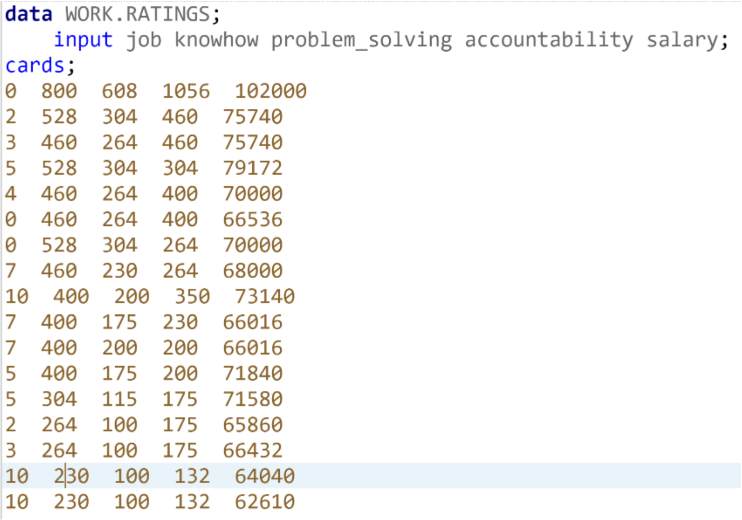
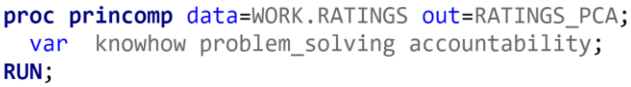
Learning Structures and Time Series Homework 2

- Akankshi Mody (am92786), MSBA Class of 2020

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
```



```
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.

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.

[Hint: You may find it helpful to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

a) Equations of principal components:

1. **Prin1 = 0.576251*knowhow + 0.584343*problem_solving + 0.571383 *accountability**
2. **Prin2 = -0.618121*knowhow – 0.145758*problem_solving + 0.772451*accountability**
3. **Prin3 = 0.534660*knowhow – 0.798310*problem_solving + 0.277201*accountability**

b) To Verify that the principal component transformation in question 1 is an orthonormal rotation of the (standardized) original three dimensions, we see if length of vectors is one and if they are orthogonal to each other.

	Eigenvectors			Eigenvectors^2			
	Prin1	**Prin2**	**Prin3**	**Prin1^2**	**Prin2^2**	**Prin3^3**	**Sum**
knowhow	0.576251	-0.61812	0.53466	0.33206522	0.382074	0.285861	1
problem_solving	0.584343	-0.14576	-0.79831	0.34145674	0.021245	0.637299	1

accountability	0.571383	0.772451	0.277201	0.32647853	0.596681	0.07684	1

	Dot Product
Prin1.Prin2	0
Prin2.Prin3	0
Prin1.Prin3	0

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.

[Hint: You may find it helpful to perform the computations in Excel. You may wish to submit Excel computations as your solution.]

Standardized knowhow	**standardized problem solving**	**standardized account - ability**	**Prin1**	**Prin2**	**Prin3**	**Original Length**	**Transformed Length**
4.35032492	5.283939792	6.116424944	9.089333	1.265434	-0.196798	9.179	9.179
2.25124045	2.122082877	2.124774933	3.751364	-0.059565	0.098557	3.753	3.753

	Cosine Angle
Original	9.247852312
Transformed	9.247849166

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.

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

Prin1	Prin2	Prin3		Variance
9.089332156	1.265430074	-0.19679536	**Prin1**	2.908081
3.75136318	-0.059567149	0.098558914	**Prin2**	0.083697
3.205857277	0.325445557	0.150108682	**Prin3**	0.008221
3.154384972	-0.866619424	-0.191059403		
2.976250274	0.015040836	0.038717021		
2.976250274	0.015040836	0.038717021		
3.001313636	-1.073555904	-0.26532051		
2.249167429	-0.636998841	0.068534492		
2.129117888	0.139605748	0.229723398		
1.517962482	-0.443303431	0.214517456		
1.55510038	-0.636406055	-0.048755753		
1.40315898	-0.598505791	0.158821625		
0.515912864	-0.178943117	0.214489617		
0.246865869	0.034604267	0.173992493		
0.246865869	0.034604267	0.173992493		
-0.068885648	-0.025666308	-0.046125216		

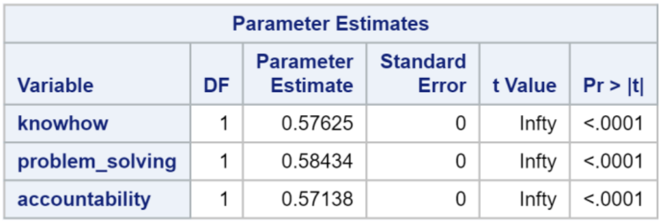
….

…..

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;
```



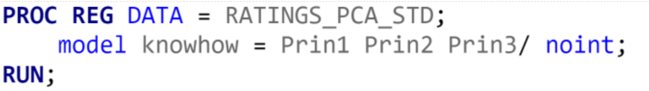
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
knowhow	1	0.57625	0	Infty	<.0001
problem_solving	1	0.58434	0	Infty	<.0001
accountability	1	0.57138	0	Infty	<.0001

Regression equation:

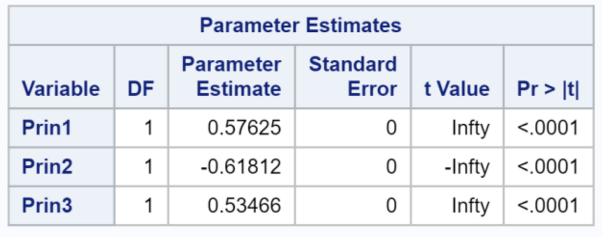
Prin1 = 0.57625*knowhow + 0.58434*problem_solving + 0.57138*accountability

This is not surprising since it is the same as the coefficients are the same as the eigenvector values.

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;
```



