Investigating Behavioral Factors Associated with Adult Obesity
A STAT 139 Final Project

Yuyue Wang, Xiangru Shu, Michelle (Chia Chi) Ho, Chengye Liu

December 13, 2017

Abstract

Obesity is an epidemic in the U.S. with significant health, economic and social costs. Elucidation of obesity risk factors can potentially aid in alleviating the pandemic. In this study, we aim to focus on investigating behavioral factors that are fairly easily controlled with the ultimate goal of calling for action to change the identified behaviors. Using the NHANES 2013-2014 data, the main factors we studied are related to alcohol use, TV consumption and selected aspects of eating habits. Using multiple linear regression to control for 12 potential confounders including race, age, marital status, smoking status and sleep hours, we found that the frequency of alcohol consumption (measured by number of days in a year a person consumes alcoholic beverage) has a slight but statistically significant negative association with obesity. Furthermore, in general, more TV consumption, more milk consumption, and more money spent on eating out are associated with higher BMI. In addition, Mexican Americans and the female gender are more prone to increases in BMI for every additional fast food meal compared to certain other races and the male gender, respectively. We also used multinomial logistic regression to build a best predictive model for predicting obesity levels ("underweight": BMI<18.5, "healthy": 18.5<BMI<25, "overweight": 25<BMI<30, "obese": BMI>30) with a cross-validated classification accuracy of 0.4904.

Introduction

Obesity in America is a concerning issue. Defined as BMI (body mass index) greater than 30 kg/m^2, obesity is linked to a myriad of human illnesses, including type II diabetes (Yaturu, 2011), certain types of cancer (Yaturu, 2013) and cardiovascular diseases (Reaven, 2008). Furthermore, there are substantial economic and social costs associated with being obese (Wang et al, 2011; Centers for Disease Control and Prevention, 2017; Wellman, 2002). Regrettably, previous studies have shown that obesity among adults in the United States has continued to increase in the past decade, leading to an alarming prevalence of 39.8% in 2015-2016 (Flegal et al, 2016; Centers for Disease Control and Prevention, 2017).

Much effort has gone into understanding risk factors associated with obesity. Both inherent (e.g. genetics, race) and acquired (e.g. sleep deprivation and smoking habits) factors have been identified (Clement et al, 2002; Patel et al, 2008; Flegal, 2007; Flegal et al, 2016). In this study, we aim to investigate and uncover additional obesity-linked behavioral factors with the ultimate goal of raising self-awareness and impetus to alter these behaviors and thus lower the overall prevalence of obesity in America. Using National Health and Nutrition Examination Survey

(NHANES) 2013-2014 data, we specifically wish to examine 3 types of behaviors that people can fairly easily control in their life:

1) Alcohol use
2) TV consumption
3) Selected aspects of eating habits, including milk consumption, money spent on grocery versus eating out or delivered food, number of fast food meals and number of meals not home prepared

Methods

Data Description. The NHANES 2013-2014 data[1] is a nationally representative sample of the noninstitutionalized U.S. civilian population. While NHANES does host data on its website, the data comes in numerous sub files. Therefore, we instead utilized a partially cleaned and merged version of this data on Kaggle.com[2], which contains thousands of variables across 6 files: demographics, diet, examination, labs, medication and questionnaire. We narrowed down to a list of 21 predictor variables, which are either confounding factors we wish to control (e.g. demographic attributes like race, age or education) or behavior factors that we wish to investigate, and a response variable, BMI (See **Table 1** for a complete list of variables and their descriptions).

Data Preprocessing and Feature Engineering. The variables we chose to include in this study reside in demographics, examination and questionnaire files. Therefore, the first thing we did was to merge data in these files based on the SEQN field, an identification number for each participant, and select the columns of interest. Next, we eliminated observations that have missing values or special values (i.e. "refused" or "don't know") in order to facilitate our analysis, and we discuss the limitations of choosing to do so in the Limitations section. Since alcohol use is an interest of this study, we chose to only include adults aged 21 and above in our analysis. We then condensed and re-grouped certain categorical variables for data cleanliness and to reduce unnecessarily large number of category levels. For instance, the 5 original categories of education (less than 9th grade, 9-11th grade, high school graduate/GED or equivalent, some college or AA degree, and college graduate or above) were re-grouped to high school or below, high school graduate, and high school or above. We also converted categorical variables to factors in R and renamed the levels, which are originally coded in numbers, to more intuitive names. Finally, we added a new feature, "*bmi_class*", to label individuals as underweight, healthy, overweight, or obese if their BMI < 18.5, or 18.5 <= BMI < 25, or 25 <= BMI < 30, or BMI >= 30, respectively. This new feature is to be used as the response variable in

[1] "NHANES 2013-2014 - Centers for Disease Control and Prevention" https://wwwn.cdc.gov/nchs/nhanes/search/nhanes13_14.aspx. Accessed 11 Dec. 2017.
[2] "National Health and Nutrition Examination Survey | Kaggle." 26 Jan. 2017, https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey. Accessed 11 Dec. 2017.

multinomial logistic regression (See **Table 2** for a complete list of preprocessed variables and their descriptions).

Exploratory Data Analysis (EDA) and Data Transformation. We visually inspected the distributions of the response and numeric predictor variables (histogram plots) to determine whether transformations are needed to correct any skewness. We also examined possible multicollinearity between numeric predictors using pair plots. Finally, we explored the relationship between the response versus categorical variables and numeric predictors using box plots and scatter plots, respectively.

Models, Model Assumptions and Comparisons. We used two categories of models to investigate the relationship between predictor variables and obesity: 1) multiple linear regression using BMI as response; and 2) multinomial logistic regression using obese levels (*bmi_class* as described above) as response. The first type of model allows simpler and more direct interpretability while the second may improve predictability by turning the problem into a more discrete one. For linear regression models, we first built a baseline model consisted of all main effects. We then explored several variations for this baseline model. Our strategy was to iteratively compare whether a specific set of additional polynomial terms or interaction terms improve the current best predictive model using Extra-Sum-of-Squares F test. We fitted a total of 8 linear regression models (details in Results section), which we visually examined model assumptions by a residual vs. fitted scatter plot and a QQ plot. For the best predictive regression model (as defined by the best out-of-sample model performance and estimated by cross-validation MSE), we used bootstrap to perform inference on the fitted coefficients. Separately, we fitted lasso, ridge and elastic net models in attempt to reduce overfitting and improve model predictive ability. For logistic regression, we fitted a baseline multinomial model with all predictors, two additional multinomial models with the predictors deemed important in linear regression and finally an ordinal model based on the best multinomial model (as measured by cross-validated prediction accuracy).

Results

Exploratory Data Analysis (EDA). We first checked the distribution of the response variable (BMI) to determine if there is any skewness to be corrected with transformation. As shown in **Figure 1**, the distribution of BMI is fairly right skewed (Figure 1, left), and such right-skewness can be corrected by applying the log transformation twice (Figure 1, right). Note that we decided to sacrifice some interpretability by choosing log(log(BMI)) (Figure 1, right) over log(BMI) (Figure 1, middle) for a more symmetrically distributed response variable. We will henceforth refer to log(log(BMI)) as BMI or our response variable for brevity unless otherwise specified.

We then inspected the distribution of the numeric predictor variables for any extreme outliers (**Figure 2**) and attempted to find possible transformations to reduce outliers (**Figure 3**). Specifically, we found that alcohol frequency (alcoholfrq), money spent on grocery (grocery), eating out (eatout), delivered food (delivery), number of meals not home prepared (meals_nothome) and number of fast food meals (meals_fastfood) all have outliers and extreme right-skewness (**Figure 2**). However, as **Figure 3** shows, not all skewness can be rescued by log transformation, so we chose to only transform those that showed improved symmetry (i.e. grocery, eatout and delivery) and leave the rest untransformed for better interpretability.

Thirdly, we visualized correlations between predictor variables and found that there exist minimal multicollinearity with the exception between meals_nothome and meals_fastfood, which have a correlation of 0.611 (**Figure 4**). We chose to still include both predictors.

Finally, we examined the association between predictor variables and the response variable. According to **Figure 5**, most of the categorical variables are at least weakly associated with BMI. For instance, Asians appear to have the least mean BMI compared to other races; individuals from top earning families may have lower BMI compared to families with lower income; people who identify themselves as often or always being depressed on average have a higher BMI than those who only sometimes or even never feel depressed; and people who are often and always struggling with sleep troubles have higher BMI than those who struggle less. These trends are consistent with previous reports (Clement et al, 2002; Patel et al, 2008; Flegal, 2007; Flegal et al, 2016). Furthermore, among those who do watch TV, more TV watching appears to be associated with higher BMI. Interestingly, people who drink at least 12 alcoholic beverages a year tend to have lower BMI than people who do not. For numeric predictors, sleep_hr appears to show a slight curvature that may indicate a quadratic relationship with the response (**Figure 6**). Other predictors show no visually striking linear relationship with the response in these marginal plots where no other confounding factors are controlled (**Figure 6**). We formally explore these potential associations or the lack thereof in our regression models in the following sections.

<u>Multiple Linear Regression Model Performance, Interpretation, Assumptions and Comparisons.</u>

Variations of Baseline Model With Forward/Backward Selection

Our baseline linear regression model (**model1**) included all the main effects without variable selection.

```
model1 = lm(log2.bmi ~ gender + age + race + edu + marriage + famsize + famincome +
alcohol12yr + alcoholfrq + log.grocery + log.eatout + log.delivery + milk + meals_nothome +
meals_fastfood + depressed + sleep_trouble + activity + tv_hrs + sleep_hr + smoke,
data=df_adult)
```

Our baseline model had an R^2 of 0.1153, mean cross-validation R^2 of 0.0881 (which is an estimate for out-of-sample performance), and cross-validation MSE of 0.0645.

