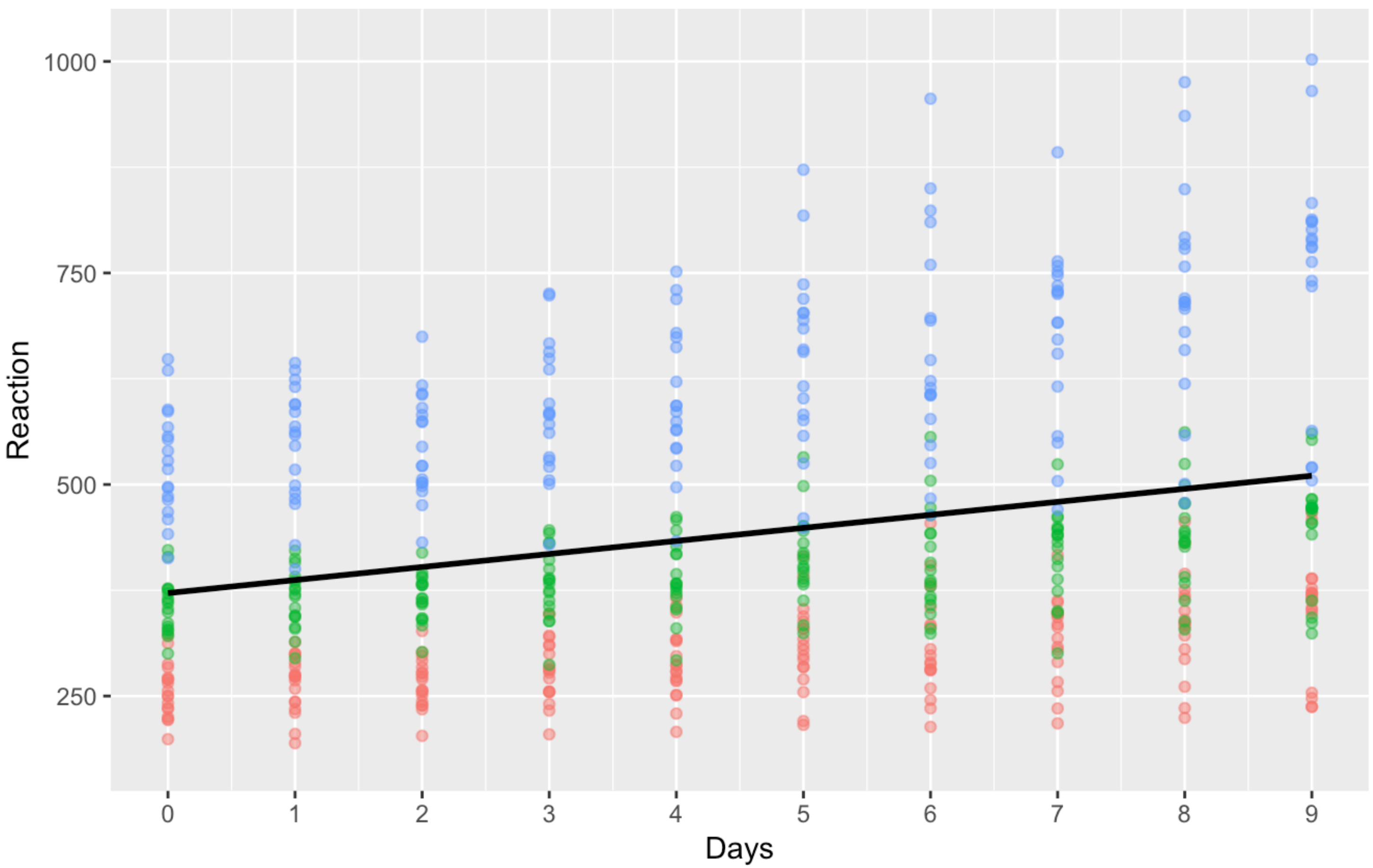
Sleep study example



Reaction	Days	Subject	Group
249.5600	0	1	group_1
258.7047	1	1	group_1
250.8006	2	1	group_1
321.4398	3	1	group_1

