Women and Work

An Analysis of Factors Affecting Female Labor Supply in 1975

STAT 448

Professor Yeon Joo Park

Group 4

Jaemin Kim

Introduction:

For final project, my group(Group 4) researched among numerous datasets to see if any specific dataset was interesting as well as available in clear format for us to easily perform analysis with. The data we confirmed to use was about women and their work conditions from Mroz, T. (1987) "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions", Econometrica, 55, 765-799. It contains data of wife's work conditions and environments, and it also provides information to compare to the conditions of husband's. It is also a basic dataset uploaded into the R repository, so the data was very credible and trustworthy. Also, it had good mix of continuous variables and categorical variables with adequate number of observation counts for our group's analysis. There are total 753 observations with 18 variables initially, 2 of them being categorical variable and 17 being continuous variable. However, I removed *work*, and w*agew* from the original dataset for my analysis because *work* variable displayed irrelevant result that went against results of other variables, and I also removed w*agew* since that data was collected in 1976 while all other variables were collected in 1975, plus w*agew* had a very similar overlapping information to *hearnw*. In addition, I converted *city* variable to *city_num* as it was in factor format originally, but I switched it to numerical. The final number of variables I chose was 16 and the detailed description of my data is listed below:

1. *hoursw*: wife's hours of work in 1975
2. *child6*: number of children less than 6 years old in household
3. *child618*: number of children between ages 6 and 18 in household
4. *agew*: wife's age
5. *educw*: wife's educational attainment, in years
6. *hearnw*: wife's average hourly earnings, in 1975 dollars
7. *hoursh*: husband's hours worked in 1975

8. *ageh*: husband's age

9. *educh*: husband's educational attainment, in years

10. w*ageh*: husband's wage, in 1975 dollars

11. *income*: family *income*, in 1975 dollars

12. *educwm*: wife's mother's educational attainment, in years

13. *educwf*: wife's father's educational attainment, in years

14. *unemprate*: unemployment rate in county of residence, in percentage points

15. *experience*: actual years of wife's previous labor market *experience*

16. *city_num*: lives in large city (SMSA) ? (Binary)

With this dataset, I decided among my group to perform Principal Component Analysis to investigate the significance and correlation between the variables, and use the correlation matrix, eigenvalues, and eigenvector table to find the optimal PCs, which will then be used to find relationships in our model.

Methods and Results:

The main goal of performing Principle Component Analysis(PCA) is to reduce data into fewer dimensions (or variables), retain large amount of information from original variables in the process of reducing variables, remove highly correlated variables and issues related to multicollinearity, and try to pick features out of the original data from chosen principal components. I will explain each step I went through to perform PCA and draw conclusion based on the results I obtained from the analysis.

After cleaning the dataset for PCA that I chose to perform, I ran a basic diagnostic of the 16 variables I have. Looking into the Simple Statistics table, I was able to find some contrasting values between wives

and husbands.

Simple Statistics										
	hoursw	child6	child618	agew	educw	hearnw	hoursh	ageh	educh	wageh
Mean	740.5763612	0.2377158035	1.353253652	42.53784861	12.28685259	2.374565339	2267.270916	45.12084993	12.49136786	7.482178752
StD	871.3142158	0.5239590379	1.319873930	8.07257401	2.28024583	3.241829080	595.566649	8.05879349	3.02080417	4.230558740

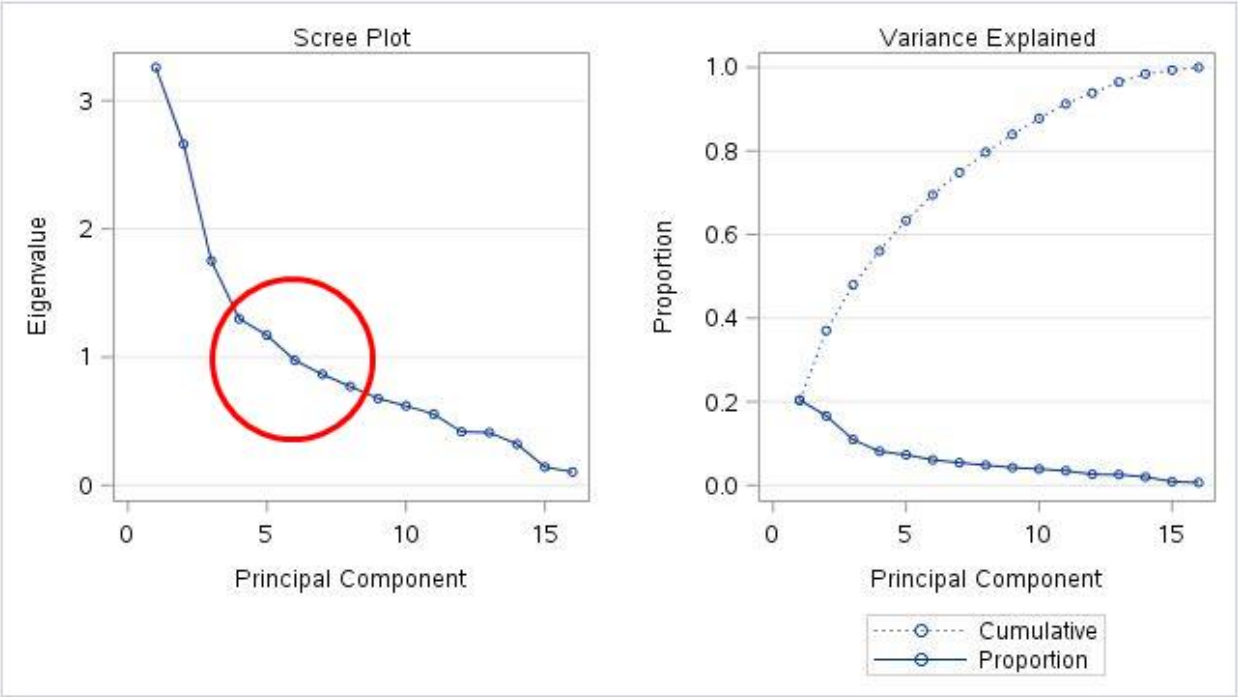
At a glance, I found out that mean values of wife's hours of work and average earning are 741 and 2.37 respectively. However, the mean values of the same variables for husbands were 2267 and 7.48. The husbands worked almost three times more than wives and earned about three times more than wives as well. Here, I could infer that husbands were the more working group and a main source of *income* for married couples in 1975. Then, I went on with further analysis using *proc princomp* function in SAS. From the correlation matrix result, I was able to see that first, the values in a diagonal line across the table were all 1 because it is a correlation-based PCA and the average eigenvalue is 1. By default, the table is ordered from most variation to least. Looking at the top left values, for example, we can see that *hoursw*(wife's working hours) is strongly negatively correlated with *child6*(Children under 6). Here, we can grasp how the variables are correlated and can see estimate trends and the strongest and weakest correlations.