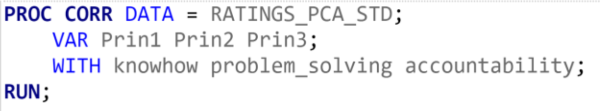
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Prin1	1	0.57625	0	Infty	<.0001
Prin2	1	-0.61812	0	-Infty	<.0001
Prin3	1	0.53466	0	Infty	<.0001

The Regression equation is as follows:

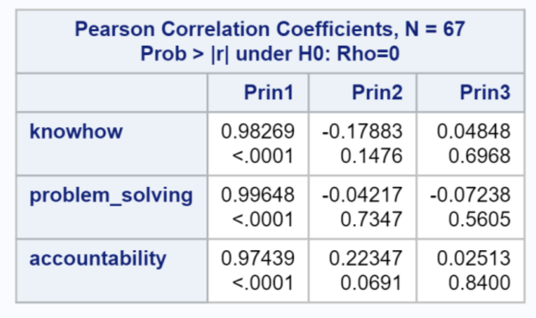
Knowhow = 0.57625 * Prin1 + -0.61812 * Prin2 + 0.53466 * Prin3

This is not surprising since the coefficients are from the eigenvalues matrix

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.



```
PROC CORR DATA = RATINGS_PCA_STD;
    VAR Prin1 Prin2 Prin3;
    WITH knowhow problem_solving accountability;
RUN;
```



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

The Pearson correlation coefficients represent the loading matrix for PCA transformation.

Interpretation:

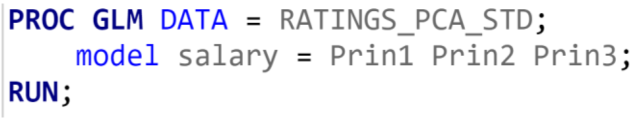
- **Prin1: PC1 has high correlation with all the 3 original variables. This means PC1 is a strong indicator of all the 3 features. This represents the job requirement will have high knowhow, problem_solving and accountability. Examples could be like Manager / Director.**

- **Prin2: PC2 has a low negative correlation with knowhow and problem_solving skills but high positive on accountability. Examples could be a fresh graduate / entry-level.**
- **Prin3: PC3 has a almost zero correlation with all 3 variables. Examples could be contract or outsourced jobs that do not need not have these skills.**

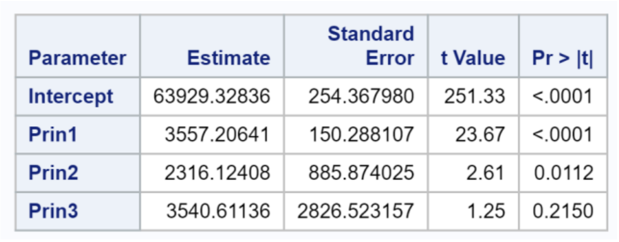
8. How many principal components would you retain …

a) Using the Kaiser rule?
Discard all the PCs with eigenvalue less than 1. Thus, we will only retain PC1.
b) Using the Joliffe rule?
Discard all the PCs with eigenvalue less than 0.7. Thus, we will only retain PC1.
c) Using the 80% rule?
PC1 explains 96.94% of the variance in the data, thus we will retain only PC1.

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**?



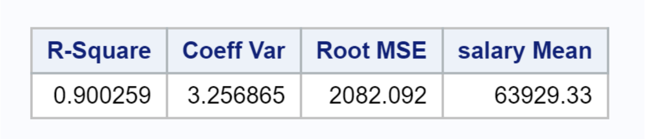
```
PROC GLM DATA = RATINGS_PCA_STD;
    model salary = Prin1 Prin2 Prin3;
RUN;
```



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	63929.32836	254.367980	251.33	<.0001
Prin1	3557.20641	150.288107	23.67	<.0001
Prin2	2316.12408	885.874025	2.61	0.0112
Prin3	3540.61136	2826.523157	1.25	0.2150

Regression Equation:
Salary = 63929 + 3557.20641 * Prin1 + 2316.12408 * Prin2 + 3540.61136 * Prin3

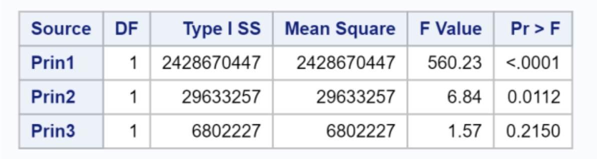
The three PCs collectively have 90.03% power in explaining salary (seen from the R-square)



R-Square	Coeff Var	Root MSE	salary Mean
0.900259	3.256865	2082.092	63929.33

10. In terms of explaining salary…
a) Which component is most useful? Second most useful? Least useful?

Type I SS is maximum for Prin1. Thus, Prin1 is the most useful in explaining the salary. Prin2 is the second most useful and Prin3 is the least useful.



Source	DF	Type I SS	Mean Square	F Value	Pr > F
Prin1	1	2428670447	2428670447	560.23	<.0001
Prin2	1	29633257	29633257	6.84	0.0112
Prin3	1	6802227	6802227	1.57	0.2150

b) Is the usefulness of the PCs for explaining salary in the order PC1 > PC2 > PC3?
Yes (From above question)

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

If we use only Prin1, the explanatory power which is retained will be

= R-square * Type I SS (Prin1) / (Sum of Type I SS for Prin1, Prin2, Prin3)
= 0.9 * 2428670447 / (2428670447 + 29633257 + 6802227)
= 88.7%

Hence, if we use only Prin1, the explanatory power that's lost
= 100 – 88.7
= 11.3%