Over the past decades, the prevalence of childhood obesity has rapidly grown. This rapid growth has made it a major public health issue. Childhood obesity is a serious risk factor for many health issues. For example, in the context of the current coronavirus pandemic, it has been shown that obesity may lead to more severe forms of coronavirus. The main issue with childhood obesity is that there is a great risk that it will, in turn, lead to adulthood obesity. The aforementioned complications associated with obesity come to support the need for an efficient preventive approach. Therefore, the aim of our study is to analyse different risk factors linked to childhood obesity in each UA we are given information for. We will in turn propose a model that can efficiently predict obesity counts. Finally, we will conclude this study by suggesting potential preventive measures to reduce childhood obesity.

EXPLORATORY ANALYSIS:

We study a subset of the NCMP data containing 2 232 annual observations. For each observation, we have 15 covariates. The latter enable us to understand the UAs through their populations (pop_under_18/pop_over_65 and pop_count), social indicators (pupil_absence, violent_offences, gender_pay_gap), economic factors (fuel_poverty, inactivity_rate, home_affordability and average_earnings) and environmental aspect (air_pollution). Other covariates like "Falls_in_old_people" or “winter_death” give us general information on the UAs. We also have information regarding the regions (9 in total) in which the UAs are located and the years in which each measurement was taken, thus enabling us to group our UAs and analyse trends over time. As far as covariates go, we noticed strong linear relationships between “violent_offences” and “sexual_offences” as well as between “average_earnings” and “home_affordability” with correlations of 0.89 and 0.68 respectively (**Figures-2-3**).

An analysis of the summary statistics shows that the mean obesity count across all regions has generally been increasing year on year from 2012-2017 (**Figure-1**). Regionwise, London seems to have the biggest issue with childhood obesity (mean of 573.8) while South East showcases the lowest counts (mean of 208.1) (**Figure-1**). More generally, across the range of covariates, London distinguishes itself from all the other regions. This is not surprising given that it is a world capital as opposed to the other regions which are more rural. We also notice the presence of several outliers. Most of them lie in the West Midlands. However, further analysis revealed that the number of obese children in this region had rocketed over the past years. Therefore, our results are of no surprise and the outliers have no reason to be removed.

The objective of our analysis is to predict Year 6 obesity. Therefore, the bigger the percentage of population under 18, the bigger we might expect the count to be, considering there are proportionally more individuals. In other respects, the increase in chronic diseases with aging reduces the ability of people aged 65+ to undertake regular physical activity, thus leading to obesity. It has been shown that households where one of the parents was obese were more prone to childhood obesity.

As regards the social indicators, obesity can lead to psychosocial issues including low self-esteem. Therefore, it seems reasonable to infer that said psychosocial issues may lead

to obese students missing school by fear of being judged. Moreover, given the somatic complications associated with obesity, obese pupils might miss school due to health obligations. In other regards, studies suggest that children exposed to violence can suffer direct biological alterations leading to obesity. If we now consider gender pay gap, it has been shown that, the rise in childhood obesity can partly be attributed to maternal employment. Therefore, it could be said that areas with lower gender pay inequalities will have more mothers in employment. This would theoretically lead to greater childhood obesity and vice versa. Finally, we have no contextual reasons to believe that sexual offences are relevant to obesity.

Let us now consider, the economic factors. The main consequence of fuel poverty is that people are unable to keep their living environment properly heated. Some studies have shown that continued exposure to cold environments can be a somatic stress factor which leads to excessive food consumption in order to compensate the lack of heat with energy (proteins). This would potentially lead to weight gain. Conversely, other studies have shown that cold could act as a form of exercise. Indeed, to preserve constant body temperature, our body compensates for the cold with similar mechanisms as when we exercise. This potentially leads to weight loss. In other regards, high economic inactivity rate implies higher unemployment. This unemployment may lead to depression in a number of cases and subsequently to behaviours that foster obesity. As said previously, children living with obese parents are in turn more likely to become obese themselves. Additionally, one could argue that economic inactivity means that there is less money to spend on healthy food, hence promoting eating habits which lead to obesity. We find that average earnings and affordability of home ownership are relevant for similar reasons.

In respect to the environmental aspect, recent studies have revealed that increased exposure to air pollution during childhood may be a significant risk factor for weight gain and in turn childhood obesity.

