Arvid Sjolander[1] / Torben Martinussen[2]

Instrumental Variable Estimation with the R Package `ivtools`

[1] Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels Väg 12A, Stockholm 171 77, Sweden, E-mail: arvid.sjolander@ki.se

[2] Department of Public Health, University of Copenhagen, Kobenhavns, Denmark, E-mail: tma@sund.ku.dk

Abstract:

Instrumental variables is a popular method in epidemiology and related fields, to estimate causal effects in the presence of unmeasured confounding. Traditionally, instrumental variable analyses have been confined to linear models, in which the causal parameter of interest is typically estimated with two-stage least squares. Recently, the methodology has been extended in several directions, including two-stage estimation and so-called G-estimation in nonlinear (e. g. logistic and Cox proportional hazards) models. This paper presents a new R package, `ivtools`, which implements many of these new instrumental variable methods. We briefly review the theory of two-stage estimation and G-estimation, and illustrate the functionality of the `ivtools` package by analyzing publicly available data from a cohort study on vitamin D and mortality.

Keywords: instrumental variables, mendelian randomization, two-stage estimation, G-estimation, statistical software

DOI: 10.1515/em-2018-0024

Received: December 12, 2018; **Revised:** March 13, 2019; **Accepted:** May 26, 2019

1 Introduction

A common aim of epidemiological studies is to estimate the causal effect of an exposure on an outcome. However, most observational studies suffer from unmeasured confounding, which makes causal inference challenging. One popular way to deal with unmeasured confounding is to use instrumental variable methods. These methods have the remarkable property of being able to infer causality in the presence of unmeasured confounding. This property, however, requires that an instrumental variable (IV) is available, for which three assumptions hold: (a) the IV is associated with the exposure, (b) the IV affects the outcome only through the exposure, and (c) the association between the IV and the outcome is unconfounded (e. g. Hernán and Robins 2006).

IV methods have a long history in econometrics, and have more recently become popular in epidemiologic research as well, in particular through so-called Mendelian randomization (MR) studies. In MR studies, the IV is a gene, or set of genes, for which the biological mechanisms are (supposedly) well understood, and the IV assumptions (a)–(c) are considered plausible.

Traditionally, IV analyses have been confined to linear models, in which the causal parameter of interest is typically estimated with two-stage least squares (TSLS). In the first stage, the exposure is regressed on the IV, and a 'predicted' exposure level is obtained for each subject. In the second stage, the outcome is regressed on these predictions, and the obtained slope in this second-stage regression is used as an estimate of the causal exposure effect.

More recently, the IV methodology has been extended in several directions. Some authors have considered two-stage estimation techniques for nonlinear models (see Vansteelandt et al. 2011; Tchetgen Tchetgen et al. 2015, and the references therein). These extensions are important, since nonlinear (e. g. logistic and Cox proportional hazards) model are by far more common than linear models in epidemiologic research. However, as noted by these authors, two-stage estimation techniques generally gives biased estimates outside the linear modeling framework. At best the bias is small, but it is possible to construct scenarios where the bias is substantial.

To circumvent this problem, an alternative technique has been proposed, called 'G-estimation'. Briefly, this technique constructs predictions of the counterfactual outcome for each subject, had the exposure been withheld. These predictions are subsequently used to solve a particular estimating equation, where the solution is an unbiased estimate of the causal exposure effect. G-estimation has been developed for both generalized linear

Arvid Sjolander is the corresponding author.

models (GLMs), Cox proportional hazards models and additive hazards models (Robins 1989, 1994; Vansteelandt and Goetghebeur 2003; Martinussen et al. 2017; Martinussen et al. 2018).

Despite these methodological developments, TSLS estimation in linear models still appears to be the most common IV analysis in applied epidemiologic research. We conjecture that this is, to a large extent, due to a lack of software. When surveying current implementations in `R` we found four packages (`AER`, `ivmodel`, `sem` and `systemfit`) with functionality for TSLS estimation in linear models, but no package for two-stage/G-estimation in nonlinear models. Surveys of other statistical software gave similar results.

The purpose of this paper is to introduce a new `R` package for IV analysis; `ivtools`. This package has three functions for causal effect estimation with IVs: `ivglm`, `ivcoxph` and `ivah`, which perform estimation in causal GLMs, causal Cox proportional hazards models and causal additive hazards models, respectively. All three functions allow for two-stage estimation and G-estimation. Furthermore, they allow control for measured covariates, as well as exposure-covariate interactions. Finally, they provide analytic standard errors, which obviates the need for time-consuming bootstrap procedures.

The paper is organized as follows. In Section 2 we introduce basic notation, definitions and assumptions. In Section 3 we define the causal models and target parameters. In Section 4 we briefly review the theory for two-stage estimation and G-estimation, and in Section 5 we illustrate how to use these estimation techniques in practice, with the `ivtools` package. For this purpose we use data from a cohort study on vitamin D and mortality. These data are publicly available, and included in the `ivtools` package. Thus, the reader can easily replicate and elaborate on the presented analyses.

We assume that the reader is already familiar with the software R, and in particular with the functions `glm` and `coxph` for fitting GLMs and Cox proportional hazards models, respectively. We emphasize that the purpose of this paper is to show how to carry out two-stage estimation and G-estimation in practice, not to compare these methods or to give practical guidelines on how to choose between them. Such comparisons have already been made by other authors, e. g. Vansteelandt et al. (2011), Tchetgen Tchetgen et al. (2015).

2 Notation, definitions and assumptions

Let Z and X be the IV and the exposure, respectively. The IV may be a scalar variable, but may also be a vector of variables (e. g. genetic markers). We consider both point outcomes (i. e. outcomes that are measured at one specific point in time), which we denote with Y, and time-to-event outcomes, which we denote with T. We allow for Z, X and Y to be arbitrary numerical (e. g. binary or continuous) variables. The methods we consider allow for T to be right-censored; however, to keep notation simple we ignore censoring in the description of these methods. Truncation poses a problem for both two-stage estimation and G-estimation; we explain why in Appendix A. Let L be a set of measured covariates that we wish to control for in the analysis. This set may include confounders for the exposure and the outcome, but we don't assume that L is sufficient for confounding control. In particular, L is the empty set if no measured covariates are available.

The Directed Acyclic Graph (DAG) in Figure 1 illustrates a possible causal structure for L, Z, X, and Y/T. Our aim is to estimate the causal effect of the exposure on the outcome, as represented by the arrow from X to Y/T. The variable U represents all unmeasured confounders for the exposure and the outcome. We attempt to use IV methods to estimate the causal exposure effect, in the presence of unmeasured confounding.

Under the DAG in Figure 1, the IV assumptions hold conditionally on L. Specifically, we have that (a) Z is associated with X, (b) Z affects Y/T only through X, and (c) the association between Z and Y/T is unconfounded, conditionally on L. We emphasize that this DAG is not unique in this aspect; other DAGs are possible for which the IV assumptions hold as well. For instance, assumption (a) also holds if Z has no causal effect on X, but Z and X rather have common causes. We also emphasize that, although it is important in real studies to verify that the IV assumptions hold, this task is beyond the scope of our paper; we refer to Glymour et al. (2012) for a thorough discussion of this topic.

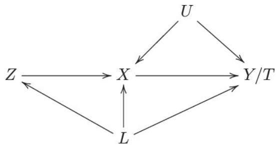

Figure 1: A causal structure under which the IV assumptions hold, conditionally on L.

3 Causal models

Unfortunately, it can be shown that the IV assumptions alone are not sufficient to estimate the causal exposure effect. Without additional assumptions, it is only possible to bound the causal effect, that is, to provide upper and lower limits for the effect (Balke and Pearl 1997). The `ivtools` package contains a function to compute such bounds. However, since our focus here is mainly on point estimation we present this function in Appendix B.

