
**EC331 Research in Applied Economics**

# Nutritional health inequalities, income, and environment in London

## Abstract

This paper aims to explain the income-related inequalities in nutritional health in London. To achieve this, it uses high resolution food consumption and nutritional health data for London neighbourhoods. Given the numerous and complex determinants of nutritional choice model selection methods are used to detect the most important predictors of nutritional health and causal relationships are explored via an IV-LASSO algorithm. Results suggest that income and the food environment is an important determinant for the health of London neighbourhoods.

***Key words: London, Health, Nutrition, Inequality***

**Word count: 4813**

Acknowledgements

The author is grateful for Prof. Andreas Markoulakis for his support and supervision, and to Prof. Luca Maria Aiello for providing the data from the Aiello et al. (2019) paper and elaborating on data processing.

# Table of Contents

# 1. Introduction

*"Your body is not a temple, it's an amusement park. Enjoy the ride"*

Anthony Bourdain (2000), *Kitchen Confidential: Adventures in the Culinary Underbelly*

Malnutrition has become a more urgent issue with the breakout of the COVID-19 pandemic. As a reaction, UK prime minister Boris Johnson stated, 'losing weight is, frankly, one of the ways that you can reduce your own risks from COVID' (Campbell and Siddique, 2020). Obesity-related costs by the NHS were as high as £6.1 billion 2014-2015, now with the added risk presented by the coronavirus tackling obesity is an urgent but complex problem (Public Health England, 2017).

In London, income-related inequalities are apparent. The understanding of these inequalities and their determinants could inform policies aiming to improve nutritional health and add to the understanding of the costs of poverty present even in London.

Given the numerous factors that could influence nutritional health, this paper aims to utilise penalised regression methods to select relevant factors and gain an understanding to the determinants of the nutritional inequalities in London. The analysis presented involves three steps (1) factors important to nutritional health are identified, (2) post-selection inference is conducted with OLS on selected variables and compared to a literature-based model, and (3) Instrumental variable estimation with LASSO selection is used to identify the income and food choice effects. This allows for the identification of causal determinants of nutritional health in London.

Additionally, the research provides an assessment of numerous datasets such as the TESCO 1.0 by Aiello et al. (2020) and research methods like the use of Google API for data collection. Furthermore, is aims to add to the literature by showcasing the importance of model selection methods when assessing complex systems such as nutritional health.

Data management and analysis were conducted in R (R Core Team, 2021), with the help of additional packages for a comprehensive list see (Appendix F.).

# 2. Literature Review

Despite societal and economic progress, malnutrition persisted globally (WHO, 2020). Illnesses linked to malnutrition and obesity such as type-2 diabetes, cancer and COVID-19 are major causes of mortality and sickness in both developing and developed countries (WHO 2000; Department of Health and Social Care, 2020). Due to the increased risk and occurrence of illnesses, obesity-related costs surpassed £5.8 billion in 2007, far greater than costs associated with smoking or alcohol use (Scarborough et al., 2011). The problem, however, is not evenly distributed. Income and poverty were found to be associated with obesity in many contexts, such as 11 OECD countries (Devaux and Sassi, 2013). The growing obesity gap by income was also observed in a longitudinal study from Germany, and several publications found consistent results in the UK (Hoebel et al., 2019; El-Sayed et al., 2012).

## 2.1 Determinants of nutritional choice

Several factors determine the nutritional choices of people and consequently nutritional health. Tastes and habits developed early in life are influenced by culture and biology (Anzman et al. 2010). Biological factors play a role in determining the physical pleasure and individual experiences from eating and behavioural issues caused by these (Volkow et al., 2011). These factors are hard to influence or measure across populations. However, numerous external factors are also crucial to nutritional health. Education, relative prices, income, and the food environment also factor into people's consumption and health choices, and policies can be applied to influence them and realise health gains (Philipson and Posner, 2008).

**Education:** Obesity and education has been linked in many contexts, such as Devaux et al. (2011), who observe a lower probability for obesity in Australia, Canada, England, and Korea with a steeper gradient for women. These inequalities remain even after accounting for the nutritional knowledge of individuals (Variyam et al., 1998). Additionally, Variyam et al. (1998) found that obesity rates could be reduced by better information or education for subgroups such as men or Black or Hispanic ethnicities.

**Price environment:** Substitution of nutritional health for higher energy density or pleasure could lead to inequalities (Drewnowski and Specter, 2004). A study on US scanner data showed that distortions in the relative price environment could lead to 40% higher prices for fruit and vegetables and, consequently, underconsumption by 15% or about a third of the gap from recommended intakes (Pancrazi et al., 2020). The authors further find that accounting for prices, there are only slight differences across income levels, and therefore the price distortions can be corrected by subsidy.

**Income:** Income can directly affect food choices as people budget for food consumption and other goods. Accounting for prices income and preferences have been found to be the major drivers of nutritional inequality, after instrumenting for local price environment using US scanner panel data (Amano-Patino, 2019). Others have observed a strong correlation between several debt measures and obesity but have failed to find a causal link in a fixed-effects model after controlling for unobserved heterogeneity (Komlos et al., 2004).

**Food environment:** Fast-food outlet density and availability also been found significant in many studies for nutritional health outcomes. Neighbourhoods with high fast-food outlet density, "food swamps", have been linked to worse nutritional outcomes (Cooksey-Stowers et al., 2017), as well as low concentrations of available healthy food options, "food deserts", (Walker et al., 2010; Ghosh-Dastidar et al., 2014). A systematic showed that studies using comprehensive food environment

description have succeeded in finding a significant association between obesity and food environment (Cobb et al., 2015).

## 2.2 The work of Aiello et al. (2019)

This paper is greatly informed by the work of Aiello et al. (2019), titled “Large-scale and high-resolution analysis of food purchases and health outcomes”. They constructed both the food consumption and health outcome data used in this research. In their research, they identify the relationships between nutritional consumption and health outcomes and some socioeconomic controls using OLS. Specifically, they have found income insignificant for health outcomes, concluding that it is not as important as suggested by the literature. Closer inspection of the data used, however, reveals that the variables used in their modelling is not descriptive of income but is a measure of income deprivation from Index of Multiple Deprivation (IMD) (Data.gov.uk, 2021). This is confirmed by comparing the variables from the IMD and the paper; it is easy to see that they are the exact same for corresponding variables. The IMD is not well suited for qualifying affluence or deprivation of an area or comparing small areas (Department of Communities and Local Government, 2015). This calls into question the counterintuitive relationships observed in the paper and illustrates the need for precise measurement.

# 3. Data

The dataset constructed describes 267 variables for 4552 Lower Super Output Areas (LSOA) in London as described by the 2011 Census (Office of National Statistics, 2019). The data is composed of primary sources: (1) The main dataset courtesy of Prof. Luca Maria Aiello, describing nutritional health and food consumption. This dataset has been used in the paper by Aiello et al. (2019), as revealed by Aiello (2021). (2) Data relating to the local food environment was constructed using the Google Places API (Google Developers, 2015). (3) Data for socioeconomic characteristics and physical environment from the London Datastore (Greater London Authority, 2014).

## 3.1 Nutritional Health

Local nutritional health is described using standardised counts of prescriptions relating to diet-related illnesses, specifically hypertension, cholesterol, and diabetes. This data is provided by Prof. Luca Aiello (2021) and is based on Prescriptions written by all NHS practices that matched to LSOA's using the number of patients a practice had in the areas. The relevant medications were selected using the OpenPrescribing API based on the British National Formulary (OpenPrescribing.net, 2017, Joint Formulary Committee, 2019).

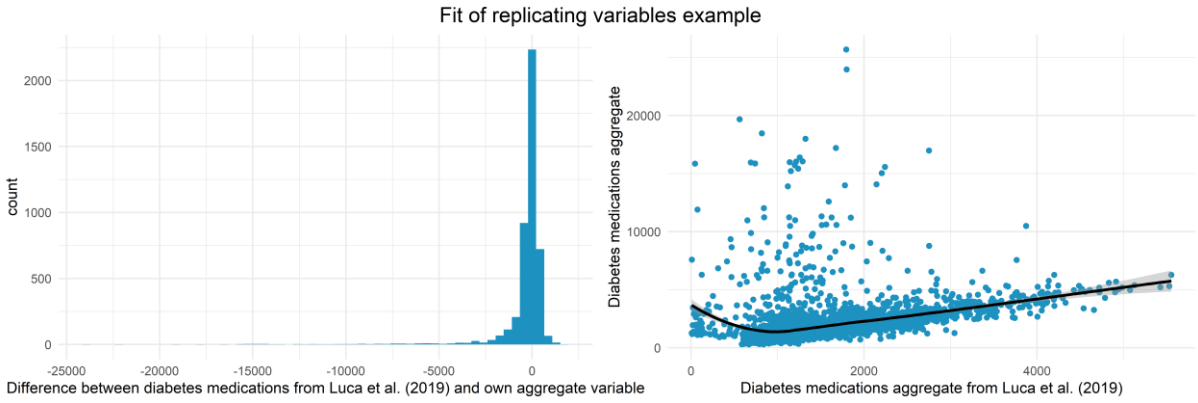



*Figure 1*

The validation and reconstruction of this data is done by replicating the described process above. However, some selection errors led to large outliers for some areas, while others fit approximately well, showcasing that the provided data accurately describes local nutritional health. Using multiple measures for nutritional health outcomes, if similar relations are observed, this reinforces the robustness of observed relationships. Additionally, prescriptions are a better measure of policy-relevant outcomes as they can be extended to measure treatment costs, and BMI has been criticized as a measurement tool for obesity (Burkhauser and Cawley, 2008). The used outcome variables exhibit no apparent spatial clustering and ample variation in the normalised counts with larger tails in all the nutritional health measures.

