Eur J Epidemiol (2016) 31:563–574
DOI 10.1007/s10654-016-0157-3

METHODS

Regression standardization with the R package stdReg

Arvid Sjölander[1]

Received: 9 February 2016 / Accepted: 30 April 2016 / Published online: 14 May 2016

Abstract When studying the association between an exposure and an outcome, it is common to use regression models to adjust for measured confounders. The most common models in epidemiologic research are logistic regression and Cox regression, which estimate conditional (on the confounders) odds ratios and hazard ratios. When the model has been fitted, one can use regression standardization to estimate marginal measures of association. If the measured confounders are sufficient for confounding control, then the marginal association measures can be interpreted as poulation causal effects. In this paper we describe a new R package, stdReg, that carries out regression standardization with generalized linear models (e.g. logistic regression) and Cox regression models. We illustrate the package with several examples, using real data that are publicly available.

Keywords Cox regression · Hazard ratio · Logistic regression · Odds ratio · Standardization

Introduction

When studying the association between an exposure and an outcome, it is common to use regression models to adjust for measured confounders. The most common models in epidemiologic research are logistic regression (for binary outcomes) and Cox regression (for time-to-event outcomes). These models are powerful and flexible, and can be adapted to various situations and sampling schemes, e.g. case–control studies, matched cohort studies and case-cohort studies.

✉ Arvid Sjölander
arvid.sjolander@ki.se

[1] Nobels väg 12 A, 171 77 Stockholm, Sweden

Logistic regression and Cox regression estimate conditional (on the confounders) odds ratios and hazard ratios, respectively. When the model has been fitted, one can use regression standardization to estimate marginal measures of association (see Rothman et al. [1, pp. 386–388, 442–445] and the references therein). This method uses the regression model to predict the risk of the outcome or the survival function, for exposed and unexposed separately, at every observed level of the measured confounders. Then, these predictions are averaged over a 'standard' confounder distribution to produce a standardized risk or survival function, for exposed and unexposed separately. Standardized survival functions are sometimes referred to as 'direct adjusted survival curves' [2]. A natural choice of 'standard' confounder distribution is the observed distribution in the sample. Finally, the standardized risks/survival functions for exposed and unexposed can be contrasted to produce standardized measures of association. If the measured confounders are sufficient for confounding control, then these standardized measures of association can be interpreted as population causal effects; they apply to a population with the 'standard' confounder distribution [1].

An appealing feature of regression standardization is that, although the underlying models estimate odds ratios and hazard ratios, the standardized measures are not restricted to these contrasts. For instance, we can use logistic regression to estimate a standardized risk difference, and we can use Cox regression to estimate a standardized difference in 5 year survival.

Another appealing feature of regression standardization is that standardized measures have simple interpretations even though the underlying models are complex. In the conventional use of logistic/Cox regression it is common to assume that the conditional odds/hazard ratio is constant across levels of the measured confounders. Arguable, this

assumption is typically not made because it is believed to hold true, but because it implies that the adjusted exposure-outcome association can be conveniently represented by a single number. In principle, the assumption can be relaxed by including interaction terms between the exposure and the measured confounders, but this makes interpretation more complicated and is therefore rarely done in practice. However, since the standardized risk difference is averaged over the confounder distribution, it remains a single number even when there are exposure-confounder interactions in the underlying model. Thus, standardization relieves the analyst from some of the pressure of having to rely on unrealistically simple models.

Despite its appeal, regression standardization is not commonly used in epidemiologic studies. We believe that this is largely due to a lack of software implementation. In this paper we present a new `R` package, `stdReg`, which carries out regression standardization with generalized linear models and Cox regression models [3]. The package is available on CRAN's webpage. The paper is organized as follows. In 'Standardization with generalized linear models' section we consider regression standardization with generalized linear models, e.g. logistic regression. In 'Standardization in case–control studies' section we consider regression standardization in case–control studies, based on logistic regression models. In 'Standardization with Cox regression models' section we consider regression standardization with Cox regression models. These sections are split into subsections labelled 'theory' and 'practice'. In the 'theory' subsections we review the theory behind regression standardization, and in the 'practice' subsections we illustrate through practical examples how regression standardization can be carried out with the `stdReg` package.

Throughout we use examples with publicly available datasets, so that the reader can replicate and elaborate on all analyses. All datasets are borrowed from the `AF` package [4]; we refer to the help files for this package and the references therein for a thorough description of the data. We assume that the reader has some familiarity with `R`, in particular with the functionality for fitting generalized linear models and Cox regression models.

Standardization with generalized linear models

Theory

Let X and Y be the exposure and outcome of interest, respectively, and let x be a specific (fixed) exposure level. Let Z be a vector of measured confounders that we wish to adjust for in the analysis. Let $E(Y|X, Z)$ be the conditional mean of Y, given X and Z. The standardized mean of Y, at $X = x$, is defined as

$$\theta(x) = E\{E(Y|X = x, Z)\},$$

where the outer mean is taken over the marginal (population) distribution of Z. If Z is sufficient for confounding control, then $\theta(x)$ can be interpreted as the counterfactual mean outcome that we would have observed, had everybody in the population been exposed to level $X = x$ [1]. Standardized means at different levels of X can be contrasted to form standardized effects. For instance, if Z is sufficient for confounding control and Y is binary (0/1), then the contrast $\theta(x_1) - \theta(x_0)$ can be interpreted as a causal risk difference, comparing levels $X = x_1$ and $X = x_0$. In practice, one would rarely believe that a set of measured confounders is sufficient for confounding control. However, in the presence of unmeasured confounding we may still view (contrasts of) $\theta(x)$ as a useful summary measure of the exposure-outcome association.

