BIOSTAT 629 001 WN 2021 Final Report

Sleep Quality in Times of COVID-19 Pandemic

Wenjing Zhou
(wenjzh@umich.edu)
University of Michigan
Department of Statistics

Tuesday 16th March, 2021

Contents

Introduction

Due to the Coronavirus disease 2019 (COVID-19) outbreak, a relatively long period of stress has been imposed on people worldwide [2]. In the United States, social distancing and staying in place order leads to people's isolation and increases people's time of staying indoors, which may have negative effects on people's mental health[1]. The goal of this project is to compare people's psychological well-being and sleep quality before and during the COVID-19 pandemic, so I focus on the information related to people's sleep quality, stress, and mood in the years 2019 and 2020.

I use the data from MIPACT (Michigan Predictive Activity and Clinical Trajectories) Study to conduct this analysis. The MIPACT study contains data about electronic health records(EHR), participant survey data, genetic information, blood pressure measurements, and Apple Watch activity and clinical data.

Data Description

0.0.1 Data Files

Here I list the data files I have checked in detail, with the variables I extract or the reasons for not being utilized.

- ActiveEnergyBurned_202004.csv, AppleExerciseTime_202004.csv :
  - all values are equal to one.
- BodyFatPercentage_202004.csv, BMI_202004.csv :
  - missing rates are 97.797% and 96.843%.
- BloodPressureDiastolic_202004.csv, BloodPressureSystolic_202004.csv:
  - ParticipantResearchID, StartDate, Value
- EHR_Demographic_202010.csv : - ParticipantResearchID, EnrollmentDate, AgeAtEnrollment, GenderName, MaritalStatusName, RaceName
- Surveys_201901.csv - Surveys_202011.csv : - ParticipantResearchID, SurveyName, SurveyStartDate, SurveyQuestion, SurveyAnswer

0.0.2 Abstract of Data Exploration

Here are some findings I'd like to highlight first.

- In Surveys_202004.csv, for the question "In the past 7 days:My sleep quality was", each person only answers this question once. Thus, if I use the sleep quality as the response in regression, then I could average other numerical predictors by month and individual, such as blood pressure and mood.
- More than 95% participants in the Surveys_202004.csv do not have BMI and bodyfat on file.
- From 2019 to 2020, people's mental states were relatively stable, but in April 2020, the population's anxiety increased significantly.
- From 2019 to 2020, overall, participants' sleep quality was getting better. From April 2020 to November 2020, people found falling asleep easier and easier, probably because they gradually got used to the new lifestyle.

0.0.3 Sleep Quality and Mental Health

In the participant survey data, four surveys are highly related to the research goal: 'Sleep Disturbance - Every 4th Quarter', 'Life Stress - Biannually', 'Sleep Disturbance - Quarterly', 'Generalized Anxiety Disorder - Quarterly'. Population's answers to the questions in these surveys are visualized according to the time in Python. From January 2019 to November 2020, there are 23 months in total, which are labeled from 0 to 22. In each month, I calculated the mean and 95% confidence interval of people's answers.

For most of the questions, there is no obvious trend or pattern. However, two phenomenons do worth noting.

In Figure 1, the four questions are about people's anxiety disorder. We can easily find people's reaction reaches the peak at the 15th month, April 2020. Michigan issued statewide stay-at-home order on March 23, 2020, so companies and schools gradually started the work-from-home policy in late March. Facing this unprecedented and unusual change, it's reasonable for people's anxiety increased sharply. As people knew more about this disease and got used to the new lifestyle, the population's anxiety decreased accordingly.

