Université Paris I, Panthéon - Sorbonne

MASTER M.A.E.F.

Cours d'Econométrie II

JEAN-MARC BARDET (UNIVERSITÉ PARIS 1, SAMM)

U.F.R. 27 et SAMM (Statistique, Analyse et Modélisation Multidisciplinaire)
Université Panthéon-Sorbonne, 90 rue de Tolbiac, 75013 Paris.

Plan du cours

1. Asymptotique du modèle linéaire
2. Sélection des variables exogènes
3. Changement de structure et modèles linéaires avec bruit corrélé ou non stationnaire
4. Modèle linéaire généralisé
5. Modèle non linéaire semi-paramétrique
6. Régression non-paramétrique

References

[1] Amemiya, T. (1985). *Advanced Econometrics*. Cambridge , MA: Harvard University Press.

[2] Barbe P. et Ledoux M. (1998) *Probabilité*. EDP Sciences.

[3] Davidson, R. et MacKinnon, J.G. (1993). Estimation and inference in econometrics. Oxford University Press.

[4] Guyon, X. (1995). Statistique et économétrie - Du modèle linéaire aux modèles non-linéaires. Ellipses.

Documents accessibles librement sur internet

- Aide-mémoire en économétrie de A. Trognon et J.M. Fournier à l'ENSAE: `http://www.ensae.fr/ParisTech/SEC02/ENSAEEconometrieCursusintegre2006.pdf`.
- Cours de R. Bourdonnais: `http://www.dauphine.fr/eurisco/eur-wp/CoursSeriesTemp-Chap*.pdf` où on peut remplacer $*$ par $1, 2, 3$ ou 4.
- Site de Ricco Rakotomalala: `http://eric.univ-lyon2.fr/~ricco/cours/index.html`
- Livre de Davidson-MacKinnon sur le site de R. Davidson: `http://russell.vcharite.univ-mrs.fr/EIE/`

Quelques sites internet intéressants

- Le site de Toulouse III: `http://wwww.lsp.ups-tlse.fr`. Regarder les documents pédagogiques.
- Le site de Paris V: `http://wwww.math-info.univ-paris5.fr`. Regarder les documents pédagogiques.
- Le site de Paris VI: `http://wwww.proba.jussieu.fr`. Regarder les documents pédagogiques.
- Le site du programme STAFAV: `http://www.math.u-psud.fr/~stafav`
- Le site de la S.F.D.S.: `http://www.sfds.fr`.
- Le site de la S.M.A.I.: `http://smai.emath.fr`. Regarder la rubrique `Logiciels` dans laquelle de nombreux logiciels de mathématiques peuvent être téléchargés (en particulier, `Scilab` et `Mupad`).
- Le site français d'où l'on peut télécharger le logiciel R: `http://cran.cict.fr`.

Introduction

1 Propriétés asymptotiques des statistiques du modèle linéaire

2 Critères de sélection de variables

3 Changement de modèle, hétéroscédasticité, erreurs corrélées et estimation par moindres carrés généralisés

3.1 Changement de structure

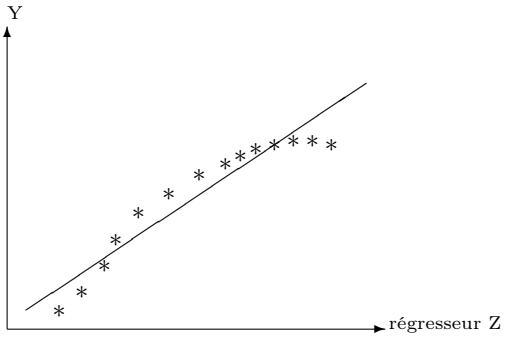
Ici on considère que le choix du modèle linéaire inialement posé n'est pas convenable, à savoir que $\mathbb{E}\widehat{Y} \neq X\theta$, donc que le modèle n'est pas $Y = X\theta + \varepsilon$. Nous allons voir tout d'abord comment détecter un tel diagnostic, puis comment le résoudre.

Diagnostic d'une nécessité de changement de structure

Essentiellement le graphe $(\widehat{Y}_i, \widehat{\varepsilon}_i)$.

Changement continu de modèle

En régression linéaire simple, la confrontation graphique entre le nuage de points (z_i, y_i) et la droite de régression de Y par Z par moindres carrés ordinaires donne une information quasi exhaustive. En voici un exemple :



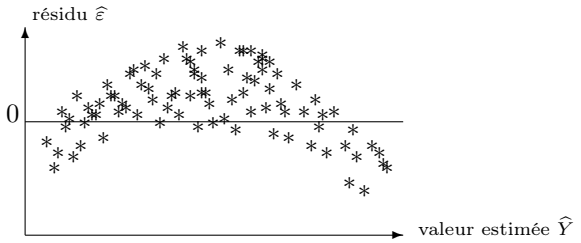
Sur ce graphique, on voit une courbure de la "vraie" courbe de régression de Y et on peut penser que le modèle est inadéquat et que le premier postulat **P1** n'est pas vérifié.