Correlation Matrix																
	hoursw	child6	child618	agew	educw	hearnw	hoursh	ageh	educh	wageh	income	educwm	educwf	unemprate	experience	city_num
hoursw	1.0000	-.2221	-.0906	-.0331	0.1060	0.4229	-.0563	-.0311	-.0097	-.0986	0.1463	0.0579	0.0137	-.0603	0.4050	-.0162
child6	-.2221	1.0000	0.0842	-.4339	0.1087	-.1229	0.0243	-.4430	0.1336	0.0324	-.0278	0.1078	0.0961	-.0091	-.1940	-.0426
child618	-.0906	0.0842	1.0000	-.3854	-.0589	-.0473	0.0994	-.3502	0.0094	-.0297	-.0195	0.0324	-.0268	0.0143	-.2995	-.0336
agew	-.0331	-.4339	-.3854	1.0000	-.1202	-.0346	-.0844	0.8881	-.1630	0.0270	0.0524	-.2346	-.1606	0.0771	0.3340	0.0964
educw	0.1060	0.1087	-.0589	-.1202	1.0000	0.3184	0.0789	-.1335	0.6120	0.2849	0.3613	0.4353	0.4425	0.0721	0.0663	0.1595
hearnw	0.4229	-.1229	-.0473	-.0346	0.3184	1.0000	-.0599	-.0318	0.1262	0.0613	0.2312	0.0903	0.0985	-.0001	0.2506	0.0654
hoursh	-.0563	0.0243	0.0994	-.0844	0.0789	-.0599	1.0000	-.0954	0.1078	-.2360	0.1281	0.0534	0.0503	-.1552	-.0994	-.1126
ageh	-.0311	-.4430	-.3502	0.8881	-.1335	-.0318	-.0954	1.0000	-.1953	0.0197	0.0405	-.2275	-.1350	0.0531	0.2719	0.0676
educh	-.0097	0.1336	0.0094	-.1630	0.6120	0.1262	0.1078	-.1953	1.0000	0.3947	0.3769	0.3245	0.3667	0.0550	-.0363	0.2333
wageh	-.0986	0.0324	-.0297	0.0270	0.2849	0.0613	-.2360	0.0197	0.3947	1.0000	0.7250	0.1267	0.1932	0.1579	-.1033	0.3234
income	0.1463	-.0278	-.0195	0.0524	0.3613	0.2312	0.1281	0.0405	0.3769	0.7250	1.0000	0.1599	0.2111	0.0615	-.0277	0.2451
educwm	0.0579	0.1078	0.0324	-.2346	0.4353	0.0903	0.0534	-.2275	0.3245	0.1267	0.1599	1.0000	0.5731	0.0184	-.0822	0.0680
educwf	0.0137	0.0961	-.0268	-.1606	0.4425	0.0985	0.0503	-.1350	0.3667	0.1932	0.2111	0.5731	1.0000	0.0585	-.0788	0.1518
unemprate	-.0603	-.0091	0.0143	0.0771	0.0721	-.0001	-.1552	0.0531	0.0550	0.1579	0.0615	0.0184	0.0585	1.0000	0.0044	0.1840
experience	0.4050	-.1940	-.2995	0.3340	0.0663	0.2506	-.0994	0.2719	-.0363	-.1033	-.0277	-.0822	-.0788	0.0044	1.0000	0.0112
city_num	-.0162	-.0426	-.0336	0.0964	0.1595	0.0654	-.1126	0.0676	0.2333	0.3234	0.2451	0.0680	0.1518	0.1840	0.0112	1.0000

