1 Introduction to Logistic Regression

Contents

Introduction

This introduction to logistic regression describes the reasons for the popularity of the logistic model, the model form, how the model may be applied, and several of its key features, particularly how an odds ratio can be derived and computed for this model.

As preparation for this chapter, the reader should have some familiarity with the concept of a mathematical model, particularly a multiple-regression-type model involving independent variables and a dependent variable. Although knowledge of basic concepts of statistical inference is not required, the learner should be familiar with the distinction between population and sample, and the concept of a parameter and its estimate.

Abbreviated Outline

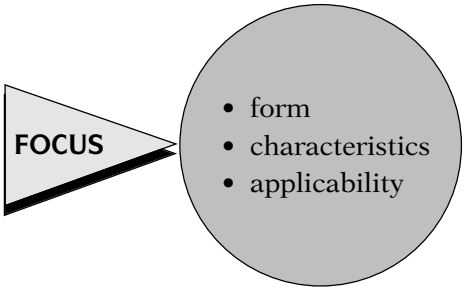
The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

Objectives

Upon completing this chapter, the learner should be able to:

1. Recognize the multivariable problem addressed by logistic regression in terms of the types of variables considered.
2. Identify properties of the logistic function that explain its popularity.
3. State the general formula for the logistic model and apply it to specific study situations.
4. Compute the estimated risk of disease development for a specified set of independent variables from a fitted logistic model.
5. Compute and interpret a risk ratio or odds ratio estimate from a fitted logistic model.
6. Identify the extent to which the logistic model is applicable to follow-up, case-control, and/or cross-sectional studies.
7. Identify the conditions required for estimating a risk ratio using a logistic model.
8. Identify the formula for the logit function and apply this formula to specific study situations.
9. Describe how the logit function is interpretable in terms of an "odds."
10. Interpret the parameters of the logistic model in terms of log odds.
11. Recognize that to obtain an odds ratio from a logistic model, you must specify **X** for two groups being compared.
12. Identify two formulae for the odds ratio obtained from a logistic model.
13. State the formula for the odds ratio in the special case of (0, 1) variables in a logistic model.
14. Describe how the odds ratio for (0, 1) variables is an "adjusted" odds ratio.
15. Compute the odds ratio, given an example involving a logistic model with (0, 1) variables and estimated parameters.
16. State a limitation regarding the types of variables in the model for use of the odds ratio formula for (0, 1) variables.

Presentation



This presentation focuses on the basic features of logistic regression, a popular mathematical modeling procedure used in the analysis of epidemiologic data. We describe the **form** and key **characteristics** of the model. Also, we demonstrate the **applicability** of logistic modeling in epidemiologic research.

I. The Multivariable Problem

We begin by describing the multivariable problem frequently encountered in epidemiologic research. A typical question of researchers is: What is the relationship of one or more exposure (or study) variables (E) to a disease or illness outcome (D)?

EXAMPLE

$D_{(0,1)}$ = CHD

$E_{(0,1)}$ = SMK

"control for"

C_1 = AGE

C_2 = RACE

C_3 = SEX

To illustrate, we will consider a dichotomous disease outcome with 0 representing **not diseased** and 1 representing **diseased.** The dichotomous disease outcome might be, for example, coronary heart disease (CHD) status, with subjects being classified as either 0 ("without CHD") or 1 ("with CHD").

Suppose, further, that we are interested in a single dichotomous exposure variable, for instance, smoking status, classified as "yes" or "no." The research question for this example is, therefore, to evaluate the extent to which smoking is associated with CHD status.

To evaluate the extent to which an exposure, like smoking, is associated with a disease, like CHD, we must often account or "control for" additional variables, such as age, race, and/or sex, which are not of primary interest. We have labeled these three control variables as C_1, C_2, and C_3.

In this example, the variable E (the exposure variable), together with C_1, C_2, and C_3 (the control variables), represent a collection of **independent** variables that we wish to use to describe or predict the **dependent** variable D.

Independent variables:
$X_1, X_2, \ldots, X_k$

More generally, the independent variables can be denoted as X_1, X_2, and so on up to X_k where k is the number of variables being considered.

X's may be E's, C's, or combinations

We have a **flexible** choice for the X's, which can represent any collection of exposure variables, control variables, or even combinations of such variables of interest.

EXAMPLE

$X_1 = E$	$X_4 = E \times C_1$
$X_2 = C_1$	$X_5 = C_1 \times C_2$
$X_3 = C_2$	$X_6 = E^2$

For example, we may have:

X_1 equal to an exposure variable E
X_2 and X_3 equal to control variables C_1 and C_2, respectively
X_4 equal to the product $E \times C_1$
X_5 equal to the product $C_1 \times C_2$
X_6 equal to E^2

The Multivariable Problem

$X_1, X_2, \ldots, X_k \Rightarrow D$

Whenever we wish to relate a set of X's to a dependent variable, like D, we are considering a **multivariable problem.** In the analysis of such a problem, some kind of **mathematical model** is typically used to deal with the complex interrelationships among many variables.

The analysis:
mathematical model

Logistic model:
dichotomous D

Logistic regression is a mathematical modeling approach that can be used to describe the relationship of several X's to a **dichotomous** dependent variable, such as D.

Logistic is most popular

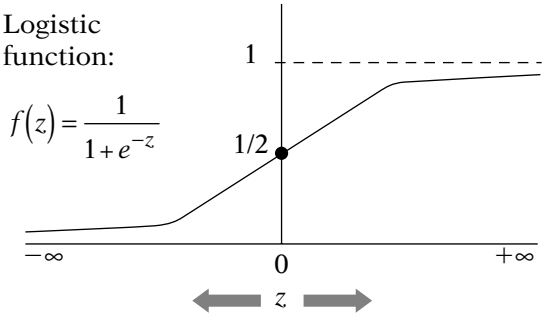
Other modeling approaches are possible also, but logistic regression is by far the most **popular** modeling procedure used to analyze epidemiologic data when the illness measure is dichotomous. We will show why this is true.

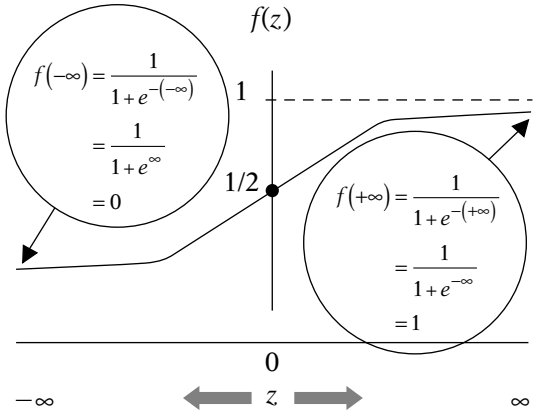
II. Why Is Logistic Regression Popular?

To explain the popularity of logistic regression, we show here the **logistic function,** which describes the mathematical form on which the **logistic model** is based. This function, called $f(z)$, is given by 1 over 1 plus e to the minus z. We have plotted the values of this function as z varies from $-\infty$ to $+\infty$.

Logistic function:

$$f(z) = \frac{1}{1+e^{-z}}$$





Notice, in the balloon on the left side of the graph, that when z is $-\infty$, the logistic function $f(z)$ equals 0.

On the right side, when z is $+\infty$, then $f(z)$ equals 1.

Range: $0 \le f(z) \le 1$

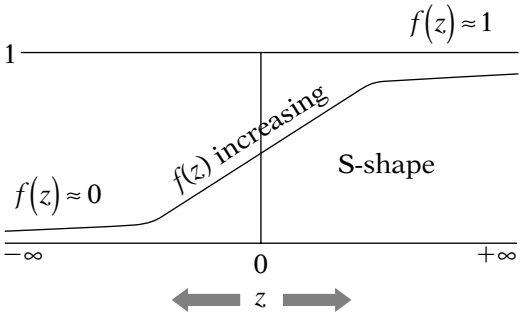
Thus, as the graph describes, the **range** of $f(z)$ is between 0 and 1, regardless of the value of z**.**

$0 \le$ probability ≤ 1
(individual risk)

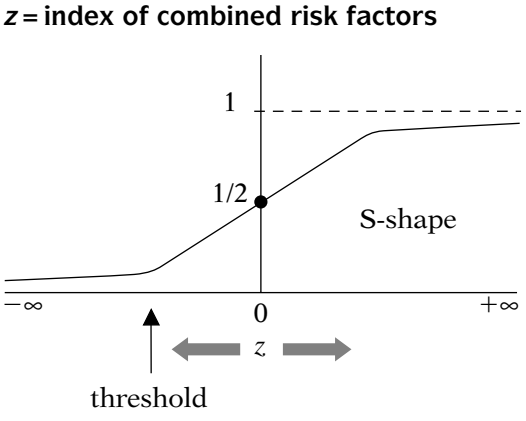
The fact that the logistic function $f(z)$ **ranges between 0 and 1** is the primary reason the logistic model is so popular. The model is designed to describe a probability, which is always some number between 0 and 1. In epidemiologic terms, such a probability gives the **risk** of an individual getting a disease.

The **logistic model**, therefore, is set up to ensure that whatever estimate of risk we get, it will always be some number between 0 and 1. Thus, for the logistic model, we can never get a risk estimate either above 1 or below 0. This is not always true for other possible models, which is why the logistic model is often the first choice when a probability is to be estimated.

Shape:



Another reason why the logistic model is popular derives from the **shape** of the logistic function. As shown in the graph, if we start at $z = -\infty$ and move to the right, then as z increases, the value of $f(z)$ hovers close to zero for a while, then starts to increase dramatically toward 1, and finally levels off around 1 as z increases toward $+\infty$. The result is an elongated, S-shaped picture.



The S-shape of the logistic function appeals to epidemiologists if the variable z is viewed as representing an index that combines contributions of several risk factors, and $f(z)$ represents the risk for a given value of z.

