

Random-effect modeling of sociolinguistic stratification

Kyle Gorman

Department of Linguistics
Institute for Research in Cognitive Science
University of Pennsylvania



Outline

- The variationist paradigm and some oppositions
- A constructed example and some problems
- Case study: (neg) in Philadelphia
 - Correlations among socioeconomic predictors
 - Speaker random intercepts and individual variation
 - Spurious results in *PLC2* (Labov, 2001)
 - Outliers and the leaders of linguistic change
- Pointers to some past and future work
- Full report(s), updated regularly:
<http://ling.upenn.edu/~kgorman/mlm.html>

The variable: (neg) in Philadelphia

- Negative concord, the use of *n*-items under the scope (C-command) of sentential negation:
 - Hit: I didn't tell John to paint **none** of these
 - Miss: I didn't tell John to paint **any** of these
- Stable sociolinguistic variable
 - So expected to be stratified across class, sex, and style, but to not show apparent- or real-time trends
- Residue of ME change, now socially marked
- Data from LCV: > 100 white Philadelphians interviewed face to face, 1973—1977 (*PLC2*)

The variationist paradigm

- Empirical approach to language, arguing against the null hypothesis that language variation is solely *free* (“optionality”)
- Account for the most variance in usage...
- ...with the fewest free parameters...
- ...usually with a species of regression
- Concerns about sample size and replicability/generalizability [funding? – KG]
- Considerable correlations between predictors

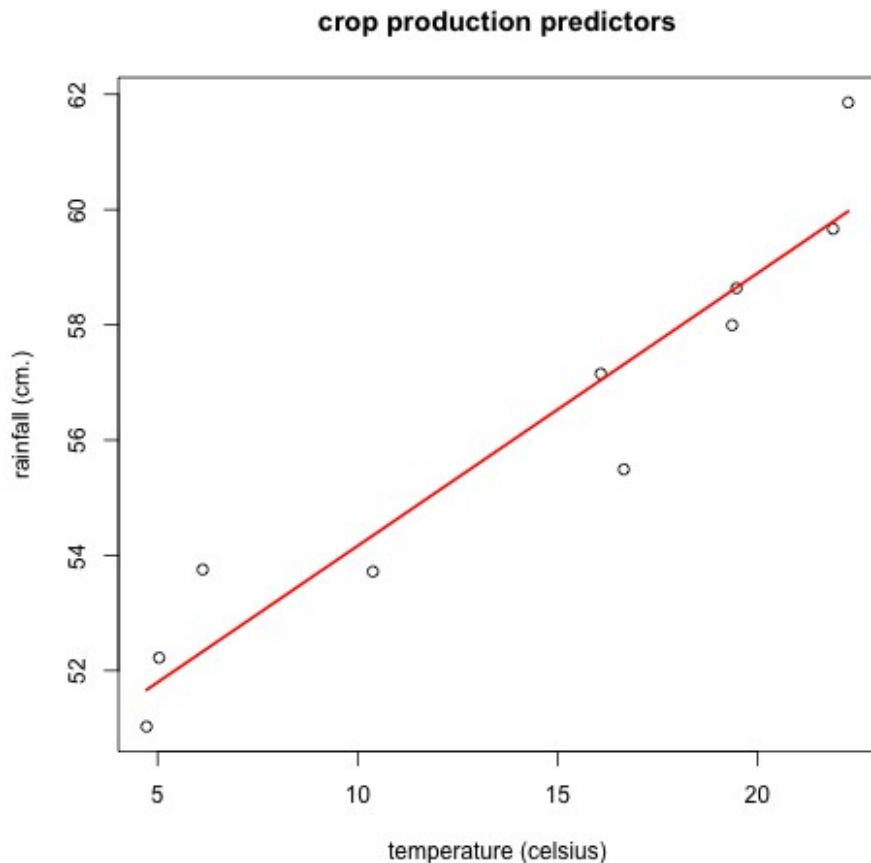
“black box” vs. software vs. statistical method

- "...hence the problems countenanced by critics of Varbrul will not be solved by substituting one black box [Varbrul – KG] for another [Rbrul? – KG], but only by a more thorough understanding of the alternatives and their *empirical consequences*." – John Paolillo [emphasis mine – KG]
- Empirical consequences of this talk? Several unintuitive results are shown to be spurious.
- "...statistical models [stepwise logistic regression – KG] and statistical software [Varbrul, which is a barebones implementation of stepwise logistic regression – KG] need to be carefully distinguished..." – John Paolillo
- No software in this talk (but I use \mathbb{R})

“Null hypothesis” vs. tested hypothesis

- “..it is also true that the assumption of no significant individual-level differences in variation has the status of a *null hypothesis*, and itself *requires evidence to refute*. Since there are different theoretical views of the actual sources of linguistic variation, one has no recourse but to settle these questions on a case-by-case basis, and this is not possible with blind application of the multi-level model...” – John Paolillo [emphasis mine – KG]
- Did you look? Untested “null hypotheses” have the status of *assumption*, and can have consequences
- HLMs the *only* appropriate way to check...
- And the model finds significant individual-level variance which results in different findings

