MVA Final Fall 06. Open notes, no book. Points (out of 200) in parentheses.

1.(15) Suppose you have a data set with two dependent variables, Y_1 and Y_2, and three independent variables, X_1, X_2, and X_3. Explain, step-by-step, how to obtain a partial correlation matrix of Y_1 and Y_2, controlling for X_1, X_2, and X_3, using regression residuals.

Solution: Fit a regression equation relating Y_1 to X_1, X_2, and X_3 getting a column of residuals, $e_1 = Y_1 - (b_{01} + b_{11}X_1 + b_{21}X_2 + b_{31}X_3)$. Fit another regression equation relating Y_2 to X_1, X_2, and X_3 getting a column of residuals, $e_2 = Y_2 - (b_{02} + b_{12}X_1 + b_{22}X_2 + b_{32}X_3)$. Now, calculate the correlation coefficient between the columns e_1, e_2. This is the partial correlation between Y_1 and Y_2, controlling for X_1, X_2, and X_3.

2. A study of student responses to questions on the TTU graduate student survey included the following variables. All survey responses are in the 1 -5 Likert scale. The data are grad student ratings of

```
FacKnowledge    =  Faculty knowledge
FacTeaching     =  Faculty teaching
FacResInClass   =  Faculty use of research in class
FacOutsideClass =  Faculty availability outside of class
FacIntSuccess   =  International success of faculty
GrAdvisAvail    =  Graduate advisor availability
GrAdvAdmin      =  Graduate advisor administrative skill
GrAdvPersonal   =  Graduate advisor personal help
GrAdvCareerAdv  =  Graduate advisor help to advance career
GrAdvPlcmt      =  Graduate advisor help in placement
```

Here is the SAS code:

```
proc princomp data=isqs6348.pgs;
   var FacKnowledge FacTeaching FacResInClass FacOutsideClass FacIntSuccess
       GrAdvisAvail GrAdvAdmin GrAdvPersonal GrAdvCareerAdv GrAdvPlcmt; run;
```

Here is some selected output:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	5.85957806	4.16341342	0.5860	0.5860
2	1.69616465	1.18571670	0.1696	0.7556
3	0.51044795	0.05013162	0.0510	0.8066
4	0.46031633	0.05371107	0.0460	0.8527
5	0.40660526	0.12012661	0.0407	0.8933
6	0.28647865	0.01403389	0.0286	0.9220
7	0.27244476	0.04491617	0.0272	0.9492
8	0.22752859	0.06433406	0.0228	0.9720
9	0.16319253	0.04594731	0.0163	0.9883
10	0.11724522		0.0117	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
FacKnowledge	0.273091	0.395137	0.432351	0.411725	-.302881
FacTeaching	0.296674	0.387501	0.249367	0.158671	-.052065
FacResInClass	0.282866	0.345842	0.244605	-.421583	0.679179
FacOutsideClass	0.299664	0.280972	-.675718	-.000788	0.161737
FacIntSuccess	0.313763	0.267767	-.417742	-.130137	-.445112
GrAdvisAvail	0.323535	-.295559	-.090535	0.439159	0.290764
GrAdvAdmin	0.336762	-.316461	0.000871	0.297661	0.178535
GrAdvPersonal	0.343815	-.305471	-.013451	0.049178	-.013682
GrAdvCareerAdv	0.349947	-.293464	0.119199	-.250745	-.186243
GrAdvPlcmt	0.332252	-.237746	0.193309	-.513697	-.262092

Eigenvectors

	Prin6	Prin7	Prin8	Prin9	Prin10
FacKnowledge	0.118130	0.552144	-.011526	0.043147	-.017159
FacTeaching	-.480731	-.654390	0.063739	-.088827	0.021777
FacResInClass	0.296804	0.043306	-.106793	-.012448	0.002186
FacOutsideClass	-.451318	0.360697	0.072077	0.049820	-.083873
FacIntSuccess	0.588544	-.294085	-.060778	0.023290	0.056380
GrAdvisAvail	0.250562	-.053508	0.531952	-.407311	0.088600
GrAdvAdmin	0.061545	-.146160	-.151681	0.770610	-.154818
GrAdvPersonal	-.123355	0.066198	-.661418	-.301311	0.487680
GrAdvCareerAdv	-.071223	0.040779	-.107735	-.293706	-.758316
GrAdvPlcmt	-.172973	0.134617	0.469521	0.224221	0.379896

2.A.(15) A portion of the data matrix looks like this:

Obs	Fac Knowledge	Fac Teaching	FacRes InClass	Fac Outside Class	...
1	3	3	3	4	...
2	4	3	4	4	...
3	4	4	3	3	...
4	3	3	4	4	...
5	4	4	3	4	...
6	5	4	3	2	...
7	2	3	1	3	...
...	...	...	...	...	...

Additional columns containing principal component scores can be calculated and included along with the rest of the data in the data set. Specifically how are the data values in the additional column corresponding to the first principal component calculated? Be detailed and specific.

Solution to 2.A: Standardize the columns of data by subtracting the sample mean and dividing by the sample standard deviation. For example, if the sample mean of Fac Knowledge is 3.67 and the standard deviation in 0.87, then standardized values are

Obs	Standardized Fac Knowledge
1	(3-3.67)/.87
2	(4-3.67)/.87
3	(4-3.67)/.87
4	(3-3.67)/.87
5	(4-3.67)/.87
6	(5-3.67)/.87
7	(2-3.67)/.87
...	...

Perform similar standardizations for the other columns, noting that the means and standard deviations will be different for every column.

The PC1 column is then obtained as 0.273091(Standardized Fac Knowledge) + 0.296674(Standardized Fac Teaching) +0.282866(Standardized Fac Research in Class) +…+0.332252(Standardized GrdAdvPlacement).

2.B.(5) What is the variance of the data values in the additional column corresponding to the first principal component?

Solution: The eigenvalues are the variances. So the variance is 5.86.

2.C.(5) An additional data column corresponding to the second principal component can also be included. What is the correlation between the data values in the two columns (PC1 and PC2)?

Solution: PC's are uncorrelated. So the correlation in 0.

2.D.(15) What do PC1 and PC2 measure? Use the 'sorting' idea: which students have the highest PC1? Which have the lowest? What then, does PC1 measure about the students? Which students have the highest PC2? Which have the lowest? What then, does PC2 measure about the students?

Solution. Students who give high ratings on all variables have high PC1; students who give low ratings on all variables have low PC1. So PC1 seems to measure the student's "overall satisfaction with teaching and advising."