• Dans le cas de la régression multiple, ce type de graphique n'est pas utilisable car il y a plusieurs régresseurs. Les différents postulats sont à vérifier sur les termes d'erreur ε_i qui sont malheureusement inobservables. On utilise leurs prédicteurs naturels, les résidus : $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$. Par exemple, pour le modèle général de régression,

$$Y_i = \mu + \beta_1 Z_i^{(1)} + \cdots + \beta_p Z_i^{(p)} + \varepsilon_i, \quad \text{pour} \;\; i = 1, \cdots, n,$$
$$\Longrightarrow \quad \widehat{\varepsilon}_i = Y_i - \widehat{\mu} - \widehat{\beta}_1 Z_i^{(1)} - \widehat{\beta}_2 Z_i^{(2)} - \cdots - \widehat{\beta}_p Z_i^{(p)} \quad \text{pour} \;\; i = 1, \cdots, n.$$

Le graphique le plus classique consiste à représenter les résidus $(\widehat{\varepsilon}_i)_i$ en fonction des valeurs prédites $(\widehat{Y}_i)_i$. Ce graphique doit être fait pratiquement systématiquement. Cela revient encore à tracer les coordonnées du vecteur $P_{[X]^\perp}.Y$ en fonction de celles de $P_{[X]}.Y$. L'intérêt d'un tel graphe réside dans le fait que si les quatre postulats **P1-4** sont bien respectés, il y a indépendance entre ces deux vecteurs qui sont centrés et gaussiens (d'après le Théorème de Cochran). Cependant, à partir de ce graphe, on ne pourra s'apercevoir que de la possible déficience des postulats **P1** et **P2**, les deux autres postulats pouvant être "contrôlés"

par d'autres représentations graphiques (voir plus loin). Concrètement, si on ne voit rien de notable sur le graphique (c'est-à-dire que l'on observe un nuage de points centré et aligné quelconque), c'est très bon signe : les résidus ne semblent alors n'avoir aucune propriété intéressante et c'est bien ce que l'on demande à l'erreur. Voyons justement maintenant un graphe résidus/valeurs prédites "pathologique" :



Dans ce cas on peut penser que le modèle n'est pas adapté aux données. En effet, il ne semble pas y avoir indépendance entre les $\widehat{\varepsilon}_i$ et les $\widehat{Y}_i$ (puisque, par exemple, les $\widehat{\varepsilon}_i$ ont tendance à croître lorsque les $\widehat{Y}_i$ sont dans un certain intervalle et croissent). Il faut donc améliorer l'analyse du problème pour proposer d'autres régresseurs pertinents, ou transformer les régresseurs $Z^{(i)}$ par une fonction de type $(\log, \sin)$, ce que l'on peut faire sans précautions particulières.

• On peut librement transformer les régresseurs $Z^{(1)}, \cdots, Z^{(p)}$ par toutes les transformations algébriques ou analytiques connues (fonctions puissances, exponentielles, logarithmiques,...), pourvu que le nouveau modèle reste interprétable. Cela peut permettre d'améliorer l'adéquation du modèle ou de diminuer son nombre de termes si on utilise ensuite une procédure de choix de modèles.

Changement discontinu de modèle

- **Test de Chow**
  Le test de Chow permet prendre en compte un éventuel changement de structure dans l'écriture du modèle (en cela ce test porte plutôt pour des modèles d'évolution temporelle). Il s'agit donc de tester:

$$H_0 : \; Y = X\,\theta + \varepsilon \qquad \text{contre} \qquad H_1 : \; Y^{(1)} = X^{(1)}\,\theta_1 + \varepsilon^{(1)} \quad \text{et} \quad Y^{(2)} = X^{(2)}\,\theta_2 + \varepsilon^{(1)}.$$

  On voit que sous H_0 on a un sous-modèle de H_1, alors que le modèle sous H_1 peut aussi s'écrire comme un modèle linéaire avec $Y = (Y^{(1)'}, Y^{(2)'})'$, $Z = (X^{(1)}, X^{(2)})$ et $\theta' = (\theta_1', \theta_2'')$. On peut donc définir un test de Fisher de sous-modèle et on a, sous H_0:

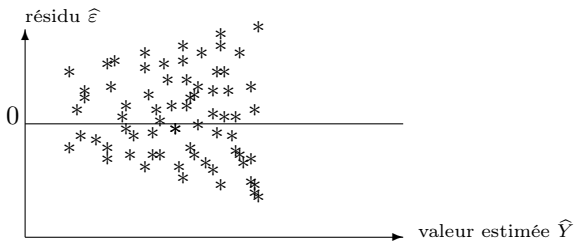
$$\widehat{F} = \frac{\frac{1}{k}}{\frac{1}{n-2k}} \frac{SC_0 - SC_1}{SC_1} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{k}\,\chi^2(k).$$

