CHAPITRE 2

Régression linéaire

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

Jonathan Jalbert – Hiver 2023

La régression linéaire a été développée bien avant l'arrivée des ordinateurs. Sa simplicité et sa performance en font néanmoins une méthode encore pertinente aujourd'hui, même si la puissance de calcul actuelle permet des modèles beaucoup plus complexes. La compréhension des modèles linéaires est à mon avis essentielle pour comprendre les modèles plus avancés. C'est pourquoi nous passerons *beaucoup* de temps à décortiquer ce modèle qui constitue un fer de lance de la statistique depuis plus d'un siècle.

La régression linéaire permet de :

— Quantifier la force de la relation entre une variable d'intérêt et un groupe de variables explicatives.

— Prédire la variable d'intérêt en fonction des variables explicatives.

Des cours entiers existent sur la régression linéaire. Puisque nous n'avons qu'un seul chapitre à y consacrer, certaines démonstrations sont omises pour se concentrer sur les caractéristiques essentielles. À la fin de ce chapitre, vous devriez être en mesure de :

— 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.

Pour illustrer la théorie, l'anomalie de la température globale de la Terre sera étudiée

en fonction des principales composantes anthropiques 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 principales composantes du cycle du carbone :

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

Ces flux de carbone, mesurés en gigatonnes, sont disponibles de 1959 à 2015. 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_1, 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. *Les premiers rudiments de la régression remontent à Gauss (1777–1855) et Legendre (1752–1833), mais le terme régression linéaire est dû à Galton (1822–1911), un scientifique britannique qui étudiait l'hérédité. Il a utilisé le terme* régression *pour décrire la tendance qu'avait la taille des fils à se rapprocher de la moyenne de la population plutôt que de celle des pères, 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 :

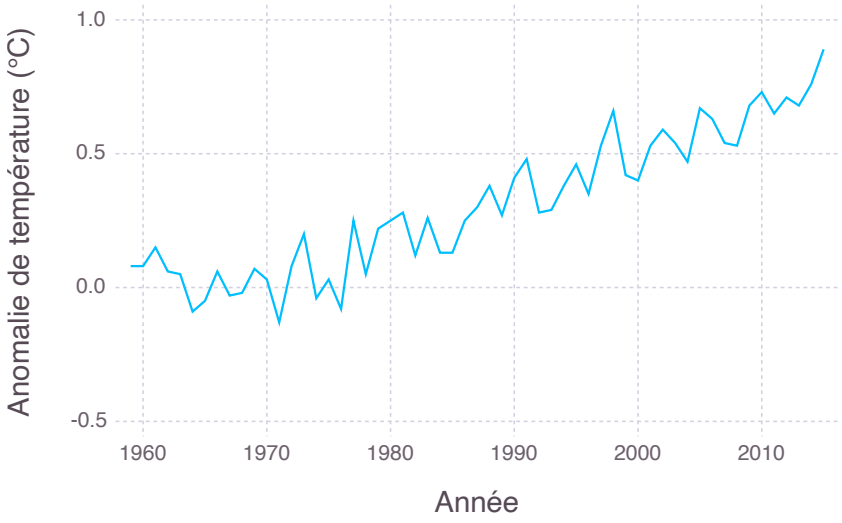
Hypothèse 1 (Linéarité) :

$$E\left(Y \mid X_1 = x_1\right) = \beta_0 + x_1\,\beta_1; \tag{2.1}$$

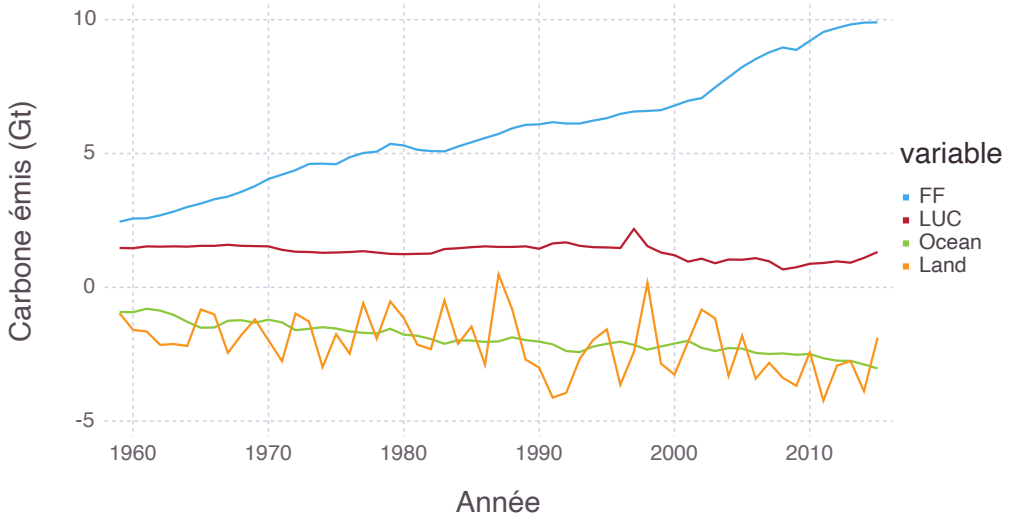
où β_0 et β_1 constituent les paramètres inconnus du modèle de régression qui devront être estimés avec les données [2]. Ces paramètres β_0 et β_1 correspondent respectivement à l'ordon-

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.

2. Dans le langage de l'apprentissage machine, on dit qu'on apprend des données lorsque l'on estime les paramètres avec les données. En statistique, on parle plutôt d'*estimation*.



(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 les principales composantes du cycle du carbone.

née à l'origine et à la pente de la droite de régression. Le paramètre β_1 s'interprète comme l'effet moyen sur Y d'un changement d'une unité de X_1.

Remarque. *L'équation* (2.1) *constitue une* relation statistique*, c'est à dire que le lien entre X_1 et Y n'est pas déterministe. Pour une valeur donnée x_1, il y a une incertitude résiduelle sur le y correspondant. Par opposition, dans une relation déterministe, il n'existe pas d'incertitude résiduelle sur y si x_1 est connu. Un exemple de relation déterministe est la conversion de température en degrés Fahrenheit (y) depuis les dégrés Celsius (x) :*

$$y = \frac{9}{5}x + 32.$$

Pour une valeur donnée x en degrés Celcius, la température exacte en degrés Fahrenheit est connue.

Soit l'échantillon aléatoire de taille n composé des couples de variable explicative et de variable d'intérêt $\{(x_{11}, y_1), (x_{21}, y_2), \ldots, (x_{n1}, y_n)\}$, où x_{i1} correspond à la i^e observation de la variable explicative X_1. L'hypothèse 1 implique le modèle statistique suivant pour chacune des variables d'intérêt Y_i :

$$Y_i = \beta_0 + x_{i1}\,\beta_1 + \varepsilon_i \text{ pour } 1 \leq i \leq n;$$

où ε_i est une erreur aléatoire d'espérance nulle, *i.e.* $\mathbb{E}(\varepsilon_i) = 0$ pour $1 \leq i \leq n$.

Une **simplification** est généralement faite en régression linéaire : on considère la ou les variables explicatives comme des constantes et non comme des variables aléatoires. Cette simplification suppose que les variables explicatives sont parfaitement mesurées et connues sans incertitude. La plupart du temps cette simplification n'est pas contraignante parce que l'incertitude sur la variable d'intérêt domine largement l'incertitude sur les variables explicatives. Dans le cas contraire, des modèles de régression plus avancés qui incorporent l'incertitude des variables explicatives pourraient être utilisés, tels les modèles Errors-in-variables.

2.1.2 Estimation des paramètres par les moindres carrés

En utilisant uniquement l'hypothèse 1, il est possible de trouver la droite qui traverse le mieux le nuage de points composé des couples de variable explicative et variable d'intérêt : $\{(x_{i1}, y_1), \ldots, (x_{in}, y_n)\}$. On doit d'abord fixer une mesure d'erreur entre la droite et les points et minimiser cette erreur pour identifier les estimation optimales de β_0 et β_1.

Posons la variable

$$e_i = y_i - \beta_0 - x_{i1}\,\beta_1,$$

correspondant à la i^e erreur observée (également appelée résidu) entre le point y_i et la droite de régression. La méthode classique en régression linéaire pour estimer β_0 et β_1 consiste à

trouver la droite qui minimise la somme des erreurs observées au carré :

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2.$$

Exercice 1

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

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}_1$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}$$

où

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1} \quad \text{et} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Remarque. *La droite ayant comme ordonnée à l'origine $\hat{\beta}_0$ et comme pente $\hat{\beta}_1$ est celle qui minimise la somme des erreurs au carré. Pour que cette droite ait un sens, la relation entre X_1 et Y doit être linéaire. Dans le cas contraire, les méthodes de la section 2.8 permettront de considérer une relation non linéaire entre X_1 et Y.*

Exemple 1: Régression linéaire simple

Pour l'anomalie de température en fonction de la quantité de carbone émise par la combustion de combustible fossile, on trouve que

$$\hat{\beta}_0 = -0.36;$$
$$\hat{\beta}_1 = 0.11.$$

La droite de régression correspondante est illustrée à la figure 2.2. L'augmentation d'une gigatonne de carbone émise par la combustion de combustible fossile augmente en moyenne l'anomalie de température de $0,11°C$.