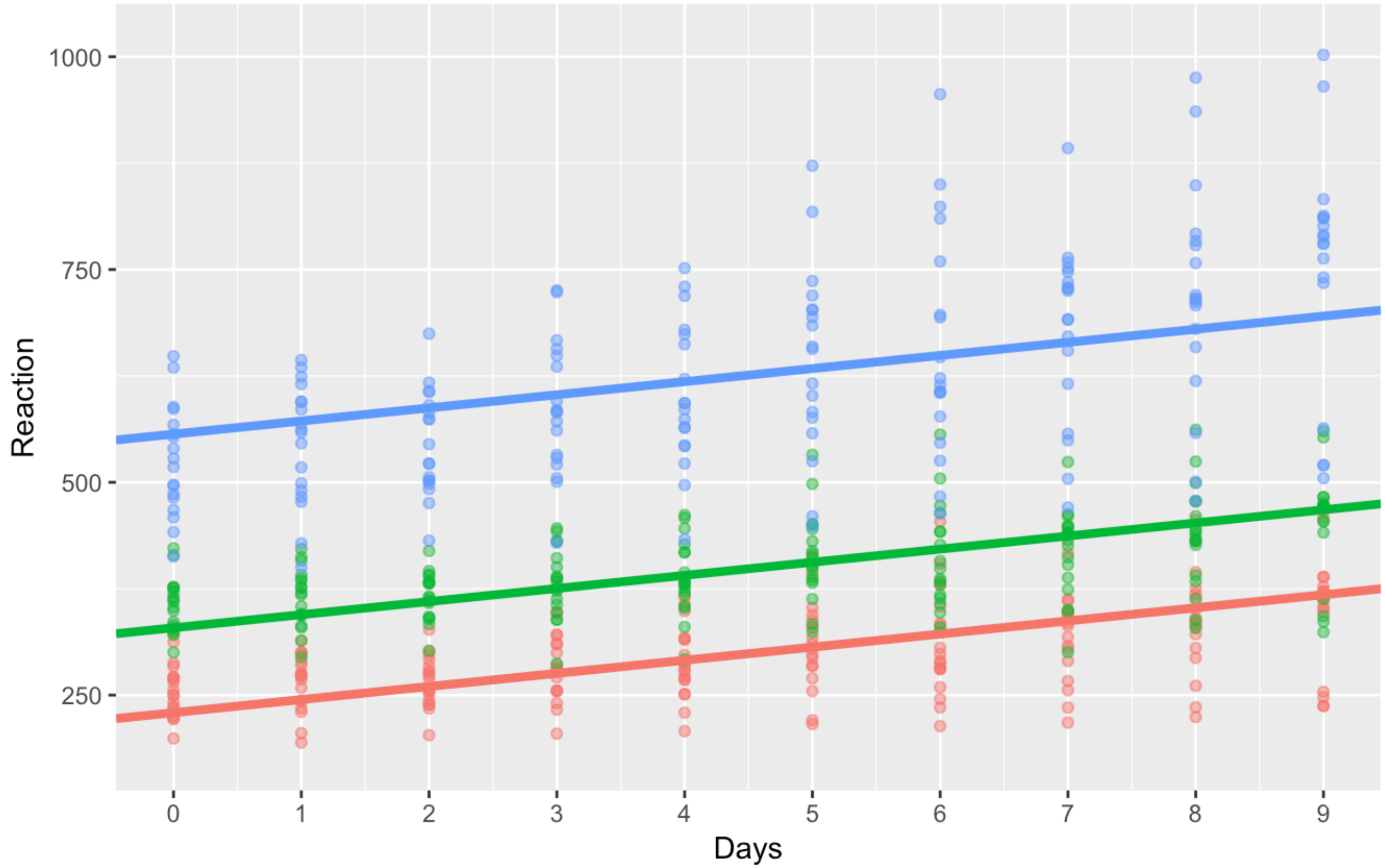
Fixed intercept, slope

```
lm(Reaction ~ Days, data = sleep_groups)
```



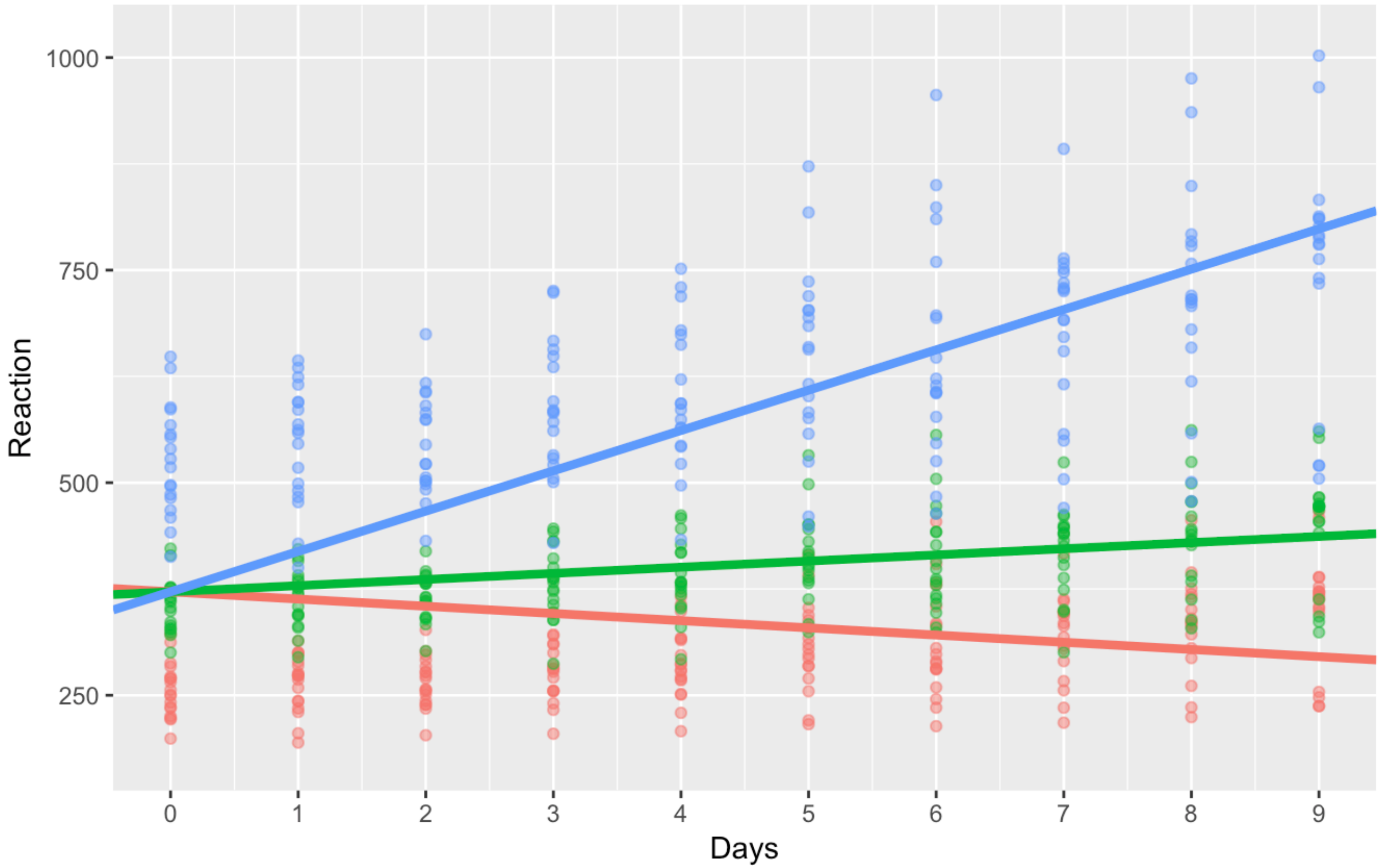
Random intercept per group

```
lmer(Reaction ~ Days + (1 | Group), data = sleep_groups)
```



