

MÀSTER D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA UPC-UB
CURSO 2018/19: EPIDEMIOLOGÍA

Klaus Langohr
klaus.langohr@upc.edu



Topic 10: Logistic Regression

LOGISTIC REGRESSION

- ① THE LOGISTIC REGRESSION MODEL
- ② ESTIMATION OF THE MODEL PARAMETERS
- ③ HYPOTHESIS TESTS
- ④ INTERPRETATION OF THE MODEL PARAMETERS
- ⑤ MODELING INTERACTION
- ⑥ LOGISTIC REGRESSION IN CASE-CONTROL STUDIES
- ⑦ BUILDING A LOGISTIC REGRESSION MODEL
- ⑧ CHECKING GOODNESS OF FIT
- ⑨ LOGISTIC REGRESSION WITH MATCHED DATA

THE LOGISTIC REGRESSION MODEL

Motivation

The Mantel-Haenszel estimator can be used to quantify the association between an outcome (D) and an exposure (E) of interest controlling for a possible confounder C .

However, the Mantel-Haenszel estimator may not be adequate if

- there are several possible confounders, C_1, \dots, C_k ,
- one of the possible confounders is a continuous variable,
- E is a continuous variable.

In these situations, **logistic regression** can be an adequate tool to analyze the degree of association between E and D .

Notation

In what follows, we denote by \mathbf{X} the model's covariate vector, which includes both E and C_1, \dots, C_k .

THE LOGISTIC REGRESSION MODEL (CONT.)

Expression of the logistic regression model

Let Y be the outcome of interest:

$$Y = \begin{cases} 1 & \text{Outcome present } (D), \\ 0 & \text{Outcome absent } (\bar{D}). \end{cases}$$

Since Y is a binary variable, a linear model such as

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_m X_m + \epsilon$$

is not meaningful. Instead, the probability for $Y = 1$ is used as the response variable in a regression model:

$$p_{\mathbf{X}} = P(Y = 1 | \mathbf{X}) = P(Y = 1 | X_1, \dots, X_m).$$

THE LOGISTIC REGRESSION MODEL (CONT.)

Expression of the logistic regression model (cont.)

Since $p_X \in (0, 1)$, logistic regression models the **logit** transformation of p_X , whose range is \mathbb{R} , as a linear combination of the independent variables X_1, \dots, X_m , the **linear predictor**:

$$\text{logit}(p_X) = \ln\left(\frac{p_X}{1 - p_X}\right) = \alpha + \beta_1 X_1 + \dots + \beta_m X_m.$$

This expression is equivalent to

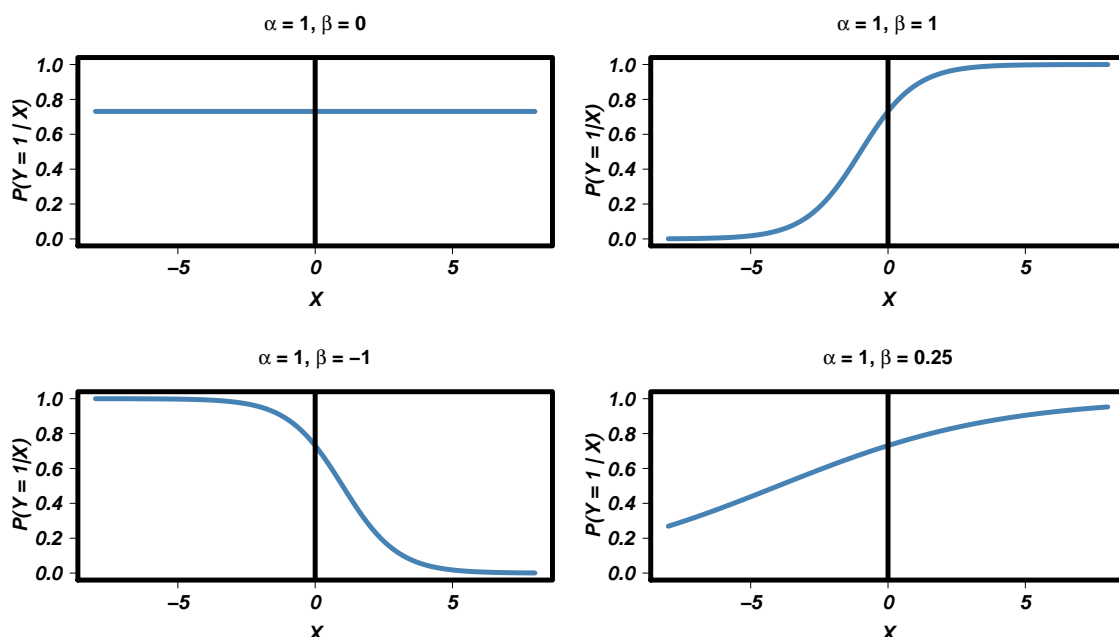
$$p_X = \frac{\exp(\alpha + \beta_1 X_1 + \dots + \beta_m X_m)}{1 + \exp(\alpha + \beta_1 X_1 + \dots + \beta_m X_m)}.$$

