Séance 5 : modèles de régression

Introduction à la sociologie quantitative, niveau 1

Samuel Coavoux

Régression linéaire simple

Motivations et intuitions

Situation de départ

Après avoir mené une enquête auprès de 395 étudiants dans un lycée portugais (Cortez, 2014), nous voudrions **modéliser très simplement comment leurs journées d'absences influencent leurs résultats en mathématiques** (à la fin de l'année, donc au troisième trimestre).

Les deux variables sont quantitatives et continues : les notes s'étalent entre 0 et 20, les journées d'absences entre 0 et 100.

Modéliser cette relation signifie que nous pourrions savoir **comment une journée supplémentaire d'absence (X) influencerait les notes (Y)**, voire même que nous pourrions « prédire » les notes des élèves selon leur nombre d'absences.

Note : la section de ces slides sur la régression linéaire a été rédigée par Gabriel Alcaras dans le cadre d'un cours fait en commun.

Limites des autres méthodes

Les méthodes statistiques abordées précédemment sont malheureusement insuffisantes pour répondre à cette question :

- Nous sommes face à des variables quantitatives continues : *un test de khi-deux nécessite des variables catégorielles* et ne peut donc être appliqué.
- Même si nous recodions nos variables continues en catégorielles, le khi-deux ne peut que nous indiquer si la relation est significative ou non, sans plus de détails.
- Calculer le coefficient de corrélation r n'est pas entièrement satisfaisant : nous ne voulons pas juste savoir si notes (Y) et absences (X) sont corrélés, mais pouvoir modéliser précisément cette relation.

Il faut donc trouver une autre méthode. **La régression linéaire propose un modèle pour expliquer des variables quantitatives continues**.

Une modélisation simple

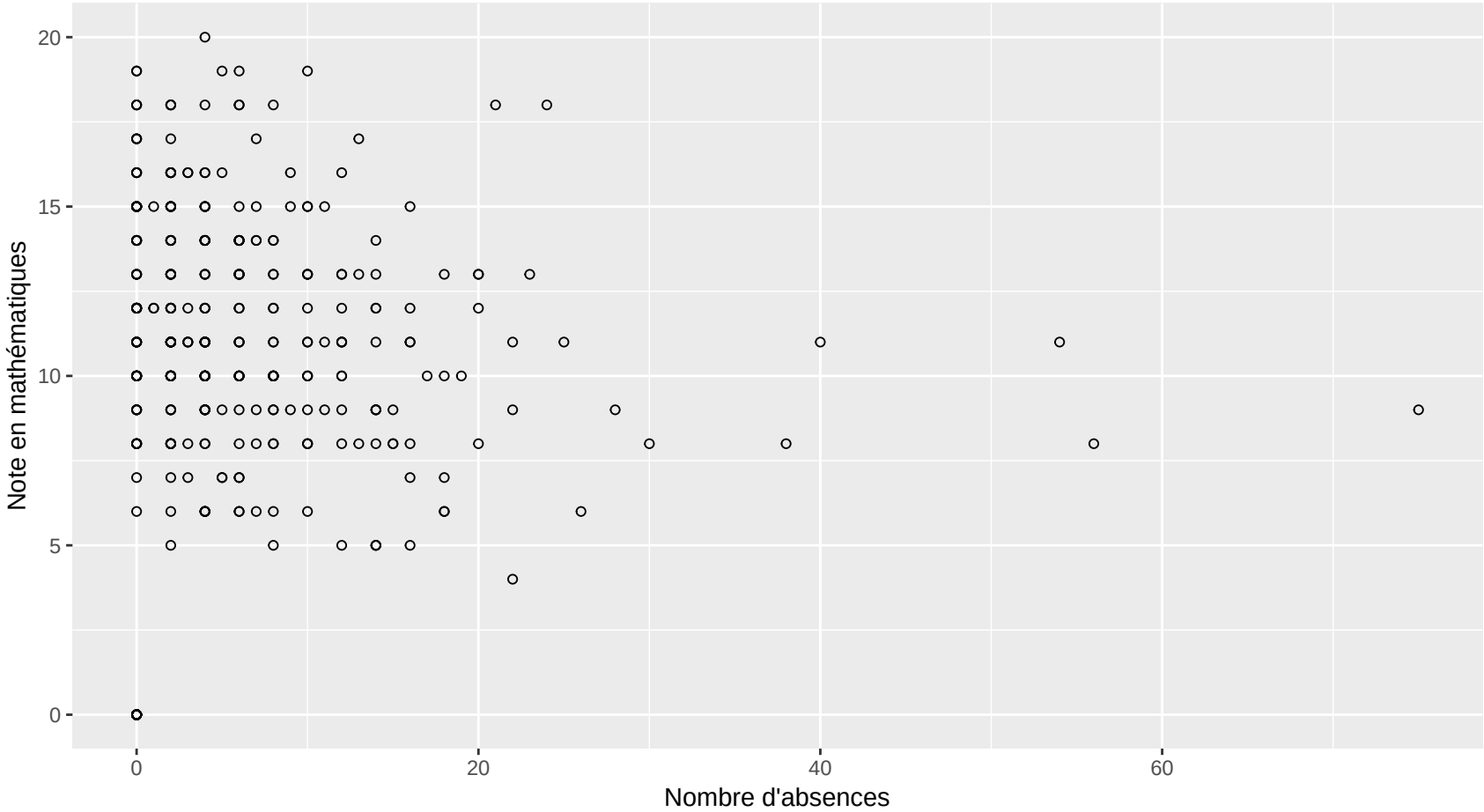
Il existe un moyen très simple de modéliser la relation entre deux variables quantitatives continues : une **équation affine** (par abus de langage, on dit souvent **linéaire**), c'est-à-dire l'équation d'une droite.

$$Y = \beta_0 + \beta_1 X_1, \text{ où :}$$

- Y est notre variable **à expliquer** (la note en mathématiques)
- X_1 est notre variable **explicative** (le nombre d'absences)
- β_0 est notre *ordonnée à l'origine* (là où notre droite croise l'axe vertical), c'est-à-dire la valeur de la note en maths quand les élèves ne s'absentent jamais)
- β_1 est le *coefficient directeur* associé à X_1 (la "pente" de notre droite), c'est-à-dire le taux de progression en maths quand les élèves s'absentent un jour de plus

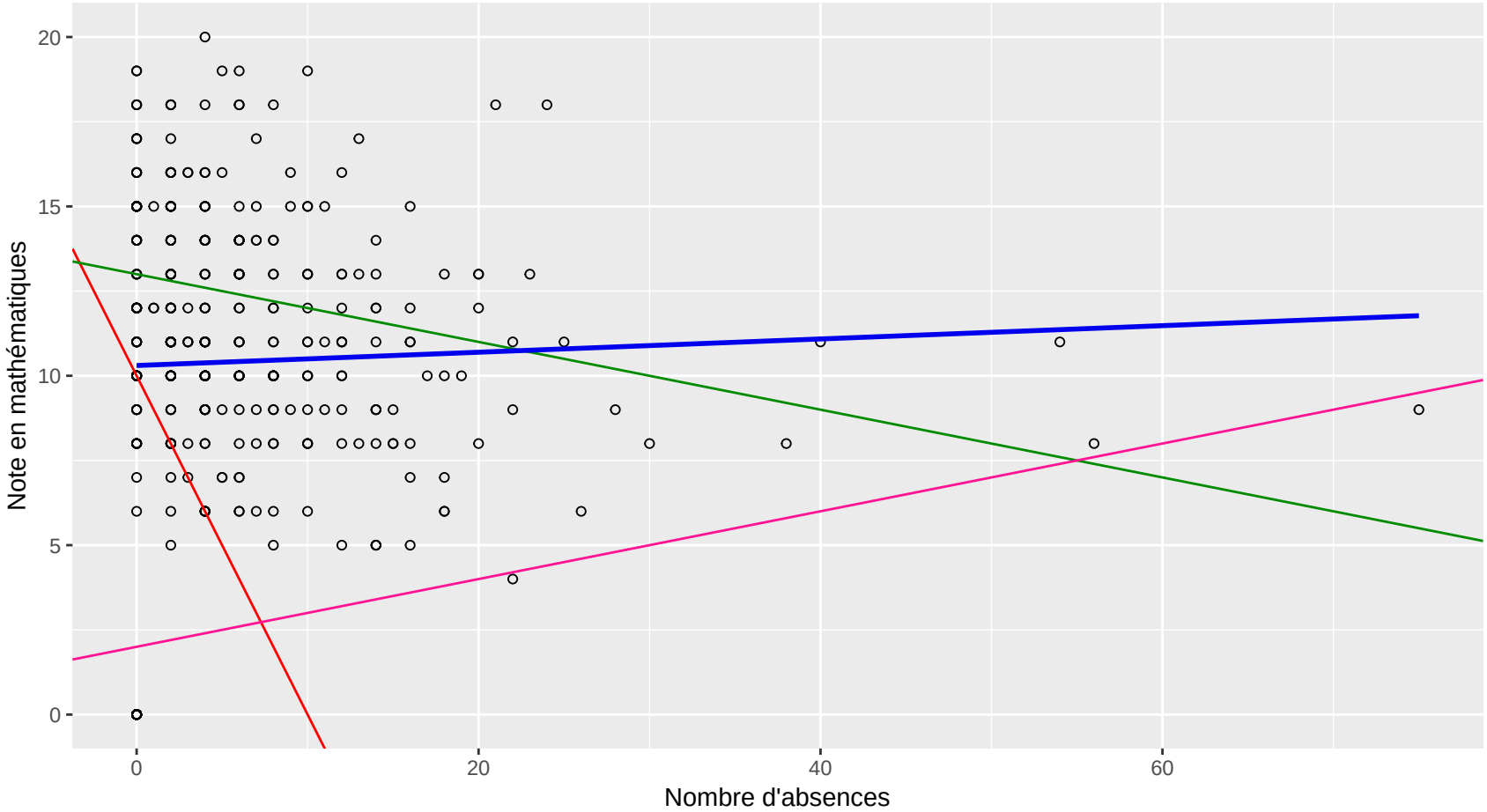
Synthétiser le nuage de points

Problème : comment synthétiser le nuage de points ci-dessous en une droite qui modélise le mieux la relation entre nos deux variables ?



