1. Introduction

Childhood obesity has become an issue of concern as it has serious physical and psychological consequences on the wellbeing of individuals from youths to adults. Through our background research, we have found that childhood obesity risk is significantly associated with economic factors like household income, homeownership. (Greves, et al., 2010) Neighbourhood violence is also an important social risk factor that has significant stress on the psychological health of children. In multiple articles, evidence shows that neighbourhood violence may influence childhood obesity by discouraging physical activity while encouraging sedentary behaviours. (An, et al., 2017) It is also the result of research that traffic pollution is positively associated with the growth in BMI in children aged 5–11 years. (Jerrett, et al., 2014) These risk factors are taken notice when doing exploratory data analysis (EDA) and building models. Given the data of obesity counts of Year 6 children from 323 Unitary Authorities (*UAs*) across the year 2011 to 2017 and some socioeconomic indicators in England, this report aims to illustrate the probable social, demographical and economic impact that may have on childhood obesity through the analysis of data and construction of statistical models. Apart from understanding the issue, our goal is also to predict Year 6 obesity counts for records that are unknown to us.

2. Data Preprocessing

At the stage of data preprocessing, we have observed that our feature set is of relatively high dimension meaning that there are many continuous covariates or categorical covariates of many levels. For example, the variable *UA* for geographical information has 323 levels. If included in our model, features of such high dimensionality are hard to handle and may cause the issue of overfitting which renders poor prediction power. We therefore decided to reduce the dimension of our feature set. We have taken measures like hierarchical clustering method (HCM) to segregate the 323 *UA*s into a fewer number of groups. Eight levels under the variable *Groups* seem to be a good choice (not too many nor too few). Principal component analysis (PCA) has also been done (details later) on some of the variables based on our background knowledge and EDA.

3. Exploratory Data Analysis (EDA)

During EDA, firstly we look at the summary statistics for all the variables regardless the *Regions* or *Groups* they are from. We observe that the larger the population count, the higher the obesity. They also have large variations (maximum value is 100 times more than the minimum value). This makes sense as the number of obese children simply cannot be greater than the total number of children. Therefore, we use obesity rate (ratio of obesity count to population count) as our response for EDA to account for this acknowledged effect. Violent and sexual offences have very right-skewed distribution meaning that most observations (below the third quantile) are small but large values grow fast as they get larger, so there is a very big difference between the third quantile and the maximum. The same situation applies to home affordability and weekly earnings. We suspect some

exponential growth here for the four variables which may need future processing when building models.

Then, comparing boxplots of obesity rate against two geographical variables: *Groups* and *Regions*, we observe *Groups* has fewer levels (5 out of 8) with extreme observations (outside whiskers) than *Regions* (7 out of 9) does. This hints us that *Groups* will be better at explaining the distribution of obesity than *Regions*. This is cross-validated from linear models using the two as the only variable. *Groups* can explain about 51% of variation while *Regions* can only do 31%. We then plot obesity rate against each numeric variable across eight groups. *UA*s from different groups are inherently different from one another (*HCM* does that) so we suspect variables would affect obesity differently in different groups. An inspection of the conditional plots suggests that pupil absence, hospital admission, population under 18, home affordability and weekly earnings (*Absence*, *Hospital*, *Pop18*, *Afford*, and *Earnings*) may affect obesity differently in different groups. In Figure1 for example, *Afford* and *Earnings* have obviously different slopes for Group 6 and 7. These give us some intuition when considering interaction terms in our model. The rest of the variables are again plotted with obesity rate as a whole. Violence and sexual offences (*Violence* and *Sexual*) show substantial non-linearity with obesity rate. Economic inactivity and fuel poverty (*Econ-inactive* and *Fuel*) show significant linearity with obesity rate. Air pollution and population above 65 (*Pollution* and *Pop65*) show some linear trend. Excess winter death (*Death*) is completely random scattered while the gender-pay-gap (*Gender*) shows some randomness. From these scatter plots, *Death* seems too irrelevant to be included in our model. *Gender* was given a try but subsequently removed due to its high p-value and what has been observed in the plot.

