Statistical Analysis on Depression

Katranitsiotis Panagiotis[1]

[1]pkatranitsiotis@gmail.com

Abstract. This report investigates the factors that affect the psychological condition depression. Several variables will be taken into account to produce accurate statistical insights using language R. More specifically, factors such as gender, age, married status, employability, other mental illnesses etc. will be used for this accomplishment. Both parametric and non-parametric methods will be performed to search which variables have an immense impact on depression. All the tests will be investigated on a global scale. Furthermore, a linear regression model of life conditions of people living on non-urban areas and depression, will be performed. Finally, logistic regression will be also used for predicting whether a person suffers from depression or not, given the correlated variables as the predictors.

Keywords: Depression; R; X-squared; T-test; Anova; Correlation; Linear Regression; Logistic Regression.

Contents

1 Introduction

No one should underestimate that depression is a widely common psychological condition that numerous people suffer from. The World Health Organization ranks depression as one of the most significant causes of disability disorders and can be found even among young people [1]. There are several factors that lead people to suffer from this highly considerable mental disorder. This is the main reason for the importance of a statistical analysis of depression, as it will be a very helpful tool both for scientists and psychiatrics to deal with it. More specifically, a research of which variables lead people to depression will be analyzed extensively.

The main purpose of this paper is to investigate depression and the ways it affects people. More specifically, several datasets will be used for this analysis. At first, it is highly important to search whether unemployment leads people to deal with depression or not. In addition, its immense impact on each gender is considerable, as it is not clarified if depression affects males and females in the same rate. Furthermore, it is essential to investigate depression comparing to other mental illnesses among the population, in order to understand if it is the most noticeable disease for the society. More specifically, depression will be compared to other mental illnesses i.e. Schizophrenia, Bipolar and Anxiety disorder. Finally, a research in non-urban areas will be performed for finding the correlations among the variables related to depression and regression models will be tried for predicting the related ones. All these research questions are essential and helpful for the effective analysis and accurate insights of depression.

In this paper several methods, both parametric and non-parametric, will be used for an accurate statistical analysis of this condition. A combination of graphs, tests such as t-test, x^2 and anova and regression models will be performed to clarify which variables are related to depression.

2 Data Processing

2.1 Pre-processed steps and chosen datasets

For this research several datasets will be used for an accurate statistical analysis on depression. The main dataset, retrieved from Kaggle, refers to a study of the life conditions of people that live in non-urban areas and whether they deal with depression or not [2]. As shown in table 1 (Appendix A), numerous factors will be considered in this survey both demographic data and factors such as the basic income and expenses, owing business, etc. However, the variable ville_id, which refers to the id of the village of its participant is not unique. As a result, the independency of the data is not occurred and thus this data will be used exclusively for the creation of the regression models. More specifically, a correlation analysis among these variables will be performed in order to understand its relations and then both linear and logistic regression will be used as the prediction models for depression.

For the X^2 test, the data used, refers to unemployment and depression [3]. More specifically, this specific dataset contains 32 variables, but for the purpose of this analysis, only the variables Unemployment and Depression will be considered. These two variables arised from this dataset are explained extensively in Table 2 (*Appendix A*). In this dataset the observations are independent and thus the tests can be performed. The main purpose of this analysis is to conclude whether depression is affected by unemployment or not.

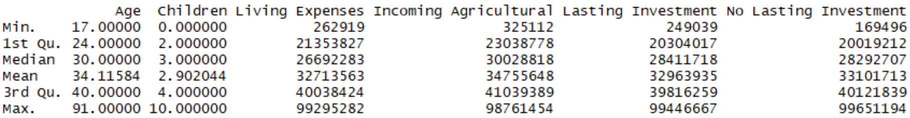
In addition, for the investigation of depression between males and females, the t-test will be used. The data for this analysis refers to the percentage of the global population that suffer from depression, categorized by gender [4]. As shown in Table 3 (*Appendix A*), the structure of this dataset is not easily manageable for the performance of the t-test. Consequently, as shown in Table 4 (*Appendix A*), a restructured of them will be used for this analysis, especially for 2017 which is its most recent year.

Furthermore, for the investigation of depression among the other types of mental problems, the ANOVA test will be performed. The dataset used for this purpose, contains variables such as the percentage of population for several illnesses, as shown in table 5 (*Appendix A*) [5]. For the purpose of this paper, only the mental illnesses will be considered i.e. Depression, Schizophrenia, Bipolar and Anxiety disorder. However, the dataset is not properly structured for this analysis. As a result, a restructure of the data was necessary for the ANOVA as shown in table 6 (*Appendix A*). More specifically, for this analysis we chose the most recent year, i.e. 2017.

All the above analysis will be explained extensively in the next chapters. Furthermore, its test will be performed with 95% confidence. Consequently, each p-value will be compared to a significance level of 0.05.

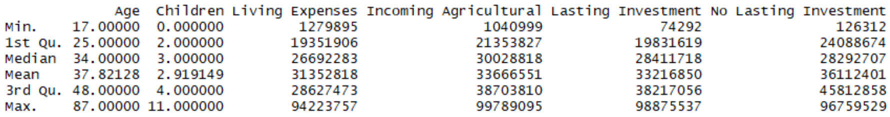
2.2 Descriptive Statistics

Several summary statistics and graphs are presented in this section in order to provide the first insights about the related variables to depression in non-urban areas. More specifically, in Figures 1 and 2, a summary of the variables of age, number of children, living expenses, incoming agricultural, lasting and non-lasting Investment are presented according to depression. No significant difference among these statistics is observed between depressed and non-depressed people. An extensive analysis of the related variables will be performed in the next chapters.



```
             Age  Children Living Expenses Incoming Agricultural Lasting Investment No Lasting Investment
Min.    17.00000  0.000000           262919                325112             249039                169496
1st Qu. 24.00000  2.000000         21353827              23038778           20304017              20019212
Median  30.00000  3.000000         26692283              30028818           28411718              28292707
Mean    34.11584  2.902044         32713563              34755648           32963935              33101713
3rd Qu. 40.00000  4.000000         40038424              41039389           39816259              40121839
Max.    91.00000 10.000000         99295282              98761454           99446667              99651194
```

Fig. 1 Central Tendency of the variables for non-depressed people



```
             Age  Children Living Expenses Incoming Agricultural Lasting Investment No Lasting Investment
Min.    17.00000  0.000000          1279895               1040999              74292                126312
1st Qu. 25.00000  2.000000         19351906              21353827           19831619              24088674
Median  34.00000  3.000000         26692283              30028818           28411718              28292707
Mean    37.82128  2.919149         31352818              33666551           33216850              36112401
3rd Qu. 48.00000  4.000000         28627473              38703810           38217056              45812858
Max.    87.00000 11.000000         94223757              99789095           98875537              96759529
```

Fig. 2 Central Tendency of the variables for depressed people

In Figure 3, histograms of Age, Number of Children, Living Expenses, Incoming Agricultural, Lasting and non-lasting investments as well as the normal distribution line of each variable, are presented. It seems that none of these variables are normally distributed. However, because of the central limit theorem, as the dataset contains more than 30 observations, the normality of the data can be assumed [6].

