Chapitre 2

Le modèle de régression linéaire multiple

2.1 Introduction : retour sur les exemples

La modélisation des données des exemples choisis par un modèle de régression linaire simple peut être critiquée : on voit bien dans certains des exemples que ce modèle, trop simpliste, n'est pas adapté...
Des variables explicatives supplémentaires à prendre en considération, et si oui, lesquelles ?
Dans chacun des exemples, on peut envisager d'introduire de nouvelles variables explicatives, quantitatives ou qualitatives (catégorielles).
Pour les données Insee, on peut introduire des informations sur les campagnes ou lois anti-tabac mises en place ; pour les données sur l'espérance de vie, on peut imaginer beaucoup de variables explicatives, et parmi les plus pertinentes, des indicateurs de richesse, des variables sur le système de santé ; pour les données Air Breizh, par ex. la nébulosité, la vitesse et la direction du vent, les températures à différentes heures de la journée, etc. Pour les données Cirad, il semble évident que la prise en compte de la racine carrée de la circonférence serait pertinente, mais on peut aussi considérer la zone de plantation par ex.

On cherche donc à généraliser le modèle précédent, en considérant non pas une mais plusieurs variables explicatives.
On ne considère pas dans ce chapitre le caractère éventuellement aléatoire des variables explicatives, quitte à conditionner sachant les valeurs de ces variables.

2.2 Modélisation

On introduit le modèle statistique suivant :

$$Y_i = \beta_0 x_{i,0} + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad \text{pour} \quad i = 1 \ldots n,$$

où

- $p \leq n$,
- Y_i est une variable aléatoire observée, appelée *variable à expliquer*,

— $x_{i,0}, x_{i,1}, \ldots, x_{i,p-1}$ sont des valeurs réelles déterministes appelées par extension directe du cas aléatoire *variables explicatives*. Souvent $x_{i,0} = 1$ pour tout $i = 1 \ldots n$, mais PAS TOUJOURS.
— $\beta_0, \beta_1, \ldots, \beta_{p-1}$ sont des paramètres réels inconnus appelés *paramètres de régression* ou *cœfficients de régression*,
— les ε_i sont des variables aléatoires, non observées, appelées *erreurs* ou *bruits*, auxquelles on impose certaines conditions complémentaires.

Les conditions standards imposées aux ε_i sont les conditions (C_1) à (C_3) vues dans le chapitre précédent i.e.
— $(C_1) : \mathbb{E}[\varepsilon_i] = 0$ pour tout $i = 1 \ldots n$ (centrage),
— $(C_2) : \text{cov}(\varepsilon_i, \varepsilon_j) = 0$ pour tout $i \neq j$ (non corrélation),
— $(C_3) : \text{var}(\varepsilon_i) = \sigma^2$ (inconnue) pour tout $i = 1 \ldots n$ (homoscédasticité).

Ce modèle est appelé *modèle de régression linéaire multiple*.

Il s'exprime de façon vectorielle :

$$Y = \mathbb{X}\beta + \varepsilon, \tag{2.1}$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} x_{1,0} & \ldots & x_{1,p-1} \\ \vdots & & \vdots \\ x_{n,0} & \ldots & x_{n,p-1} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Sous les conditions (C_1) à (C_3), on a alors :
— $\mathbb{E}[\varepsilon] = 0$ et $\mathbb{E}[Y] = \mathbb{X}\beta$,
— $\text{Var}(\varepsilon) = \text{Var}(Y) = \sigma^2 I_n$.

La matrice $\mathbb{X}$ est appelée *matrice du plan d'expérience*.
On suppose que cette matrice est de plein rang, c'est-à-dire $\text{rang}(\mathbb{X}) = p$. Ses vecteurs colonnes sont linéairement indépendants. Cela implique en particulier que la matrice symétrique $\mathbb{X}'\mathbb{X}$ est définie positive.

2.3 Exemples de modèles de régression linéaire multiple

Les variables explicatives peuvent prendre différentes formes.

