Cours 3 : Régression linéaire sous R

1 Régression linéaire en statistiques descriptives

L'objectif est de déterminer la "meilleure" fonction affine décrivant une variable y en fonction d'autres variables $x_1, x_2, \ldots, x_p$:

$$y \simeq a_1 x_1 + \cdots + a_p x_p + b.$$

La notion de "meilleure" suppose une mesure de l'erreur faite dans cette approximation. L'option la plus courante consiste à utiliser la distance ℓ^2 : somme des carrés des erreurs parmi les observations.

Si on dispose de n observations de ces variables, que l'on note

$$(x_{1,1}, x_{1,2}, \ldots, x_{1,p}, y_1), \ldots, (x_{n,1}, x_{n,2}, \ldots, x_{n,p}, y_n),$$

alors la relation affine attendue prend la forme

$$Y \simeq X\Theta,$$

où

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}, \text{ et } \Theta = \begin{pmatrix} b \\ a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

et les erreurs sont donc $\varepsilon_i = y_i - (a_1 x_{i,1} + \cdots + a_p x_{i,p} + b)$, soit

$$Y = X\Theta + \varepsilon, \qquad \text{avec } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Le problème de régression linéaire consiste ainsi à déterminer Θ tel que la quantité $\|\varepsilon\|_2^2 = \varepsilon_1^2 + \cdots + \varepsilon_n^2$ soit minimale.

Remarque. Ces relations affines recouvrent le cas d'une relation polynomiale entre 2 variables réelles (ou un polynôme plus général) : par exemple, la relation $y \simeq a + bz + cz^2$ est une relation affine entre les variables y et (z, z^2).

De même, des relations sous forme multiplicative peuvent se ramener à des relations affines via le logarithme : par exemple, la relation $y \simeq CU^\alpha e^{-\beta U}$ équivaut à $\ln y \sim \ln C + \alpha \ln U - \beta U$, c'est-à-dire une relation affine entre les variables $\ln y$ et $(\ln U, U)$.

L'utilisation de la norme ℓ^2 donne un sens géométrique à la solution : lorsque le vecteur Θ varie, le produit $X\Theta$ parcourt le sous-espace vectoriel $E = \mathrm{Im} X$, si bien que la distance $\|Y - X\Theta\|_2$ est minimale lorsque $X\Theta$ est la projection orthogonale de Y sur E. Si X est injective, ceci détermine uniquement Θ. Dans la suite, on supposera généralement X injective : cela signifie que X est de rang maximal, ou encore que ses colonnes sont indépendantes, c'est-à-dire que **les variables ne sont pas affinement liées**. Ce n'est pas toujours le cas en pratique, et l'idéal est de repérer par anticipation des liens entre les variables pour en supprimer certaines (elles sont redondantes). Par exemple, des variables "âge" et "année de naissance" sont affinement liées (leur somme est constante) ; c'est le cas aussi pour des variables représentant des proportions, dont

la somme vaut 1. La présence de variables "presque liées" peut mener à des erreurs numériques ; diverses méthodes ont été développées pour traiter ce cas automatiquement (cf. régressions *ridge*).

Proposition : Il existe un unique vecteur $\Theta \in \mathbb{R}^{p+1}$ qui minimise $\|Y - X\Theta\|_2$, c'est l'unique vecteur $\widehat{\Theta}$ tel que $\widehat{Y} := X\widehat{\Theta}$ est la projection orthogonale de Y sur $\mathrm{Im}X$ et on peut le calculer par

$$\widehat{\Theta} = ({}^tXX)^{-1}\,{}^tXY.$$

On vérifie en effet que, si X est injective, alors ${}^tXX$ est une matrice carrée inversible. Et la propriété de projection orthogonale assure que ${}^tXY = {}^tX(X\widehat{\Theta})$ (le produit scalaire de Y avec toute colonne de X est préservé lorsque Y est remplacée par la projection $X\widehat{\Theta}$).

