3. Retirement?? Retirement really COULD kill you: Researchers find those who work past 65 live longer bonus music Only The Good Die Young
Publication: Association of retirement age with mortality: a population-based longitudinal study among older adults in the USA. *Journal of Epidemiology and Community Health*.

Computing Corner: **Dose response functions (and multiple groups): Beyond Binary Treatments**
package `causaldrf` vignette: Estimating Average Dose Response Functions Using the R Package causaldrf Rnw file for vignette dot-R file for vignette
Rogosa session, causaldrf examples
also covariate balancing propensity score, package CBPS
Background publications:
The Propensity Score with Continuous Treatments
Causal Inference With General Treatment Regimes: Generalizing the Propensity Score, Journal of the American Statistical Association, Vol. 99, No. 467 (September), pp. 854-866.

bumped: graphical model (and do-calculus) applications: one resource, Identifying Causal Effects with the R Package causaleffect

Week 9 Review Questions
Computing Exercises
1. Classic "Sharp" design. Replicate the package rdd toy example: cutpoint = 0, sharp design, with treatment effect of 3 units (instead of 10). Try out the analysis of covariance (Rubin 1977) estimate and compare with rdd output and plot. Pick off the observations used in the Half-BW estimate and verify using t-test or wilcoxon.
Extra: try out also the `rdrobust` package for this sharp design.
Solution for Review Question 1
2. Systematic Assignment, "fuzzy design". Probabilistic assignment on the basis of the covariate.
i. Create artificial data with the following specification. 10,000 observations; premeasure (Y_uc in my session) gaussian mean 10 variance 1. Effect of intervention (rho) if in the treatment group is 2 (or close to 2) and uncorrelated with Y_uc. Probability of being in the treatment group depends on Y_uc but is not a deterministic step-function ("sharp design"): `Pr(treatment|Y_uc) = pnorm(Y_uc, 10,1)` . Plot that function.
ii. Try out analysis of covariance with Y_uc as covariate. Obtain a confidence interval for the effect of the treatment.
iii. Try out the fancy econometric estimators (using finite support) as in the rdd package. See if you find that they work poorly in this very basic fuzzy design example.
Extra: try out also the `rdrobust` package for this fuzzy design.
Solution for Review Question 2

self-select into level of exercise -------> health benefits

self-select into level of education ----------> career salary effects

self-select level of salt intake ------------------> effect on BP

DOSE-RESPONSE

also fish, greenery
neighborhoods

The Propensity Score with Continuous Treatments*

Keisuke Hirano
University of Miami

Guido W. Imbens
UC Berkeley and NBER

February 7, 2004

hi_est in session

1 Introduction

Much of the work on propensity score analysis has focused on the case where the treatment is binary. In this chapter we examine an extension to the propensity score method, in a setting with a continuous treatment. Following Rosenbaum and Rubin (1983) and most of the other literature on propensity score analysis, we make an unconfoundedness or ignorability assumption, that adjusting for differences in a set of covariates removes all biases in comparisons by treatment status. Then, building on Imbens (2000) we define a generalization of the binary treatment propensity score, which we label the generalized propensity score (GPS). We demonstrate that the GPS has many of the attractive properties of the binary treatment propensity score. Just as in the binary treatment case, adjusting for this scalar function of the covariates removes all biases associated with differences in the covariates. The GPS also has certain balancing properties that can be used to assess the adequacy of particular specifications of the score. We discuss estimation and inference in a parametric version of this procedure, although more flexible approaches are also possible.

We apply this methodology to a data set collected by Imbens, Rubin, and Sacerdote (2001). The population consists of individuals winning the Megabucks lottery in Massachusetts in the mid-1980's. We are interested in effect of the amount of the prize on subsequent labor earnings. Although the assignment of the prize is obviously random, substantial item and unit nonresponse led to a selected sample where the amount of the prize is no longer independent of background characteristics. We estimate the average effect of the prize adjusting for differences in background characteristics using the propensity score methodology, and compare the results to conventional regression estimates. The results suggest that the propensity score methodology leads to credible estimates, that can be more robust than simple regression estimates.

*This is a draft of a chapter for *Missing Data and Bayesian Methods in Practice: Contributions by Donald Rubin's Statistical Family*, forthcoming from Wiley. Financial support for this research was generously provided through NSF grants SES-0226164 (Hirano) and SES-0136789 (Imbens). Electronic correspondence: khirano@miami.edu, http://www.bus.miami.edu/~khirano/, imbens@econ.berkeley.edu, http://elsa.berkeley.edu/users/imbens/.

