

Proceedings of Meetings on Acoustics

Volume 19, 2013

<http://acousticalsociety.org/>

ICA 2013 Montreal Montreal, Canada 2 - 7 June 2013

Speech Communication

Session 1pSCa: Mixed Effects Modeling: Applications and Practice in Speech Research

1pSCa1. Modeling multi-level factors using linear mixed effects

Cynthia G. Clopper*

*Corresponding author's address: Department of Linguistics, Ohio State University, Columbus, OH 43210, clopper.1@osu.edu

Linear mixed-effects models of 2x2 designs are readily interpretable using treatment contrast coding, although their interpretation is not directly comparable to the interpretation of more traditional ANOVAs. In particular, the interpretation of the "main effect" term for one factor holds only for the reference level of the other factor. Interpreting models of designs involving factors with more than two levels and/or interactions involving more than two factors is more complex and even less comparable to familiar interpretations of ANOVAs. Alternative methods for analyzing these more complex designs include using different contrast coding (e.g., sum, Helmert, or custom), selecting specific reference levels for the factors, and running multiple models of the same data set with different reference levels of comparison. These methods may return quite different results, however, such that a significant factor with treatment contrast coding may not be significant with sum contrast coding or a significant interaction term in one model may not be significant when the reference levels of the relevant factors are changed. Thus, although linear mixed effects provide an opportunity to model complex designs with multiple sources of variability, this modeling requires careful consideration of model parameters to achieve the most appropriate interpretation of the data.

Published by the Acoustical Society of America through the American Institute of Physics

INTRODUCTION

Linear mixed-effects models have recently been proposed as an alternative to more traditional analyses of variance in speech science research (Baayen, Davidson, and Bates, 2008; Jaeger, 2008; Quene and van den Bergh, 2004, 2008). The primary benefits of these models that motivate their adoption are the ability to examine crossed subject and item random effects (Baayen et al., 2008; Quene and van den Bergh, 2008), the ability to straightforwardly explore both continuous and categorical independent variables in a single model (Baayen et al., 2008), and the ability to treat binary accuracy responses appropriately using logistic models, while also accounting for both subject and item variability (Jaeger, 2008; Quene and van den Bergh, 2008). These arguments in favor of mixed-effects models are compelling and the models are therefore gaining popularity in the field.

However, the process of interpreting mixed-effects models can differ substantially from the process of interpreting analyses of variance. In particular, although both analyses of variance and mixed-effects models are generalized linear models, the typical outputs produced by statistical software packages differ for the two analyses. With analysis of variance, the model output provides a series of main effect and interaction terms comparing group means, and the underlying linear model parameters are typically not presented. However, with mixed-effects models, the model output consists only of the linear model parameters themselves. Thus, the two sets of outputs must be interpreted differently, even though the mixed-effects model parameters are typically similar in name to the main effect and interaction terms produced by analysis of variance. In addition, the interpretation of mixed-effects models may be substantially impacted by seemingly trivial decisions in model-building, such as the ordering of levels within factors and the contrast structure of each factor. Although contrast structure is also relevant for analyses of variance, it emerges as part of the planned comparisons or posthoc tests that follow the omnibus analysis. In mixed-effects models, however, the contrast structure directly impacts the model parameters that are returned and interpreted. The goal of this paper is to illustrate these issues of interpretation with examples of simple 2 x 2 factorial designs as well as more complex designs involving factors with more levels and their interactions.

INTERPRETING 2 X 2 FACTORIAL DESIGNS

Although their interpretation is not directly comparable to the interpretation of more traditional ANOVAs, linear mixed-effects models of 2 x 2 designs using the default treatment contrast coding in R are readily interpretable. In our analysis of cross-dialect word recognition in noise, Clopper, Pierrehumbert, and Tamati (2010) observed a significant interaction between talker regional dialect (Midland vs. Northern American English) and lexical competition (presence vs. absence of a real word competitor) in recognition accuracy for /a/ in CVC words. As shown in Table 1, performance was less accurate for the Midland talkers when the target word did not have a real-word competitor minimal pair (e.g., *mob*, where *mab* is not a real word) than when the target word had a real-word competitor minimal pair (e.g., *mop*, where *map* is a real word)¹, whereas no effect of lexical competition was observed for the Northern talkers. As reported by Clopper et al. (2010), subject and item ANOVAs revealed a significant main effect of talker dialect and a significant talker dialect x lexical competition interaction in the subjects analysis and no significant main effects or interactions in the items analysis.

TABLE 1. Mean vowel recognition accuracy for /a/ in Clopper et al.'s (2010) cross-dialect word recognition study.

