**Quantitative Methods – Coursework 1**

**Which interventions** **appear to be the most effective to address childhood obesity in England?**

# Introduction

Childhood obesity is emerging as one of the most severe public health threats of the 21st century, the proportion of impacted children has risen tens of times over five decades (*Children: new threats to health*, 2020). These children may be at enhanced risk for a series of diseases as they grow older due to obesity (Simmonds *et al.*, 2016). In order to cope with this situation, funding was assigned by local authorities throughout England during the period from 2008 to 2018, with various interventions to mitigate the growth rate of obesity.

Therefore, the research question is which interventions seem to be the most effective to address childhood obesity in England? The objectives of this research are to use the provided data to investigate an appropriate way to express the variation of childhood obesity, and to analyze the relationship between the interventions presented and the variation of childhood obesity, then eventually assist in planning future interventions and allocating funding.

# Data

The data investigated for the research indicated the number of total cases of childhood obesity (which is called 'Condition X' in the data) and the total population of each area in 2008, 2013 and 2018 these three years, as well as the respective number of male and female in 'Condition X' and total population during the corresponding years. Annual budget was allocated to each local authority by six interventions which were: improving air quality; cleaning public spaces; training health professionals; raising awareness in schools; raising awareness through the media; and subsidising counselling services. In addition, different types of regions and local authorities were also listed. *Table 2-1* revealed the summary of the original data.

Table 2‑1 Summary of childhood obesity cases, population, intervention, region and local authority type

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Local authority area** | **2008 cases total** | **…** | **2013 pop total** | **…** | **2018 cases male** | **…** | **2018 pop female** | **Total budget** | **Clean air** | **…** | **Region** | **Local authority type** |
| 0 | Barking and Dagenham | 593 | **…** | 179030 | **…** | 230 | **…** | 103796 | 139000 | 21000 | **…** | London | London borough |
| 1 | Barnet | 539 | **…** | 344062 | **…** | 292 | **…** | 194186 | 220000 | 50000 | **…** | London | London borough |
| 2 | Barnsley | 360 | **…** | 244010 | **…** | 212 | **…** | 130828 | 160000 | 41000 | **…** | Yorkshire and the Humber | Metropolitan borough |
| **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** | **…** |
| 149 | Wolverhampton | 1442 | **…** | 252117 | **…** | 492 | **…** | 140604 | 201000 | 18000 | **…** | West Midlands | Metropolitan borough |
| 150 | Worcestershire | 2718 | **…** | 605974 | **…** | 1204 | **…** | 317961 | 470000 | 128000 | **…** | West Midlands | Non-metropolitan county |
| 151 | York | 1170 | **…** | 202851 | **…** | 518 | **…** | 114667 | 140000 | 5000 | **…** | Yorkshire and the Humber | Unitary authority |

# Methodology and Results

To investigate the research question, Python was used to modify and improve the data. For the first step, it was necessary to expressing the variation of childhood obesity properly, this value would be regarded as the dependent variable to build a multiple regression with independent variables including the interventions, regions and local authorities. Therefore, the rate difference between 2008 and 2018 was eventually decided as the dependent variable which was calculated by:

Where *2018* *X rate* and *2008* *X rate* was the rate of childhood obesity in 2018 and 2008 respectively.

And the rate of childhood obesity was calculated by (set 2008 as an example):

Where *2008 cases total* was the total number of obesity in 2008 and *2008 pop total* was the populations of areas in 2008.

Then after eliminating the irrelevant data column and encoding the regions local authorities, *Table 3-1* showed the refined data.

Table 3‑1 Improved data of childhood obesity cases rate, interventions, region code and local authority type code

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Rate difference** | **2008 X rate** | **2018 X rate** | **Total budget** | **Clean air** | **…** | **Region North West** | **…** | **Local authority type london borough** | **…** |
| 0 | 0.0008 | 0.0034 | 0.0042 | 139000 | 21000 | **…** | 0 | **…** | 1 | **…** |
| 1 | 0.0006 | 0.0016 | 0.0022 | 220000 | 50000 | **…** | 0 | **…** | 1 | **…** |
| 2 | 0.0005 | 0.0016 | 0.0020 | 160000 | 41000 | **…** | 0 | **…** | 0 | **…** |
| **…** | **…** | **…** | **…** | 161000 | 33000 | **…** | **…** | **…** | **…** | **…** |
| 149 | 0.0005 | 0.0060 | 0.0065 | 201000 | 18000 | **…** | 0 | **…** | 0 | **…** |
| 150 | 0.0002 | 0.0049 | 0.0051 | 470000 | 128000 | **…** | 0 | **…** | 0 | **…** |
| 151 | 0.0007 | 0.0060 | 0.0067 | 140000 | 5000 | **…** | 0 | **…** | 0 | **…** |

In order to check and exclude multicollinearity between these variables before applying linear regression, firstly correlation matrix was created by using Python (as shown in *Figure 3-1*), it could be noticed that there were several attributes emerging highly correlated, which indicated that a few of variables was required to be eliminated.