2.3.1 Variables explicatives quantitatives

Exemple des données Insee sur le tabac : Y_i = consommation de tabac en grammes par adulte par jour au cours de l'année i, $x_{i,0} = 1$, $x_{i,1}$ = prix relatif du tabac l'année i, $x_{i,2}$ = coût des campagnes publicitaires anti-tabac diffusées au cours de l'année i.

Exemple des données de l'OMS sur l'espérance de vie : Y_i = l'espérance de vie dans le $i$ème pays, $x_{i,0}$ = le PIB, $x_{i,1}$ = le revenu moyen par habitant, $x_{i,2}$ = le budget consacré à la santé.

Exemple des données Air Breizh : Y_i = maximum journalier de la concentration en ozone au jour i, $x_{i,0} = 1$, $x_{i,1}$ = la température à midi au jour i, $x_{i,2}$ = la température à 9 heures au jour i, $x_{i,3}$ = la nébulosité à midi au jour i, $x_{i,4}$ = la nébulosité à 9 heures, $x_{i,5}$ = la vitesse du vent au jour i...

2.3.2 Transformations de variables explicatives quantitatives

Exemple des données Cirad : Y_i = hauteur de l'eucalyptus i, $x_{i,0} = 1$, $x_{i,1}$ = la circonférence à 1m30 de l'eucalyptus i, $x_{i,2} = \sqrt{x_{i,1}}$.

On peut en fait considérer des transformations polynômiales, exponentielles, logarithmiques, trigonométriques... des variables explicatives quantitatives. Attention, ces transformations ne doivent pas faire intervenir de nouveaux paramètres inconnus !

2.3.3 Variables explicatives qualitatives

Dans le cas de variables explicatives qualitatives, on les représente sous la forme d'indicatrices.

Exemple des données Insee : $x_{i,3} = 1$ si une loi anti-tabac a été votée au cours de l'année i, 0 sinon.

Exemple des données de l'OMS : , $x_{i,3} = 1$ si le pays est dans une zone géographique particulière, 0 sinon, $x_{i,4} = 1$ si le pays est en guerre, 0 sinon...

Exemple des données Air Breizh : $x_{i_6} = 1$ si le vent a pour direction l'est, 0 sinon, $x_{i_7} = 1$ si le vent a pour direction l'ouest, 0 sinon, $x_{i_8} = 1$ si le vent a pour direction le nord, 0 sinon, $x_{i_9} = 1$ si le vent a pour direction le sud, 0 sinon, etc.

Exemple des données Cirad : $x_{i,3} = 1$ si l'eucalyptus i est situé dans le bloc A de la plantation, 0 sinon, $x_{i,4} = 1$ si l'eucalyptus i est situé dans le bloc B de la plantation, 0 sinon, etc.

2.3.4 Interactions

On peut envisager le cas où les variables explicatives interagissent entre elles. Ce phénomène est modélisé par des produits des différentes variables. Ces interactions peuvent être d'ordres variés.

Remarque : les modèles de régression linéaire multiple avec des variables explicatives qualitatives seront traités en cours d'ANOVA.

2.4 Estimateur des moindres carrés ordinaires

Comme pour la régression linéaire simple, on choisit ici comme fonction de perte la perte quadratique.

Définition 5. *L'estimateur des moindres carrés ordinaires de β dans le modèle de régression linéaire multiple (2.1) est défini par*

$$\hat{\beta} \in \underset{\beta\in\mathbb{R}^p}{argmin} \sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p-1} \beta_j x_{i,j} \right)^2 = \underset{\beta\in\mathbb{R}^p}{argmin} \|Y - \mathbb{X}\beta\|^2,$$

où $\|.\|$ est la norme euclidienne de $\mathbb{R}^n$.

On montre que

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y.$$

Preuves analytique et géométrique (c.f. chapitre précédent).
Soit $L : \beta \mapsto (Y - \mathbb{X}\beta)'(Y - \mathbb{X}\beta) = Y'Y - Y'\mathbb{X}\beta - \beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta = Y'Y - 2\beta'\mathbb{X}'Y + \beta'\mathbb{X}'\mathbb{X}\beta$.

