Kellen Bullock

Dr. Comer

4-1-2020

Exercise 6

Part A

1. Pct\_Black, Pct\_Two\_Plus, Pct\_SNAP, Pct\_FIRE\_I, Pct\_Poverty, Pct\_Unemp, Med\_HomeValue, Pct\_White,'Pct\_BlueCollar\_O, Pct\_Hispanic are the ten assigned variables used from exercises 3 and 4. Overall in the regresisons run on these variables none of them performed especially well. Meaning the r-square was not above 0.600. There were also issues concerning multicolinearity with some of the variables as indicated by the variance inflation factor. In addition to that there was also high correaltions between percent poverty / percent SNAP at 0.757 and pct\_Unemp / pct SNAP at 0.722. These variables (percent poverty, percent SNAP and percent unemployed) represent the impoverished population of Oklahoma. These variables will likely be good candidates for “collapsing” into underlying factors.
   1. 4 factors are important. This is on the basis of the fall off seen in the percent of variance after factor 4. The drop off is from 8% to 5%. This is reflected in the “elbow” of the scree plot as well. The slope of the line in the sree plot becomes a lot flatter after the fourth factor. This is also more inline with the number of factors that can collapse into one another.
   2. In the communality table the variables that have the most communality with the factors are percent tow or more races, median home value, percent Hispanic, and median age. The variables that have the least are percent employees in blue collar occupations, percent living in poverty, and percent employees in professional industries. The threshold for the high communalities and the low communalities was roughly .800.
   3. In 1st component has the highest loadings in negative percent White at -.817, Percent Black at .467, Percent Two or more races .672, Percent of workforce unemployed .820, Percent living in poverty .665, negative percent employees in blue collar occupations at -.526, and Percent employees in professional industries .691. This f factors means small subset of people that are poor from white, larger of black, almost half of two or more races, the unemployed, living in poverty, the lower half of blue collar occupations, and roughly half of the employees in professional industries.   
        
      In component 2 there are the variables negative median home value -.843, percent SNAP .736, percent in poverty .491, percent employees in blue collar occupations .575, and percent employees in professional industries -.458. What the f factor means is more well off poor people, since their home value is much higher than component 1 and working blue collar jobs but the other half living in poverty and also working the lower half of professional industries.   
        
      Component 3 is comprised of the following variables: percent Hispanic .791, and percent black .510. This component can be interpreted as the large minority group of black and Hispanic races.   
        
      In component 4 there are the following variables: percent black .463, negative percent two or more races -.457, and percent SNAP at .421. These correlations are weak but the component can be interpreted as the minority group black and the people that do not identity as two or more races on SNAP.
   4. You specifically said “**Do not do** the maps **on part A**.” on 3-31-2020 and on 4-2-2020.
2. The communality table was not very helpful because the initial values was the same as the extracted value for all variables. If I was supposing to do only 5 factors it was not entirely clear. The total variance is best explained by 3 of the components. The rational behind this is the fall off of the percent of the variance goes from 10% to additional 4%. From the scree plot there is a good deal of fall of from the slope of the percent of variance.   
     
   As for the component in the matrix, the first 3 components are the focus of this factor analysis, because of their 32.101 explanation of the variance. The first component is compromised the variables: negative percent white -.817, percent black .467, percent two or more races .672, percent of workforce unemployed .820, percent living in poverty .665, percent employees in blue collar occupations -.526, and percent employees in professional industries .691. The interpretation of this component is as follows: The impoverished that are black and another race that also are unemployed or working in the professional industry rather than the blue collar occupation.   
     
   The second component is compromised of negative median home value -.843, percent SNAP .736, percent employees in blue collar occupations .575, and negative percent employees in professional industries -.458. This component describes the well off blue collar workers that do not need SNAP and have a low median home value.  
     
   In the third component the follow variables are included: percent black .510, and percent Hispanic .791. This component is nonwhite minorities of either black or Hispanic.   
   1. The factor based model performed worse than the regular model from exercise 4. The basis for this is in the r-square values which were .552 for the factor based model and .610 for the regular variable model. The standard error was also lower in the regular model used in exercise 4 at 10.22. Compared to the factor based model that had a a standard error of 10.35.
   2. Only two of the factors were significant in the factor regression model. Both of which were at .000 significances. The other factors were not included in the model. The r-square was weaker in this factor model at .552 as opposed the .610 that was produced in the model done in exercise 4.
   3. This model does not meet all of the assumptions of regression. The residuals are not normally distributed as seen in the gap in the histogram and in the distribution in qq-plot. Another problem evident in this model was lack of explanation because the stepwise threw too many factors out and did not offer enough explanation.
   4. There are 5 significant factors that explain 71.433% of the variance. There is only a 4% increase after the fourth factor, which is not a large enough margin in my opinion. Further evidence for this is the “elbow” in the scree plot. What is meant by the “elbow” is the severe change in slope of the scree plot.
   5. The assessment from the communality table, from all 25 variables iterated 4 times shows the highest extraction from percent Hispanic .918 PCI .904, percent not completing high school .839, percent other race .836, and percent white .830. All other variables were under 0.800. The lowest extracted variable was percent commuting to work at .376. The total variance explained by these factors is 71.1443 percent. And observing the scree plot there is definitely a large change in the slope of the plot of explained variance, creating an “elbow”.   
        
