Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 14, 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. Please note, no generative AI technology was utilized in the creation of this report or the subsequent data analysis presented herein. All findings and conclusions are derived from human-driven methodologies.

Summary

In England, smoking, "vaping" and alcohol consumption are widespread habits among adults, particularly younger adults. In fact, it is estimated that around 13.9% of adults in England identified as cigarette smokers (ONS 2019a), and a similar study found that there were 10.9 alcohol-related deaths per 100,000 people living in England in 2019 (ONS 2019c). Thus, it is crucial to be aware of the predictors and consequences of these habits, which is the motivation for this analysis.

We conducted this analysis on a subset containing 8,204 adults living in England (ages 16+), who were interviewed in their homes about their demographics and their smoking and drinking habits as part of the Health Survey for England (HSE) 2019. 4,947 of these participants who consented to an at-home nurse visit had vitals such as blood pressure, weight and height measured. We investigated which features of a participant are associated with smoking habits, and which habits are associated with high blood pressure.

This is where our non-technical summary of results should go

Introduction

The mechanisms which drive lifestyle habits are part of a complex and ever-changing field in behavioural psychology. What is known, however, is that such habits are driven by cues and cravings [Anastasia Droungas (1995)](Kambouropoulos 2009). The HSE 2019 captures demographic and socioeconomic characteristics that have the potential to make such cues more visible and cravings harder to resist. For example, living in a region where smoking is more prevalent or not having a husband/wife for an accountability partner to help you quit or resist these habits (ONS 2019b).

In this report, we will use weighting variables to estimate prevalence of cigarette, e-cig, and alcohol users in our population, and compare it with the larger population of adults in England. We will then attempt to identify associations with a participants current smoking status, explore whether predictive modelling can be used to identify smokers with a high accuracy, and investigate which lifestyle habits are associated with increased values of systolic blood pressure. *This will be crucial in enabling healthcare providers to implement preventative care for people at risk of developing hypertension.*

Methodology

Exploratory Analysis

The full HSE 2019 cross-sectional data set contains 10,299 observations across thousands of variables (NatCen Social Research and Health 2019), but we only studied patients over the age of 16 among key variables. Our subset includes 8,204 participants and 19 variables, which are described in Table 1.

Table 1: Summary of the analysis variables used in our report. (D) indicates that a variable was derived.

Variable	Definition	Notation
SerialA	Respondent serial number	NA
wt_int	Weight of observation	NA
Sex	Respondent sex (M/F)	s
(D) Age35g	Respondent age grouped in 5 year bands for 16+	$a^{(5)}$
(D) ag16g10	Respondent age grouped in 10 year bands for 16+	$a^{(10)}$
(D) topqual2	Highest level qualification	t
(D) marstatD	Marital status	m
(D) qimd19	IMD score (a measure of deprivation)	q
(D) urban14b	Level of rurality	u
(D) origin2	Grouped ethnic category	o
(D) cigdyal_19	# of cigarettes smoked per day	n^{cig}
(D) cigsta3_19	Smoking status (Reg / Ex-Reg / Never-Reg)	cig
(D) NDPNow_19	Use of E-cigarettes and or NDPs	ndp
(D) DrinkYN_19	Alcohol consumed in the last 12 months	alc
dnoft_19	Frequency of alcohol consumed in the last 12 months (grouped)	NA
(D) d7many3_19	# of days alcohol was consumed in the last week	n^{alc}
(D) GOR1	Government region office number	g
(D) BMIval	Valid BMI measurement (weight estimated if 130kg)	bmi
(D) omsysval	Valid mean systolic blood pressure	sbp

Duplicates

We found 36 pairs of observations with exactly equal variables (excluding ID variables and lab measurements), but we did not remove these from our analysis because the supporting documentation didn't state a protocol for repeated visits. We assumed these were genuine observations from different participants and thus included them in our dataset. However, there were 3 pairs of observations that had duplicate lab variables also. We believed that these were true duplicates and removed one observation from each of the three pairs resulting in a dataset of 8201 observations.

