CHAPITRE 2

Introduction à la régression linéaire

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Automne 2019

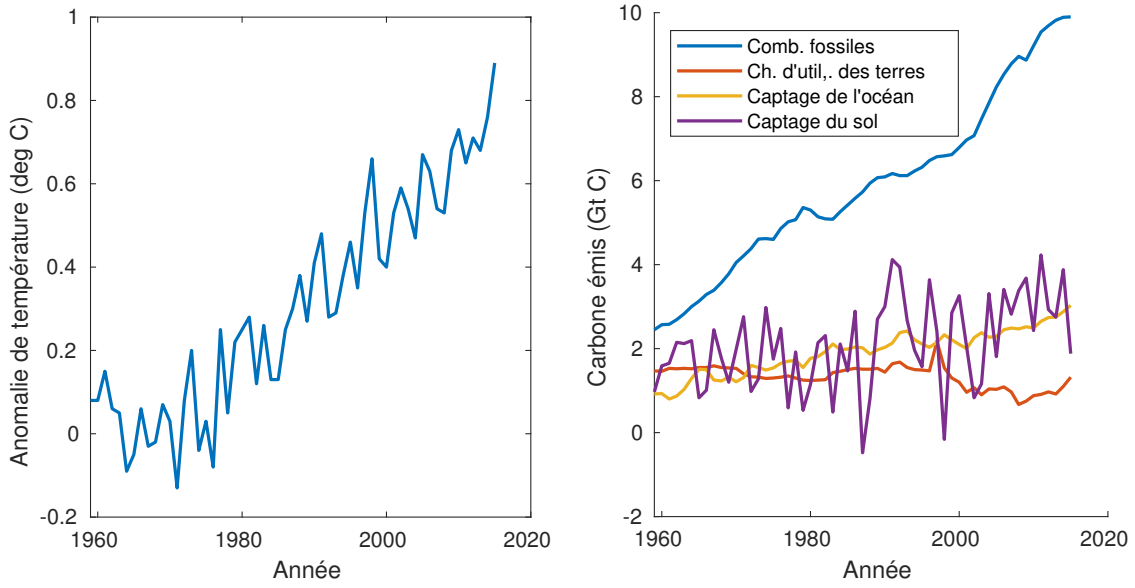
La régression linéaire est prépondérante en statistique et en apprentissage machine supervisé. Elle permet de quantifier la force de la relation entre une variable d'intérêt et un groupe de variables explicatives et de prédire la variable d'intérêt en fonction des variables explicatives.

Ce chapitre présente la base de la théorie de la régression linéaire. Les prochains chapitres seront consacrés à des méthodes plus avancées telles la régression bayésienne et la régression en composantes principales.

- Estimer les paramètres d'un modèle de régression avec la méthode des moindres de carrés
- Valider les hypothèses d'application de la régression linéaire ;
- Sélectionner les variables explicatives ;
- Détecter les valeurs suspectes.

Dans ce chapitre, l'anomalie de la température globale de la Terre sera étudiée en fonction de certaines composantes du cycle du carbone[1]. En particulier, $\boldsymbol{Y} = (Y_i : 1 \leq i \leq n)$ correspond au vecteur des $n = 57$ anomalies de températures annuelles moyennes de 1959 à 2015. Les variables explicatives considérées correspondent aux composantes suivantes du cycle du carbone :

1. Les données sur les anomalies de températures proviennent du National climate Data Center (www.ncdc.noaa.gov/) et celles sur les composantes du cycle du carbone proviennent de Boden, T.A., G. Marland, and R.J. Andres. 2016. Global, Regional, and National Fossil-Fuel CO2 Emissions. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy.



(a) (b)

FIGURE 2.1 – (a) Anomalies de température par rapport au climat de la période [1901,2000]. (b) Flux de carbone (en gigatonne de carbone par année) pour certaines des composantes du cycle du carbone.

— X_1 : émise par la combustion des combustibles fossiles ;
— X_2 : émise par le changement d'occupation des terres ;
— X_3 : captée par les océans ;
— X_4 : captée par le sol.

La figure 2.1a illustre les anomalie de températures par rapport à la période de référence [1901 , 2000] entre 1959 et 2015 et la figure 2.1b illustre les flux de carbones des variables explicatives entre 1959 et 2015.

2.1 Régression linéaire simple

Le modèle de régression linéaire **simple** correspond à celui où **une seule variable explicative**, disons X, est considérée pour expliquer la variable d'intérêt Y. Le cas où plusieurs variables explicatives sont utilisées correspond à la régression linéaire multiple qui est l'objet de la prochaine section.

Remarque. *On doit les premiers rudiments de la régression linéaire simple à Sir Francis Galton (1822-1911), un scientifique britannique qui étudia l'hérédité. Le terme* régression *a été proposé lorsqu'il étudia la taille des fils en fonction de la taille des pères. Il s'aperçut*

que la taille des fils avaient tendance à se rapprocher de la moyenne de la population, ce qu'il appela régression vers la moyenne.

