Séance 8.1: Regression linéaire simple et multiple

Visseho Adjiwanou, PhD.

SICSS - Montréal

31 October 2022

Plan de présentation

- Introduction
- Exemple: L'histoire de deux universités
- Présentation du modèle de régression
- Présentation graphique

Introduction

- Lorsque la voie de l'assignation aléatoire est bloquée, nous cherchons d'autres voies vers la connaissance causale.
- Manipulés avec habileté, les outils métriques autres que l'assignation aléatoire peuvent avoir une grande partie du pouvoir de **révélation de la causalité** d'une véritable expérience.
- Le plus élémentaire de ces outils est la **régression**, qui permet de comparer les sujets traités et les sujets témoins qui présentent les **mêmes caractéristiques observées**.
- Les concepts de régression sont fondamentaux et ouvrent la voie aux outils plus élaborés.

Introduction

- L'inférence causale basée sur la régression repose sur l'hypothèse selon laquelle, **lorsque les variables clés observées ont été rendues égales entre les groupes de traitement et de contrôle, le biais de sélection provenant de ce que nous ne pouvons pas voir est également éliminé**.

Exemple : L'histoire de deux universités

- Est-ce que le fait d'aller dans une université privée est payant?
  - Frais de scolarité et d'inscription au collège privé en 2012-2013 = environ 29 000
  - Frais de scolarité et d'inscription au collège public en 2012-2013 = environ 9 000.

Exemple : L'histoire de deux universités

- Un enseignement privé d'élite peut être meilleur à bien des égards :
  - Petites classes
  - Meilleures installations sportives
  - Corps enseignant plus distingué
  - Étudiants plus intelligents.
- Mais 20 000 dollars par année d'étude, c'est une grande différence: Est-ce la peine?

Exemple : L'histoire de deux universités

- Qu'est-ce qu'une école privée peut vous apporter?
  - Meilleur salaire plus tard
  - L'argent n'est pas tout, mais, comme l'a observé Groucho Marx : "L'argent vous libère de faire des choses que vous n'aimez pas. Comme je n'aime pas faire presque tout, l'argent est pratique."
  - Rencontrer son futur conjoint
  - Nouer des amitiés durables
- Lorsque les familles investissent 100 000 dollars supplémentaires ou plus dans le capital humain, il est probable que les gains anticipés soient plus élevés.

Exemple : L'histoire de deux universités

- Comment savoir si ça vaut la peine?
- Combien un diplômé de 40 ans, né dans le Massachusetts et ayant étudié à Harvard, aurait gagné s'il était allé à l'Université du Massachusetts (U-Mass).

Exemple : L'histoire de deux universités

- Les comparaisons des revenus entre ceux qui ont fréquenté différents types d'écoles révèlent invariablement des écarts importants en faveur des anciens élèves des collèges d'élite.

Exemple : L'histoire de deux universités

- Pourquoi ce résultat est biaisé:
  - Les diplômés de Harvard ont généralement de meilleures notes au lycée et de meilleurs résultats au SAT,
  - Ils sont plus motivés
  - Ils ont peut-être d'autres compétences et talents.
  - Ceux/celels qui parviennent à aller à Harvard constituent un groupe spécial et sélect.

==> Nous devons donc nous attendre à ce que les comparaisons de revenus entre les alma mater soient contaminées par un biais de sélection.

Exemple : L'histoire de deux universités

- Nous avons également vu que ce type de biais de sélection est éliminé par l'affectation aléatoire.
- Malheureusement, le bureau d'admission de Harvard n'est pas encore prêt à confier ses décisions d'admission à un générateur de nombres aléatoires.
- La question de savoir si la sélectivité des collèges est importante doit être résolue à l'aide des données générées par les décisions de routine en matière de candidature, d'admission et d'inscription prises par les étudiant.es et les universités de divers types.
- Données observationnelles: non issues de l'expérimentation

Exemple : L'histoire de deux universités

- Pouvons-nous utiliser ces données pour imiter l'essai randomisé que nous aimerions mener dans ce contexte ? >- Pas à la perfection, certes, mais nous pourrons peut-être nous en approcher. >- La clé de cette compréhension est le fait que de nombreuses décisions et choix, y compris ceux liés à la fréquentation universitaire, impliquent une certaine quantité de variations fortuites générées par des considérations financières, des circonstances personnelles et le calendrier.

Exemple : L'histoire de deux universités

- La sérendipité peut être exploitée dans un échantillon de candidats, qui pourraient facilement aller dans un sens ou dans l'autre.
- Est-ce que quelqu'un admis à Harvard va vraiment à son école publique locale à la place?
- Comment joue la sérendipité ?
  - Nancy a grandi au Texas, donc l'Université du Texas (UT) était son école publique. - Le campus phare de l'UT à Austin est classé "très compétitif" dans le classement de Barron, mais ce n'est pas Harvard.
  - UT est cependant beaucoup moins cher que Harvard.
  - Admise à la fois à Harvard et à l'UT, Nancy a choisi l'UT plutôt que Harvard

Exemple : L'histoire de deux universités

- Quelles sont les conséquences de la décision de Nancy d'accepter l'offre d'UT et de refuser celle d'Harvard ?
  - Les choses se sont plutôt bien passées pour Nancy malgré son choix de UT plutôt que Harvard
  - Elle est professeur d'économie dans une autre école de l'Ivy League en Nouvelle-Angleterre.
- Ce seul exemple ne permet pas de concluire. Mais, vous voyez l'idée.
- Nancy est son propre "contrefactuel", mais pas tout à fait.

Exemple : L'histoire de deux universités

- Mandy a obtenu son baccalauréat de l'Université de Virginie, son école publique d'origine, déclinant les offres de Duke, Harvard, Princeton et Stanford.
- Aujourd'hui, Mandy enseigne à Harvard.
- Un échantillon de deux est encore petit pour une inférence causale fiable.

Exemple : L'histoire de deux universités

- Nous aimerions comparer de nombreuses personnes comme Mandy et Nancy à de nombreuses autres personnes similaires qui ont choisi des collèges et des universités privés.
- A partir de comparaisons de groupes plus larges, nous pouvons espérer tirer des leçons générales.
- Cependant, l'accès à un large échantillon ne suffit pas.
  - La première et la plus importante étape dans notre effort pour isoler la composante fortuite du choix de l'école est de **maintenir constantes** les différences les plus évidentes et les plus importantes entre les élèves qui fréquentent les écoles privées et publiques.
  - De cette manière, nous espérons (mais ne pouvons pas promettre) rendre les autres choses égales (ceteris paribus).

Exemple : L'histoire de deux universités

- Voici un petit exemple numérique pour illustrer l'idée ceteris paribus.
- Supposons que les seules choses qui comptent dans la vie, du moins en ce qui concerne vos revenus, sont vos scores SAT et où vous allez à l'école.
- Considérez Uma et Harvey, qui ont tous deux des scores combinés en lecture et en mathématiques de 1 400 au SAT (sur 1600 au total).
- Uma est allée à U-Mass
- Harvey est allé à Harvard
- Nous commençons par comparer les revenus d'Uma et de Harvey. Parce que nous avons supposé que tout ce qui compte pour les revenus en plus du choix du collège est le score SAT combiné, Uma contre Harvey est une comparaison ceteris paribus.

Exemple : L'histoire de deux universités

- En pratique, bien sûr, la vie est plus compliquée.
- Ce simple exemple suggère une complication importante :
  - Uma est une jeune femme et
  - Harvey est un jeune homme.

