ENSM.SE

Axe Méthodes Statistiques et Actuariat

INTRODUCTION A LA REGRESSION

Laurent Carraro

Novembre 05

Table des matières

Présentation générale de la régression

Les pages qui suivent visent à préciser le cadre d'étude et à donner des références (livres ou cours de l'axe MSA) pour les notions qui ne seront pas étudiées ici.

PROBLEMATIQUE – CLASSIFICATION

Le but général des techniques de régression est de décrire les relations entre plusieurs variables dans un but **prédictif**, ceci à partir d'observations de ces variables.

Dans le cadre de ce cours, nous considérerons seulement le cas où nous cherchons à prévoir, ou expliquer, une variable y – appelée **variable expliquée** ou **réponse** – à l'aide d'autres variables $x_1, \ldots, x_p$ – dites **variables explicatives** ou **prédicteurs**[1].

Pour ce faire, outre d'éventuelles informations supplémentaires, on se basera toujours sur un jeu de données qui consiste en n réalisations du vecteur $(y, x_1, \ldots, x_p)$.

On se limitera également au cas où la variable y est **quantitative et continue**. Lorsque la variable y est discrète ou qualitative, on est mené bien souvent à deux types d'approches. La première, l'analyse discriminante fait partie de l'arsenal des techniques de l'analyse des données. La seconde mène à l'utilisation des modèles linéaires généralisés, dont la régression logistique est un exemple courant. Cette dernière concerne en effet le cas d'une réponse y binaire où l'on cherche en fait à prévoir la probabilité d'apparition d'un phénomène à partir de l'observations de prédicteurs quantitatifs et de l'apparition, ou non, du phénomène étudié (voir [McCullagh]).

Dans ce cadre dans lequel nous sommes , on peut chercher une relation du type :

$$y \approx f(x_1, \ldots, x_p)$$

Si la forme de la fonction f recherchée n'est pas spécifiée et est seulement déterminée à l'aide des données recueillies, on parle de **régression non paramétrique**. Sur ce sujet, on pourra consulter une des "bibles" du domaine [Hastie].

Si f est par contre de la forme $f(\theta; x_1, \ldots, x_p)$, avec f connue et θ inconnu, on parle de **régression paramétrique**[2]. Si la fonction f est linéaire **en θ**, la régression est dite linéaire et dans le cas contraire, on a affaire à un problème de régression non linéaire. Une bonne introduction à la régression non linéaire se trouve dans [Draper]. Pour un exposé plus complet, dans un esprit essentiellement applicatif, on recommande [Bates]. Les amateurs de théorie quand à eux trouveront leur bonheur dans [Antoniadis].

Enfin, une distinction très importante doit être faite à propos du mécanisme d'obtention, ou d'acquisition, des données.

Si les niveaux des prédicteurs ont été fixés par l'expérimentateur (le statisticien, l'ingénieur,…) chargé du recueil des données, on parle de **prédicteurs contrôlés** et de données recueillies selon un **plan d'expérience**. Dans ce cas, l'observation d'un effet des

[1] La littérature anglo-saxonne parle souvent de variables dépendante (réponse) et indépendantes (prédicteurs), ce que nous ne ferons pas du fait de la confusion possible avec l'indépendance des variables entre elles.

[2] Notons que dans ce cas, la fonction f est souvent obtenue par une modélisation physique, mécanique, chimique, biologique, économique, financière…

prédicteurs sur la réponse implique[3] une relation de cause à effet entre les prédicteurs et la réponse (et le vocabulaire ci-dessus prend alors tout son sens).

Si par contre, les prédicteurs sont observés en même temps que la réponse, on parle de **prédicteurs non contrôlés**. Dans ce cas, aucune relation de cause à effet ne peut être démontrée par l'expérience car des variables non observées peuvent influer à la fois les prédicteurs et la réponse.

D'un point de vue plus formalisé, dans le cas d'un plan d'expérience, les facteurs sont des variables déterministes, alors que dans le cas contraire, ils peuvent être considérés comme les réalisations de variables aléatoires.

LIEN AVEC DIVERS ENSEIGNEMENTS (TRONC COMMUN ET AXE MSA)

Analyse des données

Dans sa version la plus simple, c'est à dire l'**analyse en composantes principales** (ACP), l'analyse des données vise à la description d'un ensemble de variables, ainsi qu'aux relations pouvant exister entre ces variables. Vis-à-vis de la régression, deux différences majeures – liées entre elles - peuvent être mises en évidence :

- L'ACP ne cherche pas à **prévoir**, mais plutôt à **décrire**.
- L'ACP met toutes les variables considérées sur le même plan alors que la régression fait jouer un rôle particulier à l'une d'entre elles : la réponse.

L'**analyse discriminante** vise par contre à **expliquer** les variations d'une variable qualitative à l'aide de facteurs. Elle s'apparente donc à la régression : il s'agit de déterminer les combinaisons de facteurs qui expliquent le mieux les diverses modalités de la réponse. De plus, il existe une version **décisionnelle** de l'analyse discriminante qui permet de **prévoir** pour tout nouvel individu, à la vue seule des facteurs, la réponse y.

La plupart du temps, cette analyse s'apparente en fait à une régression non paramétrique d'une réponse quantitative y sur des prédicteurs x_1, …, x_p à partir d'une discrétisation de y. Cette technique est d'ailleurs utilisée en régression non paramétrique dans le cas d'un nombre important de prédicteurs ; il s'agit de la méthode SIR (Sliced Inverse Regression) [Chen].

Compléments de probabilités et processus aléatoires

Les cours de compléments de probabilités et de processus aléatoires [Bay] se concentrent sur le délicat problème de la définition du conditionnement en probabilités. Une des notions centrales dans ce cadre est celle d'espérance conditionnelle. Il s'agit de définir, pour des variables aléatoires Y et X_1, …, X_p la quantité :

$$E(Y/ X_1, \ldots, X_p)$$

qui vise à fournir la meilleure prévision pour Y lorsque l'on connaît les variables X_1, …, X_p. C'est précisément le problème de la régression lorsque les prédicteurs sont des variables aléatoires!

En d'autres termes, le problème général de la régression pour des prédicteurs non contrôlés coïncide avec l'évaluation expérimentale – à partir de données numériques – de l'espérance conditionnelle.

[3] A condition que l'expérience ait été conduite correctement ; par exemple par des techniques de randomisation.

Séries chronologiques

Pour prévoir le comportement futur d'une série chronologique - ou temporelle - $(y_t)_{1 \leq t \leq n}$, par exemple le cours du pétrole (!), de nombreuses techniques existent. Parmi celles-ci, il est courant d'essayer d'écrire une relation du genre :

$$y_t \approx f(\theta;t)$$

On voit par là qu'il s'agit encore d'un problème de régression, un peu particulier dans la mesure où le prédicteur est ici le temps.

Malheureusement (ou heureusement suivant les points de vue!), cette spécificité va compliquer la tâche. On verra en effet plus bas que l'une des hypothèses essentielles en régression est que les écarts entre les réponses observées et les réponses prédites, par la formule $f(x_1, \ldots, x_p)$, peuvent être considérés comme les réalisations de variables aléatoires indépendantes. Or, dans le cas d'une série temporelle, les écarts observés à des instants successifs sont en général corrélés. Il faut donc estimer cette corrélation et corriger les procédures d'estimation et d'inférence en conséquence.

Par ailleurs, il est également courant de chercher à prévoir plusieurs séries en même temps (du genre taux de chômage, inflation, PIB par exemple) à l'aide de leur passé. De même, on peut – dans une optique proche de la régression – chercher à expliquer une variable temporelle, par exemple le taux de chômage, à partir de variables considérées comme explicatives, par exemple le PIB, le coût de la main d'œuvre, etc… Du fait du caractère temporel des données recueillies, les écarts aux prévisions sont là encore corrélés dans le temps et il faut en tenir compte : c'est un un problème type d'**économétrie**.

Probabilités et Statistiques

Last but not least, le lecteur aura compris que vu le contexte - estimation et prévision dans un contexte incertain à partir de données numériques – les outils de l'inférence statistique vont être au cœur de toutes les techniques que nous allons rencontrer. L'estimation par maximum de vraisemblance, les domaines de confiance, les tests… vont donc être notre vocabulaire de base tout au long de ce cours.

Régression linéaire et non linéaire : aspect formel

On va ici définir, et pour partie rappeler, la version "probabiliste" du problème général de la régression **dans le cas où les prédicteurs sont aléatoires,** donc lorsque les prédicteurs sont non contrôlés.

Nous sommes donc dans le cas où l'on observe des réalisations de variables aléatoires : la réponse $\mathbf{Y}$ et les prédicteurs $\mathbf{X_1}$, …, $\mathbf{X_p}$.

La théorie de l'espérance conditionnelle s'interprète facilement à l'aide de l'espace $L^2(P)$ des v.a. $\mathbf{U}$ de carré intégrable, c'est à dire telles que $E(\mathbf{U^2}) < +\infty$. Cet espace est muni d'un produit scalaire :

$$<\mathbf{U}, \mathbf{V}> = E(\mathbf{UV})$$

qui en fait un espace euclidien[4].

Dans ce cadre, si $L^2(\mathcal{G})$ désigne l'espace[5] des v.a. de carré intégrable, de la forme $\mathbf{f(X_1, …, X_p)}$, **l'espérance conditionnelle** $E(\mathbf{Y}/ \mathbf{X_1}, …, \mathbf{X_p})$ s'interprète comme la projection orthogonale de $\mathbf{Y}$ sur $L^2(\mathcal{G})$.

Le problème concret auquel on est alors confronté est d'estimer cette quantité abstraite à partir de données. Si l'on se souvient du rapport entre loi conditionnelle et espérance conditionnelle[6], on s'aperçoit que l'estimation de l'espérance conditionnelle peut nécessiter l'estimation d'une loi de probabilité sur $\mathbb{R}^{p+1}$. Bien que ce ne soit pas en général la méthode utilisée[7], on devine que dès que la dimension p augmente, une estimation raisonnable de l'espérance conditionnelle ne sera possible que si l'on possède de grandes quantités de données, ce qui est malheureusement très rare.

Pour contourner ce problème, on est amené à être moins gourmand et à rechercher l'espérance conditionnelle parmi une classe plus restreinte de fonctions. Pour ce qui nous concerne, nous utiliserons les fonctions linéaires, ou plutôt affines.

💣 On désignera désormais par $\mathbf{X_0}$ la variable aléatoire (!) constante égale à 1 :

Soient $\mathbf{X_1}$, ..., $\mathbf{X_p}$, $\mathbf{Y} \in L^2(P)$. La **régression linéaire** de $\mathbf{Y}$ sur $\mathbf{X_1}$, .., $\mathbf{X_p}$ est la projection orthogonale de $\mathbf{Y}$ sur $ev\{\mathbf{X_0}, .., \mathbf{X_p}\}$. On la note $E_L(\mathbf{Y}/ \mathbf{X_1}, ..., \mathbf{X_p})$.

Notons qu'il arrive en pratique que la variable $\mathbf{X_0}$ soit omise dans l'espace qur lequel on projette : on parle alors de **régression passant par l'origine (***regression through the origin*) mais aucune notation particulière ne sera utilisée par la suite pour désigner cette quantité.

[4] Il s'agit en fait d'un espace de Hilbert, après avoir quotienté par les v.a. nulles presque sûrement...
[5] On note qu'il s'agit d'un espace vectoriel fermé.
[6] Voir l'excellent poly de mon collègue Xavier : [Bay]!
[7] Ces problèmes concernent en fait la régression non paramétrique : voir [Hastie].

L'intérêt évident est que l'on est passé ainsi d'un problème de type non paramétrique à un problème paramétrique puisque seuls p+1 coefficients doivent être connus afin de spécifier la régression linéaire. De plus, la détermination de ces coefficients se ramène à la résolution d'un système linéaire comme l'indique le résultat qui suit :

Proposition (équation de régression)

Si $E_L(\mathbf{Y}/ \mathbf{X_1}, \ldots, \mathbf{X_p}) = \sum_{i=0}^{p} a_i \mathbf{X_i}$, on a :

$$\text{(ER)} \qquad E\left[\begin{pmatrix}\mathbf{X_0}\\ .\\ .\\ .\\ \mathbf{X_p}\end{pmatrix}(\mathbf{X_0} \ldots \mathbf{X_d})\right]\begin{pmatrix}a_0\\ .\\ .\\ .\\ a_p\end{pmatrix} = \begin{pmatrix}E(\mathbf{X_0 Y})\\ .\\ .\\ .\\ E(\mathbf{X_p Y})\end{pmatrix},$$

ou de façon équivalente :

$$\text{(ER')} \qquad \Gamma_X\begin{pmatrix}a_1\\ .\\ .\\ .\\ a_p\end{pmatrix} = \begin{pmatrix}\mathrm{cov}(\mathbf{X_1},\mathbf{Y})\\ .\\ .\\ .\\ \mathrm{cov}(\mathbf{X_p},\mathbf{Y})\end{pmatrix} \text{ et } a_0 + \sum_{i=1}^{p} a_i E(\mathbf{X_i}) = E(\mathbf{Y}),$$

où Γ_X est la matrice de covariance du vecteur $\mathbf{X} = (\mathbf{X_1} \ldots \mathbf{X_p})'$.

Preuve. Elle consiste simplement à écrire que $\mathbf{Y} - E_L(\mathbf{Y}/ \mathbf{X_1}, \ldots, \mathbf{X_p})$ est orthogonal aux variables $\mathbf{X_0}, .., \mathbf{X_p}$. Pour obtenir la deuxième formulation, il suffit de soustraire à la i-ème ligne du système $E(\mathbf{X_i})$ fois la première ligne.

