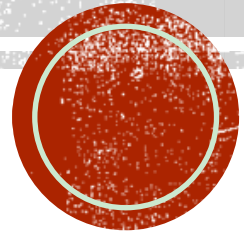
Simple Logistic regression



Motivations

In health sciences, researchers often need to investigate the *association* between *categorical (dichotomous) dependent variables* with *categorical* and/or *quantitative independent variables*.

Examples: Typical dichotomous dependent variables include disease categories (affected, unaffected), death status (dead, alive), and remission status (in remission, not in remission).

Simple and multiple linear regressions do NOT apply

For such scenarios, traditional linear regression techniques are not appropriate.

Some *fundamental assumptions are severely violated*. In particular, the *normality* assumption (on residual and thus dependent variable) does not hold any more.

Contingency analysis is insufficient

Analysis for $R \times C$ contingency could deal with some *simple situations*, where both dependent and independent variables are categorical variables.

Dichotomization of quantitative independent variables may result in loss of information and thus reduce statistical power.

Another important issue is whether a disease–exposure relationship is influenced by **confounders (covariates)**. Contingency table analysis would be able to do nothing about **adjustment of confounders**.

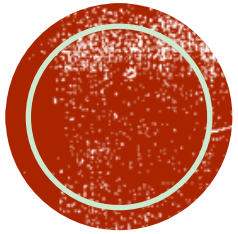
Outline

To overcome previous limitations, logistic regression is usually performed. In Part IV, we will

(1) fit logistic regression models with both category or/and continuous predictors;

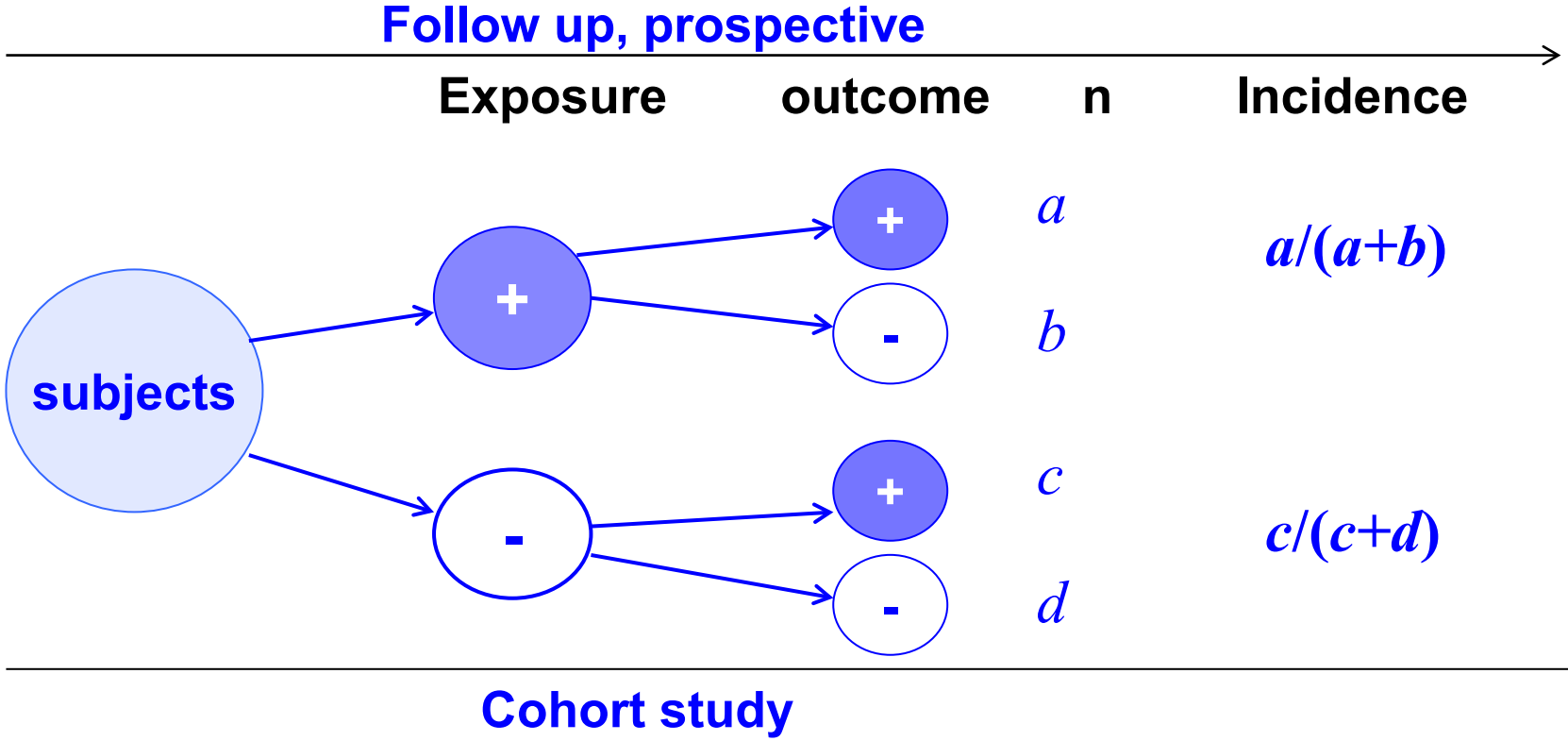
(2) test for significance of the logistic regression equation and;

