

CONTROLLING FOR CONFOUNDING IN CASE-CONTROL STUDIES

Parodi S.¹, Bottarelli E.²

Introduction

The case-control study is an observational epidemiological investigation, characterized by the measure of past exposure to risk factors, of two separate groups, named “cases” and “controls” (Breslow and Day, 1980; Kleinbaum *et al.*, 1982). For this reason it is also known as “retrospective study”. Information about past exposures and other factors possibly associated with the disease under study is generally collected by an anamnestic procedure, based on the administration of a questionnaire. Sometimes, information from genetic analyses or biological samples is also collected.

The case-control study differs from the cross-sectional study in the selection of cases and controls from two different populations. Moreover, it differs from the cohort (or follow-up) study, which gathers information on one population unaffected by the disease under study and followed up to detect the occurrence of a specific phenomenon (*e.g.*, incidence of diseases or mortality for specific causes, in general after the identification of an exposed sub-cohort).

Among the major advantages of the case-control study, which account for its large diffusion in environmental epidemiology, there are ethical reasons, because this study is directly applicable to human beings, since the exposures are not administered by an experimenter. Differently from cohort studies, it also allows: a) evaluation of the joint effect of many exposures; b) easy collection of information about many possible confounders. Moreover, it has a higher statistical power, as it includes a higher number of cases, especially when it is focused on rare diseases, that have a low incidence and can be evaluated only after a long observation period. Furthermore, it is less expensive and may be carried out in a shorter time. Despite these advantages, the case-control study has some limits, especially when compared with the cohort study: a) it is unsuited to evaluate the effect of rare exposures, even if such a limit is present only when the prevalence of exposed individuals in the cases series is low; b) it can only provide relative estimates of disease occurrence, even though it may provide incidence rates estimation, when population-based; c) it is prone to many biases, including the improper selection of either cases or (more commonly) controls (selection biases) and the different quality or completeness of information drawn from the cases compared to the controls (information biases). In particular, especially in human medicine, cases tend to correlate the onset of their disease with specific risk

¹ Epidemiology and Biostatistics Section, Scientific Directorate, G. Gaslini Children’s Hospital, Largo G. Gaslini 5, 16147 Genoa (Italy); e-mail: stefanoparodi@ospedale-gaslini.ge.it

² Università degli Studi di Parma, Dipartimento di Salute Animale. Via del Taglio 10, 43100 PARMA. e-mail: ezio.bottarelli@unipr.it

factors, thus reporting past exposures too much extensively with respect to controls. Such information bias is characteristic of case-control studies and it is known as “recall bias”. Finally, as in most observational studies and in many experimental studies, the presence of variables associated with both exposures and disease onset may induce a bias in the association estimates (confounding bias). Differently from selection and information biases, which are hard to control, the effect of confounding bias may be managed either in the phase of study design or during data analysis. The main strategies to control for confounding in case-control studies are illustrated in the next paragraphs, while their planning and implementation have been illustrated elsewhere (Parodi and Bottarelli, 2004).

Measures of association in the case-control study

In the simple case of a dichotomous exposure and in the absence of confounders, the result of a case-control study may be resumed in a 2x2 table as follows:

Table 1. Result of a case control study with exposure reported at two levels (either present or absent).

| | | Disease condition | | |
|----------|-----------|-------------------|----------|---------|
| | | Cases | Controls | Total |
| Exposure | Exposed | a | b | a+b |
| | Unexposed | c | d | c+d |
| | | a+c | b+d | a+b+c+d |

The most common estimator of association in case-control studies is the Exposure Odds Ratio (OR), which represents an estimate of the Risk OR and, therefore, is a relative risk estimator.

It may be calculated by the following simple equation:

$$\hat{OR} = \frac{a \cdot d}{b \cdot c} \quad (1)$$

Under the (null) hypothesis of no-effect of the exposure, the expected OR value equals 1, while under the hypothesis of a higher risk for the exposed group, this value will be higher. Finally, in the case of a protective effect of the exposure, OR will range between 0 and 1, and the risk factor should accordingly be called

“protective factor”.

If the sampling of cases and controls is not independently performed, in particular when controls are “matched “ to cases by some specific characteristic (like age or gender), as illustrated in the next paragraph, data will still be resumed in a 2x2 table, as follows:

Table 2. Result of a case-control study with matched data (matching ratio 1:1) and exposure at two levels (present or absent).

| | Exposed Controls | Unexposed Controls | Total |
|-----------------|---------------------|-----------------------|---------|
| Exposed Cases | A | B | A+B |
| Unexposed Cases | C | D | C+D |
| Total | A+C | B+D | A+B+C+D |

In that case, the OR estimate may be obtained looking at the couples of cases and controls with discordant appearance, *i.e.*, ignoring the exposed cases matched with the exposed controls and the unexposed cases matched with the unexposed controls. OR estimate is easily obtained as follows:

$$\hat{OR} = \frac{B}{C} \quad (2)$$

When actual data are analysed, *i.e.*, using a finite sample, it is necessary to evaluate if an OR different from 1 results either from an association between exposure and risk of disease, or from the random fluctuation related to sampling variability. Such evaluation is the object of the statistical inference, and it may be performed by calculating the confidence limits of the OR at a selected $1-\alpha$ value (in general, 95% or 90%). Some methods have been described in a previous paper (Parodi and Bottarelli, 2004). In this review, the Woolf method, which assumes a log-normal distribution for the OR, will be shortly illustrated by some numerical examples.

The estimate of the variance of the (natural) logarithm of an OR, obtained from a table like Table 1, may be calculated under the assumption of the independence of a, b, c and d figures (conditional to the number of cases and controls), as follows:

$$\hat{Var}[\ln(OR)] \cong \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad (3)$$

The corresponding confidence interval (CI) for the $\ln(OR)$ is easily computed,

and the corresponding CI of OR can be obtained by exponentiating these values:

$$95\%CI(OR) = OR \cdot e^{\mp z_{\alpha/2} \sqrt{\hat{V}ar[\ln(OR)]}} \quad (4)$$

where $z_{\alpha/2}$ is the critical value of the standardized normal distribution (*e.g.*, $z = 1.96$ for $\alpha = 0.05$). If the CI does not contain the expected OR value of 1, a statistically significant association between exposure and disease risk can be assumed.

In the case of a matched study (Table 2), the counts within A, B, C and D cells are not independent. The CI estimate of the corresponding OR may be obtained from the relation linking the binomial function to the Fisher's F distribution (Pearson and Hartley, 1966, cited in Parodi and Bottarelli, 2004). More simply, equation 4 may be applied using the following estimate of the variance of $\ln(OR)$, according to Silcocks (2005):

$$\hat{V}ar[\ln(OR)] \cong \frac{1}{B} + \frac{1}{C} \quad (5)$$

where counts B and C are extracted from Table 2.