The next table to analyze is the Eigenvalues of the Correlation Matrix. This table contains the most essential and crucial information in determining how many Principal Components(PCs) to keep. There are three main rules we learned in class that we refer to when deciding the number of PCs.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.25972210	0.59961385	0.2037	0.2037
2	2.66010824	0.90990185	0.1663	0.3700
3	1.75020639	0.45088893	0.1094	0.4794
4	1.29931746	0.12799511	0.0812	0.5606
5	1.17132235	0.19639379	0.0732	0.6338
6	0.97492856	0.11018693	0.0609	0.6947
7	0.86474163	0.09429428	0.0540	0.7488
8	0.77044735	0.09298178	0.0482	0.7969
9	0.67746558	0.05788322	0.0423	0.8393
10	0.61958235	0.06402707	0.0387	0.8780
11	0.55555528	0.13855356	0.0347	0.9127
12	0.41700172	0.00598508	0.0261	0.9388
13	0.41101664	0.08932929	0.0257	0.9645
14	0.32168735	0.18014192	0.0201	0.9846
15	0.14154543	0.03619387	0.0088	0.9934
16	0.10535156		0.0066	1.0000

First, I had to set a certain cut-off percentage from Cumulative column that determines how much of the total variation I would like to maintain. I chose 70% for my model because the usual cut-off interval is from 70% - 90% and I thought 70% is enough and that I may have too many PCs left if I have higher cut-off percentage. Second, I looked at the Eigenvalues column to keep components larger than average eigenvalues. Because we are using correlation-based PCA, the average eigenvalue is 1, so I looked for PCs that have values greater than 1. Lastly, I also tried to find an elbow point in the scree plot of eigenvalues. However, the elbow point was not clear as shown in the plot below.



From the three decision making rules, I decided to keep 6 PCs for my model as it is closest to the criteria I set above as it explains approximately 70% of the total variation of our model and is almost close to the average eigenvalue.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.25972210	0.59961385	0.2037	0.2037
2	2.66010824	0.90990185	0.1663	0.3700
3	1.75020639	0.45088893	0.1094	0.4794
4	1.29931746	0.12799511	0.0812	0.5606
5	1.17132235	0.19639379	0.0732	0.6338
6	0.97492856	0.11018693	0.0609	0.6947
7	0.86474163	0.09429428	0.0540	0.7488
8	0.77044735	0.09298178	0.0482	0.7969

Next procedure was to find if there is any relationship among the principal components we kept and among each variable. The main source to answer this question was the Eigenvectors table we obtained as one of the results from *proc princomp*. I separated my analysis into two parts: analysis of variables across all PCs in general and across individual PC. I will go through both parts one by one. Out of the complete Eigenvectors table, I only focused on *Prin1-Prin6* which contain information of the six principal

components that we want to know.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
hoursw	0.020728	0.193249	0.545652	-.270094	0.050443	0.080726	-.071490	-.031761
child6	0.183579	-.305351	-.079285	-.029618	-.191640	-.545472	0.260238	0.032237
child618	0.094956	-.305379	-.042855	-.262210	0.233281	0.537199	-.044187	-.006326
agew	-.274356	0.438510	-.174905	0.234993	0.028813	0.039297	0.022767	0.045663
educw	0.398223	0.173154	0.132736	0.172721	-.042087	-.060026	0.191277	0.084917
hearnw	0.136477	0.217850	0.425618	-.278233	0.069287	0.064719	0.029682	0.079525
hoursh	0.053340	-.122962	0.104851	0.434324	0.567370	0.162642	0.471779	-.030669
ageh	-.277355	0.421369	-.177119	0.241457	0.038089	0.078022	-.055957	0.069799
educh	0.401092	0.114400	-.043622	0.117699	0.074735	-.119839	0.243245	-.082399
wageh	0.292710	0.244081	-.368469	-.282601	0.152510	-.186964	-.235833	0.154783
income	0.305718	0.279265	-.148228	-.167946	0.449093	-.083300	-.080635	0.197345
educwm	0.344635	-.001133	0.127650	0.333425	-.261075	0.214299	-.303139	0.048375
educwf	0.350391	0.064527	0.045894	0.353421	-.261274	0.193718	-.239866	-.018297
unemprate	0.055591	0.108979	-.227257	-.227972	-.414794	0.399016	0.540710	0.447177
experience	-.121525	0.318702	0.364600	-.037483	-.152258	-.193126	0.255564	-.074949
city_num	0.158680	0.213287	-.244523	-.191401	-.115576	0.171074	0.182000	-.835779

On the table above, the blue boxes indicate relationships between wife's working hours and number of children under 6. As I stated earlier with Correlation Matrix, they have negative correlation where increase in *hoursw* leads to decrease of *child6* for most of the values. Next, the red boxes show how *income* is affected by earning of wives and husbands. There is an indication that they are related when comparing the trends of *income* to that of *hearnw* and w*ageh*. Between the earning of wife and husband, we can see that the income closely matches the pattern of husband's wage rather than wife's wage. Therefore, I was able to find out that income of families are more influenced by the wage of husbands, which matches our hypothesis made at the beginning of our analysis with simple statistics. In addition, I looked at the relationship between education attainment level of both wives and husbands and their wages. The same way as other variables, I looked for a trend and surely the education levels

matched up with wages, which allowed me to conclude that education attainment level positively affects the average wages wives and husbands earn.

Next, I looked at how the principal components are grouped and what kind of relationships I can assume from them. The Eigenvectors table I used for this analysis is the same table as what I used to find relationships between variables, but here *Prin* columns were what I was interested at. As I analyzed Association between each PC, I was surprised to find out that the components were grouped according to some characteristics.

