**Instructions:**

**There is a total of four (4) multi-part questions, with point values noted for each question.**

**Please show your calculations, or the details of your program(s) for each problem. You must supply the SAS program, and the program should be commented so that each step is clearly explained.**

**Combine all your answers/files into a single zipped file and post the zipped file to “HW\_Final\_new” in CANVAS.**

**Problem 1 - (25 points)**

Normalize and perform PCA analysis on the “PCA\_data” dataset on CANVAS. Perform the following:

* Create an output dataset “out\_PCA” containing the results of your PCA transformation for X1 to X6

Ans in the code.

* What is the sum of the variances of the normalized data?

**Ans:** 1.000 + 1.00 + 1.00 + 1.00 + 1.00 + 1.00 = **6**

* What is the sum of the eigenvalues?

**Ans:** 3.169 + 1.006 + 0.762 + 0.552 + 0.317 + 0.191 = **5.997**

* If you wanted to reduce the number of dimensions, how many principal components would you choose? Why?

**Ans:** We would choose **2** principle components if we wanted to reduce the number of dimensions. The Criteria for reducing the number of dimensions is based on Scree Plot which is created with respect to Eigenvalue table of co-relation matrix. We must choose Eigen value which is greater than 1. So only X1 and X6 have eigen values greater than 2.

* What would be the number of principal components you would choose if you want to capture around 90% of the variabilities?

**Ans:** The number of principal components I would choose if I wanted to capture around 90% of the variabilities would be **4**

* What would be the first principal component for a record with the normalized values of X1=.5, X2=.5 X3=.5 X4=.5 X5=.5 X6=.5?

**Ans:** X1=3.169\*0.5 = 1.5845

X2=1.006 \*0.5 =**0.503**

X3=0.762\*0.5 = **0.381**

X4=0.552\*0.5 =**0.276**

X5=0.317\*0.5 =**0.158**

X6=0.191\*0.5 =**0.095**

**Problem 2 - (25 points)**

Use the “out\_PCA” dataset above to establish a regression model using y as the dependent variable and Prin1 to Prin6 as independent variables. Use the following selection models and answer the corresponding questions:

1. Selection=Stepwise. What is the final model? Is it a good model?

**Ans:** The final model is good as all the variables are significant and the p-value too. The residual graph displays a 45-degree line of residuals distributed evenly. The results are shown in the code

1. Selection=MaxR. What is the best model for two predictors? Is that a good model? Why

**Ans:** MaxR gives a model according to the R2 values. For two predictors, prin1 and prin4 are two independent variables considered in the model. This is a good model since it has a normal distribution with no pattern, significant intercepts, R^2 values and VIF values are under 5,

meaning the variables are not dependent on each other.

The following two questions use the SAS dataset “Admission” on CANVAS. The dataset shows whether an applicant has been admitted to a college (admit=1), or not (admit=0). There are three predictors. The variables gre and gpa are continuous. The variable rank is categorical and takes on the values 1 through 4.

**Problem 3 - (25 points)**

* Establish a logistic regression model to predict admission (admit=1) using rank as a predictor. Using rank=1 as your base answer the following questions:
* Is this a good model?

**Ans: yes it’s a good model**

Assuming the model is a good model

* What are the odds of admission for rank=1?

**Ans: 1.89**

* What is the P(admit=1/rank=1)?

**Ans: 0.654**

* What is P(admit=1/rank=2)

**Ans: 0.5126**

* What is the odds ratio of rank=2 over rank=1?

**Ans: 0.5564**

**Problem 4 - (25 points)**

* Use hierarchical (method=average) and k means clustering methods to create two clusters for the Admission dataset using gre and gpa as clustering variables.

Ans: Done in the code

* Do Applicant 1008 and 1009 belong to the same clusters? Please explain.

**Ans:** Applicant 1008 and 1009 belong to the same cluster in Hierarchical clustering but do not belong to the same cluster in K-means Clustering.

Datasets: PCA\_data, Admission