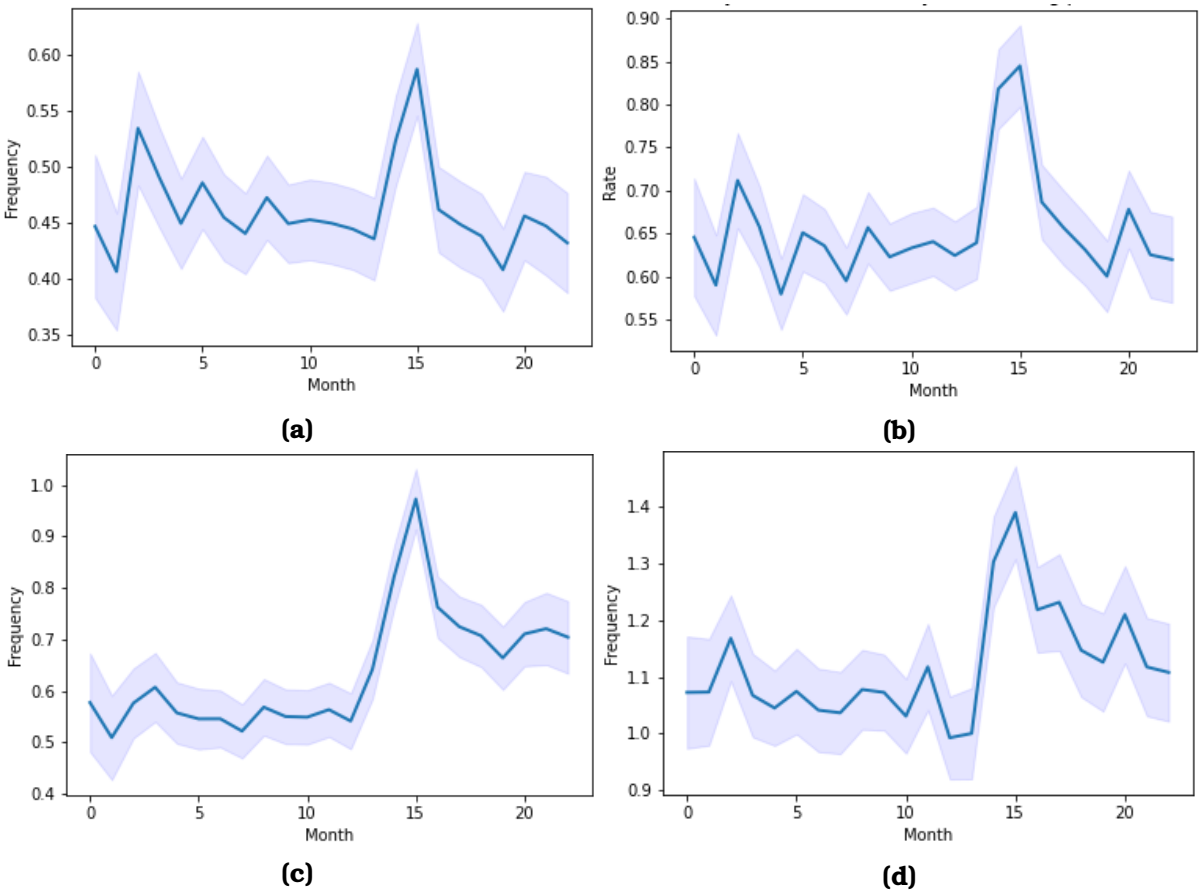

Figure 1: Mean and 95% confidence interval of people's answers to the questions: **(a)** Over the last 2 weeks, how often have you been bothered by the following problems? Not being able to stop or control worrying. **(b)** Over the last 2 weeks, how often have you been bothered by the following problems? Feeling nervous, anxious, or on edge. **(c)** Over the last 2 weeks, how often have you been bothered by the following problems? Feeling afraid as if something awful might happen **(d)** In the last month, how often have you felt that you were unable to control the important things in your life?

From Figure 2, we can see there is a long-term downward trend. This trend reveals people's confidence and sleep quality are continuously getting better in the two years. From Figure 3, we can see there is a short-term downward trend starting from the 15th month. It reflects that people's sleep quality was nearly the worst in April 2020, but gradually turned good after the stay-in-place order took effect.

