Chapitre 2

Regression Linaire

2.1 Régression Linéaire simple

2.1.1 Modèle

On note Y la variable aléatoire réelle à expliquer (ou encore de réponse, dépendante) et X la variable explicative (ou encore déterministe, de contrôle) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, l'estimation $\mathbb{E}(Y)$, est une fonction affine de X.

$$\mathbb{E}(Y) = f(X) = \beta_0 + \beta_1 X \,.$$

Pour une séquence d'observations aléatoires identiquement distribuées $\{(y_i, x_i)\,, i = 1, \ldots, n\}$, avec $n > 2$ et les x_i non tous égaux, le modèle s'écrit à partir des observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i \,, i = 1, \ldots, n$$

ou bien sous forme matricielle :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où le vecteur u contient les erreurs.

Les hypothèses relatives à ce modèle sont les suivantes :
— la distribution de l'erreur $\mathbf{u}$ est indépendante de X ou bien X est fixe.
— l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \ldots, n : E(u_i) = 0 \,, Var(u_i) = \sigma_u^2.$$

— β_0 et β_1 sont constants, il n'y a pas de rupture du modèle.
— Hypothèse complémentaire pour les inférences : $u \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Ce point important est développé dans l'appendice A

Remarque On a fait une hypothèse de linéarité ici mais en pratique cette hypothèse n'est pas toujours valide. Quand ce n'est pas le cas, il existe aussi des méthodes de régression non-paramétriques qui ne sont pas abordées dans le cours mais peuvent être très utiles. Il est aussi possible d'effectuer des transformations élémentaires sur les données, comme par exemple $y_i = \beta_0 + \beta_1 \ln x_i$ ou bien $y_i = \beta_0 + \beta_1 (x_i)^\alpha$.

2.1.2 Estimation

L'estimation des paramètres β_0, β_1, σ_u^2 peut être obtenue en minimisant la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $\{(y_i, x_i), i = 1, \ldots, n\}$, le critère des moindres carrés s'écrit :

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Pour minimiser ce critère, on pose :

$$\begin{aligned}
\bar{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\
\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} y_i \\
s_x^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
s_y^2 &= \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
s_{xy} &= \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\
r &= \frac{s_{xy}}{s_x s_y}
\end{aligned}$$

On peut alors montrer que les estimateurs de β_0 et β_1 au sens des moindres carrés sont :

$$\begin{aligned}
b_1 &= \frac{s_{xy}}{s_x^2}, \\
b_0 &= \bar{y} - b_1 \bar{x}.
\end{aligned}$$

On montre que ce sont des estimateurs sans biais et de variance minimum parmi les estimateurs fonctions linéaires des y_i. Cela signifie que pour $i \in \{0, 1\}$ alors $Biais(b_i) = \mathbb{E}(b_i) - \beta_i = 0$ et ainsi que $Var(b_i) = \mathbb{E}(b_i - \mathbb{E}(b_i)) = \mathbb{E}(b_i - \beta_i)$ est minimum. À chaque valeur x_i de X correspond la valeur estimée (ou prédite, ajustée) de Y :

$$\hat{y}_i = b_0 + b_1 x_i$$

les résidus calculés ou estimés sont :

$$e_i = y_i - \hat{y}_i$$

La variance σ_u^2 est enfin estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 \,.$$

2.1.3 Prédiction

Une fois les paramètres β_0 et β_1 estimés par b_0 et b_1, il est immédiat de prédire la valeur $\hat{y_0}$ qui a le plus de chance d'être associée à une observation x_0 avec :

$$\hat{y_0} = b_0 + b_1 x_0 \,.$$

Il est important de remarquer que le principe d'**estimation** des paramètres d'un modèle à partir de données d'apprentissage (les x_i et y_i) puis de **prédiction** de *scores/labels/variables de sortie* (ici y_0) à partir de nouvelles observations (ici x_0) est au coeur de l'apprentissage automatique.

2.1.4 Inférence

Niveau d'incertitude lié à l'estimation de b_0 et b_1

On rappel qu'une hypothese a été faite sur les résidus $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$ dans la sous-section 2.1.1 (où e est noté u). Les estimateurs b_0 et b_1 sont alors des variables aléatoires réelles. Ils ne font qu'approcher les valeurs β_0 et β_1 que l'on connaîtrait à coup sur si on disposait d'une infinité d'observations (ou si l'on contrôle le modèle). Ceci est intuitivement évident, si on compare les b_0 et b_1 obtenus sur disons 4 observations pour lesquelles e est faible avec ceux obtenus sur 3 observations avec e faible et une derniere où e est grand, ce qui peut arriver puisque $e \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$. Les valeurs de b_0 et b_1 seront différentes alors que le modèle est ses paramètres sont les mêmes.

Sous l'hypothèse de Gaussianité des résidus, on montre que

$$\frac{(n-2)s^2}{\sigma_u^2} \sim \chi^2_{(n-2)}$$

où la loi du χ^2 suit une densité de probabilité donnée appendice A. Alors, les statistiques

$$(b_0 - \beta_0) \Big/ s\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2}$$

et

$$(b_1 - \beta_1) \Big/ s\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}$$

suivent des lois de Student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres à partir de tests d'hypothèses. On va par exemple tester si le b_1 obtenu est significativement différent de 0, en fonction d'un coefficient α qui représente la probabilité avec laquelle on accepte de se tromper. Typiquement α correspond à 5% de chances de se tromper, ci

qui est raisonablement faible (voir le cours de Statistique pour aller plus loin). Notons, que si b_1 est significativement différent de 0, on peut considérer qu'il existe une relation de dépendance entre les x_i et les y_i.

Intervalles de confiance

Il est de même possible de construire des intervalles de confiance pour les valeurs de b_0 et b_1, toujours en fonction d'un niveau de confiance dépendant de α :

$$b_0 \pm s \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$
$$b_1 \pm s \left(\frac{1}{(n-1)s_x^2} \right)^{1/2} t_{n-2}(\alpha/2)$$

où $t_\nu(\alpha)$ est la distribution de Student à ν degrés de liberté (voir appendice A). En observant bien ces intervales de confiance ainsi que les distributions de Student, il est interessant de noter que plus on a d'observations n, plus les intervales de confiances sont resserés autour des b_0 et b_1 estimés. Plus on dispose d'information, moins le risque d'erreur est en effet grand par rapport aux valeurs réelles.

Attention : une inférence conjointe sur β_0 et β_1 ne peut être obtenue en considérant séparément les intervalles de confiance. La région de confiance est en effet une ellipse d'équation :

$$n(b_0-\beta_0)^2 + 2(b_0-\beta_0)(b_1-\beta_1)\sum_{i=1}^{n} x_i + (b_1-\beta_1)^2 \sum_{i=1}^{n} x_i^2 = 2s^2 \mathcal{F}_{\alpha;2,(n-2)}$$

où $\mathcal{F}_{\alpha;d_1,d_2}$ et la distribution de Fisher–Snedecor avec les paramètres d_1 et d_2 (voir appendice A).

Niveau d'incertitude lié à l'estimation d'un y_0 à partir d'un x_0

Enfin, connaissant une valeur x_0, on définit deux intervalles de confiance de prédiction à partir de la valeur prédite $\hat{y_0} = b_0 + b_1 x_0$. Le premier encadre E(Y) sachant $X = x_0$; le deuxième, encadre y_0 et est plus grand car il tient compte de la variance totale $\sigma_u^2 + Var(\hat{y_0})$:

$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2},$$
$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

2.1.5 Qualité d'ajustement

On rappelle que la variance σ_u^2 est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - b_0 + b_1 x_i)^2 .$$

et que :

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\right)^2$$
$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)\right)^2$$
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)\right)\left(y_i - \left(\frac{1}{n}\sum_{i=1}^{n}y_i\right)\right)$$

Dans l'optique de mesurer la qualité d'ajustement du modèle, il est d'usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous :

— Sum of Squares Total : $SST = (n-1)s_y^2$
— Sum of Squares Regression : $SSR = (n-1)\frac{s_{xy}^2}{s_x^2}$
— Sum of Squares Errors : $SST = (n-1)s^2$

On appelle alors *coefficient de détermination* la quantité :

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{s^2}{s_y^2} = \frac{SSR}{SST}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale. En pratique, si R^2 vaut par exemple 0.79, cela signifie que 79% de la variablilité de Y a été capturée par le modèle linéaire et que seulement 21% restent à expliquer.

2.1.6 Détection d'outliers

Le critère des moindres carrés est très sensible à des observations atypiques hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant. Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les observations influentes c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer une valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (en norme L_1), ne rien faire... , cela dépend du contexte et doit être négocié avec le commanditaire de l'étude.

Effet levier

Une première indication est donnée par l'éloignement de x_i par rapport à la moyenne $\bar{x}$. En effet, écrivons les prédicteurs y_i comme combinaisons linéaires des observations :

$$\hat{y}_i = b_0 + b_1 x_i = \sum_{j=1}^{n} h_{ij} y_j$$

avec

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

en notant $\mathbf{H}$ la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement :

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de $\hat{y_i}$.

Résidus

Différents types de résidus sont définis afin d'affiner leurs propriétés :
— Résidus : $e_i = y_i - \widehat{y_i}$
— Résidus$_i$: $e_{(i)i} = y_i - \widehat{y_{(i)i}} = \frac{e_i}{1-h_{ii}}$
où $\widehat{y_{(i)i}}$ est la prévision de y_i calculée sans la $i$ème observation (x_i, y_i).

Afin de supprimer l'influence de la variance dans les résidus, on remarque d'abord que $Var(e_i) = \sigma_u^2(1 - h_{ii})$. En supposant que $E(e_i) = 0$, les résidus peuvent alors être standardisés de deux manières. Les *résidus standardisés* r_i sont calculés avec :

$$r_i = \frac{e_i}{s\sqrt{1 - h_{ii}}}\,.$$

La standardisation ci-dessus dépend cependant de e_i dans le calcul de s (qui estime $Var(e_i)$). Une estimation non biaisée de cette variance est basée sur

$$s_{(i)}^2 = \left((n-1)s^2 - \frac{e_i^2}{1 - h_{ii}}\right)/(n-3)$$

qui ne tient pas compte de la $i$ème observation. On définit alors les *résidus studentisés* par :

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_{ii}}}$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $(n-3)$ degrés de liberté.

Il est ainsi possible de construire un test d'hypothèse pour tester la présence d'observations atypique. Plusieurs observations peuvent de même être simultanément considérées en utilisant l'inégalité de Bonferroni. En pratique, les résidus studentisés sont sonvent comparés aux bornes ± 2. Si un résidu studentisé n'est pas dans cet intervale de valeurs, il est considéré comme atypique.

Diagnostics

Un dernier indicateur couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_{(i)j}} - \widehat{y_j})^2}{2s^2} = \frac{h_{ii}}{2(1 - h_{ii})} r_i^2\,, \forall i$$

qui mesure l'influence de chaque observation i sur l'ensemble des prévisions en prenant en compte effet levier et importance des résidus.

2.2 Régression Linéaire Multiple

Les modèles classiques de régession (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage automatique. Néanmoins, ils présentent un grand intérêt compte tenu de leur robustesse, de leur stabilité face à des fluctuations d'échantillons et de leur capacité à passer à l'échelle pour des données massives. Ils restent ainsi toujours très utilisés en production notamment lorsque la fonction à modéliser est bien linéaire et qu'il serait contre productif de chercher plus compliqué.

2.2.1 Modèle

Une variable quantitative $\mathbf{Y}$ dite à expliquer (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $\mathbf{X^1}, \ldots, \mathbf{X^p}$ dites explicatives (ou encore de contrôle, endogènes, indépendantes, régresseurs, prédicteurs).

Les données sont supposées provenir d'un échantillon statistique de n observations, chacune étant dans $\mathbb{R}^{(p+1)}$ (avec $n > p+1$) :

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i)\,, i = 1, \ldots, n$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de $\mathbf{Y}$ appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, \mathbf{X^1}, \ldots, \mathbf{X^p}\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de 1s. C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \ldots + \beta_p X_i^p + \varepsilon_i\,, i = 1, 2, \ldots, n$$

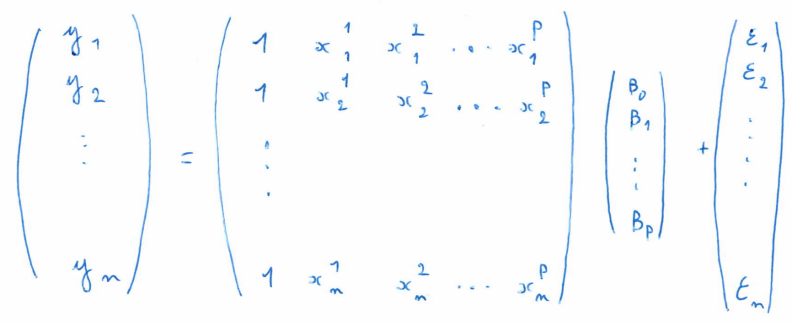
avec les hypothèses suivantes :

- Les ε_i sont des termes d'erreur indépendants et identiquement distribués, *i.e.* $E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma^2\mathbf{I}$.
- Les termes de $\mathbf{X^j}$, *i.e.* du vecteur qui contient les observations de la j^{eme} variable, sont supposés déterministes (facteurs contrôlés). Dans certain contextes, on suppose alternativement que l'erreur ε est indépendante de la distribution conjointe de $\mathbf{X^1}, ..., \mathbf{X^p}$. On écrit dans ce cas que $E(\mathbf{Y}|\mathbf{X^1}, \ldots, \mathbf{X^p}) = \beta_0 + \beta_1\mathbf{X^1} + \beta_2\mathbf{X^2} + \ldots + \beta_p\mathbf{X^p}$ et que $Var(Y|\mathbf{X^1}, \ldots, \mathbf{X^p}) = \sigma^2$.
- Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.
- En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur ε (*i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}$ de taille $(n \times (p+1))$ de terme général X_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ (c'est à dire $X_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général Y_i . En notant les vecteurs $\varepsilon = [\varepsilon_1 \ldots \varepsilon_n]'$ et $\beta = [\beta_0 \beta_1 \ldots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon\,.$$

Ce modèle est détaillé ci-dessous :



2.2.2 Estimation

Conditionnellement à la connaissance des valeurs des $\mathbf{X}^j$, les paramètres inconnus du modèle, le vecteur β et le paramètre de nuisance σ^2, sont estimés par minimisation des carrés des écarts (M.C.) ou encore par maximisation de la vraisemblance (M.V.) en considèrant l'hypothèse de la normalité de la variable d'erreur. Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Etudions l'estimation par moindres carrés. L'expression à minimiser sur $\beta \in \mathbb{R}^{p+1}$ s'écrit :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^1 - \ldots - \beta_p X_i^p)^2 = ||\mathbf{Y} - \mathbf{X}\beta\,||^2 \tag{2.1}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \tag{2.2}$$

Par dérivation matricielle de la dernière équation on obtient les équations normales :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$$

dont la solution correspond à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. Si cette hypothèse n'est pas vérifiée, il suffit en principe de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Une approche de réduction de dimension (régression ridge, Lasso, PLS...) est en pratique à mettre en oeuvre (voir Section 3.3). Alors, l'estimation des paramètres β_j est donnée par :

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{Y}$ ont pour expression :

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est connue sous le nom *hat matrix*. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace $Vect(\mathbf{X})$

engendré par les vecteurs colonnes de $\mathbf{X}$. On note alors :

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

le vecteur des résidus.

Notons finalement qu'il est possible d'inférer sur l'estimation des paramètres β_j comme dans la régression linéaire simple mais nous nous intéresserons dans ce cours à d'autres aspects de la régression multiple, notamment la sélection de modèle.

2.2.3 Prévision

Connaissant les valeurs des variables $\mathbf{X}^j$ pour une nouvelle observation : $x_0 = [x_0^1, x_0^2, \ldots, x_0^p]'$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y_0}$ de $\mathbf{Y}$ ou $E(\mathbf{Y})$ est donnée par :

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0^1 + \cdots + \widehat{\beta_p} x_0^p \,.$$

Les intervalles de confiance des prévisions de $\mathbf{Y}$ et $E(\mathbf{Y})$, pour une valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned} \widehat{y_0} &\pm t_{\alpha/2;(n-p-1)} \widehat{\sigma} (1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}, \\ \widehat{y_0} &\pm t_{\alpha/2;(n-p-1)} \widehat{\sigma} (\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}. \end{aligned}$$

Il est intéressant de remarquer que ces intervalles de confiance dépendent d'une loi de Student à $n-p-1$ degrés de liberté (voir appendice A). A dimension des observations fixée p, plus n est grand, plus les valeurs de la loi de Student seront faible, et ainsi les marges seront réduites. Cependant, plus p est proche de n, plus les marges sont élevées. En parallèle, les variances de ces prévisions, comme celles des estimations des paramètres, dépendent aussi directement du conditionnement de la matrice $\mathbf{X}'\mathbf{X}$ de par le terme $\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0$.

2.2.4 Qualité d'ajustement

Tout comme dans le modèle linéaire simple (Sous-section 2.1.5), la qualité d'ajustement du modèle peut être mesurée avec $p > 1$ variables par coefficient de détermination R^2. On note SSE la somme des carrés des résidus (sum of squared errors) :

$$SSE = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = ||e||^2 \,.$$

On définit également la somme totale des carrés (total sum of squares) par

$$SST = ||\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}||^2 = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2 \,.$$

et la somme des carrés de la régression (regression sum of squares) par

$$SSR = ||\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}||^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n\bar{\mathbf{Y}}^2 = \mathbf{Y}'\mathbf{H}\mathbf{Y} - n\bar{\mathbf{Y}}^2 = \widehat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2 \,.$$

où $\bar{\mathbf{Y}}\mathbf{1}$ est le vecteur de même taille que $\mathbf{Y}$ dont tous les termes sont égaux à la moyenne des valeurs observées de $\mathbf{Y}$. Le *coefficient de détermination* est alors le rapport

$$R^2 = \frac{SSR}{SST} = \frac{||\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}||^2}{||\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}||^2}$$

qui est donc la part de variation de $\mathbf{Y}$ expliquée par le modèle de régression. La quantité R est appelée coefficient de corrélation multiple entre $\mathbf{Y}$ et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{Y}$ et sa prévision $\widehat{\mathbf{Y}}$.

Notons que le coefficient de détermination croît avec le nombre p de variables par construction. D'une manière générale, plus un modèle est complexe plus il va pouvoir coller aux données, mais moins il sera explicable et sera généralisable.