Hierarchical regression modeling for language research

Kyle Gorman
University of Pennsylvania, kgorman@ling.upenn.edu
Hierarchical regression modeling for language research

Abstract
I demonstrate the application of hierarchical regression modeling, a state-of-the-art technique for statistical inference, to language research. First, a stable sociolinguistic variable in Philadelphia (Labov, 2001) is reconsidered, with attention paid to the treatment of collinearities among socioeconomic predictors. I then demonstrate the use of hierarchical models to account for the random sampling of subjects and items in an experimental setting, using data from a study of word-learning in the face of tonal variation (Quam and Swingley, forthcoming). The results from these case studies demonstrate that modeling sampling from the population has empirical consequences.

Disciplines
Anthropological Linguistics and Sociolinguistics | Psycholinguistics and Neurolinguistics

Comments

This technical report is available at ScholarlyCommons: http://repository.upenn.edu/ircs_reports/202
Hierarchical regression modeling for language research

Kyle Gorman
Department of Linguistics
Institute for Research in Cognitive Science
University of Pennsylvania
kgorman@ling.upenn.edu

November 1, 2009

Abstract

I demonstrate the application of hierarchical regression modeling, a state-of-the-art technique for statistical inference, to language research. First, a stable sociolinguistic variable in Philadelphia (Labov, 2001) is reconsidered, with attention paid to the treatment of collinearities among socioeconomic predictors. I then demonstrate the use of hierarchical models to account for the random sampling of subjects and items in an experimental setting, using data from a study of word-learning in the face of tonal variation (Quam and Swingley, forthcoming). The results from these case studies demonstrate that modeling sampling from the population has empirical consequences.

1 Introduction

In this document I illustrate the use of a class of generalized linear models known as hierarchical (or random-effects) models as a solution to several problems faced by researchers who wish to draw inferences from language data. While empirical scientists of language shouldn’t shy away from deeper explanations, a large degree of empirical work in linguistics is evaluated by the theory’s ability to fit the data. Accordingly, progress is achieved when more of that variability in human linguistic behavior is accounted for in a principled fashion. Hierarchical regression models conform to an understanding of the generative processes involved in the production of language data, whether in the lab or the field, as the random effects of a hierarchical model is an ideal tool to address the assumption that effects observed in a study are likely to generalize to another sample with similar properties (i.e., that the experimental findings are replicable).

1.1 Terminology

As the terminological waters have been muddied to a great degree, a note about naming is necessary at this juncture. I use the term random effects, making reference to the two different types of random effects, called random slopes and random intercepts. I favor this over the more generic (and subject to interpretation) mixed effects. I also retain the term hierarchical to refer to the class of models which

* Thanks to all the participants in the MLM reading group at the Institute for Research in Cognitive Science in Autumn 2008, and audiences at SPLUNCH, and at NWAV 38 at the University of Ottawa, where some of the results from §3 were presented. Special thanks to William Labov and Carolyn Quam for providing the data examined here. Thank you also to Delphine Dahan, Stephen Isard, Mark Liberman, Chandan Narayan, and John Trueswell. The author was funded by an NSF-IGERT training grant. Data and code used is available at the following URL: http://ling.upenn.edu/~kgorman/mlm.html
include random effects, where others have used multi-level or mixed (see Gelman and Hill, 2007, p. 2, for a discussion of these many terms).

1.2 Prospectus

In the next section (§2), I review generalized linear models operate, but demonstrate that they have serious limitations in addressing the sampling patterns common to language data, and then introduce hierarchical models as a solution; hierarchical modeling is then used to address issues of sampling, and as a tool for outlier analysis. In §3, a sociolinguistic study provides empirical evidence for the relevance of hierarchical models. A psycholinguistic study (§4) shows how hierarchical models can be used to address the language-as-fixed-effect fallacy. A final section (§5) concludes. Here I only discuss model-building and not software, though the analyses here were performed with the freely-available R statistical environment (http://www.r-project.org/).

2 Hierarchical modeling

In this section, I begin with the standard formulation of generalized fixed-effect linear modeling, and proceed to motivate extension to hierarchical models. The precise methods for fitting of these models is largely eschewed, since this procedure is done by software which performs checks for convergence (as described in Pinheiro and Bates, 2000).

2.1 Generalized linear models

Linear modeling is an explicit generative framework for statistical inference. In this approach, data is split into outcomes of interest, $Y$, and a set of predictors, $X$, which represent hypotheses about how these outcomes arise. An additive error term $\epsilon$ is also included, representing the residual variance, the variance not accounted for by the predictors. This set of overdetermined equations can be represented as an equation in one variable.

$$ Y = \beta \cdot X + \epsilon $$

Through a bit of linear algebra (in the simplest case, matrix left division), it is possible derive an optimal value for $\beta$, the parameter that measures the association between $X$ and $Y$, that minimizes $\epsilon$.

$$ \beta \sim \arg\min_{\beta} \epsilon $$

2.1.1 Extensions

To include more, $m$ in all, predictors, $\beta$ can be a vector of $m$ items, and $X$ an $m \times n$ matrix (where $n$ is the number of observations).

$$ Y = \beta_1 \cdot X_1 + \ldots + \beta_n \cdot X_n + \epsilon $$

$$ Y[1, n] = \beta[1, m] \cdot X[m, n] + \epsilon $$

For instance, in most models, an intercept, $\lambda$, is included. In the matrix formulation, this is represented by one of the columns of $X$, as a column vector consisting solely of 1’s. This permits the equation fitted
by the model to have a non-zero \( Y \)-intercept, and corresponds to the sample-level *input probability* in a logistic regression model.

\[
\begin{bmatrix}
\text{outcome}_1 & \text{outcome}_2 & \ldots & \text{outcome}_n
\end{bmatrix} = \begin{bmatrix}
\beta_1 \\
\beta_2
\end{bmatrix} \begin{bmatrix}
1 & 1 & \ldots & 1 \\
\text{observe}_1 & \text{observe}_2 & \ldots & \text{observe}_n
\end{bmatrix} + \epsilon
\] (5)

\[
Y[1,n] = \beta[1,2] \cdot \text{X}[2,n] + \epsilon
\] (6)

\[
Y = \lambda + \beta_2 \cdot \text{X} + \epsilon
\] (7)

Interacting predictors can be created by the product of two or more column vectors in \( X \). *Factors*, discrete predictors, can be created by turning observations into binary vectors and matrices of 0’s and 1’s. The equation can also be augmented with a *link function* relaxing the assumption that the relationship between \( X \) and \( Y \) is strictly linear, or the assumption that \( Y \) is continuous (e.g., in logistic regression).

\[
Y = f_{\text{link}}(\beta \cdot \text{X}) + \epsilon
\] (8)

This *generalized* linear model allows for a great deal of expressibility. Unlike an ANOVA (which is perfectly isomorphic to a linear regression), the fitted coefficients from a linear model are naturally interpretable as the effect size, allowing the researcher to ignore significant-but-miniscule effects, or attend to non-significant-but-massive effects. In addition to finding \( \beta \)'s that fit the model, it is possible to generate new \( Y \)'s by sampling from the confidence interval(s) for the values of \( \beta \) (*simulation*), or generate predictions for \( Y \), given new input values for \( X \).

