2 Important Special Cases of the Logistic Model

Contents

D.G. Kleinbaum and M. Klein, *Logistic Regression*, Statistics for Biology and Health,
DOI 10.1007/978-1-4419-1742-3_2, © Springer Science+Business Media, LLC 2010

Introduction

In this chapter, several important special cases of the logistic model involving a single (0, 1) exposure variable are considered with their corresponding odds ratio expressions. In particular, focus is on defining the independent variables that go into the model and on computing the odds ratio for each special case. Models that account for the potential confounding effects and potential interaction effects of covariates are emphasized.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completion of this chapter, the learner should be able to:

1. State or recognize the logistic model for a simple analysis.
2. Given a model for simple analysis:
   a. state an expression for the odds ratio describing the exposure–disease relationship
   b. state or recognize the null hypothesis of no exposure–disease relationship in terms of parameter(s) of the model
   c. compute or recognize an expression for the risk for exposed or unexposed persons separately
   d. compute or recognize an expression for the odds of getting the disease for exposed or unexposed persons separately
3. Given two (0, 1) independent variables:
   a. state or recognize a logistic model that allows for the assessment of interaction on a multiplicative scale
   b. state or recognize the expression for no interaction on a multiplicative scale in terms of odds ratios for different combinations of the levels of two (0, 1) independent variables
   c. state or recognize the null hypothesis for no interaction on a multiplicative scale in terms of one or more parameters in an appropriate logistic model
4. Given a study situation involving a (0, 1) exposure variable and several control variables:
   a. state or recognize a logistic model that allows for the assessment of the exposure-disease relationship, controlling for the potential confounding and potential interaction effects of functions of the control variables
   b. compute or recognize the expression for the odds ratio for the effect of exposure on disease status adjusting for the potential confounding and interaction effects of the control variables in the model
   c. state or recognize an expression for the null hypothesis of no interaction effect involving one or more of the effect modifiers in the model
   d. assuming no interaction, state or recognize an expression for the odds ratio for the effect of exposure on disease status adjusted for confounders

   e. assuming no interaction, state or recognize the null hypothesis for testing the significance of this odds ratio in terms of a parameter in the model

5. Given a logistic model involving interaction terms, state or recognize that the expression for the odds ratio will give different values for the odds ratio depending on the values specified for the effect modifiers in the model.

Presentation

I. Overview

Special Cases:

- Simple analysis $\left(\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline\end{array}\right)$
- Multiplicative interaction
- Controlling several confounders and effect modifiers

This presentation describes important special cases of the general logistic model when there is a single (0, 1) exposure variable. Special case models include simple analysis of a fourfold table, assessment of multiplicative interaction between two dichotomous variables, and controlling for several confounders and interaction terms. In each case, we consider the definitions of variables in the model and the formula for the odds ratio describing the exposure-disease relationship.

General logistic model formula:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-\left(\alpha + \sum \beta_i \mathrm{X}_i\right)}}$$

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$

α, β_i = unknown parameters

D = dichotomous outcome

Recall that the general logistic model for k independent variables may be written as P($\mathbf{X}$) equals 1 over 1 plus e to minus the quantity α plus the sum of $\beta_i X_i$, where P($\mathbf{X}$) denotes the probability of developing a disease of interest given values of a collection of independent variables X_1, X_2, through X_k, that are collectively denoted by the *bold* $\mathbf{X}$. The terms α and β_i in the model represent unknown parameters that we need to estimate from data obtained for a group of subjects on the Xs and on D, a dichotomous disease outcome variable.

$$\operatorname{logit} \mathrm{P}(\mathbf{X}) = \underbrace{\alpha + \sum \beta_i \mathrm{X}_i}_{\text{linear sum}}$$

An alternative way of writing the logistic model is called the logit form of the model. The expression for the logit form is given here.

$$\begin{aligned}\mathrm{ROR} &= \mathrm{e}^{\sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})} \\ &= \prod_{i=1}^{k} \mathrm{e}^{\beta_i (X_{1i} - X_{0i})}\end{aligned}$$

$\mathbf{X}_1$ specification of $\mathbf{X}$ for subject 1

$\mathbf{X}_0$ specification of $\mathbf{X}$ for subject 0

The general odds ratio formula for the logistic model is given by either of two formulae. The first formula is of the form e to a sum of linear terms. The second is of the form of the product of several exponentials; that is, each term in the product is of the form e to some power. Either formula requires two specifications, $\mathbf{X}_1$ and $\mathbf{X}_0$, of the collection of k independent variables X_1, X_2, ..., X_k.

We now consider a number of important special cases of the logistic model and their corresponding odds ratio formulae.

II. Special Case – Simple Analysis

$X_1 = E =$ exposure (0, 1)

$D =$ disease (0, 1)

We begin with the simple situation involving one dichotomous independent variable, which we will refer to as an *exposure* variable and will denote it as $X_1 = E$. Because the disease variable, D, considered by a logistic model is dichotomous, we can use a two-way table with four cells to characterize this analysis situation, which is often referred to as a *simple analysis*.

	$E = 1$	$E = 0$
$D = 1$	*a*	*b*
$D = 0$	*c*	*d*

For convenience, we define the exposure variable as a (0, 1) variable and place its values in the two columns of the table. We also define the disease variable as a (0, 1) variable and place its values in the rows of the table. The cell frequencies within the fourfold table are denoted as *a, b, c*, and *d*, as is typically presented for such a table.

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-(\alpha+\beta_1 E)}},$$

where $E = (0, 1)$ variable.

Note: Other coding schemes (1, −1), (1, 2), (2, 1)

A logistic model for this simple analysis situation can be defined by the expression $\mathrm{P}(\mathbf{X})$ equals 1 over 1 plus e to minus the quantity α plus β_1 times E, where E takes on the value 1 for exposed persons and 0 for unexposed persons. Note that other coding schemes for E are also possible, such as (1, −1), (1, 2), or even (2, 1). However, we defer discussing such alternatives until Chap. 3.

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \beta_1 E$$

The logit form of the logistic model we have just defined is of the form logit $\mathrm{P}(\mathbf{X})$ equals the simple linear sum α plus β_1 times E. As stated earlier in our review, this logit form is an alternative way to write the statement of the model we are using.

$\mathrm{P}(\mathbf{X}) = \Pr(D = 1|E)$

$E = 1$: $\mathrm{R}_1 = \Pr(D = 1|E = 1)$

$E = 0$: $\mathrm{R}_0 = \Pr(D = 1|E = 0)$

The term $\mathrm{P}(\mathbf{X})$ for the simple analysis model denotes the probability that the disease variable D takes on the value 1, given whatever the value is for the exposure variable E. In epidemiologic terms, this probability denotes the *risk* for developing the disease, given exposure status. When the value of the exposure variable equals 1, we call this risk $\mathbf{R}_1$, which is the conditional probability that D equals 1 given that E equals 1. When E equals 0, we denote the risk by $\mathbf{R}_0$, which is the conditional probability that D equals 1 given that E equals 0.

$$\mathrm{ROR}_{E=1\text{ vs. }E=0}=\frac{\dfrac{\mathbf{R}_1}{1-\mathbf{R}_1}}{\dfrac{\mathbf{R}_0}{1-\mathbf{R}_0}}$$

We would like to use the above model for simple analysis to obtain an expression for the odds ratio that compares exposed persons with unexposed persons. Using the terms $\mathbf{R}_1$ and $\mathbf{R}_0$, we can write this odds ratio as $\mathbf{R}_1$ divided by 1 minus $\mathbf{R}_1$ over $\mathbf{R}_0$ divided by 1 minus $\mathbf{R}_0$.

