Predicting BMI in Chinese Adolescents with Screentime*

Math 261A Project 2

Robert Yav

December 10, 2025

As the screens used to browse the internet become more widespread within the modern era, tt may be of interest to predict BMI using screen time, controlling for factors such as age, gender, and household income. We explore the use of other predictor variables to more accurately predict BMI. We source data from the China Health and Nutrition Survey (CNHS) published in 2019, and motivation from a 2024 paper by Ma J. and Shang L. on the effect of internet use on body weight in Chinese adolescents. The sample is taken from a survey of Chinese adolescents and their demographic and health data. Three main approaches were used: OLS regression using a full and base model, a full model with interaction, a transformed model, and LASSO regression. The LASSO and interaction models performed the most favorably, but the final model chosen was the base model.

1 Introduction

As the world develops, internet infrastructure and its penetration has increased dramatically in the past few decades. Screens have become commonplace, and it is not rare to see even children glued to their screens in public. (Ma and Sheng 2024) raises concerns about this observation, claiming that a rise in internet use has significantly increased BMI and obesity rates in Chinese adolescents. A healthy BMI for an adult is 18.5-24.9, underweight is below 18.5, overweight is 25-29.9, and obese is 30.0 and above. For adolescents, BMI is calculated the same, but is interpreted differently by comparing BMI to other adolescents of the same age. According to the aforementioned study, one of the hypotheses claim internet use would increase sedentary activities, mainly screen time. This claim leads to the motivation of this project: How can screen time and other lifestyle factors predict BMI, controlling for age, gender, and household income? This project examines the relationship between BMI and combined screen

*Project repository available at: [https://github.com/rubertyao/ScreentimeBMI]

time for Chinese adolescents in 2015, controlling for external factors like age, sex, and income. We then go on to include additional potential variables, gathered from the hypotheses initially put forth by (Ma and Sheng 2024), adding three new predictors, time spent exercising, time spent sedentary, and time spent asleep. This paper builds on their limitations, addressing an important point that their predictions had not accounted for the intensity of time spent with physical and sedentary activities. Using time spent on a screen, time spent exercising, time spent sedentary, and time spent asleep as predictors, and age, sex, and income as control variables, we attempt to predict BMI scores for adolescents located in China in 2015. We go ahead and build and compare three main branches of models: OLS, Inverse Square Root Transformation, and LASSO regression. To compare model performance, we used k fold-cross validation and took average RMSE and R^2 across 10 folds for each model. The final model chosen was the base model due to close relative performance and interpretation simplicity.

The remainder of the paper are as follows. Section 2 presents the data, Section 3 presents the models used, Section 4 shows the results, and Section 5 discusses the implications of our findings.

2 Data

The data set used was sourced from (Ma and Sheng 2024) on Harvard Database which merged information from the CNHS in 2019 (ma 2024). This data set tracks survey information on adolescents from China from 2005 to 2015, surveying demographic, income, education, and health information. There were 3054 observations aged 11-19, 52% of which were male. For the purposes of this analysis, and due to limitations of scope, we limit the data set to 2015, which include 459 observations, 50% of which are female. Since many of the data points in time are divisible by 10, it is highly likely that the time predictors are rounded to the nearest hour or 10th minute. While this will reduce the accuracy of the model, it is still a good indicator of the intensity of time spent doing any particular activity. This also addresses a concern in (Ma and Sheng 2024), which used binary data to answer their question, not taking into account the intensity of how much internet a particular individual consumed in comparison with sleep and exercise. Initial data visualization dropped nine observations, two of which had NA values for household income and seven because of a survey inaccuracy that recorded more than 24 hours of screen or exercise time in a day. Since these values had no influence on the analysis, they could be removed for simplicity.

