**Reporting Templates for Multilevel Regression**

Ben Meuleman, Ph.D.

Swiss Center for Affective Sciences

February 9, 2024

The following are two reporting templates for multilevel regression that you can use as an inspiration for your own reports and papers. Ideally, the Method section of your paper *always* includes a subsection on data analysis, where the important analysis steps are listed, as well as their details. The two templates below concern **(a)** trial data of an experimental factorial design, and **(b)** longitudinal data for an observational design. There are similarities but also some differences, especially with regard to the handling of random effects. References can be found at the bottom of the document.

**Caution**: Do not literally copy-paste this material into your articles. Although these sections often contain the same dry technical information, it is best to try and re/paraphrase according to your own needs and style. Be sure to also check my comments in the margin of the document.

# 1 Trial-data for an experimental factorial design

## *X.X Data analysis*

Data analysis proceeded in <X> steps **(a)** pre-processing, **(b)** descriptive analysis, and **(c)** inferential tests with multilevel regression.

**Pre-processing.** <INSERT PRE-PROCESSING DESCRIPTION>

**Descriptive analysis.** <INSERT DESCRIPTIVE ANALYSIS DESCRIPTION>

**Multilevel regression.** Trial data for the 4 outcome variables were submitted to a separate multilevel ANOVA. A fully factorial 2×2×3 design of A×B×C was included as fixed independent variables, with trial number as a fixed trial-varying covariate, self-reported mood as a fixed C-varying covariate, and gender and age as fixed subject-varying covariates. As random effects, the model included a-priori a random intercept for subject ID, a random trial number slope within subjects, and a random intercept for stimulus ID.[[1]](#footnote-1) This enabled us to take into account individual differences in the outcome due to random sampling of subjects and stimuli, as well as random trial adaptation over time in subjects. In addition, we evaluated the need for more complex random effects by adding A×B random slopes within subjects and stimuli, and then reducing these effects by minimizing Akaike’s Information Criterion (AIC).

For the present study, we opted for multilevel regression due to its flexible handling of repeated measures data, notably **(a)** trial-level data, **(b)** complex data hierarchies with more than one level of repeated data clustering, **(c)** continuous within-subject predictors (e.g., trial number and self-reported mood), **(d)** covariates at any level of data hierarchy, **(e)** missing trial data, and **(f)** the ability to separate population-level effects from individual-level effects. These features are not available in more traditional analysis approaches such as repeated measures (M)ANOVA (Fitzmaurice, Laird & Ware, 2004).

For each outcome model, a Type II ANOVA breakdown using *F*-tests was calculated to investigate significant fixed effects. The primary effect of interest was the three-way A×B×C interaction. If significant, we conducted joint *F*-tests of the two-way A×B interaction within levels of C. If significant, we conducted pairwise *t*-tests of the A effect within combinations of B and C levels. Degrees of freedom for all *F*- and *t*-tests were corrected for the random effects using Satterthwaite’s correction. All inferential tests were conducted at a reduced significance level of α = 0.005, to reduce the number of false positive tests (Benjamin et al., 2018). As effect size, we calculated partial eta squared for *F*-tests and standardized mean differences for *t*-tests. Effects were visualized using model-based effect plots.

Model adequacy and validity of statistical assumptions was checked with several model diagnostics. Overall goodness-of-fit was quantified with marginal and conditional *R2*, with the former quantifying proportion-of-variance explained by the model’s fixed effects (i.e., population effects), and the latter quantifying this proportion for fixed and random effects combined (i.e., population and individual effects). Assumption diagnostics included checks for multicollinearity (with variance inflation factors), influential cases (with Cook’s distance), non-normal residuals (with quantile-quantile plots), and heteroscedastic residuals (with partial dependence plots). In case of violations, we removed collinear effects, removed influential cases, and considered permutation repeated measures ANOVA or generalized least squares regression for non-normal or heteroscedastic residuals.

**Software.** All analyses were conducted using the R statistical software, version <X.X.X>. (R Core Team, <YEAR>), using the packages “lme4” (Bates, Maechler, Bolker & Walker, 2015) and “lmerTest” (Kuznetsova, Brockhoff & Christensen, 2017) for multilevel regression, packages “emmeans” (Lenth, 2022) and “effectsize” (Ben-Shachar, Lüdecke & Makowski, 2020) for follow-up comparisons and effect sizes, packages “car” (Fox & Weisberg, 2019) and “DHARMa” (Hartig, 2022) for model diagnostics, packages “permuco” (Frossard & Renaud, 2021) and “dfadjustSE” (Kolesár, 2023) for non-parametric and heteroscedastic regression, and package “visreg” (Breheny & Burchett, 2017) for model-based effect visualization.

# 2 Longitudinal data for an observational design

## *X.X Data analysis*

Data analysis proceeded in <X> steps **(a)** pre-processing, **(b)** descriptive analysis, and **(c)** inferential tests with multilevel regression.