### 2.1.2 Inferences

By the Central Limit Theorem, uncertainties in estimating \( \beta \)'s are characterized by normally-distributed errors. Therefore, each estimate value and the associated *standard error* can be used to perform significance testing (traditionally, a *t*-test, which converges to the *z*-test in the limit as the number of data points increases). Rejecting the null hypothesis means that there is strong evidence that the linear relationship between some column of \( X \), \( X_i \), and \( Y \) is non-zero (i.e., \( \beta \neq 0 \)). The geometric interpretation is as follows. The data is \((X_i, Y)\) pairs plotted on a Cartesian plane. If the statistical test indicates that there is enough evidence to reject the null hypothesis at some \( p \)-value, this means that the slope of the line, \( \beta_i \), which best accounts for the data given the assumption of a linear relationship between \( X_i \) and \( Y \), is non-zero with confidence \( 1 - p \).

Significance of an predictor’s effect is generally used to infer causality, though it indicates only association. It is possible to reduce the chance that the association between \( X_i \) and \( Y \) observed is in fact governed by some unmodeled \( Z \) by constructing models that includes all predictors which have a plausible casual relationship with the outcome. This way, if an association is actually due to the causal effect of \( Z \) and not of \( X_i \) (which is also correlated with \( Z \)), \( Z \) will suppress the association effects of \( X_i \), and the statistical test will fail to reject the null hypothesis of no association between \( X_i \) and \( Y \). As argued below, the effects of many such \( Z \)'s can only be estimated using random effects, and therefore adopting hierarchical models allows a researcher to consider the consequences of previously-unmodeled predictors on an inference made about a fixed effect. The residual issues, then, are simply the direction of the causal arrow inferred from any associations, and the likelihood of the finding generalizing to the population. While the former is be an ontological problem faced by any scientist, the latter is well-addressed by hierarchical models, and therefore in focus here.
2.2 Motivating random effects

A recent innovation in generalized linear modeling is the ability to fit models that include an additional stratum of predictors encoding a different type of relationship between a factor group and the outcomes (Pinheiro and Bates, 2000). Following the conventions adopted here, these effects are termed random. Models including random effects are known as hierarchical, as opposed to flat models which only include fixed effects, for two reasons. First, predictors described by random effects may be partially codetermined by certain fixed effects (see §2.2.2, §2.3.3, §3.4), so there is some nesting expressed by hierarchical models which cannot be expressed by flat models. Secondly, random effects in hierarchical models are often used to express the hierarchical nature of the sampling process described in the model. For instance, if there are multiple observations taken from a single subject, then with all else held equal, one expects these data to be more like each other than like data from different subjects. This may be a proxy for unmodeled demographic information, or simply reflect subjects’ variable experiences with their environment. In an experimental setting, if stimuli are used to elicit behaviors, data elicited by the same stimulus is also expected to be self-similar.

Given a natural method to exhaustively partition the sample into mutually-exclusive subgroups representing hierarchical sources of data (such as subject or stimulus), any biases associated which each level of the grouping can be estimated. It is desirable to model these biases for three reasons. The first is that any reduction in the residual variance supports the casual interpretation of a fixed effect, since this will result in greater reduction of variance by any truly-associated fixed effects. Secondly, calculating the size and regularity of these biases also allows interpretation of subgroup differences, a model-internal way to perform outlier analysis.

2.2.1 Simpson’s paradox

However, the need for means to account for group differences becomes most clear when considering a phenomenon usually termed, after Blyth (1972), Simpson’s paradox; in reality, the observation has a much longer history (e.g., Pearson et al., 1899; Yule, 1903; Simpson, 1951). The paradox refers to cases where sufficiently large group differences swamp regular trends shared by the whole sample, resulting in a regression line that simply models differences between the groups, and not the trend shared by the groups. A common pattern is that a set of outliers associated with a certain group create the falicious appearance of a trend (Johnson, 2009a). In an extreme case, the slope can be the opposite of the within-group regularities.

This latter scenario is demonstrated in a fake-data simulation shown in Figure 1. In this simulation, temperature is used as a predictor of crop yield, measured for many different regions. Higher temperatures lead to higher yields, but differences between different regions of the sample (environmental regulation, fossil fuel inputs, etc.) result in a large spread of input probabilities. Fitting a regression line to the whole data results in significant \( p < .001 \) negative effect of temperature. Addressing the per-group differences as a random intercept, described below, derives the correct positive temperature effect.

2.2.2 Other approaches to random biases

Before introducing random effects, consider three other approaches faced with variance of a hierarchical nature. The first is to ignore it altogether, pooling all the data from the subpopulations together, which Gelman and Hill (2007) term complete pooling. In this model, there is no method of course to inference about the subpopulation differences, nor of avoiding Simpson’s paradox effects.
A second method for treating subpopulation differences is to model them as fixed effects in a flat regression, using the levels to define a factor group as a set of predictors. This approach, however, causes a host of ill effects. As the number of levels in the group increase, so do the degrees of freedom in the model (and at a rate much faster than in a hierarchical model), and therefore failures to reject the null hypothesis are not straightforwardly interpretable as evidence of no effect since they may result from data sparsity (a Type II error). The model will be overfit, meaning that it is difficult to predict behaviors for unseen levels of the subpopulations, or to identify outlier levels. A final problem that may arise in a multiple regression of this type is singularity, when one predictor uniquely determines the value of another, in this case the different factor group vectors. Anticipating data presented in §3, if one of the predictors in a model is a binary factor for the subject’s sex, and another factor group models subject identity as a fixed effect, then the factor group for the subject determines the sex, because each subject is associated with only one sex. This creates an estimation problem for regression, since there is no way to decide which of these two equivalent predictors should be “credited” for variance lying on two or more parallel lines.

The most extreme approach is what Gelman and Hill term no pooling, in which a different model is fit to each member of the grouping factor. The result is that it is difficult to interpret the results for any of the fixed effects, since the fixed effect results are not constrained to be similar across models in any way. Furthermore, any inability to reject the null hypothesis might be the result of data sparsity induced by partitioning the data, and not lack of an consistent effect across all the models (another Type II error).
2.3 Introducing random effects

It should be clear that these approaches are insufficient, and that a generalization about the properties of certain predictors has been missed. The no-pooling model fails to address a known source of variance, the complete-pooling model faces a number of perils, and the explicit fixed-effects treatment is unpredictable. It is observed that these per-group biases have a near-infinite number of possible levels, and are expected to be (by the Law of Large Numbers) normally distributed, both for seen and unseen levels. Hierarchical models use this normality assumption as a constraint on the estimation of per-group biases.

2.3.1 Random intercepts

Now consider the simplest example of a random effects model; a model with a fixed effect, and a single random intercept. A random intercept is a function that maps members of a factor group onto points on a normal distribution. The value of this function for each level of the factor group is interpreted as an additive contribution to the predicted value of $Y$ for an observation belonging to that level. This additive effect is in addition to any whole-model intercept, if one is included. The model that results might be thought of as partial pooling, since there are model connections (both in the main effects stratum, which accounts the variance common to the sample as a whole, and the interplay between different factor levels which determines the standard deviation of the random effect) still being enforced between different levels of the sample.

$$Y = (\lambda + f_{\text{group}}(\text{level})) + \beta_2 \cdot X_2 + \epsilon$$  \hspace{1cm} (9)

This normal distribution defined by $f_{\text{group}}(\text{level})$ has a mean of 0, and a standard deviation which is the empirical standard deviation of the sample across all seen levels of the random intercept.

$$f_{\text{group}}(\text{level}) \sim \mathcal{N}(\mu = 0, \sigma = \sigma_{\text{group}})$$  \hspace{1cm} (10)

The fitting of this model produces the mapping function, with an associated confidence score (standard error) for each.