Table 2: Missing values in the training dataset

Variable	Missing Values	% Missing
omsysval	4036	61.5%
BMIVal	1519	23.2%
dnoft_19	1496	22.8%
cigdyal_19	57	0.869%
cigsta3_19	56	0.854%
NDPNow_19	53	0.808%
d7many3_19	52	0.793%
drinkYN_19	51	0.777%
topqual2	46	0.701%
origin2	29	0.442%
marstatD	1	0.0152%

Dichotomy of Factors

All variables were originally coded as numeric in our dataset, so we recoded these to factor variables accordingly. Some of the variables used for analysis were dichotomised by us for easier interpretation. We coded a binary variable indicating current smoking status, which will serve as our *primary* outcome variable. Also, we dichotomised urbanity into two levels being 'Rural' and 'Urban', and finally marital status into 'Married' and 'Not Married'. We group respondents into 'Higher Education', 'Further Education', 'A-Level equiv.', 'GCSE equiv.', 'No qualification' and 'Foreign/other'.

Missingness

The number of missing observations for each are shown in Table 2.

There is significant missingness in the lab values, which can be explained by the fact additional consent was needed from the participant to allow a nurse visit. One variable with a substantial amount of data missing was the frequency of drinking in the last 12 months, which due to the retrospective and sensitive nature of the question could be explained by either recall bias or a participant's refusal to answer. As a result, we decided to remove this variable from analysis in favour of a binary variable which simply states whether the respondent has drunk alcohol in the last 12 months. Education level, marital status and ethnicity are the only socioeconomic variables with any missing entries, with 46 observations (0.701% of the data) missing at least one of the three. These are not necessary for identifiability and as they are sensitive, we do not expect every participant to answer these questions. Therefore, we did not remove these observations.

Outliers and Data Issues

We attempted to determine whether the two lab measurement variables contained potential outliers, and as can be seen in Figure @ref(tab:output distribution plots), there were many outliers for BMI. We note that this used self-estimated weight from the participant in the calculation when it exceeded 130kg. Additionally, weight measurements were taken in inconsistent environments such as on different flooring in the participants' homes which will impact measurement accuracy. We believed this could have explained the extremely high BMI values in the range of 54.82 to 73.49. We coded such values as missing in the analysis dataset, but kept the observation.

The readings for systolic blood pressure were collected by taking an average of three readings, each five minutes apart, performed by a trained nurse who had to declare each reading to be valid. For this reason, it seemed highly unlikely that these readings were mistakes, and so we included them in our analysis. The measurement device used has been validated for use in this environment (Yechiam Ostchega PhD 2009). We did, however, note that values of systolic blood pressure that were this consistently high (>140mmHg) were indicators for hypertensive crisis, and so a part of our population may have serious underlying health conditions that could affect the generalisability of our findings (Association 2023).

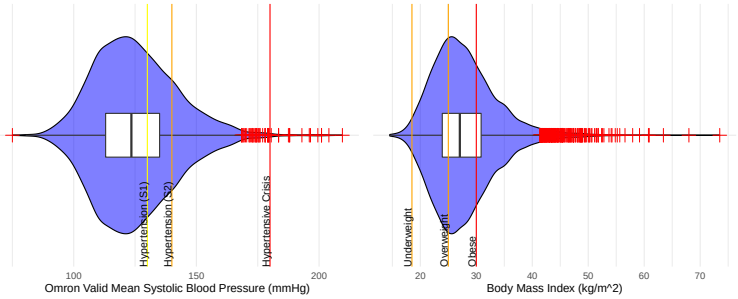

Figure 1: Distribution of BMI and Mean Systolic Blood Pressure

Analysis

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

To calculate the prevalence of these three habits we assumed 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 a habit being present for the given observation.

We used the household-level weighting variable to calculate a weighted Maximum Likelihood Estimate (MLE) of p. That is, letting w_i denote the weight of the i^{th} observation, we altered 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 calculated 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 the MLE has variance given by $\text{var}(\widehat{p}) = \frac{p(1-p)}{\sum_{i=1}^{n} w_i}$, which we estimated using $\widehat{p}$. We used large sample properties of the MLE to get a normal approximation

Table 3: Estimates and 95% Confidence Intervals for % of Population

Habit	Estimate	C.I.
Drinking	80.4%	(79.5%, 81.2%)
Smoking	16.5%	(15.7%, 17.3%)
Smoking E-cigarettes	4.28%	(3.84%, 4.72%)

and estimated 95% confidence intervals for each habit, which are shown in Table 3.

