Simple Logistic Regression

Dr. Wan Nor Arifin

Department of Biostatistics and Research Methodology
Universiti Sains Malaysia
wnarifin@usm.my / wnarifin.github.io



Last update: Nov 10, 2026

Outlines

- Introduction
- Logistic Regression Model
- Relative Risk vs Odds Ratio
- Logit Transformation
- Maximum Likelihood Estimation
- Testing Significance
- Hands-on

Learning Outcomes

- Understand the move from linear to logistic regression for a dichotomous outcome
- Identify the logistic regression model
- Understand the concept of odds ratio
- Identify the logit function

Learning Outcomes

- Calculate and interpret odds ratio from the model
- Understand maximum likelihood estimation
- Test the significance of the coefficients
- Perform simple logistic regression in SPSS and R
  - Dichotomous and numerical independent variables

Introduction

Introduction

- Outcome is categorical with two levels (dichotomous)
  - yes/no, disease/no disease
  - coded as 1/0
- In place of a numerical outcome, y in linear regression
- Simple: one independent variable, x
  - numerical or categorical

Linear vs Logistic

Recall linear regression: expected value of *Y*, or conditional mean of *Y* given *x*,

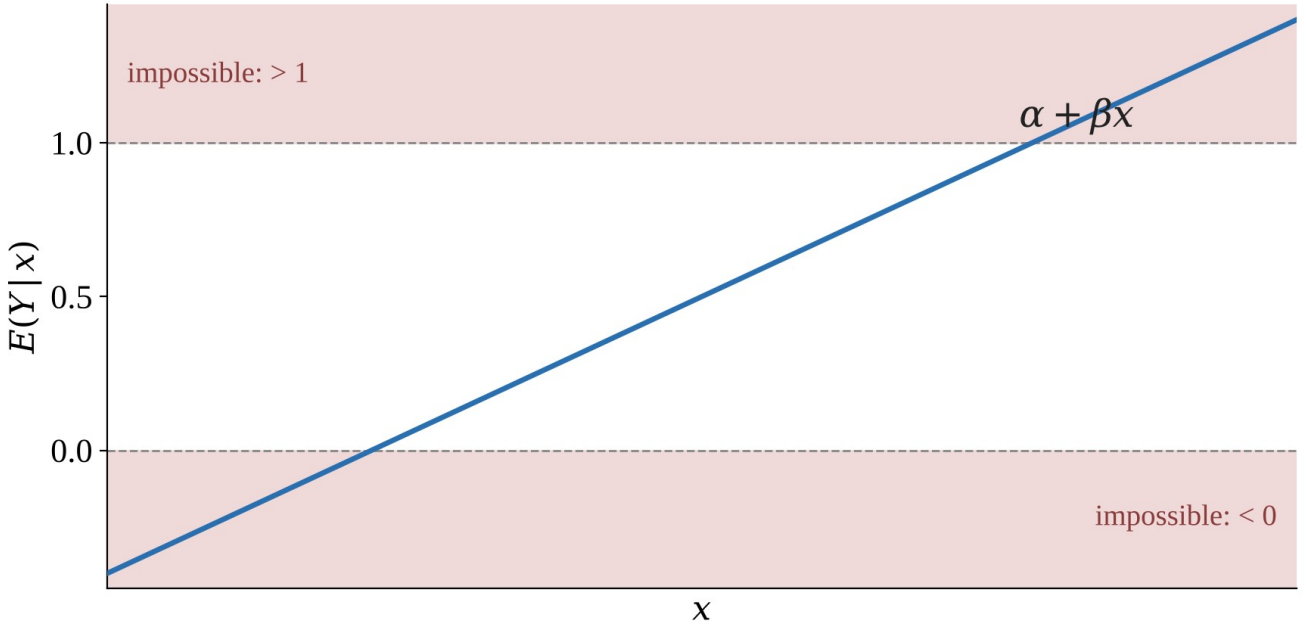
$$E(Y|x) = \alpha + \beta x$$

For a dichotomous outcome, the conditional mean is a probability,

$$0 \leq E(Y|x) \leq 1$$

Linear vs Logistic

But as x ranges from $-\infty$ to $+\infty$, so does $\alpha + \beta x$



Impossible to keep both sides equal → need a different model: **logistic regression model**