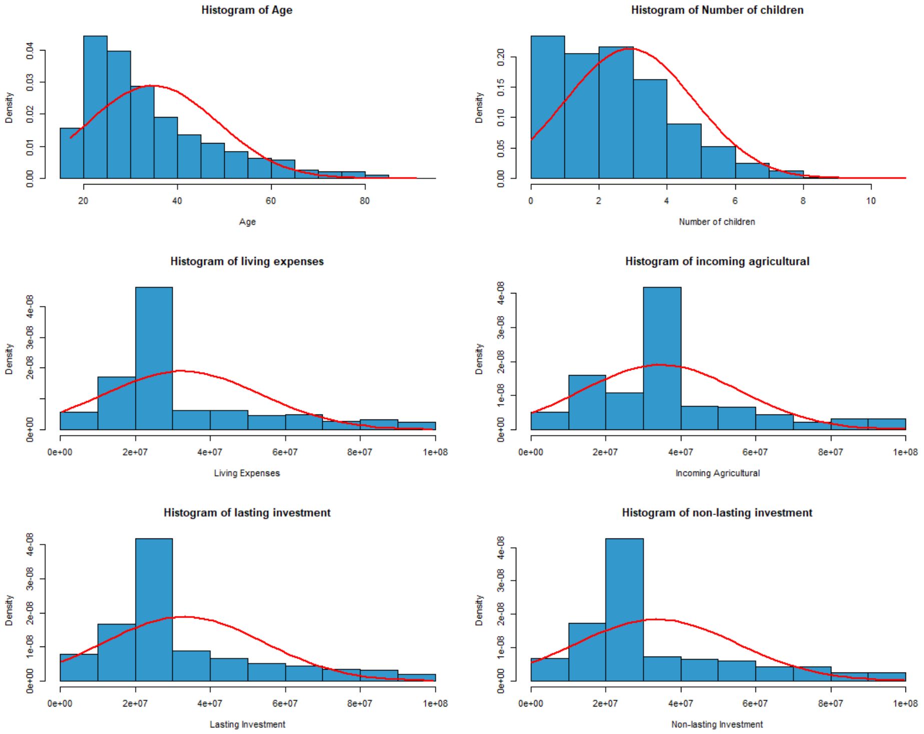


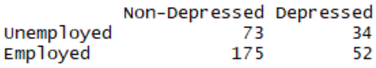
Fig. 3 Histograms of Age, Number of Children, Living Expenses, Incoming Agricultural, Lasting and non-lasting investments.

3 Analysis of Depression

3.1 X^2 test

The purpose of this section is to examine the relationship between the employment and depression. The dataset for this purpose is shown in table 2 (Appendix A) and is referred to a questionnaire about employability and depression. More specifically, the X-squared test will be performed in order to investigate if these two variables are independent.

As shown in figure 4, we observe that there is no significant difference between the employed and unemployed depressed people. More specifically, we observe that most of the depressed people are employed. In addition, most of the non-depressed people are employed and in fact this number is almost twice comparing to the unemployed non-depressed people. As a result, we expect that depression is not affected by unemployment.



```
           Non-Depressed Depressed
Unemployed            73        34
Employed             175        52
```

Fig. 4 Employability and Depression crosstabulation matrix

Nevertheless, the above table is not efficient to conclude if these two variables are independent. For this accomplishment, x-squared test is necessary for a sufficient statistical analysis of Depression and Employability.

X^2 test, is a non-parametric test as it is a distribution free tool to analyze the independency of two nominal variables. Like all non-parametric tests, x-squared does not require the equality of the variances and the homoscedacity in the data [7]. Consequently, x-squared test can be proceeded. As shown in figure 14 (*Appendix B*), the p-value of the X^2 test is higher than 0.05 which leads to the acceptance of the null hypothesis as the statistical error of rejecting it, is significantly high. Therefore, the variables Employability and Depression are independent. Consequently, despite the common belief, this analysis concludes that unemployment does not necessarily lead to depression.

3.2 T-test

The aim of this section is testing the mean value of population suffering from depression, between males and females. For this purpose, the t-test will be performed in the data on Table 4 (*Appendix A).*

More specifically, for the t-test, the independency, the homoscedacity and the normality of the data are required [8]. The independency assumption is occurred as each registration describes only one observation. In addition, considering the central limit theorem we can also assume that the data are normally distributed, as the number of observations is over exceeding the 30. Nevertheless, the normality assumption will be further investigated with the Shapiro-Wilk test [9].

Considering the QQ plot for the number of depressed people, in figure 5, it is observed that the data are slightly abstain the normal distribution. This insight is also confirmed by the Shapiro-Wilk result in figure 15 (*Appendix C*). The p-value is significantly less than 0.05 and thus the alternative hypothesis of non-normality data is occurred. Moreover, considering the Levene's test, as shown in figure 17 (*Appendix C*), it is concluded that the assumption of homogeneity does not occur. Consequently, due to the Shapiro-wilk and Levene's results, a non-parametric analysis of the mean value of depressed people is more preferable than the parametric t-test.

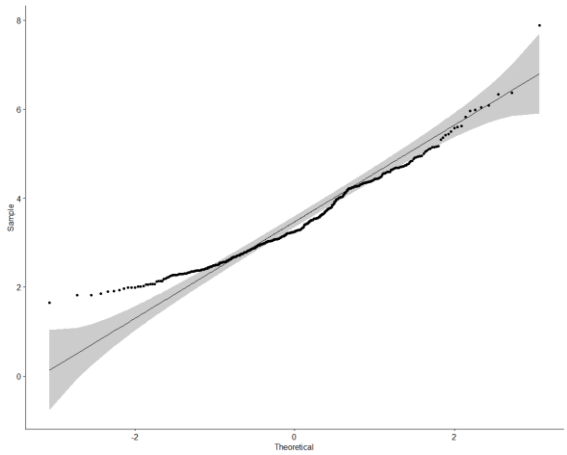


Fig. 5 QQ plot for Number of Depressed

As shown in figure 16 (*Appendix C*) and figure 6, as a first insight, it is assumed that there is a difference in the number of depressed people between males and females. More specifically, it is shown that the mean number of women with depression is larger than the men. However, this is a result considering only the specific data and thus for accurate insight about the population, further investigation is essential.

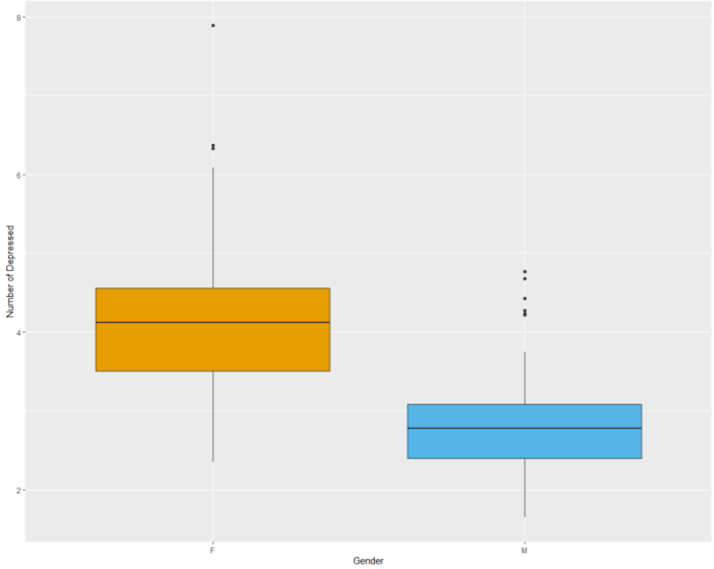


Fig. 6 Boxplot - Number of Depressed people among Gender

The non-parametric alternative of t-test is the Mann-Whitney U-test [10]. This specific test will investigate whether the percentage of population of depressed people is the same for males and females or not. The null hypothesis is that the mean value of this percentage is equal between gender and it will be tested with 95% confidence. Performing this test, in figure 18 (*Appendix C*), it is shown that the p-value is significantly less than 0.05. As a result, the statistical error of rejecting the null hypothesis is extremely low and thus the alternative one is accepted. Consequently, there is a difference in depression between males and females. In order to research deeply which gender is more likely to suffer from depression, Mann-Whitney U-test was again performed with the alternative hypothesis that mean percentage of depressed population is lower for males. As shown in figure 19 (*Appendix C*), considering the p-value, which is larger than 0.05, the null hypothesis is accepted. Consequently, the percentage of woman suffering from depression is higher than men and thus females are more likely to suffer from this mental disorder.