Confounding and effect modifying

As already briefly mentioned, in case-control study and other epidemiological investigations, it is mandatory to control for the possible effect of extraneous variables, which might influence the outcome of the analyses when associated with both the exposure levels and the risk of disease. Such a phenomenon is known as "confounding" and the related variables are called "confounders". For example, both in animals and in human beings, ageing is associated with the risk of numerous diseases (*e.g.*, incidence of most cancers). If the exposure is not homogeneously distributed among different age groups, its effect may be hidden by the effect of ageing. Differently from the other biases (*i.e.*, selection and information biases), confounding may be controlled both during the study design and at the stage of data analysis. In experimental studies such a control may be performed by the randomization process, while in observational investigations, as the case-control study, the distribution of confounders may be measured (at least partly) by the researchers, who may thus reduce their effect by adequate strategies.

Figure 1 shows an example of the effect of a hypothetical confounder measured at two levels on the basis of the occurrence of a specific feature (*e.g.*, male or female gender, Caucasian or Afro-American ethnic group, etc.). When confounding is not taken into account, no association emerges between the risk estimator (which in the case-control studies is almost invariably the OR) and the exposure (Figure 1, left). After having identified the two confounder categories (corresponding to black

and white squares in Figure 1, right), a positive association emerges.

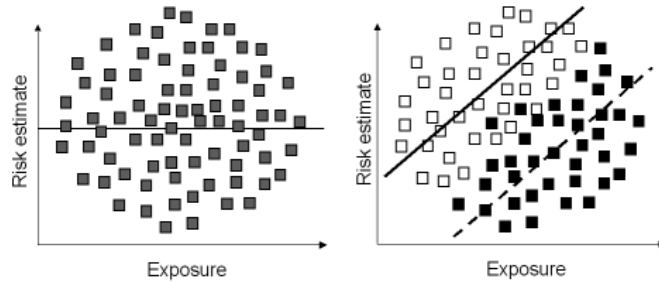


Figure 1 – Example of confounding in an epidemiological study

Another example of the possible effect of a confounder, when also the exposure is measured on a categorical scale, may be a (hypothetical) case-control study, where the proportion of exposed subjects and the disease risk (and, accordingly, the proportion of cases and controls) are associated with another categorical variable (*e.g.*, gender). In that case, the occurrence of confounding may be highlighted by comparing the estimates of association between risk and exposure obtained either using the pooled data set or after stratifying by the levels of the confounder. In particular, in a case-control study with dichotomous (*e.g.*, present/absent) exposure and confounding variable, such a comparison may be performed analyzing the following table:

Table 3. Result of a hypothetical case-control study without matching, using either pooled data (part a) or data stratified by a two-level confounder (part b and part c).

| | a) All subjects | | b) Stratum 1 | | c) Stratum 2 | |
|-----------|------------------------|----------|---|----------------|---|----------------|
| | Cases | Controls | Cases | Controls | Cases | Controls |
| Exposed | a | b | a ₁ | b ₁ | a ₂ | b ₂ |
| Unexposed | c | d | c ₁ | d ₁ | c ₂ | d ₂ |
| | OR _T =ad/bc | | OR ₁ = a ₁ d ₁ / b ₁ c ₁ | | OR ₂ = a ₂ d ₂ / b ₂ c ₂ | |

If OR_1 and OR_2 are similar, but differ from OR_T , confounding will occur. Conversely, if at least one is different from OR_T , but they also differ from one another, an interaction between the confounder and the exposure will occur, as illustrated further on in this paragraph. Finally, if OR_T , OR_1 and OR_2 are similar, there is no evidence of the presence either of confounding or of interaction.

Example 1

The following table shows the results of a hypothetical case control-study, where the only possible confounder was the sex of the subjects. Part a shows the result of the pooled analysis, while parts b and c report the results of the separate analysis for males and females, respectively. Under each OR estimate, the related 95% confidence intervals, obtained by the above described Woolf method, are reported in brackets.

| | a) All subjects | | b) Stratum 1 - Males | | c) Stratum 2 - Females | |
|-----------|-------------------------------------|----------|------------------------------------|----------|------------------------------------|----------|
| | Cases | Controls | Cases | Controls | Cases | Controls |
| Exposed | 139 | 96 | 99 | 89 | 40 | 7 |
| Unexposed | 64 | 45 | 8 | 20 | 56 | 25 |
| | OR _T = 1.0 (0.64,1.6) | | OR ₁ = 2.8 (1.2,6.6) | | OR ₂ = 2.6 (1.0,6.5) | |

Pooled analysis does not highlight any effect of the exposure, the corresponding OR estimate being close to its expected value (i.e., $OR_T=1.0$). Splitting the data set into two groups, two ORs, both above the expected value (ranging between 2.5 and 3) are observed, which suggests a positive association between risk and exposure. Please note that the confounding effect, which has completely masked the exposure effect in the pooled analysis, is due to the association between sex and both the risk indicator, i.e., the variable defining the cases and the controls, and the exposure variable. In fact, there are more cases than controls in females, whereas among males a 1:1 ratio is observed; moreover exposed males are prevalent among the exposed subjects (more than 80%), while unexposed males are only the 26% of the unexposed subjects.

OR estimates for males and females are similar, but not equal. For this reason, a formal statistical test should be performed to assess if the observed difference may merely be attributed to random noise. In such a case, the two ORs would represent an estimate of the same exposure effect, and a common estimate of OR should be calculated accordingly, together with its confidence intervals. Both procedures can be performed in an almost equivalent way, either by applying some stratified analysis methods or by a multivariate regression model. Among the former, the most largely applied technique is the Mantel-Haenszel method, while among the latter, logistic regression model is largely employed in case-control studies. Both approaches will be illustrated in the following paragraphs.

Effect modifying occurs whenever one or more variables and the exposure interact. For this reason, this phenomenon is also called “interaction”. Occurrence of interaction may be suggested by a different trend of the risk estimate within the confounder categories. In that case, due to the different association between risk and exposure in the two analysed groups, a common risk estimate, adjusted for the effect of such a variable, cannot be calculated. Then interaction and confounding should always be considered as two very different phenomena³. Occurrence of the effect modifying is in general considered as an important finding, while confounding often

regards not interesting variables. However, some exceptions may occur, for example when two (or more) important exposures act towards each other as confounders. In that case, as illustrated further on, a multivariate statistical model may be used to separate their effects.

Figure 2 shows an example of interaction or effect modifying. Splitting the subjects under study into the two categories of the effect modifier, a rise in the risk for the exposed subjects emerges in a subgroup (Figure 2, right, black squares), whereas a negative association is observed for another subgroup (white squares).

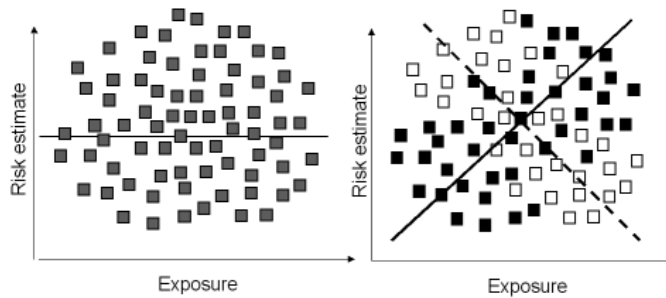


Figure 2 – Example of interaction or effect modifying

In the case of a categorical exposure variable (Table 3), if the ORs in the two strata differ, the stratification variable is suggested to modify the exposure effect, by interacting with the exposure and so changing the estimated risk of disease, as illustrated in Example 2.