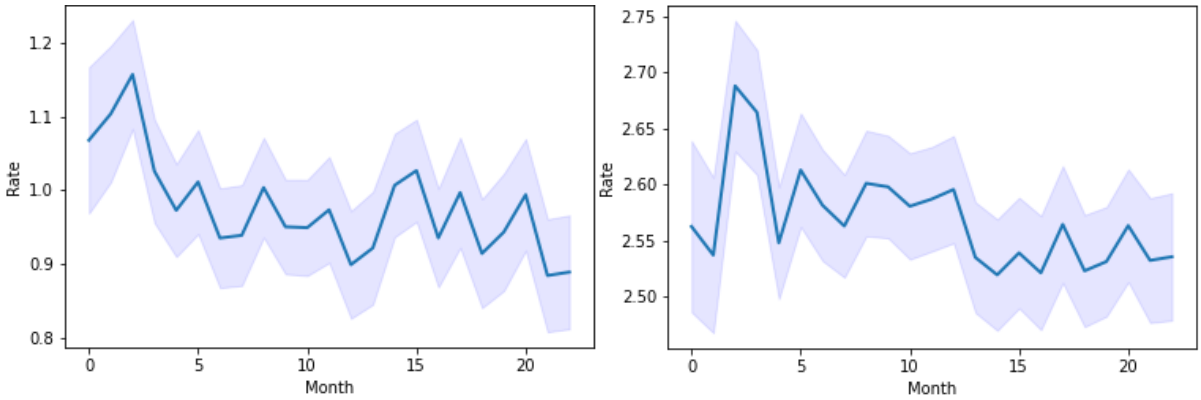

Figure 2: Mean and 95% confidence interval of people's answers to the questions: **(a)** In the last month, how often have you felt difficulties were piling up so high that you could not overcome them? **(b)** In the past 7 days: My sleep quality was: 1 - very good 2 - good 3 - fair 4 - poor 5 - very poor

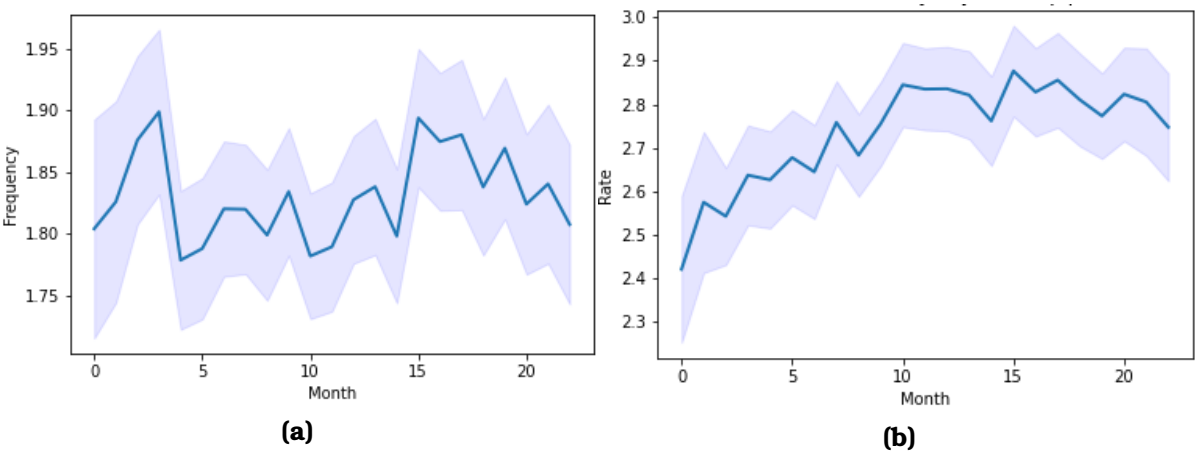

Figure 3: Mean and 95% confidence interval of people's answers to the questions: **(a)** In the past 7 days: I had difficulty falling asleep. **(b)** Please rate the current (i.e. last 2 weeks) Severity of your sleep problems: Difficulty staying asleep.

From the above exploratory data analysis, we can reach a preliminary conclusion: from 2019 to 2020, people's sleep quality and mental health are not severely affected by the appearance of Covid-19, but in April 2020, there are indeed some significant changes in people's temporary well-being and sleep quality. Thus, It worth investigating to explore the factors affecting the population's sleep quality in April 2020, the time point when people began to stay home in the pandemic. Thus, the sleep quality in April 2020 is chosen to be the outcome in the models.

0.0.4 Predictor Selection and Data Cleaning

Regarding the predictors to include in the model, I have checked the ActiveEnergyBurned, AppleExerciseTime files, as physical exercise should have some effects on people's mental state and sleep depth. However, the two files look problematic, because all entries in the "Value" column are "1", and if I sum up the values by date and person, then the sum of energy burned and exercise time is the same. Thus, the information from these two files can not be used at this time.

Here is a summary of the following data cleaning steps. I use the files mentioned in the section of 0.0.1 Data Files.

