10. Regression Linéaire

Sam Gyetvay

ECO 2273 – Économetrie I

28 novembre 2025

Un aperçu de ce qui vous attend...

Dans tous vos futurs cours d'économétrie, vous passerez la plupart (sinon la totalité) de votre temps à étudier divers modèles de régression linéaire

Malheureusement, nous avons dû consacrer presque tout notre semestre à réviser des sujets plus élémentaires et fondamentaux

- Théorie des probabilités
- Variables aléatoires
- Échantillonnage
- Estimation
- Intervalle de confiance
- Tests d'hypothèses

Aujourd'hui, nous allons jeter un coup d'œil préliminaire sur ce sujet, pour "aiguiser votre appétit" pour ce qui vous attend

Aujourd'hui

La régression linéaire est une manière de résumer la relation entre deux variables X et Y

Nous commencerons notre leçon en révisant certains concepts de la théorie des probabilités concernant la relation entre deux variables

- Distribution conjointe, distribution conditionnelle
- Covariance, corrélation

Nous introduirons ensuite et discuterons des **espérances conditionnelles** et de la **Loi des espérances itérées**

Puis nous introduirons la régression linéaire. Nous concentrerons notre discussion sur

- Comment interpréter les régressions linéaires
- Comment les estimer à l'aide de données
- Comment réaliser des tests d'hypothèses/créer des intervalles de confiance

Distributions conjointes

Pour deux variables aléatoires discrètes X et Y, une **fonction de probabilité conjointe** ou une **distribution conjointe** $p(x, y)$ donne la probabilité que X prenne une valeur spécifique et que Y prenne une valeur spécifique simultanément :

$$p(x, y) = P(X = x, Y = y)$$

Pour deux variables aléatoires continues X et Y, une **fonction de densité conjointe** ou une **distribution conjointe** $f_{X,Y}(x, y)$ décrit la densité de probabilité que X et Y prennent des valeurs spécifiques simultanément :

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f_{X,Y}(x, y)\, dy\, dx$$

Propriétés des distributions conjointes (discrètes)

Pour une fonction de probabilité conjointe $p(x, y)$ de variables aléatoires discrètes X et Y, les propriétés suivantes sont toujours vérifiées :

Non-négativité :

$$p(x, y) \geq 0 \quad \text{pour tout } x, y$$

Somme à un :

$$\sum_x \sum_y p(x, y) = 1$$

Distributions marginales : Nous pouvons retrouver la distribution de X (en ignorant Y) en faisant la somme sur toutes les valeurs de Y :

$$p(x) = \sum_y p(x, y)$$

Dans ce contexte, nous appelons $p(x)$ la **distribution marginale** de X

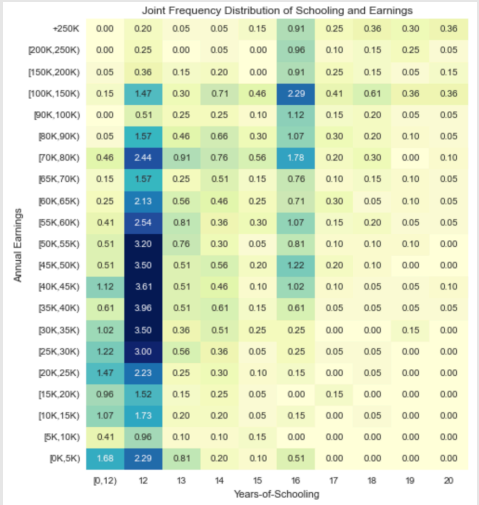
Cartes de chaleur

Parfois, les distributions conjointes peuvent être visualisées efficacement à l'aide de **cartes de chaleur**

Une carte de chaleur est une grille bidimensionnelle, où chaque cellule représente la probabilité d'une observation dans une combinaison spécifique de X et Y

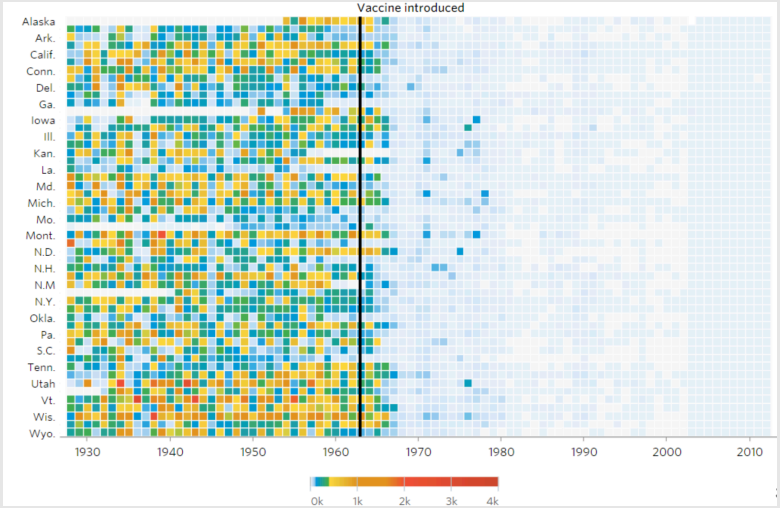
Les cellules sont ensuite colorées, généralement en utilisant des couleurs plus sombres pour les cellules avec des probabilités plus élevées

La figure à droite représente la distribution conjointe entre les années de scolarité et les revenus annuels dans le NLSY79

