UNIVERSITÉ DE NANTES

Master 2 Ingénierie Mathématique

Cours de Régression

Table des matières

1 Introduction

On veut modéliser une variable Y (variable à expliquer, réponse) en fonction d'une ou plusieurs variables explicatives $X_1, ..., X_p$ (covariables). L'objectif est de prédire ou simplement expliquer Y à partir des données disponibles $X_1, ..., X_p$. Grossièrement, on cherche une fonction f telle que $Y \approx f(X_1, ..., X_p)$. Dans ce cours, on se limitera au cas simple où f est linéaire (ou affine). La méthode varie selon si les variables sont qualitatives ou quantitatives.

- Régression linéaire : Y et $X_1, ..., X_p$ quantitatives.
  On suppose un lien linéaire entre les variables de la forme

  $$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon,$$

  où les coefficients $\beta_0, \beta_1, ..., \beta_p$ sont des réels inconnus et ϵ est un bruit correspondant à la part de Y indépendantes des variables explicatives. L'objectif principal est d'estimer les coefficients $\beta_0, \beta_1, ..., \beta_p$.

- Analyse de variance (ANOVA) : Y quantitative, $X_1, ..., X_p$ qualitatives.
  Expliquer Y revient à attribuer une valeur moyenne dans chaque classe définie à partir des valeurs de $X_1, ..., X_p$ (par ex. si X_j peut prendre k_j valeurs possibles, il existe $k_1 \times ... \times k_p$ classes différentes). On peut alors essayer d'évaluer si chaque variable explicative a une influence ou non sur Y.

- Analyse de covariance (ANCOVA) : Y quantitative, $X_1, ..., X_p$ qualitatives et quantitatives.
  Les valeurs différentes des variables explicatives qualitatives définissent des classes dans lesquelles on effectue la régression linéaire de Y sur les variables explicatives quantitatives.

Lorsque Y est qualitative et les variables explicatives $X_1, ..., X_p$ sont à la fois qualitatives et quantitatives, on peut utiliser des méthodes similaires pour modéliser un lien entre les variables. On peut par exemple cherche à évaluer la probabilité que Y appartienne à une classe conditionnellement à $X_1, ..., X_p$ en supposant une relation linéaire entre le logarithme de la probabilité et les variables explicatives (régression logistique). Cette situation ne sera cependant pas traitée dans ce cours.

De manière générale, la meilleure approximation de Y (pour le coût quadratique) par une fonction des X_j est donnée par l'espérance conditionnelle

$$f(X_1, ..., X_p) = \mathbb{E}(Y|X_1, ..., X_p),$$

qui est bien sûr inconnue en pratique. Lorsque Y admet un moment d'ordre 2, l'espérance conditionnelle minimise l'erreur quadratique $\mathbb{E}\left[(Y - f(X_1, ..., X_p))^2\right]$. Si le vecteur $(Y, X_1, ..., X_p)$ est Gaussien, on sait que l'espérance conditionnelle est une fonction affine. Dans ce cas, on peut donc se restreindre aux fonctions f linéaires en $1, X_1, ..., X_p$,

$$f(X_1, ..., X_p) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p,$$

ce qui justifie le terme régression linéaire.

2 Régression linéaire simple

Supposons dans un premier temps que l'on dispose d'une seule variable explicative X. On observe un échantillon $\{y_i, x_i\}_{i=1,...,n}$vérifiant

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n.$$

Le modèle s'écrit sous la forme matricielle

$$y = \mathrm{X}\beta + \epsilon$$

avec

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathrm{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \ \text{et} \ \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

On suppose que les variables explicatives x_i sont déterministes (non aléatoires). L'alea est uniquement dû à la présence des bruits ϵ_i. Ainsi, seuls les vecteur y et ϵ sont des réalisations des variables aléatoires, X et β sont déterministes (par ailleurs, seuls y et X sont connus du statisticien). On suppose de plus que les bruits ϵ_i sont

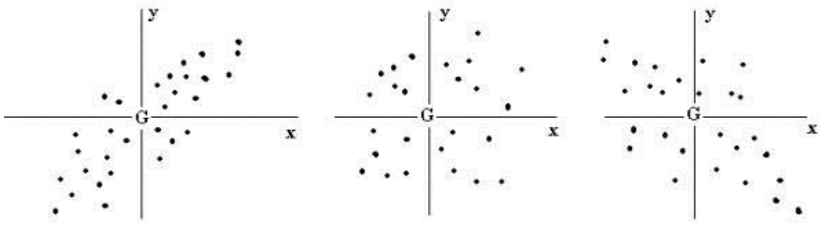
- centrés : $\mathbb{E}(\epsilon_i) = 0$,
- non-corrélés : $\forall i \neq j, \mathrm{cov}(\epsilon_i, \epsilon_j) = 0$,
- de variances égales (homoscédastiques) : $\mathrm{var}(\epsilon_i) = \sigma^2 < \infty$.

Ces hypothèses, dites *hypothèses faibles* du modèle linéaire, sont résumées par $\mathbb{E}(\epsilon) = 0$ et $\mathrm{var}(\epsilon) = \sigma^2 \mathrm{I}$. Les *hypothèses fortes* du modèle linéaires supposent en plus que les bruits sont Gaussiens et donc indépendants car non-corrélés.

Un lien affine entre des variables x et y se traduit par une corrélation linéaire non nulle. En notant $\overline{u} = \frac{1}{n}\sum_{i=1}^n u_i$ la moyenne empirique d'une variable $u \in \mathbb{R}^n$, le *coefficient de corrélation linéaire* ou coefficient de Pearson est défini par

$$\rho(x, y) = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\ \sqrt{\overline{y^2} - \overline{y}^2}},$$

où $\overline{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2$ et $\overline{y^2} = \frac{1}{n}\sum_{i=1}^n y_i^2$. La quantité $\overline{xy} - \overline{x}\ \overline{y}$ correspond à la covariance empirique entre x et y, c'est-à-dire la covariance de l'échantillon $(x_i, y_i)_{i=1,...,n}$. Les quantités $\overline{x^2} - \overline{x}^2$ et $\overline{y^2} - \overline{y}^2$ sont les variances empiriques de x et y. Le coefficient de corrélation de Pearson est compris entre -1 et 1 (on le montre avec l'inégalité de Cauchy-Schwarz). Le lien affine entre x et y est d'autant plus net que $\rho(x, y)$ est proche de 1 ou -1 (1 pour une relation croissante et -1 pour une relation décroissante).



Coefficient de corrélation proche de 1 | Coefficient de corrélation proche de 0 | Coefficient de corrélation proche de -1

La régression linéaire peut servir à modéliser certaines relations non-linéaires entre deux variables, en utilisant des transfomations de variables préalables. Par exemple une relation du type $y = \alpha x^\beta$ est

équivalente à une relation affine entre $\log(x)$ et $\log(y)$. De même, un lien exponentiel $y = \alpha e^{\beta x}$ correspond au modèle linéaire $\log(y) = \log(\alpha) + \beta \log(x)$. Une relation polynômiale $y = \mathrm{P}(x)$ où P est un polynôme de degré d équivaut à exprimer y comme une fonction linéaire des variables $1, x, x^2, ..., x^d$, qui entre donc dans le cadre de la régression linéaire (multiple).

Pour mettre en évidence l'existence d'une relation monotone (pas forcément linéaire) entre x et y du type $y_i = f(x_i) + \epsilon_i$ où f est une fonction monotone, on peut utiliser le *coefficient de corrélation de Spearman* qui se construit à partir des rangs des variables x et y. En notant $r_i \in \{1, ..., n\}$ le rang de x_i (on a donc $x_{r_1} \geqslant ... \geqslant x_{r_n}$, en privilégiant en cas d'égalité l'indice le plus petit) et s_i le rang de y_i, le coefficient de corrélation de Spearman est défini par

$$\rho_{\mathrm{S}}(x, y) = \rho(r, s) = \frac{\overline{rs} - \overline{r}\ \overline{s}}{\sqrt{\overline{r^2} - \overline{r}^2}\sqrt{\overline{s^2} - \overline{s}^2}}.$$

L'idée intuitive du coefficient de corrélation de Spearman est de dire que si Y est une fonction monotone de X alors les rangs de x et y verifient $r = s$ dans le cas d'une relation croissante, et $r = n - s$ dans le cas d'une relation décroissante. Dans les deux cas, les variables r et s sont liées linéairement.

2.1 Méthode des moindres carrés ordinaires

Une fois mis en évidence l'existence d'un lien linéaire entre deux variables x et y, l'étape suivante consiste à évaluer précisément la nature du lien. Dans le modèle

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

cela se résume simplement à estimer les paramètres β_0 et β_1 correspondent respectivement à l'ordonnée à l'origine (intercept) et à la pente de la droite qui décrit la relation linéaire. Pour estimer ces paramètres, on cherche la droite la plus "adaptée" au nuage de point. On définit l'estimateur des moindres carrés ordinaires (MCO) $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^\top$ comme le minimiseur de

$$\mathrm{R}(b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2, \quad b = (b_0, b_1)^\top \in \mathbb{R}^2.$$

Proposition 2.1 *Le minimiseur* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^\top = \arg\min_{b \in \mathbb{R}^2} \mathrm{R}(b)$ *est unique, donné par*

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \overline{x}^2} \qquad \textit{et} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

On montre que $\hat{\beta}$ est une estimateur sans biais de β. En effet $\mathbb{E}(y_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i$ par hypothèse. On a donc $\mathbb{E}(\overline{y}) = \beta_0 + \beta_1 \overline{x}$ et $\mathbb{E}(\overline{xy}) = \beta_0 \overline{x} + \beta_1 \overline{x^2}$, ce qui donne

$$\mathbb{E}(\hat{\beta}_1) = \frac{\mathbb{E}(\overline{xy} - \overline{x}\ \overline{y})}{\overline{x^2} - \overline{x}^2} = \frac{\mathbb{E}(\overline{xy}) - \mathbb{E}(\overline{x}\ \overline{y})}{\overline{x^2} - \overline{x}^2} = \frac{\beta_1(\overline{x^2} - \overline{x}^2)}{\overline{x^2} - \overline{x}^2} = \beta_1$$
$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\overline{y}) - \mathbb{E}(\hat{\beta}_1)\overline{x} = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0.$$

On peut également calculer la variance de ces estimeurs ainsi que leur covariance. On remarque tout d'abord que $\mathrm{var}(y_i) = \mathrm{var}(\epsilon_i) = \sigma^2$, $\mathrm{var}(\overline{y}) = \sigma^2/n = \mathrm{cov}(y_i, \overline{y})$ (par indépendance des y_i) et

$$\mathrm{cov}(\hat{\beta}_1, \overline{y}) = \frac{\mathrm{cov}(\overline{xy}, \overline{y}) - \overline{x}\,\mathrm{var}(\overline{y})}{\overline{x^2} - \overline{x}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i\,\mathrm{cov}(y_i, \overline{y}) - \overline{x}\,\mathrm{var}(\overline{y})}{\overline{x^2} - \overline{x}^2} = \frac{\overline{x}\sigma^2 - \overline{x}\sigma^2}{n(\overline{x^2} - \overline{x}^2)} = 0.$$

On déduit

$$\mathrm{var}(\hat{\beta}_1) = \frac{\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^n (x_i-\overline{x})y_i\right)}{(\overline{x^2}-\overline{x}^2)^2} = \frac{\frac{1}{n^2}\sum_{i=1}^n (x_i-\overline{x})^2\,\mathrm{var}(y_i)}{(\overline{x^2}-\overline{x}^2)^2} = \frac{(\overline{x^2}-\overline{x}^2)\sigma^2}{n(\overline{x^2}-\overline{x}^2)^2} = \frac{\sigma^2}{n(\overline{x^2}-\overline{x}^2)},$$

$$\mathrm{var}(\hat{\beta}_0) = \mathrm{var}(\overline{y}-\hat{\beta}_1\overline{x}) = \mathrm{var}(\overline{y}) + \overline{x}^2\,\mathrm{var}(\hat{\beta}_1) - 2\overline{x}\,\mathrm{cov}(\overline{y},\hat{\beta}_1) = \frac{\sigma^2}{n} + \frac{\overline{x}^2\sigma^2}{n(\overline{x^2}-\overline{x}^2)} = \frac{\overline{x^2}\sigma^2}{n(\overline{x^2}-\overline{x}^2)},$$

$$\mathrm{cov}(\hat{\beta}_1,\hat{\beta}_0) = \mathrm{cov}(\hat{\beta}_1,\overline{y}-\hat{\beta}_1\overline{x}) = \mathrm{cov}(\hat{\beta}_1,\overline{y}) - \overline{x}\,\mathrm{var}(\hat{\beta}_1) = \frac{-\overline{x}\sigma^2}{n(\overline{x^2}-\overline{x}^2)}.$$

On montrera par la suite que $\hat{\beta}$ est optimal parmi les estimateurs linéaires sans biais (parmi tous les estimateurs sans biais dans le cas Gaussien).

A chaque observation y_i, correspond une prévision $\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i$. L'écart $\hat{\epsilon}_i := y_i - \hat{y}_i$ entre l'observation et la valeur prédite est appelé *résidu*. C'est en quelque sorte un estimateur du bruit ϵ_i, qui lui n'est pas observé. On montre facilement que les résidus $\hat{\epsilon}_i$ sont centrés

$$\mathbb{E}(\hat{\epsilon}_i) = \mathbb{E}(y_i - \hat{y}_i) = \mathbb{E}(\beta_0 + \beta_1 x_i + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

et de moyenne empirique nulle

$$\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i) = \overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} = 0.$$

Le dernier paramètre inconnu du modèle est la variance du bruit $\sigma^2 = \mathrm{var}(\epsilon_i)$. Un estimateur sans biais de σ^2 est donné par

$$\hat{\sigma}^2 := \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2.$$

Dans le cas Gaussien $\epsilon \sim \mathcal{N}(0,\sigma^2\,\mathrm{I})$, on montrera que $\frac{n-2}{\sigma^2}\hat{\sigma}^2$ suit une loi du chi 2 à $(n-2)$ degrés de liberté, notée $\chi^2(n-2)$.