- **Test et détection de ruptures**

3.2 Hétéroscédasticité

Diagnostic d'héteroscédasticité

On représente également les $\widehat{Y}_i$ en fonction des $\widehat{\varepsilon}_i$. Par exemple :



Nature de la relation	Domaine pour Y	Transformation
$\sigma = (\text{cte})Y^k$, $k \neq 1$	$\mathbb{R}_+^*$	$Y \mapsto Y^{1-k}$
$\sigma = (\text{cte})\sqrt{Y}$	$\mathbb{R}_+^*$	$Y \mapsto \sqrt{Y}$
$\sigma = (\text{cte})Y$	$\mathbb{R}_+^*$	$Y \mapsto \log Y$
$\sigma = (\text{cte})Y^2$	$\mathbb{R}_+^*$	$Y \mapsto Y^{-1}$
$\sigma = (\text{cte})\sqrt{Y(1-Y)}$	$[0,1]$	$Y \mapsto \arcsin\left(\sqrt{Y}\right)$
$\sigma = (\text{cte})\sqrt{1-Y}.Y^{-1}$	$[0,1]$	$Y \mapsto (1-Y)^{1/2} - 1/3(1-Y)^{3/2}$
$\sigma = (\text{cte})(1-Y^2)^{-2}$	$[-1,1]$	$Y \mapsto \log(1+Y) - \log(1-Y)$

Table 1: Table des changements de variable pour la variable à expliquer

Dans ce cas la variance des résidus semble inhomogène, puisque les $\widehat{\varepsilon}_i$ ont une dispersion de plus en plus importante au fur et à mesure que les $\widehat{Y}_i$ croissent. Un changement de variable pour Y pourrait être une solution envisageable pour "rendre" constante la variance du bruit (voir un peu plus bas).

Remarque : certaines options sophistiquées utilisent plutôt des résidus réduits (Studentised residuals) qui sont ces mêmes résidus divisés par un estimateur de leur écart-type (généralement l'écart-type empirique) : cela donne une information supplémentaire sur la distribution des résidus qui doit suivre alors (toujours sous les postulats **P1-4**) une loi de Student. Cependant, on perd en capacité d'interprétation car le résidu est "adimensionnel", il n'est plus exprimé dans les unités de départ. Supposons par exemple que l'on veuille modéliser la taille (stature) d'adultes mesurée en mm. Un résidu de 5 correspond a une erreur de $5mm$ ce qui est tout-à-fait négligeable en pratique. Un résidu réduit est le plus souvent entre -2 et 2 (domaine de variation de la loi normale) sa valeur n'est pas directement interprétable.

Modifications possibles à apporter au modèle :

On ne peut envisager de transformer Y, que si les graphiques font suspecter une hétéroscédasticité. Dans ce cas, cette transformation doit obéir à des règles précises basées sur la relation suspectée entre l'écart-type résiduel σ et la réponse Y : c'est ce que précise le Tableau 3.2. Souvent ces situations correspondent à des modèles précis. Par exemple, la cinquième transformation correspond le plus souvent à des données de comptage. Dans le cas où les effectifs observés sont faibles (de l'ordre de la dizaine), on aura plutôt intérêt à utiliser un modèle plus précis basé sur des lois binomiales. Il s'agit alors d'un modèle linéaire généralisé. D'ailleurs toutes les situations issues d'une des transformations ci-dessus peuvent être traitées par modèle linéaire généralisé. Il n'entre pas dans le champ de ce cours de préciser ces modèles (on pourra consulter par exemple le livre de McCullagh et Nelder).

Notons cependant que pour des grands échantillons la transformation de Y peut suffire à transformer le modèle en un modèle linéaire classique et est beaucoup plus simple à mettre en œuvre. Par exemple, dans une étude bactériologique sur des désinfectants dentaires, on a mesuré le degré d'infection d'une racine dentaire en comptant les germes au microscope électronique. Sur les dents infectées, le nombre de germes est élevé et variable. L'écart-type est proportionnel à la racine carrée de la réponse. Une loi ayant cette propriété est la loi de Poisson, qui donne alors lieu à un modèle linéaire généralisé. Toutefois, si les décomptes sont en nombre important, travailler directement avec pour donnée la racine carrée du nombre de germe peut répondre tout aussi bien à la question.