Le vecteur $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ est bien un point critique de L puisque $\nabla L(\hat{\beta}) = 2\mathbb{X}'\mathbb{X}\hat{\beta} - 2\mathbb{X}'Y = 0$. Ce point critique correspond à un minimum. En effet, la matrice hessienne de L en $\hat{\beta}$ vaut $2\mathbb{X}'\mathbb{X}$ qui est définie positive.
On introduit maintenant, comme pour la régression linéaire simple, le sous-espace vectoriel $\mathcal{E}(\mathbb{X})$ de $\mathbb{R}^n$ engendré par les vecteurs colonnes de $\mathbb{X}$. Par définition, $\mathbb{X}\hat{\beta}$ est un vecteur de $\mathcal{E}(\mathbb{X})$ dont la distance euclidienne avec Y est la distance minimum entre Y et tout vecteur de $\mathcal{E}(\mathbb{X})$. Par conséquent, si l'on note $\Pi_{\mathbb{X}}$ la matrice de projection orthogonale sur $\mathcal{E}(\mathbb{X})$, alors $\mathbb{X}\hat{\beta} = \Pi_{\mathbb{X}}Y$. Là encore, on peut montrer que la matrice $\Pi_{\mathbb{X}}$ s'écrit aussi $\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$, d'où $\mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, puis $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$.

Proposition 1. *L'estimateur des MCO $\hat{\beta}$ est un estimateur linéaire sans biais de β, dont la matrice de variance covariance est donnée par*

$$Var(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}.$$

Preuve.
Puisque $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, il s'agit bien d'un estimateur linéaire (en Y). De plus, $\mathbb{E}[\hat{\beta}] = \mathbb{E}[(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y] = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{X}\beta = \beta$, donc $\hat{\beta}$ est sans biais. Enfin, $\text{Var}(\hat{\beta}) = \text{Var}\left((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\text{Var}(Y)\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\left(\sigma^2 I_n\right)\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1} = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.

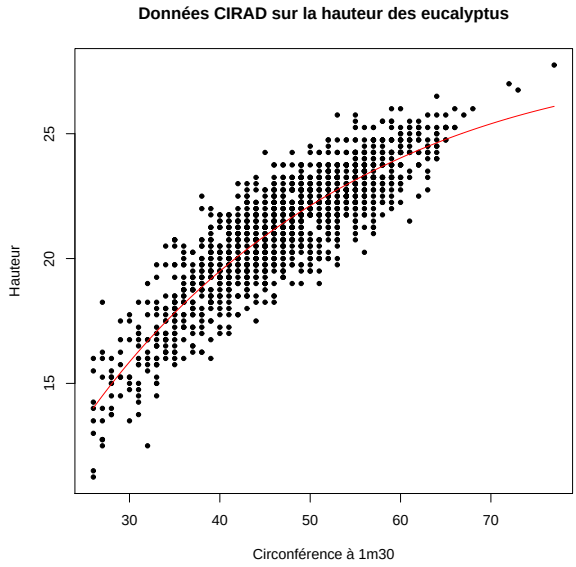
Thèorème 8 (Gauss Markov)**.** *L'estimateur $\hat{\beta}$ des moindres carrés ordinaires est l'unique estimateur linéaire sans biais de variance minimale parmi les estimateurs linéaires sans biais de β.*

Preuve (sans l'unicité).
Soit $\tilde{\beta}$ un estimateur linéaire sans biais de β. $\tilde{\beta}$ s'écrit donc $\tilde{\beta} = AY$, avec $A\mathbb{X}\beta = \beta$ pour tout β c'est-à-dire $A\mathbb{X} = I_p$.

