8.1 Logistic Regression

Mar 27, 2018

PRESENTER : JongYun Kim

UNIST Autonomous System LAB
Address. 112-#810, 50 UNIST-gil, Ulju-gun, Ulsan, 44919, South Korea
Tel. +82 52 217 2368 **Web.** https://sites.google.com/site/aslunist/

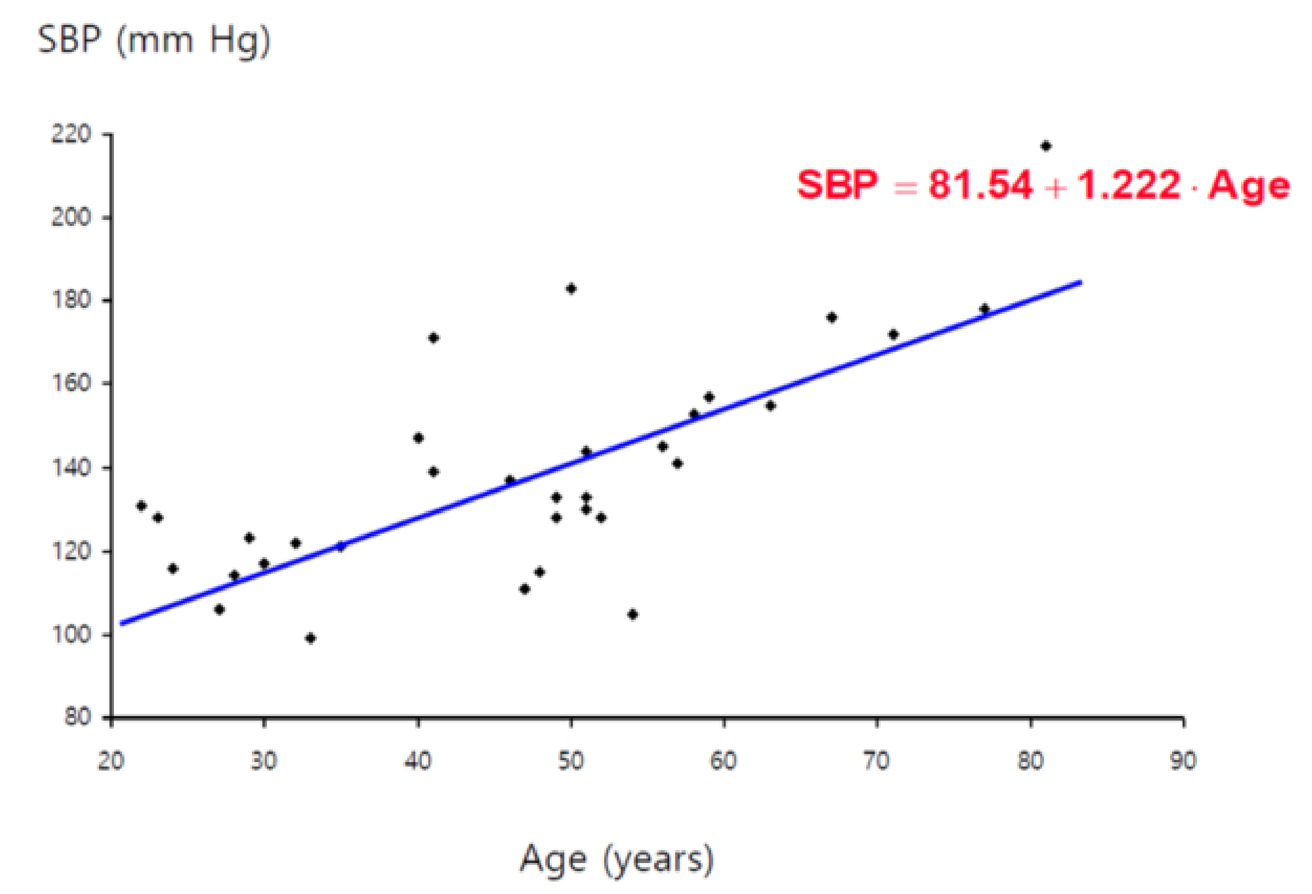
CONTENTS

Introduction

Age	SBP
22	131
23	128
24	116
27	106
28	114
29	123
30	117
32	122
33	99
35	121
40	147