2.2 Tests sur les paramètres du modèle

On s'intéresse maintenant à tester la nullité des paramètres β_0 et β_1. Théoriquement, ces tests ne sont valides que sous l'hypothèse forte de normalité des bruits. Cependant, ils restent généralement valables asymptotiquement dans le cas non-Gaussien. Dans cette partie, on se place donc dans le cas Gaussien. On note $\mathbf{1} = (1,...,1)^\top$, on sait que

$$y \sim \mathcal{N}(\beta_0\mathbf{1} + \beta_1 x, \sigma^2\,\mathrm{I}).$$

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des transformations linéaires de y, ils sont donc également Gaussiens. Connaissant leurs espérances et variances, on déduit

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\overline{x^2}\sigma^2}{n(\overline{x^2}-\overline{x}^2)}\right) \quad \text{et} \quad \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{n(\overline{x^2}-\overline{x}^2)}\right).$$

On admettra dans un premier temps les deux propriétés suivantes
– la variable aléatoire $\frac{n-2}{\sigma^2}\hat{\sigma}^2$ suit une loi $\chi^2(n-2)$.
– les variables $\hat{\sigma}^2$ et $\hat{\beta}$ sont indépendantes.
On donnera une preuve de ces affirmations dans le cas général de la régression linéaire multiple.

On rappelle que la loi de Student à d degrés de liberté, notée $\mathcal{T}_d$, est le rapport entre une loi normale standard et la racine carréee d'une loi du $\chi^2(d)$ indépendante renormalisée par d :

$$\mathcal{T}_d \overset{loi}{=} \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2(d)/d}}.$$

On a donc

$$\sqrt{n}\ \frac{\sqrt{\overline{x^2}-\overline{x}^2}}{\sqrt{\overline{x^2}}\ \sigma}(\hat{\beta}_0-\beta_0)\times\frac{\sigma}{\hat{\sigma}}=\sqrt{n}\ \frac{\sqrt{\overline{x^2}-\overline{x}^2}}{\sqrt{\overline{x^2}}\ \hat{\sigma}}(\hat{\beta}_0-\beta_0)\sim\mathcal{T}_{n-2}$$

et

$$\sqrt{n}\ \frac{\sqrt{\overline{x^2}-\overline{x}^2}}{\sigma}(\hat{\beta}_1-\beta_1)\times\frac{\sigma}{\hat{\sigma}}=\sqrt{n}\ \frac{\sqrt{\overline{x^2}-\overline{x}^2}}{\hat{\sigma}}(\hat{\beta}_1-\beta_1)\sim\mathcal{T}_{n-2}$$

On remarque que ces statistiques ont pour seules inconnues β_0 et β_1 ce qui permet de construire un test sur une valeur particulière des paramètres. Par exemple, pour tester la nullité de β_1, $\mathrm{H}_0 : \beta_1 = 0$ contre $\mathrm{H}_1 : \beta_0 \neq 0$ pour mettre en évidence l'existence ou non d'une relation affine entre X et Y, on s'intéresse à la statistique de test

$$\mathrm{T}=\sqrt{n}\ \frac{\sqrt{\overline{x^2}-\overline{x}^2}\ \hat{\beta}_1}{\hat{\sigma}},$$

qui suit une loi de Student $\mathcal{T}_{n-2}$ sous H_0. On rejettera donc l'hypothèse nulle au niveau α si $|\mathrm{T}|$ est supérieure au quantile de la loi de Student correspondant $t_{n-2}(1-\frac{\alpha}{2})$. Ce résultat permet également de construire des intervalles de confiance pour β_0 et β_1, par exemple

$$\mathbb{P}\left(\hat{\beta}_1-\frac{\hat{\sigma}\ t_{n-2}(1-\frac{\alpha}{2})}{\sqrt{n(\overline{x^2}-\overline{x}^2)}}\leqslant\beta_1\leqslant\hat{\beta}_1+\frac{\hat{\sigma}\ t_{n-2}(1-\frac{\alpha}{2})}{\sqrt{n(\overline{x^2}-\overline{x}^2)}}\right)=1-\alpha.$$

Ces tests et intervalles de confiance sont exacts dans le cas Gaussien et asymptotiquement exacts dans le cas non-Gaussien, sous des hypothèses faibles. En pratique, on peut souvent raisonnablement considérer que les tests et intervalles de confiance sont valables à partir de $n = 50$, voire $n = 30$, même après avoir rejeté l'hypothèse de normalité des résidus.

2.3 Analyse de la variance

Le vecteur $y \in \mathbb{R}^n$ peut se décomposer en une partie expliquée par x et une partie résiduelle. En construisant l'estimateur

$$\hat{\beta}=\arg\min_{b\in\mathbb{R}^2}\|y-b_0\mathbf{1}-b_1x\|^2,$$

où $\|.\|$ désigne la norme Euclidienne usuelle sur $\mathbb{R}^n$, on considère en fait la projection orthogonale $\hat{y} = \hat{\beta}_0\mathbf{1} - \hat{\beta}_1 x$ de y sur l'espace engendré par x et le vecteur constant $\mathbf{1} = (1, ..., 1)^\top$. En conséquence, le vecteur des résidus $\hat{\epsilon} = (\hat{\epsilon}_1, ..., \hat{\epsilon}_n)^\top$ est orthogonal à x et $\mathbf{1}$ (on le vérifie facilement par le calcul). En réécrivant l'égalité $y = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1 x + \hat{\epsilon}$ comme

$$y-\overline{y}\mathbf{1}=\frac{\overline{xy}-\overline{x}\ \overline{y}}{\overline{x^2}-\overline{x}^2}(x-\overline{x}\mathbf{1})+\hat{\epsilon},$$

le théorême de Pythagore entraîne

$$\|y-\overline{y}\mathbf{1}\|^2=\frac{(\overline{xy}-\overline{x}\ \overline{y})^2}{(\overline{x^2}-\overline{x}^2)^2}\|x-\overline{x}\mathbf{1}\|^2+\|\hat{\epsilon}\|^2=n\frac{(\overline{xy}-\overline{x}\ \overline{y})^2}{\overline{x^2}-\overline{x}^2}+\|\hat{\epsilon}\|^2.$$

La quantité $\|y-\overline{y}\mathbf{1}\|^2$ est appelée *somme des carrés totale* (SCT), $n(\overline{xy}-\overline{x}\ \overline{y})^2/(\overline{x^2}-\overline{x}^2)$ est la *somme des carrés expliquée* (SCE) et $\|\hat{\epsilon}\|^2=(n-2)\hat{\sigma}^2$ est la *somme des carrés résiduelle* (SCR). Le rapport

$$\mathrm{R}^2:=\frac{\mathrm{SCE}}{\mathrm{SCT}}=\frac{(\overline{xy}-\overline{x}\ \overline{y})^2}{(\overline{x^2}-\overline{x}^2)(\overline{y^2}-\overline{y}^2)}=\rho(x,y)^2$$

est appelé *coefficient de détermination*. Il est un indicateur, compris entre 0 et 1, de la qualité de la régression. Il est proche de zéro lorsque les variables y et x ne sont pas liées linéairement. Sous l'hypothèse

de normalité des bruits, il peut être utilisé pour tester l'existence d'une relation affine entre x et y, en remarquant que $1-\mathrm{R}^2 = (n-2)\hat{\sigma}^2/\|y-\overline{y}\mathbf{1}\|^2$ et

$$(n-2)\frac{\mathrm{R}^2}{1-\mathrm{R}^2} = \frac{n(\overline{xy}-\overline{x}\ \overline{y})^2}{(\overline{x^2}-\overline{x}^2)\hat{\sigma}^2} = \frac{n(\overline{x^2}-\overline{x}^2)\hat{\beta}_1^2}{\hat{\sigma}^2} = \mathrm{T}^2,$$

où T est la statistique de Student utilisée pour tester $\mathrm{H}_0 : \beta_1 = 0$ (voir précédemment). Sous H_0, la statistique de test $(n-2)\mathrm{R}^2/(1-\mathrm{R}^2)$ est donc le carré d'une loi de Student $\mathcal{T}_{n-2}$ qui correspond à une loi de Fisher à 1 et $n-2$ degrés de liberté.

2.4 Prédiction

On suppose maintenant que l'on dispose d'une nouvelle observation x_{n+1} avec laquelle on veut prédire la valeur y_{n+1} associée. Contrairement aux tests et intervalles de confiance sur les paramètres, les intervalles de prédictions traités dans cette partie ne sont valables que sous l'hypothèse de normalité des bruits. On suppose donc

$$y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \epsilon_{n+1} \sim \mathcal{N}(\beta_0 + \beta_1 x_{n+1}, \sigma^2)$$

et ϵ_{n+1} est indépendant des valeurs passées, en particulier de $\hat{\beta}$ et $\hat{\sigma}^2$. On définit naturellement le prédicteur $\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$, qui est Gaussien d'espérance et variance

$$\mathbb{E}(\hat{y}_{n+1}) = \beta_0 + \beta_1 x_{n+1}$$

$$\mathrm{var}(\hat{y}_{n+1}) = \mathrm{var}(\hat{\beta}_0) + x_{n+1}^2\,\mathrm{var}(\hat{\beta}_1) + 2x_{n+1}\,\mathrm{cov}(\hat{\beta}_0,\hat{\beta}_1) = \frac{\sigma^2(\overline{x^2} + x_{n+1}^2 - 2\overline{x}x_{n+1})}{n(\overline{x^2}-\overline{x}^2)} = \sigma^2 v$$

Par indépendance entre y_{n+1} et $\hat{y}_{n+1}$, on déduit

$$y_{n+1} - \hat{y}_{n+1} = (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_{n+1} + \epsilon_{n+1} \sim \mathcal{N}(0, \sigma^2(1+v)).$$

Puisque $\hat{\sigma}^2$ est indépendant de $\hat{\beta}$ et y_{n+1}, la statistique

$$\frac{y_{n+1}-\hat{y}_{n+1}}{\sigma\sqrt{1+v}} \times \frac{\sigma}{\hat{\sigma}} = \frac{y_{n+1}-\hat{y}_{n+1}}{\hat{\sigma}\sqrt{1+v}}$$

suit une loi de Student à $n-2$ degrés de liberté. On obtient finalement l'intervalle de confiance

$$\mathbb{P}\left(\hat{y}_{n+1} - \hat{\sigma}\sqrt{1+v}\ t_{n-2,1-\frac{\alpha}{2}} \leqslant y_{n+1} \leqslant \hat{y}_{n+1} + \hat{\sigma}\sqrt{1+v}\ t_{n-2,1-\frac{\alpha}{2}}\right) = 1-\alpha.$$

3 Régression linéaire multiple

On s'intéresse maintenant à modéliser une variable Y en fonction de plusieurs variables explicatives $\mathrm{X}_1, ..., \mathrm{X}_p$. Le modèle est une généralisation de la régression linéaire simple. On observe des réalisations indépendantes $\{y_i, x_{1,i}, ..., x_{p,i}\}_{i=1,...,n}$, avec

$$y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i} + \epsilon_i, \ i = 1, ..., n,$$

où comme précédemment, les ϵ_i sont centrés, de même variance $\sigma^2 < \infty$ et non-corrélés. Le modèle s'écrit sous forme matricielle

$$y = \mathrm{X}\beta + \epsilon,$$

avec

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \mathrm{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix}, \ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \text{ et } \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

3.1 Méthode des moindres carrés ordinaires

Comme dans le cas de la régression simple, on cherche à estimer β et σ^2. L'estimateur des MCO $\hat{\beta}$ est défini comme l'unique minimiseur de

$$\mathrm{R}(b) = \|y - \mathrm{X}b\|^2, \quad b \in \mathbb{R}^{p+1}.$$

On suppose que $p < n$ et que X est de rang $p+1$. Sous ces hypothèses, l'estimateur $\hat{\beta}$ est l'unique solution des conditions du premier ordre

$$\nabla \mathrm{R}(\hat{\beta}) = -2\mathrm{X}^\top y + 2\mathrm{X}^\top \mathrm{X}\hat{\beta} = 0 \iff \hat{\beta} = (\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top y.$$

On remarque que $\mathrm{X}^\top \mathrm{X}$ est inversible du fait que X est de plein rang.

Proposition 3.1 *L'estimateur des MCO $\hat{\beta}$ est un estimateur sans biais de β de matrice de variance*

$$\mathrm{var}(\hat{\beta}) = \sigma^2(\mathrm{X}^\top \mathrm{X})^{-1}.$$

Théorème 3.2 *(Gauss-Markov) L'estimateur des moindres carrés $\hat{\beta}$ est optimal (au sens du coût quadratique) parmi les estimateurs sans biais linéaires en y.*

L'optimalité au sens $\mathbb{L}^2$ ne nécessite pas la normalité du modèle. Un résultat plus fort est valable dans le cas Gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ où la variance de $\hat{\beta}$ atteint la borne de Cramer-Rao. L'estimateur des moindres carrés est donc optimal parmi tous les estimateurs sans biais de β dans ce cas.

