London School of Hygiene & Tropical Medicine

Cover sheet for work submitted towards Assessment 2022/23

Please attach a copy of this sheet to *each* piece of work submitted.

LSHTM Candidate Number:	221162
Module Title:	Generalised Linear models
Module Number:	2462
Module Organiser:	Chris Frost & Ruth Keogh
Timetable slot:	C1
Date Submitted:	6/2/23
Type of Assessment:	Coursework
Word count *(when a word limit is specified)*	4 pages

Declaration on Plagiarism and Cheating

By submitting work for assessment you are confirming that:

- You understand the School's definitions of plagiarism and cheating (which follow); and that failure to comply with the School's policies may be penalised
- That all work submitted is your own
- You give consent for the School to upload relevant documents and information to external services or third parties, in order to identify potential plagiarism or irregularities.

Penalties for exceeding the word count apply to both module assessments and projects, please check stated word counts.

You should consult your tutor or Module Organiser if you are in any doubt about what is permissible.

Part 1 – Introduction & Preliminary analysis

Data on 2,700 individuals were collected through an observational study. Information on age, sex, education level (a proxy for social class), smoking status, and waist to hip ratio was recorded for all 2,700 individuals. There are two outcomes of interest – a binary carotid plaque status (plaque absent or present), and a whole-number valued plaque count. Plaque count was recorded for a random 30% of the sample. Table 1 shows summary statistics.

		Carotid plaque status (N=2700)		Carotid plaque count (N=810)	
		N	Plaque present, % (N)	N	Plaque count, mean (SD)
Age	40-49	803	42.2% (339)	245	0.7 (0.9)
Age	50-59	863	60.8% (525)	256	1.3 (1.5)
Age	60-69	1,034	79.8% (825)	309	2.1 (1.7)
Sex	Female	1,550	55.4% (858)	469	1.0 (1.2)
Sex	Male	1,150	72.3% (831)	341	2.0 (1.7)
Edu.	>= Univ.	1,725	66.2% (1,142)	510	1.6 (1.6)
Edu.	< Univ.	975	56.1% (547)	300	1.2 (1.4)
Smk.	Never	1,392	57.6% (802)	424	1.1 (1.3)
Smk.	Ex	661	64.0% (423)	202	1.5 (1.6)
Smk.	Current	647	71.7% (464)	184	2.0 (1.8)

Table 1: Descriptive statistics
Note: Age has been aggregated from 5-year groups to 10-year groups.

Presence & count of carotid plaque increases with age, is more frequent (higher count) in males compared to females, and more frequent (higher) in current & ex-smokers compared to never smokers. In this sample males are significantly more likely to smoke than females, the proportion of university educated individuals is significantly lower in older age groups, and a significantly lower proportion of the older age groups smoke compared to younger groups. Higher values of waist to hip ratio (W2H) are associated with higher chances of plaque presence and higher plaque counts, with correlations of 0.25 and 0.30 respectively. There is overdispersion in plaque count, which is addressed in part 4.

Part 2 – Smoking & presence of carotid plaque

This section examines the relation between smoking and presence of carotid plaque.
Methods: The data is aggregated by the confounders age, sex, and education. A logistic regression model (Model 1) of smoking status and the confounders is fit to the aggregated data – see table 3, model 1 & table footnotes for detail. Deviance is used to assess model fit, and more complex forms of the linear predictor are assessed through likelihood ratio tests. Model 1 was than fit to the individual level data and used to estimate marginal odds ratios of carotid plaque presence in ex-smokers and current smokers, compared to never smokers. Finally, model 1 was used to estimate prevalence of carotid plaque presence if everyone in the sample never smoked.

Results: Model 1 relates smoking status to age, sex, and level of education (all encoded as categorical variables), with no interactions between any of the variables. The deviance of this model is 63.6 on 62 degrees of freedom, comparing this to a χ^2_{62} distribution gives a p-value of 0.42 – there is no evidence of poor model fit. More complex models were also considered. Three interaction terms were separately added to model 1 – an interaction between smoking and age, between age and education, and between smoking and sex. A fourth model with interactions between all variables was also tried. None of the models had

significantly better fit relative to model 1, so the additional terms were not included. p-values from likelihood ratio tests were 0.66, 0.89, 0.16, and 0.42 respectively.

Model 1 estimates that, relative to never-smokers of the same age, sex, and education, there is no significant difference in the odds of plaque presence in ex-smokers and being a current smoker significantly increases the odds of carotid plaque by 92%. See table 3, model 1 in part 5 for p-values and confidence intervals.
Model 1 was fit to the individual data and used to produce marginal probabilities of plaque presence. Marginal odds ratios are also calculated and are shown in table 2.

