Assignment #8
Daniel S Prusinski

Introduction:

Principal component analysis is a helpful non-parametric method for discovering relationships from many different variables in a data set. Factor analysis is a method for grouping variables together and identifies the latent dimensions in the variables. For this assignment, both techniques will be applied to a dataset to explore and extract information from the original data.

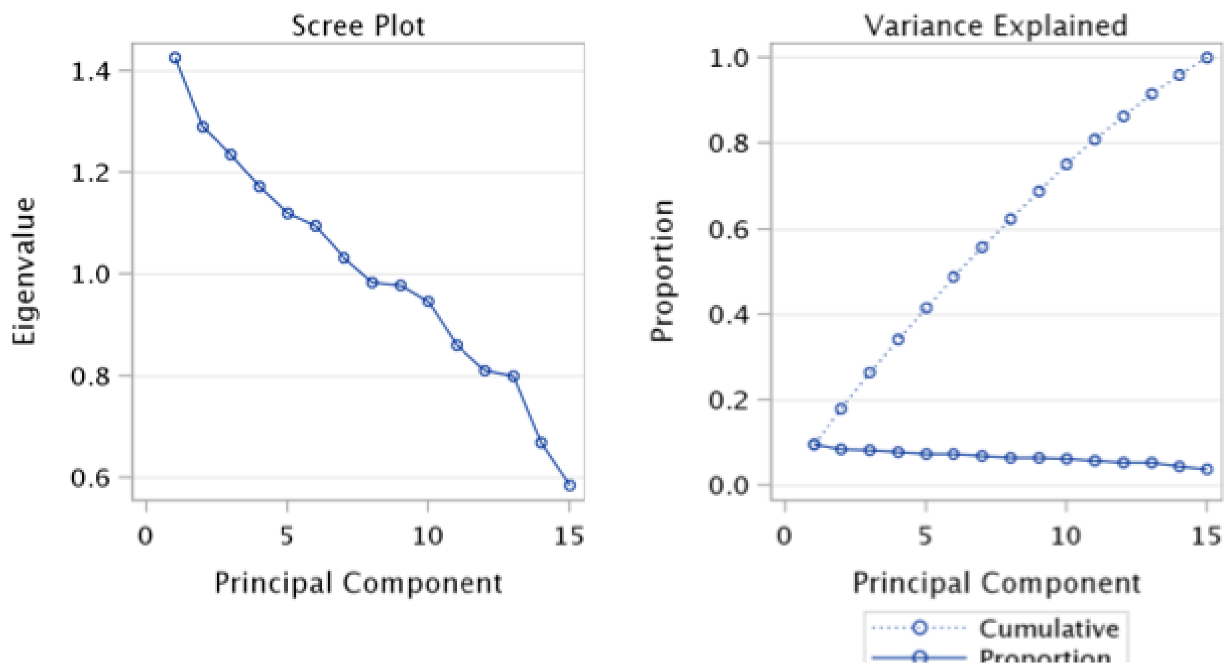
<u>Part 1: An initial Correlation Analysis:</u>

Pearson Correlation Coefficients, N = 1000 Prob > |r| under H0: Rho=0			
	x1_1	**x1_2**	**x1_3**
z1	0.79023 <.0001	0.27726 <.0001	0.31560 <.0001
	x2_1	**x2_2**	**x2_3**
z2	0.31942 <.0001	0.22605 <.0001	0.19271 <.0001
	x3_1	**x3_2**	**x3_3**
z3	0.72089 <.0001	0.20415 <.0001	0.52932 <.0001
	x4_1	**x4_2**	**x4_3**
z4	0.09325 0.0032	0.46617 <.0001	0.55083 <.0001
	x5_1	**x5_2**	**x5_3**
z5	0.77975 <.0001	0.18591 <.0001	0.52800 <.0001

The Z variables have the strongest correlation with the X variables from the first subset. Three of the five first subset variables have strong correlation coefficients. Preliminarily it looks like there are some strong correlation coefficients, but more analysis would need to be conducted to validate the linearity assumptions. Of all the variables, Z4 has the weakest correlation coefficient with the X variables. Z1, Z3, Z5 all have strong correlation coefficients with the firs variable.

