vignette

Introduction to causalglm

causalglm is an R package for robust generalized linear models and interpretable causal inference for heterogeneous (or conditional) treatment effects. Specifically, causalglm very significantly relaxes the assumptions needed for useful causal estimates and correct inference by employing semi and nonparametric models and adaptive machine-learning through targeted maximum likelihood estimation (TMLE). See the writeup causalglm.pdf for a more theoretical overview of the methods implemented in this package.

The statistical data-structure used throughout this package is $O = (W, A, Y)$ where W represents a random vector of baseline (pretreatment) covariates/confounders, A is a usually binary treatment assignment with values in $c(0, 1)$, and Y is some outcome variable. For marginal structural models, we also consider a subvector $V \subset W$ that represents a subset of baseline variables that are of interest.

The estimands supported by causalglm are

1. Conditional average treatment effect (CATE) for arbitrary outcomes: $E[Y|A = 1, W] - E[Y|A = 0, W]$
2. Conditional odds ratio (OR) for binary outcomes: $\frac{P(Y=1|A=1,W)/P(Y=0|A=1,W)}{P(Y=1|A=0,W)/P(Y=0|A=0,W)}$
3. Conditional relative risk (RR) for binary, count or nonnegative outcomes: E[Y|A=1,W]/E[Y|A=0,W]
4. Conditional treatment-specific mean (TSM) : $E[Y|A = a, W$
5. Conditional average treatment effect among the treated (CATT) : the best approximation of E[Y|A=1,W] - E[Y|A=0,W] based on a user-specified formula/parametric model among the treated (i.e. observations with $A = 1$)

causalglm also supports the following marginal structural model estimands:

1. Marginal structural models for the CATE: $E[CATE(W)|V] := E[E[Y|A = 1, W] - E[Y|A = 0, W]|V]$
2. Marginal structural models for the RR: $E[E[Y|A = 1, W]|V]/E[E[Y|A = 0, W]|V]$
3. Marginal structural models for the TSM : $E[E[Y|A = a, W]|V]$
4. Marginal structural models for the CATT : $E[CATE(W)|V, A = 1] := E[E[Y|A = 1, W] - E[Y|A = 0, W]|V, A = 1]$

causalglm consists of four main functions:

1. spglm for semiparametric estimation of correctly specified parametric models for the CATE, RR and OR
2. npglm for robust nonparametric estimation for user-specified approximation models for the CATE, CATT, TSM, RR or OR
3. msmglm for robust nonparametric estimation for user-specified marginal structural models for the CATE, CATT, TSM or RR
4. causalglmnet for high dimensional confounders W (a custom wrapper function for spglm focused on big data where standard ML may struggle)

spglm is a semiparametric method which means that it assumes the user-specified parametric model is correct for inference. This method should be used if you are very confident in your parametric model. npglm is a nonparametric method that views the user-specified parametric model as an approximation or working-model for the true nonparametric estimand. The estimands are the best causal approximation of the true conditional estimand (i.e. projections). Because of this model agnostic view, npglm provides interpretable estimates and correct inference under no conditions. The user-specified parametric model need not be correct or even a good approximation for inference! npglm should be used if you believe your parametric model is a good approximation but are not very confident that it is correct. Also, it never hurts to run both spglm and npglm for robustness! If the parametric model is close to correct then the two methods should give similar estimates. Finally, msmglm deals with marginal structural models for the conditional treatment effect estimands. This method is useful if you are only interested in modeling the causal treatment effect as a function of a subset of variables V adjusting for all the available confounders W that remain. This allows for parsimonious causal modeling, still maximally adjusting for confounding. This function can be used to understand the causal variable importance of individual variables (by having V be a single variable) and allows for nice plots (see plot_msm).

Overview of features using `estimand = "CATE"` as an example

We will begin with the conditional average treatment effect estimand (CATE) and use it to illustrate the features of causalglm. Afterwards, we will go through all the other available estimands.

We will use the following simulated data throughout this part.

```
n <- 250
W1 <- runif(n, min = -1, max = 1)
W2 <- runif(n, min = -1, max = 1)
A <- rbinom(n, size = 1, prob = plogis((W1 + W2  )/3))
Y <- rnorm(n, mean = A * (1 + W1 + 2*W1^2) + sin(4 * W2) + sin(4 * W1), sd = 0.3)
data <- data.frame(W1, W2,A,Y)
```

spglm with CATE

All methods in causalglm have a similar argument setup. Mainly, they require a formula that specifies a parametric form for the conditional estimand, a data.frame with the data, and character vectors containing the names of the variables W, A and Y. The estimand is specified with the argument *estimand* and the learning method is specified with the *learning_method* argument.

```
formula <- ~ poly(W1, degree = 2, raw = T) # A correctly specified polynomial model of degree 2
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE", # Options are CATE, RR, OR
      learning_method = "HAL" # A bunch of options. Default is a custom semiparametric Highly Adaptive
      )
```

```
## (max) epsilon: -8.456468e-04 max(abs(ED)): 7.472148e-17
```

`output` contains a `spglm` fit object. It contains estimates information and tlverse/tmle3 objects that store the fit likelihood, tmle_tasks, and target parameter objects. There are a number of extractor functions that should suffice for almost everyone. The `summary`, `coefs`, `print` and `predict` functions should be useful. They work as follows.

```
# Print tells you the object, estimand, and a fit formula/equation for the estimand
print(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.01 * (Intercept) + 0.855 * poly(W1, degree = 2, raw = T)1 + 1.98 * poly(W1, degree = 2,
```

Summary provides the coefficient estimates (tmle_est), 95% confidence intervals (lower, upper), and p-values (p_value). The coef function provides pretty much the same thing as summary.

```
summary(output)  # Summary gives you the estimates and inference
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.01 * (Intercept) + 0.855 * poly(W1, degree = 2, raw = T)1 + 1.98 * poly(W1, degree = 2, r
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se     lower    upper
## 1: CATE                    (Intercept) 1.0145391 0.05951480 0.8978922 1.131186
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8547588 0.06484824 0.7276586 0.981859
## 3: CATE poly(W1, degree = 2, raw = T)2 1.9837174 0.12604341 1.7366769 2.230758
##     Z_score p_value
## 1: 269.5342       0
## 2: 208.4085       0
## 3: 248.8454       0
```

The predict function allows you get individual-level treatment effect predictions and 95% prediction (confidence) intervals. Specifically, for each observation, the individual CATE estimate derived from the coefficient estimates is given and a 95% confidence interval + p-values for it.

```
preds <- predict(output, data = data)
preds <- predict(output) # By default, training data is used.
head(preds)
```

```
##   (Intercept) poly(W1, degree = 2, raw = T)1 poly(W1, degree = 2, raw = T)2
## 1           1                      0.7051903                     0.49729335
## 2           1                      0.5061487                     0.25618646
## 3           1                     -0.1425488                     0.02032017
## 4           1                     -0.6954722                     0.48368162
## 5           1                      0.2668601                     0.07121433
## 6           1                     -0.5013076                     0.25130935
##     CATE(W)        se   CI_left CI_right  Z-score p-value
## 1 2.6037962 1.0079686 2.4788471 2.728745 40.84416       0
## 2 1.9553757 0.8603330 1.8487277 2.062024 35.93632       0
## 3 0.9330037 0.9100303 0.8201952 1.045812 16.21054       0
## 4 1.3795658 0.9461666 1.2622777 1.496854 23.05392       0
## 5 1.3839092 0.9035073 1.2719093 1.495909 24.21843       0
## 6 1.0845687 0.7916683 0.9864325 1.182705 21.66127       0
```

It is common to want to obtain multiple fits using multiple formulas. We recommend doing this with npglm since it always provides correct interpretable inference even when these models are wrong. It is computationally expensive to recall spglm for each formula since the machine-learning is redone. Instead, we can reuse the machine-learning fits from previous calls to spglm. Due to the semiparametric nature of spglm, the way this works for spglm differs from npglm and msmglm. For spglm, you can pass a previous spglm fit object through the `data` argument with a new formula. The previous fits will then automatically be reused. The catch for spglm is that the new formula must be a subset of the original formula from the previous fit. Thus, one should first fit the most complex formula that contains all terms of interest and then call spglm with the desired subformulas. Lets see how this works. Fortunately, npglm and msmglm also allow for reusing fits and they even work across estimands and for arbitrary formulas (not just subformulas).

```
# Start with big formula
formula_full <- ~ poly(W1, degree = 3, raw = T)
output_full <- spglm(formula_full,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