Smoking status	Marginal probability	Marginal odds ratio
Never smoker	0.59 [0.57, 0.62]	REF
Ex-smoker	0.61 [0.58, 0.65]	1.10 [0.90, 1.33]
Current smoker	0.72 [0.68, 0.75]	1.74 [1.43, 2.17]

Table 2: Estimated marginal probability of plaque presence by smoking status. Notes: 95% confidence intervals shown in square brackets. Estimates are adjusted for age, sex, and education. Estimates are rounded to 2 decimals. Odds ratio CIs based on percentile bootstrap estimates.

The marginal probabilities in table 2 can be interpreted as prevalence of carotid plaque in the situation where all study participants have the same smoking status. For example, the prevalence is estimated to be 59% if no one in the study smoked. Assuming smoking has a causal effect on carotid plaque, this means that 3.5% (95% CI [1.0%, 6.0%]) of carotid plaque prevalence is due to smoking. If no one in the study ever smoked there would be between 27 and 163 fewer carotid plaque cases.

Part 3 – Waist to hip ratio & presence of carotid plaque

This section examines the effect of waist to hip ratio (W2H) on carotid plaque prevalence.
Methods: W2H is added to model 1 to form model 2. The effect of adding W2H is quantified using a likelihood ratio test, and interaction terms were examined using likelihood ratio tests. Model 2 is used to quantify the effect of increasing W2H on presence of carotid plaque. An interaction term between W2H and sex is then added to model 2, to form model 3. Model 3 is used to quantify (1) how the relation between W2H and carotid plaque prevalence varies by sex, and (2) the effect that following WHO obesity guidelines would have on carotid plaque prevalence in the study sample. Differences in the effect of W2H by sex are examined using a Wald test on the interaction term in model 3.

Results: A likelihood ratio test between models 1 and 2 was highly significant (19.9 on 1 degree of freedom, p < 0.0001), there is strong evidence that adding W2H improved model fit. There is no evidence that model fit is improved by including interaction terms between W2H and smoking, or W2H and age (likelihood ratio test p-values 0.41 and 0.83 respectively). As such, interaction terms were not added to Model 2.

Model 2 estimates that an increase in W2H of 0.1 leads to a significant increase in the odds ratio of having carotid plaque. Increasing W2H by 0.1 while holding age, sex, education, and smoking status constant increases the odds of carotid plaque presence by 30%. See table 3, model 2 for confidence intervals and p-values. Residual plots & calibration curves show no evidence of poor model fit. Including an interaction between sex and W2H does not significantly improve model fit (LR test p-value = 0.69) but is included in model 3 to quantify how the effect of W2H varies by sex. Model 3 estimates a 0.1 unit increase in W2H – holding age, sex, education, and smoking status constant – leads to a significant increase in the

odds ratio of having carotid plaque in both sexes. The odds ratio increases by 28% in females, and by 34% in males. See table 3, model 3 for confidence intervals and p-values. There is no evidence that the effect of W2H varies by sex (Wald test p-value = 0.69).

Model 3 was used to quantify the effect of reducing levels of obesity. Values of W2H which are classed by WHO as obese were reduced – males with W2H above 0.9 had their W2H value set to 0.9, and females with W2H above 0.85 had their W2H values set to 0.85. Model 3 was used as the WHO guidance varies by sex. If no one in the study was obese, the prevalence of carotid plaque presence is estimated at 60.6% (95% CI [58.6%, 62.5%]). Assuming the effect of W2H is causal, reducing obesity levels would decrease plaque prevalence by 2.0% (95% CI [0.0%, 3.9%]) and prevent up to 106 cases.

Part 4 – Smoking and waist to hip ratio & plaque count

This section looks at the 810 study participants with plaque count recorded.
Methods: Three negative binomial models of plaque count are fit – (model 4) age, sex, education, and smoking, (model 5) age, sex, education, smoking, and W2H, and (model 6) model 5 with an interaction between W2H & sex. Negative binomial models are used due to the overdispersion noted in part 1. Interactions between age & sex, and age & smoking status were tested using likelihood ratio tests.

Results: The likelihood ratio tests found evidence (p-value = 0.005) that the interaction between age and smoking improved model fit. However, the interaction term was not added in favour of model parsimony and to simplify the reporting of coefficient estimates. Table 3 shows the coefficient estimates, 95% confidence intervals, and p-values from the three models. Model 4 estimates that there is no significant difference in plaque count of ex-smokers and never-smokers of the same age, sex, and education. Current smokers have a significantly higher plaque count (50% higher on average compared to similar never-smokers). Model 5 shows that a 0.1 unit increase in W2H, while holding age, sex, education, and smoking status constant, leads to a significant increase in mean plaque count of 11%. Model 6 looks to see if this effect varies by sex. Increasing W2H leads to a significant increase in mean plaque count in females only (up 22% per 0.1 unit increase in W2H). A Wald test on the interaction term has a p-value of 0.86, there is no evidence that the effect of W2H varies by sex.