Exemple : L'histoire de deux universités

- Les femmes ayant des qualifications éducatives similaires gagnent souvent moins que les hommes
  - En raison de la discrimination ou
  - Du temps passé hors du marché du travail pour avoir des enfants.

==> Le fait que Harvey gagne 20% de plus qu'Uma peut être l'effet d'une éducation supérieure à Harvard, mais cela pourrait tout aussi bien refléter un écart salarial entre hommes et femmes généré par d'autres choses.

Exemple : L'histoire de deux universités

- Nous aimerions démêler **l'effet Harvard pur** de ces autres choses.
- C'est facile si la seule autre chose qui compte est le sexe :
  - remplacez Harvey par une étudiante de Harvard, Hannah, qui a également un SAT combiné de 1400, et comparer Uma et Hannah.

Exemple : L'histoire de deux universités

- Parce que nous recherchons des conclusions générales qui vont au-delà des histoires individuelles, nous recherchons de nombreux contrastes similaires de même sexe et de même SAT dans les deux écoles. >- Autrement dit, nous calculons la différence de revenus moyens entre les étudiants de Harvard et U-Mass avec le même sexe et les mêmes scores SAT. >- La moyenne de toutes ces différences spécifiques à un groupe entre Harvard et U-Mass est notre première tentative pour estimer l'effet causal d'une éducation à Harvard. >- Il s'agit d'un estimateur d'appariement économétrique qui contrôle - c'est-à-dire maintient fixe - les scores de sexe et SAT. >- En supposant que, sous réserve du sexe et des scores SAT, les étudiants qui fréquentent Harvard et U-Mass ont un potentiel de revenus similaire, cet estimateur capture l'effet causal moyen d'un diplôme de Harvard sur les revenus.

Matchmaker

- Les revenus ne se limitent pas au sexe, aux écoles et aux scores SAT.
- Étant donné que les décisions de fréquentation des collèges ne sont pas attribuées au hasard, nous devons contrôler tous les facteurs qui déterminent à la fois les décisions de fréquentation et les revenus ultérieurs.
- Ces facteurs incluent les caractéristiques des étudiants:

```
>- la capacité d'écriture,
>- la diligence,
>- les liens familiaux, ...
```

Matchmaker

- Les revenus ne se limitent pas au sexe, aux écoles et aux scores SAT.
- Étant donné que les décisions de fréquentation des collèges ne sont pas attribuées au hasard, nous devons contrôler tous les facteurs qui déterminent à la fois les décisions de fréquentation et les revenus ultérieurs.
- Ces facteurs incluent les caractéristiques des étudiants:

```
>- la capacité d'écriture,
>- la diligence,
>- les liens familiaux, ...
```

- Le contrôle d'un tel éventail de facteurs semble intimidant : les possibilités sont pratiquement infinies et de nombreuses caractéristiques sont difficiles à quantifier.

Matchmaker

- Stacy Berg Dale et Alan Krueger (2002) ont trouvé un raccourci intelligent et convaincant.
- Au lieu d'identifier tout ce qui pourrait avoir de l'importance pour le choix du collège et les revenus, ils travaillent avec une mesure récapitulative clé : les caractéristiques des collèges auxquels les étudiants ont postulé et ont été admis.

Matchmaker

- Considérez à nouveau l'histoire d'Uma et Harvey:
  - Tous deux ont postulé et ont été admis à U-Mass et Harvard.
  - Le fait qu'Uma ait postulé à Harvard suggère qu'elle a la motivation pour y aller, tandis que son admission à Harvard suggère qu'elle a la capacité d'y réussir, tout comme Harvey. C'est du moins ce que pense le bureau d'admission de Harvard, et ils ne sont pas facilement dupes.

Matchmaker

- Uma opte néanmoins pour une éducation à U-Mass moins chère. >- son choix pourrait être attribuable à des facteurs qui ne sont pas étroitement liés au potentiel de revenus d'Uma, >- Un oncle qui a réussi qui est allé à U-Mass >- Un meilleur ami qui a choisi U-Mass >- Uma a raté la date limite pour la Bourse du club Rotary qui aurait financé l'éducation à Harvard. >- Si de tels événements fortuits ont été décisifs pour Uma et Harvey, alors les deux font un bon match.

Matchmaker

- Dale et Krueger ont analysé un grand ensemble de données appelé College and Beyond (C&B).
- Cet ensemble de données contient des informations sur des milliers d'étudiants inscrits dans un groupe de collèges et d'universités américains modérément à hautement sélectifs >- Des informations d'enquête recueillies auprès des étudiants au moment où ils ont passé le SAT, environ un an avant leur entrée à l'université, >- et des informations recueillies en 1996, longtemps après que la plupart d'entre eux aient obtenu leur diplôme universitaire.

Matchmaker

- L'analyse se concentre ici sur les étudiants qui se sont inscrits en 1976 et qui travaillaient en 1995 .
- Collèges et Universités
  - Université privées prestigieuses: Université de Pennsylvanie, Princeton et Yale ;
  - Nombre de collèges privés plus petits: Swarthmore, Williams et Oberlin ;
  - Quatre universités publiques (Michigan, Université de Caroline du Nord, Penn State et Université de Miami en Ohio).
- Les résultats moyens (1978) au SAT dans ces écoles allaient d'un minimum de 1 020 à Tulane à un maximum de 1 370 à Bryn Mawr.
- En 1976, les frais de scolarité étaient aussi bas que 540 dollars à l'Université de Caroline du Nord et aussi élevés que 3 850

Matchmaker

- Le tableau 1 détaille une version simplifiée de la stratégie d'appariement de Dale et Krueger, dans une configuration que nous appelons la "matrice d'appariement des collèges". - Ce tableau présente les demandes d'admission, les admissions et les décisions d'inscription pour une liste (inventée) de neuf étudiants, chacun d'entre eux ayant posé sa candidature à trois écoles choisies parmi une liste imaginaire de six.
- Trois des six écoles figurant dans le tableau sont publiques (All State, Tall State et Altered State) et trois sont privées (Ivy, Leafy et Smart).
- Cinq de nos neuf élèves (numéros 1, 2, 4, 6 et 7) ont fréquenté des écoles privées. Le salaire moyen dans ce groupe est de 92 000 dollars.
- Les quatre autres, dont le revenu moyen est de 72 500 dollars,

Tableau hypothétique

	Private				Publique			
G	Et.	Ivy	Leafy	Smart	A_state	T_state	Altered	1996
A	1		Rejet	**Admis**		Admis		1300
	2		Rejet	**Admis**		Admis		1000
	3		Rejet	Admis		**Admis**		1100
B	4	**Admis**			Admis		Ad	6000
	5	Admis			Admis		**Ad**	3000
C	6		**Admis**	Admis	Admis			1150
	7		**Admis**	Admis	Admis			7500
D	8	Rejet			**Admis**	Admis		9000
	9	Rejet			Admis	**Admis**		6000

Commentaire du tableau

- Quel est l'effet du fait de fréquenter une école privée?

```
moy_private <- (130000 + 100000 + 60000 + 115000 + 75000)/5
moy_private
```

```
## [1] 96000
```

```
moy_public <- (110000 + 30000 + 90000 + 60000)/4
moy_public
```

```
## [1] 72500
```

```
Effet1_ <- moy_private - moy_public
Effet1_
```

```
## [1] 23500
```

Commentaire tableau

- Groupe A: (110000 + 100000)/2 - 110000 = -5000 >- Groupe B: 60000 - 30000 = 30000 - Cet écart suggère un avantage substantiel des écoles privées. >- Groupe C: Non informatif >- Groupe D: Non informatif

Commentaire tableau

- Les groupes A et B sont ceux où l'action se situe dans notre exemple, puisque ces groupes comprennent des élèves des écoles publiques et privées qui ont postulé et ont été admis dans le même ensemble d'écoles.
- Pour générer une estimation unique qui utilise toutes les données disponibles, nous faisons la moyenne des estimations spécifiques au groupe.
- La moyenne de -5 000 dollars pour le groupe A et de 30 000 dollars pour le groupe B est de 12 500 $.
- Il s'agit d'une bonne estimation de l'effet de la fréquentation d'une école privée sur les gains moyens, car, dans une large mesure, elle contrôle les choix et les capacités des candidats.

