Régressions Linéaires et Causalité

Pratiques de la Recherche en Économie

Florentine Oliveira

2024-11-19

Cette séance

1. Rappels: Régression linéaire simple

1. Rappels: Régression linéaire simple

1.1. Interprétation géométrique

La régression linéaire simple est une méthode statistique permettant de trouver une relation **linéaire** entre

- une **variable expliquée** (ou **variable dépendante** ou ***outcome***), y
- une **variable explicative** (ou **variable indépendante** ou **régresseur**), x

La relation linéaire entre y et x n'est pas parfaite: elle est perturbée par une **erreur** (ou **bruit** ou ***noise***), ε qui comprend tous les facteurs **non observés** qui affectent y.

Le modèle linéaire univarié s'écrit, $\forall\ i$,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

1. Rappels: Régression linéaire simple

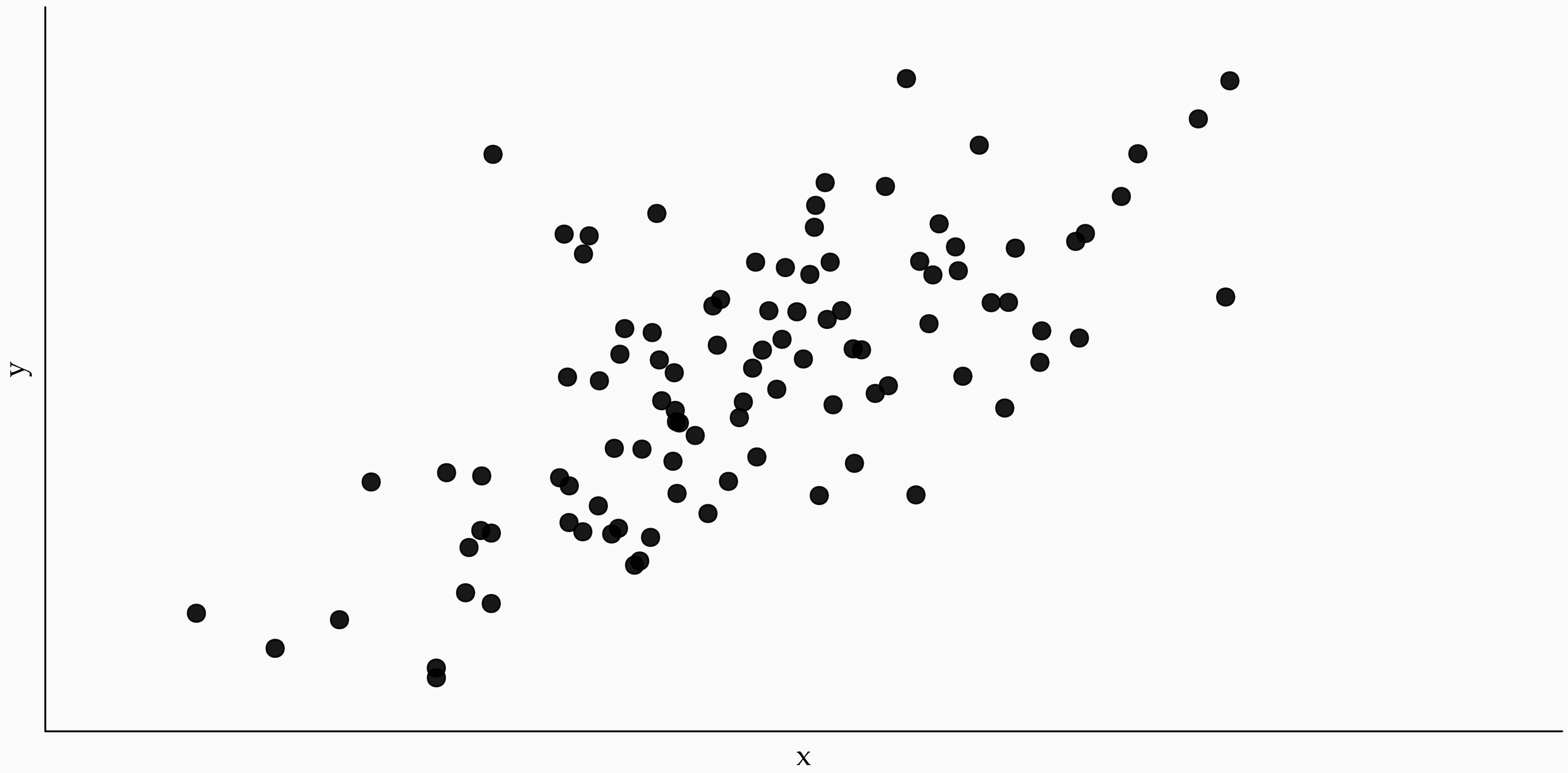
1.1. Interprétation géométrique

On considère l'échantillon suivant

1. Rappels: Régression linéaire simple

1.1. Interprétation géométrique

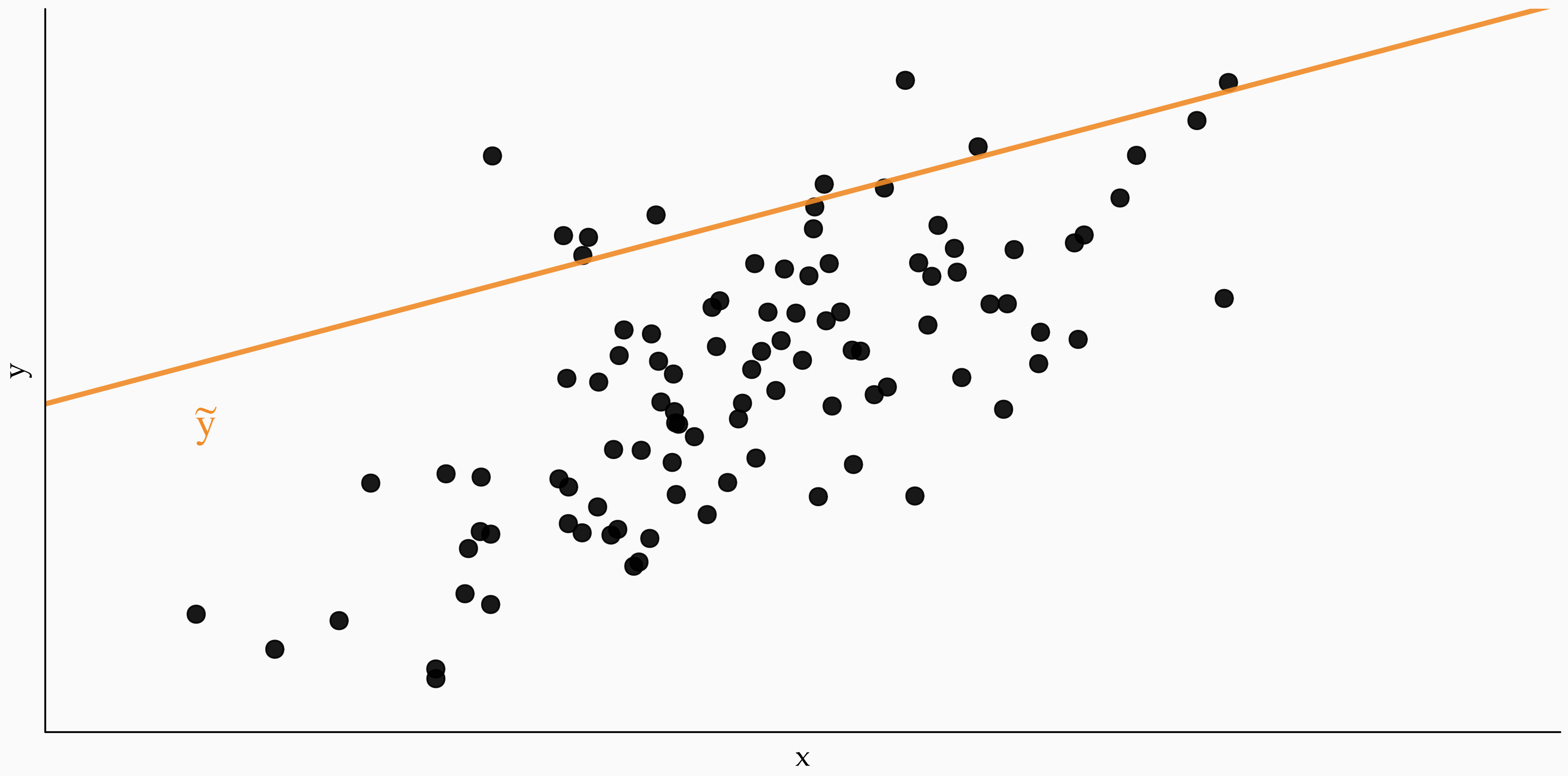
On considère l'échantillon suivant



1. Rappel: Régression linéaire simple

1.1. Interprétation géométrique

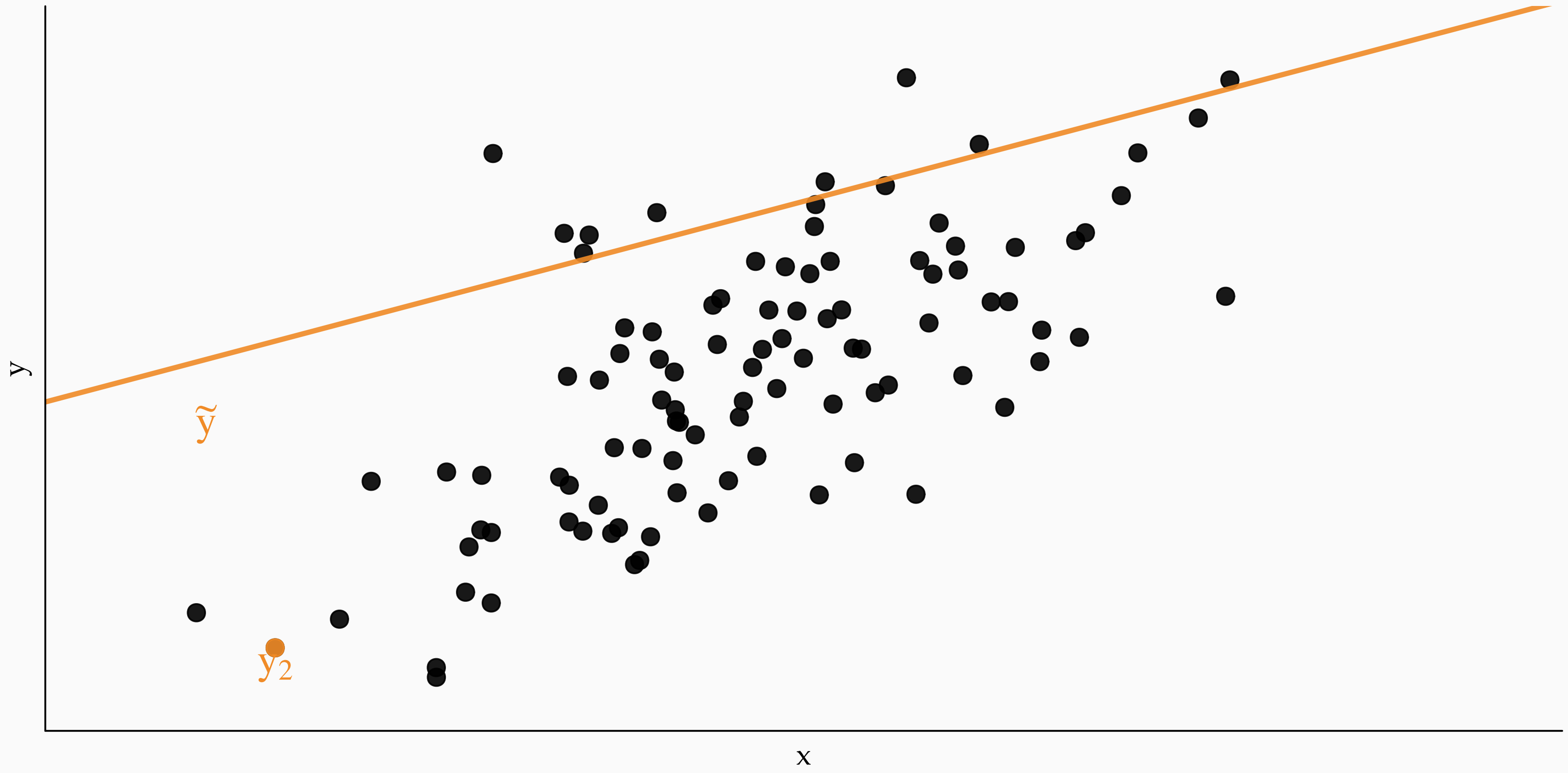
Pour toute droite $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$,



1. Rappel: Régression linéaire simple

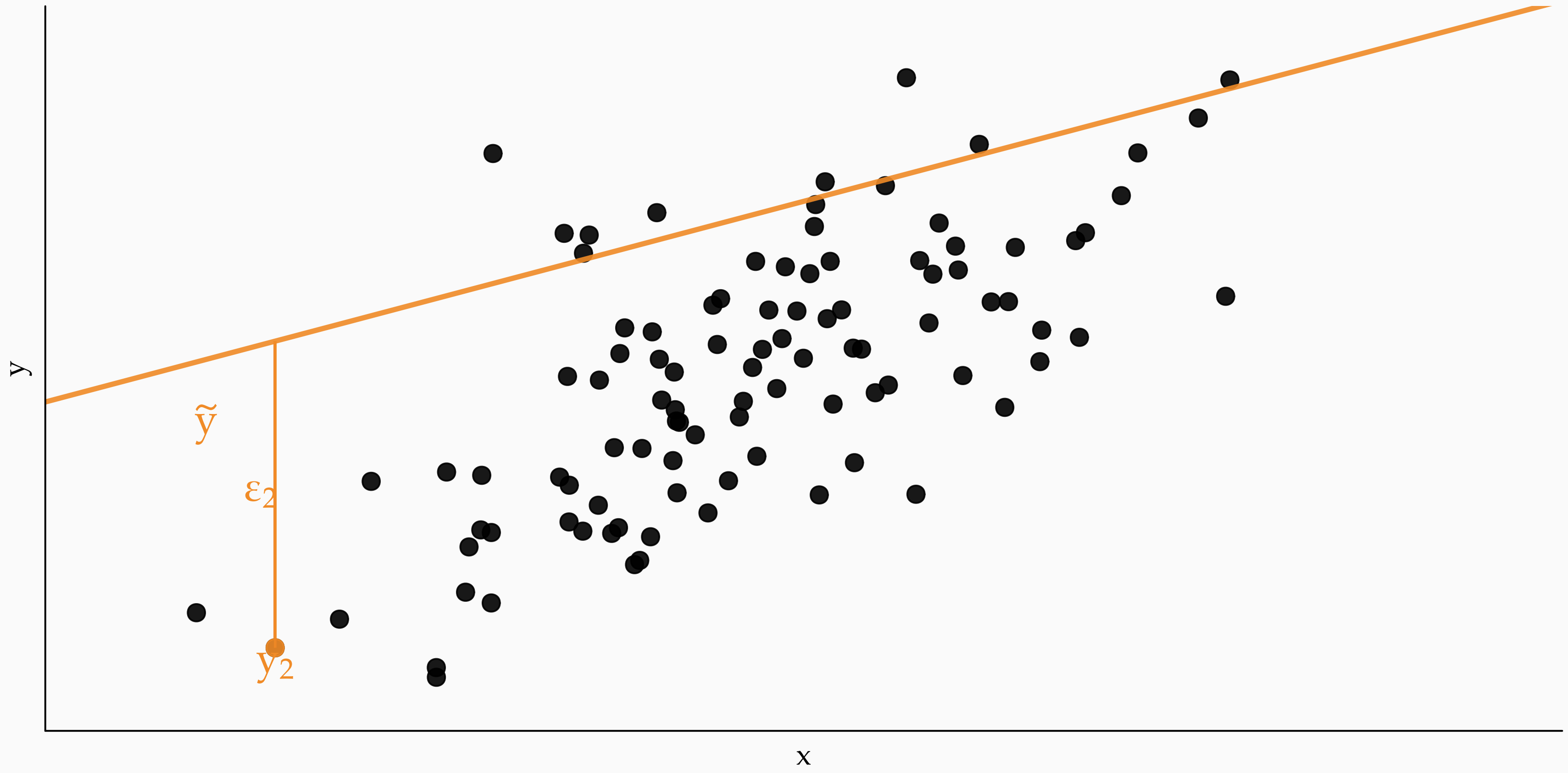
1.1. Interprétation géométrique

Pour toute droite $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$, on peut calculer les erreurs: $\varepsilon_i = y_i - \tilde{y}_i$



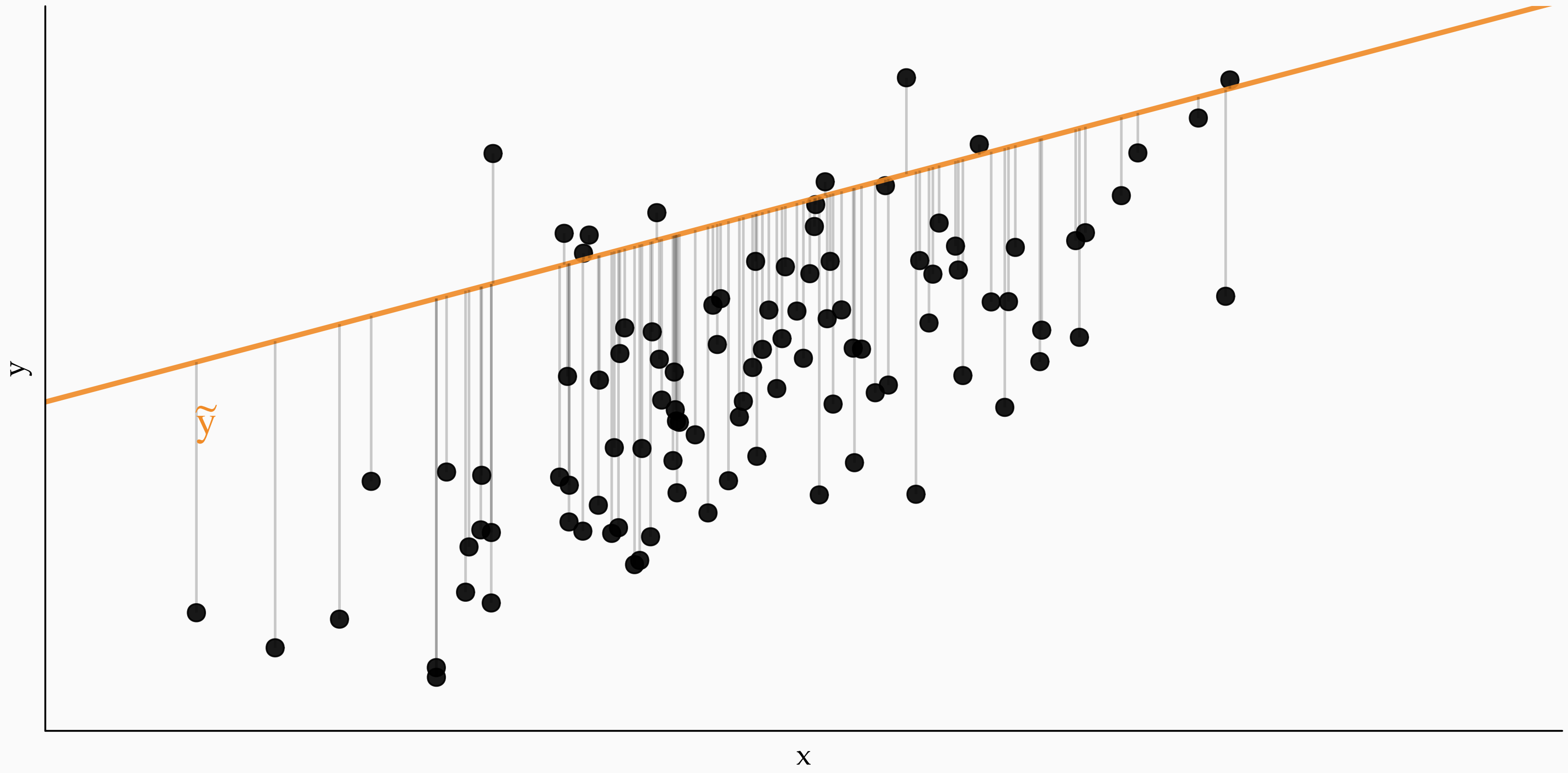
1.1. Interprétation géométrique

Pour toute droite $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$, on peut calculer les erreurs: $\varepsilon_i = y_i - \tilde{y}_i$