We go on to plot potential covariates with one another. We are not surprised to observe extreme collinearity between *Violence* and *Sexual* (correlation coefficient of 0.89) and high correlation between *Afford* and *Earnings* in Figure2 as each pair respectively describes similar aspect about a region. It is also interesting to note that *Pop18*, *Pop65* and *Pollution* show some relationship with one another as shown in Figure3. It may seem counter-intuitive at first but from our background knowledge, *Pop18* and *Pop65* are expected to relate to the working force thus hustle-and-bustle of a region which serves as the driving force for pollution. Based on our EDA and context, we would segregate the 13 numeric variables (excluding population counts) into economic factor (*Fuel*, *Econ-inactive*, *Afford*, *Gender* and *Earnings*), neighbourhood crime (*Violence* and *Sexual*), pollution factor (*Pollution*, *Pop18* and *Pop65*) and others (*Absence*, *Hospital* and *Death*). Performing PCA on correlated variables under the same risk factor category seems logical and serves as a way to reduce the dimension of our feature set. *Violence* and *Sexual* are summarized as *CrimePC* which explains 95% of their variation. *Afford* and *Earnings* are summarized as *AffordPC* which explains 84% of the variation. *Pop18*, *Pop65* and *Pollution* are summarized as *PollutePC1* and *PollutePC2* which cumulatively explain 89% of the variation. In a nutshell, *Absence*, *CrimePC*, *Hospital*, *Fuel*, *PollutePC1*, *PollutePC2*, *Econ-Inactive* and *AffordPC* are the first set of numeric candidate variables to be fitted in models. *Groups* and *Year* are treated as categorical variables in our model.

4. Model Building

At the model building phase, we start by building Ordinary Linear Models (OLMs) because from EDA, we can assume many covariates satisfy the assumption of linearity with our response, obesity counts. Although assumptions of normality and homoscedasticity seem questionable as our response is count in nature, discrete and only takes non-negative values, we will fit OLMs first and check if departure from assumptions is significant. Fitting candidate variables and potentially interesting interactions into our OLM, we can explain about 70.9% of the variation. The normal quantile-quantile (Q-Q) plot in Figure4 indeed shows heavily tailed residual distribution which indicates that given the covariates, obesity rate does not follow a normal distribution. To relax the normality assumption, generalized linear models (GLMs) with Poisson distribution and logarithm as our link function seem to be a good choice for count data because the distribution contains only non-negative integers. Also, we use obesity count as our response and include population count as an offset term in the model so we can infer and predict count directly. Using the same covariates and interaction terms as in the OLM, more variation in response can be explained by the GLM (76.1%).

GLM does seem to be a better choice. We go on checking model's residual plots and plot standardized deviance residuals (SDRs) with every covariate. SDRs roughly form a horizontal band against fitted values but are larger than expected (from -5 to +5). This hints us about overdispersion which is revealed by our calculation of the dispersion parameter (DP), 3.16. This is far from the true DP, 1 for Poisson distribution. SDRs show decreasing variance with *CrimePC* and *AffordPC*. This suggests a departure from the assumption of linearity for these two which coincides with our observation from EDA. A logarithm transformation on these two can do a decent job in randomizing the SDRs as shown in Figure5 so we carry on using transformed principal components in our model. After transformation, deviance explained by our GLM is about 76.4% but our DP continues to be high at about 3.10. Since we have already considered interaction terms, overdispersion can only be solved using a Quasi-Poisson distribution which relaxes the assumption about DP and gives more accurate confidence intervals. After doing this, covariates and interaction terms in our GLM are reconsidered based on their statistical significance and by comparing models with ANOVA.

When checking assumptions for our GLM with Quasi-Poisson distribution, we observe in normal Q-Q plot, at the positive end, observations are more positive than the theoretical values. This means the largest obesity counts are larger than expected given a normal distribution. For count data, it is expected to see such patterns. When spotting for potential outliers, observations with ID 504, 565 and 1761 seem suspicious. Looking back at the data points, the first two are both from London and in Group 6. Obesity count of 504 is lower than expected while that of 565 is higher than expected. Given the complex environment in London, it is hard to conclude that the observations are erroneous. Observation 1761 has obesity count that is lower than expected given its high absence rate, high hospital admission and low weekly earnings. We decide not to delete these extreme observations from our dataset to allow some flexibility in our model. When plotting SDRs against every numeric covariate in our model, we observe that all of them show acceptable random scatter and form horizontal bands so there is no sign of departure from the assumption of linearity. For plots of *Year* and *Groups*, average values of SDRs for each year and group lie horizontally around zero. We have also plotted SDRs against those variables that are not

included in our model which show random scatter as well so we have included all relevant covariates in our model.

