

Chapitre 4: Introduction à l'Économétrie

Christelle Dumas

Contents

1	La régression linéaire simple: modèle et estimation ponctuelle	3
1.1	Un exemple: le lien entre éducation et salaire	3
1.2	Le modèle linéaire simple	5
1.3	L'estimation du modèle linéaire simple par les moindres carrés ordinaires (MCO)	7
1.4	Résidus et mesure de dispersion	10
2	Le modèle linéaire simple: intervalles de confiance, tests sur les paramètres et analyse de la variance	11
2.1	Propriétés des estimateurs	11
2.2	Distribution d'échantillonnage de b_1	12
2.3	Estimation de β_1 par intervalle de confiance	13
2.4	Test statistique sur β_1	13
2.5	Inférence concernant le paramètre b_0	14
2.6	Inférence sur $E(Y_h)$, moyenne de la distribution conditionnelle de Y à $X = X_h$	15
2.7	Prévision d'une valeur de la variable dépendante pour une nouvelle observation de X et intervalle de prévision	16
2.8	Équation d'analyse de la variance et coefficient de détermination	17

3	Le modèle linéaire simple: écriture matricielle	19
3.1	Reposer le problème	19
3.2	Vers une représentation géométrique	21
3.3	Calcul matriciel de l'estimateur des MCO	22
3.4	Les pièges des méthodes d'analyse de régression	23

Introduction

On a vu jusqu'ici comment estimer et faire des tests d'hypothèses sur des paramètres inconnus tels que la proportion, la moyenne, la variance. On a aussi vu que l'on pouvait appliquer de telles méthodes pour analyser la force d'une corrélation entre deux variables. Ce à quoi nous allons nous intéresser maintenant est la mesure de l'effet d'une variable sur une autre et ce champ statistique s'appelle l'économétrie. On ne connaît en général pas le paramètre qui mesure l'effet d'une variable sur une autre pour l'ensemble de la population. L'objet de l'économétrie est donc de proposer une méthode pour estimer ce paramètre à l'aide d'un échantillon pour lequel les deux variables sont renseignées, et par ailleurs de donner la loi de cet estimateur afin de pouvoir effectuer des tests sur ce paramètre.

Il y a diverses raisons pour lesquelles on peut vouloir estimer l'effet d'une variable sur une autre. C'est notamment le cas si on veut évaluer l'effet d'un changement de politique (par exemple, quelles sont les implications d'une hausse du prix du tabac sur la consommation de cigarettes). D'autre part, si l'on connaît le lien entre deux variables sur un échantillon, on peut inférer cette relation sur l'ensemble de la population. Imaginons par exemple un cas où je connais l'éducation de l'ensemble des individus de ma population, mais je ne connais le salaire de ces individus que pour un échantillon d'entre eux. Je peux estimer le lien entre éducation et salaire sur mon échantillon et s'il est représentatif de la population, je peux prédire le salaire de chaque individu de ma population sur la base de son éducation. Ceci pourrait avoir un intérêt pour l'état s'il cherche à savoir quelles vont être ses rentrées budgétaires sur la base de l'impôt sur les revenus. Pour résumer, l'analyse économétrique est intéressante pour décrire les relations entre divers phénomènes, évaluer des politiques publiques et prédire des variables.

Un autre intérêt fondamental de l'économétrie est de sortir du cadre où l'on étudie uniquement le lien de deux variables; nous verrons notamment que l'on peut vouloir modéliser une variable comme dépendante de plus d'une caractéristique. Si je cherche à "expliquer" ou prédire le prix d'un logement, il va falloir non seulement que je prenne en compte sa superficie, mais aussi sa localisation. De la même façon, si je cherche à prédire la croissance pour l'année prochaine, un grand nombre de déterminants devront être pris en compte (taux de chômage, compétitivité de l'euro, salaire minimum, progrès technique, etc.). Nous verrons dans la suite du cours pourquoi ceci est important dans l'estimation au-delà du simple fait d'estimer de façon plus précise la variable d'intérêt.

Pour bien comprendre comment se situe l'économétrie dans l'inférence statistique, notez que ce que nous cherchions à faire jusqu'ici était de décrire la distribution de variables (à travers l'estimation de moyenne, variance et proportion), la plupart du temps univariée, parfois bivariée lorsque l'on s'intéresse à la corrélation. L'économétrie, elle, a pour objet d'expliquer un mécanisme. Pour cela, nous devons faire appel à des modèles. Qui dit modèle dit hypothèses: nous introduisons donc des contraintes sur la façon dont nous souhaitons lire nos données; mais l'avantage est que ceci permet d'aboutir à des conclusions plus fortes et plus intéressantes que ce que nous pouvions faire en l'absence de ces modèles. *In fine*, la pertinence et l'adéquation de ces modèles aux données peut être dans une certaine mesure évaluée.