Remarques dans le cas de la régression linéaire de Y sur une seule variable X.

- D'un point de vue pratique, on voit que la régression linéaire de **Y** sur **X** peut être réalisée à condition de connaître les quantités $E(\mathbf{X})$, $E(\mathbf{X^2})$, $E(\mathbf{Y})$, $E(\mathbf{XY})$; soit 4 nombres à estimer à partir des observations. D'une façon plus concise et pour simplifier, on peut dire que l'on peut effectuer la régression de **X** sur **Y** dès que l'on connaît espérance et matrice de covariance du vecteur (**X**, **Y**) - ce qui n'est d'ailleurs pas tout à fait exact car la variance de **Y** n'est pas a priori nécessaire. On ne conserve donc de la structure probabiliste (la loi) du vecteur (**X**, **Y**) que peu de renseignements.
- On voit facilement à partir de l'équation (ER') que **X** et **Y** sont non corrélées si et seulement si $E_L(\mathbf{Y}/ \mathbf{X}) = E(\mathbf{Y})$. Dans ce cas, notre estimation linéaire de **Y** n'est pas modifiée par l'observation de **X** ; en d'autres termes, **X** n'apporte aucune information **linéaire** sur **Y**. Il ne faut pas croire cependant que **X** n'apporte aucune information sur **Y** ;

on peut ainsi vérifier que si $\mathbf{X}$ est de loi N(0,1) et si $\mathbf{Y} = \mathbf{X}^2$, on a $E_L(\mathbf{Y}/\ \mathbf{X}) = E(\mathbf{Y})$ bien que $\mathbf{Y}$ soit très (!) liée à $\mathbf{X}$.

On a vu combien l'approche du type régression linéaire simplifiait un peu outrageusement le problème de l'évaluation de l'espérance conditionnelle. Par conséquent, il ne faut pas s'attendre en général à ce que $E_L(\mathbf{Y}/\ \mathbf{X})$ soit proche de $E(\mathbf{Y}/\ \mathbf{X})$. Cependant, dans deux cas particuliers, le calcul de l'espérance conditionnelle se ramène à l'évaluation d'une régression linéaire.

Cas des vecteurs gaussiens

Si le vecteur aléatoire $(\mathbf{X_1}, \ldots, \mathbf{X_p}, \mathbf{Y})$ est gaussien[8], on a :
$$E_L(\mathbf{Y}/\ \mathbf{X_1}, \ldots, \mathbf{X_p}) = E(\mathbf{Y}/\ \mathbf{X_1}, \ldots, \mathbf{X_p})$$

Nous renvoyons à nouveau à l'excellent cours [Bay] pour une preuve de ce résultat.

Cas de prédicteurs discrets

Afin de simplifier les écritures, nous considérerons ici uniquement le cas d'un seul prédicteur.

Si la v.a. $\mathbf{X}$ est discrète, à valeurs dans $\{x_1, \ldots, x_k\}$, on a :
$$E(\mathbf{Y}/\ \mathbf{X}) = E_L(\mathbf{Y}/\ 1_{\{\mathbf{X}=x_1\}}, \ldots, 1_{\{\mathbf{X}=x_k\}})$$

Tout vient du fait que dans ce cas, l'ensemble des fonctions de $\mathbf{X}$ coïncide avec l'espace vectoriel engendré par $1_{\{\mathbf{X}=x_1\}}, \ldots, 1_{\{\mathbf{X}=x_k\}}$. Par conséquent, les projections orthogonales sont identiques.

A noter que dans ce contexte, on effectue une régression sur l'espace vectoriel engendré par les variables $1_{\{\mathbf{X}=x_1\}}, \ldots, 1_{\{\mathbf{X}=x_k\}}$ sans ajouter la variable constante $\mathbf{X_0}$ car elle appartient déjà à cet espace comme somme des indicatrices précédentes !

Cette partie purement théorique étant achevée, passons aux données!

[8] Voir le cours de probabilités et statistiques de première année pour les rares qui auraient oublié…

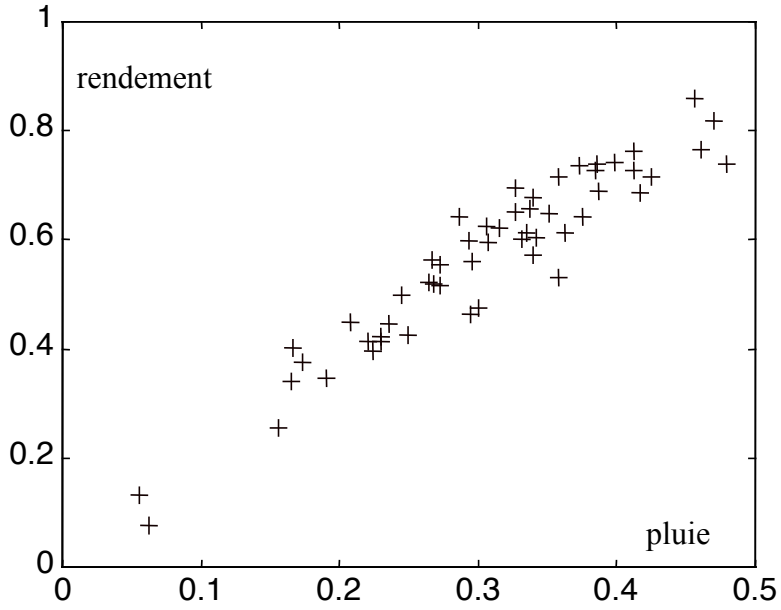
Régression linéaire – aspect empirique

EXEMPLE 1

On désire estimer l'effet sur les rendements de blé des précipitations printanières observées dans une région et on observe pour cela ces deux variables sur plusieurs parcelles et plusieurs années. On obtient ainsi :

pluie	rendement	pluie	rendement
0.0556	0.1316	0.3148	0.6208
0.0623	0.0753	0.3264	0.6511
0.1554	0.2547	0.3268	0.6957
0.1651	0.3401	0.3318	0.6000
0.1659	0.4015	0.3352	0.6144
0.1730	0.3765	0.3375	0.6557
0.1902	0.3459	0.3397	0.6772
0.2077	0.4477	0.3400	0.5731
0.2201	0.4137	0.3415	0.6037
0.2235	0.3959	0.3513	0.6483
0.2299	0.4215	0.3577	0.5318
0.2303	0.4139	0.3582	0.7146
0.2361	0.4444	0.3627	0.6121
0.2443	0.4975	0.3729	0.7362
0.2489	0.4243	0.3756	0.6432
0.2640	0.5214	0.3848	0.7275
0.2663	0.5645	0.3862	0.7388
0.2677	0.5177	0.3872	0.6904
0.2726	0.5538	0.3985	0.7433
0.2729	0.5164	0.4123	0.7279
0.2864	0.6430	0.4125	0.7623
0.2930	0.5992	0.4165	0.6851
0.2944	0.4646	0.4246	0.7142
0.2955	0.5590	0.4558	0.8607
0.2998	0.4739	0.4607	0.7654
0.3059	0.6258	0.4696	0.8190
0.3075	0.5967	0.4797	0.7395

N. B. : La pluie est en m et le rendement en fraction d'un maximum observé (donc sans unité).



⟶ points à peu près alignés.

Objectifs.

- Estimer l'équation de la droite.
- Vérifier - ou infirmer - le caractère prédictif de la variable pluie.
- Tester la validité du modèle.
- Obtenir des intervalles de confiance pour les paramètres de la droite.
- Détecter des observations aberrantes.
- Prévoir le rendement, avec incertitude associée, pour une quantité de pluie donnée.

LES MOINDRES CARRES

Expérimentalement, on observe n réalisations de la réponse y et des prédicteurs[9] x_1, …, x_p (que l'on notera y_i et $(x_j)_i$ pour $i \in \{1,..., n\}$) : ces résultats constituent les n **observations**. On cherche alors à estimer les paramètres β_1, ..., β_p qui donnent une relation approchée de la forme :

$$y \approx \sum_{j=1}^{p} x_j \beta_j$$

Pour réduire le volume des écritures, nous utiliserons systématiquement des notations matricielles :

[9] On ajoutera la plupart du temps, comme pour la régression, un prédicteur x_0 : la constante égale à 1.

Notations
$Y = [y_1 \dots y_n]'$ vecteur $n \times 1$
$X = [(x_j)_i]$ matrice $n \times p$
$\beta = [\beta_1 \dots \beta_p]'$ vecteur $p \times 1$

N.B. : lorsque les prédicteurs sont contrôlés, la matrice X prend le nom de **matrice de plan d'expérience**, puisqu'elle contient les diverses valeurs choisies pour les prédicteurs.

Pour estimer β, on cherche la valeur $\hat{\beta}$ qui minimise la somme :

$$d(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_j (x_j)_i \beta_j \right)^2 = (Y - X\beta)'(Y - X\beta)$$

Comme $\beta \mapsto d(\beta)$ est quadratique positive, son minimum est obtenu où son gradient s'annule ; ce qui donne sous forme matricielle :

Equation normale.

$$X' X \hat{\beta} = X' Y$$

soit[10]

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

Dans le cas de l'exemple 1, on obtient $\beta_0 = 0.0662$ et $\beta_{pluie} = 1.6367$. Le graphique correspondant est :

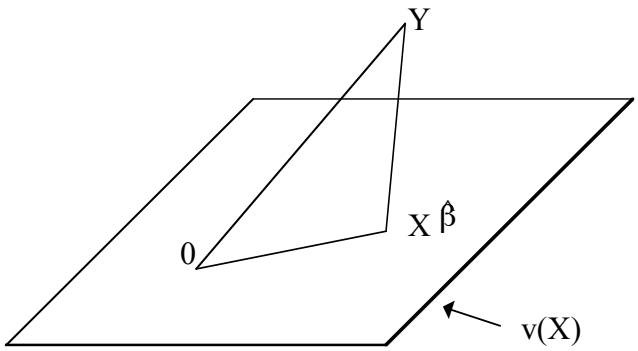
[10] Si X est de rang plein, i.e. p.

Ce graphique paraît "à vue de nez" satisfaisant mais cela reste à confirmer.

On va voir maintenant un moyen simple et géométrique d'obtenir l'équation normale qui nous permettra de plus de définir de premiers indicateurs de qualité de l'approximation.

ASPECT GEOMETRIQUE

Les n observations des diverses variables forment des vecteurs de l'espace $\mathbb{R}^n$ que l'on peut munir de la norme euclidienne. Si on appelle v(X) le sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de X, c'est à dire l'ensemble des vecteurs de la forme $X\beta$, pour $\beta \in \mathbb{R}^p$, on s'aperçoit que la quantité $d(\beta)$ à minimiser n'est rien d'autre que $\|Y - X\beta\|^2$. Par suite, le vecteur $X\hat{\beta}$ est la projection orthogonale de Y sur v(X).



Notation.

On notera désormais $\hat{Y} = X\hat{\beta}$: chaque composante $\hat{y}_i$ du vecteur $\hat{Y}$ représente la **prévision** du modèle linéaire pour l'observation numéro i (à comparer à la valeur observée y_i). Le vecteur $\hat{Y}$ est appelé **réponse estimée** et le vecteur $Y - \hat{Y}$ est le vecteur des **résidus.**

Notons que cette seule vision géométrique va nous permettre de dégager un premier outil de validation[11], encore très rustique, mais d'une utilité constante. En effet, le graphique précédent montre que le vecteur $Y - \hat{Y}$ est orthogonal à la réponse estimée $\hat{Y}$; ce qui signifie que :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,\hat{y}_i = 0$$

Comme par ailleurs, le vecteur des résidus $Y - \hat{Y}$ est orthogonal au vecteur X_0 formé de 1, la moyenne de ses composantes est nulle. On déduit de tout cela que :

$$\mathrm{cov}(Y - \hat{Y}, \hat{Y}) = 0$$

[11] Nous considérons ici le cas où la variable constante $x_0=1$ est un prédicteur.

On notera que la non corrélation précédente est toujours vérifiée, que le modèle soit satisfaisant ou non. Néanmoins, et cela se précisera avec le modèle probabiliste sous-jacent, on estime que du fait de cette non corrélation, un modèle correct doit montrer une réponse estimée non liée aux résidus[12]. Par suite, une simple représentation graphique des composantes de $Y - \hat{Y}$ contre celles de $\hat{Y}$ permettra de valider (ou plutôt d'invalider!) très rapidement la modélisation proposée. Dans le cas de l'exemple 1, on obtient ainsi :

On voit ici que les résidus semblent présenter une légère courbure ; cela nous suggère d'ajouter un terme quadratique à la régression[13] :

$$\text{rendement} \approx \beta_0 + \beta_{\text{pluie}}\ \text{pluie} + \beta_{\text{pluie}\hat{}2}\ \text{pluie}^2$$

Les résultats obtenus sont les suivants :

$$\beta_0 = -0.0425,\ \beta_{\text{pluie}} = 2.5017,\ \beta_{\text{pluie}\hat{}2} = -1.5228$$

[12] On verra plus loin le lien précis entre cette non corrélation et une indépendance probabiliste.