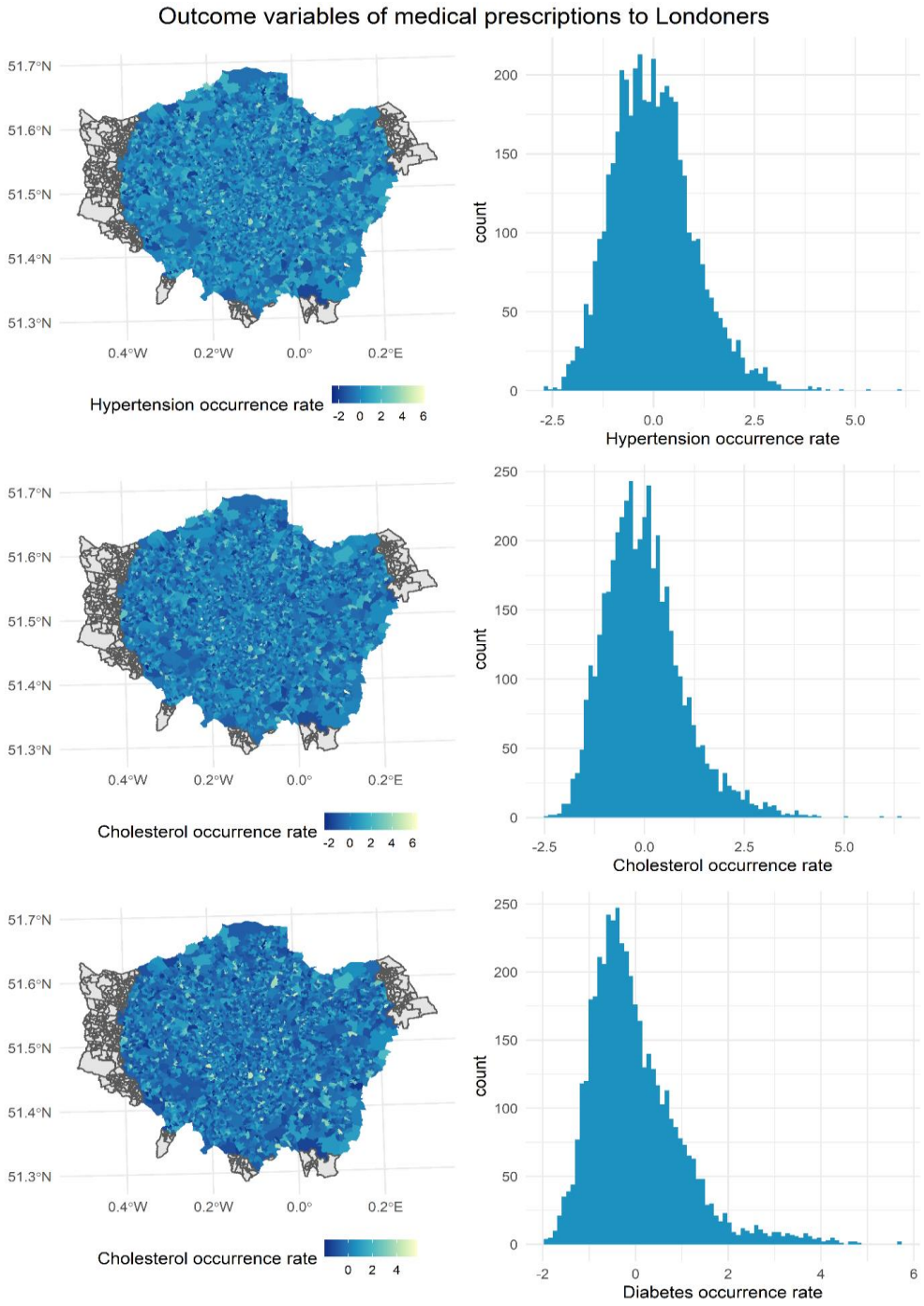



*Figure 2*

## 3.2 Food Consumption

Food choice variables are also provided by Prof. Luca Aiello (2021) and are based on purchases by TESCO Clubcard users in 2015. The dataset is comprised of different formulations of food consumption, such as the percentage of all calories from carbohydrates or the grams of fruit and vegetables consumed. Since these are based on the same purchase data by design, they are highly colinear across formulations and, in some cases, within formulations. Therefore, a choice will have to be made which formulation to use in analysing food choice. The formulations are comprehensive of food intakes and can approximate local food consumption given the market-leading position of TESCO with 29.1% in market share in 2015 (Kantar, 2021). A similar dataset has officially been

published, the TESCO 1.0 dataset (Aiello et al., 2020). However, after attempting to reproduce the regressions from Aiello et al. (2019) on the LSOA level, it is apparent (*Appendix A.*) that the TESCO 1.0 variables exhibit drastically different relationships. This can be seen by the little to no significance for caloric intake variables compared to results from the previous paper. The differences are due to filters imposed by TESCO that allowed for data publication (Aiello, 2021). This should be noted for the ease of future researchers attempting to use the TESCO 1.0 dataset.

Consequently, data provided by Prof. Luca Aiello is used in this research. This leads to 283 fewer observations than the entire London geography, as observations with low number of purchases in outer London were excluded due to low representability of local diets. On the contrary, a benefit of using purchase data from loyalty cards is that prices and the item selection are constant among establishments. Therefore, the influence of relative prices will be contained in the food choices for those shopping at TESCO. Based on this, local diets in London diverge far from WHO recommended intakes, indicating the extent of the issue (Figure 2.; WHO, 2002).

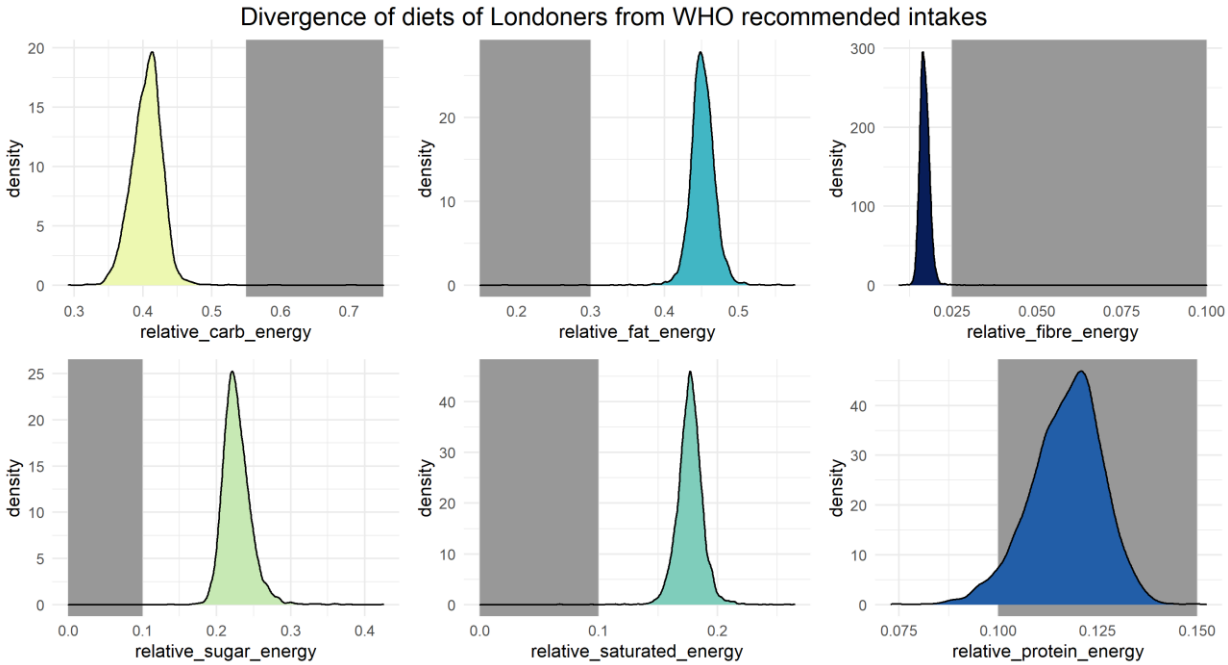



*Figure 3*

## 3.3 Food environment

To describe the food environment, the Google Places API was utilised to obtain all restaurants and grocery shops in London. Queries were based on the census geographies' centroids and the length of the polygons, describing a circular area. These had to be repeated three times as the API only allows for 20 returns at a time, 60 in total. The geographical overlap and the smaller size of the denser areas will have to account for higher densities. To comprehensively describe the food environment, queries were run for both "restaurants" and "grocery_or_supermarkets", resulting in 12040 unique restaurants and 3993 grocery stores. Using the returned description of the establishments were categorised into four price levels (0 for the missing), and the precise geolocation was recoded and used to assign them to LSOAs. Resulting in the number of establishments at each price level in the area (Figure 4.).

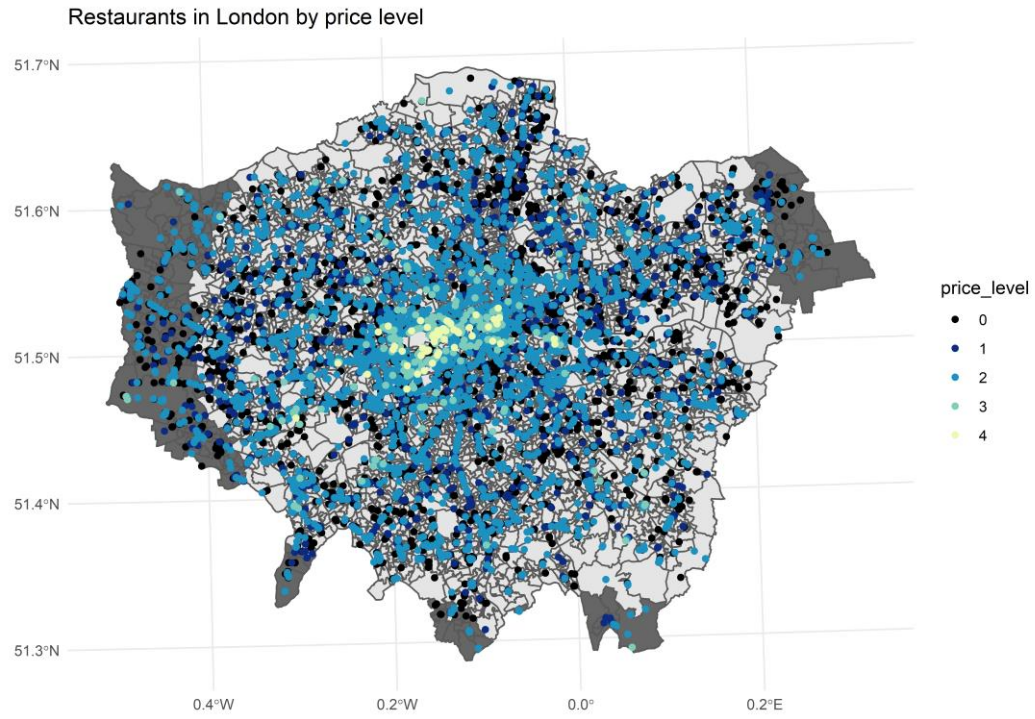



*Figure 4*

The use of the API for data collection was motivated by the feasibility study where the method was validated with in-person data collection in an urban setting in Bavaria and the lack of non-commercial data (Präger et al., 2019). Since the API limit may bias the results, the data was validated using official counts of licensed and unlicensed restaurants in boroughs (Greater London Authority, 2017). The difference between the official numbers in 2017 and the API results is uniquely high in inner London (Figure 5.). The overlap and density of smaller areas somewhat mediates this error since some LSOA's still have more than 60 establishments. None-the-less, all of these variables can be considered downward biased; this suggests that there are substantial limits to the use of the method, especially in describing high-density food environments.

