MSDS 6372 Project 2

Team members:

Alva, Albert

Bhasin, Harinder

Thacker, Matthew

Date: March 19, 2017

Introduction:

Employment is a national concern and has been defined using many different measures. Since employment is a derived measure it is important to identify the factors that influence its variance. Some of the measures include gross national product, work force and population. This study attempts to determine if the measures selected describe the variance in employment.

The data used in this study was obtained from a previous study of employment and statistical modeling (Longley, 1967). The data set consisted of 13 continuous variables reporting measures of GNP, employment and population (figure 1).

Label	Variable	Notes
Year	Calendar Year	1947 – 1962
Totem	Total Derived Employment	
Agemp	Agricultural Employment	
Selfemp	Self Employed	
Famwork	Unpaid family workers	
Domestic	Domestic	
Nonagpr	Non-agricultural private jobs	
Fedgov	Federal government	
Nonfedgv	State and local government	
Gnpdef	Gross national product deflator	(1954=100)
Gnp	Gross national product	
Unemp	Unemployment	
Armedfor	Size of armed forces	
Pop	Population 14 years and over	

Fig 1. Variables used for analysis

Descriptive Statistics:

The data consisted of 16 observations of the 13 measures and employment. There were various differences in means and standard deviations (figure 2). This was expected with the multiple units of measure. A matrixed scatterplot shows many of the variables are highly correlated. The histograms show no evidence against normality (figure 3).

The MEANS Procedure

Variable	N	Mean	Minimum	Maximum	Std Dev	Variance
year	16	1954.50	1947.00	1962.00	4.7609523	22.6666667
totemp	16	65317.00	60171.00	70551.00	3511.97	12333921.73
agemp	16	6636.75	5190.00	8256.00	930.8155922	866417.67
selfemp	16	6068.38	5670.00	6388.00	213.5699339	45612.12
famwork	16	510.0000000	396.0000000	662.0000000	101.7585377	10354.80
domestic	16	2167.56	1714.00	2626.00	318.1316224	101207.73
nonagpr	16	42819.56	37922.00	46652.00	2846.30	8101403.20
fedgov	16	2170.75	1863.00	2420.00	174.4352793	30427.67
nonfedgv	16	4944.00	3582.00	6849.00	1048.95	1100302.40
gnpdef	16	101.6812500	83.0000000	116.9000000	10.7915534	116.4576250
gnp	16	387698.44	234289.00	554894.00	99394.94	9879353659
unemp	16	3193.31	1870.00	4806.00	934.4642471	873223.43
armedfor	16	2606.69	1456.00	3594.00	695.9196044	484304.10
pop	16	117424.00	107608.00	130081.00	6956.10	48387348.93

