Joseph Roy & Yukiko Yoshizumi
University of Ottawa
CVC 2008-Ottawa
Extended Handout

Extending Variable Rule Analysis

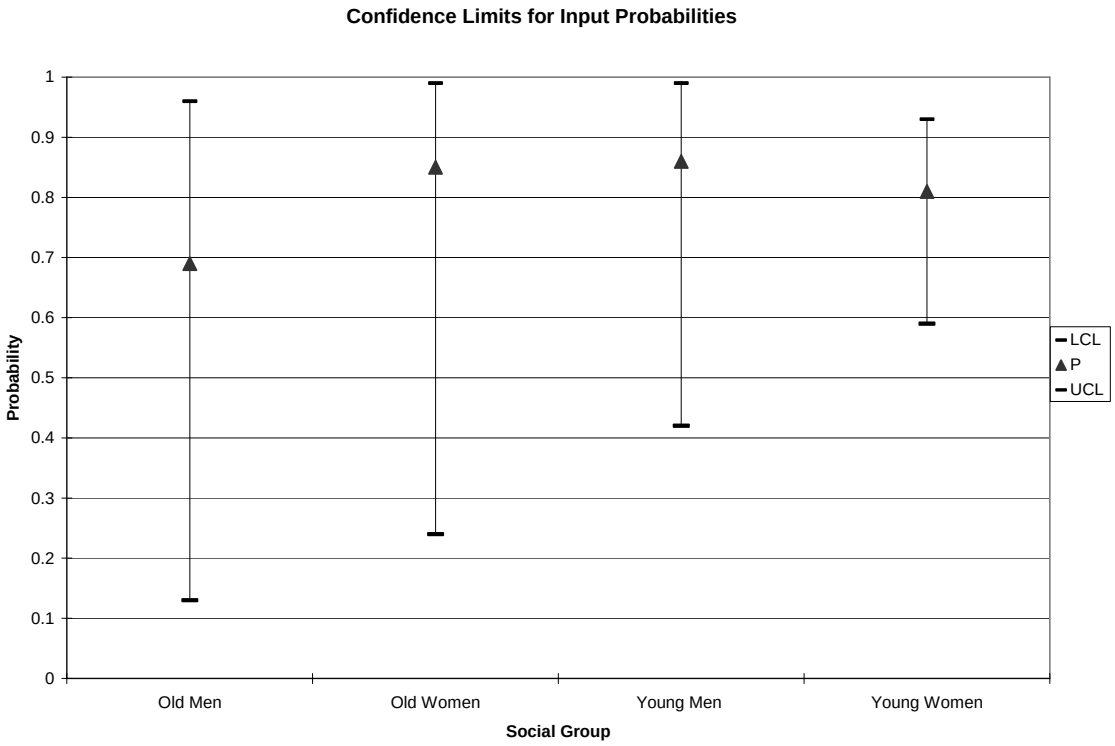

Table 1: Goldvarb Results for the selection of HAVE in stative possessives

	Old						Young					
	Women			Men			Women			Men		
	P	%	N	P	%	N	P	%	N	P	%	N
Corrected Mean	.86	84	381	.79	78	218	.79	77	368	.78	74	447
Polarity												
Negative	.23	63	64	[]	68	37	KO	100	138	.86	95	65
Affirmative	.56	88	317	[]	80	184	[]	77	368	.42	71	382
Range	33											
Subject Type												
NP	.83	97	35	[]	83	23	.76	92	38	[]	81	52
Pronoun	.46	83	346	[]	77	195	.47	75	327	[]	73	388
Range	37											
Subject Reference												
Generic	[]	91	45	.26	58	40	.34	63	67	.63	83	357
Specific	[]	83	336	.56	83	179	.54	80	298	.47	72	83
Range				30								
Object Type												
Abstract	[]	82	132	[]	77	113	[]	80	198	[]	75	263
Concrete	[]	85	244	[]	80	108	[]	73	167	[]	73	184
Range												

í

Table 2: Logistic Regression Results with Upper and Lower Confidence Limits for the factor weights with statistically significant factors shaded in with light gray.

