An Acad Bras Cienc (2021) 93(4): e20190316 DOI 10.1590/0001-3765202120190316
Anais da Academia Brasileira de Ciências | *Annals of the Brazilian Academy of Sciences*
Printed ISSN 0001-3765 I Online ISSN 1678-2690
www.scielo.br/aabc | www.fb.com/aabcjournal

MATHEMATICAL SCIENCES

Prevalence ratio estimation via logistic regression: a tool in R

LEILA D. AMORIM & RAYDONAL OSPINA

Abstract: The interpretation of odds ratios (OR) as prevalence ratios (PR) in cross-sectional studies have been criticized since this equivalence is not true unless under specific circumstances. The logistic regression model is a very well known statistical tool for analysis of binary outcomes and frequently used to obtain adjusted OR. Here, we introduce the *prLogistic* for the R statistical computing environment which can be obtained from *The Comprehensive R Archive Network*, https://cran.r-project.org/package=prLogistic. The package *prLogistic* was built to assist the estimation of PR via logistic regression models adjusted by delta method and bootstrap for analysis of independent and correlated binary data. Two applications are presented to illustrate its use for analysis of independent observations and data from clustered studies.

Key words: Logistic model, delta method, bootstrap, prevalence ratios.

INTRODUCTION

The concept of risk is fundamental in several research areas, being the measures of risk associated to the probability of occurrence of an event of interest. Particularly in Public Health and Epidemiology, two commonly used measures to estimate risk are relative risk (RR), in longitudinal studies, and prevalence ratios (PR), in cross-sectional studies. In the simplest situation, unadjusted measures can be computed easily throughout analysis of contingency tables. Another measure of association frequently reported by epidemiologists and medical researchers is the odds ratio (OR), which differs mathematically from RR and PR. It is well known that OR overestimates relative risk (RR) and prevalence ratios (PR) when the event is not rare (Stromberg 1994, Thompson et al. 1998, McNutt et al. 2003, Greenland 2004, Newcombe 2006, Tamhane et al. 2016).

When the main interest of the investigator relies on the estimation of an association or risk measure adjusted by covariates or confounders, use of statistical modeling is usually required. In epidemiology, several outcomes are binary and logistic regression models are widely applied. Using logistic regression models, one can easily estimate OR $= \exp(\beta)$, where β is a parameter related to the risk factor of interest.

However, interpretation of OR as a risk measure might be misleading in terms of how it can be interpreted. Many researchers mistakenly interpret odds as risk even when OR provides a poor approximation to PR. Phrases including terms like risk, "likelihood", "probability" and "more likely" to interpret the OR are commonly found in the literature. In certain circumstances it is possible to estimate RR or PR through their relationships to OR. Nevertheless, the computation of the

corresponding confidence intervals is not trivial and some unsuccessful methods had been proposed (Zhang & Yu 1998).

There is some debate in the literature about alternative approaches to obtain adjusted measures of PR in cross-sectional studies (Barros & Hirakata 2003, Localio et al. 2007, Petersen & Deddens 2008, Cummings 2009, Savu et al. 2010). One of the proposals is to estimate PR using logistic regression (Oliveira et al. 1997, Localio et al. 2007). A more recent discussion is about how to estimate adjusted PR for correlated data, particularly in the analysis of clustered data (Bastos et al. 2015, Santos et al. 2008).

Implementation of new statistical methods and its availability for applied researchers is other concern among data analysts. Many of the most recently proposed statistical methods can not be applied to data analysis because they are not easily accessible via standard statistical software.

Here, we introduce *prLogistic*, an R package specifically built to assist estimation of PRs in cross-sectional studies via logistic regression models for analysis of both independent and correlated data. *prLogistic* currently contains three main functions. The first one allows estimates PR using logistic models. The second function performs PR estimation using logistic models with conditional standardization. Finally, the third function estimates PR using a logistic model with marginal standardization. We provide a "how to" guide to use those functions by applying them to empirical data sets and supply insights on interpreting the outputs.

This paper is structured as follows. In Section 2 we outline the theory underlying PR estimation via logistic models while in Section 3 we describe the functions contained in our *prLogistic* package. We provide two empirical applications to illustrate the use of the *prLogistic* package in Section 4. Finally, Section 5 contains concluding remarks and directions for future research.

1 - ESTIMATION OF PREVALENCE RATIOS

The OR can be defined by the ratio of two odds, such that $\text{OR} = (p_1/(1-p_1))/(p_0/(1-p_0))$, where p_1 and p_0 denote, respectively, the prevalence of the event of interest in the exposed and non-exposed groups. The PR, on the other hand, is defined by the ratio of two proportions given by $\text{PR} = p_1/p_0$. Therefore, interpretation of these two measures is not the same, unless the event is rare, which implies $(1-p_1) \to 1$ and $(1-p_0) \to 1$.

