STAT6120 Capstone Project

Shulei Yang (sy8uu) Yisha He (yh2uq)

1. Introduction

Nutrition are crucial in maintaining functions of the human body. It is widely believed that nutrition levels vary greatly with individuals' social and economic status. This project aims to analyze the effect of household condition on an individual's nutrition consumption level by a survey of people living in a developing country from 1989 to 2011.

2. Data Description

The survey is based on household level and individual level separately. The survey was conducted every three years from 1989 to 2011. The survey on individual focuses on the nutrition level of each individual (carbohydrate, fat, kilocalories and protein consumption) and the survey on household focuses more on the wealth level (house size, dwelling size, furniture and equipment). A total of 100,714 observations with 32,083 individuals in 8,494 households are measured. 80,440 observations (80% of the entire dataset) is treated as training data and the remaining 20,274 observations (20% of the entire dataset) is treated to be testing data. Seed (999) is used to keep consistency of the output.

Variable 'kcal' is chosen to be the response variable. According to the Office of Disease Prevention and Health Promotion (ODPHP)'s website carbohydrates provide 4 calories per gram, protein provides 4 calories per gram, and fat provides 9 calories per gram. In this case, 'fat', 'carb' and 'protein' can all be represented by the variable 'kcal'.

3. Dealing with Missing Data

3.1 Numeric Variables

Regression imputation is used to deal with the missing values of 'hh_size' and 'dwelling_size'. Regression imputation is conducted with 'hh_size' and 'dwelling_size' as response respectively, and other variables except 'protein', 'carb', 'fat', 'kcal' as regressors. We imputed the missing values of 'hh_size' and 'dwelling_size' with their predicted values using other regressors.

3.2 Categorical Variables

Regarding to the codebook, we treat binary variables coded '9' or '-99' as N/As. Then we impute the missing values of binary variables as their mode category. Variables 'drinking_water_source' has multiple levels. The missing values of 'drinking_water_source' are imputed with the category 'house_tap' and category '0.0' is deleted, resulting in four categories (house_tap, yard_tap, yard_well and others).

4. Regression Analysis

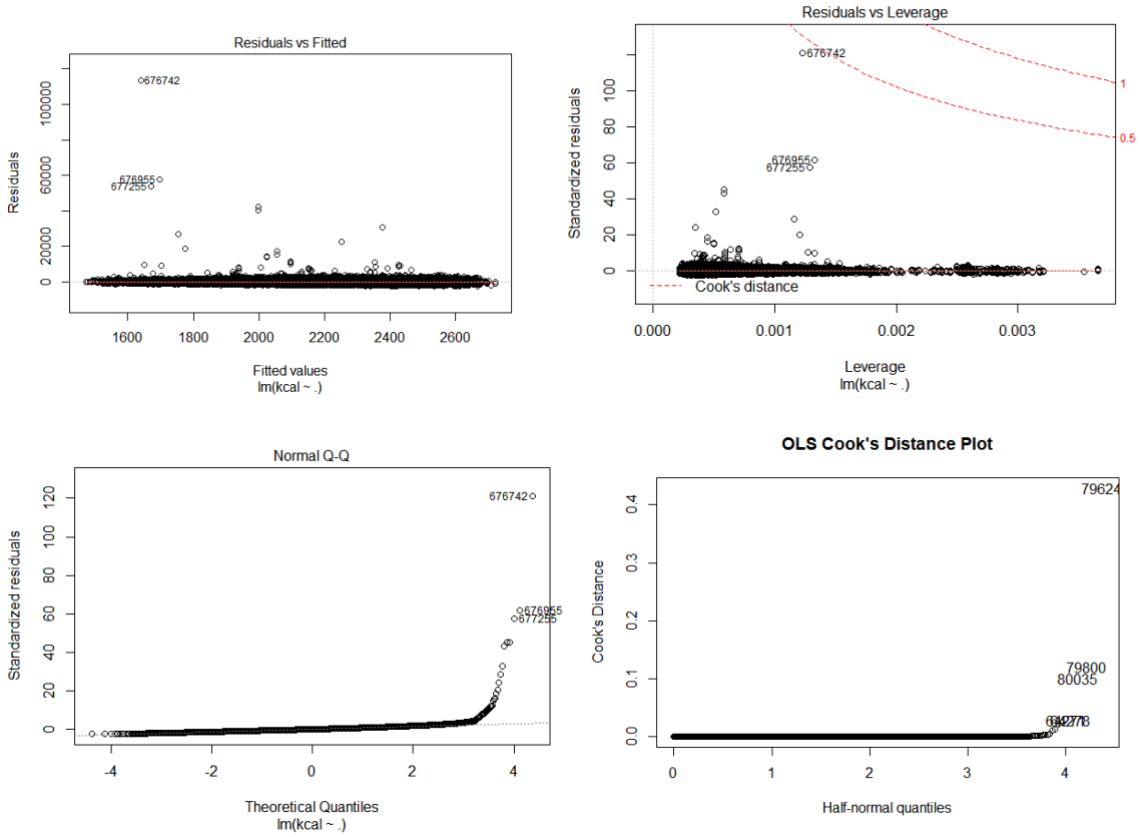
	Dependent variable:				
	OLS	OLS Outliers Removed	OLS AIC	LAD	Huber's Method
urban_ruralurban	-64.927***	-59.501***	-58.708***	-65.513***	61.218***
hh_size	-23.991***	-23.959***	-24.679***	-19.219***	-20.685***
dwelling_size	0.144***	0.196***	0.186***	0.184***	0.198***
drinking_water_sourcehouse_tap	-127.646***	-122.643***	-125.523***	-116.580***	118.565***
drinking_water_sourceother	3.193	9.221	8.441		
drinking_water_sourceyard_tap	-63.752	-56.212	-59.204	-46.555***	-51.982***
drinking_water_sourceyard_well	-64.245	-57.437	-60.332	-34.612**	-47.687***
pay_for_drinking_water1	31.953***	30.518***	31.287***	40.057***	36.770***
toilet1	14.941	13.862*	16.295**	9.190	10.565
bicycle1	42.976***	43.452***	43.142***	58.473***	52.669***
motorcycle1	35.410***	33.668***	33.040***	19.617**	26.801***
car1	-26.361*	-16.058		-3.776	-9.058
bw_tv1	-5.208	-6.486		-2.626	1.440
color_tv1	-50.493***	-52.390***	-51.129***	-30.702***	-35.792***
washing_machine1	-85.876***	-85.063***	-85.551***	-88.688***	-83.595***