La matrice $\Pi_\mathrm{X} := \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top$ utilisée dans la preuve du théorême de Gauss-Markov est la projection orthogonale sur l'image de X. On le montre simplement en vérifiant que Π_X est symétrique et vérifie $\Pi_\mathrm{X}^2 = \Pi_\mathrm{X}$ et $\mathrm{Im}(\Pi_\mathrm{X}) = \mathrm{Im}(\mathrm{X})$. Ainsi, le vecteurs des prévisions

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

est la projection orthogonale de y sur $\mathrm{Im}(\mathrm{X})$. C'est en quelque sorte la part de $y \in \mathbb{R}^n$ expliquée par les variables $\mathbf{1}, x_1, ..., x_p$ (les colonnes de X). De même, le vecteur des résidus

$$\hat{\epsilon} = y - \hat{y} = (\mathrm{I} - \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top)y = (\mathrm{I} - \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top)(\mathrm{X}\beta + \epsilon) = (\mathrm{I} - \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top)\epsilon$$

est la projection orthogonale de y sur $\mathrm{Im}(\mathrm{X})^\perp$ et par conséquent celle de ϵ puisque $y = \mathrm{X}\beta + \epsilon$. Une conséquence immédiate est que les vecteurs $\hat{\beta}$ et $\hat{\epsilon}$ sont non-corrélés. En effet,

$$\mathrm{cov}(\hat{\beta}, \hat{\epsilon}) = \mathbb{E}\left[(\hat{\beta} - \beta)\hat{\epsilon}^\top\right] = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathbb{E}\left[\epsilon\epsilon^\top\right](\mathrm{I} - \mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top) = 0.$$

La norme du vecteur des résidus permet de construire un estimateur de σ^2 par

$$\hat{\sigma}^2 := \frac{1}{n-p-1}\|y - \hat{y}\|^2 = \frac{1}{n-p-1}\|\hat{\epsilon}\|^2.$$

Proposition 3.3 *L'estimateur $\hat{\sigma}^2$ est sans biais.*

Preuve. On a vu que $\hat{\epsilon} = \Pi_{\mathrm{X}^\perp}\epsilon$ où $\Pi_{\mathrm{X}^\perp} = \mathrm{I} - \mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top$ est la matrice de projection orthogonale sur $\mathrm{Im}(\mathrm{X})^\perp$. On utilise qu'un réél est égal à sa trace et que $\mathrm{tr}(\mathrm{AB}) = \mathrm{tr}(\mathrm{BA})$:

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \mathbb{E}(\epsilon^\top\Pi_{\mathrm{X}^\perp}^\top\Pi_{\mathrm{X}^\perp}\epsilon) = \mathbb{E}(\epsilon^\top\Pi_{\mathrm{X}^\perp}\epsilon) = \mathbb{E}\left[\mathrm{tr}(\epsilon^\top\Pi_{\mathrm{X}^\perp}\epsilon)\right] = \mathbb{E}\left[\mathrm{tr}(\Pi_{\mathrm{X}^\perp}\epsilon\epsilon^\top)\right].$$

Clairement, la trace (somme des éléments diagonaux) commute avec l'espérance, d'où

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \mathrm{tr}\left[\Pi_{\mathrm{X}^\perp}\mathbb{E}(\epsilon\epsilon^\top)\right] = \mathrm{tr}\left[\Pi_{\mathrm{X}^\perp}\sigma^2\,\mathrm{I}\right] = \sigma^2\,\mathrm{tr}(\Pi_{\mathrm{X}^\perp}).$$

La trace (somme des valeurs propres) d'une matrice de projection étant égale à son rang, on obtient

$$\mathbb{E}\|\hat{\epsilon}\|^2 = \sigma^2(n-p-1) \iff \mathbb{E}\left(\hat{\sigma}^2\right) = \sigma^2. \quad \blacksquare$$

Un résultat plus fort est valable dans le cas Gaussien.

Proposition 3.4 *Dans le modèle Gaussien $\epsilon \sim \mathcal{N}(0, \sigma^2\,\mathrm{I})$, les estimateurs $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendants et vérifient*

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}) \quad \textit{et} \quad (n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1).$$

Il est intéressant de remarquer que dans le cas Gaussien, $\hat{\beta}$ est l'estimateur du maximum de vraisemblance. En revanche, l'estimateur du maximum de vraismeblance de σ^2 est différent, donné par

$$\hat{\sigma}^2_{\mathrm{MV}} = \frac{1}{n}\|y - \hat{y}\|^2 = \frac{n-p-1}{n}\hat{\sigma}^2.$$

3.2 Propriétés asymptotiques des estimateurs

On s'intéresse maintenant au comportement des estimateurs quand n tend vers l'infini. Sous les hypothèses fortes de la régression, les lois de $\hat{\beta}$ et $\hat{\sigma}^2$ sont connues ce qui permet de déduire facilement leur comportement asymptotique. La convergence de $\hat{\beta}$ dans $\mathbb{L}^2$ est soumise à la seule condition que $(\mathrm{X}^\top\mathrm{X})^{-1}$ tend vers 0. La convergence de $\hat{\sigma}^2$, que ce soit dans $\mathbb{L}^2$ ou même presque sûrement, est vérifiée dans le cas Gaussien sans hypothèse supplémentaire sur X (c'est une conséquence immédiate du théorême 3.4). On peut se demander si ces résultats restent valables sans la normalité des bruits. Un premier résultat immédiat est que sous les hypothèses faibles, $\hat{\beta}$ reste convergent dans $\mathbb{L}^2$ dès que $(\mathrm{X}^\top\mathrm{X})^{-1}$ tend vers 0. On peut également montrer que si les ϵ_i sont iid et $h_n := \max_{1\leqslant i,j\leqslant p+1}|\Pi_{\mathrm{X},ij}|$ tend vers 0, alors $\hat{\beta}$ est asymptotiquement Gaussien. Si de plus $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$ converge quand $n \to \infty$ vers une matrice inversible M, alors

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{loi} \mathcal{N}(0, \sigma^2\mathrm{M}^{-1}).$$

L'hypothèse $\frac{1}{n}\mathrm{X}^\top\mathrm{X} \to \mathrm{M}$ est souvent vérifiée en pratique. Par exemple, elle est vérifiée presque sûrement si l'échantillon $\{x_{1i}, ..., x_{pi}\}_{i=1,...,n}$ est issu de réalisations indépendantes de variables aléatoires $\mathrm{X}_1, ..., \mathrm{X}_p$ de carré intégrable. Dans ce cas, $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$ converge presque sûrement vers la matrice des moments d'ordre deux, par la loi forte des grands nombres. Ce résultat est important car il confirme que même sous les hypothèses faibles, la plupart des tests du modèle linéaire restent valables asymptotiquement.

3.3 Analyse de la variance

L'analyse de la variance consiste à diviser la variance de y en une partie expliquée par les variables $x_1, ..., x_p$ et une partie résiduelle. Il s'agit de remarquer que

- $\hat{y} = \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top y = \Pi_{\mathrm{X}} y$ est la projection orthogonale de y sur $\mathrm{Im}(\mathrm{X})$.
- $\overline{y}\mathbf{1}$ est la projection orthogonale de y sur l'espace engendré par le vecteur constant $\mathbf{1}$, noté $\mathrm{vec}\{\mathbf{1}\}$.
- $\mathrm{vec}\{\mathbf{1}\}$ étant un sous espace de $\mathrm{Im}(\mathrm{X})$, $\overline{y}\mathbf{1}$ est également la projection orthogonale de $\hat{y}$ sur $\mathrm{vec}\{\mathbf{1}\}$ (on le vérifie en remarquant que $\overline{\hat{y}} = \overline{y}$).
- $\hat{\epsilon} = y - \hat{y} = (\mathrm{I} - \mathrm{X}(\mathrm{X}^\top \mathrm{X})^{-1}\mathrm{X}^\top)y$ est la projection orthgonale de y sur $\mathrm{Im}(\mathrm{X})^\perp$.

Ainsi, en décomposant $y - \overline{y}\mathbf{1} = \hat{y} - \overline{y}\mathbf{1} + \hat{\epsilon}$, le théorême de Pythagore nous donne

$$\underbrace{\|y - \overline{y}\mathbf{1}\|^2}_{\mathrm{SCT}} = \underbrace{\|\hat{y} - \overline{y}\mathbf{1}\|^2}_{\mathrm{SCE}} + \underbrace{\|\hat{\epsilon}\|^2}_{\mathrm{SCR}}.$$

Le coefficient de détermination R^2 qui donne un indicateur de la qualité de la modélisation est défini par

$$\mathrm{R}^2 := \frac{\mathrm{SCE}}{\mathrm{SCT}} = \frac{\|\hat{y} - \overline{y}\mathbf{1}\|^2}{\|y - \overline{y}\mathbf{1}\|^2}.$$

Dans le cas univarié, on a vu que le coefficient de détermination est égal au carré du coefficient de corrélation de Pearson $\rho(x, y)$. Dans le cas multivarié, le R^2 correspond à la valeur maximale du carré du coefficient de Pearson entre y et une combinaison linéaire des variables explicatives :

$$\mathrm{R}^2 = \sup_{b \in \mathbb{R}^{p+1}} \rho(y, \mathrm{X}b)^2.$$

3.4 Tests

Le coefficient de détermination permet de tester l'existence d'une relation linéaire entre y et les variables explicatives. Précisément, l'hypothèse nulle est l'absence de relation linéaire, ce qui se traduit par H_0 : $\beta_1 = ... = \beta_p = 0$ contre $\mathrm{H}_1 : \exists j, \beta_j \neq 0$.

Proposition 3.5 *Dans le modèle Gaussien où $\epsilon \sim \mathcal{N}(0, \sigma^2)$, la statistique*

$$\mathrm{F} = \frac{n-p-1}{p} \frac{\mathrm{R}^2}{1 - \mathrm{R}^2} = \frac{n-p-1}{p} \frac{\mathrm{SCE}}{\mathrm{SCR}}$$

suit, sous $\mathrm{H}_0 : \beta_1 = ... = \beta_p = 0$, une loi de Fisher à p et $n-p-1$ degrés de liberté.

La statistique F permet donc de tester s'il existe au moins une variable pertinente parmi les variables explicatives. Elle est calculée automatiquement dans la commande `summary(lm(`$y \sim x$`))` de `R`. Ce test reste valable asymptotiquement dans le cas non-Gaussien, sous des hypothèses raisonnables.

Evidemment, l'existence d'au moins une variable explicative pertinente n'implique pas forcément que toutes sont pertinentes. Pour tester individuellement chaque variable x_j, on peut utiliser un test de Student. Précisemment, on sait que dans le modèle Gaussien, $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathrm{X}^\top \mathrm{X})^{-1})$, et en particulier,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(\mathrm{X}^\top \mathrm{X})^{-1}]_{jj}}} \sim \mathcal{T}_{n-p-1}.$$

Ce résultat permet de tester l'hypothèse nulle $\mathrm{H}_0 : \beta_j = 0$ pour vérifier la pertinence de β_j (de manière générale, on peut tester une hypothèse de la forme $\mathrm{H}_0 : \beta_j = b$) et de construire des intervalles de confiance. Le théorême suivant donne une procédure pour tester une hypothèse affine générale sur les paramètres.

Proposition 3.6 *Soient* $\mathrm{A} \in \mathbb{R}^{q\times(p+1)}$ *de rang* $q \leqslant p+1$ *et* $b \in \mathbb{R}^q$ *connus. Sous les hypothèses fortes, la statistique*

$$\mathrm{F} = \frac{(\mathrm{A}\hat{\beta} - b)^\top [\mathrm{A}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{A}^\top]^{-1}(\mathrm{A}\hat{\beta} - b)}{q\hat{\sigma}^2}$$

suit, sous $\mathrm{H}_0 : \mathrm{A}\beta = b$, *une loi de Fisher* $\mathcal{F}_{q,n-p-1}$.

Le test des hypothèses $\beta_1 = ... = \beta_p = 0$ ou encore $\beta_p = 0$ sont des cas particuliers du théorême, correspondant respectivement à $\mathrm{A} = (0, \mathrm{I}) \in \mathbb{R}^{p\times(p+1)}, b = 0$ et $\mathrm{A} = (0, ..., 0, 1), b = 0$. Les hypothèses de la forme $\beta_j = \beta_{j'}$ correspondent également à des valeurs particulières de A et b.

Ce genre d'hypothèse sur les paramètres revient à considérer un modèle contraint

$$y = \mathrm{X}_c\beta_c + \epsilon,$$

où X_c est une matrice de rang $p+1-q$ dont l'image est inclue dans Im(X) (par exemple, pour tester l'hypothèse $\mathrm{H}_0 : \beta_j = 0$, on étudie le modèle sans la variable x_j). On définit alors la statistique de test

$$\mathrm{F} = \frac{n-p-1}{q}\frac{\mathrm{SCR}_c - \mathrm{SCR}}{\mathrm{SCR}} = \frac{\|(\mathrm{I} - \Pi_{\mathrm{X}_c})\epsilon\|^2 - \|(\mathrm{I} - \Pi_{\mathrm{X}})\epsilon\|^2}{q\hat{\sigma}^2} = \frac{\|(\Pi_{\mathrm{X}} - \Pi_{\mathrm{X}_c})\epsilon\|^2}{q\hat{\sigma}^2}$$

où SCR_c désigne la somme des carrés résiduelle dans le modèle contraint. Si l'hypothèse H_0 est vraie, F suit une loi de Fisher $\mathcal{F}_{q,n-p-1}$. On obtient la même statistique de test par le théorême 3.6.

