Package 'budgetIVr'

April 16, 2025

Title Partial Identification of Causal Effects with Mostly Invalid Instruments

Version 0.1.2

Date 2025-04-12

Maintainer Jordan Penn <jordan.penn5841@gmail.com>

Description A tuneable and interpretable method for relaxing the instrumental variables (IV) assumptions to infer treatment effects in the presence of unobserved confounding. For a treatment-associated covariate to be a valid IV, it must be (a) unconfounded with the outcome and (b) have a causal effect on the outcome that is exclusively mediated by the exposure. There is no general test of the validity of these IV assumptions for any particular pre-treatment covariate. However, if different pre-treatment covariates give differing causal effect estimates when treated as IVs, then we know at least some of the covariates violate these assumptions. 'budgetIVr' exploits this fact by taking as input a minimum budget of pre-treatment covariates assumed to be valid IVs and idenfiying the set of causal effects that are consistent with the user's data and budget assumption. The following generalizations of this principle can be used in this package: (1) a vector of multiple budgets can be assigned alongside corresponding thresholds that model degrees of IV invalidity; (2) budgets and thresholds can be chosen using specialist knowledge or varied in a principled sensitivity analysis; (3) treatment effects can be nonlinear and/or depend on multiple exposures (at a computational cost). The methods in this package require only summary statistics. Confidence sets are constructed under the ``no measurement error'' (NOME) assumption from the Mendelian randomization literature. For further methodological details, please refer to Penn et al. (2024) <doi:10.48550/arXiv.2411.06913>.

License GPL (>= 3)

URL https://github.com/jpenn2023/budgetIVr

BugReports https://github.com/jpenn2023/budgetIVr/issues

Imports data.table, arrangements, MASS, Rglpk, stats

Encoding UTF-8

RoxygenNote 7.3.2

NeedsCompilation no

Author Jordan Penn [aut, cre, cph] (<https://orcid.org/0009-0002-3572-1724>)

Depends R (>= 3.5.0)

Repository CRAN

Date/Publication 2025-04-16 14:20:09 UTC

Contents

`budgetIV` *Partially identify causal effects with invalid instruments*

Description

Computes the set of possible values of a causal parameter consistent with observational data and given budget constraints. See Penn et al. (2025) for technical definitions.

Usage

```
budgetIV(
  beta_y,
  beta_phi,
  phi_basis = NULL,
  tau_vec = NULL,
  b_vec = NULL,
  ATE_search_domain = NULL,
  X_baseline = NULL,
  delta_beta_y = NULL
)
```

Arguments

`beta_y`	Either $1 \times d_Z$ matrix or a d_Z-dimensional vector representing the (estimated) cross covariance $\mathrm{Cov}(Y, Z)$.
`beta_phi`	A $d_\Phi \times d_Z$ matrix representing the (estimated) cross covariance $\mathrm{Cov}(\Phi(X), Z)$.

`phi_basis` A d_Φ-dimensional expression (separated by commas) with each term representing a component of $\Phi(X)$. The expression consists of d_X unique vars. The default value `NULL` can be used for a $d_X = d_\Phi$-dimensional linear model.

`tau_vec` A K-dimensional vector of increasing, positive thresholds representing degrees of IV invalidity. The default value `NULL` can be used for a single threshold at 0.

`b_vec` A K-dimensional vector of increasing positive integers representing the maximum number of IVs that can surpass each threshold. The default value `NULL` can be used for a single threshold at 0, with at least 50% of IVs assumed to be valid.

`ATE_search_domain` A d_X-column data.frame with column names equal to the vars in `phi_basis`. Rows correspond to values of the treatment X. The default value `NULL` can be used to generate a small d_X-dimensional grid.

`X_baseline` Either a data.frame or list representing a baseline treatment x_0, with names equal to the vars in `phi_basis`. The default value `NULL` can be used for the baseline treatment 0 for each of of the d_X vars.

`delta_beta_y` A d_Z-dimensional vector of positive half-widths for box-shaped confidence bounds on `beta_y`. The default value `NULL` can be used to not include finite sample uncertainty.

