Multivariate Stat Analysis Project-2014

----Rongmin Xia

Analysis contents:

1. Mean and Standard deviation

2. Pearson Correlation Coefficients

3. Covariance Matrix

4. Bonferroni and Simultaneous Confidential interval of single variable

5. Univariate Normality Tests per Variable

6. Chi-square distance plot

7. Outlier detection

8.Scatter plot for each pair of data

9. Principle components analysis

10. Factor analysis

11. Test hypothesis

12. Simply regression analysis

13. Canonical analysis

14. Conclusion

1. **Mean and Standard deviation (observations=62)**

	Mean	Std Dev
BL	21.723	2.881
EM	7.266	0.716
SF	5.637	1.463
BS	1.019	0.693
AFL	-0.022	0.25
LFF	39.033	14.868
FFF	26.678	17.561
ZST	1.067	0.029

2. **Pearson Correlation Coefficients**

	BL	EM	SF	BS	AFL	LFF	FFF	ZST
BL	1.000	0.914	0.984	0.988	0.648	0.735	-0.542	0.822
EM	0.914	1.000	0.942	0.875	0.537	0.609	-0.556	0.850
SF	0.984	0.942	1.000	0.975	0.681	0.764	-0.575	0.865
BS	0.988	0.875	0.975	1.000	0.706	0.796	-0.564	0.813
AFL	0.648	0.537	0.681	0.706	1.000	0.906	-0.733	0.784
LFF	0.735	0.609	0.764	0.796	0.906	1.000	-0.711	0.793
FFF	-0.542	-0.556	-0.575	-0.564	-0.733	-0.711	1.000	-0.785
ZST	0.822	0.850	0.865	0.813	0.784	0.793	-0.785	1.000

3. **Covariance Matrix**

	BL	EM	SF	BS	AFL	LFF	FFF	ZST
BL	8.303	1.887	4.147	1.972	0.466	31.489	-27.421	0.070
EM	1.887	0.513	0.988	0.434	0.096	6.483	-6.995	0.018
SF	4.147	0.988	2.140	0.988	0.248	16.626	-14.761	0.037
BS	1.972	0.434	0.988	0.480	0.122	8.204	-6.860	0.017
AFL	0.466	0.096	0.248	0.122	0.062	3.360	-3.214	0.006
LFF	31.489	6.483	16.626	8.204	3.360	221.052	-185.637	0.348
FFF	-27.421	-6.995	-14.761	-6.860	-3.214	-185.637	308.400	-0.406
ZST	0.070	0.018	0.037	0.017	0.006	0.348	-0.406	0.001

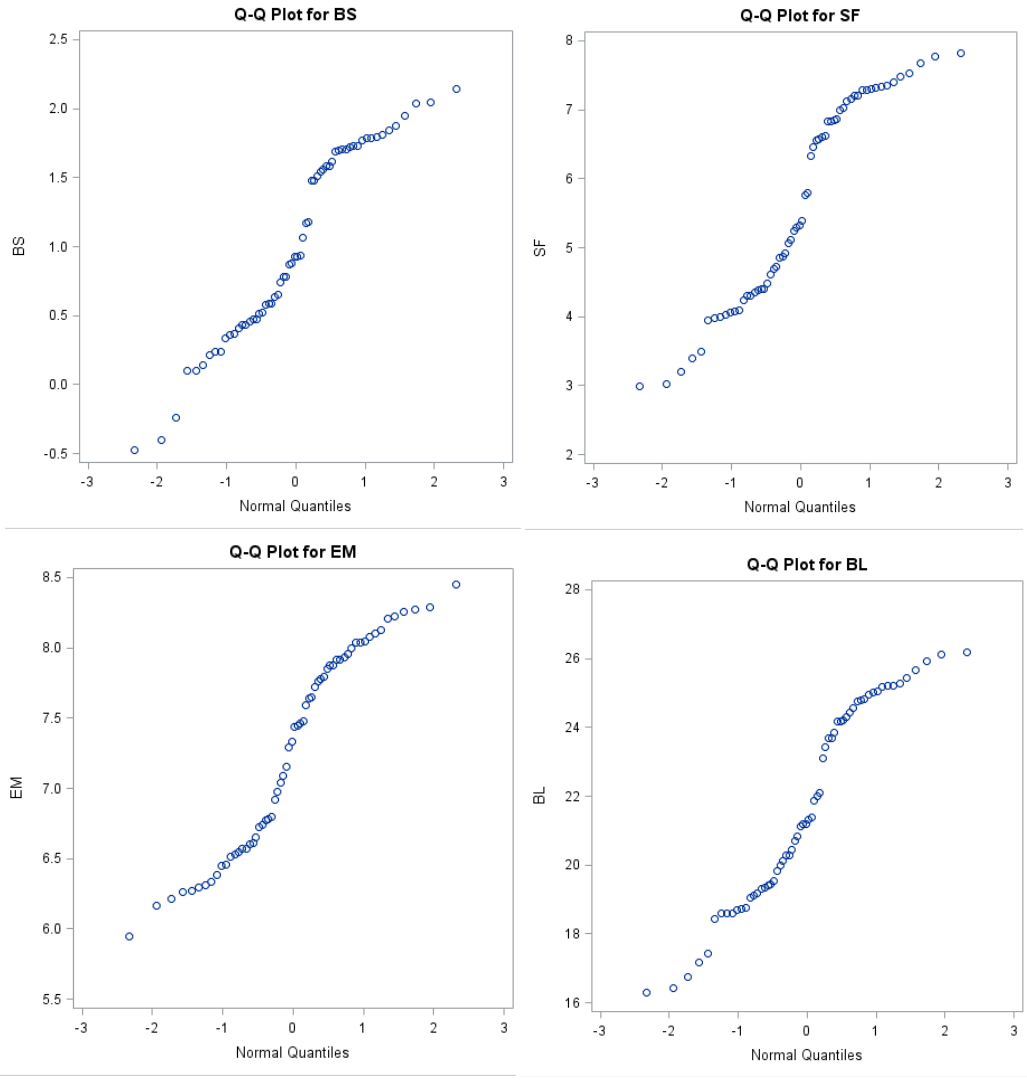
4. **Confidential interval (α=5%)**

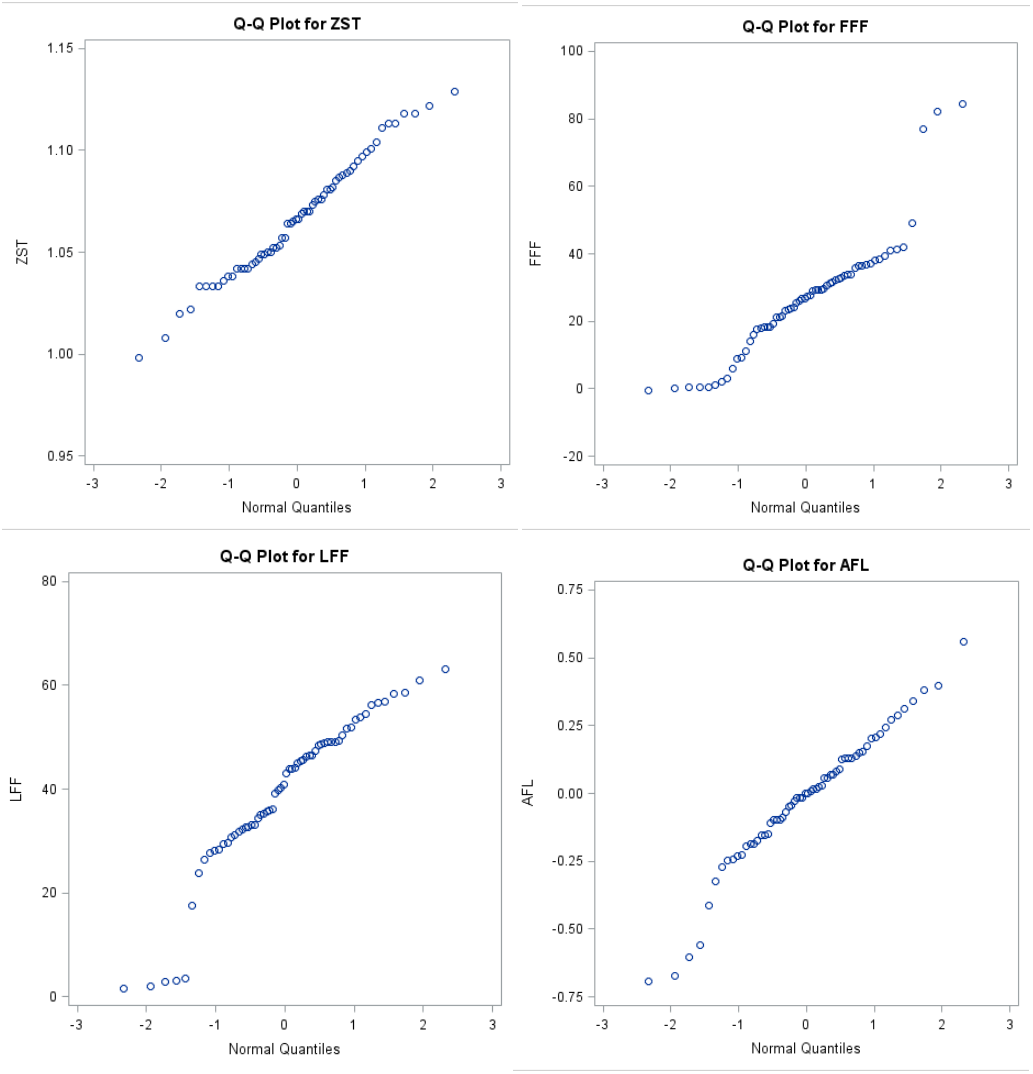
	Bonferroni		Simultaneous	
	Low	Up	Low	Up
BL	20.686	22.759	20.123	23.323
EM	7.008	7.524	6.868	7.664
SF	5.111	6.164	4.825	6.45
BS	0.769	1.268	0.634	1.404
AFL	-0.112	0.068	-0.16	0.117
LFF	33.684	44.381	30.777	47.288
FFF	32.995	20.36	16.927	36.429
ZST	1.056	1.077	1.05	1.083

5. **Univariate Normality Tests per Variable**

Variable Name	Test Statistic	P-Value
BL	0.931	0.002
EM	0.925	0.001
SF	0.914	0.000
BS	0.935	0.003
AFL	0.967	0.094
LFF	0.918	0.001
FFF	0.892	0.000
ZST	0.986	0.681