refrigerator1	-13.542	-14.440*	-13.184	-7.278	-12.245*
air_conditioner1	-35.980***	-30.346***	-22.868**	-17.496	-23.565**
sewing_machine1	-5.279	-5.778		-10.037	-3.983
electric_fan1	-21.052***	-23.176***	-25.209***	-4.858	-14.379*
wall_clock1	2.541	-0.352			
camera1	19.448*	11.491	16.174*	3.386	3.791
microwave1	30.546**	16.242		20.351*	21.590**
electric_cooking_pot1	13.315	21.246***	20.650***	24.609***	25.470***
pressure_cooker1	4.965	2.593		-0.013	-0.098
computer1	43.286***	17.324		4.271	8.231
telephone1	56.586***	63.537***	65.599***	44.547***	54.312***
dvd1	28.140***	25.272***	26.857***	44.789***	29.953***
cellphone1	61.531***	61.751***	64.585***	57.104***	57.358***
satellite_dish1	-59.585***	-41.692***	-42.802***	-42.650***	38.338***
wave	-15.452***	-15.367***	-14.767***	-17.875***	-16.364***
region21	199.685***	179.653***	177.809***	257.981***	231.910***
region23	161.791***	140.274***	137.099***	252.326***	212.973***
region31	-77.725*	-73.812**	-69.488**	5.313	-7.680
region32	376.414***	357.517***	357.437***	413.086***	388.425***
region37	242.525***	224.664***	222.714***	289.102***	272.005***
region41	233.798***	211.973***	208.709***	257.482***	242.165***
region42	391.243***	368.705***	367.051***	448.315***	414.031***
region43	359.134***	338.955***	337.958***	409.208***	387.177***
region45	154.665***	125.566***	126.032***	217.355***	180.996***
region52	271.902***	248.730***	248.403***	348.764***	311.318***
region55	-119.927	-259.441***	-259.694***	-228.327***	-217.221***
Constant	32,954.910***	32,793.100***	31,593.050***	37,627.570***	34,666.720***
Observations	80,440	80,435	80,435	80,440	80,440
R2	0.044	0.070	0.070		
Adjusted R2	0.044	0.069	0.069		
Residual Std. Error	937.429 (df = 80398)	771.556 (df = 80393)	771.565 (df = 80400)		666.731 (df = 80400)
F Statistic	90.876*** (df = 41; 80398)	146.746*** (df = 41; 80393)	176.694*** (df = 34; 80400)		
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1. Regression Output of All Discussed Models