Then, the S-shape of $f(z)$ indicates that the effect of z on an individual's risk is minimal for low z's until some **threshold** is reached. The risk then rises rapidly over a certain range of intermediate z values, and then remains extremely high around 1 once z gets large enough.

This **threshold** idea is thought by epidemiologists to apply to a variety of disease conditions. In other words, an S-shaped model is considered to be widely applicable for considering the multivariable nature of an epidemiologic research question.

SUMMARY

So, the logistic **model** is **popular** because the logistic **function**, on which the model is based, provides:

- Estimates that must lie in the range between zero and one
- An appealing S-shaped description of the combined effect of several risk factors on the risk for a disease.

III. The Logistic Model

Now, let's go from the logistic **function** to the **model**, which is our primary focus.

$$z=\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k$$

To obtain the logistic **model** from the logistic **function,** we write z as the linear sum α plus β_1 times X_1 plus β_2 times X_2, and so on to β_k times X_k, where the X's are independent variables of interest and α and the β_i are constant terms representing unknown parameters.

$$z=\underbrace{\alpha+\beta_1X_1+\beta_2X_2+\ldots+\beta_kX_k}$$

$$f(z)=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-(\alpha+\Sigma\beta_iX_i)}}$$

In essence, then, z **is an index that combines the** X's.

We now substitute the linear sum expression for z in the right-hand side of the formula for $f(z)$ to get the expression $f(z)$ equals 1 over 1 plus e to minus the quantity α plus the sum of β_iX_i for i ranging from 1 to k. Actually, to view this expression as a mathematical model, we must place it in an epidemiologic context.

Epidemiologic framework

$X_1, X_2, \ldots, X_k$ measured at T_0

$P(D=1|X_1, X_2, \ldots, X_k)$

The logistic model considers the following general **epidemiologic study framework:** We have observed independent variables X_1, X_2, and so on up to X_k on a group of subjects, for whom we have also determined disease status, as either 1 if "with disease" or 0 if "without disease."

We wish to use this information to describe the probability that the disease will develop during a defined study period, say T_0 to T_1, in a disease-free individual with independent variable values X_1, X_2, up to X_k which are measured at T_0.

The probability being modeled can be denoted by the conditional probability statement $P(D=1 \mid X_1, X_2, \ldots, X_k)$.

DEFINITION

Logistic model:

$$P(D=1|X_1, X_2, \ldots, X_k) = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

$\uparrow \quad \uparrow$

unknown parameters

The model is defined as **logistic** if the expression for the probability of developing the disease, given the X's, is 1 over 1 plus e to minus the quantity α plus the sum from i equals 1 to k of β_i times X_i.

The terms α and β_i in this model represent **unknown parameters** that we need to estimate based on data obtained on the X's and on D (disease outcome) for a group of subjects.

Thus, if we knew the parameters α and the β_i and we had determined the values of X_1 through X_k for a particular disease-free individual, we could use this formula to plug in these values and obtain the probability that this individual would develop the disease over some defined follow-up time interval.

NOTATION

$P(D=1|X_1, X_2, \ldots, X_k)$

$=P(\mathbf{X})$

For notational convenience, we will denote the probability statement $P(D=1 \mid X_1, X_2, \ldots, X_k)$ as simply $P(\mathbf{X})$ where the **bold X** is a shortcut notation for the collection of variables X_1 through X_k.

Thus, the logistic model may be written as $P(\mathbf{X})$ equals 1 over 1 plus e to minus the quantity α plus the sum $\beta_i X_i$.

Model formula:

$$P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

IV. Applying the Logistic Model Formula

EXAMPLE

$D = \text{CHD}_{(0,\,1)}$

$X_1 = \text{CAT}_{(0,\,1)}$

$X_2 = \text{AGE}_{\text{continuous}}$

$X_3 = \text{ECG}_{(0,\,1)}$

n = 609 white males

9-year follow-up

$$\text{P}(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\beta_1\text{CAT}+\beta_2\text{AGE}+\beta_3\text{ECG})}}$$

DEFINITION

fit: use data to estimate

$\alpha, \beta_1, \beta_2, \beta_3$

NOTATION

hat= ˆ

parameter ⇔ estimator

$\alpha\ \beta_1\ \beta_2 \qquad \hat{\alpha}\ \hat{\beta}_1\ \hat{\beta}_2$

Method of estimation:
maximum likelihood (ML)—
see Chapters 4 and 5

EXAMPLE

$\hat{\alpha} = -3.911$

$\hat{\beta}_1 = 0.652$

$\hat{\beta}_2 = 0.029$

$\hat{\beta}_3 = 0.342$

To illustrate the use of the logistic model, suppose the disease of interest is *D* equals CHD. Here CHD is coded 1 if a person has the disease and 0 if not.

We have three independent variables of interest: X_1=CAT, X_2=AGE, and X_3=ECG. CAT stands for catecholamine level and is coded 1 if high and 0 if low, AGE is continuous, and ECG denotes electrocardiogram status and is coded 1 if abnormal and 0 if normal.

We have a data set of 609 white males on which we measured CAT, AGE, and ECG at the start of study. These people were then followed for 9 years to determine CHD status.

Suppose that in the analysis of this data set, we consider a logistic model given by the expression shown here.

We would like to **"fit"** this model; that is, we wish to use the data set to estimate the unknown parameters α, β_1, β_2, and β_3.

Using common statistical notation, we distinguish the parameters from their estimators by putting a **hat** symbol on top of a parameter to denote its estimator. Thus, the estimators of interest here are α "hat," β_1 "hat," β_2 "hat," and β_3 "hat."

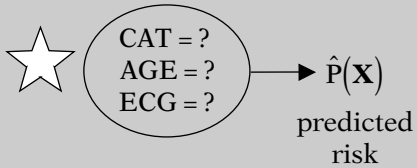
The method used to obtain these estimates is called **maximum likelihood** (ML). In two later chapters (Chapters 4 and 5), we describe how the ML method works and how to test hypotheses and derive confidence intervals about model parameters.

Suppose the results of our model fitting yield the estimated parameters shown on the left.

EXAMPLE (continued)

$$\hat{P}(\mathbf{X}) = \frac{1}{1+e^{-\left[-3.911+0.652(\text{CAT})+0.029(\text{AGE})+0.342(\text{ECG})\right]}}$$

$\hat{P}(\mathbf{X})$=?



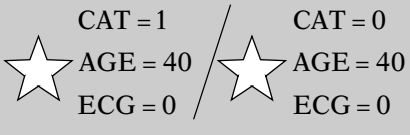
CAT = 1
AGE = 40
ECG = 0

$$\hat{P}(\mathbf{X}) = \frac{1}{1+e^{-\left[-3.911+0.652(1)+0.029(40)+0.342(0)\right]}}$$

$$= \frac{1}{1+e^{-(-2.101)}}$$

$$= \frac{1}{1+8.173}$$

$$= 0.1090$$



$$\frac{\hat{P}_1(\mathbf{X})}{\hat{P}_0(\mathbf{X})} = \frac{0.1090}{0.0600}$$

11% risk/6% risk

Our fitted model thus becomes $\hat{P}(\mathbf{X})$ equals 1 over 1 plus e to minus the linear sum -3.911 plus 0.652 times CAT plus 0.029 times AGE plus 0.342 times ECG. We have replaced P by $\hat{P}$ on the left-hand side of the formula because our estimated model will give us an estimated probability, not the exact probability.

Suppose we want to use our fitted model, to obtain the predicted risk for a **certain individual.**

To do so, we would need to specify the values of the independent variables (CAT, AGE, ECG) for this individual, and then plug these values into the formula for the fitted model to compute the estimated probability, $\hat{P}$ for this individual. This estimate is often called a "predicted risk," or simply "risk."

To illustrate the calculation of a predicted risk, suppose we consider an individual with CAT=1, AGE=40, and ECG=0.

Plugging these values into the fitted model gives us 1 over 1 plus e to minus the quantity -3.911 plus 0.652 times 1 plus 0.029 times 40 plus 0.342 times 0. This expression simplifies to 1 over 1 plus e to minus the quantity -2.101, which further reduces to 1 over 1 plus 8.173, which yields the value **0.1090**.

Thus, for a person with CAT=1, AGE=40, and ECG=0, the predicted risk obtained from the fitted model is 0.1090. That is, this person's estimated risk is about 11%.

Here, for the same fitted model, we compare the predicted risk of a person with CAT=1, AGE=40, and ECG=0 with that of a person with CAT=0, AGE=40, and ECG=0.

We previously computed the risk value of 0.1090 for the first person. The second probability is computed the same way, but this time we must replace CAT=1 with CAT=0. The predicted risk for this person turns out to be **0.0600.** Thus, using the fitted model, the person with a high catecholamine level has an **11% risk** for CHD, whereas the person with a low catecholamine level has a **6% risk** for CHD over the period of follow-up of the study.

EXAMPLE

$$\frac{\hat{P}_1(\mathbf{X})}{\hat{P}_0(\mathbf{X})} = \frac{0.109}{0.060} = 1.82 \quad \text{risk ratio } \widehat{\mathbf{(RR)}}$$

Note that, in this example, if we divide the predicted risk of the person with high catecholamine by that of the person with low catecholamine, we get a **risk ratio** estimate, denoted by $\widehat{\mathbf{RR}}$**,** of **1.82**. Thus, using the fitted model, we find that the person with high CAT has almost twice the risk of the person with low CAT, assuming both persons are of AGE 40 and have no previous ECG abnormality.

- RR (direct method)

We have just seen that it is possible to use a logistic model to obtain a risk ratio estimate that compares two types of individuals. We will refer to the approach we have illustrated above as the **direct method** for estimating RR.

Conditions for RR (direct method)
- ✓ follow-up study
- ✓ specify all *X*'s

Two conditions must be satisfied to estimate RR directly. First, we must have a **follow-up study** so that we can legitimately estimate individual risk. Second, for the two individuals being compared, we must **specify values for all the independent variables** in our fitted model to compute risk estimates for each individual.

