Determinants of the Overweight and Obesity – A Data Analysis on Population of England

Author: Yu Yuan

2020/12/21

Abstract

World Health Organization stated that the total number of people who are disturbed by overweight and obesity issues have been significantly and rapidly increased since the early 21st century, in both developing and developed countries. For investigating the potential determinants of these unpleasant problems, we focused on a data set from Health Survey for England (HSE) by the method of multilevel regression on different neighborhood areas. The results showed that socio-demographic variables, such as age and gender were significantly associated with the BMI values for England adults. Statistical evidence supports the statement that the higher the level of qualification, the less likely to be overweight and obesity.

Keywords:

Overweight and obesity, BMI, Variance Component Model, Likelihood Ratio test, Multilevel Regression.

Introduction

With the rapid development in the fields of global economy and scientific technology, the quality of human life has been continuously improved. However, the diet-related behaviors of human beings are also significantly influenced regarding the accelerated pace of life, result in a public health issues in both developing and developed countries, that is, the prevalence of overweight and obesity. People who are encountered with the problem of overweight and obesity also suffering from certain chronic and acute diseases, for instance, diabetes and heart attach. Walter et al. (2009) investigated the effect of overweight and obesity on mortality and found that its implication refers to disability-free life expectancy was associated with a higher risk in the older people [1]. Also Bjerregaard (2018) stated that childhood overweight was linked to an increased risk of type 2 diabetes in adulthood, nevertheless, the risk could be reduced by the remission of overweight before adulthood could reduce this risk [2].

BMI, abbreviated from Body mass index, was the most commonly used concept for evaluating relative overweight and obesity. It was initially proposed by Adolphe Quetelet during the 19 century [3] and rephrased as the criteria of correlation between relative weight and height by Ancel Keys et al. in 1972 [4]. According to the criteria from the Centers for Disease Control and Prevention (CDC), a person whose BMI is the range between 25 and 30 are overweight, and would be classified as obese range if he or she has a BMI greater than 30. As a fraction between weight and height, the basic formula of computing BMI is straightforward as follows:

$$BMI = \frac{Mass_{kg}}{Height_m{}^2}$$

The causes related to overweight and obesity are still a complicated issue involving multiple factors: demographic and geographic features, genetic hereditary and social culture. In general, two crucial determinants are frequently associated with overweight and obesity, that is, extravagant dietary intake and insufficient exercise. Males and younger adults were less likely to be inactive in American rural areas, result in a high prevalence of overweight and obesity, and inactive lifestyles among rural populations, according to Patterson's research [5]. Also, Correll et al. (2010) linked the phenomenon of increasing obese and overweight with social welfare policies issued from American government [6].

Consequently, recognizing the link between overweight and obesity and potential factors could be fundamentally necessary for protecting people from chronic or acute disease resulted from overweight and obesity, as well as improving their quality of daily life. In this report, we worked on a dataset from Health Survey for England (HSE) conducted in 2013, which monitored trends in the nation's health and care. The survey consists of multiple core questions and measurements in each series, such as blood pressure, height and weight measurements and analysis of blood and saliva samples. We initiated our study by the process of data cleaning and explanatory data analysis. Then we investigated the the variations of BMI across socio-demographic characteristics by a series of multilevel models. Additionally, Likelihood ratio test was applied in our research for selecting an optimal combination of covariates for modelization. The validation of our models were eventually verified by our diagnostic plots.

GitHub Repo for this study: https://github.com/cristalyu/Determinants-of-the-Obesity-and-Overweight

Preliminary Analysis

Data source and Cleaning

The dataset we focused on is from the Health Survey for England (HSE) implemented in 2013. 14,836 adults and teenagers aged above 16 from 732 neighborhood areas in were interviewed in this survey. It consists of 11 variables ranging from socio-demographic characteristics, such as age, gender and ethnicities, to geographic features, like neighborhood areas. Before we initiating our data analysis, we detected the variables features and missing values in the original dataset. From Figure 1, both quantitative and qualitative variables are included while several missing values could be observed as well. Regarding an unbiased estimation in later analysis, we conducted the data cleaning by removing these missing values with NA records or non-specific responses. 13,003 cases were eventually remained.

Descriptive Statistics

As a response variable we were interested, we found that the distribution of BMI among this surveyed population is approximately bell-shaped as shown in Figure 2, even with a slight tend of right tail. Additionally, the distribution of the mean BMI across 732 different neighborhoods area also display a roughly normal figure. Specifically, the graphical evidence already provided us a certain conclusion that the prevalence of overweight and obesity had been overwhelming in 2013 because more than half of the surveyed population processed greater BMI than the standard line of overweight, as a fact that the mean and median values of BMI are 26.97 and 26.36, respectively.