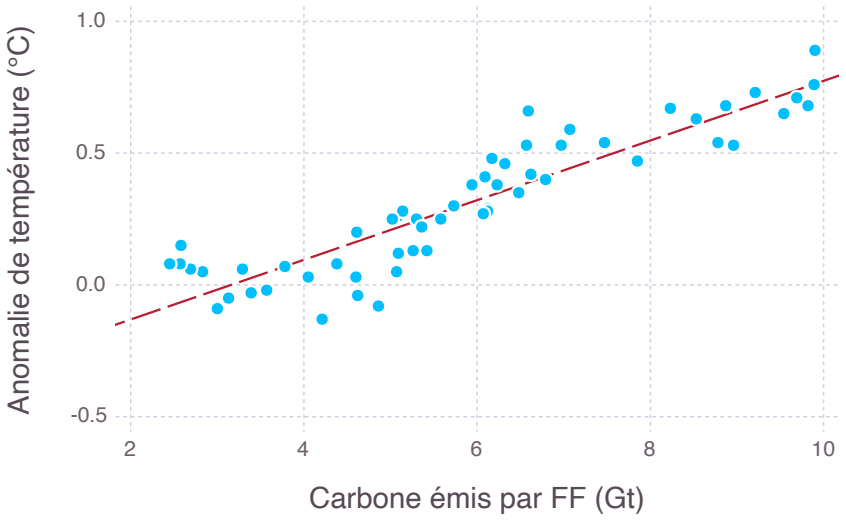


FIGURE 2.2 – Anomalies de température en fonction de la quantité de carbone émise par la combustion de combustibles fossiles superposées à la droite de régression linéaire estimée.

2.2 Régression linéaire multiple

2.2.1 Le modèle statistique

La régression linéaire multiple constitue la généralisation de la régression linéaire simple à p variables explicatives $X_1, X_2, \ldots, X_p$. 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 Y est une fonction linéaire des variables explicatives ;

Hypothèse 1 (Linéarité) :

$$\begin{aligned} \mathbb{E}\left(Y \mid X_1 = x_1, \ldots, X_p = x_p\right) &= \beta_0 + x_1\beta_1 + \ldots x_p\beta_p, \\ &= \beta_0 + \sum_{j=1}^{p} x_j\beta_j; \end{aligned} \tag{2.2}$$

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. C'est la contribution de l'effet linéaire de la variable X_j.

Soit l'échantillon aléatoire de taille n composé des $(p+1)-$uplet suivants :

$$\{(x_{11}, x_{12}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, x_{n2}, \ldots, x_{np}, y_n)\},$$

correspondant aux n variables d'intérêts accompagnées de la valeur des p variables explicatives. La variable x_{ij} correspond à la i^e observation de la j^e variable explicative. L'hypothèse 1 implique le modèle statistique suivant pour chacune des variables d'intérêt Y_i :

$$Y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i; \tag{2.3}$$

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

2.2.2 La notation matricielle

Le modèle de la régression multiple s'écrit plus simplement lorsque la notation matricielle est utilisée. L'équation (2.3) peut s'écrire de la façon compacte suivante :

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

où

$$\boldsymbol{x}_i^\top = \begin{bmatrix} 1 & x_{i1} & \ldots & x_{ip} \end{bmatrix}$$

est un colonne des p variables explicatives pour la variable Y_i augmentées d'un «1» à la première position et où

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

est le vecteur colonne des p coefficients de régressions augmentés de l'ordonnée à l'origine β_0. L'ajout du «1» aux variables explicatives est nécessaire pour intégrer l'ordonnée à l'origine dans le vecteur des coefficients de régression.

Remarque. *Dans les cours d'algèbre linéaire spécifiques au génie, les vecteurs sont dénotés avec une flèche, par exemple $\vec{u}$. En statistique, les vecteurs sont généralement dénotés par les caractères gras, par exemple* $\mathbf{u}$. *Dans la plupart des livres d'apprentissage machine que j'ai consultés, la notation statistique est celle qui est le plus souvent utilisée. C'est pourquoi c'est cette notation qui est adoptée ici.*

La notation matricielle devient encore plus utile pour exprimer le systèmes des n équations linéaires suivant :

$$Y_1 = \boldsymbol{x}_1^\top \boldsymbol{\beta} + \varepsilon_1$$

$$\vdots$$
$$Y_n = \boldsymbol{x}_n^\top \boldsymbol{\beta} + \varepsilon_n$$

en ce système matriciel :

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

avec

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} \text{—} & \boldsymbol{x}_1^\top & \text{—} \\ \text{—} & \boldsymbol{x}_2^\top & \text{—} \\ & \vdots & \\ \text{—} & \boldsymbol{x}_n^\top & \text{—} \end{bmatrix}, \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

La matrice X est généralement appelée la matrice de structure. Elle est constituée par la superposition des n vecteurs de variables explicatives. En l'écrivant sous sa forme développée :

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

on remarque que la première colonne est constituée que de «1».

Remarque. *La notation matricielle introduite dans cette section fonctionne également dans le cas de la régression linéaire simple. Dans ce dernier cas, la matrice de structure X est constituée de n lignes et de deux colonnes : une colonne de «uns» et une colonne correspondant à la seule variable explicative.*

2.2.3 Estimation des paramètres par la méthode des moindres carrés

Les paramètres $\beta_0, \beta_1 \dots, \beta_p$ sont inconnus et doivent être estimés à l'aide de échantillon aléatoire. De façon générale, le système des n équations à $p+1$ inconnus

$$\boldsymbol{y} = X\boldsymbol{\beta}$$

ne possède pas de solutions. S'il possédait une solution, les n points seraient alors parfaitement disposés sur l'hyperplan de régression et on serait plutôt dans le contexte d'une équation déterministe. Par conséquent, le vecteur $\boldsymbol{y}$ ne peut s'écrire comme une combinaison linéaire des variables explicatives $X_1, \dots, X_p$. Autrement dit, le vecteur $\boldsymbol{y}$ n'est pas dans l'espace engendré par les colonnes de la matrice de structure X, dénoté $C(X)$. À l'instar de la régression linéaire simple, les paramètres de l'hyperplan de régression peuvent être estimés en minimisant la somme des erreurs au carré :

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta} \right)^2.$$

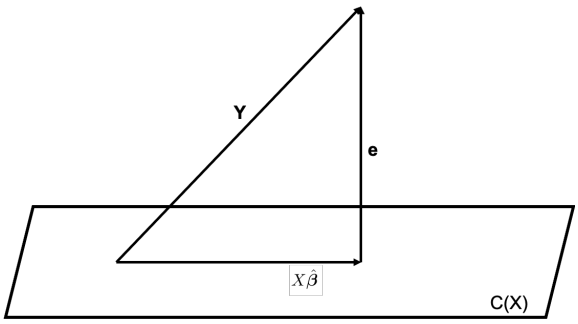


FIGURE 2.3 – Schématisation de la procédure pour identifier $\hat{\boldsymbol{\beta}}$.

De façon vectorielle, la somme des erreurs au carré correspond à la norme du vecteur des résidus au carré :

$$SS_E = ||\boldsymbol{e}||^2 = \boldsymbol{e}^\top \boldsymbol{e}.$$

L'estimation de $\boldsymbol{\beta}$ par les moindres carrés correspond à trouver la combinaison linéaire des colonnes de X qui minimise la norme au carré du vecteur des erreurs.

Avec le schéma illustré à la figure 2.3, on remarque que le vecteur des erreurs ayant la plus petite norme est celui qui est orthogonal à l'espace engendré par les colonnes de X. Cette propriété nous permettra de trouver le vecteur des estimations $\hat{\beta}$.

Estimation par les moindres carrés

Pour n'importe quel vecteur colonne de réels $\boldsymbol{\beta}$ de dimension $(p+1)$, on doit avoir que $(X\boldsymbol{\beta})^\top \boldsymbol{e} = 0$ puisque le vecteur e est orthogonal à l'espace engendré par les colonnes de X. On a alors que

$$\begin{aligned}
&(X\hat{\boldsymbol{\beta}})^\top \boldsymbol{e} = 0 \\
&\Rightarrow (X\hat{\boldsymbol{\beta}})^\top (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = 0 \\
&\Rightarrow \hat{\boldsymbol{\beta}}^\top X^\top (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) = 0 \\
&\Rightarrow \hat{\boldsymbol{\beta}}^\top \underbrace{\left(X^\top \boldsymbol{y} - X^\top X\hat{\boldsymbol{\beta}}\right)}_{\text{doit être nul}} = 0 \\
&\Rightarrow X^\top \boldsymbol{y} - X^\top X\hat{\boldsymbol{\beta}} = 0 \\
&\Rightarrow X^\top X\hat{\boldsymbol{\beta}} = X^\top \boldsymbol{y} \\
&\Rightarrow \hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y}.
\end{aligned}$$

La dernière ligne est obtenue en supposant que la matrice $X^\top X$ est inversible. Nous verrons à la section 2.9 quelles sont les conditions pour que cette matrice soit inversible

et quoi faire si elle ne l'est pas.