2.1.1 Le modèle statistique de la régression linéaire simple

L'hypothèse principale de la régression linéaire simple consiste à supposer que l'espérance de la variable d'intérêt sachant la valeur de la variable explicative s'exprime sous la forme d'une relation linéaire :

Hypothèse 1 (Linéarité) :

$$E\left(Y \mid X = x\right) = \beta_0 + \beta_1 x; \tag{2.1}$$

où β_0 et β_1 constituent les paramètres inconnus du modèle de régression. Les paramètres β_0 et β_1 correspondent respectivement à l'ordonnée à l'origine et à la pente de la droite de régression. Le paramètre β_1 s'interprète comme l'effet espéré sur Y d'un changement d'une unité de X.

Remarque. *Dans le cadre usuel de la régression, l'incertitude sur la variable explicative X est supposée négligeable par rapport à l'incertitude de la variable Y. La quantité X n'est donc pas considérée comme une variable aléatoire mais comme une constante x, facilitant ainsi les calculs.*

2.1.2 Estimation des paramètres par les moindres carrés

Les paramètres β_0 et β_1 sont inconnus. Ils doivent être estimés à l'aide d'un échantillon aléatoire. Si on possède un échantillon aléatoire de taille n : $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$, où (x_i, Y_i) dénote la variable explicative x_i correspondant à la variable d'intérêt Y_i. L'hypothèse 1 de linéarité implique le modèle suivant pour chaque Y_i :

$$Y_i = \beta_0 + \beta_1 \, x_i + \varepsilon_i;$$

où ε_i est l'erreur aléatoire d'espérance nulle entre la droite de régression et Y_i. Deux autres hypothèses sont requises pour pouvoir estimer les paramètres β_0 et β_1. L'hypothèse 2 stipule que la variance de l'erreur doit être constante pour toute les variables aléatoire Y_i :

Hypothèse 2 (Homoscédasticité de la variance) :

$$\mathbb{V}ar(\varepsilon_i) = \sigma^2 \ ; \ \text{pour } i = 1, \ldots, n.$$

En particulier, la variance de l'erreur ne doit pas varier en fonction des Y_i. La troisième hypothèse impose que les variables aléatoires Y_i doivent être mutuellement indépendantes, ce qui implique que les erreurs doivent être mutuellement indépendantes :

Hypothèse 3 (Indépendance) : Les erreurs doivent être mutuellement indépendantes, c'est-à-dire ε_i indépendante de ε_j pour tout $i \neq j$.

Nous verrons comment valider les hypothèses 1 à 3 à la section 2.3.

Avec les hypothèses 1 à 3, l'estimation de β_0 et de β_1 par la méthode du maximum de la vraisemblance ou par la méthode bayésienne est impossible puisque l'on n'a pas supposé de distribution pour la vraisemblance. En effet, nous avons que $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}ar(\varepsilon_i) = \sigma^2$ mais nous n'avons pas supposé de forme paramétrique pour la loi de probabilité des ε_i.

La méthode classique en régression linéaire pour estimer β_0 et β_1 consiste à trouver la droite qui minimise la somme des erreurs au carré, autrement dit la droite qui minimise

$$D = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \, x_i)^2$$

Exercice 1

Montrez que la droite qui minimise D possède les paramètres suivants :

$$\hat{\beta}_0 = \bar{y} - \beta_1 \, \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2.2 Régression linéaire multiple

La régression linéaire multiple constitue la généralisation de la régression linéaire simple pour plusieurs variables explicatives, disons $X_1, X_2, \ldots, X_p$. À l'instar de la régression linéaire simple, l'hypothèse principale de la régression linéaire multiple consiste à supposer que l'espérance de la variable d'intérêt sachant les variables explicatives s'exprime sous la forme linéaire ;

Hypothèse 1 (Linéarité) :

$$E\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j; \tag{2.2}$$

où $\beta_0, \beta_1, \ldots, \beta_p$ sont les paramètres inconnus du modèle de régression. Le paramètre β_0 constitue l'ordonnée à l'origine de l'hyperplan de régression. Le paramètre β_j correspond à l'effet moyen sur la variable Y lorsque la variable explicative X_j augmente d'une unité et que toutes les autres variables demeurent constantes.