We found that the usage of e-cigarette usage among adults is relatively low, making it challenging to dissect any significant trends within the data. *Maybe a wee bit more here*

How is smoking associated with socioeconomic factors and age?

Next, we worked to uncover factors that have possible associations with smoking prevalence. We started with the demographic factors of age and gender, plotting the smoking prevalence across combinations of these groups to identify any patterns. Figure 3 suggests a negative correlation between age and the proportion of current cigarette smokers, across both genders. Interestingly, the proportion of males who quit smoking in later-life was much greater than that of females (75yrs+; M: 50%, F: 31%).

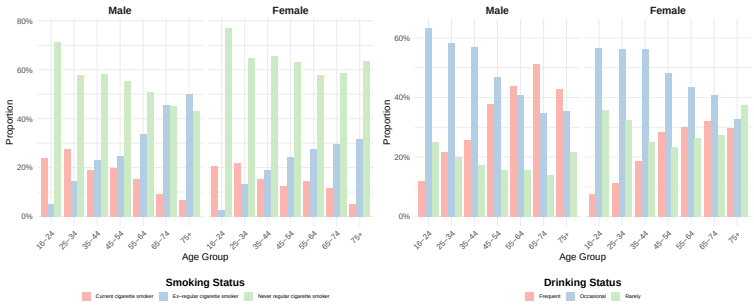

Figure 2: Smoking and drinking status proportions by age group and gender

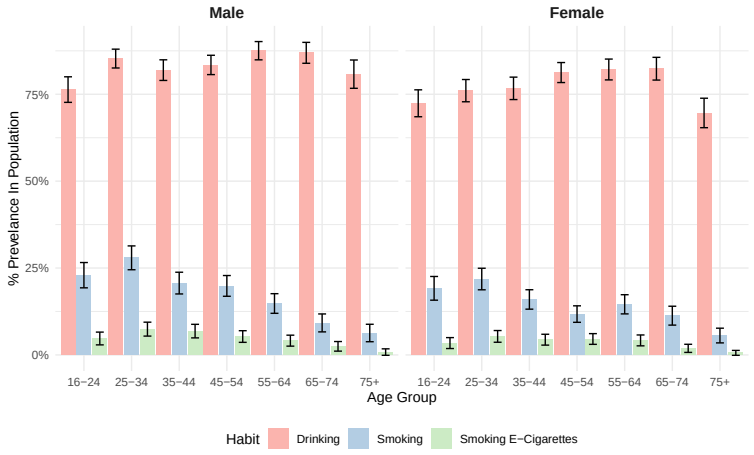

Figure 3: Estimation of the prevelance of smokers by deprivation

Table 4: Comparison of the characteristics of the training set compared with the test set.

	Proportion (%)	
Variable:Label	**Test**	**Train**
Sex:Male	46.6	44.8
Sex:Female	53.4	55.2
topqual2:No qualification	17.4	18.0
topqual2:GCSE equiv.	20.7	21.2
topqual2:A-Level equiv.	13.5	13.8
topqual2:Further Education	29.7	29.3
topqual2:Higher Education	17.6	16.4
topqual2:Foreign/Other	1.1	1.1
marstatD:Married	54.5	52.5
marstatD:Not Married	45.5	47.5
urban14b:Urban	82.2	81.0
urban14b:Not Urban	17.8	19.0
origin2:White	85.0	86.0
origin2:Black	2.6	2.8
origin2:Asian	10.1	8.5
origin2:Multiple	1.9	1.6
origin2:Other	0.5	1.0
	Mean	**Mean**
Age(Estimated)		

Additionally, males demonstrate a higher prevalence of drinking across nearly all demographic categories in comparison to females. We can clearly see that smoking prevalence increases with our deprivation variable, meaning that smokers tend to be from lower levels of deprivation. This correlation may be caused by the hefty tobacco duties the UK impose on its' residents (GOV.UK 2014). Less deprived individuals they can take-up more unnecessary, expensive habits, with smoking being one of the main candidates.

