**STA380.18  
Homework on Principal Components Analysis**

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

0 460 264 400 66536

0 528 304 264 70000

7 460 230 264 68000

10 400 200 350 73140

7 400 175 230 66016

7 400 200 200 66016

5 400 175 200 71840

5 304 115 175 71580

2 264 100 175 65860

3 264 100 175 66432

10 230 100 132 64040

10 230 100 132 62610

7 230 87 132 65002

7 230 76 115 64001

5 230 76 115 66900

5 230 87 100 63000

5 230 87 100 63780

7 200 87 100 62000

7 200 76 100 61960

7 200 76 100 62012

7 200 76 87 62300

5 200 76 87 61960

7 200 66 87 61700

7 175 66 100 61440

2 175 57 100 62220

3 175 57 100 63260

7 175 57 100 59880

2 175 57 100 62480

3 175 57 100 63000

2 175 57 100 63260

3 175 57 100 62480

4 175 57 87 62480

7 175 57 87 61440

2 175 57 87 62064

3 175 57 87 61180

2 175 57 87 59100

3 175 57 87 59620

5 175 66 76 59880

5 175 66 76 60200

7 175 57 76 60140

7 175 57 76 61700

5 175 66 66 60000

7 152 50 87 60920

7 152 50 76 59100

3 152 50 76 61700

2 152 50 76 59880

3 152 50 76 61700

5 152 50 66 59360

5 152 43 66 60660

2 152 43 66 59984

2 152 43 66 60660

3 152 43 66 60920

3 152 43 66 60920

2 152 43 66 60920

3 152 43 66 60660

3 152 43 66 60660

7 152 43 66 58320

5 152 43 66 59360

2 152 43 66 60920

3 152 43 66 60920

4 152 43 66 60660

7 152 43 57 59880

RUN;

**proc** **princomp** data=WORK.RATINGS out=RATINGS\_PCA;

var knowhow problem\_solving accountability;

**RUN**;

2. This question verifies the basic property of principal components transformations.

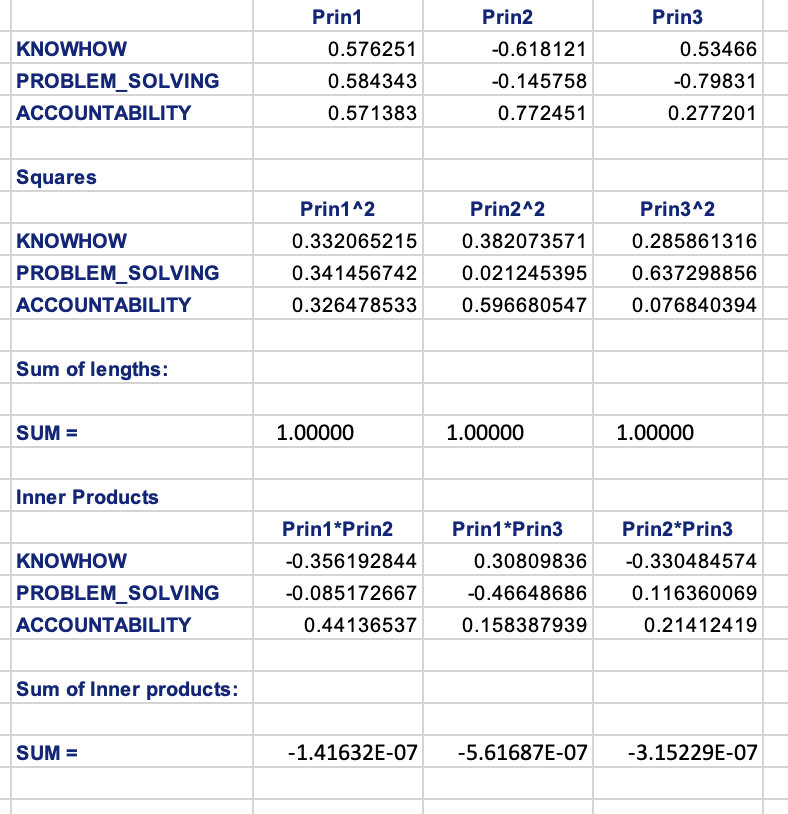
a) Write the equations of the principal components of the PCA in question 1.