```
## (max) epsilon: 5.395157e-03 max(abs(ED)): 9.182238e-16
```

```
summary(output_full)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.513 * poly(W1, degree = 3, raw = T)1 + 1.96 * poly(W1, degree = 3, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est          se     lower     upper
## 1: CATE                    (Intercept) 1.0187494 0.05923852 0.9026441 1.1348548
## 2: CATE poly(W1, degree = 3, raw = T)1 0.5127856 0.16620161 0.1870365 0.8385348
## 3: CATE poly(W1, degree = 3, raw = T)2 1.9559307 0.12113461 1.7185112 2.1933502
## 4: CATE poly(W1, degree = 3, raw = T)3 0.5957030 0.24914746 0.1073830 1.0840231
##      Z_score p_value
## 1: 271.91502       0
## 2:  48.78324       0
## 3: 255.30259       0
## 4:  37.80449       0
```

```
# This will give a warning since the term names for `poly(W1, degree = 2, raw = T)` are not a subset of
# Use argument warn = FALSE to turn this off.
subformula <- ~ poly(W1, degree = 2, raw = T)  # one less degree
output<- spglm(subformula,
      data = output_full, # replace data with output_full
      estimand = "CATE" # No need to specify the variables again.
      )
```

```
## Warning in spglm(subformula, data = output_full, estimand = "CATE"): Terms of
## new formula could not be confirmed as subsets of original formula. Make sure
## this formula is truly a subformula or else the results may be unreliable..
```

```
## (max) epsilon: -5.200079e-04 max(abs(ED)): 3.713002e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.01 * (Intercept) + 0.866 * poly(W1, degree = 2, raw = T)1 + 2.01 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param tmle_est          se     lower     upper
## 1: CATE                    (Intercept) 1.009322 0.05920473 0.8932829 1.1253612
## 2: CATE poly(W1, degree = 2, raw = T)1 0.866345 0.05913767 0.7504373 0.9822527
## 3: CATE poly(W1, degree = 2, raw = T)2 2.006462 0.11980380 1.7716514 2.2412736
##     Z_score p_value
## 1: 269.5525       0
## 2: 231.6310       0
## 3: 264.8076       0
```

```
subformula <- ~ 1 + W1  # one less degree
output<- spglm(subformula,
      data = output_full, # replace data with output_full
      estimand = "CATE", warn = FALSE # No need to specify the variables again.
      )
```

```
## (max) epsilon: 5.854659e-05 max(abs(ED)): 6.136758e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.72 * (Intercept) + 1.11 * W1
##
## Coefficient estimates and inference:
##    type       param tmle_est         se     lower    upper  Z_score p_value
## 1: CATE (Intercept) 1.721800 0.03645905 1.6503418 1.793259 746.7021       0
## 2: CATE          W1 1.106129 0.05772913 0.9929824 1.219276 302.9569       0
```

```
subformula <- ~ 1  # one less degree
output<- spglm(subformula,
      data = output_full, # replace data with output_full
      estimand = "CATE", warn = FALSE # No need to specify the variables again.
      )
```

```
## (max) epsilon: 1.828274e-06 max(abs(ED)): 1.497066e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.85 * (Intercept)
##
## Coefficient estimates and inference:
##    type       param tmle_est         se    lower    upper  Z_score p_value
## 1: CATE (Intercept) 1.850503 0.03625709 1.779441 1.921566 806.9877       0
```

```
# That was fast! Look how different the estimates are when the model is misspecified! (npglm would do b
```

Currently all learning was done with HAL (default and recommended in most cases). There are a number of other options. All methods in this package require machine-learning of $P(A = 1|W)$ (the propensity score) and $E[Y|A, W]$ (the conditinal mean outcome). For spglm, $E[Y|A, W]$ is learned in a semiparametric way. By default, the learning algorithm is provided the design matrix $cbind(W, A \cdot formula(W))$ where W is a matrix with columns being the baseline variable observations and $A \cdot formula(W)$ is a matrix with columns being the treatment interaction observations specified by the formula argument. Specifically, the design matrix is constructed as follows:

```
formula <- ~ 1 + W1
AW <- model.matrix(formula, data)
design_mat_sp_Y <- as.matrix(cbind(data[,c("W1", "W2")],AW))
head(as.data.frame(design_mat_sp_Y))
```

```
##           W1          W2 (Intercept)         W1
## 1  0.7051903  0.01077579           1  0.7051903
## 2  0.5061487 -0.72274928           1  0.5061487
## 3 -0.1425488  0.57808723           1 -0.1425488
## 4 -0.6954722  0.17189821           1 -0.6954722
## 5  0.2668601 -0.87954846           1  0.2668601
## 6 -0.5013076  0.17796000           1 -0.5013076
```

Since the design matrix automatically contains the treatment interaction terms, additive learners like glm, glmnet or gam can in principle perform well (since they will model treatment interactions). Note that the final regression fit based on this design matrix will be projected onto the semiparametric model using glm.fit to ensure all model constraints are satisfied (this is not important and happens behind the scenes).

This learning method corresponds with the default argument specification `append\_design\_matrix = TRUE`. The other option `append\_design\_matrix = FALSE` performs treatment-stratified estimation. Specifically, the machine-learning algorithm is used to learn the placebo conditional mean $E[Y|A = 0, W]$ by performing the regression of Y on W using only the observations with $A = 0$. Next, this initial estimator of $E[Y|A = 0, W]$ is used as an offset in a glm-type regression of Y on $A \cdot formula(W)$. This two-stage approach does not pool data across treatment arms and is thus not preferred.

Now that we got the nitty and gritty details out of the way. Lets use some different algorithms. We see that glm and glmnet perform very badly because of model misspecification. (The true model is quite nonlinear in the noninteraction terms). This motivates using causalglm over conventional methods like glm.

```
formula <- ~ poly(W1, degree = 2, raw = T)
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "glm"
      )
```

```
## (max) epsilon: 5.230422e-02 max(abs(ED)): 1.539047e-16
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 0.975 * (Intercept) + 0.774 * poly(W1, degree = 2, raw = T)1 + 1.73 * poly(W1, degree = 2,
##
## Coefficient estimates and inference:
##    type                          param  tmle_est        se     lower    upper
## 1: CATE                    (Intercept) 0.9752636 0.1833084 0.6159857 1.334542
## 2: CATE poly(W1, degree = 2, raw = T)1 0.7744769 0.2083289 0.3661599 1.182794
## 3: CATE poly(W1, degree = 2, raw = T)2 1.7325103 0.4228286 0.9037814 2.561239
##     Z_score p_value
## 1: 84.12200       0
## 2: 58.77993       0
## 3: 64.78604       0
```

```
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "glmnet"
      )
```

```
## (max) epsilon: -7.181157e-02 max(abs(ED)): 2.690764e-16
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.03 * (Intercept) + 0.778 * poly(W1, degree = 2, raw = T)1 + 1.58 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est        se     lower    upper
## 1: CATE                    (Intercept) 1.0265411 0.1871557 0.6597227 1.393359
## 2: CATE poly(W1, degree = 2, raw = T)1 0.7778788 0.2157166 0.3550819 1.200676
## 3: CATE poly(W1, degree = 2, raw = T)2 1.5827352 0.4354535 0.7292621 2.436208
##     Z_score p_value
## 1: 86.72481       0
## 2: 57.01620       0
## 3: 57.46938       0
```

```
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "gam"
      )
```

```
## (max) epsilon: 2.963510e-04 max(abs(ED)): 8.331530e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.867 * poly(W1, degree = 2, raw = T)1 + 1.98 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se     lower     upper
## 1: CATE                    (Intercept) 1.0221025 0.05900041 0.9064638 1.1377411
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8668868 0.05980748 0.7496663 0.9841073
## 3: CATE poly(W1, degree = 2, raw = T)2 1.9791478 0.12098571 1.7420201 2.2162754
##     Z_score p_value
## 1: 273.9110       0
## 2: 229.1801       0
## 3: 258.6510       0
```

```
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "mars"
      )
```

```
## (max) epsilon: 1.352179e-03 max(abs(ED)): 2.250977e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.01 * (Intercept) + 0.854 * poly(W1, degree = 2, raw = T)1 + 2.01 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se     lower     upper
## 1: CATE                    (Intercept) 1.0059245 0.06454145 0.8794255 1.1324234
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8537176 0.06327683 0.7296973 0.9777379
## 3: CATE poly(W1, degree = 2, raw = T)2 2.0117725 0.12794452 1.7610058 2.2625392
##     Z_score p_value
## 1: 246.4318       0
## 2: 213.3239       0
## 3: 248.6149       0
```

