

How does obesity impact healthcare utilization and spending among the elderly in the United States?

Keon Woo Kim

Student ID: 20161342

19 June 2021

Abstract

This paper aims at finding out how does obesity impact healthcare utilization and spending among the elderly in the United States by using Poisson regression and multiple linear regression models. Dependent variables were set as number of visits to doctor and hospitals on former problem and total expenditures on latter problem. BMI numerical variable was tuned to categorical variable as underweight, normal, overweight and obese, in order to analyze the characteristics of each group. The results from both problems seem to show obese impacts both on healthcare utilization and expenditures on health more than normal. In other words, controlling the degree of weight among the elderly can reduce both healthcare utilization and spending in the United States.

i. Introduction

According to Centers for Disease (CDC), obesity is classified as a common, serious, and costly disease. In US, obesity prevalence has increased from 30.5% to 42.4% through 20 years. The reason why obesity is serious is that it can incur diverse diseases related to obesity such as diabetes, heart disease and cancers. CDC has also argued that there are a lot of associations between obesity and socioeconomic status such that among men, obesity prevalence is lower in the lowest and highest income groups compared with normal income group. Obesity itself does not serious problem to personal health but also has severe impacts on society. Since obesity can cause serious disease for individuals, the government will try various efforts to prevent individuals from becoming obese, which can be as social waste of money. It is hoped that if research on the number of obese people using health care utilization and the degree of spending on health service, the government will be able to find and control factor variables for these issues so that it will not waste social costs in future obesity-related policies. Thus, in this paper, a study whether obesity impacts healthcare utilization and spending among the elderly in the United States or not will be conducted based data from MEPS (Medical Expenditure Panel Survey). Before entering this paper, I state a null hypothesis that obesity is independent from an impact on health care utilization and health care expenditure, and an alternative hypothesis as vice versa.

ii. Literature Review

The Impact of Obesity on Health Care Utilization and Expenditures in a Medicare Supplement Population

This paper shows similar thesis to the project in ‘Econometrics’. It is about how much the weight affects health care utilization and expenditures among Medicare Supplement Insureds.

Body Mass Index (BMI) is utilized as to estimate the impact of weight on health care utilization (inpatient admissions, emergency room visits, and orthopedic procedures) and expenditures including medical and pharmacy. Two statistical models, propensity weighted multivariate logistic regression models and exponential conditional mean models, are used for each different topic. Before applying the model, the approach to the variables is also divided into two by including covariates to adjust independent variables that could have impact on healthcare utilization or expenditures. The first one is only using variables that have Demographics and Socioeconomics characteristics such as age, gender race, living arrangements, and educational level. The second one is an approach that adds variables that have additional chronic conditions or health status characteristics to the aforementioned factors such as ADLs, general physical, mental health and smoking. Multivariate logistic regression model was used to estimate health care utilization. On this model, weight categories were used as predictor variables and inpatient admissions, ER visits, orthopedic procedures were used as target variables. Each partial model (Demographics and Socioeconomics) and full model (adding additional variables) were utilized. Furthermore, exponential conditional mean model was used to evaluate the health care expenditures across weight categories. Similar to the previous model, it also approaches as two ways, partial model and full model. Therefore, when we turned the experiment to the two models introduced earlier, three significant results were interpreted. First, sample characteristics were captured. High blood pressure was shown to have the most common comorbid condition. Hypertension, arthritis, diabetes and low back pain are shown to have the strongest relationships between specific conditions and weight categories. Second, on health care services utilization, obese people go to have an inpatient admission or an orthopedic procedure more often than normal people. However, with additional variables, using full model, obesity does not show statistically significant values to have higher inpatient utilization but

only remain ‘orthopedic procedure. Third, on health care expenditures across weight categories, people who have obese weight show higher expenditures than normal people only on controlling for demographics and socioeconomics. And as similar to second example, adding health/functional status variables do not have statistically significant values, but expenditures related to managing comorbid chronic conditions statistically explain the obesity-related costs well.

Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis

This paper also shows similar thesis to the project in ‘Econometrics’. It is about estimating the Medical Care Costs of Obesity in the United States. The main two purposes on this paper are to conduct a systematic review and meta-analysis of articles about medical costs and to conduct inventive statistical analysis for proving the importance of studying methodologies through the empirical analysis. This study focuses heavily on meta-analysis of articles on obesity-related costs. Five different number of models were utilized, (1. linear regression model 2.log-linear model 3.one-part log-gamma generalized linear model 4.two-part model with a logistic regression 5.extended estimating equation). Goodness of fit evaluation metrics were used to evaluate five models. Dependent variable was set as medical care costs which are the sum of payments from diverse party such as private insurers and government. And the main independent variable was set as BMI. Also, four different kinds of confounders which are (1. Demographic factors 2.Demographic + Socioeconomics factors 3. Demographic+ Socioeconomics+ Additional factors 4. All three factors mentioned + obesity-related diseases) are used. Therefore, three different number of aspects of estimating medical costs in regression are used (age groups/ statistical models/ confounder adjustment). With many combinations on diverse aspects, the experiments were executed. The effects of using different statistical models

show not significant differences among them, but extended estimating equation showed the best performance. The results on this paper are that the two most significant factors of variability on estimating the expenditures were age groups and adjustment for obesity-related comorbid conditions.

iii. Brief Outline of Model

Since the dependent variables used and the statistically expected results' approach on 1) Estimating medical utilization across weight categories and 2) Estimating health care expenditures across weight categories are different, two models for each problem were used. Both problems were executed with converting the continuous data values of BMI (Body Mass Index) into categorical ordinal data with four choices. (0: BMI < 18.5, 1: 18.5 < BMI < 25, 2: 25 < BMI < 30, 3: BMI > 30). All the results along each BMI category are compared in order to determine which group has large impacts on dependent variables.

For the first one, Poisson regression model was used because the dependent variable is set as the number of times for a particular event. Three kinds of dependent variables were used (hospital visits, doctor visits, hospital visits + doctor visits). Visiting hospital or doctor was treated as an event and the number of visiting times were treated as event times that happen, so Poisson regression model is appropriate to use.

Moreover, for the second problem, multiple linear regression, the extension of ordinary least-squares (OLS) regression involving more than one explanatory variables, was used. Satisfying and confirming the assumptions that there is a linear relationship between both the dependent and independent variables, no major correlation between independent variables and etc. were easy to proceed unlike other regression models' constraints. Scaled total expenditure, which is treated in detail on 'description of data', was used as a dependent variable.

The importance of explanatory variables identified from statistical techniques mentioned above in each problem will be compared with ‘Random Forest’ model, one of the strongest ensemble machine learning models, since it has ‘feature importance’ function to show the magnitudes of importance among the variables we can compare results. Since it is based on tree model, it does not need any data preprocessing works such as scaling unlike Poisson regression and MLR.

iv. Description of Data

There are four main explanations for the data description from data explained on Table 1. They are the distribution of data through Exploratory Data Analysis (EDA), the statistics of variables (min, max, std etc.), checking multicollinearity, and finally the data preprocessing to use it as input data for statistic models.

The continuous variable ‘BMI’ was depicted demographic and socioeconomic variables using Boxplots, illustrated in Fig1~Fig7. In addition, BMI was replaced with the categorical variable, indicating the degree of the distribution between total expenditures, doctor visits and hospital visits according to each obesity level using Boxplots, illustrated in Figure8~10. Analyzing Box-Plots shows statistically insignificant results in both cases.

Additionally, statistics which use the values of mean, standard deviation, minimum, Q1, Q2, Q3 and maximum are explained in two ways which are data with continuous BMI values and categorical BMI values. The average of the elderly counted in this sample data is 74.1, and it seems that the sample of mostly women is relatively larger than that of men. The average period of education for these elderly people is 11.6 year. Their average income and average expenditure are \$22938.3 and \$8358 each, with standard deviation values in income is similar to those in average, and in total expense is about 2 times higher than its in average. Therefore, it needs ‘scaler’ process essentially for these two variables. And the scores of mental and

physical health reported by themselves are near 2.0, and they seem to be having better evaluated than 'good'. When changing BMI into categorical variables, there seems to be many people who are overweight and obese since the average value is about 2.4. Independent of the 'BMI' results, curiously, the cholesterol level appears to have about half of the population, and relatively few people suffer from diabetes. Finally, due to the high cost of hospital visits and treatment due to the absence of medical insurance in the United States, it seems that relatively more doctors' visits are occurred than the hospital visits.