To eliminate unnecessary predictors, we used backward selection with AIC as selection criterion to construct **model1a**. Model1a had an R^2 of 0.1128, mean cross-validation R^2 of 0.0904 (an estimate for out-of-sample performance), and cross-validation MSE of 0.0644. Extra-Sum-of-Squares F test was performed to investigate whether model1a significantly improves predictive capabilities. The F-statistic comparing model1 and model1a was calculated to be 1.0044 with a corresponding p-value of 0.4303, suggesting that model1 is not significantly better than model1a by including more predictors. We proceeded our modeling based on model1a.

It appears intuitive that both sleep deprivation and excessive sleep may be associated with heavier weight, potentially forming a quadratic relationship; therefore, we added a squared term (sleep_hr2) to model1a to construct **model1b**. Model1b had an R^2 of 0.1157, mean cross-validation R^2 of 0.0925, and cross-validation MSE of 0.0643. The F-statistic comparing model1a and model1b was calculated to be 9.5821 with a corresponding p-value of 0.0020, suggesting that model1b is significantly better than model1a. Therefore, we updated model1b as our basis of our modeling

In addition, since family size (famsize) could have a bearing other predictor variables such as family income (famincome), amount of money spent on grocery (log.grocery), eating out (log.eatout) or delivered food (log.delivery), we fitted **model1c** by adding famsize*famincome, famsize*log.grocery, famsize*log.eatout and famsize*log.delivery interactions to model1b. Model1c had an R^2 of 0.1186, mean cross-validation R^2 of 0.0901 (an estimate for out-of-sample performance), and cross-validation MSE of 0.0645. The F-statistic comparing model1c and model1b was calculated to be 0.9365 with a corresponding p-value of 0.4981, suggesting that model1c is not significantly better than model1b.

To fully investigate the effect of interaction between variables, we re-fitted model1b with all two-way interactions. We conducted a step forward selection from an intercept only model using AIC as selection criterion to obtain **model1d**. Model1d had an R^2 of 0.1705, mean cross-validation R^2 of 0.1166 , and cross-validation MSE of 0.0636. The F-statistic for comparing model1d and model1b was calculated to be 3.6427 with a corresponding p-value of essentially 0, suggesting that model1d is significantly better than model1b. Therefore, our best linear regression model is model1d.

Performance of baseline model and its variations are summarized in **Table 3**; summary of all Extra-Sum-of-Squares F-test can be found in **Table 4**, and summary of the best model, model1d, can be found in **Appendix I**.

For each of these linear regression models, we checked model assumptions visually using a residual vs. fitted scatter plot and a QQ plot (**Figure 7-Figure 11**). All five models show very similar diagnostic patterns. First, the residuals on the QQ plots show minimal right-skewness, so there is no obvious violation of normality. In addition, based on the scatter plots, no apparent overall curvature is observed. However, almost every scatter plot has fanning patterns, which suggests that the assumption of constant variance is violated. Intuitively, the observations are not completely independent from each other, which we discuss in the Limitations section.

Interpretation of the Best Forward-Selected Predictive Linear Regression Model

Given the main assumption violation is non-constant variance, we used bootstrap to determine predictor significance in the best forward-selected predictive linear regression model, which is model1d from discussion above. From 500 simulation iterations, we found the following predictors have a 95% confidence interval that did not capture 0 and hence signal predictor significance: 1) *race asian*; 2) *tv_hrs 2~4hrs*; 3) *smoke no*; 4) *smoke missing*; 5) *milk sometimes*; 6) *milk varied*; 7) *famincome 25k~55k*; 8) *gender female*; 9) *log.eatout*; 10) *activity no*; 11) *alcoholfrq*; 12) *tv_hrs 5 hrs: depressed often*; 13) *tv_hrs 0 hrs: depressed always*; 14) *famincome above 100k: gender female*; 15) *race asian: gender female*; 16) *smoke missing: gender female*; 17) *race black: activity no*; 18) *race asian: activity no*; 18) *race black: meals_fastfood*; 19) *race other: meals_fastfood*; 20) *gender female: meals_fastfood*; and 21) *depressed sometimes: milk varied*.

We found many associations consistent with previous reports and our exploratory analysis. For instance, Asian Americans have significantly lower BMI compared to Mexican Americans with a difference of -2.578e-02 (p-value = 0.010872), and non-smokers have significantly higher BMI compared to smokers with a difference of 3.052e-02 (p-value = 1.60e-07).

Several behavioral factors of interest were also found to be significantly associated obesity. First, alcoholfrq (i.e. number of days participants drank alcohol in the past year) was found to be negatively associated with obesity with a slope of -1.759e-04, when controlling for the other variables in the model. In terms of TV consumption (tv_hrs), we found that watching the TV for 2~4 hours (tv_hrs2~4hrs) on averge is associated with 2.176e-02 increase in BMI compared to watching the TV for 1 hour only. Interestingly, while the main effect of tv_hrs5hrs is not significant, its significance is manifested in its interaction term with depressed often (i.e. tv_hrs5hrs:depressedoften). The positive coefficient with a magnitude of 7.323e-02 indicate that

often depressed people who consume 5 or more hours of TV, on average, have higher BMI than people are less often depressed and consume less TV, controlling for other predictors.

Eating habits that are significantly linked to BMI includes milk consumption in the past month, money spent on eating out in the past month, and number of fast food or pizza meals eaten in the past week. Specifically, milk assumption (sometimes and varied) both have a positive association with BMI, with coefficients of 8.312e-03 and 5.257e-02 respectively. These results suggest that keeping all other factors the same, if one drink milk sometimes or with a varied frequency, he is predicted to have a higher BMI compared to those who never drink milk. log(eatout) also has a positive effect on BMI with a coefficient of 3.181e-03, which suggests that, for every dollar spent on eating out in a month there is an 3.181e-03 increase in a person's BMI on average. The interaction term between race_black and meals spent on fast food has a negative coefficient of -3.007e-03, which indicates that black Americans have a lower BMI than Mexican Americans on average for every additional fast food meal. The interaction between gender(female) and meals spent on fast food has a positive coefficient of 2.647e-03, which suggests females have a higher BMI than male for every additional fast food meal.

Variable Selection and Regularization Using Elastic Net Models

Another strategy we used to perform variable selection and/or regularization was to fit lasso (**model_lasso**, alpha = 1), ridge (**model_ridge**, alpha = 0) and elastic net (**model_elastic**, alpha = 0.5) models from the glmnet package with predictors based on model1b + all two-way interactions (i.e. all main effects + sleep_hr2 + all two-way interactions). Based on their cross-validation MSE, elastic net model performed the best with a cross-validation MSE of 0.004133491 (**Table 5**). The elastic net model is also the best performing among all the linear regression models we explored. The diagnostic plots for these elastic net models share similar patterns as model1 and its variations, where non-constant variance seems to be the most prominent violation of linear regression assumptions (**Figure 12 - Figure 14**).

<u>Logistic Regression Model Performance, Interpretation and Comparisons.</u>

To attempt to improve our best predictive model, we fitted logistic regression models to a categorical response variable (*bmi_class*), which was a feature we engineered by categorizing absolute BMI values into four general classes ("underweight": BMI<18.5, "healthy": 18.5<BMI<25, "overweight": 25<BMI<30, "obese": BMI>30) based on conventional standards. We explored two types of logistic regression: multinomial and ordinal. While multinomial logistic regression is the natural model choice for classification problems with more than one level, it does have one limitation in the context of this problem of not assuming any order in the response variable. The four BMI classes have a specific order where the obese group has BMI values strictly higher than the overweight group, and so on. Therefore, we also used an ordinal regression model for comparison.

Our baseline logistic regression model (**multinomial_base_cv**) included all the main effects. In addition, we fitted two models based on model1b (main effects + the square of sleep hours) and the model1d **(**forward-selected full interaction effects), which we call **multinomial_1b_cv** and **multinomial_1d_cv**, respectively. We used the 5-folds cross validation results to evaluate their performances. Surprisingly, the base model **multinomial_base_cv** has the highest prediction accuracy 0.4904, while the other two models with more predictors have smaller accuracies, 0.4833 and 0.4681. This could possibly because multinomial_1b_cv and multinomial_1c_cv overfit the data. The classification accuracy from 5-fold cross-validation of our ordinal regression model is only 0.4126, which is smaller than the accuracy of multinomial regression treating response as nominal variable. Therefore, the best logistic regression model is our baseline multinomial model with all the main effects.

LIMITATION

One of the main limitations of our analysis is that we dropped observations with missing data and all observations with an answer of either "refused" or "unknown" in the process of data cleaning. Since we cannot assume that missing or "refused" answers were at random, dropping these observations likely resulted in some bias in our analysis. One strategy to deal with missing data is data imputation, but it is out of the scope of this course.

In addition, there likely exist some confounding factors that are not considered. For instance, daily calorie intake, genetic inheritance and other aspects of an individual may all be associated with BMI but were not measured or included to be controlled in our models. A possible strategy to deal with uncontrolled confounding factors is to gather longitudinal data and model the paired difference between 2 years. However, our data source did not allow for such study since data were collected on different individuals across the years (i.e. SEQN numbers, which are individual ID numbers, are completely different in data from different years).

Other weaknesses of our analysis include: 1) that the data in this study are likely not independent observations since individuals within the same geographic region are likely to have similar attributes such as income; and 2) the NHANES health data was collected from an observational study instead of a randomized experimental trial, so we can not draw a causal link between the predictors and BMI. We merely observe associations.

CONCLUSION

The main purpose of this study was to find fairly easily controlled behavioral factors that are associated with obesity in order to add to the existing body of understanding of obesity. In particular, we were interested in the relationship of alcohol use, TV consumption and certain aspects of eating habits like money spent on eating out with obesity.