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8
hoursw	0.020728	0.193249	0.545652	-.270094	0.050443	0.080726	-.071490	-.031761
child6	0.183579	-.305351	-.079285	-.029618	-.191640	-.545472	0.260238	0.032237
child618	0.094956	-.305379	-.042855	-.262210	0.233281	0.537199	-.044187	-.006326
agew	-.274356	0.438510	-.174905	0.234993	0.028813	0.039297	0.022767	0.045663
educw	0.398223	0.173154	0.132736	0.172721	-.042087	-.060026	0.191277	0.084917
hearnw	0.136477	0.217850	0.425618	-.278233	0.069287	0.064719	0.029682	0.079525
hoursh	0.053340	-.122962	0.104851	0.434324	0.567370	0.162642	0.471779	-.030669
ageh	-.277355	0.421369	-.177119	0.241457	0.038089	0.078022	-.055957	0.069799
educh	0.401092	0.114400	-.043622	0.117699	0.074735	-.119839	0.243245	-.082399
wageh	0.292710	0.244081	-.368469	-.282601	0.152510	-.186964	-.235833	0.154783
income	0.305718	0.279265	-.148228	-.167946	0.449093	-.083300	-.080635	0.197345
educwm	0.344635	-.001133	0.127650	0.333425	-.261075	0.214299	-.303139	0.048375
educwf	0.350391	0.064527	0.045894	0.353421	-.261274	0.193718	-.239866	-.018297
unemprate	0.055591	0.108979	-.227257	-.227972	-.414794	0.399016	0.540710	0.447177
experience	-.121525	0.318702	0.364600	-.037483	-.152258	-.193126	0.255564	-.074949
city_num	0.158680	0.213287	-.244523	-.191401	-.115576	0.171074	0.182000	-.835779

Prin1: Educational attainment for both Wife and Husband is positively large, and Age of Wife and Husband and job *experience* are negatively correlated. Here I can infer that *Prin1* represents young married couple with high education level and little job *experience* as they are at the entry level in their career.

Prin2: Age of wife and husband as well as job *experience* display positive eigenvectors whereas their children rate (both children under 6 and children between 6 and 18) is showing negative values. This shows strong opposite result from *Prin1*, allowing me to assume that this group represents old married couple who have worked for a long period of time in their careers, and their children are already adults over 18, which is why the children rate is showing large negative values.

Prin3: This column shows positively high wife's working hour and wage, but negative children rate, age of wife and husband, and *income*. I inferred that this group consists of young, married, working couples yet without children.

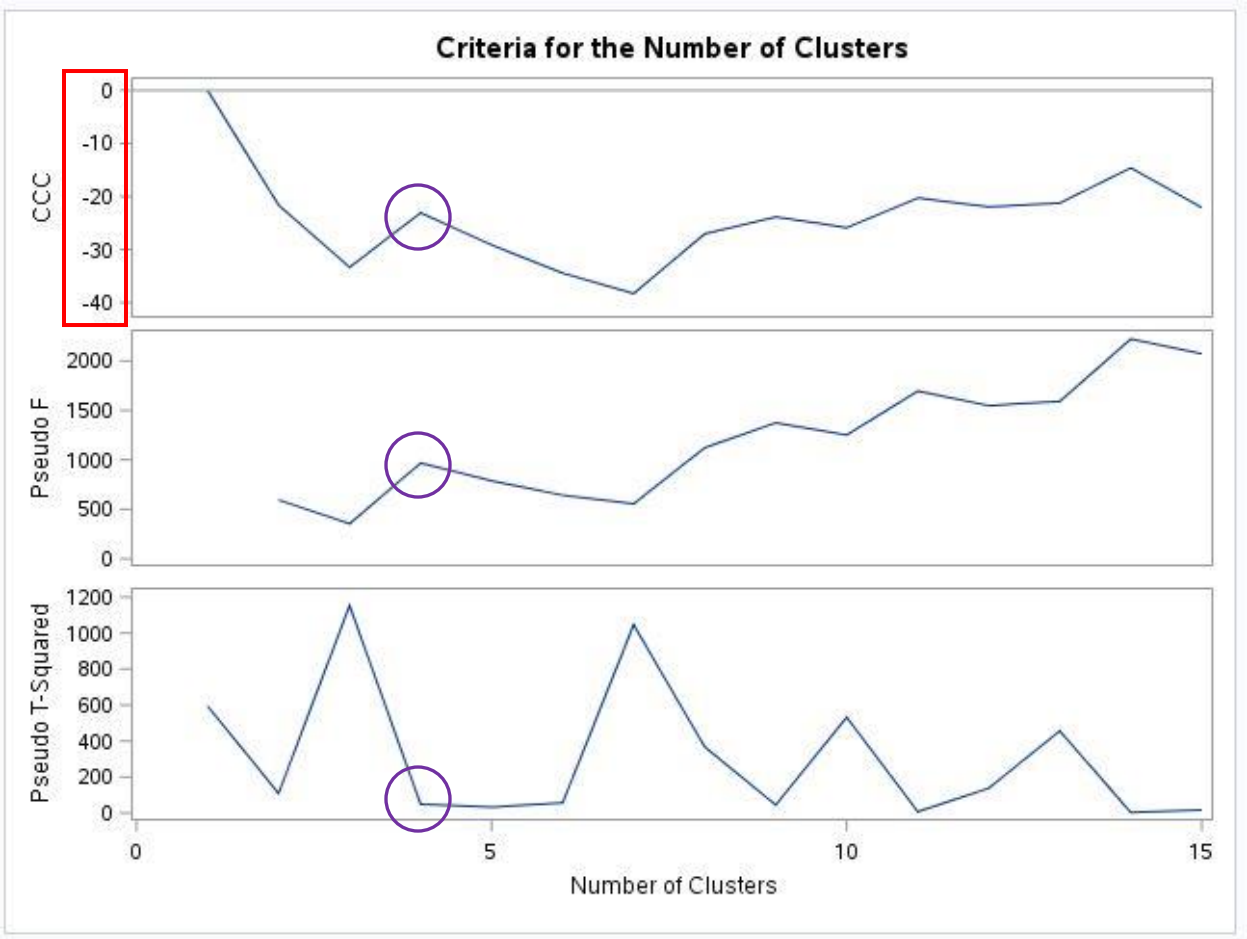
Prin4: Here we have positive age of wife and husband and high education level of wife's mother and father. Negative values include children rate and wife's working hours, which signals group with old couple and only husbands working. There may be some variations, but low working hours of wives as shown here may be an indication of housekeeping wives.