Next, one of the most important points in regression analysis is to consider multicollinearity. If a linear relationship between independent variables appears, i.e., multicollinearity appears strongly, regression analysis will not proceed properly, and the results might be poor. Thus, to solve multicollinearity problems, the degree of correlation between two or more independent variables can be found using correlation matrix. This can be seen from the Heatmap figure which has a higher correlation on darker side, attached on Figure 11. The correlation between 'srhealth' and 'education' appears relatively high as 0.4, 'srhealth' and 'phy_lim' appears relatively high as 0.42, and 'income' and 'education' appears relatively high as 0.39. We can just delete one among 'srhealth', 'education', 'phy_lim' and 'income', then find out 'correlation matrix' again to check whether there is high correlation or not. However, to accurately determine the 'multicollinearity' of each variable, another measurement method is suggested in this paper. An indicator called Variance Inflation Factor (VIF) would be utilized, usually with a value of 10 or higher, which could be determined to have multicollinearity. Methods for deleting multicollinearity using VIF include: First, if there are similar variables which have same function, remove one of the two, then, delete the variables with high VIF coefficients. After deleting, print out VIF coefficients again from updated data. Repeat this process until the VIF coefficient is not greater than 10. If there is no value greater than 10, the

regression analysis can be carried out using the remaining variables. Thus, on this experiment, ‘education, srhealth, age’ which all show high VIF coefficients are dropped from original data and rest of variables were utilized on the regression models.

Also, some simple data preprocessing was done on continuous data ‘income’ and ‘total expense’ since their standard deviation and difference between maximum value and minimum value show extremely high. Two kinds of scalers were used, log-scaler and Robust-scaler. Log-scaler was selected since Robust-scaler results show extreme which can be seen from maximum values of ‘robust scaled income’ and ‘robust scaled total expense’ as 7.9 and 35.6. Usually, if the value which has been applied on ‘Robust scaler’ is greater than 1, it can be seen as extreme outlier, but maximum values on it are much larger than 1. On the other hand, log-scaler seems to handle extreme values more smoothly than the previous one so Log-scaler.

v. Discussion of Results

Before running a regression, t-test was done on each pair of groups in BMI except for under-weight group due to small samples. The null hypothesis on t-test is followed as two group’s average total expense, doctor visits and hospital visits, followed as (1) for each of two cases that assume to have equal variances or not. And the results of them are depicted on Table 6. Assuming a significance level of 5%, statistically significant value that the average values of the two groups are not same is only in a case consisting of doctor visits because their p-values are lower than 5%.

$$H_0: \mu = 0 \quad (1)$$

Building a model which estimates healthcare utilization among the elderly in the United States while using all the demographic, socioeconomic and diseases variables is conducted with assumption not considering ‘marital’, ‘race_grp’, ‘msa’ and ‘mntl_hlth’ as independent

variables in this problem since the value of AIC in all three models show higher about 2~5%. Dependent variable was set following three kinds of approaches *dr_visits*, *hosp_vis* and *dr+hosp_vis*, so three number of Poisson regression models were executed.

Due to small number of samples in underweight group smaller than over 10 times, I think its results do not have statistically significant meanings. So, the results on ‘underweight’ sample are not utilized and analyzed on results.

The Poisson regression equation on each group having dependent variable as doctor visit or hospital visits or sum of doctor visits and hospital visits are followed as (2).

$$\log(\hat{\lambda}) = \beta_0 + \beta_1 \text{income} + \beta_2 \text{male} + \beta_3 \text{highbp} + \beta_4 \text{highchol} + \beta_5 \text{phylim} + \beta_6 \text{diabetes} + \beta_7 \text{chd} \quad (2)$$

Following results are explained on Table (2), (3) and (4). AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are used as standard evaluation on each model. The smaller values in AIC and BIC, the more stable and better performance in the model. Even though the results only using hospital visits variable as dependent variable outperforms others in each group, due to bias value in hospital visits, about 80% is biased to ‘zero’, its model does not have significantly confidence results. Therefore, I strongly believe that considering hospital visits and doctor visits at the same time which means summation values of them would have statistically significant results. Of all groups, the obese-weight group model had the lowest AIC value. AIC in obese-weight is 10647, over-weight is 14627 and normal-weight is 12297. This can be concluded that the frequency of healthcare utilization occurs frequently in obesity groups than others. Also, when analyzing p-values on obese-weight group which has summation values of doctor visits and doctor visits as dependent variable, only ‘male’ variable is seemed to have statistically insignificant values, but others are statistically significant.

