CHAPTER 22

STATISTICAL ANALYSES

STATISTICS IN LABORATORY PHONOLOGY

JOHN KINGSTON

MIXED-EFFECTS MODELS

HARALD BAAYEN

CLUSTERING AND CLASSIFICATION METHODS

CYNTHIA G. CLOPPER

The contributions in this chapter review statistical techniques appropriate for speech research. Kingston presents two in-depth case studies, discussing graphical data exploration and analysis using linear regression models. Baayen discusses the principles and applications of mixed-effects models, including consideration of continuous, binary, and count data. Clopper discusses clustering, multidimensional scaling, and factor analysis.

22.1 Statistical methods in laboratory phonology

John Kingston

22.1.1 Introduction

Statistics is one of the methods that constitute laboratory phonology. In this section, I use them to tell a clear story about two exemplary data sets. In one, the dependent variable is continuous, while in the other it is an ordinal categorical variable. Both kinds of dependent variables are commonly produced by phonological experiments; a third common kind is categorical choices between two alternatives, which Baayen (this chapter) deals with in his section. The analyses all begin with graphical exploration of the data, which is followed up by constructing linear regression models. These models are more informative than analyses of variance about how the kinds of independent variables commonly used in phonological experiments influence the dependent variable. Mixed-effects models are developed for the ordinal categorical variable to demonstrate how random effects of participants and items are accommodated.[1]

22.1.2 A linear model of Finnish vowel durations

22.1.2.1 *Graphical explorations*

The data that demonstrate the analysis of a continuous dependent variable were generously provided by Scott Myers, who with Benjamin Hansen used them to develop a phonetic explanation for the cross-linguistically common process of final vowel shortening (Myers and Hansen 2006, 2007). They consist of the total durations of Finnish vowels, as well as the durations of their voiced and voiceless portions. The potential independent variables are the phonological quantity of the vowel (short versus long), the type of syllable it occurred in (open: V, CV, or GV and closed CVN), and whether the syllable containing the vowel was word-final. A fourth independent variable was the duration of the preceding word *sanoin* 'I

[1] All the analyses presented in this section were carried out in R (R Development Core Team, 2010). Besides the base package, the principal packages used in this section are languageR (Baayen 2009), lattice (Sarkar 2010), ordinal (Christensen 2010), and lme4 (Bates and Maechler, 2010). For more comprehensive applications of R to linguistic data, see Baayen (2008) and Johnson (2008); Dalgaard (2002), Maindonald and Braun (2003), Gelman and Hill (2007), and Everitt and Hothorn (2010) are also very useful introductions.

said,' which may index speaking rate. Four speakers each produced twelve tokens in different words of the sixteen kinds of syllables.

Figure 22.1.1 displays the total durations of the short and long vowels in histograms, density plots superimposed on the histograms, and box plots.

These displays show that most long vowels last longer than most short ones, despite some overlap in their respective ranges. Two modes are visible in the distributions of both short and long vowels' durations for speakers 3 and 4, and the distributions for speakers 1 and 2 also have obvious shoulders on their upper tails. This structure raises the suspicion that the vowels' durations may be determined by another factor than their phonological quantity. The breakdown of the density plots in Figure 22.1.2 by whether the syllable containing the vowel is final or non-final confirms this suspicion. Final and non-final distributions overlap for both short and long vowels, but their distributions remain largely distinct from one another within each quantity. The lack of overlap between the notched intervals in the

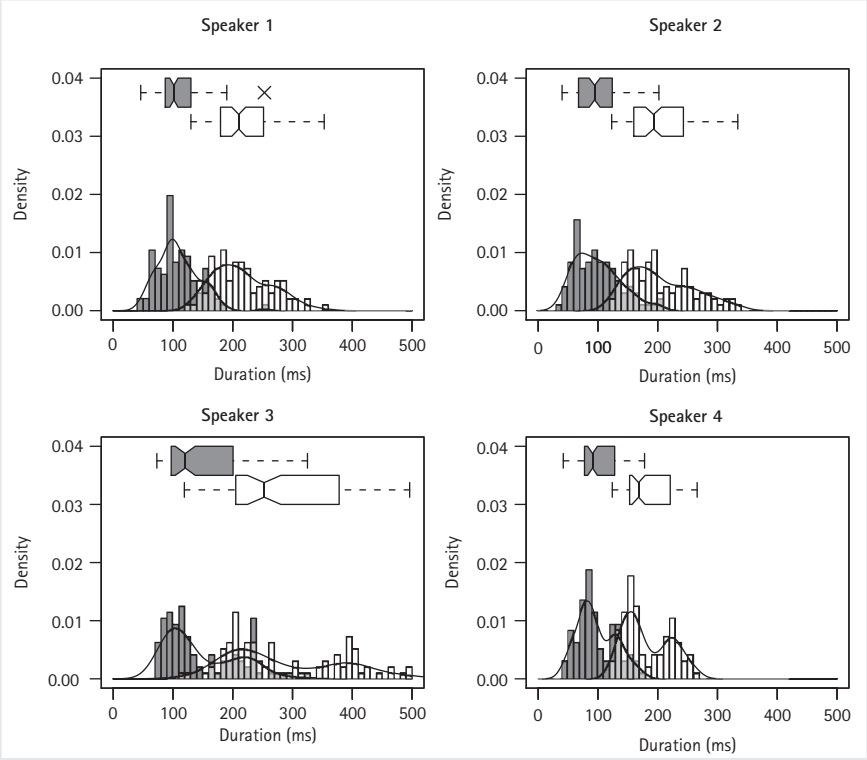

Figure 22.1.1. Histograms, density plots, and box plots of short (dark gray) and long (white) vowel durations produced by four Finnish speakers. The lighter gray bars represent durations common to short and long vowels. The Xs are outliers.

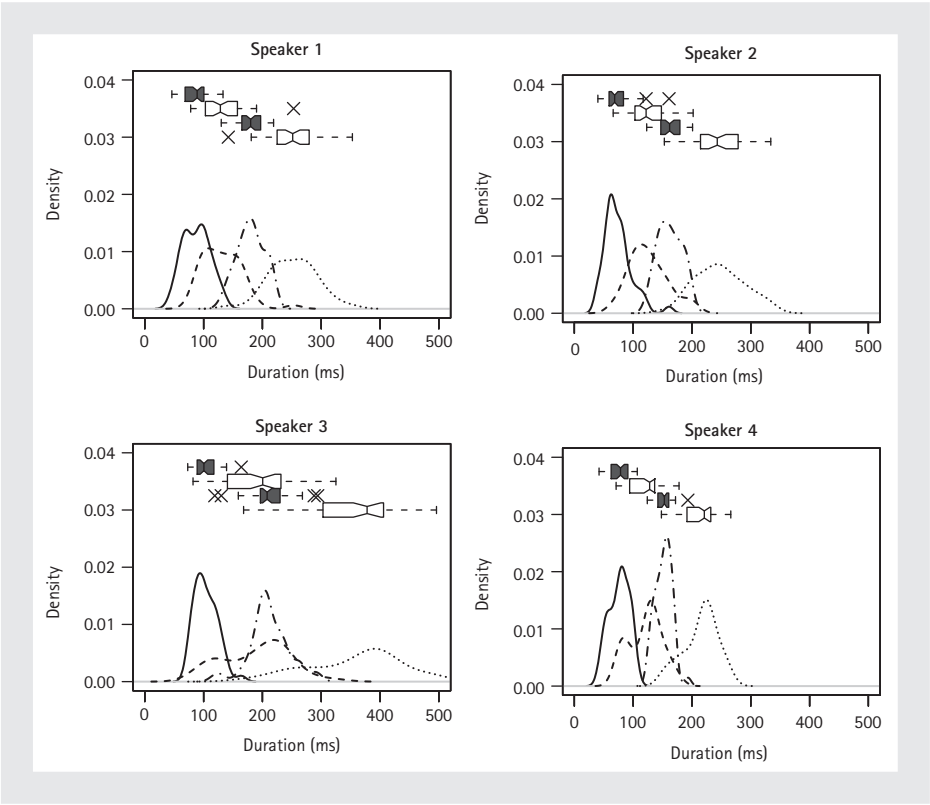

Figure 22.1.2. Density plots for short (solid, dashed) and long (dash-dot, dotted) vowels in final (dashed, dotted) and non-final (solid, dash-dot) position for the four Finnish speakers. Gray box plots represent durations in non-final position, white ones those in final position. The Xs are outliers.

box plots also indicates that the final and non-final distributions differ from one another. The density plots in this figure hint at the presence of yet further structure in the data, perhaps an influence of syllable type.

Evidence of such structure can be seen in Figure 22.1.3, where each panel plots the vowel duration for a particular speaker, vowel quantity, and position against the duration of the word *sanoin*—this figure uses the xyplot() function from the lattice package (Sarkar 2010). The durations of most vowels in open CV, GV, and V ("1–3") are longer than those in closed CVN ("X") syllables in final position for both long and short vowels, except for the short vowels produced by speaker 1 (bottom row), whose vowel durations in open and closed syllables overlap considerably. In non-final position, the distributions of vowel durations in closed and open syllables overlap for all four speakers.

The graphical exploration in Figures 22.1.1–22.1.3 has shown that besides the expected greater duration of phonologically long versus short vowels, vowels in

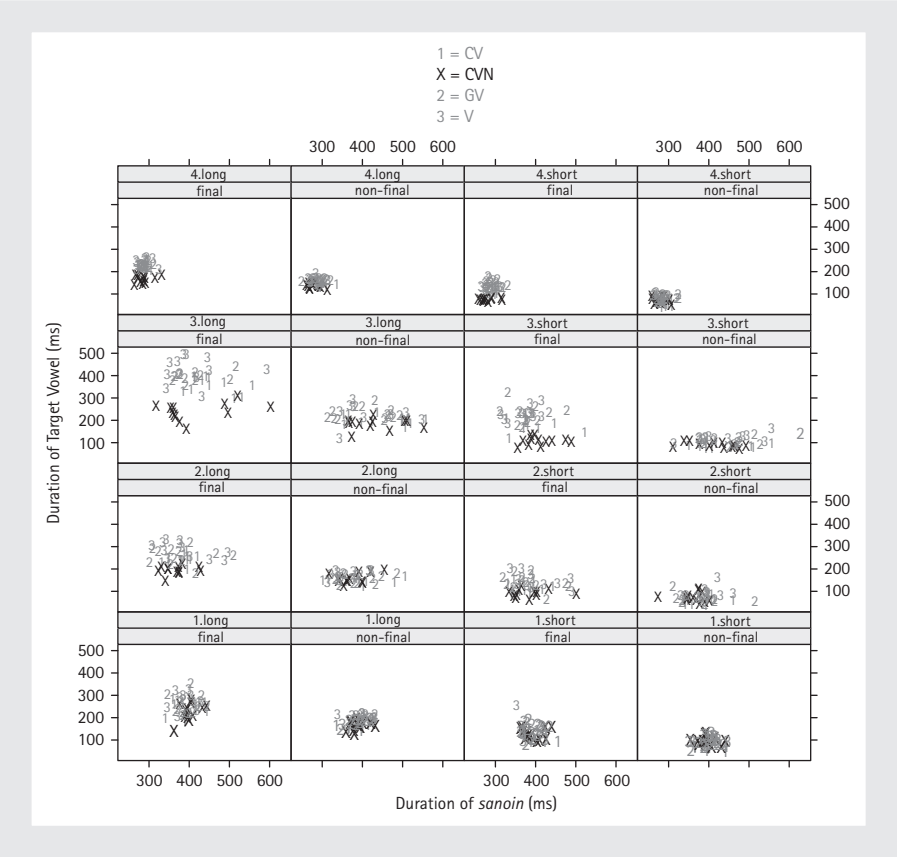

Figure 22.1.3. Vowel durations (vertical axis) by duration of the word *sanoin* in the same utterance by speaker, phonological quantity, position, and syllable type. Black "X" for closed CVN syllables, and gray "1–3" for open CV, GV, and V syllables, respectively.

final position are longer than non-final vowels, and vowels are longer in open than closed syllables when they occur in final position.

22.1.2.2 *Residuals and transforms of the dependent variable*

Unlike the duration of the word *sanoin*, the other independent variables are categorical rather than continuous. This characteristic might prompt submitting these data to an analysis of variance instead of linear regression. The reason for not doing so is that we want to know not only whether any of these independent variables significantly affects vowel durations but also the direction and size of that effect. We may even have hypotheses about the direction of these effects that we would like to test.

To examine deviations of the dependent variable's values from the predicted value, i.e. the residuals, linear regression models were first constructed with just one independent variable at a time. This examination reveals whether the dependent variable needs to be transformed before undergoing further analysis.

The three panels in each row of Figure 22.1.4 display the residuals against the fitted values (left), a QQ-plot of the residuals against the values expected if the residuals were normally distributed (middle), and the Cook's distance values for the residuals (right). Each vertical cluster of residuals in the panels on the left represents a combination of a value for the independent variable and speaker. Ideally, the vertical distances of the residual values from 0 would not vary as a function of the fitted values, but in all three panels we see that they spread out as fitted values increase. This outcome indicates the need to transform the data. The QQ-plots in the middle panels show that the positive residuals have more extreme values than expected,

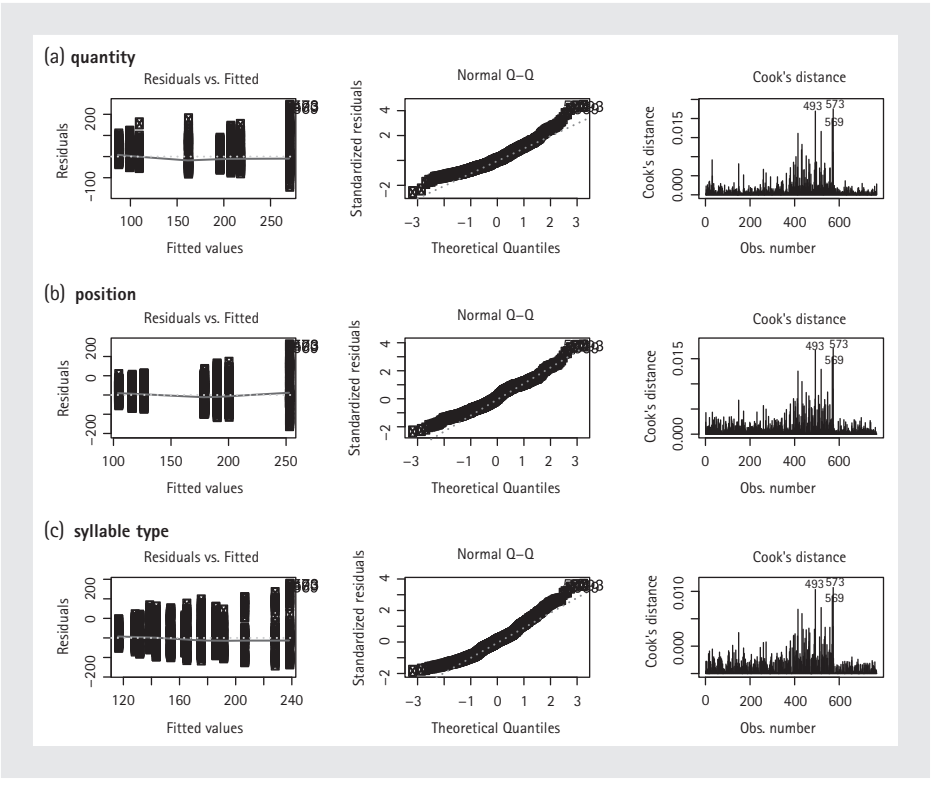

Figure 22.1.4. Residuals by fitted values (left), QQ-plots of standardized residuals by values expected if the residuals were normally distributed (middle), and Cook's distances (right) for the models in which the independent variables are (a) quantity and speaker, (b) position and speaker, and (c) syllable type and speaker.

while the values of the negative residuals are less extreme than expected. This is not surprising given the fanning outward observable in the left-hand panels. Finally, the Cook's distance values in the right-hand panels identify data points that exert "leverage" on the regression line. When data points exert considerable leverage, the analysis should be rerun omitting those data points to determine whether the independent variable's apparent influences depend on just those points. Values of 1 or more indicate influential leverage. Here, the most extreme values are 1–2 orders of magnitude smaller than 1, so none exerts particularly strong leverage, and none should be omitted.