We reserved 80% of our data to train the model, and used the remaining 20% to test the model, which was made possible due to the large size of our dataset. The training set contained 6561 observations and the test set contained 1640. *Table 4 summarises the key variables between the test and train dataset to illustrate that both are representative of the whole dataset.* This reduced the risk of overfitting and allowed us to test the predictive power of each model we proposed.

To develop a predictive model for identifying smokers, we used the binary variable for current smoking status as a response, with various socioeconomic and demographic factors as predictors. These were selected based on the associations suggested by our prior analysis.

Under the assumption that the age of participants were approximately uniformly distributed within their respective age bands, we defined the estimated age of the i^{th} observation as the midpoint of the participants respective 5-year age

Table 5: Comparison of selected model evaluations, wherein $\eta_i = \text{logit}(\mu_i)$, and Acc. refers to the model accuracy on test data.

Model	AIC	AUC	RMSE		Acc.
			Train	Test	
$\eta_i \sim a_i + a_i^2 + ms_i + q_i + u_i + o_i + t_i + s_i + q_i : u_i$	4987.5	0.759	0.342	0.324	0.861
$\eta_i \sim a_i + a_i^2 + ms_i + q_i + u_i + o_i + t_i + s_i$	4987.3	0.759	0.342	0.324	0.861
$\eta_i \sim a_i + ms_i + q_i + u_i + o_i + t_i + s_i$	5036.4	0.740	0.343	0.327	0.861
$\eta_i \sim a_i^{(5)} + ms_i + q_i + u_i + o_i + t_i + s_i$	4994.3	0.759	0.341	0.325	0.861
$\eta_i \sim a_i^{(10)} + ms_i + q_i + u_i + o_i + t_i + s_i$	5005.9	0.753	0.342	0.324	0.860

group (taking the estimation of the 90+ category as 92.5), denoted a_i. We found that age may represent a somewhat quadratic effect on the probability of smoking, leading us to include a a_i^2 term in our model. A comparison of these models are summarised in Table 5.

We selected the model based on AIC, which enabled us to balance model complexity and model fit (quantified by likelihood), which further reduced the risk of overfitting. This model is:

$$\text{logit}(\mu_i) \sim \beta_0 + \beta_1 a + \beta_2 a^2 + \beta_3 q + \alpha_j^{ms} + \alpha_k^u + \alpha_l^o + \alpha_m^t + \alpha_n^s$$

where $j \in$ {Married, Not Married}, $k \in$ {Urban, Not Urban}, $l \in$ {White, Asian, Black, Multiple, Other}, $n \in$ {Male, Female} and $m \in$ {Higher Education, A-Level equiv., Further Education, GCSE equiv., Foreign/Other, No qualification}.

To evaluate the predictive performance of our final model using our test data, Figure 4 demonstrates the predicted probability of each 'probability bin' against the mean actual outcomes. As we can see the calibration curve closely follows the line $y = x$, which is indicative of a well-calibrated model.