Details

Instrumental variables are defined by three structural assumptions: (A1) they are associated with the treatment; (A2) they are unconfounded with the outcome; and (A3) exclusively effect the outcome through the treatment. Of these, only (A1) can be tested without further assumptions. The `budgetIV` function allows for valid causal inference when some proportion (possibly a small minority) of candidate instruments satisfy both (A2) and (A3). Tuneable thresholds decided by the user also allow for bounds on the degree of invalidity for each instrument (i.e., bounds on the proportion of $\mathrm{Cov}(Y, Z)$ not explained by the causal effect of X on Z). Full technical details are included in Penn et al. (2025).

`budgetIV` assumes that treatment effects are homogeneous, which implies a structural equation of the form $Y = \theta \cdot \Phi(X) + g_y(Z, \epsilon_x)$, where θ and $\Phi(X)$ are a d_Φ-dimensional vector and vector-valued function respectively. A valid basis expansion $\Phi(X)$ is assumed (e.g., linear, logistic, polynomial, RBF, neural embedding, PCA, UMAP etc.). It is also assumed that $d_\Phi <= d_Z$, which allows us to treat the basis functions as a complete linear model (see Theil (1953), or Sanderson et al. (2019) for a modern MR focused discussion). The parameters θ capture the unknown treatment effect. Violation of (A2) and/or (A3) will bias classical IV approaches through the statistical dependence between Z and $g_y(Z, \epsilon_x)$, summarized by the covariance parameter $\gamma := \mathrm{Cov}(g_y(Z, \epsilon_x), Z)$.

`budgetIV` constrains γ through a series of positive thresholds $0 \leq \tau_1 < \tau_2 < \ldots < \tau_K$ and corresponding integer budgets $0 < b_1 < b_2 < \ldots < b_K \leq d_Z$. It is assumed for each $i \in \{1, \ldots, K\}$ that no more than b_i components of γ are greater in magnitude than τ_i. For instance, taking $d_Z = 100$, $K = 1$, $b_1 = 5$ and $\tau_1 = 0$ means assuming 5 of the 100 candidates are valid instrumental variables (in the sense that their ratio estimates $\theta_j := \mathrm{Cov}(Y, Z_j)/\mathrm{Cov}(\Phi(X), Z_j)$ are unbiased).

With `delta_beta_y = NULL`, `budgetIV` returns the identified set of causal effects that agree with both the budget constraints described above and the values of $\mathrm{Cov}(Y, Z)$ and $\mathrm{Cov}(Y, Z)$, assumed

to be exactly precise. Unlike classical partial identification methods (see Manski (1990) for a canonical example), the non-convex mixed-integer budget constraints yield a possibly disconnected solution set. Each connected subset has a different interpretation as to which of the candidate instruments Z are valid up to each threshold.

`delta_beta_y` represents box-constraints to quantify uncertainty in `beta_y`. In the examples, `delta_beta_y` is calculated through a Bonferroni correction and gives an (asymptotically) valid confidence set over `beta_y`. Under the so-called "no measurement error" assumption (see Bowden et al. (2016)), which is commonly applied in Mendelian randomization, it is assumed that the estimate of `beta_y` is the dominant source of finite-sample uncertainty, with uncertainty in `beta_x` considered negligible. With an (asymptotically) valid confidence set for `delta_beta_y`, and under the "no measurement error" assumption, `budgetIV` returns an (asymptotically) valid confidence set for θ when using just a single exposure.

Value

A `data.table` with each row corresponding to a set of bounds on the ATE at a given point in `ATE_search_domain`. Columns include: a non-unique identifier `curve_index` with a one-to-one mapping with `U`; `lower_ATE_bound` and `upper_ATE_bound` for the corresponding bounds on the ATE; a list `U` for the corresponding budget assignment; and a column for each unique variable in `ATE_search_domain` to indicate the treatment value at which the bounds are being calculated.

References

Jordan Penn, Lee Gunderson, Gecia Bravo-Hermsdorff, Ricardo Silva, and David Watson. (2024). BudgetIV: Optimal Partial Identification of Causal Effects with Mostly Invalid Instruments. *AISTATS* 2025.

Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala A Sheehan, and John R Thompson. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *Int. J. Epidemiol.* 46.6, pp. 1985–1998.

Charles F Manski. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* 80.2, pp. 219–323.

Henri Theil. (1953). Repeated least-squares applied to complete equation systems. *Centraal Planbureau Memorandum*.

Eleanor Sanderson, George Davey Smith, Frank Windmeijer and Jack Bowden. (2019). An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.* 48.3, pp. 713–727.