Figure 1. Variable features and Missing values in a subset dataset from the original

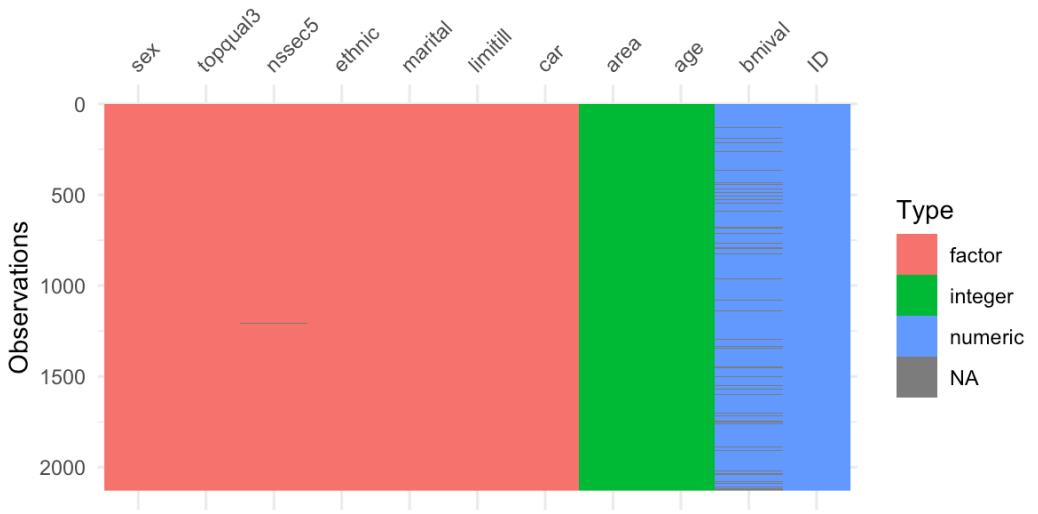


Figure 2. Distribution of BMI for the survey population of England in 2013

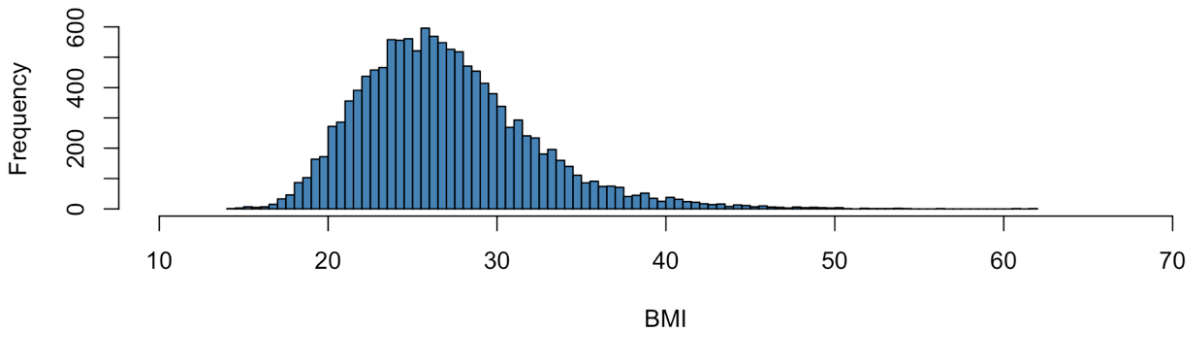
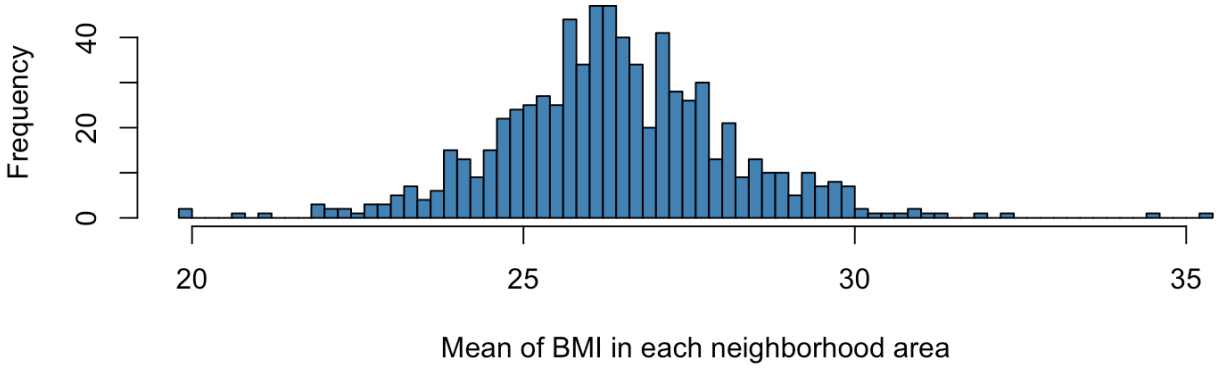