This model doesn't predict high values... answer why

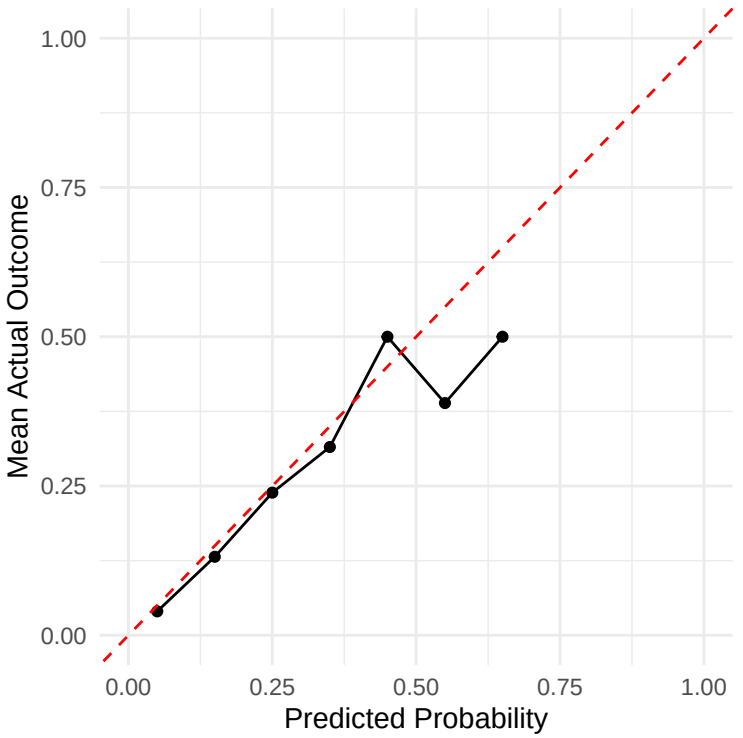

Figure 4: Calibration chart for Binomial model

Which lifestyle habits are associated with systolic blood pressure?

We chose to model the systolic blood pressure (BP) using the Inverse-Normal distribution with a $1/\mu_i^2$ link function. As we noted before this variable has a lot of outliers, is right-skewed and strictly positive, making this distribution a good choice. We also tested a Log-Normal distribution, a Normal distribution with identity link and a Gamma distribution with an inverse link but found that the Inverse-Normal distribution was able to account for the higher values of BP the best when we compared the Q-Q plots. The data we used to fit this model was the same training data set as the Binomial model. We filtered out *NA* values of BMI and BP as we only want to consider respondents that had lab measurements taken, leaving us with 2960 observations.

We studied the relationship between both age and BMI with BP, shown in Figure 5. There was a quadratic relationship between BMI and BP so we decided to include a bmi^2 term in our model. To account for possible dependencies between variables in the data, we tried fitted models with a combination of different interaction terms and found three that were likely significant. These were an interaction between BMI & Age (Medicine 2017), BMI & gender (K. et al. 2019) and Age & Sex (M. et al. 2019), which are all existing areas of research.

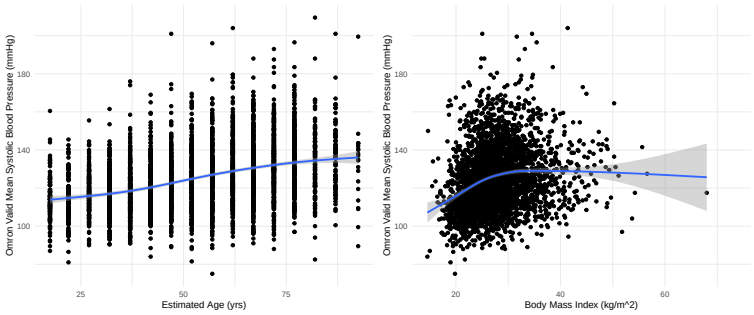

Figure 5: Relationship of BMI and Age with Mean Systolic Blood Pressure

To determine the predictors in the final model, we used a stepwise approach starting with the full model and eliminating variables based on AIC. The final model selected was

$$MODEL$$

The Q-Q plot in Figure 6 shows a good fit to the distribution and, despite some deviation at both ends of the scale, the model can capture the outliers more accurately than other distributions we tried. On our testing dataset we achieve a RMSE of 14.3mm Hg, which is relatively small considering the spread of the variable.

