Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 09, 2024

Abstract

This report provides an analysis of data related to health, age, socio-economic factors and lifestyle habits in adults (from the age of 16) from the population in England, derived from the Health Survey for England 2019.

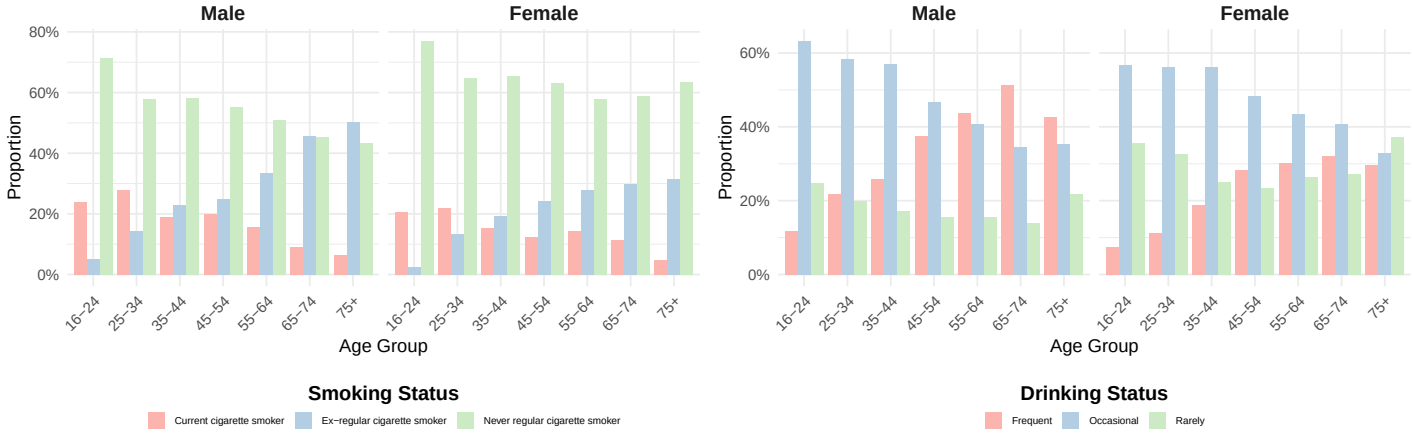
Summary (Non-Technical)

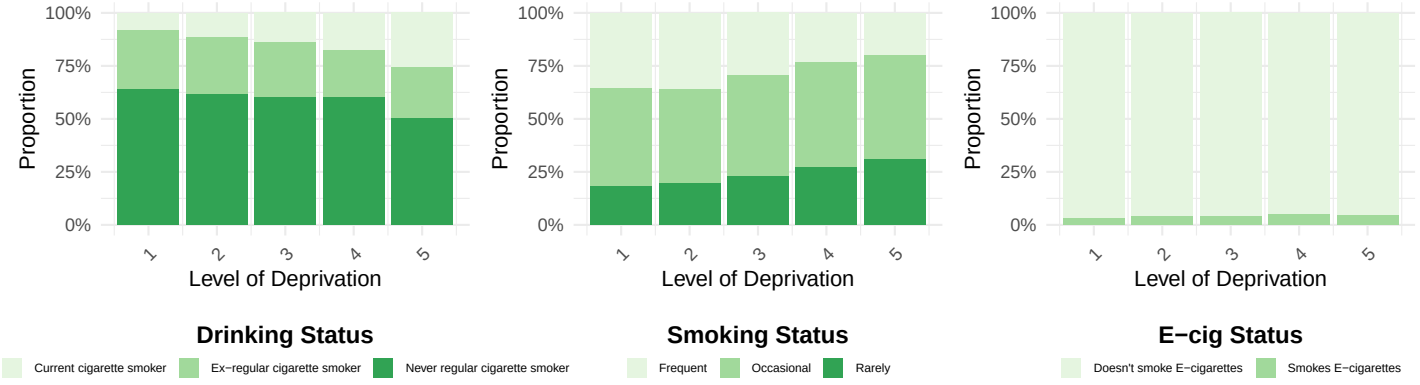
Introduction

In the UK, smoking, vaping, and alcohol consumption are widespread, particularly among youngsters. It is crucial to be aware of the consequences and dangers of these habits, and the complications they can cause in later-life. The Office for National Statistics (ONS), regarded as the foremost statistical institute in the UK, is the "go-to" for insights into public health trends. According to their estimations, approximately 14.1% of the adult population aged 18 and above were identified as cigarette smokers (ONS 2019a). Furthermore, their data revealed there were 7,565 deaths attributed to alcohol-specific causes in 2019 (ONS 2019b). These statistics alone warrant a need for an understanding of health-related behaviours and the outcome. Therefore, the aim of our study is to investigate not only the prevalence but also the severity of these habits, exploring different socioeconomic factors that may potentially contribute to each. Additionally, we will investigate the greater implications of these bad habits and how they relate to systolic blood pressure levels throughout the UK population.

