Analysis of Health Survey for England (HSE) 2019

Candidate Numbers Here

March 12, 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 are using data from the 2019 Health Survey for England (HSE) to conduct our analysis. Before proceeding with any investigations involving the data, we checked for any common errors in large datasets. Our initial step involved pre-processing the data, which entailed filtering observations to include only those above the age of 16 and removing 3 pairs of observations that were identical in every descriptive variable except for the ID, resulting in a dataset of `r` observations. As we have two different grouped age variables, we ensured they were consistent with one another. *Don't know if this final sentence needs to be included*

Need the table copied over here with the variable names/a paragraph explaining these

We also found 36 pairs with exact entries excluding lab measurements, but we do not remove these from our analysis. We suspect these may be individuals who have either initially refused a nurse visit but later changed their mind, or simply have had two nurse visits where their lab measurements vary slightly between the two. The supporting documentation doesn't state a protocol for repeated visits, so we will assume these are genuine observations from different participants and include them. *Maybe rewording this?*

I think this table could be replaced with a paragraph that may save space All variables are coded as numeric in our dataset, so we recoded all factor variables accordingly. 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 code a binary variable to group current smoking status into 'yes' or 'no', with the 'Ex-Reg' smokers falling into the 'no' category. For 'topqual2' we group respondents into 'higher education' or 'basic education' *This has changed since*. Similarly, 'urban14b' now has two levels being 'rural' and 'urban' and 'marstatD' now only has 'married' and 'not married' levels.

The number of missing observations for each are shown in Table @ref(tab:output NA table):

Table 2: Missing values in the training dataset

Variable	Missing Values	% Missing
omsysval	2972	55.7%
dnoft__19	1097	20.6%
topqual2	19	0.356%
cigsta3__19	16	0.3%
cigdyal__19	15	0.281%
NDPNow__19	14	0.262%
d7many3__19	14	0.262%
drinkYN__19	13	0.244%
origin2	5	0.0937%
marstatD	1	0.0187%

As expected, there is a lot of missingness in the lab values, particularly 'omsysval'. *Maybe some explanation of why this is expected?* When conducting our analysis of these, we will use a subset including only observations with at least one of these taken. The only other variable with a substantial amount of data missing is 'dnoft__19' and, as this is derived from a retrospective question, this may be a case of recall bias. Questions involving lifestyle habits such as alcohol consumption are considered sensitive, with many individuals potentially not being willing to answer, more likely to be the issue here. With this level of missingness in 'dnoft__19', we decide not to use it in any of our models. 'topqual2', 'origin2' and 'marstatD' are the only demographic variables with any missing entries, with 19 observations (0.356% of the data) missing at least one. As all three of these variables are not necessary for identifiability, we do not expect every participant to answer these questions due to personal preference.

Finally, we analysed the two lab measurement variables in Figure @ref(tab:output distribution plots), finding potential outliers to be present in both. The readings for 'omsysval' are collected by taking an average of three readings, each five minutes apart, performed by a trained nurse who deemed each reading to be valid. For this reason, it seems highly unlikely that these readings are mistakes, and we include them in our analysis. However, our BMI variable is estimated if the participants weight is over 130kg. These extremely high BMI values are most likely to be caused by the over-estimation of these participants weights. We deem these to be errors and code them as missing in the dataset.

