Final Project - Statistics 701

Work on this exam by yourself and be sure to reference any material you use on your exam. Only discuss the exam with me and not the other students in the class.

*If you have a pre- approved dataset you may analyze it in place of **one** on the first two problems.*

1: (*Ex 14.3 R Handbook)-The data shown in Table 14.3 were collected in a follow-up study of women patients with schizophrenia (Davis, 2002). The binary response recorded at 0, 2, 6, 8 and 10 months after hospitalization was thought disorder (absent or present). The single covariate is the factor indicating whether a patient had suffered early or late onset of her condition (age of onset less than 20 years or age of onset 20 years or above). The question of interest is whether the course of the illness differs between patients with early and late onset? Investigate this question using the GEE approach.*

i. Provide a two-page write-up (including graphs) explaining your analysis of the experiment and the conclusions you can draw from it.

ii. As a secondary component provide annotated code that replicates your analysis.

***2:** The file "y.dat" contains a dataset consisting of 150 4-variate observations. True group labels are provided for the first 135 data points in the file "idy.dat". Classify the remaining 15 observations. There are three groups in this problem.*

*i. Provide **a page** write-up (including graphs) explaining what methods you used to model the groups and how you predicted the identity of the remaining 15 observations.*

ii. *As a secondary component provide annotated code that replicates your analysis.*

3: (Vole Data)- Consider the *"microtus"* dataset in the "Flury" library in R.

Background from Airoldi_Flury_Salvioni_JTheorBiol_1995: Discrimination Between Two Species of Microtus using both Classified and Unclassified Observations.

"1. Introduction

Microtus subterraneus and M. multiplex are now considered to be two distinct species (Niethammer, 1982; Krapp, 1982), contrary to the older view of Ellerman & Morrison-Scott (1951). The two species differ in the number of chromosomes: 2n=52 or 54 for M. subterraneus, and 2n=46 or 48 for M. multiplex. Hybrids from the laboratory have reduced fertility (Meylan, 1972), and hybrids from the field, whose karyotypes would be clearly recognizable, have never been found (Krapp, 1982).

The geographic ranges of distribution of M. subterraneus and M. multiplex overlap to some extent in the Alps of southern Switzerland and northern Italy (Niethammer, 1982; Krapp, 1982). M. subterraneus is smaller than M. multiplex in most measurements, and occurs at elevations from 1000 m to over 2000 m, except in the western part of its range (for example, Belgium and Brittany), where it is found in lower elevations. M. multiplex is found at similar elevations, but also at altitudes from 200–300 m south of the Alps (Ticino, Toscana).

The two chromosomal types of M. subterraneus can be crossed in the laboratory (Meylan, 1970, 1972), but no hybrids have so far been found in the field. In M. multiplex, the two chromosomal types show a distinct distribution range, but they are morphologically indistinguishable, and a hybrid has been found in the field (Storch & Winking, 1977).

No reliable criteria based on cranial morphology have been found to distinguish the two species. Saint Girons (1971) pointed out a difference in the sutures of the posterior parts of the premaxillary and nasal bones compared to the frontal one, but this criterion does not work well in many cases. For both paleontological and biogeographical research it would be useful to have a good rule for discriminating between the two species, because much of the data available are in form of skull remains, either fossilized or from owl pellets.

The present study was initiated by a data collection consisting of eight morphometric variables measured by one of the authors (Salvioni) using a Nikon measure-scope (accuracy 1/1000 mm) and dial calipers (accuracy 1/100 mm). The sample consists of 288 specimens collected mostly in Central Europe (Alps and Jura mountains) and in Toscana. One peculiar aspect of this data set is that the chromosomes of 89 specimens were analyzed to identify the species. Only the morphometric characteristics are available for the remaining 199 specimens…"

Develop a GLM model from the 89 specimens that you can use to predict the group membership of the remaining 199 specimens'.

i. *Explain your GLM and assess the quality of the fit with the classified observations.*

   - *Use Cross Validation to predict the accuracy of your model.*

ii. *Provide a one-page write-up (including graphs) explaining your analysis of the dataset and your recommendations on the usefulness of your predictions.*

iii. *Provide predictions for the unclassified observations.*

iv. *As a secondary component provide annotated code that replicates your analysis.*