Part 2: Principal Components:

Standardizing the data before performing any type of "components" or "factor" analysis is important as it allows the variables to have an equal influence despite individual units. This does not change the ratios between different pairs of objects; rather it makes the overall interpretation more distinct (web.psych.unimelb.edu).



Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15
x1_1	0.25	-0.10	-0.20	-0.36	0.16	0.44	0.21	0.01	0.25	-0.03	0.04	0.60	0.26	-0.01	0.06
x1_2	0.16	-0.05	-0.24	-0.51	0.09	-0.09	0.15	-0.32	0.31	0.25	0.15	-0.55	-0.12	0.08	0.01
x1_3	0.08	-0.01	-0.13	-0.18	-0.18	-0.58	0.14	0.18	-0.33	0.46	0.30	0.31	0.15	0.01	0.01
x2_1	0.16	0.19	0.26	0.19	-0.42	0.31	0.18	-0.19	-0.02	-0.04	0.66	-0.13	0.19	-0.08	-0.05
x2_2	0.10	0.00	0.28	0.14	-0.12	0.20	0.62	0.19	-0.01	0.44	-0.44	-0.13	-0.05	0.08	0.05
x2_3	-0.12	-0.22	0.12	-0.04	-0.25	-0.14	-0.11	0.60	0.60	-0.07	0.10	-0.13	0.24	0.11	-0.02
x3_1	-0.13	0.21	-0.40	0.35	0.19	0.14	0.11	0.27	0.22	0.20	0.34	0.10	-0.55	0.01	0.05
x3_2	-0.17	0.27	-0.50	0.28	0.15	0.06	0.04	0.02	-0.05	0.10	-0.09	-0.25	0.68	0.02	0.04
x3_3	-0.15	0.18	0.07	0.26	-0.18	-0.30	0.02	-0.56	0.51	0.14	-0.15	0.33	0.04	0.15	0.01
x4_1	-0.54	0.22	0.29	-0.28	0.13	0.11	-0.02	0.00	-0.07	0.09	0.14	0.02	0.04	0.19	0.62
x4_2	-0.04	-0.07	0.42	0.09	0.64	0.02	-0.12	-0.02	0.06	0.33	0.22	0.01	0.17	0.07	-0.43
x4_3	0.48	-0.35	0.07	0.37	0.25	-0.15	-0.06	-0.06	0.05	0.00	0.10	-0.07	0.07	0.02	0.63
x5_1	0.39	0.51	0.08	-0.06	0.12	-0.15	0.06	0.16	-0.02	-0.30	-0.01	0.00	-0.03	0.65	-0.05
x5_2	0.28	0.57	0.17	-0.13	0.06	-0.08	-0.23	0.15	0.18	0.13	-0.14	-0.02	0.01	-0.62	0.13
x5_3	0.18	-0.01	-0.08	0.00	-0.29	0.35	-0.63	0.00	-0.07	0.49	-0.10	0.01	-0.03	0.33	0.02
sum	0.92	1.33	0.22	0.14	0.35	0.14	0.38	0.46	1.62	2.18	1.09	0.12	1.07	0.97	1.07

The principal components that have eigenvectors that explain the greatest variation in the predictor variables are: principal components 1, 2, 9, 10, 11, 13, 14, 15. I highlighted these values above, and I use these values because of the variation they explain. These components account for 85% of the variation throughout all the components. The correlation structure between the components is such that the last six of the seven components have strong correlations, while the first two components have relatively strong correlation as well.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.42537946	0.13663847	0.0950	0.0950
2	1.28874099	0.05409204	0.0859	0.1809
3	1.23464895	0.06364847	0.0823	0.2633
4	1.17100048	0.05040574	0.0781	0.3413
5	1.12059474	0.02567880	0.0747	0.4160
6	1.09491595	0.06220971	0.0730	0.4890
7	1.03270623	0.04957906	0.0688	0.5579
8	0.98312718	0.00510939	0.0655	0.6234
9	0.97801778	0.03086678	0.0652	0.6886
10	0.94715100	0.08629638	0.0631	0.7518
11	0.86085462	0.04956912	0.0574	0.8091
12	0.81128550	0.01267730	0.0541	0.8632
13	0.79860821	0.12994927	0.0532	0.9165