The choice of power transform (1) is determined by the tails of the dependent variable's distribution: a long upper tail (most outliers have high values) indicates an exponent smaller than 1 (λ)—in the limit, the exponent is 0 and the transform is the log transform—but a long lower tail indicates an exponent larger than 1.

$$(1) \quad T(x) = \frac{x^{\lambda} - 1}{\lambda}$$

Figures 22.1.1 and 22.1.2 show that the distributions' upper tails are stretched out, motivating a log transform. The left-hand panels in Figure 22.1.5 show that this transformation successfully eliminates the residual values' fanning outward as the fitted values increase, but the QQ plots in the middle panels show that the residuals' distribution now differs more from normality than before the transformation. The deviation is also different: both negative and positive residuals are less extreme than if they were normally distributed. These deviations are also much greater for the models of position (b) and syllable type (c) than quantity (a). This complementarity between the deviations in the models of position and syllable type, on the one hand, and quantity, on the other, motivates including quantity as a predictor.

22.1.2.3 *Multiple linear regression models, with and without interactions*

We begin with a model that includes all three independent variables as well as the log-transformed duration of the word *sanoin*, a main effects model (Table 22.1.1), and follow up by adding an interaction between position and syllable type (Table 22.1.2). The dependent variable is the log-transformed vowel durations.

The vowel durations do not depend significantly on the duration of *sanoin* ($t = -0.622$, $p = 0.534$; recall Figure 22.1.3), the vowel is significantly shorter when it is phonologically short ($t = -49.826$, $p < 2e-16$) or non-final ($t = -31.735$, $p < 2e-16$), but significantly longer in a GV ($t = 7.513$, $p = 1.64e-13$) or V ($t = 11.322$, $p < 2e-16$) syllable compared to a CV syllable. It is marginally shorter ($t = -1.895$, $p = 0.0585$) in a CVN syllable. To transform the predicted vowel durations back into ms, the model serves as the exponent of the base e (or practically

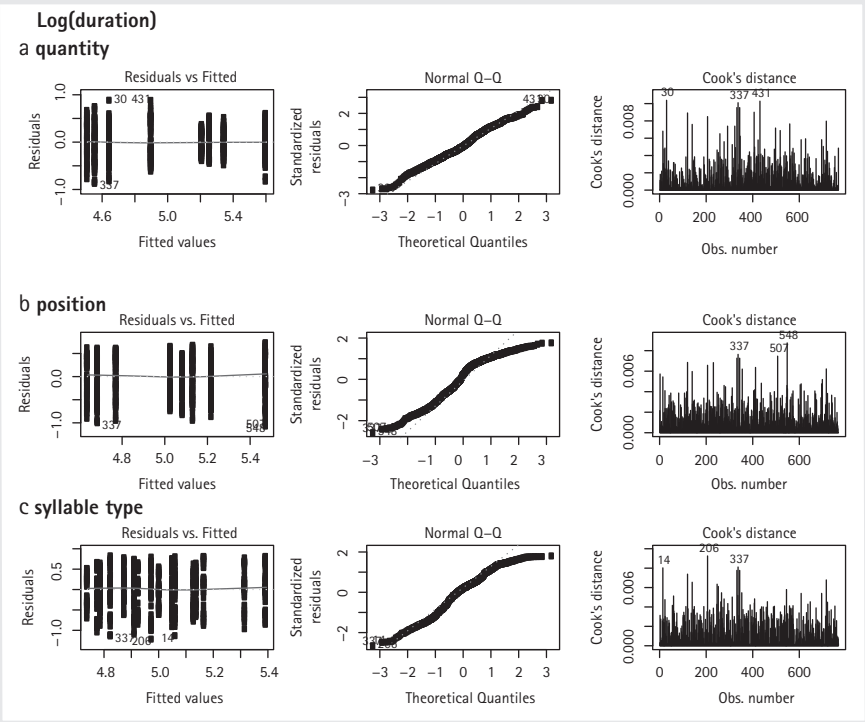

Figure 22.1.5. For log-transformed durations, residuals by fitted values (left), QQ-plots of standardized residuals by values expected if the residuals were normally distributed (middle), and Cook's distances (right) for the models in which the independent variables are (a) quantity and speaker, (b) position and speaker, and (c) syllable type and speaker.

Table 22.1.1. Predictor estimates in a linear regression model of log-transformed Finnish vowel durations in which the log-transformed duration of *sanoin*, quantity, position, syllable type, and speaker are independent variables

Predictor	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	5.74149	0.40949	14.021	< 2e-16
log(sanoin)	−0.04253	0.06835	−0.622	0.5340
short	−0.69988	0.01405	−49.826	< 2e-16
non-final	−0.44621	0.01406	−31.735	< 2e-16
CVN	−0.03764	0.01987	−1.895	0.0585
GV	0.14927	0.01987	7.513	1.64e-13
V	0.22529	0.01990	11.322	< 2e-16
Speaker 2	−0.08865	0.02004	−4.424	1.11e-05
Speaker 3	0.25689	0.02016	12.744	< 2e-16
Speaker 4	−0.14900	0.02924	−5.096	4.37e-07

Table 22.1.2. Predictor estimates in a linear regression mode of Finnish vowel durations in which the log-transformed duration of *sanoin*, quantity, position, syllable type, speaker, and the interaction between position and syllable type are independent variables

Predictor	Estimate	Std. Error	*t* value	Pr(> |*t*|)
(Intercept)	5.62074	0.38072	14.763	< 2e-16
log(sanoin)	−0.01143	0.06352	−0.180	0.857309
short	−0.69974	0.01304	−53.674	< 2e-16
non-final	−0.57721	0.02608	−22.135	< 2e-16
CVN	−0.22225	0.02607	−8.526	< 2e-16
GV	0.13628	0.02608	5.225	2.26e-07
V	0.16261	0.02610	6.230	7.75e-10
Speaker 2	−0.08744	0.01860	−4.702	3.06e-06
Speaker 3	0.25532	0.01871	13.646	< 2e-16
Speaker 4	−0.13924	0.02715	−5.128	3.73e-07
non-final:CVN	0.36968	0.03688	10.024	< 2e-16
non-final:GV	0.02655	0.03687	0.720	0.471789
non-final:V	0.12648	0.03687	3.431	0.000635

the argument of the exp() function in R). (2) shows the non-zero terms in the model and the predicted duration of a short, non-final vowel in a CVN syllable:

(2) 95.4 ms = exp(5.74149 + −0.69988 + −0.44621 + −0.03764)

This is the predicted duration for Speaker 1; to predict the duration for such a vowel for Speaker 3, one would add 0.25689 to the exponent.

The interactions are represented in Table 22.1.2 as "non-final:CVN" etc., which indicates that they are the increment or decrement in duration predicted when the vowel's position is non-final and its syllable type is CVN etc., as compared to when its position is final and/or CV.

(3) shows the duration of a short, non-final vowel in a CVN syllable predicted by this model:

(3) 89.2 ms = exp(5.62074 + −0.69974 + 0.57721 + −0.22225 + 0.36968)

22.1.2.4 *Predicting novel values*

Models are also used to predict values of the dependent variable for new cases. For this model, one cannot of course use the model to predict what duration a vowel will have if it does not belong to any of the categories defined by the independent variables other than the duration of the word *sanoin*. One can nonetheless still estimate how precisely a new token's duration is predicted, as the square root of

the sum of the squares of the standard errors of the fitted values and of the residual standard errors. This value is 0.67576, which is equivalent to just under 2 ms. The model's predictions are so precise because the *n* is large.

22.1.2.5 *Validating and bootstrapping*

Two more informative means of testing the predictive accuracy of the model assess the extent to which the model overfits the data, i.e. how much adding terms to model to improve its fit to the current data reduces its ability to predict novel data. Both begin by dividing the data into training and testing sets.

"Cross-validation" divides the data into 3–10 equal-sized *folds*, where each fold is a random sample from the original data set. Each fold serves in turn as the test set, and the remaining data as the training set. The training data is used to calculate the model, which is then used to predict the observed test values. Table 22.1.3 shows the results of cross-validation with eight folds. The R^2 value is the proportion of variance accounted for by the model. The mean squared error is the mean of the squared residuals. The intercept and slope are for the line obtained when the observed durations are regressed against the fitted values. They are necessarily 0 and 1 for the original data and the training set, but the slope may be less than 1 for the test set. For slopes less than 1, the intercept's value compensates by shifting away from 0. The values for the training and test sets are the averages across the eight folds. The optimism values are the difference between the training and test-set statistics, and they estimate the extent to which the original model overfits the data. The corrected values are obtained by subtracting the optimism values from the original values—these discounted R^2 and MSE values would be used in a conservative assessment of how well the model fits the data. For this data set and model, the optimism values are all tiny, which indicates only very slight overfitting.

Table 22.1.3. R^2, mean squared error (MSE), intercept, and slope, showing original, training, and test values, the difference between training and test values (optimism), corrected values, and the number of folds (n) in cross-validation of the final model of the Finnish vowel durations

	original	training	test	optimism	corrected	*n*
R^2	0.86851	0.86879	0.85933	0.00946	0.85905	8
MSE	0.03207	0.03198	0.03336	−0.00138	0.03345	8
Intercept	0.00000	0.00000	0.01349	−0.01349	0.01349	8
Slope	1.00000	1.00000	0.99734	0.002656	0.99734	8

Table 22.1.4. R^2, MSE, intercept, and slope, showing original, training, and test values, optimism, corrected values, and the number of folds (n) in bootstrap validation of the final model of the Finnish vowel durations

	original	training	test	optimism	corrected	n
R^2	0.86851	0.87067	0.86639	0.004274	0.86424	200
MSE	0.03207	0.03141	0.03258	−0.00118	0.03324	200
Intercept	0.00000	0.00000	0.01564	−0.01564	0.01564	200
Slope	1.00000	1.00000	0.99692	0.00308	0.99692	200

In "bootstrapping," the training set is randomly and repeatedly drawn from the original data with replacement to produce a sample of the same size. This training set consists of roughly 485 unique values for an original data set of 768 values like the one here. The samples' values are used to calculate the model, and that model is used to predict the observed values in the original full data set. These steps are repeated many times, here 200. The results in Table 22.1.4 are interpreted in the same way as those in Table 22.1.3, and like those results they show that the model overfits the data very little.

The extent of overfitting is so small because the original data sample is quite large. When overfitting is greater, these procedures would justify omitting one or more of the independent variables or their interactions from the model.

22.1.2.6 *Fitting the data*

The R^2 value for the main-effects model presented in Table 22.1.1 was 0.847, while that which included the interaction in Table 22.1.2 was 0.869, a difference of just 0.022. This increment looks small, but Figure 22.1.3 motivates including this interaction in the final model. The quantitative modeling of the data was guided by the prior graphical exploration of the influence of the independent variables and possible interactions between them, rather than by a blind desire to improve the quantitative fit to the data. The resulting fit is quite good, more than 0.85 of the variance of the data is accounted for by a model with just thirteen predictors (one of them the intercept), which is the right balance.

Similarly, no automatic procedure like step-wise regression was used to decide which variables should be kept in the model. Besides taking model interpretation away from the analyst, such procedures ignore the effects of the variable selection process in calculating standard errors and t-statistics, they produce overoptimistic estimates of standard errors and p-values, and they bias the absolute values of predictor estimates upwards—positive estimates are farther from 0 and negative estimates closer.

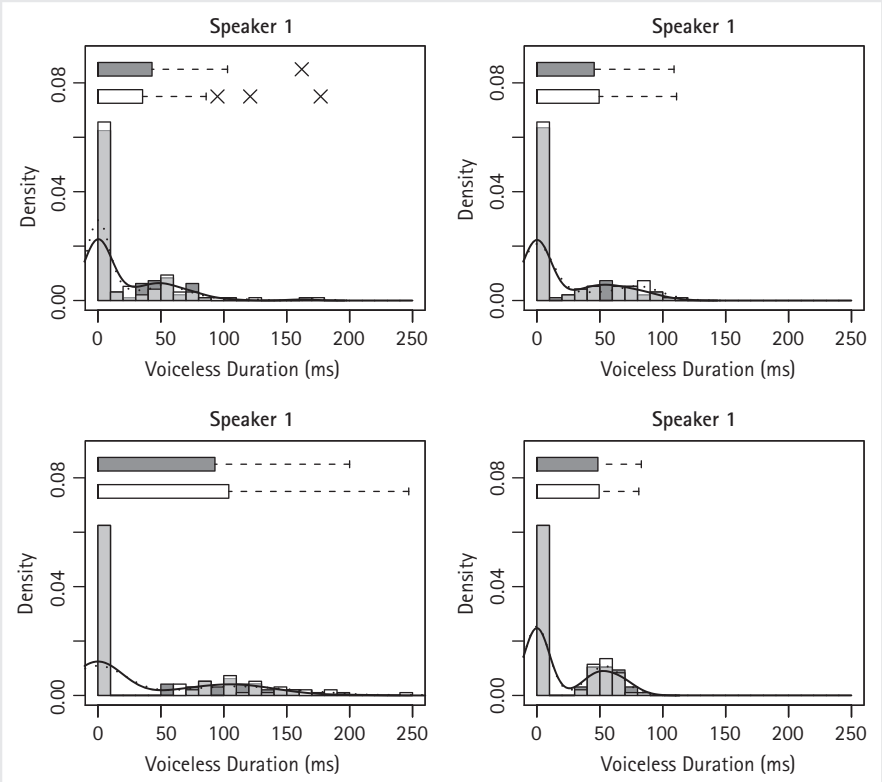

Figure 22.1.6. Histograms, density plots, and box plots of short (dark gray) and long (white) voiceless vowel durations produced by four Finnish speakers. The lighter gray bars represent durations common to short and long vowels. The Xs are outliers.