Figure 3. Distribution of BMI across 732 neighborhoods areas in England



For more comprehensive insight of this dataset, we summarized several basic descriptive statistics as shown in Table 1. There are approximately 60% of the surveyed adults and teenagers aged above 16 had issues of overweight and obesity. Besides the average BMI for females was slightly smaller than the one for males, however, the distribution of females would be more dispersedly spread due to its larger deviation. We were also aware of that BMI could vary within different levels of education, occupation, ethnicities, marital condition, medical history and way of transpiration. For example, Table 1 implied that the people owned cars had a larger averaged BMI than others; a single person seemed to be much thinner, however, it could be a confounder with the effect of age. On the other hand, the boxplot in Figure 3 also indicated that the BMI values were likely to increase with age, and the ethnicity of White population tended to process a greater BMI than the remaining minorities.

Figure 3. Distribution of BMI within each variable

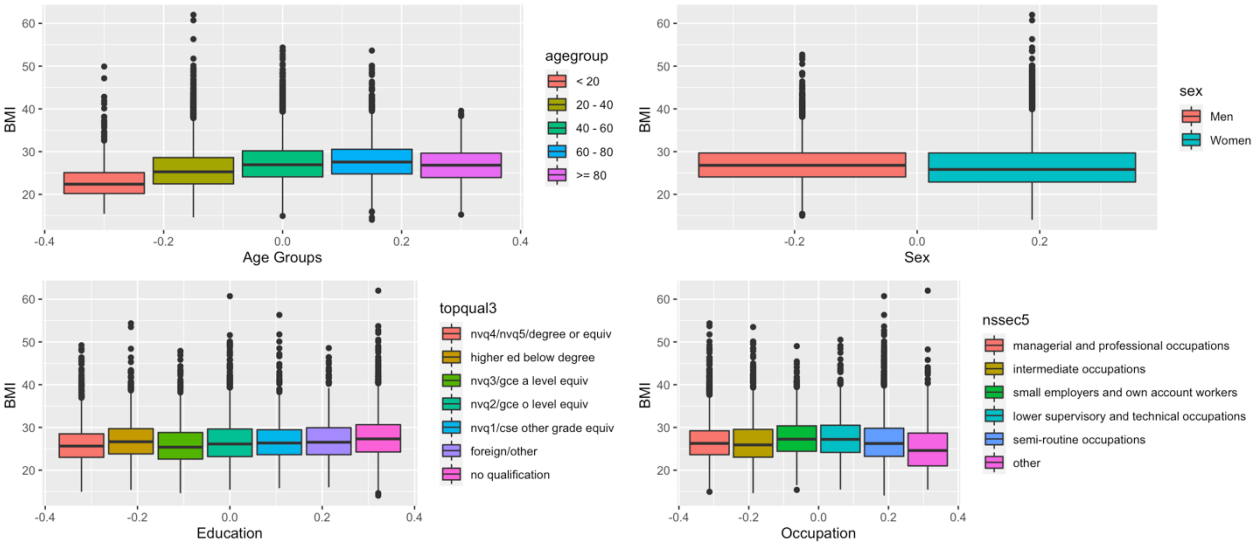


Figure 3 (Continued). Distribution of BMI within each variable

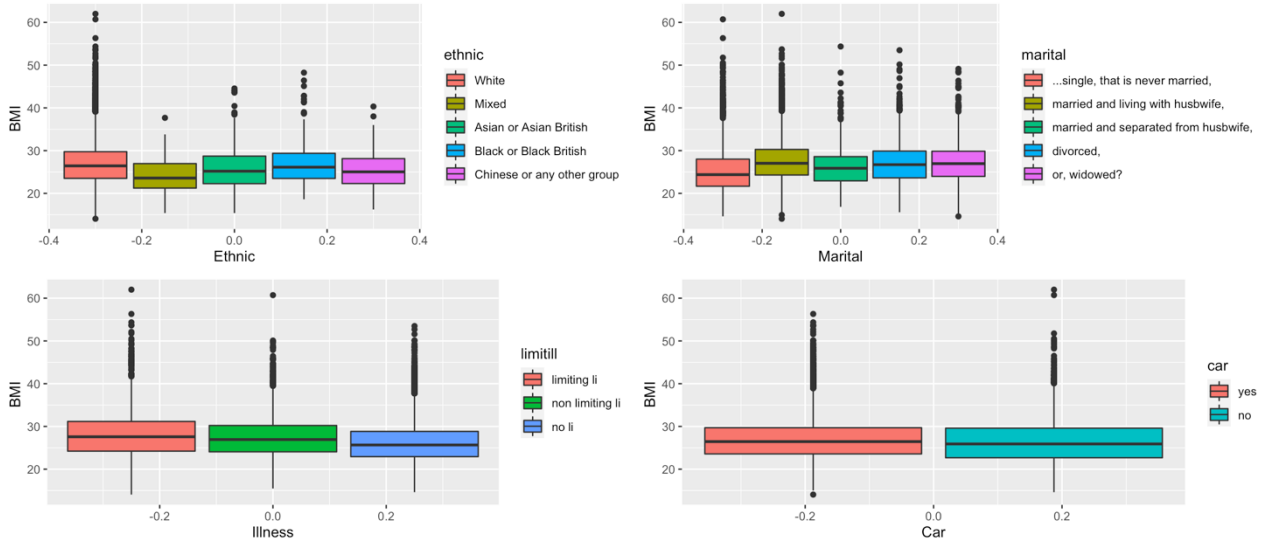


Table 1. Descriptive statistics

	Proportion	Mean	Std. Deviation	Minimum	Maximum
BMI		26.97	5.07	14.06	61.99
Underweight	1.56%				
Normal	36.77%				
Overweight	39.43%				
Obesity	22.24%				
Age		47.38	17.73	16.00	98.00
Sex					