1.1. Interprétation géométrique

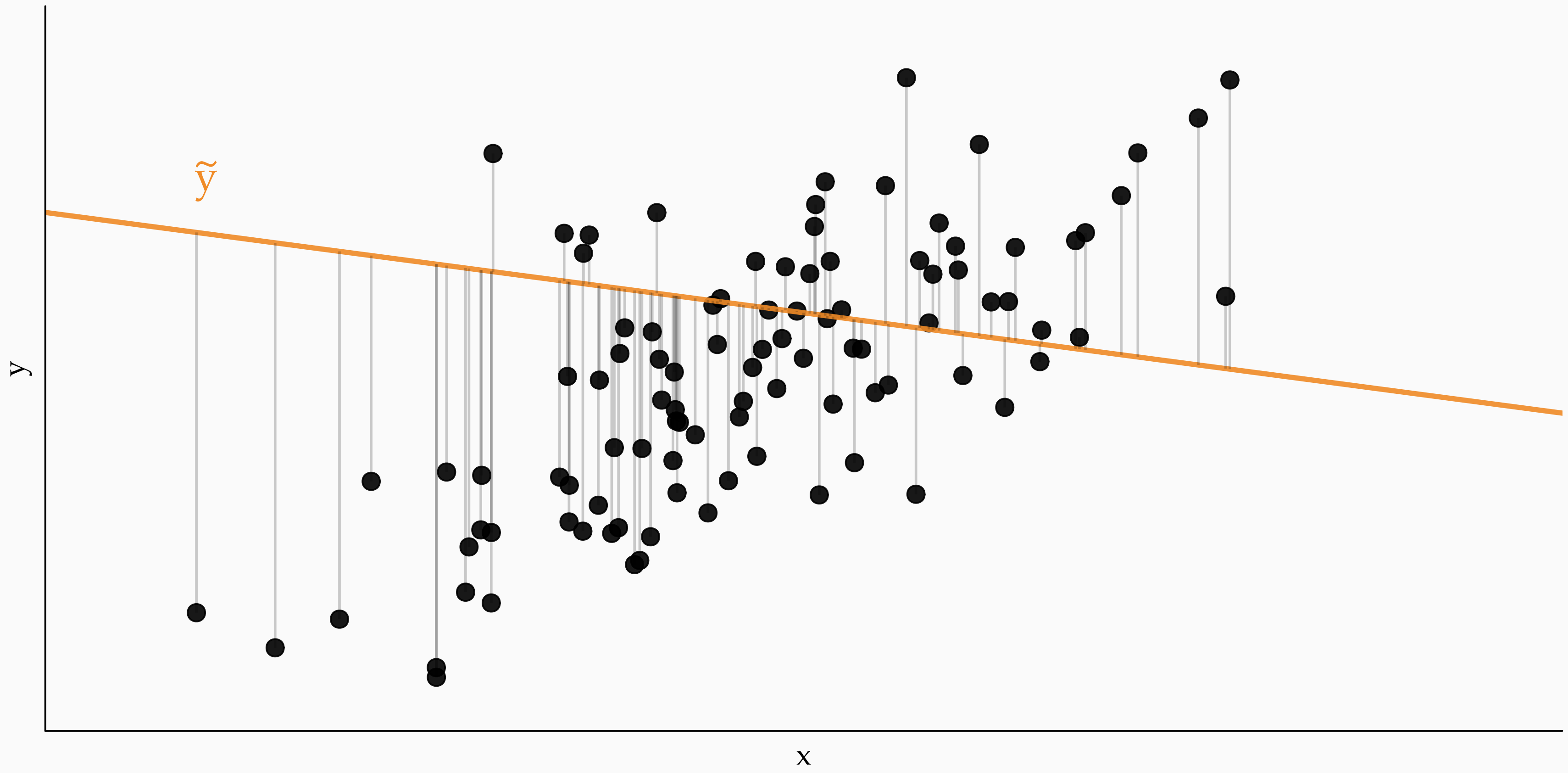
Pour toute droite $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$, on peut calculer les erreurs: $\varepsilon_i = y_i - \tilde{y}_i$



1. Rappel: Régression linéaire simple

1.1. Interprétation géométrique

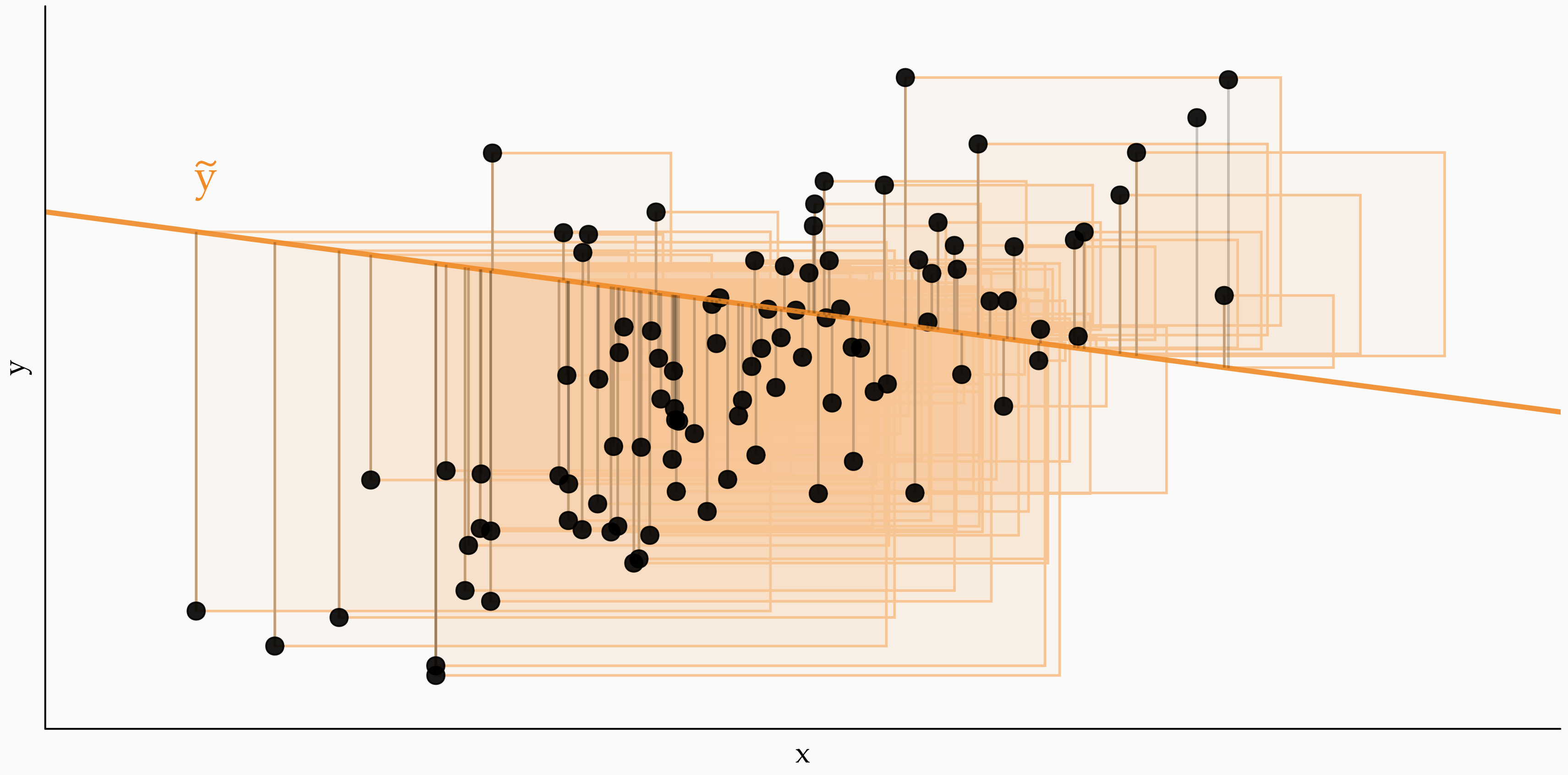
Pour toute droite $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x$, on peut calculer les erreurs: $\varepsilon_i = y_i - \tilde{y}_i$



1. Rappel: Régression linéaire simple

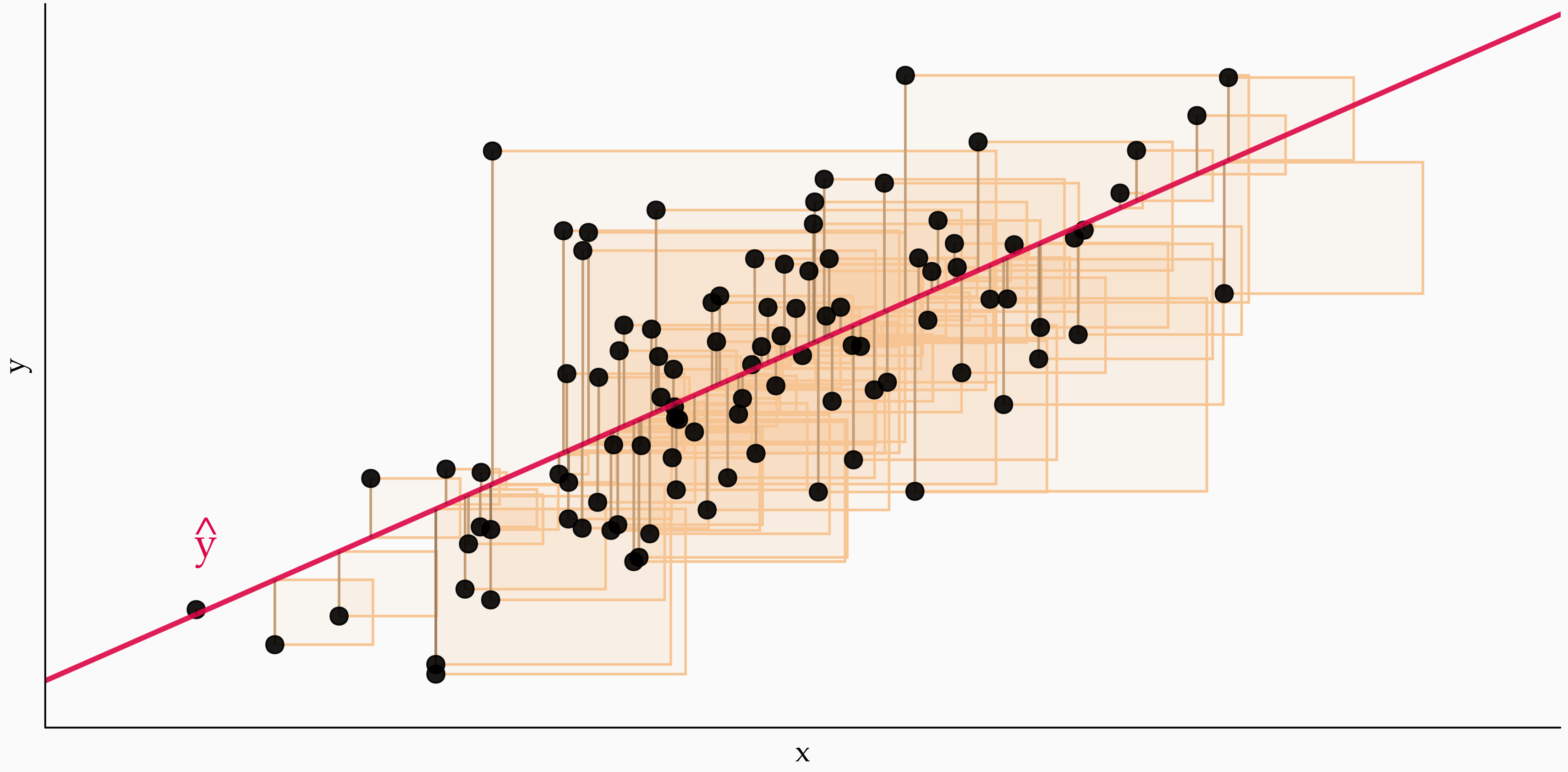
1.1. Interprétation géométrique

SCE = $\left(\sum \varepsilon_i^2\right)$: les erreurs importantes sont davantage pénalisées



1.1. Interprétation géométrique

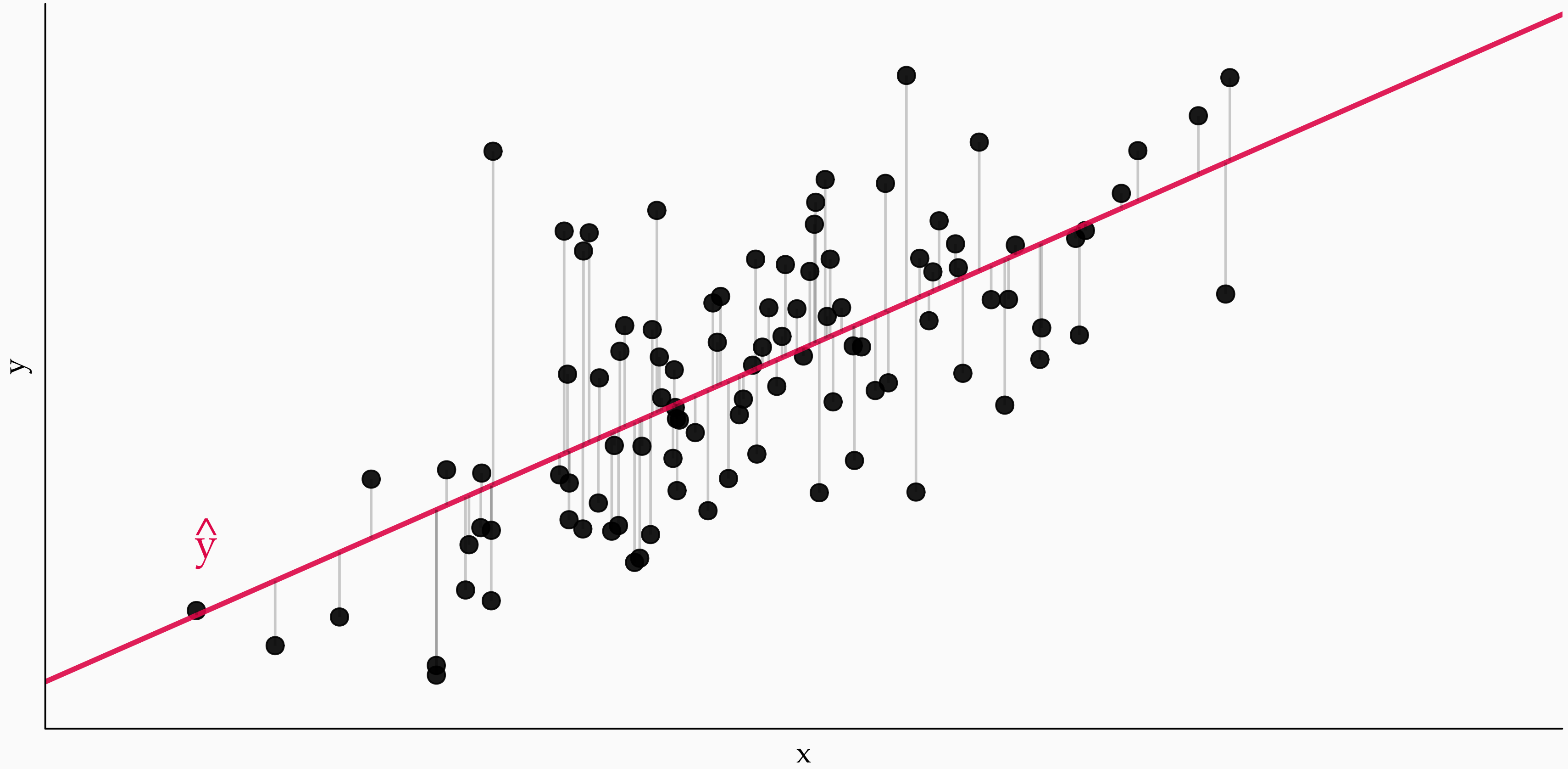
L'estimateur des MCO (*OLS*) calcule $\hat{\beta}_0$ et $\hat{\beta}_1$ qui **minimisent la SCE**.



1. Rappel: Régression linéaire simple

1.1. Interprétation géométrique

L'estimateur des MCO (*OLS*) calcule $\hat{\beta}_0$ et $\hat{\beta}_1$ qui **minimisent la SCE**.



1.2. Formule de l'estimateur MCO dans le cas univarié

L'estimateur des MCO calcule $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimmise la Somme des Carrés des Erreurs (SCE, ou *Sum of Squared Errors*) :

$$\min_{\hat{\beta}_0,\, \hat{\beta}_1} \mathrm{SCE} = \sum_{i=1}^{N} \varepsilon_i^2$$

On obtient, dans le cas univarié:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}\overline{x}$$

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)}$$

[Maths modèle univarié]()

1.3. Hypothèses

H_1 **Linéarité**: le modèle est linéaire dans les paramètres

- Formellement, $\frac{\partial y_i}{\partial x_{ik}} = \beta_k, \forall k = 1, \ldots, K$

H_2 **Échantillon Aléatoire**: l'échantillon (x_i, y_i) est un échantillon aléatoire et représentatif de la population.