To be able to estimate the causal exposure effect, one typically proceeds by making parametric modeling assumptions. Standard TSLS is often thought of as estimating the causal effect in linear structural models that condition on all unmeasured confounders U. However, such models are problematic, for two reasons. First, in real applications one would typically have limited knowledge about the unmeasured confounders. Thus, it may be difficult to interpret causal effects that are defined conditionally on these. Second, in nonlinear models it is not possible to consistently estimate causal effects that are defined conditionally on U. Thus, inference for such causal effects is essentially restricted to linear models.

For these reasons, many authors have focused on causal models that condition on Z instead of U. For point outcomes, let Y_0 be the potential outcome for a given subject, when the exposure is completely withheld, i. e. set to 0 (Rubin 1974; Pearl 2009). We consider the causal ('structural') mean model

$$\eta\{E(Y|L,Z,X)\} - \eta\{E(Y_0|L,Z,X)\} = m^T(L)X\psi, \tag{1}$$

where η is either the identity, log or logit link. The (column) vector parameter ψ measures the causal effect of a particular exposure level, among those who factually received that exposure level, conditionally on (L, Z). The (column) vector function $m(L)$ allows for interactions between X and L. To allow for a main effect, the first element of $m(L)$ would typically be the constant '1'. The causal exposure effect is assumed to be constant across levels of Z; this assumption is necessary for identifiability (Hernán and Robins 2006). For instance, suppose that X is binary (0/1), η is the identity link and $m(L) = 1$. Then $\psi = E(Y|L,Z,X=1) - E(Y_0|L,Z,X=1)$ is the causal effect of exposure level 1, as a mean difference, for those who factually received exposure level 1 and have a fixed level of (L, Z). The causal mean model in (1) was originally proposed by Robins (1989) and Robins (1994), who developed G-estimators of ψ when η is the identity link or log link. Vansteelandt and Goetghebeur (2003) later extended the theory for this model, by developing a G-estimator of ψ when η is the logit link.

Recently, analogs to the causal model in (1) were proposed for time-to-event outcomes (Martinussen et al. 2017, 2018). Similar to above, let T_0 be the potential time-to-event outcome for a given subject, when the exposure is completely withheld, i. e. set to 0, and let $\lambda(t|L,Z,X)$ and $\lambda_0(t|L,Z,X)$ be the conditional hazard functions, given (L, Z, X), for T and T_0, respectively, at time t. Martinussen et al. (2018) considered the causal Cox proportional hazards model

$$\log\{\lambda(t|L,Z,X)\} - \log\{\lambda_0(t|L,Z,X)\} = m^T(L)X\psi, \tag{2}$$

and they developed a G-estimator for ψ in this model. Martinussen et al. (2017) considered a causal additive hazards model on the form

$$\lambda(t|L,Z,X) - \lambda_0(t|L,Z,X) = XdB(t) \tag{3}$$

where $B(t)$ is a possibly unrestricted function of time, and $dB(t)$ is the first order derivative of this function. The model in (3) is somewhat different than the models in (1) and (2). On the one hand it is more restrictive, since it does not allow for exposure-covariate interactions. One the other hand it is less restrictive than the model in (2), since it does not assume a constant exposure effect over time. Martinussen et al. (2017) developed a G-estimator of $B(t)$ without making any assumptions about this function, as well as a G-estimator of $B(t)$ under the assumptions that $dB(t)$ is constant (= ψ). Under this assumption, model (3) reduces to

$$\lambda(t|L,Z,X) - \lambda_0(t|L,Z,X) = X\psi. \tag{4}$$

4 Estimation theory

4.1 Two-stage estimation

As the name suggests, two-stage estimation requires fitting two regression models. In the first stage, a regression model is fitted for the exposure, using L and Z as regressors. This model is used to create a prediction $\hat{X} = \hat{E}(X|Z,L)$ for each subject. In the second stage, another regression model is fitted for the outcome, using L and

$m^T(L)\hat{X}$ as regressors. The choice of regression model in the second stage mimics the causal target model. That is, for causal models (1), (2) and (4) we use a GLM with link function η, a Cox proportional hazards model, and an additive hazards model, respectively, in the second stage. Two-stage estimation of an unrestricted function $B(t)$ in model (3) is currently not available. The estimated coefficient vector for $m^T(L)\hat{X}$ in the second stage regression model is the two-stage estimate of ψ in the corresponding causal model. The standard error of the two-stage estimate may be obtained by stacking all estimating equations involved in the estimation process, and applying the 'sandwich formula' to the whole equation system (Stefanski and Boos 2002).

The rationale behind two-stage estimation is most easily understood in linear models, where the outcome can be thought of as a sum of two terms; the exposure X and an 'error term' ε. By confounding, the error term is associated with the exposure. Intuitively then, the first stage can be thought of as extracting the part of the variation in X that is independent of ε, since, under the IV assumptions, the IV is (conditionally) independent of the confounders (given L). In the second stage, we then use this 'exogenous' part of the variation in X to estimate the causal effect ψ.

An advantage of two-stage estimation is that it is computationally simple. It gives unbiased estimates in linear models, that is, in model (1) with identity link, in which case it reduces to standard TSLS, and in model (4) if the first stage model is a linear regression (Tchetgen Tchetgen et al. 2015). However, several authors have noted that two-stage estimation generally gives biased estimates in nonlinear models, such as model (1) with logit link, and in model (2). This bias essentially arises from the non-collapsibility of non-linear effect measures, such as the odds ratio and the hazard ratio (Greenland et al. 1999). Under some circumstances, typically when the outcome is rare, the non-collapsibility may not be severe, in which case the bias may be small. The bias can sometimes be reduced by adding the 'control function' $R = X - \hat{E}(X|Z, L)$ to the set of regressors in the second stage model (Vansteelandt et al. 2011; Tchetgen Tchetgen et al. 2015).

4.2 G-estimation

G-estimation is a fairly general estimation technique, which includes, for instance, semiparametrically efficient and doubly robust estimators. We here describe a special case of G-estimation, which is implemented in the `ivtools` package. We refer to Robins (1989, 1994); Vansteelandt and Goetghebeur (2003); Martinussen et al. (2017, 2018) for more general expositions on G-estimation in IV analyses.

As for two-stage estimation, the rationale behind G-estimation is most easily understood in linear models. Under the causal model (1) with identity link, we may define the residual $h_i(\psi) = Y_i - X_i\psi$. When the causal effect $X_i\psi$ has been removed from the factual outcome, this residual can be thought of as a prediction of the potential outcome Y_{0i}. Under the IV assumptions, this potential outcome is (conditionally) independent of the IV (given L). Thus, under the IV assumptions we may estimate ψ as the value which makes this independence happen in the sample.

Formally, for model (1) a G-estimator of ψ can be defined as the solution to the estimating equation

$$H(\psi) = \sum_{i=1}^{n} \hat{d}(L_i, Z_i)h_i(\psi) = 0, \tag{5}$$

where

$$h_i(\psi) = \left\{ \begin{array}{ll} Y_i - m^T(L_i)X_i\psi & \text{if } \eta \text{ is the identity link} \\ Y_i\text{exp}\{-m^T(L_i)X_i\psi\} & \text{if } \eta \text{ is the log link} \\ \text{expit}\{\hat{E}(Y|L_i, Z_i, X_i) - m^T(L_i)X_i\psi\} & \text{if } \eta \text{ is the logit link} \end{array} \right\},$$

and $d(L, Z)$ is an arbitrary function with the same dimension as ψ, such that $E\{d(L, Z)|L\} = 0$, see Robins (1989, 1994) and Vansteelandt and Goetghebeur (2003).

By similar reasoning, a G-estimator of ψ in model (2) can be defined as the solution to the estimating equation

$$H(\psi; t) = \sum_{i=1}^{n} \hat{d}(L_i, Z_i)h_i(\psi; t) = 0, \tag{6}$$

where t is a fixed time-point, and

$$h_i(\psi; t) = \hat{p}(T > t|Z_i, X_i, L_i)^{\exp\{-m^T(L_i)X_i\psi\}},$$

see Martinussen et al. (2018). The G-estimator in (6) depends on the choice of t. If both the causal model (2) and the model for $p(T > t|L, Z, X)$ are correct, then the choice of t may affect the precision, but not the unbiasedness, of the G-estimator. Martinussen et al. (2018) argued that t may be chosen as the time-point that minimizes the estimated variance of the G-estimate.