Male	45.68%	27.23	4.48	15.12	52.54
Female	54.32%	26.76	5.48	14.13	61.99
Education					
nvq4/nvq5/degree	16.49%	26.21	4.54	14.93	49.24
nvq3	12.12%	26.16	4.91	14.61	47.88
nvq2	24.45%	26.73	5.23	15.42	60.69
nvq1	5.40%	26.99	5.40	15.80	56.28
foreign/other	4.67%	27.19	5.22	16.00	48.50
no qualification	25.99%	28.79	5.16	14.05	61.99
below degree	10.86%	27.22	4.75	15.42	54.54
Occupation					
professional	31.94%	26.23	4.59	14.88	54.64
intermediate	12.93%	25.89	5.29	14.48	53.49
own worker	8.43%	27.31	4.83	15.85	49.02
technical	9.92%	27.19	4.73	15.32	50.52
Semi-routine	32.80%	26.22	5.39	14.54	60.71
others	3.98%	24.64	5.98	15.32	61.99
Ethnic					
White	92.66%	27.32	5.21	14.32	61.99
Mixed	0.61%	24.54	4.21	15.46	37.67

Asian	3.90%	25.84	4.54	15.65	44.55
Black	1.82%	27.23	5.21	18.32	48.24
Chinese or others	1.02%	25.65	4.54	16.32	40.32
Marital					
single	26.23%	25.12	5.21	14.54	60.71
married	54.93%	27.43	4.40	14.65	61.99
married(separated)	2.75%	26.65	5.32	16.78	54.35
divorced	8.69%	27.76	4.943	15.43	53.49
other	7.41%	27.43	4.32	14.32	49.12
Limiting illness					
limit	24.53%	28.32	5.32	14.54	61.99
non-limit	21.42%	26.54	5.54	15.54	60.71
no illness	54.05%	25.23	4.23	14.30	53.49
Car					
yes	81.77%	27.32	4.32	14.43	56.31
no	18.23%	26.69	5.65	22.32	61.99

Methods

In our case, the interviewee's data is clustered and has a hierarchical structure of two levels: socio-demographic characteristics as level 1 and neighborhood areas as level 2, respectively., then the multilevel modeling was utilized in our exploration.

