Statistics Project: Mental Health Disorders

Isotta Magistrali Gabriele Mura

9th January 2022

1. Introduction

Over the last few years “mental health” related topics have increasingly started to come to our ears, maybe because of the pandemic, which may have helped develop mental disorders due to the fact that people were forced to stay at home, or simply because talking about these kind of delicate topics is becoming less scary and more normalized.
What we are sure of is that a great share of the population, especially adults, are diminishing this type of illnesses talking about them just as a whim of young people and not treating them as a serious issue.
Therefore, we thought it could be very interesting for us, and we hope for the reader as well, to understand if there was any correlation between the share of the population presenting mental health disorders in different countries and some social, economical or political characteristics of the country itself.
In the end, what we are trying to demonstrate is that the higher the well-being of a particular country is, the higher the share of cases of mental health disorder is.
This could lead to some interesting debates about the fact that it might be easier to develop a mental illness in situations where you don’t have to worry about external threats and day-to-day survival.
Our research is done just with a statistical perspective; thus, another question we decided to ask ourself is, indeed, whether it is possible that Covid played a role in increasing the number of cases or not.
In the end, we will make some comments and considerations about our work and suggest possible topics which could be further looked into.

2. Dataset

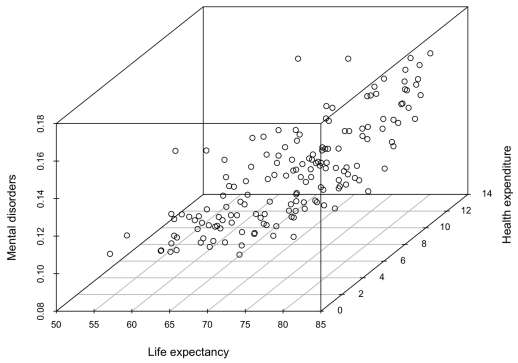
We firstly looked for data relative to the number of people presenting a mental health disorder in different countries and we reported them as a share of the total population in the countries in order to make them uniform. Afterwards, we chose some factors which, all together, we thought could describe the well-being of individuals and development of the country in an exhaustive way. We chose the year 2019 since it was the most recent year with available data on so many different countries for all our desired variables. We merged all the data in a single CSV file whose column names' explanation is the following:

Mental disorders: share of population with a mental disorder (depressive disorders, anxiety disorders, bipolar disorders, eating disorders, schizophrenia, attention-deficit/hyperactivity disorders, conduct disorders, developmental intellectual disorders, autism spectrum disorders)

GDP: GDP pro capita measured in constant 2015 US $
Life expectancy: life expectancy in years
Suicides: number of suicides per 100.000 people
Urbanization: share of population living in an urban area
Regime: 0 = not free, 1 = partially free, 2 = free (freedom in the sense of political rights and civil liberties)
Education: average number of years of schooling
Health expenditure: share of GDP allocated to healthcare
Armed conflicts: 1 = conflict, 0 = no conflict (ongoing conflicts with more than 50 casualties in 2019)
Substance use disorders: share of population with a substance use disorder (alcoholism, illicit drug dependence)

Our initial dataset is composed by 175 countries, as one can see in the attached CSV file (“Initial_dataset”).
Then, in order to visually get an idea of possible correlations between our variables, we started to plot our dependent variable against each single cofactor (see Appendix). We also chose to do that because we wanted to try to formulate more accurate hypothesis about the possible results of our study.
From the plots, we can already understand some sort of positive correlation with the variables relative to life expectancy, health expenditure and education, which seems to follow our initial idea of finding more cases in

more favorable conditions. Furthermore, the mean of the share of mental health disorders seems to be higher in situations of free political regime than the ones in which there is no freedom, while we don't see such an evident change in the war case. For what concerns GDP, it is difficult to get an idea just from the graphical visualization, because there are some countries with a consistently higher GDP than the other ones, which causes a little overlap between countries on the left. Moreover, grasping the general trend of the plots, we see that the variance is quite high, making it probably difficult to have it explained from the regression at a high percentage, but surely these are just the first and only ideas one could get.
Since the presence of outliers could have altered the results of our regression, we also realized some interactive plots to easily understand which countries deviated a lot from the regular direction of the points, and we decided to remove them. Thanks to the removal of these outliers we got our final dataset, with 160 countries ("Final_dataset").
Finally, we made one last 3D plot to get the idea of what we could get from the multivariate regression we had in mind: we clearly had to select a simpler model than the full one in order to be able to visualize it, so we only chose the two variables which, at first sight, seemed to be the most significant and, indeed, we can see a positive correlation.



