A Tutorial on Propensity Score Analysis with Semi-Continuous Treatment

Huibin Zhang[a] Ph.D. , Walter Leite[b] Ph.D.

[a]University of Tennessee, Knoxville
[b] University of Florida

Learning objectives

- Understand the benefits of performing propensity score analysis
- Understand semi-continuous treatment exposure
- Understand how to implement a propensity score analysis with a semi-continuous treatment exposure in R

Advantages of Propensity Score

- Remove selection bias due to a large number of covariates (Guo et al., 2020)
- Do not require modeling the functional form of the relationship between the outcome and covariates.
- Robust to overfitting issues (Rosenbaum & Rubin, 1983, 1984)

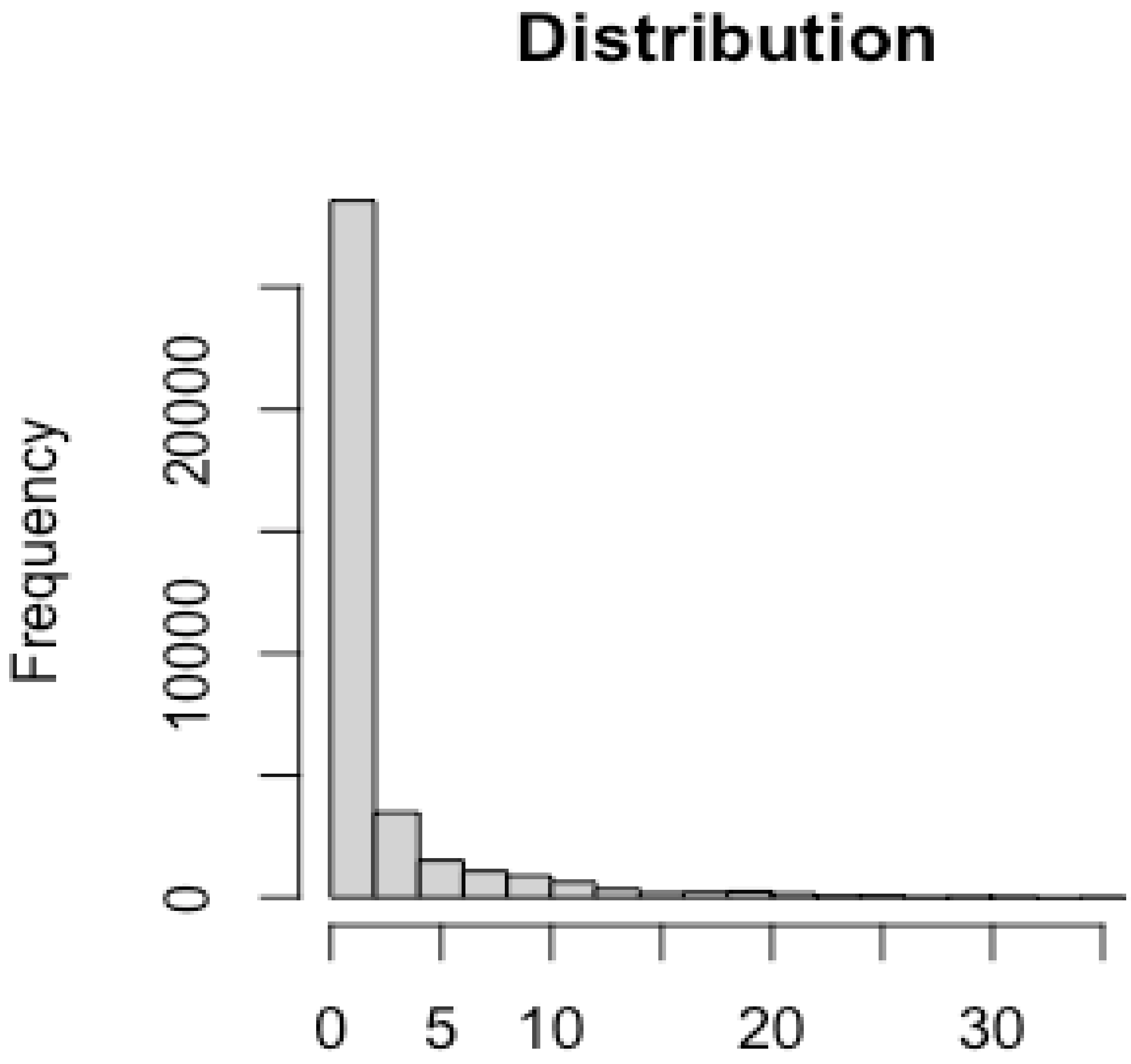
What is semicontinuous treatment exposure

- For a treatment available at different treatment doses, a large proportion of potential users have zero exposure.

