Volume 16
Issue 2 *Selected Papers from NWAV 38*

Article 9

1-1-2010

The Consequences of Multicollinearity among Socioeconomic Predictors of Negative Concord in Philadelphia

Kyle Gorman
University of Pennsylvania, kgorman@ling.upenn.edu

The Consequences of Multicollinearity among Socioeconomic Predictors of Negative Concord in Philadelphia

Abstract
This study is a reanalysis of the external predictors of the use of negative concord in Philadelphia, using archival data from the Language Change and Variation survey. It is shown that the interpretation of the effects of the various socioeconomic measures reported by Labov (2001) was biased by their multicollinearity and by per-subject differences. A new mixed-effects model with residualized socioeconomic predictors and a per-subject random intercept shows the predictive role of all four socioeconomic measures, and the per-subject estimates are used to identify the nascent leaders of linguistic change.

The Consequences of Multicollinearity among Socioeconomic Predictors of Negative Concord in Philadelphia

Kyle Gorman*

1 Introduction

This study is an attempt to apply current best practices in quantitative analysis to investigate external constraints on a sociolinguistic variable, *negative concord* in English. Negative concord is the name given to the use of negative *n*-items (e.g., *none*) under the scope of sentential negation. The examples below illustrate a minimal contrast between negative concord and the competing (and more "standard") negative polarity pattern, respectively.

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

Labov et al. (1968:267f.), Wolfram (1969:153f.) and Labov (1972) all observe this variable in African-American Vernacular English, and later work identifies it as a stable sociolinguistic variable (e.g., one with no reliable evidence of change in apparent time), though highly stigmatized, in the speech of whites as well; it has been called an English "vernacular universal" (Nevalainen 2006). It is one of the variables used to study the relationship between various socioeconomic predictors and stable linguistic variants in Labov 2001 (chapter 3; henceforth *PLC2*). This study returns to the data used in *PLC2*, tokens of this variable from 155 white speakers interviewed face-to-face by fieldworkers from the University of Pennsylvania in the Philadelphia metropolitan area from 1973–1977 as part of the Linguistic Change and Variation (LCV) project.

Section 2 describes the effect of individual speaker differences in creating an *omitted-variable bias problem* and proposes the use of per-speaker *random intercepts* as a resolution to this problem. The following section (Section 3) describes the variable and the external predictors. It is shown that the four socioeconomic measures used in LCV are highly correlated and describes the method of *residualization* to eliminate multicollinearity. Then, in Section 4, a mixed-effects model for negative concord in Philadelphia is described and interpreted. The results show that, *contra* Labov (2001), all four socioeconomic measures (as well as gender and style) are reliable, independent predictors of the use of negative concord. The per-speaker *random intercepts* are used to identify the adolescent leaders of linguistic change, as well as the most conservative speakers. A final section concludes.[1]

2 Speakers and omitted-variable bias

Regression is a statistical technique well-suited to sociolinguistics, because when used properly, it is relatively robust to large numbers of predictors and their interactions, and has been generalized to many different types of interactions between these predictors and outcomes. Despite this power, sociolinguists rarely take into account the problem of omitting predictive variables. For instance, consider a regression of the shape

$$Y \sim \beta X + \varepsilon \tag{2}$$

*Thanks to the participants in the MLM reading group at the Institute for Research in Cognitive Science, and audiences at Penn, and at NWAV 38 at the University of Ottawa. I would like to extend special thanks to William Labov for providing the data considered here and allowing me to distribute it. This paper was greatly improved by the careful comments of Daniel Ezra Johnson. Thank you also to Delphine Dahan, Stephen Isard, Roger Levy, Mark Liberman, Chandan Narayan, and John Trueswell. The author was funded by an NSF-IGERT training grant. The data (thanks to William Labov's generosity) and code used in this study is available at the following URL: `http://ling.upenn.edu/~kgorman/papers/mlm/`.

[1] This study does not discuss statistical software, but all analyses were performed with the R statistical environment (URL `http://www.r-project.org/`).

where outcome Y is predicted by X. If the results show a significant effect of X, one might reasonably infer that there is some causal relationship between X and Y. However, if there is another predictor X' which is a better predictor of Y than is X, the inference that X and Y lie in a causal relationship is invalid. It is more likely the case that X' stands in a causal relationship with Y (and perhaps X as well). This is known as the *omitted-variable bias problem*: no statistical test is appropriate for direct causal inference if an important predictor has been omitted.

