Azat Dovgeldiyev

Homework 2

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 regression adds a penalty variable which helps to minimize the overfitting. So, it is a form of regression, that regularizes or shrinks the coefficient estimates towards zero.

Ridge regression modifies regression formula and is a way to create a model when the number of predictor variables in a set exceeds the number of observations.

Lasso regression also modifies regression formula by adding a penalty parameter, but with its optimization technique, Lasso regression performs variable selection at the same time. This regression is very useful when we have high levels of multicollinearity or when we want to automate certain parts of model selection.

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

When a model is too complex, or when data has noise such as outliers and errors in data. Problems occur when we try to estimate too many parameters from the sample. Overfitting can also occur if we have highly correlated variables. To prevent our model from overfitting is remove features manually, but when we have large amount of data, we can run regularized regression, which adds a penalty parameter.

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

When there are high correlations between two or more predictor variables, multicollinearity occurs. It causes inaccuracies in the computations of the slopes (betas). An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. Multicollinearity can be detected with variance inflation factor (VIF). If the VIF is greater than 5, then there is a sign of multicollinearity, if VIF is greater than 10, then the multicollinearity is problematic. We can remove some of the highly correlated independent variables, and perform an analysis designed for highly correlated variables, such as principal component analysis (PCA).

Problem 3 (Paper review)

- **How are they applying Factoring Analysis?**

Researchers applied Factoring Analysis to determine coping intentions if cyberbullied, and the relationship between coping styles and mental health with a sample of 229 adolescents. Researchers also aimed at in validating the use of Brief Cope in a cyberbullying context. More specifically, EFA was used to determine the sub-constructs that would emerge from the adapted version of the Brief COPE when measuring coping intentions of young people in response to cyberbullying.

- **What kind of factor rotation do they use?**

Given the many inter-item correlations, principle components extraction using oblique rotation (direct oblimin) was used, as this method of rotation is considered a more accurate method when examining the complex nature of unobserved variables.

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

A total of 235 participants completed the survey (aged 12-17). 6 of these participants were excluded due to incomplete responses, resulting in a total number of participants of 229 with mean age of 14.0+- 1.2 years, and 58.6% were female. Prior to conducting EFA, the 14 categories were examined by running a reliability analysis. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO =.0816. Any coefficients below .4 were excluded.

- **Explain the breakdown of the factors and the significance of their names.**

Table 3 Correlations between the 7 coping styles, depression, anxiety and stress, and cybervictimisation frequency and cyberbullying frequency (N = 229) (Mean scores (± standard deviation) in parentheses)

	Cybervictimisation frequency (25.16 ± 8.39)	Cyberbullying frequency (22.21 ± 3.17)	Depression (5.71 ± 5.83)	Anxiety (5.30 ± 5.48)	Stress (6.28 ± 5.51)
Active coping (25.68 ± 8.09)	−.144*	−.168*	−.142*	−.108	−.092
Emotion-focused coping (14.16 ± 5.72)	.161*	.026	.381**	.322**	.412**
Coping through humour (8.35 ± 3.31)	.029	−.017	.026	.018	.076
Coping through religion (3.88 ± 2.37)	−.045	.074	.020	.053	.092
Coping through denial (3.68 ± 1.85)	−.042	.043	.031	.028	.041
Coping through substance use (2.52 ± 1.35)	.252**	.048	.301**	.256**	.261**
Coping through distraction/reframing (12.14 ± 3.91)	−.044	−.056	−.036	−.033	−.002

* denotes p < .05, ** denotes <.01 significance level

The table suggests that the younger people cyberbully or are cybervictimised, the less likely they engage in active coping strategies. Furthermore, the correlations suggest that there may be a relationship between those with an intention to use emotion-focused coping and coping though substance use and poor mental health, and that those who intend to cope actively may have more positive scores of mental health.

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

Quantitative data were analyzed through SPSS and AMOS to determine descriptive and inferential statistical relationships, using the following techniques: frequencies; correlations; independent samples t-tests; analysis of variance; and reliability analysis.

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

Analysis of variance revealed some significant differences between levels of severity and coping intentions. Tests found significant differences between mean scores for emotion-focused coping and those classified as extremely severe for levels of depression, anxiety, and stress, suggesting that higher severity levels of stress, anxiety and depression may be related to engagement in emotion focused coping.

