

Index of Lecture 3b

Page	Title
1	Introduction to logistic regression
2	Example dataset <code>mice</code>
3	Why not linear regression?
4	Logit transformation
5	Logistic regression model
6	Logistic regression for mice data
7	2×2 -table analysis
8	2×2 -table and logistic regression
9	Linear vs. logistic modelling

INTRODUCTION TO LOGISTIC REGRESSION

Logistic regression:

- binary (0/1, dichotomous) outcomes, possibly grouped to binomial outcomes (e.g., 3 positive out of 10 animals),
- first of several regression-type models not relying on normal distribution assumptions,
 - * sometimes called *generalised linear models* (glm's),
 - * model building from the predictors similar to linear regression,
 - * some common features in their analysis that distinguish them from linear regression analysis.

Today's session:

- last topics of model building (Javier from lecture 3a),
- follow-up on this week's quiz (VHM 812),
- possibly further discussion of VER Exercise 14.3,
- review of simple logistic regression (one predictor) and its relation to known analyses, in particular 2×2 -table analysis,
- computer-assisted using Stata (good facilities available, superior to many simpler programs, e.g. Minitab).

VER/MER textbooks:

- today: 16.1–5+8 with some omissions (in next lecture).

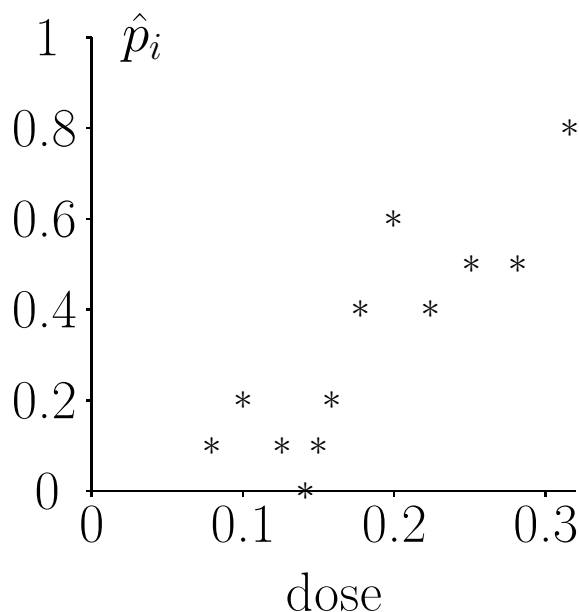
EXAMPLE DATASET MICE

Toxicity study of dose-response curve (Woodward 1941):

- lethality of different doses of chloracetic acid, measured as the mortality among 10 mice subjected to each dose,

group	dose	# mice died	# mice total	prop. died	
i	x_i	r_i	N_i	$\hat{p}_i = r_i/N_i$	$\text{logit}(\hat{p}_i)$
1	0.0794	1	10	0.1	-2.197
2	0.1000	2	10	0.2	-1.386
3	0.1259	1	10	0.1	-2.197
4	0.1413	0	10	0.0	undef.
5	0.1500	1	10	0.1	-2.197
6	0.1588	2	10	0.2	-1.386
7	0.1778	4	10	0.4	-0.405
8	0.1995	6	10	0.6	0.405
9	0.2239	4	10	0.4	-0.405
10	0.2512	5	10	0.5	0
11	0.2818	5	10	0.5	0
12	0.3162	8	10	0.8	1.386

- grouped binary data,
- statistical model:
 $r_i \sim \text{Bin}(N_i, p_i)$, and
 r_1, \dots, r_{12} independent,
- parameters: p_1, \dots, p_{12}
 (prob. of death in groups),
- question:
 how to use the doses?



WHY NOT LINEAR REGRESSION?

Regression for binary outcomes ($Y_i = 0$ or $Y_i = 1$)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i,$$

conflicts with the model assumptions:

- (1) errors ε_i are far from normally distributed (can only take two possible values¹),
- (2) with $p_i = P(Y_i = 1)$, we have $\text{Var}(Y_i) = p_i(1 - p_i)$, which is not constant when p_i is modelled by predictors,
- (3) both Y_i and p_i are bounded (do not go beyond 0 and 1) but linear predictions by the x -variables can easily give predictions outside the interval.

Regression for grouped binary outcomes (proportions r_i/N_i)

- same problems (1)–(3), although (1) is less severe,
- transformation is a possibility, usually with *variance-stabilising* transformation:
$$Y_i = \arcsin(\sqrt{r_i/N_i}), \quad \arcsin = \text{inverse sine function,}$$

– however, not recommended unless

 - * the denominators N_i are all “large” and approximately the same,
 - * the prop.’s r_i/N_i are not too extreme (close to 0 or 1),

and usually offers no advantages over logistic regression.

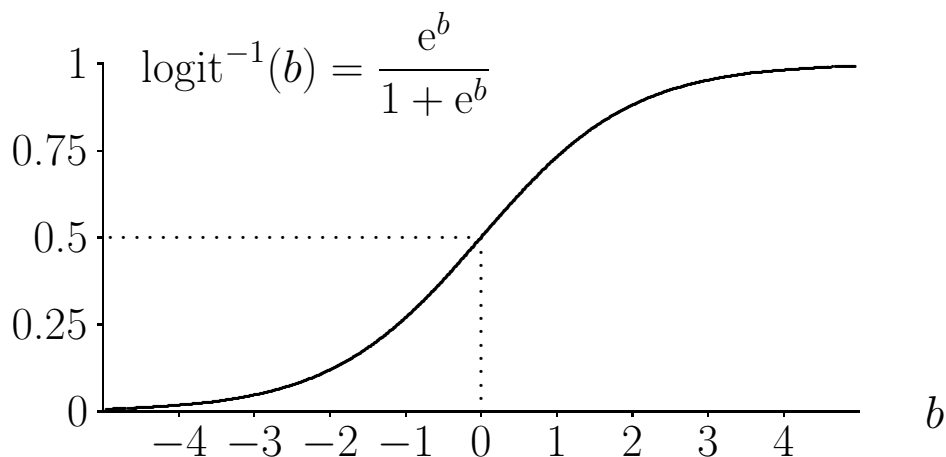
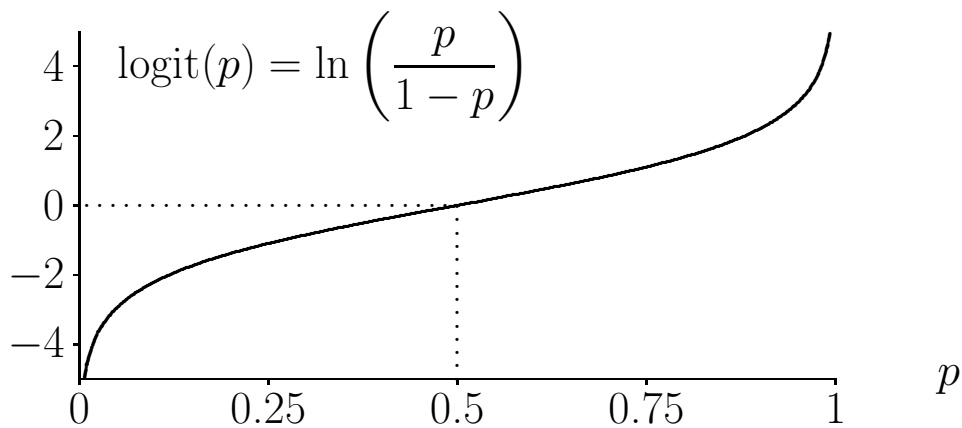
¹ Possible values for the error of obs. i are $\varepsilon_i = 1 - (\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$, and $\varepsilon_i = -(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki})$.

LOGIT TRANSFORMATION

Define for $0 < p < 1$ and any b ,

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right) \quad \text{and} \quad \text{logit}^{-1}(b) = \frac{e^b}{1+e^b} = \frac{1}{1+e^{-b}},$$

- the logit function stretches the interval $(0,1)$, excl. endpoints!, onto the entire real axis (from $-\infty$ to ∞),
- $\text{logit}(\frac{1}{2})=0$ and $\text{logit}(p)$ is increasing in p ,
- with $\text{odds}(p) = \frac{p}{1-p}$, we have $\text{logit}(p) = \ln(\text{odds}(p))$,
- logit and inverse logit functions:



LOGISTIC REGRESSION MODEL

= a different transformation approach:

- keep observations (binary/grouped binary) untransformed,
- transform probability parameter p by logit function to logit scale where linear modelling takes place, e.g.

$$\ln \left(\frac{p_i}{1 - p_i} \right) = \text{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad (1)$$

where

$$p_i = P(Y_i = 1),$$

$$Y_i = \begin{cases} 1 & \text{“success”} \\ 0 & \text{“failure”} \end{cases} \quad \text{for } i = 1, \dots, n,$$

$$x_i = \text{regression variable for observation } i.$$

Model assumptions:

- independence of all the observations (Y_i 's),
- linearity of relation (1) on logit scale.²

Grouped binary data (with N_i repl. in i th group)

- equation (1) for $p_i = \text{prob. of “success” in group } i$,
- same model as if set up as binary data (with $n = \sum_i N_i$).

Multiple logistic regression model for predictors x_1, \dots, x_k :

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}.$$

² For a single predictor model, the linearity assumption applies only to the case where x_1 is continuous.

LOGISTIC REGRESSION FOR MICE DATA

Estimates (with SE):

$$\hat{\beta}_0 = -3.57 (0.71),$$

$$\hat{\beta}_1 = 14.64 (3.33).$$

Test of $H_0: \beta_1 = 0$:

$$z = \hat{\beta}_1 / \text{SE}(\hat{\beta}_1) = 4.39$$

very sign. in $N(0,1)$

\Rightarrow strong effect of dose.

Estimated line:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{dose}$$

– on logit-scale.

Estimated curve $\hat{p}(x)$:

$$\text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{dose})$$

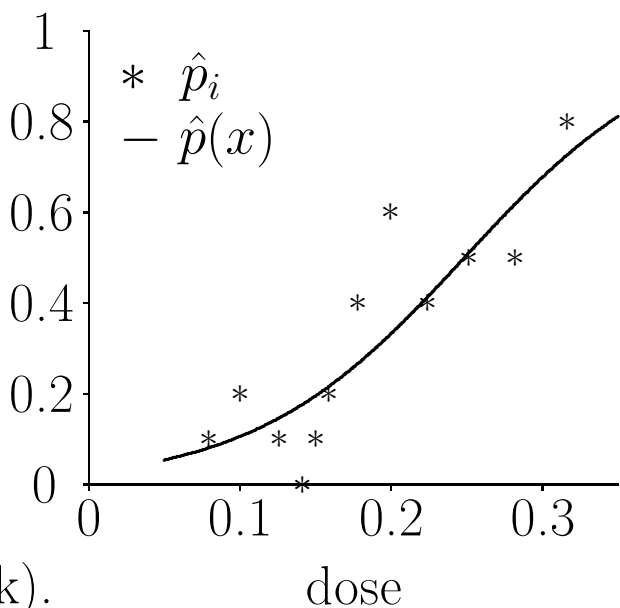
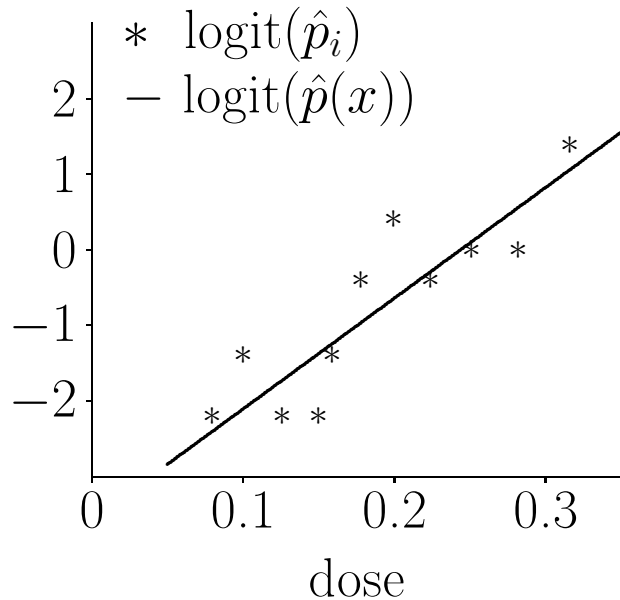
– on probability-scale.

Test of model:

(goodness-of-fit test)

$$X^2 = 8.74, P = 0.56$$

\Rightarrow no lack of fit (model ok).



Interpretation of $\hat{\beta}_1$: change in dose of a units \Rightarrow

- change in $\text{logit}(p)$ of $a\hat{\beta}_1$ units,
- change in $\text{odds}(p)$ by factor $\exp(a\hat{\beta}_1)$ (= *odds-ratio*), where $\text{odds}(p) = p/(1-p)$.

2×2 -TABLE ANALYSIS

Mice data: outcome = mortality, explanatory = dichotomous version of dose (for illustration only):

	dose>0.16		
dead	1	0	Total
1	32	7	39
0	28	53	81
Total	60	60	120

Statistical model (with binary dose predictor):

two binomial distributions $\text{Bin}(60, p_1)$ and $\text{Bin}(60, p_0)$,

– analyzed in VHM 801 by computing:

$$\begin{aligned} \hat{p}_1 &= 32/60 = 0.533, & \text{SE}(\hat{p}_1) &= \sqrt{\hat{p}_1(1-\hat{p}_1)/60} = 0.0644, \\ \hat{p}_0 &= 7/60 = 0.117, & \text{SE}(\hat{p}_0) &= \sqrt{\hat{p}_0(1-\hat{p}_0)/60} = 0.0414, \\ \hat{p}_1 - \hat{p}_0 &= 0.533 - 0.117 = 0.417, & \text{SE} &= \sqrt{0.0644^2 + 0.0414^2} = 0.077. \end{aligned}$$

Alternative ways of comparing the probabilities \hat{p}_1 and \hat{p}_0 :

relative risk : $\text{RR} = \hat{p}_1/\hat{p}_0 = 0.533/0.117 = 4.57,$

$$\begin{aligned} \text{odds-ratio : OR} &= \text{odds}(\hat{p}_1)/\text{odds}(\hat{p}_0) = [\hat{p}_1/(1-\hat{p}_1)] / [\hat{p}_0/(1-\hat{p}_0)] \\ &= [0.533/(1-0.533)] / [0.117/(1-0.117)] \\ &= 1.143/0.132 = 8.653 = (32 \cdot 53)/(28 \cdot 7). \end{aligned}$$

Advantages of these statistics (over the simple $\hat{p}_1 - \hat{p}_0$):

- multiplicative effects are more meaningful than additive for proportions bounded by 0 and 1,
- when both probabilities “close” to zero: OR \approx RR (clearly not the case in the example),
- both statistics more useful than $(\hat{p}_1 - \hat{p}_0)$ when multiple factors studied simultaneously.

2 × 2–TABLE AND LOGISTIC REGRESSION

Mice data: (same as on previous slide):

outcome = mortality, predictor = **dose2** (dichotomous version of dose):

	dose2 = (dose > 0.16)		
dead	1	0	Total
1	32	7	39
0	28	53	81
Total	60	60	120

Logistic regression model with **dose2** as predictor:

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{dose2}_i,$$

gives the estimates

$$\hat{\beta}_1 = 2.158 = \ln(8.653) = \ln(\text{OR}),$$

$$\hat{\beta}_0 = -2.024 = \text{logit}(0.117) = \text{logit}(\hat{p}_0).$$

Interpretations (valid also in multiple logistic regression):

- odds-ratio for effect of **dose2** = $e^{\hat{\beta}_1} = e^{2.158} = 8.653$,
- baseline prob. = $\text{logit}^{-1}(\hat{\beta}_0) = \text{logit}^{-1}(-2.024) = 0.117$.

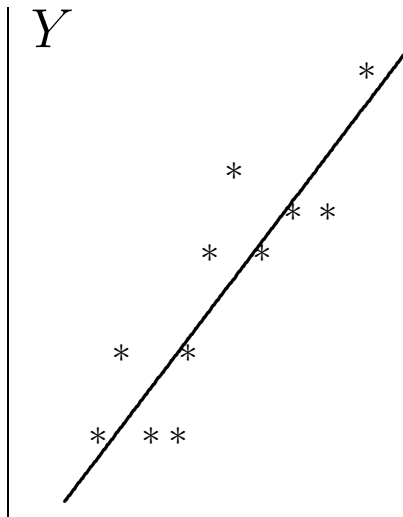
Summary:

The 2 × 2-table analysis and the logistic regression analysis are equivalent (also, the *P*-values are similar³).

³ The likelihood-ratio tests (next lecture) of the two models are identical.

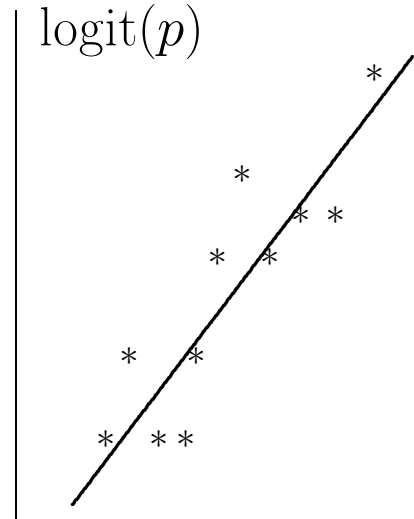
LINEAR VS. LOGISTIC MODELLING

Linear regression



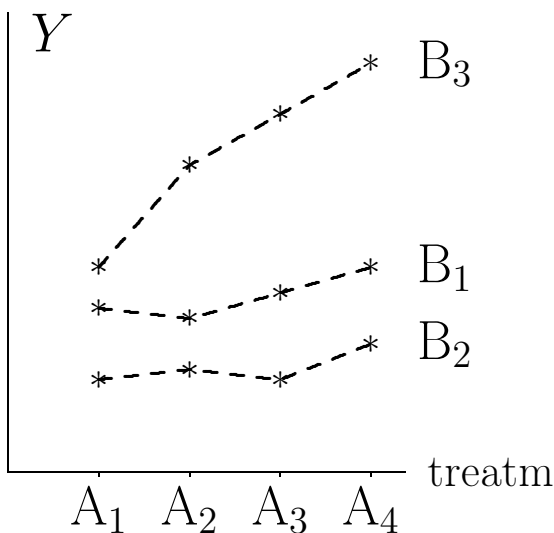
Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
 where ε_i 's $\sim N(0, \sigma^2)$

Logistic regression



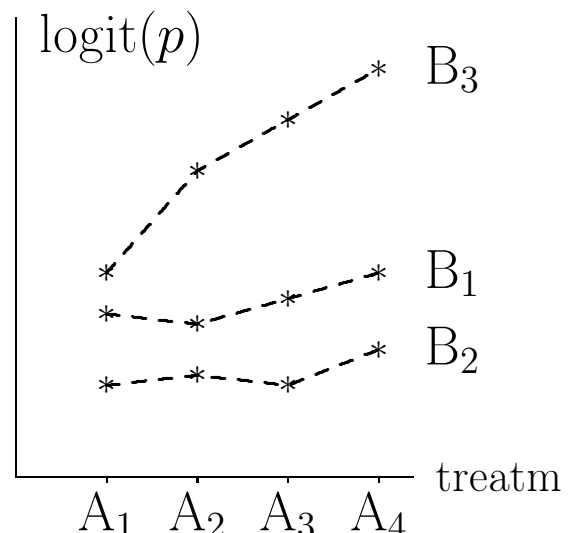
Model: $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$
 where $p_i = P(Y_i = 1)$

Factorial design



Model:
 $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
 where ε_{ij} 's $\sim N(0, \sigma^2)$

Logistic factorial design



Model:
 $\text{logit}(p_{ij}) = \mu + \alpha_i + \beta_j$
 where $p_{ij} = P(Y_{ij} = 1)$