E(D|X) for dose D as for binary Pr(T|X)

Causal Inference With General Treatment Regimes: Generalizing the Propensity Score

Kosuke IMAI and David A. VAN DYK

In this article we develop the theoretical properties of the propensity function, which is a generalization of the propensity score of Rosenbaum and Rubin. Methods based on the propensity score have long been used for causal inference in observational studies; they are easy to use and can effectively reduce the bias caused by nonrandom treatment assignment. Although treatment regimes need not be binary in practice, the propensity score methods are generally confined to binary treatment scenarios. Two possible exceptions have been suggested for ordinal and categorical treatments. In this article we develop theory and methods that encompass all of these techniques and widen their applicability by allowing for arbitrary treatment regimes. We illustrate our propensity function methods by applying them to two datasets; we estimate the effect of smoking on medical expenditure and the effect of schooling on wages. We also conduct simulation studies to investigate the performance of our methods.

KEY WORDS: Medical expenditure; Nonrandom treatment assignment; Observational studies; Return to schooling; Subclassification; Treatment effect.

1. INTRODUCTION

Establishing the effect of a treatment that is not randomly assigned is a common goal in empirical research. But the lack of random assignment means that groups with different levels of the treatment variable can systematically differ in important ways other than the observed treatment. Because these differences may exhibit complex correlations with the outcome variable, ascertaining the causal effect of the treatment may be difficult. It is in this setting that the propensity score of Rosenbaum and Rubin (1983b) has found wide applicability in empirical research; in particular, the method has rapidly become popular in the social sciences (e.g., Heckman, Ichimura, and Todd 1998; Lechner 1999; Imai 2004).

The propensity score aims to control for differences between the treatment groups when the treatment is binary; it is defined as the conditional probability of assignment to the treatment group given a set of observed pretreatment variables. Under the assumption of strongly ignorable treatment assignment, multivariate adjustment methods based on the propensity score have the desirable property of effectively reducing the bias that frequently arises in observational studies. In fact, there exists empirical evidence that in certain situations the propensity score method produces more reliable estimates of causal effects than other estimation methods (e.g., Dehejia and Wahba 1999; Imai 2004).

The propensity score is called a *balancing score* because, conditional on the propensity score, the binary treatment assignment and the observed covariates are independent (Rosenbaum and Rubin 1983b). If we further assume the conditional independence between treatment assignment and potential outcomes given the observed covariates, then it is possible to obtain unbiased estimates of treatment effects. In practice, matching or subclassification is used to adjust for the *estimated* propensity score, which is ordinarily generated by logistic regression (Rosenbaum and Rubin 1984, 1985). The advantage of using estimated propensity scores in place of true propensity scores has been discussed at length in the literature (e.g., Rosenbaum 1987; Robins, Rotnitzky, and Zhao 1995; Rubin and Thomas 1996; Heckmen et al. 1998; Hirano, Imbens, and Ridder 2003); see also Section 5.3. Indeed, even in randomized experiments where the randomization scheme specifies the true propensity score, adjusting for the estimated propensity score can reduce the variance of the estimated treatment effect. One of the principle advantages of this method is that adjusting for the propensity score amounts to matching or subclassifying on a scalar, which is significantly easier than matching or subclassifying on many covariates.

In this article we extend and generalize the propensity score method so that it can be applied to arbitrary treatment regimes. The original propensity score was developed to estimate the causal effects of a binary treatment; however, in many observational studies, the treatment may not be binary or even categorical. For example, in clinical trials, one may be interested in estimating the dose-response function where the drug dose may take on a continuum of values (e.g., Efron and Feldman 1991). Alternatively, the treatment may be ordinal. In economics, an important quantity of interest is the effect of schooling on wages, where schooling is measured as years of education in school (e.g., Card 1995). The treatment can also consist of multiple factors and their interactions. In political science, one may be interested in the combined effects of different voter mobilization strategies, such as phone calls and door-to-door visits (e.g., Gerber and Green 2000). Treatment can also be measured in terms of frequency and duration, for example, the health effects of smoking. These examples illustrate the need to extend the propensity score, a prominent methodology of causal inference, for application to general treatment regimes.

Two extensions of the propensity score have been developed to handle a univariate categorical or ordinal treatment variable. (We use the term "ordinal variable" to refer to a discrete variable that takes on ordered values, whereas a "categorical variable" is discrete with possibly unordered values.) Imbens (2000) suggested computing a propensity score for each level

Kosuke Imai is Assistant Professor, Department of Politics, Princeton University, Princeton, NJ 08544 (E-mail: *kimai@princeton.edu*). David A. van Dyk is Associate Professor, Department of Statistics, University of California, Irvine, CA 92697-1250 (E-mail: *dvd@ics.uci.edu*). The authors thank Joshua Angrist, Guido Imbens, Elizabeth Johnson, and Scott Zegar for providing the datasets used in this article. They also thank Jim Alt, Samantha Cook, Jennifer Hill, Dan Ho, Gary King, Donald Rubin, Phil Schrodt, Jas Sekhon, and Elizabeth Stuart for helpful discussions. The comments from the associate editor and anonymous referees significantly improved this article. Research support was provided by National Science Foundation grant DMS-01-04129, the U.S. Census Bureau, and the Princeton University Committee on Research in the Humanities and Social Sciences.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000001187

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.