Exploratory Analysis

We start by investigating the data





Methology

Define all variables used in analysis

Analysis

What is the prevelance of drinking, smoking and E-cig usage?

To calculate the prevalence of each habit we assume each of the n observations, $x_1, \ldots, x_n$, to be independent, identically distributed (iid) random variables (RVs) where $x_i \sim Bern(p) \, \forall i = 1, \ldots, n$ and p denotes the probability of an observation having the relevant habit. We use the household weights to calculate a weighted Maximum Likelihood Estimate (MLE) of p. That is, letting w_i denote the weight of the i^{th} observation, we alter the standard likelihood function of a Bernoulli distribution as below:

$$L(p|\mathbf{x}) = \prod_{i=1}^{n} (p^{x_i}(1-p)^{1-x_i})^{w_i}$$

From this, we calculate our weighted MLE as:

$$\widehat{p} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i}$$

It can also be shown that this MLE has variance given by $\text{var}(\widehat{p}) = \frac{p(1-p)}{\sum_{i=1}^{n} w_i}$, which we can estimate using $\widehat{p}$ and use large sample properties of the MLE to get a normal approximation and estimate 95% confidence intervals for each habit, which are shown in the table below.

Table 1: Estimates and 95% Confidence Intervals for % of Population

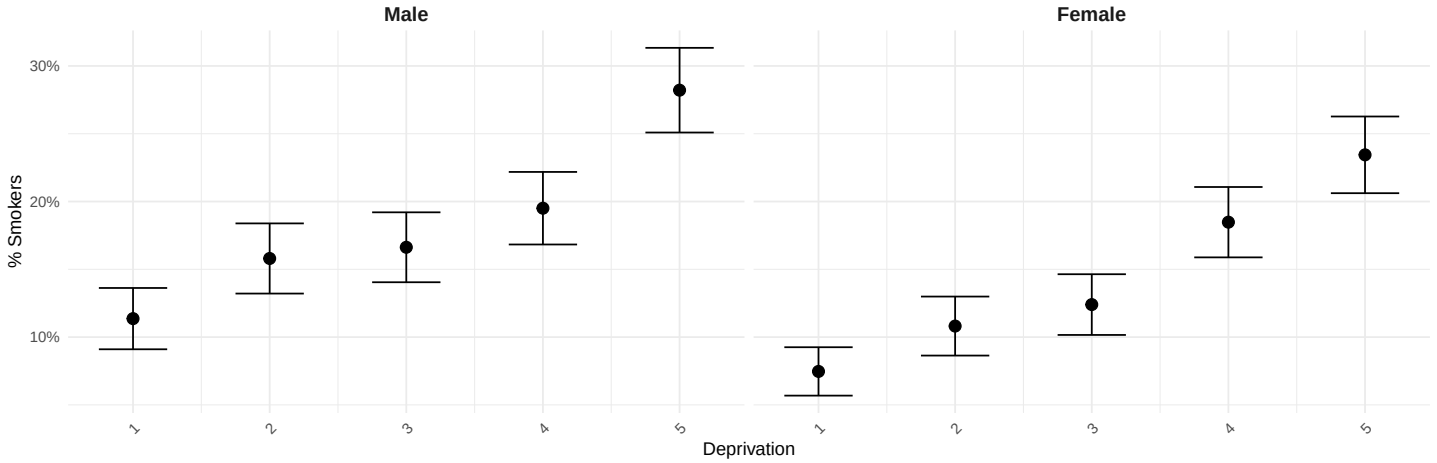
Habit	Estimate	C.I.
Drinking	80.3%	(79.5%, 81.2%)
Smoking	16.5%	(15.7%, 17.3%)
Smoking E-cigarettes	4.28%	(3.84%, 4.72%)

Interpretation of table/results - e-cig is much lower, drinking very common etc

Note that for the purposes of model building, we coded CASI/CAPI responses about alcohol and cigarette consumption in the following way:

Variable	Code	Label	Decode
NDPNow_19	1	E-cigarettes or vaping devices only	Smokes E-cigarettes
	2	Other nicotine delivery products only	Doesn't smoke E-cigarettes
	3	Both	Smokes E-cigarettes
	4	None	Doesn't smoke E-cigarettes
	-1	Not Applicable	NA
	-8	Don't know	NA
	-9	Refused	NA
dnoft_19	1	Almost every day	Frequently
	2	Five or six days a week	Frequently
	3	Three or four days a week	Frequently
	4	Once or twice a week	Occasionally
	5	Once or twice a month	Occasionally
	6	Once every couple of months	Rarely
	7	Once or twice a year	Rarely
	8	Not at all in the last 12 months	Rarely
	-1	Not Applicable	NA
	-8	Don't know	NA
	-9	Refused	NA

We can also...



How is smoking associated with socioeconomic factors and age?

