

# **Part II: Generalized Linear Models**

## **Chapter II.3**


### **Models for Binary Response**

## > To be discussed...

- > Binary Regression
- > Logistic Regression
- > Goodness of Fit for Logistic Regression
- > Infinite Estimates in Logistic Regression
- > Logistic Regression with Categorical Predictors
- > Linear Logit Model for Binary Response

# Models for Binary Response

## II.3.1 Remark (ungrouped vs. grouped data)

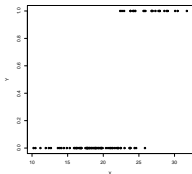
- Data for binary responses are usually *ungrouped*, in which case each observation  $y_i$ ,  $i = 1, \dots, n$ , is the realization of a single Bernoulli trial, i.e.  $y_i \in \{0, 1\}$  and is a realization of  $Y_i \sim \mathcal{B}(1, \pi_i)$ . The total sample size is then  $n_{tot} = \sum_{i=1}^n 1 = n$ .
- For *grouped* data, each observation  $y_i^*$ ,  $i = 1, \dots, n$ , corresponds to a set of sampled subjects/items which have exactly the same value for all explanatory variables. Hence, they are realizations of  $Y_i^* \sim \mathcal{B}(m_i, \pi_i)$  and thus  $y_i^* \in \{0, \dots, m_i\}$ . In this case, these  $m_i$  subjects corresponding to the  $i$ -th group are realizations of independent Bernoulli trials with the same success probability  $p_i$  (naturally, since  $p_i$  is modeled by the explanatory variables) and  $Y_i = Y_i^*/m_i$  (s. Example II.2.5). The total sample size in this case equals  $n_{tot} = \sum_{i=1}^n m_i$ .  
 Usually grouped data occur when all explanatory variables are categorical.
- A grouped data set can be converted to an ungrouped one. Unless otherwise stated explicitly, we shall consider next ungrouped data.

## > II.3.2 Remark

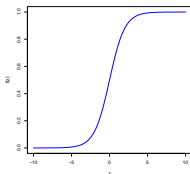
The **logistic regression model** is the most common choice for modeling a categorical response variable (usually binary or ordinal) in terms of explanatory variables, which can be continuous or/and categorical.

- Consider a binary response  $Y \sim \mathcal{B}(1, p)$  with  $Y \in \{0, 1\}$  and success probability  $p = P(Y = 1)$ .  
→ Model the dependence of  $p$  on an explanatory variable  $X$ .

$$P(Y = 1) \stackrel{?}{=} \beta_0 + \beta_1 X$$



$$f(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z} \quad \leftarrow z = \beta_0 + \beta_1 X$$



$$P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} \Rightarrow \text{logit}[P(Y = 1)] = \log \left( \frac{P(Y=1)}{1-P(Y=1)} \right) = \beta_0 + \beta_1 X$$

$\uparrow$   
odds of success

# Binary Regression

## II.3.3 Logistic Regression as a GLM

Binary response:  $Y, Y \in \{0,1\}$  – Explanatory variables:  $X_2, \dots, X_p$   
( $X_1 = 1$  coefficient of the intercept)

- 1 Random component:  $Y_i \sim \mathcal{B}(1, \pi_i)$  belongs to the exponential dispersion family)  
 $\mu_i = E(Y_i) = \pi_i$  for subject  $i, i = 1, \dots, n$ .
- 2 Systematic component (linear predictor):  $\sum_{j=1}^p \beta_j x_{ij}$   
Parameter vector:  $\beta = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$
- 3 Link function  $g$  relating linear predictor to  $\mu_i$ :  $g(\mu_i) = \beta_1 + \sum_{j=2}^p \beta_j x_{ij}$   
canonical link:  $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta_i$   
(the **logit link** is the natural parameter in exponential family representation, s. Example II.2.15)

## II.3.4 Remark

Often, the explanatory variables are considered to be  $X_1, \dots, X_{p-1}$  and the index 0 is used for the coefficient of the intercept ( $X_0 = 1$ ). In this case  $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})' \in \mathbb{R}^p$  and the linear predictor is  $\sum_{j=0}^{p-1} \beta_j x_{ij}$ .

### ➤ II.3.5 Remark

Other popular link functions for the set-up in II.3.3 and the random responses  $Y_i$ ,  $i = 1, \dots, n$ , follow (s. also Example II.2.10).