Using multiple linear regression for association estimation and bootstrap for significance determination, we made several observations. Note, the following statements are all made in the context of controlling for other predictors in our model1d, which was built by forward-selection on all main effects + sleep_hr2 + all two-way interactions. First, the frequency of alcohol consumption (measured by number of days in a year a person consumes alcoholic beverage) has a slight but statistically significant negative association with obesity. Second, in general, more TV consumption, more milk consumption, and more money spent on eating out are associated with higher BMI. Furthermore, compared to Mexican Americans, people of black and other races are less prone to increase in BMI for every additional fast food meal they eat. Compared to males, females are more prone to increase in BMI for every additional fast food meal they eat. Finally, we found significant associations that have been identified previously in our model, including that Asian Americans have significantly lower BMI compared to Mexican Americans and that non-smokers have higher BMI compared to smokers. While our analysis suggests that drinking alcohol more frequently is negatively associated with obesity, we do not recommend heavy alcohol use. However, based on our other findings people who desire or in need to lower their BMI may be advised to drink less milk, watch less TV, and spend less money eating out.

In attempt to find the best predictive regression model, we compared forward-selection, lasso, ridge and elastic net models based on all main predictors + sleep_hr2 + all two-way interactions. We found that elastic model with an alpha of 0.5 gave the best cross-validation MSE of 0.004133491. To further improve our predictive model, we explored logistic regression by classifying BMI into four categories based on conventional standards and found that multinomial logistic regression with all main predictors gave the best cross-validation accuracy of 0.4904. The moderately low classification accuracy indicates that there is still much room for improving model predictive ability. We envision several possible directions to improve our model. First, combining data from multiple years can expand our data set. Second, expand our collection of variables in order to control for more confounding factors that may have been left out in this study. Finally, we could consider other strategies to engineer our features of interest in order to better represent them in the process of modeling.

REFERENCES

Clement, Karine, et al. “Genetics of Obesity.” American Journal of PharmacoGenomics, vol. 2, no. 3, 2002, pp. 177–187., doi:10.2165/00129785-200202030-00003.

Flegal, Katherine M. “The Effects of Changes in Smoking Prevalence on Obesity Prevalence in the United States.” American Journal of Public Health, vol. 97, no. 8, 2007, pp. 1510–1514., doi:10.2105/ajph.2005.084343.

Flegal, Katherine M., et al. “Trends in Obesity Among Adults in the United States, 2005 to 2014.” *Jama*, vol. 315, no. 21, July 2016, p. 2284., doi:10.1001/jama.2016.6458.

"National Center for Health Statistics." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 13 Oct. 2017, www.cdc.gov/nchs/products/databriefs/db288.htm.

"Overweight & Obesity." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 29 Aug. 2017, www.cdc.gov/obesity/data/adult.html.

Patel, Sanjay R., and Frank B. Hu. "Short Sleep Duration and Weight Gain: A Systematic Review." *Obesity*, vol. 16, no. 3, 2008, pp. 643–653., doi:10.1038/oby.2007.118.

Reaven, Gerald M. "Insulin Resistance: the Link Between Obesity and Cardiovascular Disease." *Endocrinology and Metabolism Clinics of North America*, vol. 37, no. 3, 2008, pp. 581–601., doi:10.1016/j.ecl.2008.06.005.

Wang, Y Claire, et al. "Health and Economic Burden of the Projected Obesity Trends in the USA and the UK." *The Lancet*, vol. 378, no. 9793, 2011, pp. 815–825., doi:10.1016/s0140-6736(11)60814-3.

Wellman, Nancy S, and Barbara Friedberg. "Causes and Consequences of Adult Obesity: Health, Social and Economic Impacts in the United States." *Asia Pacific Journal of Clinical Nutrition*, vol. 11, no. s8, 2002, doi:10.1046/j.1440-6047.11.s8.6.x.

Yaturu, Subhashini. "Obesity and Type 2 Diabetes." *Journal of Diabetes Mellitus*, vol. 01, no. 04, 2011, pp. 79–95., doi:10.4236/jdm.2011.14012.

Yaturu, Subhashini. "Diabetes and Cancer." *Type 2 Diabetes*, 2013, doi:10.5772/56419.

TABLES

Table 1. Variables and variable descriptions

Variable Name	Data Type	Description	Possible values
BMXBMI (bmi) *	numeric	BMI	> 0
RIAGENDR (gender)	categorical	Gender	1 (male), 2 (female)
RIDAGEYR (age)	numeric	Age in years at screening	0 - 80
RIDRETH3 (race)	categorical	Race	1 (mexican), 2 (other hispanic), 3 (white), 4 (black), 6 (asian), 7 (other)
DMDEDUC2 (edu)	categorical	Educational level	1 (less than 9th grade), 2 (9-11th grade), 3 (high school graduate/GED), 4 (some college or AA degree), 5 (college graduate or above), 7 (refused), 9 (don't know)
DMDMARTL(marriage)	categorical	Marital status	1 (married), 2 (widowed), 3 (divorced), 4 (separated), 5 (never married), 6 (living with partner), 77 (refused), 99 (don't know)
DMDFMSIZ (famsize)	numeric	Total number of people in the Family	1 - 7
INDFMIN2 (famincome)	categorical	Total family income	1 ($ 0 to $ 4,999), 2 ($ 5,000 to $ 9,999), 3 ($10,000 to $14,999), 4 ($15,000 to $19,999), 5 ($20,000 to $24,999), 6 ($25,000 to $34,999), 7 ($35,000 to $44,999) , 8 ($45,000 to $54,999), 9 ($55,000 to $64,999), 10 ($65,000 to $74,999), 12 ($20,000 and Over), 13 (Under $20,000), 14 ($75,000 to $99,999) 15 ($100,000 and Over) 77 (refused) 99 (don't know)
ALQ101(alcohol12yr)	categorical	In any one year, has the participant had at least 12 drinks of any type of alcoholic beverage?	1 (yes), 2 (no), 7 (refused), 9 (don't know)

ALQ120Q (alcoholfrq)	numeric	How often does the participant drink alcohol over past 12 months?	0 - 365, 777 (refused), 999 (don't know)
CBD070 (grocery)	numeric	Amount of money spent on grocery during the past 30 days	> 0, or 777777 (refused), 999999 (don't know)
CBD120 (eatout)	numeric	Amount of money spent on eating out during the past 30 days	> 0 or 777777 (refused), 999999 (don't know)
CBD130 (delivery)	numeric	Amount of money spent on carry out/delivered food during the past 30 days	> 0 or 777777 (refused), 999999 (don't know)
DBQ197 (milk)	categorical	Past 30 day milk consumption	0 (never), 1 (rarely-less than once a week), 2 (sometimes-once a week or more, but less than once a day, 3 (often- >once a day) 4 (varied) 7 (refused) 9 (don't know)
DBD895 (meals_nothome)	numeric	Number of meals not home prepared during the past 7 days	0 - 21, or 5555 (more than 21), 7777 (refused), 9999 (don't know)
DBD900 (meals_fastfood)	numeric	Number of fast food or pizza meals during the past 7 days	0 - 21, or 5555 (more than 21), 7777 (refused), 9999 (don't know
DPQ020 (depressed)	categorical	Over the last 2 weeks, how often has the participant been bothered by the following problems: feeling down, depressed, or hopeless?	0 (not at all), 1 (several days), 2 (more than half of the days), 3 (nearly every day), 7 (refused), 9 (don't know)
DPQ030 (sleep_trouble)	categorical	Over the last 2 weeks, how often has the participant been bothered by the following problems: trouble falling or staying asleep, or sleeping too much?	0 (not at all), 1 (several days), 2 (more than half of the days), 3 (nearly every day), 7 (refused), 9 (don't know)
PAQ710 (tv_hrs)	categorical	Average hours spent watching TV	0 (less than 1 hr), 1 (1 hr), 2 (2 hrs), 3 (3 hrs), 4 (4 hrs), 5 (5 hrs or more) 8 (does not watch TV or videos) 77 (refused) 99 (don't know)
SLD010H (sleep_hr)	numeric	Amount of sleep on a typical weekday (hours)	> 0, or 12 (12 hours or more), 77 (refused), 99 (don't know)
SMQ040 (smoke)	categorical	Does the participant now smoke cigarettes?	1 (every day), 2 (some days), 3 (not at all),

			7 (refused), 9 (don’t know)
PAQ665 (activity)	categorical	In a typical week, does the participant do any moderate-intensity activity for at least 10 minutes continuously?	1 (yes), 2 (no), 7 (refused), 9 (don’t know)

* Response variable

Table 2. Preprocessed variables and variable descriptions

Variable Name	Data Type	Description	Possible values
BMXBMI (bmi) *	numeric	BMI	> 0
bmi_class *	categorical	Obese levels	underweight, healthy, overweight, obese, and class 3 obesity
RIAGENDR (gender)	categorical	Gender	male, female
RIDAGEYR (age)	numeric	Age in years at screening	0 - 80
RIDRETH3 (race)	categorical	Race	mexican, other hispanic, white, black, asian, other
DMDEDUC2 (edu)	categorical	Educational level	below hs, hs, above hs
DMDMARTL(marriage)	categorical	Marital status	Never_married, married, divorced, widowed
DMDFMSIZ (famsize)	numeric	Total number of people in the Family	1 - 7
INDFMIN2 (famincome)	categorical	Total family income	below 25k, 25k~55k, 55k~75k, 75k~100k, above 100k
ALQ101(alcohol12yr)	categorical	In any one year, has the participant had at least 12 drinks of any type of alcoholic beverage?	yes, no
ALQ120Q (alcoholfrq)	numeric	How often does the participant drink alcohol over past 12 months?	0 - 365
CBD070 (grocery)	numeric	Amount of money spent on grocery during the past 30 days	> 0
CBD120 (eatout)	numeric	Amount of money spent on eating out during the past 30 days	> 0
CBD130 (delivery)	numeric	Amount of money spent on carry out/delivered food during the past 30 days	> 0
DBQ197 (milk)	categorical	Past 30 day milk consumption	never, rare, sometimes, often, varied
DBD895 (meals_nothome)	numeric	Number of meals not home prepared during the past 7 days	0 - 21
DBD900 (meals_fastfood)	numeric	Number of fast food or pizza meals during the past 7 days	0 - 21
DPQ020 (depressed)	categorical	Over the last 2 weeks, how often has the participant been bothered by the following problems: feeling down,	no, sometimes, often, always

		depressed, or hopeless?	
DPQ030 (sleep_trouble)	categorical	Over the last 2 weeks, how often has the participant been bothered by the following problems: trouble falling or staying asleep, or sleeping too much?	no, sometimes, often, always
PAQ710 (tv_hrs)	categorical	Average hours spent watching TV	0hr, 1hr, 2-4hrs, 5hrs
SLD010H (sleep_hr)	numeric	Amount of sleep on a typical weekday (hours)	> 0
SMQ040 (smoke)	categorical	Does the participant now smoke cigarettes?	everyday, sometimes, no, missing
PAQ665 (activity)	categorical	In a typical week, does the participant do any moderate-intensity activity for at least 10 minutes continuously?	yes, no