Looking through the data set, it is important to disclose that the merged data (Ma and Sheng 2024) provide to the Harvard Dataverse doesn't contain a data dictionary, and looking through the CNHS data sets do not have clear definitions for the obtained variable names either. This leads us to believe that the study renamed certain columns when merging the data sets, which have not been shared. However, we can infer what certain columns are due to their names and values. The main variables we use as a base model in our study is as follows:

bmi: The Body Mass Index (BMI) of an adolescent.

screen time: The average amount of combined screen time an adolescent receives, calculated from previous study by summing TV, video game, and computer screen times.

age: the age of an individual.

gender: the gender of an individual.

ln(income): the log scaled income of an individual's family. Since this is the only variable for income and is generally right skewed, it is best to leave this variable transformed.

2.1 Hypotheses

(Ma and Sheng 2024) make four hypotheses on what can increase body weight with internet use. By using these hypotheses, we can obtain potential variables that can be used in variable selection to improve prediction accuracy in our model.

2.1.1 Hypothesis 1: Internet use increases the body weight in adolescents through sedentary activities.

exercise time: The average amount of combined physical activity, in exercise time, calculated in minutes.

sedentary time: The average amount of combined sedentary activity time, calculated in minutes.

2.1.2 Hypothesis 2: Internet use increases the body weight by crowding out sleep time.

sleep time: The average amount of sleep time, calculated in hours.

2.1.3 Hypothesis 3: Internet use increases the body weight through adverse health behaviors.

This includes health education, along with alcohol and cigarette consumption. Following exploratory analysis, the variables associated with alcohol and cigarette consumption contain mainly N/A values, which raises issues in the analysis. Since it is illegal for a child to smoke or drink in China, and the legal drinking and smoking age is above the age limit for the average child in the study, it is likely that the values obtained do not accurately represent the population. For health, since there is no data dictionary and the health variable seems to be an arbitrary score, it is in the analysis' best interest to leave it out of prediction, pending further clarification from the original study. Due to these concerns, we drop this hypothesis for future studies to consider.

2.1.4 Hypothesis 4: Internet use increases the body weight through food consumption behavior.

This includes junk food, tv, vegetable, and fruit food consumption scores. Since these scores also contain majority null values, we drop this hypothesis for future studies as well.

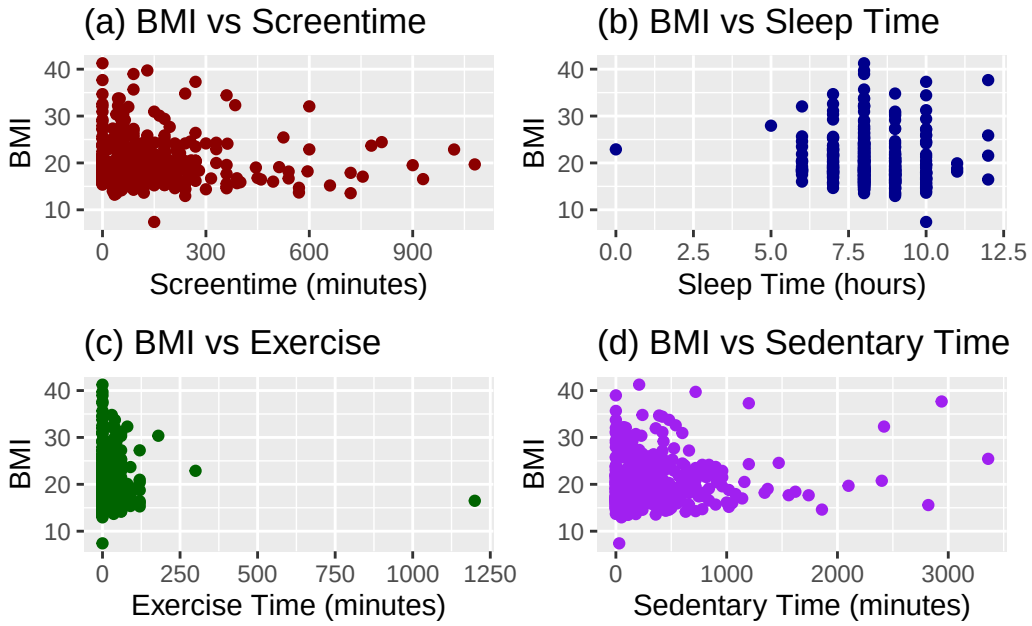

Figure 1: Initial Data visualization, Scatter plot of (a) BMI as explained by Screentime, (b) BMI as explained by Sleep Time, (c) BMI as explained by Exercise Time, and (d) BMI as explained by Sedentary Time

3 Methods

3.1 Base Model

We fit the base multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 ln(X_{4i}) + \varepsilon_i$$

β_0 Intercept

β_1 represents the expected point change in BMI per point change in minutes spent on a screen a day holding other variables constant.

