Xiaoyu Fu ISYE 6420 Apr. 21, 2023

Establishing a Bayesian Linear Regression Model for the Prediction of Mental Health of Chinese High School Students

1. Introduction

The importance of mental health in adolescence should not be underestimated. No family wants to experience the tragedy of losing a child to mental health problems. Factors such as obesity, family conditions, and personal habits may all affect mental health. Adolescence from different areas of the world may also be affected by different factors. For example, in China, because of the one child policy, currently the majority of teenagers are the only child of the household. Such a child has become a social norm so called “little emperor”. “Little emperors” have the whole attention of the whole family that they live under greater pressure, and their physical and mental health are of concern. To address this issue, Peng et al. (2019) conducted research to study the relationship between physiological factors such as obesity, family factors such as number of siblings, and mental health.

Keeping track of teenager’s mental health status might not be an easy thing to do for the school and healthcare providers. Compared with physical exams, a person’s mental health is very subjective. In most cases, it is difficult to be diagnosed by simply observing daily behavior. Diagnosing via self-reporting questionnaires also cannot always be accurate. However, testing other physical parameters such as weight, height, cholesterol level can easily get quantitative results that are relatively more objective. In addition, other social factors that might also have an impact on a teenager’s mental health can also be quantitatively represented. For example, in the particular study conducted by Peng et al. looked into the influence of the number of siblings.

To more accurately diagnose a teenager's mental health status with physical exam results, objective family status, and personal habits, based on the data collected by Peng et al., I used Bayesian linear regression to use different factors (fasting blood glucose, triglycerides, BMI, number of siblings, father’s accompaniment in elementary

school, mother's accompaniment in elementary school, father's educational level, mother's educational level, family financial status, sleeping hours, skipped breakfast) and created a model for the prediction of the depression and anxiety score.

2. The Data

Data was collected by Peng et al. from a random cohort of 1348 high school students in Guangzhou, China. The data was first processed by the authors to delete values that are abnormal, and it still contains missing values throughout. Not all assessment parameters are selected for this specific Bayesian analysis. Some redundant parameters contain information that can be better represented by another parameter also collected in the dataset. For example, the height and weight are not being analyzed separately because BMI could cover the information contained in the two variables.

For this specific analysis, the PYMC package was used for all Bayesian analysis.

3. The Process

3.1. Fitting each dataset into Normal distribution

For the first step to build the model, all the data are fitted into Normal distribution by Markov Chain Monte Carlo (MCMC) sampling to estimate the posterior distribution of the model parameters. The mu and sigma for each parameter are then copied into the Bayesian Linear Regression model for further use.

Specifically, here are the prior of each parameter:

Depression score (yi1) ~ N(38.894, 8.081)
Anxiety score (yi2) ~ N(32.829, sigma=6.282)
fasting blood glucose ~ N(4.6, 0.44)
triglycerides ~ N(4.2, 0.74)
BMI~ N(19, 2.9)
Number of Sibling ~N(1.436, 0.132)
father's accompaniment in elementary school ~ N(5, 0.8)
mother's accompaniment in elementary school ~ N(5, 0.8)

father's educational level ~ N(5.5, 4.7)
mother's educational level ~ N(4.7, 4.6)
family financial status ~ N(3, 0.5)
sleeping hours ~ N(2.3, 0.7)
skipping breakfast ~ N(1.3, 0.8)

3.2. Bayesian Linear Regression

To predict depression and anxiety levels, 11 parameters were proposed as the possible factors. The linear model is set as the following:

*Depression score = g0 + g5 * fasting blood glucose + g6 * triglycerides + g9 * BMI + g13 * number of siblings +g14 * father's accompaniment in elementary school + g15 * mother's accompaniment in elementary school + g16 * father's educational level + g17 * mother's educational level + g18 * family financial status + g19 * sleeping hours + g20 * skipping breakfast*

*Anxiety score = g0 + g5 * fasting blood glucose + g6 * triglycerides + g9 * BMI + g13 * number of siblings +g14 * father's accompaniment in elementary school + g15 * mother's accompaniment in elementary school + g16 * father's educational level + g17 * mother's educational level + g18 * family financial status + g19 * sleeping hours + g20 * skipping breakfast*

The dataset contains missing values. These missing values are imputed by the PYMC.

3.3. Posterior Prediction Evaluation

To validate the effectiveness of the model, posterior prediction evaluation was conducted for both models. Specifically, the package of posterior_predictive in PYMC was used to plot the posterior predictive results and the actual observation in the dataset.