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
14	0.66865894	0.08434898	0.0446	0.9610
15	0.58430996		0.0390	1.0000

Part 3: Factor Analysis:

The method of factor analysis performed in Example 1 is Maximum Likelihood Factor Analysis. The error message generated in the SAS log states that communality is greater than 1, thus this method cannot be used. The SAS user guide stipulates that the ML method cannot be used with a single correlation matrix, which is why another method needs to be used.

Preliminary Eigenvalues: Total = 3.061678 Average = 0.20411187				
	Eigenvalue	Difference	Proportion	Cumulative
1	0.97613665	0.21857792	0.3188	0.3188
2	0.75755873	0.25536063	0.2474	0.5663
3	0.50219811	0.11031692	0.1640	0.7303
4	0.39188118	0.09560732	0.1280	0.8583
5	0.29627386	0.07995004	0.0968	0.9550
6	0.21632382	0.05336783	0.0707	1.0257
7	0.16295599	0.05028800	0.0532	1.0789
8	0.11266800	0.01639943	0.0368	1.1157
9	0.09626857	0.04203941	0.0314	1.1472
10	0.05422916	0.06995174	0.0177	1.1649
11	-.01572258	0.03991642	-0.0051	1.1597
12	-.05563900	0.03190769	-0.0182	1.1416
13	-.08754669	0.05246228	-0.0286	1.1130
14	-.14000896	0.06588989	-0.0457	1.0673
15	-.20589885		-0.0673	1.0000

Significance Tests Based on 1000 Observations			
Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	105	400.9678	<.0001
HA: At least one common factor			

Chi-Square without Bartlett's Correction	403.32283
Akaike's Information Criterion	193.32283
Schwarz's Bayesian Criterion	-321.99148
Tucker and Lewis's Reliability Coefficient	0.00000

Part 3: Example 2:

The method of factor analysis performed in Example 2 is the Unweighted Least Squares Method with Heywood approach which allows communality to exceed 1 and the iteration process to continue. The error message generated means there are too many factors in the set for the process to work and a unique solution to be calculated. It is my opinion that we have too many factors for this calculation to return without any errors. If we reduce the factors in the problem then perhaps we will not have SAS output that has errors.

Eigenvalues of the Reduced Correlation Matrix: Total = 5.80062438 Average = 0.38670829				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.12128560	0.11790914	0.1933	0.1933
2	1.00337646	0.16916826	0.1730	0.3663
3	0.83420819	0.06165979	0.1438	0.5101
4	0.77254841	0.07390229	0.1332	0.6433
5	0.69864612	0.18714614	0.1204	0.7637
6	0.51149998	0.14287112	0.0882	0.8519
7	0.36862886	0.12144146	0.0635	0.9155
8	0.24718740	0.10372523	0.0426	0.9581
9	0.14346217	0.04367811	0.0247	0.9828
10	0.09978406	0.09796679	0.0172	1.0000
11	0.00181727	0.00141753	0.0003	1.0003
12	0.00039974	0.00035794	0.0001	1.0004
13	0.00004180	0.00076161	0.0000	1.0004

Eigenvalues of the Reduced Correlation Matrix: Total = 5.80062438 Average = 0.38670829				
	Eigenvalue	Difference	Proportion	Cumulative
14	-.00071981	0.00082205	-0.0001	1.0003
15	-.00154186		-0.0003	1.0000