- Visualisation

H_3 **Exogeneité/Identification**: X est exogène

- Formellement, $\mathbb{E}(\varepsilon_i | x) = 0$

H_4 **Variation**: il y a suffisamment de variation dans x.

- Dit autrement, chaque variable explicative apporte une information qui lui est propre
- Formellement, les explicatives ne sont pas colinéaires (cas univarié: $x_i \neq$ constante)
- 🚨 Outliers

1.3. Hypothèses

H_5 **Les erreurs ε_i sont sphériques** :

- H_{5a} **Homoscédasticité** : la variance est constante : $\forall i$, $\mathbb{V}(\varepsilon_i|x) = \mathbb{E}(\varepsilon_i^2|x) = \sigma^2$
  - Visualisation
- H_{5b} **Absence d'autocorrélation** : $\mathbb{E}(\varepsilon_i\varepsilon_j|x) = 0, \forall i \neq j$

Propriété: Normalité (asymptotique): sous hypothèse que l'échantillon (x_i, y_i) est iid, l'esitmateur $\hat{\beta}$ suit une loi normale:

$$\hat{\beta} \sim \mathcal{N}(\beta, \mathbb{V}(\hat{\beta}))$$

Sous ces hypothèses, l'estimateur des MCO est BLUE (Best Linear Unbiased Estimator) (cf démonstrations faites en cours).

1.4. Implémentation sur `R`

- Calcul de la variance empirique

```
var(x)
```

- Calcul de la covariance empirique

```
cov(x,y)
```

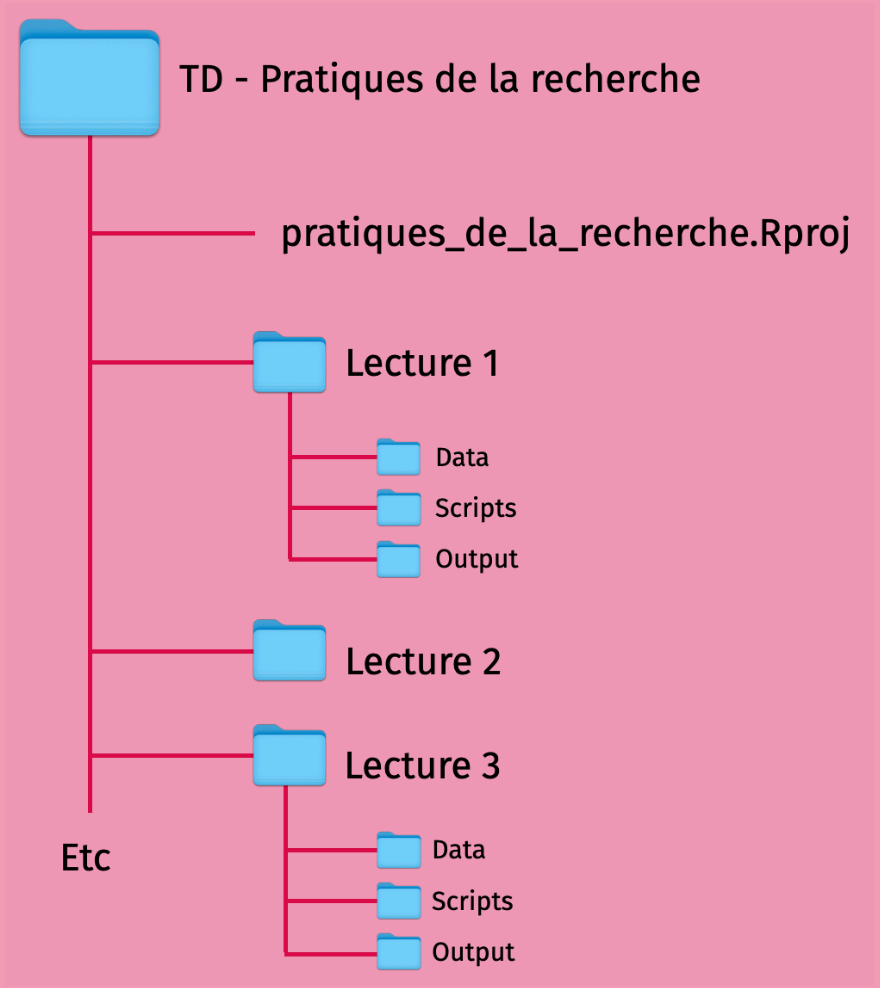
- Régression linéaire (simple)

```
lm(variable dépendante ~  variable indépendante, data = data.frame)
```

- Résultats de l'estimation visibles avec la commande `summary`

Application

Projet R



Mise à jour structure projet R :

- un seul projet pour l'ensemble du TD
- un dossier par TD

Application

05:00

Taille de la fratrie et performances scolaires

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `df`

2) Effectuer les opérations ne nettoyage de données suivantes:

- ne conserver que les enfants âgés d'au moins 25 ans
- ne conserver que les enfants dont la mère avait entre 16 et 49 ans à leur naissance
- ne conserver que les familles pour lesquelles le nombre d'observations au sein de la famille est égale à la taille de la fratrie

Application

10:00

Taille de la fratrie et performances scolaires

Intuition

3) Selon vous, quelle est la relation entre **taille de la fratrie** et **performances scolaires** ?

Sur `R` : se familiariser avec les données

4) Calculer quelques statistiques descriptives de la **taille de la fratrie** (`family_size`) et du **nombre d'années d'études** (`education`):

- moyenne, écart-type, corrélation entre les deux variables

5) Représenter le nuage de points qui définit la relation entre **taille de la fratrie** et **nombre d'années d'études**

6) Calculer les estimateurs de β_0 et β_1 du modèle $\text{Nb Années études} = \beta_0 + \beta_1 \text{Taille Fratrie} + \varepsilon$, à la main et directement via la fonction `lm`

Solution

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `data`

```
# Import data
df = readRDS("data/simulated_data_black_et_al_2005.rds")
```

Solution

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `data`

```
# Import data
df = readRDS("data/simulated_data_black_et_al_2005.rds")
```

2) Effectuer les opérations ne nettoyage de données suivantes:

- ne conserver que les enfants âgés d'au moins 25 ans
- ne conserver que les enfants dont la mère avait entre 16 et 49 ans à leur naissance
- ne conserver que les familles pour lesquelles le nombre d'observations au sein de la famille est égale à la taille de la fratrie

```
df = df %>%
  filter(age_2000 >= 25,
         (mother_age_2000 - age_2000 <= 49) & (mother_age_2000 - age_2000 >= 16)) %>%
  group_by(family_id) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(family_size == n)
```

Solution

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `df`

```
# Import data
df = readRDS("data/simulated_data_black_et_al_2005.rds")
```

2) Effectuer les opérations ne nettoyage de données suivantes:

- **ne conserver que les enfants âgés d'au moins 25 ans**
- ne conserver que les enfants dont la mère avait entre 16 et 49 ans à leur naissance
- ne conserver que les familles pour lesquelles le nombre d'observations au sein de la famille est égale à la taille de la fratrie

```
df = df %>%
  filter(age_2000 >= 25,
         (mother_age_2000 - age_2000 <= 49) & (mother_age_2000 - age_2000 >= 16)) %>%
  group_by(family_id) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(family_size == n)
```

Solution

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `df`

```
# Import data
df = readRDS("data/simulated_data_black_et_al_2005.rds")
```

2) Effectuer les opérations ne nettoyage de données suivantes:

- ne conserver que les enfants âgés d'au moins 25 ans
- **ne conserver que les enfants dont la mère avait entre 16 et 49 ans à leur naissance**
- ne conserver que les familles pour lesquelles le nombre d'observations au sein de la famille est égale à la taille de la fratrie

```
df = df %>%
  filter(age_2000 >= 25,
         (mother_age_2000 - age_2000 <= 49) & (mother_age_2000 - age_2000 >= 16)) %>%
  group_by(family_id) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(family_size == n)
```

Solution

Cleaning

1) Importer la base de données `simulated_data_black_et_al_2005.rds` que vous nommerez `df`

```
# Import data
df = readRDS("data/simulated_data_black_et_al_2005.rds")
```

2) Effectuer les opérations ne nettoyage de données suivantes:

- ne conserver que les enfants âgés d'au moins 25 ans
- ne conserver que les enfants dont la mère avait entre 16 et 49 ans à leur naissance
- **ne conserver que les familles pour lesquelles le nombre d'observations au sein de la famille est égale à la taille de la fratrie**

```
df = df %>%
  filter(age_2000 >= 25,
         (mother_age_2000 - age_2000 <= 49) & (mother_age_2000 - age_2000 >= 16)) %>%
  group_by(family_id) %>%
  mutate(n = n()) %>%
  ungroup() %>%
  filter(family_size == n)
```

Solution

Intuition

3) Selon vous, quelle est la relation entre **taille de la fratrie** et **performances scolaires** ?

Solution

Intuition

3) Selon vous, quelle est la relation entre **taille de la fratrie** et **performances scolaires** ?

- Relation théorique (Becker (1960), Becker and Lewis (1973), Becker and Tomes (1976)):
    - arbitrage entre la quantité et la *qualité* des enfants au sein d'une famille.

Solution

Intuition

3) Selon vous, quelle est la relation entre **taille de la fratrie** et **performances scolaires** ?

- Relation théorique (Becker (1960), Becker and Lewis (1973), Becker and Tomes (1976)):
  - arbitrage entre la quantité et la *qualité* des enfants au sein d'une famille.
- Relations empiriques (Black, Devereux and Salvanes (2005)):
  - données norvégiennes exhaustives
  - corrélation négative entre la taille des fratries et le nombre d'années de scolarisation moyen des enfants

Solution

4) Calculer quelques statistiques descriptives de la **taille de la fratrie** (`family_size`) et du **nombre d'années d'études** (`education`):

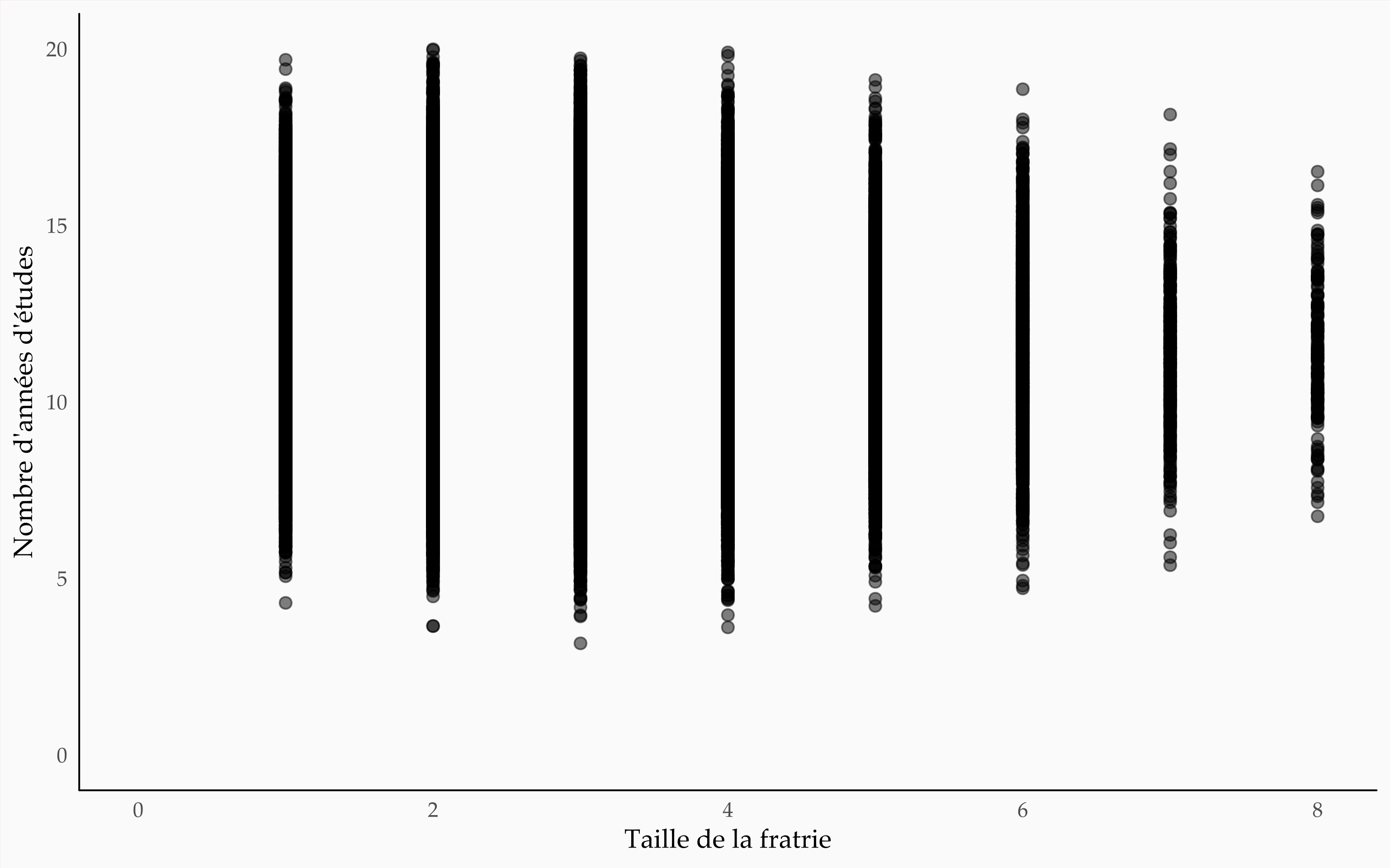
- moyenne
- écart-type et variance
- corrélation entre les deux variables

```
df %>%
  summarise(
    across(c(family_size, education), ~ round(mean(.), 2), .names = "mean_{•col}"),
    across(c(family_size, education), ~ round(sd(.), 2), .names = "sd_{•col}"),
    across(c(family_size, education), ~ round(var(.), 2), .names = "var_{•col}"),
    correlation = cor(family_size, education, use = "complete.obs")
  )
```

```
## # A tibble: 1 × 7
##   mean_family_size mean_education sd_family_size sd_education var_family_size
##              <dbl>          <dbl>          <dbl>        <dbl>           <dbl>
## 1             2.71           12.0           1.08         2.02            1.16
## # ℹ 2 more variables: var_education <dbl>, correlation <dbl>
```

Solution

5) Représenter le nuage de points qui définit la relation entre **taille de la fratrie** et **nombre d'années d'études**



Solution

6) Calculer les estimateurs de β_0 et β_1 du modèle $\textbf{Nb Années études} = \beta_0 + \beta_1 \textbf{Taille Fratrie} + \varepsilon$, à la main et directement via la fonction `lm`

```r
# 1: Calcul à la main
beta_1 = cov(df$family_size, df$education)/var(df$family_size)
beta_1
```

```
## [1] -0.1003702
```

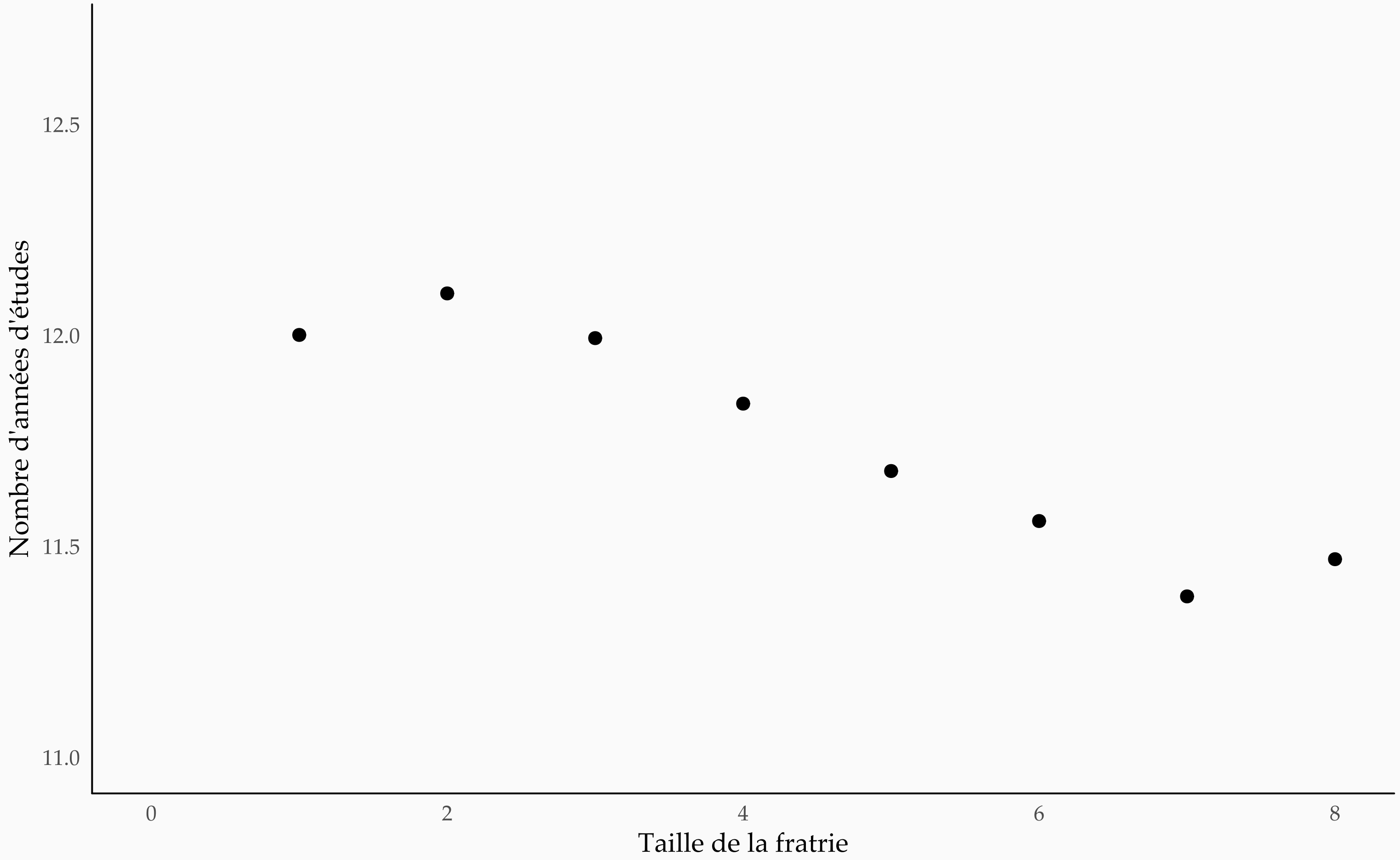
```r
beta_0 = mean(df$education) - beta_1*mean(df$family_size)
beta_0
```

```
## [1] 12.26441
```

