410-57, Assignment #6
Anamitra Bhattacharyya

INTRODUCTION

This assignment specifically deals with the use of Principal Components Analysis (PCA) as a method of dimension reduction and as a remedial measure for multicollinearity in Ordinary Least Squares regression. The raw data set being used in this assignment comprises the daily returns from 20 individual stocks from a variety of market sectors. The overall goal of the assignment is to compare regression models utilizing either all 20 stocks to build a regression model *versus* one that uses 8 stocks derived from PCA analysis. The analysis in this assignment also acts to detect, evaluate and minimize the effects of multicollinearity in multiple regression analysis.

RESULTS

1. Data Prep

Sort stock price data by date and calculate the log returns of twenty individual stocks *versus* the Vanguard (VV) index fund, and add it to the output table.

return_AA	return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_DPS	return_GS	return_HAL	return_HES	return_HON	return_HUN	return_JPM	return_KO	return_MMM	return_MPC	return_PEP	return_SLB	return_WFC	return_XOM	response_VV
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
0.023556	0.001723	0.009946	-0.001723	0.010906	0.005357	0.005462	-0.006523	0.028008	0.010222	-0.000900	-0.008073	-0.000858	-0.006293	0.008230499	0.010421	0.005107	-0.007590	0.004562	0.000232531	0.001202439
-0.009569	0.082555	-0.013874	-0.009850	-0.006829	0.006324	0.006206	-0.001690	-0.016074	-0.024015	0.001080	-0.005079	0.020672	-0.004890	-.004524356	-0.056044	-0.007822	-0.021653	0.015978	-.003027130	0.003256494
-0.021599	-0.020817	0.008621	-0.007267	-0.014234	0.005954	-0.006985	-0.012341	0.012080	-0.020699	-0.007403	0.008114	-0.009009	-0.006364	-.005144475	-0.008179	-0.012613	-0.004269	-0.002761	-.007490672	-.002055499
0.027989	0.014458	0.006223	0.010836	0.008435	-0.000330	0.000000	0.013503	0.011370	0.008472	0.008302	-0.006079	-0.001698	0.000000	0.005979449	-0.022358	0.005186	0.015227	0.012363	0.004454350	0.002226600
0.002121	0.055828	0.007148	-0.003935	0.015176	0.021864	0.002593	0.037721	0.026497	0.028757	0.016753	0.035932	0.021024	0.006076	0.005113884	0.027713	-0.001066	0.027658	0.003747	0.002569795	0.009196250
0.019927	0.035559	-0.034068	-0.011899	0.003388	0.014421	-0.012245	0.014438	-0.026497	-0.010644	-0.002123	0.037522	0.016779	-0.018632	-.006306917	0.032759	-0.009949	-0.008374	0.007115	-.007494180	0.001185938
0.030677	-0.011713	-0.038990	-0.026325	0.016772	0.035322	-0.004993	0.014430	-0.018543	-0.016181	0.012847	0.040709	0.005169	-0.007374	0.006069641	-0.007630	-0.006017	-0.006578	-0.000338	-.004004245	0.002367666
-0.013178	-0.026867	-0.005607	0.010613	0.006218	-0.016724	-0.003695	-0.022482	-0.023010	-0.008101	-0.008605	0.016187	-0.025561	-0.008323	-.008101069	0.011878	-0.003410	-0.024840	0.000000	0.001650749	-.004231915
-0.004090	-0.019863	-0.006686	0.005921	0.002888	0.019178	-0.005036	-0.013019	-0.002360	0.014745	0.008080	-0.032641	-0.028521	0.005359	0.007507632	0.015323	0.003874	-0.005161	0.007067	0.009497638	0.002541297

Conclusion: Created a sorted list of sorted daily stock prices and calculated a log of the ratio of today's price against yesterday's price.

2. Correlation between individual stocks and the market index

Perform a correlation using the Pearson correlation between the log of returns for each of the individual twenty stocks and the log return of the VV index fund. A table of results showing R-squared values for the correlations and the respective p-values is displayed below.