Standardized Scoring Coefficients										
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7	Factor8	Factor9	Factor10
x1_1	0.04729	-0.02285	0.13403	-0.00592	-0.08310	0.61290	0.07241	0.12276	-0.00961	0.04621
x1_2	0.01639	-0.01132	0.02279	-0.00083	-0.02547	0.12943	-0.00658	-0.18049	0.16229	0.03607
x1_3	0.00632	-0.00791	0.00438	0.01158	-0.02086	-0.01463	-0.02375	-0.16504	0.15683	-0.14697
x2_1	0.01449	0.00415	0.01885	0.00554	0.01942	-0.03664	-0.11101	0.25769	0.07587	0.03394
x2_2	0.01335	0.00695	0.01208	0.01471	-0.01147	-0.00273	-0.09001	0.21632	0.10623	-0.17522
x2_3	-0.18283	-0.03243	-0.07328	0.83117	0.18170	0.05674	0.11532	0.01340	0.00962	0.00071
x3_1	-0.00234	-0.01771	-0.02882	-0.03516	0.00878	-0.02976	0.29760	0.09751	0.02624	-0.03900
x3_2	-0.00466	-0.02705	-0.04722	-0.06570	0.01708	-0.04727	0.37177	0.03236	0.02727	-0.01827
x3_3	0.00412	-0.00083	-0.03549	-0.00388	0.02157	-0.08865	0.01917	0.05443	0.16193	0.18672
x4_1	-0.04867	-0.00406	-0.17331	-0.05123	0.15978	0.02355	-0.08885	0.00158	-0.11854	0.05936
x4_2	-0.05423	0.92624	0.01029	-0.02638	0.22472	0.08374	0.08759	-0.02329	0.06587	-0.04041
x4_3	0.09119	0.10247	0.40710	0.16441	-0.43321	-0.19544	0.03550	0.00638	-0.07354	0.08239
x5_1	0.92964	0.04143	-0.19050	0.21910	0.18622	-0.01890	0.03584	-0.04673	-0.12269	-0.02624
x5_2	0.01120	-0.00651	0.00387	-0.00629	0.04700	0.00015	-0.01823	0.01645	0.17261	0.06887
x5_3	0.04284	-0.11648	0.59837	-0.06147	0.59785	-0.06930	0.03067	-0.04320	-0.03060	-0.01896

Part 4: Factor Analysis:

What is wrong with this factor analysis is we only used 5 of the 15 factors. We are not generating a complete picture of all the factors we are supposed to be analyzing. By changing the nfactors to 5, SAS was able to compute without any errors. The VARIMAX rotational method is an orthogonal method that maximizes the sum of variances for the factor matrix, and simplifies the columns of the factor matrix. This method reduces the number of variables and produces uncorrelated variables as well.

Factor Pattern					
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
x1_1	-0.03642	0.12602	0.00315	-0.02541	0.30710
x1_2	-0.04783	0.06825	0.01944	-0.04121	0.34917
x1_3	-0.05750	0.03644	0.00990	-0.02195	0.03695
x2_1	-0.03481	0.05830	0.09203	-0.08623	-0.15146
x2_2	0.02957	0.04987	0.01036	-0.05946	-0.08302
x2_3	-0.00344	-0.04455	-0.10032	-0.08189	-0.04925
x3_1	-0.02657	-0.06177	0.04647	0.26562	-0.01603
x3_2	-0.08712	-0.14773	0.12392	0.59806	0.02647
x3_3	-0.00118	-0.06692	0.03996	0.03740	-0.14676
x4_1	0.23260	-0.64413	0.13520	-0.16702	0.03246
x4_2	0.98892	0.11508	-0.00312	0.09017	0.02879
x4_3	0.08273	0.46049	-0.15986	0.04359	-0.08355
x5_1	-0.03098	0.21404	0.44628	-0.03638	-0.01242
x5_2	0.02968	0.14116	0.53388	-0.07522	-0.01347
x5_3	-0.04701	0.08236	0.01935	-0.01471	0.00761

Variance Explained by Each Factor				
Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
1.0602758	0.7725813	0.5666561	0.4952539	0.2816066

Orthogonal Transformation Matrix					
	1	**2**	**3**	**4**	**5**
1	0.98566	0.09119	0.03338	-0.13726	-0.01430
2	0.05442	-0.91773	0.30214	-0.16528	0.19027
3	-0.03709	0.27023	0.94830	0.15016	-0.06144
4	0.15472	-0.20977	-0.09099	0.95784	-0.07951
5	0.01409	0.18008	-0.00613	0.11764	0.97647

