Stat 503X

EXAM 1

Name: SOLUTION

Student Number:

There are 5 equally weighted questions, and you have 50 minutes to complete the exam. You should spend no more than 10 minutes on each question. This exam is also closed notes. You should only need a pen or pencil to complete it.

1. (10pts) Explain the method called “jittering”, and how it might be used for exploratory data analysis.

   Jittering a variable means adding a small amount of noise, some random quantity from a known distribution, such as uniform or normal, to the data value. It is useful in exploratory data analysis when visually exploring categorical data using scatterplots (includes dotplots, grand tours of scatterplots). In categorical data there can be many data points with the same variable values, so jittering allows spreading these values out so we can see how many points have the same values, and hence allow us examine the distribution of the data.
2. *(10pts)* What visual cues would you use to detect cluster structure if you were using a grand tour on high-dimensional data set?

*There are two visual cues here: motion of the points across views and separation of points in some views. Points that are clustered will likely move together in the same motion pattern as a grand tour progresses. Also, if the clusters are separated or unconnected then they will appear separated in some views displayed by the grand tour. Changes in density can also be an indicator of clustering, where the differences between clusters are due in part to variance differences.*

3. *(10pts)* Which of these are valid interpretations of the mosaic plot of the tips data on the next page? Circle agree or disagree with the interpretation.

- More males pay the bill than females. *(Agree)*)
- Relatively higher numbers of males pay the bill on Fridays, otherwise the numbers are similar on other days. *(Disagree)*
- The distribution of female bill payers is different to males over days. *(Agree)*
- Sex and day of the week are independent variables. *(Disagree)*
- Amongst males, relatively more smoking parties on Thursday and Friday with numbers dropping off to Sunday *(Agree)*
- In smoking parties where a female has paid the bill, there is a similar pattern to males from day to day, except that Friday has the larger number. *(Agree)*
- Thursday is a big day for males during the day, but there are almost no males on Friday. *(Disagree)*
- The restaurant is closed Saturday and Sunday during the day. *(Disagree)*
- At night there is a predominance of smokers. *(Disagree)*
- No bill for a party of size 6 was paid by a female. *(Disagree)*
4. *(10pts)* This data was collected on two species (Orange and Blue), both males and females of crabs in Australia. There are five measured variables: Frontal Lobe, Rear Width, Claw Length, Carapace Width, Body Depth. From the figures below, the primary question is to work out which variables contribute to the separation of species and sex. In the grand tour view all the variables were standardized first, and then projection pursuit was used to get this view.

(a) *(3)* Describe what you learn about the data from the grand tour plot.

It appears that the two species are easily separated, and within the species it is possible to separate the sexes with a high degree of accuracy. It looks like all variables contribute to the separation of species, with Rear Width separating sexes more than the other variables.

(b) *(3)* Describe what you learn about the data from the dotplots and pairwise scatterplots combined.

The dot plots are very difficult to interpret. There are some location differences between the groups, but no clear separations. The reasons are clear when we look at the pairwise plots: there is strong correlation between the different measurements. Smaller crabs are harder to distinguish from each other based on these measurements, but the separations increase as size gets larger.

Males can be separated from females in plots of CL vs RW (males have a higher CL:RW ratio), and similarly in plots of BD vs RW, and CW vs RW.

The two species can be separated in the plots of BD vs CW, and CW vs FL.

It looks like it is almost possible to separate all four groups using just RW and FL.

(c) *(4)* Which variables contribute to the separation of species and sex?

In separating sex males have higher CL:RW, CW:RW, BD:RW ratio. This makes sense because of differences in reproduction. It is clear then that RW has a strong discriminatory power for the variable sex for both species, when used in relation with CL, BD, and CW.

In separating the species the ratio of BD: CW, and FL: CW, is roughly the same for both species, but the Blue species is a little smaller in both these measurements on average. That is, for equal values of CW, the Blue species has smaller BD and FL measurements.
5. (10pts) This data was collected on a field managed by Iowa State University. The primary question is how corn yield is related to the soil nutrients:

- B  Boron (parts per million)
- Ca  Calcium (parts per million)
- Cu  Copper (parts per million)
- Fe  Iron (parts per million)
- K  Potassium (parts per million)
- Mg  Magnesium (parts per million)
- Mn  Manganese (parts per million)
- Na  Sodium (parts per million)
- P  Phosphorus (parts per million)
- Zn  Zinc (parts per million)

In particular, how can we get higher values of the yield by controlling the values of the soil nutrients.

(a) (6) Briefly describe how you would approach to answering the primary question. (Like the Suggested approaches part of the assignments, give methods, and what parts of the question they might address.

*Summary statistics to obtain location-scale information about each variable. This would allow us to answer questions such as “What is the average yield?”.*

*Histograms and scatterplots to explore the distributions of each variable, and the relationship between yield and the nutrient variables, and correlations amongst the nutrient variables. It also might allow us to detect outliers, clustering or collinearity of the independent variables. If there were non-normal distributions, or non-linear relationships with yield we would explore transformations to fix the problem before conducting linear regression. This would allow us to answer questions such as “Which soil nutrient variables are most strongly related with yield?” or “How are the soil nutrient variables related to one another?” or “Are there erroneous data values?”.*

*Grand tour of the nutrient variables would allow us to explore them for multicollinearity and outliers, and to answer questions such as “How are the soil nutrient variables related to one another?”*  

*Multiple linear regression would allow us to pick out the variables that are the most linearly related with yield, and obtain numerical values for the quantities. It would allow us to answer such questions as “Which soil nutrient variables are most strongly related with yield?” or “What amount of Copper should I maintain in my field to get best yield.”*
(b) (4) From the Figure below, describe the relationship between Yield and Boron. (High, medium and low values of Yield are brushed differently, but this is irrelevant for answering the question.) Explain why simple linear regression would have missed this important relationship.

As Boron increases the variation in yield decreases. This is important because it says that to gain a consistently high yield you should maintain high levels of Boron. That is, you can’t ignore the Boron content in the soil if you want consistently good yield. Linear regression would completely ignore this relationship because it assumes constant variance, and would read this association as not significant.