3. Multivariate Linear Regression

As we already presented in the introduction, the purpose of our study is to try to explain the variable relative to the share of mental health disorders in a country through some potentially explanatory variable, which are our cofactors. Therefore, we decided to perform a multivariate linear regression (for more precise details, you can always refer to the attached R script).
We firstly chose the whole model, which is the one with all the 9 covariates. As we can see from the R output, some of the variables seem to be significant, since their p-value is below the 0.05 significance level. On the other side, as we expected, the values of the R-squared and adjusted R-squared are not so high, in particular they are respectively around 56% and 53%, which means that this percentage of the variance is explained. This is not optimal per se, but we will discuss it further later on.
Afterwards, we decided it would be good to dig deeper in this first regression, trying to apply some selection methods to find a possibly better explanatory model.

3.1. Model Selection

Given the fact that sometimes smaller models can result in better predictions and can be better adjusted to additional data, we chose to perform the following model selection methods in order to compare the results we would obtain and choose the one which, for us, was the overall best.
What we did is, firstly, both the step-down and step-up methods, hoping to obtain a smaller and at the same time better model; afterwards, a penalty method, to see how bigger and smaller models compare to each other.
The step-up and step-down methods seem to agree in the model selected since, in the end, the result is the same, namely a model with the following four variables: GDP, life expectancy, health expenditure and armed conflicts.
As one can see, all the cofactors are significant at the 0.05 level, with a positive correlation for all of them. What is also interesting is that this result contradicts the hypothesis about armed conflicts we stated initially

just looking at the plots, because this variable seems to be, in fact, significant in the prediction of mental disorder cases.
On the other side, the R-squared value is lower than the one of the whole model, which accounts for a smaller percentage of the variance explained.
In order to explore a further possibility, we had a brief discussion about which penalty method to choose: from what we studied in class both the AIC and BIC seemed to be accurate, but since we are interested in the explanation of our dependent variables and we do not really care about the selection of the best covariates, we decided to apply the Akaike Information Criterion (AIC).
This resulted in a bigger model than the one we obtained from the step-wise methods: indeed, this last model included 7 cofactors against the 4 of the first two selections.
This last model has the biggest percentage of variance explained and the lowest AIC value. Nonetheless, computing on R the AIC value of our step-wise model, we realize it is not significantly higher and it is also lower than the one of the whole initial model (again, see the R script or the Appendix with the R output).
Furthermore, what the step-down and step-up methods tell us, is that when trying to consider a bigger model than the one with the 4 covariates, the null hypothesis is retained (in the former case separately in the bigger model and in the latter separately and one at a time); this means that, when considering a single additional covariate, we don't have enough statistical evidence of them being different from zero. Thus, from the perspective of the step-wise methods, the 3 additional covariates of the AIC model are potentially not useful.
This, together with the fact that smaller models are always preferred when talking about adaptation and manageability, we decided to keep the 4-variables model.

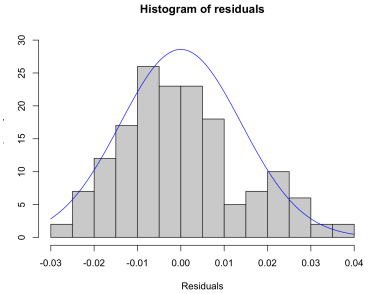
```
Call:
lm(formula = Mental.disorder ~ GDP + Life.expectancy + Health.expenditure +
    Armed.conflicts, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.025314 -0.009471 -0.001861  0.007295  0.036083

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.357e-02  1.427e-02   1.652 0.100597
GDP                3.204e-07  9.129e-08   3.510 0.000587 ***
Life.expectancy    1.194e-03  2.074e-04   5.759 4.42e-08 ***
Health.expenditure 1.332e-03  5.151e-04   2.586 0.010627 *
Armed.conflicts    9.990e-03  2.967e-03   3.367 0.000958 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01413 on 155 degrees of freedom
Multiple R-squared:  0.5228,    Adjusted R-squared:  0.5105
F-statistic: 42.46 on 4 and 155 DF,  p-value: < 2.2e-16
```

