Regression Discontinuity Design (RDD)

Pratiques de la Recherche en Économie

Florentine Oliveira
2025-04-11

1. Intuition et définitions

De nombreux traitements sont définis selon une **règle/un seuil**.

Par exemple:

- le revenu à partir duquel un individu peut bénéficier d'une prestation sociale
- l'âge d'entrée à l'école, âge légal pour avoir droit de vote, consommer de l'alcool (majorité)
- moyenne au bac requise pour pouvoir candidater à certaines écoles

La régression sur discontinuité, ou ***Regression Discontinuity Design***, exploite ce(s) seuil(s) pour estimer l'effet causal du traitement.

Intuition: **exogeneité locale**

- les individus proches du seuil sont raisonnablement comparables
- cependant ceux au-dessus du seuil sont traîtés alors que ceux en dessous ne le sont pas
- la discontinuité crée une **quasi-expérience** au voisinage du seuil de discontinuité
    - autour du seuil, l'allocation au traitement est *as good as random*

1. Intuition & définitions

Formellement

Si l'on revient au framework des outcomes potentiels:

$$Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i)$$

Maintenant,

$$D_i = 1\{X_i \geq c\}$$

où X_i est appelée ***running/forcing variable***, c'est à dire la variable (*continue*) sur laquelle s'applique le critère de sélection dans le traitement.

i.e. la probabilité pour l'individu i d'être traîté passe de 0 à 1 au seuil de discontinuité c

Deux types de régressions sur discontinuité:

- **sharp**: la probabilité de traitement devient certaine au seuil c (ex: la consommation légale d'alcool (D_i) à partir d'un certain âge (X_i))
- **fuzzy**: la probabilité de traitement augmente au seuil c mais ne passe pas nécessairement à 1 (ex:)

1. Intuition & définitions

Bandwith: intervalle autour du seuil de discontinuité dans lequel on conserve les observations pour estimer l'effet du traitement.

$\Longrightarrow$ Arbitrage:

- un bandwith trop étroit peut limiter le nombre d'observations
- un bandwith trop large peut inclure des observations moins comparables

Forme fonctionnelle: désigne la spécification de la relation entre l'outcome et la forcing variable dans le modèle de régression. Elle peut être linéaire, polynomiale, etc.

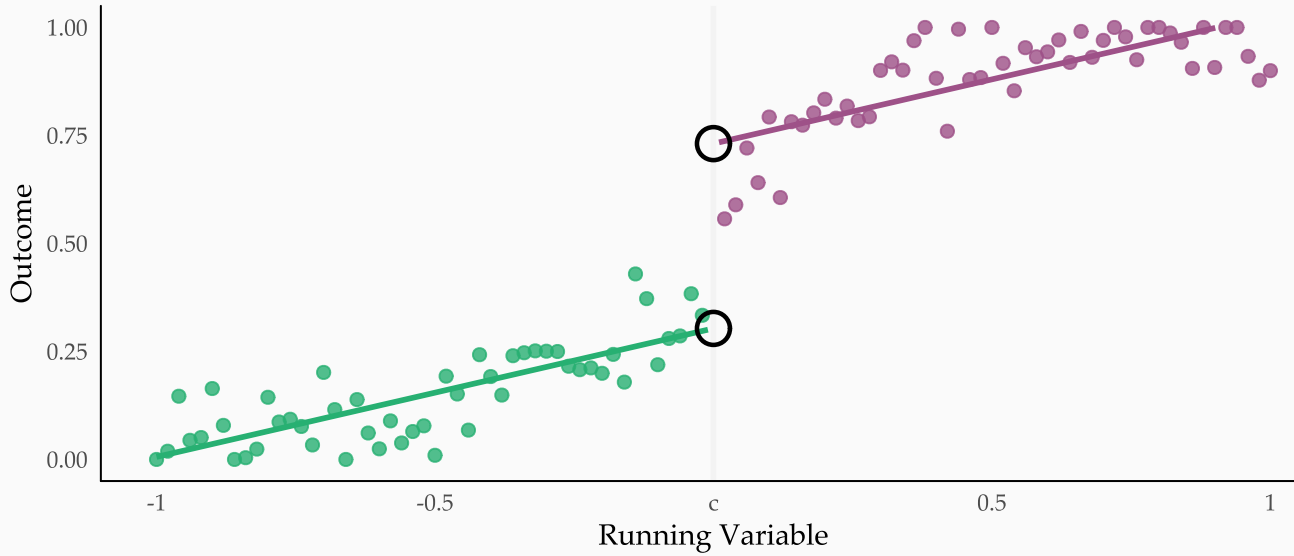
2. Sharp RDD

2. Sharp RDD

2.1. Définition de l'estimateur

L'estimateur de l'effet causal du traitement D_i sur Y_i revient alors à comparer la moyenne de l'outcome Y_i de part et d'autre du seuil c:

$$\begin{aligned}\beta_{\text{RDD}}^{\text{sharp}} &= \lim_{x \to c^+} \mathbb{E}(Y_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_i | X_i = x) \\ &= \lim_{x \to c^+} \mathbb{E}(Y_{1i} | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_{0i} | X_i = x)\end{aligned}$$

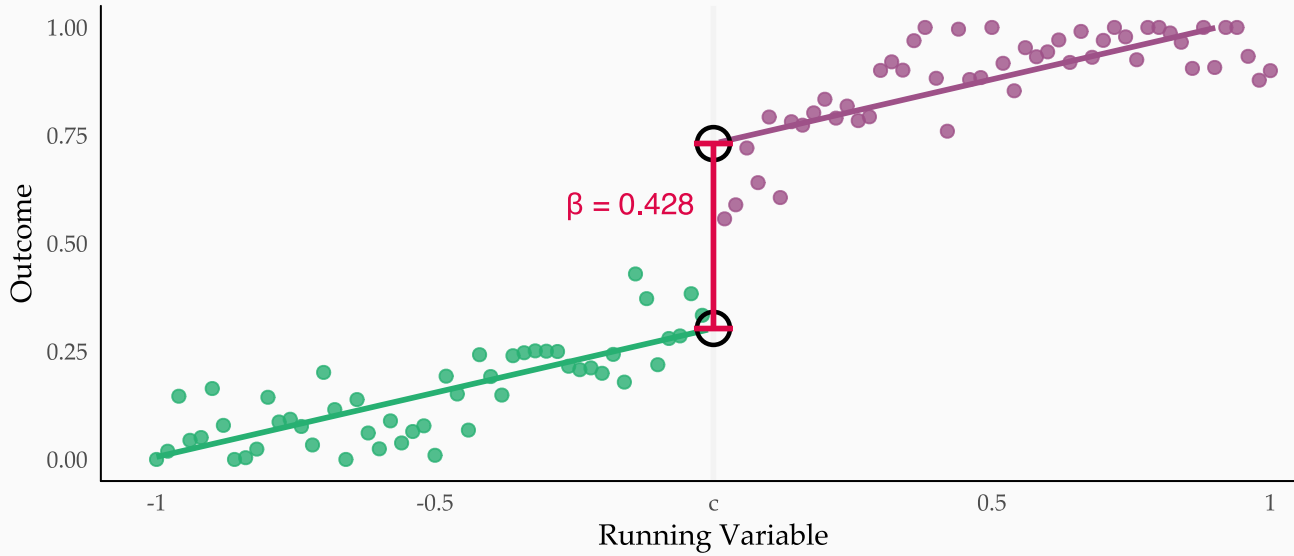


2. Sharp RDD

2.1. Définition de l'estimateur

L'estimateur de l'effet causal du traitement D_i sur Y_i revient alors à comparer la moyenne de l'outcome Y_i de part et d'autre du seuil c:

$$\begin{aligned}\beta_{\text{RDD}}^{\text{sharp}} &= \lim_{x \to c^+} \mathbb{E}(Y_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_i | X_i = x) \\ &= \lim_{x \to c^+} \mathbb{E}(Y_{1i} | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_{0i} | X_i = x)\end{aligned}$$