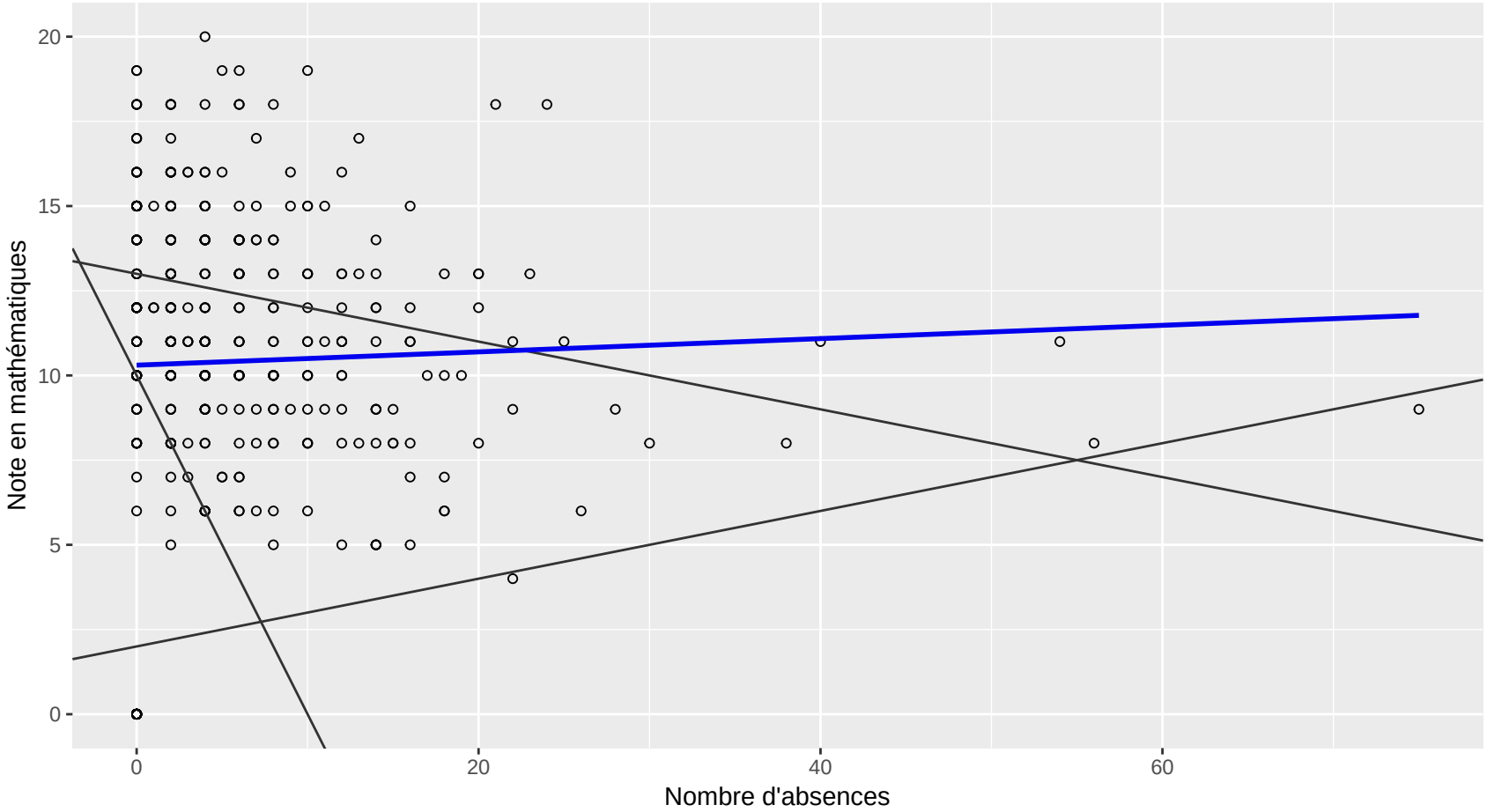
Intuition géométrique (1/3)

Quelle droite modélise le mieux le nuage de points ?



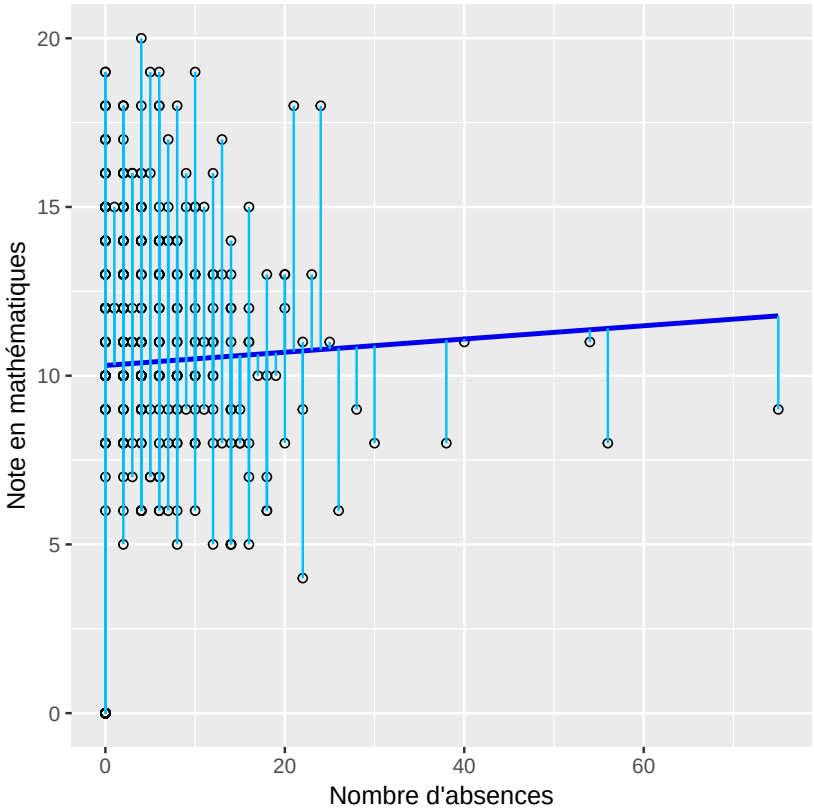
Intuition géométrique (2/3)

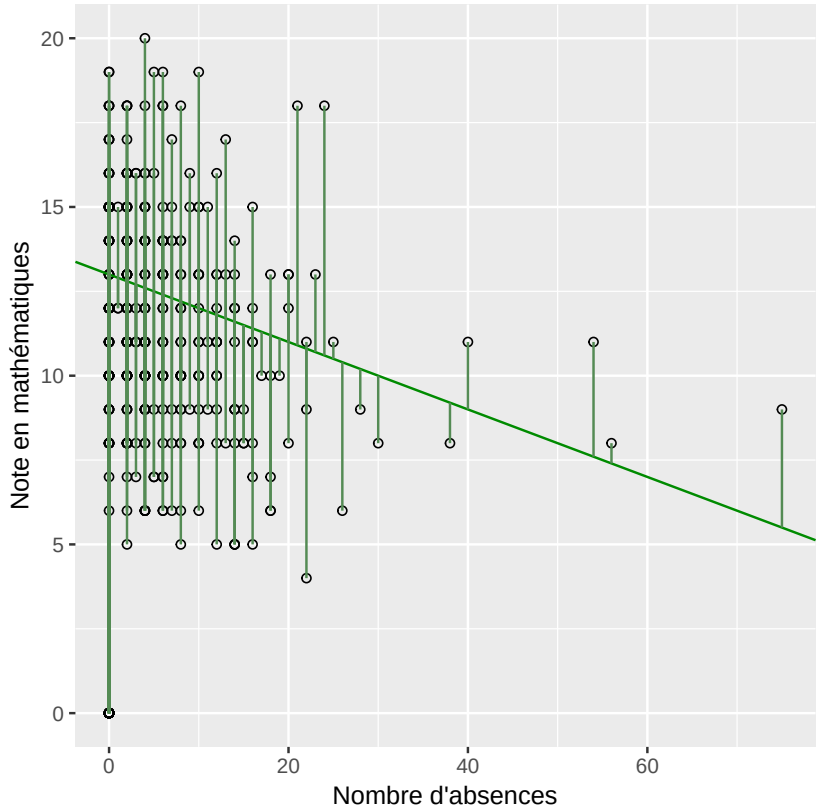
La droite bleue est la meilleure. Pourquoi ?



Intuition géométrique (3/3)

Approximativement, c'est celle qui *minimise la distance verticale* entre la droite et tous les points. En additionnant les barres verticales bleues, on trouverait une plus faible valeur que pour les barres verticales vertes.





La régression linéaire simple

Formalisation

Soient deux variables X_1 (resp. Y), comptant chacune n observations x_i^1 (resp. y_i). **Notre modèle de régression linéaire propose d'estimer (ou de prédire) Y grâce à X_1** tel que :

$$\hat{Y} = \beta_0 + \beta_1 X_1$$

où $\hat{Y}$ est notre estimation (ou prédiction) pour Y.

En pratique, **même le meilleur modèle ne sera pas parfait** : il y aura toujours des erreurs telles que

$$Y = \hat{Y} + \epsilon = \beta_0 + \beta_1 X_1 + \epsilon$$

où ϵ est **l'erreur** (ou le **résidu**), qui mesure l'écart entre les données observées Y et prédites $\hat{Y}$. Évidemment, le meilleur modèle est celui qui minimise les erreurs ϵ.

Les Moindres Carrés Ordinaires

Le critère pour déterminer le **meilleur** modèle avec ces variables sera de trouver les coefficients β_0 et β_1 grâce à la méthode des **moindres carrés ordinaires (MCO)**.

