**STA280 Unsupervised Learning**

**Homework on Principal Components Analysis**

***Submission by Mahika Bansal (mb62835)***

1. **Create a SAS dataset called WORK.RATINGS that contains the data in the job ratings.txt file. Assign the SAS names JOB, KNOWHOW, PROBLEM\_SOLVING, ACCOUNTABILITY, SALARY, respectively, to the five variables as they appear from left to right in the file. Extract the principal components of the three dimensions that were rated by the management consulting firm. Use the default (standardized) version of the extraction. Your answer for question 1 is your SAS code only:**

Ans. \*q1 - dataset creation and extraction of principal components;

data work.ratings;

infile '/home/u60740963/job ratings.txt' firstobs=2;

input JOB KNOWHOW PROBLEM\_SOLVING ACCOUNTABILITY SALARY;

cards;

run;

proc standard data=work.ratings mean=0 std=1 out=zratings;

var KNOWHOW PROBLEM\_SOLVING ACCOUNTABILITY;

run;

proc princomp data=zratings out=ratings\_PC;

var KNOWHOW PROBLEM\_SOLVING ACCOUNTABILITY;

run;

1. **This question verifies the basic property of principal components transformations.**

**a) Write the equations of the principal components of the PCA in question 1.**

**b) Verify that the principal component transformation in question 1 is an orthonormal rotation of the (standardized) original three dimensions by showing that the rotation matrix satisfies the definition of an orthonormal transformation.**

Ans. a) The principal components can be written as:

The principal components can be written as:

**PC1 = 0.576251\*KNOWHOW+0.584343\*PROBLEM\_SOLVING+0.571383\*ACCOUNTABILITY**

**PC2 = -0.618121\*KNOWHOW+-0.145758\*PROBLEM\_SOLVING+0.772451\*ACCOUNTABILITY**

**PC3 = 0.53466\*KNOWHOW+-0.79831\*PROBLEM\_SOLVING+0.277201\*ACCOUNTABILITY**

This can be derived from the results from SaS program with the respective Eigenvectors:

Graphical user interface, application

Description automatically generated

b) The orthornormality can be verified by computing dot product and length of the principal components. These conditions were satisfied as, the dot products were ~0 and vectors were of unit length, as we can see in the snapshots given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **DotProduct-Components** | **PC1,PC2** | **PC2,PC3** | **PC1,PC3** |
| Result | -1.4163E-07 | -3.15229E-07 | -5.61687E-07 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Components** | **PC1** | **PC2** | **PC3** |
| Result | 1.000000245 | 0.999999756 | 1.000000283 |

1. **This question partially verifies the geometry-preserving property of principal components transformations.**

**a) Rotate the first two jobs in the text file by calculating their principal component scores.**

**b) The rotated scores for the two jobs in part (a) are each a vector of three scores. Verify that the lengths of these two vectors are the same as the lengths of the original (but standardized) ratings vectors of the two jobs.**

**c) Verify that the angle between these two rotated vectors is the same as the angle between the original unrotated vectors.**

Ans. Based on the calculations,



The length and the angle between the vectors seems to be equal.

1. **Obtain the principal components scores for all 67 jobs. Calculate the variances of the three sets of scores and verify that the variances are equal to the eigenvalues of the PC transformation.**

Ans. After calculating the variance of 3 components for the 67 obs, and comparing it with eigenvalues of the PCs:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Var-PC1** | | **Var-PC2** | | **Var-PC3** | |
| **2.908082561** | | **0.08369733** | | **0.008221497** | |
| **Eigenvalues of the Correlation Matrix** | | | | | |
|  | **Eigenvalue** | | **Difference** | **Proportion** | **Cumulative** |
| **1** | 2.90808114 | | 2.82438377 | 0.9694 | 0.9694 |
| **2** | 0.08369737 | | 0.07547588 | 0.0279 | 0.9973 |
| **3** | 0.00822149 | |  | 0.0027 | 1.0000 |

The variance and eigenvalues seem to be almost equal to each other.

1. **Find the regression equation that results from regressing PRIN1 on the three ratings knowhow, problem\_solving, and accountability after the ratings have been standardized and without an intercept.2 Are you surprised by the equation?**

Ans.

