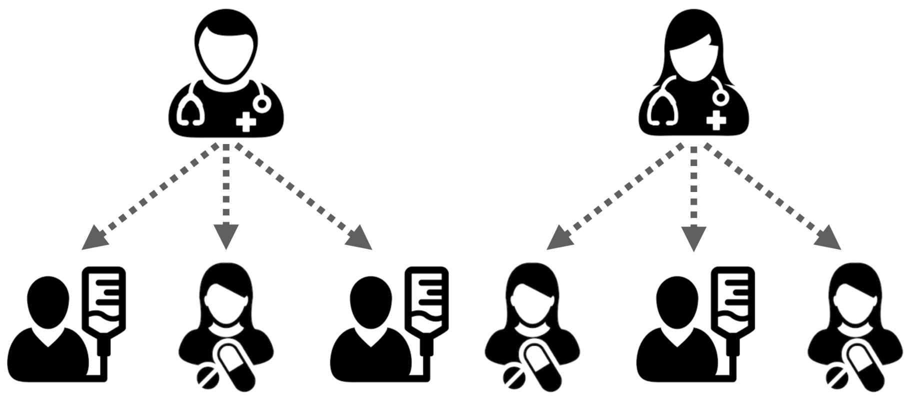
Mixed Effects Models

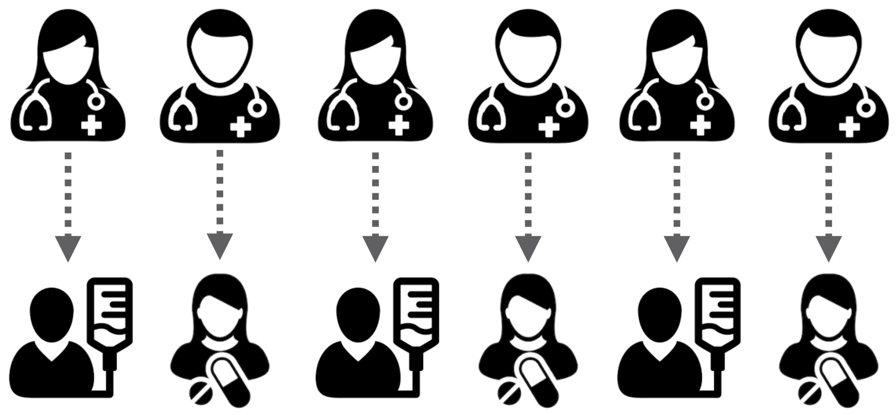
Statistics with R

Basel R Bootcamp

April 2019

Data Collection Example



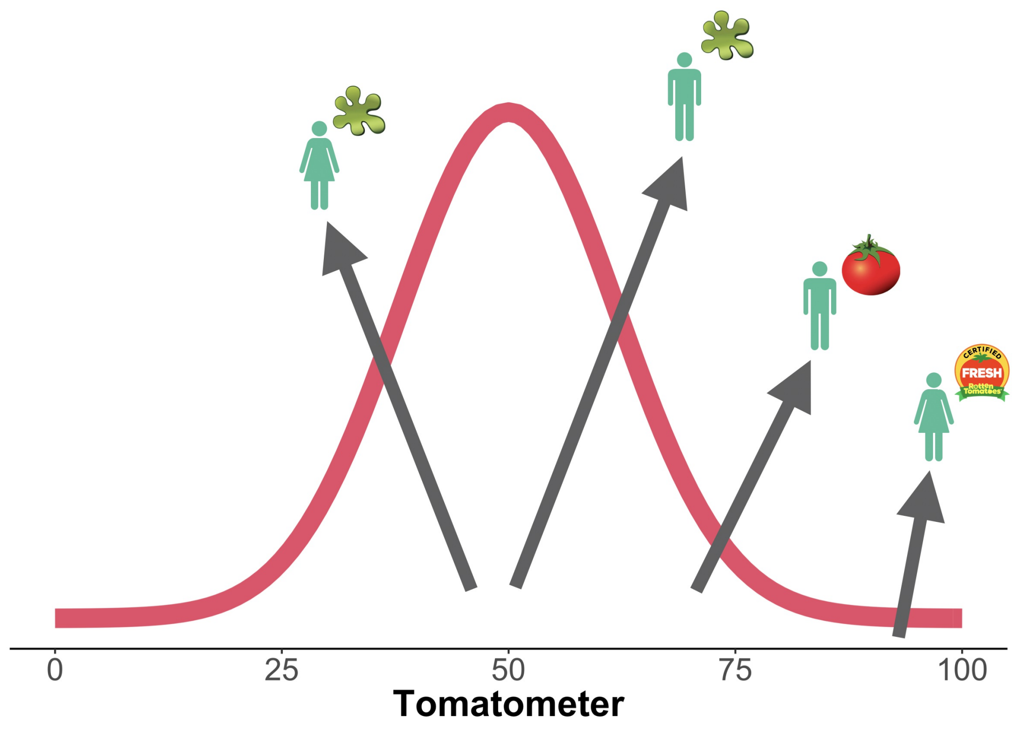


IID Assumption

Previous regression models assumed that the data points are independent and identically distributed (**iid**).

- **Independent**: The occurrence of one value does not affect the occurrence of another value.
- **Identically distributed**: All data points stem from the same probability distribution.

Previously considered models are **not robust** to violations of the independence assumption.