A generalized linear model is a model for $E(Y|X, Z)$ on the form

$$\eta\{E(Y|X, Z)\} = h(X, Z; \beta) \quad (1)$$

where $\eta(\cdot)$ is a suitable link function. The identity link gives linear regression, the log link gives Poisson regression, and the logit link gives logistic regression. The parametric function $h(X, Z; \beta)$ is often assumed to be a linear in X and Z, e.g $\eta\{E(Y|X, Z)\} = \beta_0 + \beta_1 X + \beta_2 Z$, but it may also, for instance, include interaction terms, higher order terms or splines. Provided that the sample is taken randomly from the population, a generalized linear model can be used estimate standardized means as follows. First, the model is fitted to obtain an estimate of the parameter vector β. Then, for each subject i with confounder vector Z_i, $i = 1, 2, \ldots, n$, we use $\eta^{-1}\{h(X = x, Z_i; \hat{\beta})\}$ as a prediction of $E(Y|X = x, Z_i)$. Finally, these predictions are averaged to obtain an estimate of $\theta(x)$:

$$\hat{\theta}(x) = \sum_{i=1}^{n} \eta^{-1}\{h(X = x, Z_i; \hat{\beta})\}/n. \quad (2)$$

In Appendix 1 we use the 'sandwich formula' [5] to derive the asymptotic distribution for the estimate $\hat{\theta}(x)$, and for estimated contrasts such as $\hat{\theta}(x_1) - \hat{\theta}(x_0)$.

Practice

In this section we use the dataset `clslowbwt`. This dataset contains information on 487 births among 188 women. Here we are interested in the following variables: `lbw` (a binary indicator of whether the newborn has low birthweight, defined as birthweight $\leq$2500 g) `smoker` (a binary indicator of whether the mother smoked during pregnancy), `race` (race of the mother, coded as white, black or other), `age` (age of the mother), and `id` (a unique

identification number for each mother). We aim to estimate the association between smoking during pregnancy and low birthweight, adjusted for mother's race and age.

A conventional analysis fits the logistic regression model

$$\begin{aligned}&\text{logit}\{\Pr(\texttt{lbw}=1|\texttt{smoker},\texttt{race},\texttt{age})\}\\&\quad=\beta_0+\beta_1\texttt{smoker}+\beta_2 I(\texttt{race}=\text{''}\texttt{black}\})\\&\qquad+\beta_3 I(\texttt{race}=\}\texttt{other}\})+\beta_4\texttt{age},\end{aligned}\tag{3}$$

where $I(A)$ is the indicator variable taking value 1 if A is true, 0 otherwise. In R, the model is fitted by typing

```
> fit1 <- glm(formula=lbw~smoker+race+age, family="binomial",
   data=clslowbwt)
```

which gives the output

```
> summary(fit1)

             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.35946    0.53281  -2.551  0.01073 *
smoker        0.60080    0.21693   2.770  0.00561 **
race2. Black -0.85852    0.32331  -2.655  0.00792 **
race3. Other -0.70449    0.24624  -2.861  0.00422 **
age           0.02399    0.01785   1.344  0.17900
```

The output indicates that the odds of having low birthweight is about $\exp(0.6) \approx 1.8$ times higher among children born to smokers as compared to children born to non-smokers.

The model in (3) is perhaps unrealistically simple, since it assumes that the effect of smoking is the same regardless of age and race. Thus, we next consider the following model, which includes both main effects and all pair-wise interactions between the predictors:

$$\begin{aligned}&\text{logit}\{\Pr(\texttt{lbw}=1|\texttt{smoker},\texttt{race},\texttt{age})\}\\&\quad=\beta_0+\beta_1\texttt{smoker}+\beta_2 I(\texttt{race}=\}\texttt{black}\})\\&\qquad+\beta_3 I(\texttt{race}=\}\texttt{other}\})+\beta_4\texttt{age}\\&\qquad+\beta_5\texttt{smoker}\times I(\texttt{race}=\}\texttt{black}\})+\beta_6\texttt{smoker}\\&\qquad\times I(\texttt{race}=\}\texttt{other}\})+\beta_7\texttt{smoker}\times\texttt{age}\\&\qquad+\beta_8 I(\texttt{race}=\}\texttt{black}\})\times\texttt{age}\\&\qquad+\beta_9 I(\texttt{race}=\}\texttt{other}\})\times\texttt{age}.\end{aligned}\tag{4}$$

The model is fitted by typing

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
   family="binomial", data=clslowbwt)
```

which gives the output

```
> summary(fit2)

                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.91927    0.99462   1.930 0.053649 .
smoker              -3.81160    1.16321  -3.277 0.001050 **
race2. Black        -4.79348    1.69498  -2.828 0.004683 **
race3. Other        -2.99103    1.28846  -2.321 0.020265 *
age                 -0.08611    0.03519  -2.447 0.014415 *
smoker:race2. Black  1.43758    0.76748   1.873 0.061050 .
smoker:race3. Other  1.00778    0.54283   1.857 0.063377 .
smoker:age           0.14763    0.04153   3.555 0.000378 ***
race2. Black:age     0.10954    0.06340   1.728 0.084040 .
race3. Other:age     0.06936    0.04706   1.474 0.140518
```

We observe that the interaction term between smoker and age is highly significant, with a p value <0.001. Thus, it seems like the model in (3), which only includes main effects, may indeed be too simplistic. On the other hand, the more realistic model in (4) is harder to interpret and communicate, since the smoking effect in this model is captured by four parameters (one main effect plus three interaction terms). We note that the standard errors and p-values in the output may not be entirely correct, as the glm function assumes that all observations are independent. This assumption is not likely to hold in the clslowbwt dataset, since some women have given birth to multiple children.

To perform regression standardization with the model in (4), we use the function stdGlm from the stdReg package. The function takes a fitted model as input, together with the data frame that was used to fit the model, and standardizes to the confounder distribution in the data frame. This is done by typing