- Use package dplyr in R to extract **BMI, body fat, age at enrollment, gender, marital status, race, mood, stress, sleep quality, systolic and diastolic blood pressure**; order the data frames by ParticipantResearchID and StartDate to check the datasets.
- Use string manipulation to change the age at enrollment to the actual age.
- Found each individual only has one response of sleep quality. Then average the predictors including BMI, body fat, mood, stress, and blood pressure by the individual.
- Left join the sleep quality with predictors on ParticipantResearchID. The shape of the data frame is (1362, 12) at this step.
- Check the proportion of missingness in each column. See Table 1.
- Drop the columns of BMI and body fat.
- Drop all rows still containing NA's, which accounts for about 10% of the total.
- The data is ready for analysis with the shape (1196, 10).

Variable	Missing Proportion (%)	Variable	Missing Proportion (%)
bodyfat	97.797	systolic	4.993
BMI	96.843	race	0.367
marital status	7.048	sleep quality	0
mood	5.14	age	0
stress	5.14	gender	0
diastolic	4.993	ID	not applicable

Table 1: Missing proportion of variables

0.0.5 Analysis of Predictors

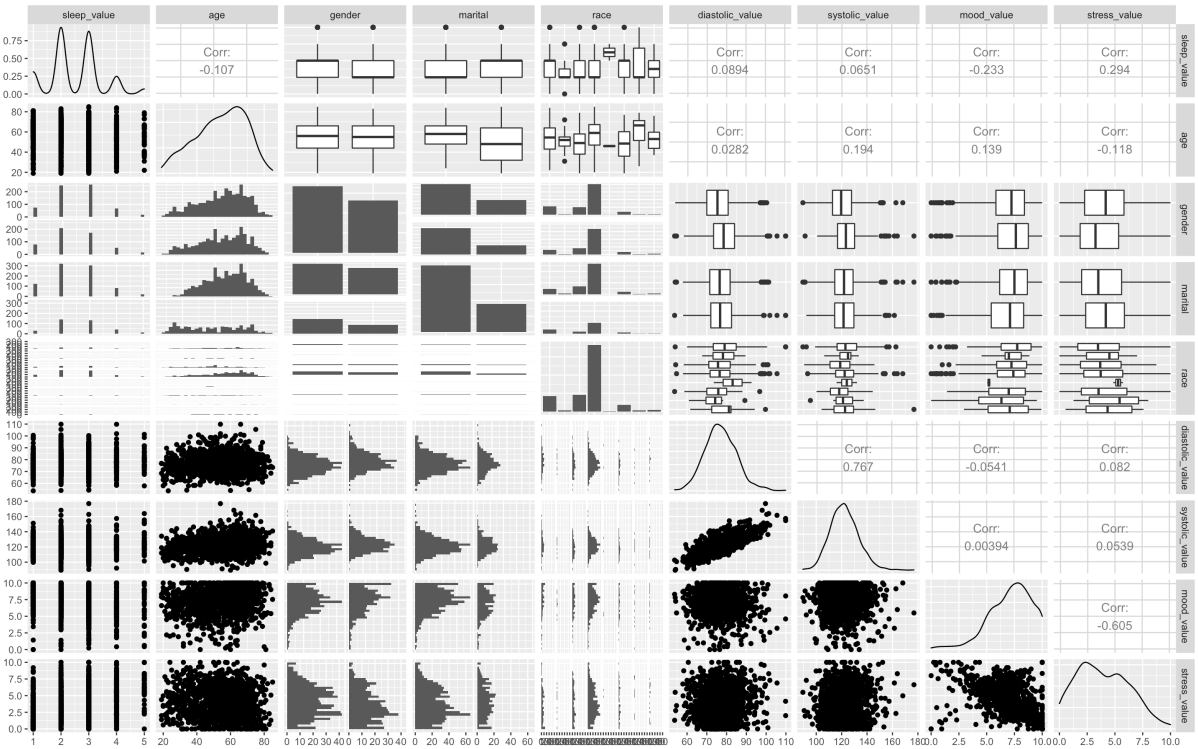

Figure 4: Visualization of covariates. (Left part: scatterplots of each pair of variables; Right: Pearson correlation; diagonal: variable distribution.)

From Figure 6, we can first see that systolic blood pressure and diastolic blood pressure are highly correlated (Pearson correlation: 0.767), so only the diastolic blood pressure will be kept in the regression model, as it has a larger correlation with the response variable.