- RR (indirect method)
  - ✓ OR
  - ✓ assumptions

If either of the above conditions is not satisfied, then we cannot estimate RR directly. That is, if our study design is not a follow-up study *or* if some of the *X*'s are not specified, we cannot estimate RR directly. Nevertheless, it may be possible to estimate RR **indirectly.** To do this, we must first compute an **odds ratio,** usually denoted as **OR,** and we must make some assumptions that we will describe shortly.

- OR: direct estimate from
  - ✓ follow-up
  - ✓ case-control
  - ✓ cross-sectional

In fact, **the odds ratio (OR),** not the risk ratio (RR), *is the only measure of association* **directly estimated** *from a logistic model (without requiring special assumptions), regardless of whether the study design is* **follow-up, case-control,** *or* **cross-sectional.** To see how we can use the logistic model to get an odds ratio, we need to look more closely at some of the features of the model.

V. Study Design Issues

★ Follow-up study orientation

$X_1, X_2, \ldots, X_k \Rightarrow \boldsymbol{D}_{(0,\,1)}$

An important feature of the logistic model is that it is defined with a **follow-up study orientation.** That is, as defined, this model describes the probability of developing a disease of interest expressed as a function of independent variables presumed to have been measured at the start of a fixed follow-up period. For this reason, it is natural to wonder whether the model can be applied to case-control or cross-sectional studies.

✓ **case-control**
✓ **cross-sectional**

The answer is **yes:** logistic regression can be applied to study designs other than follow-up.

Breslow and Day (1981)
Prentice and Pike (1979)

robust conditions
case-control studies

robust conditions
cross-sectional studies

Two papers, one by **Breslow and Day** in 1981 and the other by **Prentice and Pike** in 1979 have identified certain **"robust" conditions** under which the logistic model can be used with case-control data. "Robust" means that the conditions required, which are quite complex mathematically and equally as complex to verify empirically, apply to a large number of data situations that actually occur.

The reasoning provided in these papers carries over to **cross-sectional studies** also, though this has not been explicitly demonstrated in the literature.

Case control:

Follow-up:

In terms of **case-control** studies, it has been shown that even though cases and controls are selected first, after which previous exposure status is determined, the analysis may proceed as if the selection process were the other way around, as in a follow-up study.

Treat case control like follow-up

In other words, even with a case-control design, one can pretend, when doing the analysis, that the dependent variable is disease outcome and the independent variables are exposure status plus any covariates of interest. When using a logistic model with a case-control design, you can treat the data as if it came from a follow-up study, and still get a *valid* answer.

LIMITATION

case-control and
cross-sectional studies:

~~individual risk~~

✓ OR

Although logistic modeling is applicable to case-control and cross-sectional studies, there is one important **limitation** in the analysis of such studies. Whereas in follow-up studies, as we demonstrated earlier, a fitted logistic model can be used to predict the risk for an individual with specified independent variables, this model cannot be used to predict individual risk for case-control or cross-sectional studies. In fact, *only estimates of* **odds ratios** *can be obtained for case-control and cross-sectional studies.*

Simple Analysis

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

Risk: only in follow-up
OR: case-control or cross-sectional

The fact that only odds ratios, not individual risks, can be estimated from logistic modeling in case-control or cross-sectional studies is not surprising. This phenomenon is a carryover of a principle applied to simpler data analysis situations, in particular, to the simple analysis of a 2×2 table, as shown here.

For a 2×2 table, **risk estimates** can be used *only* if the data derive from a follow-up study, whereas only **odds ratios** are appropriate if the data derive from a case-control or cross-sectional study.

$\widehat{\text{OR}} = ad/bc$

To explain this further, recall that for 2×2 tables, the odds ratio is calculated as $\widehat{\text{OR}}$ equals a times d over b times c, where a, b, c, and d are the cell frequencies inside the table.

Case-control and cross-sectional studies:

$$= \frac{\hat{\text{P}}(E=1 \mid D=1) \,/\, \hat{\text{P}}(E=0 \mid D=1)}{\hat{\text{P}}(E=1 \mid D=0) \,/\, \hat{\text{P}}(E=0 \mid D=0)}$$

In case-control and cross-sectional studies, this OR formula can alternatively be written, as shown here, as a ratio involving probabilities for exposure status conditional on disease status.

$\hat{\text{P}}(E=1 \mid D=1)$
$\hat{\text{P}}(E=1 \mid D=0)$ → $\text{P}(E \mid D)$ (general form)

In this formula, for example, the term $\hat{\text{P}}(E{=}1 \mid D{=}1)$ is the estimated probability of being exposed, given that you are diseased. Similarly, the expression $\hat{\text{P}}(E{=}1 \mid D{=}0)$ is the estimated probability of being exposed given that you are not diseased. All the probabilities in this expression are of the general form $\text{P}(E \mid D)$.

Risk: $\text{P}(D|E)$
↓

$$\text{RR} = \frac{\hat{\text{P}}(D=1 \mid E=1)}{\hat{\text{P}}(D=1 \mid E=0)}$$

In contrast, in follow-up studies, formulae for risk estimates are of the form $\text{P}(D \mid E)$, in which the exposure and disease variables have been switched to the opposite side of the "given" sign.

For example, the risk ratio formula for follow-up studies is shown here. Both the numerator and denominator in this expression are of the form $\text{P}(D \mid E)$.

Case-control or cross-sectional studies:

~~P(D|E)~~

✓ $P(E|D) \Rightarrow$ risk

$$\hat{P}(\mathbf{X}) = \frac{1}{1 + e^{-\left(\hat{\alpha} + \Sigma \hat{\beta}_i X_i\right)}}$$

estimates

Case control:

~~$\hat{\alpha}$~~ $\Rightarrow \hat{P}($~~$\mathbf{X}$~~$)$

Follow-up:

$\hat{\alpha} \Rightarrow \hat{P}(\mathbf{X})$

Case-control and cross-sectional:

✓ $\hat{\beta}_i$, $\widehat{OR}$

Thus, in case-control or cross-sectional studies, risk estimates cannot be estimated because such estimates require conditional probabilities of the form $P(D|E)$, whereas only estimates of the form $P(E|D)$ are possible. This classic feature of a simple analysis also carries over to a logistic analysis.

There is a simple **mathematical explanation** for why predicted risks cannot be estimated using logistic regression for case-control studies. To see this, we consider the parameters α and the β's in the logistic model. To get a predicted risk $\hat{P}(\mathbf{X})$ from fitting this model, we must obtain valid estimates of α and the β's, these estimates being denoted by "hats" over the parameters in the mathematical formula for the model.

When using logistic regression for case-control data, the parameter α cannot be validly estimated without knowing the sampling fraction of the population. Without having a "good" estimate of α, we cannot obtain a good estimate of the predicted risk $\hat{P}(\mathbf{X})$ because $\hat{\alpha}$ is required for the computation.

In contrast, in follow-up studies, α can be estimated validly, and, thus, $P(\mathbf{X})$ can also be estimated.

Now, although α cannot be estimated from a case-control or cross-sectional study, the β's can be estimated from such studies. As we shall see shortly, the β's provide information about odds ratios of interest. Thus, even though we cannot estimate α in such studies, and therefore cannot obtain predicted risks, we can, nevertheless, obtain estimated measures of association in terms of odds ratios.

Note that if a logistic model is fit to case-control data, most computer packages carrying out this task will provide numbers corresponding to all parameters involved in the model, including α. This is illustrated here with some fictitious numbers involving three variables, X_1, X_2, and X_3. These numbers include a value corresponding to α, namely, -4.5, which corresponds to the constant on the list.

EXAMPLE

Printout

Variable	Coefficient
constant	$-4.50 = \hat{\alpha}$
X_1	$0.70 = \hat{\beta}_1$
X_2	$0.05 = \hat{\beta}_2$
X_3	$0.42 = \hat{\beta}_3$

~~α~~

EXAMPLE (repeated)

Printout

Variable	Coefficient
constant	$-4.50 = \hat{\alpha}$
X_1	$0.70 = \hat{\beta}_1$
X_2	$0.05 = \hat{\beta}_2$
X_3	$0.42 = \hat{\beta}_3$

However, according to mathematical theory, the value provided for the constant does not really estimate α. In fact, this value estimates some other parameter of no real interest. Therefore, an investigator should be forewarned that, even though the computer will print out a number corresponding to the constant α, the number will not be an appropriate estimate of α in case-control or cross-sectional studies.

SUMMARY

	Logistic Model	$\hat{P}(\mathbf{X})$	OR
Follow-up	✓	✓	✓
Case-control	✓	X	✓
Cross-sectional	✓	X	✓

We have described that the logistic model can be applied to case-control and cross-sectional data, even though it is intended for a follow-up design. When using case-control or cross-sectional data, however, a key limitation is that you cannot estimate risks like $\hat{P}(\mathbf{X})$, even though you can still obtain odds ratios. This limitation is not extremely severe if the goal of the study is to obtain a valid estimate of an exposure–disease association in terms of an odds ratio.

VI. Risk Ratios Versus Odds Ratios

OR

vs. ? follow-up study

RR

The use of an odds ratio estimate may still be of some concern, particularly when the study is a follow-up study. In follow-up studies, it is commonly preferred to estimate a risk ratio rather than an odds ratio.

EXAMPLE

$$\widehat{RR} = \frac{\hat{P}(CHD = 1 \mid CAT = 1,\ AGE = 40,\ ECG = 0)}{\hat{P}(CHD = 1 \mid CAT = 0,\ AGE = 40,\ ECG = 0)}$$

Model:

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \beta_1 CAT + \beta_2 AGE + \beta_3 ECG)}}$$

We previously illustrated that a risk ratio can be estimated for follow-up data provided all the independent variables in the fitted model are specified. In the example, we showed that we could estimate the risk ratio for CHD by comparing high catecholamine persons (that is, those with CAT=1) to low catecholamine persons (those with CAT=0), given that both persons were 40 years old and had no previous ECG abnormality. Here, we have specified values for all the independent variables in our model, namely, CAT, AGE, and ECG, for the two types of persons we are comparing.