3.3 Anova

In this section, the percentage of population who suffer from various mental disorders is investigated. The Analysis of Variance, tests the statistically significance of the mean value of different groups [11]. More specifically, the ANOVA test will be performed to search if the mean value of the population dealing with mental health problems differs among Schizophrenia, Bipolar disorder, Anxiety and Depression, which are the most important and common disorders. In order to proceed with the parametric test, it is highly important that the data must be independent, drawn from a population with a normal distribution and have homoscedacity i.e. have equal residual variance [12]. Comparing these mental disorders, we will be able to conclude whether depression is the one that most people suffer from and generally which mental problems are the ones that plague most the society.

As shown in figure 7, it is assumed that the most common mental health disorders, are Depression and Anxiety disorder and in fact in a very close percentage. In addition, Bipolar disorder follows and finally Schizophrenia with a significant difference in the percentage of population, comparing to the other ones.

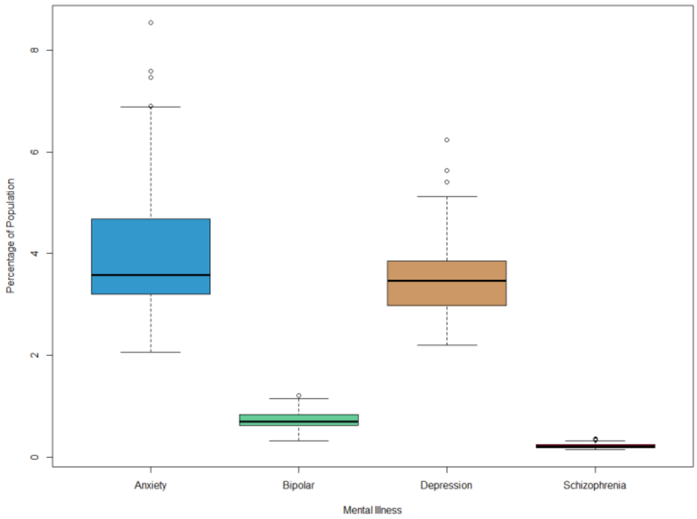


Fig. 7 Boxplot for each Mental Illness.

In order to provide an accurate statistical analysis of this, the ANOVA will be performed. Firstly, the independence of the data is valid as each registration represents one observation in this dataset. Furthermore, due to the central limit theorem, we can assume that the normality restriction is also occurred. However, for more accuracy the Shapiro-Wilk will be performed to test the normality of the data. Based on figure 20 (*Appendix D*) and figure 8, the normality hypothesis is rejected as the p-value of the Shapiro-Wilk test is significantly low.

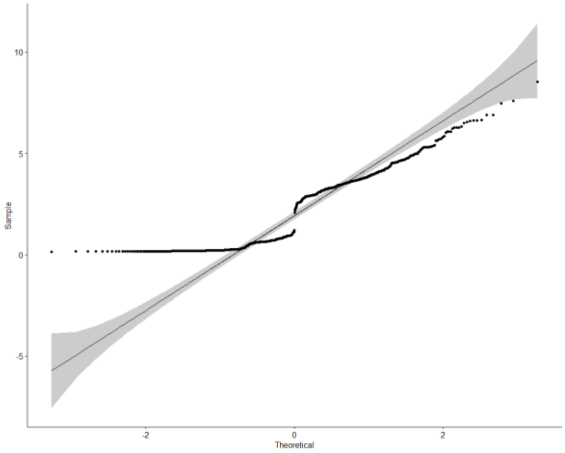
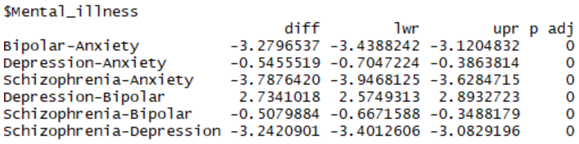


Fig. 8 QQ plot - Percentage of Population

Finally, for the investigation of the homoscedacity, the Leven's test will be performed. Executing this test, the hypothesis of the equal variance of population of each mental illness, will be checked. Based on figure 21(*Appendix D*), this hypothesis is also rejected, as the p-value is significantly less than 0.05, indicating that the variance of the population differs for each disorder. However, the ratio of the largest/smallest group is less than 1.5 and thus ANOVA can still be performed. Considering the central limit theorem, we will accept the normality of the data, as each of our group has more than 30 observations and each size of group is equal and thus all three assumptions of anova are valid. Nevertheless, because of the Shapiro-Wilk non-normality result, we will also perform the non-parametric counterpart of anova, the Kruskal-Wallis test [13].

Performing the ANOVA test, figure 22 (*Appendix D*), we observe that the p-value is significantly low, which means that the null-hypothesis of equality of the population mean, differs among the mental illnesses. As a result, we conclude that there is a difference in their appearance on the society. In order to observe exactly their difference among them, the Tukey test is performed, as shown in the figure 9. More specifically, as the purpose of this paper is the analysis of depression, we observe that this disorder is more common than Bipolar Disorder but also less common than the Anxiety Disorder.



```
$Mental_illness
                                diff        lwr        upr p adj
Bipolar-Anxiety           -3.2796537 -3.4388242 -3.1204832     0
Depression-Anxiety        -0.5455519 -0.7047224 -0.3863814     0
Schizophrenia-Anxiety     -3.7876420 -3.9468125 -3.6284715     0
Depression-Bipolar         2.7341018  2.5749313  2.8932723     0
Schizophrenia-Bipolar     -0.5079884 -0.6671588 -0.3488179     0
Schizophrenia-Depression -3.2420901 -3.4012606 -3.0829196     0
```

Fig. 9 Difference of each mean value per Mental Illness

For the non-parametric analysis, Kruskal-Wallis test will be used. As shown in figure 23 (*Appendix D*), we extract the same result as the anova. The p-value is significantly less than 0.05 which leads to the rejection of the null hypothesis. Consequently, each Mental Illness does not affect the population in the same rate.

3.4 Correlation

Considering the above insights, the statistical analysis was performed in a global scale. However, depression in not only appeared in megacities but also in non-urban areas. In this section, a data referring to the life conditions of people that live in non-urban areas, will be used. An investigation of how these conditions are correlated and the way that they affect depression, is essential for accurate insights referring to this highly important mental condition.

From technical perspective, a subset of this dataset of considering only the depressed people's life conditions, will be used to investigate the correlation among these life conditions. The most common method for this purpose is Pearson correlation. Pearson correlation number, r, shows the strength of the linear relation among the variables [14]. More specifically, it is a number between -1 and 1, with the edges describing the ultimate linear relationship.

In order, to proceed with the Pearson Collinearity test, the linearity of the data must occur. Considering the pairwise plot of figure 24 (*Appendix E*), which shows the relation among the life conditions of depressed people, it seems that there are not any variables related linearly. As a result, for the investigation of the collinearity, non-parametric method should be used. More specifically, a non-parametric alternative to Pearson, is Spearman correlation which does not require the linearity assumption [15].

