Fondements statistiques de la méthode expérimentale en technologie éducative

Mattia A. Fritz

01/04/2023

Résumé

La méthode expérimentale est étroitement liée à la modélisation des données notamment avec des finalités d'inférence statistique. Le rapport entre les données, la modélisation de celles-ci, et les conclusions qu'on peut tirer des tests statistiques effectués est cependant complexe et régit par différentes approches philosophiques et méthodologiques. À complement du document sur les fondements empiriques de la méthode expérimentale, ce document s'intéresse donc aux fondements statistiques de cette méthode. Il propose d'abord une introduction à la modélisation des données et aux probabilités. Ensuite, les statistiques *fréquentistes* (ou *classiques*) et les statistiques Bayésiennes sont illustrées dans les grandes lignes. En guise de conclusion, le document propose une brève analyse sur les avantages des statistiques Bayésiennes dans un contexte pédagogique.

Introduction

Très peu des chercheurs en sciences sociales s'intéressent aux statistiques. Elles sont souvent vécues comme un *mal nécessaire* pour pouvoir bénéficier des avantages épistémologiques fournis par la méthode expérimentale. Cette perspective se reflète souvent dans les manuels de méthode expérimentale qui associent les différents design expérimentaux avec les tests statistiques correspondantes, afin d'illustrer comment les résultats de ces tests doivent être interprétés en relation aux hypothèses opérationnelles d'abord, et théoriques ensuite. Il en suit que les statistiques ressortent de cette démarche comme une liste de recettes à appliquer machinalement selon les *ingrédients* du plan expérimental: si j'ai n variables indépendantes de type t et m variables dépendantes de type u, alors il faut utiliser le test x. Cette démarche mécanique est d'ailleurs corroborée par des diagrammes qu'on peut facilement trouver dans des manuels ou en ligne, dont l'objectif est précisément d'accompagner les chercheurs à travers un nombre de bifurcations avant de trouver, enfin, le test nécessaire à leurs besoin. Malheureusement, cette approche ne semble pas donner ses fruits, car il existe désormais

plusieurs témoignages dans la littérature scientifique qui montrent comme les statistiques sont mal comprises et utilisées dans les contributions expérimentales (Bakker & Wicherts, 2011; Greenland et al., 2016; Nickerson, 2000; Nuijten et al., 2016; Singmann et al., 2023). Même les ressources pédagogiques qui sont censées former les nouveaux chercheurs présentent souvent des erreurs dans l'exposition de concepts clé (Cobb, 2007; Maurer et al., 2019; McElreath, 2020).

Cette contribution utilise une approche différente caractérisée en premier lieu par la séparation entre les fondements empiriques et les fondements statistiques de la méthode expérimentale. Ce document s'intéresse à ces dernier et le fait de manière à dévoiler des concepts importants qui, dans une approche machinale, sont cachés ou traités comme s'il n'existaient pas d'alternatives (Lakens, 2021; McElreath, 2020; Rodgers, 2010). L'objectif de cette contribution est de sensibiliser des étudiant-es en technologie éducative sans un background en méthode expérimentale à l'importance que les statistiques jouent dans la démarche expérimentale. Pour ce faire, le texte présente d'abord une brève introduction à la modélisation des données, c'est-à-dire à l'utilisation de techniques mathématiques et probabilistes pour manier et représenter des données empiriques. Ensuite, le document illustre deux approches et philosophies différentes aux statistiques: l'approche *fréquentiste* ou *classique*, dominante en sciences sociales depuis plusieurs décennies, et l'approche Bayésienne, qui commencent à s'introduire dans certaines contributions. Les deux approches sont illustrées dans les grandes lignes avec l'objectif de formuler qu'est-ce qu'on peut *vraiment* demander aux tests statistiques, et comment leur réponses peuvent nous aider dans la démarche explicative de la méthode expérimentale. En guise de conclusion, ce document prend position par rapport aux deux approches et suggère que les statistiques Bayésienne devraient être enseignées en priorités dans les cours de méthodologie. Cette position est brièvement argumentée.

1 Modélisation des données

Les chercheurs récoltent des données empiriques dans la tentative de pouvoir en tirer des informations utiles qui, souvent, dépassent le cadre spécifique des données récoltées et visent plutôt à des affirmations ou principes avec une portée plus étendue. Ce mécanisme qui permet de passer du particulier à un cadre plus généralisé est souvent identifié avec le terme d'**inférence**. On peut identifier différents objectifs relatifs à l'inférence (Gelman et al., 2021; McElreath, 2020; Rodgers, 2010).

Réduction des données à des indicateurs représentatifs. Grâce à la représentation en forme d'indicateurs on peut *inférer* des caractéristiques d'un ensemble de données. Un exemple d'inférence de ce type sont la moyenne et l'écart type tirés d'un nombre

d'observations. Quand on indique que les notes obtenues dans un cours ont une moyenne de $M = 3.5$ ($SD = 0.8$) on peut par exemple inférer que la plupart des étudiant-es n'ont pas validé le cours, car la moyenne est inférieur au barème de 4. En même temps, l'écart type qui est représentatif de la dispersion des données indique que certain-es étudiant-es ont sûrement dépassé ce barème, car si on ajoute un écart type à la moyenne, on obtient 4.3, et en ajoutant deux écarts types on obtient 5.1. Avec ce calcul on peut également inférer que le système de notation de l'enseignant-e est assez sévère, car il y a peu de chances qu'une étudiant-e ait obtenu une très bonne note.

Évaluer le degré d'association entre des variables. On peut utiliser un jeu de données pour établir une mesure d'intérêt et utiliser d'autres variables récoltées pour faire des comparaisons depuis lesquelles inférer, par exemple, si les observations avec un certain type de valeurs sur une variable sont associées à des valeurs différentes sur la mesure d'intérêt. Par exemple, dans un jeu de données qui s'intéresse à la mesure du sentiment de bien être des étudiant-es dans un cursus de bachelor, on peut comparer si la perception de bien être diffère entre la première, la deuxième et la troisième année. Si c'est le cas, on peut inférer la présence d'une association entre les deux variables.

Prédire des évènements futurs. Des données récoltées peuvent être utilisées pour produire un *algorithme* qui prend des variables comme Input et produisent la mesure d'intérêt comme Output. Une fois cet algorithme de conversion obtenu avec les données observées, on peut entrer des nouvelles données pour estimer la mesure d'intérêt. Dans ce mécanisme, on infère que le comportement des nouvelles données sera similaire au *comportement* des données observées. Cette inférence est traduite par l'algorithme lui-même. Ce principe est notamment à la base des techniques de *machine learning*.

Déterminer l'effet d'une intervention. À travers des données qui varient systématiquement sur une ou plusieurs variables dont les valeurs ont été fixées par les chercheurs, on peut inférer des mécanismes contre-factuels du type: que se serait-il passé si la personne avait été attribuée à une autre valeur/modalité de l'intervention? Dans ce contexte, l'objectif de l'inférence est en général double: (1) déterminer si l'effet est présent dans le *micro-monde* des données observées, et (2) éstimer à quel point on peut être confiant que l'effet puisse se réproduire au *macro-monde*.