4. Results

4.1. Results of Parameters

Tabel 1. Prediction parameter of depression score of linear regression

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk
g0	351.486	80.999	190.207	507.849	1.405	1.011	3312.0
g5	-17.221	10.748	-38.386	3.866	0.167	0.134	4151.0
g6	-60.546	6.691	-73.736	-47.308	0.078	0.057	7277.0
g9	8.755	1.632	5.354	11.814	0.026	0.019	3864.0
g13	-52.005	10.452	-72.993	-31.713	0.142	0.102	5386.0
g14	32.499	5.310	21.870	42.498	0.075	0.054	5084.0
g15	-41.667	6.653	-54.234	-27.793	0.090	0.065	5504.0
g16	-18.030	5.156	-28.100	-7.972	0.081	0.058	4055.0
g17	19.161	5.721	7.989	30.132	0.085	0.065	4559.0
g18	-9.024	7.104	-23.597	4.779	0.108	0.094	4263.0
g19	-6.599	8.625	-23.893	9.991	0.139	0.140	3842.0
g20	39.452	6.999	25.504	52.648	0.084	0.064	6931.0

Tabel 2. R hat of parameters of depression score of linear regression

	ess_tail	r_hat
g0	2262.0	1.0
g5	2377.0	1.0
g6	2089.0	1.0
g9	2437.0	1.0
g13	2258.0	1.0
g14	2484.0	1.0
g15	2327.0	1.0
g16	2438.0	1.0
g17	2394.0	1.0
g18	2140.0	1.0
g19	1925.0	1.0
g20	2021.0	1.0

Tabel 3. Prediction parameter of anxiety score of linear regression

	mean	sd	hdi_2.5%	hdi_97.5%	mcse_mean	mcse_sd	ess_bulk
g0	341.029	81.309	191.098	501.760	1.403	1.014	3372.0
g5	-17.515	10.698	-39.017	2.791	0.174	0.144	3810.0
g6	-59.732	6.590	-72.355	-47.009	0.065	0.047	10197.0
g9	8.829	1.708	5.393	12.101	0.026	0.019	4160.0
g13	-50.972	9.935	-70.370	-32.403	0.130	0.096	5769.0
g14	33.004	5.305	23.039	43.306	0.074	0.054	5163.0
g15	-41.848	6.668	-54.768	-29.509	0.093	0.066	5161.0
g16	-17.654	5.076	-27.307	-7.924	0.074	0.053	4658.0
g17	18.982	5.723	8.215	29.933	0.090	0.064	4069.0
g18	-8.850	7.004	-21.871	5.305	0.110	0.093	3988.0
g19	-7.639	8.626	-25.283	8.541	0.160	0.138	2889.0
g20	38.813	6.708	26.229	52.131	0.078	0.058	7388.0

Tabel 4. R hat of parameters of anxiety score of linear regression

	ess_tail	r_hat
g0	2236.0	1.0
g5	1990.0	1.0
g6	2531.0	1.0
g9	2629.0	1.0
g13	2575.0	1.0
g14	2604.0	1.0
g15	2403.0	1.0
g16	2502.0	1.0
g17	2291.0	1.0
g18	2291.0	1.0
g19	2017.0	1.0

g20	2085.0	1.0

Table 1 to 4 are the summarized results of Bayesian Linear Regression parameters for both depression score and anxiety score. From Table 2 and Table 4, the r_hat score for all parameters is 1.0, which indicates that all MCMC evaluations converge with large enough sampling numbers.

Taking a closer look into Table 1, the credible interval of g6, g9, g13, g14, g15, g16, g17, and g20 excludes 0, indicating that these parameters have a stronger effect on the predicted depression score. The top 3 parameters with the highest absolute value of the means are g6, g13, and g15, which are the parameters for triglycerides, number of siblings, mother's accompaniment in elementary school, respectively.

For anxiety score prediction, the same sets of the parameters, g6, g9, g13, g14, g15, g16, g17, and g20, excludes 0. The top 3 parameters with the highest absolute value of the means are g6, g13, and g15, which are the parameters for triglycerides, number of siblings, mother's accompaniment in elementary school, respectively. Both results are consistent with depression scores.

From the above analysis, we can conclude that among all the analyzed factors, the top 3 factors that have more impact on the mental health status are triglycerides value, number of siblings, and mother's accompaniment in elementary school. All three parameters are negative, indicating a negative relationship with the predicted value. Specifically, higher levels of triglycerides, more siblings, and more mother's accompaniment in elementary school results in lower depression and anxiety score.

