# Estimating Average Dose Response Functions Using the R Package `causaldrf`

Douglas Galagate, Joseph L. Schafer

November 30, 2015

**Abstract**

This chapter describes the R package `causaldrf` for estimating average dose response functions (ADRF). The R package contains functions to estimate ADRFs using parametric and non-parametric models when the data contains a continuous treatment variable. The `causaldrf` R package is flexible and can be used on data sets containing treatment variables from a range of probability distributions.

*Keywords:* Causal Inference; Propensity Score; Generalized Propensity Score; Propensity Function; Average Dose Response Function.

# 1 Introduction

In this chapter, we provide examples to illustrate the flexibility and the ease of use of the `causaldrf` R package, which estimates the average dose response function (ADRF) when the treatment is continuous. The `causaldrf` R package also provides methods for estimating average potential outcomes when the treatment is binary or multi-valued. The user can compare different methods to understand the sensitivity of the estimates and a way to check robustness. The package contains new estimators based on a linear combination of a finite number of basis functions Schafer and Galagate (2015). In addition, `causaldrf` includes functions useful for model diagnostics such as assessing common support and for checking covariate balance. This package fills a gap in the R package space and offers a range of existing and new estimators described in the statistics literature such as Schafer and Galagate (2015), Bia et al. (2014), Flores et al. (2012), Imai and Van Dyk (2004), Hirano and Imbens (2004), and Robins et al. (2000).

The `causaldrf` R package is currently available on the Comprehensive R Archive Network (CRAN). The R package contains 12 functions for estimating the ADRF which are explained in more detail in Chapters 2, 3, and in the documentation files for the package `https://cran.r-project.org/web/packages/causaldrf/index.html`. The user can choose which estimator to apply based on their particular problems and goals.

This chapter is organized as follows. In Section 2, we introduce a simulated dataset from Hirano and Imbens (2004) and Moodie and Stephens (2012) and apply functions from `causaldrf` to estimate the ADRF. In Section 3, we use data from the National Medical Expenditures Survey (NMES) to show the capabilities of `causaldrf` in analyzing a data set containing weights. Section 4 contains data from the Infant Health and Development Program (IHDP) and applies methods from `causaldrf` to the data. Conclusions are presented in Section 5.

# 2 An Example Based on Simulated Data

This section demonstrates the use of the `causaldrf` package by using simulated data from Hirano and Imbens (2004) and Moodie and Stephens (2012). This simulation constructs an ADRF with an easy to interpret functional form, and a means to clearly compare the performance of different estimation methods.

Let $Y_1(t)|X_1, X_2 \sim \mathcal{N}\left(t + (X_1 + X_2)e^{-t(X_1+X_2)}, 1\right)$ and $X_1, X_2$ be unit exponentials, $T_1 \sim \exp(X_1 + X_2)$. The ADRF can be calculated by integrating out the covariates analytically (Moodie and Stephens, 2012),

$$\mu(t) = E(Y_i(t)) = t + \frac{2}{(1+t)^3} \tag{1}$$

This example provides a setting to compare ADRF estimates with the true ADRF given in Equation 1. In this simulation, our goal is to demonstrate how to use the functions. We introduce a few of the estimators and show their plots.

First, install **causaldrf** and then load the package:

```r
library (causaldrf)
```

The data is generated from:

```r
set.seed(301)
hi_sample <- function(N){
  X1 <- rexp(N)
  X2 <- rexp(N)
  T <- rexp(N, X1 + X2)
  gps <- (X1 + X2) * exp(-(X1 + X2) * T)
  Y <- T + gps + rnorm(N)
  hi_data <- data.frame(cbind(X1, X2, T, gps, Y))
  return(hi_data)
}


hi_sim_data <- hi_sample(1000)
head(hi_sim_data)
```

```
##          X1         X2           T        gps          Y
## 1 0.1942127 0.18045487 4.718463128 0.06395528  4.1426651
## 2 1.4441432 0.60652576 0.168123100 1.45266708  0.9888306
## 3 5.6393370 0.17758343 0.005784747 5.62444109  5.2284042
## 4 0.5079408 0.45976378 0.350261484 0.68950725 -0.3301777
## 5 0.2282938 0.71565806 0.431730712 0.62800127  1.8360819
## 6 1.1539278 0.09854209 0.786804283 0.46751158  1.4745739
```

Below is code for a few different estimators of the ADRF. The first is the additive spline estimator from Bia et al. (2014). This estimator fits a treatment model to estimate the GPS. Next, additive spline bases values are created for both the treatment and the GPS. The outcome is regressed on the treatment, GPS, treatment bases, and GPS bases. After the outcome model is estimated, each treatment grid value and set of covariates is plugged in to the model which corresponds to imputed values for each unit at that particular treatment value. The imputed values are averaged to get the estimated ADRF at that treatment value. Repeating this process for many treatment values, `grid_val`, traces out the estimated ADRF.

The arguments are: `Y` for the name of the outcome variable, `treat` for the name of the treatment variable, `treat_formula` for the formula used to fit the treatment model, `data` for the name of the data set, `grid_val` for a vector in the domain of the treatment for where the outcome is estimated, `knot_num` for the number of knots for the spline fit, and `treat_mod` for the treatment model that relates treatment with the covariates.

