Package: aciccomp2016 (via r-universe)

May 28, 2026

Version 0.1-0

Date 2017-05-20

Title Atlantic Causal Inference Conference Competition 2016 Simulation

Depends R (>= 3.2-2)

Imports stats, methods, utils

Suggests testthat

Description Generate simulation data.

License GPL (>= 2)

NeedsCompilation no

Biarch yes

LazyData yes

Repository https://mrcieu.r-universe.dev

Date/Publication 2020-07-08 20:56:10 UTC

RemoteUrl https://github.com/vdorie/aciccomp

RemoteRef HEAD

RemoteSha 282d26659b2d3d6fd060dde6d32feeb8f1e8ab5a

RemoteSubdir 2016

Contents

`constants_2016` *Constants Used in DGP for ACIC Competition 2016*

Description

Returns or sets elements of a named list containing all of the constants required to run the data generating processes for the 2016 ACIC Competition.

Usage

```
constants_2016(...)
```

Arguments

`...`	Options from the list below.

Details

Returns default values or sets them, as appropriate. Minimal error checking is performed.

Value

`RSP_INPUT_SCALE`
: Scaling factor applied to covariates before evaluating the response function.

`RSP_OUTPUT_SHAPE_1`
: The first shape parameter in a beta-prime used to generate the output scale of the response function.

`RSP_OUTPUT_RATE`
: The inverse scale parameter in a beta-prime used to generate the output scale of the response function.

`RSP_OUTPUT_SHAPE_2`
: The second shape parameter in a beta-prime used to generate the output scale of the response function.

`TRT_INPUT_SCALE`
: Scaling factor applied to covariates before evaluating the treatment assignment function.

`TRT_OUTPUT_SCALE`
: Scaling factor applied to result of the treatment assignment function.

`TRT_BIAS_SCALE`
: Approximate scale for treatment biasing functions when `overlap` parameter is not `"full"`.

`RSP_SIGMA_Y`
: Scale of noise added to response.

`BF_CONSTANT_SCALE`
: Scale of constant base function parameter.

`BF_LINEAR_SCALE`
: Scale of linear base function parameter.

BF_QUADRATIC_SHAPE_1
: First shape parameter used to generate quadratic base function root parameter.

BF_QUADRATIC_SHAPE_2
: Second shape parameter used to generate quadratic base function root parameter.

BF_QUADRATIC_RATE
: Rate parameter used to generate quadratic base function root parameter.

BF_QUADRATIC_SCALE
: Scale of quadratic base function parameter.

BF_CUBIC_SHAPE
: Shape parameter used to generate cubic base function root parameters.

BF_CUBIC_RATE
: Rate parameter used to generate cubic base function root parameters.

BF_CUBIC_SCALE
: Scale of cubic base function parameter.

BF_CONTINUOUS_SCALE
: Scale parameter shared by continuous base functions.

BF_STEP_SHAPE
: Shape parameter for step base functions.

BF_STEP_CONSTANT_SCALE
: Scale of step-wise constant base function parameter.

BF_STEP_LINEAR_SCALE
: Scale of piece-wise linear base function parameter.

BF_SIGMOID_SHAPE_1
: First shape parameter used to generate sigmoid base function parameters.

BF_SIGMOID_RATE_1
: First rate parameter used to generate sigmoid base function parameters.

BF_SIGMOID_SHAPE_2
: Second shape parameter used to generate sigmoid base function parameters.

BF_SIGMOID_RATE_2
: Second rate parameter used to generate sigmoid base function parameters.

BF_QUANTILE_SHAPE_1
: First shape parameter used to generate quantile base function cutoff.

BF_QUANTILE_SHAPE_2
: Second shape parameter used to generate quantile base function cutoff.

BF_TWEAK_SIGN_PROB
: Probability of changing sign when copy/modifying base function.

BF_TWEAK_NORMAL_SCALE
: Scale of normal noise added to unconstrained base function parameters when copy/modifying.

BF_TWEAK_GAMMA_SHAPE
: Shape parameter of positive noise added to constrained base function parameters when copying/modifying.

BF_TWEAK_GAMMA_RATE
: Rate parameter of positive noise added to constrained base function parameters when copying/modifying.

TRT_BF_DF
: Base function degrees of freedom when generating treatment assignment mechanism.

RSP_BF_DF
: Base function degrees of freedom when generating response surface.

TRT_LINEAR_SCALE_SHAPE_1
: First scale parameter used to generate overall scale of treatment assignment mechanism.

TRT_LINEAR_SCALE_SHAPE_2
: Second scale parameter used to generate overall scale of treatment assignment mechanism.

TRT_LINEAR_SCALE_RATE
: Rate parameter used to generate overall scale of treatment assignment mechanism.

RSP_EXP_SCALE_SHAPE
: Shape parameter used when generating scale factor for exponential functions.