[13] On prendra garde ici au fait qu'une courbure sur ce graphique s'interprète ici en courbure sur le modèle du fait que rendement estimé et pluie sont liés de manière approximativement linéaire. On verra des outils plus adaptés pour compléter le modèle avec l'analyse complète des résidus.

et le tracé des résidus est reproduit ci-dessous :

On voit ici que ce modèle paraît bien plus satisfaisant dans le mesure où le tracé des résidus face à la réponse estimée ressemble au tracé de deux variables indépendantes. Par contre certains résidus semblent un peu grands mais il faudra attendre l'approche inférentielle avant de pouvoir répondre correctement à cette interrogation.

Pour revenir aux aspects plus théoriques de la régression, le lecteur perspicace n'aura pas manqué de noter l'analogie entre cette situation et l'étude de la régression linéaire "abstraite" que nous avons étudié précédemment. Il s'agit en fait de bien plus qu'une analogie puisque, dans le cas de prédicteurs non contrôlés, la régression linéaire est la limite, lorsque le nombre d'observations tend vers l'infini, de l'estimation par moindres carrés.

En effet, si $(y_i;(x_1)_i, \ldots, (x_p)_i)_{1\leq i\leq n}$ est un échantillon de taille n des variables aléatoires $(\mathbf{Y}, \mathbf{X_1}, \ldots, \mathbf{X_p})$, le terme d'indice (k,l) de la matrice X' X est :

$$\sum_{i=1}^{n} (x_k)_i \, (x_l)_i$$

De même, le terme d'indice k du vecteur X' Y est :

$$\sum_{i=1}^{n} (x_k)_i \, y_i$$

On en déduit, en divisant par n l'équation normale et en utilisant la loi des grands nombres, que l'équation normale tend vers l'équation :

$$E(\mathbf{X'\,X})\,\beta = E(\mathbf{X'\,Y})$$

Ce système n'est autre que l'équation de régression[14].

Ainsi, l'espace $L^2(P)$ peut être considéré comme l'espace limite des espaces euclidiens $\mathbb{R}^n$. D'un point de vue symétrique, on peut également voir la méthode des moindres carrés comme la version échantillonnée de la régression linéaire abstraite.

[14] On notera que l'on a considéré ici pour simplifier l'écriture une régression sans la constante $X_0=1$, qui peut être l'une des variables X_j.

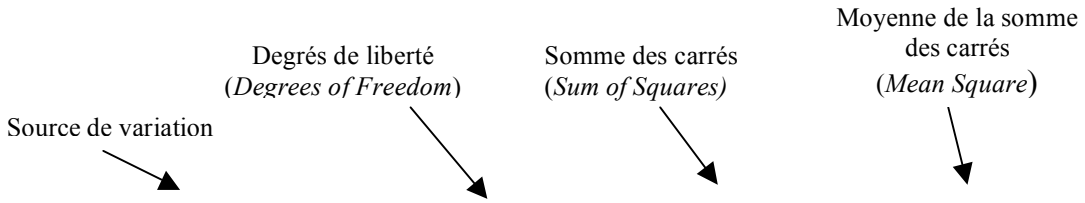
ANALYSE DE VARIANCE

La vision géométrique précédente permet en outre de définir simplement un indicateur de "qualité" pour le modèle estimé.

Tous les logiciels effectuant des calculs de régression, même le très rustique Excel[15], donnent comme sortie un tableau appelé table d'analyse de variance ou table d'ANOVA (ANalysis Of VAriance).

Celle-ci débute de la manière suivante (on verra par la suite comment interpréter la table complète) :

💣 On considère ici le cas où la variable constante $x_0 = 1$ est contenue dans l'ensemble des prédicteurs qui sont donc au total au nombre de p+1.



Source	DF	SS	MS
Regression	p	SSR	SSR/p
Error	n-p-1	SSE	SSE/(n-p-1)
Total	n-1	SST	
R^2	SSR/SST		

Ce tableau mérite quelques commentaires !

- Les diverses sommes des carrés sont composées comme suit :
  - La quantité SST représente la somme des carrés des écarts à la moyenne $\bar{y}$:

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

  - La quantité SSR représente la somme des carrés des écarts à la moyenne lorsque l'on remplace les observations y_i par les prévisions obtenues à l'aide de la régression :

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

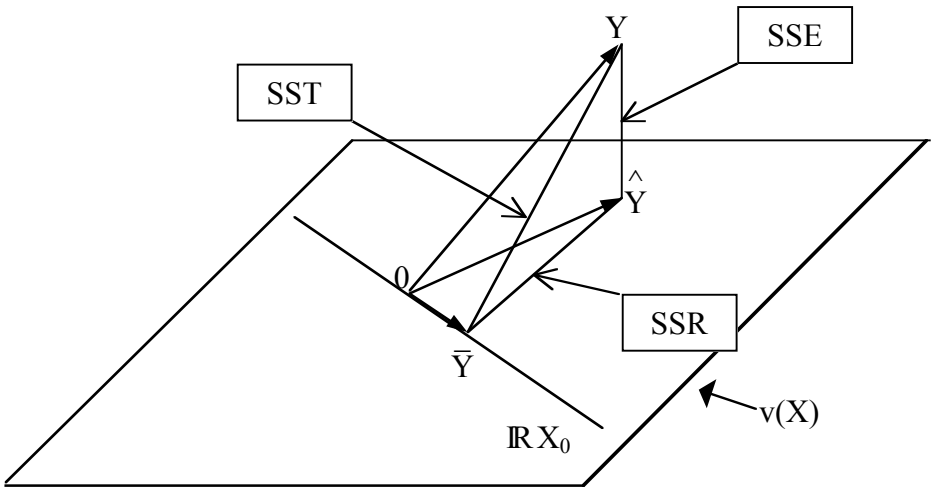
  - La quantité SSE représente la somme des carrés des écarts, ou erreurs (*error*) $y_i - \hat{y}_i$:

[15] Voir en annexe l'utilisation de Excel pour la régression linéaire.

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

- Le nombre de degrés de liberté indique le nombre de variables indépendantes (au sens algébrique du terme) qui permettent d'évaluer les sommes des carrés correspondantes (par exemple n-1 pour l'estimation de SST).
- Les quantités MSR et MSE visent à estimer les variances des écarts considérés pour le calcul de SSR et SSE. Elles sont obtenues en divisant la somme des carrés par le nombre de degrés de liberté correspondants. A noter que la quantité MST n'est en général pas reproduite par les logiciels : c'est l'estimation sans biais de la variance des données y_i.

Toutes ces quantités s'interprètent de manière géométrique (**prendre garde au fait que les diverses sommes de carrés sont les carrés des longueurs représentées ci-dessous**), comme le montre le graphique suivant, où l'on a représenté les vecteurs Y, $\hat{Y}$ et $\overline{Y} = \overline{y}$ (1 … 1)' :



De plus, du fait que $Y - \hat{Y}$ est orthogonal à v(X), il est clair que le triangle $Y\hat{Y}\overline{Y}$ est rectangle en $\hat{Y}$. Par suite, on obtient la très importante :

Formule d'analyse de variance

$$SST = SSR + SSE$$

De la relation précédente découle la définition du :

Coefficient de détermination

Le coefficient de détermination R^2 est défini par : $R^2 = \frac{SSR}{SST}$

Notons que R^2 est également (voir le graphique) le carré du cosinus de l'angle entre les vecteurs $Y - \bar{Y}$ et $\hat{Y} - \bar{Y}$ et mesure donc – de façon a-dimensionnelle – la proximité de Y et $\hat{Y}$.

 Il convient de bien comprendre à quoi peut servir le coefficient de détermination, et surtout à quoi il ne peut pas servir!

Le coefficient R^2 mesure presque[16] la proportion de variance expliquée par la régression. Par exemple, si $R^2 \approx 0$, les prédicteurs sont sans effet sur la réponse et si $R^2 \approx 1$, ils expliquent au contraire complètement la réponse y. En fait, ce coefficient mesure un rapport du type signal/bruit et permet donc de quantifier la réduction du niveau de bruit apportée par le modèle. Nous verrons plus loin comment quantifier les symboles $\approx$ qui précèdent.

Disons pour résumer que R^2 est un premier indicateur d'intérêt du modèle considéré, mais pas plus!

Par exemple, il ne faut surtout pas utiliser R^2 pour **valider** le modèle linéaire proposé. Pour ce faire, nous avons déjà vu que l'examen des résidus était plus adapté : nous irons plus avant sur ce sujet par la suite.

Enfin, il n'est pas recommandé d'utiliser R^2 pour **comparer** des modèles. Par exemple, si on hésite entre $y \approx \beta_0 + \beta_1 x_1$ et $y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$, le coefficient R^2 nous dira toujours de choisir le second modèle car son R^2 sera plus important (on projette sur un espace plus grand), même si la variable x_2 est sans rapport avec la réponse y !

Pour comparer des modèles, en pénalisant les modèles les plus complexes, il existe de nombreux indicateurs. Parmi ceux-ci, le coefficient de détermination ajusté découle simplement de notre table d'analyse de variance :

Coefficient de détermination ajusté

Le coefficient de détermination ajusté R^2-adj est défini par : $R^2\text{-adj} = 1 - \frac{MSE}{MST}$

Notons que R^2 et R^2-adj sont liés fonctionnellement :

$$1 - R^2\text{-adj} = \frac{n-1}{n-p-1}(1 - R^2)$$

Cette formule indique notamment que R^2-adj est toujours inférieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs. A noter également que si le nombre de prédicteurs augmente, le coefficient R^2-adj peut devenir négatif !

[16] Il faudrait en fait diviser les quantités SSR et SST par leur nombre de degrés de liberté respectifs, et non pas par n comme sous-entendu ici : voir le R^2 ajusté.

Notons enfin que les coefficients R^2 et R^2-adj sont peu différents lorsque le nombre n d'expériences est grand devant le nombre p de prédicteurs.

Les tables d'analyse de variance obtenues pour l'exemple 1 sont données ci-dessous :

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ pluie

Source	DF	SS	MS
Regression	1	1.279	1.279
Error	52	0.141	0.003
Total	53	1.420	
R^2	0.901	R^2-adj	0.899

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ pluie $+ \beta_{pluie\wedge 2}$ pluie^2

Source	DF	SS	MS
Regression	2	1.298	0.649
Error	51	0.122	0.002
Total	53	1.420	
R^2	0.914	R^2-adj	0.911

On observe que là encore, le modèle quadratique semble légèrement meilleur que le modèle linéaire. De ce fait, et surtout du fait que les résidus sont mieux répartis, on adoptera pour l'instant ce modèle de régression.

INSUFFISANCES

On voit à travers les notions présentées précédemment – et ceci est particulièrement vrai du coefficient de détermination – que la plupart de nos analyses restent de nature essentiellement qualitatives.

En effet, mise à part la procédure d'estimation des paramètres par moindres carrés, les objectifs que nous nous sommes fixés en début de chapitre (vérification du caractère prédictif de la variable pluie, validité du modèle construit, détection d'observations aberrantes, ...) paraissent largement hors de portée. Disons pour être plus précis qu'il est toujours possible de fixer par exemple un seuil pour le coefficient R^2 en deçà duquel on considère que la variable pluie n'est pas influente mais qu'une méthode systématique d'obtention de ce seuil est pour l'instant hors de portée.

C'est l'intervention de modèles probabilistes, et de procédures statistiques associées, qui va permettre d'aller plus loin. Avant cela, étudions un peu en détail le cas , de loin le plus courant, où plusieurs prédicteurs sont présents :

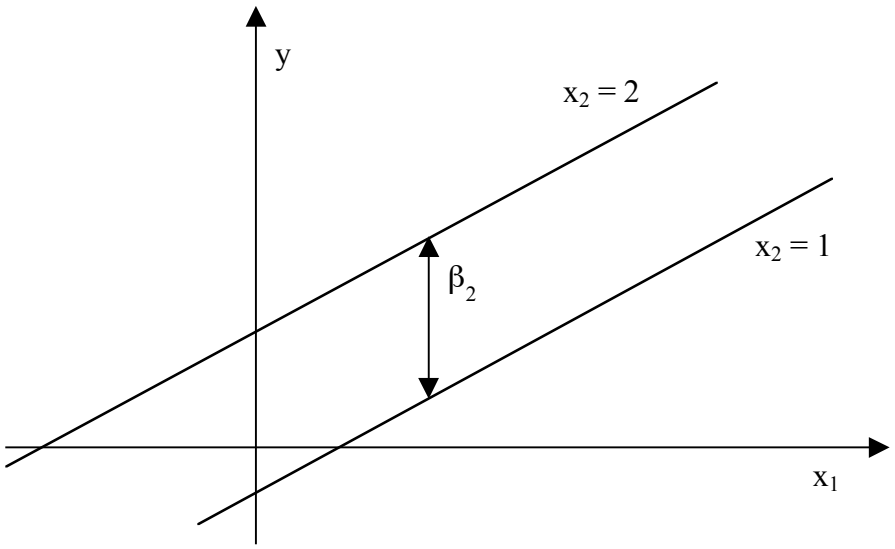
PREDICTEURS MULTIPLES

Nous donnons ici quelques notions simples qui visent à faciliter la modélisation :

Considérons tout d'abord une régression qui relie la réponse y à deux prédicteurs quantitatifs x_1 et x_2 :