EXAMPLE (continued)

$$\widehat{RR} = \frac{\hat{P}(CHD = 1 \mid CAT = 1,\ AGE = 40,\ ECG = 0)}{\hat{P}(CHD = 1 \mid CAT = 0,\ AGE = 40,\ ECG = 0)}$$

AGE uspecified but fixed

ECG unspecified but fixed

Nevertheless, it is more common to obtain an estimate of a risk ratio or odds ratio without explicitly specifying the control variables. In our example, for instance, it is typical to compare high CAT with low CAT persons keeping the control variables like AGE and ECG fixed but unspecified. In other words, the question is typically asked, What is the effect of the CAT variable controlling for AGE and ECG, considering persons who have the same AGE and ECG *regardless* of the values of these two variables?

Control variables unspecified:

$\widehat{OR}$ directly

$\widehat{RR}$ indirectly
provided $\widehat{OR} \approx \widehat{RR}$

When the control variables are generally considered to be fixed, but **unspecified,** as in the last example, we can use logistic regression to obtain an estimate of the odds ratio **directly,** but we cannot estimate the risk ratio. We can, however, stretch our interpretation to obtain a risk ratio **indirectly** provided we are willing to make certain assumptions. The key **assumption** here is that the odds ratio provides a good approximation to the risk ratio.

$\widehat{OR} \approx \widehat{RR}$ if rare disease

From previous exposure to epidemiologic principles, you may recall that one way to justify an odds ratio approximation for a risk ratio is to assume that the disease is rare. Thus, if we invoke the **rare disease assumption,** we can assume that the odds ratio estimate from a logistic regression model approximates a risk ratio.

Rare disease	OR	RR
yes	✓	✓
no	✓	?

If we cannot invoke the rare disease assumption, we cannot readily claim that the odds ratio estimate obtained from logistic modeling approximates a risk ratio. The investigator, in this case, may have to review the specific characteristics of the study before making a decision. It may be necessary to conclude that the odds ratio is a satisfactory measure of association in its own right for the current study.

VII. Logit Transformation

OR: Derive and Compute

Having described why the odds ratio is the primary parameter estimated when fitting a logistic regression model, we now explain how an odds ratio is derived and computed from the logistic model.

Logit

To begin the description of the odds ratio in logistic regression, we present an alternative way to write the logistic model, called the **logit form** of the model. To get the **logit** from the logistic model, we make a transformation of the model.

$$\text{logit } P(\mathbf{X}) = \ln_e\left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right]$$

where

$$P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

The **logit transformation,** denoted as **logit** P(**X**), is given by the natural log (i.e., to the base *e*) of the quantity P(**X**) divided by one minus P(**X**), where P(**X**) denotes the logistic model as previously defined.

This transformation allows us to compute a number, called **logit** P(**X**), for an individual with independent variables given by **X**. We do so by:

(1) $P(\mathbf{X})$
(2) $1-P(\mathbf{X})$
(3) $\frac{P(\mathbf{X})}{1-P(\mathbf{X})}$
(4) $\ln_e\left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right]$

(1) computing P(**X**) and
(2) 1 minus P(**X**) separately, then
(3) dividing one by the other, and finally
(4) taking the natural log of the ratio.

EXAMPLE

(1) $P(\mathbf{X}) = 0.110$
(2) $1-P(\mathbf{X}) = 0.890$
(3) $\frac{P(\mathbf{X})}{1-P(\mathbf{X})} = \frac{0.110}{0.890} = 0.123$
(4) $\ln_e\left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right] = \ln(0.123) = -2.096$
i.e., logit $(0.110) = -2.096$

For example, if P(**X**) is 0.110, then

1 minus P(**X**) is 0.890,

the ratio of the two quantities is 0.123,

and the log of the ratio is −2.096.

That is, the **logit** of 0.110 is −2.096.

$$\text{logit } P(\mathbf{X}) = \ln_e\left[\frac{P(\mathbf{X})}{1-P(\mathbf{X})}\right] = ?$$

$$P(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

Now we might ask, **what general formula do we get when we plug the logistic model form into the logit function? What kind of interpretation can we give to this formula? How does this relate to an odds ratio?**

Let us consider the formula for the logit function. We start with P(**X**), which is 1 over 1 plus *e* to minus the quantity α plus the sum of the $\beta_i X_i$.

$$1-\mathrm{P}(\mathbf{X}) = 1 - \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

$$= \frac{e^{-(\alpha+\Sigma\beta_i X_i)}}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

Also, using some algebra, we can write $1-\mathrm{P}(\mathbf{X})$ as:

e to minus the quantity α plus the sum of $\beta_i X_i$ divided by one over 1 plus *e* to minus α plus the sum of the $\beta_i X_i$.

$$\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})} = \frac{\dfrac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}}{\dfrac{e^{-(\alpha+\Sigma\beta_i X_i)}}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}}$$

If we divide $\mathrm{P}(\mathbf{X})$ by $1-\mathrm{P}(\mathbf{X})$, then the denominators cancel out,

$$= e^{(\alpha+\Sigma\beta_i X_i)}$$

and we obtain *e* to the quantity α plus the sum of the $\beta_i X_i$.

$$\ln_e\left[\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})}\right] = \ln_e\left[e^{(\alpha+\Sigma\beta_i X_i)}\right]$$

$$= \underbrace{(\alpha+\Sigma\beta_i X_i)}_{\text{linear sum}}$$

We then compute the natural log of the formula just derived to obtain:

the linear sum α plus the sum of $\beta_i X_i$.

Thus, the **logit** of $\mathrm{P}(\mathbf{X})$ simplifies to the **linear sum** found in the denominator of the formula for $\mathrm{P}(\mathbf{X})$.

Logit form:

$$\text{logit } \mathrm{P}(\mathbf{X}) = \alpha + \Sigma\beta_i X_i$$

where

$$\mathrm{P}(\mathbf{X}) = \frac{1}{1+e^{-(\alpha+\Sigma\beta_i X_i)}}$$

For the sake of convenience, many authors describe the logistic model in its logit form rather than in its original form as $\mathrm{P}(\mathbf{X})$. Thus, when someone describes a model as **logit** $\mathrm{P}(\mathbf{X})$ equal to a linear sum, we should recognize that a logistic model is being used.

logit $\mathrm{P}(\mathbf{X})$? OR

$$\frac{\mathrm{P}(\mathbf{X})}{1-\mathrm{P}(\mathbf{X})} = \text{odds for individual } X$$

Now, having defined and expressed the formula for the logit form of the logistic model, we ask, **where does the odds ratio come in?** As a preliminary step to answering this question, we first look more closely at the definition of the logit function. In particular, the quantity $\mathrm{P}(\mathbf{X})$ divided by $1-\mathrm{P}(\mathbf{X})$, whose log value gives the **logit,** describes the **odds** for developing the disease for a person with independent variables specified by **X**.

$$\text{odds} = \frac{P}{1-P}$$

In its simplest form, an **odds** is the ratio of the probability that some event will occur over the probability that the same event will not occur. The formula for an odds is, therefore, of the form *P* divided by $1-P$, where *P* denotes the probability of the event of interest.

EXAMPLE

$P = 0.25$

$$\text{odds} = \frac{P}{1-P} = \frac{0.25}{0.75} = \frac{1}{3}$$

$\frac{1}{3} \leftarrow$ event occurs
$\frac{}{3} \leftarrow$ event does not occur

3 to 1 event will not happen

$$\text{odds}: \left[\frac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right] \text{ vs. } \frac{P}{1-P}$$

describes risk in logistic model for individual **X**

$$\text{logit P}(\mathbf{X}) = \ln_e\left[\frac{\text{P}(\mathbf{X})}{1-\text{P}(\mathbf{X})}\right]$$
$$= \text{log odds for individual } \mathbf{X}$$
$$= \alpha + \Sigma\beta_i X_i$$

EXAMPLE

all $X_i = 0$: logit P(**X**) = ?

0

$$\text{logit P}(\mathbf{X}) = \alpha + \Sigma\beta_i X_i$$
$$\text{logit P}(\mathbf{X}) \Rightarrow \alpha$$

INTERPRETATION

(1) α = log odds for individual with all $X_i = 0$

For example, if P equals 0.25, then $1-P$, the probability of the opposite event, is 0.75 and the **odds** is 0.25 over 0.75, or one-third.

An **odds** of one-third can be interpreted to mean that the probability of the event occurring is one-third the probability of the event not occurring. Alternatively, we can state that the **odds** are **3 to 1** that the event will not happen.

The expression P(**X**) divided by $1-$P(**X**) has essentially the same interpretation as P over $1-P$, which ignores **X.**

The main difference between the two formulae is that the expression with the **X** is more specific. That is, the formula with **X** assumes that the probabilities describe the risk for developing a disease, that this risk is determined by a logistic model involving independent variables summarized by **X**, and that we are interested in the odds associated with a particular specification of **X**.

Thus, the logit form of the logistic model, shown again here, gives an expression for the **log odds** of developing the disease for an individual with a specific set of X's.

And, mathematically, this expression equals α plus the sum of the $\beta_i X_i$.

As a simple example, consider what the **logit** becomes when all the X's are 0. To compute this, we need to work with the mathematical formula, which involves the unknown parameters and the X's.

If we plug in 0 for all the X's in the formula, we find that the logit of P(**X**) reduces simply to α.

Because we have already seen that any logit can be described in terms of an **odds,** we can interpret this result to give some meaning to the parameter α.

One interpretation is that α gives the **log odds** for a person with zero values for all X's.

EXAMPLE (continued)

(2) α = log of background odds

LIMITATION OF (1)

All $X_i = 0$ for any individual?

↓

AGE $\neq 0$

WEIGHT $\neq 0$

(2) α = log of background odds

DEFINITION OF (2)

background odds: ignores all X's

$$\text{model}: P(\mathbf{X}) = \frac{1}{1+e^{-\alpha}}$$

α ✓

β_i?

