Ecol 145 Assignment 1
Dahl Winters
1/20/06

Questions

1. *In the full data set there are 48 different reefs (identified by the variable REEF_NAME) that were visited repeatedly (roughly once a year) over a 6-year period. Some reefs were missed in some years. Thus we have what are called unbalanced data. Give me one line of R code whose output makes it easy to identify which reefs were missed one or more times. Using the output identify those reefs. (**Hint**: The data set is an example of what an epidemiologist would call a person-period data set, or better yet here, a coral-period data set. Thus in the data set each coral reef has multiple records, one for each measurement time. One of the functions we used during Tuesday's class will do the job to answer this question.)*

   ```
   table(corals$REEF_NAME)
   ```

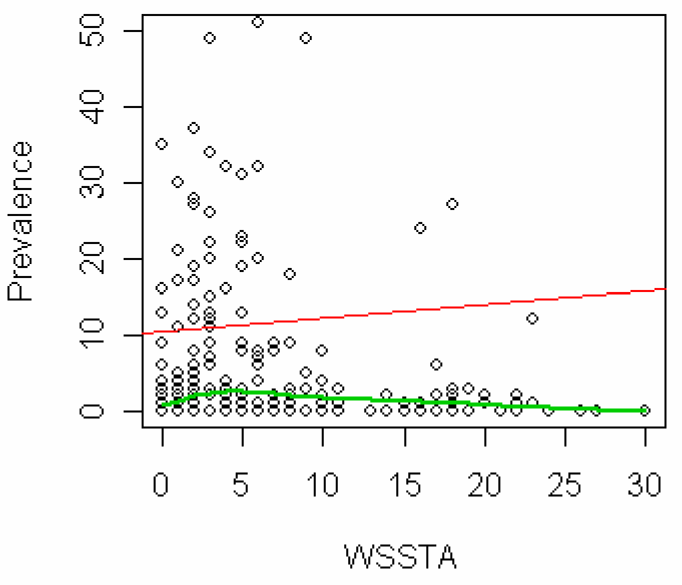
   - This gives a count of the number of times each reef was visited. Those reefs without 6 entries must have been missed one or more times. There were only 2 such reefs, DECAPOLIS and THETFORD, both of which were visited 5 times.
   (Just a comment: there are actually 7 years (1998-2004) in the data, so if we go exactly by the year column by doing `table(corals$REEF_NAME,corals$YEAR)`, then every reef is missed at least once because there are never more than 6 years recorded for any reef. This is because the data was collected from the end of 1998 to the beginning of 2004. If this is right, maybe the problem could be rephrased to say 6 sampling periods instead of 6 years. I figured this was meant after doing Question 3 where the 6 sampling periods are clearly visible.)

   > *BONUS: Can you figure out how to get R to list only the reefs that were not visited every year? This can be done with one additional line of code.*
   > ***Hint 1**: The output of every R function can be assigned to a variable.*
   > ***Hint 2**: The method we used to locate the hotspots (i.e., the way we subsetted the data in coloring points on a scatter plot) is relevant here.*

   ```
   > names(table(corals$REEF_NAME))[table(corals$REEF_NAME)<6]
   [1] "DECAPOLIS" "THETFORD"
   ```

2. *Plot disease prevalence (PREV_1) versus the temperature metric (WSSTA) using all six years of data (ignoring the data structure, i.e., treat all 280 observations as if they were independent. Add a linear regression line to the plot. Add a lowess smooth to the plot. Comment on what you see. You may need to change the vertical scale to see anything at all.*

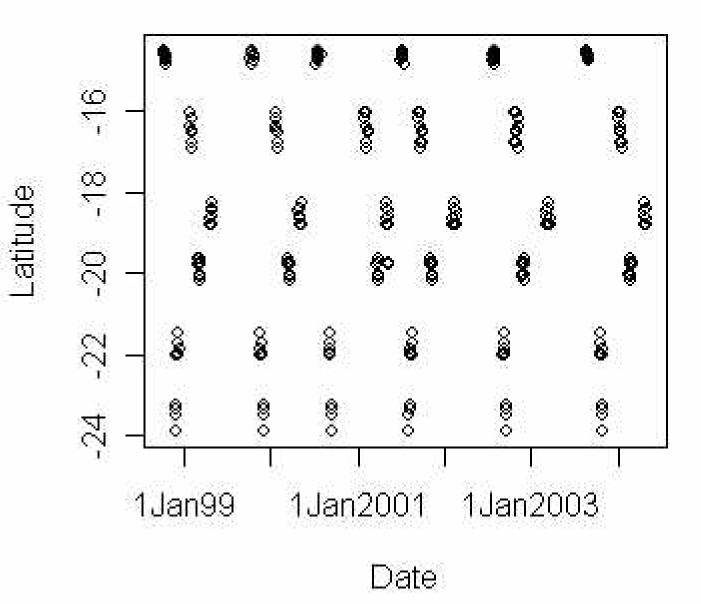
   ```
   plot(corals$WSSTA, corals$PREV_1, xlab='WSSTA', ylab='Prevalence',
   ylim=c(0,50))
   abline(lm(PREV_1~WSSTA, data=corals), col=2)
   summary(lm(PREV_1~WSSTA, data=corals))
   ```



- The linear regression line's intercept is not negative, which is good. The R2 is 0.001009, and the p-value is 0.5966, which is significant (above 0.3).
- The lowess curve does not match the linear regression, which indicates problems with the linear regression. According to the important trends in the data characterized by the nonparametric lowess curve, if there is less than a 5-degree increase in sea surface temperature, disease prevalence increases, but then it decreases steadily as the temperature change increases further. However, the linear regression indicates that prevalence should increase steadily as temperature increases.
- The data itself is clustered toward the left and bottom of the plot. Most samples had low prevalence at low temperature changes; very few had high prevalence at high temperature changes.