Example:

- Time watching videos in online platform (Berry, 2017)
- Number of logins and number of completed quizzes for Algebra Nation (Leite et al., 2022)

Example of semi-continuous exposure: The Math Nation virtual learning environment



Video Views in Math Nation

Theoretical background

Propensity Score

The propensity score is also a balancing score, making covariates distributions between treatment and control groups similar to a randomized experiment to reduce selection bias (Austin, 2011).

- Binary treatment： $P(T = 1|X)$ (Rosenbaum & Rubin, 1983)
- Continuous treatment：
  - $r(T_i, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{\sqrt{2\sigma^2}}(T_i - \beta_0 - \beta X_i)^2\}$ (Hirano & Imbens, 200
- Semi-continuous treatment：
  - $smGPS^1 = E(T|X;\beta) = \pi(X;\beta_1)\,\varpi(X;\beta_2)$ (Hocagil et al, 2021)

Semi-Continuous Outcome Models

- When to use?
  - Regular count analysis cannot recover true parameter due to zero inflation (Young et al., 2020)

- Common distributions of nonzero part ：
  - Poisson distribution (Zelterman, 2006)
  - Negative-Binomial distribution (Agresti, 2002)

- Overdispersion:
  - variance of data bigger than mean of data (Yau et al., 2003)

Zero-Inflated Model

Lambert (1992) proposes that zeros in the semi-continuous data come from two latent classes-"structural zeros" and "sampling zeros".

Using number of logins to Math Nation, Structural zeros indicate students never used the system, while sampling zeros indicate students did use the system but the system did not count it due to internet problems during the observation period.

Zero-Inflated Models

The general equation of the Zero-Inflated model is defined as (Cameron & Trivedi, 2013):

$$\mathrm{p}(Y=y)=\begin{cases}\pi+(1-\pi)p(y=0;\mu) & y=0,\\(1-\pi)p(y;\mu) & y>0\end{cases}$$

$$\mathrm{In}(\mu)=\gamma\mathrm{X}, logit(\pi)=\beta X$$

$$\boldsymbol{smGPS_{ZIP}}=[1-(\pi|X)](\mu|\mathrm{X})$$

$$(\pi|X)=\frac{e^{\beta X}}{1+e^{\beta X}},\ (\mu|\mathrm{X})=\mathrm{e}^{\gamma\mathrm{X}}$$

$\boldsymbol{smGPS_{ZINB}}=[1-(\pi|X)](\mu|\mathrm{X})$ (Cameron & Trivedi, 2013)

$$(\pi|X)=\frac{e^{\beta X}}{1+e^{\beta X}},\ (\mu|\mathrm{X})=\mathrm{e}^{\gamma\mathrm{X}}$$

Illustration with real data

The real data comes from Math Nation (Leite et al., 2022). The number of observations is 37,550.

The semi-continuous treatment: number of recommended videos watched by students in each section.

There are 33 covariates, 25 of which are dummy variables. The outcome is 10-question quiz for that section.

Analysis Flowchart

1. Set Up R Environment
2. Propensity Score Estimation
3. Covariate Balance Check
4. ATE Estimation

Set Up R Environment

- Install countreg package to run zero-inflated Model, which is not available in Cran:
  - install.packages("countreg", repos="http://R-Forge.R-project.org")

- Install AER package, to check if there is overdispersion
  - install.packages(”AER”)

Propensity Score Estimation

Run dispersion test to select appropriate distribution for zero-inflated model:

- library(AER)
- rd <- glm(followed ~ ., data = newdata, family = poisson)
- dispersiontest(rd,trafo=1)

```
> dispersiontest(rd,trafo=1)# above 1

        Overdispersion test

data:  rd
z = 49.011, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
4.457496
```

Generalized Propensity Score Estimation with Zero-Inflated Negative Binomial Model

```
Library(countreg)
tryzi=zeroinfl(followed ~ pretest + yearsteaching + yearsAN + ANTotalTime +
minority + lowses + tquestion + mengagement  +coursetype_1+ clusterid_1
+clusterid_2 + clusterid_3 +
clusterid_4 + clusterid_5 + clusterid_6 + clusterid_7 + clusterid_8 +
clusterid_9 + clusterid_10 + clusterid_11 + clusterid_12 +    clusterid_13 +
clusterid_14 + clusterid_15 + clusterid_16 +  clusterid_17 + clusterid_18 +
clusterid_19  + districtname_1 + districtname_2 + sectionid_1 + sectionid_2 +
sectionid_3 , data=newdata, dist = "negbin", link="logit")

outzi=summary(tryzi)
zczi=outzi$coefficients$zero[,1] # parameters for binary part
nzczi=outzi$coefficients$count[,1] # parameters for nonzero part
```

