Biostatistics (MATH11230)
Introduction to Logistic Regression

Vanda Inácio

University of Edinburgh

Semester 1, 2022/2023

Logistic regression models to relate exposure to disease

General context

- Logistic regression models allow us to study the effect of several risk factors.
- Further, logistic regression models allow us to expand greatly the scope we have been working with so far, as they can handle binary/categorical risk factors as well as risk factors that are measured on a continuous scale.
- Logistic regression models are also a valuable tool when we want to control for confounding variables. We have learned about stratification, but this may not be viable if it is necessary to stratify on many potential confounding variables simultaneously. By using regression models, we also open the door to the possibility of controlling for continuous confounding variables (without having to categorise them).
- These slides follow partially Chapter 12 of Jewell (2003).

Logistic regression models to relate exposure to disease

Naive approaches and its drawbacks

- $\hookrightarrow$ We will now be labelling the exposure variable of interest with X instead of E and it can either represent a binary risk factor, one that has several (more than two) discrete categories or a risk factor measured on a continuous scale.

- $\hookrightarrow$ This is mainly for consistency with standard treatment of regression models, and it also reinforces the possibility that the exposure of interest may now be measured on a continuous scale.

- $\hookrightarrow$ The simplest model that we can think of is the linear model, under which we would write

$$p_x = Pr(D \mid X = x) = \beta_0 + \beta_1 x. \tag{1}$$

- $\hookrightarrow$ As the name says, the model in (1) assumes that as the exposure level $X = x$ changes, the risk of D, as measured by $Pr(D \mid X = x)$, changes linearly in x.

Logistic regression models to relate exposure to disease

Naive approaches and its drawbacks

- However there is a structural drawback about the use of the linear model for binary outcome data.

- Whatever the values of the parameters β_0 and β_1, at some values in the range of X, either low values or high values, the model in (1) may predict values of $p_x < 0$ and $p_x > 1$, which are not valid for risks.

- An alternative specification to the linear model would be the **log linear model** that assumes a linear relationship between the log risk of D and the exposure, that is

$$\log(p_x) = \log\{\Pr(D \mid X = x)\} = \beta_0 + \beta_1 x, \qquad (2)$$

or, equivalently,

$$p_x = \Pr(D \mid X = x) = e^{\beta_0 + \beta_1 x}.$$

- However, the risk $e^{\beta_0 + \beta_1 x}$ can still exceed one for any nonzero value of β_1 with large (or small, depending on the sign of β_1) values of X.

Regression models relating exposure to disease

The (simple) logistic regression model

- The **simple logistic regression model** relates p_x to x through the following equation:

$$p_x = \Pr(D \mid X = x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}} = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{3}$$

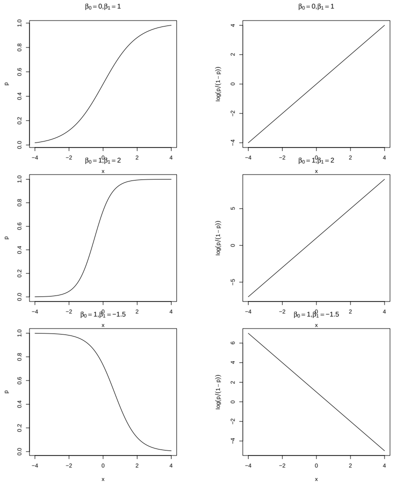
- Alternatively, the above relationship in (3) can be expressed in terms of the log odds associated with p_x

$$\log\left(\frac{p_x}{1 - p_x}\right) = \log(\text{odds for } D \mid X = x) = \beta_0 + \beta_1 x. \tag{4}$$

- The key assumption underlying the model in (4) is that the log odds of D changes linearly with changes in X.

- Because the term $e^{-(\beta_0+\beta_1 x)}$ is always positive, the risk p_x in (3) must lie between 0 and 1 for any values of β_0 and β_1 and for any level of the exposure.

- Thus, the logistic regression model does not suffer from the structural drawback shared by both the linear and log linear model (predicting, at some exposures, negative or above one risks).



$\beta_0=0, \beta_1=1$
$\beta_0=0, \beta_1=1$
$\beta_0=1, \beta_1=2$
$\beta_0=1, \beta_1=2$
$\beta_0=1, \beta_1=-1.5$
$\beta_0=1, \beta_1=-1.5$
p
log(p/(1−p))
x

Regression models relating exposure to disease

The (simple) logistic regression model

$\hookrightarrow$ The value $\beta_1 = 0$ represents no relationship between the risk of D and exposure level, i.e., independence between D and X.