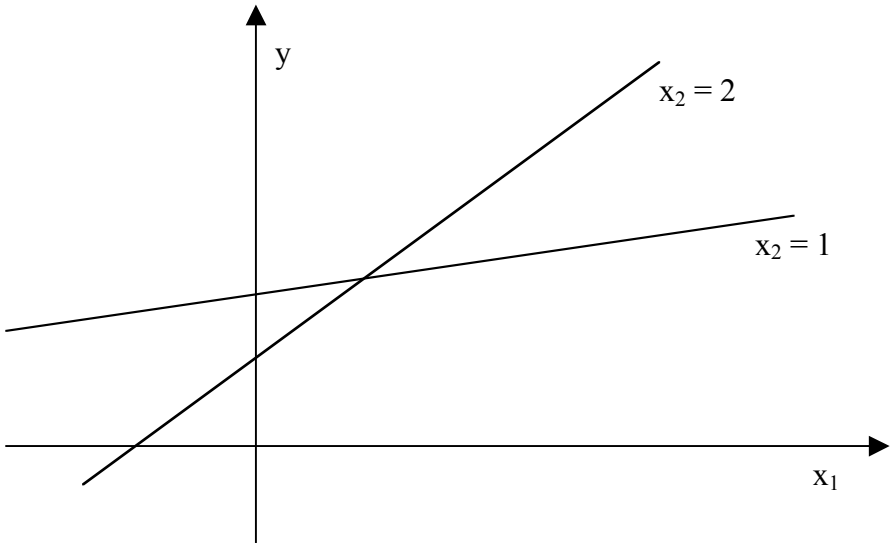
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Un tel modèle, universellement connu (!), masque des hypothèses sous-jacentes. Il suppose en effet - entre autres - que l'influence marginale de la variable x_1 sur la réponse y n'est pas modifiée par le niveau pris par le deuxième prédicteur x_2 (idem en échangeant x_1 et x_2). En d'autres termes, lorsque x_2 est fixé, les droites de régression donnant y en fonction de x_2 sont parallèles :



Si par contre, on veut modéliser une interaction (au niveau des effets sur la réponse y) entre les prédicteurs x_1 et x_2, on peut ajouter un terme produit $x_1 x_2$ au modèle pour obtenir une **régression linéaire avec interaction** :

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{1,2} x_1 x_2$$



L'intérêt d'un tel modèle est notamment d'être linéaire en chaque prédicteur – **lorsque les autres sont fixés**. Il convient de bien faire la différence entre un tel modèle et un modèle quadratique : bien que mathématiquement, un tel modèle soit quadratique, on l'interprète comme un modèle linéaire étendu qui permet la prise en compte d'interactions.

Ne pas confondre interaction entre deux prédicteurs et corrélation entre ces deux mêmes prédicteurs : ce sont deux notions sans **aucun** rapport entre elles.

Venons-en au cas de prédicteurs discrets : on a vu au chapitre portant sur l'aspect formel de la régression que dans le cas de tels prédicteurs, la régression générale s'évalue simplement à partir d'une régression linéaire. Considérons par exemple le cas d'une réponse y qui dépend d'un seul prédicteur x, supposé discret, de modalités $a_1,\ldots, a_k$.

Prenons un exemple : une compagnie d'assurances américaine s'inquiète des coûts de maintenance informatique de ses agences et d'interroge notamment sur le lien éventuel entre l'Etat dans lequel l'agence est située et le coût observé. Pour 3 états et 10 agences choisies au hasard dans chaque état, on obtient les résultats suivants (le nombre indiqué est le coût par machine observé en $) :

Kansas	Kentucky	Texas
198	563	385
126	314	693
443	483	266
570	144	586
286	585	178
184	377	773
105	264	308
216	185	430
465	330	644
203	354	515

En notant y la variable coût, on pourra alors écrire un modèle du type :

$$y \approx \beta_{Ka}\, 1_{\{x=Ka\}} + \beta_{Ke}\, 1_{\{x=Ke\}} + \beta_{Te}\, 1_{\{x=Te\}}$$

On peut également paramétrer ce modèle sous une forme différente ; par exemple en ajoutant la constante :

$$y \approx \beta_{Ka} + (\beta_{Ke} - \beta_{Ka})\, 1_{\{x=Ke\}} + (\beta_{Te} - \beta_{Ka})\, 1_{\{x=Te\}}$$

On prendra garde au fait, déjà signalé plus haut, que l'on ne peut pas ajouter la constante en gardant les 3 variables indicatrices. On parviendrait alors à un modèle non identifiable, c'est à dire dont les paramètres ne peuvent pas être estimés (numériquement, la matrice X'X devient non inversible). On notera également que les paramètres associés aux divers prédicteurs changent alors d'interprétation (par exemple, la nullité du coefficient associé au prédicteur $1_{\{x=Ke\}}$ a trait à l'identité des réponses moyennes observées entre les états du Kansas et du Kentucky).

Signalons pour finir que l'on peut à partir de ce formalisme modéliser l'interaction entre un prédicteur qualitatif et un autre prédicteur (qualitatif ou quantitatif) en multipliant comme précédemment les prédicteurs concernés.

Régression linéaire - aspect probabiliste

Afin de pallier aux insuffisances détectées au chapitre précédent, la démarche consiste à enrichir le modèle en le transformant en modèle probabiliste. Nous répondrons ainsi aux questions posées en utilisant le formalisme statistique. Evidemment, les hypothèses sous-jacentes devront être contrôlées si l'on désire que la démarche proposée ait quelque fondement !

LE MODELE LINEAIRE

On reprend donc ici le chapitre qui précède en ajoutant quelques hypothèses sur les écarts au modèle.

Une réponse (aléatoire) y est une fonction linéaire d'un certain nombre de variables x^1, x^2,..., x^p déterministes ou aléatoires[17] modulo une erreur (aléatoire) ε supposée additive :

$$y = \beta_0 x^0 + \beta_1 x^1 + ...+ \beta_p x^p + \varepsilon$$

En tenant compte des n observations faites pour la réponse et les prédicteurs, on obtient :

Notations
$Y = [y_1 ... y_n]'$ vecteur n×1 des réponses
$X = [x^j_i]$ matrice n×(p+1)
$\beta = [\beta_0 ... \beta_p]'$ vecteur (p+1)×1 des paramètres
$\varepsilon = [\varepsilon_1 ... \varepsilon_n]'$ vecteur n×1 des écarts au modèle

et le modèle s'écrit alors sous la forme :

$$Y = X \beta + \varepsilon$$

où ε_1, ..., ε_n sont les réalisations de v. a. $\boldsymbol{\varepsilon_1}$, ..., $\boldsymbol{\varepsilon_n}$ indépendantes de même loi normale $N(0,\sigma^2)$

Un tel modèle est appelé **modèle linéaire**

En utilisant les notations précédentes, l'expression vectorielle du modèle linéaire est particulièrement concise :

$Y = X \beta + \varepsilon$,
où ε est la réalisation d'un vecteur aléatoire $\boldsymbol{\varepsilon}$ de loi $N(0,\sigma^2 Id)$

[17] auxquelles on ajoutera toujours pour simplifier la "variable constante" (!) égale à 1, notée x^0.

Commentaires

- Le formalisme que nous allons développer reste valable si la constante x^0 n'est pas utilisée. Dans ce cas, tous les résultats qui suivent doivent être adaptés en conséquence.
- Le terme statistique consacré pour l'écart entre réponse et modèle est celui de **résidus** : on parle du vecteur ε des résidus et du résidu ε_i associé à la i-ème observation.
- Très souvent, le vecteur ε des résidus est considéré dans la littérature comme une erreur de mesure sur la réponse y. Cette vision par trop simpliste est en fait très restrictive et on n'hésite à assimiler ces résidus à une erreur de modélisation, voire à un mélange entre ces deux types d'erreurs[18], **dans la mesure où la suite des résidus ε_i se comporte comme une suite de v. a. indépendantes de loi $N(0,\sigma^2)$**.
- Toujours dans ce registre, si les résidus se réduisent quelquefois à une erreur de modélisation, on peut s'attendre dans de tels cas à ce que leur comportement se rapproche de celui d'une suite de v.a. dépendantes (du fait de la régularité supposée du modèle régissant le phénomène) et il faut alors tenir compte de ces corrélations. On flirte alors avec les techniques de régression non paramétrique et à la géostatistique.
- Encore sur cette question, si les résidus observés (on verra plus loin comment les estimer correctement) montrent une variance apparente constante, on parle d'**homoscédasticité** et dans le cas contraire d'**hétéroscédasticité**. Dans ce dernier cas, on est souvent amené à transformer la réponse afin de stabiliser la variance. On renvoie au cours de séries temporelles de Olivier [Roustant], au livre consacré à la régression linéaire [Jobson], ou encore au difficile mais passionnant [Hastie] pour des notions sur les procédés de stabilisation de la variance.
- Enfin, l'hypothèse de normalité des résidus peut étonner a priori. Disons sur ce sujet que dans un grand nombre de situations, on observe des résidus à peu près symétriques autour de 0, avec un histogramme en forme de "cloche". Un test de normalité montre alors dans la plupart des cas que l'hypothèse normale n'est pas en désaccord flagrant avec nos observations et cette hypothèse est conservée dans toute l'analyse. Cependant, lorsque par exemple les résidus sont très dissymétriques, il faut faire une autre hypothèse sur la loi sous-jacente. Une alternative dans ce cas est d'utiliser les techniques de modèles linéaires généralisés (voir [McCullagh]) qui consistent à utiliser d'autres lois de probabilités, choisies parmi une famille, appelée la famille exponentielle, qui est bien connue par ailleurs pour ses bonnes propriétés vis-à-vis de l'estimation statistique (existence d'estimateurs efficaces,…).
- On trouve souvent dans la littérature des hypothèses plus faibles sur les résidus. Dans cette approche, on suppose que les résidus ε_1, ..., ε_n sont les réalisations de v. a. $\boldsymbol{\varepsilon_1}$, ..., $\boldsymbol{\varepsilon_n}$ d'espérance nulle, de variance σ^2, non corrélées. Il s'agit d'un affaiblissement notable des hypothèses précédentes qui ne nous permettra pas de réaliser entièrement l'inférence désirée ; nous l'appellerons **modèle linéaire - version faible**. On notera au fur et à mesure du texte les moments où l'inférence est possible sous cette forme faible des hypothèses.

[18] Il s'agit en fait du cas le plus courant.

FONCTION COUT, HYPOTHESES SUR LE BRUIT ET ESTIMATION DES PARAMETRES

D'un point de vue statistique, le problème d'estimation du vecteur β des paramètres peut être traité par la théorie classique de l'estimation. Plus précisément, les réponses observées y_i sont les réalisations indépendantes de variables aléatoires de loi $N((X\beta)_i,\sigma^2)$. La vraisemblance de l'échantillon $(y_1,\ldots, y_n)$ observé est donc :

$$L(y_1,\ldots, y_n ; \beta,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i-(X\beta)_i)^2}{2\sigma^2}\right]$$

soit :

$$L(y_1,\ldots, y_n ;\beta,\sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[- \sum_{i=1}^{n} \frac{(y_i-(X\beta)_i)^2}{2\sigma^2} \right]$$

On en déduit que l'estimation du vecteur β par maximum de vraisemblance est équivalente à la procédure des moindres carrés.

Mais la conclusion importante de ce simple calcul est surtout que l'on voit ici de quelle façon les hypothèses probabilistes sur les résidus influencent la fonction coût qui sera utilisée pour estimer β. Par exemple, si l'écart-type du résidu ε_i dépend de i – notons le σ_i, le critère à minimiser sera de la forme :

$$\sum_{i=1}^{n} \frac{(y_i-(X\beta)_i)^2}{2\sigma_i^2}$$

On parle alors de moindres carrés pondérés. Dans le même esprit (voir le cours de séries temporelles [Roustant]), si les résidus sont corrélés, la matrice de covariance des résidus n'est plus diagonale et la fonction coût à minimiser est une forme quadratique non réduite à une somme de carrés : ce sont les moindres carrés généralisés. Enfin, si la loi des résidus n'est pas normale, la fonction coût n'est même plus quadratique et la minimisation de cette fonction nécessite la mise en œuvre d'un algorithme itératif d'optimisation. Notons cependant que dans ce cas, des considérations probabilistes peuvent aider au choix de l'algorithme ; c'est le cas par exemple pour les modèles linéaires généralisés (voir [McCullagh]).

INFERENCE SUR LES PARAMETRES

Dans toute cette partie, on notera $\hat{\beta}$ l'estimation (ou l'estimateur suivant le contexte!) de β par moindres carrés.

Estimation de la variance σ^2.

On pose $\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \frac{SSE}{n-p-1} = MSE$

Théorème de Gauss-Markov.

Sous les hypothèses du modèle linéaire - version faible – on a :

$\hat{\beta}$ est BLUE (best linear unbiaised estimate), c'est à dire que :

- (linear) $\hat{\beta}$ est une fonction linéaire du vecteur des données Y
- (unbiaised) $E(\hat{\beta}) = \beta$
- (best) si $\tilde{\beta}$ est un autre estimateur linéaire non biaisé, on a :

$\forall\ \alpha \in \mathbb{R}^p$, $var(\alpha' \hat{\beta}) \leq var(\alpha' \tilde{\beta})$.

Enfin, on a : $Cov(\hat{\beta}) = \sigma^2 (X' X)^{-1}$

$\hat{\sigma}^2$ est un estimateur non biaisé de σ^2

Eléments de preuve.