Now moving onto the remaining covariates. As far as obesity is concerned, there is no contextual evidence to support the relevance of emergency hospital admissions due to falls in people aged 65 and over. Conversely, it can be argued that excess winter deaths are relevant in a similar fashion as fuel poverty. We can add that it also gives an indication on the overall quality of the healthcare system in the region. The better the healthcare system the less obesity we can expect to see.

MODEL BUILDING:

As a starting point, we decided to create a linear model including all the covariates at our disposal. Our first step was to deal with the “UA” covariate. Given that it had such a large number of categories, we thought it would not be sensible to incorporate it in our models. Hence, we tried to summarise this covariate in a number of ways including hierarchical clustering. We tried different clustering cut-off points which gave us a different number of categories every time. However, the results of this process were never as satisfying as using the “Regions” covariate. Therefore, we ended up grouping UAs according to regions. The

summary of this model confirmed our qualitative analysis. Indeed, covariates that we did not suspect to be linked to obesity presented p-values greater than 0.05 (i.e. falls_in_old_people = 0.23). Some other ones were more contradictory. Indeed, “fuel_poverty” and “pop_under_18” also presented high p-values. F-tests revealed that we preferred the nested model each time we removed one of these covariates. Moreover, given that the contextual evidence to back these covariates relied on early stage studies we decided to remove them. Finally, we added an interaction term between “violent_offences” and “average_earnings”. Indeed, violent offences affect obesity, however it can be said that relatively poorer areas are more prone to such violence. Hence, people living in said areas might develop a form of mental resistance to the effects of violence as they are confronted to it on a regular basis. Conversely, violent offences might have a stronger effect on people who earn more as they’re not “used” to it. We then plotted the results of that model. There was no form of clear pattern in the “residuals vs fitted values” plot or in the “standard residuals vs predicted values” plot. This suggested that the assumption of homoskedasticity is verified. Further investigation through a Breusch-Pagan test revealed the existence of heteroskedasticity, however the plots do not suggest that the issue is significant enough that it would justify acting to resolve it. The main problem arises when we have a look at the normal Q-Q plot (**Figure-4**). Indeed, we can see that the residuals are not normally distributed (heavy tail on the plot) thus violating one of the main assumptions of a linear model. This led us to choose a generalized linear model. Indeed, given that our response variable is a count (obesity count), it seems reasonable to resort to models that rely on a Poisson distribution. Moreover, given the plots in **Figure-6** (2 examples) it is legitimate to use the log link function so as to represent the log of the expected obesity counts as a linear function of the other covariates. Indeed, we observe that plotting the log of the obesity counts against the covariates gives light to a clearer linear relationship than when we just plot the obesity counts against those same covariates

We then decided to plot the results of this model (Poisson GLM with log link function and all the covariates). Looking at the residuals vs fitted values plot, another issue arises. The residuals range from -5 to 5. Given that we are relying on a Poisson distribution, if the model is correct, the residuals should lie between -2 and 2. Moreover, the variance of the Pearson residuals was 3.44 when, if the model was correct, it should have been 1. This is a sign of overdispersion. This could have been due to the fact that there are other factors that influence our response variable. However, we do not have any other covariates at our disposal. Therefore, we tried adding the interaction between “violent_offences” and “average_earnings”. This did not have the desired effect, as the residuals still exceeded the (-2,2) range and the variance of the Pearson residuals only fell to 3.02. Hence, we decided to resort to a Quasi-poisson distribution to account for the overdispersion.

As per the inferences of our exploratory analysis, we started our covariate selection process on the basis of a Quasi-poisson GLM with a log link function. At this stage, we decided to use the “Year” covariate as a factor instead of a continuous covariate. This caused the residual deviance of the GLM model to go from 6071.3 to 5916. We then started our covariate selection process. As for the first linear model we fit, we decided to remove