Logistic Regression Model

Logistic Regression Model

Linear predictor,

$$z = \alpha + \beta x$$

Expected value of Y given x is the probability of the outcome,

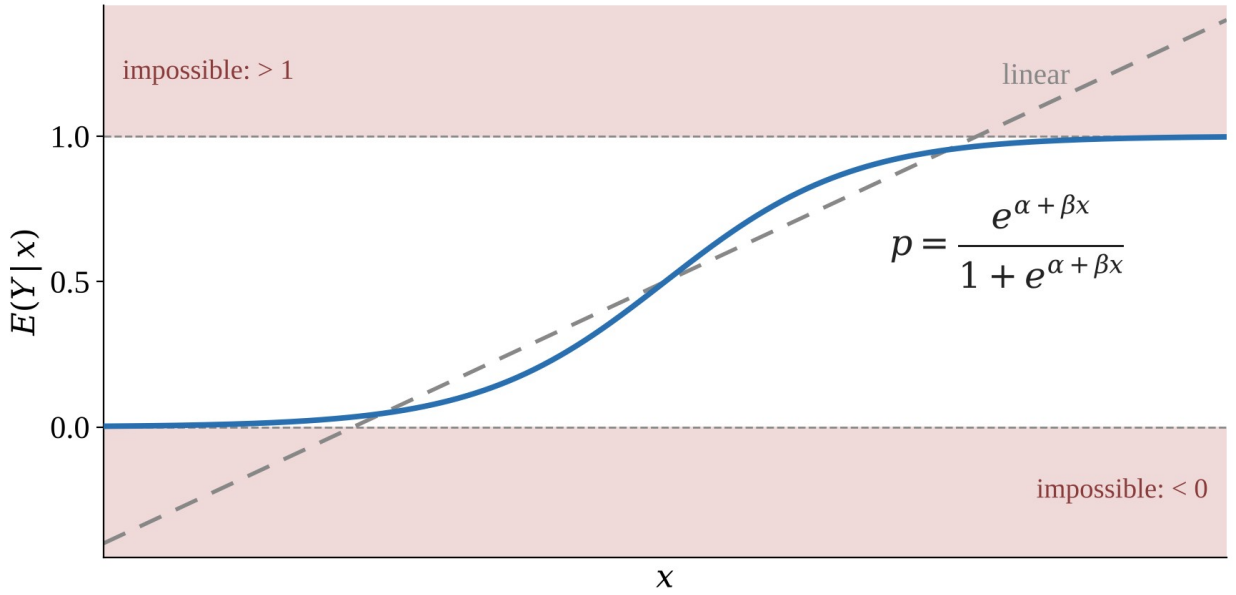
$$E(Y|x) = P(Y = 1|x) = p$$

$$p = \frac{e^z}{1+e^z} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Logistic Regression Model

Now, the probability of having a disease, p is kept between 0 and 1,

$$0 \leq p \leq 1$$



Relative Risk vs Odds Ratio

Cohort Study

Smoking (exposure) and lung cancer (outcome),

Table 1: Lung cancer vs smoking

	Lung cancer	No lung cancer	Total
Smoking	20	12	32
No smoking	95	73	168
Total	115	85	200

Relative risk of developing lung cancer: smoker vs non-smoker

Relative Risk

Risk of lung cancer,

$$\text{Risk}_{\text{smoking}} = \frac{20}{20+12} = \frac{20}{32} = 0.625$$

$$\text{Risk}_{\text{no smoking}} = \frac{95}{95+73} = \frac{95}{168} = 0.565$$

Relative risk,

$$\text{RR} = \frac{\text{Risk}_{\text{smoking}}}{\text{Risk}_{\text{no smoking}}} = \frac{0.625}{0.565} = 1.106$$