Let Y be the binary outcome, where $Y = 1$ if the outcome is a "success", whatever your definition, and $Y = 0$ otherwise. The probability of success is assumed to depend on known covariates, i.e., we consider the logistic regression model

$$\text{E}(Y|\mathbf{X}) = P(Y = 1|\mathbf{X}) = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)},$$

where $\mathbf{X} = [X_1, X_2, ..., X_k]$ is a matrix of independent variables and $\beta = (\beta_0, \ldots, \beta_k)^\top$ is a vector of model parameters.

We are interested in obtaining an expression for estimating PR as a function of β. For instance, suppose that we are evaluating the effect of a binary exposure X_1 (0/1) on the occurrence of Y after adjustment by $k-1$ independent variables $(X_2, ..., X_k)$. In this case, PR is given by

$$\mathrm{PR} = \frac{1 + \exp\{-\beta_0 - \beta_2 X_2 - \cdots - \beta_k X_k\}}{1 + \exp\{-\beta_0 - \beta_1 - \beta_2 X_2 - \cdots - \beta_k X_k\}}.$$

Note that, in this expression, PR is also function of the values of the independent variables included in the model, differently from OR.

1.1 - Standardization Procedures

Some standardization procedures for effect measures based on regression models had been proposed in the literature (Wilcosky & Chambless 1985, Lane & Nelder 1982), being the two most commonly used methods called conditional and marginal standardization. For the conditional standardization procedure, a reference or baseline value (for instance, the mean for continuous variables) of each variable included in the model is defined and, thus, the prevalence for each group ($X_1 = 1$ e $X_1 = 0$) is calculated. Using conditional standardization via logistic regression, the investigator is interested in finding the PR comparing exposure status (present/absent, i.e., $X_1 = 1/X_1 = 0$) at a fixed level of covariates.

For the marginal standardization procedure, on the other hand, the prevalence is computed, for each group (say p_{1i} and p_{0i}, respectively, for exposed and non exposed groups) using the individual values for the covariates and later getting the average value among all observations. Therefore, the marginal standardized prevalences for the exposed and non exposed subjects are given, respectively, by averaging p_{1i} and p_{0i} across all i subjects.

As an example consider data on n subjects with a binary exposure X_1 (1 = exposed, 0 = non-exposed), and a continuous variable X_2. Using the conditional standardization procedure, the adjusted PR is given by

$$\mathrm{PR} = \frac{1 + \exp\{-\beta_0 - \beta_2 \overline{X}_2)}{1 + \exp\{-\beta_0 - \beta_1 - \beta_2 \overline{X}_2\}},$$

where $\overline{X}_2$ denotes the mean of X_2. If X_2 were a binary covariate, the researcher should set specific value for computing PR (for instance $X_2 = 0$).

For the marginal method the adjusted PR is defined by

$$\mathrm{PR} = \frac{\frac{1}{n}\sum_i (1/\{1 + \exp(-(\beta_0 + \beta_1 + \beta_2 X_{2i}))\})}{\frac{1}{n}\sum_i (1/\{1 + \exp(-(\beta_0 + \beta_2 X_{2i}))\})}, \quad (1)$$

where the sum is over all n subjects.

1.2 - Inference for Prevalence Ratios

Methods for obtaining confidence intervals for PR include delta method and bootstrap. The delta method is a general technique for asymptotic distribution of random variable functions that is based on approximation by Taylor series (Bishop et al. 2007). Let $(X_1, X_2, ..., X_k)$ be a k-dimensional random

vector and $h(X_1, X_2, ..., X_k)$ be a function defined on an open subset of k-dimensional space to real values. We assume that $h(\cdot)$ is differential and $\mathrm{E}(X_i) = \mu_i$. Thus

$$\mathrm{VAR}(h(X_1, X_2, ..., X_k)) \approx \sum_{i=1}^{k} \left(\frac{\partial h}{\partial \mu_i}\right)^2 \cdot \mathrm{VAR}(X_i) + 2 \sum \sum_{i<j}^{k} \frac{\partial^2 h}{\partial \mu_i \mu_j} \mathrm{COV}(X_i, X_j),$$

which involves partial derivates of the function of interest. For the estimation of the variance of PR, $\log(\mathrm{PR})$ is asymptotically normally distributed and we use the delta method for estimating $\mathrm{VAR}(\log(\mathrm{PR}))$, where

$$\widehat{\mathrm{VAR}}(\log(\widehat{\mathrm{PR}})) \approx X_* \hat{\Sigma} X_*',$$

