Prediction for the Obesity Prevalence of Children in London

1. Introduction (Literature Review)

1.1. Background

Childhood weight status is an important public health indicator, as it provides insight into the nutritional well-being of young populations and highlights potential healthcare challenges. In England, for instance, the National Child Measurement Programme (NCMP) collects children's height and weight data (calculate BMI) in Reception (aged 4–5) and Year 6 (aged 10–11) to determine the prevalence of underweight, healthy weight, overweight, and obesity. NHS Digital (2022) claimed that, on the base of recent NCMP data, around 10.1% of Reception-age children and 23.4% of Year 6 children were classified as obese in the 2021–22 school year.

1.2. Significance of Predicting Childhood Obesity Prevalence

Early identification of obesity trends is crucial for enabling targeted interventions, such as nutritional education, physical activity programs, and family-based support, to prevent weight-related health issues from becoming deeply established (Public Health England, 2022). Robust forecasts of obesity rates also support governments in planning for healthcare services, school lunch programs, and neighbourhood initiatives promoting healthy living (WHO, 2022). Childhood obesity is strongly associated with serious medical conditions later in life, including type II diabetes, cardiovascular disease, musculoskeletal disorders, and mental health concerns, highlighting the importance of early intervention to prevent or mitigate these outcomes (World Health Organization, 2016; Viner and Cole, 2005). Additionally, tracking overweight and obesity rates among children helps inform and adjust national school food and fitness policies to improve population health outcomes (HM Government, 2016). Although some prediction models already exist, their accuracy and coverage of socio-economic factors are somewhat low.

2. Research Question

Is that possible to predict children prevalence of obesity using socio-economic variables in London?

3. Data

Although conduct analysis on ward scale may be more appropriate than on borough, but the time scale of the data collected among wards are different from each other. Thus, I used the dataset among boroughs. Due to missing data, 'City of London' is considered as an outlier.

Variable	Type	Description
Prevalence (obesity) (%)	Numeric	The prevalence of obesity for two child group (reception, year6) of LAs. Used as dependent variables in regression.
Number (**N**)	Numeric	The number of children measured.
Count of outlets (**C**)	Numeric	The total number of fast-food outlets in each borough.
Rate per 100,000 population ($\boldsymbol{R_1}$)	Numeric	Fast-food outlet density, calculated by the formula: $\boldsymbol{R_1 = 10^5 \times \frac{C}{P}}$.
Earnings per hour (£)	Numeric	Average earning for each people per hour in each borough.
Percentage of people worked with NVQ4+ (%)		The percentage of people who has a level NVQ4 or a higher qualification among people age 16-64.
PTAI value	Numeric	An index to represent the public transport accessibility of each borough.
Population (**P**)	Numeric	The total population of each borough.
Population per hectare	Numeric	The population density using unit of hectare for

		each borough.
Ratio of children per 100,000 population ($\boldsymbol{R_2}$)	Numeric	Children distribution density, calculated by the formula: $\boldsymbol{R_2 = 10^5 \times \frac{N}{P}}$.
IMD - Average rank	Numeric	Rank of average IMD score.
IMD – Average score	Numeric	Poverty index of the region.

Table 1 Key variables

4. Methodology

Variance Inflation Factor analysis (VIF):

Variance Inflation Factor (VIF) quantifies how much the variance of a regression coefficient is inflated due to multicollinearity among the independent variables. High VIF values (threshold=5 in the analysis) indicate problematic correlation among predictors. Used to drop variables for **multilinear** model.

Simple linear regression and multilinear regression:

- **Simple Linear Regression:** Models the relationship between one independent variable (x) and one dependent variable (y) with a straight line.
- **Multiple Linear Regression:** Extends this idea to multiple independent variables to predict a single continuous dependent variable.

Ridge regression and Lasso regression:

Both Ridge and Lasso are regularization techniques that penalize large coefficients to reduce overfitting and improve model generalization.

- **Ridge Regression** (ℓ_2 regularization): Adds a penalty proportional to the sum of the squared coefficients. It tends to shrink coefficients but usually does not set any to zero.
- **Lasso Regression** (ℓ_1 regularization): Adds a penalty proportional to the absolute values of the coefficients. It can shrink coefficients to zero, effectively performing feature selection.