In this example we fit the correct treatment model so that the GPS is correctly specified with a gamma distribution.

```
add_spl_estimate <- add_spl_est(Y = Y,
                                treat = T,
                                treat_formula = T ~ X1 + X2,
                                data = hi_sim_data,
                                grid_val = quantile(hi_sim_data$T,
                                                    probs = seq(0, .95, by = 0.01)),
                                knot_num = 3,
                                treat_mod = "Gamma",
                                link_function = "inverse")
```

The next estimator is based on the generalized additive model. This method requires a treatment formula and model to estimate the GPS. The estimated GPS values are used to fit an outcome regression. The outcome, `Y`, is regressed on two things: the treatment, `T`, and spline basis terms from the GPS fit.

```
gam_estimate <- gam_est(Y = Y,
                        treat = T,
                        treat_formula = T ~ X1 + X2,
                        data = hi_sim_data,
                        grid_val = quantile(hi_sim_data$T,
                                    probs = seq(0, .95, by = 0.01)),
                        treat_mod = "Gamma",
                        link_function = "inverse")
```

The Hirano-Imbens estimator also requires two models. The first model regresses the treatment, `T`, on a set of covariates to estimate the GPS values. The second step requires fitting the outcome, `Y`, on the observed treatment and fitted GPS values. The summary above shows the fit of both the treatment model and outcome model. Also shown is the estimated outcome values on the grid of treatment values, `quantile_grid`.

```
hi_estimate <- hi_est(Y = Y,
                      treat = T,
                      treat_formula = T ~ X1 + X2,
                      outcome_formula = Y ~ T + I(T^2) +
                        gps + I(gps^2) + T * gps,
                      data = hi_sim_data,
                      grid_val = quantile(hi_sim_data$T,
                                  probs = seq(0, .95, by = 0.01)),
                      treat_mod = "Gamma",
                      link_function = "inverse")
```

This last method, importance sampling, fits the treatment as a function of the covariates, then calculates GPS values. The GPS values are used as inverse probability weights in the regression of `Y` on `T` (Robins et al., 2000). The estimated parameters correspond to coefficients for a quadratic model of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$. In this example, the estimator is restricted to a quadratic fit.

```
iptw_estimate <- iptw_est(Y = Y,
                          treat = T,
                          treat_formula = T ~ X1 + X2,
                          numerator_formula = T ~ 1,
                          data = hi_sim_data,
                          degree = 2,
                          treat_mod = "Gamma",
                          link_function = "inverse")
```

The true ADRF and 4 estimates are plotted in Figure 1.

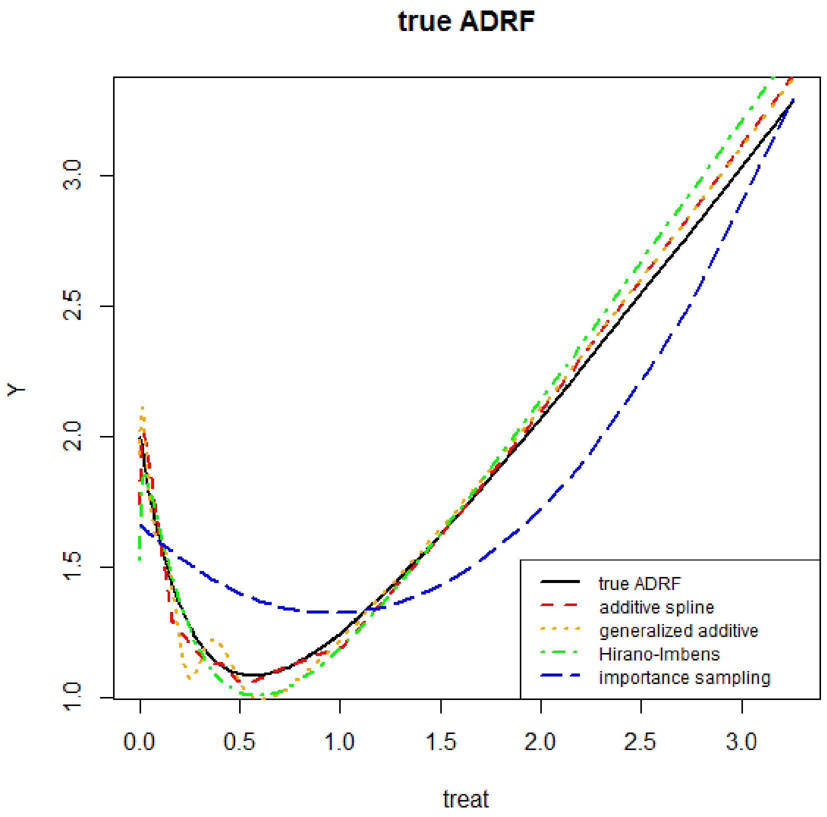

Figure 1: True ADRF along with estimated curves.

# 3 Analysis of the National Medical Expenditures Survey

## 3.1 Introduction

The 1987 National Medical Expenditures Survey (NMES) includes information about smoking amount, in terms of the quantity packyears, and medical expenditures in a representative sample of the U.S. civilian, non-institutionalized population (U.S. Department of Health and Human Services, Public Health service, 1987). The 1987 medical costs were verified by multiple interviews and other data from clinicians and hospitals.

