Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

causalglm: A tlverse R package for interpretable and robust causal inference for heterogeneous treatment effects using generalized linear working models and targeted machine-learning.

Lars van der Laan
University of Washington, Seattle/Statistics

Abstract

Generalized linear models are widely-used and interpretable methods for learning heterogeneous treatment effects. However, generalized linear models methods, as commonly used in practice, make strong parametric assumptions on components of the data-generating distribution that are not directly of interest for the causal question at hand. Notably, in real-world settings, the relation between confounders and the outcome is almost always unknown and may be quite complex. As a result, causal inference based on parametric generalized linear models can be very biased and even misleading, especially when the model is not chosen a priori. Moreover, in high dimensional settings, generalized linear models and their inference breaks down and alternative methods like lasso and elastic-net regression do not easily provide inference. In this article, we present the R package **causalglm** that implements nonparametrically robust causal inference for user-specified generalized linear working (or approximate) models for heterogeneous treatment effects. The user-specified parametric model is viewed as an approximation of the true nonparametric treatment-effect estimand and an interpretable causal estimand is defined as the nonparametric projection of this true estimand onto the working model, thereby allowing for estimates that are causally interpretable and valid inference under a nonparametric statistical model. These methods utilize targeted maximum likelihood estimation and therefore can leverage adaptive machine-learning algorithms for estimation and variable selection of nonparametric nuisance components of the data-distribution. For binary, categorical and continuous treatments, this package supports nonparametric causal inference for user-specified parametric working models of the estimands: the conditional average treatment effect, the conditional relative risk, the conditional odds ratio, and more. Nonparametric-robust causal inference for working marginal structural models are also supported. In addition, a specialized lasso-based method is implemented, allowing for robust post-confounder-selection causal inference in very high dimensions. These methods are implemented using the powerful **tlverse** machine-learning (**sl3**) and generalized targeted learning (**tmle3**) ecosystem.

Keywords: keywords, not capitalized, Java.

1. Introduction to causalglm

In the search for answers to causal questions, assuming parametric models can be dangerous. With even a seemingly small amount of confounding and model misspecificaton, they can give biased answers. One way of mitigating this challenge is to only parametrically model the feature of the data-generating distribution that you care about. It is not even necessary to assume your parametric model is correct! Instead, view it as a "working" or "approximation" model and define your estimand as the best causal approximation of the true nonparametric estimand with respect to your parametric working model. This allows for causal estimates and robust inference under no parametric assumptions on the functional form of any feature of the data generating distribution. Alternatively, you can assume a semiparametric model that only assumes the parametric form of the relevant part of the data distribution is correct. Let the data speak for itself and use machine-learning to model the nuisance features of the data that are not directly related to your causal question. It is in fact possible to get robust and efficient inference for causal quantities using machine-learning. Why worry about things that don't matter for your question? It is not worth the risk of being wrong.

causalglm is an all-purpose and user-friendly package for interpretable and robust causal inference for heterogeneous treatment effects using generalized linear working model and machine-learning. Unlike parametric methods based on generalized linear models and semiparametric methods based on partially-linear models, the methods implemented in **causalglm** do not assume that any user-specified parametric models are correctly specified. That is, **causalglm** does not assume that the true data-generating distribution satisfies the parametric model. Instead, the user-specified parametric model is viewed as an approximation or "working model", and an interpretable estimand is defined as a projection of the true conditional treatment effect estimand onto the working model. Moreover, **causalglm**, unlike **glm**, only requires a user-specified parametric working model for the causal estimand of interest. All nuisance components of the data-generating distribution are left unspecified and data-adaptively learned using machine-learning. Thus, **causalglm** provides not only nonparametrically robust inference but also provides estimates (different from **glm**) for causally interpretable estimands that maximally adjust for confounding. To allow for valid inference with the use of variable-selection and machine-learning, Targeted Maximum Likelihood Estimation (van der Laan, Rose, 2011) is employed.

By providing inference under a fully nonparametric statistical model, as opposed to a parametric or even semiparametric statistical model, **causalglm** gains the following features:

1. No model selection bias: Users can specify as many incompatible parametric working models for the treatment effect estimands as desired and obtain valid inference even when the models are incorrect.