Part 5 – Discussion and conclusions

A series of models have been used in this report to estimate the causal effect of smoking and waist to height ratio on carotid plaque. Table 3 (next page) summarises these models.

Model 1 found that current smokers have a significantly higher odds of carotid plaque presence compared to never-smokers, and that there was no significant difference in the odds of plaque presence for ex-smokers compared to never-smokers. Model 4 looked at the effect of smoking on plaque count and found similar results – on average current smokers have significantly higher plaque count compared to never-smokers, and there is no significant difference in plaque count in never-smokers and ex-smokers. The remaining models examined the relation between W2H and carotid plaque. Increasing W2H led to significant increases in the odds of carotid plaque being present, and carotid plaque count (see models 2 and 5). The effect of W2H was then allowed to vary by sex, to examine if there are sex differences in the effect of W2H. No evidence of sex differences was found in either model 3 or 6. Model 6 found that increasing W2H caused a significant increase in plaque count in females but not in males. This may be due to a non-linear effect of W2H on carotid plaque count, where increasing W2H from small values causes a larger change in

plaque count compared to increasing W2H from large values of W2H. In this sample, females have lower W2H than males, which may explain the differences saw in model 6.

(a) Effect on carotid plaque prevalence – Logistic regression models					
Model	**Variable**	**Odds ratio**	**95% CI**	**p-value**	**Adjusted for**
1	Ex-smoker	1.12	[0.89, 1.40]	0.34	Age, sex, education
	Current smoker	1.92	[1.52, 2.43]	<0.0001	
2	W2H	1.30	[1.16, 1.46]	<0.0001	Age, sex, education, smoking
3	W2H – females	1.28	[1.11, 1.47]	0.001	
	W2H – males	1.34	[1.11, 1.62]	0.003	
(b) Effect on carotid plaque count – Negative binomial regression models					
Model	**Variable**	**Δ plaque**	**95% CI**	**p-value**	**Adjusted for**
4	Ex-smoker	1.05	[0.88, 1.24]	0.59	Age, sex, education
	Current smoker	1.50	[1.27, 1.78]	<0.0001	
5	W2H	1.11	[1.02, 1.22]	0.02	Age, sex, education, smoking
6	W2H – females	1.22	[1.08, 1.37]	0.001	
	W2H – males	1.01	[0.89, 1.15]	0.84	

Table 3: Effect estimates.
Notes: Δ plaque is the multiplicative change in mean plaque count. Effect estimates in models 1,4 relative to never-smokers. Effect estimates in models 2,3,5,6 are for an increase of 0.1 in W2H. All effect estimates in the table are conditional on the adjustment variables. Coefficients for adjustment variables not shown as they do not have a causal interpretation. Models 3,6 have an interaction between W2H and sex, all other models have no interactions. Model 1 fit on grouped data, all others on ungrouped data.

Being a current smoker causes the largest increase in plaque presence & count. Public health teams interested in reducing plaque prevalence can use this as an evidence base for setting up smoking cessation services. Prevalence estimates at the end of sections 2 & 3 show that the effect of reducing smoking in the population is likely to have a slightly larger, but similar, effect than reducing obesity levels. These results can be combined with other local knowledge – e.g., how likely the population is to engage with stop smoking or weight management services – when deciding which public health interventions to prioritise.

The data in this study is observational, which adds some limitations to our results. There are two sources of bias: (1) there may be unmeasured confounding which biases our coefficient estimates, e.g., a social factor which is not associated with education level. (2) The causal DAG we are assuming may be mis-specified. A potential DAG is shown below, this doesn't have W2H on the direct causal path between smoking and plaque. If this is the true DAG, then the effect of W2H would be correctly estimated by adjusting for either age and smoking or sex and smoking, neither of which were considered in this report. By also adjusting on smoking in model 5 we may be introducing bias into our model and getting spurious results.
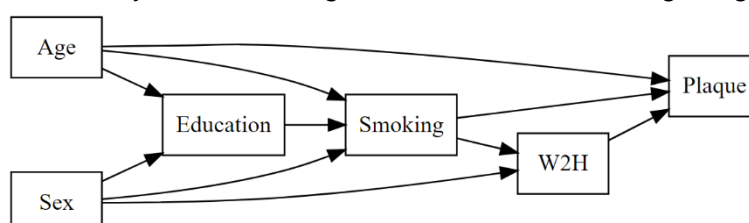
These sources of bias cannot be addressed through an RCT, as the 'interventions' (weight gain and smoking) would be unethical to randomly assign. Other observational studies – such as replication studies in different populations or running cohort studies to examine effects over time – may improve the strength of evidence found in this report.