L'estimation des paramètres de l'hyperplan de régression est remarquable. D'une part, une formule explicite très simple existe pour ceux-ci :

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y},$$

et, d'autre part, des algorithmes numériques efficaces existent pour le calcul de $\hat{\boldsymbol{\beta}}$.

Exemple 2: Régression linéaire multiple

Pour l'anomalie de température en fonction des différentes composantes du cylcle du carbone, on trouve que

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.60 \\ 0.17 \\ 0.19 \\ 0.13 \\ 0.04 \end{bmatrix}$$

Avec ce modèle, on peut affirmer que l'augmentation d'une gigatonne de carbone émise par la combustion de combustible fossile augmente en moyenne l'anomalie de température de $0,17°C$ si toutes les autres variables sont constantes.

2.2.4 Prédiction 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 *prédiction*. L'estimation de Y_0, dénotée $\hat{Y}_0$ sachant $\boldsymbol{x}_0$ est tout simplement le point correspondant sur le plan de régression :

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

Exemple 3: Prédiction de la variable d'intérêt

Supposons que l'on prévoit pour l'année prochaine l'émission de 10 Gt de carbone par la combustion de combustible fossile et 1.5 Gt par le changement d'occupation des terres ainsi que l'absorption de 1 Gt par les océans et 1 Gt par les strates végétales. On a donc que

$$\boldsymbol{x}_0^\top = \begin{bmatrix} 1 & 10 & 1.5 & -1 & -1 \end{bmatrix}.$$

L'anomalie de température alors attendue est de

$$\hat{Y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}} \approx 1.22°C.$$

2.3 Indice de qualité du modèle de régression linéaire

Lorsque les paramètres du modèle de régression sont estimés, on peut vérifier si la relation linéaire possède un bon pouvoir prédictif pour Y.

2.3.1 Décomposition de la variabilité

Dans cette section, on montre que la variabilité totale de la variable d'intérêt Y

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

peut être décomposée en deux termes : la variabilité expliquée par la régression

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

et la variabilité résiduelle (ou variabilité de l'erreur)

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

On aura donc que

$$SS_T = SS_R + SS_E.$$

Commençons par écrire la variabilité totale comme le carré de la norme du vecteur $\boldsymbol{y}$ centré :

$$SS_T = ||\boldsymbol{y} - \bar{y}\mathbf{1}||^2,$$

où $\mathbf{1}$ dénote le vecteur colonne composé de «un» de dimension appropriée. Dans cette norme, on peut ajouter et retrancher le vecteur des estimations de $\boldsymbol{y}$:

$$\begin{aligned} SS_T &= ||\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1}||^2, \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1}), \\ &= \underbrace{(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\boldsymbol{y} - \hat{\boldsymbol{y}})}_{SS_E} + \underbrace{(\hat{\boldsymbol{y}} - \bar{y}\mathbf{1})^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1})}_{SS_R} + 2(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}). \end{aligned}$$

Il faut donc montrer que

$$(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}) = \boldsymbol{e}^\top (\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}) = 0.$$

Par la construction illustrée à la figure 2.2, on peut constater ce résultat. On peut aussi le démontrer de la façon suivante :

$\boldsymbol{e}^\top \hat{\boldsymbol{y}} = 0$

$$
\begin{aligned}
\boldsymbol{e}^\top \hat{\boldsymbol{y}} &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^\top \hat{\boldsymbol{y}} \\
&= \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^\top \hat{\boldsymbol{y}} \\
&= \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right)^\top X\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^\top X^\top X\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}} - \left\{\left(X^\top X\right)^{-1} X^\top \boldsymbol{y}\right\}^\top X^\top X\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}} - \boldsymbol{y}^\top X\left(X^\top X\right)^{-1} X^\top X\hat{\boldsymbol{\beta}} \\
&= \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}} - \boldsymbol{y}^\top X\hat{\boldsymbol{\beta}} \\
&= 0.
\end{aligned}
$$

$\boldsymbol{e}^\top \mathbf{1} = 0$

$$
\begin{aligned}
& X^\top X\,\hat{\boldsymbol{\beta}} = X^\top \boldsymbol{y} \\
& \Rightarrow X^\top \boldsymbol{y} - X^\top X\,\hat{\boldsymbol{\beta}} = \mathbf{0} \\
& \Rightarrow X^\top \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\right) = \mathbf{0} \\
& \Rightarrow X^\top \boldsymbol{e} = \mathbf{0} \\
& \Rightarrow \mathbf{1}^\top \boldsymbol{e} = 0.
\end{aligned}
$$

Exercice 2

Montrer ou expliquer comment on passe de l'avant dernière ligne à la dernière ligne de la démonstration précédente.

2.3.2 Le coefficient de détermination

La décomposition de la variabilité précédente permet de définir un indice de qualité naturel pour le modèle de régression : le coefficient de détermination R^2. Il correspond à la proportion de la variabilité expliquée par le modèle de régression sur la variabilité totale :

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

La valeur du coefficient de détermination est comprise entre 0 et 1. Un modèle de régression qui ne possède pas de pouvoir prédictif supérieur à la moyenne de Y aura un R^2 nul. Un modèle pour lequel on peut prédire parfaitement la valeur de Y aura un R^2 de 1. Par conséquent, plus le R^2 est près de 1, meilleur est le modèle de régression.

Exemple 4: Coefficient de détermination

Dans le cas de la régression linéaire multiple de l'exemple 2, on a que

$$SS_T \approx 3.85;$$
$$SS_R \approx 3.33;$$
$$SS_E \approx 0.52.$$

Le coefficient de détermination est donc de $R^2 \approx 0.87$. Autrement dit, 87% de la variabilité de l'anomalie de température est expliquée par les 4 composantes du cycle du carbone considérées.

2.4 Hypothèses supplémentaires

Jusqu'à maintenant, nous n'avons eu recours qu'à l'hypothèse 1 pour développer le modèle de régression. Nous avons pu trouver l'hyperplan qui minimise la somme des erreurs au carré et calculer le coefficient de détermination de la régression. En supposant des hypothèses supplémentaires, des propriétés additionnelles du modèle de régression linéaire peuvent être développées.

La deuxième hypothèse concerne la variance de l'erreur et s'énonce comme suit :

Hypothèse 2 (Homoscédasticité de la variance) :

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

Pour chacune des observations, on suppose que la variance de l'erreur est la même. La troisième hypothèse concerne l'indépendance des variables aléatoires Y_i, ce qui implique l'indépendance des erreurs :

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

On suppose donc que chacune des observations ont été obtenues de façon indépendantes.

Les hypothèses 1 et 2 nous indiquent que les erreurs sont d'espérance nulle et de variance σ^2. La dernière hypothèse concerne la distribution de ces erreurs :

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

Nous verrons à la section 2.7 comment les quatre hypothèses peuvent être vérifiées.

2.5 Propriétés des estimateurs

Avant de développer les propriétés des estimateurs, il convient de rappeler la définition de la matrice de covariance d'un vecteur aléatoire.

2.5.1 Préliminaire : la matrice de covariance d'un vecteur aléatoire

Pour un vecteur aléatoire, par exemple le vecteur colonne $\boldsymbol{Y}$ de taille n, sa matrice de covariance, dénotée par $\mathbb{V}ar(\boldsymbol{Y})$, est définie comme suit :

$$\mathbb{V}ar(\boldsymbol{Y}) = \mathbb{C}ov(\boldsymbol{Y}, \boldsymbol{Y}) = \mathbb{E}\left[\{\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\}\{\boldsymbol{Y} - \mathbb{E}(\boldsymbol{Y})\}^\top\right].$$

En développant le produit, on touve que

$$\mathbb{V}ar(\boldsymbol{Y}) = \mathbb{E}\left(\boldsymbol{Y}\boldsymbol{Y}^\top\right) - \mathbb{E}\left(\boldsymbol{Y}\right)\mathbb{E}\left(\boldsymbol{Y}^\top\right)$$

$$= \begin{bmatrix} \mathbb{E}\left(Y_1^2\right) - \{\mathbb{E}\left(Y_1\right)\}^2 & \mathbb{E}\left(Y_1Y_2\right) - \mathbb{E}\left(Y_1\right)\mathbb{E}\left(Y_2\right) & \dots & \mathbb{E}\left(Y_1Y_n\right) - \mathbb{E}\left(Y_1\right)\mathbb{E}\left(Y_n\right) \\ \mathbb{E}\left(Y_2Y_1\right) - \mathbb{E}\left(Y_2\right)\mathbb{E}\left(Y_1\right) & \mathbb{E}\left(Y_2^2\right) - \{\mathbb{E}\left(Y_2\right)\}^2 & \dots & \mathbb{E}\left(Y_2Y_n\right) - \mathbb{E}\left(Y_2\right)\mathbb{E}\left(Y_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}\left(Y_nY_1\right) - \mathbb{E}\left(Y_n\right)\mathbb{E}\left(Y_1\right) & \mathbb{E}\left(Y_nY_2\right) - \mathbb{E}\left(Y_n\right)\mathbb{E}\left(Y_2\right) & \dots & \mathbb{E}\left(Y_n^2\right) - \{\mathbb{E}\left(Y_n\right)\}^2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{V}ar(Y_1) & \mathbb{C}ov(Y_1, Y_2) & \dots & \mathbb{C}ov(Y_1, Y_n) \\ \mathbb{C}ov(Y_2, Y_1) & \mathbb{V}ar(Y_2) & \dots & \mathbb{C}ov(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}ov(Y_n, Y_1) & \mathbb{C}ov(Y_n, Y_2) & \dots & \mathbb{V}ar(Y_n) \end{bmatrix}$$

La matrice de covariance de $\boldsymbol{Y}$ est donc une matrice
- de dimensions $(n \times n)$;
- symétrique, puisque $\mathbb{C}ov(Y_i, Y_j) = \mathbb{C}ov(Y_j, Y_i)$;
- dont les éléments sur la diagonale correspondent aux variances de chacune des composantes du vecteur ;
- et dont les éléments hors diagonale correspondent aux différentes covariances entre les composantes du vecteur.