1 La régression linéaire simple: modèle et estimation ponctuelle

1.1 Un exemple: le lien entre éducation et salaire

Prenons un étudiant qui vient d'avoir son bac et qui se demande si cela vaut la peine de continuer à l'université ou pas. Il a donc le choix d'aller travailler avec son diplôme du bac en poche ou de continuer ses études, auquel cas il doit payer une faible somme pour s'inscrire en L1, L2, L3, etc. Pour déterminer s'il a intérêt à poursuivre ses études, il doit connaître le salaire qu'il gagnerait avec son diplôme de baccalauréat et le salaire qu'il gagnerait s'il avait un diplôme de licence, maîtrise, ou plus en poche. Dans la mesure où il ne peut

pas essayer ces différentes situations, il va se baser sur les taux de salaires qui prévalent pour d'autres étudiants, qui lui sont comparables, et qui ont atteint ces différents niveaux. Il compile donc un échantillon d'observations pour lesquelles il dispose du nombre d'années d'études post-bac et du premier salaire obtenu par ces individus. Les résultats sont compilés dans le tableau suivant:

Table 1: Nombre d'années d'études universitaires et salaires mensuels

Années d'étude	Salaire	Années d'étude	Salaire
0	1250	4	1450
2	1290	5	1490
4	1420	0	1200
6	1530	6	1550
8	1580	8	1600
1	1245	3	1400
3	1360	2	1310
5	1440	1	1230
1	1290	4	1500
2	1340	1	1350
3	1340		

La représentation du nuage de points indique une relation croissante entre le nombre d'années d'éducation et le salaire. Une liaison linéaire entre le nombre d'années d'éducation et le salaire semble plausible. Ce qui va nous intéresser est la pente de cette relation positive, à savoir comment une année d'éducation se transmet en différentiel de salaire.

Variable dépendante et variable explicative Cette recherche de la relation linéaire entre deux variables va nous permettre d'obtenir un outil de prévision: on pourra estimer, à l'aide de cette équation, les valeurs d'une variable à partir des valeurs prises par l'autre variable. Cependant, il faut d'abord convenir de la variable que nous allons exprimer en fonction de l'autre. Ce choix est important et permettra d'identifier la variable "dépendante" ou "expliquée", que nous notons Y et la variable "indépendante" ou "explicative" que nous notons X . Sur le plan purement théorique, on peut bien

souvent établir une droite de régression de Y par rapport à X ou de X par rapport à Y , mais la plupart du temps, il y a un sens dicté par la théorie économique ou le simple bon sens. Dans notre cas, il est logique de choisir le salaire comme variable dépendante et le nombre d'années d'éducation comme variable indépendante ou explicative. L'inverse n'aurait simplement pas de sens: quelle serait la signification du paramètre de la pente? "la façon dont se transmet un salaire en nombre d'années d'éducation"? Par conséquent, le choix d'expliquer les fluctuations d'une variable par une autre doit être dicté par l'aspect pratique, physique ou économique du phénomène étudié.

1.2 Le modèle linéaire simple

Lorsque l'on cherche à établir le lien entre deux variables, deux questions fondamentales se posent:

- quel modèle statistique semble le plus approprié pour décrire la forme de relation entre les variables concernées? devrions-nous utiliser une forme linéaire, parabolique, exponentielle?
- en admettant un modèle particulier comme plausible, comment peut-on, avec les données de l'étude, calculer les estimations des paramètres du modèle avec le plus de justesse possible?

Le premier modèle que nous allons traiter est le modèle linéaire simple, qui s'énonce comme suit:

Étant donné n couples d'observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ et en supposant que la relation entre Y et X est linéaire alors le modèle s'écrit:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n$$

où:

- Y est la variable dépendante (ou expliquée) ayant un caractère aléatoire dont les valeurs sont conditionnées par celles de la variable explicative X et la composante aléatoire ϵ ; Y_i représente la i -ième observation de Y .
- β_0 et β_1 sont les paramètres du modèle de régression

- X est la variable explicative; on la considère comme une valeur certaine
- ϵ dénote la fluctuation aléatoire non observable attribuable à un ensemble de facteurs ou de variables non pris en considération dans le modèle. Cette fluctuation aléatoire n'est pas expliquée par le modèle et se reflète sur la variable dépendante.

Pourquoi appeler ce modèle “linéaire simple”? Le terme “simple” indique la présence d’une seule variable explicative. Le terme “linéaire” ne se réfère qu’aux paramètres du modèle. Ainsi, le modèle $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$ est aussi dit linéaire car il est une combinaison linéaire des paramètres. De même, si le modèle estimé était $Y = \beta_0 + \beta_1 \ln(X) + \epsilon$, on dirait aussi qu’il est linéaire (i.e. il ne s’agit pas de linéarité en X). Un exemple de modèle non linéaire pourrait être: $Y = \beta_0 + X^{\beta_1} + \epsilon$. Nous ne verrons dans ce cours que des modèles linéaires.

Composantes du modèle linéaire simple On a vu que la seule composante aléatoire du modèle provient du terme ϵ . On peut donc décomposer le modèle comme étant la somme de deux composantes:

- une composante non aléatoire $\beta_0 + \beta_1 X_i$ attribuable aux valeurs certaines prises par la variable explicative X : pour chaque valeur X_i , la résultante $\beta_0 + \beta_1 X_i$ est une valeur fixe;
- une composante aléatoire ϵ_i , qui tient compte du caractère aléatoire de Y . En effet, on a vu que la relation entre X et Y n’est pas parfaite (les points ne sont pas parfaitement alignés, la corrélation entre les deux est différente de 1); on explique la différence entre la composante non aléatoire $\beta_0 + \beta_1 X_i$ et Y par un terme aléatoire, non lié à X . La composante ϵ_i représente la fluctuation aléatoire des valeurs individuelles Y_i autour de la valeur centrale $\beta_0 + \beta_1 X_i$ et Y et cette fluctuation aléatoire peut provenir de déterminants inobservés.

Par conséquent, pour chaque valeur particulière prise par la variable indépendante X , la variable dépendante Y est caractérisée par une certaine distribution de probabilité.