$\hookrightarrow$ When β_1 is positive, the risk of D increases as exposure increases.

$\hookrightarrow$ In turn, when β_1 is negative, the risk of D decreases as the level of exposure increases.

Regression models relating exposure to disease

The (simple) logistic regression model

- Let us now turn our attention to the interpretation of the parameters β_0 and β_1.
- The intercept β_0 is the log odds of D when $X = 0$.
- In order to understand the interpretation of the slope β_1, let us consider two exposure levels separated by one unit on the scale of X, say $X = x + 1$ and $X = x$.
- The log odds ratio comparing such two exposure groups is

$$\begin{aligned}
\log(\mathrm{OR}) &= \log\left(\frac{\text{odds of } D \mid X = x+1}{\text{odds of } D \mid X = x}\right) \\
&= \log\left(\frac{p_{x+1}/(1-p_{x+1})}{p_x/(1-p_x)}\right) \\
&= \log(p_{x+1}/(1-p_{x+1})) - \log(p_x/(1-p_x)) \\
&= [\beta_0 + \beta_1 \times (x+1)] - [\beta_0 + \beta_1 \times x] \\
&= \beta_1.
\end{aligned}$$

Regression models relating exposure to disease

The (simple) logistic regression model

- ↪ Thus β_1 is the log odds ratio associated with comparing two exposure groups whose exposure differs by one unit on the scale of X, i.e., the log odds ratio associated with a unit increase in X.
- ↪ Note that this odds ratio, associated with a unit increase in X, does not depend on the choice of the baseline value X from which this unit increase is measured.
- ↪ For instance, considering the familiar situation where X can only take two values, say $X = 1$ (exposed) and $X = 0$ (unexposed), β_0 is the log odds of D amongst the unexposed and β_1 is the log odds ratio of D comparing the exposed to the unexposed.

Regression models relating exposure to disease

The (simple) logistic regression model

- Now suppose that we are interested in describing the relationship between the risk of infant mortality and birth weight (X), measured in grams. We do not need to dichotomise birth weight (e.g., as low and normal) as we did before when considering a similar example.

- In this context, β_0 corresponds to the log odds of infant mortality for a baby with zero grams as birth weight ($X = 0$). This is, obviously, extrapolating beyond the range of values of X in the population.

- Also, in this context, β_1 gives the log odds ratio of infant mortality comparing babies with birth weights that differ 1g (e.g., 2001g to 2000g, or 2501g to 2500g, or...).

- In this example, none of the parameters has a useful interpretation as zero birth weight is impossible and an increase in birth weight of 1 g is too small to expect any meaningful difference in the (log) odds ratio.

Regression models relating exposure to disease

The (simple) logistic regression model

- Recentring the exposure variable(s) X is a trick commonly used in regression modelling in such cases.
- For instance, recentring the scale of birth to be, say $X = \text{birthweight} - 2500\text{g}$ gives a more useful interpretation of the intercept β_0.
- With respect to the slope β_1, one can simply rescale birth weight in terms of 100g, for instance.
- Both of these changes lead to a new scale given by

$$X^* = (X - 2500)/100 = (\text{birthweight} - 2500)/100.$$

- The resulting model

$$\log\left(\frac{p_{x^*}}{1 - p_{x^*}}\right) = \beta_0 + \beta_1 x^*,$$

has a more interpretable intercept (log odds of infant mortality at 2500g) and slope (log odds ratio of infant mortality associated with an increase of 100g in birth weight) parameter.

Regression models relating exposure to disease

The (multiple) logistic regression model

- Let us now suppose that we have several risk factors that we wish to relate to the risk for D, say $X_1, \ldots, X_k$.

- At given exposure levels, say $X_1 = x_1, \ldots, X_k = x_k$, we use $p_{x_1,\ldots,x_k}$ to denote $\Pr(D \mid X_1 = x_1, \ldots, X_k = x_k)$.