Examples

```
data(simulated_data_budgetIV)

beta_y <- simulated_data_budgetIV$beta_y

beta_phi_1 <- simulated_data_budgetIV$beta_phi_1
beta_phi_2 <- simulated_data_budgetIV$beta_phi_2

beta_phi <- matrix(c(beta_phi_1, beta_phi_2), nrow = 2, byrow = TRUE)
```

```
delta_beta_y <- simulated_data_budgetIV$delta_beta_y

tau_vec = c(0)
b_vec = c(3)

x_vals <- seq(from = 0, to = 1, length.out = 500)

ATE_search_domain <- expand.grid("x" = x_vals)

phi_basis <- expression(x, x^2)

X_baseline <- list("x" = c(0))

solution_set <- budgetIV(beta_y = beta_y,
                         beta_phi = beta_phi,
                         phi_basis = phi_basis,
                         tau_vec = tau_vec,
                         b_vec = b_vec,
                         ATE_search_domain = ATE_search_domain,
                         X_baseline = X_baseline,
                         delta_beta_y = delta_beta_y)
```

`budgetIV_scalar` *Efficient partial identification of a scalar causal effect parameter with invalid instruments*

Description

Partial identification and coverage of a causal effect parameter using summary statistics and budget constraint assumptions. See Penn et al. (2025) for technical definitions.

Usage

```
budgetIV_scalar(
  beta_y,
  beta_phi,
  tau_vec = NULL,
  b_vec = NULL,
  delta_beta_y = NULL,
  bounds_only = TRUE
)
```

Arguments

`beta_y`	A d_Z-dimensional vector representing the (estimated) cross covariance $\mathrm{Cov}(Y, Z)$.
`beta_phi`	A d_Z-dimensional vector representing the (estimated) cross covariance $\mathrm{Cov}(\Phi(X), Z)$.

`tau_vec` A K-dimensional vector of increasing, positive thresholds representing degrees of IV invalidity. The default value `NULL` can be used for a single threshold at 0.

`b_vec` A K-dimensional vector of increasing positive integers representing the maximum number of IVs that can surpass each threshold. The default value `NULL` can be used for a single threshold at 0, with at least 50% of IVs assumed to be valid.

`delta_beta_y` A d_Z-dimensional vector of positive half-widths for box-shaped confidence bounds on `beta_y`. The default value `NULL` can be used to not include finite sample uncertainty.

`bounds_only` A boolean `TRUE` or `FALSE`. `TRUE` will store overlapping intervals in the confidence set as a single interval, while `FALSE` will store different intervals for different values of `budget_assignment` (see return value of Penn et al. (2025) for further details). The default is `TRUE`.

If TRUE (default), the output consists only of disjoint bounds. Otherwise, if FALSE, the output consists of bounds for possibly touching intervals (but never overlapping), as well as the budget assignment corresponding to each bound.

Details

Instrumental variables are defined by three structural assumptions: (A1) they are associated with the treatment; (A2) they are unconfounded with the outcome; and (A3) they exclusively effect the outcome through the treatment. Assumption (A1) has a simple statistical test, whereas for many data generating processes (A2) and (A3) are unprovably false. The `budgetIV` and `budgetIV_scalar` algorithms allow for valid causal inference when some proportion, possibly a small minority, of candidate instruments satisfy both (A2) and (A3).

`budgetIV` & `budgetIV_scalar` assume a homogeneous treatment effect, which implies the separable structural equation $Y = \theta\Phi(X) + g_y(Z, \epsilon_x)$. The difference between the algorithms is that `budgetIV_scalar` assumes $\Phi(X)$ and θ take scalar values, which is exploited for super-exponential computational speedup and allows for causal inference with thousands of candidate instruments Z. Both methods assume ground truth knowledge of the functional form of $\Phi(X)$, e.g., a linear, logistic, Cox hazard, principal component based or other model. The parameter θ captures the unknown treatment effect. Violation of (A2) and/or (A3) will bias classical IV approaches through the statistical dependence between Z and $g_y(Z, \epsilon_x)$, summarized by the covariance parameter $\gamma := \mathrm{Cov}(g_y(Z, \epsilon_x), Z)$.