Example 2

Let's now suppose that gender acts as an effect modifier of the exposure effect in a case-control study. Part a of the following table illustrates the pooled analysis of the whole data set, while parts b and c refer to the results of a separate analysis for males and females, respectively. Under each OR estimate, the related 95% confidence intervals, obtained by the Woolf method, are reported in brackets.

| | a) All subjects | | b) Stratum 1 - Males | | c) Stratum 2 - Females | |
|-----------|------------------------------------|----------|------------------------------------|----------|-------------------------------------|----------|
| | Cases | Controls | Cases | Controls | Cases | Controls |
| Exposed | 182 | 108 | 124 | 101 | 58 | 7 |
| Unexposed | 78 | 87 | 38 | 58 | 40 | 29 |
| | OR _T = 1.9 (1.3,2.8) | | OR ₁ = 1.8 (1.1,3.0) | | OR ₂ = 6.0 (2.4,15.1) | |

A statistically significant association emerges between exposure and the risk of disease, which remains almost unchanged when the analysis is restricted to

³ However, the same variable may behaves both as a confounder and an effect modifier.

males. On the contrary, in females the exposure impact seems to be enhanced, the related OR being three times higher than the corresponding figure in males. Then, these results suggest that an interaction between sex and exposure exists. However, the possibility that the different risk observed by gender was simply due to the high variability of the OR estimated among females (who represent the smallest group under analysis) could not be completely ruled out. An incorrect, but very common, approach to assess whether the ORs between the two sexes are significantly different, consists in comparing the related 95% confidence intervals, which in the case of example 2 overlap. Conversely, a correct approach includes a formal statistical test applied to a stratified analysis or using a regression model, as illustrated in the following paragraphs.

The occurrence of interaction may cause a reversal of the relative risk in the two categories of the effect modifier. In that case the interaction is called “qualitative”, while when the risks estimated in the two exposure categories are both either above or below 1, the interaction is called “quantitative”, because the effect modifier influences the strength of the association, but not its direction.

In some cases it is possible to make a distinction between a “synergistic” interaction, which occurs when the presence of the effect modifier enhances the impact of the exposure, and an “antagonist” interaction, when the effect modifier reduces the effect of the exposure. An example of synergistic interaction is the effect of the simultaneous exposure to asbestos and smoking habit on the risk of developing lung cancer, especially observed in occupational studies. In fact, workers exposed to both factors showed a much higher risk than workers exposed to only one factor, and such joint effect was clearly higher than the sum of the effects of the two single exposures (Nelson and Kelsey, 2002).

Main methods to control for confounding

The main strategies to control for confounding in observational epidemiological investigations and, in particular, in case-control studies, are: restriction, matching, stratification and fitting of regression models. Such methods are not mutually exclusive, because they may be combined to control for the effect of many confounders.

Restriction simply consists in the exclusion from the study of some individuals (for example, people who have an illness closely related to the disease characterizing the cases, as HIV seropositives in a study about pesticides exposure and risk of lymphomas). Restriction does not imply particular problems from a statistical point of view, but it represents a very delicate decision to be adopted in the phase of study design, because it unavoidably reduces the representativeness of the population under study.

Matching is a method for confounding control typical of case-control

studies. It consists on the selection of controls with characteristics homogeneous to those of the cases. A fixed number of controls is selected at random for each case, among the population (*e.g.*, people living in the same area or hospitalized in the same institution) showing one or more similar characteristics, like age, gender or region of residence. Matching may then be considered as a variant of the restriction, applied only to control selection. This procedure allows control for some confounders by enhancing the statistical power of the investigation. For such a reason, matching allows the execution of studies based on a smaller number of individuals than studies where controls are selected by simple randomization. Its major limit consists in its incapability of producing estimates of the effect of the confounders involved in matching. Moreover, non independent sampling of the statistical units must be taken into account during data analysis. Specific statistical methods have to be adopted accordingly, such as the conditional logistic regression model.

Individual matching is performed at the level of each subject, while frequency matching is based on groups of individuals. The latter is rather unusual. Moreover, matching is called “artificial” when the possible confounders for the matching process are selected by the researcher. On the contrary, it is called “natural” when cases and controls are matched on the basis of some natural characteristic, allegedly associated with most of the possible confounders. Examples of natural matching are those studies carried out on couples of twins, one affected by a specific disease (thus belonging to the group of cases) and the other unaffected (belonging to the controls).

Stratified analysis represents quite a simple method to control for the effect of one or more confounders. It consists in the measure of the association between the exposure and the risk of a specific disease within each level (also called “stratum”) of the confounder, as illustrated in Table 4. Strata may be the different levels of some confounder, when it is categoric (*e.g.*, sex, race), or groups obtained by aggregating the individuals in classes on the basis of the values of a continuous variable (*e.g.*, age). At the end of the stratification procedure, the homogeneity of the effects estimated per each stratum must be tested to assess whether interaction occurs. If interaction does not occur, a common measure of association should be obtained and its statistical significance should be evaluated by another formal test. As illustrated in the next paragraph, estimates of a common effect may be obtained as appropriate weighted means of the stratum-specific estimates.

Finally, regression models, in particular logistic regression, represent the most largely employed method to control for confounding in case-control studies, also providing an estimate of the joint effects of many confounders and exposure variables. Furthermore, risk estimates may also be obtained when a certain degree of data dispersion across the cells occurs (*i.e.*, when there are many sub-groups of exposures or confounders, like observation periods, age classes and gender). In that case, estimation procedure is based on the assumption of a mathematical relation between the variable under study (the risk of developing a disease) and the variables describing either the different exposures or the possible confounders.

Stratified analysis in case-control studies: the Mantel-Haenszel Odds Ratio

Mantel-Haenszel (MH) estimators, introduced in Epidemiology at the end of the 1950s (Mantel and Haenszel, 1959) have been largely applied in epidemiological studies (Silcocks, 2005). Recently they have been in part replaced by the generalized linear models (GLM), of which they represent, in a certain sense, a specific case, because they are largely based on the same statistical theory (likelihood theory).

In this paragraph, a brief mention will be made about the MH method to estimate a common OR, adjusted for the effect of possible confounders, in case-control studies. A characteristic of the MH OR (shared by all MH estimators) is the capability of producing consistent estimates even when one or more strata of the confounder lack observations.

Let K be the number of confounder strata in a case-control study. The association between exposure and risk of disease may be resumed in a table similar to Table 4. For each j stratum, an OR_j may be accordingly estimated as the product $(a_j d_j)/(c_j b_j)$ and the inference about its value can be made under an approximate log-normal distribution assumption, by applying the above illustrated Woolf method for the estimate of its variance, as follows:

$$\hat{Var}(\log OR_j) \cong \frac{1}{a_j} + \frac{1}{b_j} + \frac{1}{c_j} + \frac{1}{d_j} \quad (5)$$

It can be immediately noted that an estimate of the common OR obtained by averaging the stratum-specific OR_j is expected to yield not real values when 0 occurs in at least one c_j or b_j cell (*i.e.*, when there are neither exposed controls nor unexposed cases in at least one stratum of confounder). The same happens for the corresponding variance of its logarithm.