Substitute $\mathrm{P}(\mathbf{X})=\dfrac{1}{1+e^{-\left(\alpha+\sum\beta_i X_i\right)}}$ into ROR formula:

To compute the odds ratio in terms of the parameters of the logistic model, we substitute the logistic model expression into the odds ratio formula.

$$E=1:\ \mathbf{R}_1=\frac{1}{1+e^{-(\alpha+[\beta_1\times 1])}}$$
$$=\frac{1}{1+e^{-(\alpha+\beta_1)}}$$

For E equal to 1, we can write $\mathbf{R}_1$ by substituting the value E equals 1 into the model formula for P($\mathbf{X}$). We then obtain 1 over 1 plus e to minus the quantity α plus β_1 times 1, or simply 1 over 1 plus e to minus α plus β_1.

$$E=0:\ \mathbf{R}_0=\frac{1}{1+e^{-(\alpha+[\beta_1\times 0])}}$$
$$=\frac{1}{1+e^{-\alpha}}$$

For E equal to zero, we write $\mathbf{R}_0$ by substituting E equal to 0 into the model formula, and we obtain 1 over 1 plus e to minus α.

$$\mathrm{ROR}=\frac{\dfrac{\mathbf{R}_1}{1-\mathbf{R}_1}}{\dfrac{\mathbf{R}_0}{1-\mathbf{R}_0}}=\frac{\dfrac{1}{1+e^{-(\alpha+\beta_1)}}}{\dfrac{1}{1+e^{-\alpha}}}$$

algebra

$$=\boxed{e^{\beta_1}}$$

To obtain ROR then, we replace $\mathbf{R}_1$ with 1 over 1 plus e to minus α plus β_1, and we replace $\mathbf{R}_0$ with 1 over 1 plus e to minus α. The ROR formula then simplifies algebraically to e to the β_1, where β_1 is the coefficient of the exposure variable.

General ROR formula used for other special cases

We could have obtained this expression for the odds ratio using the general formula for the ROR that we gave during our review. We will use the general formula now. Also, for other special cases of the logistic model, we will use the general formula rather than derive an odds ratio expression separately for each case.

General:

$$\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} = \mathrm{e}^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

Simple analysis:

$$k = 1, \mathbf{X} = (X_1), \beta_i = \beta_1$$

The general formula computes ROR as e to the sum of each β_i times the difference between X_{1i} and X_{0i}, where X_{1i} denotes the value of the ith X variable for group 1 persons and X_{0i} denotes the value of the ith X variable for group 0 persons. In a simple analysis, we have only one X and one β; in other words, k, the number of variables in the model, equals 1.

group 1: $\mathbf{X}_1 = E = 1$
group 0: $\mathbf{X}_0 = E = 0$

$$\mathbf{X}_1 = (X_{11}) = (1)$$
$$\mathbf{X}_0 = (X_{01}) = (0)$$

For a simple analysis model, group 1 corresponds to exposed persons, for whom the variable X_1, in this case E, equals 1. Group 0 corresponds to unexposed persons, for whom the variable X_1 or E equals 0. Stated another way, for group 1, the collection of Xs denoted by the *bold* $\mathbf{X}$ can be written as $\mathbf{X}_1$ and equals the collection of one value X_{11}, which equals 1. For group 0, the collection of Xs denoted by the *bold* $\mathbf{X}$ is written as $\mathbf{X}_0$ and equals the collection of one value X_{01}, which equals 0.

$$\begin{aligned}\mathrm{ROR}_{\mathbf{X}_1,\,\mathbf{X}_0} &= \mathrm{e}^{\beta_1 (X_{11} - X_{01})} \\ &= \mathrm{e}^{\beta_1 (1-0)} \\ &= \mathrm{e}^{\beta_1}\end{aligned}$$

Substituting the particular values of the one X variable into the general odds ratio formula then gives e to the β_1 times the quantity X_{11} minus X_{01}, which becomes e to the β_1 times 1 minus 0, which reduces to e to the β_1.

SIMPLE ANALYSIS SUMMARY

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-(\alpha + \beta_1 E)}}$$
$$\mathrm{ROR} = \mathrm{e}^{\beta_1}$$

In summary, for the simple analysis model involving a (0, 1) exposure variable, the logistic model P(**X**) equals 1 over 1 plus e to minus the quantity α plus β_1 times E, and the odds ratio that describes the effect of the exposure variable is given by e to the β_1, where β_1 is the coefficient of the exposure variable.

$$\widehat{\mathrm{ROR}}_{\mathbf{X}_1,\,\mathbf{X}_0} = \mathrm{e}^{\hat{\beta}_1}$$

We can estimate this odds ratio by fitting the simple analysis model to a set of data. The estimate of the parameter β_1 is typically denoted as $\hat{\beta}_1$. The odds ratio estimate then becomes e to the $\hat{\beta}_1$.

	$E = 1$	$E = 0$
$D = 1$	a	b
$D = 0$	c	d

$\widehat{\text{ROR}} = e^{\hat{\beta}} = ad/bc$

The reader should not be surprised to find out that an alternative formula for the estimated odds ratio for the simple analysis model is the familiar a times d over b times c, where a, b, c, and d are the cell frequencies in the fourfold table for simple analysis. That is, e to the $\hat{\beta}_1$ obtained from fitting a logistic model for simple analysis can alternatively be computed as ad divided by bc from the cell frequencies of the fourfold table.

Simple analysis: does not need computer

Other special cases: require computer

Thus, in the simple analysis case, we need not go to the trouble of fitting a logistic model to get an odds ratio estimate as the typical formula can be computed without a computer program. We have presented the logistic model version of simple analysis to show that the logistic model incorporates simple analysis as a special case. More complicated special cases, involving more than one independent variable, require a computer program to compute the odds ratio.

III. Assessing Multiplicative Interaction

We will now consider how the logistic model allows the assessment of interaction between two independent variables.

$X_1 = A = (0, 1)$ variable

$X_2 = B = (0, 1)$ variable

Interaction: equation involving RORs for combinations of A and B

Consider, for example, two (0, 1) X variables, X_1 and X_2, which for convenience we rename as A and B, respectively. We first describe what we mean conceptually by interaction between these two variables. This involves an equation involving risk odds ratios corresponding to different combinations of A and B. The odds ratios are defined in terms of risks, which we now describe.

$$\begin{aligned} R_{AB} &= \text{risk given } A, B \\ &= \Pr(D = 1 \mid A, B) \end{aligned}$$

Let R_{AB} denote the risk for developing the disease, given specified values for A and B; in other words, R_{AB} equals the conditional probability that D equals 1, given A and B.

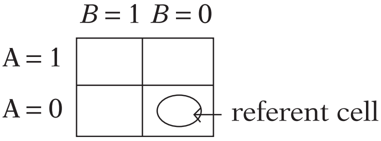
	$B = 1$	$B = 0$
A = 1	R_{11}	R_{10}
A = 0	R_{01}	R_{00}

Because A and B are dichotomous, there are four possible values for R_{AB}, which are shown in the cells of a two-way table. When A equals 1 and B equals 1, the risk R_{AB} becomes R_{11}. Similarly, when A equals 1 and B equals 0, the risk becomes R_{10}. When A equals 0 and B equals 1, the risk is R_{01}, and finally, when A equals 0 and B equals 0, the risk is R_{00}.

Note: above table not for simple analysis.

