# Biostatistics (MATH11230)

## Confounding and interaction within logistic regression

Vanda Inácio

University of Edinburgh


Semester 1, 2022/2023

# Refresher of the concepts of confounding and interaction

- ↪ Here we will study how to check for any confounding or interaction/effect modification using logistic regression.
- ↪ As we have learned, these are two ways an extraneous/third variable may affect the relationship between outcome and exposure.
- ↪ **Confounding** exists when the estimated relationship of interest changes when we add a third variable.
- ↪ **Interaction** or **effect modification** exits when the relationship between two variables is different for different levels of a third variable.

# Assessment of confounding using logistic regression models

## General context

- Let us start by the confounding issue and how to deal with it using logistic regression.
- For simplicity, we focus on two risk factors, say $X_1$ and $X_2$, and assume that there is no (multiplicative) interaction between $X_1$ and $X_2$.
- The ideas presented throughout generalise directly to the situation where there are more than two risk factors being analysed.
- Let us consider the following two logistic regression models:

$$\log\left(\frac{p_{x_1}}{1-p_{x_1}}\right) = \beta_0 + \beta_1 x_1, \tag{1}$$

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \beta_2 x_2. \tag{2}$$

- Here $p_{x_1} = \Pr(D \mid X_1 = x_1)$ and $p_{x_1,x_2} = \Pr(D \mid X_1 = x_1, X_2 = x_2)$.

# Assessment of confounding using logistic regression models

## General context

- ↪ As the notation suggests, there is a distinction between the two slope coefficients, $\beta_1$ and $\tilde{\beta}_1$, associated with the variable $X_1$ in these models.
- ↪ Because $X_2$ is not involved in the model (1), the coefficient $\beta_1$ is interpreted as the log odds ratio associated with a unit increase in the scale of $X_1$, ignoring any potential confounding role of $X_2$ (or any other variable, for that matter).
- ↪ In turn, $\tilde{\beta}_1$ in model (2) gives the log odds ratio associated with a unit increase in $X_1$, *holding the variable $X_2$ fixed*.
- ↪ The log odds ratio given by $\tilde{\beta}_1$ in model (2), measures the effect of a unit increase in $X_1$, in any population group that shares a common value of $X_2$.
- ↪ Hence, this measure of association of $X_1$ and $D$ directly accounts for the possible confounding effect of $X_2$.

# Assessment of confounding using logistic regression models

## General context

- ↪ Note that for simple categorical exposures, the maximum likelihood estimate of $\tilde{\beta}_1$ is an alternative estimator to the Mantel-Haenszel approach for estimating a common (log) odds ratio, accounting for the presence of a possible confounding variable.
- ↪ Maximum likelihood is a slightly different technique and so will usually yield similar, but not identical, results to the Mantel-Haenszel approach.
- ↪ Further, for arbitrary exposure variables, comparison of the corresponding estimated odds ratios, $e^{\widehat{\beta}_1}$ and $e^{\widehat{\tilde{\beta}}_1}$, allows us to assess the confounding of the $X_1$–$D$ relationship induced by $X_2$, assuming that models (1) and (2) adequately describe the exposure effects.
- ↪ As a rule of thumb, and as in the Mantel-Haenszel approach, if the ratio

$$\frac{e^{\widehat{\beta}_1} - e^{\widehat{\tilde{\beta}}_1}}{e^{\widehat{\beta}_1}},$$

is larger than 10% in size, then most epidemiologists would agree that there is evidence of confounding.

# Assessment of confounding using logistic regression models

## General context

- ↪ Note that in association studies (where the goal is to estimate the underlying relationship between a disease outcome and a set of exposure variables, our goal here) all variables that are clinically thought to be confounders should be studied by including them in the model.
- ↪ If a potential confounder changes to an important degree the estimate of the OR (e.g., 10% rule of thumb) or the confidence interval, the variable should be included in the model regardless of whether its coefficient is significant or not.
- ↪ If it does change the coefficient, decisions should be made on believability of results and statistical significance of the variable.
- ↪ I found the following short video about confounding by Sir David Spiegelhalter very instructive

`https://www.youtube.com/watch?v=j8J2L_g76c4`

# Assessment of confounding using logistic regression models

## Example

- ↪ As an example, we use the Western Collaborative Group Study data (that arises from a population-based design).
- ↪ We will study, this time, the effect of smoking on the risk of coronary heart disease (CHD).
- ↪ The variable `Ncigs0` in the dataset has information on the number of cigarettes smoked per day by each individual. To keep matters simple, we create a binary variable, that takes the value 0 if the individual does not smoke any cigarettes per day and 1 otherwise.
- ↪ After estimating the logistic regression model with binary smoking status as the only exposure, we obtain that the estimate of the corresponding regression coefficient is 0.6299.
- ↪ This means that smokers have $e^{0.6299} = 1.877$ times the odds of CHD than non-smokers. The corresponding 95% CI is $(1.445, 2.440)$.