Table 4. Stratified analysis of data in a case-control study.

| Stratum (e.g., age classes) | Exposure condition | Cases | Controls | Total |
|--------------------------------|-----------------------|----------|----------|----------|
| 1 | Exposed | a_1 | b_1 | n_{11} |
| | Unexposed | c_1 | d_1 | n_{01} |
| | Total | m_{11} | m_{01} | n_1 |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| j | Exposed | a_j | b_j | n_{1j} |
| | Unexposed | c_j | d_j | n_{0j} |
| | Total | m_{1j} | m_{0j} | n_j |
| ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... |
| K | Exposed | a_K | b_K | n_{1K} |
| | Unexposed | c_K | d_K | n_{0K} |
| | Total | m_{1K} | m_{0K} | n_K |

Conversely, the MH OR (OR_{MH}) is obtained by the following formula:

$$\hat{OR}_{MH} = \frac{\sum_{j=1}^K \frac{a_j d_j}{n_j}}{\sum_{j=1}^K \frac{b_j c_j}{n_j}} \quad (6)$$

The corresponding test for the null hypothesis: $OR_{MH} = 1$, known as the “Mantel-Haenszel chi squared”, is expressed in the following equation:

$$\chi^2_{MH} = \frac{\left[\sum_{j=1}^K a_j - E\left(\sum_{j=1}^K a_j \right) \right]^2}{\sum_{j=1}^K \hat{Var}(a_j)} \quad (7)$$

where:

$$\hat{Var}(a_j) = \frac{m_{0j} m_{1j} n_{0j} n_{1j}}{n_j^2 (n_j - 1)} \quad (8)$$

In conclusion, the MH χ^2 test is equivalent in assessing whether the sum of a_j exposed cases significantly differs from its expected value.

A consistent estimate of the variance of $\ln(OR_{MH})$ can be obtained using the following formula, proposed by Robins, Breslow and Greenland (Robins et al, 1986; Silcocks P, 2005) :

$$\hat{Var}[\ln(\hat{OR}_{MH})] \cong \frac{\sum_{j=1}^K \frac{a_j d_j}{n_j} \cdot \frac{a_j + d_j}{n_j}}{2 \left(\sum_{j=1}^K \frac{a_j d_j}{n_j} \right)^2} + \frac{\sum_{j=1}^K \left(\frac{b_j c_j}{n_j} \cdot \frac{a_j + d_j}{n_j} + \frac{b_j + c_j}{n_j} \cdot \frac{a_j d_j}{n_j} \right)}{2 \left(\sum_{j=1}^K \frac{a_j d_j}{n_j} \right) \left(\sum_{j=1}^K \frac{b_j c_j}{n_j} \right)} + \frac{\sum_{j=1}^K \frac{b_j c_j}{n_j} \cdot \frac{b_j + c_j}{n_j}}{2 \left(\sum_{j=1}^K \frac{b_j c_j}{n_j} \right)^2} \quad (9)$$

Assuming an approximate log-normal distribution of the OR_{MH} , its corresponding confidence intervals at a selected $1-\alpha$ level may be obtained by the following equation, similar to equation 4:

$$\left[\hat{OR}_{MH} e^{\mp z_{\alpha/2} \sqrt{V\hat{ar}[\ln(\hat{OR}_{MH})]}} \right] \tag{10}$$

OR_j homogeneity, *i.e.*, the occurrence of interaction, may be formally tested via the following χ^2 test with K-1 degrees of freedom (Breslow and Day, 1980):

$$\chi^2 = \sum_{j=1}^K \frac{[\ln(\hat{OR}_j)]^2}{V\hat{ar}[\ln(\hat{OR}_j)]} - \frac{\left\{ \sum_{j=1}^K \frac{\ln(\hat{OR}_j)}{V\hat{ar}[\ln(\hat{OR}_j)]} \right\}^2}{\sum_{j=1}^K \frac{1}{V\hat{ar}[\ln(\hat{OR}_j)]}} \tag{11}$$

where the variance of ln(OR_j) is obtained by equation 5.

Example 3

Data from examples 1 and 2 may be re-analysed by the MH method to provide a common estimate of the ORs, adjusted for the effect of the confounding variable (Sex). A test checking for a possible effect modifying of such a variable is also performed.

As regards the data in the Example 1, the common estimate of relative risk (*i.e.*, the OR_{MH}) is obtained as follows:

$$\hat{OR}_{MH} = \frac{\frac{99 \cdot 20}{216} + \frac{40 \cdot 25}{128}}{\frac{89 \cdot 8}{216} + \frac{7 \cdot 56}{128}} = 2.67$$

Applying equation 9, the variance of ln(OR_{MH}) is:

$$V\hat{ar}[\ln(\hat{OR}_{MH})] = \frac{\frac{99 \cdot 20}{216} \cdot \frac{99+20}{216} + \frac{40 \cdot 25}{128} \cdot \frac{40+25}{128} + \frac{89 \cdot 8}{216} \cdot \frac{99+20}{216} + \frac{89+8}{216} \cdot \frac{99 \cdot 20}{216} + \frac{7 \cdot 56}{128} \cdot \frac{40+25}{128} + \frac{7+56}{128} \cdot \frac{40 \cdot 25}{128}}{2 \cdot \left(\frac{99 \cdot 20}{216} + \frac{40 \cdot 25}{128} \right)^2} + \frac{\frac{89 \cdot 8}{216} \cdot \frac{89+8}{216} + \frac{7 \cdot 56}{128} \cdot \frac{7+56}{128}}{2 \cdot \left(\frac{89 \cdot 8}{216} + \frac{7 \cdot 56}{128} \right)^2} = 0.1051$$

Furthermore, applying equation 10, the 95% confidence intervals of OR_{MH} are obtained as follows:

$$\left[2.67 \cdot e^{\mp 1.96 \sqrt{0.1051}} \right] \rightarrow [1.41; 5.04]$$

Finally, the homogeneity of the stratum-specific ORs (indicated as OR_1 for the males and OR_2 for the females, respectively, accordingly to Example 1) may be tested by equation 11. However, before applying such a formula, the variance of the corresponding logarithms must be estimated, by applying equation 12:

$$\hat{V}ar[\ln(OR_1)] \cong \frac{1}{99} + \frac{1}{89} + \frac{1}{20} + \frac{1}{8} = 0.1963$$

$$\hat{V}ar[\ln(OR_2)] \cong \frac{1}{40} + \frac{1}{7} + \frac{1}{25} + \frac{1}{56} = 0.2257$$

Putting such estimates into equation 11, the χ^2 test for homogeneity is obtained:

$$\chi^2 = \frac{[\ln(2.78)]^2}{0.1963} + \frac{[\ln(2.55)]^2}{0.2257} - \frac{\left[\frac{\ln(2.78)}{0.1963} + \frac{\ln(2.55)}{0.2257} \right]^2}{\left(\frac{1}{0.1963} + \frac{1}{0.2257} \right)} = 0.018$$

The observed value is clearly lower than the critical value of χ^2 distribution with 1 degree of freedom (i.e., 3.84). As a consequence, it can be stated that there is no evidence of an interaction between sex and exposure.