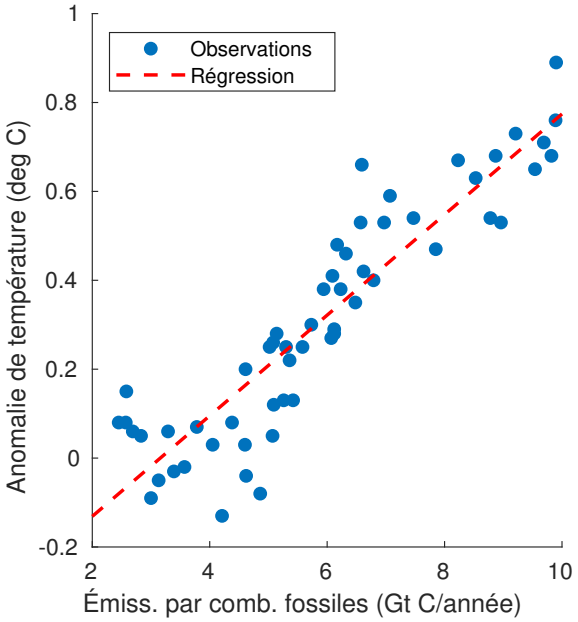


FIGURE 2.2 – Anomalies de température en fonction de la quantité de carbone émise par la combustion de combustibles fossiles superposées à la droite de régression linéaire estimée.

En supposant que les variables explicatives ne sont pas des variables aléatoires, l'hypothèse 1 implique le modèle statistique suivant :

$$Y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon; \tag{2.3}$$

où ε est un terme d'erreur aléatoire d'espérance nulle. L'erreur ε modélise la distance entre l'hyperplan de régression et la variable d'intérêt Y.

Le modèle de régression multiple s'écrit plus simplement lorsque la notation matricielle est utilisée. D'abord, intégrons l'ordonnée à l'origine β_0 dans un vecteur colonne avec les p coefficients de régression :

$$\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \ldots, \beta_p).$$

Soit le vecteur ligne des variables explicatives additionné d'un «1» à la première position :

$$\boldsymbol{x} = (1, x_1, \ldots, x_p).$$

Alors l'équation (2.3) s'écrit sous la forme suivante :

$$Y = \boldsymbol{x}\boldsymbol{\beta} + \varepsilon. \tag{2.4}$$

Si on possède un échantillon aléatoire composé de n couples $(\boldsymbol{x}_i, y_i)$ pour $1 \leq i \leq n$ où x_i correspond aux p variables explicatives de Y_i additionné d'un «1» pour l'ordonnée à l'origine :

$$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip}),$$

on a que

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i, \text{ pour } i = 1, \ldots, n.$$

Soit le vecteur colonne, $\boldsymbol{Y}^\top = (Y_1, \ldots, Y_n)$ et le vecteur colonne des erreurs $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \ldots, \varepsilon_n)$, l'ensemble des n équations de la forme de l'équation (2.4) pour l'échantillon aléatoire s'écrit de façon compacte :

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.5}$$

où X correspond à la matrice des superposant les n vecteurs de variables explicatives :

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}.$$

2.2.1 Estimation des paramètres par la méthodes des moindres carrés

Les paramètres $\beta_0, \beta_1 \ldots, \beta_p$ sont inconnus et doivent être estimés à l'aide d'un échantillon aléatoire. À l'instar de la régression linéaire simple, deux autres hypothèses sont requises pour pouvoir estimer les paramètres. L'hypothèse 2 stipule que la variance de l'erreur doit être constante pour toute les variables aléatoire Y_i :

Hypothèse 2 (Homoscédasticité de la variance) :

$$\mathbb{V}ar(\varepsilon_i) = \sigma^2 \text{ ; pour } i = 1, \ldots, n.$$

La troisième hypothèse impose que les variables aléatoires Y_i doivent être mutuellement indépendantes, ce qui implique que les erreurs doivent être mutuellement indépendantes :

Hypothèse 3 (Indépendance) : Les erreurs doivent être mutuellement indépendantes, c'est-à-dire ε_i indépendante de ε_j pour tout $i \neq j$.

Encore une fois, une estimation par maximum de la vraisemblance ou par le théorème de Bayes est impossible puisque nous n'avons pas supposé de distribution pour l'erreur ε. La méthode la plus simple consiste à trouver l'hyperplan de régression qui passe dans le nuage de points $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq n\}$ en minimisant la somme des erreurs au carré :

$$D = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon} = (\boldsymbol{Y} - X\boldsymbol{\beta})^\top(\boldsymbol{Y} - X\boldsymbol{\beta}).$$

En utilisant le calcul différentiel matriciel et en particulier les deux identités suivantes :

$$\frac{\partial\left(\boldsymbol{\beta}^\top X^\top \boldsymbol{Y}\right)}{\partial \boldsymbol{\beta}} = X^\top \boldsymbol{Y}$$
$$\frac{\partial\left(\boldsymbol{\beta}^\top X^\top X\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = 2X^\top X\boldsymbol{\beta},$$

on peut montrer que l'hyperplan de régression qui minimise la somme des erreurs au carré D possède les paramètres :

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}.$$

Remarque. *La notation matricielle introduite dans cette section fonctionne également dans le cas de la régression linéaire simple. Pour la régression linéaire simple, la matrice X possède n lignes et deux colonnes : une colonne de «1» et une colonne correspondant à la variable explicative.*

2.2.2 Estimation de la variance de l'erreur

L'hypothèse 1 et 2 de la régression linéaire stipulent respectivement que $\mathbb{E}(\varepsilon_i) = 0$ et $\mathbb{V}ar(\varepsilon_i) = \sigma^2$. Posons

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \varepsilon_i^2,$$

alors on peut montrer que

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2.$$

Par conséquent, $\hat{\sigma}^2$ est un estimateur de σ^2. En remplaçant ε_i par les erreurs observées e_i, aussi appelées résidus :

$$e_i = y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}},$$

on obtient donc une estimation de σ^2.