Pearson Correlation Coefficients, N = 501
Prob > |r| under H0: Rho=0

	return_AA	return_BAC	return_BHI	return_CVX	return_DD	return_DOW	return_DPS	return_GS	return_HAL	return_HES	return_HON	return_HUN	return_JPM	return_KO	return_MMM	return_MPC	return_PEP	return_SLB	return_WFC	return_XOM
response_VV	0.63241 <.0001	0.65019 <.0001	0.57750 <.0001	0.72090 <.0001	0.68952 <.0001	0.62645 <.0001	0.44350 <.0001	0.71216 <.0001	0.59750 <.0001	0.61080 <.0001	0.76838 <.0001	0.58194 <.0001	0.65785 <.0001	0.59980 <.0001	0.76085 <.0001	0.47312 <.0001	0.50753 <.0001	0.69285 <.0001	0.73357 <.0001	0.72111 <.0001

Conclusion: All the twenty comparisons *versus* the VV index show good correlations (highest is HON). The probabilities are low, indicting to reject the null hypothesis and suggesting all the correlations are significant.

3. SAS data formats: wide and long

SAS has two data formats 'wide' and 'long', this section transforms from the wide to the long format, and re-names the first data column to 'correlations', and the second data column is the stock ticker symbol. The output is shown below:

Obs	correlation	tkr
1	0.63241	AA
2	0.65019	BAC
3	0.57750	BHI
4	0.72090	CVX
5	0.68952	DD
6	0.62645	DOW
7	0.44350	DPS
8	0.71216	GS
9	0.59750	HAL
10	0.61080	HES
11	0.76838	HON
12	0.58194	HUN
13	0.65785	JPM
14	0.59980	KO
15	0.76085	MMM
16	0.47312	MPC
17	0.50753	PEP
18	0.69285	SLB
19	0.73357	WFC
20	0.72111	XOM

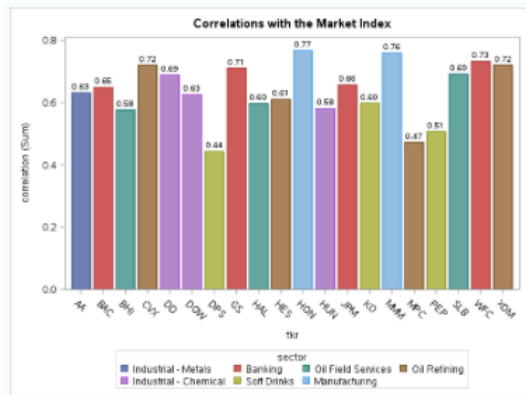
Conclusion: Using the transpose command in SAS it is possible to switch between wide and long data formats.

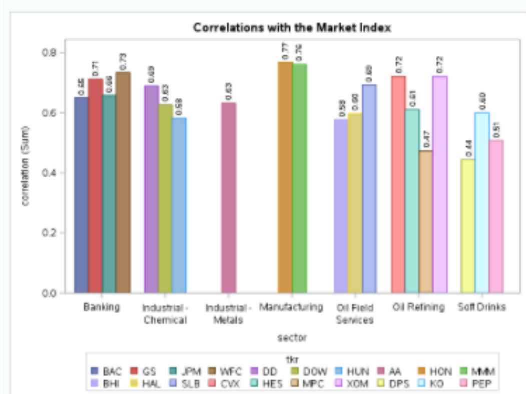
4. Visualization of correlations

In this section we merge a table of the stock ticker and correlation with a table of the stock ticker and business sector it is in. Pre-merging of the tables, they are both sorted by the ticker symbol s that they are in the same order. The resultant output from the merged table is below:

Obs	correlation	tkr	sector
1	0.63241	AA	Industrial - Metals
2	0.65019	BAC	Banking
3	0.57750	BHI	Oil Field Services
4	0.72090	CVX	Oil Refining
5	0.68952	DD	Industrial - Chemical
6	0.62645	DOW	Industrial - Chemical
7	0.44350	DPS	Soft Drinks
8	0.71216	GS	Banking
9	0.59750	HAL	Oil Field Services
10	0.61080	HES	Oil Refining
11	0.76838	HON	Manufacturing
12	0.58194	HUN	Industrial - Chemical
13	0.65785	JPM	Banking
14	0.59980	KO	Soft Drinks
15	0.76085	MMM	Manufacturing
16	0.47312	MPC	Oil Refining
17	0.50753	PEP	Soft Drinks
18	0.69285	SLB	Oil Field Services
19	0.73357	WFC	Banking
20	0.72111	XOM	Oil Refining

It is possible to visualize the correlations between the individual stocks and the VV index fund.