- Probit (underlying normal latent variable):  $\Phi^{-1}(\pi_i) = \beta_1 + \sum_{j=2}^p \beta_j x_{ij}$
- Complementary log-log:  $\log[-\log(1 - \pi_i)] = \beta_1 + \sum_{j=2}^p \beta_j x_{ij}$

# Tolerance Distribution Justification for Binary Regression

**Framework:** Toxicological experiments (explanatory variable  $X$  is the dosage of a drug)

Bliss: "The sigmoid dosage-mortality curve, ..., is interpreted as a cumulative normal frequency distribution of the variation among the individuals of a population in their susceptibility to a toxic agent, which susceptibility is inversely proportional to the logarithm of the dose applied." \*

## II.3.6 Remark

Suppose that the binary response  $Y$  is determined by a continuous underlying explanatory variable  $T$  (tolerance) so that for  $i = 1, \dots, n$ :

$$Y_i = 1 \quad \text{if-f} \quad x \geq T_i \quad \text{and} \quad Y_i = 0, \quad \text{else,}$$

where  $T_i$  is the tolerance of the  $i$ -th person to a drug dosage equal to  $x$ . Then, for fixed  $X = x$ ,

$$P(Y = 1|x) = \pi(x) = P(T \leq x) = G(x) = F(\beta_1 + \beta_2 x),$$


for some 'standard' cdf and suggests models of the form

$$F^{-1}[\pi(x)] = \beta_0 + \beta_1 x,$$

for some cdf  $F$  (i.e. motivates to consider a GLM with link function  $g$  the inverse of this cdf).

\*C.I.Bliss (1935). The calculation of the dosage-mortality curve, Annals of Applied Biology, 22:134-167.

### ▶ II.3.7 Special Binary Regression Models

①  $F(x) = \frac{e^x}{1+e^x}$   standard logistic regression model:

$$P(Y = 1|x) = \pi(x) = P(T \leq x) = F(\beta_1 + \beta_2 x) = G(x) = \frac{e^{\beta_1 + \beta_2 x}}{1 + e^{\beta_1 + \beta_2 x}}$$

②  $F(x) = \Phi(x)$   probit model:

$$\pi(x) = \Phi(\beta_1 + \beta_2 x)$$



# Simple Logistic Regression: Interpretation

Linearized form of the **simple logistic regression** model (**logit transform**):

$$\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_1 + \beta_2 x$$

## II.3.8 Remark (the role of $\beta_2$ )

➤ odds =  $\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_1 + \beta_2 x} = e^{\beta_1} (e^{\beta_2})^x$  ; i.e. multiplicative effect

➤ Monotonicity

$$\beta_2 > 0 \quad : \quad \pi(x) \uparrow \text{ as } x \rightarrow \infty$$

$$\beta_2 < 0 \quad : \quad \pi(x) \downarrow \text{ as } x \rightarrow \infty$$

$$\beta_2 = 0 \quad : \quad \pi(x) \text{ constant}$$

➤  $\frac{\partial \pi(x)}{\partial x} = \beta_2 \pi(x) [1 - \pi(x)]$

Thus locally,  $\pi(x)$  is approximated by a line with corresponding slope 'close' to  $x_0$ :

$$\pi(x) \approx \pi(x_0) + \beta_2 \pi(x_0) (1 - \pi(x_0)) (x - x_0) .$$

Slope is proportional to  $\beta_2$  and steepest ( $\beta_2/4$ ) at  $x$ -value where  $\pi(x) = 0.50$ ; this  $x$  value is  $x = -\beta_1/\beta_2$ , known as *median effective level*.

### ► II.3.9 Remark (interpretation of $\beta_2$ )

For two levels of  $x$ , denoted by  $x_1$  and  $x_2$ ,

$$\text{odds ratio} = \frac{\pi(x_1)/[1 - \pi(x_1)]}{\pi(x_2)/[1 - \pi(x_2)]} = \frac{e^{\beta_1 + \beta_2 x_1}}{e^{\beta_1 + \beta_2 x_2}} = e^{\beta_2(x_1 - x_2)}$$

- For  $x_1 - x_2 = 1$ , the odds of a success at  $x = x_1$  are  $e^{\beta_2}$  times the odds of success at  $x = x_2$ , i.e., odds multiply by  $e^{\beta_2}$  for every 1-unit increase in  $x$ .
- $\beta_2 = 0 \iff \text{odds ratio} = 1 \iff \text{no effect}$  of  $x$  on  $Y$

### ► II.3.10 Remark (generalization to multiple logistic regression model)

$$\log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q$$

In this case,  $\exp(\beta_j)$  represents the odds ratio between  $Y$  and two levels of  $x_j$  that are 1-unit apart ( $j = 2, \dots, p$ ), adjusting for all other predictors in the model.