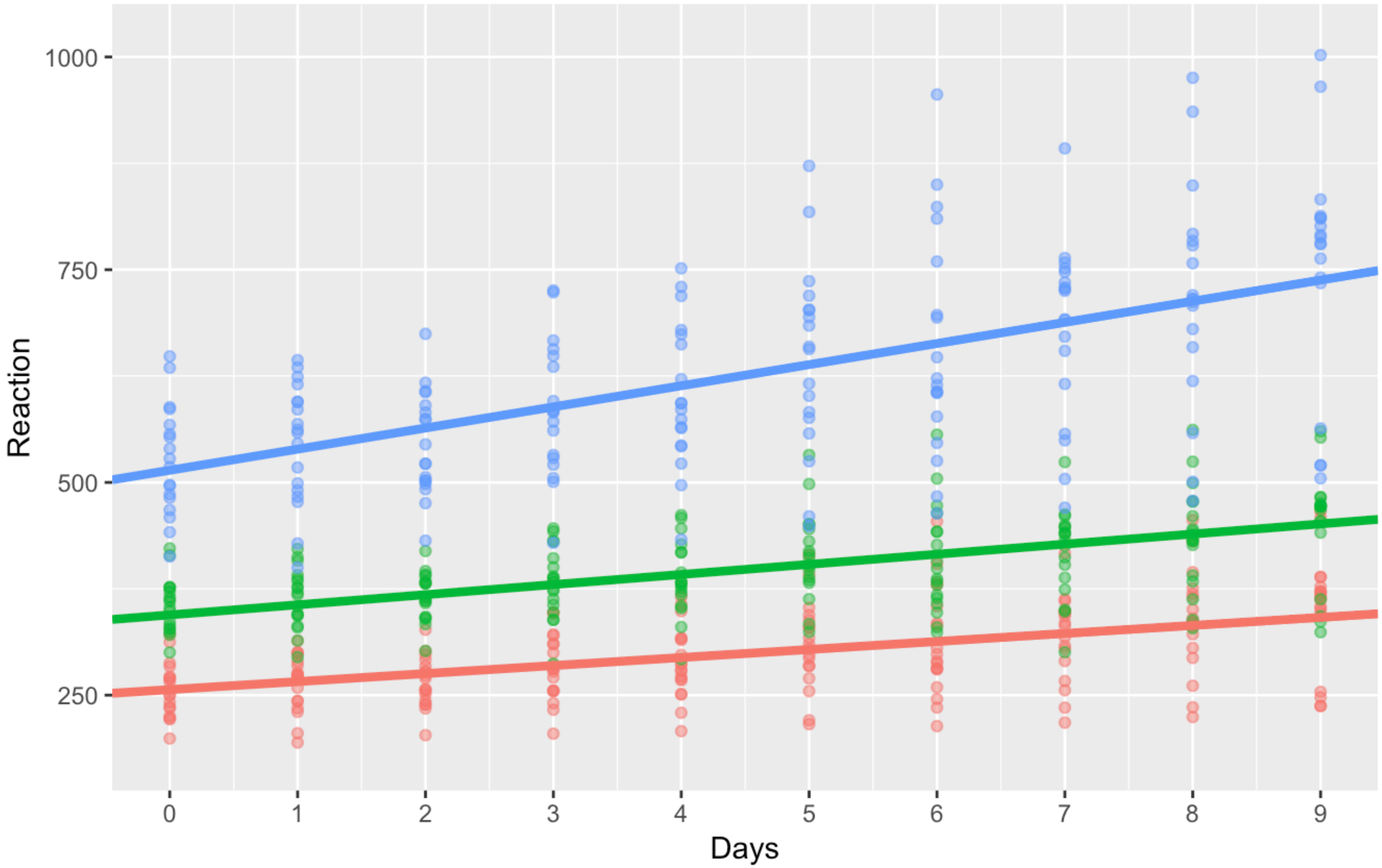
Random slope per group

```
(Reaction ~ Days + (0 + Days | Group)
```