Exemple 5: Matrice de covariance des variables d'intérêts

Avec les hypothèses 2 et 3, la matrice de covariance des variables d'intérêts $\boldsymbol{Y}$ est la suivante :

$$\mathbb{V}ar(\boldsymbol{Y}) = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n,$$

où I_n dénote la matrice identité de dimension n.

Soit la matrice A réelle de dimensions $(d \times n)$. On pourrait montrer que la matrice de covariance de l'ensemble des d combinaisons linéaires $A\boldsymbol{Y}$ s'exprime de la façon suivante :

$$\mathbb{V}ar(A\boldsymbol{Y}) = A\mathbb{V}ar(\boldsymbol{Y})A^\top.$$

2.5.2 Estimateur de la variance de l'erreur

En utilisant les hypothèses 1 à 3, on pourrait montrer qu'une estimation non biaisée de la variance de l'erreur σ^2 est donné par :

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2. \tag{2.4}$$

Le $(p+1)$ qui est retiré au dénominateur provient de la perte des degrés de liberté de liberté due à l'estimation des $(p+1)$ coefficients de régression. Dans un cours sur la régression, c'est le genre de résultat que l'on démontrerait.

Exemple 6: Estimation de l'erreur

Dans le cas de la régression linéaire de l'exemple 2, on a que

$$\hat{\sigma}^2 \approx 0.0099.$$

2.5.3 Estimateur des coefficients de régression

En supposant que les variables explications sont des constantes et en supposant l'hypothèse 1 satisfaite, on peut montrer que l'estimateur $\hat{\boldsymbol{\beta}}$ est sans biais, c'est-à-dire que $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}\left\{\left(X^\top X\right)^{-1} X^\top \boldsymbol{Y}\right\} \\
&= \left(X^\top X\right)^{-1} X^\top \mathbb{E}(\boldsymbol{Y}) \quad \text{(par la simplication)}\,; \\
&= \left(X^\top X\right)^{-1} X^\top X\boldsymbol{\beta} \quad \text{(par l'Hypothèse 1)}\,; \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Ce n'est pas tous les estimateurs de $\boldsymbol{\beta}$ qui sont sans biais. Nous en verrons d'ailleurs plusieurs qui sont biaisés dans les prochains chapitres lorsque nous étudierons la régression Ridge, la régression Lasso et plus généralement la régression bayésienne.

En utilisant en plus les hypothèses 2 et 3, on peut montrer que la matrice de covariance de $\hat{\boldsymbol{\beta}}$ est la suivante :

$$\mathbb{V}ar(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(X^\top X\right)^{-1}.$$

Exercice 3

Montrer que $\mathbb{V}ar(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(X^\top X\right)^{-1}$.

Dans un cours sur la régression, on pourrait aussi montrer que la relation suivante entre la vraie valeur du coefficient et régression et son estimation :

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

où v_{j+1} correspond au $(j+1)^e$ élément sur la diagonale de la matrice $(X^\top X)^{-1}$ et $t_\nu(\mu, \sigma)$ correspond à la loi de Student à ν degrés de liberté, de localisation μ et d'échelle σ. On peut développer avec cette équation un test d'hypothèses pour vérifier si la variable explicative X_j possède un pouvoir prédictif significatif :

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

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

On peut également développer un intervalle de confiance pour la vraie valeur β_j.

Exercice 4

Montrer que l'intervalle de confiance de niveau $1-\alpha$ est donné par l'équation suivante :

$$\left[\hat{\beta}_j - q(\alpha/2)\sqrt{\hat{\sigma}^2 v_{j+1}}\ ,\ \hat{\beta}_j + q(\alpha/2)\sqrt{\hat{\sigma}^2 v_{j+1}}\right],$$

où $q(\alpha)$ est le quantile d'ordre $(1-\alpha)$ de la loi de $t_{(n-p-1)}(0,1)$.
Aussi, que peut-on conclure si la valeur 0 est incluse dans l'intervalle de confiance ?

Exemple 7: Intervalle de confiance pour β_1

La figure 2.4 illustre la distribution de $\hat{\beta}_1$ ainsi que l'intervalle de confiance de niveau 95% correspondant :

$$\mathbb{P}\left\{\beta_1 \in (0.12\,,\,0.22)\right\} = 0.95.$$

Remarque. *Même dans le cas où les résidus ne sont pas distribués selon la loi normale, l'intervalle de confiance ainsi obtenu constitue une très bonne approximation. C'est pourquoi il est automatiquement fourni dans toute bonne librairie de régression linéaire.*

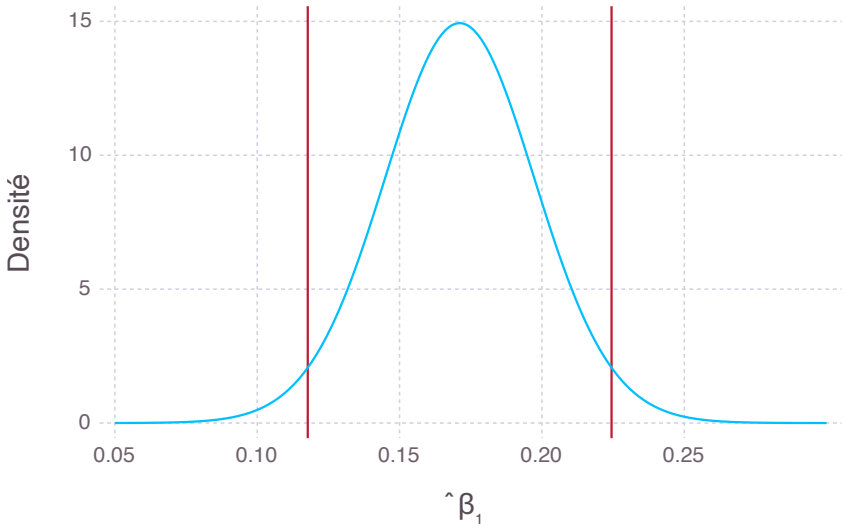


FIGURE 2.4 – Densité de $\hat{\beta}_1$ et intervalle de confiance de niveau 95% correspondant.

2.5.4 Prédictions

Le fait que l'estimateur de $\boldsymbol{\beta}$ obtenu par la méthode des moindres carrés soit sans biais implique que la prédiction de Y_0 pour le vecteur de variables explicatives $\boldsymbol{x}_0$ par $\hat{Y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$ est aussi sans biais :

$$\mathbb{E}(\hat{Y}_0) = \mathbb{E}(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^\top \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^\top \boldsymbol{\beta} = Y_0.$$

La variance des prédictions est obtenue de la façon suivante :

$$\mathbb{V}ar(\hat{Y}_0) = \mathbb{V}ar(\boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}) = \boldsymbol{x}_0^\top \mathbb{V}ar(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_0 = \sigma^2 \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0.$$

On peut alors montrer que

$$\frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2 \left\{1 + \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0\right\}}} \sim t_{(n-p-1)}(0,1).$$

Exercice 5

Montrer que l'intervalle de confiance de niveau $1-\alpha$ est donné par l'équation suivante :

$$\left[\hat{Y}_0 - q(\alpha/2)\sqrt{\hat{\sigma}^2 \left\{1 + \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0\right\}} \ , \ \hat{Y}_0 + q(\alpha/2)\sqrt{\hat{\sigma}^2 \left\{1 + \boldsymbol{x}_0^\top (X^\top X)^{-1} \boldsymbol{x}_0\right\}}\right],$$