β_2 represents the expected point change in BMI per point change in age while holding other variables constant.

β_3 represents the expected point change in BMI for a male while holding other variables constant.

β_4 represents the expected point change in BMI per point change in log household income while holding other variables constant.

X_{ji} represents the i^{th} adolescent's j^{th} predictor in the model. Therefore, X_{1i} is the i^{th} adolescent's minutes in screen time, X_{2i} is the i^{th} adolescent's age, X_{3i} is the i^{th} adolescent's gender,and X_{4i} is the i^{th} adolescent's log household income.

Y_i represents the i^{th} adolescent's BMI as a response in the model.

ϵ_i represents independent, uncorrelated, normally distributed error terms with mean 0 and constant variance σ^2.

For inference, initial visualizations of the errors with a residual versus fitted values plot showed no clear violations of homoskedasticity. A visual of the qqplot shows tails at the end of the qqplots, which is an indication of a violation in the normality of errors. With a high enough sample size, this doesn't present as a problem at this time. Since our goal is prediction, the assumption we care about the most is linearity, which upon closer look using a residual vs fitted values plot, shows no clear violation.

3.2 Full Model

We compare the base model with the full model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 ln(X_{4i}) + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \varepsilon_i$$

β_5 represents the expected point change in BMI per point change in minutes spent exercising a day while holding other variables constant.

β_6 represents the expected point change in BMI per point change in spent sedentary a day while holding other variables constant.

β_7 represents the expected point change in BMI per point change in hours spent asleep a day while holding other variables constant.

X_{5i} is the i^{th} adolescent's minutes spent exercising, X_{6i} is the i^{th} adolescent's minutes spent sedentary, and X_{7i} is the i^{th} adolescent's hours spent asleep.

3.2.1 Full Model with Interactions

We consider interactions between time spent on a screen with time spent exercising and time spent sleeping. We consider these interactions because Hypothesis 1 and 2 from (Ma and Sheng 2024) reason that higher screen time "crowds out" potential time spent sleeping and exercising, thereby reducing their respective totals and increasing weight gain.

We consider the model

$$Y_i = \beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i}+\beta_4 ln(X_{4i})+\beta_5 X_{5i}+\beta_6 X_{6i}+\beta_7 X_{7i}+\beta_8 (X_{1i}X_{6i})+\beta_9 (X_{1i}X_{7i})+\varepsilon_i$$

Where β_8 represents the interaction effect between minutes spent on a screen and minutes spent exercising and β_9 represents the interaction effect between minutes spent on a screen and hours spent sleeping.

3.3 Inverse Square Root Transformed Models

We attempt boxcox transformation in order to justify an appropriate transformation for better prediction.

Boxcox transformations take the form

$$y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda}$$

when $\lambda \neq 0$, and

$$y_i^{(\lambda)} = ln(y_i)$$

when $\lambda = 0$.

Boxcox transformation found the optimal result to be $\lambda = \frac{-2}{3}$.

This corresponds to a transformation of the form $\frac{1}{\sqrt{Y}}$.

Therefore, we can compare model performance of both the transformed base and full models.

3.3.1 Inverse Square Root Transformed Base Model:

$$\frac{1}{\sqrt{Y_i}} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 ln(X_{4i}) + \varepsilon_i$$

3.3.2 Inverse Square Root Transformed Full Model:

$$\frac{1}{\sqrt{Y_i}} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 ln(X_{4i}) + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \varepsilon_i$$

Where Y_i is transformed, so each slope interpretation β is modified to an expected inverse square root point change in BMI, holding all other variables constant.