Although the correlation between mood and stress is -0.605, none of them are considered being removed, because these two variables have the two largest correlations with the response variable.

Regarding the distributions of these variables, most numerical variables' distributions are bell-shaped, while the people reporting "good" sleep quality are almost as many as people who report "fair" sleep quality. It is also noticeable that the distribution of age differs in different marital statuses, so I think adding the interaction between these two is worth trying.

Methods and Results

0.0.6 Linear Regression Model

As mentioned previously, people's sleep quality in April 2020 looks unusual, so in order to analyze the factors influencing the sleep quality, and considering the fact that each person only has one record of sleep quality in this month, I build a linear regression model. In MIPACT dataset, "1" represents good sleep quality, and "5" represents poor sleep quality.

$$\begin{aligned}\text{sleep quality} = \beta_0 + \text{mood} * \beta_1 + \text{stress} * \beta_2 + \text{gender} * \beta_3 + \\ \text{age} * \beta_4 + \text{race} * \beta_5 + \text{marital} * \beta_6 + \text{diastolic} * \beta_7\end{aligned} \tag{1}$$

Among 7 predictors, there are four having significant relationships with the response. We can see older people and Asians are likely to have a better sleep quality in April 2020, probably because they have more previous knowledge of the COVID-19, and thus less mentally influenced by the outbreak. Stress also has a strong relationship to the sleep quality in this model, and it makes sense that a better mood and less stress will contribute to good sleep quality.

Predictors	Coefficients	P Value
age	-0.00543	0.0060 **
raceAsian	-0.31586	0.0012 **
mood	-0.03920	0.0201 *
stress	0.08932	1.8e-10 ***

Table 2: Summary of the linear regression model.
(See full version in the Appendix.)

0.0.7 Linear Model with Interactions

From the correlation analysis, the distribution of age differs apparently in different marital statuses, and from the first model, age has a significant influence on the response, so I add the interactions between marital status and the numerical predictors. There is some research [3] revealing the blood pressure is associated with gender, so the interaction between these two is added to the regression model too. The formula of the mixed model is shown below.

$$\begin{aligned}\text{sleep quality} = &\beta_0 + \text{mood} * \beta_1 + \text{stress} * \beta_2 + \text{gender} * \beta_3 \\ &+ \text{age} * \beta_4 + \text{race} * \beta_5 + \text{marital} * \beta_6 + \text{diastolic} * \beta_7 \\ &+ \text{marital:age} * \beta_8 + \text{marital:mood} * \beta_9 + \text{marital:stress} * \beta_{10} \\ &+ \text{marital:diastolic} * \beta_{11} + \text{gender:diastolic} * \beta_{12}\end{aligned} \tag{2}$$

From the summary of the mixed model, we can see that the effects of stress, age, and race of Asians are nearly the same to that in the simple linear model. And marital status is also significant in this model. Especially, in the unmarried population, younger people tend to have a better sleep quality, and high diastolic blood pressure may lead to bad sleep quality.

Predictors	Coefficients	P Value
stress	0.09201	5.7e-08 ***
age	-0.00974	0.00016 ***
raceAsian	-0.31854	0.00110 **
maritalUnmarried.	-1.72704	0.00775 **
age:maritalUnmarried	0.00938	0.01172 *
maritalUnmarried:diastolic	0.01581	0.02118 *

Table 3: Summary of the linear regression model with interactions. (See full version in the Appendix.)

Summary and Discussion

In summary, people's psychological well-being and sleep quality are relatively stable and have the trend of getting better in the years 2019 and 2020. In April 2020, the COVID-19 caused a certain degree of anxiety, and the population had

the worst quality. to study the factors affecting people's sleep quality this year, I build two linear regression models and find that age, race, stress and mood, marital status and diastolic blood pressure have significant relationships with sleep quality. It is helpful to add the interactions to the linear regression model because the adjusted R^2 increases and more predictors are significant.

There are some limitations of the current models. First, in Surveys_202004.csv, as each person only has one record of sleep quality, I can not build longitudinal models. I plan to use more data files in the next step, from 2017 to 2020, and build two mixed effect models: one is studying the factors influencing sleep quality before COVID-19 begun, and one is afterward. Secondly, the information on physical activity, medication, and dietary data are not included in the models. I plan to contact the researchers of the MIPACT study to understand the energy data and exercise data, and if workable, I will analyze more predictors' effects on sleep quality.