Examples of Data Where the IID Assumption Is Violated

- Test of student performance in the PISA study (**students within classes within schools within countries**).
- Test of a drug in a clinical trial over multiple sites (**patients within doctors within sites**).
- Test of the best design or taste of different versions of a product with **multiple ratings from the same persons**.

from wikipedia.org

from mssociety.org.uk

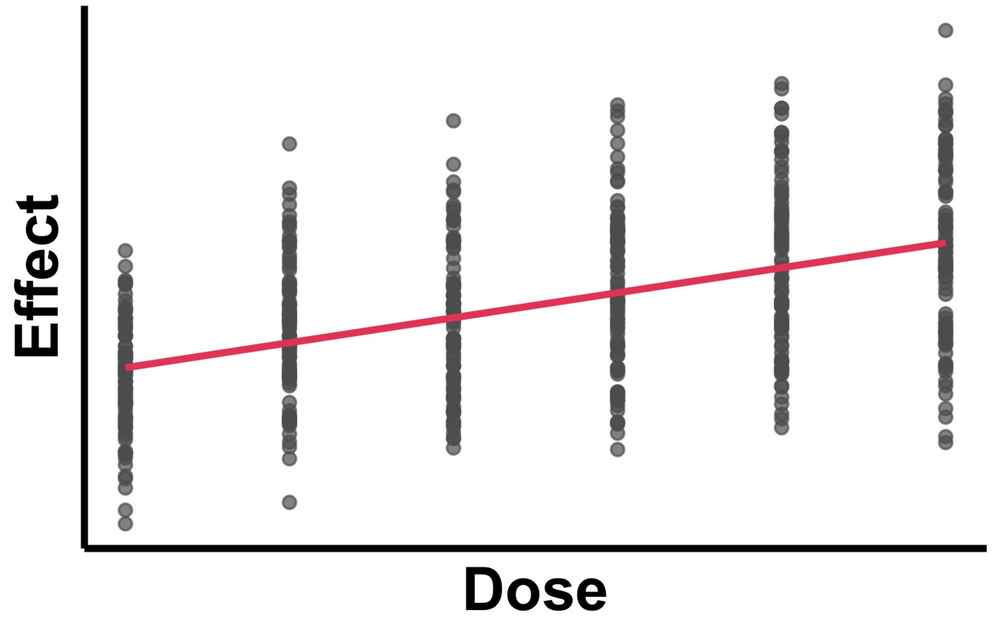
from londonist.com

Example - Clinical Trial Over Multiple Sites

Let's assume we want to evaluate the effect of a drug (in different doses). We have data of a clinical trial over three sites.

Here's how we would run a regression over all sites (but is it a good idea to do so?):

```
mod <- lm(formula = Effect ~ Dose,
          data = clinical_data)
```



Example - Clinical Trial Over Multiple Sites

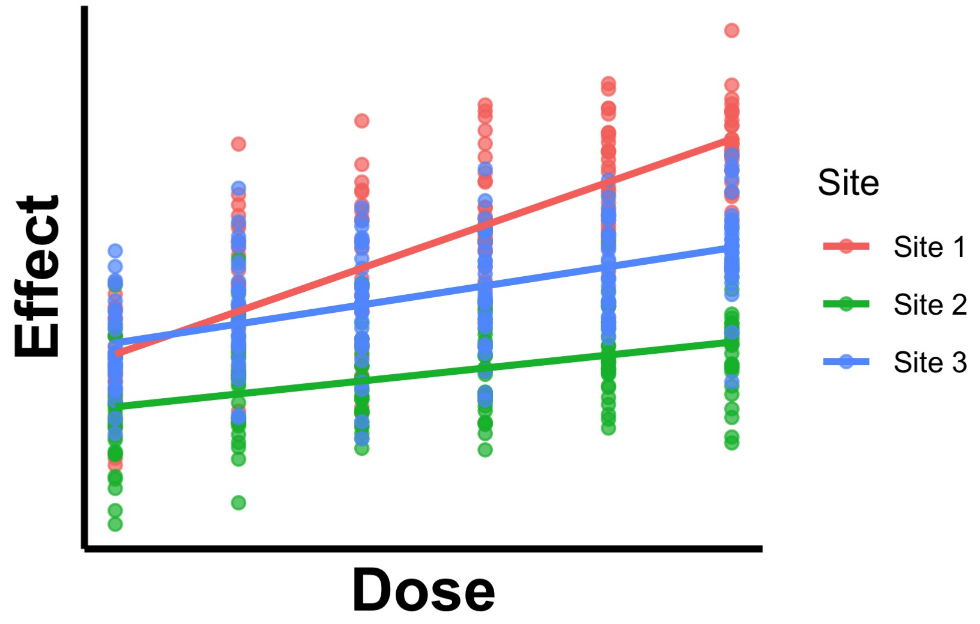
Let's assume we want to evaluate the effect of a drug (in different doses). We have data of a clinical trial over three sites.

Here's how we would run one regression per site (but is it a good idea to do so?):

```
mod_s1 <- lm(formula = Effect ~ Dose,
             data = site1_data)

mod_s2 <- lm(formula = Effect ~ Dose,
             data = site2_data)

mod_s3 <- lm(formula = Effect ~ Dose,
             data = site3_data)
```



Mixed Effects Model

A better way of analysing data where the IID assumption is violated than running separate regressions is to use mixed effects models.

Mixed effects models combine two kinds of effects that serve different purposes:

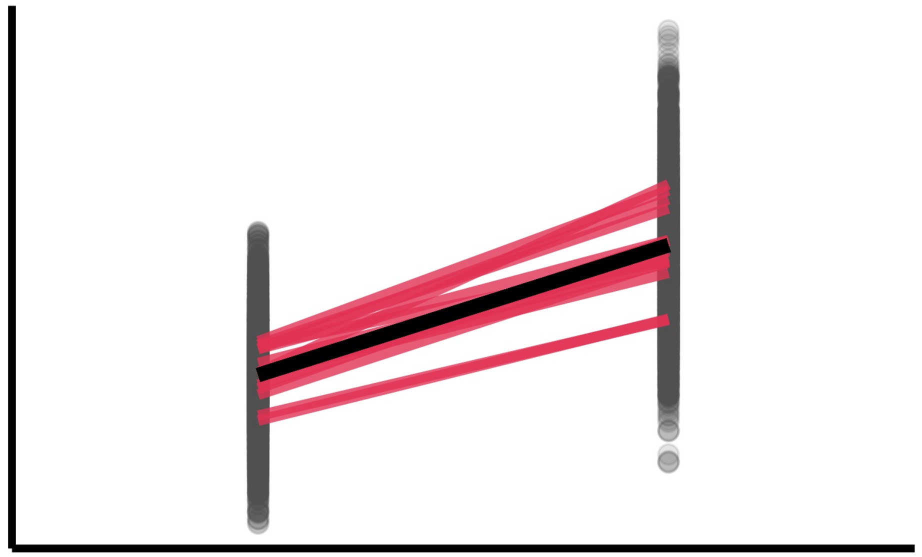
Fixed effects

- The main effects/ interactions we want to quantify.