A final note on the decision process during analysis about whether or not to use the generalized additive model (GAM) for this dataset. Firstly, from EDA, we would consider most of our covariates show monotonic relationships with our response meaning that they either positively or negatively affect obesity. There could be incidences where a certain risk factor affects obesity differently (change signs). However, we believe this happens as a result of factors interacting with one another which has been accounted for in our model by adding the interaction terms. Secondly, if we choose to use GAM, we put ourselves at the risk of poor interpretation on those "smooth" covariates as there are no coefficients to interpret on. We see this as a tradeoff between interpretation and prediction. Last but not least, using GAM on count data seems odder to me than on time series data for example and not to mention the risk of overfitting.

5. Conclusion and Limitation

Together with judgement and analysis discussed above, we arrive at our final model which is a GLM with Quasi-Poisson distribution and a logarithm link function having 2 categorical covariates (*Groups* and *Year*), 7 numeric covariates (*Absence*, *Hospital*, *Fuel*, *PollutePC1*, *Econ-inactive*, *LogCrimePC* and *LogEarnPC*) and 2 interaction terms (*Groups* with LogEarnPC and *PollutePC1* with LogEarnPC). Together, the model is able to explain about 75.8% of the variation in our data which we believe is a decent amount. Table1 summarizes some of the interesting coefficients together with their standard errors and p-values of our final model. We choose to record them in the link scale so the effects are additive meaning a plus sign indicates a positive effect and a minus sign indicates a negative effect. The positive effect of pupil absence, pollution factor and crime factor on obesity are significant as shown by their higher weighted coefficients compared to other factors. Their significant effects on childhood obesity are expected and could be explained. Pupil absence could be an indicator of poor childhood health condition which encourages less outdoor activities thus contributing to obesity. Pollution and crime factors are shared views from previous academic research which has been discussed in introduction. *Hospital*, *Econ-inactive* and *Fuel* are expected to have less weighted positive effect on childhood obesity. The economic well-being of a household indicated by factor *LogEarnPC* would have a significant negative effect on childhood obesity shown by its high weighted coefficient. The better-off a family is, the less likely they would have obese children. This makes sense if we link to the less nutritious but cheap junk food that financially difficult families may feel tempted to feed their children. In the time domain, we may expect a little dip in obesity count in the year 2012 but after that, we would expect obesity count to increase compared to 2011 with a peak in 2015. Geographically, Group 7 is expected to have extremely high obesity count compared to the rest of the groups. Looking back at summaries of Group 7, we realize that all 56 observations are from London so given the high pollution, high crime rate and great financial stress there, we are expected to see such high obesity count. As we look at interactions between *LogEarnPC* and *PollutePC1*, the coefficient is positive while the coefficient for *LogEarnPC* is negative. This interestingly suggests that at a place where air pollution is significantly high, household earnings may positively affect childhood obesity. Group 6 with high air pollution could be seen as an illustration of this dominated effect of

air pollution on obesity whereby the coefficient of interaction between Group 6 and *LogEarnPC* is positive.

As seen above, our model has successfully shed light upon the socio-economic factors that may potentially affect childhood obesity in England. However, there are some limitations to it. There are only seven observations in Group 8. This inevitably gives large p-values and wide confidence intervals as there are too few data points for solid inference. This could cause poor prediction for observations found in this group. We have naturally assumed a log link function for the data in our GLM but it is not certain that the effects on obesity count is multiplicative in reality. In the end, after everything we have tried, we are still worried that our model overfits the training data and gives poor prediction performance.

6. References

An R, Yang Y, Hoschke A, Xue H, Wang Y. Influence of neighbourhood safety on childhood obesity: a systematic review and meta-analysis of longitudinal studies. Obes Rev. 2017;18(11):1289–1309. doi:10.1111/obr.12585

Greves Grow HM, Cook AJ, Arterburn DE, Saelens BE, Drewnowski A, Lozano P. Child obesity associated with social disadvantage of children's neighborhoods. Soc Sci Med. 2010;71(3):584–591. doi:10.1016/j.socscimed.2010.04.018

Jerrett M, McConnell R, Wolch J, et al. Traffic-related air pollution and obesity formation in children: a longitudinal, multilevel analysis. Environ Health. 2014;13:49. Published 2014 Jun 9. doi:10.1186/1476-069X-13-49