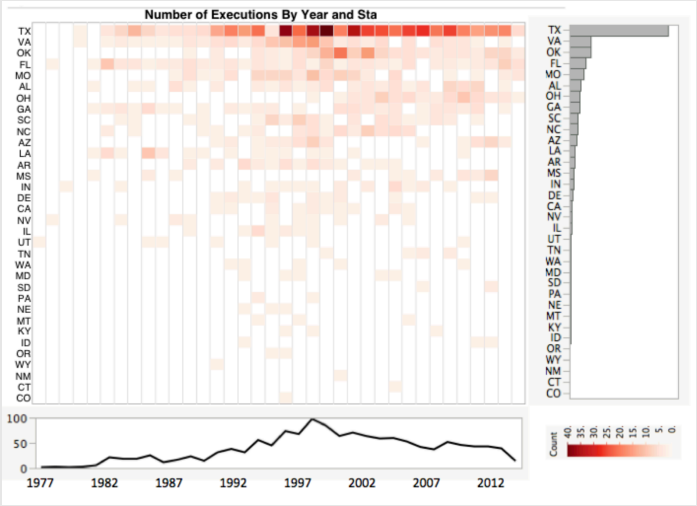


Source :

`https://eml.berkeley.edu/~bgraham/Teaching/Ec240a_Fall2015/Ec240a_Python_Notebook_1.html`



Source : https://graphics.wsj.com/infectious-diseases-and-vaccines/



Source : https://www.ericzwick.com/heatmap/heatmaps.pdf

Trop d'informations ?

La distribution conjointe vous dit *tout* sur la relation entre deux variables

Mais savoir tout est trop. Le cerveau humain recherche la simplicité

Les cartes de chaleur, bien qu'esthétiques, sont rarement utilisées en économie pour des raisons pratiques

- Il faut diviser les variables continues en intervalles. Quels intervalles utiliser ?
- L'ordre des variables discrètes non numériques est arbitraire

Nous voulons des moyens plus simples de résumer la relation entre Y et X

Résumé de la relation entre Y et X

Nous avons déjà vu deux statistiques qui résument la relation entre deux variables

- Covariance
- Corrélation

La covariance et la corrélation sont symétriques : $Cov(Y, X) = Cov(X, Y)$, $\rho_{X,Y} = \rho_{Y,X}$

Mais parfois, nous voulons comprendre les relations de manière non symétrique

- X est une variable qui prédit Y
- X est une variable qui affecte la valeur de Y
- Y est la “variable dépendante”
- X est la “variable indépendante”

L’espérance conditionnelle et la régression linéaire sont des moyens de représenter ces types de relations

Rappel : Covariance

La **covariance** est

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

L'espérance est prise par rapport à la distribution conjointe des variables aléatoires X et Y

- Si X et Y sont discrètes :

$$Cov(X, Y) = \sum_x \sum_y (x - E[X])(y - E[Y])\, p(x, y)$$

- Si X et Y sont continues :

$$Cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E[X])(y - E[Y])\, f_{X,Y}(x, y)\, dy\, dx$$

Corrélation

La **corrélation** est

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)\,Var(Y)}}$$

La corrélation est toujours comprise entre -1 et 1

- Corrélation = 0 : aucune relation (linéaire) entre X et Y
- Corrélation = -1 : relation linéaire négative parfaite entre X et Y
- Corrélation = +1 : relation linéaire positive parfaite entre X et Y

Espérance conditionnelle

L'**espérance conditionnelle** de Y sachant X décrit la valeur attendue de Y lorsque X prend une valeur particulière.

X et Y discrètes

$$E[Y \mid X = x] = \sum_{y} y \cdot P(Y = y \mid X = x)$$

X et Y continues

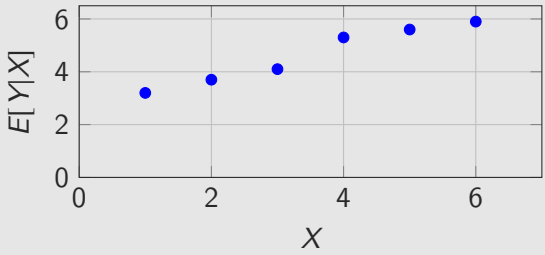
$$E[Y \mid X = x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dy$$

L'espérance conditionnelle est une fonction décrivant la valeur moyenne ou prévue de Y lorsque X prend n'importe quelle valeur

Espérance conditionnelle : exemple # 1

Si Y = nombre de pièces dans une maison, X = nombre de résidents, $E[Y|X]$ est le nombre moyen de pièces dans une maison avec X résidents

X	$E[Y\|X]$
1	3.2
2	3.7
3	4.1
4	5.3
5	5.6
6	5.9



Espérance conditionnelle : exemple #2

Si Y = salaire horaire, X = genre ($X = 0$ homme, $X = 1$ femme), alors $E[Y|X]$ est le revenu moyen au sein de chaque genre.