Rotated Factor Pattern					
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
x1_1	-0.02876	-0.05749	0.04028	-0.00356	0.32620
x1_2	-0.04561	0.00978	0.03907	-0.00019	0.35671
x1_3	-0.05793	-0.02475	0.02025	-0.01332	0.04497
x2_1	-0.05002	-0.04100	0.11250	-0.09146	-0.13511
x2_2	0.02110	-0.04275	0.03180	-0.07747	-0.06791
x2_3	-0.01546	0.02178	-0.10096	-0.09146	-0.04384
x3_1	0.00960	0.00821	0.00044	0.27337	-0.05100
x3_2	-0.00560	0.04043	0.01539	0.63094	-0.05618
x3_3	-0.00256	0.03783	0.01513	0.03578	-0.16145
x4_1	0.16381	0.68977	-0.04364	-0.06132	-0.08922
x4_2	0.99548	-0.03001	0.05643	-0.06547	0.02889
x4_3	0.11810	-0.48245	-0.01316	-0.07954	0.01121
x5_1	-0.04125	-0.07327	0.49024	-0.00042	0.00451
x5_2	0.00531	0.03079	0.55685	-0.02087	-0.01354
x5_3	-0.04474	-0.07019	0.04296	-0.01745	0.02376

Part 5: Correlation Analysis:

From the correlation analysis, it seems PCA produce orthogonal components seeing that the correlation matrix is one for each component and zero for any relationship between the components. I am surprised by this, seeing that I expected VARIMAX to produce a correlation matrix of no collinearity. ULS & VARIMAX does not produce orthogonal components. From reading, it would appear that VARIMAX always produces orthogonal results. ULS can move between orthogonal and collinear outputs. So, these results confuse my reading and SAS code.

Pearson Correlation Coefficients, N = 1000
Prob > |r| under H0: Rho=0

	Prin1	Prin2	Prin3	Prin4	Prin5
Prin1	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin2	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin3	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000
Prin4	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000
Prin5	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000

Pearson Correlation Coefficients, N = 1000
Prob > |r| under H0: Rho=0

	Factor1	Factor2	Factor3	Factor4	Factor5
Factor1	1.00000	0.06170 0.0511	-0.00014 0.9966	0.07213 0.0225	0.04585 0.1474
Factor2	0.06170 0.0511	1.00000	-0.03145 0.3204	0.04422 0.1623	-0.01731 0.5846
Factor3	-0.00014 0.9966	-0.03145 0.3204	1.00000	0.00036 0.9908	0.00679 0.8303
Factor4	0.07213 0.0225	0.04422 0.1623	0.00036 0.9908	1.00000	0.01275 0.6873
Factor5	0.04585 0.1474	-0.01731 0.5846	0.00679 0.8303	0.01275 0.6873	1.00000

Pearson Correlation Coefficients, N = 1000 Prob > |r| under H0: Rho=0					
	Factor1	**Factor2**	**Factor3**	**Factor4**	**Factor5**
Factor1	1.00000	-0.01296 0.6822	0.04533 0.1520	-0.05409 0.0874	0.03654 0.2484
Factor2	-0.01296 0.6822	1.00000	-0.04702 0.1373	0.01039 0.7429	-0.13908 <.0001
Factor3	0.04533 0.1520	-0.04702 0.1373	1.00000	-0.00127 0.9681	0.01732 0.5844
Factor4	-0.05409 0.0874	0.01039 0.7429	-0.00127 0.9681	1.00000	-0.06599 0.0369
Factor5	0.03654 0.2484	-0.13908 <.0001	0.01732 0.5844	-0.06599 0.0369	1.00000

Part 6: Regression Models:

Analysis of Variance Model 1					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	982.63196	196.52639	19490.1	<.0001
Error	994	10.02288	0.01008		
Corrected Total	999	992.65484			