# Assessment of confounding using logistic regression models

## Example

- Now, let us averiguate if the estimated association between smoking status and CHD was confounded by age.
- From a conceptual perspective, age satisfies the three criteria for being a potential confounder variable:
  - Age is likely a predictor of CHD. As age increases, the likelihood of CHD probably increases.
  - Age could also be positively or negatively associated with smoking, depending on the population.
  - It does not make sense for age to be in the causal pathway between smoking and CHD.

# Assessment of confounding using logistic regression models

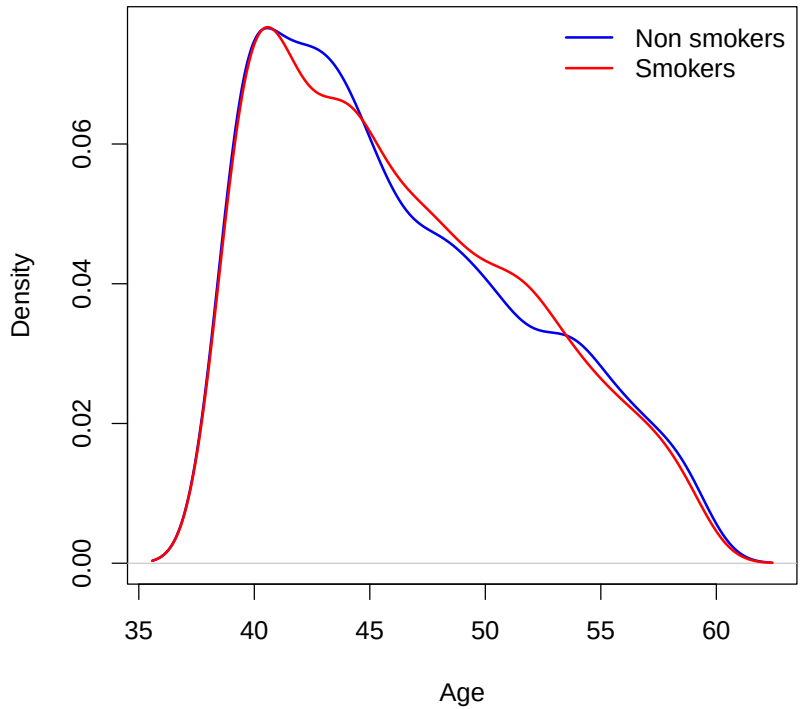
## Example

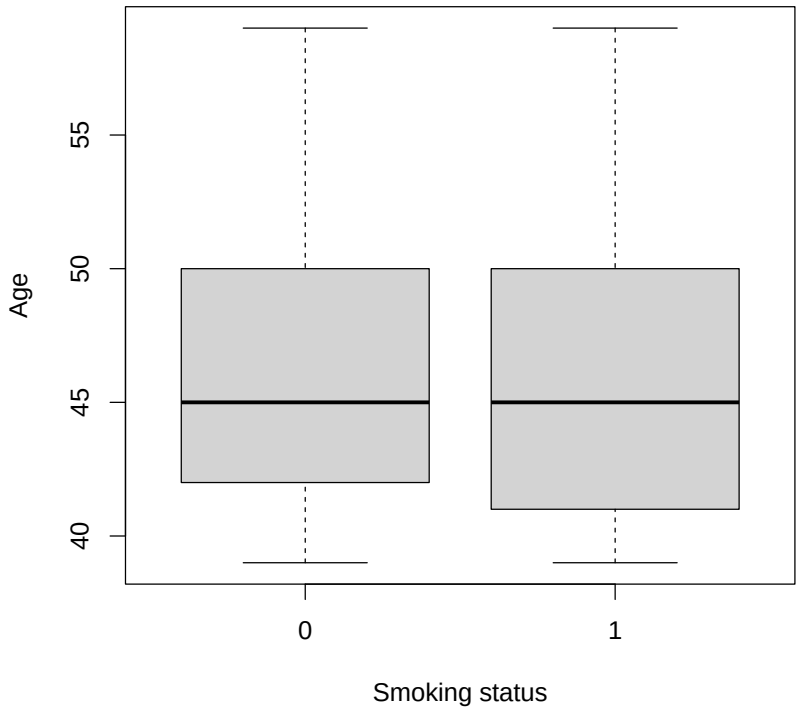
- ↪ Let us start by checking, based on the data, the association between age and smoking status (that is, between the exposure (binary) and the confounding variable (continuous)).
- ↪ We can create two groups of age: one for smokers and another one for non-smokers.
- ↪ We can then use t-tests or look at density plots (or boxplots) of the distributions. I prefer the second option as looking only at the means may be misleading (e.g., two distributions can have the same mean but still be quite different).
- ↪ A t-test comparing the two age groups leads to not reject the null hypothesis that the means of the two groups are the same. The mean age is similar for smokers and non-smokers.