3.4 LASSO Model

We implement Least Absolute Squares Selection Operator (LASSO) regression, which introduces a tuning parameter λ that penalizes additional parameter terms. We select λ and use k-fold cross validation with 10 folds. LASSO essentially performs variable selection by penalizing parameter values to 0.

For LASSO regression, we select parameter values that minimize

$$\sum_{i=1}^{n}(\mathbf{Y}_i - \mathbf{X}_i^T)^2 + \lambda \sum_{j=1}^{p-1}|\beta_j|$$

where λ is a tuning parameter. LASSO found $\lambda = 0.3$, and an Inverse Transformed LASSO found $\lambda = 1$. Upon further inspection, the full model's choice of parameter $\lambda = 0.3$ set almost all β slope coefficients to 0 but age, and the transformed model's choice of $\lambda = 1$ only leaves the intercept. While RMSE values are still reported in the final table, we drop LASSO from model consideration due to a lack of interpretability.

3.5 Cross Validation

By performing k-fold cross validation, we can estimate model performance by comparing RMSE and R^2 values. k-fold Cross validation splits the data into k folds, and uses all but one fold of our training data to fit the model, and uses the final fold to test the model and make predictions. The process is then repeated for all folds and the average fold root mean squared error across k folds calculated is as follows, where Y is the observed value and $\hat{y}$ is the predicted value:

$$RMSE = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(Y_j - \hat{Y}_j)^2}$$

In the cases where models used transformations on the response variable, it was necessary to transform the values back to compare RMSEs across untransformed and transformed values.

$$RMSE_{adj} = \frac{1}{k}\sum_{i=1}^{k}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(Y_j^{-2} - \hat{Y}_j^{-2})^2}$$

We also take a look at the coefficient of multiple determination, or R^2 to assess the model performance, which is the proportionate reduction in variation in Y associated with the use of a particular model.

$$R^2 = \frac{SSR}{SSTO}$$

Where SSR is the sum of squares regression and SSTO is the total sum of squares.

3.6 Limitations of Analysis

It is important to note that this survey data is not completely reliable. For example, the time in minutes recorded is most likely approximated and not the actual time spent doing various activities (eg., sleeping, screen time, physical activity). A null value majority in the smoking, drinking, and food consumption categories have prevented the use of those variables in the model, but can likely be imputed with estimates which fall outside of the scope of this analysis. [ma_effect_2024] uses the full dataset, which includes data from 2005-2015. Since the scope of the analysis is only on 2015, a majority of the information in the dataset is lost. Additionally, no data dictionary has been supplied, therefore many of the variable values needed to be interpreted, since some of the merged variable names have been renamed by the study and definitions of them are not available on the Harvard Database or the CHS database.

3.7 Software

We implement the analysis using R (R Core Team 2025). Plots done with (Wickham 2016) and (Pedersen 2025). Data manipulation done with (Wickham et al. 2019). Boxcox transformations done with (Venables and Ripley 2002). Predictive modeling and cross validation done with (Kuhn and Max 2008). Lasso and Ridge Regression done with (Tay, Narasimhan, and Hastie 2023) Tables done with (Xie 2025).

4 Results

Below is a table to the cross validation results for every model considered in the final model selection.

Model	RMSE	RMSEadj	RSquared
Base Model	4.3610	0.0000	0.1082
Full Model	4.4530	0.0000	0.0869
Full Model with Interaction	4.4377	0.0000	0.0934
Transformed Base Model	0.0216	4.3807	0.1301
Transformed Full Model	0.0217	4.4184	0.1141
Transformed Full Model with Interaction	0.0219	4.4359	0.1209
Lasso Model	4.3998	0.0000	0.1078
Transformed Lasso Model	0.0233	4.6985	0.0036

Figure 2: Summary table of cross validation results, reporting the RMSE, adjusted RMSE for transformation, and the coefficient of multiple determination.