Also, building a model which estimates spending among the elderly in the United States while using only the demographics and socioeconomics variables or all the variables including additional ones as independent variables has similar procedures with previous model. Assumption on this problem is also same that, it does not consider ‘marital’, ‘race_grp’, ‘msa’ and ‘mntl_hlth’ as independent variables in this problem because model’s R-squared values are higher if not using them. The only two differences between this model and the previous one are type of regression models and dependent variables. Dependent variable is set as total expense and the model used in this problem is a multiple linear regression model.

Due to small number of samples in underweight group smaller than over 10 times, p-value on underweight group’s independent variables are high which close to 1.0. I think its results do not have statistically significant meanings. So, the results on ‘underweight’ sample is not also utilized and analyzed on results in this problem.

The multiple linear regression equation on each group having dependent variable as total expense is followed as (3).

$$TotalExpense = \beta_0 + \beta_1 income + \beta_2 male + \beta_3 highbp + \beta_4 highchol + \beta_5 phylim + \beta_6 diabetes + \beta_7 chd + \varepsilon \quad (3)$$

Following results is explained on Table (5). R-squared, AIC and BIC are used to evaluate the model’s performance. Although the value of r-squared is the highest in over-weight group among all groups, since all r-squared values in each model is lower than 10% it cannot be a plausible metrics to evaluate. Thus, by comparing the values in AIC and BIC, we can see that obese-weight group show the lowest value which means obese-group can be seen as a group that best describe total expenditure related to healthcare. Also, when analyzing p-values on

obese-weight group, only 'income', 'diabetes' and 'phy_lim' variables are seemed to have statistically significant values.

Random forest, one of the strongest machine learning ensemble models, was used to estimate both on health utilization and total expenditures among the elderly in the United States. Unlike regression models I used, Random Forest showed the best accuracy over 86% on both problems, so I assume that it has near-answer results and compare it with the above results. Checking the variable importance of random forest, 'srhealth', 'race_grp', 'msa', 'mntl_hlth' and 'marital' are insignificant variables same as my feature selection decision. However, I removed out 'age' as insignificant variable in VIF, random forest treats it significant on both problems about 13% to describe the models. Also, in my regression models, income was treated as insignificant variable due to high p-value, but random forest treats it significant on both problems about 20% to describe the models.

vi. Conclusion

According to the results on this research, the null hypothesis mentioned in introduction that obese does not have any impact on healthcare utilization and expenditures related to health service is rejected. Even though models' fitness in each model is low, it does not mean that its statistics are all wrong and useless. I strongly believe that since testing's main purpose is whether to find out obesity impacts on healthcare utilization and spending, focusing only on its purpose not the regression model's performance but statistics is plausible and appropriate on this paper.

vii. References

Centers for Disease Control and Prevention. 2021. *Obesity is a Common, Serious, and Costly Disease*. [online] Available at: <<https://www.cdc.gov/obesity/data/adult.html>> [Accessed 14 June 2021].

Musich, S., MacLeod, S., Bhattarai, G., Wang, S., Hawkins, K., Bottone, F. and Yeh, C., 2016. The Impact of Obesity on Health Care Utilization and Expenditures in a Medicare Supplement Population. *Gerontology and Geriatric Medicine*, 2, p.233372141562200.

Kim, D. and Basu, A., 2016. Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. *Value in Health*, 19(5), pp.602-613.

Statsmodels.org. 2021. *Introduction — statsmodels*. [online] Available at: <<https://www.statsmodels.org/devel/index.html>> [Accessed 14 June 2021].

