# Determining What Causes Poor Mental Health

**Daniel Jin**

## Introduction:

Social, psychological, and mental well-being makes up our mental health.[2] Mental health is a big part of our overall health. About 14% of the diseases in the world are attributed from mental illnesses. Mental illnesses interact with other health conditions which illustrates the importance of mental health. [1] Some of the known risk factors for poor mental health includes: poor physical health, poor diet, high stress, smoking, excessive working, poor socio-economic status, getting little sleep, and biological factors. [2,3] Our objective in this study is to determine what other risk factors exist for poor mental health. Specifically, what predictors exist that can help us predict how many days of poor mental health a person would experience in any given month. Also, part of the study is to verify some of the already known risk factors for poor mental health to see if it is congruent to what we know.

## Description of Dataset:

The original dataset was taken from the 2014 Behavioral Risk Factor Surveillance System located at http://www.cdc.gov/brfss/annual_data/annual_2014.html. This dataset contains 464664 observations and 279 variables collected through phone surveys throughout the USA. The response variable is 'Number of days mental health not good' which ranges from 0 to 30, spanning a month's worth of time. The following variables were used in the analysis: 'Interview Month', 'Number of days physical health not good', 'Average hours of sleep per day', 'Education level' (1=Never attended or only kindergarten, 2=grades 1-8, 3=grades 9-11, 4=grade 12 or GED, 5=college 1-3, 6= college4+), 'Frequency of smoking'(1=not at all, 2=some days, 3=everyday), 'Average hours of work per week', 'Race' (1=White and not Hispanic, 2=Black and not Hispanic, 3=Other race and not Hispanic, 4=Multiracial and not Hispanic, 5=Hispanic), 'Age Group' (1=ages 18-24, 2=ages 25-34, 3=ages 35-44, 4=ages 45-54, 5=ages 55-64, 6=ages 65+), and BMI. Missing observations and 'refusal to answer' or similar observations were all removed from the dataset prior to analysis. The resulting dataset contains 3095 observations. SAS was used to create a random sample of 10 observations is shown on Table 1. SAS was also used throughout the rest of the analysis.

**TABLE 1: Random sample of 10 observations**

| Obs | Interview Month | Number of days physical health not good | Number of days mental health not good | Average hours of sleep per day | Education Level | Frequency of Smoking | Average hours of work per week | Race | Age Group | BMI |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 07 | 0 | 0 | 7 | 4 | 3 | 53 | 1 | 3 | 2903 |
| **2** | 01 | 7 | 0 | 6 | 4 | 1 | 40 | 1 | 4 | 2421 |
| **3** | 02 | 0 | 0 | 4 | 5 | 3 | 70 | 1 | 3 | 3085 |
| **4** | 10 | 0 | 0 | 7 | 4 | 3 | 50 | 1 | 2 | 2483 |
| **5** | 03 | 10 | 0 | 8 | 5 | 3 | 18 | 3 | 6 | 3071 |
| **6** | 03 | 2 | 0 | 6 | 6 | 3 | 45 | 1 | 5 | 2391 |
| **7** | 01 | 0 | 0 | 6 | 5 | 3 | 40 | 1 | 4 | 3877 |
| **8** | 04 | 0 | 0 | 6 | 4 | 1 | 40 | 1 | 2 | 3399 |
| **9** | 09 | 0 | 10 | 6 | 4 | 1 | 30 | 1 | 2 | 2929 |
| **10** | 10 | 0 | 0 | 7 | 5 | 3 | 40 | 1 | 2 | 3519 |

Several categorical variables were re-coded into dummy variables which include 'Interview Month' (JAN, FEB, MAR, APR, MAY, JUN, JUL, AUG, SEP, OCT, NOV, DEC), 'Education Level' (NEVER, MS, SomeHS, HSGrad, SomeCol, ColGrad), 'Frequency of smoking' (NONE, SOME, EVERYDAY), Race (WHITE, BLACK, OTHER, MULTI, HISPANIC), 'Age Group' (A18_24, A25_34, A35_44, A45_54, A55_64, A65PLUS) prior to fitting regression models.

## Statistical Analysis:

Descriptive statistics and frequency counts are presented below.

