

How does allocation of funding to London local authorities affect obesity cases in 2018

CASA 0005 Quantitative Methods Coursework 1

November 15, 2020

1. Introduction

NHS claimed almost one in five Year 6 children in the UK was found to be obese in 2018, and sadly this number is not dropping during the past few years. The origins of childhood obesity stem from various aspects, including lifestyle, genetic and environmental issues. The government has been taking considerable forms of actions to tackle this problem, from sugar reduction to advertising and promotions. This study investigates how government has been allocating their funding, in particular, to local authorities in London.

2. Data

The data used in this study contains population, obesity cases, total budget, and allocation of funding for local authorities across London in 2018. An illustration of data employed is shown in Table (1).

	local authorities	total obesities	total population	obesity density	...	total budget	...
0	Barking and Dagenham	763	181779	0.420	...	139000	...
1	Barnet	773	355955	0.217	...	220000	...
⋮							

Table 1: Illustration of data used, list of column names include:

Names of local authority areas; total obesity cases in each area; total population; obesity density per 100 people (obesity cases divided by population multiplied by 100); total budget allocated (in pounds); percent of budget spent on improving air quality, cleaner environment, health training, raising school awareness, media awareness and subsidizing counselling.

Only one observation is considered an outlier due to its relatively small scale, City of

London data, therefore dropped from the dataset. Its population is below the average population in London boroughs by 97%, having this data in the linear regression plot of obesity cases vs. total budget spent lowers the regression coefficient from 0.437 to 0.349.

3. Methodology

Three approaches were taken to investigate the criteria of funding:

- i. A linear line was fitted using `scipy.stats.linregress()` between total budget spent and obesity density.
- ii. A more outlier-robust linear approach, called Random Sample Consensus (RANSAC) was used. This method compliments the ordinary least squares methods by adding detections of outliers and accord them to have no influence on the parameters of the model (Fischler and Bolles, 1981). In `sklearn.linear_model.RANSACRegressor()`, outliers are classified as those whose residual exceed the median absolute deviation of dependent variables (Pedregosa, F. et al., 2011).
- iii. Finally, a non-linear approach was taken. Polynomial regression of various degrees was fitted to the data using `numpy.polyfit()` in Python.

4. Results

- i. Logically one might assume that the more obesity cases discovered per unit population, the more funding would be allocated. The linear line fitted to total budget spent vs. obesity density yields a coefficient of determination of 0.002, which means that only 0.2% of the variance of dependent variable (total budget) is explained by independent variable (obesity density). Log transformations were also applied to the variables, associated parameters are shown in Table (2).

	unchanged data	$\log(x)$ and y	$\log(y)$ and x	$\log(x)$ and $\log(y)$
slope	3×10^{-9}	0.001	3×10^{-7}	0.0574
constant	-0.002	-0.005	0.47	-0.259
R^2	0.037	0.032	0.29	0.29
Pearson correlation coefficient	0.192	0.18	0.538	0.537
p value	0.29	0.32	0.001	0.002
relationship implied	linear	exponential	exponential	power

Table 2: coefficients related to the relationship between original and log-transformed total budget spent and obesity density

As the maximum value of R^2 being under 0.3 in Table (2), the amount of funding allocated to local authorities is extremely weakly correlated with the obesity cases per unit population by the power law, thus indicating that this relationship is meaningless.

ii. The lack of a meaningful linear relationship could be induced by number of outliers involved. The outliers in this case originate from changes in human behaviour, the government may identify some boroughs need more attention due to other reasons that cannot be demonstrated from statistical measures. Building from the results in i, log transformed data has been used to plot figure 1, where in **RANSAC**, a considerable proportion of data has been classified as outliers due to the randomness and large variance of human behaviours.

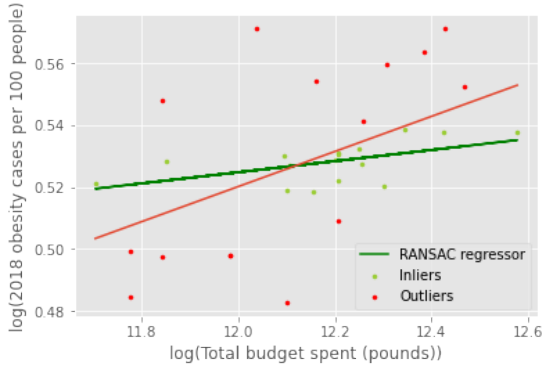


Figure 1: $\log(\text{total budget})$ vs. $\log(\text{obesity density})$ plot, linear relationship obtained from Ordinary Least Squares (red) and RANSAC (green). RANSAC is only fitting to inliers, while OLS is fitting to all data.

