Comparing apples and oranges: Weights of factors with different numbers of levels

Vsevolod Kapatsinski, Irina Shport & Susan Guion-Anderson
vkapatsi@uoregon.edu, ishport@uoregon.edu, guion@uoregon.edu
University of Oregon

The problem and the solution

- Two possible cues to a phonological contrast
- More generally, two independent variables with a binary dependent variable (multiple logistic regression)
- We want to know the weights of cues relative to each other
  - E.g., Escudero et al. (2009), Morrison (2005a): spectrum vs. duration as cues to vowel contrast in Dutch
  - Shport (2011): F0 peak location vs. F0 fall as cues to pitch accent in Japanese
- More generally, relative importance of predictors
  - E.g., Kapatsinski (2009): conjunct constituent type vs. register as predictors of conjunction choice in Russian

We need to know the weights of cues relative to each other and whether predictors have different numbers of levels.

- The cues have different numbers of levels
- Predictors have different numbers of levels
- What we show:
  - State-of-the-art methods (varbrule range, coefficient comparisons) are biased if the predictors are nominal and have different numbers of levels (cf. also Morrison 2005)
  - We can estimate how biased they are
  - And therefore how much of a difference in cue/factor weight is attributable to differences in numbers of levels

Example: Shport (2011)

1. F0 = 100% syllable 1 prominence responses
2. 0.5 = No reliance on the cue
3. 0 = 100% syllable 2 prominence responses

Which cue do the listeners rely on more? F0 peak location OR F0 fall magnitude?

General algorithm (with R code excerpts)

numSpeakers = Number of Speakers, same as actual study nRep = Number of Replications–same large number nVal1 = Number of values of first factor nVal2 = Number of values of second factor nPerCell = Number of stimuli per subject with the same values on factors 1 and 2

for (i in 1:nRep) #for each replication
  # the first predictor
  fakeIV1 = factor(rep(0:nVal1-1,each=nPerCell*subj*nVal1))

  # the second predictor
  fakeIV2 = factor(rep(0:nVal2-1,each=nPerCell*subj*nVal2)) #sample in stimuli)

for (j in 1:numSpeakers) #for each speaker in the simulated sample
  # Get a sample of nSims random 1’s and 0’s: a simulated subject
  # Since zeros and ones are chosen at random, this is a subject
  subj1 = sample(nVal1, nVal1, replace = TRUE)

  # who is drawn from a population where the effects of both predictors
  # on the dependent variable are equal to zero
  p = rnorm(n=1, mean=0, sd=1)

  # Extract factor weights from the results of the regression
  factor.weights = coef(fakeIV1 + fakeIV2, family = binomial(link = "logit",
                                                                   data = fake.sim))

  # Estimate factor weights from the results of the regression
  # The factor weights are handy because they scale between 0 and 1
  # As a result we get how important the predictors are for this individual subject
  # On average, all should be around 0.5 because the predictors have no effect
  # on the population from which subjects are drawn
  factor.weights = as.factor(fakeIV1 + fakeIV2, family = binomial(link = "logit",
                                                                          data = fake.sim))

  # For each factor group, we get the difference between highest factor
  # and highest factor weight (range)
  # This is the standard error of estimating predictor importance in quantitative
  # sociolinguistics (e.g., Tagliamonte 2006)

  # We compare the within-subject importance measures for predictor 1
  # vs. within-subject importance measures for predictor 2 using a t-test
  if (t.test(SubjCoeff[1], SubjCoeff[2], var.equal = TRUE))

  # We can see now if the value we obtained in actual study is within the range of the simulated 1’s. If it is not, we can say that the factors differ in importance at the 0.1Rep level.

For full code, see pages.uoregon.edu/vkapatsi/CueWeightScript.R

Two nominal predictors

Individual subjects:

- 1000 simulated subjects, nVal1 = 3, nVal2 = 8
- Result: Predictor with more levels tends to have a larger coefficient
- We know how much larger:
  - Range of factor group with 3 levels must be greater than range of the factor group with 8 levels to conclude that the factor group with 3 levels is more important.
  - Range of factor group with 8 levels must exceed the range of the factor group with 3 levels by 0.35 to conclude that the factor group with 8 levels is more important.

Note: bias increases if subjects favor one value of the DV over another; The graph below is for subjects who on average do not have a bias in favor of one value of the DV.

Two numerical / ordinal predictors

Individual subjects:

- 1000 simulated subjects, nVal1 = 3 (0, 1, 2), nVal2 = 8 (0, 1, 2, 3, 4, 5, 6, 7)
- Result: Very little bias

Two numerical predictors with same number of levels but where there are equal numbers of stimuli having each value on predictor 1 but not on predictor 2:

- Predictor 1: 10 stimuli with each value
- Predictor 2: 5 stimuli with one value, 15 stimuli with other value
- Result: No observable bias

Conclusion

- Numerical predictors with different numbers of levels, scales → not to worry, R will take care of it.
- Categorical/nominal predictors → range biased in favor of predictor with more levels
- We can estimate how biased it is for a particular dataset (need to know whether subjects are biased in favor of a response)
- Therefore, we can test whether the observed difference in range between predictors could be due to differences in numbers of levels.

So, we can still compare them! Yay!

References