2.3.2 Random slopes

A simple extension to this variety of model is a random effect which is not simply an intercept, but rather added to the the $\beta_2$ coefficient for a given predictor $X_2$, with the result that the random effect function maps from each factor group level to a normally-distributed ($\mu = 0$) effect to be multiplied by a fixed effect predictor in addition to the whole-model $\beta_2$. This effect is termed a random slope, since the coefficient determines a distribution across slopes of the fixed-effect $\beta_2$. The resulting model adds a function $f_{\text{group},X_2}$ which interacts linearly with (possibly a subset of) the predictor matrix $X$.

$$Y = \lambda + (\beta_2 + f_{\text{group},X_2}(\text{level})) \cdot X_2 + \epsilon$$  \hspace{1cm} (11)

Note that while the two normal distributions have similar properties, the mapping functions $f_{\text{group}}$ and $f_{\text{group},X_2}$ need not be equivalent if they are included in the same model. Though it is not strictly necessary to include a random slope’s predictor as a fixed effect (as done above), this is generally the correct behavior if one is not sure that all subpopulations have the same relationship between $X$ and $Y$; it may be that some level level of group shows more or less of an association between $X$ and $Y$ than another level, for instance if observations from a certain subject show a greater effect of condition or frequency.
2.3.3 Properties of random effects estimates

Random effects of hierarchical models generate *shrinkage estimates* of these biases, meaning that they systematically underestimate the biases appearing in the experimental data. Somewhat counter-intuitively, this is desirable because of the well-known phenomenon of *regression towards the mean*. For instance, in a lexical decision task, “in replication studies with the same subjects, the extremely slow subjects will be faster, and the extremely fast subjects will be slower responders.” (Baayen, 2008, p. 302). Treating these subjects as fixed effects will overestimate the size of the per-subject effects. Replicability is therefore less likely on two accounts: the experimenter does not have a good estimate of the variability associated with individual subjects, and per-subject effects have overestimated the per-subject variance obscuring regularities common to the sample as a whole.

It is important to note that random intercepts help also avoid the singularity problem (§2.2.2). The random effects vector is assigned a penalty which favors fixed effects over codetermined random effects, a fact noted by Douglas Bates in a post to the R help mailing list dated July 2009 (https://stat.ethz.ch/pipermail/r-help/2009-July/203509.html). This in turn is tied into the fact that hierarchical models fully *nest* their similar fixed-effects. If there simply are no effects of the random groupings included, the model could assign an estimate of zero to each group, thus mimicking the flat model. This property of nesting is a strength of hierarchical models: they make fewer assumptions about the real world.

2.4 Adoption

Adoption of these models in linguistics has been variable. A recent issue of the experimental *Journal of Memory and Language* (Vol. 59, #4, Special Issue on Emerging Data Analysis and Inferential Techniques) is dedicated to such best practices in analysis (e.g., Forster, 2008; Baayen et al., 2008; Jaeger, 2008; Barr, 2008), as was the 2009 WOMM pre-session workshop at the CUNY Conference on Sentence Processing (e.g., Jaeger and Kuperman, 2009; Baayen, 2009; Barr and Frank, 2009); efforts are clearly underway to encourage psycholinguists to adopt these techniques. In §4, I attempt to document their use on experimental designs of arbitrary complexity.

However, in sociolinguistics, things are somewhat farther behind. Jaeger and Staum (2005) is an early use of hierarchical modeling to account for the experimental results of a study of language variation. Daniel Johnson’s 2007 University of Pennsylvania dissertation makes use of hierarchical models of a merger. Johnson is also the creator of *Rbrul* (Johnson, 2009b), a package for R that supports hierarchical models in the style of the classic *VARBRUL* program, which performs logistic regression with only categorical predictors. Though *Rbrul* is sure to prove a useful tool for sociolinguists, it is one black box replacing another, unless users educate themselves about other best practices, such as the treatment of correlated variables (§3.2) and the proper interpretation of effects.

At NWAV 38, held at the University of Ottawa in October 2009, Sali Tagliamonte hosted a workshop entitled “Using statistical tools to explain linguistic variation”, where some of the authors presented arguments against the adoption of hierarchical models. Though it would require a paper unto itself to address this issue fully, the relevant observation here is that the use of hierarchical models is not just a pleasantry, motivated by progress in the field of statistics, but has empirical consequences, demonstrated in the following two case studies.
A stable sociolinguistic variable in Philadelphia: (neg)

The following data is taken from one of the largest urban dialect studies, the Linguistic Change and Variation (LCV) project, which interviewed white speakers in the greater Philadelphia metropolitan area between 1973 to 1977. Though the study had several different samples (including a phone survey), here I focus on data from 156 speakers interviewed face-to-face by fieldworkers from the Linguistics Lab at the University of Pennsylvania. According to the Wall Street Journal, then-Vice Presidential candidate George H.W. Bush asked, in disgust, why anyone would care about how people speak in Philadelphia (Janda and Joseph, 2003, p. 122). However, this data has proved priceless in the study of linguistic variation, and I am most grateful to William Labov for providing it to me.

This section focuses on one of what Labov (2001, henceforth, PLC2) terms stable sociolinguistic variables. It has been argued by variationists that variable components of language production may be governed by a component of linguistic knowledge that consists of grammatical rules/forms that compete for application. These variable rules are associated with their probability of use. One set of conditioning factors not considered here are linguistic, relating to contextual and grammatical effects. In particular, the stable variables vary in their rates of usage across speakers’ style and audience, and gender, age, and socioeconomic status help determine a speaker’s base rate of usage, but they do not show signs of change in rates of usage across apparent (i.e., no significant and regular differences between older and younger speakers in a single time slice) or real time (in longitudinal “restudies”). Many of these stable variables are the residue of language change that did not go to completion, but became socially marked within the speech community.

3.1 Introducing (neg)

One of the best-studied sociolinguistic variables (and historical changes) in the history of English is the variable use of negative n-items (such as none) under the scope of sentential negation, termed negative concord (NC). This was the norm in Middle English, but at some point in the Early Modern period, the opposing tendency to license negative polarity items like any under sentential negation arose, becoming the prestige norm it is today (Kallel, 2004, 2007). The use of NC today carries on at some non-zero frequency in nearly every dialect of English, which is why Nevalainen (2006) calls it a “vernacular universal”. (12a) below illustrates a [+NC] system, as in AAVE or Early Middle English, and (12b) is a [-NC] system, as in Modern Standard English (examples are from Labov 1972, p. 783).

(12) (a) I didn’t tell John to paint none of these.
(b) I didn’t tell John to paint any of these.

Wolfram (1969) noted stratification of this variable, termed (neg), across sex and socioeconomic status (in African-American Vernacular English), and subsequent work identified it as a stable sociolinguistic variable, as it shows no reliable evidence of stratification across age (i.e., change in apparent time). However, one difficulty in investigating social conditioning on variables like (neg) is that there are strong positive correlations between the various socioeconomic measures used in the LCV survey. Chapter 2 of PLC2 makes a strong case for the relevance of a wide variety of socioeconomic measures in Philadelphia.

(13) Measures of socioeconomic prestige:
(a) Occupation
(b) Social mobility
(c) Residential value
Correlation  Kappa  VIF$_{Sc1}$  VIF$_{Sc2}$

| .71  | 14  | 1.4  | 1.4 |

Table 1: Several measures of collinearity between subject and parental education are shown here. Correlation is bounded by [-1, 1], with 1 indicating perfect positive correlation. The kappa statistic is computed after Belsey et al. (1980); $\kappa > 10$ indicates non-trivial collinearity as a rule of thumb. The variable inflation factor (VIF: Davis et al., 1986) analysis indicates “the inflation in size of the confidence ellipse or ellipsoid for the coefficients of the term in comparison with what would be obtained for orthogonal data”, and is calculated for the two educational predictors of (neg) in a logistic model (it is not necessarily equivalent for both members).