with $X_* = \hat{q}_1 X_1 - \hat{q}_0 X_0$, $\hat{q}_1 = 1 - \hat{p}_1$, $\hat{q}_0 = 1 - \hat{p}_0$, $X_1 = [1, 1, X_2, ..., X_k]$, $X_0 = [1, 0, X_2, ..., X_k]$, and $\hat{\Sigma}$ is the covariance matrix of the model parameters (β's) (Oliveira et al. 1997). Using the delta method, the adjusted $(1 - \alpha)$% confidence intervals (CIs) for PR are defined by

$$\exp(\log(\widehat{\mathrm{PR}}) \pm z_{\alpha/2} \widehat{\mathrm{se}}(\log(\widehat{\mathrm{PR}}))),$$

where $\log(\widehat{\mathrm{PR}})$ is the estimate for adjusted $\log(\mathrm{PR})$, $\widehat{\mathrm{se}}(\log(\widehat{\mathrm{PR}}))$ is the estimate of standard-error for $\log(\mathrm{PR})$ and $z_{\alpha/2}$ is the quantile of a standard normal distribution, α being the significance level.

The bootstrap approach, in its turn, is based on resampling with replacement for estimation of functions of interest (Efron & Tibshirani 1993). For instance, we can consider 1,000 bootstrap replicates to produce a bootstrap distribution of PR values. Using bootstrap estimates for sample variance (Davison & Hinkley 1997), the first order normal approximation bootstrap $(1 - \alpha)$% CI is

$$\exp(\log(\widehat{\mathrm{PR}^*}) \pm z_{\alpha/2} \widehat{\mathrm{se}^*}(\log(\widehat{\mathrm{PR}^*}))),$$

where $\log(\widehat{\mathrm{PR}^*})$ is the bootstrap estimate for adjusted $\log(\mathrm{PR})$ and $\widehat{\mathrm{se}^*}(\log(\widehat{\mathrm{PR}^*}))$ is the bootstrap estimate for the standard error of $\log(\widehat{\mathrm{PR}})$. An alternative approach, using bootstrap percentile interval (Fox 2015), considers the empirical quantiles of bootstrap estimates for defining the interval. In such situation, the interval limits are given, for instance, by percentiles 2.5 and 97.5 when we are interested in the 95% CI.

1.3 - Random-Effects Logistic Model

Random-effect models, also known as mixed models or multilevel models, are often used for modeling correlated data (Diggle et al. 1994, Hox et al. 2017). Such data arises from the sampling design, including the use of cluster sampling, where individuals are nested in geographic areas or institutions, such as schools or companies, and the use of longitudinal studies, which investigates changes in repeated measures of the outcome over time for the same sampling unit. These models can be applied for analysis of a variety of outcomes types: continuous, binary, polytomous, counts, time-to-event, etc. Here we focus on the analysis of binary outcomes. These models make adjustments for non-observed individual characteristics, which reflect a natural heterogeneity among subjects. Let Y_{ij} be the binary outcome variable at cluster j for subject i, and denote X_1 and X_2 two independent variables. The random-effects logistic model can be defined by

$$\mathrm{logit}[\mathrm{P}(Y_{ij} | X_{1ij}, X_{2ij}, u_{oj})] = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + u_{oj},$$

where $u_{0j} \sim N(0, \varsigma^2)$ represents a cluster specific random effect. Using the estimates for the parameters of this model (β's), we can obtain PR as defined previously. Interpretation of regression coefficients from the random-effects logistic model has to be done by conditioning on the random effects (Larsen et al. 2000, McCulloch & Searle 2001).

2 - THE R PACKAGE PRLOGISTIC

The *prLogistic* package is implemented under the FLOSS (Free/Libre Open Source Software) paradigm in the **R** system for statistical computing (R Development Core Team 2021) and it is available from the Comprehensive **R** Archive Network (CRAN) at https://cran.r-project.org/package=prLogistic and the experimental updates at `GitHub` repository https://github.com/Raydonal/prLogistic. It takes advantage of functionality developed in other packages as *epiR* (Stevenson et al. 2013), *boot* (Canty 2002) and *lme4* (Bates et al. 2015, 2021).

The main functionalities of the **R** package *prLogistic* (Ospina & Amorim 2021) are:

- The function `prLogisticDelta` estimates prevalence ratios (PRs) and their confidence intervals using logistic models. The estimation of standard errors for PRs is done using the delta method. The function `prLogisticDelta` allows estimation of PRs using two standardization procedures: conditional or marginal (Wilcosky & Chambless 1985).

  A typical form used with `glm()` function is included in the formula argument as response ~ terms, where the response is the (binary) response vector and terms is a series of variables which specifies a linear predictor for the response. The `prLogisticDelta` assumes a binomial family associated with it. The `lmer()` function is used when a vertical bar character `|` separates an expression for a model matrix and a grouping factor. The output returned by `prLogisticDelta` contains prevalence ratio and its 95% confidence intervals.

