Ling 236 Homework #5
Due Feb 20, 2002

1 and only question.

Here's some linguistic data where one could not too unreasonably hope to apply a multiple linear regression analysis – for a phonology problem no less! (Edward Flemming rightly chided me in an early lecture for casually saying that there's no use for linear regression in linguistics….) The direct source of the data and description is Naomi Nagy. This problem illustrates how categorical variables (of the kind used in Labov 1969, or from ANOVA studies as in Keller 2000) can be 'coded up' in a fashion that allows them to be used in a standard multiple linear regression analysis. This assignment can definitely be completed using either WebStat 2.0 or Excel (providing you have the 'Data Analysis' pack installed, or can get hold of an Office CD). A link to WebStat 2.0 can be found off the course website:

http://nlp.stanford.edu/~manning/courses/ling236/handouts/ling236-statnotes5.html

Kansas City is located in an area on the borderline between the South Midlands dialect, characteristic of the rural areas located to the south of it, and the Northern Cities dialect, characteristic of cities like Buffalo, Detroit and Chicago. In the Northern Cities, /æ/ (low front vowel 'ash') is rising (toward [e]) (mid front tense vowel) as part of the Northern Cities chain shift (Labov 1994, etc.).

In 1980 Melanie Lust studied the distribution of the low front vowel /æ/ before nasals in Kansas City in terms of a scale of height ranging from 100 [æ] to 900 [e]. The choice of numbers in this scale is essentially arbitrary (big numbers were chosen so as to avoid having to deal with decimals!). Here is the information she presents:

Age	**60+**		**40-59**		**20-30**		**< 20**	
Sex	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Social Status								
High	345	163	347	575	---	365	416	697
Mid	429	442	509	388	400	417	475	592
Low	384	358	232	353	326	594	269	486

Lust had a rather small sample of speakers, which caused some perturbation in the data, and a missing cell in the top row, but the data is still basically analyzable.

a. Draw the most revealing graph(s) you can to sort out what is going on here.

b. Do multiple linear regressions of the data. The way one does this is by turning the categorical variables into numeric variables by means of indicator (Bernoulli) variables that just take values 0 and 1. For example, one can have a variable for 60+, for 40-59, for Female, and so on. For instance one could represent one of the data points as:

Height	60+	40-59	20-30	Male	Mid	High
345	1	0	0	1	0	1

Important: one should not define such an indicator variable for *every* value of a dimension/factor group as the regression problem then becomes underdetermined. In my discussion of Labov last time (and in what happens in ANOVA), an extra constraint is added that the average of the effects of the different values contributed by a dimension should be 0. There's no easy way to tell a standard linear regression package that (AFAIK), but we can get the same effect of adding a constraint by leaving out one of the values. The left out value will have a zero for each listed value of the factor, and so its value will be determined just by other factors and the intercept. [We could then posthoc readjust the weights and the intercept.] As well as throwing everything in, you might wish to try just taking a slice of the data, say the lower class group and doing a regression of vowel height versus age group and gender within that slice. The missing data can be dealt with simply by leaving out a row of the complete table.

c. Comment on your analysis, making sure to answer the following questions:
    i. Is this variable of /æ/ height socially stratified? That is, do the different social classes appear to exhibit different speech patterns?
    ii. Is there evidence of an ongoing change? If so, what appears to be its direction and which group(s) appear(s) to be leading the change? (If you know a little sociolinguistics, it might be helpful to consider which of the two dialects in contact here would likely be the more prestigious.)
    iii. Do the two sexes show any differences in their behavior with respect to this variable?
    iv. How well does the linear regression model appear to fit the data? Are there particular cells or dimensions that it works badly on? Why?
    v. *Bonus:* Is there anything one could do to improve the fit of the model?