As far as data from Example 2 are concerned, the OR_{MH} estimate is:

$$\hat{O}R_{MH} = \frac{\frac{124 \cdot 58}{321} + \frac{58 \cdot 29}{134}}{\frac{101 \cdot 38}{321} + \frac{7 \cdot 40}{134}} = 2.49$$

and the variance of its logarithm:

$$\begin{aligned} \hat{V}ar[\ln(\hat{OR}_{MH})] &\cong \frac{\frac{124 \cdot 58}{321} \cdot \frac{124+58}{321} + \frac{58 \cdot 29}{134} \cdot \frac{58+29}{134} + \frac{101 \cdot 38}{321} \cdot \frac{124+58}{321} + \frac{101+38}{321} \cdot \frac{124 \cdot 58}{321} + \frac{7 \cdot 40}{134} \cdot \frac{58+29}{134} + \frac{7+40}{134} \cdot \frac{58 \cdot 29}{134}}{2 \cdot \left(\frac{124 \cdot 58}{321} + \frac{58 \cdot 29}{134} \right)^2} + \frac{\frac{101 \cdot 38}{321} \cdot \frac{101+38}{321} + \frac{7 \cdot 40}{134} \cdot \frac{7+40}{134}}{2 \cdot \left(\frac{124 \cdot 58}{321} + \frac{58 \cdot 29}{134} \right) \cdot \left(\frac{101 \cdot 38}{321} + \frac{7 \cdot 40}{134} \right)} \\ &+ \frac{\frac{101 \cdot 38}{321} \cdot \frac{101+38}{321} + \frac{7 \cdot 40}{134} \cdot \frac{7+40}{134}}{2 \cdot \left(\frac{101 \cdot 38}{321} + \frac{7 \cdot 40}{134} \right)^2} = 0.0462 \end{aligned}$$

Finally, the related 95% confidence interval of OR_{MH} is:

$$\left[2.49 \cdot e^{\mp 1.96 \cdot \sqrt{0.0462}} \right] \rightarrow [1.63; 3.79]$$

The variances of the logarithm of the estimated stratum-specific ORs are:

$$\hat{V}ar[\log(OR_1)] \cong \frac{1}{124} + \frac{1}{101} + \frac{1}{58} + \frac{1}{38} = 0.0615$$

$$\hat{V}ar[\log(OR_2)] \cong \frac{1}{58} + \frac{1}{7} + \frac{1}{29} + \frac{1}{40} = 0.2196$$

Putting such estimates in equation 11, the value of the test for the OR homogeneity is obtained as follows:

$$\chi^2 = \frac{[\log(1.87)]^2}{0.0615} + \frac{[\log(6.01)]^2}{0.2196} - \frac{\left[\frac{\log(1.87)}{0.0615} + \frac{\log(6.01)}{0.2196} \right]^2}{\left(\frac{1}{0.0615} + \frac{1}{0.2196} \right)} = 4.849$$

The observed value exceeds the critical one for $\alpha = 0.05$, then it may be concluded that there is evidence of an interaction between sex and exposure, then the common OR estimate (OR_{MH}) is not suitable to describe the impact of the exposure. Separate analysis by gender should instead be adopted.

An introduction to the generalized linear models

The analysis of epidemiological data often makes use of statistical models, which represent a reduction and an analogy of the actual world (Piccolo, 2000).

In general, the application of a statistical model needs three conceptual phases:

1) model specification: a relation between the variables under study is hypothesized and made explicit in a mathematical way. A statistic (stochastic) model differs from a deterministic one, because the relation between the considered variables is probabilistic;

2) parameter estimation: estimates of resuming variables, called “parameters”, which are the object of scientific investigation, may be obtained by means of specific procedures from the observed sample (*e.g.*, time trend of incidence rates, estimate of a relative risk between two groups, as the OR, and so on). Parameters of a statistical model applied to epidemiological data may be used both to estimate the occurrence of a phenomenon, like the incidence or the prevalence of specific diseases (estimates of disease occurrence) and its association with some risk factors (estimates of association or effect);

3) checking for goodness-of-fit of the model: it comprises a group of procedures aimed at assessing the general fit of the model to experimental data, looking for any violation of the assumptions underlying the parameters estimate, and identifying extreme values (called “outliers”) that may have influenced parameters estimates.

In spite of this importance, point 3 will not be discussed in this paper for lack of space.

Among statistical models applied to Epidemiology, regression models are largely employed for their conceptual simplicity. In fact, a regression model relates one variable (*e.g.*, the odds of exposure, whose ratio between cases and controls represents the OR) with one or more variables, in general (but not only) exposure measures and confounders. A dependent variable is then defined, which is assumed to be associated with one or more independent variables. The dependent variable is also called “the response” and independent variables “the predictors”.

During the last decades, generalized linear models (GLM), which include the logistic regression model (commonly applied to case-control studies) have largely spread in Epidemiology. Theory at the basis of GLM is very complex and the reader interested in it may refer to Dobson (1990) or McCullagh and Nelder (1989). In Appendix I a short mention of the likelihood theory, which is at the basis of such a statistical modelling, is provided. However, in spite of their complexity and the need for methods of sophisticated automatic computation, some GLMs may fruitfully be applied to epidemiological data, just knowing some of their statistical properties and without focusing on theoretical details.

A GLM is defined by specifying the following characteristics: (Dobson, 1990; Piccolo, 2000):

1) error function: it must belong to a group of functions, called “the exponential family”, which includes Normal, Binomial, Poisson and many other

distributions;

2) variance of the response variable: it is put in connection with the mean value μ by a variance function:

$$Var(Y) = \phi v(\mu) \quad (11)$$

The variance of the response variable Y is assumed to be proportional either to its mean value μ or to its transformation v. The proportionality constant ϕ is called “scale parameter”;

3) the expected (mean) value of the response variable depends on a linear function of parameters:

$$E(Y) = f(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p) \quad (12)$$

The linear function of parameters: $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ is called “linear predictor”, and it forms the deterministic component of the model, while the function f, which relates the deterministic part with the stochastic one, is called “link function”.

The Logistic Regression Model

The logistic regression model represents one of the most largely applied techniques to model binary variables. It is largely used in Epidemiology, especially in case-control studies, to obtain OR estimates, adjusted for the effect of confounding variables (Breslow and Day, 1980). It is also employed in cross-sectional studies to model prevalence OR. Furthermore, it can also be used in clinical investigation, allowing the modelling of the association between a specific condition (*e.g.*, ill vs. healthy) and classification variables (*e.g.*, occurrence of a clinical pathologic value).

The logistic model was firstly applied to an epidemiological framework by Jerome Cornfield, who is considered as one of the fathers of the modern case-control study. Cornfield was the first to demonstrate, both from a theoretical point of view and using data from a cohort study, the equivalence between the exposure OR (equation 1) and the relative risk (Cornfield, 1951). Afterward, the logistic model had a great evolution, towards models for the analysis of matched data (conditional logistic regression) and for ordinal data (continuation ratio model, cumulative odds model). A complete treatment of the different applications of the logistic model and its variants may be found in the book by Hosmer and Lemeshow (2000).

The simplest formulation of an unconditional logistic regression model assumes the following relation between the response variable Y, which may only take

either 0 or 1 values (*e.g.*, corresponding to the disease status, *i.e.*, case or controls) and a set of predictors x :

$$E(Y|\underline{\beta}, \underline{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}} \quad (13)$$

This model also assumes the independence of the observations, and it is accordingly applied to case-control studies without matching. A variant of this model exists, that is usually applied to studies with matching.