(3) estimate ***odds ratio*** as the *most important* ***comparison measure for disease risk***.



RR/OR

COHORT STUDY



Cohort study

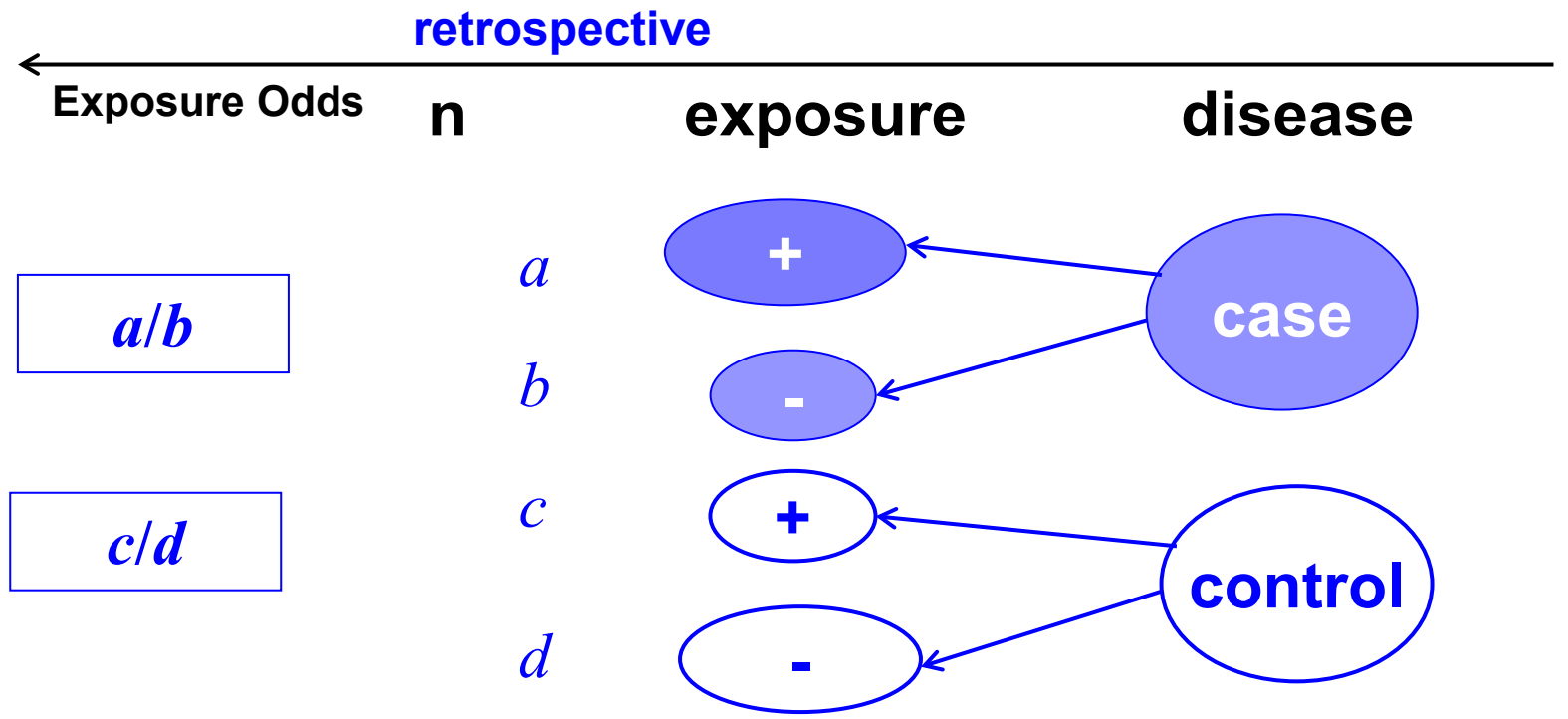
$$RR = \frac{I_e}{I_0} = \frac{a/n_1}{c/n_0}$$