**Tables 2-7: Descriptive Statistics**

| Education Level | Frequency |
|---|---|
| 1 | 2 |
| 2 | 36 |
| 3 | 181 |
| 4 | 935 |
| 5 | 903 |
| 6 | 1038 |

| Interview Month | Frequency |
|---|---|
| 01 | 348 |
| 02 | 431 |
| 03 | 295 |
| 04 | 229 |
| 05 | 243 |
| 06 | 169 |
| 07 | 166 |
| 08 | 248 |
| 09 | 292 |
| 10 | 279 |
| 11 | 244 |
| 12 | 151 |

| Frequency of Smoking | Frequency |
|---|---|
| 1 | 965 |
| 2 | 340 |
| 3 | 1790 |

| Age Group | Frequency |
|---|---|
| 1 | 112 |
| 2 | 373 |
| 3 | 501 |
| 4 | 800 |
| 5 | 933 |
| 6 | 376 |

| Race | Frequency |
|---|---|
| 1 | 2501 |
| 2 | 394 |
| 3 | 62 |
| 4 | 38 |
| 5 | 100 |

| Variable | Obs | Mean | StDev | Min | Median | Max |
|---|---|---|---|---|---|---|
| Number of days mental health not good | 3095 | 3.27 | 7.63 | 0 | 0 | 30 |
| Number of days physical health not good | 3095 | 2.79 | 6.86 | 0 | 0 | 30 |
| Average hours of sleep per day | 3095 | 6.79 | 1.25 | 1 | 7 | 16 |
| Average hours of work per week | 3095 | 42.09 | 14.13 | 0 | 40 | 96 |
| BMI | 3095 | 2808.16 | 567.17 | 1412 | 2728 | 6147 |

Histograms and a scatter plot matrix of continuous variables are presented below in Figure 1. A correlation matrix is presented in Table 8.

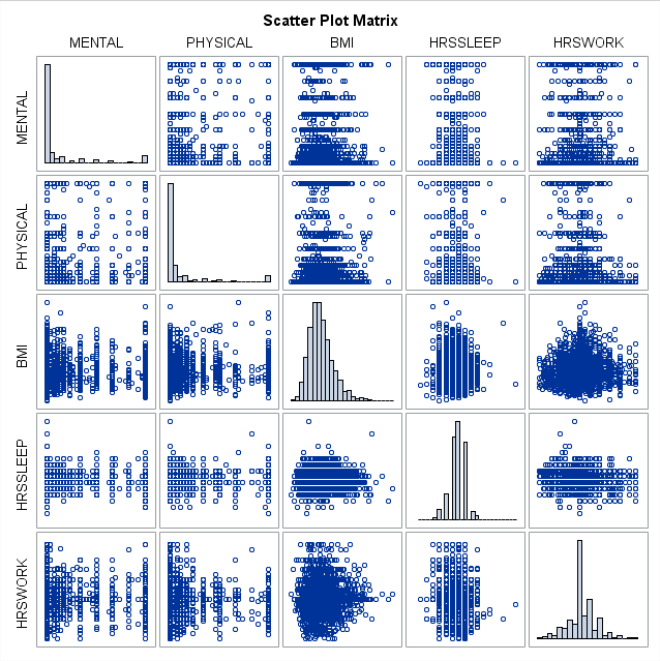


**Figure 1: Histograms and Scatter plot Matrix**

**Table 8: Correlation matrix**

| | MENTAL | PHYSICAL | BMI | HRSSLEEP | HRSWORK |
|---|---|---|---|---|---|
| **MENTAL** | 1.00000 | 0.25544<br><.0001 | 0.06662<br>0.0002 | -0.15859<br><.0001 | -0.02937<br>0.1023 |
| **PHYSICAL** | 0.25544<br><.0001 | 1.00000 | 0.07765<br><.0001 | -0.06071<br>0.0007 | -0.08180<br><.0001 |
| **BMI** | 0.06662<br>0.0002 | 0.07765<br><.0001 | 1.00000 | -0.06004<br>0.0008 | 0.03075<br>0.0871 |
| **HRSSLEEP** | -0.15859<br><.0001 | -0.06071<br>0.0007 | -0.06004<br>0.0008 | 1.00000 | -0.13152<br><.0001 |
| **HRSWORK** | -0.02937<br>0.1023 | -0.08180<br><.0001 | 0.03075<br>0.0871 | -0.13152<br><.0001 | 1.00000 |

We see that the normality assumption is satisfied for BMI, HRSSLEEP, and HRSWORK, but not for MENTAL or PHYSICAL. No clear pattern is observed from any of the scatter plots, and looking over to Table 8 we see the same thing; no moderate or strong linear relationships between any of the variables. There is a weak positive correlation between PHYSICAL and MENTAL (r = 0.25544) and weak negative correlation between MENTAL and HRSSLEEP (r=-0.15859). Collinearity is clearly not a problem here. A full correlation matrix of all the variables in the dataset reveals that collinearity is not a problem as well, but is not shown here due to the size.