4.1 Ordinary least squares (OLS) Regression

Ordinary least squares (OLS) model is applied to the training set. Table 1 displays the regression results. From the residual plots of OLS regression (Figure 1), it is clear to see that heavy tails exist in the dataset and more robust methods should be used accordingly. From the cook's distance plot, we targeted outliers as the five observations with the highest residuals. After manually removed the most significant five outliers in the dataset, we applied OLS model on the new dataset and the OLS regression results with outliers removed can be found on Table 1. Both residual standard error and F statistics significantly increased when outliers are removed.



4.2 Ordinary least squares (OLS) Regression with Selected Regressors

Overfitting problem may exist when too many regression methods are used in the regression. Akaike information criterion (AIC) is applied to select the optimal number of predictors that best explains the regression model. Using stepwise AIC method, seven regressors (car, bw_tv, sewing_machine, wall_clock, microwave, pressure_cooker, computer) are removed. From regression result using AIC selection criteria (column 3 of Figure 1), residual standard error is almost unchanged compared to the OLS model (column 2 of Figure 1) but F statistics further improved. ANOVA analysis is used to see whether it is valid to remove these seven regressors implied by AIC. F statistic of the AVOVA test is 1.26 with p value of 0.26. As the test statistic is not significant at 5% level, it is valid to say that the model with fewer regressors are as good as the original models. In this case, we can remove the car, bw_tv, sewing_machine, wall_clock, microwave, pressure_cooker and computer regressors in OLS model.

4.3 Least Absolute Deviations (LAD) Model

As normality assumption is likely to be violated in the OLS regression, more robust regressions are considered. LAD is more robust to heavy-tailed distributions since it penalizes by the proportion of fitted error and does not put extra penalty on large deviations. The result on LAD is also shown in Table 1 (column 4). Most of the coefficients of LAD model are consistent with the AIC model but more regressors are found insignificant with LAD model.

4.4 Huber's Method

We also consider define the loss function by Huber's method. The result is shown in Table 1 (column 5). Comparing with the other models, the residual standard error is lowest using Huber's method.

4.5 Discussion of Regression Results

From the regression results illustrated in Table 1, several factors have significant effect on individuals' nutrition level. Regional difference is very significant. Compared to people in region 11, people living in all the other regions (except region 31 and 55) consumed extra 150 kilocalories in previous three days of the survey period. From the AIC model, urban people consume an average of 58.71 less kilocalories than rural people. The expected kilocalories level will decrease by 24.68 if the household increase the number of individuals by one. The average kilocalories level of an individual will increase 0.186 if the house size increases by one square meters. Individuals whose households are with high economic status (pay for drinking water and have toilet, bicycle, motorcycle, camera, electric cooking pot, telephone, DVD, cellphone, satellite dish etc.) are more likely to have higher kilocalories level than households with low economic status.

4.6 Comparing Prediction Accuracy

We apply the estimated coefficients of the discussed models onto the testing dataset and calculated the mean squared prediction error (MSPE) of each modelling procedure. From Table 2 we can see that OLS with outliers removed, OLS with AIC model selection and Huber's method all successfully have lower MSPE than the original OLS model. By comparing the five models, we can find that the OLS model with outliers removed has the best prediction accuracy among all models. The OLS model with predictors selected by AIC method has also significantly lower MSPE than other methods. However, the improvement in MSPE is not significant and in general we have a large MSPE, implying the general prediction power is weak.

MEAN SQUARED PREDICTION ERRORS

OLS	656617.3
OLS (OUTLIERS REMOVED)	655557
OLS (AIC)	655588
LAD	658165.1
HUBER'S METHOD	656237.2

5. Conclusion

From our analysis, we observed a clear declining pattern of nutrition consumption level from 1989 to 2011, which may cause by a healthier lifestyle people are pursuing. And individuals from households with more members have lower nutrition level on average. Also, people lived in the same region seems to have a similar eating habit. Comparing the five models applied in the analysis, OLS model with outliers removed has the best prediction power. However, the prediction power is weak in general, which implies the model needs further improvement. Nonlinear relationship could exist and more variables such as individual's wealth level could be added to improve model fitting.