	$B = 1$	$B = 0$
$A = 1$	R_{11}	R_{10}
$A = 0$	R_{01}	R_{00}

Note that the two-way table presented here does not describe a simple analysis because the row and column headings of the table denote two independent variables rather than one independent variable and one disease variable. Moreover, the information provided within the table is a collection of four risks corresponding to different combinations of both independent variables, rather than four cell frequencies corresponding to different exposure-disease combinations.



$$\text{OR}_{11} = \text{odds}(1, 1)/\text{odds}(0, 0)$$
$$\text{OR}_{10} = \text{odds}(1, 0)/\text{odds}(0, 0)$$
$$\text{OR}_{01} = \text{odds}(0, 1)/\text{odds}(0, 0)$$

Within this framework, odds ratios can be defined to compare the odds for any one cell in the two-way table of risks with the odds for any other cell. In particular, three odds ratios of typical interest compare each of three of the cells to a *referent cell*. The referent cell is usually selected to be the combination A equals 0 and B equals 0. The three odds ratios are then defined as OR_{11}, OR_{10}, and OR_{01}, where OR_{11} equals the odds for cell 11 divided by the odds for cell 00, OR_{10} equals the odds for cell 10 divided by the odds for cell 00, and OR_{01} equals the odds for cell 01 divided by the odds for cell 00.

$$\text{odds }(A,B) = R_{AB}/(1 - R_{AB})$$

$$\text{OR}_{11} = \frac{R_{11}/(1-R_{11})}{R_{00}/(1-R_{00})} = \frac{R_{11}(1-R_{00})}{R_{00}(1-R_{11})}$$

$$\text{OR}_{10} = \frac{R_{10}/(1-R_{10})}{R_{00}/(1-R_{00})} = \frac{R_{10}(1-R_{00})}{R_{00}(1-R_{10})}$$

$$\text{OR}_{01} = \frac{R_{01}/(1-R_{01})}{R_{00}/(1-R_{00})} = \frac{R_{01}(1-R_{00})}{R_{00}(1-R_{01})}$$

As the odds for any cell A,B is defined in terms of risks as R_{AB} divided by 1 minus R_{AB}, we can obtain the following expressions for the three odds ratios: OR_{11} equals the product of R_{11} times 1 minus R_{00} divided by the product of R_{00} times 1 minus R_{11}. The corresponding expressions for OR_{10} and OR_{01} are similar, where the subscript 11 in the numerator and denominator of the 11 formula is replaced by 10 and 01, respectively.

$$\text{OR}_{\text{AB}} = \frac{R_{\text{AB}}(1-R_{00})}{R_{00}(1-R_{\text{AB}})}$$

$$A = 0, 1; \quad B = 0, 1$$

In general, without specifying the value of A and B, we can write the odds ratio formulae as OR_{AB} equals the product of R_{AB} and 1 minus R_{00} divided by the product of R_{00} and $1 - R_{AB}$, where A takes on the values 0 and 1 and B takes on the values 0 and 1.

DEFINITION

$$\underset{\text{no interaction on a multiplicative scale}}{\mathrm{OR}_{11} = \mathrm{OR}_{10} \underset{\nwarrow \text{ multiplication}}{\times} \mathrm{OR}_{01}}$$

Now that we have defined appropriate odds ratios for the two independent variables situation, we are ready to provide an equation for assessing interaction. The equation is stated as OR_{11} equals the product of OR_{10} and OR_{01}. If this expression is satisfied for a given study situation, we say that there is "no interaction on a *multiplicative* scale." In contrast, if this expression is not satisfied, we say that there is evidence of interaction on a multiplicative scale.

Note that the right-hand side of the "no interaction" expression requires *multiplication* of two odds ratios, one corresponding to the combination 10 and the other to the combination 01. Thus, the scale used for assessment of interaction is called multiplicative.

No interaction:

$$\underset{\substack{\uparrow \\ \mathrm{OR}_{11}}}{\begin{pmatrix}\text{effect of} \\ A \text{ and } B \\ \text{acting} \\ \text{together}\end{pmatrix}} = \underset{\substack{\uparrow \\ \mathrm{OR}_{10} \times \mathrm{OR}_{01} \\ \text{multiplicative} \\ \text{scale}}}{\begin{pmatrix}\text{combined} \\ \text{effect of} \\ \text{A and B} \\ \text{acting} \\ \text{separately}\end{pmatrix}}$$

When the no interaction equation is satisfied, we can interpret the effect of both variables A and B acting together as being the same as the combined effect of each variable acting separately.

The effect of both variables acting together is given by the odds ratio OR_{11} obtained when A and B are both present, that is, when A equals 1 and B equals 1.

The effect of A acting separately is given by the odds ratio for A equals 1 and B equals 0, and the effect of B acting separately is given by the odds ratio for A equals 0 and B equals 1. The combined separate effects of A and B are then given by the product OR_{10} times OR_{01}.

no interaction formula:
$\mathrm{OR}_{11} = \mathrm{OR}_{10} \times \mathrm{OR}_{01}$

Thus, when there is no interaction on a multiplicative scale, OR_{11} equals the product of OR_{10} and OR_{01}.

EXAMPLE

	$B = 1$	$B = 0$
$A = 1$	$R_{11} = 0.0350$	$R_{10} = 0.0175$
$A = 0$	$R_{01} = 0.0050$	$R_{00} = 0.0025$

$$\text{OR}_{11} = \frac{0.0350(1 - 0.0025)}{0.0025(1 - 0.0350)} = 14.4$$

$$\text{OR}_{10} = \frac{0.0175(1 - 0.0025)}{0.0025(1 - 0.0175)} = 7.2$$

$$\text{OR}_{01} = \frac{0.0050(1 - 0.0025)}{0.0025(1 - 0.0050)} = 2.0$$

As an example of no interaction on a multiplicative scale, suppose the risks R_{AB} in the fourfold table are given by R_{11} equal to 0.0350, R_{10} equal to 0.0175, R_{01} equal to 0.0050, and R_{00} equal to 0.0025. Then the corresponding three odds ratios are obtained as follows: OR_{11} equals 0.0350 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0350, which becomes 14.4; OR_{10} equals 0.0175 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0175, which becomes 7.2; and OR_{01} equals 0.0050 times 1 minus 0.0025 divided by the product of 0.0025 and 1 minus 0.0050, which becomes 2.0.

$$\text{OR}_{11} \overset{?}{=} \text{OR}_{10} \times \text{OR}_{01}$$

$$14.4 \underset{\substack{\uparrow \\ \text{Yes}}}{\overset{?}{=}} \underbrace{7.2 \times 2.0}_{14.4}$$

To see if the no interaction equation is satisfied, we check whether OR_{11} equals the product of OR_{10} and OR_{01}. Here we find that OR_{11} equals 14.4 and the product of OR_{10} and OR_{01} is 7.2 times 2, which is also 14.4. Thus, the no interaction equation is satisfied.

$B = 1$	$B = 0$
$R_{11} = 0.0700$	$R_{10} = 0.0175$
$R_{01} = 0.0050$	$R_{00} = 0.0025$

$\text{OR}_{11} = 30.0$

$\text{OR}_{10} = 7.2$

$\text{OR}_{01} = 2.0$

$$\text{OR}_{11} \overset{?}{=} \text{OR}_{10} \times \text{OR}_{01}$$

$$30.0 \underset{\substack{\uparrow \\ \text{No}}}{\overset{?}{=}} 7.2 \times 2.0$$

In contrast, using a different example, if the risk for the 11 cell is 0.0700, whereas the other three risks remained at 0.0175, 0.0050, and 0.0025, then the corresponding three odds ratios become OR_{11} equals 30.0, OR_{10} equals 7.2, and OR_{01} equals 2.0. In this case, the no interaction equation is not satisfied because the left-hand side equals 30 and the product of the two odds ratios on the right-hand side equals 14. Here, then, we would conclude that there is interaction because the effect of both variables acting together is more than twice the combined effect of the variables acting separately.

EXAMPLE (continued)

Note: "=" means approximately equal ($\approx$)
e.g., $14.5 \approx 14.0 \Rightarrow$ no interaction

Note that in determining whether or not the no interaction equation is satisfied, the left- and right-hand sides of the equation do not have to be exactly equal. If the left-hand side is approximately equal to the right-hand side, we can conclude that there is no interaction. For instance, if the left-hand side is 14.5 and the right-hand side is 14, this would typically be close enough to conclude that there is no interaction on a multiplicative scale.

REFERENCE
multiplicative interaction vs. additive interaction
Epidemiologic Research, Chap. 19

A more complete discussion of interaction, including the distinction between *multiplicative interaction* and *additive interaction*, is given in Chap. 19 of *Epidemiologic Research* by Kleinbaum, Kupper, and Morgenstern (1982).