Scikit-learn.org. 2021. *scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation*. [online] Available at: <<https://scikit-learn.org/stable/index.html>> [Accessed 14 June 2021].

viii. Appendix

Figure 1 BMI by age

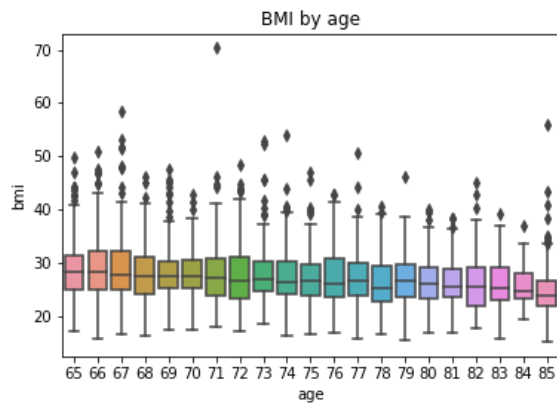


Figure 2 BMI by education

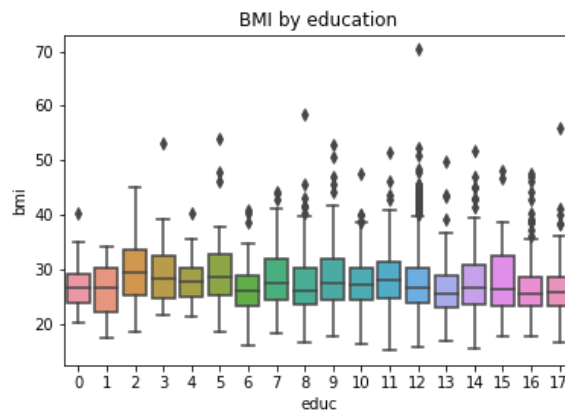


Figure 3 BMI by gender (0=female, 1=male)

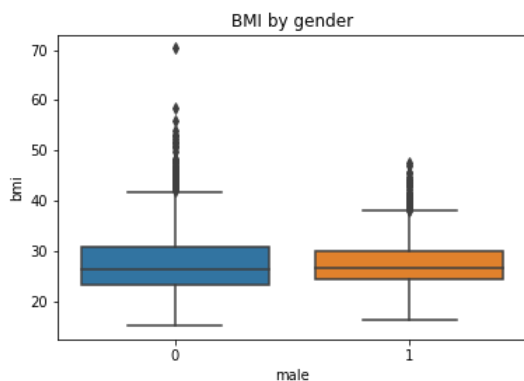


Figure 4 BMI by marital (1=married, 2=widowed, 3=divorced, 4=never married)

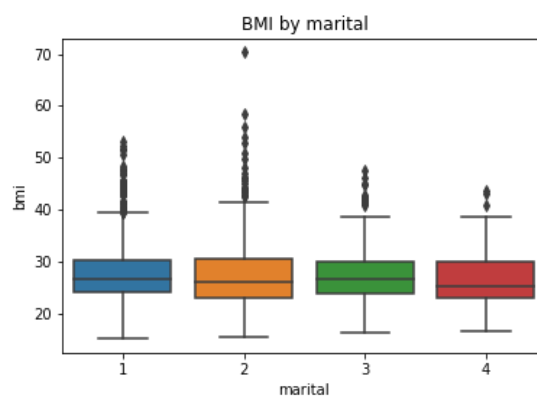


Figure 5 BMI by MSA (0=not live, 1=live)

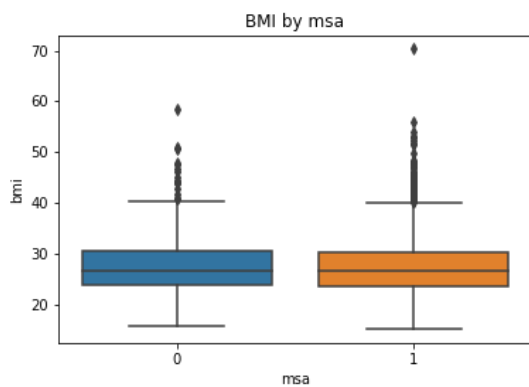


Figure 6 BMI by grouped race (1=White non-Hispanic, 2=black non-Hispanic, 3=other non-Hispanic, 4=Hispanic)

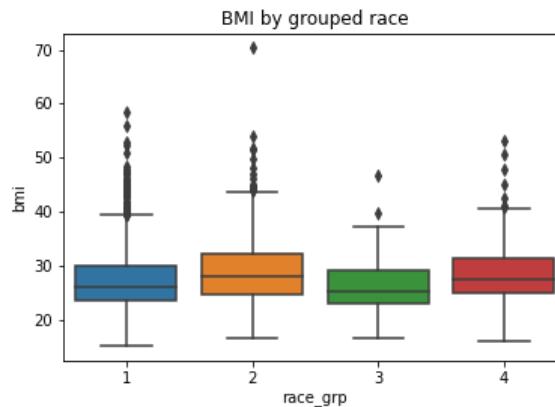


Figure 7 BMI by high blood pressure (1=Yes 0=No)