(d) Education level of subject
(e) Education level of parents

Sociological fieldwork carried out before the LCV fieldwork determined that these measures are closely correlated with perceptions of prestige in Philadelphia and other Anglophone speech communities.

### 3.2 Correlated socioeconomic predictors

Unfortunately, it is not possible to use these additive predictors in a regression model without giving rise to partial collinearity. Partial collinearity occurs when two predictors have a strong correlation, which makes model-fitting procedures unreliable, as these assume independence among the predictors (the problem of attributing variance accounted for; §2.2.2). It may result in extra-large standard errors, or even worse, counter-directed results for closely-correlated predictors that should have similar parameter values. Subsampling the data may reveal a lot of “flips” in the signs on the $\beta$'s (Jaeger and Kuperman, 2009), and bootstrap analyses may reveal a great degree of fragility in the fit. In other word, collinearity is a potential source of error, both the Type I error of rejecting the null hypothesis of no main effect due to subject-level variance, and the Type II error of failing to reject the null hypothesis due to unmodeled group effects washing out the regularities common to each group. Figure 2 shows scatterplots of the 5-way interactions of the 5 socioeconomic indicators recorded by the LCV fieldworkers: the existence of considerable correlations should be obvious.

There are many ways to evaluate the degree of correlation and what kind of problem it may cause for model fitting: a few diagnostics are shown in Table 1 for the collinearity between Sc1, the level of speaker education, and Sc2, the reported education level of the speakers’ father, in the LCV data. These results suggest a high degree of collinearity is present. For example, the VIF values suggest that the confidence in $\beta$’s resulting from fitting a model including these two correlated predictors is about one and a half times smaller than it would be if the predictors were orthogonal (i.e., not at all correlated).

The relevance of individual socioeconomic measure to cultural behaviors, like the use of a stigmatized sociolinguistic variable, has been debated in the sociological literature; specifically, whether a single measure of status in the community determines social behaviors more or less independently of the effects of others (i.e., socioeconomic prestige is multi-dimensional), or in fact these measures are just a projection onto a single scale of prestige in the community (i.e., social class), which is the ultimate predictor of social behaviors. For instance, Krueger (2004) and Cutler et al. (2008) both consider the relevance of
Figure 2: Pairs plot for socioeconomic predictors; the red lines indicate a smooth fit to the data and the top right shows Pearson correlation coefficients.
various socioeconomic measures to health and healthcare, concluding that the complexity of relationship between the individual measures leads to a rejection of the single-dimension approach. In sociolinguistics, a contrast can be drawn between LCV, which uses many measures of class, and the Linguistic Marketplace index, a single-scale subjective operationalization of the class theory of Pierre Bourdieu (1979) which used in the Montreal survey. Without a method for reducing the partial collinearity of correlated socioeconomic variables, however, it is difficult to determine the association between the socioeconomic predictors and usage of a variable.

3.2.1 Previous approaches to collinearity

The approach to the correlation of predictors in PLC2 is empirically-driven; Labov performs separate regressions with each of the individual measures of socioeconomic status as predictors, and compares each of these models to a unitary measure, SEC, the averaged sum of the score of three of the unitary measures (see below for more sophisticated ways for combining collinear measures). Labov’s findings (p. 117) are that this combined measure is the best predictor (in terms of variance accounted for) across the stable variables studied, followed by the occupation measure (this holds for both careful and casual speech, which he considers separately). This is a useful result for accounting for variance due to socioeconomic class, but it remains to be seen whether the multi-dimensional approach to social class could better describe the data. To test this, the measures must be combined into $X$, the predictor matrix, in an orthogonal fashion.

3.2.2 Principal component analysis of SEC

One method of generating orthogonal predictors out of the correlated socioeconomic measures is to use a dimensionality reduction technique like principal component analysis (PCA). PCA finds a rotated space for the matrix containing all the correlated predictors as column vectors; this space is such that the first column accounts for the most variance of any linear combination of the columns in the input matrix, the second column accounts for the largest amount of remaining variance, and so on. An additional constraint on this computation is that the column vectors must be orthogonal. Once this rotated matrix has been found, the first few column vectors can be extracted and by proxy can be used as orthogonal predictors representing the combination of the input vectors. Of course, this greatly complicates model interpretation since the effect of none of the input predictors can be fully isolated from other correlated predictors. A biplot, which shows the individual samples in the first two dimensions of the PC space, is a way to visualize the relationship between the input vectors (shown in red in Figure 3) and the PC scaling and rotation of this space. The result of a random-effects regression including the first two PC vectors as predictors indicates significant effects for the first principal component, partially containing the two schooling predictors, and the second principal component, partially containing occupation and residential value. Still, it is difficult to determine the effect size of PC predictors, given that for any given input predictor, some of it is contained in each principal component. The proper interpretation should add these effects for each component in the final model together, controlling for differences in input scale.

3.2.3 Residualization of SEC correlated predictors

There exists another approach to this same problem which will allow direct consideration of the individual contributions of each of the five SEC measures. The technique is brute-force but effective. Iteratively, correlated portions of predictors are removed from each of the five socioeconomic predictors, resulting in five new orthogonal predictors; this is known as residualization. To begin, a single predictor ($X_i$)
is selected as a baseline; this enters the new orthogonal matrix as the first column. The second predictor ($X_j$) is replaced by $X_j'$, the residuals from fitting a linear model defined by $X_j \sim X_i$, written $\text{residuals}(X_j; X_i)$. This quantity represents the portion of $X_j$ not linearly explained by $X_i$. If there is a third predictor ($X_k$), then the process becomes slightly more complex. It is not appropriate to use residuals from $X_k \sim X_j$ to create $X_k'$, since the resulting model will fail to account for any collinearity existing between $X_i$ and $X_k$. Nor is it appropriate to use the residuals of $X_k \sim X_i + X_j$, since $X_i$ and $X_j$ are already known to be correlated and this fit will have the same correlational issues. For this reason, the correct approach is to compute the residuals from fitting $X_k \sim X_i + \text{residuals}(X_j; X_i)$. This procedure can be continued iteratively for any finite predictor set.

$$X_j' = \text{residuals}(X_j \sim X_i) \quad (14)$$

$$X_k' = \text{residuals}(X_k \sim X_i + \text{residuals}(X_j \sim X_i)) \quad (15)$$

This may render these simple measures difficult to interpret in terms of scale, however. One simple solution is to adopt an assumption (validated by exploratory analysis; see PLC2, p. 62) that the predictor values (after residualization) are normally distributed, and perform standardization, projecting them onto a space where they can easily be compared with each other, and with any binary predictors added to the model (Gelman and Hill, 2007, p. 57). This is done by subtracting out the mean and divided by the standard deviation times two.

Figure 3: Left: a biplot of the socioeconomic predictors in PC space. Right: the amount of variance accounted for by each principal component.
Table 2: Cross-tabulation for sex and style effects on (neg)

<table>
<thead>
<tr>
<th>Style</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Casual</td>
</tr>
<tr>
<td>n-word</td>
<td>360</td>
</tr>
<tr>
<td>any-word</td>
<td>579</td>
</tr>
</tbody>
</table>

\[ s(X) = \frac{X - \mu_x}{2\sigma_x} \]  

The interpretation of these values, once the model is fit, is straightforward: the prediction for \( Y \) changes by an addition of one \( \beta_i \) for every \( \sigma_i \) (standard deviation) change in \( X_i \), with all else held equal.