* response variable

Table 3. Model Performance of Base Model and Its Variations

Models	R-Squared	Cross-Validation Mean R-Squared	Cross-Validation Mean Squared Error
model1	0.1153	0.0889	0.0643
model1a	0.1142	0.0925	0.0644
model1b	0.1174	0.0938	0.0645
model1c	0.1705	0.1184	0.0634

Table 4. Results of Extra-Sum-of-Squares F Test

Models	Res.Df	RSS	Df	Sum.of.Sq	F	Pr(>F)
1 vs. 1a	2924	11.928	NA	NA	NA	NA
	2916	11.896	8	0.03278	1.0044	0.4303
1a vs. 1b	2924	11.928	NA	NA	NA	NA
	2923	11.889	1	0.038976	9.5821	0.001983 ***
1b vs. 1c	2923	11.889	NA	NA	NA	NA
	2913	11.851	10	0.038099	0.9365	0.4981
1b vs. 1d	2923	11.889	NA	NA	NA	NA
	2871	11.153	52	0.73587	3.6427	< 2.2e-16 ***

Table 5. Model Performance of Elastic Models

Models	Cross-Validation Mean Squared Error
model_lasso	0.004177477
model_ridge	0.004142151
model_elastic	0.004133491

FIGURES

Figure 1. Distribution and Transformation of Response Variable BMI

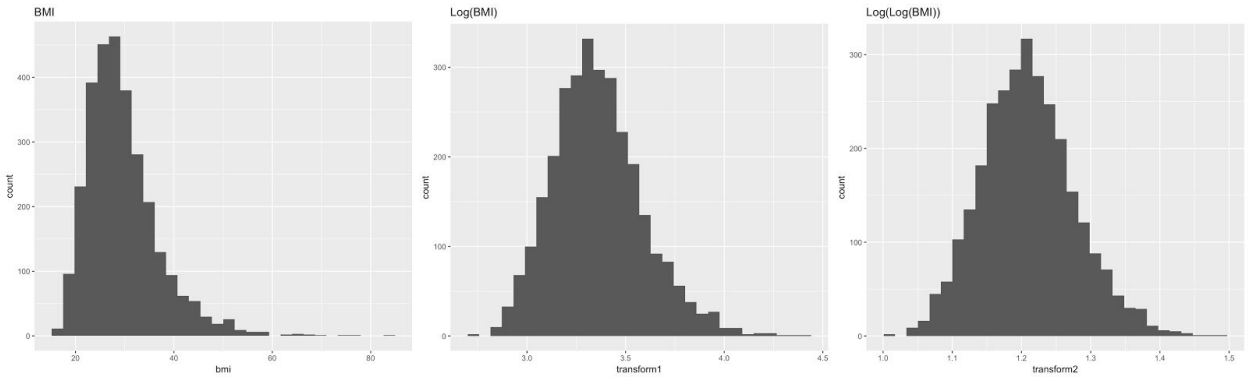


Figure 2. Distribution of Numeric Predictor Variables