Transformation de Box-Cox Lorsque les Y_i sont des variables positives, on peut également utiliser une transformation continue de la variable Y sous la forme suivante:

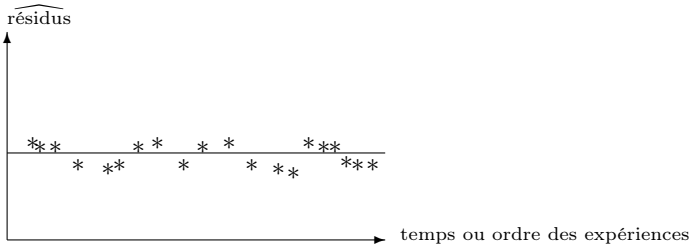
$$\tau(Y_i, \lambda) := \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{pour } \lambda \neq 0 \\ \log Y_i & \text{pour } \lambda = 0 \end{cases}$$

où λ est un réel a priori inconnu.
Numériquement, on peut à partir d'une grille de valeurs de λ calculer la variance des résidus $\widehat{\sigma}^2_{\widehat{\varepsilon},\lambda}$ pour chaque valeur de λ. On choisira alors:

$$\widehat{\lambda} = \text{Argmin}_{\lambda \in [\lambda_{\max}, \lambda_{\max}]} \widehat{\sigma}^2_{\widehat{\varepsilon},\lambda}.$$

Diagnostic de non-indépendance du bruit

Un graphe pertinent pour s'assurer de l'indépendance des résidus entre eux et celui des résidus estimés $\widehat{\varepsilon}_i$ en fonction de l'ordre des données (lorsque celui-ci a un sens, en particulier s'il représente le temps). Par exemple, on peut obtenir le graphe suivant :



Un graphique comme celui ci-dessus est potentiellement suspect car les résidus ont tendance à rester par paquets lorsqu'ils se trouvent d'un côté ou de l'autre de 0. On pourra confirmer ces doutes en faisant un test de runs. Ce test est basé sur le nombre de runs, c'est-à-dire sur le nombre de paquets de résidus consécutifs de même signe. Sur le graphique ci-dessus, il y a 8 runs. On trouve les références de ce test dans tout ouvrage de tests non-paramétriques ou dans un livre comme celui de Draper et Smith. Voir également un exercice.

Par ailleurs, si les erreurs sont corrélées suivant certaines conditions (par exemple si ce sont des processus ARMA), il est tout d'abord possible d'obtenir encore des résultats quand à l'estimation des paramètres, mais il existe également des méthodes de correction (on peut penser par exemple à des estimations par moindres carrés généralisés ou pseudo-généralisés; voir par exemple les livres d'Amemiya, Green, Guyon ou Jobson).

Définition des MCG

On suppose le modèle linéaire général

$$Y = X\,\theta + \varepsilon, \qquad \text{avec} \qquad \mathbb{E}\varepsilon = 0. \tag{1}$$

On suppose également connue la matrice de variance-covariance des erreurs (qui n'est donc pas forcément diagonale), qui sera notée $\Sigma = \mathbb{E}\big[\varepsilon\varepsilon'\big]$, que l'on suppose de rang n (X est supposée de rang p). On considère maintenant, en lieu et place de la distance euclidienne classique dans $\mathbb{R}^n$ utilisée pour les moindres carrés ordinaires, la distance définie par la norme :

$$\|U - V\|_\Sigma^2 = (U - V)'\,\Sigma^{-1}\,(U - V).$$

Cette distance est donc associée à un produit scalaire dans $\mathbb{R}^n$, $< Z_1, Z_2 >= Z_1'\,\Sigma^{-1}\,Z_2$. L'estimateur $\widehat{\theta}_G$ de θ par **moindres carrés généralisés** minimise $\|Y - X\cdot\theta\|_\Sigma$ pour $\theta \in \mathbb{R}^p$. On montre que

$$\widehat{\theta}_G = (X'\,\Sigma^{-1}\,X)^{-1} X'\,\Sigma^{-1}\,Y$$

En effet, les moindres carrés généralisés reviendront à minimiser $\|Y - X\cdot\theta\|_\Sigma$, donc à trouver la projection Σ-orthogonale de Y sur $[X]$. Le projeté Σ-orthogonal $P_{[X]}(Y) = M\cdot Y$ se définit par le fait que $P_{[X]}(Y) \in [X]$ et $(Y - P_{[X]}(Y)) \perp [X]$. En écrivant que $P_{[X]}(Y) = X\cdot\widehat{\theta}_G$ on a déjà $P_{[X]}(Y) \in [X]$. Ensuite, pour chaque colonne $Z^{(i)}$ de X (qui est supposé de rang k), ce qui définit un vecteur de $\mathbb{R}^n$, on doit avoir $Z^{(i)'}\cdot\Sigma^{-1}\cdot(Y - X\cdot\widehat{\theta}_G) = 0$, donc en "concaténant" ces k équations, on retrouve les équations "normales" :

$$X'\cdot\Sigma^{-1}\cdot(Y - X\cdot\widehat{\theta}_G) = 0, \quad \text{soit} \quad X'\cdot\Sigma^{-1}\cdot Y = X'\cdot\Sigma^{-1}\cdot X\cdot\widehat{\theta}_G$$

On en déduit donc que $\widehat{\theta}_G = (X'\cdot\Sigma^{-1}\cdot X)^{-1}\cdot X'\cdot\Sigma^{-1}\cdot Y$.