Johnson et al. (2003) analyzed the NMES to estimate the fraction of disease cases and the fraction of the total medical expenditures attributable to smoking for two disease groups. Imai and Van Dyk (2004) emulate the setting by Johnson et al. (2003) but estimated the effect of smoking amount on medical expenditures. Johnson et al. (2003) and Imai and Van Dyk (2004) conducted a complete case analysis by removing units containing missing values. Both Johnson et al. (2003) used multiple imputation techniques to deal with the missing values, but did not find significant differences between that analysis and the complete case analysis. Complete case analysis with propensity scores will lead to biased causal inference unless the data are missing completely at random (D'Agostino Jr and Rubin, 2000). Regardless of this drawback, the analysis in this section uses the complete case data to illustrate the different statistical methods available for estimating the ADRF relating smoking amount and medical expenditures.

This example is analyzed in this section because the treatment variable, smoking amount, is a continuous variable. The data is restricted to that used in Imai and Van Dyk (2004) with 9708 observations and 12 variables. For each person interviewed, the survey collected information on age at the time of the survey, age when the person started smoking, gender, race (white, black, other), marital status (married, widowed, divorced, separated, never married), education level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, or West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always) (Imai and Van Dyk, 2004). The data is available in the `causaldrf` package.

Our goal is to understand how the amount of smoking affects the amount of medical expenditures. Johnson et al. (2003) use a measure of cumulative exposure to smoking that combines self-reported information about frequency and duration of smoking into a variable called *packyear*

$$\textit{packyear} = \frac{\text{number of cigarettes per day}}{20} \times (\text{number of years smoked}) \tag{2}$$

*packyear* can also be defined as the number of packs smoked per day multiplied by the number of years the person was a smoker. The total number of cigarettes per pack is normally 20.

Determining the effect of smoking on health has a long history. Scientists cannot ethically assign smoking amounts randomly to people because of the potential negative effects, so observational data analysis is needed to understand the relationship. The rest of this section will focus on the relationship between smoking amount and medical expenditures.

The NMES oversampled subgroups of the population in order to reduce variances of the estimates. Oversampling reduces the variances of the estimates by increasing the sample size of the target sub-population disproportionately (Singh et al., 1994). The U.S. Department of Health and Human Services oversampled Blacks, Hispanics, the poor and near poor, and the elderly and persons with functional limitations (Cohen, 2000).

## 3.2 Data

Load **nmes_data** into the workspace with

```
data("nmes_data")
dim (nmes_data)

## [1] 9708   12

summary(nmes_data)

##    packyears         AGESMOKE        LASTAGE           MALE       
##  Min.   :  0.05   Min.   : 9.00   Min.   :19.0   Min.   :0.0000  
##  1st Qu.:  6.60   1st Qu.:16.00   1st Qu.:32.0   1st Qu.:0.0000  
##  Median : 17.25   Median :18.00   Median :45.0   Median :1.0000  
##  Mean   : 24.48   Mean   :18.39   Mean   :47.1   Mean   :0.5159  
##  3rd Qu.: 34.50   3rd Qu.:20.00   3rd Qu.:62.0   3rd Qu.:1.0000  
##  Max.   :216.00   Max.   :70.00   Max.   :94.0   Max.   :1.0000  
##  RACE3    beltuse  educate  marital  SREGION  POVSTALB
##  1: 633   1:2613   1:2047   1:6188   1:2047   1:1034  
##  2:1496   2:2175   2:2451   2: 771   2:2451   2: 470  
##  3:7579   3:4920   3:3386   3:1076   3:3386   3:1443  
##                    4:1824   4: 333   4:1824   4:3273  
##                             5:1340            5:3488  
##                                                       
##     HSQACCWT         TOTALEXP       
##  Min.   :  908   Min.   :     0.0  
##  1st Qu.: 4975   1st Qu.:    90.0  
##  Median : 7075   Median :   406.1  
##  Mean   : 8072   Mean   :  2042.0  
##  3rd Qu.:10980   3rd Qu.:  1350.3  
##  Max.   :35172   Max.   :175096.0  
```

The dataset **nmes_data** is a data frame with 9708 rows and 12 variables with summaries of the variables given above. Six of the variables are numeric and the other six are categorical.

The outcome variable is the total amount of medical expenditures, `TOTALEXP`, the treatment is the amount of smoking, `packyears`. The data set contains weights, `HSQACCWT`, that can be used to upweight estimates to the population of interest which is the set of people who smoke and above the age of 18. This analysis demonstrates the capability of `causaldrf` by estimating the ADRF with and without weights. In Figure 3, we plot the estimated ADRFs, their 95% confidence bands, and the 95% confidence bands without weigths.

### 3.3 Common support

The data set is restricted to observations that overlap and have a common support. Units outside of the common support are removed. See Figure 2. The preliminary steps of analysis are omitted such as cleaning and making sure the data overlap.

From Bia et al. (2014), we use the formula

$$CS = \cap_{q=1}^{K}\{i : \hat{R}_i^q \in [max\{min_{j:Q_j=q}\hat{R}_j^q, min_{j:Q_j\neq q}\hat{R}_j^q\}, min\{max_{j:Q_j=q}\hat{R}_j^q, max_{j:Q_j\neq q}\hat{R}_j^q\}]\}$$

to get the common support. For 3 subclasses, the sample is reduced to 8732 units in the common support.