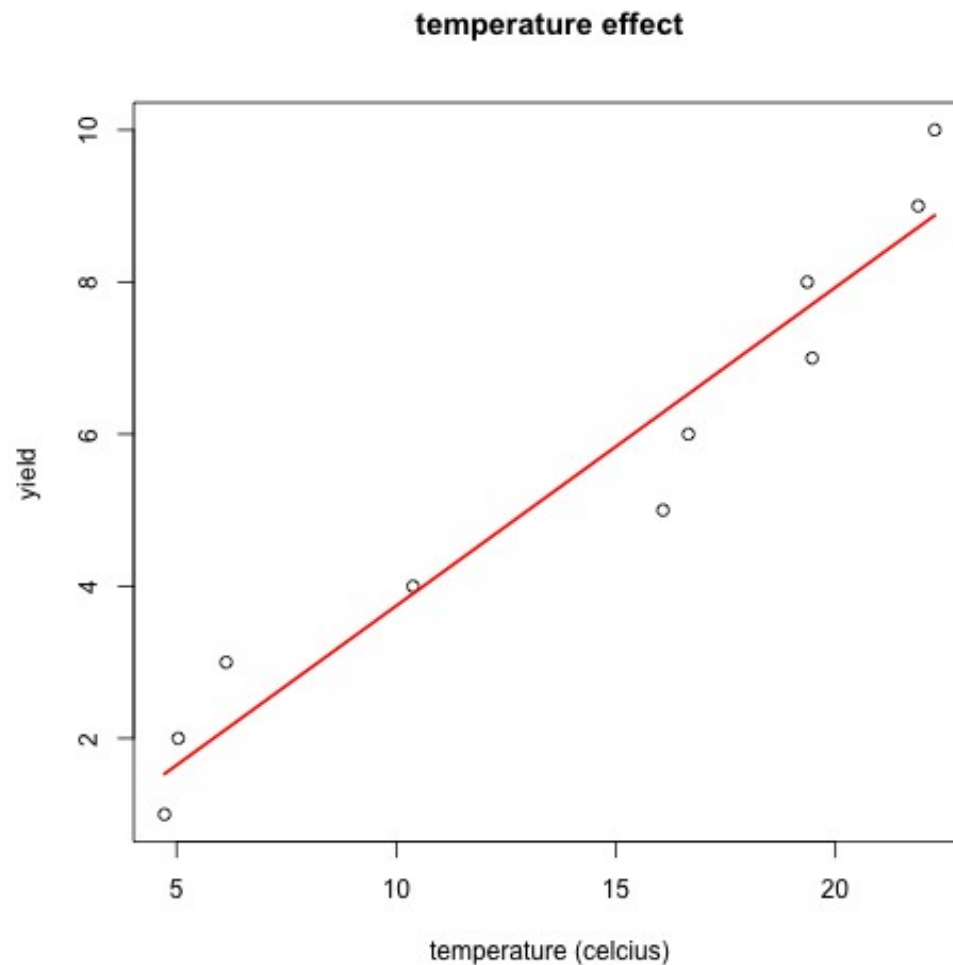
Constructed example: crop growth



- “Researchers interested in crop production measured per-country yields...
- ...and two predictors: avg. temperature and rainfall”
- But it's wet in the hot places...and vis versa, so they're correlated ($r = .93$)

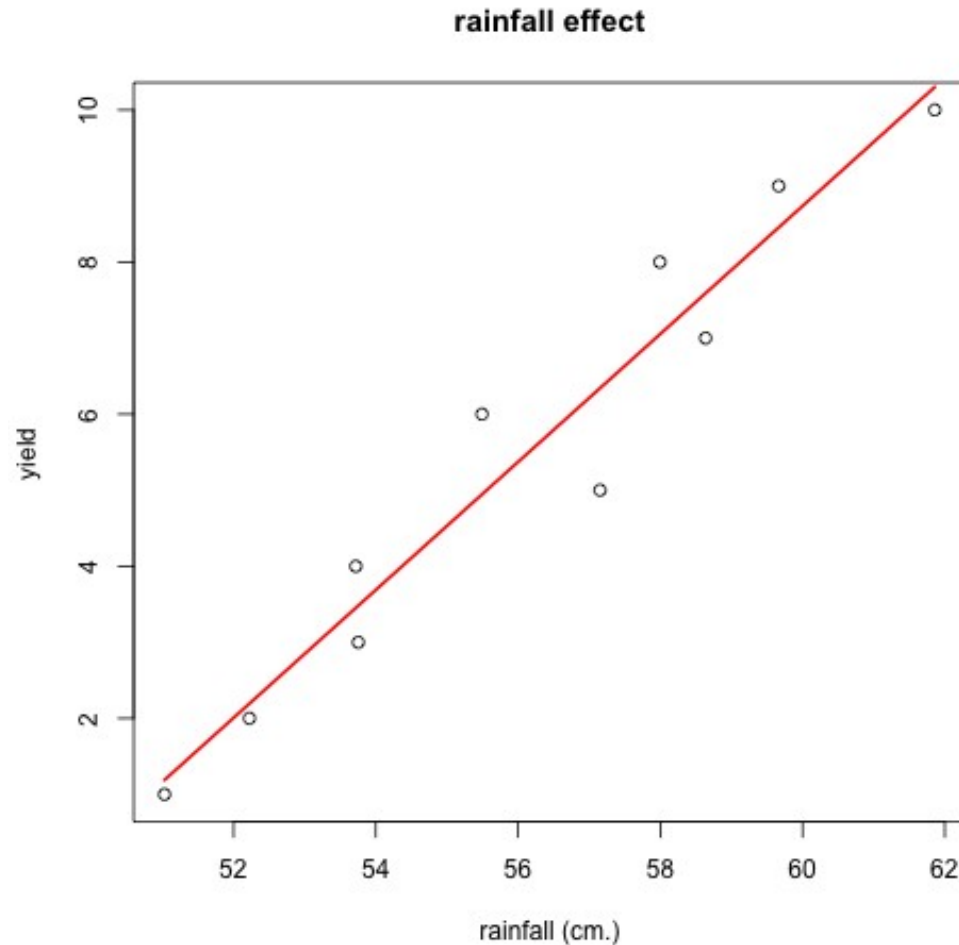
Temperature effect

- temperature a reliable predictor of yield ($p < .001$)



Rainfall effect

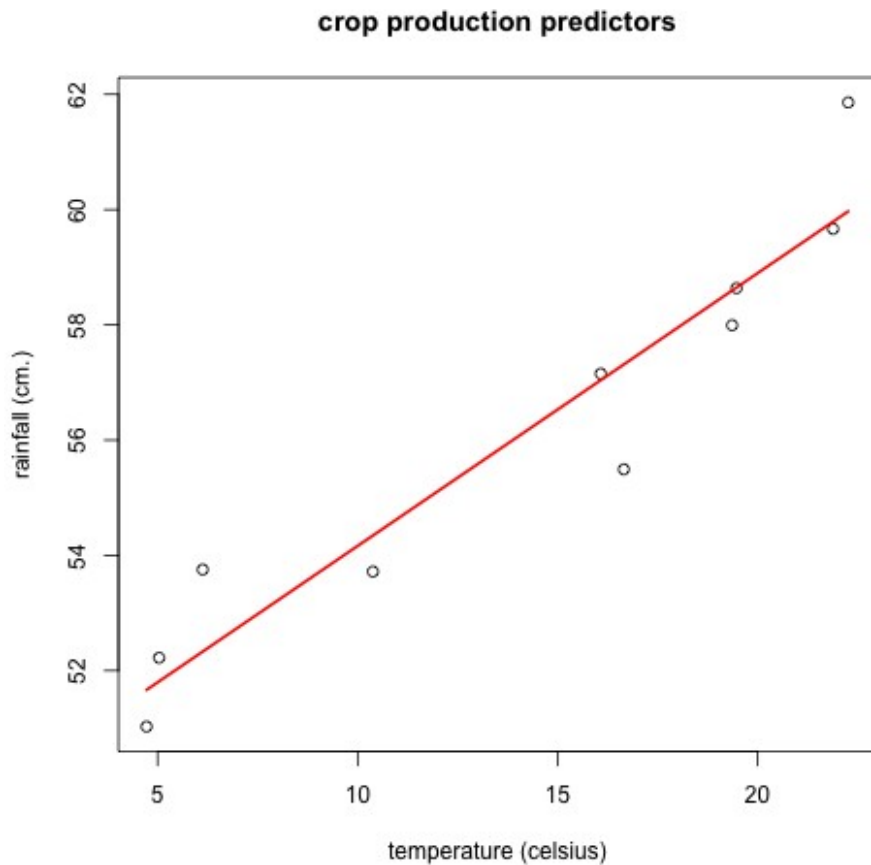
- `rainfall` a reliable predictor of `yield` ($p < .001$)



