chap10-standardization

yousri elfattah

2024-10-08

# Chapter 10

# Causal Inference with Observational Data

## Standardization Model

**Load required libraries**.

The CausalModels library is used to fit causal models like standardization.

# Load libraries  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(haven)  
library(CausalModels)  
library(ggplot2)

## Loading the dataset:

This code will output the first few rows and the structure of your dataset into text files, making it easier for you to read and review the results.

1. **Loading the Dataset**:
   * read\_dta("cattaneo2.dta") reads the Stata dataset into R.
2. Viewing the first few rows of data:
   * as.data.frame(data) converting data to a standard data frame for full column display.
   * head(data\_df) Inspect the first few rows and column names.
3. **Capturing str(data) Output**:
   * capture.output(str(data)) captures the output of the str(data) function as a character vector.
   * writeLines(str\_output, "str\_data\_output.txt") writes this output to a text file named str\_data\_output.txt.
4. **Printing Confirmation**:
   * cat statements print confirmation messages to the console, letting you know the outputs have been successfully saved.

# Load necessary library  
library(haven) # for read\_dta function  
  
# Load the dataset  
data <- haven::read\_dta("cattaneo2.dta")  
  
# Convert to standard data frame for full column display  
data\_df <- as.data.frame(data)  
  
# Inspect the first few rows and column names  
head(data\_df)

## bweight mmarried mhisp fhisp foreign alcohol deadkids mage medu fage fedu  
## 1 3459 1 0 0 0 0 0 24 14 28 16  
## 2 3260 0 0 0 1 0 0 20 10 0 0  
## 3 3572 1 0 0 1 0 0 22 9 30 9  
## 4 2948 1 0 0 0 0 0 26 12 30 12  
## 5 2410 1 0 0 0 0 0 20 12 21 14  
## 6 3147 0 0 0 0 0 0 27 12 40 12  
## nprenatal monthslb order msmoke mbsmoke mrace frace prenatal birthmonth  
## 1 10 30 2 0 0 1 1 1 12  
## 2 6 42 3 0 0 0 0 1 7  
## 3 10 17 3 0 0 1 1 1 3  
## 4 10 34 2 0 0 1 1 1 1  
## 5 12 0 1 0 0 1 1 1 3  
## 6 9 0 1 0 0 1 1 1 4  
## lbweight fbaby prenatal1  
## 1 0 0 1  
## 2 0 0 1  
## 3 0 0 1  
## 4 0 0 1  
## 5 1 1 1  
## 6 0 1 1

# Check the structure of the dataset to see data types  
# Use capture.output to capture the output of str(data)  
str\_output <- capture.output(str(data))  
writeLines(str\_output, "str\_data\_output.txt")  
  
# Print to console to verify  
cat("Output of str(data) saved to str\_data\_output.txt\n")

## Output of str(data) saved to str\_data\_output.txt

## Initialize CausalModels Package

The function CausalModels::init\_params must be run first before any other function can run.

Arguments:

1. **Outcome** - the outcome variable (must be continuous). We set outcome = bweight, which is the “infant birthweight (grams)”
2. **Treatment**—the treatment with the causal effect of interest on the outcome. We set treatment = mbsmoke, which is “1 if the mother smoked.” It has a value of one if the mother is a smoker and 0 if she is a nonsmoker.
3. **Covariates** - a list/vector of covariate names for confounding adjustment. We set the confounders to the list of variables as follows:
   * **“mmarried”**: “1 if mother married” takes the value 0 indicating “notmarried”, and one indicating “married.”
   * **“prenatal1”**: “1 if first prenatal visit in 1 trimester” takes the value “Yes”-“No”
   * **“fbaby”**: “1 if first baby” takes the value “Yes”-“No”
   * **“medu”**: “mother’s education attainment”
4. **Data** - is the data frame containing the variables in the model.
5. **Simple** - a boolean indicator to build a default formula with interactions. If true, interactions will be excluded. If false, interactions will be included. By default, simple is set to false.

The line

data$mbsmoke <- as.factor(data$mbsmoke)

converts the data$mbsmoke variable in the **cattaneo2** dataset into a factor. The mbsmoke variable indicates whether a mother smoked or not. By converting it to a factor, you inform R that this variable should be treated as categorical, which is crucial for models that handle categorical predictors differently than continuous variables. It allows R to properly interpret **mbsmoke** as levels of a categorical variable in the regression model you are running. The as.factor function converts a variable into a factor, an R data type used to represent categorical data.

# Initialize the parameters for the CausalModels package  
data$mbsmoke <- as.factor(data$mbsmoke)  
confounders <- c("mmarried", "prenatal1", "fbaby","medu")  
CausalModels::init\_params(outcome = bweight, treatment = mbsmoke, covariates = confounders, data)

## Successfully initialized!  
##   
## Summary:  
##   
## Outcome - bweight   
## Treatment - mbsmoke   
## Covariates - [ mmarried, prenatal1, fbaby, medu ]   
##   
## Size - 4642 x 23   
##   
## Default formula for outcome models:   
## bweight ~ mbsmoke + mmarried + (mbsmoke \* mmarried) + I(mmarried \* mmarried) + prenatal1 + (mbsmoke \* prenatal1) + I(prenatal1 \* prenatal1) + fbaby + (mbsmoke \* fbaby) + I(fbaby \* fbaby) + medu + (mbsmoke \* medu) + I(medu \* medu)   
##   
## Default formula for propensity models:   
## mbsmoke ~ mmarried + I(mmarried \* mmarried) + prenatal1 + I(prenatal1 \* prenatal1) + fbaby + I(fbaby \* fbaby) + medu + I(medu \* medu)

## Fit standardization model

Purpose: Fit a standardization model to the dataset data. data: The dataset containing your variables of interest. n.boot = 100: Specifies 100 bootstrap samples to calculate standard errors. simple = TRUE: Indicates that a simplified version of the model should be fitted.