`budgetIV` & `budgetIV_scalar` constrain γ through a series of positive thresholds $0 \leq \tau_1 < \tau_2 < \ldots < \tau_K$ and corresponding integer budgets $0 < b_1 < b_2 < \ldots < b_K \leq d_Z$. It is assumed for each $i \in \{1, \ldots, K\}$ that no more than b_i components of γ are greater in magnitude than τ_i. For instance, taking $d_Z = 100$, $K = 1$, $b_1 = 5$ and $\tau_1 = 0$ means assuming 5 of the 100 candidates are valid instrumental variables (in the sense that their ratio estimates $\theta_j := \mathrm{Cov}(Y, Z_j)/\mathrm{Cov}(\Phi(X), Z_j)$ are unbiased).

With `delta_beta_y = NA`, `budgetIV` & `budgetIV_scalar` return the identified set of causal effects that agree with both the budget constraints described above and the values of $\mathrm{Cov}(Y, Z)$ and $\mathrm{Cov}(Y, Z)$, assumed to be exactly precise. Unlike classical partial identification methods (see Manski (1990) ofr a canonical example), the non-convex mixed-integer budget constraints yield a possibly disconnected identified set. Each connected subset has a different interpretation as to which of the candidate instruments Z are valid up to each threshold. `budgetIV_scalar` returns these interpretations alongside the corresponding bounds on θ.

When `delta_beta_y` is not null, it is used as box-constraints to quantify uncertainty in `beta_y`. In the examples, `delta_beta_y` is calculated through a Bonferroni correction and gives an (asymptotically) valid confidence set over `beta_y`. Under the so-called "no measurement error" (NOME) assumption (see Bowden et al. (2016)) which is commonly applied in Mendelian randomisation, it is assumed that the estimate of `beta_y` is the dominant source of finite-sample uncertainty, with uncertainty in `beta_x` entirely negligible. With an (asymptotically) valid confidence set over `delta_beta_y` and under the "no measurement error" assumption, `budgetIV_scalar` returns an (asymptotically) valid confidence set for θ.

Value

A data.table with each row corresponding to bounds on the scalar causal effect parameter θ corresponding to a particular budget assignment U (see Penn et al. (2025)). The return table has the following rows: a logical `is_point` determining whether the upper and lower bounds are equivalent; numerical `lower_bound` and `upper_bound` giving the lower and upper bounds; and a list `budget_assignment` giving the value of U for each candidate instrument. `budget_assignment` will only be returned if `bounds_only == FALSE` as input by the user.

A list of two entries: `intervals`, which is a two-column matrix with rows corresponding to disjoint bounds containing plausible values of θ; and `points`, which is a one-column matrix consisting of lone plausible values of θ—relevant when using $\tau_1 = 0$.

References

Jordan Penn, Lee Gunderson, Gecia Bravo-Hermsdorff, Ricardo Silva, and David Watson. (2024). BudgetIV: Optimal Partial Identification of Causal Effects with Mostly Invalid Instruments. *arXiv* preprint, 2411.06913.

Jack Bowden, Fabiola Del Greco M, Cosetta Minelli, George Davey Smith, Nuala A Sheehan, and John R Thompson. (2016). Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I^2 statistic. *Int. J. Epidemiol.* 46.6, pp. 1985–1998.

Charles F Manski. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* 80.2, pp. 219–323.

Examples

```
data(Do_et_al_summary_statistics)

candidatesHDL = Do_et_al_summary_statistics[Do_et_al_summary_statistics$pHDL <= 1e-8, ]

candidate_labels <- candidatesHDL$rsID
d_Z <- length(candidate_labels)

beta_x <- candidatesHDL$betaHDL

beta_y <- candidatesHDL$betaCAD

SE_beta_y <- abs(beta_y) / qnorm(1-candidatesHDL$pCAD/2)

alpha = 0.05
```

```
delta_beta_y <- qnorm(1 - alpha/(2*d_Z))*SE_beta_y

feasible_region <- budgetIV_scalar(
                                    beta_y = candidatesHDL$betaCAD,
                                    beta_phi = beta_x,
                                    tau_vec = c(0),
                                    b_vec = c(30),
                                    delta_beta_y = delta_beta_y,
                                    bounds_only = FALSE
                                    )
```

`Do_et_al_summary_statistics` *Summary statistics from Do et al. (2013)*

Description

Common variants associated with plasma triglycerides and risk for coronary artery disease. Pre-processed and harmonized summary statistics from a Mendelian randomization analysis, including summary statistics for variants' association with plasma triglyceride levels, serum HDL levels, serum LDL levels and risk of coronary artery disease (CAD). Dataset previously applied in the mode-based estimate approach of Hartwig et al. (2017). Each row of the dataset corresponds to a single genetic variant (single nucleotide polymorphism) found to be associated with either the HDL, LDL, or triglyceride biomarkers across a population of 180,000 (HDL, LDL) or 86,000 (triglyceride) individuals. Got further biological and statistical details, see Do et al. (2013).