- The function `prLogisticBootCond()` estimates prevalence ratios (PRs) and bootstrap confidence intervals using logistic models with conditional standardization. The estimation of standard errors for PRs is given through bootstrapping. The fitted model object can be obtained using `glm()` function for binary responses when unit samples are independent. The `lmer()` function should be used for correlated binary responses. The output returned by `prLogisticBootCond` contains prevalence ratios and their 95% bootstrap confidence intervals with conditional standardization. Both normal and percentile bootstrap confidence intervals are presented.

- The function `prLogisticBootMarg()` estimates prevalence ratios (PRs) and bootstrap confidence intervals using logistic models with marginal standardization. The estimation of standard errors for PRs is given through bootstrapping. The fitted model object can be obtained using `glm()` function for binary responses when unit samples are independent. The `lmer()` function should be used for correlated binary responses. The output returned by `prLogisticBootMarg` contains prevalence ratios and their 95% bootstrap confidence intervals with marginal standardization. Both normal and percentile bootstrap confidence intervals are presented.

For the functions `prLogisticDelta`, `prLogisticBootCond` and `prLogisticBootMarg`, confidence intervals of $(1 - \alpha)\%$ for PRs are available for standard logistic regression and for random-effects

logistic models (Santos et al. 2008). If categorization for predictors is other than (0,1), `factor()` should be considered.

3 - APPLICATIONS

We illustrate the use of the R implemented functions `prLogisticDelta()`, `prLogisticBootCond()` and `prLogisticBootMarg()`, which are available in *prLogistic* package. We describe two datasets that are used in the examples. The first example is related to data from randomized clinical trials to evaluate the impact of intervention programs on drug use reduction (dataset UIS), which contains independent observations. The second example, on the other hand, is an observational clustered study about primary education in Thailand (dataset Thailand).

3.1 - The Umaru Impact Study

Dataset UIS contains information from randomized trials related to treatment for drug abuse obtained by the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study (UIS). The study aimed to compare treatment programs of different durations in the reduction of drug abuse and in the prevention of high-risk HIV behavior. The variables on the dataset available at *prLogistic* package are age at enrollment, intravenous (IV) drug use history at admission, race, treatment group, treatment site, and patient's status at the end of the treatment program (Hosmer Jr et al. 2013).

We load the package *prLogistic* and dataset UIS, and look at the first 5 rows of the data:

```
R> library("prLogistic")
R> data("UIS")
R> head(UIS)

ID Age DrugUse race trt site drugFree
1   0       0    0   1    0        0
2   0       0    0   1    0        0
3   0       0    0   1    0        0
4   1       0    0   0    0        0
5   1       1    1   1    0        1
```

Dataset UIS contains the following subset of the variables from the original study:

- `ID` is the patient identification code.
- `Age` is the age at enrollment (in years) recoded to 1 = 32 years or younger; 0 = otherwise.
- `DrugUse` is the IV drug use history at admission (1 = never; 0 = previous or recent).
- `race` is the patient's race (1 = other; 0 = white).
- `trt` is the treatment group (1 = long; 0 = short).
- `site` is the treatment site (1 = B; 0 = A).

- `drugFree` is an indicator of returning to drug use prior to the scheduled end of the treatment program (1 = remained drug free; 0 = otherwise).

We describe the outcome variable using the following:

```
R> prop.table(table(drugFree))

drugFree
0         1
0.7443478 0.2556522
```

We noted that about 26% of the patients remained drug free at the end of the treatment program. The description of the outcome according to treatment group is given by:

```
R> prop.table(table(drugFree, trt),2)

                    trt
drugFree         0         1
0            0.7854671 0.7027972
1            0.2145329 0.2972028
```

Since the outcome is relatively common, ORs do not approximate prevalence ratios. We fit the following logistic regression model

$$\log\left\{\frac{P(Y_i=1)}{P(Y_i=0)}\right\} = \beta_0 + \beta_1 \texttt{Age}_i + \beta_2 \texttt{DrugUse}_i + \beta_3 \texttt{trt}_i. \quad (2)$$

We estimate the prevalence ratios (PRs) in model (2) with 95% confidence intervals using delta method and conditional standardization through:

```
R> prLogisticDelta(drugFree ~ Age + DrugUse + trt, data = UIS)

95% Confidence Interval using Delta method
              Estimate    2.5%    97.5%
Age            0.62482   0.4410  0.88526
DrugUse        1.83711   1.3817  2.44269
trt            1.42095   1.0580  1.90845
```