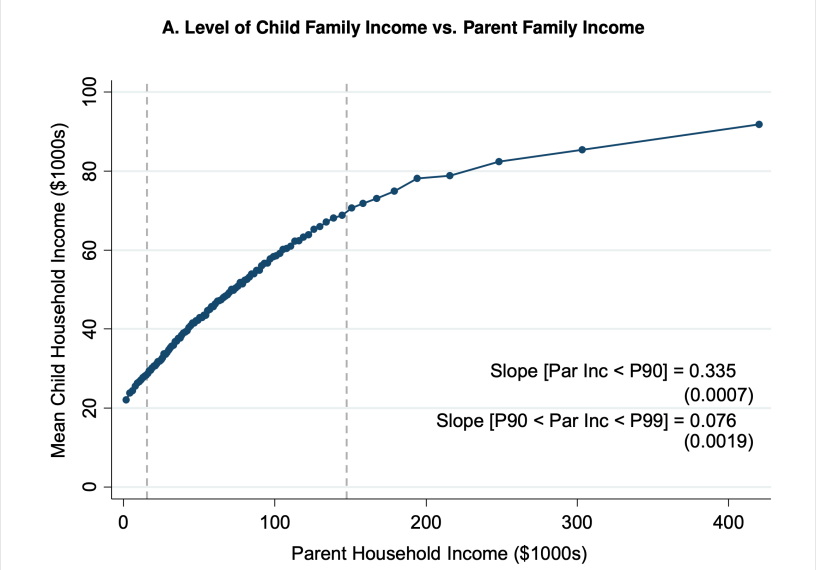
Période	$E[Y\|0]$	$E[Y\|1]$	$E[Y\|1] - E[Y\|0]$	$E[Y\|1]/E[Y\|0]$	$P(X = 1)$
1997-2001	24.32	19.51	-4.81	0.802	0.48
2013-2017	26.85	22.96	-4.89	0.855	0.50

Espérance conditionnelle : exemple #3

Supposons que nous menions une expérience où $X = 1$ si une personne bénéficie d'une assurance santé et $X = 0$ autrement. Laissez $Y = 1$ si la personne meurt dans l'année, $Y = 0$ autrement. Alors $E[Y|X]$ est le taux de mortalité dans chaque groupe.

$E[Y\|0]$	$E[Y\|1]$	$E[Y\|1] - E[Y\|0]$	$E[Y\|1]/E[Y\|0]$
0.012	0.008	-0.004	0.667

Espérance conditionnelle : exemple #4



Calcul des moyennes conditionnelles

Une moyenne conditionnelle est la version échantillonnale d'une espérance conditionnelle. Nous calculons simplement la moyenne de Y séparément pour chaque valeur de X. Dans cet exemple, nous calculons une moyenne pondérée.

Y	X	w
n_rooms	n_residents	n_households
1	1	10
2	1	12
3	1	14
4	1	16
5	1	18
1	2	20
2	2	22
3	2	24
4	2	26
5	2	28

Calcul des moyennes conditionnelles

Une moyenne conditionnelle est la version échantillonnale d'une espérance conditionnelle. Nous calculons simplement la moyenne de Y séparément pour chaque valeur de X. Dans cet exemple, nous calculons une moyenne pondérée.

Y	X	w
n_rooms	n_residents	n_households
1	1	10
2	1	12
3	1	14
4	1	16
5	1	18

$$E[Y \mid X = 1] = \frac{1}{\sum w_i} \sum w_i r_i$$

$$\sum w_i = 10 + 12 + 14 + 16 + 18 = 70$$

$$\sum w_i r_i = 1 \cdot 10 + 2 \cdot 12 + 3 \cdot 14 + 4 \cdot 16 + 5 \cdot 18$$
$$= 220$$

$$E[Y \mid X = 1] = \frac{220}{70} = 3.14$$

Calcul des moyennes conditionnelles

Une moyenne conditionnelle est la version échantillonnale d'une espérance conditionnelle. Nous calculons simplement la moyenne de Y séparément pour chaque valeur de X. Dans cet exemple, nous calculons une moyenne pondérée.

Y n_rooms	X n_residents	w n_households
1	2	20
2	2	22
3	2	24
4	2	26
5	2	28

$$E[Y \mid X = 2] = \frac{1}{\sum w_i} \sum w_i r_i$$

$$\sum w_i = 20 + 22 + 24 + 26 + 28 = 120$$

$$\sum w_i r_i = 1 \cdot 20 + 2 \cdot 22 + 3 \cdot 24 + 4 \cdot 26 + 5 \cdot 28 = 300$$

$$E[Y \mid X = 2] = \frac{300}{120} = 2.5$$

Loi des espérances itérées

C'est un résultat très utile qui apparaît partout en économétrie. Il stipule que l'espérance d'une espérance conditionnelle est l'espérance inconditionnelle :

$$E[E[Y|X]] = E[Y]$$

Voici une preuve (juste pour Sam et les nerds, pas à l'examen) :

$$\begin{aligned}
E[E[Y|X]] &= \int E[Y|X = x] f_X(x)\, dx \\
&= \int \left(\int y \cdot f_{Y|X}(y|x)\, dy \right) f_X(x)\, dx \\
&= \int \int y \cdot f_{Y|X}(y|x) f_X(x)\, dy\, dx \\
&= \int \int y \cdot f_{Y,X}(y, x)\, dy\, dx \\
&= \int y \cdot \left(\int f_{Y,X}(y, x)\, dx \right) dy = \int y \cdot f_Y(y)\, dy = E[Y]
\end{aligned}$$

Loi des espérances itérées : exemple

Revenons à notre exemple de l'écart salarial entre les genres mentionné précédemment

Période	$E[Y\|0]$	$E[Y\|1]$	$E[Y\|1] - E[Y\|0]$	$E[Y\|1]/E[Y\|0]$	$P(X = 1)$
1997-2001	24.32	19.51	-4.81	0.802	0.48
2013-2017	26.85	22.96	-4.89	0.855	0.50

Nous pouvons utiliser la LIE pour calculer le salaire moyen de chaque période. Faisons-le pour 1997-2001 :

$$\begin{aligned} E[E[Y|X]] &= E[Y|X = 0] \cdot P(X = 0) + E[Y|X = 1] \cdot P(X = 1) \\ &= 24.32 \cdot 0.52 + 19.51 \cdot 0.48 = 22.01 \end{aligned}$$

La LIE indique que le salaire moyen est une moyenne pondérée du salaire moyen des femmes et des hommes, où les poids sont les parts des femmes et des hommes

Trop d'informations ?