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*

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

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.

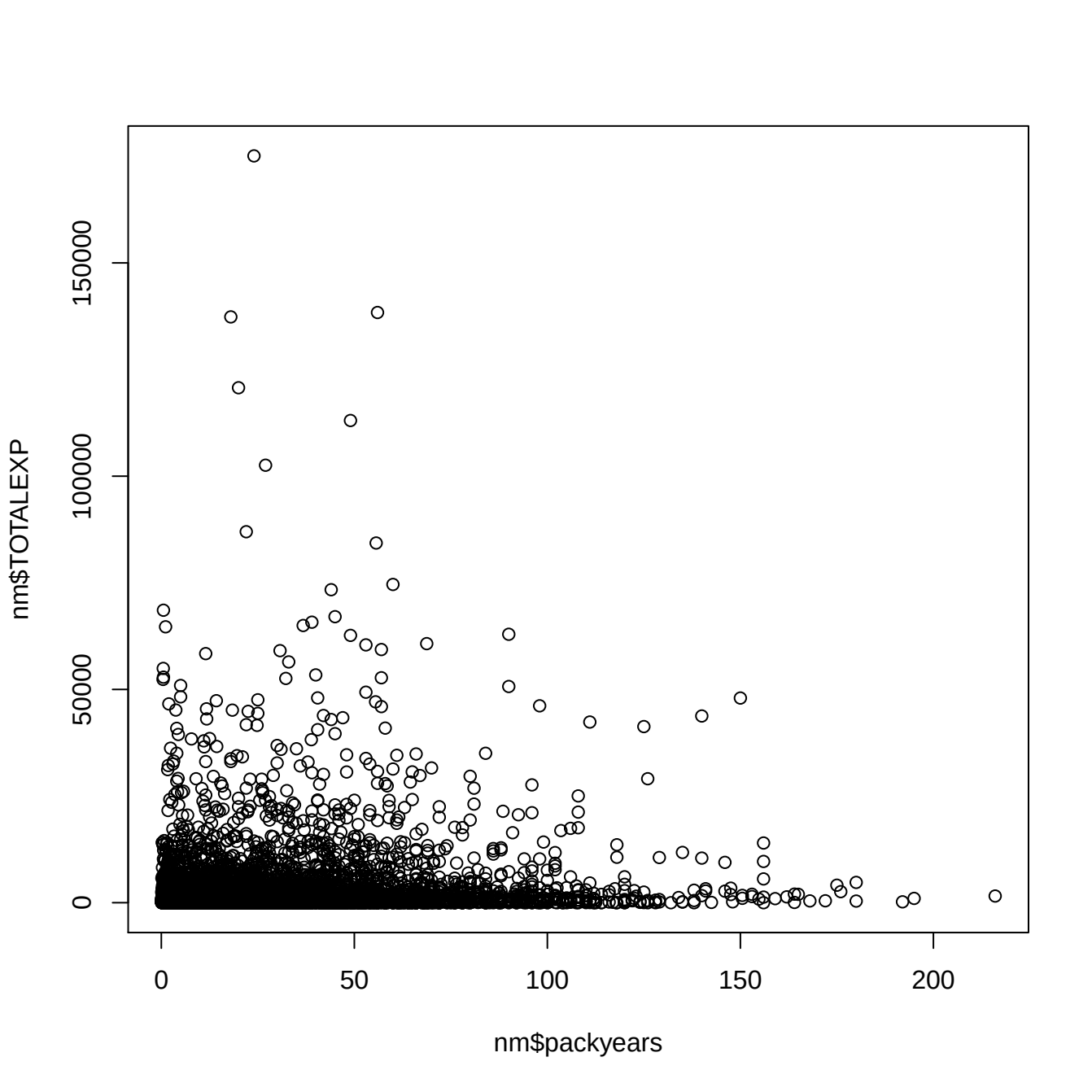
Determining the effect of smoking on health has a long history. Scientists cannot ethically assign smoking amounts randomly to people because of the potential negative effects, so