$X_1, X_2, \ldots, X_i, \ldots, X_k$

fixed varies fixed

EXAMPLE

CAT changes from 0 to 1;

AGE = 40, ECG = 0

fixed

logit $P(\mathbf{X}) = \alpha + \beta_1 CAT + \beta_2 AGE + \beta_3 ECG$

A second interpretation is that α gives the **log** of the **background,** *or* **baseline, odds.**

The first interpretation for α, which considers it as the **log odds** for a person with 0 values for all X's, has a serious limitation: There may not be any person in the population of interest with zero values on all the X's.

For example, no subject could have zero values for naturally occurring variables, like age or weight. Thus, it would not make sense to talk of a person with zero values for all X's.

The second interpretation for α is more appealing: to describe it as the **log** of the **background,** *or* **baseline, odds.**

By background odds, we mean the odds that would result for a logistic model without any X's at all.

The form of such a model is 1 over 1 plus e to minus α. We might be interested in this model to obtain a baseline risk or odds estimate that ignores all possible predictor variables. Such an estimate can serve as a starting point for comparing other estimates of risk or odds when one or more X's are considered.

Because we have given an interpretation to α, can we also give an interpretation to β_i? Yes, we can, in terms of either **odds** or **odds ratios.** We will turn to odds ratios shortly.

With regard to the odds, we need to consider what happens to the logit when only one of the X's varies while keeping the others fixed.

For example, if our X's are CAT, AGE, and ECG, we might ask what happens to the logit when CAT changes from 0 to 1, given an AGE of 40 and an ECG of 0.

To answer this question, we write the model in **logit form** as $\alpha + \beta_1 CAT + \beta_2 AGE + \beta_3 ECG$.

EXAMPLE (continued)

(1) CAT = 1, AGE = 40, ECG = 0

$$\text{logit } P(\mathbf{X}) = \alpha+\beta_1 1+\beta_2 40+\beta_3 0$$
$$=(\alpha+\beta_1+40\beta_2)$$

(2) CAT = 0, AGE = 40, ECG = 0

$$\text{logit } P(\mathbf{X}) = \alpha+\beta_1 0+\beta_2 40+\beta_3 0$$
$$=(\alpha+40\beta_2)$$

$$\text{logit } P_1(\mathbf{X}) - \text{logit } P_0(\mathbf{X})$$
$$= (\alpha+\beta_1+40\beta_2) - (\alpha+40\beta_2)$$
$$=(\beta_1)$$

NOTATION

$\triangle$ = change

$\beta_1 = \triangle$ logit = $\triangle$ log odds when $\triangle$ CAT = 1, AGE and ECG fixed

The first expression below this model shows that when CAT=1, AGE=40, and ECG=0, this logit reduces to $\alpha + \beta_1 + 40\beta_2$.

The second expression shows that when CAT=0, but AGE and ECG remain fixed at 40 and 0, respectively, the logit reduces to $\alpha + 40\,\beta_2$.

If we subtract the **logit for CAT=0** from the **logit for CAT=1,** after a little arithmetic, we find that the difference is β_1, the coefficient of the variable CAT.

Thus, letting the symbol $\triangle$ denote change, we see that β_1 represents the change in the logit that would result from a unit change in CAT, when the other variables are fixed.

An equivalent explanation is that β_1 represents the *change in the log odds that would result from a one unit change* in the variable CAT when the other variables are fixed. These two statements are equivalent because, by definition, a *logit* is a *log odds,* so that the difference between two logits is the same as the difference between two log odds.

$$\text{logit } P(\mathbf{X}) = \alpha+\Sigma\beta_i X_i$$

More generally, using the logit expression, if we focus on any coefficient, say β_L, for i=L, we can provide the following interpretation:

$i = L$:

$\beta_L = \triangle \ln(\text{odds})$

when $= \triangle X_L = 1$, other X's fixed

β_L represents the change in the log odds that would result from a one unit change in the variable X_L, when all other X's are fixed.

SUMMARY

logit $P(\mathbf{X})$

α = background log odds

β_i = change in log odds

In summary, by looking closely at the expression for the logit function, we provide some interpretation for the parameters α and β_i in terms of odds, actually *log odds*.

Now, how can we use this information about logits to obtain an **odds ratio,** rather than an odds? After all, we are typically interested in measures of association, like odds ratios, when we carry out epidemiologic research.

VIII. Derivation of OR Formula

Any **odds ratio,** by definition, is a ratio of two odds, written here as $\mathbf{odds_1}$ divided by $\mathbf{odds_0}$, in which the subscripts indicate two individuals or two groups of individuals being compared.

$$\text{OR} = \frac{\text{odds}_1}{\text{odds}_0}$$

EXAMPLE

(1) CAT = 1, AGE = 40, ECG = 0

(0) CAT = 0, AGE = 40, ECG = 0

Now we give an example of an odds ratio in which we compare two groups, called group 1 and group 0. Using our CHD example involving independent variables CAT, AGE, and ECG, group 1 might denote persons with CAT=1, AGE=40, and ECG=0, whereas group 0 might denote persons with CAT=0, AGE=40, and ECG=0.

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

More generally, when we describe an odds ratio, the two groups being compared can be defined in terms of the bold **X** symbol, which denotes a general collection of X variables, from 1 to k.

(1) $\mathbf{X_1} = (X_{11}, X_{12}, \ldots, X_{1k})$

(0) $\mathbf{X_0} = (X_{01}, X_{02}, \ldots, X_{0k})$

Let $\mathbf{X}_1$ denote the collection of X's that specify group 1 and let $\mathbf{X}_0$ denote the collection of X's that specify group 0.

EXAMPLE

$\mathbf{X}$ = (CAT, AGE, ECG)

(1) $\mathbf{X}_1$ = (CAT = 1, AGE = 40, ECG = 0)
(0) $\mathbf{X}_0$ = (CAT = 0, AGE = 40, ECG = 0)

In our example, then, k, the number of variables, equals 3, and

X is the collection of variables CAT, AGE, and ECG,
$\mathbf{X}_1$ corresponds to CAT=1, AGE=40, and ECG=0, whereas
$\mathbf{X}_0$ corresponds to CAT=0, AGE=40 and ECG=0.

NOTATION

$$\text{OR}_{\mathbf{X}_1, \mathbf{X}_0} = \frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0}$$

Notationally, to distinguish the two groups $\mathbf{X}_1$ and $\mathbf{X}_0$ in an **odds ratio,** we can write $\text{OR}_{\mathbf{X}_1, \mathbf{X}_0}$, which equals the **odds** for $\mathbf{X}_1$ *divided by* the *odds* for $\mathbf{X}_0$.

We will now apply the logistic model to this expression to obtain a general odds ratio formula involving the logistic model parameters.

$$P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \Sigma\beta_i X_i)}}$$

Given a logistic model of the general form P(**X**),

$$(1)\ \text{odds}: \frac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}$$

we can write the **odds** for **group 1** as $P(\mathbf{X}_1)$ divided by $1 - P(\mathbf{X}_1)$

$$(0)\ \text{odds}: \frac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}$$

and the **odds** for **group 0** as $P(\mathbf{X}_0)$ divided by $1 - P(\mathbf{X}_0)$.

$$\frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0} = \frac{\dfrac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}}{\dfrac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}} = \text{ROR}_{\mathbf{X}_1, \mathbf{X}_0}$$

To get an odds ratio, we then divide the first odds by the second odds. The result is an expression for the odds ratio written in terms of the two risks $P(\mathbf{X}_1)$ and $P(\mathbf{X}_0)$, that is, $P(\mathbf{X}_1)$ over $1 - P(\mathbf{X}_1)$ divided by $P(\mathbf{X}_0)$ over $1 - P(\mathbf{X}_0)$.

We denote this ratio as **ROR**, for **risk odds ratio,** as the probabilities in the odds ratio are all defined as risks. However, we still do not have a convenient formula.

$$\text{ROR} = \frac{\dfrac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)}}{\dfrac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)}} \qquad P(\mathbf{X}) = \frac{1}{1 + e^{-(\alpha + \Sigma\beta_i X_i)}}$$

Now, to obtain a convenient computational formula, we can substitute the mathematical expression 1 over 1 plus e to minus the quantity $(\alpha + \Sigma\beta_i X_i)$ for P(**X**) into the **risk odds ratio** formula above.

$$(1)\ \frac{P(\mathbf{X}_1)}{1 - P(\mathbf{X}_1)} = e^{(\alpha + \Sigma\beta_i X_{1i})}$$

For group 1, the **odds** $P(\mathbf{X}_1)$ over $1 - P(\mathbf{X}_1)$ reduces algebraically to e to the linear sum α plus the sum of β_i times X_{1i}, where X_{1i} denotes the value of the variable X_i for group 1.

$$(0)\ \frac{P(\mathbf{X}_0)}{1 - P(\mathbf{X}_0)} = e^{(\alpha + \Sigma\beta_i X_{0i})}$$

Similarly, the odds for group 0 reduces to e to the linear sum α plus the sum of β_i times X_{0i}, where X_{0i} denotes the value of variable X_i for group 0.

$$\text{ROR}_{\mathbf{X}_1, \mathbf{X}_0} = \frac{\text{odds for } \mathbf{X}_1}{\text{odds for } \mathbf{X}_0} = \frac{e^{(\alpha + \Sigma\beta_i X_{1i})}}{e^{(\alpha + \Sigma\beta_i X_{0i})}}$$

To obtain the **ROR**, we now substitute in the numerator and denominator the exponential quantities just derived to obtain e to the group 1 linear sum divided by e to the group 0 linear sum.

$$\text{Algebraic theory}: \frac{e^a}{e^b} = e^{a-b}$$

$$a = \alpha + \beta_i X_{1i},\ b = \alpha + \beta_i X_{0i}$$

The above expression is of the form e to the a divided by e to the b, where a and b are linear sums for groups 1 and 0, respectively. From algebraic theory, it then follows that this ratio of two exponentials is equivalent to e to the difference in exponents, or e to the a minus b.

$$\text{ROR} = e^{(\alpha+\Sigma\beta_i X_{1i})-(\alpha+\Sigma\beta_i X_{0i})}$$

We then find that the **ROR** equals e to the difference between the two linear sums.

$$= e^{\left[\alpha-\alpha+\Sigma\beta_i(X_{1i}-X_{0i})\right]}$$

In computing this difference, the α's cancel out and the β_i's can be factored for the ith variable.

$$= e^{\Sigma\beta_i(X_{1i}-X_{0i})}$$

Thus, the expression for **ROR** simplifies to the quantity e to the sum β_i times the difference between X_{1i} and X_{0i}.