Comme il a été indiqué dans les fondements empiriques de la méthode expérimentale, les expériences appartiennent donc à ce dernier cas de figure. Cependant, l'inférence depuis des données empiriques s'appuie en général sur les mêmes instruments: des **modèles mathématiques**. Ceci est souvent source de confusion, surtout dans un contexte introductif aux méthodes dits *quantitatifs*. En effet, très concrètement, on utilise souvent les mêmes logiciels et les mêmes *fonctions/tests* à l'intérieur de ces logiciels indépendamment de comment les

données ont été créées (e.g. observation, simulation ou expérience). Il en résulte une compréhensible difficulté à cerner les différences qui ne résident en effet pas dans les modèles eux-mêmes, mais plutôt dans les connaissances scientifiques relatives à:

- Les relations causales entre les variables impliquées dans le modèle, qui peuvent notamment être explicitées avec un modèle structurel de causalité sous forme de *Directed Acyclic Graph* (DAG) (Bareinboim & Pearl, 2016; Cinelli et al., 2020; Pearl, 2000; Pearl et al., 2016; Pearl & Mackenzie, 2018);
- Le processus génératif des données, c'est-à-dire sous quelles conditions les données ont été produites (Maxwell et al., 2017; McElreath, 2020).

Sur la base de ces informations, les modèles mathématiques et les indicateurs que ces modèles produisent doivent être interprétés de manière conforme à ce qu'ils permettent ou ne permettent pas d'inférer. À ce propos, cette section illustre d'abord en quoi consiste la modélisation des données et quels sont ces avantages. L'étape suivante indroduit la *famille* de modèles la plus fréquente en science sociale: la modélisation linéaire. Ensuite, cette famille de modèle sera centrée plus spécifiquement dans le cadre des expériences, notamment en ce qui concerne la perspective contre-factuelle. Enfin, elle introduit le concept d'inférence statistique qui sera ensuite décliné dans les deux approches fréquentiste et Bayésien.

1.1 Pourquoi modéliser des données?

L'une des questions souvent inexplorée dans les manuels de méthodologie *quantitative* est la suivante: pourquoi avons-nous besoins des statistiques en premier lieu? La vie des chercheurs – et des étudiant-es! – seraient tellement plus simple sans elles. Il est donc légitime de s'attendre à ce que leur association pratiquement indissoluble avec la recherche *quantitative* soit justifiée par un apport exceptionnel en termes épistémologiques.

Malheureusement, ce n'est pas vraiment le cas, au moins selon cette contribution qui adopte une attitude désenchantée. Le rôle prépondérant des statistiques dans la recherche s'explique principalement par deux raisons:

- La complexité des phénomènes étudiées qui sont souvent dépendantes d'un large éventail de facteurs qui s'influencent mutuellement. Dans cette complexité, même des patterns *stables* peuvent ne pas se produire à chaque fois, mais seulement *la plupart des fois*. Par exemple, il existe un lien de causalité stable entre les heures passées à étudier et la réussite à un examen. Mais il se peut que, parfois, un-e étudiant-e qui a beaucoup étudié rencontre un échec, et un-e étudiant-e qui n'a pas beaucoup étudié réussisse néanmoins l'examen.

- La tendance intrinsèque aux êtres humains à voir des liens de cause à effet lorsqu'en réalité il ne s'agit que d'épiphénomènes circonstantielles, dont l'occurrence est indépendant des causes supposées par la personne.

Ces deux raisons sont brièvement développées par la suite. En guise de conclusion de cette partie introductive sur la modélisation, une définition formelle d'un modèle applicable dans le contexte statistique sera fournie.

1.1.1 Vivre – et faire de la recherche – dans un monde pseudo-déterministe

1.1.2 Se protéger de ses propres biais

1.1.3 Qu'est-ce qu'un modèle statistique

1.2 La modélisation linéaire

Il existe différentes manières pour modéliser des données. Dans les sciences sociales, les modèles les plus utilisés appartiennent à la *famille* des modélisations linéaires. Ces modèles se caractérisent par le fait que la variable d'intérêt (le outcome ou la mesure) peut être représenté par ce qu'on appelle une équation de régression. L'équation de régression dans les modèles linéaires correspond à l'équation représentant une pente dans un plan cartésien:

$$Y = Intercepte + Peinte \times X$$

Dans les manuels statistiques cette équation est plus souvent représentée de la manière suivante:

$$\mathrm{y_i} = \beta_0 + \beta_1(\mathrm{x_i}) + \epsilon_i$$

L'explication des composantes de cette version de l'équation est la suivante:

- $\mathrm{y_i}$ est la mesure de la variable Y pour l'observation i dans le jeu de données, c'est-à-dire $\mathrm{y_1}$, $\mathrm{y_2}$, … $\mathrm{y_n}$. Malheureusement, en statistique on fait souvent la distinction entre notations qui pour les non-statisticiens ne sont pas très marquée, comme la distinction entre minuscule et majuscule ou entre lettre latine et grecque.
- β_0 correspond à l'intercepte, c'est-à-dire la valeur de Y lorsque toutes les éventuelles variables sur la droite de l'équation (dans ce cas seulement X) équivalent à 0.
- β_1 correspond au paramètre (ou coefficient de régression) pour la variable X. Ce coefficient est commun à toutes les valeurs observées pour X, c'est-à-dire $\mathrm{x_i}$.

- ϵ_i correspond à ce qu'on appelle le résidu de l'équation. Il s'agit d'une valeur de *compensation* dû au fait que les paramètres β_0 et β_1 sont communs à toutes les observations, mais pratiquement jamais les paramètres du modèles permettent d'arriver *exactement* à la valeur de y_i correspondante.

Contrairement à ce qui est souvent indiqué, une régression linéaire ne veut pas forcément dire *rectilinéaire*. En effet, on peut par exemple créer des courbes en ajoutant des valeurs exponentielles aux variables prédictives:

$$\checkmark \; y_i = \beta_0 + \beta_1(x_i^2) + \epsilon_i$$

Par contre, il n'est pas possible d'utiliser des exponentiels pour les coefficients de régression. Par exemple, cette équation ne serait pas considérée comme un modèle linéaire:

$$\times \; y_i = \beta_0 + \beta_1^{x_i} + \epsilon_i$$

Ce préambule très technique, qui sera développé davantage dans les exemples plus bas, sert à ce point pour indiquer que les modèles linéaires sont des modèles qu'on peut facilement accommoder pour prendre en ligne de compte plusieurs types de relations entre les variables prédictives et la mesure de outcome. Cette flexibilité se traduit malheureusement dans la littérature scientifique avec des noms d'analyses différentes qui sont en réalité de cas spéciaux de la modélisation linéraire. Ce tableau propose une liste de ces *cas spéciaux*.