Cette formule est l'analogue multidimensionnel de la projection sur une droite, et prend la même forme : la projection orthogonale de y sur la droite engendrée par x est θx où

$$\theta = \frac{1}{\|x\|^2}(y|x) = ({}^txx)^{-1}\,{}^txy.$$

Le cas $p = 1$ (relation affine entre deux variables réelles) est très fréquent, on parle d'ailleurs de régression linéaire simple si $p = 1$ et multiple si $p > 1$. Pour $p = 1$, on a calcule explicitement

$${}^tXX = n\begin{pmatrix}1 & \overline{x}\\ \overline{x} & \overline{x^2}\end{pmatrix} \qquad \text{d'où} \qquad ({}^tXX)^{-1} = \frac{1}{nS_x}\begin{pmatrix}\overline{x^2} & -\overline{x}\\ -\overline{x} & 1\end{pmatrix} \text{ et } {}^tXY = n\begin{pmatrix}\overline{y}\\ \overline{xy}\end{pmatrix},$$

où $S_x = \overline{x^2} - (\overline{x})^2$ (variance empirique), et $\overline{y}$, $\overline{x}$, $\overline{x^2}$, $\overline{xy}$ représentent des moyennes. D'où

$$\widehat{\Theta} = \frac{1}{S_x}\begin{pmatrix}\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}\\ -\overline{x}\,\overline{y} + \overline{xy}\end{pmatrix} = \begin{pmatrix}b\\ a\end{pmatrix}$$

où on retiendra que $a = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)}$ et b est tel que $\overline{y} = a\overline{x} + b$.

La dernière propriété est générale (p quelconque) : si on note $\overline{x} = (\overline{x}_1, \ldots, \overline{x}_p)$, alors $\overline{y} = \overline{x}\widehat{\Theta}$. En effet, comme $\widehat{Y}$ est la projection de Y sur $\mathrm{Im}X$ qui inclut la droite $\mathbb{R}\mathbf{1}$, on a égalité des produits scalaires $(Y|\mathbf{1}) = (\widehat{Y}|\mathbf{1})$ donc Y et $\widehat{Y}$ ont même moyenne, or la moyenne de $\widehat{Y} = X\widehat{\Theta}$ est $\overline{x}\widehat{\Theta}$ par linéarité.

Comment apprécier la taille de l'erreur ? Les erreurs $\varepsilon_i = y_i - \widehat{y}_i$ s'appellent les *résidus*. La méthode de projection minimise la quantité

$$\mathrm{RSS} := \|Y - \widehat{Y}\|^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\varepsilon_i^2$$

RSS signifie "Residual Sum of Squares". Pour apprécier la valeur numérique de RSS, on la compare à d'autres quantités similaires. On définit ainsi

$$\mathrm{ESS} := \|\widehat{Y} - \overline{y}\mathbf{1}\|^2 = \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$$

ESS signifie "Explained Sum of Squares". C'est la variance des projections $\widehat{y}_i$. Les projections suivant exactement une relation affine, cette variance s'interprète comme la part de la variance des observations qui est due à la relation affine (et donc "expliquée" par le modèle). Enfin, par analogie on note

$$\mathrm{TSS} := \|Y - \overline{y}\mathbf{1}\|^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \mathrm{Var}(y)$$

TSS signifie "Total Sum of Squares". C'est la variance des observations.

Par théorème de Pythagore dans le triangle $(Y, \widehat{Y}, \overline{y}\mathbf{1})$, dont les côtés ont pour longueurs $\sqrt{\text{ESS}}$, $\sqrt{\text{TSS}}$ et $\sqrt{\text{RSS}}$, vu que $\mathbf{1}$ est dans l'image de X et donc orthogonal à $Y - \widehat{Y}$, on a

$$\text{TSS} = \text{ESS} + \text{RSS}.$$

Cela justifie (les termes étant positifs) de voir ESS comme la "partie" de la variance des y_i due à la relation affine, et la proportion associée est notée

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \in [0, 1]$$

R^2 est le *coefficient de détermination de Pearson* (communément appelé "le R2") de la régression. Cette quantité a l'avantage d'être adimensionnelle. Une valeur de R^2 proche de 1 signifie une bonne explication des données par la relation affine.