The logistic model can be formulated as a GLM assuming for the linear predictor the following (usual) formula: $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, for the function error the binomial distribution (because the response variable is binary) and as link function the logit function, which corresponds to the logarithm of the odd of Y (*i.e.*, the probability of Y divided by its complement to 1). In the framework of the case-control study, the response variable is in general coded 0 when the subject belongs to the controls and 1 when the subjects belongs to the cases. Letting the predictor E represent the exposure of interest, which assumes the values of 0 and 1, respectively in the absence and in the presence of the exposure, an estimate of OR adjusted for the possible effect of other covariates may be obtained by exponentiating the corresponding estimated coefficient $\hat{\beta}_1$. In fact, when only one confounder C is included in the model:

$$\hat{Y} = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 C)}} \quad (14)$$

Applying the logit transformation to the Y variable, the following equation is obtained:

$$\text{logit}(\hat{Y}) = \ln\left(\frac{\hat{Y}}{1 - \hat{Y}}\right) = \ln\left(\frac{\frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 C)}}}{1 - \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 C)}}}\right) = \ln\frac{1}{e^{-(\hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 C)}} = \hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 C$$

The expected value of the logit of Y for the unexposed group (corresponding to E=0), will be:

$$\ln\left(\frac{\hat{Y}_{NE}}{1 - \hat{Y}_{NE}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 C_{NE} = \hat{\beta}_0 + \hat{\beta}_2 C_{NE} \quad (15)$$

where the NE subscript in the response and confounder variables indicates that the logit estimate is performed within the unexposed subgroup.

Likewise, the expected value of logit in the exposed subgroup (*i.e.*, $E = 1$), will be:

$$\ln\left(\frac{\hat{Y}_E}{1-\hat{Y}_E}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 C_E = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 C_E \quad (16)$$

Subtracting equation 15 from equation 16:

$$\log\left(\frac{\hat{Y}_E}{1-\hat{Y}_E}\right) - \log\left(\frac{\hat{Y}_{NE}}{1-\hat{Y}_{NE}}\right) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 C_E - \hat{\beta}_0 - \hat{\beta}_2 C_{NE} = \hat{\beta}_1 + \hat{\beta}_2 (C_E - C_{NE})$$

Finally, for equal values of C confounder in the two subgroups (*i.e.*, when $C_E = C_{NE}$), and remembering that the difference between the logarithm of two values equals the logarithm of their ratio, the following equation is obtained:

$$\ln\left(\frac{\frac{\hat{Y}_E}{1-\hat{Y}_E}}{\frac{\hat{Y}_{NE}}{1-\hat{Y}_{NE}}}\right) = \hat{\beta}_1 \longrightarrow \exp(\hat{\beta}_1) = OR \quad (17)$$

Inference about Maximum Likelihood Estimators

After obtaining some maximum likelihood estimates (MLE) from a logistic regression model (as shortly illustrated in Appendix I), it is necessary to test whether such estimates fit a specific hypothesis (*e.g.*, in general, whether the coefficients for the exposures are different from 0, which implies that the corresponding OR significantly differs from 1). A method similar to that applied to linear regression models consists in dividing the difference between an estimated coefficient and its expected value, under a specific (null) hypothesis, by the standard error of such a difference (corresponding to the standard error of the estimated coefficient). In the GLM, this test (Wald test) has an asymptotic approximate standardized normal distribution:

$$z = \frac{\hat{\beta} - \beta_{H0}}{\sqrt{VAR(\hat{\beta})}} \quad (18)$$

The Wald test also makes it possible to obtain the confidence intervals for the parameter, by the usual way:

$$CI(\hat{\beta}; 1 - \alpha): \left[\hat{\beta} - z_{\alpha} \sqrt{VAR(\hat{\beta})}; \hat{\beta} + z_{\alpha} \sqrt{VAR(\hat{\beta})} \right] \quad (19)$$

where z_{α} represents the critical value of the standardized normal distribution at a specific α value. The so obtained confidence intervals may be used to make statistical inference.

Another test, more accurate than the Wald test, is based on the ratio between the likelihood of the MLE estimate of the OR and the related likelihood that will be obtained under the null hypothesis. That is briefly illustrated in Appendix I.

Testing the interaction by a logistic regression modelling

If variable M was an effect modifier, different OR values would correspond to different levels of M. A new variable (interaction variable), obtained as the product between the exposure E and M, may be introduced in the model to test such interaction:

$$\log it(\hat{Y}) = \hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 M + \hat{\beta}_3 EM \quad (20)$$

The logit in the exposed subgroup is:

$$\ln\left(\frac{\hat{Y}_E}{1 - \hat{Y}_E}\right) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 M + \hat{\beta}_3 M \quad (21)$$

while the logit among the unexposed is:

$$\log\left(\frac{\hat{Y}_{NE}}{1 - \hat{Y}_{NE}}\right) = \hat{\beta}_0 + \hat{\beta}_2 M \quad (22)$$

Subtracting equation 22 from equation 21 the following relation between the $\ln(\text{OR})$ and M is obtained:

$$\ln \left(\frac{\frac{\hat{Y}_E}{1 - \hat{Y}_E}}{\frac{\hat{Y}_{NE}}{1 - \hat{Y}_{NE}}} \right) = \ln(OR) = \hat{\beta}_1 + \hat{\beta}_3 M \quad (23)$$

Applying equation 23, different OR values may be obtained, corresponding to different values of the effect modifier M. Let (for simplicity) M be a binary variable, assuming value 0 when the effect modifier is absent and 1 when it is present. Two different ORs estimates will be accordingly calculated, i.e.: $\hat{OR}(M = 0) = e^{\hat{\beta}_1}$, and $\hat{OR}(M = 1) = e^{\hat{\beta}_1 + \hat{\beta}_3}$.

When $\beta_3 = 0$ there is no interaction between M and E, then the occurrence of effect modifying may be assessed by testing the statistical significance of this coefficient, *e.g.*, by Likelihood Ratio Test. Contrary to interaction, confounding occurrence cannot be assessed by a test on the related coefficient (*i.e.*, β_2 in the equation 14). In fact, this coefficient just measures the association between the variable C and the predictor Y, but C behaves as a confounder only if it is also associated with the exposure. Formally testing the confounding occurrence is feasible in theory, but generally useless, because confounding does not represent an interesting phenomenon for the researcher, who just wants to reduce its impact. Conversely, effect modifying and its interpretation often represent an issue of great interest from a bio-medical point of view.

Example 4

The analysis of data used in Example 1 may also be performed by applying a logistic regression model (in the present example, using STATA for Windows statistical package). Including only the exposure variable among the predictors (univariate model) the estimate of the pooled OR is obtained without adjusting for the confounding effect of Sex.

```

Logit estimates                                     Number of obs   =       344
                                                    LR chi2(1)      =         0.01
                                                    Prob > chi2     =       0.9394
Log likelihood = -232.82188                         Pseudo R2      =       0.0000

```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|------|------------|-----------|------|-------|----------------------|
| Esp | 1.018066 | .2397501 | 0.08 | 0.939 | .6416851 1.615215 |

The software automatically provided the OR estimates obtained by exponentiating the estimates of the coefficients for each predictor. Moreover, it also provided the related 95% confidence intervals (i.e., 0.64;1.62), by exponentiating the confidence intervals of the estimated coefficient, obtained by the Wald method (equation 19). Please note that results of the analysis by the logistic regression model are almost identical to those obtained by the Woolf method.