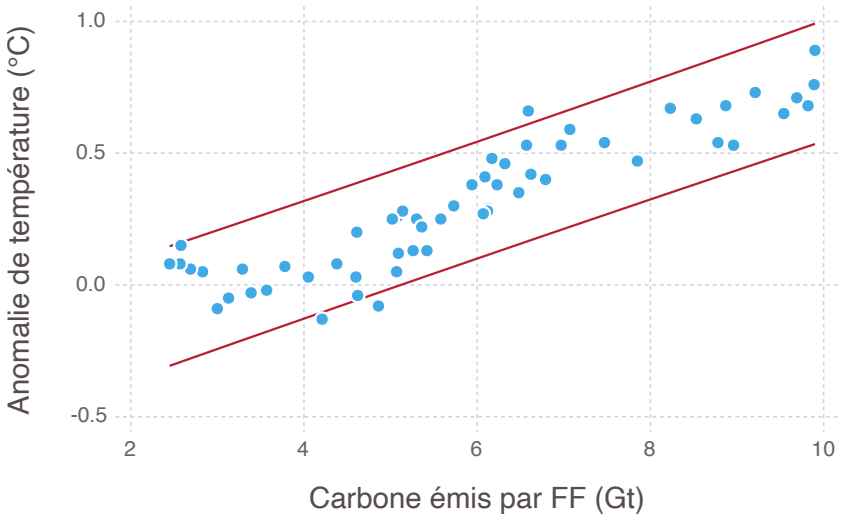


FIGURE 2.5 – Intervalle de confiance à 95% pour les prédictions effectuées avec le modèle de régression linéaire simple n'utilisant que X_1.

où $q(\alpha)$ est le quantile d'ordre $(1-\alpha)$ de la loi de $t_{(n-p-1)}(0,1)$.
Aussi, que peut-on conclure si la valeur observée n'est pas incluse dans l'intervalle de confiance ?

Exemple 8: Intervalle de prédiction

Revenons à l'exemple de la régression linéaire simple en utilisant que X_1 pour prédire l'anomalie de température Y. L'intervalle de confiance de niveau 95% pour les prédictions du modèles en fonction de x_1 est illustré à la figure 2.5.

En apprentissage machine, une mesure populaire pour évaluer la qualité des prédictions d'un modèle est l'erreur quadratique moyenne (*mean squared error*) définie ainsi :

$$\text{MSE}(\hat{Y}_0) = \mathbb{E}\left\{(\hat{Y}_0 - Y_0)^2\right\}$$

L'erreur quadratique moyenne peut être estimée avec les erreurs observées sur un ensemble de validation ou de test (voir la section 2.11) mais on peut aussi l'estimer à l'aide des propriétés du modèle. On peut montrer que

$$\text{MSE}(\hat{Y}_0) = \mathbb{V}ar(\hat{Y}_0) + \left\{\text{Biais}(\hat{Y}_0)\right\}^2.$$

Alors, dans le cas de l'estimation par les moindres carrés,

$$\mathrm{MSE}(\hat{Y}_0) = \mathbb{V}ar(\hat{Y}_0) = \sigma^2 \, \boldsymbol{x}_0^\top \left(X^\top X\right)^{-1} \boldsymbol{x}_0.$$

Exercice 6

Montrer que l'erreur quadratique moyenne peut s'exprimer ainsi :

$$\mathrm{MSE}(\hat{Y}_0) = \mathbb{V}ar(\hat{Y}_0) + \left\{\mathrm{Biais}(\hat{Y}_0)\right\}^2.$$

Indice : ajouter et retrancher $\mathbb{E}(\hat{Y}_0)$.

Un thèorème célèbre en statistique stipule que de toutes les estimations non biaisées de Y_0, celle qui possède la plus petite erreur quadratique moyenne est celle estimée avec $\hat{\boldsymbol{\beta}}$, l'estimation de $\boldsymbol{\beta}$ obtenue par les moindres carrés. Il s'agit du théorème de Gauss-Markov. Autrement dit, si on cherche une estimation non biaisée de Y_0, alors il n'existe pas d'estimation plus précise au sens du MSE que celle obtenue avec $\hat{\boldsymbol{\beta}}$. C'est un résultat formidable !

Remarque. *Parfois, il peut être avantageux de tolérer un petit biais pour diminuer la variance de la prédiction et ultimement réduire l'erreur quadratique moyenne. C'est ce que tente de faire la régression Ridge et Lasso que nous verrons plus tard.*

2.6 Test sur l'importance de la régression

Tester l'importance de la régression consiste à vérifier si le modèle de régression explique une partie significative de la variabilité de Y. Ceci 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 $j \in \{1, \ldots, p\}$.

Intuitivement, on s'attend à ce que le modèle de régression soit significatif si la variabilité expliquée par le modèle SS_R est grande par rapport à la variabilité résiduelle SS_E. La statistique du test exploite justement ces deux quantités et elle s'exprime ainsi :

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

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

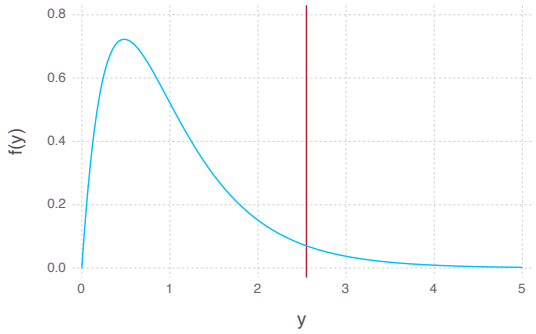


FIGURE 2.6 – Densité de la loi de $Fisher(4, 52)$. La ligne rouge correspond au seuil de quantile d'ordre 95%.

Exemple 9: Test de l'importance de la régression

Dans le cas de la régression linéaire de l'exemple 2, on a que $p = 4$ et $n = 57$. Si l'hypothèse nulle est vraie, alors la statistique est distribuée selon la loi de $Fisher(4, 52)$ dont la densité est illustrée à la figure 2.6. La statistique observée du test est de 83.7, ce qui est supérieur à 2.55, le quantile d'ordre 95% de la loi de Fisher. Puisque cette statistique est très peu commune lorsque l'hypothèse nulle est vraie, cela suggère de rejeter l'hypothèse nulle.

2.7 Vérification des hypothèses de la régression linéaire

Nous avons vu que quatre hypothèses doivent être satisfaites afin que toutes les propriétés mentionnées jusqu'à maintenant de la régression linéaire puissent s'appliquer. Dans cette section, nous verrons comment vérifier si ces hypothèses sont satisfaites.

2.7.1 Vérification des hypothèses 1 et 2 pour la régression linéaire simple

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.). Si la dispersion des points autour de la droite ne semble pas constante le long de la droite, cela suggère que l'hypothèse 2 n'est pas satisfaite.

2.7.2 Vérification des hypothèses 1 et 2 pour la régression linéaire multiple

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. Le nuage de points illustrant les résidus e_i en fonction des estimations de Y_i, soit $\hat{Y}_i = \boldsymbol{x}_i\hat{\boldsymbol{\beta}}$ permet de vérifier les hypothèses 1 et 2. Si les deux hypothèses sont satisfaites, alors les résidus devraient former un nuage de points de forme rectangulaire centré autour de l'axe des abscisses. Si le nuage de point n'est pas centré autour de 0, cela suggère que l'hypothèse 1 n'est pas satisfaite et si le nuage a plutôt une forme conique, cela suggère que l'hypothèse 2 n'est pas satisfaite.

Remarque. *Cette méthode étudiant les résidus fonctionne également pour la régression linéaire simple.*

Exemple 10: Vérification des hypothèses 1 et 2

La figure 2.7 illustrent les résidus en fonction des estimations pour le modèle de l'exemple 2. Bien que le nuage de points soit bien distribué autour de 0, la forme n'est pas parfaite : on peut remarquer une légère diminution de la variance lorsque $\hat{Y}$ augmente. Il est tout de même raisonnable de supposer que les hypothèses de linéarité et d'homoscédasticité sont satisfaites.

2.7.3 Vérification de l'hypothèse 3

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.

2.7.4 Vérification de l'hypothèse 4

L'hypothèse de normalité des résidus peut être vérifié à l'aide d'un diagramme quantile-quantile comparant les quantiles empiriques des résidus aux quantiles de la loi normale. Si les points du diagramme s'alignent sur une droite, cela signifie que la loi normale est une distribution raisonnable pour ceux-ci.

Exemple 11

La droite de Henry pour les résidus du modèle de régression est illsutrée à la figure 2.8. Puisque tous les points s'alignent presque parfaitement sur la droite, l'hypothèse de normalité des erreurs semble tout à fait raisonnable.