On veut trouver les **coefficients** β_0 **et** β_1 tels que :

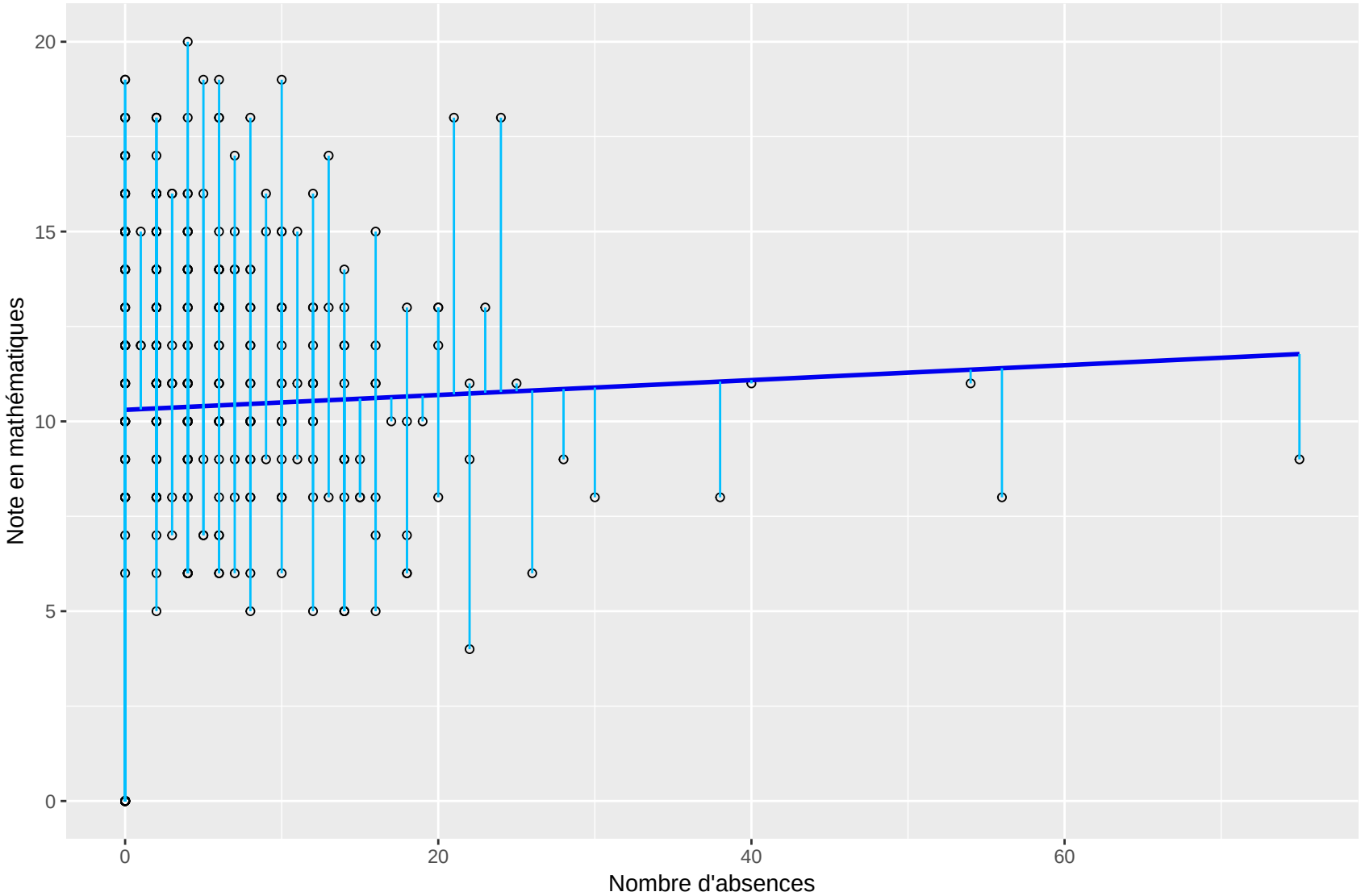
$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

c'est-à-dire tels qu'ils **minimisent la somme des carrés des distances entre les valeurs** y_i **observées et les valeurs** $\hat{y}_i$ **prédites**, ou encore qu'ils minimisent les erreurs ϵ_i.

Les distances $y_i - \hat{y}_i$ sont représentées par les barres verticales bleues ci-contre.

Il existe deux méthodes pour trouver β_0 et β_1, une mathématique et l'autre algorithmique, qui sortent du cadre de ce cours.

Les Moindres Carrés Ordinaires



Interprétation graphique des résultats

Revenons à notre exemple : **nous voulions estimer la note de mathématiques de lycéens à partir de leur nombre d'absences, grâce à une régression linéaire simple**.

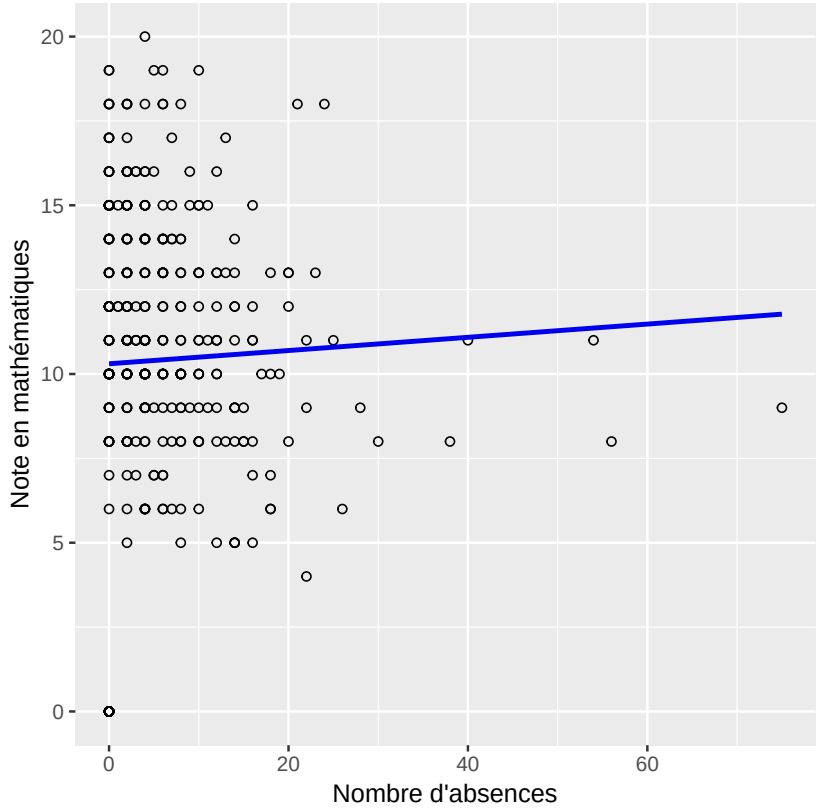
La méthode des MCO nous donne les résultats suivants :

$$\beta_0 = 10,3 \text{ et } \beta_1 = 0,02$$

D'où l'équation suivante :

$$\hat{Y} = 10,3 + 0,02 \times X$$

Comment interpréter les coefficients β_0 et β_1 ?

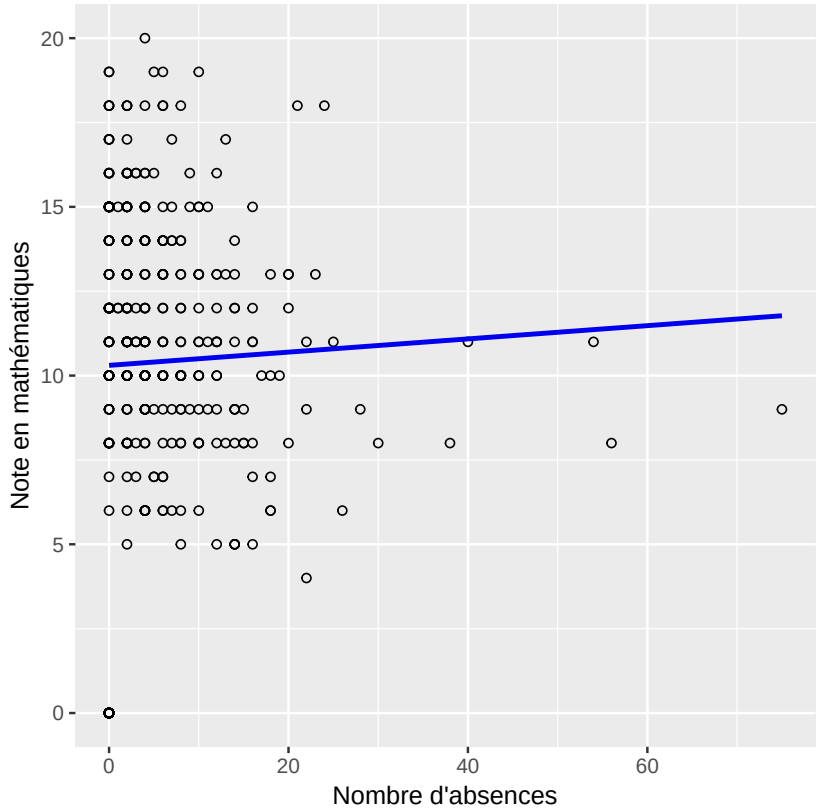


Interprétation des coefficients

Comment interpréter les coefficients β_0 et β_1 ?

- **Les élèves qui ne sont jamais absents ont en moyenne $10,3$ en maths**
- **Chaque jour d'absence supplémentaire augmente en moyenne la note en maths de $0,02$ points**

L'absence a donc un effet quasiment nul sur les résultats en mathématiques, voire très légèrement positif.