Hypothèses du modèle linéaire simple Dans la mesure où l'on n'observe pas les ϵ , il faut ajouter des conditions restrictives pour définir le rôle qu'ils vont jouer. En effet, comme on ne connaît ni β_0 ni β_1 , tant qu'on n'ajoute pas des conditions sur les ϵ , il existe une infinité de solutions au modèle (pas seulement pour estimer les β mais même pour les définir). Par exemple, on pourrait fixer $\beta_0 = \beta_1 = 0$ et $\epsilon = Y$. Il est donc nécessaire d'ajouter des hypothèses sur les ϵ pour préciser la signification du modèle. D'une certaine façon, cela revient à préciser quelle droite de régression est la plus appropriée pour représenter le nuage de points.

Hypothèses 1 *On suppose que les ϵ_i sont des variables aléatoires normales et indépendantes de moyenne $E(\epsilon) = 0$ et de variance identique $V(\epsilon_i) = \sigma^2$ pour toutes les valeurs de X .*

Ceci a pour implication que pour toute valeur particulière X_i , la variable dépendante Y est une v.a. distribuée d'après une loi normale de moyenne $E(Y_i) = \beta_0 + \beta_1 X_i$ et de variance $V(Y_i) = \sigma^2$. Inclure graphique.

1.3 L'estimation du modèle linéaire simple par les moindres carrés ordinaires (MCO)

Qu'est-ce que les MCO? Maintenant que nous avons complètement défini le modèle linéaire simple (et nous avons vu qu'il était autant défini par l'équation de régression que par les hypothèses), il convient de se demander comment on va pouvoir estimer les β à partir d'un échantillon. L'idée générale est que l'on veut que le terme aléatoire (ϵ) soit le plus petit possible. On veut obtenir l'estimation de la droite de régression $E(Y_i) = \beta_0 + \beta_1 X_i$. Notons cette estimation $\hat{Y} = b_0 + b_1 X_i$ (droite de régression empirique) où b_0 est un estimateur de β_0 sur la base de l'échantillon et b_1 un estimateur de β_1 . b_0 représente alors l'ordonnée à l'origine et b_1 la pente de la droite de régression. Voir graphique.

Une première solution serait de minimiser la somme des écarts entre Y et la valeur prédite $\hat{Y} = b_0 + b_1 X_i$. Cependant, des écarts négatifs compenseraient des écarts positifs et avoir *in fine* d'assez grands écarts entre Y et sa valeur prédite. Une deuxième option consisterait à minimiser la somme des valeurs absolues de ces écarts pour pallier ce problème; c'est effectivement une bonne

solution qui donne lieu à une méthode d'estimation utilisée mais que nous n'allons pas retenir car la fonction valeur absolue n'est pas dérivable en 0, ce qui complique la résolution.

La troisième option, qui est celle que nous allons retenir est celle de la méthode des moindres carrés ordinaires (MCO ou OLS en anglais pour Ordinary Least Square); elle consiste à minimiser la somme des écarts au carré. Elle revient donc à chercher b_0 et b_1 tels qu'ils minimisent l'expression:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

L'écart e_i est appelé le résidu de la i -ième observation et la somme de carrés $\sum e_i^2$ la somme des carrés résiduelle ou variation résiduelle. Elle nous permettra d'obtenir une mesure de l'ampleur de l'éparpillement des observations Y_i autour de la droite de régression. Plus les points seront serrés autour de la droite de régression empirique, plus la valeur de $\sum e_i^2$ sera faible.

Estimation par MCO Pour minimiser l'expression

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

par rapport à b_0 et b_1 , on a recours aux dérivées partielles. La minimisation impose que les dérivées premières soient nulles et que les dérivées secondes soient positives.

$$\frac{\partial (\sum e_i^2)}{\partial b_0} = -2 \sum (Y_i - b_0 - b_1 X_i)$$

$$\frac{\partial (\sum e_i^2)}{\partial b_1} = -2 \sum (Y_i - b_0 - b_1 X_i) X_i$$

En simplifiant, on obtient 2 conditions à satisfaire:

$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^n X_i (Y_i - b_0 - b_1 X_i) = 0$$

En réarrangeant ces deux expressions en fonction des deux inconnues b_0 et b_1 , on obtient deux équations dites équations normales:

$$\begin{aligned} nb_0 + b_1 \sum X_i &= \sum Y_i \\ b_0 \sum X_i + b_1 \sum X_i^2 &= \sum X_i Y_i \end{aligned}$$

La résolution de ces deux équations fournit les expressions algébriques pour les estimateurs ponctuels b_0 et b_1 :

$$\begin{aligned} b_1 &= \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{n \sum X_i^2 - (\sum X_i)^2} \\ b_0 &= \frac{(\sum X_i^2)(\sum Y_i) - (\sum X_i)(\sum X_i Y_i)}{n \sum X_i^2 - (\sum X_i)^2} \end{aligned}$$

Le premier terme indique la pente de la droite tandis que le second indique l'ordonnée à l'origine.

Les dérivées secondes $\frac{\partial^2(\sum e_i^2)}{\partial b_0^2} = 2n > 0$ et $\frac{\partial^2(\sum e_i^2)}{\partial b_1^2} = 2 \sum X_i^2 > 0$ assurent que les expressions précédentes pour b_0 et b_1 minimisent la somme des carrés résiduelle.