This implies that X_i is a risk factor (protective factor) for Y , if $\beta_i > 0$ ($\beta_i < 0$). Y and X_i are independent, if $\beta_i = 0$.

THE LOGISTIC REGRESSION MODEL (CONT.)

The logistic curve

The logistic regression model assumes that the relation between $P(Y = 1)$ and a continuous variable X is modeled by a logistic curve:



THE LOGISTIC REGRESSION MODEL (CONT.)

Categorical independent variables (Factors)

Dummy coding is used when categorical variables are included in a regression model. Dummy codes are a series of numbers assigned to indicate group membership in any mutually exclusive and exhaustive category.

For example, if one of the regressors, X_k say, is a categorical variable with s levels, $s - 1$ dummy variables are included in the model:

$$X_{k_1} = \begin{cases} 1 & X_k = 2 \\ 0 & \text{otherwise} \end{cases}, \dots, X_{k_{s-1}} = \begin{cases} 1 & X_k = s \\ 0 & \text{otherwise} \end{cases}.$$

Any of the s levels can be chosen as the reference category. However, if X_k is an ordinal variable, for ease of model interpretation, it is preferable to choose $X_k = 1$ or $X_k = s$ as reference level, as long as **the number of observations is not too small**.

ESTIMATION OF THE MODEL PARAMETERS

Maximum likelihood estimation of the parameters

The model parameters can be estimated using maximum likelihood estimation: given a sample of independent observations, (y_i, \mathbf{x}_i) , $i = 1, \dots, n$, the expression of the likelihood function is the following:

$$\begin{aligned} L(\alpha, \beta | Y, \mathbf{X}) &= \prod_{i=1}^n P(Y = y_i | \mathbf{x}_i) f(\mathbf{x}_i) \propto \prod_{i=1}^n P(Y = y_i | \mathbf{x}_i) \\ &= \prod_{i=1}^n P(Y = 1 | \mathbf{x}_i)^{\delta_i} P(Y = 0 | \mathbf{x}_i)^{1-\delta_i} \\ &= \prod_{i=1}^n \frac{\exp(\alpha + \beta' \mathbf{x}_i)^{\delta_i}}{1 + \exp(\alpha + \beta' \mathbf{x}_i)}, \end{aligned}$$

where $\beta = (\beta_1, \dots, \beta_m)'$ and $\delta_i = 1$ if $Y_i = 1$ and zero otherwise.

In R, the functions `glm` (package `stats`) and `lrm` (`rms`; Harrell 2019) can be used to fit a logistic regression model.

ESTIMATION OF THE MODEL PARAMETERS (CONT.)

Confidence intervals for the model parameters

According to the properties of the maximum likelihood estimators, under the large sample condition:

$$\hat{\theta}_{\text{ML}} \stackrel{\text{asym.}}{\sim} \mathcal{N}(\theta, \text{Var}(\hat{\theta}_{\text{ML}})),$$

where θ stands for any of the model parameters $\alpha, \beta_1, \dots, \beta_m$.

The variance $\text{Var}(\hat{\theta}_{\text{ML}})$ is the corresponding element of the diagonal of the inverse of the Fisher information matrix and can be estimated using the observed Fisher information matrix.

Hence, the confidence interval for θ is:

$$\text{CI}(\theta; 1 - \alpha) = \hat{\theta}_{\text{ML}} \mp z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\hat{\theta}_{\text{ML}})},$$

where $z_{1-\alpha/2}$ is the $(1 - \frac{\alpha}{2})$ -quantile of the standard normal distribution.