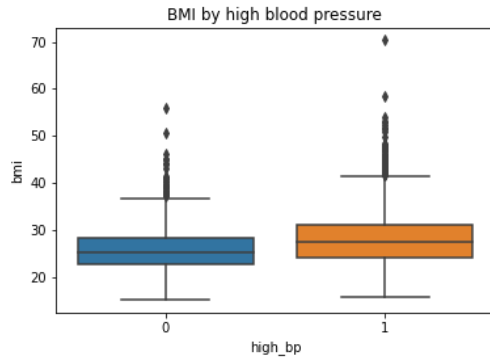


Figure 8 Doctor visit by BMI (0=under-weight, 1=normal, 2=over-weight, 3=obese)

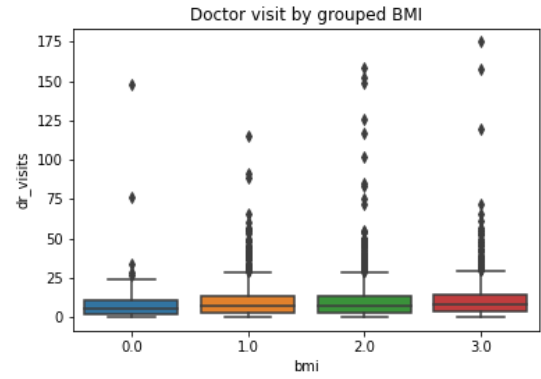


Figure 9 Hospital visit expense by BMI (0=under-weight, 1=normal, 2=over-weight, 3=obese)

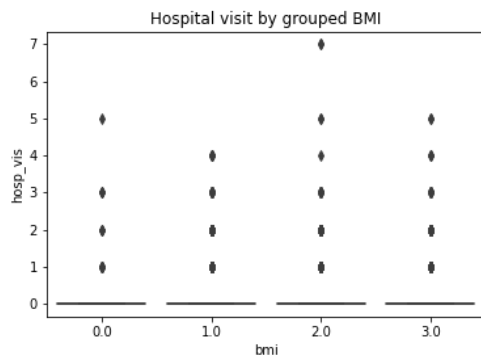


Figure 10 Total expense by BMI (0=under-weight, 1=normal, 2=over-weight, 3=obese)

