Grau d'Estadística
Estadística Mèdica

Francesc Miras

francesc.miras@upc.edu

Universitat de Barcelona – Universitat Politècnica de Catalunya

28th November, 2018

Today's programme

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. Checking goodness of fit

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, $C_1, \ldots, C_k$,
- 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 $\boldsymbol{X}$ the model's covariate vector, which includes both E and $C_1, \ldots, C_k$.

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

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

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

$$p = \mathrm{P}(Y = 1|\boldsymbol{X}) = \mathrm{P}(Y = 1|X_1, \ldots, X_m).$$

The logistic regression model (cont.)

Expression of the logistic regression model (cont.)

Since $p \in [0,1]$, logistic regression models the **logit** transformation of p, whose range is $\mathbb{R}$, as a linear combination of the independent variables $X_1, \ldots, X_m$:

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

This expression is equivalent to

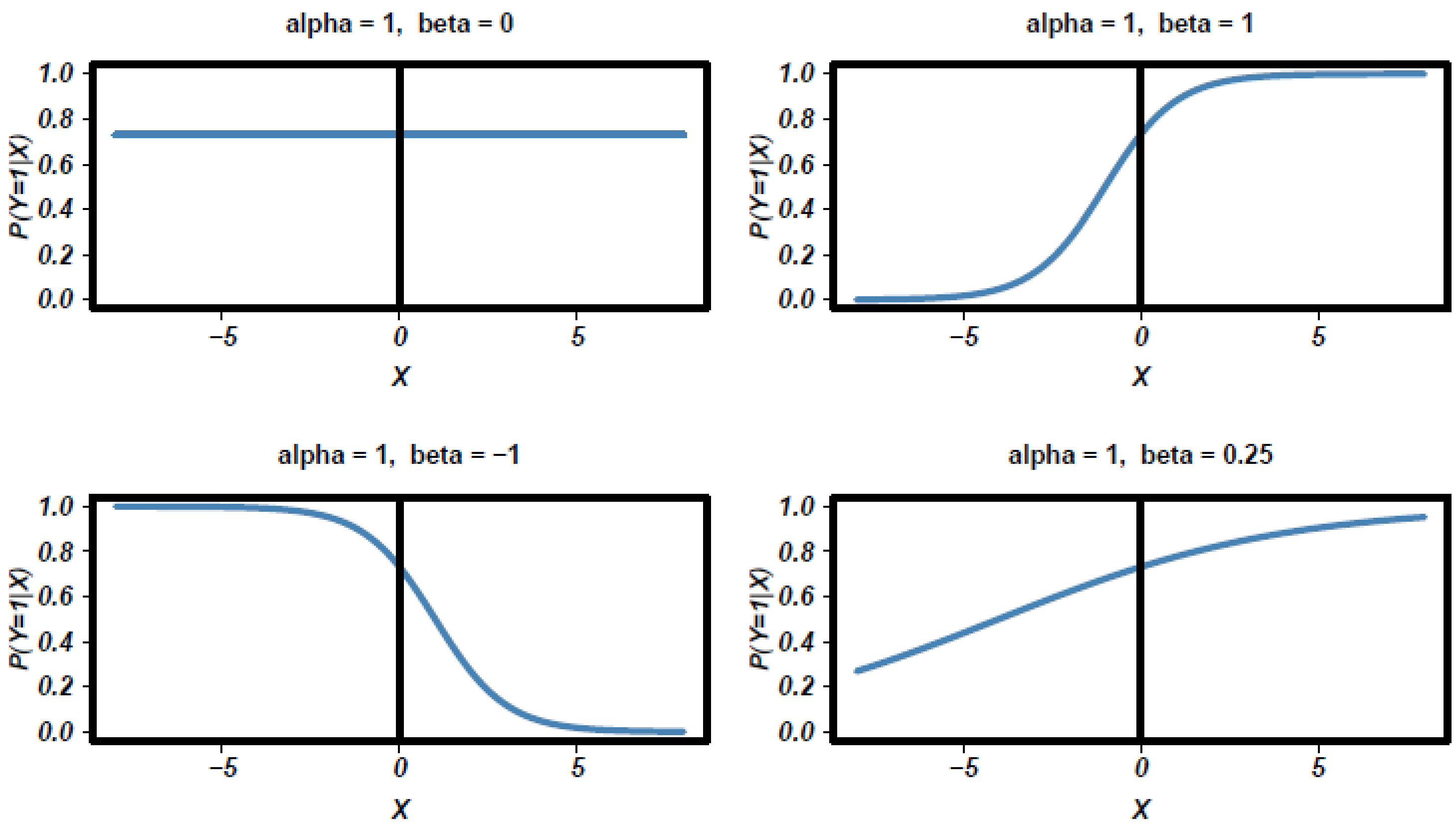
$$p = \frac{\exp(\alpha + \beta_1 X_1 + \cdots + \beta_m X_m)}{1 + \exp(\alpha + \beta_1 X_1 + \cdots + \beta_m X_m)}.$$

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 can be 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:

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

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**.