HYPOTHESIS TESTS

The Wald test

The **Wald test** can be used to test whether a specific covariate X_k is associated with Y in presence of the remaining covariates. The corresponding hypothesis is

$$H_0: \beta_k = 0 \quad \text{vs.} \quad H_1: \beta_k \neq 0.$$

This test takes advantage of the MLE's properties shown above using either of the following two as test statistic

$$\frac{\hat{\beta}_k}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}} \stackrel{H_0}{\sim} \mathcal{N}(0, 1),$$

$$\frac{\hat{\beta}_k^2}{\widehat{\text{Var}}(\hat{\beta}_k)} \stackrel{H_0}{\sim} \chi_1^2.$$

HYPOTHESIS TESTS (CONT.)

The likelihood ratio test

The **likelihood ratio test** permits to check the joint association of several covariates with Y in a logistic regression model, that is, to test the hypothesis

$$H_0: \beta_1 = \dots = \beta_s = 0 \quad \text{vs.} \quad H_1: \exists j: \beta_j \neq 0.$$

The test statistic is the difference of the **deviances** of the model under the null hypothesis and the full model:

$$\text{Dev}_{H_0} - \text{Dev}_{\text{full}} = -2 \cdot \ln \left(\frac{L(\hat{\alpha}_{H_0}, \hat{\beta}_{H_0} | Y, \mathbf{X})}{L(\hat{\alpha}, \hat{\beta} | Y, \mathbf{X})} \right) \stackrel{H_0}{\sim} \chi_s^2.$$

This test can be used to check whether there is an association between Y and a factor with more than two levels.

INTERPRETATION OF THE MODEL PARAMETERS

Interpretation of the parameters

One of the reasons for the popularity of logistic regression is the fact that the model parameters can be interpreted in terms of the (log) odds ratio. For example, in the paper of Martín-Santos *et al.* (2012), the logistic regression model applied is summarized not only in terms of $\hat{\beta}$, but also in terms of odds ratios.

For what reason?

Let X_k be a binary covariate of a logistic regression model. The odds ratio associated with $X_k = 1$ and adjusted for all other covariates can be expressed as:

$$\text{OR}_{X_k} = \frac{\text{odds}(Y = 1 | X_1, \dots, X_k = 1, \dots, X_m)}{\text{odds}(Y = 1 | X_1, \dots, X_k = 0, \dots, X_m)} = \exp(\beta_k).$$


INTERPRETATION OF THE PARAMETERS (CONT.)

Hence, the estimator of adjusted odds ratio, OR_{X_k} , and the corresponding confidence interval are:

$$\begin{aligned}\widehat{OR}_{X_k} &= \exp(\hat{\beta}_k), \\ \text{CI}(OR_{X_k}; 1 - \alpha) &= \exp\left(\hat{\beta}_k \mp z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}\right) \\ &= \widehat{OR}_{X_k} \cdot \exp\left(\mp z_{1-\alpha/2} \cdot \sqrt{\widehat{\text{Var}}(\hat{\beta}_k)}\right).\end{aligned}$$

If X_k is a continuous variable, for example age, the odds ratio associated with comparing two exposure levels that differ by c units is

$$OR_{X_{k,c}} = \frac{\text{odds}(Y = 1 | X_1, \dots, X_k = x + c, \dots, X_m)}{\text{odds}(Y = 1 | X_1, \dots, X_k = x, \dots, X_m)} = \exp(c \cdot \beta_k).$$

 Which is the confidence interval of $OR_{X_{k,c}}$?

INTERPRETATION OF THE PARAMETERS (CONT.)

Odds ratio associated with two covariates

Let X_1 and X_2 be two binary covariates with model-based odds ratios

$$OR_1 = \exp(\beta_1) \quad \text{and} \quad OR_2 = \exp(\beta_2).$$

The odds ratio associated with comparing the exposure levels

$X_1 = X_2 = 1$ and $X_1 = X_2 = 0$ is

$$OR_{X_1, X_2} = \frac{\text{odds}(Y = 1 | X_1 = 1, X_2 = 1, X_3, \dots, X_m)}{\text{odds}(Y = 1 | X_1 = 0, X_2 = 0, X_3, \dots, X_m)} = \exp(\beta_1 + \beta_2).$$

That is, the odds ratio is the product of OR_1 and OR_2 :

$$OR_{X_1, X_2} = \exp(\beta_1 + \beta_2) = \exp(\beta_1) \exp(\beta_2) = OR_1 \cdot OR_2.$$

 Which is the corresponding confidence interval?

INTERPRETATION OF THE PARAMETERS (CONT.)