2. Sharp RDD

2.2. Hypothèse d'identification

Hypothèse d'identification: $\mathbb{E}(Y_{1i}|X_i = x)$ **et** $\mathbb{E}(Y_{0i}|X_i = x)$ **sont continues en** x

Donc. $\beta_{\text{RDD}}^{\text{sharp}} = \mathbb{E}(Y_i|X_i = c) - \mathbb{E}(Y_i|X_i = c) = \mathbb{E}(Y_{1i} - Y_{0i}|X_i = c)$

L'estimateur $\beta_{\text{RDD}}^{\text{sharp}}$ est un estimateur local de l'effet moyen du traitement (***LATE***).

NB: on n'a pas fait l'hypothèse ici d'assignation aléatoire du traitement D_i (et donc X_i).

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Motivation:

- Peu d'évidence sur les effets du congé maternité rémunéré sur la santé mentale des mères... alors même que c'est la première justification de l'existence de ce congé
- Effet ambigu:
  - Effets positifs si l'emploi augmente le stress ou réduit le temps que la femme consacre à s'occuper d'elle-même et à se remettre des effets physiques de l'accouchement
  - Effets négatifs si cela permet à la mère d'avoir davantage d'interactions sociales et cela augmente le revenu

Contexte:

- Introduction d'un congé payé maternité en Norvège, au 1er Juillet 1977
  - avant la réforme, aucun congé rémunéré; seulement 12 semaines de congé non rémunérés
  - après la réforme: 4 mois de congés payés et à 12 mois de congés non rémunérés

Question de recherche: quel effet l'introduction du congé payé a t-elle eu sur la santé mentale des mères ?

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Question: pourquoi ne peut-on pas simplement comparer la santé mentale moyenne des mères ayant recours au congé maternité et celle des mères n'y ayant pas recours?

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Question: pourquoi ne peut-on pas simplement comparer la santé mentale moyenne des mères ayant recours au congé maternité et celle des mères n'y ayant pas recours?

Endogeneité

- certaines caractéristiques inobservables peuvent affecter à la fois le recours au congé maternité et la santé mentale des mères
- causalité inversée: si la santé mentale des mères impacte le recours au congé maternité

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Stratégie d'identification: Regression Discontinuity Design!!

Question: quelle est la running variable? quel est le seuil?

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Stratégie d'identification: Regression Discontinuity Design!!

Question: quelle est la running variable? quel est le seuil?

- Date de naissance de l'enfant
- 1er juillet 1977

Les mères ayant accouché de leur enfant **avant le 1er juillet 1977** ne bénéficient pas de l'introduction du congé maternité rémunéré. Celles ayant accouché **le 1er juillet 1977 ou après** peuvent en bénéficier.

2. Sharp RDD

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Stratégie d'identification: Regression Discontinuity Design!!

Question: quelles sont les hypothèses d'identification?

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

Stratégie d'identification: Regression Discontinuity Design!!

Question: quelles sont les hypothèses d'identification?

- **Pas de manipulation de la running variable**: ici, cela signifie que les mères ne peuvent pas stratégiquement choisir d'accoucher avant ou après le 1er juillet 1977
- **Continuité de la running variable au point de discontinuité**: ici, cela signifie que le nombre de naissances est continu au point de discontinuité

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy

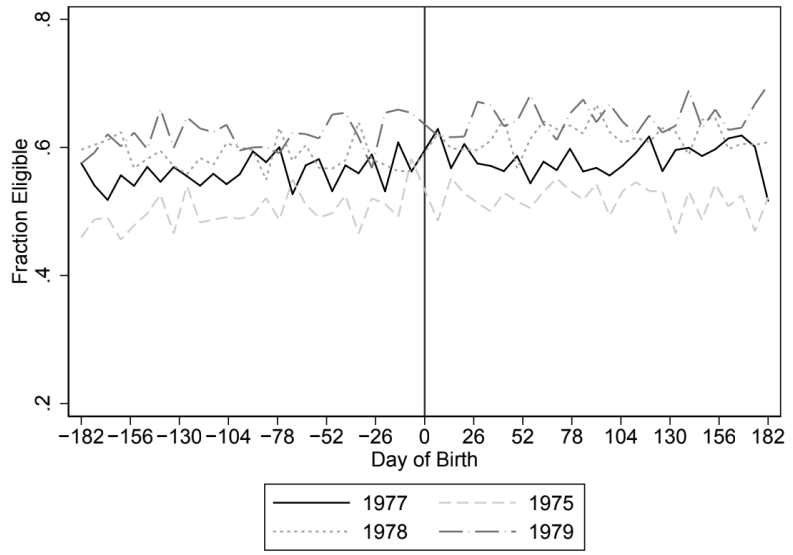


Figure 1. Proportion of Mothers Eligible for Paid Maternity Leave

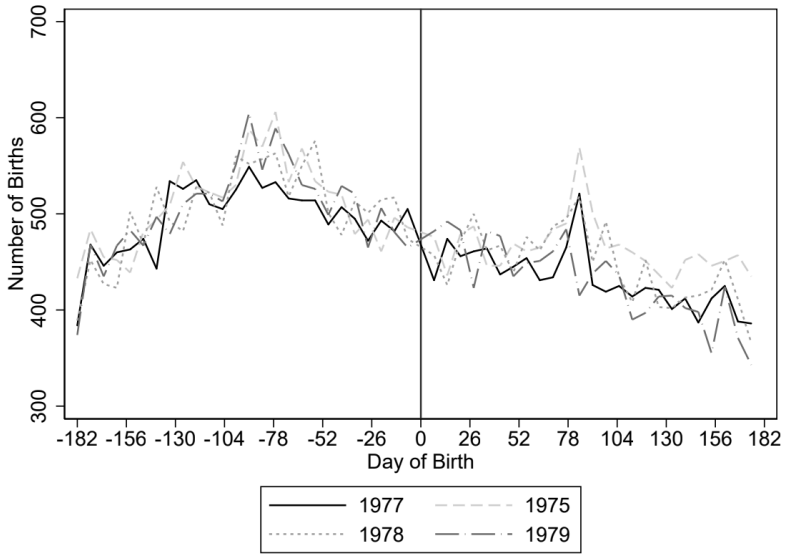
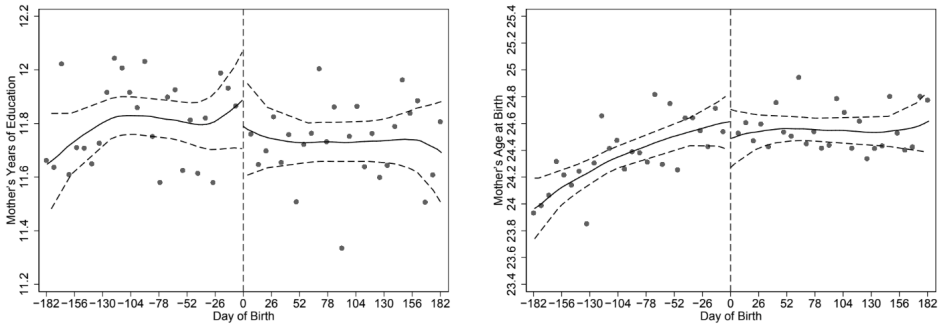