2. Causal estimands: The estimands estimated can be meaningfully interpreted as causal parameters (e.g. a best linear approximation) even when the working model is incorrect. Moreover, the working-model-based estimands are invariant to the level of confounding in the data. Two different studies with different levels of confounding and different treatment-assignment probabilities will estimate the same target parameter.

3. Adaptive variable-selection techniques and machine-learning can be used to adjust for confounding, thus allowing for maximal confounding bias reduction and robust inference in both low, high and very high dimensions.

causalglm consists of the following functions.

1. 'npglm' for robust nonparametric inference of user-specified working-models for conditional treatment effects with binary and categorical treatments.
2. 'msmglm' for robust nonparametric inference of user-specified working marginal structural models for marginalized conditional treatment effects with binary and categorical treatments.
3. 'contglm' for robust nonparametric inference of user-specified working models for conditional treatment effects with continuous treatments.
4. 'spglm' for semiparametric inference of assumed-to-be correctly-specified parametric models for conditional treatment effects with binary treatments.
5. 'causalglmnet' for semiparametric inference of assumed-to-be correctly-specified parametric models for conditional treatment effects in high dimensions with binary treatments.

2. Data-Structure and treatment-effect estimands

We will mainly consider the point-treatment data-structure $O = (W, A, Y)$ where W represents a vector of baseline covariates (i.e. possible confounders), A is a binary, categorical or continuous treatment assignment, and Y is some outcome variable. As an example, for a given observation O, W could be measurements: age, sex, a risk-score, location, income; A could be categorical and take the value a if the individual receives treatment a and 0 if they do not receive any treatment; and Y is a binary or continuous variable that captures the effect of the treatment. Additionally, for the marginal structural model estimands, we will also consider a subvector of baseline variables $V \subset W$. For the goal of assessing heterogeneity of the treatment effect, there are a number of popular estimands that are implemented in this package:

The conditional average treatment effect (CATE):

$$CATE_a(w) := E[Y|A = a, W = w] - E[Y|A = 0, W = w], \tag{1}$$

which is an additive measure of the effect of the treatment $(A = a)$ relative to no treatment $(A = 0)$.

The conditional odds ratio (OR) for when Y is binary:

$$OR_a(w) := \frac{P(Y = 1|A = a, W = w)/P(Y = 0|A = 1, W = w)}{P(Y = 1|A = 0, W = w)/P(Y = 0|A = 0, W = w)} \tag{2}$$

The conditional relative risk (RR) for when Y is nonnegative (e.g. a binary or count variable):

$$RR_a(w) := \frac{E[Y|A = a, W = w]}{E[Y|A = 0, W = w]}, \tag{3}$$

which is a relative measure of the effect of the treatment $(A = a)$ relative to no treatment $(A = 0)$.

In some application, non-contrast measures like the conditional treatment-specific mean (TSM) may be of interest:

$$TSM_a(w) := E[Y|A = a, W = w]. \tag{4}$$

Often, we are only interested in the treatment effect as a function of a subset V of the collected variables W. In such settings, we would still like to adjust for all variables W to minimize confounding bias but only wish to model the treatment effect as a function of V. Marginal structural models provide a rich class of causal estimands that accomplish this. Key estimands considered in this package are

The V-specific conditional average treatment effect (V-CATE):

$$E[CATE_a(W)|V = v] := E\left\{E[Y|A = a, W] - E[Y|A = 0, W] \mid V = v\right\}. \tag{5}$$

The V-specific conditional average treatment effect among the treated (V-CATT):

$$E[CATE_a(W)|V = v, A = a] := E\left\{E[Y|A = a, W] - E[Y|A = 0, W] \mid V = v, A = a\right\}. \tag{6}$$

The V-specific conditional relative risk (V-RR) for when Y is nonnegative (e.g. a binary or count variable):

$$E[RR_a(W)|V = v] := \frac{E[E[Y|A = a, W] \mid V = v]}{E[E[Y|A = 0, W] \mid V = v]}. \tag{7}$$