3.5 Prediction

On observe un nouveau jeu de variables $x_{1,n+1}, ..., x_{p,n+1}$ et on cherche à prédire la valeur y_{n+1} correspondante. On note $\mathrm{X}_{n+1} = (1, x_{1,n+1}, ..., x_{p,n+1})$. Sous l'hypothèse de normalité (qui est essentielle ici), la prédiction $\hat{y}_{n+1} = \mathrm{X}_{n+1}\hat{\beta}$ suit une loi normale $\mathcal{N}(\mathrm{X}_{n+1}\beta, \sigma^2\mathrm{X}_{n+1}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}_{n+1}^\top)$ et est indépendante de $y_{n+1} = \mathrm{X}_{n+1}\beta + \epsilon_{n+1}$. On montre alors facilement que

$$\frac{\hat{y}_{n+1} - y_{n+1}}{\hat{\sigma}\sqrt{\mathrm{X}_{n+1}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}_{n+1}^\top}} \sim \mathcal{T}_{n-p-1},$$

ce qui permet de construire un intervalle de prédiction, qui n'est valable que sous l'hypothèse de normalité.

3.6 Vérification des hypothèses

La pluparts des résultats de la régression linéaires reposent sur les hypothèses de normalité, homoscédasticité et non-corrélations des résidus. Il est donc utile de pouvoir vérifier la validité de ces hypothèses.

- Normalité : Pour vérifier si les bruits ϵ_i sont Gaussiens, on effectue un test de normalité sur les résidus $\hat{\epsilon}_i$. En effet, la normalité de ϵ entraîne la normalité de $\hat{\epsilon}$. Plusieurs tests existent comme le test de Shapiro-Wilk (commande `shapiro.test` sour `R`) ou encore le test de Lilliefors (commande `lillie.test` du package `nortest`). Le diagramme quantile-quantile (ou qq-plot) permet également de vérifier graphiquement la normalité des résidus.

- Homoscédasticité (ou homogénéité) : Le test de Breusch-Pagan (commande `bptest` du package `lmtest`) permet de tester si la variance des bruits est constante. On peut également utiliser le test de White (commande `white.test` du package `bstats`). Graphiquement, l'hétéroscédasticité du bruit se traduit par une répartition d'ampleurs inégales du nuage de points autour de la droite de régression.

– Non-corrélation : Le test de Breusch-Godfrey (commande `bgtest` du package `lmtest`) permet de tester une corrélation à l'ordre 1 ou supérieur des bruits ϵ_i. Pour tester une corrélation à l'ordre 1, on peut également utiliser la statistique de Durbin-Watson (commande `dwtest` du package `lmtest`), définie par

$$\mathrm{D} = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}.$$

On montre facilement que D est comprise entre 0 et 4 mais sa loi sous l'hypothèse nulle de non-corrélation n'est pas une loi usuelle. Une règle de décision couramment utilisée est de conclure qu'il n'y a pas de corrélation entre ϵ_i et ϵ_{i+1} si la statistique de Durbin-Watson est comprise entre 1 et 3.

3.7 Détection d'observations atypiques

Les observations atypiques (outliers) sont les observations qui s'éloignent particulièrement de la valeur attendue estimée par le modèle. Il existe plusieurs outils permettant de détecter des observations atypiques. Une fois une valeur aberrante détectée, on peut choisir de la supprimer (par exemple si on conclut que celle-ci est dûe à une erreur de mesure) afin d'améliorer l'estimation.

– Effet levier : Même si les bruits ϵ_i sont homoscédastiques, les résidus $\hat{\epsilon}_i$ n'ont généralement pas les mêmes variances. En effet, en notant $h_{ij}, i, j = 1, ..., p+1$ les entrées de la matrice de projection $\Pi_{\mathrm{X}} = \mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top$, l'égalité $\hat{\epsilon} = (\mathrm{I} - \Pi_{\mathrm{X}})\epsilon$ entraîne

$$\mathrm{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii}).$$

Les valeurs h_{ii} permettent donc de détecter les points y_i qui sont éloignés de la prédiction $\hat{y}_i$ en moyenne. La matrice Π_{X} est parfois appelée *hat-matrice* (et notée H) du fait que $\mathrm{H}y = \hat{y}$.

– Etude des résidus : Une observation atypique y_i peut être détectée à partir du résidu $\hat{\epsilon}_i$. Pour prendre en compte le fait que les résidus n'ont pas la même variance, on peut effectuer deux types de normalisation. Le *résidu standardisé* $\hat{\epsilon}_i$ correspond à la valeur

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.$$

Sous les hypothèses fortes, la loi de r_i ne dépend pas des paramètres β et σ^2. Cependant, le calcul exact de sa loi est difficile car $\hat{\epsilon}_i$ et $\hat{\sigma}^2$ ne sont pas indépendants. Pour obtenir une normalisation qui suit approximativement une loi de Student, on définit le *résidu studentisé* par

$$\mathrm{T}_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}},$$

où $\hat{\sigma}^2_{(i)}$ est l'estimateur de σ^2 obtenu sans la i-ème observation. L'estimateur $\hat{\sigma}^2_{(i)}$ est donc indépendant de ϵ_i et la statistique T_i suit approximativement une loi de Student $\mathcal{T}_{n-p-2}$. Les observations pour lesquelles $|\mathrm{T}_i|$ est supérieur à 2 peuvent donc être considérées comme atypiques.

– Distance de Cook : La distance de Cook de y_i mesure l'écart entre la vraie prédiction $\hat{y}$ et celle obtenue en enlevant la i-ème observation $(y_i, x_{1i}, ..., x_{pi})$. Elle permet donc d'évaluer l'impact de la i-ème observation sur la régression. Soit $\mathrm{X}_{(-i)}$ la matrice obtenue en enlevant la i-ème ligne $\mathrm{X}_i = (1, x_{1i}, ..., x_{pi})$ de la matrice X. On suppose ici que $n > p+1$ pour que $\mathrm{X}_{(-i)}$ soit de plein rang. On note $\hat{\beta}^{(i)}$ l'estimateur des moindres carrés obtenu à partir de l'échantillon sans la i-ème observation :

$$\hat{\beta}^{(i)} = (\mathrm{X}_{(-i)}^\top\mathrm{X}_{(-i)})^{-1}\mathrm{X}_{(-i)}^\top y_{(-i)}$$

où $y_{(-i)} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)^\top$. De même, on note $\hat{y}^{(i)} = \mathrm{X}\hat{\beta}^{(i)}$ le vecteur de prédiction associé. On définit alors la distance de Cook de l'observation i par

$$\mathrm{D}_i = \frac{1}{p+1}\,\frac{\|\hat{y}^{(i)} - \hat{y}\|^2}{\hat{\sigma}^2}.$$

Une grande distance de Cook D_i signifie que l'ajout de la i-ème observation modifie considérablement la régression. En pratique, on peut utiliser le seuil $\mathrm{D}_i > 1$ pour décider que l'observation y_i est influente.

Le calcul de la distance de Cook d'une observation où de $\hat{\sigma}^2_{(i)}$ ne nécessite pas de refaire tous les calculs de la régression dans le modèle sans l'observation i, comme on peut voir dans la proposition suivante.

Proposition 3.7 *La distance de Cook* D_i *vérifie*

$$\mathrm{D}_i = \frac{1}{p+1}\frac{h_{ii}}{(1-h_{ii})^2}\frac{\hat{\epsilon}_i^2}{\hat{\sigma}^2} = \frac{1}{p+1}\frac{h_{ii}}{1-h_{ii}}r_i^2.$$

De plus, l'estimateur de la variance $\hat{\sigma}^2_{(i)}$ *obtenu dans le modèle sans la* i*-ème observation est donné par*

$$\hat{\sigma}^2_{(i)} = \frac{1}{n-p-2}\left[(n-p-1)\hat{\sigma}^2 - \frac{\hat{\epsilon}_i^2}{1-h_{ii}}\right].$$

On voit en particulier que la distance de Cook permet de synthétiser l'information sur les données influentes contenue dans l'effet levier (via le terme $h_{ii}/1-h_{ii}$) et le résidu standardisé r_i.

3.8 Multicolinéarité

On observe des problèmes de multicolinéarité lorsqu'au moins une variable explicative est très corrélée aux autres régresseurs. Cela signifie d'une certaine façon que l'information apportée par cette variable est redondante car contenue dans les autres variables. Mathématiquement, la multicolinéarité conduit à des valeurs propres de $\mathrm{X}^\top\mathrm{X}$ proches de zéro. Dans ce cas, l'estimateur des moindres carrés $\hat{\beta}$ n'est pas performant car sa variance $\sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}$ explose.

Un moyen simple de détecter si une variable x_j est corrélée au reste des régresseurs est d'effectuer une régression linéaire de x_j sur les autres variables explicatives $x_k, k \neq j$. On peut alors calculer le coefficient de détermination R_j^2 correspondant et vérifier s'il est proche de 1. Le *variance inflation factor* (VIF) est défini par

$$\mathrm{VIF}(x_j) = \frac{1}{1-\mathrm{R}_j^2},$$

le coefficient R_j^2 étant toujours strictement inférieur à 1 lorsque X est de plein rang. On conclut généralement à un problème de multicolinéarité pour x_j si $\mathrm{VIF}(x_j) > 5$, ou de manière équivalente si $\mathrm{R}_j^2 > 0.8$.

En pratique, il faut toujours vérifier en premier lieu les problèmes de multicolinéarité lorsque l'on efectue une régression linéaire. Cela implique de calculer le VIF pour chaque variable explicative. Si jamais plusieurs variables ont un VIF supérieur à 5, on supprime seulement la variable dont le VIF est le plus élevé. Puis, on réitère la procédure dans le modèle sans la variable supprimée jusqu'à ce que toutes les variables explicatives aient des VIFs acceptables.

3.9 Moindres carrés généralisés

On s'intéresse maintenant à la situation où les hypothèses de non-corrélation des bruits n'est pas vérifiée. On suppose ici que ϵ est centré de matrice de variance $\sigma^2\mathrm{V}$, où V est une matrice définie positive connue. Dans ce modèle, le théorème de Gauss-Markov n'est plus valable.

Proposition 3.8 *Dans le modèle de régression linéaire* $y = \mathrm{X}\beta + \epsilon$ *avec* $\mathbb{E}(\epsilon) = 0$ *et* $\mathrm{var}(\epsilon) = \sigma^2\mathrm{V}$*, l'estimateur linéaire sans biais de* β *de variance minimale est donné par*

$$\hat{\beta}_{\mathrm{G}} = (\mathrm{X}^\top\mathrm{V}^{-1}\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{V}^{-1}y.$$

L'estimateur $\hat{\beta}_{\mathrm{G}}$ est appelé l'estimateur des moindres carrés généralisés (MCG). Sa variance

$$\mathrm{var}(\hat{\beta}_{\mathrm{G}}) = (\mathrm{X}^\top \mathrm{V}^{-1} \mathrm{X})^{-1}$$

est donc minimale parmi les estimateurs linéaires sans biais. Dans le cas Gaussien, on vérifie facilement que $\hat{\beta}_{\mathrm{G}}$ est l'estimateur du maximum de vraisemblance.

4 Analyse de variance et de covariance

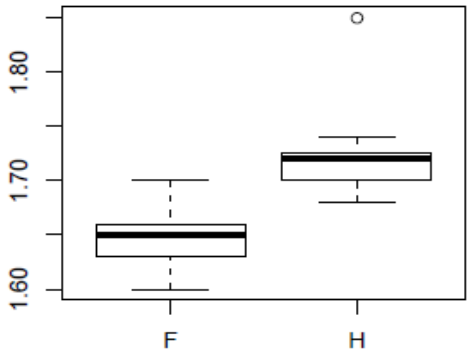
L'analyse de variance (ANOVA) a pour objectif d'expliquer une variable aléatoire Y quantitative à partir de variables explicatives qualitatives, appelées *facteurs*. On compare les moyennes empiriques des observations de Y pour les différentes modalités prises par les facteurs.

4.1 Analyse de variance à un facteur

Soit $\mathrm{J} \geqslant 2$ modalités $\mathrm{A}_1, ..., \mathrm{A}_\mathrm{J}$, on observe J échantillons indépendants $(y_{11}, ..., y_{n_1 1}), ..., (y_{1\mathrm{J}}, ..., y_{n_\mathrm{J}\mathrm{J}})$ de tailles $n_1, ..., n_\mathrm{J}$ suivant le modèle

$$y_{ij} = \mu_j + \epsilon_{ij},\ j = 1, ..., \mathrm{J}\ ,\ i = 1, ..., n_j.$$

On note $n = \sum_{j=1}^{\mathrm{J}} n_j$. L'observation y_{ij} correspond à une réalisation de Y dans la modalité A_j. Par exemple, on veut évaluer la taille moyenne d'une population en fonction du sexe. On a donc une variable quantitative y (la taille) et une variable explicative qualitative (le sexe) comprenant deux modalités. On peut représenter graphiquement les données par des boîtes à moustaches.