# Logistic Regression: Estimation of $\beta$

## ► II.3.11 Remark

The logit link is the *canonical* link for a binomial response and from II.2.19 and II.2.24 we have the following results that apply to grouped or ungrouped data.

- Likelihood Equations for  $\beta$  (s. Remark II.2.21):

$$\sum_{i=1}^n m_i (y_i - \pi_i) x_{ij} = 0, \quad j = 1, \dots, p,$$

i.e. the sufficient statistics of  $\beta_j$ 's are equated to their expected values.

- Covariance matrix of  $\hat{\beta}$ :

For the information matrix for GLMs with canonical link holds  $\mathcal{I}_F = \mathcal{I}_F^{obs} = \mathbf{X}' \mathbf{W}_c \mathbf{X}$  with  $\mathbf{W}_c = \text{diag}(w_1, \dots, w_n)$ , where  $w_i = \frac{b''(\vartheta_i)}{a(\phi; i)}$ . In this case (s. Example II.2.5):

$w_i = \frac{\text{Var}(Y_i)}{[a(\phi; i)]^2} = \frac{\pi_i(1-\pi_i)/m_i}{1/m_i^2} = m_i \pi_i(1 - \pi_i)$  and thus the *estimated* covariance matrix of  $\hat{\beta}$  is

$$\text{Cov}(\hat{\beta}) = \mathcal{I}_F^{-1} = (\mathbf{X}' \text{diag}[m_i \hat{\pi}_i(1 - \hat{\pi}_i)] \mathbf{X})^{-1}.$$

- Standard iterative methods apply for solving the system of likelihood equations for the logistic regression model. Furthermore, the Newton-Raphson and the Fischer scoring algorithms coincide (s. Remark II.2.29).

### ► II.3.12 Example (cancer remission)

Predictors for cancer remission

$Y$ : cancer remission (1 = yes, 0 = no)

$x$ : labeling index (LI)

**Data:** 27 Bernoulli outcomes or 14 binomials

(for example: 2 observations with LI=8 and  $Y=0$ );

same likelihood function and ML estimates either way.

```
LI <- c(8,8,10,10,12,12,12,14,14,14,16,16,16,18,20,20,20,
22,22,24,26, 28,32,34,38,38,38)
y <- c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,1,0,0,1,1,0,1,1,1,0)
logit.fit <- glm(y ~ LI, family=binomial(link=logit))
```

↪ What is saved under 'logit.fit'?  names(logit.fit)

### II.3.13 Example (cancer remission) - continues

```
summary(logit.fit)
```

```
Call:    glm(formula = y ~ LI, family = binomial(link = logit))
```

```
Deviance Residuals:
```

Min	1Q	Median	3Q	Max
-1.9448	-0.6465	-0.4947	0.6571	1.6971

```
Coefficients:
```

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.77714	1.37862	-2.740	0.00615 **
LI	0.14486	0.05934	2.441	0.01464 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 34.372 on 26 degrees of freedom
```