Considering the Spearman Correlation number, in figure 25 (*Appendix E*), we observe that the maximum relation among variables, is the one of the gained asset and the lasting investment, which is approximately 0.35. Performing the correlation test, as shown in figure 26 (*Appendix E*), it is concluded that there is a relation between these two variables. This is because the statistical error of rejecting the null hypothesis of independence is significantly low and thus the alternative hypothesis is accepted. The Spearman correlation number and the significantly small number of the p-value (4.241e-08) also indicates that there is a relation between durable asset and lasting investment, but it is a weak one.

Comparing to the Spearman correlation of these two variables but on non-depressed people, on figure 27 (*Appendix E*), we observe that in this situation the 'ρ' is approximately 0.2, which leads to a weaker relation than in the depression case. The investment of each person and the asset of his business are highly important for a financial convenience. As a result, the difference between these two numbers was anticipated, as the economic factor has an immense impact on the people's psychology. Consequently, it was expected that the variables gained asset and lasting investment were more related on the depressed people.

4 Regression Models

4.1 Linear Regression

As examined on the correlation chapter and based on the pairwise plot given on figure 24 (*Appendix E*), it seemed that there were not any strong linear relations among the variables for depressed people. However, the strongest linear correlation based on figure 28 (*Appendix F*), is the one between durable asset and lasting investment, which is 0.42.

As shown in the scatterplot on figure 10, these two variables seem to have a weak linear relation. Nevertheless, the Pearson correlation number indicated that this is the strongest linear relationship in the data. Consequently, a linear regression model will be performed in order to predict the lasting investment for the depressed people given the durable asset as the predictor.

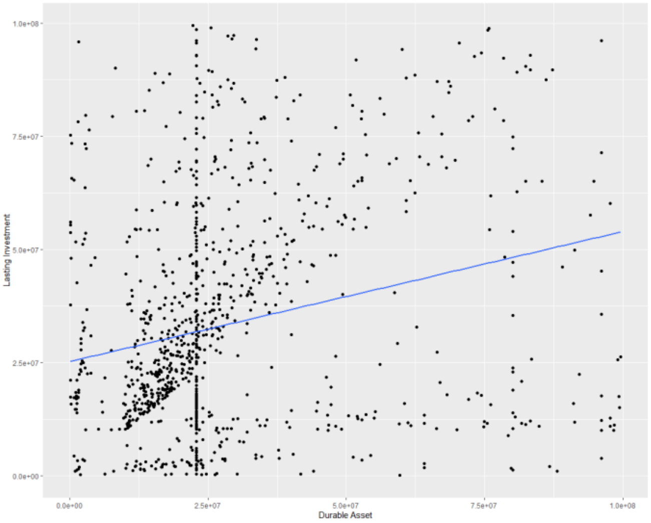


Fig. 10 Scatterplot of durable_asset and lasting_investmnet (for depressed people)

Fitting a simple linear model with durable asset as the predictor and lasting investment as the response is shown on figure 29 (*Appendix F*). More specifically, the R^2 is approximately 0.17. In this case of simple linear regression model, the R^2 is the same with the Pearson Correlation number. Consequently, a low R-squared number indicates a not so accurate prediction model, as it was expected.

Based on figure 30 (*Appendix F*), we observe the coefficients of the linear predictive model and thus the estimated lasting investment is calculated by the following equation:

lasting_investment = 1.966e+07 + 0.4737*durable_asset

However, to consider this linear regression model as valid, a variety of assumptions must occur. For this purpose, we should consider the diagnostic plots as shown in figure 11.

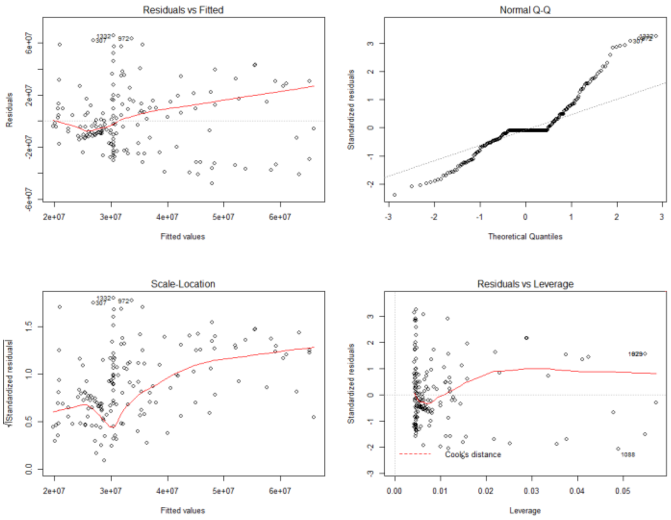


Fig. 11 Diagnostic Plots

- On the Residual vs Fitted, the line slightly represents the data. This was anticipated as the linear relationship of the variables is weak.
- On the Normal Q-Q Plot, it is shown that the data are not quite normally distributed. However, because of the central limit theorem, as the dataset contains more than 30 observations, we can assume the normality of the data.
- On the Scale-Location plot, it is indicated that there is not homoscedacity. Otherwise, for an equal variance among the data, the red line should be horizontal.
- In the Residuals vs Leverage plot, it is observed that Cook's distance lines are not shown. This results in both no cases of high Cook's distance and not influential cases in the predictive model.

4.2 Logistic Regression

Logistic regression is a widely machine learning technique for predicting a binary variable [16]. In this report, this method will be used to create a model for finding related variables to depression as well as for predicting depression on non-urban places given several binary variables as input. More specifically, the response of the model will be the variable Depressed which indicates whether a person suffers from a depression or not. In order to create an accurate prediction model, the dataset will be split into training and test sets. The model will be trained in the first set and the test data will be used on the model as the new data, so its accuracy can be computed. The predictors for this model will be the following variables:

- Sex, which indicates the gender of the participant.
- Married, indicating the married status of the participant.
- Incoming salary, which indicates if the participant has a job income.

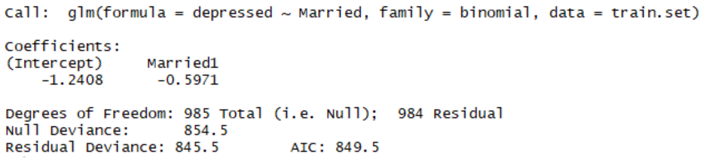
- Incoming own farm, which indicates if the participant has an income for an owned farm.
- Incoming Business, which indicates if the participant has an income for an owned business.
- Labor primary, which indicates whether the participant has a primary job or not.

All these variables are explained extensively in Table 1 (*Appendix A*).

At first the split of the dataset is required to create the model in the training data and using the test data as the predicting set, which will indicate the accuracy of the model. As shown in figure 31 (*Appendix G*), the training data will constitute a sample of the 70% of the data and the remaining 30% as the test one.

Performing logistic regression on the above variables, as shown in figure 32 (*Appendix G*), we conclude that not all the variables on non-urban areas are statistically significant for predicting depression. More specifically, the important variable (with 95% importance) is only the variable Married. Variables that referred to income were expected not to be significant, as the X^2 test performed in 3.1 indicated that depression and employment are independent.