Autre expression pour b_0 et b_1 On montre que b_1 peut se réécrire de la façon suivante:

$$b_1 = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

et que b_0 se déduit facilement de la première équation normale, une fois b_1 calculé:

$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Application

- Estimation par MCO des coefficients de l'équation:

$$\text{Salaire} = \beta_0 + \beta_1 \text{Education} + \epsilon$$

- tracé de la droite de régression:

$$\widehat{\text{Salaire}} = b_0 + b_1 \text{Education}.$$

- Interprétation des coefficients.

Remarques

- Seule la droite des moindres carrés (ou droite de régression) assure que $\sum(Y_i - \hat{Y}_i)^2$ est minimale. Cette droite est unique pour l'échantillon observé.
- La droite de régression passe toujours par le point (\bar{X}, \bar{Y}) puisque pour $X_i = \bar{X}$,

$$\hat{Y}_i = \bar{Y} + b_1(X_i - \bar{X}) = \bar{Y}.$$

1.4 Résidus et mesure de dispersion

Un des intérêts majeurs des MCO est de fournir une estimation des paramètres du modèle β_0 et β_1 mais ceci ne nous indique finalement pas si le modèle retenu est satisfaisant ou non. En effet, il est toujours possible d'estimer ces coefficients, même si finalement on n'observe qu'une relation ténue entre les variables Y et X . Pour juger de cela, on peut regarder à quel point la prédiction est proche de la valeur observée. En effet, si l'on prédit une variable \hat{Y} sur la base de la variable X et que celle-ci n'est que peu reliée à la variable Y alors la distance entre la variable prédite (\hat{Y}) et la valeur observée (Y) sera importante. Si au contraire ma prédiction est systématiquement (i.e. pour l'ensemble des observations) très proche de la valeur observée alors on pourra considérer que le modèle est satisfaisant.

Nous avons déjà introduit cette différence entre \hat{Y} et Y via les résidus. En effet,

$$e_i = Y_i - \hat{Y}_i$$

est appelé le i -ième résidu ou résidu pour la i -ième observation. On sait que les MCO minimisent la somme des carrés des résidus, mais il faut regarder *in fine* si cette somme de carrés de résidus est faible ou forte. Comme déjà précisé, la variance des Y_i autour de la droite de régression est la même que la variance des résidus (σ^2) et une estimation de celle-ci s'obtient à partir des résidus calculés. En effet, l'estimation de σ^2 s'obtient de la variance résiduelle que nous notons s^2 et qui consiste à diviser la somme des carrés résiduelle ($\sum(Y_i - \hat{Y}_i)^2$) par $(n - 2)$, le nombre de degrés de liberté restants (une fois estimés β_0 et β_1):

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n - 2} = \frac{\sum(Y_i - b_0 - b_1 X_i)^2}{n - 2}$$

Application. L'écart-type des résidus, noté σ , donne une estimation de la dispersion des Y autour de la droite de régression et a les mêmes unités que la variable dépendante Y . La somme des résidus, $\sum e_i$ doit être nulle.

2 Le modèle linéaire simple: intervalles de confiance, tests sur les paramètres et analyse de la variance

2.1 Propriétés des estimateurs

Les estimateurs b_0 et b_1 de β_0 et β_1 respectivement ont des propriétés importantes.

Proposition 2 *Les estimateurs b_0 et b_1 de β_0 et β_1 respectivement sont sans biais et efficaces, ainsi:*

- $E(b_0) = \beta_0$ et $E(b_1) = \beta_1$;
- *parmi tous les estimateurs sans biais de β_0 et β_1 , b_0 et b_1 ont la plus petite variance.*

Ces deux estimateurs sont donc les meilleurs estimateurs de b_0 et b_1 .

Démonstration:

$$E(b_1) = E \left[\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \right]$$

or $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ d'où $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{\epsilon}$. Par conséquent, les X_i étant certains, on obtient:

$$\begin{aligned} E(b_1) &= \frac{1}{\sum (X_i - \bar{X})^2} \cdot \sum (X_i - \bar{X}) E(\beta_0 + \beta_1 X_i + \epsilon_i - (\beta_0 + \beta_1 \bar{X} + \bar{\epsilon})) \\ &= \frac{1}{\sum (X_i - \bar{X})^2} \cdot \sum (X_i - \bar{X}) E(\beta_1 (X_i - \bar{X}) + (\epsilon_i - \bar{\epsilon})) \\ &= \frac{1}{\sum (X_i - \bar{X})^2} \cdot \sum (X_i - \bar{X}) [\beta_1 (X_i - \bar{X})] \end{aligned}$$

car $E(\epsilon_i) = E(\bar{\epsilon}) = 0$. Enfin,

$$\begin{aligned} E(b_1) &= \beta_1 \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} \\ &= \beta_1 \end{aligned}$$

Par ailleurs,

$$\begin{aligned} E(b_0) &= E(\bar{Y} - b_1 \bar{X}) \\ &= E(Y) - E(b_1) \bar{X} \\ &= \beta_0 + \beta_1 \bar{X} + E(\bar{\epsilon}) - \beta_1 \bar{X} \\ &= \beta_0 + 0 = \beta_0 \end{aligned}$$

2.2 Distribution d'échantillonnage de b_1

Pour établir un intervalle de confiance sur un paramètre de régression ou pour exécuter un test statistique sur l'un ou l'autre des paramètres β_0 ou β_1 , nous devons caractériser la distribution d'échantillonnage de son estimateur; il faut donc en connaître la forme, la moyenne (déjà fait) et la variance.