### 3.4 Covariate balance

One of the main goals of fitting a treatment model is to balance the covariates. The GPS or the PF provide a way to balance the covariates. Comparisons of the balance of the covariates before and after adjusting for the GPS or the PF are shown in the following results:

```
t(p_val_bal_cond)

##             Estimate  Std. Error    t value  Pr(>|t|)
## AGESMOKE 0.002649140 0.052968507 0.05001349 0.9601128
## LASTAGE  0.156568519 0.119806490 1.30684505 0.1912998
## MALE     0.006755654 0.005139165 1.31454310 0.1886980

t(p_val_bal_no_cond)

##             Estimate  Std. Error   t value      Pr(>|t|)
## AGESMOKE -0.64598664 0.047692883 -13.54472  2.225653e-41
## LASTAGE   5.11758526 0.140218443  36.49723 1.483453e-271
## MALE      0.05582729 0.004566978  12.22412  4.385267e-34
```

The last column displays the p-value of regressing each of the continuous covariates on the outcome variable, packyears, before and after conditioning on the PF. The first three rows show the p-values after conditioning on the PF, while the last three rows show the p-values when there is no conditioning.

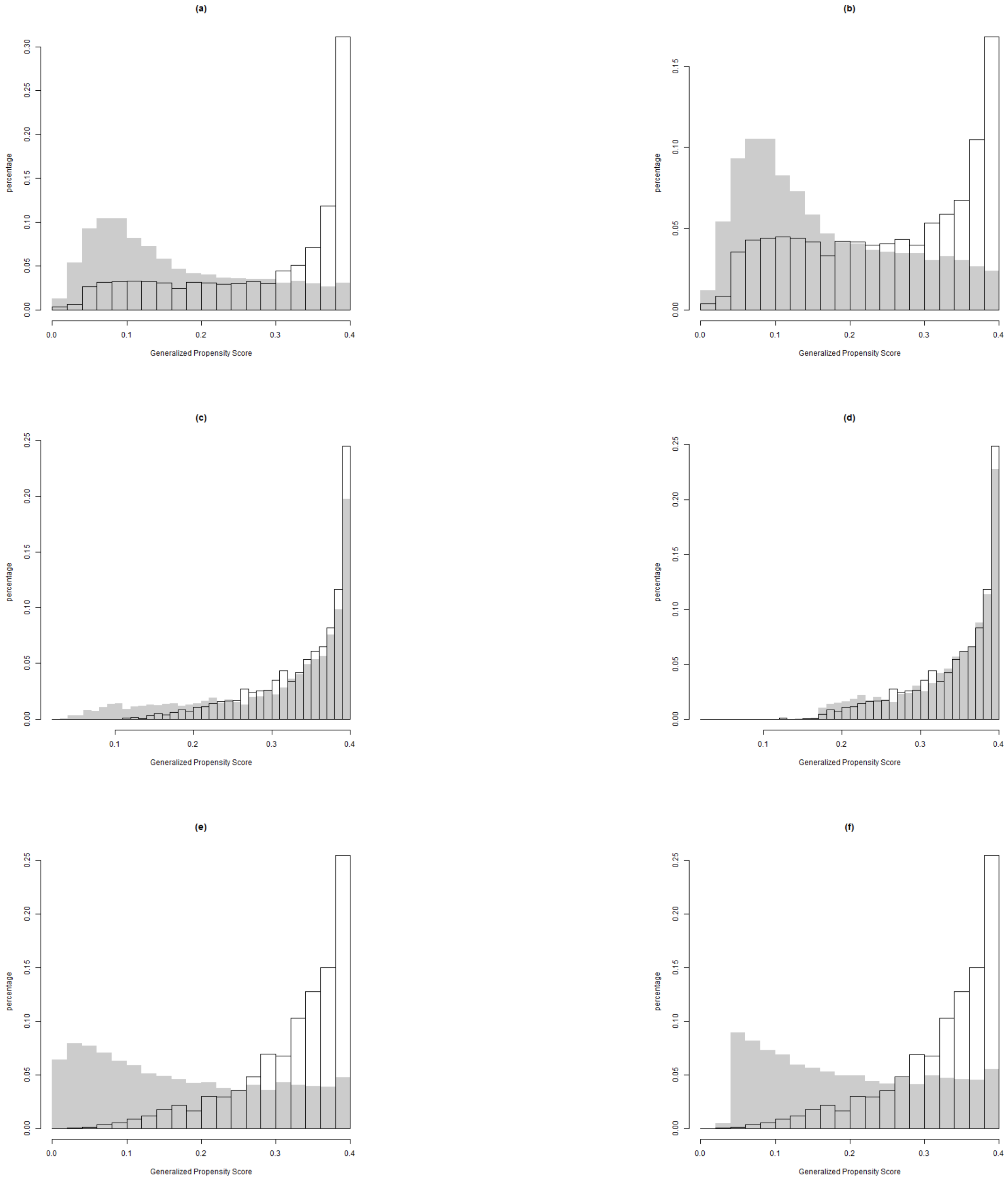

Figure 2: Common support restriction. Shaded bars represent units not in tercile, while white bars represent units in the tercile. (a) Compares group 1 vs others before deleting non-overlapping units. (b) Compares group 1 vs others after deleting non-overlapping units. (c) Compares group 2 vs others before deleting non-overlapping units. (d) Compares group 2 vs others after deleting non-overlapping units. (e) Compares group 3 vs others before deleting non-overlapping units. (f) Compares group 3 vs others after deleting non-overlapping units.