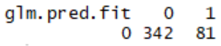
Consequently, a new logistic regression model will be created exclusively with the variable Married, as shown in figure 12.



```
Call:  glm(formula = depressed ~ Married, family = binomial, data = train.set)

Coefficients:
(Intercept)     Married1
    -1.2408      -0.5971

Degrees of Freedom: 985 Total (i.e. Null);  984 Residual
Null Deviance:      854.5
Residual Deviance: 845.5        AIC: 849.5
```

Fig. 12 Logistic Regression for Depression ~ Married

In order to compute the accuracy of this new model, the test set will be used as the prediction data. More specifically, the accuracy is the calculation of the true positive predictions divided by the total ones. The confusion matrix of the predictions is shown below, in figure 13. It is observed that the model was unable to predict depression cases. However, the accuracy of this model is $342/(342 + 81) = 0.81$, which leads to a high effective prediction model of depression.



```
glm.pred.fit   0   1
           0 342  81
```

Fig. 13 Confusion Matrix for the new logistic regression model

5 Conclusion

The purpose of this paper was to provide a variety of statistical techniques in order to investigate the variables that affect depression. From the X^2 test, it is concluded that despite the common belief, employability and depression are independent variables. As a result, unemployment does not necessarily lead to depression. Furthermore, from the Mann-Whitney U-test, it is observed that depression does not affect males and females in the same rate. More specifically, it is concluded that women are more possible to suffer from depression, as the mean percentage of females dealing with this mental illness is larger than the males one.

Moreover, it was highly important to investigate whether depression is the most common mental disease that society faces. Executing the Anova and Kruskal-Wallis tests among the mental disorders, it is concluded that each mental disorder appears in the society with a different rate. More accurately, the most common ones are Anxiety disorder and Depression, then Bipolar disorder and Schizophrenia as the least appeared illness.

Finally, life conditions of depressed people that live on non-urban areas were investigated. Collinearity test and both linear and logistic regression were performed. It was concluded that the durable asset of a depressed person is related to the lasting investment, considering the Spearman correlation number. In addition, a linear regression model was created with durable asset as the response and the lasting investment as the predictor. The R^2 of this model is 0.17, which describes a low efficient regression model, as the linear relation of these two variables is weak. Last but not least, the logistic regression was performed, indicating that Married status is highly related to depression.

Taken everything into consideration, all the above statistical analysis was performed to provide accurate and helpful insights about depression.

References

1. Christopher, J., Murray, L., Evidence Based Health Policy. *Lessons from the Global Burden of Disease Study*. Science, 274, p.740-2008 (1996).
2. Kaggle, (2019) *Depression*. [Online] Available from https://www.kaggle.com/diegobabativa/depression. [Accessed: 27th April 2021].
3. Kaggle, (2019) *Unemployment and mental illness survey*. [Online] Available from https://www.kaggle.com/michaelacorley/unemployment-and-mental-illness-survey. [Accessed: 27th April 2021].
4. Our World in Data, (2017) *Prevalence of Depression – Males vs Females*. [Online] Available from https://ourworldindata.org/grapher/prevalence-of-depression-males-vs-females?tab=table. [Accessed: 27th April 2021].
5. Our World in Data, (2017) *Prevalence by mental and substance use disorder*. [Online] Available from https://ourworldindata.org/grapher/prevalence-by-mental-and-substance-use-disorder?tab=table&country=~OWID_WRL. [Accessed: 27th April 2021].
6. Chang, H-J., "Determination of Sample Size in Using Central Limit Theorem for Weibull Distribution", *Information and Management Sciences*. 17 (3), pp.31-46 (2006).
7. McHugh, M. L., "The Chi-square test of Independence", *Lessons in Biostatistics*. Biochemia Medica, pp.143-149 (2013).
8. Kim, T. K., "T test as a parametric statistic*", Korean Journal of Anesthesiology*. 68 (6), p.540, (2015).
9. Ghasemi, A., Zahediasl, S., "Normality tests for Statistical Analysis: A Guide for non-staticians", *International Journal of Endocrinology and Metabolism.* KOWSAR. 10(2), pp.486-489, (2012).
10. McKnight, P. E., Najab, J. "Mann-Whitney U-test", *The Corsini Encyclopedia of Psychology*. (2010).
11. Kim, T. K., "Understanding one-way ANOVA using conceptual figures". *Korean Journal of Anesthesiology*. 70(1), pp.22-26, (2017).
12. Stahle, L., Wold, S., "Analysis of Variance (ANOVA)", *Chemometrics and Intelligent Laboratory Systems.* Elsevier. 6(4), pp.259-272. (1989).
13. McKnight, P. E., Najab, J. "Kruskal-Wallis test", *The Corsini Encyclopedia of Psychology*. (2010).
14. Bonesty, J., Chen, J., Huang, Y., Cohen, I., "Pearson Correlation Coefficient", *Springer Topics in Signal Processing*. Springer, pp.1-4. (2009).
15. Myers, L., Sirois, M. J. "Spearman Correlation Coefficients, Differences between", *Encyclopedia of Statistical Sciences*. Wiley Online Library. (2006)
16. LaValley, M. P. "Logistic Regression". *Circulation.* 117(18), pp.2395-2399, (2008).

Appendix A: Description of the variables

The dataset of Table 1 consists of 1,429 records with 23 variables about depression in non-urban areas.

Table 1. Description of Variables

	Variable Name	Type	Description
1.	Survey_id	Id	The Id of each Survey (unique)
2.	Ville_id	Id	The Id of each village (not unique)
3.	Sex	Categorical	The gender of the participant. 0: Male 1: Female
4.	Age	Numeric	The age of each participant.
5.	Married	Categorical	Whether the participant is married (1) or not (0)
6.	Number_children	Numeric	The amount of children of each person
7.	Educational_level	Categorical	An ordinal variable that describes the educational level, from 1 to 19
8.	Total_members	Numeric	Number of total members of each family.
9.	Gained_asset	Numeric	The gained asset of each participant.
10.	Durable_asset	Numeric	The durable asset of each participant.
11.	Save_asset	Numeric	The save asset of each participant.
12.	Living_expenses	Numeric	The living expenses of each participant.
13.	Other_expenses	Numeric	The other expenses that each participant has.
14.	Incoming_salary	Categorical	Whether the person has a salary income (1) or not (0)
15.	Incoming_own_farm	Categorical	The participant 0: Does not have income from an own farm 1: Has an income from an own farm
16.	Incoming_business	Categorical	Whether the person has a salary income from a business (1) or not (0)
17.	Incoming_no_business	Categorical	Whether the person has a non-business income (1) or not (0)

18.	Incoming_agricultural	Numeric	The agricultural income of its participant.
19.	Farming_expenses	Integer	The total amount of farming expenses.
20.	Labor_primary	Categorical	Whether the participant has a primary job (1) or not (0).
21.	Lasting_investment	Numeric	The lasting investment of each participant.
22.	No_lasting investment	Numeric	The income of no-lasting investment of each participant.
23.	Depressed	Categorical	Whether the person suffers from depression (1) or not (0).

The dataset of Table 2 is consisted of 334 records with 31 variables and refers to a questionnaire about employment and mental disorders.

Table 2. Unemployment and Mental Illness

	Variable Name	Type	Description
1.	I am currently employed at least part-time	Categorical	If the participant is employed (1) or not (0)
2.	Depression	Categorical	Whether the participant suffers from depression (1) or not (0)

The dataset of Table 3 is consisted of 47,858 records with 6 variables and refers to the percentage of population dealing with depression between males and females.

Table 3. Depression between Gender

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
2.	Code	Character	The three-letter code of each entity
3.	Year	Integer	Year of each registration
4.	Prevalence - Depressive disorders - Sex: Male - Age: Standardized (Percent)	Numeric	The male percent of population that deal with depression
5.	Prevalence - Depressive disorders - Sex: Female	Numeric	The female percent of population that deal with depression

	- Age: Standardized (Percent)		
6.	Total population (Gapminder, HYDE & UN)	Numeric	The number of the total population of each Entity

Table 4. Depression between Gender (restructured)

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
3.	Year	Integer	Year of each registration
4.	Percentage of population	Numeric	The percentage of the depressed population.
5.	Gender	Categorical	The gender of each participant: - Male - Female

The dataset of Table 5 is consisted of 6,468 records with 10 variables and refers to the percentage of population dealing with each disorder.