Conclusion: A bar plot can be generated of ticker symbol (x-axis) *versus* correlation (y-axis; response), shown in the upper panel above. The stocks can also be grouped together by virtue of those stocks that are in the same market sector, as shown in the bottom panel above.

5. Principal components

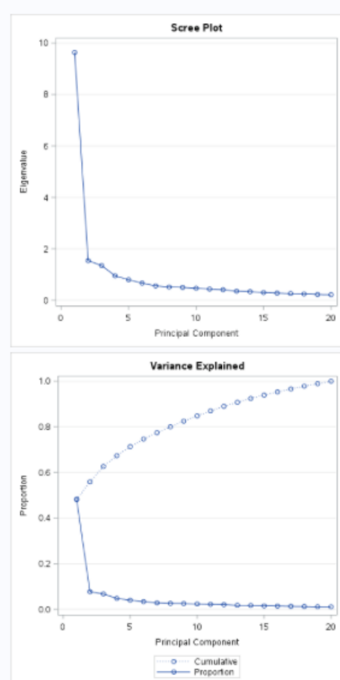
This step calculates the principal components for the data set using the SAS PRINCOMPS procedure.

How many principal components do you think that we should keep? Why?

There are numerous decision rules that could be used to select the principal components, the Kaiser rule for instance, utilizes the number of principal components that have eigen values > 1. In this case this would equate to 3 principal components (see correlation matrix table below, #s 1-3 have eigen values > 1). One can also use a 'scree plot' (see output graph, upper panel below) to plot out the sorted eigenvalues against the principal component number. The elbow coincides with the first three principal components and is consistent with the Kaiser rule method for selecting the principal component to choose.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	9.63645075	8.09792128	0.4818	0.4818
2	1.53852947	0.19109235	0.0769	0.5587
3	1.34743712	0.39975791	0.0674	0.6261
4	0.94767921	0.15217268	0.0474	0.6735
5	0.79550653	0.12909860	0.0398	0.7133
6	0.66640793	0.10798740	0.0333	0.7466
7	0.55842052	0.04567198	0.0279	0.7745
8	0.51274854	0.01590728	0.0256	0.8002
9	0.49684126	0.03250822	0.0248	0.8250
10	0.46433304	0.03089374	0.0232	0.8482
11	0.43343929	0.02568332	0.0217	0.8699
12	0.40775598	0.05667006	0.0204	0.8903
13	0.35108592	0.01597897	0.0176	0.9078
14	0.33510695	0.03813712	0.0168	0.9246
15	0.29696984	0.02068234	0.0148	0.9394
16	0.27628750	0.01692712	0.0138	0.9532
17	0.25936037	0.01730228	0.0130	0.9662
18	0.24205809	0.02020002	0.0121	0.9783
19	0.22185807	0.01013445	0.0111	0.9894
20	0.21172363		0.0106	1.0000

The PRINCOMP Procedure



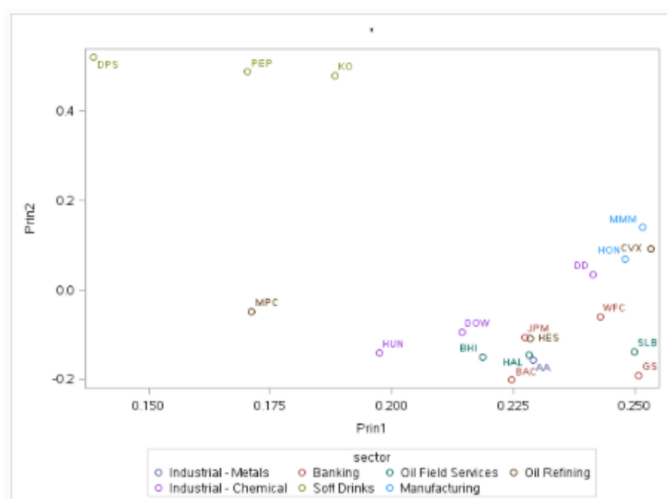
Later in the assignment we will use the first eight principal components. Why eight?