### 3.3 Age, Sex, and Style Effects

As a stable sociolinguistic variable, (neg) is expected to show an adolescent peak, be used less by women, and less in formal styles.

#### 3.3.1 Age

Though there is no strong evidence that (neg) is undergoing any major changes across apparent time in Philadelphia, there are differences rates of NC correlated with age in the sample, a process known as *age-grading*. Unfortunately, these are not linear, or even roughly monotonic. As is common, adolescent speakers show a peak in their use of stigmatized stable variable (Tagliamonte and D’Arcy, 2009), with frequency being much lower in middle age. There is also a less-understood, less-studied, but reliable effect that older speakers use stigmatized variants with higher frequency than the middle-aged population, an effect also observed for NC (PLC2, p. 103): “...speakers over 60 show a decided increase. The most remarkable rise for this oldest group is shown for (neg).”

Observing these two points of inflection, in adolescents, middle aged-speakers, and older speakers, this continuous pattern can be modeled using a restricted cubic spline with a small number of knots (Harrell, 2001, p. 16–24; Baayen, 2008, p. 194). Here I choose four, to allow for several more-or-less steady states (i.e., adolescence, adulthood, and old age). One can compare this to Labov’s solution, which is to dichotomize the data into six age-range by creating a five-level fixed effect predictor for age. Such dichotomization is a common solution to predictors which have a non-monotonic relationship with the outcome. However, it is somewhat distasteful, since the sparsity effect is unpredictable in the error it may generate, since these are conditioned upon the number of bins and the locations of the breaks (Cohen, 1983). The dichotomization of age used in PLC2 resulted in non-significant estimates of all the age predictors in the hierarchical model, motivating the use of the restricted spline. The number of knots used for the spline doesn’t seem to bear heavily on the result: three and seven knots were used in earlier analyses with similar results.
3.3.2 Sex

Speaker sex has a well-known effect in stable variation: Eckert (1989) and Labov (1990), among others, note that stratification of stable variables across sex results in women producing fewer of the stigmatized variant than men, all else held equal. These effects are shown the cross-tabulation in Table 2 for the LCV (neg) data (n = 1755).

3.3.3 Style

It is known that speaker style is also an important predictor of linguistic behavior. As expected in stable sociolinguistic variation, the stigmatized variant is used more during casual style. Therefore, the LCV fieldworkers coded whether the n-word or any-word produced under the scope of negation was during casual speech (Style A) or a more careful, formal style (Style B). Labov’s approach is to treat (neg) in separate regressions, one for each of the two styles. Here, I’ve merged the two into one model while maintaining this factor coding, under the hypothesis that other predictors are independent of style, a finding supported in PLC2 with one exception (and anticipating my conclusion, one argued here to be the result of Type II error). See Table 2 for a cross-tabulation of style effects with sex.

3.4 A hierarchical model for (neg)

So far, I’ve sketched a model which includes additive, well-understood external predictors, though the treatment of age and the correlations between the socioeconomic predictors is novel in sociolinguistics. Exploratory analysis left me with no reason to believe there were any interactions between age, sex, style, and the socioeconomic predictors for (neg), and therefore interaction terms were not included. At this point, consider the modeling of the effects of individual speakers. The appropriate way to treat such behaviors is to accept the assumption that speakers have differential behaviors which are normally distributed biases; this is of course the definition of a random intercept with per-speaker levels. The study includes many tokens from each speaker, making this intercept easy to estimate in a hierarchical model. If instead one chose to treat this with a fixed effect, singularity would result, since sex codetermines the various vectors of the individual-level factor group (e.g., speaker “Celeste S.” implies female).

3.4.1 Model specification

All the fixed effects are summarized below.

(17) (a) speaker sex (binary)
(b) speaker style (binary)
(c) a per-subject random intercept
(d) an intercept, representing the overall response bias
(e) a four-knot restricted cubic spline of subject age (continuous)
(f) standardized measure of subject occupation level (continuous)
(g) standardized residualized family residence value (continuous)
(h) standardized residualized subject education level (continuous)
(i) standardized residualized parental education level (continuous)
(j) standardized residualized mobility score (3-point impressionistic scale)

This entire right side of the equation was wrapped with the logistic (inverse logit) link function. This is appropriate for binomial data (coded as 0’s or 1’s), which results from the forced choice between
exactly two options, in this case the use of [+:NC] words like “none” under the scope of negation, versus the use of [−NC] words like “any” in the same environment. The use of logistic regression, as this model is known, is known in sociolinguistics under the name variable rule analysis (VARBRUL) analysis (though VARBRUL as a software package implements logistic regression with only categorical predictors, whereas the model described here includes many continuous predictions). The flat model was compared against a hierarchical model with a per-speaker random intercept.

Results are shown in Table 3. Since the curve of the logistic function is steepest at its center, where the slope is \( \beta/4 \), an upper bound on the difference in the probability of a hit outcome corresponding to a unit change in the predictor is given by this informal divide-by-four rule for logistic regression. For the continuous scales of socioeconomic status, each unit is the size of a standard deviation because of standardization (§3.2.3).

3.4.2 Age

In the flat model, the age spline effects (three in all) are significant, suggesting the existence of age stratification as a cause of variation in (neg). However, this result disappears in the hierarchical model. Once again, the fact that the sample is a single time-slice, and therefore a small number of samples per subject, and one age per subject, conspires to make it difficult to assert the regularity of age in this sample, though the flat model results are suggestive; a denser sample (in particular, one that includes longitudinal data) would perhaps presumably be sufficient to reject it. Age is plotted against random-speaker intercept in Figure 4; no obvious trends manifest themselves, though it is interesting to note the high amount of variance associated with adolescence is more prominent than any peak in the use of this stable variable.

3.4.3 Sex

There is a regular effect of speaker sex, with women using as much as 25% less (neg) by the divide-by-four rule.
3.4.4 Style

There is a style effect of approximately the same size and reliability: speakers in formal style use up to 25% less (neg).

3.4.5 SEC

Higher occupation, residence value, and subject and parental education all hinder the use of (neg). Since these have been projected onto a standardized scale, they can be interpreted directly: occupation is the largest and most regular effect, followed by smaller effects of residence and education. An increase occupation level by a single standard deviation is predicted to decrease the probability of the use of (neg) as much as half. The effects of the two education and residence value predictors are on the order of as much of a 25% change per standard deviation. It should be noted that the sum of these effects is less than linear: the divide-by-four rule interprets the effect at the middle of the logistic curve, which is the steepest point, and therefore the point of fastest change. Figure 5 shows the slope of the regression line mapped back onto the seven-point scale. It appears that the assumption of logistic change in the rate of (neg) usage across the occupation scale is a good fit but too steep at either end. This could be addressed with additional transformations before fitting the model, but the effect is likely to be small.

This particular result contrasts with the findings in PLC2, which does not reject the null hypothesis for all these predictors of (neg). Labov’s “first regression” (p. 99) finds an effect for SEC, age, and an additional effect for subjects from South Philadelphia (in comparison to surveys done in suburban King of Prussia and urban Overbrook and Fishtown; unfortunately, this coding was unavailable). Labov’s “second regression” (p. 117) included all the collinear socioeconomic measures, and finds a regular effect for occupation and the combined SEC measure in both casual and careful style. Indicative of the collinearity
problem, the careful speech regression in PLC2 finds an unexpected positive effect for occupation ($\beta = 7.19, p < .001$), which is not consistent with the normal findings for a stable sociolinguistic variable, which would be expected to be used less frequently by speakers with higher occupational status. This is an artifact of collinearity, as can be seen from Figure 5, where the logistic regression line (obtained from the hierarchical model described in Table 3) for occupation is shown, occupation is negatively correlated with use of NC across both styles.