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

*Prin2 =* -.618121(**KNOWHOW)** -.145758(**PROBLEM\_SOLVING)** + 0.772451(**ACCOUNTABILITY)**

*Prin3 =* 0.534660(**KNOWHOW)** -.798310(**PROBLEM\_SOLVING)** + 0.277201(**ACCOUNTABILITY)**

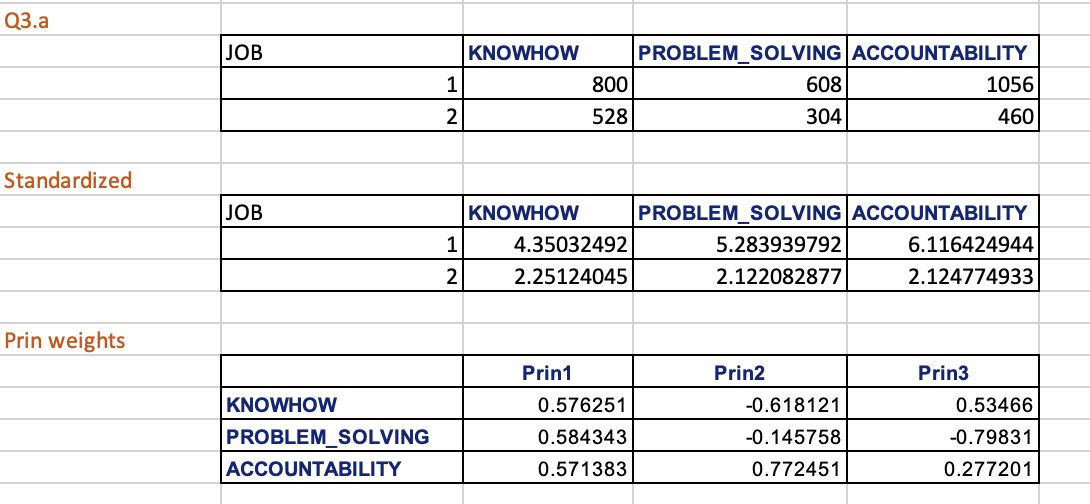
b) Verify that the principal component transformation in question 1 is an orthonormal rotation of the (standardized) original three dimensions by showing that the rotation matrix satisfies the definition of an orthonormal transformation.

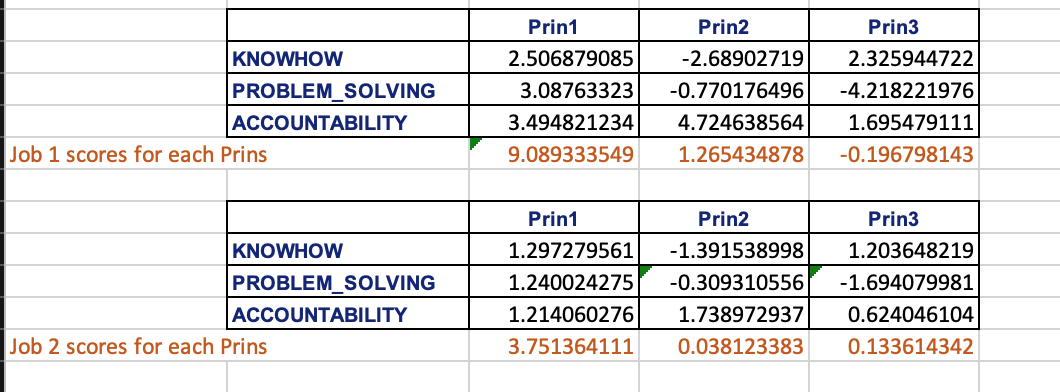
****

*We see that the sum of lengths is 1 and sum of inner products is 0 hence verifying principal component transformation in question 1 is an orthonormal rotation of the (standardized) original three dimensions.*