L'espérance conditionnelle vous dit tout sur la manière dont la valeur moyenne de Y change avec X

Mais savoir tout est trop. Le cerveau humain recherche la simplicité

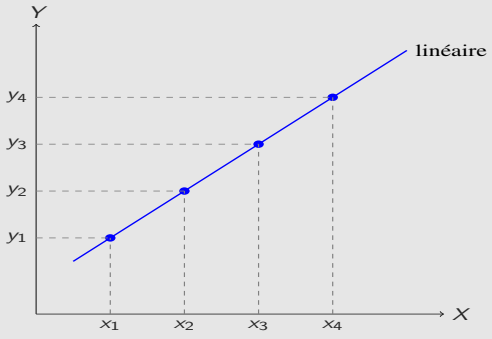
Les espérances conditionnelles sont utilisées occasionnellement en économie

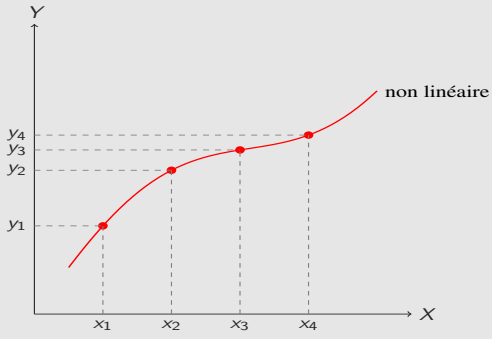
- Comme avec les cartes de chaleur : que faites-vous lorsque X est continu ?
- Réponse nerd : utiliser des méthodes de noyau « non paramétriques »

Les économistes préfèrent résumer la relation entre Y et X en utilisant des régressions linéaires

Linéarité

Linéaire signifie que la pente est constante : $\frac{y_2-y_1}{x_2-x_1} = \frac{y_4-y_3}{x_4-x_3}$ pour n'importe quelles valeurs de y et x





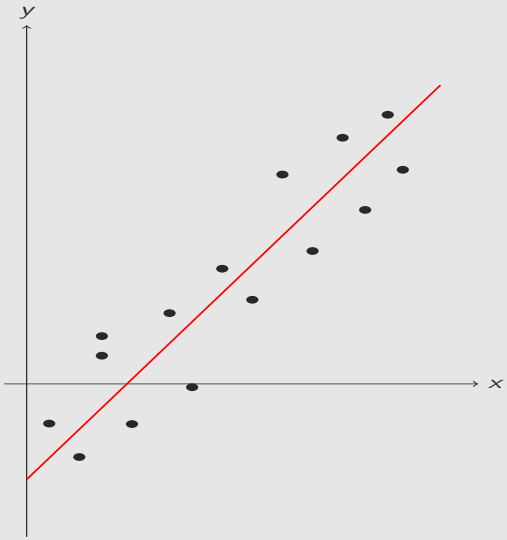
Régression linéaire simple

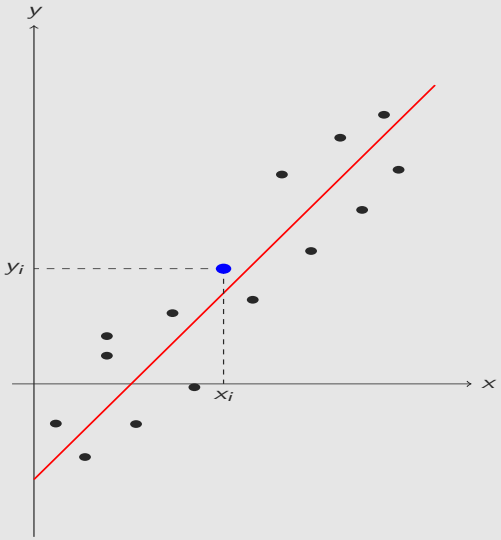
Une régression linéaire simple est un modèle de la forme

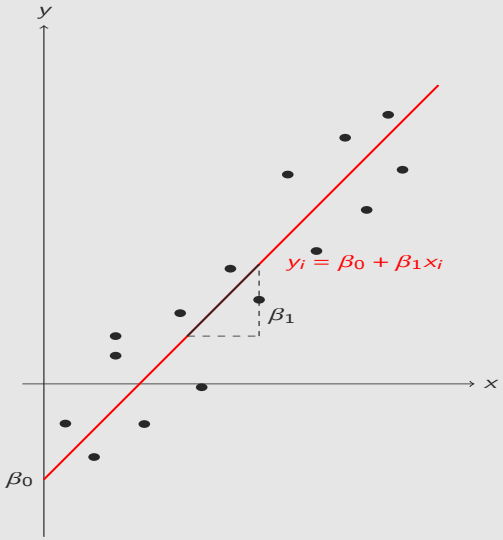
$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

