Group1: COVID-1

Qingyuan Guo
Xianru Liu
Di Qi
Dawei Wang

Group Updated and Proposal

The first two line analysis is factor analysis and Lasso regression. In this report, factor analysis will be our main explanation. We ran the Bartlett test and KMO that shows the p-value is less than zero which means that the data pass the null hypothesis and MSA are adequate which is 87%. The individual MSA for each variable is almost over 80% except the uninsured children, below 18 years old, over 65 age, and the birth rate per 1k is from 65% to 72%.
In PCA analysis, we extract 4 factors because PC4 has captured variance of 70%. We are looking for the factor analysis, not the dimensional deduction; so we have decided to choose to extract 4 factors. After we use 4 factors, we get 4 factors and the interesting thing is in RC1 and RC4 that has a negative value which is civilian labor. This was interesting since, at RC1 and 4, most variables are related to lifestyle and the social environment. These variables could be interpreted with heart disease. But civilian labor is the factor that is neither related to lifestyle nor the social environment. Hence, this is the only variable we discovered and couldn't group it.

Area_Rucc describes the area classification scheme that distinguishes the metro and non-metro area by the degree of urbanization and then divides them by population into more subtle classes.

econ_economictypology describes the typology classification of each area that is distinguished by *County Typology Codes.* Six different classes represent the major industry of this area.

Area_UrbanInfluence distinguishes metropolitan counties by population size of their metro area and nonmetropolitan counties, then distinguishes more classes by the size of the largest city or town and proximity to metro and micropolitan areas.

We would like to use all independent variables, plus heart disease mortality as parameters to predict the classification, the economic typology, and the urban influence level of each area. By applying linear discriminant analysis, we hope to use the economic, social, and health data to see if different areas have distinguished features.

After a delicate observation, I found the classification could be re-arranged by the following list and the data in our set was changed to the new standard accordingly.
Metro 1, Metro 2, Metro 3, City A, City N, SmallCity A, SmallCity N, Rural A, Rural N
The difference between Metro 1 to 3 is population. For other city levels, A means adjacent to the metro, and N means not adjacent to the metro. The size of the city or area from largest to the smallest is Metro, City, Small City, and rural.

The following prediction on the training set of Area_Rucc shows only 49.17%

	Predicted (cv)								
Actual	City_A	City_N	Metro_1	Metro_2	Metro_3	Rural_A	Rural_N	SmallCity_A	SmallCity_N
City_A	0.3218	0.0805	0.0805	0.1839	0.1264	0.0000	0.0000	0.1609	0.0460
City_N	0.1698	0.3396	0.0377	0.1132	0.1887	0.0000	0.0000	0.0566	0.0943
Metro_1	0.0300	0.0037	0.6854	0.1985	0.0824	0.0000	0.0000	0.0000	0.0000
Metro_2	0.1250	0.0560	0.1121	0.4224	0.2069	0.0000	0.0000	0.0690	0.0086
Metro_3	0.2105	0.0263	0.0789	0.1974	0.3816	0.0000	0.0000	0.0921	0.0132
Rural_A	0.0000	0.0000	0.0000	0.0000	0.0000	0.6667	0.0000	0.0000	0.3333
Rural_N	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5000	0.5000	0.0000
SmallCity_A	0.1596	0.0106	0.0426	0.1383	0.0745	0.0213	0.0106	0.4681	0.0745
SmallCity_N	0.0667	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3333	0.4000

By observing The confusion matrix, the prediction results tend to be in the seem city class. If we combine the same city level, the accuracy would be increased. The result is listed below: accuracy now increase to 73.22%

	Predicted (cv)			
Actual	City	Metro	Rural	SmallCity
City	0.4505	0.3604	0.0000	0.1892
Metro	0.1314	0.8144	0.0016	0.0525
Rural	0.0000	0.0000	0.4000	0.6000
SmallCity	0.2193	0.1754	0.0088	0.5965

Giving up the inside subclass could help us get a better understanding of the features of different city levels and know the difference between different areas.