Il est tout d'abord évident que $\hat{\beta}$ est linéaire en Y puisque $\hat{\beta} = (X' X)^{-1} X'Y$

Par ailleurs, du fait que $\hat{\beta} = (X' X)^{-1} X' (X\beta + \boldsymbol{\varepsilon}) = \beta + (X' X)^{-1} X'\boldsymbol{\varepsilon}$, on en déduit tout d'abord que $\hat{\beta}$ est sans biais (car $\boldsymbol{\varepsilon}$ est centré), puis que la matrice de covariance de $\hat{\beta}$ s'évalue comme suit :

$cov(\hat{\beta}) = E\left((\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right) = (X' X)^{-1} X' E(\boldsymbol{\varepsilon}\ \boldsymbol{\varepsilon}') X (X' X)^{-1} = \sigma^2 (X' X)^{-1}$

De la même façon, on montre que $cov(Y-\hat{Y}) = \sigma^2 (Id - X (X' X)^{-1} X')$. Le calcul de l'espérance de SSE en découle[19] (et donc le non biais de $\hat{\sigma}^2$).

[19] On utilise pour cela le fait élémentaire suivant : si **u** est un vecteur aléatoire centré de matrice de covariance Γ, on a : $E(\mathbf{u}\ \mathbf{u}') = Trace(\Gamma)$.

Enfin, nous laissons le meilleur morceau aux bons soins du lecteur (vérifier que $\hat{\beta}$ est le "meilleur"). Indiquons seulement ici les phases de la démonstration de ce résultat. En écrivant $\tilde{\beta}$ sous la forme $\tilde{\beta} = MY$ et en utilisant le non biais de $\tilde{\beta}$, on montre tout d'abord que $MX = Id$. On évalue alors la quantité $E\left((\hat{\beta}-\tilde{\beta})(\hat{\beta}-\beta)'\right)$, d'où l'on déduit facilement que $\hat{\beta} - \tilde{\beta}$ est non corrélé avec $\hat{\beta}$. Il en est donc de même de $\alpha'(\hat{\beta} - \tilde{\beta})$ et $\alpha' \tilde{\beta}$ et l'inégalité sur les variances en découle alors simplement.

Remarque

On notera qu'au facteur σ^2 près la matrice de covariance de $\hat{\beta}$ ne dépend que de la matrice du plan d'expérience. Ce résultat est **essentiel** pour les applications puisqu'il est à la base de la théorie des **plans d'expérience** qui concerne le cas de variables explicatives contrôlées. Sans vouloir aborder cette très riche théorie, signalons l'idée qui sous-tend cette dernière. On a vu dans ce qui précède de quelle manière utiliser "au mieux" les résultats de l'expérience pour estimer les paramètres du modèle (moindres carrés ou maximum de vraisemblance). Par contre, dans la mesure où l'on contrôle les prédicteurs, on peut mener notre réflexion plus en amont. En d'autres termes, on peut se demander s'il y a moyen de fixer les niveaux des prédicteurs de manière à estimer "au mieux" le vecteur β. A cette question tentent de répondre de nombreuses techniques ; toutes ont pour point commun de raisonner à partir de la matrice de covariance $\sigma^2 (X' X)^{-1}$, le but étant d'essayer de la rendre la "plus petite possible". Par exemple, on va chercher à rendre les coefficients extra-diagonaux nuls (non corrélation des estimations) : on parle alors de plans d'expérience orthogonaux. On peut également s'attacher à rendre les termes diagonaux les plus petits possibles (variances d'estimation faibles). On touche alors la très riche théorie de l'optimalité avec les plans D-optimaux, A-optimaux ... Pour davantage de détails sur ce domaine très riche d'applications, on pourra consulter par exemple [Benoist].

Venons-en maintenant aux premiers résultats inférentiels qui vont nous permettre de tester des hypothèses, détecter des prédicteurs non significatifs,…

Loi des estimateurs et statistiques pivotales.

(i) Le vecteur $\hat{\beta}$ est de loi normale $N(\beta, \sigma^2(X' X)^{-1})$.

(ii) $\dfrac{(n-p-1)\,\hat{\sigma}^2}{\sigma^2} = \dfrac{\|Y - X\hat{\beta}\|^2}{\sigma^2}$ est de loi χ^2_{n-p-1} et $\hat{\sigma}$ est indépendant de $\hat{\beta}$.

(iii) Pour tout $j \in \{0, \ldots, p\}$, en notant c_j le j-ème terme diagonal de la matrice $(X' X)^{-1}$, la variable $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{c_j}\,\hat{\sigma}}$ est de loi de Student t_{n-p-1}.

(iv) **Sous l'hypothèse** $H_0 : \beta_1 = \ldots = \beta_p = 0$, l'écart relatif[20] $\dfrac{\|X^0 \bar{y} - X \hat{\beta}\|^2}{p \hat{\sigma}^2} = \dfrac{MSR}{MSE}$ est de loi de Fisher-Snedecor F^p_{n-p-1}.

Ces résultats permettent d'interpréter les tables d'analyse de variance complètes fournies par tout logiciel mettant en œuvre la régression linéaire[21].

La table complète est du type suivant :

Table d'analyse de variance

Source	DF	SS	MS	F	p
Regression	p	SSR	SSR/p	MSR/MSE	p value
Error	n-p-1	SSE	SSE/(n-p-1)		
Total	n-1	SST			

R^2 SSR/SST

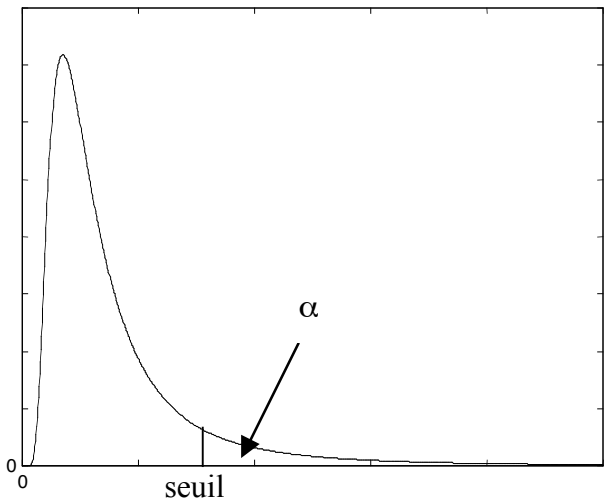
Nous venons de voir que la loi de la quantité F était connue sous l'hypothèse :

$H_0 : \beta_1 = \ldots = \beta_p = 0$

Cette hypothèse signifie que nos prédicteurs sont sans influence réelle (en tout cas telle qu'elle a été modélisée!) sur la réponse. Un test est donc bâti sur le résultat précédent : on oppose l'hypothèse nulle à l'hypothèse alternative :

$H_1 : \exists\, j \in \{1,\ldots,p\},\ \beta_j \neq 0$

Intuitivement, nous rejetterons l'hypothèse nulle lorsque la somme des carrés expliquée par la régression est grande. En d'autres termes, la région critique de ce test est de la forme {F > seuil}. L'interprétation graphique de ce test est donnée ci-dessous :

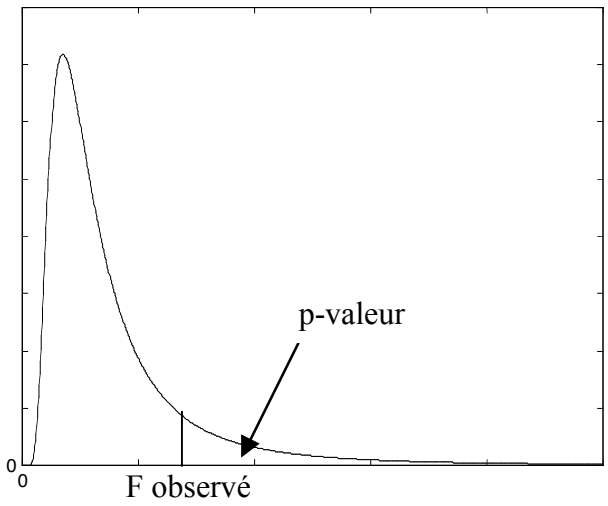


et notre règle de décision est alors la suivante. Si la quantité F observée dépasse seuil, on rejette l'hypothèse H_0 (au niveau α) et dans le cas contraire, on conserve H_0.

[20] On note X^0 le vecteur colonne formé des valeurs de la variable x^0, i.e. de 1.

[21] On renvoie toujours à l'annexe concernant l'utilisation de Excel pour la régression.

Cependant, la plupart des logiciels ne demandent pas de fixer a priori un niveau α pour effectuer le test : ils donnent en sortie le niveau à partir duquel notre décision aurait changé : c'est la notion de **p-valeur**. Le graphique qui suit rappelle la signification de cette notion :



En d'autres termes, la p-valeur du test est la probabilité, si H_0 était la bonne hypothèse, d'avoir observé une valeur pour F qui ait dépassé le F que nous avons observé. Au niveau décisionnel, on rejettera donc H_0 lorsque la p-valeur est faible. L'avantage d'une telle approche vis-à-vis d'un test "classique" est de ne pas fixer un niveau α a priori. En effet, lorsque l'on effectue un test à un niveau de 5% et que l'on rejette H_0, on ne sait pas quelle aurait été notre décision pour un niveau de 4%. Ici, par contre, une p-valeur de 1,5% nous dit que nous aurions rejeté au niveau 5% et accepté au niveau 1%. Une telle approche permet en quelque sorte de faire le test pour tous les niveaux à la fois[22].

Afin de concrétiser les choses, voici ce que donne la table d'analyse de variance complète dans le cas de l'exemple 1 :

Modèle rendement $\approx \beta_0 + \beta_{pluie}$ pluie

Source	DF	SS	MS	F	p
Regression	1	1.279	1.279	472.9	0
Error	52	0.141	0.003		
Total	53	1.420			
R^2	0.901	R^2-adj	0.899		

On voit sur cet exemple que la valeur observée F est tellement grande que la p-valeur correspondante est un zéro machine. Une telle situation est assez courante car la plupart du temps, les prédicteurs que l'on choisit ont globalement un effet sur la réponse ! Ce test doit donc être considéré essentiellement comme un garde-fou et la lecture d'une p-valeur faible nous assure que les efforts de modélisation que nous allons faire ensuite ont une chance d'être couronnés de succès.

[22] Néanmoins, cette notion simple cache un peu son jeu puisque un peu de réflexion montre que la p-valeur est, sous H_0, la réalisation d'une v.a. uniforme sur [0,1]…

Par contre, les tables qui suivent – associées en général à la table d'analyse de variance – sont d'une grande importance pratique :

Variable	Coeff	Std[23]	t value	p value
Intercept	$\hat{\beta}_0$	$\sqrt{c_0}\,\hat{\sigma}$	$Coeff_0/Std_0$	
x_1	$\hat{\beta}_1$	$\sqrt{c_1}\,\hat{\sigma}$	$Coeff_1/Std_1$	
.	.	.	.	
.	.	.	.	
.	.	.	.	
x_p	$\hat{\beta}_p$	$\sqrt{c_p}\,\hat{\sigma}$	$Coeff_p/Std_p$	

Il convient d'interpréter correctement ces tables.

- Les prédicteurs sont nommés dans la première colonne (le prédicteur constant est souvent appelé **intercept** par les logiciels anglo-saxons).
- La colonne suivante donne les estimations des coefficients associés à chaque prédicteur.
- L'écart-type Std est l'écart-type d'estimation du coefficient (cf. th. de Gauss-Markov).
- La t value est simplement le rapport entre l'estimation du coefficient et son écart-type. **Si le coefficient β_j est nul**, ce rapport est de loi de Student t_{n-p-1} (cf. proposition précédente, point (iii)).
- La p-valeur de la dernière colonne est la probabilité, pour une variable de loi t_{n-p-1}, de dépasser en valeur absolue la t value observée pour la j-ème variable. En d'autres termes, cette p-valeur permet de tester l'hypothèse nulle $H_0 : \beta_j = 0$.

Revenons dans cet esprit sur l'exemple 1 :

Modèle rendement ≈ $\beta_0 + \beta_{pluie}$ pluie

Variable	Coeff	Std	t value	p value
Intercept	0.0662	0.02405	2.752	0.00813
pluie	1.637	0.07526	21.75	0

On voit ici que les deux variables utilisées sont très significatives puisque les p-valeurs sont toutes deux très faibles.

Voyons ce qu'il en est du modèle de degré 2 :

Modèle rendement ≈ $\beta_0 + \beta_{pluie}$ pluie + $\beta_{pluie\hat{}2}$ pluie²

Variable	Coeff	Std	t value	p value
Intercept	-004246	0.04512	-0.9411	0.3511
pluie	2.502	0.3187	7.85	2.494e-10
pluie²	-1.523	0.547	-2.784	0.007518

Il apparaît très clairement que le terme quadratique est considéré comme étant très significatif. Par contre, le terme constant est douteux. Nous le conserverons ici du fait que la plupart des logiciels considèrent ce terme comme faisant partie des prédicteurs et que les tables d'analyse de variance s'y réfèrent mais on pourrait l'ôter a priori[24].

A partir de là, on peut se demander si l'ajout d'une puissance, i.e. l'utilisation d'un terme cubique, est susceptible d'améliorer la modélisation. C'est pourquoi on reproduit la table qui suit :

[23] Pour la définition des coefficients c_j, voir la proposition précédente p. 24.
[24] Une analyse des résidus montrerait qu'un tel modèle est satisfaisant.