Variance Component Model

For confirming the the variation of BMI across neighborhood areas, we initiated our modelization with a variance component model as follows:

variance component model:

$$BMI_{ij} = \beta_0 + u_j + \varepsilon_{ij}$$

where BMI_{ij} is the BMI value for ith individual resident in jth neighborhood area, u_j is the deviation of BMI values of jth neighborhood area from average, ε_{ij} is error of model fitting. The variation of BMI values across neighborhood areas could be explained by variance partition coefficient, which could be computed as:

$$\rho = \frac{{\sigma_u}^2}{{\sigma_u}^2 + {\sigma_\varepsilon}^2}$$

where ${\sigma_u}^2$ is the variance between neighborhood areas (level 2), and ${\sigma_\varepsilon}^2$ is the variance generated within neighborhood areas (level 1). If there did exist the effect between different neighborhood areas that contribute to the variation of BMI, then the ρ should be significantly greater than 0. For verifying this effect on level 2, a Likelihood Ratio (LR) test had been implemented between a variance component model and a single level linear model in which only contains an intercept term as follows:

Null model:

$$BMI_{ij} = \beta_0 + \varepsilon_{ij}$$

The LR test statistic with the null hypothesis that the $u_j = 0$ could be computed by the logarithm of the fraction on the likelihood values. That is,

$$LR\ statistics = -2ln\left(\frac{LogLik(Null\ Model)}{LogLik(VC\ Model)}\right)$$

Then, the LR statistic should follow a Chi-square distribution, ${\chi^2}_{df=1}$, if the null hypothesis is true. If there is sufficient evidence to reject the null hypothesis, then the application of multilevel modeling is reasonable for this dataset, otherwise, considering the effect of neighborhoods areas as a hierarchical level might be unnecessary in this case.

Random Intercept Model

For certain hierarchical data, that is, confirmed variation of BMI across different neighborhood areas, a random intercept model could be derived from a variance component model by controlling fixed effects on level 1, which has a basic form as follows:

Random intercept model:

$$BMI_{ij} = \boldsymbol{X}_i\beta + u_j + \varepsilon_{ij}$$

where $\boldsymbol{X}_i$ is the covariate matrix containing all the fixed effects by socio-demographic variables in our dataset, for example, the age (group), gender, education background and occupation levels. In order to select the "best" combination of covariates for modelization, we applied the following algorithms with LR tests.

1. Initiate with variance component model without any explanatory variable;
2. Add a single fixed effect of an independent variable on variance component model (or concise model with less covariates) to generate a random intercept model as a candidate model;
3. Compute the LR statistic between variance component model (or concise model with less covariates) and the candidate model in step 2;
4. Compare the p-value of LR test with a significance level $\alpha = 0.05$, that is, if p-value <0.05, it provides sufficient evidence to reject the null hypothesis that variance components model (or concise model with less covariates) performs better. Then we select the new model as the criterion model for next comparison. Otherwise, we keep the simpler model and consider to add another covariate if p-value is greater than 0.05;
5. Repeat the steps of 2 – 4 until we have a candidate model without a p-value of LR test being greater than 0.05 or the number of independent variables is greater than 5.

Results

By applying R programming, we have the variance partition coefficient equal to 0.0233, implying that the differences across neighborhood areas only accounted for 2.23% of the

variation in BMI. Although this number seems to be inappreciable, the result by LR test (LR statistic = 23510.22) provided significantly sufficient evidence to reject the null hypotheses that there was not any random effect across neighborhood areas at a level α =0.05. Therefore, the BMI did vary across the different areas in England and the utilization of multiple modeling could be appropriate in this scenario.

Table 2. LR tests for variable selection

Model	Explanatory Variables	Log Likelihood	*df*	p-values
0	NULL	-39539.57	3	*NA*
1	… + Sex	-39533.43	4	0.0005 (vs. model 0)
2	… + Sex + Age	-39298.28	5	0.0000 (vs. model 1)
3	… + Sex + Age + Qualification	-39275.75	11	0.0000 (vs. model 2)
4	… + Sex + Age + Qualification + Occupation	-39259.84	16	0.0000 (vs. model 3)
5	… + Sex + Age + Qualification + Occupation + Car	-39244.98	17	0.0000 (vs. model 4)

From Table 2, the *model 5* containing variables *sex, age, qualification, occupation* and *car* is the "best" subset model by the algorithm and LR test method that we have discussed above.