“falls_in_old_people” and “sexual_offences” because of their lack of contextual relevance which was supported by their high p-values (0.35 & 0.23 respectively). Additionally, in regards to “sexual_offenses” its high correlation with “violent_offences” comforted us in our choice. Given the correlations established earlier, we attempted a PCA analysis with average_earnings, home_affordability, violent_offences and sexual_offences. However, the resulting PCs did not improve our model. Moving on, we removed “fuel_poverty” and “pop_under_18” for the same reasons that led us to remove them from our original linear model. We ran Chi-squared tests when removing each of these covariates one at a time. The results of said tests coincided with our previous findings. A harder choice was to remove “winter_death”. Indeed, it had a decent p-value when we did a Chi-squared test (0.01) which indicated that the model including it was significantly better than the one without it. However, our judgment call was that there was not enough current literature regarding the effect of excess winter deaths on obesity to include it. At this stage, our residual deviance was 5949.8. We then added interactions. The first one was between “violent_offences” and “average_earnings” which we explained before. The second was between “regions” and “air_pollution”. Indeed, it has been shown that air pollution affects obesity. However, it can be argued that some regions are so polluted (e.g. London) that their inhabitants “adapt” to this air quality. It would make sense that someone who has lived their whole life in a city like London would be less sensitive to air pollution than someone who has lived in the countryside and experienced fresh air for their whole life. These two interactions brought the residual deviance down to 5060.6. Moreover, Chi-squared tests on both of these interactions revealed that the model including them fit the data better than the one which did not. At this point, to reinforce our understanding of the relationship between the covariates and the response, we decided to fit a GAM and plot the estimated smooth of “home_affordability”, which we were not sure had a linear relationship with the response variable (**Figure-5**). As you can see, the relationship is fairly linear, hence there is no need for a non-parametric term. However, the bands are very wide for large values mainly due to the fact that we do not have many observations with high values for this covariate. Therefore, we tried a log transformation on this covariate and the results were very satisfying (**Figure-5**).

We then produced diagnostic plots for our model (**Figure-7**). We observe a fairly symmetric distribution around 0 with no apparent patterns in the “residuals vs fitted values” plot. Moreover, when plotting the standardized residuals against our final set of covariates individually, we also observe a symmetric distribution around 0 with no apparent patterns. This implies that there should not be any additional dependence to the covariates that our model does not capture (**Figures 8-9-10**). Hence, we are satisfied with the relative fit of our model.

CONCLUSION:

Hence, the final model we retained is a Quasi-poisson GLM with a log link function. The covariates are “regions”, “Year” (factor), “pupil_absence”, “gender_pay_gap”, “inactivity_rate”, “violent_offences”, “average_earnings”, the log of “home_affordability”, “pop_over_65”, “air_pollution” and interactions between “violent_offences” and “average_earnings” and between “regions” and “air_pollution”. Based on our statistical analysis, we would infer that the most important drivers of childhood obesity are pupil

absence (positive relationship), average earnings (negative relationship) and violent offences (negative relationship). This makes sense contextually and logically. In this regard, one potential measure to tackle childhood obesity could be to implement seminars on the subject of body shaming from a young age.

In regards to limitations, it can be said that we could have considered more informative covariates. Indeed, indicators of how much exercise kids are undertaking or the proportion of time spent by children on their computers might have helped us to have a model with more predictive power. Moreover, we also realised that obesity trends are observed over decades rather than shorter periods of time. Therefore, access to data over a larger range of time might be a sensitive request for an ulterior analysis.

REFERENCES:

Bahreynian, M., Qorbani, M., Khaniabadi, B., Motlagh, M., Safari, O., Asayesh, H. and Kelishadi, R., 2017. *Association Between Obesity And Parental Weight Status In Children And Adolescents*. NCBI. <:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5463282/>.

Cawley, J., 2020. *The Economics Of Childhood Obesity | Health Affairs*. Healthaffairs.org. <https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2009.0721>.

Miller, C. and Rodger, J., 2018. *The Number Of Severely Obese Kids In The West Midlands Is Skyrocketing*. BirminghamLive. <https://www.birminghammail.co.uk/news/midlands-news/number-severely-obese-children-west-15066979?fbclid=IwAR3jp7ojWO9qxX6d99BeuTLDZhNql6mEQMKLNJr7rhO01fhM6Dw4axlgefk>.

Theall, K., Pia Chaparro, M., Denstel, K., Bilfield, A. and Drury, S., 2019. *Childhood Obesity And The Associated Roles Of Neighborhood And Biologic Stress*. ScienceDirect. <https://www.sciencedirect.com/science/article/pii/S2211335519300348#!>.

