Biostatistics (MATH11230)

Vanda Inácio

University of Edinburgh

Semester 1, 2021/2022

Introducing interaction into the multiple logistic regression model

General context

- For the methodological exposition, we keep following 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 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:

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

↪ The interpretation of the intercept coefficient 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 see.

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, 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 ≥ 40 and 0 if age is below 40).

Introducing interaction into the multiple logistic regression model

- 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 and women who do not, is just β_1.

- In 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 $\beta_1 + \beta_3$.

- Thus, the parameter β_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 β_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 at or above 45 (median) and 0 if below 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.
- The baseline or reference group is formed by those who weigh 150 lb or less.

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, similar calculations show that the log odds ratio comparing the two levels of X_1 (age ≥ 45 vs age < 45) is
  - β_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 $\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ 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 we have used in the last lecture. Recall we were interested in the effect of smoking status (dichotomised) on the risk of coronary heart disease (CHD) and we checked whether age 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

$$\log\left(\frac{p_{x_1,x_2}}{1-p_{x_1,x_2}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2). \tag{2}$$

- To check whether age is an effect modifier we test

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

- 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.

Introducing interaction into the multiple logistic regression model

- ↪ 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 sat 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 easy have access to such covariance. After fitting the model in (2) with the aid of the `glm` function, we can use the command `vcov` evaluated on an object that contains our fitted logistic regression model.