Root MSE	0.10042	**R-Square**	0.9899
Dependent Mean	0.11645	**Adj R-Sq**	0.9899
Coeff Var	86.22825		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	0.11645	0.00318	36.67	<.0001	0
z1	1	0.09522	0.00320	29.79	<.0001	1.01194
z2	1	0.48585	0.00318	152.84	<.0001	1.00109
z3	1	-0.19936	0.00319	-62.51	<.0001	1.00769
z4	1	0.83565	0.00319	261.68	<.0001	1.01031
z5	1	0.00050284	0.00319	0.16	0.8746	1.00573

The above model is the true model, and produces a nearly perfect straight line or correlation coefficient. The p-values for the coefficients look fine except for Z5. The VIFS are low enough to not warrant concern for collinearity. With such a strong R-squared I would want to further investigate. There is a chance that this model is over fit, given that my R-squared is large. In order to validate this model, along with the other models, I would need to conduct a goodness of fit analysis to verify that the assumptions are satisfied.

Analysis of Variance Model 2					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	308.95149	61.79030	89.83	<.0001
Error	994	683.70336	0.68783		
Corrected Total	999	992.65484			

Root MSE	0.82936	**R-Square**	0.3112
Dependent Mean	0.11645	**Adj R-Sq**	0.3078
Coeff Var	712.17556		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	0.11645	0.02623	4.44	<.0001	0
Prin1	1	0.18065	0.02198	8.22	<.0001	1.00000
Prin2	1	-0.19736	0.02311	-8.54	<.0001	1.00000
Prin3	1	0.34014	0.02361	14.40	<.0001	1.00000
Prin4	1	0.07354	0.02425	3.03	0.0025	1.00000
Prin5	1	0.23780	0.02479	9.59	<.0001	1.00000

Model 2 is based on PCA, and produces a much weaker R-square and Adjusted R-squared than model 1. The p-values for the coefficients look fine, but some coefficient values are very small. The VIFS are low, which I would expect given that this is the PCA method. With a relatively weak R-squared I would want to further investigate my PCA. In order to further validate this model I would need to conduct a goodness of fit analysis to verify that the assumptions are satisfied.

Model 3 is based on factor analysis, and produces a stronger R-square and Adjusted R-squared than model 2. The p-values for the coefficients look fine except for X2_2. The VIFS are low, which placates my concern of collinearity. The R-squared and preliminary statistical tests warrant further analysis. In order to further validate this model I would need to conduct a goodness of fit analysis to verify that the assumptions are satisfied.

Analysis of Variance Model 3 – Backward					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	440.76370	55.09546	98.93	<.0001
Error	991	551.89115	0.55690		
Corrected Total	999	992.65484			

Root MSE	0.74626	**R-Square**	0.4440
Dependent Mean	0.11645	**Adj R-Sq**	0.4395
Coeff Var	640.82034		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.14335	0.02377	6.03	<.0001	0
x2_1	1	0.17520	0.02471	7.09	<.0001	1.01261
x2_2	1	0.05191	0.02504	2.07	0.0384	1.00979
x2_3	1	0.13048	0.02330	5.60	<.0001	1.00148
x3_1	1	-0.16006	0.02397	-6.68	<.0001	1.00309
x3_3	1	-0.15809	0.02440	-6.48	<.0001	1.00305
x4_1	1	0.17533	0.02671	6.57	<.0001	1.15093
x4_2	1	0.31867	0.02418	13.18	<.0001	1.06263
x4_3	1	0.43626	0.02481	17.59	<.0001	1.15063

Conclusion:

After assessing PCA and Factor Analysis to the original model, I do not see a great enough change in our model to warrant using these methods for this situation. From this example, I do not see better results which can happen when using PCA and factor analysis. In my opinion, these methods can be most beneficial in the presence of collinear data. Had the original data suffered from collinearity, these models would have proved to be very valuable. But, because this data did not suffer from collinearity there was not much gain from employing these methods.