```
output <- spglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "xgboost"
      )
```

```
## (max) epsilon: 2.231328e-02 max(abs(ED)): 1.355860e-16
```

```r
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand CATE with formula:
## CATE(W) = 1.03 * (Intercept) + 0.872 * poly(W1, degree = 2, raw = T)1 + 1.84 * poly(W1, degree = 2,
##
## Coefficient estimates and inference:
##    type                           param  tmle_est          se     lower    upper
## 1: CATE                     (Intercept) 1.0320400 0.08909077 0.8574253 1.206655
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8723124 0.09408987 0.6878996 1.056725
## 3: CATE poly(W1, degree = 2, raw = T)2 1.8449992 0.18025392 1.4917080 2.198290
##     Z_score p_value
## 1: 183.1613       0
## 2: 146.5883       0
## 3: 161.8384       0
```

npglm with CATE

npglm is a model-robust version of spglm that we personally recommend (at least as a robustness check). npglm works similarly to spglm. Fitting and extractor functions are pretty much the same.

```r
formula <- ~ poly(W1, degree = 2, raw = T)
output <- npglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

```
## (max) epsilon: 1.979347e-02 max(abs(ED)): 1.049369e-16
```

```r
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.877 * poly(W1, degree = 2, raw = T)1 + 1.97 * poly(W1, degree = 2,
##
## Coefficient estimates and inference:
##    type                           param  tmle_est          se     lower     upper
## 1: CATE                     (Intercept) 1.0165898 0.05636563 0.9061152 1.1270644
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8766666 0.06122802 0.7566619 0.9966713
## 3: CATE poly(W1, degree = 2, raw = T)2 1.9665575 0.11488849 1.7413802 2.1917348
##     Z_score p_value
## 1: 285.1684       0
## 2: 226.3885       0
## 3: 270.6451       0
```

```r
head(predict(output))
```

```
##   (Intercept) poly(W1, degree = 2, raw = T)1 poly(W1, degree = 2, raw = T)2
## 1           1                      0.7051903                     0.49729335
## 2           1                      0.5061487                     0.25618646
## 3           1                     -0.1425488                     0.02032017
## 4           1                     -0.6954722                     0.48368162
## 5           1                      0.2668601                     0.07121433
## 6           1                     -0.5013076                     0.25130935
##     CATE(W)        se   CI_left CI_right  Z-score p-value
## 1 2.6127626 0.9526119 2.4946756 2.730850 43.36646       0
## 2 1.9641188 0.8167146 1.8628779 2.065360 38.02484       0
## 3 0.9315828 0.8665045 0.8241698 1.038996 16.99889       0
## 4 1.3580803 0.8905022 1.2476925 1.468468 24.11351       0
## 5 1.3905843 0.8540957 1.2847095 1.496459 25.74310       0
## 6 1.0713244 0.7636106 0.9766663 1.165983 22.18294       0
```

npglm can reuse fits across both formulas and estimands with no restrictions. This is because the conditional mean and propensity score are learned fully nonparametrically (the previous semiparametric learning method no longer applies). The nice thing about npglm is that all models are viewed as approximations and thus each model below is interpretable as the best approximation. The intercept model is actually a nonparametric estimate for the marginal ATE! (See writeup.) Additionally, the inference for each model is correct (we don't require correctly specified parametric models!).

```
formula <- ~ 1 # We can start with simplest model. npglm does not care.
output_full <- npglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

```
## (max) epsilon: 3.741621e-04 max(abs(ED)): 4.257705e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.877 * poly(W1, degree = 2, raw = T)1 + 1.97 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se     lower     upper
## 1: CATE                    (Intercept) 1.0165898 0.05636563 0.9061152 1.1270644
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8766666 0.06122802 0.7566619 0.9966713
## 3: CATE poly(W1, degree = 2, raw = T)2 1.9665575 0.11488849 1.7413802 2.1917348
##     Z_score p_value
## 1: 285.1684       0
## 2: 226.3885       0
## 3: 270.6451       0
```

```
formula <- ~  1 + W1
output <- npglm(formula,
      output_full,
      estimand = "CATE"
      )
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: -1.668826e-03 max(abs(ED)): 4.562323e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 1.71 * (Intercept) + 0.943 * W1
##
## Coefficient estimates and inference:
##    type       param  tmle_est         se     lower    upper  Z_score p_value
## 1: CATE (Intercept) 1.7062659 0.05048261 1.6073218 1.805210 534.4104       0
## 2: CATE          W1 0.9431718 0.09357859 0.7597612 1.126582 159.3618       0
```

```r
formula <- ~ poly(W1, degree = 2, raw = T)
output <- npglm(formula,
      output_full,
      estimand = "CATE"
      )
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: 1.555718e-02 max(abs(ED)): 7.659151e-17
```

```r
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.872 * poly(W1, degree = 2, raw = T)1 + 1.94 * poly(W1, degree = 2, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se     lower     upper
## 1: CATE                    (Intercept) 1.0222328 0.05514049 0.9141594 1.1303062
## 2: CATE poly(W1, degree = 2, raw = T)1 0.8722691 0.05693900 0.7606707 0.9838675
## 3: CATE poly(W1, degree = 2, raw = T)2 1.9444051 0.11090501 1.7270353 2.1617749
##     Z_score p_value
## 1: 293.1225       0
## 2: 242.2204       0
## 3: 277.2079       0
```

```r
formula <- ~ poly(W1, degree = 3, raw = T)
output <- npglm(formula,
      output_full,
      estimand = "CATE"
      )
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: 1.576478e-02 max(abs(ED)): 1.412259e-15
```

```r
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 1.02 * (Intercept) + 0.597 * poly(W1, degree = 3, raw = T)1 + 1.94 * poly(W1, degree = 3, 
##
## Coefficient estimates and inference:
##    type                          param  tmle_est         se      lower
## 1: CATE                    (Intercept) 1.0211578 0.05509527 0.91317306
## 2: CATE poly(W1, degree = 3, raw = T)1 0.5969236 0.15672941 0.28973965
## 3: CATE poly(W1, degree = 3, raw = T)2 1.9356982 0.11069377 1.71874235
## 4: CATE poly(W1, degree = 3, raw = T)3 0.4582765 0.22842937 0.01056317
##        upper   Z_score p_value
## 1: 1.1291425 293.05463       0
## 2: 0.9041076  60.21966       0
## 3: 2.1526540 276.49320       0
## 4: 0.9059898  31.72091       0
```

causalglmnet with CATE

causalglmnet is a wrapper for spglm that uses the LASSO with glmnet for all estimation. This is made for high dimensional settings. It is used in the same way as spglm.

```r
formula <- ~ poly(W1, degree = 3, raw = T)
output <- causalglmnet(formula,
       data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE"
      )
```

```
## (max) epsilon: 2.455080e+00 max(abs(ED)): 4.284295e-15
```

```
summary(output)
```

```
## A causalglm fit object obtained from causalglmnet for the estimand CATE with formula:
## CATE(W) = 0.992 * (Intercept) + 1.06 * poly(W1, degree = 3, raw = T)1 + 1.68 * poly(W1, degree = 3, 
##
## Coefficient estimates and inference:
##    type                          param   tmle_est        se      lower
## 1: CATE                    (Intercept)  0.9915036 0.1623367  0.6733294
## 2: CATE poly(W1, degree = 3, raw = T)1  1.0619949 0.4556319  0.1689727
## 3: CATE poly(W1, degree = 3, raw = T)2  1.6759669 0.3640296  0.9624819
## 4: CATE poly(W1, degree = 3, raw = T)3 -0.4487612 0.7091628 -1.8386948
##        upper  Z_score p_value
## 1: 1.3096777 96.57117       0
## 2: 1.9550171 36.85346       0
## 3: 2.3894518 72.79452       0
## 4: 0.9411724 10.00551       0
```

msmglm with CATE

msmglm is for learning marginal structural models (e.g. marginal estimands like the ATE, ATT, and marginal relative risk). It operates in the same way as npglm. It is also a nonparametrically robust method that does not require correct model specification and estimates the best approximation. The only difference is that the marginal covariate(s) of interest V need to be specified. It also has a useful plotting feature that displays 95% confidence bands (only if V is one-dimensional). This method is used if you have many confounders W for which to adjust but only care about the treatment effect association with a subset of variables V. This can be used to build causal predictors that only utilize a handful of variables.