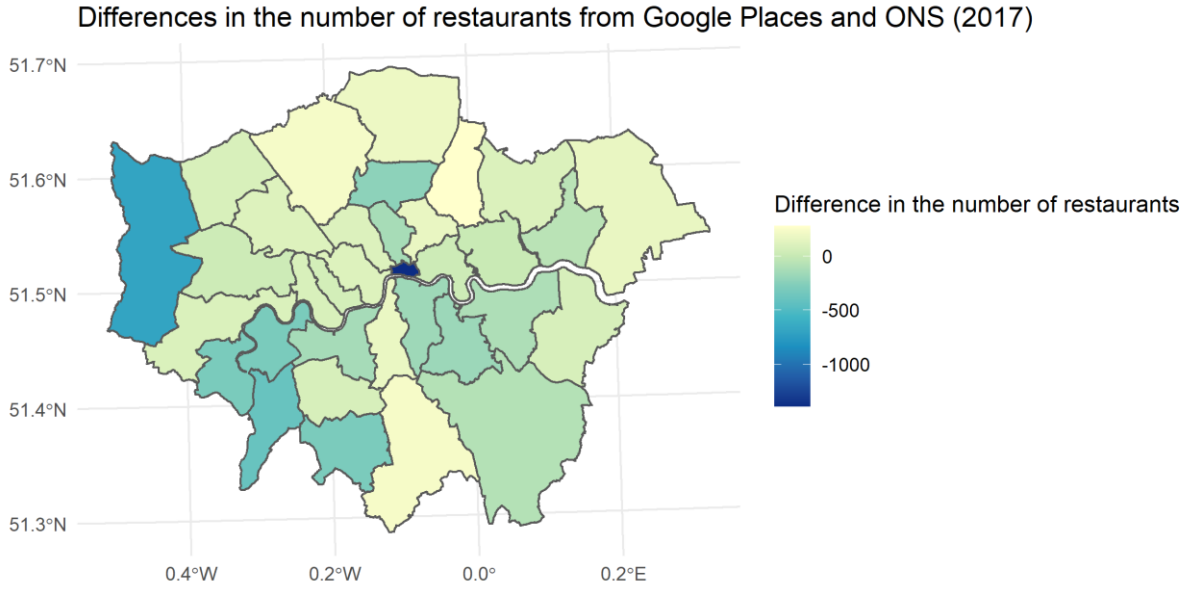



*Figure 5*

## 3.4 Socioeconomic & Environment

Variables describing the socioeconomic characteristics of residents and living environment come from the London Datastore and are based on the 2011 census from the Office of National Statistics (GLA, 2014; ONS, 2019). These contain a large number of income measures as well as socioeconomic descriptors such as the socioeconomic characteristics of the areas such as the percentage of people who describe themselves as "% White", "% Buddhist" or have "No_qualifications". Data also describes several measures for the living environment in these areas such as the type of buildings "Flat_maisonette_or_apartment", public transport accessibility "..4.6..good.access." or the number of traffic accidents "X2011_Total". These allow understanding of the relation between socioeconomic status and health outcomes. As shown by plotting the average log income of areas against the disease occurrence rate (Figure 6.), the negative relations showcase the large income-related inequalities present in London. Negative correlations are significant and large for all diseases hypertension (correlation = -0.4714815, p-value = <0.001), cholesterol (correlation = -0.4308745, p-value = <0.001), and diabetes (correlation = -0.5782858, p-value = <0. 001).

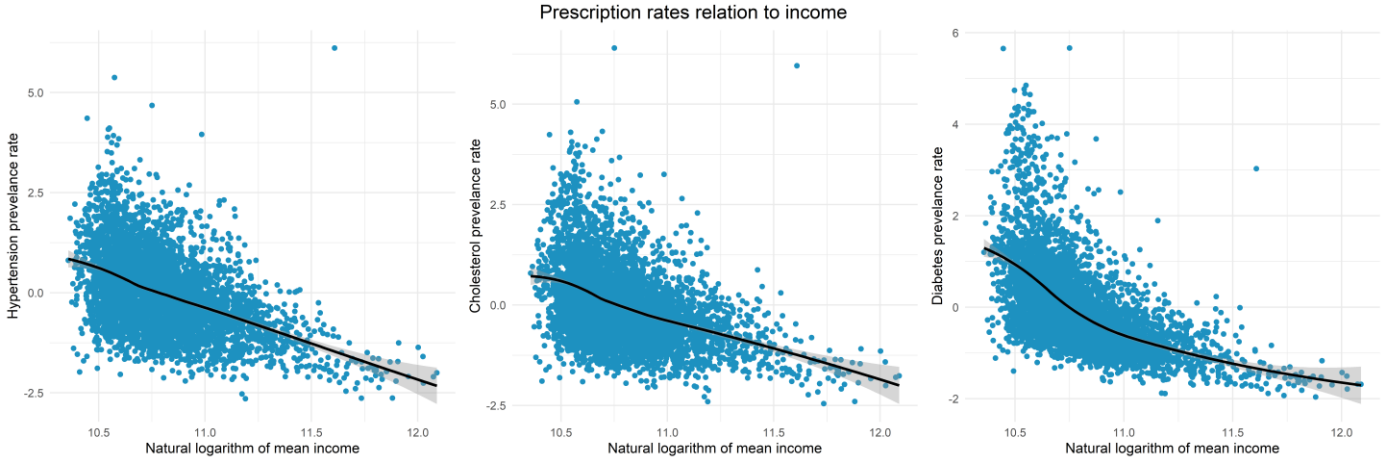



*Figure 6*

# 4. Methodology

This research aims to explain these health inequalities present among Londoners by examining the impact of income, food environments and food choice and showcase the potential of machine learning methods for such complex interaction cases.

1. Given the large number of potential factors influencing health outcomes, various regularization methods are utilised to identify variables best describing health outcomes while avoiding over-specification.
2. Post-Selection models are specified and compared to a naïve model motivated by the literature.
3. To account for the possible complex endogeneity of food choice, an Instrumental Variable LASSO model is utilised, to identify the effect of income and food choice.

## 4.1 Model selection

Given the number of potentially relevant factors to nutritional health from income, the food environment, education culture and more, the risk of overfitting and misidentifying relevant regressors is considerable. To systematically select variables for modelling, several penalised linear models were utilised, specifically, Ridge, LASSO, Adaptive LASSO, Elastic Net, and Adaptive Elastic Net. These regressions all utilise various kinds of penalty terms to the coefficients shrinking them towards zero. For instance, the coefficients (*Equation 1.*) for a ridge regression are downward biased by the second expression where λ is the parameter setting the strength of the L2 penalty imposed on the coefficients. Adding this bias to the regression coefficients allows correct large collinearities present in the data, specifically between food choice variables identifying the relevant regressors reliably (Schroeder et al., 1990). However, Ridge regression does not perform model selection as coefficients cannot be zeroed given the quadratic term from the model, for that L1 penalty is needed, such as in the Lasso regression. The LASSO regressions using L1 penalty is robust against the numerous complex outliers present in the data but are particularly sensitive to collinearities (Chong and Jun, 2005).

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left[ \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{K} \beta_j^2 \right], \lambda \geq 0$$

*Equation 1.*

To get the best of both methods, I use Elastic Net, an approach that mixes between L1 and L2 penalties (*Equation 2.*). The Elastic Net allows for the inclusion of highly collinear variables, outliers, and model selection.

$$\beta_{\text{ElasticNet}} = \arg\min_{\beta} \left[ \sum_{i=1}^{N} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{K} \left|\beta_j\right| + \lambda_2 \sum_{j=1}^{K} \beta_j^2 \right], \lambda_1 \geq 0, \lambda_2 \geq 0$$

*Equation 2.*

Since the penalisation introduces bias into the estimation to correct this bias and aid selection, a second-round penalised regression can be performed where the penalty term depends on results from (*Equation 2.*). This is called the Adaptive Elastic net (*Equation* 3; Zou and Zhang, 2009). This method applies a higher penalty for variables with low coefficients from the first stage while reducing the penalty for those variables with high coefficients despite the penalty, removing some of the bias on significant parameters.

$$\beta_{\text{AdaptiveElasticNet}} = \arg\min_{\beta}\left[\sum_{i=1}^{N}(y_i - x_i\beta)^2 + \lambda_1\sum_{j=1}^{K}\omega_j\left|\beta_j\right| + \lambda_2\sum_{j=1}^{K}\beta_j^2\right], \lambda_1 \geq 0, \lambda_2 \geq 0, \omega_j = \frac{1}{\left|\hat{\beta}_{j,\,\text{ElasticNet}}\right|}$$

*Equation 3.*

To choose optimal penalisation parameters, 10-fold cross-validation of the regression is performed where a randomly selected $9/10^{th}$ of the data is used for sample estimation and $1/10^{th}$ for evaluation of model performance. This procedure is repeated 10 times, and the penalty terms selected are those that minimise the mean squared error of the model. Adaptive Elastic Net is utilised for selection as it is robust against outliers as well as multicollinearity; the selection is then confirmed, comparing results between regularisation methods.