Prin5: Large positive values are husband's working hour and *income*, and most prominent negative values are wife's education level as well as education level of wife's mother and father. Here we can see that wife's family was in an environment where they received low level of education, and husband is wealthy and supporting his family, having large positive *income* value.

Prin6: Large positive *child618* and *unemprate*, and large negative education of wife and husband and *child6*. The associations here is less noticeable compared to other components. The unemployment rate in that region is low as well as the education level wives and husbands received. We can vaguely guess that this group represents rural areas or underdeveloped regions.

These analysis over each principal component gave us some underlying information about the features of our mroz dataset. The first few principal components are usually the most likely to be interpretable because they pick up the more prominent features in the data. In my dataset, I was able to find strong associations and grouping in each PC, but the interpretation became less visible and more vague as we

move on to the last few components. To step on further analysis of data, I tried to perform cluster analysis based on the PCs I kept in my PCA.



As the very first procedure, I generated ccc plot, pseudo F plot, and pseudo T-squared plot to determine number of clusters for my data. The graph may look fine initially with peak point at 4 clusters for ccc and pseudo F plot, and a drop for pseudo T-Squared plot at 4 clusters as well. However, there is a big error in the ccc plot, which is that all values are negative. Warren Sarle, a SAS analyst who created the ccc plot analysis, states that if all values of the ccc are negative and have decreasing point for two or more cluster, then the distribution is probably unimodal or long-tailed. Therefore, I could not proceed further with the cluster analysis as there may be highly irregularly distributed clusters and I need to consider using different algorithm, which goes out of the range materials learned in class.

Conclusion:

By performing Principal Component Analysis, we are bound to lose some interpretability because we are no longer looking at our original variables (variables that we know the meaning of). Instead, we have aggregate measures which may represent underlying features in the data that we can't concretely define (and we may not be able to even describe what those features are) in our initial original dataset. Through my analysis, I successfully reduced the dimensions, or the number of variables, removed highly correlated, multicollinearity issues, and found some underlying information about our dataset that wasn't visible without the grouping of principal components. With the basic understanding of my mroz dataset from simple statistics, I performed the rest of the PCA to find the optimal PCs to keep and investigated how the variables and PCs are associated. From the analysis over Eigenvectors table, I discovered that even though the groups are unknown in its nature, we obtained some useful information about our dataset.

In conclusion, datawise, I found out that wife's working hours have negative correlation with number of children under 6, education attainment level affects wages of both wife and husband, and *income* of families in 1975 depended more on husband's earnings than wife's earnings. On top of that, the 6 principal components I kept contained respective characteristics, where one of the PCs may represent young, highly educated couple that has negative correlation to number of children, and other PCs possibly representing old couple with different level of education, work *experience*, and/or wife's mother and father's education affecting wife's education and wage. This PCA analysis was personally very interesting for me because I did not expect to find such various hidden observations from our original dataset.

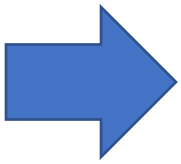
Appendix:

- Description of original Mroz dataset

Variables	Description
Work:	participation in 1975 ?
Hoursw:	wife's hours of work in 1975
Child6:	number of children less than 6 years old in household
Child618:	number of children between ages 6 and 18 in household
Agew:	wife's age
Educw:	wife's educational attainment, in years
Hearnw:	wife's average hourly earnings, in 1975 dollars
Wagew:	wife's wage reported at the time of the 1976 interview
Hoursh:	husband's hours worked in 1975
Ageh:	husband's age
Educh:	husband's educational attainment, in years
Wageh:	husband's wage, in 1975 dollars
Income:	family income, in 1975 dollars
Educwm:	wife's mother's educational attainment, in years
Educwf:	wife's father's educational attainment, in years
Unemprate:	unemployment rate in county of residence, in percentage points
City:	lives in large city (SMSA) ?
Experience:	actual years of wife's previous labor market experience

- PROC CONTENTS – I initially had warning errors when running SAS to perform PCA and later found out that PCA cannot have factor values in its analysis, so I converted the factor-type variables to numerical.