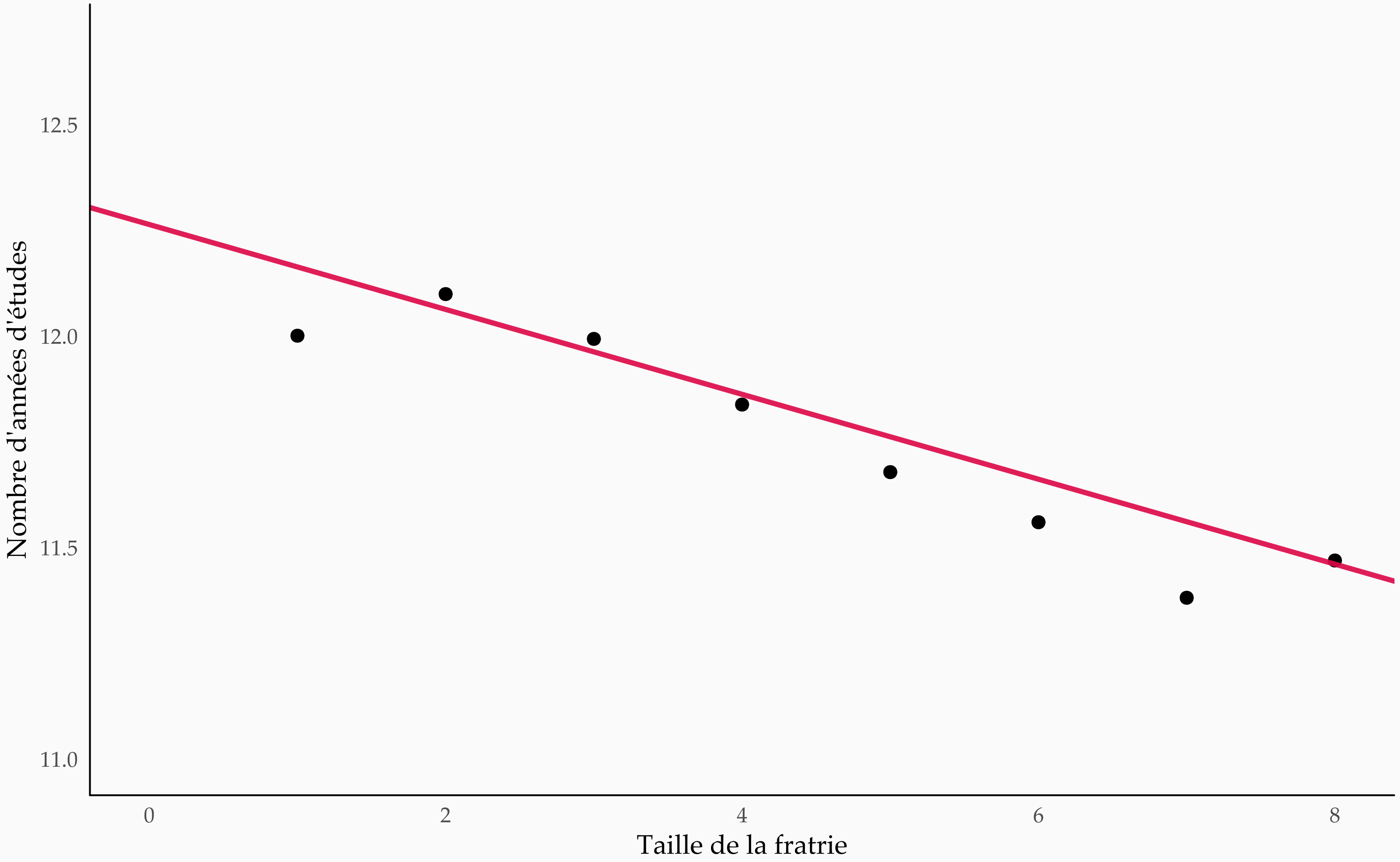
```r
# 2: Via lm
summary(lm(education ~ family_size, data = df))$coefficients
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 12.2644114 0.013739444 892.64249  0.00000e+00
## family_size -0.1003702 0.004707536 -21.32118 1.00352e-100
```

Taille de fratrie et nombre d'années d'études



Taille de fratrie et nombre d'années d'études



Taille de fratrie et nombre d'années d'études

Question: Que se passe t-il quand on considère `family_size` comme une variable discrète et non continue?

Taille de fratrie et nombre d'années d'études

Question: Que se passe t-il quand on considère `family_size` comme une variable discrète et non continue?

hint: utiliser `as.factor(family_size)`

Taille de fratrie et nombre d'années d'études

Question: Que se passe t-il quand on considère `family_size` comme une variable discrète et non continue?

hint: utiliser `as.factor(family_size)`

Question: Est-ce que $\hat{\beta}_1$ représente l'**effet causal** de la taille de la fratrie sur le nombre d'années d'études?

Taille de fratrie et nombre d'années d'études

Question: Que se passe t-il quand on considère `family_size` comme une variable discrète et non continue?

hint: utiliser `as.factor(family_size)`

Question: Est-ce que $\hat{\beta}_1$ représente l'**effet causal** de la taille de la fratrie sur le nombre d'années d'études?

Black, Devereux and Salvanes (2005) : montrent qu'il existe un autre prédicteur de la performance scolaire, corrêlé à la taille de la fratrie

Taille de fratrie et nombre d'années d'études

Question: Que se passe t-il quand on considère `family_size` comme une variable discrète et non continue?

hint: utiliser `as.factor(family_size)`

Question: Est-ce que $\hat{\beta}_1$ représente l'**effet causal** de la taille de la fratrie sur le nombre d'années d'études?

Black, Devereux and Salvanes (2005) : montrent qu'il existe un autre prédicteur de la performance scolaire, corrêlé à la taille de la fratrie

le rang de naissance

Recap: Régression linéaire simple

Data: Données observationnelles

Hypothèse d'identification: $\mathbb{E}(\varepsilon_i|x) = 0$, i.e. x n'est pas corrêlée au terme d'erreur ε

- dit autrement, x est exogène, i.e. il n'y a pas de variable omise/biais de sélection

Modèle: pour tout individu i,

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Estimateur de l'effet causal de x sur Y:

$$\hat{\beta}_1 = \frac{Cov(x, y)}{\mathbb{V}(x)}$$

Implémentation sur `R`

- `lm` pour estimer les paramètres du modèle
- `summary` pour afficher le résultat de l'estimation
- `coeftest`, argument `vcov = vcovHC(fit, type = 'HC0')` pour obtenir des se robustes à l'hétéroscédasticité
- `stargazer` ou `modelsummary` pour exporter les résultats en une table $\LaTeX$

2. Causalité

2. Causalité

2.1. Corrélation *vs* Causalité

L'hypothèse d'exogeneité implique que l'estimateur des MCO est non biaisé

- donc que $\hat{\beta}$ représente l'effet causal de x sur y

Cependant, cette hypothèse est très forte et rarement vérifiée

Biais de Variable Omise/Biais de sélection (*Omitted Variable Bias - OVB*) : il existe un biais de variable omise lorsqu'une variable qui n'est pas inclue dans le set de variables explicatives (et donc $\in \varepsilon$),
1) affecte y
2) est correlée à x_i

Exemple canonique de l'équation de Mincer: on cherche à estimer l'effet d'une année de scolarisation supplémentaire sur le salaire: $\text{Wage} = \alpha + \beta \text{Education} + \varepsilon$

- Problème: l'abilité, la motivation, ne sont pas observées
- Dans ce cas, le paramètre β n'estime pas l'effet *causal* de l'éducation sur le salaire, mais une *corrélation*

2.2. Potential Outcomes Framework

Neyman, 1923 & **Rubin, 1974**: cadre conceptuel qui aide à penser la causalité. On s'intéresse à la relation entre deux variables:

- une variable d'outcome Y_i
- une variable de traitement (que l'on suppose binaire par simplicité),

$$D_i = \begin{cases} 1 \text{ si l'individu } i \text{ est traîté} \\ 0 \text{ si l'individu } i \text{ n'est pas est traîté} \end{cases}$$

On cherche estimer l'**effet de D_i sur Y_i**, par exemple:

- l'effet d'avoir un master sur le salaire (*returns to education*)
- l'effet d'une peine de prison sur la probabilité de récidive
- l'effet d'un médicament sur la santé d'un patient
- l'effet d'appartenir à une fratrie de plus de 2 enfants sur la réussite scolaire

2.2. Potential Outcomes Framework

Chaque individu i a deux outcomes potentiels:

- Y_{1i} si $D_i = 1$, l'outcome en cas de traitement
- Y_{0i} si $D_i = 0$, l'outcome en l'absence de traitement

L'effet causal du traitement pour chaque individu i est simplement la différence entre l'outcome en cas de traitement et l'outcome en l'absence de traitement:

$$\delta_i = Y_{1i} - Y_{0i}$$

L'espérance des δ_i donne l'effet **moyen** du traitement (**A**verage **T**reatment **E**ffect):

$$\text{ATE} = \mathbb{E}(\delta_i) = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{0i})$$

2.2. Potential Outcomes Framework

🚨 **Problème fondamental de l'inférence causale: il n'est pas possible d'observer à la fois Y_{1i} et Y_{0i}** 🚨

On observe uniquement $Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i)$:

- quand l'individu i est traîté, i.e. $D_i = 1$, on observe uniquement $Y_i = Y_{1i}$
- quand l'individu i n'est pas traîté, i.e. $D_i = 0$, on observe uniquement $Y_i = Y_{0i}$

$\implies$ on observe **deux groupes**: le groupe des individus **traités** et le groupe des individus **non traîtés (ou témoins ou contrôles)**

2.2. Potential Outcomes Framework

Question: que peut-on faire à partir des données que l'on observe sur ces deux groupes ?

2.2. Potential Outcomes Framework

Question: que peut-on faire à partir des données que l'on observe sur ces deux groupes ?

Réponse: calculer la différence entre l'outcome moyen des individus traîtés et l'outcome moyen des individus non traîtés,

$$\Delta = \mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)$$

2.2. Potential Outcomes Framework

Question: est-ce que Δ = ATE, l'effet causal moyen du traitement ?

2.2. Potential Outcomes Framework

Question: est-ce que Δ = ATE, l'effet causal moyen du traitement ?

Réponse: si le traitement n'est pas corrêlé à l'outcome

Intuition:

- si le **le traitement est indépendant de l'outcome**, i.e. si le groupe de contrôle est comparable au groupe de traîtés, ou formellement si $(Y_{1i}, Y_{0i}) \perp D_i$,
- alors il n'y a pas de biais de sélection dans le traitement,
- donc **la différence entre l'outcome moyen du groupe des individus traîtés et celui des individus non traîtés estime l'effet causal du traitement (ATE)**:

$$\Delta = \mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0) = \text{ATE} \ \text{ iff } \ (Y_{1i}, Y_{0i}) \perp D_i$$

Maths

2.2. Potential Outcomes Framework

Problème : $\mathbb{E}(Y_{0i}|D_i = 1) = \mathbb{E}(Y_{0i}|D_i = 0)$ (= absence de sélection) est une hypothèse forte. Souvent les groupes traités et non traités ne sont pas comparables (étudiants qui décident de faire un master sont peut-être plus motivés, les individus incarcérés sûrement plus dangereux, etc, **inobservable**).

Différents *types* de sélection:

- **Auto-selection**:
    - si les gains espérés du traitement sont corrélés à l'outcome
    - si les coûts liés au traitement sont hétérogènes
- **Selection par les entités qui délivrent le traitement**:
    - si seuls les individus ayant un outcome initial faible sont traîtés
    - ou l'inverse

Application

04:00

Voici le code pour simuler un jeu de données. Il comprend 10 000 individus, deux variables d'outcome potentiel Y_0 et Y_1, une variable de traitement D:

```
set.seed(123) # pour la reproductibilité
n = 10000 # nombre d'individus
ATE = 3

# Outcomes potentiels
Y0 = rnorm(n, mean = 10, sd = 2)  # outcome potentiel en cas de traitement
Y1 = Y0 + rnorm(n, mean = ATE, sd = 1)  # outcome potentiel en l'absence de traitement

# Traitement non aléatoire
D = ifelse(Y0 > median(Y0), 1, 0)

# Outcome observé
Y = Y1*D + Y0*(1-D)
```

Calculer:

- Δ
- l'ATT
- le biais de sélection

Solution

```
head(data.frame(Y0, Y1, D, Y))
```

```
##          Y0       Y1 D         Y
## 1  8.879049 14.24977 0  8.879049
## 2  9.539645 12.37283 0  9.539645
## 3 13.117417 17.04438 1 17.044378
## 4 10.141017 12.57287 1 12.572865
## 5 10.258575 13.48367 1 13.483666
## 6 13.430130 17.56212 1 17.562116
```

Solution

```r
SD = mean(Y1[D == 1]) - mean(Y0[D == 0])
ATT = mean(Y1[D==1] - Y0[D==1])
bias = mean(Y0[D == 1]) - mean(Y0[D == 0])

# Afficher les résultats
data.frame(
  Delta = SD,
  ATT = ATT,
  Biais_de_selection = bias
)
```

```
##      Delta      ATT Biais_de_selection
## 1 6.177746 2.995577           3.182169
```

3. Randomized Controlled Trials (RCTs)

3. Randomized Controlled Trials (RCTs)

3.1. Suppression du biais de sélection

La **randomisation** permet d'éliminer le biais de sélection en allouant aléatoirement les individus au groupe de contrôle et au gorupe de traitement.

Formellement, cela signifie que $(Y_{1i}, Y_{0i}) \perp D_i \implies \mathbb{E}(Y_{0i}|D_i = 1) = \mathbb{E}(Y_{0i}|D_i = 0)$.

Donc,

$$\begin{aligned}\text{Biais de Sélection} &= \mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0) \\ &= 0\end{aligned}$$

Les expériences aléatoires contrôlées, ou ***Randomized Controlled Trials (RCT)***, permettent ainsi d'estimer l'effet causal d'un traitement

- Très répandues en médecine
- De plus en plus répandues (et reconnues) en économie pour l'évaluation des politiques publiques (Prix Nobel par Esther Duflo, Abhijit Banerjee et Michael Kremer en 2019)

3.2. Exemple: le projet STAR

Krueger, A. B. "Experimental estimates of education production functions.", QJE 1999

Le projet **STAR** (*Student-Teacher Achievement Ratio*) est un exemple très connu d'expériementation qui a pour but d'estimer l'**effet causal de la taille des classes sur les performances scolaires des élèves**.

Principaux éléments:

- 11 600 élèves de l'état du Tennessee ont participé à l'expérimentation
- l'expériementation a débutée l'année scolaire 1985-1986 et concerne des élèves de la GS au CE2
- **Trois groupes de traitement**:
  - assignement à une classe de petite taille, de 13 à 17 élèves
  - assignement à une classe de taille moyenne, de 22 à 35 élèves (= groupe de contrôle)
  - assignement à une classe de taille moyenne, de 22 à 35 élèves + aide d'un professeur à temps plein
- élèves et enseignants répartis aléatoirement, à l'échelle d'une école, dans ces trois types de classes
- chaque année les compétences en maths et lecture des élèves sont évaluées

Application: le projet STAR

05:00

La réplication des résultats de cette expérimentation est possible grâce à la mise à disposition des données directement sur `R`, à l'aide du package `AER` (pour Applied Econometrics with `R`). [Variables details]

```
#install.packages("AER")
library(AER)

data(STAR)

head(STAR)
```

```
##      gender ethnicity   birth        stark star1        star2        star3
## 1122 female      afam 1979 Q3         <NA>  <NA>         <NA>      regular
## 1137 female      cauc 1980 Q1        small small        small        small
## 1143 female      afam 1979 Q4        small small regular+aide regular+aide
## 1160   male      cauc 1979 Q4         <NA>  <NA>         <NA>        small
## 1183   male      afam 1980 Q1 regular+aide  <NA>         <NA>         <NA>
## 1195   male      cauc 1979 Q3         <NA>  <NA>      regular      regular
##      readk read1 read2 read3 mathk math1 math2 math3   lunchk lunch1   lunch2
## 1122    NA    NA    NA   580    NA    NA    NA   564     <NA>   <NA>     <NA>
## 1137   447   507   568   587   473   538   579   593 non-free   free non-free
## 1143   450   579   588   644   536   592   579   639 non-free   <NA> non-free
## 1160    NA    NA    NA   686    NA    NA    NA   667     <NA>   <NA>     <NA>
## 1183   439    NA    NA    NA   463    NA    NA    NA     free   <NA>     <NA>
## 1195    NA    NA    NA   644    NA    NA    NA   648     <NA>   <NA> non-free
```

Application: le projet STAR

Intuition:

1) En l'absence de randomisation, pourquoi simplement comparer les résultats moyens des élèves de petites classes et de classes de taille moyenne ne suffit pas à estimer un effet causal de la taille des classes ?

Code

NB: Aide pour calculer les percentiles.