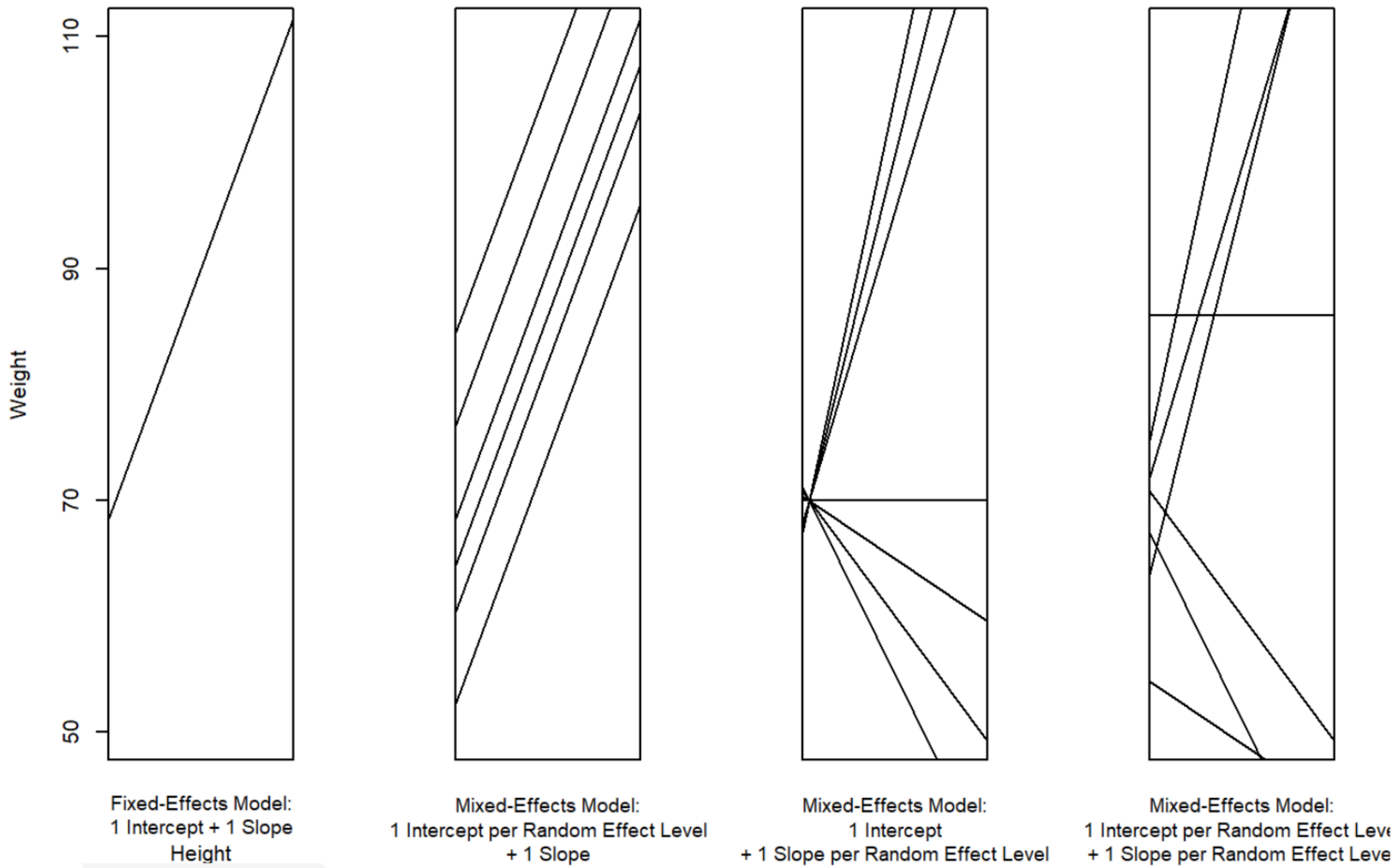


Random slope + intercept per group

```
(Reaction ~ Days + (1 + Days | Group)
```



Mixed effects models



`y~1+x`

`y~1+x+(1|g)`

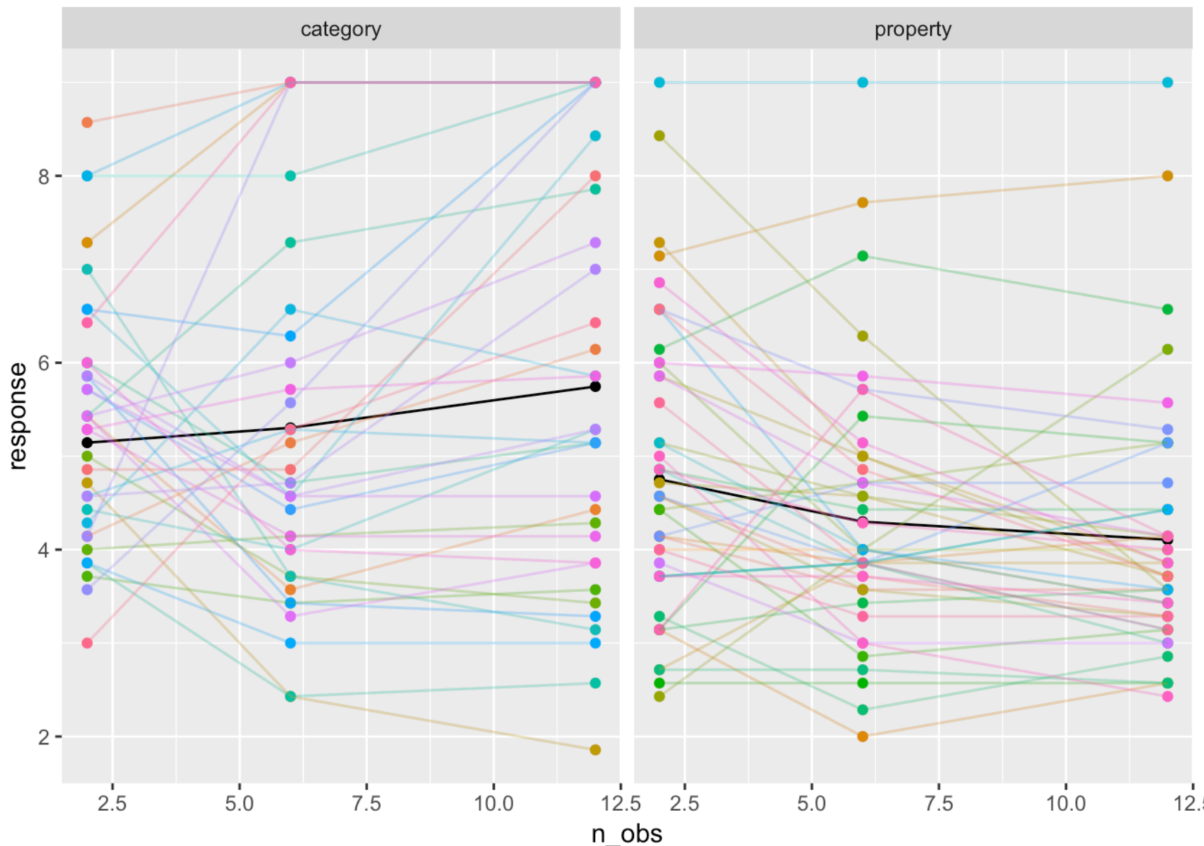
`y~1+x+(0+x|g)`

`y~1+x+(1+x|g)`

Exercise

`modestframes` data

id	age	condition	n_obs	response
1	36	category	2	5.857143
1	36	category	6	5.285714
1	36	category	12	4.857143
2	46	category	2	5.285714