Le cas $p = 1$ est particulier. On a $\widehat{Y} = aX + (\overline{y} - a\overline{x})\mathbf{1}$ d'où en développant

$$\text{RSS} = \|Y - \overline{y}\mathbf{1} - a(X - \overline{x}\mathbf{1})\|^2 = nS_y - 2anS_{xy} + a^2 nS_x$$

avec $S_y = \frac{1}{n}\sum_i (y_i - \overline{y})^2$, $S_{xy} = \frac{1}{n}\sum_i (x_i y_i - \overline{x}\,\overline{y})$ et $S_x = \frac{1}{n}\sum_i (x_i - \overline{x})^2$ (de sorte que $a = \frac{S_{xy}}{S_x}$ avec la formule déjà vue), puis

$$\text{RSS} = nS_y - 2n\frac{S_{xy}^2}{S_x} + n\frac{S_{xy}^2}{S_x} = nS_y\left(1 - \frac{S_{xy}^2}{S_x S_y}\right) = \text{TSS}\left(1 - \text{Corr}(x, y)^2\right)$$

donc R^2 est le carré de la corrélation entre les variables x et y :

$$R^2 = \text{Corr}(x, y)^2.$$

En particulier, on note la symétrie entre x et y dans cette formule. Cela signifie que le R^2 est **le même** dans la régression de y par x et dans la régression de x par y.

2 Dans R

Dans R, on utilise la commande `lm` (pour "linear model"). L'usage courant est `lm(formula, data)` où `data` est un data frame, et `formula` est une expression de la forme `y~x` (sans guillemets) ou plus généralement `y~x1+x2+x3`, où `y`,`x`,`x1`,`x2`,`x3` sont les noms de variables du data frame. Dans la formule, on peut de plus utiliser des fonctions de variables sous la forme `I(...)` : par exemple `I(log(y))=x+I(x^2)+I(x^3)` pour exprimer $\log(y)$ comme fonction affine de x, x^2, x^3 (régression polynomiale).

On peut aussi exécuter `lm(y~x)` où `x` et `y` sont deux vecteurs.

Pour récupérer les résultats, on posera `reg=lm(...)`, et on obtient alors le vecteur des coefficients par `reg$coefficients` (ici, `reg` est une liste, dont on extrait la composante `coefficients`). Dans ce vecteur, l'ordonnée à l'origine s'appelle `(Intercept)`. Le vecteur des projections $\widehat{Y}$ est `reg$fitted.values`.

La commande `summary(reg)` (où `reg` est le résultat d'un appel à `lm`) effectue divers calculs (dont on parlera plus loin) et en particulier donne la valeur de R^2, appelée `Multiple R-squared`. Si `s=summary(reg)`, alors `s$r.squared` renvoie cette valeur.

Graphiquement, avec `ggplot`, on ajoute une régression linéaire par `+geom_smooth(method="lm")`. On ajoute souvent l'option `se=F` pour ne pas afficher la zone de confiance.

3 Régression linéaire et statistiques inférentielles

On fait l'hypothèse suivante : dans l'expression $y_i = x_i\Theta + \varepsilon_i$, les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont les réalisations de variables aléatoires indépendantes et de même loi gaussienne $\mathcal{N}(0, \sigma^2)$ (en particulier, elles ont même

variance). Ainsi, $\varepsilon \sim \mathcal{N}(0, I_n)$, et donc $Y \sim \mathcal{N}(X\Theta, \sigma^2 I_n)$ a pour densité

$$y \mapsto \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\|y - X\Theta\|^2}.$$

Cette densité est maximale par rapport à la variable Θ lorsque $\|y - X\Theta\|$ est minimale. Ainsi, la régression linéaire $\widehat{\Theta}$ est l'**estimateur du maximum de vraisemblance** de Θ.