Proposition 3 *Sous l'hypothèse que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$, alors:*

$$b_1 \rightsquigarrow \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right).$$

Par conséquent, les fluctuations de l'écart-réduit $Z = \frac{b_1 - \beta_1}{\sigma / \sqrt{\sum (X_i - \bar{X})^2}}$ suivent une loi normale centrée réduite.

On a vu que si σ^2 était inconnu, on peut l'estimer en sommant les carrés des résidus estimés. Dans le cas d'un petit échantillon, remplacer le vrai paramètre de variance par son estimation est susceptible de changer la loi, comme on l'a déjà vu auparavant. En effet, si on note s^2 l'estimation de σ^2 , dont la formule suit:

$$s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2} = \frac{\sum (Y_i - b_0 - b_1 X_i)^2}{n - 2}$$

alors l'écart-réduit

$$T = \frac{b_1 - \beta_1}{s/\sqrt{\sum(X_i - \bar{X})^2}}$$

suit une loi de Student à $n - 2$ degrés de liberté (on perd 2 degrés de liberté du fait de l'estimation de β_0 et β_1 , b_0 et b_1 qui exhibent une relation linéaire avec les variables X et Y).

2.3 Estimation de β_1 par intervalle de confiance

Maintenant que nous connaissons la distribution de l'estimateur b_1 de β_1 , nous sommes en mesure de fournir une estimation par intervalle de confiance. Dans la mesure où la distribution de l'estimateur est une Student, on verra qu'il y a peu de différence avec l'estimation par intervalle de confiance d'une moyenne dans un environnement où la population est distribuée normalement, l'échantillon petit et la variance inconnue. Nous ne présentons ici que le cas avec petit échantillon. Si vous êtes dans le cas d'un grand échantillon, la loi de Student est approximativement une loi normale et les quantiles de cette dernière peuvent être utilisés.

Dans le cas d'un échantillon de petite taille, prélevé de populations Y_i distribués selon une loi normale de moyenne conditionnelle $E(Y_i) = \beta_0 + \beta_1 X_i$ et de variance σ^2 inconnue, alors l'intervalle de confiance de niveau $(1 - \alpha)$ s'écrit:

$$IC_\alpha = \left[b_1 - t_{\alpha/2; n-2} \cdot \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}}; b_1 + t_{\alpha/2; n-2} \cdot \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}} \right]$$

où s est donné par la formule ci-dessus.

Application.

2.4 Test statistique sur β_1

De la même façon, il nous est maintenant aisé de tester des hypothèses sur β_1 . Dans la grande majorité des cas, ce qui nous intéresse de tester est la significativité du coefficient, ce qui revient à tester s'il est nul ou non. Pourquoi est-ce si important? simplement parce que lorsqu'on trouve un effet

de 0.12 par exemple, il n'est pas clair de savoir si cet effet est suffisamment important pour être considéré comme non nul et donc positif ou si la valeur positive provient seulement de l'aléa de l'échantillonnage. Bien entendu on peut tester l'égalité de β_1 à toute autre valeur, c'est seulement moins courant. Prenons donc pour exemple le cas où l'hypothèse nulle est $H_0 : \beta_1 = 0$ et où l'hypothèse alternative est bilatérale $H_1 : \beta_1 \neq 0$.

Sous l'hypothèse nulle, la statistique T se réécrit:

$$T = \frac{b_1}{s/\sqrt{\sum(X_i - \bar{X})^2}}$$

et suit toujours une Student à $n - 2$ degrés de libertés. Par conséquent, la prise de décision suivra la règle suivante:

Rejet de H_0 si et seulement si $t > t_{n-2;\alpha/2}$ ou $t < -t_{n-2;\alpha/2}$

La règle de décision peut se réécrire en fonction de l'estimation b_1 comme suit:

Rejet de H_0 si et seulement si $b_1 > t_{n-2;\alpha/2} \cdot s/\sqrt{\sum(X_i - \bar{X})^2}$
ou $b_1 < -t_{n-2;\alpha/2} \cdot s/\sqrt{\sum(X_i - \bar{X})^2}$

Application.

Remarque: on peut également effectuer, selon les besoins de l'analyse, un test unilatéral sur β_1 .

2.5 Inférence concernant le paramètre b_0

Il est moins fréquent d'effectuer de l'inférence sur le paramètre b_0 . En effet, il arrive que le paramètre de constante n'ait pas beaucoup de sens économique; ce pourra notamment être le cas si X ne prend jamais de valeurs proches de 0. Dans notre cas, cependant, le paramètre b_0 représente le salaire que peut attendre un individu n'ayant pas poursuivi ses études post-bac. Dans la logique de déterminer ce qui est optimal en termes de temps d'études pour l'individu, ce paramètre est d'importance. Donnons donc quelques éléments pour être en mesure de l'estimer par intervalle ou de faire des tests dessus.

Proposition 4 Sous l'hypothèse que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$, alors:

$$b_0 \rightsquigarrow \mathcal{N} \left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right] \right).$$

Ainsi, si l'on pose $\sigma(b_0) = \sigma \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]^{1/2}$, les fluctuations de l'écart-réduit $Z = \frac{b_0 - \beta_0}{\sigma(b_0)}$ suivent une loi normale centrée réduite.

Si la taille de l'échantillon est petite et que la variance est inconnue, alors les fluctuations de l'écart-réduit $t = \frac{b_0 - \beta_0}{s(b_0)}$ sont celles de la loi de Student à $n - 2$ degrés de liberté avec $s(b_0) = s \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]^{1/2}$.

