Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #19 (ruptures & discontinuités)

Maths of Causal Inference

n individuals are either treated ($t_i = 1$) or not ($t_i = 0$).
We observe outcome y_i for covariates $\mathbf{x}_i$.
We want to study potential outcomes $y_i(1)$ and $y_i(0)$

	turnout				
	$y_i(1)$	$y_i(0)$	t_i	$x_{1,i}$	$x_{2,i}$
1	y_1	?	1	$x_{1,1}$	$x_{2,1}$
2	?	y_2	0	$x_{1,2}$	$x_{2,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	?	1	$x_{1,n}$	$x_{2,n}$

The causal effect is $y_i(1) - y_i(0)$

Maths of Causal Inference

- Average Treatment Effect $\mathbb{E}\Big[Y(1) - Y(0)\Big]$
- Sample Average Treatment Effect $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\Big[y_i(1) - y_i(0)\Big]$

Assumption : $(Y(1), Y(0)) \perp\!\!\!\perp T$

Crude estimator (difference in means), $\displaystyle\widehat{\tau} = \frac{1}{n_1}\sum_{i:t_i=1} y_i - \frac{1}{n_0}\sum_{i:t_i=0} y_i$,

or

$$\widehat{\tau} = \sum_{i=1}^{n}\frac{t_i y_i}{n_1} - \frac{(1-t_i)y_i}{n_0}$$

Then $\mathbb{E}\Big[\widehat{\tau}\Big] = \mathbb{E}\Big[Y(1) - Y(0)\Big]$

- Local Average Treatment Effect $\mathbb{E}\big[Y(1) - Y(0)|\mathbf{X} = \mathbf{x}\big]$

The Propensity Score is the probability to receive the treatment $\pi(\mathbf{x}) = \mathbb{P}[T = 1|\mathbf{X} = \mathbf{x}]$

Assumption: Balancing property $T \perp\!\!\!\perp \mathbf{X} \mid \pi(\mathbf{X})$

Assumption: Exogeneity $(Y(1), Y(0)) \perp\!\!\!\perp T \mid \pi(\mathbf{x}), \forall\mathbf{x}$

Consider here

$$\widehat{\tau} = \sum_{i=1}^{n} \frac{t_i y_i}{n\widehat{\pi}(\mathbf{x}_i)} - \frac{(1 - t_i)y_i}{n(1 - \widehat{\pi}(\mathbf{x}_i))}$$

Discontinuity

Introduced in Regression-discontinuity analysis: An alternative to the ex post facto experiment , to quantify the effects of college scholarships on later students' achievements

- X is SAT score
- Binary treatment T,

  receipt of scholarship,

$$T_i = \mathbf{1}(X_i \geq c) = \begin{cases} 1 \text{ if } X_i \geq c \\ 1 \text{ if } X_i < c \end{cases}$$

- Outcome Y

  (e.g. subsequent earnings)

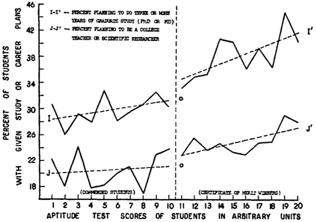


Fig. 3. Regression of study and career plans on exposure determiner.

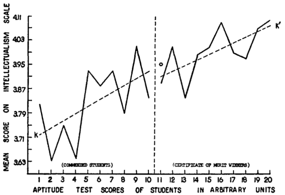
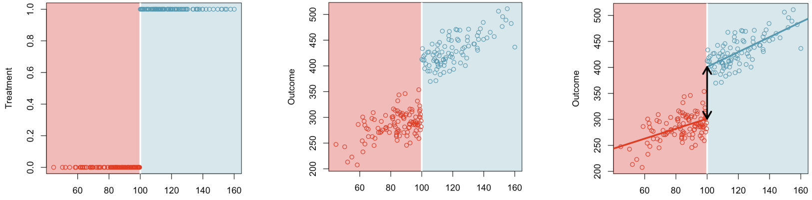


Fig. 4. Regression of attitudes toward intellectualism on exposure determiner.

Discontinuity

D_i and Y_i against X_i, and two regressions,

$$Y_i = \begin{cases} \alpha^+ + \beta^+ X_i \text{ if } X_i \geq c \text{ (i.e. } D_i = 1) \\ \alpha^- + \beta^- X_i \text{ if } X_i < c \text{ (i.e. } D_i = 0) \end{cases}$$



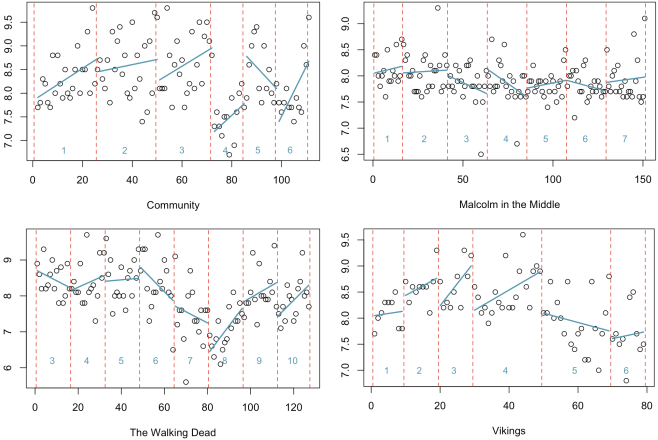
Here the Local Average Treatment Effect is

$$\mathbb{E}\Big[Y(1) - Y(0)|X = c\Big] = (\alpha^+ - \alpha^-) + (\beta^+ - \beta^-) \cdot c$$

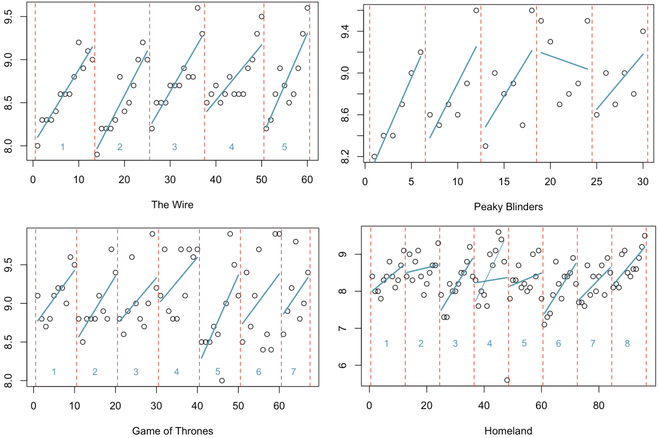
Why only linear regression ? See Regression Discontinuity Designs

TV shows

```
> download.file("https://github.com/nazareno/imdb-series/raw/master/data/series_from_imdb.csv",
> destfile="series_from_imdb.csv")
> base = read.csv("series_from_imdb.csv")
```



TV shows




```
> sbase = base[base$series_name=="Game of Thrones",]
> sbase12 = sbase[sbase$season%in%c(1,2),]
> seuil = sbase12$series_ep[which(diff(sbase12$season)
  !=0)]+.5
```

TV shows

```
> s = function(x) (x-seuil)*(x >seuil)
> reg = lm(UserRating~series_ep+s(series_ep)+I(
    series_ep>seuil), data=sbase12)
> library(rdd)
> sbase12 = sbase[sbase$season%in%c(1,2),]
> lmr = RDestimate(UserRating~series_ep, data=sbase12,
     cutpoint=mean(range(sbase12$series_ep)))
```