Polynomial regression (add poly features):

Polynomial Regression extends linear models by adding polynomial terms (e.g., x^2, x^3, etc.) to capture non-linear relationships between the features and the target variable, which could be used in all the linear regression models (**multi, ridge, lasso** etc.)

Support Vector Regressor (SVR):

SVR is a derivative of the Support Vector Machine (SVM). Instead of finding a hyperplane that separates classes, SVR fits a function $f(x)$ that deviates from the target values by at most ϵ for most data points, while also seeking to keep the model as "flat" as possible.

Key Hyperparameters of SVR:

1. **C (Penalty Parameter)**
   - Determine the balance between model flatness and the amount of deviation that can be tolerated greater than ϵ.
   - **C = [0.1, 1, 10, 100, 1000] in the analysis.**
2. **ϵ (Epsilon Parameter)**
   - Defines the width of the ϵ-tube; points within this tube are not penalized.
   - **ϵ = [0.001, 0.01, 0.1] in the analysis.**
3. **Kernel function**
   - The kernel functions allow SVR to operate in high-dimensional feature spaces without explicit transformations.
   - **Kernel = ['linear', 'poly', 'rbf', 'sigmoid'] in the analysis.**
4. **Gamma (γ)**
   - Controls the influence of a single training example.
   - **γ = [0.01, 0.1] in the analysis**

Grid Search method:

Grid Search is a technique to systematically work through multiple combinations of hyperparameters

to determine which set provides the best model performance. It exhaustively tests all specified parameter values. Used to find the best hyperparameters of model **SVR**.

5. Results and Initial Discussion

First, we presented the overall obesity prevalences of the two child groups over the years (2007-2018) through London borough (**Figure 1**).

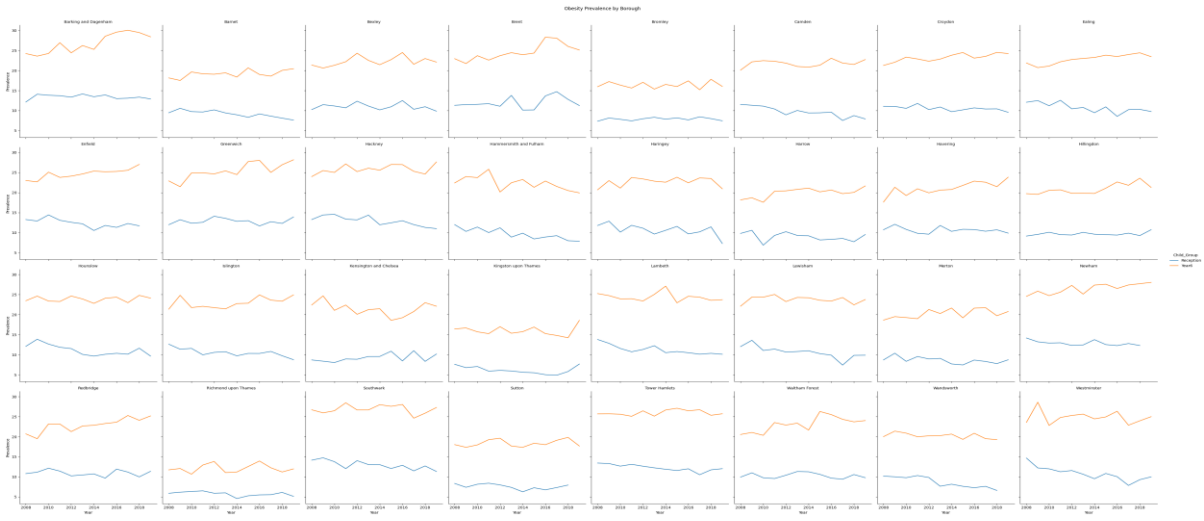

Figure 1 Obesity prevalence by boroughs

We could find that the obesity rate of the rep group is generally lower than that of year6. Correspondingly, the result of paired t-test for each borough also shows the same conclusion. Then, we performed ANOVA-tests on the two child groups. For **Reception** group, ANOVA Statistic: **40.16**, P-Value: **5.53e-96**, for **Year6** group, ANOVA Statistic: **74.05**, P-Value: **4.47e-134**, both reject the null hypothesis, which means there is a significant difference in obesity prevalence among boroughs (**Figure 2**). Thus, Therefore, when we build a regression model, it is better to include all regions.