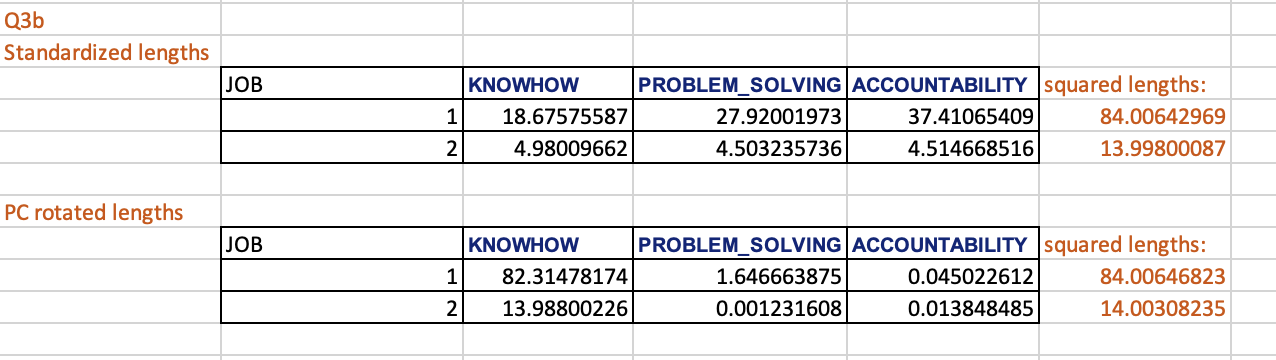
3. This question partially verifies the geometry-preserving property of principal components transformations.