```
> fit.std <- stdGlm(fit=fit2, data=clslowbwt, X="smoker",
   clusters="id")
```

The argument X is mandatory, and specifies the name of the exposure variable. The argument clusters is optional, and specifies the name of a cluster identification variable. By specifying this argument we ensure that the standard errors of the estimates are corrected for within-cluster correlations. The results are summarized by typing

```
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0    0.279     0.0406    0.199    0.358
1    0.407     0.0555    0.298    0.516
```

This produces a table with the standardized risk for low birthweight, for the unexposed (smoker=0) and the exposed (smoker=1), together with standard errors and 95 % confidence intervals. If race and age are sufficient for confounding control, we may conclude that 28 % of all newborns would have had low birthweight, had no mother smoked, and that 41 % of all newborns would have had low birthweight, had all mothers smoked.

To obtain the standardized risk difference we type

```
> summary(fit.std, contrast="difference", reference=0)

  Estimate Std. Error lower 95 upper 95
0    0.000     0.0000  0.00000    0.000
1    0.128     0.0681 -0.00544    0.262
```

which gives the risk difference for the unexposed and exposed, using unexposed as the reference level. The contrast argument can also be set to ''ratio'', which then gives the standardized risk ratio. The summary function has an additional argument, transform, which allows for a log-, logit-, and odds-transformation of the standardized risks. The transformation is applied before taking the contrast. By combining the contrast and transform arguments, all common (and some less common) association measures can be obtained. For instance, to obtain the standardized odds ratio we type

```
> summary(fit.std, transform="odds", contrast="ratio",
   reference=0)

  Estimate Std. Error lower 95 upper 95
0     1.00      0.000    1.000     1.00
1     1.77      0.537    0.721     2.83
```

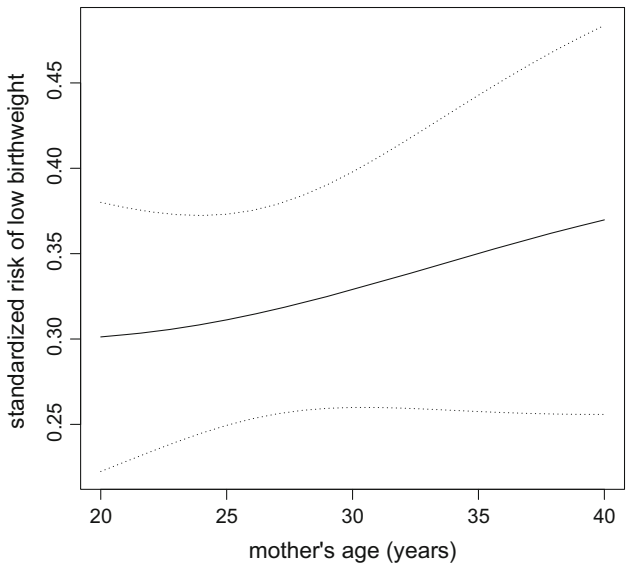

Fig. 1 Standardized risk of low birthweight as a function of mother's age

The stdGlm is not limited to binary exposures. Suppose that we want to fit the model in (4), but consider age as the exposure. We then type

```
fit.std2 <- stdGlm(fit=fit2, data=clslowbwt, X="age",
   clusters="id", x=20:40)
```

The x argument specifies at what exposure levels to standardize; we here chose to standardize at ages 20 to 40, in steps of 1 year. The x argument defaults to 'all levels' when the exposure is categorical/binary. When the exposure is continuous, it may be more convenient to summarize the results in a plot. This can be done by typing