This problem frequently arises in sociolinguistic surveys which use standard regression techniques. Since the 1970s, sociolinguists have deployed linear (Lennig 1978) and logistic (Sankoff and Labov 1979) regression to study the internal and external constraints on language variation. The well-known Varbrul program implements one such model, stepwise logistic regression with categorical predictors.[2] These *fixed-effects* models, however, do not provide an appropriate solution to the *nesting*. Sociolinguists regularly collect many tokens from individual speakers, but also wish to model demographic effects (age, sex, gender) which subdivide the sample. Therefore, the identity of speaker is in a nesting relationship with gender; e.g., for every token from speaker "Celeste S.", the value of the gender predictor is female. While it is certainly possible to include both gender and speaker in a fixed-effect regression model, the estimates obtained are of little use because the assumption that the predictor values are independent, necessary for obtaining accurate estimates, has been violated. Therefore, the interpretation of gender or per-speaker effects is invalid.

It is very difficult to avoid this issue with a fixed-effect model without introducing new problems. One approach, more or less standard in sociolinguistics, is known as *complete pooling* (Gelman and Hill 2007:chapter 12). In a complete-pooling model, speaker differences are at best modeled indirectly, e.g., by any demographic predictors, such as sex, age, or gender. This results in two problems. First, as Johnson (submitted) shows, failing to account for differences in speakers' average usage rates of a variable (i.e., their *input probabilities*) results in systematic underestimation of the size of fixed effects in a logistic regression model. Secondly, averaging over speakers in this fashion may result in a spurious interaction first noted by Pearson et al. (1899), and termed Simpson's paradox by Blyth (1972); this paradox has been observed in just about any field where inferential statistics are used (e.g., Yule 1903, Simpson 1951, Wagner 1982, Wainer 1986, Wardrop 1995, Tu et al. 2005). One of the best-known examples comes from Bickel et al. (1975), who were recruited to assess whether Berkeley graduate admissions showed a gender bias. Analyzing the data as a group, the authors found that a female applicant was significantly less likely to be admitted than a male applicant. However, analysis by individual department showed that in most departments which had a significant trend towards either women or men, it was that a female applicant was significantly *more* likely to be accepted than a male one. How could this be? Bickel et al. conclude that women are simply more likely to apply to competitive departments (particularly in the humanities) than men. It is not that the group statistic was incorrect, but simply that the significance interaction of gender and admission was only an indirect association caused by a more innocuous interaction between gender and department choice. Another instance of this paradox is illustrated with a fake-data simulation in Figure 1; in this data set, age is positively correlated with the F1 realization of a vowel, but this association is obscured by speaker differences.

In another approach, *no pooling*, separate models are fit for each speaker. This will be of no use in determining the effect of socioeconomic status, gender, etc., since these values are constant for individuals. Further, it is inconsistent with the expectation that speakers in a speech community share grammar-internal constraints, and social evaluations, of linguistic variables (e.g., Guy 1980).

Estimating speaker effects directly in the model, at the exclusion of gender (i.e., as *fixed effects*), also has detrimental effects. First, it may exhaust the model's degrees of freedom and overfit the data, but secondly it complicates the interpretation of gender effects. Ideally, a model should be able to distinguish between a true effect of female gender and a few outlying speakers who just happen to

[2] A problem with Varbrul and related software which is not dealt with here is that many predictors (e.g., age, income) and outcomes (e.g., acoustic measures) in sociolinguistics are continuous, but Varbrul requires that the outcomes be *binomial* (e.g., two-valued), and predictors themselves all categorical values, which has a detrimental effect on the ability of a model to reject the null hypothesis of a true association (Cohen 1983). Varbrul's inability to fit continuous predictors or outcomes is not a conscious choice by the creators, but an accident of history (Johnson 2009, Gorman 2009b).

Figure 1: Averaging over the speakers (left panel) produces a spurious positive correlation between age and raising (i.e., lower F1), but fitting the groups separately to the same data (right panel) appears to indicate that the speakers are lowering (i.e., higher F1) in apparent time.

be female creating the appearance of a gender difference, but the two cannot be teased apart unless they can be modeled simultaneously; otherwise, a crucial variable has been omitted.