# Assessment of confounding using logistic regression models

## Example

↪ The densities/boxplots in the two groups are basically indistinguishable.



# Assessment of confounding using logistic regression models

## Example

- ↪ Another approach to investigate a possible association between smoking status and age is to build a logistic regression model that looks at if we can use age to predict smoling status. If age can predict the smoking status, then there is an association between the two variables.

- ↪ By fitting such a model (please see the Supplementary Materials), the estimated odds ratio of smoking associated to an year increase of age is 1.002 (95% CI: (0.989, 1.014)). Again, we arrive at the conclusion that there is no association between age and smoking status.

- ↪ The Wald test also does not lead to rejection of the null hypothesis that the age coefficient is zero.

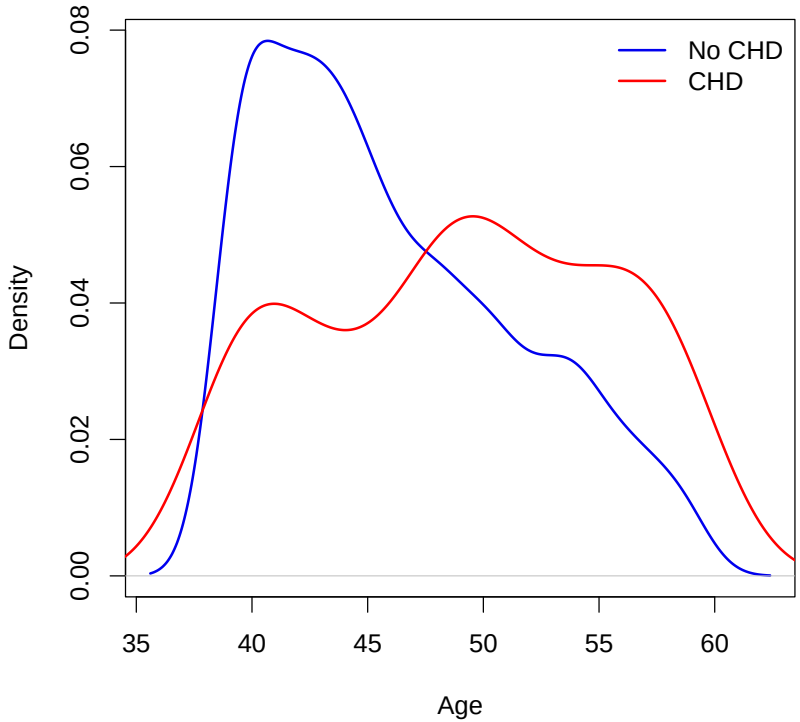
# Assessment of confounding using logistic regression models

