Biostatistics

V. Inácio de Carvalho & M. de Carvalho

University of Edinburgh

PRR Plano de Recuperação e Resiliência

GOVERNO DE PORTUGAL

Financiado pela União Europeia NextGenerationEU

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Logistic regression with case-control data

Example

- MacMahon et al. (1981) reported on a traditional case-control study of pancreatic cancer and its relationship to various lifestyle habits including consumption of tobacco, alcohol, tea, and coffee.
- The table below gives the resulting data on coffee drinking and incidence of pancreatic cancer, for men and women separately.

		Coffee Drinking (Cups per Day)				
Sex	Disease Status	0	1–2	3–4	≥ 5	Total
Men	Case	9	94	53	60	216
	Controls	32	119	74	82	307
Women	Case	11	59	53	28	151
	Controls	56	152	80	48	336
	Total	108	424	260	218	1010

- Cases were recruited from 11 hospitals in Boston and Rhode Island over a 5-year period.
- Controls were then sampled from the patient populations of physicians who treated the selected pancreatic cancer cases, excluding patients who had any pancreatic disease, or who suffered from other smoking or alcohol-related conditions.

Logistic regression with case-control data

Example

↪ Let us consider the following logistic regression model:

$$\log\left(\frac{p_{x_1,x_2,x_3,x_4}}{1-p_{x_1,x_2,x_3,x_4}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,$$

where

$$X_1 = \begin{cases} 1, & \text{1–2 cups of coffee per day,} \\ 0, & \text{otherwise.} \end{cases}$$

$$X_2 = \begin{cases} 1, & \text{3–4 cups of coffee per day,} \\ 0, & \text{otherwise.} \end{cases}$$

$$X_3 = \begin{cases} 1, & \geq 5 \text{ cups of coffee per day,} \\ 0, & \text{otherwise.} \end{cases}$$

$$X_4 = \begin{cases} 1, & \text{female,} \\ 0, & \text{male.} \end{cases}$$

Logistic regression with case-control data

Example

- What is the interpretation of the parameters in this case?
- β_0 is the log odds of pancreatic cancer for a man who does who does not drink any coffee.
- β_1 is the log odds ratio of pancreatic cancer comparing a subject who drinks 1-2 cups of coffee per day to one who does not drink any, holding gender constant. To see why

$$\begin{aligned}\beta_1 &= \log\left(\frac{p_{1,0,0,x_4}/(1-p_{1,0,0,x_4})}{p_{0,0,0,x_4}/(1-p_{0,0,0,x_4})}\right)\\ &= [\beta_0+\beta_1\times 1+\beta_2\times 0+\beta_3\times 0+\beta_4\times x_4]-[\beta_0+\beta_1\times 0+\beta_2\times 0+\beta_3\times 0+\beta_4\times x_4],\end{aligned}$$

  where x_4 can either be 0 or 1 (but it needs to take the same value).

- An analogous interpretation follows for β_2 and β_3.
- β_4 is the log odds ratio comparing a female to a male, holding coffee consumption constant.

Logistic regression with case-control data

Example

↪ Fitting the model using the `glm` function in `R` (see Supplementary Materials file) we obtain

Parameter	Estimate	OR (95% CI)
β_0	—	
β_1	0.867	2.379 (1.405, 4.029)
β_2	1.073	2.923 (1.691, 5.051)
β_3	0.990	2.691 (1.536, 4.716)
β_4	−0.404	0.668 (0.513, 0.870)

↪ We can see that for instance, the odds of pancreatic cancer, when holding gender constant, for subjects who drink 3 or 4 cups of coffee per day almost three fold those of subjects who do not drink any coffee a day. Because 1 is not included in the 95% CI, there is evidence of a (positive) significant association.

↪ Also, the odds of pancreatic cancer for females, when holding coffee consumption constant, are roughly 0.7 times the odds of male subjects.

Logistic regression with case-control data

Example

↪ In addition, because calculations involving the intercept will be distorted, estimation of the risk of pancreatic cancer at any particular exposure level is not possible, since this involves $\widehat{\beta}_0$

$$\log\left(\frac{\widehat{p}_{x_1,x_2,x_3,x_4}}{1-\widehat{p}_{x_1,x_2,x_3,x_4}}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4$$

$$\widehat{p}_{x_1,x_2,x_3,x_4} = \frac{e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2+\widehat{\beta}_3 x_3+\widehat{\beta}_4 x_4}}{1+e^{\widehat{\beta}_0+\widehat{\beta}_1 x_1+\widehat{\beta}_2 x_2+\widehat{\beta}_3 x_3+\widehat{\beta}_4 x_4}}.$$

Tests for significance of model coefficients

- In what follows we list some facts, without proving them. Any textbook covering logistic regression will cover the proofs in detail.

- Two likelihood based approaches to testing significance of model coefficients are popular:

  - The Wald test for a single coefficient.

  - The likelihood ratio method.

