Part II: Examining the factors that influence depression level of people

Student No. : D16129273
Name : Han Tang
Programme Code: DT228A DA
Option Chosen: Option A
version of R: version 3.5.1 (2018-07-02)
library_used: pastects, ggplot2, psych, semTools, car, dplyr, userfriendlyscience, ppcor, olsrr, lmSupport, strargazer, dplyr

Introduction

Previous exploration in Part I proved the relationship between the sleepiness scale and the level of depression. The reasons behind being depressed can be multiple and complex. The research of the factors that influence depression level of people is useful not only for people under depression but can instruct people the way to stay relaxed and peaceful. For government and health institute, this research can also help to filter the group of people under depression to monitor their psychological condition and help them precisely. This paper will identify additional factors which could be related to the depression level. A multiple linear regression model capable of predicting people's level of depression will be built by the factors identified before.

Additional research on the topic has found that insomnia, anxiety, and depression are intercorrelated (Markus Jansson-Fröjmark, Karin Lindblom, 2008). Their research suggested that there is a bidirectional relationship between depression, anxiety and insomnia and they are intertwined over time.

Other research also noted that nicotine, which is the major compound in a cigarette, can affect the level of depression (Picciotto Marina R., Brunzell Darlene H., Caldarone Barbara J., 2002). Their studies suggested that the receptors of nicotine in brains are involved in people's behavior related to stress response, anxiety, and depression. However, the mechanism is complicated and still uncleared.

The WHO World Health Survey also suggested that the level of depression is also related with a health condition (Saba M., Somnath C., Emese V., Ajay T., Vikram P., Bedirhan U., 2007). The researchers surveyed more than 200000 participants from 60 countries all over the world. Their studies noted that an average of between 9.3% and 23.0% participants with one or more chronic physical disease had comorbid depression, which is statistical significantly ($p < 0.0001$) higher than the likelihood of having depression but with no chronic physical disease. Their research also suggested that a high level of depression can also worsen mean health score.

Another study (Joan M. G., Rebecca F., Stephen A. S., Michael M., 2002) also indicated that people's control at home and work can influence their level of depression significantly. Meanwhile, the effect of control at life on their level of depression also varies by gender and their social class.

I will conduct my study over a dataset with 271 samples and check the normality distribution of each variable. Afterward, this research will look at the influence of the factors such as trouble

falling asleep, level of anxiety, history of smoking, general health rate, highest educational level achieved, marital status and gender over the level of depression. I will combine the most significant factors to produce a multilinear regression model. Afterward, I will test whether this model can predict the level of depression precisely.

.

Methodology

The dataset is from a questionnaire about sleep quality and the effect of sleep on daily life. This dataset has 55 variables and gathered from 271 individuals.

Records with the missing value of the interest of variables are removed from the dataset as it can change the shape of the distribution and influence the multilinear regression model if missing values are replaced to zero or any other value.

For this analysis, a significant of 0.05 was chosen.

The variables of interest are ‘depress’, ‘anxiety’,‘getsleeprec’, ‘smoke’, ‘healthrate’, ‘edlevel’, and ‘sex’. These variables are extracted from the original dataset and formed a new dataset. Detailed descriptions (the normality distribution, outliers, etc) are covered in the part the Dataset.

The Dataset

I formed a new dataset composed of the variables of interest and remove all the records have missing data. There are 260 records left after all records with missing value are removed, which is still sufficient for conducting my multilinear regression model.

Variable of interest

depress (n = 260, M = 3.51, SD = 2.96): The variable about the level of depression is the dependent variable, which is measured by the depression scale of the subject surveyed. Its full name is the Total HADS Depression Score. It continuously ranges from 0 to 21 and the higher the score is, the more severe depression the subject is in.

anxiety (n = 260, M = 6.35, SD = 3.5): The variable about the level of anxiety is a continuous independent variable, which is measured by the anxiety scale of the subject surveyed. Its full name is the Total HADS Anxiety Score. It continuously ranges from 0 to 21 and the higher the score is, the more severe anxiety the subject is in.

getsleprec (n = 260) is a nominal variable indicating 1 (Have problem getting to sleep, n = 102) and 0 (Do not have problem getting to sleep, n = 158). Subjects chose ‘yes’ in the question ‘Do you have trouble falling asleep?’ were classified as 1.