Application: calcul d'un intervalle de confiance et test $H_0 : b_0 = 0$

2.6 Inférence sur $E(Y_h)$, moyenne de la distribution conditionnelle de Y à $X = X_h$

Nous savons que dans le cas du modèle linéaire simple, la moyenne des Y_i , à $X = X_i$ est donnée par

$$E(Y_i | X = X_i) = \beta_0 + \beta_1 X_i.$$

Cette quantité peut être considérée comme la valeur moyenne des Y_i pour l'ensemble des unités de la population dont la valeur prise par la variable explicative est X_i .

L'estimation ponctuelle de l'espérance $E(Y_h | X = X_h)$, sur la base de l'échantillon, s'obtient de la droite de régression:

$$\hat{Y}_h = b_0 + b_1 X_h.$$

Il convient de décrire la distribution de cette quantité pour pouvoir l'estimer par intervalle.

Proposition 5 Sous l'hypothèse que $\epsilon_i \rightsquigarrow \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$, alors:

$$\hat{Y}_h = b_0 + b_1 X_h \rightsquigarrow \mathcal{N} \left(\beta_0 + \beta_1 X_h, \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \right).$$

Les fluctuations de l'écart-réduit $Z = \frac{\hat{Y}_h - E(Y_h)}{\sigma(\hat{Y}_h)}$ suivent donc une loi normale centrée réduite où $\sigma(\hat{Y}_h) = \sigma \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]^{1/2}$.

Si l'échantillon est petit (et que la variance σ^2 des erreurs est inconnue), alors l'écart-réduit $t = \frac{\hat{Y}_h - E(Y_h)}{s(\hat{Y}_h)}$ suit une Student à $n - 2$ degrés de liberté, où $s(\hat{Y}_h) = s \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]^{1/2}$.

Application au calcul d'un intervalle de confiance.

2.7 Prédiction d'une valeur de la variable dépendante pour une nouvelle observation de X et intervalle de prédiction

L'objectif d'une étude de régression est non seulement d'obtenir des estimations de la moyenne des Y_i ($E(Y_i)$) pour diverses valeurs X_i mais également de fournir des prévisions concernant les valeurs éventuelles de la variable dépendante Y pour de nouvelles observations. C'est notamment une des motivations que nous avons avancé en introduction : prédiction de la croissance, prédiction de la pauvreté sur la base de certains indicateurs.

La prédiction d'une valeur éventuelle de Y pour une nouvelle observation X_h est obtenue de la droite de régression empirique:

$$Y_{h(p)} = b_0 + b_1 X_h.$$

L'estimation de la prédiction est donc la même que l'estimation de l'espérance conditionnelle de Y sachant $X = X_h$. Cependant, un aspect diffère. En effet, l'erreur de prédiction peut provenir de deux sources:

- l'erreur d'estimation de la moyenne $E(Y_h|X = X_h)$, c'est-à-dire l'écart entre $E(Y_h)$ et \hat{Y}_h ;
- l'erreur présente dans toute valeur individuelle de Y , c'est-à-dire l'écart entre Y_h et $E(Y_h)$.

L'écart de prévision, pour une nouvelle observation X_h (dont la réalisation, inconnue est en fait Y_h), s'écrit

$$d_h = Y_h - \widehat{Y}_h.$$

Puisque cette nouvelle observation n'a pas servi à établir la droite de régression, Y_h et \widehat{Y}_h sont indépendants. Par conséquent, la variance de l'erreur prévisionnelle vaut:

$$V(d_h) = V(Y_h - \widehat{Y}_h) = V(Y_h) + V(\widehat{Y}_h) = \sigma^2 + \sigma^2(\widehat{Y}_h)$$

où σ^2 représente la variance de la distribution des valeurs individuelles de Y et $\sigma^2(\widehat{Y}_h)$ la variance des fluctuations d'échantillonnage de \widehat{Y}_h . Au final,

$$\sigma^2(d_h) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

qui peut être estimée par:

$$s^2(d_h) = s^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

Application: intervalle de prévision de Y à $X = X_h$. Si l'étudiant cherche à estimer quels pourraient être ses salaires à différents niveaux d'études, il doit bien prendre en compte le fait que 1) les estimateurs de la droite de régression peuvent ne pas être exacts 2) sa propre réalisation peut différer de la moyenne (ou espérance).

2.8 Équation d'analyse de la variance et coefficient de détermination

Analyser la variance de Y va nous permettre de juger du pouvoir explicatif du modèle statistique proposé (ici, le modèle linéaire simple). La variabilité de la variable Y peut être décomposée comme suit:

- une variation attribuable à la régression (puisque \widehat{Y} varie avec les valeurs prises par X) et
- une variation résiduelle

En effet, puisque $Y_i = \hat{Y}_i + \epsilon_i$, on peut montrer que

$$V_{emp}(Y) = V_{emp}(\hat{Y}) + V_{emp}(\epsilon).$$

Démonstration:

$$\begin{aligned} V_{emp}(Y) &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \frac{1}{n} \sum_{i=1}^n ((y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) \\ &= \frac{1}{n} \sum_{i=1}^n e_i^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n e_i(\hat{y}_i - \bar{y}) \end{aligned}$$

or

- le premier terme est égal à la variance empirique de e puisque la moyenne des résidus estimés est égale à 0 par construction;
- le second terme est égal à la variance empirique de \hat{y} puisque la moyenne des prédictions est égal à la moyenne des valeurs ($\bar{y} = \bar{\hat{y}}$);
- le dernier terme vaut 0 car le résidu estimé est orthogonal à la valeur prédite: $\sum e_i(\hat{y}_i - \bar{y}) = \sum e_i \hat{y}_i - \bar{y} \sum e_i$.