The `ivtools` package allows for two choices of $d(L, Z)$:

$$d(L,Z) = m(L)\{Z - E(Z|L)\} \tag{7}$$

and

$$d(L,Z) = m(L)\{E(X|L,Z) - E(X|L)\}. \tag{8}$$

The function in (7) has the advantage of requiring less modeling, i. e. it only requires a model for $E(Z|L)$ whereas the function in (8) requires models for both $E(X|L,Z)$ and $E(X|L)$. However, the function in (7) requires Z to be a scalar variable, whereas function in (8) allows for Z to a vector of arbitrary length. In the real data analysis (Section 5) we only present G-estimation with (7), and in Appendix C we repeat the analysis with (8).

The estimating functions in (5) and (6) require additional regression modeling. Specifically, they require estimates of $E(Z|L)$ (for $d(L, Z)$ in (7)), $E(X|L,Z)$ and $E(X|L)$ (for $d(L, Z)$ in (8)), $E(Y|L,Z,X)$ (for (5)) with logit link) and $p(T > t|L,Z,X)$ (for (6)), which are obtained by regression models. To avoid bias due to model misspecification, these models may be saturated, if possible.

The standard error of the G-estimator estimate may be obtained by stacking all estimating equations involved in the estimation process, and applying the 'sandwich formula' to the whole equation system (Stefanski and Boos 2002).

Martinussen et al. (2017) derived G-estimators of $B(t)$ and ψ in models (3) and (4), respectively. These estimators follow similar intuition as the G-estimators of ψ in models (1) and (2). However, they are more complex, as they are defined as recursive estimators of counting process integrals, rather than solutions to estimating equations. We refer to Martinussen et al. (2017) for details on these G-estimators.

We end this section with a remark on terminology. We have here defined 'G-estimator' as the class of estimators which make the potential outcome Y_0 (or T_0) independent of Z in the sample, conditionally on L. Under this definition, all estimators in this section are indeed 'G-estimators'. However, there appears to be no uniform agreement on the definition, and some authors seem to reserve the term for estimators that solely rely on a model for the IV. Under this more restrictive definition, the estimators of ψ in model (1) with logit link and model (2) are not G-estimators, since they additionally require models for $E(Y|L,Z,X)$ and $p(T > t|L,Z,X)$, respectively.

5 Real data analysis

5.1 The Vitamin D data

All functions that we use for the real data analysis are included in the `ivtools` and `survival` packages; we load these as usual by typing

```
> library(ivtools)
> library(survival)
```

We use a data set borrowed from Martinussen et al. (2018), on a cohort study of vitamin D and mortality. To estimate the causal effect of vitamin D on mortality, these authors carried out an IV analysis, using mutations in the filaggrin gene as an instrument. These mutations have been shown to be associated with higher vitamin D status (Skaaby et al. 2013). For ethical reasons, we were not able to make the original data publicly available. Thus, we have mutilated the original data slightly, by adding random noise to each variable. These mutilated data are included in the `ivtools` package, as the `VitD` data frame. We load these data by typing

```
> data(VitD)
```

The `VitD` data frame contains 2571 subjects and 5 variables: `age` (at baseline), `filaggrin` (binary indicator of whether filaggrin mutations are present) `vitd` (vitamin D level at baseline, measured as serum 25-OH-D (nmol/L)), `time` (follow-up time), and `death` (indicator of whether the subjects died during follow-up). The median follow-up time is 16.1 years, and 23.5 % of all subjects died during follow-up.

In Sections 5.2 and 5.3 we analyze death during follow-up as a binary point outcome. This analysis is mainly for illustration; since the exact time of death is known it is more natural to view the outcome as a time-to-event. We thus proceed in Sections 5.4 and 5.5 with more appropriate time-to-event analyses. We control for age at baseline in all analyses. We follow Martinussen et al. (2018), and define the vitamin D exposure as a continuous variable centered around 20 and normed with 20:

```
> VitD$vitd_std <- (VitD$vitd-20)/20
```

The analyses below require fitting several regression models. We use the naming convention `fitA.B`, where `A` is the name of the left-hand side variable in the model, and `B` are the names of the right-hand side variables in the model. For instance, `fitY.LX` is a model with Y on the LHS and (L, X) on the RHS, whereas `fitY.LZX` is a model with Y on the LHS and (L, Z, X) on the RHS.

5.2 Two-stage estimation with a point outcome

We start by using standard logistic regression to estimate the association between vitamin D and mortality, while controlling for baseline age. We fit the model, store the results in an object called `fitY.LX`, and summarize by typing

```
> fitY.LX <- glm(formula=death˜age+vitd_std, family="binomial",
  data=VitD)
> summary(fitY.LX)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.891271   0.370382 -21.306  < 2e-16 ***
age          0.120501   0.005835  20.653  < 2e-16 ***
vitd_std    -0.168094   0.043036  -3.906 9.39e-05 ***
```

We observe that higher vitamin D levels are associated with lower mortality, with a decrease of 0.17 units in the log odds of death during follow-up, for each 1 unit increase in standardized vitamin D. This association is statistically significant, with a p-value equal to 9.39×10^{-5}.

Since we only controlled for age in the logistic regression, we may worry that the observed association is partly or fully due to unmeasured confounding. We thus proceed with an IV analysis, fitting the causal model (1) with η being the logit link and $m(L) = 1$. For this purpose we use the `ivglm` function in the `ivtools` package, which allows for both two-stage estimation and G-estimation.

To carry out two-stage estimation we need one model for vitamin D, regressed on age and filaggrin (the first stage model), and one model for death, regressed on age and vitamin D (the second stage model). For the latter, we use the logistic regression model fitted above. For the former, we use an ordinary linear regression. We fit this model, store the results in an object called `fitX.LZ`, and summarize by typing

```
> fitX.LZ <- glm(formula=vitd_std˜age+filaggrin,
  data=VitD)
> summary(fitX.LZ)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.588441   0.139318  18.579  < 2e-16 ***
age         -0.006791   0.002471  -2.748  0.00603 **
filaggrin    0.279163   0.100703   2.772  0.00561 **
```

We observe that presence of mutations in the filaggrin gene are associated with higher vitamin D levels, as expected. We now use two-stage estimation to fit the causal model (1), by typing

```
> fitIV_ts <- ivglm(estmethod="ts", fitX.LZ=fitX.LZ, fitY.LX=fitY.LX,
  data=VitD)
```

The `estmethod` argument specifies the desired estimation method; `"ts"` or `"g"` for two-stage estimation and G-estimation, respectively. For two-stage estimation, three additional arguments must be specified; `fitX.LZ`, fitY.LX and data, which specify the exposure model, the outcome model and the data frame, respectively. The exposure and outcome models may be arbitrary GLMs, as fitted by the glm function. The ivglm function uses the exposure model to create predictions, and re-fits the outcome model with the true exposure levels replaced by these predictions. The link function η and the interaction function $m(L)$ in the target causal model (1) are implicitly defined by the outcome model fitY.LX.

We summarize as usual, by typing

```
> summary(fitIV_ts)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.632786   2.415038  -1.918   0.0551 .
age          0.111355   0.008835  12.603   <2e-16 ***
vitd_std    -1.394216   0.925283  -1.507   0.1319
```

We observe that the estimated effect of vitamin D is quite substantial, with a decrease of 1.39 units in the log odds of death during follow-up, for each 1 unit increase in standardized vitamin D. However, this effect is not statistically significant.