Exercice de lecture

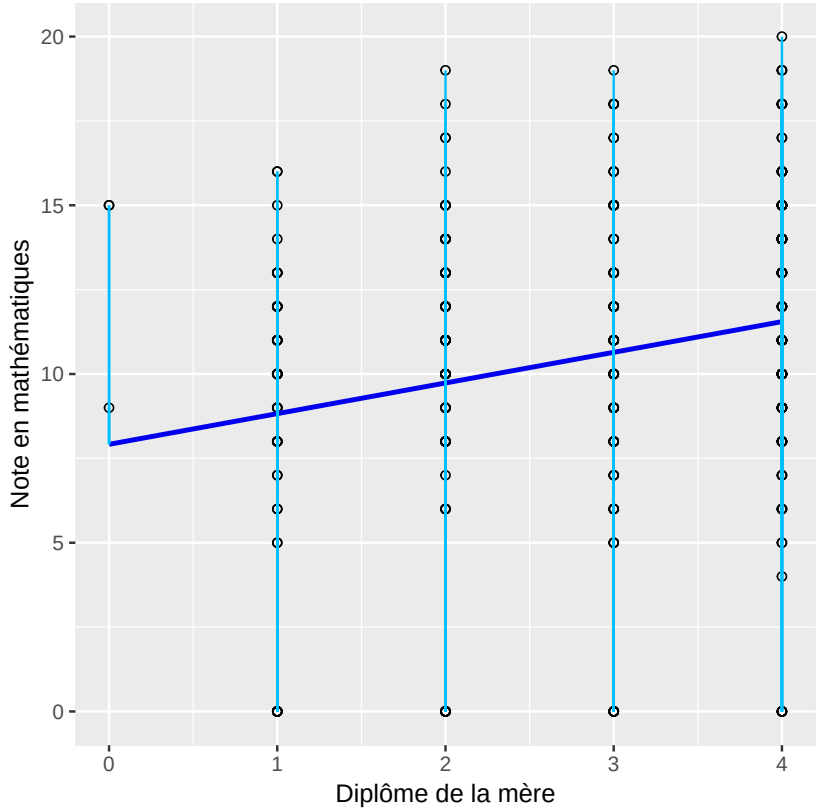
Après avoir rencontré peu de succès avec les absences, **nous essayons d'estimer la note en mathématiques à partir du niveau d'études de la mère** (0 : aucun, 1 : primaire, 2 : collège, 3 : lycée, 4 : supérieur).

On trouve les valeurs suivantes :

$$\beta_0 = 7,9 \ , \ \beta_1 = 0,9$$

Exercices :

- Écrivez l'équation de la droite (bleu foncé)
- Interprétez les coefficients, à la fois sur le graphique et en rapport avec les variables
- À quoi correspondent les segments bleu clair ?



Goodness of fit et R^2

Nous voulons maintenant **déterminer si notre droite est correctement ajustée à notre nuage de points** (*goodness of fit*). En particulier, nous voudrions savoir si la variance de Y provient davantage de notre modèle explicatif $\hat{Y}$ ou des erreurs ϵ du modèle.

On calcule l'indicateur R^2 de la sorte :

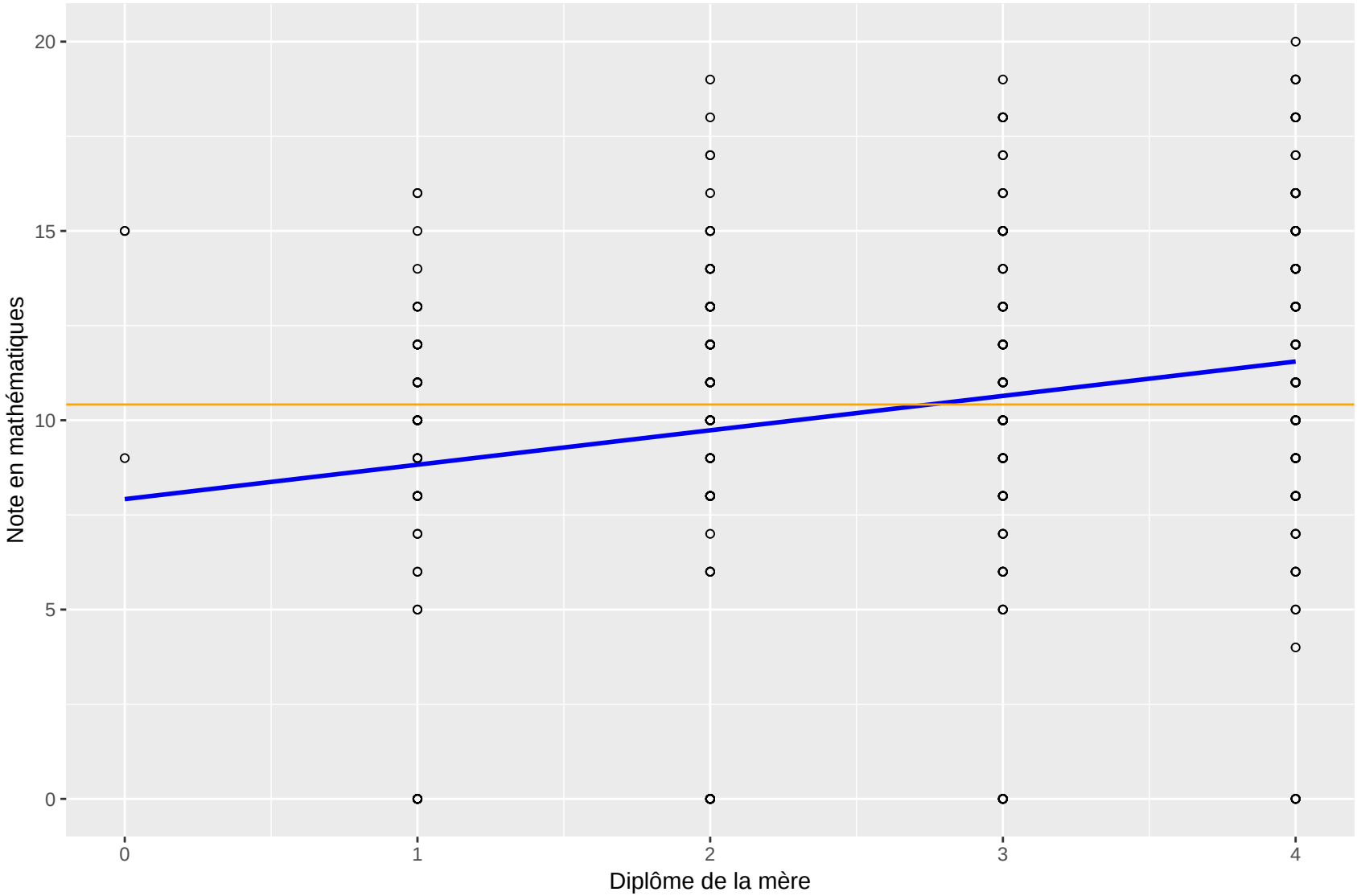
$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

où $\bar{y}$ est la moyenne de Y (en jaune ci-contre).

Un R^2 de 1 signifierait que la droite serait parfaitement ajustée (points répartis sur une droite), un R^2 de 0 que la droite est totalement inajustée (points répartis aléatoirement ou selon un modèle non-linéaire).

Le R^2 ne garantit pas la qualité du modèle ou sa pertinence, il ne fait qu'évaluer son ajustement.

Goodness of fit et R^2

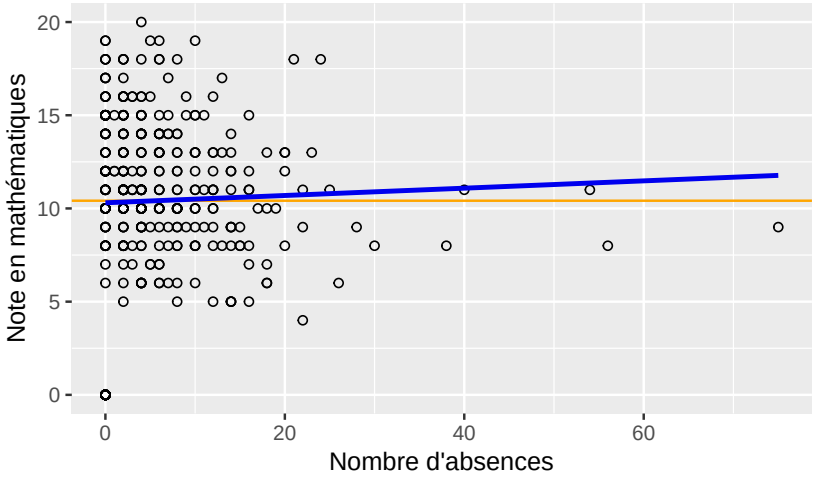


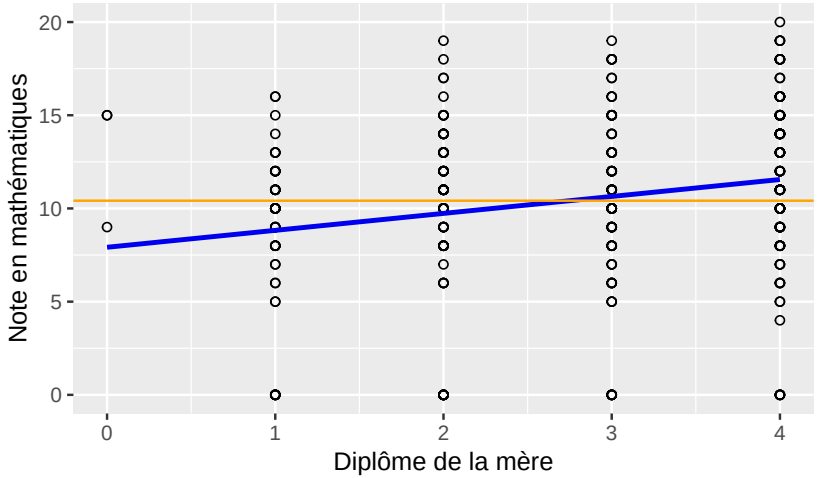
Exercice de lecture

Interprétez les R^2 suivants :

- Note en mathématiques & absences : $R^2 = 0,001$
- Note en mathématiques & diplôme de la mère : $R^2 = 0,05$

Ces résultats sont-ils étonnants (comparer à la répartition des nuages de points ci-dessous) ?





Différence entre régression et corrélation

On peut se poser la question : en quoi la régression est-elle différente d'un coefficient de corrélation r ?

- Point commun : ni le coefficient de corrélation ni la régression ne répondent directement à la question de la causalité
- Différence 1 : r sert à évaluer la force du lien entre deux variables, la régression fournit un modèle qui permet d'évaluer Y à partir de X.
- Différence 2 : pour le coefficient de corrélation, peu importe l'ordre des variables. Si X et Y sont interchangés, r conserve la même valeur. À l'inverse, une régression a une variable **dépendante** (*à expliquer*) et une variable **indépendante** (*explicative*), et les coefficients β seraient différents si les variables étaient inversées.