Usage

```
data(Do_et_al_summary_statistics)
```

Format

A data frame with 185 rows and 14 variables:

Details

`X` A unique identifier from 1 to 185.

`rsID` A unique string specifying each SNP using the rsID format.

`chr` String specifying the chromosomal position of each SNP.

`a1` Character specifying one allele of the SNP (all 185 SNPs are assumed to be biallelic).

`a2` Character specifying the other allele of the SNP.

`betaLDL` Effect size (linear regression) for association between SNP allele and LDL.

`pLDL` p-value for testing association between SNP allele and LDL.

`betaHDL` Effect size (linear regression) for association between SNP allele and HDL.

`pHDL` p-value for testing association between SNP allele and HDL.

`betaTri` Effect size (linear regression) for association between SNP allele and triglyceride.

`pTri` p-value for testing association between SNP allele and triglyceride.

`betaCAD` Effect size (logistic regression) for association between SNP allele and CAD.

`pCAD` p-value for testing association between SNP allele and CAD.

References

Ron Do et al. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet.* 45.11, pp. 1345–52.

Fernando Pires Hartwig, George Davey Smith, and Jack Bowden. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46.6, pp. 1985–1998.

Examples

```
# Extracting relevant summary statistics to investigate the causal effect of HDL on CAD risk.

data(Do_et_al_summary_statistics)

candidatesHDL = Do_et_al_summary_statistics[Do_et_al_summary_statistics$pHDL <= 1e-8, ]

candidate_labels <- candidatesHDL$rsID
d_Z <- length(candidate_labels)

beta_x <- candidatesHDL$betaHDL

beta_y <- candidatesHDL$betaCAD

SE_beta_y <- abs(beta_y) / qnorm(1-candidatesHDL$pCAD/2)

# For confidence set in budgetIV/budgetIV_scalar.
alpha = 0.05
delta_beta_y <- qnorm(1 - alpha/(2*d_Z))*SE_beta_y
```

---

simulated_data_budgetIV

Simulated summary statistics with invalid instruments and nonlinear treatment effect

---

Description

Example dataset from the nonlinear simulation study using 6 candidate instruments, 3 of which are invalid with violation of IV assumptions (A2) and (A3). See Appx. C.2 of Penn et al. (2025) for technical details or visit the source code for reproducibility, both referenced below. The ground truth causal effect is $\Phi^*(X) = (X - 0.25)^2 - 0.25^2$. β_Φ is taken with respect to the basis functions $\Phi = (X, X^2)$.

Usage

```
data(simulated_data_budgetIV)
```

Format

A data frame with 6 rows and 4 columns.

Details

`beta_y` Components of the estimator $\mathrm{Cov}(Y, Z)$.

`beta_phi_1` Components of the estimator $\mathrm{Cov}(\Phi_1(X), Z)$.

`beta_phi_2` Components of the estimator $\mathrm{Cov}(\Phi_2(X), Z)$.

`delta_beta_y` Components of the standard error $\mathrm{Se}(\mathrm{Cov}(Y, Z))$.

Source

The code that generated this dataset was written by the authors and can be found in https://github.com/jpenn2023/budgetIVr/blob/main/paper/simulate_nonlinear_data.R. The dataset is saved as "my_dat R = 0.5 SNR_y = 1.csv".

References

Jordan Penn, Lee Gunderson, Gecia Bravo-Hermsdorff, Ricardo Silva, and David Watson. (2024). BudgetIV: Optimal Partial Identification of Causal Effects with Mostly Invalid Instruments. *arXiv* preprint, 2411.06913.

Examples

```
data(simulated_data_budgetIV)

beta_y <- simulated_data_budgetIV$beta_y

beta_phi_1 <- simulated_data_budgetIV$beta_phi_1
beta_phi_2 <- simulated_data_budgetIV$beta_phi_2

d_Z <- length(beta_phi_1)

beta_phi <- matrix(c(beta_phi_1, beta_phi_2), nrow = 2, byrow = TRUE)

delta_beta_y <- simulated_data_budgetIV$delta_beta_y
```

Index