	Old						Young					
	Women			Men			Women			Men		
	P	LCL	UCL	P	LCL	UCL	P	LCL	UCL	P	LCL	UCL
Corrected Mean	.85	.24	.99	.69	.13	.96	.81	.59	.93	.86	.42	.99
Polarity												
Negative	.44	.39	.50	.46	.54	.50				.60	.81	.50
Affirmative	.58	.55	.50	.59	.51	.51				.40	.19	.50
Range	15	8	0	6	2	1				20	62	0
Subject Type												
NP	.58	.71	.50	.52	.56	.49	.55	.65	.51	.53	.58	.50
Pronoun	.41	.29	.50	.46	.42	.51	.44	.34	.49	.47	.42	.50
Range	17	43	0	6	14	2	11	31	2	5	16	0
Subject Reference												
Generic	.52	.59	.50	.36	.47	.47	.48	.50	.49	.53	.58	.50
Specific	.47	.39	.50	.62	.52	.52	.51	.48	.51	.46	.42	.50
Range	5	21	0	25	5	5	4	2	2	6	16	0
Object Type												
Abstract	.50	.44	.50	.50	.53	.48	.49	.51	.49	.50	.52	.50
Concrete	.50	.54	.50	.48	.45	.52	.50	.47	.51	.50	.48	.50
Range	0	9	0	2	8	4	1	4	2	0	4	0

R Statistical Program: http://www.r-project.org/

R code used in the analysis of QEC Data:

```
#setwd is the command to tell R where to look for your data files.
setwd("C:\\Documents and Settings\\Joseph Roy\\Desktop\\Research\\Stative Possessives
VARBRUL")


qep.data <- read.csv(file="qep.data.csv", header=T)

#R's version of recoding
qep.data$polarity[qep.data$polarity != "A"] <- 1
qep.data$subject[qep.data$subject == "Z"] <- NA
qep.data$reference[qep.data$reference == "Z"] <- NA
qep.data$object[qep.data$object == "Z"] <- NA

# QEC Data

qep.old <- qep.data[qep.data$age == "O",]
qep.oldmen <- qep.data[qep.data$age == "O" & qep.data$sex == "M",]
qep.oldwomen <-qep.data[qep.data$age == "O" & qep.data$sex == "F",]

qep.young <- qep.data[qep.data$age == "Y",]
qep.youngmen <- qep.data[qep.data$age == "Y" & qep.data$sex == "M",]
qep.youngwomen <- qep.data[qep.data$age == "O" & qep.data$sex == "F",]

# Young Women have a KO for Negative Contexts

qep.youngwomen <- qep.data[qep.data$polarity != 1,]

#Logistic Regression is performed by glm procedure (and binomial family--> the logit
link is the canonical (default and most interpretable link).
qep.om <- glm(variant ~ subject + reference + object +
polarity,binomial,data=qep.oldmen, na.exclude)
qep.ow <- glm(variant ~ subject + reference + object +
polarity,binomial,data=qep.oldwomen)

qep.ym <- glm(variant ~ subject + reference + object +
polarity,binomial,data=qep.youngmen)
qep.yw <- glm(variant ~ subject + reference + object,binomial,data=qep.youngwomen)
```

#Print out the Results
summary(qep.om)
summary(qep.ow)
summary(qep.ym)
summary(qep.yw)

How we generated the VRA results for the logistic regression and the upper and lower confidence intervals:

(1) From the logistic regression, we held each factor fixed and iterated through all other possible combinations of factor levels to generate all of the probabilities for that factor level:

	Intercept	Subj Pron	Refer Spec	Polarity Aff	Object C				Average	Factor Weight
	1.17	-1.33	-0.11	0.00	-0.18					
Neg	0.39	0.71	0.74	0.42	0.76	0.43	0.74	0.73	0.62	0.53

(2) I then averaged through each factor level, and centered the averages.
(3) I followed the same procedure for the set of upper and lower confint intervals.

References by Category

General Categorical Statistics

Agresti, Alan. 2002. *Categorical Data Analysis*, 2nd Edition. Hoboken, New Jersey: John Wiley & Sons.

Hosmer, David & Stanley Lemeshow. *Applied Logistic Regression*, 2nd Edition. Hoboken, New Jersey: John Wiley & Sons.

Hardin, James W. & Joseph M. Hilbe. 2001. *Generalized Linear Models and Extensions*. College Station, Texas: Stata Press.

R Programming Language

Manning, Christopher. 2007. *Logistic Regression (with R).* ms. Downloaded: 2 June 2008. (http://nlp.stanford.edu/~manning/courses/ling289/logistic.pdf)

Vernables, W.N. and B.D. Ripley. 2002. Modern Applied Statistics with S. 4th Edition. New York: Springer. [While this book has "S" in the title, it is applicable to R—S is the commercial form of R.]

Quantitative Linguistics

Bod, Rens, Jennifer Hay, and Stefanie Jannedy (eds). 2003. *Probabilistic Linguistics*. Cambridge, Mass: MIT Press.

Bayley, Robert. 2002. *The Quantitative Paradigm*. In the Handbook of Language Variation and Change.

Labov, William. 2001. *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

2005. Quantitative Reasoning in Linguistics. In Ulrich Ammon, Norbert Dittmarr, Klaus J. Mattheier and Peter Trudgill (Eds) *An International Handbook of the Science of Language and Societ*y, Vol. 2. New York: Mouton de Gruyter.

Cedegren, Henrietta and David Sankoff. 1974. Variable Rules: Performance as a Statistical Reflection of Competence. *Language*, 50(2):333-355.

Godfrey, Elizabeth and Sali Tagliamonte. 1999. Another piece for the verbal –s story: Evidence from Devon in southwest England. *Language Variation and Change*, 11:87-121.

Guy, Gregory R. 1988. Advanced Varbrul Analysis. In Kathleen Farrara, Becky Brown, Keith and John Baugh (eds), *Linguistic Change and Contact*. Pp. 124-136.

Mendoza-Denton, Norma, Jennifer Hay, and Stefanie Jannedy. 2003. Probabilistic sociolinguistics: Beyond variable rules. In Bod, Hay, Jannedy (2003).

Paolillo, John. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*. Stanford: Center for the Study of Language and Information Publications.

Sigley, Robert. 2003. The importance of interaction effects. *Language Variation and Change*, 15: 227-253.

Sankoff, David (ed.). 1978. *Linguistic Variation: Models and Methods*. New York: Academic Press.

Saito, Hidetoshi. 1999. Dependence and interaction in frequency data analysis in SLA research. Studies *in Second Language Acquisition*, 21:453-75.

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. New York: Cambridge University Press.

Young, Richard, and Brian Yandell. 1999. Top-down versus bottom up analysis. *Studies in Second Language Acquisition*, 21:477-88.

Advanced Categorical Data Analysis

Agresti, Alan, James G. Booth, James P. Hobert, & Brian Caffo. 2000. Random Effects Modeling of Categorical Data Response. *Sociological Methodology*, 30:27-80.

Carey, Vincent, Scott L. Zeger, and Peter Diggle. 1993. Modeling multivariate binary data with alternating logistic regressions. *Biometrika*, 80: 517-26.

Hardin, James W. & Joseph M. Hilbe 2002. *Generalized Estimation Equations*. New York: Chapman & Hall/CRC Press.

Horton, Nicholas J., Judith D. Bebchuk, Cheryl L. Jones, Stuart R. Lipsitz, Paul J. Catalano, Gwedolyn E. P. Zahner and Garret M. Fitmaurice. 1999. Goodness-of-fit for GEE: An example with mental health service utilization. *Statistics in Medicine*, 18:213-22.

Liang, Kung-Yee & Scott L. Zeger. 1986. Longitudinal Data Analysis using generalized linear models. *Biometrika*, 73:13-22.

Molenberghs, Geert and Geert Verbeke. 2005. *Models for Discrete Longitudinal Data*. New York, NY: Springer.

Segal, Mark Robert. 1992. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, 87:407-18.