The model constant

The interpretation of the model constant (intercept) α is related to the probability of $Y = 1$ in the case of an individual with zero values in all covariates:

$$p_0 = P(Y = 1 | \mathbf{X} = \mathbf{0}) = \frac{\exp(\alpha)}{1 + \exp(\alpha)} \iff \frac{p_0}{1 - p_0} = \exp(\alpha).$$

To provide the model constant with a relevant meaning, continuous covariates may be centered, for example, around their means:

$$X^* = X - \bar{X}.$$

Hence,

$$p_0 = P(Y = 1 | X^* = 0) = P(Y = 1 | X = \bar{X}) = \frac{\exp(\alpha)}{1 + \exp(\alpha)}.$$

Note: The interpretation given is valid in cohort and cross-sectional studies, but **not** in case-control studies.

EXAMPLE WITH CROSS-SECTIONAL DATA

Liver Enzyme Alterations in HCV-Monoinfected and HCV/HIV-Coinfected Patients

Klaus Langohr^{1,2}, Arantza Sanvisens¹, Daniel Fuster¹, Jordi Tor¹, Isabel Serra¹, Celestino Rey-Joly¹, Inmaculada Rivas³ and Roberto Muga^{*,1}

¹Department of Internal Medicine, Hospital Universitari Germans Trias i Pujol, Badalona, Spain

²Human Pharmacology and Clinical Neurosciences Research Unit. Institut Municipal d'Investigació Mèdica / IMIM, Barcelona, Spain

³Municipal Centre for Drug Abuse Treatment (Centro Delta), Badalona, Spain

Logistic regression model among HCV+/HIV– IDUs:

- Y : Elevated ALT (alanine aminotransferase) level
- X_1 : Gender; X_2 : Body mass index; X_3 : Age
- Cross-sectional data

EXAMPLE WITH CROSS-SECTIONAL DATA

```
> lrm(alt_high ~ sexe + age + imc, hivnegdat)
```

	Coef	S.E.	Wald Z	Pr(> Z)
Intercept	-1.3266	1.5612	-0.85	0.3955
sexe=Male	1.8008	0.5308	3.39	0.0007
age	-0.1000	0.0343	-2.91	0.0036
imc	0.1441	0.0656	2.20	0.0280

Interpretation of model parameters:

```
> round(exp(1.8008), 2)
```

```
[1] 6.05
```

Estimated **prevalence odds ratio** comparing male to female IDUs of same age and BMI.

EXAMPLE WITH CROSS-SECTIONAL DATA

```
> lrm(alt_high ~ sexe + age + imc, hivnegdat)
```

	Coef	S.E.	Wald Z	Pr(> Z)
Intercept	-1.3266	1.5612	-0.85	0.3955
sexe=Male	1.8008	0.5308	3.39	0.0007
age	-0.1000	0.0343	-2.91	0.0036
imc	0.1441	0.0656	2.20	0.0280

Interpretation of model parameters:

```
> round(exp(- 1.3266 - 0.1 * 30 + 0.1441 * 22)/  
+ (1 + exp(- 1.3266 - 0.1 * 30 + 0.1441 * 22)), 2)
```

```
[1] 0.24
```

Estimated **prevalence** of elevated ALT of 30 years old female IDU with BMI of 22 kg/m².



Incidence of and risk factors for nodding off at scientific sessions

Kenneth Rockwood, David B. Hogan, Christopher J. Patterson; for the Nodding at Presentations (NAP) Investigators

CMAJ • DEC. 7, 2004; 171 (12)

1443

© 2004 Canadian Medical Association or its licensors

Logistic regression model for **risk** of nodding off at lectures:

- Y: Nod at scientific sessions
- Covariates: Environmental, audiovisual, or speaker-related factors
- Prospective data

EXAMPLE WITH DATA FROM COHORT STUDIES

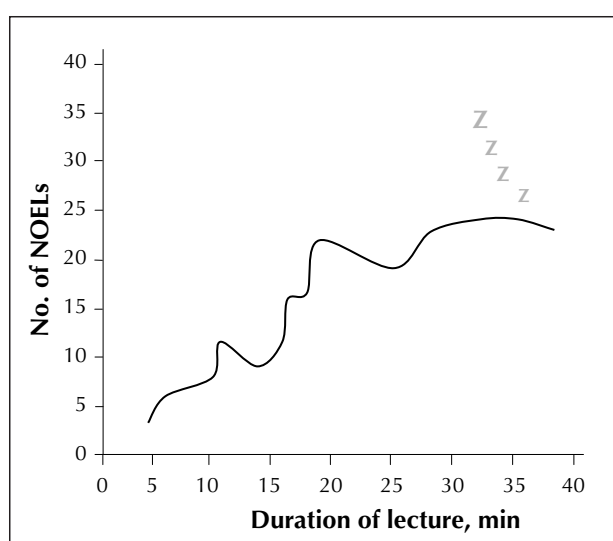


Fig. 1: Special incidence density curve, showing number of nodding-off events per lecture (NOELs) per 100 attendees over length of time of presentation.

