**Assignment #6: Principal Components in Predictive Modeling (80 points)**

**Data:** The data for this assignment is the stock portfolio data set. This data will be made available by your instructor.

Currently there is no up to date SAS reference for multivariate analysis and the new SAS SG graphical output other than the SAS User’s Manual. Since the SAS User’s Manual can be difficult for a new SAS user, we have provided some SAS information in the course supplements, and we will include more SAS example code and code snippets in the multivariate analysis portion of this course. Some code will be explicit and complete, and some code will only be pseudo code or code snippets. The student is expected to recognize the difference by the time that we reach this section of the class.

**Assignment Instructions:**

For this assignment we will use Principal Components Analysis as a method of dimension reduction and as a remedial measure for multicollinearity in Ordinary Least Squares regression.

1. Let’s begin with some data prep. Our raw data consists of daily closing stock prices for twenty stocks and a large-cap index fund from Vanguard (VV). We can use the log-returns of the individual stocks to explain the variation in the log-returns of the market index. We will explore this concept using both linear regression and principal components analysis.

Here is some code to get you started.

**data** temp**;**

set mydata.stock\_portfolio\_data**;**

**run;**

**proc** **sort** **data**=temp**;** by date**;** **run;** **quit;**

**data** temp**;**

set temp**;**

\* Compute the log-returns - log of the ratio of today's price

to yesterday's price;

\* Note that the data needs to be sorted in the correct

direction in order for us to compute the correct return;

return\_AA = log**(**AA/lag1**(**AA**));**

return\_BAC = log**(**BAC/lag1**(**BAC**));**

\* Continue to compute the log-returns for all of the stocks;

\* Name the log-return for VV as the response variable;

response\_VV = log**(**VV/lag1**(**VV**));**

**run;**

**proc** **print** **data**=temp**(**obs=**10);** **run;** **quit;**

1. Now let’s look at the correlations between the individual stocks and the market index.

\* We can use ODS TRACE to print out all of the data sets available to ODS for a particular SAS procedure.;

\* We can also look these data sets up in the SAS User's Guide in the chapter for the selected procedure.;

\*ods trace on;

ods output PearsonCorr=portfolio\_correlations**;**

**proc** **corr** **data**=temp**;**

\*var return: with response\_VV;

var return\_:**;**

with response\_VV**;**

**run;** **quit;**

\*ods trace off;

**proc** **print** **data**=portfolio\_correlations**;** **run;** **quit;**

1. SAS has two ‘types’ of data sets or data formats – long and wide. If you do not already know this difference, then we will illustrate it here. We can use PROC TRANSPOSE to transform one format to the other.

Note that the output data set portfolio\_correlations has much more output than we need. All that we want are the correlations between the returns.

**data** wide\_correlations**;**

set portfolio\_correlations **(**keep=return\_:**);**

**run;**

\* Note that wide\_correlations is a 'wide' data set and we need a 'long' data set;

\* We can use PROC TRANSPOSE to convert data from one format to the other;

**proc** **transpose** **data**=wide\_correlations out=long\_correlations**;**

**run;** **quit;**

**data** long\_correlations**;**

set long\_correlations**;**

tkr = substr**(**\_NAME\_,**8**,**3);**

drop \_NAME\_**;**

rename COL1=correlation**;**

**run;**

**proc** **print** **data**=long\_correlations**;** **run;** **quit;**

1. Can we visualize these correlations? When working with a large amount of predictor variables, it can be helpful to use visualizations instead of tables. As a visualization we will create a colored bar plot of the correlations.

\* Merge on sector id and make a colored bar plot;

**data** sector**;**

input tkr $ **1**-**3** sector $ **4**-**35;**

datalines**;**

AA Industrial - Metals

BAC Banking

BHI Oil Field Services

CVX Oil Refining

DD Industrial - Chemical

DOW Industrial - Chemical

DPS Soft Drinks

GS Banking

HAL Oil Field Services

HES Oil Refining

HON Manufacturing

HUN Industrial - Chemical

JPM Banking

KO Soft Drinks

MMM Manufacturing

MPC Oil Refining

PEP Soft Drinks

SLB Oil Field Services

WFC Banking

XOM Oil Refining

VV Market Index

**;**

**run;**

**proc** **print** **data**=sector**;** **run;** **quit;**

**proc** **sort** **data**=sector**;** by tkr**;** **run;**

**proc** **sort** **data**=long\_correlations**;** by tkr**;** **run;**

**data** long\_correlations**;**

merge long\_correlations **(**in=a**)** sector **(**in=b**);**

by tkr**;**

if **(**a=**1)** and **(**b=**1);**

**run;**

**proc** **print** **data**=long\_correlations**;** **run;** **quit;**

\* Make Grouped Bar Plot;