2.2.3 Prévision de la variable d'intérêt

Un des buts de la régression linéaire consiste à estimer la variable d'intérêt Y en fonction des variables explicatives, ce qui est appelé *prévision*. L'estimation de Y_0, dénotée $\hat{Y}_0$ sachant $\boldsymbol{x}_0$ est tout simplement le point correspondant sur le plan de régression :

$$\hat{Y}_0 = \boldsymbol{x}_0 \hat{\boldsymbol{\beta}}.$$

2.2.4 Inclusions de variables qualitatives

Des variables explicatives qualitatives peuvent être intégrées dans un modèle de régression linéaire. Par exemple, si deux opérateurs sont en charge de recueillir les données, on peut vérifier si les opérateurs ont une influence sur la mesure en intégrant la variable suivante indicatrice suivante dans les variables explicatives :

$$X = \begin{cases} 0 & \text{si l'observation provient de l'opérateur 1;} \\ 1 & \text{si l'observation provient de l'opérateur 2.} \end{cases}$$

Dans le cas où il y aurait trois opérateurs, deux variables explicatives seraient nécessaires :

X_1	X_2	
0	0	si l'observation provient de l'opérateur 1 ;
1	0	si l'observation provient de l'opérateur 2 ;
0	1	si l'observation provient de l'opérateur 3.

On peut procéder de la sorte pour un nombre quelconque d'opérateurs ou pour n'importe quelle variable explicative qualitative.

2.3 Validation des hypothèses de la régression linéaire

Les paramètres du modèle de régression linéaire nécessite les trois premières hypothèses pour être estimés avec la méthode des moindre carrés présentée dans ce chapitre. Dans le cas de la régression linéaire simple, l'hypothèse 1 peut être vérifiée en traçant la droite de régression dans le nuage de points $\{(x_i, y_i) : i = 1, \ldots, n\}$. Si une droite coupe le nuage, comme c'est le cas à la figure 2.2, alors l'hypothèse de linéarité est raisonnable. Notez que la forme du nuage de points peut donner une indication sur la forme de la relation entre X et Y (linéaire, quadratique, etc.)

Dans le cas de la régression multiple, les hypothèses 1 et 2 peuvent être validées en analysant les erreurs observées entre l'hyperplan de régression et les observations, aussi appelés les *résidus de régression.*

$$e_i = y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}, \text{ pour } i = 1, \ldots, n.$$

Un nuage de points illustrant les résidus e_i en fonction des estimations de Y_i, soit $\hat{Y}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$. Les résidus devraient former un nuage de points de forme rectangulaire centré autour de l'axe des abscisses. Une forme conique du nuage de points suggère la violation de l'hypothèse 2 tandis qu'une forme parabolique ou sinusoïdale suggère la violation de l'hypothèse 1. Dans ce dernier cas, une transformation des données pourraient s'avérer utile (voir la prochaine section).

En pratique, il s'avère impossible de vérifier l'hypothèse 3 lorsque seulement les données sont disponibles. Seule une planification d'expérience adéquate permet de s'assurer que les données constituent un échantillon aléatoire.

2.4 Transformation de variable

Parfois, la relation linéaire entre X et Y n'est pas adéquate ou optimale. Que ce soit parce que le nuage de points $\{(x_i, y_i) : i = 1, \ldots, n\}$ indique une autre forme ou bien parce que les résidus e_i ne sont pas constant en fonction de $\hat{Y}_i$, une transformation de la variable Y peut être utile pour linéariser la relation. Par exemple, on peut tenter la transformation suivante :

$$\ln Y = \beta_0 + \beta_1 x + \varepsilon. \tag{2.6}$$

Les techniques de la régression linéaire simple peuvent alors être utilisées pour estimer le modèle aux données transformées $\{(x_i, \ln y_i) : i = 1, \ldots, n\}$. Les transformations les plus courantes sont $\ln Y$, $\sqrt{Y}$ et Y^2.

2.5 Régression polynomiale

Le modèle de régression linéaire exprimé à l'équation (2.5) permet également de modéliser une relation polynomiale entre une variable explicative X et la variable réponse Y. Par exemple, si la vraie relation entre X et Y est un polynôme d'ordre 2 :

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon,$$

alors il peut être écrit comme un modèle de régression linéaire multiple

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

où $X_1 = X$ et $X_2 = X^2$. L'estimation par les moindres carrés développée pour la régression linéaire peut être utilisée. Cela est vrai pour n'importe quel degré de polynôme.