**Pre-processing.** <INSERT PRE-PROCESSING DESCRIPTION>

**Descriptive analysis.** <INSERT DESCRIPTIVE ANALYSIS DESCRIPTION>

**Multilevel regression.** The longitudinal data across the 10 time points for the outcome was submitted to a multilevel regression. This model was chosen for its flexibility in handling longitudinal data, especially when **(a)** time is a continuous variable, **(b)** there are time-varying covariates, and **(c)** data suffer from missing time points (e.g., due to longitudinal dropout). Handling such challenges is not possible in more traditional analysis approaches such as repeated measures (M)ANOVA, which require strictly categorical time, only subject-varying covariates, and subjects without missing time points.

Model fitting proceeded in two stages, **(I)** random effects selection, and **(II)** fixed effects selection (Fitzmaurice, Laird & Ware, 2004). During random effects selection, we fitted a model with Y as the outcome, and predictor effects of Time, A, B, C, D, E, and F, with an interaction between Time and A. The Time, A, B and C effects were considered confirmatory, whereas the D, E, and F effects were considered exploratory. Conditional on these fixed effects, we tested different random effects structures by fitting different random slopes for the time-varying predictors A, B and C, within a random intercept for subjects. The optimal structure among these was the smallest one that minimized Akaike’s Information Criterion (AIC), within a margin of 2 AIC points.

Once the optimal random slopes were selected, data analysis proceeded to fixed effects selection. In this stage, a conventional Type II ANOVA breakdown of effects with *F*-tests was conducted. The primary effect of interest was the two-way interaction of Time × A. If significant, we conducted *t*-tests to compare the Time slope between levels of A. For predictor B, a significant *F*-test was followed up with pairwise t-tests between levels of B. Finally, the exploratory effects D, E and F were removed from the final model, if non-significant. Degrees of freedom for all *F*- and *t*-tests were corrected for the random effects using Satterthwaite’s correction. All inferential tests were conducted at a reduced significance level of α = 0.005, to reduce the number of false positive tests (Benjamin et al., 2017). As effect size, we calculated partial eta squared for *F*-tests and standardized mean differences for *t*-tests. Effects were visualized using model-based effect plots.

Model adequacy and validity of statistical assumptions was checked with several model diagnostics. Overall goodness-of-fit was quantified with marginal and conditional *R2*, with the former quantifying proportion-of-variance explained by the model’s fixed effects (i.e., population effects), and the latter quantifying this proportion for fixed and random effects (i.e., population and individual effects). Assumption diagnostics included checks for multicollinearity (with variance inflation factors), influential cases (with Cook’s distance), non-normal residuals (with quantile-quantile plots), and heteroscedastic residuals (with partial dependence plots). In case of violations, we removed collinear effects, removed influential cases, and considered quantile multilevel regression or generalized least squares regression for non-normal or heteroscedastic residuals.

**Software.** All analyses were conducted using the R statistical software, version <X.X.X>. (R Core Team, <YEAR>), using the packages “lme4” (Bates, Maechler, Bolker & Walker, 2015) and “lmerTest” (Kuznetsova, Brockhoff & Christensen, 2017) for multilevel regression, packages “emmeans” (Lenth, 2022) and “effectsize” (Ben-Shachar, Lüdecke & Makowski, 2020) for follow-up comparisons and effect sizes, packages “car” (Fox & Weisberg, 2019) and “DHARMa” (Hartig, 2022) for model diagnostics, packages “lqmm” (Geraci, 2014) and “dfadjustSE” (Kolesár, 2023) for robust and heteroscedastic regression, and package “visreg” (Breheny & Burchett, 2017) for model-based effect visualization.

# 3 References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67(1)*, 1–48.

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5(56)*, 2815–2822.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2(1)*, 6–10.

Breheny, P., and Burchett, W. (2017). Visualization of regression models using visreg. *The R Journal*, *9*, 56–71

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004). *Applied Longitudinal Analysis*. Hoboken: Wiley-Interscience

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (Third Edition). Thousand Oaks CA: Sage.

Frossard, J., & Renaud, O. (2021). Permutation tests for regression, ANOVA, and comparison of signals: The permuco package. *Journal of Statistical Software*, *99(15)*, 1–32

Geraci, M. (2014). Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software*, *57(13)*, 1–29.

Hartig, F. (2022). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. R package version 0.4.5.

Kolesár, M. (2023). *dfadjust: Degrees of freedom adjustment for robust standard errors*. R package version 1.0.5.

Kuznetsova, A., Brockhoff, P.B., & Christensen, R.H.B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82(13)*, 1–26.

Lenth, R. (2022). *emmeans: Estimated marginal means, a.k.a. least-squares means*. R package version 1.7.4–1.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

1. In R formula notation: Y~A\*B\*C+Trial+Mood+Gender+Age+(1+Trial|SubID)+(1|StimID) [↑](#footnote-ref-1)