Together, though...

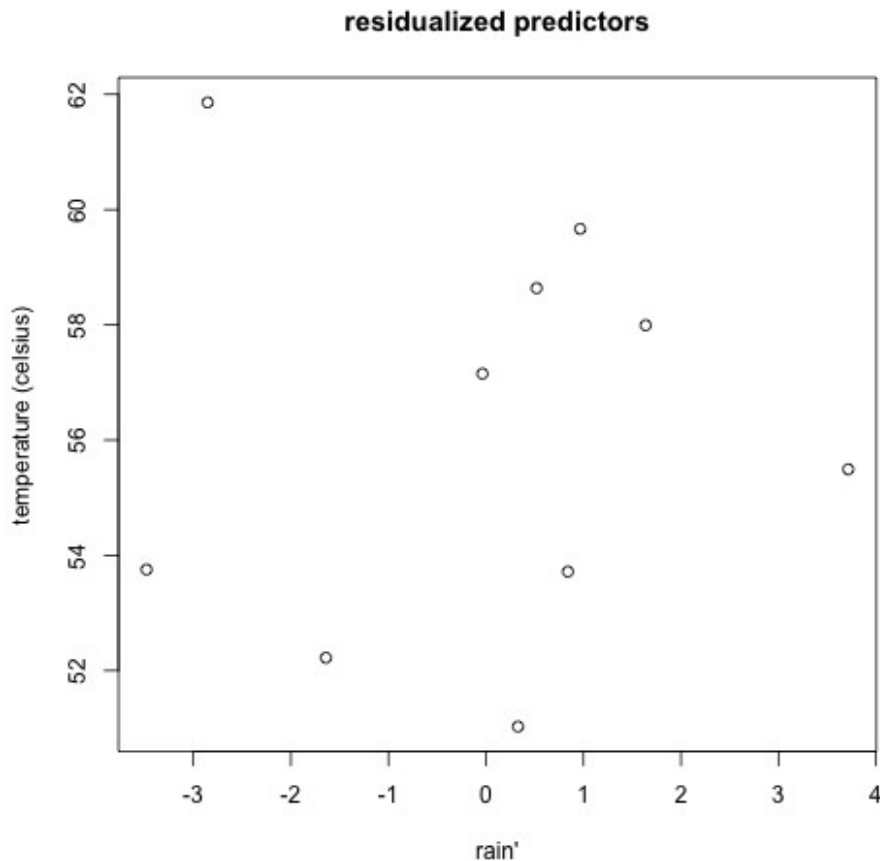
- There is no effect of *either* predictor ($p > .1$ for both)
- I know this is wrong, because I made up the data. They're both linearly related to each other, and the outcome, with a bit of noise.
- Regression simply cannot handle large correlations among predictors
- What we need is to somehow *orthogonalize* the predictors in an interpretable way

Residualization



- Adjust `rainfall` by subtracting out what's predicted by `temperature`
- `rainprime`: the portion of `rainfall` not predicted by `temperature` alone

Residualization

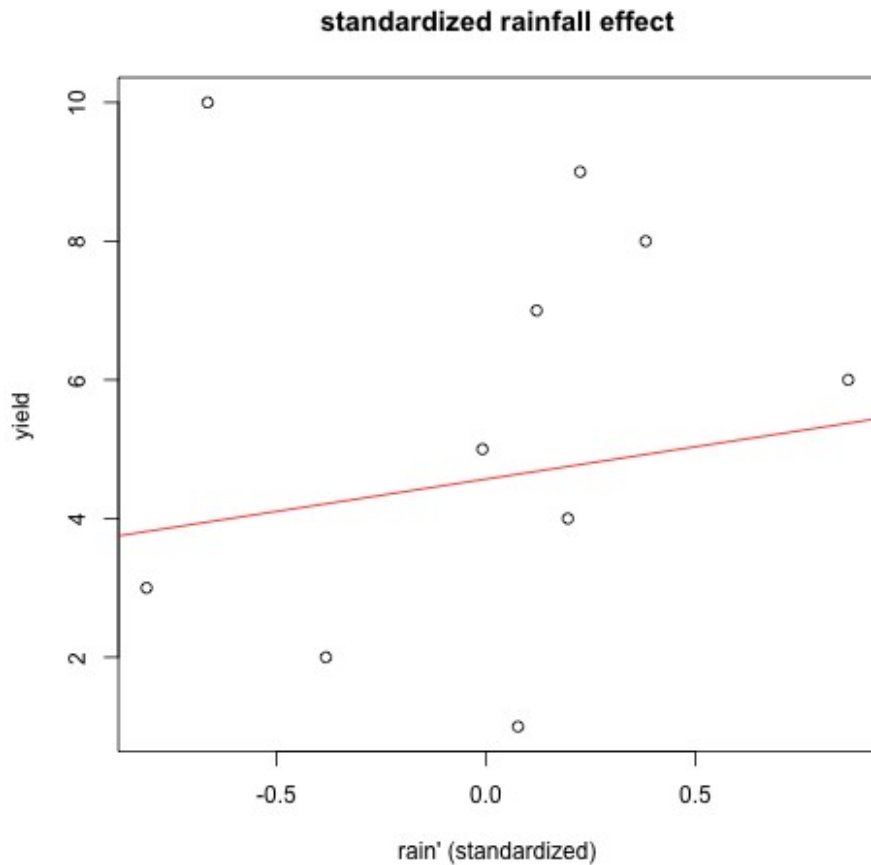


- Adjust `rainfall` by subtracting out what's predicted by `temperature`
- `rainprime`: the portion of `rainfall` not predicted by `temperature` alone

Standardization

- What do units of `rainprime` mean?
- My approach: project all predictors onto a common scale, one they also share w/ factors
- Actual method: subtract out the mean and divide by the standard deviation x 2
- Does not usually have an effect on prediction, just interpretation...