Model 5:

$$BMI_{ij} = \beta_0 + \beta_1 Sex_{i1} + \beta_2 Age_{i2} + \beta_3 Education_{i3} + \beta_4 Occupation_{i4} + \beta_8 Car_{i8} + u_j + \varepsilon_{ij}$$

The regression results of the *Model 5* are summarized in Table 3, with a baseline of Male adults who own a car, qualification and occupation are degree or equivalent, and professional, separately. The regression results implicate that the fixed effect of gender, age and process of car are significantly associated with the variation of BMI. Specifically, the older people tend to be sensitive to accumulate fat than the youths. The females are averagely thinner than males, resulted from the difference in physiology between women and men, or the slogan of current culture "thin is beauty". Likewise, people who own cars would be at a higher risk of being overweight and obesity than those who do not have cars. Additionally, we noticed that the people without any qualification or lower qualification are more likely to obtain a higher BMI than those with degree or equivalent qualification. However, the interpretation for occupation could be ambiguous since several terms estimated with insignificant coefficients. Furthermore, we depicted the predicted marginal effect for these variables with 95% confidence interval in Figure 4, the expected BMI values display a random variation across different levels in each covariate.

	Estimate	Std. Error	t value	p-value
(Intercept)	26.674	0.122	218.779	0.000

Sex Women	-0.232	0.091	-2.547	0.011
age	0.832	0.051	16.353	0.000
topqual3higher ed below degree	0.645	0.173	3.727	0.000
topqual3nvq3/gce a level equiv	0.059	0.172	0.343	0.732
topqual3nvq2/gce o level equiv	0.602	0.152	3.948	0.000
topqual3nvq1/cse other grade equiv	0.586	0.231	2.539	0.011
topqual3foreign/other	0.323	0.244	1.325	0.185
topqual3no qualification	0.994	0.173	5.754	0.000
nssec5intermediate occupations	-0.030	0.153	-0.193	0.847
nssec5small employers and own account worker	0.281	0.177	1.589	0.112
nssec5lower supervisory and technical occupation	0.415	0.169	2.450	0.014
nssec5semi-routine occupations	-0.073	0.131	-0.555	0.579
nssec5other	-0.801	0.251	-3.194	0.001
Car no	-0.652	0.119	-5.456	0.000