# Fit standardization model  
std\_model <- standardization(data, n.boot = 100, simple = T)  
std\_model

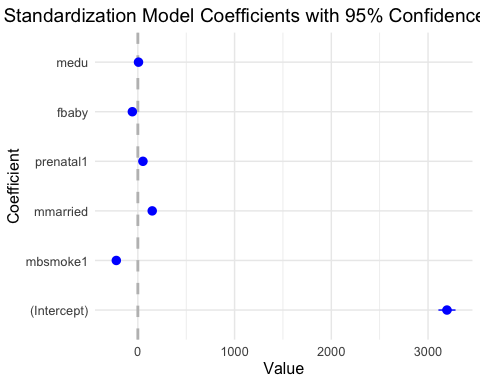
##   
## Call: glm(formula = bweight ~ mbsmoke + mmarried + prenatal1 + fbaby +   
## medu, family = family, data = data)  
##   
## Coefficients:  
## (Intercept) mbsmoke1 mmarried prenatal1 fbaby medu   
## 3195.673 -222.610 148.512 52.567 -56.873 6.802   
##   
## Degrees of Freedom: 4641 Total (i.e. Null); 4636 Residual  
## (9284 observations deleted due to missingness)  
## Null Deviance: 1.555e+09   
## Residual Deviance: 1.464e+09 AIC: 71960  
##   
## Average treatment effect of mbsmoke:  
## Estimate - -222.6099   
## SE - 20.56428   
## 95% CI - ( -262.9152 , -182.3047 )

## Using the coefplot package to visualize the coefficients along with their 95% confidence intervals:

**Explanation**:

1. **Load Packages**: Ensure both **coefplot** and **broom** are installed and loaded.
2. **Creating Tidy Data Frame**: Using broom::tidy() on your model object keeps it clean, but now, it isn’t explicitly required by coefplot as coefplot can directly use the model object. For an introduction to broom, see <https://cran.r-project.org/web/packages/broom/vignettes/broom.html>
3. **Plotting with coefplot**:
   * Directly pass the model object to **coefplot**.
   * intercept=TRUE to include the intercept term in the plot.
   * outerCI=1.96 to define 95% confidence intervals.

# Install and load necessary packages  
#install.packages("coefplot")  
#install.packages("broom")  
library(coefplot)  
library(broom)  
  
# Create a tidy data frame for the model coefficients  
coef\_df <- tidy(std\_model$model)  
  
# Use coefplot for visualization  
coefplot(std\_model$model, intercept=TRUE, outerCI=1.96) +  
 ggtitle("Standardization Model Coefficients with 95% Confidence Intervals") +  
 theme\_minimal() +  
 theme(  
 text = element\_text(size = 12),  
 plot.title = element\_text(hjust = 0.5)  
 )



## Print Summary of Average Treatment Effect (ATE)

Print the summary of the Average Treatment Effect (ATE) estimated by the model.

print(std\_model$ATE.summary)

## Beta SE 2.5 % 97.5 %  
## 1 -222.6099 20.56428 -262.9152 -182.3047

### Session Info

sessionInfo()

## R version 4.4.1 (2024-06-14)  
## Platform: aarch64-apple-darwin20  
## Running under: macOS 15.0.1  
##   
## Matrix products: default  
## BLAS: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib   
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib; LAPACK version 3.12.0  
##   
## locale:  
## [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8  
##   
## time zone: America/Los\_Angeles  
## tzcode source: internal  
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] broom\_1.0.6 coefplot\_1.2.8 ggplot2\_3.5.1 CausalModels\_0.2.0  
## [5] haven\_2.5.4 dplyr\_1.1.4   
##   
## loaded via a namespace (and not attached):  
## [1] sandwich\_3.1-1 utf8\_1.2.4 generics\_0.1.3 tidyr\_1.3.1   
## [5] stringi\_1.8.4 lattice\_0.22-6 hms\_1.1.3 digest\_0.6.37   
## [9] magrittr\_2.0.3 evaluate\_0.24.0 grid\_4.4.1 useful\_1.2.6.1   
## [13] mvtnorm\_1.3-1 fastmap\_1.2.0 plyr\_1.8.9 Matrix\_1.7-0   
## [17] backports\_1.5.0 survival\_3.7-0 causaldata\_0.1.3 multcomp\_1.4-26   
## [21] purrr\_1.0.2 fansi\_1.0.6 scales\_1.3.0 TH.data\_1.1-2   
## [25] codetools\_0.2-20 cli\_3.6.3 rlang\_1.1.4 munsell\_0.5.1   
## [29] splines\_4.4.1 geepack\_1.3.11.1 withr\_3.0.1 yaml\_2.3.10   
## [33] tools\_4.4.1 reshape2\_1.4.4 tzdb\_0.4.0 colorspace\_2.1-1   
## [37] forcats\_1.0.0 boot\_1.3-31 vctrs\_0.6.5 R6\_2.5.1   
## [41] zoo\_1.8-12 lifecycle\_1.0.4 stringr\_1.5.1 MASS\_7.3-61   
## [45] pkgconfig\_2.0.3 pillar\_1.9.0 gtable\_0.3.5 Rcpp\_1.0.13   
## [49] glue\_1.7.0 highr\_0.11 xfun\_0.47 tibble\_3.2.1   
## [53] tidyselect\_1.2.1 rstudioapi\_0.16.0 knitr\_1.48 farver\_2.1.2   
## [57] htmltools\_0.5.8.1 labeling\_0.4.3 rmarkdown\_2.28 readr\_2.1.5   
## [61] compiler\_4.4.1