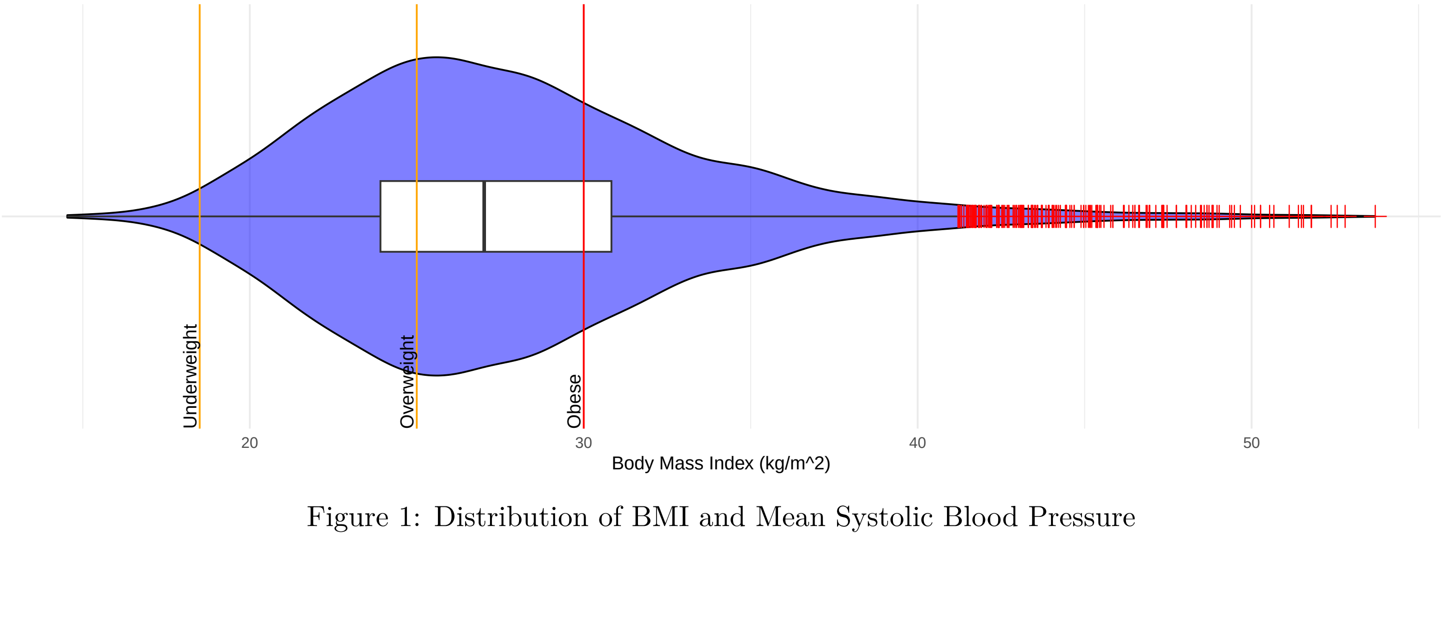

Figure 1: Distribution of BMI and Mean Systolic Blood Pressure

Methology

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, $x_1, \dots, x_n$, to be independent, identically distributed (iid) random variables (RVs) where $x_i \sim Bern(p) \, \forall i = 1, \dots, n$ 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 Table @ref(tab:output estimates table).

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

Habit	Estimate	C.I.
Drinking	81.8%	(80.9%, 82.7%)
Smoking	16.7%	(15.8%, 17.6%)
Smoking E-cigarettes	4.4%	(3.91%, 4.89%)

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

We segmented our data by gender, age brackets and level of deprivation and regrouped the levels of each lifestyle factor variables into more manageable buckets *This bit should read better, but I'm not sure the changes - just needs to be a sentence to introduce the plots*