Remarque : Si les observations sont conjointement gaussiennes, c'est-à-dire que le vecteur ε suit une loi $\mathcal{N}_n(0,\Sigma)$, on montre facilement que $\widehat{\theta}_G$ est l'estimateur du maximum de vraisemblance de θ.

Par ailleurs, on a $\mathbb{E}(\widehat{\theta}_G) = (X' \cdot \Sigma^{-1} \cdot X)^{-1} \cdot X' \cdot \Sigma^{-1} \cdot X \cdot \theta = \theta$: l'estimateur $\widehat{\theta}_G$ est non biaisé.
Enfin,

$$\begin{aligned} \mathrm{cov}(\widehat{\theta}_G) &= (X' \cdot \Sigma^{-1} \cdot X)^{-1} \cdot X' \cdot \Sigma^{-1} \cdot \Sigma \cdot \left((X' \cdot \Sigma^{-1} \cdot X)^{-1} \cdot X' \cdot \Sigma^{-1}\right)' \\ &= (X' \cdot \Sigma^{-1} \cdot X)^{-1} \cdot X' \cdot \Sigma^{-1} \cdot \Sigma \cdot \Sigma^{-1} \cdot X \cdot (X' \cdot \Sigma^{-1} \cdot X)^{-1} \\ &= (X' \cdot \Sigma^{-1} \cdot X)^{-1}. \end{aligned}$$

En utilisant le Théorème de Gauss-Markov (voir exercice), on peut également montrer la proposition suivante:

Proposition. *Pour le modèle* (1) *avec ε un vecteur aléatoire centré tel que $cov(\varepsilon) = \mathbb{E}(\varepsilon\,\varepsilon') = \Sigma$, où Σ est une matrice de rang n (X est une matrice (n,p) de rang p), $\widehat{\theta}_G$ est l'estimateur de θ non biaisé et linéaire ayant la plus petite matrice de variance-covariance (au sens de la relation d'ordre entre matrice définie positive). En particulier, sa matrice de variance-covariance est plus petite ou égale à celle de l'estimateur par moindres carrés ordinaires.*

Application: estimation de la matrice de covariance et Pseudo-MCG

En pratique il est difficile de connaître a priori la matrice de variance-covariance des erreurs. Donc l'estimateur des moindres carrés généralisés n'est généralement pas utilisable tel quel. Cependant il est parfois possible d'utiliser une estimation de la matrice de variance-covariance des erreurs à partir d'une estimation de celle des résidus $\widehat{\varepsilon}$ par moindres carrés ordinaires. Si $\widehat{\Sigma}$ est cette estimation, et si $\widehat{\Sigma} \simeq \Sigma$ avec $\widehat{\Sigma}$ inversible, on peut approcher $\widehat{\theta}_G$ par estimateur $\widetilde{\theta}_G$ appelé **estimateur par moindres carrés pseudo-généralisés** défini par :

$$\widetilde{\theta}_G = (X'\,\widehat{\Sigma}^{-1}\,X)^{-1} X'\,\widehat{\Sigma}^{-1}\,Y.$$

On verra un exemple concret d'utilisation d'un tel estimateur dans la section suivante.
Mais dans le cas général comment penser pouvoir estimer Σ qui est une matrice symétrique de taille n, donc qui contient a priori $(n(n+1)/2$ éléments distincts à partir d'un échantillon de taille n? Ce n'est pas envisageable!

Cas particulier des modèles linéaires avec bruit autoregressif

On suppose que l'erreur du modèle linéaire suit un processus ARMA(p,q) stationnaire causal et d'ordre 2. On est donc dans le cadre d'un modèle de bruit homoscédastique mais corrélé. Quelle démarche utiliser alors pour estimer θ? On proposera plutôt une démarche en 3 temps:

1. On applique un estimateur par moindres carrés ordinaires. On peut alors montrer que sous l'hypothèse classique, c'est-à-dire $\max_{1\leq i\leq n} |H_{ii}| \underset{n\to+\infty}{\longrightarrow} 0$, $\widehat{\theta} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \theta$. On en déduit $(\widehat{\varepsilon}_i)_i$. On peut alors montrer que $\widehat{\varepsilon}_i \underset{n\to+\infty}{\xrightarrow{\mathbb{L}^2}} \varepsilon_i$ pour tout $i \in \mathbb{N}$.

2. On estime paramétriquement les paramètres du modèle ARMA(p,q) (typiquement par la méthode de Whittle). Si on note $\alpha = (a_1, \cdots, a_p, b_1, \cdots, b_q, \sigma^2)$ l'ensemble des paramètres, on montre que $\widehat{\alpha} \underset{n\to+\infty}{\xrightarrow{\mathcal{P}}} \alpha$.