```
formula <-  ~ poly(W1, degree = 3, raw = T)
output <- msmglm(formula,
      data,
      V = "W1",
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

```
## (max) epsilon: -1.277054e-01 max(abs(ED)): 1.377662e-15
```

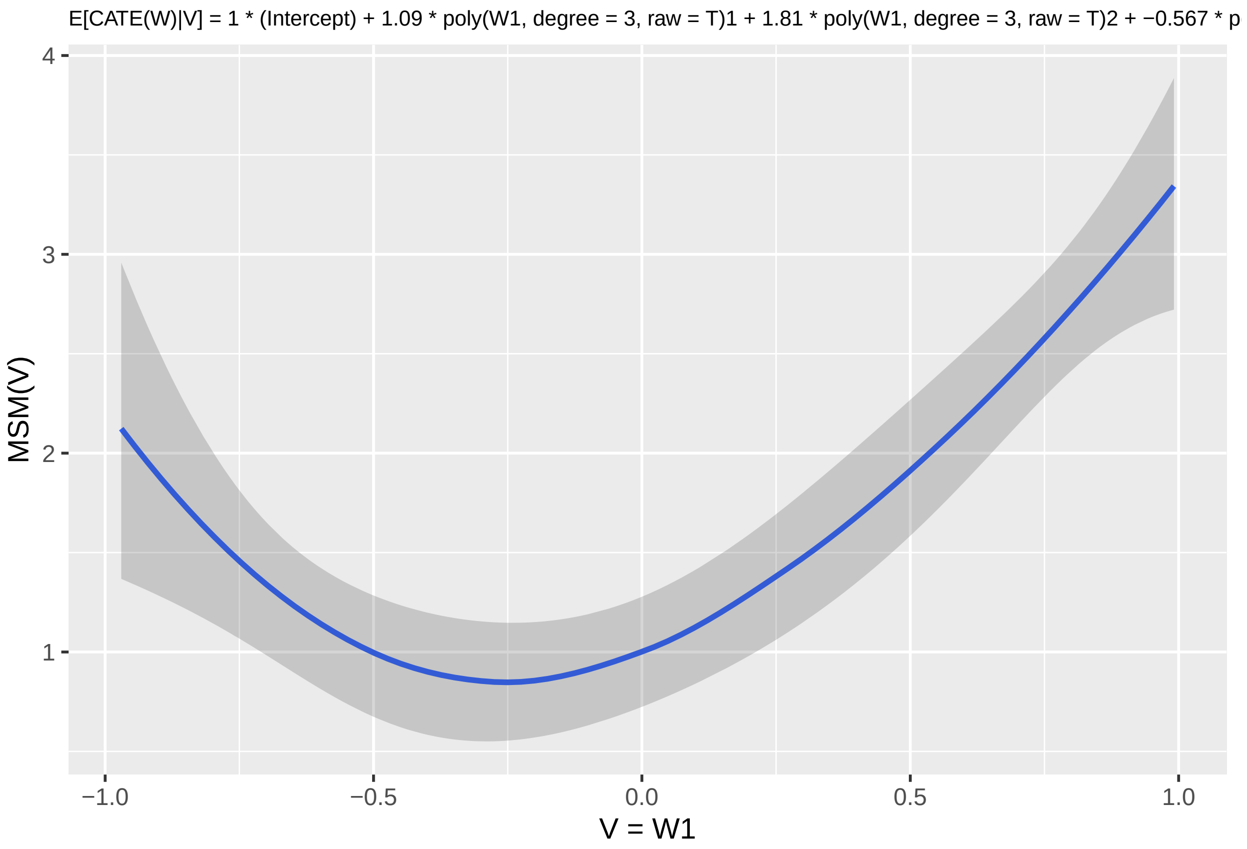
```
summary(output)
```

```
## A causalglm fit object obtained from msmglm for the estimand CATE with formula:
## E[CATE(W)|V] = 1 * (Intercept) + 1.09 * poly(W1, degree = 3, raw = T)1 + 1.81 * poly(W1, degree = 3, 
##
## Coefficient estimates and inference:
##    type                          param   tmle_est        se      lower
## 1: CATE                    (Intercept)  1.0010426 0.1412946  0.7241103
## 2: CATE poly(W1, degree = 3, raw = T)1  1.0888919 0.3987863  0.3072852
## 3: CATE poly(W1, degree = 3, raw = T)2  1.8076034 0.3243272  1.1719338
## 4: CATE poly(W1, degree = 3, raw = T)3 -0.5669925 0.6163691 -1.7750538
##        upper   Z_score p_value
## 1: 1.2779749 112.02038       0
## 2: 1.8704986  43.17323       0
## 3: 2.4432729  88.12311       0
## 4: 0.6410687  14.54476       0
```

```
plot_msm(output)
```

```
## Loading required package: ggplot2

## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
formula <-  ~ 1 + W1 # Best linear approximation
output <- msmglm(formula,
      data,
      V = "W1",
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

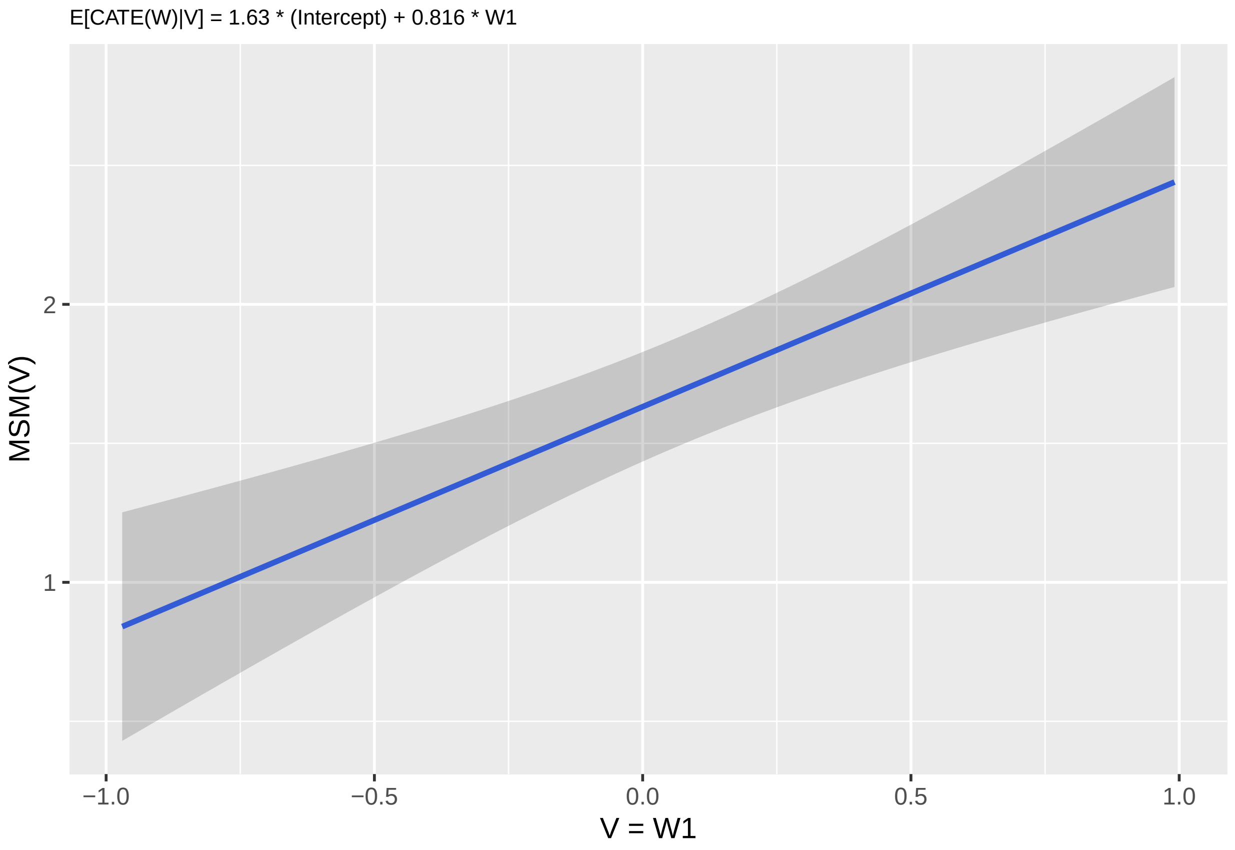
```
## (max) epsilon: -6.038813e-03 max(abs(ED)): 8.905376e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from msmglm for the estimand CATE with formula:
## E[CATE(W)|V] = 1.63 * (Intercept) + 0.816 * W1
##
## Coefficient estimates and inference:
##    type       param  tmle_est        se     lower     upper   Z_score p_value
## 1: CATE (Intercept) 1.6316915 0.1005671 1.4345836 1.828799 256.53827       0
## 2: CATE          W1 0.8155119 0.1780295 0.4665805 1.164443  72.42831       0
```