Table 3. Regression estimation on multilevel *Model 5*

Figure 4. Predicted marginal effect on *Model 5*

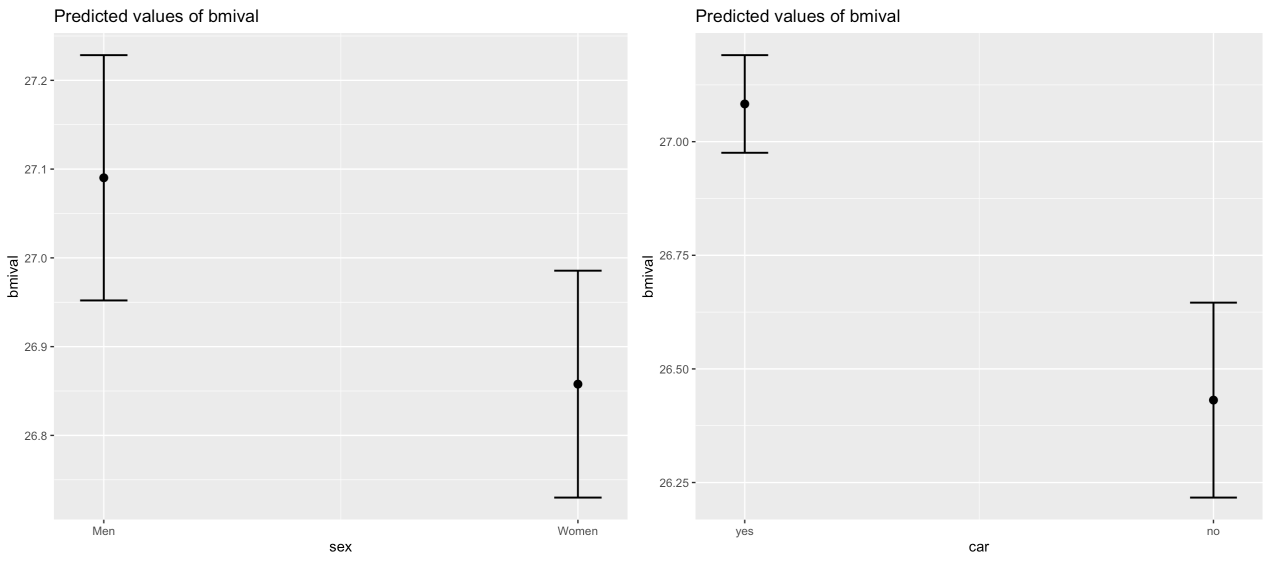
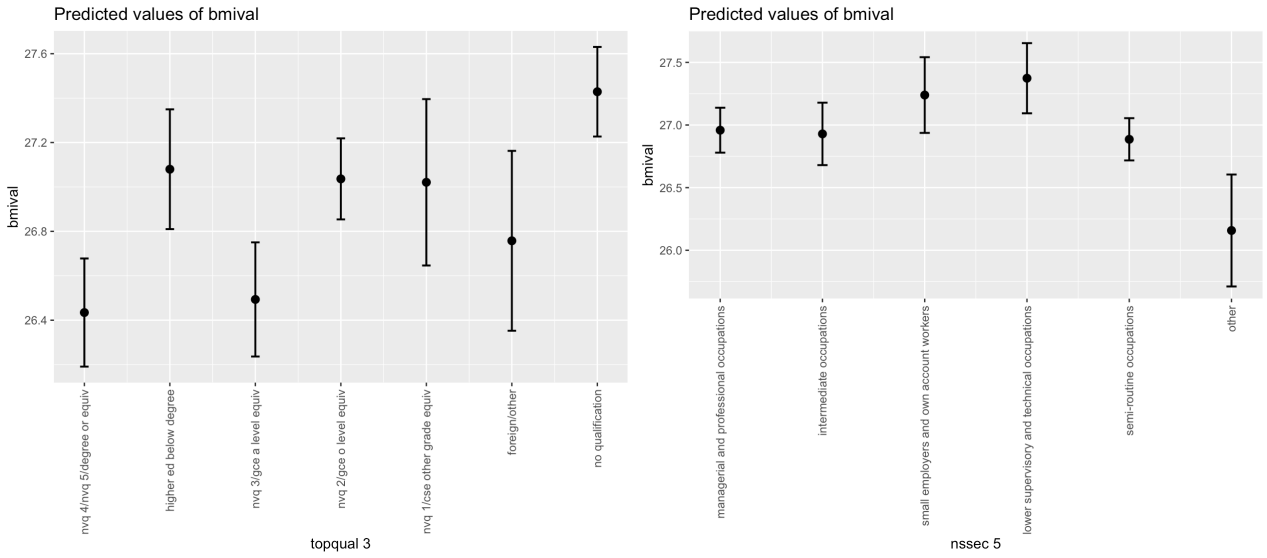
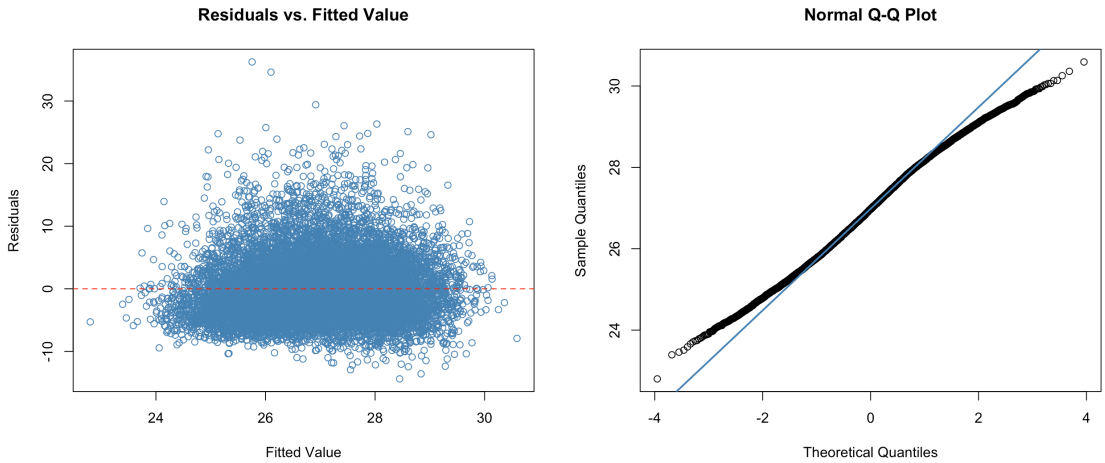


Figure 4 (Continued). Predicted marginal effect on *Model 5*



Moreover, the diagnostic plots for *Model 5* show that the residuals are randomly dispersed around the horizontal line of $y = 0$ without any distinct patterns. Besides, the majority of dots are affiliated with the diagonal straight line in normal Q-Q plot such that the normality assumption could be roughly satisfied, although several sample quantiles are deviated at the head and tail. Therefore, the model we selected by our algorithm and LR test is moderately appropriate for fitting and interpreting the dataset. Other than these facts, the variance partition coefficient in *Model 5* gradually dropped to 1.78% if we successively added an individual level socio-demographic variable on the variance component model, implying that our selected variables did did help to explain the variation of BMI across neighborhood areas.