The summary method for "ivglm" objects produces the same statistics as the summary method for "glm" objects. It would often be desirable to compute confidence intervals as well. This can be done with the confint method:

```
> confint(fitIV_ts)
                  2.5%      97.5%
(Intercept) -9.36617460  0.1006019
age          0.09403776  0.1286714
vitd_std    -3.20773707  0.4193051
```

We remind the reader that the two-stage estimate of ψ in model (1) is generally biased, unless η is the identity link. The bias can often be reduced by adding the 'control function' $R = X - \hat{E}(X|Z, L)$ to the set of regressors in the second stage model (Vansteelandt et al. 2011). To use this control function we set the optional argument ctrl equal to TRUE:

```
> fitIV_ts <- ivglm(estmethod="ts", fitX.LZ=fitX.LZ, fitY.LX=fitY.LX,
  data=VitD, ctrl=TRUE)
> summary(fitIV_ts)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.54793    2.44037  -1.864 0.062375 .
age          0.11186    0.00891  12.553  < 2e-16 ***
vitd_std    -1.45115    0.93737  -1.548 0.121598
R           -0.16528    0.04689  -3.525 0.000423 ***
```

We observe that the control function approach gives a slightly larger estimated effect of vitamin D.

5.3 G-estimation with a point outcome

G-estimation with $d(L, Z) = Z - E(Z|L)$ requires one model for filaggrin, regressed on age, and one model for death, regressed on age, filaggrin and vitamin D. We use logistic regression models for these, fitted as

```
> fitZ.L <- glm(formula=filaggrin˜age, family="binomial",
  data=VitD)
> fitY.LZX <- glm(formula=death˜age+filaggrin+vitd_std,
  family="binomial", data=VitD)
```

We now use G-estimation to the fit the causal model (1) with a logit link, by typing

```
> fitIV_g <- ivglm(estmethod="g", X="vitd_std", fitZ.L=fitZ.L,
  fitY.LZX=fitY.LZX, data=VitD, link="logit")
```

The `X` argument specifies the names of the exposure; this must be specified when `estmethod="g"`. The `fitZ.L` and `fitY.LZX` arguments specify the fitted models for the IV and the outcome, respectively. Finally, the `link` argument specifies the desired link function η in model (1); either `"identity"`, `"log"` or `"logit"`. For the identity and log links, G-estimation does not require a model for the outcome, in which case the argument `fitY.LZX` is not used.

We summarize and compute confidence intervals, by typing

```
> summary(fitIV_g)
         Estimate Std. Error z value Pr(>|z|)
vitd_std   -1.423      1.844  -0.772     0.44
> confint(fitIV_g)
             2.5%   97.5%
vitd_std -5.036987 2.191156
```

We observe that the estimated effect of vitamin D is similar for G-estimation as for two-stage estimation, with a decrease of 1.42 units in the log odds of death during follow-up, for each 1 unit increase in standardized vitamin D. However, as for two-stage estimation the effect is not statistically significant.

The two stage estimate and the G-estimate are very similar (1.39 vs 1.42), but the standard errors are strikingly different (0.93 vs 1.84). We investigate the possible explanations for this discrepancy in Appendix D, where we do a bootstrap analysis.

Several authors have noted that the estimating eq. (5) does not always have a solution (Vansteelandt et al. 2011; Burgess et al. 2014). The `ivglm` function prints a warning if no solution is found, but it may also be desirable to manually inspect the estimating function $H(\psi)$ in (5), to make sure that it is equal to 0 at the reported solution. For this purpose we may use the `estfun` function, which computes $H(\psi)$ for a range of values for ψ around the G-estimate. We type

```
> H <- estfun(object=fitIV_g)
> plot(H)
```

which gives the plot in Figure 2, where the vertical dashed line indicates the G-estimate. We observe that $H(\psi)$ is indeed equal to 0 at the reported G-estimate −1.423.

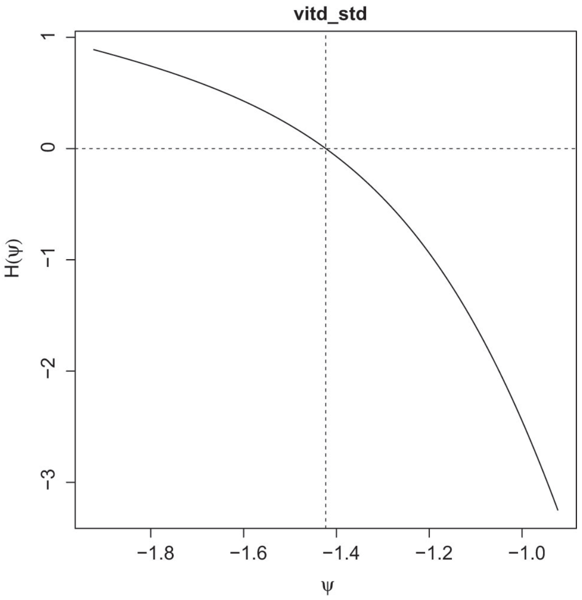

Figure 2: Estimating function for the Vitamin D data, using G-estimation under the causal logistic model (1).

So far, we have considered causal models without interactions, i. e. models with $m(L) = 1$. For two-stage estimation, interactions are defined implicitly by the fitted outcome model. For G-estimation, interactions have to be defined explicitly, by using an optional argument `formula` in the `ivglm` function. This argument specifies $m(L)$ as the right-hand side of a formula on the usual R-format. For instance, to allow for an interaction between vitamin D and age, we type

```
> fitIV_g <- ivglm(estmethod="g", X="vitd_std", fitZ.L=fitZ.L,
  fitY.LZX=fitY.LZX, data=VitD, link="logit", formula=~age)
Warning message:
In ivglm(estmethod = "g", Z = "filaggrin", X = "vitd_std", Y = "death",  :
  No solution to the estimating equation was found
```

Here we observe that, for this more complex model, no solution to the estimating equations was found. However, the performance of numerical equation solvers may often be sensitive to specific control parameter settings, such as the choice of starting values and maximum number of iterations. For G-estimation, `ivglm` and `ivcoxph` internally use the `nleqslv` function from the `nleqslv` package to solve the estimating equations. This function has a wide range of control parameters (see the help page), which can be specified in the call to `ivglm` and `ivcoxph`, and are then passed on to `nleqslv`. For instance, the starting values for the equation solver in `nleqslv` are specified by an argument `x`, which defaults to 0 for all elements of ψ . Specifying starting values −7 and 0.1 for the main effect and the interaction term resolves the convergence problem in this particular case:

```
> fitIV_g <- ivglm(estmethod="g", X="vitd_std", fitZ.L=fitZ.L,
  fitY.LZX=fitY.LZX, data=VitD, link="logit", formula=~age,
  x=c(-7,0.1))
> summary(fitIV_g)
             Estimate Std. Error z value Pr(>|z|)
vitd_std     -9.19943    5.33863  -1.723   0.0849 .
vitd_std:age  0.13007    0.07741   1.680   0.0929 .
```

Figure 3 shows $H(\psi)$ for the main effect and the interaction term, as computed by `estfun`. When the causal model has more than one parameter, `estfun` computes $H(\psi)$ over a range of values for each parameter separately, at the G-estimate for the remaining parameters. We observe that $H(\psi)$ is indeed equal to 0 at the reported G-estimates −9.20 and 0.13 for the main effect and the interaction, respectively.

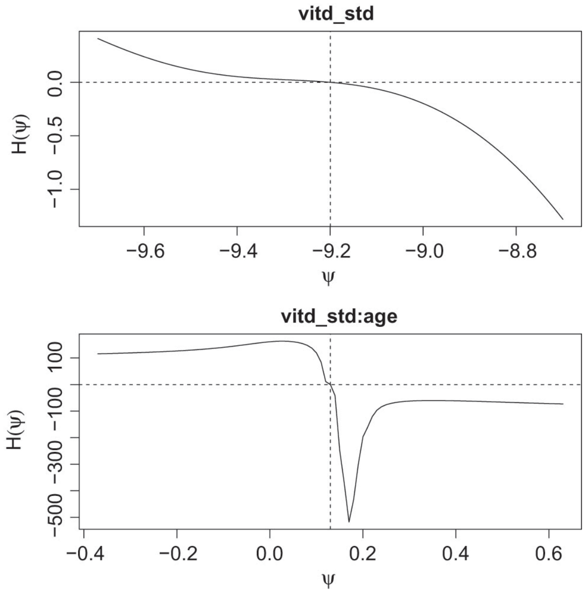

Figure 3: Estimating function for the Vitamin D data, using G-estimation under the causal logistic model (1), with an interaction term between vitamin D and age.