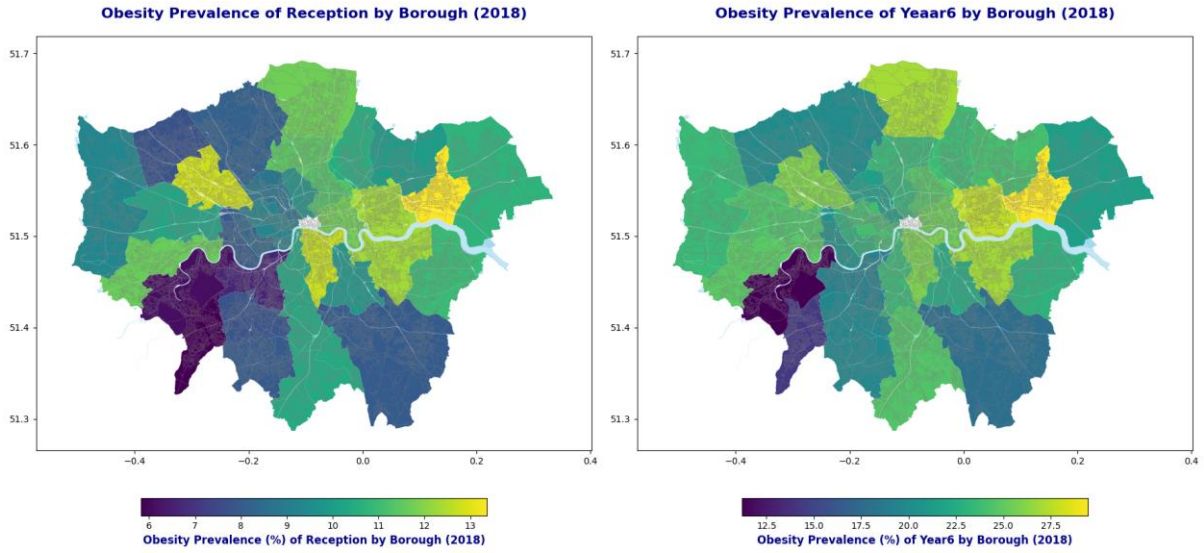

Figure 2 Obesity prevalence by child groups

After this step, the analysis focused on the Reception child group.

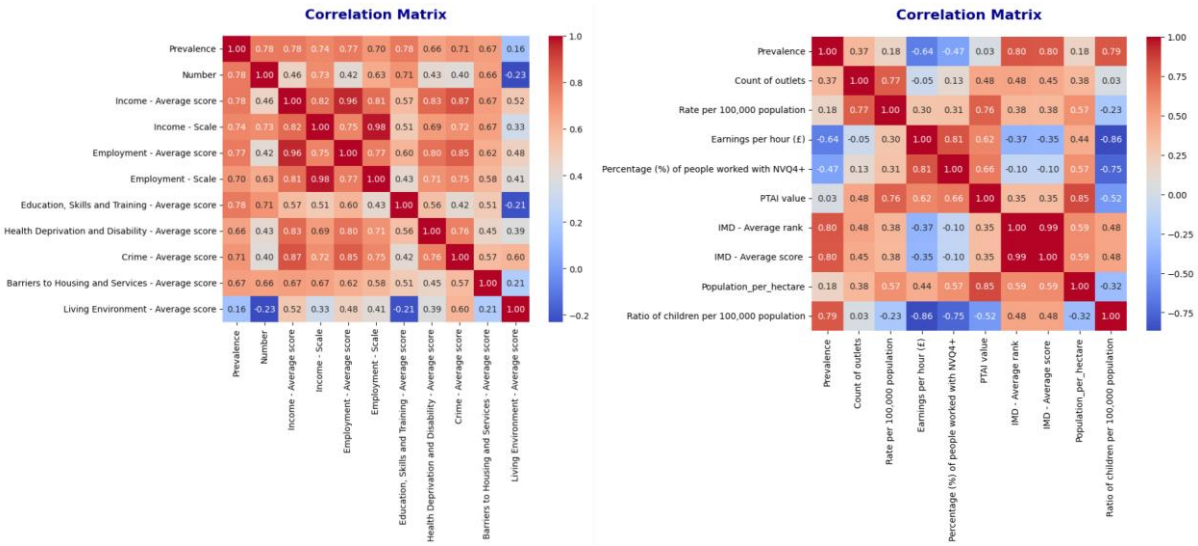

Figure 3 Correlation matrix

On the other hand, NHS England (2017) demonstrated that there is a very strong correlation between deprivation1 and obesity. The **IMD scores**, which is used to measure poverty, were obtained by weighting the variables in the correlation matrix on the left side of the (**Figure 3**). The correlation matrix on the right and side (**Figure 3**) contains all the independent variables I used in this analysis.