Table 1: Risk factors for nodding off at lectures

Factor	Odds ratio (and 95% CI)
Environmental	
Dim lighting	1.6 (0.8–2.5)
Warm room temperature	1.4 (0.9–1.6)
Comfortable seating	1.0 (0.7–1.3)
Audiovisual	
Poor slides	1.8 (1.3–2.0)
Failure to speak into microphone	1.7 (1.3–2.1)
Circadian	
Early morning	1.3 (0.9–1.8)
Post prandial	1.7 (0.9–2.3)
Speaker-related	
Monotonous tone	6.8 (5.4–8.0)
Tweed jacket	2.1 (1.7–3.0)
Losing place in lecture	2.0 (1.5–2.6)

Note: CI = confidence interval.

among attendees on when the speaker's prefatory comments

MODELING INTERACTION

Interaction of two binary covariates

Let X_1 and X_2 be two binary variables in a logistic regression model and suppose we are interested to check whether interaction exists with respect to the probability of disease. The model to be fitted is the following:

$$\text{logit}(P(Y = 1|X_1, X_2)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2. \quad (1)$$

To check for **multiplicative** interaction, one can either calculate $CI(\beta_3; 1 - \alpha)$ or test the hypothesis

$$H_0: \beta_3 = 0 \quad \text{vs.} \quad H_1: \beta_3 \neq 0,$$

by means of the Wald test. In case H_0 is not rejected, multiplicative interaction is not considered statistically significant and the term $\beta_3 X_1 \cdot X_2$ can be removed from Model (1).

Additive interaction can be assessed by estimating the $RERI_{OR}$.

MODELING INTERACTION (CONT.)

By contrast, if $\beta_3 X_1 \cdot X_2$ is kept in the model, the odds ratio associated with one variable varies across the levels of the other.

For example, the OR associated with X_2 is

$$OR_{X_2} = \begin{cases} \exp(\beta_2) & \text{if } X_1 = 0 \\ \exp(\beta_2 + \beta_3) & \text{if } X_1 = 1 \end{cases}.$$

The corresponding $(1 - \alpha) \cdot 100\%$ confidence intervals are:

$$\begin{cases} \exp\left(\hat{\beta}_2 \mp z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_2}\right) & \text{if } X_1 = 0 \\ \exp\left(\hat{\beta}_2 + \hat{\beta}_3 \mp z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\sigma}_{\hat{\beta}_2}^2 + \hat{\sigma}_{\hat{\beta}_3}^2 + 2 \cdot \hat{\sigma}_{\hat{\beta}_2, \hat{\beta}_3}}\right) & \text{if } X_1 = 1 \end{cases},$$

where $\hat{\sigma}_{\hat{\beta}_i}^2 = \widehat{\text{Var}}(\hat{\beta}_i)$, $i \in \{2, 3\}$, and $\hat{\sigma}_{\hat{\beta}_2, \hat{\beta}_3} = \widehat{\text{Cov}}(\hat{\beta}_2, \hat{\beta}_3)$.

MODELING INTERACTION (CONT.)

Interaction of 2 factors with more than 2 levels

If X_1 and X_2 are two categorical covariates with k and l levels, resp., the interaction terms to be included in the model are of type

$$\beta_{ij} X_{1_i} \cdot X_{2_j}, i = 1, \dots, k-1, j = 1, \dots, l-1,$$

where X_{1_i} and X_{2_j} are the dummy variables corresponding to X_1 and X_2 . To check for interaction, use the likelihood ratio test for the hypothesis

$$H_0: \beta_{11} = \dots = \beta_{(k-1)(l-1)} = 0 \quad \text{vs.} \quad H_1: \exists i, j: \beta_{ij} \neq 0.$$

 Assume, X_1 and X_2 have two and three levels, respectively:

- 1 Which is the expression of the logistic regression model including both variables and their interaction?
- 2 Which is the OR associated with comparing level $X_2 = 3$ with $X_2 = 2$?