Logistic model variables:

$$\left.\begin{aligned} X_1 &= A_{(0,1)} \\ X_2 &= B_{(0,1)} \end{aligned}\right\}\text{main effects}$$

$X_3 = A \times B$ interaction effect variable

We now define a logistic model that allows the assessment of multiplicative interaction involving two (0, 1) indicator variables A and B. This model contains three independent variables, namely, X_1 equal to A, X_2 equal to B, and X_3 equal to the product term A times B. The variables A and B are called main effect variables and the product term is called an interaction effect variable.

$$\begin{aligned} \text{logit P}(\mathbf{X}) = \alpha &+ \beta_1 A + \beta_2 B \\ &+ \beta_3 A \times B, \end{aligned}$$

where

$$\begin{aligned} \text{P}(\text{X}) &= \text{risk given } A \text{ and } B \\ &= R_{AB} \end{aligned}$$

The logit form of the model is given by the expression logit of P($\mathbf{X}$) equals α plus β_1 times A plus β_2 times B plus β_3 times A times B. P($\mathbf{X}$) denotes the risk for developing the disease given values of A and B, so that we can alternatively write P($\mathbf{X}$) as R_{AB}.

$$\beta_3 = \ln_e\left[\frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}}\right]$$

For this model, it can be shown mathematically that the coefficient β_3 of the product term can be written in terms of the three odds ratios we have previously defined. The formula is β_3 equals the natural log of the quantity OR_{11} divided by the product of OR_{10} and OR_{01}. We can make use of this formula to test the null hypothesis of no interaction on a multiplicative scale.

H_0 no interaction on a multiplicative scale

$$\Leftrightarrow H_0 : \text{OR}_{11} = \text{OR}_{10} \times \text{OR}_{01}$$

$$\Leftrightarrow H_0 : \frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}} = 1$$

$$\Leftrightarrow H_0 : \ln_e\left(\frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}}\right) = \ln_e 1$$

$$\Leftrightarrow H_0 : \beta_3 = 0$$

One way to state this null hypothesis, as described earlier in terms of odds ratios, is OR_{11} equals the product of OR_{10} and OR_{01}. Now it follows algebraically that this odds ratio expression is equivalent to saying that the quantity OR_{11} divided by OR_{10} times OR_{01} equals 1, or equivalently, that the natural log of this expression equals the natural log of 1, or, equivalently, that β_3 equals 0. Thus, the null hypothesis of no interaction on a multiplicative scale can be equivalently stated as β_3 equals 0.

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 A + \beta_2 B + \beta_3 AB$$

H_0: no interaction $\Leftrightarrow \beta_3 = 0$

Test result	*Model*
not significant	$\Rightarrow \alpha + \beta_1 A + \beta_2 B$
significant	$\Rightarrow \alpha + \beta_1 A + \beta_2 B + \beta_3 AB$

In other words, a test for the no interaction hypotheses can be obtained by testing for the significance of the coefficient of the product term in the model. If the test is not significant, we would conclude that there is no interaction on a multiplicative scale and we would reduce the model to a simpler one involving only main effects. In other words, the reduced model would be of the form logit $P(\mathbf{X})$ equals α plus β_1 times A plus β_2 times B. If, on the other hand, the test is significant, the model would retain the β_3 term and we would conclude that there is significant interaction on a multiplicative scale.

MAIN POINT:

Interaction test ⇒ test for product terms

A description of methods for testing hypotheses for logistic regression models is beyond the scope of this presentation (see Chap. 5). The main point here is that we can test for interaction in a logistic model by testing for significance of product terms that reflect interaction effects in the model.

EXAMPLE

Case-control study

ASB = (0, 1)	variable for asbestos exposure
SMK = (0, 1)	variable for smoking status
D = (0, 1)	variable for bladder cancer status

As an example of a test for interaction, we consider a study that looks at the combined relationship of asbestos exposure and smoking to the development of bladder cancer. Suppose we have collected case-control data on several persons with the same occupation. We let *ASB* denote a (0,1) variable indicating asbestos exposure status, *SMK* denote a (0, 1) variable indicating smoking status, and D denote a (0, 1) variable for bladder cancer status.

EXAMPLE (continued)

$$\text{logit}\,(\mathbf{X}) = \alpha + \beta_1\text{ASB} + \beta_2\text{SMK} + \beta_3\text{ASB} \times \text{SMK}$$

To assess the extent to which there is a multiplicative interaction between asbestos exposure and smoking, we consider a logistic model with ASB and SMK as main effect variables and the product term ASB times SMK as an interaction effect variable. The model is given by the expression logit P(**X**) equals α plus β_1 times ASB plus β_2 times SMK plus β_3 times ASB times SMK. With this model, a test for no interaction on a multiplicative scale is equivalent to testing the null hypothesis that β_3, the coefficient of the product term, equals 0.

$\mathbf{H}_0$: no interaction (multiplicative)
$\Leftrightarrow \text{H}_0 : \beta_3 = 0$

Test Result	*Conclusion*
Not Significant	No interaction on multiplicative scale
Significant ($\hat{\beta}_3 > 0$)	Joint effect > combined effect
Significant ($\hat{\beta}_3 < 0$)	Joint effect < combined effect

If this test is not significant, then we would conclude that the effect of asbestos and smoking acting together is equal, on a multiplicative scale, to the combined effect of asbestos and smoking acting separately. If this test is significant and $\hat{\beta}_3$ is greater than 0, we would conclude that the joint effect of asbestos and smoking is greater than a multiplicative combination of separate effects. Or, if the test is significant and $\hat{\beta}_3$ is less than zero, we would conclude that the joint effect of asbestos and smoking is less than a multiplicative combination of separate effects.

IV. The *E, V, W* Model – A General Model Containing a (0, 1) Exposure and Potential Confounders and Effect Modifiers

We are now ready to discuss a logistic model that considers the effects of several independent variables and, in particular, allows for the control of confounding and the assessment of interaction. We call this model the *E, V, W* model. We consider a single dichotomous (0, 1) exposure variable, denoted by *E*, and *p* extraneous variables C_1, C_2, and so on, up through C_p. The variables C_1 through C_p may be either continuous or categorical.

The variables:
E = (0, 1) exposure
$C_1, C_2, \ldots, C_p$ continuous or categorical

EXAMPLE

$$D = \text{CHD}_{(0,1)}$$
$$E = \text{CAT}_{(0,1)}$$
$$\text{Control variables}\ \left\{\begin{array}{l} C_1 = \text{AGE}_{\text{continous}} \\ C_2 = \text{CHL}_{\text{continous}} \\ C_3 = \text{SMK}_{(0,1)} \\ C_4 = \text{ECG}_{(0,\,1)} \\ C_5 = \text{HPT}_{(0,\,1)} \end{array}\right.$$

As an example of this special case, suppose the disease variable is coronary heart disease status (CHD), the exposure variable *E* is catecholamine level (CAT), where 1 equals high and 0 equals low, and the control variables are AGE, cholesterol level (CHL), smoking status (SMK), electrocardiogram abnormality status (ECG), and hypertension status (HPT).

EXAMPLE (continued)

1 E : CAT
5 *C*s : AGE, CHL, SMK, ECG, HPT

We will assume here that both AGE and CHL are treated as continuous variables, that SMK is a (0, 1) variable, where 1 equals ever smoked and 0 equals never smoked, that ECG is a (0, 1) variable, where 1 equals abnormality present and 0 equals abnormality absent, and that HPT is a (0, 1) variable, where 1 equals high blood pressure and 0 equals normal blood pressure. There are, thus, five *C* variables in addition to the exposure variable CAT.

Model with eight independent variables:

2 E × *C*s : CAT × CHL
CAT × HPT

We now consider a model with eight independent variables. In addition to the exposure variable CAT, the model contains the five *C* variables as potential confounders plus two product terms involving two of the *C*s, namely, CHL and HPT, which are each multiplied by the exposure variable CAT.