2) Représenter graphiquement la densité de `avg_perc` pour le groupe *small* et le groupe *regular + regular-with-aide* (Reproduction de la Figure 1, Panel Kindergarten, page 509)

3) Estimer les paramètres du modèle : $Y_i = \beta_0 + \beta_1 \mathbf{Small}_i + \beta_2 \mathbf{RegularAide}_i + \varepsilon_i$ (Reproduction de la Table 5, Colonne 1 page 512)

Data Cleaning

In each grade level the regular and regular/aide students were pooled together, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A separate percentile distribution was generated for each subject test (e.g. Math-STA, Reading-SAT, Word-SAT, etc). For each test I then determined where in the distribution of the regular-class students every student in the small classes would fall, and students in the small classes were assigned these percentiles scores. Finally, to summarize overall achievement, the average of the three SAT percentile rankings was calculated.

```r
STAR = STAR %>%
  mutate(
    group = ifelse(as.character(stark) == "regular+aide", "regular", as.character(stark)),
    readk_perc = case_when(
      group == "regular" ~ ecdf(readk[group == "regular"])(readk) * 100,
      group == "small"   ~ ecdf(readk[group == "regular"])(readk) * 100),
    mathk_perc = case_when(
      group == "regular" ~ ecdf(mathk[group == "regular"])(mathk) * 100,
      group == "small"   ~ ecdf(mathk[group == "regular"])(mathk) * 100),
    avg_perc = (readk_perc + mathk_perc)/2
  )
```

Data Cleaning

*In each grade level the **regular and regular/aide students were pooled together**, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A separate percentile distribution was generated for each subject test (e.g. Math-STA, Reading-SAT, Word-SAT, etc). For each test I then determined where in the distribution of the regular-class students every student in the small classes would fall, and students in the small classes were assigned these percentiles scores. Finally, to summarize overall achievement, the average of the three SAT percentile rankings was calculated.*

```
STAR = STAR %>%
  mutate(
    group = ifelse(as.character(stark) == "regular+aide", "regular", as.character(stark)),
    readk_perc = case_when(
      group == "regular" ~ ecdf(readk[group == "regular"])(readk) * 100,
      group == "small"   ~ ecdf(readk[group == "regular"])(readk) * 100),
    mathk_perc = case_when(
      group == "regular" ~ ecdf(mathk[group == "regular"])(mathk) * 100,
      group == "small"   ~ ecdf(mathk[group == "regular"])(mathk) * 100),
    avg_perc = (readk_perc + mathk_perc)/2
  )
```

Data Cleaning

In each grade level the regular and regular/aide students were pooled together, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A ***separate percentile distribution was generated for each subject test (e.g. Math-STA, Reading-SAT, Word-SAT, etc). For each test I then determined where in the distribution of the regular-class students every student in the small classes would fall****, and students in the small classes were assigned these percentiles scores. Finally, to summarize overall achievement, the average of the three SAT percentile rankings was calculated.*

```
STAR = STAR %>%
  mutate(
    group = ifelse(as.character(stark) == "regular+aide", "regular", as.character(stark)),
    readk_perc = case_when(
      group == "regular" ~ ecdf(readk[group == "regular"])(readk) * 100,
      group == "small"   ~ ecdf(readk[group == "regular"])(readk) * 100),
    mathk_perc = case_when(
      group == "regular" ~ ecdf(mathk[group == "regular"])(mathk) * 100,
      group == "small"   ~ ecdf(mathk[group == "regular"])(mathk) * 100),
    avg_perc = (readk_perc + mathk_perc)/2
  )
```

Data Cleaning

In each grade level the regular and regular/aide students were pooled together, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A separate percentile distribution was generated for each subject test (e.g. Math-STA, Reading-SAT, Word-SAT, etc). For each test I then determined where in the distribution of the regular-class students every student in the small classes would fall, and ***students in the small classes were assigned these percentiles scores****. Finally, to summarize overall achievement, the average of the three SAT percentile rankings was calculated.*

```
STAR = STAR %>%
  mutate(
    group = ifelse(as.character(stark) == "regular+aide", "regular", as.character(stark)),
    readk_perc = case_when(
      group == "regular" ~ ecdf(readk[group == "regular"])(readk) * 100,
      group == "small"   ~ ecdf(readk[group == "regular"])(readk) * 100),
    mathk_perc = case_when(
      group == "regular" ~ ecdf(mathk[group == "regular"])(mathk) * 100,
      group == "small"   ~ ecdf(mathk[group == "regular"])(mathk) * 100),
    avg_perc = (readk_perc + mathk_perc)/2
  )
```

Data Cleaning

In each grade level the regular and regular/aide students were pooled together, and students were assigned percentile scores based on their raw test scores, ranging from 0 (lowest score) to 100 (highest score). A separate percentile distribution was generated for each subject test (e.g. Math-STA, Reading-SAT, Word-SAT, etc). For each test I then determined where in the distribution of the regular-class students every student in the small classes would fall, and students in the small classes were assigned these percentiles scores. Finally, to ***summarize overall achievement, the average of the three SAT percentile rankings was calculated****.*

```
STAR = STAR %>%
  mutate(
    group = ifelse(as.character(stark) == "regular+aide", "regular", as.character(stark)),
    readk_perc = case_when(
      group == "regular" ~ ecdf(readk[group == "regular"])(readk) * 100,
      group == "small"   ~ ecdf(readk[group == "regular"])(readk) * 100),
    mathk_perc = case_when(
      group == "regular" ~ ecdf(mathk[group == "regular"])(mathk) * 100,
      group == "small"   ~ ecdf(mathk[group == "regular"])(mathk) * 100),
    avg_perc = (readk_perc + mathk_perc)/2
  )
```

Solution

Intuition

1) En l'absence de randomisation, pourquoi simplement comparer les résultats des élèves de petites classes et de classes de taille moyenne ne suffit pas à estimer un effet causal de la taille des classes ?

Solution

Intuition

1) En l'absence de randomisation, pourquoi simplement comparer les résultats des élèves de petites classes et de classes de taille moyenne ne suffit pas à estimer un effet causal de la taille des classes ?

Within School sorting: L'allocation des élèves et professeurs **au sein** des écoles n'est pas aléatoire

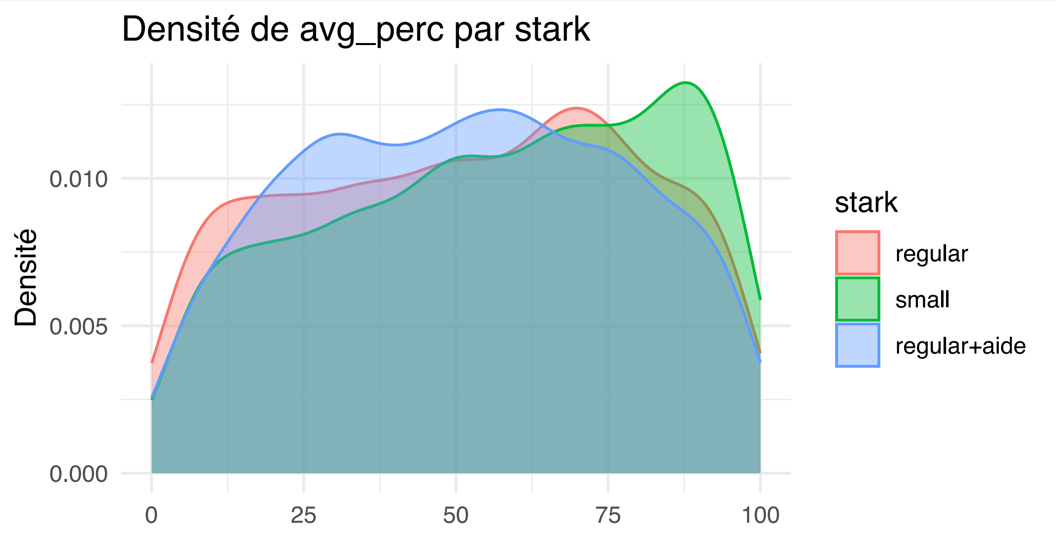
- les élèves les plus en difficultés peuvent-être volontairement placés dans des classes plus petites
  - auquel cas l'abilité de l'élève, que l'on observe pas, est correlée à la taille des classes, et est également intrinsèquement liée à ses performances scolaires
  - donc l'effet de la taille des classes peut refléter l'effet de l'abilité initiale
- les enseignants les plus expérimentés peuvent vouloir préférer enseigner dans les classes les plus petites au sein des écoles

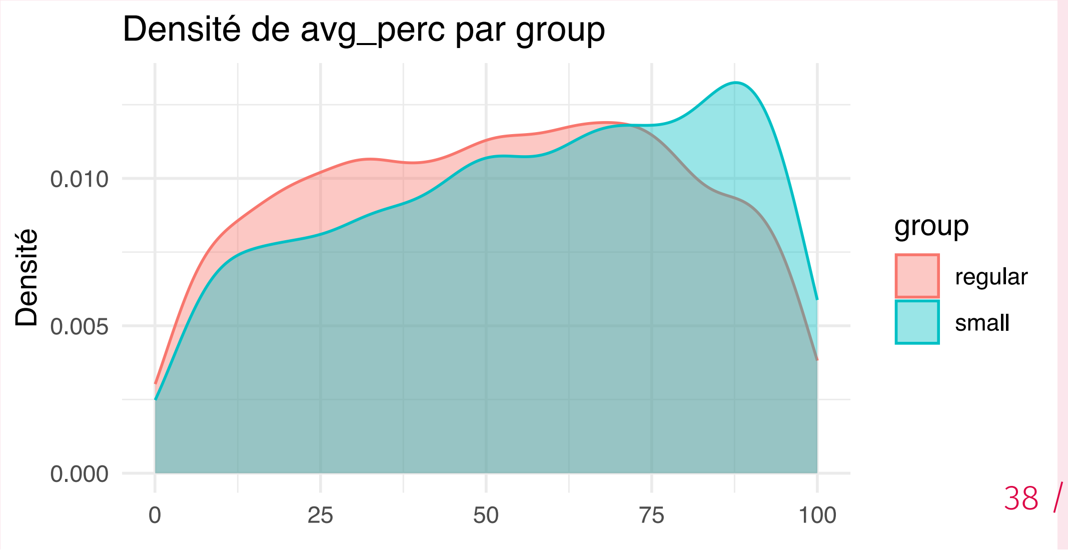
Solution

2) Représenter graphiquement la densité de `avg_perc` pour le groupe *small* et le groupe *regular + regular-with-aide* (Reproduction de la Figure 1, Panel Kindergarten, page 509)

```
STAR %>%
  ggplot(aes(x = avg_perc, color = stark, fill = stark)) +
  geom_density(alpha = 0.4) +
  labs(title = "Densité de avg_perc par stark",
       x = "",
       y = "Densité") +
  theme_minimal()
```

```
STAR %>%
  ggplot(aes(x = avg_perc, color = group, fill = group)) +
  geom_density(alpha = 0.4) +
  labs(title = "Densité de avg_perc par group",
       x = "",
       y = "Densité") +
  theme_minimal()
```





Solution

3) Estimer les paramètres du modèle : $Y_i = \beta_0 + \beta_1 \textbf{Small}_i + \beta_2 \textbf{RegularAide}_i + \varepsilon_i$ (Reproduction de la Table 5, Colonne 1 page 512)

```
STAR = STAR %>%
  mutate(smallk = ifelse(as.character(stark == "small"), 1, 0),
         regularaidek = ifelse(as.character(stark == "regular+aide"), 1, 0))

summary(lm(avg_perc ~ smallk + regularaidek, data = STAR)) # coeftest(lm(avg_perc ~ smallk + regularaidek, data = STA
```

```
##
## Call:
## lm(formula = avg_perc ~ smallk + regularaidek, data = STAR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.484 -22.186   1.383  22.693  48.677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   51.4352     0.6033  85.249  < 2e-16 ***
## smallk         4.9178     0.8854   5.554 2.91e-08 ***
## regularaidek  -0.1125     0.8493  -0.133    0.895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Randomized Controlled Trials (RCTs)

3.3. Limites

Bien qu'il s'agisse d'une stratégie empirique très ***clean***, les expériences aléatoires comportent des limites:

- **Coût**:
  - Financier: coût de mise en place de la politique (ex: projet STAR = $12 million)
  - Durée: design de l'expérimentation, validation par l'ERB, pilote, durée du traitement, analyse des résultats (ex: projet STAR a duré 4 ans)
- **Validité externe**: les expérimentations sont souvent réalisées à une échelle très locale. Dans quelle mesure les résultats se généralisent à d'autres contextes? Quid du passage à l'échelle?
- Ethique:
  - une partie de la population "privée" du traitement
  - Quel accompagnement après le traitement?

Recap: Randomized Controlled Trial

Data: Données expérimentales

Hypothèse d'identification:

- Intuition: allocation aléatoire du statut de traitement
- Formellement: $(Y_{1i}, Y_{0i}) \perp D_i$

Modèle: pour tout individu i,

$$Y_i = \alpha + \delta D_i + \varepsilon$$

Estimateur de l'effet du traitement:

- Différence entre l'outcome moyen du groupe des individus traîtés et celui du groupe de contrôle
- $\hat{\delta} = \mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)$

Implémentation sur `R`:

- Balancing Tests: test de différences de moyennes
- Estimation de l'effet du traitement: `lm(y ~ D, data = data)`

Application

02:00

Voici le code utilisé dans la précédente application auquel on ajoute une variable `D_random` qui distribue aléatoirement le traitement D.

```
set.seed(123) # pour la reproductibilité
n = 10000 # nombre d'individus
ATE = 3
Y0 = rnorm(n, mean = 10, sd = 2)  # outcome potentiel en cas de traitement
Y1 = Y0 + rnorm(n, mean = ATE, sd = 1)  # outcome potentiel en l'absence de traitement
D = ifelse(Y0 > median(Y0), 1, 0)
Y = Y1*D + Y0*(1-D)

# Traitement aléatoire
D_random = rbinom(n, 1, 0.5)
Y_random = Y1*D_random + Y0*(1-D_random)
```

Calculer:

- Δ avec `D`
- l'ATT
- le biais de sélection
- Δ avec `D_random`

Solution

```
SDO = mean(Y[D == 1]) - mean(Y[D == 0])
ATT = mean(Y1[D_random == 1] - Y0[D_random == 1])
bias = mean(Y0[D_random == 1]) - mean(Y0[D_random == 0])
SD_random = mean(Y1[D_random == 1]) - mean(Y0[D_random == 0])

# Afficher les résultats
data.frame(
  Delta = SD,
  ATT = ATT,
  Biais_de_selection = bias,
  Random_diff = SD_random
)
```