22.1.2.7 *Voiceless durations*

We now turn briefly to the durations of the voiceless portions of these vowels (Figure 22.1.6). They are of interest because Myers and Hansen propose that final vowels shorten as a result of being partially devoiced (Myers and Hansen 2006, 2007; Baayen 2008). Figure 22.1.6 looks very different from Figure 22.1.1: the short and long distributions are no longer even approximately discrete but instead overlap completely, and now there is a very strong mode at 0, which represents all those vowels which have no voiceless portion. Figure 22.1.7 reveals that no portion of any vowel is voiceless in non-final position or closed syllables. Position and CVN syllables can therefore be left out of modeling the voiceless portions' durations.

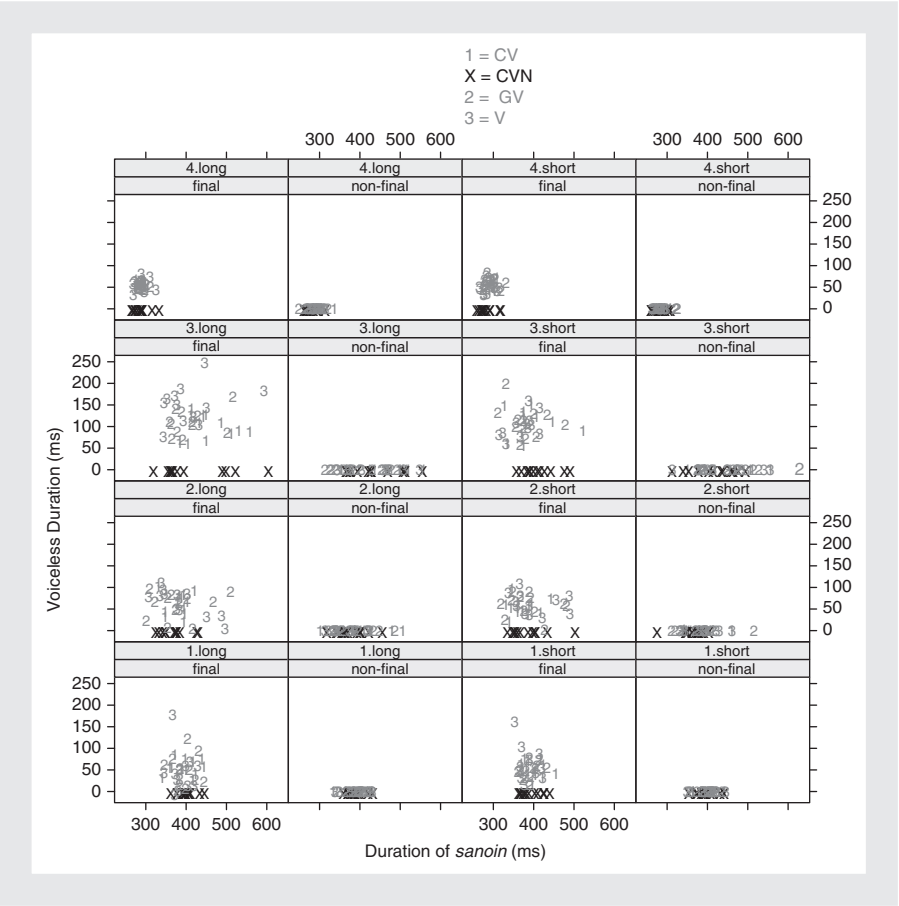

Figure 22.1.7. Durations of voiceless portions of vowels (vertical axis) by duration of the word *sanoin* duration in the same utterance by speaker, phonological quantity, position, and syllable type. Black "X" for closed CVN syllables, and gray "1, 2, 3" for open CV, GV, and V syllables, respectively.

The predictor estimates for the model of the durations of the voiceless portions of the vowels in Table 22.1.5 show that none of the variables have a significant effect on these durations, except that they are significantly longer for speaker 3 than speaker 1. In Table 22.1.6, the log-transformed duration of the voiced portion of the vowel has been added to the model as a predictor. The predictor estimates show that the duration of the voiceless portion shortens significantly as the duration of the voiced portion lengthens ($t = -2.6858$, $p = 0.00767$). A more interesting finding is that the duration of the voiceless portion is now also significantly shorter when the vowel is phonologically short ($t = -2.4136$, $p = 0.01644$). In other words, the voiceless portion's duration varies inversely with the voiced portion's duration but directly with the vowel's phonological quantity.

Table 22.1.5. Predictor estimates in a linear regression model of the durations of the voiceless portions of Finnish vowels in final position in which the log-transformed duration of *sanoin*, quantity, syllable type, and speaker are independent variables

Predictor	Estimate	Std. Error	*t* value	Pr(> |*t*|)
Intercept	6.273839	1.92506	3.2590	0.001256
log(sanoin)	−0.404738	0.32082	−1.2616	0.208150
short	0.008546	0.06327	0.1351	0.892658
GV	−0.084840	0.07708	−1.1007	0.271997
V	−0.028269	0.07742	−0.3651	0.715298
Speaker 2	0.115028	0.08965	1.2831	0.200533
Speaker 3	0.871518	0.08930	9.7598	0.000000
Speaker 4	0.058198	0.13202	0.4408	0.659667

Table 22.1.6. Predictor estimates in a linear regression model of the durations of the voiceless portions of Finnish vowels in final position in which the log-transformed duration of *sanoin*, log-transformed duration of the voiced portion of the vowel, quantity, syllable type, and speaker are independent variables

Predictor	Estimate	Std. Error	*t* value	Pr(> |*t*|)
Intercept	8.64921	2.09935	4.1199	0.00005
log(sanoin)	−0.44829	0.31772	−1.4110	0.15940
log(voiced duration)	−0.41354	0.15397	−2.6858	0.00767
short	−0.38244	0.15845	−2.4136	0.01644
GV	0.01427	0.08470	0.1685	0.86630
V	0.07546	0.08576	0.8799	0.37970
Speaker 2	0.11330	0.08867	1.2778	0.20240
Speaker 3	1.00419	0.10119	9.9234	0.00000
Speaker 4	0.02268	0.13124	0.1728	0.86290

22.1.2.8 *Collinearity and principal components*

There is a problem here, however: the voiced portion's duration and its phonological quantity are probably not independent of one another. The voiced portion is expected to be longer when the vowel is phonologically long. An analysis not shown here confirms this expectation. The covariation of the the voiced portion's duration and the vowel's phonological quantity is an example of collinearity. The extent of collinearity here is 8.89, which is greater than negligible (0–6), but less than moderate (around 15) and much less than severe (values greater than 30). Although collinearity is slight enough here that no remedy is required, this example can nonetheless be used to show how to proceed when collinearity is more

severe. One could simply leave out one of the collinear variables, but which one? A more principled approach uses principal components analysis to eliminate the collinearity.

Principal components analysis uses the covariation (collinearity) between the values of individual variables to construct a new composite variable that combines their influences on the data's structure. The first principal component accounts for the largest proportion of the variance in the data's structure, the second for the next largest proportion, and so on. One ends up with as many principal components as the original number of variables, but they are now rank-ordered in terms of how much of the variance they account for. A common rule of thumb is to disregard any principal components that account for less than 0.05 of the variance.

The principal components replace the collinear variables in constructing a new model of the data. Table 22.1.7 lists the statistics of the two principal components extracted when this method was applied to the quantity and voiced durations for the vowels whose voiceless durations were modeled above. The first principal component accounts for just over 0.93 of the variance, and the second for just under 0.07. Table 22.1.8 shows the "loadings" of these two principal components on the original variables, quantity and voiced duration. The signs of these loadings are positive for the loading of PC1 on both quantity and voiced duration, which captures the fact that these two variables co-vary directly: a long vowel has a longer voiced duration. The opposite signs of the loadings for PC2 capture the weaker inverse variation between voiced duration and phonological quantity, namely, that the voiced duration is shorter in long vowels.

Table 22.1.7. Standard deviations and proportions of variance accounted for by the first two principal components extracted from the covariation between phonological quantity and voiced duration for final vowels in CV, GV, and V syllables

	PC1	PC2
Standard deviation	1.365	0.3702
Proportion of variance	0.931	0.0685

Table 22.1.8. Loadings of the first two principal components on phonological quantity and voiced duration

	PC1	PC2
quantity	0.7071068	−0.7071068
log-transformed voiced duration	0.7071068	0.7071068

Table 22.1.9. Predictor estimates in a linear regression model of the durations of the voiceless portions of Finnish vowels in final position in which the log-transformed duration of *sanoin*, two principal components representing the log-transformed duration of the voiced portion of the vowel and quantity, syllable type, and speaker are independent variables

Predictor	Estimate	Std. Error	*t* value	Pr(> |*t*|)
Intercept	6.44336	1.90150	3.3886	0.00080
log(sanoin)	−0.44829	0.31772	−1.4110	0.15937
PC1	−0.02465	0.02358	−1.0454	0.29674
PC2	−0.29555	0.11335	−2.6073	0.00962
GV	0.01427	0.08470	0.1685	0.86634
V	0.07546	0.08576	0.8799	0.37969
Speaker 2	0.11330	0.08867	1.2778	0.20240
Speaker 3	1.00419	0.10119	9.9234	0.00000
Speaker 4	0.02268	0.13124	0.1728	0.86292

Table 22.1.9 presents the linear model of the voiceless durations with PC1 and PC2. The estimate for PC1 is not significant ($t = -1.0454$, $p = 0.29674$), while that for PC2 is ($t = -2.6073$, $p = 0.00962$). This outcome is not surprising in light of Table 22.1.6, which showed that the voiceless duration was shorter when the vowel was phonologically short and when its voiced duration was longer. This outcome also indicates that we could leave PC1 out of the final model of the voiceless durations. That model accounts for the same proportion of the variance in the voiceless durations as that using quantity and voiced durations as independent variables, 0.294, and does so with one less explanatory predictor.

Besides the slightness of the collinearity in this example, there is another, more general reason to hesitate to apply this method: by collapsing the influences of two or more of the original variables into a single variable, principal components analysis can obscure rather than illuminate the analysis. In some instances, those variables may simply be alternative ways of measuring the same psychologically real linguistic property; then, principal components analysis reveals that underlying reality. But in this instance, the original analysis with quantity and voiced duration as distinct variables provides a more straightforward description and explanation of what influences the duration of the vowel's voiceless portion.

22.1.2.9 *Summary*

This section has presented linear models of continuous dependent variables, the total durations of Finnish vowels and of their voiceless portions. Graphical explorations preceded and guided model construction. The models were then criticized,

by examining their residuals, through cross-validation and bootstrapping, and assessing the extent to which the independent variables were collinear. These critiques led to the log transformation of the dependent variable to bring the residuals into line.

22.1.3 Mixed-effects models of ordinal dependent variables: First try

22.1.3.1 *Random versus fixed effects*

Myers and Hansen presumed that they had drawn a random yet representative sample from the population of Finnish speakers. Presuming that the sample is representative does not entail that one member's data values will not differ from other potential members, but only that the differences will be idiosyncratic rather than systematic. The expected idiosyncrasies were built into the models by representing speakers 2–4 with their own predictor, whose value showed how their vowel durations differed overall from default speaker 1. Interactions were excluded between the speaker predictors and any of the others because I tacitly assumed that the effects of phonological quantity, position, and syllable type would not differ substantially between speakers. That assumption represents a fundamental difference between kinds of effects, random versus fixed effects.

Speaker is a random effect in that the speakers are a random sample from the population of possible speakers. We would not expect to *repeat* the idiosyncrasies of one sample of speakers in another. Fixed effects (also referred to as "conditions" or "treatments") such as phonological quantity etc. are repeatable, in that they can be applied to another sample (see also Baayen, this chapter, for further discussion of how random effects differ from fixed effects).

Models which combine random and fixed effects are called "mixed-effects" or simply "mixed" models. Mixed models of ratings are presented here. Ratings are an example of an ordered or "ordinal" categorical variable. The R packages, lme4 and ordinal, used in carrying out these analyses are described in Bates and Maechler (2010) and Christensen (2010). Before beginning these analyses, contrast coding of categorical variables with more than two values must be discussed (see also Baayen's section in this chapter).

22.1.3.2 *Categorical predictors with more than two values: Contrasts*

A common practice when a categorical predictor has more than two values has been to determine its significance overall by means of an analysis of variance, and then to run post-hoc tests comparing pairs of predictor values. Because there is a substantial danger of getting a spuriously significant result when one runs multiple

Table 22.1.10. Recoding of a four-valued predictor (A–D) into three contrasts, using treatment, Helmert, or polynomial recoding

Predictor	Treatment			Helmert			Polynomial		
	T1	T2	T3	H1	H2	H3	Linear	Quadratic	Cubic
A	0	0	0	−1	−1	−1	−3	1	−1
B	1	0	0	1	−1	−1	−1	−1	3
C	0	1	0	0	2	−1	1	−1	−3
D	0	0	1	0	0	3	3	1	1

tests, the α value must be corrected to α/m, where m is the number of post-hoc comparisons—this is the "Bonferroni" correction. Applying this correction can make it difficult to achieve significance if the number of comparisons is large.

A better solution is to recode the original predictor with a set of contrasts that embody the comparisons one wants to do. All ways of recoding categorical predictors require the contrasts be orthogonal; that is, for k predictor values (AKA "treatments"), there are only $k - 1$ contrasts. Otherwise, treatments excluded from a contrast are assigned a value of 0, treatments that are grouped together are assigned the same sign, and those which are contrasted are assigned opposite signs. In some kinds of contrasts, the values assigned to included treatments also sum to 0. Table 22.1.10 illustrates the recoding of a four-valued predictor (A-D) for treatment, Helmert, and polynomial contrasts.

Treatment recoding is identical to simply comparing each non-default treatment to the default treatment (as was done for syllable type in the Finnish vowel duration analysis). The first contrast in Helmert recoding compares the second treatment (B) with the first (A) and excludes the other treatments (C and D), the second compares the third treatment (C) with the mean of the first and second (A, B) and excludes the fourth (D), and the third compares the fourth treatment (D) with the mean of the first, second, and third (A, B, C). By judiciously ordering the treatments, one can obtain the comparisons one wants with Helmert recoding. Polynomial recoding is only appropriate when the original predictor's values are ordered but cannot be assigned a value along a scale. The contrasts model the predictor's effect as linear, quadratic, cubic functions, etc., where the highest order of the polynomial equals one less than the number of contrasts.