Random effects

- Account for the dependencies in the data.
- The effects of the population we want to generalize over.

```
LMM <- lmer(formula = y ~ FE1 + FE2 + # Fixed
                (FE1|RE) + (FE2|RE), # Random
              data = df) # Data
```

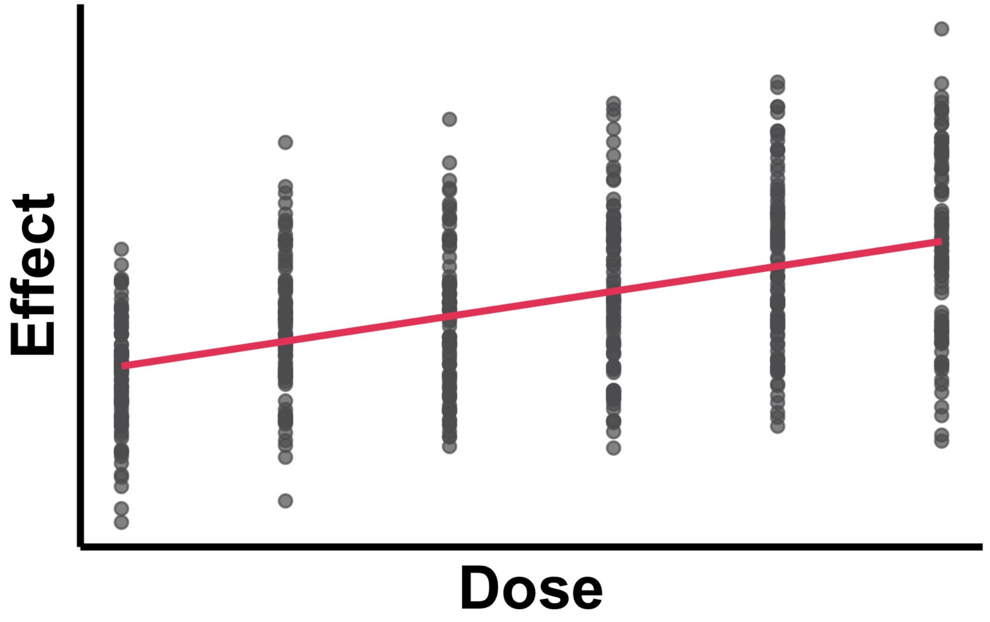


Fixed Effects

Fixed effects represent the overall effect (main effect or interaction) of a predictor. For example the effect of a treatment.

So far we only considered "fixed effects only" models (e.g., simple regression, t-tests, ANOVA).

```
# continuous predictor
mod <- lm(formula = Effect ~ Dose,
          data = clinical_data)
```



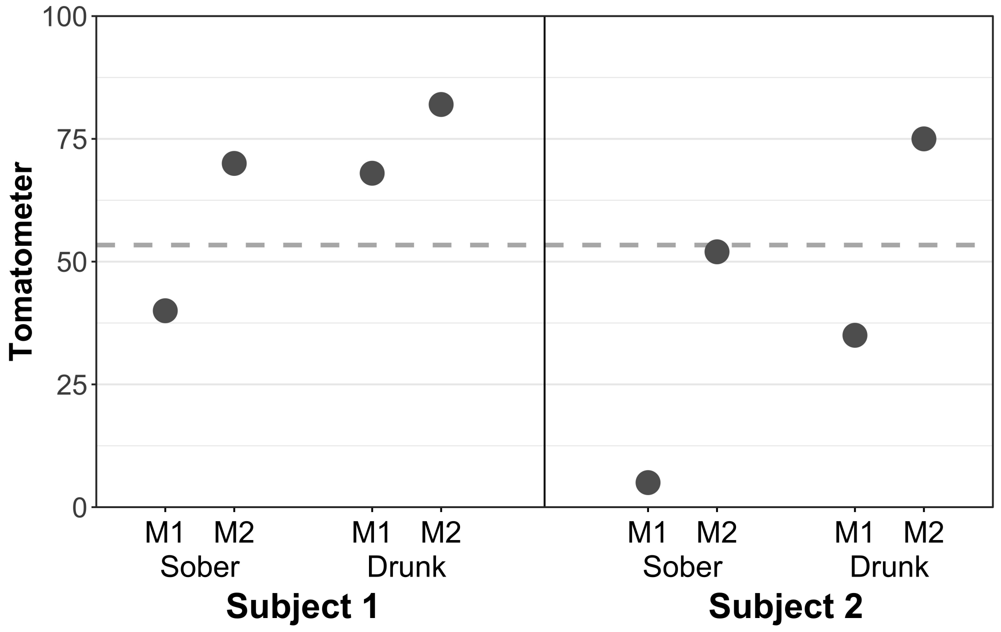
Example Data

In this introduction we will work with the following data:

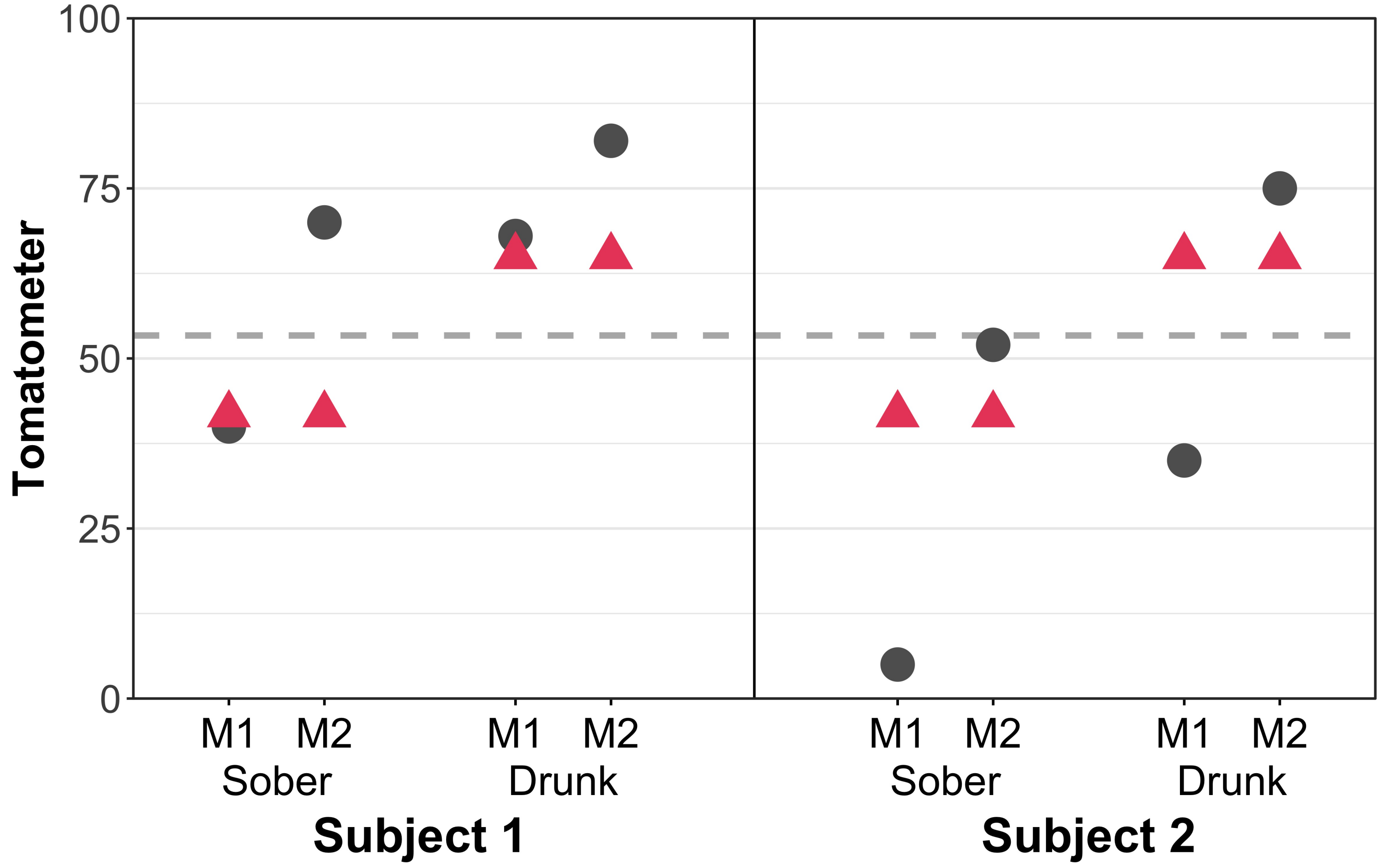
You are a researcher and want to gauge whether alcohol influences movie ratings on **Rotten Tomatoes**.

You let two subjects rate two movies once when they are sober and once when they've had 4 beers (we'll call this drunk; more alcohol would count as binge drinking and you don't get ethics approval for that.)

- Repeated measures (i.e., within subjects design)
- The same two movies at all time points for both participants



Fixed Effects Only Model

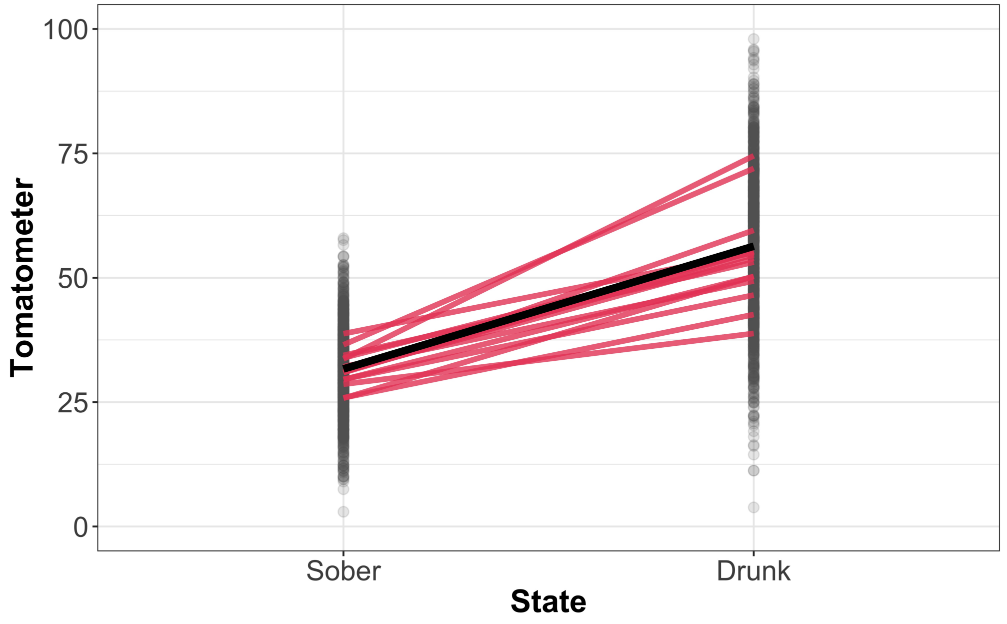


Random Effects

With random effects, we can **account for "random"" variability** in the data that comes from different sources that introduce dependency in the data. We are usually not interested in the concrete instantiation of a particular level.