SAS Code:

```
*********************************************
Assignment 8 Version1
Daniel Prusinski
11/25/2012
*********************************************;

*****Part 1: An Initial Correlation Analysis*****;

libname mydata    '/courses/u_northwestern.edu1/i_833463/c_3505/SAS_Data/' access=readonly;
title ;

proc contents data=mydata.factor_data; run; quit;
proc print data=mydata.factor_data (obs=5); run; quit;

data temp;
      set mydata.factor_data;


*****I am still working on the macro,
but at least the correlation matrix is done*****;

%macro corr_matrix (k);

proc corr data=temp plots=matrix;
var x&k._1 x&k._2 x&k._3;
with z&k.;
run;

%mend corr_matrix;

%corr_matrix(k=1);
%corr_matrix(k=2);
%corr_matrix(k=3);
%corr_matrix(k=4);
%corr_matrix(k=5);
****************
*****Part 2******
****************;
```

```
proc standard data=temp mean=0 std=1 out=temp_std;
var z1 z2 z3 z4 z5
      x1_1 x1_2 x1_3
      x2_1 x2_2 x2_3
      x3_1 x3_2 x3_3
      x4_1 x4_2 x4_3
      x5_1 x5_2 x5_3 ;
run;

data zdata;
      set temp_std;
      keep y z1 z2 z3 z4 z5;
run;

data xdata;
      set temp_std;
      drop y z1 z2 z3 z4 z5;
run;

ods graphics on;
proc princomp data=xdata out=xdata_pca outstat=pca_stats plots=(scree);
run;
ods graphics off:

ods graphics on; proc princomp data=xdata out=xdata_pca outstat=pca_stats plots=(scree); run; ods graphics
off;

*****************
*****Part 3******
*****************;

ods graphics on;
proc factor data=xdata method=ml out=xdata_ml outstat=ml_stats
mineigen=0 priors=max nfactors=15 score scree ;
run; ods graphics off;

ods graphics on; proc factor data=xdata method=uls heywood out=xdata_uls
outstat=uls_stats mineigen=0 priors=max nfactors=15 score scree ;
run; ods graphics off;

*****************
```

```
*****Part 4******
*****************;

ods graphics on;
proc factor data=xdata method=uls heywood out=xdata_uls outstat=uls_stats
mineigen=0 priors=max nfactors=5 score scree ;
run; ods graphics off;

ods graphics on; proc factor data=xdata method=uls heywood rotate=varimax
out=xdata_varimax outstat=varimax_stats mineigen=0
priors=max nfactors=5 score scree ;
run; ods graphics off;

*****************
*****Part 5******
*****************;
proc corr data=xdata_pca;
var prin1 prin2 prin3 prin4 prin5;
run;

proc corr data=xdata_uls;
var factor1 factor2 factor3 factor4 factor5;
run;

proc corr data=xdata_varimax;
var factor1 factor2 factor3 factor4 factor5;
run;

*****************
*****Part 6******
*****************;

data pca_data;
set xdata_pca (keep= prin1 prin2 prin3 prin4 prin5);
id_nbr = _n_;
run;

data varimax_data;
set xdata_varimax (keep= factor1 factor2 factor3 factor4 factor5);
id_nbr = _n_;
run;
```

```
data zdata;
set zdata; id_nbr = _n_;
run;

proc sort data=pca_data;
by id_nbr; run;

proc sort data=varimax_data;
by id_nbr; run;

proc sort data=zdata;
by id_nbr; run;

data model_data;
retain id_nbr;
merge zdata pca_data varimax_data;
by id_nbr; run;

* True model;
proc reg data=model_data;
model Y = z1 z2 z3 z4 z5 / vif;
run; quit;

* PCA model; proc reg data=model_data;
model Y = prin1 prin2 prin3 prin4 prin5 / vif;
run; quit;

proc reg data=model_data;
model Y = factor1 factor2 factor3 factor4 factor5 / vif;
run; quit;

proc reg data=temp;
model Y = x1_1 x1_2 x1_3
x2_1 x2_2 x2_3
x3_1 x3_2 x3_3
x4_1 x4_2 x4_3
x5_1 x5_2 x5_3
 / selection=backward vif;
 run; quit;
```