To remedy the normality violation for PHYSICAL, the variable is re-coded as a dummy variable. PHYSICALBAD=1 if 1 or more days physical health not good in a month, PHYSICALBAD=0 if 0 days physical health not good.

Next, simple linear regression models are fit for each possible predictor for 'Number of days mental health not good', and each model is summarized below in Table 9.

**Table 9: Simple linear regression models summary**

| Possible predictor | Reference group | Overall P value | $R^2$ |
|---|---|---|---|
| **Interview month** | JAN | 0.1625 | 0.0014 |
| **BMI** | N/A | 0.0002 | 0.0041 |
| **Race** | WHITE | 0.0492 | 0.0018 |
| **Average hours of work per week** | N/A | 0.1023 | 0.0005 |
| **Education level** | ColGrad | <.0001 | 0.0109 |
| **Age group** | A65PLUS | <.0001 | 0.0131 |
| **Frequency of smoking** | NONE | <.0001 | 0.0249 |
| **Average hours of sleep per day** | N/A | <.0001 | 0.0248 |
| **PHYSICALBAD** | N/A | <.0001 | 0.0395 |

We see that all of the models are significant at the level of alpha = 0.05 except 'Interview month' and 'Average hours of work per week'. The coefficient of determination values are very low for all of the models indicating that those variables alone aren't good at explaining the variance observed in 'number of days of mental health not good', so model fit is not good based on $R^2$ value.

It is worth noting that multiracial people differed significantly from Whites (p = .0067) in the Race model, people that completed some high school and people that completed some college differed significantly from college graduates (p <.0001 and p = .0066 respectively) in the 'Education level' model, smoking on some days and everyday differed significantly from not smoking at all (both p <.0001) in the 'Frequency of smoking' model, and every age group differed significantly from people aged 65 or higher (p < .0001 for all except A54_64 = .006) in the 'Age group' model.

Residual plots of the three continuous variables are shown below in Figures 2-4.

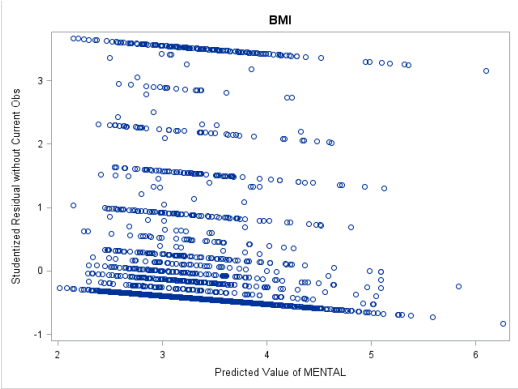


**Figure 2: Residuals of BMI model**

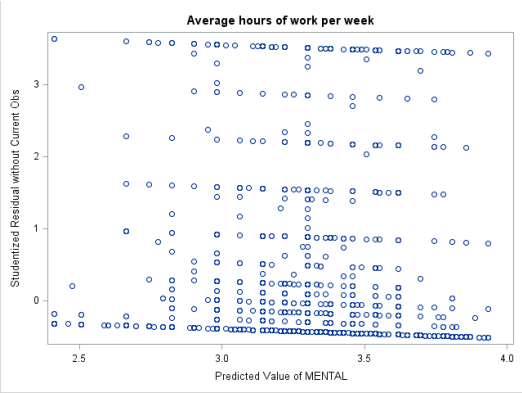


**Figure 4: Residuals of work Model**

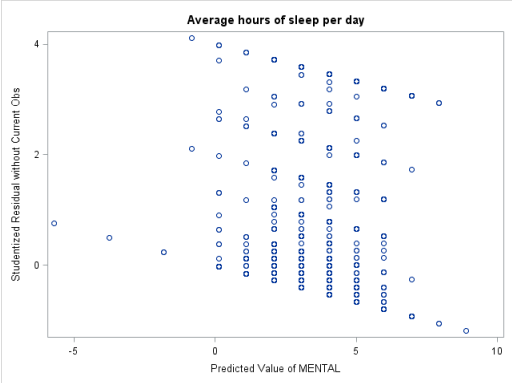


**Figure 3: Residuals of sleep Model**

It is clear that the equal variance assumption is violated in the BMI and sleep models. Various transformations were tried, but were unable to remedy the problem. The same assumption is easily satisfied in the work model. No clear linear relationship is seen in any of the scatter plots in Figure 1 of MENTAL versus any of the other possible predictors, so that assumption may be violated as well.