Outcome



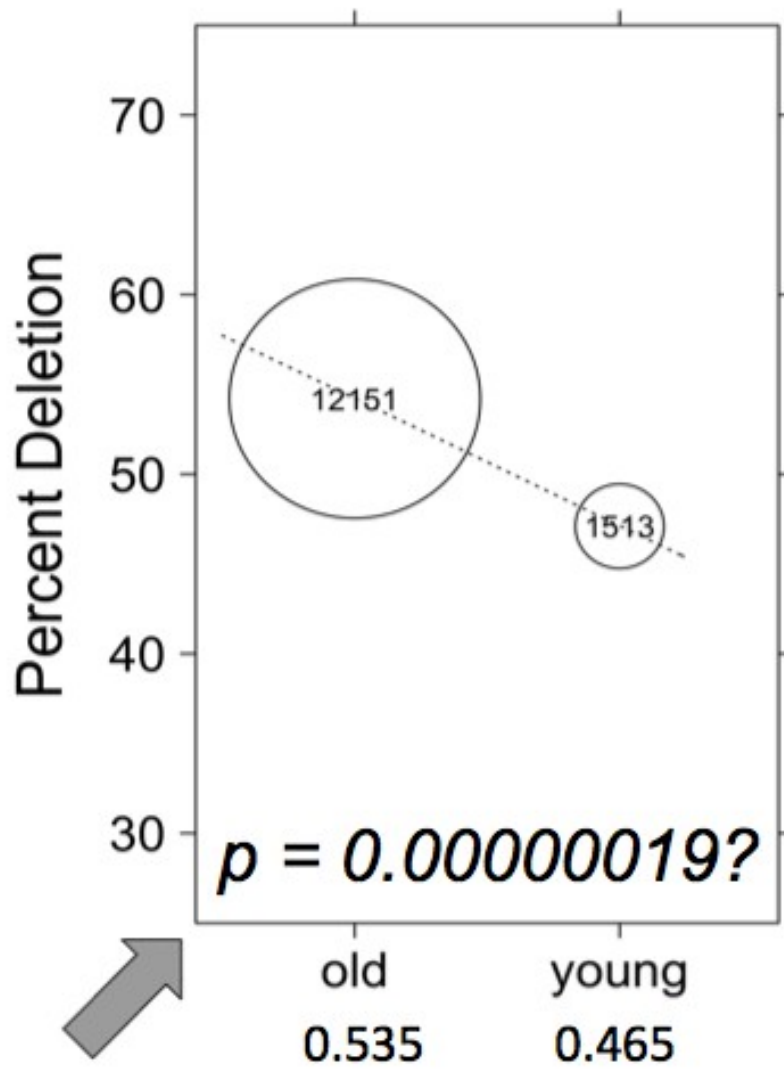
- Large effect of temperature on yield ($p < .001$)
- Small effect of rainprime left over ($p = .05$)

Simpson's paradox

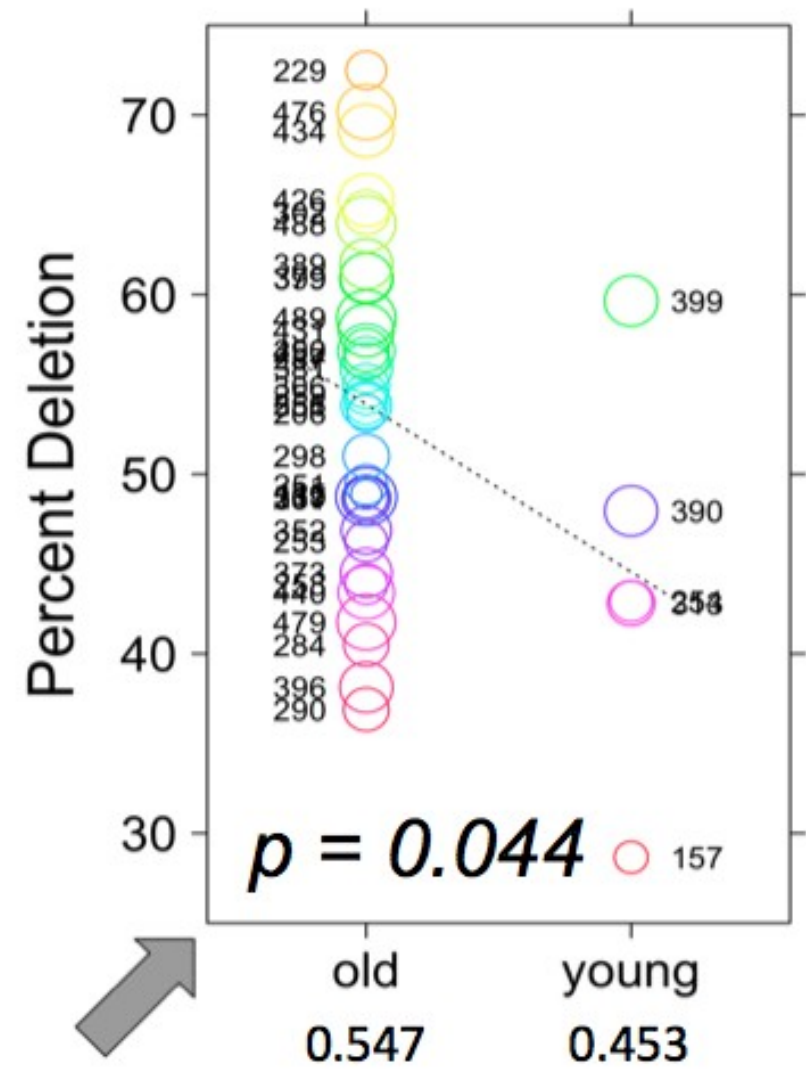
- Regardless of beliefs about individual variation, we expect that all other things equal, two data from the *same* individual will be more similar than two data from *different* individuals
- But sometimes the differences *between* individuals will swamp the regularities shared across the whole population...the result is not *generalizable* to the population
- We have accounted for more variance, but missed a generalization