3. *We discovered in the data set that covered only one year that the samples were taken in such way that observations near each other spatially also were sampled at about the same time of year. When all six years of data are examined, does this pattern continue? (Provide evidence with a graph.)*

```
plot(as.date(as.character(co
rals$DATE)), corals$LAT_DD,
xlab='Date',
ylab='Latitude')
```



Yes, the pattern continues. It is especially clear when looking at samples for 2004. All samples at a particular latitude are taken all around the same month, with samples at different latitudes being done at different times of the year.

4. *In Tuesday's class we used the summary function of R to view detailed regression results from a linear model (lm) object. When summary is applied to other kinds of objects, the information*

you get is different. Try using summary on the WSSTA variable. Explain what each element in the output represents.

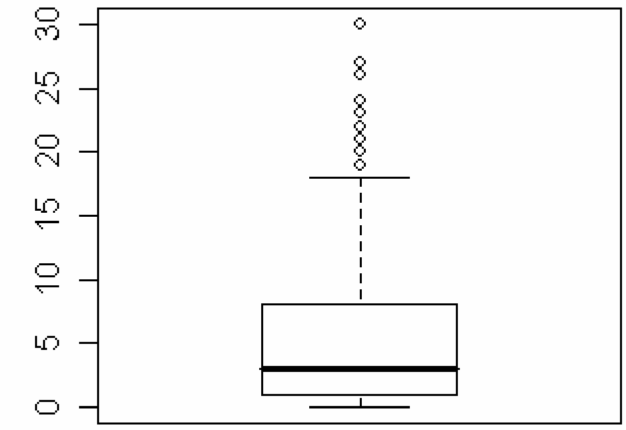
```
summary(corals$WSSTA)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   1.000   3.000   6.064   8.000  30.000 
```

- Min. and Max: the lowest and highest amounts of temperature change, respectively.
- 1st Qu. and 3rd Qu: the value of the 1st and 3rd quartiles. If we ordered all WSSTA values from 0 to 30 and created 4 equal divisions, the quartiles would be the WSSTA values at these divisions. (I tested this idea after doing **`sorted<-sort(corals$WSSTA)`** to see what the quartile values would be; I tried **`help(summary)`** but that didn't give me any information on what these elements meant.) I imagine the 1st and 3rd quartiles could be useful in determining how much skew the distribution has.
- Median and Mean: the median and mean of all values in the WSSTA column.

5. *A useful summary plot for comparing distributions across groups when there are too many data points to plot individually is the boxplot. In R the command is boxplot(variable) where the argument, that which goes inside the parentheses, can be the variable you wish a boxplot of. Produce a boxplot of WSSTA. Using either the web, a textbook, or whatever, explain everything you see in the boxplot. Your answer to question 4 may be helpful here in interpreting the boxplot. If you're totally at a loss as to what you're seeing, here is a journal article to look at (available online at UNC).*

   Reese, R. Allan. 2005. Boxplots. Significance ***2****(3): 144-145.*

```
boxplot(corals$WSSTA)
```



- This boxplot (also box-and-whisker plot) is a way of showing the shape of the distribution of WSSTA values. The line at the end of the top whisker is the maximum value, and the line at the bottom of the bottom one is the minimum value. The bottom and top edges of the box are the 1st and 3rd quartiles, respectively (which also can be thought of as the 25th and 75th percentiles, so that the box contains the middle 50% of the data). The thick line in the box is the median value.
- The boxplot is good for instantly telling if a distribution is skewed, and whether there are any outliers in the data. Since the median line in the box is not equidistant from the top and bottom edges of the box, the WSSTA data is skewed. Since the median is less than the mean (looking not at the boxplot but at the values found in question 4), the data is skewed to the right. Suspected outliers are any points beyond the whiskers; there are 9 of those in our data.
- Sources: http://www.stats.gla.ac.uk/steps/glossary/presenting_data.html#box, http://www.netmba.com/statistics/plot/box/.

6. *What percentage of the prevalence values (PREV_1) are zero? Ideally, write one line of R code to do the entire calculation for you.*
   Hint 1*: the sum function can be used to add up a list of numbers. The operator for division is / and for multiplication is *.*
   Hint 2*: If x is a vector then we can extract, e.g., the third element of x using the notation x[3]*

   ```
   sum(corals$PREV_1==0)/sum(table(corals$PREV_1))*100
   ```

   I didn’t use Hint 2, but the above gives me 38.92857%.