```
plot(fit.std2, xlab="mother's age (years)",
   ylab="standardized risk of low birthweight")
```

which gives the plot in Fig. 1. This plot shows that the risk of low birthweight (solid line) increases from 0.3 at age 20, to 0.37 at age 40, in an almost linear fashion. By default, the plot displays 95 % pointwise confidence intervals (dashed lines). Like the summary function, the plot function has optional arguments transform and contrast, which can be used to plot, for instance, the standardized risk difference and the odds ratio as a function of age.

We end this section by noting that, although we have restricted attention to logistic regression, other models can be used as well. The stdGlm function accepts all models that have been fitted with the glm function, e.g. linear and log-linear (Poisson) models.

Standardization in case–control studies

Theory

Standardization is more difficult when data are collected under a 'biased' sampling scheme, as in case–control studies. If the sampling probabilities are known, then standardization can be carried out as described above, by using weights that 'corrects' for the sampling scheme. These weights should be used both when fitting the generalized linear model in (1) and when averaging the predictions in (2). In case–control studies, the generalized linear model in (1) would typically be a logistic regression, but other choices are possible, e.g. probit regression.

Different weights are possible, and the choice of weights may affect the efficiency of the resulting estimators [7]. We here consider weights for matched case–control studies that are conceptually simple and easy to compute. Let $p(Y = 1)$ and $p^*(Y = 1)$ be the population and sample probability of the outcome, respectively. Let M_i be the value of the matching variable observed for subject i. Let $p(M_i|Y = 0)$ and $p(M_i|Y = 1)$ be the probability of M_i among the controls and the cases in the population, respectively. When the controls are matched to the cases, we assign the following weight to subject i:

$$w_i = \begin{cases} \dfrac{p(Y=1)}{p^*(Y=1)} & \text{if subject } i \text{ is a case} \\ \dfrac{p(Y=0)}{p^*(Y=0)}\dfrac{p(M_i|Y=0)}{p(M_i|Y=1)} & \text{if subject } i \text{ is a control} \end{cases}$$

These weights are mathematically equivalent to the weights proposed in Theorem 3 by van der Laan [7]. Weights for unmatched case–control studies can be obtained as a special case, by setting the ratio $p(M_i|Y = 0)/p(M_i|Y = 1)$ equal to 1.

Practice

In this section we use the dataset `singapore`. This dataset contains information on 80 male cases of oesophageal cancer, who were collected from hospitals in Singapore during 1970–1972. Each case was individually matched to 4 controls on age within 5 year intervals. We are interested in the following variables: `Oesophagealcancer` (the binary case–control indicator), `Everhotbev` (a binary indicator of whether the subject drinks beverages at 'burning hot' temperatures on a daily basis), `Age` (the subject's age), `Dial` (dialect group; 1 for Hokhien/Teochew and 0 for Cantonese/Other), `Samsu` (a binary indicator of whether the subject consumes Samsu wine on a daily basis), `Cigs` (number of cigarettes smoked per day), and `Set` (a unique identification number for each matched set). We aim to estimate the association between intake of beverages at 'burning hot' temperatures and oesophageal cancer, adjusted for age, dialect group, intake of Samsu wine and smoking.

A conventional analysis fits the conditional regression model

$$\begin{aligned}&\texttt{logit}\{\texttt{Pr}(\texttt{Oesophagealcancer}=1|\texttt{Everhotbev},\\&\texttt{Dial},\texttt{Samsu},\texttt{Cigs},\texttt{Set})\}\\&=\beta_{\texttt{Set}}+\beta_1\texttt{Everhotbev}+\beta_2\texttt{Age}+\beta_3\texttt{Dial}\\&\quad+\beta_4\texttt{Samsu}+\beta_5\texttt{Cigs},\end{aligned}$$

where $\beta_{\texttt{Set}}$ is a set-specific intercept. The variable `Age` is 'absorbed' by the intercept, since the study is matched on age. In `R`, the model is fitted by the `clogit` function in the `survival` package:

```
> fit1 <- clogit(formula=Oesophagealcancer~Everhotbev
   +Dial+Samsu+Cigs+strata(Set), data=singapore)
```

which gives the output

```
> summary(fit1)

              coef exp(coef) se(coef)     z Pr(>|z|)
Everhotbev 1.17646   3.24286  0.28807 4.084 4.43e-05 ***
Dial       1.22815   3.41492  0.32024 3.835 0.000126 ***
Samsu      0.48273   1.62049  0.28588 1.689 0.091303 .
Cigs       0.01102   1.01108  0.00956 1.152 0.249263
```

The output indicates that the odds of getting oesophageal cancer is about 1.18 times higher for those who drink beverages at 'burning hot' temperatures as compared to those who do not.

To carry out standardization we use weights, as descibed in 'Theory' section. From the 1:4 matched design, we have that $p^*(Y=1)$ is equal to 1/5. The incidence of oesophageal cancer, in the population of male Chinese in Singapore, was 19.3 per 100,000 person-years at the time when data were collected [8]; we use this as an estimate of $p(Y=1)$. We note that by using the incidence at this step in the weights, the standardized risks that we obtain should be interpreted as incidences as well. Due to the matched design, $p(M_i|Y=1)$ equals the probability of M_i among the cases in the sample, up to sampling variability. We don't know the age distribution among the controls in the population, so we vary $p(M_i|Y=0)$ in a sensitivity analysis. Towards this end we assume that age has a normal distribution in the population, in both cases and controls. We further assume that these normal distributions have the same standard deviation, s, but possibly different means, equal to m and $m-d$, respectively. The parameter d measures how much younger, on average, the controls are in the population as compared to the cases. Under these assumptions, the ratio $p(M_i|Y=0)/p(M_i|Y=1)$ becomes a ratio of two normal densitites evaluated at M_i, with standard deviation equal to s and means equal to $m-d$ and m, respectively. We vary d over a range of values and carry

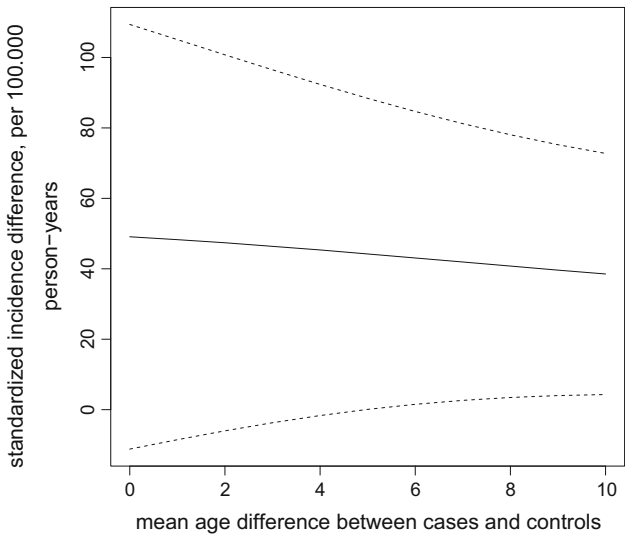

Fig. 2 Standardized incidence difference as a function of the population mean difference in age between cases and controls

out the standardization for each value separately. The following code shows the analysis for $d = 5$. As before we use a model that includes both main effects and all pairwise interactions between the predictors.