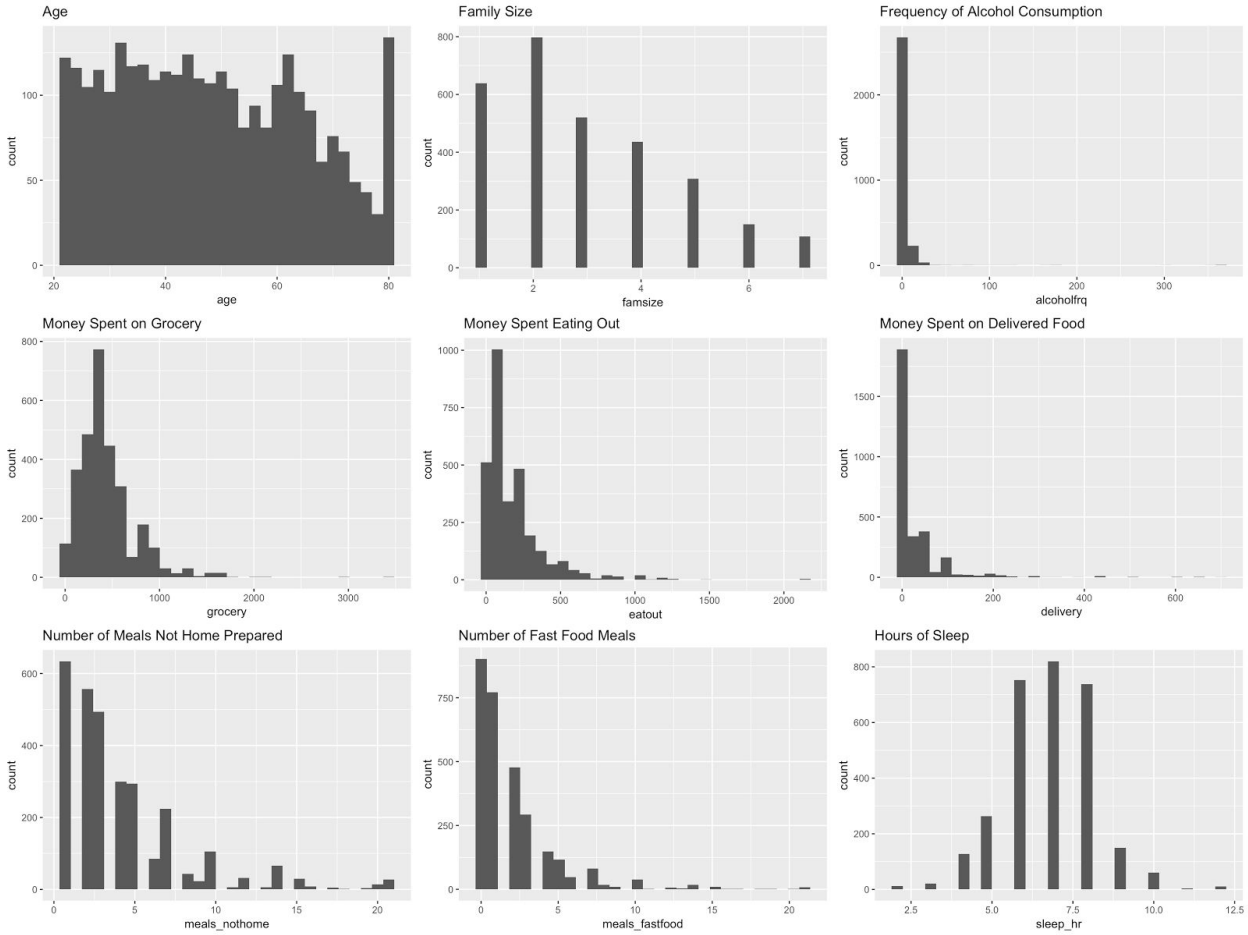


Figure 3. Exploration of Transformation of Numeric Predictors

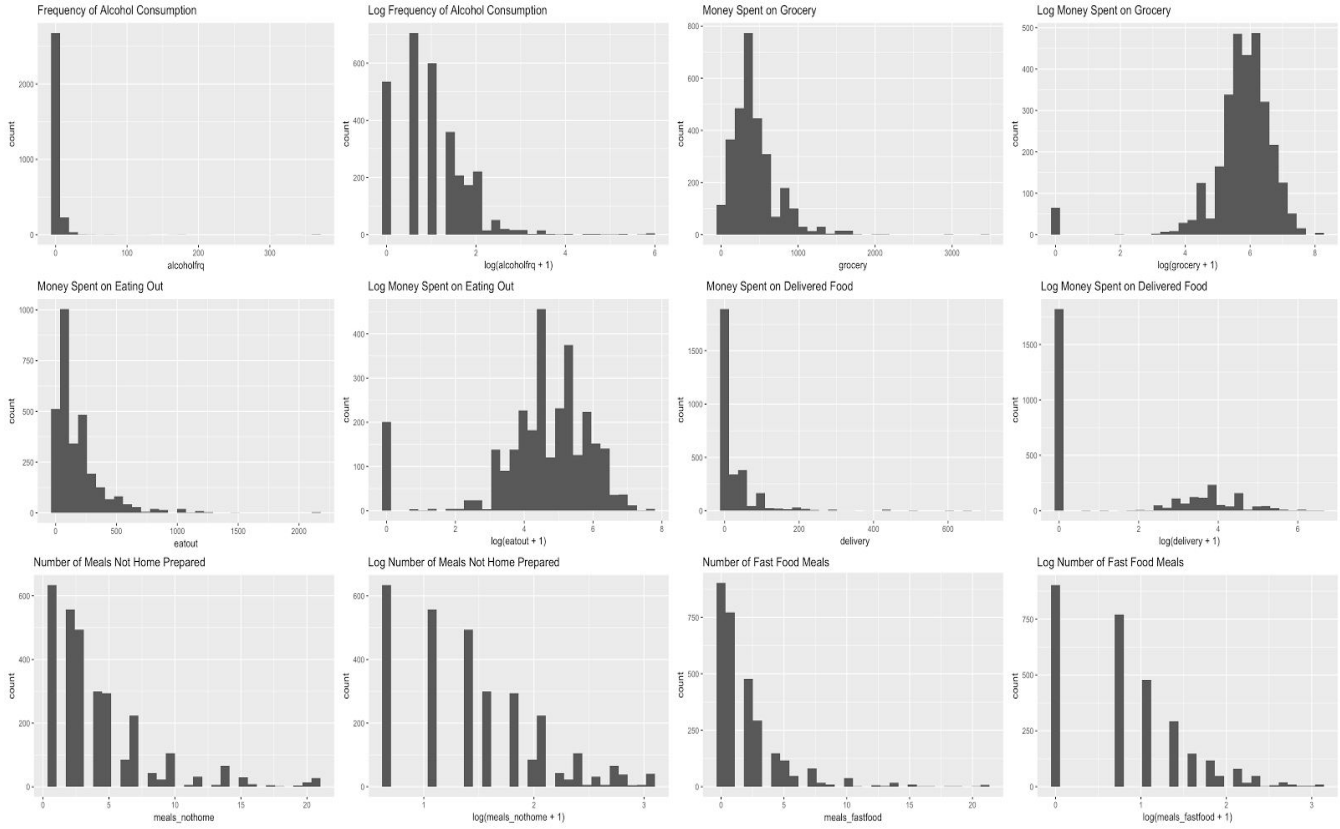


Figure 4. Pairplot

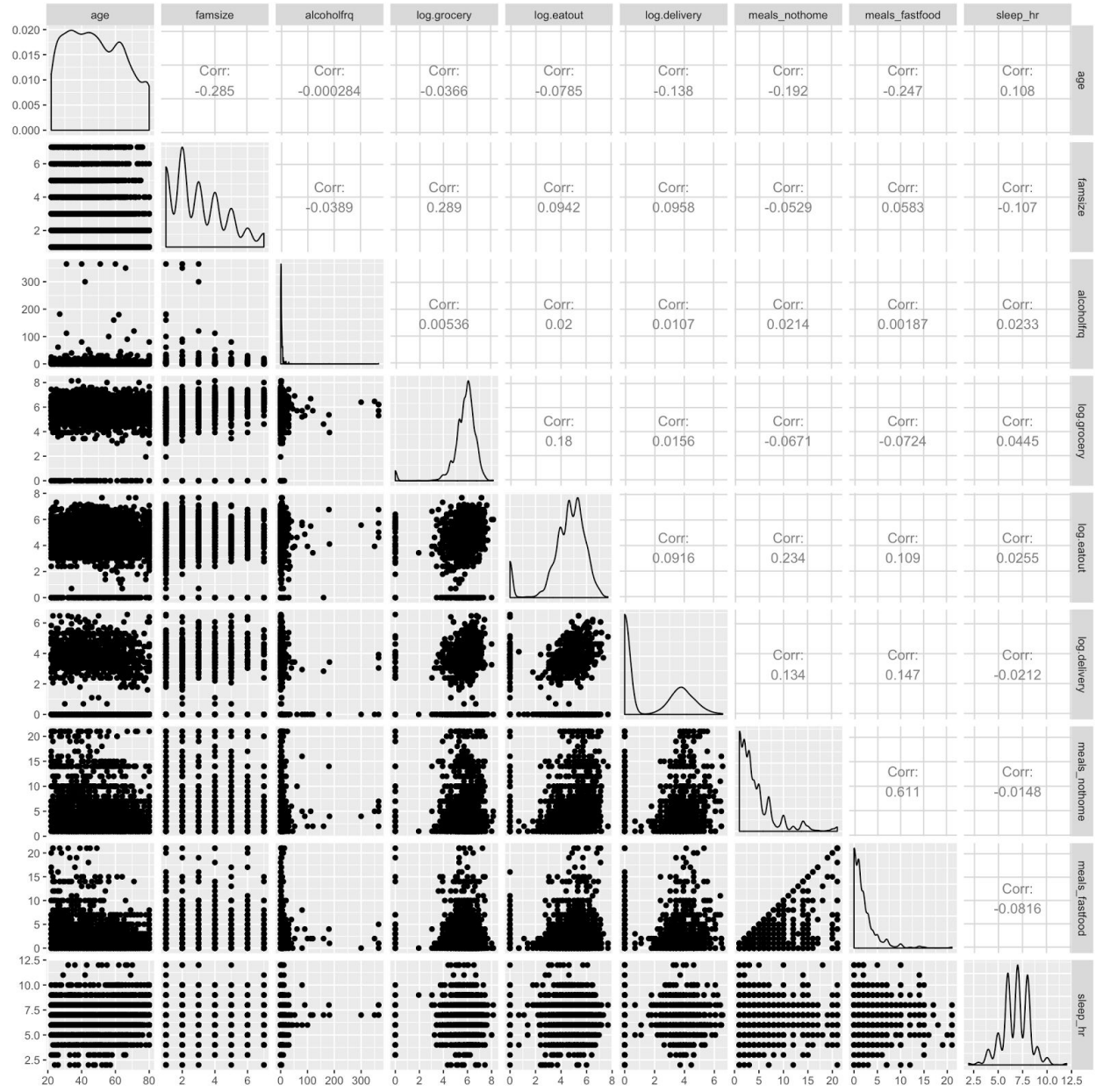


Figure 5. Response vs. Categorical Variables

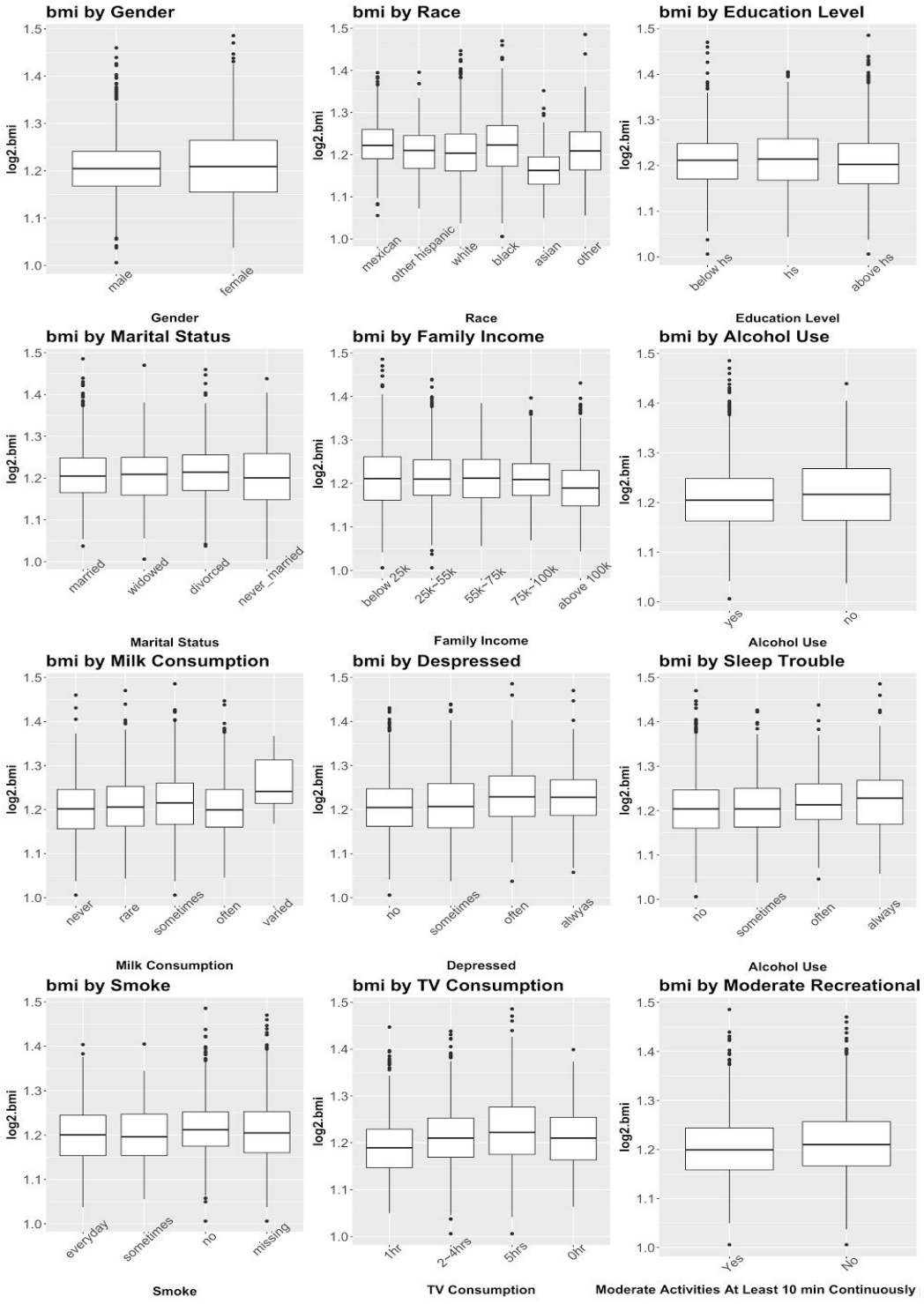


Figure 6. Response vs. Numeric Predictors

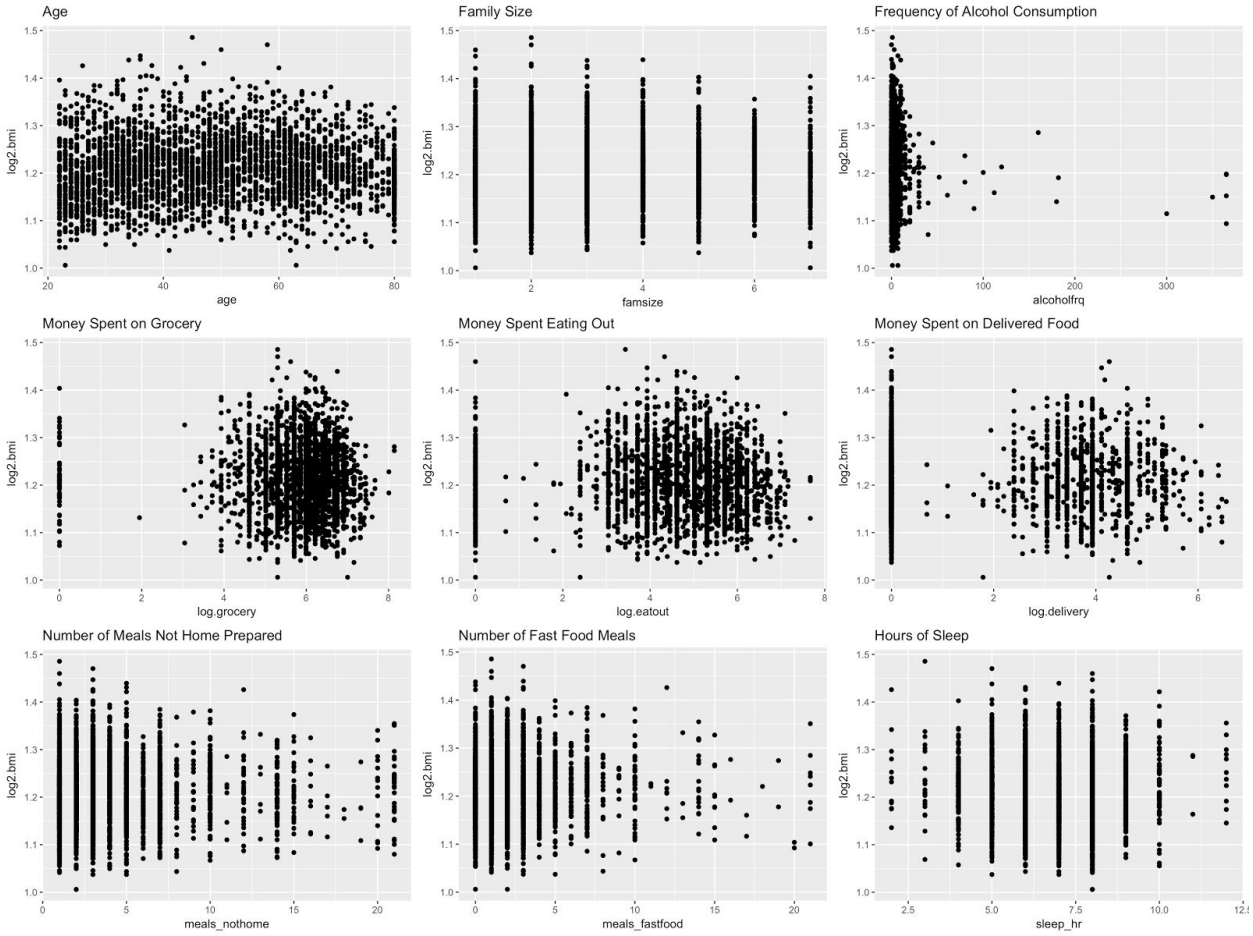


Figure 7. Model 1 Assumption Check

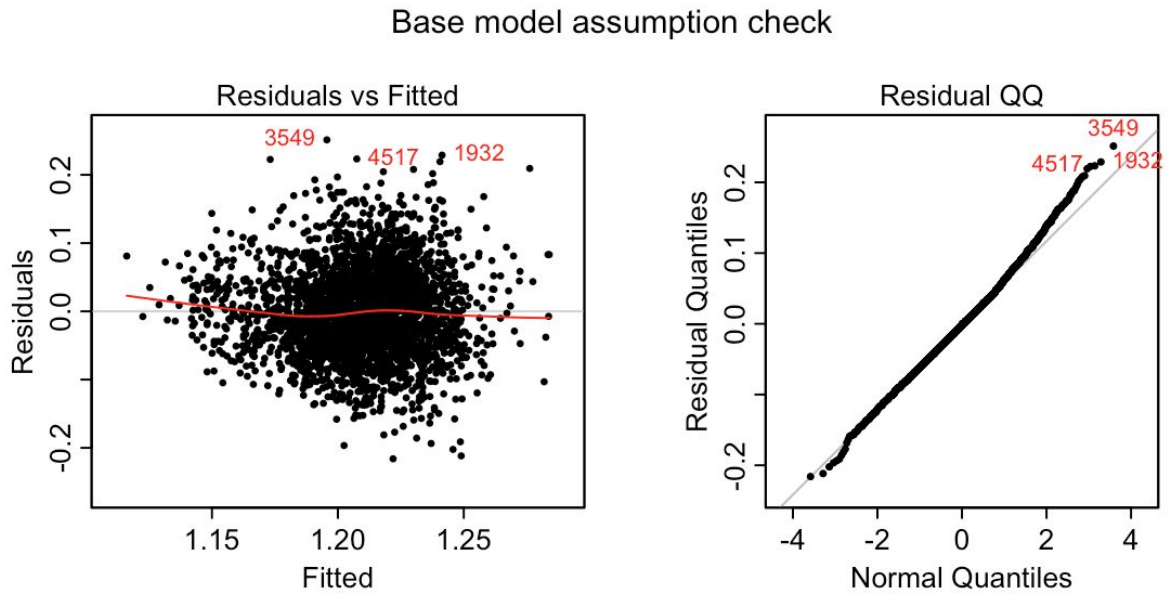


Figure 8. Model 1a Assumption Check

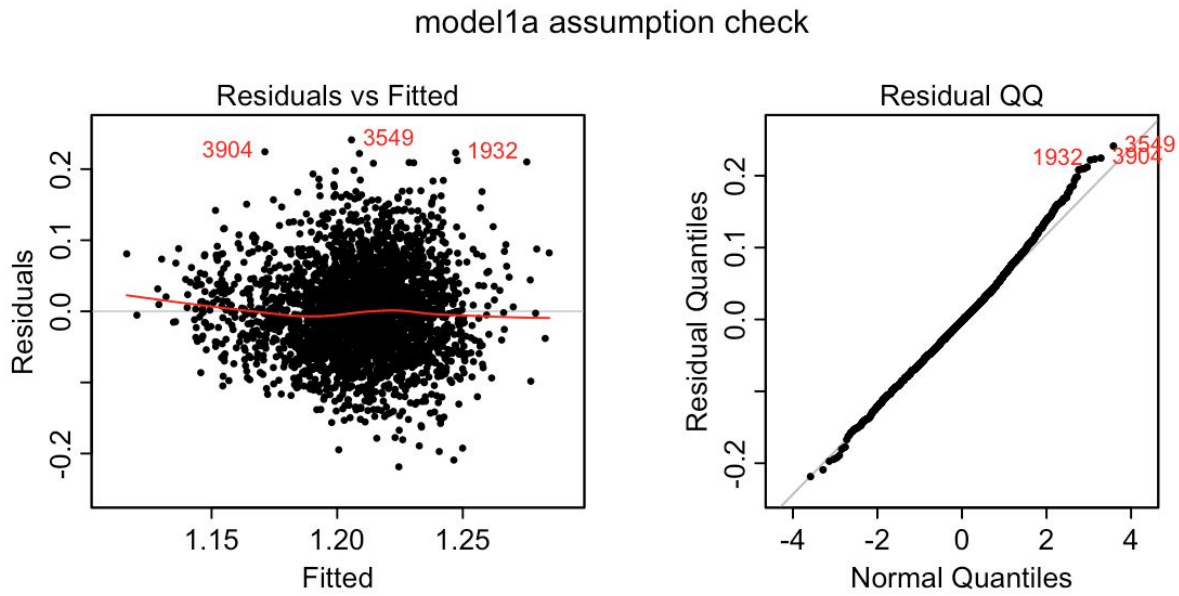


Figure 9. Model 1b Assumption Check

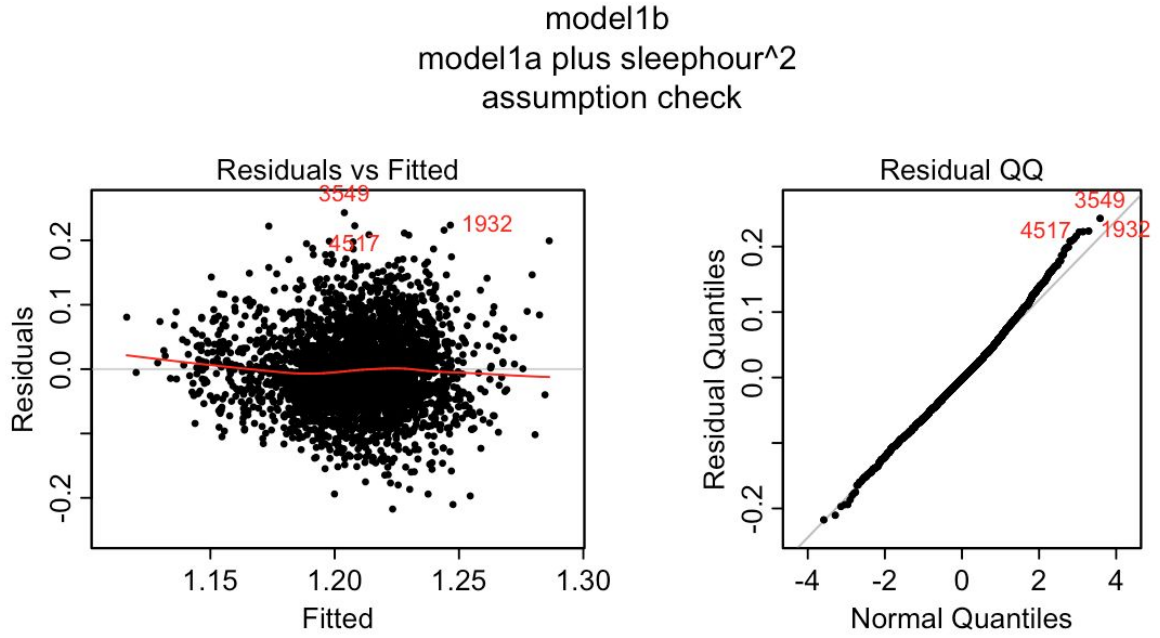


Figure 10. Model 1c Assumption Check

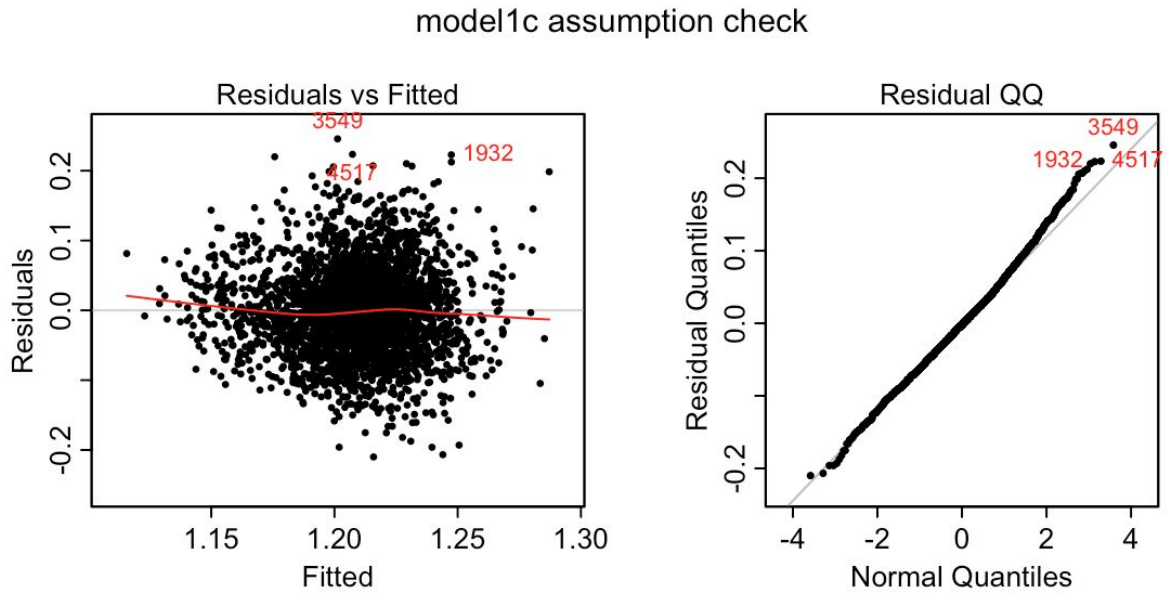


Figure 11. Model 1d Assumption Check

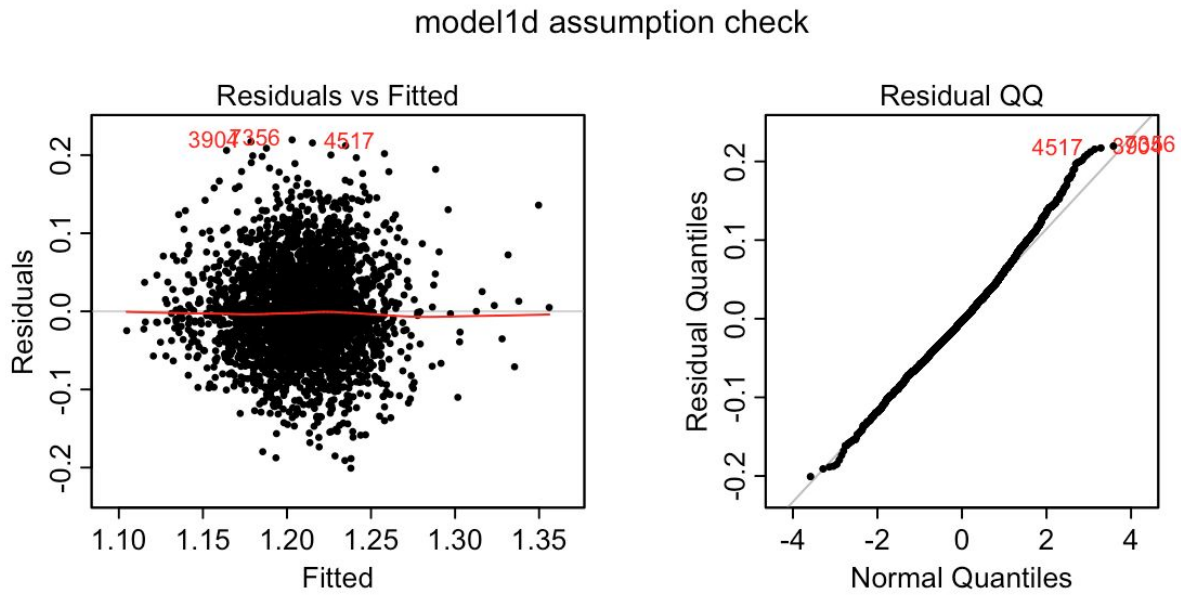


Figure 12. Model Lasso Assumption Check

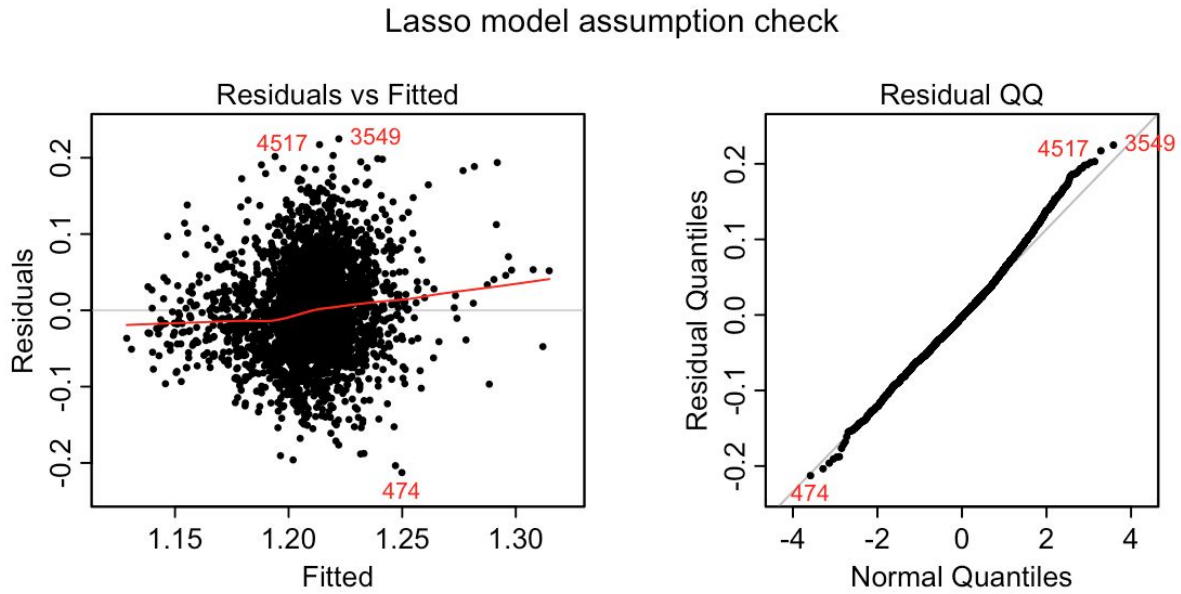


Figure 13. Model Ridge Assumption Check

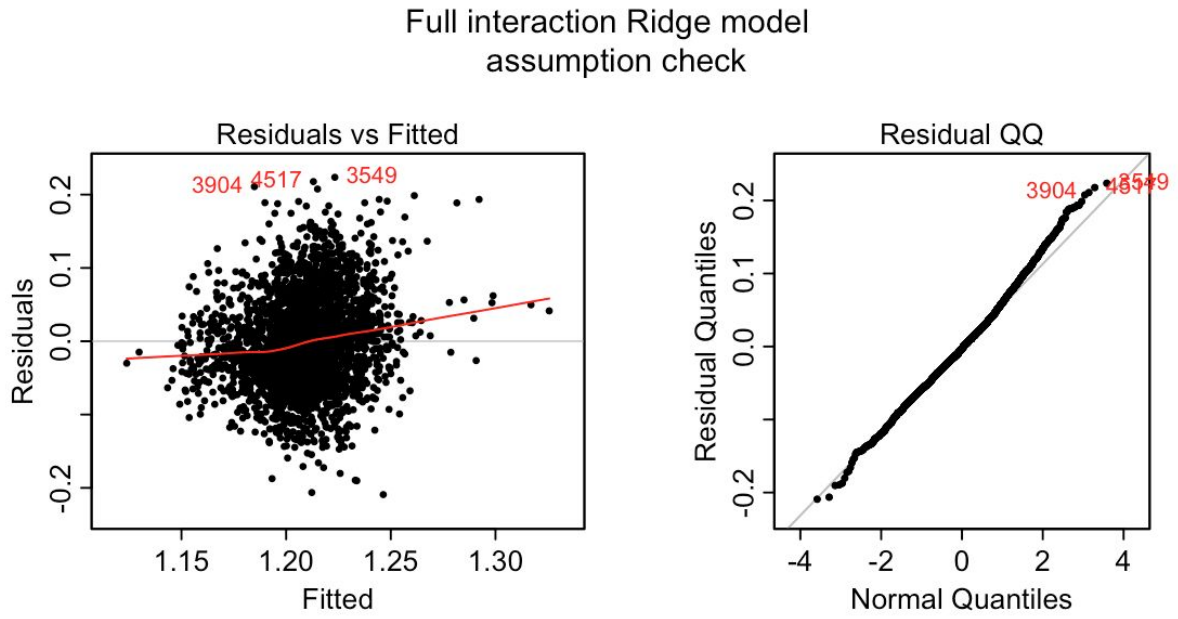
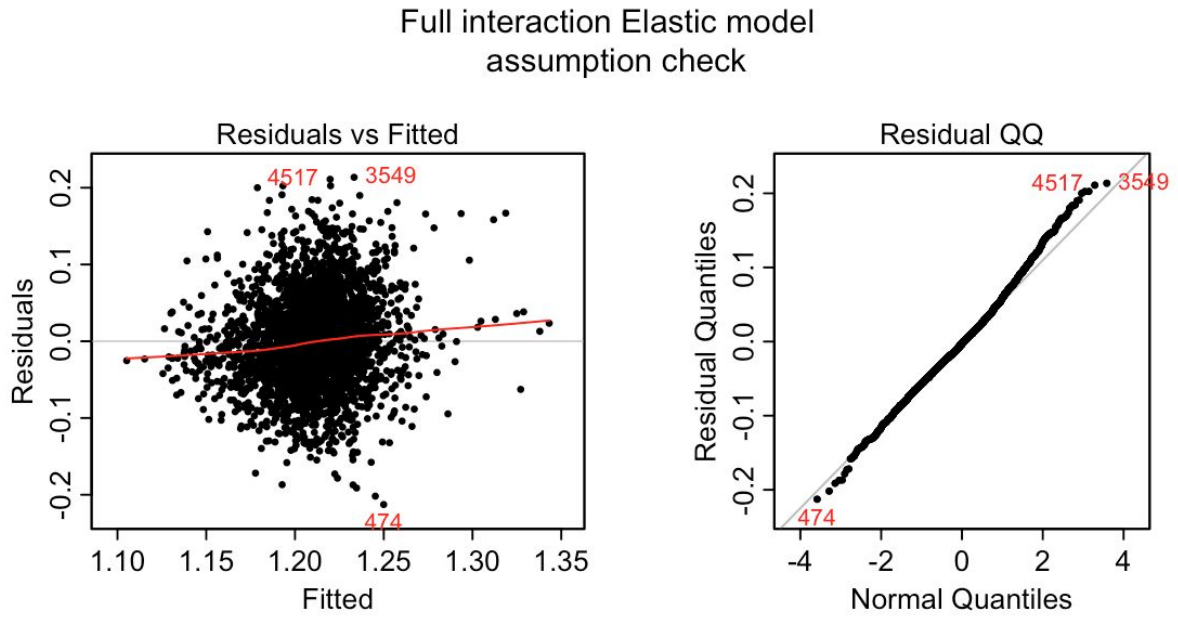