```
plot_msm(output)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
# This gives a nonparametric estimate for the marginal ATE
formula <-  ~ 1
output <- msmglm(formula,
      data,
      V = "W1",
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      learning_method = "HAL"
      )
```

```
## (max) epsilon: 2.330897e-03 max(abs(ED)): 2.831069e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from msmglm for the estimand CATE with formula:
## E[CATE(W)|V] = 1.64 * (Intercept)
##
## Coefficient estimates and inference:
##    type       param tmle_est        se    lower    upper Z_score p_value
## 1: CATE (Intercept) 1.641342 0.1053016 1.434955 1.847729 246.453       0
```

Learning other estimands.

All of the vignette discussed so far can be applied to other estimands by specifying a different "estimand" argument.

Let us begin with npglm (msmglm acts in the same exact way). Both npglm and msmglm support the CATE, OR, RR, CATT and TSM

```
n <- 250
W1 <- runif(n, min = -1, max = 1)
W2 <- runif(n, min = -1, max = 1)
A <- rbinom(n, size = 1, prob = plogis((W1 + W2  )/3))
Y <- rnorm(n, mean = A * (1 + W1 + 2*W1^2) + sin(4 * W2) + sin(4 * W1), sd = 0.3)
data <- data.frame(W1, W2,A,Y)
# CATE
formula = ~ poly(W1, degree = 2, raw = TRUE)
output <- npglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE")
```

```
## (max) epsilon: 1.635120e-02 max(abs(ED)): 5.171905e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 0.933 * (Intercept) + 1.03 * poly(W1, degree = 2, raw = TRUE)1 + 2.13 * poly(W1, degree =
##
## Coefficient estimates and inference:
##    type                             param  tmle_est         se     lower
## 1: CATE                       (Intercept) 0.9325288 0.05619489 0.8223889
## 2: CATE poly(W1, degree = 2, raw = TRUE)1 1.0336420 0.06241303 0.9113147
## 3: CATE poly(W1, degree = 2, raw = TRUE)2 2.1324552 0.12590256 1.8856908
##       upper  Z_score p_value
## 1: 1.042669 262.3828       0
## 2: 1.155969 261.8574       0
## 3: 2.379220 267.8030       0
```

```
# CATT, lets reuse fit
output <- npglm(formula,
      output,
      estimand = "CATT")
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: 2.975476e-02 max(abs(ED)): 7.272551e-16
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATT with formula:
## CATT(W) = 0.938 * (Intercept) + 1.03 * poly(W1, degree = 2, raw = TRUE)1 + 2.12 * poly(W1, degree =
##
## Coefficient estimates and inference:
##    type                             param  tmle_est         se     lower
## 1: CATT                       (Intercept) 0.9383892 0.05909138 0.8225722
## 2: CATT poly(W1, degree = 2, raw = TRUE)1 1.0293169 0.06380592 0.9042596
## 3: CATT poly(W1, degree = 2, raw = TRUE)2 2.1151107 0.13675927 1.8470675
##       upper  Z_score p_value
## 1: 1.054206 251.0897       0
## 2: 1.154374 255.0693       0
## 3: 2.383154 244.5380       0
```

```
# TSM, note this provides a list of npglm objects for each level of `A`.
outputs <- npglm(formula,
      output,
      estimand = "TSM")
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: 2.242597e-02 max(abs(ED)): 1.118827e-16
```

```
summary(outputs[[1]])
```

```
## A causalglm fit object obtained from npglm for the estimand TSM with formula:
## TSM(W) = 1.11 * E[Y_{A=1}]: (Intercept) + 1.26 * E[Y_{A=1}]: poly(W1, degree = 2, raw = TRUE)1 + 1.9
##
## Coefficient estimates and inference:
##    type                                         param tmle_est         se
## 1:  TSM                       E[Y_{A=1}]: (Intercept) 1.109638 0.09066914
## 2:  TSM E[Y_{A=1}]: poly(W1, degree = 2, raw = TRUE)1 1.264778 0.10388745
## 3:  TSM E[Y_{A=1}]: poly(W1, degree = 2, raw = TRUE)2 1.921307 0.20389092
##        lower    upper  Z_score p_value
## 1: 0.9319297 1.287346 193.5048       0
## 2: 1.0611625 1.468394 192.4958       0
## 3: 1.5216882 2.320926 148.9940       0
```

```
summary(outputs[[2]])
```

```
## A causalglm fit object obtained from npglm for the estimand TSM with formula:
## TSM(W) = 0.177 * E[Y_{A=0}]: (Intercept) + 0.232 * E[Y_{A=0}]: poly(W1, degree = 2, raw = TRUE)1 + -
##
## Coefficient estimates and inference:
##    type                                         param   tmle_est         se
## 1:  TSM                       E[Y_{A=0}]: (Intercept)  0.1773610 0.09221203
## 2:  TSM E[Y_{A=0}]: poly(W1, degree = 2, raw = TRUE)1  0.2320967 0.10655505
## 3:  TSM E[Y_{A=0}]: poly(W1, degree = 2, raw = TRUE)2 -0.2119410 0.21403104
##           lower     upper  Z_score p_value
## 1: -0.003371269 0.3580932 30.41169       0
## 2:  0.023252626 0.4409408 34.44014       0
## 3: -0.631434144 0.2075521 15.65699       0
```

Both the OR and RR estimands provide the original coefficient estimates and their exponential transforms. This is because the parametric model/formula is actually for the log RR and log OR (that is log-linear models). The predict function gives the exponential of the linear predictor (so actually predicts the OR and RR).

```
# odds ratio
n <- 250
W <- runif(n, min = -1,  max = 1)
A <- rbinom(n, size = 1, prob = plogis(W))
Y <- rbinom(n, size =  1, prob = plogis(A + A * W + W + sin(5 * W)))
data <- data.frame(W, A, Y)
output <-
  npglm(
    ~1+W,
    data,
    W = c("W"), A = "A", Y = "Y",
    estimand = "OR"
  )
```

```
## risk_change: -5.871322e-05 (max) epsilon: 2.499999e-02 max(abs(ED)): 3.494115e-02
## risk_change: -1.139163e-05 (max) epsilon: 1.654493e-02 max(abs(ED)): 1.352857e-03
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand OR with formula:
## log OR(W) = 0.845 * (Intercept) + 0.922 * W
##
## Coefficient estimates and inference:
##    type       param  tmle_est        se      lower    upper  psi_exp lower_exp
## 1:   OR (Intercept) 0.8451672 0.2937378  0.2694516 1.420883 2.328367 1.3092463
## 2:   OR           W 0.9220205 0.4996805 -0.0573353 1.901376 2.514365 0.9442774
##    upper_exp  Z_score p_value
## 1:  4.140774 45.49385       0
## 2:  6.695102 29.17549       0
```

```
output <-
  spglm(
    ~1+W,
    data,
    W = c("W"), A = "A", Y = "Y",
    estimand = "OR"
  )
```

```
## risk_change: -1.956900e-05 (max) epsilon: 2.106300e-02 max(abs(ED)): 9.852812e-04
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand OR with formula:
## log OR(W) = 0.89 * (Intercept) + 0.857 * W
##
## Coefficient estimates and inference:
##    type       param  tmle_est        se      lower    upper  psi_exp lower_exp
## 1:   OR (Intercept) 0.8903096 0.2922255  0.3175581 1.463061 2.435884 1.3737690
## 2:   OR           W 0.8569111 0.4923800 -0.1081359 1.821958 2.355872 0.8975056
##    upper_exp  Z_score p_value
## 1:  4.319161 48.17180       0
## 2:  6.183956 27.51727       0
```