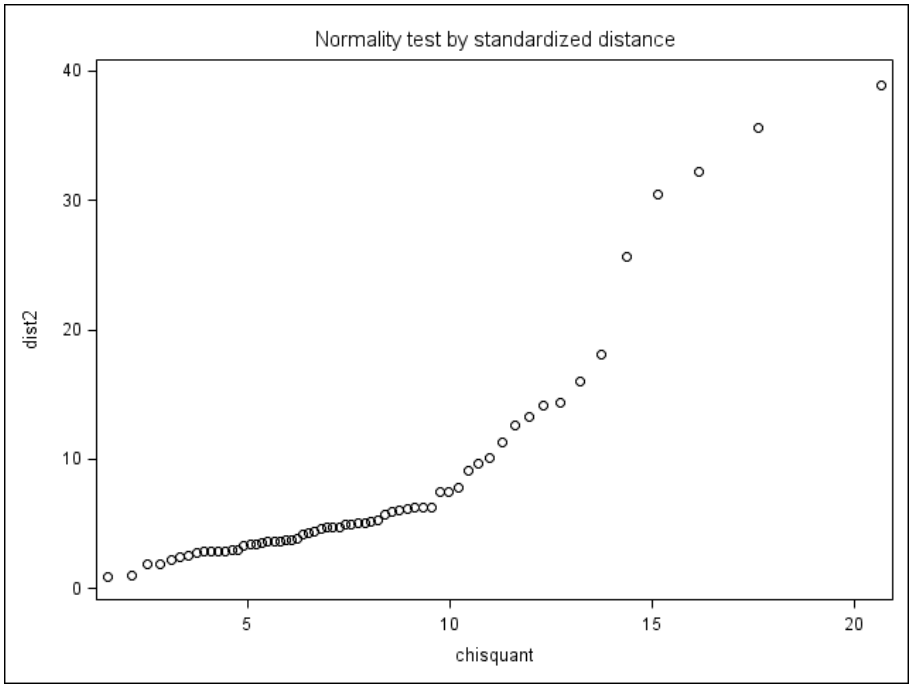
QQ plot for each variable





Only AFL and ZST follow normal distribution, all others are not normal distribution.

6. Chi-square distance plot



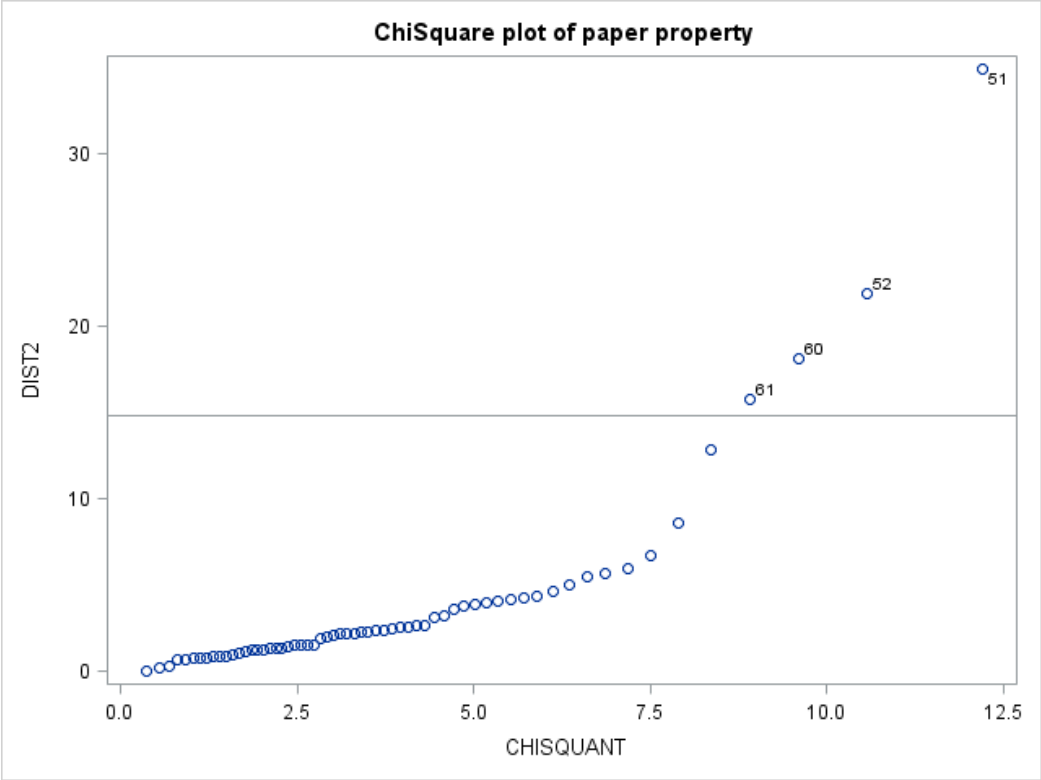
Chi-square plot doesnot follow a straight line, especially some of all the points are far away from (0,0). Thus, we can think that this data set does not a normal distribution, and have some suspicious outliers.

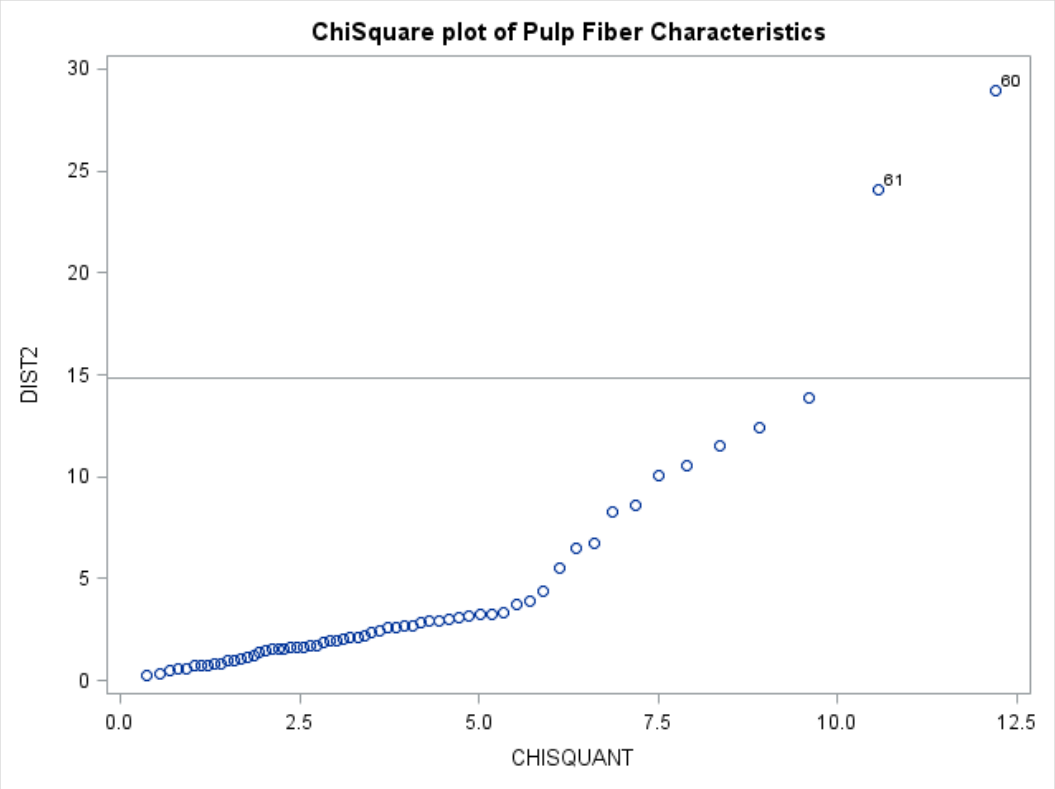
7. Outlier detection

Obs	BL	EM	SF	BS	zBL	zEM	zSF	zBS	Chi-D
	:	:	:	:	:	:	:	:	
51	22.007	8.259	7.322	1.169	0.099	1.386	1.152	0.217	**34.937**
52	21.115	7.913	6.557	0.928	-0.211	0.903	0.629	-0.131	**21.916**
53	26.194	8.454	7.816	1.113	**1.552**	**1.658**	1.489	**1.625**	5.442
54	25.674	8.208	7.534	1.104	1.371	1.314	1.296	1.482	4.095
55	25.93	8.1	7.669	1.111	1.460	1.164	1.389	1.469	2.309
56	21.39	7.475	5.294	1.113	-0.116	0.291	-0.235	-0.207	3.877
57	18.441	6.652	3.946	1.02	-1.139	-0.857	-1.156	-1.268	2.177
58	16.441	6.315	2.997	1.008	**-1.833**	-1.328	**-1.805**	**-2.047**	5.681
59	16.294	6.572	3.017	0.998	**-1.884**	-0.969	**-1.791**	**-2.160**	8.611
60	20.289	7.719	4.866	1.081	-0.498	0.632	-0.527	-1.125	**18.105**
61	17.163	7.086	3.396	1.033	**-1.582**	-0.251	**-1.532**	**-1.811**	**15.743**