smoke (n = 260) is a nominal variable indicating 1 (Smoke, n = 33) and 0 (Do not smoke, n = 227). Subject chose ‘Yes’ in the question ‘Do you smoke?’ were classified as 1. According to the EuroStats, the proportion of people smoke ranging from 26% in Greece to 7% in Sweden in the European Union. The distribution of this variable in my dataset stays in this range and closes to the

average value over all the Europe, which indicates that this distribution can basically represent the situation in the Europe.

healthrate (n = 260) is an ordinal variable indicating the rate of general health, ranging from 1 (very poor) to 10 (very good). For the purpose of building a multilinear regression model this variable will be treated as a numerical scale and standardised. However, for the purpose of simulating the situation in real life precisely, this range of this ordinal variable is narrowed from 1-10 to 1-2, which means the subject who rates 10 on their general health could be only 2 times more healthy than those who rates 1 on their general health condition (which could have been 10 times, which may can not represent the real life condition).

edlevel (n = 260) is an ordinal variable indicating the highest education level achieved by the subject, ranging from 1 (primary school level of education) to 5 (postgraduate degree level of education). The higher the value is, the higher education level the subject gets. For the purpose of building a multilinear regression model this variable will be treated as a numerical scale. The mode of the variable is 5 (Postgraduate degree level of education, n = 128). The second most common value of the variable is 4 (Undergraduate degree level of education, n = 70). According to the EuroStats, 80% of the adults age from 25 to 54 in the EU had completed at least an upper secondary level of education, which indicates that the distribution of this variable in this dataset can basically represent the situation in the Europe.

sex (n = 260) is a categorical variable indicating the gender of each subject. The gender make up of the population is Female (sex = 0, n = 141) and Male (sex = 1, n = 119). According to the EuroStats, the sex ratio (Male to Female) in the Europe is approximately 1.05, which is quite different from the proportion in the dataset. However, the conclusion we got from the multilinear regression model can still have facticity.

Normality of numeric variables

The numeric variables were standardised for building the multilinear regression model and the test of skewness and kurtosis is conducted.

Table1. Skew and kurtosis of numeric variable

	"anxiety"	"depress"	"healthrate"
"skew (g1)"	0.54	0.93	-0.87
"Excess Kur (g2)"	-0.02	0.03	0.57

Table1 highlights the result, both the kurtosis and skewness of the variable anxiety, the variable depress and the variable health rate falls within the range of -2/+2. These three numeric variables are normally distributed and suitable for our model of multilinear regression model.

Correlation and Difference Tests

Table2. Correlation Test of depress

	"df"	"r"	"p_value"
"anxiety"	258	0.59	"<.001"
"healthrate"	258	-0.31	"<.001"

The relationship between the Total HADS Depression score and the Total HADS Anxiety score was investigated using a Pearson correlation. A strong positive correlation was found ($r = 0.59$, $n = 258$, $p < .001$).

The relationship between the Total HADS Depression score and the General Health Rate was investigated using a Pearson correlation. A moderate negative correlation was found ($r = -0.31$, $n = 258$, $p < .001$).

Table3. Correlation Test of anxiety and health rate

"df"	"r"	"p_value"
258	-0.29	"<.001"

The relationship between the Total HADS Anxiety score and the General Health Rate was investigated using a Pearson correlation. A small negative correlation was found ($r = -0.29$, $n = 258$, $p < .001$).

A Student T Test compare the nominal variable getselrec versus depression. A Student T Test was used to test the difference from the group ‘yes’ and ‘no’ with respect to level of depression.

A Student T Test compare the nominal variable getselrec versus anxiety. A Student T Test was used to test the difference from the group ‘yes’ and ‘no’ with respect to level of depression.

A Student T Test compare the nominal variable getselrec versus healthrate. A Student T Test was used to test the difference from the group ‘yes’ and ‘no’ with respect to level of depression. However, the difference in the variance of different groups is too large as it did not pass the levene’s test.

A Student T Test compare the nominal variable smoke versus depression. A Student T Test was used to test the difference from the group ‘yes’ and ‘no’ with respect to level of depression. However, the result of levene’s test reaches a p-value of 0.51, which means there is a difference in the variance of different groups. The variable smoke is abandoned finally in order to conduct my multilinear regression model precisely.

