DSC324/424
Assignment #2 (DUE SUNDAY, July 31st, 2022 by Midnight)

Deliverables: Turn in your answers in a single PDF file. Use KnitR or Copy any R output relevant to your answer into your Word document and explain your answer thoroughly and include a copy of the full analysis in your report along with your conclusions. Also, provide your R code files.

Problem 1 (10 points) Answer each of the following questions:

a) What are regularized regressions? What are the differences between ridge and lasso regressions?

Regularized regressions models are penalized to avoid overfitting. Ridge regression has a penalty that is equal to the sum of the squares of the coefficients. Lasso regression has a penalty equal to the sum of the absolute values of the coefficients.

b) What are some causes of overfitting? How do we diagnose and treat overfitting in regression models?

Overfitting can be caused by many things: (1) too many features, (2) too few data points, and (3) highly correlated features. To diagnose overfitting you look at training error and test error; the training error is much lower than the test error. There are a few ways to treat overfitting: (1) simplify the model, (2) increase the amount of data, and (3) reduce the number of features.

c) What is multicollinearity? How do we diagnose and treat multicollinearity in regression models?

Multicollinearity - when multiple features or variables are highly correlated. To diagnose multicollinearity, look at the correlation matrix and the variance inflation factor (VIF). To treat multicollinearity, remove one of the correlated features or use regularization.

Problem 2 (10 Points): Have 1 Group member post the answers to the below questions **to the final project forum under the discussion section of D2L**:

DONE

- Project Team: Group Members
- Data:
  - Subject Area or Field of Interest
  - Source of Data (provide link to data)
  - Specific dataset(s)
  - description of its scope (# metric variables, #categorical variables, #samples, multiple related tables?)
  - Technology group plans to use for Project (i.e. Python, R, SPSS, Tableau, etc.) ○ How do you plan to use the technology?

  - In addition, as you are forming your groups, remember the following requirements for

datasets and groups:

a. Your group should have 4-5 people in it.

b. Please to make sure to have 1 liaison person for the group, who can submit assignments and ask me questions on behalf of the group.

c. Your dataset should be a real and rich dataset with at least 15 to 20 variables metric (continuous). It should have at least (10 * #var, but better yet 20 * #var) samples (we will see that some techniques like PCA require this for significance/stability). You will need a large sample size if you have a large n umber of variables. See me if you have any doubts about your dataset.

Problem 3 (Paper review) (10 Points) An academic paper from a conference or Journal will be posted to the Homework 2 content section of D2L. Review the paper and evaluate their usage of Factor Analysis. In particular address the following: **(See article on Psychological, social and economic impact of COVID 19 on the working population of India: Exploratory factor analysis approach)**

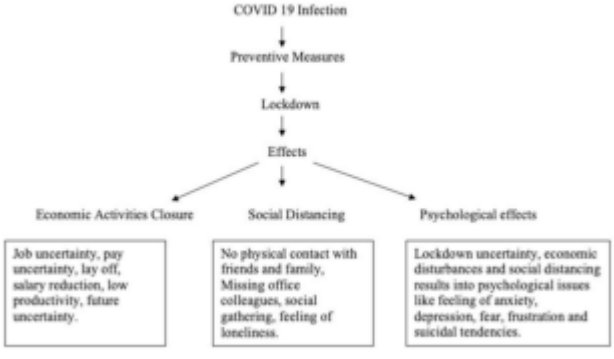
• How are they applying Factoring Analysis?

They apply exploratory factoring analysis because there are no standard tools to measure psycho-social, economic, and work related issues caused due to the COVID-19 pandemic (4, 4.1 Preliminary analysis).

• What kind of factor rotation do they use?

They are using the oblimin rotation method. On page 4 in section 4.1. Preliminary analysis states " Principal component analysis (PCA) and Oblique rotation (Direct Oblimin) are used, as the extraction and factor rotation method, as the items were not independent". Oblique rotation is the same as oblimin rotation.

• How many factors do they concentrate on in their analysis? How did they arrive at these number of factors?