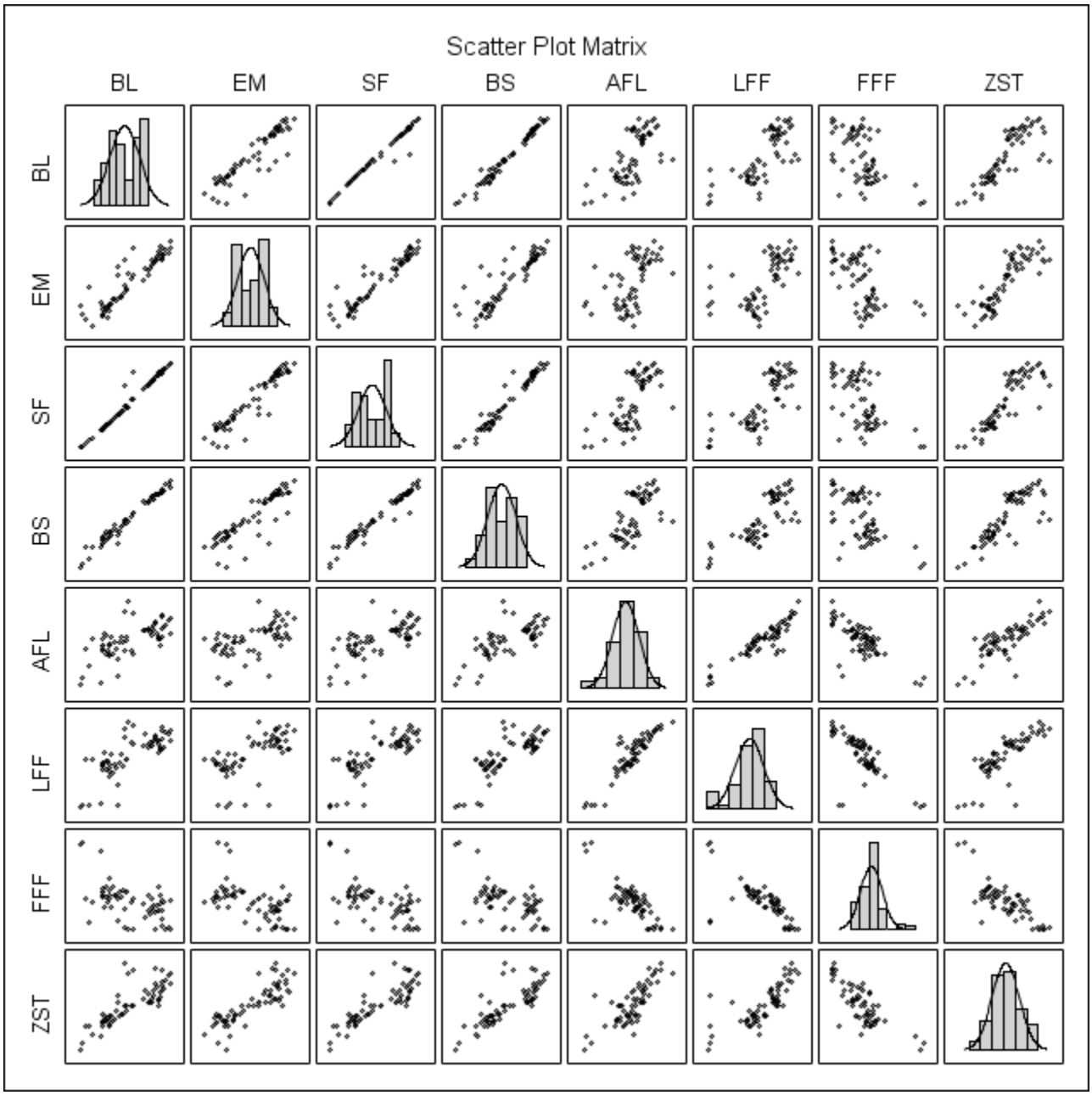
Obs	AFL	LFF	FFF	ZST	zAFL	zLFF	zFFF	zZST	Chi-D
	:	:	:	:	:	:	:	:	
49	0.314	56.627	2.925	1.118	1.346	1.183	-1.353	**1.736**	3.216
50	0.217	53.458	0.511	1.122	0.957	0.970	-1.490	**1.871**	4.389
51	0.381	60.993	0	1.118	**1.614**	1.477	-1.519	**1.736**	3.243
52	0.397	58.429	1.147	1.129	**1.678**	1.305	-1.454	**2.109**	5.482
53	0.289	56.755	0.407	1.113	1.245	1.192	-1.496	**1.566**	2.676
54	0.202	56.111	0.407	1.104	0.897	1.149	-1.496	1.261	2.717
55	0.273	53.847	2.023	1.111	1.181	0.996	-1.404	1.499	2.633
56	0.558	63.035	-0.391	1.113	**2.323**	**1.614**	**-1.541**	**1.566**	6.734
57	-0.672	3.448	76.878	1.02	**-2.606**	**-2.393**	**2.859**	**-1.587**	**11.533**
58	-0.605	2.845	84.554	1.008	**-2.337**	**-2.434**	**3.296**	**-1.994**	**12.404**
59	-0.694	1.515	81.988	0.998	**-2.694**	**-2.523**	**3.150**	**-2.333**	**10.579**
60	-0.559	2.054	8.786	1.081	**-2.153**	**-2.487**	-1.019	0.481	**28.930**
61	-0.415	3.018	5.855	1.033	**-1.576**	**-2.422**	-1.186	-1.146	**24.104**

Here n=64 and p=4, 1.5 of standardized value might be considered larger. All the possible outlier was marked as red color. For the paper property, the suspicious points have observation number 53, 58 59 and 61. For the pulp fiber characteristics, 49-61 are all need to be further verified, because the value of FFF(points 49-56) is extremly away from its confidential interval CI(33.684, 44.381). same situation happens on LFF, the value of points(57-61) is much smaller than its CI(20.36, 32.995,). Based on current research, observation 49-61 are considered as suspicious outliers. Next step we will process the chi-square plot to verify our pre-judgement. The chi-D is the generalized squared distance, listed in the above form. The observation 51-52, 57-61 have large distances.


ChiSquare plot of paper property
DIST2
CHISQUANT
0
10
20
30
0.0
2.5
5.0
7.5
10.0
12.5
51
52
60
61


ChiSquare plot of Pulp Fiber Characteristics
DIST2
CHISQUANT
0
5
10
15
20
25
30
0.0
2.5
5.0
7.5
10.0
12.5
60
61

8. Scatter plot for each pair of data



Since there are two properties, paper and fiber characteristics. We did the principle component analysis respectively.

9. Principle components analysis

1. Paper Characteristics

- Correlation Matrix

Correlation Matrix				
	BL	EM	SF	BS
BL	1.000	0.914	0.984	0.988
EM	0.914	1.000	0.942	0.875
SF	0.984	0.942	1.000	0.975
BS	0.988	0.875	0.975	1.000

- Eigenvalue

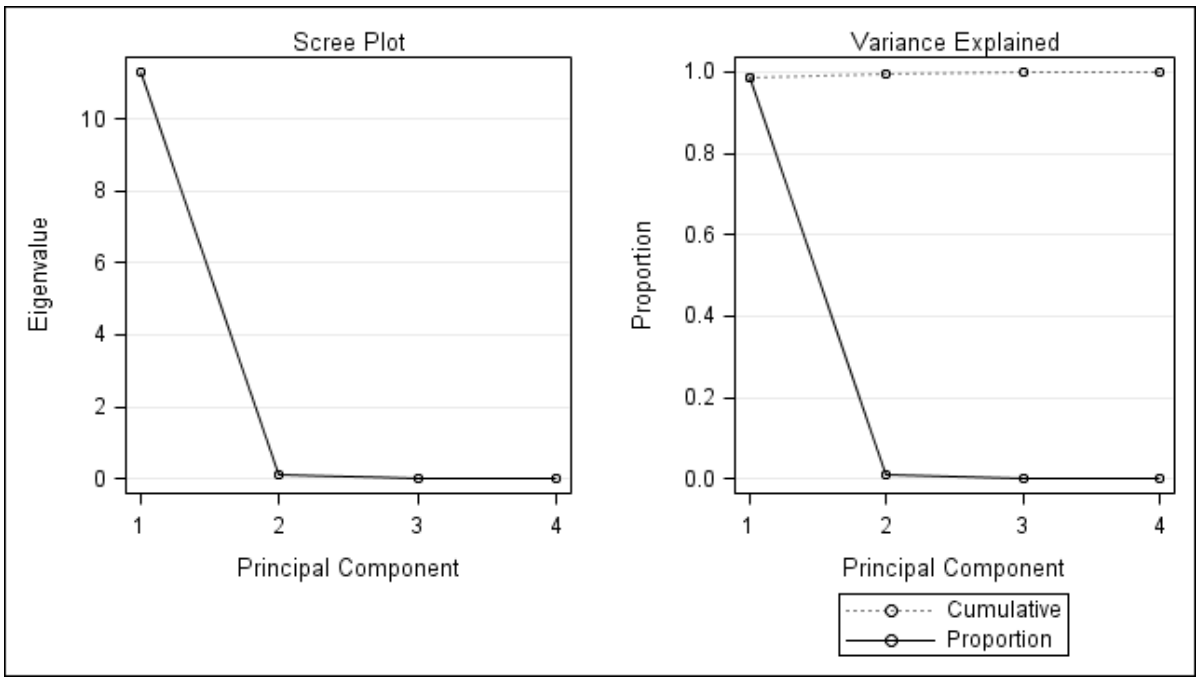
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.840	3.699	0.960	0.960
2	0.140	0.128	0.035	0.995
3	0.013	0.005	0.003	0.998
4	0.008		0.002	1.000

- Eigenvectors

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
BL	0.506	-0.261	0.565	-0.597
EM	0.485	0.819	0.194	0.237
SF	0.508	-0.020	-0.800	-0.318
BS	0.500	-0.510	0.053	0.698

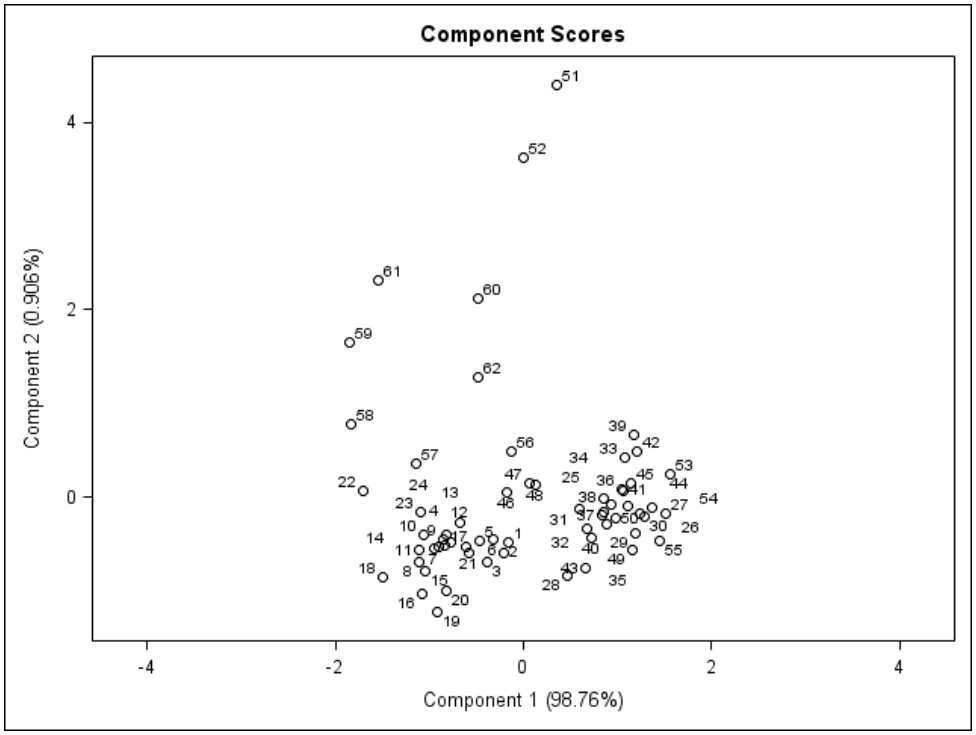
- Analysis of component

Based on the eigenvector form, Pinciple component one is much larger than others, suggests that one principal component effectively summarzes the paper properties data. This opinion also can be seen in the scree plot(elbow appears at 2nd component). This First principle component explains 98.8% of the total variance. Therefore the four vector can be expressed by one index variable labeled as an index of paper strength.