The statistical software also provided the log-likelihood of the model, the total number of observations, the LR test performed comparing the fitted model with the corresponding model including only the intercept (indicated as “LR chi2”), the related statistical significance (reported as “Prob > chi2”) and, finally, the pseudo-R², which represents a goodness-of-fit statistic, analogous to the R² of the linear regression model. Fitting the model with both predictors (called “main effect model”, to distinguish it from the interaction model and other more complex models), the exposure OR estimate, adjusted for the effect of the sex variable, is obtained.

```

Logit estimates                                     Number of obs =      344
                                                    LR chi2(2)      =      32.36
                                                    Prob > chi2     =      0.0000
Log likelihood = -216.64513                       Pseudo R2      =      0.0695
    
```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|-------|------------|-----------|------|-------|----------------------|
| Esp | 2.672364 | .8662628 | 3.03 | 0.002 | 1.415707 5.044495 |
| Sesso | 5.341609 | 1.731241 | 5.17 | 0.000 | 2.830046 10.08209 |

Please note that including the Sex predictor in the model, the estimate of exposure effect (i.e., the OR for Esp variable, which represents the exposure) changes from 1.0 to 2.7 and reaches statistical significance (as highlighted by the 95% confidence interval: 1.42; 5.04, which does not include the expected value of 1). This finding indicates that the exposure effect was masked by the different proportion of males and females among cases and controls and by the association of the Sex variable with the exposure, which is made evident by the OR estimated value for the variable itself (5.3, statistically significant). Please note that results of this analysis completely overlap those obtained by stratifying analysis using the MH method (first part of Example 3).

Finally, to assess the occurrence of effect modifying, as illustrated above, it is sufficient to introduce in the model an interaction term, equivalent to the product of the variables corresponding to the main effect (i.e., Sex and Exposure).

```

Logit estimates                                     Number of obs =      344
                                                    LR chi2(3)      =      32.38
                                                    Prob > chi2    =      0.0000
Log likelihood = -216.63631                       Pseudo R2      =      0.0695

```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|--------------|------------|-----------|-------|-------|----------------------|
| _IEsp_1 | 2.780899 | 1.232214 | 2.31 | 0.021 | 1.166859 6.627534 |
| _ISesso_1 | 5.6 | 2.702295 | 3.57 | 0.000 | 2.174895 14.41908 |
| _IEspXSes_~1 | .9173366 | .5959521 | -0.13 | 0.894 | .2567684 3.277297 |

The OR estimates, automatically provided by the output of the program, are not easily interpretable, because the occurrence of interaction indicates that two separated estimates of ORs should be made in each stratum of the effect modifier. Moreover, to obtain the statistical significance of the interaction, 95% confidence interval for the related exponentiated coefficient (incorrectly indicated as OR) may be compared with the expected value under the null hypothesis (i.e., 1). Otherwise, a LR test may be performed, as briefly described at the end of Appendix I.

```

Logistic: likelihood-ratio test                    chi2(1)      =      0.02
                                                    Prob > chi2 =      0.8944

```

Please note that the result (i.e., $\chi^2 = 0.02$, test for the interaction) completely overlap that obtained by MH method (Example 3).

The results of the same approach applied to the data of Example 2, where a statistically significant interaction emerged, are provided below.

Univariate analysis: model with only one variable for the exposure:

```

Logit estimates                                     Number of obs =      455
                                                    LR chi2(1)      =      10.27
                                                    Prob > chi2    =      0.0014
Log likelihood = -305.58979                       Pseudo R2      =      0.0165

```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|------|------------|-----------|------|-------|----------------------|
| Esp | 1.87963 | .371523 | 3.19 | 0.001 | 1.275927 2.768973 |

Main effect model for the control of confounding due to the Sex variable:

```

Logit estimates                               Number of obs   =      455
                                                LR chi2(2)      =      39.46
                                                Prob > chi2     =      0.0000
Log likelihood = -290.99137                    Pseudo R2      =      0.0635
    
```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|-------|------------|-----------|------|-------|----------------------|
| Esp | 2.505299 | .5396665 | 4.26 | 0.000 | 1.642488 3.821351 |
| Sesso | 3.437588 | .8269118 | 5.13 | 0.000 | 2.145346 5.508208 |

Model with an interaction term to assess the effect modifying of Sex:

```

Logit estimates                               Number of obs   =      455
                                                LR chi2(3)      =      44.69
                                                Prob > chi2     =      0.0000
Log likelihood = -288.37866                    Pseudo R2      =      0.0719
    
```

| CaCo | Odds Ratio | Std. Err. | z | P> z | [95% Conf. Interval] |
|--------------|------------|-----------|------|-------|----------------------|
| _IEsp_1 | 1.873893 | .4647959 | 2.53 | 0.011 | 1.15243 3.047016 |
| _ISesso_1 | 2.105263 | .675786 | 2.32 | 0.020 | 1.122205 3.949487 |
| _IEspXses_~1 | 3.205703 | 1.699638 | 2.20 | 0.028 | 1.134026 9.061989 |

LR test for the statistical significance of the interaction term:

```

Logistic: likelihood-ratio test                chi2(1)        =      5.23
                                                Prob > chi2    =      0.0223
    
```

The advantages of applying a logistic regression model rather than the MH method, become evident when the effect of many exposure variables have to be modelled in the presence of many confounders. Furthermore, the logistic model allows an estimate of the effect of continuous variables, if any, whereas stratifying analysis, like the MH method, does not. However, the MH method may be applied by a pocket calculator, while fitting a logistic regression model needs dedicated statistical software.

Appendix I

An outline of likelihood theory applied to logistic regression model

The concept of likelihood is central in theoretical statistics, because it provides the basis of the estimator theory, which plays a fundamental role in statistical inference. Likelihood is defined as the probability of some unknown parameters given the observed data, under an hypothesis of probabilistic distribution⁴.

In epidemiology, binomial function may be employed to model the distribution of the proportion of sick subjects (either prevalent or new cases of a specific disease, the latter called “incident” cases) in different study designs. In particular, in case-control studies, a binomial distribution may be assumed for the proportion of cases within the different categories of exposure and confounding variables. Binomial function has the following expression for the probability of observing y events (for example, cases of a disease) in n trials (*e.g.*, the number of subjects within an exposure group):

$$P(Y = y|n, \pi) = \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}$$

The likelihood of $\pi|y, n$ is obtained as a function of the y observed events, as follows:

$$L(\pi|y, n) = \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}$$

Since n and y are fixed, likelihood being calculated from observed data, the likelihood of a binomial distribution may be more simply expressed in the following way:

$$L(\pi|y, n) = \pi^y (1-\pi)^{n-y}$$

(A.1)

Formally, an estimated parameter $\hat{\pi}$ is a maximum likelihood estimator if:

$$L(Y|\hat{\pi}) > L(Y|\pi^*)$$

where π^* represents any other parameter.