For comparison, note that the odds ratios can be estimated using logistic regression as:

```
R> fit.logistic <- glm(drugFree ~ Age + DrugUse + trt,
family=binomial, data=UIS)
R> cbind(exp(fit.logistic$coefficients),
exp(confint(fit.logistic)))[2:4,]

        Estimate     2.5%     97.5%
Age     0.5716754 0.3790377 0.8557963
DrugUse 2.3179035 1.5475867 3.4904855
trt     1.5864326 1.0803760 2.3406993
```

Note that both ORs and PRs were estimated using logistic regression. Considering the previous results, we can observe, for instance, that patients who never used IV drugs before admission were more likely to remain drug free than patients with previous/recent IV drug use history ($\widehat{\mathrm{PR}}$ = 1.84 (95% CI: 1.38; 2.44); $\widehat{\mathrm{OR}}$ = 2.32 (95% CI: 1.54; 3.48)).

We can obtain 95% bootstrap confidence intervals for PR with conditional standardization using:

```
R> prLogisticBootCond(fit.logistic, data = UIS)

95% Confidence Interval using Bootstrap Method
                      Normal      Percentile
         Estimate  2.5%    97.5%   2.5%   97.5%
Age      0.62482 0.40227 0.82575 0.40285 0.90058
DrugUse  1.83711 1.19974 2.37142 1.33865 2.55240
trt      1.42095 0.85847 1.81893 1.11409 2.12350
```

Note that function `prLogisticBootCond` provides two bootstrap confidence intervals: (a) one based on normal theory, which is often approximately the case in sufficiently large samples, (b) the other using empirical quantiles of the bootstrap estimates to form the interval.

Based on the results of the conditional standardization, the probability of remaining drug-free by the end of the treatment program is 42% larger for those participating in the long treatment group compared to those in the short group among patients with more than 32 years old and who used IV drug previously or recently before to admission.

Similarly, we can estimate PRs using marginal standardization with delta method:

```
> prLogisticDelta(drugFree ~ Age + DrugUse + trt,
data = UIS, pattern="marginal")

95% Confidence Interval using Delta method
         Estimate  2.5%   97.5%
Age      0.67209 0.49201 0.91809
DrugUse  1.81927 1.36511 2.42453
trt      1.39428 1.04434 1.86148
```

or with bootstrap confidence intervals:

```
> prLogisticBootMarg(fit.logistic, data = UIS)

95% Confidence Interval using Bootstrap Method
                      Normal       Percentile
         Estimate  2.5%   97.5%    2.5%   97.5%
Age      0.67209 0.43077 0.87387 0.49765 0.93893
DrugUse  1.81927 1.31717 2.39688 1.32944 2.42177
trt      1.39428 0.97643 1.85325 1.02391 1.96567
```

Considering the results with marginal standardization (population-averaged), the probability of remaining drug free at the end of the treatment program, assuming that all patients were in the long

group, is 39% larger than that same probability when all patients were assumed to be in the short group. For comparison purposes, we plot these estimates (see Figure 1).

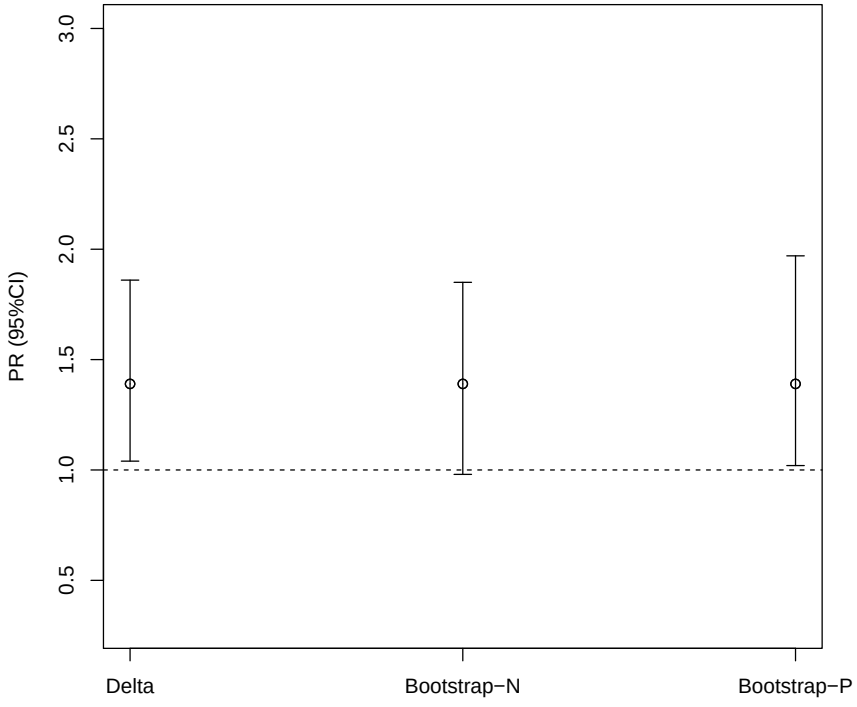

Figure 1. Marginal Prevalence Ratios (PR) and corresponding 95% CIs defined by three methods: (a) Delta; (b) Bootstrap-Normal approximation; (c) Boostrap-Percentile Interval.

