**Unsupervised Learning HW1 – Principle Components Analysis**

Jake Johnson (jpj753)

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.

**data WORK.RATINGS;**

**input JOB KNOWHOW PROBLEM\_SOLVING ACCOUNTABILITY SALARY;**

**cards;**

**0 800 608 1056 102000**

**2 528 304 460 75740**

**3 460 264 460 75740**

**5 528 304 304 79172**

**4 460 264 400 70000**

**Rest of data….**

**RUN;**

**proc princomp data=Work.ratings out=PCA\_VALUES;**

**var KNOWHOW PROBLEM\_SOLVING ACCOUNTABILITY;**

**RUN;**

**Table

Description automatically generated**

1. Write PCA equations from question 1 and verify that they are orthonormal transformation of the original data. Computations to confirm orthonormal transformation in Excel.

**Prin1 = 0.576251\*KNOWHOW + 0.584343\*PROBLEM\_SOLVING + 0.571383\*ACCOUNTABILITY**

**Prin2 = -0.61812\*KNOWHOW – 0.14576\*PROBLEM\_SOLVING + 0.772451\*ACCOUNTABILITY**

**Prin3 = 0.53466\*KNOWHOW – 0.79831\*PROBLEM\_SOLVING + 0.22\*ACCOUNTABILITY**

**Verify orthonormal transformation by first verifying unit vector and then that it is orthogonal. Done in excel.**

1. 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. Computations done in excel.

|  |  |  |
| --- | --- | --- |
| **Verify Length** | |  |
|  | **Original** | **After Transform** |
| **Job1** | 9.17910675 | 9.179108851 |
| **Job2** | 3.753130402 | 3.753131274 |
|  |  |  |
| **Verify Angle** |  |  |
|  | **Original** | **After Transform** |
|  | 0.987002391 | 0.9870024 |

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. Computations done in Excel.

|  |  |  |
| --- | --- | --- |
| **Variance** |  |  |
| **Prin1** | **Prin2** | **Prin3** |
| 2.908083 | 0.083697 | 0.008221 |

Table

Description automatically generated

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. Are you surprised by the equation?

Regression Equation Below From Excel Calculation

**PRIN1 = -9.45104E-16 + 0.576251 \* KNOWHOW + 0.584343\*PROBLEM\_SOLVING + 0.571383 \* ACCOUNTABILITY**

**No I am not surprised by this equation because the coefficients are the same as the eigenvectors we calculated in SAS written out exactly the same in Question 2. The only difference is the intercept term but it is basically zero as it is 10 raised to the -16th.**

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?

**KNOWHOW = 5.58831E-16 + 0.576250931(PRIN1) -0.618121051(PRIN2) + 0.534659826 (PRIN3)**

**This is not surprising because the KNOWHOW regression is dependent on the three eigenvector values of the principal components as we saw in the SAS calculations and again the intercept term is basically 0.**

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.

Table

Description automatically generated

**Prin1 (Difficulty) – This component scores high when all three variables are also high meaning the job requires knowhow, problem solving, and accountability. A job that scores high on Prin1 is difficult and one that scores low on Prin1 is not very difficult.**

**Prin2 (Managerial) – This component is positive for accountability and negative for knowhow with problem solving not being much of a factor. This could be a manager towards the top of the chain who has accountability for issues but not much actual knowledge of technical skills**

**Prin3 (SME) – This component is slightly positive for knowhow and accountability and slightly negative for problem solving. This could be a subject matter expert who doesn’t have much responsibility problem solving for larger tasks.**

1. How many principal components would you retain … a) Using the Kaiser rule? b) Using the Joliffe rule? c) Using the 80% rule?

**Kaiser Rule – Just one, PRIN1 has an eigenvalue above 1**

**Joliffe Rule – Just one, PRIN1 has eigenvalue above 0.7**

**80% rule – Just one, PRIN1 captures 96.94% of the variance**

1. 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?

**Salary = 63929.33 + 3557.20517(Prin1) + 2316.128(Prin2) + 3540.612(Prin3)**

**The three principal components explain 90% of the variance in 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?

**a) Based on the p values from the regression, PC1 is the most useful, PC2 is the second most useful, and PC3 is the least useful.**

**b) Yes, this is the correct order.**

**c) If only Prin1 is used only about 1.4% of explanatory power is lost. Still 88.6% of the variance in salary is captured.**