There are three factors that are concentrated on in the analysis: (1) psychological effects, (2) social distancing and (3) economic activities closure. Social distance was derived

from Social identity theory. The psychological need theory or self-determination theory supports psychological effects. Lastly, economic activities closure was derived from the economic uncertainty principle.

• Explain the breakdown of the factors and the significance of their names.

Table 2
Values of communalities extraction and factor loading for psycho-social measures.

	Communalities	Component 1	Component 2
I'm worried about myself and my family members who may be reached byCOVID-19.	0.72	0.85	
I'm worried about myself and my colleagues who may be reached by COVID-19.	0.69	0.85	
I see the possibility that COVID-19 will break out in the area where I live and work.	0.56	0.75	
Uncertainty of the situation makes me feel worried about my family.	0.64	0.75	
Even after the lockdown period, I will avoid using public transport.	0.75		0.87
I will try to avoid social gatherings in the coming few months.	0.78		.088
I'm afraid of the uncertainty of the situation.	0.54		0.63
I freak out while going out for groceries.	0.58		0.71
I easily get irritated these days.	0.69		0.82

Table 3
Values of communalities extraction and factor loading for work-related Economic measures.

	Communalities	Component 1	Component 2	Component 3
I feel uncertain about the future of my job.	0.75	0.86		
I think my salary, bonus and other benefits will be reduced in the near future.	0.53	0.74		
My organization has assured the Employees regarding the regularity of salary.	0.61	0.48		
I feel that my organization will help its employees to attain financial stability.	0.62	0.57		
I'm doubtful about my future in this job.	0.77	0.88		
Work from home provides flexibility to some extent.	0.66		0.79	
Work from home helps in maintaining a balance between work and household chores.	0.78		0.88	
Work from home has decreased my productivity at work.	0.67		0.76	
I find it difficult to concentrate on work while working from home.	0.69		0.81	
Now I'm able to manage work and home easily.	0.54		0.48	
I think I will be able to continue working here.	0.64			0.71
Safety is given high priority by the management.	0.73			0.85
When COVID-19 broke out the company immediately established a pandemic prevention committee.	0.56			0.75
Management of my organization	0.65			0.61

Table 3 (*continued*)

	Communalities	Component 1	Component 2	Component 3
was neutral about the employees' safety at the time of the outbreak of COVID-19				
Work is given utmost priority in my organization even in the current situation.	0.56			0.37
Safety rules and procedures are strictly followed by the management.	0.65			0.80

Factors were further broken down into two factors: (1) psycho-social effects and (2) economic status.

• How do they evaluate the stability of the components (i.e. factorability)?

In the study they use the Bartlett's test and KMO measurements. “For psycho-social measures, the KMO value to check sampling adequacy was 0.713, and Bartlett’s test value of sphericity was found significant at 0.01 levels” (4). “For measuring economic status and work-related, the KMO value to check sampling adequacy was 0.735, and Bartlett’s test value of sphericity was significant at 0.01 levels” (4). However, there did not seem to be any reliability analysis.

• Do they use these factors in later analysis, such as regression? If so, what do they discover?

In the study target use a one-way ANOVA test to draw later analysis.They find the tier 1 cities differ significantly between tier 2 and tier 3 cities. That being tier 2 and 3 are more financially stable than tier 1.

• What overall conclusions does Factor Analysis allow them to draw?

The study finds that women are more socially vulnerable than men. Working individuals are among those who face the most severe repercussions. They also find that tier 1 cities differ significantly between tier 2 and tier 3 cities. That being tier 2 and 3 are more financially stable than tier 1. The overall conclusion they draw is that the Indian Government “has to take drastic measures, keeping in mind the overall behavioral, psycho-social, financial, and economic impact of the COVID19 pandemic” (6).

Problem 4 (Principal Component Analysis - 20 points): The data given in the file ‘Big5.csv’ are 5-point Likert items taken from the Big Five Personality Test web-based personality assessment. Techniques, such as Principal Component Analysis (PCA), can be used to determine different types of personalities. There are 19,719 subjects in the file and 50 variable items as follows:

E1 I am the life of the party.