The development of *mixed-effects* (or *hierarchical*) models (e.g., Pinheiro and Bates 2000) allows for *partial pooling*, in which demographic predictors (like gender) can be simultaneously estimated with per-speaker effects without the estimation and interpretation problems associated with nesting. A mixed-effects model includes standard *fixed-effects* predictors of the sort familiar to sociolinguistics, along with an additional stratum of *random effects*. The size of these random effects is disfavored over that of fixed effects, so that variance which could be accounted for by either by a random effect (e.g., speaker), or a fixed effect which nests it (e.g., gender) is always assigned to the latter. The simplest type of mixed-effects model augments regression with a *random intercept*. In this model structure, in addition to the fixed effect predictor(s) X, a normal distribution is estimated; each "level" (or label, in this case speaker) maps onto a point on this normal distribution.

$$Y \sim \beta X + \mathcal{N}(0, \sigma^2)\, Z + \varepsilon \tag{3}$$

This model maps each speaker onto a value on a normal distribution centered at zero. This allows speakers to differ in their input probabilities, but in no other fashion (cf. Bickerton 1971, Guy 1980), and simultaneously, for the effect of nesting predictors to be estimated. Because of the biases imposed on the estimation of individual speaker differences by the mixed-effects model, the size of individual speaker differences is often somewhat smaller than the empirically measured differences in the sample. Somewhat counter-intuitively, this *shrinkage* is desirable because of what is called *regression towards the mean*. For instance, considering lexical decision tasks, Baayen (2008:302) notes that "in replication studies with the same subjects, the extremely slow subjects will be faster, and the extremely fast subjects will be slower responders." Another useful random intercept is that of word; a per-word random intercept provides an elegant way to tease apart word-level effects from word frequency effects (e.g., Johnson submitted). More complex random designs can be used for experimental studies (Baayen et al. 2008, Gorman 2009a).

3 Multicollinearity of socioeconomic predictors

The LCV survey relied on a recent sociological study of the Philadelphia area for the different measures of socioeconomic status. Occupation (Occ), and property value of residence (Res), were

coded on a seven-point scale. Education level was coded in two variables, one corresponding to the number of years the speaker spent in school (Sc1), and the other to the number of years in school of the speaker's father (Sc2), both ranging from 1–23 years.

As is often the case, however, these socioeconomic measures are closely correlated; the Pearson correlation between the occupation, residence, and education measures is shown in Figure 2. This creates a problem for determining which socioeconomic measures are the best predictors of linguistic behaviors, but more importantly, this *multicollinearity* has a detrimental effect on all varieties of regression models; regression assumes that all predictors are perfectly *orthogonal* (that is, uncorrelated), and pronounced multicollinearity among predictors makes model estimates extremely unpredictable, unstable, and often contrary to the empirically-observed trend.

There are two common approaches to multicollinearity. One is to combine the collinear predictors in some fashion to create a new measure, which can be thought of as implementing a unitary measure of the class spectrum (e.g., Bourdieu 1977, Sankoff and Laberge 1978). The simplest form of this approach, adopted in *PLC2*, is to add the measures to make a single score. However, this is not appropriate for addressing the relative contributions of these different measures, as it does not take the differential correlations and regularities into account. A more sophisticated technique is the transformation of the various socioeconomic measures into a new orthogonal coordinate system, a technique known as *principal component analysis* (cf. Gorman 2009a:12). This technique results in a complete elimination of multicollinearity, but interpreting the individual contribution of the different socioeconomic measures in the resulting model is most difficult.

Another class of approaches is more suitable to the treatment of social status as a multidimensional system, and allows, to a first approximation, for the consideration of the variables as partially independent. This technique, common in psycholinguistics, is known as residualization. Iteratively, the portions of collinear predictors are subtracted out, resulting in a new, perfectly orthogonal, set of socioeconomic measures. A single predictor X_i is selected as a baseline and remains unchanged. Given a linear regression like (2), the residuals are set equal to the observed errors.

