Perform principal component analysis

We perform PCA on the explanatory variables by putting our dataset into the PRINCOMP function (appendix) to create the Principal Components (PC) that are uncorrelated with each other.

In order to select the number of components to keep, there are three methods that will be employed: PC with an eigenvalue greater than 1.0 (first school of thought), result from parallel analysis (second school of thought) and the rule of the thumb decision to keep components that total up to 80-90% variance.

First School of Thought: Choosing Principal Components with eigenvalue greater than one

It can be seen in the table below that only the first principal component (PC1) has an eigenvalue greater than 1 and will be kept in the model. However, only using one component would only explain 47.31% of the total variance, which is not enough for a reliable predictive accuracy. Thus, only using the first school of thought is not suitable and it is necessary to investigate if the second school of thought may be more suitable.

Second School of Thought: Choosing Principal Component using Parallel Analysis

Since SAS studios does not have a specific function for parallel analysis it is necessary to employ the factor procedure using the principal option. The result provides the factors that we can keep with an asterisk in the specific PC's row, as that PC's observed eigenvalue is greater than the corresponding 95th percentile.
In the table below, the parallel analysis agrees with the first school of thought in keeping only PC1.

Rule of Thumb

Since both schools of thought have determined that only PC1 should be kept, but has been determined to be too strict and retain only 47.31% of the variance, the rule of thumb will be used. The number of components that will be kept depends on the data and how big of an R-square needed. However, an acceptable R-square, or rather the variance, should come to a total value of 80-90% of the variance explained. (Shalizi, 2012, pp. 353-354)

It can be seen in figure 20 that the first three components have a cumulative value of 0.85 (85%). Hence, the first three components collectively explain 85% of the total variance of the original variables.

Scree Plot

To support this decision to keep the three components, the Scree plot can be used to visually identify which components should be kept. According to Shalizi (2012, pp. 354), folklore recommends looking for the "base of the cliff" or "elbow" in the graph. In the Scree Plot below, there is a clear elbow after the third component, which confirms that three components explains the majority of the original dataset's variance.

Run regression model using principal components

Using the three principal components in the regression function, the result is a regression model with three principal component variables that is statistically significant ($p < 0.0001$) with an F-value = 22.60 and explains 19.95% of the variance in the model (r-square = 0.1995).

Determine the usefulness of the model

In order to determine the usefulness of the PCR model, its performance is compared to the original multiple regression model. The PCR model uses three components that explains 85% of the variance in the predictors and yields R-square = 0.1995. The original regression model with all five variables yielded R-square = 0.2114. The original model explains a greater proportion of the variance in drinks consumed, thus the risk of a loss of information outweighs the benefit of mitigating multicollinearity. According to Kiers and Smilde (2006), there will be an improved performance of the prediction rule if the method (PCR) optimizes not only variance explained in the explanatory variables but also in the predictor variables . Hence, for this dataset, the original multiple regression model is preferable and will be used despite the multicollinearity present in order to prioritize prediction accuracy.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.36555394	1.38875466	0.4731	0.4731
2	0.97679928	0.06319366	0.1954	0.6685
3	0.91360562	0.41992606	0.1827	0.8512
4	0.49367956	0.24331797	0.0987	0.9499
5	0.25036159		0.0501	1.0000

Table 15: Eigenvalues of principal components

Parallel Analysis: NSims=1000 Seed=1234		
	Observed Eigenvalue	Simulated Crit Val
1	2.3656	1.2452*
2	0.9768	1.1257
3	0.9136	1.0413
4	0.4937	0.9741
5	0.2504	0.9038
* Retained Dimension (Obs > Crit, alpha=0.05)		

Table 16: Principal components in parallel analysis

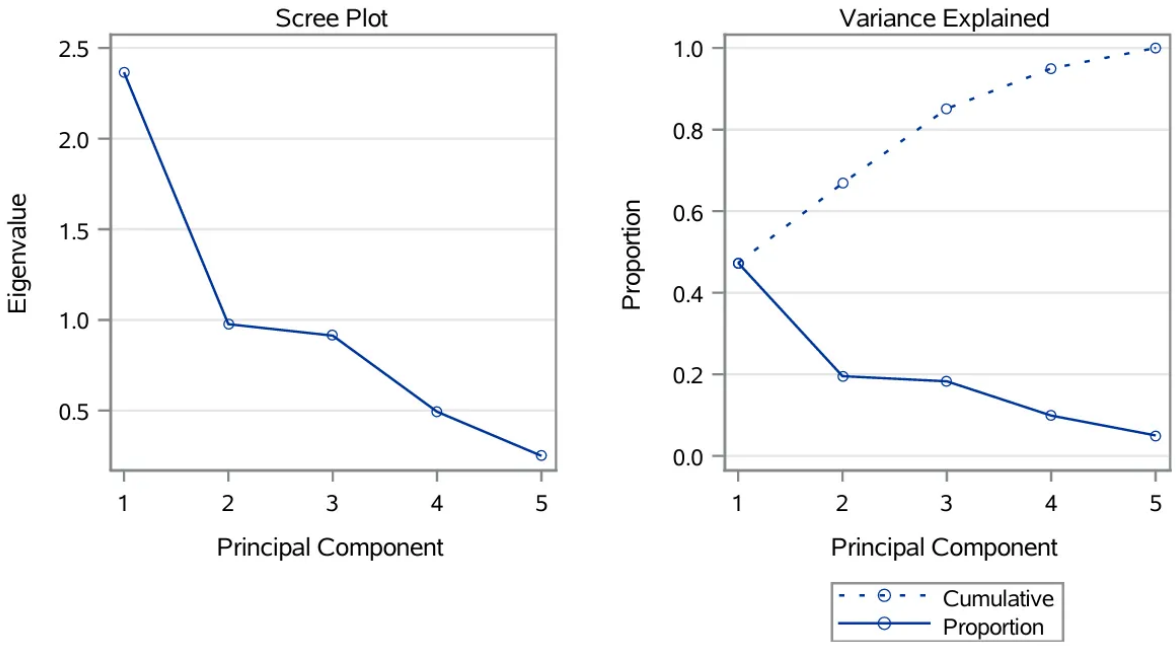

Figure 7: Scree Plot

Root MSE	2.95386	**R-Square**	0.1995
Dependent Mean	3.40942	**Adj R-Sq**	0.1907
Coeff Var	86.63811		

Table 17

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	3.40942	0.17780	19.18	<.0001	0
z1	1	0.76145	0.11581	6.57	<.0001	1.00000
z2	1	-0.35373	0.18023	-1.96	0.0507	1.00000
z3	1	0.84800	0.18636	4.55	<.0001	1.00000

Table 18: Principal components and the variance they explain

Code

Principal Component Regression Code

```
/* Assigning Library to Dataset */
libname mydata '/home/u64159750/My folders/Group Assignment/';

/* Deriving principal components from dataset using first school
of thought*/
ODS GRAPHICS ON;
proc princomp data=mydata.group_1_train
out=alcohol_blood_test prefix=z outstat=test1 plots=all;
var mcv alkphos sgpt sgot gammagt;
run;
ODS GRAPHICS OFF;

/* Deriving principal components from dataset using parallel
analysis */
proc factor data=mydata.group_1_train
method=principal parallel(nsims=1000 seed=1234);
var mcv alkphos sgpt sgot gammagt;
run;

/* Running regression model with principal components */
proc reg data=alcohol_blood_test;
model drinks = z1 z2 z3/VIF;
title 'Principal Component Regression with drinks as response';
run;
```