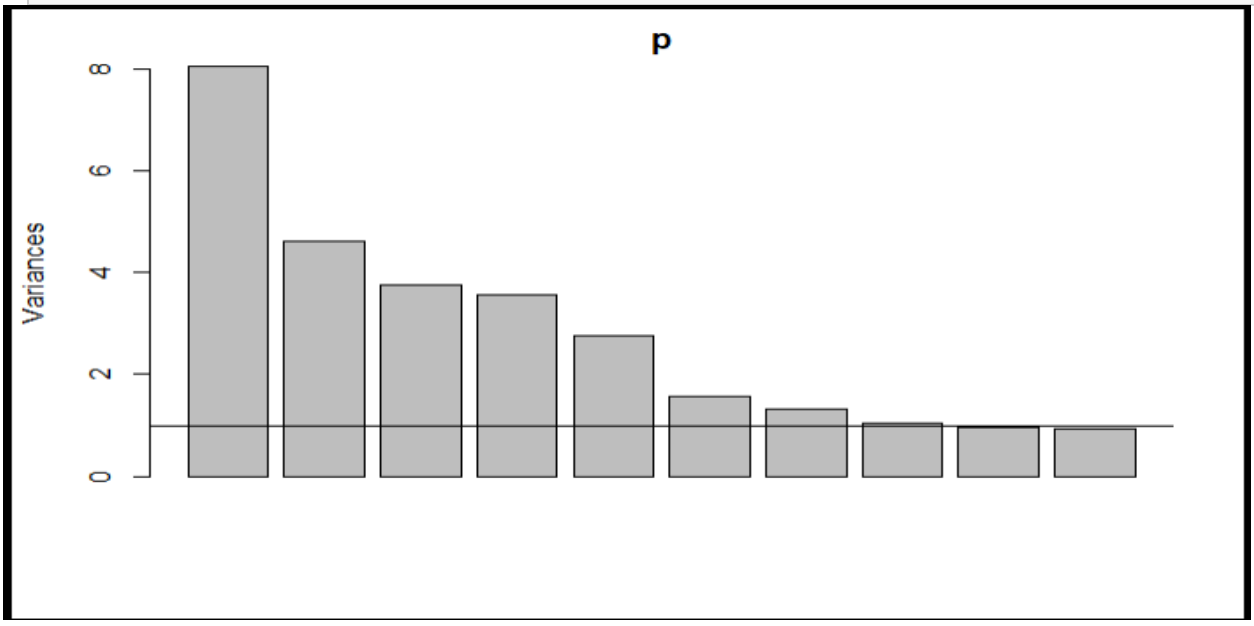
• What overall conclusions does Factor Analysis allow them to draw?

Results regarding cyberbullying involvement and coping intentions revealed that whilst young people intended to cope actively, the more young people were cybervictimised or cyberbullied others the less likely they were to do so. Young people who were more likely to use unproductive coping strategies were also more likely to have higher scores of depression, anxiety, and stress.

Problem 4 (Principal Component Analysis - 20 points):

A) How many components are need 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?

```
B) p <- prcomp( big_five, center = T, scale=T)
C) plot(p)
D) abline(1,0)
```



```
Parallel analysis suggests that the number of factors =  10  and the number o
f components =  7
```

```
table(p2$values>1)
```

```
##
## FALSE  TRUE
##    42     8
```

Parallel analysis suggests 7 components, while eigenvalues suggest 8, and when we look at the scree plot we could select 8 components.

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

```
# Contributions of variables to PC1
fviz_contrib(p4, choice = "var", axes = 1, top = 10)
```

```
## Loadings:
##     RC1    RC2    RC3    RC5    RC4    RC7    RC8    RC6
## E1   0.695
## E2  -0.727
## E3   0.658
## E4  -0.754
## E5   0.741
## E6  -0.627
## E7   0.747
## E8  -0.631
## E9   0.660
## E10 -0.700
## N1           0.723
## N2          -0.589
## N3           0.651
## N5           0.575
## N6           0.762
## N7           0.739
## N8           0.768
## N9           0.730
```

```
## N10         0.661
## A2                  0.568
## A4                  0.812
## A5                 -0.667
## A6                  0.675
## A7                 -0.617
## A8                  0.653
## A9                  0.754
## C1                         0.644
## C2                        -0.602
## C4                        -0.599
## C5                         0.670
## C6                        -0.651
## C7                         0.605
## C8                        -0.531
## C9                         0.676
## C10                        0.520
```

From the table above we can interpret that all variable names in RC1 are associated with positive attitudes, I would name it positive-attitude.

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). The highest value is 0.692 and the lowest value is -0.754.

```
scores <- p2$scores
scores_1 <- scores[,1]


min_score <- min(scores_1)
min_score
```

```
## [1] -2.858513
```

```
max_score <- max(scores_1)
max_score
```

```
## [1] 2.895168
```

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 Factor4 Factor5 Factor6 Factor7 Factor8
## E1   0.655
## E2  -0.699
## E3   0.647
## E4  -0.715
## E5   0.730
## E6  -0.598
## E7   0.730
## E8  -0.552
## E9   0.593
## E10 -0.660
## N1           0.733
## N2          -0.569
## N3           0.648
## N5           0.531
## N6           0.730
## N7           0.633                                         0.515
## N8           0.667                                         0.600
## N9           0.692
## N10          0.573
## A2                   0.534
## A4                   0.791
## A5                  -0.651
## A6                   0.596
## A7                  -0.608
## A8                   0.583
## A9                   0.712
```

There does not seem to be significant changes in factor analysis, although coefficients are slightly changed.