```
Residual deviance: 26.073 on 25 degrees of freedom
```

```
AIC: 30.073
```

```
Number of Fisher Scoring iterations: 4
```

### ► II.3.14 Example (cancer remission) - continues

ML fit of logistic regression model for  $\pi = P(\text{remission})$  is

$$\log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = \hat{\beta}_1 + \hat{\beta}_2 x = -3.777 + 0.145x$$

Prediction equation:

$$\hat{\pi}(x) = \frac{\exp(-3.777 + 0.145x)}{1 + \exp(-3.777 + 0.145x)}$$

Thus, at  $\bar{x} = 20.1$ ,

$$\hat{\pi}(\bar{x}) = \frac{\exp(-3.777 + 0.145 \cdot 20.1)}{1 + \exp(-3.777 + 0.145 \cdot 20.1)} = 0.296$$

The incremental rate of change at  $x = 20.1$  is

$$\hat{\beta}_2 \hat{\pi}(\bar{x}) [1 - \hat{\pi}(\bar{x})] = 0.14486(0.296)(0.704) = 0.030$$

Furthermore,  $\hat{\pi} = 0.50 \Leftrightarrow \log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = 0 = \hat{\beta}_1 + \hat{\beta}_2 x \Leftrightarrow x = -\frac{\hat{\beta}_1}{\hat{\beta}_2} = 26.0$

### ► II.3.15 Example (cancer remission) - continues

**Interpretation of  $\hat{\beta}_2 = 0.145$ :** for each unit change in LI, the estimated odds of remission are multiplied by  $\exp(0.145) = 1.16$ , i.e., 16% increase when LI  $\uparrow$  1.

E.g., at  $x = 26$ ,  $\hat{\pi} = 0.498$  (odds = 0.990)

at  $x = 27$ ,  $\hat{\pi} = 0.534 = 0.990 \cdot 1.16$  (odds = 1.145)

i.e., odds ratio =  $\frac{0.534/(1-0.534)}{0.498/(1-0.498)} = 1.16$

Simpler effect measures: Change in  $\hat{\pi}$  from minimum to maximum value of  $x$

(Here, as LI goes from 8 to 38,  $\hat{\pi}$  increases from 0.07 to 0.85)

**Remark:** With multiple predictors, can proceed to similar interpretations for a specific predictor setting all the other predictors equal to their means, modes (or at specific values of interest/levels).

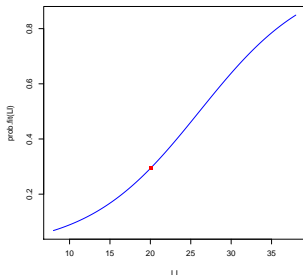
Estimated expected probabilities:

```
prob.fit <- function(x){  
  exp(logit.fit$coefficients[1]+ logit.fit$coefficients[2]*x)/  
  (1+exp(logit.fit$coefficients[1]+ logit.fit$coefficients[2]*x)) }  
prob.fit(LI)      # the same as: logit.fit$fitted.values  
prob.fit(mean(LI))
```

### ➤ II.3.16 Example (cancer remission) - continues

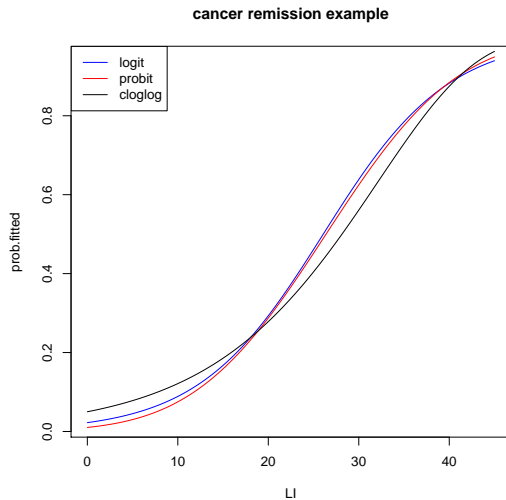
Plot of the estimated expected probabilities:

```
low <- min(LI); up <- max(LI)
plot(prob.fit, col='blue', from=low, to=up, xlab='LI', ylab='prob.fit(LI)')
x0 <- mean(LI); y0 <- prob.fit(mean(LI))
points(x0, y0, pch=15, col='red')
```





## II.3.17 Example (cancer remission) - continues



### II.3.18 Example (cancer remission) - continues

#### Tests of Significance and CIs

👉 Test of no effect ( $H_0 : \beta_2 = 0$ )

SE = Standard Error

➤  $z = \hat{\beta}_2 / SE = \frac{0.145}{0.059} = 2.45$  ( $z^2 = 5.96 \sim \chi_1^2$ , under  $H_0$ , called **Wald** statistic)

Strong evidence of a positive association between cancer remission and labeling index ( $p$ -value = 0.015).

👉 Confidence Interval (CI) for  $\beta_2$     **significant if the interval does not contain 0**

➤ 95% Wald CI:  $\hat{\beta}_2 \pm 1.96(SE) = (0.029, 0.261)$

(based on inverting the test above, e.g., the 95% CI is the set of  $\beta_2$  not rejected at the 5% level in testing  $H_0$  against  $H_1 : \beta_2 \neq 0$ )

### II.3.19 Remark

- ➊ Beyond the Wald test statistic, there exist also the likelihood-ratio and the score test statistics. The three types of tests are asymptotically equivalent, when  $H_0$  is true.
- ➋ There exist other types of CIs, based on inverting the likelihood ratio and the score tests.
- ➌ Methods extend to inference for multiple parameters.