Case-control Study

Case-control or cross-sectional study,

Table 2: Smoking vs lung cancer

	Smoking	No smoking	Total
Lung cancer	20	95	115
No lung cancer	12	73	85
Total	32	168	200

Inappropriate to calculate relative risk due to the study design

→ calculate **odds** and **odds ratio**

Odds Ratio

Odds of being a smoker,

$$\text{Odds}_{\text{lung cancer}} = \frac{20}{95} = 0.211$$

$$\text{Odds}_{\text{no lung cancer}} = \frac{12}{73} = 0.164$$

Odds ratio,

$$\text{OR} = \frac{\text{Odds}_{\text{lung cancer}}}{\text{Odds}_{\text{no lung cancer}}} = \frac{0.211}{0.164} = 1.287$$

Approximates the relative risk, RR = 1.106 (closer when the outcome is rare)

Odds Ratio

Odds ratio calculated from the cohort study (Table 1),

$$\text{OR} = \frac{20}{12} \div \frac{95}{73} = \frac{20 \cdot 73}{12 \cdot 95} = 1.281$$

= odds ratio from Table 2 (slight difference due to rounding error)

The use of odds ratio is appealing,

- applicable to all study designs
- obtained from the logistic regression model coefficients

Logit Transformation

Logit Function

Logit link function applied on the logistic regression model,

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

Linear, ranges from $-\infty$ to $+\infty$

Allows calculation of the odds, and subsequently the odds ratio

Logit Function

Given,

$$p=\frac{e^z}{1+e^z}$$

then,

$$1-p=1-\frac{e^z}{1+e^z}=\frac{1+e^z-e^z}{1+e^z}=\frac{1}{1+e^z}$$

Logit Function

Odds,

$$\frac{p}{1-p}=\frac{e^z}{1+e^z}\div\frac{1}{1+e^z}=\frac{e^z\left(1+e^z\right)}{1+e^z}=e^z$$

Log odds,

$$ln\left(\frac{p}{1-p}\right)=ln\left(e^z\right)=z=\alpha+\beta x$$

Logit Function

Thus,

$$logit(p) = \alpha + \beta x$$

is the *ln(odds)*

Not really useful in the current form

→ what we need is the **odds ratio**

Odds Ratio: Binary x

Odds ratio, OR when $x = 0, 1$,

$$\text{OR} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{e^{z_1}}{e^{z_0}}$$

$$= \frac{e^{\alpha+\beta(1)}}{e^{\alpha+\beta(0)}} = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

Odds Ratio: Numerical x

OR when x is a numerical variable, e.g. age,

$$\text{OR} = \frac{e^{\alpha+\beta x_1}}{e^{\alpha+\beta x_0}} = e^{\beta(x_1 - x_0)}$$

where x_1 and x_0 are any two values of x

Interested in the change (difference) between the values, $\Delta = x_1 - x_0$,

$$\text{OR} = e^{\beta\Delta}$$

Odds Ratio: Numerical x

For 1 unit change in x,

$$\text{OR} = e^{\beta}$$

Same formula as when x is binary

For c units change in x, e.g. an increase of 10 mmHg in SBP,

$$\text{OR} = e^{c\beta} = e^{10\beta}$$

Maximum Likelihood Estimation

Estimation Method

- Linear regression: ordinary least squares (OLS)
  - α and β minimise the sum of squared deviations of observed Y from predicted Y
- Logistic regression model is nonlinear
  - OLS cannot be used
- Maximum likelihood (ML) estimation
  - to estimate the unknown parameters α and β

Probability vs Likelihood

One subject, outcome $y = 0, 1$, follows a **Bernoulli** distribution,

$$P(Y=y|p)=p^{y}(1-p)^{1-y}$$

Probability: p fixed, data vary

- $p = 0.6$: $P(Y = 1) = 0.6$, $P(Y = 0) = 0.4$ → sum to 1

Likelihood: data fixed, p varies

- observed $y = 1$: $l(p) = p$, e.g. $l(0.3) = 0.3$, $l(0.8) = 0.8$ → not sum to 1