As noted above, one of the other criterion for selecting the principal components is to use as many components such that it would explain > 80% of the variation. The cumulative 'scree plot' shown in the output graph above (lower panel) visualizes the cumulative proportion of variance *versus* the principal component number. One 'rule of thumb' for selecting the number of principal components is to use as many as explains 80% of the variance, in this case, that value would satisfied by the first eight principal components, hence eight principal components are chosen later in this assignment.

We will also plot the first two principal components from the principal components analysis. When we plot them, we can see relationships in the data. Do we see any groupings (or clusters) in the plot of the first two principal components? Any surprises?

When we look at the first two principal components (based on the scree plot method) and add the sector information. The table below (left) shows the eigen vectors of the first two principal components. When we plot the eigen values of principal components 1 versus 2 (see plot below), we can see some clustering taking place. For example, the soft drinks stocks (*e.g.* DPS, PEP, COK) have higher coefficients in the eigen vectors while the banking stocks (*e.g.* BAC, GS, WFC, JPM) have negative values. This results in the soft drinks stocks segregating at the opposite end of the plot (below right) *versus* the banking and/or chemical (*e.g.* HUN, DOW, DD) sectors.

Obs	Prin1	Prin2	tkr	sector
1	0.22895	-0.15741	AA	Industrial - Metals
2	0.22464	-0.20081	BAC	Banking
3	0.21862	-0.15017	BHI	Oil Field Services
4	0.25320	0.09116	CVX	Oil Refining
5	0.24140	0.03434	DD	Industrial - Chemical
6	0.21430	-0.09442	DOW	Industrial - Chemical
7	0.13862	0.52040	DPS	Soft Drinks
8	0.25075	-0.19111	GS	Banking
9	0.22812	-0.14601	HAL	Oil Field Services
10	0.22843	-0.10832	HES	Oil Refining
11	0.24796	0.06892	HON	Manufacturing
12	0.19747	-0.14011	HUN	Industrial - Chemical
13	0.22735	-0.10656	JPM	Banking
14	0.18820	0.47871	KO	Soft Drinks
15	0.25149	0.13961	MMM	Manufacturing
16	0.17114	-0.04824	MPC	Oil Refining
17	0.17021	0.48760	PEP	Soft Drinks
18	0.24973	-0.13757	SLB	Oil Field Services
19	0.24282	-0.06011	WFC	Banking



If we had chose the first eight principal components and performed a correlation, we can see from the correlation matrix (see table below) that the eigenvalues off the diagonal are all zero indicating no multicollinearity and that the principal components are all orthogonal to each other.

The CORR Procedure

8 Variables:	Prin1 Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Prin1	501	0	3.10426	0	-10.76983	8.77566
Prin2	501	0	1.24037	0	-4.53262	3.87247
Prin3	501	0	1.16079	0	-4.62132	3.82720
Prin4	501	0	0.97349	0	-3.82800	3.74994
Prin5	501	0	0.89191	0	-2.67790	3.65876
Prin6	501	0	0.81634	0	-3.02537	3.10657
Prin7	501	0	0.74728	0	-3.02789	3.14927
Prin8	501	0	0.71606	0	-3.02497	3.52884

Pearson Correlation Coefficients, N = 501 Prob > |r| under H0: Rho=0								
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
Prin1	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	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	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	0.00000 1.0000	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	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin5	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000
Prin6	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000	0.00000 1.0000
Prin7	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000	0.00000 1.0000
Prin8	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	0.00000 1.0000	1.00000

6. Principal components in regression modeling

For cross-validation purposes the assignment created a train:test split of the data. The sample data set was split into a 70:30 training/test split. Using 70% of the data identified as the training data set, and we can 'test' each model by examining the predictive accuracy on the 30% of the data. We will use the response variable 'train_response' when fitting our models.