- Outlier detection

The plot below of the scores on the first two sample principal components since we only consider first component, all the points is within the range of 2, so it does not indicate any obvious outliers.



2. Pulp Fiber Characteristics

- Correlation Matrix

Correlation Matrix				
	AFL	LFF	FFF	ZST
AFL	0.062	3.360	-3.214	0.006
LFF	3.360	221.052	-185.637	0.348
FFF	-3.214	-185.637	308.400	-0.406
ZST	0.006	0.348	-0.406	0.001

- Eigenvalue

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	455.478	381.453	0.860	0.860
2	74.026	74.016	0.140	0.999
3	0.010	0.010	0.000	1.000
4	0.000		0.000	1.000

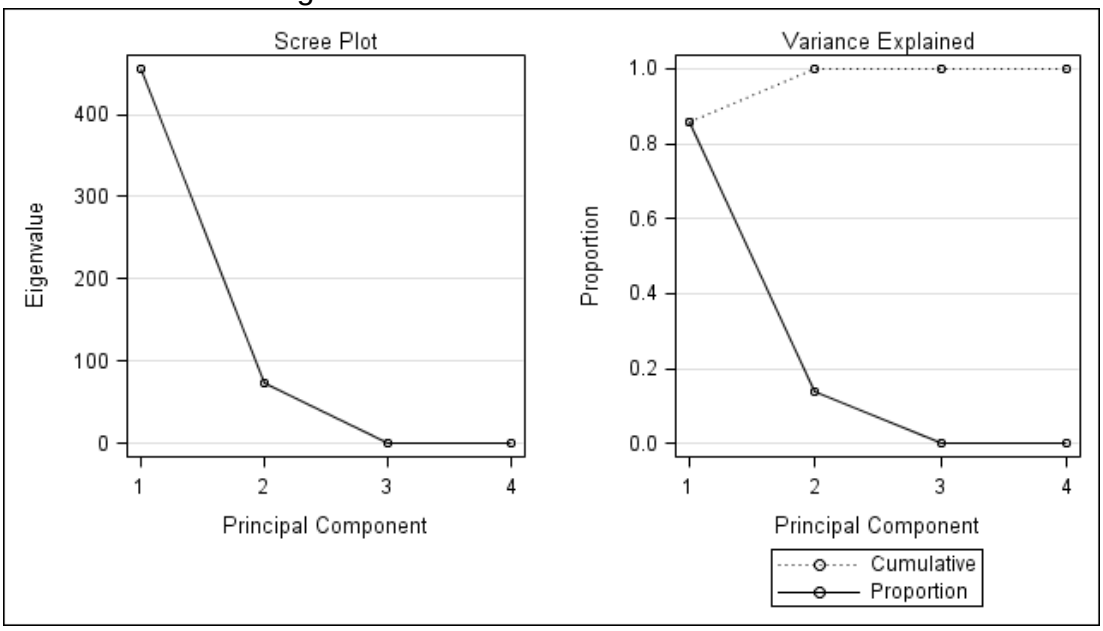
- Eigenvectors

Eigenvectors				
	Prin1	Prin2	Prin3	Prin4
AFL	-0.010	0.009	1.000	-0.019
LFF	-0.621	0.784	-0.013	-0.001
FFF	0.784	0.621	0.003	0.001
ZST	-0.001	0.000	0.019	1.000

- Analysis of component

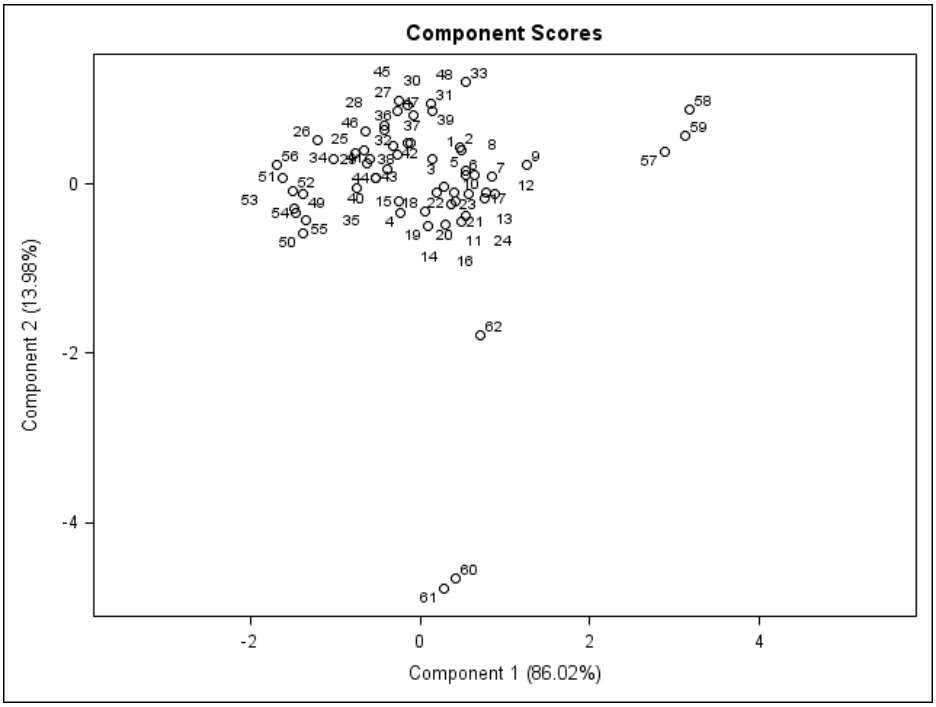
Based on the eigenvector form, Pinciple 1st and 2nd component are much larger than others, suggests that one or (one and two) principal component effectively summarizes the paper properties data. This opinion also can be seen in the scree plot(elbow appears at 2nd component). This First principle component explains 86% of the total variance, in the contrast, the first two has occupied 99% of the total variance. Further checking the eigen vectors, the 2nd principle component has similar pattern with 1st principle component,

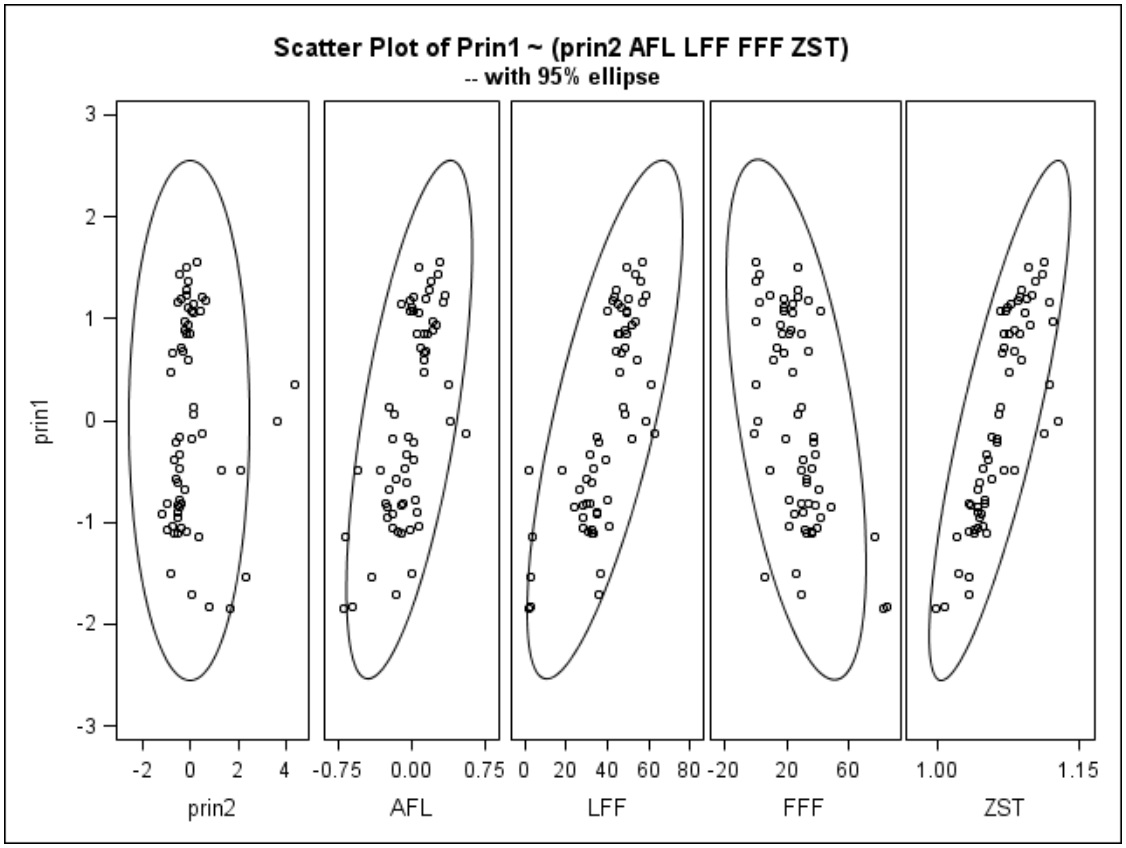
therefore, we think those four vectors also can be expressed by one index variable labeled as an index of fiber strength.



- Outlier detection

  The plot below of the scores on the first two sample principal components. The points 57-61 is a little far from mean, which indicate the possibility of outliers. In the scatter plots(with 95% confidence region) of principle component with variables also indicate these points (57-61) are suspicious.





