

## Index of Lecture 4a

| Page | Title   |
|------|---|
| 1    | Practical information                         |
| 2    | Dataset <code>nocardia</code> (VER)           |
| 3    | Case-control study and logistic regression    |
| 4    | Two-way table and logistic regression         |
| 5    | Odds-ratio in multiple logistic regression    |
| 6    | Statistical inference for logistic regression |
| 7    | Likelihood function                           |
| 8    | Maximum likelihood estimation                 |
| 9    | Likelihood-based inference                    |
| 10   | LR-tests in logistic regression               |

## PRACTICAL INFORMATION

Today's session includes:

- continuation of review for logistic regression,
- continuation of prediction using `margins` command, covering logistic regression (material from Lecture 3a),
- VER Exercise 15 (model building, `pig_adg` data): discussion/review planned.

Today's lecture:

- multiple logistic regression:
  - \* more on interpretation of coefficients,
  - \* statistical inference,
- new topics for logistic regression:
  - \* “likelihood”, in particular its use for maximum likelihood estimation and likelihood-ratio tests,
  - \* relation to case-control studies.
- textbook (VER/MER) reading:  
16.1–8 fully covered after this lecture.

|                        |
|------------------------|
| DATASET NOCARDIA (VER) |
|------------------------|

- subset of a real dataset on Nocardia mastitis in Nova Scotia dairy herds collected in 1989,
- case-control study design with 54 case and control herds,
  - \* 54 (all!) case herds included in study,
  - \* 54 non-case herds randomly selected from population of herds,
- purpose of study: evaluate the association between various exposure variables and case-control status of the herds.

| Variable | Description                                   | Values                                     |
|----------|---|--|
| id       | farm id                                       | (nominal)                                  |
| casecont | herd status for Nocardia mastitis             | 0/1 (control/case)                         |
| dcpct    | percent of dry cows treated                   | 0–100 %                                    |
| dneo     | use of dry-cow product containing neomycin    | 0/1 (no/yes)                               |
| dclox    | use of dry-cow product containing cloxacillin | 0/1 (no/yes)                               |
| dbarn    | barn type for dry cows                        | 1 = freestall<br>2 = tiestall<br>3 = other |
| numcow   | number of cows milked                         | 16–190                                     |
| ...      | ...   | ...  |

## CASE-CONTROL STUDY AND LOGISTIC REGRESSION

Nocardia data: outcome=dclox(!), explanatory=casecont:

| casecont | dclox |    | Total | Prop. exposed |
|----------|-------|----|-------|---------------|
|          | 1     | 0  |       |               |
| 1        | 8     | 46 | 54    | 0.148         |
| 0        | 19    | 35 | 54    | 0.352         |
| Total    | 27    | 81 | 108   |               |

Statistical model:

two binomial distributions  $\text{Bin}(54, p_1)$  and  $\text{Bin}(54, p_0)$  for case and control populations, respectively.

Odds-ratio (OR) for comparison of exposure in case and control populations, or for comparison of risk of exposed and non-exposed herds

$$\text{OR} = \text{odds}(0.148) / \text{odds}(0.352) = 8 \cdot 35 / (19 \cdot 46) = 0.320,$$

$\Rightarrow$  cloxacillin treatment protective against Nocardia mast.

Logistic regression:  $\text{logit}(p_i) = \beta_0 + \beta_1 \text{dclox}_i$ , gives:

$$\hat{\beta}_1 = -1.138 = \ln(0.320) = \ln(\text{OR}),$$

$$\hat{\beta}_0 = 0.273 \quad \leftarrow \text{meaningless/useless!}$$

Same OR from  $2 \times 2$ -table analysis and logistic regression.<sup>1</sup>

---

<sup>1</sup> A statistical result states that under the assumption that the sampling proportions in case and control populations are independent of the predictors, a case-control study can be analysed by the same logistic regression model as if the design had been a cohort study, except that the estimated intercept is meaningless.