Table 5. Percentage of population for several Illnesses

	Variable Name	Type	Description
1.	Entity	Character	The entity of each registration (unique)
2.	Code	Character	The three-letter code of each entity
3.	Year	Integer	Year of each registration
4.	Schizophrenia (%)	Numeric	The percentage of population of each entity suffering from schizophrenia
5.	Bipolar disorder (%)	Numeric	The percentage of population of each entity suffering from bipolar disorder
6.	Eating disorders (%)	Numeric	The percentage of population of each entity suffering eating disorders
7.	Anxiety disorders (%)	Numeric	The percentage of population of each entity suffering from anxiety disorder
8.	Drug use disorders (%)	Numeric	The percentage of population of each entity suffering from drug use disorder

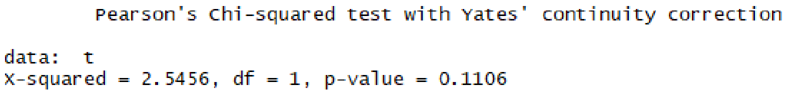
9.	Depression (%)	Numeric	The percentage of population of each entity suffering from depression
10.	Alcohol use disorders (%)	Numeric	The percentage of population of each entity suffering from alcohol use disorder

Table 6 is the transformed data of Table 5, exclusively with mental illnesses and it is restructured in order to create a categorical variable indicating the Mental Illness of each participant.

Table 6. Percentage of population exclusively for each Mental Illness (restructured)

	Variable Name	**Type**	**Description**
1.	Entity	Character	The entity of each registration (unique)
3.	Year	Integer	Year of each registration
4.	Percentage of population	Numeric	The number of the total population of each Entity
5.	Mental Illness	Categorical	The mental illness that each participant suffers: - Schizophrenia - Bipolar disorder - Anxiety disorders - Depression

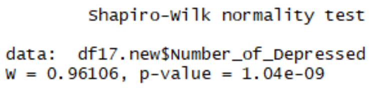
Appendix B: X-Squared



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  t
X-squared = 2.5456, df = 1, p-value = 0.1106
```

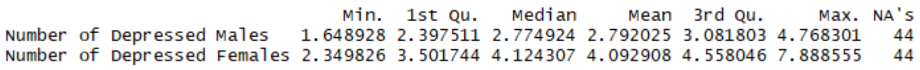
Fig. 14 X^2 for testing the independency of Depression and Employability.

Appendix C: T-test



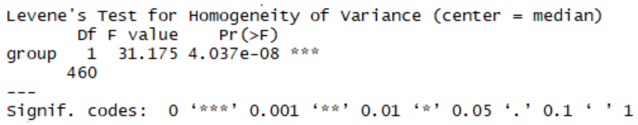
```
        Shapiro-Wilk normality test

data:  df17.new$Number_of_Depressed
W = 0.96106, p-value = 1.04e-09
```

Fig. 15 Shapiro-Wilk normality test for Number of Depressed



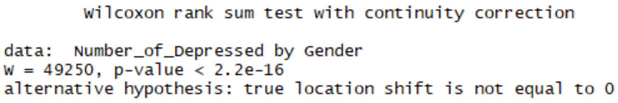
```
                              Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
Number of Depressed Males   1.648928 2.397511 2.774924 2.792025 3.081803 4.768301   44
Number of Depressed Females 2.349826 3.501744 4.124307 4.092908 4.558046 7.888555   44
```

Fig. 16 Descriptive Statistic - Number of Depressed people between Gender



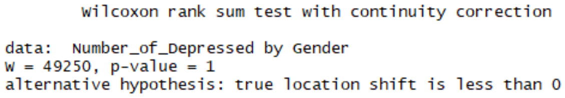
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)    
group   1  31.175 4.037e-08 ***
      460                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 17 Levene's test for the percentage of population



```
        Wilcoxon rank sum test with continuity correction

data:  Number_of_Depressed by Gender
W = 49250, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

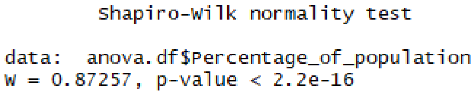
Fig. 18 Mann-Whitney U-test



```
        Wilcoxon rank sum test with continuity correction

data:  Number_of_Depressed by Gender
W = 49250, p-value = 1
alternative hypothesis: true location shift is less than 0
```

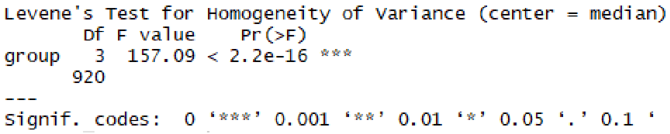
Fig. 19 Mann-Whitney U-test with alternative=less

Appendix D: Anova



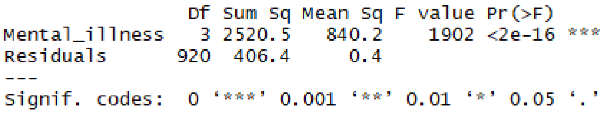
```
        Shapiro-Wilk normality test

data:  anova.df$Percentage_of_population
W = 0.87257, p-value < 2.2e-16
```

Fig. 20 Shapiro-wilk normality test for the Percentage of Population



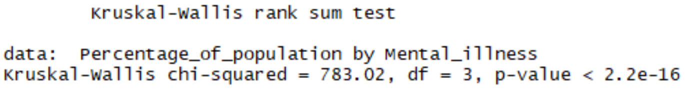
```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)    
group   3  157.09 < 2.2e-16 ***
      920                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 21 Leven’s homoscedacity test for the Population in each Mental Disorder



```
                Df Sum Sq Mean Sq F value Pr(>F)    
Mental_illness   3 2520.5   840.2    1902 <2e-16 ***
Residuals      920  406.4     0.4                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 22 Anova test for the population in each mental illness



```
        Kruskal-Wallis rank sum test

data:  Percentage_of_population by Mental_illness
Kruskal-Wallis chi-squared = 783.02, df = 3, p-value < 2.2e-16
```