$$residual(X;Y) := \hat{\varepsilon} \tag{4}$$

This value is just the portion of Y which is not linear with X; as a consequence, it is perfectly orthogonal to X. This is mostly commonly applied between sets of multicollinear predictors. To remove the collinearity between two predictors X_i and X_j, one can select one (X_i) as a baseline and replace X_j with $X_j{}' := residual(X_j;X_i)$. The procedure when there are more than two multicollinear predictors (e.g., X_k), sometimes known as *partialization*, is only slightly more complex. It is not appropriate to use $residual(X_k;X_j)$ for $X_k{}'$, since the resulting vector will not have eliminated any partial collinearity between X_i and X_k. Nor is it appropriate to use $residual(X_k;X_i+X_j)$, since X_i and X_j are partially collinear as well. Instead, the correct approach is to use $residual(X_k;X_i+X_j{}')$. This can continue indefinitely:

$$X_i{}' := X_i \tag{5}$$
$$X_j{}' := residual(X_j;X_i) \tag{6}$$
$$X_k{}' := residual(X_k;X_i+X_j{}') \tag{7}$$
$$X_l{}' := residual(X_l;X_i+X_j{}'+X_k{}') \tag{8}$$
$$X_m{}' := residual(X_l;X_i+X_j{}'+X_k{}'+X_l{}') \tag{9}$$

This may render the set of X's difficult to interpret in scale terms, however. One simple solution is to adopt an assumption, validated by exploratory analysis (*PLC2*: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, defined as

$$X_s := \frac{X-\mu_X}{\sigma_X} \tag{10}$$

The interpretation of standardized predictors is straightforward: the prediction for Y changes by

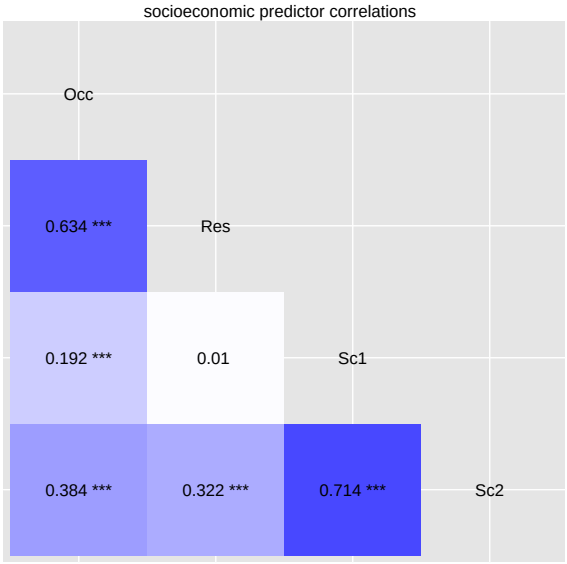

Figure 2: Pearson correlations and values of significance between the socioeconomic predictors.

an addition of one β_i for every σ_i (standard deviation) change in X_i, with all else held equal.[3] Residualization and standardization were applied to all the socioeconomic predictors.

4 Modeling results

Exploratory analysis of the LCV data revealed no additional interactions between the external predictors of age, gender, style, and the socioeconomic measures beyond those already removed with residualization above. The use of negative concord vs. negative polarity (e.g., *any*-words) was modeled with a logistic regression. More (neg), as the variable is written, indicates a higher rate of negative concord in this context. As shown in *PLC2*, age is not monotonically related to (neg); rather, the usage of (neg) shows peaks in adolescence and later life, with a steady state during middle adulthood. For this reason, the age-predictor was wrapped with a restricted cubic spline with four knots (e.g., Harrell 2001:16). As expected of a stigmatized stable variant, men use more (neg) than women, and there is more (neg) in casual speech (Style A) than in formal speech (Style B). These two effects were modeled with simple binary factor predictors. In *PLC2*, separate regressions were fit to the two speech styles, but exploratory analysis indicated that there were no interactions between style and other predictors, so these models can be safely combined into a single model. The fixed-effects logistic model is compared with a mixed-effects model which is identical, except for a per-speaker random intercept. These results are summarized in Table 1. The large decrease in the *Akaike Information Criterion* (AIC) from the fixed-effects model to the mixed-effects model indicates that the additional parameters present in the latter model greatly increase the quality of fit.[4] For this reason, the interpretations given below are those derived from the mixed-effects model.

[3]Gelman and Hill (2007:57) explain that it may be better to replace this denominator with $2\sigma_X$, since this allows a rough comparison between continuous predictors and binary predictors which have roughly equivalent frequencies of outcomes in the sample, because a binomial distribution with $p = .5$ will have the same range and standard deviation as continuous predictors scaled in such a fashion. This method was used for this study.