Model comparison

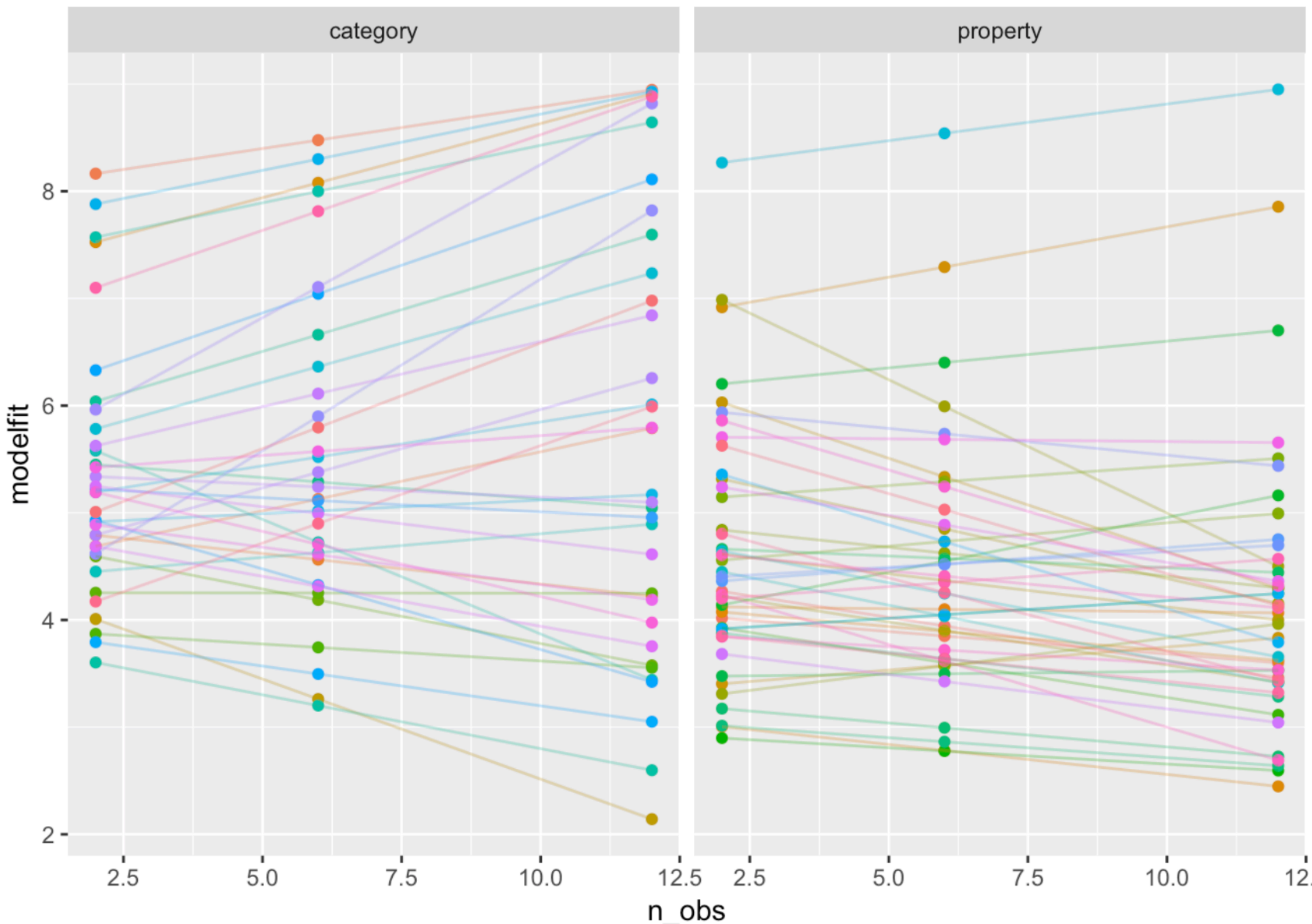
```
modest1: response ~ 1 + (1 | id)
modest2: response ~ condition + n_obs + (1 | id)
modest3: response ~ condition + n_obs + (1 + n_obs | id)
```

```
anova(modest1, modest2, modest3)
```

	Df	AIC	BIC	logLik	deviance	Chisq	Chi Df	Pr(>Chisq)	
modest1	3	2354.9	2368.5	-1174.5	2348.9				
modest2	5	2333.5	2356.1	-1161.8	2323.5	25.403	2	3.046e-06	***
modest3	7	2270.4	2302.0	-1128.2	2256.4	67.164	2	2.603e-15	***