Le modèle d'analyse de variance peut s'écrire comme un modèle de régression linéaire multiple avec comme variable à expliquer $y = (y_{11}, ..., y_{n_1 1}, y_{12}, ..., y_{n_2 2}, ..., y_{1\mathrm{J}}, ..., y_{n_\mathrm{J}\mathrm{J}})^\top \in \mathbb{R}^n$ et comme variables explicatives les indicatrices des modalités $x_1 = \mathbf{1}_{\mathrm{A}_1} = (1, ..., 1, 0, ..., 0)^\top, x_2 = \mathbf{1}_{\mathrm{A}_2} = (0, ..., 0, 1, ..., 1, 0, ..., 0)^\top$ etc... En notant $\epsilon = (\epsilon_{11}, ..., \epsilon_{n_1 1}, ..., \epsilon_{1\mathrm{J}}, ..., \epsilon_{n_\mathrm{J}\mathrm{J}})^\top$ et $\mathrm{X} = (\mathbf{1}_{\mathrm{A}_1}, ..., \mathbf{1}_{\mathrm{A}_\mathrm{J}}) \in \mathbb{R}^{n \times \mathrm{J}}$, on a bien

$$y = \mathrm{X}\beta + \epsilon,$$

avec $\beta = (\mu_1, ..., \mu_\mathrm{J})^\top$. Le modèle ne contient pas la constante car ajouter celle-ci rendrait le modèle sur-identifié avec une matrice X qui ne serait pas de plein rang (le vecteur constant $\mathbf{1}$ est égal à la somme des colonnes x_j). On peut cependant reparamétrer le modèle en prenant comme variables explicatives $\mathbf{1}, \mathbf{1}_{\mathrm{A}_2}, ..., \mathbf{1}_{\mathrm{A}_\mathrm{J}}$ de manière à retrouver un modèle de régression linéaire avec constante. Avec cette paramétrisation, on doit définir $\beta_0 = \mu_1$ et $\beta_j = \mu_{j+1} - \mu_1$ pour $j = 1, ..., \mathrm{J} - 1$.

L'estimation des paramètres μ_j se fait naturellement par les moyennes empiriques sur chaque modalité

$$\hat{\mu}_j = \overline{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

Proposition 4.1 *L'estimateur $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_\mathrm{J})^\top$ est l'estimateur des moindres carrés du modèle de régression linéaire correspondant. Il ne dépend pas du choix de la paramétrisation.*

En supposant les ϵ_{ij} centrés, non-corrélés et de même variance σ^2, l'estimateur de σ^2 est celui de la régression linéaire, donné par

$$\hat{\sigma}^2 = \frac{1}{n-\mathrm{J}} \sum_{j=1}^{\mathrm{J}} \sum_{i=1}^{n_j} (y_{ij} - \hat{\mu}_j)^2.$$

La renormalisation se fait par $n-\mathrm{J}$ et non $n-\mathrm{J}-1$ car le modèle ne contient pas la constante.

On s'intéresse maintenant à des tests d'hypothèses. On se place maintenant sous l'hypothèse forte de normalité des bruits ϵ_{ij}. Premièrement, une question naturelle dans ce modèle est de savoir si les modalités ont une influence sur la variable y. Cela revient à tester l'égalité simultanée des moyennes μ_j, soit l'hypothèse $\mathrm{H}_0 : \mu_1 = ... \ \mu_{\mathrm{J}}$.

Proposition 4.2 *La statistique*

$$\mathrm{F} = \frac{n-\mathrm{J}}{\mathrm{J}-1} \frac{\sum_{j=1}^{\mathrm{J}} n_j (\overline{y}_{.j} - \overline{y})^2}{\sum_{j=1}^{\mathrm{J}} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{.j})^2}$$

suit, sous $\mathrm{H}_0 : \mu_1 = ... = \mu_{\mathrm{J}}$ *une loi de Fisher* $\mathcal{F}_{\mathrm{J}-1, n-\mathrm{J}}$.

Dans le cas particulier de l'analyse de variance, la décomposition $\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR}$ s'écrit donc

$$\underbrace{\sum_{j=1}^{\mathrm{J}} \sum_{i=1}^{n_j} (y_{ij} - \overline{y})^2}_{\mathrm{SCT}} = \underbrace{\sum_{j=1}^{\mathrm{J}} n_j (\overline{y}_{.j} - \overline{y})^2}_{\mathrm{SCE}} + \underbrace{\sum_{j=1}^{\mathrm{J}} \sum_{i=1}^{n_j} (y_{ij} - \overline{y}_{.j})^2}_{\mathrm{SCR}}$$

Dans ce cadre, la quantité $\frac{1}{n}\mathrm{SCT}$ est parfois appelée la *variance totale*, $\frac{1}{n}\mathrm{SCE}$ la *variance interclasses* et $\frac{1}{n}\mathrm{SCR}$ la *variance intraclasses*.

On peut également être amené à tester seulement l'égalité entre deux moyennes μ_j et $\mu_{j'}$. Dans ce cas, on peut utiliser un test de Student.

Proposition 4.3 *La statistique*

$$\mathrm{T} = \frac{\hat{\mu}_j - \hat{\mu}_{j'}}{\hat{\sigma}\sqrt{\frac{1}{n_j} + \frac{1}{n_{j'}}}}$$

suit, sous $\mathrm{H}_0 : \mu_j = \mu_{j'}$ *une loi de Student* $\mathcal{T}_{n-\mathrm{J}}$.

On a vu comment tester l'égalité simultanée de toutes les moyennes et l'égalité de deux moyennes. On s'intéresse maintenant à tester une hypothèse de la forme $\mathrm{H}_0 : \mu_{j_1} = \mu_{j'_1}, ..., \mu_{j_q} = \mu_{j'_q}$ où $(j_1, j'_1), ..., (j_q, j'_q)$ est une collection de paires d'indices différents de $\{1, ..., \mathrm{J}\}$ (attention, on impose évidemment $j_k \neq j'_k$ mais pas nécessairement $j_k \neq j_l$). Pour tester ce genre d'hypothèses, deux procédures sont envisageables. Un premier moyen est d'appliquer la correction de Bonferroni. En notant T_k la statistique de test pour l'hypothèse $\mathrm{H}_0 : \mu_{j_k} = \mu_{j'_k}$, $k = 1, ..., q$, on utilise que, pour un niveau $a \in]0, 1[$ donné,

$$\mathbb{P}\left(\exists k, |\mathrm{T}_k| > t_{n-\mathrm{J}}(1-\tfrac{a}{2})\right) = \mathbb{P}\left(\bigcup_{k=1}^{q} \left\{|\mathrm{T}_k| > t_{n-\mathrm{J}}(1-\tfrac{a}{2})\right\}\right) \leqslant \sum_{k=1}^{q} \mathbb{P}\left(|\mathrm{T}_k| > t_{n-\mathrm{J}}(1-\tfrac{a}{2})\right) = qa,$$

où $t_{n-\mathrm{J}}(.)$ désigne la fonction quantile de la loi de Student $\mathcal{T}_{n-\mathrm{J}}$. On déduit que la procédure de test qui consiste à rejeter $\mathrm{H}_0 : \mu_{j_1} = \mu_{j'_1}, ..., \mu_{j_q} = \mu_{j'_q}$ s'il existe un k pour lequel $|\mathrm{T}_k| > t_{n-\mathrm{J}}(1-\frac{\alpha}{2q})$ a une erreur de première espèce inférieure à α.

On peut également effectuer un test exact pour cette hypothèse en utilisant l'écriture du modèle de régression linéaire. On remarque en effet que l'hypothèse $\mathrm{H}_0 : \mu_{j_1} = \mu_{j'_1}, ..., \mu_{j_q} = \mu_{j'_q}$ correspond à un cas

particulier du théorême 3.6, où A est une matrice de taille J $\times$ q et $b = 0$.

Le test de Bartlett (commande `bartlett.test` sous `R`) permet de vérifier l'hypothèse d'homoscédasticité $H_0 : \sigma_1^2 = ... = \sigma_J^2$ où σ_j^2 représente la variance de Y dans la modalité A_j. Le test de Bartlett nécessite la normalité des données, qui peut être testée au préalable par un moyen classique (test de Shapiro-Wilk par exemple). Si les données ne sont pas Gaussiennes, le test d'homogénéité de Levene (commande `levene.test` de la library `car`) est préférable. Il est construit à partir des variables $z_{ij} = |y_{ij} - \overline{y}_{.j}|$, en considérant la statistique

$$F = \frac{n - J}{J - 1} \frac{\sum_{j=1}^{J} n_j (\overline{z}_{.j} - \overline{z})^2}{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (z_{ij} - \overline{z}_{.j})^2}$$

qui suit approximativement sous $H_0 : \sigma_1^2 = ... = \sigma_J^2$, une loi de Fisher $\mathcal{F}_{J-1,n-J}$.

Le test de Brown-Forsythe utilise la médiane au lieu de la moyenne $\overline{y}_{.j}$ pour construire z_{ij}, ce qui peut parfois conduire à un test plus robuste quand la loi des observations est trop différente de la loi normale.

4.2 Analyse de variance à deux facteurs

On suppose maintenant que la variable y dépend de deux facteurs, notés A et B ayant respectivement J et K modalités. En présence de plusieurs facteurs, le problème de l'intéractions entre les facteurs apparaît. On observe les observations suivantes

$$y_{ijk} = \mu + \alpha_j + \delta_k + \gamma_{jk} + \epsilon_{ijk}, j = 1, ..., J,\ k = 1, ..., K,\ i = 1, ..., n_{jk}$$

où les ϵ_{ijk} sont iid de loi $\mathcal{N}(0, \sigma^2)$. Le nombre n_{jk} est le nombre d'observations qui sont simultanément de modalités A_j et B_k. On note $n_{j.} = \sum_{k=1}^{K} n_{jk}$, $n_{.k} = \sum_{j=1}^{J} n_{jk}$ et $n = \sum_{j=1}^{J} n_{j.} = \sum_{k=1}^{K} n_{.k}$. Parmi les paramètres, μ représente l'effet général, α_j l'effet du niveau j du premier facteur, δ_k l'effet du niveau k du second facteur et γ_{jk} l'effet de l'intéraction entre les niveaux j et k des deux facteurs.

L'effet d'intéraction existe quand le niveau d'un facteur modifie l'influence de l'autre facteur sur la variable Y. Dans l'exemple utilisé précédemment, la taille moyenne dans une population est modélisée en tenant compte du sexe. Si l'on ajoute un deuxième facteur (par exemple l'âge séparé en trois modalités "enfant", "adolescent"' et "adulte"), on peut évaluer l'intéraction entre les facteurs en mesurant par exemple si l'écart moyen de taille entre hommes et femmes et le même chez les adolescents et chez les adultes.

Le modèle $y_{ijk} = \mu + \alpha_j + \delta_k + \gamma_{jk} + \epsilon_{ijk}$ est sur-identifié, on impose donc les contraintes sur les paramètres

$$\sum_{k=1}^{K} \alpha_k = \sum_{j=1}^{J} \delta_j = 0\ ,\ \sum_{k=1}^{K} \gamma_{jk} = 0,\ \forall j = 1, ..., J\ ,\ \sum_{j=1}^{J} \gamma_{jk} = 0,\ \forall k = 1, .., K.$$

On peut vérifier que ces contraintes diminuent de J+K+1 le nombre de degrés de liberté des paramètres. Dans ce modèle, il y a J $\times$ K paramètres à estimer, autant que de classes formées par les différents croisements des modalités A_j et B_k. En notant

$$\overline{y}_{.jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}\ ,\ \overline{y}_{..k} = \frac{1}{n_{.k}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} \overline{y}_{ijk}\ ,\ \overline{y}_{.j.} = \frac{1}{n_{j.}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} \overline{y}_{ijk} \quad \text{et} \quad \overline{y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} y_{ijk},$$

les paramètres sont estimés par

$$\hat{\mu} = \overline{y}\ ,\ \hat{\alpha}_j = \overline{y}_{.j.} - \overline{y}\ ,\ \hat{\delta}_k = \overline{y}_{..k} - \overline{y} \quad \text{et} \quad \hat{\gamma}_{jk} = \overline{y}_{.jk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y}.$$

La prédiction $\hat{y}_{ijk}$ est naturellement la moyenne sur la classe $\mathrm{A}_j \cap \mathrm{B}_k$, $\hat{y}_{ijk} = \hat{\mu} + \hat{\alpha}_j + \hat{\delta}_k + \hat{\gamma}_{jk} = \overline{y}_{.jk}$. Comme pour l'analyse de variance à un facteur, le modèle peut s'écrire comme un modèle de régression linéaire particulier. On peut par exemple considérer le modèle équivalent

$$y = \sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \beta_{jk} \mathbf{1}_{\mathrm{A}_j \cap \mathrm{B}_k} + \epsilon,$$

où $\mathbf{1}_{\mathrm{A}_j \cap \mathrm{B}_k} \in \mathbb{R}^n$ est l'indicatrice des modalités A_j et B_k simultanément. Les paramètres correspondants sont les coefficients $\beta_{jk} = \mu + \alpha_j + \delta_k + \gamma_{jk}$ pour $j = 1, ..., \mathrm{J}$ et $k = 1, ..., \mathrm{K}$. Les estimateurs $\hat{\mu}$, $\hat{\alpha}_j$, $\hat{\delta}_k$ et $\hat{\gamma}_{jk}$ correspondent ici aussi à l'estimation des moindres carrés. On a en particulier $\hat{\beta}_{jk} = \overline{y}_{.jk} = \hat{\mu} + \hat{\alpha}_j + \hat{\delta}_k + \hat{\gamma}_{jk}$.