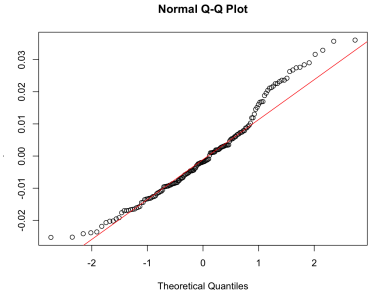
Step-wise methods model

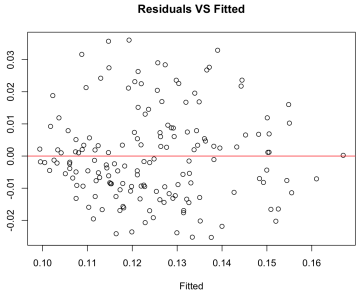
```
Step:  AIC=-1361.77
Mental.disorder ~ GDP + Life.expectancy + Urbanization + factor(Regime) +
    Education + Health.expenditure + factor(Armed.conflicts)

                          Df Sum of Sq      RSS     AIC
<none>                                 0.028771 -1361.8
- Urbanization             1 0.0006523 0.029423 -1360.2
- Education                1 0.0007411 0.029512 -1359.7
- factor(Regime)           2 0.0012744 0.030046 -1358.8
- Health.expenditure       1 0.0013852 0.030156 -1356.2
- GDP                      1 0.0014075 0.030179 -1356.1
- factor(Armed.conflicts)  1 0.0016709 0.030442 -1354.7
- Life.expectancy          1 0.0045711 0.033342 -1340.2
```

AIC model

Finally, we performed some graphical and statistical tests to check the homoscedasticity and normality assumptions on the residuals.
For the normality part, we decided to use the Kolmogorov-Smirnov and Shapiro-Wilk tests, the last one justified from the fact that we have quite a big sample size. From the Kolmogorov-Smirnov test we get a p-value of around 0.2, which seems to be quite good for our purpose since it is greater than the 0.05 level of significance, allowing us to retain the null hypothesis of the distribution being equal to the normal one.
From the Shapiro-Wilk test, instead, we get a really small p-value, but this could be due to the fact that having a great number of data points, and being the Shapiro-Wilk test more accurate the bigger the sample size is, even the smallest deviation from the chosen distribution is detected.
Our normality assumption can, indeed, be justified from the graphical methods: the histogram of the residuals follows quite well the bell-shaped curve of the normal distribution and in the QQ-plot the residuals have a trend that overall follows the straight red line.
Finally, regarding the homoscedasticity, we can see from the plot of residuals VS fitted that the constant variance assumption is not evidently violated.