Fig 2. Descriptive statistics of the variables

Fig 3. Scatterplot and histogram of variables

Pearson Correlation Coefficients, N = 16
Prob > |r| under H0: Rho=0

	year	totemp	agemp	selfemp	famwork	domestic	nonagpr	fedgov	nonfedgv	gnpdef	gnp	unemp	armedfor	pop
year	1.00000	0.97133 <.0001	-0.97511 <.0001	0.49751 0.0499	0.92486 <.0001	0.96322 <.0001	0.93718 <.0001	0.67327 0.0043	0.97925 <.0001	0.99115 <.0001	0.99527 <.0001	0.66826 0.0047	0.41725 0.1078	0.99395 <.0001
totemp	0.97133 <.0001	1.00000	-0.94308 <.0001	0.41607 0.1089	0.92555 <.0001	0.94963 <.0001	0.98266 <.0001	0.70604 0.0022	0.93860 <.0001	0.97090 <.0001	0.98355 <.0001	0.50250 0.0473	0.45731 0.0749	0.96039 <.0001
agemp	-0.97511 <.0001	-0.94308 <.0001	1.00000	-0.36156 0.1688	-0.84653 <.0001	-0.91948 <.0001	-0.93877 <.0001	-0.77999 0.0004	-0.93325 <.0001	-0.98244 <.0001	-0.97724 <.0001	-0.58245 0.0179	-0.55771 0.0248	-0.96431 <.0001
selfemp	0.49751 0.0499	0.41607 0.1089	-0.36156 0.1688	1.00000	0.58931 0.0163	0.54015 0.0308	0.25001 0.3504	-0.19593 0.4671	0.64349 0.0072	0.45995 0.0730	0.46656 0.0685	0.78530 0.0003	-0.53769 0.0317	0.54664 0.0284
famwork	0.92486 <.0001	0.92555 <.0001	-0.84653 <.0001	0.58931 0.0163	1.00000	0.93444 <.0001	0.86227 <.0001	0.47731 0.0615	0.93049 <.0001	0.91093 <.0001	0.91523 <.0001	0.66700 0.0048	0.20338 0.4500	0.92022 <.0001
domestic	0.96322 <.0001	0.94963 <.0001	-0.91948 <.0001	0.54015 0.0308	0.93444 <.0001	1.00000	0.89713 <.0001	0.60163 0.0137	0.95707 <.0001	0.95417 <.0001	0.96114 <.0001	0.65845 0.0055	0.30474 0.2511	0.95561 <.0001
nonagpr	0.93718 <.0001	0.98266 <.0001	-0.93877 <.0001	0.25001 0.3504	0.86227 <.0001	0.89713 <.0001	1.00000	0.78842 0.0003	0.87185 <.0001	0.94646 <.0001	0.95573 <.0001	0.37968 0.1469	0.59909 0.0142	0.91505 <.0001
fedgov	0.67327 0.0043	0.70604 0.0022	-0.77999 0.0004	-0.19593 0.4671	0.47731 0.0615	0.60163 0.0137	0.78842 0.0003	1.00000	0.56149 0.0236	0.72165 0.0016	0.71081 0.0020	0.05248 0.8469	0.89350 <.0001	0.64108 0.0074
nonfedgv	0.97925 <.0001	0.93860 <.0001	-0.93325 <.0001	0.64349 0.0072	0.93049 <.0001	0.95707 <.0001	0.87185 <.0001	0.56149 0.0236	1.00000	0.96283 <.0001	0.97336 <.0001	0.73599 0.0012	0.24911 0.3522	0.99121 <.0001
gnpdef	0.99115 <.0001	0.97090 <.0001	-0.98244 <.0001	0.45995 0.0730	0.91093 <.0001	0.95417 <.0001	0.94646 <.0001	0.72165 0.0016	0.96283 <.0001	1.00000	0.99159 <.0001	0.62063 0.0103	0.46474 0.0697	0.97916 <.0001
gnp	0.99527 <.0001	0.98355 <.0001	-0.97724 <.0001	0.46656 0.0685	0.91523 <.0001	0.96114 <.0001	0.95573 <.0001	0.71081 0.0020	0.97336 <.0001	0.99159 <.0001	1.00000	0.60426 0.0132	0.44644 0.0830	0.99109 <.0001
unemp	0.66826 0.0047	0.50250 0.0473	-0.58245 0.0179	0.78530 0.0003	0.66700 0.0048	0.65845 0.0055	0.37968 0.1469	0.05248 0.8469	0.73599 0.0012	0.62063 0.0103	0.60426 0.0132	1.00000	-0.17742 0.5109	0.68655 0.0033
armedfor	0.41725 0.1078	0.45731 0.0749	-0.55771 0.0248	-0.53769 0.0317	0.20338 0.4500	0.30474 0.2511	0.59909 0.0142	0.89350 <.0001	0.24911 0.3522	0.46474 0.0697	0.44644 0.0830	-0.17742 0.5109	1.00000	0.36442 0.1662
pop	0.99395 <.0001	0.96039 <.0001	-0.96431 <.0001	0.54664 0.0284	0.92022 <.0001	0.95561 <.0001	0.91505 <.0001	0.64108 0.0074	0.99121 <.0001	0.97916 <.0001	0.99109 <.0001	0.68655 0.0033	0.36442 0.1662	1.00000

Fig 4. Correlation of variables

The Person Correlation Coefficient Matrix (figure 4) shows most variables correlated but the correlations seem logical. Agricultural employment was negatively correlated with all other variables. As the other sectors of employment increased, agricultural employment decreased. Agricultural employment decreased every year of the observations.

Due to the data being highly correlated in nature and the correlations seem linear, principle components analysis using Proc PRINCOMP procedure in SAS followed by the Proc PCL procedure using the correlation matrix and multiple linear regression on the components.