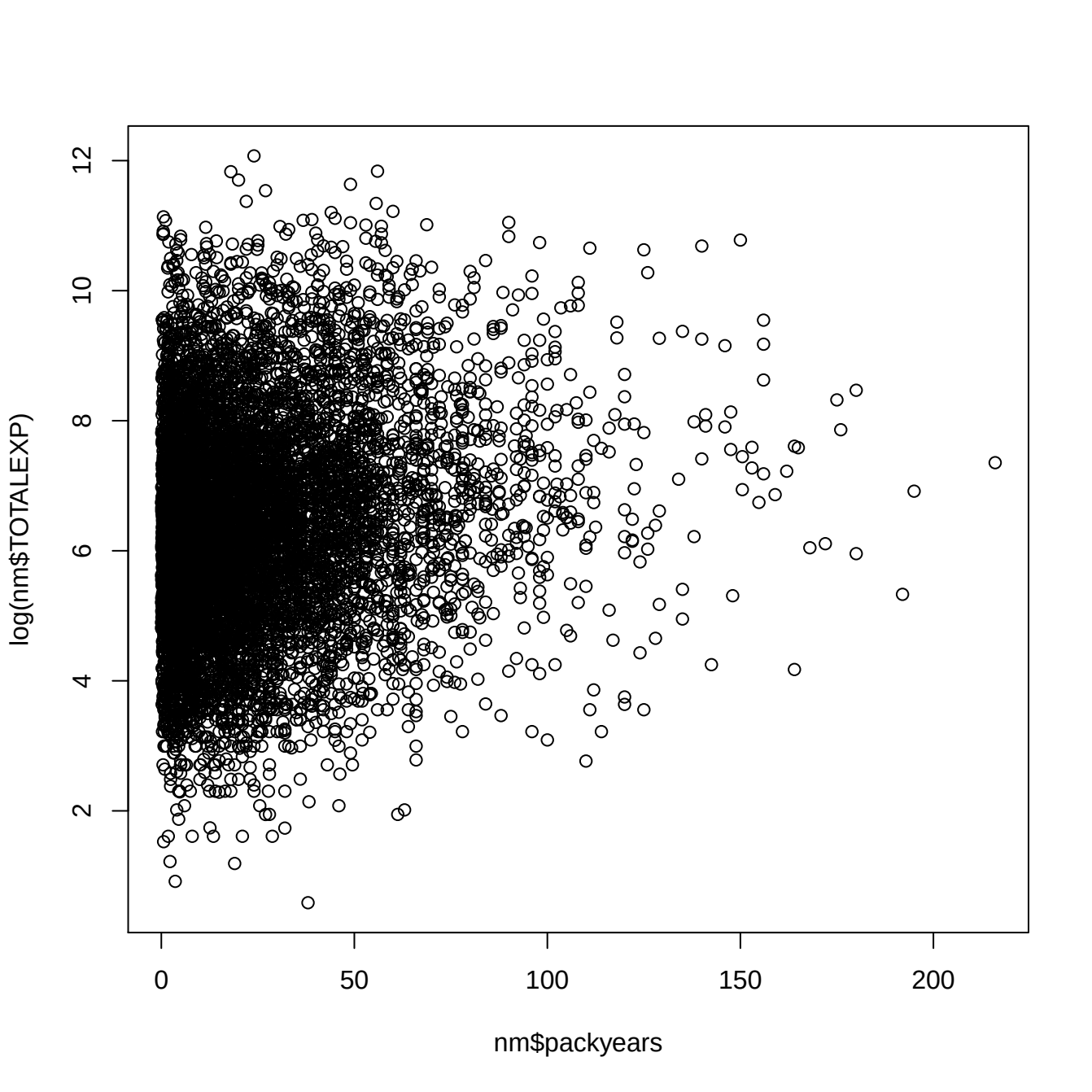
```
R version 3.2.2 (2015-08-14) -- "Fire Safety"
> install.packages("causaldrf")
> library(causaldrf)

> data(nmes_data)
> dim(nmes_data)
[1] 9708   12
> nm = nmes_data
> dim(nm)
[1] 9708   12

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

> plot(nm$packyears, nm$TOTALEXP)
# not much dose response evident??
> plot(nm$packyears, log(nm$TOTALEXP))
>
```


nm$TOTALEXP
0
50000
100000
150000
0
50
100
150
200
nm$packyears


log(nm$TOTALEXP)
2
4
6
8
10
12
0
50
100
150
200
nm$packyears

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).

Package 'causaldrf'

November 30, 2015

Type Package

Title Tools for Estimating Causal Dose Response Functions

Version 0.3

Date 2015-11-27

Description Functions and data to estimate causal dose response functions given continuous, ordinal, or binary treatments.

License MIT + file LICENSE

LazyData TRUE

Depends R(>= 3.1.2)

Imports mgcv, splines, stats, survey,

Suggests BayesTree, dplyr, foreign, Hmisc, knitr, MASS, nnet, reshape2, rmarkdown, sas7bdat, testthat, tidyr

VignetteBuilder knitr

NeedsCompilation no

Author Douglas Galagate [cre], Joseph Schafer [aut]

Maintainer Douglas Galagate <galagated@gmail.com>

Repository CRAN

Date/Publication 2015-11-30 17:18:48

R topics documented:

`add_spl_est` *The additive spline estimator*

Description

This function estimates the ADRF with an additive spline estimator described in Bia et al. (2014).