A Student T Test compare the nominal variable sex versus depression. A Student T Test was used to test the difference from the group ‘Male’ and ‘Female’ with respect to level of depression. However, the result of levene’s test reaches a p-value of 0.29, which means there is a difference in the variance of different groups. The variable sex is abandoned finally in order to conduct my multilinear regression model precisely.

A one-way between-groups analysis of variance was conducted to explore the impact of highest education level achieved on the level of depression. Participants were divided into five groups according to their highest level of education (1 = primary; 2 = secondary; 3 = trade; 4 = undergrad; 5 = postgrad).

A one-way between-groups analysis of variance was conducted to explore the impact of highest education level achieved on the level of anxiety. Participants were divided into five groups according to their highest level of education (1 = primary; 2 = secondary; 3 = trade; 4 = undergrad; 5 = postgrad).

A one-way between-groups analysis of variance was conducted to explore the impact of highest education level achieved on the general health rate. Participants were divided into five groups according to their highest level of education (1 = primary; 2 = secondary; 3 = trade; 4 = undergrad; 5 = postgrad).

Table4. Student T Test Result

	"depress"	"anxiety"
"getsleprec"	"p-value = 0.019"	"p-value < .001"

The result of Student T Test shows that there is a statistical significant difference between the participants with problem of getting asleep and those without this problem in the level of their anxiety.

Table5. The result of one-way between-groups test of edlevel on depress

	SS <fctr>	Df <fctr>	MS <fctr>	F <fctr>	p <fctr>
Between groups (error + effect)	6.84	4	1.71	1.73	.144
Within groups (error only)	252.16	255	0.99	*NA*	*NA*

```
### Post hoc test: Tukey

    diff  lwr   upr  p adj
2-1 0.39  -1.61 2.38 .984
3-1 0.9   -1.09 2.9  .725
4-1 0.56  -1.4  2.52 .934
5-1 0.42  -1.53 2.37 .976
3-2 0.52  -0.19 1.22 .260
4-2 0.18  -0.42 0.77 .926
5-2 0.04  -0.52 0.59 1.000
4-3 -0.34 -0.94 0.25 .515
5-3 -0.48 -1.04 0.07 .120
5-4 -0.14 -0.55 0.27 .875
```

Post-hoc comparisons using the Turkey HSD test indicated that there is only statistical significant difference between the group 3 (trade level of highest education, M = 4.73, SD = 3.2) and the group 5 (postgraduate level of highest education, M = 3.26, SD = 3.03) over the level of depression.

Table6. The result of one-way between-groups test of edlevel on anxiety

	SS <fctr>	Df <fctr>	MS <fctr>	F <fctr>	p <fctr>
Between groups (error + effect)	7.02	4	1.76	1.78	.134
Within groups (error only)	251.98	255	0.99	*NA*	*NA*

```
### Post hoc test: Tukey

    diff  lwr   upr  p adj
2-1 0.36  -1.63 2.36 .987
3-1 0.61  -1.38 2.6  .918
4-1 0.29  -1.67 2.25 .994
5-1 0.11  -1.84 2.06 1.000
3-2 0.25  -0.46 0.95 .870
4-2 -0.07 -0.67 0.52 .997
5-2 -0.25 -0.81 0.3  .720
4-3 -0.32 -0.92 0.28 .580
5-3 -0.5  -1.05 0.05 .098
5-4 -0.18 -0.59 0.23 .739
```

Post-hoc comparisons using the Turkey HSD test indicated that there is only statistical significant difference between the group 3 (trade level of highest education, M = 4.73, SD = 3.2) and the group 5 (postgraduate level of highest education, M = 3.26, SD = 3.03) over the level of anxiety.

Table7. The result of one-way between-groups test of edlevel on healthrate

	SS	Df <fctr>	MS <fctr>	F <fctr>	p <fctr>
Between groups (error + effect)		4	6.53	7.15	<.001
Within groups (error only)		255	0.91	*NA*	*NA*

```
### Post hoc test: Tukey

    diff  lwr   upr  p adj
2-1 -1.37 -3.29 0.55 .288
3-1 -1.16 -3.08 0.76 .461
4-1 -0.88 -2.77 1    .697
5-1 -0.51 -2.38 1.36 .945
3-2 0.21  -0.47 0.89 .913
4-2 0.48  -0.09 1.06 .141
5-2 0.86  0.33  1.39 <.001
4-3 0.27  -0.3  0.85 .683
5-3 0.65  0.12  1.18 .008
5-4 0.38  -0.01 0.77 .065
```

Post-hoc comparisons using the Turkey HSD test indicated that there is only statistical significant difference between the group 2 (secondary level of highest education, M = 3.16, SD = 2.58) and the group 5 (postgraduate level of highest education, M = 3.26, SD = 3.03) over general health rate.