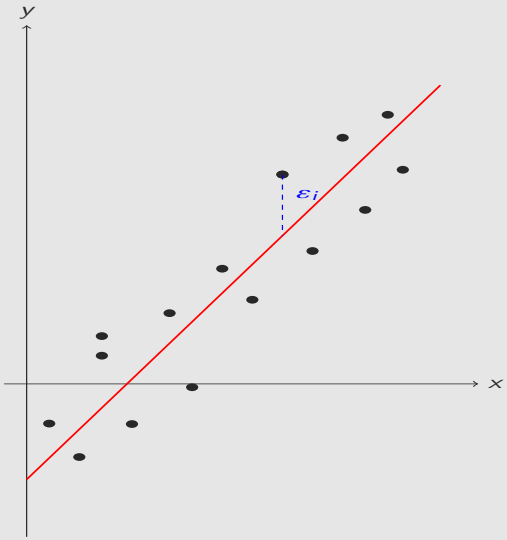
- $i \in \{1, \ldots, n\}$ indexe les observations
- y_i est la **variable dépendante**
- x_i est la **variable indépendante**
- ε_i est l'**erreur**
- β_0 et β_1 sont les **coefficients**
  - β_0 est l'**ordonnée à l'origine** ou **constante**
  - β_1 est la **pente**

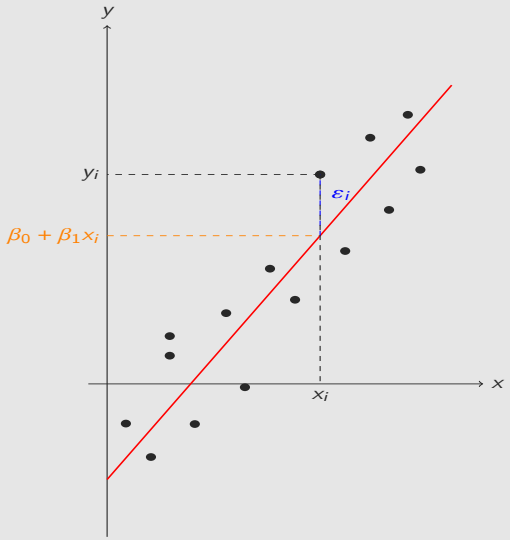
"Simple" fait référence au fait qu'il n'y a qu'une seule variable dépendante x_i


y
x


y
y_i
x
x_i


y
x
$y_i = \beta_0 + \beta_1 x_i$
β_1
β_0


y
x
ε_i


y
y_i
ε_i
$\beta_0 + \beta_1 x_i$
x
x_i

Interprétation #1 : prédiction

Nous pouvons interpréter l'équation de régression linéaire

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

comme un modèle qui **prédit** la valeur de y_i en utilisant la valeur de x_i

Supposons qu'une valeur de x_i vous soit donnée mais pas celle de y_i, et qu'on vous demande de prédire la valeur de y_i. Selon le modèle de régression linéaire, vous devriez prédire y_i =

Interprétation #1 : prédiction

Nous pouvons interpréter l'équation de régression linéaire

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

comme un modèle qui **prédit** la valeur de y_i en utilisant la valeur de x_i

Supposons qu'une valeur de x_i vous soit donnée mais pas celle de y_i, et qu'on vous demande de prédire la valeur de y_i. Selon le modèle de régression linéaire, vous devriez prédire $y_i = \beta_0 + \beta_1 x_i$

Supposons maintenant qu'après avoir fait votre prédiction, on vous donne la vraie valeur de y_i. Alors $\varepsilon_i =$

Interprétation #1 : prédiction

Nous pouvons interpréter l'équation de régression linéaire

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

comme un modèle qui **prédit** la valeur de y_i en utilisant la valeur de x_i

Supposons qu'une valeur de x_i vous soit donnée mais pas celle de y_i, et qu'on vous demande de prédire la valeur de y_i. Selon le modèle de régression linéaire, vous devriez prédire $y_i = \beta_0 + \beta_1 x_i$

Supposons maintenant qu'après avoir fait votre prédiction, on vous donne la vraie valeur de y_i. Alors $\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$ est l'**erreur de prédiction**

Exemple : prédire la valeur du taux de chômage ce mois-ci en utilisant la valeur du taux de chômage le mois dernier

Interprétation #2 : description

Nous pouvons également interpréter la régression linéaire comme un modèle de l'**espérance conditionnelle** de y_i étant donné x_i, $E[Y|X = x_i]$

Rappelez-vous que $E[Y|X = x_i]$ est la valeur moyenne de Y parmi les individus ayant $X = x_i$. L'espérance conditionnelle est une manière d'exprimer la relation entre Y et X comme une fonction qui vous donne la valeur moyenne de Y pour chaque valeur de X

La régression linéaire peut être interprétée comme une approximation de l'espérance conditionnelle

$$y_i = E[y_i|x_i] + (y_i - E[y_i|x_i])$$

si $E[y_i|x_i] = \beta_0 + \beta_1 x_i$, alors

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

Interprétation #3 : causalité

Nous pouvons également interpréter la régression linéaire comme un modèle de l'**effet causal** de x_i sur y_i

Supposons que nous ayons une observation (y_i, x_i). Par exemple, y_i est la note de l'étudiant i dans ECO2273 et x_i le nombre d'heures qu'ils ont passé à étudier. Alors ε_i représente l'effet de tous les autres facteurs qui déterminent les notes, tels que l'intelligence, la fréquentation des cours et les pots-de-vin financiers au professeur.

