# Prediction of Life Expectancy From Stress Level Using Lifestyle Survey Data

Huiyan(Livia) Li

December 22, 2020

## Abstract

According to the World Health Organization, mental health issues have become an emerging health crisis worldwide over the last decade (Volkov, 2020). In the first part of this study, one of the significant evaluation criteria of mental wellness - "stress level" will be predicted by utilizing a work-life balance survey data obtained from Kaggle and a multiple linear regression model(MLR). The data contains 12,757 survey responses with 23 attributes describing people's style of living. The MLR model's validity was assessed by penalized-likelihood criteria such as AIC and BIC. In the second part of this study, a propensity score matching technique was adopted to simulate the effects of stress level on life expectancy. This study aims to identify and understand some of the critical factors of stress levels, then try to discover some practical lifestyle practice advice to help reduce stress levels, improve mental health, and ultimately increase life expectancy.

Keywords: MLR, AIC, BIC, Propensity Score, Causal Inference, Observational Study, Mental Health, Stress, Life Expectancy.

## Introduction

According to the World Health Organization, mental illness is increasing at an alarming rate in the past decade all over the world and has had a significant impact on all aspects of patients' daily life, from school or work performance to relationships with others (Volkov, 2020). Stress and its deduced anxiety disorder have been considered one of the most common mental health problems (Arthur, 2005). For many years, researchers have been trying to develop models for various aspects of stress (Avison & Gotlib, 1994), which is crucial because we need to understand the factors that cause stress to resolve the issue and thus decrease mental illness cases.

Although the experimental method(randomized control trial) is often considered the ideal method for evaluating cause and effect relationship (Black, 1996), however, there are certain drawbacks of this method as well, for example, ethical issues in doing experiments, real life funding issues, technical issues etc. Therefore, more attention starts to fall on using observational data to make causal inferences between randomly allocated treatment and control groups through propensity scores (Austin, 2011). Although there are some drawbacks with the propensity scores method, it is a relatively safer method compared with other methods using observational data as it eliminates the confounding issue of observational data(Austin, 2011).

A simple linear regression model is no longer suitable when facing with more than one independent predictor variable for a single dependent variable. In contrast, the multiple linear regression model was invented for such a situation (Nathans et al.,2012). Multiple Linear Regression (MLR) model could help find some possible important predictor factors for a specific outcome, such as the stress level in this paper. However, the Multiple Linear Regression Model might face some issues when there are many independent variables, and these variables are intercorrelated, which would lead to "Multicollinearity" (Nathans et al.,2012). To check the validity of our MLR model selection, two primary penalized model selection criteria: the Bayesian information criterion (BIC) and Akaike's information criterion (AIC), are introduced, and some useful information could be deduced to form an ultimate model(Kuha, 2004).

There are two phrases/research questions I plan to tackle in this study. In the first phase of this study, I would like to investigate how personal lifestyle influences the number of daily stress people felt. I would first build a Multiple Linear Regression Model based on a work-life balance survey data obtained from Kaggle (Dalat,2020) that consists of 12,757 survey responses data with 23 variables describing people's style of living, including individually reported stress level. Then I will use penalized-likelihood criteria such as AIC and BIC to evaluate the MLR model's validity. In phase two of this paper, I am interested in testing if there is a causal relationship between a person's stress level and his/her life expectancy; a propensity score matching technique was adopted in this study.

# Methodology

## DATA

The data used in this study is extracted from a work-life balance survey data obtained from Kaggle (Dalat,2020) that consists of 12,757 survey responses data with 23 variables describing people's style of living, which includes: self-reported daily level of stress, number of daily fruits and veggies intake, number of places visited past year, number of close friends, number of times supporting others, number of people interacted every day, number of achievement in the past year, number of donation times, BMI range, number of todos completed, number of times feeling "flow" daily, number of daily steps, number of years ahead with a clear vision, daily sleep hours, number of days of vacations lost in a year, number of times shouting at someone per day, having sufficient income or not, number of awards received in life, number of hours spent for passion daily, number of meditation practice daily, age and gender.

In Phrase one of the study, all 12,757 survey responses data and all 23 of the variables listed above were adopted. (i.e. data frame 1)

In Phrase two of the study, all 12,757 survey responses data, but only seven variables including the self-reported daily level of stress, daily sleep hours, number of days of vacations lost in a year, number of times shouting at someone per day, having sufficient income or not, number of meditation practice daily, and gender (i.e. data frame 2). Please see Table 1 (Yoshida, 2020) below:

```
##
##                                          Overall
##   n                                      12756
##   DAILY_STRESS = medium to high stress (%)  7498 (58.8)
##   SLEEP_HOURS (mean (SD))                 7.04 (1.20)
##   LOST_VACATION (mean (SD))               2.83 (3.67)
##   DAILY_SHOUTING (mean (SD))              2.92 (2.69)
##   SUFFICIENT_INCOME (mean (SD))           1.73 (0.44)
##   DAILY_MEDITATION (mean (SD))            6.25 (3.03)
##   GENDER = Male (%)                       5042 (39.5)
```