There doesn't appear to be any curvilinear relationships in any of the three plots above, so no quadratic terms will be used.

Forward Stepwise Selection criteria is used initially with the maximum model including all three continuous variables (Average hours work per week, Average hours sleep per day, and BMI). All three variables are kept in the model. It is interesting how the work variable is significant in the multiple linear regression model, but not the simple linear regression model containing only work as a predictor. This is because the effects of the other predictors, BMI and sleep affects it.

Next, we threw in the other variables one at a time and computed partial F statistics to decide if it should be added in the model at alpha = 0.05. Below are Table 10 that summarizes this step and Table 11 with the final model. We checked for two possible interaction terms that made sense.

**Table 10: Model building**

| Variable | Add? |
|---|---|
| **Interview month** | No |
| **Age group** | Yes |
| **Race** | No |
| **Education level** | Yes |
| **Frequency of smoking** | Yes |
| **PHYSICALBAD** | Yes |
| **Interview month** | No |
| **Race** | No |
| **Sleep * Work** | No |
| **Sleep * PHYSICALBAD** | Yes |

**Table 11: Final Model**

| Parameter Estimates | | | | | |
|---|---|---|---|---|---|
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > \|t\|** |
| **Intercept** | **1** | 3.09830 | 1.34600 | 2.30 | 0.0214 |
| **HRSWORK** | **1** | -0.02004 | 0.00962 | -2.08 | 0.0372 |
| **HRSSLEEP** | **1** | -0.56284 | 0.13809 | -4.08 | <.0001 |
| **BMI** | **1** | 0.00076356 | 0.00023469 | 3.25 | 0.0012 |
| **A18_24** | **1** | 3.26265 | 0.81309 | 4.01 | <.0001 |
| **A25_34** | **1** | 1.01907 | 0.55650 | 1.83 | 0.0672 |
| **A35_44** | **1** | 1.11014 | 0.51342 | 2.16 | 0.0307 |
| **A45_54** | **1** | 1.07831 | 0.46837 | 2.30 | 0.0214 |
| **A55_64** | **1** | 0.65814 | 0.45211 | 1.46 | 0.1456 |
| **NEVER** | **1** | -3.16274 | 5.16014 | -0.61 | 0.5400 |
| **MS** | **1** | 0.62506 | 1.24124 | 0.50 | 0.6146 |
| **SomeHS** | **1** | 2.30096 | 0.59815 | 3.85 | 0.0001 |
| **HSGrad** | **1** | -0.25026 | 0.33982 | -0.74 | 0.4615 |
| **SomeCol** | **1** | 0.14562 | 0.33774 | 0.43 | 0.6664 |
| **SOME** | **1** | 1.35226 | 0.44589 | 3.03 | 0.0024 |
| **EVERYDAY** | **1** | 1.97698 | 0.31028 | 6.37 | <.0001 |
| **PHYSICALBAD** | **1** | 6.63373 | 1.48624 | 4.46 | <.0001 |
| **SleepPHYSICALBAD** | **1** | -0.56309 | 0.21640 | -2.60 | 0.0093 |

**Adjusted $R^2$ = 0.0921 F = 19.46**

There doesn't appear to be any extreme values in the parameter estimates or standard errors, they are all stable. Negative parameter estimates indicates that that variable is associated with observing less days of poor mental health in any given month. It is surprising how HRSWORK has a negative sign, meaning that the more you work per week results in less days of poor mental health, whereas we