$$\begin{aligned}
\text{Var}(AY) &= \text{Var}\left(A(I_n - \Pi_{\mathbb{X}} + \Pi_{\mathbb{X}})Y\right) \\
&= \text{Var}(A(I_n - \Pi_{\mathbb{X}})Y + A\Pi_{\mathbb{X}}Y) \\
&= A(I_n - \Pi_{\mathbb{X}})\sigma^2 I_n (I_n - \Pi_{\mathbb{X}})A' + 2A(I_n - \Pi_{\mathbb{X}})\sigma^2 I_n \Pi_{\mathbb{X}} A' + \text{Var}\left(A\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) \\
&= \sigma^2 A(I_n - \Pi_{\mathbb{X}})A' + 0 + \text{Var}\left((\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y\right) \\
&= \sigma^2 A(I_n - \Pi_{\mathbb{X}})A' + \text{Var}\left(\hat{\beta}\right).
\end{aligned}$$

Puisque la matrice $A(I_n - \Pi_{\mathbb{X}})A'$ est symétrique réelle positive (rappel sur la relation d'ordre partielle entre matrices symétriques réelles), on en conclut que $\hat{\beta}$ est de variance minimale parmi les estimateurs linéaires sans biais.

FIGURE 2.1 – Données Cirad : représentation de la courbe de régression obtenue



2.5 Valeurs ajustées, résidus

Définition 6. *Le vecteur aléatoire* $\hat{Y} = \Pi_{\mathbb{X}} Y = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$ *est appelé le* vecteur des valeurs ajustées.
Le vecteur $\hat{\varepsilon} = Y - \hat{Y} = (I_n - \Pi_{\mathbb{X}})Y$ *est appelé le* vecteur des résidus.
La matrice $\Pi_{\mathbb{X}}$ *est parfois appelée la matrice "chapeau" (hat matrix en anglais), et souvent notée dans ce cas H. Ses cœfficients sont notés* $h_{i,j}$.

Remarque : $\hat{\varepsilon}$ est orthogonal au vecteur $\hat{Y}$. Il correspond au projeté orthogonal de Y sur $\mathcal{E}(\mathbb{X})^{\perp}$.
Représentation géométrique.

On peut ensuite montrer (facilement) le résultat suivant.

Proposition 2. $\mathbb{X}'\hat{\varepsilon} = 0$, $\mathbb{E}[\hat{\varepsilon}] = 0$ *et* $Var(\hat{\varepsilon}) = \sigma^2(I_n - \Pi_{\mathbb{X}})$.

Les résidus sont centrés, corrélés et hétéroscédastiques.

2.6 Somme des carrés résiduelle et estimation ponctuelle de la variance

Comme dans le modèle de régression linéaire simple, le vecteur des résidus peut servir à l'estimation ponctuelle de la variance σ^2.
On introduit la *somme des carrés résiduelle* : SCR $=\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \|\hat{\varepsilon}\|^2$.
On a $\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \mathbb{E}\left[\hat{\varepsilon}'\hat{\varepsilon}\right] = \mathbb{E}\left[tr\left(\hat{\varepsilon}'\hat{\varepsilon}\right)\right]$ (astuce de la trace).
Or pour toutes matrices A, B, C, $tr(ABC) = tr(CAB) = tr(BCA)$, d'où :
$\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \mathbb{E}\left[tr\left(\hat{\varepsilon}\hat{\varepsilon}'\right)\right] = tr\mathbb{E}\left[\hat{\varepsilon}\hat{\varepsilon}'\right] = tr(\mathrm{Var}(\hat{\varepsilon})) = \sigma^2 tr(I_n - \Pi_{\mathbb{X}}) = \sigma^2(n - tr(\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'))$.
On a donc $\mathbb{E}\left[\|\hat{\varepsilon}\|^2\right] = \sigma^2(n - tr(\mathbb{X}'\mathbb{X}(\mathbb{X}'\mathbb{X})^{-1})) = \sigma^2(n - tr(I_p)) = \sigma^2(n-p)$.