[4]There is a debate in the statistical literature (e.g. Vaida and Blanchard 2005) about the appropriateness of AIC comparison for mixed-effects models. It is unclear how many free parameters a mixed-effects model has, since the normality of random effects values imposes some clustering on these estimates. The per-subject random effect group here is treated as a single parameter (i.e., standard deviation) nuisance variable.

	Fixed-effects model			Mixed-effects model		
	Estimate	Std. Error	p(>|z|)	Estimate	Std. Error	p(>|z|)
(Intercept)	−0.9935	0.0612	<2E-16	−1.5794	0.1802	<2E-16
Occ'_s	−1.0200	0.1238	<2E-16	−2.1233	0.3660	6.6E-16
Res'_s	−0.7224	0.1298	2.6E-08	−1.1660	0.3624	0.0013
$\text{Sc1}'_s$	−0.8032	0.1156	3.7E-12	−1.2453	0.2767	6.8E-06
$\text{Sc2}'_s$	−0.9662	0.1327	3.2E-13	−1.1916	0.3715	0.0013
Style(CASUAL)	0.4065	0.1170	2.7E-12	0.5184	0.1478	2.4E-12
Style(FORMAL)	−0.4065			−0.5184		
Gender(MALE)	0.2927	0.0596	9.1E-07	0.5971	0.1832	0.0011
Gender(FEMALE)	−0.2927			−0.5971		
AIC	1878.239			1581.197		

Table 1: Results ($n = 1755$) for the fixed-effects and mixed-effects models of (neg).

4.1 Socioeconomic effects

Higher occupation, residence value, and speaker and parental education all result in less (neg). Since these continuous predictors have been projected onto a standardized scale, they can be compared directly: occupation is the largest and most regular effect, followed by the smaller effects of residualized residence value and education. An increase in occupation level by a single standard deviation (approximately one level out of seven) is predicted to decrease the probability of the use of (neg) as much as half, with all else held equal. The effects of parental education (Sc2), speaker education (Sc1), and residence value are all highly significant, though their effects are somewhat smaller.

This particular result contrasts with the findings in *PLC2*, which is unable to reject the null hypothesis that these predictors are unassociated with use of (neg). Labov's "first regression" (*PLC2*:99) finds an effect for the combined SEC measure, age, and an additional effect for speakers from South Philadelphia (in comparison to speakers from suburban King of Prussia and urban Overbrook and Fishtown; this coding was not available in this study, however). Labov's "second regression" (*PLC2*:117) includes all the multicollinear socioeconomic measures at once, and finds a regular effect for occupation and the combined SEC measure in both casual and careful style. Indicative of multicollinearity, as well as pronounced per-speaker differences, the careful speech (i.e., Style B) regression in *PLC2* finds a very large, unexpected *positive* effect for occupation ($\beta = 7.19$, $p < .001$). This is inconsistent with the expected findings for a stable sociolinguistic variable; one expects speakers with higher socioeconomic status to use it less, not more. That the finding in *PLC2* is spurious can be seen from side-by-side plots of the logistic regression line by occupation, and the observed per-style, per-occupation level means in Figure 3. Labov also finds an effect ($p = .03$) of speaker's education (Sc1), but only for casual speech.

These results should lead to a reappraisal of the various contributions of socioeconomic measures to the use of stable linguistic variants, and seems to support a model in which social class is treated as multivariate. For instance, Labov writes that "[o]ccupation is most closely linked to family background and tends to be the strongest determinant of linguistic patterns established early in life…" (*PLC2*:114). This is contrasted to educational status, a measure which coalesces with age and which "is linked more closely with superposed variables that are acquired later in life…" Labov provides no interpretation of a similar type for residence value, but presumably it represents a behavior somewhere between life-long occupation and education, insofar as it is a lagging indicator of status. The model shows a strong significant effect of residence value on (neg), and that speaker and parent education measures have a significant effect on (neg) in both styles. A final important result is that this model is in conflict with Labov's claim that parental education has no reliable effect on the use of this variable (bracketed labels standardized for consistency):