Graphical user interface

Description automatically generated

Figure 3‑1 Correlation matrix

Therefore, Python was used to remove the attributes (*Figure 3-2*) with highest VIF by creating a function. Then a multiple linear regression model was built and fitted in *Figure 3-3*, the R2 and Adjusted R2 were 0.628 and 0.580 respectively.

Figure 3‑2 The eliminated variables

Text

Description automatically generatedTable

Description automatically generated

Figure 3‑3 OLS regression results

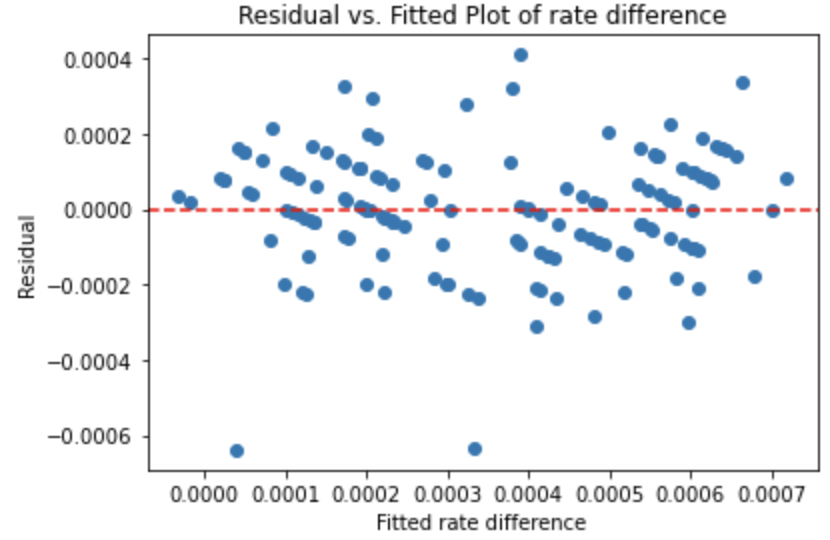
Subsequently, residual analysis was applied as shown in *Figure 3-4*. However, it was observed there were two residuals (red circle) sticking out from the overall pattern of residuals, which could be regarded as outliers.

Table

Description automatically generated

Figure 3‑4 Residual analysis plot

Figure 3‑5 OLS regression results after removing outliers (a)



Then the multiple regression model was rebuilt after removing outliers from the data frame (*Figure 3-5*), the R2 and Adjusted R2 were 0.610 and 0.560 respectively.

Finally, *Figure 3-6* revealed the coefficient and p-value of each variable to decide the most effective intervention. It could be noticed that clean air, health training and school awareness had significant influence on the rate difference, which showed that the rate difference decreasing with the increase of these variables. However, hypothesis testing was applied by assuming the significance level was 0.05, only school awareness's p-value was below the significance level. Thus, school awareness appeared to be the most effective intervention.

Graphical user interface, text, table

Description automatically generated

Figure 3‑6 OLS regression results after removing outliers (b)

The data manipulation refers to Python code link:

<https://github.com/joeylizh/QM_Coursework_1/blob/main/Coursework_1.ipynb>

# Discussion

Firstly, correlation matrix could not be used for dealing with multicollinearity, it could only detect whether the variables were highly correlated. Nevertheless, the R2 did not boost a lot after eliminating, and even reduced after removing outliers, which might indicate there were other variables was not considered, and might also related to coding technique issues.

Subsequently, it could be observed from the residual plot (*Figure 3-4*) the residuals were stochastically distribute around 0-residual line, which illustrated the linear relationship between the ratio difference and the variables, and showed the residuals were independent. Furthermore, since plenty of points were distributed close to the 0 line and the density of points reduced, the regression model was considered as normally distributed. And when the fitted value changing, horizontal band could be roughly drawn around the 0 line for most of the residuals which demonstrated equal variance.

From the regression results (*Figure 3-5*), R2 interpreted 61% of the total variance of the ratio difference and the results indicated that removing the outliers did not improve the model, while the R2 variation was within an acceptable range.

# Conclusion

To conclude, the research firstly introduced the risk of childhood obesity and reducing the childhood obesity

Future research topic may focus on adding more variables to improve the model and investigate other factors that can mitigate the childhood obesity rate difference.

(Word Count: )

**Reference**

*Children: new threats to health*. (2020). Available at: https://www.who.int/news-room/fact-sheets/detail/children-new-threats-to-health (Accessed: 12 November 2021).

Simmonds, M., Llewellyn, A., Owen, C. G. and Woolacott, N. (2016). ‘Predicting adult obesity from childhood obesity: a systematic review and meta-analysis’. *Obesity Reviews*. John Wiley & Sons, Ltd, 17 (2), pp. 95–107. doi: 10.1111/obr.12334.