Let,
'0' = test set
'1' = training set

Obs	response_VV	u	train	train_response
1	.	0.75040	0	.
2	0.001202439	0.32091	1	0.001202439
3	0.003256494	0.17839	1	0.003256494
4	-.002055499	0.90603	0	.
5	0.002226600	0.35712	1	0.002226600
6	0.009196250	0.22111	1	0.009196250
7	0.001185938	0.78644	0	.
8	0.002367666	0.39808	1	0.002367666
9	-.004231915	0.12467	1	-.004231915
10	0.002541297	0.18769	1	0.002541297

7. Fitting the regression model

Using the training set data (70% split) with 338 observations, there is a statistically significant fit (F-statistic) and the p-value is low (see table below). The adjusted R-square value is high (>89%), so most (>89%) of the variance is explained using these variables.

Training set (70% of the data)

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	502
Number of Observations Used	338
Number of Observations with Missing Values	164

Analysis of Variance

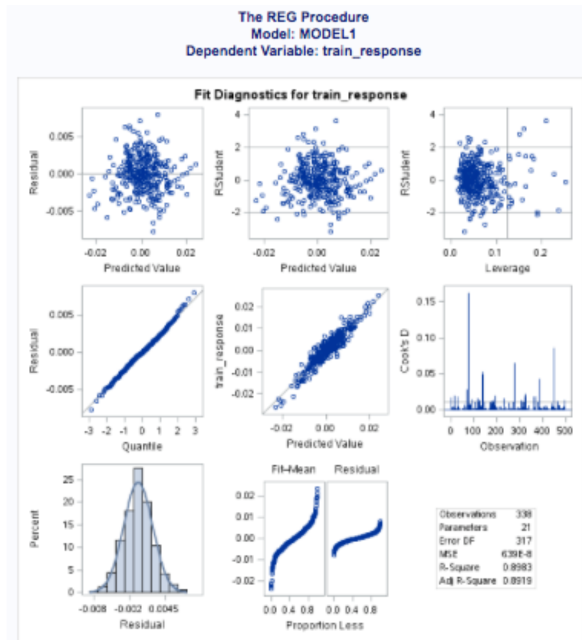
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	0.01790	0.00089510	140.04	<.0001
Error	317	0.00203	0.00000639		
Corrected Total	337	0.01993			

Root MSE	0.00253	R-Square	0.8983
Dependent Mean	0.00061635	Adj R-Sq	0.8919
Coeff Var	410.18453		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00008640	0.00014092	0.61	0.5403	0
return_AA	1	0.01769	0.01317	1.34	0.1802	2.11490
return_BAC	1	0.03198	0.01165	2.75	0.0064	3.10927
return_BHI	1	-0.00111	0.01323	-0.08	0.9333	2.62997
return_CVX	1	0.04907	0.02536	1.93	0.0539	3.07524
return_DD	1	0.04674	0.02037	2.29	0.0224	2.51406
return_DOW	1	0.03642	0.01162	3.14	0.0019	1.88893
return_DPS	1	0.03670	0.01679	2.19	0.0295	1.54768
return_GS	1	0.04849	0.01555	3.12	0.0020	3.10450
return_HAL	1	0.00948	0.01466	0.65	0.5184	3.08758
return_HES	1	0.00359	0.01092	0.33	0.7425	2.10199
return_HON	1	0.12213	0.01924	6.35	<.0001	2.73505
return_HUN	1	0.02712	0.00836	3.24	0.0013	1.79852
return_JPM	1	0.00902	0.01708	0.53	0.5979	3.36439
return_KO	1	0.07903	0.02226	3.55	0.0004	1.93633
return_MMM	1	0.09796	0.02646	3.70	0.0003	2.98277
return_MPC	1	0.01673	0.00809	2.07	0.0394	1.32999
return_PEP	1	0.02911	0.02231	1.30	0.1929	1.68825
return_SLB	1	0.03776	0.01709	2.21	0.0279	3.13690
return_WFC	1	0.07587	0.01848	4.10	<.0001	2.59492
return_XOM	1	0.05467	0.02697	2.03	0.0435	2.98393

Goodness-of-Fit (GOF): training set



<u>The variance inflation factors (VIFs) are not excessively high, indicating that multicollinearity is *not* a major issue for the training set</u>. Recall those VIFs > 5 need to be noted and values > 10 are serious. The values shown here (table left) seem to be in an acceptable range, mostly below 3. The GOF is also quite good, with the QQ plot (above) showing the observations mostly consist with a normal distribution and the residual analysis suggests a random distribution of points, with no geometric shapes apparent.