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	VAR1	Char	4	$4.	$4.
11	ageh	Num	8	BEST12.	BEST32.
6	agew	Num	8	BEST12.	BEST32.
4	child6	Num	8	BEST12.	BEST32.
5	child618	Num	8	BEST12.	BEST32.
18	city	Char	5	$5.	$5.
12	educh	Num	8	BEST12.	BEST32.
7	educw	Num	8	BEST12.	BEST32.
16	educwf	Num	8	BEST12.	BEST32.
15	educwm	Num	8	BEST12.	BEST32.
19	experience	Num	8	BEST12.	BEST32.
8	hearnw	Num	8	BEST12.	BEST32.
10	hoursh	Num	8	BEST12.	BEST32.
3	hoursw	Num	8	BEST12.	BEST32.
14	income	Num	8	BEST12.	BEST32.
17	unemprate	Num	8	BEST12.	BEST32.
13	wageh	Num	8	BEST12.	BEST32.
9	wagew	Num	8	BEST12.	BEST32.
2	work	Char	4	$4.	$4.

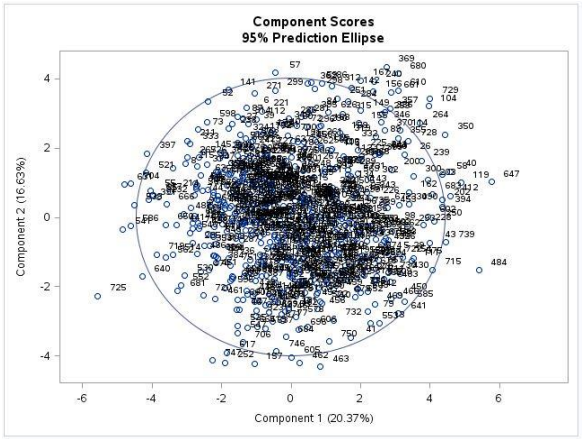


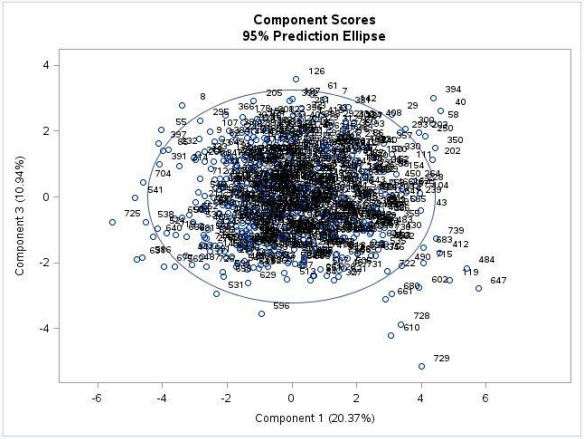
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
8	ageh	Num	8	BEST12.	BEST32.
4	agew	Num	8	BEST12.	BEST32.
2	child6	Num	8	BEST12.	BEST32.
3	child618	Num	8	BEST12.	BEST32.
16	city_num	Num	8		
9	educh	Num	8	BEST12.	BEST32.
5	educw	Num	8	BEST12.	BEST32.
13	educwf	Num	8	BEST12.	BEST32.
12	educwm	Num	8	BEST12.	BEST32.
15	experience	Num	8	BEST12.	BEST32.
6	hearnw	Num	8	BEST12.	BEST32.
7	hoursh	Num	8	BEST12.	BEST32.
1	hoursw	Num	8	BEST12.	BEST32.
11	income	Num	8	BEST12.	BEST32.
14	unemprate	Num	8	BEST12.	BEST32.
10	wageh	Num	8	BEST12.	BEST32.

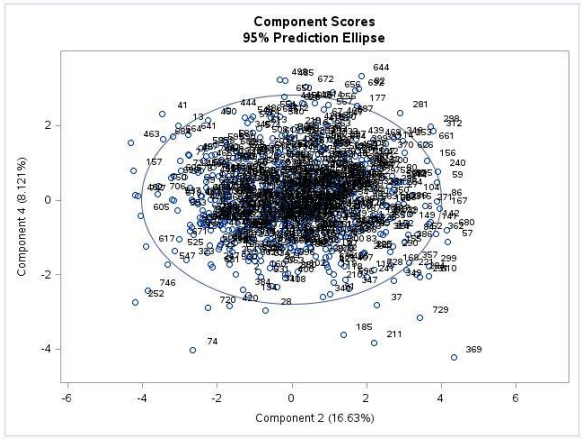
- Complete view of Eigenvectors table (In my actual report, I only displayed prin1 – prin6 portion out of the whole table)