Figure 2. Number of Children Born to Eligible Mothers

2.3. Exemples

Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy



(a) Years of Education

(b) Mother's Age at Birth

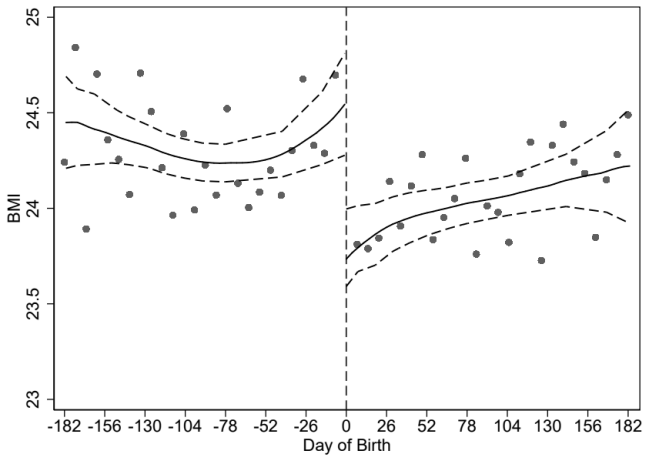
(c) Mother's Income in 1975

(d) Married at Birth (Dummy)

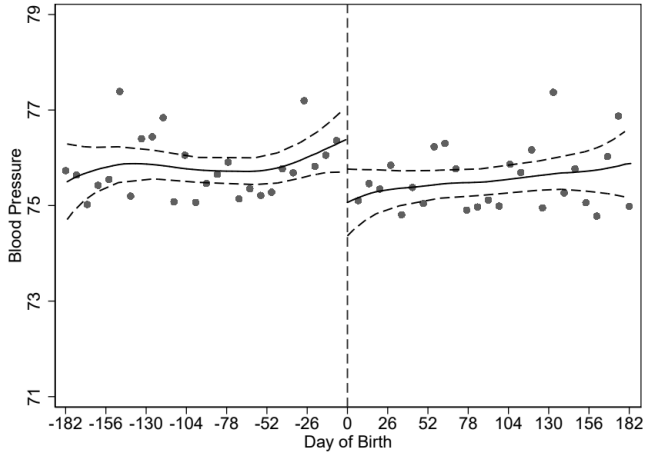
FIGURE 3. MOTHERS' PRE-REFORM CHARACTERISTICS

2.3. Exemples

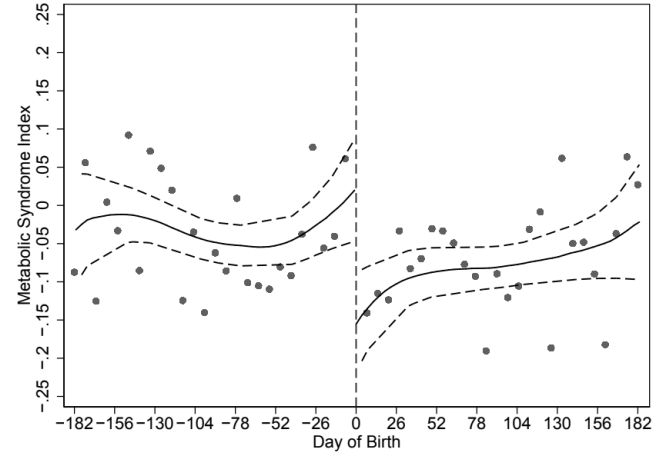
Bütikofer, A., Riise, J., & M. Skira, M. (2021). *The impact of paid maternity leave on maternal health*. AEJ: Economic Policy



(a) Body Mass Index



(d) Blood Pressure (Diastolic)



(g) Metabolic Syndrome Index

2.3. Exemples

Canaan, S. (2022) . *Parental leave, household specialization and children's well-being*. Labour Economics.

- Impact de l'allongement de la durée du congé parental en France en 1994: allocation mensuelle pouvant aller jusqu'aux trois ans de l'enfant
  - avant réforme, parents éligibles à partir du 3ème enfant
  - après la réforme, parents éligibles à partir du 2ème enfant
  - $\simeq$ augmentation de la durée du congé parental de 3 ans
- Effets :
  - Négatifs sur l'emploi des mères: éloignées plus longtemps du marché du travail et baisse du salaire
  - Négatifs sur la spécialisation des tâches au sein du ménage: les hommes ne prennent pas ce congé parental et travaillent davantage
  - Négatifs sur le développement verbal des enfants