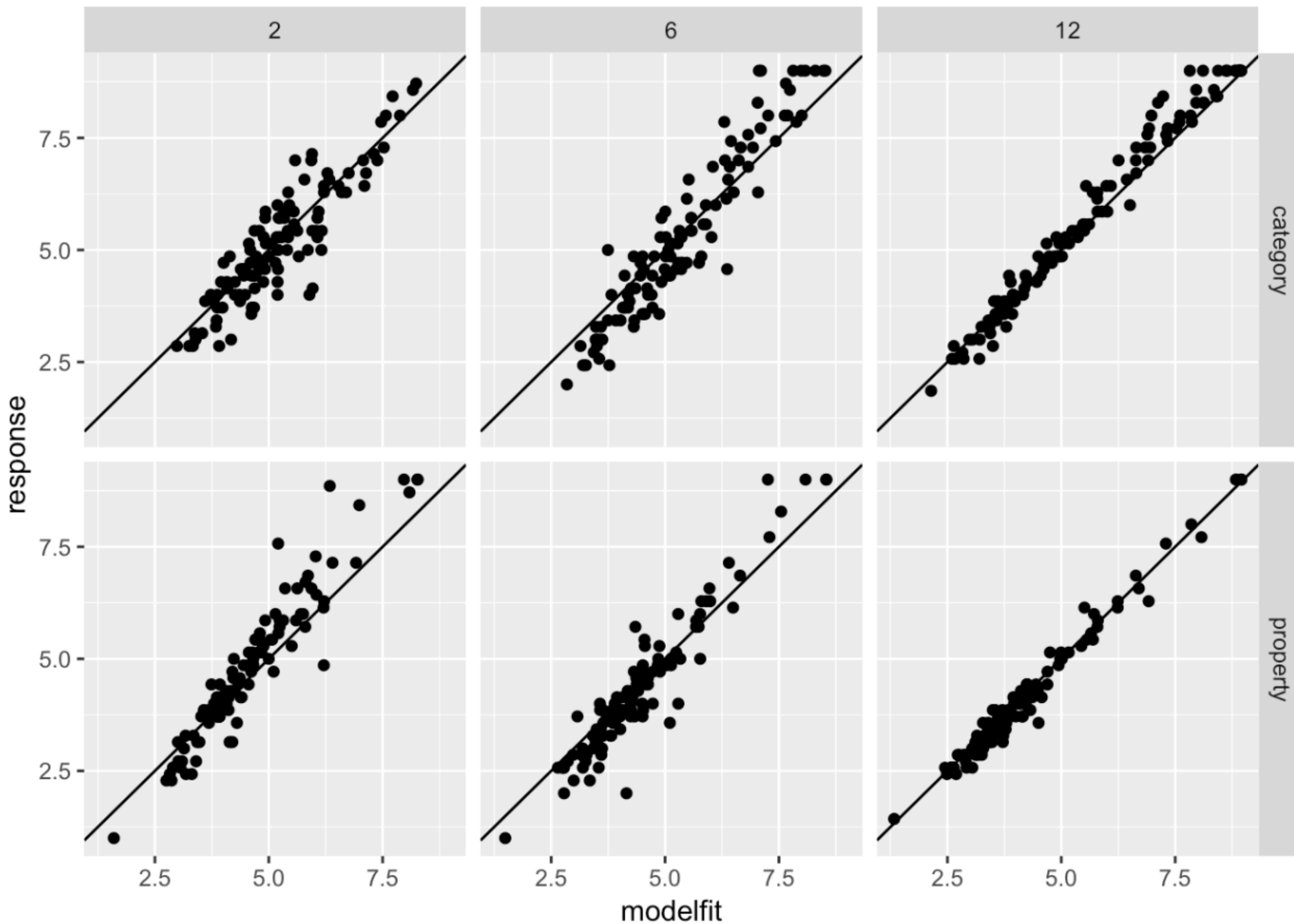
Model checking: individuals

```
modestframes$modelfit <- predict(modest3)
modestframes$residuals <- residuals(modest3)
```



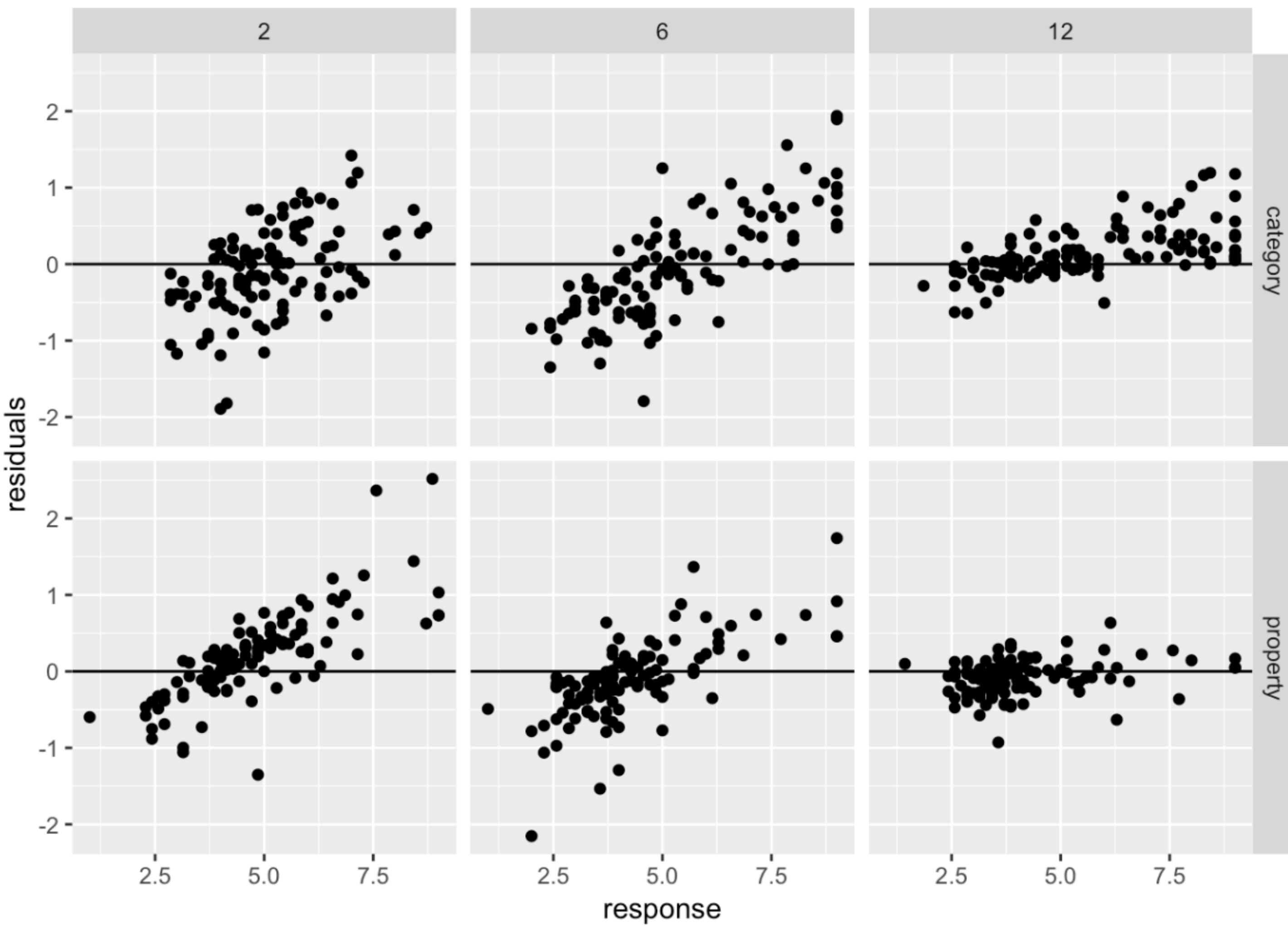
Model checking: predictions

```
modestframes$modelfit <- predict(modest3)
modestframes$residuals <- residuals(modest3)
```



