

Index of Lecture 4a: Inference for logistic regression

Page	Title
1	Practical information
2	Dataset nocardia (VER)
3	Case-control study and logistic regression
4	Two-way table and logistic regression
5	Odds-ratios 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
11	Predictions in logistic regression
12	Scale-dependence of predictions with averaging/weighting

PRACTICAL INFORMATION

News/Schedule:

- lab session on Monday (Jan 30), focused on logistic regression,
- deadline for first home assignment getting close (Feb 3, both courses).

Today's session includes:

- continuation of prediction using margins command (both linear and logistic regression),
- review/discussion of Model-building Exercise (VER 15),
- first material on 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 historical exposure variables and the subsequent case-control status of the herds.

Variable	Description	Values
id	farm id	(nominal; 108 values)
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 2 = tiestall 3 = other
numcow	number of cows milked	16 – 190
...

CASE-CONTROL STUDY AND LOGISTIC REGRESSION

Nocardia data:

outcome=dclox(!),

explanatory=casecont:

casecont	dclox		Total	Proportion exposed
	1	0		
1	8	46	54	0.148
0	19	35	54	0.352
Total	27	81	108	

Statistical model: 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

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

⇒ cloxacillin treatment seems protective against Nocardia mastitis.

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

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

$$\hat{\beta}_0 = 0.273 \quad \leftarrow \text{meaningless for the population!}$$

Bottom line: also here the **same OR** from 2×2 -table analysis and logistic regression.¹

¹ A statistical result states that if the sampling proportions in the 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 does not refer to the population; Breslow & Day (1980), Section 6.3.

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 by a Pearson X^2 (or χ^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:

$$\hat{\beta}_0 = 0.526 \quad \leftarrow \text{meaningless for the population!}$$

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

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

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

ODDS-RATIOS IN MULTIPLE LOGISTIC REGRESSION

Basic fact: in an additive² 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 units³ by predictor x_1 (say) is given by $\text{OR} = e^{\beta_1 a}$,
no matter the values of all other predictors.⁴

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 same values of dclox , dcpct ,

○ **compute logit scale** $\text{dneo} = 1 : \text{logit}(\hat{p}_1) = -2.98 + 2.21 - 1.41 \text{dclox} + 0.023 \text{dcpct}$,
predicted probabilities: $\text{dneo} = 0 : \text{logit}(\hat{p}_0) = -2.98 + 0 - 1.41 \text{dclox} + 0.023 \text{dcpct}$,

○ **convert logit probabilities to odds:**

$$\begin{aligned} \text{dneo} = 1 : \text{odds}(\hat{p}_1) &= e^{-2.98+2.21-1.41 \text{dclox}+0.023 \text{dcpct}} = e^{-2.98} e^{2.21} e^{-1.41 \text{dclox}} e^{0.023 \text{dcpct}}, \\ \text{dneo} = 0 : \text{odds}(\hat{p}_0) &= e^{-2.98-1.41 \text{dclox}+0.023 \text{dcpct}}, \end{aligned}$$

○ **compute odds-ratio:** $\text{OR} = \frac{\text{odds}(\hat{p}_1)}{\text{odds}(\hat{p}_0)} = \frac{e^{-2.98} e^{2.21} e^{-1.41 \text{dclox}} e^{0.023 \text{dcpct}}}{e^{-2.98} e^{-1.41 \text{dclox}} e^{0.023 \text{dcpct}}} = e^{2.21} = 9.14$.

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

³ Recall, the logistic command gives the OR for a 1-unit change (**maybe inappropriate** for continuous predictors).

⁴ However, just as in multiple linear regression, the other predictors must **be the same** in the two scenarios compared.