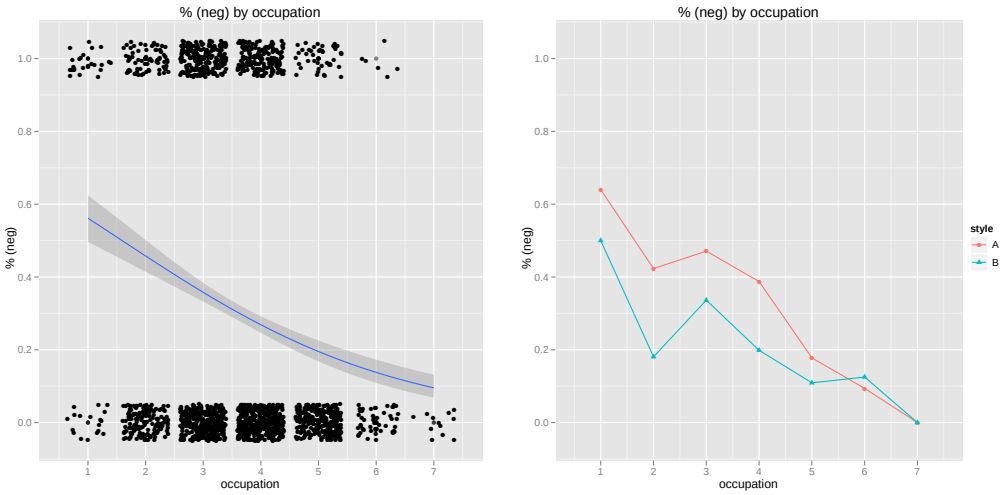

Figure 3: The rate of (neg) across both speech styles with fitted regression line.

> Significant or not, there was no single case in which the [Sc2] index was more highly correlated with the linguistic variable than the [Sc1] index, and none in which the [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 use of linguistic variables is determined by how much schooling they have received, not the general educational milieu of the family.

The finding here, in contrast, is that Sc2 (parental education) has a somewhat larger and more regular effect on (neg) than does Sc1 (speaker education).

4.2 Age effects

In the fixed model, the three parameters in the age spline were found to be significant, suggesting age stratification in (neg). However, this result is not replicated by the mixed-effects model. Gorman (2009a:17f.) shows that there is no correlation between the per-speaker intercepts and age, and a bootstrap validation (ibid.) excludes the age spline in 10% of the iterations, indicating that the age effect cannot be separated from chance.

4.3 Style effects

As expected, casual speech contains significantly more (neg) than formal speech.

4.4 Gender effects

There is a significant effect of speaker gender, with men using considerably more (neg) than women. Interestingly, the coefficient for this effect is twice the size of that in the fixed model, suggesting much of the gender effect was in fact conflated with the low (neg) usage of certain outlier male speakers, the high (neg) usage of certain outlier female speakers, or both.

4.5 Speaker effects

The mixed model estimates an additional intercept and standard error for each speaker. The identification of the archetypical speakers who lead language change is the primary goal of *PLC2*. So far, the random intercept has been treated as a "nuisance parameter," a quantity estimated solely to

avoid the problems caused by omitting speaker differences from the model, but as an estimate of the speaker differences that remain once all other external factors in the model have been controlled for, the per-speaker intercepts directly address this question. Labov pleads for more attention to individual speaker differences in the interstitial notes of *Social Stratification of English in New York City* (Labov 2006:157): "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." This, of course, lies in stark contrast to the view that variationists should simply assume there are no meaningful differences between speakers of the same speech community (cf. Gorman 2009b). However, it is not necessarily the case that advanced outliers are leaders: Labov (2006:158) says, of *PLC2*, that "it is not the exceptional but the prototypical individuals who are in focus. In general, trying to explain exceptional cases is a dangerous procedure, unless we put the same effort into studying unexceptional cases."