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT}$$

$$\underbrace{+\gamma_1\text{AGE}+\gamma_2\text{CHL}+\gamma_3\text{SMK}+\gamma_4\text{ECG}+\gamma_5\text{HPT}}_{\text{main effects}}$$

$$\underbrace{+\ \delta_1\text{CAT}\times\text{CHL}\ +\ \delta_2\text{CAT}\times\text{HPT}}_{\text{interaction effects}}$$

The model is written as logit P(**X**) equals α plus β times CAT plus the sum of five main effect terms γ_1 times AGE plus γ_2 times CHL and so on up through γ_5 times HPT plus the sum of δ_1 times CAT times CHL plus δ_2 times CAT times HPT. Here the five main effect terms account for the potential confounding effect of the variables AGE through HPT and the two product terms account for the potential interaction effects of CHL and HPT.

Parameters:
α, β, γs, and δs instead of α and βs,

where
β: exposure variable
γs: potential confounders
δs: potential interaction variables

Note that the parameters in this model are denoted as α, β, γs, and δs, whereas previously we denoted all parameters other than the constant α as β_is. We use β, γs, and δs here to distinguish different types of variables in the model. The parameter β indicates the coefficient of the exposure variable, the γs indicate the coefficients of the potential confounders in the model, and the δs indicate the coefficients of the potential interaction variables in the model. This notation for the parameters will be used throughout the remainder of this presentation.

The general *E, V, W* Model

single exposure, controlling for C_1, C_2, . . . , C_p

Analogous to the above example, we now describe the general form of a logistic model, called the *E, V, W* model, that considers the effect of a single exposure controlling for the potential confounding and interaction effects of control variables C_1, C_2, up through C_p.

E, V, W Model

$k = p_1 + p_2 + 1 =$ no. of variables in model
$p_1 =$ no. of potential confounders
$p_2 =$ no. of potential interactions
$1 =$ exposure variable

The general *E, V, W* model contains p_1 plus p_2 plus 1 variables, where p_1 is the number of potential confounders in the model, p_2 is the number of potential interaction terms in the model, and 1 denotes the exposure variable.

CHD EXAMPLE

$p_1 = 5$: AGE, CHL, SMK, ECG, HPT

$p_2 = 2$: CAT × CHL, CAT × HPT

$p_1 + p_2 + 1 = 5 + 2 + 1 = 8$

In the CHD study example above, there are p_1 equals to five potential confounders, namely, the five control variables, and there are p_2 equal to two interaction variables, the first of which is CAT × CHL and the second is CAT × HPT. The total number of variables in the example is, therefore, p_1 plus p_2 plus 1 equals 5 plus 2 plus 1, which equals 8. This corresponds to the model presented earlier, which contained eight variables.

- $V_1, \ldots, V_{p_1}$ are potential confounders
- Vs are functions of Cs

In addition to the exposure variable E, the general model contains p_1 variables denoted as V_1, V_2 through V_{p_1}. The set of Vs are functions of the Cs that are thought to account for confounding in the data. We call the set of these Vs *potential confounders*.

e.g., $V_1 = C_1, V_2 = (C_2)^2, V_3 = C_1 \times C_3$

For instance, we may have V_1 equal to C_1, V_2 equal to $(C_2)^2$, and V_3 equal to $C_1 \times C_3$.

CHD EXAMPLE

$V_1 =$ AGE, $V_2 =$ CHL, $V_3 =$ SMK,
$V_4 =$ ECG, $V_5 =$ HPT

The CHD example above has five Vs that are the same as the Cs.

Following the Vs, we define p_2 variables that are product terms of the form E times W_1, E times W_2, and so on up through E times W_{p_2}, where W_1, W_2, through W_{p_2}, denote a set of functions of the Cs that are *potential effect modifiers* with E.

- $W_1, \ldots, W_{p_2}$ are potential effect modifiers
- Ws are functions of Cs

e.g., $W_1 = C_1$, $W_2 = C_1 \times C_3$

For instance, we may have W_1 equal to C_1 and W_2 equal to C_1 times C_3.

CHD EXAMPLE

$W_1 =$ CHL, $W_2 =$ HPT

The CHD example above has two Ws, namely, CHL and HPT, that go into the model as product terms of the form CAT × CHL and CAT × HPT.

REFERENCES FOR CHOICE OF Vs AND Ws FROM Cs

- Chap. 6: Modeling Strategy Guidelines
- *Epidemiologic Research*, Chap. 21

It is beyond the scope of this chapter to discuss the subtleties involved in the particular choice of the *V*s and *W*s from the *C*s for a given model. More depth is provided in a separate chapter (Chap. 6) on modeling strategies and in Chap. 21 of *Epidemiologic Research* by Kleinbaum, Kupper, and Morgenstern.

Assume: *V*s and *W*s are *C*s or subset of *C*s

EXAMPLE

$C_1 = \text{AGE}, C_2 = \text{RACE}, C_3 = \text{SEX}$

$V_1 = \text{AGE}, V_2 = \text{RACE}, V_3 = \text{SEX}$

$W_1 = \text{AGE}, W_2 = \text{SEX}$

$p_1 = 3, p_2 = 2, k = p_1 + p_2 + 1 = 6$

In most applications, the *V*s will be the *C*s themselves or some subset of the *C*s and the *W*s will also be the *C*s themselves or some subset thereof. For example, if the *C*s are AGE, RACE, and SEX, then the *V*s may be AGE, RACE, and SEX, and the *W*s may be AGE and SEX, the latter two variables being a subset of the *C*s. Here the number of *V* variables, p_1, equals 3, and the number of *W* variables, p_2, equals 2, so that *k*, which gives the total number of variables in the model, is p_1 plus p_2 plus 1 equals 6.

NOTE
Ws ARE SUBSET OF Vs

EXAMPLE

~~$V_1 = \text{AGE}, V_2 = \text{RACE}$~~
~~$W_1 = \text{AGE}, W_2 = \text{SEX}$~~

Note, as we describe further in Chap. 6, that you cannot have a *W* in the model that is not also contained in the model as a *V*; that is, *W*s have to be a subset of the *V*s. For instance, we cannot allow a model whose *V*s are AGE and RACE and whose *W*s are AGE and SEX because the SEX variable is not contained in the model as a *V* term.

$$\begin{aligned}\text{logit}\,P(\mathbf{X}) = \alpha + \beta E + \gamma_1 V_1 + \gamma_2 V_2 \\ + \cdots + \gamma_{p_1} V_{p_1} + \delta_1 E W_1 \\ + \delta_2 E W_2 + \cdots + \delta_{p_2} E W_{p_2},\end{aligned}$$

A logistic model incorporating this special case containing the *E, V*, and *W* variables defined above can be written in logit form as shown here.

where
β = coefficient of *E*
γs = coefficient of *V*s
δs = coefficient of *W*s

Note that β is the coefficient of the single exposure variable *E*, the γs are coefficients of potential confounding variables denoted by the *V*s, and the δs are coefficients of potential interaction effects involving *E* separately with each of the *W*s.

$$\text{logit}\,P(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

We can factor out the *E* from each of the interaction terms, so that the model may be more simply written as shown here. This is the form of the model that we will use henceforth in this presentation.

Adjusted odds ratio for $E = 1$ vs. $E = 0$ given $C_1, C_2, \ldots, C_p$ fixed

We now provide for this model an expression for an adjusted odds ratio that describes the effect of the exposure variable on disease status adjusted for the potential confounding and interaction effects of the control variables C_1 through C_p. That is, we give a formula for the risk odds ratio comparing the odds of disease development for exposed vs. unexposed persons, with both groups having the same values for the extraneous factors C_1 through C_p. This formula is derived as a special case of the odds ratio formula for a general logistic model given earlier in our review.

$$\text{ROR} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right)$$

For our special case, the odds ratio formula takes the form ROR equals e to the quantity β plus the sum from 1 through p_2 of the δ_j times W_j.

- γ_i terms not in formula

Note that β is the coefficient of the exposure variable E, that the δ_j are the coefficients of the interaction terms of the form E times W_j, and that the coefficients γ_i of the main effect variables V_i do not appear in the odds ratio formula.