22.1.3.3 *Experiment design and a first look at the results*

The data are dissimilarity ratings obtained in an ERP study carried out by Mara Breen, Lisa Sanders, and me. The participants were presented with two syllables on each trial, the first was the "prime" and the second the "target," and their task was to rate how dissimilar the target was to the prime on a four-point scale, where

Table 22.1.11. Voiced and unvoiced prime-target pairs for Legal, Illegal, and Absent primes in Identity, Control, and Test trials

Prime status	Legal		Illegal		Absent	
Trial type	Prime	Target	Prime	Target	Prime	Target
Identity	gw	gw	gl	gl	gw	gw
Control	kw	gw	kl	gl	tw	gw
Test	dw	gw	dl	gl	bw	gw
Identity	kw	kw	kl	kl	tw	tw
Control	gw	kw	gl	kl	gw	tw
Test	tw	kw	tl	kl	pw	tw

1 corresponded to maximally similar, 4 to maximally dissimilar and 2 and 3 to lesser values of similarity and dissimilarity. Ratings are an example of an ordinal dependent variable. The values of ordinal variables are ordered like those of continuous variables such as vowel duration, but they are also categorical; that is, values such as 1.5, $\sqrt{2}$, 3.14159, etc. are not possible.

Table 22.1.11 shows that primes in test trials consisted of syllables beginning with consonant clusters that are legal in English, e.g. [dw, tw], absent but perhaps not illegal, e.g. [bw, pw], or illegal [dl, tl] (see Moreton 2002, for justification of this classification). The consonant clusters in the targets were always legal, as were all clusters in primes in identity and control trials. Multiple tokens of each syllable were used to compose 100 distinct trials for each of the eighteen possible prime-target combinations. Responses were collected from eighteen native speakers of English.

Figures 22.1.8 and 22.1.9 display the mean dissimilarity ratings across participants and items for the three kinds of primes and the three trial types, collapsed across voicing. Figure 22.1.8 averages across items, while Figure 22.1.9 averages across participants. Both figures show that responses cluster near 1 on identity trials (top rows) and near 4 on Control and Test trials (middle and bottom rows), except when the prime is Illegal (bottom right), where ratings instead cluster near 2. The noticeably greater spread of values in Figure 22.1.8 than Figure 22.1.9 shows that ratings differed more between participants than items—there are also few if any ratings of 2–3 for items.

22.1.3.4 *Mixed-effects model with Helmert contrasts*

Averaging across participants (Figure 22.1.8) or items (Figure 22.1.9) is the first step in what has until recently been standard practice in psycholinguistics, namely, carrying out a by-participants (= by-subjects) analysis, in which participants are treated as random effects, and then a by-items analysis, in which items are treated as a random effect (Clark 1973; Forster and Dickinson 1976). In this approach, a predictor's effect is treated as significant only if it is significant in both analyses. The

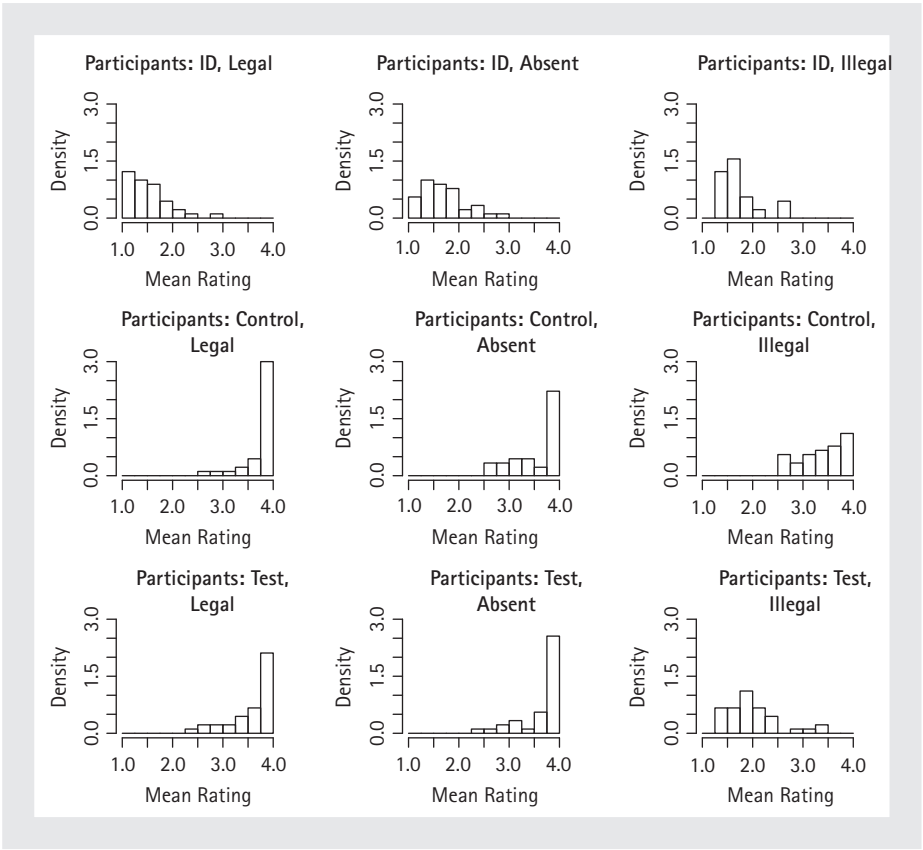

Figure 22.1.8. Mean dissimilarity ratings by participants and by the status of the prime and type of the trial.

development of mixed-effects models has largely superseded the need to carry out separate analyses, because both participants and items can be treated as random effects within a single analysis. However, software to implement such models has not yet been developed for cases like this one where the dependent variable is an ordinal variable, so I carry out separate participant and items analyses here.

Averaging across items or participants replaced the categorical integer values of the original ratings with non-integer values. To turn these means back into integers, they were first multiplied by 4, to preserve to some extent the distinctions between the means, and then rounded to the nearest integer. The combined effect of these two operations is to turn the continous 1–4 scale produced by the averaging into a 4–16 integer scale.

Because the dependent variable is a frequency rather than the measure of some quantity, it is transformed further. The original frequencies or more precisely the ratios of ratings of 4 versus ratings greater than 4, of ratings of 4–5 versus ratings

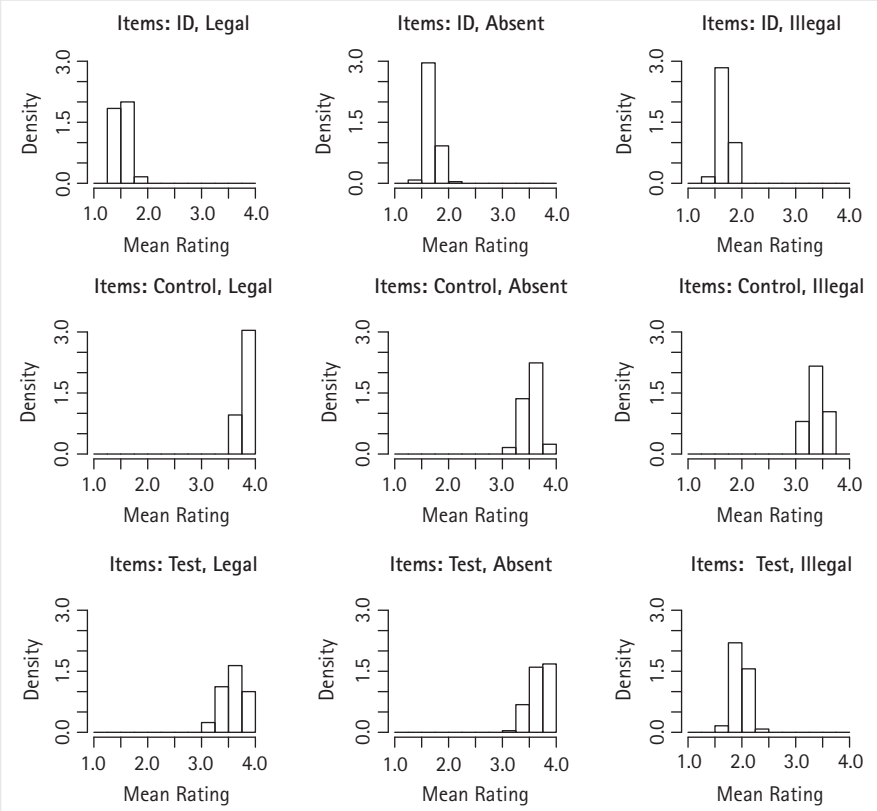

Figure 22.1.9. Mean dissimilarity ratings by items and by the status of the prime and type of the trial.

greater than 5, . . . , to ratings of 4–15 versus ratings of 16 are transformed into log odds ratios, aka logits (see also Baayen, this chapter). The model for frequency data is therefore a logistic rather than linear regression.

In both analyses, voicing, prime status, and trial type served as independent variables, with prime status and trial type recoded as Helmert contrasts (Table 22.1.12). The signs of the Helmert contrasts for status are opposite to those given in Table 22.1.10 because we expect lower dissimilarity ratings for identity than control or test trials. The models also included all pair-wise interactions between prime status and trial type.

The results of the two analyses are displayed in Tables 22.1.13 and 22.1.14.[2] The variance and standard deviation are much larger for participants, 6.1251 and 2.4749,

[2] The estimates in these tables are logits and can be converted back into odds ratios by using the products of their values with the values of the corresponding Helmert contrasts as exponents of e.

Table 22.1.12. Recoding of the three-valued status and type predictors as Helmert contrasts

Status	S1	S2	Type	T1	T2
Legal	1	1	Identity	−1	−1
Absent	−1	1	Control	1	−1
Illegal	0	−2	Test	0	2

Table 22.1.13. Predictor estimates in an ordinal logistic regression by-participants model of dissimilarity ratings in which S1, S2, T1, and T2 are Helmert contrasts representing prime status and trial type

Predictor	Estimate	Std. Error	*z* value	Pr(> |*z*|)
S1	−0.0264	0.1444	−0.1826	0.8550765
S2	1.3181	0.1176	11.2042	< 2.22e-16
T1	6.1420	0.4453	13.7923	< 2.22e-16
T2	1.0297	0.1022	10.0732	< 2.22e-16
Voiced	−0.5804	0.2192	−2.6473	0.0081131
S1:T1	1.1080	0.1862	5.9515	2.6570e-09
S1:T2	−0.2933	0.1037	−2.8284	0.0046778
S2:T1	0.5924	0.0967	6.1233	9.1631e-10
S2:T2	1.0392	0.0958	10.8460	< 2.22e-16

Table 22.1.14. Predictor estimates in an ordinal logistic regression by-items model of dissimilarity ratings in which S1, S2, T1, and T2 are Helmert contrasts representing prime status and trial type

Predictor	Estimate	Std. Error	*z* value	Pr(> |*z*|)
S1	−0.0988	0.0949	−1.0413	0.2977301
S2	11.0210	0.5798	19.0083	< 2.22e-16
T1	50.8508	2.6350	19.2979	< 2.22e-16
T2	6.3377	0.3154	20.0957	< 2.22e-16
Voiced	−1.1321	0.1464	−7.7338	1.0439e-14
S1:T1	1.7209	0.1297	13.2682	< 2.22e-16
S1:T2	−0.2041	0.0663	−3.0788	0.0020783
S2:T1	0.9032	0.0699	12.9237	< 2.22e-16
S2:T2	10.6456	0.5739	18.5485	< 2.22e-16

Table 22.1.15. Values of interaction terms for each combination of prime status and trial type contrasts

Interaction	Leg:ID	Leg:Cntl	Leg:Test	Abs:ID	Abs:Cntl	Abs:Test	Ill:ID	Ill:Cntl	Ill:Test
S1:T1	−1	1	0	1	−1	0	0	0	0
S1:T2	−1	−1	2	1	1	−2	0	0	0
S2:T1	−1	1	0	−1	1	0	2	−2	0
S2:T2	−1	−1	2	−1	−1	2	2	2	−4

than for items, 0.3082 and 0.5552. These outcomes confirm what Figures 22.1.8 and 22.1.9 had already shown: ratings differ more between participants than items.

The results of the two analyses are otherwise strikingly similar: all the predictors significantly influence the dissimilarity ratings, except for S1, the contrast that represents the comparison between legal and absent primes; ratings are predicted to be higher for legal and absent than illegal test primes (S2), on control than identity trials (T1), and on test than control and identity trials (T2); and voiced pairs were judged to be less dissimilar than voiceless ones.

The interactions are also strikingly similar in the two analyses. Table 22.1.15 lists the values which when multiplied by the estimates in Table 22.1.13 or 22.1.14 yield the predicted effects on the odds ratios of the dissimilarity ratings for each combination of prime status and trial type. For example, multiplying postive S2:T2 estimates by the value −4 for the illegal test combination predicts a dramatic drop in the dissimilarity ratings, which can be observed in Figure 22.1.8 and even more dramatically in Figure 22.1.9.

22.1.3.5 *Summary*

In this section, two mixed-effects ordinal logistic regression models of dissimilarity ratings were presented in which two three-valued predictors were recoded as Helmert contrasts to rule out the need for post-hoc tests comparing subsets of predictor values. Each model included a single random effect, either of participants or items, and the dependent variable was the log odds ratios of the dissimilarity ratings for each interval along the ordinal scale averaged over the other random effect.

22.1.4 Mixed-effects models of ordinal dependent variables: Second try

Here, I reanalyze these data by treating the dissimilarity ratings as a series of three binomially distributed variables, 1 versus 2–4, 1–2 versus 3–4, and 1–3 versus 4, in

Table 22.1.16. Random effects for logistic regression models of the partition into a series of two-valued variables: 1 versus 2–4, 1–2 versus 3–4, and 1–3 versus 4

	1:234		12:34		123:4	
Random effects	Variance	Std. Dev.	Variance	Std. Dev.	Variance	Std. Dev.
items	2.622	1.619	1.517	1.232	1.120	1.058
participants	1.848	1.359	1.556	1.247	2.461	1.569

order to include both random effects in the model at once.[3] Three mixed-effects models are constructed, one for each these new variables. The random effects in these models are participants and items, while the fixed effects are the same as those just used, coded once again as Helmert contrasts.

Table 22.1.16 lists the random effects on the intercepts of these three models. The standard deviations differ relatively little between the models, which shows that the range of differences between items and participants is roughly the same for all three. These values are no longer noticeably smaller for items than participants.