## 4.2 Post-Selection OLS

For the selected variables common among the three Adaptive Elastic Net specifications, post estimation is performed via Ordinary Least Squares. This is done to obtain confidence intervals for estimates and compare the performance of the model selection methods with a literature motivated approach. Additionally, to confirming the results from the Adaptive Elastic Net models, the post-selection OLS has several desirable qualities for identification and bias (Belloni and Chernozhukov, 2013).

## 4.3 LASSO-IV

Clearly, food choice is endogenous to health outcomes as people choose what to eat as an input to their health outcomes. A set of 26 instruments are considered that describe food environmental characteristics of the neighbourhoods, such as type of houses, number of traffic accidents, and house prices in the area, restaurant, and grocery store numbers (Appendix B. for a comprehensive summary). Instrument selection is further reinforced by Ohri-Vachaspati and Leviton (2010) and Saelens and Glanz (2009), both of whom assess a variety of environmental instruments for health outcomes used in the literature.

To deal with the complex endogeneity of food choice a post regularised instrumental variable approach called the IV-LASSO is used (Chernozhukov et al., 2015). This method allows for automatic categorisation and selection of exogenous controls and instruments and capturing more complex relationships. This method is based on the standard two-stage least squared methods with several lasso models preceding it selecting and categorising exogenous and instrumental variables.

$$1. \quad y_i = d_i\alpha + x_i'\beta + \varepsilon_i,$$
$$2. \quad d_i = x_i'\gamma + z_i'\delta + u_i,$$

*Equation 4.*

The algorithm works the following way, given the stylised form of the problem (Equation 4.). Where $y_i$ are the nutritional health outcomes, $d_i$ are endogenous variables of food consumption and income, $x_i$ are exogenous socioeconomic controls, and $z_i$ are the set of instruments. (1) The regression is performed of endogenous parameters on exogenous and instrumental variables via LASSO $d_i = x_i'\gamma + z_i'\delta + u_i$ to get coefficients $\hat{\gamma}$ , $\hat{\delta}$ , and fitted values $\hat{d}_i$. (2) Regression of outcome variables on exogenous variables is performed with LASSO $y_i = x_i'\theta + p_i^y$ giving the results $\hat{\theta}$ and $p_i^y$. (3) LASSO is used to regress the fitted values from the first regression on exogenous variables $\hat{d} = x_i'\hat{\gamma} + z_i'\hat{\delta} = x_i'\vartheta - v_i$ to obtain $v_i$ and $\hat{\vartheta}$. To identify the effect of food choice $\hat{\alpha}$ two stage least squares estimation is used $p_i^y = p_i^d\alpha \mid v_i$ where the outcome variable is $p_i^y \coloneqq y_i - x_i'\hat{\theta}$ the endogenous variables are $p_i^d \coloneqq d_i - x_i'\hat{\vartheta}$ and instruments are $v_i \coloneqq x_i'\hat{\gamma} + z_i'\hat{\delta} - x_i'\hat{\vartheta}$. The method enables selection of only relevant instruments negating the need for instrument relevance checks, however instrument exogeneity still can be questioned (Chernozhukov et al., 2015).

# 5. Results & Discussion

First, given the highly collinear nature of different formulations of the food consumption data, regularisation methods were used to choose between different formulations (*Appendix C.*). Fraction-of-nutrients consumption and categorical-items formulations were selected for further analysis, based on adjusted $R^2$ values from Adaptive Elastic Net results while using significantly fewer variables than other formulations. The big differences in the estimates between formulations is interesting since the literature includes a wide variety of data formulations, and the most appropriate formulation is rarely discussed, possibly biasing results.

## 5.1 Model selection

Several models were fitted using all possible regularisation methods for all three outcome variables and two formulations of food consumption, ensuring robustness. Out of the 176 relative-nutrition-based variables, 11 were selected in all Adaptive Elastic Net models, 31 variables were selected from the items-based formulation from 184 variables. Results from the Adaptive Elastic Net selection are presented in (Figure 7.) showing the coefficient estimates for the three outcome variables for the items-based and the relative-nutrient based data. Interestingly more variables are selected in the items-formulation, but this only leads to marginal improvement in adjusted $R^2$. Size of coefficients and variables selected are common amongst regularisation methods used, allowing for confidence in the selection. In both methods, other than the nutritional variables, Log of Income is identified as an important negative predictor for the three illnesses, and nutritional variables are correctly selected as the largest coefficients. Additionally, food environmental variables such as “Restaurants without price level” are identified in both data formulations despite the downward bias along with some common socioeconomic controls. Log Income and the food environment have a strong effect on nutritional health outcome while controlling for food choices.

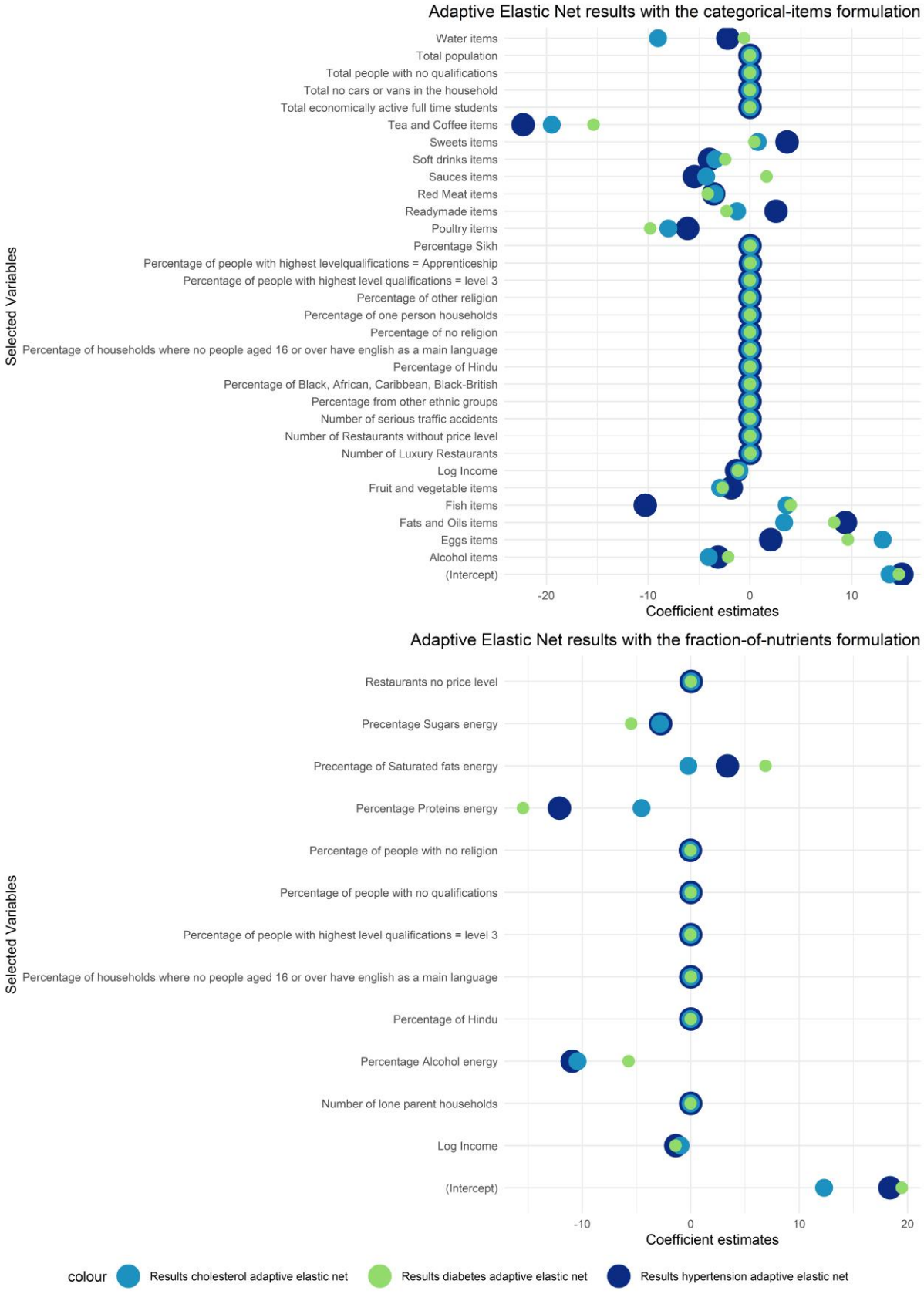



*Figure 7. Note: sizes and colours allow for better visibility*

## 5.2 Post-selection OLS

To reduce remaining biases and aid interpretation, OLS models were fitted to selected variables. Multicollinearity also remains an issue in the categorical-items formulations as measured by Variance Inflation Factors (up to 6 for "Fruit and vegetables" compared to 3 for Log Income in the nutrient formulation) remains high (Craney and Surles, 2002). Highly collinear variables could cause issues with the late IV-LASSO estimation, and dropping these variables would invalidate the model selection, given that the categorical-items adjusted $R^2$ are only marginally higher, results for this model are not presented. As heteroscedasticity is detected (Breusch-Pagan test p-value<0.001, White's test for heteroscedasticity p-value<0.001) in the residuals heteroscedasticity robust standard errors are used. (*Table 1.)* presents the results from the Post selection OLS for the variables common from Adaptive Elastic Nets.