Figure 5. Diagnostic plots for *Model 5*



Discussions

In this report, we focused on a clustered and hierarchical dataset from Health Survey for England and explored several determinants that might be associated with the BMI values across different neighborhood areas. The variation of BMI within in level 2, neighborhood areas, was initially detected by the Likelihood Ratio test between a variance component model and a null model. That is, the average BMI in each neighborhood areas were significantly different with each other, and this index seemed to be impacted by the social environment. Then we established a random intercept model by a specific algorithm of selecting variables, the "best" subset model contained 5 fixed effects: sex, age, qualification, occupation and car. The regression results indicated that male adults generally had a greater BMI values than female adults in England. Besides, the way of transportation seemed to be related with the BMI as well, specifically, people who drove cars tended to have greater BMI values, probably resulted from their physical inactivity. Moreover, we found that the qualification level had an impact on the BMI, people obtained a senior qualification, like a degree or equivalent, were more likely to maintain a lower BMI values.

Although interesting facts and conclusions have been found, further work need to be considered for improving the integrity of this study. For instance, we could transform the continuous response, BMI value, as a dichotomous variable. Specifically, we could have $y = 1$, if BMI ≥ 25; otherwise $y = 0$. Besides, the regression results from R provided a warning message about singularity, a Bayesian model might be available to solve this problem. Additionally, since the original dataset contained a large number of neighborhood areas across the whole England, we may consider a spatial random effect regarding as a geographic factor, and a BYM model (Besag, York and Mollié model) could be an option for further study.

Reference

[1] Walter, S., Kunst, A., Mackenbach, J., Hofman, A., & Tiemeier, H. (2009). Mortality and disability: the effect of overweight and obesity. International journal of obesity, 33(12), 1410-1418.

[2] Bjerregaard, L. G., Jensen, B. W., Ängquist, L., Osler, M., Sørensen, T. I., & Baker, J. L. (2018). Change in overweight from childhood to early adulthood and risk of type 2 diabetes. New England Journal of Medicine.

[3] Garabed Eknoyan (2008). Adolphe Quetelet (1796–1874)—the average man and indices of obesity, Nephrology Dialysis Transplantation, 23(1), 47–51.

[4] Keys, A., Fidanza, F., Karvonen, M. J., Kimura, N., & Taylor, H. L. (1972). Indices of relative weight and obesity. Journal of chronic diseases, 25(6-7), 329-343.

[5] Patterson, P. D., Moore, C. G., Probst, J. C., & Shinogle, J. A. (2004). Obesity and physical inactivity in rural America. The Journal of Rural Health, 20(2), 151-159.

[6] Correll, M. (2010). Getting fat on government cheese: the connection between social welfare participation, gender, and obesity in America. Duke J. Gender L. & Pol'y, 18, 45.

[7] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4), 348-354.

[8] Stafford, M., Brunner, E. J., Head, J., & Ross, N. A. (2010). Deprivation and the development of obesity: a multilevel, longitudinal study in England. American journal of preventive medicine, 39(2), 130-139.

[9] McTigue, K. M., Cohen, E. D., Moore, C. G., Hipwell, A. E., Loeber, R., & Kuller, L. H. (2015). Urban neighborhood features and longitudinal weight development in girls. American journal of preventive medicine, 49(6), 902-911.

[10] Tierney N (2017). "visdat: Visualising Whole Data Frames." _JOSS_, *2*(16), 355. doi:10.21105/joss.00355 (URL: https://doi.org/10.21105/joss.00355), <URL:http://dx.doi.org/10.21105/joss.00355>.

[11] Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S. Y., Freimer, N. B., ... & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nature genetics, 42(4), 348-354.

[12] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[13] Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr

[14] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[15] Lüdecke D (2020). _sjPlot: Data Visualization for Statistics in Social Science_. R packageversion 2.8.6, <URL: https://CRAN.R-project.org/package=sjPlot>.

[16] Sarkar, Deepayan (2008) Lattice: Multivariate Data Visualization with R. Springer, New York. ISBN 978-0-387-75968-5