Random effects:

- are often part of what we want to generalize over
- are always categorical variables (**factors**)
- are thought to be randomly drawn from the population



Singmann & Kellen, 2017

Random Effects - Intercepts and Slopes

We can differentiate two kinds of random effects:

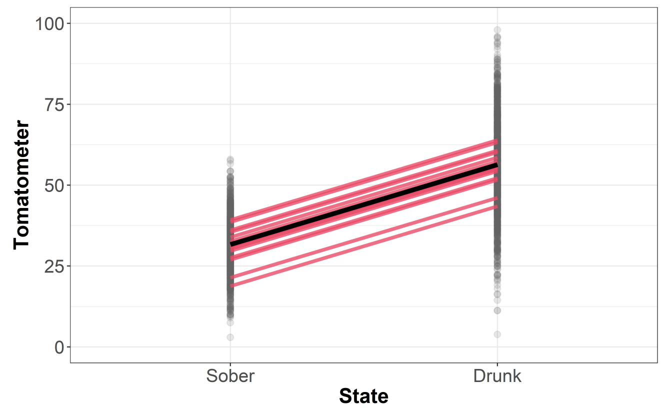
Random Intercepts

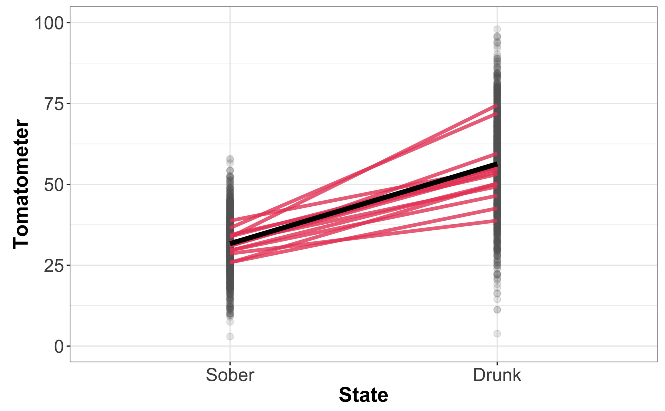
- Estimated average deviations of each random effects factor level from the grand mean.

Random Slopes

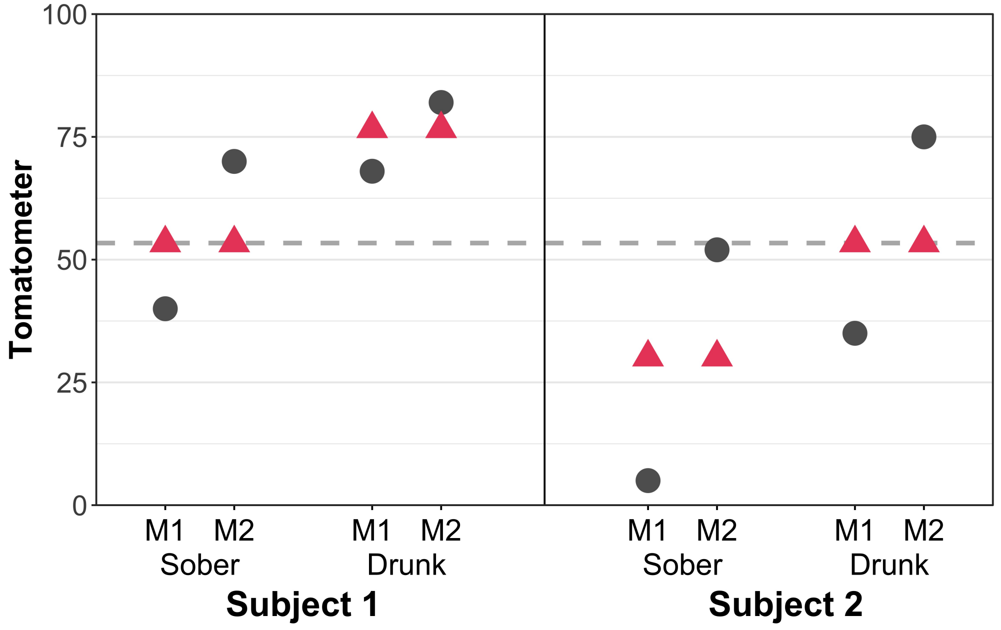
- Estimated deviations of each random effects factor level from the fixed effect slopes.

Note: When estimating random intercepts, only **one additional parameter** is estimated: The **variance** of the normal distribution the random intercepts are thought to stem from.