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

We thus have a general exponential formula for the risk odds ratio from a logistic model comparing any two groups of individuals, as specified in terms of $\boldsymbol{X}_1$ and $\boldsymbol{X}_0$. Note that the formula involves the β_i's but not α.

$$e^{a+b} = e^a \times e^b$$

We can give an equivalent alternative to our ROR formula by using the algebraic rule that says that the exponential of a sum is the same as the product of the exponentials of each term in the sum. That is, e to the a plus b equals e to the a times e to the b.

$$e^{\sum_{i=1}^{k} z_i} = e^{z_1} \times e^{z_2} \times \cdots e^{z_k}$$

More generally, e to the sum of z_i equals the product of e to the z_i over all i, where the z_i's denote any set of values.

NOTATION

$$= \prod_{i=1}^{k} e^{z_i}$$

$$z_i = \beta_i(X_{1i}-X_{0i})$$

We can alternatively write this expression using the product symbol Π, where Π is a mathematical notation which denotes the product of a collection of terms.

Thus, using algebraic theory and letting z_i correspond to the term β_i times $(X_{1i} - X_{0i})$,

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

we obtain the **alternative formula** for **ROR** as the product from i=1 to k of e to the β_i times the difference $(X_{1i} - X_{0i})$

$$\prod_{i=1}^{k} e^{\beta_i(X_{1i}-X_{0i})}$$

That is, Π of e to the β_i times $(X_{1i} - X_{0i})$ equals e to the β_1 times $(X_{11} - X_{01})$ multiplied by e to the β_2 times $(X_{12} - X_{02})$ multiplied by additional terms, the final term

$$= e^{\beta_1(X_{11}-X_{01})}\, e^{\beta_2(X_{12}-X_{02})} \ldots e^{\beta_k(X_{1k}-X_{0k})}$$

being e to the β_k times $(X_{1k} - X_{0k})$.

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

- Multiplicative

The **product formula** for the **ROR**, shown again here, gives us an interpretation about how each variable in a logistic model contributes to the odds ratio.

In particular, we can see that each of the variables X_i contributes jointly to the odds ratio in a **multiplicative** way.

For example, if

EXAMPLE

$e^{\beta_2(X_{12}-X_{02})} = 3$

$e^{\beta_5(X_{15}-X_{05})} = 4$

$3 \times 4 = 12$

e to the β_i times $(X_{1i} - X_{0i})$ is

3 for variable 2 and

4 for variable 5,

then the joint contribution of these two variables to the odds ratio is **3** $\times$ **4**, or **12**.

Logistic model $\Rightarrow$ multiplicative OR formula

Thus, the product or Π formula for **ROR** tells us that, when the logistic model is used, the contribution of the variables to the odds ratio is **multiplicative.**

Other models $\Rightarrow$ other OR formulae

A model different from the logistic model, depending on its form, might imply a different (for example, an additive) contribution of variables to the odds ratio. An investigator not willing to allow a multiplicative relationship may, therefore, wish to consider other models or other OR formulae. Other such choices are beyond the scope of this presentation.

IX. Example of OR Computation

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

Given the choice of a logistic model, the version of the formula for the **ROR**, shown here as the exponential of a sum, is the most useful for computational purposes.

EXAMPLE

X = (CAT, AGE, ECG)
(1) CAT = 1, AGE = 40, ECG = 0
(0) CAT = 0, AGE = 40, ECG = 0

For example, suppose the **X**'s are CAT, AGE, and ECG, as in our earlier examples.

Also suppose, as before, that we wish to obtain an expression for the odds ratio that compares the following two groups: **group 1** with CAT=1, AGE=40, and ECG=0, and **group 0** with CAT=0, AGE=40, and ECG=0.

$\mathbf{X}_1$ = (CAT = 1, AGE = 40, ECG = 0)

For this situation, we let $\mathbf{X}_1$ be specified by CAT=1, AGE=40, and ECG=0,

EXAMPLE (continued)

$\mathbf{X}_0$ = (CAT = 0, AGE = 40, ECG = 0)

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

$$= e^{\beta_1(1-0)+\beta_2(40-40)+\beta_3(0-0)}$$

$$= e^{\beta_1+0+0}$$

$= e^{\beta_1}$ ← coefficient of CAT in

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{ECG}$$

$$\text{ROR}_{\mathbf{X}_1,\mathbf{X}_0} = e^{\beta_1}$$

(1) CAT = 1, AGE = 40, ECG = 0
(0) CAT = 0, AGE = 40, ECG = 0

$$\text{ROR}_{\mathbf{X}_1,\mathbf{X}_0} = e^{\beta_1}$$

= an "adjusted" OR

AGE and ECG:

- fixed
- same
- control variables

e^{β_1}: population ROR

$e^{\hat{\beta}_1}$: estimated ROR

and let $\mathbf{X}_0$ be specified by CAT=0, AGE=40, and ECG=0.

Starting with the general formula for the **ROR**, we then substitute the values for the $\mathbf{X}_1$ and $\mathbf{X}_0$ variables in the formula.

We then obtain **ROR** equals e to the β_1 times $(1 - 0)$ plus β_2 times $(40 - 40)$ plus β_3 times $(0 - 0)$.

The last two terms reduce to 0,

so that our final expression for the **odds ratio** is e to the β_1, where β_1 is the coefficient of the variable CAT.

Thus, for our example, even though the model involves the three variables CAT, ECG, and AGE, the odds ratio expression comparing the two groups involves only the parameter involving the variable CAT. Notice that of the three variables in the model, the variable CAT is the only variable whose value is different in groups 1 and 0. In both groups, the value for AGE is 40 and the value for ECG is 0.

The formula e to the β_1 may be interpreted, in the context of this example, as an **adjusted odds ratio.** This is because we have derived this expression from a logistic model containing two other variables, namely, AGE, and ECG, in addition to the variable CAT. Furthermore, we have fixed the values of these other two variables to be the same for each group. Thus, e to β_1 gives an odds ratio for the effect of the CAT variable **adjusted** for AGE and ECG, where the latter two variables are being treated as **control variables.**

The expression e to the β_1 denotes a population odds ratio parameter because the term β_1 is itself an unknown population parameter.

An estimate of this population odds ratio would be denoted by e to the $\hat{\beta}_1$. This term, $\hat{\beta}_1$, denotes an **estimate** of β_1 obtained by using some computer package to fit the logistic model to a set of data.

X. Special Case for (0, 1) Variables

Adjusted OR = e^{β}
where β = coefficient of (0, 1) variable

Our example illustrates an important special case of the general odds ratio formula for logistic regression that applies to (0, 1) variables. That is, an **adjusted odds ratio** can be obtained by exponentiating the coefficient of a (0, 1) variable in the model.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{ECG}$$

(adjusted)

In our example, that variable is CAT, and the other two variables, AGE and ECG, are the ones for which we adjusted.

X_i(0, 1): adj. ROR = $e^{\beta i}$

(controlling) for other X's

More generally, if the variable of interest is X_i, a (0, 1) variable, then e to the β_i, where β_i is the coefficient of X_i, gives an adjusted odds ratio involving the effect of X_i adjusted or controlling for the remaining X variables in the model.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{ECG}$$

(ECG: adjusted)

ECG (0, 1): adj. ROR = $e^{\beta 3}$

controlling for CAT and AGE

Suppose, for example, our focus had been on **ECG,** also a (0, 1) variable, instead of on CAT in a logistic model involving the same variables CAT, AGE, and ECG.

Then e to the β_3, where β_3 is the coefficient of ECG, would give the adjusted odds ratio for the effect of ECG, controlling for CAT and AGE.

SUMMARY

X_i is $(0, 1)$: ROR = e^{β_i}

General OR formula:

$$\text{ROR} = e^{\sum_{i=1}^{k} \beta_i (X_{1i} - X_{0i})}$$

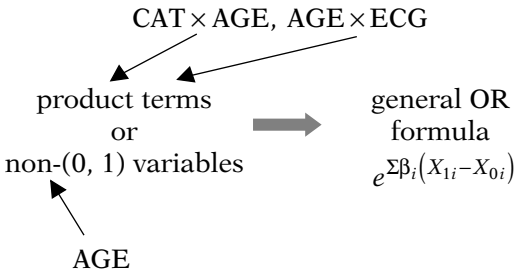
Thus, we can obtain an adjusted odds ratio for each (0, 1) variable in the logistic model by exponentiating the coefficient corresponding to that variable. This formula is much simpler than the general formula for ROR described earlier.

EXAMPLE

$$\text{logit } P(\mathbf{X}) = \alpha + \beta_1 \text{CAT} + \beta_2 \text{AGE} + \beta_3 \text{ECG}$$

(main effect variables)

Note, however, that the example we have considered involves only **main effect variables,** like CAT, AGE and ECG, and that the model does not contain product terms like CAT $\times$ AGE or AGE $\times$ ECG.



When the model contains product terms, like CAT × AGE, or variables that are not (0, 1), like the continuous variable AGE, the simple formula will not work if the focus is on any of these variables. In such instances, we must use the general formula instead.

Chapters

✓1. Introduction
2. Important Special Cases

This presentation is now complete. We suggest that you review the material covered here by reading the summary section. You may also want to do the practice exercises and the test which follows. Then continue to the next chapter entitled, "Important Special Cases of the Logistic Model."