When checking the model performances in the table under different seeds, RMSE values for each of the models fluctuated, changing interpretation of which model performed the most favorably. Since RMSE only changed minimally throughout the models, it is likely that performance does not change drastically between models. However, the Base Model, Full Model with Interaction, and LASSO Models often performed the best in terms of RMSE. Using a transformation often decreased RMSE from the full model, and performing LASSO with a transformed response variable often performed the worst.

Generally, model performance was low in terms of R^2. This is evident in how model selection under different seeds changed, as each "best" model performed equally as poorly to each other. Note that under back-transformed RMSE, the non transformed models are 0, since these models did not need to be adjusted for transformed error.

4.1 Final Model

We compare OLS vs Transformation vs LASSO while including interaction terms. It was initially found that the LASSO model found the best root mean square error using cross validation. However, as mentioned in Section 3, the algorithm chose a final model that only included age, which drops the main predictor we are interested in. Using other methods with the full model performed worse, with only slightly increased performance in R^2. It is important to note that the RMSE minimally changed between models, so each method only slightly changes performance. We choose the base model over the LASSO model because it sacrifices the least in interpretability while also maintaining relatively similar performance. The estimated intercept term is 10.2, which has little interpretability, but is technically possible: a female child born within the year that has no screen time and no family income. The estimated slope parameter for screen time is -.0013, which represents an expected -0.0013 change in BMI for each additional point change in screen time, holding all other variables constant. The estimated slope parameter for age is 0.6395, meaning that for each point

change in age, the expected increase in BMI is .6395, holding all other variables constant. The estimated slope parameter for gender is .7414, meaning that for a boy, the expected BMI is to be increased by .7414, holding all other variables constant. The estimated slope parameter for log income is 0.089, meaning that for each point increase in log income, BMI is expected to raise by 0.089, holding all other variables constant.

5 Discussion

Although we have selected the model that most accurately predicts bmi in our analysis, since all of the R^2 values for our models were low, all models performed poorly overall. Note that comparing R^2 values for the transformed models to the other models cannot be performed. In practice, we could not accurately predict BMI using screentime, age, sex, income, exercise time, sedentary time, sleep time, and age. This was hinted when the LASSO variable selection dropped almost all of the predictor variables, showing a low predictive accuracy when using the variables. Since the RMSE values were all greater than 4, and the BMI ranges for underweight, healthy, overweight, and obese are all around 6.5, our models have a strong likelihood of misclassifying, for example a healthy adolescent as overweight. (Ma and Sheng 2024) used a binary variable as their main predictor in their empirical analysis, while the extension done by this paper uses a continuous variable, screen time. This distinction could explain the negative coefficient found for screen time, whereas in the previous study, internet use was found to significantly increase BMI.

5.1 Potential Weaknesses and Extensions

Data values were inferred due to a lack of a data dictionary. This makes the report prone to error from data interpretation. Since the data was sourced from a national survey, there are many sources of data recording error, which we treat at the high end, where time values exceeded possibly daily limits. Inaccuracies towards the lower bound were not considered. Another weakness was that we did not consider scaling and centering the predictors. Standardizing variables can help avoid round off errors and make comparisons between estimated slope parameters easier.

One possible extension is the inclusion of Hypothesis 3 and 4 in Section 2. We did not include these hypotheses in the model synthesizing process due to the large amount of N/A values in the rows, but with interpolation, further variable selection could be performed. Since this analysis only included adolescents from 2015, the model could be extended towards children in the other years, controlling for individual fixed effects.

References

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

ma, junqi. 2024. "Internet use and Chinese adolescents' body weight." Harvard Dataverse. https://doi.org/10.7910/DVN/FTTF6A.

Ma, Junqi, and Li Sheng. 2024. "The Effect of Internet Use on Body Weight in Chinese Adolescents: Evidence from a Nationally Longitudinal Survey." *PLOS ONE* 19 (12): e0311996. https://doi.org/10.1371/journal.pone.0311996.

Pedersen, Thomas Lin. 2025. *Patchwork: The Composer of Plots*. https://doi.org/10.32614/CRAN.package.patchwork.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie. 2023. "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2025. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.