MODELING INTERACTION (CONT.)

Interaction including continuous covariates

Let X_1 be a binary and X_2 a continuous variable. The expression of the model including both and their interaction may be the same as Model (1):

$$\text{logit}(p_{\mathbf{x}}) = \ln \left(\frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}} \right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2.$$

Thus, the odds ratio associated with a c units increase in X_2 is

$$\text{OR}_{X_2, c} = \begin{cases} \exp(c \cdot \beta_2) & \text{if } X_1 = 0 \\ \exp(c \cdot (\beta_2 + \beta_3)) & \text{if } X_1 = 1 \end{cases}.$$

The **log** odds ratio associated with X_1 is a **linear** function of X_2 :

$$\text{OR}_{X_1} = \exp(\beta_1 + \beta_3 \cdot X_2) \iff \ln(\text{OR}_{X_1}) = \beta_1 + \beta_3 \cdot X_2.$$

Note: The interaction of X_1 and X_2 may be modeled in another way.

LOGISTIC REGRESSION IN CASE-CONTROL STUDIES

In **case-control studies** it is possible to estimate $P(\mathbf{X}|Y)$ but not $P(Y = 1|\mathbf{X})$. Nevertheless, it is possible to fit a logistic regression model whose parameters β_1, \dots, β_m have the same interpretation as in cohort studies.

Let Z be an indicator variable of whether a person is included in the study sample or not. In addition, let

$$\pi_i = P(Z = 1|Y = i, \mathbf{X}) = P(Z = 1|Y = i), i \in \{0, 1\},$$

be the (unknown) probabilities for possible controls and cases to be sampled. Thus,

$$\begin{aligned} P(Y = 1|Z = 1, \mathbf{X}) &= \frac{P(Z = 1|Y = 1, \mathbf{X})P(Y = 1|\mathbf{X})}{\sum_{i \in \{0,1\}} P(Z = 1|Y = i, \mathbf{X})P(Y = i|\mathbf{X})} \\ &= \frac{\pi_1 P(Y = 1|\mathbf{X})}{\pi_1 P(Y = 1|\mathbf{X}) + \pi_0 P(Y = 0|\mathbf{X})}. \end{aligned}$$

CASE-CONTROL STUDIES (CONT.)

That is,

$$P(Y = 1|Z = 1, \mathbf{X}) = \frac{\pi_1 P(Y = 1|\mathbf{X})/P(Y = 0|\mathbf{X})}{\pi_1 P(Y = 1|\mathbf{X})/P(Y = 0|\mathbf{X}) + \pi_0} \quad (2)$$

Plugging the logistic regression model expression into equation (2) gives

$$P(Y = 1|Z = 1, \mathbf{X}) = \frac{\pi_1 \exp(\alpha + \beta' \mathbf{X})}{\pi_1 \exp(\alpha + \beta' \mathbf{X}) + \pi_0} = \frac{\exp(\alpha^* + \beta' \mathbf{X})}{1 + \exp(\alpha^* + \beta' \mathbf{X})},$$

where $\alpha^* = \alpha + \ln(\pi_1/\pi_0)$.

That is, a logistic regression model can be fitted to case-control data using MLE for the parameter estimation. The interpretation of β is the same as in a cohort study, whereas the interpretation of the model constant, which depends on the (unknown) sampling probabilities π_0 and π_1 , is different.

MODEL BUILDING

General considerations

- George E. P. Box¹:

“Essentially, all models are wrong, but some are useful.”

- Potential risk factors of interest (and, hence, confounders), such as sex and age, should be included in an epidemiologic regression model.
- However there should be at least five events per variable in the model. For more information, see:
Vittinghoff, E., C. McCulloch (2007). Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology* 165, 710–718.
- Be aware of possible collinearity between predictor variables.
- Both statistical and research-specific aspects should be considered.

¹From: http://en.wikiquote.org/wiki/George_E._P._Box

MODEL BUILDING (CONT.)

General considerations (cont.)

- Stepwise regression such as forward selection or backward elimination are very controversial:

For example, Hosmer and Lemeshow (2000) state that it

“... can provide a fast and effective means to screen a large number of variables,...”,

whereas others argue that, apart from many technical problems,

*“It allows us to **not** think about the problem.”*

For several (negative) opinions, see the comments under
<http://core.ecu.edu/psyc/wuenschk/stathelp/Stepwise-Voodoo.htm>

- How to include a continuous variable? See, e.g.,: Becher, H. (2002). Analysis of continuous covariables in epidemiological studies: Dose-response modelling and confounder adjustment. *Biometrical Journal* 44(6), 683–699.