Same expression, read as a function of the parameter,

$$l(\theta)=l(\theta|y)=P(Y=y|x,\theta)$$

y: observed outcomes (the data); x: predictor values (fixed)

Likelihood Function

For *n* independent subjects, the likelihood of the observed outcomes given the predictor values,

$$l(\theta)=P(Y_1=y_1,\cdots,Y_n=y_n|x,\theta)=\prod_{i=1}^{n}P(Y_i=y_i|x_i,\theta)$$

$$=\prod_{i=1}^{n}p(x_i)^{y_i}\left[1-p(x_i)\right]^{1-y_i}$$

where,

$$p(x_i)=P(Y_i=1|x_i)=\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}},\qquad \theta=(\alpha,\beta)$$

Log Likelihood

Log likelihood,

$$L(\theta)=ln\left[l(\theta)\right]=\sum_{i=1}^{n}\left\{y_i ln\left[p(x_i)\right]+(1-y_i)ln\left[1-p(x_i)\right]\right\}$$

Products become sums → easier to compute and to maximise

Likelihood Function

ML estimators maximise the log likelihood,

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

ln is an increasing function → maximising $L(\theta)$ = maximising $l(\theta)$

Found iteratively by computer software.

Likelihood Function

Table 2: Smoking vs lung cancer

	Smoking	No smoking	Total
Lung cancer	20	95	115
No lung cancer	12	73	85
Total	32	168	200

Example, using data in Table 2,

$$l(\theta)=p(1)^{20}\times\left[1-p(1)\right]^{12}\times p(0)^{95}\times\left[1-p(0)\right]^{73}$$

$l(\theta)$ is maximised at $p(1) = 20/32 = 0.625$ and $p(0) = 95/168 = 0.565$ based on Table 2.

Likelihood: Example

Smokers: 20 with lung cancer, 12 without,

$$p(1)^{20}\left[1-p(1)\right]^{12}$$

$p(1)$	0.300	0.500	0.625	0.800
l	4.8×10^{-13}	2.3×10^{-10}	$\mathbf{6.4 \times 10^{-10}}$	4.7×10^{-11}

Non-smokers: 95 with lung cancer, 73 without,

$$p(0)^{95}\left[1-p(0)\right]^{73}$$

$p(0)$	0.300	0.500	0.565	0.800
l	1.0×10^{-61}	2.7×10^{-51}	$\mathbf{1.1 \times 10^{-50}}$	5.9×10^{-61}

Maximum at the observed risks, $p(1) = 20/32$ and $p(0) = 95/168$

Likelihood: Bernoulli

Each subject: Bernoulli (binomial with $n = 1$); for independent subjects,

$$l(\theta) = \prod_{i=1}^{n} p(x_i)^{y_i} \left[1 - p(x_i)\right]^{1-y_i}$$

All 32 smokers share the same $p(1)$; 20 have $y = 1$ and 12 have $y = 0$,

$$\prod_{i \in \text{smokers}} p(1)^{y_i} \left[1 - p(1)\right]^{1-y_i} = p(1)^{\Sigma y_i} \left[1 - p(1)\right]^{32 - \Sigma y_i}$$

$$= p(1)^{20} \left[1 - p(1)\right]^{12}$$

$\rightarrow$ the smokers' part of $l(\theta)$; likewise $p(0)^{95}[1 - p(0)]^{73}$ for non-smokers

Only the counts (20 of 32, 95 of 168) matter, not the order of subjects

Likelihood: Binomial

Number of cases among smokers, $K = 20$ of $n = 32 \rightarrow$ **binomial**,

$$P(K=20|p(1))=\binom{32}{20}p(1)^{20}\left[1-p(1)\right]^{12}$$

= C(32, 20) × smokers' part of $l(\theta)$

For Table 2, smokers and non-smokers together,

$$P(K_1=20, K_0=95|\theta)=\binom{32}{20}\binom{168}{95}l(\theta)$$

The constants (i.e. C(...) parts) do not involve $\theta \rightarrow$ same maximum, same estimates