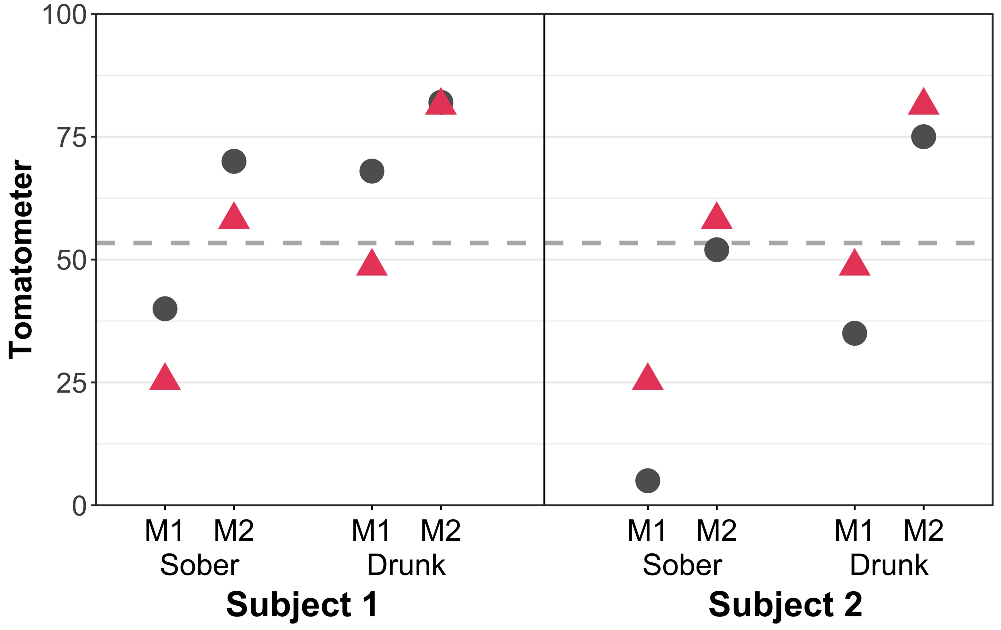




Random Intercepts Model

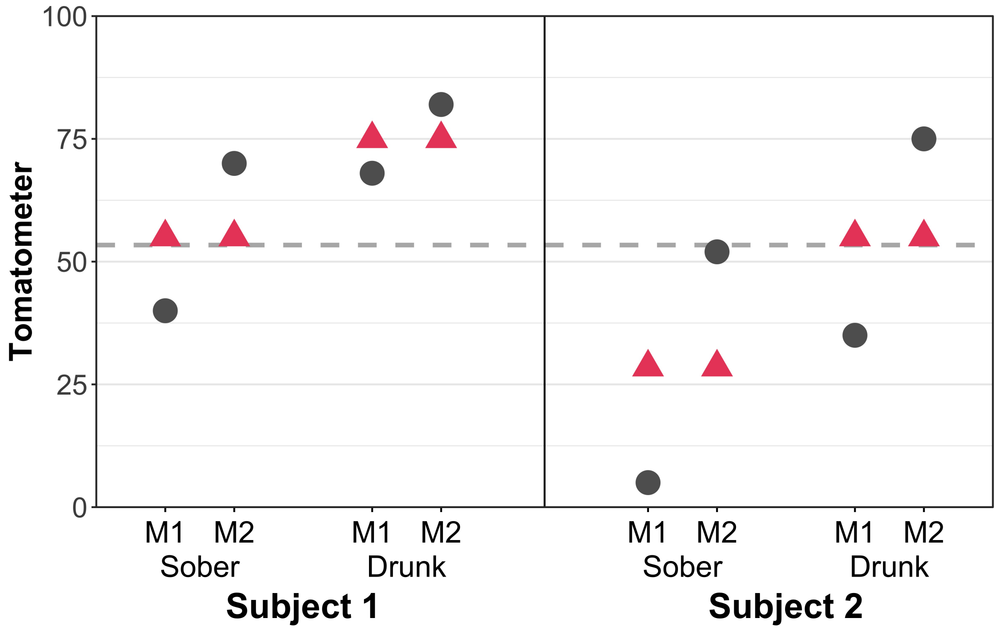


By-Subject Random Intercepts

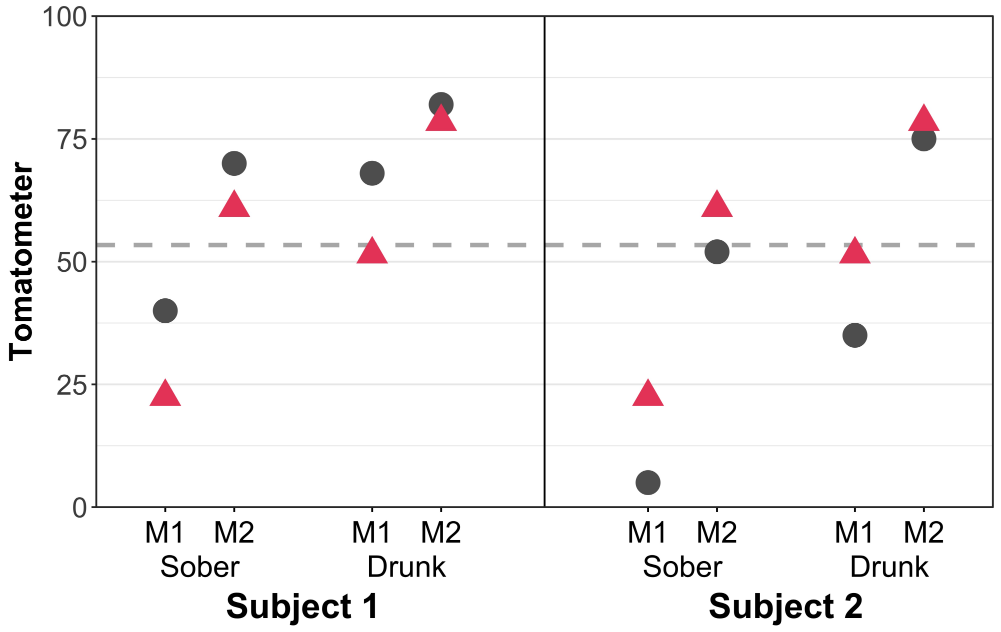


By-Movie Random Intercepts

Random Intercepts and Slopes Model

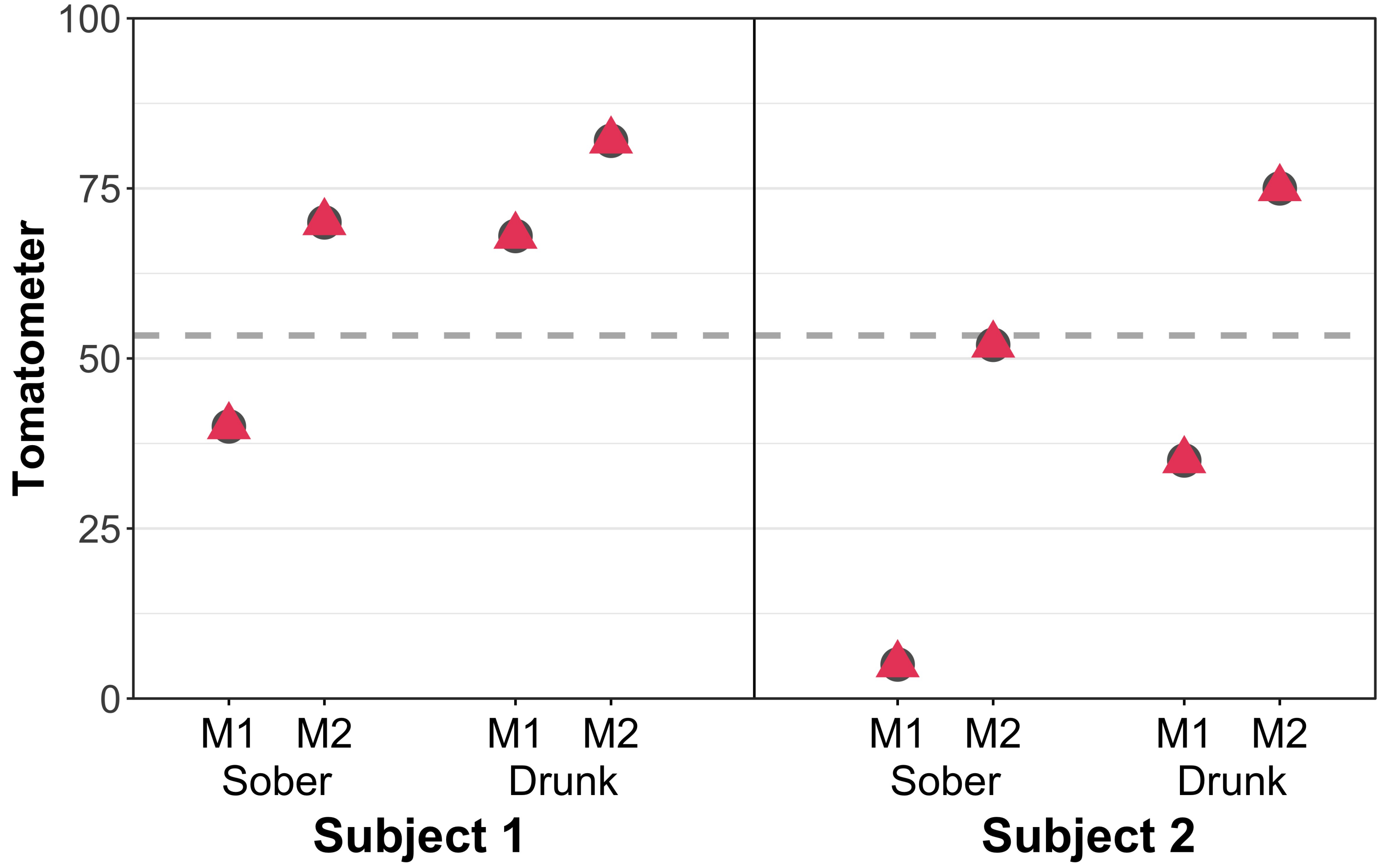


By-Subject Random Intercepts and Slopes



By-Movie Random Intercepts and Slopes

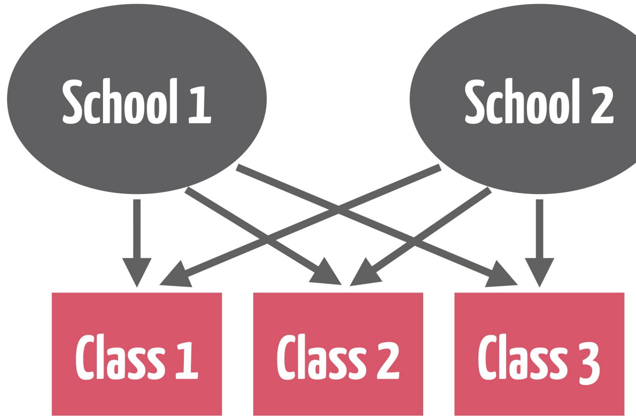
Crossed Random Effects Model



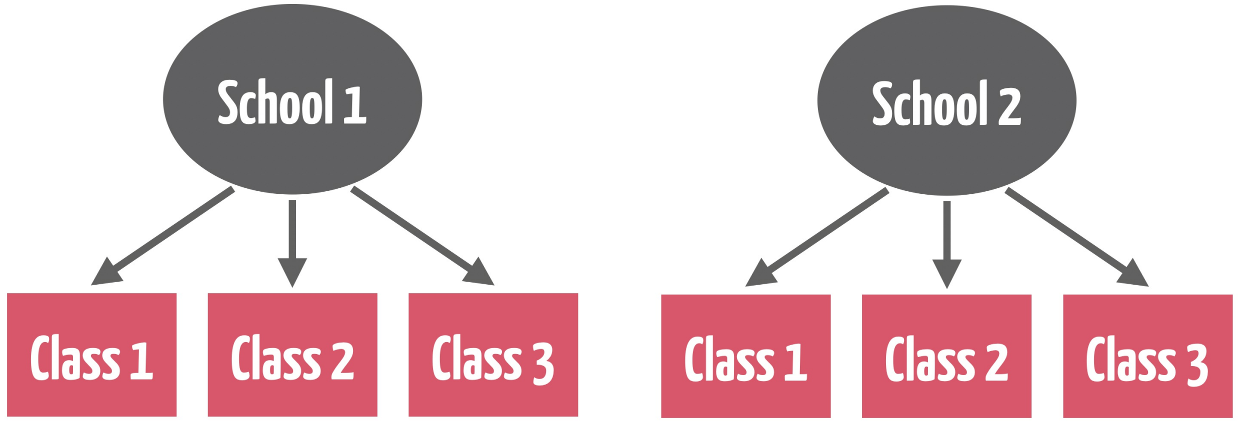
Mixed Effects Model - Equations (In Case You Like Them)

$$
\begin{aligned}
&y_{n,m,d} = \beta_0 + S_{0,n} + I_{0,m} + (\beta_\delta + S_{\delta,n} + I_{\delta,m})X_{n,m,d} + \epsilon_{n,m,d} \\
&n = 1, 2, \ldots, N, \quad m = 1, 2, \ldots, M, \quad d = 1, 2 \\
&\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2) \\
&\begin{pmatrix} S_0 \\ S_\delta \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{S_0}^2 & \rho_{S_0,S_\delta} \\ \rho_{S_\delta,S_0} & \sigma_{S_\delta}^2 \end{bmatrix}\right) \\
&\begin{pmatrix} I_0 \\ I_\delta \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{I_0}^2 & \rho_{I_0,I_\delta} \\ \rho_{I_\delta,I_0} & \sigma_{I_\delta}^2 \end{bmatrix}\right)
\end{aligned}
$$

Crossed vs. Nested Random Effects



Crossed Random Effects



Nested Random Effects

Crossed random effects are all multiple random effects structures that are not nested.