*Table 1*

**Post selection regression results**

| | *Dependent variable:* | | |
|---|---|---|---|
| | Hypertension rate | Cholesterol rate | Diabetes rate |
| | (1) | (2) | (3) |
| Log Income | -1.856*** | -1.527*** | -1.554*** |
| | (0.080) | (0.083) | (0.065) |
| Total population | 0.0001* | 0.0003*** | 0.001*** |
| | (0.0001) | (0.0001) | (0.00004) |
| % Hindu | 0.002 | -0.009*** | 0.006*** |
| | (0.002) | (0.002) | (0.002) |
| % No religion | -0.024*** | -0.034*** | -0.042*** |
| | (0.002) | (0.002) | (0.001) |
| Number of houses owned outright | 0.002*** | 0.0002 | -0.004*** |
| | (0.0002) | (0.0002) | (0.0002) |
| All lone parent households with dependent children | -0.007*** | -0.009*** | -0.010*** |
| | (0.001) | (0.001) | (0.0004) |
| % highest level of qualification level 3 qualifications | -0.006 | -0.013*** | -0.027*** |
| | (0.004) | (0.005) | (0.004) |
| Restaurants no price level | 0.044*** | 0.043*** | 0.024*** |
| | (0.006) | (0.006) | (0.005) |
| % Saturated fats energy | -5.555*** | -3.144** | 0.440 |
| | (1.185) | (1.230) | (0.961) |
| % Sugars energy | -1.136 | -2.076** | -5.034*** |
| | (0.993) | (1.031) | (0.806) |

| | | | |
|---|---|---|---|
| % Proteins energy | -6.410* | -7.218** | -9.780*** |
| | (3.507) | (3.641) | (2.846) |
| % Fibres energy | -79.242*** | -84.673*** | -69.552*** |
| | (11.975) | (12.430) | (9.716) |
| % Alcohol energy | -6.208 | -10.069* | -2.624 |
| | (5.839) | (6.061) | (4.738) |
| Diversity of nutrient calories | -3.826 | -2.856 | -3.136 |
| | (3.734) | (3.875) | (3.029) |
| Constant | 26.880*** | 22.948*** | 23.701*** |
| | (2.473) | (2.567) | (2.007) |
| Observations | 4,552 | 4,552 | 4,552 |
| $R^2$ | 0.424 | 0.379 | 0.621 |
| Adjusted $R^2$ | 0.422 | 0.377 | 0.620 |
| Residual Std. Error (df = 4537) | 0.760 | 0.789 | 0.617 |
| F Statistic (df = 14; 4537) | 238.418*** | 198.005*** | 530.297*** |

*Note: Robust standard errors in brackets*

*p**p***p<0.01

Only a few variables are found insignificant expectedly with the selected controls, showing the significance of socioeconomic explanatory variables. Here "Log Income" and "Restaurants with no price level" are found to be significant for all outcome variables. Notably, restaurants that the API did not provide price levels for have been found to be positively associated with worse nutritional health outcomes due to their influence on food choices as found in the literature (Public Health England, 2017). This is especially notable since these measures of the food environment are downward biased but clearly important factors to local health. Education effects remain relatively small compared to nutritional, income or food environment effects, bringing into question policies aiming to tackle nutritional health through education.

This relationship is further explored by comparing the selected model with a "naïve" model that is based on the literature and includes income and all education, and food environment variables along with religion and ethnicity controls (Appendix D.). In this model, the significance of nutritional variables falls for all outcome measures because of their interaction with the controls. This result showcases the issue with naïve specifications unguided control selection could lead to concluding that income matters more for hypertension rates or that local education level effects are approximately similar with the strongest protective effect for a high percentage of level 1 qualifications (Ons.gov.uk, 2013). Contrary to consistent income effects for nutritional health outcomes with some evidence to the significance of education level of the area. Note that the selected model only performs marginally worse in terms of adjusted $R^2$ than naïve models, which contain significantly more variables.

## 5.3 IV-LASSO results

To tackle the endogeneity issue of food choice and income IV-LASSO models were used to identify the effect of food choice over and above what is explained by socioeconomic and environmental factors. The 28 instrument and 36 exogenous variables are used in this regression. Instruments are the environmental descriptors of the areas that can be argued to be exogenous to food choice given that food choice occurs on a very different timeframe than the choice of environment (Appendix D., for summary statistics of instruments).

*Table 2. Note: Robust standard errors*

**Lasso Instrumental variable results Cholesterol**

| | Estimate | Std. Error | t-statistics | p-value |
|---|---|---|---|---|
| Total energy consumption | 0.003 | 0.004 | 0.712 | 0.476 |
| % Carbs energy | -3.565 | 13.265 | -0.269 | 0.788 |
| % Fats energy | 13.773 . | 8.165 | 1.687 | 0.092 |
| % Saturated fats energy | -3.703 | 6.395 | -0.579 | 0.563 |
| % Sugars energy | -9.049 | 19.733 | -0.459 | 0.647 |
| % Proteins energy | 33.100 | 39.328 | 0.842 | 0.400 |
| % Fibres energy | 105.081 | 581.315 | 0.181 | 0.857 |
| % Alcohol energy | 22.824 | 262.032 | 0.087 | 0.931 |
| Log Income | -1.430 *** | 0.438 | -3.265 | 0.001 |
| (Intercept) | -0.070 | 0.070 | -0.993 | 0.321 |

$^{*}p$ $^{**}p$ $^{***}p<0.01$

Results from the IV-LASSO (Table 2.) show that while the choice of how much fats a person consumes is a significant positive predictor of cholesterol the rest of nutritional choice can be explained by socioeconomic factors to the extent that they become insignificant. The results are similar for hypertension and diabetes for the importance of income, but no food choice variables are found significant (Appendix F.).

Income seems to be an important predictor of nutritional health, accounting for socioeconomic setting, neighbourhood characteristics and food choice. This showcases both the importance of income and the environment; while local income does causally predict better nutritional health, the environment can explain nutritional intakes to the extent that local nutritional choice loses significance. This also means that the risk of obesity is greater for those in lower-income areas, adding to costs associated with low income. These results reinforce those showing the importance of income for nutritional health, such as Amano-Patino (2020), and those showing the importance of local environments for nutritional choice, such as (Ghosh-Dastidar et al., 2014).

Policies such a subsidising food expenditure of low-income areas, such as the one proposed by Amano-Patino (2020), would lead to health gains for London areas particularly affected by worse

nutritional outcomes. Furthermore, policies aiming to tackle the food environmental inequalities via zoning and growing availability of healthy food options or nudging could be a viable way to influence food choice and, therefore, nutritional health (Rozin et al., 2011).

# 6. Limitations and Extensions

There are numerous limitations to the results presented, such as the limit on the Google API, which biases the Restaurant estimators. To properly identify the effect of the food environment, a better description of the environment is needed. A more detailed query procedure can be developed for the Google API where the areas with high number of observations in the first round are split and queried again, hopefully resulting in higher resolution. Commercial datasets such as Leadsdeposit (2021) or other API's such as Tripadvisor (2021) can be utilised as they might provide a better description of the areas.

The use of the TESCO data might have biased the results as it embeds selection of TESCO Clubcard customers involving several confounders, such as the exclusion of people who do choose not to shop there or shop for select items from the store. The use of environmental controls somewhat mediates this issue; however, the effects can be further explored in future research.

Additionally, the model could be extended in several ways such as incorporating the local price environment with relation to the nutritional consumption data. This would be possible through reaggregation of purchase records, including prices this time, and better specifying the Google API query allowing for better identification of income effects. However, these factors can be argued to be controlled by the use of purchases from one chain and controlling for the food environment.

Addressing these issues would allow for the identification of the relative importance of income and the environment allowing for better specification of policies. A different approach would be correcting for the errors via spatial modelling, which allows for neighbouring areas characteristics to affect each other, thereby correcting for potential spatial autocorrelation.

# 7. Conclusion

This research aimed to explain the nutritional inequalities present in London with the use of high-resolution nutritional consumption and nutritional health. Several penalised regressions were explored to select relevant confounders to nutritional health, showcasing the strength of the methods for dealing with such complex settings. LASSO Instrumental Variable algorithm was used with the use of environmental instruments the identify to causal effect of income and food choice for nutritional health. As a result, nearly all nutritional variables were found insignificant, while income remained a significant negative predictor on nutritional health. The results showcase the importance of the environment and income for the nutritional health of Londoners.

Policies aiming to grow the food budget of Londoners and improving the food environment can lead to health gains in London. Education and food choice seem to have a relatively small effect on nutritional health. Therefore, policies aiming to improve nutritional health through these channels are unlikely to succeed.

# 8. Appendices

## 8.1 Appendix A:

**Reproducing Aiello et al. (2019) & comparing with TESCO 1.0**

| | *Dependent variable:* | | | | | |
|---|---|---|---|---|---|---|
| | Hypertension rate | Cholesterol rate | Diabetes rate | Hypertension rate | Cholesterol rate | Diabetes rate |
| | Luca et al. (2019) | Luca et al. (2019) | Luca et al. (2019) | TESCO 1.0 | TESCO 1.0 | TESCO 1.0 |
| Income | -0.164** | 0.070 | 0.644*** | -0.110 | -398.769*** | 434.188*** |
| | (0.069) | (0.078) | (0.082) | (0.073) | (144.947) | (156.255) |
| Education | 0.006*** | 0.006*** | 0.004*** | 0.012*** | 11.366*** | 11.905*** |
| | (0.001) | (0.001) | (0.001) | (0.001) | (1.156) | (1.246) |
| Average age | 0.012*** | 0.011*** | -0.007*** | 0.015*** | -28.991*** | -54.573*** |
| | (0.001) | (0.001) | (0.002) | (0.001) | (2.735) | (2.948) |
| % females | -1.771*** | -2.271*** | -4.179*** | -1.980*** | -3,101.236*** | -6,551.127*** |
| | (0.163) | (0.184) | (0.194) | (0.176) | (348.998) | (376.227) |
| Energy carb | 0.017*** | 0.016*** | 0.023*** | -0.001 | 0.866 | 2.065 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (1.548) | (1.669) |
| Energy fat | 0.007*** | 0.009*** | 0.015*** | 0.001 | 0.900 | 2.566 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (1.874) | (2.020) |
| Energy sugar | -0.008*** | -0.011*** | -0.025*** | 0.002 | -0.240 | -4.513 |
| | (0.001) | (0.002) | (0.002) | (0.001) | (2.969) | (3.200) |
| Energy protein | -0.041*** | -0.048*** | -0.097*** | 0.002 | -1.382 | -5.019 |
| | (0.004) | (0.004) | (0.004) | (0.004) | (7.763) | (8.369) |
| Energy fibre | -0.146*** | -0.137*** | -0.251*** | 0.004 | 70.685* | 51.213 |
| | (0.022) | (0.025) | (0.026) | (0.021) | (40.796) | (43.979) |
| Constant | 1.275*** | 1.685*** | 4.049*** | 1.186*** | 3,859.941*** | 6,333.462*** |
| | (0.118) | (0.132) | (0.140) | (0.135) | (267.421) | (288.286) |
| Observations | 4,550 | 4,550 | 4,550 | 4,550 | 4,550 | 4,550 |
| $R^2$ | 0.291 | 0.260 | 0.460 | 0.140 | 0.107 | 0.277 |
| Adjusted $R^2$ | 0.290 | 0.259 | 0.459 | 0.138 | 0.106 | 0.275 |
| Residual Std. Error (df = 4540) | 0.245 | 0.276 | 0.291 | 0.270 | 534.189 | 575.867 |
| F Statistic (df = 9; 4540) | 206.995*** | 177.492*** | 429.913*** | 81.817*** | 60.633*** | 193.184*** |