Il s'en suit que $\widehat{\Theta} = ({}^tXX)^{-1}{}^tXY$ est également gaussien, d'espérance Θ et de matrice de covariance

$$\Gamma_{\widehat{\Theta}} = \sigma^2\, {}^t\big(({}^tXX)^{-1}{}^tX\big)({}^tXX)^{-1}{}^tX = \sigma^2({}^tXX)^{-1}.$$

En particulier, on en déduit que les coefficients suivent des lois gaussiennes si on connaît σ^2, et des lois de Student si on les normalise avec leur variance empirique :

Proposition : $\widehat{\Theta}$ est indépendant de $\widehat{\sigma}^2 = \frac{\|Y - \widehat{Y}\|^2}{n-p-1}$, et on a $(n-p-1)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1}$. En notant $\widehat{a}_0 (= \widehat{b}), \widehat{a}_1, \ldots, \widehat{a}_p$ les coefficients de $\widehat{\Theta}$, on a, pour tout i, $\frac{\widehat{a}_i - a_i}{\sqrt{\widehat{\sigma}^2({}^tXX)^{-1}_{i,i}}} \sim t_{n-p-1}$

La dernière partie permet notamment d'obtenir des intervalles de confiance sur les coefficients (`confint(reg,level=.95)`), et de faire des tests de l'hypothèse "a_i=0" (résultats fournis par `summary`, à droite des coefficients). Ces tests sont importants : ils permettent de savoir si une variable joue un rôle significatif dans le modèle.

Enfin, il est possible de tester si plusieurs coefficients sont simultanéments nuls. On note $\widehat{Y}_q$ la projection de Y sur l'espace engendré seulement par les q premières variables de X (de sorte que $\widehat{Y}_p = \widehat{Y}$ et $\widehat{Y}_0 = \overline{y}\mathbf{1}$). Alors $\frac{\|\widehat{Y}_p - \widehat{Y}_q\|^2}{\widehat{\sigma}^2(p-q)} \sim F(p-q, n-p-1)$. Dans `R`, ce test peut s'effectuer en calculant deux modèles de régression `reg.total=lm(...)` (avec toutes les variables) et `reg.partiel=lm(...)` (sans certaines variables), puis en exécutant `anova(reg.total, reg.partiel)`.

Cas particulier pour tester la nullité de tous les coefficients $a_1 = \cdots = a_p = 0$, $\frac{\|\widehat{Y} - \overline{y}\mathbf{1}\|^2}{p\widehat{\sigma}^2} \sim F(p, n-p-1)$. C'est le test effectué directement à la dernière ligne affichée par `summary(reg)`.

4 Exercices

4.1 Exercice : Hauteur et poids

Lien entre taille (`height`) et poids (`weight`) dans les données `women`.

- Ouvrir les données `women` (inclus dans R). Regarder l'aide sur ces données.
- Proposer une régression linéaire
- Faire une représentation graphique
- Proposer une régression plus fidèle
- Comparer les R^2.

4.2 Exercice : Taux de criminalité

Influence de diverses variables sur le taux de criminalité selon les états des USA.

- Charger les données `state.x77` (incluses dans R). Regarder l'aide sur ces données.
- Créer un dataframe avec les variables "Murder", "Population", "Illiteracy", "Income" et "Frost".
- Calculer les corrélations entre ces variables (`cor`). Quel lien entre "Frost" et "Murder" ?
- Proposer un modèle linéaire pour "Murder" en fonction des autres paramètres, puis un modèle réduit. Tester sa compatibilité avec les données.

4.3 Exercice : Taux d'ozone et vitesse du vent

On considère le jeu de données `airquality` disponible dans R.

1. Charger les données et comprendre d'où elles émanent.
2. Définir un nouveau data.frame `air` obtenu en enlevant les lignes où manquent des données "Ozone" ou "Wind".
3. Utiliser `lm` pour étudier le lien entre taux d'ozone et vitesse du vent. Peut-on affirmer un lien linéaire fort entre Ozone et Vent ?
4. Tracer la droite de régression sur le nuage de points.

4.4 Exercice : Identifiabilité

On considère le jeu de données "ocde.csv" à télécharger, et charger sous le nom `ocde`. On y trouve, en 1960, pour les pays de l'OCDE, le revenu par habitant `PCINC`, la part (en pourcents) de la population active travaillant dans l'agriculture `AGR`, dans l'industrie `IND` et dans les services `SER`.