The V-specific conditional treatment-specific mean (V-TSM) may be of interest:

$$E[TSM_a(W) \mid V = v] := E[E[Y|A = a, W] \mid V = v]. \tag{8}$$

2.1. Conventional estimators using parametric generalized linear models and semiparametric generalized partially-linear models

In order to estimate the estimands of the previous section, parametric generalized linear models are often employed (e.g. the R package **glm**). For the CATE, the following linear regression model is often used:

$$E[Y|A, W = w] = \beta_1^T \underline{f}_1(w) \cdot A + \beta_2^T \underline{f}_2(w),$$

where $\underline{f}_1(w)$, $\underline{f}_2(w)$ are user-specified vector-valued functions that encode the parametric model. This model is equivalent to assuming both the nuisance linear model

$$E[Y|A = 0, W = w] = \beta_2^T \underline{f}_2(w)$$

and target linear model

$$CATE(w) = \beta_1^T \underline{f}_1(w).$$

An immediate drawback of this model is that strong parametric assumptions are made not only on the target estimand $CATE(w)$ but also on the orthogonal nuisance function $E[Y|A=0,W]$ which is not meaningfully related to the estimand of interest\$. As a consequence, unnecessary assumptions are made that do not provide any meaningful benefit in interpretability and can only bias the estimates for the target estimand.

A more robust way of estimating the CATE is to assume a semiparametric model. Specifically, we could assume the partially-linear linear-link regression model which only assumes that

$$CATE(w) = \beta^T \underline{f}(w).$$

Thus, only the relevent feature of the data-generating distribution, the actual estimand, is modeled parametrically. This assumption is equivalent to the regression model:

$$E[Y|A, W=w] = A \cdot \beta^T \underline{f}(w) + h_0(w),$$

where $h_0(w) := E[Y|A=0,W]$ is left unspecified. While the semiparametric model makes assumptions much weaker than the parametric model, it still makes strong parametric assumptions on the target estimand. In particular, the semiparametric still requires apriori correctly specifying a parametric model for the estimand for valid inference and therefore suffers from issues like model selection bias. Also, it is not easy to generalize the estimators based on these models to marginal structural model estimands.

Before going to the other estimands, it will be useful to generalize the parametric and semiparametric model considered above for general link functions. For a given monotone link function $g : \mathbb{R} \to \mathbb{R}$, the generalized linear model assumes

$$g(E[Y|A, W=w]) = \beta_1^T \underline{f}_1(w) \cdot A + \beta_2^T \underline{f}_2(w),$$

where $\underline{f}_1(w)$, $\underline{f}_2(w)$ are user-specified vector-valued functions that encode the parametric model. The generalized partially linear model assumes

$$g(E[Y|A, W=w]) = A \cdot \beta^T \underline{f}(w) + g(E[Y|A=0,W]),$$

where $\underline{f}(w)$ is a user-specified vector-valued function that encodes a parametric model for the estimand of interest and $E[Y|A=0,W]$ is left unspecified.

The CATE-based models we just considered correspond with the identity link function $g(x) := x$. The conditional odds ratio can be estimated using the parametric logistic regression model and semiparametric partially-linear logistic regression model, which correspond with the logistic link $g(p) = \frac{p}{1-p}$. The parametric logistic regression model is equivalent to assuming both the nuisance logistic model

$$\text{logit}\left\{E[Y|A=0, W=w]\right\} = \beta_2^T \underline{f}_2(w)$$

and target log-linear model

$$\log OR(w) = \beta_1^T \underline{f}_1(w).$$

The semiparametric method only assumes that $\log OR(w) = \beta^T \underline{f}(w)$.