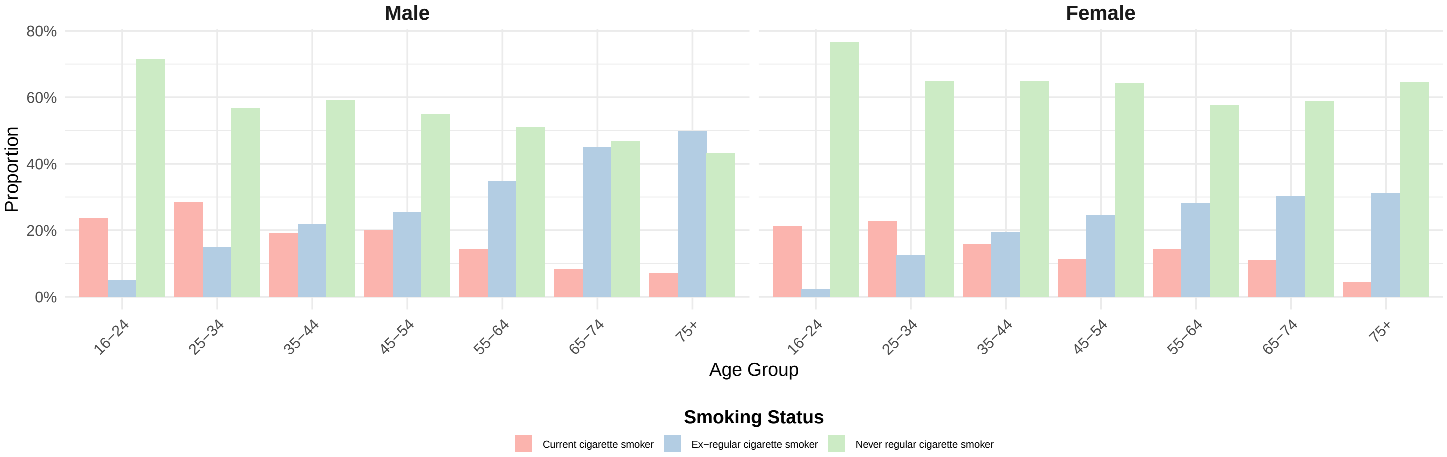

Figure 2: Smoking and drinking status proportions by age group and gender

I will change this plot to include all three (smoking, drinking, ecigs) in one graph, similarly needs a sentence intro

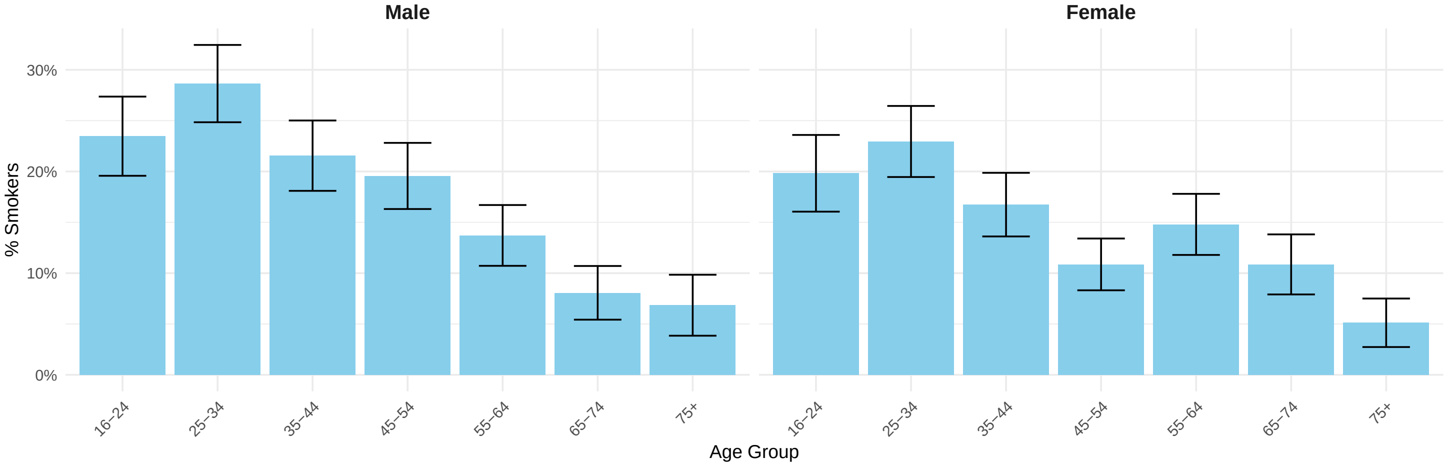

Figure 3: Estimation of the prevelance of smokers by deprivation

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 reword?* We noted an inverse correlation between age and the proportion of 'occasional' drinking, while 'frequent' drinking showed a positive correlation. Additionally, males demonstrate a higher prevalence of drinking across nearly all demographic categories in comparison to females. 'Active' smoking appears to follow a similar relationship to 'occasional' drinkers, with the number of 'active' smokers decreasing with age. Interestingly, the proportion of males who quit smoking in later-life is consistently greater than that of females.

This probably needs rewording It is important to mention the variable 'Level of Deprivation' ranges from 1 (meaning most deprived) to 5 (least deprived). 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.

How is smoking associated with socioeconomic factors and age?

Before attempting to fit any suitable models, we split the data into a 80/20 test train data. This reduces the risk of overfitting and allows us to test the predictive power of each model.

This leaves us with 5335 observations in our training dataset. *Summary of NA of this? maybe just important variables, i.e. smoking binary*

We use the binary smoker/non-smoker variable as a response and fit… *Finish this paragraph!*

Reword intro here This suggests that modelling age as continuous may be more effective. To achieve this, we defined the estimated age of the i^{th} observation as the midpoint of their respective 5-year age bracket (taking the estimation of the 90+ category as 92.5), denoted a_i. We believe 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 @ref(tab:output model selection table).