In the LCV sample, 28 out of 155 speakers use (neg) at a rate significantly different from the mean (at $p = .05$). Speaker intercepts are plotted in Figure 4. The outlier with the lowest rate of (neg) is "Ed D." (all names here are pseudonyms), one of the oldest and most conservative Irish-American speakers in the sample (Labov, p.c.). Ed D. does not produce a single token of negative concord, and 26 *any* tokens. On the other end of the spectrum, "Barbara C." produces a remarkable 18 tokens of (neg) without a single *any*-word in the scope of negation. Barbara C., interviewed at age 16, is identified as one of the highest users of (neg), as well as a leader of some incoming vowel changes from below. Socially, she is identified as upwardly-mobile, a well-connected opinion leader, a pronounced anti-conformist, and one who rejects the dominant racist ethos of her Irish-American Fishtown neighborhood. Even more extreme is 14-year-old "Theresa M.", who produces 13 (neg) tokens to 1 *any* token. She too is identified by Labov as a non-conformist. The one "leader" identified in *PLC2* who is not regarded as a (neg) outlier is "Celeste S." Her productions of (neg) put her slightly below average for her peer group. This is consistent with the orthodoxy about the leaders of change: Celeste S.'s profound influence and social standing in her community as an interior-class, middle-age woman predict leadership of incoming change from below and avoidance of stigmatized variants. Her role as a leader in other variables is clear, both in many incoming vowel changes, and as Labov observes (*PLC2*:374 and p.c.), in her high use of (dh) (i.e., the realization of /ð/ as [d]), a stable variable which, unlike (neg), lies below the level of conscious awareness in this community. The same is true of her use of (eyC), the lowering of checked /ey/ (e.g., in *cake*), a "new and vigorous" change at the time of the LCV survey, in which Celeste S. is far ahead of her peers. Her status as leader, however, would appear to be intimately tied to her avoidance of stigmatized variables like (neg). If it is correct to regard these leaders as the source of the *adolescent peak*, the increase in use of stable variables during adolescence, followed by a reduction in use in adulthood (e.g., Tagliamonte and D'Arcy 2009), one would predict that Celeste S., who self-describes as an adolescent non-conformist in the model of Barbara C. and Theresa M., was a high user of (neg) at an earlier age.

5 Conclusion

As this study shows, the effect of speaker differences and multicollinearity may obscure the relationship between linguistic variables and external constraints on their use. These associations are revealed by using per-speaker random effects in a mixed-effects model and residualization of correlated predictors, respectively. The results obtained may refine our understanding of socioeconomic predictors of linguistic behavior, and of the multidimensionality of social class. The speaker random effect has been shown to not simply be a nuisance variable, but also a measure of individual speakers' conformity to the population's sociolinguistic patterns. A mixed-effects model with speaker effects turns the hypothesis that there are no meaningful speaker differences, once the external predictors have been accounted for, into a null hypothesis, subject to statistical test. To ignore speaker differences, however, is to elevate it to an assumption, one with the power to render statistics based on it meaningless in the many cases where it is invalid.

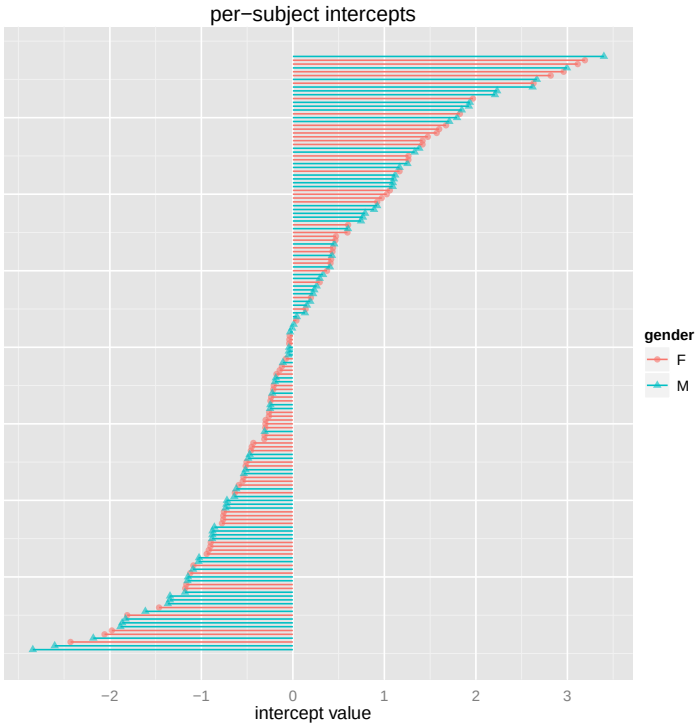

Figure 4: Per-speaker random intercepts estimated by the mixed-effects model

References

Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge: Cambridge University Press.