### 3.5 Estimating the ADRF

The `causaldrf` R package contains a variety of estimators. Below is code for 4 other estimators that can account for weights. Although the true ADRF is not a polynomial, we will illustrate methods that are restricted to polynomial form of up to degree 2.

The prima facie estimator is a basic estimator that regresses the outcome `Y` on the treatment `T` without taking covariates into account. The prima facie estimator is unbiased if the data comes from a simple random sample; otherwise it will likely be biased. The model fit is $Y \sim \alpha_0 + \alpha_1 t + \alpha_2 t^2$.

```
pf_estimate <- reg_est(Y = TOTALEXP,
                       treat = packyears,
                       covar_formula = ~ 1,
                       data = full_data_orig,
                       degree = 2,
                       wt = full_data_orig$HSQACCWT,
                       method = "same")
pf_estimate

##
## Estimated values:
## [1] 1128.5947250   36.8409486   -0.1348346
```

The regression prediction method generalizes the prima facie estimator and takes the covariates into account (Schafer and Galagate, 2015).

```
reg_estimate <- reg_est(Y = TOTALEXP,
                        treat = packyears,
                        covar_formula = ~ LASTAGE + LASTAGE2 +
                          AGESMOKE + AGESMOKE2 + MALE + beltuse +
                          educate + marital + POVSTALB + RACE3,
                        covar_lin_formula = ~ 1,
                        covar_sq_formula = ~ 1,
                        data = full_data_orig,
                        degree = 2,
                        wt = full_data_orig$HSQACCWT,
                        method = "different")
reg_estimate

##
## Estimated values:
## [1] 1619.329529   23.260395   -0.109507
```

The propensity spline prediction method adds spline basis terms to the regression prediction method. This method is similar to that of Little and An (2004) and Schafer and Kang (2008), but for the continuous treatment setting (Schafer and Galagate, 2015).

```
spline_estimate <- prop_spline_est(Y = TOTALEXP,
                                   treat = packyears,
                                   covar_formula = ~ LASTAGE + LASTAGE2 +
                                     AGESMOKE + AGESMOKE2 + MALE + beltuse +
                                     educate + marital + POVSTALB + RACE3,
                                   covar_lin_formula = ~ 1,
                                   covar_sq_formula = ~ 1,
                                   data = full_data_orig,
                                   e_treat_1 = full_data_orig$est_treat,
                                   degree = 2,
                                   wt = full_data_orig$HSQACCWT,
                                   method = "different",
                                   spline_df = 5,
                                   spline_const = 4,
                                   spline_linear = 4,
                                   spline_quad = 4)
spline_estimate

##
## Estimated values:
## [1] 1583.0374335   30.5793023   -0.1980041
```

This last method fits a spline basis to the estimated PF values and then regresses the outcome on both the basis terms and the treatment to estimate the ADRF. This is described in Imai and Van Dyk (2004) and Schafer and Galagate (2015). The estimated parameters correspond to coefficients for a quadratic model of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$.

```
ivd_estimate <- prop_spline_est(Y = TOTALEXP,
                                treat = packyears,
                                covar_formula = ~ 1,
                                covar_lin_formula = ~ 1,
                                covar_sq_formula = ~ 1,
                                data = full_data_orig,
                                e_treat_1 = full_data_orig$est_treat,
                                degree = 2,
                                wt = full_data_orig$HSQACCWT,
                                method = "different",
                                spline_df = 5,
                                spline_const = 4,
                                spline_linear = 4,
                                spline_quad = 4)
ivd_estimate

##
## Estimated values:
## [1] 1487.99099309   24.89207005   -0.05530696
```