Table 4: Comparison of selected model evaluations

Linear Predictor	Train AIC	Test AUC	Train RMSE	Test RMSE	Test Accuracy
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i + q_i : u_i$	3961.9	0.727	0.337	0.342	0.849
$\text{logit}(\mu_i) \sim a_i + a_i^2 + m_i + q_i + u_i + o_i + t_i + s_i$	3965.8	0.728	0.337	0.341	0.850
$\text{logit}(\mu_i) \sim a_i + m_i + q_i + u_i + o_i + t_i + s_i$	4006.0	0.721	0.338	0.342	0.847
$\text{logit}(\mu_i) \sim a_i^{(5)} + m_i + q_i + u_i + o_i + t_i + s_i$	3978.6	0.730	0.336	0.342	0.847
$\text{logit}(\mu_i) \sim a_i^{(10)} + m_i + q_i + u_i + o_i + t_i + s_i$	3977.2	0.726	0.337	0.342	0.849

We selected the model based on AIC, which should help us find balance between model complexity and how well the model fits the data. This approach helps avoid overfitting meaning that the model generalises well to new data, without any loss of practicality.

To test the predictive performance of our model using our test data, we plot the predicted probability of each 'probability bin' against the mean actual outcomes and obtained the calibration chart in Figure @ref(tab:output calibration chart). As we can see the calibration curve closely follows the line $y = x$, which is indicative of a well-calibrated model.

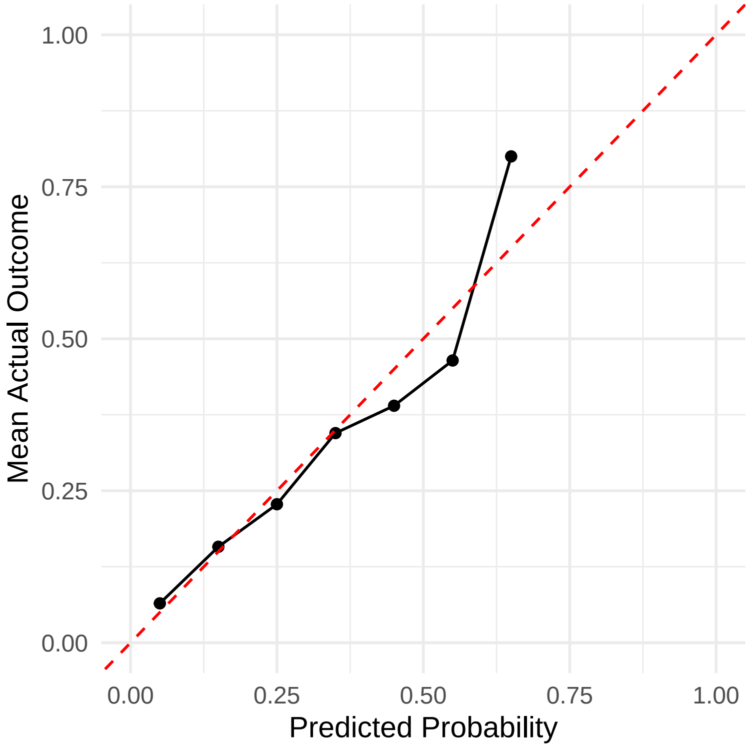

Figure 4: Calibration chart for Binomial model

We show a forest plot of the odds ratio in Figure *Need code for this* for each variable in our model. Note, it is important to state the reference level of each factor variable, so we have a baseline for comparison among other levels. The intercept (roughly 0.25) can be interpreted as the probability of being a smoker if all factor variables are at their reference level. Any blue variables, i.e. age, imd and urban, increase the probability by a scale of its corresponding odds ratio and all red variables are associated with a decreased probability of being a smoker.

Limitations of model -

Which lifestyle habits are associated with systolic blood pressure?

I will do this tomorrow!