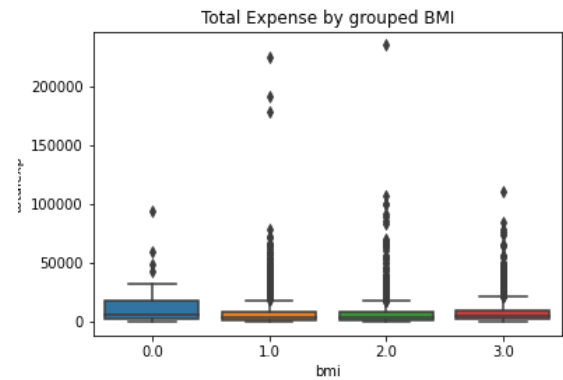


Figure 11 Data Correlation Matrix

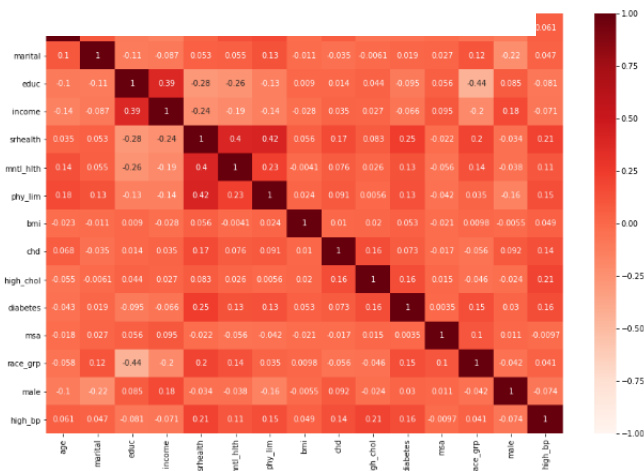


Table 1 Explanations on data variables

Variable Name	Definition
age	Age in yearss
male	Dummy variabe, =1 if male, 0 otherwise
race_grp	Race categorical vairable, =1 if the respondent is whitenon-Hispanic, =2 if black non-Hispanic, =3 if other non-Hispanic, =4 if Hispanic
marital	Marital status categorial variable, =1 if married, =2 if widowed, =3 if divorced or sepearted, =4 if never married
income	Annual family income
educ	Years of education
msa	Dummy variable, =1 if live in MSA, =0 otherwise
bmi	Body mass index, weight in kilograms/(height in cm-squared)
hihg_bp	Dummy variabe, =1 if respondent has high blood pressure, =0 otherwise
high_chol	Dummy variabe, =1 if respondent has high blood pressure, =0 otherwise
phy_lim	Dummy variabe, =1 if respondent has high blood pressure, =0 otherwise,
diabetes srhealth	chd Dummy variabe, =1 if respondent has diabetes, =0 otherwise / Dummy variabe, =1 if respondent has chronic heart disease, =0 otherwise, Self reported health status, =1 if excellent, =2 if very good, =3 if good, =4 if fair, =5 if poor
mntl_hlth	
dr_visits	Doctor visits in 2005. This variable is an integer value (0,1,2...)
hosp_vis	Hospital visits in 2005. This variable is an integer value (0,1,2...)
total exp	Total expenditures on medical care in 2005. This variable measures the value of medical care consumed by the respondent, not what they paid out of pocket.

Table 2 (Poisson Regression Result of Equation (1)- BMI normal weight group)

Dependent Variables	Doctor Visits Coefficient (std err), p-value	Hospital Visits Coefficient (std err), p-value	Doctor + Hospital Visits Coefficient (std err), p-value
intercept	1.87 (0.03), 0.00	-2.54 (0.18), 0.00	1.88 (0.02), 0.00
income	0.00 (0.00), 0.00	0.00 (0.00), 0.20	0.00 (0.00), 0.00
male	-0.05 (0.02), 0.03	0.40 (0.13), 0.002	-0.03 (0.02) 0.10
high_bp	0.07 (0.02), 0.001	0.32 (0.15), 0.04	0.08 (0.02), 0.00
high_chol	0.16 (0.02), 0.00	0.33 (0.14), 0.02	0.17 (0.02), 0.00
phy_lim	0.32 (0.02), 0.00	1.13 (0.14), 0.00	0.34 (0.02), 0.00
diabetes	0.21 (0.03), 0.00	0.05 (0.17), 0.78	0.21 (0.03), 0.00
chd	0.23 (0.03), 0.00	0.61 (0.16), 0.00	0.24 (0.03), 0.00
Results	Observations: 1029 AIC: 12274.73 BIC: 1568	Observations: 1029 AIC: 1201.02 BIC: -6282	Observations: 1029 AIC: 12297.46 BIC: 1554

Table 3 (Poisson Regression Result of Equation (1)- BMI over-weight group)

Dependent Variables	Doctor Visits	Hospital Visits	Doctor + Hospital Visits
	Coefficient (std err), p-value	Coefficient (std err), p-value	Coefficient (std err), p-value
intercept	1.88 (0.03), 0.00	-2.10 (0.18), 0.00	1.90 (0.03), 0.00
income	0.00 (0.00), 0.00	0.00 (0.00), 0.16	0.00 (0.00), 0.001
male	-0.08 (0.02), 0.00	-0.16 (0.13), 0.24	-0.08 (0.02) 0.00
high_bp	0.06 (0.02), 0.009	0.04 (0.15), 0.77	0.06 (0.02), 0.009
high_chol	0.17 (0.02), 0.00	0.26 (0.14), 0.06	0.17 (0.02), 0.00
phy_lim	0.38 (0.02), 0.00	0.66 (0.13), 0.00	0.39 (0.02), 0.00
diabetes	0.49 (0.02), 0.00	0.53 (0.13), 0.00	0.49 (0.02), 0.00
chd	0.24 (0.02), 0.00	0.88 (0.14), 0.00	0.26 (0.02), 0.00
Results	Observations: 1104	Observations: 1104	Observations: 1104
	AIC: 14573 BIC: 2953	AIC: 1315 BIC: -6780	AIC: 14627 BIC: 2966