2.4. Estimation sur R

Étape 1: centrer la *running variable*

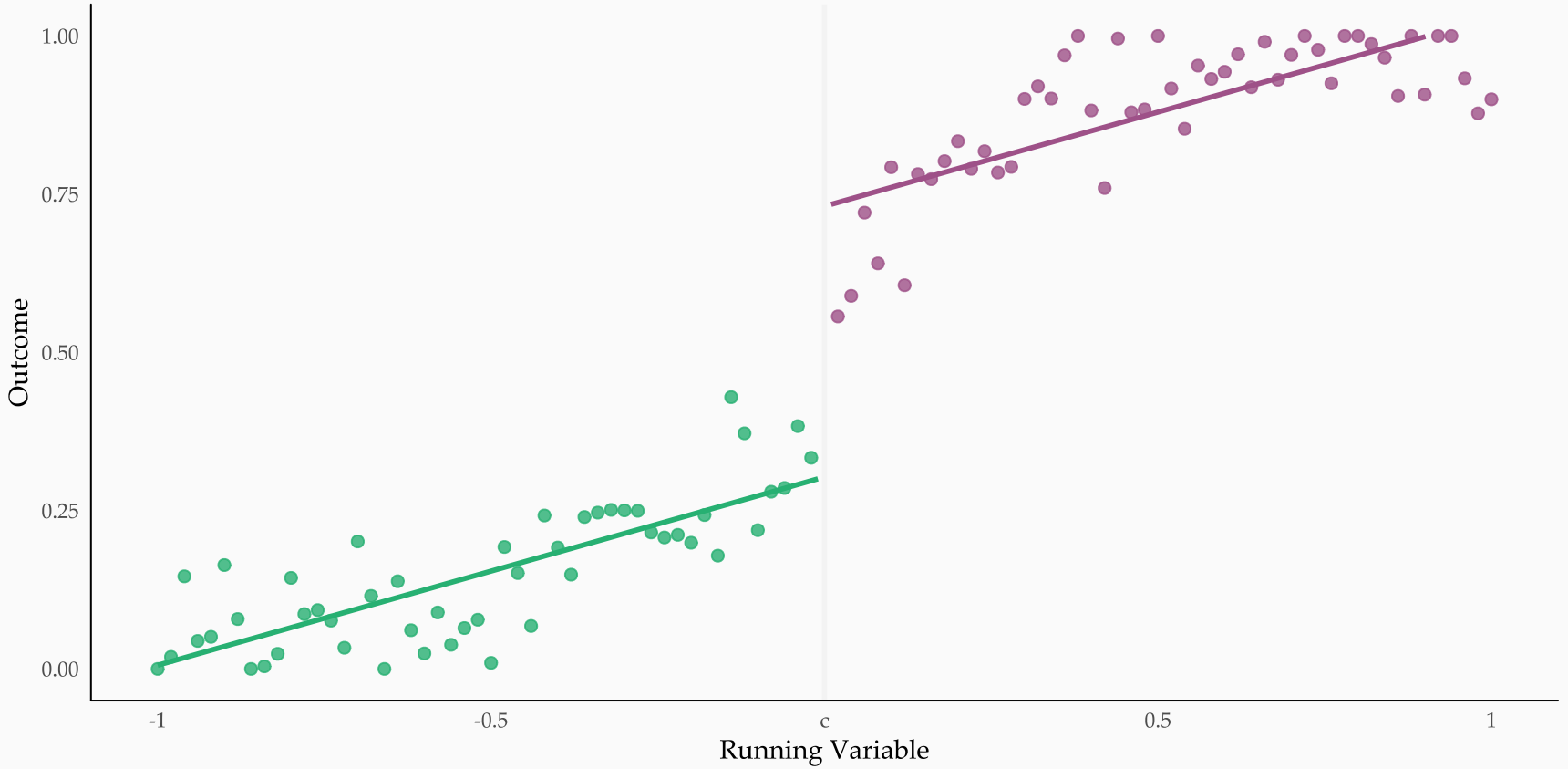
- $\tilde{X}_i = X_i - c$

Étape 2: choisir le modèle à estimer

- Linéaire avec pentes communes: $Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \varepsilon_i$
- Linéaire avec pentes différentes: $Y_i = \alpha + \beta(D_i \times \tilde{X}_i) + \delta \tilde{X}_i + \eta D_i + \varepsilon_i$
- Quadratique: $Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \lambda \tilde{X}_i^2 + \eta(D_i \times \tilde{X}_i) + \nu(D_i \times \tilde{X}_i)^2 + \varepsilon_i$

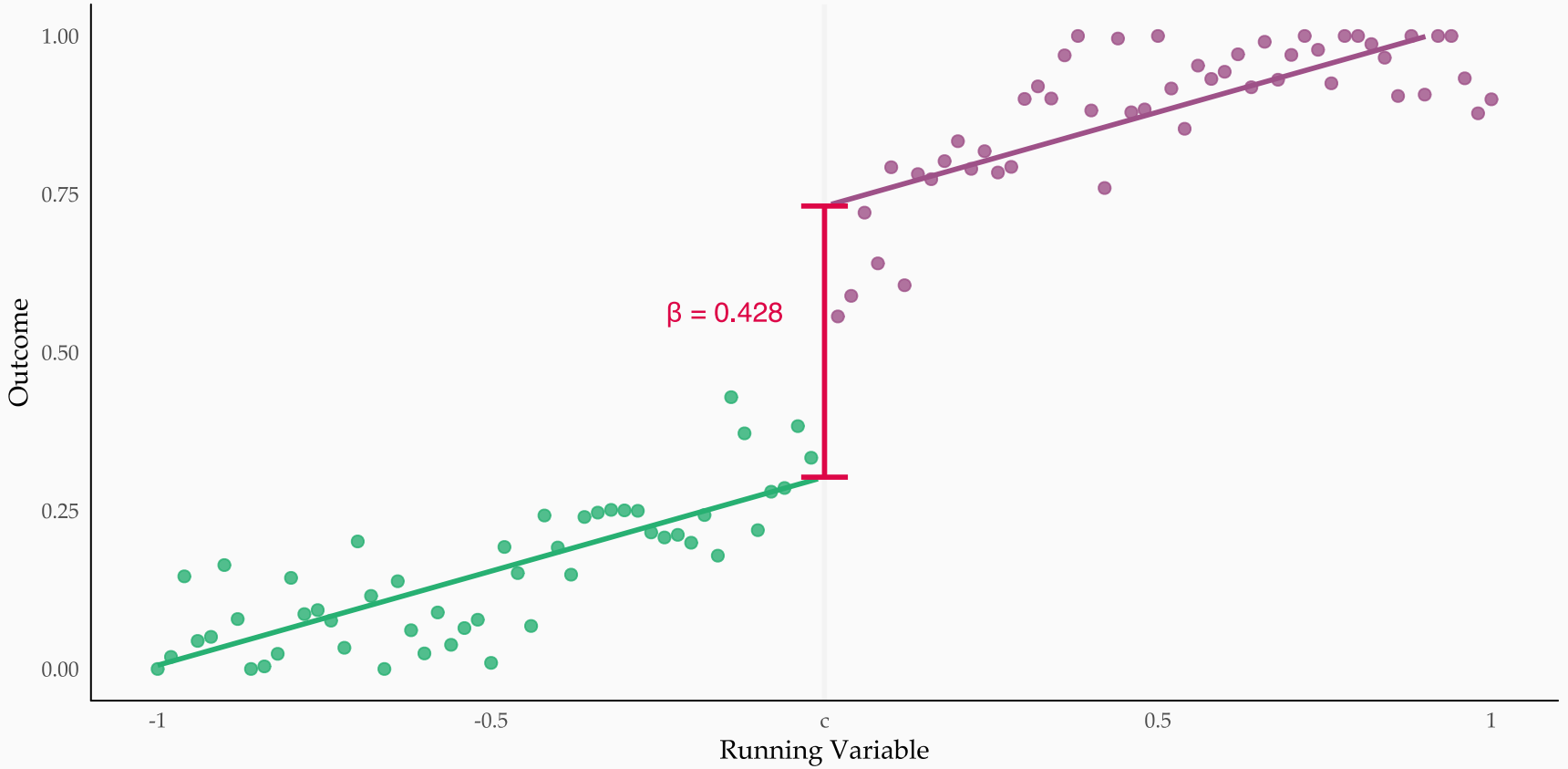
2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \varepsilon_i$$