- It is straightforward to extend the simple logistic regression model to accommodate k risk variables: we simply add linear terms to the right-hand side of the model in (4)

$$\log\left(\frac{p_{x_1,\ldots,x_k}}{1 - p_{x_1,\ldots,x_k}}\right) = \log(\text{odds of } D \mid X_1 = x_1, \ldots, X_k = x_k) = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k. \quad (5)$$

- Expressing this is terms of the risk yields

$$p_{x_1,\ldots,x_k} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}}.$$

Regression models relating exposure to disease

The (multiple) logistic regression model

- Holding $X_2, \ldots, X_k$ fixed, the pattern that describes how the risk of D changes as X_1 (alone) changes is still a logistic curve as the ones illustrated in slide 13.

- This is obviously true if each risk factor (not only X_1) is examined in turn, keeping constant the other variables of the model.

- How should we interpret $\beta_0, \beta_1, \ldots, \beta_k$?

- As for the model with only one risk factor, if $X_1 = X_2 = \ldots = X_k = 0$,we have that

$$\log\left(\frac{p_{0,\ldots,0}}{1 - p_{0,\ldots,0}}\right) = \beta_0.$$

- Hence, β_0 is just the log odds of D at the baseline level where all the risk variables are at zero, on their respective scales.

Regression models relating exposure to disease

The (multiple) logistic regression model

- For the slope parameters, $\beta_1, \ldots, \beta_k$, let us consider the comparison of two groups whose risk factor X_1 differs by one unit on the scale of X_1, and who share identical values for all other risk variables $X_2, \ldots, X_k$.

- That is, one group has risk variables given by $X_1 = x_1 + 1, X_2 = x_2, \ldots, X_k = x_k$ and the other group has $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$.

- Then the multiple logistic regression model in (5) tells us that the difference in log odds of D in these two groups is simply given by

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

- That is, β_1 is the log odds ratio associated with a unit increase in the scale of X_1, holding all other risk variables in the model constant.

Regression models relating exposure to disease

The (multiple) logistic regression model

- In general, β_j is the log odds ratio associated with a unit increase in the scale of X_j, holding all other variables in the model fixed.

- Note that that the log odds ratio (and, by consequence, the odds ratio) associated with changes in X_j is not affected by the values (held fixed) of the other variables in the model.

- That is, the multiple logistic regression model in (5) assumes that there is no (multiplicative) interaction between X_j and the other variables. We will later extend the model to allow for interactions/effect modifiers.

- Knowledge of the slope parameters, $\beta_1, \ldots, \beta_k$, allow us to compute the odds ratio comparing any two groups with specified risk factors. If our reference group has $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$ and the second group has $X_1 = x_1^*, X_2 = x_2^*, \ldots, X_k = x_k^*$, then the log odds ratio comparing the risk in these two groups is

$$\beta_1(x_1^* - x_1) + \beta_2(x_2^* - x_2) + \ldots + \beta_k(x_k^* - x_k),$$

  and the corresponding odds ratio is

$$\mathrm{OR} = e^{\beta_1(x_1^* - x_1) + \beta_2(x_2^* - x_2) + \ldots + \beta_k(x_k^* - x_k)}.$$

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- $\hookrightarrow$ We have already seen how a two-state exposure variable can be accommodated through the use of a binary variable X.
- $\hookrightarrow$ We can accommodate discrete exposure variables by using indicator (or *dummy*) variables.
- $\hookrightarrow$ For an exposure variable with K distinct levels, one level is first chosen as the baseline or reference group. We refer to this level as level 0.
- $\hookrightarrow$ The other $K-1$ levels are referred to as level 1, level 2, and so on up to level $K-1$.
- $\hookrightarrow$ We then define $K-1$ binary exposure variables as follows:
    - $\hookrightarrow$ $X_1 = 1$ if an individual's exposure is at level 1, and $X_1 = 0$ otherwise.
    - $\hookrightarrow$ $X_2 = 1$ if an individual's exposure is at level 2, and $X_2 = 0$ otherwise.
    - $\hookrightarrow$ $\ldots$
    - $\hookrightarrow$ $X_{K-1} = 1$ if an individual's exposure is at level $K-1$, and $X_{K-1} = 0$ otherwise.
- $\hookrightarrow$ For an individual at the baseline level, $X_1 = X_2 = \ldots = X_{K-1} = 0$.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- Having defined the indicator variables, we can use the multiple logistic regression model

$$\log\left(\frac{p_{x_1,\ldots,x_{K-1}}}{1-p_{x_1,\ldots,x_{K-1}}}\right)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_{K-1}x_{K-1}.$$

- Note that the intercept β_0 is the risk when $X_1 = X_2 = \ldots = X_{K-1} = 0$, that is, the log odds of D in the baseline exposure group.