Modèle rendement ≈ $\beta_0 + \beta_{pluie}$ pluie + $\beta_{pluie\hat{}2}$ pluie² + $\beta_{pluie\hat{}3}$ pluie³

Variable	Coeff	Std	t value	p value
Intercept	-0.01225	0.07243	-0.1691	0.8664
pluie	2.031	0.9363	2.169	0.03488
pluie²	0.4484	3.721	0.1205	0.9046
pluie³	-2.419	4.516	-0.5357	0.5945

On voit ici des résultats pour le moins surprenants puisque, mis à part le prédicteur pluie, toutes les variables sont considérées comme non significatives. Concrètement, cela signifie qu'il y a globalement trop de variables dans notre modèle, mais de façon apparemment étrange, la variable pluie^3 semble être la dernière à ôter. Tout cela vient du fait que nous travaillons avec des prédicteurs très corrélés et que dans ce cas, les résultats inférentiels sont à interpréter avec précaution car la non significativité apparente d'un prédicteur peut venir d'un autre avec qui il est très corrélé. Il s'agit ici d'un problème de non identifiabilité du modèle. Ici, la matrice donnant les coefficients de corrélations entre les variables pluie, pluie^2 et pluie^3 est la suivante :

1.0000	0.9750	0.9292
0.9750	1.0000	0.9872
0.9292	0.9872	1.0000

On voit que les corrélations entre les variables – notamment entre pluie^2 et pluie^3 - est très forte ; ce qui explique les résultats observés précédemment.

Un exemple

Cet exemple est un exemple historique puisque le terme régression est né de ces préoccupations. Une des théories défendues par Francis Galton était que la race humaine avait tendance à "régresser vers la moyenne" si aucune politique volontariste (eugénisme) n'était mise en place. En d'autres termes, sans entrer dans des débats philosophiques difficiles, il estimait que la variabilité de l'espèce diminuait avec le temps.

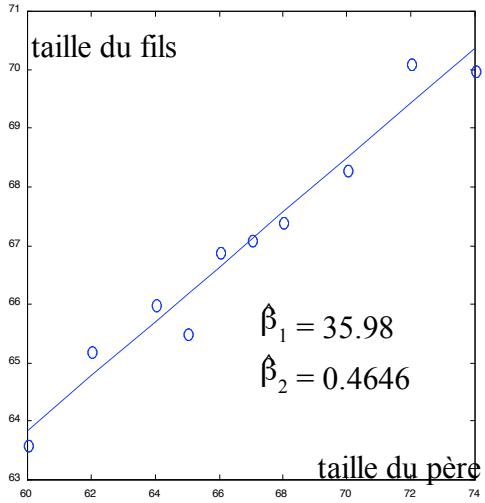
Pour vérifier cette hypothèse, il comparait la taille d'enfants mâles avec celle de leur père. Les données correspondantes, reproduites ci-dessous, ont été ensuite étudiées par Karl Pearson (les tailles sont données en pouces!) :

taille père	60	62	64	65	66	67	68	70	72	74
taille fils	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

L'ajustement d'un modèle linéaire du type :

$$taille\_fils = \beta_1 + \beta_2\ taille\_père$$

donne :



La question de Galton est de savoir si β_2 est inférieur ou supérieur à 1. On effectue donc le test :

$H_0 : \beta_2 \geq 1$ contre $H_1 : \beta_2 < 1$.

La région critique de niveau 1% est donnée par[25] :

$$\hat{\beta}_2 < 1 - t_8^{-1}(0.99) \frac{\hat{\sigma}}{\sigma_{\text{taille-père}} \sqrt{10}}$$

dans le cas présent, la région critique est donnée par : $\hat{\beta}_2 < 0.9045$. On rejette donc H_0 et l'hypothèse de Galton semble validée.

N.B. les données sont suspectes (chiffres ronds pour les pères, une décimale pour les fils)... Ce sont assurément des données qui proviennent d'une base plus riche dans laquelle des regroupements ont été faits. En conséquence, nos conclusions sont basées sur des données assez douteuses, et donc ne sont pas des plus affirmées !

SELECTION DE MODELES

Dans la même esprit que les tables d'analyse de variance, il est possible de comparer des modèles emboîtés (en anglais *nested models*) afin de voir si la réduction de variance due au passage à un modèle plus riche est significative.

De façon précise, deux modèles emboîtés sont de la forme suivante :

Modèle complet

$$y \approx \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q + \ldots + \beta_p x_p$$

Modèle réduit

$$y \approx \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q \text{ (avec } q < p)$$

[25] on note $t_n^{-1}(1-\alpha)$ le quantile d'ordre $1-\alpha$ de la loi de Student t_n.

La démarche consiste à partir du modèle complet que l'on suppose valide et à voir si la suppression des prédicteurs x_{q+1}, …, x_p ne détériore pas significativement le modèle. Choisir entre ces deux modèles revient à choisir entre les deux hypothèses :

$H_0 : \beta_{q+1} = \ldots = \beta_p = 0$
$H_1 : \exists\, j \in \{q+1,\ldots,p\}, \beta_j \neq 0$

Pour cela, nous allons construire un test de Fisher basé sur l'analyse de variance. On notera SSR_C, MSR_C, … les quantités qui concernent le modèle complet et SSR_R, MSR_R, … les quantités correspondantes pour le modèle réduit.

Analyse de variance pour des modèles emboîtés

Sous l'hypothèse $H_0 : \beta_{q+1} = \ldots = \beta_p = 0$, la statistique $F = \dfrac{(SSE_R - SSE_C)/(p-q)}{SSE_C/(n-p-1)}$ suit une loi de Fisher $F_{p-q,n-p-1}$.

On utilise ce test de la même manière que le test F de significativité de l'ensemble des prédicteurs. En d'autres termes, on rejettera l'hypothèse nulle lorsque la quantité F est trop grande – cas où l'accroissement de variance due au passage au modèle réduit est grand. Les logiciels qui effectuent ce test mènent à raisonner comme précédemment à partir des p-valeurs.

Remarques

- Si l'on veut tester la nullité d'un coefficient β_j, on peut utiliser une variable de Student comme précédemment ou le test de Fisher pour les modèles emboîtés. Rassurons tout de suite le lecteur : dans ce cas, la quantité F n'est rien d'autre que le carré de la t-valeur et par conséquent, les deux tests donneront la même p-valeur.
- Le test de significativité de l'ensemble des prédicteurs est également un cas particulier de ce test puisqu'il concerne le cas où l'hypothèse nulle est $\beta_1 = \ldots = \beta_p = 0$. Il est aisé de vérifier que les quantités appelées F dans chacun de ces deux tests sont en fait une seule et même quantité (donc la notation F est sans ambiguïté!).
- On a vu plus haut que les t-tests sur chaque régresseur permettaient de sélectionner les régresseurs influents. Il faut prendre garde ici au fait qu'à ces multiples tests sont associés des niveaux et que le niveau d'un grand nombre de tests n'est absolument pas maîtrisé. C'est la raison pour laquelle on effectue ce test après avoir fait un certain nombre de sélections/suppressions afin de maîtriser le niveau de confiance associé globalement à toutes nos décisions.

PREVISIONS

L'obtention de la loi des estimateurs permet également de donner l'intervalle de confiance pour la valeur prédite $\hat{y}_{new}$ lorsque les variables explicatives prennent la valeur $x_{new} = (x^0_{new}, x^1_{new}, \ldots, x^p_{new})$. Notons $\hat{y}_{new}$ la quantité :

$$\hat{y}_{new} = x_{new}\,\hat{\beta}$$

On tire facilement de ce qui précède que l'estimateur $\hat{y}_{new}$ est de loi normale, d'espérance $x_{new}\,\beta$, de variance $\sigma^2\, x_{new}'\,(X'\,X)^{-1}\,x_{new}$, d'où l'apparition, après estimation de la variance, de lois de Student :

Intervalle de confiance pour la réponse espérée[26].

Un intervalle de confiance de niveau α pour $x_{new}\,\beta$ est donné par :

$$[x_{new}\hat{\beta} - s_1(x_{new})t^{-1}_{n-p-1}(1-\alpha/2),\ x_{new}\hat{\beta} + s_1(x_{new})t^{-1}_{n-p-1}(1-\alpha/2)],$$

$$\text{où } s_1(x_{new}) = \hat{\sigma}\sqrt{x_{new}'(X'\,X)^{-1}x_{new}}.$$

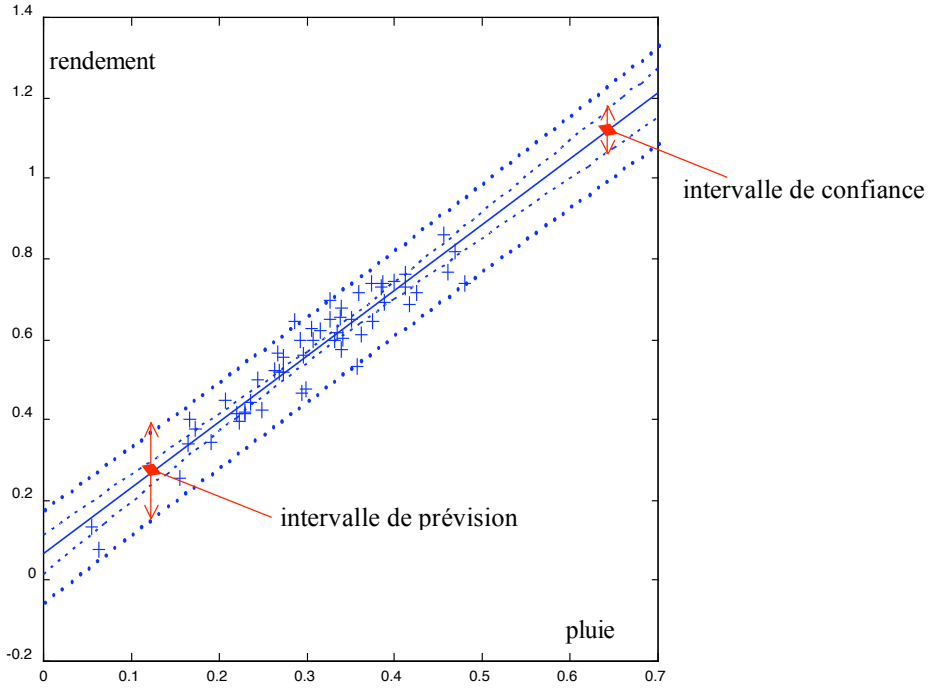
Intervalle de prévision pour la réponse.

Un intervalle de prévision de niveau α pour la réponse y_{new} lorsque $x=x_{new}$ est donné par :

$$[x_{new}\hat{\beta} - s_2(x_{new})t^{-1}_{n-p-1}(1-\alpha/2),\ x_{new}\hat{\beta} + s_2(x_{new})t^{-1}_{n-p-1}(1-\alpha/2)],$$

$$\text{où } s_2(x_{new}) = \hat{\sigma}\sqrt{1+x_{new}'(X'\,X)^{-1}x_{new}}.$$

Retour sur l'exemple 1



Intervalles de confiance à 95% pour la valeur prédite
et de prévision à 95% pour la réponse

[26] On suppose toujours qu'il y a p+1 prédicteurs, dont la constante x_0 ; remplacer dans le cas contraire p+1 par p.

Avant d'aller plus loin, donnons quelques mises en garde.


On remarquera que l'amplitude des intervalles augmente lorsque l'on s'éloigne de la moyenne des quantités de pluie observées ; ce qui est plutôt rassurant dans la mesure où on arrive rapidement dans des zones où nous n'avons pas d'expérience. Néanmoins, il faut bien se garder de faire des prévisions dans de telles zones (par exemple ici lorsque pluie>0.6) car nous n'avons **aucun** moyen de savoir si l'approximation linéaire est encore valable.


L'interprétation de l'intervalle de prévision est délicate : il ne faut surtout pas conclure que si 100 prévisions sont faites pour des valeurs différentes des prédicteurs, environ 95 parmi elles seront dans la bande en question. Cette affirmation serait valide si les coefficients du modèle étaient connus parfaitement. Or, l'incertitude sur la prévision a deux causes bien différentes : l'aléa naturel de la réponse autour de sa moyenne et l'erreur faite en estimant cette moyenne (estimation des paramètres).


L'intervalle de confiance pour la valeur prédite ne donne pas une bande de confiance dans laquelle la droite est située, comme tendrait à nous l'indiquer le graphique précédent. A la limite, un tel graphique ne devrait pas être reproduit car son interprétation n'est valable que pour une nouvelle valeur de x_{new} **fixée**. Il est possible d'obtenir une telle bande de confiance, mais il est alors nécessaire d'utiliser le domaine de confiance pour β.

Bande de confiance pour la surface de régression.

La bande de confiance de niveau α pour la surface de régression est donnée par :

$$[x\hat{\beta} - s_3(x)\, F^{-1}_{p,n-p-1}(1-\alpha),\ x\hat{\beta} + s_3(x)\, F^{-1}_{p,n-p-1}(1-\alpha)],$$

$$\text{où } s_3(x) = \hat{\sigma}\sqrt{p\, x(X'X)^{-1}x'}$$

ANALYSE DES RESIDUS – VALIDATION

C'est l'outil principal de vérification des hypothèses (modèle linéaire, indépendance, normalité, ...).

Le vecteur des erreurs est estimé par $\hat{\varepsilon} = Y - \hat{Y}$, où $\hat{Y} = X\hat{\beta}$.

Ce vecteur $\hat{\varepsilon}$ est appelé vecteur des **résidus estimés** ou des **résidus bruts**.

Première idée de validation : si $\hat{\varepsilon}_i$ est la i-ème composante de $\hat{\varepsilon}$, le tracé des $\hat{\varepsilon}_i$ doit ressembler à celui d'un bruit blanc.

On va voir sur un exemple simulé que cette idée n'est pas tout à fait la bonne. Pour cela, on construit un modèle du type suivant :

$y = 20*x+1.5+0.2*\boldsymbol{\varepsilon}$, où $\boldsymbol{\varepsilon}$ est une v.a. de loi N(0,1).