3. On définit alors $\widehat{\Sigma} = \Sigma(\widehat{\alpha})$ où $\mathrm{cov}(\varepsilon) = \Sigma(\alpha)$. On définit alors l'estimateur par moindres carrés pseudo-généralisés $\widetilde{\theta}_G$, qui asymptotiquement a une covariance inférieure à celle de $\widehat{\theta}$.

Cas particulier des modèles linéaires avec bruit hétéroscédastique non corrélé

On suppose donc que l'on a le modèle linéaire général $Y = X\theta + \varepsilon$ avec ε centré et $\mathrm{cov}(\varepsilon) = diag((\sigma_i^2)_{1\leq i\leq n})$ c'est-à-dire une matrice diagonale avec des σ_i^2 sur la diagonale. On définit $\widehat{\varepsilon}$, le vecteur des résidus obtenus après une régression linéaire par moindres carrés ordinaires.
On l'a dit, il n'est pas possible d'estimer la matrice de covariance lorsqu'elle est aussi générale. Au lieu de l'estimer pour chaque coordonnées, on se contentera de l'estimer "globalement". On considèrera ainsi:

$$\widehat{\Sigma} = diag(\widehat{\varepsilon}_i^2 * (1 - H_{ii})^{-1}.$$

Le terme $(1-H_{ii})^{-1}$ vient du fait que lorsque $\sigma_i^2 = \sigma^2$ pour tout i, donc le cas homoscédastique, alors $\text{var}(\widehat{\varepsilon}_i) = (1-H_{ii})\sigma^2$.

Remarque: On utilise parfois $\widehat{\Sigma} = diag(\widehat{\varepsilon}_i * (1-H_{ii})^{-2}$: c'est le procédé Jackknife (couteau suisse).

4 Régréssion logistique

4.1 Position du problème

On suppose dans cette partie que les Y_i sont des variables qualitatives, et pour commencer que les Y_i sont des variables à deux modalités, que nous noterons 0 et 1. On retrouve ce genre de variables dans de nombreux exemples, par exemple lorsque Y_i mesure l'obtention ou non d'un crédit (économie), la mort ou la vie (biologie, pharmacologie),...
On suppose donc que l'on connaît $Y_1, \cdots, Y_n$ et que des variables $X_1, \cdots, X_p$ sont des variables potentiellement explicatives de Y. Ces variables peuvent être quantitatives ou bien qualitatives, auquel cas on les remplace par des indicatrices. Comme les Y_i ne prennent pour valeurs que 0 ou 1, on ne peut utiliser un modèle linéaire "habituel", $Y = X\theta + \varepsilon$. On cherchera plutôt un modèle reliant les probabilités que $Y = 0$ et$Y = 1$ avec les variables potentiellement explicatives. Plus concrètement, on note

$$p_i = P(Y_i = 1) \qquad \text{et donc} \qquad 1 - p_i = P(Y_i = 0).$$

L'idée sera ainsi d'écrire que:

$$g(p_i) = \theta_0 + \theta_1 X_1^i + \cdots + \theta_p X_p^i \qquad \text{pour tout } i \in \{1, \cdots, n\},$$

où g est une fonction réelle monotone qui va de $[0,1]$ dans $\mathbb{R}$. On en déduit donc que

$$p_i = g^{-1}\big(\theta_0 + \theta_1 X_1^i + \cdots + \theta_p X_p^i\big).$$

Les modèles les plus utilisés de cette fonction g sont les suivants:

1. Fonction **probit**: $g^{-1}(x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt$;
2. Fonction **logit**: $g^{-1}(x) = \dfrac{e^x}{1+e^x} \quad \Longrightarrow \quad g(p) = \ln\big(\dfrac{p}{1-p}\big)$;
3. Fonction **log-log**: $g^{-1}(x) = 1 - \exp(-e^x) \quad \Longrightarrow \quad g(p) = \ln\big(-\ln(1-p)\big)$;

On peut trouver des légitimations à l'utilisation des 2 premières fonctions:

1. Fonction **probit**: Si on considère le modèle classique $Z = X\theta + \varepsilon$ avec $\varepsilon \overset{\mathcal{L}}{\sim} \mathcal{N}(0, I_n)$, Z étant une variable dite latente, et $Y = \mathbb{I}_{Z \geq 0}$. Alors on peut montrer que $p_i = g^{-1}((X\theta)_i)$.
2. Fonction **logit**: Si on considère le modèle classique $Z = X\theta + \varepsilon$ avec $\varepsilon_i \overset{\mathcal{L}}{\sim}$, Z étant une variable dite latente, et $Y = \mathbb{I}_{Z \geq 0}$. Alors on peut montrer que $p_i = g^{-1}((X\theta)_i)$.