- By definition, the first slope coefficient β_1 is the log odds ratio associated with a unit increase in X_1, holding $X_2, \ldots, X_{K-1}$ fixed.

- But the only way to increase X_1 by one unit and keep all other variables constant is to move from the baseline exposure level ($X_1 = 0$, all other X_ks are zero) to level 1 ($X_1 = 1$, all other X_ks are zero).

- Thus, β_1 is the log odds ratio comparing exposure level 1 to the baseline level 0.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- ↪ Similarly, β_j is the log odds ratio comparing level j to the baseline level 0 for $j = 1, \ldots, K-1$.

- ↪ Knowledge of the slope coefficients associated with the indicator variables again permit us to compute the log odds ratio that compare groups other than the baseline group.

- ↪ For example, the odds ratio comparing those in, say, level 3 (for which $X_3 = 1$ and $X_1 = X_2 = X_4 = \ldots = X_{K-1} = 0$), to those, say, in level 1 (for $X_1 = 1$ and $X_2 = X_3 = \ldots = X_{K-1} = 0$), is simply $e^{\beta_3 - \beta_1}$.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

↪ To make ideas concrete, let us look at data from the Western Collaborative Group Study, that conducted a form of population based study and that, among other information, collected information about body weight and incidence of coronary heart disease (CHD) (Jewell, 2003, p 194).

		CHD Event		
		D	not D	
Body weight (lb)	≤150	32	558	590
	150+–160	31	505	536
	160+–170	50	594	644
	170+–180	66	501	567
	>180	78	739	817
		257	2897	3154

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- In this example, the exposure variable has five categories and so we need four indicator variables.
- Let the baseline or reference group be formed by those who weigh 150 lb or less.
- The four indicator variables are defined as follows:
  - $X_1 = 1$ if body weight is 150^+ to 160 lb, and $X_1 = 0$ otherwise.
  - $X_2 = 1$ if body weight is 160^+ to 170 lb, and $X_2 = 0$ otherwise.
  - $X_3 = 1$ if body weight is 170^+ to 180 lb, and $X_3 = 0$ otherwise.
  - $X_4 = 1$ if body weight is $>$ 180 lb, and $X_4 = 0$ otherwise.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- The multiple logistic regression model takes the following form

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

- Here β_0 gives the log odds of CHD in the baseline group (≤ 150) lb.

- β_1 is the log odds ratio comparing the group who weigh between 150 and 160 lb to those who weigh less than or equal to 150 lb.

- β_2 is the log odds ratio comparing the group who weigh between 160 and 170 lb to those who weigh less than or equal to 150 lb.

- β_3 is the log odds ratio comparing the group who weigh between 170 and 180 lb to those who weigh less than or equal to 150 lb.

- Finally, β_4 is the log odds ratio comparing those who weigh more than 180 lb to those who weigh less than or equal to 150 lb.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

- $\hookrightarrow$ Further, and as an example, the log odds ratio comparing those who weigh more than 180 lb ($X_4 = 1$ and $X_1 = X_2 = X_3 = 0$) to those who weigh between 160 and 170lb ($X_2 = 1$ and $X_1 = X_3 = X_4 = 0$) is $\beta_4 - \beta_2$.

- $\hookrightarrow$ We will see how to estimate the parameters of the logistic regression model in the next lecture but for this simple example only involving a discrete exposure variable the data arranged in a 2×2 contingency table allow estimating the five parameters.

- $\hookrightarrow$ For example, β_0 is the log odds of CHD in the baseline group ($\leq$ 150) lb: $\log((32/590)/(558/590))$ = -2.859.

- $\hookrightarrow$ Analogously, the log odds of CHD in the group with body weight between 150 and 160lb, i.e., $X_1 = 1$, is $\log((31/536)/(505/536)) = -2.791$.

- $\hookrightarrow$ Thus, the log odds ratio comparing this weight level to the baseline group is just the difference in these log odds: $-2.791 - (-2.859) = 0.068$.

Regression models relating exposure to disease

The (multiple) logistic regression model: indicator variables for discrete exposures

↪ The estimates of the remaining parameters are obtained exactly in the same manner.

Parameter	Estimate	OR
β_0	-2.859	$--$
β_1	0.068	1.070
β_2	0.384	1.468
β_3	0.832	2.298
β_4	0.610	1.840