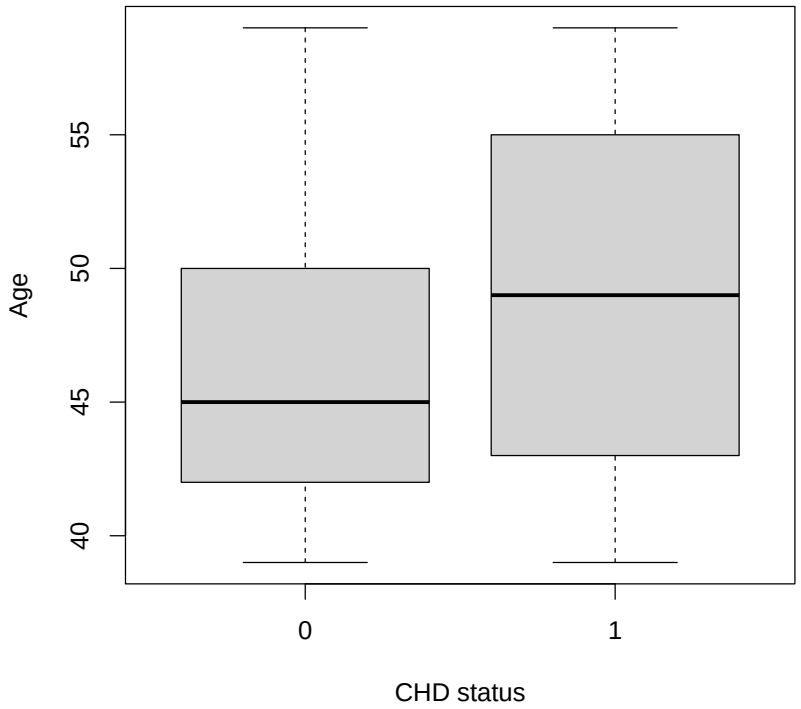
## Example

- ↪ We have just seen that for this specific dataset age does not seem to be associated with smoking status and so we should not expect it to be confounded the association found between smoking status and CHD.
- ↪ Neverthless, let us proceed and check whether age is associated with the CHD outcome in the unexposed group (i.e., in the group of non-smokers).
- ↪ By creating two groups of age in the unexposed group, one for those with CHD and another one for those without CHD, the corresponding t-test leads us to reject the null hypothesis of equal means in the two groups.
- ↪ There is a significant difference in mean age by CHD status in the unexposed group. Those with CHD tend to be slightly older than those without CHD.

# Assessment of confounding using logistic regression models

## Example



# Assessment of confounding using logistic regression models

## Example

- ↪ In a similar vein to what we have done before, we can also fit a logistic regression model to investigate whether age can predict the CHD outcome in the non-smokers group.
- ↪ The estimated odds ratio of CHD associated with a year increase on age is 1.1 (95% CI: $(1.057, 1.134)$). There is thus a **very** slight association between age and CHD outcome (for non-smokers).
- ↪ The Wald test also does lead to rejection of the null hypothesis that the age coefficient is zero (but look at estimated effect size...).

# Assessment of confounding using logistic regression models

## Example

- ↪ To sum up: age was significantly associated with CHD in the non-smokers group but not with smoking.
- ↪ Let us look now at the change in the smoking coefficient when age is added to the model.
- ↪ Recall that in our simple logistic regression model the estimated odds ratio was 1.877 $(1.445, 2.440)$.
- ↪ When age is added to the model, the estimated odds ratio of CHD comparing smokers to non-smokers, after adjusting/controlling for age, is 1.893 $(1.454, 2.465)$.
- ↪ There is not much of a change compared with the unadjusted (for age) odds ratio. Further, the aforementioned rule of thumb is less than 1% in size.
- ↪ So based on the lack of odds ratio change and on the bivariate associations, it does not look like age is a confounder.

# Assessment of confounding using logistic regression models

## Example

- ↪ But we should also check if including age makes our model better by increasing predictive power.
- ↪ This could provide some evidence to keep age in the model!
- ↪ We can test this by conducting a Wald test/likelihood ratio test (the reduced model is nested within the full model) or by looking at the AIC/BIC.
- ↪ The AIC/BIC of the second model is lower than the AIC/BIC of the first one, suggesting that the second model is better.
- ↪ The null hypothesis of the Wald test is that the regression coefficient associated with age is zero. The p-value obtained is basically zero, and so there is strong evidence in favour of rejecting the null hypothesis.
- ↪ We would arrive at the same conclusion by doing a likelihood ratio test.

# Assessment of confounding using logistic regression models

## Example

- So, what do we do? Should we keep age in the model or not?
- We could argue either way!!
- We saw that age did not meet the criteria for confounding numerically and the smoking coefficient did not change much. Based on these reasons, we might not want to keep age in the model.
- But we saw that conceptually age could be a potential confounder, and we also see from the likelihood ratio test and AIC/BIC that it improved the model. Often, age is also adjusted for in epidemiological studies for face validity. These are reasons to keep it in the model.
- Personally, and for the reasons mentioned in the previous point, I would keep age in the model.

