An Analysis of Screentime's effect on BMI*

Math 261A Project 2

Robert Yav

December 6, 2025

The purpose of this paper is to analyze the relationship between BMI and screentime, 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 includes 459 observations of Chinese adolescents and their demographic and health data.

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 J.] raises concerns about this observation, claiming that a rise in internet use has significantly increased BMI and obesity rates in Chinese adolescents. 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 time for Chinese adolescents in 2015, controlling for external factors like age, sex, and income.

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.

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

Data

The data set used was sourced from Ma K. and Shang L. in 2024 (Ma and Sheng 2024) 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 three observations, 2 of which had NA values for household income and 1 because of a survey inaccuracy that recorded more than 24 hours of screen 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 J. and Shang L. (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.

Ma K. and Shang L. (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. 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. Hypothesis 2: Internet use increases the body weight by crowding out sleep time. sleep time: The average amount of sleep time, calculated in hours. 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. 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.

Methods

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 average minutes spent on a screen 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.

$X_j i$ represents the i^{th} adolescent's j^{th} predictor in the model.

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.

```
#plot(bmi_fit,which=2)
```

Initial visualizations of the errors 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.

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 average minutes spent exercising while holding other variables constant.

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

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

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

Boxcox transformations take the form

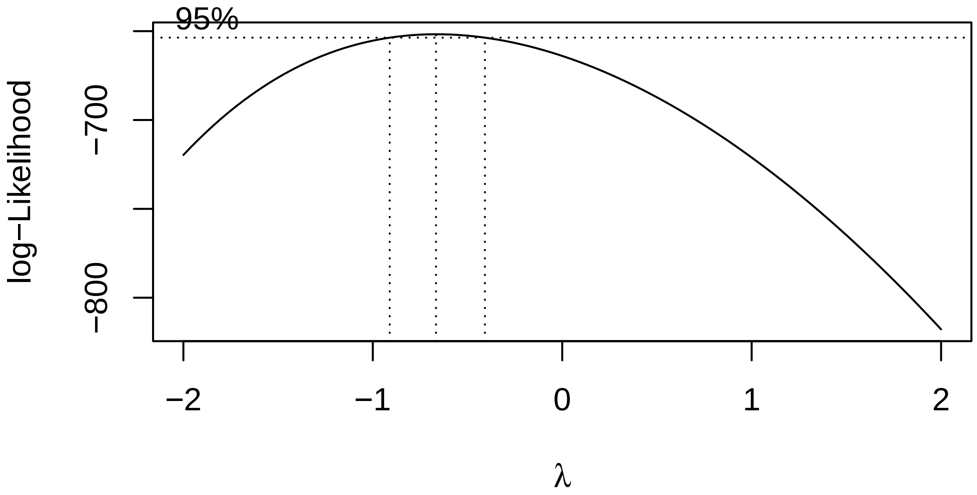
$$y_i^{(\lambda)} = \frac{y_i^{\lambda} - 1}{\lambda}$$

when $\lambda \neq 0$, and

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

when $\lambda = 0$.

```
bc_result=boxcox(bmi_fit)
```



```
optimal_lambda_index = which.max(bc_result$y)
bc_result$x[optimal_lambda_index]
```

```
[1] -0.6666667
```

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. Transformed base model:

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

```
tbmi_fit=lm((1/sqrt(bmi))~screeTIME+age+gender+lnincome, data = bmi_data)
#plot(tbmi_fit)
```

Transformed full model:

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

We implement Ridge and LASSO regression, both of which introduces a tuning parameter λ that penalizes additional parameter terms. We select λ and use k-fold cross validation with 10 folds.

For ridge regression, we select parameter values that minimize

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

For LASSO regression, we select parameter values that minimize

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

where λ is a tuning parameter. LASSO essentially performs variable selection by penalizing parameter values to 0. Since ridge regression found the best performing lambda value $= 0$, we drop the model since it is statistically the same as the full model. LASSO found λ=.5.

```
train_control <- trainControl(method = "cv", number = 10,savePredictions="final")
r_model = train(bmi~screeTIME+age+gender+lnincome+t_exe+all_seditry+sleep, data = bmi_data, m
l_model = train(bmi~screeTIME+age+gender+lnincome+t_exe+all_seditry+sleep, data = bmi_data, m
```

By performing k-fold cross validation, we can estimate model performance by comparing RMSE and R^2 values.

```
b_model = train(bmi~screeTIME+age+gender+lnincome, data = bmi_data, method = "lm", trControl
f_model = train(bmi~screeTIME+age+gender+lnincome+t_exe+all_seditry+sleep, data = bmi_data, m
bt_model = train((1/sqrt(bmi))~screeTIME+age+gender+lnincome, data = bmi_data, method = "lm"
ft_model = train((1/sqrt(bmi))~screeTIME+age+gender+lnincome+t_exe+all_seditry+sleep, data =
fti_model = train((1/sqrt(bmi))~screeTIME+age+gender+lnincome+t_exe+all_seditry+sleep+screeTI
lti_model = train(1/sqrt(bmi)~screeTIME+age+gender+lnincome, data = bmi_data, method = "lasso

#For RMSE, switching for each model
#print(b_model)
#print(f_model)
#print(bt_model)
#print(ft_model)
#print(fti_model)
#print(l_model)
#print(lti_model)

#For Untransformed RMSE
```

```r
preds=lti_model$pred #Switching for each model
sqrt(mean((preds$pred^(-2)-preds$obs^(-2))^2))
```

```
[1] 4.591553
```

Base Model: RMSE 4.4538 R^2 .1331

Full Model: RMSE 4.4943 R^2 .1103

Transformed Base Model: RMSE 0.0220 Un-transformed RMSE 4.4648 R^2 0.1324

Inverse Square Root Transformed Full Model: RMSE 0.0222 Un-transformed RMSE 4.5681 R^2 0.0173

Inverse Square Root Transformed Full Model with Interaction between screentime and exercise, screentime and sleep: RMSE 0.0229 Un-transformed RMSE 4.7318 R^2 0.0176

LASSO with $\lambda = 0.5$: RMSE 4.4751 R^2 0.1257

LASSO with Inverse Square Root Transformation and $\lambda = 0.9$: RMSE 0.0220 Un-transformed RMSE 4.5668 R^2 0.1448

We implement the analysis using R (R Core Team 2025).

Results

We compare OLS vs Transformation vs LASSO while including interaction terms. As it turns out, the simplest model found the best root mean square error using cross validation, with RMSE 4.4538. LASSO with an inverse square root transformation and $\lambda = 0.9$ found the highest R^2 value of .1448.

Discussion

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). As discussed in the data section, we could not

(Previous study uses internet use Y/N while we use screentime as main predictor.) REMOVE LATER # References

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.

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