Talker Dialect	Competitor	No Competitor
Midland	74%	60%
North	59%	59%

A reanalysis of the data using a logistic mixed-effects model with subject and item as crossed random effects confirms the significant talker dialect x lexical competition interaction. Table 2 shows the results of two mixed-effects analyses of these data to illustrate the effect of factor levels on output and interpretation. In the model on the left, the default treatment coding in R was used, which means that factor levels were coded in alphabetical order and the first level alphabetically of each factor served as the reference level. Thus, the reference level of the lexical

¹ Although this effect is in the opposite direction of what would be expected, it reflects the acoustic difference between the two sets of words. The vowel /a/ in words in the No Competitor condition was acoustically more similar to, and therefore more confusable with, the vowel /ʌ/ than the vowel /a/ in words in the Competitor condition.

competition factor was “no competitor” and the reference level of the talker dialect factor was “Midland”. Given this factor specification, the intercept captures performance on the Midland talkers in the No Competitor condition, the talker dialect coefficient captures the difference between the Midland and Northern talkers in the No Competitor condition, and the competitor coefficient captures the difference between the No Competitor and Competitor conditions for the Midland talkers. In this model, the talker dialect coefficient is very small and not significant, because listeners performed similarly on the two talker dialects in the No Competitor condition. In the model on the right, the reference level of the lexical competition factor was changed to “word competitor” so that the talker dialect coefficient captures the difference between the Midland and Northern talkers in the Competitor condition. The coefficient is much larger, although still not significant, reflecting the larger effect of dialect in the Competitor condition than the No Competitor condition (see Table 1). The results of both models differ from the results obtained in the subjects ANOVA, in which a main effect of talker dialect was observed. This difference could be the result of including both subject and item as random effects in the mixed-effects model, given that the talker dialect effect was also not significant in the items ANOVA. However, the mixed-effects model parameters also do not lend themselves to an interpretation of main effects similar to those obtained in ANOVAs. In particular, the talker dialect coefficients in the two models shown in Table 2 reflect simple effects of talker dialect in the No Competitor (Model 1) and Competitor (Model 2) conditions. Thus, unlike in the original ANOVAs reported by Clopper et al. (2010), the overall effect of talker dialect, independent of lexical competition condition, cannot be directly assessed in these mixed-effects models.

TABLE 2. Logistic mixed-effects analyses of Clopper et al.’s (2010) /a/ accuracy data.

Model 1	Coefficient (p-value)	Model 2	Coefficient (p-value)
Intercept	.46	Intercept	1.20
Competitor (Word)	.74 (p = .17)	Competitor (None)	-.74 (p = .17)
Talker Dialect (North)	.02 (p = .98)	Talker Dialect (North)	-.73 (p = .23)
Word Competitor x North	-.75 (p = .006)	No Competitor x North	.74 (p = .006)

INTERPRETING MORE COMPLEX DESIGNS

Careful consideration of how factor levels are coded is even more critical when the design involves factors with more than two levels and/or interactions in designs that are larger than 2 x 2. In addition to carefully selecting specific reference levels for the factors and running multiple models of the same data set with different reference levels of comparison, as demonstrated above for the 2 x 2 design, different contrast coding schemes (e.g., sum, Helmert, or custom contrasts) may be fruitfully explored. Crucially, changing the coding scheme fundamentally changes the comparisons that the model parameters capture. For example, the treatment contrast coding used above compares each factor level to the specified reference level, whereas sum contrast coding compares each factor level to the grand mean of all levels. Therefore, two models containing the same factors as predictor variables, but with different coding schemes, may return quite different results, such that a significant factor with treatment contrast coding may not be significant with sum contrast coding or a significant interaction term in one model may not be significant when the reference levels of the relevant factors are changed.

Factors with More than Two Levels

Table 3 shows mean speaking rates in read passages for the 10 talkers from each of the six dialect regions in the Nationwide Speech Project corpus (Clopper and Pisoni, 2006)². Midland talkers exhibited the slowest speaking rate overall and New England talkers exhibited the fastest rate. Linear mixed-effects models with treatment coding of dialect as a fixed effect and talker and passage as random effects revealed significant effects of talker dialect on speaking rate. However, these effects were highly dependent on the reference level of the talker dialect factor. As noted above, in R, the reference level of a factor defaults to the first level alphabetically, which in this case was Mid-Atlantic. When Mid-Atlantic was the reference level of the talker dialect factor, none of the dialect comparisons were significant. However, when Midland was selected as the reference level of the talker dialect factor, a significant difference between the slow Midland and the fast New England dialects was obtained ($\beta = .37$, pMCMC = .01). When New England was selected as the reference level of the talker dialect factor, significant

² The speaking rate analysis of these data was conducted in collaboration with Rajka Smiljanic.

differences between the fast New England and the slower Midland and Southern dialects were obtained ($\beta < -.33$, $pMCMC < .03$ for both comparisons).

