MULTIVARIATE DATA ANALYSIS

FINAL PROJECT

Mental Health of Workforce

Author:
Minzhao LIU , Xuan LUO

Dec. 15, 2010

Contents

Abstract

This article tried to figure out the relationship between mental health and personal demographic, family and working information. Dataset was brought by a national study of United States in 2008. Multivariate regression, canonical correlation analysis, discrimination analysis and cluster analysis were adopted and reasonable interpretations were made by results from four methods. Some defects are also mentioned in the last section of the article.

Key words: Multivariate regression analysis, canonical correlation analysis, classification and discrimination analysis, cluster analysis, mental health, workforce.

1 Introduction

1.1 Motivation and Background

This study intended to investigate the relationship between mental health of workforce and factors from their family and professional life. The World Health Organization says more than four hundred fifty million people suffer from poor mental health, most of which are depression and schizophrenia. Realizing the effect from family and work on the mental health may help psychologists and doctors solve people's mental health problems.

1.2 Data Description

To address this puzzle, the Families and Work Institute conducted a national study of the United States workforce forming the 2008 National Study of the Changing Workforce (NSCW) dataset. A total of 3502 (with 2769 wage and salaried) interviews were completed with a nationwide cross-section of employed adults between Nov. 2007 and Apr. 2008. Interviews, which averaged 50 minutes in length (47 minutes for substantive questions and 3 minute for eligibility screening), were conducted by telephone using a computer-assisted telephone interviewing (CATI) system. Calls were made to a regionally stratified unclustered random probability sample generated by random-digit-dial methods. Up to 60 calls were made to each telephone number that appeared to represent a potentially eligible household- busy signal obtained, voice mail pickup, or answered by a non-eligible with some indication of a potential eligible in household. When eligibility were identified and requested callbacks, additional calls were made. If twenty-five consecutive calls were made to numbers where there were no answers and busy-signals, these numbers were considered nonresidential, non-working numbers, or non-voice communication numbers. Three to five attempts were made to convert each initial refusal.

For each subject, there are 4 parts of information. Demographic information includes age, gender, and education level of the subject. Mental health information contains three indexes: sleep, stress, and depression. Family factor consists of number of kids under 18 at home, whole family

income in 2008, and subjects satisfaction with life. As for work life factor, workplace effectiveness, coworker support, satisfaction with job, and difference in work hours between practice and ideal are considered. Here is the brief introduction for these four categories information.

1. **Demographic Information:**

   - *Age:* Age in years when the subject was interviewed.
   - *Gender:* Female/Male is coded as (1/0).
   - *Education Level:* There are six levels of education, listed from low to high as '<High school diploma', 'High school or GED', 'Some college, no degree', 'Associate degree', '4-year college degree', 'Grad or prof degree'. The six levels of education is coded as from 1 to 6.

2. **Health Indicators:**

   - *Stress:* Stress is calculated from answer to five questions related to one's stress condition, including "qpw3: How often have you felt nervous and stressed?", "qpw4: How often have you felt that you were unable to control the important things in your life?", "qpw6: How often have you felt that difficulties were piling up so high that you could not overcome them?", "qpw7: How often have you felt that things were going your way?", and "qpw17: Not thinking about work, how stressful has your personal and family life been in recent months? -extremely stressful, very stressful, somewhat stressful, not very stressful, or not stressful at all?". As for the calculated value of stress, it runs from 1 to 5, with the smaller the value, the less stressful for the subject.
   - *Sleep:* Sleep is also calculated from answer to three questions related to one's sleep, including "qpw2: How often have you had trouble sleeping to the point that it affected your performance on and off the job?", "qpw7a: How often have you had trouble falling asleep when you go to bed?", and "qpw7b: How often have you awakened before you wanted to and had trouble falling back asleep?" As for the calculated value of sleep, it runs from 1 to 5, with the higher the value, the worse the sleep problem for the subject.
   - *Depression:* Depression is based on answers to two questions, which are "qpw8: During the past month, have you been bothered by feeling down, depressed, or hopeless?" and "qpw9: During the past month, have you been bothered by little interest or pleasure in doing things?". The value of Depression are: 2,1,0 respective to how many yes in your answer. So basically, value 2 means high level of depression, and value 0 means low level of depression, with value 1 being in the middle. Hence, the smaller the value, the less depressed is the subject.