Supposons maintenant que nous remontions le temps, et que nous obligions l'étudiant i à étudier pendant $x_i + 1$ heures, en maintenant tout le reste (ε_i) constant. Nous observons alors leur note **contrefactuelle** y_i^*

Dans ce contexte, β_1 est l'**effet causal** d'une heure supplémentaire d'étude

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ y_i^* &= \beta_0 + \beta_1 (x_i + 1) + \epsilon_i \\ y_i^* - y_i &= \beta_1 \end{aligned}$$

Quelle interprétation utiliser ?

Une régression linéaire est un modèle. Elle n'a pas d'interprétation inhérente. Comment l'interpréter est un jugement qui doit être fait en fonction du contexte

Une régression linéaire peut être interprétée comme une prédiction chaque fois que nous essayons de prévoir une valeur inconnue, comme une valeur future (taux de chômage, taux d'inflation, prix des actions, etc.)

Une régression linéaire peut être interprétée de manière descriptive comme représentant la valeur moyenne de Y pour une valeur donnée de X de manière assez générale

Une régression linéaire ne peut être interprétée de manière causale que dans des circonstances très particulières, comme lorsque X a été défini de manière aléatoire dans le cadre d'une expérience

Moindres carrés

Une fois que nous avons un ensemble de données $(y_i, x_i)_{i=1}^n$ et que nous voulons estimer une régression linéaire, comment procédons-nous pour estimer β_0 et β_1 ?

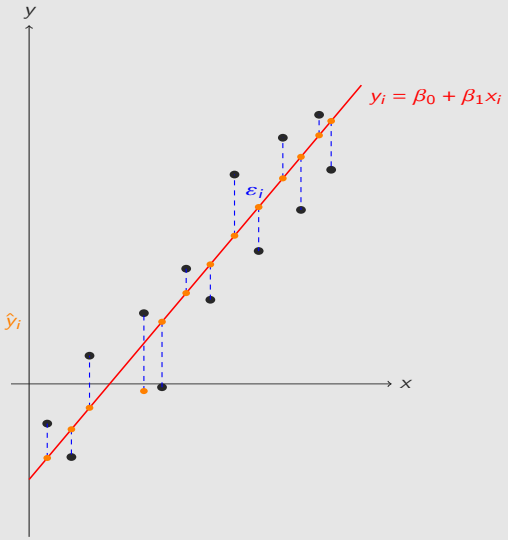
Nous utilisons la **méthode des moindres carrés**

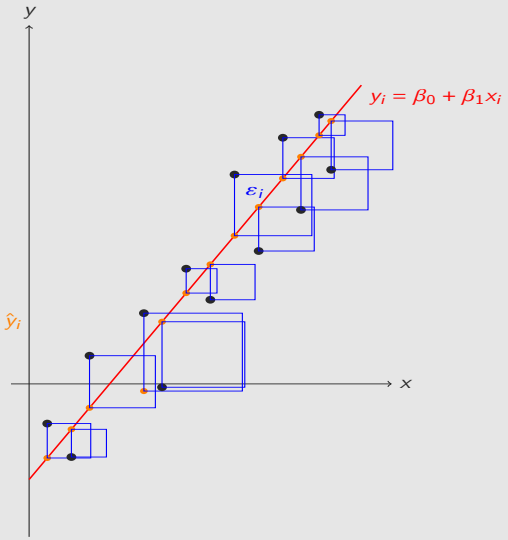
Les moindres carrés nous disent de trouver les valeurs de $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent la **somme des résidus au carré**

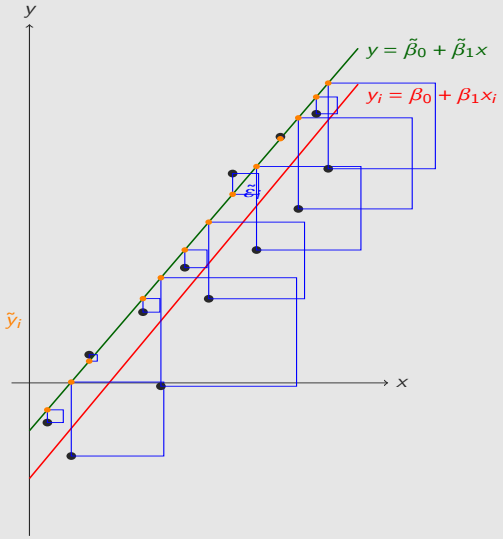
$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

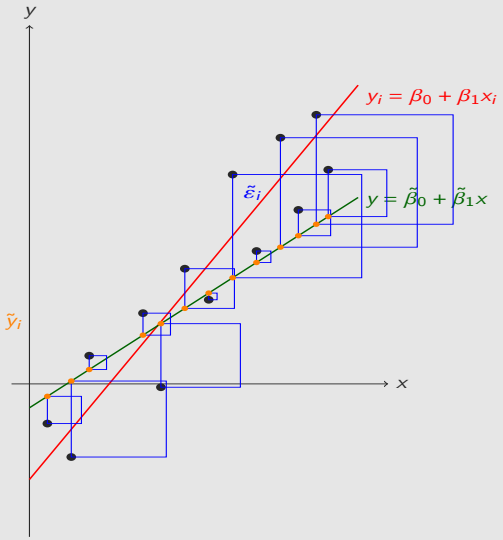
Nous prenons le carré parce que nous voulons éviter de surestimer et de sous-estimer y_i