4.2. Posterior Prediction Evaluation Result

The model prediction accuracy was analyzed by plotting posterior predictive results in blue lines, and the actual observed data in black lines.

Figure 1. Anxiety Model Posterior Prediction Evaluation

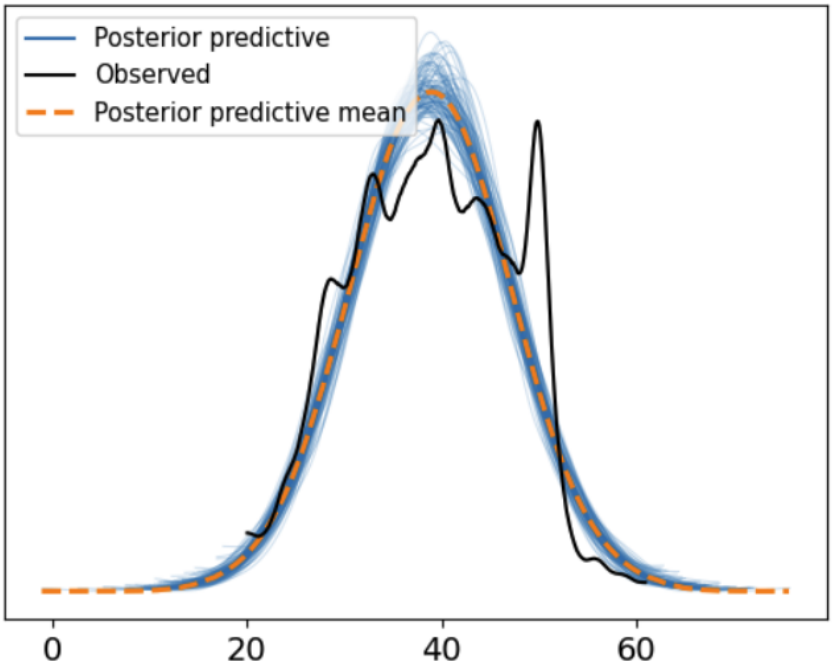
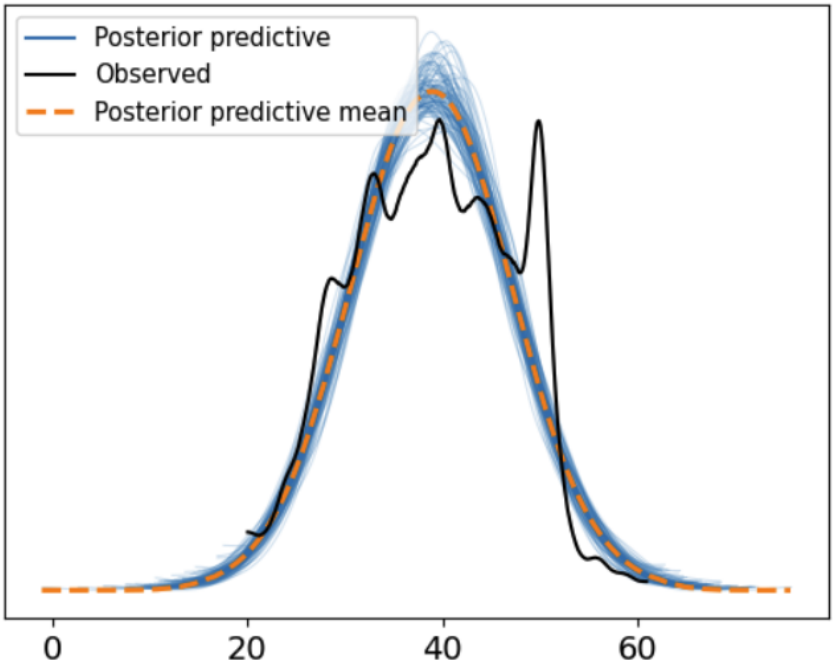


Figure 2. Depression Model Posterior Prediction Evaluation



For the setting of the analysis, both the observed data and parameters are fitted into Normal distribution. However, the observed data are not processed into a formatted

distribution. Nevertheless, the overall trend shows good fitting with observed data points, suggesting the effectiveness of the model.

5. Conclusion

In conclusion, the two proposed models can be used to effectively predict a high school student’s mental health status by objectively collecting the student's physical and family data. The models can be used as an aid when diagnosing a student's mental health status.

6. References

Peng et al. "Imbalance in Obesity and Mental Health among Little Emperors in China." Figshare, 2019, doi: 10.6084/m9.figshare.7977686.v1.