3. **Family Factors:**

- *Number of kids under 18 at home.*
- *Whole family income in 2008:* This is a categorical variable, having family income "Q1: $<$ \$37,440" with value 1, "Q2: \$37,441-\$67,514" with value 2, "Q3: \$67,515-\$104,000" with value 3, and "Q4: $>$\$104,000" with value 4.
- *Satisfaction with life:* This is a categorical variable, having "Very dissatisfied", "Somewhat dissatisfied", "Somewhat satisfied", and "Very satisfied" assigned to value 1 to 4. The larger the value, the higher satisfaction with one's current life.

4. **Workplace Factors:**

- *Satisfaction with job:* This is a categorical variable based on answers to three questions, "qwc38: All in all, how satisfied are you with your job–very satisfied, somewhat satisfied, not too satisfied, or not satisfied at all(1-4)", "qwc41: Knowing what you know now, if you had to decide all over again about the job, what will you do? – Take same job again without hesitation, have second thoughts, or definitely not take job (1-3)", "qwc41a: What will you tell your good friend if he or she was interested in working in a job like yours?– Strongly recommend it, have doubts about recommending it, or advise against it (1-3)". The value runs from -3 to 1, with the larger the value, the more satisfied the subject with his/her job.
- *Workplace effectiveness:* This is a continuous variable, it measures the effectiveness of workplace in five aspects: autonomy, economic security, climates of respect, work life fit, and job challenge & learning. The larger the value, the more effective of the workplace.
- *Coworker support:* This a discrete variable having value from 1 to 4 measuring the support from coworkers, from less supportive to strongly supportive.
- *Difference in work hours:* between practice work hours and ideal work hours: continuous variable

1.3 Methods to Use and Outline

There are around 500 variables in the original dataset, and those 4 categories information are of great interest, so variables introduced above are selected as candidates. Although great dimension reduction had been made, there was still potential significant relationship concealed within the remaining collection of random variables. Multivariate statistical analysis is such a useful tool which can simultaneously analyze huge data and incorporate information into the statistical analysis about the relationships between all the variables.

The outline of the paper is as follows. Multivariate regression can deal with many inputs and many outputs, which is suitable for our dataset. It is provided in section 2.1. In Section 2.2,

canonical correlation analysis will be presented which take into account the correlation between two groups of variables (within and between groups). Section 2.3 provides linear and quadratic discrimination analysis method used to classify people into different mental health conditions by their demographic, family and working information. Another approach, cluster analysis, is used to explore the complete separation of mental health condition from all observations in Section 2.4. And Section 3 makes the conclusion and lists some discussion for further research. Tables and figures will be presented in the appendix.

2 Statistical Methods and Results

2.1 Multivariate Regression Analysis

Multivariate linear regression is a natural extension of multiple linear regression, where least squared method is used to study to what extent the behavior of s output variables $\boldsymbol{Y} = (Y_1, \ldots, Y_s)^{'}$ are influenced by a set of r input variables $\boldsymbol{X} = (X_1, \ldots, X_r)^{'}$. In our scenario, we set mental health category ('sleep', 'stress', 'depression') as the output responses, and put all the other covariates as predictors.

In addition, as to model selection, stepwise procedure was adopted, specifically backward direction, and decision was made based on *Wilk's lambda*. Here's the final model :

$$(\text{Sleep}, \text{Stress}, \text{Depression}) \sim \text{Age} + \text{Sat.job} + \text{Work.Eff} + \text{Edu} + \text{Sat.life} + \text{Gender} + \text{Numkid}$$

By common sense, we may think that since people spend at least eight hours with coworkers each workday, so gaining support from coworkers may make people feel released, hence good for ones mental health. However, it turned out that its effect was not significant at all. Similar case for difference in work hours, so it seems that although people complain sometimes that they have a heavy workload, it actually doesn't impact their psychological condition. The most interesting one is that we get the conclusion that money actually doesn't make people sleep better and feel happier.

From Table 1, it was found that except for satisfaction on job and number of kids at home, all other covariates have the same direction of effect on all three indexes of mental health. Some other easy inference can be made as follows:

- as employee gets older, the mental health improves.
- female employee suffered worse mental problems than male employee.
- high workplace effectiveness and the satisfaction with one's life greatly improve mental health.