Detailed Outline

I. The multivariable problem (pages 4–5)

A. Example of a multivariate problem in epidemiologic research, including the issue of controlling for certain variables in the assessment of an exposure–disease relationship.

B. The general multivariate problem: assessment of the relationship of several independent variables, denoted as *X*'s, to a dependent variable, denoted as *D*.

C. Flexibility in the types of independent variables allowed in most regression situations: A variety of variables is allowed.

D. Key restriction of model characteristics for the logistic model: The dependent variable is dichotomous.

II. Why is logistic regression popular? (pages 5–7)

A. Description of the logistic function.

B. Two key properties of the logistic function: Range is between 0 and 1 (good for describing probabilities) and the graph of function is S-shaped (good for describing combined risk factor effect on disease development).

III. The logistic model (pages 7–8)

A. Epidemiologic framework

B. Model formula: $P(D = 1 \mid X_1,\ldots, X_k) = P(\mathbf{X}) = 1/\{1 + \exp[-(\alpha + \Sigma\beta_i X_i)]\}$.

IV. Applying the logistic model formula (pages 9–11)

A. The situation: independent variables CAT (0, 1), AGE (constant), ECG (0, 1); dependent variable CHD(0, 1); fit logistic model to data on 609 people.

B. Results for fitted model: estimated model parameters are $\hat{\alpha}=-3.911$, $\hat{\beta}_1(\text{CAT})=0.65$, $\hat{\beta}_2(\text{AGE})=0.029$, and $\hat{\beta}_3(\text{ECG})=0.342$.

C. Predicted risk computations:
$\hat{P}(\mathbf{X})$ for CAT=1, AGE=40, ECG=0: 0.1090,
$\hat{P}(\mathbf{X})$ for CAT=0, AGE=40, ECG=0: 0.0600.

D. Estimated risk ratio calculation and interpretation: 0.1090/0.0600=1.82.

E. Risk ratio (RR) vs. odds ratio (OR): RR computation requires specifying all *X*'s; OR is more natural measure for logistic model.

V. Study design issues (pages 11–15)

A. Follow-up orientation.

B. Applicability to case-control and cross-sectional studies? Yes.

C. Limitation in case-control and cross-sectional studies: cannot estimate risks, but can estimate odds ratios.

D. The limitation in mathematical terms: for case-control and cross-sectional studies, cannot get a good estimate of the constant.

VI. Risk ratios versus odds ratios (pages 15–16)

A. Follow-up studies:

i. When all the variables in both groups compared are specified. [Example using CAT, AGE, and ECG comparing group 1 (CAT=1, AGE=40, ECG=0) with group 0 (CAT=0, AGE=40, ECG=0).]

ii. When control variables are unspecified, but assumed fixed and rare disease assumption is satisfied.

B. Case-control and cross-sectional studies: when rare disease assumption is satisfied.

C. What if rare disease assumption is not satisfied? May need to review characteristics of study to decide if the computed OR approximates an RR.

VII. Logit transformation (pages 16–22)

A. Definition of the logit transformation:
logit $P(\mathbf{X}) = \ln_e[P(\mathbf{X})/(1-P(\mathbf{X}))]$.

B. The formula for the logit function in terms of the parameters of the logistic model: logit $P(\mathbf{X}) = \alpha + \Sigma\beta_i X_i$.

C. Interpretation of the logit function in terms of odds:

i. $P(\mathbf{X})/[1-P(\mathbf{X})]$ is the odds of getting the disease for an individual or group of individuals identified by **X**.

ii. The logit function describes the "log odds" for a person or group specified by **X**.

D. Interpretation of logistic model parameters in terms of log odds:

i. α is the log odds for a person or group when all X's are zero—can be critiqued on grounds that there is no such person.

ii. A more appealing interpretation is that α gives the "background or baseline" log odds, where "baseline" refers to a model that ignores all possible X's.

iii. The coefficient β_i represents the change in the log odds that would result from a one unit change in the variable X_i when all the other X's are fixed.

iv. Example given for model involving CAT, AGE, and ECG: β_1 is the change in log odds corresponding to one unit change in CAT, when AGE and ECG are fixed.

VIII. Derivation of OR formula (pages 22–25)

A. Specifying two groups to be compared by an odds ratio: $\mathbf{X}_1$ and $\mathbf{X}_0$ denote the collection of X's for groups 1 and 0.

B. Example involving CAT, AGE, and ECG variables:
$\mathbf{X}_1$=(CAT=1, AGE=40, ECG=0), $\mathbf{X}_0$=(CAT=0, AGE=40, ECG=0).

C. Expressing the risk odds ratio (ROR) in terms of P(**X**):

$$\text{ROR} = \frac{(\text{odds for } \mathbf{X}_1)}{(\text{odds for } \mathbf{X}_0)} = \frac{\text{P}(\mathbf{X}_1)/1-\text{P}(\mathbf{X}_1)}{\text{P}(\mathbf{X}_0)/1-\text{P}(\mathbf{X}_0)}.$$

D. Substitution of the model form for P(**X**) in the above ROR formula to obtain general ROR formula:
ROR=exp[$\Sigma\beta_i(X_{1i} - X_{0i})$] = Π\{exp[$\beta_i(X_{1i} - X_{0i})$]\}

E. Interpretation from the product (Π) formula: The contribution of each X_i variable to the odds ratio is **multiplicative.**

IX. Example of OR computation (pages 25–26)

A. Example of ROR formula for CAT, AGE, and ECG example using X_1 and X_0 specified in VIII B above:
ROR=exp(β_1), where β_1 is the coefficient of CAT.

B. Interpretation of exp(β_1): an adjusted ROR for effect of CAT, controlling for AGE and ECG.

X. Special case for (0, 1) variables (pages 27–28)

A. General rule for (0, 1) variables: If variable is X_i, then ROR for effect of X_i controlling for other X's in model is given by the formula ROR=exp(β_i), where β_i is the coefficient of X_i.

B. Example of formula in A for ECG, controlling for CAT and AGE.

C. Limitation of formula in A: Model can contain only main effect variables for X's, and variable of focus must be (0, 1).

KEY FORMULAE

[exp(a)=e^a for any number a]

LOGISTIC FUNCTION: $f(z)$=1/[1+exp($-z$)]

LOGISTIC MODEL: P(**X**)=1/\{1+exp[$-(\alpha+\Sigma\beta_i X_i)$]\}

LOGIT TRANSFORMATION: logit P(**X**)= $\alpha+\Sigma\beta_i X_i$

RISK ODDS RATIO (general formula):
$\text{ROR}_{\mathbf{X}_1, \mathbf{X}_0}$: =exp[$\Sigma\beta_i(X_{1i} - X_{0i})$]=$\Pi$\{exp[$\beta_i(X_{1i} - X_{0i})$]\}

RISK ODDS RATIO [(0, 1) variables]: ROR=exp(β_i) for the effect of the variable X_i *adjusted* for the other X's

Practice Exercises

Suppose you are interested in describing whether social status, as measured by a (0, 1) variable called SOC, is associated with cardiovascular disease mortality, as defined by a (0, 1) variable called CVD. Suppose further that you have carried out a 12-year follow-up study of 200 men who are 60 years old or older. In assessing the relationship between SOC and CVD, you decide that you want to control for smoking status [SMK, a (0, 1) variable] and systolic blood pressure (SBP, a continuous variable).

In analyzing your data, you decide to fit two logistic models, each involving the dependent variable CVD, but with different sets of independent variables. The variables involved in each model and their estimated coefficients are listed below:

Model 1

VARIABLE	COEFFICIENT
CONSTANT	−1.1800
SOC	−0.5200
SBP	0.0400
SMK	−0.5600
SOC × SBP	−0.0330
SOC × SMK	0.1750

Model 2

VARIABLE	COEFFICIENT
CONSTANT	−1.1900
SOC	−0.5000
SBP	0.0100
SMK	−0.4200

1. For each of the models fitted above, state the form of the logistic model that was used (i.e., state the model in terms of the unknown population parameters and the independent variables being considered).

   Model 1:

   Model 2:

2. For each of the above models, state the form of the estimated model in logit terms.

   Model 1: logit P($\mathbf{X}$)=

   Model 2: logit P($\mathbf{X}$)=

3. Using Model 1, compute the estimated risk for CVD death (i.e., CVD=1) for a high social class (SOC=1) smoker (SMK=1) with SBP=150. (You will need a calculator to answer this. If you don't have one, just state the computational formula that is required, with appropriate variable values plugged in.)

4. Using Model 2, compute the estimated risk for CVD death for the following two persons:

   Person 1: SOC=1, SMK=1, SBP=150.
   Person 2: SOC=0, SMK=1, SBP=150.
   (As with the previous question, if you don't have a calculator, you may just state the computations that are required.)

   Person 1:

   Person 2:

5. Compare the estimated risk obtained in Exercise 3 with that for person 1 in Exercise 4. Why aren't the two risks exactly the same?

6. Using Model 2 results, compute the risk ratio that compares person 1 with person 2. Interpret your answer.

7. If the study design had been either case-control or cross-sectional, could you have legitimately computed risk estimates as you did in the previous exercises? Explain.

8. If the study design had been case-control, what kind of measure of association could you have legitimately computed from the above models?

9. For Model 2, compute and interpret the estimated odds ratio for the effect of SOC, controlling for SMK and SBP? (Again, if you do not have a calculator, just state the computations that are required.)

10. Which of the following general formulae is *not* appropriate for computing the effect of SOC controlling for SMK and SBP in *Model 1*? (Circle one choice.) Explain your answer.

    a. $\exp(\beta_S)$, where β_S is the coefficient of SOC in model 1.
    b. $\exp[\Sigma\beta_i(X_{1i} - X_{0i})]$.
    c. $\Pi\{\exp[\beta_i(X_{1i} - X_{0i})]\}$.

Test

True or False (Circle T or F)

T F 1. We can use the logistic model provided all the independent variables in the model are continuous.

T F 2. Suppose the dependent variable for a certain multivariable analysis is systolic blood pressure, treated continuously. Then, a logistic model should be used to carry out the analysis.

