Hello everyone,

The pre-live assignment for this week is purely dedicated to PCA as well as my plans for the live session.

We are going to use the data from problem 17.16 in the text book to start out as a simple example.  So take a look there for some reference to the variables.  Do not do the questions in the book, do the questions below.

The SAS code attached contains the data and some basic starting code to run.

1. Take a look at the summary statistics, correlation matrix, and scatterplot matrix.
2. The PROC PRINCOMP code conducts PCA using the “cov” option which calculates the components using the covariance matrix estimated from the data.
   1. Examine the eigen values, eigen vectors, and the scree plot for this particular run.  What is the scree plot telling us here?  Comment on the “weirdness” of the eigen vector for the first principle component.  Any thoughts on why this occurred?
      1. **The scree plot indicates that only the first principal component should be included.**
      2. **The weirdness appears due to the extreme variance of the income variable.**
   2. There is an option “out=pca” included that creates a new data set that contains the principle component variables along with the original variables.  Use this data set to create summary statistics for the principle component variables.  (Use proc print to see the data set if you’re not sure what the variable names are).  In particular, look at the variance for each of the principle component variables in the summary table.  Do those values look familiar?  Hopefully this will be a wow that’s interesting moment 
      1. Table

         Description automatically generated
   3. Compare the scatterplot matrix of the raw values and that of the principle components.  What happens to the trends that we see in the original variables?
      1. Diagram

         Description automatically generated
      2. Diagram

         Description automatically generated
3. Run another princomp procedure with the “cov” option removed.
   1. Compare the eigenvalue, eigenvector, scree plot with the first run.  How does the interpretation of the first principle component change?
      1. **The eigenvalues are much more standardized in the 2nd run.**
   2. Check out the component profile plot. This plot is used to help provide interpretation of the new variables that PCA creates.