Crop example, continued

- DEJ had a similar linguistic example, where an “apparent-time” effect on (td)-deletion actually *just* came from an older speaker who never deleted...



age w/ no random effect

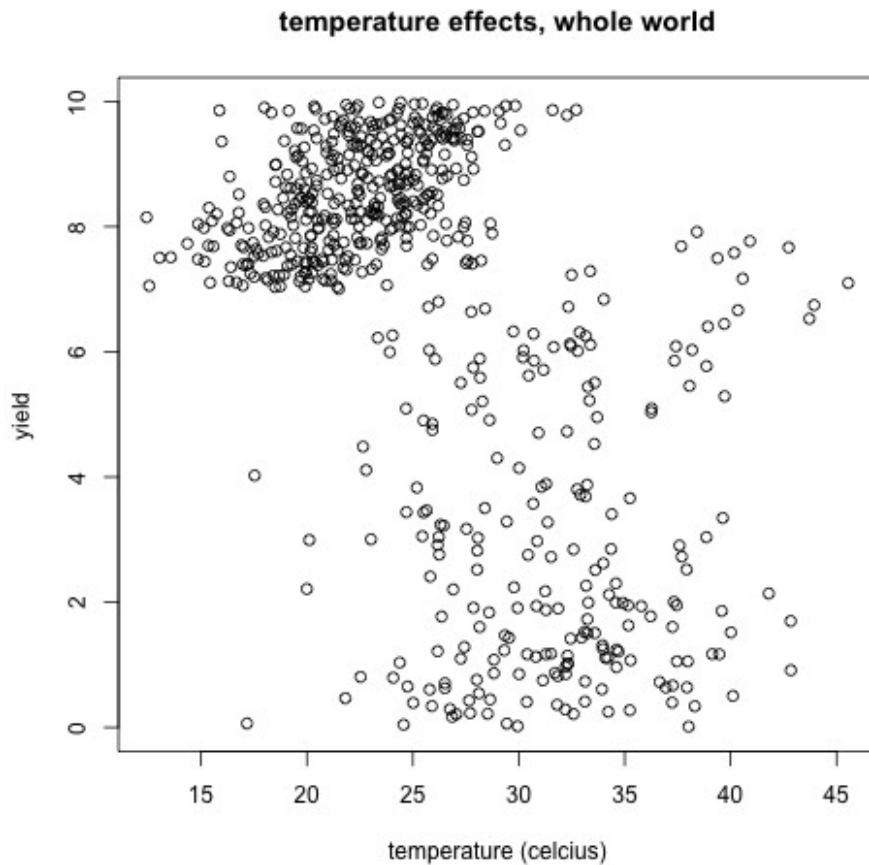


age + random intercept: speaker

Crop example, continued

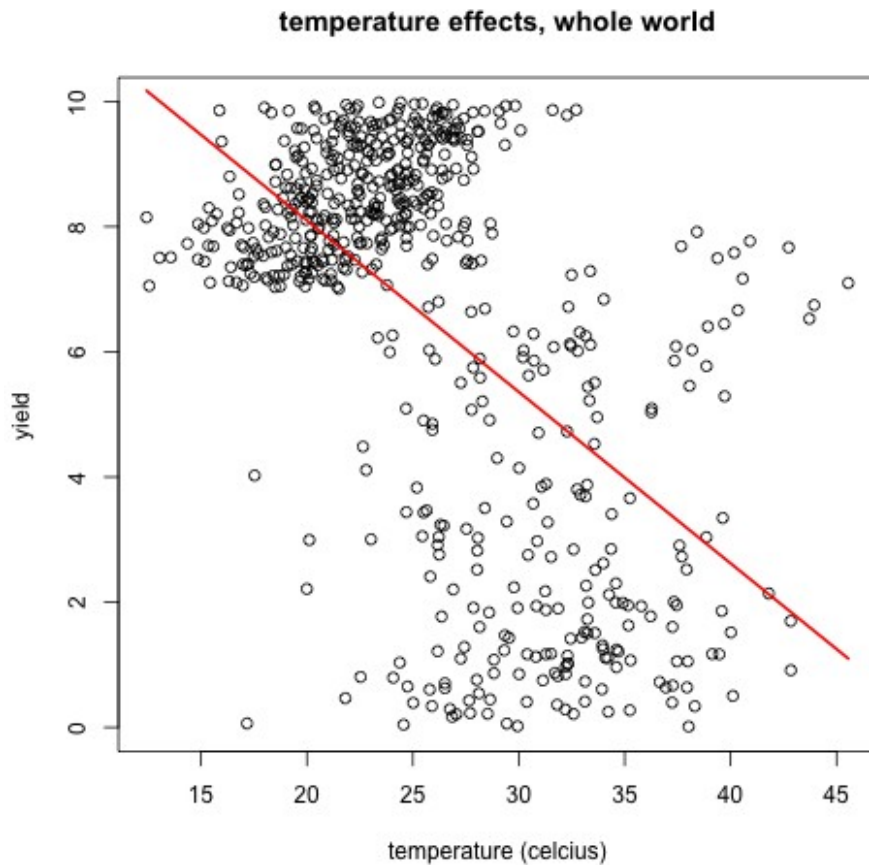
- Here's one where differences between different sources of data actually result in the *opposite* spurious conclusion
- Focusing on temperature, let's say these measurements are drawn from 25 different countries, which differ drastically in fertility and agricultural practices (e.g. fossil fuel inputs)
- Turns out those countries who use less fossil fuels tend to be near the equator...

