Index of Lecture 4a: Inference for logistic regression

Practical Information

Major news:

- the first of two asynchronous lectures this week,
- Friday's session will offer review and discussion of logistic regression exercises; you may start to work on VER16:1,
- also another in-person lab session planned for Monday, February 7 (maybe again move to 9-12?).

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,
- also continuation of prediction using `margins` command (logistic regression),
- 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:

	dclox			
casecont	1	0	Total	Proportion exposed
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

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

$\Rightarrow$ cloxacillin treatment seems protective against Nocardia mastitis.

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

$$\begin{aligned} \hat{\beta}_1 &= -1.138 = \ln(0.320) = \ln(\text{OR}), \\ \hat{\beta}_0 &= 0.273 \quad \leftarrow \text{meaningless for the population!} \end{aligned}$$

Bottom line: also here the same OR from $2\times2-$table analysis and logistic regression.[1]

[1] 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 does not refer to the population.

Two-way Table and Logistic Regression

Nocardia data:
outcome=dbarn,
explanatory=casecont:

	dbarn			
casecont	1:freestall	2:tiestall	3:other	Total
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:

$$\text{tiestall vs. freestall}: \ \text{OR} = 13 \cdot 29/(22 \cdot 38) = 0.451,$$
$$\text{other vs. freestall}: \ \text{OR} = 13 \cdot 3/(22 \cdot 3) = 0.591.$$

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

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

gives the estimates:

$$\hat{\beta}_0 = 0.526 \quad \leftarrow \text{meaningless for the population!}$$
$$\hat{\beta}_1 = -0.796 = \ln(0.451) = \ln(\text{OR}) \ \text{ for tiestall vs. freestall,}$$
$$\hat{\beta}_2 = -0.526 = \ln(0.591) = \ln(\text{OR}) \ \text{ for other vs. freestall.}$$

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 additive[2] multiple logistic regression model,

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

the odds-ratio for an increase of a units[3] by predictor x_1 (say) is given by $\text{OR} = \text{e}^{\beta_1 a}$, no matter the values of all other predictors.[4]

Illustration by Nocardia data and the model,

$$\begin{aligned} \text{logit}(p_i) &= \beta_0 + \beta_1\,\texttt{dneo}_i + \beta_2\,\texttt{dclox}_i + \beta_3\,\texttt{dcpct}_i, \\ \text{logit}(\hat{p}) &= -2.98 + 2.21\,\texttt{dneo} - 1.41\,\texttt{dclox} + 0.023\,\texttt{dcpct}, \end{aligned}$$

- compare herds with $\texttt{dneo}=1$ and $\texttt{dneo}=0$ (i.e., $a=1$) and same values of $\texttt{dclox}, \texttt{dcpct}$,
- compute logit scale predicted probabilities:

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

- convert logit probabilities to odds:

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

- compute odds-ratio:

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

[2] Assuming that the predictors $x_1, \ldots, x_k$ 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.

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

- approximate z-tests of simple hypotheses, e.g. of $\mathrm{H}_0 : \beta_1 = b$ vs. $\mathrm{H}_a : \beta_1 \neq b$,

$$z = (\hat{\beta}_1 - b)/\mathrm{SE}(\hat{\beta}_1) \ \approx\ \mathrm{N}(0,1) \ \text{ under } \mathrm{H}_0,$$

- 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,

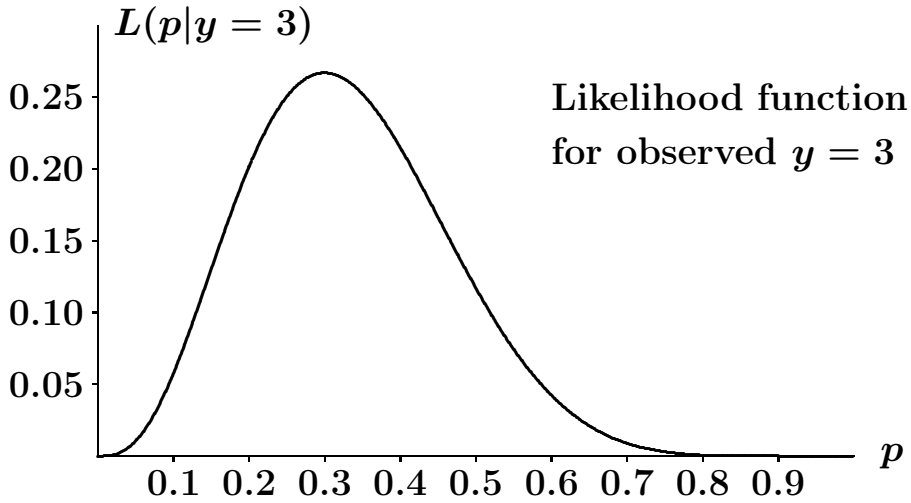
- consider one group of 10 mice subjected to a particular dose, and denote by
  - Y the number of dead mice, assume we observed $Y = 3$,
  - p the probability of mice dying at this dose,
- the probability distribution of Y is binomial $(10, p)$ with values

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

  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.