5.4 Two-stage estimation with a time-to-event outcome

We start by using standard Cox proportional hazards regression to estimate the association between vitamin D and mortality. We fit the model, store the results in an object called `fitT.LX`, and summarize by typing

```
> fitT.LX <- coxph(formula=Surv(time, death)˜age+vitd_std,
  data=VitD)
> summary(fitT.LX)
              coef exp(coef)  se(coef)      z Pr(>|z|)
age       0.099563  1.104688  0.004612 21.586  < 2e-16 ***
vitd_std -0.143115  0.866654  0.034684 -4.126 3.69e-05 ***
```

We observe that higher vitamin D levels are associated with lower mortality, with a decrease of 0.14 units in the log hazard of death during follow-up, for each 1 unit increase in standardized vitamin D. This association is statistically significant, with a p-value equal to 3.69×10^{-5}.

As before, we may worry that the observed association is partly or fully due to unmeasured confounding. We thus proceed with an IV analysis, fitting the causal model (2) with $m(L) = 1$. For this purpose we use the `ivcoxph` function in the `ivtools` package, which allows for both two-stage estimation and G-estimation.

To carry out two-stage estimation we need one model for vitamin D, regressed on age and filaggrin (the first stage model), and one model for death, regressed on age and vitamin D (the second stage model). For the latter, we use the Cox proportional hazards model fitted above. For the former, we use the same model `fitX.LZ` as in Section 5.2. We now use two-stage estimation to the fit the causal model (2), by typing

```
> fitIV_ts <- ivcoxph(estmethod="ts", fitX.LZ=fitX.LZ, fitT.LX=fitT.LX,
  data=VitD)
```

For two-stage estimation, the `ivcoxph` function has almost identical syntax to the `ivglm` function. An exception is that the argument specifying the outcome model is called `fitT.LX` instead of `fitY.LX`, to emphasize that the outcome is a time-to-event. The exposure model may be an arbitrary GLM, as fitted by the `glm` function, and the outcome model must be a Cox proportional hazards model, as fitted by the `coxph` function.

We summarize and compute confidence intervals by typing

```
> summary(fitIV_ts)
          Estimate Std. Error z value Pr(>|z|)
age       0.092530   0.006189  14.950   <2e-16 ***
vitd_std -1.071472   0.616576  -1.738   0.0822 .
> confint(fitIV_ts)
               2.5 %    97.5 %
age       0.08039898 0.1046609
vitd_std -2.27993775 0.1369945
```

We observe that the estimated effect of vitamin D is quite substantial, with a decrease of 1.07 units in the log hazard of death during follow-up, for each 1 unit increase in standardized vitamin D. However, this effect is not statistically significant. We remind the reader that the two-stage estimate of ψ in model (2) is generally biased. The bias can often be reduced by adding the 'control function' $R = X - \hat{E}(X|Z, L)$ to the set of regressors in the second stage model (Tchetgen Tchetgen et al. 2015). Using the control function approach gives a slightly larger estimate:

```
> fitIV_ts <- ivcoxph(estmethod="ts", fitX.LZ=fitX.LZ, fitT.LX=fitT.LX,
  data=VitD, ctrl=TRUE)
> summary(fitIV_ts)
          Estimate Std. Error z value Pr(>|z|)
age       0.092779   0.006154  15.075  < 2e-16 ***
vitd_std -1.142065   0.613819  -1.861 0.062802 .
R        -0.141177   0.037016  -3.814 0.000137 ***
```

The Cox proportional hazards model is by far the most commonly used model for time-to-event outcomes in epidemiologic research. However, the `ivtools` package also allows for IV analyses with additive hazards models. To fit an additive hazards model we use the `ah` function in the `ivtools` package. This function is essentially a wrapper around the `ahaz` function in the `ahaz` package, which has a less standard R input/output interface. We fit the model and summarize by typing

```
> fitT.LX <- ah(formula=Surv(time, death)~age+vitd_std, data=VitD)
> summary(fitT.LX)
           Estimate  Std. Error  Z value  Pr(>|z|)
age       1.478e-03  8.024e-05  18.425  < 2e-16 ***
vitd_std -1.807e-03  4.633e-04  -3.899 9.66e-05 ***
```

Again we observe a protective, statistically significant effect of high vitamin D levels. However, this may be partly or fully due to unmeasured confounding. To carry out two-stage estimation with the fitted additive hazards model, together with the previously fitted model `fitX` for the exposure, we use the `ivah` function:

```
> fitIV_ts <- ivah(estmethod="ts", fitX.LZ=fitX.LZ, fitT.LX=fitT.LX,
  data=VitD)
> summary(fitIV_ts)
          Estimate  Std. Error  z value  Pr(>|z|)
age       0.001360   0.000105  12.946   <2e-16 ***
vitd_std -0.019294   0.010924  -1.766   0.0774 .
> confint(fitIV_ts)
                2.5%    97.5%
age       0.00115369 0.001565354
vitd_std -0.04070457 0.002117029
```

We remind the reader that, since we have used a linear regression model at the first stage, the two-stage estimate of ψ is consistent when using an additive hazards model at the second stage, as opposed to when using a Cox proportional hazards model (Tchetgen Tchetgen et al. 2015).

5.5 G-estimation with a time-to-event outcome

G-estimation with $d(L, Z) = Z - E(Z|L)$ requires one model for filaggrin, regressed on age, and one model for death, regressed on age, filaggrin and vitamin D. For the former, we use the same logistic regression `fitZ.L` as in Section 5.3. For the latter, we use a Cox proportional hazards model, fitted as

```
> fitT.LZX <- coxph(formula=Surv(time, death)~age+filaggrin+vitd_std,
  data=VitD)
```

We now use G-estimation to the fit the causal model (2), by typing

```
> fitIV_g <- ivcoxph(estmethod="g", X="vitd_std", fitZ.L=fitZ.L,
  fitT.LZX=fitT.LZX, data=VitD)
```

For G-estimation, the `ivcoxph` function has similar syntax as the `ivglm` function. The two exceptions are that the outcome is called `T` instead of `Y`, and the the outcome model is called `fitT` instead of `fitY`. We note that `T` is the name of the follow-up time variable, not the event indicator. We also note that the estimating eq. (6) always requires an estimate of $p(T > t|L, Z, X)$ obtained from an outcome model, and thus `fitT.LZX` must be specified. This model can be either a Cox proportional hazards model or a set of nonparametric Kaplan-Meier curves, as fitted by the `coxph` and `survfit` functions, respectively.

We summarize and compute confidence intervals by typing

```
> summary(fitIV_g)
```

```
Equation solved at t = 16.46264


Coefficients:
           Estimate Std. Error z value Pr(>|z|)
vitd_std  -0.8628     0.3620  -2.384   0.0171 *
> confint(fitIV_g)
              2.5%    97.5%
vitd_std -1.57216 -0.15334
```

We observe that the estimated effect of vitamin D is smaller for G-estimation than for two-stage estimation, with a decrease of 0.86 units in the log hazard of death during follow-up, for each 1 unit increase in standardized vitamin D. This effect is also statistically significant. We inspect the estimating function by typing

```
> H <- estfun(object=fitIV_g)
> plot(H)
```

which gives the plot in Figure 4, where the vertical dashed line indicates the G-estimate. We observe that $H(\psi)$ is indeed equal to 0 at the reported G-estimate -0.86.