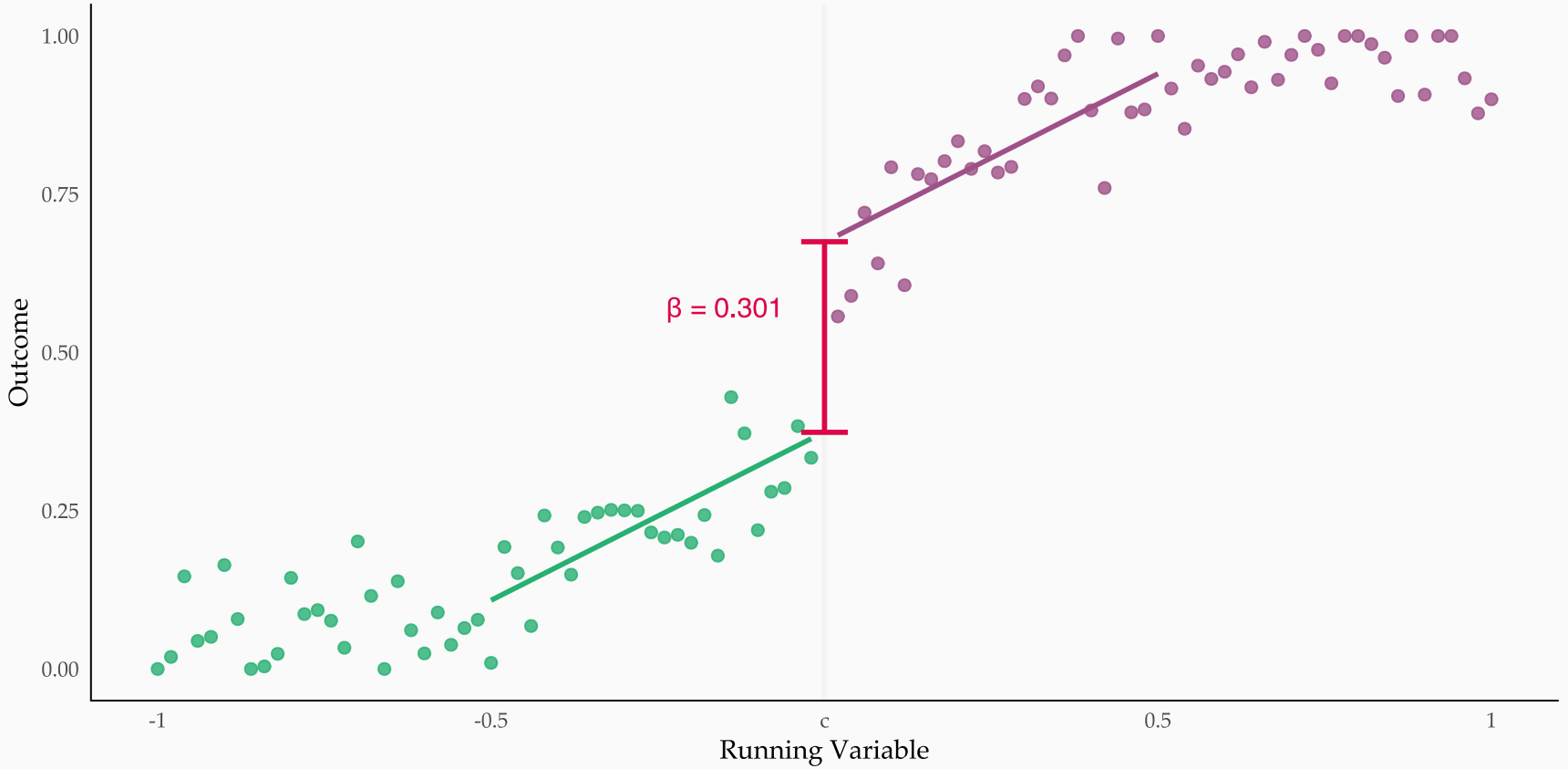
2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \varepsilon_i$$



2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \varepsilon_i$$

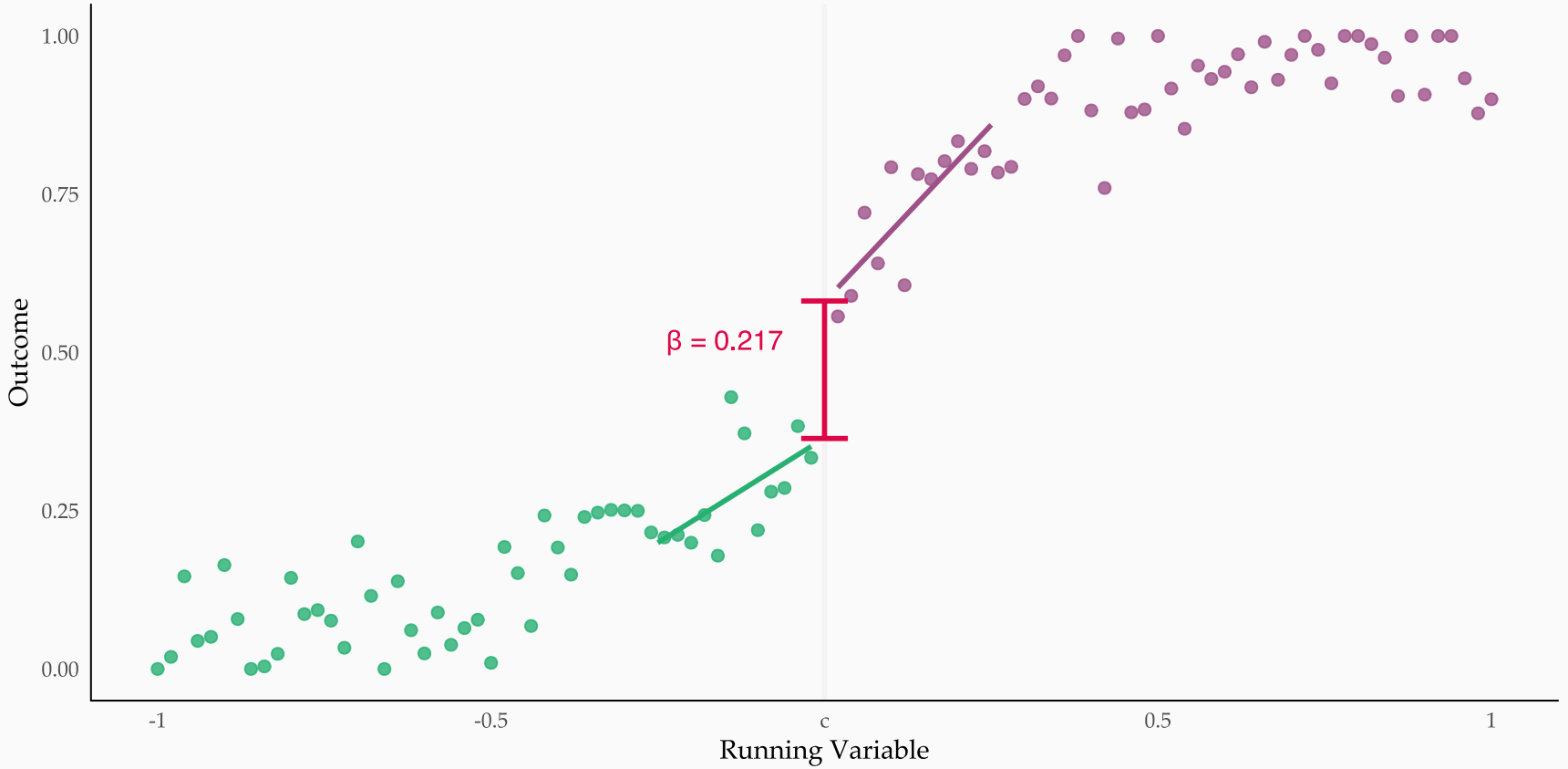


2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta(D_i \times \tilde{X}_i) + \delta\tilde{X}_i + \eta D_i + \varepsilon_i$$