Age	SBP
41	139
41	171
46	137
47	111
48	115
49	133
49	128
50	183
51	130
51	133
51	144

Age	SBP
52	128
54	105
56	145
57	141
58	153
59	157
63	155
67	176
71	172
77	178
81	217



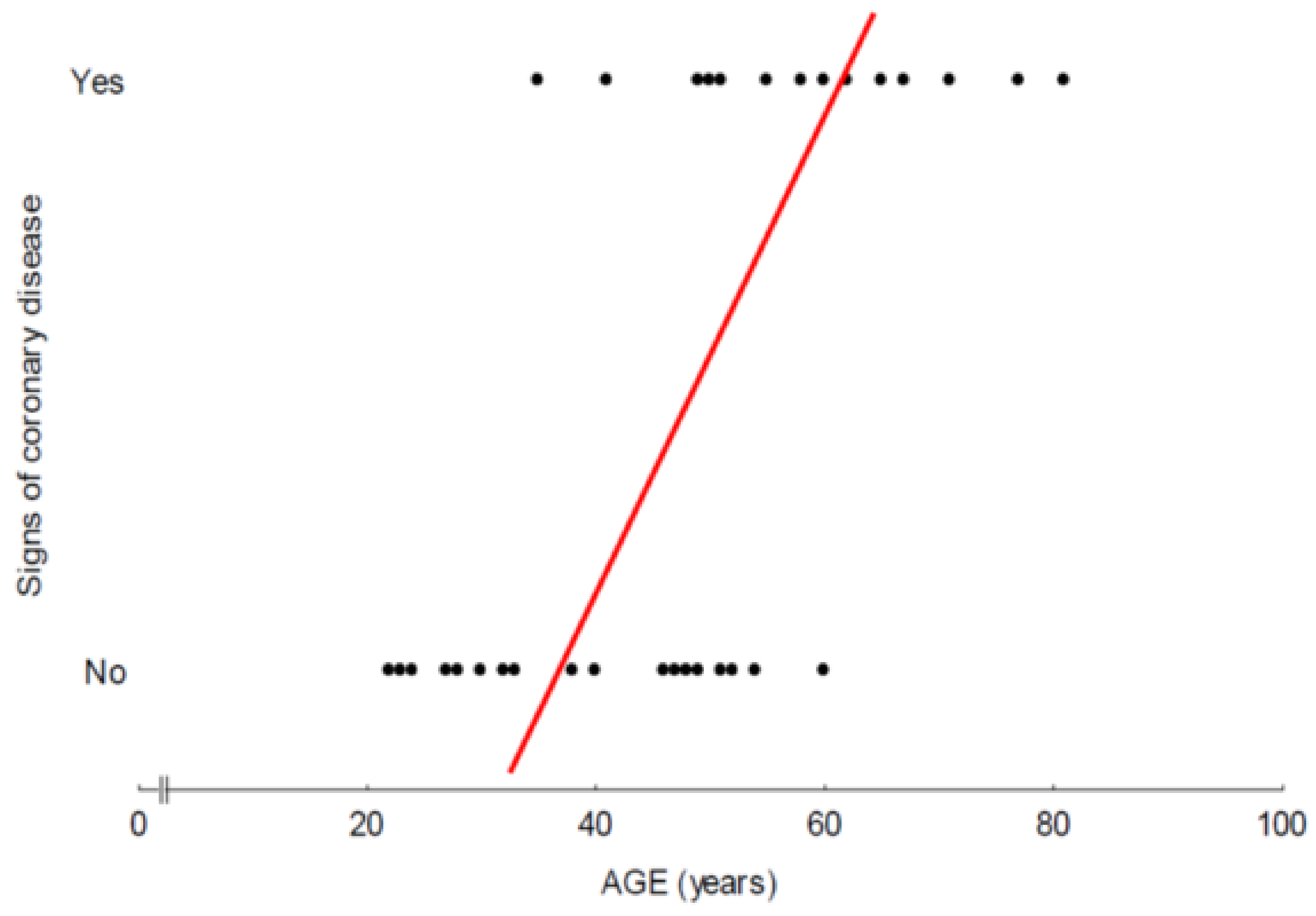
- Linear Regression
- $\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon$
- Estimates continuous variable