Fig. 23 Kruskal-Wallis test for the population in each mental illness

Appendix E: Correlation

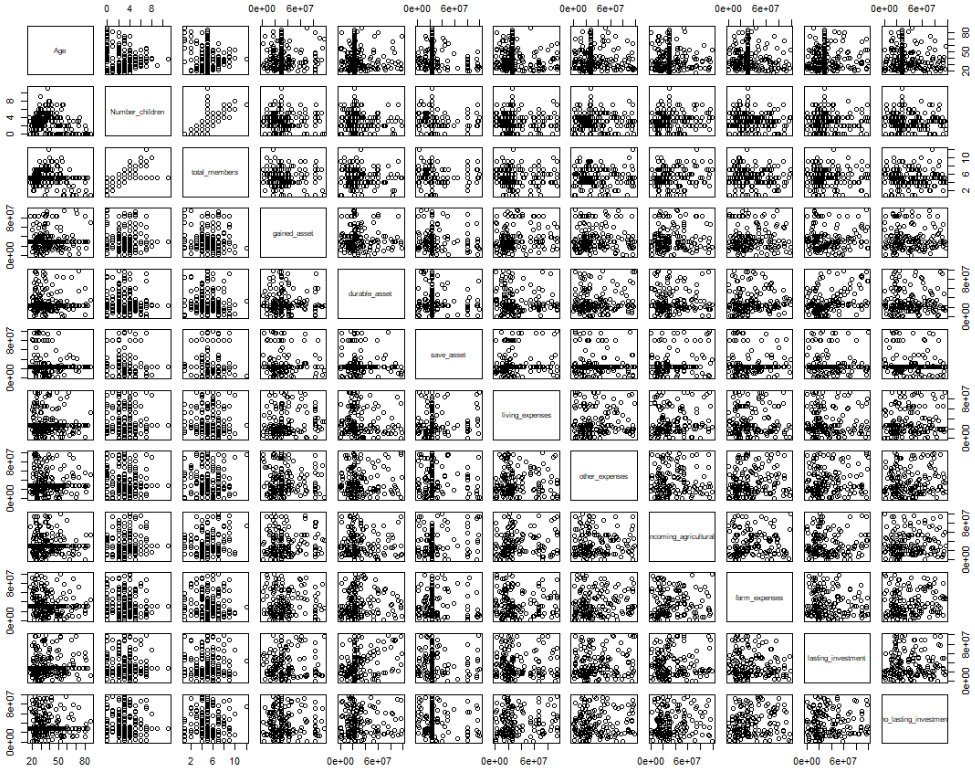
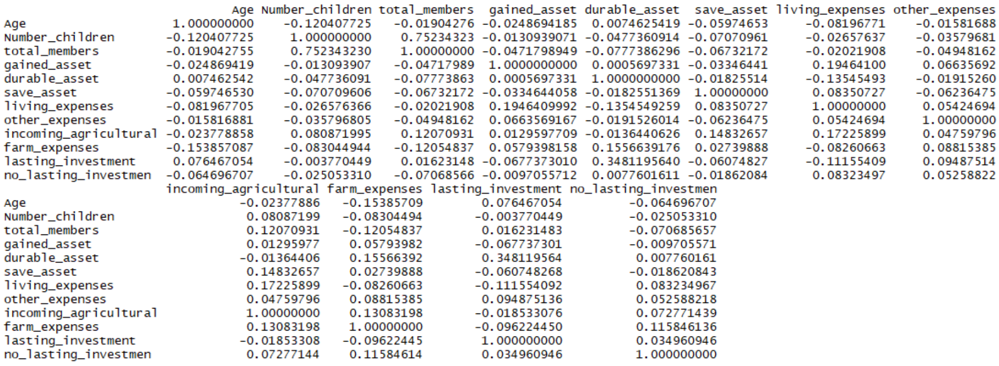
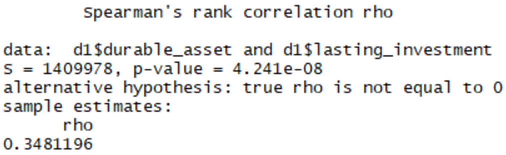


Fig. 24 Pairwised Scatterplot of all the variables of depressed people



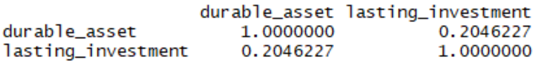
```
                              Age Number_children total_members  gained_asset durable_asset  save_asset living_expenses other_expenses
Age                    1.000000000    -0.120407725   -0.01904276 -0.0248694185  0.0074625419 -0.05974653     -0.08196771    -0.01581688
Number_children       -0.120407725     1.000000000    0.75234323 -0.0130939071 -0.0477360914 -0.07070961     -0.02657637    -0.03579681
total_members         -0.019042755     0.752343230    1.00000000 -0.0471798949 -0.0777386296 -0.06732172     -0.02021908    -0.04948162
gained_asset          -0.024869419    -0.013093907   -0.04717989  1.0000000000  0.0005697331 -0.03346441      0.19464100     0.06635692
durable_asset          0.007462542    -0.047736091   -0.07773863  0.0005697331  1.0000000000 -0.01825514     -0.13545493    -0.01915260
save_asset            -0.059746530    -0.070709606   -0.06732172 -0.0334644058 -0.0182551369  1.00000000      0.08350727    -0.06236475
living_expenses       -0.081967705    -0.026576366   -0.02021908  0.1946409992 -0.1354549259  0.08350727      1.00000000     0.05424694
other_expenses        -0.015816881    -0.035796805   -0.04948162  0.0663569167 -0.0191526014 -0.06236475      0.05424694     1.00000000
incoming_agricultural -0.023778858     0.080871995    0.12070931  0.0129597709 -0.0136440626  0.14832657      0.17225899     0.04759796
farm_expenses         -0.153857087    -0.083044944   -0.12054837  0.0579398158  0.1556639176  0.02739888     -0.08260663     0.08815385
lasting_investment     0.076467054    -0.003770449    0.01623148 -0.0677373010  0.3481195640 -0.06074827     -0.11155409     0.09487514
no_lasting_investmen  -0.064696707    -0.025053310   -0.07068566 -0.0097055712  0.0077601611 -0.01862084      0.08323497     0.05258822
                      incoming_agricultural farm_expenses lasting_investment no_lasting_investmen
Age                             -0.02377886   -0.15385709        0.076467054         -0.064696707
Number_children                  0.08087199   -0.08304494       -0.003770449         -0.025053310
total_members                    0.12070931   -0.12054837        0.016231483         -0.070685657
gained_asset                     0.01295977    0.05793982       -0.067737301         -0.009705571
durable_asset                   -0.01364406    0.15566392        0.348119564          0.007760161
save_asset                       0.14832657    0.02739888       -0.060748268         -0.018620843
living_expenses                  0.17225899   -0.08260663       -0.111554092          0.083234967
other_expenses                   0.04759796    0.08815385        0.094875136          0.052588218
incoming_agricultural            1.00000000    0.13083198       -0.018533076          0.072771439
farm_expenses                    0.13083198    1.00000000       -0.096224450          0.115846136
lasting_investment              -0.01853308   -0.09622445        1.000000000          0.034960946
no_lasting_investmen             0.07277144    0.11584614        0.034960946          1.000000000
```

Fig. 25 Spearman correlation (life conditions of depressed people)



```
        Spearman's rank correlation rho

data:  d1$durable_asset and d1$lasting_investment
S = 1409978, p-value = 4.241e-08
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3481196
```

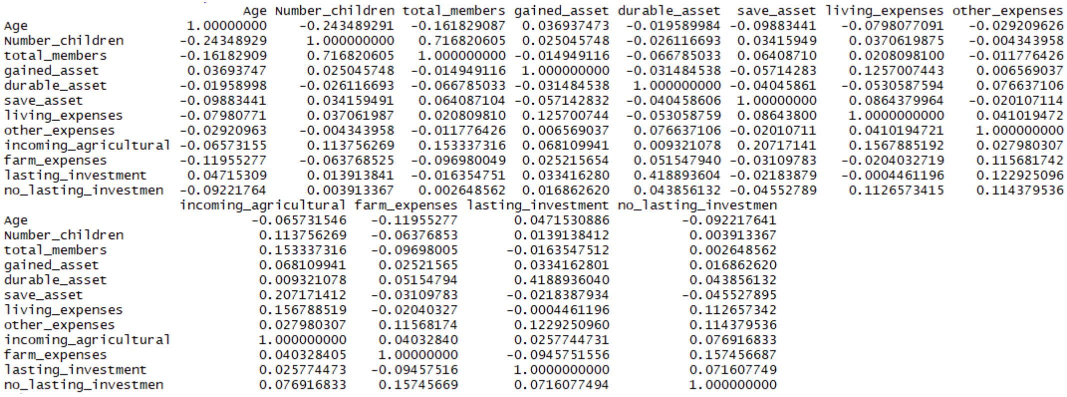
Fig. 26 Spearman Correlation test on durable asset and lasting investment



```
                   durable_asset lasting_investment
durable_asset          1.0000000          0.2046227
lasting_investment     0.2046227          1.0000000
```