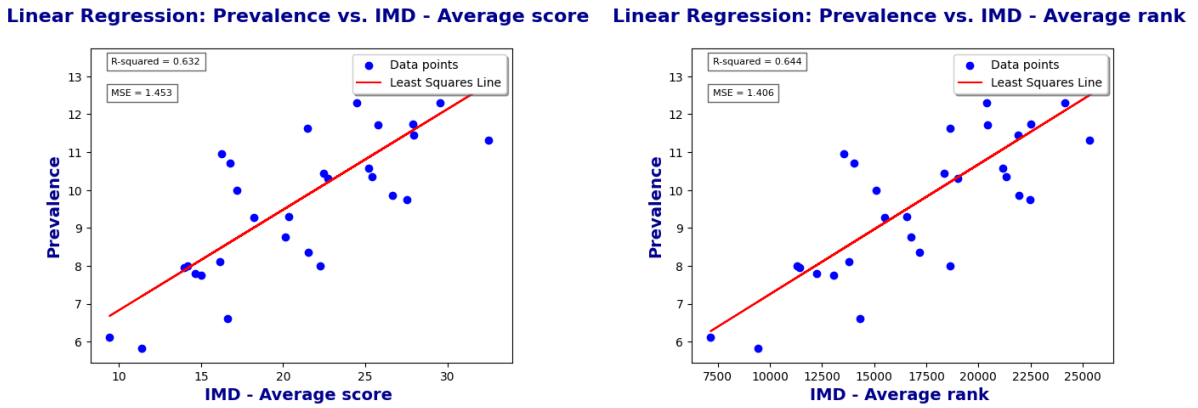

Figure 4 Simple linear regression model for IMD score and rank

In (**Figure 4**), I established two simple linear programming models with **IMD-Average score** and **IMD-Average rank** as independent variables and Prevalence of obesity among child group reception as dependent variable. The **MSE** and **R-squared** values are **0.632, 1.543; 0.644, 1.406.**

Although the two simple linear regression models established previously proved that IMD is related to obesity rate, the fitting results were not very ideal. Therefore, the next step is to predict obesity rates again by adding independent variables and using more complex models.

In the further analysis, the dataset was truncated and partitioned. The data from **2013 to 2017** were used as **training** sets, and the data from **2018** were used as **test** sets. For multiple linear regression models, VIF was used to select independent variables to eliminate multicollinearity. Finally, the three variables **'IMD - Average rank', 'IMD - Average score' and 'PTAI value'** were removed.

When it comes to (**Figure 5**), the results of four linear regression (**Multi, Ridge, Lasso, SVR**) models were aggregated, consisting of a scatter plot of the actual values and the predicted values, and a residual distribution plot.