- 0/1 (Bernoulli) or grouped (binomial) data → family = binomial in R

Likelihood: Logistic Model

Maximising $l(\alpha, \beta)$ of the logistic regression model for Table 2 gives,

Non-smokers ($x = 0$): α is the baseline ln(odds),

$$\hat{\alpha} = ln\left(\frac{p(0)}{1-p(0)}\right) = ln\left(\frac{95}{73}\right) = 0.263$$

Smokers ($x = 1$): ln(odds) = $\alpha + \beta$, so β is the change in ln(odds) attributable to smoking,

$$\hat{\beta} = ln\left(\frac{p(1)}{1-p(1)}\right) - \hat{\alpha} = ln\left(\frac{20}{12}\right) - ln\left(\frac{95}{73}\right) = 0.247$$

$$e^{\hat{\beta}} = e^{0.247} = 1.281 = \text{OR}$$

Testing Significance

Testing Significance

- Likelihood ratio test, G
- Wald test, W

Likelihood Ratio Test

$$G = -2\left[L(\text{model without } x) - L(\text{model with } x)\right]$$

$$G = -2(L_0 - L_1)$$

G follows chi-square distribution, P-value is

$$P\left[\chi^2(df) > G\right]$$

Degrees of freedom, df = difference in number of parameters between the models

Likelihood Ratio Test

Alternatively, using deviance, D (−2 log likelihood in SPSS),

$$G = D(\text{model without } x) - D(\text{model with } x)$$

$$G = D_0 - D_1$$

Wald Test

$$W = \frac{\hat{\beta}}{\widehat{SE}(\hat{\beta})}$$

W follows standard normal distribution, two-tailed P-value is

$$P(|z| > W)$$

More suitable for testing a single variable

Hands-on

Dataset

- slog.sav / slog.csv
  - modified from a dataset, courtesy of AP Dr. Kamarul Imran Musa
- Dependent variable (DV): *cad* (1: Yes, 0: No)
- Independent variables (IV):
  - categorical: *gender* (1: Male, 0: Female)
  - numerical: *sbp* (systolic blood pressure)

SPSS

1. **Analyze → Regression → Binary Logistic...**
2. In **Logistic Regression** window, **Dependent:** *cad*, **Covariates:** *gender*
3. Click **Categorical...** button
   - place *gender* under **Categorical Covariates:**
   - **Change Contrast → Reference Category: First → Change → Continue**

SPSS

1. Click **Options...** button
   - choose **Iteration history** and **CI for exp(B)** → **Continue**
2. Click **OK**
3. Repeat the same steps for *sbp*

R: Categorical x

```
data = read.csv("slog.csv")

# gender, categorical
table(gender=2-data$gender, cad=2-data$cad)
lreg = glm(cad ~ gender, data = data,
           family = binomial(link = "logit"))
summary(lreg)
cbind(coef(lreg), confint(lreg))
exp(cbind(coef(lreg), confint(lreg)))
```

R: Numerical x

```
# sbp, numerical
lreg1 = glm(cad ~ sbp, data = data,
            family = binomial(link = "logit"))
summary(lreg1)
cbind(coef(lreg1), confint(lreg1))
exp(cbind(coef(lreg1), confint(lreg1)))

# an increase in 10mmHg in sbp
exp(cbind(coef(lreg1)[[2]]*10, confint(lreg1)[[2]]*10,
          confint(lreg1)[[4]]*10))
```

References

Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. I. (2008). *Analysis of multivariate social science data*. USA: CRC Press.

Hosmer, D., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley Series in Probability and Statistics. USA: John Wiley & Sons, Inc.

Kleinbaum, D., & Klein, M. (2002). *Logistic regression: A self-learning text* (2nd ed.). Statistics for Biology and Health. USA: Springer New York.