2.6 Tests d'hypothèses et intervalles de confiance

Lorsque les paramètres du modèle de régression sont estimés, on peut vérifier si la relation linéaire entre les variables explicatives et la variable d'intérêt est significative. On peut également tester si chaque variable explicatives explique de façon significative la variable d'intérêt. Dans le premier cas, on dit que l'on teste l'importance de la régression et dans le deuxième cas, on dit que l'on teste l'importance des variables explicatives. On peut également développer un intervalle de confiance pour le plan de régression et pour la prévision de la variable d'intérêt.

Pour effectuer les tests d'hypothèses et développer les intervalles de confiance, une quatrième hypothèse est nécessaire sur les erreurs ε_i :

Hypothèse 4 (Normalité) : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, pour $i = 1, \ldots, n$.

Cette dernière hypothèse peut être vérifiée à l'aide d'un test de normalité effectué sur les résidus $(e_i : 1 \leq i \leq n)$ de la régression.

2.6.1 Test de l'importance de la régression

Tester s'il existe une relation linéaire significative entre les variables explicatives et la variable d'intérêt revient à tester si au moins une des variables explicatives possède un pouvoir prédictif significatif :

$\mathcal{H}_0$: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$;

$\mathcal{H}_1$: $\beta_j \neq 0$ pour au moins un $j \in \{1, \ldots, p\}$.

L'idée du test consiste à décomposer la variabilité de Y, dénotée $SS_T = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$ en deux portions, une portion expliquée par les variables explicatives $SS_R = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_i)^2$ et la portion résiduelle $SS_E = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. La régression est significative si la portion de la variabilité expliquée par les variables explicatives est grande par rapport à la variabilité résiduelle. On peut montrer que la statistique

$$F_0 = \frac{SS_R/p}{SS_E/(n-p-1)}$$

est distribuée selon la loi de Fisher à p degrés de liberté au numérateur et $(n-p-1)$ degrés de liberté au dénominateur si $\mathcal{H}_0$ est vraie. On rejette alors $\mathcal{H}_0$ au seuil α si F_0 est plus grand que le quantile d'ordre $1-\alpha$ de la loi de Fisher à p degrés de liberté au numérateur et $(n-p-1)$ degrés de liberté au dénominateur.

Le rapport entre la variabilité expliquée et la variabilité totale, dénotée R^2 et appelé le coefficient de détermination, est d'ailleurs un indice de la qualité du modèle de régression :

$$R^2 = \frac{SS_R}{SS_T}.$$

On a que $0 \leq R^2 \leq 1$. Si $R^2 \to 1$, alors une grande partie de la variabilité de Y est expliquée par les variables explicatives X. Si $R^2 \to 0$, alors les variables explicatives X n'expliquent qu'un très petite partie de la variabilité de Y. On cherche généralement un modèle avec le plus grand coefficient de détermination possible.

2.6.2 Intervalle de confiance sur un coefficient de régression

Dans un cours sur la régression, on pourrait montrer que :

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{n-p-1}(0, 1^2);$$

où c_{jj} correspond à l'élément (j,j) de la matrice $(X^\top X)^{-1}$. De cette équation, on peut développer un test d'hypothèses pour tester

$\mathcal{H}_0$: $\beta_j = 0$;

$\mathcal{H}_1$: $\beta_j \neq 0$.

ou bien développer un intervalle de confiance pour β_j. Si la valeur 0 est incluse dans l'intervalle de confiance, alors la variable explicative X_j n'explique pas de façon significative Y.

2.6.3 Intervalle de prévision

On a vu précédemment que la meilleur estimation de Y_0 correspondante aux variables explicatives $\boldsymbol{x}_0$ était le point sur l'hyperplan de régression. Avec l'hypothèse 4 de normalité, on peut calculer un intervalle de confiance pour Y_0, la vraie valeur en utilisant la relation suivante :

$$\frac{\hat{Y}_0 - Y}{\sqrt{\hat{\sigma}^2 \left\{1 + \boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top\right\}}} \sim t_{(n-p-1)}(0, 1).$$

2.7 Comparaison de modèles

Dans un monde idéal, toutes les variables pouvant potentiellement expliquer en partie la variable réponse devraient être intégrées au modèle de régression. Toutefois, cette approche n'est pas efficace d'un point de vue numérique. Plus on ajoute de variables explicatives, plus il devient difficile d'estimer les paramètres de régression associés à ces variables. D'une part, l'attribution d'une variation de la réponse à une variable explicative devient de plus en plus difficile à mesure que le nombre de variables explicative augmente. D'autre part, un problème d'instabilité numérique peut survenir dû à la multicolinéarité (voir section 2.9) Un modèle de régression optimal est un compromis entre le pouvoir prédictif de la variable réponse et la qualité d'ajustement des paramètres. Le modèle optimal conserve le nombre minimal de variables explicatives tout en maximisant le pouvoir prédictif sur la réponse. Pour ce faire, on ne retient que les variables explicatives qui ont un pouvoir **prédictif significatif** sur la variable réponse. Deux méthodes pour sélectionner les variables explicatives seront présentées dans cette section. La première méthode est la plus répandue en statistique. Il s'agit de comparer les coefficient de détermination des différents modèles. L'autre méthode, plus répandue en apprentissage machine, consiste à comparer le pouvoir prédictif des modèles.