Table 1: Multivariate regression coefficients from final model. Except for satisfaction of job and number of kids at home covariates, all other predictors have the same direction of effect on all three aspects of mental health.

	stress	sleep	depress
(Intercept)	4.829	4.129	2.111
age	-0.005	-0.002	-0.003
as.factor(sat.job)-2	0.0140	-0.034	0.173
as.factor(sat.job)-1	0.047	-0.154	0.169
as.factor(sat.job)0	-0.010	-0.213	0.092
as.factor(sat.job)1	-0.147	-0.306	0.075
work.eff	-0.171	-0.142	-0.128
as.factor(edul)2	-0.262	-0.232	-0.133
as.factor(edul)3	-0.337	-0.365	-0.184
as.factor(edul)4	-0.372	-0.367	-0.211
as.factor(edul)5	-0.287	-0.410	-0.207
as.factor(edul)6	-0.321	-0.342	-0.251
as.factor(sat.lifel)2	-0.594	-0.329	-0.340
as.factor(sat.lifel)3	-1.209	-0.645	-0.968
as.factor(sat.lifel)4	-1.803	-1.058	-1.327
sex	0.230	0.248	0.0562
numkid	0.0678	0.002	-0.004

- education impacts one's psychological condition, but the difference is not significant once you attain certain level of education.

2.1.1 Assumption Check

Pairwise plot and QQ-plot are used to check the normality of residuals. From the Figure 1 and Figure 2 , the multivariate regression residuals behaved pretty well. The normality assumption can be hold.

2.2 Canonical Correlation Analysis

Canonical variate and correlation analysis (CVA or CCA) is a method for studying linear relationships between two vector variates, here mental health variables vs the other three categories. The goal is to analyze correlation of $\boldsymbol{X}$ and $\boldsymbol{Y}$, i.e., to find out what aspects of demographic, family and work information are related to mental health symptoms.

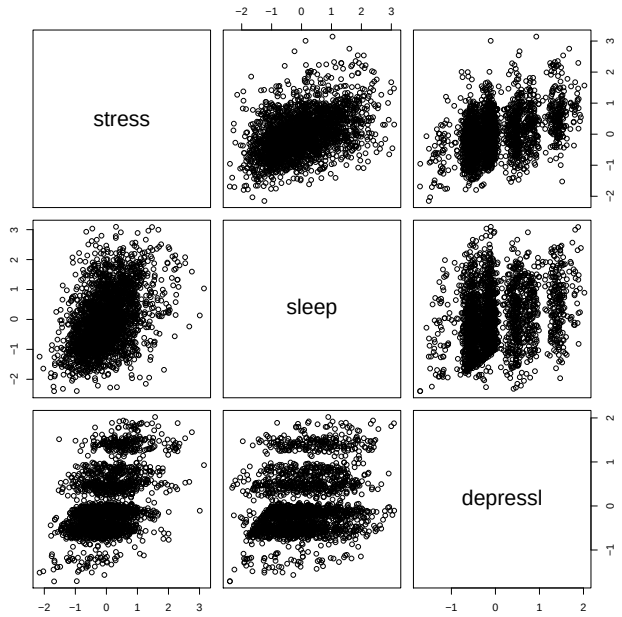

Figure 1: Pairwise plot for multivariate regression residuals show the multivariate normality assumption can be hold.

Table 2: CCA: Sample correlations

	1st Cor	2nd Cor	3rd Cor
All	0.640	0.144	0.108
Family	0.622	0.138	0.091
Work	0.427	0.104	0.085
Multi	0.639	0.137	0.102

Four canonical correlation analysis models were fitted as follows: While always setting mental health variables as vector $\boldsymbol{Y}$, first, $\boldsymbol{X}$ are all the covariates . Secondly, set another collection variables as family plus demographic information. Thirdly, compare $\boldsymbol{Y}$ with collection of working and demographic information, and finally with all variables from final model of multivariate regression.

From Figure 3, 4, 5, Plots of model with all factors and model with variates from final multivariate regression model have a clear linear trend. However, there is no obvious pattern for the 2nd pair and 3rd pair of canonical variates, since they do not explain much of the correlation.

