STAT 501 **Assignment #6** **Name _____**
Spring 2005

Reading Assignment: Johnson &Wichern, Chapter 10, 11 and 12

Written Assignment: Due Wednesday, April 27, in class.

Final Exam: Thursday, May 5, 12:00-2:00 p.m.

1. Suppose that $n_1 = 11$ and $n_2 = 12$ observations are sampled from two different bivariate normal distributions that have a common covariance matrix Σ and possibly different mean vectors $\underset{\sim}{\mu}_1$ and $\underset{\sim}{\mu}_2$. The sample mean vectors and pooled covariance matrix are:

$$\underset{\sim}{\overline{X}}_1 = \begin{bmatrix} -1 \\ -2 \end{bmatrix} \qquad \underset{\sim}{\overline{X}}_2 = \begin{bmatrix} 4 \\ 5 \end{bmatrix} \qquad S = \begin{bmatrix} 13 & 6 \\ 6 & 22 \end{bmatrix}$$

   (a) Use the Hotelling two sample T^2-statistic to test for a difference in the population mean vectors. Report

   T^2 = _____ F = _____ d.f. = _____ p-value = _____

   (b) Report the estimate of the formula for Fisher's linear discriminant function. Explain how this function is used to classify.

   (c) Consider an observation $\underset{\sim}{X}_0 = [\ 2,\ 1\ ]$ on a new experimental unit. Was this unit more likely to have come from population 1 or population 2? (Assume equal misclassification costs and equal prior probabilities.)

   (d) Classify the unit in Part (c) assuming prior probabilities .65 and .35 of observing a unit from populations 1 and 2, respectively. Also, assume the cost of misclassifying a unit from population 1 into population 2 is ten times greater than the cost of misclassifying a unit from population 2 into population 1.

2. One part of developing a plant disease management system for growers of fruits and vegetables is the development of a model for predicting dew events. Many fungal and bacterial pathogens are active on plant surfaces only when free water is present. The primary sources of water on plant surfaces are dew and rainfall. Reliable information on rainfall is readily available from the National Weather Service weather stations. The development of a model for the occurrence of dew would enable growers to estimate the total length of time that plant surfaces are wet and help them minimize the use of pesticides for controlling fungal and bacterial pathogens that could damage their crops.

Hourly data on dew events was collected at various locations in Iowa, Illinois, and Nebraska during the growing season. At each location, dew events were monitored by either CR-10 or CR21-X wetness monitors located on mowed turf grass. These monitors were programmed to continuously record data from electrical wetness sensors at one minute intervals. A determination was made for each hour of the night between 7 p.m. and 9 a.m. Daylight hours, between 9 a.m. and 7 p.m. were excluded because dew is unlikely to be present on plant surfaces during those hours. Hours with rain events were also excluded from this data file. An hour was classified as a dew event if it was not a rain event and the monitor recorded wetness for at least 30 minutes during the hour.

The wetness monitors also recorded information on temperature, humidity, wind speed and other weather variables. This information is not included in your data set, however, because the objective is to predict dew duration from data available from local weather stations associated with the National Weather Service. Farmers have been reluctant to use wetness monitors in their fields because of the purchase cost, maintenance cost and calibration problems. Farmers can obtain hourly data on temperature, wind speed and relative humidity from local weather stations, and the goal is to predict dew duration from such data. Although there is at least one such weather station in each county in Iowa, the location of the wetness monitor may be as far 30 miles from the nearest weather station where air temperature, wind speed and relative humidity are measured. This can be a source of error because weather conditions can be slightly different at the weather station than they are at the location of the wetness monitor. For example is could be raining on the wetness monitor but not raining at the weather station, or air temperature at the weather station might be measured at a point several feet higher than the location of the sensor plates on the wetness monitor. In the evening air cools as it passes across the earth and air temperature at 5 feet above ground tends to be higher than air temperatures 1 foot above the ground. Also, temperature at the location of the wetness monitor could be slightly higher than temperature at the nearest weather station if the weather station is in a valley and the moisture monitor is on higher ground, or vice versa. Actual wind speeds could be different if either the wetness monitor or the weather station is in a more sheltered location. These and other sources of variability will prevent you from being able to construct a perfectly accurate classification model for hourly dew events.

Dew forms when air temperature cools. Cool air cannot hold as much water as warm air. The temperature at which the water in the air begins to condense to form dew is called the dew point. The dew point is one of the variables in your data set. It is a function of relative humidity, a measure of the amount of water in the air. The dew point is higher when relative humidity is higher. For your data, dew will form and be present at air temperatures slightly above the dew point. This results from measuring air temperature at a point higher above the ground than the location of the wetness monitor (and the leaves on the plants). Consequently, it may be important to consider some deviation between dew point and air temperature in your model. Furthermore, high wind speeds may either prevent dew from forming or help to dry up existing dew.

Along with the total amount of time that dew is present, the data contain hourly information on air temperature (ºC), wind speed (km/sec), relative humidity (%), and dew point (ºC).

The data are stored in the file dewa.dat. There is one line for each hour at each location, with 8 numbers on each line in the following order.

Location	A location code
Day	Number of days from the beginning of the year
Hour	Hour (700 denotes 7 a.m., 2100 denotes 9 p.m.)
X1	Air temperature (ºC)
X2	Relative humidity (%)
X3	Wind speed (km/sec)
X4	Dew point (ºC)
Wet	wetness indicator (0 = dry, 1 = wet)

Since we are trying to construct a model for the entire region to classify each hour as either a dew event or a non-dew event, and do not use location as an explanatory variable in your analysis. Although you could create new explanatory variables as functions of the original variables, use only the original values of air temperature, relative humidity, wind speed and dew point in the following analysis. Also consider only equal priors and equal misclassification costs. This will minimize the effort needed to answer the questions. You could explore these data further this summer.

(a) Apply linear discriminant analysis to values of air temperature (X1), relative humidity (X2), wind speed (X3) and dew point (X4) to construct a classification rule for distinguishing wet hours from dry hours. Report a formula for your classification rule. Report resubstitution and cross validation estimates of misclassification probabilities for your classification rule.

(b) Apply logistic regression to the same four variables to construct a classification rule for distinguishing wet hours from dry hours. Classify into population 1 if the estimated conditional probability of coming from population 1, given the measurements on air temperature (X1), relative humidity (X2), wind speed (X3) and dew point (X4), exceeds 0.5, otherwise classify into population 2. Report a formula for your classification rule. Report resubstitution and cross validation estimates of misclassification probabilities.

(c) Use the rpart ( ) function in S-PLUS to construct a classification tree. Report a diagram for your tree and explain how you will use it to classify evening hours as either dew events or non-dew events. Report cross validation estimates of misclassification probabilities.

(d) Give any additional suggestions you may have for fitting a good classification rule. (This is an optional part of the assignment and you may report your suggestions without actually trying them out on the data.)