## TWO-WAY TABLE AND LOGISTIC REGRESSION

Nocardia data: outcome=dbarn, explanatory=casecont:

| casecont | dbarn       |            |         | Total |
|----------|-------------|------------|---------|-------|
|          | 1:freestall | 2:tiestall | 3:other |       |
| 1        | 22          | 29         | 3       | 54    |
| 0        | 13          | 38         | 3       | 54    |
| Total    | 35          | 67         | 6       | 108   |

Simple statistical model and analysis:

comparison of case and control populations of herds with respect to distribution of barn types  $\sim \chi^2$ -test.

Multiple odds-ratios by focusing only on two dbarn categories at a time, e.g. involving freestall barn type:

$$\text{tiestall vs. freestall : OR} = 13 \cdot 29 / (22 \cdot 38) = 0.451,$$

$$\text{other vs. freestall : OR} = 13 \cdot 3 / (22 \cdot 3) = 0.591.$$

Logistic regression with dbarn as a categorical predictor and freestall as the reference category:

$$\text{logit}(p_i) = \beta_0 + \beta_1(\text{dbarn}=2)_i + \beta_2(\text{dbarn}=3)_i,$$

gives the estimates:

$$\hat{\beta}_0 = 0.526 \quad \leftarrow \text{meaningless/useless!}$$

$$\hat{\beta}_1 = -0.796 = \ln(0.451) = \ln(\text{OR}) \quad \text{for tiestall vs. freestall,}$$

$$\hat{\beta}_2 = -0.526 = \ln(0.591) = \ln(\text{OR}) \quad \text{for other vs. freestall.}$$

Tests of dbarn effect give  $P$ -values around 0.17 with both approaches (i.e.,  $\chi^2$ -test and logistic regression).

## ODDS-RATIO IN MULTIPLE LOGISTIC REGRESSION

### Basic fact:

In an additive<sup>2</sup> multiple logistic regression model,

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

the odds-ratio for an increase of  $a$  units by predictor  $x_1$  (say) is given by  $OR = e^{\beta_1 a}$ , *no matter the values of all other predictors* (as long as they are equal in the two scenarios whose probabilities we compare).

Illustration by Nocardia data and the model,

$$\text{logit}(p_i) = \beta_0 + \beta_1 \text{dneo}_i + \beta_2 \text{dclox}_i + \beta_3 \text{dcpct}_i,$$

$$\text{logit}(\hat{p}) = -2.98 + 2.21 \text{dneo} - 1.41 \text{dclox} + 0.023 \text{dcpct},$$

- compare herds with  $\text{dneo} = 1$  and  $\text{dneo} = 0$  (i.e.,  $a = 1$ ) and any (but same!) values of other predictors,
- compute predicted probabilities on logit scale,

$$\text{dneo} = 1 : \text{logit}(\hat{p}_1) = -2.98 + 2.21 - 1.41 \text{dclox} + 0.023 \text{dcpct},$$

$$\text{dneo} = 0 : \text{logit}(\hat{p}_0) = -2.98 + 0 - 1.41 \text{dclox} + 0.023 \text{dcpct},$$

- convert logit probabilities to odds,

$$\begin{aligned} \text{dneo} = 1 : \text{odds}(\hat{p}_1) &= e^{-2.98+2.21-1.41 \text{dclox}+0.023 \text{dcpct}}, \\ &= e^{-2.98} e^{2.21} e^{-1.41 \text{dclox}} e^{0.023 \text{dcpct}}, \end{aligned}$$

$$\text{dneo} = 0 : \text{odds}(\hat{p}_0) = e^{-2.98-1.41 \text{dclox}+0.023 \text{dcpct}},$$

- compute odds-ratio,

$$OR = \frac{\text{odds}(\hat{p}_1)}{\text{odds}(\hat{p}_0)} = \frac{e^{-2.98} e^{2.21} e^{1.41 \text{dclox}} e^{0.023 \text{dcpct}}}{e^{-2.98} e^{1.41 \text{dclox}} e^{0.023 \text{dcpct}}} = e^{2.21} = 9.14,$$

- recall that the `logistic` command always gives the OR for a one unit change (possibly inappropriate for continuous predictors).