Proposition 4.4 *Si le plan d'expérience est équilibré, c'est-à-dire que $n_{jk} = \frac{n}{\mathrm{JK}}$ pour tout j, k (le nombre d'observations dans chaque classe est identique), on a la décomposition suivante*

$$\underbrace{\sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y})^2}_{\mathrm{SCT}} = \underbrace{\frac{n}{\mathrm{J}} \sum_{j=1}^{\mathrm{J}} (\overline{y}_{.j.} - \overline{y})^2}_{\mathrm{SCA}} + \underbrace{\frac{n}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} (\overline{y}_{..k} - \overline{y})^2}_{\mathrm{SCB}} + \underbrace{\frac{n}{\mathrm{JK}} \sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} (\overline{y}_{.jk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y})^2}_{\mathrm{SCAB}}$$
$$+ \underbrace{\sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{.jk})^2}_{\mathrm{SCR}}$$

Afin de tester l'influence du facteur A, on considère la statistique

$$\mathrm{F_A} = \frac{n - \mathrm{JK}}{\mathrm{J} - 1} \frac{\frac{n}{\mathrm{J}} \sum_{j=1}^{\mathrm{J}} (\overline{y}_{.j.} - \overline{y})^2}{\sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{.jk})^2} = \frac{n - \mathrm{JK}}{\mathrm{J} - 1} \frac{\mathrm{SCA}}{\mathrm{SCR}}$$

qui suit, sous $\mathrm{H}_0 : \alpha_1 = ... = \alpha_{\mathrm{J}} = 0$, une loi de Fisher $\mathcal{F}_{\mathrm{J}-1, n-\mathrm{JK}}$. La procédure est bien sûr également valable pour tester l'influence du facteur B par l'hypothèse $\mathrm{H}_0 : \delta_1 = ... = \delta_{\mathrm{K}} = 0$. Enfin, pour tester la présence d'interaction entre les facteurs et l'hypothèse $\mathrm{H}_0 : \gamma_{jk} = 0, \forall j, k$, la statistique

$$\mathrm{F_{AB}} = \frac{n - \mathrm{JK}}{(\mathrm{J} - 1)(\mathrm{K} - 1)} \frac{\frac{n}{\mathrm{JK}} \sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} (\overline{y}_{.jk} - \overline{y}_{.j.} - \overline{y}_{..k} + \overline{y})^2}{\sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{.jk})^2} = \frac{n - \mathrm{JK}}{(\mathrm{J} - 1)(\mathrm{K} - 1)} \frac{\mathrm{SCAB}}{\mathrm{SCR}}$$

suit sous H_0 une loi de Fisher $\mathcal{F}_{(\mathrm{J}-1)(\mathrm{K}-1), n-\mathrm{JK}}$. Lorsque le plan n'est pas équilibré, la décomposition de la proposition 4.4 n'est plus valable. Pour tester l'influence de chaque facteur séparemment, l'idée reste la même. Pour le facteur A par exemple, la statistique

$$\mathrm{F_A} = \frac{n - \mathrm{JK}}{\mathrm{J} - 1} \frac{\sum_{j=1}^{\mathrm{J}} n_{j.} (\overline{y}_{.j.} - \overline{y})^2}{\sum_{j=1}^{\mathrm{J}} \sum_{k=1}^{\mathrm{K}} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{.jk})^2}$$

permet également de tester $\mathrm{H}_0 : \alpha_1 = ... = \alpha_{\mathrm{J}} = 0$ et sa loi reste identique au cas équilibré. En revanche, tester l'interaction entre les facteurs n'est plus aussi directe. En effet, les vecteurs $\Pi_{\mathrm{A}}(y - \overline{y})$ et $\Pi_{\mathrm{B}}(y - \overline{y})$ ne sont pas nécessairement orthogonaux quand le plan est déséquilibré. Le moyen le plus simple pour tester la présence d'interaction dans le modèle d'analyse de variance à deux facteurs est sans doute d'utiliser la représentation par le modèle de régression linéaire. Le modèle s'écrit

$$y = \mathrm{X}\beta + \epsilon,$$

où $\mathrm{X} = (\mathbf{1}_{\mathrm{A}_1 \cap \mathrm{B}_1}, ..., \mathbf{1}_{\mathrm{A}_1 \cap \mathrm{B}_\mathrm{K}}, ..., \mathbf{1}_{\mathrm{A}_\mathrm{J} \cap \mathrm{B}_1}, ..., \mathbf{1}_{\mathrm{A}_\mathrm{J} \cap \mathrm{B}_\mathrm{K}})$ et $\beta = (\beta_{11}, ..., \beta_{1\mathrm{K}}, ..., \beta_{\mathrm{J}1}, ..., \beta_{\mathrm{JK}})^\top$. L'absence d'interaction signifie que les observations sont issues du modèle contraint

$$y = \mathrm{X}_c \beta_c + \epsilon,$$

où $X_c = (\mathbf{1}_{A_1}, ..., \mathbf{1}_{A_J}, \mathbf{1}_{B_1}, ..., \mathbf{1}_{B_{K-1}})$ et $\beta_c = (\alpha_1, ..., \alpha_J, \delta_1, ..., \delta_{K-1})^\top$ (la variable $\mathbf{1}_{B_K}$ n'est pas incluse pour rendre le modèle identifiable). Le modèle initial contient JK paramètres et le modèle contraint n'en contient plus que $J + K - 1$, on a donc $q = JK - (J + K - 1) = (J - 1)(K - 1)$ pour les notations du théorême 3.6. On regarde alors à la statistique

$$F_{AB} = \frac{n - JK}{(J-1)(K-1)} \frac{SCR_c - SCR}{SCR},$$

qui suit sous l'hypothèse nulle d'absence d'interaction une loi de Fisher $\mathcal{F}_{(J-1)(K-1),n-JK}$. Cette statistique n'a pas d'expression simplifiée quand le plan n'est pas équilibré.

4.3 Analyse de covariance

Lorsque les variables explicatives contiennent également des variables quantitatives, on parle d'analyse de covariance (ANCOVA). On définit alors un modèle de régression linéaire entre Y et les variables explicatives quantitatives pour chaque classe déterminée par les différentes modalités des facteurs. Pour simplifier, on considère la situation où on dispose d'un seul facteur A à J modalités et une variable explicative quantitative X, mais l'idée se généralise facilement à plusieurs variables. Le modèle s'écrit

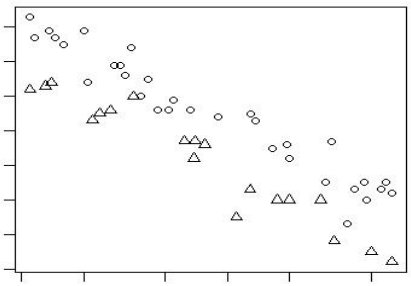
$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}, \ j = 1, ..., J, \ i = 1, ..., n_j$$

où les ϵ_{ij} sont iid de loi $\mathcal{N}(0, \sigma^2)$. La forme matricielle est donnée par

$$y = X\beta + \epsilon,$$

où $X = (\mathbf{1}_{A_1}, x_1, ..., \mathbf{1}_{A_J}, x_J)$ et $\beta = (\beta_{01}, \beta_{11}, ..., \beta_{0J}, \beta_{1J})^\top$. Ici, le support des variables colonnes $x_1, ..., x_J$ est restreint à leurs modalités, $x_1 = (x_{11}, ..., x_{1n_1}, 0, ..., 0)^\top$, $x_2 = (0, ..., 0, x_{21}, ..., x_{2n_2}, 0, ..., 0)^\top$ etc...

Pour représenter graphiquement les données, on peut utiliser un nuage de points en distinguant les points (par des formes ou couleurs différentes) selon leur modalité, par exemple comme sur le graphique suivant.



On distingue ainsi les données selon l'appartenance à la première modalité (cercles) ou à la seconde modalité (triangles). L'analyse de covariance permet alors de tester plusieurs hypothèses, en comparant le modèle à des modèles contraints n'intégrant que l'effet de X, que l'effet de A ou que l'effet d'interaction.

- Pour tester l'interaction entre les variables explicatives A et X, on considère dans un premier temps l'hypothèse nulle $H_0^{(1)} : \beta_{11} = ... = \beta_{1J}$. S'il n'y a pas d'interaction, les pentes de la régression de Y sur X sont toutes identiques. Le modèle contraint s'écrit donc $y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij}$ qui comprend $J + 1$ paramètres à estimer. On a ici $q = JK - J - 1$ et la statistique de test

$$F^{(1)} = \frac{n - JK}{JK - J - 1} \frac{SCR_c^{(1)} - SCR}{SCR}$$

suit, sous $H_0^{(1)} : \beta_{11} = ... = \beta_{1J}$, une loi de Fisher $\mathcal{F}_{JK-J-1,n-JK}$.

– Si l'interaction n'est pas significative, on peut alors vouloir tester l'effet de X dans le modèle contraint précédent par le biais de l'hypothèse plus forte $\mathrm{H}_0^{(2)} : \beta_1 = 0$. Le modèle contraint s'écrit alors $y_{ij} = \beta_{0j} + \epsilon_{ij}$. C'est le modèle d'analyse de variance à un facteur, qui comprend J paramètres à estimer, le modèle initial étant ici $y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij}$. La statistique de test est donc

$$\mathrm{F}^{(2)} = \frac{n-\mathrm{JK}+\mathrm{J}+1}{\mathrm{J}(\mathrm{K}-1)} \frac{\mathrm{SCR}_c^{(2)} - \mathrm{SCR}_c^{(1)}}{\mathrm{SCR}_c^{(1)}},$$

et suit sous $\mathrm{H}_0^{(2)}$ une loi de Fisher $\mathcal{F}_{\mathrm{J(K-1)},n-\mathrm{JK+J+1}}$.

– Enfin, si aucun des deux tests précédents n'est significatif, on peut vouloir tester l'effet du facteur A. On considère comme modèle initial $y_{ij} = \beta_{0j} + \epsilon_{ij}$ et comme modèle contraint $y_{ij} = \beta_0 + \epsilon_{ij}$ pour tester l'hypothèse $\mathrm{H}_0^{(3)} : \beta_{01} = ... = \beta_{0\mathrm{J}}$. La statistique de test est donc

$$\mathrm{F}^{(3)} = \frac{n-\mathrm{J}(\mathrm{K}-1)}{\mathrm{J}-1} \frac{\mathrm{SCR}_c^{(3)} - \mathrm{SCR}_c^{(2)}}{\mathrm{SCR}_c^{(2)}},$$

qui suit sous $\mathrm{H}_0^{(3)}$ une loi de Fisher $\mathcal{F}_{\mathrm{J-1},n-\mathrm{J(K-1)}}$.

5 Sélection de modèle

Les méthodes de sélection de modèle ont pour objectif d'identifier les variables pertinentes du modèle, c'est-à-dire celles qui apportent de l'information sur Y. Si une des variables explicatives X_j n'apporte pas d'information supplémentaire sur Y (ce qui se traduit par $\beta_j = 0$), il est évidemment préférable de ne pas l'inclure dans le modèle. Il est donc judicieux de s'autoriser à éliminer des variables du modèles initial si celles-ci sont jugées non pertinentes. Paradoxalement, la suppression d'une variable X_j peut parfois être bénéfique même si celle-ci est corrélée avec Y. Pour s'en convaincre, on peut remarquer par exemple que l'erreur quadratique de $\hat{\beta}_j$ vaut $\mathrm{var}(\hat{\beta}_j) = \sigma^2[(\mathrm{X}^\top\mathrm{X})^{-1}]_{jj}$ alors que l'erreur commise en excluant x_j du modèle, ce qui revient à estimer β_j par zéro, est $\mathbb{E}(\beta_j - 0)^2 = \beta_j^2$. Il est donc préférable, au sens du coût quadratique, de ne pas inclure la variable x_j dans le modèle si $\sigma^2[(\mathrm{X}^\top\mathrm{X})^{-1}]_{jj} > \beta_j^2$. Cette remarque est également valable pour la variable constante $\mathbf{1}$, qui doit être considérée comme une variable explicative comme les autres, pouvant être supprimée du modèle si elle est jugée non pertinente.

La sélection de modèle présente plusieurs avantages. Premièrement, elle permet de fournir une interprétation sur l'existence ou non d'un lien entre les variables (par exemple, observer expérimentalement que la vitesse d'un objet en chute libre ne dépend pas directement de sa masse). Deuxièmement, diminuer la dimension du modèle permet de diminuer la variance de la prédiction. Enfin, limiter le nombre de variables est un bon moyen d'éviter le sur-ajustement des données.

Le modèle global contient donc $p+1$ variables explicatives, et on peut considérer comme modèle potentiel tout modèle obtenu avec un sous-ensemble $m \subseteq \{\mathbf{1}, x_1, ..., x_p\}$ de ces variables. Il y a donc 2^{p+1} modèles possibles. Pour un modèle $m \subseteq \{\mathbf{1}, x_1, ..., x_p\}$, on note Π_m le projecteur orthogonal sur l'espace engendré par les variables $x_j \in m$ et $|m|$ le cardinal de m. Le prédicteur correspondant au modèle m est donc $\hat{y}^{(m)} = \Pi_m y$.

5.1 Sélection par tests d'hypothèse

Pour chaque variable explicative, la question se pose de savoir s'il est préférable de l'inclure ou non au modèle. Il existe de nombreuses méthodes permettant de sélectionner les variables pertinentes. Un premier outil pratique permettant de juger de la pertinence de x_j est de tester l'hypothèse $\mathrm{H}_0 : \beta_j = 0$. Il existe alors plusieurs façons de faire, qui ne conduisent pas forcément au même modèle final.

- Méthode descendante (backward elimination) : On part du modèle comprenant toutes les variables explicatives. A chaque étape, la variable ayant la plus grande p-value du test de Student (ou de Fisher) de nullité du paramètre est supprimée du modèle si la p-value est supérieure à un seuil choisi à l'avance (généralement 0.10 ou 0.05). Attention, il est important d'éliminer les variables une par une. La procédure s'arrête lorsque toutes les variables sont significatives.

- Méthode ascendante (forward selection) : On effectue la régression linéaire de y sur chacune des variables explicatives séparément. On conserve la variable la plus pertinente, c'est-à-dire, celle dont la p-value est la plus faible. On réitère le procédé en introduisant les variables une par une et en ne conservant que la variable dont l'apport est le plus significatif. On s'arrête dès qu'aucune des variables pas encore introduites n'est jugée significative.