  (blood pressure in the example)

Age	CD
22	0
23	0
24	0
27	0
28	0
30	0
30	0
32	0
33	0
35	1
38	0

Age	CD
40	0
41	1
46	0
47	0
48	0
49	1
49	0
50	1
51	0
51	1
52	0

Age	CD
54	0
55	1
58	1
60	1
60	0
62	1
65	1
67	1
71	1
77	1
81	1



- What about estimating **categorical (count) variable or probability (proportion)** ??
- Could be **{0, 1}** or **[0, 1]**
- If applying linear regression, ……?!

Odds and Logistic Transformation

Logit parameter

$$Odds = \frac{P(A)}{P(A^c)} = \frac{\pi}{1-\pi}$$ $\Rightarrow$ in the range [0, ∞]

$$\lambda = \log\{odds\} = \log\frac{\pi}{1-\pi}$$ $\Rightarrow$ in the range **[−∞, ∞]**

$$\pi = \alpha_0 + \alpha_1 \boldsymbol{X} \xrightarrow{\text{Logistic transformation}} \log\left\{\frac{\pi}{1-\pi}\right\} = \alpha_0 + \alpha_1 \boldsymbol{X}$$

Now, we can consider count or proportion data as holding linear regression frames

Example : dose response

Group i

10 mice $n = 10$

Dose $x_i = i$ (1~N=11)

injection

y_i = # of mice dying in i-th group

$p_i = y_i / n$

We are going to model y_i as independent binomials

$$y_i \overset{\text{ind}}{\sim} \text{Bi}(n_i, \pi_i) \qquad \text{for } i = 1, 2, \ldots, N$$

Assume that the logit follows linear function of dose

$$\lambda_i = \log\left\{\frac{\pi_i}{1 - \pi_i}\right\} = \alpha_0 + \alpha_1 x_i$$

Example : dose response

Group i

10 mice $n = 10$

Dose $x_i = i$

(1~N=11)