# Introducing interaction into the multiple logistic regression model

## General context

- ↪ Let us now turn attention to the issue of effect modification/interaction. For the methodological exposition, we follow Jewell (2003, chapter 14).
- ↪ All logistic regression models with more than one exposure variable that we have considered so far assume no interaction amongst the exposure variables.
- ↪ We will now learn how to extend the multiple logistic regression model to allow for the possibility of interaction effects.

# Introducing interaction into the multiple logistic regression model

- For simplicity, we begin with the simplest situation where interest focuses on the impact of two risk factors, say $X_1$ and $X_2$, on an outcome $D$.

- As noted, the model

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \log(\text{odds of } D \mid X_1 = x_1, X_2 = x_2)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

  assumes no interaction between $X_1$ and $X_2$.

- To incorporate an interaction, we simply need to add to this model an additional derived covariate, $X_3$, defined by $X_1 \times X_2$.

# Introducing interaction into the multiple logistic regression model

↪ We then fit the following model:

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2). \quad (3)$$

↪ The interpretation of the intercept coefficient, $\beta_0$, remains as before, namely, the log odds of $D$ when both $X_1$ and $X_2$ are zero.

↪ The interpretation of the slope coefficients is, however, somewhat different, as we will now learn.

# Introducing interaction into the multiple logistic regression model

- Consider two groups of individuals whose risk factor $X_1$ differs by one unit on the scale of $X_1$ and who share identical values for the other risk variable $X_2$.
- That is, suppose that one group has risk variables given by $X_1 = x_1 + 1$ and $X_2 = x_2$, and the other group has levels $X_1 = x_1$ and $X_2 = x_2$.
- Then, the logistic regression model in (1) indicated that the differences in the log odds of $D$ of these two groups is simply

$$[\beta_0 - \beta_1(x_1 + 1) + \beta_2 x_2 + \beta_3(x_1 + 1)x_2] - [\beta_0 - \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2] = \beta_1 + \beta_3 x_2.$$

- This, as before, is the log odds ratio associated with a unit increase in $X_1$, but now this log odds ratio depends on the fixed level of $X_2$.
- In other words, this log odds ratio is modified by $X_2$. This is exactly what we wanted when adding the interaction term in the model, that the effect of $X_1$ on $D$ is modified by the levels of $X_2$.

# Introducing interaction into the multiple logistic regression model

- Suppose that both $X_1$ and $X_2$ are binary and coded with values 0 and 1 to describe their two levels.

- For a concrete example, when studying breast cancer incidence, let $X_1$ denote the use of oral contraceptives (1: yes, 0: no) and $X_2$ be the woman's age (1 if age is $\geq 40$ and 0 if age is below 40).

- We can interpret that when $X_2 = 0$ (women below 40), the log odds ratio comparing the two levels of $X_1$, women who take oral contraceptives to those who do not, is just $\beta_1$. To see why

$$\begin{aligned}
&\log(\text{odds breast cancer} \mid \text{use contraceptive and age} < 40) - \log(\text{odds breast cancer} \mid \text{no contraceptive and age} < 40) \\
&= [\beta_0 + \beta_1 \times 1 + \beta_2 \times 0 + \beta_3 \times 1 \times 0] - [\beta_0 + \beta_1 \times 0 + \beta_2 \times 0 + \beta_3 \times 0 \times 0] \\
&= \beta_1
\end{aligned}$$

- On the other hand, when $X_2 = 1$ (women at or above 40 years), the log odds ratio comparing the two levels of $X_1$ is $\beta_1 + \beta_3$. To see why

$$\begin{aligned}
&\log(\text{odds breast cancer} \mid \text{use contraceptive and age} \geq 40) - \log(\text{odds breast cancer} \mid \text{no contraceptive and age} \geq 40) \\
&= [\beta_0 + \beta_1 \times 1 + \beta_2 \times 1 + \beta_3 \times 1 \times 1] - [\beta_0 + \beta_1 \times 0 + \beta_2 \times 1 + \beta_3 \times 0 \times 1] \\
&= \beta_1 + \beta_3
\end{aligned}$$

# Introducing interaction into the multiple logistic regression model

- ↪ Thus, the parameter $\beta_3$ measures the difference in the log odds ratio associated with $X_1$ (use of oral contraceptives) between women at or above 40 years and women younger than 40 years (the $X_2 = 1$ and $X_2 = 0$ strata).