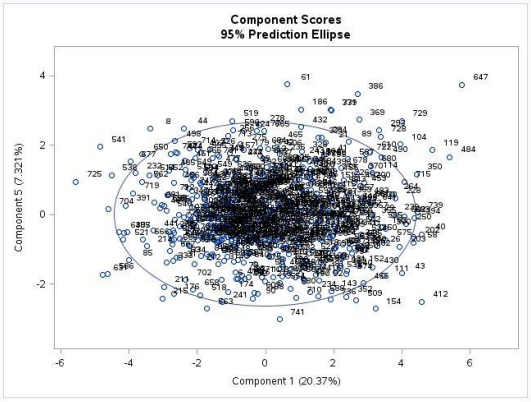
Eigenvectors																
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13	Prin14	Prin15	Prin16
hoursw	0.020728	0.193249	0.545652	-.270094	0.050443	0.080726	-.071490	-.031761	0.342880	0.206895	-.146626	0.108704	0.582966	0.028582	0.198670	0.050334
child6	0.183579	-.305351	-.079285	-.029618	-.191640	-.545472	0.260238	0.032237	0.137568	-.153296	0.522676	0.112917	0.339679	0.127186	0.053767	0.007308
child618	0.094956	-.305379	-.042855	-.262210	0.233281	0.537199	-.044187	-.006326	-.325917	0.290484	0.522902	-.013857	0.128462	0.001086	0.022723	0.029977
agew	-.274356	0.438510	-.174905	0.234993	0.028813	0.039297	0.022767	0.045663	-.108215	-.024766	0.221415	0.103279	0.188484	0.104847	-.037548	0.722290
educw	0.398223	0.173154	0.132736	0.172721	-.042087	-.060026	0.191277	0.084917	-.397512	-.013793	-.045996	0.179381	0.134785	-.708207	0.048858	0.014226
hearnw	0.136477	0.217850	0.425618	-.278233	0.069287	0.064719	0.029682	0.079525	-.272134	-.668936	0.149890	-.052484	-.202431	0.258863	0.064136	0.017350
hoursh	0.053340	-.122962	0.104851	0.434324	0.567370	0.162642	0.471779	-.030669	0.310556	-.097408	0.026398	-.012189	-.116877	0.023190	0.298917	0.014096
ageh	-.277355	0.421369	-.177119	0.241457	0.038089	0.078022	-.055957	0.069799	-.122484	-.077452	0.233938	0.078781	0.282404	0.081705	0.135259	-.673536
educh	0.401092	0.114400	-.043622	0.117699	0.074735	-.119839	0.243245	-.082399	-.405117	0.311959	-.301167	-.078892	0.150055	0.578186	-.087660	-.038917
wageh	0.292710	0.244081	-.368469	-.282601	0.152510	-.186964	-.235833	0.154783	0.084853	0.122869	0.021177	-.057953	-.192578	-.007961	0.657964	0.071787
income	0.305718	0.279265	-.148228	-.167946	0.449093	-.083300	-.080635	0.197345	0.313805	-.003600	0.122882	0.031653	-.002032	-.067838	-.631952	-.075770
educwm	0.344635	-.001133	0.127650	0.333425	-.261075	0.214299	-.303139	0.048375	0.177486	0.067625	0.134310	0.619129	-.242227	0.221628	0.010730	-.009914
educwf	0.350391	0.064527	0.045894	0.353421	-.261274	0.193718	-.239866	-.018297	0.186853	-.083824	0.159391	-.701634	0.142986	-.024012	-.006674	0.041060
unemprate	0.055591	0.108979	-.227257	-.227972	-.414794	0.399016	0.540710	0.447177	0.207606	-.024555	-.108636	0.007261	0.000440	0.042207	-.007381	-.020484
experience	-.121525	0.318702	0.364600	-.037483	-.152258	-.193126	0.255564	-.074949	0.045230	0.489662	0.374123	-.164101	-.448811	-.035782	-.034033	-.073012
city_num	0.158680	0.213287	-.244523	-.191401	-.115576	0.171074	0.182000	-.835779	0.143878	-.145422	0.037268	0.100546	-.005025	-.067163	-.015158	-.022710

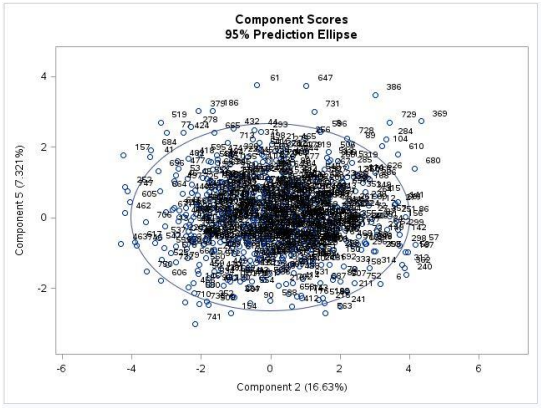
- Score plots for PCA of 6 PCs I chose – Following the order of procedures we used in class when performing PCA, I generated Score plot to see if there is any new discovery I can make from them. However, the score plots did not seem to have any distinct pattern/trends, outliers, or groupings visible in all of the score plots which compared mixed pairs of prin1 to prin6.