Table 1: Cas spéciaux de la modélisation linéaire présents dans la littérature scientifique

Nom de l'analyse	Outome	Variable prédictive
Régression simple	1 continue	1 continue
Régression multiple	1 continue	1 continue ou plus
t-test à groupes indépendants	1 continue	1 catégorielle avec 2 modalités
ANOVA simple	1 continue	1 catégorielle avec plus de 2 modalités
ANOVA factorielle	1 continue	2 catégorielles ou plus
ANCOVA	1 continue	1 catégorielle ou plus et 1 continue ou plus

Table 1: Cas spéciaux de la modélisation linéaire présents dans la littérature scientifique *(continued)*

Nom de l'analyse	Outome	Variable prédictive
Corrélation de Pearson	1 continue standardisée	1 continue standardisée
t-test avec un seul groupe	1 continue	Intercepte seulement
t-test apparié	Différence entre 2 continues	Intercepte seulement
t-test de Hotelling	2 continues ou plus	1 catégorielle avec 2 modalités
MANOVA	2 continues ou plus	1 catégorielle avec plus de 2 catégorielles
Régression multiple multivariée	2 continues ou plus	1 catégorielle ou plus et 1 continue ou plus

Comme le tableau l'indique, ces cas spécieaux présupposent que la mesure soient toujours représentée sur une échelle continue poténtiellement infinie sur les côtés. Ceci n'est cependant pas toujours le cas, par exemple lorsque la mesure est exprimée sur des échèlles de Lickert ou sur une échelle avec des limites inférieurs et/ou supérieurs. Dans le reste de cette partie, le texte propose des exemples concrets de différentes typologies de modèles linéaires dans lesquels le modèle suivante est une version plus flexible du précédent et peut donc être appliqué à des données plus complexes. Les modèles sont dans l'ordre:

- Modèle linéaire simple
- Modèle linéaire multiple
- Modèle linéaire généralisée
- Modèle linéaire généralisée mixte

1.2.1 Modèle linéaire simple

Comme indiqué plus haut, le modèle linéaire simple présuppose que la variable outcome est le résultat de l'addition entre l'intercepte, la peinte d'une variable prédictive, et le résidu:

$$\mathrm{y_i} = \beta_0 + \beta_1(\mathrm{x_i}) + \epsilon_i$$

Voici un jeu de données avec 10 observations. Chaque observation se compose simplement de la variable prédictive x et de la variable outcome y.

Table 2: Jeu de données pour une régression linéaire simple.

i	x	y
1	16.64	101.99
2	13.69	107.25
3	16.61	125.49
4	20.00	130.52
5	15.70	96.45
6	12.81	80.77
7	10.00	50.00
8	11.78	63.24
9	14.84	102.25
10	19.31	200.00

Lorsqu'on demande à un logiciel d'analyse statistique comme par exemple R de mener une régression linéaire simple sur ces données, le résultat qu'on obtient sera le suivant.

Table 3: Tableau des paramètres d'une régression linéaire simple dans une perspective fréquentiste.

Predictor	*b*	95% CI	*t*	*df*	*p*
Intercept	-63.44	[-148.49, 21.61]	-1.72	8	.124
X	11.18	[5.67, 16.69]	4.68	8	.002

Le résultat se réfère à une analyse de type fréquentiste, mais pour l'instant cet aspect n'a pas d'importance. Ce qui est à noter depuis les résultats sont les deux coefficients attribués aux paramètres β_0 = -63.44 et β_1 = 11.18. Il en résulte qu'on peut écrire les résultats de notre modèle également sous forme d'équation, mais cette fois-ci avec les coefficients intégrés.

$$\hat{\mathrm{y}} = -63.44 + 11.18(\mathrm{x}) \tag{1}$$

Cette équation diffère de celle générique présentée plus haut de deux manières. En premier lieu, y a été remplacé par $\hat{\mathrm{y}}$. Lorsque un élément à un *chapeau* en statistique, cela signifie qu'il s'agit d'une estimation liée à l'utilisation d'un modèle. Donc $\hat{\mathrm{y}}$ n'est pas une valeur observée, mais une valeur calculée sur la base des paramètres inférés par le modèle. La deuxième différence consiste dans la disparition du résidu ϵ. En effet, en s'agissant d'une prédiction, on ne peut pas savoir à quel point ce valeur s'éloigne de la *vraie* valeur. Par contre, on peut

calculer cette distance si on compare les valeurs observées du jeu de données avec les valeurs qui seraient prédites par le modèle lorsque x a exactement la même valeur de l'observation dans le jeu de données. La figure suivante montre ce principe graphiquement.

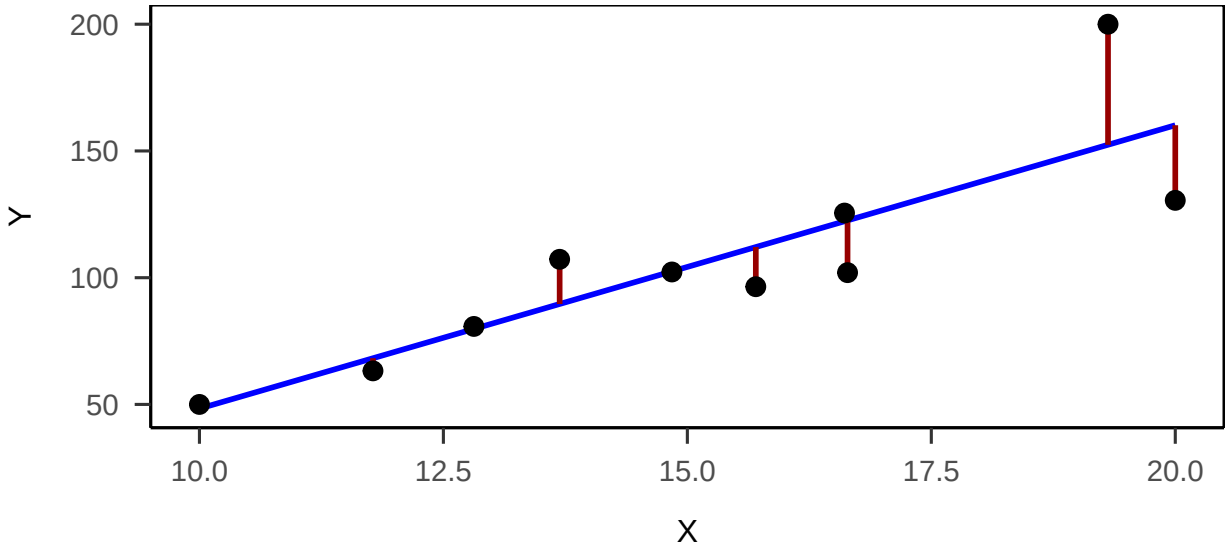


Figure 1: Représentation graphique d'une régression linéaire simple. La ligne bleu représente les valeurs prédites par le modèle. Les segments rouges entre les points et la ligne bleu sont les résidus des données observées.

