Assignment 2

Due: 6/24/2018

Student name: Lavinia Wang #1473704

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

a) What are the advantages and disadvantages of using ridge regression and lasso regression? How are these regressions different?

Type	Advantage	Disadvantage
Ridge regression	Ridge regression allows you to analyze data even when severe multicollinearity is present and helps prevent overfitting. This type of model reduces the large, problematic variance that multicollinearity causes by introducing a slight bias in the estimates.	Ridge regression is not able to shrink coefficients to exactly zero. As a result, it cannot perform variable selection.
LASSO regression	LASSO performs variables selection. This penalty allows coefficients to shrink towards exactly zero. LASSO usually results into sparse models, that are easier to interpret.	Using cross validation to find the optimal regularization coefficient, lambda can be just as expensive as stepwise selection techniques.

Difference: Ridge regression can't zero out coefficients; thus, you either end up including all the coefficients in the model, or none of them. In contrast, the LASSO does both parameter shrinkage and variable selection automatically.

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

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. Too few data points for the number of parameters could cause overfitting.

To avoid overfitting your model in the first place, collect a sample that is large enough so you can safely include all of the predictors, interaction effects, and polynomial terms that your response variable requires. Cross-validation can detect overfit models by determining how well your model generalizes to other data sets by partitioning your data. This process helps you assess how well the model fits new observations that weren't used in the model estimation process.

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

Multicollinearity is a phenomenon in which one predictor variable in a multiple regression model is correlated with other predictors.

One way to measure multicollinearity is the variance inflation factor (VIF), which assesses how much the variance of an estimated regression coefficient increases if your predictors are

correlated. If no factors are correlated, the VIFs will all be 1. Another way is to check the correlation matrix in predictor variables. If the correlation between two variables are strong, it is more appropriate to include just one of the correlated variables. If multicollinearity is a problem in your model, try removing highly correlated predictors from the model or using PCA that cut the number of predictors to a smaller set of uncorrelated components.

Problem 2 (Paper review 2) (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.

- How are they applying Factoring Analysis?

Exploratory Factor Analysis is utilized to reduce the security variables to a smaller set of summary variables (security factors) and to explore the underlining hypothetical structure. It is applied to distinguish the structure of relationship between the variables and respondent. EFA was also conducted to understand the measurement and significance of the variables from the survey.

- What kind of factor rotation do they use?

In this analysis, they tried both orthogonal (varimax) and oblique(oblimin). In the first rotation (varimax), the pattern matrix explains the factor loadings after the rotation. The second rotation (oblique) was performed to obtain the component correlation matrix that shows the relationship between factors extracted in general.

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

9 factors. In order to determine how many numbers of factors (or components) are extracted, eigenvalues (or Kaiser criterion) and scree plot are two sets of information that can be referred. The first method, the Kaiser's criterion or eigenvalues will extract and maintain the factors that obtain the value of eigenvalues more than 1 to be included in next investigations. Using the scree plot, the point at which there is a drastic change of direction and becomes horizontal recommends the number of factors.

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

Factor	Variables	Loadings	Interpretation
1	Cloud Security Policy Cloud Security Procedures Cloud Security Practices	0.861 0.930 0.902	Security Implementation for Cloud Storage Aspects
2	Identity Management Authentication Access Management Authorization Secure API Standards authenticating user accounts Secure Access Communication Channel	0.685 0.798 0.794 0.793 0.684 0.617	Confidentiality Aspects

Assignment 2
Due: 6/24/2018
Student name: Lavinia Wang #1473704

		0.789	
3	Encryption Key management Data ownership Data stewardship Data deletion Data protection for sensitive data	0.784 0.786 0.763 0.772 0.818 0.802	Integrity Aspects
4	Accessibility to data stored Backup of data stored Recovery of data Verify data authenticity	0.800 0.830 0.823 0.866	Availability Aspects
5	Bind and Validation Between the Identities Time Stamp Geographical Location as an Authentication Restrict the storage of user data to specific countries or geographic locations	0.866 0.860 0.465 0.841	Non-repudiation Aspects
6	Cryptographic Protection Origin authentication Verification assurances Anti-counterfeit/Anti-tampering Authenticity of session	0.821 0.765 0.636 0.587 0.714	Authenticity Aspects
7	Multi-Failure Disaster Recovery Capability Monitoring service continuity System Maintenance System Monitoring Malicious Code Protection Mechanism	0.741 0.830 0.799 0.728 0.822	Reliability Aspects
8	Conformance to External Responsibility Roles Mechanisms to put internal security policies in effect Transparency and participation Security Functionality Security Assurance	0.842 0.846 0.840 0.840 0.601	Accountability Aspects
9	Audit Policy Audit Record logs Automated audit logs Report Generation Original content or time of audit records	0.729 0.816 0.865 0.853 0.685	Auditability Aspects

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

The future work will test the structure identified in a constrained manner using confirmatory factor analysis (CFA). CFA will validate how much the 43 item indicators explains the nine constructs (security factors) through a CFA measurement model.