Proposition 3. *Un estimateur sans biais de la variance σ^2 est donné par $\widehat{\sigma^2} = SCR/(n-p) = \|\hat{\varepsilon}\|^2/(n-p)$.*

2.7 Equation d'analyse de la variance, cœfficient de détermination

On a défini la *somme des carrés résiduelle* : SCR $=\|\hat{\varepsilon}\|^2$.
On introduit maintenant la *somme des carrés totale* : SCT $=\|Y - \bar{Y}\mathbb{1}\|^2$ si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, ou la *somme des carrés totale sans constante* : $\mathrm{SCT}_{sc}=\|Y\|^2$ si le vecteur $\mathbb{1}$ n'est pas l'un des vecteurs colonnes de la matrice $\mathbb{X}$.
On introduit aussi la *somme des carrés expliquée* : SCE $=\|\hat{Y} - \bar{Y}\mathbb{1}\|^2$ si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, ou la *somme des carrés expliquée sans constante* : $\mathrm{SCE}_{sc}=\|\hat{Y}\|^2$ si $\mathbb{1}$ n'est pas l'un des vecteurs colonnes de la matrice $\mathbb{X}$.
On a, par le théorème de Pythagore : SCT=SCE+SCR (équation d'analyse de la variance) si $\mathbb{1}$ est l'un des vecteurs colonnes de la matrice $\mathbb{X}$, $\mathrm{SCT}_{sc}=\mathrm{SCE}_{sc}$+SCR dans tous les cas.

Définition 7. *Le* cœfficient de détermination R^2 *est défini par :*

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

Le cœfficient de détermination sans constante R^2_{sc} *est défini par :*

$$R^2_{sc} = \frac{SCE_{sc}}{SCT_{sc}} = 1 - \frac{SCR}{SCT_{sc}}.$$

Interprétations géométriques dans les deux cas. Interprétations des cas limites.

Proposition 4. *Le cœfficient de détermination croît avec le nombre de variables explicatives p.*

Conséquence : on ne peut pas utiliser ce critère comme critère de comparaison entre deux modèles dont les nombres de variables explicatives diffèrent... Idée du R^2 *ajusté* comme critère de comparaison dans ce cas.

2.8 Prédiction

A partir d'une nouvelle valeur explicative $x_{n+1} = (x_{n+1,0}, \ldots, x_{n+1,p-1})$, on souhaite prédire une nouvelle observation d'une variable $Y_{n+1} = \beta_0 x_{n+1,0} + \ldots + \beta_{p-1} x_{n+1,p-1} + \varepsilon_{n+1} = x_{n+1}\beta + \varepsilon_{n+1}$,

avec $\mathbb{E}[\varepsilon_{n+1}] = 0$, $\text{var}(\varepsilon_{n+1}) = \sigma^2$ et $\text{cov}(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour tout $i = 1 \dots n$ i.e. Y_{n+1} non corrélée avec les Y_i, $i = 1 \dots n$, utilisées pour construire $\hat{\beta}$.

Pour cela, on introduit $\hat{Y}^p_{n+1} = x_{n+1}\hat{\beta}$.

L'erreur de prédiction est définie par $\hat{\varepsilon}^p_{n+1} = Y_{n+1} - \hat{Y}^p_{n+1}$ (inconnue).
Elle est centrée, de variance égale à $\text{var}(\hat{\varepsilon}^p_{n+1}) = \text{var}(x_{n+1}\beta + \varepsilon_{n+1} - x_{n+1}\hat{\beta}) = \text{var}(\varepsilon_{n+1}) + x_{n+1}\text{Var}(\hat{\beta})x'_{n+1} = \sigma^2(1 + x_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x'_{n+1})$.

On remarque par ailleurs que : $\text{var}(\hat{\varepsilon}^p_{n+1}) = \mathbb{E}\left[\left(Y_{n+1} - \hat{Y}^p_{n+1}\right)^2\right]$ appelée aussi *erreur quadratique moyenne de prédiction* (EQMP), qu'on utilisera plus tard pour faire de la sélection de variables ou de modèle.