Régression linéaire multiple

La régression linéaire multiple : 2 variables indépendantes

Motivations

En pratique, il serait très étonnant qu'un phénomène puisse être expliqué uniquement par une variable. Nous voulons au contraire avoir **un modèle qui prédise Y avec au moins deux variables X_1 et X_2**. Cela nous permettrait de :

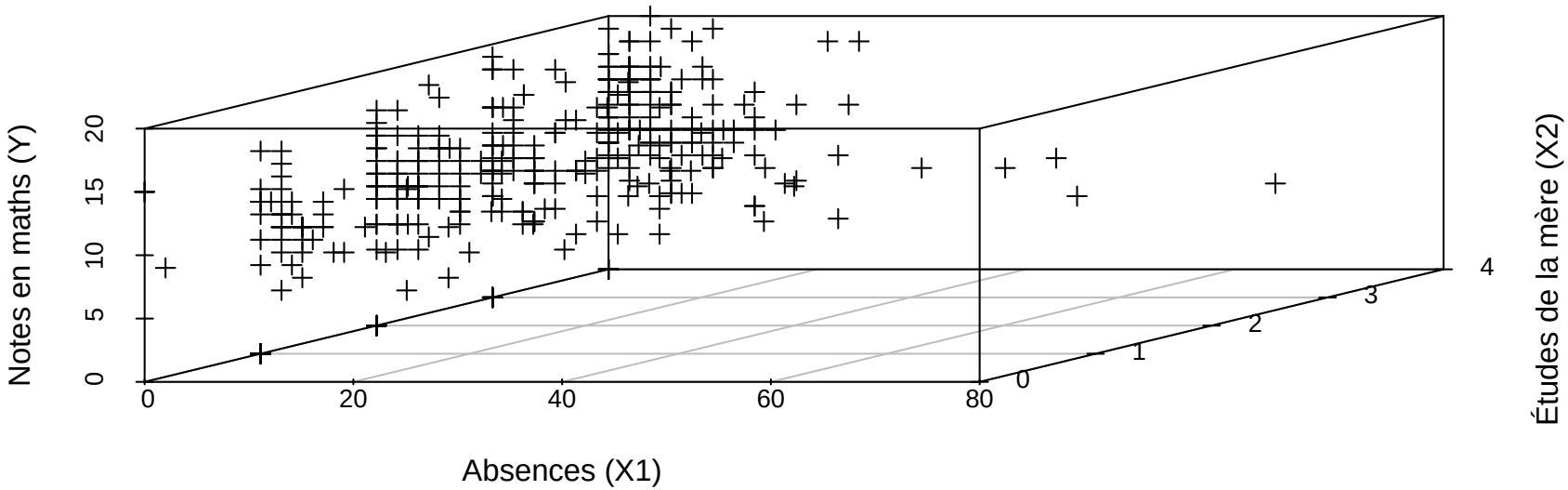
- **améliorer notre modèle** en le complexifiant, nous espérons ainsi augmenter la part de la variance expliquée (la qualité de son ajustement évalueé par R^2)
- **comparer** les influences respectives de nos variables indépendantes (explicatives)
- **contrôler** les effets de variable cachée, en intégrant à notre modèle au moins une variable supplémentaire (*endogénéité*).

Dans notre exemple, imaginons que nous souhaitions expliquer les notes en mathématiques (Y) par les absences (X_1) et par le diplôme de la mère (X_2).

Nuage de points pour trois variables

On peut représenter notre nuage de points dans un espace à 3 dimensions (Y, X_1, X_2), contrairement à notre plan pour deux variables précédemment (Y, X_1).

Intuition : en deux dimensions, la régression nous donnait l'équation d'une droite. Qu'en sera-t-il en trois dimensions ?



Formalisation

Soit une variable Y (respectivement X_1, X_2), avec n observations y_i (resp. x_i^1, x_i^2).

Notre modèle de régression linéaire multiple permet d'estimer notre variable dépendante Y grâce aux deux variables indépendantes X_1, X_2, tel que :

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \text{ , soit}$$

$$\hat{Y} = \beta_0 + \sum_{i=1}^{2} \beta_i X_i$$

décrivant ainsi un **plan de régression** (voir slide suivante) et non plus une droite.

Formalisation

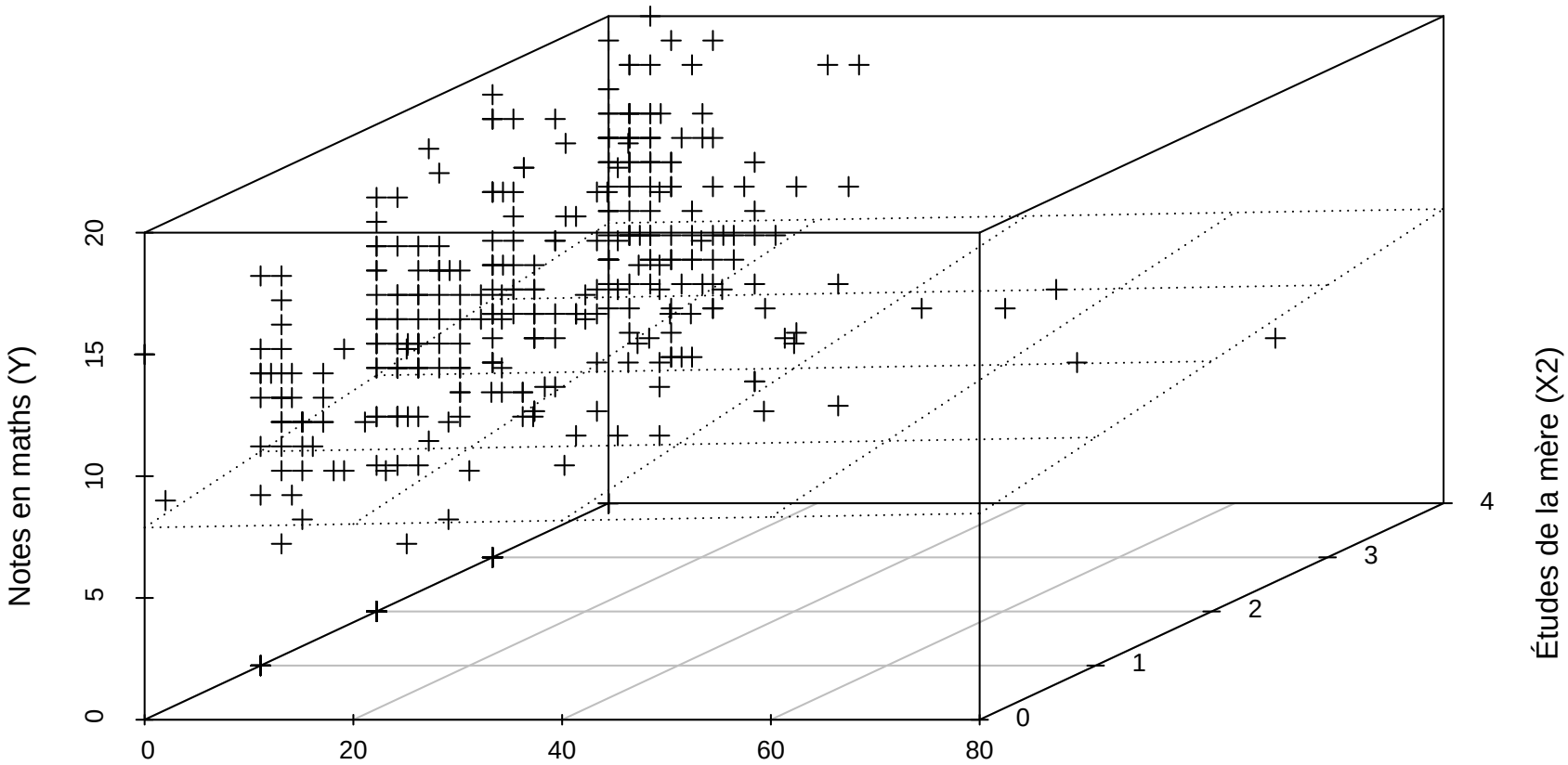
Notons que :

- notre prédiction n'est jamais parfaite : $Y = \hat{Y} + \epsilon$
- le meilleur modèle pour ces variables est toujours déterminé par la méthode des MCO
- les X_i peuvent être continues, discrètes ou catégorielles (moyennant un recodage)

Plan de régression

Dans notre exemple, l'équation du plan de régression est la suivante :

$$\hat{Y} = 7,9 + 0,007X_1 + 0,9X_2$$

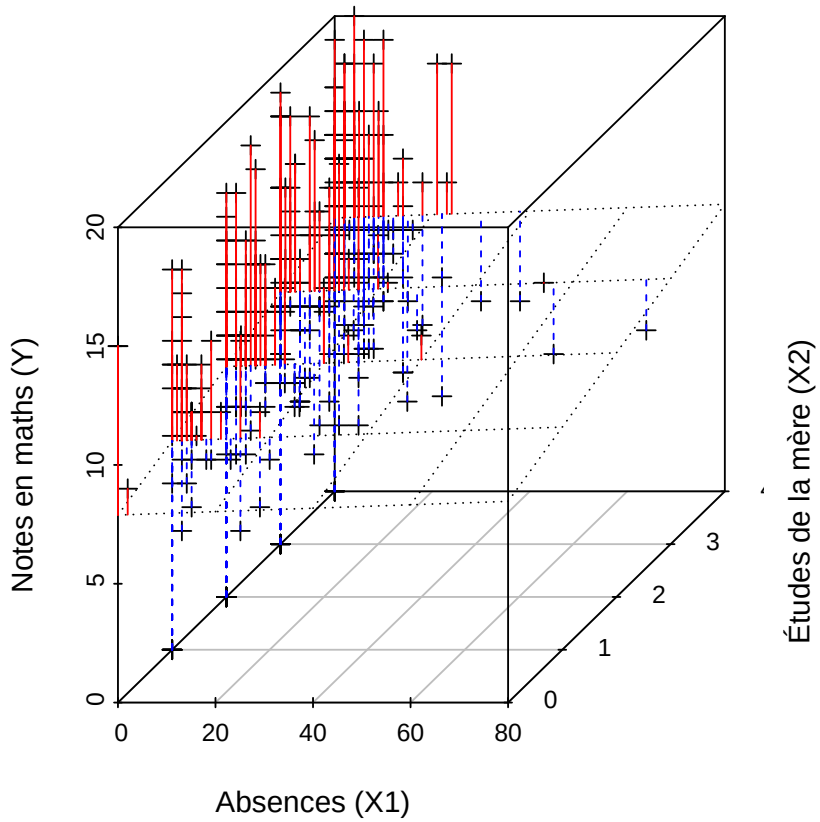


Représentation des MCOs en 3 dimensions

On étend simplement notre optimisation à β_2 :

$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

On a représenté ci-contre les distances $y_i - \hat{y}_i$ par des segments rouges lorsque la distance était positive et bleus quand elle était négative.