Tests for significance of model coefficients

Wald test

- Consider a logistic regression model with k predictors and the test for significance:

$$H_0 : \beta_j = 0, \qquad j = 1, \ldots, k.$$

- H_0 is tested against the two-sided alternative $H_0 : \beta_j \neq 0$.
- Akin to the technique used to calculate confidence intervals for β_j, the Wald method simply computes the test statistic

$$z_{\beta_j} = \frac{\widehat{\beta}_j - 0}{\widehat{\mathrm{SE}}(\widehat{\beta}_j)}.$$

- Under the null hypothesis, the Wald statistic z_{β_j} will approximately follow, in large samples, a standard normal distribution.
- Often the Wald statistic is squared and then compared to a χ^2 distribution with one degree of freedom.
- This test is equivalent to checking whether the value 0 is contained at the $100(1-\alpha)\%$ confidence interval (in `R` we need to be careful and use `confint.default` instead of `confint`).

Tests for significance of model coefficients

Wald test

- ↪ Note that the p-values are testing whether the corresponding coefficients could really be zero given that the other terms remain in the model (i.e. are nonzero).

- ↪ Dropping one term (i.e., setting it to zero) will change the estimates of the other coefficients and hence their p-values.

Tests for significance of model coefficients

Likelihood ratio method

- The likelihood ratio (or deviance) test is used to test the null hypothesis that any subset of the regression coefficients is equal to zero.
- For a given dataset, consider two regression models, labelled as Model A and Model B for convenience.
- Further, suppose that model A is nested within model B, that is, model A is a special case of model B, described by setting some of the coefficients of model B to zero.
- For instance, in the coronary heart disease (CHD) example (last set of slides), where we have used weight discretised in five categories (≤ 150, $150^+ - 160$, $160^+ - 170$, $170^+ - 180$, ≥ 180), one possible model A might be

$$\log\left(\frac{p_{x_1,x_2,x_3,x_4}}{1-p_{x_1,x_2,x_3,x_4}}\right) = \beta_0.$$

Tests for significance of model coefficients

Likelihood ratio method

- This model, of course, claims that CHD incidence is the same for all levels of body weight, that is, body weight and CHD are independent.
- Thus, model A corresponds to our null hypothesis and model B to the alternative model.
- Model A is nested within model B; we can see that model A is a special case of model B since it corresponds to setting $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Tests for significance of model coefficients

Likelihood ratio method

- ↪ Consider the calculation of the maximised log likelihood under model A.
- ↪ In our running CHD example, this involves calculation of $\widehat{\beta}_0$ and evaluation of the log likelihood at $\widehat{\beta}_0$.
- ↪ Consider also the calculation of the maximised log likelihood under model B.
- ↪ This now involves calculation of $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, and $\widehat{\beta}_4$, and evaluation of the log likelihood at these values.
- ↪ Note that the intercept estimates $\widehat{\beta}_0$ will be different under the two models and have different interpretations.
- ↪ The maximised log likelihood under model B will necessarily be larger than under model A since it is a more general model.

Tests for significance of model coefficients

Likelihood ratio method

- Under the null hypothesis that the two models are equivalent, i.e., that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, the test statistic is given by

  $2 \times$ (the maximised log likelihood in model B $-$ the maximised log likelihood in model A).

- Note that this test statistic is equivalent to

$$2 \log \left(\frac{\text{the maximised likelihood in model B}}{\text{the maximised likelihood in model A}} \right) = 2\{\log L_B(\widehat{\beta}) - \log L_A(\widehat{\beta})\}.$$

- The test statistic follows a sampling distribution approximated by a χ^2 distribution in large samples.
- The appropriate number of degrees of freedom for the χ^2 distribution is $p_B - p_A$, where p_A and p_B denote the number of parameters in model A and model B, respectively.
- In the CHD example, model A has one parameter and model B has five parameters, and so the null hypothesis sampling distribution for the likelihood ratio test is $\chi^2_{(4)}$.
- Likelihood ratio tests are similar to partial F-tests in the sense they compare the full model with a restricted model where the explanatory variables of interest are omitted.

Tests for significance of model coefficients

Likelihood ratio method

- We will now introduce the concept of **deviance** which is a key concept in logistic regression (and, more broadly, in generalised linear models).
- Intuitively, it measures the deviance of the fitted logistic regression model with respect to a perfect model for the sample.
- This perfect model, known as the **saturated** model, is the model that perfectly fits the data, in the sense that the fitted (estimated) probabilities equal the observed responses.
- Remember the discussion we have just had (in the Supplementary Materials) about grouped and ungrouped data and note that the saturated model differs in the two cases.
- For ungrouped binary data, the saturated model has fitted probability $\tilde{p}_{\mathbf{x}_i} = \tilde{p}_i = d_i$, where d_i is either zero or one.
- On the other hand, for grouped binary data, the saturated model has fitted probability given by $\tilde{p}_k = d_k / n_k$ for all n_k observations at a particular category k, where d_k is the number of cases for n_k individuals (cases plus nondiseased subjects) in a particular category k.