2.9 Estimation par intervalles de confiance et tests d'hypothèses asymptotiques

Pour construire des intervalles de confiance ou des tests d'hypothèses sur β, on a besoin de connaître la loi de $(\hat{\beta} - \beta)$. Dans le cas général, on ne fait aucune hypothèse sur la loi de ε, donc on peut éventuellement chercher à en avoir une connaissance approximative lorsque n est très grand.

2.9.1 Théorèmes limites

Pour chaque taille d'échantillon n, on précisera la dépendance en n à l'aide de $^{(n)}$: Y sera ainsi noté $Y^{(n)}$, $\mathbb{X}$, $\hat{\beta}$, $\hat{\varepsilon}$ et $\widehat{\sigma^2}$ deviennent respectivement $\mathbb{X}^{(n)}$, $\hat{\beta}^{(n)}$, $\hat{\varepsilon}^{(n)}$ et $\widehat{\sigma^2}^{(n)}$.

Thèorème 9. *Si les ε_i sont maintenant supposées i.i.d. et si $\mathbb{X}^{(n)'}\mathbb{X}^{(n)}/n \to_{n\to+\infty} A$ définie positive, alors*

$$\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right) \to^{(\mathcal{L})} \mathcal{N}(0, \sigma^2 A^{-1}).$$

Remarque : σ^2 étant inconnu, on l'estime par $\widehat{\sigma^2}^{(n)}$.

Thèorème 10. *Si les ε_i sont supposées i.i.d. alors $\widehat{\sigma^2}^{(n)} \to^{(P)} \sigma^2$.*

Le lemme de Slutsky permet de conclure que si en plus $\mathbb{X}^{(n)'}\mathbb{X}^{(n)}/n \to_{n\to+\infty} A$, la loi de $\sqrt{nA/\widehat{\sigma^2}^{(n)}}\left(\hat{\beta}^{(n)} - \beta\right)$ est asymptotiquement gaussienne centrée réduite.

2.9.2 L'idée du bootstrap non paramétrique

La loi de $\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)$ étant inconnue, une méthode de rééchantillonnage permettra de "recréer" à partir des Y_i de nouvelles variables dont la loi conditionnelle sachant les Y_i est proche en un certain sens de cette loi inconnue.
Une méthode de rééchantillonnage classique est celle du bootstrap non paramétrique qu'on peut décrire de la façon suivante.

On suppose que le vecteur $\mathbb{1}$ est l'un des vecteurs colonnes de $\mathbb{X}^{(n)}$.

1. Calcul de l'estimateur des MCO de β, $\hat{\beta}^{(n)}$ à partir de $Y^{(n)}$, puis du vecteur des résidus $\hat{\varepsilon}^{(n)}$.
2. Tirage de n éléments notés $(\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)$, appelés résidus bootstrapés pris au hasard avec remise dans $\{\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n\}$.
3. A partir de $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)'$, calcul de $Y^* = \mathbb{X}^{(n)}\hat{\beta}^{(n)} + \hat{\varepsilon}^*$.
4. Calcul de l'estimateur bootstrapé : $\hat{\beta}^{(n)*} = \left(\mathbb{X}^{(n)'}\mathbb{X}^{(n)}\right)^{-1}\mathbb{X}^{(n)'}Y^*$.

Si d désigne une distance sur les lois de probabilité, alors :

$$d\left(\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)|Y^{(n)}\right), \mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)\right)\right) \to^{(P)}_{n\to+\infty} 0.$$

Puisque les variables $\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)$ se calculent à partir de $Y^{(n)}$, on peut simuler empiriquement la loi $\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)*} - \hat{\beta}^{(n)}\right)|Y^{(n)}\right)$ qui "approche" la loi $\mathcal{L}\left(\sqrt{n}\left(\hat{\beta}^{(n)} - \beta\right)\right)$. On peut ainsi déterminer des quantiles empiriques, etc.

2.10 Exercices

Exercice 1 : Questions de cours

On considère le modèle de régression linéaire multiple

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

où le vecteur Y à valeurs dans $\mathbb{R}^n$ représente la variable à expliquer, $\mathbb{X}$ est une matrice réelle de taille $n \times p$ de rang p, $\beta \in \mathbb{R}^p$ (inconnu) et ε est le vecteur des bruits à valeurs dans $\mathbb{R}^n$.