Note that 95% CIs are very similar for the three methods for estimating marginal prevalence ratios in this example (see Figure 1). The major difference is that statistical significance is not reached when using bootstrap with Normal approximation.

3.2 - Education in Thailand

Data are from a large national survey on primary education in Thailand, including information for 8,582 sixth graders nested within 411 schools (Raudenbush & Bhumirat 1992). The binary outcome variable "`rgi`" indicates whether a student has repeated a grade during primary education. The predictor variables in the dataset are the child's sex and the child's pre-primary education. Every level-1 record corresponds to a student. The level-2 is defined by schools.

We load the *prLogistic* package and the dataset `Thailand`, and explore the dataset:

```
R> library("prLogistic")
R> data("Thailand")
R> head(Thailand)
```

```
   schoolid sex pped rgi
1     10101   0    1   0
2     10101   0    1   0
3     10101   0    1   0
4     10101   0    1   0
5     10101   0    1   0
```

Dataset Thailand contains the following variables from the original study:

- `schoolid` is the school identification.
- `sex` is the student's sex (1 = boy, 0 = girl).
- `ped` is an indicator for pre-primary education (1 = yes, 0 = no).
- `rgi` is an indicator whether a student has ever repeated a class (1 = yes, 0 = no).

The distribution of the outcome variable (repeated grade indicator – `rgi`) is given by:

```
R> prop.table(table(rgi))

rgi
    0         1
0.8549289 0.1450711
```

Due to clustering, the following random-effects logistic model was fitted:

$$\text{logit}[\text{P}(\texttt{rgi}_{ij}|\texttt{sex}_{ij}, \texttt{pped}_{ij}, u_{0j})] = \beta_0 + \beta_1 \texttt{sex}_{ij} + \beta_2 \texttt{pped}_{ij} + u_{0j}.$$

Using a random intercept logistic model without covariates, the intraclass correlation coefficient (ICC) is 0.33, indicating an important effect of clustering that has to be considered in the analysis. Thus, the random effects logistic model is indicated for this data analysis instead of traditional logistic regression, which assumes the independence of observations.

The conditional PR using delta method can be obtained by adding the option `cluster="TRUE"` in the function `prLogisticDelta` as follows:

```
R> prLogisticDelta(rgi ~ sex + pped + (1|schoolid),
data = Thailand, cluster=TRUE)

95% Confidence Interval using Delta method
      Estimate  2.5%   97.5%
sex   1.61311 1.43065 1.81883
pped  0.56198 0.47788 0.66087
```

To get the marginal estimates using this function, we should also include the option `pattern="marginal"` as showed below:

```
R> prLogisticDelta(rgi ~ sex + pped + (1|schoolid), data = Thailand,
pattern="marginal", cluster=TRUE)

95% Confidence Interval using Delta method
      Estimate  2.5%   97.5%
sex   1.63125 1.44159 1.8459
pped  0.57276 0.48905 0.6708
```

We also can obtain estimates for a different confidence level. For instance, suppose we are interested in computing a 90% confidence interval for PR. In this case we use

```
R> prLogisticDelta(rgi ~  sex + pped + (1|schoolid), data = Thailand,
conf=0.90, cluster=TRUE)

90% Confidence Interval using Delta method
      Estimate   5%     95%
sex   1.61311 1.4585 1.78407
pped  0.56198 0.4905 0.64387
```

For obtaining bootstrap confidence intervals for PR we use

```
R> library("lme4")
R> ML <- lmer(rgi ~  sex + pped  + (1|schoolid),
family = binomial, data = Thailand)
R> prLogisticBootCond(ML, data = Thailand)

95% Confidence Interval using Bootstrap Method
                  Normal        Percentile
      Estimate  2.5%    97.5%   2.5%   97.5%
sex   1.61312 1.30682 1.75578 1.48510 1.94841
pped  0.56198 0.49795 0.66575 0.45234 0.62857
```

or

```
R> prLogisticBootMarg(ML, data = Thailand)

95% Confidence Interval using Bootstrap Method
                   Normal      Percentile
      Estimate  2.5%   97.5%    2.5%   97.5%
sex   1.63126 1.33466 1.77736 1.48676 1.91974
pped  0.57276 0.51853 0.67374 0.45166 0.62215
```

respectively, for conditional and marginal standardization. Alternatively, we could obtain these estimates using the following syntax:

```
R> prLogisticBootCond(lmer(rgi~  sex + pped  + (1|schoolid),
family = binomial, data = Thailand),
```

```
data=Thailand)

95% Confidence Interval using Bootstrap Method
                  Normal        Percentile
      Estimate  2.5%   97.5%    2.5%   97.5%
sex   1.61312 1.34484 1.76708 1.45750 1.88915
pped  0.56198 0.47754 0.68699 0.45217 0.67638
```