$$I_e = \frac{a}{a+b} = a/n_1$$

$$I_0 = \frac{c}{c+d} = c/n_2$$

	case	control	total	incidence
Exposure	a	b	n_1=a+b	a/ n_1
Non-exposure	c	d	n_0=c+d	c/ n_0

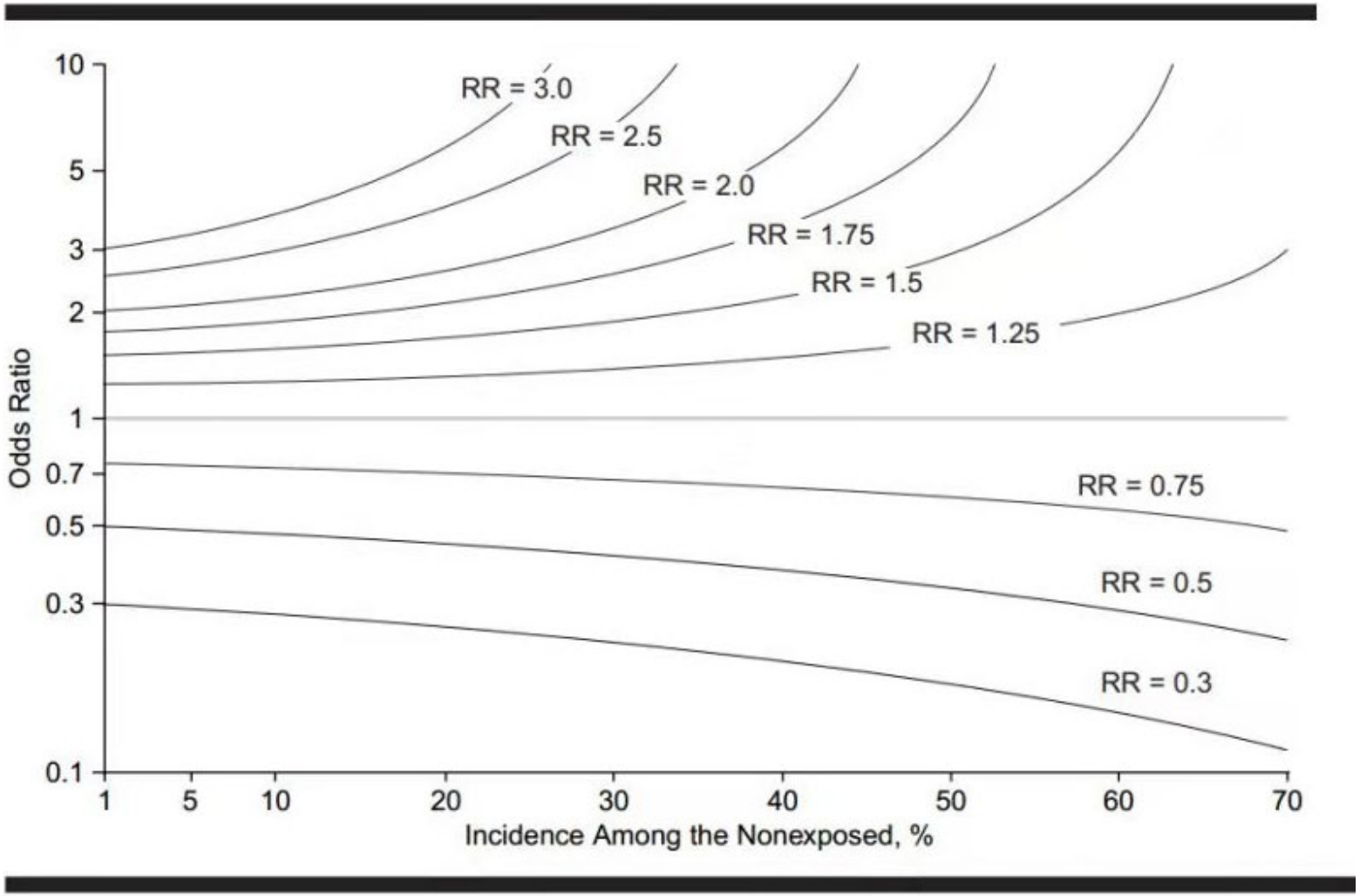
CASE CONTROL STUDY



$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

Case-control study

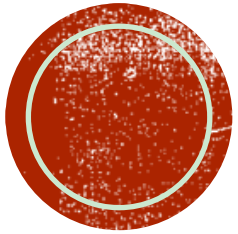
	case	control	total
Exposure	a	b	n_1=a+b
Non-exposure	c	d	n_0=c+d



The relationship between risk ratio (RR) and odds ratio by incidence of the outcome.

When incidence among nonexposed is <5%,

OR≈RR



SIMPLE LOGISTIC REGRESSION

Simple Logistic regression Model