Table 4 (Poisson Regression Result of Equation (1)- BMI obese-weight group)

Dependent Variables	Doctor Visits	Hospital Visits	Doctor + Hospital Visits
	Coefficient (std err), p-value	Coefficient (std err), p-value	Coefficient (std err), p-value
intercept	1.97 (0.03), 0.00	-2.13 (0.25), 0.00	1.98 (0.03), 0.00
income	0.00 (0.00), 0.00	0.00 (0.00), 0.06	0.00 (0.00), 0.00
male	0.02 (0.02), 0.34	-0.11 (0.16), 0.47	0.02 (0.02) 0.42
high_bp	0.06 (0.03), 0.043	0.29 (0.21), 0.17	0.06 (0.03), 0.03
high_chol	0.07 (0.02), 0.003	0.33 (0.16), 0.04	0.08 (0.02), 0.001
phy_lim	0.28 (0.02), 0.00	0.80 (0.16), 0.00	0.29 (0.02), 0.00
diabetes	0.21 (0.02), 0.00	-0.05 (0.15), 0.71	0.20 (0.02), 0.00
chd	0.20 (0.03), 0.00	0.61 (0.17), 0.00	0.21 (0.03), 0.00
Results	Observations: 775	Observations: 775	Observations: 775
	AIC: 10570 BIC: 2588	AIC: 1024.34.48 BIC: -4417	AIC: 10646.55 BIC: 2641

Table 5 (Multiple Linear Regression Result of Equation – dependent variable: total expense)

Data source	Normal-weight group	Over-weight group	Obese-weight group
	Coefficient (std err), t, p-value	Coefficient (std err), t, p-value	Coefficient (std err), t, p-value
intercept	5.93 (0.77), 7.71, 0.00	6.22 (1.02), 6.13, 0.00	6.50 (1.42), 4.57, 0.00
income	0.00 (0.00), 2.88, 0.004	0.00 (0.00), 0.76, 0.45	0.00 (0.00), 2.34, 0.02
male	-0.29 (0.69), -0.41, 0.68	-0.83 (0.84), -1.00, 0.32	0.18 (1.07), 0.17, 0.87
high_bp	0.72 (0.70), 1.02, 0.31	0.46 (0.90), 0.51, 0.61	0.64 (1.27), 0.50, 0.61
high_chol	1.62 (0.68), 2.37, 0.02	1.68 (0.84), 2.00, 0.046	0.80 (1.03), 0.77, 0.44
phy_lim	3.55 (0.72), 4.96, 0.00	4.39 (0.89), 4.97, 0.00	3.18 (1.01), 3.15, 0.002
diabetes	2.37 (1.03), 2.30, 0.02	6.33 (1.01), 6.23, 0.00	2.39 (1.07), 2.24, 0.03
chd	2.93 (1.09), 2.69, 0.07	3.31 (1.19), 2.79, 0.005	2.77 (1.50), 1.85, 0.07
Results	Observations: 1029 R-squared: 0.060, AIC:7769, BIC:7809	Observations: 1104 R-squared: 0.088, AIC:8864, BIC:8904	Observations: 775 R-squared: 0.037, AIC:6246, BIC:6283

Table 6 (5% significance level T-test results on total expense / doctor visits / hospital visits among groups)

Types	Pair for two groups	Tstatistics (equal variance)	p-values (equal variance)	Tstatistics (! equal variance)	p-values (! equal variance)
Total Expense	Normal, Overweight	0.55	0.59	0.54	0.59
	Normal, Obese	-1.14	0.25	-1.81	0.24
	Overweight, Obese	-1.79	0.07	-1.83	0.07
Doctor Visits	Normal, Overweight	-1.46	0.15	-1.47	0.14
	Normal, Obese	-2.70	0.006	-2.61	0.009
	Overweight, Obese	-1.19	0.23	-1.20	0.23
Hospital Visits	Normal, Overweight	0.16	0.87	0.17	0.87
	Normal, Obese	-0.98	0.33	-0.97	0.33
	Overweight, Obese	-1.11	0.27	-1.11	0.27
Doctor Visits + Hospital Visits	Normal, Overweight	-1.43	0.15	-1.44	0.15
	Normal, Obese	-2.73	0.006	-2.64	0.008
	Overweight, Obese	-1.23	0.22	-1.23	0.22