Assignment #8: Multivariate Analysis (30 points)

Data Directory: Data can be accessed on the SAS OnDemand server using this libname statement.

```
libname mydata        '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
```

Data Set: mydata.factor_data

Data Description: Simulated factor data. The z-variables are the true or latent factors, the x-variables are the observed predictor variables, and Y is the response variable. Note that there are five z-variables (Z1, Z2, Z3, Z4, and Z5) and each z-variable has three observed x-variables (X1_1, X1_2, and X1_3 are observed variables for the latent variable Z1).

Assignment Instructions:

Part 1: An Initial Correlation Analysis

Begin your analysis by writing a macro function `%corr_matrix(k=1)` to compute the correlation matrix between each z-variable and its observed x-variables. Note that a macro variable needs to be closed by a period when it is used in a naming convention, e.g. x&k._1 will be interpreted as x1_1 when k=1. Comment on the correlation structure for each group of variables.

Part 2: Compute the Principal Components

Standardize the variables using PROC STANDARD and create two working data sets zdata and xdata. Why do we need to standardize the data before we perform any type of "components" or "factor" analysis?

```
proc standard data=temp mean=0 std=1 out=temp_std;
var z1 z2 z3 z4 z5
      x1_1  x1_2  x1_3
      x2_1  x2_2  x2_3
      x3_1  x3_2  x3_3
      x4_1  x4_2  x4_3
      x5_1  x5_2  x5_3 ;
run;


data zdata;
      set temp_std;
      keep y z1 z2 z3 z4 z5;
run;


data xdata;
      set temp_std;
      drop y z1 z2 z3 z4 z5;
run;
```

Compute the principal components using the x-variables using these options.

```
ods graphics on;
proc princomp data=xdata out=xdata_pca outstat=pca_stats plots=(scree);
run;
ods graphics off;
```

How many principal components do you recommend keeping? How are you making this decision? What is the correlation structure between the principal components?

Part 3: Examples of Problems with Factor Analysis

Now let's consider using factor analysis. In Part 2 we applied principal components to our x-variables. PCA will always produce a set of components that we can use to reduce the dimension of the model and improve multicollinearity deficiencies. Factor analysis will not always produce a usable set of factors. Consider these two examples.

Example 1:

```
ods graphics on;
proc factor data=xdata method=ml out=xdata_ml outstat=ml_stats
mineigen=0 priors=max nfactors=15 score scree ;
run;
ods graphics off;
```

What method of factor analysis is performed in Example 1, and what does the error message mean? (Hint: You will need to consult the chapter on PROC FACTOR in the pdf version of the SAS User's Guide.)

Example 2:

```
ods graphics on;
proc factor data=xdata method=uls heywood out=xdata_uls
outstat=uls_stats
mineigen=0 priors=max nfactors=15 score scree ;
run;
ods graphics off;
```

What method of factor analysis is performed in Example 2, and what does the error message mean? (Hint: You will need to consult the chapter on PROC FACTOR in the pdf version of the SAS User's Guide.)

Part 4: A Factor Analysis that "Worked"

Run the following factor analysis with these options.

```
ods graphics on;
proc factor data=xdata method=uls heywood out=xdata_uls outstat=uls_stats
      mineigen=0 priors=max nfactors=5 score scree ;
run;
ods graphics off;
```

This factor analysis "worked" in the sense that SAS did not return any errors. However what is wrong with this factor analysis and what option did we employ to allow SAS to return an answer? (Hint: You will need to consult the chapter on PROC FACTOR in the pdf version of the SAS User's Guide.)

Now let's rotate the factor components using the *VARIMAX* rotation. What does the varimax rotation do?

```
ods graphics on;
proc factor data=xdata method=uls heywood rotate=varimax out=xdata_varimax
      outstat=varimax_stats
      mineigen=0 priors=max nfactors=5 score scree ;
run;
ods graphics off;
```

Part 5: A Correlation Analysis of the Components

Examine and comment the correlation structure of the components produced by each of our three examples. Which methods produced orthogonal components and which did not? Do some methods always produce orthogonal components? Can some methods produce orthogonal components sometimes, but non-orthogonal components other times?

```
proc corr data=xdata_pca;
var prin1 prin2 prin3 prin4 prin5;
run;

proc corr data=xdata_uls;
var factor1 factor2 factor3 factor4 factor5;
run;

proc corr data=xdata_varimax;
var factor1 factor2 factor3 factor4 factor5;
run;
```

Part 6: Fit Regression Models

Let's fit regression models to our principal components and our factors and compare the results against the model selected by backwards variable selection. (The sample code will create the data set and fit the regression models.) In this example have we improved our model in any way by using PCA or FA?

Will using PCA or FA always provide us with better modeling results? When can these methods be most beneficial?

```
* Create a single data set for fitting regression models;
data pca_data;
      set xdata_pca (keep= prin1 prin2 prin3 prin4 prin5);
      id_nbr = _n_;
run;

data varimax_data;
      set xdata_varimax (keep= factor1 factor2 factor3 factor4 factor5);
      id_nbr = _n_;
run;

data zdata;
      set zdata;
      id_nbr = _n_;
run;

proc sort data=pca_data; by id_nbr; run;
proc sort data=varimax_data; by id_nbr; run;
proc sort data=zdata; by id_nbr; run;

data model_data;
      retain id_nbr;
      merge zdata pca_data varimax_data;
      by id_nbr;
run;

* True model;
proc reg data=model_data;
model Y = z1 z2 z3 z4 z5 / vif;
run; quit;

* PCA model;
proc reg data=model_data;
model Y = prin1 prin2 prin3 prin4 prin5 / vif;
run; quit;

* Factor model;
proc reg data=model_data;
model Y = factor1 factor2 factor3 factor4 factor5 / vif;
run; quit;


proc reg data=temp;
model Y =   x1_1  x1_2  x1_3
      x2_1  x2_2  x2_3
      x3_1  x3_2  x3_3
      x4_1  x4_2  x4_3
      x5_1  x5_2  x5_3
      / selection=backward vif;
run; quit;
```

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.

Treat each Part of the assignment as a separate subsection in your report. Part 1 will contain the results and discussion of the initial correlation analysis. Part 2 will contain the results from the principal components analysis. Part 3 will explain why the two PROC FACTOR examples failed. Part 4 will contain the results from the VARIMAX rotation and a discussion of the VARIMAX rotation. Part 5 will contain a correlation analysis of the components, and Part 6 will contain the output and a comparison of three regression models. The document should be submitted in pdf format.