Baayen, R. Harald, Doug Davidson, and Douglas Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59:390–412.

Bickel, Peter J., Eugene A. Hammel, and J. William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187:398–404.

Bickerton, Derek. 1971. Inherent variability and variable rules. *Foundations of Language* 7:457–492.

Blyth, Colin. 1972. On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 67:364–366.

Bourdieu, Pierre. 1977. L'économie des échanges linguistiques. *Langue Française* 34:17–34.

Cohen, Jacob. 1983. The cost of dichotomization. *Applied Psychological Measurement* 7:249–253.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gorman, Kyle. 2009a. Hierarchical regression modeling for language research. Technical Report 09-02, University of Pennsylvania Institute for Research in Cognitive Science. URL `http://repository.upenn.edu/ircs_reports/202/`.

Gorman, Kyle. 2009b. On `VARBRUL` – Or, The Spirit of '74. Ms., University of Pennsylvania. URL `http://ling.auf.net/lingBuzz/001080`.

Guy, Gregory R. 1980. Variation in the group and the individual: The case of final stop deletion. In *Locating Language in Time and Space*, ed. W. Labov, 1–35. New York: Academic Press.

Harrell, Frank. 2001. *Regression Modeling Strategies*. New York: Springer.

Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3:359–383.

Johnson, Daniel Ezra. Submitted. Progress in regression: Why sociolinguistic data needs mixed models.

Labov, William. 1972. Negative attraction and negative concord in English grammar. *Language* 48:773–818.

Labov, William. 2001. *Principles of Linguistic Change: Social Factors*. Malden, MA: Wiley-Blackwell.

Labov, William. 2006. *The Social Stratification of English in New York City*. Cambridge: Cambridge University Press, 2nd edition.

Labov, William, Paul Cohen, Clarence Robins, and John Lewis. 1968. A study of the non-standard English of Negro and Puerto Rican speakers in New York City. Volume 1: Phonological and grammatical analysis. Cooperative Research Project No. 3288, Columbia University, New York.

Lennig, Matthew. 1978. Acoustic Measurements of Linguistic Change: The Modern Paris Vowel System. Doctoral Dissertation, University of Pennsylvania.

Nevalainen, Terttu. 2006. Negative concord as an English "vernacular universal": Social history and linguistic typology. *Journal of English Linguistics* 34:257–279.

Pearson, Karl, Alice Lee, and Leslie Bramley-Moore. 1899. Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Translations of the Royal Statistical Society, Series A* 173:534–539.

Pinheiro, José, and Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Sankoff, David, and William Labov. 1979. On the uses of variable rules. *Language in Society* 8:189–222.

Sankoff, Gillian, and Suzanne Laberge. 1978. The linguistic market and the statistical explanation of variability. In *Linguistic Variation: Models and Methods*, ed. D. Sankoff, 239–250. New York: Academic Press.

Simpson, Edward. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* 13:238–241.

Tagliamonte, Sali, and Alexandra D'Arcy. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language* 85:55–105.

Tu, Yu-Kang, Robert West, George T. Ellison, and Mark S. Gilthrope. 2005. Why evidence for the fetal origins of adult disease might be a statistical artifact: The "reversal paradox" for the relation between birth weight and blood pressure in later life. *American Journal of Epidemiology* 161:27–32.

Vaida, Florian, and Suzette Blanchard. 2005. Conditional Akaike information for mixed-effects models. *Biometrika* 92:351–370.

Wagner, Clifford H. 1982. Simpson's paradox in real life. *The American Statistician* 36:46–48.

Wainer, Howard. 1986. Minority contributions to the SAT score turnaround: An example of Simpson's paradox. *Journal of Educational and Behavioral Statistics* 11:239–244.

Wardrop, Robert L. 1995. Simpson's paradox and the Hot Hand in baseball. *The American Statistician* 49:24–28.

Wolfram, Walt. 1969. *A Sociolinguistic Description of Detroit Negro Speech*. Arlington, VA: Center for Applied Linguistics.

Yule, G. Udny. 1903. Notes on the theory of association of attributes in statistics. *Biometrika* 2:121–134.

Department of Linguistics
University of Pennsylvania
619 Williams Hall
255 South 36th St.
Philadelphia, PA 19104
kgorman@ling.upenn.edu