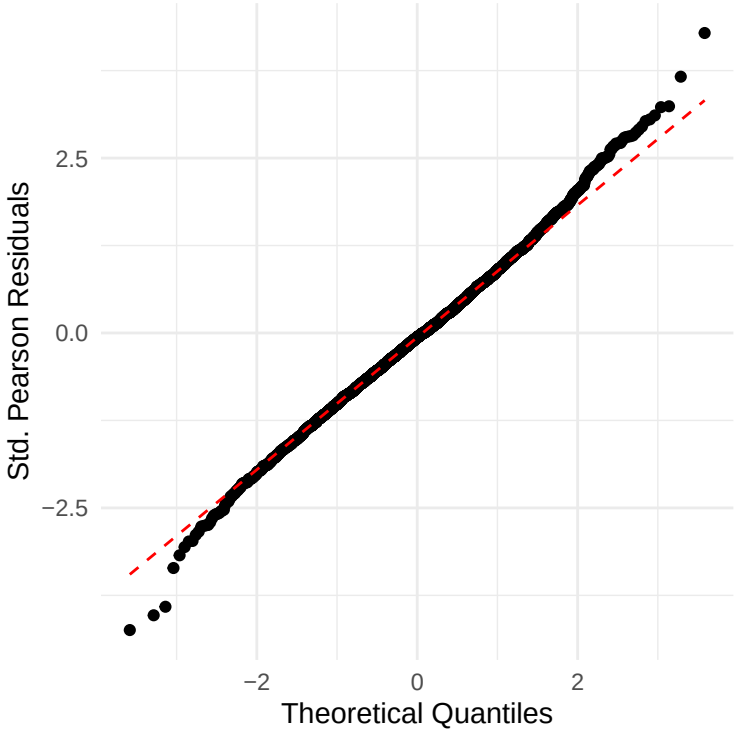

Figure 6: Q-Q plot of Inverse-Normal model residuals

We found that, on average, young females have lower blood pressure than young men but as they age, their blood pressure increases at a faster rate. Controlling for other lifestyle choices included in the model, a female's blood pressure can be expected to exceed an equivalent male's at around 72 years old. Figure 7 shows this relationship. One of the biggest influences on BP was smoking status, and we found smoking was associated with higher blood pressure for both males and females, also shown in Figure 8. This effect is largest for males and a typical male can expect a 1.69% increase in BP by taking up smoking.

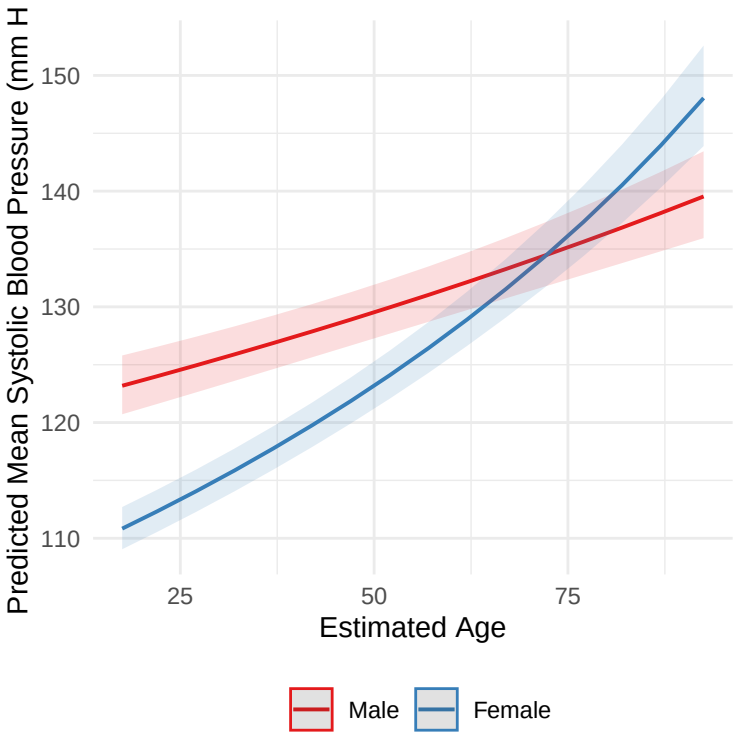

Figure 7: Marginal effects of Age on Systolic Blood Pressure

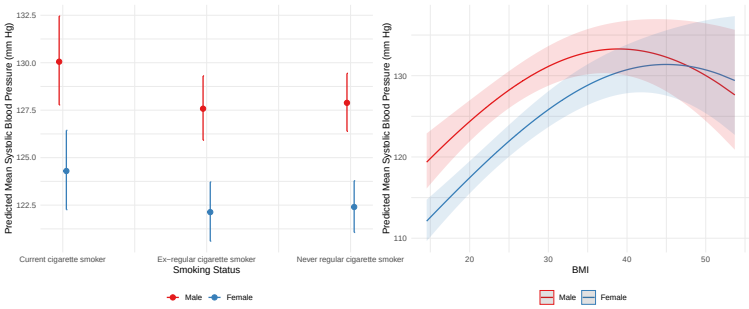

Figure 8: Marginal effects of Smoking Status and BMI on Systolic Blood Pressure

Drinking was also related to higher BP, and we observe a linear relationship in Figure 9 with the number of days that alcohol was consumed in the previous week. On average, one extra day of drinking led to an increase of 0.74‘ mm Hg.