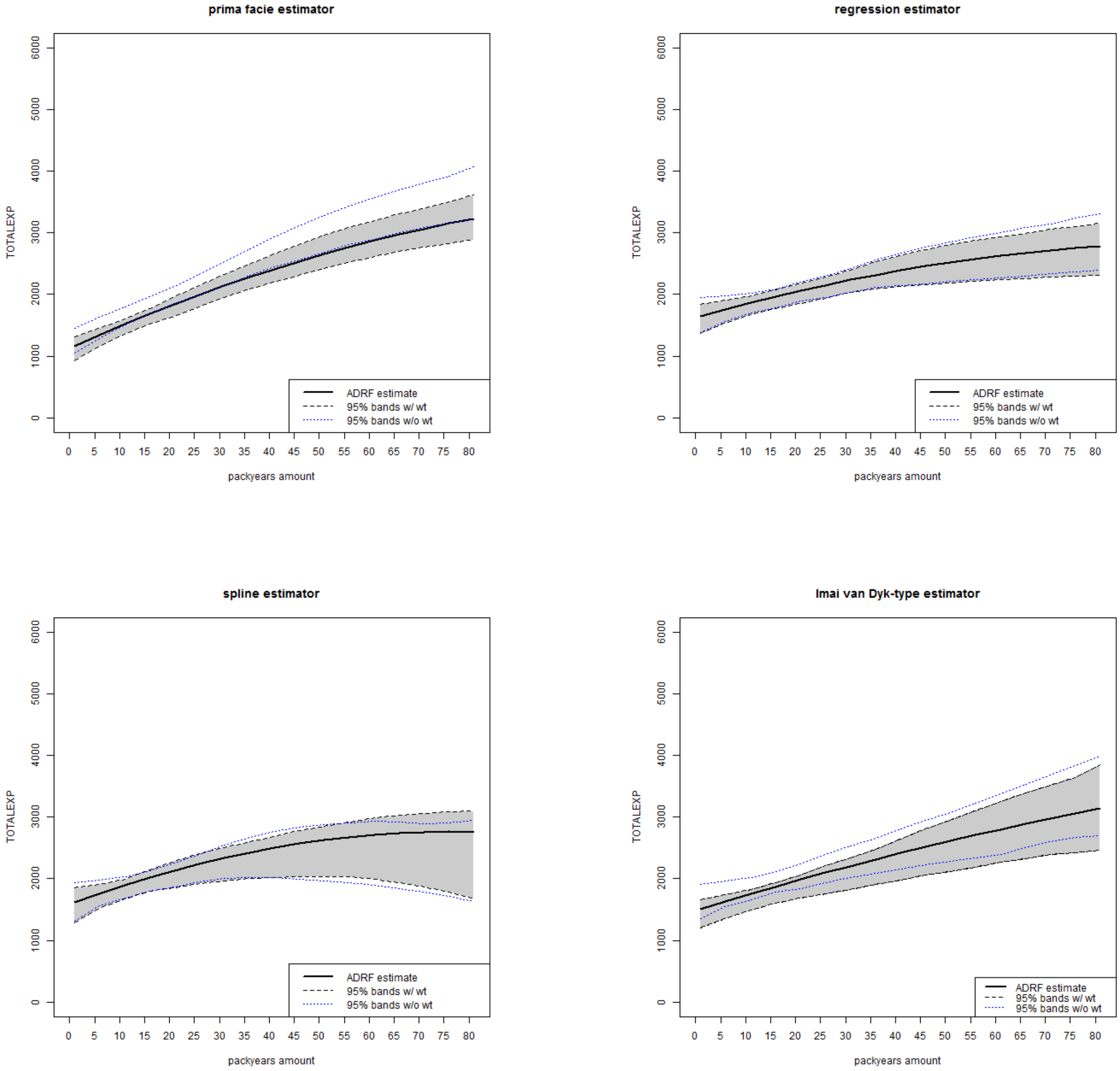

Figure 3: Estimated dose-response functions using 4 different methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process from the beginning.

### 3.6 Discussion

These four methods estimate the ADRF in a structured way and assumes the true ADRF is a linear combination of a finite number of basis functions. Figure 3 shows an overall rising amount of `TOTALEXP` as `packyear` increases. Recall that in this example, the four estimators are restricted to fitting the ADRF as a polynomial of up to degree 2. Fitting more flexible models may give slightly different curves. The next section analyzes a different data set and will fit other flexible estimators such as BART, which allows for flexible response surfaces to estimate the ADRF.

## 4 Analysis of the Infant Health and Development Program

### 4.1 Introduction

The next example on the Infant Health and Development Program is described by Gross (1992):

> The Infant Health and Development Program (IHDP) was a collaborative, randomized, longitudinal, multisite clinical trial designed to evaluate the efficacy of comprehensive early intervention in reducing the developmental and health problems of low birth weight, premature infants. An intensive intervention extending from hospital discharge to 36 months corrected age was administered between 1985 and 1988 at eight different sites. The study sample of infants was stratified by birth weight (2,000 grams or less, 2,001-2,500 grams) and randomized to the Intervention Group or the Follow-Up Group.

The intervention (treatment) group received more support than the control group. In addition to the standard pediatric follow-up, the treatment group also received home visits and attendance at a special child development center. Although the treatment was assigned randomly, families chosen for the intervention self-selected into different participation levels (Hill, 2011). Therefore, restricting our analysis to families in the intervention group and their participation levels leads to an observational setting.

In this section, even though families are randomly selected for intervention, we restrict our analysis on those selected for the treatment. These families choose the amount of days they attend the child development centers and this makes the data set, for practical purposes, an observational data set. We apply our methods on this subset of the data to estimate the ADRF for those who received the treatment.

We analyze this data set because the treatment variable, number of child development center days, is analyzed as a continuous variable. The data set we use comes from Hill (2011).

## 4.2 Data

Part of this data set is included in the supplement in Hill (2011), but does not include all the needed variables. The continuous treatment is available through the data repository at `icpsr.umich.edu`. To get the data, go to `http://www.icpsr.umich.edu/icpsrweb/HMCA/studies/9795?paging.startRow=51` and download DS141: Transport Format SAS Library Containing the 59 Evaluation Data Files - Download All Files (27.9 MB). After downloading the .zip file, extract the data file named "09795-0141-Data-card_image.xpt" to a folder and set the R working directory to this folder. The following instructions describe how to extract the continuous treatment variable.