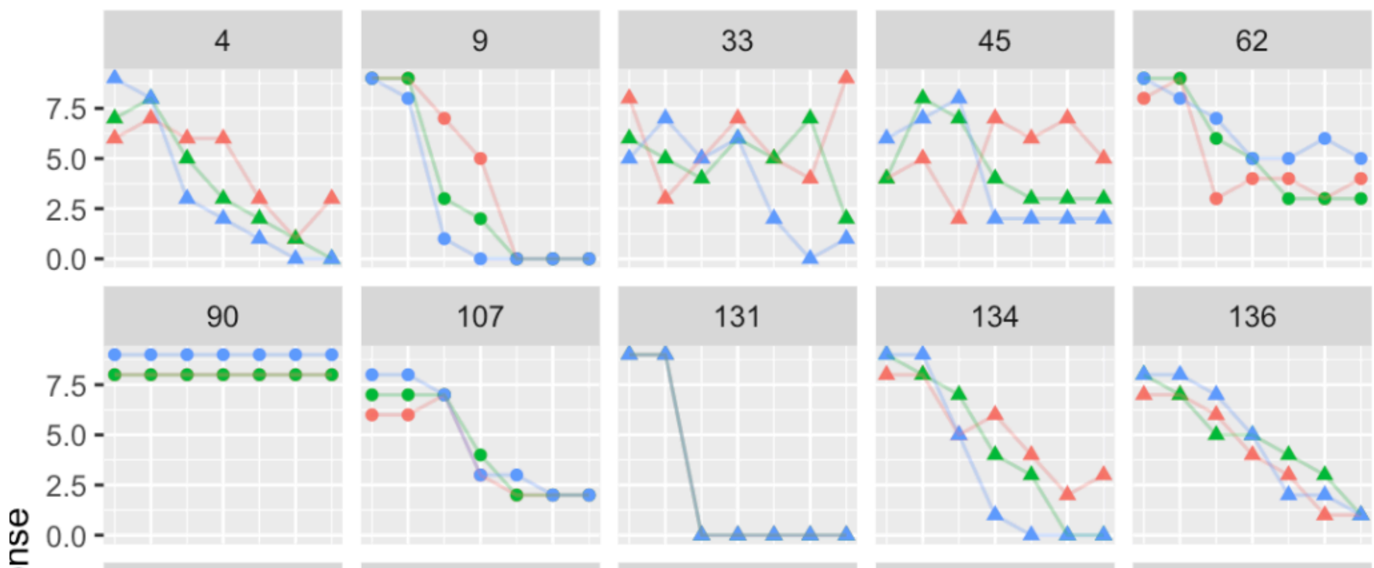
Model checking: residuals

```
modestframes$modelfit <- predict(modest3)
modestframes$residuals <- residuals(modest3)
```



`frames` data

id	condition	sample_size	n_obs	test_item	response
1	category	small	2	1	8
1	category	small	2	2	7
1	category	small	2	3	6
1	category	small	2	4	6
1	category	small	2	5	5