- Formula assumes E is (0, 1)
- Formula is modified if E has other coding, e.g., (1, −1), (2, 1), ordinal, or interval (see Chap. 3 on coding)

Note also that this formula assumes that the dichotomous variable E is coded as a (0, 1) variable with E equal to 1 for exposed persons and E equal to 0 for unexposed persons. If the coding scheme is different, for example, (1, −1) or (2, 1), or if E is an ordinal or interval variable, then the odds ratio formula needs to be modified. The effect of different coding schemes on the odds ratio formula will be described in Chap. 3.

Interaction:

$$\text{ROR} = \exp\left(\beta + \Sigma \boxed{\delta_j W_j}\right)$$

- $\delta_j \neq 0 \Rightarrow$ OR depends on W_j
- Interaction $\Rightarrow$ effect of E differs at different levels of Ws

This odds ratio formula tells us that if our model contains interaction terms, then the odds ratio will involve coefficients of these interaction terms and that, moreover, the value of the odds ratio will be different depending on the values of the W variables involved in the interaction terms as products with E. This property of the OR formula should make sense in that the concept of interaction implies that the effect of one variable, in this case E, is different at different levels of another variable, such as any of the Ws.

- Vs not in OR formula but Vs in model, so OR formula controls confounding:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \Sigma \enclose{circle}{\gamma_i} V_i + E\, \Sigma \enclose{circle}{\delta_j} W_j$$

Although the coefficients of the V terms do not appear in the odds ratio formula, these terms are still part of the fitted model. Thus, the odds ratio formula not only reflects the interaction effects in the model but also controls for the confounding variables in the model.

No interaction:

$$\text{all } \delta_j = 0 \Rightarrow \text{ROR} = \underset{\substack{\uparrow \\ \text{constant}}}{\exp(\beta)}$$

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \underset{\substack{\uparrow \\ \text{confounding} \\ \text{effects adjusted}}}{\sum \gamma_i V_i}$$

In contrast, if the model contains no interaction terms, then, equivalently, all the δ_j coefficients are 0; the odds ratio formula thus reduces to ROR equals to e to β, where β is the coefficient of the exposure variable E. Here, the *odds ratio is a fixed constant*, so that its value does not change with different values of the independent variables. The model in this case reduces to logit P($\mathbf{X}$) equals α plus β times E plus the sum of the main effect terms involving the Vs and contains no product terms. For this model, we can say that e to β represents an odds ratio that *adjusts for the potential confounding effects* of the control variables C_1 through C_p defined in terms of the Vs.

EXAMPLE

The model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta\text{CAT} \underbrace{+\gamma_1\text{AGE} + \gamma_2\text{CHL} + \gamma_3\text{SMK} + \gamma_4\text{ECG} + \gamma_5\text{HPT}}_{\text{main effects}} \underbrace{+\ \text{CAT}(\delta_1\text{CHL} + \delta_2\text{HPT})}_{\text{interaction effects}}$$

$$\text{logit P}(X) = \alpha + \beta\text{CAT} \underbrace{+\gamma_1\text{AGE} + \gamma_2\text{CHL} + \gamma_3\text{SMK} + \gamma_4\text{ECG} + \gamma_5\text{HPT}}_{\text{main effects: confounding}} \underbrace{+\ \text{CAT}(\delta_1\text{CHL} + \delta_2\text{HPT})}_{\text{product terms: interaction}}$$

$$\text{ROR} = \exp(\beta + \delta_1\text{CHL} + \delta_2\text{HPT})$$

As an example of the use of the odds ratio formula for the E, V, W model, we return to the CHD study example we described earlier. The CHD study model contained eight independent variables. The model is restated here as logit P($\mathbf{X}$) equals α plus β times CAT plus the sum of five main effect terms plus the sum of two interaction terms.

The five main effect terms in this model account for the potential confounding effects of the variables AGE through HPT. The two product terms account for the potential interaction effects of CHL and HPT with CAT.

For this example, the odds ratio formula reduces to the expression ROR equals e to the quantity β plus the sum δ_1 times CHL plus δ_2 times HPT.

EXAMPLE (continued)

$\text{ROR} = \exp(\hat{\beta} + \hat{\delta}_1\text{CHL} + \hat{\delta}_2\text{HPT})$

- varies with values of CHL and HPT

AGE, SMK, and ECG are adjusted for confounding

In using this formula, note that to obtain a numerical value for this odds ratio, not only do we need estimates of the coefficients β and the two δs, but we also need to specify values for the variables CHL and HPT. In other words, once we have fitted the model to obtain estimates of the coefficients, we will get different values for the odds ratio depending on the values that we specify for the interaction variables in our model. Note, also, that although the variables AGE, SMK, and ECG are not contained in the odds ratio expression for this model, the confounding effects of these three variables plus CHL and HPT are being adjusted because the model being fit contains all five control variables as main effect *V* terms.

n = 609 white males from Evans County, GA 9-year follow up

Fitted model:

Variable	*Coefficient*
Intercept	$\hat{\alpha} = -4.0497$
CAT	$\hat{\beta} = -12.6894$
AGE	$\hat{\gamma}_1 = 0.0350$
CHL	$\hat{\gamma}_2 = -0.0055$
SMK	$\hat{\gamma}_3 = 0.7732$
ECG	$\hat{\gamma}_4 = 0.3671$
HPT	$\hat{\gamma}_5 = 1.0466$
CAT × CHL	$\hat{\delta}_1 = 0.0692$
CAT × HPT	$\hat{\delta}_2 = -2.3318$

To provide numerical values for the above odds ratio, we will consider a data set of 609 white males from Evans County, Georgia, who were followed for 9 years to determine CHD status. The above model involving CAT, the five *V* variables, and the two *W* variables was fit to this data, and the fitted model is given by the list of coefficients corresponding to the variables listed here.

$$\widehat{\text{ROR}} = \exp(\underbrace{-12.6894}_{\text{exposure coefficient}} + \underbrace{0.0692\text{CHL} - 2.3318}_{\text{interaction coefficient}}\text{ HPT})$$

Based on the above fitted model, the estimated odds ratio for the CAT, CHD association adjusted for the five control variables is given by the expression shown here. Note that this expression involves only the coefficients of the exposure variable CAT and the interaction variables CAT times CHL and CAT times HPT, the latter two coefficients being denoted by δs in the model.

EXAMPLE (continued)

$\widehat{\text{ROR}}$ varies with values of CHL and HPT

effect modifiers

This expression for the odds ratio tells us that we obtain a different value for the estimated odds ratio depending on the values specified for CHL and HPT. As previously mentioned, this should make sense conceptually because CHL and HPT are the only two effect modifiers in the model, and the value of the odds ratio changes as the values of the effect modifiers change.

- CHL = 220, HPT = 1

$$\begin{aligned}\widehat{\text{ROR}} &= \exp[-12.6894 + 0.0692(220) \\ &\quad - 2.3318(1)] \\ &= \exp(0.2028) = \boxed{1.22}\end{aligned}$$

To get a numerical value for the odds ratio, we consider, for example, the specific values CHL equal to 220 and HPT equal to 1. Plugging these into the odds ratio formula, we obtain e to the 0.2028, which equals 1.22.

- CHL = 200, HPT = 0

$$\begin{aligned}\widehat{\text{ROR}} &= \exp[-12.6894 + 0.0692(200) \\ &\quad - 2.3318(0)] \\ &= \exp(1.1506) = \boxed{3.16}\end{aligned}$$

As a second example, we consider CHL equal to 200 and HPT equal to 0. Here, the odds ratio becomes e to 1.1506, which equals 3.16.

CHL = 220, HPT = 1 $\Rightarrow \widehat{\text{ROR}} = 1.22$
CHL = 200, HPT = 0 $\Rightarrow \widehat{\text{ROR}} = 3.16$

controls for the confounding effects of AGE, CHL, SMK, ECG, and HPT

Thus, we see that depending on the values of the effect modifiers we will get different values for the estimated odds ratios. Note that each estimated odds ratio obtained adjusts for the confounding effects of all five control variables because these five variables are contained in the fitted model as *V* variables.