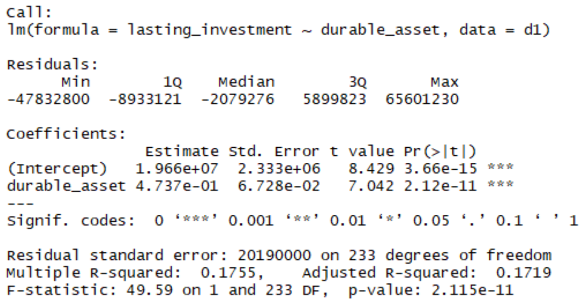
Fig. 27 Spearman Correlation on durable asset and lasting investment for non-depressed people

Appendix F: Linear Regression



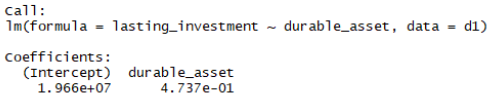
```
                             Age Number_children total_members gained_asset durable_asset  save_asset living_expenses other_expenses
Age                   1.00000000    -0.243489291  -0.161829087  0.036937473  -0.019589984 -0.09883441   -0.0798077091   -0.029209626
Number_children      -0.24348929     1.000000000   0.716820605  0.025045748  -0.026116693  0.03415949    0.0370619875   -0.004343958
total_members        -0.16182909     0.716820605   1.000000000 -0.014949116  -0.066785033  0.06408710    0.0208098100   -0.011776426
gained_asset          0.03693747     0.025045748  -0.014949116  1.000000000  -0.031484538 -0.05714283    0.1257007443    0.006569037
durable_asset        -0.01958998    -0.026116693  -0.066785033 -0.031484538   1.000000000 -0.04045861   -0.0530587594    0.076637106
save_asset           -0.09883441     0.034159491   0.064087104 -0.057142832  -0.040458606  1.00000000    0.0864379964   -0.020107114
living_expenses      -0.07980771     0.037061987   0.020809810  0.125700744  -0.053058759  0.08643800    1.0000000000    0.041019472
other_expenses       -0.02920963    -0.004343958  -0.011776426  0.006569037   0.076637106 -0.02010711    0.0410194721    1.000000000
incoming_agricultural -0.06573155    0.113756269   0.153337316  0.068109941   0.009321078  0.20717141    0.1567885192    0.027980307
farm_expenses        -0.11955277    -0.063768525  -0.096980049  0.025215654   0.051547940 -0.03109783   -0.0204032719    0.115681742
lasting_investment    0.04715309     0.013913841  -0.016354751  0.033416280   0.418893604 -0.02183879   -0.0004461196    0.122925096
no_lasting_investmen -0.09221764     0.003913367   0.002648562  0.016862620   0.043856132 -0.04552789    0.1126573415    0.114379536
                     incoming_agricultural farm_expenses lasting_investment no_lasting_investmen
Age                            -0.065731546   -0.11955277       0.0471530886         -0.092217641
Number_children                 0.113756269   -0.06376853       0.0139138412          0.003913367
total_members                   0.153337316   -0.09698005      -0.0163547512          0.002648562
gained_asset                    0.068109941    0.02521565       0.0334162801          0.016862620
durable_asset                   0.009321078    0.05154794       0.4188936040          0.043856132
save_asset                      0.207171412   -0.03109783      -0.0218387934         -0.045527895
living_expenses                 0.156788519   -0.02040327      -0.0004461196          0.112657342
other_expenses                  0.027980307    0.11568174       0.1229250960          0.114379536
incoming_agricultural           1.000000000    0.04032840       0.0257744731          0.076916833
farm_expenses                   0.040328405    1.00000000      -0.0945751556          0.157456687
lasting_investment              0.025774473   -0.09457516       1.0000000000          0.071607749
no_lasting_investmen            0.076916833    0.15745669       0.0716077494          1.000000000
```

Fig. 28 Pearson Correlation Number



```
Call:
lm(formula = lasting_investment ~ durable_asset, data = d1)

Residuals:
      Min        1Q    Median        3Q       Max 
-47832800  -8933121  -2079276   5899823  65601230 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.966e+07  2.333e+06   8.429 3.66e-15 ***
durable_asset 4.737e-01  6.728e-02   7.042 2.12e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20190000 on 233 degrees of freedom
Multiple R-squared:  0.1755,	Adjusted R-squared:  0.1719 
F-statistic: 49.59 on 1 and 233 DF,  p-value: 2.115e-11
```

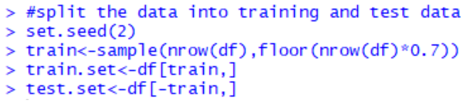
Fig. 29 Linear Regression model



```
Call:
lm(formula = lasting_investment ~ durable_asset, data = d1)

Coefficients:
  (Intercept)  durable_asset  
    1.966e+07      4.737e-01  
```

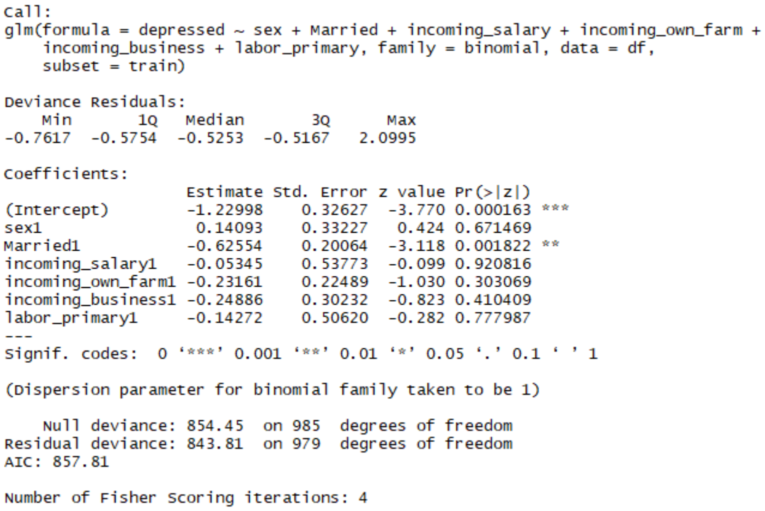
Fig. 30 Coefficients of the Linear Regression Model

Appendix G: Logistic Regression



```
> #split the data into training and test data
> set.seed(2)
> train<-sample(nrow(df),floor(nrow(df)*0.7))
> train.set<-df[train,]
> test.set<-df[-train,]
```

Fig. 31 Split the data into training and test set.



```
Call:
glm(formula = depressed ~ sex + Married + incoming_salary + incoming_own_farm +
    incoming_business + labor_primary, family = binomial, data = df,
    subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7617  -0.5754  -0.5253  -0.5167   2.0995

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.22998    0.32627  -3.770 0.000163 ***
sex1                0.14093    0.33227   0.424 0.671469
Married1           -0.62554    0.20064  -3.118 0.001822 **
incoming_salary1   -0.05345    0.53773  -0.099 0.920816
incoming_own_farm1 -0.23161    0.22489  -1.030 0.303069
incoming_business1 -0.24886    0.30232  -0.823 0.410409
labor_primary1     -0.14272    0.50620  -0.282 0.777987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 854.45  on 985  degrees of freedom
Residual deviance: 843.81  on 979  degrees of freedom
AIC: 857.81

Number of Fisher Scoring iterations: 4
```

Fig. 32 Logistic regression for depression.