En faites, c'est exactement avec ce type de processus que les paramètres du modèle linéaire sont calculés en premier lieu. En effet, la régression linéaire simple consiste à trouver la *ligne* qui passe à travers les observations et minimisent la distance avec toutes les observations du jeu de données. Le tableau suivant reprend à ce propos les observations, mais ajoute trois autres colonnes:

- $\hat{y}$: la valeur prédite par le modèle pour une observation x avec la même valeur du jeu de donnée
- $y - \hat{y}$: la différence entre la valeur observée et la valeur prédite
- $(y - \hat{y})^2$: la différence élevée au carré. L'utilisation de l'exponentiel sert deux objectifs: (1) éviter que dans la somme de toutes les différences les valeurs positives et négatives s'annulent, et (2) donner plus de poids aux distances plus extrêmes, comme par exemple la dernière, dont la différence de 47.56 unités devient de 2262.41.

Table 4: Jeu de données avec valeur prédite et différence (normale et au carré) entre la valeur observée et prédite

i	x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
1	16.64	101.99	122.60	-20.60	424.44
2	13.69	107.25	89.61	17.64	311.16
3	16.61	125.49	122.25	3.25	10.54
4	20.00	130.52	160.14	-29.62	877.52
5	15.70	96.45	112.08	-15.64	244.56
6	12.81	80.77	79.79	0.98	0.96
7	10.00	50.00	48.35	1.65	2.73
8	11.78	63.24	68.23	-4.99	24.91
9	14.84	102.25	102.48	-0.23	0.05
10	19.31	200.00	152.44	47.56	2262.41

La somme de $(y - \hat{y})^2$ est notamment utilisée comme mesure pour établir à quel point le modèle s'encastre avec les données observées. Le plus cette somme est élevée, le moins le modèle est en adéquation avec les données. Est-ce que la somme de 4,159.27 est élevée? C'est difficile à dire sans une mesure de comparaison. À cet effet, on peut par exemple utiliser la moyenne $\overline{Y} = 105.80$ comme modèle linéaire alternative. La moyenne peut être tout à fait considérée un modèle linéaire qui a seulement l'intercepte. L'intercepte correspond justement à la moyenne de la variable considérée. Le tableau suivant reprend la même structure du précédent, mais en utilisant la valeur fixe de la moyenne pour calculer la distance des observations.

Table 5: Jeu de données avec valeur prédite et différence (normale et au carré) entre la valeur observée et la moyenne de Y

i	x	y	$\hat{y}$	$y - \overline{Y}$	$(y - \overline{Y})^2$
1	16.64	101.99	105.8	-3.80	14.46
2	13.69	107.25	105.8	1.46	2.12
3	16.61	125.49	105.8	19.70	388.05
4	20.00	130.52	105.8	24.72	611.06
5	15.70	96.45	105.8	-9.35	87.41
6	12.81	80.77	105.8	-25.03	626.41
7	10.00	50.00	105.8	-55.80	3113.18
8	11.78	63.24	105.8	-42.56	1811.11

Table 5: Jeu de données avec valeur prédite et différence (normale et au carré) entre la valeur observée et la moyenne de Y *(continued)*

i	x	y	$\hat{y}$	$y - \overline{Y}$	$(y - \overline{Y})^2$
9	14.84	102.25	105.8	-3.55	12.58
10	19.31	200.00	105.8	94.20	8874.42

La somme des distances au carré en utilisant $\overline{Y}$ comme modèle linéaire est de 15,540.79, c'est à dire 11,381.52 plus du modèle dérivé de la régression linéaire simple. On peut notamment inférer depuis cette différence qu'en connaissant la valeur de x, on peut avoir une meilleure idée de la valeur de y. En termes formelles: $\mathbb{P}(Y|X) \neq \mathbb{P}(Y)$.

Comme indiqué plus haut, le modèle linéaire permet de prendre en compte d'autres *formes* de régression, comme par exemple une courbe. Cependant, il faut considérer deux aspects très importants dans la modélisation: le sous-ajustement et le sur-ajustement d'un modèle. La figure ci-dessous montre, sur la gauche, un modèle qui est bien ajusté aux données, car la ligne passe bien au milieu des différents points. Sur la droite, en revanche, la ligne rouge est *sous-ajustée* car elle infère une association linéaire qui, dans les données, ne semblent pas être présente. La ligne violet, au contraire, est très probablement sur-ajustée, car elle se superpose parfaitement aux données. Il est difficile de croire, cependant, que ce modèle puisse s'adapter à d'autres données que celles-ci. La modélisation doit toujours faire face à un équilibre précaire entre la précision *in-sample* et l'utilité *out-of-sample* (McElreath, 2020).

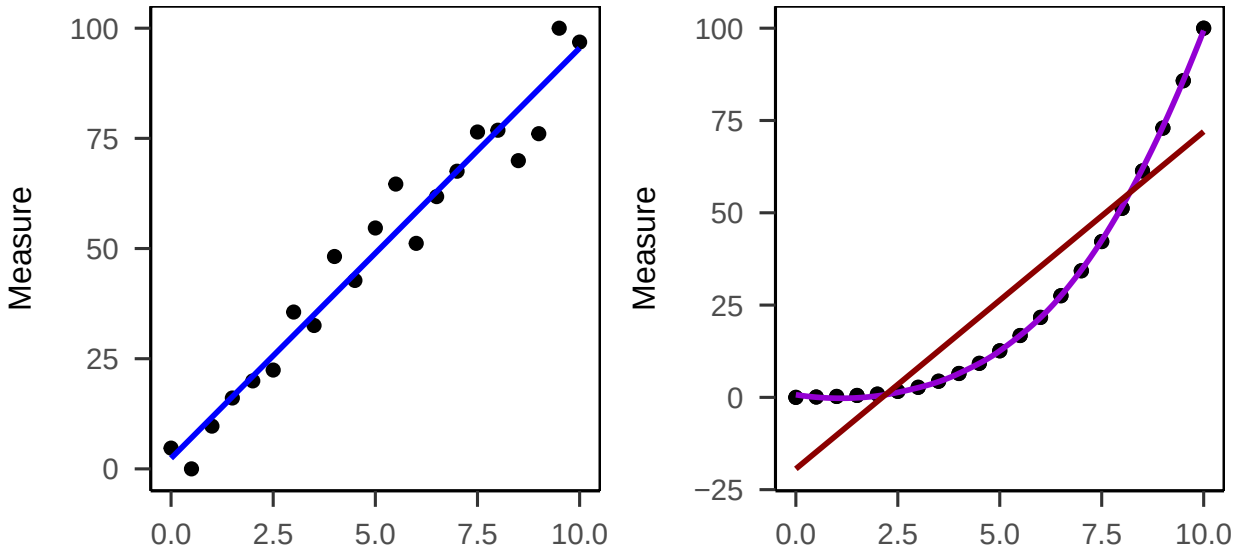


Figure 2: Représentation graphique d'un modèle qui est bien ajusté (gauche), un modèle sous-ajusté (ligne rouge à droite) et sur-ajusté (ligne violet à droite)