On estime les coefficients du modèle à partir de valeurs du prédicteur x et de la réponse (simulée selon l'équation précédente). Les valeurs utilisées pour x mènent à la matrice de plan d'expérience X qui suit :

$$X = \begin{bmatrix} 1 & 0.05 \\ 1 & 0.06 \\ 1 & 0.08 \\ 1 & 0.1 \\ 1 & 0.13 \\ 1 & 0.13 \\ 1 & 0.14 \\ 1 & 0.14 \\ 1 & 0.14 \\ 1 & 0.15 \\ 1 & 0.16 \\ 1 & 0.18 \\ 1 & 0.18 \\ 1 & 0.19 \\ 1 & 0.21 \\ 1 & 0.24 \\ 1 & 0.28 \\ 1 & 0.32 \\ 1 & 0.4 \end{bmatrix}$$

Dans ce cadre simulé, il est tout à fait clair que les résidus doivent se comporter comme la réalisation de 19 v.a. indépendantes de loi N(0,1).

Afin de vérifier si c'est effectivement le cas, on effectue un certain nombre de simulations (donc de régressions) et on trace le graphique des résidus obtenus au cours de ces n régressions. Voici la figure que l'on obtient pour une répétition de 20 régressions :

On voit clairement ici que les résidus estimés ne sont absolûment pas de variance constante. En conclusion :

Si les hypothèses du modèle linéaire sont valables, les résidus estimés ne sont pas un bruit blanc puisque de variance non constante !

Ce résultat peut sembler paradoxal, mais un peu de réflexion suffit pour se convaincre que, dans le cas de notre exemple, les observations qui correspondent aux valeurs extrêmes du prédicteur sont "avantagées" par la régression, i.e. que leurs résidus sont plus petits. Nous reviendrons sur ces questions à propos des observations dites influentes.

On peut en fait calculer explicitement la variance des résidus estimés :

Matrice chapeau

La **matrice chapeau** H (H comme *hat*) est la matrice qui "met le chapeau" sur Y :

$$\hat{Y} = H\,Y \text{ et } \hat{\varepsilon} = (Id - H)\,Y,$$

$$\text{avec } H = X\,(X'\,X)^{-1}\,X'$$

N.B.: H est la matrice de la projection orthogonale sur l'espace des prédicteurs v(X) et (Id - H) est celle de la projection orthogonale sur $v(X)^{\perp}$.

Loi des résidus bruts

$\hat{\varepsilon}$ est de loi normale $N(0, \sigma^2 (Id - H))$.

On notera que ce résultat avait été partiellement établi lors de la démonstration du théorème de Gauss Markov. La preuve ici est particulièrement aisée puisqu'il suffit de connaître la loi de l'image d'un vecteur gaussien par une application linéaire. Il est en tout cas clairement établi que la variance de $\hat{\varepsilon}_i$ est non constante, égale à $\sigma^2 (1 - h_{ii})$, d'où la :

Définition.

$T_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$ est appelé **résidu standardisé**.

Remarque

Le procédé de standardisation des résidus permet d'obtenir des variables de même loi. Par contre, les résidus standardisés ne sont pas des variables indépendantes, même dans le cas où σ est connu : cela découle simplement de la proposition précédente. On peut d'ailleurs s'en apercevoir directement en remarquant par exemple que, du fait de la prise en compte du prédicteur constant x_0, la somme des résidus bruts est nulle. Cependant, dès que le nombre n d'observations devient grand devant le nombre p (ou p+1) de paramètres, cette dépendance peut être considérée comme négligeable.

Exemple 2 (cf. [Antoniadis])

Le temps nécessaire (variable temps en mn) pour approvisionner un parc de distributeurs de boissons est a priori fonction de deux variables explicatives : le nombre de bouteilles à charger (variable nb) et la distance à parcourir pour la personne (variable dist en m). On effectue 25 observations qui sont données ci-après. Des analyses graphiques élémentaires mène au modèle plausible suivant :

$$temps = \beta_0 + \beta_1\, nb + \beta_2\, dist + \varepsilon$$

On obtient les estimations :

$\hat{\beta}' = [2.353 \;\; 1.6159 \;\; 0.01437]$
$\hat{\sigma} = 3.255$

nb	dist	temps
7	560	16.7
3	220	11.5
3	340	12.0
4	80	14.9
6	150	13.8
7	330	18.1
2	110	8.0
7	210	17.8
30	1460	79.2
5	605	21.5
16	688	40.3
10	215	21.0
4	255	13.5
6	462	19.8
9	448	24.0
10	776	29.0
6	200	15.4
7	132	19.0
3	36	9.5
17	770	35.1
10	140	17.9
26	810	52.3
9	450	18.8
8	635	19.8
4	150	10.7

La table d'analyse de variance obtenue est la suivante :

Table d'analyse de variance

Source	df	SS	MS	F	p
Regression	2	5544	2772	261.7	4.441e-016
Error	22	233	10.59		
Total	24	5777			

Coefficients

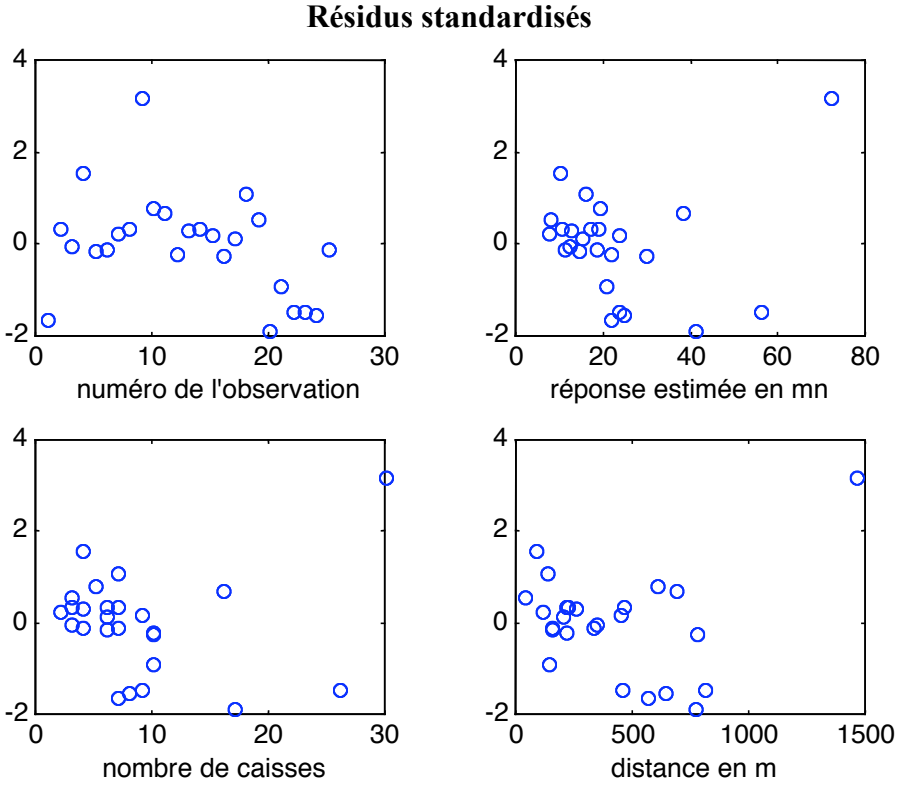
Root MSE	3.255	R-square	0.9597
		R-sq(adj)	0.956

Paramètres estimés

Predictor	Coeff	Stdev	t-ratio	p
intercept	2.353	1.095	2.149	0.04292
nb	1.615	0.1705	9.474	3.199e-009
dist	0.01437	0.003608	3.984	0.0006273

Les prédicteurs semblent tous très significatifs et on va chercher à voir si le modèle construit peut être validé.

La première chose à faire est de réaliser divers tracés des résidus standardisés : tracés des T_i séquentiels ou "contre" les variables explicatives ou encore contre la réponse estimée (rappel : $\hat{Y}$ et $\hat{\varepsilon}$ sont indépendantes) : on s'attend au tracé un bruit.



On observe une mauvaise répartition des résidus (notamment les graphiques des résidus contre les prédicteurs nombre et distance) et l'observation n°9 semble suspecte ; est-elle aberrante?

Problème :

Pour quantifier l'écart de T_i à la normale, il faut connaître sa loi. Or, $\hat{\varepsilon}_i$ est de loi normale et $(n-p-1)\hat{\sigma}^2/\sigma^2$ de loi du χ^2 mais ces deux variables ne sont pas indépendantes.

➔ nécessité d'estimer $\hat{\sigma}^2$ sans la i-ème donnée.

On note $X_{(i)}$, $\hat{\beta}_{(i)}$, $\hat{\sigma}_{(i)}$, ... les quantités X, $\hat{\beta}$, $\hat{\sigma}$ lorsque la i-ème observation est supprimée.

Résidus studentisés

$T_i^* = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$ est appelé **résidu studentisé**.

Introduire une telle quantité est sans doute théoriquement satisfaisant, mais il faut avoir conscience que l'évaluation des résidus studentisés nécessite a priori la calcul de n régressions linéaires supplémentaires ! Heureusement, il n'est pas nécessaire de recommencer tout le travail à chaque fois comme le montre le résultat qui suit :

Estimation d'une régression en ôtant une observation

(i) En notant x_i la i-ème ligne de la matrice X, on a les relations :

$$\hat{\beta}_{(i)} = \hat{\beta} - (X' X)^{-1} x_i' \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

$$(n-p-2)\,\hat{\sigma}_{(i)}^2 = (n-p-1)\,\hat{\sigma}^2 - \frac{\hat{\varepsilon}_i^2}{1 - h_{ii}}$$

$$T_i^* = T_i \sqrt{\frac{n-p-2}{n-p-1-T_i^2}}$$

(ii) T_i^* est de loi de Student t_{n-p-2}.

(iii) T_i^* est le résidu standardisé correspondant à la prévision pour Y_i faite sans la connaissance de la i-ème donnée.

Remarque

Du fait du point (iii) de la proposition précédente, T_i^* est également appelé **résidu par validation croisée**.

Tracé séquentiel des résidus studentisés avec bande de confiance à 95%

↪ Le résidu correspondant à l'observation n° 9 est aberrant.

On se pose alors des questions sur cette observation. Un retour sur les données montre qu'elle correspond à un domaine pour la variable dist non couvert par les autres observations (dist pour cette observation est à peu près le double des plus grandes valeurs observées pour les autres observations). On est sans doute en présence d'un domaine où le modèle linéaire construit n'est pas valable. Pour voir ce qu'il en est, on examine les résultats obtenus en enlevant l'observation n° 9.

En ôtant cette observation, on obtient avec 24 observations les estimations suivantes :

$\hat{\beta}' = [4.4472 \;\; 1.4977 \;\; 0.0103]$
$\hat{\sigma} = 2.3741$

On remarque que les estimations sont grandement modifiées : l'observation n° 9 était à la fois aberrante et influente.

Le tracé des résidus studentisés est alors le suivant :

Résidus studentisés en ôtant l'observation n°9

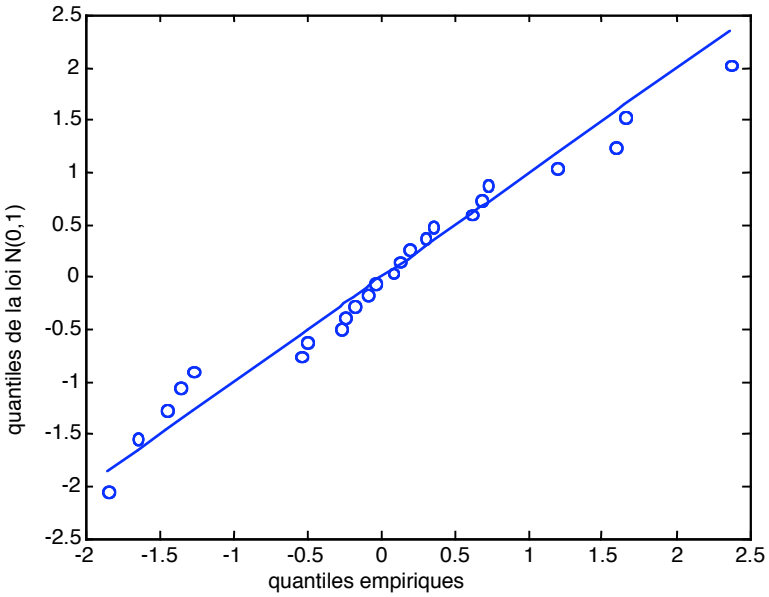
Les résidus semblent corrects. On peut alors passer à un test plus formel. Déjà, on remarque que les résidus studentisés doivent suivre[27], si le modèle est valable, une loi de

[27] On rappelle que l'on néglige les corrélations entre les résidus. Par suite, on considère que les résidus studentisés observés constituent un échantillon de taille n d'une loi de Student t_{20}.

Student t_{20}. Or, la différence entre cette loi et la loi normale N(0, 1) est très peu sensible, comme le montre le graphique suivant.

Par suite, on va chercher à tester si les résidus studentisés peuvent être considérés comme un échantillon de taille 24, provenant d'une loi normale N(0,1).

La première chose à faire est de tracer la droite de Henri. On supperpose au graphique la première bissectrice car, contrairement aux cas rencontrés en général en statistiques, les paramètres de la loi normale n'ont pas ici à être estimés : l'espérance doit être nulle et la variance égale à 1.