Since we used the correlation matrix to do the cca, it makes sense to compare the coefficients in $\boldsymbol{H}$ for $\boldsymbol{Y}$ vector. As we can see from the Table 3 , whenever the satisfaction with life is included in the Y vector, it has the largest absolute value of all variates, hence it is strongly correlated with

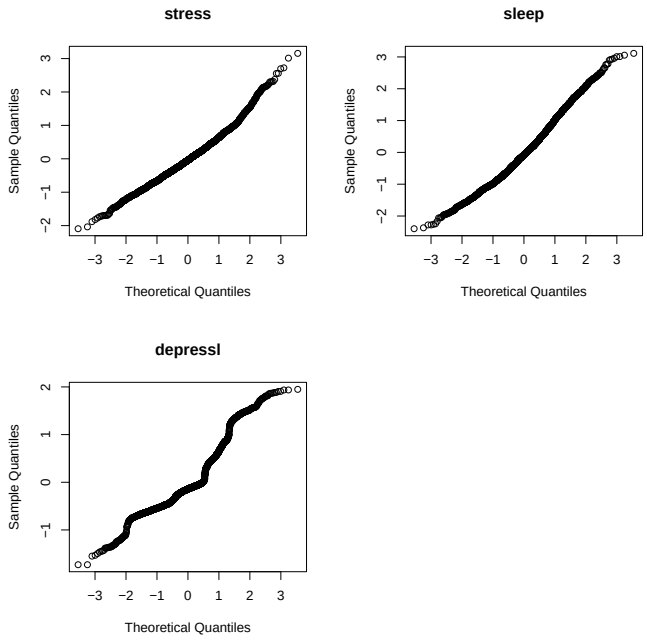

Figure 2: QQ-Plot for multivariate regression residuals

the employees mental health. For age, sex, satisfaction with job, and workplace effectiveness, they are also fairly correlated with mental health, but when satisfaction with job is excluded from the Y vector, their correlation with mental health get stronger.

For coworker support, income, difference in work hours, we got the same result as that from the multivariate regression, they were least correlated with ones mental health.

2.3 Classification and Discrimination

Another interesting approach might be from classification and discrimination. If people can be labeled into 3 groups: low stress , median stress, high stress, what we are interested in is whether we could find a good classifier using demographic, family and working information that will separate their mental health conditions as much as possible. If that is the case, then this method can really be adopted by psychologists and doctors to mental health diagnostic.

Before finding the classifier, we need to have a standard to classify people into 3 distinct groups in which they would have low, median, high extent of stress (sleep trouble, depression). Since 'stress' variable took value $1, 2, 3, 4, 5$, thus person having 'stress' (1, 2.33) would be low stress one, (2.33, 3.66) as median stress one, and the other would be high risk stress people. The same setting also hold for 'sleep' variable. As to the 'depression' covariate, it 's already a 3-level categorical variable, so there's no need to do the discretization.

In this report, the most 2 basic types of classifiers: linear and quadratic combinations of the

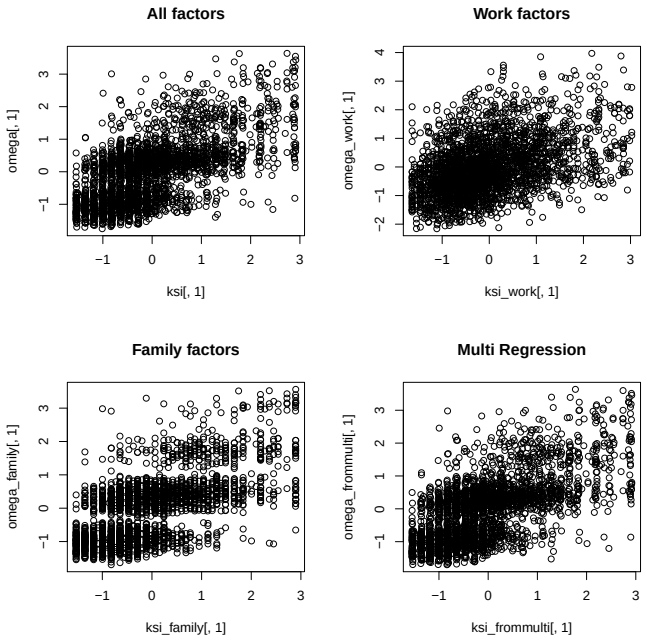

Figure 3: Pairs of canonical variates (1st)

predictor variables are used. The results are as follows:

As we can see from Figure 6, 7, and 8, LDA for 'Depression' has the best performance and according to Table 4 and 5, leave-one-out cross validation correction would have a little bit higher apparent error rate than those without cross validation. Meanwhile, linear discrimination analysis seemed to have a better overall performance than quadratic discrimination analysis, the only exception is that for 'depression' variable, LDA could not tell the person with mediate depression, while QDA method did a better job, though not so good as well.