1.2.2 Modèle linéaire multiple

Le modèle linéaire multiple, connu aussi comme modéle linéaire général, est simplement une extension de la régression linéaire simple à plusieurs variables prédictives. Dans l'exemple précédent, le modèle consistait dans une seule variable numérique, mais le modèle linéaire multiple accepte également des variables prédictives binaires ou catégorielles. Par exemple, dans le jeu de données suivante, la variable outcome Y est calculée sur la base de l'effet additive entre les variables numériques X et W, plus la variable catégorielle Z qui possède trois modalités: *Faible*, *Moyenne* et *Forte*. L'équation de la régression linéaire multiple est donc la suivante:

$$\mathrm{y_i} = \beta_0 + \beta_1(\mathrm{x_i}) + \beta_2(\mathrm{w_i}) + \beta_3(\mathrm{z_{iMoyenne}}) + \beta_4(\mathrm{z_{iForte}}) + \epsilon_i$$

Il est utile de remarquer comme dans cette équation n'apparaît pas la modalité faible de la variable Z. Ceci s'explique par un mécanisme adopté souvent en régression linéaire qui consiste à attribuer à la première modalité d'une variable catégorielle une sorte de valeur de base. Ensuite, on attribue aux autres modalités une *dummy* variable, c'est-à-dire une variable qui assume les valeurs 0 ou 1. Dans ce cas, si l'observation appartient à la modalité *Moyenne*, alors $\mathrm{z_{iMoyenne}}$ sera 1 et $\mathrm{z_{iForte}}$ sera 0. Si la variable appartient au contraire à *Forte*, les valeurs seront inversés. Ceci à la conséquence d'additionner dans le calcul seulement le coefficient de la variable avec valeur 1 et d'annuler le coefficient de la variable avec valeur 0. En effet, multiplier n'importe quel coefficient par 1 signifie ajouter une fois ce coefficient au calcul, tandis que le multiplier par 0 signifie ajouter 0 au calcul. Le tableau suivant montre les premières 10 observations d'un jeu de données qui compte $N = 200$.

Table 6: Extrait des premières 10 observations d'un jeu de données avec plusieurs variables prédictives.

i	x	w	z	y
1	11.917	9.929	Moyenne	51.840
2	17.232	10.799	Moyenne	97.255
3	13.968	14.466	Moyenne	72.491
4	6.645	14.047	Faible	24.949
5	15.063	3.767	Moyenne	79.507
6	17.450	11.484	Forte	93.480
7	10.855	5.691	Forte	71.177
8	2.428	8.581	Forte	17.317
9	14.297	1.634	Faible	47.803

Table 6: Extrait des premières 10 observations d'un jeu de données avec plusieurs variables prédictives. *(continued)*

i	x	w	z	y
10	14.671	7.361	Forte	80.809

Le tableau de la régression multiple doit à ce moment proposer les coefficients pour X, Y, ainsi que pour les deux modalités *Moyenne* et *Forte* de Z.

Table 7: Tableau des paramètres d'une régression linéaire multiple dans une perspective fréquentiste.

Predictor	*b*	95% CI	*t*	*df*	*p*
Intercept	-3.25	[-8.84, 2.35]	-1.14	195	.254
X	4.73	[4.36, 5.10]	25.20	195	$< .001$
W	0.11	[-0.23, 0.44]	0.62	195	.537
ZMoyenne	5.13	[1.79, 8.48]	3.02	195	.003
ZForte	7.69	[4.40, 10.98]	4.62	195	$< .001$

Encore une fois, les résultats de ce tableau font référence à l'approche fréquentiste, mais ces sont seulement les coefficients qui nous intéressent pour le moment. L'équation de régression avec les coefficients estimés sera donc la suivante:

$$\hat{\text{y}} = -3.25 + 4.73(\text{x}) + 0.11(\text{w}) + 5.13(\text{z}_{\text{Moyenne}}) + 7.69(\text{z}_{\text{Forte}}) \tag{2}$$

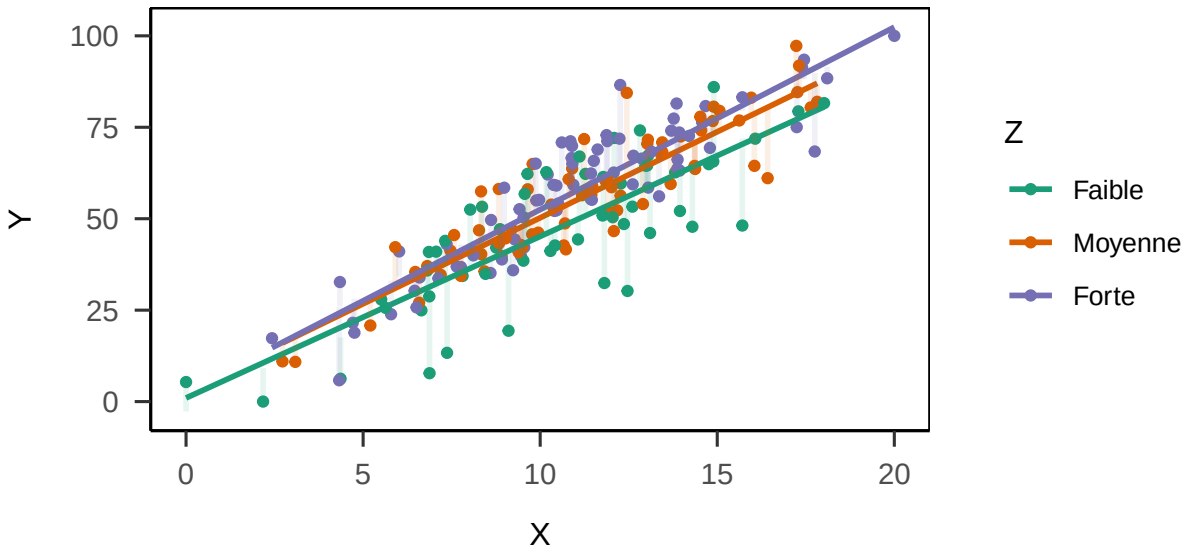
Un élément d'intérêt particulier dans la régression linéaire multiple concerne l'interprétation des coefficients les uns par rapport aux autres. En effet, dans ce type de modèle, la contribution additive de chaque coefficient est à interpréter comme une contribution unique par rapport aux autres coefficients. Voici comment interpréter chaque coefficient en termes de changements sur la variable outcome Y:

- β_0 Intercept(-3.25) : Lorsque toutes les variables prédictives (X, W et Z) sont égales à zéro, la valeur prédite de la variable Y est de -3,25. Cependant, cette interprétation peut ne pas être saillante si la valeur zéro pour l'une des variables prédictives n'est pas plausible dans le contexte des données, ou même si une variable négative de l'outcome Y ne fait pas de sens (e.g. temps négatif).
- β_1 x(4.73) : Deux observations qui diffèrent d'une unité sur la variable X, tout en maintenant les autres variables (w et z) constantes, diffèrent sur la variable outcome Y de 4,73 unités.