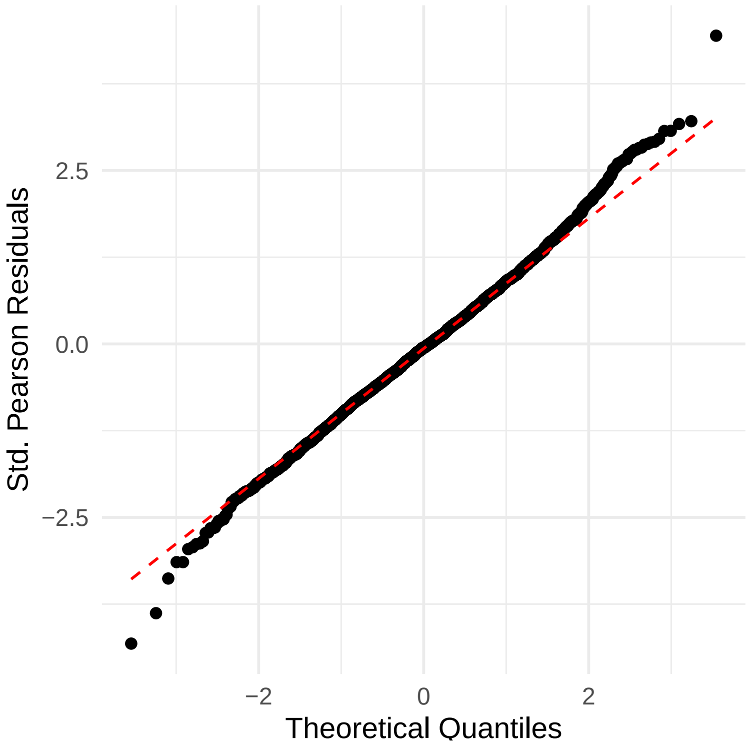

Figure 5: Q-Q plot of Inverse-Normal model residuals

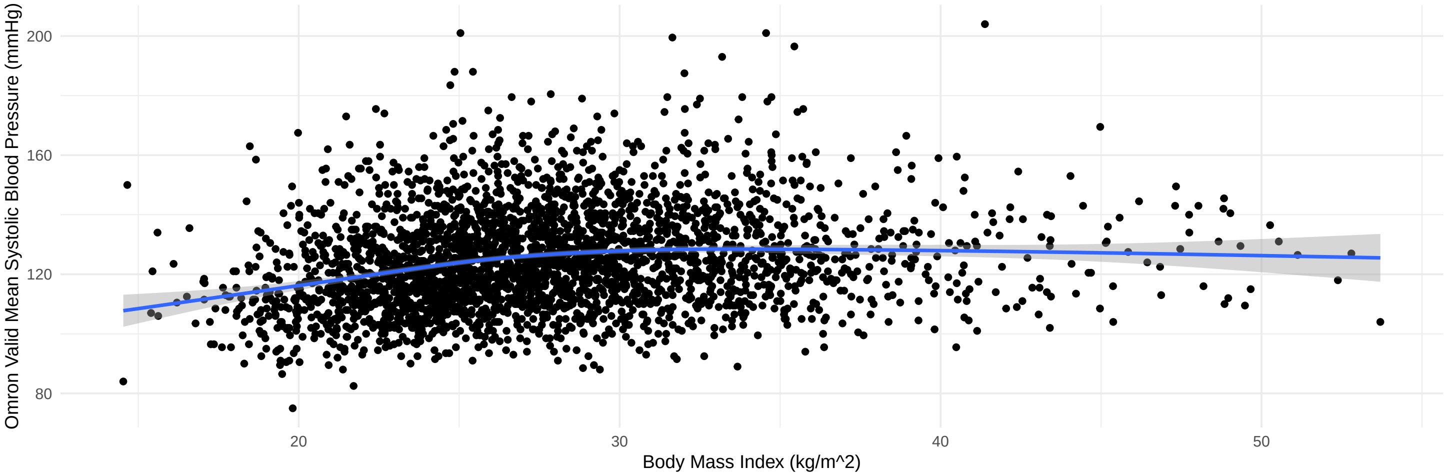

Figure 6: Relationship of BMI and Age with Mean Systolic Blood Pressure

Results/Conclusion

Our analysis showed that alcohol consumption is extremely common among UK adults, with approximately 81.8% being consumers, while smoking and vaping rates are lower at 16.7% and 4.4%, 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 16.7). *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.

NatCen Social Research and Health (2019) *?*

References

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

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.