2.4 Cluster Analysis

Cluster analysis is also known as class discovery, which basically does a data segmentation job. Our motivation for using cluster analysis is taking another way(cluster) to verify if there 's a complete separation for mental health condition among people. After all the observations have been divided into K clusters, if we are lucky and our data segmentation method does well, we would have found there's significant difference of mental health condition between these K clusters. For cluster analysis, K must be chosen in advance, here $K = 3$ according to discrimination analysis.

As we can see from Table 6, 7, although mean and median can not represent the whole characteristic of the distribution of 'sleep' and 'stress' variables, they are still quite different between clusters. Nevertheless, we still can not say cluster analysis did a good job for helping discriminate people's mental health, since apparent error rate still remains high for 'sleep' and 'stress' vari-

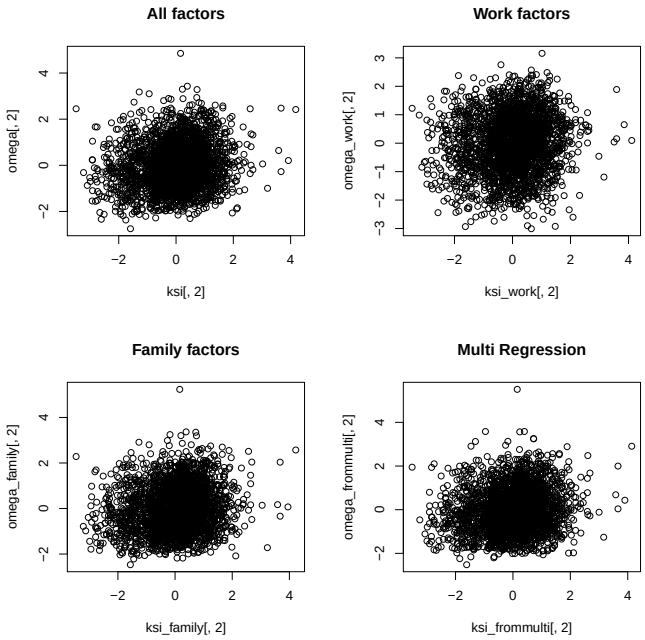

Figure 4: Pairs of canonical variates (2nd)

ables. The situation was even worse for 'depression' variable, as the distributions of 'depression' in 3 clusters are quite similar to each other.

3 Conclusion and Discussion

Through this article, multivariate regression analysis and canonical correlation analysis seem to have a better, more clear explanation on the relationship between mental health and demographic, family and workforce information. By discrimination analysis, we had a better result for 'depression ' variable, while for the other variables, those classifiers did not work so well. Cluster analysis could roughly give a big picture, but no reliable inference can be made from that.

Although significant relationship has been found through the project, there are still some further work to do. First, as to the missing data, we omitted about 5% of the original data, which might give rise to biased results. Second, in the multivariate linear regression, we only used the stepwise procedure for variable selection, multi-collinearity problems were omitted. Some variables deleted through stepwise procedure may be better predictors, but its effect might be covered by bad predictors which lead to its elimination in variable selection procedure. Third, some of the variables are ordinal or discrete data, although we took them as continuous variables, it's better to take into account the latent variable behind those ordinal and discrete covariates. In addition, *Factor Analysis* may also be used to figure out the most influential factor on mental health.

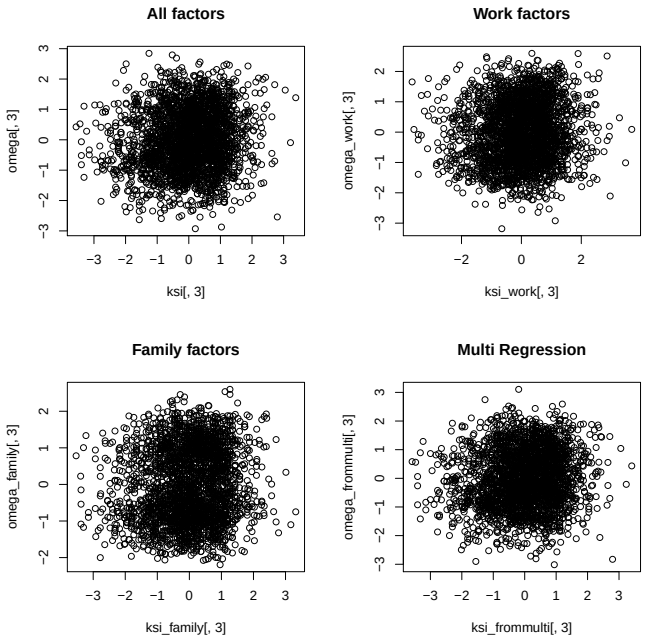

Figure 5: Pairs of canonical variates (3rd)

4 Acknowledge

We thank Izenman's book [1] and Dr. Park's notes, as well as his code, web resources [2].