- β_2 w(0.11) : Deux observations qui diffèrent d'une unité sur la variable W, les autres variables (x et z) restant constantes, diffèrent de 0,11 unité sur la variable outcome Y.
- β_3 zMoyenne(5.13) : Il s'agit d'une variable catégorielle avec comme catégorie de référence la modalité *Faible*. Dans ce cas, une observation avec modalité *Moyenne* diffère d'une observation avec modalité *Faible* de 5,13 unités sur la variable outcome Y, tout en maintenant les autres variables (X et W) constantes.
- β_4 zForte(7.69) : De même, le coefficient de 7,69 représente la différence sur la variable outcome Y entre une observation avec modalité *Forte* et une avec modalité *Faible*, tout en gardant les autres variables (X et W) constantes.
- On peut récupérer le coefficient qui détermine la différence entre une observation avec modalité *Forte* et une avec modalité *Moyenne* à travers la substraction 7.69 (zForte) - 5.13 (zMoyenne) = 2.56. En d'autres termes, deux observations qui gardent X et W constantes diffèrent de 2.56 unités sur la variable outcome Y lorsque l'une est avec modalité *Forte* et l'autre modalité *Moyenne*.

La comparaison entre les modalités de Z est de quelque sorte plus compliquée à comprendre, car la valeur de *Faible* n'est pas vraiment explicitée. En effet, la valeur de Y pour les observations avec modalité *Faible* correspond tout simplement à l'effet additive de X et W, car zMoyenne et zForte sont annulés par la multiplication par 0.

L'interprétation des coefficients d'une régression linéaire multiple nécessite de beaucoup de pratique avant d'être maîtrisée. Mais ces coefficients jouent un rôle fondamental dans l'inférence et il était donc nécessaire d'en illustrer le mécanisme. D'ailleurs, une régression linéaire multiple est également plus difficile à visualiser graphiquement par rapport à la régression linéaire simple. Ici nous proposons un graphique qui ne prend pas en compte la variable W juste pour des propos illustratifs. Le graphique propose trois lignes de régression en correspondance aux trois modalités de la variable Z.



Nous verrons par la suite que l'interprétation de ces coefficients sera plus saillante lorsque les données seront issues d'une expérience, tandis que dans cette partie nous nous référons de manière générale aux modèles linéaires indépendamment de leur application concrète. Avec des exemples dans lesquels les variables assument des connotations plus concrètes (e.g. heures d'études, attribution à des interfaces différentes d'un logiciel, etc.), aussi l'interprétation des coefficients en résultera facilitée.

1.2.3 Modèle linéaire généralisé

Le modèle linéaire généralisé est plus flexible du modèle linéaire de la régression multiple car il arrive à mieux accomoder des variables outcomes Y de différentes types. En effet, grâce à ce modèle plus flexible, on peut notamment dépasser certaines limites sur la mesure d'intérêt notamment sur les aspects suivants.

Utiliser des mesures avec des bornes. La régression linéaire multiple présuppose que, au moins potentiellement, la variable outcome peut assumer toute valeur de $-\infty$ à ∞. Cependant, on peut s'intéresser à des mesures qui ont des limites inférieure et/ou supérieure, par exemple lorsqu'on s'intéresse à un évènement qui s'avère ou pas (variable binaire). Dans ce cas, le modèle linéaire généralisé permet d'effectuer une régression linéaire dite logistique.

Utiliser des mesures discrètes. Encore une fois au moins sur la carte, la régression linéaire multiple s'attend à ce que la variable d'outcome dispose d'une fonction de densité de probabilité, c'est-à-dire que ses valeurs s'étalent sur un continuum. Ceci n'est notamment pas le cas dans des mesures de type décompte, comme le nombre d'erreurs dans une dictée ou le nombre de phonèmes émis par un-e élève pendant une scéance. Le modèle linéaire généralisée permet de modéliser la variable outcome par exemple avec une distribution de

Poisson, utilisée souvent avec des décomptes.

Utiliser des mesures ordinales. Un cas particulier des mesures discrètes consiste dans les échelles ordinales, par exemple les échelles de Lickert qui représente souvent le degré d'accord sur une échelle de 1 à 5 (ou de 1 à 7) dans laquelle une extrémité représente *pas d'accord du tout* et l'autre *tout à fait d'accord*. Ce cas, qui sera abordé dans l'exemple plus bas, peut s'appuyer dans le modèle linéaire généralisé sur une régression ordinale.

Cette flexibilité est rendue possible par le fait que le modèle linéaire généralisé utilise une fonction de transformation (une fonction de *link* en jargon statistique) qui permet de faire passer la variable d'outcome dans un format continu et potentiellement infini. De manière formelle, le modèle s'exprime de la manière suivante:

$$g(Y) = X\beta$$

La fonction $g()$ est celle qui s'occupe de transformer le lien entre les variables prédictives, représentée ici par un *set* X qui correspond à $\{X_0, X_1, X_2, ...X_n\}$, et la variable outcome Y dans un format propice à la modélisation linéaire, quand cette relation, sans la fonction $g()$ ne serait pas linéaire. Cette équation permet également d'identifier comme la régression linéaire multiple est en effet un cas particulier du modèle linéaire généralisé qui s'avère lorsque $g(Y) = Y$, c'est-à-dire quand il n'y a pas de transformation de la variable outcome Y.

Cette flexibilité garantie par le modèle linéaire généralisé comporte néanmoins certaines conséquences qu'il faut considérer pour pouvoir utiliser ce modèle est interpréter correctement les résultats qu'on peut obtenir. Dans le reste de cette partie, nous proposons à cet effet un exemple avec la variable outcome Y de type catégorielle ordinale, plus spécifiquement une échelle de Lickert de 1 à 5. Dans l'équation générique suivante, la variable outcome Y est en lien avec deux variables prédictives X (catégorielle avec deux modalités A et B) et W qui est continue.

$$\begin{aligned}
\log\left[\frac{P(1 \geq 2)}{1 - P(1 \geq 2)}\right] &= \alpha_1 + \beta_1(\mathrm{x_B}) + \beta_2(\mathrm{w}) \\
\log\left[\frac{P(2 \geq 3)}{1 - P(2 \geq 3)}\right] &= \alpha_2 + \beta_1(\mathrm{x_B}) + \beta_2(\mathrm{w}) \\
\log\left[\frac{P(3 \geq 4)}{1 - P(3 \geq 4)}\right] &= \alpha_3 + \beta_1(\mathrm{x_B}) + \beta_2(\mathrm{w}) \\
\log\left[\frac{P(4 \geq 5)}{1 - P(4 \geq 5)}\right] &= \alpha_4 + \beta_1(\mathrm{x_B}) + \beta_2(\mathrm{w})
\end{aligned} \tag{3}$$

Le tableau suivant affiche les premières 10 observations d'un jeu de données simulé. Le nombre d'observations totales est de $N = 20$.