For estimation of the conditional relative risk, the link function is chosen to be $g(x) = \log x$, which corresponds with the parametric and partially-linear log-linear models. The parametric log-linear model is equivalent to assuming both the nuisance log-limear model

$$\log\left\{E[Y|A=0, W=w]\right\} = \beta_2^T \underline{f}_2(w)$$

and target log-linear model

$$\log RR(w) = \beta_1^T \underline{f}_1(w).$$

The semiparametric partially-linear log-linear model only assumes that

$$\log RR(w) = \beta^T \underline{f}(w)$$

and leaves $E[Y|A = 0, W]$ unspecified. Both the conditional odds ratio and relative risk parametric and semiparametric models suffer from similar issues as the CATE models.

3. Nonparametric working-model-based estimands for model-free inference

In this section, we introduce nonparametrically-defined estimands based on projections of the true conditional treatment-effect estimands onto parametric working-models. We consider the estimands based on binary and categorical treatments separately from the estimands based on continuous treatments.

3.1. Working-model-based nonparametric estimands for binary or categorical treatments

Let a be a given treatment level of A that is of interest. Consider the least-squares projection estimand,

$$\beta_{a,CATE} := \text{argmin}_\beta \, E\left(CATE_a(W) - \beta^T \underline{f}(W)\right)^2$$

where $\beta_{a,CATE}^T \underline{f}(w)$ is the best linear approximation of the true $CATE_a(W)$ relative to the average squared-error loss. Clearly, when the parametric working-model $\beta^T \underline{f}(w)$ is correctly specified, $\beta_{a,CATE}$ reduces to the same estimand considered by the parametric and semiparametric methods. When the working-model is incorrect, the causal interpretation of $\beta_{a,CATE}$ as an approximation is clear. Importantly, $\beta_{a,CATE}$ is invariant with respect to the treatment-assignment probability $P(A = a|W)$, which crucially ensures this nonparametric estimand is not affected by confounding.

This same estimand also has the property that it reduces to the best approximation of a $V-specific$ marginal structural working model when the parametric form is lower-dimensional and satisfies $\underline{f}(W = w) \equiv \underline{f}(V = v)$. To see this, note that minimizing $E\left(CATE_a(W) - \beta^T \underline{f}(V)\right)^2$ is equivalent to minimizing $E\left\{\left(\beta^T \underline{f}(V)\right)^2 - 2\beta^T \underline{f}(V) CATE_a(W)\right\}$. Applying the law of iterated conditional expectations and applying the equivalence in reverse, we find

$$\text{argmin}_\beta \, E\left(CATE_a(W) - \beta^T \underline{f}(V)\right)^2 = \text{argmin}_\beta \, E\left(E[CATE_a(W)|V] - \beta^T \underline{f}(V)\right)^2.$$

Thus, the estimand $\beta_{a,CATE}$ has the powerful property of automatically reducing to the least-squares projection of the true V-specific marginal structural CATE model onto the user-specified working model $\beta^T \underline{f}(V = v)$. This allows for the development of a single estimator/method that simultaneously allows for estimates and inference for both conditional estimands and marginal structual model estimands.

A notable special case if when $\underline{f}(w) := 1$ is taken as the intercept model. We then find

$$\beta_{a,CATE} = \text{argmin}_\beta \, E\left(CATE_a(W) - \beta\right)^2 = E[CATE(W)] = E[E[Y|A = a, W]] - E[E[Y|A = 0, W]]$$

which is exactly the nonparametric marginal average treatment effect estimand!

Based on the least-squares projection, we can define natural working-model estimands for the conditional treatment-specific mean (TSM) and the conditional average treatment effect among the treated:

$$\beta_{a,CATT} := \text{argmin}_\beta \, E\left\{\left(CATE_a(W) - \beta^T \underline{f}(W)\right)^2 \mid A = a\right\}, \tag{9}$$

$$\beta_{a,TSM} := \text{argmin}_\beta \, E\left(E[Y|A = a, W] - \beta^T \underline{f}(W)\right)^2. \tag{10}$$