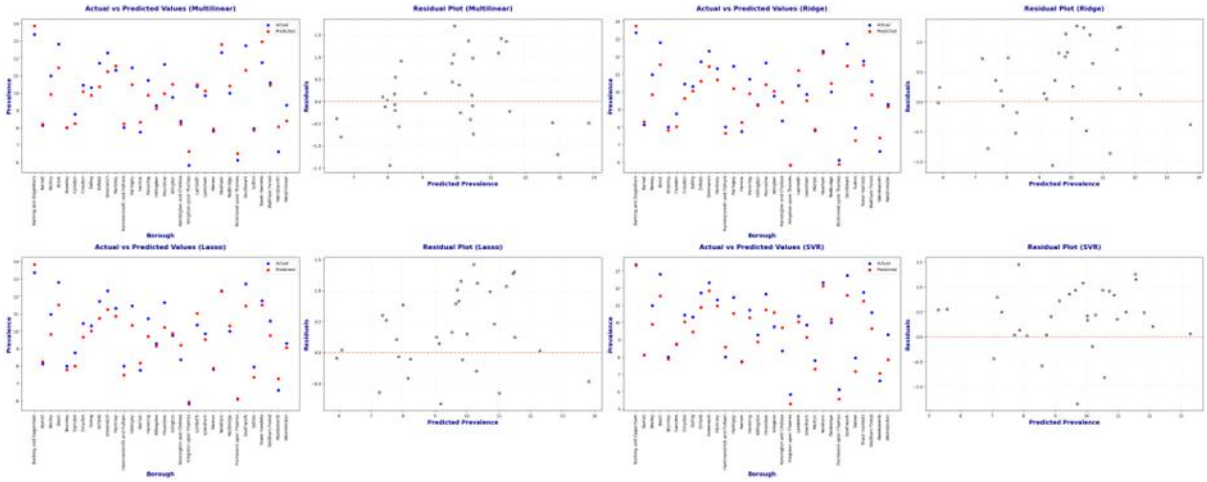

Figure 5 Actual vs Predict scatter and residual plots for each regression model

The graph (**Figure 6**) below shows the coefficients of all the features in different linear regression model.