- ↪ Further, testing the null hypothesis $H_0 : \beta_3 = 0$ against the alternative $H_A : \beta_3 \neq 0$ provides a test of the evidence for heterogeneous odds ratios, that is, for interaction.

- ↪ Similarly, in model (1), the log odds ratio comparing women with an age equal or above 40 years to women younger than 40 is $\beta_2$ for those who do not take oral contraceptives and $\beta_2 + \beta_3$ for women who take oral contraceptives.

# Introducing interaction into the multiple logistic regression model

- Now suppose that the risk factor can assume several discrete levels.
- For a concrete example, let us revisit the Western collaborative group study data.
- Let $X_1$ be the dichotomised age (1 if $\geq 45$ (median) and 0 if $< 45$) and let $X_2$, $X_3$, $X_4$, and $X_5$ capture the five categories of body weight, that is,
  - $X_2 = 1$ if body weight is $150^+$ to 160 lb, and $X_2 = 0$ otherwise.
  - $X_3 = 1$ if body weight is $160^+$ to 170 lb, and $X_3 = 0$ otherwise.
  - $X_4 = 1$ if body weight is $170^+$ to 180 lb, and $X_4 = 0$ otherwise.
  - $X_5 = 1$ if body weight is $> 180$ lb, and $X_5 = 0$ otherwise.

# Introducing interaction into the multiple logistic regression model

↪ To allow for the possibility that the log odds ratio associated with age is different in each of the five strata defined by the levels of body weight, we must add 4 product terms to obtain the analogous of model (1), each extra derived covariate comprising the product of $X_1$ and one of the 4 indicator variables, $X_2$, $X_3$, $X_4$, and $X_5$.

↪ That is, we fit the following model:

$$\log\left(\frac{p_{x_1,x_2,x_3,x_4,x_5}}{1-p_{x_1,x_2,x_3,x_4,x_5}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5.$$

# Introducing interaction into the multiple logistic regression model

- With this model, $\beta_0$ is the log odds of CHD for a person younger than 45 years old and weighing less than 150 lb.
- The log odds ratio comparing the two levels of $X_1$ (age $\geq 45$ vs age $< 45$) is
  - $\beta_1$ for the reference group of weight (i.e., weight below 150 lb), $X_2 = X_3 = X_4 = X_5 = 0$.
  - $\beta_1 + \beta_6$ for those in the weight category $150^+$ to 160 ($X_2 = 1$ and $X_3 = X_4 = X_5 = 0$).
  - $\beta_1 + \beta_7$ for those in the weight category $160^+$ to 170 ($X_3 = 1$ and $X_2 = X_4 = X_5 = 0$).
  - $\beta_1 + \beta_8$ for those in the weight category $170^+$ to 180 ($X_4 = 1$ and $X_2 = X_3 = X_5 = 0$).
  - $\beta_1 + \beta_9$ for those in the weight category $> 180$ ($X_5 = 1$ and $X_2 = X_3 = X_4 = 0$).

# Introducing interaction into the multiple logistic regression model

- ↪ This achieves the goal of permitting a different odds ratio for age (dichotomised) at each level of weight.

- ↪ Testing the null hypothesis $\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (through a likelihood ratio test) assesses the homogeneity, or lack thereof, of these 5 odds ratios for age, thus providing a test of interaction between age and body weight.

# Introducing interaction into the multiple logistic regression model

- Lastly, we consider one further situation: when both risk factors $X_1$ and $X_2$ are measured on a continuous scale.

- One possible logistic regression model for $X_1$ and $X_2$ that permits interaction is given by

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2).$$

- Now the log odds ratio associated with a unit increase in $X_1$ is given by $\beta_1 + \beta_3 x_2$ at a fixed level of the second risk factor $X_2 = x_2$.

- Thus, this model allows the odds ratio associated with $X_1$ to vary across the levels of $X_2$, but only according to a linear trend on the scale of $X_2$.

- As a variant of this model, we may wish to fit $X_2$ as a 'main effect' invoking its continuous scale, but use an indicator for a categorised version of $X_2$ in the interaction terms to avoid the trend assumption in the interactive effects.

# Introducing interaction into the multiple logistic regression model

- As a final comment, it is, in principle, possible to examine higher order interaction terms involving three risk factors, say, $X_1$, $X_2$, and $X_3$.

- A second order interaction term examines the extent to which the nature of the interaction, or effect modification, between $X_1$ and $X_2$ is itself modified by the levels of $X_3$.