Commentaire tableau

- La moyenne simple des différences de traitement-contrôle dans les groupes A et B n'est pas la seule comparaison bien contrôlée qui peut être calculée à partir de ces deux groupes.
- Par exemple, nous pourrions construire une moyenne pondérée qui reflète le fait que le groupe B comprend deux étudiants et que le groupe A en comprend trois.
- La moyenne pondérée dans ce cas est calculée comme suit (3/5 * -5 000) + (2/5 * 30 000) = 9 000.
- En mettant l'accent sur des groupes plus grands, ce schéma de pondération utilise les données plus efficacement et peut donc générer un résumé statistiquement plus précis de l'écart entre les revenus privés et publics.

Commentaire tableau

- Le point le plus important dans ce contexte est la nature des comparaisons appariées sous-jacentesdes, pommes avec des pommes et des oranges avec des oranges.
- Les pommes du groupe A sont comparées aux autres pommes du groupe A, tandis que les oranges du groupe B ne sont comparées qu'aux autres oranges du groupe B.
- En revanche, les comparaisons naïves qui comparent simplement les revenus des élèves des écoles privées et publiques génèrent un écart beaucoup plus important de 19 500 dollars lorsqu'elles sont calculées en utilisant les neuf élèves du tableau.
- Même calculées pour les cinq élèves des groupes A et B, les comparaisons non contrôlées génèrent un écart de 20 000 dollars (20 = 110 + 100 + 60)/3 - (110 + 30)/2).

Commentaire tableau

- Ces comparaisons non contrôlées beaucoup plus importantes reflètent un biais de sélection :
  - les étudiants qui postulent et sont admis dans des écoles privées ont des revenus plus élevés partout où ils ont finalement choisi d'aller.
- La preuve d'un biais de sélection émerge d'une comparaison des revenus moyens entre (plutôt qu'au sein des) groupes A et B.
- Les revenus moyens dans le groupe A, où les deux tiers postulent aux écoles privées, sont d'environ 107 000 dollars.
- Les revenus moyens dans le groupe B, où les deux tiers postulent pour les écoles publiques, ne sont que de 45 000 dollars.
- Nos comparaisons au sein des groupes révèlent qu'une grande partie de ce manque à gagner n'est pas liée aux décisions de fréquentation universitaire des étudiants.
- La différence entre les groupes s'explique par une combinaison

Conclusion

1. Nous avons estimé 5 mesures de l'effet de l'école privée. Mais, pas toutes ces mesures sont pas une bonne estimation de l'effet. C'est quoi un bon estimateur?
2. Quel outil disposons-nous pour estimer cet effet correctement?

Régression

- La régression est l'outil que les maîtres s'emparent en premier, ne serait-ce que pour servir de référence à des stratégies empiriques plus élaborées.
- Bien que la régression soit une chose aux multiples splendeurs, nous la considérons comme un entremetteur automatisé. Plus précisément, les estimations de régression sont des moyennes pondérées de comparaisons appariées multiples du type construit pour les groupes dans notre matrice d'appariement stylisée.

Régression

- Les ingrédients clés de la recette de régression sont :
  - la variable dépendante, dans ce cas, les revenus de l'étudiant i plus tard dans la vie, également appelée variable de résultat (notée Yi) ;
  - la variable de traitement, en l'occurrence une variable fictive qui indique les étudiants ayant fréquenté un collège ou une université privée (notée Pi) ; et
  - un ensemble de variables de contrôle, en l'occurrence des variables qui identifient des ensembles d'écoles dans lesquelles les élèves ont postulé et ont été admis.

Régression

- Dans notre matrice d'appariement, les cinq élèves des groupes A et B apportent des données utiles, tandis que les élèves des groupes C et D peuvent être écartés. >- Une seule variable indiquant les élèves du groupe A nous indique dans lequel des deux groupes se trouvent les élèves restants. >- La variable que nous appellerons Ai, est notre seul contrôle. >- Notez que Pi et Ai sont des variables dichotomiques, c'est-à-dire qu'elles sont égales à 1 pour indiquer des observations dans un état ou une condition spécifique, et à 0 sinon. >- Le modèle de régression dans ce contexte est une équation reliant la variable de traitement à la variable dépendante tout en maintenant les variables de contrôle fixes en les incluant dans le modèle.

Régression

- Avec une seule variable de contrôle, Ai, la régression d'intérêt peut s'écrire :

$$Y_i = \alpha + \beta * P_i + \gamma * A_i + \epsilon_i$$

Régression

- L'analyse de régression attribue des valeurs aux paramètres du modèle de manière à rendre Y chapeau aussi proche que possible de Y.
- Ceci est accompli en choisissant des valeurs qui minimisent la somme des résidus au carré, ce qui conduit au surnom de moindres carrés ordinaires (OLS) pour les estimations résultantes.
- En exécutant cette minimisation dans un échantillon particulier, on dit que nous estimons les paramètres de régression.

$$\alpha = 40000$$

$$\beta = 10000$$

$$\gamma = 60000$$

Régression - forme générale

Définition

- Le modèle de **régression linéaire simple** peut être utilisé pour étudier la relation entre deux variables, la variable dépendante (Y) et la variable indépendante (X), comme l'exemple dont nous venons de parler.
- On parle de **régression linéaire multiple** dans le cas où il y a au moins 2 **variables indépendantes**.

Spécification

- Nous avons {Yi, Xi}, un échantillon de Y et X
- Nous sommes intéressés à **"expliquer Y en termes de X"** ou à **"étudier comment Y varie avec les changements de X"**
- Modèle

$$Y = \alpha + \beta X + \epsilon$$

- Y = variable dépendante | variable à expliquer
- X = variable indépendante | variable explicative | prédicteur
- (α, β) = coefficients à déterminer (on dit à estimer) | paramètres du modèle
- ϵ = erreurs| termes d'erreur de moyenne nulle (unobserved error / disturbance error)

Interprétation

$$Y = \alpha + \beta X + \epsilon$$

ou une formulation alternative:

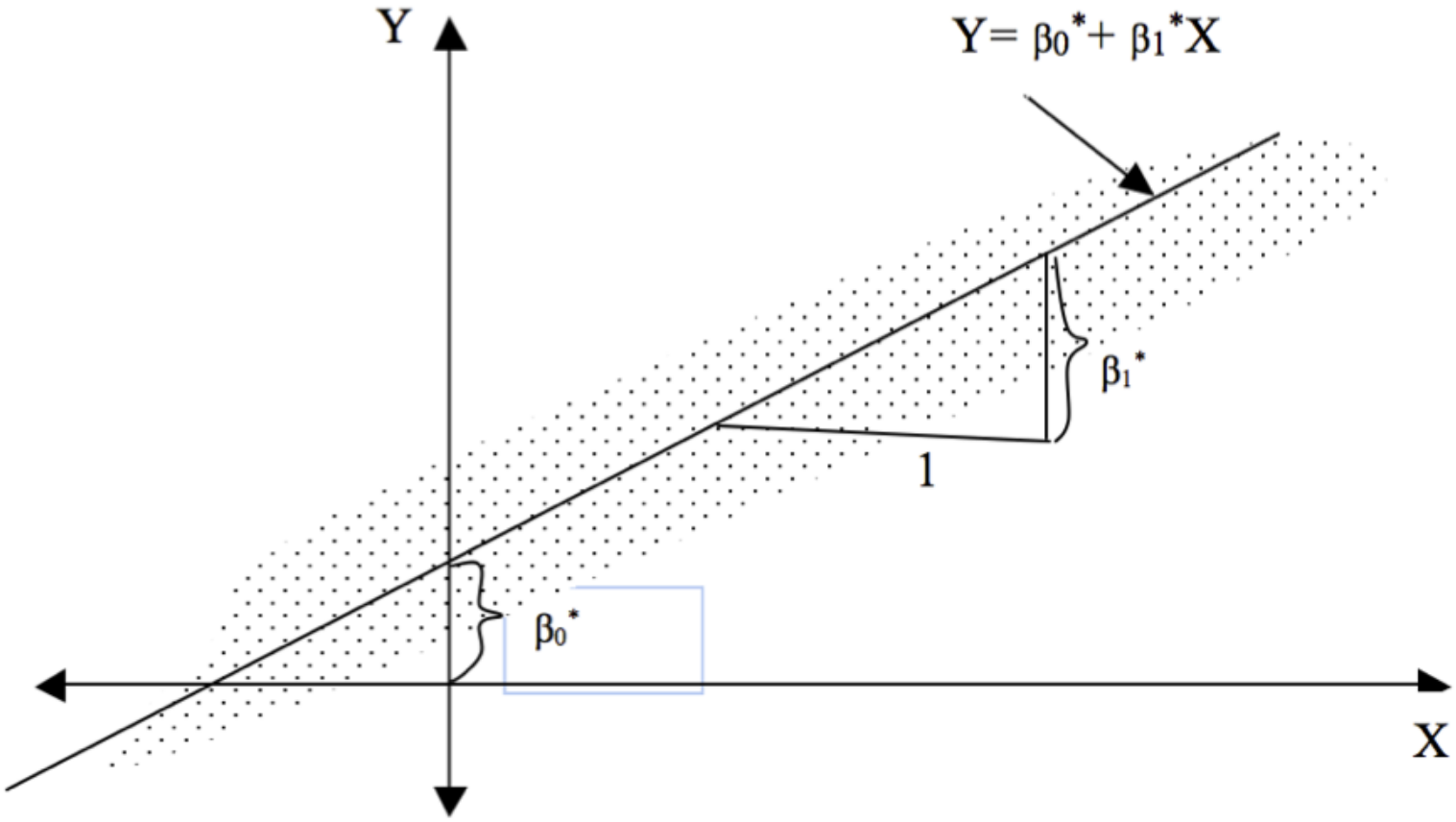
$$E(Y|X) = \alpha + \beta X$$

- $\alpha + \beta X$: moyenne de Y étant donnée la valeur de X
- α : la valeur de Y quand X est zéro
- β : augmentation de Y associée à une augmentation d'une unité de X

D'où vient le epsilon

1. Omission de l'influence d'innombrables évènements fortuits
   - Autres covariables importantes (influences systématiques) Etat nutritionnel de la mère
   - Autres petites variables non significatives avec une très légère influence irrégulière
2. Erreur de mesure
   - Dans la variable dépendante
   - Dans la variable indépendante (plus problématique)
3. Indétermination humaine
   - Le comportement humain est tel que les actions entreprises dans des circonstances identiques différeront de manière aléatoire

Spécification



- β_0 = intersection à l'origine (intercept)

Méthodes d'estimation

- Il existe de nombreux estimateurs présentant des caractéristiques différentes susceptibles de résoudre l'équation 1.
- La tâche de l'économètre est de trouver le meilleur **estimateur**.
- Les deux approches les plus importantes sont:
  1. Méthodes des moindres carrés: dans le cas d'une régression linéaire simple, il s'agit de trouver la meilleure ligne qui décrit de manière appropriée le nuage de points $\{Yi, Xi\}$.
  2. Approche du maximum de vraisemblance

Méthodes d'estimation

- Estimer les paramètres du modèle à partir des données $\{X_i, Y_i\}$
- $(\hat{\alpha}, \hat{\beta})$: Coefficients estimés
- $\hat{Y} = \hat{\alpha} + \hat{\beta}X$: Valeur prédite (predicted/fitted value)
- $\hat{epsilon} = Y - \hat{Y}$: Résidus (residuals)

Méthode d'estimation: moindres carrées ordinaires

- Minimiser la somme des carrées des résidus (SSR) :

$$SSR = \sum_{i=1}^{n} \hat{\epsilon}^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} + \beta \hat{X}_i)^2$$

- Solution
- Coefficients estimés :

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- la droite des moindres carrées passe toujours par les points $(\bar{X}, \bar{Y})$
- $\hat{Y} = \bar{Y}$
- la moyenne des résidus est toujours égale à zéro

Régression linéaire multiple

Introduction

- La régression linéaire simple ne permet pas de déduire une causalité: la réalité est plus complexe
- Permet de comprendre le concept de régression
- En cas de plus d'une variable indépendante, on parle de régression linéaire multiple

Spécification

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

Où ϵ_i suit une loi normale de moyenne 0 et de variance σ^2. On a k indépendantes variables pour n observations $(Y_i, X_{11}, X_{12}, ..., X_{1k}), ..., (Y_n, X_{n1}, X_{n2}, ..., X_{nk})$.

Exemple: - Y peut être le poids à la naissance - X1 l'age de la mère à la naissance de l'enfant - X2 le sexe de l'enfant

Spécification

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \epsilon_i$$

- Votre tâche: estimer l'effet de chaque variable X spécifique sur Y, en contrôlant l'effet des autres.
- Cette équation peut être récrite :

$$Y_1 = \alpha + \beta_1 X_{11} + \beta_2 X_{21} + ... + \beta_k X_{k1} + \epsilon_1$$
$$Y_2 = \alpha + \beta_1 X_{12} + \beta_2 X_{22} + ... + \beta_k X_{k2} + \epsilon_2 \ ...$$
$$Y_n = \alpha + \beta_1 X_{1n} + \beta_2 X_{2n} + ... + \beta_k X_{kn} + \epsilon_n$$

Cette façon d'écrire les équations est difficile à manipuler:

Notation matricielle: $Y = X\beta + \epsilon$

Spécification

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}; \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Estimation des paramètres

Hypothèses

		Expression mathématique		
	Hypothèses	**Linéaire simple**	**Linéaire multiple**	**Violations**
H1	Variable dépendante fonction linéaire d'un ensemble spécifique de variables indépendantes, plus une perturbation	$Y_i = ß_0 + ß_1X_i + \varepsilon_i$, $i = 1$ à N	$Y = Xß + \varepsilon$	Mauvaises variables indépendantes, Non linéarité, Paramètres changeant
H2	L'espérance mathématique de l'erreur est nulle	$E(\varepsilon_i) = 0$, pour tout i	$E\varepsilon = 0$	Intercept biaisé
H3	La variance de l'erreur est constante quel que soit i (homoscédasticité)	$Var(\varepsilon_i) = \sigma^2$	$E\varepsilon\varepsilon' = \sigma^2 I$	Hétéroscédasticité
	Les erreurs sont non corrélées (ou encore indépendantes)	$E(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$		Erreurs autocorrélées
H4	Les variables indépendantes sont observées sans erreur	X_i fixé	X fixé	Erreurs dans les variables X, Auto-régression, Équations simultanées
H5	Absence de colinéarité entre les variables indépendantes	$\sum(X_i - X)^2 \neq 0$	Rang de X = K<= N	Colinéarité parfaites
	Le nombre d'observations est supérieur au nombre des séries explicatives	N>= K		

Estimation des paramètres

- Les paramètres inconnus:
  - k (beta) + 1 (alpha) paramètres
  - σ^2
- Estimation par les moindres carrés ordinaires ou la méthode des maximums de vraisemblance: Plus difficile à estimer.