injection

y_i = # of mice dying in i-th group

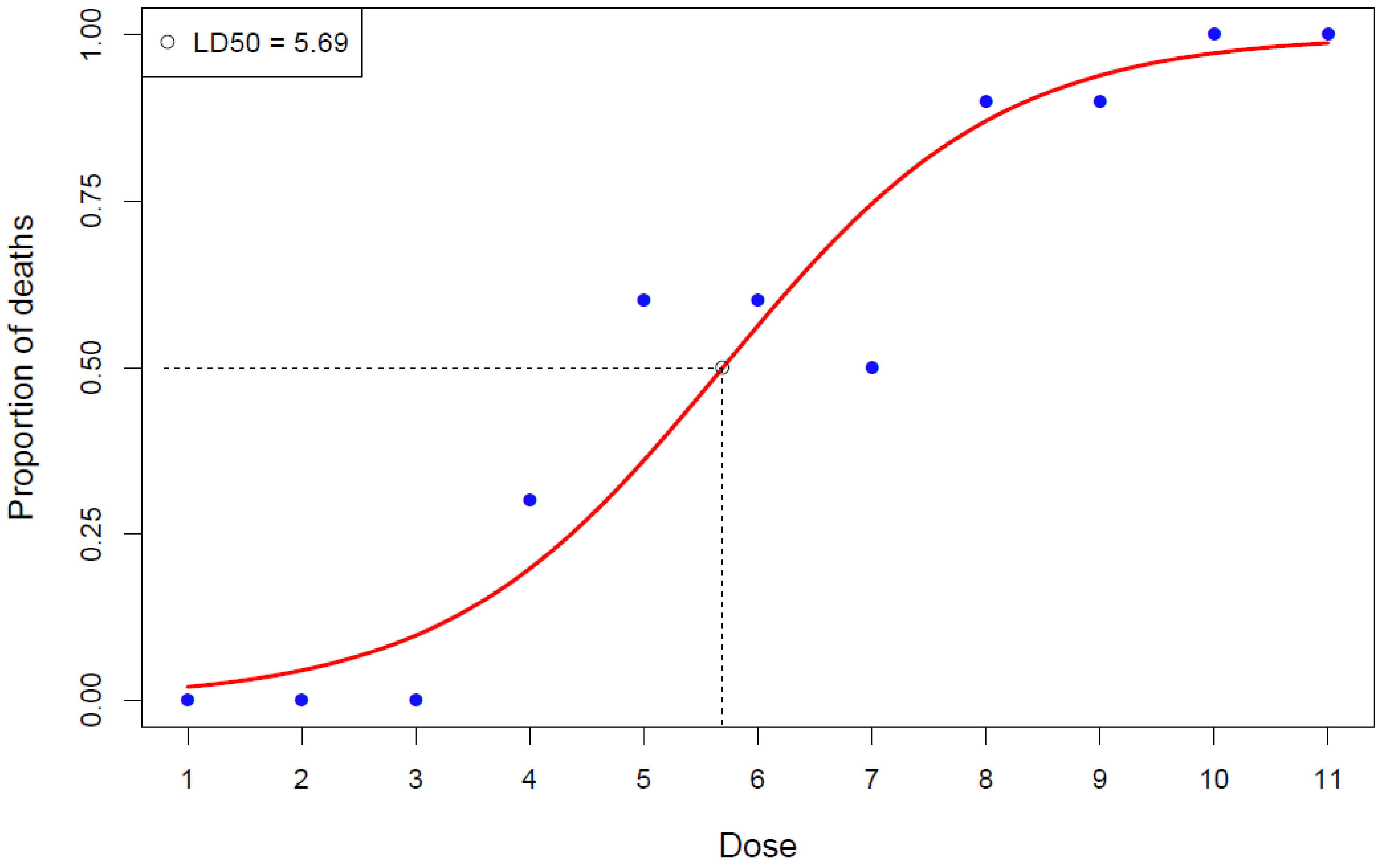
$p_i = y_i/n$ **(blue dots)**

MLE provides $(\hat{\alpha}_0, \hat{\alpha}_1)$ which yields the following equation

$$\hat{\lambda} = \log\left\{\frac{\hat{\pi}}{1-\hat{\pi}}\right\} = \hat{\alpha}_0 + \hat{\alpha}_1 \boldsymbol{x}$$