The code of this research can be found at
https://github.com/wenjzh/Health_data_science_project

Bibliography

[1] Desana Kocevska, Tessa F Blanken, Eus JW Van Someren, and Lara Rösler. Sleep quality during the covid-19 pandemic: not one size fits all. *Sleep medicine*, 76:86–88, 2020.

[2] Markku Partinen. Sleep research in 2020: Covid-19-related sleep disorders. *The Lancet Neurology*, 20(1):15–17, 2021.

[3] Jane F Reckelhoff. Gender differences in the regulation of blood pressure. *Hypertension*, 37(5):1199–1208, 2001.

Appendix

```
Call:
lm(formula = sleep_value ~ ., data = analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1407 -0.5761 -0.0602  0.5331  2.9606

Coefficients:
                                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 2.21473    0.34714    6.38  2.5e-10 ***
age                                        -0.00543    0.00197   -2.75   0.0060 **
genderMale                                 -0.05819    0.05412   -1.08   0.2825
maritalUnmarried                            0.01960    0.05980    0.33   0.7432
raceAmerican Indian or Alaska Native       -0.33300    0.27735   -1.20   0.2301
raceAsian                                  -0.31586    0.09755   -3.24   0.0012 **
raceCaucasian                              -0.13850    0.07742   -1.79   0.0739 .
raceNative Hawaiian and Other Pacific Islander  0.52502    0.63279    0.83   0.4069
raceOther                                  -0.12717    0.12653   -1.01   0.3151
racePatient Refused                        -0.05248    0.29131   -0.18   0.8571
raceUnknown                                -0.21574    0.22091   -0.98   0.3290
diastolic_value                             0.00523    0.00513    1.02   0.3090
systolic_value                              0.00240    0.00377    0.64   0.5247
mood_value                                 -0.03920    0.01684   -2.33   0.0201 *
stress_value                                0.08932    0.01389    6.43  1.8e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.888 on 1181 degrees of freedom
Multiple R-squared:  0.112,     Adjusted R-squared:  0.102
F-statistic: 10.7 on 14 and 1181 DF,  p-value: <2e-16
```

Figure 5: Summary of simple linear model

```
Call:
lm(formula = sleep_value ~ mood_value + stress_value + gender +
    age + race + marital + diastolic_value + marital:age + marital:mood_value +
    marital:stress_value + marital:diastolic_value + +gender:diastolic_value,
    data = analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2448 -0.5977 -0.0396  0.5452  3.0957

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   2.75009    0.45676    6.02  2.3e-09 ***
mood_value                                   -0.04115    0.02136   -1.93  0.05431 .
stress_value                                  0.09201    0.01684    5.46  5.7e-08 ***
genderMale                                    0.57442    0.51202    1.12  0.26215
age                                          -0.00974    0.00257   -3.79  0.00016 ***
raceAmerican Indian or Alaska Native         -0.35911    0.27647   -1.30  0.19423
raceAsian                                    -0.31854    0.09733   -3.27  0.00110 **
raceCaucasian                                -0.14422    0.07755   -1.86  0.06318 .
raceNative Hawaiian and Other Pacific Islander  0.38514    0.63205    0.61  0.54240
raceOther                                    -0.12527    0.12630   -0.99  0.32150
racePatient Refused                           0.00701    0.29153    0.02  0.98083
raceUnknown                                  -0.23459    0.22060   -1.06  0.28782
maritalUnmarried                             -1.72704    0.64748   -2.67  0.00775 **
diastolic_value                               0.00535    0.00501    1.07  0.28646
age:maritalUnmarried                          0.00938    0.00372    2.52  0.01172 *
mood_value:maritalUnmarried                   0.00848    0.03453    0.25  0.80613
stress_value:maritalUnmarried                -0.00498    0.02945   -0.17  0.86563
maritalUnmarried:diastolic_value              0.01581    0.00685    2.31  0.02118 *
genderMale:diastolic_value                   -0.00801    0.00657   -1.22  0.22346
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.884 on 1177 degrees of freedom
Multiple R-squared:  0.123,     Adjusted R-squared:  0.109
F-statistic: 9.13 on 18 and 1177 DF,  p-value: <2e-16
```

Figure 6: Summary of linear model with interactions