```
##      Delta      ATT Biais_de_selection Random_diff
## 1 6.177746 2.995971        -0.04824477    2.947726
```

Sources

Econometrics with R

Project STAR: Student-Teacher Achievement Ratio in AER

Causal inference: The Mixtape, Scott Cunningham

Florian Oswald

Edward Rubin

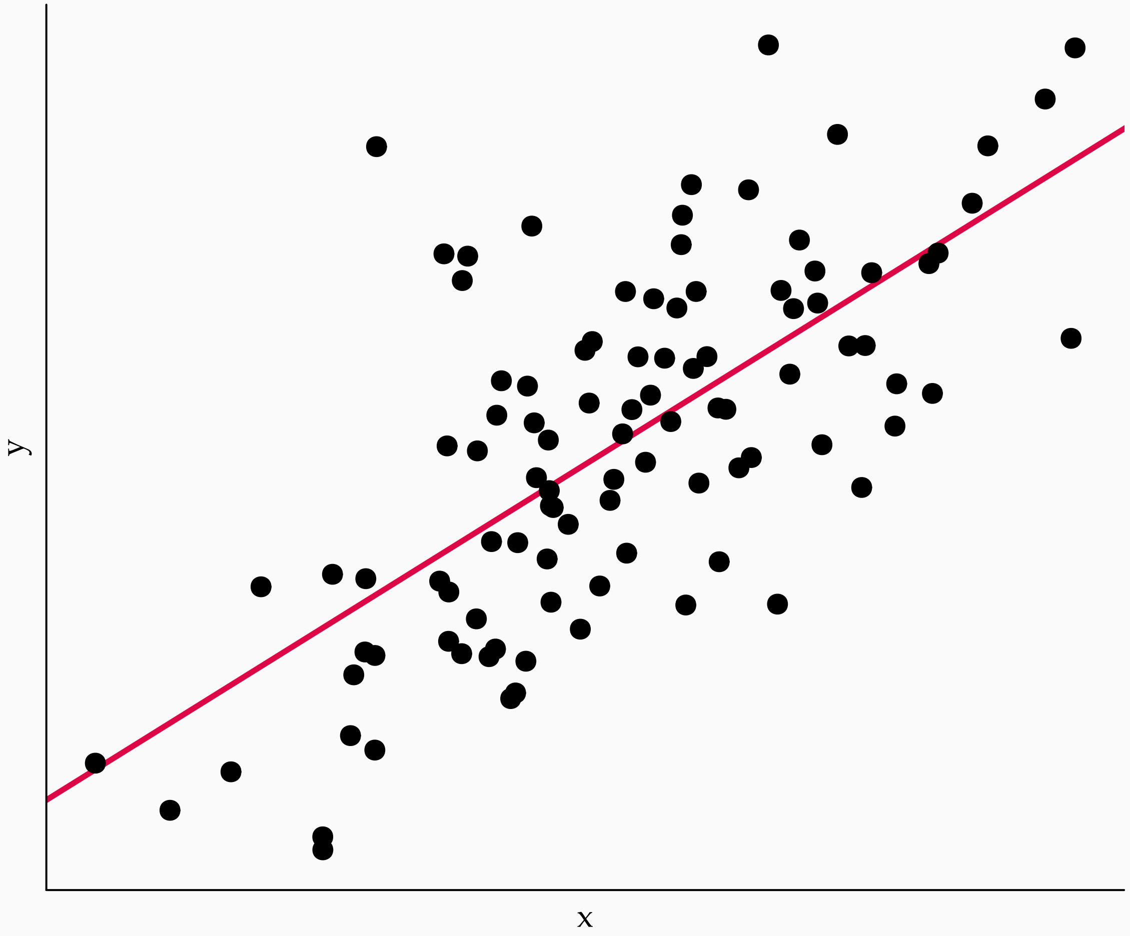
Scott Cunningham

Scientific Research and Methodology, Peter K. Dunn

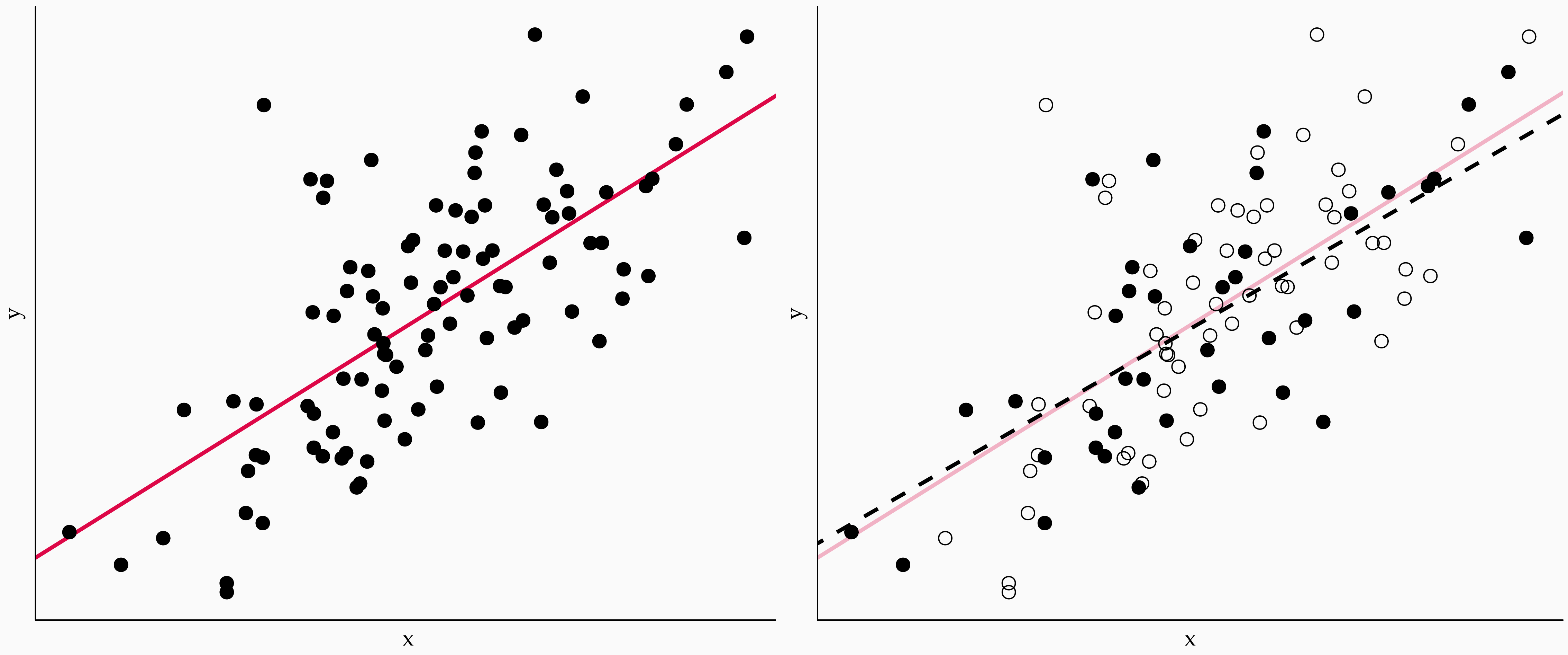
Économétrie: méthodes et applications. Bruno Crépon et Nicolas Jacquemet

Annexe

$\hat{\beta}$ est une variable aléatoire

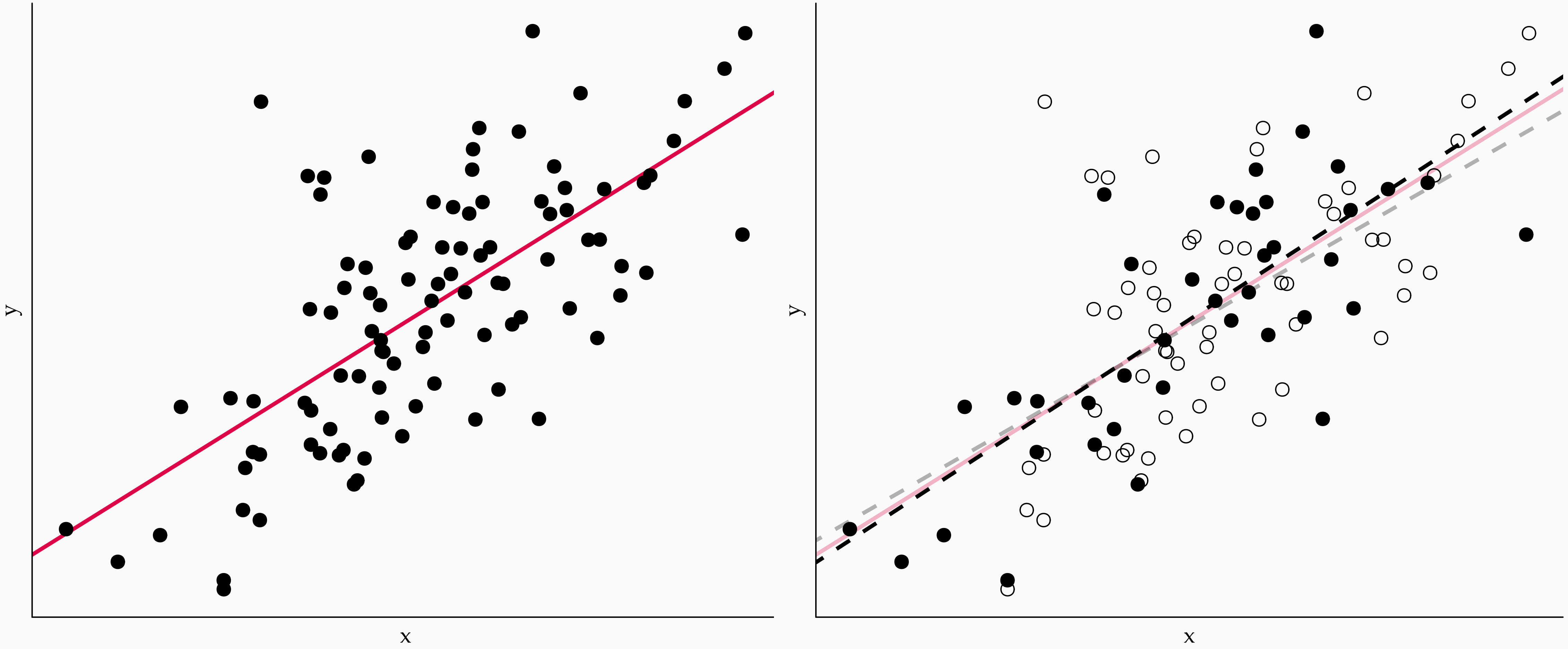


$\hat{\beta}$ est une variable aléatoire



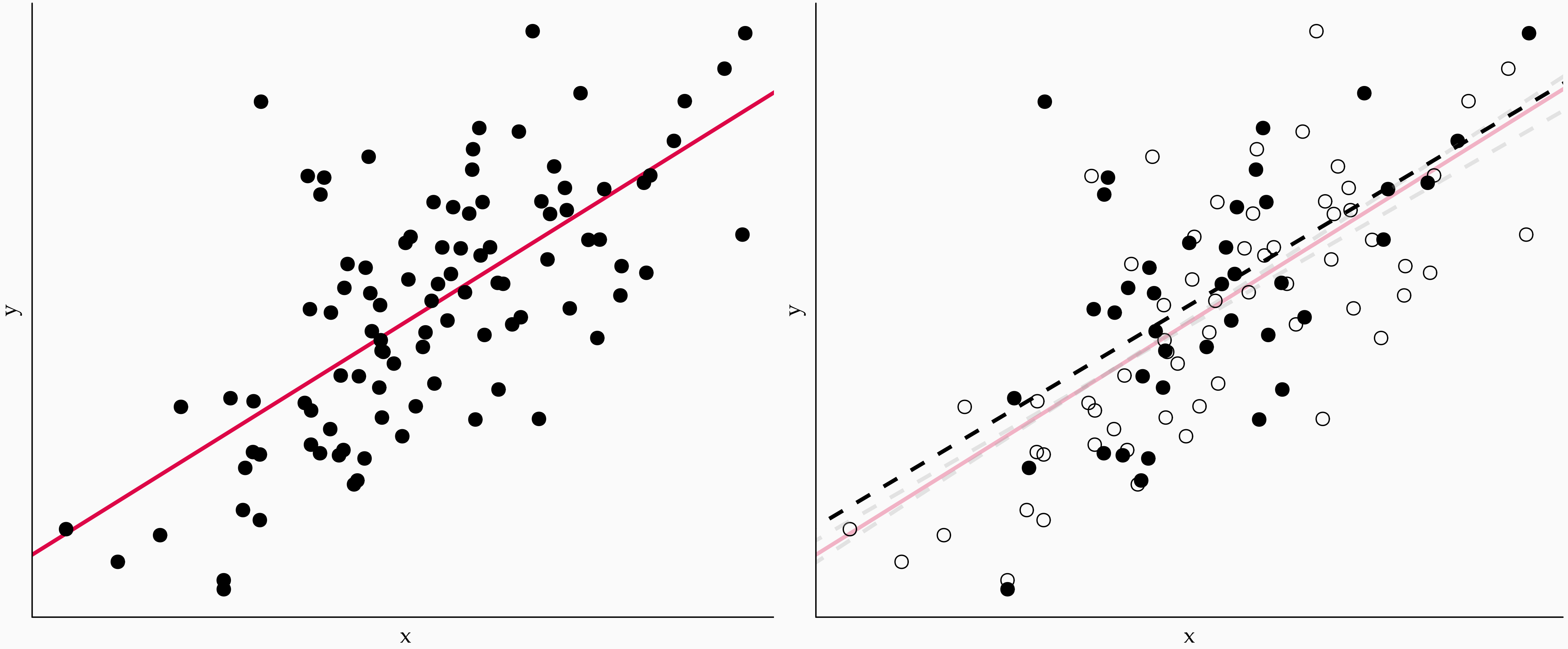
[Tiré du cours d'Edward Rubin]

$\hat{\beta}$ est une variable aléatoire



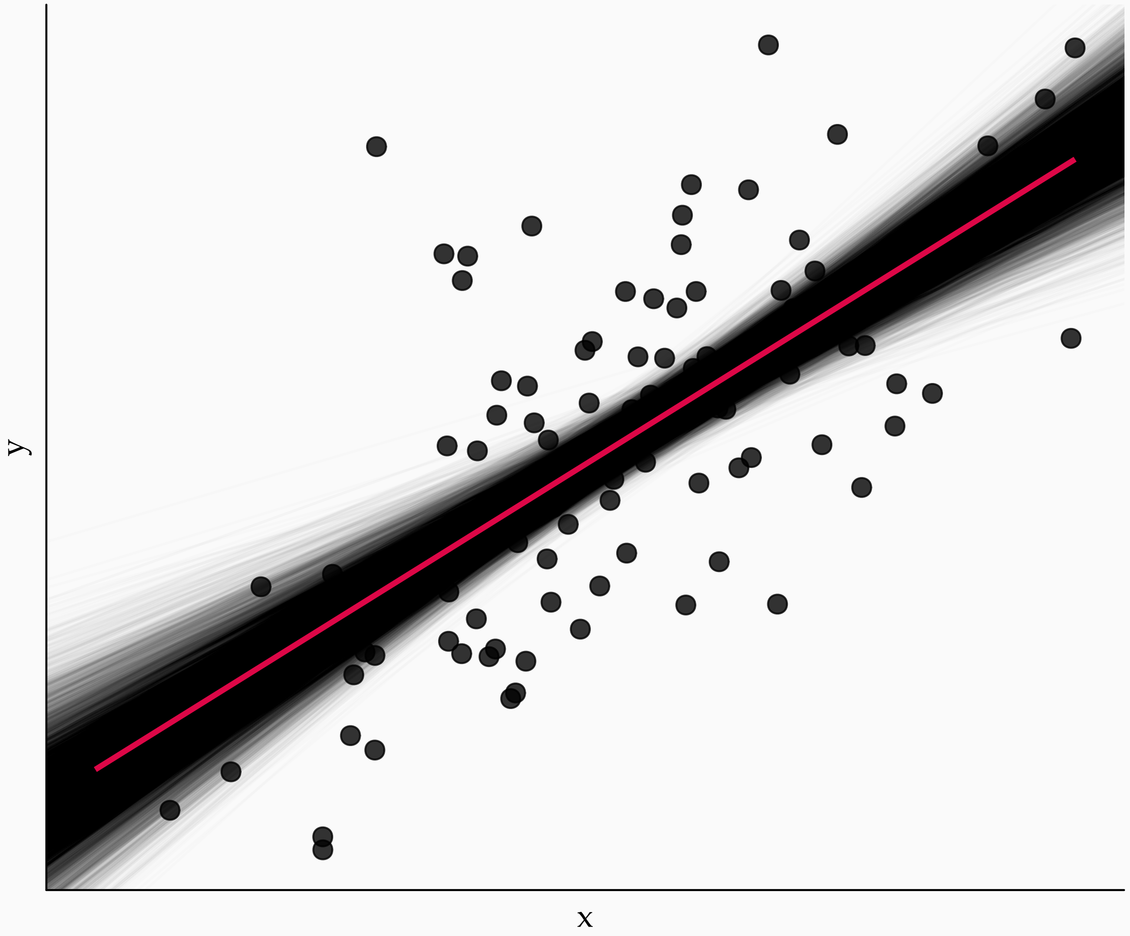
[Tiré du cours d'Edward Rubin]

$\hat{\beta}$ est une variable aléatoire



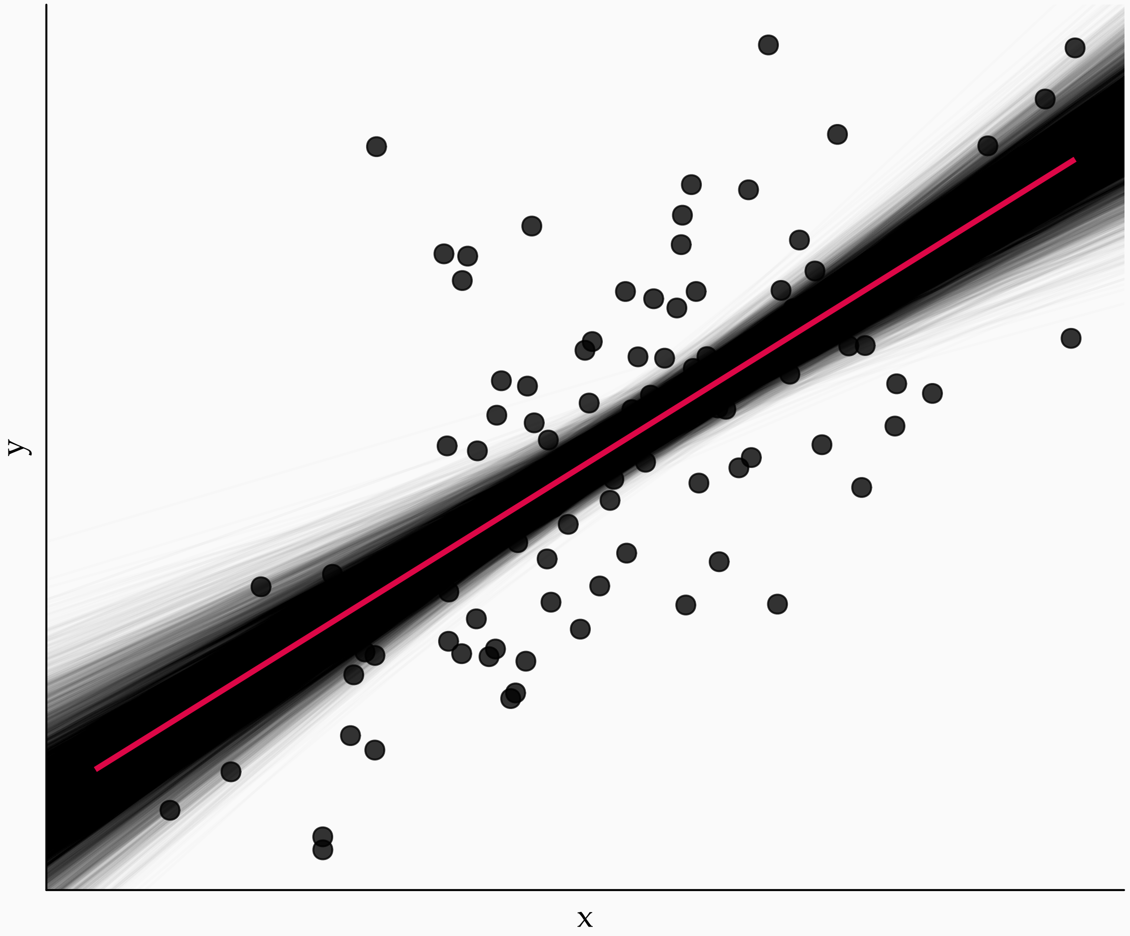
[Tiré du cours d'Edward Rubin]

$\hat{\beta}$ est une variable aléatoire



[Tiré du cours d'Edward Rubin]

$\hat{\beta}$ est une variable aléatoire



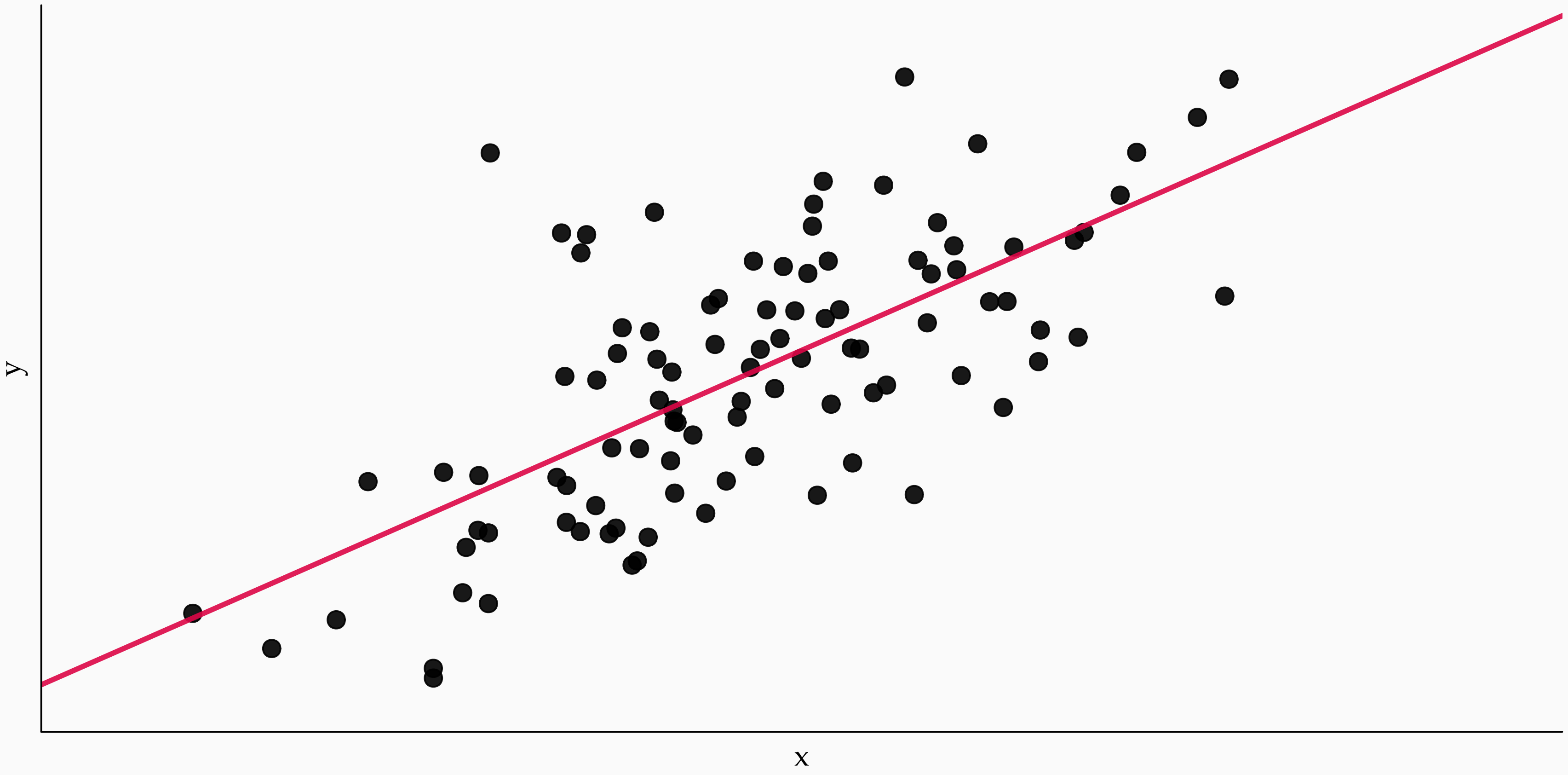
- En **moyenne**, les droites de régressions sur les échantillons sont très proches de la droite de régression sur l'ensemble de la population
- Mais certaines en sont très éloignées
- **$\hat{\beta}$ est une variable aléatoire : sa valeur est propre à l'échantillon sur lequel il est estimé**

$\implies$ Tout l'enjeu pour l'économètre est d'assurer que l'échantillon est aléatoire et/ou représentatif de telle sorte à ce que $\hat{\beta}$ soit proche de β

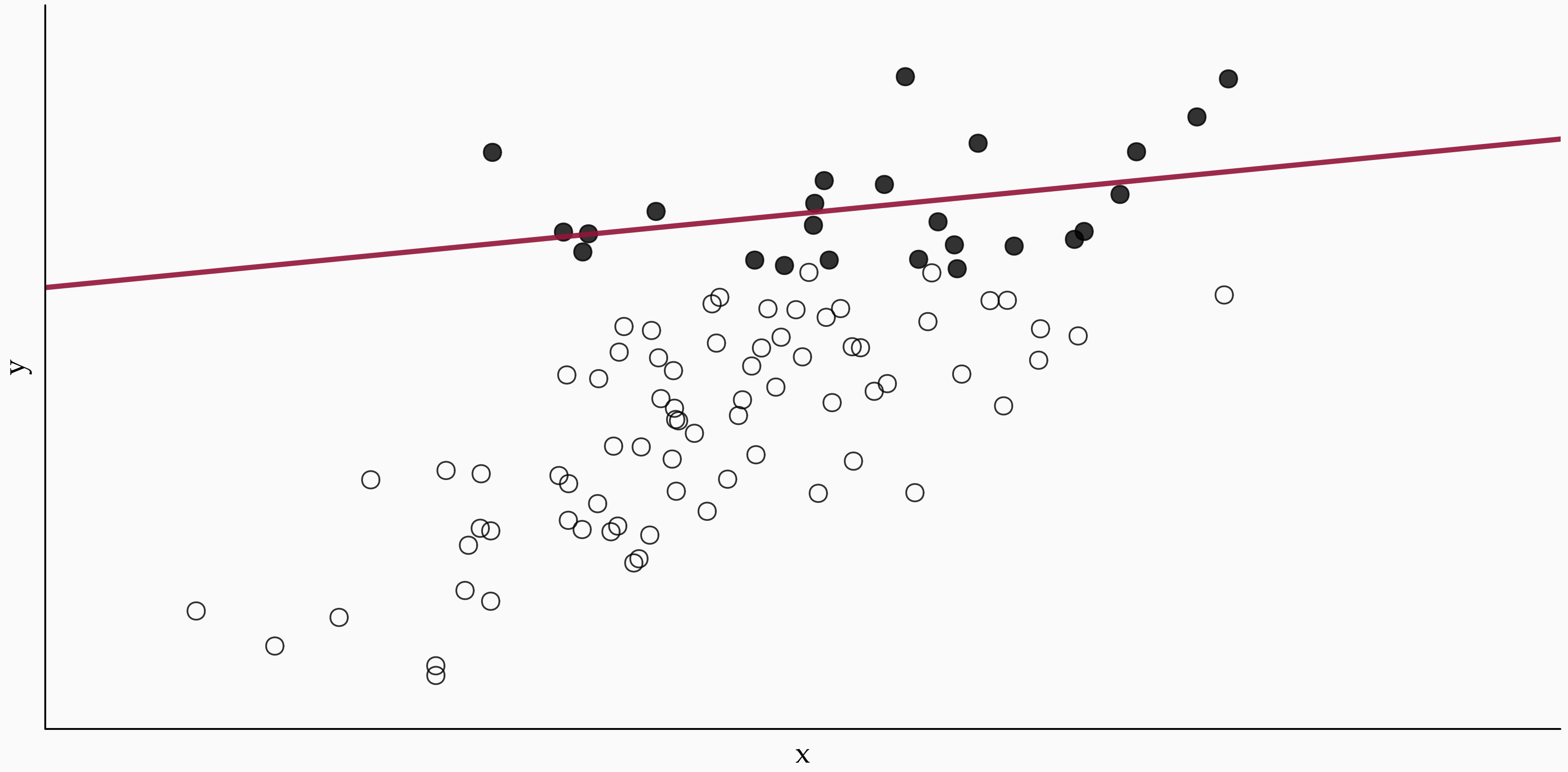
🚨 Bien lire la description de l'échantillon et utiliser les variables de **pondération** lorsque cela est nécessaire!

[Tiré du cours d'Edward Rubin]

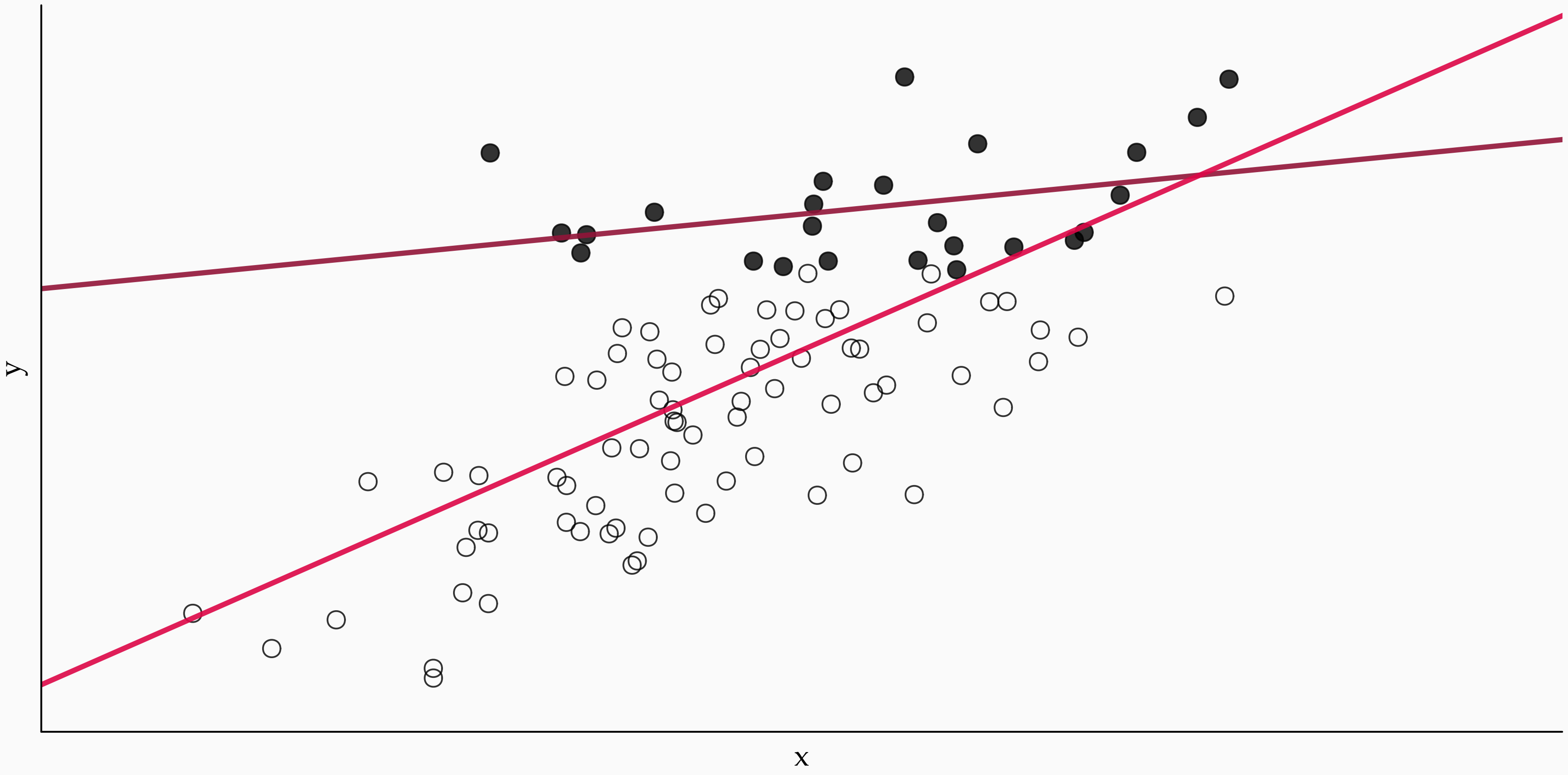
Échantillon non représentatif/aléatoire



Échantillon non représentatif/aléatoire

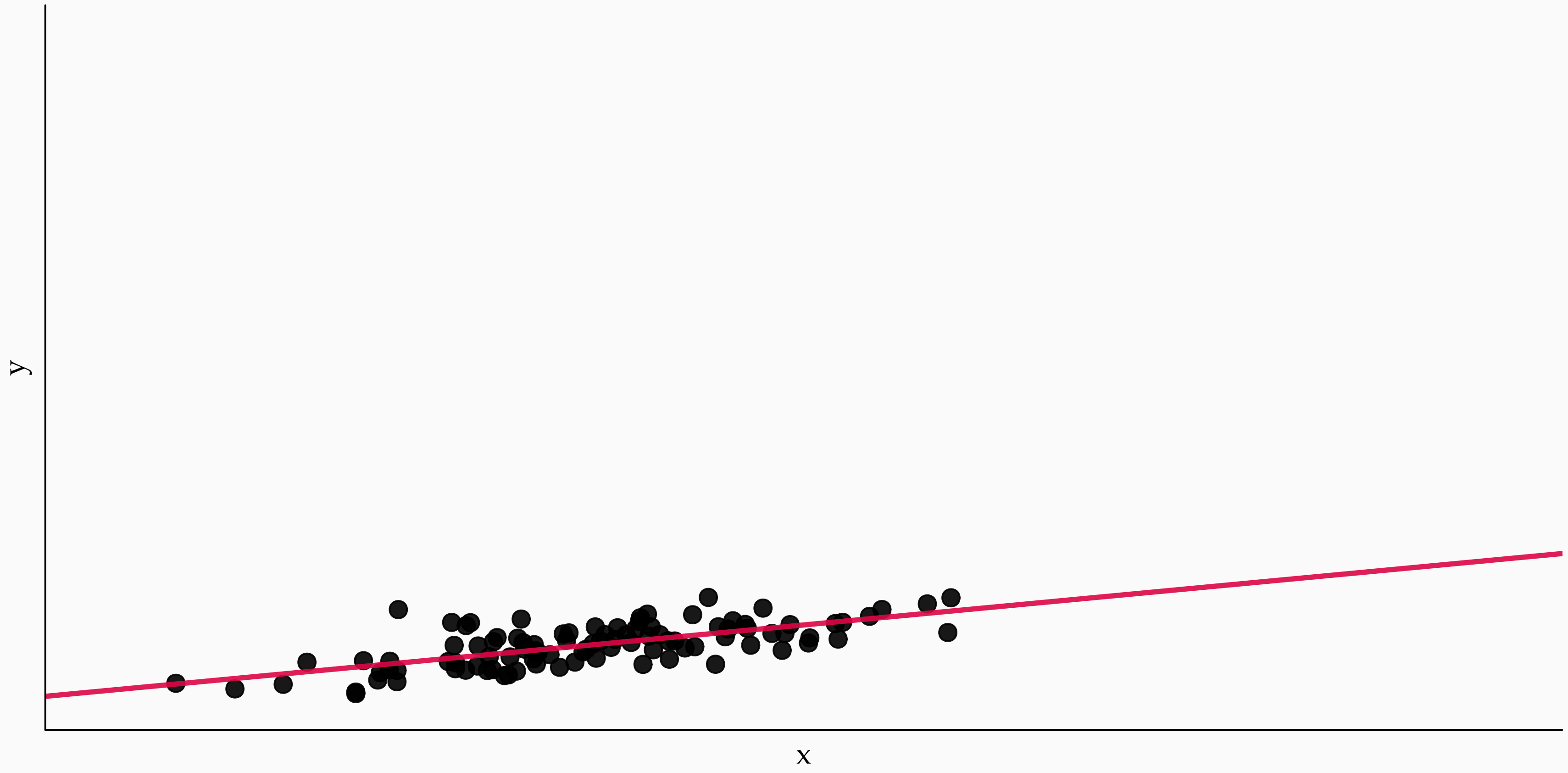


Échantillon non représentatif/aléatoire

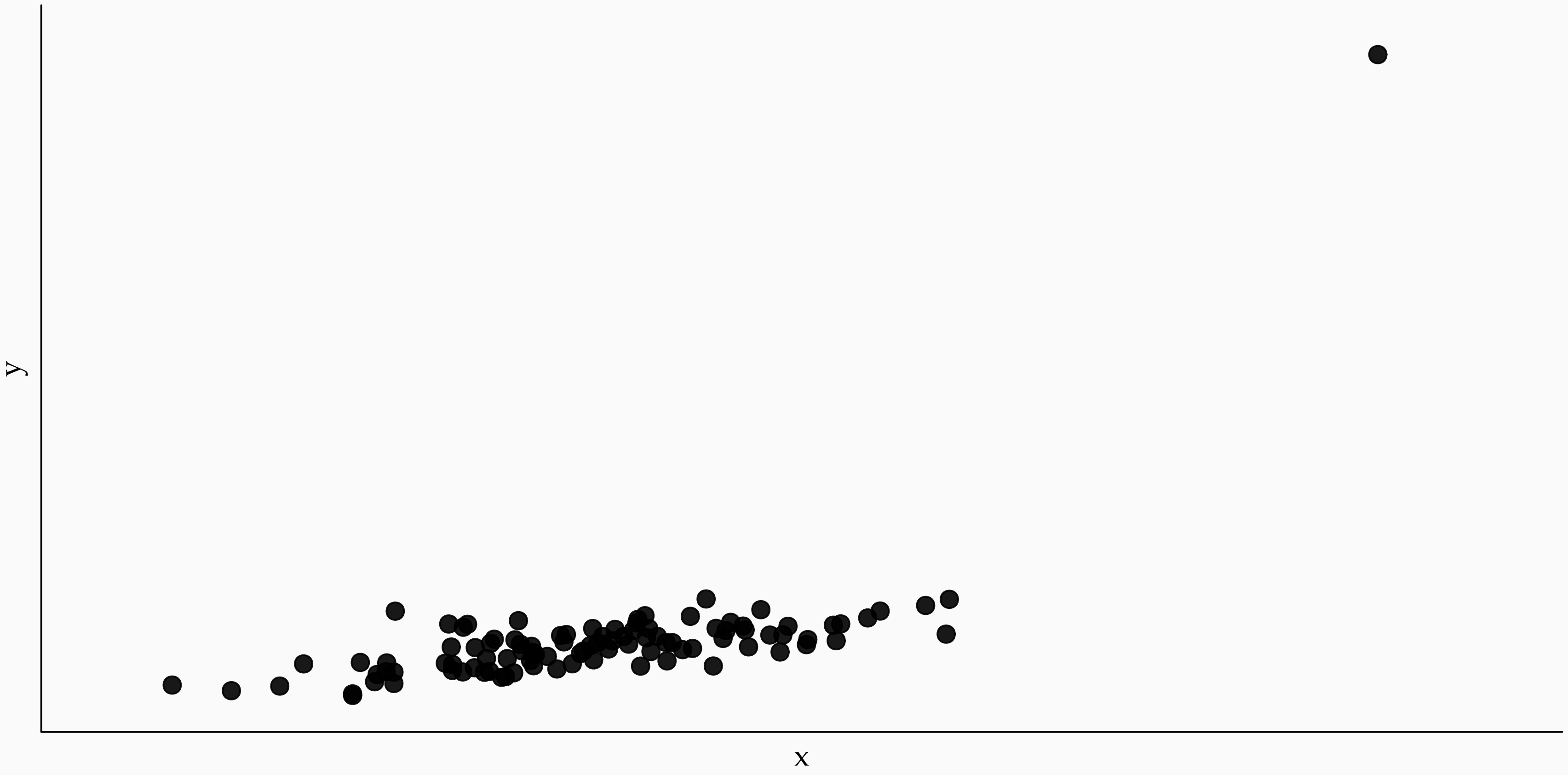


[Back]

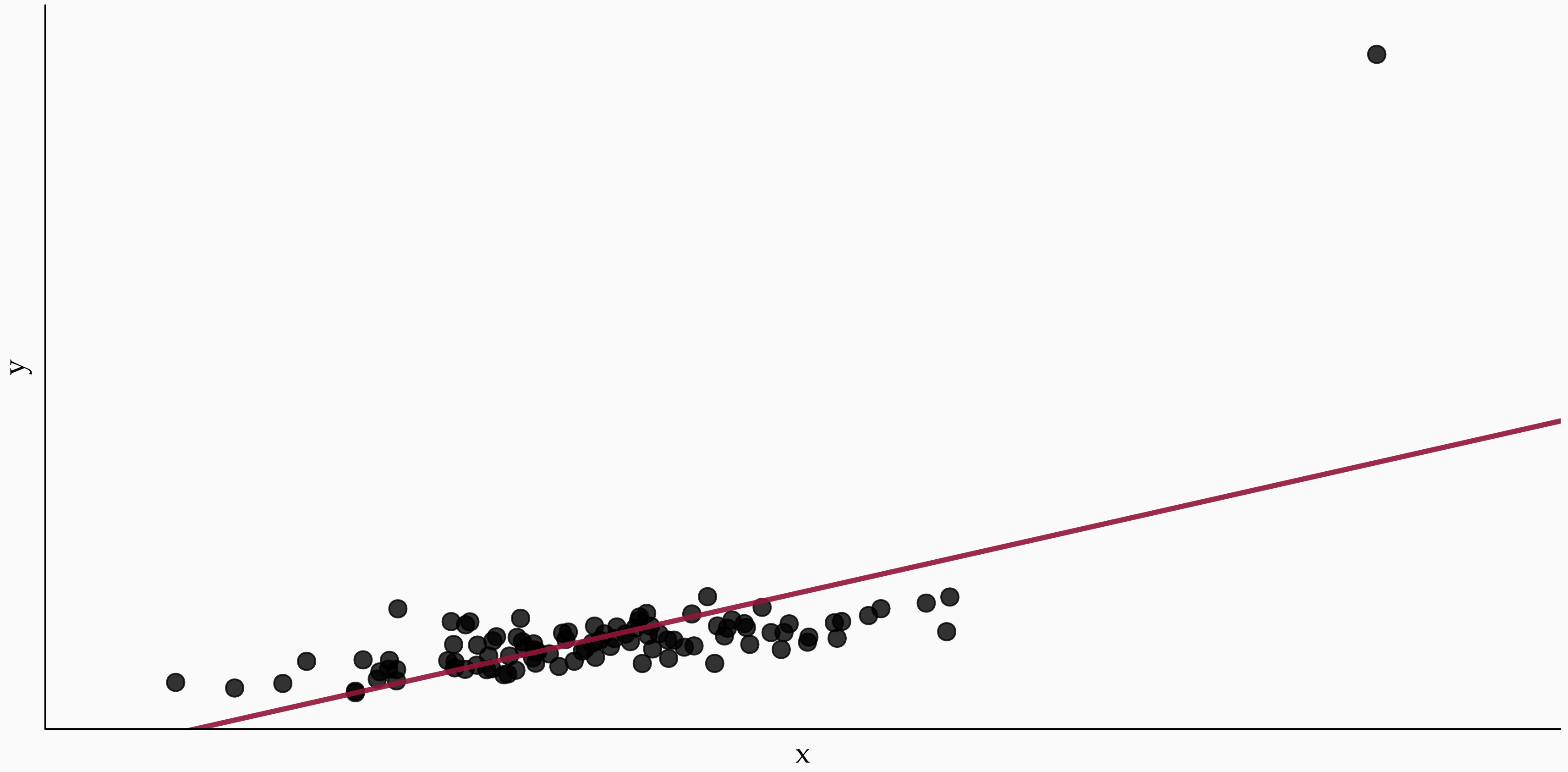
Outliers



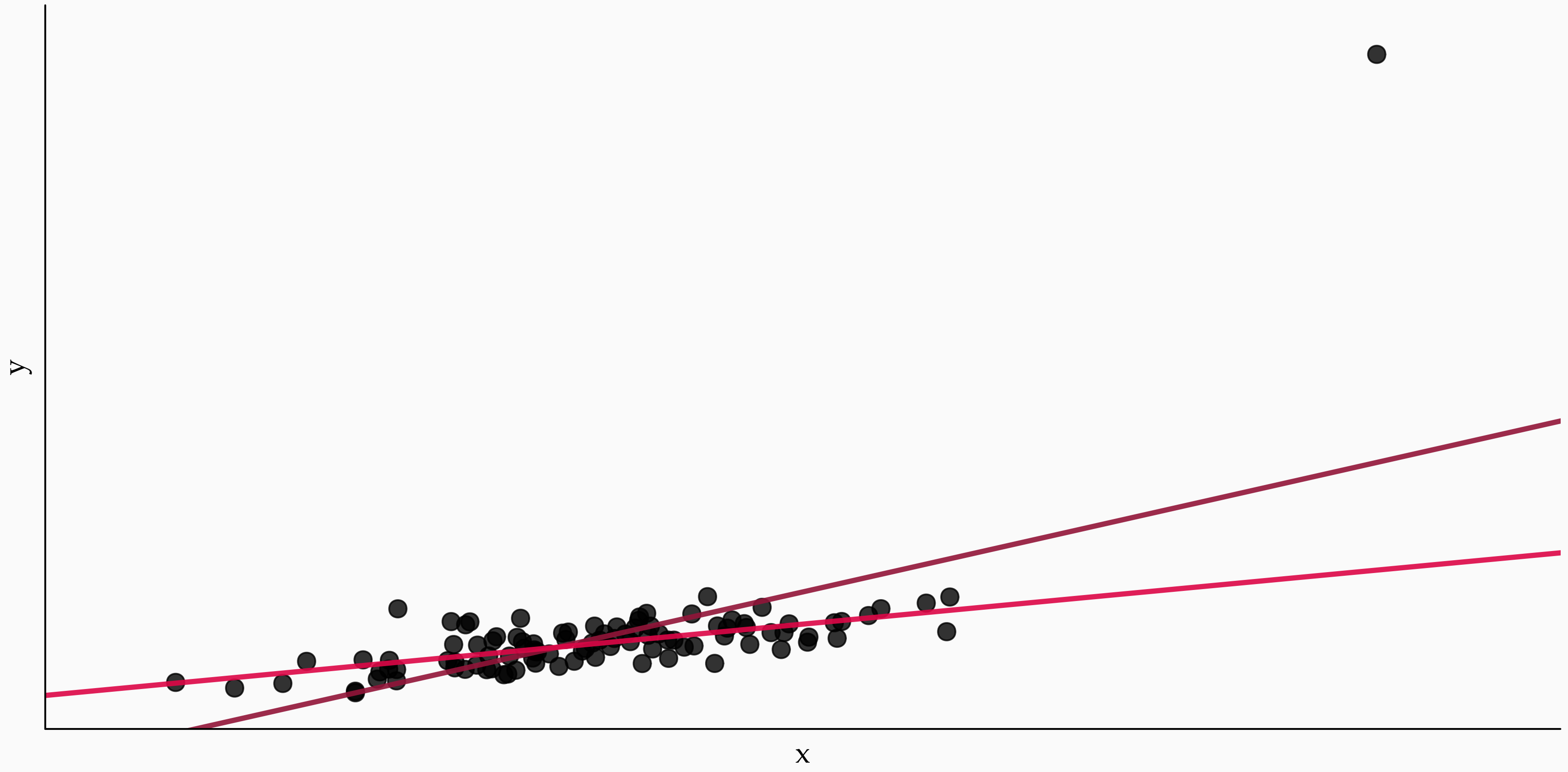
Outliers



Outliers



Outliers



Traitement des outliers

Solution 1: Supprimer les outliers

- Identifier les outliers:
  - à partir de l'**écart-type**: *lorsque la distribution des données est relativement symétrique*. Une observation éloignée de plus de 3 $\times$ écart-type de la moyenne peut être considérée comme une valeur abberrante
  - à partir de l'**écart interquartile**: peut être considérée comme outlier toute observation non incluse dans l'intervalle $[Q_1 - k(Q_3 - Q_1) \ ; \ Q3 + k(Q3 - Q1)]$ où $k > 0$. On détecte des outliers *moyens* pour $k = 1.5$, et *extrêmes* pour $k = 3$