Crop example, continued



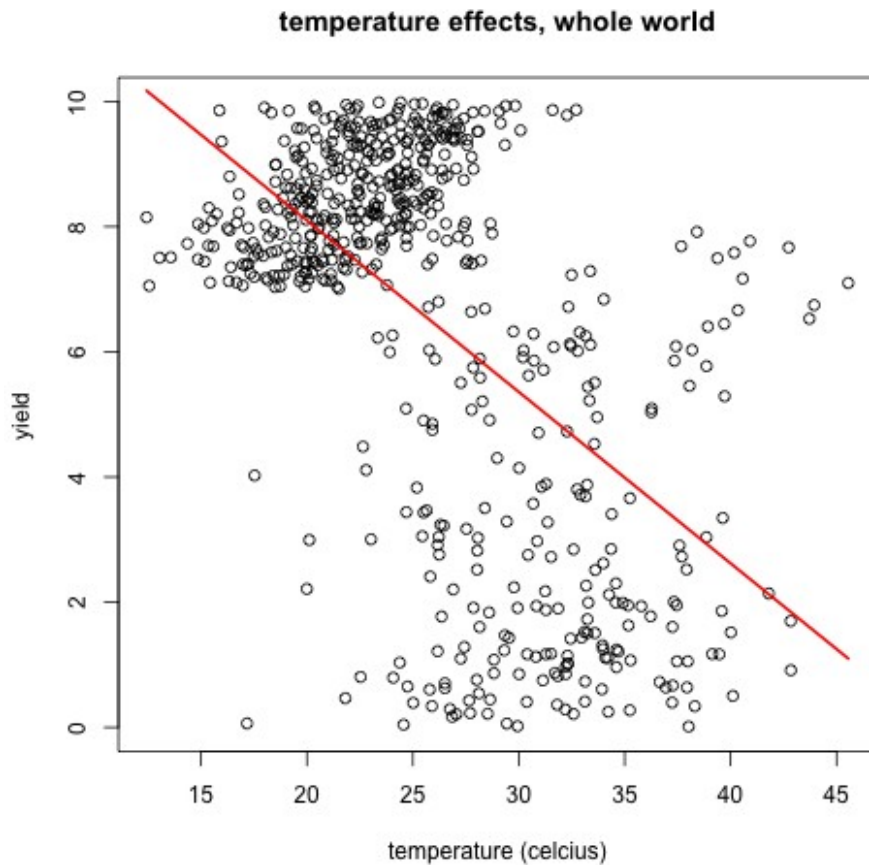
- When you run a regression, the effect of temperature is significant ($p < .001$)

Crop example, continued



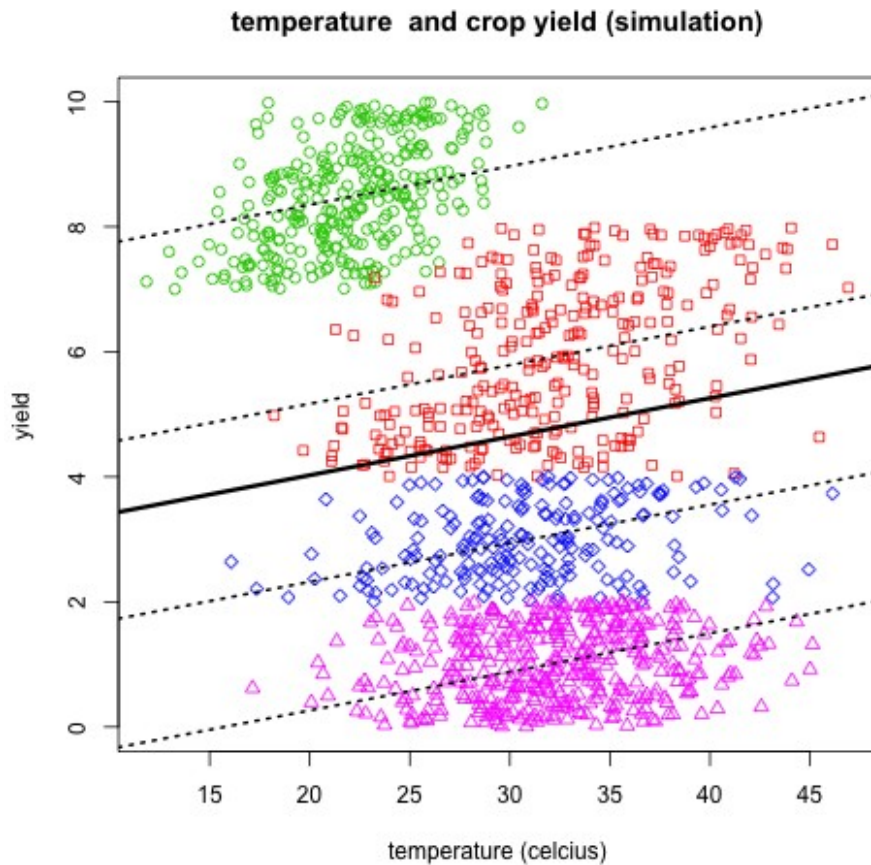
- When you run a regression, the effect of temperature is significant ($p < .001$)
- But, effect of temperature on yield is *negative*
- This isn't how I set up the data...

Crop example, continued



- Inside each group of data (1st world, 2nd world, etc.), there is a regular and positive effect of temperature
- But group differences have swamped it...

Crop example, continued



- We've missed the generalization common across all groups
- And in fact resulted in the opposite conclusion
- But if our model knew about the different groups...

Singularity and individuals

- Singularity: when one predictor uniquely determines the value of another
 - If you have a binary factor `Sex`, and a `Subject` factor group, then every time you know the name, you know the sex: e.g. “Celeste S.” implies “female”
 - Regression doesn't know which to choose!
- For this reason, including a factor group `Subject` as a main effect is infeasible
- Separate regressions for `Subject` and `Sex` undermine any claims of causality...

Random intercepts

- A function that maps group labels onto a points on a normal (Gaussian) distribution
 - With mean of zero
 - And empirically measured standard deviation
- An additive effect in a regression model
- “You do not encounter a singularity in solving for the conditional means of the random effects and the conditional estimates of the fixed effects because there is a penalty assigned to the size of the random effects vector.” – Douglas Bates [co-creator of HLMs – KG]

Per-subject random intercept

- How different is subject i from the (abstract) mean subject?
- Actually a *shrinkage estimate*, which is good:
 - no individual variation unless strong evidence to the contrary
 - regression towards the mean: “in replication studies with the same subjects, the extremely slow subjects will be faster, and the extremely fast subjects will be slower responders.” – Harald Baayen