MODEL BUILDING (CONT.)

A possible model building strategy for a causal model

Hosmer and Lemeshow (2000) recommend the following strategy:

- Fit simple logistic models for all variables of interest and keep those with a significance level of $\alpha < 0.25$.
- Fit a model with all variables retained in the previous step.
- Variables may now be removed, if the following conditions hold:
 - ▶ they are not statistically significant,
 - ▶ the estimated parameter is similar to the one of the simple model,
 - ▶ the coefficients of the other variables do hardly change when removing this variable from the model.
- Check, whether the previously discarded variables are now significant.
- Check how the continuous variables should be included in the model.
- Consider possible interactions of the included variables.
- Assess the final model for its goodness of fit.

CHECKING GOODNESS OF FIT

The Hosmer-Lemeshow (HL) Test

General idea: Under the hypothesis of a correct model specification, the number of events predicted by the model is expected to be similar to the ones observed.

The HL test orders the subjects according to their predicted risk for disease, $\hat{p}_{\mathbf{X}} = \hat{P}(Y = 1|\mathbf{X})$, and divides them into 5 to 10 groups of (nearly) the same size. Within each of these g groups, the observed number of events, O_k , $k = 1, \dots, g$, is compared to the expected number E_k :

$$E_k = \sum_{i=1}^{N_k} \hat{p}_{\mathbf{X}_i} = \sum_{i=1}^{N_k} \hat{P}(Y = 1|\mathbf{X}_i) = \sum_{i=1}^{N_k} \frac{\exp(\hat{\alpha} + \hat{\beta}'\mathbf{X}_i)}{1 + \exp(\hat{\alpha} + \hat{\beta}'\mathbf{X}_i)},$$

where N_k is the size of group k . The test statistic of the HL test is

$$\chi_{\text{HL}}^2 = \sum_{k=1}^g \frac{(O_k - E_k)^2}{E_k} \stackrel{H_0}{\sim} \chi_{g-2}^2.$$

CHECKING GOODNESS OF FIT (CONT.)

The Hosmer-Lemeshow Test: comments

- Implementation in R: `HLtest` (`vcdExtra`; Friendly 2017).
- The groups may be defined by the different **covariate patterns**, if these are not too many.
- An important disadvantage of the HL test is that it depends on the number of groups and how subjects are assigned to these in the case of ties. In addition, it is not a very powerful test.
- For an extensive review and comparison of several goodness of fit tests, see Hosmer *et al.* (1997).
- The test proposed by Le Cessie and van Houwelingen is implemented in the R function `residuals.lrm` (package `rms`). For details, see: le Cessie, S., J.C. van Houwelingen (1991): A Goodness-of-Fit Test for Binary Regression Models, Based on Smoothing Methods. *Biometrics* 47, 1267–1282.
- In addition to single summary statistics, different regression diagnostics exist. See, e.g., Hosmer & Lemeshow (2000), Chapter 5.3.

LOGISTIC REGRESSION WITH MATCHED DATA

Logistic regression including a matching factor

If the data come from a matched case-control study, the use of logistic regression must take into account this sampling scheme. This is done by including a stratum variable into the model that indicates to which matched set the data belong. In the case of paired matching with a total of N pairs, this implies the inclusion of $N - 1$ dummy variables I_l :

$$\text{logit}(P(Y = 1 | \mathbf{X}, i^{th} \text{ level of the matching factor})) = \alpha + \beta' \mathbf{X} + \sum_{l=1}^{N-1} \gamma_l I_l = \alpha_i^* + \beta' \mathbf{X},$$

where $I_i = 1$ and $I_l = 0, \forall l \neq i$. Hence, $\alpha_i^* = \alpha + \gamma_i$.

Problem: Too many parameters \implies ML estimation cannot be used!

Solution: The **conditional likelihood function!**

LOGISTIC REGRESSION WITH MATCHED DATA (CONT.)