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

Through CFA measurement analysis, the author will be able to verify the factor structure of a set of indicators and this allows the author to define the relationship between a set of measured variables and a set of latent variables. Moreover, verification of construct validation

and construct reliability is completed through the measurement model. Results obtained from measurement model will further be used to test the structural model and perform path analysis.

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

EFA allow the authors to summarize data and define the structure. By implying EFA the identified structure has shown a structure of a model with 9 interpretable factors.

Problem 3 (10 points-Data Ethics or Data Integrity): Using Google Scholar, locate **a journal article**, which discusses data ethics or data integrity in terms of big data in your field of interest. **Write a summary** of the journal article and how it utilizes data ethics or data integrity in **two to three paragraphs**. **Cite the paper in APA format.**

In this paper, the author introduced the topic by telling a story of a girl who mistakenly posted birthday party invitation publicly which lead to the reconsideration of traditional ethical conceptions with the emergence of Big Data. The traditional ethical principles with regard to moral responsibility of the individual are agreed upon causality, knowledge and choice. In general, however, the ethics of Big Data is towards an impersonal ethics based on consequences for others as Big Data has a huge global representation, represents reality digitally more naturally and emphasizes correlation.

The nature of hyper-networked societies exacerbates the collateral damage caused by actions within the network of Big Data collectors, Big Data utilizers, and Big Data generators. Privacy, and propensity are studied by different groups for various purposes, which are some of the challenges Bid Data is facing. Big Data might induce certain changes to traditional assumptions of ethics regarding individuality, free will, and power and have consequences in many areas that is taken for granted. In conclusion, people at all ages need to be educated about the unintended consequences of their digital footprints. Ethicists will have to continue to discuss how we can and how we want to live in a world and how we can prevent the abuse of Big Data as a new found source of information and power.

Andrej Zwitter (November 20, 2014). *Big Data ethics.* Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/2053951714559253

Problem 4 (Principal Component Analysis - 20 points): The data given in the file 'bfi.csv' is the16 multiple choice ability items taken from the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. Techniques such as Principal Component Analysis (PCA) can be used to determine different types of personalities. There are 2,800 subjects in the file and 28 variable items as follows.

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?

Assignment 2

Due: 6/24/2018

Student name: Lavinia Wang #1473704

After removing the missing values, the original 2800 observations have 2436 left and 25 variables. In order to explain 100% of total variance, we need to include them all, i.e. 25 variables. From the scree plot, we can identify 6 values that are above 1. Thus we can conclude there are 6 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?

RC2 = 0.837*N1+0.835*N2+0.795*N3+0.617*N4+0.608*N5

RC3 = 0.653*C1+0.738*C2+0.678*C3-0.689*C4-0.625*C5

RC5 = -0.662*A1+0.749*A2+0.709*A3+0.547*A4+0.582*A5

RC1 = 0.417*N4+0.730*E1+0.729*E2-0.585*E4-0.515*E5-0.421*E3+0.431*O4

RC6 = 0.576*E3 +689*O1+0.662*O3+0.434*O4

RC4 = 0.662*O2+0.704*O5

RC2 represents those who have frequent mood swings, get angry/irritated easily, feel blue or panic easily, which will be named as “emotional”. RC3 represents those who are exciting in work, pursue perfection, follow plans, don’t do things in a half-way manner and value time, which will be named as “efficient”. RC5 represents those who care about other’s feelings and wellbeing, know how to comfort others, love children and make people feel at ease, which will be named as “empathy”. RC1 represents those who often feel blue, quiet, find it difficult to approach others, hard to make friends, passive, don’t know how to captivate people and spend time reflecting on things, which will be named as “diffident”. RC6 represents those who know to captivate people, are full of ideas, carry the conversation to a higher level and spend time reflecting on things, which will be named as “dominant”. RC4 represents those who avoid difficult reading material and will not probe deeply into a subject, which will be named as “dastard”.

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).