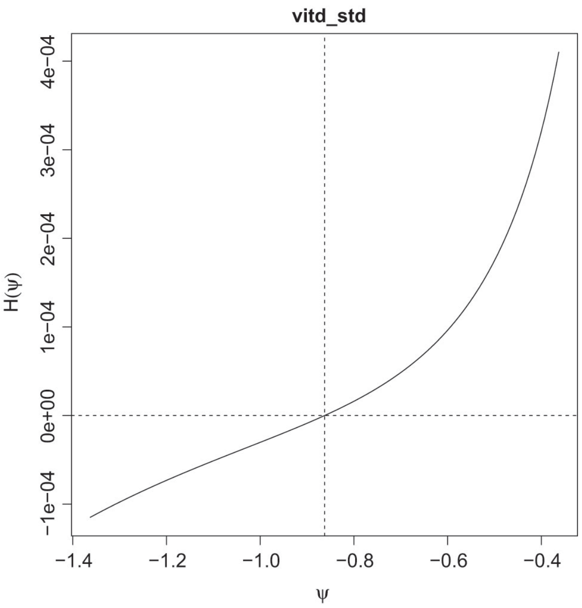

Figure 4: Estimating function for the Vitamin D data, using G-estimation under the causal Cox proportional hazards model (2).

The summary output tells us that the estimation equation is solved at $t = 16.46$. By default, `ivcoxph` solves the equation at the value of t that minimizes the estimated variance of the G-estimate, as suggested by Martinussen et al. (2018). This behavior can be over-ridden by specifying a specific time-point, through an optional argument `t`. Figure 5 shows the G-estimate (solid line) as a function of t, together with 95 % confidence limits (dashed lines). We observe that, for this particular case, the estimate is not very sensitive to the choice of t; it increases from -1.10 at $t = 10$ to -0.84 at $t = 20$. We further observe that the estimate does not change when $t > 17$; this is because there are no deaths in the data set beyond this point. The estimating equation in (6) only depends on t through $\hat{p}(T > t|Z_i, X_i, L_i)$, which is flat beyond the last observed event when obtained from a Cox proportional hazards model.

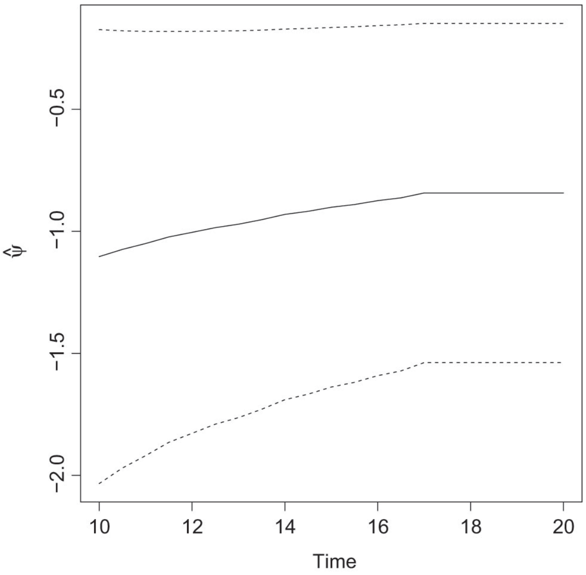

Figure 5: G-estimate (solid line) as a function of t, together with 95 % confidence limits (dashed lines).

To carry out G-estimation under the causal additive hazards model (3) we type

```
> fitIV_g <- ivah(estmethod="g", X="vitd_std", T="time",
  fitZ.L=fitZ.L, data=VitD, event="death")
```

The syntax for ivah is identical to the syntax for ivcoxph, except for the additional arguments T and event, which specify the name of the time-to-event variable and event indicator, respectively. To display the G-estimate of $B(t)$ we type

```
> plot(fitIV_g)
```

which gives the plot in Figure 6. The solid curve is the estimate of $B(t)$, the dashed curves are point-wise lower and upper 95 % confidence limits. The solid straight line corresponds to a constant exposure effect, $dB(t) = \psi$. From Figure 6, the assumption of a constant exposure effect seems fairly reasonable. Summarizing the model fit gives:

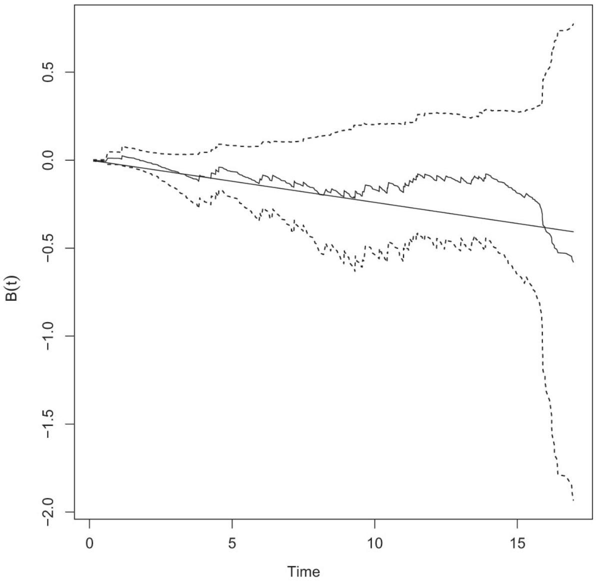

Figure 6: Estimate of $B(t)$ for the Vitamin D data, under the causal additive hazards model (3).

```
> summary(fitIV_g)
Test for non-significant exposure effect. H_0: B(t)=0

         Supremum-test pval
vitd_std               0.4

Goodness-of-fit test for constant effects model
         Supremum-test pval
                      0.51

Constant effect model

         Estimate Std. Error z value Pr(>|z|)
vitd_std -0.02400    0.02669  -0.899    0.369
```

This model summary is somewhat different than the other summaries we have presented above, which reflects that the causal model (3) is also somewhat different than the causal models (1) and (2). On the top, two p-values are displayed. The first p-value tests the null hypothesis of no exposure effect, $B(t) = 0$, and the second p-value tests the null hypothesis of a constant exposure effect, $dB(t) = \psi$. Both p-values are high, and thus we observe no strong evidence against either of the null hypotheses. On the bottom, the usual summary statistics are given for ψ, assuming a constant effect.

6 Discussion

In this paper we have presented a new R package, `ivtools`, which implements several important tools for IV analyses. We have illustrated the functionality of this package by analyzing data from a cohort study on vitamin D and mortality. These data are publicly available and included in the `ivtools` package, which makes it easy for the reader to replicate and elaborate on the presented analyses.

An advantage of the `ivtools` package is that it provides analytic standard errors, which obviates the need for time-consuming bootstrap procedures. However, we caution the reader that these standard errors are asymptotic, and Wald-type confidence intervals and p-values based on these may not be accurate in small-to-moderate samples where the normal approximation doesn't hold. For these cases it is often more appropriate to use bootstrap methods. We illustrate how this can be done with the `boot` package, in Appendix D.

Causal inference is a rapidly evolving branch of statistics, and new methods are constantly developed. In our experience, practitioners often don't have time or adequate skills to program themselves. Thus, the availability of off-the-shelf software is crucial for disseminating these new methods and ensure that they are used in applied research. We hope that the `ivtools` package will facilitate the practical use of IV methods, and we intend to keep the package up-to-date as new methods are developed in the future.

Funding

This work was supported by the Swedish Research Council (Grant Number: 2016-01267).

A Left-truncation

Let $\widetilde{T}$ be the random truncation time. Thus, a subject is only included in the sample if $T > \widetilde{T}$. This selection causes problems for both two-stage estimation and G-estimation. For two-stage estimation we wish to estimate $E(X|L, Z)$ in the first stage. However, when the first stage regression model is fitted to a left-truncated sample, it estimates $E(X|L, Z, T > \widetilde{T})$, which is generally different from $E(X|L, Z)$, even if the truncation is independent.

Similarly, for a left-truncated sample the G-estimator ψ in model (2) solves the estimating equation

$$H(\psi; t) = \sum_{i=1}^{n} \hat{d}(L_i, Z_i) h_i(\psi; t) I(T > \widetilde{T}) = 0, \tag{9}$$

where $h_i(\psi;t) = \hat{p}(T > t|Z_i, X_i, L_i)^{\exp\{-m^T(L_i)X_i\psi\}}$. Replacing the estimates $\hat{d}(L_i, Z_i)$ and $\hat{p}(T > t|Z_i, X_i, L_i)$ with the true values we have that

$$\begin{aligned} E\{H(\psi;t)\} &= nE\left[d(L,Z)p(T > t|L,Z,X)^{\exp\{-m^T(L)X\psi\}}p(T > \tilde{T}|L,Z,X)\right] \\ &= nE\left\{d(L,Z)p(T_0 > t|L,Z,X)p(T > \tilde{T}|L,Z,X)\right\}, \end{aligned}$$

which is generally different from 0 even if the truncation is independent; thus, the estimating equation is generally biased.