Every level of the nested factor only appears within a single level of the higher order factor.

Selecting the Random Effects Structure

Specifying the correct random effects structure is important. Failure to do so can result in inflated Type I errors or loss of power.

It has been argued that one should specify the **maximal random effects structure justified by the design** (maximal model; Barr et al., 2014).

But specifying the random effects structure when it is not apparent in the data can be overconservative. We will use a compromise and run a **backwards selection procedure** (see Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

from shutterstock.com

Fitting Mixed Effects Models in R

Several packages exist to run mixed effects models in R. Here are some of them:

Package	Function
lme4	`glmer()`, `lmer()`
afex	`mixed()`
rstanarm	`stan_glmer()`, `stan_lmer()`

```
# mixed model using lme4::lmer()

# fixed effects
LMM_out <- lmer(formula = Reaction ~ Days +
# by-ID random slopes and intercepts
                  (Days|Subject),
# define data
                 data = sleepstudy)

# show output
summary(LMM_out)
```

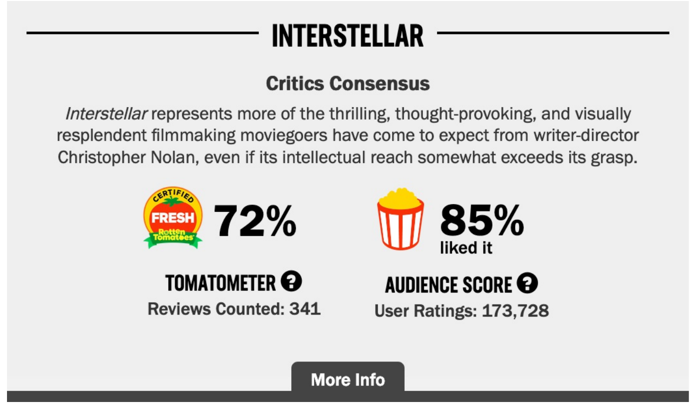
Specifying Random Effects in lme4

Random effects are specified with a new element to the formulas:

<table>
<tr><th>Formula</th><th>Meaning</th></tr>
<tr><td><code>(1|S)</code></td><td>Random intercepts by S</td></tr>
<tr><td><code>(1|S) + (1|I)</code></td><td>Random intercepts by S and I</td></tr>
<tr><td><code>(X1|S)</code> OR <code>(1 + X1|S)</code></td><td>Random intercepts by S and random slopes for X1 by S with correlations</td></tr>
<tr><td><code>(X1 * X2|S)</code></td><td>Random intercepts by S and random slopes for X1, X2, and their interaction <code>X1:X2</code> by S with correlations</td></tr>
<tr><td><code>(0 + X1|S)</code></td><td>Random slopes for X1 by S, no random intercepts</td></tr>
<tr><td><code>(X1||S)</code></td><td>Random intercepts by S and random slopes for X1 by S no correlations</td></tr>
<tr><td><code>(1|S/C)</code> OR <code>(1|S) + (1|S:C)</code></td><td>Random intercepts by S and C with C nested under S</td></tr>
</table>

Fitting Mixed Effects Models in R - An Example

We want to predict the **Tomatometer** rating with **State** as fixed effect and with by-Subjects (**ID**) and by-Item (**Movie**) random slopes and intercepts.



from rottentomatoes.com

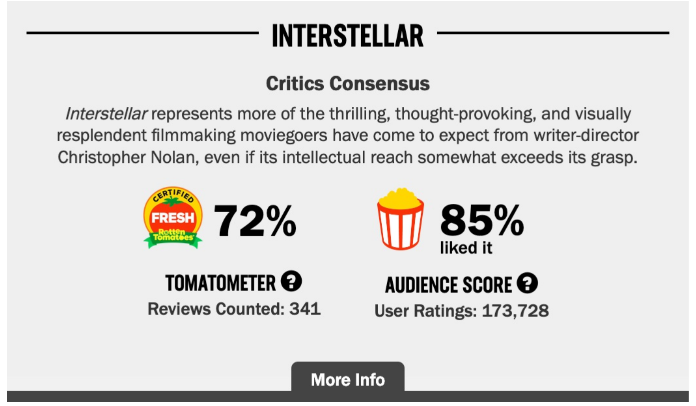
```
# mixed model using lme4::lmer()

# fixed effects
LMM_out <- XXX(formula = XXX ~ XXX +
# by-ID random slopes and intercepts
                   (XXX) +
# by-Movie random slopes and intercepts
                   (XXX),
# define data
                 data = tomato)

# show output
summary(LMM_out)
```

Fitting Mixed Effects Models in R - An Example

We want to predict the **Tomatometer** rating with **State** as fixed effect and with by-Subjects (**ID**) and by-Item (**Movie**) random slopes and intercepts.



from rottentomatoes.com

```r
# mixed model using lme4::lmer()

# fixed effects
LMM_out <- lmer(formula = Tomatometer ~ State +
# by-ID random slopes and intercepts
                  (State|ID) +
# by-Movie random slopes and intercepts
                  (State|Movie),
# define data
                data = tomato)

# show output
summary(LMM_out)
```

`lmer()` Output

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: Tomatometer ~ State + (State | ID) + (State | Movie)
   Data: tom
Control: lmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
 41815.4  41875.7 -20898.7  41797.4     5986

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.6559 -0.6658  0.0043  0.6756  4.3224

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 ID       (Intercept) 19.39    4.404
          StateDrunk  42.90    6.550    0.13
 Movie    (Intercept) 24.87    4.987
          StateDrunk  29.89    5.467    -0.05
 Residual             52.47    7.244
Number of obs: 5995, groups:  ID, 200; Movie, 15

Fixed effects:
            Estimate Std. Error t value
(Intercept)   29.631      1.331   22.25
StateDrunk    25.831      1.497   17.25

Correlation of Fixed Effects:
           (Intr)
StateDrunk -0.042
```

Fit Indices

Random Effect Coefficients

Fixed Effect Coefficients

Practical