Table 22.1.17 lists the estimates of the fixed effects with their standard errors and z-scores; a * follows when the associated p value is less than 0.05. The model predicts that the probability of the lower dissimilarity rating(s) in each partition increases for illegal compared to legal and absent primes (S2 = −2 versus 1), this probability decreases in control compared to identity trials (T1 = 1 versus −1) and in test compared to control and identity trials (T2 = 2 versus −1), and it decreases in test trials compared to identity and control trials for legal and absent primes (S2 : T2 = 2 versus −1), while increasing for test trials compared to identity and control trials for illegal primes (S2 : T2 = −4 versus 2). This analysis is noticeably more conservative than the earlier ones in that only the S2:T2 interaction is significant. Averaging across items or participants hid variation in the random effects that remains exposed in this analysis. This reanalysis thus shows that the traditional approach's insistence on effects being significant in both by-participants (by-subjects) and by-items analyses is not sufficient protection against rejecting the null hypothesis when there is a good chance after all that it's true (Type 1 error).

Ordinal logistic regression was illustrated in this section by treating the dissimilarity ratings as a three-step series of binary and thus binomially distributed dependent variables, and including both participants and items as random

[3] The ordinal logistical regression models just illustrated partition the data similarly.

Table 22.1.17. Fixed effects for a series of logistic regression models in which a four-valued ordinal variable is partitioned into a series of two-valued variables: 1 versus 234, 12 versus 34, and 123 versus 4. $* = p < 0.05$

	1:234				12:34				123:4			
Predictor	Estimate	Std. Error	z		Estimate	Std. Error	z		Estimate	Std. Error	z	
(Intercept)	−1.984	0.501	−3.963	*	−0.613	0.415	−1.477		0.322	0.447	0.719	
S1	−0.031	0.333	−0.095		−0.015	0.253	−0.058		−0.055	0.218	−0.253	
S2	−0.447	0.192	−2.329	*	−0.558	0.146	−3.818	*	−0.647	0.0126	−5.138	*
T1	−2.773	0.333	−8.326	*	−2.706	0.254	−10.656	*	−2.667	0.219	−12.164	*
T2	−0.491	0.192	−2.559	*	−0.498	0.146	−3.408	*	−0.477	0.126	−3.793	*
Voiced	0.175	0.543	0.323		0.264	0.413	0.638		0.399	0.0356	1.121	
S1:T1	−0.414	0.408	−1.014		−0.395	0.311	−1.272		−0.413	0.268	−1.545	
S1:T2	0.099	0.235	0.420		0.112	0.179	0.627		0.136	0.154	0.881	
S2:T1	−0.218	0.235	−0.927		−0.204	0.179	−1.140		−0.254	0.154	−1.646	
S2:T2	−0.375	0.136	−2.768	*	−0.421	0.103	−4.078	*	−0.459	0.089	−5.165	*

effects. With the random effects unconstrained, fewer interactions turned out significant.

22.1.5 Concluding remarks

In this chapter, I have tried to illustrate how statistical tools can be used to explore and test hypotheses concerning two kinds of data commonly encountered while studying phonology in the laboratory: continuous variables and ordinal scales. My goal throughout has been to focus on the practicalities of carrying out such explorations and tests. My hope is that these illustrations provide sufficient guidance that you may see how to adapt them to the data you are trying to analyze. To understand the practicalities of these analyses, you will also need to study the sources cited at the beginning of this chapter. The payoff for doing so is enormous, as the statistical tools used here provide many insights into the data.

22.2 Mixed-effects models

Harald Baayen

22.2.1 Introduction

Consider an experiment in which the duration of the first vowel in a word is studied. It is expected that this duration is determined in part by the number of syllables following in the same word, in part by whether the vowel is in an open syllable (vs. closed syllable), in part by the position of the word in the sentence, by the speech rate, and possibly by the frequency of the word. If our interest is in the generality of vowel shortening, different vowels will be studied, in different words, and produced by different speakers. For this type of experiment, mixed models are an excellent choice.

In this example, the factor `Syllable Type` (with levels *open syllable* and *closed syllable*) is a fixed-effect factor, as its two levels exhaust all possible values that the predictor `Syllable Type` can take. By contrast, the factor `Speaker` is a random-effect factor, as its levels, identifiers for the different speakers, are randomly sampled from a much larger population of speakers. `Word` is another random-effect factor, as the words sampled for the experiment represent only a small proportion of the words known to the speakers (see also Section 22.2.6 for how to define fixed-effect vs. random-effect factors).

Classical analysis of variance and regression analysis run into problems for data sets combining fixed- and random-effect factors, especially when more than one random-effect factor has to be brought into the analysis. Often, researchers aggregate their data to obtain means or proportions for subjects (averaging over items) or for items (averaging over subjects, see also Kingston, this chapter). In psycholinguistics, the work by Clark (1973) and Forster and Dickinson (1976) led to the practice of averaging both over subjects and over items, with an effects accepted as significant only if it reaches significance both 'by subjects' and 'by items'. Mixed-effect models provide the researcher with a more sophisticated tool for analyzing repeated measures data that is both more flexible, more powerful, and more insightful.

22.2.2 Basic concepts

Let X_1 denote the fixed-effect factor `Syllable Type` and let X_2 represent the covariate `Frequency` of occurrence. Suppose that ten vowels are selected, and that the question of interest is whether the duration of the k-th vowel, Y_k, can be predicted from `Syllable Type` (*open* versus *closed syllable*) and `Frequency`. The linear model decomposes the dependent variable into a weighted sum:

$$(1) \quad Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_{12} X_{1k} X_{2k} + \epsilon_k, \quad k = 1, 2, \ldots, 10.$$

Fixed-effect factors are coded numerically using dummy coding, such that a factor with n levels contributes $n - 1$ predictors to the model. Of the many ways in which factors can be coded numerically, *treatment coding* is the most straightforward and the most easy to interpret, especially in the case of analysis of covariance. One level of the factor is selected as default or reference level. Although the selection of the reference level can be guided by theoretical considerations, technically, any level can serve as reference level. For the two-level factor `Syllable Type`, treatment coding adds one extra predictor, X_1 in (1), consisting of ones and zeroes. Observations for the reference level, say *closed syllable*, are assigned a zero, and observations for the other, contrasting level (*open syllable*) are assigned a one. As a consequence, the β weight for `Syllable Type` represents the *difference* (or contrast) between the group mean for the vowels in an open syllable and the group mean for the vowels in a closed syllable. This β weight, although technically a slope for a "degenerate" numerical predictor (consisting only of zeroes and ones), is referred to as a contrast coefficient.

The model defined in (1) includes an interaction term for `Syllable Type` by `Frequency`. This interaction allows for the possibility that two different regression lines are required for `Frequency`, one for vowels in closed syllables and a different one for vowels in open syllables. As a consequence, two intercepts and two slopes have to be defined. With treatment coding, the regression line for the reference level (*closed syllable*) is specified by the intercept β_0 and the slope for frequency β_2. The

coefficients of the regression line for *open syllables* is obtained by *adjusting* these slopes and intercepts (by β_1 and β_{12}) respectively (see Table 22.2.1) to make them precise for the data points with the vowels in open syllables. In summary, for a fixed-effect factor, one level is selected as the baseline, and coefficients are invested to adjust slopes and intercepts for the other levels of the factor.

When dealing with a random-effect factor, it does not make sense to select one—arbitrary—level (e.g. a given speaker, or a specific word) as reference level: Such a reference level is unlikely to be representative of the population sampled. Therefore, mixed models dispense with fixing a reference level and contrasts for random-effect factors. Instead, the β coefficients for the intercept, covariates, and fixed-effect factors are taken to represent the population average for each of the populations sampled by the random-effect factors. For any given random-effect factor, adjustments are implemented to allow precise predictions for the individual units sampled, such as the individual speakers in an experiment or corpus. These adjustments (technically referred to as Best Linear Unbiased Predictors or BLUPS) are assumed to follow a normal distribution with mean zero and some unknown standard deviation (to be estimated from the data). Instead of investing $n - 1$ coefficients for a simple main effect for a random-effect factor with n levels (e.g. n speakers), only one parameter is invested, a standard deviation characterizing the spread of the adjustments.

By way of example, consider a data set in which vowels are elicited in m words from n speakers, and that a simple main-effects model is appropriate. A first model,

$$(2)\quad \begin{aligned} &Y_{ij} = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]X_{1j} + [\beta_2 + b_{2i}]X_{2j} + \epsilon_{ij}, \\ &i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m, \\ &b_{0i} \sim \mathcal{N}(0, \sigma_1),\ b_{1i} \sim \mathcal{N}(0, \sigma_2),\ b_{2i} \sim \mathcal{N}(0, \sigma_3),\ \epsilon_{ij} \sim \mathcal{N}(0, \sigma), \end{aligned}$$

calibrates the model, for each speaker i, for that speaker's speech rate (through the adjustments b_{0i} to the intercept β_0), as well as for that speaker's sensitivity to the type of syllables (through the adjustments b_{1i} to the contrast coefficient β_1) and for that speaker's specific sensitivity to frequency of occurrence (through the

Table 22.2.1. Treatment coding in analysis of covariance: the contrast coefficients β_1 and β_{12} specify the differences in intercept and slope between the vowels in *open* and *closed* syllables

β_0	the intercept (group mean) for the reference-level *closed syllable*
$\beta_0 + \beta_1$:	the intercept (group mean) for *open syllables*
β_2	the slope for frequency for vowels *in closed syllables*
$\beta_2 + \beta_{12}$:	the slope for frequency for vowels *in open syllables*

adjustments b_{2i} to the slope β_2). Each of the sets of adjustments $b_{.i}$ is assumed to be normally distributed with zero mean. In other words, a random-effect factor (whether speaker, word, text, or syllable) is represented as a source of random variation around the population parameters $\{\beta\}$. This is the sense in which a random-effect factor is "random."

Model (2) is incomplete, in that it does not take into account that the words in which the vowels are embedded are repeated across speakers. To incorporate word as a second random-effect factor, (2) has to be modified as follows,

$$
\begin{aligned}
(3)\quad & Y_{ij} = [\beta_0 + b_{0i} + b_{0j}] + [\beta_1 + b_{1i} + b_{1j}]X_{1j} + [\beta_2 + b_{2i}]X_{2j} + \epsilon_{ij}, \\
& i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m; \\
& b_{0i} \sim \mathcal{N}(0, \sigma_1), \ b_{1i} \sim \mathcal{N}(0, \sigma_2), \ b_{2i} \sim \mathcal{N}(0, \sigma_3), \\
& b_{0j} \sim \mathcal{N}(0, \sigma_4), \ b_{1j} \sim \mathcal{N}(0, \sigma_5), \ \epsilon \sim \mathcal{N}(0, \sigma),
\end{aligned}
$$

with crossed random effects for speaker and word. Adjustments to the intercept are often referred to as random intercepts. Similarly, adjustments to slopes are known as random slopes. In the case of adjustments to a contrast coefficient, one can speak of random contrasts. In (3), there are by-speaker random intercepts (b_{0i}) as well as by-word random intercepts (b_{0j}). Likewise, there are both by-speaker and by-word random contrasts (b_{1i}, b_{1j}). The model includes random slopes for frequency only for speaker (b_{2i}). It is not possible to include as well by-word random slopes for frequency, as this would lead to an unsolvable confound with frequency itself, which is a word property. In other words, it is only possible to include by-subject random slopes and contrasts for item properties, and by-item random slopes and contrasts for subject properties. For instance, speakers may require adjustments to the slope of the frequency effect, while words may require adjustments to the slope of the effect of aging (see e.g. Baayen and Milin 2010).

Whenever in addition to random intercepts, one or more random slopes (or contrasts) are associated with a given random-effect factor, the possibility arises that the random intercepts and random slopes (or contrasts) are correlated. Assuming multivariate normality, the full specification of the random effects for (3) is therefore given by the matrices

$$
(4)\quad M_{\text{speaker}} = \begin{bmatrix} \sigma_1 & r_{12} & r_{13} \\ r_{21} & \sigma_2 & r_{23} \\ r_{31} & r_{32} & \sigma_3 \end{bmatrix}, \qquad M_{\text{word}} = \begin{bmatrix} \sigma_4 & r_{45} \\ r_{54} & \sigma_5 \end{bmatrix},
$$

where $r_{kl} = r_{lk}$ specifies the correlation of the adjustments k and l estimated for the population of speakers or the population of words. In other words, the adjustments for a given random-effect factor are assumed to be multivariate normal with zero means and unknown standard deviations and correlations.

22.2.3 Advantages of mixed-effects models

Mixed-effects models offer many advantages compared to the classical linear model using dummy coding for random-effect factors (see also Kingston, this chapter). First, a fitted mixed model provides straightforward predictions for unseen levels of random-effect factors. For an unseen speaker and an unseen word, all $b_{\cdot\cdot}$ are set to zero, and predictions based on model (3) for a given position X_1 and frequency X_2 reduce to

(5) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$

For a specific speaker i that contributed observations to the data and an unseen word, more precise predictions can be obtained using the by-subject random-effect adjustments:

(6) $Y_i = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}] X_1 + [\beta_2 + b_{2i}] X_2.$

Similarly, when the identity of the word is known, even more precise predictions are available by adding in the by-word random intercepts and slopes. For comparison: the classical linear model only provides predictions for the subjects and items sampled in the data, and models with many interactions involving subjects and items may not even be able to estimate all relevant coefficients.

Second, the mixed-effects model allows for fine-grained hypotheses about the random-effects structure of the data. For every data set, it is an empirical question whether all the terms in matrices such as shown in (4) contribute to a significantly better fit of the model to the data. The possibility of including or excluding correlation parameters is not available in the classical linear model, but turns out to be an important tool for understanding, for instance, individual differences between the subjects participating in experiments. In chronometric studies, for instance, one may find that subjects with a large positive adjustment to the intercept reveal a large negative adjustment to the slope of frequency of occurrence. Such a negative correlation suggests that slow responders (with large intercepts) carry the frequency effect (see e.g. Baayen and Milin 2010 for examples).

Third, mixed-effects models are better able to directly model heteroskedasticity. A fundamental assumption of the linear model is that the residual errors have the same variance across all conditions in the data. In many actual data sets, this assumption of homoskedasticity is violated. For instance, the duration of a vowel might be more variable for a sample of non-native speakers than for a sample of native speakers. Given a fixed-effect factor distinguishing between native and non-native speakers, each set of speakers can be assigned its own standard deviation for the by-subject random intercepts, thereby modeling the heteroskedasticity directly (instead of correcting p-values post-hoc for non-sphericity).

Fourth, mixed-effects models can handle autocorrelational structure in data elicited from subjects over time, whether obtained from a stretch of speech or in

an experimental context. Human behavior is consistent over time, and this often gives rise to autocorrelations in language data. For instance, although there are fluctuations in speech rate, the speech rate at time t is likely to be very similar to the speech rate at the immediately preceding timesteps $t-1$, $t-2$,.... If the sequence of responses elicited from a given subject constitutes an autocorrelated time series, then it is essential to bring this autocorrelation into the model. If ignored, the residual errors will enter into autocorrelations, violating the assumption of independence of the residual errors, and giving rise to suboptimal conclusions about significance. The simplest way in which autocorrelations can be brought into a mixed model is by including as a separate predictor the response at the preceding point in time. For detailed discussion of experimental longitudinal effects, the reader is referred to Baayen and Milin (2010).