Model comparison

linframes1:

```
response ~ condition + n_obs + (1 + n_obs|id)
```

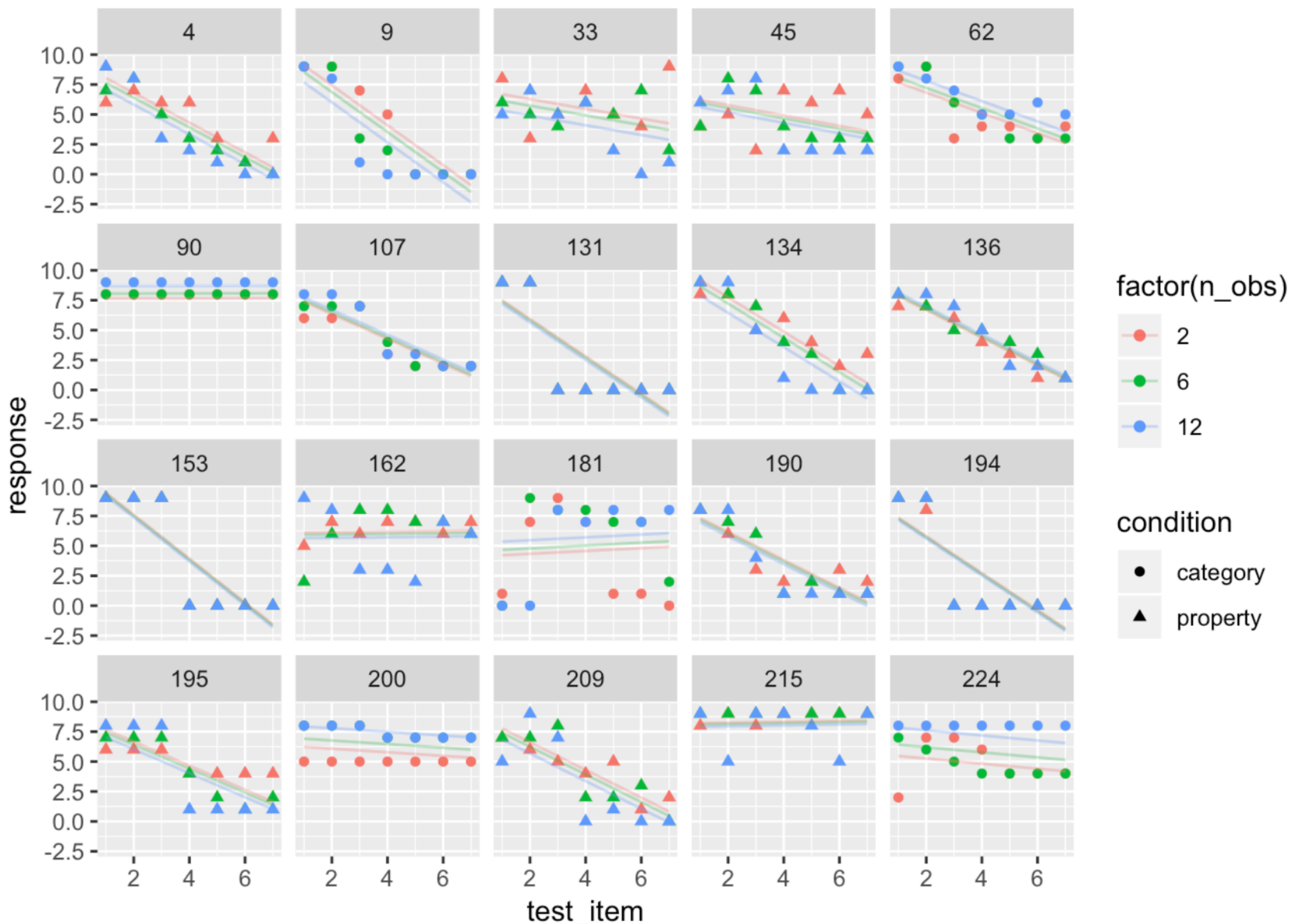
linframes2:

```
response ~ condition + n_obs + test_item + (1 + n_obs + test_item |id)
```

```
anova(linframes1, linframes2)
```

```
           Df   AIC   BIC   logLik deviance Chisq Chi Df Pr(>Chisq)
linframes1  7 23128 23173 -11556.8    23114
linframes2 11 19734 19805  -9855.8    19712  3402      4  < 2.2e-16 ***
```

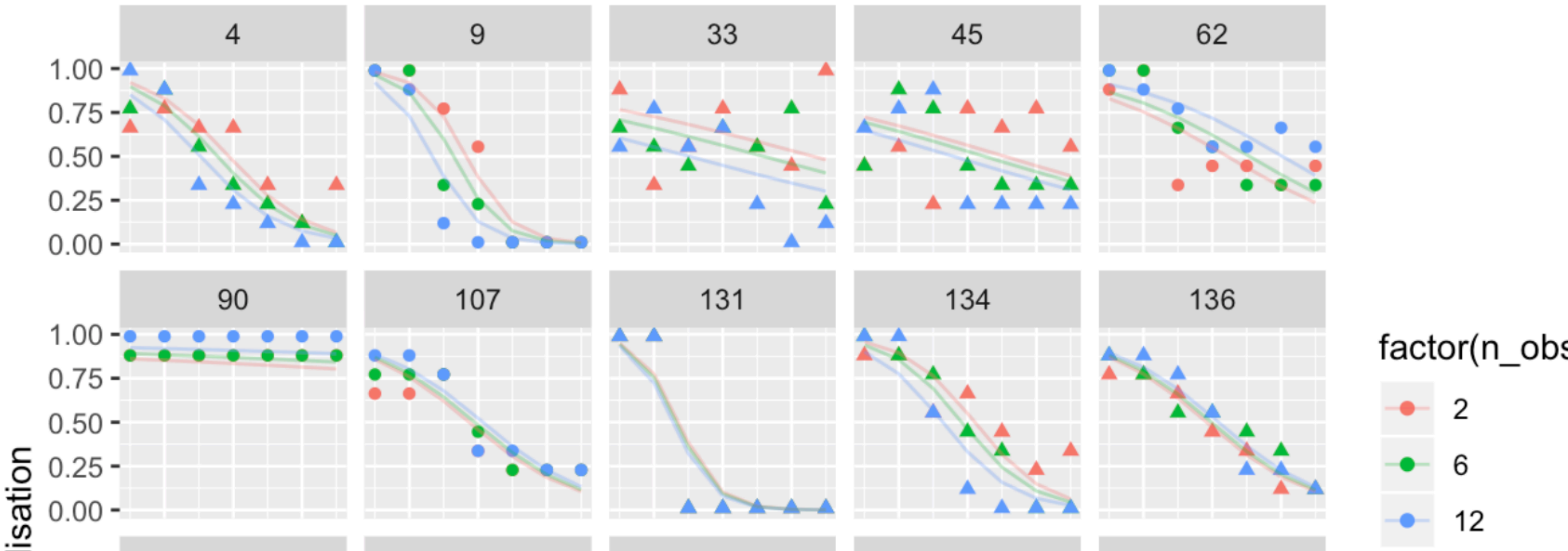
Model checking: individuals



Generalized linear mixed models

Map `response` to `generalization` (between 0 and 1)

```
logitmod <- glmer(
  formula = generalisation ~ condition + test_item + n_obs + (1 + test_item + n_obs|id),
  family = gaussian(link = "logit"),
  data = glmerframes)
```



What to write up?

- The actual paper reported Bayes factors computed using JASP

- See `analysis_samplesize.Rmd` (in `samplingframes/analysis`) for more