4. Is Covid significant?

Now we move on to the last question we decided to ask ourselves, namely the one regarding the pandemic. Being forced to stay at home, and in general the huge threat we had to face in the past two years, caused a lot of speculations about the difficulties it might have caused at a relational and personal level. We decided to dig deeper into this matter, restricting our focus to the United States, since it was the only country for which we were able to find data also about more recent years.
The dataset ("Pre_post_covid_dataset") contains the 50 states (plus the District of Columbia) of the USA, along with the data relative to the share of adults with any mental illness (AMI) in 2016, 2019, 2022.
We decided to perform two different t-tests: one with the samples about the years 2019 and 2022 and the other with the ones regarding 2016 and 2019. Our first choice was determined by the fact that we wanted to choose the most recent year before Covid, and then the first year afterwards where things had already had time to settle a little bit; specularly, we chose the year 2016.
What we are trying to prove is that the increase in the number of cases between 2019 and 2022 is significant, while the one between 2016 and 2019 is not: this would be a very interesting result, which we are going to discuss further afterwards.
(From now on: H_0: null hypothesis; H_1: alternative hypothesis)
In the first case we consider as our H_0 that the mean of cases regarding 2019 is bigger than the one in 2022; if H_0 gets rejected through the test, we will obtain significant statistical evidence supporting H_1: the mean of the number of cases increased after the pandemics.
In the same way, in the 2016-2019 case, we consider as H_0 that the mean of cases in 2019 is lower than the one in 2016: if H_0 is retained it means we don't have statistically significant evidence to support H_1: the mean of cases is higher than three years before.
We computed the variances in the three cases and, although the two relative to 2016 and 2019 don't differ much, the ones relative to the other two years seem to be quite different. This, along with the fact that our sample size is quite big, brought us to choose the asymptotic t-test.
As we suspected, in the 2019-2022 case we get a test which is greater than the quantile of the normal, at a 0.95 confidence level, allowing us to reject the null hypothesis and to conclude that the mean of cases in 2022 is higher than the one in 2019 (actually, this result is significant at an even higher confidence level, really close to 1).
What is quite surprising is that, conversely, in the other case the test is lower than the bound given by the 0.95 quantile of the normal, which doesn't allow us to reject the null: we do not have enough statistical evidence of the fact that the variation in the sample mean is significant. Even exchanging the two hypothesis, thus trying to prove that the mean of 2016 is higher that the one in 2019, doesn't provide us with enough statistical evidence to reject the null. Actually, if you try to change the level of significance, you can see that it is more diminishing than increasing, in the sense that, in order to reject H_0, in the diminishing case you get a confidence level of 0.69, whereas in the increasing one you would get a confidence level of 0.30, which is way lower.
Anyway, what this means in general is that the trend between 2016 and 2019 is not significantly changing from a statistical point of view, while the trend between 2019 and 2022 significantly increased.

5. Conclusions, limitations and further research questions

We wanted to conclude by saying a few words about the results we obtained and suggesting ways to make them better or further explore the topic.

As we said before, we obtained statistical evidence of the fact that that there exist a positive correlation between our factors and the dependent variable. In particular, our model composed by the cofactors regarding life expectancy, GDP, presence of armed conflicts and health expenditure seems to be the one which better describes and adapts to our datapoints.
What this can show is that that a higher safety and well-being of the country, despite providing a secure environment for what concerns the physical conditions of an individual, can result in a higher risk of developing mental health illnesses. The main limitation of our study is that, as we can see in the R code output, the variance explained is only around 1/2, which means that our datapoints are not all well explained by the regression.

An idea which could be a potential solution to try to improve this result, is to perform other regression models, like ANOVA, on different subsets of covariates: this way, one could try to understand if considering some kind of interaction effect between the parameters could result in a better explanatory model.
It is also true that any medical condition, and even more if it is a mental disorder, can be related to an infinite variety of factors, which can be personal or social, given by inner thoughts or by one's surroundings' conditions, so it is very difficult to get a purely statistical explanation which can be considered clear and exhaustive.
Furthermore, another difficulty regarding medical conditions of the population, which can also be considered as a limitation of our dataset, is that it is always difficult to collect specific and accountable data, especially in countries where awareness and/or screening methods are more limited.
Nonetheless, what we think could be interesting for further research, is to repeat our work on other years before 2019 and try to see whether the result changes, how it changes, or if one can find a stronger or weaker correlation, to get an idea of how this phenomenon is evolving throughout time. This way, one could try to understand whether some other factors could have had an impact, like the growth of social media.
Another thing which we found quite surprising is that suicides and substance use disorders don't seem to be significantly correlated to mental disorders. A possible explanation could be that we are taking into consideration a wide range of mental illnesses all together, while those kind of issues could be related to some specific mental health disorders more than others, as well as impact some age-ranges or gender more than others. Thus, it would be interesting to study them separately, in order to be more accurate and precise at selecting our predictive variables.

Moving on to the part regarding the pre- and post-covid situations, again we can just see a statistical demonstration of the fact that the mean of the number of cases is significantly higher in 2022 with respect to 2019, which cannot be said for the years 2019 and 2016.
Actually, we have no proof of the fact that Covid itself boosted the number of people with mental health disorders, since the amount of things that changed between the two periods of time cannot be quantified, neither can they be analyzed separately.
However, we can make a further assumption: we are actually studying an overall amount of years which is relatively small, so we can consider a lot of external factors to remain constant in time and focus ourselves on bigger and more impactful events as the only variables which could have made a significant difference, like, indeed, the pandemic and its effects at a political, economical and social level.
Again, additional research could be made, going for example further back in time, or expanding the area of research outside the USA.