References

[1] A.J. Izenman. *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer Verlag, 2008.

[2] Trevor Park. Sta 6707 multivariate data analysis. http://www.stat.ufl.edu/~tpark/STA6707/, Fall 2010.

Table 3: CCA: Coefficients of $\boldsymbol{H}$

	All	Work	Family	Multi
sat.job	-0.115	-0.416		-0.113
work.eff	-0.156	-0.507		-0.168
cowkspt	-0.013	-0.042		
difwkhours	-0.012	-0.006		
incml	-0.053		-0.066	
sat.life	-0.809		-0.936	-0.818
numkid	0.099		0.088	0.093
age	-0.107	-0.263	-0.122	-0.115
sex	0.171	0.333	0.161	0.175
edul	-0.037	-0.176	-0.044	-0.059

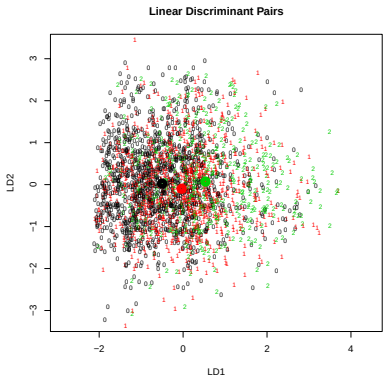

Figure 6: Worst behavior of LDA for 'Sleep'

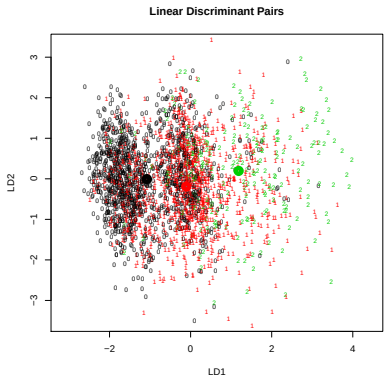

Figure 7: Mediate behavior of LDA for 'Stress'

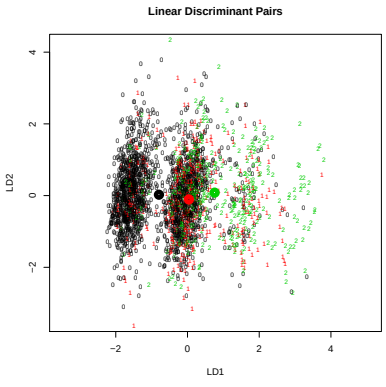

Figure 8: Best behavior of LDA for 'Depression'

Table 4: Confusion table from LDA, QDA and also those with leave-one-out cross-validation

		Predict											
Term	TRUE	LDA			LDA(CV)			QDA			QDA(CV)		
		0	1	2	0	1	2	0	1	2	0	1	2
sleep	0	1008	162	43	1003	164	46	951	191	71	936	205	72
	1	594	254	66	597	248	68	538	282	93	561	248	104
	2	202	142	87	202	146	83	191	109	131	195	125	111
stress	0	815	336	10	813	338	10	833	308	20	818	323	20
	1	346	775	57	350	771	57	372	742	64	384	723	71
	2	17	127	74	17	135	66	22	120	76	22	126	70
depression	0	1655	0	66	1654	0	67	1606	25	90	1601	29	91
	1	365	0	79	365	0	79	340	31	73	344	19	81
	2	229	0	163	229	0	163	205	11	176	209	21	162

Table 5: Apparent error rate from 4 models

Term	LDA	LDA(CV)	QDA	QDA(CV)
sleep	0.47	0.478	0.467	0.494
stress	0.349	0.355	0.354	0.37
depression	0.289	0.289	0.291	0.303

Table 6: Sleep Mean and Median for 3 Groups

	A	B	C
Mean	2.87	2.15	2.47
Median	3	2	2.33

Table 7: Stress Mean and Median for 3 Groups

	A	B	C
Mean	3.14	2.16	2.63
Median	3	2	2.6

Table 8: Depression Table

	Depr		
Group	0	1	2
A	168	76	135
B	950	143	73
C	603	225	184