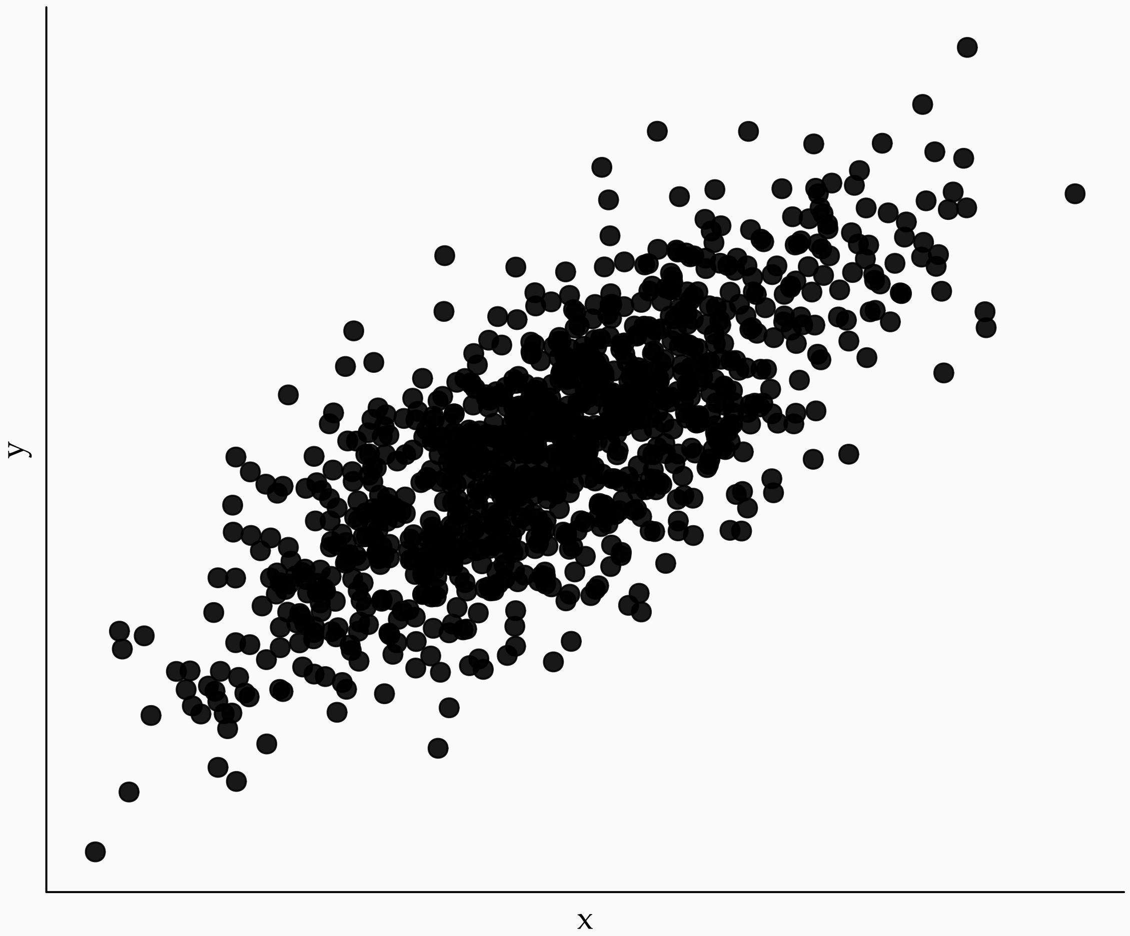
Solution 2: Windsoring: remplacer les outliers par la valeur du 99ème percentile de la variable

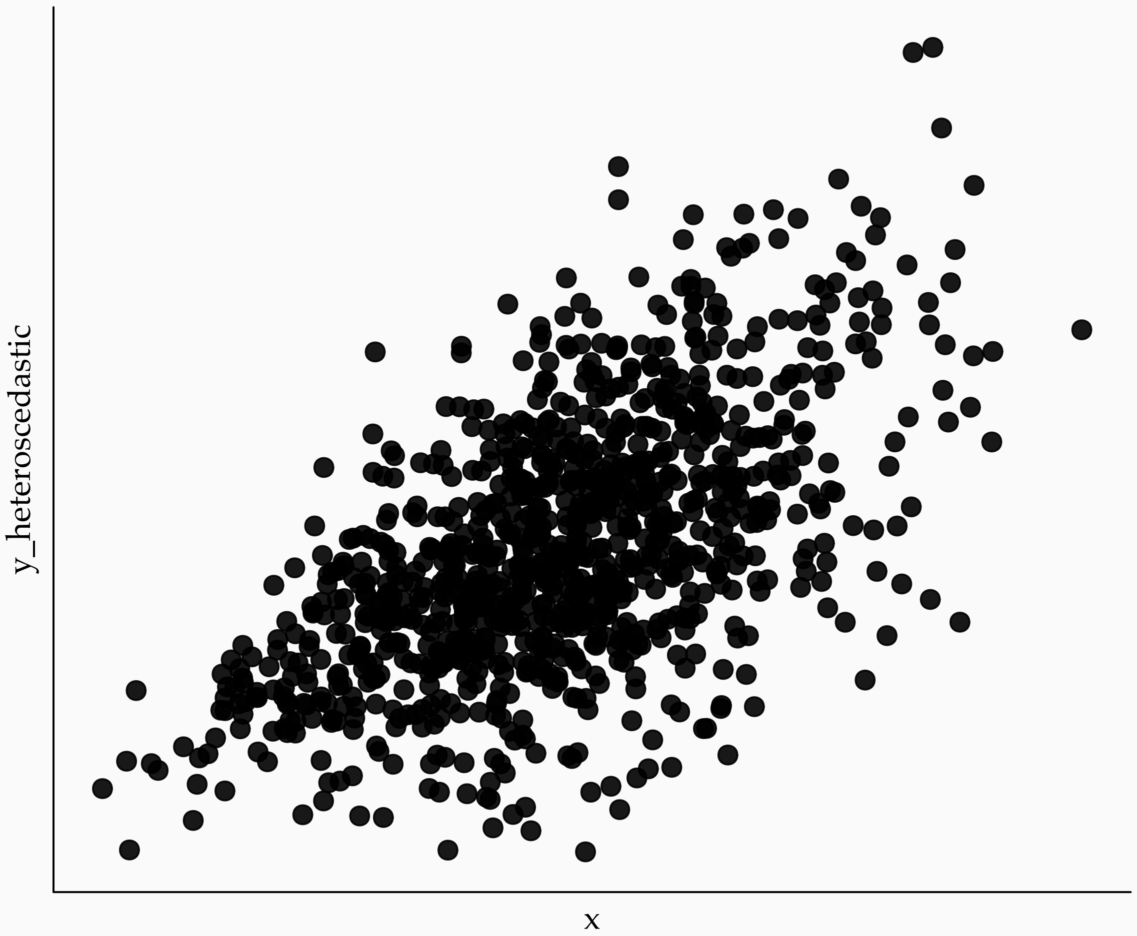
Solution 3: utiliser le log de la variable

Solution 4: ne rien faire. Parfois certaines observations/individus sont très éloignés de la moyenne.

Back

Hétéroscédasticité





[Back](#)

Calcul de l'estimateur des MCO dans le cas univarié

On a : $\mathrm{SCE} = \sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i^2 - 2y_i\hat{\beta}_0 - 2y_i\hat{\beta}_1 x_i + \hat{\beta}_0{}^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1{}^2 x_i^2 \right)$

Les conditions de premier ordre de la minimisation sont:

$$\frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_0} = 0 \ \textbf{(1)} \text{ et } \frac{\partial \mathrm{SSE}}{\partial \hat{\beta}} = 0 \ \textbf{(2)}$$

Pour (1):

$$\begin{aligned} \frac{\partial \mathrm{SSE}}{\partial \hat{\beta}_0} = 0 \qquad & \Longrightarrow \sum_{i=1}^{N} \left(2\hat{\beta}_0 + 2\hat{\beta}_1 x_i - 2y_i \right) = 2N\hat{\beta}_0 + 2\hat{\beta}_1 \sum_{i=1}^{N} x_i - 2\sum_{i=1}^{N} y_i = 2N\hat{\beta}_0 + 2\hat{\beta} N\overline{x} - 2N\overline{y} = \ 0 \\ & \Longrightarrow \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad (3) \end{aligned}$$

où $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{n}$ et $\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ sont les moyennes de x et y sur notre échantillon de taille n.

Pour (2):

$$\frac{\partial \text{SSE}}{\partial \hat{\beta}} = 0 \quad \Longrightarrow \sum_{i=1}^{N}\left(2\hat{\beta}_0 x_i + 2\hat{\beta}_1 x_i^2 - 2y_i x_i\right) = 2\hat{\beta}_0 N\overline{x} + 2\hat{\beta}_1 \sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} y_i x_i = 0 \qquad (4)$$

En remplaçant $\hat{\beta}_0$ par sa valeur définie dans (3), on obtient:

$$2N\left(\overline{y} - \hat{\beta}_1\overline{x}\right)\overline{x} + 2\hat{\beta}\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} y_i x_i = 0$$

en développant,

$$2N\overline{y}\,\overline{x} - 2N\hat{\beta}\overline{x}^2 + 2\hat{\beta}\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} y_i x_i = 0 \quad \Longrightarrow \quad 2\hat{\beta}\left(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\right) = 2\sum_{i=1}^{N} y_i x_i - 2N\overline{y}\,\overline{x}$$

$$\Longrightarrow \hat{\beta} = \frac{\sum_{i=1}^{N} y_i x_i - N\overline{y}\,\overline{x}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2} = \frac{Cov(x,y)}{Var(x)}$$

