STAT 4620/5620: DATA ANALYSIS

2006 Fall

Instructor: Hong Gu, Office: Chase 101, Phone: 494-7161, E-mail: hgu@mathstat.dal.ca

Lectures: 2:35pm-3:25pm Mon. Wed. Fri. Place: Chase 227

Office Hours: MWF. 12:30 -1:30

Prerequisite: Stat 3340, Stat 3460 , or instructor's consent

Textbook: Categorical Data Analysis, Second Edition by Agresti (Wiley).

Other recommended reference books:

"Generalized Linear Models" by Nelder & McCullagh (2nd Ed, Chapman Hall),

"Generalized Additive Models" by Hastie & Tibshirani (Chapman Hall)

"Modern Applied Statistics in S" by Venables and Ripley (4th ed Springer)

"Generalized Linear models" by R. H. Myers , D. Montgomery and G. Vining.

Assessment:

30 % Homework assignments (probably 5 to 6 assignments)

30 % Project (Detailed requirement will be given in the middle of the term)

40 % Final Exam (open book, take home data analysis exam)

Course Outline:

1. Review of the Linear Regression Models.
2. Logistic Regression .
3. Poisson Regression.
4. Logit Models for Multinomial responses.

5. Contingency tables.
6. Generalized Linear Models .
7. Gamma response.
8. Smoothing and non-parametric regression.
9. Generalized Additive models.

University Policy on Intellectual Honesty

Data Analysis 4620/5620 Final Examination - Fall 2006

Instructions:

1. **This is an open book, take-home exam. You must work independently and not collaborate with other classmates or anyone else. Any suspected plagiarism will be investigated.**

2. **You must send in your solutions by emailing to me no later than 10pm Nov. 11, unless other agreement has been reached beforehead.**

3. **For numerical analysis questions, you need to submit the codes and the output as well as the explainations or comments about the output to address the questions. Handing in only the computer output is not sufficient. If your explanations or comments are made right below the program output, You should try to highlight those parts.**

4. **Partial marks can be given for partially correct answers.**

5. **Should you have problems in comprehending the questions, email me or call me at any time during the day. You may also leave a message on my answering machine, and I'll reply ASAP. My email address: hgu@mathstat.dal.ca, and phone: 4455306 (this is my home phone number, I won't be on campus on Nov. 11) .**

6. **There are 4 questions given below. Each of the first 3 questions includes 5 small questions. For those students registered as 4620, you may only answer the questions 1-3 with each of those small questions worth of 5 marks. Answering the 5th question may get you some bonus points. For those students registered as 5620, you need to answer all 4 questions. Each of those small questions in question 1-4 is worth of 5 marks and the last question is worth of 10 marks. Your marks will be rescaled finally to have full mark as 100.**

Question 1.

These are the data from Wilner, Walkley and Cook (1955) on the effect of racial attitudes on segregation and

integration of public housing:

			Sentiment	
Proximity	Contact	Norms	fav	unfav
close	frequent	favorable	77	32
		unfavorable	30	36
	infrequent	favorable	14	19
		unfavorable	15	27
distant	frequent	favorable	43	20
		unfavorable	36	37
	infrequent	favorable	27	36
		unfavorable	41	118

You can get a file in the usual character and numeric formats from housing.dat or housing.raw, respectively. The latter codes the factor levels in order of appearance as follows:

- Proximity: 1 = close, 2=distant
- Contact: 1 = frequent, 2=infrequent
- Norms: 1=favorable, 2=unfavorable

(a) Use the logit link within logistic regression to analyze these data. Choose the model that you think the most proper for this data.

(b) Test the goodness of fit of your model, interprete.

(c) Interprete parameter estimates.

(d) Which combination of the values for Proximity, Contact and Norms is most likely to give the Sentiment favorable answer? Find the estimated probability for this combination of the values of Proximity, Contact and Norms.

(e) Calculate a 95% confidence interval for the probability of giving the Sentiment favorable answer at the combination values of the predictors in (d).

Question 2.

Smoking and Lung Cancer

This dataset has information on lung cancer deaths by age and smoking status. The file in "raw" format has four columns:

- age: in five-year age groups coded 1 to 9 for 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79, 80+.
- smoking status: coded 1 = doesn't smoke, 2 = smokes cigars or pipe only, 3 = smokes cigarrettes and

cigar or pipe, and 4 = smokes cigarrettes only,

- population: in hundreds of thousands, and
- deaths: number of lung cancer deaths in a year.

The file is also available in "dat" format with variable names, row names and string labels for age and smoking status.

(a) Use logistic regression to analyze this data.

(b) Interprete the results and draw the proper conclusions from your logistic regression.

(c) Repeat the analysis using Poisson regression.

(d) Interprete the results and draw the proper conclusions from your poisson regression.

(e) Discuss the two analyses, which one do you prefer to use on this data? Why?

Question 3.

The Method Choice Data

The method choice data from Brazil are available in a file containing three columns:

- Age group: 15-19, 20-24, 25-29, 30-34, 35-39 or 40-44
- Method: sterilization, efficient, inefficient, or not_using
- Frequency: the number of women in each age/method combination.

As usual, the file is available in two formats: brazil.dat codes the factors using character labels, and brazil.raw uses numeric codes (the age groups are coded 1-6 and the methods are coded 1=not_using, 2=inefficient, 3=efficient, 4=sterilization).

(a) Fit a multinormial baseline category logit model to predict the “Method” using “Age group”.

(b) Interpret the above fitted model in (a), compare the fitted values with the observed values using graphical display(s).

(c) Fit a cumulative logit model to predict the “Method” using “Age group”.

(d) Interpret the above fitted model in (c), compare the fitted values with the observed values using graphical display(s).

(e) Compare the above two fittings, which one do you think more proper to this data set? Why? (hint: AIC can be used to compare two different models).

Question 4.

Leukaemia survival times data

The data gives the times to death (in weeks) and white blood cell counts for two groups of leukaemia patients, AG-positive and AG-negative. There are 33 observations on three variables.

- WBC: The white blood cell count.
- Time: The time to death (in weeks)
- AG: AG is 1 for AG-positive patients, and is 2 for AG-negative patients.

The file is available here.

Use proper approach in GLM to predict The time to death for the leukaemia patients, draw conclusions from your analysis.