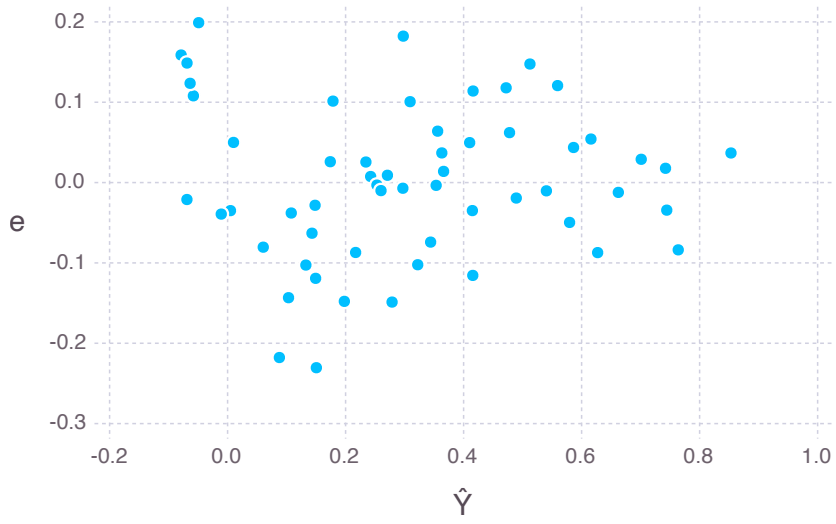


FIGURE 2.7 – Résidus en fonction de estimations pour la régression linéaire multiple utilisant X_1, X_2, X_3 et X_4.

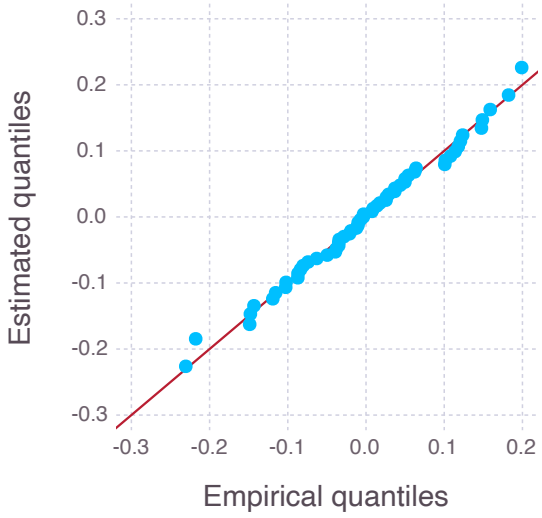


FIGURE 2.8 – Droite de Henry pour les résidus du modèle de régression multiple.

2.7.5 Que faire si l'une ou l'autre des hypothèses n'est pas satisfaite ?

Si l'hypothèse 1 supposant la linéarité n'est pas satisfaite, alors la régression polynomiale peut être tentée (voir section 2.8). Aussi, les modèles linéaires généralisés que nous verrons au prochain chapitre peuvent constituer une alternative.

Si l'hypothèse 2 d'homoscédasticité n'est pas satisfaite, alors une transformation de la variable d'intérêt Y peut éventuellement régler le problème. Par exemple, on peut tenter la transformation suivante :

$$\ln Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.5}$$

Les techniques usuelles de la régression linéaire sont tout simplement utilisées sur les données transformées $\{(x_i, \ln y_i) : i = 1, \ldots, n\}$. Il convient de vérifier si la transformation a réglé le problème. Les transformations les plus courantes sont $\ln Y$, $\sqrt{Y}$ et Y^2.

Si l'hypothèse 4 de normalité des résidus n'est pas satisfaite, soit une transformation de la variable d'intérêt Y ou soit l'utilisation des modèles linéaires généralisés peut éventuellement régler le problème. Cela dit, cette hypothèse est nécessaire pour effectuer des tests d'hypothèses et construire des intervalles de confiance. Alors si c'est la prédiction qui vous intéresse plus que l'estimation, alors cette hypothèses est peut-être un peu moins importante.

2.7.6 Identification des données suspectes

On appelle *données suspectes* les observations peu communes selon le modèle de régression ajusté. Une donnée suspecte peut correspondre à une donnée aberrante (erreur de mesure, erreur de transcription, etc.) ou à une donnée extrême réellement observé. Seul un expert connaissant les données pourra déterminer si une donnée suspecte est une donnée aberrante ou un extrême. Les données suspectes présentes dans un jeu de données peuvent occasionner d'importantes conséquences dans une analyse de régression. C'est pourquoi il est important de les identifier de façon automatique avec un critère afin d'étudier leurs répercussions sur le modèle de régression.

Si on suppose le modèle de régression comme valide, c'est à dire que toutes les hypothèses sont satisfaites, alors on peut définir une méthode pour détecter les données suspectes. Soit les résidus *studentisés* s_i définis de la façon suivante :

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

où h_i est le i^e élément de la diagonale de la matrice $H = X\left(X^\top X\right)^{-1} X^\top$. Si $|s_i| > 3$, alors on définit y_i comme une valeur suspecte qu'il convient d'étudier.

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 le fait que la donnée suspecte constitue bien une erreur, la décision est alors plus difficile et repose sur des considérations pragmatiques.

2.8 Transformation des variables explicatives

2.8.1 Inclusion de variable qualitatives

Des variables explicatives qualitatives peuvent être intégrées dans un modèle de régression linéaire. Par exemple, la variable *couleur de la voiture* peut être introduite dans un modèle de régression pour calculer la prime d'assurance. Cette variable qualitative peut être intégrée dans le modèle en utilisant une combinaison de variables indicatrices. Supposons d'abord que la variable ne peut prendre que deux couleurs : rouge ou vert. Alors la variable indicatrice suivante peut être définie :

$$X = \begin{cases} 0 & \text{si la couleur est rouge;} \\ 1 & \text{si la couleur est vert.} \end{cases}$$

Dans le cas où il y aurait trois couleur (par exemple rouge, vert et bleu), deux variables indicatrices sont nécessaires pour inclure cette variable explicative qualitative :

X_1	X_2	
0	0	si la couleur est rouge ;
1	0	si la couleur est vert ;
0	1	si la couleur est bleu.

On peut procéder de façon analogue pour une variable qualitative à k catégories, $k-1$ variables indicatrices seront nécessaires pour correctement encoder cette information dans le modèle de régression.

2.8.2 Régression polynomiale

Le modèle de régression linéaire multiple 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 de dimension 2 :

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

où $X_1 = X$ et $X_2 = X^2$. L'estimation par les moindres carrés développée pour la régression linéaire peut être utilisée avec cette dernière formulation. On peut ainsi troquer les degrés du polynôme de la relation entre X et Y pour des dimensions additionnelles dans le modèle de régression linéaire multiple. Cela est vrai pour n'importe quel degré de polynôme.

2.9 Multicolinéarité

2.9.1 Définition et cause

Pour développer l'estimateur des coefficients de $\boldsymbol{\beta}$ par la méthode des moindres carrés, nous avons supposé que la matrice $\left(X^\top X\right)^{-1}$ était inversible pour obtenir :

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

Si la matrice n'est pas inversible, alors il est impossible d'obtenir le vecteur des estimations des coefficients de régression $\hat{\boldsymbol{\beta}}$. La matrice $\left(X^\top X\right)^{-1}$ n'est pas inversible lorsque son déterminant est nul. Cela se produit lorsqu'une colonne est une combinaison linéaire des autres colonnes. Cette situation n'est pas vraiment problématique. En effet, lors de l'estimation des paramètres avec votre librairie favorite, le calcul de l'inverse sera impossible et vous recevrez un message d'erreur. Il suffira de retirer la ou les colonnes redondantes de la matrice de structure X.

Le problème que l'on appelle multicolinéarité survient lorsqu'une colonne de X est *presque* une combinaison linéaire des autres colonnes, mais pas exactement. Dans ce cas, le déterminant de la matrice ne sera pas nul, mais presque nul. Votre librairie favorite ne détectera pas de problème et effectuera une opération numériquement dangereuse : diviser par un très petit nombre.

D'un point de vue statistique, la multicolinéarité survient lorsque l'information apportée par les différentes variables explicatives est redondante. C'est le cas par exemple lorsque l'année de naissance et l'âge d'un patient en mois sont des variables explicatives d'un modèle de régression linéaire. L'information est probablement suffisamment redondante pour causer des problèmes de multicolinéarité. Parfois la situation est plus subtile car c'est une combinaison de variables qui est redondante avec une autre variable.

2.9.2 Conséquences

La multicolinéarité affecte la précision du calcul de l'inverse de la matrice $\left(X^\top X\right)$. Le premier effet notable concerne la précision des estimations des coefficients de régression $\hat{\boldsymbol{\beta}}$. Puisque les variables sont redondantes, il n'existe plus une forme unique pour minimiser la somme des erreurs au carré. Autrement dit, plusieurs vecteurs $\hat{\boldsymbol{\beta}}$ peuvent donner le même SS_E. Cela se manifeste par une grande variance des estimations. Le problème d'identifiabilité des paramètres empêche également l'interprétation physique des estimations des coefficients de régression.

L'instabilité numérique occasionnée par le calcul de l'inverse de $\left(X^\top X\right)$ a aussi pour conséquence d'invalider les intervalles de confiance et les tests d'hypothèses développés dans les sections précédentes. Les intervalles de confiance risquent en effet d'être exagérément grands puisqu'ils impliquent d'une façon ou une autre le terme $\left(X^\top X\right)^{-1}$ dont le calcul numérique est très instable.

Néanmoins, la multicolinéarité influence peu les prédictions ponctuelles obtenues avec le modèle de régression, *i.e.* $\hat{Y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$. La raison est que même s'il existe plusieurs formulations possibles du modèle, le modèle retenu constitue un des modèles qui minimise la somme des erreurs au carré.