Results

Multilinear Regression model will be built to predict the value of the level of depression in this part. A baseline model will be conducted firstly, then additional independent variables will be added to this model gradually to build new model. The fit and usefulness of these models will also be checked in this part.

Hypothesis

1. There is no significant prediction of the level of depression by the level of anxiety and general health rate.
2. There is no significant prediction of the level of depression by the level of anxiety, general health rate and trouble falling asleep.
3. There is no significant prediction of the level of depression by the level of anxiety, general health rate, trouble falling asleep and the highest education level achieved.

Model 1

A multilinear regression was conduct to predict a participant's level of depression based on the continuous variable level of anxiety and the variable general health rate. A significant regression equation was found ($F(257) = 77.499$, $p < .001$) with an adjusted R^2 of 0.371. The predicted value of depress of participant is equal to 0.548(anxiety) - 0.159(healthrate), where both anxiety and health rate are both standardised continuous variables. The students standardised level of depression increases 0.548 standard deviations for each standard deviation increase in level of anxiety and decreases 0.159 standard deviations for each standard deviation increase in general health rate. Both anxiety ($p < .001$) and health rate ($p < .001$) were significant predictors of level of depression.

Table8. The First Multilinear Regression Model

	Dependent variable:
	depress
anxiety	0.548***
	(0.051)
healthrate	-0.159***
	(0.051)
Constant	-0.000
	(0.049)
Observations	260
R2	0.376
Adjusted R2	0.371
Residual Std. Error	0.793 (df = 257)
F Statistic	77.499*** (df = 2; 257)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fit and usefulness check of Model 1

Table9. The result of F-test on Model 1

Analysis of Variance Table

Response: sleepdata$depress

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
sleepdata$anxiety	1	91.438	91.438	145.4524	< 2.2e-16	***
sleepdata$healthrate	1	6.001	6.001	9.5452	0.002225	**
Residuals	257	161.562	0.629			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The P-value of F-test is less than our significant level we set, which indicates that the model 1 provides a better fit than the intercept-only model or predicting values of the outcome by using the mean.

Figure1. Model 1: Density Plot of model residuals

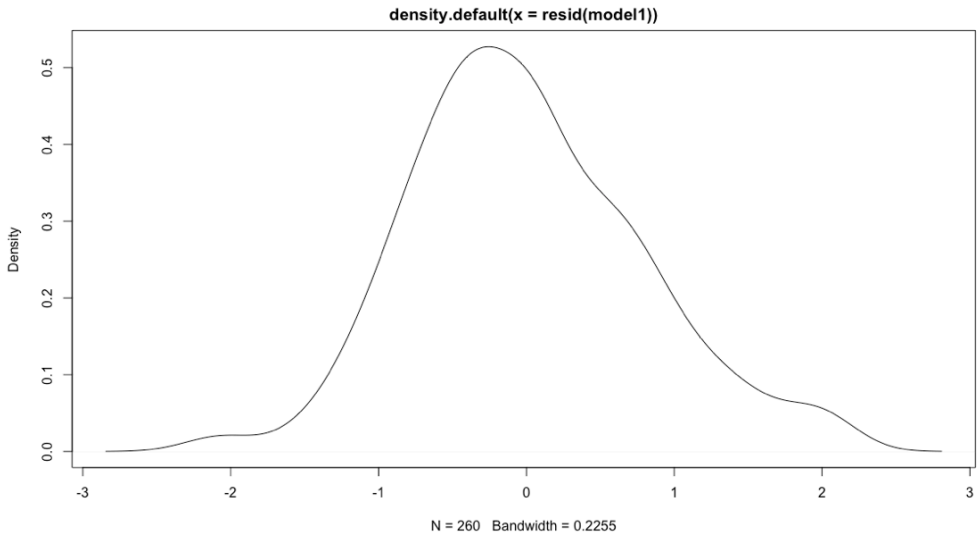
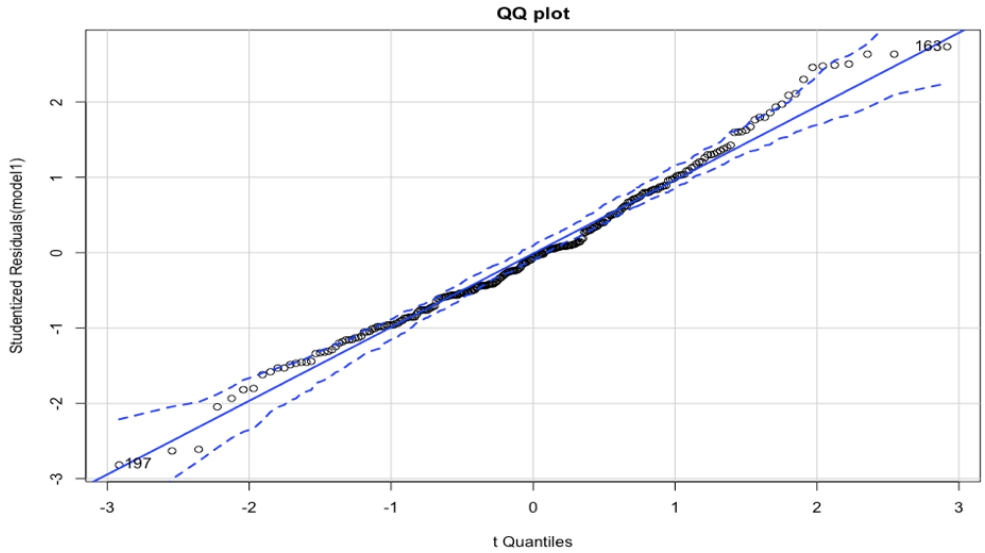