Fifth, the estimates provided by the mixed-effects model for the adjustments to the population parameters (the BLUPS) are *shrinkage estimates*. A danger inherent in fitting a statistical model to the data is overfitting. By way of example, consider a sample of subjects for which speech rate is recorded. Some subjects will have a faster speech rate than others. The more extreme the speech rate of a given subject is, the less likely it is that in a replication study the speech rate of that subject will be equally extreme (or even more extreme). It is much more likely that in the replication study the speech rate of this subject will have "regressed" or "shrunk" towards the mean. Mixed models anticipate this regression towards the mean and implement estimates for the BLUPS that shrink the adjustments in the direction of the mean. As a consequence, predictions for replication studies with the same subjects or items will be more precise.

Sixth, more than two random-effect factors can be included in the model. Returning to the above example, one possible design is to embed the same vowel in different carrier words. In such a design, vowels are repeated independently of the words, and hence the vowel should be considered as a potential third random-effect factor.

Finally, mixed-effects models tend to be better able to detect effects as significant. Baayen et al. (2008) show, on the basis of simulation studies for several experimental designs, that mixed-effects models offer a slight increase in power without giving rise to inflated Type I error rates, when compared with traditional analyses based on subject and/or item means. More important than the (generally small) increase in power is the much greater flexibility offered by mixed-effects models for bringing into the model specification various sources of variability that are unavailable when working with subject or item means. Even though longitudinal autocorrelational structure is as such often not of specific interest to the researcher, by taking it into account in the statistical model, the data become less noisy, and the effects of actual interest are more likely to reach significance (see e.g. De Vaan et al. 2007, as well as Baayen and Milin 2010).

22.2.4 Generalized linear mixed models (GLMMs)

Thus far, we have considered a dependent variable, duration, that is real-valued, and for which a model assuming normally distributed (Gaussian) errors is reasonable. Two commonly encountered dependent variables require special attention. First, instead of being continuous, the outcome of an experimental observation can be binary: true versus false, correct versus incorrect, success versus failure, present versus absent, etc. This kind of dependent variable is referred to as a binary, or binomial response variable. Second, a response variable can represent how often a phenomenon occurs in a given time window. In this case, we are dealing with count data.

For binary response variables, the traditional approach is to aggregate over trials (by subjects, or by items) to obtain proportions. Subsequently, analysis of variance or multiple regression is applied with these proportions as dependent variable. Three problems arise with this kind of analysis. First, instead of the variance being independent of the mean, the variance changes systematically with the mean, reaching a maximum when the proportion equals 0.5. This violates the assumption of homoskedastic variance that is fundamental to standard regression and analysis of variance. Second, proportions are bounded between 0 and 1, but the linear model assumes the dependent variable can assume any real value. The generalized linear model deals with these problems by taking as dependent variable not the proportion P,

$$P = \frac{\#\text{ successes}}{\#\text{ successes} + \#\text{ failures}}, \tag{7}$$

but the log odds ratio (or logit)

$$L = \log \frac{\#\text{ successes}}{\#\text{ failures}}. \tag{8}$$

The log odds ratio ranges from minus infinity to plus infinity, and thus circumvents the problem with the boundedness of proportions. (An alternative to the logit link function that can be attractive for researchers familiar with signal detection theory is the probit link function.) The generalized linear model also implements different options for how the variance changes with the mean. For binary dependent variables, the appropriate variance function is that of a binomial random variable. Given the log odds (or logit) as *link function* and binomial variance, it becomes possible to obtain for each individual observation a good estimate of the probability of a success (or a failure).

A response variable may also represent counts. For example, for a series of interviews of the same length, the number of syllable deletions can be extracted. Just as the normal distribution is often appropriate for measurement data, the Poisson distribution tends to be an approximation for count data. The Poisson distribution has a single parameter, λ, which represents the rate at which a phenomenon occurs.

For one specific syllable, for instance, the rate at which it is deleted might be five times in an interview. For another syllable, the deletion rate might be ten times in an interview. Typical for count data is that the variability in the counts increases with the count itself. The Poisson distribution captures this well, as its single parameter λ represents both the mean and the variance. Thus, a greater mean rate is automatically paired with a greater variance. The generalized linear model for count data takes as dependent variable not the count itself, but its logarithmic transformation. This is the link function for count data. In addition, it uses the Poisson distribution to model how the variance changes with the mean.

The generalized linear model has been extended to incorporate random-effect factors in addition to fixed-effect factors. Crucially, generalized linear mixed-effects models, or GLMMs, do not require any prior aggregation into proportions, as the ambition is to provide estimates of the likelihood of a success (or failure), or the rate at which a phenomenon occurs (in the case of count data), for each individual observational unit.

22.2.5 Significance in mixed-effects models

The significance of covariates and fixed-effect factors can be evaluated in two ways. One option is to test whether slopes or contrasts are significantly different from zero. For non-Gaussian GLMMs, evaluation is based on Z-scores and associated p-values. For Gaussian models, the relevant t-tests run into the problem that there is no good analytical solution for the appropriate degrees of freedom. For large data sets, the upper bound for the degrees of freedom, the number of observations minus the number of fixed-effect parameters, often provides a good approximation. Informally, an absolute t-value exceeding 2 is a robust indicator of significance for $\alpha = 0.05$.

As an alternative to the t-test, a Bayesian method estimating the posterior distribution of the parameters can be used to obtain 95 percent credible intervals for the coefficients, as well as estimates of the probability of values more extreme than those actually observed. For data sets with at least several hundreds of observations, these probabilities are very similar to the probabilities obtained with the t-test based on the upper bound for the degrees of freedom. For smaller samples, the Bayesian probabilities are more precise. Informally, the Bayesian method can be conceptualized as generating a long series of parameter estimates as might be observed in replication studies. For each simulated replication study, a new set of parameters (intercept, slopes, contrasts, standard deviations, correlations) is generated. One can then inspect the distribution of a given parameter, for instance, the contrast coefficient for `Syllable Type`. If the observed contrast has a value that is extreme for the distribution of simulated contrasts, it is more likely to be significant.

A second option for evaluating significance of a predictor is to compare a model with and a model without a given predictor in order to ascertain whether the

parameters invested for this predictor lead to a non-trivial increase in goodness of fit. For mixed-effects models fitted to measurement data, a likelihood ratio test is appropriate. When two models are compared that differ with respect to the presence or absence of a factor or covariate, then both models should be fitted using maximum likelihood. In case the models have exactly the same factors and covariates in their model specification, but differ with respect to their random-effects structure, the two models are best fitted with relativized maximum likelihood.

The test statistic used by the likelihood ratio test is two times the difference between the log likelihood of the model with more parameters and the log likelihood of the model with fewer parameters. This test statistic follows a chi-squared distribution with as degrees of freedom the difference in the number of parameters. For this test to be precise, the models entering into the comparison should be nested, i.e. the full set of parameters of the model with fewer parameters should be a subset of the set of parameters of the model with more parameters. For generalized linear mixed models, an analysis of deviance test is the functional equivalent of the likelihood ratio test.

22.2.6 Working with mixed models

Mixed models are implemented in a range of software packages (e.g. SPSS, SAS, MLwiN, ASReml, S-Plus) and can be programmed within WinBUGS as well. Open-source software for carrying out mixed-effects modeling is available in `R` (the de-facto standard in statistical computing, freely available at <http://www.r-project.org>) using the `lme4` package by Bates and Maechler (2009).

When working with mixed models, several questions may arise. First, there are cases where it is not immediately self-evident whether a factor is to be modeled as fixed or random. Consider an experiment targeting the duration of English front high and mid vowels. Let `Vowel` denote the pertinent factor with as its four levels the four targeted vowels. Is `Vowel` fixed or random? English has fourteen vowels, so we are dealing with a sample of vowels. On the other hand, the population of vowels is quite small. In this example, `Vowel` is best modeled as a fixed-effect factor. The front high and mid vowels do not constitute a random sample from the population of vowels. The focus of the study is on specifically the four high and mid front vowels, with no aims to generalize beyond these four vowels to, e.g., back vowels or diphthongs.

Second, for a classical linear model fitted to a data set, an R-squared (or adjusted R-squared) value is generally reported. This R-squared specifies the proportion of the variance accounted for by the model (see Kingston, this chapter, for an example). For mixed models, an R-squared is often not reported, because it is no longer a good measure for understanding the contribution of the linguistic variables to explaining the variance: Parts, often very substantial parts, of the variance

are explained by the random-effect factors. In chronometric studies, for instance, linguistic predictors sometimes contribute less than 1 percent to the R-squared (Baayen 2008). If required, the R-squared can be calculated by squaring the correlation coefficient for the observed and expected values of the dependent variable in the case of Gaussian and Poisson models, and the index of concordance (Harrell 2001) for binomial models.

22.2.7 Selected studies using mixed models

Mixed-effects models are a relatively recent development in statistics, and do not have a long history of use in language studies. In psycholinguistics, mixed-effects models are rapidly becoming the new standard for data analysis with repeated measures. Quené and van den Bergh (2008), Baayen et al. (2008), and Jaeger (2008), all in a special issue in the *Journal of Memory and Language*, provide non-technical introductions, with Quené and van den Bergh discussing an example from phonetics, Baayen et al. presenting simulations of data sets as encountered in psycholinguistics, and Jaeger focusing on generalized linear mixed-effect models for binary data. Chapters 1 and 4 of Pinheiro and Bates (2000) are also highly recommended for introductory reading. Examples of psycholinguistic studies of auditory comprehension using mixed models are Baayen et al. (2007), Ernestus and Baayen (2007), and Balling and Baayen (2008). For application of mixed models to corpus-based data, see Ernestus et al. (2006), Janda et al. (2010), and Keune et al. (2005).

22.2.8 Concluding remarks

Mixed-effects models provide the researcher with a powerful tool for understanding the structure of quantitative data. Mixed models are robust with respect to unequal numbers of observations in different cells of one's experimental design. This is a useful property not only for the statistical analysis of experimental data, where observations may be lost due to errors, hesitations, or false starts, but also to observational data sets compiled from corpora, for which unbalanced distributions tend to be the norm.

However, mixed-effects models also have their limitations that come with the assumption that the correct model is linear or additive, and that the modeling problem is sparse in the sense that only a few predictors are assumed to be involved. An excellent complementary tool, especially for high-dimensional observational data, is the random forest technique (Strobl et al. 2009). For highly unbalanced data, random forests may yield fits that are as good or better than those provided by mixed-effects models, as observed by Tagliamonte and Baayen (2010) for a sociolinguistic

data set. As each method has its own strengths and weaknesses, statistical analysis often profits from the insights and perspectives offered by different techniques.

22.3 Clustering and classification methods

Cynthia G. Clopper

22.3.1 Introduction

Clustering, multidimensional scaling (MDS), and factor analysis are all data reduction methods that can be used to visualize and interpret the relationships between variables in high-dimensional spaces. Unlike most of the statistical analyses described in this chapter, clustering, MDS, and factor analysis do not involve rejection of a null hypothesis and do not return a p-value or other metric for assessing statistical significance. The researcher is therefore responsible for selecting and interpreting an appropriate model. Clustering analyses produce a tree (dendrogram) visualization of similarity data, allowing for the identification of hierarchical structure and/or subsets (clusters) of data within the larger set. Multidimensional scaling analyses produce a spatial representation of similarity data in one or more dimensions, in which distance in the space corresponds to dissimilarity, and allows for the identification of the primary dimensions of similarity. Factor analyses identify correlations among variables, allowing for the reduction of the data set to a smaller number of hidden, or unobserved, factors.

22.3.2 Research questions and data types

Clustering and MDS are well suited for exploring the similarity structure of a set of items, including identifying subgroups of similar items and the dimensions along which similarity is defined. In the domain of laboratory phonology, similarity may be defined in terms of perception or production, and may be computed over linguistic units, such as segments, words, or phrases, or over indexical units, such as talkers, dialects, or languages. The perceptual data used in clustering and MDS analyses are typically either confusion matrices of identification responses or explicit similarity rating or classification judgments. Items that are highly confusable, rated as highly similar, or classified together are interpreted as perceptually more similar than items that are less confusable, rated as less similar, or classified

separately. Clustering analyses have been used to examine the relationship between phonological features and perceptual confusions among vowels (Warner 2003) and consonants (Zhang et al. 1982), the phonetic similarity of unfamiliar languages (Bradlow et al. 2007), and the effects of native language on the perceptual similarity of linguistic tones (Gandour 1983) and regional dialects (Clopper and Bradlow 2009). The production data used in clustering and MDS analyses are typically distance metrics, such as difference scores, Euclidean distances in a multidimensional space, or Levenshtein distances, calculated from a set of acoustic (e.g. Heeringa et al. 2009) or phonetic (e.g. van de Velde and van Hout 1999; Heeringa et al. 2009) features.

MDS analyses have also been used to examine the effects of phonological structure and linguistic experience on the perceptual similarity of vowels (Fox 1983; Warner 2003), consonants (Goldstein 1977; Iverson and Kuhl 1996; Harnsberger 2001), tones (Gandour 1983; Francis et al. 2008), intonation contours (Grabe et al. 2003), talkers (Kreiman and Papcun 1991), dialects (Clopper and Pisoni 2007; Heeringa et al. 2009), and languages (Stockmal et al. 2000; Bradlow et al. 2007). Clustering and MDS techniques can also be used together to simultaneously explore the subgroupings of items within the larger set and the dimensions of similarity. For example, Warner (2003) used clustering to examine the hierarchical structure of phonological features in perceptual vowel similarity and MDS to determine the primary dimensions of similarity.

Factor analyses are used for data reduction in projects involving large numbers of independent variables that are correlated with one another. In the domain of laboratory phonology, these variables may be acoustic, articulatory, and/or perceptual. Factor analyses have been used to explore the relationships among different acoustic measures of the glottal source spectrum (Kreiman et al. 2007) and vowel variation across talkers (van Nierop et al. 1973), genders (Bachorowski and Owren 1999), and dialects (Clopper and Paolillo 2006), as well as articulatory measures of vowel production (Story 2005), and factors affecting lexical access in production (Bates et al. 2001). Bates et al. (2001) used factor analysis to reduce a set of fifteen intercorrelated variables related to lexical access to a smaller set of four interpretable factors representing the frequency, length, phonetic content, and meaning of the target word. Factor analysis results are often used in further statistical analyses to show the relationship between the underlying factors and other variables of interest. For example, the results of factor analyses on variable productions of consonants and vowels have been used to predict accentedness ratings (van Bezooijen and van Hout 1985) and to identify social categories such as age, ethnicity, gender, and social class (Horvath and Sankoff 1987).