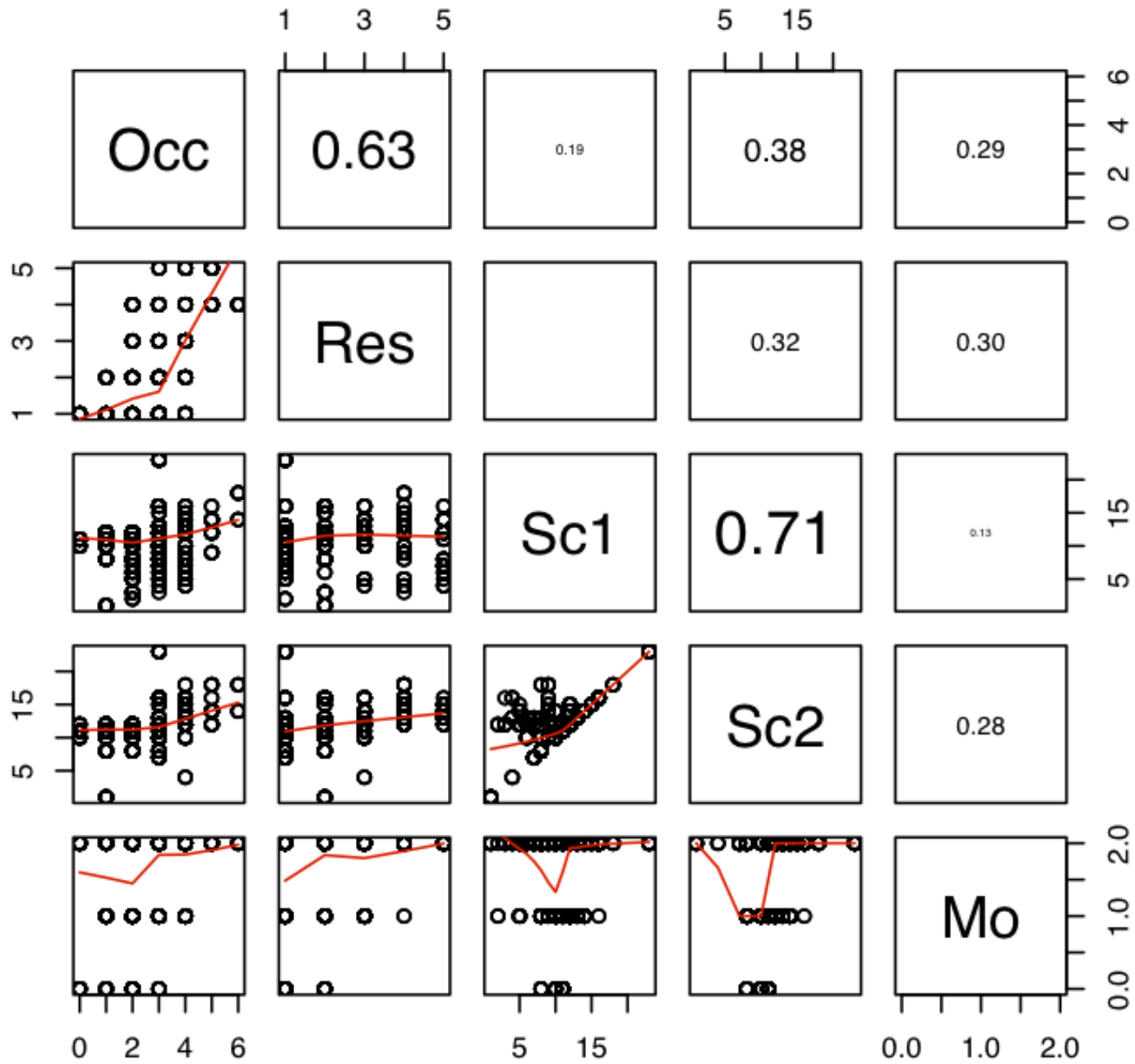
Back to (neg)

- There are some weird findings in *PLC2*
- Are we sure about the claims about SEC effects on the use of this stable variable? Some are rejected
- Particularly, some weird things with occupation

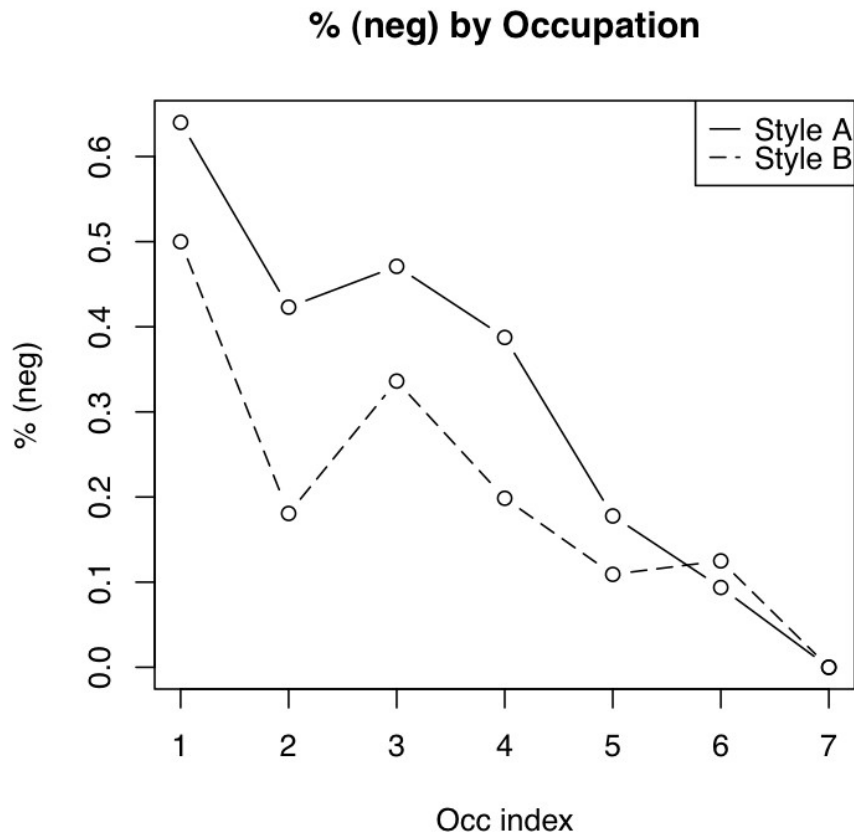
Socioeconomic measures in LCV

- Education of subject and of subject's father
- 7-point occupation scale
- Residential value
- Social mobility (up, down, stable) as rated by the fieldworkers
- High correlations...
 - Pearson's r as much as .71 (perfect corr. = 1)
 - k as much as 14 ($k > 10$ indicates strong corr.)
 - VIF = 1.4 (confidence interval inflated)