```
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand OR with formula:
## log OR(W) = 0.89 * (Intercept) + 0.857 * W
##
## Coefficient estimates and inference:
##    type       param  tmle_est        se      lower    upper  psi_exp lower_exp
## 1:   OR (Intercept) 0.8903096 0.2922255  0.3175581 1.463061 2.435884 1.3737690
## 2:   OR           W 0.8569111 0.4923800 -0.1081359 1.821958 2.355872 0.8975056
##    upper_exp  Z_score p_value
## 1:  4.319161 48.17180       0
## 2:  6.183956 27.51727       0
```

```
# relative risk
n <- 250
```

```
W <- runif(n, min = -1,  max = 1)
A <- rbinom(n, size = 1, prob = plogis(W))
Y <- rpois(n, lambda = exp( A * (1 + W + 2*W^2)  + sin(5 * W)))
data <- data.frame(W, A, Y)
formula = ~ poly(W, degree = 2, raw = TRUE)
output <-
  npglm(
    formula,
    data,
    W = "W", A = "A", Y = "Y",
    estimand = "RR",
    verbose = FALSE
  )
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand RR with formula:
## log RR(W) = 1.03 * (Intercept) + 0.825 * poly(W, degree = 2, raw = TRUE)1 + 1.75 * poly(W, degree = 2
##
## Coefficient estimates and inference:
##    type                            param  tmle_est        se     lower    upper
## 1:   RR                      (Intercept) 1.0338565 0.1922153 0.6571215 1.410591
## 2:   RR poly(W, degree = 2, raw = TRUE)1 0.8246791 0.2538949 0.3270542 1.322304
## 3:   RR poly(W, degree = 2, raw = TRUE)2 1.7541499 0.5113477 0.7519268 2.756373
##     psi_exp lower_exp upper_exp  Z_score p_value
## 1: 2.811889  1.929231  4.098379 85.04375       0
## 2: 2.281149  1.386877  3.752056 51.35716       0
## 3: 5.778533  2.121083 15.742640 54.24009       0
```

```
output <-
  spglm(
    formula,
    data,
    W = "W", A = "A", Y = "Y",
    estimand = "RR",
    verbose = FALSE
  )
summary(output)
```

```
## A causalglm fit object obtained from spglm for the estimand RR with formula:
## log RR(W) = 1.02 * (Intercept) + 0.84 * poly(W, degree = 2, raw = TRUE)1 + 1.81 * poly(W, degree = 2
##
## Coefficient estimates and inference:
##    type                            param  tmle_est        se     lower    upper
## 1:   RR                      (Intercept) 1.0168779 0.1405426 0.7414195 1.292336
## 2:   RR poly(W, degree = 2, raw = TRUE)1 0.8396594 0.1862784 0.4745605 1.204758
## 3:   RR poly(W, degree = 2, raw = TRUE)2 1.8103224 0.3848236 1.0560819 2.564563
##     psi_exp lower_exp upper_exp   Z_score p_value
## 1: 2.764550  2.098913  3.641284 114.40129       0
## 2: 2.315578  1.607308  3.335953  71.27065       0
## 3: 6.112418  2.875084 12.994976  74.38137       0
```

```
output <-
  msmglm(
    formula,
    data,
    V = "W",
    W = "W", A = "A", Y = "Y",
    estimand = "RR",
    verbose = FALSE
  )
summary(output)
```

```
## A causalglm fit object obtained from msmglm for the estimand RR with formula:
## log E[RR(W)|V] = 1.03 * (Intercept) + 0.825 * poly(W, degree = 2, raw = TRUE)1 + 1.75 * poly(W, degr
##
## Coefficient estimates and inference:
##    type                            param  tmle_est        se     lower    upper
## 1:   RR                      (Intercept) 1.0337841 0.1927680 0.6559659 1.411602
## 2:   RR poly(W, degree = 2, raw = TRUE)1 0.8246887 0.2547962 0.3252973 1.324080
## 3:   RR poly(W, degree = 2, raw = TRUE)2 1.7542823 0.5132320 0.7483661 2.760198
##     psi_exp lower_exp upper_exp  Z_score p_value
## 1: 2.811685  1.927003  4.102524 84.79398       0
## 2: 2.281171  1.384442  3.758726 51.17609       0
## 3: 5.779298  2.113544 15.802979 54.04503       0
```

Custom learners with sl3

We refer to the documentation of the tlverse/sl3 package for how learners work. To specify custom learners for the propensity score use the argument sl3_learner_A and to specify custom learners for the outcome conditional mean use the argument sl3_learner_Y. For spglm, keep in mind the argument "append_design_matrix" when choosing learners. A good rule of thumb for spglm is to think of sl3_learner_Y as a learner for $E[Y|A=0,W]$. For msmglm and npglm, the learning is fully nonparametric and the regression is performed how you would expect (a standard design matrix containing W and A is passed to the learner). For msmglm and npglm, make sure the learner models interactions, specifically treatment interactions, as these are crucial for fitting the conditional treatment effect estimands well.

```
library(sl3)
lrnr_A <- Lrnr_gam$new()
lrnr_Y <- Lrnr_xgboost$new(max_depth = 4)
lrnr_Y <- Lrnr_cv$new(lrnr_Y, full_fit = TRUE) #cross-fit xgboost

n <- 250
W1 <- runif(n, min = -1, max = 1)
W2 <- runif(n, min = -1, max = 1)
A <- rbinom(n, size = 1, prob = plogis((W1 + W2  )/3))
Y <- rnorm(n, mean = A * (1 + W1 + 2*W1^2) + sin(4 * W2) + sin(4 * W1), sd = 0.3)
data <- data.frame(W1, W2,A,Y)
# CATE
formula = ~ poly(W1, degree = 2, raw = TRUE)
output <- npglm(formula,
      data,
      W = c("W1", "W2"), A = "A", Y = "Y",
      estimand = "CATE",
      sl3_Learner_A = lrnr_A,
      sl3_Learner_Y = lrnr_Y)
```

```
## (max) epsilon: 8.897193e-02 max(abs(ED)): 3.306799e-16
```

Other arguments

See the documentation for other arguments for all methods. We note that the remaining arguments will likely not be needed for the average user.

Effects of categorical treatments with npglm and msmglm

For `msmglm` and `npglm`, the CATE, CATT, TSM and RR can be learned for categorical treatments relative to a control treatment. To do this, you need to specify the arguments treatment_level and control_level. The estimands are then user-specified parametric models in W for

$$W \mapsto E[Y|A = a, W] - E[Y|A = 0, W]$$

$$W \mapsto E[Y|A = a, W]$$

$$W \mapsto E[Y|A = a, W]/E[Y|A = 0, W]$$

where a is the specified treatment level.

```
n <- 250
V <- runif(n, min = -1, max = 1)
W <- runif(n, min = -1, max = 1)
A <- rbinom(n, size = 1, prob = 0.66*plogis(W))
A[A==1] <- 2
A[A==0] <- rbinom(n, size = 1, prob = plogis(W))
```

```
## Warning in A[A == 0] <- rbinom(n, size = 1, prob = plogis(W)): number of items
## to replace is not a multiple of replacement length
```

```
table(A)
```

```
## A
##  0  1  2
## 84 93 73
```

```
Y <- rnorm(n, mean = A * (1 + W  ) + W , sd = 0.5)
data <- data.table(W,A,Y)

output_init <- npglm(~1+W, data, W = "W", A = "A", Y = "Y", estimand = "CATE", learning_method = "mars"
```

```
## (max) epsilon: 9.326478e-04 max(abs(ED)): 5.108414e-17
```

```
summary(output_init)
```

```
## A causalglm fit object obtained from npglm for the estimand CATE with formula:
## CATE(W) = 0.944 * (Intercept) + 0.975 * W
##
## Coefficient estimates and inference:
##    type       param  tmle_est         se     lower    upper  Z_score p_value
## 1: CATE (Intercept) 0.9440812 0.06428732 0.8180804 1.070082 232.1956       0
## 2: CATE           W 0.9752912 0.10972950 0.7602253 1.190357 140.5338       0
```

```
output <- msmglm(~1+W, data, V = "W", W = "W", A = "A", Y = "Y", estimand = "CATE", learning_method = "r
```

```
## (max) epsilon: 9.326478e-04 max(abs(ED)): 5.108414e-17
```

```
summary(output)
```

```
## A causalglm fit object obtained from msmglm for the estimand CATE with formula:
## E[CATE(W)|V] = 0.944 * (Intercept) + 0.975 * W
```