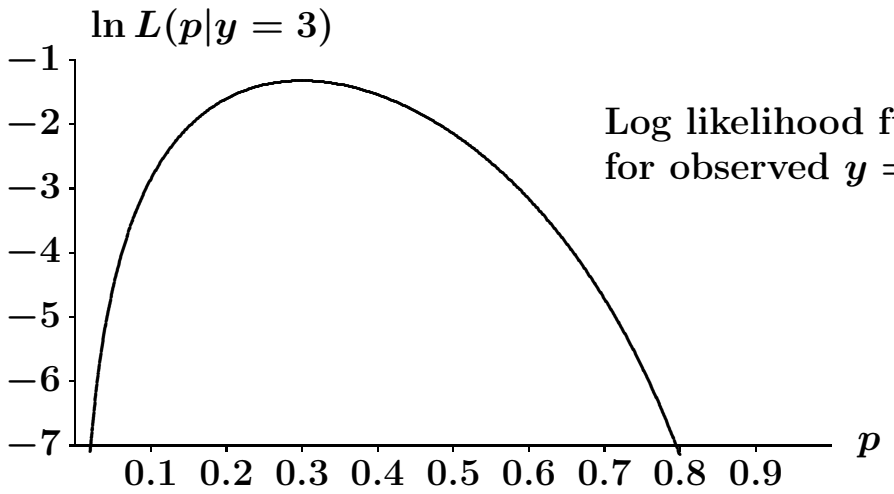
Maximum Likelihood Estimation

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 $\rightarrow \ldots \rightarrow$ 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,

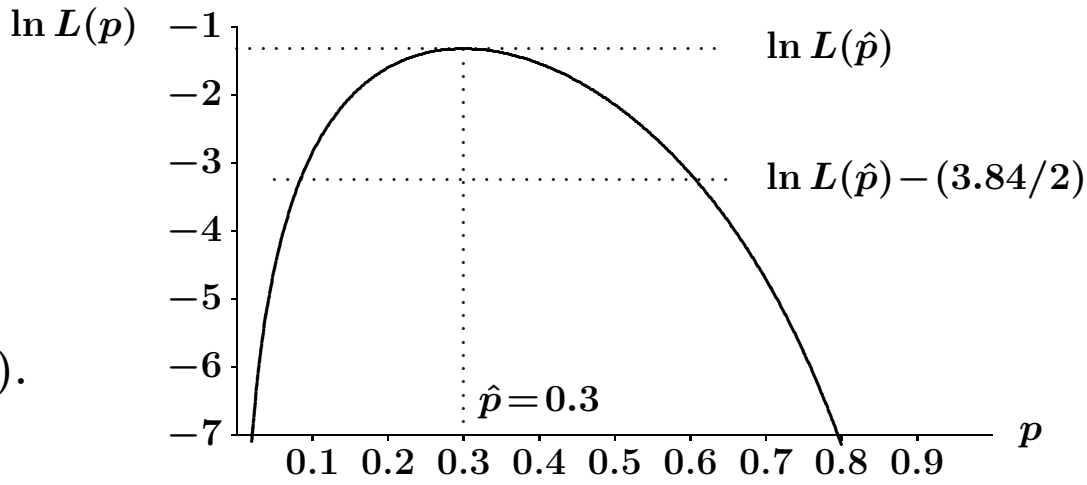
- scenarios with low likelihood seem little plausible (observed data unlikely),
- scenarios with likelihood close to optimal seem plausible,
- statistical theory (based on large samples):
  - differences in $2 \ln L \approx \chi^2$-distributions,
  - likelihood-ratio (LR) test, denoted G^2, for comparing "full" and "reduced" 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 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),

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

- 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,[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 recommended[12]).

[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`:

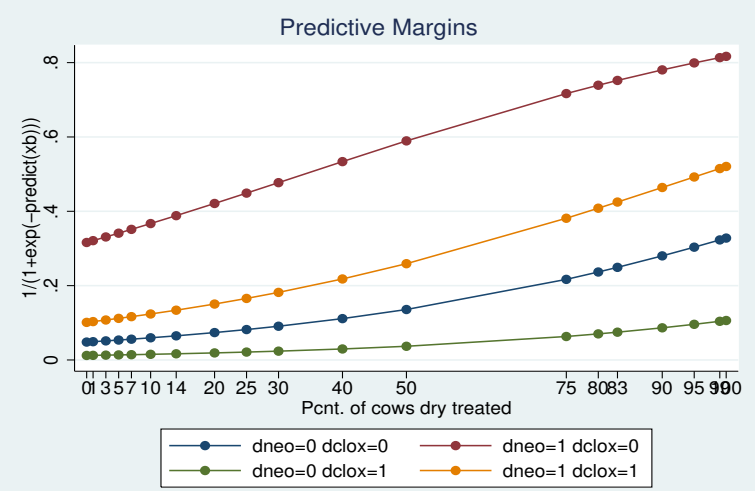
$$\text{logit}(\hat{p}) = -2.984 + 0.023\,\texttt{dcpct} + 2.212\,\texttt{dneo} - 1.412\,\texttt{dclox},$$

- 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 scale[13]

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

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



non-linear relation!

non-parallel curves!

Computing the predictions:

- for actual or new observations,
- Stata: `margins` and `marginsplot` commands.[14]

[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

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

and the predictions for `dcpct` = 0:

dneo	dclox	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),
- logit($\hat{p}$) (averaged and then transformed): simpler, because backtransformation of a "natural" (modelled) value from the additive scale.