associate work to be typically stressful which in turns causes poor mental health. This is the only result that is not congruent to what we know (excessive working increases chance of poor mental health). HRSSLEEP has a negative sign, which is not a surprise since we associate more sleep with good mental health. BMI has a fairly low beta estimate, but since it is multiplied by numbers in the thousands, it can still translate to something clinically useful. For an example, someone with a BMI of 2000 would be in the range of normal healthy weight, but someone with a BMI of 4000 would be severely obese. The person with BMI of 4000 would experience about 1.5 more days of poor mental health than the person with a BMI of 2000. This doesn't strike any surprise since people tend to be more conscious of their self appearance at higher BMI's resulting in more likely to experience poor mental health due to pressures they experience. A18_24 has the highest beta coefficient out of the various dummy variables for the age groups. This also makes sense because this age bracket tends to be in college and graduate school which can be much more stressful when compared to people aged 65 or higher. None of the age groups are negative which means that everyone age 64 or below are more likely to experience more days of poor mental health in any given month compared to people aged 65 or higher. At the age bracket 65 or higher people tend to be retired and don't have to worry about a whole lot so it makes sense. It is interesting how people who have never gone to school or only completed kindergarten has a beta coefficient of -3.16 indicating that they experience about 3 less days of poor mental health compared to college graduates. But, in the descriptive statistics we see that only 2 people have this education level so probably we don't have enough data to make complete sense of this beta coefficient. We see that people with some high school completed experience much more days of poor mental health compared to college graduates. This may be due to still trying to obtain their GED's or unable to find jobs due to never graduating high school. It is interesting to see that people who have only completed high school tend to experience fewer days of poor mental health when compared to college graduates. Next, we see that both people that smoke on some days and people that smoke everyday tend to have more days of poor mental health when compared to people who never smoke. We also see that people that smoke everyday have a higher beta coefficient meaning that they have more days of poor mental health per month. One of the reasons people smoke is because of stress, which in turn is having poor mental health. Some of it might also be attributed to the stress of trying to quit smoking but are unable to resulting in poorer mental health. PHYSICALBAD is the variable with the highest parameter estimate in the model which basically means that if someone has even one day of poor physical health, they experience about 6.6 more days of poor mental health compared to people with 0 days of poor physical health. Next, we have the only interaction term between Sleep and PHYSICALBAD with a negative coefficient. This makes sense because poor physical health includes things like having an injury or fatigue and sleep helps alleviate both of these problems. Overall, the model is significant, but we have a very low adjusted $R^2$ value indicating that our fit is extremely poor.

Next, we need to check model diagnostics. Figure 5 and 6 are shown below.

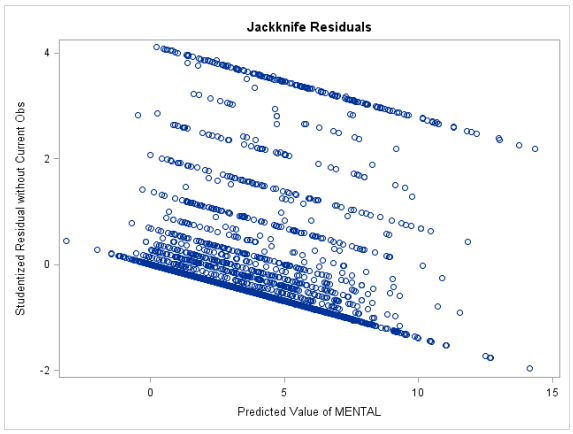


**Figure 5: Jackknife residual plot**

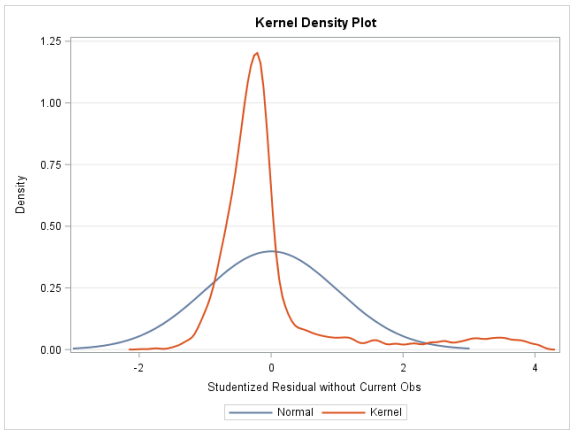


**Figure 6: Kernel Density plot**

It is clear that the equal variance assumption is not satisfied by looking at Figure 5 and the normality assumption is not satisfied by looking at Figure 6. We may have suspected this by our low Adjusted $R^2$ value in the final model which shows poor fit. Several common transformations were tried to satisfy regression assumptions, but none of them worked. Since the assumptions of regressions are satisfied, none of the conclusions we draw from the data are valid.