Clustering and MDS analyses require square (N × N) matrices, where N is the number of items in the data set and the value of any given cell is a pairwise distance, similarity, or dissimilarity measure for the pair of items represented by that cell. For the examples of clustering and MDS analyses discussed in this section, the data set

was a square talker similarity matrix obtained from 22 listeners in an unpublished auditory free classification task (e.g. Clopper and Bradlow 2009). The stimulus materials included 20 male talkers (five from each of four American English regional dialects) producing the sentence *She had your dark suit in greasy wash water all year*. A subset of the 20 × 20 talker similarity matrix is shown in Table 22.3.1. The possible values in the cells range from 0 (for pairs of talkers who were not classified together by any of the listeners) to 22 (for pairs of talkers who were classified together by all of the listeners). Thus, larger numbers (e.g. Midland1 and Midland4) indicate greater perceptual similarity than smaller numbers (e.g. Midland5 and North2). The similarity between any talker and himself is 0 and the similarities are symmetric (North1 to North2 equals North2 to North1). Most implementations of clustering and MDS analyses assume that the distance between any item and itself is 0, and that the distance relationships between items are symmetric, although asymmetric similarities are theoretically possible, particularly for perceptual similarity data (e.g. North Korea is more similar to China than China is to North Korea, Tversky and Gati 1982).

Factor analyses require rectangular (N × M) matrices, where N is the number of items in the data set, M is the number of variables, and the number of variables is smaller than the number of items (M < N). For the factor analysis example discussed in this section, the data set was the rectangular matrix shown in Table 22.3.2. The 20 × 6 matrix includes six acoustic measures for each of the 20 talkers in the example free classification task. The measures were selected to reflect phonetic differences between the dialects (see Clopper and Bradlow 2009), including r-lessness in New England (Rhotic = F_3 midpoint – F_3 offset of /a/ in *dark*); intrusive /r/ in the South (No Intrusive R = F_3 midpoint of /a/ in *wash*); pronunciation of *greasy* as [grizi] in the South (Greazy = proportion of voicing of /s/

Table 22.3.1. A 10 × 10 square matrix showing the perceptual similarity of the five Northern (N) talkers and the five Midland (M) talkers in the sample free classification data.

	N1	N2	N3	N4	N5	M1	M2	M3	M4	M5
North1	0	3	6	11	7	6	10	9	6	8
North2	3	0	5	4	3	4	8	3	6	2
North3	6	5	0	5	10	8	3	4	8	4
North4	11	4	5	0	8	11	5	9	7	8
North5	7	3	10	8	0	5	5	6	9	8
Midland1	6	4	8	11	5	0	6	11	12	8
Midland2	10	8	3	5	5	6	0	5	6	5
Midland3	9	3	4	9	6	11	5	0	6	10
Midland4	6	6	8	7	9	12	6	6	0	6
Midland5	8	2	4	8	8	8	5	10	6	0

Table 22.3.2. A 20 × 6 rectangular matrix showing the values for each of the six acoustic variables for the twenty talkers in the sample free classification data

Talker	Rhotic (Hz)	No Intrusive R (Hz)	Greazy (%)	Greasy Duration (s)	/u/ Retraction (Hz)	Speaking Rate (s)
NewEngland1	139	2350	0.00	0.367	926	4.46
NewEngland2	324	2095	0.22	0.347	393	4.23
NewEngland3	324	2709	0.00	0.286	347	4.88
NewEngland4	265	2533	0.06	0.398	885	4.45
NewEngland5	266	2589	0.20	0.302	487	4.56
North1	343	2213	0.07	0.352	432	4.30
North2	619	2378	0.00	0.390	487	4.20
North3	244	2190	0.08	0.370	796	3.83
North4	453	2356	0.07	0.358	730	4.72
North5	464	2312	0.00	0.297	299	4.49
Midland1	188	2412	0.00	0.327	841	4.21
Midland2	542	2334	0.00	0.331	398	4.23
Midland3	321	2334	0.00	0.334	520	4.33
Midland4	332	2235	0.00	0.380	465	4.00
Midland5	465	2412	0.00	0.353	332	4.34
South1	576	2423	0.00	0.385	420	4.02
South2	376	2113	0.00	0.363	166	4.84
South3	487	2445	1.00	0.197	487	4.24
South4	465	2190	1.00	0.242	244	4.19
South5	465	2257	1.00	0.249	420	5.08

in *greasy*, Greasy Duration = duration of /s/ in *greasy*); /u/ fronting in the Midland and South (/u/ Retraction = F2 midpoint of /u/ in *suit* normalized to F2 of /i/ in *year*); and speaking rate (Speaking Rate = duration of the sentence). Factor analysis variables must be numeric and continuous, but, like the variables in Table 22.3.2, do not need to share the same scale.

22.3.3 Clustering

Two different approaches to clustering, hierarchical and additive similarity, have been used in laboratory phonology and related fields. Both hierarchical and additive similarity models build trees iteratively by identifying the most similar items in the matrix, grouping them together, and then recalculating the matrix by treating the grouped items as a single unit. In the matrix in Table 22.3.1, the cell with the highest value represents the two talkers with the greatest similarity in the set (Midland1 and Midland4), and those two talkers would be grouped together in the first iteration. For hierarchical clustering analyses, different algorithms have been

developed for recalculating the similarity matrix at each iteration. These different methods define the similarity between clusters and individual items in different ways, and can, therefore, produce different results. For example, the Ward and complete methods are compact methods and tend to produce many small clusters that are later joined together. The single method is a chaining method and tends to add single items to existing clusters. Hierarchical clustering algorithms have been implemented in R and SPSS. Baayen (2008) provides examples of hierarchical clustering and R. Everitt et al. (2001) provide a comprehensive introduction to clustering analyses, with an entire chapter dedicated to hierarchical methods.

Figure 22.3.1 shows hierarchical clustering solutions for the talker similarity matrix obtained from the free classification task. The solution using the Ward method is on the left and the solution using the single method is on the right. Distance between items is represented in these figures by the height on the y-axis at which the items are connected. The height values reflect the model distance between items and do not have inherent units. Higher connections indicate more dissimilar objects, whereas lower connections indicate more similar objects. The distance between NewEngland2 and NewEngland3 in the left panel of Figure 22.3.1 is about 5, whereas the distance between NewEngland2 and North1 is about 60.

In clustering analyses, the researcher interprets the clusters by deciding where in the tree to make the cut between objects within a cluster and objects between clusters. The Ward method solution shown on the left in Figure 22.3.1 could be interpreted as showing three clusters with a break at Height $\approx$ 40 or as showing four clusters with a break at Height $\approx$ 30. The single method solution shown on the right in Figure 22.3.1 clearly shows three clusters with a break at Height $\approx$ 17. If we interpret the Ward method solution as having three clusters, the overall structure of the two solutions is similar with Southern, New England, and mixed Midland and Northern clusters. The structures of the Midland and Northern clusters exhibit the primary difference between the two methods: the Ward method clearly separated

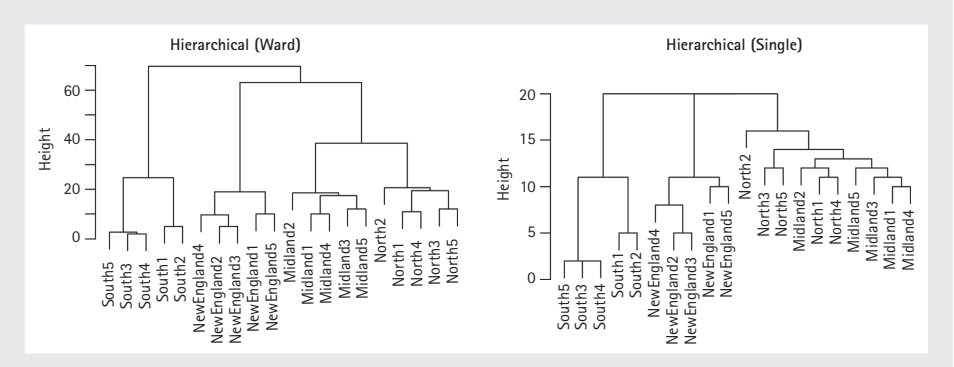

Figure 22.3.1. Hierarchical clustering solutions using the Ward (left) and single (right) clustering methods for the perceptual talker similarity data.

the Midland and Northern talkers into two smaller clusters, whereas the single method produced longer chains of mixed Midland and Northern talkers. The structures of the Southern and New England clusters are virtually identical in the Ward and single method solutions.

Given the strict hierarchical structure of these solutions, for any two clusters, all intracluster distances will be smaller than all intercluster distances. In addition, because the New England cluster is attached to the Northern cluster at about 60 in the Ward model, the distance between any New England talker and any Northern talker is also 60. Thus, for any two clusters, all intercluster distances are equal. These two distance relationships (for any two clusters, all intracluster distances are shorter than all intercluster distances, and all intercluster distances are equal) are intrinsic to hierarchical clustering and therefore hold for all hierarchical clustering solutions, but are intuitively false for many kinds of real data. In the free classification data, some Southern talkers may be more similar to the Midland talkers than others, but hierarchical clustering models cannot capture those differences.

Figure 22.3.2 shows the results of an additive similarity analysis of the free classification data. The additive similarity tree was obtained using Corter's (1982)

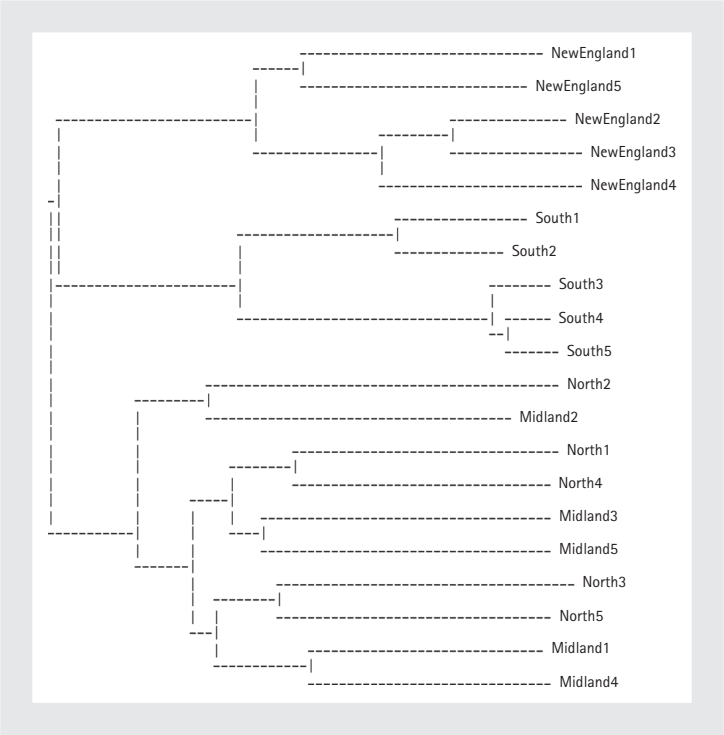

Figure 22.3.2. Additive similarity solution for the perceptual talker similarity data.

ADDTREE program, an implementation of Sattath and Tversky's (1977) Additive Similarity Tree model. Distance between items is represented by the lengths of the horizontal branches connecting the items. Longer branches indicate more dissimilar objects, whereas shorter branches indicate more similar objects. Thus, NewEngland2 and NewEngland3 are the most similar of the New England talkers, because the branches connecting them are shorter than the branches connecting any other pair of New England talkers. As in the hierarchical clustering solutions, the distance between NewEngland2 and North1 is larger than the distance between NewEngland2 and NewEngland3. However, unlike in the hierarchical models, the distance between NewEngland2 and North1 is shorter than the distance between NewEngland3 and North1, because NewEngland2 is closer to the root of the tree than NewEngland3. The overall structure of the additive similarity model is similar to the hierarchical models, with New England, Southern, and mixed Northern and Midland clusters. In the additive similarity model, the Northern and Midland talkers are mixed, similar to the single method hierarchical solution, but the structure is more compact and no chaining is observed. The structure of the New England and Southern clusters is highly similar across the three solutions.

The selection of the clustering model to interpret is based on considerations of the interpretability of the solution as well as the relationship between the data set and the model assumptions. In Figure 22.3.1, the Ward method might be preferred because the separate clusters of Midland and Northern talkers are highly interpretable, and it is less clear how to interpret the chaining of talkers in the single method solution. The additive similarity solution includes a mixed Midland and Northern cluster, but captures the relative similarity of talkers across clusters better than the hierarchical clustering solutions. In the additive similarity solution, South1 and South2 are more similar to the other dialects than South3, South4, and South5, whereas in the hierarchical models, the Southern talkers are all equally similar to the other talkers. Additive similarity clustering is more appropriate for modeling the similarity structure of data that do not exhibit the intracluster and intercluster distance relationships assumed by hierarchical clustering, but can also be used with data where those relationships hold. If hierarchical clustering is used, the linkage method should be chosen based on the interpretability of the solution.

22.3.4 Multidimensional scaling (MDS)

The two most common MDS models in laboratory phonology and related fields are non-metric MDS and individual differences scaling (INDSCAL). MDS analyses are iterative procedures that attempt to maximize the monotonicity of the relationship between the input similarity data and the output distance space. In non-metric and INDSCAL analyses, similarities in the data matrix are rank-ordered (from most to

least similar), and monotonicity is achieved if the rank ordering of the input data is preserved in the rank-ordered output distances (from closest to farthest away). Smaller distances in the MDS solution therefore correspond to greater similarity in the data matrix than larger distances. For the data in Table 22.3.1, Midland1 and Midland4 should be closer in the MDS space than Midland5 and North2. The lack of monotonicity in the output model is reflected in the stress (or badness-of-fit) of the model. Lower stress indicates better model fit than higher stress. Metric MDS analyses are also possible, but they treat the input matrix as ratio data, rather than ordinal data, and are therefore less flexible with respect to the kinds of data that they can be used to model. Kruskal and Wish (1978) provide an excellent introduction to the conceptual and numerical foundations of MDS.

The number of dimensions returned by MDS models is specified by the researcher. However, as a general rule, the number of items in the analysis should be greater than four times the number of dimensions. For an MDS analysis of the 20×20 talker similarity matrix obtained from the free classification task, the maximum number of dimensions is 4. As the number of dimensions increases, the number of parameters in the model also increases, and the fit of the model will improve. The number of dimensions to interpret is selected by the researcher by considering the relative fit and interpretability of models with different numbers of dimensions. The goal is to select a model with a small number of dimensions that has low stress and is interpretable. A scree plot is typically produced to examine the relationship between stress and dimensionality, as shown on the left in Figure 22.3.3 for four independent non-metric MDS analyses of the talker similarity data from the free classification task. The dimension selected for interpretation is usually at the elbow in the scree plot. That is, the selected dimensionality should substantially reduce stress from the next lowest dimension, but not be substantially worse than the next highest dimension. In Figure 22.3.3, the elbow is at two dimensions. The space is interpretable in two dimensions, so the two-dimensional space was selected for interpretation.