Par conséquent, la variance totale (de Y) se décompose en variance expliquée ($V_{emp}(\hat{Y})$) et variance résiduelle ou inexpliquée ($V_{emp}(\epsilon)$). Pour mémoire, rappelons que la procédure qui consistait à minimiser la somme des erreurs, consiste à minimiser la variance inexpliquée. Pour une variabilité donnée de la variable dépendante, on a donc cherché à maximiser la partie expliquée du nuage de points.

Définition 6 On appelle coefficient de détermination et on note R^2 la quantité:

$$R^2 = \frac{V_{emp}(\hat{y})}{V_{emp}(y)} = 1 - \frac{V_{emp}(\epsilon)}{V_{emp}(y)}.$$

Notons immédiatement quelques caractéristiques de cette statistique:

- par construction, elle est comprise entre 0 et 1: $0 \leq R^2 \leq 1$;
- plus R^2 est proche de 1, plus la part inexpliquée de la variance totale est petite (proche de 0); par conséquent, R^2 proche de 1 correspond à un bon ajustement du nuage de points par la droite.

Proposition 7

$$R^2 = \text{Corr}(X, Y)$$

On peut donc évaluer l'adéquation d'un modèle linéaire simple en calculant son coefficient de détermination; si le R^2 est élevé, cela signifie que la corrélation linéaire entre la variable dépendante et la variable indépendante est élevée, ce qui justifie l'utilisation d'un tel modèle.

Remarque: le terme constant est nécessaire dans l'estimation pour que l'analyse de la variance soit valide. Si vous omettez le terme β_0 dans le modèle, le R^2 calculé ne se situe plus nécessairement entre 0 et 1 et n'est plus interprétable.

3 Le modèle linéaire simple: écriture matricielle

Avant de passer au modèle linéaire multiple, il est utile de voir l'écriture matricielle du modèle linéaire simple car nous aurons besoin de cette représentation matricielle lorsque nous introduirons plus d'une variable explicative.

3.1 Reposer le problème

Notons y le vecteur qui rassemble l'ensemble des valeurs y_i pour l'ensemble des observations:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

De même, on peut noter $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$, c le vecteur constitué de 1, $c = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, et $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$. Ainsi, le modèle linéaire simple se réécrit:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\Leftrightarrow y = \beta_0 c + \beta_1 x + \epsilon$$

Par ailleurs, si l'on définit la matrice X (de taille $(n, 2)$) comme égale à $(c \ x)$ et le paramètre β comme $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ alors le modèle est encore:

$$\begin{aligned} y &= X\beta + \epsilon \\ &= (c \ x) \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \epsilon \\ &= c\beta_0 + x\beta_1 + \epsilon \\ &= \beta_0 c + \beta_1 x + \epsilon \end{aligned}$$

car β_0 et β_1 sont des réels.

Le problème des MCO est ensuite de trouver $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \widehat{\beta}$ tels que l'on minimise

$$\sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

Rappels d'algèbre Rappel 1: il est possible de multiplier la matrice A avec la matrice B si leurs tailles sont de la forme: A de taille (m, n) et B de taille (n, p) . Leur produit est alors une matrice de taille (m, p) .

Rappel 2: On peut définir une application bilinéaire, appelée produit scalaire,

de la façon suivante: soient

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad \text{et} \quad v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix},$$

alors la fonction produit scalaire s'écrit:

$$(u|v) = \sum_{i=1}^n u_i v_i.$$

La norme associée à ce produit scalaire est définie comme suit:

$$\|u\|^2 = (u|u) = \sum_{i=1}^n u_i^2$$

3.2 Vers une représentation géométrique

Ainsi, si l'on utilise ces notations, le problème des MCO consiste à minimiser $\|y - \beta X\|^2$ sur β . On pourra noter par ailleurs que quand β décrit \mathcal{R}^2 , βX décrit le sous-espace vectoriel F de \mathcal{R}^n engendré par le vecteur c et le vecteur x ; ce sous-espace est de dimension 2 car x n'est pas colinéaire à c par hypothèse (équivalent à dire que x n'est pas constant). Ceci revient à dire que l'on cherche dans ce sous-espace le vecteur $z = Xb = X\hat{\beta}$ qui minimise la distance à y .

Un sous-espace de dimension 2 est un plan (au sens usuel du terme). Imaginons que nous soyons dans un espace de dimension 3; alors le vecteur de F qui minimise la distance à y est la projection orthogonale de y sur F . Représentation graphique. La différence entre y et sa projection orthogonale ($\hat{y} = X\hat{\beta}$) est alors orthogonale au plan F ; il s'agit du résidu estimé $e = y - \hat{y}$. L'orthogonalité peut s'exprimer en utilisant le produit scalaire:

$$\begin{aligned} e \perp F = \text{Vect}(X) &\Leftrightarrow (e|c) = 0 \quad \text{et} \quad (e|x) = 0 \\ &\Leftrightarrow \sum_{i=1}^n e_i = 0 \quad \text{et} \quad \sum_{i=1}^n e_i x_i = 0 \\ &\Leftrightarrow X'e = \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} (c|e) \\ (x|e) \end{pmatrix} = 0 \end{aligned}$$

3.3 Calcul matriciel de l'estimateur des MCO

Nous cherchons, à partir de là, à exprimer l'estimateur des MCO de façon matricielle. On a vu deux conditions. La première est que la prédiction appartient au plan (espace vectoriel) généré par X , la seconde que le résidu est orthogonal à ce plan. Ces deux conditions s'écrivent:

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ X'(y - \hat{y}) &= 0\end{aligned}$$

Note: si l'on pense β comme constitué de deux paramètres inconnus, il s'agit ici de résoudre un système de deux équations (chacune des conditions est de dimension 1 et fournit donc une seule équation) à deux inconnues.

