LING82100: homework 5

(Due 4/22)

1 Multiple regression

The *Atlas of North American English* (Labov et al. 2006) analyzed vowel productions by over 400 native English speakers using a phone survey. The file `ANAE-O.tsv`[1] contains vowel measures—F1, an acoustic correlate of vowel "height", and F2, an acoustic correlate of vowel "frontness"—for the (o) and (oh) vowels (as in *cot* and *caught*, respectively), aggregated by speaker. While these two vowels are historically distinct, many English speakers do not produce a clear contrast. The data also includes speaker's `age` in years, self-reported `gender`, and `dialect` region. You will examine the effects of age, gender, and dialect region on degree of merger.

1.1 What to turn in

1. First, compute the dependent variable, the *Euclidean distance* between the (o) and (oh) vowels in F1 × F2 space, and add it back into the data frame. Here the distance is given by

$$d = \sqrt{(\mathrm{F1}_o - \mathrm{F1}_{oh})^2 + (\mathrm{F2}_o - \mathrm{F2}_{oh})^2}$$

   and is just the "straight line" distance between the two points; you may also recognize it as a special case of the Pythagorean formula.

2. Fit a linear regression to the distance d using

   (a) age (centered),

   (b) gender (sum-coded), and

   (c) dialect (sum-coded)

   as independent variables.

3. Report the statistical results for your model. Your report should include

   (a) the coefficient and results of the likelihood-ratio hypothesis test for age,

   (b) whether merger increases or decreases with age (if significant),

   (c) the coefficient and results of the likelihood-ratio hypothesis test for gender,

   (d) whether men or women are more merged (if significant),

[1] `http://www.wellformedness.com/courses/LING82100/Data/ANAE-O.tsv`

(e) the results of the likelihood-ratio hypothesis test for dialect,

(f) the Tukey HSD results for dialect.

4. What dialect is least merged? How do you know?

1.2 Hints

- There are many ways to compute the Euclidean distance but probably the easiest is to just write it out in R roughly as it appears in the equation.
- You do not need to report the coefficients for dialect; there are just too many.
- Do not forget to apply sum coding to the appropriate variable(s).
- Do not forget to center the appropriate variable(s).
- To answer the last question, compute estimated means for each dialect.

1.3 Stretch goals

- Instead of centering age, standardize it, then repeat the experiment. What if anything changes?
- Research an alternative distance metric called *Mahalanobis distance*. Then compute the distance between (o) and (oh) using the using the `mahalanobis` function, and repeat the experiment with this as your new dependent variable. What if anything changes?

2 Multicollinearity

Labov (2001: chap. 3) studies the use of negative concord (e.g., "I didn't tell John to paint *none* of these") in a survey of 155 white speakers in the Philadelphia metropolitan area, using data collected by the *Language Change and Variation* (LCV) study (1973–77). The dependent variable is binomial: use of negative concord vs. the use of the negative polarity (e.g., *any* instead of *none*) under the scope of sentential negation. Independent variables include age in years, speech style, self-reported gender, and four measures of socioeconomic status:

- `occ`: a 7-point scale of occupational prestige adapted from a prior sociological survey,
- `res`: property value of the family, rounded to the nearest $1,000 USD,
- `sc1`: number of years of education (ranges from 1–23), and
- `sc2`: number of years of father's education (ranges from 1–23).

Following Labov, we assume that all four are essentially interval measures, and that all four measures are measuring a single construct: social class. The file `LCV-SES.tsv`[2] contains (simulated, but realistic) socioeconomic measures for the speakers interviewed. You will examine, and then correct, multicollinearity among these four socioeconomic variables.

[2] `http://www.wellformedness.com/courses/LING82100/Data/LCV-SES.tsv`

2.1 What to turn in

1. For each of the four variables, measure its (Pearson r) correlation between each (unordered) pair of the four variables. Do these constitute non-trivial multicollinearity?
2. Standardize all four social class measures and prove standardization was successful.
3. Using the standardized measures, iteratively residualize all variables.
4. Prove residualization was successful.

Make sure to turn in both code and results.

2.2 Hints

- You don't need to report p-values.
- You can use `cor` here; `cor.test` is overkill because we're not doing hypothesis testing.
- Correlation is symmetrical; i.e., $cor(x, y) = cor(y, x)$ for all x and y.
- Ask yourself: if some variable is standardized, what properties would you expect it to have?
- Just pick one of the four variables as your baseline for residualization; I used `occ`.
- Ask yourself: if some set of variables have been residualized, what do we expect their Pearson correlation to be?
- Don't lose your notes and implementation here; we'll return to this problem in a few weeks.

2.3 Stretch goals

- As much as possible, automate the previous steps as much as possible, so you don't have to copy and paste code, using `for`-loops and similar constructs.
- Use *variable inflation factor* to measure the overall multicollinearity between the four socioeconomic status measures.
- Visualize multicollinearity before standardization and residualization.

For both of these, you may have to do some research.

References

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

Labov, William, Sharon Ash, and Charles Boberg. 2006. *The Atlas of North American English: Phonetics, Phonology and Sound Change.* New York: Mouton de Gruyter.