Using all the data with 501 observations, there is a statistically significant fit (F-statistic) and the p-value is low (see table below). The adjusted R-square value is high (>87%), so most (>87%) of the variance is explained using these variables.

All data set (100% of the data)

Goodness-of-Fit (GOF): all data

The REG Procedure
Model: MODEL1
Dependent Variable: response_VV

Number of Observations Read	502
Number of Observations Used	501
Number of Observations with Missing Values	1

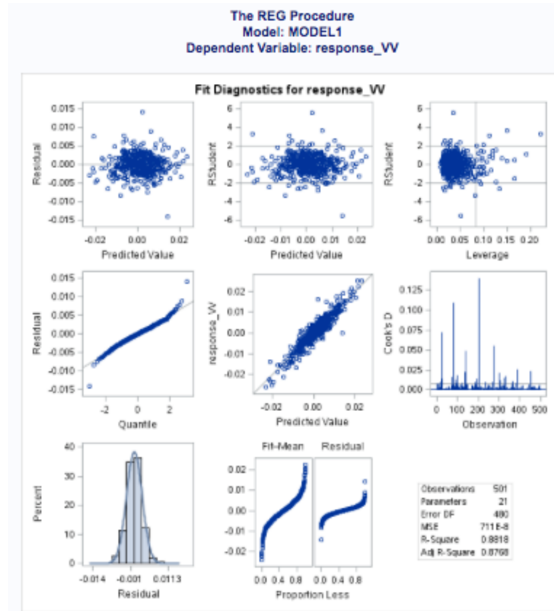
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	0.02543	0.00127	178.96	<.0001
Error	480	0.00341	0.00000711		
Corrected Total	500	0.02884			

Root MSE	0.00267	R-Square	0.8818
Dependent Mean	0.00075200	Adj R-Sq	0.8768
Coeff Var	354.45666		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00009953	0.00012133	0.82	0.4124	0
return_AA	1	0.01538	0.01040	1.48	0.1399	2.02627
return_BAC	1	0.02723	0.00970	2.81	0.0052	2.68654
return_BHI	1	0.01604	0.01161	1.38	0.1678	2.65137
return_CVX	1	0.05742	0.02068	2.78	0.0057	2.92019
return_DD	1	0.01003	0.01625	0.62	0.5371	2.44149
return_DOW	1	0.03600	0.01069	3.37	0.0008	1.96589
return_DPS	1	0.05659	0.01493	3.79	0.0002	1.52563
return_GS	1	0.03434	0.01358	2.53	0.0118	3.19781
return_HAL	1	-0.00198	0.01210	-0.16	0.8703	2.91992
return_HES	1	0.00439	0.00969	0.45	0.6504	2.09606
return_HON	1	0.10707	0.01608	6.66	<.0001	2.45588
return_HUN	1	0.02867	0.00722	3.97	<.0001	1.74391
return_JPM	1	0.02224	0.01329	1.67	0.0948	2.87596
return_KO	1	0.09425	0.01847	5.10	<.0001	1.98165
return_MMM	1	0.10928	0.02202	4.96	<.0001	2.68494
return_MPC	1	0.01079	0.00702	1.54	0.1251	1.37671
return_PEP	1	0.02092	0.02034	1.03	0.3043	1.72066
return_SLB	1	0.04851	0.01453	3.34	0.0009	3.25898
return_WFC	1	0.07738	0.01580	4.90	<.0001	2.53263
return_XOM	1	0.05797	0.02301	2.52	0.0121	2.94983



<u>As above, using all the data, the variance inflation factors (VIFs) are not high, indicating that multicollinearity is *not* a major issue for the training set</u>. The values shown here (table left) seem to be in an acceptable range, mostly below 3. The GOF is also quite good, with the QQ plot (above) showing the observations mostly consist with a normal distribution and the residual analysis suggests a random distribution of points, with no geometric shapes apparent.

8. Comparing regression models

This aspect of the assignment involved fitting a regression model using the rotated predictor variables (Principal Component Scores). This involved fitting a regression model using 8 selected principal components and VV as the response variable.