E2 I don't talk a lot.

E3 I feel comfortable around people.

E4 I keep in the background.

E5 I start conversations.

E6 I have little to say.

E7 I talk to a lot of different people at parties.

E8 I don't like to draw attention to myself.

E9 I don't mind being the center of attention.

E10 I am quiet around strangers.

N1 I get stressed out easily.

N2 I am relaxed most of the time.

N3 I worry about things.

N4 I seldom feel blue.

N5 I am easily disturbed.

N6 I get upset easily.
N7 I change my mood a lot.

N8 I have frequent mood swings.

N9 I get irritated easily.

N10 I often feel blue.

A1 I feel little concern for others.

A2 I am interested in people.

A3 I insult people.

A4 I sympathize with others' feelings.

A5 I am not interested in other people's problems. A6

I have a soft heart.

A7 I am not really interested in others.

A8 I take time out for others.

A9 I feel others' emotions.

A10 I make people feel at ease.

C1 I am always prepared.

C2 I leave my belongings around.

C3 I pay attention to details.

C4 I make a mess of things.

C5 I get chores done right away.

C6 I often forget to put things back in their proper place. C7

I like order.

C8 I shirk my duties.

C9 I follow a schedule.

C10 I am exacting in my work.

O1 I have a rich vocabulary.

O2 I have difficulty understanding abstract ideas.

O3 I have a vivid imagination.

O4 I am not interested in abstract ideas.

O5 I have excellent ideas.

O6 I do not have a good imagination.
O7 I am quick to understand things.

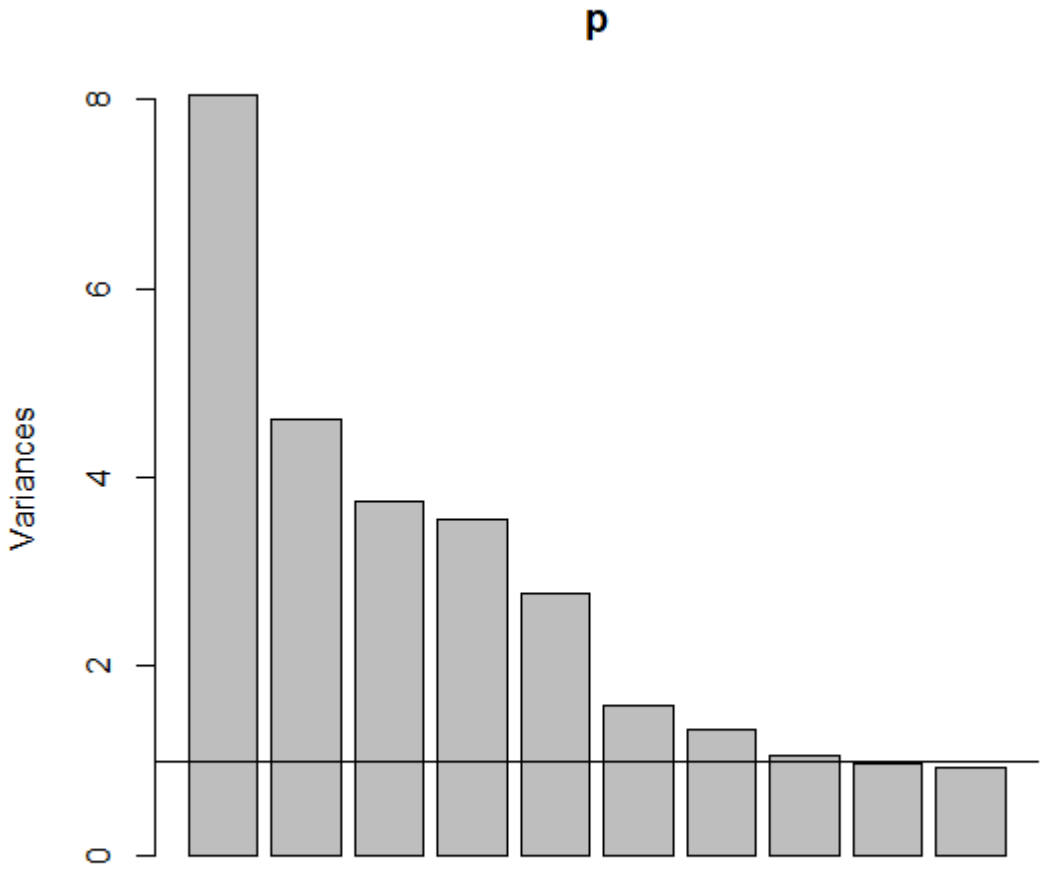
O8 I use difficult words.