Estimation

- On démontre que :

$\beta^* = (X'X)^{-1}(X'Y)$

Variance-covariance de $\beta^* = \sigma^2(X'X)^{-1}$

Mais encore une fois, σ^2 n'est pas connu. Il est remplacé par:

$s^2 = e'e/(T-k)$

avec (e = Y-Y')

Estimation avec R

Estimation sous R: forme générale

lm(**formula**, **data**, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, . . .)

- lm : pour dire que vous estimez un modèle linéaire. A l'intérieur, vous devez spécifier :

1. Formula: elle comprend trois éléments :

- la variable dépendante: le premier élément dans la parenthèse. Il doit s'agir d'une variable qui est continue.
- le tilde ($\sim$) qui sépare la variable dépendante des variables indépendantes
- les éléments après le $\sim$, ce sont les variables indépendantes. Elles doivent être séparées par des $+$.

Estimation sous R: forme générale

2. Data: Vous devez ensuite spécifier les données sur lesquelles vous faites votre régression.

Ces deux éléments sont les plus importants: ils sont obligatoires.

Estimation sous R: forme générale

```
?lm()
```

Solution graphique

Exemple : Violence et accès à l'information

Nom D	escription
`beat_goesout`	Pourcentage de femmes dans chaque pays qui pensent a le droit de battre sa femme si elle sort sans le lui dir
`beat_burnfood`	Pourcentage de femmes dans chaque pays qui pensent a le droit de battre sa femme si elle brûle sa nourriture
`no_media`	Pourcentage de femmes dans chaque pays qui ont rare à un journal, une radio ou une télévision.
`sec_school`	Pourcentage de femmes dans chaque pays ayant un niv d'éducation secondaire ou supérieur.
`year`	Année de l'enquête
`region`	Région du monde
`country`	pays

Dressons la table

```
rm(list = ls())


library(tidyverse)
```

```
## -- Attaching packages ---------------------------------
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2

## Warning: package 'tibble' was built under R version 3.6.

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2
```

Quelques informations sur les données

```
head(dhs_ipv)
```

```
## # A tibble: 6 x 8
##    ...1 beat_burnfood beat_goesout sec_school no_media c
##   <dbl>         <dbl>        <dbl>      <dbl>    <dbl> <
## 1     1           4.4         18.6       25.2      1.5 A
## 2     4           4.9         19.9       67.7      8.7 A
## 3     5           2.1         10.3       67.6      2.2 A
## 4     6           0.3          3.1       46        6.4 A
## 5     7          12.1         42.5       74.6      7.4 A
## 6     8          NA           NA         24       41.9 R
```

Quelques informations sur les données

```
glimpse(dhs_ipv)
```

```
## Rows: 151
## Columns: 8
## $ ...1           <dbl> 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1
## $ beat_burnfood <dbl> 4.4, 4.9, 2.1, 0.3, 12.1, NA, NA,
## $ beat_goesout   <dbl> 18.6, 19.9, 10.3, 3.1, 42.5, NA, 1
## $ sec_school     <dbl> 25.2, 67.7, 67.6, 46.0, 74.6, 24.0
## $ no_media       <dbl> 1.5, 8.7, 2.2, 6.4, 7.4, 41.9, 45.
## $ country        <chr> "Albania", "Armenia", "Armenia", "
## $ year           <dbl> 2008, 2000, 2005, 2010, 2006, 2004
## $ region         <chr> "Middle East and Central Asia", "M
```

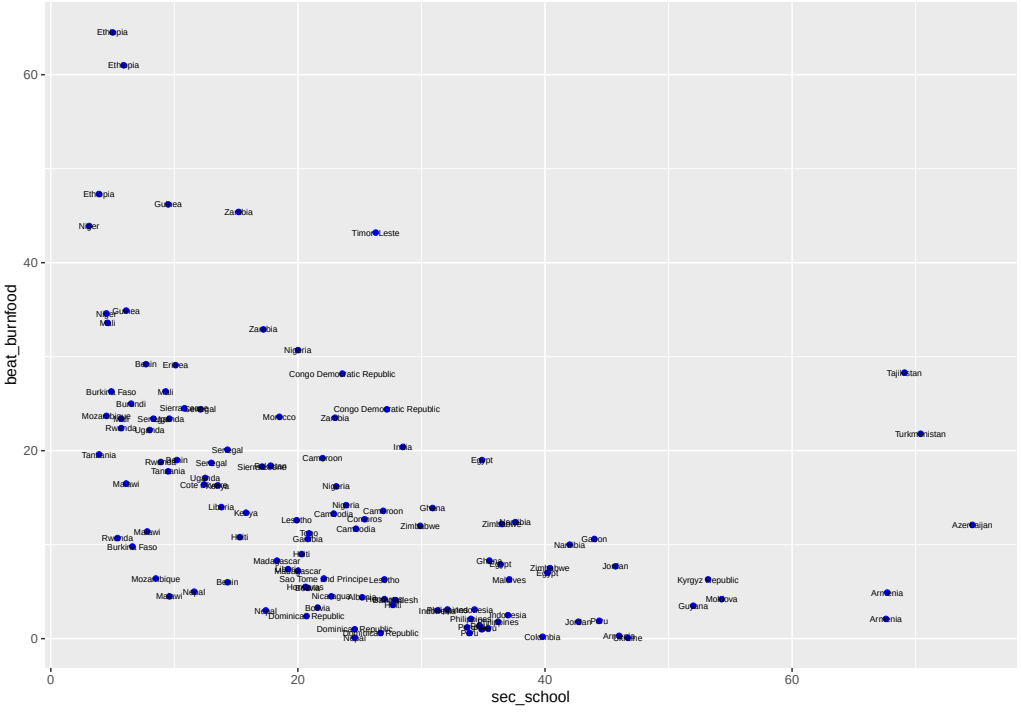
Quelques informations sur les données

```
summary(dhs_ipv)
```

```
##       ...1        beat_burnfood    beat_goesout      sec_
##  Min.   :  1.00   Min.   : 0.10   Min.   : 0.30   Min.
##  1st Qu.: 40.50   1st Qu.: 4.50   1st Qu.:11.85   1st Qu
##  Median : 79.00   Median :11.85   Median :28.10   Median
##  Mean   : 80.53   Mean   :15.04   Mean   :28.60   Mean
##  3rd Qu.:119.50   3rd Qu.:22.25   3rd Qu.:42.08   3rd Qu
##  Max.   :160.00   Max.   :64.50   Max.   :82.70   Max.
##                   NA's   :31      NA's   :27      NA's
##     no_media         country              year         re
##  Min.   : 0.80   Length:151         Min.   :1999   Lengt
##  1st Qu.:11.25   Class :character   1st Qu.:2004   Class
##  Median :29.15   Mode  :character   Median :2007   Mode
##  Mean   :28.40                      Mean   :2007
##  3rd Qu.:43.23                      3rd Qu.:2011
```