Principal Component Analysis

The PCA started with generating the eigenvectors for the components and a Scree plots of the components to begin to visualize the contribution of each component identified. These plots and tables were used to select the components of the model used in the final regression model used to test the hypothesis that these variables explain the variance in the employment number. The SAS PROC PRINCOMP procedure was used.

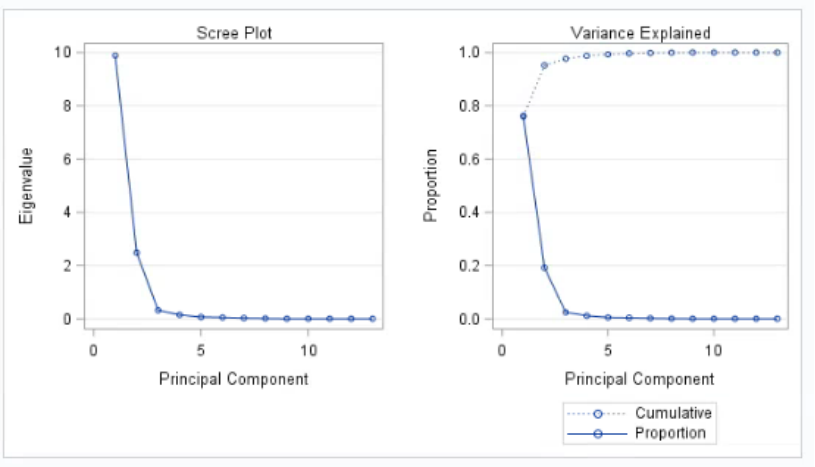

Fig 5. Scree Plot and variance explained

Though the Scree plot has an obvious 'elbow' at 3 components, there is a less apparent bend at 2 components. The proportion of variance graph showed little benefit of the third component since the difference between the 3rd and 4th components was only an additional 0.0567 to the variance explained by the model. The increase flattened out through the balance of the components. In order to more clearly determine which components to select the PROC PLS with cross validation and the PCR method was run.

The PLS Procedure

Percent Variation Accounted for by Principal Components

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	76.0589	76.0589	94.7834	94.7834
2	19.1069	95.1659	0.7655	95.5489
3	2.4595	97.6254	3.8894	99.4383

Fig 6. Accounted Variation

Figure 6 shows explained variation is at 94.78% with only the first factor. The second factor only adds 0.7655% more to the explanation of the dependent variable. By the third component, the addition drops off after point 1 for the model effects and drops precipitously after point 3 for the dependent variables. Given this data the first two principal components were selected to be used as factors in multiple linear regression.

	Prin1	Prin2
agemp	-.310843	0.094222
armedfor	0.134339	-.559584
selfemp	0.157520	0.529900
unemp	0.208306	0.385931
fedgov	0.219255	-.433135
famwork	0.295257	0.119281
nonagpr	0.298677	-.164830
domestic	0.307911	0.057099
nonfedgv	0.311431	0.113035
pop	0.315674	0.039661
gnpdef	0.316284	-.030874
gnp	0.316862	-.025345
year	0.317463	0.006525

Fig 7. Eigenvector for principle components

Based on the major factors in each component we would base these as being representative of the following: Prin1 is seems to be loading on economic measures gross national product, the population and the year (figure 7). All are important factors when determining employment. Higher GNP, larger population and year all increase employment. Prin2 seems to be loading on types of employment. Armed Forces and Government employment decrease the employment measure while self-employment and unemployment both increased employment.

Multiple Regression

The model being tested is employement = $\beta_0 + \beta_1$ Prin1 + β_2 Prin2. The coefficient β_0 represents the mean employment given the data while the other three represent the magnitude and direction (negative or positive) of each principal component used in the model. The F-test resulting from this model is significant (p-value of <0.0001) showing that our model is apporpriate for explaining a large amount of the variation. But the parameter estimates showed principle 2 did not add to the model with the confidence interval of the estimate crossing zero.

As a result of the initial findings, the model was changed to Employment = β0 + β1 Prin1. The resulting model yielded a higher F value than the model with 2 components 254.37 compared to 139.53 with 2 components. This shows the model explains the variation in the employment.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	175357592	175357592	254.37	<.0001
Error	14	9651234	689374		
Corrected Total	15	185008826			

Fig 8. Overall F-test

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	65317	207.57135	314.67	<.0001	64872	65762
Prin1		1	1087.35136	68.17655	15.95	<.0001	941.12721	1233.57552

Fig 9. Parameter Estimates for regression model

The parameter estimate is significant, Prin1 has a p-value of <0.0001 (fig 9). The estimate shows a change in prin1 changes the employment measure by 1087.35.

