

## Stat 191 Final Exam

Due: March 19 12 P.M. Please submit paper copy to my soffice.

**Instructions:** There are 4 problems, please submit each problem on separate sheet(s) and put your name on each sheet. Please turn in all of your R code and R output. **Submit only the plots and output that support your methods and conclusions.** For each question, explain clearly (i) your objectives, (ii) the hypotheses that you are testing, (iii) the statistical procedure, and (iv) your conclusions.

**Honor Code:** Please respect the honor code in completing this exam. You can use books and computers, but not other people.

1. The following table refers to the effect on political party identification of gender and race.

		Party Identification		
Gender	Race	Democrat	Republican	Independent
Male	White	132	176	127
	Black	42	6	12
Female	White	172	129	130
	Black	56	4	15

- (a) Analyze this data using a multinomial logit model treating party identification as a response. Determine the best fitting model. State your model selection criterion.
  - (b) Interpret the model you chose for part (a), and use it to quantify the effect of race and gender on party identification. Be specific, use log odds ratios, and give confidence intervals using Gaussian approximation. Do you believe in these confidence intervals? Why or why not?
2. Suppose that  $X$  and  $Y$  are conditionally independent given  $Z$ , and that  $X$  and  $Z$  are marginally independent.
    - (a) Show that  $X$  is jointly independent of  $Y$  and  $Z$ .
    - (b) Show that  $X$  and  $Y$  are marginally independent.

- (c) Show that if  $X$  and  $Z$  are conditionally (rather than marginally) independent, then  $X$  and  $Y$  are still marginally independent.
3. The data file `HeartDisease.txt` contains information collected by a health insurance company on 788 of its subscribers who had made claims resulting from coronary heart disease. The columns of the table are:

Variable Name	Description
ID	1-788
Total cost	Total cost of claims by subscriber (dollars)
Age	Age of subscriber (years)
Gender	Gender of subscriber (1 if male; 0 otherwise)
Interventions	Total number of interventions or procedures performed
Drugs	Number of drugs prescribed
Emergency room visits	Number of emergency room visits
Complications	Number of other complications during treatment
Comorbidities	Number of other diseases during period
Duration	Number of days of duration of treatment condition

Treat the number of emergency room visits as the response and the other variables as potential predictors.

- (a) Fit a Poisson regression model to the data. Use the deviance residuals and leverage to assess the adequacy of the Poisson regression model. Are there any outliers in the data? You may want to drop these outliers for the rest of this question.
- (b) Using the AIC and step-wise search, decide which variables should be included in the model.
- (c) Write your own R function to perform  $K$ -fold cross-validation, and use it to assess the prediction error of your model.
- (d) Analyze the same data using linear regression by appropriately transforming the response variable. Do you get similar results?
4. **Speed of Evolution.** How fast can evolution occur in nature? Are evolutionary trajectories predictable or idiosyncratic? To answer these questions, R.B. Huey et al. (*Rapid evolution of a geographic cline in size in an introduced fly*, *Science* 287:308-9, 1990) studied the development of a fly *Drosophila subobscura* that had accidentally been introduced from the Old World into North America (NA) around 1980. In Europe (EU), characteristics of the flies wings follow a “cline” a

steady change with latitude. One decade after introduction, the NA population had spread throughout the continent, but no such cline could be found. After two decades, Huey and his team collected flies from 11 locations in western NA and native flies from 10 locations in EU at latitudes ranging from 35-55 degrees N. They maintained all samples in uniform conditions through several generations to isolate genetic differences from environmental differences. Then they measured about 20 adults from each group. (Data contained in **Evolution.txt**)

- (a) Construct a scatter plot of average wing size against latitude, in which the four groups defined by continent and sex are coded differently. Do these suggest that the wing sizes of the NA flies have evolved toward the same cline as in EU?
- (b) Construct a multiple linear regression model with wing size as the response, with latitude as a linear explanatory variable, and with indicator variables to distinguish the sexes and continents. Construct the model in such a way that one parameter measures the difference between the slopes of the wing size v. latitude regressions of NA and EU for females, one measures the same difference for males, one measures the difference between the intercepts of the regressions of NA and EU for females, and one measures the same difference for males.
- (c) The authors of that study concluded that although the wing size of North American flies was converging rapidly to the same cline as exhibited by the European flies, the means by which the cline is achieved is different in the North American population. As evidence that the means of convergence is different, they concluded that there was a marked difference between the NA and the EU patterns of the basal length-to-wing size ratios versus latitude (in females). Fit a multiple linear regression, which allows for different slopes and different intercepts. In a single F-test, evaluate the evidence against there being a single straight line that describes the cline on both continents. If you conclude there is a difference, is the difference one of slope alone? of intercept alone? or of both?
- (d) Return to the basic question of whether the wing sizes in NA flies have established a cline similar to their EU ancestors. Using the model developed in (b), answer these questions: (i) Is there a

- non-zero slope to the cline of NA females? (ii) Is there a non-zero slope to the cline of NA males? (iii) Is there a difference between the clines of NA and EU females, and if so, what is its nature? and (iv) repeat (iii) for males?
- (e) Repeat the above analysis using weighted regression. Do the results differ? Why is this preferable to using each fly as a separate case?