```
> case <- singapore$Oesophagealcancer
> ctrl <- 1-case
> p <- 19.3/100000
> p.star <- 1/5
> Mi <- singapore$Age
> m <- mean(Mi)
> s <- sd(Mi)
> d <- 5
> weights <- case*p/p.star+
   ctrl*(1-p)/(1-p.star)*dnorm(Mi,m-d,s)/dnorm(Mi,m,s)
> fit2 <- glm(formula=Oesophagealcancer~(Everhotbev+Age+
   Dial+Samsu+Cigs)^2, family="binomial", data=singapore,
   weights=weights)
> fit.std <- stdGlm(fit=fit2, X="Everhotbev", data=singapore,
   clusters="Set", case.control=TRUE)
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0 0.000128   1.69e-05 9.48e-05 0.000161
1 0.000570   2.19e-04 1.41e-04 0.000999
```

The standardized incidences of oesophageal cancer are 12.8 and 57.0 cases per 100,000 person-years, for those who do and do not drink beverages at 'burning hot' temperatures, respectively.

We make some remarks. First, note that we have used an ordinary logistic regression model here, with explicit adjustment for the matching variable `age`, instead of a conditional logistic regression model. The reason for this is that conditional logistic regression does not produce estimates of the set-specific intercepts, which hampers standardization. Second, note that there is no need (or possibility) to explicitly input the weights to the `stdGlm` function; it retrieves the weights from the fitted model (the `fit2` object), and uses them in the standardization. Third, note the argument `case.control` in the `stdGlm` function. Setting this to `TRUE` does not affect the estimates, but corrects the standard errors for the fact that the number and cases and controls are fixed by design [9, Appendix E].

The plot in Fig. 2 shows the standardized incidence difference as a function of the population mean difference in age between cases and controls, together with 95 % pointwise confidence intervals. We observe that the incidence difference is not overly sensitive to the population mean difference in age; it decreases from 49 to 39 cases per 100,000 person-years when the mean difference is increased from 0 to 10 years. The 95 % confidence intervals are very wide, implying that the uncertainty in the incidence difference is dominated by sampling variability.

Standardization with Cox regression models

Theory

As before, let X and Z be the exposure and measured confounders, respectively, and let x be a specific (fixed)

exposure level. Let T be a time-to-event outcome, e.g. time to death. Let $S(t|X, Z)$ and $\lambda(t|X, Z)$ be the conditional survival function and hazard function, given X and Z, respectively. The standardized survival function, at $X = x$, is defined as

$$\theta(t, x) = E\{S(t|X = x, Z)\},$$

where the outer mean is taken over the marginal (population) distribution of Z. If Z is sufficient for confounding control, then $\theta(t, x)$ can be interpreted as the counterfactual survival function that we would have observed, had everybody in the population been exposed to level $X = x$ [1]. Standardized survival functions at different levels of X can be contrasted to form standardized effects. For instance, if Z is sufficient for confounding control, then the contrast $\theta(t, x_1) - \theta(t, x_0)$ can be interpreted as a causal survival function difference, comparing levels $X = x_1$ and $X = x_0$. In practice, one would rarely believe that a set of measured confounders is sufficient for confounding control. However, in the presence of unmeasured confounding we may still view (contrasts of) $\theta(t, x)$ as a useful summary measure of the exposure-outcome association.

A Cox regression model is a model for $\lambda(t|X, Z)$ on the form

$$\lambda(t|X, Z) = \lambda_0(t)\exp\{\mathrm{h}(\mathrm{X}, \mathrm{Z}; \beta)\},$$

where $\lambda_0(t)$ is the unspecified bazeline hazard. Often, the parametric function $h(X, Z; \beta)$ is assumed to be a linear in X and Z, e.g $\lambda(t|X, Z) = \lambda_0(t)\exp(\beta_0 + \beta_1 \mathrm{X} + \beta_2 \mathrm{Z})$, but it may also, for instance, include interaction terms, higher order terms or splines. Provided that the sample is taken randomly from the population, a Cox regression model can be used estimate standardized survival functions as follows. First, the model is fitted to obtain an estimate of the parameter vector β. An estimate of the cumulative baseline hazard $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is obtained by Breslow's estimator [10]. Then, for each subject i with confounder vector Z_i, $i = 1, 2, \ldots, n$, we use $\exp[-\hat{\Lambda}_0(t)\exp\{\mathrm{h}(\mathrm{X} = \mathrm{x}, \mathrm{Z_i}; \hat{\beta})\}]$ as a prediction of $S(t|X = x, Z_i)$. Finally, these predictions are averaged to obtain an estimate of $\theta(t, x)$:

$$\hat{\theta}(t, x) = \sum_{i=1}^{n} \exp[-\hat{\Lambda}_0(\mathrm{t})\exp\{\mathrm{h}(\mathrm{X} = \mathrm{x}, \mathrm{Z_i}; \hat{\beta})\}]/\mathrm{n}.$$

In Appendix 1 we use the 'sandwich formula' [5] to derive the asymptotic distribution for the estimate $\hat{\theta}(t, x)$, and for estimated contrasts such as $\hat{\theta}(t, x_1) - \hat{\theta}(t, x_0)$.

Practice

In this section we use the dataset `rott2`. This dataset contains information on 2982 women with primary breast cancer from the Rotterdam tumor bank in the Netherlands. The follow-up time ranges from 1 to 231 months. We are interested in the following variables: `rf` (the time, measured in months, that the patient is under study), `rfi` (an indicator of whether the patient experienced death or relapse before censoring), `chemo` (an indicator of whether the patient received chemotherapy, coded as 'yes' or 'no'), `age` (the patient's age at surgery), `meno` (menopausal status, coded as 0 for pre and 1 for post), `size` (tumor size in three classes: '$\leq$20mm', '>20–50 mm' and '$>$ 50mm'), `grade` (tumor grade; 2 or 3), `nodes` (the number of positive lymph nodes, ranging from 0 to 34), `pr` (progesterone receptors, fmol/l), and `er` (oestrogen receptors, fmol/l). We aim to estimate the association between chemotherapy and relapse-free survival time, adjusted for age, menopausal status, tumor size, tumor grade, lymph nodes, progesterone and oestrogen receptors.

A conventional analysis fits the Cox regression model

$$\begin{aligned}&\lambda(t|\mathtt{chemo}, \mathtt{age}, \mathtt{meno}, \mathtt{size}, \mathtt{grade}, \mathtt{nodes}, \mathtt{pr}, \mathtt{er})\\ &= \lambda_0(t)\exp\{\beta_1 \mathtt{chemo} + \beta_2 \mathtt{age} + \beta_3 \mathtt{meno} + \beta_4 \mathtt{size}\\ &+ \beta_5 \mathtt{grade} + \beta_6 \exp(-0.12\mathtt{nodes}) + \beta_7 \mathtt{pr} + \beta_8 \mathtt{er}\}.\end{aligned}$$

In this model we used the transformation $\exp(-0.12\mathtt{nodes})$, since previous analyses of this dataset have shown that this transformation gives a better model fit [11]. In `R`, the model is fitted by the `coxph` function in the `survival` package:

```
fit1 <- coxph(formula=Surv(rf,rfi)~chemo+age+meno+size+
   grade+I(exp(-0.12*nodes))+pr+er, data=rott2)
```

which gives the output

```
> summary(fit1)

                       exp(coef) exp(-coef) lower .95 upper .95
chemoyes                  0.7546     1.3251    0.6550    0.8695
age                       0.9842     1.0161    0.9774    0.9910
menopre                   0.8761     1.1414    0.7352    1.0441
size>20-50mmm             1.3346     0.7493    1.1899    1.4969
size>50mm                 1.6172     0.6184    1.3581    1.9258
grade                     1.4153     0.7066    1.2460    1.6076
I(exp(-0.12 * nodes))     0.1567     6.3811    0.1293    0.1899
pr                        0.9999     1.0001    0.9997    1.0001
er                        0.9999     1.0001    0.9997    1.0002
```

The output indicates that the hazard of death or relapse is about 1.33 times higher for those who did not recieve chemotherapy as compared to those who did recieve chemotherapy.

To perform regression standardization with a Cox model we use the function `stdCox` from the `stdReg` package. This function estimates standardized survival functions, based on a user-specified Cox regression model. The function takes a fitted model as input, together with the data frame that was used to fit the model, and standardizes to the observed confounder distribution in the data frame. We use a model that includes both main effects and all pair-wise interactions between the predictors. The model is fitted by typing

```
> fit2 <- coxph(formula=Surv(rf,rfi)~(chemo+age+meno+
   size+grade+I(exp(-0.12*nodes))+pr+er)^2, data=rott2)
```

and the standardization is carried out by typing

```
> fit.std <- stdCoxph(fit=fit2, data=rott2, X="chemo")
```

The argument `X` is mandatory, and specifies the name of the exposure variable. There is an optional argument `t` that we have not used here. This argument specifies a vector of time points at which standardization is carried out. It defaults to all observed event times. When there are many observed event times, the computing time can be dramatically reduced by specifying a limited number of time point through the `t` argument. In addition, there is an optional argument `clusters`, which should be specified for clustered data.

It is convenient to summarize the results in a plot. This is done by typing

```
> plot(fit.std, xlab="time (months)",
   ylab="standardized survival probability")
```

which gives the plot in Fig. 3. This plot shows that the standardized survival function for those who did receive chemotherapy (red solid line) and did not receive chemotherapy (black solid line), together with pointwise 95% confidence intervals (dashed lines). The survival for those who did receive chemotherapy is well above the survival for those who did not. For instance, the standardized survival probabilities at 5 years (60 months) are 65 and 57 % for those who did and did not receive chemotherapy, respectively.

To plot the difference in standardized survival functions, with 'no chemotherapy' as the reference, we type