⁴ Sometimes, likelihood is defined up to a specific constant, basically for computational reasons.

Applying equation A.1 to data from a case-control study, an estimate of the likelihood in each subgroup under analysis (e.g., older exposed males, or older unexposed females, *etc.*) is easily obtained.

According to the theorem of independent probabilities, the likelihood of the logistic model may simply be obtained as the product of the likelihoods of each subgroup. For many reasons, including some computational ones, instead of directly modelling the likelihood function, its logarithmic transformation (log-likelihood) is commonly used. According to a known property of the logarithm function, log-likelihood of the whole model will be equal to the sum of the log-likelihoods of each subgroup. Finally, the relationship linking the log-likelihood to the β parameters to be estimated is obtained by replacing the unknown parameter π with the logistic function (equation 13).

After having obtained (not in a trivial way) the estimate of the coordinates of the point of the maximum of the likelihood function, values of the β coefficients corresponding to this point are calculated, which provide the maximum likelihood estimates (MLEs) of such parameters.

Formally, the log-likelihood of a logistic model (and of any other model with a binomial error) is:

$$l = \sum_j y_j \ln \hat{\pi}_j + (m_j - y_j) \ln(1 - \hat{\pi}_j) \quad (\text{A.2})$$

where j indicates the different subgroups, m_j the corresponding number of observed subjects and y_j the number of the cases.

Moreover, in the logistic model, the following relation between the π parameter and the linear predictor is assumed (according to equation 13):

$$\hat{\pi}_j = E(Y | \underline{\beta}, \underline{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p)}}$$

MLEs of β coefficients and, accordingly, those of the unknown parameter π are obtained by identifying the maximum value of the function l .

The difference between two log-likelihoods corresponds to the logarithm of the ratio of related likelihood functions. Based on this property, a statistical test (called the ‘‘Likelihood Ratio Test’’) may be obtained to assess the statistical significance of one or more selected coefficients in a GLM, including logistic regression models. In fact, the double of the difference between the log-likelihood of the model containing the variable, whose effect has to be tested, and the model without such variable, follows (asymptotically) a χ^2 distribution with 1 degree of freedom under the null hypothesis of no effect (*i.e.*, when the corresponding coefficient does not differ from

O). If the test is performed on more than one variable, the assumption of χ^2 distribution still holds, while the corresponding degrees of freedom will be equal to the number of the variables left out from the model. Differently from many other statistical tests applied to epidemiological and bio-medical research, likelihood ratio test does not employ any estimate of the variance of the parameters.

Acknowledgements

This work was partly supported by a grant from the Fondazione Italiana Neuroblastoma. We thank Dr Anna Capurro for revising the English.

References

BRESLOW N.E., DAY N.E.: Statistical Methods in Cancer Research – Volume 1 – The analysis of case-control studies. IARC Scientific Publications N. 32, Lyon, 1980.

CORNFIELD J.: A method for estimating comparative rate from clinical data; Applications to cancer of the lung, breast and cervix - J. Natl. Cancer Inst., 11:1269-75, 1951;.

DOBSON A.J.: An introduction to generalized linear models. Chapman & Hall, New York, 1990.

HOSMER W., LEMESHOW S.: Applied Logistic Regression – Second Edition. John Wiley and Son, New York, 2000.

KLEINBAUM D.G., KUPPER L.L., MORGENSTERN H.: Epidemiologic research: principles and quantitative methods. John Wiley & Sons, Inc., New York, 1982.

MANTEL N., HAENSZEL, W.: Statistical aspects of the analysis of data from retrospective studies of disease. - J Natl. Cancer Inst., 22: 719-748, 1959.

MCCULLAGH P., NELDER J.A.: Generalized Linear Models - Chapman and Hall, 2nd edition, New York, 1989.

NELSON H.H., KELSEY K.T. The molecular epidemiology of asbestos and tobacco in lung cancer - Oncogene.21(48), 7284-8, 2002.

PARODI S., BOTTARELLI E.: Introduzione allo studio caso-controllo in epidemiologia - Annali Fac. Medic. Vet. di Parma, XXIV, 209, 2004.

PEARSON E.S., HARTLEY H.O.: Biometrika Tables for Statisticians, Vol.I (3rd

Edition), Cambridge University Press, Cambridge, (UK), 1966.

PICCOLO D.: Introduzione ai modelli statistici. In: Statistica. Il Mulino, Bologna 2000, pagg. 827-921.

ROBINS J., BRESLOW N., GREENLAND S.: Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models - Biometrics 1986, 42:311-323.

SILCOCKS P.: An easy approach to the Robins-Breslow-Greenland variance estimator - Epidemiologic Perspectives & Innovations, 2:9, 2005.

Key words

Case control studies, bias, confounders, logistic regression, Maximum Likelihood Estimates

SUMMARY

The case-control study is one of the main observational epidemiological investigations. It collects and compares information about past exposures of two different groups: the cases, which are subjects (or animals) affected by the disease under study, and the controls, which are unaffected by the disease. Such a study design is characterized by numerous advantages in terms of statistical power and feasibility (low cost and short execution time), and it can be applied to make causal inference, because, under proper conditions, it can assess a causal relation between one or more exposures and the risk of developing selected diseases. In spite of such undeniable advantages, the case-control study is more prone than prospective investigations to many biases, which may be classified into three categories: selection bias, information bias and confounding bias. The first two may be controlled only by selecting a proper study design, while confounding bias may be partly overcome by (a) matching or restriction in the phase of study design, and (b) statistical modelling in the phase of data analysis. In this paper the main methods to control for confounding in case-control studies are illustrated. In particular, some techniques of stratified (Mantel-Haenszel method) and multivariate (logistic regression) analysis are described. Many examples are provided using simulated data sets.

RIASSUNTO

Lo studio caso-controllo rappresenta uno dei principali studi epidemiologici osservazionali. La sua caratteristica principale risiede nel reperimento dell'informazione pregressa sull'avvenuta esposizione in due gruppi distinti a

confronto: i casi, ovvero soggetti affetti dalla patologia in esame e i controlli, ovvero soggetti non affetti dalla patologia di interesse. Tale disegno di studio presenta notevoli vantaggi in termini di potenza statistica, fattibilità (bassi costi e brevi tempi di realizzazione) ed è in grado di compiere inferenza causale, ovvero, sotto opportune condizioni, di stabilire la presenza di un nesso causale tra una o più esposizioni e il rischio di contrarre una determinata patologia. A fronte degli innegabili vantaggi, rispetto agli studi prospettici lo studio caso-controllo risulta maggiormente vulnerabile a fattori di distorsione, denominati bias, che possono essere classificati come: bias di selezione, di informazione e da confondimento. Mentre i primi due possono essere contrastati solamente mediante un opportuno disegno di studio, i bias da confondimento possono essere almeno parzialmente controllati in fase di disegno dello studio, mediante appaiamento (*matching*) o restrizione, oppure, in fase di analisi dei dati, adottando un opportuno modellamento statistico. Il presente lavoro riassume i principali metodi per il controllo del confondimento, illustrando tecniche di analisi stratificata (metodo di Mantel-Haenszel) e di analisi multivariata (modello di regressione logistica), e fornendo alcuni esempi su data set simulati.