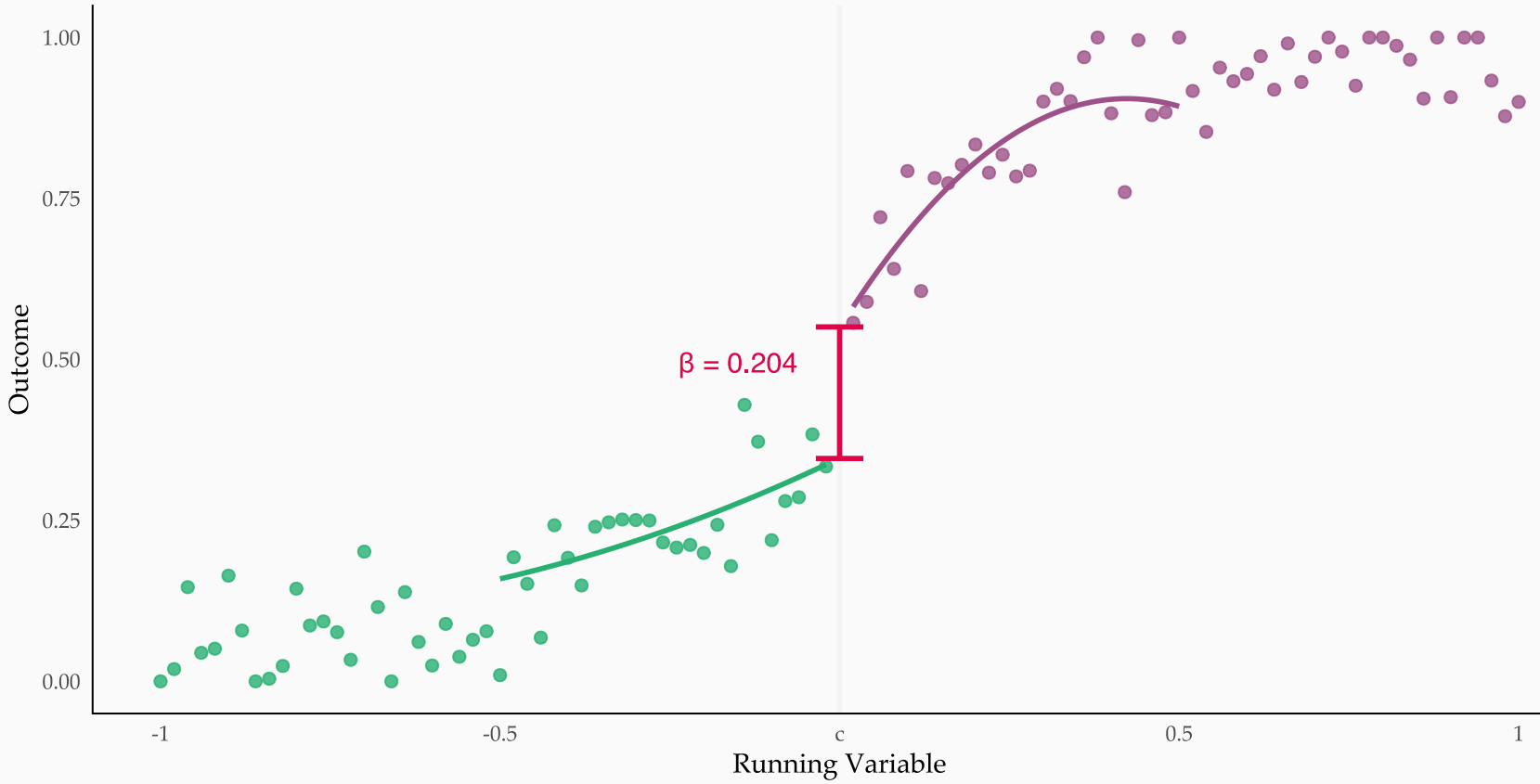
2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta(D_i \times \tilde{X}_i) + \delta\tilde{X}_i + \eta D_i + \varepsilon_i$$



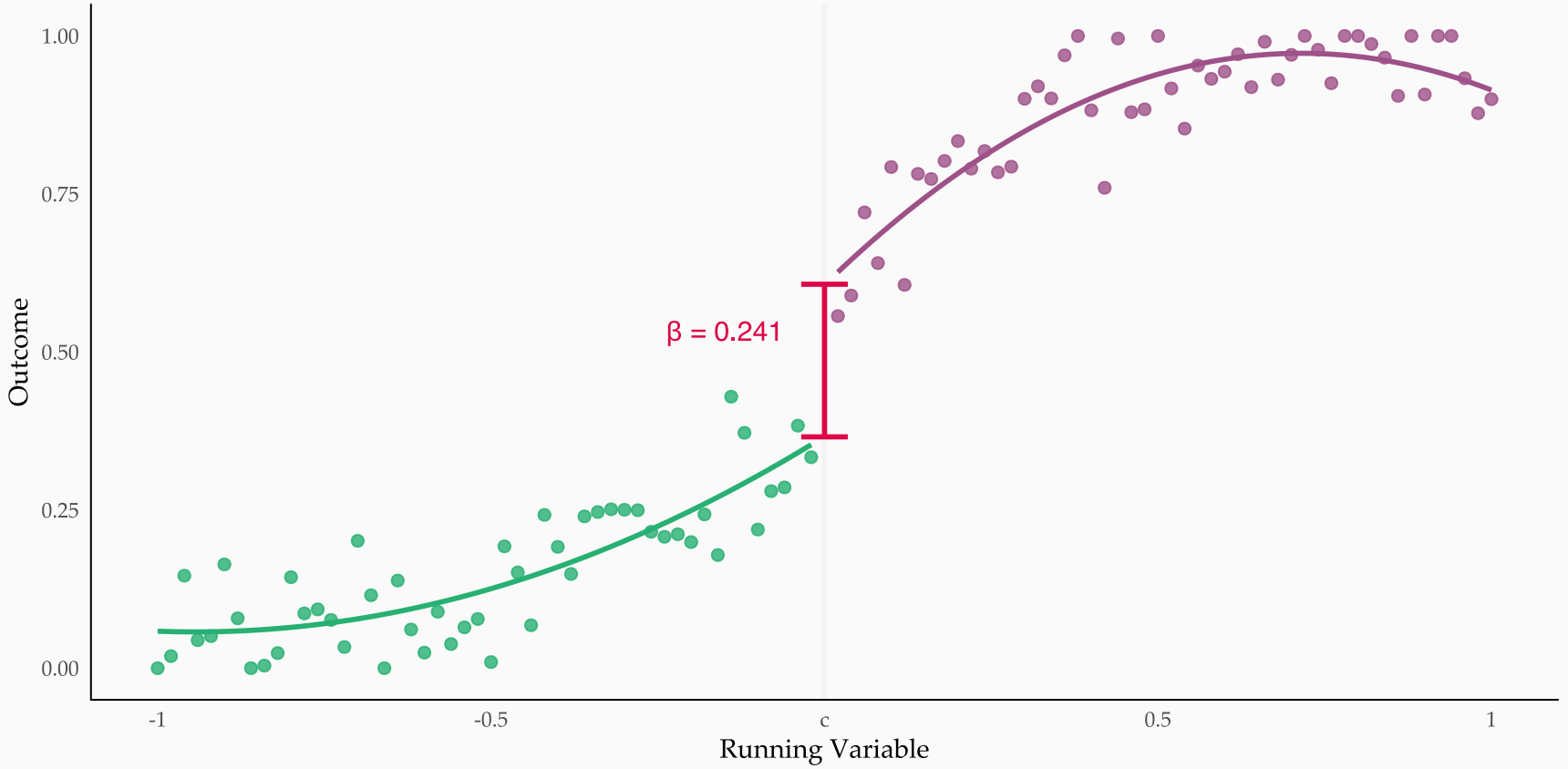
2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \lambda \tilde{X}_i^2 + \eta(D_i \times \tilde{X}_i) + \nu(D_i \times \tilde{X}_i)^2 + \varepsilon_i$$



2.5. Importance de la forme fonctionnelle

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \lambda \tilde{X}_i^2 + \eta(D_i \times \tilde{X}_i) + \nu(D_i \times \tilde{X}_i)^2 + \varepsilon_i$$



Application: Carpenter and Dobkin (2011)

Question de Recherche: quel est l'effet causal de la consommation d'alcool sur la mortalité des jeunes?

Application: Carpenter and Dobkin (2011)

Question de Recherche: quel est l'effet causal de la consommation d'alcool sur la mortalité des jeunes?

Question: pourquoi ne peut-on pas simplement comparer?