One solution to this problem is to fit the first stage model, and solve the estimating equation, for only those subjects who are not truncated, i. e. for those subjects for which $\widetilde{T} = 0$. The first stage model thus estimates $E(X|L,Z,\widetilde{T} = 0) = E(X|L,Z)$, if $\widetilde{T} \perp X|L,Z$.

Similarly, the G-estimator solves the estimating equation

$$H(\psi;t) = \sum_{i=1}^{n} \hat{d}(L_i, Z_i)h_i(\psi;t)I(\widetilde{T} = 0) = 0. \tag{10}$$

Replacing the estimates $\hat{d}(L_i, Z_i)$ and $\hat{p}(T > t|Z_i, X_i, L_i)$ with the true values we have that

$$\begin{aligned} E\{H(\psi;t)\} &= nE\left[d(L,Z)p(T > t|L,Z,X)^{\exp\{-m^T(L)X\psi\}}p(\tilde{T} = 0|L,Z,X)\right] \\ &= nE\left\{d(L,Z)p(T_0 > t|L,Z,X)p(\tilde{T} = 0|L,Z,X)\right\} \\ &= nE\left\{d(L,Z)p(T_0 > t|L)p(\tilde{T} = 0|L)\right\} = 0, \end{aligned}$$

where the third equality follows if $\widetilde{T} \perp (Z,X)|L$ and $T_0 \perp Z|L$.

B Nonparametric bounds

Balke and Pearl (1997) showed that the IV assumptions are not sufficient to identity the causal exposure effect. However, they showed that the causal effect can be bounded, using a certain linear programming technique. These bounds are implemented in the `ivtools` package, by the `ivbounds` function. In addition to the bounds, the `ivbounds` function evaluates the so-called IV inequality

$$\max_x \sum_y \max_z p_{yx.z} \leq 1, \tag{11}$$

where $p_{yx.z} = p(Y = y, X = x|Z = z)$. Balke and Pearl (1997) showed that, in order for the IV assumptions to hold, the IV inequality must be satisfied.

As before, we use the `VitD` data for illustration. A limitation of the linear programming technique is that it requires both the IV, exposure and outcome to be categorical, with relatively few levels. The `ivbounds` function allows for binary or ternary IVs, but requires that both the exposure and the outcome are binary. As vitamin D levels below 30 nmol/L indicate vitamin D deficiency (Martinussen et al. 2018), we use this level as a cutoff for defining a binary exposure:

```
> VitD$vitd_bin <- as.numeric(VitD$vitd>30)
```

To compute the bounds and summarize the result we type

```
> fit <- ivbounds(data=VitD, Z="filaggrin", X="vitd_bin", Y="death")
> summary(fit)
The IV inequality is violated at the following conditions:
[1] p01.1+p11.0<=1


Symbolic bounds:
   lower                   upper
p0 p10.0+p11.0-p00.1-p11.1 p01.0+p10.0+p10.1+p11.1
p1 p11.0                   1-p01.1
```

```
Numeric bounds:
       lower  upper
p0   0.05297 0.9215
p1   0.21540 0.2010
CRD -0.70608 0.1481
CRR  0.23375 3.7953
COR  0.02339 4.4986
```

In this summary, `p0` and `p1` represent the counterfactual probabilities $p(Y_0 = 1)$ and $p(Y_1 = 1)$, respectively. `CRD`, `CRR` and `COR` represent the causal risk difference, the causal risk ratio and the causal odds ratio, respectively, defined as

$$\begin{aligned}\mathrm{CRD} &= p(Y_1 = 1) - p(Y_0 = 1)\\ \mathrm{CRR} &= p(Y_1 = 1)/p(Y_0 = 1)\\ \mathrm{COR} &= \frac{p(Y_1 = 1)/p(Y_1 = 0)}{p(Y_0 = 1)/p(Y_0 = 0)}\end{aligned}$$

The summary output tells us that the IV inequality (11) is violated at the condition $p_{01.1}+p_{11.0} \leq 1$. Indeed, we also observe that the lower bound for $p(Y_1 = 1)$ is above the upper bound. This may be explained by sampling variability, but may also indicate that the IV assumptions are not valid for this particular scenario.

Violations of the IV assumptions may occur when continuous variables are dichotomized, and whether or not violations occur may depend on the chosen cutoff. When using the cutoff 35 nmol/L, the IV inequality is not violated, and all lower bounds are below the corresponding upper bounds:

```
> VitD$vitd_dbin <- as.numeric(VitD$vitd>35)
> fit <- ivbounds(data=VitD, Z="filaggrin", X="vitd_dbin", Y="death")
> summary(fit)
The IV inequality is not violated


Symbolic bounds:
   lower upper
p0 p10.0 p01.0+p10.0+p10.1+p11.1
p1 p11.0 1-p01.1


Numeric bounds:
       lower  upper
p0   0.03870 0.9025
p1   0.19983 0.2423
CRD -0.70271 0.2036
CRR  0.22141 6.2595
COR  0.02697 7.9411
```

In the above examples, data were provided at a subject level, as an R data frame. However, the `ivbounds` function also accepts a named vector (p 00.0, ... , p 11.1):

```
> freq <- xtabs(formula=~death+vitd_bin+filaggrin, data=VitD)
> p <- prop.table(freq, margin=3)
> p
, , filaggrin = 0


     vitd_bin
death          0          1
    0 0.05384939 0.70761464
    1 0.02313841 0.21539756


, , filaggrin = 1

     vitd_bin
death          0          1
    0 0.01030928 0.79896907
    1 0.01546392 0.17525773
> names(p) <- c("p00.0", "p10.0", "p01.0", "p11.0",
  "p00.1", "p10.1", "p01.1", "p11.1")
> fit <- ivbounds(data=p)
> summary(fit)
The IV inequality is not violated
Symbolic bounds:
   lower upper
p0 p10.0 p01.0+p10.0+p10.1+p11.1
p1 p11.0 1-p01.1
Numeric bounds:
       lower  upper
p0   0.03870 0.9025
p1   0.19983 0.2423
CRD -0.70271 0.2036
CRR  0.22141 6.2595
COR  0.02697 7.9411
```

C G-estimation with $d(L,Z) = m(L)\{E(X|L,Z) - E(X|L)\}$

In this section we repeat the main effects analysis from Sections 5.3 and 5.5, using $d(L,Z) = m(L)\{E(X|L,Z) - E(X|L)\}$. For a point outcome we type

```
> fitX.LZ <- glm(formula=vitd_std~age+filaggrin, data=VitD)
> fitX.L <- glm(formula=vitd_std~age, data=VitD)
> fitY.LZX <- glm(formula=death~age+filaggrin+vitd_std, family="binomial",
  data=VitD)
> fitIV_g <- ivglm(estmethod="g", fitX.LZ=fitX.LZ, fitX.L=fitX.L,
  fitY.LZX=fitY.LZX, data=VitD, link="logit")
> summary(fitIV_g)
         Estimate Std. Error z value Pr(>|z|)
vitd_std   -1.426      1.859  -0.767    0.443
```

We note that, when $d(L,Z) = m(L)\{E(X|L,Z) - E(X|L)\}$ we don't have to specify the argument `X`, since the name of the exposure variable can be obtained from the left-hand side of model `fitX.LZ` (or model `fitX.L`). We observe that both the estimate and the standard error are very similar to those obtained with $d(L,Z) = m(L)\{Z - E(Z|L)\}$.