1. Comparer les coefficients de détermination des ajustements linéaires suivants : `PCINC~AGR+IND+SER` et `PCINC~IND+SER`. Cela suggère-t-il quelque chose sur l'influence de la part d'emplois agricoles sur le revenu moyen ? Regarder les coefficients.
2. Ajouter au data.frame `ocde` une variable égale à la somme des 3 proportions. L'observer, et l'utiliser pour définir une variable `SER2` (toujours dans le data.frame `ocde`) qui modifie légèrement `SER` pour avoir toujours une somme de 100 %. Que donne l'ajustement selon `PCINC~AGR+IND+SER2` ?

Les colinéarités sont souvent plus subtiles et posent des difficultés.

4.5 Exercice : Analyse des résidus

Pour que l'interprétation des intervalles de confiances ou des tests soit pertinente, il faut vérifier les hypothèses statistiques : les résidus sont indépendants, centrés, gaussiens, de même variance. Regardons un exemple.

On considère le jeu de données `LifeCycleSavings` qui, pour 50 pays, relève le niveau d'épargne `sr` en fonction de la proportion de population âgée de moins de 15 ans `pop15`, de plus de 75 ans `pop75`, du revenu réel disponible par personne `dpi` et du taux de croissance du revenu disponible `ddpi` (tout ceci, en moyenne sur les années 1960–1970).

1. Approcher linéairement le niveau d'épargne à l'aide des autres variables. Garder le résultat dans une variable `reg`.
2. Calculer avec `rstandard(reg)` et représenter graphiquement les résidus standardisés (sous hypothèse gaussienne, ils suivent une loi de student t_{n-p-1} où n est le nombre d'observations et p le nombre de variables de la régression). Donner un `boxplot` de la distribution des résidus, et (séparément) un graphe avec les numéros des observations en abscisse et les résidus en ordonnée. Ajouter sur ce 2è graphe deux lignes horizontales (`gg_hline(intercept= )`) aux quantiles 5% et 95% de la loi normale standard (ou, mieux, de la loi de student à `df`$=n-p-1$ degrés de liberté, calculés avec `qt(p=0.05 ,df= )`). À quels pays correspondent les valeurs en-dehors de ces lignes, et sont-elles vraiment aberrantes ?
3. Pour tester la normalité (ou "studentité") des résidus, on peut représenter leur QQ-plot qui compare quantiles théoriques (loi de student, ou loi normale approximativement) et observés. On l'obtient (pour la loi normale) via `geom_qq()` en ayant spécifié l'esthétique `sample` (égale ici aux résidus standardisés). Ajouter `geom_qq_line()` pour avoir la droite `y=x`. Expliquer, commenter. On peut aussi utiliser, de façon plus quantitative que visuelle, le test de Shapiro-Wilk par exemple : utiliser la fonction `shapiro` pour effectuer ce test de normalité.

4.6 Exercice : Contre-exemples

Taper `data(anscombe)` pour charger ce jeu de données. C'est un `data.frame` qui comprend 8 variables `x1`,`y1`,...,`x4`,`y4` qui correspondent en fait à 4 jeux de données factices à étudier à titre d'exemple ou contre-exemple.

1. On va d'abord réorganiser ce data.frame de façon plus logique.

- Définir un data.frame `exemples` qui contient deux variables `x` et `y`, et dont les observations sont la concaténations de `x1`,...,`x4` pour `x`, et idem pour `y`. *Utiliser `c( , )` pour concaténer des vecteurs.*
- Ajouter une variable `n` à `exemples`, qui donne le numéro du jeu de données initial (de 1 à 4). *On pourra utiliser `rep(x,n)` qui répète un vecteur `x` `n` fois.*