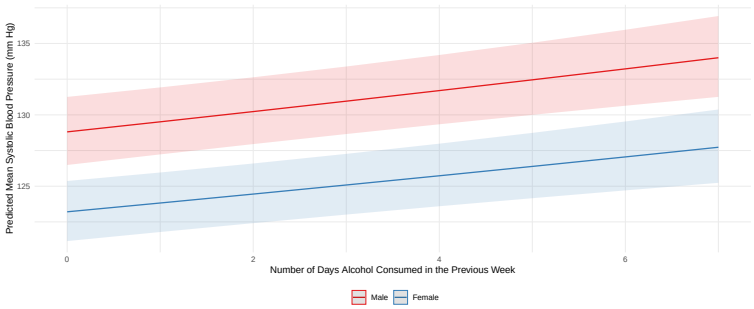

Figure 9: Marginal effects of alcohol consumption on Systolic Blood Pressure

Results/Conclusion

Our analysis showed that alcohol consumption is extremely common among UK adults, with approximately 80.4% being consumers, while smoking and vaping rates are lower at 16.5% and 4.3%, respectively. Older males show the highest tendencies for alcohol consumption, with 'frequent' drinking the most prevalent among males aged 65-74. Unlike 'frequent' alcohol consumption, the prevalence of smoking decreases with age. Individuals aged 16-24 appear to be the worst offenders when it comes to smoking cigarettes, indicating a significant issue within youth culture. Moreover, we found an interesting relationship between deprivation levels and smoking and drinking behaviours. Smoking prevalence seemed to increase as deprivation levels decrease, while alcohol consumption appeared to do the opposite.

To help us uncover the underlying socio-economic factors that may drive the prevalence of smoking, we first looked at the most 'comparable' respondent type in our dataset. This 'respondent' is a white, single male who has no qualifications and lives in an urban area. We found that the probability of our hypothetical respondent being a smoker is approximately 25% *Where is this from?* (notably higher than the population average of 16.5). *Maybe worth talking about confidence intervals here - is value outside 95%* The only socio-economic variables to certainly increase this probability is the deprivation level our respondent falls under. The rest of the socio-economic variables we have access to, appear to decrease the likelihood of smoking. For example, if our respondent is any of the following: Asian, black, married, or went on to further education, the chances of them being a smoker is at-least halved.

References

al., Katsuya et. 2019. "The Investigation of Sex Differences in the Effect of Body Mass Index." https://www.hindawi.com/journals/ijhy/2019/1360328/.

al., Martins et. 2019. "The effect of gender on age-related blood pressure changes and the prevalence of isolated systolic hypertension among older adults." https://pubmed.ncbi.nlm.nih.gov/11605350/.

Anastasia Droungas, Anna Rose Childress, Ronald N. Ehrman. 1995. "Effect of smoking cues and cigarette availability on craving and smoking behavior." https://www.sciencedirect.com/science/article/pii/030646039500029C.

Association, American Heart. 2023. "Understanding Blood Pressure Readings." https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

GOV.UK. 2014. "Tax on shopping and services." https://www.gov.uk/tax-on-shopping/alcohol-tobacco.

Kambouropoulos, Nicolas. 2009. "'Cue reward salience' predicts craving in response to alcohol cues." https://www.sciencedirect.com/science/article/abs/pii/S0191886908003127.

Medicine, National Library of. 2017. "Weight and Body Mass Index in Old Age: Do They Still Matter?" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5704942/.

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. "Adult smoking habits in the UK: 2019." https://shorturl.at/itPQU.

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

Yechiam Ostchega PhD, Tatiana Nwankwo MS, RN. 2009. "Assessing the Validity of the Omron HEM-907XL Oscillometric Blood Pressure Measurement Device in a National Survey Environment." https://onlinelibrary.wiley.com/doi/full/10.1111/j.1751-7176.2009.00199.x.