Usage

```
add_spl_est(Y,
            treat,
            treat_formula,
            data,
            grid_val,
            knot_num,
            treat_mod,
            link_function,
            ...)
```

Arguments

`Y`	is the the name of the outcome variable contained in `data`.
`treat`	is the name of the treatment variable contained in `data`.
`treat_formula`	an object of class "formula" (or one that can be coerced to that class) that regresses `treat` on a linear combination of `X`: a symbolic description of the model to be fitted.
`data`	is a dataframe containing `Y`, `treat`, and `X`.
`grid_val`	contains the treatment values to be evaluated.
`knot_num`	is the number of knots used in outcome model
`treat_mod`	a description of the error distribution to be used in the model for treatment. Options include: `"Normal"` for normal model, `"LogNormal"` for lognormal model, `"Sqrt"` for square-root transformation to a normal treatment, `"Poisson"` for Poisson model, `"NegBinom"` for negative binomial model, `"Gamma"` for gamma model.

3.5 Estimating the ADRF

FACE in Holland 1988

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
```

```
                        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.

poor performance in sim

```
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.

Usage

```
hi_est(Y,
       treat,
       treat_formula,
       outcome_formula,
       data,
       grid_val,
       treat_mod,
       link_function,
       ...)
```

Arguments

`Y`	is the the name of the outcome variable contained in `data`.
`treat`	is the name of the treatment variable contained in `data`.
`treat_formula`	an object of class "formula" (or one that can be coerced to that class) that regresses `treat` on a linear combination of `X`: a symbolic description of the model to be fitted.