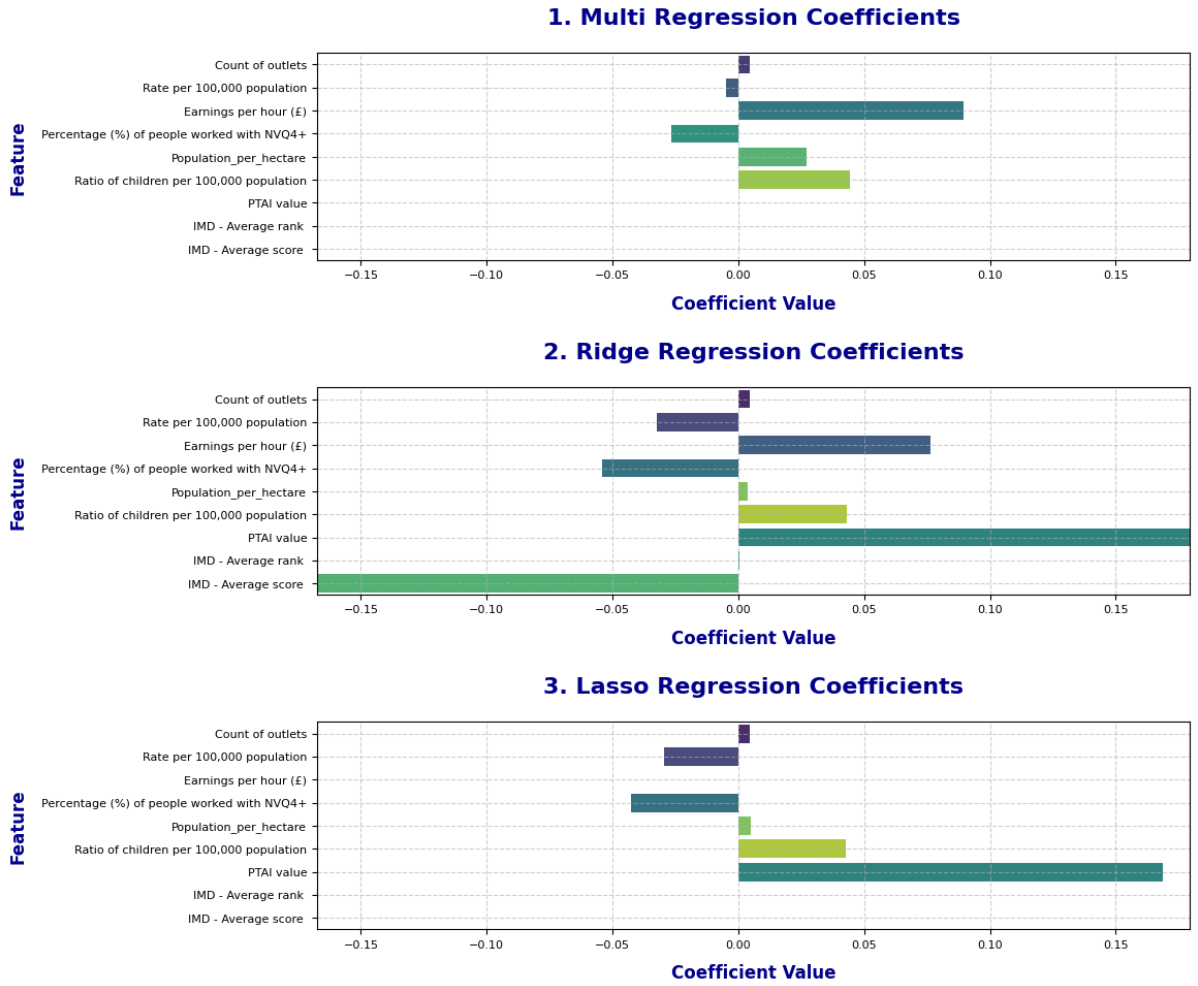

Figure 6 Regression coefficients of each linear regression model

After added polynomial features for each model, changed the hyperparameter degree in the list **[1, 2, 3, 4, 5]**, explored the different performance with different degree, which is displayed in (**Figure 7**). Then, used **Grid Search Method** to find the **best hyperparameters. {'C': 10, 'epsilon': 0.1, 'gamma': 0.1 'kernel': 'rbf'}**

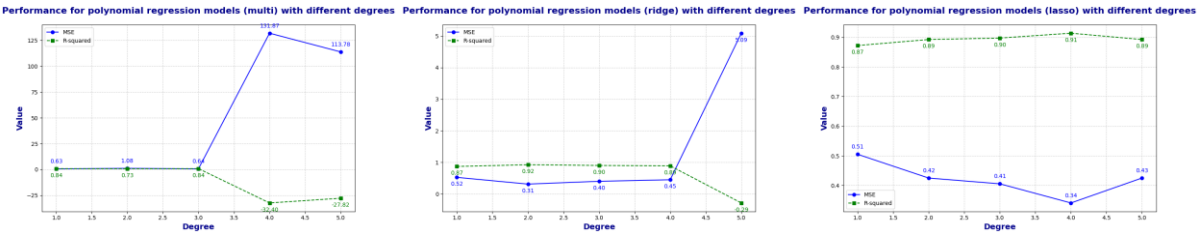

Figure 7 Performance of each polynomial regression model with different degrees

Finally, after all models had been run, the final result graph is obtained (**Figure 8**).