Cela minimise la distance entre les données et $\beta_0 + \beta_1 x_i$, parfois appelée la **ligne de meilleur ajustement**


y
x
$y_i = \beta_0 + \beta_1 x_i$
ε_i
$\hat{y}_i$


y
x
$y_i = \beta_0 + \beta_1 x_i$
ε_i
$\hat{y}_i$


y
x
$y = \tilde{\beta}_0 + \tilde{\beta}_1 x$
$y_i = \beta_0 + \beta_1 x_i$
$\tilde{y}_i$
β_{1i}


y
x
$y_i = \beta_0 + \beta_1 x_i$
$y = \tilde{\beta}_0 + \tilde{\beta}_1 x$
$\tilde{\varepsilon}_i$
$\tilde{y}_i$

Estimations MCO des coefficients

Il existe des formules pour les estimations MCO des coefficients β_0 et β_1

$$\hat{\beta}_1 = \frac{Cov(y,x)}{Var(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Ce sont les valeurs des coefficients qui minimisent la somme des erreurs au carré

Interprétation de la formule pour $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{Cov(y,x)}{Var(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

La covariance mesure dans quelle mesure Y et X "évoluent ensemble"

- si y_i est au-dessus (en dessous) de $\bar{y}$ lorsque x_i est au-dessus (en dessous) de $\bar{x}$, covariance positive
- si y_i est en dessous (au-dessus) de $\bar{y}$ lorsque x_i est au-dessus (en dessous) de $\bar{x}$, covariance négative

La variance de x_i mesure combien X "varie"

- Si x_i est toujours proche de sa moyenne, petite variance
- si x_i est souvent loin de sa moyenne, grande variance

Le rapport des deux nous indique combien Y varie pour un mouvement typique de X

Relation entre $\hat{\beta}_1$ et ρ_{XY}

L'estimation MCO du coefficient $\hat{\beta}_1$ est étroitement liée à la mesure de corrélation ρ_{XY}

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

$$\hat{\beta}_1 = \frac{Cov(y,x)}{\sqrt{Var(x)\,Var(x)}} = \frac{Cov(y,x)}{\sqrt{Var(x)\,Var(x)}} \cdot \frac{\sqrt{Var(y)}}{\sqrt{Var(y)}} = \frac{Cov(y,x)}{\sqrt{Var(x)\,Var(y)}} \cdot \frac{\sqrt{Var(y)}}{\sqrt{Var(x)}} = \rho_{y,x} \cdot \frac{\sigma_y}{\sigma_x}$$

La pente estimée est la corrélation multipliée par le rapport des erreurs-types !

Si $\sigma_y = \sigma_x$, alors $\rho_{y,x} = \hat{\beta}_1$!

Interprétation de la formule pour $\hat{\beta}_0$

La formule pour l'ordonnée à l'origine $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

peut être réarrangée en

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

ce qui implique que la droite de meilleur ajustement passe toujours par $(\bar{y}, \bar{x})$!

Choisir $\hat{\beta}_0$ de cette manière assure que la droite de meilleur ajustement est au milieu du "nuage de points"

Valeurs ajustées et résidus

Après avoir effectué une régression et estimé $\hat{\beta}_0$ et $\hat{\beta}_1$, nous pouvons les insérer dans l'équation pour obtenir les **valeurs ajustées** ou **valeurs prédites**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

ce sont les valeurs le long de la ligne de meilleur ajustement pour chaque point de nos données

Nous pouvons également calculer les **résidus** pour chaque point

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

les résidus sont l'écart entre les valeurs réelles y_i et les valeurs ajustées pour chaque point de nos données

Les résidus des moindres carrés sont toujours de moyenne nulle

Lorsque nous effectuons une régression, les résidus ont toujours une moyenne égale à zéro

$$\frac{1}{n}\sum_i \hat{\varepsilon}_i = 0$$

Cela est par construction. Cela provient de la manière dont nous avons défini $\hat{\beta}_0$:

$$\begin{aligned}\frac{1}{n}\sum_i \hat{\varepsilon}_i &= \frac{1}{n}\sum_i y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= \frac{1}{n}\sum_i y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i = \frac{1}{n}\sum_i (y_i - \bar{y}) - \hat{\beta}_1 \frac{1}{n}\sum_i (x_i - \bar{x})\end{aligned}$$

Quelle part de la variance de Y est expliquée par X ?