TABLE 3. Mean speaking rate (syllables/s) in read passages in the NSP corpus (Clopper and Pisoni, 2006).

Talker Dialect	Speaking Rate
New England	5.73
Mid-Atlantic	5.62
North	5.54
Midland	5.36
South	5.40
West	5.50

Thus, unlike in the reanalysis of Clopper et al.'s (2010) word recognition data, the interpretation of the results is critically tied to the selected reference levels of each factor when treatment coding is used. In this analysis, in which the talker dialect factor had six levels, when the reference level was set to the alphabetically-determined default, no effects of talker dialect emerged. However, when the reference level was specified to correspond to the observed minimum or maximum mean values so that comparisons with the most extreme speaking rates could be examined, significant results were obtained. In addition, as noted above, no overall measure of the talker dialect effect is returned and each model coefficient can only be interpreted relative to the intercept. Given that the number of possible within-factor comparisons is limited to one fewer than the number of levels in the factor, multiple models are needed to explore all possible pairwise comparisons within a given factor.

One alternative to specifying the reference level of each factor using treatment coding and running multiple models to compare the factor levels to each other is to use a different set of contrasts, such as sum contrasts, in which the intercept corresponds to the grand mean, or Helmert contrasts, in which each level is compared to the mean of all subsequent levels combined. Given that the levels of the dialect factor for the data shown in Table 3 are not intrinsically ordered, Helmert contrasts would be difficult to interpret because "subsequent" levels of the factor are not uniquely defined. However, a sum contrast analysis would allow us to determine if any of the talker dialects differ significantly in speaking rate from the sample of talkers as a whole. A linear mixed-effects model with sum contrast coding of talker dialect as a fixed effect and talker and passage as random effects revealed one significant effect of dialect on speaking rate. As suggested by the previous models with treatment coding, New England emerged as significantly faster than the overall mean ($\beta = .21$, $pMCMC = .02$). Thus, both treatment contrasts and sum contrasts suggest that the New England speaking rate is significantly faster than the speaking rate of other dialect regions in the United States. However, unlike the treatment contrast analysis, the sum contrast analysis does not provide specific information about which varieties have a significantly slower speaking rate than New England.

Like the treatment contrast analysis, the sum contrast analysis also only returns coefficients for one fewer than the number of levels in a factor group. Although the N^{th} coefficient is equal to the negative sum of the $N-1$ coefficients and can therefore be calculated by hand from the model output, the statistical significance of the N^{th} coefficient cannot be directly determined. For example, a linear-mixed effects model of the speaking rate data with sum contrast coding of talker dialect in which the coefficient for the New England level is unspecified does not return any significant talker dialect coefficients. Thus, for factors with more than two levels, an analysis with sum contrasts may require the construction of two models so that coefficients for all levels of a given factor can be compared to the grand mean, just as an exploration of specific comparisons with treatment contrasts will typically require the construction and interpretation of multiple different models so that all relevant comparisons can be examined. As in traditional posthoc analyses, this process of constructing multiple models of the same data set should be undertaken cautiously so that the probability of Type I errors does not increase dramatically.

Interactions Involving Factors with More than Two Levels

In addition to talker dialect and lexical competition, Clopper et al.'s (2010) word recognition study also examined the effects of listener dialect on performance. In the ANOVAs that we reported, a significant interaction between listener dialect and lexical competition was observed for recognition accuracy for /ɔ/. As shown in Table 4, listener dialect was a three-level factor that distinguished among Northern, Midland, and Mobile listeners, where Mobile listeners were defined as those who had lived in more than one dialect region prior to age 18. The significant interaction reflects the more accurate performance for the Northern listeners than the Midland and Mobile listeners

for /ɔ/ in words with a real-word competitor (e.g., *caught*, where *cot* is a real word). The listener dialect groups did not differ in the No Competitor condition for /ɔ/ (e.g., *fought*, where *fot* is not a real word).

TABLE 4. Mean vowel recognition accuracy for /ɔ/ in Clopper et al.'s (2010) cross-dialect word recognition study.

Listener Dialect	Competitor	No Competitor
Midland	68%	83%
North	81%	79%
Mobile	63%	86%

A reanalysis of these data using logistic mixed-effects models with subject and item as crossed random effects confirms the listener dialect x lexical competition interaction, but only for some specifications of the factor reference levels using treatment coding. The default model with No Competitor as the reference level of the lexical competition factor and Midland as the reference level of the listener dialect factor reveals no significant main effects or interactions because the Midland listeners' performance was intermediate to the Northern and Mobile listeners' performance in both lexical competition conditions. However, models with Competitor as the reference level of the lexical competition factor and either North or Mobile as the reference level of the listener dialect factor reveal the significant listener dialect x lexical competition interaction, as shown in Table 5.