Labov also finds an effect ($p = .03$) of subject’s education (Sc1), but only for the casual speech regression. The model presented here may lead to a re-evaluation of the claimed differences between these predictors. For instance, Labov writes that “occupation is most closely linked to family background and tends to be the strongest determinant of linguistic patterns established early in life…” (PLC2, p. 114). This is in contrast to educational status, a more mercurial measure in adolescence which coalesces with age and is “is linked more closely with superposed variables that are acquired later in life…” (p. 114). No such interpretation is given for residence value, but presumably it represents a behavior somewhere between life-long occupation and education, insofar as it has characteristics of permanence but is also a lagging indicator of status. Thus it’s interesting to observe that, contra Labov, that there is a strong significant effect of residence value on (neg), and that subject education is a significant effect on (neg) in both styles for the hierarchical model. A final important result is that this model leads us to reject Labov’s claim that parental education has no reliable effect on the use of this variable: in fact, it is a larger and more regular effect than subject education, insofar as two highly-correlated ($r = .71$) predictors remain interpretable after residualization. This finding contrasts with Labov’s.

“Significant or not, there was no single case in which the [Sc2] index was more highly
correlated with the linguistic variable than the \([\text{Sc1}]\) index, and non in which the \([\text{Sc2}]\) index was itself significant. The conclusion is clear. In this community, and perhaps elsewhere as well, the effect of education is cumulative. Children’s used of linguistic variables is determined by how much schooling they have received, not the general educational milieu of the family.”

(PLC2, p. 115, bracketed labels standardized for consistency)

Upward mobility has a predicted negative \(\beta\), indicating that it is correlated with lower use of \(n\)-words under negation. In the flat model this effect is significant, but that result is not reproduced by the hierarchical model. This result suggests that the effect of upward mobility cannot be asserted to generalize to a random population, as it cannot be reliably distinguished from overfitting.

Upward mobility has a predicted negative \(\beta\), indicating that it is correlated with lower use of \(n\)-words under negation. In the flat model this effect is significant, but that result is not reproduced by the hierarchical model. At the risk of interpreting a negative result, I consider it plausible that mobility is similar to the quantity identified with the per-subject random intercept: that which explains how subjects diverge from the conditioning placed upon them by external factors governing language use. Regardless, this result suggests that the effect of upward mobility cannot be asserted to generalize to a random population, as it cannot be distinguished from overfitting.

3.4.6 Model validation

The rejection of the null hypotheses in the hierarchical model not accomplished by the models in PLC2 represent errors eliminated by the treatment of correlated predictors and accounting for speaker differences with a speaker-level random effect. In turn, failing to reject the null hypothesis for age and mobility effects might well be Type II errors resulting from this more powerful treatment, or the PLC2 models, which assert these to be significant, make spurious Type I conclusions resulting from failing to address speaker differences. It is difficult to say which, though a backward stepdown bootstrap validation (200 iterations) performed on the flat model is suggestive: 50 iterations (25%) resulted in the exclusion of 1 or more parameters (44 resulted in the exclusion of 1 parameter, 6 in 2), with age and mobility being by far the most likely predictors to be excluded, suggesting they are unreliable effects. In turn, evaluating the posterior distribution of the hierarchical model with a Markov chain Monte Carlo simulation using 100,000 draws produced a model near-identical to the model fit by maximum likelihood, suggesting the integrity of the hierarchical model is considerable compared to the significances assigned to the age and mobility effects in the flat model. Therefore, I conclude that there is no strong evidence to reject the null hypothesis of no age or mobility effects.

3.4.7 Model comparison

It is worthwhile to demonstrate that there has been an improvement not only in the power of inference, but also that the quality of model fit has improved once the increase in number of parameters has been taken into account. As a first approximation, the use of simulations and bootstrap procedures demonstrate the stability of the hierarchical model compared to the flat model, but it is desirable to quantify this improvement. Since there is a strong dependence between the random effects parameters (i.e., the function mapping from group levels to the estimated normal distribution), the number of actual parameters is significantly greater than the number of effective parameters, which itself is determined both by the number of groups and their members, and the degree to which the data can be approximated by complete-pooling and no-pooling models. Because of this complication, simple methods for demonstrating
goodness of fit, like weighting the log-likelihood values by the number of parameters, as in the Akaike Information Criterion (AIC), are anticonservative with respect to the additional power derived from including random effects as opposed to a fixed-effect factor group with the same number of levels. Still, the AIC for the hierarchical model ($AIC = 1588$) is much smaller, indicating a better fit, than the similar model without per-subject random intercepts ($AIC = 1878$). This is essentially saying that in a stepwise regression, familiar to sociolinguists, the subject-level random effect would not be dropped from the model.

### 3.4.8 Outlier analysis and the leaders of linguistic change

The hierarchical model generates random intercept values and associated standard errors for each speaker. The interpretation of these per-subject intercept values is straightforward: it tells us the degree to which a subject’s use of (neg) is different from an abstract mean subject. For any random effect, be it an intercept or a slope, it is possible to evaluate whether any level is significantly different from the (abstract) mean level by using the estimate for that level of the random effect, and the associated standard error. The null hypothesis, that the subpopulation bias is nil, can be evaluated using a simple $z$-test.

The identification of archetypical speakers driving language change is the primary goal of PLC2. I address the question of whether speakers identified as outliers, once the stratifying predictors like age, sex, socioeconomic status, and style are taken into account, are the same speakers identified as leaders of change. If this is the case, a per-subject random intercept could be a data-driven answer to Labov’s plea in the interstitial notes for the second edition of *Social Stratification of English in New York City*:

> “Many aspects of the NYC study influenced linguists’ later work, but one aspect did not. There are no people in most of the sociolinguistic studies that followed—just means, charts, and trends. Although I have campaigned to bring people back into the field of sociolinguistics there has been only a limited response on this front.” (Labov, 2006, p. 157)

I take this to be a strong refutation of the view that ‘no individual variation should simply be assumed in sociolinguistics; instead the interpretation of individual level differences continues to be a critical part of the study of linguistic variation. However, the precise question how leaders are to be identified in the population is somewhat more complex. Labov (2006, p. 158) writes that, in PLC2, “it is not the exceptional but the prototypical individuals who are in focus.” This may cast doubt on the use of outlier analysis to identify leaders, since these leaders exist embedded in social networks and patterns of change. Labov continues with a warning: “In general, trying to explain exceptional cases is a dangerous procedure, unless we put the same effort into studying unexceptional cases.” Shrinkage estimates of random effects anticipate this concern; random effects require considerable evidence to overcome the bias that speakers are unexceptional.