---

<sup>2</sup> Assuming that the predictors  $x_1, \dots, x_k$  do *not* represent interaction or polynomial regression terms.

## STATISTICAL INFERENCE FOR LOGISTIC REGRESSION

Estimation by maximum-likelihood method (next slides).<sup>3</sup>

Wald confidence intervals and tests:

based on estimates  $\hat{\beta}_1$  (say) and standard errors  $SE(\hat{\beta}_1)$ :

- approximate  $(1-\alpha)$  confidence interval using a standard normal reference distribution, e.g.

$$95\% \text{ CI for } \beta_1 : \hat{\beta}_1 \pm z^* SE(\hat{\beta}_1), \quad z^* = 1.96 \quad (z_{1-\alpha/2}),$$

- approximate  $z$ -tests of simple hypotheses, e.g. of  $H_0 : \beta_1 = b$  vs.  $H_a : \beta_1 \neq b$  by the  $z$ -statistic

$$z = (\hat{\beta}_1 - b) / SE(\hat{\beta}_1) \approx N(0, 1) \quad \text{under } H_0,$$

- multiple Wald tests possible as well (using software),
- beware that Wald procedures do not work when either estimates or their standard errors are “extreme”.<sup>4</sup>

Likelihood-based inference: likelihood-ratio test (next slides) and (profile) likelihood confidence interval,

- also approximate, but generally considered more precise than Wald procedures although difference often small,
- confidence intervals available in Stata using `logprof` or `pllf` add-on commands (but not core part of course).

---

<sup>3</sup> Quasi-likelihood estimation is the name used for the procedure when the model contains “overdispersion” or “underdispersion” (to be discussed in a later lecture).

<sup>4</sup> Occurs with perfectly fitted categories, or more generally, (quasi-)separation of parameters, see e.g. Heinze & Schemper (2002), *Statist. Med.* **21**, 2409-2419.

## LIKELIHOOD FUNCTION

Simple example: binomial distribution,

- consider one group of 10 mice subjected to a particular dose, and denote by
  - \*  $Y$  the no. of dead mice, assume we observed  $Y = 3$ ,
  - \*  $p$  the probability of mice dying at this dose,

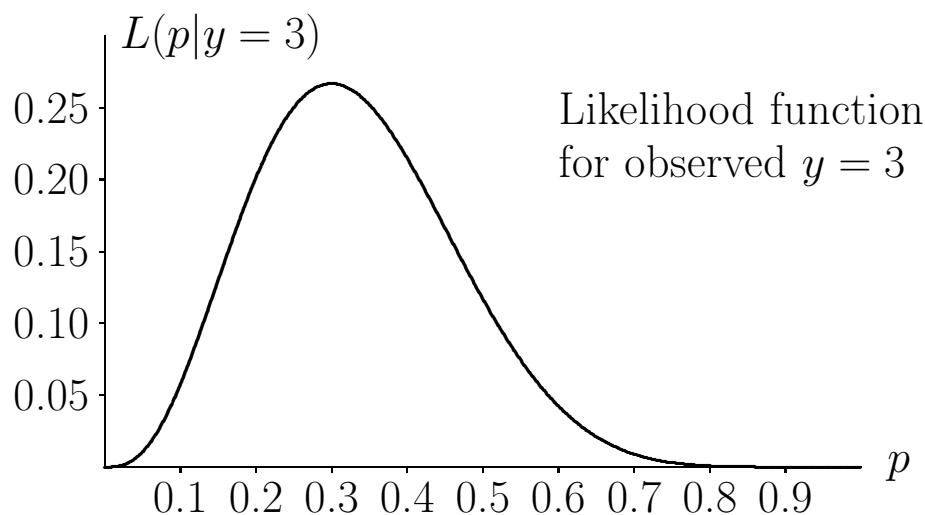
- the prob. distribution of  $Y$  is  $\text{Bin}(10, p)$  with values