Back

ATE

On a $\delta_i = Y_{1i} - Y_{0i}$ que l'on peut réécrire

$$Y_{1i} = \delta_i + Y_{0i} \quad (1)$$

Prenons la différence entre l'outcome moyen des individus traîtés et l'outcome moyen des individus non traîtés:

$$\begin{aligned}\Delta &= \mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) \\ &= \mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)\end{aligned}$$

En remplaçant Y_{1i} par sa valeur décrite en (1),

$$\begin{aligned}\Delta &= \mathbb{E}(\delta_i + Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0) \\ &= \underbrace{\mathbb{E}(\delta_i|D_i = 1)}_{=\text{ ATT}} + \underbrace{\mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)}_{=\text{ Selection Bias}}\end{aligned}$$

Donc $\Delta = \text{ATT} + \text{Selection Bias}$.

ATE

Si D_i n'est pas corrêlé à l'outcome, formellement si $(Y_{1i}, Y_{0i}) \perp D_i$,

Alors

$$\mathbb{E}(\delta_i | D_i = 1) = \mathbb{E}(\delta_i)$$

Et

$$\mathbb{E}(Y_{0i} | D_i = 1) = \mathbb{E}(Y_{0i} | D_i = 0)$$

Donc

$$\Delta = ATE$$

Back

Variables STAR

- `gender` : genre de l'élève, `male` ou `female`
- `ethnicity` : ethnicité de l'élève, `cauc` (caucasien), `afam` (afro-americain), `asian`, hispanic, `amindian` (amerindien), `other`
- `birth` : sous la forme Année de naissance Trimestre de naissance (eg 1998 Q2)
- `stark` à `star3` : groupe de traitement (`small` ou `regular-with-aide`) ou contrôle (`regular`) pour chaque classe du kindergarten (GS) à la grade 3 (CE2). Si `NA`, alors l'élève ne fait pas encore parti/a quitté l'expérience
- `readk` à `read3` : score en lecture, pour chaque classe (k,1,2,3)
- `mathk` à `math3` : score en maths pour, chaque classe (k,1,2,3)
- `lunchk` à `lunch3` : dummy qui indique si l'élève est élègible aux repas gratuits (= proxy pour l'origine sociale), pour chaque classe (k,1,2,3)
- `schoolk` à `school3` : type d'école (`inner-city`, `suburban`, `rural` or `urban`), pour chaque classe (k,1,2,3)
- `degreek` à `degree3` : plus haut niveau de diplôme du professeur (`bachelor`, `master`, `specialist`, `phd`), pour chaque classe (k,1,2,3)
- `ladderk` à `ladder3` : degré d'expérience/statut du professeur (`level1`, `level2`, `level3`, `apprentice`, `probation`, `pending`), pour chaque classe (k,1,2,3)
- `experiencek` à `experience3` : nombre d'années d'expérience du professeur, pour chaque classe (k,1,2,3)
- `tethnicityk` à `tethnicity3` : ethnicité du professeur, `cauc` (caucasien), `afam` (afro-americain), `asian`
- `systemk` à `system3` : identifiant du système scolaire
- `schoolidk` à `schoolid3` : identifiant de l'école

Back