```
plot(fit.std, contrast="difference", reference="no",
   xlab="time (months)",
   ylab="difference in standardized survival probability")
```

which gives the plot in Fig. 4. We observe that the 95% confidence interval for the difference includes 0 everywhere, so that the observed difference is not statistically significant.

Discussion

In this paper we have described a new `R` package for regression standardization; `stdReg`. The `stdReg` package allows for standardization with generalized linear models and Cox regression models, which are the most commonly used models in epidemiologic research. We have demonstrated that it can accomodate various sampling schemes, such as longitudinal studies with repeated measures and case–control studies. We have made an effort to optimize the code so that it runs fast and smoothly, even for large datasets with millions of observations. In particular, all standard errors are coded with analytic formulas based on the theory for M-estimation and the 'sandwich formula' [5], thereby avoiding time consuming numerical approximations and bootstrap procedurs.

The `stdReg` package standardizes to the observed confounder distribution in the sample. This is a natural choice, which leads to a simple interpretation of the standardized measures as population causal effects (provided that the observed confounders are sufficient for confounding control). Other choices could be relevant though, for instance, when one wants to study the effect in a subpopulation, e.g. females. In this case it would be appropriate to standardize to the confounder distribution in the subpopulation of interest. We plan to extend the `stdReg` package in the future, to allow for different choices of 'standard' confounder distributions.

The `stdReg` package carries out standardization with regression models for the outcome. Standardization can also be carried out with regression models for the exposure [12, 13]. A more sophisticated approach combines an outcome model with an exposure model to produce a doubly robust estimator of the standardized measure of interest [14, 15]. A possible extension of the package is to allow for standardization with exposure models and doubly robust estimation.

The `stdCoxph` function only allows for time-stationary exposures and confounders. When the exposure and confounders are time-varying, usual regression techniques do not estimate parameters that can be interpreted as causal effects, even when the measured confounders are sufficient for confounding control [16]. Instead, standardization must be based on the g-formula [16], which is currently not implemented in the `stdReg` package.

An appealing feature of standardized measures is that they have a simple interpretation even though the underlying model is complex. To illustrate this, we have consistently used models with both main effects and all pair-wise interactions between the predictors. However, as the number of parameters in the underlying model increases, the statistical uncertainty in the standardized measures increases as well. Indeed, in all our examples

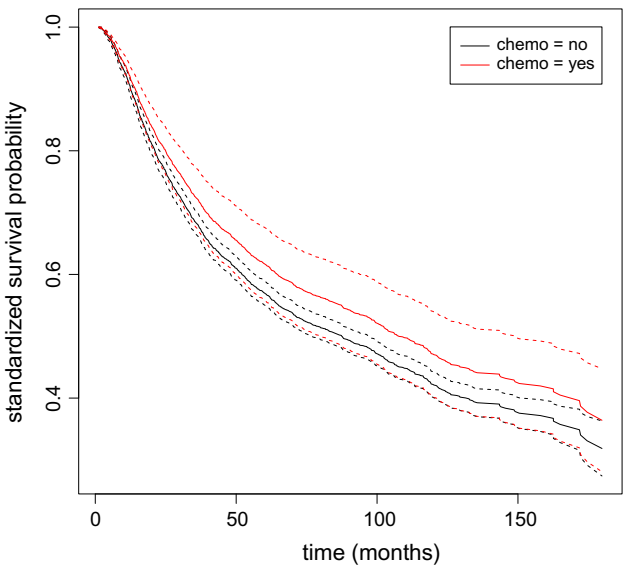

Fig. 3 Standardized survival functions for those who did receive chemotherapy (*red solid line*) and did not receive chemotherapy (*black solid line*)

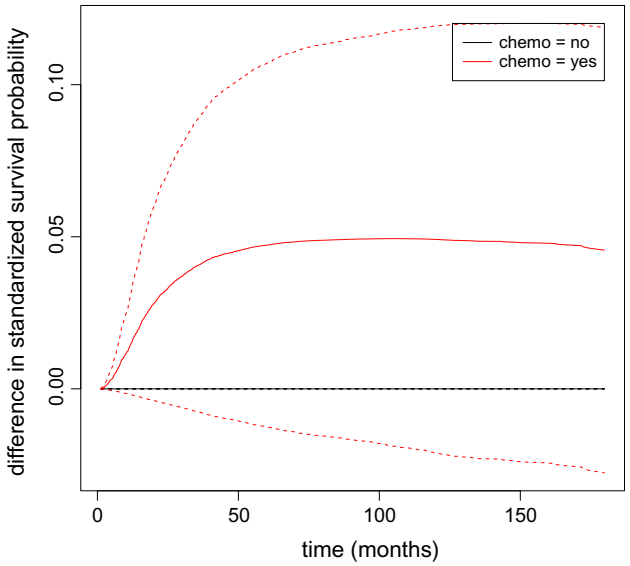

Fig. 4 Difference in standardized survival functions with 'no chemotherapy' as the reference

the confidence intervals are quite wide, and no effect is statistically significant. In practice, the level of model complexity must thus be chosen to strike a reasonable balance between robustness and efficiency.

Appendix 1: Asymptotic distribution for standardized measures

For generalized linear models, let x_0 and x_1 be fixed constants. Let $\psi = g\{\theta(x_0), \theta(x_1)\}$ be a function of $\theta(x_0)$ and $\theta(x_1)$, e.g. $\theta(x_1) - \theta(x_0)$. Define $v = \{\beta, \theta(x_0), \theta(x_1), \psi\}$. The estimator $\hat{v} = [\hat{\beta}, \hat{\theta}(x_0), \hat{\theta}(x_1), g\{\hat{\theta}(x_0), \hat{\theta}(x_1)\}]$ is an M-estimator [5] that solves the estimating equation

$$\sum_{i=1}^{n} U_{v,i}(v) = \sum_{i=1}^{n} \begin{bmatrix} U_{\beta,i}(\beta) \\ U_{\theta(x_0),i}\{\beta, \theta(x_0)\} \\ U_{\theta(x_1),i}\{\beta, \theta(x_1)\} \\ U_{\psi,i}\{\theta(x_0), \theta(x_1), \psi\} \end{bmatrix} = 0,$$

where $U_{\beta,i}(\beta)$ is the contribution to the maximum likelihood score function from subject i, $U_{\theta(x),i}\{\beta, \theta(x)\} = \eta^{-1}\{h(X = x, Z_i; \beta)\} - \theta(x)$ for $x = x_1$ and $x = x_0$, and $U_{\psi,i}\{\theta(x_0), \theta(x_1), \psi\} = g\{\theta(x_0), \theta(x_1)\} - \psi$.