The right panel of Figure 22.3.3 shows the two-dimensional space produced by the MDS analysis of the talker similarity data. Non-metric MDS algorithms have been implemented in R and SPSS. The implementation in R is based on Kruskal's (1964) method, whereas the implementation in SPSS is based on Takane et al.'s (1977) ALSCAL model. In non-metric MDS analyses, interpretation of the space and the dimensions of similarity is not restricted to the dimensions returned by the model. The perceptual similarity space in Figure 22.3.3 could be rotated clockwise approximately 30° prior to interpretation, so that one dimension clearly separated the Southern and Northern talkers, and the other dimension clearly separated the Midland and New England talkers. In addition, while most implementations of non-metric MDS analyses center the space at the origin (0, 0), the space can be reflected across either axis and the scale of the space is arbitrary.

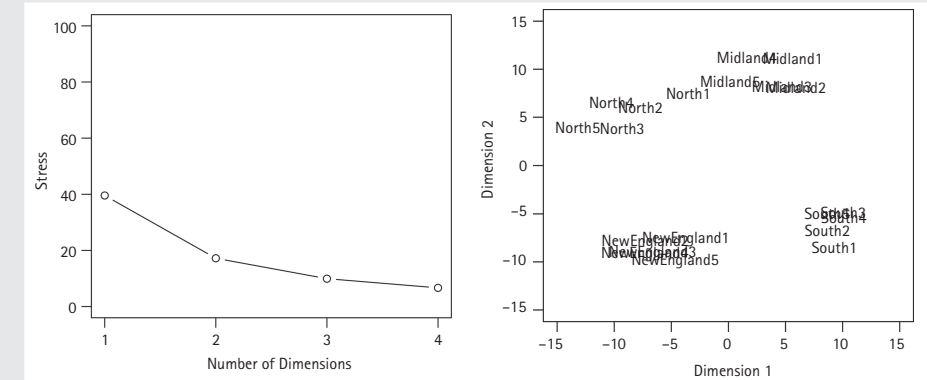

Figure 22.3.3. Scree plot (left) and two-dimensional non-metric MDS solution (right) for the perceptual talker similarity data.

The perceptual similarity space shown on the right in Figure 22.3.3 is also interpretable without rotation. The talkers from the two northern dialects (North and New England) are to the left of Dimension 1, whereas the talkers from the two non-northern dialects (Midland and South) are to the right of Dimension 1. The talkers from the two more stereotyped dialects (New England and South) are to the bottom of Dimension 2, whereas the talkers from the two less stereotyped dialects (North and Midland) are to the top of Dimension 2. As in any statistical analysis, the interpretation of the MDS solution is driven not only by the results themselves, but also by our knowledge and understanding of the data and how they were collected. Thus, the two dimensions of the unrotated similarity space are interpreted as reflecting two important aspects of regional dialect variation in the United States: geography (northern vs. non-northern) and stereotypes (more vs. less). The MDS solution is also consistent with the clustering analyses, and shows separate groups of New England and Southern talkers, but a more mixed group of Northern and Midland talkers.

The interpretation of the dimensions of an MDS solution can be confirmed by regression analyses demonstrating the relationship between the values of the items along a given dimension and some other measure related to the interpretation of that dimension, such as perceptual judgments of voice quality (Kreiman and Papcun 1991) or theoretical vowel features (Fox 1983). The interpretation of the dimensions in Figure 22.3.3 could be supplemented by correlating the values along each dimension for each talker with the acoustic measures shown in Table 22.3.2 to determine which acoustic properties are perceptually salient in the free classification task.

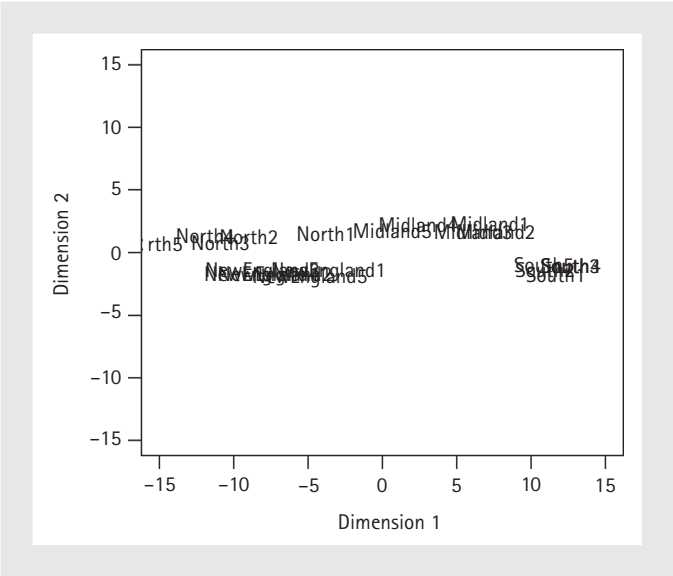

Figure 22.3.4. Hypothetical listener-specific INDSCAL solution with a stretched x-axis and a compressed y-axis.

Given that most MDS models produce solutions that can be rotated, reflected, and rescaled, it is not possible to directly compare two or more MDS solutions. However, INDSCAL analyses can be used to compare solutions across different participants, participant groups, or experimental conditions. For example, INDSCAL has been used to model the effects of native language on the perceptual similarity of tones (Gandour 1983; Francis et al. 2008), and the effect of native dialect on the perceptual similarity of vowels (Fox 1974) and regional dialects (Clopper and Pisoni 2007). The INDSCAL model was develop by Carroll and Chang (1970) and has been implemented in SPSS and Praat. The INDSCAL model accepts a series of square matrices (one per participant, group, or condition) and returns a single similarity space for the set of items, as well as weights for each dimension for each input matrix. The weights reflect the relative strength of each dimension for each participant, group, or condition, and can be visualized as stretching or shrinking the space. If some listeners attended more to geography than stereotypes in the free classification task, an INDSCAL model would return large Dimension 1 weights and smaller Dimension 2 weights for those listeners. Conceptually, the space would be stretched along the x-axis and compressed along the y-axis, as shown in Figure 22.3.4. Unlike non-metric MDS analyses with a single input matrix, INDSCAL solutions cannot be rotated and must be interpreted with respect to the dimensions that the model returns.

22.3.5 Factor analysis

The two most common factor analysis models in laboratory phonology and related fields are factor analysis and principal components analysis. Principal components analysis is a subtype of factor analysis, and the primary difference between them is that principal components analysis uses a single error term to represent all of the variables, whereas factor analysis assigns a different error term to each variable. When the scales of the variables differ (e.g. Hertz vs. seconds in Table 22.3.2), it may be inappropriate to assign the same error term to the distributions of all of the variables. Thus, factor analysis is more appropriate for modeling the structure of data sets that include variables with different scales and/or variances. Factor analysis, including principal components analysis, has been implemented in R and SPSS. Baayen (2008) and Johnson (2008) provide examples of factor analysis and principal components analysis and R. Kim and Mueller (1978a,b) provide a brief, but complete, introduction to factor analysis.

Like MDS solutions, factor analysis solutions can be reflected and rotated. In order to find a unique solution, however, the rotation method must be specified in advance in the analysis, and the resulting space cannot be rotated to improve interpretability. Rotation methods include varimax rotation, which maximizes the variance of the loadings for each factor; quartimax rotation, which maximizes the variance of the loadings for each variable; and oblique rotation, which permits non-orthogonal factors. Varimax rotation is the most commonly used rotation method in laboratory phonology and related fields (e.g. Clopper and Paolillo 2006) because it is more useful for data reduction than quartimax rotation and easier to interpret than oblique rotation. Principal components analysis solutions have a default varimax rotation. Thus, the results of factor analysis with varimax rotation and the results of principal components analysis are typically similar.

The number of factors to interpret is selected by the researcher by considering the eigenvalues and interpretability of the different factors. Typically, factors with eigenvalues greater than 1 are interpreted, although the number of factors to interpret can also be selected by considering the variance accounted for by each factor. The goal is to select a model with a small number of factors that together account for a large proportion of the variance and are interpretable. Eigenvalues can be converted to variance accounted for by dividing each eigenvalue by the total number of input variables. A scree plot can then be produced to examine the relationship between variance accounted for and number of factors. As in the interpretation of MDS solutions, the elbow in the scree plot can be used to select the number of factors to interpret. In the factor analysis with varimax rotation of the six acoustic measures in Table 22.3.2, the analysis returned three factors with eigenvalues greater than 1. However, in the scree plot in Figure 22.3.5, the elbow is at two factors and the third factor was difficult to interpret, so the first two factors were selected for interpretation. The first factor accounts for 38 percent of the variance

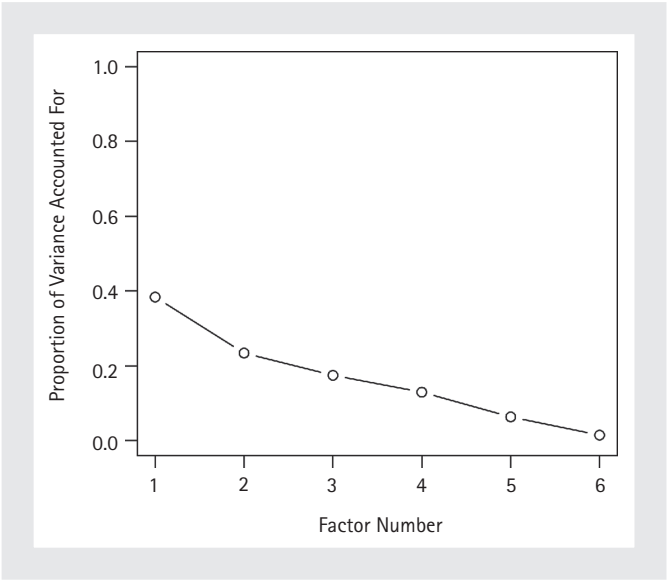

Figure 22.3.5. Scree plot for the factor analysis of the acoustic measures from the free classification stimulus materials.

and the second factor accounts for an additional 23 percent of the variance, for a total of 61 percent of the variance accounted for by the first two factors.

A factor analysis returns factor loadings for each of the variables and, optionally, factor scores for each of the items. The factor loadings of the six variables in the free classification task for the first two factors are shown on the left in Figure 22.3.6. Higher absolute factor loadings indicate greater association between that variable and that factor. For example, Greasy Duration was strongly negatively associated with Factor 1, whereas Greazy was strongly positively associated with Factor 1, suggesting that the variables Greasy Duration and Greazy were strongly negatively correlated. Factor 1 can be interpreted as representing the Southern pronunciation of *greasy* as [grizi]. The variables that were strongly associated with Factor 2 are /u/ Retraction and Rhotic. Rhotic was negatively associated with Factor 2, whereas /u/ Retraction was positively associated with Factor 2, suggesting that backed /u/ productions in *suit* were correlated with r-less productions of *dark* in the sentence analyzed. Factor 2 can be interpreted as representing the New England features of r-lessness and non-fronted /u/s. The other two variables, No Intrusive R and Speaking Rate, were not strongly associated with either factor. Thus, the factor analysis reduced the set of six intercorrelated acoustic variables to two factors that can be interpreted with respect to co-occurring phonetic variation for Southern and New England talkers.

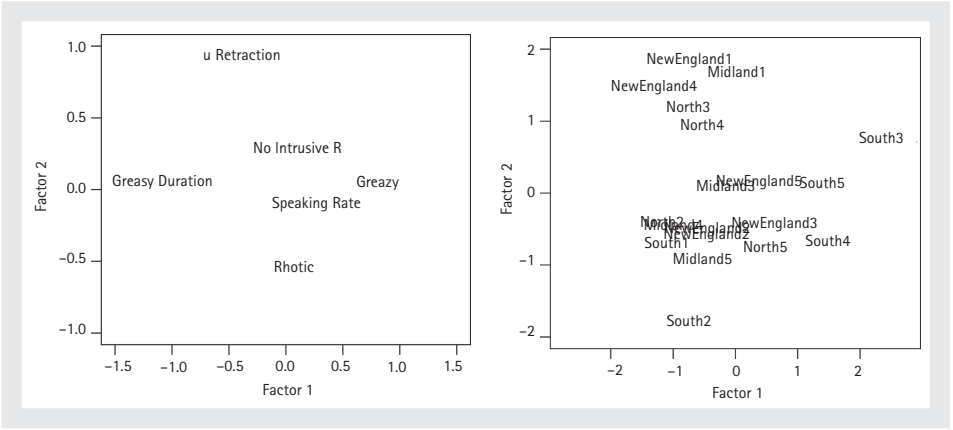

Figure 22.3.6. Factor loadings (left) and factor scores (right) from the two-factor analysis of the acoustic measures from the free classification stimulus materials.

The factor scores of the twenty talkers for the first two factors are shown on the right in Figure 22.3.6. The three talkers who pronounced *greasy* as [grizi] (South3, South4, South5) have high scores on Factor 1. Most of the other talkers, including the other two Southern talkers, have scores below 0 on Factor 1, indicating pronunciation of *greasy* as [grisi]. The talker with the most fronted /u/ (South2) has the lowest score on Factor 2. The talkers with the least constriction of /r/ (i.e. the most r-less productions) in *dark* (NewEngland1, NewEngland4) and with the most retracted /u/s in *suit* (Midland1, North3, North4) have the highest scores on Factor 2. Given that the factor analysis was based on acoustic data and the clustering and MDS analyses were based on perceptual data, the results of the three analyses cannot be directly compared. However, the three Southern talkers who pronounced *greasy* as [grizi] had high Factor 1 scores in the factor analysis and were grouped together in all three clustering analyses and in the MDS analysis, suggesting that [grizi] may be a perceptually salient dialect marker. In general, however, the two-dimensional factor space based on acoustic measures is quite different from the two-dimensional MDS space based on perceptual classification judgments, suggesting that the acoustic measures included in the factor analysis did not fully capture the information available to the listeners in the free classification task.

22.3.6 Summary and future directions

Clustering, MDS, and factor analysis methods have been fruitfully applied to research questions in laboratory phonology and related fields. Clustering and MDS analyses have been used mostly with perception data to explore the perceptual similarity of segmental, suprasegmental, and indexical properties of speech, whereas

factor analysis has been used mostly with production data to explore correlations among acoustic and articulatory measures. However, clustering and MDS analyses could also be used with production data if appropriate similarity metrics could be developed for comparing acoustic or articulatory measures (e.g. Heeringa et al. 2008). In addition, factor analysis could be used with perception data to explore the relationships between different types of tasks and/or responses to the same stimulus materials under different conditions.