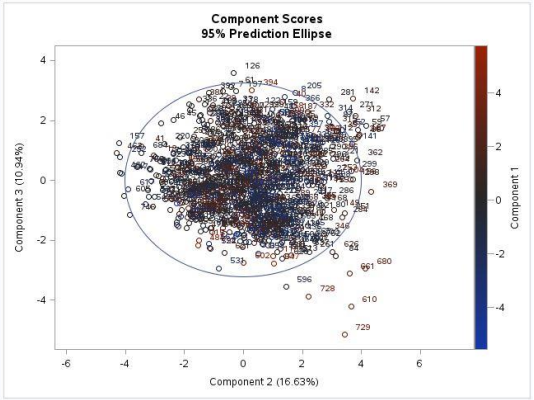




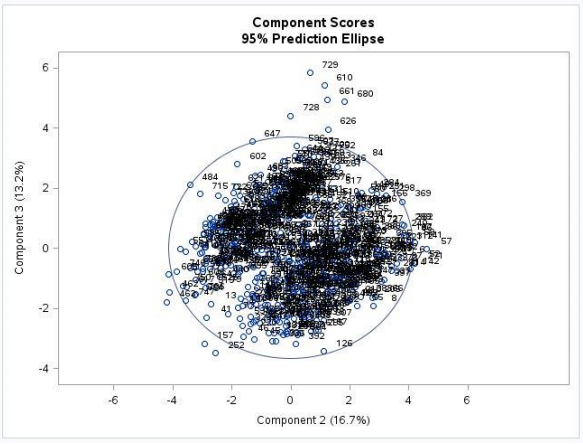


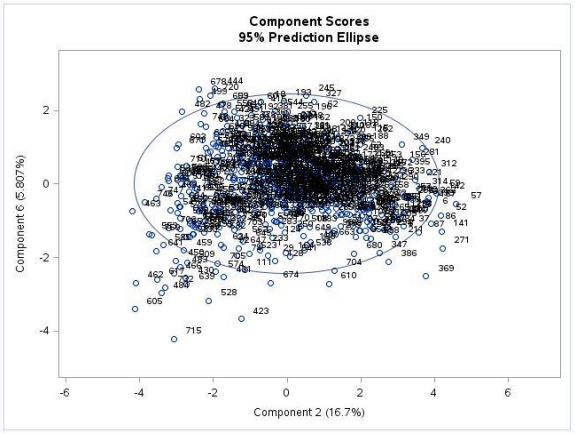


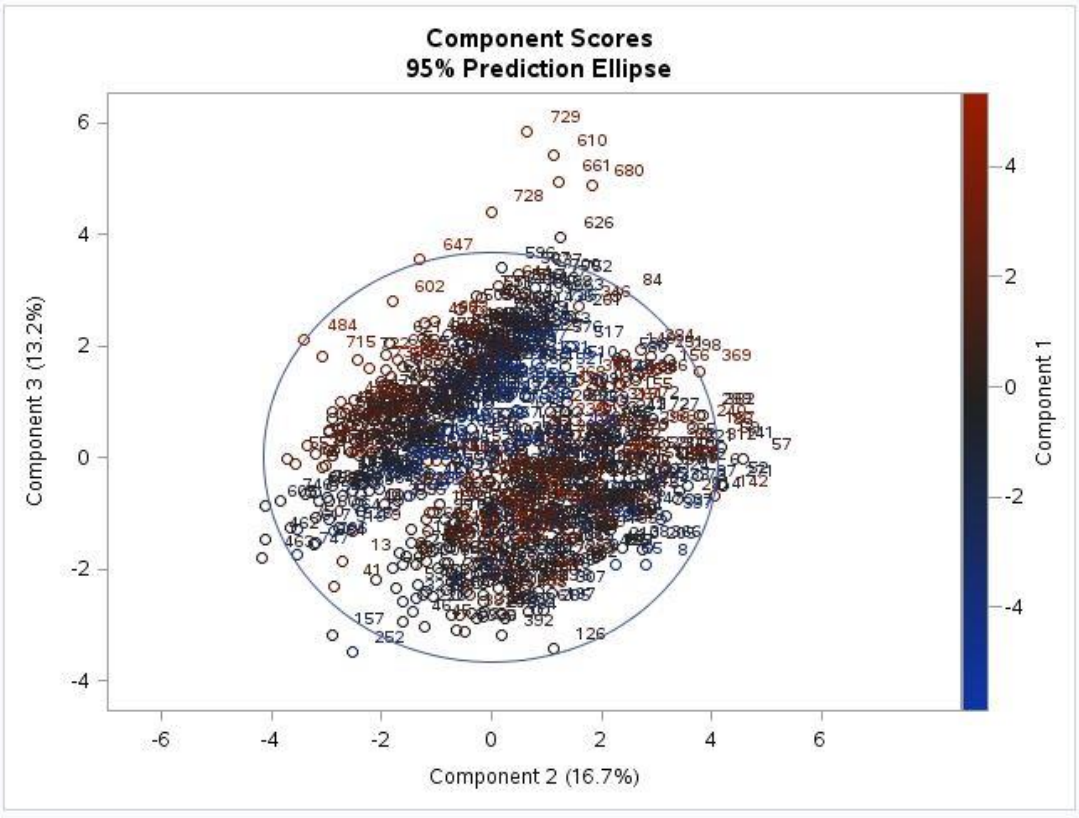




I tried to find the relationship between employment rate and unemployment rate as suggested in my presentation. To accommodate this method, I added work variable in my model (since our dataset does not contain a specific variable for employment variable) as ‘0’ for ‘yes’ and ‘1’ for ‘no’ to see if there is any change in the output of score plots.







There seems to be a little clearer distinction of two clusters in score plots of PC2 and PC3, PC2 and PC6, and PC1 and PC2 and PC3. I could see some points out of the circle boundary, symbolizing high PC1, PC2, and PC3 values. However, I could not gather any specific interpretation from the results partly because the score plots are comparing between two, or three, PCs at a time rather than directly comparing the original variables(work and unemprate). Therefore, the score plots reflect points of the combined values

of variables within each component, which may not display clear comparison between employment and unemployment rate.