      In the first component there is aa very large representation of percent white, percent native American, percent of population not in the workforce, percent of workforce unemployed, percent receiving SNAP assistance, PCI, percent living in poverty, percent registered republican (negative), and percent voting for Fallin (negative). What this component seems to be explaining is the poor unemployed democrats seeking welfare.   
        
      The second component has high explanation of percent Asian, median age, percent employees in blue collar occupations, percent employees in professional industries, percent housing units vacant, and median home value. What this component is likely expressing is a minority of the population that is better off than the poorest but not by a large margin. This is based on the percent vacant housing units, Asian race, blue collar, and FIRE\_I.   
        
      The third component is compromised of percent other race, percent Hispanic and percent not completing high school. This component describes the uneducated minorities in the state.  
        
      For the fourth component there is a very light composition of population density, percent black, (negative percent commuting to work alone, and percent households vacant. This factor seems to be just the remainder of what is left of this group of variables. None of the correlations are above .450.   
        
      Finally, the fifth component is comprised of percent employees in service occupations, and percent pacific islander. It would appear that the fifth component explains the minority of people that identify as pacific islander in service occupations.

Part B

1. Percent poverty, percent SNAP, percent unemployed, percent two plus, and percent black will be the variables to most likely to cluster. This is based on the regional boxplots indicating that there are more of percent SNAP, percent poverty, and percent unemployment and percent black in the areas. Percent two plus is a bit of an outlier for the clustering by region but all of the variables had very skewed histograms from exercises one and three at different scales. This would very much indicate that clustering is occurring at those regions of counties, tracts, and school districts. As for the maps done on several of these variables there does not appear to be intense clustering at the county level but since the cluster analysis test will be run a more quantitative evaluation should prove that they are indeed clustered.
2. 1. The variables I am using will be percent poverty, percent SNAP, percent unemployed, percent two plus, and percent black. These are “important” and will most likely cluster based on the hunches detailed above. None of these variables are transformed. I believe 3 clusters is optimal because the clusters are relativity the same sizes. 2 are almost the same side but 1 of them is slightly bigger. However, at the lower levels of the dendogram there are too many clusters to really indicate clustering.
   2. 
      1. The g-clusters are contiguous in the third category. There are two main clusters of the third category in the southwest and north east parts of the state. For the second category the continuity is there but to a much less extent. There is a large cluster to the west, and smaller ones throughout the middle of the state. As for the first category there is the one big contiguous group in the southeastern part of the state, but there are many small clusters scattered elsewhere.
      2. The first cluster is majority percent poverty, percent unemployed, and percent SNAP. This makes sense because the southeastern part of Oklahoma and OKC are two to the highest percent poverty rates in the state. The second category seems to be a blend of percent black, percent two or more races and percent SNAP. This is logical because minority races are clustered in the western part of the state and suffer from low income. The third category seems to be a blend of percent unemployed, percent SNAP, and percent Black. Based on the factor analysis and the locations on the map there does seem to be lower rates of employment in the southwestern region of Oklahoma. Tulsa county does seem like an outlier in this circumstance.
   3. 
      1. The first cluster is not contiguous at all but the second and third are. The second cluster permeates through the interior of the state connecting the western to the eastern boarder. The third cluster covers the largest area of the state and is almost entirely through the southwestern and north western regions.
      2. According to the ANOVA table, all variables are contributing significantly to the clustering. The first cluster is majority percent poverty, percent unemployed, and percent SNAP. This makes sense because the southeastern part of Oklahoma and OKC are two to the highest percent poverty rates in the state. The second category seems to be a blend of percent black, percent two or more races and percent SNAP. This is logical because minority races are clustered in the western part of the state and suffer from low income. The third category seems to be a blend of percent unemployed, percent SNAP, and percent Black. Based on the factor analysis and the locations on the map there does seem to be lower rates of employment in the southwestern region of Oklahoma. Tulsa county does seem like an outlier in this circumstance.

* 1. There is a much higher level of clustering in the k-means as opposed to the the Hierarchical cluster analysis. The third clustering in the k-means almost completely covers the entire state where as the Hierarchical cluster analysis third category only covers the southwestern and north western predominately. What is evident from both maps is the very similar factor analysis groups and how they cover almost the same region on each map.