Table 8: Aperçu de 10 observations dans un jeu de données avec la variable outcome Y en forme d'échelle de Lickert (de 1 à 5)

i	x	w	outcome
1	A	15.173	1
2	A	11.345	1
3	B	29.442	4
4	B	11.015	2
5	A	15.105	2
6	A	15.336	1
7	A	24.628	5
8	A	21.348	3
9	B	23.016	5
10	B	20.982	1

En appliquant une régression logistique ordinale aux données on obtient les coefficients de la régression suivants.

Table 9: Résultats de la régression ordinale avec coefficients pour les variables prédictives et les interceptes pour les réponses possibles

	Estimate	Std..Error	z.value	p.value
xB	1.962	0.937	2.093	0.036
w	0.198	0.093	2.123	0.034
1|2	2.446	1.603	1.526	0.127
2|3	3.713	1.669	2.224	0.026
3|4	5.020	1.835	2.735	0.006
4|5	6.597	2.118	3.115	0.002

Comparé au tableau de la régression linéaire multiple, on s'aperçoit surtout de la présence des 4 *seuils de coupure* entre les 5 options de l'échelle de Lickert. Ces valeurs n'ont pas une application directe dans l'interprétation des coefficients, mais sont importantes pour déterminer l'ensemble du modèle, ainsi que pour définir la probabilité que chaque catégorie de l'échelle soit choisi. En effet, il y a une forme d'effet d'accumulation entre les catégories

ordinales, car la catégorie précédente est de quelque sorte *contenue* dans la suivante. Pour cette raison, le modèle calcule les *seuils* pour définir à quel moment c'est plus probable qu'une observation *passe à la vitesse supérieure*.

Au contraire, les deux variables prédictive obtiennent un coefficient tout à fait similaire à la régression linéaire multiple. En ayant X seulement deux modalités, le coefficient concerne la modalité B, tandis que la modalité A est prise comme ligne de base. Il faut cependant faire attention à interpréter ces coefficients exactement de la même manière que dans le cadre de la régression linéaire multiple. En effet, dans le modèle linéaire généralisé la fonction de transformation de la variable outcome X est valable sur l'ensemble des variables prédictive. Par conséquent, il faut prendre en compte l'ensemble du modèle, car les coefficients individuellement ne produise pas la même information que dans la régression linéaire multiple. L'équation avec les coefficients qui résulte de la modélisation des données prend donc en compte à la fois les coefficients des variables prédictives et les 4 seuils entre les catégories ordinales de l'échelle de Lickert.

$$\begin{aligned}
\log\left[\frac{P(1 \geq 2)}{1 - P(1 \geq 2)}\right] &= 2.45 + 1.96(\mathrm{x_B}) + 0.2(\mathrm{w}) \\
\log\left[\frac{P(2 \geq 3)}{1 - P(2 \geq 3)}\right] &= 3.71 + 1.96(\mathrm{x_B}) + 0.2(\mathrm{w}) \\
\log\left[\frac{P(3 \geq 4)}{1 - P(3 \geq 4)}\right] &= 5.02 + 1.96(\mathrm{x_B}) + 0.2(\mathrm{w}) \\
\log\left[\frac{P(4 \geq 5)}{1 - P(4 \geq 5)}\right] &= 6.6 + 1.96(\mathrm{x_B}) + 0.2(\mathrm{w})
\end{aligned} \tag{4}$$

La valeur des coefficients également plus difficile à interpréter, car il ne sont pas en unité de la variable outcome Y. Dans le cas spécifique de la régression logistique ordinale, les coefficients sont exprimées en *log-odds.* En français on les appellent plutôt "logarithme des cotes" ou "logarithme des rapports de probabilité".Il s'agit d'une transformation mathématique qui sert pour définir la probabilité d'un évènement par rapport à son contraire. Un *log-odds* de 0 correspond à une probabilité équitable entre l'évènement et son contraire. Une valeur de 1 donne plus de chances à l'évènement, tandis qu'une valeur de -1 donne plus de chance à son contraire. On peut donc transformer les *log-odds* dans des probabilité qui sont plus simple à cerner, comme c'est le cas dans le graphique suivant.

Dans cette représentation graphique on peut noter plusieurs éléments d'intérêts au niveau de l'inférence depuis les données:

- Pour chacune des valeurs de l'échelle de Lickert on peut voir la probabilité qu'elle ait été choisi en fonction des deux modalités de X. De plus, la variable W, étant continue,

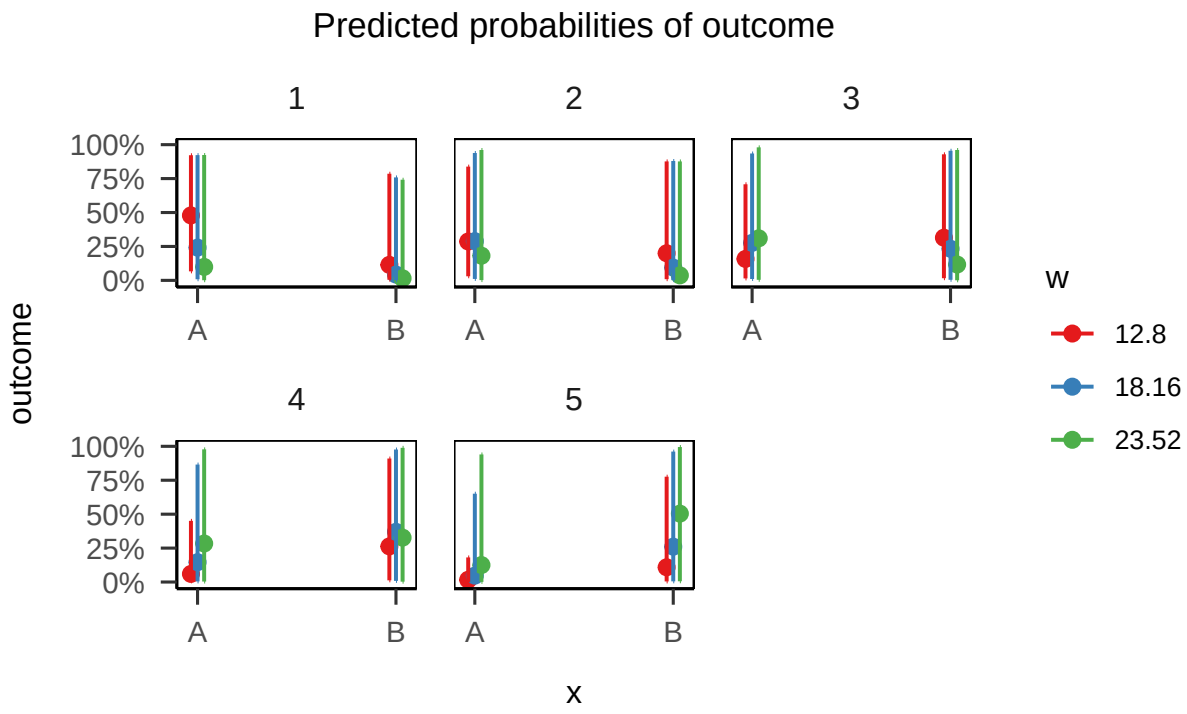


Figure 3: Représentation graphique d'une regréssion ordinale dans laquel les log-odds de la variable outcome ont été rétransformées en probabilité de choisir une réponse entre 1 et 5 en fonction des variables prédictives X et W

est stratifiée en trois valeurs, une faible (rouge), une moyenne (bleu) et un élevée (verte). Ces trois valeurs sont ensuite disposées pour chaque modalité de X.