# Model

## Model1 - MLR Model

### Model Specifics

The full MLR model is as follows: $\hat{DailyStress} = 3.908267 - 0.011526 FruitsVeggies - 0.011306 PlacesVisited - 0.027389 CoreCircle + 0.019954 SupportingOthers + 0.027164 SocialNetwork - 0.014529 Achievement + 0.006631 Donation + 0.100439 BMIRange - 0.026703 TodoCompleted - 0.032269 Flow + 0.007280 DailySteps - 0.015418 LiveVision - 0.088685 SleepHours + 0.050166 LostVacations + 0.123530 DailyShouting - 0.262382 SufficientIncome + 0.020776 PersonalAwards - 0.028141 TimeforPassion - 0.048747 DailyMeditation - 0.016778 Age36to50 - 0.040489 Age51orMore - 0.110432 AgeLessthan20 - 0.311667 GenderMale$. Where the intercept is 3.908267, showing when all other $\beta$ equals to 0, the amount of daily stress a person in our

dataset would feel. Taking one of the most significant variable which has the lowest p value (<0.05) and largest t-value - Daily Shouting as example, the $\beta_{dailyshouting} = 0.123530$, meaning that when the number of times shouting to someone per day changes by 1 unit, the corresponding average change in the stress level is 0.123530 unit holds all other variables constant.

#### Backward Elimination with AIC

Here, I use backward Elimination with AIC to search for a better model. The final fitted model selected by this method is as follows: $\hat{DailyStress} = 3.900678 - 0.011660PlacesVisited - 0.027259CoreCircle + 0.020644SupportingOthers + 0.027147SocialNetwork - 0.014372Achievement + 0.103920BMIRange - 0.027076TodoCompleted - 0.032321Flow + 0.006423DailySteps - 0.015298LiveVision - 0.089654SleepHours + 0.050243LostVacations + 0.123543DailyShouting - 0.261218SufficientIncome + 0.020456PersonalAwards - 0.028368TimeforPassion - 0.048990DailyMeditation - 0.018407Age36to50 - 0.0430989Age51orMore - 0.111020AgeLessthan20 - 0.310447GenderMale$. This model is choosen because it has the lowest value of AIC.

#### Backward Elimination with BIC

Here, I use backward Elimination with BIC to search for a better model. The final fitted model selected by this method is as follows: $\hat{DailyStress} = 3.894142 - 0.028637CoreCircle + 0.020488SupportingOthers + 0.026546SocialNetwork - 0.015789Achievement + 0.109714BMIRange - 0.026586TodoCompleted - 0.032058Flow - 0.015788LiveVision - 0.092708SleepHours + 0.050924LostVacations + 0.124095DailyShouting - 0.268032SufficientIncome + 0.019677PersonalAwards - 0.028736TimeforPassion - 0.049936DailyMeditation - 0.304862GenderMale$. This model is choosen because it has the lowest value of AIC.

### Model2 - Logistic Regression Model & Propensity Score Matching

#### Model Specifics

In this Phrase, I first estimated the propensity score of our treatment group through a logistic regression model with "feeling stressed" as the dependent variable. Gender, daily meditation hours, if having sufficient income or not, number of daily shouting to someone, number of lost vacations and daily sleep hours as independent/predictor variables. In this Phrase, feeling stressed is our treatment, and life expectancy is our outcome of interest. Then I matched our treatment group and comparison group through the nearest neighbour matching method. Lastly, I evaluated the outcome (i.e. life expectancy) by running a regression on matched samples for unbalanced covariates.

## Results

In Phase 1, I produced three different models: the full model obtained through Multiple Linear Regression Method, A fitted model through AIC, and a fitted model through BIC. I chose the BIC model as my final model because it covers all three models' most significant factors and eliminated ones with a relatively larger p-value(not as significant ones). From the results of all three models, we found that six variables including gender, daily meditation hours, if having sufficient income or not, number of daily shouting to someone, number of lost vacations and daily sleep hours all have a p-value < 2e-16, which is much lower than 0.05, suggesting that they are some of the most significant variables in predicting stress level. Therefore, I took all six of them into account in phrase 2 - assigning a probability of feeling stressed. The table below(on the last page) is the regression result of average life expectancy with given factors. From the table, we can see that feeling_stressed is a very significant(p<0.001) predictor variable for average life expectancy. People feeling stressed tend to have around nine years less life expectancy than those not. Since R^2 is equal to 0.449, which shows that 44.9% of the variation in life expectancy can be explained with all the predictor variables in this model.