*Note:* *p**p***p<0.01

## 8.2 Appendix B:

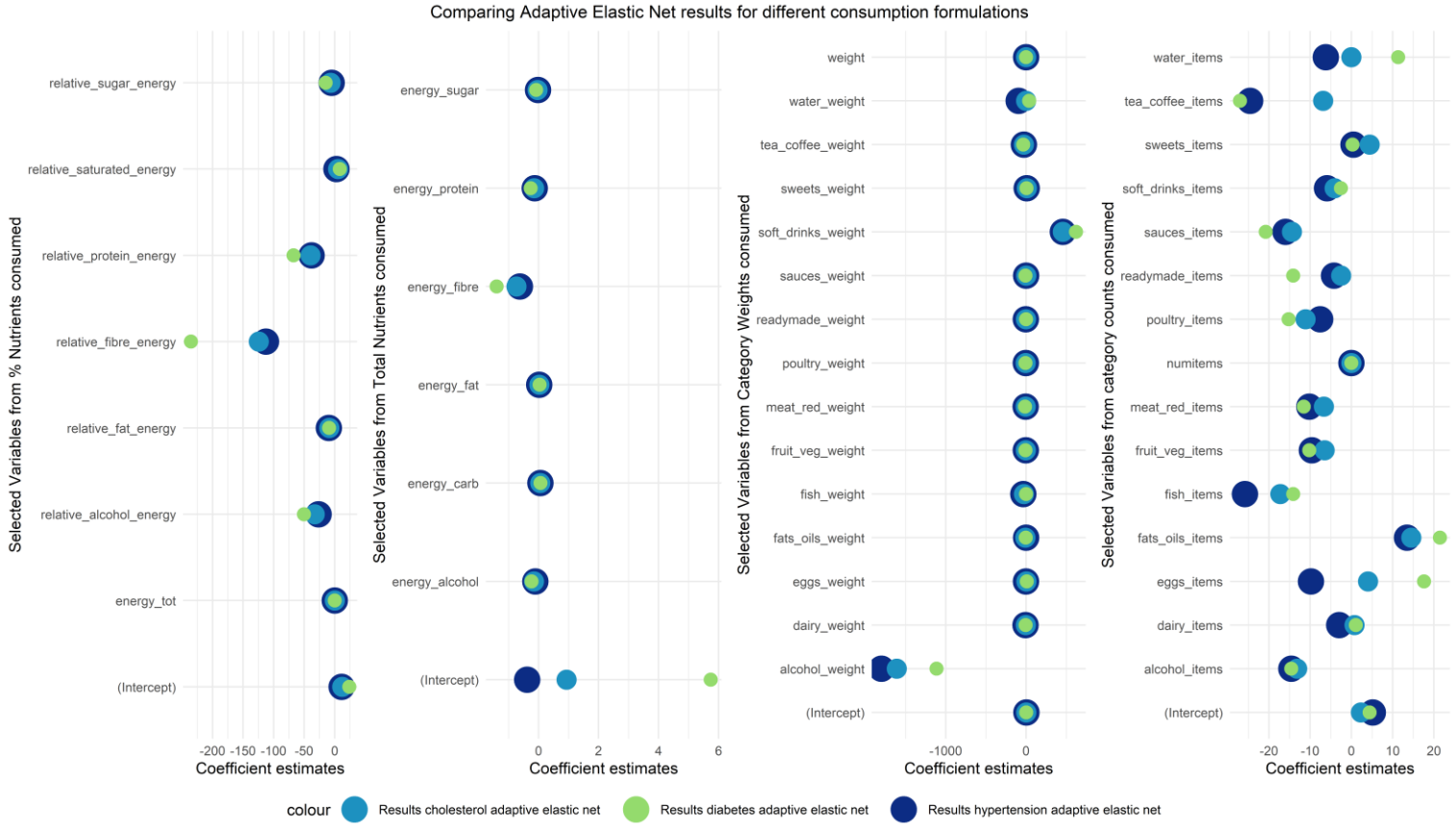
**Summary statistics of instruments used in IV-LASSO estimation**

| Statistic | N | Mean | St. Dev. | Min | Pctl(25) | Pctl(75) | Max |
|---|---|---|---|---|---|---|---|
| % Houses owned outright | 4,552 | 21.426 | 12.421 | 0.200 | 11.200 | 30.000 | 61.200 |
| % Owned with a mortgage or loan | 4,552 | 27.404 | 11.707 | 0.800 | 17.975 | 36.400 | 59.700 |
| % Socially rented | 4,552 | 23.949 | 20.383 | 0.000 | 7.000 | 37.100 | 90.900 |
| % Privately rented | 4,552 | 24.706 | 12.781 | 2 | 14.7 | 33.1 | 88 |
| % Whole house or bungalow detached | 4,552 | 6.175 | 9.606 | 0.000 | 1.600 | 6.300 | 94.900 |
| % Whole house or bungalow semi detached | 4,552 | 19.409 | 19.337 | 0.000 | 4.875 | 27.625 | 97.200 |
| % Whole house or bungalow terraced including end terrace | 4,552 | 24.131 | 18.618 | 0.000 | 8.700 | 36.900 | 88.300 |
| % Flat maisonette or apartment | 4,552 | 50.215 | 29.298 | 0 | 24.6 | 75.9 | 100 |
| House median prices | 4,552 | 332,611.200 | 219,474.000 | 0 | 217,000 | 372,625 | 3,377,000 |
| Number of house sales | 4,552 | 18.523 | 13.119 | 0 | 10 | 24 | 221 |
| PATL average | 4,552 | 3.835 | 1.596 | 0.300 | 2.700 | 4.800 | 8.000 |
| PATL poor access 1 – 0 % | 4,552 | 21.746 | 31.634 | 0.000 | 0.000 | 37.500 | 100.000 |
| PATL average access 2 - 3 % | 4,552 | 48.515 | 36.335 | 0.000 | 10.500 | 83.625 | 100.000 |
| PATL good access 4 - 6 % | 4,552 | 29.739 | 39.276 | 0 | 0 | 64.6 | 100 |
| Restaurants no price level | 4,552 | 1.043 | 2.038 | 0 | 0 | 1 | 26 |
| Restaurants cheap | 4,552 | 0.484 | 1.097 | 0 | 0 | 1 | 12 |
| Restaurants medium | 4,552 | 0.866 | 2.087 | 0 | 0 | 1 | 39 |
| Restaurants medium-high | 4,552 | 0.050 | 0.370 | 0 | 0 | 0 | 15 |
| Restaurants luxury | 4,552 | 0.013 | 0.164 | 0 | 0 | 0 | 6 |
| Grocery no price level | 4,552 | 0.486 | 1.002 | 0 | 0 | 1 | 11 |
| Grocery cheap | 4,552 | 0.175 | 0.455 | 0 | 0 | 0 | 4 |
| Grocery medium | 4,552 | 0.177 | 0.509 | 0 | 0 | 0 | 14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Grocery medium-high | 4,552 | 0.003 | 0.053 | 0 | 0 | 0 | 1 |
| Population density | 4,552 | 100.717 | 62.538 | 1 | 56 | 135 | 678 |
| Number of fatal traffic accidents | 4,552 | 0.032 | 0.185 | 0 | 0 | 0 | 2 |
| Number of serious traffic accidents | 4,552 | 0.558 | 1.156 | 0 | 0 | 1 | 34 |
| Number of slight traffic accidents | 4,552 | 5.517 | 8.089 | 0 | 1 | 7 | 252 |

## 8.3 Appendix C:

### Comparing different formulations of food-consumption



## 8.4 Appendix D:

**Naïve model results**

| | *Dependent variable:* | | |
|---|---|---|---|
| | Hypertension rate | Cholesterol rate | Diabetes rate |
| | (1) | (2) | (3) |
| Total energy consumption | 0.002 | 0.001 | -0.001 |
| | (0.002) | (0.002) | (0.001) |
| % Carbs energy | 6.160** | 7.672*** | 3.562* |
| | (2.878) | (2.948) | (2.036) |
| % Fat energy | 4.316 | 5.643* | 1.233 |
| | (3.052) | (3.120) | (2.209) |
| % Saturated fats energy | -0.043 | 0.700 | 1.830 |
| | (2.343) | (2.294) | (1.822) |
| % Sugars energy | -1.150 | -1.824* | -2.196*** |
| | (0.953) | (0.952) | (0.734) |
| % Proteins energy | 7.669** | 11.512*** | 5.297** |
| | (3.554) | (3.536) | (2.485) |
| % Fibres energy | -0.533 | -17.194 | -19.134** |
| | (12.504) | (13.559) | (9.574) |
| Log Income | -1.224*** | -0.876*** | -0.776*** |
| | (0.114) | (0.114) | (0.081) |
| % No qualifications | -0.011 | 0.003 | -0.021*** |
| | (0.006) | (0.006) | (0.005) |