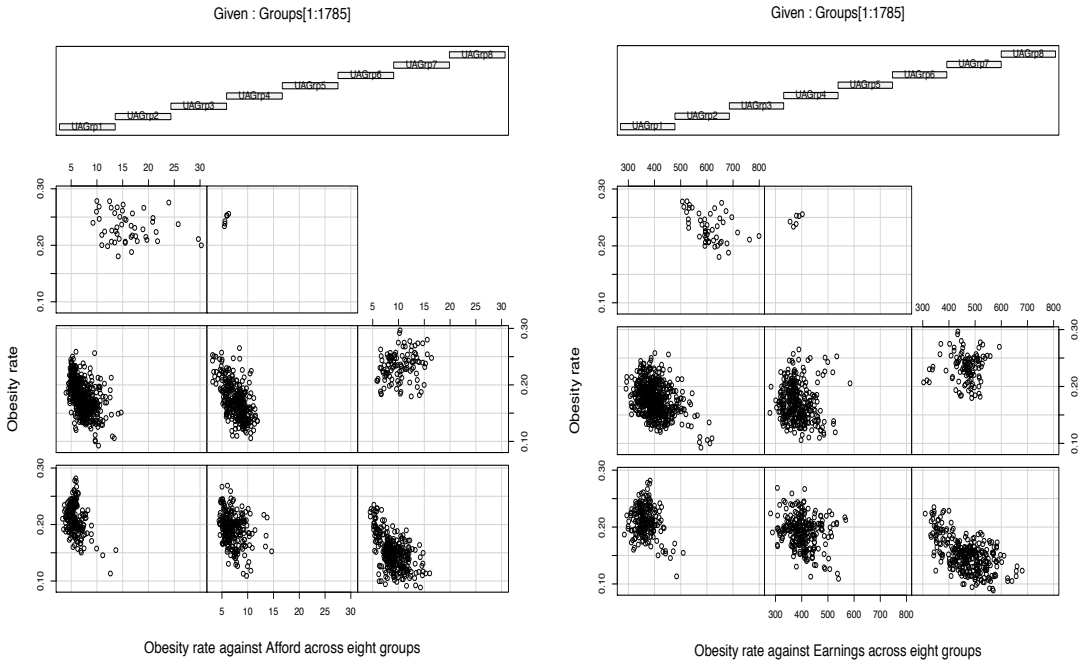

Figure1: conditional plots of obesity rate against Afford and Earnings across eight groups

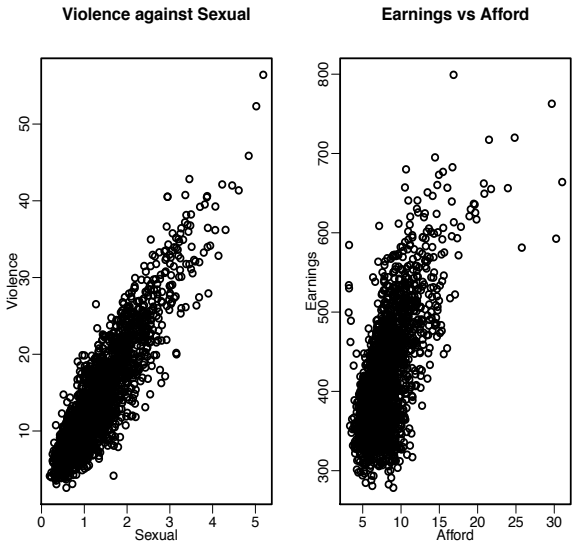

Figure2: plots of Violence against Sexual and Earnings against Afford to show collinearity

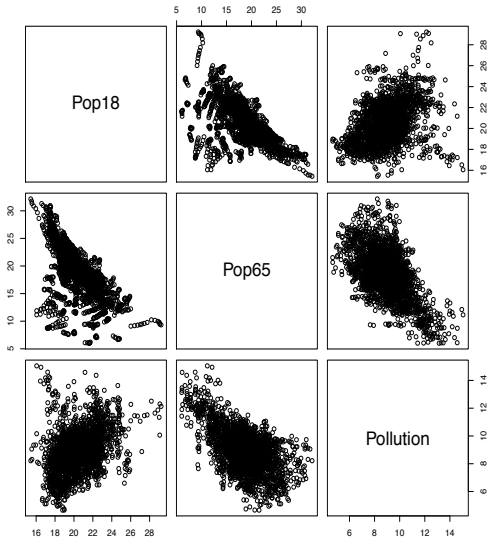

Figure3: plot of Pop18, Pop65 and Pollution to show the relationship amongst them