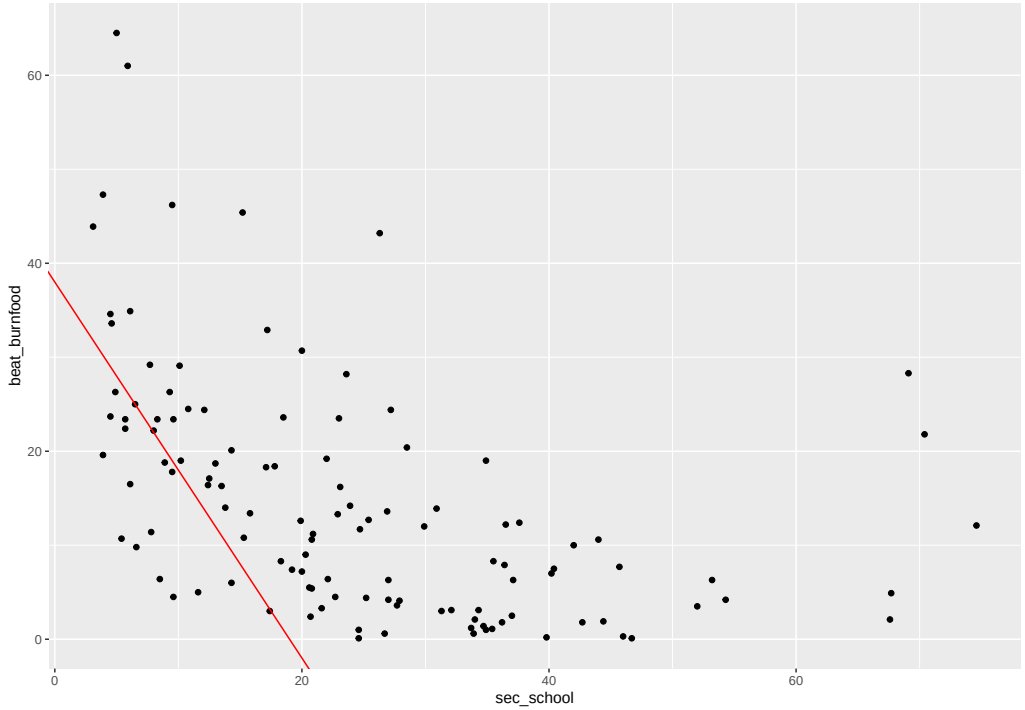
Association entre beat_burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood), color
  geom_text(aes(x = sec_school, y = beat_burnfood, label =
```



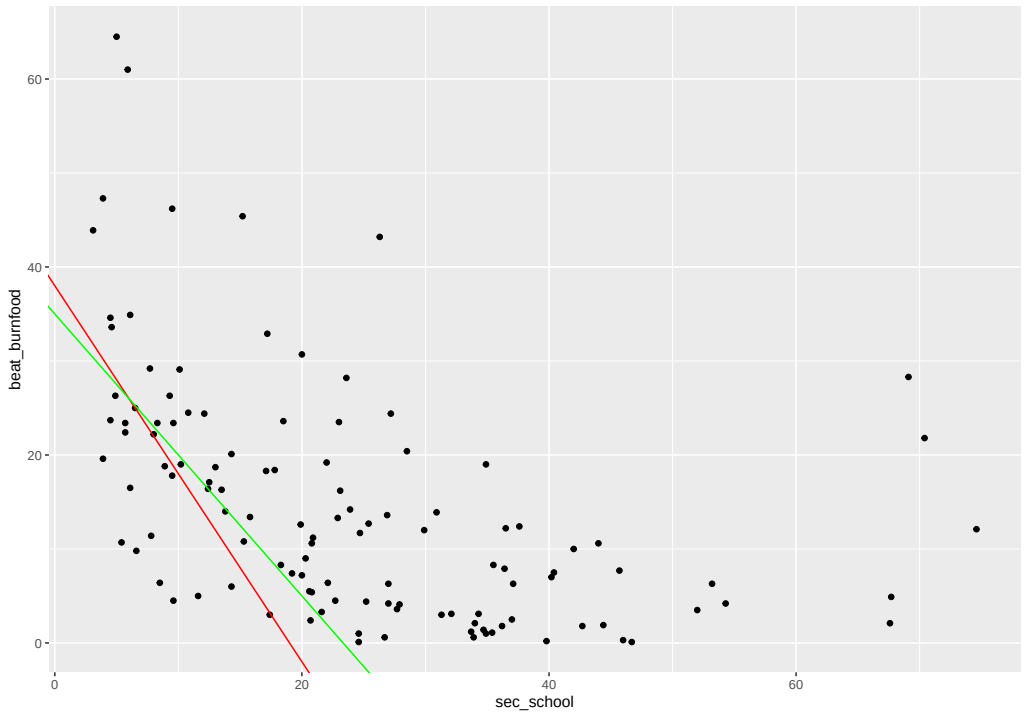
Association entre beat-burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(aes(intercept = 38, slope = -2), color = "red
```



Association entre beat-burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(aes(intercept = 38, slope = -2), color = "red
  geom_abline(aes(intercept = 35, slope = -1.5), color = "g
```



Association entre beat-burnfood et niveau d'éducation

- Maintenant, simulons un ensemble de droites pour voir si la meilleure candidate figure parmi elles.
- Pour ce faire, je crée 150 intersections (intercept) et 150 pentes (slopes), selon une loi uniforme.
- On parle de loi uniforme quand tous les éléments ont les mêmes probabilités d'être choisis. (Dire qu'une variable aléatoire X suit une loi uniforme sur l'intervalle [a, b]) signifie que sa densité f est la fonction constante égale à 1/(b-a).

Association entre beat-burnfood et niveau d'éducation

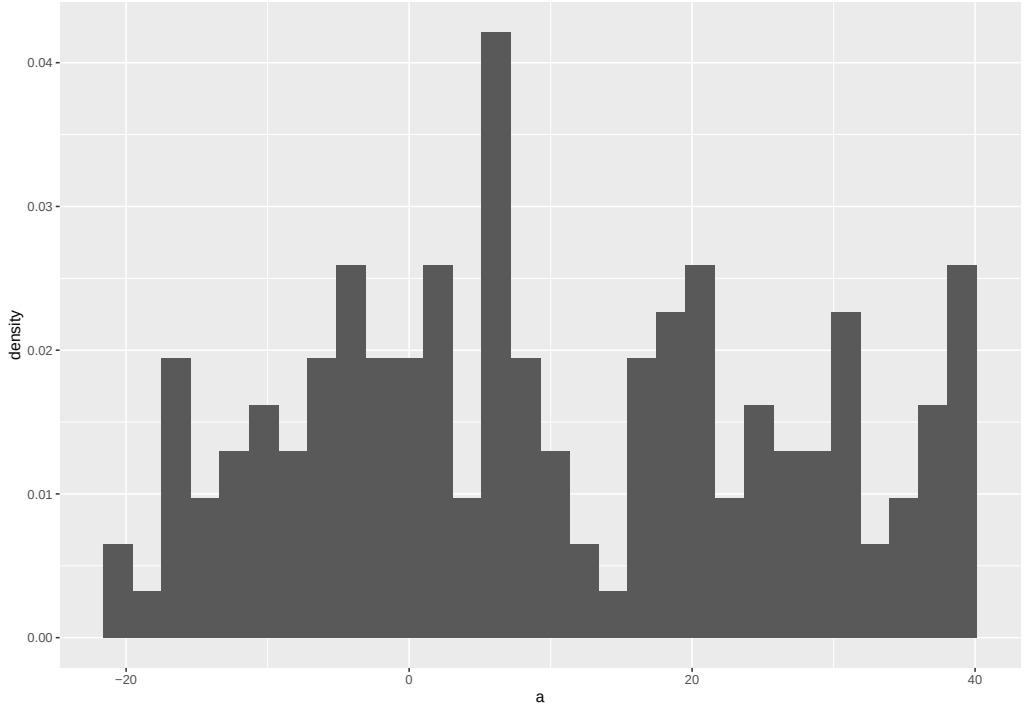
```
#library(modelr)

set.seed(2001)
models <- tibble(
  a = runif(150, -20, 40),
  b = runif(150, -5, 5)
)
models
```