10.Factor analysis

- **Covariance Matrix**

	BL	EM	SF	BS
BL	8.303	1.887	4.147	1.972
EM	1.887	0.513	0.988	0.434
SF	4.147	0.988	2.140	0.988
BS	1.972	0.434	0.988	0.480

Same as the previous principle component analysis, we did the factor analysis for paper and pulp fiber property respectively.

1. Paper Characteristics

-

Factor	Eigenvalue	Difference	Proportion	Cumulative
1	3.813	3.720	0.980	0.980
2	0.093	0.096	0.024	1.004
3	-0.003	0.009	-0.001	1.003
4	-0.013		-0.003	1.000

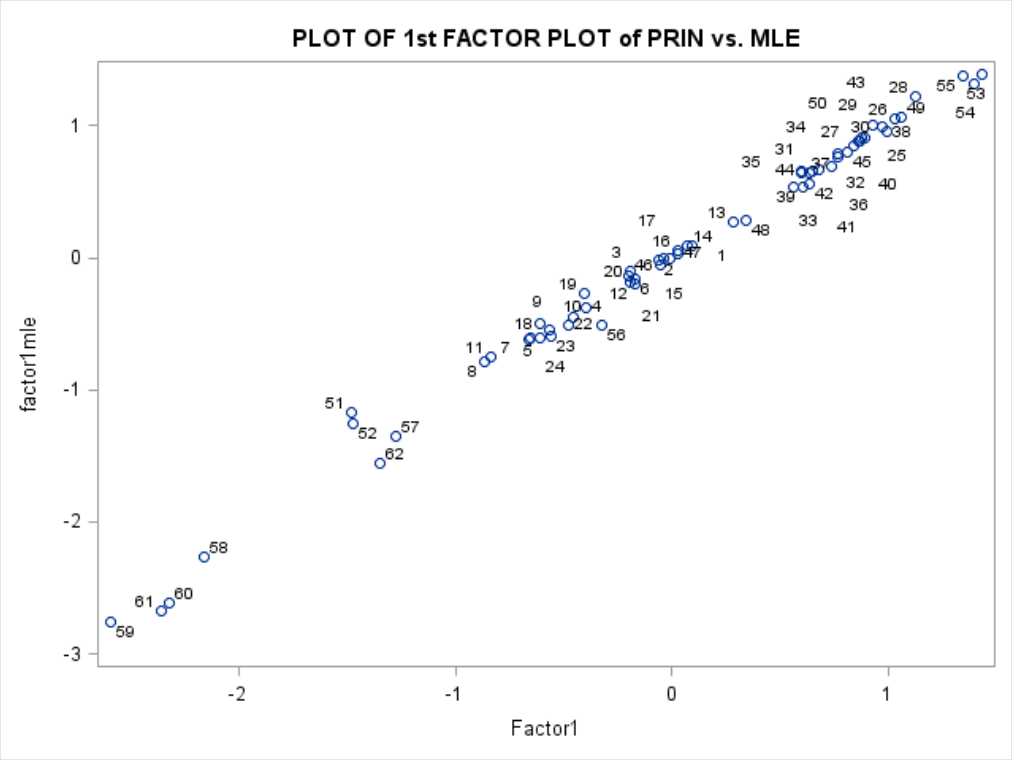
- Factor pattern based on two factors analysis of principle component

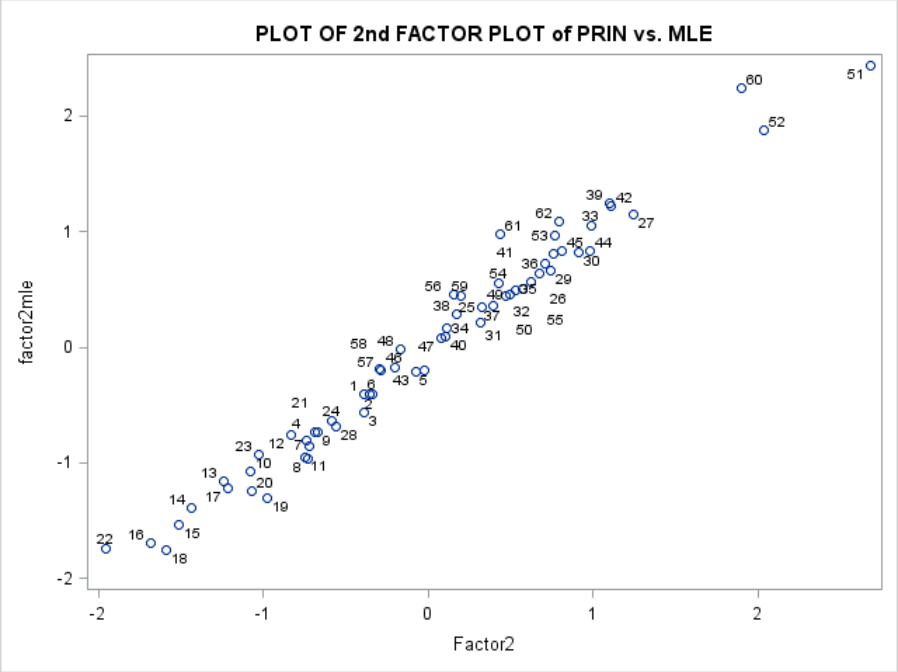
	initial		rotated	
	Factor1	**Factor2**	**Factor1**	**Factor2**
BL	0.993	-0.025	0.787	0.608
EM	0.992	0.063	0.531	0.814
SF	0.99	-0.136	0.724	0.682
BS	0.928	0.283	0.845	0.526

- Result of initial and rotated MLE of Two Factors loadings

	initial		rotated	
	Factor1	**Factor2**	**Factor1**	**Factor2**
BL	0. 98756	0. 10942	0. 79684	0. 59353
EM	0. 87467	0. 45736	0. 52358	0. 83671
SF	0. 97451	0. 19645	0. 74160	0. 66203
BS	1. 00000	0. 00000	0. 86288	0. 50540

- Plot of factors based on two method Principle component vs. MLE





- Factor pattern based on one factor principle component and MLE analysis

Factor1	PRIN	MLE
BL	0.992	0.998
BS	0.940	0.916
SF	0.994	0.986
EM	0.980	0.989

The first factor explains about 98% of the varance and all varables load highly and about equally on this factor. Therefore, these four factors can be represented by this first factor, and we will do further research based on this one factor analysis. This factor might be called a "paper properties index." The principle component and MLE factor analysis provide very similar results.

- Outlier detection:

Observations 60 and 61, and 58 and 59 may be outliers.

2. Pulp Fiber Characteristics

Eigenvalues of the Reduced Correlation Matrix: Total = 3.0598459 Average = 0.76496147				
Factor	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**
1	3.358	3.008	0.839	0.839
2	0.349	0.149	0.087	0.927
3	0.201	0.108	0.050	0.977
4	0.092		0.023	1.000

The first and second factor explains 84% and 8.7% of the total variance respectively. Thus, those variables can be reduced to 1 or 2 factors.

- Un-rotated factor pattern of two factors

	Factor1	**Factor2**
AFL	0.936	0.256
LFF	0.933	0.288
ZST	0.917	-0.150
FFF	-0.878	0.423

Variance Explained by Each Factor

Final Communality Estimates: Total = 3.707					
AFL	LFF	FFF	ZST	**Factor1**	**Factor2**
0.942	0.953	0.949	0.863	3.358	0.349

- Rotated factor pattern of two factors

	Factor1	**Factor2**
AFL	*0.744*	0.543
LFF	*0.895*	0.441
ZST	*0.537*	0.708
FFF	-0.397	*-0.806*