And we finally obtain the **linear logistic regression curve**

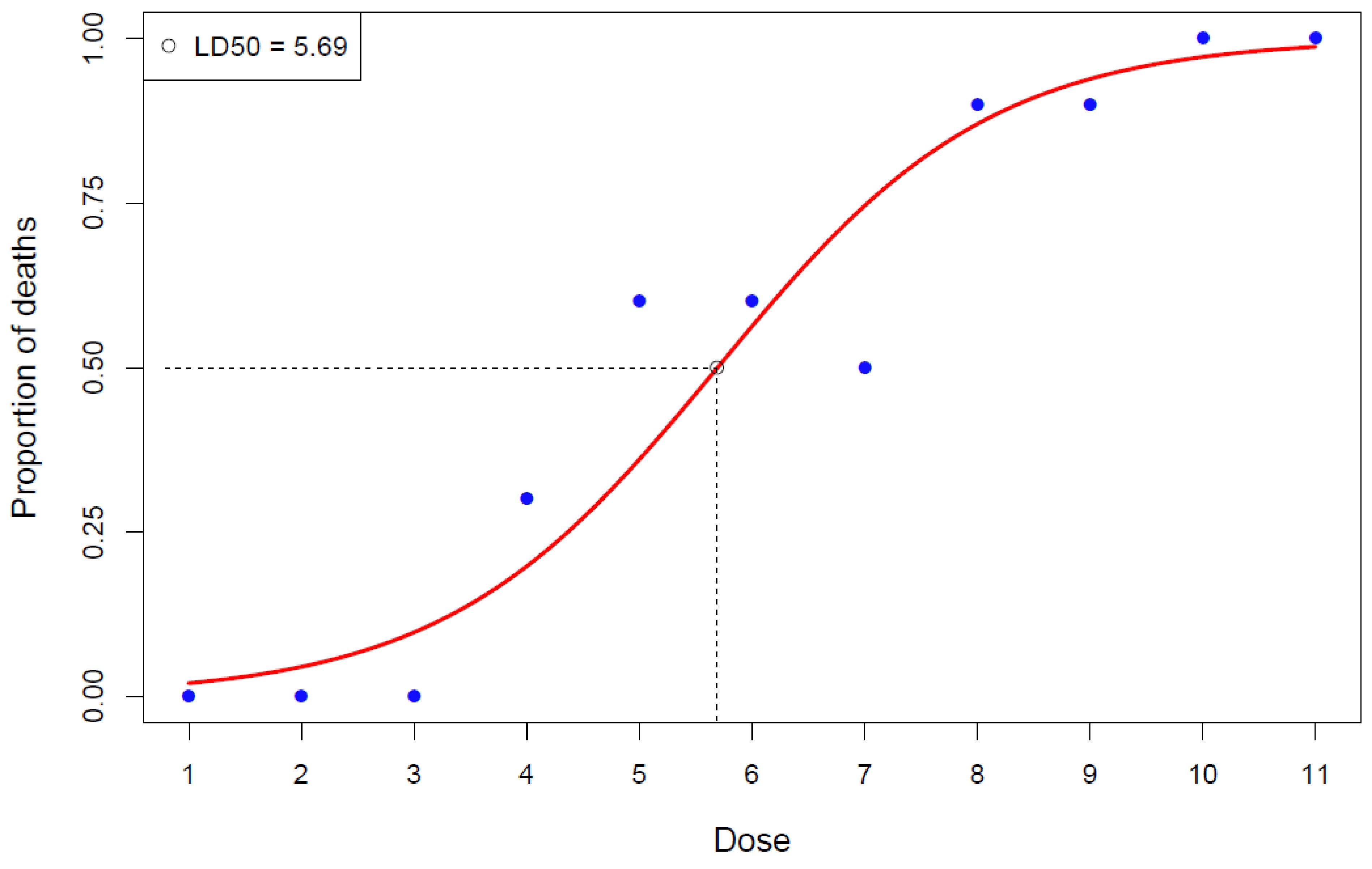
$$\hat{\pi}(x) = \left(1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 x)}\right)^{-1}$$



Note : dose response

- The regression **has reduced error** (see the table on the right side)
- It is ture unless logit linear model seriously goes wrong
- λ is NOT restricted to the range [0, 1]
- **Able to utilize exponential family properties!**

x	1	2	3	4	5	6	7	8	9	10	11
sd $\hat{\pi}(x)$	.015	.027	.043	.061	.071	.072	.065	.050	.032	.019	.010
sd p_i	.045	.066	.094	.126	.152	.157	.138	.106	.076	.052	.035



Merging Into Exponential Family

The probability density function of Bi(n, y) is given

$$\binom{n}{y}\pi^{y}(1-\pi)^{n-y} = e^{\lambda y - n\psi(\lambda)}\binom{n}{y}$$

→ **one parameter exponential family**

(see chapter 5.5 ; eq 5.54 or 5.46)

