MATH-661: Generalized Linear Models
Final Exam
Due Monday May 7, 2018

INSTRUCTIONS

No collaboration or discussion is permitted on the final exam. If you need clarifications, you can contact me but you are not allowed to ask anyone else. Please note that clarifications are limited to ambiguities in the wording of questions. This exam is intended to demonstrate your grasp of the material, so no help will be provided.

Please fill your name and sign the following honor pledge:
I, _____, pledge that I have not violated the Georgetown University honor code (see http://gervaseprograms.georgetown.edu/honor/). The work I am submitting for this exam is completely my own. I have not communicated with anyone and have not allowed any other student to use or borrow portions of my work. I understand that if I violate this honesty pledge, I will be reported for academic dishonesty to the Honor Council.
Signature : _____

- Show the details of your work in order to get full credit for correct answers, and partial credit for incorrect answers if you are on the right track.
- Provide interpretations and conclusions in the context of the problem.
- Include the relevant R code and output for each question, when applicable.
- The exam must be e-mailed to `mgt26@georgetown.edu` by 11:59 pm on Monday May 7, 2018.

Part I: Chronic respiratory disease [25 points]

Table 1 summarizes the data from an epidemiological study of chronic respiratory disease. Researchers collected information on subjects' exposure to general pollution (low or high), exposure to pollution in their jobs (yes or no), and their smoking status (current smoker, ex-smoker, non-smoker). The measured response is chronic respiratory disease status classified into four categories:

- 1 – no symptoms
- 2 – cough or phlegm less than 3 months a year
- 3 – cough or phlegm more than 3 months a year
- 4 – cough and phlegm plus shortness of breath more than 3 months a year

			Response level				
Air pollution	Job exposure	Smoking status	1	2	3	4	Total
Low	No	Non	158	9	5	0	172
		Ex	167	19	5	3	194
		Current	307	102	83	68	560
	Yes	Non	26	5	5	1	37
		Ex	38	12	4	4	58
		Current	94	48	46	60	248
High	No	Non	94	7	5	1	107
		Ex	67	8	4	3	82
		Current	184	65	33	36	318
	Yes	Non	32	3	6	1	42
		Ex	39	11	4	2	56
		Current	77	48	39	51	215

Table 1: Chronic respiratory disease data

1. Fit a proportional odds cumulative logit model with pairwise interaction effects for all covariates and assess its goodness of fit. Use low air pollution, no job exposure and non-smoker as reference group.
2. Use a likelihood ratio test to check the proportional odds assumption in the model above.
3. Use a likelihood ratio test to determine whether to include or not the interaction terms in the proportional odds cumulative logit model.
4. In the following questions, use the main effects cumulative logit proportional odds model:
   (a) Interpret each of the three intercepts.

(b) Which variables appear to be associated with chronic respiratory disease? Interpret the regression coefficients for the covariates with significant association.

(c) What are the estimated probabilities of falling in each of the different response categories for a current smoker with job exposure to pollution and high general air pollution exposure? Show the details of your calculations manually.

(d) For each covariate pattern, provide the predicted number of people falling in each of the response levels.

Part II: Number of plant species in the Galápagos [25 points]

The 30 islands in the Galápagos archipelago have long been studied by botanists, zoologists and biologists to learn about species survival and the process of natural selection in an almost experimental setting. The islands are essentially uninhabited by humans and all experience the same surrounding climate. Yet some species of birds, plants and mammals thrive on only a few or even just one of the islands. In addition, some islands have a wide variety of species, while others are not nearly as biodiverse. We are interested in investigating which variables may be related to the number of plant species in the archipelago islands.

The data `Galapagos.txt` posted on Canvas contain information on plant species on the Galápagos islands. The variables in the data correspond to

- `island` – name of island
- `species` – island total observed plant species count
- `endemics` – island endemic plant species count
- `area` – island area (km^2)
- `elevation` – island elevation (meters)
- `nearest` – distance in km from the island to its nearest neighbor (adjacent island)
- `scruz` – distance in km from the island to the largest island (Santa Cruz)
- `adjacent` – area of the adjacent island

1. **Exploratory data analysis**

   (a) Provide a histogram and summary statistics for the observed counts of total plant species. Discuss the distribution.

   (b) Create plots of the logarithm of the observed counts of total plant species, log(`species`), versus each of the five potential covariates: `area, elevation, nearest, scruz, adjacent`.

(c) Repeat the previous question using the logarithm of each of the covariates. Which variables appear to be related to log(`species`)?

Caution: Always check for any zero value before using a logarithm transformation. A quick fix is to add a small non-zero number, e.g., consider $x + 0.1$ instead of x.

2. **Model building & diagnostics**

(a) Fit a Poisson model with all five covariates on the log scale. Which covariates appear to have a significant effect on species counts?

i. Evaluate the goodness-of-fit of this model.

ii. Examine the standardized residuals. Explain whether or not they suggest the presence of overdispersion?

iii. Fit a negative binomial model with all five covariates on the log scale. Provide the point estimate and 95% confidence interval for the dispersion parameter. Which covariates appear to have a significant effect on species counts?

iv. Use a quasi-likelihood approach with an inflated quadratic function using all five covariates on the log scale. What is the estimated dispersion parameter? Which covariates appear to have a significant effect on species counts?

(b) Calculate the pairwise sample correlation between the covariates on the log scale and comment on whether or not multicollinearity may be an issue.

(c) Perform stepwise selection for the Poisson model with all covariates on the log scale.

i. Evaluate the goodness-of-fit of the selected Poisson model.

ii. Examine the standardized residuals and identify potential outliers.

(d) Perform stepwise selection for the negative binomial model with all covariates on the log scale.

i. Evaluate the goodness-of-fit of the selected negative binomial model.

ii. Examine the standardized residuals and identify potential outliers.