| | | | |
|---|---|---|---|
| % Highest level of qualification level = level 1 qualifications | -0.036*** | -0.030*** | -0.045*** |
| | (0.009) | (0.010) | (0.007) |
| % Highest level of qualification level = level 2 qualifications | -0.005 | -0.020** | -0.056*** |
| | (0.009) | (0.009) | (0.007) |
| % Highest level of qualification level = level 3 qualifications | -0.012** | -0.014* | -0.038*** |
| | (0.008) | (0.008) | (0.006) |
| % Highest level of qualification level 4 qualifications and above | -0.018 | -0.017*** | -0.035*** |
| | (0.006) | (0.006) | (0.004) |
| % Christian | 0.123 | 0.122 | 0.145 |
| | (0.119) | (0.123) | (0.094) |
| % Buddhist | 0.033 | 0.036 | 0.029 |
| | (0.121) | (0.124) | (0.095) |
| % Hindu | 0.133 | 0.113 | 0.151 |
| | (0.119) | (0.123) | (0.094) |
| % Jewish | 0.110 | 0.114 | 0.144 |
| | (0.119) | (0.123) | (0.094) |
| % Muslim | 0.142 | 0.140 | 0.172* |
| | (0.119) | (0.123) | (0.094) |
| % Sikh | 0.136 | 0.125 | 0.169* |
| | (0.119) | (0.123) | (0.094) |
| % Other religion | 0.149 | 0.123 | 0.140 |
| | (0.120) | (0.124) | (0.094) |

| | | | |
|---|---|---|---|
| % No religion | 0.097 | 0.099 | 0.140 |
| | (0.119) | (0.123) | (0.094) |
| % Religion not stated | 0.082 | 0.084 | 0.119 |
| | (0.119) | (0.123) | (0.094) |
| % White | -3.896*** | -3.516*** | -2.756** |
| | (1.192) | (1.280) | (1.294) |
| % Mixed multiple ethnic groups | -0.233 | -0.272* | -0.186 |
| | (0.162) | (0.165) | (0.126) |
| % Asian, Asian-British | -0.167 | -0.198 | -0.128 |
| | (0.161) | (0.164) | (0.126) |
| % Black, African, Caribbean, Black-British | -0.175*** | -0.221 | -0.132 |
| | (0.161) | (0.164) | (0.126) |
| Total Population | -0.0001 | -0.00004 | -0.00002 |
| | (0.00005) | (0.0001) | (0.00004) |
| % Other ethnic group | -0.229 | -0.260 | -0.195 |
| | (0.161) | (0.164) | (0.126) |
| % BAME | -3.737*** | -3.313*** | -2.615** |
| | (1.197) | (1.285) | (1.297) |
| Restaurants no price level | 0.028** | 0.028*** | 0.016*** |
| | (0.007) | (0.007) | (0.005) |
| Restaurants cheap | 0.025*** | 0.022* | 0.024** |
| | (0.013) | (0.013) | (0.011) |
| Restaurants medium | 0.001 | 0.001 | -0.003 |
| | (0.006) | (0.006) | (0.005) |

| | | | |
|---|---|---|---|
| Restaurants medium-high | -0.047 | -0.063** | -0.034 |
| | (0.031) | (0.031) | (0.021) |
| Restaurants luxury | 0.088*** | 0.060 | 0.058 |
| | (0.050) | (0.050) | (0.040) |
| Grocery cheap | 0.037* | 0.036 | 0.011 |
| | (0.027) | (0.027) | (0.022) |
| Grocery medium | -0.017 | -0.020 | -0.030* |
| | (0.022) | (0.023) | (0.017) |
| Grocery medium-high | 0.059 | 0.040 | 0.091 |
| | (0.135) | (0.155) | (0.145) |
| Constant | 388.186*** | 345.093*** | 270.813** |
| | (120.074) | (128.816) | (129.834) |
| Observations | 4,552 | 4,552 | 4,552 |
| $R^2$ | 0.498 | 0.473 | 0.695 |
| Adjusted $R^2$ | 0.493 | 0.469 | 0.693 |
| Residual Std. Error (df = 4514) | 0.712 | 0.729 | 0.554 |
| F Statistic (df = 37; 4514) | 120.789*** | 109.673*** | 278.418*** |

*Note: Robust standard errors in parentheses*
*% Alcohol energy is excluded for collinearity*

*p**p***p<0.01

## 8.5 Appendix E:

**Lasso Instrumental variable results Hypertension**

| | Estimate | Std. Error | t-statistics | p-value |
|---|---|---|---|---|
| Total energy consumption | -0.001 | 0.004 | -0.287 | 0.774 |
| % Carbs energy | -13.015 | 15.242 | -0.854 | 0.393 |
| % Fats energy | 14.358 | 9.586 | 1.498 | 0.134 |
| % Saturated fats energy | 4.098 | 7.831 | 0.523 | 0.601 |
| % Sugars energy | -14.814 | 22.612 | -0.655 | 0.512 |
| % Proteins energy | -11.763 | 43.448 | -0.271 | 0.787 |
| % Fibres energy | 512.837 | 663.629 | 0.773 | 0.440 |
| % Alcohol energy | -142.850 | 308.356 | -0.463 | 0.643 |
| Log Income | -1.761 ** | 0.562 | -3.132 | 0.002 |
| (Intercept) | -0.006 | 0.077 | -0.079 | 0.937 |

**Lasso Instrumental variable results Diabetes**

| | Estimates | Std. errors | t-statistics | p-value |
|---|---|---|---|---|
| Total energy consumption | -0.009 | 0.006 | -1.544 | 0.123 |
| % Carbs energy | 5.990 | 22.295 | 0.269 | 0.788 |
| % Fats energy | -13.335 | 13.943 | -0.956 | 0.339 |
| % Saturated fats energy | 5.751 | 12.143 | 0.474 | 0.636 |
| % Sugars energy | 2.367 | 33.105 | 0.071 | 0.943 |
| % Proteins energy | -87.566 | 63.209 | -1.385 | 0.166 |
| % Fibres energy | -573.708 | 956.519 | -0.600 | 0.549 |
| % Alcohol energy | 259.109 | 447.758 | 0.579 | 0.563 |
| Log Income | -1.795* | 0.859 | -2.089 | 0.037 |
| (Intercept) | 0.148 | 0.110 | 1.345 | 0.179 |

## 8.6 Appendix F:

**Summary of packages used and references.**

| Package name | What it is used for | Reference |
|---|---|---|
| tidyverse | For data wrangling | Wickham et al., 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686 |
| here | For file management | Müller, K., 2020. here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here |
| sf | For managing spatial data | Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal 10 (1), 439-446, https://doi.org/10.32614/RJ-2018-009 |
| rgdal | For transforming spatial projection of geocoded data | Bivand, R., Keitt, T., and Rowlingson, B., 2021. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-23. https://CRAN.R-project.org/package=rgdal |
| janitor | For renaming variables | Firke, S., 2021. janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor |
| skimr | For data exploration | Waring, E., Quinn, M., McNamara, A., Arino de la Rubia, E., Zhu, H., and Ellis, S., 2020. skimr: Compact and Flexible Summaries of Data. R package version 2.1.2. https://CRAN.R-project.org/package=skimr |
| reshape2 | For the melt function producing quick plots | Wickham, H., 2007. Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. URL http://www.jstatsoft.org/v21/i12/. |
| googleway | For accessing the Google API | Cooley, D., 2020. googleway: Accesses Google Maps APIs to Retrieve Data and Plot Maps. R package version 2.7.3. https://CRAN.R-project.org/package=googleway |
| OpenprescribingR | For accessing the OpenPrescribing API | Taylor, F., 2021. openprescribingR: Load OpenPrescribing data directly into R. R package version 0.0.0.9000. https://github.com/fergustaylor/openprescribingR/ |
| glmnet | For penalised regressions | Friedman, J., Hastie, T., and Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL https://www.jstatsoft.org/v33/i01/. |
| coefplot | For extracting coefficients from cv.glmnet | Lander, J. P., 2021. coefplot: Plots Coefficients from Fitted Models. R package version 1.2.7. https://CRAN.R-project.org/package=coefplot |
| hdm | For IV-LASSO estimation | Chernozhukov, V., Hansen, C., Spindler, M., 2016. hdm: High-Dimensional Metrics R Journal, 8(2), 185-199. URL https://journal.r-project.org/archive/2016/RJ-2016-040/index.html. |

| | | |
|---|---|---|
| lmtest | For evaluating regression results | Zeileis, A., Hothorn, T., 2002. Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/ |
| caret | For tuning penalty terms | Kuhn, M., (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret |
| skedastic | For White's test for heteroscedasticity | Farrar, T. J., 2020. skedastic: Heteroskedasticity Diagnostics for Linear Regression Models. R Package version 1.0.0. University of the Western Cape. Bellville, South Africa. https://github.com/tjfarrar/skedastic |
| sandwich | For producing heteroscedasticity robust standard errors | Zeileis, A., Köll, S., Graham, N., 2020. "Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R." _Journal of Statistical Software_, *95*(1), 1-36. doi: 10.18637/jss.v095.i01 (URL: https://doi.org/10.18637/jss.v095.i01). |
| broom | For accessing regression results | Robinson, D., Hayes, A., and Couch, S., 2021. broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.4. https://CRAN.R-project.org/package=broom |
| xtable | For producing tables from cv.gmlnet results | Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J., 2019. xtable: Export Tables to LaTeX or HTML. R package version 1.8-4. https://CRAN.R-project.org/package=xtable |
| stargazer | For presenting regression results | Hlavac, M., 2018. stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer |
| ggplot2 | For producing the plots presented | Wickham, H., 2016ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. |
| ggpubr | For combining plots and titling | Kassambara, A., 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr |

# 4. Bibliography

## 9.1 Data sources

Aiello, L.M., 2021. Email communications. November 2020 – March 2021

Aiello, L.M., Quercia, D., Schifanella, R. and Del Prete, L., 2020. Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London. Scientific Data, 7(1), pp.1-11. [online] Available at: <https://www.nature.com/articles/s41597-020-0397-7> [Accessed 26 April 2021].

Google Developers. 2015. *Google Places API*. [online] Available at: <https://developers.google.com/maps/documentation/places/web-service/search> [Accessed 26 April 2021].