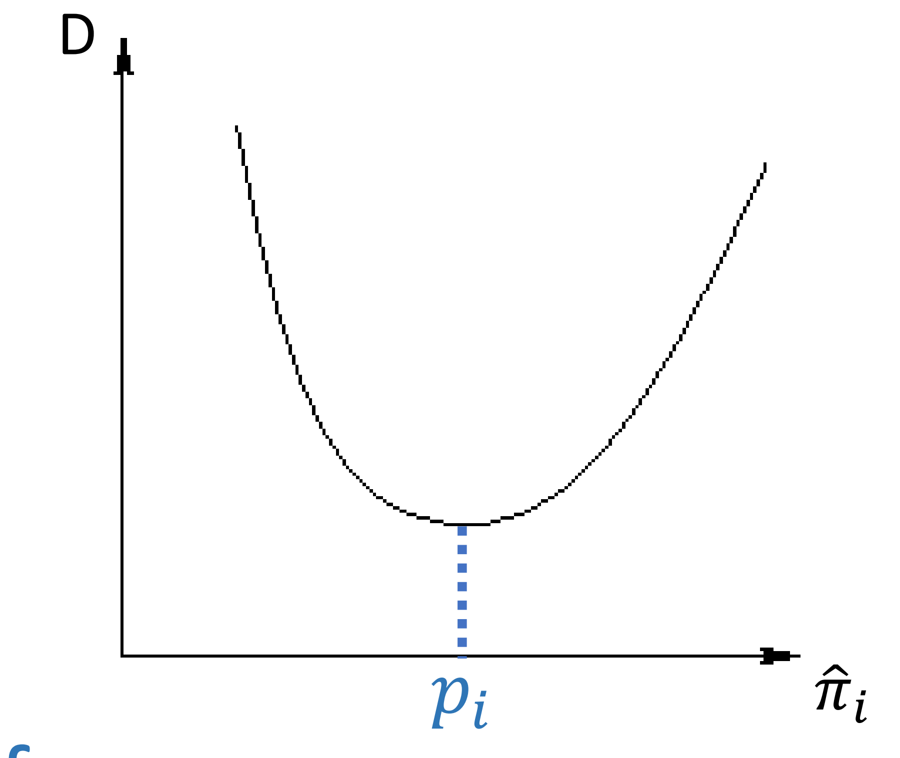
, where $\psi(\lambda) = \log\{1 + e^{\lambda}\}$

The independence of the data gives the probability density of full data set $\boldsymbol{y}$ as a function of (α_0, α_1),

$$\begin{aligned} f_{\alpha_0,\alpha_1}(\boldsymbol{y}) &= \prod_{i=1}^{N} e^{\lambda_i y_i - n_i\psi(\lambda_i)}\binom{n_i}{y_i} \\ &= e^{\alpha_0 S_0 + \alpha_1 S_1} \cdot e^{-\sum_1^N n_i\psi(\alpha_0+\alpha_1 x_i)} \cdot \prod_1^N \binom{n_i}{y_i} \end{aligned}$$

, where $S_0 = \sum_{i=1}^{N} y_i$ and $S_1 = \sum_{i=1}^{N} x_i y_i$

Merging Into Exponential Family



Suppose that the deviance is given as follows

$$D(p_i, \hat{\pi}_i) = 2n_i \left[p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (1-p_i) \log\left(\frac{1-p_i}{1-\hat{\pi}_i}\right) \right]$$

The deviance gives us the intuition : it is zero at $\hat{\pi}_i = p_i$, otherwise it increases as $\hat{\pi}_i$ departs further from p_i

The logistic regression MLE value $(\hat{\alpha}_0, \hat{\alpha}_1)$ has to do with minimizing the total deviance between p_i and $\hat{\pi}_i = \pi_{\alpha_0,\alpha_1}(x_i)$

$$(\hat{\alpha}_0, \hat{\alpha}_1) = \underset{(\alpha_0,\alpha_1)}{\arg\min} \sum_{i=1}^{N} D\left(p_i, \pi_{\alpha_0,\alpha_1}(x_i)\right)$$

Example : cell infusion

Let π_{ij} denote the true probability of thriving of ratio i during time period j

And take logistic regression

$$\lambda_{ij} = \log\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \mu + \alpha_i + \beta_i$$