O9 I spend time reflecting on things.

O10 I am full of ideas.

A) How many components are needed to explain 100% of total variation for this data? How many components are determined from the scree plot? What number of components would you use in the model?

There are 50 components that are needed to 100% describe total variation. The scree plot determines there are 10 components. The model should only see about 3 or 4 components in the model.



B) For the number of components in part A, give the formula for each component and a brief interpretation after rotating the components. What names might you give for each of the components?

Loadings:

	RC1	RC2	RC3
E1	0.645		
E2	-0.676		
E3	0.674		
E4	-0.660		
E5	0.737		
E6	-0.648		
E7	0.720		
E8	-0.531		
E9	0.602		
E10	-0.623		
A2	0.559		
A7	-0.550		
N1		0.687	
N3		0.659	
N5		0.545	
N6		0.720	
N7		0.651	
N8		0.674	
N9		0.596	
N10		0.623	
A4			0.523
C4		0.455	-0.501
C5			0.535
C8			-0.513
C9			0.550
N2		-0.487	
N4			
A1			
A3			-0.494
A5			
A6			
A8			0.453
A9			0.479
A10	0.481		
C1			0.494
C2			-0.426
C3			
C6			-0.488
C7			0.467
C10			0.429
O1			
O2			
O3			
O4			
O5			
O6			
O7			
O8			
O9			
O10			

	RC1	RC2	RC3
SS loadings	6.611	5.244	4.557
Proportion Var	0.132	0.105	0.091
Cumulative Var	0.132	0.237	0.328

C) What subjects have the highest and lowest values for each principal component (only include the number of components specified in part A. For each of those subjects, give the principal component scores (again only for the number of components specified in part A).

RC1 - Highest: 2.845205 | Lowest: -3.801185

RC2 - Highest: 3.05435 | Lowest:-3.77084

RC3 - Highest: 3.31042 | Lowest: -4.84888

D) Finally, run a common factor analysis on the same data. What difference, if any, do you find? Does the factor analysis change your ability to interpret the results practically?

Loadings:

	Factor1	Factor2	Factor3
E1	0.670		
E2	-0.686		
E3	0.665		
E4	-0.698		
E5	0.744		
E6	-0.605		
E7	0.745		
E8	-0.553		
E9	0.615		
E10	-0.655		
N1		0.633	
N3		0.542	
N5		0.546	
N6		0.723	
N7		0.736	
N8		0.764	
N9		0.673	
N10		0.624	
C4		0.509	
A4			0.767
A5			-0.608
A6			0.593
A7	-0.377		-0.555
A8			0.572
A9			0.704
N2		-0.450	
N4		-0.328	
A1			-0.408
A2	0.404		0.482
A3			-0.416
A10	0.375		0.382
C1			
C2			
C3			
C5			
C6		0.338	
C7			
C8		0.364	
C9			
C10			
O1			
O2			
O3			
O4			
O5			
O6			
O7			
O8			
O9			
O10			

	Factor1	Factor2	Factor3
SS loadings	5.462	5.077	3.765
Proportion Var	0.109	0.102	0.075
Cumulative Var	0.109	0.211	0.286

When running a factor analysis you find that most of the data is the same based on the rotation of the components in previous questions. There still needs to be variables removed like in the rotation. This will be taken care of in further iterations of the work.