The conditional likelihood function

If the covariate patterns \mathbf{x}_1 and \mathbf{x}_2 are observed in the case-control pair i , the (conditional) probability that \mathbf{x}_1 is the covariate pattern of the case ($Y = 1$) is

$$\begin{aligned} & \frac{P(\mathbf{X} = \mathbf{x}_1|Y = 1)P(\mathbf{X} = \mathbf{x}_2|Y = 0)}{P(\mathbf{X} = \mathbf{x}_1|Y = 1)P(\mathbf{X} = \mathbf{x}_2|Y = 0) + P(\mathbf{X} = \mathbf{x}_2|Y = 1)P(\mathbf{X} = \mathbf{x}_1|Y = 0)} \\ &= \frac{P(Y = 1|\mathbf{X} = \mathbf{x}_1)P(Y = 0|\mathbf{X} = \mathbf{x}_2)}{P(Y = 1|\mathbf{X} = \mathbf{x}_1)P(Y = 0|\mathbf{X} = \mathbf{x}_2) + P(Y = 1|\mathbf{X} = \mathbf{x}_2)P(Y = 0|\mathbf{X} = \mathbf{x}_1)} \\ &= \frac{P(Y = 1|\mathbf{X} = \mathbf{x}_1)/P(Y = 0|\mathbf{X} = \mathbf{x}_1)}{P(Y = 1|\mathbf{X} = \mathbf{x}_1)/P(Y = 0|\mathbf{X} = \mathbf{x}_1) + P(Y = 1|\mathbf{X} = \mathbf{x}_2)/P(Y = 0|\mathbf{X} = \mathbf{x}_2)} \\ &= \frac{e^{\alpha_i^* + \beta' \mathbf{x}_1}}{e^{\alpha_i^* + \beta' \mathbf{x}_1} + e^{\alpha_i^* + \beta' \mathbf{x}_2}} = \frac{1}{1 + e^{\beta'(\mathbf{x}_2 - \mathbf{x}_1)}} \end{aligned}$$

LOGISTIC REGRESSION WITH MATCHED DATA (CONT.)

The conditional likelihood function (cont.)

Hence, the likelihood contribution of each of the N **independent** matched case-control pairs is

$$\frac{1}{1 + e^{\beta'(\mathbf{x}_2 - \mathbf{x}_1)}}.$$

That implies that the model constant α cannot be estimated with matched case-control data.

Notes:

- If $\mathbf{x}_1 = \mathbf{x}_2$, no information is contributed to the estimation of β .
- If only one (exposure) variable is included in the model, $\hat{\beta} = \ln(\frac{n_2}{n_3})$, where n_2 (n_3) is the number of pairs with exposed case (control) and non-exposed control (case).

In R: Functions `clogit` (survival) and `clogistic` (Epi; Carstensen 2019).

REFERENCES

-  Carstensen, B., M. Plummer, E. Laara, M. Hills (2019). Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.35. URL <https://CRAN.R-project.org/package=Epi>.
-  Friendly M. (2017). vcdExtra: 'vcd' Extensions and Additions. R package version 0.7-1. [https://CRAN.R-project.org/package = vcdExtra](https://CRAN.R-project.org/package=vcdExtra)
-  Harrell, F. Jr (2019). rms: Regression Modeling Strategies. R package version 5.1-3. URL: <https://CRAN.R-project.org/package=rms>
-  Hosmer, D., T. Hosmer, S. Le Cessie, and S. Lemeshow (1997). A Comparison of Goodness-of-fit Tests for the Logistic Regression Model. *Statistics in Medicine* 16, 965–980.
-  Hosmer, D. and S. Lemeshow (2000). *Applied Logistic Regression*. Second Edition. New York: John Wiley & Sons.
-  Jewell, N. (2004). *Statistics for Epidemiology*. Chapman & Hall/ CRC.

REFERENCES

-  Kaestle, C., C. Halpern, W. Miller, and C. Ford (2005). Young Age at First Intercourse and Sexually Transmitted Infections in Adolescents and Yound Adults. *American Journal of Epidemiology* 161, 774–780.
-  Kreienbrock, L. and S. Schach (1995). *Epidemiologische Methoden*. Gustav Fischer Verlag Stuttgart-Jena.
-  Martín-Santos, R., E. Gelabert, S. Subirà, A. Gutiérrez-Zotes, K. Langohr, M. Jover, M. Torrens, R. Guillamat, F. Mayoral, F. Cañellas, J.L. Iborra, M. Gratacòs, J. Costas, I. Gornemann, R. Navinés, M. Guitart, M. Roca, R. De Frutos, F. Vilella, M. Valdés, L. García-Estévez, J. Sanjuán (2012). Research Letter: Is neuroticism a risk factor for postpartum depression? *Psychological Medicine*, 42 (7), 1559–1565.