2. Représenter ces données comme un nuage de points, en affichant chaque jeu dans une sous-fenêtre différente.
3. Ajouter la représentation de la régression linéaire dans chaque cas. On pourra utiliser le paramètre `fullrange=TRUE` de `geom_smooth` pour ne pas restreindre la droite au plus petit intervalle contenant les abscisses des données.
4. Parmi les nuages de points affichés, est-ce qu'un ajustement par une droite est-il toujours judicieux ? D'un point de vue théorique, à quelles conditions simples sur deux familles de couples de données unidimensionnelles peut-on conclure qu'elles ont la même droite de régression linéaire ?
5. Charger le fichier `datasaurus.csv`. Il s'agit d'une extension du contre-exemple précédent. Ici les données sont déjà sous un format pratique, où la variable `dataset` représente l'identifiant de l'un des 13 jeux de données considérés. Représenter comme précédemment les nuages de points et la droite de régression linéaire de chacun de ces jeux de données dans des sous-fenêtres différentes.

4.7 Exercice : Calcul "manuel" des formules

Une entreprise fixe des prix différents pour un produit particulier dans huit régions différentes des États-Unis. Elle souhaite étudier la liaison éventuelle entre le nombre de ventes (variable Y) et le prix du produit (variable X). Pour les $n = 8$ régions, on observe les valeurs $(x_1, y_1), \ldots, (x_n, y_n)$ de (X, Y) suivantes :

x_i	420	380	350	400	440	380	450	420
y_i	5.5	6.0	6.5	6.0	5.0	6.5	4.5	5.0

1. Représenter le nuage de points $(x_1, y_1), ..., (x_n, y_n)$. Suggère-t-il une liaison linéaire entre Y et X ?
2. On adopte alors le modèle de régression linéaire simple : $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Les paramètres β_0, β_1 sont des réels inconnus. On considère la forme matricielle usuelle : $Y = X\beta + \epsilon$, avec $\beta = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$. Définir dans R la matrice X associée.
3. Calculer $\widehat{\beta}$ par la formule $\widehat{\beta} = (X'X)^{-1}X'Y$. (Utiliser `solve` pour obtenir l'inverse, `t()` pour la transposée, et `%*%` pour le produit matriciel)
4. Vérifier que $\widehat{\beta} = \begin{pmatrix}b \\ a\end{pmatrix}$ avec $a = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$ et $\overline{y} = a\overline{x} + b$.
5. Tracer la droite de régression sur le nuage de points.
6. Calculer "à la main" le coefficient de détermination.
7. Retrouver les estimations précédentes avec la commande `summary` exécutée sur le résultat de la commande `lm`.

4.8 Exercice : Volume des arbres

Avec `data(trees)`, charger les données `trees` qui donnent le volume `Volume`, la hauteur `Height` et la circonférence `Girth` de 31 arbres. On cherche à comprendre la dépendance du volume en la hauteur et la circonférence.

1. Comparer la performance des modèles suivants (on regardera notamment le R^2, et on pourra faire un graphe) : `Volume~Height`, `Volume~Girth`, `Volume~I(Girth^2)`.
2. On souhaite trouver une relation du type Volume $\simeq C\text{Height}^{\alpha}\text{Girth}^{\beta}$. Quelle régression effectuer, et quels α et β trouve-t-on ?

4.9 Exercice :

On veut étudier la liaison éventuelle entre le taux de fibre oxydative (variable x) et la teneur en lipides dans la chair de lapins (variable y). Pour $n = 9$ échantillons de chair de lapins, on observe les valeurs suivantes :

x_i	3	4	4	17	24	45	55	68	73
y_i	0.9	1.3	1.0	2.4	2.8	4.4	5.2	6.3	6.6

1. Représenter le nuage de points. À partir de celui-ci, expliquer pourquoi on peut envisager l'existence d'une liaison linéaire entre Y et X.
2. Donner l'équation de la droite de régression avec la commande `lm`. Vérifier que celle-ci passe par le point de coordonnées $(\overline{x}, \overline{y})$. Tracer ce point de façon visible (avec un autre `+geom_point`, ou avec `+annotate("point",x= ,y= )`).
3. Avec `predict(reg,df)` où `reg` est le résultat de `lm` et `df` est un data.frame contenant une variable `x`, donner la prédiction de Y lorsque X = 28.