In conclusion, we just wanted to say that our whole project purely shows statistical correlation and doesn't imply any form of causation, but one thing can be observed from this: if one could argue that mental disorders are just an invention of young people, maybe something just for them to be listened to or to be noticed, surely, what cannot be said is that these phenomena are completely uncorrelated to the environment that surrounds us, or that they can be disconnected from the outline of a country itself; so they are indeed, and consequently should be considered as, a global and serious issue, even more than the majority of people seems to be actually considering them.

Appendix

References

Datasets:
https://ourworldindata.org/grapher/share-with-mental-and-substance-disorders?tab=table&time=latest

https://ourworldindata.org/grapher/total-healthcare-expenditure-gdp?tab=table&time=latest

https://ourworldindata.org/global-education#years-of-schooling

https://ourworldindata.org/life-expectancy

https://ourworldindata.org/grapher/political-regime-fh?time=2019&country=ARG~AUS~BWA~CHN

https://ourworldindata.org/urbanization#share-of-populations-living-in-urban-areas

https://ourworldindata.org/suicide

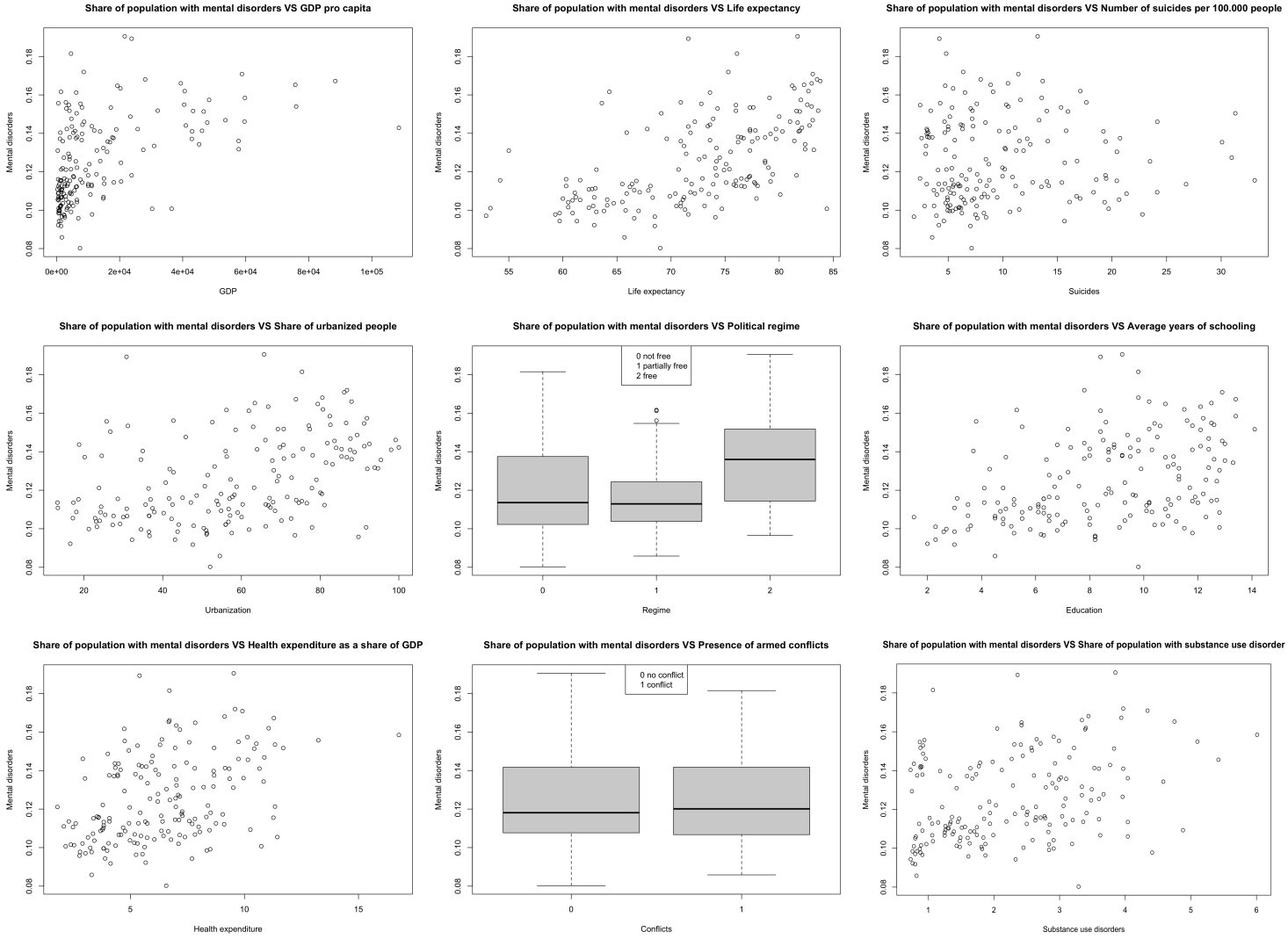
https://en.wikipedia.org/wiki/List_of_armed_conflicts_in_2019