Let Y be a *dichotomous* dependent variable, represented as

$$Y = \begin{cases} 1 & \text{for a } \textit{success} \text{ (e.g., affected subject),} \\ 0 & \text{for a failure (e.g., unaffected subject).} \end{cases}$$

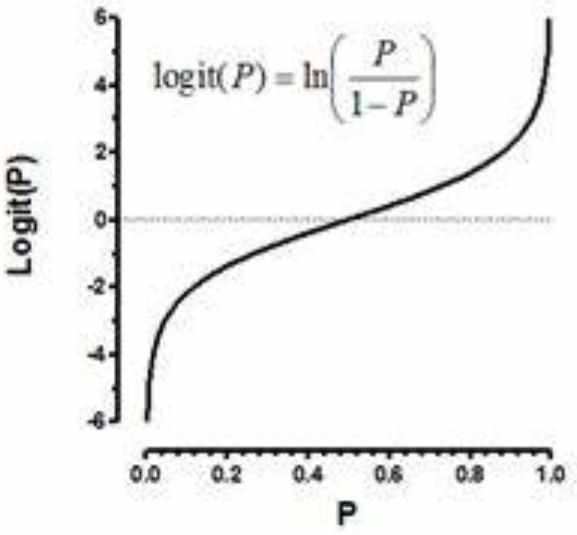
Let X be an independent variable (either quantitative or categorical).

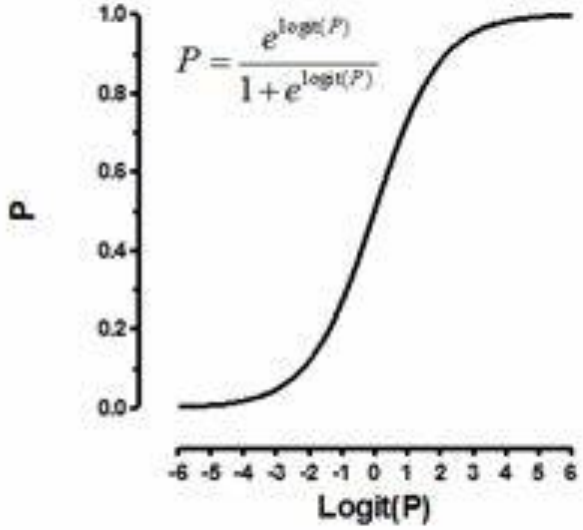
$P(Y=1) \in [0, 1]$

Definition 1 (logit transformation): For a *success probability* $p \in (0,1)$, the **logit transformation** is defined as

$$\text{logit}(p)=\ln\left(\frac{p}{1-p}\right)$$

where ln(.) is the ***natural logarithm***.