Interprétation des coefficients et du modèle

Les MCO nous donnent les coefficients suivants :

$$\beta_0 = 7,9 \ , \ \beta_1 = 0,007 \ , \ \beta_2 = 0,9 \ , \ R^2 = 0,05$$

Ce qui signifie :

- En moyenne, les élèves qui ne sont jamais absents et dont la mère n'a aucun niveau d'études ont $7,9$ en maths.
- ***Toutes variables du modèle tenues égales par ailleurs*, une journée d'absence supplémentaire se traduit en moyenne par une augmentation de $0,007$ de la note en maths.**
- ***Toutes variables du modèle tenues égales par ailleurs*, un niveau d'étude supplémentaire de la mère se traduit en moyenne par une augmentation de $0,9$ de la note en maths.**

Interprétation des coefficients et du modèle

Dans le cadre d'un raisonnement *toutes choses égales par ailleurs* (ce qui est, techniquement, un abus de langage), on parle parfois d'identifier les **effets purs** d'une variable : c'est l'effet d'une variable quand toutes les autres variables sont **contrôlées**, c'est-à-dire qu'aucune autre variable ne varie.

Nota Bene : pour l'instant, nous sommes toujours dans le cadre d'un modèle descriptif !

Que penser de l'ajustement R^2 ?

Comment identifier les variables significatives ?

Si nous voulons sélectionner les variables explicatives significatives d'un modèle de régression, **il existe un test statistique (appelé F) qui permet de rejetter l'hypothèse nulle grâce à une p-value**.

Les détails de ce test ne seront pas abordés dans le cours. *Grosso modo*, il s'agit d'évaluer la contribution d'une variable à l'explication de la variance de Y en la pondérant par les degrés de liberté du système (la « taille » de notre problème).

Traditionnellement, les modèles de régression donnent leur significativité avec des astérisques. Habituellement :

- $p < 0.001$: ***
- $p < 0.005$: **
- $p < 0.01$: *

Exercice de lecture

	G3
absences	0.007
	(0.028)
Medu	0.903***
	(0.207)
Constant	7.890***
	(0.619)
N	395
R^2	0.047
Adjusted R^2	0.042
Residual Std. Error	4.483 (df = 392)
F Statistic	9.733*** (df = 2 ; 392)

Notes : ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

La régression linéaire multiple : n variables indépendantes

Exemple

Imaginons maintenant que nous souhaitions contrôler davantage de variables dans notre modèle.

Ainsi, on pourrait penser que **les études du père ont elles aussi une influence sur les résultats en mathématiques**, ou bien que l'influence des études de la mère sur les notes en maths diffèrent selon que l'élève est une fille ou un garçon. Bref, on voudrait pouvoir raisonner « tous choses égales par ailleurs » avec davantage de variables.

En plus des absences et du niveau d'études de la mère, nous allons donc contrôler le niveau d'études du père, l'âge et le sexe (5 variables **indépendantes** ou *explicatives* X_i, qualitatives ou quantitatives) pour prédire les notes en mathématiques (1 variable quantitative **dépendante** ou *à expliquer*).

Formalisation

Notre modèle de régression linéaire multiple permet d'estimer notre variable dépendante Y grâce aux cinq variables indépendantes $X_1, X_2, ..., X_5$, tel que :

$$\hat{Y} = \beta_0 + \sum_{i=1}^{5} \beta_i X_i$$

Avec de nombreuses variables indépendantes, on adopte souvent une notation matricielle, et l'équation s'écrit :

$$\hat{Y} = \beta X$$

d'où

$$Y = \hat{Y} + \epsilon = \beta X + \epsilon$$

Formalisation

La méthode des MCO est toujours la même. Les coefficients β se trouvent en résolvant :

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Exercice de lecture

- Quelles sont les variables significatives ?
- Comment interpréter les coefficients ?
- Comment comprendre notre R^2 ?

Exercice de lecture

	Note en mathématiques (T3)
	G3
absences	0.027
	(0.029)
Medu	0.717^{***}
	(0.265)
Fedu	0.081
	(0.264)
age	-0.488^{***}
	(0.182)
sexM	0.812^{*}
	(0.450)
Constant	15.856^{***}
	(3.196)
Observations	395
R^2	0.073
Adjusted R^2	0.061

Interprétation des modèles

Précautions d'interprétation

Conditions des MCO

Les MCO nécessitent de valider un certain nombre d'hypothèses :

1. Linéarité de la modélisation
2. Absence d'autocorrélation des variables explicatives (X_1 et X_2 ne doivent pas être "trop" corrélés linéairement, sinon il faut abandonner une variable ou utiliser une méthode légèrement différente)
3. Homéoscédasticité : les résidus sont répartis de manière homogène
4. Absence d'autocorrélation des résidus
5. Normalité des résidus
6. Absence de corrélation des variables explicatives et du résidu (dans le modèle théorique)

On va revenir surtout sur les points 2 et 3.

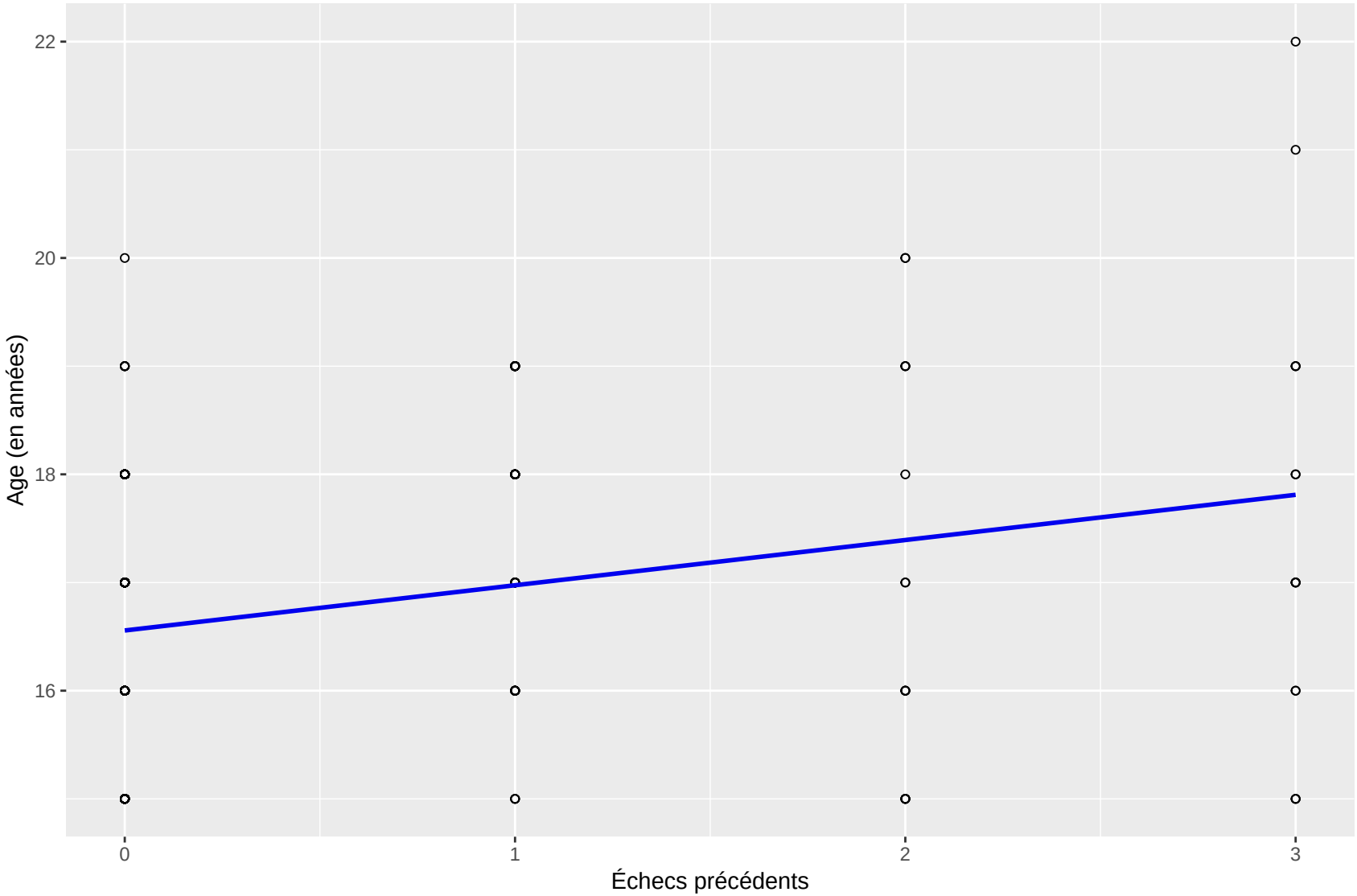
Autocorrélation des variables explicatives

Pour que le modèle soit valide, **les variables explicatives ne doivent pas être significativement corrélées linéairement entre elles**. Si c'est le cas, il faut soit abandonner certaines variables, soit utiliser une variation de la méthode des MCO.

Dans notre exemple, l'âge et le nombre de fois qu'un élève a "échoué" en maths les années précédentes (moins de 10 de moyenne) sont linéairement corrélés, car les élèves ayant le plus échoué sont aussi souvent les plus vieux (probablement à cause des redoublements).

En faisant une régression sur les échecs à partir de l'âge, on trouve en effet un coefficient de 0,14 ($p \approx 0$). On pourrait ainsi se passer de la variable "échec" et se contenter de la variable "âge".