Propensity Score Model

Parameters for binary part β

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.161e+01	4.151e-01	27.978	< 2e-16	***
pretest	8.988e-02	1.655e-02	5.432	5.57e-08	***
yearsteaching	-6.344e-02	2.790e-03	-22.741	< 2e-16	***
yearsAN	1.117e-02	6.408e-03	1.743	0.081291	.
ANTotalTime	5.809e-04	7.106e-05	8.174	2.99e-16	***
minority	-1.305e-02	1.956e-03	-6.673	2.51e-11	***
lowses	1.359e-02	1.734e-03	7.840	4.50e-15	***
tquestion	-5.066e-02	1.681e-03	-30.128	< 2e-16	***
mengagement	-2.970e+00	1.311e-01	-22.655	< 2e-16	***
coursetype_1	1.188e+00	5.783e-02	20.544	< 2e-16	***
clusterid_1	6.732e-02	9.638e-02	0.699	0.484850	
clusterid_2	3.769e-01	1.059e-01	3.560	0.000371	***
clusterid_3	9.812e-01	1.075e-01	9.124	< 2e-16	***
clusterid_4	5.786e-01	1.030e-01	5.619	1.92e-08	***
clusterid_5	-1.867e-02	8.977e-02	-0.208	0.835227	
clusterid_6	9.411e-01	1.052e-01	8.946	< 2e-16	***
clusterid_7	6.637e-02	8.994e-02	0.738	0.460573	
clusterid_8	-2.159e-01	8.961e-02	-2.410	0.015962	*
clusterid_9	1.385e-01	8.139e-02	1.701	0.088893	.
clusterid_10	-4.724e-01	8.899e-02	-5.308	1.11e-07	***
clusterid_11	1.294e-01	7.554e-02	1.713	0.086634	.
clusterid_12	9.991e-03	7.173e-02	0.139	0.889216	
clusterid_13	8.577e-01	7.599e-02	11.287	< 2e-16	***
clusterid_14	4.769e-01	7.012e-02	6.801	1.04e-11	***
clusterid_15	1.437e-01	7.022e-02	2.047	0.040676	*
clusterid_16	2.083e-01	7.202e-02	2.892	0.003834	**
clusterid_17	2.076e-01	7.393e-02	2.808	0.004986	**
clusterid_18	1.034e-02	7.508e-02	0.138	0.890481	
clusterid_19	-1.673e-01	7.174e-02	-2.332	0.019698	*
districtname_1	1.303e+00	5.255e-02	24.798	< 2e-16	***
districtname_2	-8.830e-01	1.034e-01	-8.539	< 2e-16	***
sectionid_1	-8.614e-01	4.342e-02	-19.841	< 2e-16	***
sectionid_2	-8.951e-01	3.920e-02	-22.830	< 2e-16	***
sectionid_3	2.603e-01	3.656e-02	7.121	1.07e-12	***

Propensity Score Model

Parameters for nonzero part γ

```
Count model coefficients (negbin with log link):
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.664e+00  2.463e-01  -6.756 1.42e-11 ***
pretest        -3.852e-02  8.262e-03  -4.662 3.13e-06 ***
yearsteaching   1.975e-02  1.439e-03  13.724  < 2e-16 ***
yearsAN        -2.540e-02  3.305e-03  -7.684 1.54e-14 ***
ANTotalTime    -4.268e-04  5.094e-05  -8.378  < 2e-16 ***
minority       -4.840e-03  1.135e-03  -4.263 2.02e-05 ***
lowses          7.967e-03  9.609e-04   8.291  < 2e-16 ***
tquestion       3.655e-02  9.137e-04  40.006  < 2e-16 ***
mengagement     7.933e-01  7.737e-02  10.253  < 2e-16 ***
coursetype_1   -1.068e-01  4.274e-02  -2.500 0.012435 *
clusterid_1    -3.462e-02  5.535e-02  -0.625 0.531729
clusterid_2    -2.535e-01  6.225e-02  -4.073 4.65e-05 ***
clusterid_3     1.903e-01  6.319e-02   3.012 0.002597 **
clusterid_4    -2.216e-02  6.084e-02  -0.364 0.715613
clusterid_5    -1.020e-01  4.905e-02  -2.080 0.037536 *
clusterid_6     7.359e-02  5.958e-02   1.235 0.216794
clusterid_7     1.018e-01  5.007e-02   2.034 0.041932 *
clusterid_8     5.544e-02  4.339e-02   1.278 0.201411
clusterid_9     1.111e-01  4.332e-02   2.566 0.010291 *
clusterid_10   -1.805e-02  4.473e-02  -0.404 0.686483
clusterid_11    3.276e-02  3.958e-02   0.828 0.407728
clusterid_12    1.258e-01  3.715e-02   3.386 0.000708 ***
clusterid_13    6.059e-02  4.104e-02   1.476 0.139840
clusterid_14   -1.396e-02  3.689e-02  -0.378 0.705097
clusterid_15   -1.664e-01  3.606e-02  -4.616 3.92e-06 ***
clusterid_16    9.848e-02  3.595e-02   2.739 0.006159 **
clusterid_17   -1.422e-01  3.791e-02  -3.750 0.000177 ***
clusterid_18   -7.332e-02  3.781e-02  -1.939 0.052445 .
clusterid_19   -1.501e-02  3.625e-02  -0.414 0.678836
districtname_1 -1.888e-01  3.359e-02  -5.620 1.91e-08 ***
districtname_2 -4.839e-01  4.598e-02 -10.526  < 2e-16 ***
sectionid_1    -3.704e-01  2.327e-02 -15.919  < 2e-16 ***
sectionid_2    -3.631e-01  2.036e-02 -17.840  < 2e-16 ***
sectionid_3     6.601e-02  1.776e-02   3.718 0.000201 ***
```