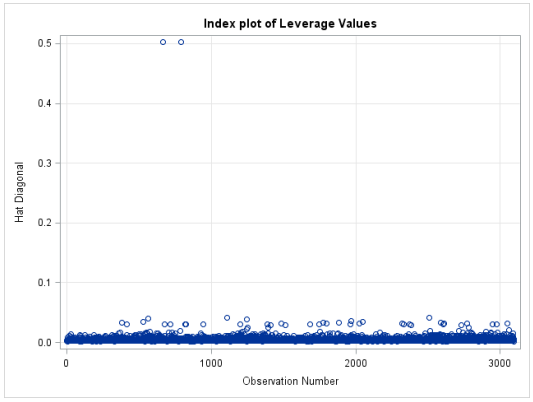


**Figure 7: Leverage index**

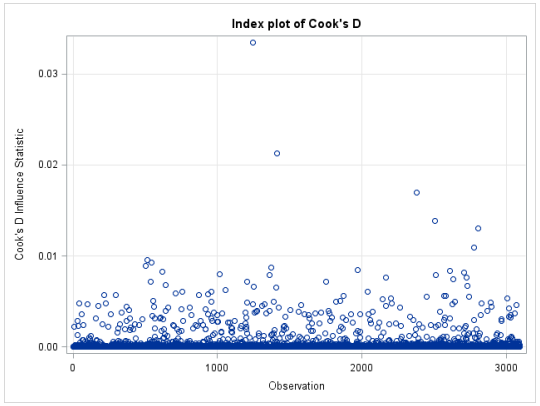


**Figure 8: Cook's distance index**

Using the cutoff point of about 0.001 for Cook's distance we see that there are over a hundred influential points. We will ignore them since it would be tedious to look at every observation to check if it is clinically possible. Instead, we will look at the very extreme observations. On Figure 7, there are two extreme points (Observation 794 and 665). After looking at the observations, there is nothing particularly odd about any of the values in the variables other than that these are the two observations encompassing the 'Never' category under Education. We would probably combine the 'Never' dummy variable with the 'MS' category, one step up, but since our regression assumptions are not met it does not make sense to do so. Removing these points would probably cause a great shift to the beta coefficients since these are the only two observations.

On Figure 8, there are another two extreme points (Observation 1248 and 1414). Observation 1248 is someone who works 60 hours per week and sleeps 10 hours per day, but they have 30 days of poor

mental health. According to the model, this person should have much less days of poor mental health because of the negative betas associated with work and sleep which attributes to the high cook's d value. But we don't see anything wrong with his data so we will not remove that point. Observation 1414 is another person with 30 days of poor mental health. This person has 0 days of poor physical health and relatively normal values for everything else. Since physical health has a fairly large positive beta, this goes against the actual value versus the predicted value. Again, there is nothing wrong with the data so we will not remove this point.

**Conclusion:**

In the final model, we see that there several additional predictors we didn't know about before that may attribute to poor mental health such as age group and education level. There were some predictors that went against what we currently know (working more decreases number of days of poor mental health). However, since the underlying assumptions of multiple linear regression were not met, these conclusions are not valid. We can't draw any conclusions from our final model until they are addressed. There needs to be some non-linear transformations done to some or all of the variables in the final model in order to satisfy those assumptions (linear relationship, equal variance, and normality of residuals). A transformation of the response variable may also be needed. If we were to do the analysis again, we would combine the 'Never' dummy variable with the level above it, 'MS', since there are only two observations in it that turned out to be influential points.

The large sample size may have attributed to the significance of some of the variables when they are not actually a good predictor of poor mental health. In the future study, we would have a much smaller sampling size.

It is also worth nothing that the data may not be accurate since most of the values are reported by the respondent. There is no way to accurately gauge exactly how many days of poor mental health an individual experienced in the past month. For a future analysis, it may be a better idea to create a new study design and collect new data while finding a better way to gauge mental health. A possible idea would be to use a categorical response variable such as a variable that simply codes poor mental health or not as a binary variable. Instead of multiple linear regression, logistic regression would be used to analyze the data.

**References:**

1. Prince, Martin, Vikram Patel, Shekhar Saxena, Mario Maj, Joanna Maselko, Michael R. Philips, and Atif Rahman. "No Health without Mental Health."The Lancet 370.9590 (2007): 859-77. The Lancet. Elsevier, 4 Sept. 2007. Web. 26 Apr. 2016.

2. "What Is Mental Health?" MentalHealth.gov. U.S. Department of Health & Human Services. Web. 26 Apr. 2016. <https://www.mentalhealth.gov/basics/what-is-mental-health/>.

3. "Risks Factors for Poor Mental Health Wellness." Mental Health Wellness Week. Freedom From Fear, 2009. Web. 26 Apr. 2016. <http://www.mhww.org/risks.html>.