```
## # A tibble: 150 x 2
##          a       b
##      <dbl>   <dbl>
##  1  25.5     4.48
##  2  16.5    -0.682
##  3  -6.90    3.67
```

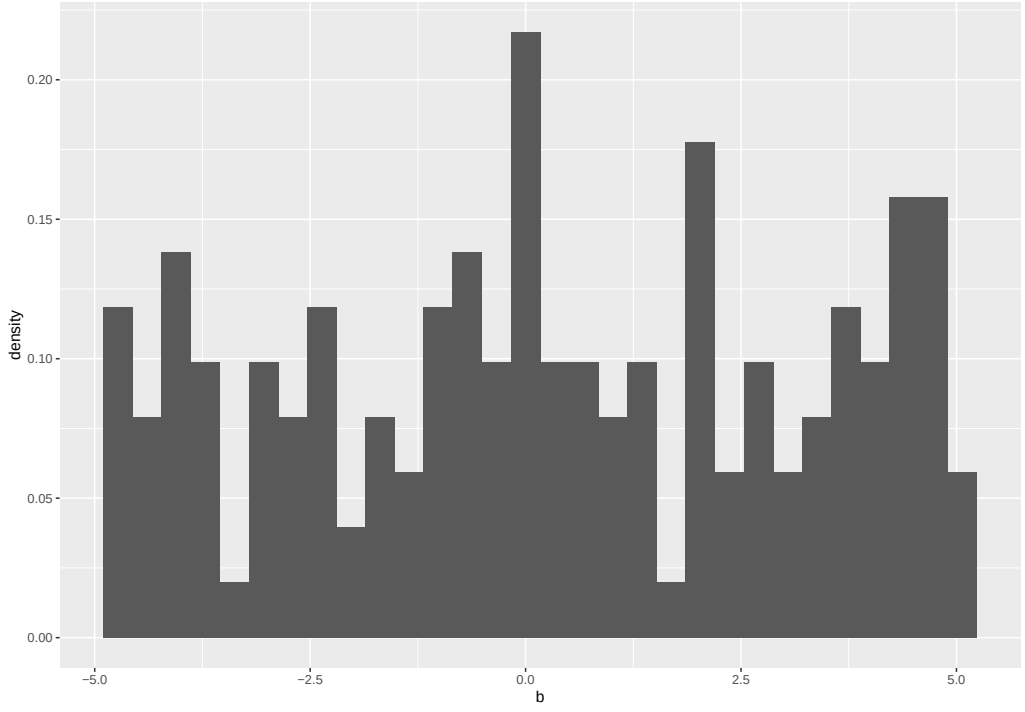
Voyons les distributions de a

```
ggplot(models, aes(x = a, y = ..density..)) +
  geom_histogram()
```



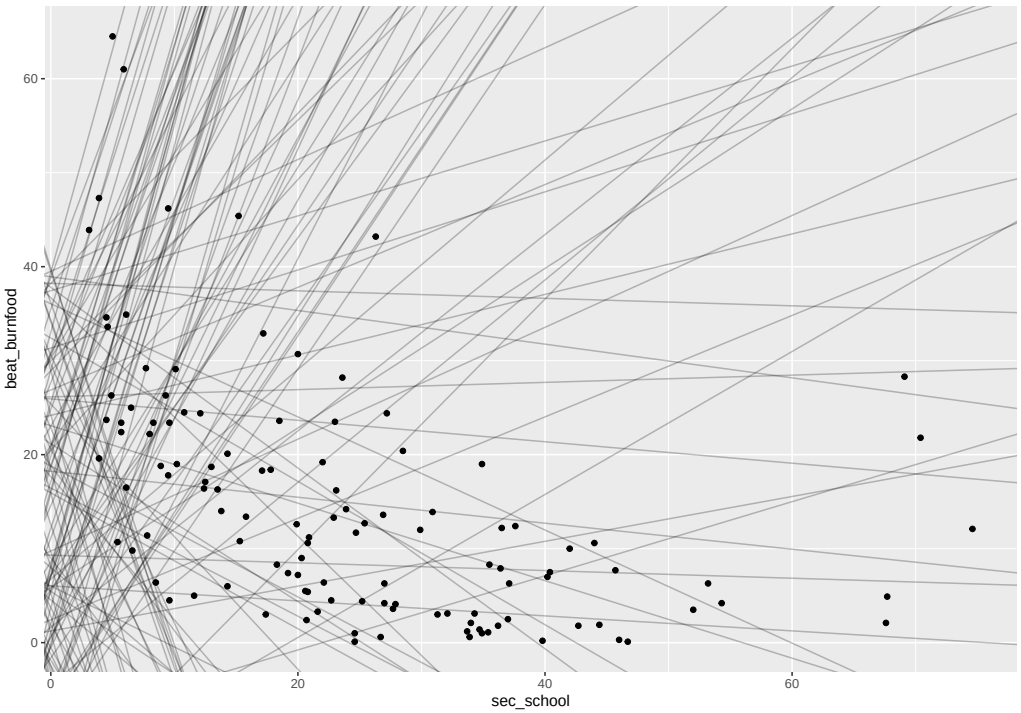
Voyons les distributions de b

```
ggplot(models, aes(x = b, y = ..density..)) +
  geom_histogram()
```



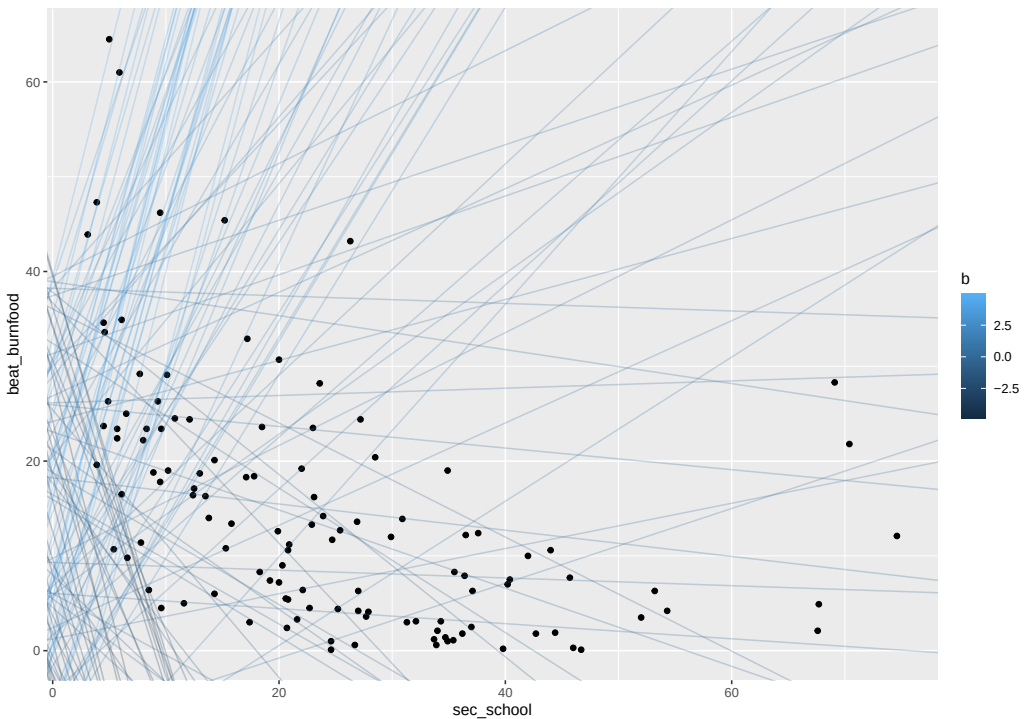
Association entre beat-burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(data = models, aes(intercept = a, slope = b)
```