Root MSE	830.28541	R-Square	0.9478
Dependent Mean	65317	Adj R-Sq	0.9441
Coeff Var	1.27116		

Fig 10. R^2 table

The R2 value of the model was .9478. This high R^2 shows the variables associated with principle one, economic factors, population and year, explain almost all of the variance in employment.

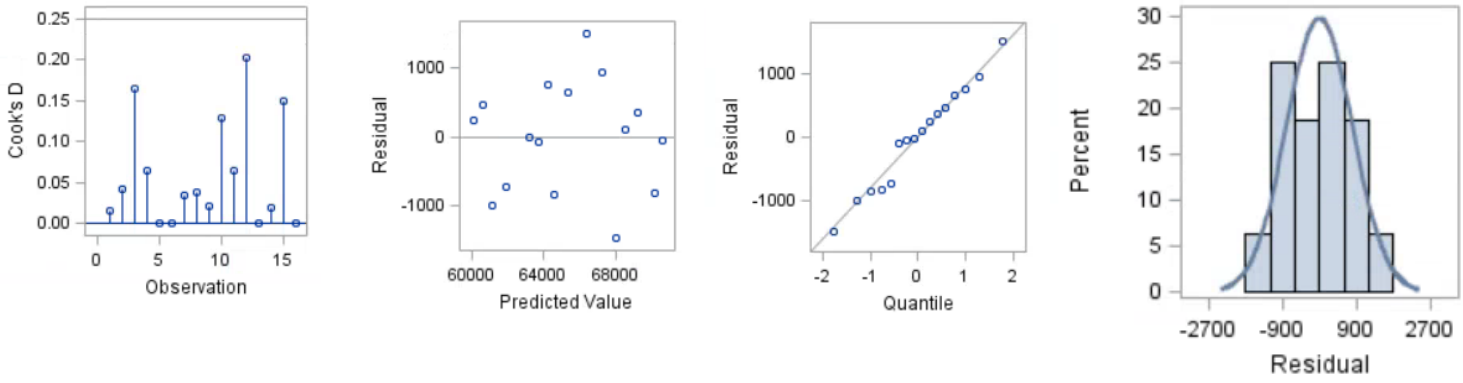

Fig 11 Model Residuals and Cook's D

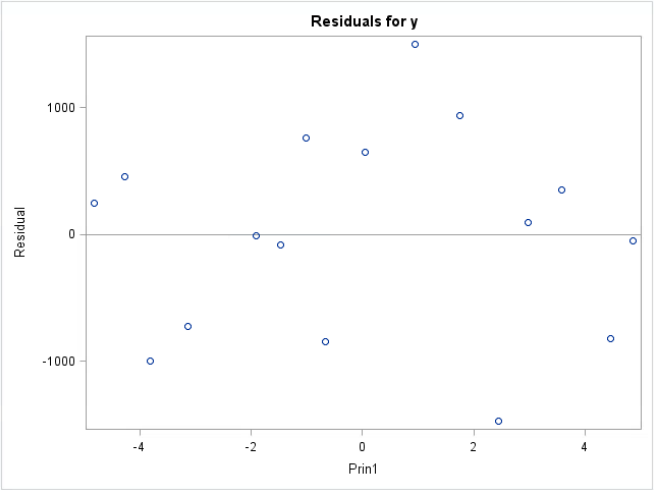

Fig 12 Residuals for component 1

The residuals in figure 11 show no observation has a high Cooks d. The predicted values shows no pattern thus good independence. The QQ plot and the histogram show a normal distribution. The residuals of the component show no evidence of non-consistency.

Conclusions

Most of the variation in employment can be explained with one principle component. Employment is highly related to the population, year and the health of the economy. One can look over the years at the differences in employment based on the number of workers available and how well the economy is producing. As production and population increase, employment increases. It is logical that these two variables along with year can explain over 94 percent of the variance in the number employed. The number of jobs needed and the number of people to fill them explain how many people are employed.

References

J. Longley (1967) "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User", Journal of the American Statistical Association, vol. 62. September, pp. 819-841.