T F 3. One reason for the popularity of the logistic model is that the range of the logistic function, from which the model is derived, lies between 0 and 1.

T F 4. Another reason for the popularity of the logistic model is that the shape of the logistic function is linear.

T F 5. The logistic model describes the probability of disease development, i.e., risk for the disease, for a given set of independent variables.

T F 6. The study design framework within which the logistic model is defined is a follow-up study.

T F 7. Given a fitted logistic model from case-control data, we can estimate the disease risk for a specific individual.

T F 8. In follow-up studies, we can use a fitted logistic model to estimate a risk ratio comparing two groups provided all the independent variables in the model are specified for both groups.

T F 9. Given a fitted logistic model from a follow-up study, it is not possible to estimate individual risk as the constant term cannot be estimated.

T F 10. Given a fitted logistic model from a case-control study, an odds ratio can be estimated.

T F 11. Given a fitted logistic model from a case-control study, we can estimate a risk ratio if the rare disease assumption is appropriate.

T F 12. The logit transformation for the logistic model gives the log odds ratio for the comparison of two groups.

T F 13. The constant term, α, in the logistic model can be interpreted as a baseline log odds for getting the disease.

T F 14. The coefficient β_i in the logistic model can be interpreted as the change in log odds corresponding to a one unit change in the variable X_i that ignores the contribution of other variables.

T F 15. We can compute an odds ratio for a fitted logistic model by identifying two groups to be compared in terms of the independent variables in the fitted model.

T F 16. The product formula for the odds ratio tells us that the joint contribution of different independent variables to the odds ratio is additive.

T F 17. Given a (0, 1) independent variable and a model containing only main effect terms, the odds ratio that describes the effect of that variable controlling for the others in the model is given by e to the α, where α is the constant parameter in the model.

T F 18. Given independent variables AGE, SMK [smoking status (0, 1)], and RACE (0, 1), in a logistic model, an adjusted odds ratio for the effect of SMK is given by the natural log of the coefficient for the SMK variable.

T F 19. Given independent variables AGE, SMK, and RACE, as before, plus the product terms SMK $\times$ RACE and SMK $\times$ AGE, an adjusted odds ratio for the effect of SMK is obtained by exponentiating the coefficient of the SMK variable.

T F 20. Given the independent variables AGE, SMK, and RACE as in Question 18, but with SMK coded as (1, -1) instead of (0, 1), then e to the coefficient of the SMK variable gives the adjusted odds ratio for the effect of SMK.

21. Which of the following is *not* a property of the logistic model? (Circle one choice.)

   a. The model form can be written as $P(\mathbf{X})=1/\{1+\exp[-(\alpha+\Sigma\beta_i X_i)]\}$, where "exp{·}" denotes the quantity e raised to the power of the expression inside the brackets.

   b. logit $P(\mathbf{X})=\alpha+\Sigma\beta_i X_i$ is an alternative way to state the model.

   c. $ROR=\exp[\Sigma\beta_i(X_{1i}-X_{0i})]$ is a general expression for the odds ratio that compares two groups of **X** variables.

   d. $ROR=\Pi\{\exp[\beta_i(X_{1i}-X_{0i})]\}$ is a general expression for the odds ratio that compares two groups of **X** variables.

   e. For any variable X_i, $ROR=\exp[\beta_i]$, where β_i is the coefficient of X_i, gives an adjusted odds ratio for the effect of X_i.

Suppose a logistic model involving the variables D=HPT[hypertension status (0, 1)], X_1=AGE(continuous), X_2=SMK(0, 1), X_3=SEX(0, 1), X_4=CHOL (cholesterol level, continuous), and X_5=OCC[occupation (0, 1)] is fit to a set of data. Suppose further that the estimated coefficients of each of the variables in the model are given by the following table:

VARIABLE	COEFFICIENT
CONSTANT	−4.3200
AGE	0.0274
SMK	0.5859
SEX	1.1523
CHOL	0.0087
OCC	−0.5309

22. State the form of the logistic model that was fit to these data (i.e., state the model in terms of the unknown population parameters and the independent variables being considered).
23. State the form of the *estimated* logistic model obtained from fitting the model to the data set.
24. State the estimated logistic model in logit form.
25. Assuming the study design used was a follow-up design, compute the estimated risk for a 40-year-old male (SEX=1) smoker (SMK=1) with CHOL=200 and OCC=1. (You need a calculator to answer this question.)
26. Again assuming a follow-up study, compute the estimated risk for a 40-year-old male nonsmoker with CHOL=200 and OCC=1. (You need a calculator to answer this question.)
27. Compute and interpret the estimated risk ratio that compares the risk of a 40-year-old male smoker to a 40-year-old male nonsmoker, both of whom have CHOL=200 and OCC=1.
28. Would the risk ratio computation of Question 27 have been appropriate if the study design had been either cross-sectional or case-control? Explain.
29. Compute and interpret the estimated odds ratio for the effect of SMK controlling for AGE, SEX, CHOL, and OCC. (If you do not have a calculator, just state the computational formula required.)
30. What assumption will allow you to conclude that the estimate obtained in Question 29 is approximately a risk ratio estimate?
31. If you could not conclude that the odds ratio computed in Question 29 is approximately a risk ratio, what measure of association is appropriate? Explain briefly.
32. Compute and interpret the estimated odds ratio for the effect of OCC controlling for AGE, SMK, SEX, and CHOL. (If you do not have a calculator, just state the computational formula required.)
33. State two characteristics of the variables being considered in this example that allow you to use the $\exp(\beta_i)$ formula for estimating the effect of OCC controlling for AGE, SMK, SEX, and CHOL.
34. Why can you not use the formula $\exp(\beta_i)$ formula to obtain an adjusted odds ratio for the effect of AGE, controlling for the other four variables?

Answers to Practice Exercises

1. *Model 1:* $\hat{P}(\mathbf{X})=1/(1+\exp\{-[-1.18-0.52(SOC)+0.04(SBP)-0.56(SMK) -0.033(SOC\times SBP)+0.175(SOC\times SMK)]\})$.

   Model 2: $\hat{P}(\mathbf{X})=1/(1+\exp\{-[-1.19-0.50(SOC)+0.01(SBP)+0.42(SMK)]\})$.

2. *Model 1:* $\text{logit}\ \hat{P}(\mathbf{X}) =-1.18-0.52(SOC)+0.04(SBP)-0.56(SMK) -0.033(SOC\times SBP)+0.175(SOC\times SMK)$.

   Model 2: $\text{logit}\ \hat{P}(\mathbf{X})=-1.19-0.50(SOC)+0.01(SBP)-0.42(SMK)$.

3. For SOC=1, SBP=150, and SMK=1, $\mathbf{X}=(SOC, SBP, SMK, SOC\times SBP, SOC\times SMK)=(1, 150, 1, 150, 1)$ and

   Model 1 $\hat{P}(\mathbf{X})=1/(1+\exp\{-[-1.18-0.52(1)+0.04(150)-0.56(1) -0.033(1\times 150)-0.175(1\times 1)]\})$.
   $=1/\{1+\exp[-(-1.035)]\}$
   $=1/(1+2.815)$
   $=0.262$

4. For *Model 2, person 1* (SOC=1, SMK=1, SBP=150):

   $\hat{P}(\mathbf{X})=1/(1+\exp\{-[-1.19 - 0.50(1) + 0.01 (150) - 0.42(1)]\})$
   $=1/\{1 + \exp [- (-0.61)]\}$
   $=1/(1 + 1.84)$
   $=0.352$

   For *Model 2, person 2* (SOC=0, SMK=1, SBP=150):

   $\hat{P}(\mathbf{X})=1/(1 + \exp\{-[-1.19 - 0.50(0) + 0.01(150) - 0.42(1)]\})$
   $=1/\{1 + \exp[-(-0.11)]\}$
   $=1/(1 + 1.116)$
   $=0.473$

5. The risk computed for *Model 1* is 0.262, whereas the risk computed for *Model 2, person 1* is 0.352. Note that both risks are computed for the same person (i.e., SOC=1, SMK=150, SBP=150), yet they yield different values because the models are different. In particular, *Model 1* contains two product terms that are not contained in *Model 2,* and consequently, computed risks for a given person can be expected to be somewhat different for different models.

6. Using *model 2* results,

$$\text{RR(1 vs. 2)} = \frac{\text{P(SOC=0, SMK=1, SBP=150)}}{\text{P(SOC=1, SMK=1, SBP=150)}}$$

$= 0.352/0.473 = 1/1.34 = 0.744$

This estimated risk ratio is less than 1 because the risk for high social class persons (SOC=1) is less than the risk for low social class persons (SOC=0) in this data set. More specifically, the risk for low social class persons is 1.34 times as large as the risk for high social class persons.

7. No. If the study design had been either case-control or cross-sectional, risk estimates could not be computed because the constant term (α) in the model could not be estimated. In other words, even if the computer printed out values of -1.18 or -1.19 for the constant terms, these numbers would not be legitimate estimates of α.

8. For case-control studies, only odds ratios, not risks or risk ratios, can be computed directly from the fitted model.

9. $\widehat{\text{OR}}$(SOC=1 vs. SOC=0 controlling for SMK and SBP)

$=e^{\hat{\beta}}$, where $\hat{\beta}=-0.50$ is the estimated coefficient of SOC in the fitted model

$=\exp(-0.50)$
$=0.6065 = 1/1.65.$

The estimated odds ratio is less than 1, indicating that, for this data set, the risk of CVD death for high social class persons is less than the risk for low social class persons. In particular, the risk for low social class persons is estimated as 1.65 times as large as the risk for high social class persons.

10. Choice (a) is *not* appropriate for the effect of SOC using model 1. Model 1 contains interaction terms, whereas choice (a) is appropriate only if all the variables in the model are main effect terms. Choices (b) and (c) are two equivalent ways of stating the general formula for calculating the odds ratio for any kind of logistic model, regardless of the types of variables in the model.