Calculation of Semi-Generalized Propensity Score (Semi-GPS)

pnzi=exp(zczi[1]+ as.matrix(newdata1[,c(2:9, 12:36)]) %*% zczi[2:34]) # this is $e^{\beta X}$

pzi=exp(nzczi[1]+ as.matrix(newdata1[,c(2:9, 12:36)]) %*% nzczi[2:34]) # this is $e^{\gamma X}$

GPSzi=pzi/(1+pnzi) # this is **Semi-GPS**

summary(GPSzi)

Min	Median	Mean	Max
0.0044	1.4660	2.025	17.7521

Covariate Balance Check

Fit one regression for each covariate with semi-continuous treatment as the outcome and Semi-GPS and the covariate as predictors.

Standardized regression coefficients can be used as a measure of the effect sizes of the covariates on treatment (Leite, 2017).

Covariate balance is achieved if the effect sizes of the covariates are smaller than 0.05 (WWC, 2022)

Covariate Balance with R

```
covariatesname=names(newdata[, c(2:9, 12:36)])
balancetable=data.frame()
for (var in 1:length(covariatesname)){
  balformula=paste("followed~GPSzi+", covariatesname[var], sep="")
  maxeff=max(abs(coef(lm(balformula, newdata))[-(1:2)]))
  balancetable=rbind(balancetable, c(var, maxeff))
}
names(balancetable)=c("variable", "coef")
balancetable$variable=covariatesname
balancetable$coef=balancetable$coef/sd(newdata$followed)
```

Covariate Balance Check Result

27 covariates' standard coefficients are smaller than 0.05, but 6 covariates' standard coefficients are larger than 0.05 (See table)

variable	mengagement	coursetype_1	clusterid_1	clusterid_7	districtname_1	districtname_2
coef	0.0790	0.0501	0.0984	0.0600	0.0564	0.0665

ATE Estimation with R

- **Get covariates list which are not balanced from last step**

  which(balancetable$coef>0.05)

- **Add those unbalanced covariates into the outcome model to get ATE**

  outzinb=glm(**posttest~followed+GPSzi+mengagement+coursetype_1+clusterid_1 + clusterid_7+districtname_1 + districtname_2**, data=newdata)

  resultzinb=summary(outzinb)

  resultzinb$coefficients[2,]

ATE Estimate

Estimate	Std. Error	t value	Pr(>|t|)
7.5638*e-03	1.2103*e-03	6.2494	4.1656*e-09

Each recommended video watched could increase 0.0076 standard deviation unit of the quiz score

Link to example code and data

bit.ly/semi-GPS

Contact Information:
Dr. Huibin Zhang
hzhan117@utk.edu

Take aways

Only one published study has investigated how to apply PSA with semi-continuous treatment (Hocagil et al, 2021).

However, that paper is a simulation study paper, which does not provide a guideline for potential users to apply PSA.

There is a need for more studies to inform applied researchers on how to use the semi-GPS

Key References

- Hocagil, T., Cook, R. J., Jacobson, S. W., Jacobson, J. L., & Ryan, L. M. (2021). Propensity score analysis for a semi-continuous exposure variable: a study of gestational alcohol exposure and childhood cognition. *Journal of the Royal Statistical Society*. 184 (4), p.1390–1413, https://doi.org/10.1111/rssa.12716

- Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14

- Leite, W. L. (2017). Practical propensity score methods using R. Sage Publishing.