Figure2. Model1 : QQ Plot of model residuals



An analysis of Cooks distance was carried out, which showed that the data contained no extreme outliers and none with Cook's distance > 1 as outlined in (Field (2013)) (Max Cooks Distance = 0.128). Tests of if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Anxiety, VIF= 0.92; Healthrate, VIF= 0.92) and were within acceptable levels (VIF <2.5) as outlined in (Tarling (2008)). The density plot of standardised residuals indicated that the data contained normally distributed errors, as did the normal Q-Q plot of standardised residuals, which showed all points were close to the line and plotted in Figure 2. The data also meets the assumption of non-zero variances of the predictors.

Model 2

A multiple linear regression was calculated to predict a participant's level of depression based on the nominal variable getsleprec . The nominal variables were recoded as follows; getsleprec (1= yes, 0= no).

Table9. Comparison between the first model and the second model

	Dependent variable:	
	depress	
	(1)	(2)
anxiety	0.548***	0.554***
	(0.051)	(0.054)
healthrate	-0.159***	-0.158***
	(0.051)	(0.052)
getsleprec1		-0.036
		(0.106)
Constant	-0.000	0.014
	(0.049)	(0.064)
Observations	260	260
R2	0.376	0.376
Adjusted R2	0.371	0.369
Residual Std. Error	0.793 (df = 257)	0.794 (df = 256)
F Statistic	77.499*** (df = 2; 257)	51.528*** (df = 3; 256)
Note:		*p<0.1; **p<0.05; ***p<0.01

However, the comparison between the second model we get and the first model indicates that model2 does not provide a better fit than the first model. The variable interaction between level of anxiety and trouble getting asleep was abandoned for the purpose of building a precise multilinear regression model.

Model 3

A multiple linear regression was calculated to predict a participant's level of depression based on the continuous variables level of anxiety and health rate and nominal variable edlevel. The nominal variables were recoded as follows; edlevel (1= primary school, 2 = secondary school, 3 = trade training/ post secondary training , 4 = undergraduate degree, 5 = postgraduate degree).

Table10. The Third Multilinear Regression Model

	Dependent variable:
	depress
anxiety	0.543***
	(0.052)
healthrate	-0.178***
	(0.054)
edlevel2	-0.054
	(0.582)
edlevel3	0.367
	(0.581)
edlevel4	0.248
	(0.569)
edlevel5	0.272
	(0.564)
Constant	-0.237
	(0.561)
Observations	260
R2	0.389
Adjusted R2	0.374
Residual Std. Error	0.791 (df = 253)
F Statistic	26.813*** (df = 6; 253)
Note:	*p<0.1; **p<0.05; ***p<0.01

A statistical regression model was found ($F(253) = 26.813$, $p < .01$) with an adjusted $R2$ of 0.374. The interpretation of this model are in the table11.