4.2 Mise en place concrète

Concrètement, on écrit le modèle de régression suivant:

$$(g(Y_i))_{1 \leq i \leq n} = X\theta + \varepsilon$$

et on estime θ par moindres carrés ordinaires. Bien-sûr les $g(Y_i)$ ne prennent que 2 valeurs, mais on peut supposer que X est de rang p donc on peut estimer θ. On en déduit alors les valeurs prédites:

$$\widehat{p}_i = g^{-1}\big((X\widehat{\theta})_i\big),$$

ce qui permet d'avoir une estimation de la probabilité que Y_i soit égale à 1.
On en déduit également pour un nouvel individu $n+1$ une prédiction de $P(Y_{n+1}=1$:

$$\widehat{p}_{n+1} = g^{-1}\big((X\widehat{\theta})_{n+1}\big).$$

Cependant, une telle méthode est loin d'être optimale. On va plutôt revenir à une estimation par maximum de vraisemblance: les Y_i sont des variables de Bernoulli de paramètres p_i. On a la relation $p_i = g^{-1}\big((X\theta)_i\big)$ et la vraisemblance s'écrit:

$$L_\theta(Y_1,\cdots,Y_n) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i},$$

(on a un modèle de type multinomial). En passant au logarithme, on obtiendra ainsi:

$$\widehat{\theta} = \mathrm{Argmax}_{\theta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} Y_i \log\big[g^{-1}\big((X\theta)_i\big)\big] + (1-Y_i)\log\big[1-g^{-1}\big((X\theta)_i\big)\big].$$

La question se pose de savoir comment déterminer $\widehat{\theta}$, car cette optimisation n'admet pas dans le cas général de solution explicite. Une méthode possible est l'utilisation de l'algorithme de Newton-Raphson, qui sert usuellement à déterminer numériquement une solution x_0 de l'équation $f(x)=0$ en se plaçant avec une condition initiale pas "trop" éloignées de cette solution. Pour ce faire, on utilise une suite récurrente (u_n) définie par:

$$u_{n+1} = u_n - \frac{f(u_n)}{f'(u_n)}.$$

Si on a choisie une condition initiale trop éloignée de x_0 telle que $f'(x)=0$ dans ce voisinage de x_0, il y a un risque que la méthode n'aboutisse pas. On peut montrer que la vitesse de cette méthode est quadratique car on montre à l'aide d'un développement de Taylor-Lagrange d'ordre 2 de f en x_0, que si $M_1 = \min_{x\in I}|f'(x)|$ et $M_2 = \max_{x\in I}|f''(x)|$, alors

$$|u_n - x_0| \le \frac{M_2}{2M_1}|u_{n-1}-a|^2 \le \Big(\frac{M_2}{2M_1}\Big)^{2^n-1}|u_0-a|^{2^n}.$$

Donc la convergence est dite quadratique et si u_0 est suffisamment proche de a (si $\frac{M_2}{2M_1}|u_0-a|<1$).

On peut appliquer une telle méthode pour trouver un extremum d'une fonction régulière puisque celui-ci satisfera $f'(x)=0$. Plus précisément, comme θ est de dimension $p+1$, on doit utiliser une méthode permettant d'obtenir une solution de l'équation:

$$\frac{\partial}{\partial\theta_i}\log(L_\theta(Y_1,\cdots,Y_n)=0 \qquad \text{pour tout } i=0,\cdots,p.$$

Ainsi on définira la suite $(\theta^{(n)})$ telle que:

$$\theta^{(n+1)} = \theta^{(n)} - \Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(L_{\theta^{(n)}}(Y_1,\cdots,Y_n)\Big)^{-1}_{ij}\Big(\frac{\partial}{\partial\theta_i}\log(L_{\theta^{(n)}}(Y_1,\cdots,Y_n)\Big)_i.$$

Les différentes dérivées partielles sont pénibles à calculer (dépendantes de la fonction g choisie et de ses dérivées, ainsi que de X), mais ce calcul est possible et conduit même à une résolution par moindres carrés pondérés (voir le logiciel).

Pour terminer, en ayant supposé les variables Y_i indépendantes et comme le modèle est régulier dès que g est de classe $\mathcal{C}^2$, on peut montrer que les résultats usuels sur l'estimateur par maximum de vraisemblance s'applique (attention les variables ne sont pas identiquement distribuées...) et on obtient que:

$$\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(L_{\widehat{\theta}}(Y_1,\cdots,Y_n)\Big)^{1/2}_{ij}\big(\widehat{\theta}-\theta\big) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{p+1}\big(0\,,\,I_{p+1}\big), \tag{2}$$

(avec le Théorème de Slutsky). Ceci permet d'obtenir des intervalles de confiance sur les paramètres et également des tests.

4.3 Qualité d'une régression logistique, tests d'adéquation et de significacité

Une fois que l'on obtient $\widehat{\theta}$ comment savoir si:

1. le modèle de régression est acceptable?
2. toutes les variables sont réellement significatives?
3. comment choisir entre `logit`, `probit`,...