Speaker outliers are identified as those for whom it is possible to reject the null hypothesis (at $p = .05$) that they are different from an abstract mean speaker in their NC productions. 28 of the 155 speakers are significantly different from the mean speaker. At the very bottom of the list is “Ed D.” (all names here are pseudonyms), who Labov identified (in a post-hoc personal communication) as being one of the oldest and most conservative Irish-American speaker in the sample, and who produces no negative concord in 26 tokens, a shocking degree of regularity. On the other end of the spectrum, “Barbara C.” produces a remarkable 18 tokens of (neg) without a single any-word to be heard. In PLC2, Barbara C., interviewed at age 16, is identified as an profound anti-conformist, an early rejector of the dominant racist ethos of her Irish-American Fishtown neighborhood, upwardly-mobile, a well-connected opinion
leader, influential in her peer group, and extreme in her high production of both stigmatized stable forms and advanced tokens of incoming vowel changes from below. An even more extreme (neg) user is 14-year-old “Theresa M.” (13 out of 14 tokens), also subject to exegesis by Labov, and considered a possible leader of linguistic change in the same mold as Barbara C. The one leader identified in PLC2 who does not appear as an outlier in this model is “Celeste S.”, a pronounced leader of linguistic change earlier in her life (and an outlier for some incoming changes); her productions are slightly below average (though non-significant) for (neg). This is perhaps to be expected in later life: Celeste S.’s profound influence and social standing in her community as an interior-class middle-age woman would predict that she avoids stigmatized variables, while leading in her use of incoming forms. If such leaders account for the adolescent peak, one would predict a Celeste S. to have been a (neg) outlier at a younger age, much like young Barbara C. and Theresa M. at the time of the survey. Thus there is a divorce between outliers and leaders of changes, though this is the result of the social status of different variables; Celeste S. is a leader of incoming changes, and as Labov observes (PLC2, p. 374) she also has one of the highest rates of (dh), a stable variable which is below the level of conscious awareness in the population, but her role as a leader is closely tied to her avoidance of stigmatized variables like (neg). Regardless, the utility of this style of inference to sociolinguistics should be clear.

4 Subject and item effects in an experimental setting

Another case where hierarchical modeling is useful is in psycholinguistic experimentation. Clark (1973) was one of the first (that honor goes to Coleman, 1964) to note that in addition to subject sampling, psycholinguists regularly use language stimuli (henceforth, items, which Clark calls as a group language) drawn from a near-infinite set of possible stimuli, often crossing subjects with items, and both items and subjects with conditions, meaning that in experimental designs, psycholinguists face complex hierarchies of data collection and problems of generalization to the population. Failing to test whether results are sensitive to sampling biases is what Clark called the language-as-fixed-effect fallacy. Before showing the hierarchical approach to this problem, it is worth considering how Clark’s concerns were addressed by the psycholinguistics community.

4.1 $F_1$ and $F_2$ analysis

Clark’s particular suggestion is that experimenters should compute the quasi-$F$ ratio, written $F'$, in ANOVA analysis. In a within-subjects design, where subjects are exposed to multiple conditions, this quantity is computed by constructing a repeated-measures ANOVA which includes the fixed effects of subject, crossed with both item and condition.

\[ Y = \lambda + \beta_s \cdot \text{subject} + \beta_i \cdot \text{item} + \beta_c \cdot \text{condition} + \beta_{sc} \cdot \text{subject} \times \text{condition} + \beta_{si} \cdot \text{subject} \times \text{item} + \epsilon \quad (18) \]

$F'$ is then the ratio of the sum of mean square (MS) of condition and subject \times item over the sum of the mean square of item and condition \times subject, which is $F$-distributed with degrees of freedom determined by the mean squares (Raaijmakers et al., 1999).
\[ \text{nmr} = MS_c + MS_{si} \]  
\[ \text{dnm} = MS_i + MS_{sc} \]  
\[ F_I = \frac{\text{nmr}}{\text{dnm}} \]  
\[ df(F_I) = \frac{(\text{nmr} + \text{dnm})^2}{(\text{nmr}^2/df(\text{nmr}) + \text{dnm}^2/df(\text{dnm}))} \]

In practice though, it is difficult to compute this quantity because of the difficulty of calculating the large subject-by-item cross. Therefore, as Clark shows, a conservative approximation of \( F_I \) is provided by a quantity called \( \min F_I \).

\[ \min F_I = \frac{MS_c}{MS_{es} + MS_i} = \frac{F_1 F_2}{F_1 + F_2} \]

As should be clear, this quantity can be calculated by averaging over separate subject and item analyses (see Forster and Dickinson, 1976), though many critiques have been leveled against this criterion for generalizability. Baayen (2008, p. 289) shows that this is anticonservative in some designs and sample sizes. And though it is possible to derive a \( p \)-value from this model, the result will be drastically overfit and unable to generalize to new data. This is because these models fail to account for the anticipated regression towards the mean, overestimating the subject and item biases, and providing no estimate of how a new subject or item would behave. A further problem is that ANOVA analysis is not appropriate for binomial outcomes (Jaeger, 2008), even with transformations suggested in earlier literature, whereas linear models have the option to deploy the logistic link function. Finally, a technique for transferring \( F \)-analysis to experimental designs of arbitrary complexity has simply not been developed; these analyses are simply too baroque for many experiments and users.

### 4.2 Random effects approach to experimental generalizability

Let’s now consider how subject and items can be treated with random effects. Once the fixed effects are determined for the model, the experimental design must be analyzed for random sources of variance. In standard psycholinguistic designs, these come from two sources, which define the random effects groupings: subject and items. However, subjects and items may interact with each other, and the conditions, as well. In any given study, subjects may be better at worse at the task at hand, but also in different conditions, or with different items, as a result of their differential experience with the random elements of the experiment. The ideal choice of grouping is made under considerations of interpretability and the relative size of the groups, though Baayen (2004) defends the use of subject instead of items as the fundamental grouping in experimental work, and that approach is adopted here. However, items too have their hierarchical characteristics: even inside a conditions, some stimuli may be easier for all subjects, in different conditions (if stimuli appear in multiple conditions), or for some subjects. Of course, these effects are not estimable when a significant number of the cells defined by the cross are not filled, and when this occurs, the inestimable random effect should discarded in further models. For instance, if there is only one observation per subject, this is insufficient to calculate a subject effect: the same is true for an item-given-subject effect, for example, if each subject does not respond to an item multiple times; the matrix must not be too sparse. Likely estimable sources are summarized in Figure 6 (though this is not intended to be exhaustive). A final step is taken from standard model fitting: goodness-of-fit measures can be used to identify significant random effects, and to reject models with additional random...
Subject-grouped effects: are there multiple observations from each subject?
- Yes: (1 | Subject)
  * Are there stimuli presented multiple times for each subject?
    · Yes: (Item | Subject)
  * Are subjects exposed to multiple conditions?
    · Yes: (Condition | Subject)

Item-grouped effects: are there stimuli presented multiple times?
- Yes: (1 | Item)
  * Are there stimuli presented in multiple conditions?
    · Yes: (Condition | Item)

Figure 6: Random experimental sources of variance which are estimable (cf. Clark, 1973, his Table 2); (1 | x) is a random intercept conditioned on x, and (y | x) is random slope of y conditioned on x.

effects that don’t affect replicability.

4.3 An example: the irrelevance of pitch contrasts in English

Quam and Swingley (forthcoming) investigated whether English-learning children know that pitch cannot contrast words in English, as it does in languages with lexical tone, such as Thai. Participants, either adults or children, were taught a novel word which was always pronounced with a salient pitch contour, then were tested on either a change in pitch contour or a change in the vowel. This study addresses a broader question of whether children’s (and adults’) representations of words give weight to all acoustic dimensions equally, or instead ignore dimension that are not used contrastively in the native phonology, such as pitch in English.

Subjects were taught a word paired with a novel item (deebo \[\text{di.bo}\]) in a labeling story. In the adult study, conditions were within-subjects, whereas the study with 2-year-old subjects was between-subjects. At trial time, subjects either heard the original word (with its original pitch), a pitch change, or a change of the primary-stressed vowel (dahbo \[\text{dA.bo}\]) with the original pitch.