Table11. Interpretation of the model 3

	constant	anxiety	health rate	edlevel2	edlevel3	edlevel4	edlevel5	dependent variable = depress
Secondary school level of education	-0.237	0.543	-0.178	-0.054				0.074
trade training/ post secondary level of education	-0.237	0.543	-0.178		0.367			0.495
undergraduate level of education	-0.237	0.543	-0.178			0.248		0.376
postgraduate level of education	-0.237	0.543	-0.178				0.272	0.4

The participant's level of depress increases 0.543 standard deviations for each standard deviation increase in level of anxiety and decreases 0.178 standard deviations for each standard deviation increase in health rate. Participants of achieving post secondary level of education , of achieving undergraduate level of education and of achieving postgraduate level of education can expect to score 0.367, 0.248 and 0.272 when compared to participants of achieving primary level of education. However, the variable edlevel and the constant were not significant predictor of the level of depression. The rest variables are significant predictors of the level of depression.

Fit and usefulness check of Model 3

Table12. F-test Result on the Model 3

```
Analysis of Variance Table

Response: sleepdata$depress
                     Df  Sum Sq Mean Sq  F value    Pr(>F)
sleepdata$anxiety     1  91.438  91.438 146.1170 < 2.2e-16 ***
sleepdata$healthrate  1   6.001   6.001   9.5888  0.002178 **
sleepdata$edlevel     4   3.238   0.809   1.2936  0.273016
Residuals           253 158.324   0.626
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the result of F-test, we can also find that the P-value of the variable level of anxiety and the general health rate were less than the significant level. But we can not prove that the model 3 provides a better fit than model 1 as the P-value of the variable level of education achieved is above the significant level.

An analysis of Cooks distance was carried out, which showed that the data contained no extreme outliers and none with Cook's distance >1 as outlined in (Field (2013)) (Max Cooks Distance = 0.06). Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (Anxiety, VIF= 1.10; Healthrate, VIF= 1.19; edlevel, VIF= 1.12) and were within acceptable levels (VIF < 2.5) as outlined in (Tarling (2008)). The density plot of standardised residuals indicated that the data contained normally distributed errors, as did the normal Q-Q plot of standardised residuals, which showed all points were close to the line. The data also meets the assumption of non-zero variances of the predictors.

