## Research Question

What factors explain the prevalence of childhood obesity in local authorities in 2018?

## Data

The data includes total population and the total number of cases of childhood obesity by local authority areas in 3 time periods – 2008, 2013 and 2018. To account for different population sizes of local authorities, the number of cases was normalised to 100,000 people. *Table 1* shows descriptive statistics relating to the total number of cases per 100,000 over the years. The summary shows that on average, the prevalence of childhood obesity has increased over the years. It also suggests the presence of outliers, as the observed maximum number of cases lies outside the upper Tukey fence.

*Table 1: Descriptive statistics of number of cases per 100,000*

|  |  |  |  |
| --- | --- | --- | --- |
|  | 2008 | 2013 | 2018 |
| min | 74.00 | 41.93 | 101.39 |
| 25th percentile | 214.84 | 226.89 | 253.54 |
| median | 291.45 | 298.48 | 318.21 |
| mean | 309.73 | 320.98 | 344.67 |
| 75th percentile | 365.03 | 381.07 | 406.70 |
| max | 824.13 | 849.61 | 872.88 |
| IQR | 150.19 | 154.81 | 153.16 |
| standard deviation | 136.61 | 138.64 | 138.26 |

The data also includes the total budget allocated for tackling the condition and its distribution across six policy areas. All of these were recalculated per person, to account for different sizes of local authorities. We can observe from *figure 1a* that the data contains a spending outlier, the City of London, which is not a typical local authority due to its low population and was removed from the data in the subsequent analysis. Apart from that, total budget per person does not seem to be related to the number of cases per 100,000 in 2008 (*figure 1b*).

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated*Figure 1: Number of cases per 100,000 in 2008 vs total budget per person*

|  |  |
| --- | --- |
| *a)* | *b)* |

## Methodology

To explore the factors which best explain the prevalence of childhood obesity in 2018, a multiple regression was fitted in the form

, where *cases\_total\_prop\_2018* denotes the number of cases per 100,000 people in 2018, *cases\_total\_prop\_*2008 the same in 2008, *female\_prop\_change* indicates the change in the proportion of females between 2008 and 2018 and variables ending in *\_pp* stand for budgets per person allocated to the various interventions. Total budget per person was not included because it would result in problems with multicollinearity, as it is completely determined by the sum of budgets allocated to the different interventions.

Including the prevalence of childhood obesity as one of the predictors makes it an instance of an autoregressive model. The reason for including it is that by looking at *figure 2*, we can see that the number of cases in 2018 is strongly and linearly related with the number of cases in 2008, which suggests that the number of cases in 2008 is a significant explanatory factor and appropriate for use in a linear regression model.

*Figure 2: Number of cases per 100,000 in 2008 vs 2018*

*Chart, scatter chart

Description automatically generated*

Such model specification also helps us better assess the effectiveness of the various interventions by reducing the potential for omitted variable bias. Omitted variable bias would result if there were some unobserved characteristics of the local authorities, which influence both the dependent variable (number of cases) and the independent variable (the budget allocated to a particular intervention). An example could be that authorities that put more value on their education system would have lower prevalence of childhood obesity, as well as higher spending on school awareness, creating the appearance of effectiveness of higher school awareness spending. By controlling for the number of cases in 2008, we simultaneously control for time-invariant characteristics of the authorities that produced it, such as the value put on education.

## Results

*Table 2: Summary of the fitted model*

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Variable: | cases\_total\_prop\_2018 | R-squared: | 0.983 |
| Model: | OLS | **Adj. R-squared:** | 0.982 |
|  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | t | P>|t| | [0.025 | 0.975] |
| const | 61.6162 | 15.465 | 3.984 | 0.000 | 31.046 | 92.187 |
| cases\_total\_prop\_2008 | 0.9945 | 0.011 | 88.022 | 0.000 | 0.972 | 1.017 |
| female\_prop\_change | 480.2027 | 55.134 | 8.710 | 0.000 | 371.213 | 589.192 |
| total\_budget\_pp | -21.4954 | 16.824 | -1.278 | 0.203 | -54.753 | 11.762 |
| clean\_air\_pp | -26.6858 | 20.755 | -1.286 | 0.201 | -67.715 | 14.343 |
| clean\_environ\_pp | -21.4054 | 19.334 | -1.107 | 0.270 | -59.624 | 16.813 |
| health\_training\_pp | -52.2263 | 39.526 | -1.321 | 0.189 | -130.361 | 25.909 |
| school\_awareness\_pp | -72.0710 | 17.605 | -4.094 | 0.000 | -106.873 | -37.269 |
| media\_awareness\_pp | 100.1211 | 32.882 | 3.045 | 0.003 | 35.120 | 165.123 |
| sub\_counselling\_pp | 50.7720 | 51.315 | 0.989 | 0.324 | -50.668 | 152.212 |

*Figure 3: Residuals vs fitted plot*

*Chart, scatter chart

Description automatically generated*

## Interpretation and Discussion

The summary of the fitted model (*table 2*) shows very high R-squared and adjusted R-squared of 0.983 and 0.982, respectively. Such result is not unexpected considering *figure 2*, which suggests that the number of cases per 100,000 in 2008 is highly correlated with the number of cases in 2018. This is also reflected in the p-value of the coefficient for *cases\_total\_prop\_2008* which is close to zero, meaning that the variable is a significant predictor of the number of cases in 2018 and we can reject the null hypothesis that its gradient is equal to zero. The estimated coefficient is 0.99, which means that controlling for the other independent variables (keeping them equal to zero), having one case of childhood obesity per 100,000 in 2008 is associated with having 0.99 cases in 2018. This nevertheless cannot be interpreted as saying that the prevalence of cases has decreased over the years because the other independent variables were not zero in practice.

The change in the proportion of females is also significant with a p-value close to zero, suggesting a significant difference in the prevalence of obesity between males and females. On average, a 1% increase in authority’s proportion of females is associated with 4.8 more cases per 100,000 inhabitants.

Considering the effect of the interventions, most of them are not significantly associated with a change in the number of cases and we cannot reject the null hypothesis that their coefficients are equal to zero. Exceptions are school awareness and media awareness, with increased funding for school awareness associated with number of cases lower by 0.72 per 100,000 with each extra 0.01 pounds per person. In contrast, extra 0.01 pounds per person spent on media awareness was associated with the number of cases per 100,000 higher by 1. The results imply that school awareness has been the most effective and media awareness the least effective of the interventions.

Inspecting the residuals vs fitted values plot in *figure 3* to assess the validity of linear regression assumptions, it seems that the errors are independent (there is no observable pattern), normally distributed around zero and have equal variance for all x values. This corroborates the validity of using a multiple linear regression model for understanding the data.

## Conclusion