(24) (a) Where’s the dēēbo?
(b) Where’s the dēēbo?
(c) Where’s the dāhhbo?

This was presented in the visual world with the target and a competitor object, and proportions of fixations were recorded during the online period for adults and infants. The fixed effect of these three conditions, each represented by one of the above prompts (18) heard at trial time, is the main effect. Look preference was also coded as a binary outcome (this is coded as a hit if and only if more than 50% of time was spent looking at the target deebo). A cross-tabulation of binarized look preference is shown in Table 4. A glance at this table seems to indicate that the vowel change condition favors looks to the competitor, whereas adult subjects appear to be ignoring pitch changes.

4.3.1 Model construction

Since there are multiple subjects and multiple stimuli inside each condition, it is possible to estimate random intercepts for both subject and item, and a random slope for condition given subject (in this
Table 4: Cross-tabulation of binarized looking preference in the 3 conditions

<table>
<thead>
<tr>
<th></th>
<th>target looks</th>
<th>competitor looks</th>
</tr>
</thead>
<tbody>
<tr>
<td>original word</td>
<td>188</td>
<td>4</td>
</tr>
<tr>
<td>vowel change</td>
<td>44</td>
<td>76</td>
</tr>
<tr>
<td>new pitch</td>
<td>114</td>
<td>6</td>
</tr>
</tbody>
</table>

Figure 7: Boxplot of percentage looks to target across trials
study, the items correspond to the particular pitch taught and location of presentation). The outcome in question here is the percentage spent looking at the target. The proportions of look times are transformed into log odds using the logit function (e.g., Barr, 2008). Since some trials have zero looks to target, and others the ceiling value of 1 (only looks to target), proportions were remapped to [0.025, 0.975].

\[
\text{logit}(x) = \log \frac{x}{1-x} = \log(x) - \log(1-x)
\] (25)

A series of linear regressions was then fit to this log-odds outcome data, using condition (a 3-way factor) as a main effect. The first model contains no random effects. For a second hierarchical model, a subject-level stratum contains an intercept and an condition-by-subject slope, as well as an per-item intercept.

(26) (a) a per-item intercept
   (b) a per-subject intercept
   (c) a condition-by-subject slope

### 4.3.2 Results

For these models, \( p \)-values are calculated using Markov Chain Monte Carlo techniques to inference about the posterior distribution. This is needed because it it is not possible determine the number of effective parameters in a hierarchical linear regression, a quantity needed to parameterize the statistical test for the significance of the fixed effects, an issue Douglas Bates addressed in a post to the R help mailing list (https://stat.ethz.ch/pipermail/r-help/2006-May/094785.html). The relative quality of fit between the models is evaluated using ANOVA model comparison.

The vowel-change and pitch-change conditions are compared to the original word condition in a treatment contrast scheme. The flat regression model finds that the vowel-change is significantly different than the original word (\( p < .001 \)), but identifies no effect for the pitch-change. In addition to AIC, the related deviance information criterion (DIC: Spiegelhalter et al., 2002), a generalization of AIC which takes into account the smaller number of effective parameters in a hierarchical model, is appropriate for comparing two hierarchical models to each other. AIC and DIC for the models are given in Table 5; combining the results for the two information criteria and the ANOVA comparisons indicates that the model with a random subject intercept and a random treatment-given-subject slope is the most likely to generalize to a new sample, and therefore this model is selected.

The results of the models also appear in Table 5. Once again, the results suggest that once generalizability is taken into account, there is strong evidence that the vowel-change trials cause many more looks to competitor (60% more by the divide-by-four rule) than the presentation of the original word stimuli. However, there is no strong evidence that there is a difference in fixations to competitor for the pitch-change trial. This is consistent with the analysis of Quam and Swingley. Using a paired \( t \)-test, the authors determined that “adults universally showed no effect of the pitch change, fixating the target (deebo) object equally in response to the trained pronunciation and the pitch change. They also universally showed sensitivity to the vowel change; all participants fixated the target less in response to the vowel change than in response to the trained pronunciation.”

### 4.3.3 Outlier analysis

Three different patterns of subject outliers were identified. Two subjects (#31, 33) are identified as having fewer looks to target across trials (both \( \beta = -.61, p = .02 \)); these effects persist under both treatment
Table 5: Above: AIC and DIC for hierarchical models of look proportions by their random effects; lower values indicates better quality of fit given the additional parameters, and significant ($p < .001$ in both cases) improvements according to ANOVA model comparison are starred. Below: coefficients and $p$-values for the selected model (with subject and condition-given-subject random effects), computed from a 100,000-sample MCMC evaluation.

<table>
<thead>
<tr>
<th>SUBJ, COND × SUBJ</th>
<th>vowel-change</th>
<th>pitch-change</th>
</tr>
</thead>
<tbody>
<tr>
<td>β</td>
<td>−3.247</td>
<td>−.120</td>
</tr>
<tr>
<td>$p$</td>
<td>&lt;.001</td>
<td>.477</td>
</tr>
</tbody>
</table>

Table 6: Subject-effect outliers in the Quam and Swingley study; those conditions expected to be significantly different from the mean for the outliers are starred (*) in the predicted-means stratum.

and contrast coding, suggesting that these two subjects have differential strategies in monitoring the visual world. Analyzing the empirical means in comparison with the random effects predictions (which have the tendency to regress towards the mean) reveals a more-complex picture (see Table 4.3.3). In fact, these two subjects tend to have fewer looks to target in the original and pitch-change trials, but slightly more (non-significantly so, according to the model) in the vowel-change trial. Therefore, it appears these two subjects have search strategies which generate fewer looks to the target, except in the vowel-change trial where, like other subjects, they are attending more to the distractor object: the result is that their cross-trial looks are closer to 50%. The model (erroneously) predicts that they will have fewer looks to competitor during the vowel-change trial, a model-internal compromise.

A second group of subjects (#5, 7, 20, 21, 26) has significantly greater looks to competitor (all $p < .02$) during both the change (pitch and vowel) trials. This seems to implicate a different search strategy yet: in both change trials, these subjects spent significantly more time searching in the competitor space, even though the assumption that they treated the vowel-change trial as contrastive and ignored the
pitch-change as irrelevant to lexical identity is both validated for the study as a whole in the main effects analysis above and true for the empirical means for this group of outliers (see Table 4.3.3). Many possible interpretations for this behavior are available, and can only be decided on examining their behavior across the 200–2000 ms window. This effect is many times larger for the vowel-change trials than for the pitch-change trial (where the effect is miniscule but significant), suggesting that this group of five subjects are those who spent a much larger portion of the window fixating on the competitor object during these trials.

Identifying these subjects in this way and not simply “squashing” outliers allows the researcher to directly assert the study’s replicability in a way not fully available before the development of hierarchical models. Since these subjects remain in the model, this style of outlier analysis does not bleed the dataset or require ad hoc exclusions of data. This is desirable since all these subjects follow the main-effect pattern observed across the larger subject pool, but simply have some interesting patterns of behavior illuminated by the subject-level stratum.

5 Conclusion

The results demonstrate that generalizability is an important concern in inferencing from statistical models. In concert with other best practices, such as those highlighted above, hierarchical models provide language researchers with tools to better understand the relationships between predictors and outcomes. Random effects have positive empirical consequences for statistical inference.

References


Dale J. Barr and Austin Frank. Analyzing multinomial and time-series data. Talk given at WOMM presession to the 22nd CUNY Conference on Sentence Processing, 2009.