Explanation of why we split into test/train dataset.

This leaves us with 6563 observations in our training dataset. The number of missing observations for each are shown:

Table 3: Missing values in the training dataset

Variable	Missing Values	% Missing
omsysval	3235	49.3%
BMIVal	1201	18.3%
dnoft_19	1180	18%
cigdyal_19	39	0.594%
cigsta3_19	38	0.579%
d7many3_19	37	0.564%
drinkYN_19	36	0.549%
NDPNow_19	35	0.533%
topqual2	28	0.427%
origin2	19	0.29%
marstatD	1	0.0152%

Explanation of first model - binomial with logit link

Table 4: Comparison of selected model evaluations

Linear Predictor	Train AIC	Test AUC	Train RMSE	Test RMSE	Test Accuracy
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i + q_i : u_i$	4959.9	0.751	0.341	0.329	0.858
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i$	4961.4	0.752	0.341	0.329	0.859
$\text{logit}(\mu_i) \sim a_i + m_i + q_i + u_i + o_i + t_i + s_i$	5021.8	0.739	0.343	0.330	0.857
$\text{logit}(\mu_i) \sim a_i^{(5)} + m_i + q_i + u_i + o_i + t_i + s_i$	4970.9	0.750	0.340	0.329	0.859
$\text{logit}(\mu_i) \sim a_i^{(10)} + m_i + q_i + u_i + o_i + t_i + s_i$	4987.5	0.752	0.341	0.329	0.858

Explain why we selected the model we did - lowest AIC

Some model diagnostics

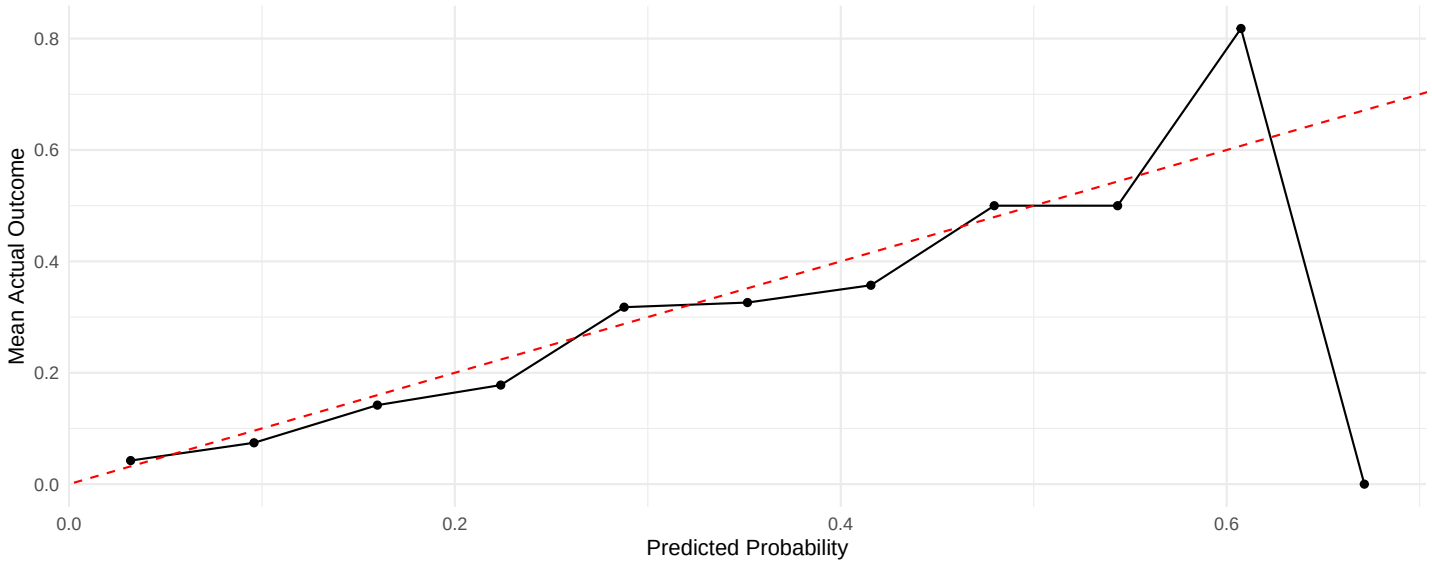


Figure 1: Calibration chart for binomial model

Interpretation of model coefficients etc

Limitations of model

Which lifestyle habits are associated with systolic blood pressure?

Question Three

Results/Conclusion

NatCen Social Research and Health (2019)

References

NatCen Social Research, Department of Epidemiology, University College London, and Public Health. 2019. "Health Survey for England." http://doi.org/10.5255/UKDA-SN-8860-1.

ONS. 2019a. "Adult smoking habits in the UK: 2019." https://shorturl.at/qQW27.

———. 2019b. "Alcohol-specific deaths in the UK: registered in 2019." https://shorturl.at/gqxY1.