$$p(y) = \binom{10}{y} p^y (1-p)^{10-y}, \quad y = 0, \dots, 10,$$

and  $p(y)$  is the probability of observing  $y$  dead mice,

- the likelihood function is this function viewed as a function of the unknown parameter  $p$  and taking the observed data ( $y$ ) as fixed,

$$L(p) = L(p|y) = \binom{10}{y} p^y (1-p)^{10-y}, \quad 0 \leq p \leq 1,$$



In general, the likelihood function is the probability (in continuous models: density value) of the observed data viewed as a function of the unknown parameters.

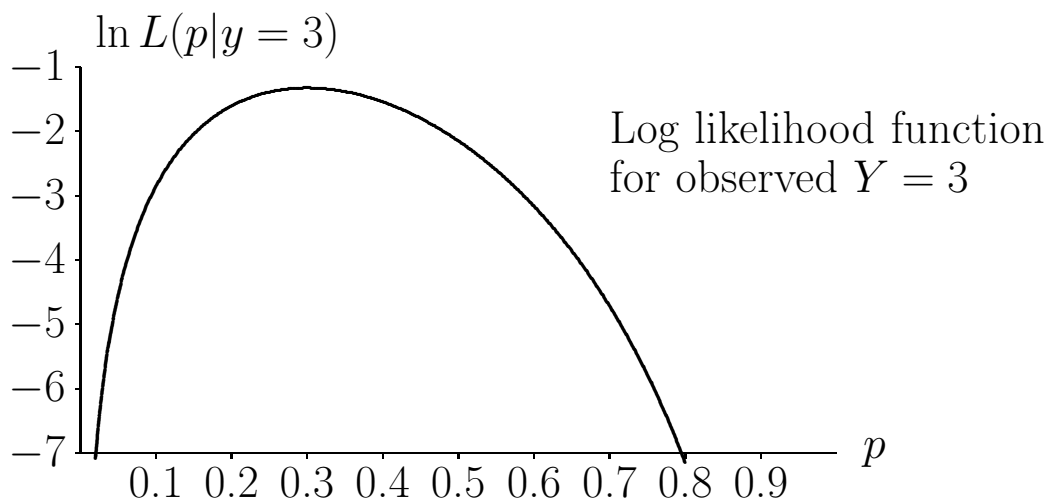
## MAXIMUM LIKELIHOOD ESTIMATION

Idea: choose as our estimate the value of the parameter which *maximizes* the likelihood function = the maximum likelihood estimate (MLE),

- intuitively plausible (“make the data as probable as possible”),
- general procedure applicable to all parametric models,<sup>5</sup>
- easy to compute analytically in many models,
- leads to estimates with good theoretical properties, in particular in large samples.

Computing the MLE in complex models (also logistic reg.):

- iterative procedure: starting value  $\rightarrow$  improved value  $\rightarrow$  improved value  $\rightarrow \dots \rightarrow$  no further improvement possible (convergence  $\sim$  maximum found, or failure),
- convenient and common to work with  $\ln L$  instead of  $L$ .



---

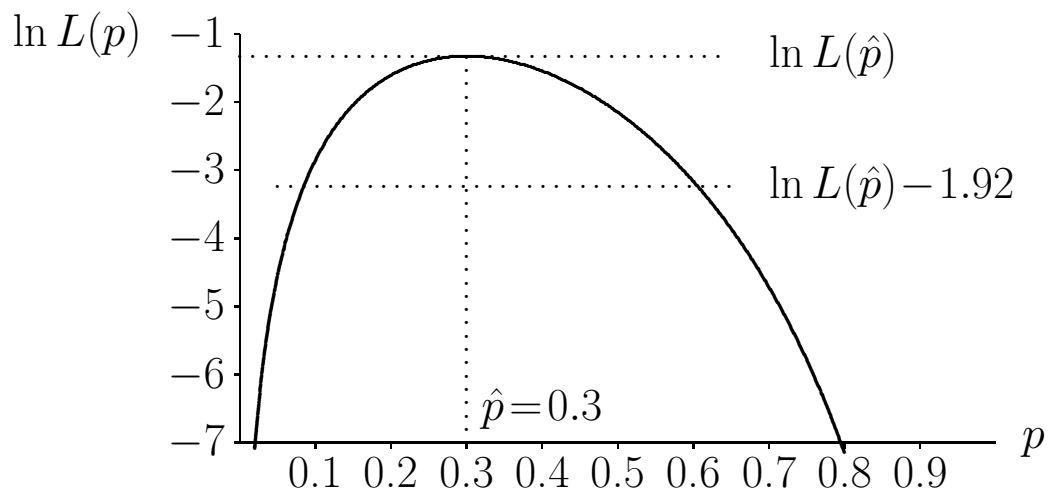
<sup>5</sup> In linear (regression) models, least-squares estimates are also MLEs.

## LIKELIHOOD-BASED INFERENCE

Idea: use likelihood function as our “evidence” against specific parameter values (or hypotheses),

- $p$ 's with low likelihood seem little plausible (observed data unlikely),
- $p$ 's with likelihood close to optimal seem plausible,
- statistical theory (based on large samples):
  - \* differences in  $2 \ln L \approx \chi^2$ -distribution,
  - \* likelihood-ratio (LR) test, denoted  $G^2$ , for comparing “full” and “reduced” model (submodel  $\sim H_0$ ),
 
$$G^2 = 2(\ln \hat{L}_{\text{full}} - \ln \hat{L}_{\text{red}}) \approx \chi^2(\text{df}) \quad \text{under } H_0,$$
 where df = difference in no. of parameters between the models, and  $\hat{L}$  = optimal likelihood values.

Example: testing  $H_0 : p = p_0$  ( $p_0$  known)



Interpretation: no evidence against  $p_0$ -values in the range  $(0.08, 0.61)$  at the 5% significance level (so it's a 95% CI).

## LR-TESTS IN LOGISTIC REGRESSION

Example I: comparing models for `nocardia` data,

| Model                         | $2 \ln L$ | params | change prev. model |    |        |
|-------------------------------|-----------|--------|--------------------|----|--------|
|                               |           |        | $G^2$              | df | $P$    |
| <code>dneo,dclox,dcpct</code> | -107.99   | 4      | —                  | —  | —      |
| <code>dcpct</code>            | -138.15   | 2      | 30.16              | 2  | <0.001 |
| (intercept only)              | -149.72   | 1      | 11.57              | 1  | 0.001  |

- both model reductions strongly significant,
- test against null model directly in Stata listing.

Example II: goodness-of-fit test for `mice` data,

| Model                         | $2 \ln L$ | params | change prev. model |    |      |
|-------------------------------|-----------|--------|--------------------|----|------|
|                               |           |        | $G^2$              | df | $P$  |
| <code>dose categorical</code> | -117.64   | 12     | —                  | —  | —    |
| <code>dose continuous</code>  | -127.89   | 2      | 10.25              | 10 | 0.42 |

- no evidence against linear relation of dose (logit scale),
- categorical model has one “perfectly fitted category” (dose = 0.1413)  $\Rightarrow$  care is needed in Stata.

(Technical) Deviance:

= difference in  $2 \ln L$  between actual and “saturated” model,<sup>6</sup>

- can be used to compute  $G^2$  for LR-test instead of  $2 \ln L$ ,<sup>7</sup>
- can be used for goodness-of-fit test for grouped data if “saturated model” defined properly (not recommended<sup>8</sup>).

<sup>6</sup> “Saturated” model: one parameter for every observation or every distinct group of predictor values; different uses exist within and between softwares. . . .

<sup>7</sup> In my view, there is no real advantage in using the deviance instead of  $2 \ln L$ .

<sup>8</sup> It is safer to compute the LR-test from the two model fits.