`outcome_formula`	is the formula used for fitting the outcome surface. gps is one of the independent variables to use in the outcome_formula. ie. `Y ~ treat+ I(treat^2) + gps + I(gps^2) + treat * gps` or a variation of this. Use `gps` as the name of the variable representing the gps in `outcome_formula`.
`data`	is a dataframe containing `Y`, `treat`, and `X`.
`grid_val`	contains the treatment values to be evaluated.
`treat_mod`	a description of the error distribution to be used in the model for treatment. Options include: `"Normal"` for normal model, `"LogNormal"` for lognormal model, `"Sqrt"` for square-root transformation to a normal treatment, `"Poisson"` for Poisson model, `"NegBinom"` for negative binomial model, `"Gamma"` for gamma model, `"Binomial"` for binomial model.
`link_function`	For `treat_mod = "Gamma"` (fitted using glm) alternatives are "log" or "inverse". For `treat_mod = "Binomial"` (fitted using glm) alternatives are "logit", "probit", "cauchit", "log" and "cloglog".
`...`	additional arguments to be passed to the outcome lm() function.

Details

Hirano (2004) (HI) introduced this imputation-type method that includes a GPS component. The idea is to fit a parametric observable (outcome) model, which includes the estimated GPS as a covariate, to impute missing potential outcomes.

The method requires several steps. First, a model is used to relate treatment to the recorded covariates. For example, $T_i|\mathbf{X}_i \sim \mathcal{N}(\mathbf{X}_i^T\boldsymbol{\beta}, \sigma^2)$ and then estimate the $\boldsymbol{\beta}$ parameters. Next, the GPS for each unit is estimated

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:

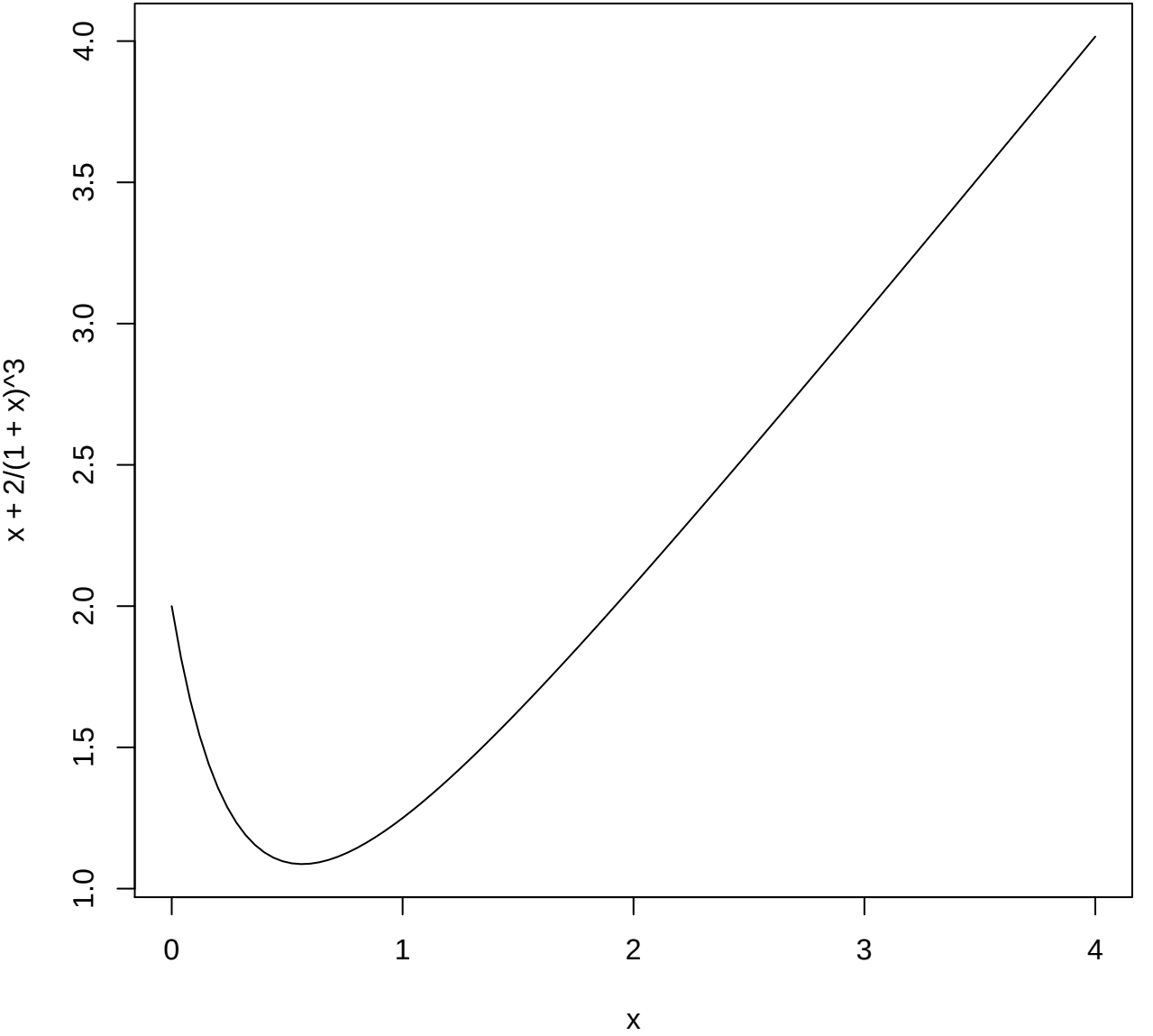
```
library (causaldrf)
```

The data is generated from:

```
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)
```

my curve-- plot,looks like hazard for alcohol on cvd, nurses



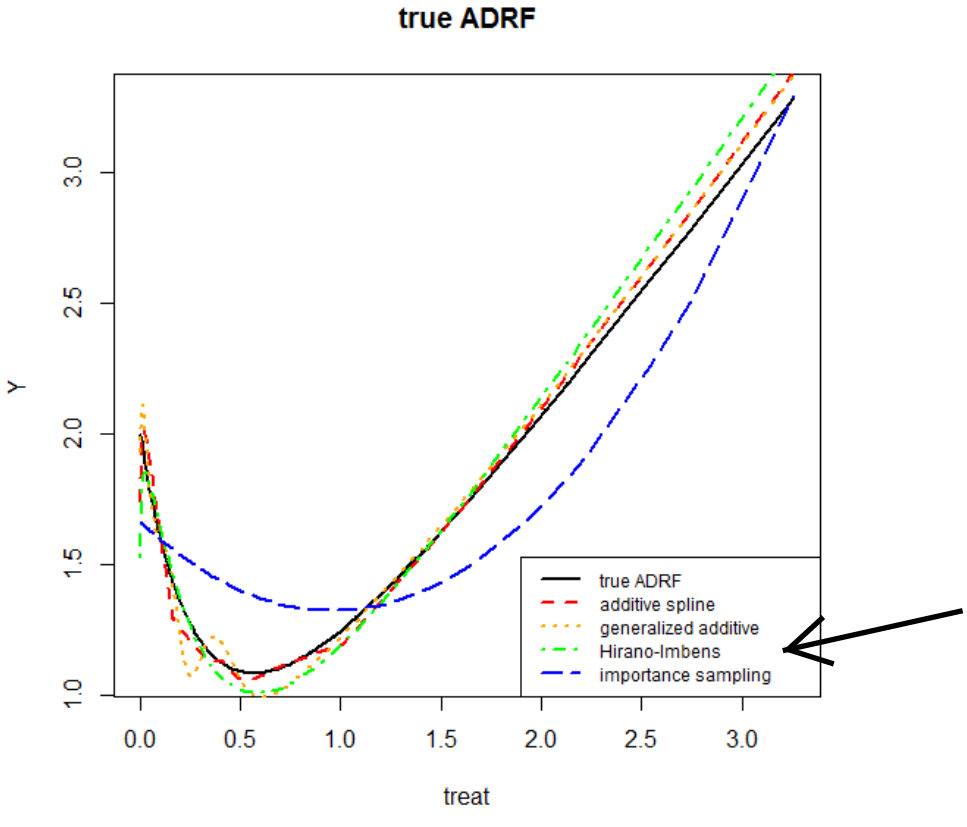
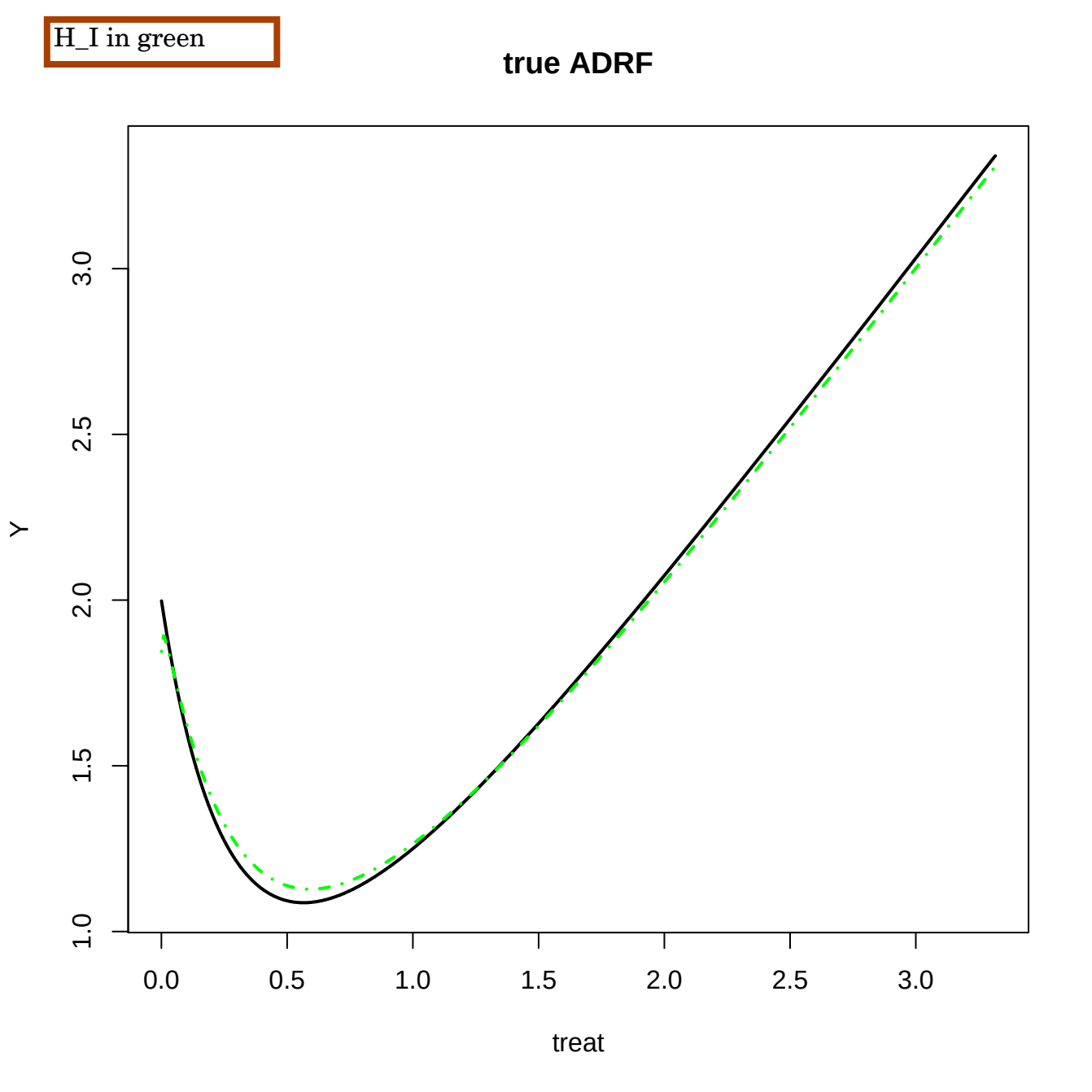


Figure 1: True ADRF along with estimated curves.

see linked .Rnw file for construction of this summary plot


H_I in green
true ADRF
Y
1.0
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
treat