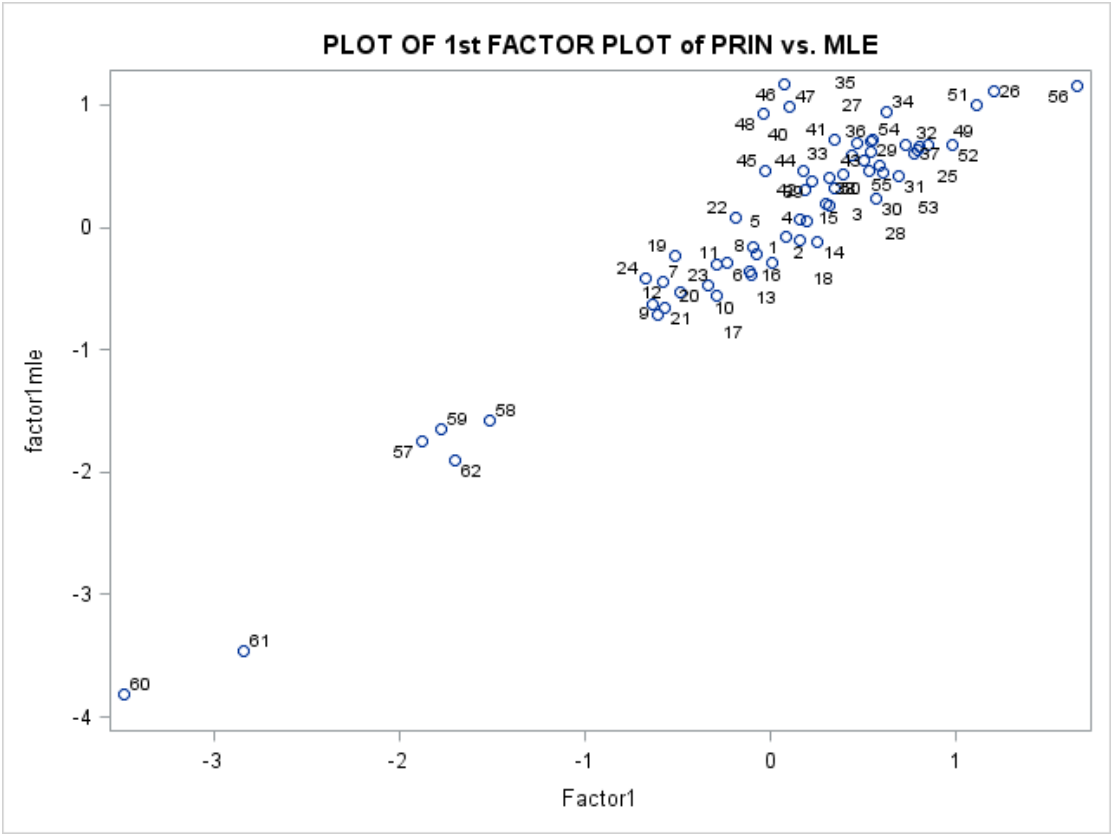
Variance Explained by Each Factor

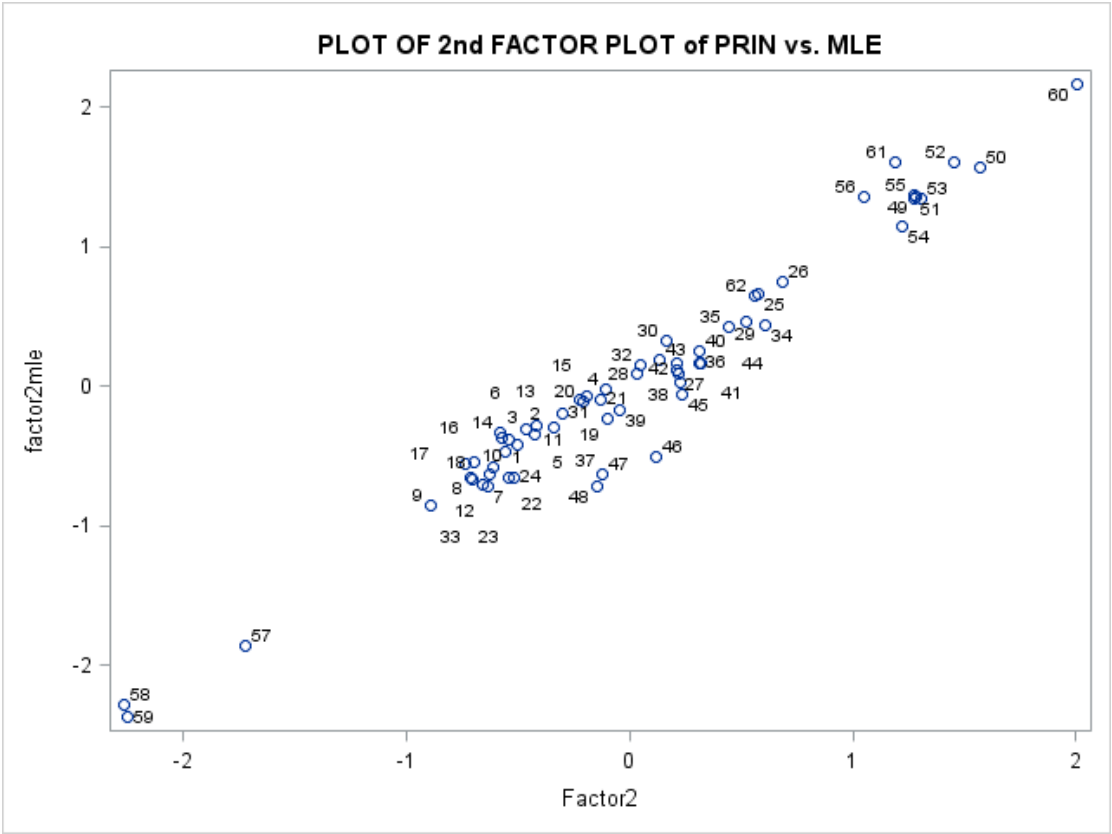
Final Communality Estimates: Total = 3.707					
AFL	LFF	FFF	ZST	**Factor1**	**Factor2**
0.942	0.953	0.949	0.863	2.018	1.689

- Result of initial and rotated MLE of Two factor loadings

	initial		rotated	
	Factor1	**Factor2**	**Factor1**	**Factor2**
AFL	*0.91091*	*-0.13864*	*0.74458*	*0.54276*
LFF	*0.99735*	*0.02097*	*0.89484*	*0.44092*
FFF	*-0.72408*	*0.53273*	*-0.39718*	*-0.80644*
ZST	*0.80286*	*-0.38148*	*0.53679*	*0.70849*

- Plot of factors based on two method Principle component vs. MLE





Examining the unrotated loadings for both solution methods, we see that the second factor explains little (about 10%) of the remaining variance. Also, this factor has moderate to very small loadings on all the variables with the possible exception of variable FFF. If retained, this factor might be

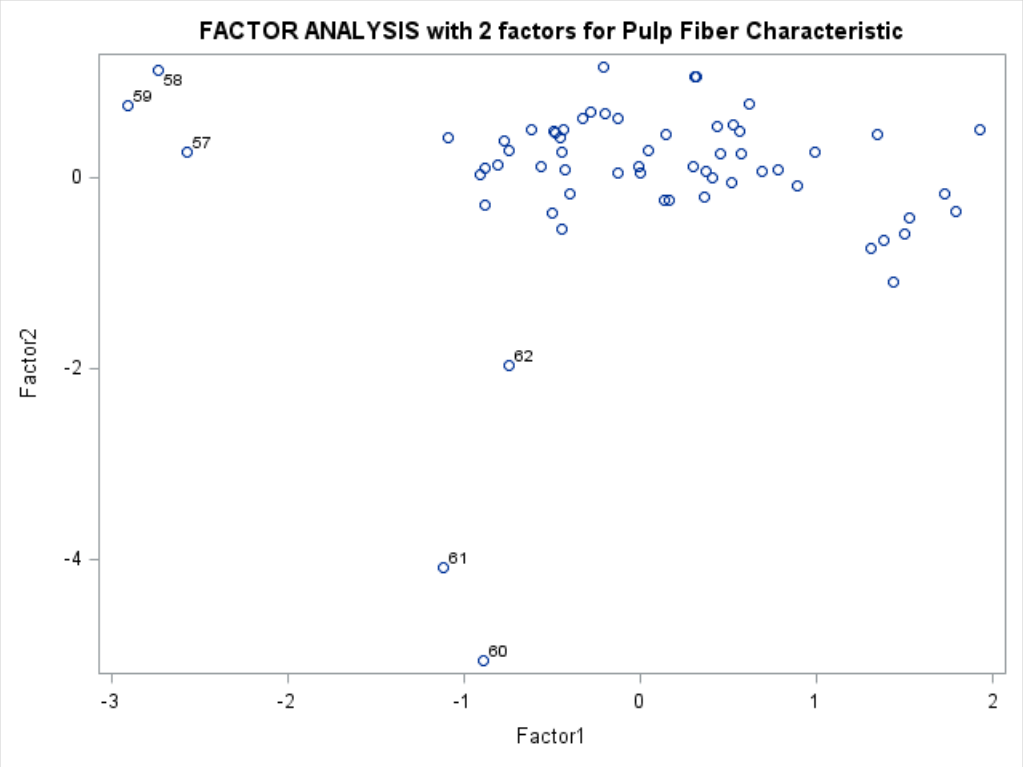
called a "fine fiber" of "quality" factor. Using the rotated loadings, the second factor looks much like the first factor for both solution methods. That is, this factor appears to be a contrast between variable FF and the group of variables AF, LFF and ZST. To summarize, there seems to be no gain in understanding from adding a second factor to the model. A one factor model appears be sufficient in this case. Also, the comparison plot of Principle component vs. MLE provide some information of possible outliners which are away from origin point, such as #60 and #61.

- Standardized Scoring Coefficients from one factor analysis

Factor1	PRIN	MLE
AFL	0.932	0.949
LFF	0.928	0.945
FFF	-0.839	-0.784
ZST	0.880	0.847

Since the first factor explains 84% of the total variance and represents a contrast between FFF (with a negative loading) and the AFL, LFF and ZST (with positive Loadings). AFL, LFF and ZST may all have to do with paper strength, while FFF may have something to do with paper quality. Therefore, we can make a conclusion that Pulp Fiber Characteristics can expressed as one factor(strength-quality).

- Outlier detection:

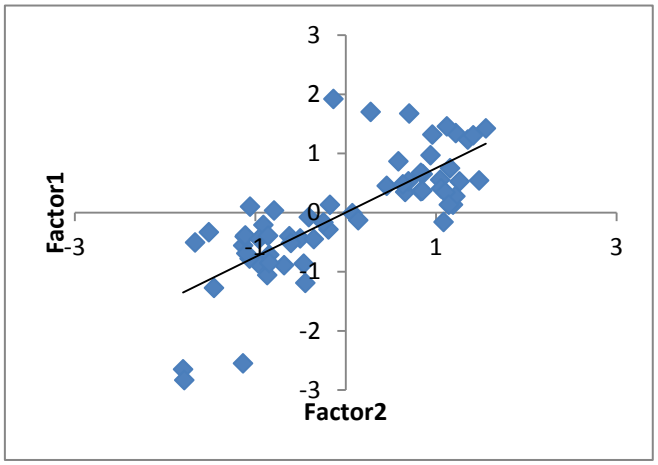


However, plots of the factor scores for two factors suggest observations 60, 61 and, perhaps, observations 57, 58 and 59 may be outliers.

11.Test hypothesis

There are two group parameters to describe paper quality, Pulp Fiber Characteristics and Paper Characteristics. According to factor and principle component analysis, each of them can be expressed as one factor or component. Are those two factor can be treated equally? If so, this dataset can be simplified as one parameter. First we draw the scatter plot, and it seems a positive high correlation.

We hypothesis that these two factors are same, both of them show some property of the paper quality. T-test is conducted for the comparison of paper factor and Fiber factor, with assuming H0=0 and equal variance. The p-value of T-test is 0.993. Therefore, there are no reject to our hypothesis, we can select one of them as the paper quality index, and here I prefer the factor of (BL EM SF BS) group.