Making sure the working directory contains "09795-0141-Data-card_image.xpt", the next step is to load the `Hmisc` package to read sas export files.

```
library(Hmisc)
mydata <- sasxport.get("09795-0141-Data-card_image.xpt")
data_58 <- mydata[[58]]
ihdp_raw <- data_58
# restricts data to treated cases
treated_raw <- ihdp_raw[which(ihdp_raw$tg == "I"),]
# continuous treatment variable
treat_value <- treated$cdays.t
```

The continuous treatment variable is merged with the data given in the supplement by Hill (2011) to create the data set for this section.

A few more steps are needed to clean and recode the data. We collect a subset of families eligible for the intervention and restrict the data set to families that use the child development centers at least once. The data set contains the outcome variable, `iqsb.36`, which is the measured iq of the child at 36 months. The treatment variable is the number of days the child attended the child development center divided by 100, `ncdctt` (i.e. `ncdctt` = 1.5 means 150 days in the child developement center). We select the covariates using a stepwise procedure to simplify the analysis.

## 4.3 Common support

```
overlap_temp <- overlap_fun(Y = iqsb.36,
                            treat = ncdctt,
                            treat_formula = t_formula,
                            data = data_set,
                            n_class = 3,
                            treat_mod = "Normal")

median_list <- overlap_temp[[2]]
```

```
overlap_orig <- overlap_temp[[1]]
overlap_3 <- overlap_temp[[3]]
fitted_values_overlap <- overlap_3$fitted.values
```

## 4.4 Covariate balance

Balance is evaluated similarly to the NMES example.

## 4.5 Estimating the ADRF

The BART estimator fits a rich outcome model on the treatment and covariates to create a flexible response surface (Hill, 2011). The flexible response surface imputes the missing potential outcomes. The estimated potential outcomes are averaged to get the estimated ADRF over a grid of treatment values.

```
bart_estimate <-  bart_est(Y = iqsb.36,
                             treat = ncdctt,
                             outcome_formula = iqsb.36 ~ ncdctt + bw +
                             female + mom.lths +
                             site1 + site7 + momblack +
                             workdur.imp,
                             data = full_data_orig,
                             grid_val = grid_treat)
```

The next method is described in Flores et al. (2012) and uses inverse weights to adjust for the covariates. First a treatment model is fit and GPS values are estimated. This is a method that uses weights to locally regress the outcome on nearby points. This is a local linear regression of the outcome, `iqsb.36`, on the treatment, `ncdctt`, with a weighted kernel. The weighted kernel is weighted by the reciprocal of the GPS values.

```
iw_estimate <- iw_est(Y = iqsb.36,
                       treat = ncdctt,
                       treat_formula = ncdctt ~ bw + female + mom.lths +
                        site1 + site7 + momblack +
                        workdur.imp,
                       data = full_data_orig,
                       grid_val = grid_treat,
                       bandw = 2 * bw.SJ(full_data_orig$ncdctt),
                       treat_mod = "Normal")
```

This next method, the Nadaraya-Watson based estimator, is similar to the inverse weighting method in the previous code chunk, but uses a local constant regression.