Maximum likelihood estimation of the parameters

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

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

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

In R, functions `glm` (package `stats`) and `lrm` (`rms`; Harrell 2016) 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 estimation, under the large sample condition:

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

where θ stands for any of the model parameters $\alpha, \beta_1, \ldots, \beta_m$.
The variance $\mathrm{Var}(\hat{\theta}_{\mathrm{ML}})$ 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:

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

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

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

$$\mathrm{H}_0\colon\ \beta_k = 0 \quad \text{vs.} \quad \mathrm{H}_1\colon\ \beta_k \neq 0.$$

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

$$\frac{\hat{\beta}_k}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_k)}} \overset{\mathrm{H}_0}{\sim} \mathcal{N}(0,1),$$

$$\frac{\hat{\beta}_k^2}{\widehat{\mathrm{Var}}(\hat{\beta}_k)} \overset{\mathrm{H}_0}{\sim} \chi_1^2.$$

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

$$\mathrm{H}_0\colon\ \beta_1 = \cdots = \beta_s = 0 \quad \text{vs.} \quad \mathrm{H}_1\colon\ \exists j : \beta_j \neq 0.$$

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

$$\mathrm{Dev}_{\mathrm{H}_0} - \mathrm{Dev}_{\mathrm{full}} = -2 \cdot \ln\left(\frac{L(\hat{\alpha}_{\mathrm{H}_0}, \hat{\beta}_{\mathrm{H}_0} | Y, \boldsymbol{X})}{L(\hat{\alpha}, \hat{\beta} | Y, \boldsymbol{X})}\right) \overset{\mathrm{H}_0}{\sim} \chi^2_s.$$

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

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

Interpretation of the parameters (cont.)

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

$$\widehat{\text{OR}}_{X_k} = \exp(\hat{\beta}_k),$$

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

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

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

✍ Which is the confidence interval of $\text{OR}_{X_{k,c}}$?

Interpretation of the parameters (cont.)

The model constant

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

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

To provide the model constant with a relevant meaning, continuous covariates may be centered, for example, around their means:
$X^* = X - \bar{X}$.
Hence,

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

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

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:

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

Hence, to check for interaction, one can either calculate $\text{CI}(\beta_3; 1 - \alpha)$ or test the hypothesis

$$\text{H}_0\colon\ \beta_3 = 0 \quad \text{vs.} \quad \text{H}_1\colon\ \beta_3 \neq 0,$$

by means of the Wald test.

In case, H_0 cannot be rejected, interaction is not statistically significant and the term $\beta_3 X_1 \cdot X_2$ may be removed from model (1).

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 depends on the levels of the other.
For example, the OR associated with X_2 is

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

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

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

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

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

$$\mathrm{H}_0\colon \; \beta_{11} = \cdots = \beta_{(k-1)(l-1)} = 0 \quad \text{vs.} \quad \mathrm{H}_1\colon \; \exists i, j : \beta_{ij} \neq 0.$$

✍ 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):

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

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

$$\text{OR}_{X_{2,c}} = \begin{cases} \exp(c \cdot \beta_2) & \text{if } X_1 = 0 \\ \exp\left(c \cdot (\beta_2 + \beta_3)\right) & \text{if } X_1 = 1 \end{cases}.$$

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

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

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 $\mathrm{P}(\boldsymbol{X}|Y)$ but not $P(Y=1|\boldsymbol{X})$. Nevertheless, it is possible to fit a logistic regression model whose parameters $\beta_1, \ldots, \beta_m$ 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

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

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

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

Case-control studies (cont.)

That is,

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

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

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

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.

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, $\hat{p} = \hat{\mathrm{P}}(Y = 1|\boldsymbol{X})$, 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, $k = 1, \ldots, g$, is compared to the expected number E_k:

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

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

$$\chi^2_{\mathrm{HL}} = \sum\nolimits_{k=1}^{g} \frac{(O_k - E_k)^2}{V_k} \overset{\mathrm{H}_0}{\sim} \chi^2_{g-2}.$$

Checking goodness of fit (cont.)

The Hosmer-Lemeshow Test: comments

- Implementation in R: `HLtest` (`vcdExtra`; Friendly 2016).
- 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.

References

- 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.
- Martín-Santos, R., E. Gelabert, S. Subirà, A. Gutiérrez-Zotes, K. Langohr, M. Jover *et al.* (2012). Research Letter: Is neuroticism a risk factor for postpartum depression? *Psychological Medicine*, 42 (7), 1559–1565.
- Kreienbrock, L. and S. Schach (1995). *Epidemiologische Methoden.* Gustav Fischer Verlag Stuttgart·Jena.