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0 | **R-Square** | 1.0000 |
| **Dependent Mean** | 3.24782E-16 | **Adj R-Sq** | 1.0000 |
| **Coeff Var** | 0 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **KNOWHOW** | 1 | 0.57625 | 0 | Infty | <.0001 |
| **PROBLEM\_SOLVING** | 1 | 0.58434 | 0 | Infty | <.0001 |
| **ACCOUNTABILITY** | 1 | 0.57138 | 0 | Infty | <.0001 |

The regression coefficients and eigenvectors for Prin 1 are same, which makes sense.

1. **Find the regression equation that results from regressing (standardized) KNOWHOW on the three principal components without an intercept. Are you surprised by the equation?**

Ans.

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0 | **R-Square** | 1.0000 |
| **Dependent Mean** | 3.67865E-16 | **Adj R-Sq** | 1.0000 |
| **Coeff Var** | 0 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Prin1** | 1 | 0.57625 | 0 | Infty | <.0001 |
| **Prin2** | 1 | -0.61812 | 0 | -Infty | <.0001 |
| **Prin3** | 1 | 0.53466 | 0 | Infty | <.0001 |

The eigenvectors of Knowhow are same as the regression coefficients, which should be the case.

1. **Write the loadings matrix, structured with components as columns and variables as rows. Using the loadings matrix, try to interpret meanings for the three principal components**.

Ans.

| **Pearson Correlation Coefficients, N = 67 Prob > |r| under H0: Rho=0** | | | |
| --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** |
| **KNOWHOW** | 0.98269  <.0001 | -0.17883  0.1476 | 0.04848  0.6968 |
| **PROBLEM\_SOLVING** | 0.99648  <.0001 | -0.04217  0.7347 | -0.07238  0.5605 |
| **ACCOUNTABILITY** | 0.97439  <.0001 | 0.22347  0.0691 | 0.02513  0.8400 |

Loading matrix shows that the variables are highly correlated to Prin1 owing to high correlation between the variables themselves and Prin1 explaining the relationship. Looking at the p-values, Prin2 and Prin3 seem to be insignificant whereas Prin1 can be considered a good measure for evaluating the jobs.

1. **How many principal components would you retain …**

**a) Using the Kaiser rule?**

**b) Using the Joliffe rule?**

**c) Using the 80% rule?**

Ans.

| **Eigenvalues of the Correlation Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 2.90808114 | 2.82438377 | 0.9694 | 0.9694 |
| **2** | 0.08369737 | 0.07547588 | 0.0279 | 0.9973 |
| **3** | 0.00822149 |  | 0.0027 | 1.0000 |

1. By Kaiser’s Rule: PCs with Eigenvalue > 1: Thus, PC1 selected
2. By Joliffe Rule: PCs with Eigenvalue > 0.7: Thus, PC1 selected
3. By 80% Rule: Cumulative upto 0.8: Thus, PC1 selected
4. **Find the regression equation that results from regressing salary on the three principal components with intercept. How much explanatory power do the three PCs collectively have in explaining salary?**

Ans.

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2082.09165 | **R-Square** | 0.9003 |
| **Dependent Mean** | 63929 | **Adj R-Sq** | 0.8955 |
| **Coeff Var** | 3.25686 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 63929 | 254.36798 | 251.33 | <.0001 |
| **Prin1** | 1 | 3557.20641 | 150.28811 | 23.67 | <.0001 |
| **Prin2** | 1 | 2316.12408 | 885.87403 | 2.61 | 0.0112 |
| **Prin3** | 1 | 3540.61136 | 2826.52316 | 1.25 | 0.2150 |

The R-Square and adjusted R-Square are ~0.9, thus the PCs have ~90% explanatory power in explaining the salary.

1. **In terms of explaining salary…**

**a) Which component is most useful? Second most useful? Least useful?**

**b) Is the usefulness of the PCs for explaining salary in the order PC1 > PC2 > PC3?**

**c) How much explanatory power is lost if one uses only PRIN1 to explain salary?**

Ans.

a) Based on the analysis, PC1 seems to be most useful, followed by PC2. P-value is not significant for PC3.

b) Yes, it’s in the order PC1 > PC2 > PC3

c)

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2182.26114 | **R-Square** | 0.8870 |
| **Dependent Mean** | 63929 | **Adj R-Sq** | 0.8852 |
| **Coeff Var** | 3.41355 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | 1 | 63929 | 266.60563 | 239.79 | <.0001 |
| **Prin1** | 1 | 3557.20641 | 157.51847 | 22.58 | <.0001 |

The R-square is ~88.7%, thus we have lost ~1.3% of explanatory power with regressing against only PC1.