2.7.1 Coefficient de détermination

Supposons que l'on souhaite comparer les deux modèles de régression suivants qui ne se distinguent que par les variables explicatives utilisées :

$$\mathcal{M}_1 : Y = \beta_0 + X_1\beta_1 + \varepsilon;$$
$$\mathcal{M}_2 : Y = \beta_0 + X_2\beta_2 + \varepsilon.$$

On souhaite donc choisir le modèle avec la meilleure variable explicative. Dans ce cas, le meilleur modèle est celui avec le plus grand coefficient de détermination R^2. En effet, on cherche le modèle qui explique la plus grande partie de la variabilité de Y.

Supposons maintenant que l'on veuille choisir entre les modèles précédents et le suivant :

$$\mathcal{M}_3 : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Puisque les modèles n'ont pas tous la même dimension, le coefficient de détermination n'est pas le meilleur critère car il ne pénalise pas pour l'ajout de variables explicatives non significatives. Il convient alors d'utiliser le coefficient de détermination ajusté :

$$R^2_{aj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}.$$

À l'instar du coefficient de détermination, le R^2_{aj} est borné entre 0 et 1 et $R^2_{aj} \to 1$ indique une forte adéquation avec le modèle de régression. Pour comparer des modèles de différentes dimensions, il convient de sélectionner celui avec le plus grand R^2_{aj}.

2.7.2 Puissance de prévision

En apprentissage machine, il est commun de partitionner le jeu de données en «training dataset» et en «test dataset». On se sert de l'échantillon d'entraînement pour estimer les paramètres du modèle de régression et on estime les prévisions pour les données de l'échantillon test. Puisque l'on possède les vraies données de l'échantillon test, on choisit le modèle qui donne les meilleures prévisions.

Dans le cadre de ce chapitre, nous effectuerons une validation croisée qui exploite de façon plus efficace le jeu de données. La validation croisée consiste à estimer la valeur Y_i avec les $(n-1)$ données restantes. Dénotons par $\hat{Y}_{-i}$ l'estimation de Y_i avec les $n-1$ données restantes. On répète cette procédure pour toutes les données et on peut calculer un coefficient de détermination de prévision :

$$R^2_{prev} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_{-i})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Plus le R^2_{prev} est près de 1, plus le modèle de régression obtient des prévision près de la vraie valeur. On sélectionnera le modèle qui possède le plus grand R^2_{prev}.

On pourrait penser que calculer le R^2_{prev} sera très long parce que l'on devrait estimer n fois les paramètres de la régression puis estimer Y_i. Ce n'est toutefois pas le cas. Posons $\tilde{e}_i = Y_i - \hat{Y}_{-i}$ l'erreur observée par validation croisée. On peut montrer que cette erreur peut être obtenue à partir du modèle complet :

$$\tilde{e}_i = \frac{e_i}{1 - h_{ii}},$$

où h_{ii} est l'élément (i,i) de la matrice $H = X(X^\top X)^{-1}X^\top$ du modèle complet. Une démonstration simple de se résultat se trouve en ligne à l'adresse suivante : https://robjhyndman.com/hyndsight/loocv-linear-models/. Pour calculer le R^2_{prev} d'un modèle de régression linéaire, il suffit d'estimer les paramètres du modèle avec toutes les données puis de calculer

$$R^2_{prev} = 1 - \frac{\sum_{i=1}^n \left(\frac{e_i}{1-h_{ii}}\right)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}.$$

2.8 Données suspectes

Les données aberrantes, qui peuvent correspondent à des erreurs de mesure ou à des erreurs de transcription, peuvent occasionner d'importantes conséquences dans une analyse de régression. Dans un premier temps, un critère permettant de détecter les données suspectes est nécessaire. Une valeur est suspecte si elle est très peu probable considérant le modèle de régression. Les résidus studentisés, définis à partir des résidus de la régression et de la matrice H de la façon suivante :

$$s_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}},$$

permettent d'identifier les valeurs suspectes. En effet si $|s_i| > 3$, alors y_i est une valeur suspecte.

Lorsque les données suspectes sont identifiées, on doit choisir si on les conserve ou si on les retire du jeu de données. S'il est clair que la donnée est une erreur ou que celle-ci est non représentative du jeu de données, il est préférable de la retirer. Si on ne peut exclure que la données suspecte est une erreur, il faut alors la conserver.