Both these working model estimands similarly reduce to projections of the true marginal structural model estimands for lower dimensional working-models. Specifically, we have the identities

$$\text{argmin}_\beta \, E\left\{\left(E[CATE_a(W) - \beta^T \underline{f}(V)\right)^2 \mid A = a\right\}$$

$$= \text{argmin}_\beta \, E\left\{\left(E[CATE_a(W)|V, A = a] - \beta^T \underline{f}(V)\right)^2 \mid A = a\right\},$$

and

$$\text{argmin}_\beta \, E\left(E[Y|A = a, W] - \beta^T \underline{f}(V)\right)^2 = \text{argmin}_\beta \, E\left(E[E[Y|A = a, W]|V] - \beta^T \underline{f}(V)\right)^2.$$

Next, we consider the nonparametric working-model-based estimands for conditional relative risk (RR) which will utilize projections based on the poisson (log-linear) log-likelihood loss function. Specifically, define the poisson-likelihood-type projection estimand:

$$\beta_{a,RR} := \text{argmin}_\beta \, E\left\{E[Y|A = 0, W] \cdot e^{\beta^T \underline{f}(W)} - E[Y|A = a, W] \cdot \beta^T \underline{f}(W)\right\}. \tag{11}$$

It can be verified when $\beta^T \underline{f}(W)$ is correctly specified for the conditional relative risk that $\beta_{a,RR}$ indeed reduces to the true coefficient vector. While this estimand is motivated by the poisson and log-linear loss functions, we note that it is well-defined nonparametrically and we make no assumptions on the error distribution of our outcome. One reason for choosing this projection to define the estimand is because it reduces to the projection of a marginal structural model for the relative treatment effect when $\underline{f}(W = w) \equiv \underline{f}(V = v)$ is lower-dimensional. Specifically, we have the identity

$$\text{argmin}_\beta \, E\left\{E[Y|A = 0, W] \cdot e^{\beta^T \underline{f}(V)} - E[Y|A = a, W] \cdot \beta^T \underline{f}(V)\right\}$$

$$= \text{argmin}_\beta \, E\left\{E[E[Y|A = 0, W]|V] \cdot e^{\beta^T \underline{f}(V)} - E[E[Y|A = a, W]V] \cdot \beta^T \underline{f}(V)\right\},$$

where the right-hand side is the log-linear projection of the V-specific marginalized conditional relative risk function $\frac{E[E[Y|A=a,W]V]}{E[E[Y|A=0,W]V]}$ onto the working-model $\beta^T \underline{f}(V = v)$. Under the special case where the intercept working model $\underline{f}(w) := 1$ is used, we have

$$\beta_{a,RR} = \frac{E[E[Y|A = a, W]]}{E[E[Y|A = 0, W]]}$$

, which is the nonparametric marginal relative risk.

Finally, we present a working-model for the conditional odds ratio based on the logistic-link log-likelihood projection. First, define the working-model

$$P_\beta(Y=1|A=a,W) := \text{expit}\left\{\beta^T \underline{f}(W) + \text{logit}(P(Y=1,A=0,W))\right\}.$$

Consider the estimand,

$$\beta_{a,OR} = \text{argmin}_\beta -1 \cdot E\Big\{ P(Y=1|A=a,W)\log\{P_\beta(Y=1|A=a,W)\}$$

$$+P(Y=0|A=a,W)\log\{1-P_\beta(Y=1|A=a,W)\}\Big\}, \tag{12}$$

which is the log-likelihood projection of $P(Y=1|A=a,W)$ onto the working model $P_\beta(Y=1|A=a,W)$. Since $P_\beta(Y=1|A=0,W) = P_\beta(Y=1|A=0,W)$ is correctly specified, this projection is targeted towards $\beta^T\underline{f}(w)$ which is a log-linear working model for the conditional odds ratio $OR_a(w)$. This estimand only reduces to a V-specific marginal structural working model conditional odds ratio estimand when $P(Y=1|A=0,W) = P(Y=1|A=0,V)$, which may not be true in practice but is still a useful property to have.