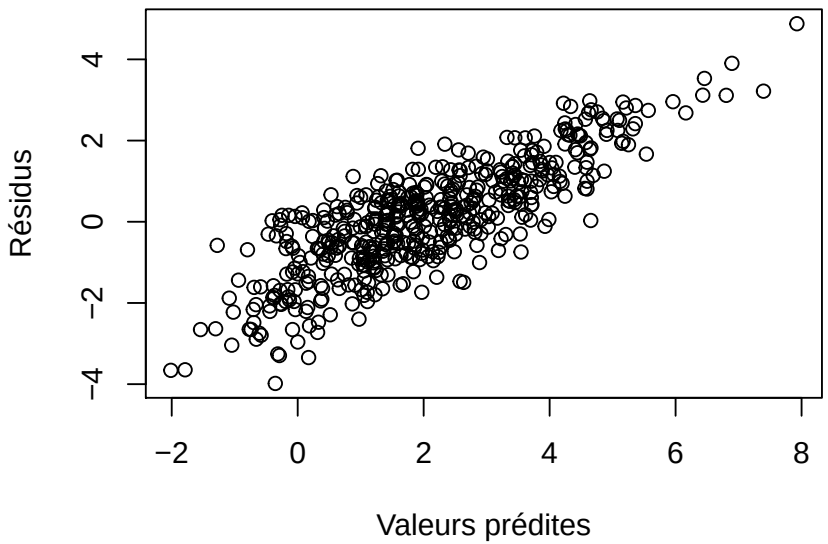
Autocorrélation des variables explicatives

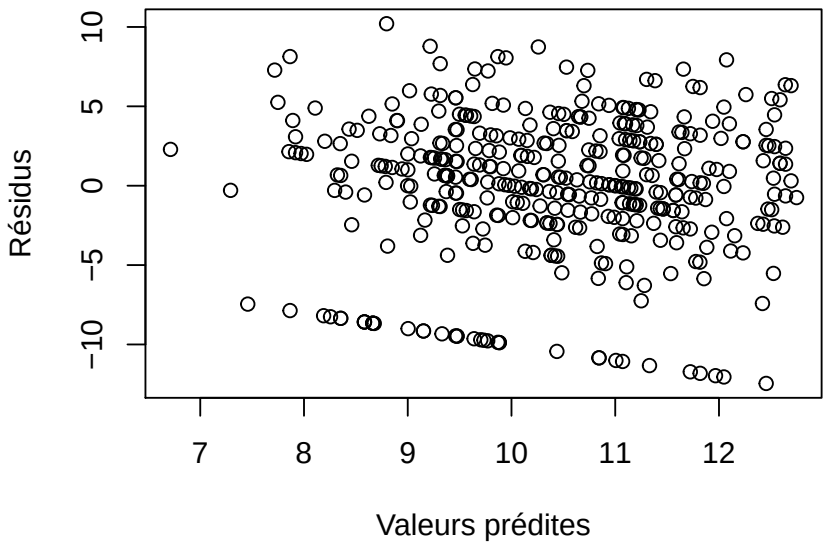


Hétéroscédasticité

Pour que les MCO fonctionnent correctement, **il faut que la variance des résidus soit homogène, et ne soit pas plus élevée pour une sous-population en particulier**.

À gauche, un cas d'hétéroscédasticité ; à comparer avec les résidus de la régression précédente à droite.





Endogénéité inobservée

L'endogénéité inobservée recoupe le problème dit « des variables cachées ».

Cela signifie que le modèle présenté pourrait ne pas refléter correctement la "réalité", s'il n'intègre pas une (ou plusieurs) variable(s) capitale(s) qui influencent simultanément les autres variables du modèle.

Autrement dit, nous pourrions sélectionner une variable comme significative à tort, n'ayant pas pu contrôler l'impact d'une autre variable importante.

Par exemple, comparons les conclusions que nous pouvons tirer de ces deux modèles :

Endogénéité inobservée : Avec le diplôme du père

	Note en mathématiques (T3)
	G3
Fedu	0.642^{***}
	(0.210)
Constant	8.797^{***}
	(0.576)
Observations	395
R^2	0.023
Adjusted R^2	0.021
Residual Std. Error	4.534 (df = 393)
F Statistic	9.352^{***} (df = 1 ; 393)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Endogénéité inobservée : Avec le diplôme des deux parents

	Note en mathématiques (T3)
	G3
Medu	0.836***
	(0.264)
Fedu	0.118
	(0.265)
Constant	7.821***
	(0.648)
Observations	395
R^2	0.048
Adjusted R^2	0.043
Residual Std. Error	4.482 (df = 392)
F Statistic	9.802*** (df = 2 ; 392)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Sélection de modèle

Principes de la sélection

Concrètement, **pour trouver le meilleur modèle de régression linéaire, on teste différents modèles avec des variables différentes** (ajout de nouvelles variables, création de nouvelles variables à partir des variables déjà connues, etc.).

Il faut ensuite choisir le meilleur modèle en gardant l'équilibre entre deux principes :

- le modèle **le mieux ajusté** est, en un certain sens, le meilleur (mesuré par R^2)
- le modèle **le plus simple** est, en un autre sens (rasoir d'Ockham), le meilleur

Autrement dit, **il faut éviter de prendre un modèle trop simple** (problèmes d'endogénéité inobservée, faible ajustement) **ou un modèle trop ajusté** (*overfitting*) qui sacrifierait la simplicité.

Dans notre exemple suivi, nous allons essayer différents modèles pour essayer de retenir le meilleur.

Modèle "Medu + Age + Sexe"

	Note en mathématiques (T3)
	G3
Medu	0.793^{***}
	(0.207)
age	-0.460^{***}
	(0.177)
sexM	0.779^{*}
	(0.448)
Constant	15.547^{***}
	(3.122)
Observations	395
R^2	0.071
Adjusted R^2	0.064
Residual Std. Error	4.434 (df = 391)
F Statistic	9.908^{***} (df = 3 ; 391)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Modèle "T1"

	Note en mathématiques (T3)
	G3
G1	1.106^{***}
	(0.042)
Constant	-1.653^{***}
	(0.475)
Observations	395
R^2	0.642
Adjusted R^2	0.641
Residual Std. Error	2.743 (df = 393)
F Statistic	705.842^{***} (df = 1 ; 393)
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Modèle "T1 + Medu + Age + Sexe"

	Note en mathématiques (T3)
	G3
G1	1.083^{***}
	(0.042)
Medu	0.154
	(0.129)
age	-0.375^{***}
	(0.108)
sexM	0.234
	(0.274)
Constant	4.328^{**}
	(1.950)
Observations	395
R^2	0.657
Adjusted R^2	0.653
Residual Std. Error	2.699 (df = 390)
F Statistic	186.402^{***} (df = 4 ; 390)

Synthèse et sélection

Modèle	R^2
Absences	0,001
Études de la mère	0,05
Absences + Études de la mère	0,05
Études de la mère + Age + Sexe	0,06
Absences + Études de la mère + Études du père + Age + Sexe	0,07
T1	0,64
T1 + Études de la mère + Age + Sexe	0,65

- Quels modèles sont peu intéressants scientifiquement ? Quels sont les modèles les moins ajustés ?
- Que penser de l'emploi de la variable T1 ?
- **Quel modèle retenir ?**

Remarques générales autour de la régression

Régression vers la moyenne

Francis Galton est considéré comme le père de la régression linéaire (1886). Il en a fait le premier usage "moderne" et lui a donné son nom actuel, même si la méthode avait déjà été utilisée auparavant, notamment par Laplace au XVIIIème siècle.

Cousin de Darwin, fondateur de l'eugénisme, Galton avait notamment employé cette méthode pour décrire le phénomène de « régression vers la moyenne », qu'il décrivait comme un phénomène de régression vers la « médiocrité ». Ainsi, la taille des enfants de deux parents de grande taille était plus proche de la moyenne de la population.

Ce phénomène est utile pour interpréter des séries temporelles : **lorsqu'on mesure une variable (chômage, rendement d'une entreprise, etc.) une première fois et que la valeur s'avère être extrêmement faible ou élevée, il faut s'attendre à ce que la seconde mesure soit plus proche de la moyenne**.

Interprétation de la régression vers la moyenne

Daniel Kahneman ("prix Nobel" d'économie en 2002) donne l'exemple suivant comme une mauvaise interprétation de la régression vers la moyenne.

Dans une base aérienne militaire en Israël, Kahneman discute avec l'instructeur de vol qui lui tient ces propos :

> Lorsque je félicite un cadet après une bonne manœuvre, ses performances sont en général plus mauvaises la fois suivante. Lorsque je punis un cadet après une mauvaise manœuvre, la suivante est presque toujours meilleure. La punition est dont plus efficace que la récompense.

Comment la régression vers la moyenne invalide-t-elle cette conclusion ?

Critiques du raisonnement "toutes choses égales par ailleurs"

Les méthodes de régression (y compris linéaire) sont très souvent utilisées en économétrie, ainsi que dans la sociologie anglo-saxonne. Cette méthode fait davantage polémique en France. Selon une citation que François Simiand attribuait à Maurice Halbwachs, **« cette méthode conduit à étudier et comparer les comportements d'un renne au Sahara et d'un chameau au pôle Nord »** [Desrosières, 2001].

Cette critique souligne le fait que, dans le monde social, les choses sont rarement égales par ailleurs. Si les variables explicatives d'une régression intègrent la profession et le revenu, le modèle suppose qu'un ouvrier puisse gagner 10 000€ ou qu'un cadre gagne le salaire minimum. Autre exemple, si les variables explicatives intègrent le revenu et le niveau de responsabilité professionnelle des femmes (pour raisonner « à poste égal »), le modèle suppose que les hommes et les femmes ont potentiellement autant de chances d'avoir des postes à haute responsabilité, ce qui ne décrit nullement la réalité sociale.