Training set (70% of the data)

The REG Procedure
Model: MODEL1
Dependent Variable: train_response

Number of Observations Read	502
Number of Observations Used	338
Number of Observations with Missing Values	164

Analysis of Variance

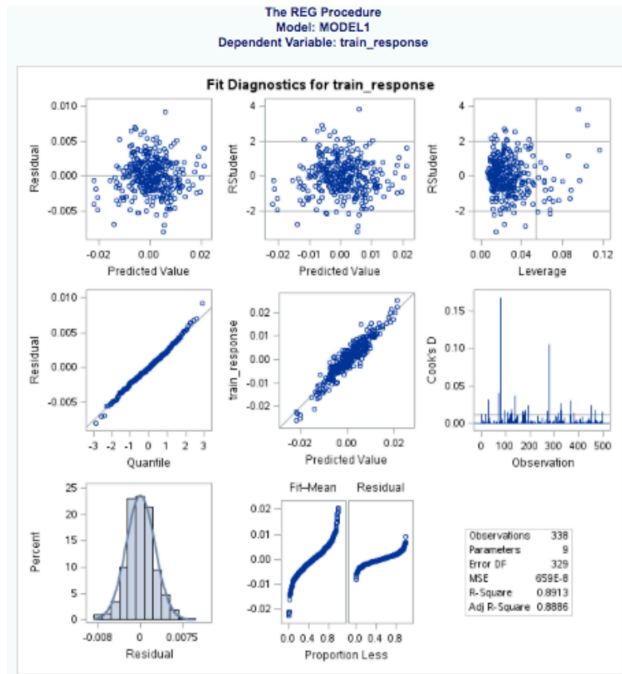
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	0.01776	0.00222	337.13	<.0001
Error	329	0.00217	0.00000659		
Corrected Total	337	0.01993			

Root MSE	0.00257	R-Square	0.8913
Dependent Mean	0.00061635	Adj R-Sq	0.8886
Coeff Var	416.36522		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00075978	0.00014045	5.41	<.0001	0
Prin1	1	0.00231	0.00004519	51.05	<.0001	1.00527
Prin2	1	0.00032245	0.00011425	2.82	0.0051	1.00868
Prin3	1	0.00070635	0.00012322	5.73	<.0001	1.00861
Prin4	1	0.00030481	0.00014536	2.10	0.0368	1.00636
Prin5	1	-0.00017356	0.00015516	-1.12	0.2641	1.00297
Prin6	1	0.00000315	0.00017108	0.02	0.9853	1.00766
Prin7	1	-0.00010331	0.00018604	-0.56	0.5791	1.02315
Prin8	1	-0.00040760	0.00020293	-2.01	0.0454	1.02271

Goodness-of-Fit (GOF): training set



Using 70% of the data in the training set produced a statistically significant F-statistic value. <u>The VIFs for the training for set using 8 principal components were all ideal in the VIF= 1 range</u>, suggesting no multicollinearity issues. The GOF analysis was similarly good, with residual scatter and QQ plots being reasonably optimal.

Using all the data, the results from the regression analysis showed the following results:

All data set (100% of the data)

The REG Procedure
Model: MODEL1
Dependent Variable: response_VV

Number of Observations Read	502
Number of Observations Used	501
Number of Observations with Missing Values	1

Analysis of Variance

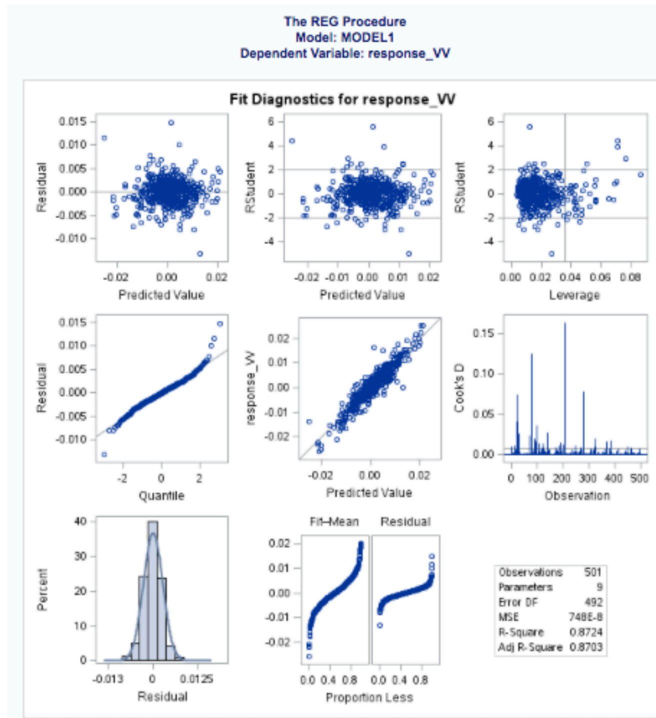
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	0.02516	0.00315	420.52	<.0001
Error	492	0.00368	0.00000748		
Corrected Total	500	0.02884			