MLE and the data set $\{p_{ij}\}$ give estimation $\hat{\pi}_{ij}$ as follows

$$\hat{\pi}_{ij} = \frac{1}{1+e^{-(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j)}}$$

i Ratio \ j Time	1	2	3	4	5
1	5/31 .11	3/28 .25	20/45 .42	24/47 .54	29/35 .75
2	15/77 .24	36/78 .45	43/71 .64	56/71 .74	66/74 .88
3	48/126 .38	68/116 .62	145/171 .77	98/119 .85	114/129 .93
4	29/92 .32	35/52 .56	57/85 .73	38/50 .81	72/77 .92
5	11/53 .18	20/52 .37	20/48 .55	40/55 .67	52/61 .84

data p_{ij}

$\hat{\pi}_{ij}$ *estimated*

Cell infusion data; human cell colonies infused with mouse nuclei in five ratios over 1 to 5 days and observed to see whether they did or not thrive.

Example : spam filter

George labeled N=4601 emails whether spam or ham(*i.e.* non-spam)
He used 57 words as predictors in the table.

x_{ij} : relative frequency of keyword j in email i
π_{ij} : true probability that email i is spam
λ_i : the logit transformation of π_{ij}

$$\lambda_{ij} = \log\left\{\frac{\pi_{ij}}{1-\pi_{ij}}\right\} = \alpha_0 + \sum_{j=1}^{57} \alpha_j x_{ij}$$

- Then you are able to predict whether future emails are spam or ham by using these keywords
- The table provides the estimated $\hat{\alpha}_j$ and its *se* value (by MLE)
- It seems that '*free*' and '*your*' are good spam predictors
  large $\hat{\alpha}_j$ and small se; large z-value
- The occasional very large $\hat{\alpha}_j$ may bother MLE

$\frac{\hat{\alpha}_j}{se}$

	Estimate	se	z-value		Estimate	se	z-value
intercept	−12.27	1.99	−6.16	lab	−1.48	.89	−1.66
make	−.12	.07	−1.68	labs	−.15	.14	−1.05
address	−.19	.09	−2.10	telnet	−.07	.19	−.35
all	.06	.06	1.03	857	.84	1.08	.78
3d	3.14	2.10	1.49	data	−.41	.17	−2.37
our	.38	.07	5.52	415	.22	.53	.42
over	.24	.07	3.53	85	−1.09	.42	−2.61
remove	.89	.13	6.85	technology	.37	.12	2.99
internet	.23	.07	3.39	1999	.02	.07	.26
order	.20	.08	2.58	parts	−.13	.09	−1.41
mail	.08	.05	1.75	pm	−.38	.17	−2.26
receive	−.05	.06	−.86	direct	−.11	.13	−.84
will	−.12	.06	−1.87	cs	−16.27	9.61	−1.69
people	−.02	.07	−.35	meeting	−2.06	.64	−3.21
report	.05	.05	1.06	original	−.28	.18	−1.55
addresses	.32	.19	1.70	project	−.98	.33	−2.97
free	.86	.12	7.13	re	−.80	.16	−5.09
business	.43	.10	4.26	edu	−1.33	.24	−5.43
email	.06	.06	1.03	table	−.18	.13	−1.40
you	.14	.06	2.32	conference	−1.15	.46	−2.49
credit	.53	.27	1.95	char;	−.31	.11	−2.92
your	.29	.06	4.62	char(	−.05	.07	−.75
font	.21	.17	1.24	char_	−.07	.09	−.78
000	.79	.16	4.76	char!	.28	.07	3.89
money	.19	.07	2.63	char$	1.31	.17	7.55
hp	−3.21	.52	−6.14	char#	1.03	.48	2.16
hpl	−.92	.39	−2.37	cap.ave	.38	.60	.64
george	−39.62	7.12	−5.57	cap.long	1.78	.49	3.62
650	.24	.11	2.24	cap.tot	.51	.14	3.75

THANK YOU

Q&A