RSP_EXP_SCALE_RATE
: Rate parameter used when generating scale factor for exponential functions.

RSP_EXP_WEIGHT_SHAPE
: Shape parameter used when generating relative weight factor for exponential functions.

RSP_EXP_WEIGHT_RATE
: Rate parameter used when generating relative weight factor for exponential functions.

RSP_TE_MEAN
: Expected value for population average treatment effect.

RSP_TE_SCALE
: Scale factor for population average treatment effect.

RSP_TE_DF
: Degrees of freedom for population average treatment effect.

SPARSE_COVARIATE_WEIGHT
: Weight of inclusion for sparse, discrete covariates.

CONTINUOUS_COVARIATE_WEIGHT
: Weight of inclusion for continuous covariates.

DEFAULT_COVARIATE_WEIGHT
: Default weight of inclusion for covariates.

TRT_BASELINE_SHIFT
: Function used to derive a scale when generating a baseline treatment probability from `root.trt`

BASE_FUNCTION_DIST_LIN
: Base function distribution containing only linear functions.

BASE_FUNCTION_DIST_POLY
: Base function distribution containing linear, quadratic, and cubic functions.

BASE_FUNCTION_DIST_STEP
: Base function distribution containing linear, step-wise constant, and piece-wise linear functions.

BASE_FUNCTION_DIST_EXP
: Base function distribution containing third order polynomials to be used in exponential functions.

`dist.lin`
: Function distribution for purely linear treatment or response.

`dist.int`
: Function distribution with linear terms and interactions.

`dist.pure.poly`
: Function distribution with quadratic terms and no interactions.

`dist.poly`
: Function distribution with cubic terms and interactions.

dist.step	Function distribution with linear terms, step-wise constant terms, and interactions.
dist.exp	Function distribution with quadratic terms and interactions appropriate for use with exponential link functions.
dist.bias1	Function distribution over treatment assignment biasing functions.
dist.bias2	Function distribution over treatment assignment biasing functions.
dist.hetero.med	Function distribution specifying interaction retention probabilities for medium degrees of treatment effect heterogeneity.
dist.hetero.high	Function distribution specifying interaction retention probabilities for high degrees of treatment effect heterogeneity.

Author(s)

Vincent Dorie: <vdorie@gmail.com>.

---

dgp_2016 *Data Generating Process for the 2016 ACIC Competition*

Description

Applies the data generating process used in the Atlantic Causal Inference Competition of 2016.

Usage

```
dgp_2016(x, parameters, random.seed,
         constants = constants_2016(),
         extraInfo = FALSE)
```

Arguments

x	Input data in the form of a data frame, most likely input_2016.
parameters	A named list containing elements in the form of parameters_2016, a row of the same object, or an integer specifying which row of parameters_2016 is to be used; see that page for details.
random.seed	A list of arguments to be used in a call to set.seed or an integer between 1 and 100 specifying the random seed associated with an iteration from the competition.
constants	A named list containing elements as returned by constants_2016; see there for details.
extraInfo	Boolean determining if additional information is to be returned, including the treatment and control response surfaces, the transformed input data, and whether or not a simulation would have been deemed interesting enough to include in the competition.

Details

Creates a causal inference problem by taking the input `x` and using the passed in `parameters` to generate a treatment assignment mechanism (probability of treatment for each individual), response surface (expected value under treatment and control), and finally observed data. The `parameters` provide high-level controls to adjust the result for causal inference features that may be of interest, while `constants` control at a lower level the parameters of generated functions.

Generalized Additive Functions: The 2016 competition used a unique set of software that was internally described as "Generalized Additive Functions" (GAFs). A GAF consists of many small functions applied to various features/columns of the input that are added together or interacted with each other. The complete sum may then be passed through a link function to achieve a result in a transformed space. The small functions are randomly derived from a library of functions, so that the general features of the result can vary according to high level parameters.

This package reproduces GAFs as they were used in the 2016 contest without the intention that they be further applied. It may be possible to use `dgp_2016` with different input data and changes to the `constants` should propogate, however these extensions will not be widely supported.

Value

A named list containing:

`z`	Vector of treatment assignments. If `extraInfo` is `FALSE`, `z` contains 0s and 1s. If `TRUE`, `z` is a factor with levels `ctl` and `trt`.
`y`	Vector of observed response variables, $y(z)$.
`y.0`	Vector of response variables under the control condition, $y(0)$.
`y.1`	Vector of response variables under the treatment condition, $y(1)$.
`mu.0`	Vector of expected response under the control condition, $E[Y(0)]$.
`mu.1`	Vector of expected response under the treatment condition, $E[Y(1)]$.
`e`	Vector of propensity scores, $P(Z = 1)$.
`f.z`	Optional - the GAF for the treatment assignment mechanism.
`f.y`	Optional - the GAF for the response surface.
`x`	Optional - the transformed input passed to `f.z` and `f.y`.
`valid`	Optional - boolean if the simulation would be rejected as "uninteresting".