Du système découle:

$$X'y = X'\hat{y} = X'X\hat{\beta};$$

si la matrice $X'X$ est inversible alors

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Montrons que $X'X$ est effectivement inversible. Tout d'abord, il faut vérifier qu'il s'agit bien d'une matrice carrée: X est de dimension $(n, 2)$ donc X' est de dimension $(2, n)$ et le produit des deux est de dimension $(2, 2)$. Par ailleurs,

$$X'X = \begin{pmatrix} c' \\ x' \end{pmatrix} \begin{pmatrix} c & x \end{pmatrix} = \begin{pmatrix} c'c & c'x \\ x'c & x'x \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n} \sum x_i^2 \end{pmatrix}$$

dont le déterminant vaut

$$\frac{1}{n} \sum x_i^2 - \bar{x}^2 = V_{emp}(x) \neq 0$$

tant que la variable x n'est pas constante (hyp.). La matrice $X'X$ est donc bien inversible.

Par conséquent, l'expression ci-dessus fournit l'estimateur de β écrit de façon matricielle. On peut vérifier que l'on retrouve bien les expressions données en début de chapitre.

3.4 Les pièges des méthodes d'analyse de régression

Extrapolation avec une équation de régression Question: peut-on prédire le salaire d'un individu avec 9 ans d'études sur la base de la droite de régression? le salaire d'un individu n'ayant pas atteint le bac?

La régression qui utilise des données avec des individus ayant de 0 à 8 ans d'études après-bac est en mesure de fournir des prévisions de salaires pour des individus ayant des niveaux d'études dans cet intervalle. Il se peut tout à fait que la relation entre niveau d'étude et salaire ne soit pas la même en-dehors de ce support. En effet, la relation peut être linéaire pour un certain intervalle et présenter un tout autre comportement en dehors du champ observé. Tant que nous ne disposons pas de données supplémentaires, nous ne pouvons rien dire sur la relation en-dehors du support.

L'utilisation de la droite de régression en-dehors du support est une extrapolation, qu'il vaut mieux éviter de faire. Par ailleurs, même si des considérations théoriques ou pratiques nous conduisent à le faire, il faut bien voir que la marge d'erreur augmente au fur et à mesure que l'on s'éloigne de la valeur moyenne de la variable explicative.

Validation du modèle L'obtention de nouvelles données permet de vérifier la validité d'une équation de régression: en effet, il suffit de comparer leur réalisation à la prédiction qu'on aurait faite sur la base des données précédentes.

Si l'on dispose de données suffisamment importantes, il est aussi possible de mimer cet afflux de nouvelles données en écartant de l'estimation de la droite de régression une portion d'entre elles (choisies aléatoirement). Par la suite, on regarde pour ces observations disponibles si la prédiction obtenue est proche de la réalisation; si c'est le cas, le modèle est satisfaisant (notamment, la moyenne des erreurs doit être proche de 0).

Notez bien que cela n'aurait pas de sens de faire ce genre de vérifications à partir d'observations utilisées pour estimer la droite de régression. En effet, la droite de régression passe par définition par le milieu du nuage de points et, à moins de s'être trompé dans le calcul, la moyenne des erreurs est nulle pour tout sous-échantillon.

Relation causale Par ailleurs, nous avons introduit ce chapitre en arguant qu'il allait enfin être possible de mesurer l'effet d'une variable sur une

autre, c'est-à-dire d'identifier une relation causale de l'une sur l'autre. Cependant, cela n'est pas si simple. Dans le modèle que nous avons présenté, la variable explicative est une variable certaine (non aléatoire), ce qui est une hypothèse très forte. En effet, si l'on revient sur notre exemple, il est assez clair que le nombre d'années d'éducation relève d'un choix et donc d'un mécanisme décisionnel. Quelles sont les conséquences de ceci? si certains facteurs déterminent à la fois la variable explicative et la variable expliquée, il est fort possible que l'effet causal que l'on attribue à la variable explicative (augmentation du nombre d'années d'éducation implique hausse du salaire) soit en fait dû à ces autres facteurs non pris en compte. Imaginons un monde où le choix du nombre d'années d'éducation soit déterminé par la CSP des parents et que par ailleurs, plus les parents ont une position sociale élevée, plus l'enfant trouve un emploi bien rémunéré via les réseaux de ses parents. Dans ce cas-là, on observera effectivement que les individus avec plus d'années d'études sont mieux rémunérés, mais cela ne signifie pas pour autant que cela est une relation causale: tout est peut-être le fait d'une CSP élevée des parents...

La façon dont on traite ce genre de problèmes ne sera que partiellement traité dans ce cours mais il importe de garder un oeil critique lorsqu'on lit les résultats de régression pour penser à ce genre de mécanismes.