https://ourworldindata.org/drug-use#prevalence-of-substance-use-disorders

https://ourworldindata.org/grapher/national-gdp?tab=table

https://mhanational.org/issues/state-mental-health-america

Initial plotting



R-output: regression with all factors

```
Call:
lm(formula = Mental.disorder ~ GDP + Life.expectancy + Suicides +
    Urbanization + factor(Regime) + Education + Health.expenditure +
    factor(Armed.conflicts) + Substance.use.disorders, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.024853 -0.008152 -0.000773  0.006115  0.037571

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.468e-02  1.614e-02   1.530  0.12825
GDP                       2.574e-07  9.750e-08   2.640  0.00919 **
Life.expectancy           1.249e-03  2.688e-04   4.646 7.39e-06 ***
Suicides                 -5.627e-05  2.109e-04  -0.267  0.79000
Urbanization              1.168e-04  6.568e-05   1.779  0.07735 .
factor(Regime)1          -5.829e-03  3.019e-03  -1.931  0.05543 .
factor(Regime)2           1.223e-03  3.384e-03   0.362  0.71823
Education                -1.095e-03  6.813e-04  -1.607  0.11013
Health.expenditure        1.465e-03  5.615e-04   2.609  0.01001 *
factor(Armed.conflicts)1  8.762e-03  3.065e-03   2.859  0.00486 **
Substance.use.disorders  -1.460e-04  1.497e-03  -0.098  0.92244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01389 on 149 degrees of freedom
Multiple R-squared:  0.557,     Adjusted R-squared:  0.5272
F-statistic: 18.73 on 10 and 149 DF,  p-value: < 2.2e-16
```

R-output: step-wise regression model

```
Call:
lm(formula = Mental.disorder ~ GDP + Life.expectancy + Health.expenditure +
    Armed.conflicts, data = data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.025314 -0.009471 -0.001861  0.007295  0.036083

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.357e-02  1.427e-02   1.652 0.100597
GDP                3.204e-07  9.129e-08   3.510 0.000587 ***
Life.expectancy    1.194e-03  2.074e-04   5.759 4.42e-08 ***
Health.expenditure 1.332e-03  5.151e-04   2.586 0.010627 *
Armed.conflicts    9.990e-03  2.967e-03   3.367 0.000958 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01413 on 155 degrees of freedom
Multiple R-squared:  0.5228,    Adjusted R-squared:  0.5105
F-statistic: 42.46 on 4 and 155 DF,  p-value: < 2.2e-16
```

R-output: AIC regression model

```
Step:  AIC=-1361.77
Mental.disorder ~ GDP + Life.expectancy + Urbanization + factor(Regime) +
    Education + Health.expenditure + factor(Armed.conflicts)

                          Df Sum of Sq      RSS     AIC
<none>                                0.028771 -1361.8
- Urbanization             1 0.0006523 0.029423 -1360.2
- Education                1 0.0007411 0.029512 -1359.7
- factor(Regime)           2 0.0012744 0.030046 -1358.8
- Health.expenditure       1 0.0013852 0.030156 -1356.2
- GDP                      1 0.0014075 0.030179 -1356.1
- factor(Armed.conflicts)  1 0.0016709 0.030442 -1354.7
- Life.expectancy          1 0.0045711 0.033342 -1340.2
```

Residuals check of the final 4-variables model