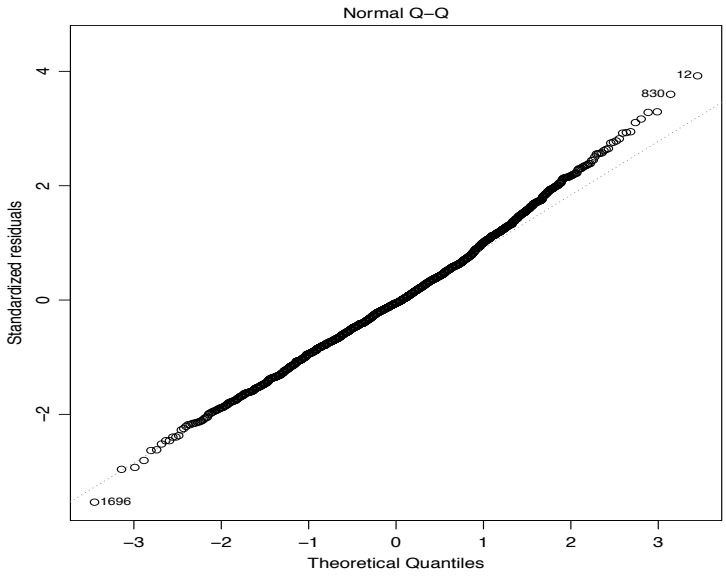

Figure4: QQ-plot of OLMs to check the assumption of normality

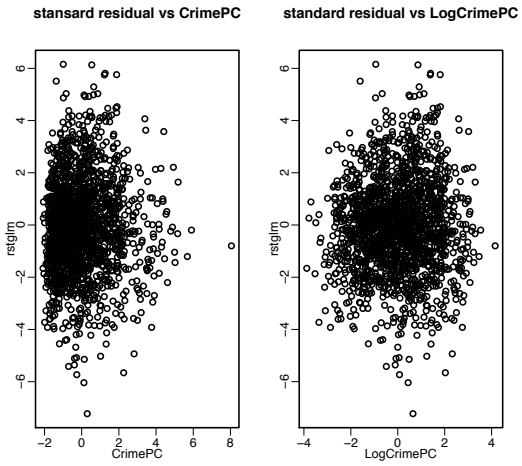

Figure5: plots of SDRs against CrimePC and SDRs against it after a logarithm transformation

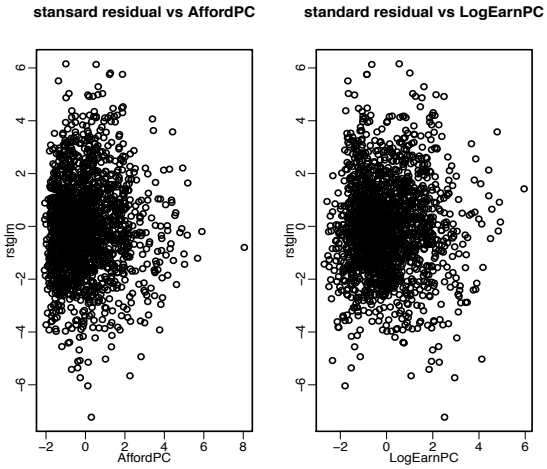

Figure6: plots of SDRs against AffordPC and SDRs against it after a logarithm transformation

Coefficient	Estimate	Std.Error	p-value
intercept	-2.25	0.047	<2e-16
factor(Year)2012	-0.00765	0.00930	0.411
factor(Year)2015	0.102	0.0122	<2e-16
GroupsUAGrp6	0.0527	0.0166	0.00148
GroupsUAGrp7	0.281	0.0517	6.44e(-08)
Absence	0.0615	0.00815	7.48e(-14)
PollutePC1	0.0528	0.00275	<2e-16
LogCrimePC	0.0183	0.00336	6.12e(-08)
LogEarnPC	-0.0584	0.00889	6.83e(-11)
GroupsUAGrp6:LogEarnPC	0.0251	0.0113	0.0270
PollutePC1:LogEarnPC	0.0185	0.00243	3.43e(-14)

Table1: partial summary of our final GLM

Our group, ICA group I, have contributed fairly during this project.