Allie Touchstone

STA280: Unsupervised Learning

2/16/22

Homework 1: Homework on Principal Components Analysis

**Problem 1:**

\* Problem 1;

**PROC** **IMPORT** DATAFILE = "/home/u60719183/job ratings.xlsx"

OUT = WORK.RATINGS

DBMS = XLSX

REPLACE;

**RUN**;

**PROC** **PRINCOMP** DATA = ratings OUT = ratings\_pc;

VAR knowhow problem\_solving accountability;

**RUN**;

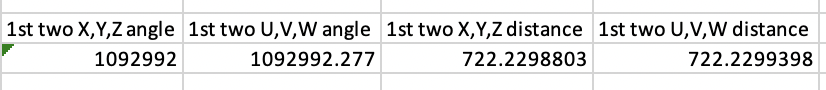
**Problem 2:**

Table

Description automatically generated

These computations show that the principal components (PC) are orthonormal. As well as that when the matrix is transposed the orthonormal holds.

**Problem 3:**



For the image, x,y,z are the original values and u,v,w are the standardized values. This shows that even when the jobs are rotated, the principal components hold.

**Problem 4:**

The first image is the calculated variances from the 67 jobs of the three sets of score, the second image is the eigenvalues that SAS computed. The eigenvalues and variances are the same (with minimal rounding errors)

| **Eigenvalues of the Correlation Matrix** | |
| --- | --- |
|  | **Eigenvalue** |
| **1** | 2.90808114 |
| **2** | 0.08369737 |
| **3** | 0.00822149 |

Text

Description automatically generated

**Problem 5:**

After standardizing in SAS, the regression equation we obtain from the following image is:

| **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 |

This equation is not surprising as the equation wants to linearly model the dependent variable, which is what this equation does.

**Problem 6:**

This is still an orthonormal transformation, so it was expected for this to be a linear equation.

| **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 |

**Problem 7:**

Based on these correlation coefficients, it appears that the first PC is highly correlated with all three variables. PC 1 can be assumed to be the higher-level jobs. PC 2 has negative correlations in everything expect accountability, these jobs are more like paper-pushers, they are needed but the main qualification of the job is showing up. PC 3 is tricky, there are not high correlations in either direction. These jobs are most likely unnecessary jobs that are mostly there to make other jobs easier, such as someone who passes out the mail, makes copies or sends faxes all day. They are helpful, but don’t heavily rely any of the variables we are looking at.

| **Pearson Correlation Coefficients, N = 67 Prob > |r| under H0: Rho=0** | | | |
| --- | --- | --- | --- |
|  | **Prin1** | **Prin2** | **Prin3** |
| **knowhow** | 0.98269 | -0.17883 | 0.04848 |
| **problem\_solving** | 0.99648 | -0.04217 | -0.07238 |
| **accountability** | 0.97439 | 0.22347 | 0.02513 |

**Problem 8:**

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

1. In the Kaiser rule you would only retain PC 1 as the Kaiser rule only considers PC with a eigenvalue greater than one.
2. The Joliffe rule would retain PC 1 as it’s the only PC with an eigenvalue greater than 0.7.
3. The 80% rule once again only retains PC 1 as its variance explains more than 80% of the data.

**Problem 9:**

The regression equation with an intercept and all three principal components for salary is:

**The REG Procedure**

**Model: MODEL1**

**Dependent Variable: salary**

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

| **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 explanatory power is 90.03%, this is given to us by the R-Square value.

**Problem 10:**

The R-Square reported for each of the out puts below shows that prin1 contains the most explanatory power for salary with 88.70%. This is followed by prin2 with 1.08% explanatory power, lastly prin3 has 0.25% explanatory power.

In short:

1. For salary: PC1 > PC2 > PC3
2. Yes, see above.
3. PC2 and PC3 have a combined explanatory power of 1.33% (1.08% + 0.25%), so that is how much is lost when we only use PC1.

**Prin 1:**

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

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

**Prin 2:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 6455.27135 | **R-Square** | 0.0108 |
| **Dependent Mean** | 63929 | **Adj R-Sq** | -0.0044 |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | 63929 | 788.63691 | 81.06 | <.0001 |
| **Prin2** |  | 1 | 2316.12408 | 2746.54442 | 0.84 | 0.4022 |

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Label** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | Intercept | 1 | 63929 | 791.95370 | 80.72 | <.0001 |
| **Prin3** |  | 1 | 3540.61136 | 8800.14647 | 0.40 | 0.6888 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 6482.42044 | **R-Square** | 0.0025 |
| **Dependent Mean** | 63929 | **Adj R-Sq** | -0.0129 |

**Prin 3:**