- D'après la répartition des catégories en fonction de X, on peut noter comme pour les valeurs 1 et 2 de l'échelle de Lickert, en général il y a une probabilité plus élevé d'avoir ces valeurs avec la modalité A. Cela signifie que dans cette modalité, les personnes ont choisi plus souvent ces deux valeurs faibles. En ce qui concerne la valeur neutrale 3, les proportions sont plus ou moins les mêmes dans les deux conditions. Enfin, pour les valeurs 4 ou 5, la probabilité est plus élevée pour la modalité B. Depuis cette représentation on peut corroborer pourquoi le coefficient X_B dans le tableau de la régression est positif.

- Pour chaque modalité de X, en même temps, il y a trois probabilités pour les valeurs faible, moyenne et élevé de W. Dans ce cas également on peut voir comme la valeur faible de W (rouge) a des probabilités plus élevée dans les évaluations de 1 et 2 sur l'échelle de Lickert, tandis que ces probabilités sont plus élevée sur les valeurs de 4 et 5 pour la valeur élevée de W (verte). Cette répartition corrobore donc la valeur positive du coefficient de W.

Depuis ces informations on peut donc inférer que l'évaluation des personnes dans la modalité B de X a été plus positive que pour la modalité A, et que cette évaluation a été d'autant plus positive si la valeur sur la variable W était élevée. Dans un exemple concret, on peut imaginer que X correspond à deux versions d'une interface de logiciel et que W correspond au temps en minutes passées par les personnes à utiliser l'interface. La variable Y pourrait être l'évaluation utilisabilité sur une échelle UX avec un seul item. À ce moment, on peut inférer que l'interface B est jugée meilleure, surtout lorsque les participantes l'ont explorée pour plus longtemps.

Une fois que les grandes lignes du modèle linéaire générale ont été abordées, il est utile de faire une comparaison avec le même modèle qui utilise une régression linéaire multiple. Il arrive souvent dans les contributions *quantitatives* en sciences sociales de voir que des mesures de type Lickert sont analysées avec des cas spéciaux de la régression linéaire multiple (t-test, ANOVA, …). À cet effet, avec quelques passages techniques épargnés ici dans l'explication mais disponible dans le code source du document, il a été possible de récupérer pour des finalités purement pédagogiques la différence entre les évaluations des modalités A et B de X sur la même échelle de Lickert. La différence obtenue est de 1.22 points. Deux personnes qui ont passé le même temps sur l'interface, donc, diffèrent de plus d'un point d'évaluation sur l'échelle de Lickert. À titre de comparaison, le tableau suivant montre les résultats du tableau de régression linaire multiple avec les mêmes données utilisées plus haut.

Table 10: Tableau des paramètres d'une régression linéaire multiple lorsque la variable outcome ordinale est considéré comme continue

Predictor	*b*	95% CI	*t*	*df*	*p*
Intercept	0.31	[-1.68, 2.30]	0.33	17	.744
XB	1.11	[-0.02, 2.24]	2.07	17	.054
W	0.12	[0.01, 0.23]	2.29	17	.035

On peut noter comme le coefficient appliqué à la variable X, qui est dans ce cas dans la même métrique de la variable outcome Y, est de 1.11. Si cette différence d'environ 0.1 peut faire la différence dépend en réalité des objectifs de la recherche et du degré de précision souhaité par les chercheurs. Si vous avez néanmoins l'habitude fréquentiste de regarder la p-valeur associée aux coefficients, vous aurez probablement remarqué que dans cette analyse la p-valeur > 0.05. Au contraire, dans la régression logistique ordinale, le même coefficient apparaît à p-valeur < 0.05. Il s'agit en toute vérité d'une coincidence issue de la simulation, mais il se peut en effet qu'une application d'un modèle plus appropriée pour le type de données puisse proposer des avantages dans la précision des estimations. Bien évidemment,

ceci s'applique comme principe général est non pas exclusivement si on *flirte* avec le seuil conventionnel de $p < 0.05$.

1.2.4 Modèle linéaire généralisée mixte

1.3 La modélisation d'outcomes potentielles

1.4 L'inférence statistique

2 Statistiques *fréquentistes*

3 Statistiques Bayésiennes

4 Conclusion

Références

Bakker, M., & Wicherts, J. M. (2011). The (Mis)Reporting of Statistical Results in Psychology Journals. *Behavior Research Methods*, *43*(3), 666-678. https://doi.org/10.3758/s13428-011-0089-5

Bareinboim, E., & Pearl, J. (2016). Causal Inference and the Data-Fusion Problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345-7352. https://doi.org/10.1073/pnas.1510507113

Cinelli, C., Forney, A., & Pearl, J. (2020). A Crash Course in Good and Bad Controls. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3689437

Cobb, G. W. (2007). The Introductory Statistics Course: A Ptolemaic Curriculum. *Technology Innovations in Statistics Education*, *1*(1), 1-16.

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and Other Stories*.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations. *European Journal of Epidemiology*, *31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science*, 1745691620958012. https://doi.org/10.1177/1745691620958012

Maurer, K., Hudiburgh, L., Werwinski, L., & Bailer, J. (2019). Content Audit for P-Value Principles in Introductory Statistics. *The American Statistician*, *73*(sup1), 385-391. https://doi.org/10.1080/00031305.2018.1537890

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: a model comparison perspective* (Third edition). Routledge.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2 éd.). Taylor; Francis, CRC Press.

Nickerson, R. S. (2000). Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. *Psychological Methods*, *5*(2), 241-301. https://doi.org/10.1037/1082-989X.5.2.241

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The Prevalence of Statistical Reporting Errors in Psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Pearl, J. (2000). *Causality: models, reasoning, and inference*. Cambridge University Press.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic

Books.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, *65*(1), 112. https://doi.org/10.1037/a0018326

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., Gronau, Q. F., Kalish, M. L., McMullin, S. D., Navarro, D. J., & Shiffrin, R. M. (2023). Statistics in the Service of Science: Don't Let the Tail Wag the Dog. *Computational Brain & Behavior*, *6*(1), 64-83. https://doi.org/10.1007/s42113-022-00129-2