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

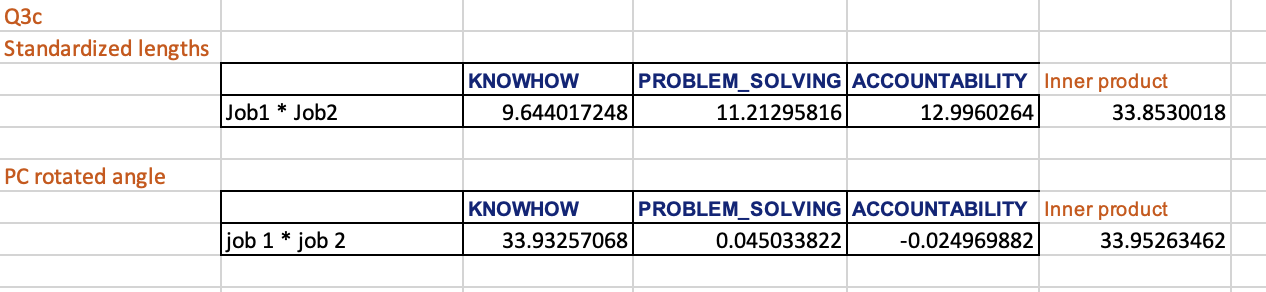




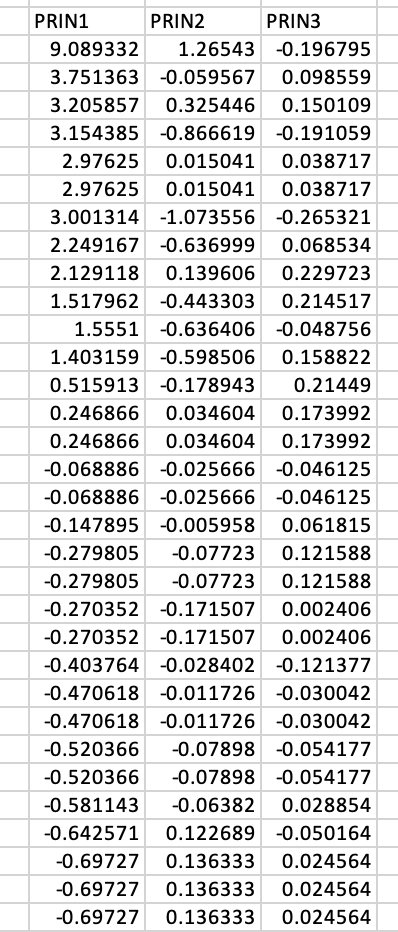
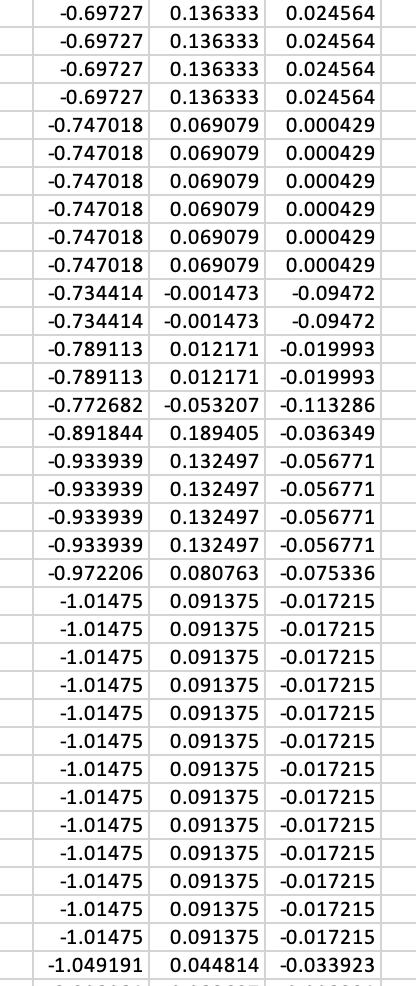
b)  The rotated scores for the two jobs in part (a) are each a vector of three scores. Verify that the lengths of these two vectors are the same as the lengths of the original (but standardized) ratings vectors of the two jobs.



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



4. 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.

** **

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variances =** | | **2.90808114** | | **0.08369737** | **0.00822149** |
| **Eigenvalue** | **Difference** | |
| **1** | 2.90808114 | |
| **2** | 0.08369737 | |
| **3** | 0.00822149 | |

*We see that the Eigen values of the PC transformation and variances of the three sets of scores are equal*

5. 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?

**proc** **stdize** data=RATINGS\_PCA out=RATINGS\_PCA\_STD;

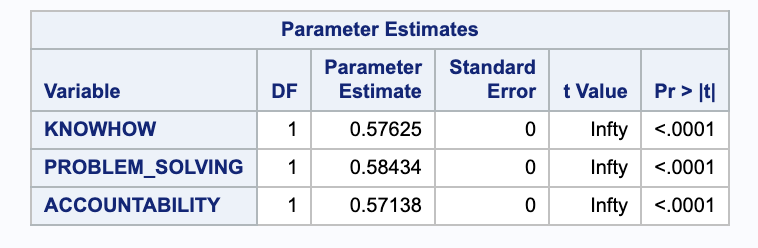
var knowhow problem\_solving accountability;

**RUN**;

**proc** **reg** data=RATINGS\_PCA\_STD;

model PRIN1 = knowhow problem\_solving accountability / noint;

**RUN**;



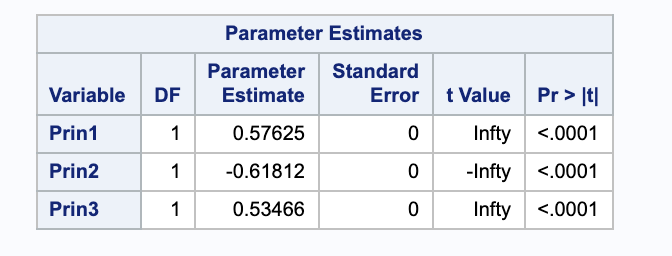
*The regression equation is the same as PRIN1 eigen vector values. This is not surprising since the PCs are linear functions of the ratings variables.*

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

**proc** **reg** data=RATINGS\_PCA\_STD;

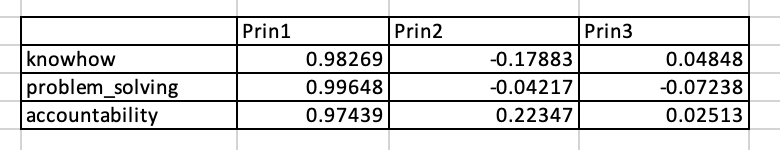
model knowhow = PRIN1 PRIN2 PRIN3 / noint;

**RUN**;

**

*The regression equation is the same as the reverse rotation from PRIN to ratings. This is not surprising since the reverse rotation expresses the ratings as a linear function of PRINs.*

7. 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.