econ_economictypology describes the pillar industry of each area and by conducting all LDA on this variable, we get the prediction accuracy about 63.74% and this may give us some valid information with further analysis

Actual	Predicted (cv) Farm	Fed	Manu	Mining	Nonsp	Recr
Farm	0.3333	0.3333	0.1111	0.0000	0.2222	0.0000
Fed	0.0000	0.5857	0.0857	0.0286	0.2714	0.0286
Manu	0.0000	0.0000	0.4615	0.0000	0.5385	0.0000
Mining	0.0000	0.0833	0.0000	0.3750	0.5417	0.0000
Nonsp	0.0030	0.1134	0.1343	0.0388	0.6627	0.0478
Recr	0.0000	0.0862	0.0000	0.0172	0.2931	0.6034

Second, we also use all independent variables, plus heart disease mortality as parameters in the training dataset to predict the Area_UrbanInfluence level of each area.
The following prediction on the training set of Area_UrbanInfluence shows 60.2%.

Actual	Predicted (cv) M1	M2	M3	M4	M5	N1	N2	N3	N4	N5	N7
M1	0.743	0.224	0.013	0.013	0.008	0.000	0.000	0.000	0.0	0.000	0.000
M2	0.137	0.581	0.053	0.119	0.060	0.026	0.005	0.005	0.0	0.000	0.014
M3	0.086	0.257	0.457	0.086	0.029	0.029	0.000	0.000	0.0	0.000	0.057
M4	0.017	0.267	0.017	0.500	0.100	0.100	0.000	0.000	0.0	0.000	0.000
M5	0.017	0.200	0.100	0.100	0.467	0.033	0.000	0.033	0.0	0.033	0.017
N1	0.097	0.161	0.065	0.065	0.129	0.419	0.000	0.032	0.0	0.032	0.000
N2	0.125	0.125	0.000	0.000	0.000	0.000	0.375	0.250	0.0	0.000	0.125
N3	0.000	0.000	0.000	0.000	0.000	0.167	0.000	0.833	0.0	0.000	0.000
N4	0.000	0.000	0.000	0.000	0.000	0.400	0.000	0.000	0.6	0.000	0.000
N5	0.000	0.167	0.000	0.333	0.000	0.000	0.000	0.000	0.0	0.500	0.000
N7	0.000	0.000	0.000	0.111	0.000	0.111	0.000	0.000	0.0	0.000	0.778

As we have the Annotation of this data, M and N just the shortcut to present the whole names of all variables. Based on this confusion matrix, we can see that the high accuracies are M1(74.3%)N3(83.3%） and N7（77.8%）, they are all above 70%. Those three represent *Large-in a metro area with at least 1 million residents or more*, *Noncore adjacent to micro area and contains a town of 2,500-19,999 residents* and *Noncore adjacent to a large metro area*. However there are also some low accurancies predictions such as M3,M5,N1,N2. Those four represent *Micropolitan adjacent to a large metro area*,*Micropolitan not adjacent to a metro area*,*Non core adjacent to a small metro with town of at least 2,500 residents and Non core adjacent to a small metro and does not contain a town of at least 2,500 residents.* Those are all under 50%. The common of the high accurancies is all of them with high populations or near large metro areas. For example, the inaccuracy percentage of M1 are assigned to M2-M5, so the common characters of them are near metro areas and with medium or even high percentage of residents. so it is reasonable to assign like this. The common of the high accurancies is Non core adjacent and with less residents. Another example is the lowest accuracy which is N2(37.5%). The inaccuracy percentage of N2 is assigned to M1,M2,N3 and N7. So the common characters of them are

small metro area and Non core adjacent, but except M1, it's a little bit suppressed to assign the inaccuracy to M1, but it's just a small portion, so overall, the accuracy of the confusion matrix is reasonable.

Through the analysis for the three category variables, Area_Rucc, econ_economictypology and Area_UrbanInfluence. The accuracy for each will be 73.22%, 63.74% and 60.2%. The percentage is acceptable, so it is appropriate to apply Linear discriminant analysis as our third line of analysis.