(i) Fit a logistic model

Understanding the Problem:

We have a case-control study with:

- **Cases:** 1000 males with lung cancer.
- **Controls:** 1000 males without lung cancer.

The exposure variable is smoking:

- Among cases: 452 smokers, 548 non-smokers.
- Among controls: 215 smokers, 785 non-smokers.

Logistic Regression Model:

In logistic regression, we model the log-odds of the outcome (having lung cancer) as a linear function of the predictor (smoking).

Let:

- $Y = 1$ for cases (lung cancer), $Y = 0$ for controls.
- $X = 1$ for smokers, $X = 0$ for non-smokers.

The logistic model is:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X$$

Estimating Parameters:

In a case-control study, we cannot directly estimate $P(Y = 1|X)$ because the proportion of cases and controls is fixed by design. However, we can estimate the odds ratio, which is the exponentiated coefficient β_1.

The odds ratio (OR) can be calculated from the 2x2 table:

$$OR = \frac{a \cdot d}{b \cdot c} = \frac{452 \cdot 785}{215 \cdot 548}$$

Calculating:

$$a = 452, \quad b = 215, \quad c = 548, \quad d = 785$$

$$OR = \frac{452 \times 785}{215 \times 548}$$

First, compute numerator and denominator:

- Numerator: $452 \times 785 = 354,820$
- Denominator: $215 \times 548 = 215 \times 500 + 215 \times 48 = 107,500 + 10,320 = 117,820$

Now, divide:

$$OR = \frac{354,820}{117,820} \approx 3.01$$

Thus, $\beta_1 = \log(OR) \approx \log(3.01) \approx 1.102$.

For β_0, it's the log-odds of lung cancer when $X = 0$ (non-smokers). In case-control studies, the intercept is not directly interpretable as it depends on the sampling ratio of cases to controls.

However, if we proceed to fit the model based on the given data (even though it's not standard for case-control studies to interpret intercept directly), we can think of the logistic regression as:

$$\log\left(\frac{P(Y = 1|X)}{P(Y = 0|X)}\right) = \beta_0 + \beta_1 X$$

From the data:

- For $X = 1$ (smokers):

$$\log\left(\frac{452}{215}\right) = \beta_0 + \beta_1$$

- For $X = 0$ (non-smokers):

$$\log\left(\frac{548}{785}\right) = \beta_0$$

So:

$$\beta_0 = \log\left(\frac{548}{785}\right) \approx \log(0.698) \approx -0.359$$

$$\beta_1 = \log\left(\frac{452}{215}\right) - \beta_0 = \log(2.102) + 0.359 \approx 0.743 + 0.359 = 1.102$$

This matches our earlier calculation of β_1.

Final Logistic Model:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = -0.359 + 1.102X$$

(ii) Determine the probability of having lung cancer in both groups

In a case-control study, the absolute probabilities $P(Y = 1|X)$ cannot be directly estimated because the proportion of cases and controls is fixed by design. However, if we assume that the sample proportions reflect the population (which is not typically the case in case-control studies), we can proceed cautiously.

Assuming the sample reflects the population:

- Total smokers: $452 + 215 = 667$
  - $P(Y = 1|X = 1) = \frac{452}{667} \approx 0.678$ or 67.8%
- Total non-smokers: $548 + 785 = 1333$
  - $P(Y = 1|X = 0) = \frac{548}{1333} \approx 0.411$ or 41.1%

But this is not correct for case-control studies because the marginal distribution of Y is fixed by design. The correct approach is to recognize that probabilities cannot be directly estimated from case-control data without additional information (like disease prevalence).

However, if the question is asking for the observed proportions in the given sample (even though it's not population-representative), then:

- For smokers: $\frac{452}{667} \approx 0.678$
- For non-smokers: $\frac{548}{1333} \approx 0.411$

But this is not the population probability.

Alternative Interpretation:
Perhaps the question is asking for the predicted probabilities from the logistic model, even though the intercept is not directly interpretable. Using the model:

$$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

From (i), $\beta_0 \approx -0.359$, $\beta_1 \approx 1.102$.

- For smokers ($X = 1$):

$$\text{logit} = -0.359 + 1.102 = 0.743$$

$$P(Y = 1|X = 1) = \frac{e^{0.743}}{1 + e^{0.743}} \approx \frac{2.102}{3.102} \approx 0.678$$

- For non-smokers ($X = 0$):

$$\text{logit} = -0.359$$

$$P(Y = 1|X = 0) = \frac{e^{-0.359}}{1 + e^{-0.359}} \approx \frac{0.698}{1.698} \approx 0.411$$

These match the sample proportions, but again, this is not the population probability.

Given the ambiguity, the most likely interpretation is that the question is asking for the sample proportions, acknowledging that these are not population estimates.

Final Answer:

- Probability of having lung cancer for smokers: ≈ 0.678 or 67.8%
- Probability of having lung cancer for non-smokers: ≈ 0.411 or 41.1%

(iii) Compute log-odds for smokers and non-smokers

From the logistic model:

$$\text{log-odds} = \log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = \beta_0 + \beta_1 X$$

- For smokers ($X = 1$):

$$\text{log-odds} = -0.359 + 1.102 = 0.743$$

- For non-smokers ($X = 0$):

$$\text{log-odds} = -0.359$$

Alternatively, directly from the data:

- For smokers:

$$\text{log-odds} = \log\left(\frac{452}{215}\right) \approx \log(2.102) \approx 0.743$$

- For non-smokers:

$$\text{log-odds} = \log\left(\frac{548}{785}\right) \approx \log(0.698) \approx -0.359$$

Final Answer:

- Log-odds for smokers: ≈ 0.743
- Log-odds for non-smokers: ≈ -0.359

(iv) Obtain the odds ratio for lung cancer

The odds ratio (OR) compares the odds of lung cancer in smokers to the odds in non-smokers.

From the 2x2 table:

$$OR = \frac{a \cdot d}{b \cdot c} = \frac{452 \times 785}{215 \times 548}$$

Calculating:

$$452 \times 785 = 354,820$$

$$215 \times 548 = 117,820$$

$$OR = \frac{354,820}{117,820} \approx 3.01$$

Alternatively, from the logistic model:

$$OR = e^{\beta_1} = e^{1.102} \approx 3.01$$

Final Answer:

The odds ratio for lung cancer comparing smokers to non-smokers is approximately 3.01.

Summary of Answers:

(i) The fitted logistic model is:

$$\log\left(\frac{P(Y = 1|X)}{1 - P(Y = 1|X)}\right) = -0.359 + 1.102X$$

where $X = 1$ for smokers and $X = 0$ for non-smokers.

(ii) The estimated probabilities (from the sample) are:

- Smokers: ≈ 0.678 or 67.8%
- Non-smokers: ≈ 0.411 or 41.1%

(iii) The log-odds are:

- Smokers: ≈ 0.743
- Non-smokers: ≈ -0.359

(iv) The odds ratio is approximately 3.01.