For a time-to-event outcome we type

```
> fitT.LZX <- coxph(formula=Surv(time, death)~age+filaggrin+vitd_std,
  data=VitD)
> fitIV_g <- ivcoxph(estmethod="g", fitX.LZ=fitX.LZ, fitX.L=fitX.L,
    fitT.LZX=fitT.LZX, data=VitD)
> summary(fitIV_g)
Equation solved at t = 16.46367
```

```
Coefficients:
         Estimate Std. Error z value Pr(<|z|)
vitd_std  -0.8632     0.3620  -2.384   0.0171 *
```

Again, we observe that both the estimate and the standard error are very similar to those obtained with $d(L, Z) = m(L)\{Z - E(Z|L)\}$.

D Bootstrap standard errors

Bootstrap confidence intervals and p-values can easily be computed with the `boot` package. To illustrate this, we repeat the two-stage estimation and G-estimation from Sections 5.2 and 5.3, using bootstrap. First, we load the `boot` package and set a seed:

```
> library(boot)
> set.seed(1)
```

We next define a function to bootstrap, and carry out the bootstrap with 10000 replicates:

```
> bootfun <- function(data, indicies){
  dd <- data[indicies, ]
  #two-stage estimation
  fitX.LZ <- glm(formula=vitd_std˜ age+filaggrin, data=dd)
  fitY.LX <- glm(formula=death˜age+vitd_std, family="binomial",
    data=dd)
  fitIV_ts <- ivglm(estmethod="ts", fitX.LZ=fitX.LZ, fitY.LX=fitY.LX,
    data=dd, vcov.fit=FALSE)
  est_ts <- fitIV_ts$est["vitd_std"]
  #G-estimation
  fitZ.L <- glm(formula=filaggrin˜age, family="binomial", data=dd)
  fitY.LZX <- glm(formula=death˜age+filaggrin+vitd_std,
    family="binomial", data=dd)
  fitIV_g <- ivglm(estmethod="g", X="vitd_std", fitZ.L=fitZ.L,
    fitY.LZX=fitY.LZX, data=dd, link="logit", vcov.fit=FALSE)
  est_g <- fitIV_g$est["vitd_std"]
  return(c(est_ts, est_g))
}
> bb <- boot(data=VitD, statistic=bootfun, R=10000)
```

We obtain bootstrap standard errors by typing

```
> apply(bb$t, MARGIN=2, FUN=sd, na.rm=TRUE)
[1] 38.440136  1.900573
```

The bootstrap standard error for the G-estimator (=1.90) agrees well with the analytic standard error computed in Section 5.3 (=1.84). However, the bootstrap standard error for the two-stage estimator is huge (=38.44), whereas the analytic standard error computed in Section 5.2 is very modest (0.93).

To understand why this happens we look closer into the bootstrap distributions. Figure 7 shows QQ-plots and histograms of the bootstrap distributions for the two-stage estimator (left column) and G-estimator (right column). The vertical axes for the QQ-plots extend to the maximal and minimal bootstrap replicates, but the horizontal axes for the histograms are truncated at $(-10, 10)$. From the QQ-plot we observe that there are extreme outliers in the bootstrap distribution for the two-stage estimators, which inflate the bootstrap standard error and are not accounted for in the analytic calculation. Thus, the analytic standard error for the two-stage estimator does not agree well with the true standard error in this particular case.

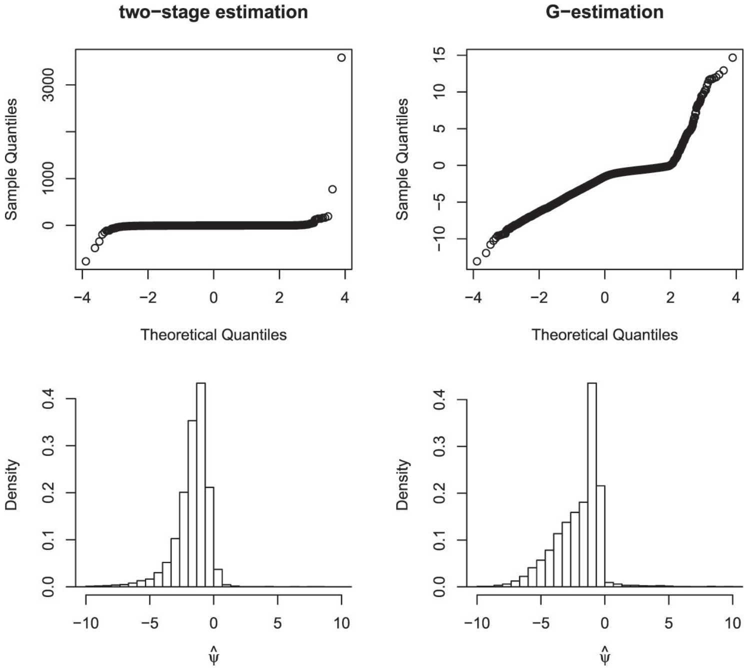

Figure 7: QQ-plots and histograms for bootstrap distributions of the two-stage estimator and the G-estimator.

From the histograms we observe that both bootstrap distributions decline steeply just above 0, which doesn't agree well with the normal distribution. Thus, Wald-type confidence intervals may not have accurate coverage in this case. To compute non-parametric 95 % confidence intervals we type

```
> apply(bb$t, MARGIN=2, FUN=quantile, probs=c(0.025, 0.975), na.rm=TRUE)
              [,1]          [,2]
2.5% -5.9488033   -6.21231895
97.5% 0.1211937   -0.03633081
```

These confidence intervals are fairly similar; however the G-estimate is statistically significant (at 5 % significance level), whereas the two-stage estimate is not.

References

Balke, A., and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. Journal of the American Statistical Association, 92:1171–1176.

Burgess, S., Granell, R., Palmer, T., Sterne, J., and Didelez, V. (2014). Lack of identification in semiparametric instrumental variable models with binary outcomes. American Journal of Epidemiology, 180:111–119.

Glymour, M., Tchetgen Tchetgen, E., and Robins, J. (2012). Credible mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. American Journal of Epidemiology, 175:332–339.

Greenland, S., Robins, J., and Pearl, J. (1999). Confounding and collapsibility in causal inference. Statistical Science, 14:29–46.

Hernán, M., and Robins, J. (2006). Instruments for causal inference: an epidemiologist's dream? Epidemiology, 17:360–372.

Martinussen, T., Nørbo Sørensen, D., and Vansteelandt, S. (2018). Instrumental variables estimation under a structural Cox model. Biostatistics, 20:65–79.

Martinussen, T., Vansteelandt, S., Tchetgen Tchetgen, E., and Zucker, D. (2017). Instrumental variables estimation of exposure effects on a time-to-event endpoint using structural cumulative survival models. Biometrics, 73:1140–1149.

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. 2nd Edition. New York: Cambridge University Press,

Robins, J. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In: Health Service Research Methodology: A Focus on AIDS, L. Sechrest, H. Freeman, and A. Mulley (Eds.), 113–159. US Public Health Service, National Center for Health Services Research.

Robins, J. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. Communications in Statistics - Theory and Methods, 23:2379–2412.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66:688–701.

Skaaby, T., Husemoen, L., Martinussen, T., Thyssen, J., Melgaard, M., Thuesen, B., Pisinger, C., Jørgensen, T., Johansen, J., Menné, T., et al. (2013). Vitamin d status, filaggrin genotype, and cardiovascular risk factors: a mendelian randomization approach. PloS one, 8:e57647.

Stefanski, L., and Boos, D. (2002). The calculus of M-estimation. The American Statistician, 56:29–38.

Tchetgen Tchetgen, E., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental variable estimation in a survival context. Epidemiology, 26:402–410.

Vansteelandt, S., Bowden, J., Babanezhad, M., and Goetghebeur, E. (2011). On instrumental variables estimation of causal odds ratios. Statistical Science, 26:403–422.

Vansteelandt, S., and Goetghebeur, E. (2003). Causal inference with generalized structural mean models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65:817–835.

Supplementary Material: The online version of this article offers supplementary material (DOI:https://doi.org/10.1515/em-2018-0024).