Author(s)

Vincent Dorie: <vdorie@gmail.com>.

References

Dorie V., Hill J., Shalit U., Scott M. and Cervone D. (2017) Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, preprint arXiv https://arxiv.org/abs/1707.02641.

Examples

```
## Not run:
# to test a method
ate <- matrix(NA, 77, 100)
for (i in seq_len(77)) {
  for (j in seq_len(100)) {
    sim <- dgp_2016(input_2016, i, j)
    df <- input_2016
    df$y <- sim$y
    df$z <- sim$z
    fit <- lm(y ~ ., df)
    ate[i,j] <- coef(fit)["z"]
  }
}

## undocumented features, getting closest approximate linear model
sim <- dgp_2016(input_2016, 1, 1, extraInfo = TRUE)

e <- aciccomp:::evaluate(sim$f.z, sim$x)
x.z.approx <- aciccomp:::evaluateGeneralizedAdditiveFunctionToDataframe(sim$f.z, sim$x)

x.temp <- sim$x
x.temp$.z <- sim$z
x.y.approx <- aciccomp:::evaluateGeneralizedAdditiveFunctionToDataframe(sim$f.y, x.temp)

## End(Not run)
```

`input_2016` *Input Data for the 2016 ACIC Competition*

Description

Input data used in the 2016 Atlantic Causal Inference Competition, taken from the Collaborative Perinatal Project.

Usage

```
input_2016
```

Format

A data frame consisting of 4802 observations and 58 covariates. The columns have been de-identified from their original source, but correspond to possible confounders, instruments, and uncorrelated variables from a hypothetical twin study on the impact of birthweight on IQ.

Details

The variable in the original CPP are:

- mom_age
- mar_status
- mom_cigs_per_day
- mom_years_smoked
- mom_height
- mom_weight_prior
- mom_num_cardio_cond
- mom_num_pulm_cond
- mom_num_hema_cond
- mom_num_endocrine_cond
- mom_num_veneral_cond
- mom_num_urin_cond
- mom_num_gyne_cond
- mom_num_neur_cond
- mom_num_obst_compl
- mom_num_infect_dis
- mom_work_status
- mom_years_educ
- family_income
- housing_density
- mom_birth_place
- consanguinity
- socio_eco
- mom_race
- age_menarche
- dias_blood_pres
- mom_weight_birth
- dad_age
- dad_years_educ
- num_premes
- num_abortions
- num_prior_pregs
- num_stillbirths
- bayley_mental
- bayley_motor

- placental_weight
- cord_length
- sex
- apgar_1m_total
- apgar_5m_total
- bottle_feed_days
- breast_feed_days
- child_bilirubin
- child_hematocrit
- child_hemoglobin
- child_num_neur_abn
- child_num_cns_cond
- child_num_muscoskel
- child_num_resp_abn
- child_num_cardio_abn
- child_num_liver_abn
- child_num_hemo_cond
- child_num_infect
- child_num_synd
- child_num_endo_dis
- child_num_proc
- head_size_1yr
- gest_delivery

Source

Niswander, K. R. and Gordon, M. (1972) The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke: the women and their pregnancies. Philadelphia, PA: W.B. Saunders Company https://www.archives.gov/research/electronic-records/nih.html

---

parameters_2016 *Parameters Data for the 2016 ACIC Competition*

Description

Data set containg the parameters used to generate data for the 2016 Atlantic Causal Inference Conference competition.

Usage

```
parameters_2016
```

Format

A data frame describing 77 scenarios that vary across 6 features.

1. `model.trt` - Function distribution over the treatment assignment mechanism. Can be `"linear"`, `"polynomial"`, or `"step"`.
2. `root.trt` - Baseline probability of receiving treatment.
3. `overlap.trt` - Term that controls the addition of overlap-penalizing terms that forcibly exclude observations from the treatment group by carving out hyper-rectangles of the covariate space and assigning their treatment probability to 0. Can be `"full"` for complete overlap, `"one-term"` for adding a single function as described above, or `"two-term"` for adding two. Two-terms were not used in the competition and is not thoroughly tested.
4. `model.rsp` - Function distribution over the response surface. Can be `"linear"`, `"polynomial"`, `"step"`, or `"exponential"`.
5. `alignment` - A numeric value that determines the degree to which terms from the treatment assignment function appear in response surface function.
6. `te.hetero` - A term that controls the degree of treatment effect heterogeneity. Can be `"none"` for parallel surfaces, `"med"` or `"high"`. Higher heterogeneity is achieved by selectively interacting terms from the response surface with a treatment indicator.

Source

Original release.

References

Dorie V., Hill J., Shalit U., Scott M. and Cervone D. (2017) Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition, preprint arXiv https://arxiv.org/abs/1707.02641.

Index