Le modèle linéaire généralisé : la régression logistique

Principes de la régression logistique

Généralisation du modèle linéaire

Le modèle de régression linéaire fonctionne sous un certain nombre de conditions que l'on a énuméré (homeoscedasticité des résidus, absence d'autocorrélation, etc.), ainsi que des conditions sur la nature des variables inclues :

- Les variables indépendantes sont quantitatives ou dichotomiques
- La variable dépendante est continue et illimitée

Généralisation du modèle linéaire

On peut pourtant vouloir prédire une variable dépendante qui ne répond pas à ces conditions. C'est en particulier le cas lorsque l'on souhaite prédire une variable dichotomique (qui peut prendre deux valeurs). Dans ce cas, ce que l'on cherche à modéliser n'est pas la valeur d'une variable quantitative, mais la probabilité d'une catégorie.

Une première possibilité est de considérer la variable dichotomique à prédire comme une variable numérique valant 0 ou 1, et de prédire sa valeur par une régression linéaire simple ou multiple (**modèle de probabilité linéaire**).

Cette solution pose plusieurs problèmes, parmi lesquelles :

- la valeur prédite va très souvent dépasser l'intervalle [0, 1], et n'aura donc aucun sens en tant que probabilité ;
- la relation entre la variable dépendante et les variables indépendantes n'est pas linéaire.

Solution, étape 1 : rapport des probabilités

Comment, alors, produire un modèle pour prédire une variable dichotomique ?

Le premier problème est celui de l'étendu de la variable à prédire. Une probabilité est contenue dans l'intervalle [0, 1]. Cependant, dans le cas d'une variable dichotomique, on peut facilement calculer le rapport des probabilités.

Soit la variable dépendante Y. On cherche à modéliser la probabilité $P(Y = 1)$. Le rapport des probabilités est alors égal à :

$$\frac{P(Y = 1)}{1 - P(Y = 1)}$$

Ce rapport est compris entre dans $[0, +\infty]$.

Solution, étape 2 : transformation logarithmique

Reste que le rapport des probabilités est nécessairement positif, et est donc une variable finie, contrairement à ce que demande le modèle OLS. Une façon habituelle, en statistique, de contourner ce problème est de faire subir une transformation logarithmique à une variable. Le logarithme néperien d'une variable définie sur $]0, +\infty]$ est en effet défini sur $[-\infty, +\infty]$.

On qualifie de "logit" de Y le logarithme népérien du rapport des probabilités de Y.

$$ln(\frac{P(Y=1)}{1-P(Y=1)})$$

Le modèle logit

Le modèle logit est la forme la plus courante de régression logistique. Elle consiste à estimer le logit de Y sous la forme d'une équation affine de k variables indépendantes.

$$logit(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

On retrouve ici un modèle linéaire, si ce n'est que la variable dépendante a été transformée.

Logit, rapport de probabilité et probabilités

À partir de là, on peut retrouver la définition de la probabilité.

$$logit(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

$$odds(Y = 1) = e^{ln(odds(Y=1))} = e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}$$

$$P(Y = 1) = \frac{e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}{1 + e^{\alpha+\beta_1 X_1+\beta_2 X_2+...+\beta_k X_k}}$$

Estimation des paramètres

Comme dans le cas de la régression linéaire, il s'agit d'estimer les paramètres $\beta_1, \beta_2, ..., \beta_k$. Contrairement à la régression linéaire, on emploie pas la méthode des MCO dans un modèle logit, mais des méthodes dites d'optimisation du maximum de vraisemblance.

En somme, il s'agit de trouver l'ensemble de paramètres $\beta_1, \beta_2, ..., \beta_k$ qui maximisent la log-vraisemblance du modèle (qui la probabilité des données observée pour un ensemble de paramètres). Les détails de ce calculs peuvent être ignorés.

Situation de référence

Les variables indépendantes peuvent être catégorielles ou continue. Dans le premier cas, l'une des modalités de la variable est considérée comme la modalité de référence. Son coefficient β est donc fixé à zero. Mesurer l'effet de cette variable sur le logit revient alors à considérer l'effet d'une modification de cette modalité.

Lecture des paramètres

Les paramètres β sont des estimations qu'il faut donc, en toute rigueur, noter $\hat{\beta}$. Il s'agit des coefficients estimés par le modèle. Comme il s'agit d'estimations, on ne peut pas avoir de certitude que les β véritables sont biens égaux aux $\hat{\beta}$.

On calcule donc l'intervalle de confiance de $\hat{\beta}$:

$$[\hat{\beta} - 1.96 * \hat{\sigma_{\hat{\beta}}}; \hat{\beta} + 1.96 * \hat{\sigma_{\hat{\beta}}}]$$

.

On réalise par ailleurs un test de Student sur chacun des $\hat{\beta}$, par lequel on teste l'hypothèse nulle $\hat{\beta} \approx 0$.

Lecture des paramètres

Les paramètres désignent constituent l'effet d'une modalité sur le logit. Ils sont difficiles à interpréter, car, contrairement à la régression linéaire, il ne s'agit pas d'un effet direct sur la probabilité, mais sur le logit.

Tout ce que l'on peut dire, c'est que :

- si le coefficient est significativement différent de zéro, alors la variable produit un effet sur la variable dépendante
- si le coefficient est négatif, alors cet effet est négatif (avoir cette modalité plutôt que la modalité de référence décroit la probabilité de Y)
- si le coefficient est positif, alors cet effet est positif (avoir cette modalité plutôt que la modalité de référence positif la probabilité de Y)

Odds ratio

Pour faciliter la lecture des coefficiens, on les transforme souvent en odds-ratio (OR). Il suffit pour cela de prendre l'exponentielle du coefficient.

L'odds-ratio décrit le rapport des probabilités. Il se lit ainsi : toutes variables du modèle tenues égales par ailleurs, le fait d'avoir la modalité test plutôt que la modalité de référence multiplie par OR la probabilité de Y.

Estimation de la robustesse du modèle

Il n'y a pas de R^2 pour les régressions logistiques.

À la place, de nombreux statisticiens ont proposés des manières de calculer des pseudo-R^2. Tous ont en commun de comparer la variance totale et la variance résiduelle, la différence entre les deux étant la part de la variance expliquée par le modèle.

On emploie parfois le Akaike Information Criterion, basé sur la vraisemblance, pour comparer des modèles logit entre eux.

Lecture et interprétation d'une régression logistique

Une modélisation de variable dichotomique

En pratique, un modèle logit ne peut être employé que pour modéliser **une variable dichotomique**. Si l'on peut l'employer pour des variables catégorielles, il faut cependant que les catégories ne soient que deux.

On dispose d'archives de procès et l'on souhaite modéliser la peine reçue par l'accusé en fonction de variables sociodémographiques (par exemple pour mettre en évidence un biais envers certains accusés). La variable dépendante, la peine, contient trois modalités : "acquitement", "peine inférieure aux réquisitions", "peine supérieure aux réquisitions".

Une modélisation de variable dichotomique

On ne peut pas produire de modèle logit pour prédire ces trois modalités ensembles. Une solution est de dichotomiser la variable

- On commence par rassembler les deux dernières modalités, de sorte que l'on a une variable avec seulement deux alternatives, acquitement ou condamnation. On peut alors ajuster un modèle logit mesurant la probabilité d'être condamné.
- On réduit ensuite à la sous-population des condamnés. On a plus que deux alternatives : peine inférieure ou supérieure aux réquisitions. On peut alors ajuster un modèle logit mesurant la probabilité d'une peine sévère.

Données : survivre au naufrage du Titanic

Dans les données suivantes, on cherche à modéliser la probabilité, pour un passager du Titanic, d'avoir survécu au naufrage. On prend en compte trois variables indépendantes :

- Age (enfant/adulte). Reférence : adulte ;
- Sexe (homme/femme). Référence : homme ;
- Classe (première/deuxième/troisième). Référence : première classe.

On va donc mesurer :

- β_1 : l'effet sur $logit(Survie)$ du fait d'être un enfant plutôt qu'un adulte.
- β_2 : l'effet sur $logit(Survie)$ du fait d'être une femme plutôt qu'un homme.
- β_3 : l'effet sur $logit(Survie)$ du fait d'être en deuxième classe plutôt qu'en première.
- β_4 : l'effet sur $logit(Survie)$ du fait d'être en troisième classe plutôt qu'en première.

Données : survivre au naufrage du Titanic

PClass	Age	Sex	Survived	SexCode	Children
1st	29	Femme	Survécu	1	Adulte
1st	2	Femme	Pas survécu	1	Enfant
1st	30	Homme	Pas survécu	0	Adulte
1st	25	Femme	Pas survécu	1	Adulte
1st	0.92	Homme	Survécu	0	Enfant
1st	47	Homme	Survécu	0	Adulte

Quels déterminants de la survie ?

On commence par regarder les liens entre la variable dépendante et les variables indépendantes, afin de déterminer nos hypothèses.

	Pas survécu	Survécu	Total
Adulte	61.36	38.64	100
Enfant	39.58	60.42	100
All	58.6	41.4	100

Quels déterminants de la survie ?

	Pas survécu	Survécu	Total
Homme	83.29	16.71	100
Femme	33.33	66.67	100
All	65.7	34.3	100

Quels déterminants de la survie ?

	Pas survécu	Survécu	Total
1st	40.06	59.94	100
2nd	57.35	42.65	100
3rd	80.59	19.41	100
All	65.7	34.3	100

Modèle logit

	Dependent variable :
	Survived
ChildrenEnfant	1.258^{***}
	(0.284)
SexFemme	2.598^{***}
	(0.200)
PClass2nd	-0.888^{***}
	(0.236)
PClass3rd	-2.079^{***}
	(0.243)
Constant	-0.513^{***}
	(0.169)
Observations	756
Akaike Inf. Crit.	713.307
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Modèle logit – avec odds-ratio

	Dependent variable :
	Survived
ChildrenEnfant	3.519***
	(1.001)
SexFemme	13.435***
	(2.683)
PClass2nd	0.412***
	(0.097)
PClass3rd	0.125***
	(0.030)
Constant	0.599***
	(0.101)
Observations	756
Akaike Inf. Crit.	713.307
Note :	*p<0.1 ; **p<0.05 ; ***p<0.01

Efficacité du modèle

Probabilité moyenne prédite par le modèle pour les deux groupes.

Pas survécu	Survécu
0.2534	0.6414