```
   ##### Week 9 CC, dose-response functions

R version 3.2.2 (2015-08-14) -- "Fire Safety"
> install.packages("causaldrf")
> library(causaldrf)

> data(nmes_data)
> dim(nmes_data)
[1] 9708   12
> nm = nmes_data
> dim(nm)
[1] 9708   12

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

> plot(nm$packyears, nm$TOTALEXP)
# not much dose response evident??
> plot(nm$packyears, log(nm$TOTALEXP))
>
##########################################################################
## Artificial data example
# true dose-response
> ?curve
> curve(x + 2/(1 + x)^3, 0,4)  # plot shown in CC materials

> # I read in the supplied sim data rather than create it
> data(hi_sim_data)
> ?hi_sim_data
starting httpd help server ... done

> sim = hi_sim_data #simplify my typing
> head(sim)
         X1         X2         T        gps         Y
1 2.8762787 0.52729990 0.2223654 1.59678021 1.0393833
2 0.4875109 0.18797037 0.5856397 0.45479046 0.5546942
3 0.7407761 0.22908956 0.3763913 0.67324450 0.3727181
4 0.6561316 1.29076597 2.0496851 0.03599808 3.2768007
5 0.2495930 1.02818788 1.0473338 0.33516304 1.6742432
6 0.3888915 0.07456587 1.4867275 0.23268359 1.8758940
```

```
> # run the imbens estimator
> 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 = sim,
+                       grid_val = quantile(hi_sim_data$T,
+                                           probs = seq(0, .95, by = 0.01)),
+                       treat_mod = "Gamma",
+                       link_function = "inverse")
> summary(hi_estimate)

Estimated values:
 [1] 1.844885 1.892549 1.885924 1.872564 1.859171 1.843147 1.833042 1.821880
 [9] 1.810837 1.789285 1.771371 1.753162 1.740413 1.715755 1.697731 1.681189
[17] 1.654436 1.641968 1.623282 1.601950 1.589434 1.568637 1.559566 1.548223
[25] 1.531785 1.513656 1.489437 1.475388 1.454679 1.445420 1.429200 1.419578
[33] 1.398187 1.377899 1.366134 1.349114 1.333631 1.311605 1.295041 1.282153
[41] 1.266689 1.249200 1.236905 1.225240 1.213712 1.207096 1.198966 1.194443
[49] 1.186686 1.181235 1.173873 1.162262 1.155129 1.145633 1.141084 1.135167
[57] 1.132318 1.129410 1.127848 1.127688 1.128432 1.130149 1.133215 1.136869
[65] 1.141953 1.146036 1.156082 1.165157 1.174271 1.188479 1.207558 1.222418
[73] 1.236831 1.253292 1.269864 1.284218 1.317895 1.338323 1.366325 1.412826
[81] 1.446591 1.483681 1.532863 1.613043 1.668634 1.738570 1.809313 1.875484
[89] 1.977283 2.064272 2.206672 2.328933 2.515111 2.726571 2.860258 3.308236

Treatment Summary:

Call:
glm(formula = formula_t, family = Gamma(link = link_function),
    data = samp_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2278  -1.0433  -0.3420   0.3783   2.8016

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01262    0.02422  -0.521    0.602
X1           0.99870    0.06560  15.223   <2e-16 ***
X2           0.97901    0.06462  15.151   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.9431856)

    Null deviance: 1888.7  on 999  degrees of freedom
Residual deviance: 1135.7  on 997  degrees of freedom
AIC: 1187

Number of Fisher Scoring iterations: 6
```

```
lm(formula = outcome_formula, data = tempdat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0078 -0.6943  0.0184  0.6469  3.1011

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.094e-01  1.331e-01  -0.822    0.411
T            1.022e+00  5.696e-02  17.949   <2e-16 ***
I(T^2)      -2.309e-03  4.504e-03  -0.513    0.608
gps          1.069e+00  1.045e-01  10.232   <2e-16 ***
I(gps^2)    -5.769e-06  2.115e-02   0.000    1.000
T:gps        3.214e-01  3.088e-01   1.041    0.298
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 994 degrees of freedom
Multiple R-squared:  0.7032,    Adjusted R-squared:  0.7017
F-statistic: 471.1 on 5 and 994 DF,  p-value: < 2.2e-16

# Vignette plot compares handfull of estimates, here's HI

> x <- hi_sim_data$T
> quantile_grid <- quantile(x, probs = seq(0, .95, by = 0.01))
>  # quantile_grid <- quantile_grid[1:100]
>  true_hi_fun <- function(t){t + 2/(1 + t)^3}
> plot(quantile_grid,
+       true_hi_fun(quantile_grid),
+       pch = ".",
+       main = "true ADRF",
+       xlab = "treat",
+       ylab = "Y",
+       col = "black")
>
>  lines(quantile_grid,
+        true_hi_fun(quantile_grid),
+        col = "black",
+        lty = 1,
+        lwd = 2)
> lines(quantile_grid,
+        hi_estimate$param,
+        lty = 4,
+        col = "green",
+        lwd = 2)
> # adapt code from vignette .Rnw to get this reduced plot
```