Choice of *W* values depends on investigator

EXAMPLE

TABLE OF POINT ESTIMATES $\widehat{\text{ROR}}$

	HPT = 0	HPT = 1
CHL = 180	0.79	0.08
CHL = 200	3.16	0.31
CHL = 220	12.61	1.22
CHL = 240	50.33	4.89

In general, when faced with an odds ratio expression involving effect modifiers (*W*), the choice of values for the *W* variables depends primarily on the interest of the investigator. Typically, the investigator will choose a range of values for each interaction variable in the odds ratio formula; this choice will lead to a table of estimated odds ratios, such as the one presented here, for a range of CHL values and the two values of HPT. From such a table, together with a table of confidence intervals, the investigator can interpret the exposure–disease relationship.

EXAMPLE

No interaction model for Evans County data (n = 609)

$$\begin{aligned}\text{logit P}(\mathbf{X}) = \alpha &+ \beta\text{CAT} \\ &+ \gamma_1\text{AGE} + \gamma_2\text{CHL} \\ &+ \gamma_3\text{SMK} + \gamma_4\text{ECG} \\ &+ \gamma_5\text{HPT}\end{aligned}$$

As a second example, we consider a model containing no interaction terms from the same Evans County data set of 609 white males. The variables in the model are the exposure variable CAT, and five *V* variables, namely, AGE, CHL, SMK, ECG, and HPT. This model is written in logit form as shown here.

EXAMPLE (continued)

$\widehat{\text{ROR}} = \exp\left(\hat{\beta}\right)$

Because this model contains no interaction terms, the odds ratio expression for the CAT, CHD association is given by e to the $\hat{\beta}$, where $\hat{\beta}$ is the estimated coefficient of the exposure variable CAT.

Fitted model:

Variable	*Coefficient*
Intercept	$\hat{\alpha} = -6.7747$
CAT	$\hat{\beta} = 0.5978$
AGE	$\hat{\gamma}_1 = 0.0322$
CHL	$\hat{\gamma}_2 = 0.0088$
SMK	$\hat{\gamma}_3 = 0.8348$
ECG	$\hat{\gamma}_4 = 0.3695$
HPT	$\hat{\gamma}_5 = 0.4392$

$\widehat{\text{ROR}} = \exp(0.5978) = 1.82$

When fitting this no interaction model to the data, we obtain estimates of the model coefficients that are listed here.

For this fitted model, then, the odds ratio is given by e to the power 0.5978, which equals 1.82. Note that this odds ratio is a fixed number, which should be expected, as there are no interaction terms in the model.

EXAMPLE COMPARISON

	Interaction model	*No interaction model*
Intercept	−4.0497	−6.7747
CAT	−12.6894	0.5978
AGE	0.0350	0.0322
CHL	−0.0055	0.0088
SMK	0.7732	0.8348
ECG	0.3671	0.3695
HPT	1.0466	0.4392
CAT × CHL	0.0692	–
CAT × HPT	−2.3318	–

In comparing the results for the no interaction model just described with those for the model containing interaction terms, we see that the estimated coefficient for any variable contained in both models is different in each model. For instance, the coefficient of CAT in the *no interaction* model is 0.5978, whereas the coefficient of CAT in the *interaction* model is – 12.6894. Similarly, the coefficient of AGE in the no interaction model is 0.0322, whereas the coefficient of AGE in the interaction model is 0.0350.

Which model? Requires *strategy*

It should not be surprising to see different values for corresponding coefficients as the two models give a different description of the underlying relationship among the variables. To decide which of these models, or maybe what other model, is more appropriate for this data, we need to use a *strategy* for model selection that includes carrying out tests of significance. A discussion of such a strategy is beyond the scope of this presentation but is described elsewhere (see Chaps. 6 and 7).

This presentation is now complete. We have described important special cases of the logistic model, namely, models for

SUMMARY

1. Introduction

✓ 2. Important Special Cases

- simple analysis
- interaction assessment involving two variables
- assessment of potential confounding and interaction effects of several covariates

We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

3. Computing the Odds Ratio

All of the special cases in this presentation involved a (0, 1) exposure variable. In the next chapter, we consider how the odds ratio formula is modified for other codings of single exposures and also examine several exposure variables in the same model, controlling for potential confounders and effect modifiers.

Detailed Outline

I. **Overview** (page 45)

A. Focus:

- Simple analysis
- Multiplicative interaction
- Controlling several confounders and effect modifiers

B. Logistic model formula when $\mathbf{X} = (X_1, X_2, \ldots, X_k)$:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-\left(\alpha + \sum\limits_{i=1}^{k} \beta_i X_i\right)}}.$$

C. Logit form of logistic model:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \sum_{i=1}^{k} \beta_i X_i.$$

D. General odds ratio formula:

$$\mathrm{ROR}_{\mathbf{X}_1, \mathbf{X}_0} = \mathrm{e}^{\sum\limits_{i=1}^{k} \beta_i (X_{1i} - X_{0i})} = \prod_{i=1}^{k} \mathrm{e}^{\beta_i (X_{1i} - X_{0i})}.$$

II. **Special case – Simple analysis** (pages 46–49)

A. The model:

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1 + \mathrm{e}^{-(\alpha + \beta_1 E)}}$$

B. Logit form of the model:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \beta_1 E$$

C. Odds ratio for the model: $\mathrm{ROR} = \exp(\beta_1)$

D. Null hypothesis of no *E, D* effect: H_0: $\beta_1 = 0$.

E. The estimated odds ratio $\exp(\hat{\beta})$ is computationally equal to *ad*/*bc* where *a, b, c,* and *d* are the cell frequencies within the four-fold table for simple analysis.

III. **Assessing multiplicative interaction** (pages 49–55)

A. Definition of no interaction on a multiplicative scale: $\mathrm{OR}_{11} = \mathrm{OR}_{10} \times \mathrm{OR}_{01}$,
where OR_{AB} denotes the odds ratio that compares a person in category *A* of one factor and category *B* of a second factor with a person in referent categories 0 of both factors, where *A* takes on the values 0 or 1 and *B* takes on the values 0 or 1.

B. Conceptual interpretation of no interaction formula: The effect of both variables *A* and *B* acting together is the same as the combined effect of each variable acting separately.

C. Examples of no interaction and interaction on a multiplicative scale.

D. A logistic model that allows for the assessment of multiplicative interaction:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta_1 A + \beta_2 B + \beta_3 A \times B$$

E. The relationship of β_3 to the odds ratios in the no interaction formula above:

$$\beta_3 = \ln\left(\frac{\text{OR}_{11}}{\text{OR}_{10} \times \text{OR}_{01}}\right)$$

F. The null hypothesis of no interaction in the above two factor model: H_0: $\beta_3 = 0$.

IV. The *E, V, W* model – A general model containing a (0, 1) exposure and potential confounders and effect modifiers (pages 55–64)

A. Specification of variables in the model: start with *E*, $C_1, C_2, \ldots, C_p$; then specify potential confounders $V_1, V_2, \ldots, V_{p_1}$, which are functions of the *C*s, and potential interaction variables (i.e., effect modifiers) $W_1, W_2, \ldots, W_{p_2}$, which are also functions of the *C*s and go into the model as product terms with *E*, i.e., $E \times W_j$.

B. The *E, V, W* model:

$$\text{logit P}(\mathbf{X}) = \alpha + \beta E + \sum_{i=1}^{p_1} \gamma_i V_i + E \sum_{j=1}^{p_2} \delta_j W_j$$

C. Odds ratio formula for the *E, V, W* model, where *E* is a (0, 1) variable:

$$\text{ROR}_{E\,=\,1\text{ vs. }E\,=\,0} = \exp\left(\beta + \sum_{j=1}^{p_2} \delta_j W_j\right)$$

D. Odds ratio formula for *E, V, W* model if no interaction: ROR = exp(β).

E. Examples of the *E, V, W* model: with interaction and without interaction

Practice Exercises

True or False (Circle T or F)

T F 1. A logistic model for a simple analysis involving a (0, 1) exposure variable is given by logit P(**X**) $= \alpha + \beta E$, where E denotes the (0, 1) exposure variable.