1. Quelles sont les conditions standards imposées au vecteur des bruits ? Expliquer comment l'analyse du modèle est facilitée par ces conditions.
2. Rappeler les définitions de l'estimateur des moindres carrés ordinaires de β, de la valeur ajustée de Y, puis du vecteur des résidus. Quelle est l'interprétation géométrique de ces vecteurs aléatoires ?
3. Proposer un calcul matriciel de l'estimateur des moindres carrés ordinaires, et préciser les propriétés de cet estimateur. Retrouver à partir du calcul matriciel les estimateurs des moindres carrés ordinaires obtenus lorsque le modèle est un modèle de régression linéaire simple.
4. Le vecteur des résidus $\hat{\varepsilon}$ a-t-il des propriétés analogues à celles de ε ?
5. Donner un estimateur naturel de la variance du modèle. Cet estimateur est-il sans biais ?
6. Peut-on prévoir l'évolution de la somme des carrés résiduelle avec l'ajout d'une variable explicative au modèle ?
7. Préciser l'équation d'analyse de la variance et son interprétation géométrique.
8. Donner la définition du cœfficient de détermination R^2, ainsi que son interprétation géométrique. Discuter des cas limites, et de l'utilisation du R^2 comme mesure de la qualité explicative du modèle.
9. Comment peut-on mesurer la qualité prédictive du modèle ?
10. Peut-on construire des régions de confiance pour β sans faire d'hypothèse sur la loi de ε ?

Exercice 2 : Données Cirad sur la hauteur des eucalyptus

On reprend les données du Cirad présentées en cours, donnant 1429 mesures de la circonférence à 1 mètre 30 du sol et de la longueur du tronc d'eucalyptus d'une parcelle plantée. On a représenté le nuage de points sur le graphique fourni en Annexe 1.1.

1. On cherche à expliquer la longueur du tronc d'un eucalyptus comme une fonction affine de la circonférence du tronc, à une erreur aléatoire près.
a) Ecrire le modèle de régression correspondant, de façon analytique puis de façon vectorielle, en veillant à bien poser les hypothèses.
b) Les valeurs calculées sur les observations des estimateurs des moindres carrés ordinaires des cœfficients de régression sont égales à 9.04 et 0.26, celle du cœfficient de détermination R^2 à 0.7683, et celle de la somme des carrés résiduelle à 2051.457. Représenter la droite de régression obtenue sur le graphique fourni.
2. On cherche maintenant à expliquer, à une erreur aléatoire près, la longueur du tronc d'un eucalyptus comme une fonction linéaire des variables explicatives suivantes : 1, la circonférence et la racine carrée de la circonférence.

a) Ecrire le modèle de régression correspondant, de façon analytique puis de façon vectorielle, en veillant à bien poser les hypothèses : on notera Y le vecteur modélisant les longueurs des troncs des eucalyptus, β le vecteur des cœfficients de régression et $\mathbb{X}$ la matrice du plan d'expérience, supposée de plein rang. Si y désigne l'observation de Y, on a

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} 1429 & 67660 & 9791.6 \\ 67660 & 3306476 & 471237.9 \\ 9791.6 & 471237.9 & 67660 \end{pmatrix}, \quad \mathbb{X}'y = \begin{pmatrix} 30312.5 \\ 1461695.8 \\ 209685.6 \end{pmatrix}, \quad \text{SCR} = 1840.247.$$

b) Donner la valeur de l'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β calculé sur les observations. Représenter sur le graphique fourni la courbe de régression obtenue.

c) Donner l'expression d'un estimateur $\widehat{\sigma^2}$ sans biais de la variance du modèle. Donner les valeurs de cet estimateur et d'un estimateur sans biais de la matrice de variance-covariance de $\hat{\beta}$ calculés sur les observations.

d) Calculer les valeurs de la somme des carrés expliquée puis du cœfficient de détermination R^2 sur les observations. Comparer ce dernier résultat à la valeur du R^2 dans le modèle de régression linéaire simple. Que peut-on en conclure ?