Figure3. Model3： Density Plot of model residuals

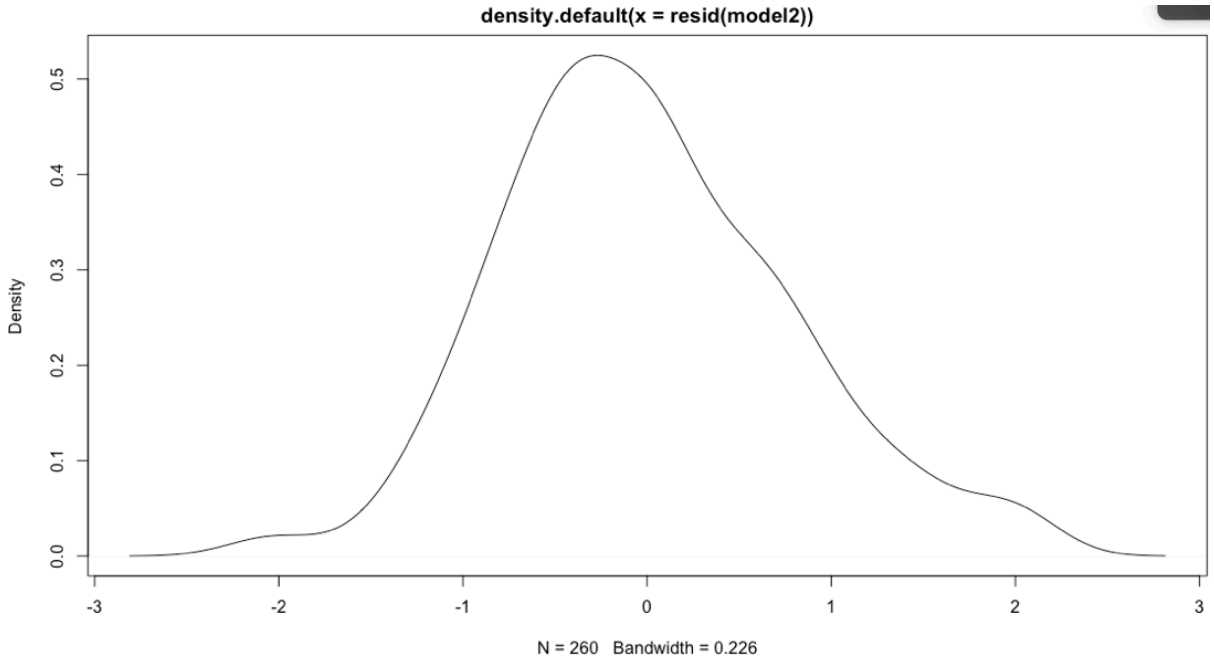
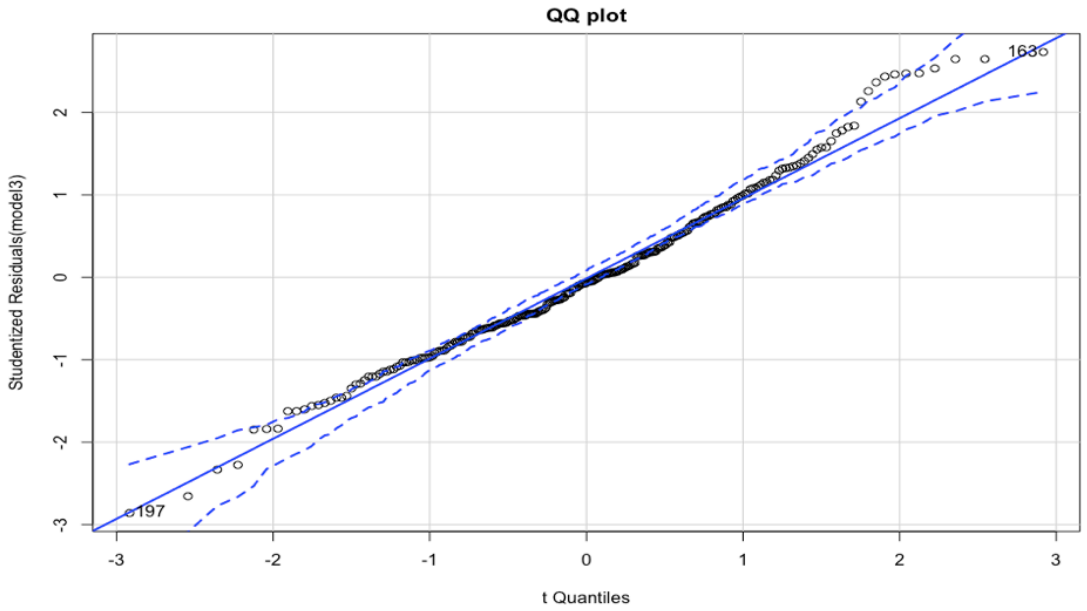


Figure4. Model3 : QQ Plot of model residuals



Discussion

The purpose of this paper is to build a multilinear regression model can predict the level of depression in population so that people with depression can be classified and helped. This model can be meaningful both for individuals and organisations such as governments, etc.

According to this model, the level of anxiety and general health rate are correlated with the level of depression. However, there is no statistical significant evidence indicates that variables gender, highest education level achieved, history of smoking, having problem get to sleep were related to the level of depression.

Three multilinear regression models are built in this paper. However, it could barely be proved that the successive models can improve the predicability of the dependent variable compare to the baseline model.

Further studies should be conducted to identify more nominal variables related with the level of depression to help this model predicting the depression situation over population precisely.

Reference

Markus J., Karin L. (2008). A bidirectional relationship between anxiety and depression, and insomnia? A prospective study in the general population. *Journal of Psychosomatic Research,* 64, 443–449.

Marina R. P., Darlene H. B., Barbara J. C. (2002). Effect of nicotine and nicotinic receptors on anxiety and depression. *neuroreport,* 13, 1097-1106.

Saba M., Somnath C., Emese V., Ajay T., Vikram P., Bedirhan U. (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. *The Lancet,* 370, 851-858.

Joan M. G., Rebecca F., Stephen A. S., Michael M., (2002). The importance of low control at work and home on depression and anxiety: do these effects vary by gender and social class? *Social Science & Medicine,* 54, 783-798.