Application: Carpenter and Dobkin (2011)

Question de Recherche: quel est l'effet causal de la consommation d'alcool sur la mortalité des jeunes?

Question: pourquoi ne peut-on pas simplement comparer?

Biais de sélection/OVB:

- tout ce qui n'est pas observable et qui impacte à la fois la consommation d'alcool et la mortalité

Application: Carpenter and Dobkin (2011)

Question de Recherche: quel est l'effet causal de la consommation d'alcool sur la mortalité des jeunes?

Question: pourquoi ne peut-on pas simplement comparer?

Biais de sélection/OVB:

- tout ce qui n'est pas observable et qui impacte à la fois la consommation d'alcool et la mortalité

Carpenter and Dobkin (2011): utilisent l'âge minimum légal à partir duquel un individu est autorisé à consommer de l'alcool (MLDA: Minimum Legal Drinking Age)

Contexte: US

- MLDA: 21 ans

Importer les données:

```
library(masteringmetrics)
data("mlda", package = "masteringmetrics")
```

Application: Carpenter and Dobkin (2011)

```
## # A tibble: 6 × 19
##   agecell   all allfitted internal internalfitted external externalfitted
##     <dbl> <dbl>     <dbl>    <dbl>          <dbl>    <dbl>          <dbl>
## 1    19.1  92.8      91.7     16.6           16.7     76.2           75.0
## 2    19.2  95.1      91.9     18.3           16.9     76.8           75.0
## 3    19.2  92.1      92.0     18.9           17.1     73.2           75.0
## 4    19.3  88.4      92.2     16.1           17.3     72.3           74.9
## 5    19.4  88.7      92.3     17.4           17.4     71.3           74.9
## 6    19.5  90.2      92.5     17.9           17.6     72.3           74.9
## # ℹ 12 more variables: alcohol <dbl>, alcoholfitted <dbl>, homicide <dbl>,
## #   homicidefitted <dbl>, suicide <dbl>, suicidefitted <dbl>, mva <dbl>,
## #   mvafitted <dbl>, drugs <dbl>, drugsfitted <dbl>, externalother <dbl>,
## #   externalotherfitted <dbl>
```

Application: Carpenter and Dobkin (2011)

1) Quelle est la running variable? Quel est le seuil? Quelle est l'hypothèse d'identification?

2) Construisez la variable `above21` qui vaut 1 pour toutes les classes d'âge ≥ 21

3) Représenter graphiquement l'évolution **linéraire et quadratique** de la mortalité liée aux accidents de la route (variable `mva`) et de la mortalité globale (`all`) autour du seuil de discontinuité. *Hint: utilisez les commandes* `geom_smooth()` *de* `ggplot` *pour tracer des droites de régressions sur un graphique et* `poly(x,2)` *pour un polynôme de degré 2 de la variable* x.

4) **Pour le prochain cours**: Estimez l'effet d'atteindre le MLDA sur la mortalité globale (linéaire avec pente identique, linéaire avec pente différente, et quadratique).

Solution : Carpenter and Dobkin (2011)

1) Quelle est la **running variable**? Quel est le seuil? Quelle est l'hypothèse d'identification?

Solution : Carpenter and Dobkin (2011)

1) Quelle est la **running variable**? Quel est le seuil? Quelle est l'hypothèse d'identification?

Running Variable: Age

Solution : Carpenter and Dobkin (2011)

1) Quelle est la running variable? Quel est le **seuil**? Quelle est l'hypothèse d'identification?

Running Variable: Age

Solution : Carpenter and Dobkin (2011)

1) Quelle est la running variable? Quel est le **seuil**? Quelle est l'hypothèse d'identification?

Running Variable: Age

Seuil: 21 ans

Solution : Carpenter and Dobkin (2011)

1) Quelle est la running variable? Quel est le seuil? Quelle est l'**hypothèse d'identification**?

Running Variable: Age

Seuil: 21 ans

Solution : Carpenter and Dobkin (2011)

1) Quelle est la running variable? Quel est le seuil? Quelle est l'**hypothèse d'identification**?

Running Variable: Age

Seuil: 21 ans

Hypothèse d'identification: Autour du seuil, l'allocation du traitement est *as good as random*, i.e. les individus autour du seuil de leur 21ème anniversaire, ne diffèrent en moyenne que de par leur accès ou on à l'alcool

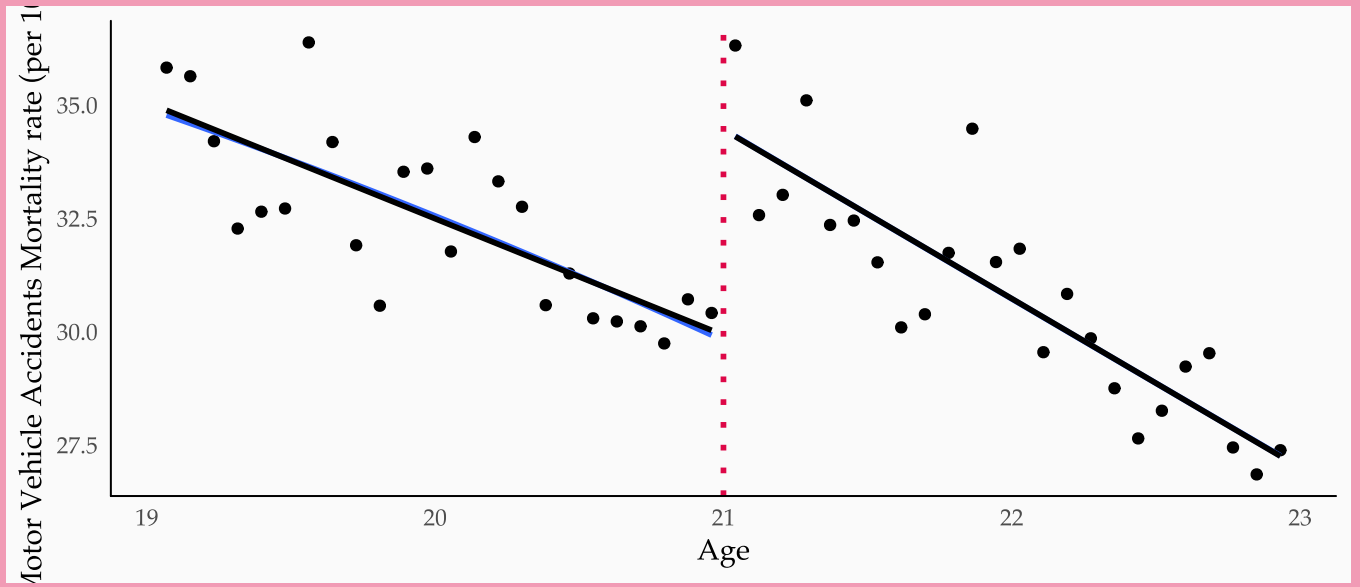
Application: Carpenter and Dobkin (2011)

2) Construisez la variable `above21` qui vaut 1 pour toutes les classes d'âge ≥ 21

```
mlda = mlda %>%
  mutate(above21 = ifelse(agecell >= 21, 1, 0))
```