Tests for significance of model coefficients

Likelihood ratio method

- A perfect fit sounds good but the saturated model does not smooth the data or has the advantages of parsimony that a simpler model has, such as a better estimate of the true relation between the disease outcome/response and exposure variables.

- However, it serves as a baseline for constructing the likelihood ratio statistics that compares it to the chosen model

$$\text{deviance} = 2 \log \left(\frac{\text{maximum likelihood for saturated model}}{\text{maximum likelihood for chosen model}} \right) = 2(\log L_s - \log L(\widehat{\beta})),$$

  where $\log L_s$ and $\log L(\widehat{\beta})$ denote the log likelihood of the saturated model and the log likelihood of the chosen model, respectively.

- Since the saturated model is more general than the chosen model, $\log L_s \geq \log L(\widehat{\beta})$ and so the deviance is always non-negative.

Tests for significance of model coefficients

Likelihood ratio method

- Let us go back to the situation where we want to compare two models A and B, and model A is a particular case of model B (in statistical terminology, we say that model A is nested on model B).
- We have that

$$\begin{aligned}\text{deviance A} - \text{deviance B} &= 2(\log L_s - \log L_A(\widehat{\boldsymbol{\beta}})) - 2(\log L_s - \log L_B(\widehat{\boldsymbol{\beta}})) \\ &= 2(\log L_B(\widehat{\boldsymbol{\beta}}) - \log L_A(\widehat{\boldsymbol{\beta}})),\end{aligned}$$

  which is exactly the likelihood ratio statistic we had before.

- The advantage of using the deviance is that in `R` we obtain the deviances in the summary of the output of the `glm` function.

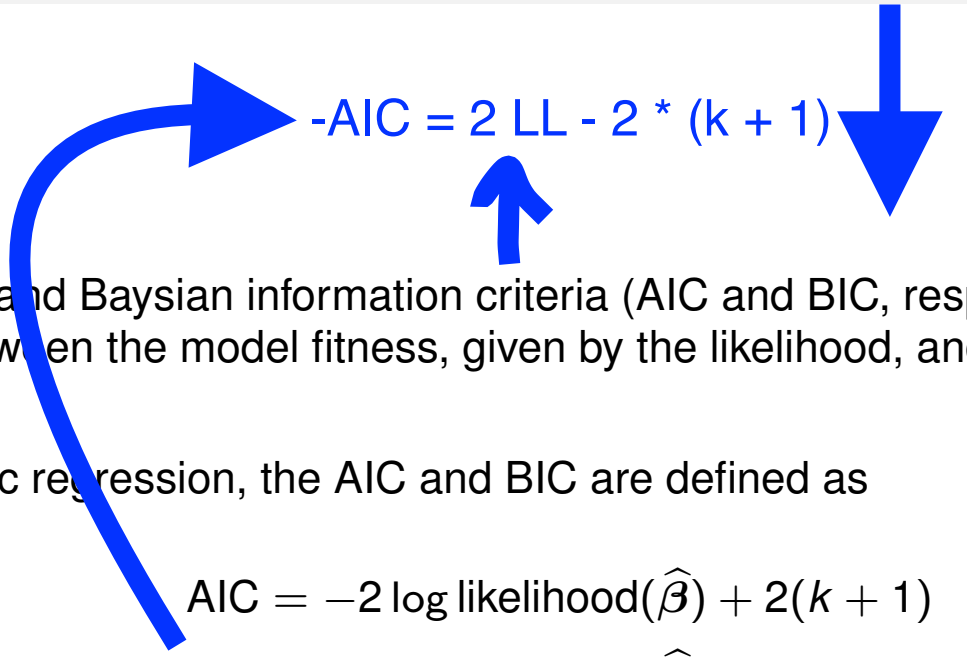
Tests for significance of model coefficients

Likelihood ratio method

- In our running CHD example, the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, yields a test statistic of 21.40.

- The corresponding p-value is 0.0003. We therefore reject the null hypothesis.

AIC/BIC



↪ The Akaike and Baysian information criteria (AIC and BIC, respectively) are based on a balance between the model fitness, given by the likelihood, and its complexity.

↪ In the logistic regression, the AIC and BIC are defined as

$$\text{AIC} = -2\log\text{likelihood}(\widehat{\beta}) + 2(k+1)$$
$$\text{BIC} = -2\log\text{likelihood}(\widehat{\beta}) + \log(n) \times (k+1).$$

↪ Here $k+1$ denote the total number of parameters (k exposure variables plus one intercept).

AIC/BIC

- When $n \geq 8$, $\log(n) \geq 2$ and so the penalty term in the BIC is greater than the penalty term in the AIC.
- Thus, in those circumstances, the BIC penalises model complexity more heavily than the AIC, thus favouring simpler models than the AIC.
- When comparing several models, the one with the lowest AIC/BIC is to be preferred.
- An advantage of the AIC/BIC to the likelihood ratio test is that they allow to compare non-nested models.