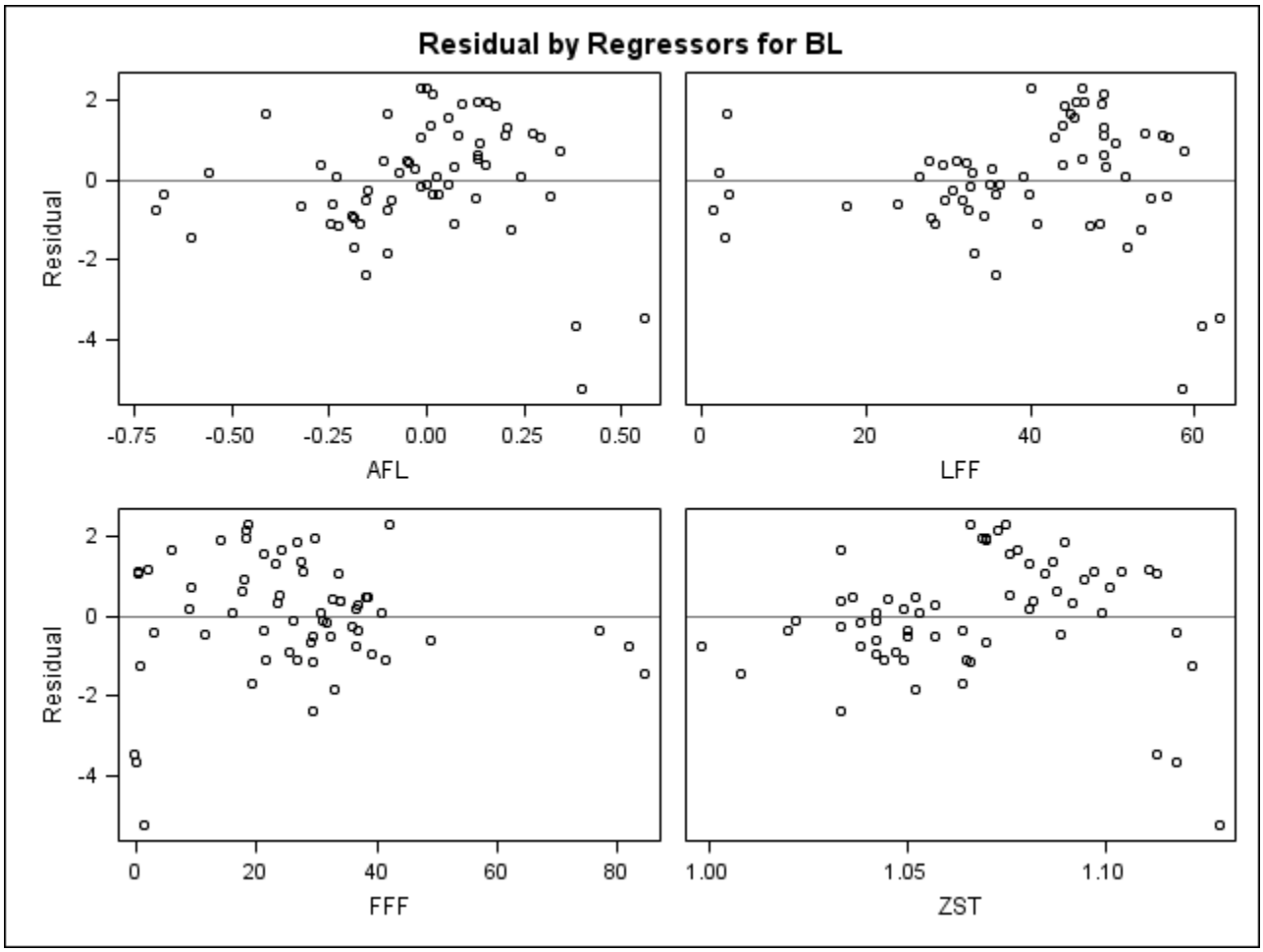
t-Test: Two-Sample Assuming Equal Variances		
	factor1	factor2
Mean	0.003131	0.00467
Variance	1.011205	0.973406
Observations	62	62
Pooled Variance	0.992306	
Hypothesized Mean Difference	0	
df	122	
t Stat	-0.00853	
P(T<=t) one-tail	0.496603	
t Critical one-tail	1.657651	
P(T<=t) two-tail	0.993206	
t Critical two-tail	1.97993	

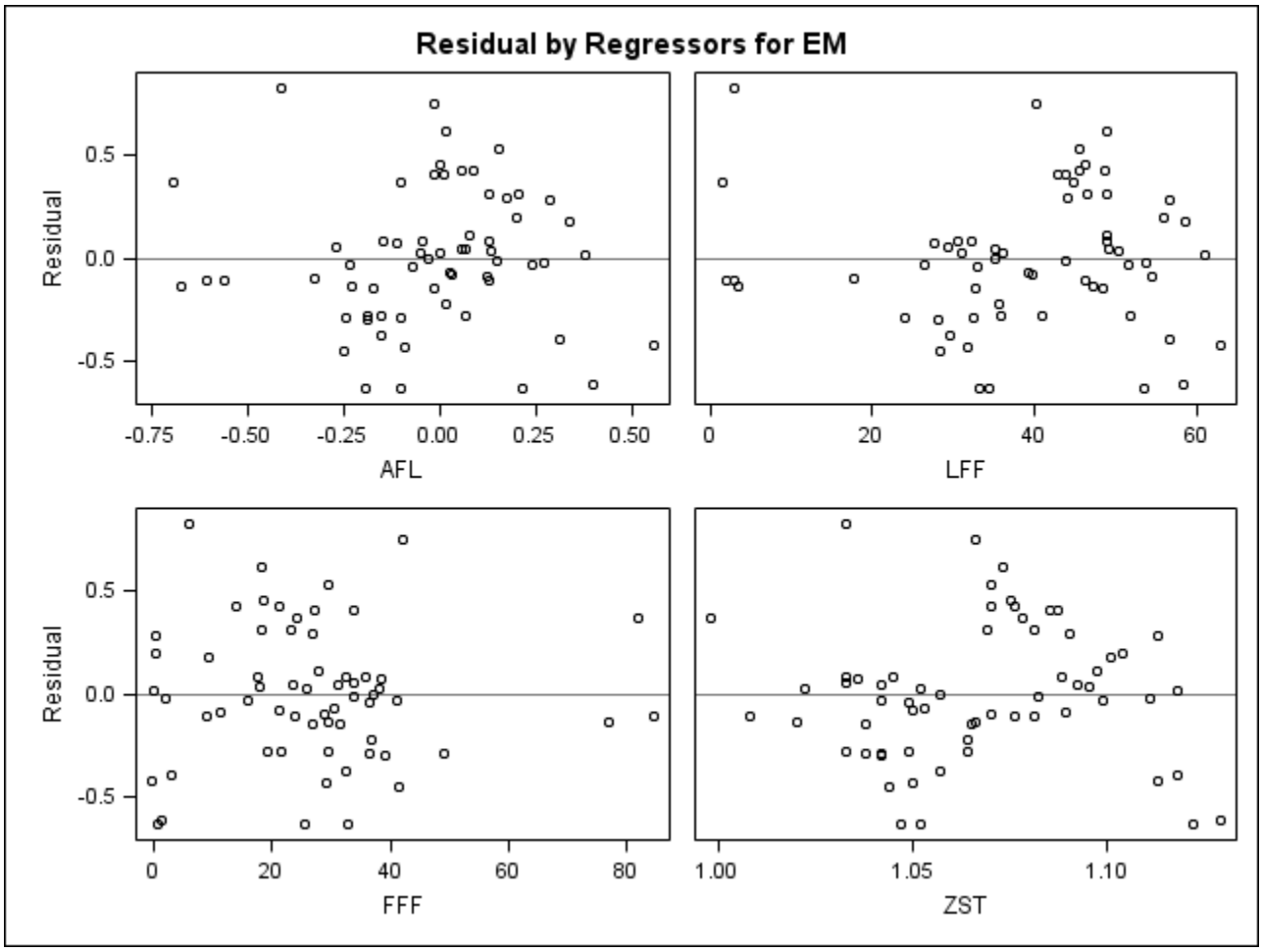
12.Simply regression analysis

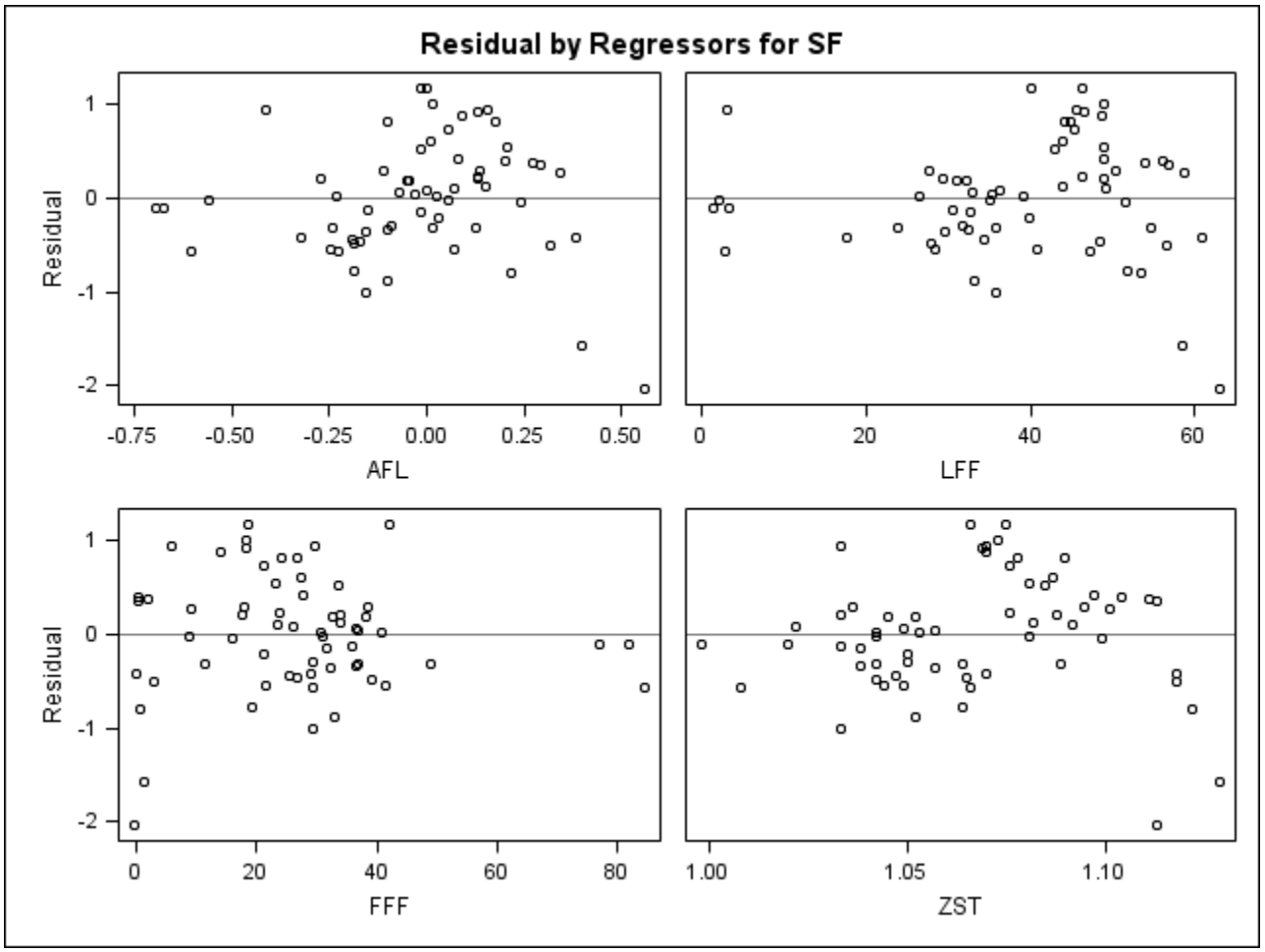
The table below summarizes the results of the individual regressions. Here the AFL, LFF, FFF and ZST treated as independent variable.