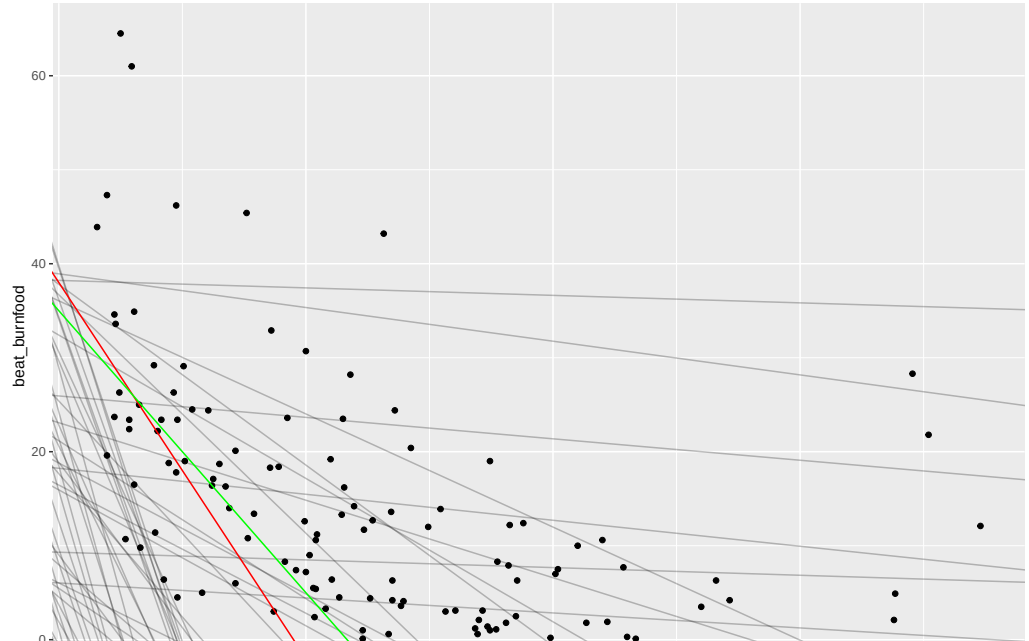
Association entre beat-burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(data = models, aes(intercept = a, slope = b,
```



Association entre beat-burnfood et niveau d'éducation

```
ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(data = models %>% filter(b < 0), aes(intercep
  geom_abline(aes(intercept = 38, slope = -2), color = "re
  geom_abline(aes(intercept = 35, slope = -1.5), color = "g
```

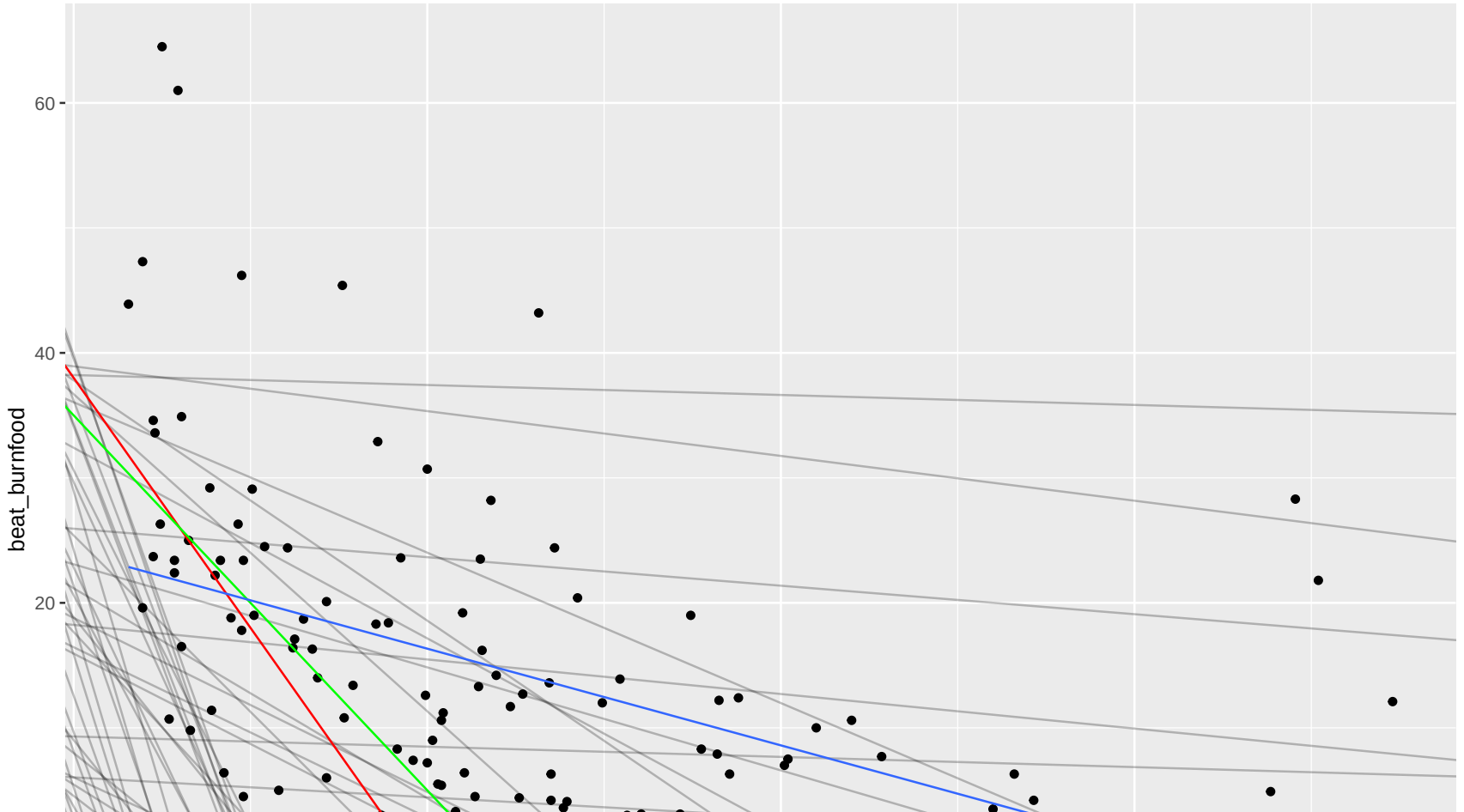


Association entre beat-burnfood et niveau d'éducation

```
graph1 <- ggplot(dhs_ipv) +
  geom_point(aes(x = sec_school, y = beat_burnfood)) +
  geom_abline(data = models %>% filter( b <0 ), aes(interce
  geom_abline(aes(intercept = 38, slope = -2), color = "red
  geom_abline(aes(intercept = 35, slope = -1.5), color = "g
  geom_smooth(aes(x = sec_school, y = beat_burnfood), metho
```

Association entre beat-burnfood et niveau d'éducation

```
graph1
```



Labo

Interprétation des résultats

- Variables continues
- Variables dichotomiques
- Variables catégorielles