Véronique, P. and Delbecque, C., 2020. *Le Surpoids Et L'obésité, Facteurs De Risque D'infections Sévères Au Coronavirus*. LExpress.fr. <https://www.lexpress.fr/actualite/societe/sante/le-surpoids-et-l-obesite-facteurs-de-risque-d-infections-severes-au-coronavirus_2123265.html>.

Zhu, P., Zhang, Z., Huang, X., Shi, Y., Khandekar, N., Yang, H., Liang, S., Song, Z. and Lin, S., 2018. *Cold Exposure Promotes Obesity And Impairs Glucose Homeostasis In Mice Subjected To A High-Fat Diet*. NCBI. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6131648/>

Figure1:

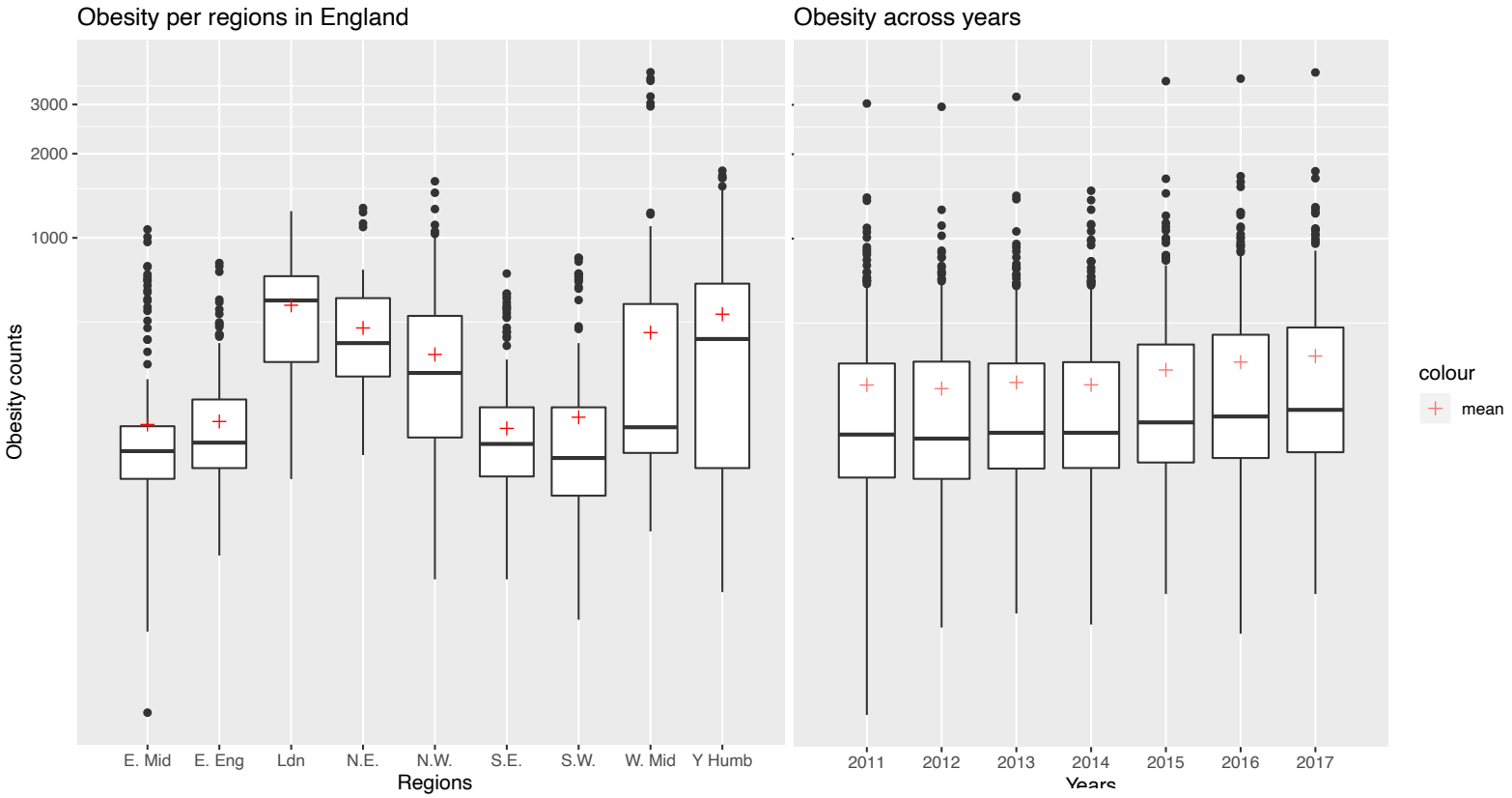


Figure 2:

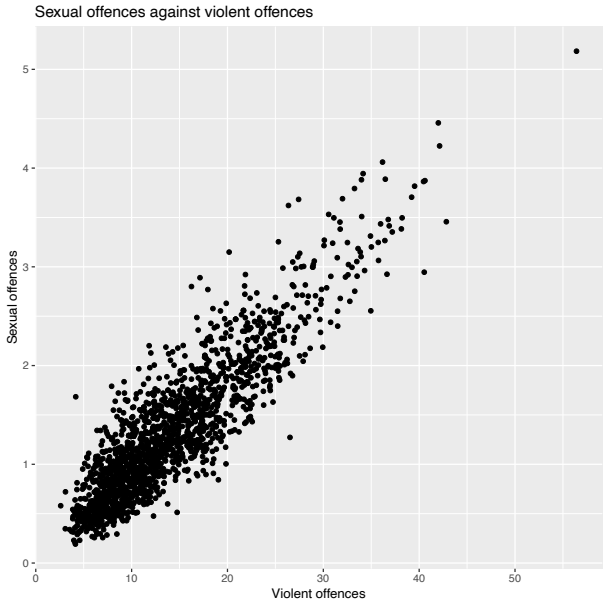


Figure 3:

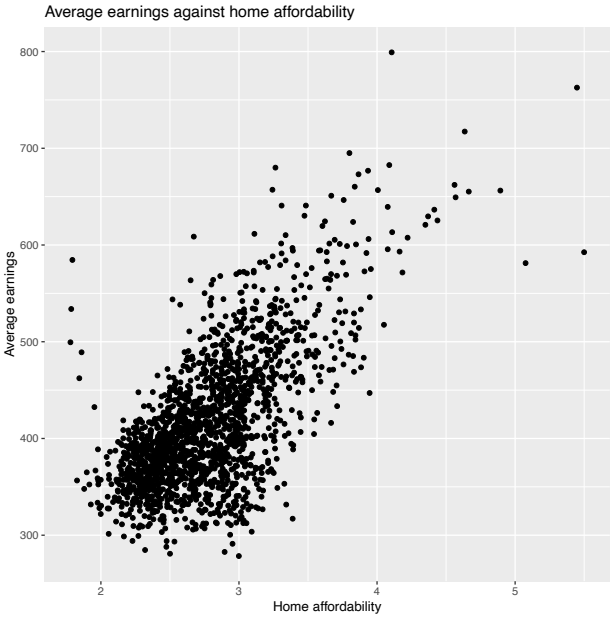


Figure 4:

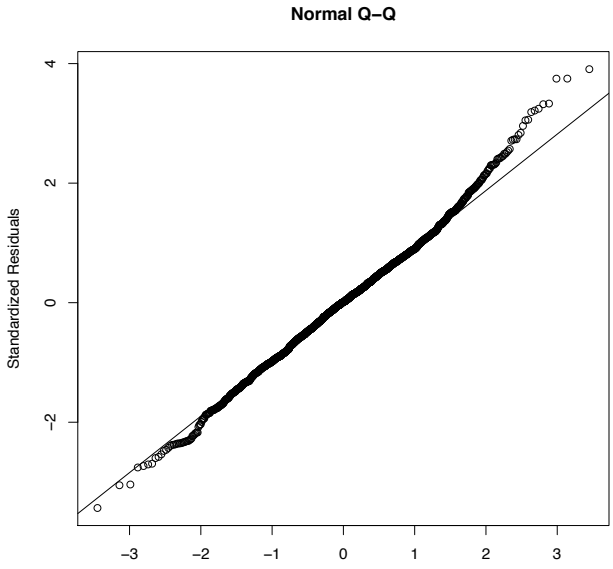


Figure 5:

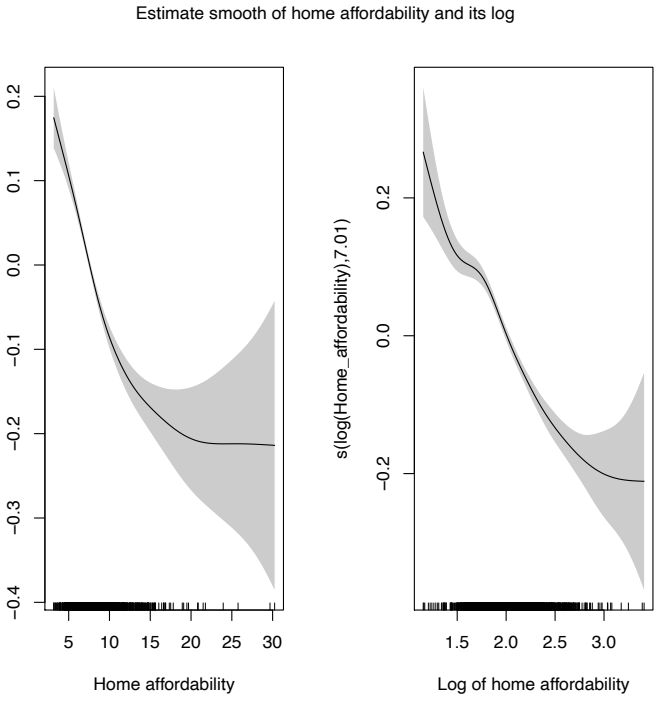


Figure 6:

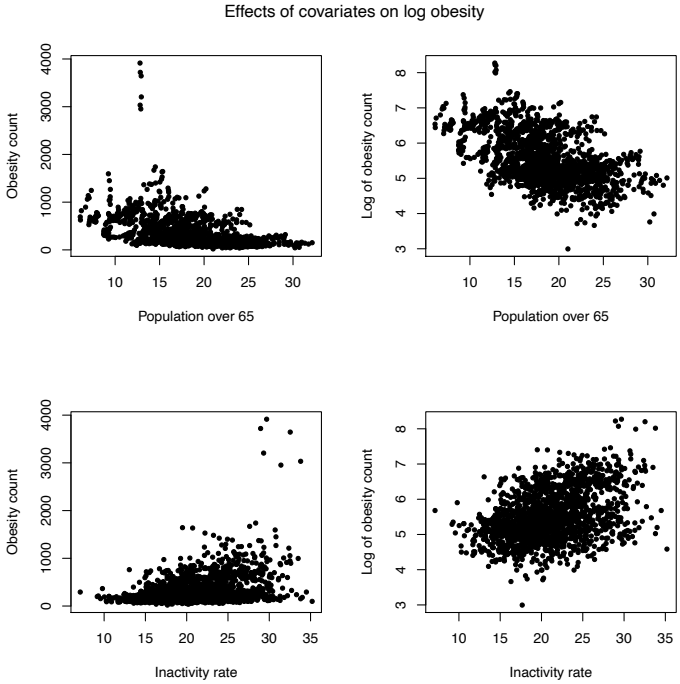


Figure 7:

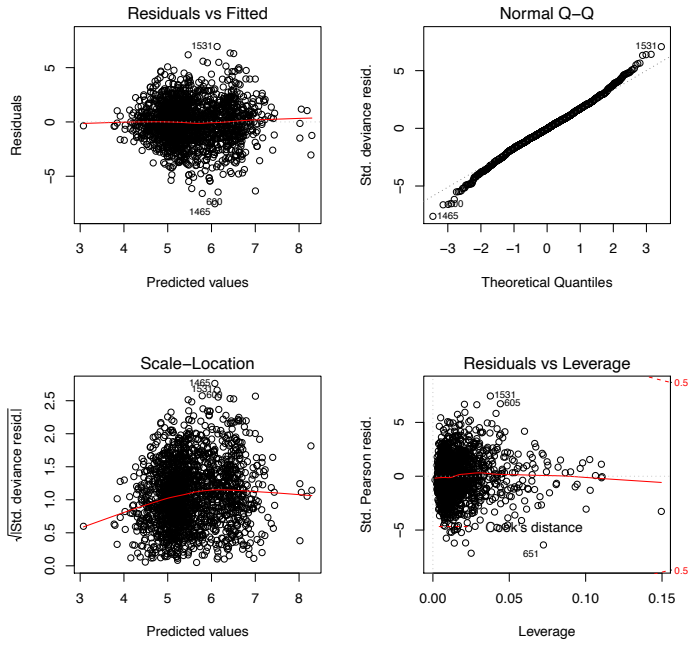


Figure 8:

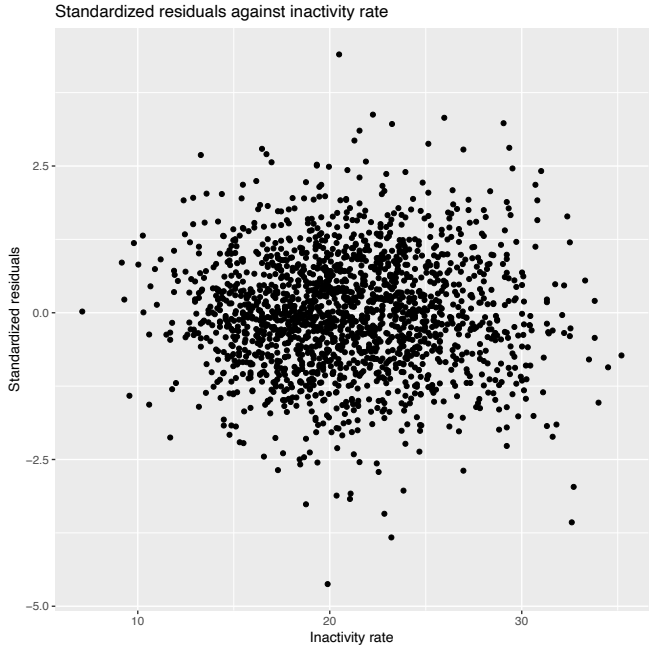


Figure 9:

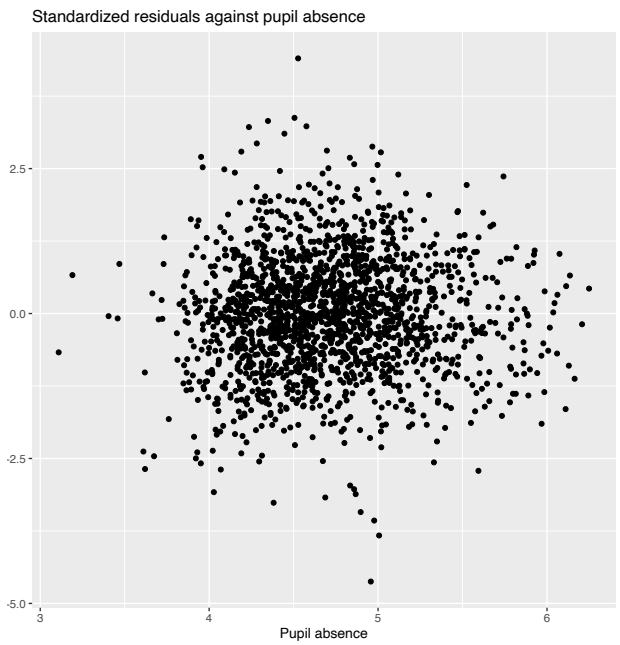
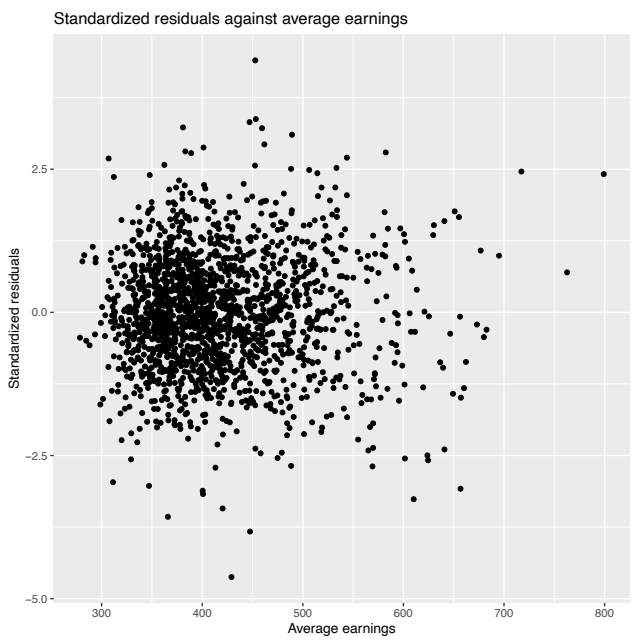


Figure10:



CONTRIBUTIONS:

Axel Tagnon and Ryan Bouguerra contributed equally.