2.9.3 Détection

Pour vérifier la présence de multicolinéarité dans un problème de régression, la régression linéaire peut être utilisée. On peut en effet vérifier si la variable explicative X_1 pour être expliquée par les autres variables explicatives dénotée par $X_{-1} = [X_2, \ldots, X_p]$. La variable d'intérêt de la régression est maintenant X_1. Le coefficient de détermination de cette régression R_1^2 est utilisée pour déterminer le niveau de la multicolinéarité. Si le coefficient de détermination est petit, alors il est raisonnable de supposer que la variable X_1 n'est pas linéairement dépendante des autres. Au contraire, si le coefficient de détermination est élevé, la variable X_1 est redondante par rapport aux autres variables et elle risque de causer de la multicolinéarité.

Pour définir un seuil pour lequel la muticolinéarité devient problématique, il est suggéré dans la littérature de calculer le facteur d'inflation de la variance VIF (pour *Variance Inflation Factor*). Pour la variable X_1, le VIF est défini de la façon suivante :

$$\text{VIF}_1 = \frac{1}{1 - R_1^2}.$$

S'il n'y a pas de multicolinéarité, le facteur d'inflation de la variance est de 1. Il tend vers l'infini dans le cas multicolinéaire. La multicolinéarité devient problématique lorsque le facteur d'inflation de la variance est supérieur à 10. Bien que cette valeur est arbitraire, elle donne néanmoins un seuil pour lequel il devient important de se soucier de la multicolinéarité. Le facteur d'inflation de la variance doit être calculé pour chacune des variables explicatives ($\text{VIF}_j : 1 \leq j \leq p$).

Exemple 12

Les facteurs d'inflation de la variance pour le modèle de régression multiple sont les suivants :

$$\text{VIF} = \begin{bmatrix} 17.7 & 2.3 & 12.8 & 1.3 \end{bmatrix}^\top.$$

Il y a présence de multicolinéarité dans les variables explicatives car au les VIF associées aux variables X_1 et X_3 sont supérieurs à 10. En particulier, il est possible de retrouver les valeurs de X_1 assez précisément en utilisant les autres variables explicatives. En effet, on peut prédire X_1 avec une régression utilisant les autres variables explicatives dont le coefficient de détermination est aussi élevé que 0.94.

2.9.4 Solutions

Dans le cas où l'on contrôle l'acquisition des données, il est parfois possible de limiter la multicolinéarité en recueillant les données dans un contexte différent. Cette situation se présente toutefois rarement avec les jeux de données modernes observationnels. Dans ce dernier cas, l'élimination des variables explicatives redondantes est une solution privilégiée ainsi que l'utilisation des modèles de régression biaisés (régression Ridge, régression Lasso et régression bayésienne).

2.10 Sélection des variables explicatives

2.10.1 Motivation

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 explicatives 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.

En somme, la recherche du sous-ensemble des meilleures variables explicatives est principalement motivée par trois raisons ;

Qualité des prédictions : en ne conservant que les meilleures variables explicatives, on ne gaspille pas l'information de nos données pour estimer des paramètres de régression non-significatifs. La variance des estimateurs est ainsi réduite et cela à un effet sur la précision des prédictions.

Interprétation : en ne considérant que les variables avec les effets les plus importants sur Y, il peut être plus facile d'interpréter et de comprendre le modèle.

Limiter la multicolinéarité : si on ne conserve que les variables les plus importantes, le risque de multicolinéarité est moindre mais encore présent.

2.10.2 Comparaison de modèles

La sélection du meilleur sous-ensemble de variables explicatives s'effectue par la comparaison des différents modèles de régression linéaire induit par les variables explicatives

choisies. 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 unidimensionnel avec la *meilleure* variable explicative. Un des critère possible pour la comparaison, et probablement le plus populaire, est le coefficient de détermination. Le meilleur modèle sera parmi ces deux sera celui qui possède le plus grand coefficient de détermination.

Supposons maintenant que l'on veuille choisir entre les trois modèles suivants :

$$\mathcal{M}_1 : Y = \beta_0 + X_1\beta_1 + \varepsilon;$$
$$\mathcal{M}_2 : Y = \beta_0 + X_2\beta_2 + \varepsilon;$$
$$\mathcal{M}_3 : Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

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. En effet, le coefficient de détermination du modèle $\mathcal{M}_3$ ne peut pas être plus petit que celui du modèle $\mathcal{M}_1$ et $\mathcal{M}_2$ même si l'une ou l'autre des variables X_1 ou X_2 est non significative (à moins d'instabilité numérique). Il convient alors d'utiliser le coefficient de détermination ajusté :

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

À l'instar du coefficient de détermination, le R^2_{aj} est borné entre $-\infty$ 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}.

En somme, la sélection des variables explicatives s'effectuent en calculant le coefficient de détermination ajusté pour les modèles construits à partir de tous les sous-ensembles possibles des variables explicatives. Le modèle retenu est celui qui maximise le coefficient de détermination ajusté.

Exemple 13

Avec les 4 variables explicatives, il y a 16 sous-modèles possibles. Le tableau 2.1 compile les coefficients de régression ajusté pour chacun des sous-modèles en ordre décroissant. Dans cet exemple, le meilleur modèle correspond à celui avec toutes les variables explicatives.

2.10.3 Méthode pas-à-pas ascendante

La recherche exhaustive du meilleur modèle telle que décrite à la section précédente n'est possible qu'avec un nombre limité de variables explicatives. En utilisant le très effi-

TABLE 2.1 – Coefficient de détermination ajusté pour chacun des 16 sous-modèles de régression possibles.

FF	LUC	$Ocean$	$Land$	R^2_{aj}
1	1	1	1	0.858023
1	1	0	1	0.854417
1	0	0	1	0.846293
1	0	1	1	0.843668
1	1	0	0	0.834963
1	1	1	0	0.834311
1	0	0	0	0.828575
1	0	1	0	0.825495
0	1	1	0	0.750086
0	1	1	1	0.749665
0	0	1	0	0.742698
0	0	1	1	0.740949
0	1	0	1	0.270453
0	1	0	0	0.262078
0	0	0	1	0.073492
0	0	0	0	0.017544

cace algorithme *leaps and bounds* développé par Furnival & Wilson (1974), il est possible d'effectuer la recherche exhaustive du meilleur modèle pour plus de 30 variables explicatives, ce qui fait 2^{30} modèles possibles. Pour plus de 30 variables explicatives, la recherche exhaustive n'est plus possible. Il faut alors recourir à des approches algorithmiques. L'une de ces approches est la régression pas-à-pas ascendante.

Dans la régression pas-à-pas ascendante (ou *forward stepwise regression* en anglais), à chacun des pas, l'ajout d'une variable explicative dans le modèle est considéré. La variable qui augmente le plus le critère choisi est intégrée dans le modèle ; ce critère est souvent le coefficient de détermination ajusté ou le score Z. Cette procédure est répétée jusqu'à ce que l'ajout d'une variable n'augmente plus le critère de qualité. La régression pas-à-pas ascendante produit une suite de modèles emboîtés : les modèles de plus petites dimensions sont inclus dans les modèles de plus grandes dimensions.

Les méthodes de régression algorithmiques telle que la régression pas-à-pas ascendante ne garantit pas de sélectionner le meilleur sous-ensemble possible des variables explicatives. En effet, le meilleur modèle de dimension deux peut ne pas inclure le meilleur modèle de dimension 1, ce qui est un problème pour la procédure algorithmique qui produit une suite de modèles emboîtés. Bien que le modèle sélectionné par la procédure itérative peut ne pas être le meilleur, il constitue néanmoins un bon modèle.

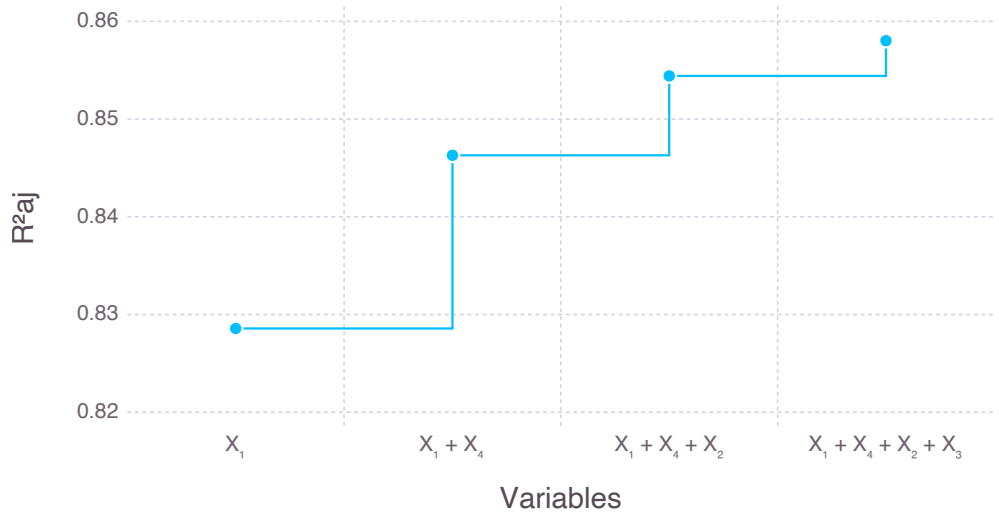


FIGURE 2.9 – Variables explicatives de chacune des itérations de la procédure de régression pas-à-pas ascendante et coefficients de détermination ajusté correspondants.