T F 2. The odds ratio for the exposure–disease relationship in a logistic model for a simple analysis involving a (0, 1) exposure variable is given by β, where β is the coefficient of the exposure variable.

T F 3. The null hypothesis of no exposure–disease effect in a logistic model for a simple analysis is given by H_0: $\beta = 1$, where β is the coefficient of the exposure variable.

T F 4. The log of the estimated coefficient of a (0, 1) exposure variable in a logistic model for simple analysis is equal to ad/bc, where a, b, c, and d are the cell frequencies in the corresponding fourfold table for simple analysis.

T F 5. Given the model logit P(**X**) $= \alpha + \beta E$, where E denotes a (0, 1) exposure variable, the *risk* for exposed persons ($E = 1$) is expressible as e^{β}.

T F 6. Given the model logit P(**X**) $= \alpha + \beta E$, as in Exercise 5, the *odds* of getting the disease for exposed persons ($E = 1$) is given by $e^{\alpha+\beta}$.

T F 7. A logistic model that incorporates a multiplicative interaction effect involving two (0, 1) independent variables X_1 and X_2 is given by logit P(**X**) $= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$.

T F 8. An equation that describes "no interaction on a multiplicative scale" is given by $OR_{11} = OR_{10}/OR_{01}$.

T F 9. Given the model logit P(**X**) $= \alpha + \beta E + \gamma$SMK $+ \delta E \times$ SMK, where E is a (0, 1) exposure variable and SMK is a (0, 1) variable for smoking status, the null hypothesis for a test of no interaction on a multiplicative scale is given by H_0: $\delta = 0$.

T F 10. For the model in Exercise 9, the odds ratio that describes the exposure disease effect controlling for smoking is given by $\exp(\beta + \delta)$.

T F 11. Given an exposure variable E and control variables AGE, SBP, and CHL, suppose it is of interest to fit a model that adjusts for the potential confounding effects of all three control variables considered as main effect terms and for the potential interaction effects with E of all three control variables. Then the logit form of a model that describes this situation is given by logit P($\mathbf{X}$) = $\alpha + \beta E + \gamma_1\text{AGE} + \gamma_2\text{SBP} + \gamma_3\text{CHL} + \delta_1\text{AGE} \times \text{SBP} + \delta_2\text{AGE} \times \text{CHL} + \delta_3\text{SBP} \times \text{CHL}$.

T F 12. Given a logistic model of the form logit P($\mathbf{X}$) = $\alpha + \beta E + \gamma_1\text{AGE} + \gamma_2\text{SBP} + \gamma_3\text{CHL}$, where E is a (0, 1) exposure variable, the odds ratio for the effect of E adjusted for the confounding of AGE, CHL, and SBP is given by exp(β).

T F 13. If a logistic model contains interaction terms expressible as products of the form EW_j where W_j are potential effect modifiers, then the value of the odds ratio for the E, D relationship will be different, depending on the values specified for the W_j variables.

T F 14. Given the model logit P($\mathbf{X}$) = $\alpha + \beta E + \gamma_1\text{SMK} + \gamma_2\text{SBP}$, where E and SMK are (0, 1) variables, and SBP is continuous, then the odds ratio for estimating the effect of SMK on the disease, controlling for E and SBP is given by exp(γ_1).

T F 15. Given E, C_1, and C_2, and letting $V_1 = C_1 = W_1$ and $V_2 = C_2 = W_2$, then the corresponding logistic model is given by logit P($\mathbf{X}$) = $\alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_2 + E(\delta_1 C_1 + \delta_2 C_2)$.

T F 16. For the model in Exercise 15, if $C_1 = 20$ and $C_2 = 5$, then the odds ratio for the E, D relationship has the form exp($\beta + 20\delta_1 + 5\delta_2$).

Test

True or False (Circle T or F)

T F 1. Given the simple analysis model, logit P(**X**) $= \phi + \psi Q$, where ϕ and ψ are unknown parameters and Q is a (0, 1) exposure variable, the odds ratio for describing the exposure–disease relationship is given by $\exp(\phi)$.

T F 2. Given the model logit P(**X**) $= \alpha + \beta E$, where E denotes a (0, 1) exposure variable, the *risk* for unexposed persons ($E = 0$) is expressible as $1/\exp(-\alpha)$.

T F 3. Given the model in Question 2, the *odds* of getting the disease for unexposed persons ($E = 0$) is given by $\exp(\alpha)$.

T F 4. Given the model logit P(**X**) $= \phi + \psi\text{HPT} + \rho\text{ECG} + \pi\text{HPT} \times \text{ECG}$, where HPT is a (0, 1) exposure variable denoting hypertension status and ECG is a (0, 1) variable for electrocardiogram status, the null hypothesis for a test of no interaction on a multiplicative scale is given by H_0: $\exp(\pi) = 1$.

T F 5. For the model in Question 4, the odds ratio that describes the effect of HPT on disease status, controlling for ECG, is given by $\exp(\psi + \pi\text{ECG})$.

T F 6. Given the model logit P(**X**) $= \alpha + \beta E + \phi\text{HPT} + \psi\text{ECG}$, where E, HPT, and ECG are (0, 1) variables, then the odds ratio for estimating the effect of ECG on the disease, controlling for E and HPT, is given by $\exp(\psi)$.

T F 7. Given E, C_1, and C_2, and letting $V_1 = C_1 = W_1$, $V_2 = (C_1)^2$, and $V_3 = C_2$, then the corresponding logistic model is given by logit P(**X**) $= \alpha + \beta E + \gamma_1 C_1 + \gamma_2 C_1{}^2 + \gamma_3 C_2 + \delta E\text{C}_1$.

T F 8. For the model in Question 7, if $\text{C}_1 = 5$ and $C_2 = 20$, then the odds ratio for the E, D relationship has the form $\exp(\beta + 20\delta)$.

Consider a 1-year follow-up study of bisexual males to assess the relationship of behavioral risk factors to the acquisition of HIV infection. Study subjects were all in the 20–30 age range and were enrolled if they tested HIV negative and had claimed not to have engaged in "high-risk" sexual activity for at least 3 months. The outcome variable is HIV status at 1 year, a (0, 1) variable, where a subject gets the value 1 if HIV positive and 0 if HIV negative at 1 year after start of follow-up. Four risk factors were considered: consistent and correct condom use (CON), a (0, 1) variable; having one or more sex partners in high-risk groups (PAR), also a (0, 1) variable; the number of sexual partners (NP); and the average number of sexual contacts per month (ASCM). The primary purpose of this study was to determine the effectiveness of consistent and correct condom use in preventing the acquisition of HIV infection, controlling for the other variables. Thus, the variable CON is considered the exposure variable, and the variables PAR, NP, and ASCM are potential confounders and potential effect modifiers.

9. Within the above study framework, state the logit form of a logistic model for assessing the effect of CON on HIV acquisition, controlling for each of the other three risk factors as both potential confounders and potential effect modifiers. (Note: In defining your model, *only* use interaction terms that are two-way products of the form $E \times W$, where E is the exposure variable and W is an effect modifier.)
10. Using the model in Question 9, give an expression for the odds ratio that compares an exposed person (CON = 1) with an unexposed person (CON = 0) who has the same values for PAR, NP, and ASCM.

Answers to Practice Exercises

1. T
2. F: OR = e^{β}
3. F: H_0: $\beta = 0$
4. F: $e^{\beta} = ad/bc$
5. F: risk for $E = 1$ is $1/[1 + e^{-(\alpha+\beta)}]$
6. T
7. T
8. F: $OR_{11} = OR_{10} \times OR_{01}$
9. T
10. F: OR = $\exp(\beta + \delta\text{SMK})$
11. F: interaction terms should be E × AGE, E × SBP, and E × CHL
12. T
13. T
14. T
15. T
16. T

Springer

http://www.springer.com/978-1-4419-1741-6

Logistic Regression
A Self-Learning Text
Kleinbaum, D.G.; Klein, M.
2010, XVII, 702 p., Hardcover
ISBN: 978-1-4419-1741-6