**BIFX 503: Statistics for Bioinformatics**

**Homework Set #4**

**Due: November 30, 2021**

*Instructions:*

*Use R to complete this assignment. Assignment is to be submitted via Blackboard.*

The data **Sitka89** (from the MASS package) contains data from a study of Sitka spruce trees. 54 trees were grown in ozone-enriched chambers and 25 trees were controls. Their size (height times diameter squared, on log scale) was measured eight times at approximately monthly intervals. You will use repeated measures regression to determine how ozone affects tree growth.

1. In this dataset, what is the dependent variable? The subject variable? The within- and between-subject variables?

Size is the dependent variable. Tree is the subject variable. Time is the within-subject variable. Treatment is the between-subject variable.

1. Plot tree growth over time separately for the ozone and control groups, using interaction or other plot (e.g. spaghetti plot, lattice plot). Is there evidence of interaction?

An interaction plot using the means setting to separately compare the ozone and control groups. The lines do not cross but also not perfectly parallel. This seems to indicate that there is either slight interaction, or none.

Chart

Description automatically generated

Chart, line chart

Description automatically generated

1. Run a repeated measures regression model having time, treatment, and their interaction as independent variables. Is the interaction term significant?

No, the interaction term between treatment and time is only 0.76, far above the threshold for significance. This matches our predictions with the interaction plot to some extent, though we might have expected the lines to be more parallel, give how far above the threshold our interaction term turned out to be.

Text

Description automatically generated

1. Run a second repeated measures regression model having Day and Treatment as the only independent variables. How does the model fit compare with the interaction model? What do these models tell you about how ozone affects tree growth?

Time remains the most significant factor, while the difference between treatments changed from barely insignificant to barely significant. These models seems to indicate that there is no time-ozone interaction. This is reinforced by the graph. It is therefore likely to be the cause of the size difference between treatment groups. However, there does appear to be a significant difference in their size. This could be due to unknown confounding variables that disproportionately effected one group over the other, leading to such differences.

Text

Description automatically generated with medium confidence

The R dataset **Pima.tr** (from the MASS package) contains data on 200 women of Pima Indian heritage living near Phoenix, Arizona. All women were tested for diabetes according to World Health Organization criteria. The dataset contains whether or not they met WHO criteria for diabetes (variable type=Yes for diabetes, No otherwise), their plasma glucose concentration (variable glu), as well as variables npreg (number of pregnancies), bp (diastolic blood pressure), skin (triceps skin fold thickness), bmi (body mass index) and age.

1. First, examine the distributions of each of the *continuous* variables in this dataset (i.e. exclude type). Generate a histogram and QQ plot for each one, and comment on the distributional shape. Which ones look normal? For the skewed distributions, choose a transformation that will normalize them, and regenerate the plots for the transformed variables.

Glu, bp, and bmi all appear very normal. Meanwhile, npreg, skin, ped, and age all appear to be abnormal. Squaring appeared to fix skin, and reduce skewness of the others. Log10 appeared to fix ped, break npreg, and reduced skewness of age.

Initial graphs:

Chart, histogram, box and whisker chart

Description automatically generatedGraphical user interface, chart, line chart

Description automatically generated

Squared graphs:

Chart

Description automatically generated

Log10 graphs:

Chart, histogram

Description automatically generated

1. Create a dataset that contains the seven numeric predictor variables npreg, glu, bp, skin, bmi, ped, and age. For the variables that were transformed to be more normal in question 1, use the transformed instead of the original variables. Perform a principal components analysis (PCA) on these 7 variables:
   1. Scale the 7 variables.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. Produce a scatterplot matrix and calculate the correlation matrix using the 7 scaled variables. Do any variables appear highly related to each other?

Glu seems to be positively correlated with almost everything. Most correlated with log\_age. Barely correlated with log\_ped.

Bp seems to be correlated with almost everything. Most correlated with log\_age. But barely and negatively correlated with log-ped.

Bmi seems to be positively correlated with almost everything except sqrt\_npreg. Most correlated with sqrt\_skin. Least correlated with log\_age.

Sqrt\_npreg seems to be correlated with almost everything except bmi. Most correlated with log\_age. Least and negatively correlated with log\_ped.

Sqrt\_skin seems to be positively correlated with almost everything. Most correlated with bmi. Least correlated with log\_ped.

Log\_ped seems to be barely correlated with anything. Most correlated with bmi. Barely and negatively correlated with bp, sqrt\_npreg, log\_age.

Log\_age seems to be correlated with almost everything. Most correlated with sqrt\_npreg. Barely and negatively correlated with log\_ped.

Graphical user interface

Description automatically generated with low confidence

Table

Description automatically generated

* 1. Perform PCA, generate a scree plot and output the proportion of variance for each component. How many components seem adequate to describe the data? (In other words … how many components are necessary to capture most of the variance?)

It appears that the first 3 components account for about 69% of the variance, a clear majority. Including 4 components raises this to about 81%, and 5 to about 91%. However we must keep in mind the set only has 7 variables. As indicated by the screen plot, the graph seems to level out after component 3, so the best number of components to use would be 3.

A picture containing text

Description automatically generated

Chart, line chart

Description automatically generated

1. Perform a hierarchical cluster analysis on the 7 variables you used for the PCA:
2. Calculate the distance matrix on the (unscaled) variables

Calendar

Description automatically generated

1. Perform hierarchical clustering and plot the dendrogram. How many clusters are apparent?

There appears to be 4 distinct clusters all cut just below a height of 40.

Diagram

Description automatically generated

1. Perform a k-means cluster analysis on the same 7 variables, specifying two clusters. Link the cluster assignments back to the original dataset using cbind().
2. Cross-tabulate the Type variable (diabetes yes or no) with the cluster assignment from the k-means clustering.

Calendar

Description automatically generated

A picture containing application

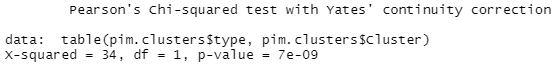
Description automatically generated

1. What proportion of women with diabetes (type=Yes) are in each of the two clusters?

There were 24 yes in cluster 1, and 44 in cluster 2. This means that cluster 1 had about 35% of the women with diabetes, while cluster 2 had about 64%.

1. Perform a chi square test to see if cluster assignment and diabetes are associated.

The p-value is way below the threshold for significance. There is clearly an association here between cluster assignment and diabetes type.



1. How well do you think the clustering algorithm did at classifying women with diabetes?

About 2/3rds of the diabetes women were put in cluster 2, and over 3/4ths of non-diabetes in cluster 1. This is good and better than I was expecting!