# Discussion

## Summary

In this study, I have examined two research questions that I wanted to tackle: 1) how personal lifestyle influences the number of daily stress people felt. 2) Testing if there is a causal relationship between a person's stress level and his/her life expectancy. The method/model used for the first phrase is through three models, one full model using a multiple linear regression model, one fitted model with AIC and one fitted model with BIC. The final selected model is the fitted model with BIC. The method/model used for the second phrase is through a logistic regression model and propensity score matching method. In this phrase, feeling stressed is our treatment, and life expectancy is our outcome of interest. I first estimated our treatment group's propensity score through a logistic regression model, feeling stressed as the dependent variable and six significant variables deduced in phrase one as independent variables. Then I matched our treatment group and comparison group through the nearest neighbour matching method. Lastly, I evaluated the outcome (i.e. life expectancy) by running a regression on matched samples for unbalanced covariates.

## Conclusion

In conclusion, for the first research question, we can conclude that some lifestyle criteria such as number of close friends, number of times supporting others, number of people interact every day, number of achievement in the past year, BMI range, number of todos completed, number of times feeling "flow" daily, number of years ahead with a clear vision, daily sleep hours, number of days of vacations lost in a year, number of times shouting at someone per day, having sufficient income or not, number of awards received in life, number of hours spent for passion daily, number of meditation practice daily and gender all could affect a person's stress level to certain degrees, especially the gender, daily meditation hours, if having sufficient income or not, number of daily shouting to someone, number of lost vacations and daily sleep hours. For the second research question, we can conclude a strong causal relationship between a person's life expectancy and if they are feeling stressed daily or not. If my model is correct, people who are feeling stressed daily could have around nine years shorter life expectancy than those who don't feel stressed daily.

## Weakness & Next Steps

My model's weakness in the first phrase is that I didn't check if there is "Multicollinearity" between my predictor variables. Especially when there is a large number of predictor variables, and there are likely correlations between them. In the next step, I will try to address this issue. Moreover, I didn't check the MLR model assumptions for the final model selected, the fitted model with BIC. In the next step, I will plot four diagnostic plots for this model and see if any violation of MLR assumptions such as normality, constant variance, etc.

My model's weakness in the second phrase is that I didn't evaluate the quality of my matching, i.e. if the treatment group and comparison group are balanced. I will try to use graphical comparison, percent bias reduction, or t-test for evaluation in the next step.

## References


- Alexander, R. (2020, November 5). Telling Stories With Data: Difference in differences. Telling Stories With Data. https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html
- Arthur, A. R. (2005). When stress is mental illness: A study of anxiety and depression in employees who use occupational stress counselling schemes. Stress And Health: Journal Of The International Society For The Investigation Of Stress, 21(4), 273-280.
- Avison, W., & Gotlib, I. H. (Eds.). (1994). Stress and mental health: Contemporary issues and prospects for the future. Springer Science & Business Media.
- Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3), 399-424.

- Black, N. (1996). Why we need observational studies to evaluate the effectiveness of health care. Bmj, 312(7040), 1215-1218.
- Dalat, Y.(2020). Lifestyle_and_Wellbeing_Data (Version 7)[Kaggle data file]. Retrieved from https://www.kaggle.com/ydalat/lifestyle-and-wellbeing-data
- Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. Sociological methods & research, 33(2), 188-229.
- Nathans, L. L., Oswald, F. L., & Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. Practical Assessment, Research, and Evaluation, 17(1), 9.
- Step function | R Documentation. (stats v3.6.2). R-core R-core@R-project.org. Retrieved December 22, 2020, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step
- Volkov, S. (2020). World Health Organization - Mental health. https://www.who.int/health-topics/mental-health.
- Yoshida, K. (2020, July 25). Introduction to tableone. Cran.r-Project.Org. https://cran.r-project.org/web/packages/tableone/vignettes/introduction.html


## Appendix

Github Repo link: https://github.com/liviasidealab/LifeExpectancy-vs-Stress

| | (1) |
|---|---|
| (Intercept) | 75.140 *** |
| | (2.562) |
| GENDERMale | 0.012 |
| | (0.460) |
| DAILY_MEDITATION1 | 0.211 |
| | (0.866) |
| DAILY_MEDITATION2 | -0.201 |
| | (0.773) |
| DAILY_MEDITATION3 | -0.826 |
| | (0.771) |
| DAILY_MEDITATION4 | 0.324 |
| | (0.803) |
| DAILY_MEDITATION5 | -0.064 |
| | (0.787) |
| DAILY_MEDITATION6 | -1.439 |
| | (1.365) |
| DAILY_MEDITATION7 | 0.638 |
| | (1.098) |
| DAILY_MEDITATION8 | -3.343 * |
| | (1.496) |
| DAILY_MEDITATION9 | -1.430 |
| | (1.735) |
| DAILY_MEDITATION10 | -1.827 * |
| | (0.822) |
| SUFFICIENT_INCOME2 | -0.111 |
| | (0.359) |
| DAILY_SHOUTING1 | 0.667 |
| | (0.773) |
| DAILY_SHOUTING2 | 0.248 |
| | (0.787) |
| DAILY_SHOUTING3 | 0.267 |
| | (0.815) |