Figure 14. Model Elastic Assumption Check



Appendix I: Summary of Model1d

	Estimate	Std. Error	t.value	Pr(>|t|)
(Intercept)	1.14834313	0.01214535	94.55002147	0
raceother hispanic	-0.006377078	0.010505721	-0.607010095	0.543892212
racewhite	-0.011492358	0.007842363	-1.465420256	0.142915796
raceblack	0.011104239	0.008874569	1.251242682	0.210947896
raceasian	-0.025784828	0.010117903	-2.548436044	0.010872355
raceother	0.021003104	0.015622528	1.344411366	0.178921612
tv_hrs2~4hrs	0.021764665	0.009357162	2.325989987	0.020088387
tv_hrs5hrs	0.007280244	0.011531263	0.63134839	0.527863014
tv_hrs0hr	0.011844219	0.026731224	0.443085548	0.657737241
smokesometimes	-0.005032631	0.010178061	-0.494458762	0.621020055
smokeno	0.030522704	0.005809894	5.253573617	1.60E-07
smokemissing	0.033140266	0.005636156	5.879940796	4.58E-09
depressedsometimes	-0.019441888	0.009464716	-2.054143731	0.040052136
depressedoften	-0.011781235	0.017696835	-0.665725557	0.505640016
depressedalwyas	0.007267109	0.020834102	0.348808342	0.727258778
milkrare	0.005320384	0.004487318	1.185649014	0.235859057
milksometimes	0.008312277	0.004003246	2.07638433	0.037947188
milkoften	0.000119223	0.00395034	0.030180361	0.975925309
milkvaried	0.052567145	0.022423325	2.344306384	0.019130324
famincome25k~55k	0.010410592	0.004465136	2.331528663	0.019794354
famincome55k~75k	0.006884158	0.005673419	1.213405503	0.225074641
famincome75k~100k	0.007329284	0.00612046	1.197505481	0.231208451
famincomeabove 100k	0.002492509	0.005162621	0.482799208	0.629275123

genderfemale	0.035515408	0.009896689	3.588615105	0.000337998
log.eatout	0.003180576	0.001595162	1.993889358	0.046258591
activityNo	0.018149444	0.006949935	2.611455439	0.009062711
alcoholfrq	-0.000175864	6.58E-05	-2.674191609	0.007533418
marriagewidowed	-0.024781505	0.018863011	-1.31376194	0.189031297
marriagedivorced	0.009005726	0.007572689	1.189237459	0.234444584
marriagenever_married	-0.001729803	0.007144258	-0.242124959	0.808700632
meals_fastfood	0.000820541	0.001206495	0.680102608	0.496494329
sleep_troublesometimes	8.22E-05	0.002936483	0.028006816	0.977658661
sleep_troubleoften	0.008353956	0.004982889	1.676528542	0.093743581
sleep_troublealways	0.009882544	0.004490231	2.200898664	0.027822533
tv_hrs2~4hrs:depressedsometimes	0.008323091	0.007515722	1.107423969	0.268203439
tv_hrs5hrs:depressedsometimes	0.015766402	0.009635027	1.636362956	0.101873253
tv_hrs0hr:depressedsometimes	0.039962953	0.022438776	1.780977423	0.075021841
tv_hrs2~4hrs:depressedoften	0.0155541	0.015754205	0.987298299	0.323579621
tv_hrs5hrs:depressedoften	0.073227052	0.020114741	3.640467084	0.000276948
tv_hrs0hr:depressedoften	0.002633958	0.029188686	0.090239006	0.928103584
tv_hrs2~4hrs:depressedalwyas	-0.016178745	0.01734526	-0.932747323	0.351028878
tv_hrs5hrs:depressedalwyas	-0.018618305	0.018838576	-0.988307439	0.323085386
tv_hrs0hr:depressedalwyas	-0.065118526	0.065781489	-0.989921752	0.322295787
famincome25k~55k:genderfemale	-0.010373729	0.006295879	-1.647701351	0.099523398
famincome55k~75k:genderfemale	-0.013001696	0.008367247	-1.553879687	0.120323338
famincome75k~100k:genderfemale	-0.017102615	0.008864065	-1.929432347	0.053775644
famincomeabove 100k:genderfemale	-0.028530629	0.007043838	-4.050438022	5.25E-05
raceother hispanic:genderfemale	-0.0140219	0.010841218	-1.293387898	0.195980987

racewhite:genderfemale	-0.008346095	0.007753607	-1.076414534	0.281832348
raceblack:genderfemale	0.008942294	0.008740191	1.023123477	0.306335667
raceasian:genderfemale	-0.024339206	0.010826057	-2.24820596	0.024638637
raceother:genderfemale	-0.020778943	0.015694039	-1.324002306	0.18560768
smokesometimes:genderfemale	-0.002045232	0.013337349	-0.153346196	0.878136049
smokeno:genderfemale	-0.004893253	0.007584803	-0.645139138	0.518888572
smokemissing:genderfemale	-0.023182069	0.007008283	-3.307810009	0.000951876
raceother hispanic:activityNo	-0.013730366	0.010770629	-1.274797014	0.202484257
racewhite:activityNo	-0.007820591	0.007702967	-1.015269912	0.310062733
raceblack:activityNo	-0.022079498	0.008837875	-2.498281235	0.012535225
raceasian:activityNo	-0.030170451	0.010616315	-2.841894819	0.004516277
raceother:activityNo	-0.014047442	0.015118412	-0.929161208	0.352883661
smokesometimes:marriagewidowed	0.0281959	0.035535114	0.793465866	0.427572019
smokeno:marriagewidowed	-0.003128824	0.020625256	-0.151698682	0.879435255
smokemissing:marriagewidowed	0.033816896	0.020228085	1.671779457	0.094676792
smokesometimes:marriagedivorced	0.022250747	0.017845238	1.246873038	0.212545766
smokeno:marriagedivorced	-0.010472073	0.009742777	-1.074855094	0.282529932
smokemissing:marriagedivorced	-0.015559214	0.009094118	-1.710909566	0.087205832
smokesometimes:marriagenever_married	0.028176216	0.015273879	1.844732164	0.065179453
smokeno:marriagenever_married	0.004362082	0.010320353	0.422667889	0.67256924
smokemissing:marriagenever_married	-0.014943052	0.008293389	-1.801802847	0.071681295
raceother hispanic:meals_fastfood	0.000909923	0.001921807	0.473472397	0.635912188
racewhite:meals_fastfood	0.000809797	0.001332153	0.607885502	0.543311496
raceblack:meals_fastfood	-0.003007486	0.001431694	-2.100649453	0.035758749
raceasian:meals_fastfood	-0.001690251	0.002510907	-0.673163476	0.500897476

raceother:meals_fastfood	-0.004747154	0.002426747	-1.956179982	0.050540786
genderfemale:meals_fastfood	0.002647252	0.000928094	2.852352764	0.004370728
depressedsometimes:milkrare	0.018300303	0.010465494	1.748632484	0.080461437
depressedoften:milkrare	-0.006828871	0.020098661	-0.339767453	0.734056513
depressedalwyas:milkrare	0.036559306	0.021060942	1.735881767	0.082692028
depressedsometimes:milksometimes	0.007194845	0.009558269	0.752735194	0.45167077
depressedoften:milksometimes	0.030484781	0.019039378	1.601143756	0.1094551
depressedalwyas:milksometimes	-0.015145339	0.019373751	-0.7817453	0.434428736
depressedsometimes:milkoften	0.015248893	0.009384161	1.624960739	0.104280688
depressedoften:milkoften	0.021086897	0.017672793	1.193184198	0.232895834
depressedalwyas:milkoften	0.034176609	0.019053613	1.793707542	0.072965059
depressedsometimes:milkvaried	0.100308355	0.067267219	1.491192242	0.136020889
tv_hrs2~4hrs:log.eatout	-0.001933534	0.001862707	-1.03802375	0.299346412
tv_hrs5hrs:log.eatout	0.003813556	0.002369531	1.609413701	0.107635852
tv_hrs0hr:log.eatout	-0.000370674	0.005890377	-0.062928698	0.949827663