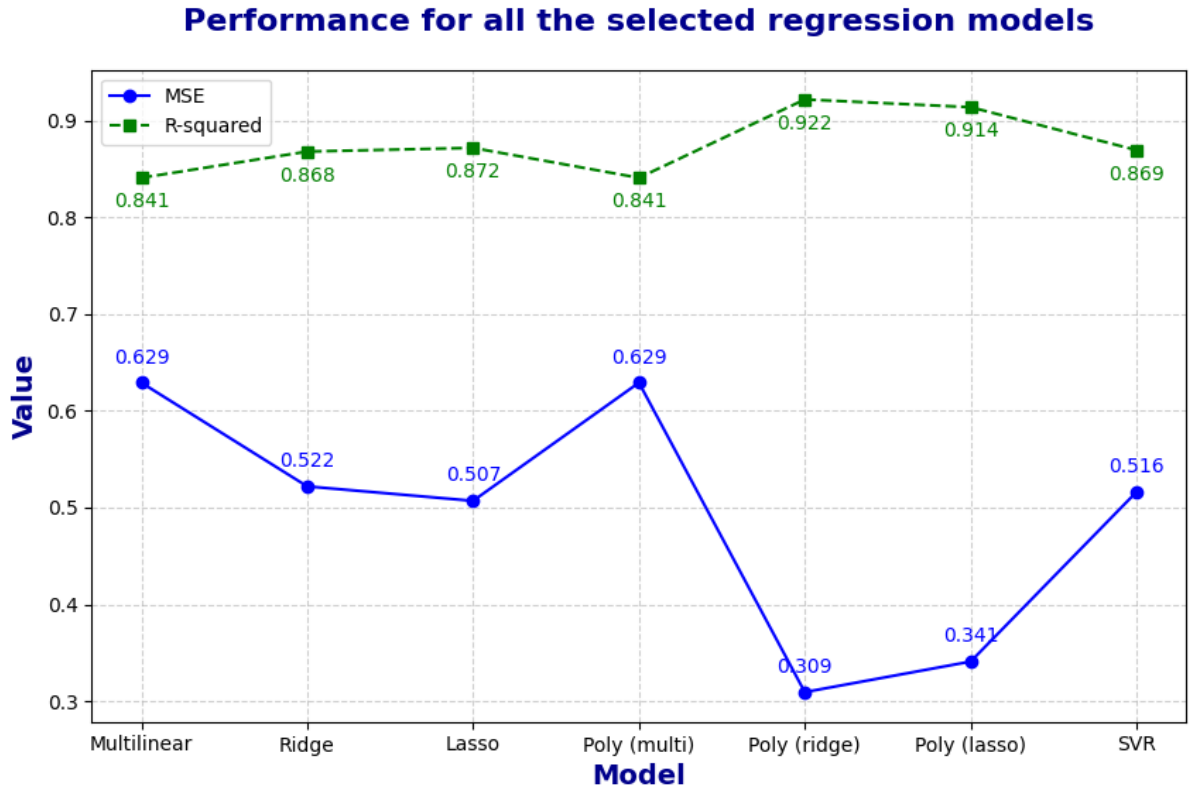

Figure 8 Overall performance of all the regression models

6. Further Discussion

Analysis results show significant spatial and socioeconomic variation in childhood obesity rates across London boroughs. Regression results indicate that poverty (based on **IMD score and rank**) is strongly and **positively** associated with obesity prevalence in school-age children. This finding is consistent with existing research highlighting poverty as a key determinant of health disparities. However, despite the strong relationship, simple linear regression models do not fully capture variation in obesity prevalence, suggesting that more complex models are needed.

Adding multiple predictors, such as **fast-food restaurant density** and **PTAI**, improves prediction accuracy. However, removing **IMD**-related variables due to multicollinearity highlights the complexity of disentangling the impact of socioeconomic factors. Regularization techniques such as **Ridge** and **Lasso** further refine the model by balancing bias and variance, while **SVR** provides robust predictions by capturing nonlinear patterns.

Interestingly, polynomial features slightly improved the models' performance, suggesting that interactions between socio-economic variables may contribute to obesity trends. However, **overfitting** risks increase with higher polynomial degrees, highlighting the importance of model simplicity for

generalization.

One limitation of this study is the exclusion of the **City of London** and reliance on **borough-level** data, which may overlook more subtle spatial variations. Future research could use **ward-level** data or other geographic aggregations to improve predictions.

7. Conclusions

This study confirms that socio-economic factors significantly influence childhood obesity prevalence in London, with deprivation being a dominant predictor. Advanced regression techniques provided robust insights, but challenges like multicollinearity still exist. These findings highlight the need for targeted public health strategies that address both individual behaviors and systemic inequalities. Governments should prioritize resource allocation to poorer regions and consider multifaceted interventions that combine environmental and socioeconomic reforms to reduce childhood obesity.

WORD COUNT: 1740

STUDENT NUMBER: 22222869
GITHUB: https://github.com/wbwhaha/QM_Write_Investigation

References:

Christiansen, T., Richelsen, B. and Bruun, J.M. (2004). Monocyte chemoattractant protein-1 is produced in isolated adipocytes, associated with adiposity and reduced after weight loss in morbid obese subjects. *International Journal of Obesity*, 29(1), pp.146–150. doi:https://doi.org/10.1038/sj.ijo.0802839.

Gov.UK (2016). *Childhood obesity: a plan for action*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/childhood-obesity-a-plan-for-action.

NHS ENGLAND (2022). *National Child Measurement Programme, England, 2021/22 School Year*. [online] NHS Digital. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2021-22-school-year.

NHS England (2017). *National Child Measurement Programme, England - 2017/18 School Year [PAS] - NHS Digital*. [online] NHS Digital. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2017-18-school-year.

Public Health England (2014). *National child measurement programme: operational guidance*. [online] GOV.UK. Available at: https://www.gov.uk/government/publications/national-child-measurement-programme-operational-guidance.

World Health Organization (2016). *Report of the Commission on Ending Childhood Obesity*. [online] www.who.int. Available at: https://www.who.int/publications/i/item/9789241510066.

World Health Organization (2017). Childhood Overweight and Obesity. *World Health Organization*. [online] doi:https://doi.org/entity/dietphysicalactivity/childhood/en/index.html.