For Cox regression models, let x_0, x_1 and t be fixed constants. Let $\psi = g\{\theta(t, x_0), \theta(t, x_1)\}$ be a function of $\theta(t, x_0)$ and $\theta(t, x_1)$, e.g. $\theta(t, x_1) - \theta(t, x_0)$. Define $v = \{\beta, \Lambda_0(t), \theta(t, x_0), \theta(t, x_1), \psi\}$. The estimator $\hat{v} = [\hat{\beta}, \hat{\Lambda}_0(t), \hat{\theta}(t, x_0), \hat{\theta}(t, x_1), g\{\hat{\theta}(t, x_0), \hat{\theta}(t, x_1)\}]$ is an M-estimator [5] that solves the estimating equation

$$\sum_{i=1}^{n} U_{v,i}(v) = \sum_{i=1}^{n} \begin{bmatrix} U_{\beta,i}(\beta) \\ U_{\Lambda_0(t),i}\{\beta, \Lambda_0(t)\} \\ U_{\theta(t,x_0),i}\{\beta, \Lambda_0(t), \theta(t, x_0)\} \\ U_{\theta(t,x_1),i}\{\beta, \Lambda_0(t), \theta(t, x_1)\} \\ U_{\psi,i}\{\theta(t, x_0), \theta(t, x_1), \psi\} \end{bmatrix} = 0,$$

where $U_{\beta,i}(\beta)$ is the contribution to the Cox partial likelihood score function from subject i, $U_{\Lambda_0(t),i}\{\beta, \Lambda_0(t)\}$ is the contribution to the estimating function for Breslow's estimator of the cumulative baseline hazard from subject i, $U_{\theta(t,x),i}\{\beta, \Lambda_0(t), \theta(t,x)\} = \exp[-\Lambda_0(t)\exp\{h(X = x, Z_i; \beta)\}] - \theta(t,x)$ for $x = x_1$ and $x = x_0$, and $U_{\psi,i}\{\theta(t,x_0), \theta(t,x_1), \psi\} = g\{\theta(t,x_0), \theta(t,x_1)\} - \psi$.

For both generalized linear models and Cox regression models it now follows from standard theory for M-estimators [5] that $n^{1/2}(\hat{v} - v)$ is asymptotically normal with mean 0 and variance given by the 'sandwich formula'

$$\Sigma = E'\left\{\frac{\partial U_{v,i}(v)}{\partial v}\right\}^{-1} \text{var}\{U_{v,i}(v)\} E\left\{\frac{\partial U_{v,i}(v)}{\partial v}\right\}^{-1}. \quad (5)$$

A consistent estimate of the variance of $\hat{v}$ is obtained by replacing v in (5) with $\hat{v}$, and the population moments in (5) by their sample counterparts.

The sandwich formula assumes that $U_{v,i}(v)$ and $U_{v,i'}(v)$ are independent, for $i \neq i'$. When data are clustered, as in the example in 'Standardization with generalized linear models' section, we may define $U_{v,i}(v) = \sum_{j=1}^{n_i} U_{v,ij}(v)$, where $U_{v,ij}(v)$ is the contribution to the estimating equation from subject j within cluster i, and n_i is the total number of subjects in cluster i. Provided that the clusters are independent we thus have that $U_{v,i}(v)$ and $U_{v,i'}(v)$ are independent as well, for $i \neq i'$, so that the sandwich formula still applies.

References

1. Rothman K, Greenland S, Lash T. Mod Epidemiol. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
2. Gail M, Byar D. Variance calculations for direct adjusted survival curves, with applications to testing for no treatment effect. Biom J. 1986;28(5):587–99.
3. Sjölander AAF. stdReg: Regression Standardization. R package version 0.1. 2016.
4. Dahlqwist E, Sjölander AAF. Model-based estimation of confounder-adjusted attributable fractions. R package version 0.1 2015.
5. Stefanski L, Boos D. The calculus of M-estimation. Am Stat. 2002;56(1):29–38.
6. Breslow N, Day N. Statistical methods in cancer research. The analysis of case–control studies, vol. 1. Lyon: IARC/WHO; 1980.

7. van der Laan M. Estimation based on case–control designs with known prevalence probability. Int J Biostat. 2008;4(1):a17.
8. De Jong U, Breslow N, Hong G, Ewe J, Sridharan M, Shanmugaratnam K. Aetiological factors in oesophageal cancer in singapore chinese. Int J Cancer. 1974;13(3):291–303.
9. Sjölander A, Vansteelandt S, Humphreys K. A principal stratification approach to assess the differences in prognosis between cancers caused by hormone replacement therapy and by other factors. Int J Biostat. 2010;6(1):a20.
10. Breslow N. Discussion of the paper by D. R. Cox. J R Stat Soc B. 1972;34(2):216–7.
11. Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. Biom J. 2007; 49(3):453–73.
12. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology. 2003;14(6):680–6.
13. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. Comput Methods Progr Biomed. 2004;75(1):45–9.
14. Robins J. Robust estimation in sequentially ignorable missing data and causal inference models. Proc Am Stat Assoc. 2000;1999: 6–10.
15. Bai X, Tsiatis A, O'Brien S. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. Biometrics. 2013;69(4):830–9.
16. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model. 1986;7(9):1393–512.