```
##
## Coefficient estimates and inference:
##    type       param  tmle_est          se     lower    upper  Z_score p_value
## 1: CATE (Intercept) 0.9440812 0.06428732 0.8180804 1.070082 232.1956       0
## 2: CATE           W 0.9752912 0.10972950 0.7602253 1.190357 140.5338       0
```

```
# Reuse fits
output <- npglm(~1+W, output_init , estimand = "CATT",   treatment_level = 2, control_level = 0)
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: 5.227004e-03 max(abs(ED)): 2.471796e-16
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand CATT with formula:
## CATT(W) = 2.04 * (Intercept) + 1.99 * W
##
## Coefficient estimates and inference:
##    type       param tmle_est          se    lower    upper  Z_score p_value
## 1: CATT (Intercept) 2.038966 0.07161189 1.898609 2.179322 450.1889       0
## 2: CATT           W 1.993721 0.11784906 1.762741 2.224700 267.4904       0
```

```
output <- npglm(~1+W, output_init , estimand = "TSM",   treatment_level = c(0,1,2))
```

```
## [1] "Reusing previous fit..."
## (max) epsilon: -2.182467e-03 max(abs(ED)): 1.001976e-16
```

```
lapply(output, summary)
```

```
## A causalglm fit object obtained from npglm for the estimand TSM with formula:
## TSM(W) = -0.00475 * E[Y_{A=0}]: (Intercept) + 1.1 * E[Y_{A=0}]: W
##
## Coefficient estimates and inference:
##    type                   param     tmle_est          se       lower      upper
## 1:  TSM E[Y_{A=0}]: (Intercept) -0.004750394 0.04657472 -0.09603517 0.08653439
## 2:  TSM           E[Y_{A=0}]: W  1.102218007 0.07487634  0.95546308 1.24897294
##       Z_score p_value
## 1:   1.612684 0.10681
## 2: 232.751718 0.00000
## A causalglm fit object obtained from npglm for the estimand TSM with formula:
## TSM(W) = 0.939 * E[Y_{A=1}]: (Intercept) + 2.08 * E[Y_{A=1}]: W
##
## Coefficient estimates and inference:
##    type                   param  tmle_est          se     lower    upper
## 1:  TSM E[Y_{A=1}]: (Intercept) 0.9390751 0.04435351 0.8521438 1.026006
## 2:  TSM           E[Y_{A=1}]: W 2.0776290 0.08015592 1.9205263 2.234732
##     Z_score p_value
## 1: 334.7668       0
## 2: 409.8287       0
## A causalglm fit object obtained from npglm for the estimand TSM with formula:
## TSM(W) = 2.04 * E[Y_{A=2}]: (Intercept) + 3.09 * E[Y_{A=2}]: W
##
## Coefficient estimates and inference:
##    type                   param tmle_est          se    lower    upper  Z_score
## 1:  TSM E[Y_{A=2}]: (Intercept) 2.044341 0.05305615 1.940353 2.148329 609.2388
## 2:  TSM           E[Y_{A=2}]: W 3.086342 0.08583553 2.918108 3.254577 568.5216
##    p_value
## 1:       0
## 2:       0
```

```
## $`E[Y_{A=0}]`
##    type                   param     tmle_est         se       lower      upper
## 1:  TSM E[Y_{A=0}]: (Intercept) -0.004750394 0.04657472 -0.09603517 0.08653439
## 2:  TSM           E[Y_{A=0}]: W  1.102218007 0.07487634  0.95546308 1.24897294
##       Z_score p_value
## 1:   1.612684 0.10681
## 2: 232.751718 0.00000
##
## $`E[Y_{A=1}]`
##    type                   param  tmle_est         se     lower    upper
## 1:  TSM E[Y_{A=1}]: (Intercept) 0.9390751 0.04435351 0.8521438 1.026006
## 2:  TSM           E[Y_{A=1}]: W 2.0776290 0.08015592 1.9205263 2.234732
##     Z_score p_value
## 1: 334.7668       0
## 2: 409.8287       0
##
## $`E[Y_{A=2}]`
##    type                   param tmle_est         se    lower    upper  Z_score
## 1:  TSM E[Y_{A=2}]: (Intercept) 2.044341 0.05305615 1.940353 2.148329 609.2388
## 2:  TSM           E[Y_{A=2}]: W 3.086342 0.08583553 2.918108 3.254577 568.5216
##    p_value
## 1:       0
## 2:       0
##
## $estimand
##    Length     Class      Mode
##         1 character character
##
## $levels_A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.5     1.0     1.0     1.5     2.0
```

```r
n <- 250
V <- runif(n, min = -1, max = 1)
W <- runif(n, min = -1, max = 1)
A <- rbinom(n, size = 1, prob = 0.66*plogis(W))
A[A==1] <- 2
A[A==0] <- rbinom(n, size = 1, prob = plogis(W))
```

```
## Warning in A[A == 0] <- rbinom(n, size = 1, prob = plogis(W)): number of items
## to replace is not a multiple of replacement length
```

```r
table(A)
```

```
## A
##  0  1  2
## 74 88 88
```

```r
Y <- rpois(n, lambda = exp( A * (1 + W)  + sin(5 * W)))
data <- data.table(W,A,Y)

output_init <- npglm(~1+W, data, W = "W", A = "A", Y = "Y", estimand = "RR", learning_method = "gam", t
```

```
## risk_change: -2.042496e-02 (max) epsilon: 2.499999e-02 max(abs(ED)): 3.446271e-01
## risk_change: -3.301082e-03 (max) epsilon: 1.226953e-02 max(abs(ED)): 1.287985e-02
## risk_change: -1.241172e-05 (max) epsilon: 3.873672e-04 max(abs(ED)): 3.182013e-03
```

```
summary(output_init)
```

```
## A causalglm fit object obtained from npglm for the estimand RR with formula:
## log RR(W) = 1.04 * (Intercept) + 0.755 * W
##
## Coefficient estimates and inference:
##    type       param  tmle_est        se     lower    upper  psi_exp lower_exp
## 1:   RR (Intercept) 1.0370146 0.1619504 0.7195976 1.354432 2.820783  2.053607
## 2:   RR           W 0.7554482 0.3253503 0.1177734 1.393123 2.128565  1.124989
##    upper_exp   Z_score p_value
## 1:  3.874558 101.24482       0
## 2:  4.027408  36.71331       0
```

```
output <- npglm(~1+W, output_init , estimand = "RR",   treatment_level = 2, control_level = 0)
```

```
## [1] "Reusing previous fit..."
## risk_change: -8.517069e-02 (max) epsilon: 2.164899e-02 max(abs(ED)): 2.073420e-01
## risk_change: -5.398276e-04 (max) epsilon: 1.481582e-03 max(abs(ED)): 4.226499e-02
## risk_change: -1.335957e-04 (max) epsilon: 4.493303e-04 max(abs(ED)): 6.560898e-02
## risk_change: -4.859419e-05 (max) epsilon: 4.812885e-04 max(abs(ED)): 1.415951e-02
## risk_change: -1.561265e-05 (max) epsilon: 1.322815e-04 max(abs(ED)): 2.289777e-02
## risk_change: -5.871227e-06 (max) epsilon: 1.733479e-04 max(abs(ED)): 5.041533e-03
## risk_change: -2.030976e-06 (max) epsilon: 4.482966e-05 max(abs(ED)): 8.246303e-03
## risk_change: -7.635452e-07 (max) epsilon: 6.346268e-05 max(abs(ED)): 1.829839e-03
```