The ratio $\frac{p}{1-p}$ is the ***odds in favor of success***, and logit(p) is also called the log odds

$-\infty < \text{logit}(p) < +\infty$ for $0 < p < 1$.

DEFINITION 2 (SLRM-SIMPLE LOGISTIC REGRESSION MODEL):

For a specific value x of X, let

$$p(x) = \Pr(Y = 1|X = x),$$

then

$$1 - p(x) = \Pr(Y = 0|X = x).$$

The SLRM is defined by

$$\text{logit}\big(p(x)\big) = \beta_0 + \beta_1 x.$$

Equivalently,

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

Note that $0 \leq p(x) \leq 1$ for any specific value x of X. The independent variable X can be either quantitative or categorical.

In the SLRM, β_1 represents **the average change in the log odds** for **every one-unit change in x**.

Inverting with an exponential function, we see the odds in favor of success represented as a function of x:

$$\text{ODDS}(x) = \frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

Often, OR=we **compare the odds in favor of success** ($Y = 1$) at two distinct values of an independent variable X.

Definition 3. (OR – Odds Ratio): Odds ratio (*OR*) between the odds at *two fixed values* x_i, x_j of X is defined as

$$\mathrm{OR} = \frac{\mathrm{ODDS}(x_i)}{\mathrm{ODDS}(x_j)}.$$

OR = 1: the probability of success is the same for individuals at $X = x_i$ and $X = x_j$.

OR > 1: the probability of success is greater for those with $X = x_i$ than those with $X = x_j$.

OR < 1: the probability of success is greater for those with $\mathrm{X} = x_j$ than those with $X = x_i$.

Under the SLRM, we have

$$OR = \frac{ODDS(x_i)}{ODDS(x_j)} = \frac{e^{\beta_0+\beta_1 x_i}}{e^{\beta_0+\beta_1 x_j}} = e^{\beta_1(x_i-x_j)}.$$

The ***Maximum Likelihood Estimation Method*** is needed to fit the logistic regression model, i.e., compute the maximum likelihood estimates (MLEs) of the regression coefficients.

In general, this MLE method requires an *intensive iterative process for optimization*. We will rely on statistical packages for this purpose. ***Using the MLEs, we can estimate odds and odds ratio accordingly***.

2. Connection with Contingency-Table Analysis

In particular, if X is a categorical variable at **two categories** with dummy variable coding:

X = 1 for category 1;

X = 0 for the other category.

Then it follows that

$$\text{OR} = \frac{\boldsymbol{ODDS(X=1)}}{\boldsymbol{ODDS(X=0)}} = \frac{e^{\beta_0+\beta_1\times 1}}{e^{\beta_0+\beta_1\times 0}} = e^{\beta_1}.$$

	case	control
Exposure (x=1)	a	b
Non-exposure (x=0)	c	d

$$\text{ODDS}(x=1) = \frac{p(x=1)}{1-p(x=1)} = \frac{a/(a+b)}{1-a/(a+b)} = a/b$$

$$\text{ODDS}(x=0) = \frac{p(x=0)}{1-p(x=0)} = \frac{c/(c+d)}{1-c/(c+d)} = c/d$$

$$OR = \frac{odds(x=1)}{odds(x=0)} = \frac{ad}{bc}$$

1. We can ***estimate*** the *OR* relating Y to X in either of two equivalent ways:

a. We can ***estimate*** the OR directly from the 2 × 2 table: **(ad)/(bc).**

b. We can set up a logistic-regression model of the form

$$\log[p/(1-p)] = \beta_0 + \beta_1 X,$$

where p = probability of Y=1 given X=1 and where we estimate the OR by $\exp(\widehat{\beta}_1)$.

2. For simple random samples (prospective or cross-sectional studies), we can estimate the Pr(Y=1|X=1) and Pr(Y=1|X=0) in either of two equivalent ways:

a. From the 2 × 2 table, we have

$$\Pr(Y = 1|X = 1) = \frac{a}{a + c},$$
$$\Pr(Y = 1|X = 0) = \frac{b}{b + d}.$$

a. From the logistic-regression model,

$$\Pr(Y = 1|X = 1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1}}, \Pr(Y = 1|X = 0) = \frac{e^{\hat{\beta}_0}}{1 + e^{\hat{\beta}_0}}.$$

EXAMPLE 1

Hosmer and Lemeshow (1989) present an example regressing the **presence** or **absence** of Coronary Heart Disease (CHD) on ***Age***, for 100 subjects. CHD was coded as 1 for present and 0 for absent (**Table 1**).

The 0's and 1's obviously group into two parallel lines (**Figure 1**), demonstrating the dichotomous nature of CHD. Any linear trend?

Table 1. Age and CHD status of 100 subjects

Table 19.1: Age and Coronary Heart Disease (CHD) Status of 100 Subjects.

ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD
1	1	20	0	35	3	38	0	68	6	51	0
2	1	23	0	36	3	39	0	69	6	52	0
3	1	24	0	37	3	39	1	70	6	52	1
4	1	25	0	38	4	40	0	71	6	53	1
5	1	25	1	39	4	40	1	72	6	53	1
6	1	26	0	40	4	41	0	73	6	54	1
7	1	26	0	41	4	41	0	74	7	55	0
8	1	28	0	42	4	42	0	75	7	55	1
9	1	28	0	43	4	42	0	76	7	55	1
10	1	29	0	44	4	42	0	77	7	56	1
11	2	30	0	45	4	42	1	78	7	56	1
12	2	30	0	46	4	43	0	79	7	56	1
13	2	30	0	47	4	43	0	80	7	57	0
14	2	30	0	48	4	43	1	81	7	57	0
15	2	30	0	49	4	44	0	82	7	57	1
16	2	30	1	50	4	44	0	83	7	57	1
17	2	32	0	51	4	44	1	84	7	57	1
18	2	32	0	52	4	44	1	85	7	57	1
19	2	33	0	53	5	45	0	86	7	58	0
20	2	33	0	54	5	45	1	87	7	58	1

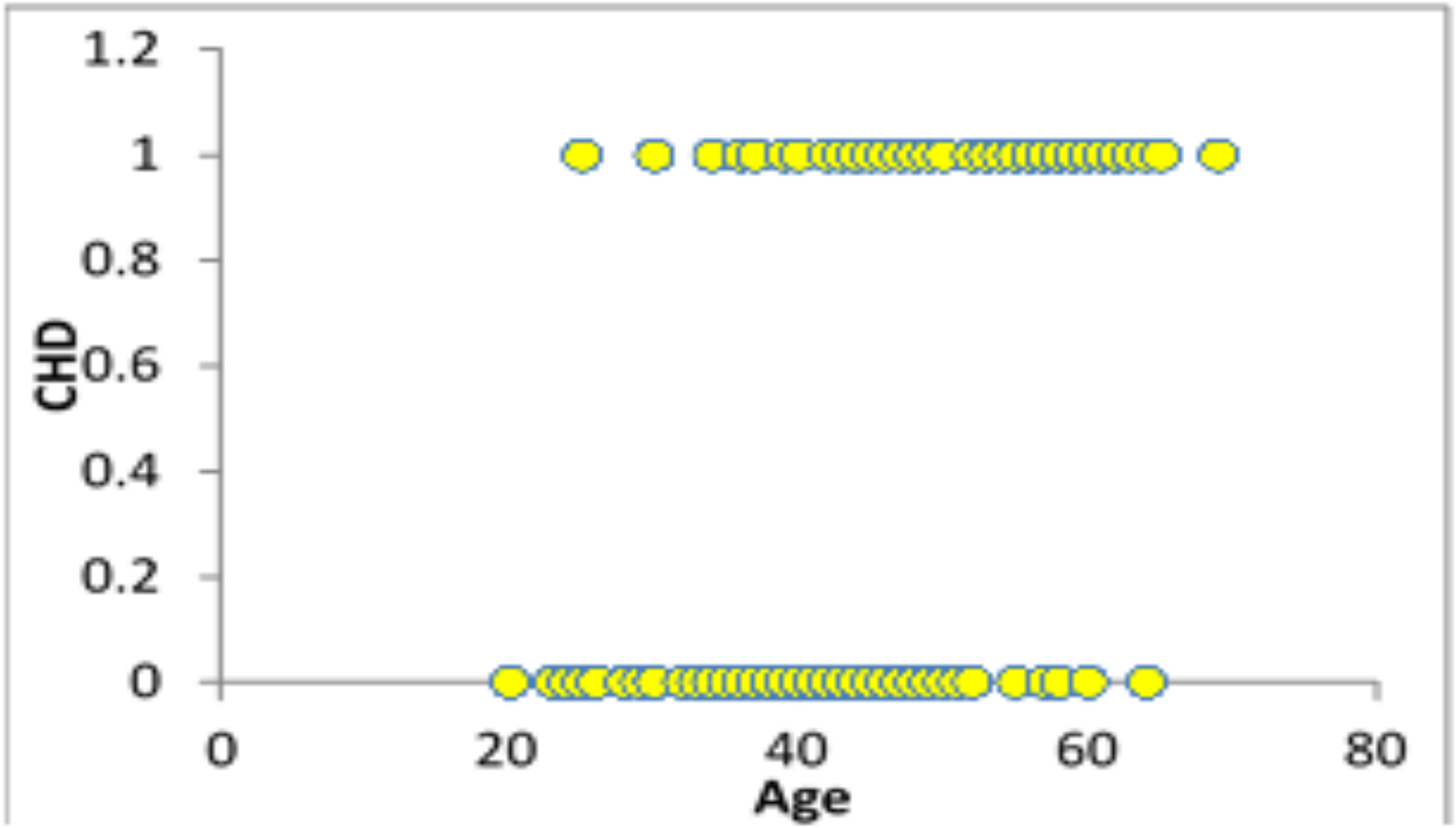


Figure 1. Age and CHD status of 100 subjects

EXAMPLE 1

To get *a pictorial representation* of the *probability* of CHD being present as AGE increases, the authors created *age groups* (AGRP) and the *proportion of subjects with CHD* in each age group (**Table 2**), where a clear increasing likelihood of CHD emerges as AGE increases.

Table 2. Frequency Table of Age group by CHD status of 100 subjects

Age Group	n	CHD Absent	CHD Present	Mean (Proportion)
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46

Note that the trend also represents *an S-shaped curve* (**Figure 2**, where the proportion with CHD are plotted against the mid-point of each age interval).

Figure 2 demonstrates the *fitted curve* and the *observed proportion* with CHD. The logistic model provides a pretty good fit to the data. **(Source:** Hosmer, David W., Lemeshow, Stanley, *Applied Logistic Regression*, John Wiley & Sons, Inc., 1989).

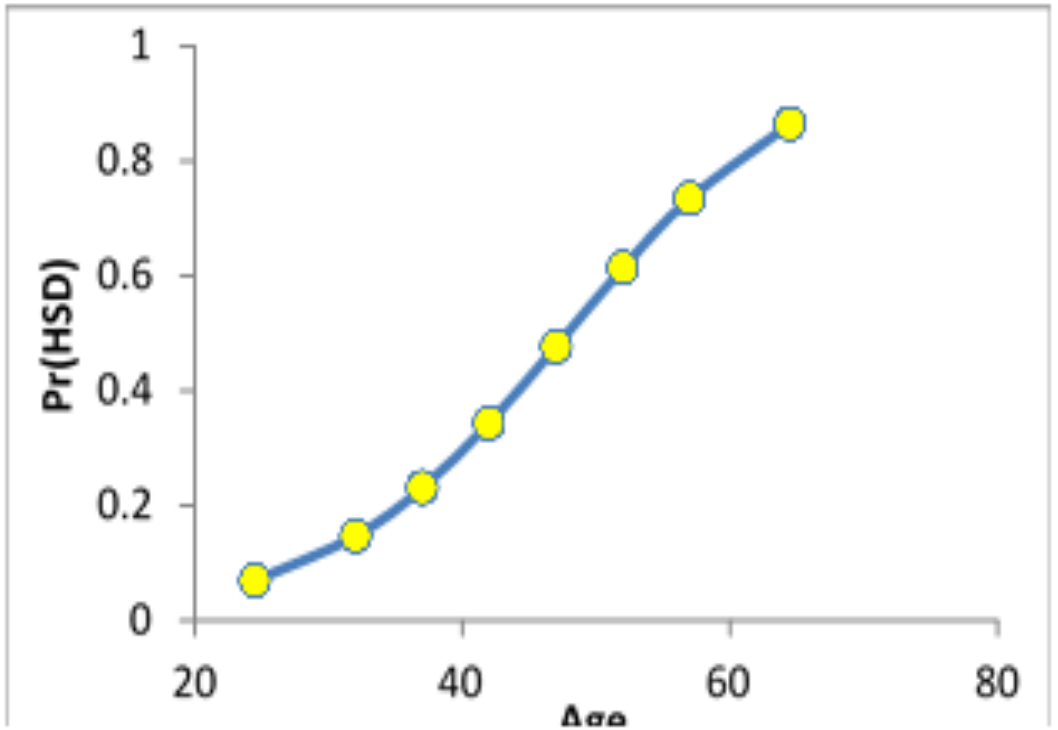


Figure 2. predicted probability of CHD based on logistic regression

The logistic regression model has this S-shape and would be an appropriate choice to model this trend.

Hosmer and Lemeshow present the MLEs $\hat{\beta}_0 = -5.31$ and $\hat{\beta}_1 = 0.111$, and present the *logistic model fit* to this data as

$$\hat{p}(x) = \frac{e^{-5.310+0.111x}}{1+e^{-5.31+0.111x}}$$

Example 2 (APID - Acute Pelvic inflammatory Disease): Daniel presents an example of a logistic regression application from a study of smoking and APID, the data (Table 3) were reported by Scholes et al (1992). Calculate the odds in favor of disease for a smoker versus a nonsmoker.

Table 3. Smoking and APID status of 425 subjects

Ever Smoked?	*Cases*	*Controls*	*Total*
Yes	77	123	200
No	54	171	225
Total	131	294	425

Sources: 1. Daniel, Wayne W. (1999) *Biostatistics: A*

Solution: Let X represents smoking categories (coded as 1 for an ever smoker and 0 for a never smoker), and Y represents status of APID (coded as Y=1 for a case and Y = 0 for a control). Under the simple logistic regression model, **the odds ratio can be calculated as** $\boldsymbol{OR = e^{\beta_1}}$.

Using PROC LOGIST in SAS, the fitted logistic regression model is

$$\hat{p}(x) = \frac{e^{-1.1527+0.6843x}}{1 + e^{-1.1527+0.6843x}}$$

The **estimated odds ratio** is $e^{0.6843}$ =**1.9824.** Smokers have almost ***two times higher risk*** of developing APID than *do* nonsmokers.

Alternatively, the odds ratio can be estimated per **Contingency Table Analysis** as below.

$$\text{ODDS(smoker)} = \frac{\Pr(\text{Affected}|\text{Smoker})}{\Pr(\text{Unaffected}|\text{Smoker})} = \frac{77/200}{123/200} = \frac{77}{123}.$$

$$\text{ODDS(nosmoker)} = \frac{\Pr(\text{Affected}|\text{Nonsmoker})}{\Pr(\text{Unaffected}|\text{Nonsmoker})} = \frac{54/225}{171/225} = \frac{54}{171}.$$

$$\text{OR} = \frac{77/123}{54/171} = \frac{77 \times 171}{54 \times 123} = 1.9824$$

PRACTICE PROBLEMS

1. In SLRM, let both dependent variable Y and independent variable X are categorical variables with two categories. Is there any way to *estimate the regression coefficients without iteration*?

2. **(Cardiovascular Disease)** A study looked at the effects of *oral contraceptives (OC) use* on heart disease in women 40 to 44 years of age. The researchers found that among 5000 current OC users at baseline, 13 women developed a ***myocardial infarction (MI)*** over a 3-year period, whereas among 10,000 non-OC users, 7 developed an MI over a 3-year period (**Table 4**). Estimate the *OR* in favor of MI for an OC user compared with a non-OC user.

Table 4: 2×2 contingency table for the OC–MI data

	MI status over 3 years		
OC-use group	Yes	No	Total
OC users	13	4987	5000
Non-OC users	7	9993	10000
Total	20	14980	15000

Table 4: 2×2 contingency table for the OC–MI data

	MI status over 3 years		
OC-use group	Yes	No	Total
OC users	13	4987	5000
Non-OC users	7	9993	10000
Total	20	14980	15000