La réponse à cette question est donnée par le R^2 (“R carré”)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Le terme $\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$ est la fraction de variance que le modèle ne peut pas expliquer

- Numérateur : somme des carrés des résidus (SSR)
- Dénominateur : variance de y_i

Ainsi, un moins ce terme est la fraction de variance que le modèle *peut* expliquer

Inférence pour les coefficients de régression

Les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ sont basées sur un échantillon particulier

Nous ne connaissons pas la valeur réelle de β_0 et β_1 dans la population

Si nous avions tiré un échantillon différent, nous aurions obtenu des estimations différentes de $\hat{\beta}_0$ et $\hat{\beta}_1$

Tout comme nous l'avons fait avec les moyennes d'échantillon, nous pouvons calculer des erreurs standard, former des intervalles de confiance, calculer des valeurs p, et réaliser des tests d'hypothèses sur nos coefficients de régression $\hat{\beta}_0$ et $\hat{\beta}_1$. Cela nous permet d'exprimer l'incertitude de nos estimations

Lorsque X est une variable binaire

Supposons que la variable indépendante soit binaire :

$$X_i \in \{0, 1\}, \qquad Y_i = \beta_0 + \beta_1 X_i + u_i.$$

Dans ce cas, $\hat{\beta}_1$ mesure la **différence dans la valeur attendue de** Y entre les deux groupes :

$$\hat{\beta}_1 = \hat{E}[Y \mid X = 1] - \hat{E}[Y \mid X = 0].$$

Et l'intercept $\hat{\beta}_0$ est la moyenne pour le groupe avec $X = 0$

$$\hat{\beta}_0 = \hat{E}[Y \mid X = 0].$$

Erreur standard de $\hat{\beta}_1$

L'erreur standard de $\hat{\beta}_1$ est

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (x_i - \bar{x})^2}}$$

Tout comme pour les moyennes d'échantillons, nous pouvons utiliser cette erreur standard pour former des intervalles de confiance, calculer des valeurs de p et réaliser des tests d'hypothèses

Intervalle de confiance à 95% pour $\hat{\beta}_1$

Un intervalle de confiance à 95% pour le coefficient de pente est

$$\hat{\beta}_1 \pm t_{n-2,\,0.975}\,\hat{\sigma}_{\hat{\beta}_1},$$

où

$$\hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{1}{n-2} \cdot \frac{\sum_i \hat{\varepsilon}_i^2}{\sum_i (x_i - \bar{x})^2}}$$

La valeur critique $t_{n-2,\,0.975}$ est le 97,5e percentile de la distribution t avec $n-2$ degrés de liberté

Lorsque n est grand, nous pouvons utiliser l'approximation $t_{n-2,\,0.975} \approx 2$

Test d'hypothèse de $\beta_1 = 0$ au niveau de 5%

Nous voulons tester

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0.$$

Statistique de test :

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}.$$

Règle de décision (niveau 5%, bilatéral) :

$$\text{Rejeter } H_0 \quad \text{si} \quad |t| > t_{n-2,\,0.975}.$$

La valeur critique $t_{n-2,\,0.975}$ est le 97,5e percentile de la distribution t avec $n-2$ degrés de liberté. Lorsque n est grand, nous pouvons utiliser l'approximation $t_{n-2,\,0.975} \approx 2$

Valeur p pour le test $H_0 : \beta_1 = 0$

Pour le test d'hypothèse

$$H_0 : \beta_1 = 0 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0,$$

la statistique de test est

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}.$$

La valeur p, sous H_0, d'observer une statistique t aussi extrême (en valeur absolue) que celle calculée

Pour un test au niveau de 5%, rejeter H_0 si $p < 0,05$

Résultats de régression dans STATA

```
. reg gnppc safewater
```

Source	SS	df	MS		
				Number of obs	= 37
				F(1, 35)	= 34.84
Model	1.8971e+09	1	1.8971e+09	Prob > F	= 0.0000
Residual	1.9059e+09	35	54453421.5	R-squared	= 0.4989
				Adj R-squared	= 0.4845
Total	3.8030e+09	36	105639058	Root MSE	= 7379.3

gnppc	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
safewater	406.3531	68.84412	5.90	0.000	266.5921	546.1141
_cons	-23217.96	5338.333	-4.35	0.000	-34055.35	-12380.57

Comprendre la sortie de régression de Stata

`regress gnppc safewater` $\rightarrow$ $gnppc_i = \beta_0 + \beta_1\, safewater_i + u_i$

- `Nombre d'obs` $\rightarrow$ taille de l'échantillon n.
- `Coefficient`
  - `safewater` $\rightarrow \hat{\beta}_1$, `_cons` $\rightarrow \hat{\beta}_0$
- `Err. std.` — erreurs standard estimées $\hat{\sigma}_{\hat{\beta}_0}$ et $\hat{\sigma}_{\hat{\beta}_1}$
- `t` — statistiques t pour tester $H_0 : \beta_j = 0$ (bilatéral)

$$t_j = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\hat{\beta}_j}}$$

- `P>|t|` — valeurs p bilatérales pour les tests t ci-dessus
- `[Intervalle conf. 95%]` —

$$\hat{\beta}_j \pm t_{n-2,\,0.975}\, \hat{\sigma}_{\hat{\beta}_j}.$$

- `R-carré` - $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$

Semaine prochaine

Il n'y aura pas de nouveau matériel vendredi prochain (5 déc.). Nous avons terminé. Hourra !

Je posterai un grand nombre de problèmes d'entraînement sur le site web du cours vers le milieu de la semaine prochaine (environ mercredi, 3 déc.). Je ferai une annonce lorsque je le ferai.

Vendredi prochain (5 déc.), je ferai une révision de tout le matériel, répondrai aux questions, clarifierai ce que vous devez savoir pour l'examen, et résoudrai certains (mais pas tous) des problèmes d'entraînement.

Le lundi suivant (9 déc.), Oumar et/ou Sam résoudront quelques problèmes d'entraînement supplémentaires lors de la séance.

Le vendredi suivant (12 déc.) est l'examen final, en classe. Il sera similaire à l'examen de mi-session, mais avec quelques problèmes supplémentaires.