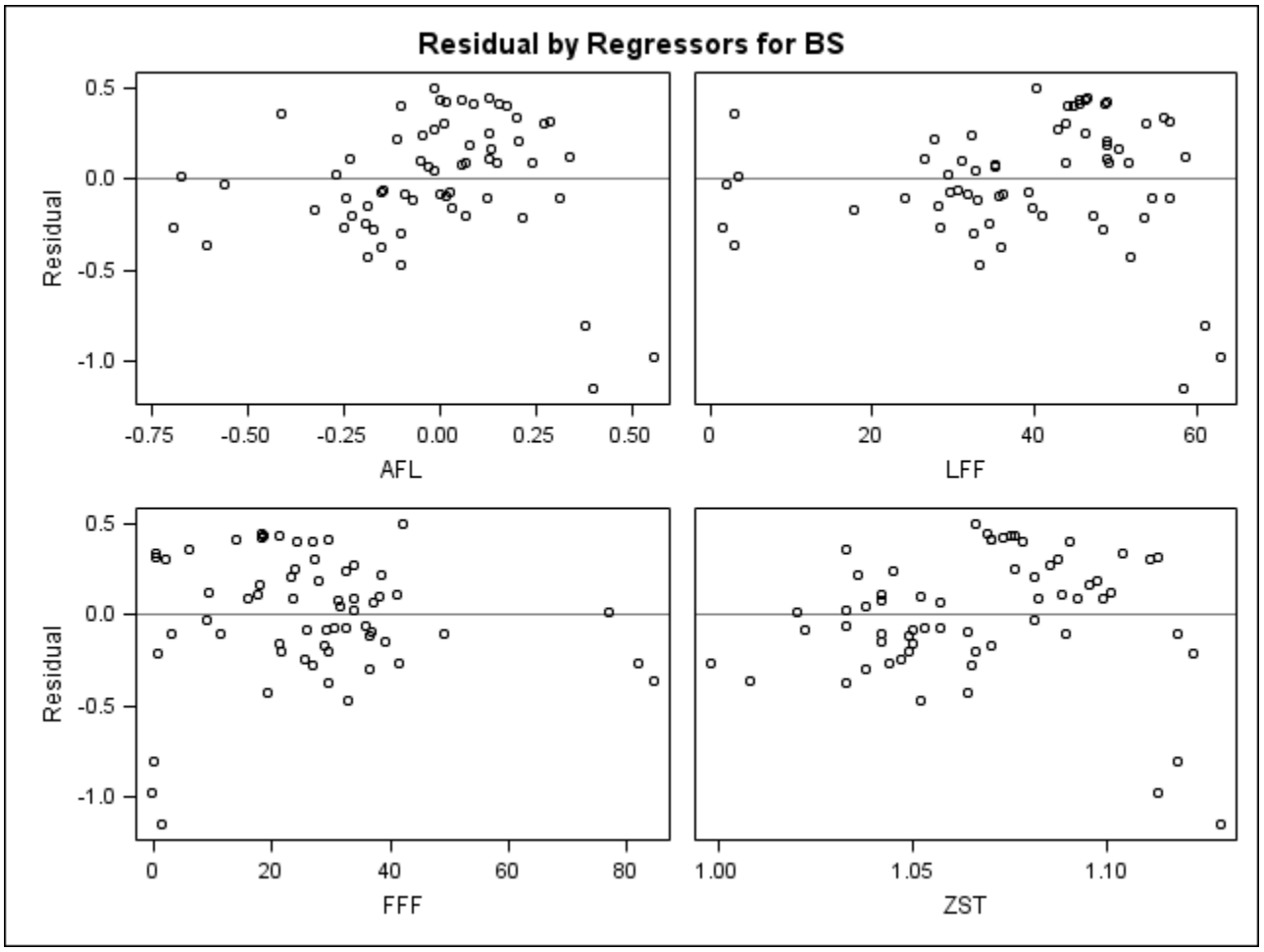
	(Intercept)	AFL	LFF	FFF	ZST	Adj R-Sq	MSE
BL	**-74.2**	**-3.12**	**0.1**	**0.05**	**85.08**	0.7297	1.4980
EM	**-24**	**-1.18**	**0.01**	**0.01**	**28.75**	0.7704	0.3433
SF	**-45.8**	**-1.49**	**0.05**	**0.03**	**45.8**	0.8043	0.6472
BS	**-17.7**	**-0.55**	**0.03**	**0.01**	**16.22**	0.7466	0.3488

Residual Analysis


Residual by Regressors for BL
Residual
AFL
LFF
FFF
ZST


Residual by Regressors for EM
Residual
AFL
LFF
FFF
ZST





Observations with large standardized residuals (outliers) include #51, #52 and #56. Observations

with high leverage include #57, #58, #60 and #61. Apart from the outliers, the residuals plots look good.

13. Canonical analysis

We have two set data, by observing the covariance value of whole dataset, there are highly correlated. Therefore, it is better to treat the two dataset as one entire data using the canonical analysis than separated factor analysis.

	Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)		Test of H0:				
		Eigenvalue	Cumulative	Likelihood Ratio	F Value	Num DF	Den DF	Pr>F
1	0.9173	5.3089	0.7175	0.0486	17.50	16	165.61	<.0001
2	0.8169	2.0063	0.9886	0.3066	9.31	9	134.01	<.0001
3	0.2653	0.0758	0.9989	0.9217	1.16	4	112	0.3305
4	0.0917	0.0085	1.0000	0.9916	0.48	1	57	0.4898

Here, Test of H0: The canonical correlations in the current row and all that follow are zero. Based on the canonical correlation, the first two factors have much higher value than others, and they contribute 71.8% and 27.1% to total variance respectively. Therefore, it maybe two factors is enough for current question, and it also verified by the p-value of Test of H0, only first two are significant.

Now we investigated the Correlations coefficient analysis, the two-set standard canonical variable can be written as

Paper quality U1 = -1.505*BL-0.212*EM +2.000*SF+0.676*BS

Fiber characteristic V1 = -0.159 *AFL +0.632* LFF+0.325* FFF+0.818* ZST

Paper quality U2 = -3.496*BL-1.543*EM +1.076*SF+3.768*BS

Fiber characteristic V2 = 0.689 *AFL +1.003* LFF+0.005* FFF-1.562* ZST

However, by further research the Correlations Between original variables and their canonical variables, and corresponding cross canonical correlations between two datasets, we find that (U1,V1) has much higher correlations with their component variables than (U2,V2), (U3,V3), and (U4,V4). Therefore, we prefer to use the (U1,V1) as main factor to represent this dataset.

- **Canonical Structure**

Correlations Between the BL-BS and Their Canonical Variables					Correlations Between the AFL-ZST and Their Canonical Variables				
	paper1	paper2	paper3	paper4		fiber1	fiber2	fiber3	fiber4
BL	*0.9351*	-0.1261	-0.0534	-0.3270	AFL	*0.8166*	0.3683	0.1661	0.4122
EM	*0.8869*	-0.4280	0.1306	-0.1148	LFF	*0.9056*	0.3848	0.1779	-0.0126
SF	*0.9767*	-0.1453	-0.0307	-0.1549	FFF	*-0.6496*	0.0123	-0.7309	-0.2087
BS	*0.9518*	0.0147	0.0127	-0.3061	ZST	*0.9395*	-0.2307	0.1851	0.1730

Correlations Between the BL-BS and the Canonical Variables of the AFL-ZST					Correlations Between the AFL-ZST and the Canonical Variables of the BL-BS				
	fiber1	fiber2	fiber3	fiber4		paper1	paper2	paper3	paper4
BL	*0.8578*	-0.1030	-0.0142	-0.0300	AFL	*0.7491*	0.3009	0.0441	0.0378
EM	*0.8136*	-0.3496	0.0346	-0.0105	LFF	*0.8307*	0.3144	0.0472	-0.0012
SF	*0.8960*	-0.1187	-0.0081	-0.0142	FFF	*-0.5959*	0.0100	-0.1940	-0.0191
BS	*0.8731*	0.0120	0.0034	-0.0281	ZST	*0.8618*	-0.1885	0.0491	0.0159

The first canonical variants seem good summary measures of their respective sets of variables. The correlations between canonical variables and their component variables match to the result of one factor analysis. Moreover, the first canonical variants, which might be labeled a "paper characteristic index" and "a pulp fiber strength-quality index", are highly correlated. There is a strong association between an index of pulp fiber characteristics and an index of the characteristics of paper made from them.

For the second and third canonical variable, only 2 moderately large correlation between the canonical variant and it's component variables is the correlation (0.428) between paper 2 with EM, and (0.7309) between fiber3 and FFF. Since we already test the H0 in the above tables, then the interaction between fiber3 and its component variables can be ignored. If we keep the fiber2 canonical variable, then it might be a "fiber length/strength" measure.

14. Conclusion

This dataset consists of paper quality and pulp fiber characteristics, both of them can use one factor for each of them, e.g. fiber characteristics factor and paper properties factor.

Paper quality U1 = -1.505*BL-0.212*EM +2.000*SF+0.676*BS

Fiber characteristic V1 = -0.159 *AFL +0.632* LFF+0.325* FFF+0.818* ZST

This dataset does not follow normal distribution, based on the result of normality test and QQ plot. Furthermore, we think some of the observations from #50-#61 are possible to be outlier, especially the #51, #52, #60 and #61. We highly suggest to do further research for these observations.