2.9 Multicolinéarité

La multicolinéarité est le problème qui survient lorsque les variables explicatives sont corrélées linéairement entres elles. Par exemple, si on peut écrire une colonne de la matrice des variables explicatives X comme une combinaison linéaire des autres colonnes, alors le déterminant de la matrice $(X^\top X)$ est nul et cette dernière n'est pas inversible. On ne peut donc pas estimer les paramètres par la méthode des moindres carrés. Même lorsque les colonnes ne sont pas exactement liées linéairement mais suffisamment pour que le déterminant de la matrice $(X^\top X)$ soit près de 0, les estimations des paramètres de régression deviennent très instables numériquement. Dans cette section, des méthodes pour détecter et retirer les variables responsables de la multicolinéarité sont présentées. Les prochains chapitres présentent deux méthodes possibles permettant de traiter avec la multicolinéarité en conservant toutes les variables explicatives. Il s'agit de la régression en composantes principales et la régression bayésienne.

2.9.1 Détection de la multicolinéarité avec le facteur d'inflation de la variance

Pour mesurer le niveau de dépendance linéaire de la variable explicative X_j avec les autres variables explicatives $X_{-j} = (X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$, on peut effectuer une régression linéaire où la variable réponse est X_j et les variables explicatives X_{-j}. On définit le **facteur d'inflation de la variance** (VIF) de la façon suivante à l'aide du coefficient

de détermination $\tilde{R}_j^2$ de cette régression :

$$VIF_j = \frac{1}{1 - \tilde{R}_j^2}.$$

Le facteur d'inflation de la variance augmente lorsque le $\tilde{R}_j^2$ augmente.

Dans la littérature, il est suggéré de retirer les variables explicatives pour lesquelles le facteur d'inflation de la variance est supérieur à 10. Cette valeur est arbitraire mais elle donne néanmoins un seuil pour lequel il devient important de se soucier de la multicolinéarité.

2.9.2 Détection de la multicolinéarité avec les valeurs propres

Vous vous rappelez certainement que le déterminant d'une matrice est égale au produit des valeurs propres de la matrice. Dans le cas de la matrice de $(X^\top X)$ de taille $(p+1) \times (p+1)$, le déterminant s'écrit en fonction des $(p+1)$ valeurs propres $\lambda_0, \ldots, \lambda_p$:

$$|X^\top X| = \prod_{j=0}^{p} \lambda_j.$$

La matrice n'est pas $(X^\top X)$ n'est pas inversible si son déterminant est nul et donc si au moins une valeur propre λ_j est nulle.

En présence de multicolinarité, au moins une valeur propre sera bien plus petite que les autres. De façon arbitraire, la multicolinéarité devient importante lorsque le facteur

$$\phi_j = \sqrt{\frac{\max\limits_{j \in \{0,\ldots,p\}} \lambda_j}{\lambda_j}}$$

dépasse 30. Il est alors recommandé de rejeter les variables explicatives correspondantes.

2.10 Exercices

1. Déterminer si les énoncés suivants sont vrais ou faux.
   a) Le modèle comprenant toutes les variables explicatives disponibles a toujours un R^2 supérieur à celui des modèles incluant moins de variables explicatives.
   b) Le modèle comportant toutes les variables explicatives disponibles a toujours une estimation de la variance de l'erreur $\hat{\sigma}^2$ supérieur à celui des modèles incluant moins de variables explicatives.
   c) Les estimations $\hat{\beta}_1, \ldots, \hat{\beta}_p$ ont tous la même variance échantillonnale.

d) Si certaines variables explicatives sont très corrélées entre elles, les estimations des coefficient de régression peuvent changer beaucoup selon les variables incluses dans le modèle.

e) Si les observations sont indépendantes, alors les coefficients de régression sont mutuellement indépendants.

f) Les estimations de β_0 et de β_1 seront les mêmes avec le modèle $Y = \beta_0 + X_1\beta_1 + \varepsilon$ qu'avec le modèle $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$.

2. Montrer que le coefficient de détermination R^2 peut aussi s'écrire sous la forme suivante :

$$R^2 = 1 - \frac{SS_E}{SS_T}.$$

3. Aux États-Unis, décrocher un baccalauréat dans les meilleures universités peut coûter jusqu'à 300 000 USD. On veut savoir s'il existe un lien entre les frais de scolarité annuels et le salaire médian des diplômés en mi-carrière. Le jeu de données `tuition_vs_salary.csv` contient les frais de scolarité annuels de 12 université américaines choisie arbitrairement et les salaires annuels médians en mi-carrières des diplômés [2].

   a) Tracer le nuage de points entre les frais de scolarité (Tuition) et le revenu médian annuel des diplômés en mi-carrière (Salary). Est-ce qu'une relation linéaire semble raisonnable ?

   b) Quelles sont les estimations ponctuelles obtenues par la méthode des moindres carrés des coefficients de régression ?

   c) En vous basant sur les résidus de la régression, est-ce que les hypothèses 1 et 2 vous semblent raisonnables ?

   d) En vous basant sur la droite de Henry des résidus, est-ce que l'hypothèse 4 de normalité semble raisonnable ?

   e) Existe-t-il une relation linéaire significative entre les frais de scolarité et le salaire médian ? Justifiez.

   f) Quel est le coefficient de détermination de la régression ?

   g) Est-ce que les réponses aux questions (e) et (f) sont incompatibles ?