3.2. Robust working-model-based nonparametric estimands for continuous treatments

When A represents a continuous or ordered numeric treatment, we utilize working-models that also model the treatment-dependence of the conditional treatment effect estimand. For computational reasons, it is beneficial to solely consider least-squares-type projections for all estimands. Specifically, for a given link function $g:\mathbb{R}\to\mathbb{R}$, we consider working models of the form

$$g(E[Y|A=a,W=w]) - g(E[Y|A=0,W=w]) = 1(a>0)\cdot\beta_0^T\underline{f}_0(w) + a\cdot\beta_1^T\underline{f}_1(w),$$

where we assume that the treatment A is nonnegative and $A=0$ encodes the control arm. The first term models discontinuous treatment effects associated with being either treated or not. The second term models continuous or dosage-dependent treatment effects. We define a rich class of nonparametrically-defined working-model-based estimands via the least-squares projection

$$(\beta_{0,g},\beta_{1,g}) := \text{argmin}_{\beta_0,\beta_1} E\left\{g(E[Y|A,W]) - g(E[Y|A=0,W]) - 1(A>0)\cdot\beta_0^T\underline{f}_0(W) + A\cdot\beta_1^T\underline{f}_1(W)\right\}^2.$$

For the identity link $g(x) := x$, we obtain as estimand the least-squares projection of the true CATE

$$(\beta_{0,CATE},\beta_{1,CATE}) := \text{argmin}_{\beta_0,\beta_1} E\left\{CATE_A(W) - 1(A>0)\cdot\beta_0^T\underline{f}_0(W) + A\cdot\beta_1^T\underline{f}_1(W)\right\}^2.$$

For the log link $g(x) = \log x$, we obtain as estimand the least-squares projection of the true conditional log RR

$$(\beta_{0,RR},\beta_{1,RR}) := \text{argmin}_{\beta_0,\beta_1} E\left\{\log RR_A(W) - 1(A>0)\cdot\beta_0^T\underline{f}_0(W) + A\cdot\beta_1^T\underline{f}_1(W)\right\}^2.$$

For the log-odds link $g(p) = \log p - \log(1-p)$, we obtain as estimand the least-squares projection of the true conditional log OR

$$(\beta_{0,OR}, \beta_{1,OR}) := \text{argmin}_{\beta_0,\beta_1} E\left\{\log OR_A(W) - 1(A>0)\cdot \beta_0^T \underline{f}_0(W) + A \cdot \beta_1^T \underline{f}_1(W)\right\}^2.$$

4. Model-robust causal inference with npglm and msmglm

npglm is one of the main functions of **causalglm** and implements statistically efficient estimators for the nonparametrically-defined working-model-based conditional treatment effect estimands defined in Section 3.1. *msmglm* is a specialized wrapper function for *npglm* that focuses on marginal structural working model estimation and provides additional plotting features for such models. Both methods have user-friendly front-end and only require the following arguments:

1. *formula* : An R formula object that specifies a parametric working model for the conditional treatment-effect estimand.
2. *data*: A data.frame containing the baseline, treatment and outcome variables.
3. W: A character vector containing the names of the variables in *data* for which to adjust (i.e. possible confounders).
4. A: A character string giving the name of the treatment variable of interest in *data*.
5. Y: A character string giving the name of the outcome variable of interest in *data*.
6. *estimand*: A character string specifying the estimand - *CATE, OR, RR, CATT, TSM*.

Additionally, the following optional arguments are useful.

1. *learning_method* : A character string specifying one of the built-in machine-learning algorithms for nuisance function estimation. The nuisance functions are $P(A=1|W)$ and $E[Y|A,W]$.
2. *treatment_level*: A value/level of the treatment A that encodes the treatment arm of interest. (useful for categorical treatments)
3. *control_level*: A value/level of the treatment A that encodes the control arm. (useful for categorical treatments)
4. *sl3_Learner_A*: A **sl3** learner object that specifies a custom estimator for the treatment-assignment probability $P(A=1|W)$. (Overrides *learning_method*)
5. *sl3_Learner_Y*: A **sl3** learner object that specifies a custom estimator for the true outcome conditional mean $E[Y|A,W]$. (Overrides *learning_method*)

Affiliation:

Lars van der Laan
University of Washington, Seattle/Statistics
University of Washington, Seattle
Department of Statistics
E-mail: vanderlaanlars@yahoo.com
URL: https://tlverse.org/causalglm/

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd