

Homework #5 (the last one!)

Handed out Wed Feb 16
Due Mon Feb 28

This problem makes reference to files in

<http://www.stanford.edu/class/linguist235/sample-final-project/>.

1. One potential aspect of parallelism in coordinate NPs involves the genitive alternation (e.g., *France's president* vs. *the president of France*). The files

genitive-trees-brown
genitive-trees-wsj

contain binary coordinate NPs where each conjunct contains either a possessive pre-modifier or an *of* PP post-modifier. The trees are shown in Penn Treebank (Lisp S-expression) notation: for linguists, this is essentially the same notation as is often used with square brackets to represent tree structure compactly on one line. The part-of-speech of words is also shown similarly.

- (a) Show the contingency tables of possessive modification for Brown and Wall Street Journal coordinate NPs. The file is divided into sections for each combination of pre- and post-modification. One way to partially automate this is to cut the file into separate files for each section by hand and then to find the length of the four sections with the Unix `wc` program (or many other things can tell you the number of lines in a file).
- (b) Test the null hypothesis of independent possessive realization for each corpus.
- (c) Rosenbach (2002) has shown that length of the possessive modifier plays a significant role in determining whether the possessive is realized prenominal or postnominal. This means that realization of possessives in coordinate NPs could potentially be an artifact of correlated modifier length. The file

possessive-matrix

contains entries of the form

left-realization right-realization left-mod-length right-mod-length

for coordinate NPs in the the Wall Street Journal. “pre” in this matrix means prenominal modification, and “post” means postnominal modification. Possessive modifier length is measured in number of words.

Design a logistic regression model of possessive realization in coordinate NPs that incorporates the effect of possessor length. The general model, in which the realization of conjunct NP possessors is not independent, is a case of *multinomial* logistic regression, where there are more than two outcomes (in our case, there are four outcomes – pre/pre, pre/post, post/pre, post/post). SPSS has an interface for multinomial logistic regression right in the Analyze | Regression menu, just under Binary Logistic. In the general model, assume that left and right conjunct length have independent effects on the likelihood of possible outcomes. What is the data likelihood for this model?

- (d) Design an alternative model in which possession is realized independently in left and right conjuncts, and the only independent variable for a conjunct’s possession type is the weight of that possessor. What is the data likelihood for this model? (Hint: to get the data likelihood for all the data of the model, you will have to do two separate binary logistic regressions, one for left conjuncts and one for right conjuncts, and combine their likelihoods.)
- (e) The general model contains the independence model; how many free parameters does each have? Test the independence model against the general model using a G^2 test. Can we reject the null hypothesis that possessor realization is independent between conjuncts? (Remember, the G^2 test relies on the fact that for models $M_0 \subset M_1$, $-2 \ln(L_1 - L_0)$ is chi-square distributed, where L_i is the likelihood of the data under M_0 .)

Appendix on using SPSS for multinomial logistic regression and data likelihood.

The relevant menu selection is Analyze | Regression | Multinomial Logistic. This interface differs somewhat from the Binary Logistic interface. Within the Multinomial Logistic interface, you want to move the dependent variable (the response) into the “Dependent” box. You then want to move your nominal independent variables into the “Factor(s)” box, and your continuous independent variables into the “Covariates” box.

Unfortunately, the output for Multinomial Logistic doesn’t seem to give a simple way of finding the data likelihood for your fitted model. We suggest the following way of getting this information. In the Multinomial Logistic interface, click on “Save”, and in the new interface check the “Predicted category probability” box. Then click Continue and, back at the main interface, click OK. The Data Editor will now have a new variable for each case, which is the predicted probability for the observed response value.

To turn this into a log likelihood, use the menu selection Transform | Compute; under Function group select “Arithmetic” and then “Ln” under Function, and then stick the predicted probability variable name inside LN in the Numeric Expression box. The log likelihood of the entire dataset is then the sum of the log likelihoods for each case.¹ You can calculate the sum under Analyze | Descriptive Statistics | Descriptives; move your log probability variable name into the “Variable(s)” box, click on Options, and check on Sum.

Once you have a G^2 statistic comparing the joint and independence models, you can look up the significance value as follows. Go to Transform | Compute; under Function

¹As usual, we ignore the irrelevant combinatorial constant part of the likelihood.

group choose “CDF & Noncentral CDF”, and under Functions and Special Variables select “Cdf.Chisq”. Enter your G^2 value and degrees of freedom as the two arguments of CDF.CHISQ under the Numeric Expression box, name this new variable, and click OK. The resulting value is $1-p$ (where p is the significance). Unfortunately, the result comes up in the Data Editor, but that’s the best we’ve got so far!

Finally, in order to do the joint realization model you will need to define a new variable that is a combination of left and right conjunct possessor types. The easiest way to do that is probably through the Transform | Compute interface; under Functions select String and use the Concat function. This allows you to create a new joint variable with variables “prepre”, “prepost”, “postpre”, “postpost”.

(An alternative to the latter part of the instructions is that once you’ve gotten the individual likelihoods, you can copy the column to Excel and work there. You can do the G^2 test using the CHIDIST function (under Statistical) in Excel. You cannot do logistic regressions in Excel, though....)