	Highest Value Subject	Principal Component Score	Lowest Value Subject	Principal Component Score
RC1	E1	0.730	E4	-0.585
RC2	N1	0.837	N5	0.608
RC3	C2	0.738	C4	-0.689
RC4	O5	0.704	O2	0.662
RC5	A2	0.749	A1	-0.662
RC6	O1	0.689	O4	0.434

Assignment 2
Due: 6/24/2018
Student name: Lavinia Wang #1473704

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:
   RC2    RC3    RC5    RC1    RC6    RC4
N1  0.837
N2  0.835
N3  0.795
N4  0.617                0.417
N5  0.608
C1         0.653
C2         0.738
C3         0.678
C4        -0.689
C5        -0.625
A1               -0.662
A2                0.749
A3                0.709
A4                0.547
A5                0.582
E1                       0.730
E2                       0.729
E4                      -0.585
E5                      -0.515
E3                      -0.421  0.576
O1                              0.689
O3                              0.662
O2                                     0.662
O5                                     0.704
O4                       0.431  0.434

                 RC2   RC3   RC5   RC1   RC6   RC4
SS loadings    3.094 2.594 2.570 2.547 2.088 1.610
Proportion Var 0.124 0.104 0.103 0.102 0.084 0.064
Cumulative Var 0.124 0.227 0.330 0.432 0.516 0.580
```

Figure 1: PCA loadings from RStudio

```
Loadings:
   Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
N1  0.819
N2  0.805
N3  0.710
N4  0.561
N5  0.512
E1         -0.583
E2         -0.674
E3          0.566
E4          0.674
E5          0.506
C1                  0.557
C2                  0.677
C3                  0.548
C4                 -0.638
C5                 -0.546
A1                         -0.530
A2                          0.663
A3                          0.601
O3                                  0.572
O5                                 -0.572
A4                          0.401
A5          0.435           0.461
O1                                  0.483
O2                                 -0.494
O4

                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6
SS loadings       2.722   2.652   2.075   1.672   1.512   0.614
Proportion Var    0.109   0.106   0.083   0.067   0.060   0.025
Cumulative Var    0.109   0.215   0.298   0.365   0.425   0.450
```

Figure 2: Factor analysis loadings from RStudio

Compared two outputs, three components remain the same which are the N group (factor 1), the C group (factor 3) and the A group (factor 4). Factor 2 has all E group plus A5, and represent those who talk a lot, easy to approach others, know how to captivate people, make friends easily, take charge and make people feel at ease, which will be named as “sophisticated”. Factor 5 includes all O group but O4, and represent those who are full of ideas, carry the conversation to a higher level, probe deeply into a subject and confront difficult reading material, which will be named as “pioneer”. In factor analysis, factor 6 does not exist. So there are only 5 factors in the output.

R code attached:

```
# Student name: Lavinia Wang
# ASsignment 2

# Problem 4:
library(psych)

##############################################################################
# Set up dataset
##############################################################################
# Read in the csv "bfi" dataset.
bfi = read.csv(file = "C:/Users/cher0_000/Desktop/CSC424/HW 2/bfi.csv", heade
r = TRUE)
str(bfi)

# Check frequencies and missing values for all variables or a specific variab
le
```

Assignment 2

Due: 6/24/2018

Student name: Lavinia Wang #1473704

```
describe(bfi)

# Pull out just the numeric fields and place price at the front
# Change Variable Names

bfi_sub = bfi[, c(2:26)]
plot(bfi_sub)

# Check data types
str(bfi_sub)

# list rows of data that have missing values
bfi_sub[!complete.cases(bfi_sub),]

# create new dataset without missing values
newbfi <- na.omit(bfi_sub)
newbfi
describe(newbfi)

# Compute the correlation matrix and visualize it
cor.bfi = cor(newbfi)
cor.bfi
corrplot(cor.bfi, method="ellipse")

# Look at the size of the numeric data
dim(newbfi)

# Run a correlation test to see how correlated the variables are.  Which corr
elations are significant
library(psych)
options("scipen"=100, "digits"=5)
round(cor(newbfi[,]), 2)
MCorrTest = corr.test(newbfi[,], adjust="none")
MCorrTest

M = MCorrTest$p
M

# Now, for each element, see if it is < .01 (or whatever significance) and se
t the entry to
# true = significant or false
MTest = ifelse(M < .01, T, F)
MTest

# Now lets see how many significant correlations there are for each variable.
  We can do
# this by summing the columns of the matrix
colSums(MTest) - 1  # We have to subtract 1 for the diagonal elements (self-c
orrelation)

# Compute covariance matrix
cov(newbfi[,])

# Initial PCA
```

Assignment 2

Due: 6/24/2018

Student name: Lavinia Wang #1473704

```
p = prcomp(newbfi, center=T, scale=T)
plot(p)
abline(1, 0)
summary(p)
print(p)
p$rotation
biplot(p)

rawLoadings = p$rotation %*% diag(p$sdev, nrow(p$rotation), nrow(p$rotation))
print(rawLoadings)
v = varimax(rawLoadings)
ls(v)
v

# Use psych package to run PCA
p2 = psych::principal(newbfi, rotate="varimax", nfactors=6, scores=TRUE)
p2
print(p2$loadings, cutoff=.4, sort=T)
p2$loadings
p2$values
p2$communality
p2$rot.mat

v$loadings

# RUN factor analysis to compare results
fit = factanal(newbfi, 6)
print(fit$loadings, cutoff=.4, sort=T)
summary(fit)
```