Socioeconomic pairs



Spurious occupation effects

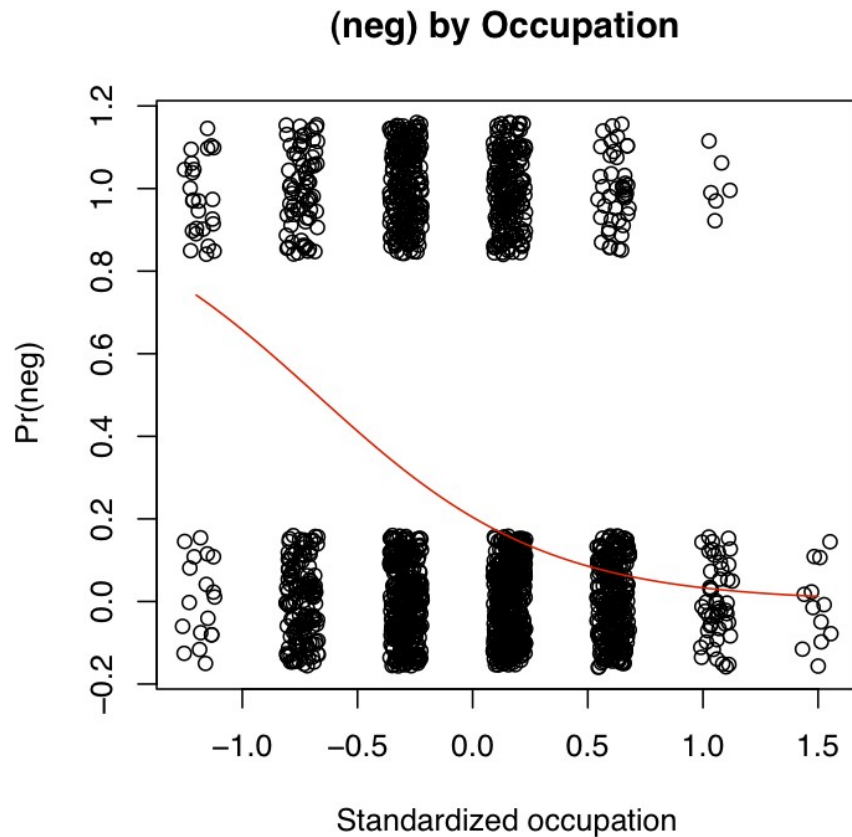


- Finds a *positive* effect for occupation in casual style, meaning higher probability of this *stigmatized* variable with higher-status occupations?
- ...but a *negative* effect for careful style, which makes more sense

So, I fit a new model

- I residualized all the SEC collinearities
- I also include sex, age (restricted cubic spline), and style predictors
- And a per-subject random intercept...
- ...which I hope will help me find the leaders of linguistic change (that's the stated goal of *PLC2* anyways...)

Hierarchical model findings



- Largest effect of occupation, increasing it by one s.d. decreases the $p(\text{neg})$ by half
- Large residence, sex, and style effects
- Age and mobility effects non-significant (bootstrap agrees)

Spurious rejection of S_{c2} effects

- Both education effects come out highly significant: more education means less (neg)
- Contra *PLC2*: “Significant or not, there was no single case in which the [S_{c2}] index was more highly correlated with the linguistic variable than the [S_{c1}] index, and none in which the [S_{c2}] 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.*” – William Labov [bracketed labels standardized for consistency; emphasis mine – KG]

Why this matters

- Some variance attributed to main effect predictors is now attributed to subject effects
- Some variance previously unaccounted for is now attributed to main effect predictors
- Asserts the study is replicable, assuming the next sample has the same subject-level variance

The leaders of linguistic change

- “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.” – William Labov (*SSENYC*, interstitial notes to 2nd ed., 2006)
- *PLC2* highlights individual speakers as leaders
- What is the relationship between leaders and subject-effect outliers?

Outlier analysis

- By-subject random intercept tested for significant outliers ($p = .05$); no data thrown out
- Oldest and most conservative Irish-American speaker, “Ed D.”, is a low-(neg) outlier
- 16-year-old “Barbara C.”, an upwardly-mobile anti-conformist who rejects the dominant racist ethos of her Fishtown neighborhood, well-connected opinion leader, is a high-(neg) outlier
- Middle-aged “Celeste S.”, now comfortably situated, is not a (neg) outlier, but is a vowel outlier

Other applications

- Studies of continuous variables, like (ay0)
- Laboratory studies:
 - F_1 (subject) effects w/ a random intercept
 - F_2 (item) effects w/ a random intercept
 - $F_1 \times F_2$ w/ a item-given-subject random slope
 - Better than F' (quasi- F) analysis for addressing the *language-as-fixed-effect* fallacy; Clark 1973); see Baayen et al. 2008

F_1 -analysis

- In a *between-subjects* design, there may be subject differences, but in a balanced design they're unlikely to affect the fixed-effect conclusion
- In a *within-subjects* design, the situation is much more precarious
- Standard treatment: *repeated measures ANOVA* with *error stratum* for fixed-effect-given-subject
- This quantity is F -distributed w/ d.f. = MS

F_2 -analysis

- If you repeat stimuli (particularly in what you might call a *within-treatment* design), you may have the *language-as-fixed-effect* fallacy, which also needs to be accounted for in the same way
- But subjects and items can interact: John doesn't know what a *deebo* is, but Mary does
- Compute a quantity called F' , which is ugly

Quasi- F'

- In practice, F' is too big and can't be computed
- Clark (1973) suggests $\min F'$, which is a conservative approximation from averaging separate subject and item analyses
- Thanks to simulations, we know it's a mess (Baayen, 2008)
- No regression towards the mean ;(
- What to do with more complicated designs?

Experimental design w/ ran-effects

- Subject-grouped effects:
 - are there multiple observations from each subject? (`1 | Subject`)
 - Are there stimuli presented multiple times for each subject? (`Item | Subject`)
 - Are subjects exposed to multiple conditions? (`Condition | Subject`)
- Item-grouped effects: are there stimuli presented multiple times? (`1 | Item`)
 - Are there stimuli presented in multiple conditions? (`Condition | Item`)

Quam and Swingley (forthcoming)

- Turns out there are significant subject effects
- Condition-given-subject effects are interesting too
- Item effects are miniscule, so they're not analyzed
- Much-improved model fit

For more info...

- Gelman and Hill (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: CUP.
- Previous use of HLMs in sociolinguistics:
 - Jaeger and Staum (2005), NWAU talk
 - Johnson (2007), U. Penn dissertation
- <http://www.r-project.org/>
- <http://ling.upenn.edu/~kgorman/mlm.html>

This site includes two reports, one short and non-technical, and one long and heavy on R code...

Acknowledgements

- William Labov and collaborators for the data
- Members of the Penn MLM Reading Group in Autumn 2008, and audiences at SPLUNCH
- John Trueswell, Mark Liberman and Steve Isard
- *Author was funded by an NSF-IGERT training grant and a University of Pennsylvania Ben Franklin Scholarship*