STATISTICAL INFERENCE FOR LOGISTIC REGRESSION

Estimation by maximum-likelihood method (next slides).⁵

Wald confidence intervals and tests: based on estimates $\hat{\beta}_1$ (say) and standard errors:

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

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

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

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

- **multiple Wald tests** possible as well (using software),
- beware that Wald procedures **do not work** when either the estimates or their standard errors are “extreme”.⁶

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

- also approximate, but generally considered **more precise** than Wald procedures, even if the difference is often small,
- **confidence intervals available in Stata using logprof or pllf add-on commands.**⁷

⁵ Quasi-likelihood estimation is the name used for the procedure when the model contains “overdispersion” or “underdispersion” (to be discussed in a later lecture).

⁶ Occurs with perfectly fitted categories, or more generally, (quasi-)separation of parameters, see e.g. Heinze & Schemper (2002), *Statistics in Medicine* 21, 2409-2419; for an example, see slide 4aL-10.

⁷ Likelihood-based confidence intervals are less commonly used, and not in the core part of the course.

LIKELIHOOD FUNCTION

Simple example: binomial distribution,

- consider one group of 10 mice subjected to a particular dose, and denote by
 - * Y the number of dead mice, assume we observed $Y = 3$,
 - * p the probability of mice dying at this dose,

- the probability distribution of Y is **binomial (10, p)** with values

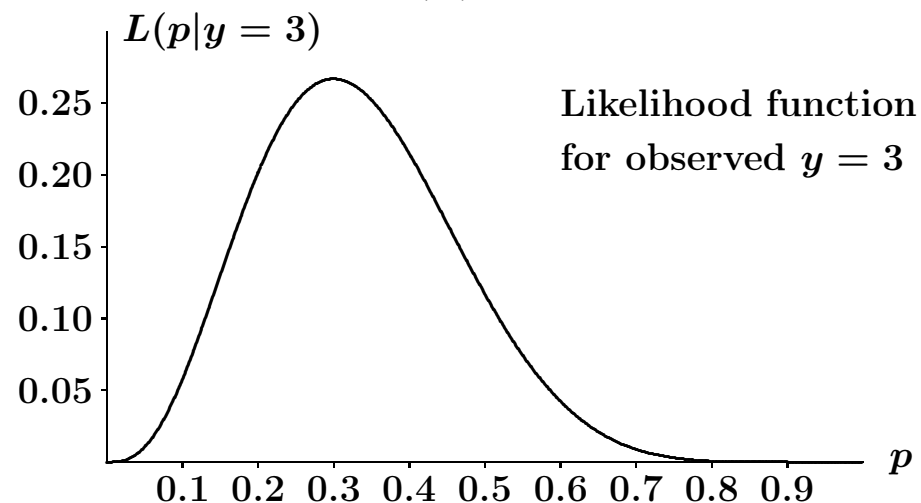
$$p(y) = \binom{10}{y} p^y (1 - p)^{10-y}, \quad y = 0, \dots, 10,$$

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

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

$$\begin{aligned} L(p) &= L(p|y) \\ &= \binom{10}{y} p^y (1 - p)^{10-y}, \end{aligned}$$

for $0 \leq p \leq 1$.



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.

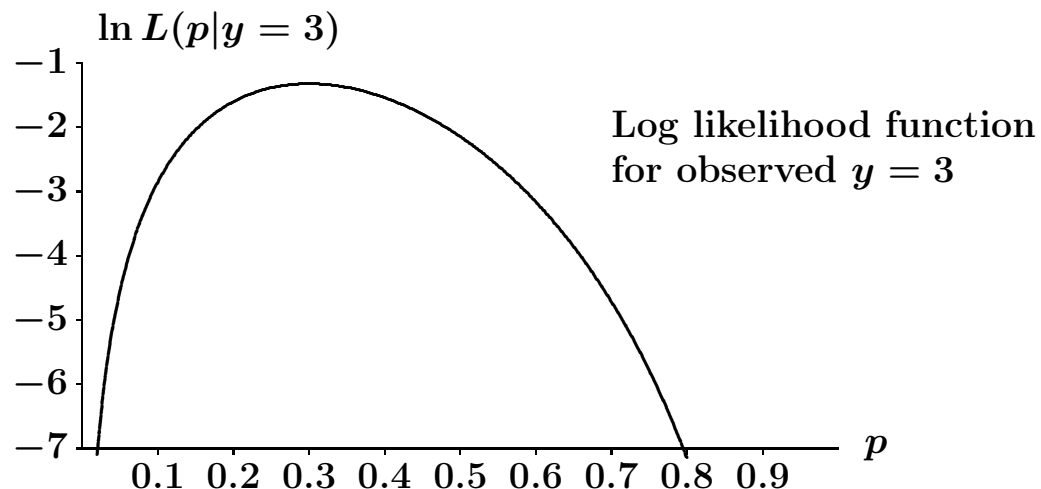
MAXIMUM LIKELIHOOD ESTIMATION

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,⁸
- **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 (including logistic regression):

- **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 .



⁸ In linear (regression) models, least-squares estimates are also MLEs.

LIKELIHOOD-BASED INFERENCE

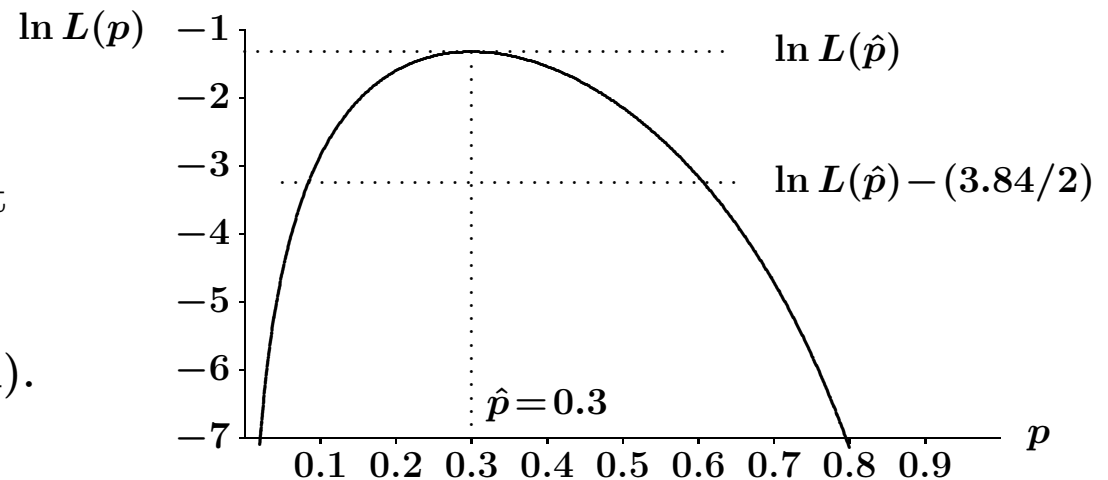
Idea: use likelihood function as our “evidence” against specific scenarios/hypotheses,

- scenarios with low likelihood seem little plausible (observed data unlikely),
- scenarios with likelihood close to optimal seem plausible,
- **statistical theory** (based on large samples):
 - * differences in $2 \ln L \approx \chi^2$ -distributions,
 - * **likelihood-ratio** (LR) test, denoted G^2 , for comparing “full” and “reduced” models (reduced model $\sim H_0$), $G^2 = 2(\ln \hat{L}_{\text{full}} - \ln \hat{L}_{\text{red}}) \approx \chi^2(\text{df})$ under H_0 ,

where df = difference in number of parameters between the models, and \hat{L} = optimal likelihood values.

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

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



LR-TESTS IN LOGISTIC REGRESSION

Example I: comparing models for nocardia data,

- both model reductions strongly significant.⁹

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

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

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

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

(Technical) **Deviance** = difference in $2 \ln L$ between actual and “saturated” model,¹⁰

- can be used to compute G^2 for LR-test instead of $2 \ln L$,¹¹
- can be used for **goodness-of-fit test** for grouped data if “saturated model” defined properly (not recommended¹²).

⁹ Note that the test against the **null** (intercept only) model is shown in the Stata output.

¹⁰ “Saturated” model: one parameter for every observation or every distinct group of predictor values; different usages exist within and between softwares....

¹¹ In my view, there is no real advantage in using the deviance instead of $2 \ln L$.

¹² It is safer to compute the LR-test from the two model fits.

PREDICTIONS IN LOGISTIC REGRESSION

Predictions/presentation of effects on probability scale:

- easier to understand probabilities than OR's,
- more complex behaviour than on logit scale due to non-linear backtransformation.

Illustration for Nocardia data and effect of dcpct:

$$\text{logit}(\hat{p}) = -2.984 + 0.023 \text{ dcpct} + 2.212 \text{ dneo} - 1.412 \text{ dclox},$$

- OR for 1% “change” in dcpct = $e^{0.023} = 1.023$,
- OR for 10% “change” in dcpct = $e^{0.023 \cdot 10} = e^{0.23} = 1.25$.

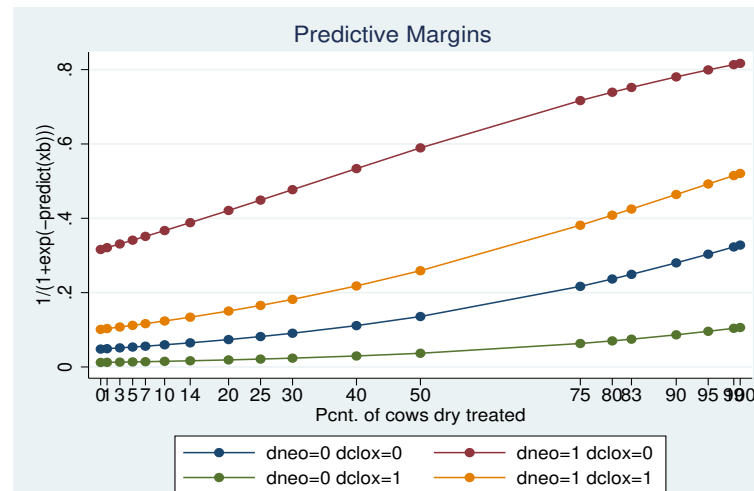
or on probability scale¹³

$$\begin{aligned} \hat{p} &= \text{logit}^{-1}(\text{logit}(\hat{p})) \\ &= 1/(1 + e^{-\text{logit}(\hat{p})}), \end{aligned}$$

can be plotted against dcpct values in a suitable range, for fixed values of dneo,dclox:

Computing the predictions:

- for actual or new observations,
- **Stata**: margins and marginsplot commands.¹⁴



non-linear
relation!

non-parallel
curves!

¹³ For demonstration purposes only; in a case-control design predicted probabilities **do not make sense**, because the proportion of cases and controls is controlled.

¹⁴ Similar to linear models, but see next page for discussion of averaging or weighting.

SCALE-DEPENDENCE OF PREDICTIONS WITH AVERAGING/WEIGHTING

Main message: in models involving transformations (e.g. logistic regression), any averaging of predictions involves a choice of scale:

- **different results** (even after transformation to same scale) and **interpretations**.

Example: consider again the logistic “predictive” equation

$$\text{logit}(\hat{p}) = -2.984 + 0.023 \text{ dcpct} + 2.212 \text{ dneo} - 1.412 \text{ dclox},$$

and the **predictions**

for $\text{dcpct} = 0$:

dneo	dclox	$\text{logit}(\hat{p})$	\hat{p}
0	0	-2.984	0.0481
0	1	-4.397	0.0123
1	0	-0.772	0.316
1	1	-2.184	0.101
weighted*		-1.821	0.198
(transformed)		0.139	-1.399

* based on data counts for the 4 categories of (dneo,dclox): 22, 12, 59, 15

Interpretations (of resulting values on probability scale):

- \hat{p} (averaged): \sim averaged probabilities, might correspond to a population (if data are representative of such),
- $\text{logit}(\hat{p})$ (averaged and then transformed): simpler, because backtransformation of a “natural” (modelled) value from the additive scale.