TABLE 5. Logistic mixed-effects analyses of Clopper et al.'s (2010) /ɔ/ accuracy data.

Model 1	Coefficient (p-value)	Model 2	Coefficient (p-value)
Intercept	.47	Intercept	1.38
Competitor (None)	1.83 (p = .03)	Competitor (None)	.14 (p = .87)
Listener Dialect (Midland)	.38 (p = .34)	Listener Dialect (Midland)	-.54 (p = .25)
Listener Dialect (North)	.91 (p = .07)	Listener Dialect (Mobile)	-.91 (p = .07)
No Competitor x Midland	-.65 (p = .33)	No Competitor x Midland	1.05 (p = .08)
No Competitor x North	-1.69 (p = .03)	No Competitor x Mobile	1.69 (p = .03)

In Model 1, Mobile was the reference level of the listener dialect factor and significant coefficients were observed for lexical competition condition and the lexical competition x listener dialect interaction for Northern listeners. This combination of results captures the significant improvement in performance for the Mobile listeners in the No Competitor condition relative to the Competitor condition as well as the lack of improvement for the Northern listeners across the two lexical competition conditions because their performance in the Competitor condition was already excellent (see Table 4). In Model 2, North was the reference level of the listener dialect factor and a significant coefficient was observed only for the lexical competition x listener dialect interaction for Mobile listeners. This combination of results similarly captures the large increase in performance for the Mobile listeners across lexical competition conditions that was not observed for the Northern listeners. Note that the difference between the Mobile and Northern listeners' performance in the Competitor condition is marginally significant in both models (p = .07) and that the lexical competition x listener dialect interaction is marginally significant (p = .08) for the Midland listeners in Model 2, but not Model 1 (p = .33), consistent with our interpretation of the overall pattern as reflecting a lexical competition effect for the Midland and Mobile listeners, but not the Northern listeners.

Thus, the selection of the reference level of individual factors is essential both for interpreting the effects of those factors, as shown above for the six-level dialect factor in the speaking rate analysis, and for interpreting the interactions among factors, as shown here for the listener dialect x lexical competition interaction in the word recognition analysis. In addition, as noted above, the "main effect" terms in mixed-effects models reflect simple effects in which all other factors have their reference value and therefore cannot be interpreted like main effects in ANOVAs. In this analysis, for example, the Competitor coefficient was significant in Model 1, but not Model 2, reflecting the simple effect of lexical competition condition for the Mobile listeners (Model 1) and the lack of an effect of lexical competition condition for the Northern listeners (Model 2).

CONCLUSIONS

Taken together, these examples suggest that although linear mixed-effects analyses provide an opportunity to model complex designs with multiple sources of variability, this modeling requires careful consideration of model parameters to achieve the most appropriate interpretation of the data. These considerations include the selection of factor reference levels and contrast coding schemes, which can have a substantial impact on the model parameters.

Although this variability in model output may superficially seem to affect the results, careful interpretation of model coefficients in terms of the comparisons that they capture should ultimately reveal a coherent understanding of the data. In particular, both “main effects” and null results obtained in mixed-effects models must be very cautiously interpreted within the context of the factor specifications and other model parameters, so that significant effects are neither over-interpreted nor overlooked.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation (BCS-1056409 to Ohio State University). I would like to thank Jeonghwa Shin, Rory Turnbull, and Kodi Weatherholtz for helpful discussions about contrast coding in R.

REFERENCES

- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). “Mixed-effects modeling with crossed random effects for subjects and items,” *J. Mem. Lang.* **59**, 390-412.
- Clopper, C. G., Pierrehumbert, J. B., and Tamati, T. N. (2010). “Lexical neighborhoods and phonological confusability in cross-dialect word recognition in noise,” *Lab. Phonol.* **1**, 65-92.
- Clopper, C. G., and Pisoni, D. B. (2006). “Effects of region of origin and geographic mobility on perceptual dialect categorization,” *Lang. Var. Change* **18**, 193-221.
- Jaeger, T. F. (2008). “Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models,” *J. Mem. Lang.* **59**, 434-446.
- Quene, H., and van den Bergh, H. (2004). “On multi-level modeling of data from repeated measures designs: A tutorial,” *Speech Commun.* **43**, 103-121.
- Quene, H., and van den Bergh, H. (2008). “Examples of mixed-effects modeling with crossed random effects and with binomial data,” *J. Mem. Lang.* **59**, 413-425.