Droite de Henri des résidus studentisés

A première vue, le graphique paraît satisfaisant. On envisage alors un test, en l'occurrence le test de Kolmogorov d'adéquation à la loi N(0,1).

La valeur de la statistique D_n est donnée par : $D_n = 0.1071$. La lecture des tables (pour n = 24) montre que cette valeur est inférieure à tous les seuils (pour 80%, le seuil est de 0.21205). On valide donc notre modèle.

OBSERVATIONS INFLUENTES ET ABERRANTES

On a vu précédemment comment détecter une observation dite aberrante à partir du tracé des résidus studentisés. On a vu également que dans le cas de l'exemple 2, l'observation n°9 – considérée comme aberrante - était de plus très influente puisque sa suppression changeait radicalement l'estimation du modèle. Il convient de bien faire la part des choses entre ces deux notions :

- Une observation est dite **aberrante** si elle n'est pas en accord avec le modèle ajusté.
- Une observation est dite **influente** si elle influence fortement l'ajustement du modèle.

On a vu plus haut comment détecter des observations aberrantes. Pour ce qui concerne les observations influentes, nous allons donner deux outils de détection courants : le levier et les distances de Cook.

Avant d'aller plus loin, on gardera en tête qu'une détection de valeurs aberrantes **ne doit pas être faite sans une étude simultanée sur l'influence des observations**.

L'exemple fictif qui suit nous permettra de nous en convaincre :

On fait une régression d'une réponse y en fonction d'un prédicteur à l'aide de 6 observations :

$$Y = \begin{bmatrix} 1.1 \\ 1.4 \\ 2.2 \\ 2.6 \\ 2.8 \\ 7.1 \end{bmatrix} \quad \text{et } X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 9 \end{bmatrix}$$

Le résultat de l'ajustement est fourni par le graphique qui suit :

On voit que le résidu de l'observation correspondant à une valeur de 5 pour x est grand et risque d'être détecté comme valeur aberrante[28]. Cependant, il apparaît également assez clairement que l'observation obtenue pour x = 9 n'est pas en accord avec le modèle. C'est donc plutôt cette observation qu'il faudrait détecter et éliminer.

Le premier outil, le **levier** (en anglais *leverage*), est en rapport avec la matrice chapeau H. On a vu plus haut que cette matrice est la matrice de projection orthogonale sur l'espace v(X) engendré par les prédicteurs. Par suite, la réponse estimée pour la i-ème observation est donnée par :

$$\hat{y}_i = \sum_{i=1}^{n} h_{ij}\, y_j$$

Par suite, le coefficient h_{ii} – appelé **levier** - apparaît comme une mesure de l'influence de la i-ème observation sur sa propre prédiction. On comprend donc que lorsque ce terme est grand, l'observation en question peut être considérée comme influente. On notera cependant que cet indicateur ne dépend pas des réponses observées mais seulement des prédicteurs.

Pour préciser ce que signifie "grand", on utilise les identités suivantes, qui découlent du fait que H est une matrice de projection orthogonale :

$$\forall\ i,\ 1/n \leq h_{ii} \leq 1$$

$$\sum_{i=1}^{n} h_{ii} = p+1$$

Dans le cas de l'exemple fictif qui précède, on obtient la figure suivante :

On voit que le levier de la sixième observation est fort ; il faut donc regarder tout particulièrement la réponse observée pour cette observation.

Dans le cas de l'exemple 2, le graphique est reproduit ci-après.

[28] Ici, un tracé des résidus studentisés détecterait néanmoins les deux observations correspondant à x = 5 et x = 9.

Leviers obtenus pour l'exemple des distributeurs de boissons

On voit ici que deux observations se distinguent : l'observation n°9 et l'observation n°22. Il semble que l'observation n°22 soit influente mais non aberrante. Un examen rapide du tableau de données montre que cette observation correspond à des valeurs importantes pour les prédicteurs. Sans information supplémentaire, on peut qualifier cette observation de potentiellement suspecte.

Un deuxième outil de détection essaye à la fois de mesurer plus globalement l'influence de la i-ème observation sur la régression et de tenir compte des réponses observées. L'idée de base est de mesurer l'effet de la suppression de la i-ème observation sur l'ensemble des résultats obtenus.

Si $\hat{Y}_{(i)}$ est la réponse estimée pour les n observations à partir de toutes les observations, sauf la i-ème, on pose :

Distance de Cook

La distance de Cook D_i correspondant à la i-ème observation est définie par :

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})' (\hat{Y} - \hat{Y}_{(i)})}{(p+1)\,\hat{\sigma}^2}$$

Notons que l'expression définissant la distance de Cook est a priori la plus intuitive (écart quadratique) mais se prête assez mal au calcul. En pratique, on utilise l'expression équivalente suivante :

$$D_i = \frac{T_i^2}{p+1} \frac{h_{ii}}{1-h_{ii}}$$

On remarquera en outre que cette nouvelle expression fait apparaître la distance de Cook sous un jour nouveau : une observation considérée comme influente par la distance de

Cook est une observation pour laquelle à la fois le résidu standardisé et le levier sont importants.

La grandeur à partir de laquelle une observation est considérée comme influente n'est pas facile à déterminer : le seuil varie selon les auteurs. En pratique, l'ordre de grandeur à partir duquel on considérera une observation influente est de l'ordre de l'unité. En effet, il est très courant de rencontrer des valeurs beaucoup plus grandes : un seuil très précis n'est donc pas nécessaire.

Si on reprend l'exemple 2 des distributeurs de boissons, les distances de Cook obtenues sont les suivantes :

On voit que l'observation n°9 se distingue très nettement !

Terminons ce cours par des considérations essentiellement pratiques (en anglais *practicalities*).

Practicalities

Nous reprenons ici les diverses notions et techniques abordées, dans une optique d'application immédiate.

PLAN D'ETUDE D'UN PROBLEME DE REGRESSION

Traiter de cette question est un exercice très délicat ; tellement délicat même que très peu de livres se risquent à l'aborder. La raison en est simple : tout schéma d'étude d'un problème de régression se verra contredit par un exemple bien (mal ?) choisi. Considérant que le mieux est l'ennemi du bien, nous allons néanmoins tenter de décrire une méthodologie qui, bien qu'étant insuffisante dans certains cas, permet au moins de guider la démarche du débutant.

Nous distinguerons 3 étapes : statistique descriptive, statistique inférentielle, prévisions. On notera que la première étape est en général largement sous-traitée, alors qu'il s'agit en fait de la plus importante qui va conditionner toute la suite.

1. Statistique descriptive

- **Représentations monodimensionnelles**

Sous forme d'histogramme ou de lissage par noyau, elles permettent de saisir très rapidement la dispersion de chaque variable, ainsi que la présence de données "extrêmes" qui sont susceptibles d'influencer, voire de biaiser, l'analyse.

- **Représentations bidimensionnelles**

L'observation de corrélations entre les prédicteurs permet de savoir a priori si l'analyse inférentielle ultérieure peut être faussée. Elle peut être complétée par une analyse de la matrice des corrélations. En cas de corrélation forte, une ACP peut être mise en œuvre afin de travailler sur des prédicteurs décorrélés, mais on perd alors le sens de chacun des prédicteurs.. La considération du contexte peut également aider. Par exemple, si deux prédicteurs sont le chiffre d'affaires et le nombre d'employés d'une société, on peut définir deux prédicteurs a priori moins corrélés : le CA par employé et le nombre d'employés.

L'observation de l'influence de chaque prédicteur sur la réponse permet déjà d'avoir une idée d'un modèle (droite, parabole…). Elle permet également de voir si l'on est a priori dans le bon cadre inférentiel (résidus centrés de variance constante). Si la variance ne semble pas constante, on peut transformer la réponse (transformations de Cox i.e. Log ou puissance vues en séries temporelles) ou modifier le modèle, par exemple : $y = ax+b+\sqrt{x}\,\varepsilon$, où ε est de variance constante. On notera que cette dernière approche revient finalement à modifier réponse et prédicteurs (diviser la relation précédente par $\sqrt{x}$).

- **Examen des interactions**

Elle permet de deviner a priori si l'on doit ajouter des termes d'interaction dans le modèle. On rappelle que pour cela, il faut essayer de tracer la réponse y en fonction d'un prédicteur x_1 lorsque un autre prédicteur x_2 est fixé. Compte tenu du caractère discret des données, on regroupera les données en classes selon les valeurs du prédicteur x_2. Dans un

premier temps, on peut simplement découper les données en deux classes : les points pour lesquels x_2 est faible et ceux pour lesquels x_2 est fort. On représente ensuite sur un même graphique les deux nuages de points avec x_1 en abscisse et y en ordonnée.

Tout ceci aboutit à une première proposition de modèle

2. Statistique inférentielle

- **Estimation des paramètres du modèle**

  Très simple, et donnée par tout logiciel.

- **Corrélations des paramètres estimés $\hat{\beta}$**

  Cette étude permet de vérifier que les estimations sont les plus indépendantes possibles. Si ce n'est pas le cas, les tables d'ANOVA qui suivent seront faussées.

- **ANOVA**

  Les tables d'ANOVA visent essentiellement à sélectionner les prédicteurs influents. La comparaison de modèles concurrents peut alors s'effectuer à l'aide d'un test de Fisher pour des modèles emboîtés ou l'examen de R^2_{adj} et $\hat{\sigma}$ s'ils ne le sont pas.

- **Examen des résidus**

  Les divers tracés de résidus permettent de s'assurer que l'on n'a pas oublié de composante importante dans le modèle. On vérifie ensuite que les hypothèses faites sur la distribution des résidus pour mener à bien l'ANOVA (normalité notamment) sont correctes. D'autres tracés (levier, distance de Cook…) sont consacrés à la mise en évidence de données aberrantes et/ou influentes (voir cours).

3. Prévisions

Elles s'effectuent in fine, lorsque le modèle est validé. On renvoie là encore au cours pour les résultats nécessaires.

Terminons ces considérations par une étude de cas… en exercice.

ETUDE DE CAS – LES PLUIES EN CALIFORNIE

Ces données proviennent de [Mendenhall], p. 714.

Il s'agit de mettre en relation la quantité de pluie tombée dans diverses villes de Californie et diverses caractéristiques de la position de la ville dans le pays.

Le recueil des données dans 30 villes californiennes donne le tableau suivant, où :

- Pluie désigne la quantité de pluie tombée en 1980. Unité = inch
- Altitude est l'altitude de la ville considérée. Unité = foot
- Latitude est la latitude de la ville considérée. Unité = degré
- Distance est la distance de la ville à la mer. Unité = miles

Ville	Pluie	Altitude	Latitude	Distance
Eureka	39.57	43	40.8	1
Red Bluff	23.27	41	40.2	97
Thermal	18.2	4152	33.8	70
Fort Bragg	37.48	74	39.4	1
Soda Springs	49.26	6752	39.3	150
San Francisco	21.82	52	37.8	5
Sacramento	18.07	25	38.5	80
San Jose	14.17	95	37.4	28
Giant Forest	42.63	6360	36.6	145
Salinas	13.85	74	36.7	12
Fresno	9.44	331	36.7	114
Pt. Piedras	19.33	57	35.7	1
Paso Robles	15.67	740	35.7	31
Bakersfield	6	489	35.4	75
Bishop	5.73	4108	37.3	198
Mineral	47.82	4850	40.4	142
Santa Barbara	17.95	120	34.4	1
Susanville	18.2	4152	40.3	198
Tule Lake	10.03	4036	41.9	140
Needles	4.63	913	34.8	192
Burbank	14.74	699	34.2	47
Los Angeles	15.02	312	34.1	16
Long Beach	12.36	50	33.8	12
Stockton	8.26	125	37.8	74
Blythe	4.05	268	33.6	155
San Diego	9.94	19	32.7	5
Daggett	4.25	2105	34.1	85
Death Valley	1.66	-178	36.5	194
Crescent City	74.87	35	41.7	1
Colusa	15.95	60	39.2	91

NB : l'étude peut se mener très simplement ici, mais la considération d'une variable géographique, et donc l'étude d'une carte, risque de chambouler fortement les résultats !

Index

Bibliographie

[Antoniadis]
Antoniadis A., Berruyer J., Carmona R. : *Régression non linéaire et applications*, Economica (1992).

[Bates]
Bates D.M., Watts D.G. : *Nonlinear regression analysis and its applications*, Wiley (1998).

[Bay]
Bay X. : *Conditionnement en probabilités et processus stochastiques*, Cours de 2ème année de l'ENSM.SE (2004).

[Benoist]
Benoist D., Tourbier Y., Germain-Tourbier S. : *Plans d'expériences : construction et analyse*, Technip (1994).

[Chen]
Chen C.H., Li K.C. : *Can SIR be as popular as multiple linear regression ?*, Statistica Sinica **8**, p. 289-316 (1998).

[Draper]
Draper N., Smith H. : *Applied regression analysis, second edition*, Wiley (1981).

[Jobson]
Jobson J. D.: *Applied multivariate data analysis, vol. 1 : regression and experimental design*, Springer Verlag (1991).

[Hastie]
Hastie. T.J., Tibshirani R.J. : *Generalized additive models*, Chapman and Hall (1991).

[McCullagh]
Mc Cullagh P., Nelder J.A. : *Generalized linear models*, Chapman and Hall (1989).

[Mendenhall]
W. Mendenhall, T. Sincich, *A second course in statistics – regression analysis*, Prenctice Hall (1996).

[Roustant]
Roustant O. : *Introduction aux séries chronologiques*, cours de 2ème année de l'ENSM.SE (2004).