\* p. 48 Statistical Graphics Procedures By Example;

ods graphics on**;**

title 'Correlations with the Market Index'**;**

**proc** **sgplot** **data**=long\_correlations**;**

format correlation **3.2;**

vbar tkr / response=correlation group=sector groupdisplay=cluster datalabel**;**

**run;** **quit;**

ods graphics off**;**

1. Compute the principal components for the return data. How many principal components do you think that we should keep? Why? Later in the assignment we will use the first eight principal components. Why eight? We will also plot the first two principal components from the principal components analysis. When we plot them, we can see relationships in the data. Do we see any groupings (or clusters) in the plot of the first two principal components? Any surprises?

**data** return\_data**;**

set temp **(**keep= return\_:**);**

\* What happens when I put this keep statement in the set statement?;

\* Look it up in The Little SAS Book;

**run;**

**proc** **print** **data**=return\_data**(**obs=**10);** **run;**

ods graphics on**;**

**proc** **princomp** **data**=return\_data out=pca\_output outstat=eigenvectors plots=scree**(**unpackpanel**);**

**run;** **quit;**

ods graphics off**;**

\* Notice that PROC PRINCOMP produces a lot of output;

\* How many principal components should we keep?;

\* Do the principal components have any interpretability?;

\* Can we display that interpretability using graphics?;

**proc** **print** **data**=pca\_output**(**obs=**10);** **run;**

**proc** **print** **data**=eigenvectors**(**where=**(**\_TYPE\_='SCORE'**));** **run;**

\* Display the two plots and the Eigenvalue table from the output;

**data** pca2**;**

set eigenvectors**(**where=**(**\_NAME\_ in **(**'Prin1','Prin2'**)));**

drop \_TYPE\_ **;**

**run;**

**proc** **print** **data**=pca2**;** **run;**

**proc** **transpose** **data**=pca2 out=long\_pca**;** **run;** **quit;**

**proc** **print** **data**=long\_pca**;** **run;**

**data** long\_pca**;**

set long\_pca**;**

format tkr $**3.;**

tkr = substr**(**\_NAME\_,**8**,**3);**

drop \_NAME\_**;**

**run;**

\*\* Add sector to this dataset;

**data long\_pca;  
set long\_pca;  
merge long\_pca(in=x) sector(in=y);  
by tkr;  
if x=1;  
run;**

**proc** **print** **data**=long\_pca**;** **run;**

\* Plot the first two principal components;

ods graphics on**;**

**proc** **sgplot** **data**=long\_pca**;**

scatter x=Prin1 y=Prin2 / datalabel=tkr group=sector**;**

**run;** **quit;**

ods graphics off**;**

1. Now let’s move towards using principal components in regression modeling. Let’s split the data into a train and test data sets. Again, we will use the standard SAS trick of augmenting our data with a new response variable so that we can keep all of our data in one data set.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\* Create a training data set and a testing data set from the PCA output;

\* Note that we will use a SAS shortcut to keep both of these 'datasets' in one data set that we will call cv\_data (cross-validation data). ;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**data** cv\_data**;**

merge pca\_output temp**(**keep=response\_VV**);**

\* No BY statement needed here. We are going to append a column in its current order;

\* generate a uniform(0,1) random variable with seed set to 123;

u = uniform**(123);**

if **(**u < **0.70)** then train = **1;**

else train = **0;**

if **(**train=**1)** then train\_response=response\_VV**;**

else train\_response=**.;**

**run;**

**proc** **print** **data**=cv\_data**(**obs=**10);** **run;**

1. Fit a regression model using all of the individual stocks with train\_response as the response variable. Have SAS output the predicted values and the variance inflation factors. Perform a goodness-of-fit analysis for this model and analyze the variance inflation factors. Does this model have a multicollinearity problem? Compute the mean square error and the mean absolute error for the training and testing samples.
2. Fit a regression model using the first eight principal components using train\_response as the response variable. Have SAS output the predicted values and the variance inflation factors. Perform a goodness-of-fit analysis for this model and analyze the variance inflation factors. Does this model have a multicollinearity problem? Compute the mean square error and the mean absolute error for the training and testing samples. Compare these two models, which model fits better.

**Assignment Document:**

All assignment reports should conform to the standards and style of the report template provided to you. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary results or information. The document should be submitted in pdf format. Name your file Assignment6\_LastName.pdf.