Exemple 14: Régression pas-à-pas ascendante

La figure2.9 illustre les itérations de la procédure pas-à-pas ascendante pour le problème de la régression multiple avec les 4 composantes du cycle du carbone. À la première itération, le meilleur modèle unidimensionnel est celui utilisant la variable X_1. À la deuxième itération, le meilleure modèle bidimensionnel qui inclut la variable X_1 est obtenu en ajoutant la variable X_4. Cette procédure est répétée jusqu'à ce que le coefficient de détermination ajusté n'augmente plus. Dans le cas de cet exemple, toutes les variables explicatives sont sélectionnées.

2.11 Qualité des prédictions

2.11.1 Motivation

Jusqu'à maintenant, nous avons estimé la qualité des modèles de régression sur la base de l'*estimation statistique*. Cette évaluation utilise astucieusement les données pour estimer les paramètres et évaluer la qualité du modèle. Les modèles statistiques peuvent aussi être évalués sur la base de la prédiction. Ce type d'évaluation est maintenant très répandue, notamment en raison de l'existence de modèles prédictifs non statistique. L'évaluation

d'un modèle statistique par l'estimation ou la prédiction constituent les deux côtés d'une médaille ; elles partagent certes le modèle statistique mais le mode d'évaluation est bien différent.

2.11.2 Partitionnement du jeu de données

Pour évaluer la qualité d'un modèle de régression linéaire sur la base de la prédiction dans un cas idéal, l'ensemble des données devraient être partitionnées en trois sous-ensembles tel qu'illustré à la figure 2.10a :

Ensemble d'entraînement : Données utilisées pour l'estimation des paramètres du modèle.

Ensemble de validation : Données utilisées pour évaluer la qualité des prédictions du modèles. Ces données ne sont pas utilisées pour estimer les paramètres ; elles servent uniquement à évaluer les prédictions des différents modèles. C'est sur cet ensemble que la sélection de modèle s'effectue.

Ensemble de test : Données utilisées pour estimer l'erreur de prédiction du modèle choisi. Ces données ne devraient jamais être utilisées avant d'avoir choisi le modèle final sinon l'erreur de prédiction sera sous-estimée.

Soit $\{(\boldsymbol{x}_i, y_i) : 1 \leq i \leq n_V\}$ les n_V données de l'ensemble de validation. Dénotons par $\hat{Y}_{-i}$ la prédiction de y_i dans l'ensemble de validation. L'indice $-i$ rappelle que cette donnée n'a pas été utilisée pour l'estimation des paramètres. On choisira un modèle qui optimise une mesure de qualité des prédictions. Choisissons ici l'erreur quadratique moyenne, bien qu'il en existe d'autres. L'erreur quadratique moyenne de prédiction est définie ainsi :

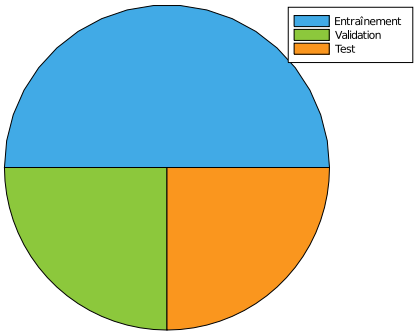
$$\text{MSE} = \frac{1}{n_V} \sum_{i=1}^{n_V} \left(\hat{Y}_{-i} - y_i \right)^2 .$$

Le modèle choisi est celui qui minimise l'erreur quadratique moyenne.

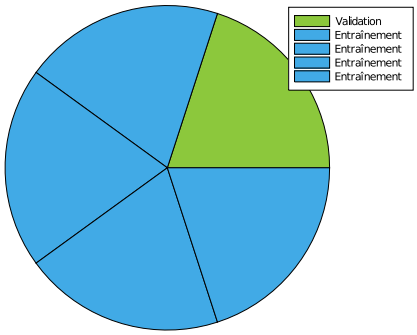
2.11.3 Validation croisée à k blocs

Le scénario décrit à la section précédente est idéal dans les situations où la taille d'échantillon est très grande. Dans les situations où la taille d'échantillon est modeste, ignorer la moitié des observations pour l'estimation des paramètres peut avoir des conséquences importantes sur la qualité d'estimation. La validation croisée permet d'utiliser les ensembles de d'entraînement et de validation astucieusement pour maximiser l'information que l'on extrait des données.

La validation croisée à k blocs partitionne l'ensemble d'entraînement et de validation en k sous-ensembles tel qu'illustré à la figure 2.10b. Chacun de ces sous-ensembles joue de façon itérative le rôle d'échantillon de validation et les autres sous-ensembles l'ensemble d'entraînement. Pour la validation croisée à k blocs, k itérations sont nécessaires pour



(a) Partitionnement pour évaluer la qualité des prédictions.



(b) Partitionnement pour la validation croisée à 5 blocs.

FIGURE 2.10 – Partitionnement des données.

construire l'ensemble des prédictions du modèles $(\hat{Y}_{-i} : 1 \leq i \leq n)$. Ensuite, la qualité des prédictions peut être évaluée.

Le cas où $k = n$ correspond au *leave-one-out cross-validation*. Dans ce dernier cas, à chacune des itérations, une donnée est retirée du jeu de données et elle est prédite par le modèle à l'aide des $(n-1)$ autres données. Le problème avec cette approche est l'échantillon d'entraînement ne change presque pas à chaque itération. On a donc tendance à avoir une grande incertitude sur la vraie erreur de prédiction. C'est pourquoi il est recommandé dans la littérature de prendre $k = 5$ ou $k = 10$. De cette façon, les échantillons d'entraînement à chacune des itérations de la validation croisée sont moins semblables.

2.12 Exercices

1. L'estimation des paramètres du plan de régression est donné par l'équation $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$. Pour le calcul informatique, on préfère généralement ne pas calculer la matrice $(X^\top X)$ parce que l'opération est complexe en temps ($O(n^3)$). Aussi, on préfère généralement éviter de calculer l'inverse d,une matrice notamment en raison de la complexité en temps et des opérations instables numériquement. Les librairies effectuant la régression linéaire utilisent plutôt la décomposition QR de la matrice X. En utilisant la décomposition QR de X, *i.e.* $X = QR$, montrez que

   a) $\hat{\boldsymbol{\beta}} = R^{-1} Q^\top \boldsymbol{Y}$;

   b) $H = X \left(X^\top X\right)^{-1} X^\top = QQ^\top$.

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[3]. Répondez aux questions suivantes en utilisant Julia et le fichier `tuition_vs_salary.csv` disponible sur Moodle.

   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) Donnez une interprétation de $\hat{\beta}_1$.

   d) Donnez une interprétation de $\hat{\beta}_0$.

   e) Quel est le coefficient de détermination de la régression ? Comment pouvez-vous l'interpréter ?

4. 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. On considère le modèle de régression linéaire suivant :

$$Y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \varepsilon_i.$$

   a) Estimez les coefficients de régression.

   b) Estimez la variance de l'erreur σ^2.

3. 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

c) Testez l'importance de la régression. Quelle est la statistique observée du test ? Quelle est la valeur-p de la statistique observée ? Est-ce que l'hypothèse nulle est rejetée ?

d) Calculez les intervalles de confiance de à 95% des coefficients de régression. Est-ce qu'il y a une variable explicative qui semble non significative ? Justifiez votre réponse.

e) Calculez un intervalle de confiance à 95% pour le pourcentage de matière grasse d'une patiente possédant les variables explicatives suivantes :

$$x_1 = 24.0 \qquad x_2 = 50.0 \qquad x_3 = 26.0$$

5. Le jeu de données `notes.csv` compile les notes obtenues aux contrôles 1 et 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) Estimez les intervalles de confiance de niveau 95% des coefficients de régression.

   c) Est-ce que la session semble être une variable explicative ? Justifiez votre réponse.

   d) 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.

6. 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 ?

7. Reprenez le jeu de données `bodyfat.csv`. Vérifiez s'il y a présence de multicolinéarité en calculant le facteur d'inflation de la variance pour chacune des variables explicatives.

8. 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 : le 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).
   On veut identifier les variables qui expliquent le mieux la tension artérielle des patients parmi celles énumérées.

   a) Est-ce qu'il y a présence de multicolinéarité ?

   b) En utilisant toutes les variables explicatives, vérifiez si les hypothèses de la régression sont satisfaites.

   c) Quel est le meilleur sous-ensemble des variables explicatives pour expliquer la tension artérielle des patients ? Justifiez votre réponse.

9. 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) En exlcluant le cas où il y a présence de multicolinéarité, 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 $Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon$.