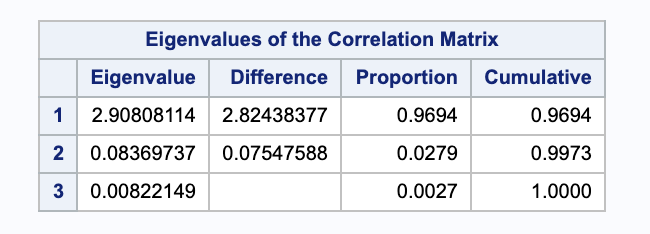


PRIN1 loads equally on all the ratings. Prin1 represents the most measure of a job’s requirements.

PRIN2 and PRIN3 however, have low correlations with all the other ratings. Jobs with high scores on PRIN2 have high accountability but low knowhow. This might mean these are managerial jobs. Jobs with high scores on PRIN3 have high knowhow but little problem solving. This might describe consultant jobs who have high domain knowledge.

8. How many principal components would you retain ...

1. a)  Using the Kaiser rule? -> PRIN1
2. b)  Using the Joliffe rule? -> PRIN1
3. c)  Using the 80% rule? -> PRIN1

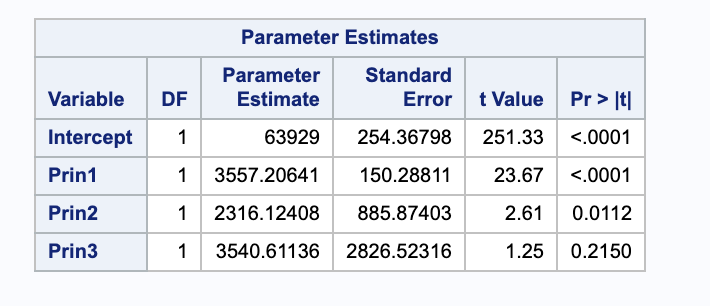


9. Find the regression equation that results from regressing **salary** on the three principal components with intercept. How much explanatory power do the three PCs collectively have in explaining **salary**?

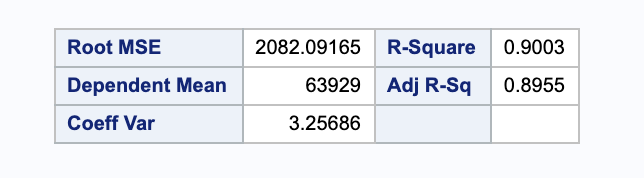
**proc** **reg** data=ratings\_PCA;

model salary = PRIN1 PRIN2 PRIN3;

**RUN**;



From the table below, we see that the R-square is 0.9003. Hence 90% of the variance is explained by the 3 PCs.



10. In terms of explaining **salary**...

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

Most useful in terms of explaining salary is PC1 since the t-stat value is the highest.

Second most useful is PC2 which has the second highest t-stat and the least useful is PC3 which has the least t-stat and has a high p-value.

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

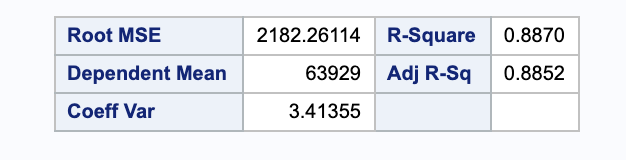
Yes, the usefulness of the PCs for explaining salary is in the order PC1>PC2>PC3

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

**proc** **reg** data=ratings\_PCA;

model salary = PRIN1;

**RUN**;



New R2 = 0.8870

Old R2 = 0.9003

Explanatory power lost = 0.9003 - 0.8870 = 0.0133