Greater London Authority (GLA). 2014. *LSOA Atlas – London Datastore*. [6 October 2014] Available at: <https://data.london.gov.uk/dataset/lsoa-atlas> [Accessed 26 April 2021].

Greater London Authority (GLA). 2017. *Number of public houses, licenced clubs, restaurants and takeaways by Borough.* [1 January 2017] Available at: <https://data.london.gov.uk/dataset/pubs-clubs-restaurants-takeaways-borough> [Accessed 26 April 2021].

Office of National Statistics. 2019. *Census geography - Office for National Statistics*. [3 October 2019] Available at: <https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography> [Accessed 26 April 2021].

OpenPrescribing.net, 2017. EBM DataLab, University of Oxford, Available at: <https://openprescribing.net/api/ > [Accessed 26 April 2021].

## 9.2 Literature

Aiello, L.M., Schifanella, R., Quercia, D. and Del Prete, L., 2019. Large-scale and high-resolution analysis of food purchases and health outcomes. EPJ Data Science, 8(1), p.14.

Amano-Patiño, N., 2020. *Nutritional Inequality: The Role of Prices, Income, and Preferences* (No. 1909). Faculty of Economics, University of Cambridge.

Anzman, S.L., Rollins, B.Y. and Birch, L.L., 2010. Parental influence on children's early eating environments and obesity risk: implications for prevention. *International journal of obesity*, *34*(7), pp.1116-1124.

Basu, S., Yoffe, P., Hills, N. and Lustig, R.H., 2013. The relationship of sugar to population-level diabetes prevalence: an econometric analysis of repeated cross-sectional data. *PloS one*, *8*(2), p.e57873.

Belloni, A., Chernozhukov, V. and Hansen, C., 2011. Lasso methods for gaussian instrumental variables models. [online] Available at: <https://arxiv.org/abs/1012.1297> [Accessed 29 April 2021].

Burkhauser, R.V. and Cawley, J., 2008. Beyond BMI: the value of more accurate measures of fatness and obesity in social science research. *Journal of health economics*, *27*(2), pp.519-529.

Campbell, D., Walker, P. and Siddique, H., 2020. *Boris Johnson to unveil £10m ad campaign to cut obesity in England*. [online] The Guardian. Available at: <https://www.theguardian.com/society/2020/jul/25/boris-johnson-to-unveil-10m-ad-campaign-to-cut-obesity-in-england> [Accessed 29 April 2021].

Chernozhukov, V., Hansen, C. and Spindler, M., 2015. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, *7*(1), pp.649-688.

Chong, I.G. and Jun, C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and intelligent laboratory systems*, *78*(1-2), pp.103-112.

Cobb, L.K., Appel, L.J., Franco, M., Jones-Smith, J.C., Nur, A. and Anderson, C.A., 2015. The relationship of the local food environment with obesity: a systematic review of methods, study quality, and results. *Obesity*, *23*(7), pp.1331-1344.

Cooksey-Stowers, K., Schwartz, M.B. and Brownell, K.D., 2017. Food swamps predict obesity rates better than food deserts in the United States. *International journal of environmental research and public health*, *14*(11), p.1366.

Craney, T.A. and Surles, J.G., 2002. Model-dependent variance inflation factor cutoff values. *Quality Engineering*, *14*(3), pp.391-403.

Data.gov.uk. 2021. *English Indices of Deprivation 2015 - LSOA Level - data.gov.uk*. [online] Available at: <https://data.gov.uk/dataset/8f601edb-6974-417e-9c9d-85832dd2bbf2/english-indices-of-deprivation-2015-lsoa-level> [Accessed 23 April 2021].

Department of Communities and Local Government, 2015. *English Indices of Deprivation 2015-Statistical Release-Main Findings.* [online] Available at: <https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015> [Accessed 23 April 2021].

Department of Health and Social Care. 2020. *New obesity strategy unveiled as country urged to lose weight to beat coronavirus (COVID-19) and protect the NHS*. [online] Available at: <https://www.gov.uk/government/news/new-obesity-strategy-unveiled-as-country-urged-to-lose-weight-to-beat-coronavirus-covid-19-and-protect-the-nhs> [Accessed 26 April 2021].

Devaux, M., Sassi, F., Church, J., Cecchini, M. and Borgonovi, F., 2011. Exploring the relationship between education and obesity. *OECD Journal: Economic Studies*, *2011*(1), pp.1-40.

Devaux, M. and Sassi, F., 2013. Social inequalities in obesity and overweight in 11 OECD countries. *The European Journal of Public Health*, *23*(3), pp.464-469.

Dietz, W. and Santos-Burgoa, C., 2020. Obesity and its implications for COVID-19 mortality. *Obesity*, *28*(6), pp.1005-1005.

Drewnowski, A. and Specter, S.E., 2004. Poverty and obesity: the role of energy density and energy costs. *The American journal of clinical nutrition*, *79*(1), pp.6-16.

El-Sayed, A.M., Scarborough, P. and Galea, S., 2012. Unevenly distributed: a systematic review of the health literature about socioeconomic inequalities in adult obesity in the United Kingdom. *BMC public health*, *12*(1), pp.1-12.

Ghosh-Dastidar, B., Cohen, D., Hunter, G., Zenk, S.N., Huang, C., Beckman, R. and Dubowitz, T., 2014. Distance to store, food prices, and obesity in urban food deserts. *American journal of preventive medicine*, *47*(5), pp.587-595.

Hoebel, J., Kuntz, B., Kroll, L.E., Schienkiewitz, A., Finger, J.D., Lange, C. and Lampert, T., 2019. Socioeconomic Inequalities in the Rise of Adult Obesity: A Time-Trend Analysis of National Examination Data from Germany, 1990-2011. *Obesity Facts*, *12*(3), pp.344-357.

Holsten, J.E., 2009. Obesity and the community food environment: a systematic review. *Public health nutrition*, *12*(3), pp.397-405.

Joint Formulary Committee, 2019. B*NF 78: September 2019-March 2020*. London: Pharmaceutical Press.

Kantar, 2021, Worldpanel grocery share data. [online] Available at: <https://uk.kantar.com/consumer/shoppers/2015/march-kantar-worldpanel-uk-grocery-share/> [Accessed 26 April 2021].

Kim, T.J. and von dem Knesebeck, O., 2018. Income and obesity: what is the direction of the relationship? A systematic review and meta-analysis. *BMJ open*, *8*(1), p.e019862.

Komlos, J., Smith, P.K. and Bogin, B., 2004. Obesity and the rate of time preference: is there a connection?. *Journal of biosocial science*, *36*(2), pp.209-219.

Leadsdeposit, 2021. (Probably) The Best Restaurant database on the internet [online] Available at: <https://leadsdeposit.com/restaurant-database/> [Accessed 26 April 2021].

Office for National Statistics. 2013. *2011 Census*. [online] Available at: <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/keystatisticsandquickstatisticsforlocalauthoritiesintheunitedkingdom/2013-12-04> [Accessed 29 April 2021].

Ohri-Vachaspati, P. and Leviton, L.C., 2010. Measuring food environments: a guide to available instruments. *American Journal of Health Promotion*, *24*(6), pp.410-426.

Pancrazi, R., van Rens, T. and Vukotic, M., 2020. How Distorted Food Prices Discourage a Healthy Diet. [online] Available at: <https://www.thijsvanrens.com/DFPDHD/>

Philipson, T. and Posner, R., 2008. *Is the obesity epidemic a public health problem? A decade of research on the economics of obesity* (No. w14010). National Bureau of Economic Research.

Präger, M., Kurz, C., Böhm, J., Laxy, M. and Maier, W., 2019. Using data from online geocoding services for the assessment of environmental obesogenic factors: a feasibility study. *International journal of health geographics*, *18*(1), pp.1-13.

Public Health England. 2017. *Health matters: obesity and the food environment*. [online] Available at: <https://www.gov.uk/government/publications/health-matters-obesity-and-the-food-environment/health-matters-obesity-and-the-food-environment--2> [Accessed 29 April 2021].

Rozin, P., Scott, S., Dingley, M., Urbanek, J.K., Jiang, H. and Kaltenbach, M., 2011. Nudge to nobesity I: Minor changes in accessibility decrease food intake. *Judgment and Decision Making*, *6*(4), pp.323-332.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Saelens, B.E. and Glanz, K., 2009. Work group I: Measures of the food and physical activity environment: instruments. *American journal of preventive medicine*, *36*(4), pp.S166-S170.

Scarborough, P., Bhatnagar, P., Wickramasinghe, K.K., Allender, S., Foster, C. and Rayner, M., 2011. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006–07 NHS costs. *Journal of public health*, *33*(4), pp.527-535. <https://doi.org/10.1093/pubmed/fdr033>

Schroeder, M.A., Lander, J. and Levine-Silverman, S., 1990. Diagnosing and dealing with multicollinearity. *Western journal of nursing research*, *12*(2), pp.175-187.

Tripadvisor, 2021. *Content API* [online] Available at: <https://developer-tripadvisor.com/content-api/> [Accessed 26 April 2021]. [Online]

Townshend, T. and A. Lake (2017). Obesogenic environments: current evidence of the built and food environments. Perspectives in Public Health 137(1), 38–44.

Variyam, J.N., Blaylock, J.R., Smallwood, D.M. and Basiotis, P.P., 1998. *USDA's Healthy Eating Index and Nutrition Information* (No. 33588). United States Department of Agriculture, Economic Research Service.

Volkow, N., Wang, G.J., Fowler, J.S., Tomasi, D. and Baler, R., 2011. Food and drug reward: overlapping circuits in human obesity and addiction. *Brain imaging in behavioral neuroscience*, pp.1-24.

Walker, R.E., Keane, C.R. and Burke, J.G., 2010. Disparities and access to healthy food in the United States: A review of food deserts literature. Health & place, 16(5), pp.876-884.

WHO. 2002. *Population nutrient intake goals for preventing diet-related chronic diseases*. [online] Available at: <https://www.who.int/nutrition/topics/5_population_nutrient/en/> [Accessed 26 April 2021].

WHO. 2020. Obesity and Overweight Fact Sheet [online] <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight> [Accessed 26 April 2021].

Zou, H. and Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, *37*(4), p.1733.