2. Les données sur les frais de scolarité ont été récupérées du site https://www.forbes.com/value-colleges/list/ et les données sur les salaires du site https://www.payscale.com/college-salary-report

4. Le jeu de données `visco.csv` contient la résistance au cisaillement (en kPa) d'un composé de caoutchouc en fonction de la température de durcissement (en degré Celcius).
   a) Tracez la résistance au cisaillement en fonction de la température. Est-ce qu'une relation linéaire est appropriée ?
   b) Estimez les paramètres de régression du modèle quadratique, *i.e.* $Y_i = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon$. Quel est le coefficient de détermination ajusté ?
   c) Estimez les paramètres de régression du modèle cubique, *i.e.* $Y_i = \beta_0 + x\beta_1 + x^2\beta_2 + x^3\beta_3 + \varepsilon$. Quel est le coefficient de détermination ajusté ?
   d) En fonction du coefficient de détermination ajusté, quel est le meilleur modèle ?

5. Le jeu de données `notes.csv` compile les notes obtenues au contrôle 1, au contrôle 2 et au final des 91 étudiants inscrits dans ma section du cours MTH2302B pour les sessions A2017 et H2018. On souhaite déterminer s'il existe une relation linéaire entre la note du final (Y) et les variables explicatives suivantes :
   — x_1 : notes au CP1 ;
   — x_2 : notes au CP2 ;
   — x_3 : session.
   Posons
   $$x_3 = \begin{cases} 0 & \text{si l'étudiant a suivi le cours durant la session A2017 ;} \\ 1 & \text{si l'étudiant a suivi le cours durant la session A2018.} \end{cases}$$
   Le modèle de régression linéaire est le suivant :
   $$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \varepsilon_i, \text{ pour } i = 1, \ldots, 91.$$
   a) Estimez les coefficients de régression.
   b) En vous basant sur les résidus de la régression, est-ce que les hypothèses 1, 2 et 4 vous semblent raisonnables ?
   c) Estimez les intervalles de confiance de niveau 95% des coefficients de régression.
   d) Y a-t-il des variables non-significatives ?
   e) Supposons qu'un étudiant de la session H2018 a obtenu 13/20 et 15/20 lors des contrôles partiels mais qu'il ne s'est pas présenté au final. Obtenez une estimation de sa note au final s'il s'était présenté ainsi qu'un intervalle de confiance à 95% sur cette estimation.

6. Il est assez difficile et inconfortable pour les patients de mesurer le pourcentage de matière grasse de celui-ci. En effet, cette mesure implique d'immerger le patient

dans un cylindre gradué rempli d'eau. Par conséquent, on souhaite savoir si on peut prédire le pourcentage de gras Y avec trois mesures beaucoup plus simples à obtenir :
— x_1 : l'épaisseur des plis de la peau des triceps (en mm) ;
— x_2 : le tour de cuisse (en mm) ;
— x_3 : la circonférence du bras en (mm).
Les mesures du fichier `bodyfat.csv` proviennent de 20 femmes en bonne santé, âgées entre 20 et 34 ans. Il y a en tout $2^3 = 8$ modèles de régression possibles avec ces 3 variables explicatives.

a) Identifiez le modèle de régression linaire qui prédit le mieux le pourcentage de gras. Utilisez le coefficient de détermination de prévision R^2_{prev} comme critère.

b) Identifiez le modèle de régression linaire qui explique le mieux la variabilité de Y. Utilisez le coefficient de détermination ajusté R^2_{aj} comme critère.

7. Le jeu de données `bloodpressure.csv` contient les tensions artérielles Y mesurée en mm de Hg de 20 patients souffrant d'hypertension en fonction de
— x_1 : leur âge (en années) ;
— x_2 : leur poids (en kg) ;
— x_3 : la surface de leur corps (BSA, en m^2) ;
— x_4 :temps écoulé depuis le début de leur hypertension (en années) ;
— x_5 : leur pouls au repos (en battements par minutes) ;
— x_6 : leur niveau de stress (de 0 à 100).
Les médecins veulent déterminer les variables qui sont significative pour expliquer la tension artérielle.

a) Calculez le facteur d'inflation de la variance VIF pour chacune des variables explicatives.

b) Selon le facteur d'inflation de la variance, y a-t-il présence de multicolinéarité ? Le cas échéant, quelles variables devrions nous retirer ?

c) Calculez à l'aide d'un logiciel les valeurs propres de la matrice $(X^\top X)$.

d) Avec les valeurs propres obtenues, y a-t-il présence de multicolinarité ? Le cas échéant, quelles variables devrions nous retirer ?