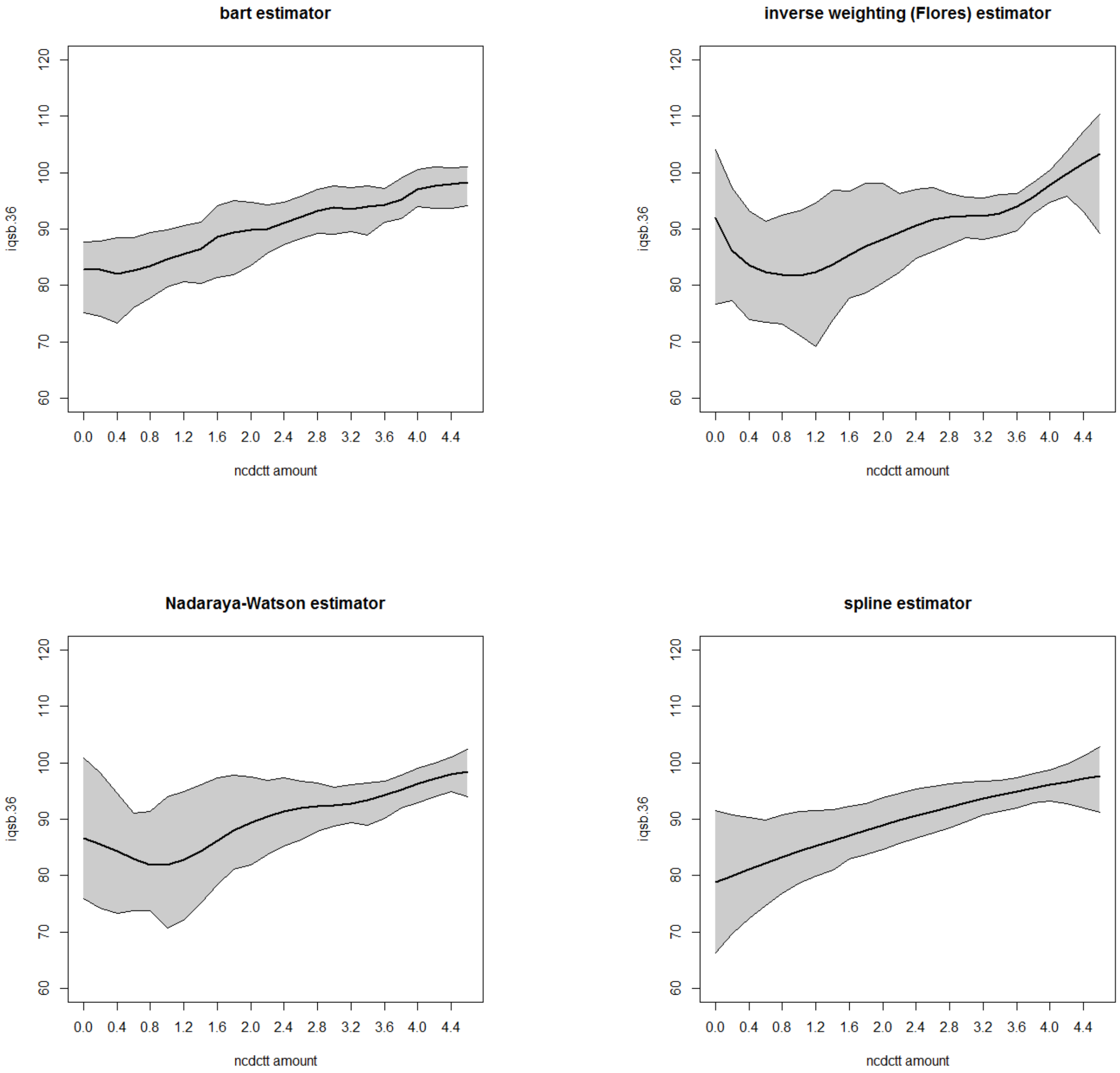

Figure 4: Estimated dose-response functions using 4 different methods with 95% pointwise standard errors. The standard errors are estimated by bootstrapping the entire estimation process from the beginning.

```
nw_estimate <- nw_est(Y = iqsb.36,
                      treat = ncdctt,
                      treat_formula = ncdctt ~ bw + female + mom.lths +
                        site1 + site7 + momblack +
                        workdur.imp,
                      data = full_data_orig,
                      grid_val = grid_treat,
                      bandw = 2 * bw.SJ(full_data_orig$ncdctt),
                      treat_mod = "Normal")
```

The propensity spline estimator is a generalization of the prima facie and regression prediction method in Schafer and Galagate (2015). In this example, the estimator is restricted to a polynomial of up to degree 2 of the form $\hat{\mu}(t) = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2$.

```
spline_estimate <- prop_spline_est(Y = iqsb.36,
                                   treat = ncdctt,
                                   covar_formula = ~ bw + female +
                                     mom.lths + site1 + site7 +
                                     momblack + workdur.imp,
                                   covar_lin_formula = ~ 1,
                                   covar_sq_formula = ~ 1,
                                   data = full_data_orig,
                                   e_treat_1 = full_data_orig$est_treat,
                                   degree = 2,
                                   wt = NULL,
                                   method = "different",
                                   spline_df = 5,
                                   spline_const = 2,
                                   spline_linear = 2,
                                   spline_quad = 2)
```

## 4.6 Discussion

The plots in Figure 4 show the estimated relationship of IQ at 36 months, `iqsb.36`, on number of days in the child development care center, `ncdctt`. The inverse weighting and Nadaraya-Watson show a decreasing trend for `ncdctt` $\in (0, 0.8)$, but an increasing trend for `ncdctt` $> 0.8$. These estimators are jagged because of the bandwidth selection. In this example, we use twice the Sheather-Jones bandwidth estimate to select the bandwidth. Picking a larger bandwidth will give smoother estimates. The BART and propensity spline estimators have a generally increasing trend.

# 5 Conclusion

In this chapter, we have demonstrated how to estimate ADRFs using different statistical techniques using the R package `causaldrf`, both for simulated and real data, by correcting for confounding variables. `causaldrf` can accommodate a wide array of treatment models, is user friendly, and does not require extensive programming. This contribution of the R package `causaldrf` will make ADRF estimation more accessible to applied researchers. In future updates of the package, the functions will be adapted to an even wider range of problems.

# References

Bia, M., Flores, C. A. F., Flores-Lagunes, A., and Mattei, A. (2014). A stata package for the application of semiparametric estimators of dose-response functions. *The Stata journal*, 14(3):580–604.

Cohen, S. B. (2000). *Sample Design of the 1997 Medical Expenditure Panel Survey, Household Component*. US Department of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality.

D'Agostino Jr, R. B. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association*, 95(451):749–759.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Review of Economics and Statistics*, 94(1):153–171.

Gross, R. T. (1992). *Infant Health and Development Program (IHDP): Enhancing the Outcomes of Low Birth Weight; Premature Infants in the United States, 1985-1988*. Inter-university Consortium for Political and Social Research.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1).

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, 99(467).

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics*, 112(1):135–151.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14:949–968.

Moodie, E. E. and Stephens, D. A. (2012). Estimation of dose–response functions for longitudinal data using the generalised propensity score. *Statistical methods in medical research*, 21(2):149–166.

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Schafer, J. and Galagate, D. (2015). Causal inference with a continuous treatment and outcome: alternative estimators for parametric dose-response models. *Manuscript in preparation.*

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279.

Singh, R. P., Petroni, R. J., Allen, T. M., et al. (1994). Oversampling in panel surveys. *Bureau of the Census.*