- However, such higher order interaction effects are rarely studied with epidemiological data due mainly to two reasons:

  1. It is difficult to interpret them.

  2. There is reduced power (probability of detecting an effect, if there is a true effect) to assess them.

# Introducing interaction into the multiple logistic regression model

- ↪ Let us revisit the example involving the Western Collaborative Group Study that we have used to illustrate the assessment of confounding.
- ↪ Recall we were interested in the effect of smoking status (dichotomised) on the risk of coronary heart disease (CHD) and we checked whether age (measured on a continuous scale) was a confounder of such an association.
- ↪ We have concluded that age was not a confounder but what if age is an effect modifier?
- ↪ We are therefore interested in answer the question: does the effect smoking has on CHD depends on age?

# Introducing interaction into the multiple logistic regression model

- Letting $X_1$ be the binary variable denoting smoking status and taking the value 1 if the individual smokes, at least, one cigarette per day and 0 if the individual does not smoke at all. Let the continuous effect of age be captured by $X_2$.

- We fit the following regression model

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2).$$

- Note that $\beta_0$ (log odds of CHD for a non-smoker of 0 (!!!) years old) is not meaningful unless we center the age variable.

- The log odds ratio comparing smokers to non-smokers is $\beta_1 + \beta_3 x_2$ for a fixed value of age $X_2 = x_2$.

- To check whether age is an effect modifier we test

$$H_0 : \beta_3 = 0, \quad \text{vs} \quad H_A : \beta_3 \neq 0.$$

# Introducing interaction into the multiple logistic regression model

- This hypothesis can be tested either through a Wald test or a likelihood ratio test. Both yielded a p-value of about 0.25.
- This suggests that age is not an effect modifier. The effect of smoke on the odds of CHD does not appear to vary depending on age.
- Based on the model fit, we concluded that there is **no** significant interaction between smoking and age.
- The following is assuming that the interaction effect was significant for demonstration purposes only.
- This means that we want to report on the effect of smoking for different ages.
- Anything we say about the effect of smoking on CHD needs to be age-specific!!

# Introducing interaction into the multiple logistic regression model

- Let us compute the estimated log odds ratio associated with smoking (i.e., comparing smokers to non-smokers) for an individual who is 50 years old.

- The coefficient estimates based on model (2) are as follows:

$$\widehat{\beta}_0 = -7.07999, \quad \widehat{\beta}_1 = 1.91472, \quad \widehat{\beta}_2 = 0.09077, \quad \widehat{\beta}_3 = -0.02639.$$

- We thus have that the required estimated log odds ratio is given by

$$\begin{aligned}
&[\widehat{\beta}_0 + \widehat{\beta}_1 \times 1 + \widehat{\beta}_2 \times 50 + \widehat{\beta}_3 \times 1 \times 50] - [\widehat{\beta}_0 + \widehat{\beta}_1 \times 0 + \widehat{\beta}_2 \times 50 + \widehat{\beta}_3 \times 0 \times 50] \\
&= \widehat{\beta}_1 + 50 \times \widehat{\beta}_3 \\
&= 1.91472 + 50 \times (-0.02639) \\
&= 0.59522.
\end{aligned}$$

- Among those aged 50, smokers have $e^{0.59522} = 1.81$ times the odds of CHD compared to non-smokers.

# Introducing interaction into the multiple logistic regression model

- We shall note that the confidence interval for this log odds ratio depends on the sampling variance of $\widehat{\beta}_1 + 50 \times \widehat{\beta}_3$.

- This is just

$$\widehat{\text{var}}(\widehat{\beta}_1) + \widehat{\text{var}}(50\widehat{\beta}_3) + 2\widehat{\text{cov}}(\widehat{\beta}_1, 50\widehat{\beta}_3) = \widehat{\text{var}}(\widehat{\beta}_1) + (50^2)\widehat{\text{var}}(\widehat{\beta}_3) + 2 \times 50 \times \widehat{\text{cov}}(\widehat{\beta}_1, \widehat{\beta}_3).$$

- The third term appears because the estimates of $\widehat{\beta}_1$ and $\widehat{\beta}_3$ are correlated.

- In `R` we can easily have access to such covariance. After fitting the model in (2) with the aid of the `glm` function, we can use the command `vcov` (which returns the entire covariance matrix) evaluated on an object that contains our fitted logistic regression model (see its implementation in the Supplementary Materials).