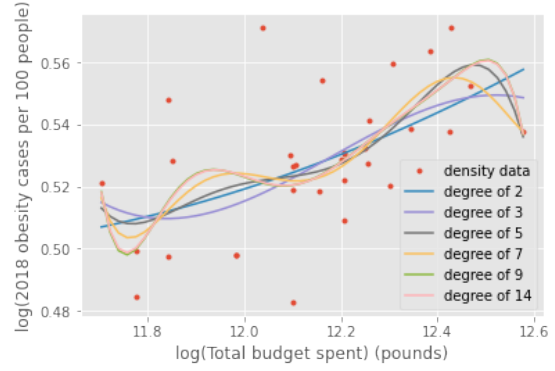


Figure 2: Fitting various degrees of polynomials to $\log(\text{total budget})$ vs. $\log(\text{obesity density})$ plot. Both logs refer to the natural log.

Although RANSAC increased R^2 from 0.29 to 0.39, its p value has surmounted past the acceptable range (0.05). Visually, the rise of R^2 can be explained by the excessive amount of outliers removed. Additionally, the classification of outliers is entirely statistical and without contextual reasons, it is difficult to examine whether the outliers it has removed could otherwise play an essential part on the overall trend. To summarise, the criteria of identifying outliers in RANSAC approach is stricter and less stable, thus making it a less reliable method.

iii. Selected degrees of 2 to 14 of polynomial regression are shown in Figure (2), and associated R^2 values are displayed in Table (3).

degrees of polynomial fitted	2	3	5	6	9	14
Coefficient of determination R^2	0.292	0.309	0.353	0.364	0.373	0.373

Table 3: coefficients related to the relationship between original and log-transformed total budget spent and obesity density

It can be spotted that as the degree of polynomial fitted increases, R^2 increases. One can deliberately dial the degree up and manipulate r squared to the range accepted, but this would not be a valid approach logically. If so, the model yielded would be highly influenced by the noise in the dataset, rather than demonstrating the data's underlying real trend. Furthermore, R^2 starts to converge when hitting degree of 6, and a warning message is generated when fitting above this, suggesting that `Polyfit()` may be poorly conditioned. The R^2 value that optimises the bias-variance trade-off is near 0.37, still not enough to indicate an appropriate polynomial relationship.

5. Conclusion

Logically, the obesity problem should be alleviated with more budget spent on tackling it. The drawback of this concept is the lack of a reference level, which arises from an existing annual increase of obesity density along with population expansion. The reference level of obesity density increase could be calculated from different budget spent in previous years and associated obesity densities. The alternative approach to fit a linear relationship by excluding outliers using RANSAC has unfortunately failed, due to its unstable and outlier-robust nature. Last but not least, there is no appropriate polynomial relationship that could conserve bias-variance trade-off.

One can conclude that there is no strong statistical relationship between total budget and obesity density. Another approach to examine the criteria of allocating funding is to consider the relationship of total budget with population and number of obesity cases respectively. Interestingly, there is strong power law relationship when fitting between the above variables. Additionally, no multiple linear relationship is found between areas of spending the budget and obesity density by plotting the correlation matrix. The above relationships have been omitted from this study for brevity. Further research could be conducted to investigate the reasons of such behaviours.

6. Appendix

Python code used to generate the figures and calculate the relevant data, as well as L^AT_EX source code used to generate this report can be found at: https://github.com/Tiana125/CASA0005_QM_CW1

References

Fischler, M. A. and Bolles, R. C. (1981). ‘Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography’. *Communications of the ACM*, 24 (6), pp. 381–395. doi: 10.1145/358669.358692.

NHS. (2018). National Child Measurement Programme, England 2018/19 School Year. Available at: <https://digital.nhs.uk/data-and-information/publications/statistical/national-child-measurement-programme/2018-19-school-year>.

Pedregosa, F., Varoquaux, G., Gramfort, A. and Michel, V. (2011). ‘Scikit-learn: Machine Learning in Python’. *Journal of Machine Learning Research*, (12), pp. 2825–2830.