```
summary(output)
```

```
## A causalglm fit object obtained from npglm for the estimand RR with formula:
## log RR(W) = 2.06 * (Intercept) + 1.56 * W
##
## Coefficient estimates and inference:
##    type       param tmle_est        se     lower    upper  psi_exp lower_exp
## 1:   RR (Intercept) 2.057365 0.1545036 1.7545432 2.360186 7.825321  5.780806
## 2:   RR           W 1.561382 0.4256357 0.7271515 2.395613 4.765403  2.069178
##    upper_exp   Z_score p_value
## 1:  10.59292 210.54388       0
## 2:  10.97492  58.00176       0
```

Effects of a continuous treatment with contglm

The function `contglm` supports treatment effects for continuous treatments. Currently, the CATE, OR and RR estimands are supported. Specifically, `contglm` computes estimates and nonparametric inference for the best approximation of the true CATE $E[Y|A=a,W] - E[Y|A=0,W]$ (for instance) with respect to the parametric working model $E[Y|A=a,W] - E[Y|A=0,W] = 1(a>0) \cdot \beta^T \underline{f}(W) + a \cdot \beta^T \underline{g}(W)$ where $\underline{f}(W)$ and $\underline{g}(W)$ are user-specified parametric models. $\underline{f}(W)$ is specified with the argument `formula\_binary` and captures the treatment effect caused by being treated or not treated $(1(A>0))$. $\underline{g}(W)$ is specified with the argument `formula_continuous` and captures the treatment effect caused by dosage of continuous effects in the treatment A. Note A should be a nonnegative treatment value with $A=0$ being the placebo group and $A>0$ being a continuous or ordered numeric dose value.

Thus, unlike other functions, both the argument `formula\_continuous` and `formula\_binary` need to be specified.

For the OR and RR, the models are

$$\log OR(a,W) := \log P(Y=1|A=a,W)/P(Y=0|A=a,W) - \log P(Y=1|A=0,W)/P(Y=0|A=0,W)$$
$$= 1(a>0) * formula\_binary(W) + a * formula\_continuous(W)$$

and

$$\log RR(a,W) := logE[Y|A=a,W] - \log E[Y|A=0,W]$$
$$= 1(a>0) * formula\_binary(W) + a * formula\_continuous(W)$$

```r
# CATE
n <- 500
W <- runif(n, min = -1, max = 1)
Abinary <- rbinom(n , size = 1, plogis(W))
A <- rgamma(n, shape = 1, rate = exp(W))
A <- A * Abinary
Y <- rnorm(n, mean =   (A > 0) + A * (1 + W) + W , sd = 0.5)
data <- data.table(W, A, Y)

# Model is CATE(A,W) = formula_binary(W) 1(A > 0) + A * formula_continuous(W)

out <- contglm(
  formula_continuous = ~ 1 + W,
  formula_binary = ~ 1,
  data = data,
  W = "W", A = "A", Y = "Y",
  estimand = "CATE"
)
```

```
## risk_change: -7.832334e-05 (max) epsilon: 6.377232e-03 max(abs(ED)): 2.629171e-03
```

```r
summary(out)
```

```
## A causalglm fit object obtained from contglm for the estimand CATE with formula:
## contCATE(A,W) = 0.942 * 1(A>0)*(Intercept) + 1.11 * A*(Intercept) + 1.14 * A*W
##
## Coefficient estimates and inference:
##        type                param tmle_est         se     lower    upper  Z_score
## 1: contCATE 1(A>0)*(Intercept) 0.941754 0.05995716 0.8242401 1.059268 351.2218
## 2: contCATE      A*(Intercept) 1.113929 0.03424815 1.0468038 1.181054 727.2862
## 3: contCATE                A*W 1.142921 0.04996823 1.0449846 1.240857 511.4546
##    p_value
## 1:       0
## 2:       0
## 3:       0
```

```r
# The CATE predictions are now a function of `A`
### head(predict(out))
```

```r
# OR
# Model is log OR(a,W) =
# log P(Y=1|A=a,W)/P(Y=0|A=a,W) - log P(Y=1|A=0,W)/P(Y=0|A=0,W)
# ~ 1(a>0) * formula_binary(W) + a * formula_continuous(W)
n <- 5000
W <- runif(n, min = -1, max = 1)
```

```r
Abinary <- rbinom(n ,size = 1, plogis(W))
A <- pmin(rgamma(n, shape = 1, rate = exp(W)),1)
A <- A * Abinary
quantile(A)
```

```
##         0%        25%        50%        75%       100%
## 0.00000000 0.00000000 0.02410456 0.57093876 1.00000000
```

```r
Y <- rbinom(n, size = 1,  plogis((A>0) + A * (1 + W  ) + W))
data <- data.table(W,A,Y)
out <- contglm(formula_continuous = ~1+W, formula_binary = ~1, estimand = "OR", data =data, W = "W", A =
```

```
## risk_change: -1.343434e-04 (max) epsilon: 8.283047e-03 max(abs(ED)): 1.114440e-01
## risk_change: -2.193672e-05 (max) epsilon: 3.777421e-03 max(abs(ED)): 5.454534e-02
## risk_change: -6.440931e-06 (max) epsilon: 2.067093e-03 max(abs(ED)): 2.319922e-02
## risk_change: -3.037521e-06 (max) epsilon: 1.183851e-03 max(abs(ED)): 7.150142e-03
```

```r
summary(out)
```

```
## A causalglm fit object obtained from contglm for the estimand OR with formula:
## log contCATE(W) = 0.843 * 1(A>0)*(Intercept) + 1.33 * A*(Intercept) + 1.78 * A*W
##
## Coefficient estimates and inference:
##        type              param tmle_est        se     lower    upper  psi_exp
## 1: contCATE 1(A>0)*(Intercept) 0.842786 0.1253133 0.5971764 1.088396 0.842786
## 2: contCATE      A*(Intercept) 1.328392 0.1984484 0.9394399 1.717343 1.328392
## 3: contCATE                A*W 1.781928 0.2234092 1.3440535 2.219802 1.781928
##    lower_exp upper_exp  Z_score p_value
## 1: 0.5971764  1.088396 475.5597       0
## 2: 0.9394399  1.717343 473.3295       0
## 3: 1.3440535  2.219802 563.9933       0
```

```r
# The OR predictions are now a function of `A`
#head(predict(out))
```

```r
# RR
# Model is log RR(a,W) =
# log E[Y|A=a,W]   - log E[Y|A=0,W]
# ~ 1(a>0) * formula_binary(W) + a * formula_continuous(W)
n <- 1000
W <- runif(n, min = -1, max = 1)
Abinary <- rbinom(n ,size = 1, plogis(W))
A <- pmin(rgamma(n, shape = 1, rate = exp(W)), 1)
A <- A * Abinary
quantile(A)
```

```
##          0%         25%         50%         75%        100%
## 0.000000000 0.000000000 0.008349402 0.612888205 1.000000000
```

```r
Y <- rpois(n,  exp((A>0) + A * (1 + W  ) + W))
table(Y)
```

```
## Y
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
## 236 201 113  78  58  58  42  20  18  11  20  22  16   7   7  11   5   9   9   5
##  20  21  22  23  24  25  26  27  29  30  31  32  33  34  35  36  37  38  44  45
##   4   2   5   6   3   2   5   4   1   2   2   4   2   1   1   1   2   1   1   2
##  49  52  54
##   1   1   1
```

```
data <- data.table(W,A,Y)
out <- contglm(formula_continuous = ~1+W, formula_binary = ~1, data =data, W = "W", A = "A", Y = "Y",
               estimand = "RR")
```

```
## risk_change: -3.438387e-04 (max) epsilon: 4.182924e-02 max(abs(ED)): 3.713731e-02
## risk_change: -1.377427e-04 (max) epsilon: 4.999999e-02 max(abs(ED)): 2.698683e-02
## risk_change: -5.919196e-05 (max) epsilon: 2.848788e-02 max(abs(ED)): 2.054077e-02
## risk_change: -3.118575e-05 (max) epsilon: 1.687983e-02 max(abs(ED)): 1.615589e-02
## risk_change: -2.029738e-05 (max) epsilon: 1.949830e-02 max(abs(ED)): 1.325095e-02
```

```
summary(out)
```

```
## A causalglm fit object obtained from contglm for the estimand RR with formula:
## log contCATE(W) = 0.929 * 1(A>0)*(Intercept) + 1.07 * A*(Intercept) + 0.963 * A*W
##
## Coefficient estimates and inference:
##        type              param  tmle_est         se     lower    upper
## 1: contCATE 1(A>0)*(Intercept) 0.9292312 0.06422653 0.8033495 1.055113
## 2: contCATE      A*(Intercept) 1.0737265 0.06550409 0.9453409 1.202112
## 3: contCATE                A*W 0.9633438 0.11223102 0.7433750 1.183313
##      psi_exp lower_exp upper_exp  Z_score p_value
## 1: 0.9292312 0.8033495  1.055113 457.5192       0
## 2: 1.0737265 0.9453409  1.202112 518.3526       0
## 3: 0.9633438 0.7433750  1.183313 271.4366       0
```

```
# The RR predictions are now a function of `A`
#head(predict(out))
```