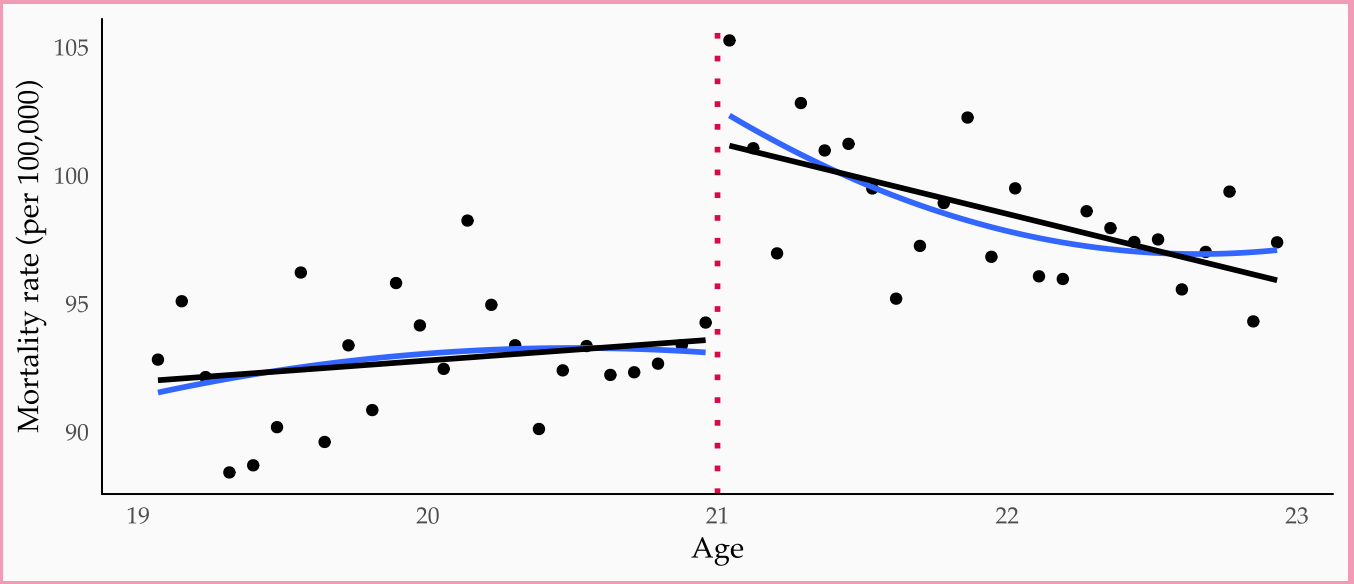
Application: Carpenter and Dobkin (2011)

```
mlda %>%
  ggplot(aes(x = agecell, y = mva)) +
  geom_point() +
  geom_smooth(aes(group = above21), se = FALSE, method = "lm", formula = y ~ poly(x, 2)) +
  geom_smooth(aes(group = above21), se = FALSE, method = "lm", formula = y ~ x, color = "black") +
  geom_vline(xintercept = 21, color = "#dd0747", linetype = "dotted", linewidth = 1) +
  labs(y = "Motor Vehicle Accidents Mortality rate (per 100,000)", x = "Age") +
  theme_minimum
```



Application: Carpenter and Dobkin (2011)

```
mlda %>%
  ggplot(aes(x = agecell, y = all)) +
  geom_point() +
  geom_smooth(aes(group = above21), se = FALSE, method = "lm", formula = y ~ poly(x, 2)) +
  geom_smooth(aes(group = above21), se = FALSE, method = "lm", formula = y ~ x, color = "black") +
  geom_vline(xintercept = 21, color = "#dd0747", linetype = "dotted", linewidth = 1) +
  labs(y = "Mortality rate (per 100,000)", x = "Age") +
  theme_minimum
```



3. Fuzzy RDD

3. Fuzzy RDD

Les **fuzzy RDD** exploitent un changement discontinu dans l'assignation au traitement D_i au seuil de c.

Contrairement au *sharp* RDD où la probabilité d'être traité passe de 0 à 1 lorsque X_i passe le seuil de discontinuité c, dans un cas de *fuzzy* RDD, la probabilité de traitement est continue au voisinage de c:

$$0 < \lim_{x \to c^+} \mathbb{P}(D_i = 1 | X_i = x) - \lim_{x \to c^-} \mathbb{P}(D_i = 1 | X_i = x) < 1$$

On a maintenant deux effets lorsque X_i franchit le seuil c:

1. L'effet sur l'**outcome**: $\lim_{x \to c^+} \mathbb{E}(Y_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_i | X_i = x)$
2. L'effet sur la **probabilité de traitement**: $\lim_{x \to c^+} \mathbb{E}(D_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(D_i | X_i = x)$

L'estimateur de l'effet du traitement dans le cas d'une *fuzzy* RDD est donc:

$$\beta_{\text{RDD}}^{\text{fuzzy}} = \frac{\lim_{x \to c^+} \mathbb{E}(Y_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_i | X_i = x)}{\lim_{x \to c^+} \mathbb{E}(D_i | X_i = x) - \lim_{x \to c^-} \mathbb{E}(D_i | X_i = x)}$$

3. Fuzzy RDD

La formule de l'estimateur nous rappelle...

3. Fuzzy RDD

La formule de l'estimateur nous rappelle...

Celle de l'estimateur IV! Ici, l'instrument $Z_i = 1\{X_i \geq c\}$!!

3. Fuzzy RDD

La formule de l'estimateur nous rappelle...

Celle de l'estimateur IV! Ici, l'instrument $Z_i = 1\{X_i \geq c\}$!!

Angrist, J. and Lavy, V. (1999). *Using Maimonides' rule to estimate the effect of class size on scholastic achievement*. QJE

- Règle de Maïmonide: en Israël, règle stipulant qu'au-delà de 40 élèves par classe, l'école doit scinder la classe en deux $\Longrightarrow$ crée une discontinuité dans la taille des classes dès que la taille des classes dépasse 40 élèves, 80 élèves, 120 élèves.
- En pratique : la règle n'est pas toujours appliquée de manière stricte (certaines écoles de petite taille n'ont pas les moyens matériels de scinder la classe, d'autres peuvent décider de scinder avant d'atteindre 40 élèves) $\Longrightarrow$ le fait d'être au-dessus du seuil de 40 élèves augmente fortement la probabilité que la classe soit effectivement scindée **mais ne passe pas de 0 à 1**

Recap: Regression Discontinuity Design (RDD)

Data: Données observationnelles

Hypothèse d'identification:

- Intuition: individus comparables autour du seuil de discontinuité
- Formellement: $\mathbb{E}(Y_{1i}|X_i = x)$ et $\mathbb{E}(Y_{0i}|X_i = x)$ sont continues en x

Modèle Sharp RDD: pour tout individu i,

$$Y_i = \alpha + \beta D_i + \delta \tilde{X}_i + \varepsilon_i \quad \text{où} \quad D_i = 1\{X_i \geq c\}$$

Estimateur de l'effet du traitement:

- Différence entre les valeurs prédites juste au-dessus et juste en-dessous du seuil
- $\hat{\beta}^{\text{sharp}}_{\text{RDD}} = \lim_{x \to c^+} \mathbb{E}(Y_i|X_i = x) - \lim_{x \to c^-} \mathbb{E}(Y_i|X_i = x)$

Implémentation sur `R`:

- Visualisation d' l'outcome autour du seuil avec `ggplot`
- Tests de manipulation: `DCdensity()` du package `rdd` (test de McCrary)
- Estimation: `lm(y ~ D + X + I(D*X), data = data)` ou packages `rdrobust`, `rddensity`

Sources

Causal inference: The Mixtape, Scott Cunningham

Mixtape Sessions: RDD, Scott Cunningham

Mastering Metrics: RDD

RDD, Edward Rubin