

Ling 235 – sample final project

Here is a sample problem set, based on the problem of parallelism in coordinate noun phrases. The problems associated with the Cedergren data from problem set #1 are also good examples for the final project.

1. In the online directory <http://www.stanford.edu/class/linguist235/sample-final-project/>, the following files contains lists of binary coordinate NPs from the Brown corpus with PP modifiers in both, only left, only right, and neither conjunct:

- both-conj-brown
- left-conj-brown
- neither-conj-brown
- right-conj-brown

Express this set of coordinate NPs as a 2×2 contingency table. Calculate the odds ratio for PP modification. Test the null hypothesis of independence between left and right PP modification using X^2 , G^2 , and Fisher's exact tests. Do these tests lead to qualitatively similar conclusions? Is this surprising?

2. The following files contain similar lists for the parsed Wall Street Journal corpus:

- both-conj-wsj
- left-conj-wsj
- neither-conj-wsj
- right-conj-wsj

Quantitatively compare the Brown corpus and the Wall Street Journal in terms of the strength of association for PP modification of conjuncts. Which corpus shows a stronger parallelism effect?

3. One of the problems with this contingency-table view of coordinate NPs is that it conflates speaker choice of NP *contents* with choice of NP *order*. If the choice of NP order is made contingent on NP contents, then tests for statistical independence between NP1 and NP2 may reject the null hypothesis even if there is no parallelism effect.

For the following questions, assume a model of coordinate NP generation as follows:

- I. First, the speaker chooses to generate a coordinate NP.

II. Next, the speaker chooses an unordered NP pair $\{NP_a, NP_b\}$.

III. Finally, the speaker chooses a linear order for the NP pair.

- (a) Qualitatively describe what would constitute a parallelism effect in this model. In which step of NP generation would parallelism be relevant?
 - (b) Construct a hypothetical 2x2 contingency table for PP modification in which there is no parallelism effect, but Fisher's exact test rejects the null hypothesis of independence of PP modification for NP1 and NP2.
 - (c) What linguistic phenomena could give rise to such a contingency table pattern?
 - (d) Construct a statistical test for parallelism that is not confounded by linear order choice effects. Define the null and alternative hypotheses, describe how to estimate its parameters, and calculate a p -value. (Hint: no need to construct an exact test in this case; either the X^2 or likelihood-ratio test is simplest.) Apply this test to the data for the contingency table you constructed, and to the Brown and WSJ data for PP modification.
4. Another 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.
- (a) Show the contingency tables of possessive modification for Brown and Wall Street Journal coordinate NPs, and test the null hypothesis of independent possessive realization for each corpus.
 - (b) Rosenbach (2002) has shown that length of the possessive modifier plays a significant role in determining whether the possessive is realized pre-nominally or post-nominally. 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.

Calculate the mean left and right modifier lengths, and use a t -test to test whether left and right modifiers are similarly distributed in length. Are they? Explain your results.

- (c) Calculate the correlation coefficient r between left-modifier and right-modifier lengths. Can we safely reject the null hypothesis that $r = 0$ (i.e., modifier lengths are uncorrelated)?
- (d) 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 – 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? 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 overall data likelihood models, 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: 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

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

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 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”.