- Méthode pas-à-pas (stepwise selection) : On part soit du modèle global, soit du modèle sans variables et on évalue à chaque fois la significativité de chaque variable supprimée ou rajoutée au modèle. On s'arrête dès que le modèle ne peut être modifié sans améliorer la significativité.

L'utilisation des tests pour la sélection de variables ne permet que de comparer, à chaque étape, deux modèles emboîtés (c'est-à-dire pour lesquels un des modèles contient toutes les variables de l'autre). Ces critères sont limités car ils permettent seulement de juger de la pertinence de chaque variable individuellement. Or, le choix du meilleur modèle doit tenir compte de la significativité des variables et de leurs interactions. Il est donc souvent préférable d'utiliser un critère universel qui permet de comparer des modèles de manière plus globale.

5.2 Coefficient de détermination

Le coefficient de détermination d'un modèle m est la quantité

$$\mathrm{R}^2(m) = \frac{\|\hat{y}^{(m)} - \overline{y}\mathbf{1}\|^2}{\|y - \overline{y}\mathbf{1}\|^2},$$

qui évalue la part de y expliquée par le modèle. Utiliser le coefficient de détermination pour comparer plusieurs modèles (en choisissant le modèle avec le R^2 le plus grand) va conduire à choisir le modèle complet, qui colle le plus aux données. Ce critère ne tient pas compte d'un possible sur-ajustement et pour cette raison, n'est souvent pas approprié pour la sélection de modèle.

5.3 Coefficient de détermination ajusté

Si on s'intéresse à la relation stochastique sous-jacente

$$\mathrm{Y} = \mathbb{E}(\mathrm{Y}|\mathrm{X}) + \epsilon$$

entre les variables, on peut considérer que le coefficient de détermination R^2 est une estimation du R^2 théorique défini par

$$\mathrm{R}^2_{th} = \frac{\mathrm{var}(\mathbb{E}(\mathrm{Y}|\mathrm{X}))}{\mathrm{var}(\mathrm{Y})} = \frac{\mathrm{var}(\mathrm{Y}) - \mathrm{var}(\epsilon)}{\mathrm{var}(\mathrm{Y})} = 1 - \frac{\mathrm{var}(\epsilon)}{\mathrm{var}(\mathrm{Y})}.$$

En remplaçant les quantité $\mathrm{var}(\epsilon)$ et $\mathrm{var}(\mathrm{Y})$ par leurs estimateurs sans biais du modèle, on obtient le coefficient de détermination ajusté

$$\mathrm{R}^2_a(m) = 1 - \frac{\|y - \hat{y}^{(m)}\|^2/(n - |m|)}{\|y - \overline{y}\mathbf{1}\|^2/(n-1)} = \frac{(n-1)\mathrm{R}^2(m) - |m| + 1}{n - |m|}.$$

Le R^2 ajusté quantifie la part du modèle expliquée par les variables explicatives en tenant compte du nombre de variables utilisées, privilégiant les modèles contenant peu de variables. On choisit le modèle dont le R^2_a est le plus élevé. Ce critère est beaucoup plus judicieux que le R^2 classique, qui lui privilégiera toujours le modèle contenant toutes les variables.

5.4 Cp de Mallows

Dans une optique de prédiction, on peut considérer que le meilleur modèle m est celui qui minimise l'erreur de prédiction

$$r(m) := \mathbb{E}\|\mathrm{X}\beta - \Pi_m y\|^2.$$

L'erreur $r(m)$ est inconnue en pratique mais elle peut être estimée.

Proposition 5.1 *L'erreur quadratique vaut*

$$r(m) = \|(\mathrm{I} - \Pi_m)\mathrm{X}\beta\|^2 + \sigma^2|m|.$$

Elle est estimée sans biais par

$$\hat{r}(m) := \|y - \hat{y}^{(m)}\|^2 + (2|m| - n)\hat{\sigma}^2.$$

L'écriture du risque $r(m) = \|(\mathrm{I} - \Pi_m)\mathrm{X}\beta\|^2 + \sigma^2|m|$ est appelée *décomposition biais-variance*. Le carré du biais est la partie déterministe $\|(\mathrm{I} - \Pi_m)\mathrm{X}\beta\|^2$. C'est elle qui pose le plus de problème pour évaluer le meilleur modèle. La variance $\mathbb{E}\|\Pi_m\epsilon\|^2 = \sigma^2|m|$ ne pose pas de problème majeur pour le choix du modèle.

Le Cp de Mallows d'un modèle m est défini par

$$\mathrm{Cp}(m) = \frac{\|y - \hat{y}^{(m)}\|^2}{\hat{\sigma}^2} + 2|m| - n.$$

On sélectionne le modèle dont le Cp de Mallows est le plus faible. On voit bien par la proposition 5.1 que cela revient à chercher le modèle qui minimise l'estimateur sans biais du risque $\hat{r}(m)$.

5.5 Critère AIC

Le critère d'information d'Akaike (AIC) est construit à partir de la vraisemblance du modèle et nécessite donc de connaître la loi du bruit ϵ, que l'on ne suppose pas forcément normale ici (historiquement, la motivation derrière le critère d'Akaike est de minimiser la divergence de Kullback avec la vraie loi des observations, ce qui dans ce cadre est équivalent à maximiser la vraisemblance). Soit $f(\sigma^2, .)$ la densité de ϵ (qui dépend du paramètre inconnu σ^2), la vraisemblance associée au modèle de régression est

$$f(\sigma^2, \epsilon) = f(\sigma^2, y - \mathrm{X}\beta) := \mathrm{V}(\sigma^2, \beta, \mathrm{X}, y).$$

Lorsqu'on fait de la sélection de modèle, on ne cherche pas à exprimer la vraisemblance comme une fonction des paramètres mais plutôt comme une fonction du modèle m. Si seule la loi du bruit est connue, évaluer la vraisemblance d'un modèle m nécessite d'estimer les paramètres, ce qui entraîne un biais. Pour évaluer la log-vraisemblance, ce biais est asymptotiquement de l'ordre du nombre de paramètres à estimer, à savoir $|m|$. Le critère AIC, qui utilise une version asymptotiquement débiaisée de l'estimateur de la log-vraisemblance, est défini par

$$\mathrm{AIC}(m) = 2|m| - 2\log(\mathrm{V}(\hat{\sigma}^2_m, \hat{\beta}_m, \mathrm{X}, y)),$$

où $\hat{\sigma}^2_m$ et $\hat{\beta}_m$ sont les estimateurs du maximum de vraisemblance de σ^2 et β pour le modèle m. Le meilleur modèle est celui qui minimise le critère AIC. Dans le cas Gaussien, les critères AIC et Cp de Mallows sont équivalents.

Le critère AIC se justifie asymptotiquement mais pas pour des échantillons de petites tailles. Il existe une version corrigée du critère, plus adaptée aux petits échantillons,

$$\mathrm{AIC}_c(m) = \mathrm{AIC}(m) + \frac{2|m|(|m|+1)}{n - |m| - 1}.$$

5.6 Critère BIC

Le critère BIC (Bayesian Information Criterion), défini par

$$\mathrm{BIC}(m) = 2|m|\log(n) - 2\log(\mathrm{V}(\hat{\sigma}^2_m, \hat{\beta}_m, \mathrm{X}, y))$$

est une version modifiée du critère AIC motivée par l'utilisation d'un a priori sur le paramètre β. Schwarz, qui a introduit ce critère, a montré que l'influence de l'a priori était négligeable asymptotiquement ce qui justifie que le critère n'en dépend pas. Le facteur $\log(n)$ dans la pénalité a pour conséquence de favoriser, plus que les autres critères, les modèles avec moins de paramètres.

5.7 Critère PRESS de validation croisée

La validation croisée est un des moyens les plus efficaces de juger de la qualité d'un modèle. Le principe de la validation croisée est d'estimer les paramètres à partir d'un sous-échantillon des données et d'évaluer leurs performances de prédiction sur les données mises de côté. La version la plus simple est le critère PRESS (prediction error sum of square), pour lequel une seule observation est laissée de côté. Soit $\hat{y}^{(i)}_{m,i}$ la prédiction de y_i estimée dans le modèle m à partir des données sans y_i, le critère PRESS est défini par

$$\mathrm{PRESS}(m) = \sum_{i=1}^{n} (\hat{y}^{(i)}_{m,i} - y_i)^2.$$

On retient bien sûr le modèle avec le PRESS le plus faible. D'après le lemme 3.8, on montre que le critère PRESS est également donné par

$$\mathrm{PRESS}(m) = \sum_{i=1}^{n} \frac{\hat{\epsilon}^2_{m,i}}{(1 - h_{m,ii})^2},$$

où $\hat{\epsilon}_{m,i}$ est le résidu de la i-ème observation, estimé dans le modèle m et $h_{m,ii}$ l'entrée diagonale de la matrice de projection sur l'espace engendré par les variables $x_j \in m$. Ce critère produit en général de très bons résultats.

En pratique, on chosit un de ces critères pour retenir un modèle final. Si le nombre de variables explicatives dans le modèle complet est grand, le calcul du critère pour les 2^{p+1} sous-modèles peut vite devenir très coûteux. Dans ce cas, on peut se contenter de faire une recherche pas-à-pas du meilleur modèle en enlevant et ajoutant les variables les plus pertinentes une par une, ce qui permet de ne calculer le critère que pour un petit nombre de modèles. Cette approche est nettement moins coûteuse en temps de calcul mais ne garantit pas de sélectionner le meilleur modèle pour le critère choisi.

6 Méthodes robustes d'estimation

Dans le modèle de régression linéaire Gaussien $y = \mathrm{X}\beta + \epsilon$, on a vu que l'estimateur des moindres carrés est le meilleur estimateur sans biais de β. Il s'avère que l'on peut souvent améliorer l'estimation en recherchant un estimateur de β biaisé, pour lequel l'erreur quadratique est plus faible que celle des moindres carrés. Les méthodes de sélection de modèle peuvent conduire à un estimateur biaisé de β. Par exemple, si la variance de $\hat{\beta}_j$ est élevée, il peut être préférable d'estimer β_j par zéro, même si celui-ci est non nul. Cela entraîne un biais qui est compensé par une plus forte diminution de la variance. Dans cette section, on s'intéresse à d'autres méthodes d'estimation, plus robustes aux problèmes de multicolinéarité et de sur-ajustement.

Les méthodes décrites dans cette section ne nécessite pas que la matrice de régression X soit de plein rang. On relache donc cette hypothèse dorénavant. En particulier, le nombre d'observations peut être inférieur au nombre de variables explicatives.

6.1 Analyse en composantes principales

L'analyse en composantes principales (ACP) recherche les directions qui résument le mieux l'information des variables explicatives. On travaille avec les variables standardisées

$$w_j = \frac{x_j - \overline{x}_j \mathbf{1}}{\sqrt{\overline{x_j^2} - \overline{x}_j^2}} \ , \ j = 1, ..., p.$$

L'idée est de faire une régression sur des combinaisons linéaires bien choisies des variables w_j, de manière à optimiser l'information tout en réduisant la dimension du modèle. Une *composante* est une combinaison linéaire $c_\lambda = \sum_{j=1}^p \lambda_j w_j \in \mathbb{R}^n$ des variables standardisées telle que $\sum_{j=1}^p \lambda_j^2 = 1$. On note W la matrice construite à partir des variables standardisées

$$\mathrm{W} = \begin{pmatrix} w_{11} & \dots & w_{p1} \\ \vdots & \ddots & \vdots \\ w_{1n} & \dots & w_{pn} \end{pmatrix},$$

et $r = \mathrm{rang}(\mathrm{W}) = \dim(\mathrm{Im}(\mathrm{W}))$. Une composante est un vecteur $c_\lambda = \mathrm{W}\lambda$ avec $\|\lambda\|^2 = 1$.

Du point de vue de l'ACP, l'information d'une composante est donnée par sa norme. La première composante principale $c_1 = \mathrm{W}\lambda_1$ est la composante de norme maximale (définie au signe près),

$$\lambda_1 = \arg \max_{\substack{\lambda \in \mathbb{R}^p \\ \|\lambda\|=1}} \|\mathrm{W}\lambda\|^2 = \arg \max_{\substack{\lambda \in \mathbb{R}^p \\ \|\lambda\|=1}} \lambda^\top \mathrm{W}^\top \mathrm{W}\lambda.$$

C'est la direction qui est privilégiée par l'ACP. Si celle-ci n'est pas unique, on en choisit une arbitrairement parmi les maximiseurs. La composante obtenue $c_1 = \mathrm{W}\lambda_1 = \sum_{j=1}^p \lambda_{1j} w_j$ détermine la variable explicative privilégiée par l'ACP. La deuxième composante principale $c_2 = \mathrm{W}\lambda_2$ est la composante orthogonale à c_1 de norme maximale

$$\lambda_2 = \arg \max_{\substack{\lambda \in \mathbb{R}^p \\ \|\lambda\|=1 \\ \lambda_1^\top \lambda = 0}} \lambda^\top \mathrm{W}^\top \mathrm{W}\lambda.$$

On choisira de l'intégrer ou non au modèle, suivant l'information supplémentaire apportée. On construit ensuite la troisième composante principale orthogonale aux deux premières de la même façon, et ainsi de suite jusqu'à obtenir r composantes.

Théorème 6.1 *Les vecteurs $\lambda_1, ..., \lambda_r$ sont des vecteurs propres de la matrices $\mathrm{W}^\top\mathrm{W}$ associés aux valeurs propres non nulles $\gamma_1, ..., \gamma_r$ classées dans l'ordre croissant. De plus, les composantes principales $c_1, ..., c_r$ sont des vecteurs propres de la matrice $\mathrm{W}\mathrm{W}^\top$, qui ont pour valeurs propres $\gamma_1, ..., \gamma_r$. En particulier, les composantes principales forment une base orthogonale de* $\mathrm{Im}(\mathrm{W})$.