Similar syntax could be used with function `prLogisticBootMarg` to obtain estimates with marginal standardization.

We could also modify the number of bootstrap replications and the confidence level using:

```
R> prLogisticBootCond(ML, data = Thailand, conf=0.90, R=45)

90% Confidence Interval using Bootstrap Method
                  Normal        Percentile
      Estimate   5%     95%      5%     95%
sex   1.61312 1.33242 1.74580 1.47331 1.94644
pped  0.56198 0.49857 0.66187 0.44942 0.64782
```

Both predictors are significantly associated to whether a student has repeated a grade during primary education, i.e., in a given school the boys have 63% higher probability of repetition than girls (PR = 1.63 [95% CI = 1.49; 1.92]) and a child has 43% less probability if he/she received pre-primary education (PR = 0.57 [95 %CI = 0.45; 0.62]). These results were based on marginal standardization procedure with bootstrap-percentile 95% confidence interval. Similar conclusions were reached by using any of the methods described here.

4 - CONCLUSION AND FUTURE WORK

We have shown how logistic regression models can be implemented to estimate the prevalence ratios and their confidence intervals using our *prLogistic* package, in situations where the observations are independent or when data comes from clustered studies. The package can accommodate the information of conditional and marginal standardization, commonly used in epidemiology, as well as either delta method or bootstrap resampling for the obtention of confidence intervals. Our package is easily used and does not involve extensive programming. Our contribution of the package *prLogistic* will make these methodologies more accessible to applied researchers. In future updates of the package, the functions will implement generalized estimating equations (GEE), a marginal approach for longitudinal/clustered data. It is, however, worth mentioning that most analysis in Epidemiology and Public Health involves only categorical variables. Future implementation might consider the extension of the procedures to incorporate other types of variables, nonlinear and interaction terms between covariates and include the sampling design via the R package *survey* (Lumley 2004, Oberski 2014).

Acknowledgments

This work was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant 305305/2019-0 and Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) /Brazil. We are also thankful to two anonymous referees whose comments and suggestions led to a much improved manuscript.

REFERENCES

BARROS A & HIRAKATA V. 2003. Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio. BMC Med Res Methodol 3: 21-33.

BASTOS LS, OLIVEIRA RDVCD & VELASQUE LDS. 2015. Obtaining adjusted prevalence ratios from logistic regression models in cross-sectional studies. Cad Saúde Pública 31: 487-495.

BATES D, MÄCHLER M, BOLKER B & WALKER S. 2015. Fitting Linear Mixed-Effects Models Using lme4. J Stat Softw 67(1): 1-48. doi:10.18637/jss.v067.i01.

BATES D, MAECHLER M, BOLKER B, WALKER S, CHRISTENSEN RHB, SINGMANN H, DAI B, SCHEIPL F, GROTHENDIECK G & GREEN P. 2021. Package 'lme4'. R Version 1.1-27.1.

BISHOP YM, FIENBERG SE & HOLLAND PW. 2007. Discrete multivariate analysis: theory and practice. Springer Science & Business Media, p. 481-490.

CANTY AJ. 2002. Resampling methods in R: the boot package. R News 2: 3-8.

CUMMINGS P. 2009. Methods for estimating adjusted risk ratios. Stata J 9(2): 175-196.

DAVISON A & HINKLEY D. 1997. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press, p. 522-540.

DIGGLE P, LIANG K & ZEGER S. 1994. Analysis of Longitudinal Data. New York: Oxford University Press, p. 141-179.

EFRON B & TIBSHIRANI R. 1993. An Introduction to the Bootstrap. New York: Chapman & Hall, p. 45-57.

FOX J. 2015. Applied regression analysis and generalized linear models. Sage Publications, p. 647-668.

GREENLAND S. 2004. Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies. Am J Epidemiol 160(4): 301-305.

HOSMER JR DW, LEMESHOW S & STURDIVANT RX. 2013. Applied logistic regression. vol. 398. J Wiley & Sons, p. 104-116.

HOX JJ, MOERBEEK M & VAN DE SCHOOT R. 2017. Multilevel analysis: Techniques and applications. Routledge, p. 103-147.