Root MSE	0.00273	R-Square	0.8724
Dependent Mean	0.00075200	Adj R-Sq	0.8703
Coeff Var	363.66783		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.00075200	0.00012218	6.15	<.0001	0
Prin1	1	0.00227	0.00003940	57.55	<.0001	1.00000
Prin2	1	0.00045862	0.00009860	4.65	<.0001	1.00000
Prin3	1	0.00050372	0.00010536	4.78	<.0001	1.00000
Prin4	1	0.00015525	0.00012563	1.24	0.2171	1.00000
Prin5	1	-0.00018175	0.00013713	-1.33	0.1856	1.00000
Prin6	1	0.00008231	0.00014982	0.55	0.5830	1.00000
Prin7	1	-0.00003961	0.00016367	-0.24	0.8088	1.00000
Prin8	1	-0.00035470	0.00017080	-2.08	0.0383	1.00000

Goodness-of-Fit (GOF): all data



Using the entire data also produced a statistically significant F-statistic value (table above). Once again the VIFs using all the data set using 8 principal components were all ideal in the VIF= 1 range, suggesting no multicollinearity issues. The GOF analysis was similarly good, with residual scatter and QQ plots being reasonably optimal.

To complete the cross validation analysis, two models were compared using either the

(a) Model 1: using all 20 variables, in a regression model
(b) Model 2: using the 8 PCA regression

Model 1:

Obs	train	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	1	164	.000009306	.002144904
2	1	1	338	.000005994	.001902032

The FREQ Procedure

train=0

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	15	9.15	15	9.15
Grade 2	5	3.05	20	12.20
Grade 3	144	87.80	164	100.00

The FREQ Procedure

train=1

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	31	9.17	31	9.17
Grade 2	17	5.03	48	14.20
Grade 3	290	85.80	338	100.00

Obs	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	502	.000006807	.001949987

Model 2:

Obs	train	_TYPE_	_FREQ_	MSE_2	MAE_2
1	0	1	164	.000009677	.002179249
2	1	1	338	.000006410	.001975239

The FREQ Procedure

train=0

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	16	9.76	16	9.76
Grade 2	8	4.88	24	14.63
Grade 3	140	85.37	164	100.00

The FREQ Procedure

train=1

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	31	9.17	31	9.17
Grade 2	13	3.85	44	13.02
Grade 3	294	86.98	338	100.00

Obs	_TYPE_	_FREQ_	MSE_1	MAE_1
1	0	502	.000006807	.001949987

Both models 1 and 2 produce similar quality grades of prediction (Grade 1 within ± 10%; Grade 2 within ± 15%; Grade 3 the rest). The mean absolute error (MAE) is comparable for both models. However, model 2 uses the PCA approach and utilizes less predictors.

Finally, for model 2 using all the data (see table below, upper panel) the grade proportions are similar to the results shown above. Further comparing the MSE (mean square of the error) and MAE for models 1 and 2 (see table below, bottom panel) for the training set (train=1) and test set (train=0).

The FREQ Procedure

Prediction_Grade	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Grade 1	52	10.36	52	10.36
Grade 2	23	4.58	75	14.94
Grade 3	427	85.06	502	100.00

Obs	train	MSE_1	MAE_1	MSE_2	MAE_2
1	0	.000009306	.002144904	.000009677	.002179249
2	1	.000005994	.001902032	.000006410	.001975239

CONCLUSION

From the final comparison of the models 1 (using 20 variables) and 2 (using 8 principal components) while the metrics for comparison are similar, in terms of MSE and MAE, model 2 is preferred as it is more parsimonious since it utilizes 8 variables compared to all 20 variables.