3. Quelle valeur peut-on prédire pour la longueur du tronc d'un eucalyptus dont la circonférence à 1m30 du sol est de 48cm dans chaque modèle ? Estimer la variance de l'erreur de prédiction correspondante dans les deux modèles. Commenter les résultats.

Exercice 3 : Rôle de la constante dans le modèle

Soit $\mathbb{X}$ une matrice $n \times p$ de rang p. Soit $\hat{Y}$ le projeté orthogonal sur l'espace engendré par les vecteurs colonnes de $\mathbb{X}$ d'un vecteur Y de $\mathbb{R}^n$.
Montrer que $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$ si l'un des vecteurs colonnes de $\mathbb{X}$ est le vecteur $\mathbb{1} = (1, \ldots, 1)'$.

Exercice 4 : Cœfficient de détermination et modèles emboîtés

Soit $\mathbb{Z}$ une matrice $n \times q$ de rang q, dont le premier vecteur colonne est $\mathbb{1}$, et $\mathbb{X}$ une matrice $n \times p$ de rang p composée des q vecteurs colonnes de $\mathbb{Z}$ et de $p - q$ autres vecteurs linéairement indépendants ($q \leq p \leq n$). On considère les deux modèles de régression linéaire multiple suivants :

$$\begin{aligned} Y &= \mathbb{Z}\beta + \varepsilon \\ Y &= \mathbb{X}\tilde{\beta} + \tilde{\varepsilon}, \end{aligned}$$

où ε et $\tilde{\varepsilon}$ vérifient les conditions standards d'un modèle de régression linéaire multiple. Comparer les cœfficients de détermination R^2 dans les deux modèles. Discuter de l'utilisation du R^2 pour la sélection de modèle ou de variables explicatives.

Exercice 5 : Régression sur variables explicatives orthogonales

On considère le modèle de régression linéaire multiple :

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

où $Y \in \mathbb{R}^n$, $\mathbb{X}$ est une matrice réelle de taille $n \times p$ ($p \leq n$) composée de p vecteurs colonnes orthogonaux, $\beta = (\beta_0, \ldots, \beta_{p-1})' \in \mathbb{R}^p$, $\varepsilon \in \mathbb{R}^n$ vérifiant les conditions standards. L'estimateur des moindres carrés ordinaires de β est noté $\hat{\beta}^{(\mathbb{X})} = (\hat{\beta}_0^{(\mathbb{X})}, \ldots, \hat{\beta}_{p-1}^{(\mathbb{X})})'$.

Soit U la matrice des q premières colonnes de $\mathbb{X}$ et V la matrice des $p-q$ dernières colonnes de $\mathbb{X}$. On définit à partir de ces matrices deux nouveaux modèles de régression linéaire multiple, l'un à q variables explicatives, l'autre à $p-q$ variables explicatives. Les estimateurs des moindres carrés ordinaires de β obtenus dans ces modèles sont respectivement notés $\hat{\beta}^{(U)} = (\hat{\beta}_0^{(U)}, \ldots, \hat{\beta}_{q-1}^{(U)})'$ et $\hat{\beta}^{(V)} = (\hat{\beta}_q^{(V)}, \ldots, \hat{\beta}_{p-1}^{(V)})'$. Les sommes des carrés expliquées dans les trois modèles sont notées SCE($\mathbb{X}$), SCE(U) et SCE(V).

1. Montrer que SCE($\mathbb{X}$)=SCE(U)+SCE(V).
2. Montrer que pour $0 \leq j \leq q-1$, $\hat{\beta}_j^{(U)} = \hat{\beta}_j^{(\mathbb{X})}$ et que pour $q \leq j \leq p-1$, $\hat{\beta}_j^{(V)} = \hat{\beta}_j^{(\mathbb{X})}$.