Du fait de la standardisation des variables explicatives, le modèle de l'ACP peut s'écrire

$$y - \overline{y}\mathbf{1} = \sum_{j=1}^{r} \alpha_j c_j + (\epsilon - \overline{\epsilon}\mathbf{1}).$$

La variable à expliquer est le vecteur recentré $y - \overline{y}\mathbf{1}$, les paramètres à estimer sont les coefficients α_j et les variables explicatives sont les composantes principales $c_1, ..., c_r$. On peut choisir de ne retenir dans le modèle que les composantes principales les plus pertinentes. Il y a deux avantages majeurs à utiliser les composantes principales comme variables explicatives. Premièrement, les composantes principales sont orthogonales, ce qui permet de juger de la pertinence de chacune des composantes sans tenir compte des autres (contrairement au cas général où la pertinence d'une variable doit être testé en présence des autres variables du modèle). Deuxièmement, les composantes principales sont classées par ordre d'importance, la sélection de variables se fait donc en choisissant un rang k à partir duquel les composantes $c_j, j > k$ sont jugées non-pertinentes (on peut éventuellement prendre $k = 0$). Cela réduit à un maximum de $p+1$ modèles à tester. On peut retenir par exemple le modèle avec le PRESS le plus faible.

Une fois le seuil k déterminé, les paramètres $\alpha_1, ..., \alpha_k$ sont estimés par les moindres carrés. Du fait de l'orthogonalité des composantes, l'estimateur des moindres carrés s'exprime très simplement. En effet, en notant $\mathrm{C}_{(k)} = (c_1, ..., c_k)$, on vérifie facilement que

$$\hat{\alpha}_{(k)} = (\mathrm{C}_{(k)}^\top \mathrm{C}_{(k)})^{-1} \mathrm{C}_{(k)}^\top (y - \overline{y}\mathbf{1}) = \begin{pmatrix} \frac{1}{\gamma_1}\langle c_1, y - \overline{y}\mathbf{1}\rangle \\ \vdots \\ \frac{1}{\gamma_k}\langle c_k, y - \overline{y}\mathbf{1}\rangle \end{pmatrix}.$$

On obtient la prédiction

$$\hat{y}_{(k)}^{\mathrm{ACP}} = \overline{y}\mathbf{1} + \sum_{j=1}^{k} \hat{\alpha}_j c_j = \overline{y}\mathbf{1} + \sum_{j=1}^{k} \frac{1}{\gamma_j}\langle c_j, y - \overline{y}\mathbf{1}\rangle c_j = \overline{y}\mathbf{1} + \sum_{j=1}^{k} \frac{1}{\gamma_j}\langle \hat{y}, c_j\rangle c_j.$$

Si $k = r$, la prédiction obtenue par l'ACP est celle des moindres carrés car les composantes c_j et la constante $\mathbf{1}$ engendrent entièrement $\mathrm{Im(X)}$. De plus, on voit dans la formule de $\hat{\alpha}_{(k)}$ que l'ajout ou le retrait d'une composante c_j dans le modèle retenu ne modifie pas la valeur des estimateurs $\hat{\alpha}_{j'}$ pour $j' \neq j$, ce qui s'explique simplement par l'orthogonalité des composantes.

La décomposition de $\hat{y}$ dans la base $c_1, ..., c_r$ permet de privilégier les directions de faibles variances. En effet, les variances des projections de $\hat{y}$ dans les directions $c_j/\|c_j\|$ sont classées par ordre croissant,

$$\mathrm{var}\left\langle \hat{y}, \frac{c_j}{\|c_j\|}\right\rangle = \frac{1}{\gamma_j^2}\,\mathrm{var}\left\langle \frac{c_j}{\|c_j\|}, y - \overline{y}\mathbf{1}\right\rangle = \frac{1}{\gamma_j^2}\frac{c_j^\top}{\|c_j\|}\,\mathrm{var}(y - \overline{y}\mathbf{1})\frac{c_j}{\|c_j\|} = (n-1)\frac{\sigma^2}{\gamma_j}.$$

Classer les composantes principales par valeurs propres γ_j décroissantes est donc un moyen de privilégier les directions de faibles variances pour la prédiction.

6.2 Moindres carrés partiels

Comme l'ACP, la régression par moindres carrés partiels, ou régression PLS (partial least squares) fait intervenir les variables explicatives standadisées w_j. Le principe de la régression PLS est similaire à celui de l'ACP. L'objectif est de chosir des directions à privilégier pour définir le modèle, de manière à maximiser l'information tout en minimisant la dimension du modèle.

Les composantes de la régression PLS sont choisies en maximisant la corrélation avec y. Précisément, on définit la première composante $c_1 = \mathrm{W}\lambda_1$ par

$$\lambda_1 = \arg\max_{\substack{\lambda \in \mathbb{R}^p \\ \|\lambda\|=1}} \left\langle y - \overline{y}\mathbf{1}, \mathrm{W}\lambda\right\rangle^2.$$

Contrairement à l'ACP où les composantes ne sont déterminées qu'en fonction de l'information des variables explicatives, la régression PLS tient compte également de la réponse pour construire les composantes. L'idée est de maximiser l'information apportée par les variables explicatives et leurs interactions avec y. La deuxième composante de la régression PLS est construite de la même façon, en imposant qu'elle soit orthogonale à la première. On définit donc $c_2 = \mathrm{W}\lambda_2$ avec

$$\lambda_2 = \arg \max_{\substack{\lambda \in \mathbb{R}^p \\ \|\lambda\|=1 \\ \lambda_1^\top \mathrm{W}^\top \mathrm{W}\lambda = 0}} \langle y - \overline{y}\mathbf{1}, \mathrm{W}\lambda \rangle^2.$$

Le procédé peut être itéré jusqu'à obtenir r composantes, qui sont orthogonales par construction. On choisit alors un seuil k à partir duquel les composantes de sont plus intégrées au modèle. On peut utiliser ici aussi le critère PRESS pour choisir le meilleur seuil k.

6.3 Régression Ridge

Lorsqu'il y a des corrélations entre les variables explicatives, la matrice $\mathrm{X}^\top\mathrm{X}$ a des valeurs propres proches de zéro et l'estimateur des moindres carrés $\hat{\beta}$ n'est pas satisfaisant, du fait d'une forte variance. Pour contrôler la variance de l'estimateur, un moyen simple est de pénaliser les grandes valeurs de $\hat{\beta}$. C'est le principe de la régression ridge, on définit l'estimateur par

$$\hat{\beta}_\kappa^{ridge} \in \arg\min_{b \in \mathbb{R}^{p+1}} \|y - \mathrm{X}b\|^2 + \kappa\|b\|^2,$$

où $\kappa \geqslant 0$ est un paramètre à déterminer.

Théorème 6.2 *Si $\kappa > 0$, l'estimateur $\hat{\beta}_\kappa^{ridge}$ est unique donné par*

$$\hat{\beta}_\kappa^{ridge} = (\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}\mathrm{X}^\top y.$$

Il est également l'unique solution du problème d'optimisation sous contrainte

$$\hat{\beta}_\kappa^{ridge} = \arg\min_{b \in \mathbb{R}^{p+1}} \|y - \mathrm{X}b\|^2 \quad \text{sous la contrainte} \quad \|b\|^2 \leqslant \tau,$$

où $\tau = \|(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}\mathrm{X}^\top y\|^2$.

L'estimateur Ridge est donc solution de deux problèmes d'optimisation duaux : le minimiseur du critère pénalisé ou du critère sous contrainte.

Contrairement à l'estimateur des moindres carrés, l'estimateur Ridge est bien défini même si la matrice des régresseurs n'est pas de plein rang. Par ailleurs, l'estimateur Ridge est biaisé, mais de variance plus faible que l'estimateur des moindres carrés lorsque celui-ci existe. Précisemment,

$$\begin{aligned}
\mathrm{biais}(\hat{\beta}_\kappa^{ridge}) &= \mathbb{E}(\hat{\beta}_\kappa^{ridge} - \beta) = \left[(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}\mathrm{X}^\top\mathrm{X} - \mathrm{I}\right]\beta = -\kappa(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}\beta \\
\mathrm{var}(\hat{\beta}_\kappa^{ridge}) &= (\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}\mathrm{X}^\top\,\mathrm{var}(y)\mathrm{X}(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1} = \sigma^2(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-2}\mathrm{X}^\top\mathrm{X}
\end{aligned}$$

On a utilisé pour la dernière ligne le fait que $(\mathrm{X}^\top\mathrm{X} + \kappa\,\mathrm{I})^{-1}$ et $\mathrm{X}^\top\mathrm{X}$ commutent car elles sont diagonalisables dans la même base. Si on s'intéresse à l'erreur quadratique, appliquer la trace à la décomposition biais-variance $\mathbb{E}[(\hat{\beta}_\kappa^{ridge} - \beta)(\hat{\beta}_\kappa^{ridge} - \beta)^\top] = \mathrm{var}(\hat{\beta}_\kappa^{ridge}) + \mathrm{biais}(\hat{\beta}_\kappa^{ridge})\mathrm{biais}(\hat{\beta}_\kappa^{ridge})^\top$ donne

$$\mathbb{E}\|\hat{\beta}_\kappa^{ridge} - \beta\|^2 = \mathrm{tr}\left[\mathrm{var}(\hat{\beta}_\kappa^{ridge})\right] + \|\mathrm{biais}(\hat{\beta}_\kappa^{ridge})\|^2.$$

Soit $\phi_0, ..., \phi_p$ une base orthonormée de vecteurs propres de $\mathrm{X}^\top\mathrm{X}$ et $\gamma_0 \geqslant ... \geqslant \gamma_p$ les valeurs propres associées (qui, on rappelle, sont positives), les termes précédents s'écrivent

$$\mathrm{tr}\left[\mathrm{var}(\hat{\beta}_\kappa^{ridge})\right] = \sigma^2 \sum_{j=0}^{p} \frac{\gamma_j}{(\gamma_j + \kappa)^2} \quad \text{et} \quad \|\mathrm{biais}(\hat{\beta}_\kappa^{ridge})\|^2 = \kappa^2 \sum_{j=0}^{p} \frac{\langle \beta, \phi_j \rangle^2}{(\gamma_j + \kappa)^2}.$$

On voit bien que la variance de l'estimateur Ridge une fonction décroissante de κ alors que le carré du biais est une fonction croissante. Le paramètre κ permet donc de faire un compromis entre biais et variance. En pratique, le choix du paramètre κ se fait généralement par validation croisée. On a dans les cas extrêmes :

- Si X est de plein rang, la limite en zéro, $\lim_{\kappa\to 0}\hat{\beta}_\kappa^{ridge}$, est égale à l'estimateur des moindres carrés $\hat{\beta}$. Si X n'est pas de plein rang, la limite $\lim_{\kappa\to 0}\hat{\beta}_\kappa^{ridge}$ existe et est égale à l'image de y par l'inverse généralisé (ou opérateur de Moore-Penrose) de X,

$$\lim_{\kappa\to 0}\hat{\beta}_\kappa^{ridge} = \mathrm{X}^\dagger y.$$

- En l'infini, on voit facilement que la limite $\lim_{\kappa\to+\infty}\hat{\beta}_\kappa^{ridge}$ est nulle.

6.4 Régression lasso

La régression lasso (least absolute shrinkage and selection operator) est basée sur la même idée que la régression ridge, en remplaçant la norme Euclidienne de la pénalité par la norme ℓ_1. L'estimateur lasso est donc défini par

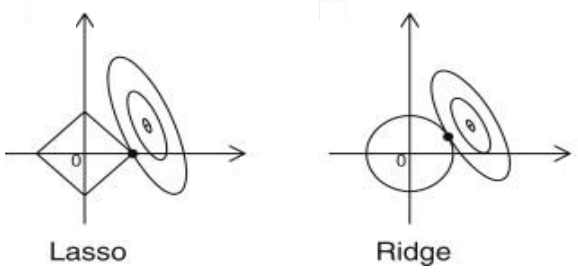
$$\hat{\beta}_\kappa^{lasso} = \arg\min_{b\in\mathbb{R}^{p+1}} \|y-\mathrm{X}b\|^2 + \kappa\sum_{j=0}^{p}|b_j|,$$

où $\kappa \geqslant 0$ est un paramètre à déterminer. Comme pour le Ridge, l'estimateur lasso est également solution du problème dual

$$\hat{\beta}_\kappa^{lasso} = \arg\min_{b\in\mathbb{R}^{p+1}} \|y-\mathrm{X}b\|^2 \quad \text{sous la contrainte} \quad \sum_{j=0}^{p}|b_j| \leqslant \tau,$$

pour un $\tau = \tau(\kappa) > 0$.

Le lasso est principalement utilisé pour construire un estimateur parcimonieux (dont certaines composantes sont nulles). Une interprétation graphique de ce phénomène est donnée dans le dessin suivant, qui compare les régressions Ridge et lasso.



Les ellipses représentent les courbes de niveaux de la fonction $b \mapsto \|y-\mathrm{X}b\|^2$. La solution au problème d'optimisation sous contraintes est le point d'intersection avec la boule pour chaque norme. On voit que l'estimateur lasso a une composante nulle, contrairement à l'estimateur Ridge.

Du fait de la sparsité de la solution, la méthode lasso peut être utilisée dans une optique de sélection de variables. En revanche, il n'existe pas de formule analytique pour l'estimateur lasso (contrairement au Ridge), le calcul de l'estimateur se fait numériquement par des algorithmes d'optimisation convexe.

7 Régression non-paramétrique