LANE P & NELDER J. 1982. Analysis of covariance and standardization as instances of prediction. Biometrics 38: 613-621.

LARSEN K, PETERSEN JH, BUDTZ-JØRGENSEN E & ENDAHL L. 2000. Interpreting Parameters in the Logistic Regression Model with Random Effects. Biometrics 56: 909-914.

LOCALIO AR, MARGOLIS DJ & BERLIN JA. 2007. Relative risks and confidence intervals were easily computed indirectly from multivariate logistic regression. J Clin Epidemiol 60: 874-882.

LUMLEY T. 2004. Analysis of complex survey samples. J Stat Softw 9(1): 1-19.

MCCULLOCH C & SEARLE S. 2001. Generalized, Linear, and Mixed Models. New York: Wiley & Sons Inc, p. 57-68.

MCNUTT LA, WU C, XUE X & HAFTNER JP. 2003. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. Am J Epidemiol 157(10): 940-943.

NEWCOMBE RG. 2006. A deficienty of the odds ratio as a measure of effect size. Stat Med 25: 4235-4240.

OBERSKI D. 2014. lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models. J Stat Softw 57(1): 1-27. doi:10.18637/jss.v057.i01.

OLIVEIRA NFD, SANTANA VS & LOPES AA. 1997. Ratio of proportions and the use of the delta method for confidence interval estimation in logistic regression. Rev Saúde Pública 31: 90-99.

OSPINA R & AMORIM LD. 2021. *prLogistic:* Estimation of Prevalence Ratios using Logistic Models. URL https://cran.r-project.org/package=prLogistic.
R package version 1.2.

PETERSEN M & DEDDENS J. 2008. A comparison of two methods for estimating prevalence ratios. BMC Med Res Methodol 8: 9-18.

R DEVELOPMENT CORE TEAM. 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL http://www.R-project.org.

RAUDENBUSH SW & BHUMIRAT C. 1992. The distribution of resources for primary education and its consequences for educational achievement in Thailand. J Educ Res 44: 143-164.

SANTOS C, FIACCONE R, OLIVEIRA N, CUNHA S, BARRETO M, DO CARMO M, MONCAYO A, RODRIGUES L, COOPER P & AMORIM L. 2008. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. BMC Med Res Methodol 8(1): 80.

SAVU A, LIU Q & YASUI Y. 2010. Estimation of relative risk and prevalence ratio. Stat Med 29: 2269-2281.

STEVENSON M, NUNES T, SANCHEZ J, THORNTON R, REICZIGEL J, ROBISON-COX J & SEBASTIANI P. 2013. epiR: An R package for the analysis of epidemiological data. R package version 09 - 43, p. 1-197.

STROMBERG U. 1994. Prevalence odds ratio versus prevalence ratio. Occupational Environmental Medicine 51: 143-144. doi:10.1136/oem.51.2.143.

TAMHANE AR, WESTFALL AO, BURKHOLDER GA & CUTTER GR. 2016. Prevalence odds ratio versus prevalence ratio: choice comes with consequences. Stat Med 35(30): 5730-5735.

THOMPSON ML, MYERS JE & KRIEBEL D. 1998. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med 55(4): 272-277.

WILCOSKY T & CHAMBLESS L. 1985. A comparison of direct adjustment and regression adjustment of epidemiologic measures. J Chronic Dis 34: 849-856.

ZHANG J & YU K. 1998. What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. JAMA 280: 1690-1691.

How to cite
AMORIM LD & OSPINA R. 2021. Prevalence ratio estimation via logistic regression: a tool in R. An Acad Bras Cienc 93: e20190316. DOI 10.1590/0001-3765202120190316.

Manuscript received on March 21, 2019;
accepted for publication on July 11, 2019

LEILA D. AMORIM[1]
https://orcid.org/0000-0002-1112-2332

RAYDONAL OSPINA[2]
https://orcid.org/0000-0002-9884-9090

[1]Departamento de Estatística, Universidade Federal da Bahia, Instituto de Matemática e Estatística, Av. Ademar de Barros, s/n, Campus de Ondina, 40170-110 Salvador, BA, Brazil
[2]Departamento de Estatística, CASTLab, Universidade Federal de Pernambuco, Cidade Universitária, Av. Prof. Moraes Rego, 1235, 50740-540 Recife, PE, Brazil

Correspondence to: **Leila Amorim**
E-mail: leiladen@ufba.br

Author contributions

The current paper was jointly developed by the two authors. The first author proposed the research topic and the prototype scripts and drafted the paper. The package and analytical implementations were derived by the second author and checked by the first author. The empirical application was carried out jointly by the two authors. The manuscript was written by the two authors.