Matrice de confusion En premier lieu, suite à une régression logistique on peut définir la matrice de confusion, qui n'est d'ailleurs pas spécifique à la méthode de régression logistique, mais plutôt à une méthode permettant d'obtenir une explication d'une variable qualitative:

$$M = \begin{pmatrix} \text{1 prédit, 1 en réalité} & \text{0 prédit, 1 en réalité} \\ \text{1 prédit, 0 en réalité} & \text{0 prédit, 0 en réalité} \end{pmatrix}.$$

Le pourcentage d'erreurs commises se déduit facilement d'une telle matrice. On pourra donc comparer différentes méthodes avec une telle matrice et évidemment on préférera une méthode minimisant le taux d'erreur.

Remarque importante: Pour mieux "comparer" différentes modélisation on préférera découper la base de données en 2 bases différentes, une base dite d'apprentissage, sur laquelle on apprend le modèle, donc on estime les différents paramètres, et une base dite base de test sur laquelle on comparera les différentes modélisations, dont les paramètres ont été estimés sur la base d'apprentissage.

Le soucis d'une telle matrice de confusion est la difficulté de trouver le niveau d'erreur à partir duquel le modèle est oui ou non satisfaisant. On préférera plutôt des tests naturellement issus de la statistique inférentielle.

Test de rapport de vraisemblance Comme nous sommes dans un cadre paramétrique, il est possible d'utiliser un tel test pour tester des hypothèses telles que:

$$H_0 :\ \theta_{i_1} = 0,\ \theta_{i_2} = 0, \cdots, \theta_{i_m} = 0 \qquad \text{contre} \qquad H_1 :\ \text{Le modèle est complet,}$$

où m, et les $i_1, \cdots, i_m$ sont choisis par l'expérimentateur. En particulier on pourra tester si chacune des variables est significative (on teste à chaque fois la nullité d'un seul coefficient) ou si globalement le modèle est satisfaisant (on teste le cas où tous les θ_i sauf θ_0 sont nuls).
De tels problèmes de test reviennent donc à tester un sous-modèle contre le modèle complet. On rappelle que le test du rapport de vraisemblance consiste à calculer:

$$\widehat{T} = \frac{\text{Vraisemblance maximisée sous } H_0}{\text{Vraisemblance maximisée sous } H_1}.$$

Lorsque les variables sont supposées êtres indépendantes les unes les autres on montre que pour un modèle régulier (ce qui est le cas ici lorsque la fonction g est de classe $\mathcal{C}^2$) et sous l'hypothèse H_0:

$$-2\log\left(\widehat{T}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(p+1-m).$$

On peut utiliser aisément un tel test dans le cas de la régression logistique.

Test de Wald Toujours dans un cadre paramétrique, et en utilisant la normalité asymptotique des coefficients estimés, on peut directement tester si ces coefficients sont nuls ou non. C'est le principe d'un test de Wald et cela permet de tester les mêmes hypothèses que le test du rapport de vraisemblance:

$$H_0 :\ \theta_{i_1} = 0,\ \theta_{i_2} = 0, \cdots, \theta_{i_m} = 0 \qquad \text{contre} \qquad H_1 :\ \text{Le modèle est complet,}$$

On définit ainsi:

$$\widetilde{T} = \left(\widehat{\theta}_{i_1}, \cdots, \widehat{\theta}_{i_m}\right)\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(L_{\widehat{\theta}}(Y_1, \cdots, Y_n)\right)^{-1}_{i,j=i_1,\cdots,i_m}\begin{pmatrix}\widehat{\theta}_{i_1}\\ \vdots \\ \widehat{\theta}_{i_m}\end{pmatrix}.$$

Alors, sous l'hypothèse H_0, on déduit facilement de (2),

$$\widetilde{T} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2(m).$$

Sélection de modèles Il est aussi possible d'utiliser les critères de sélection de modèles déjà vus pour choisir parmi les potentielles variables explicatives celles que l'on choisira pour effectuer des prédictions. En particulier, on utilisera le critère BIC qui se définit de la manière suivante:

$$BIC(m) = -2\,\log(\text{Vraisemblance maximisée pour le modèle } m) + \log n\,|m|,$$

et l'on choisira $\widehat{m} = \text{Argmin}_{m\in\mathcal{M}} BIC(m)$.

4.4 Extension dans le cas de plusieurs modalités: régression polytomique

On suppose maintenant que la variable Y est une variable qualitative pouvant prendre plusieurs modalités, soit J modalités. Dans le cas $J = 2$, la régression logistique suffit. Dans le cas où les modalités sont ordonnées (par exemple, niveau 1 à 4 pour la gravité d'une maladie) une extension assez simple de la régression logistique peut-être envisagée: on découpe l'intervalle $[0, 1]$ en J parties égales ordonnées, chacune de ses parties correspondant à l'affectation ordonnée d'une modalité.
Dans le cas de modalités non ordonnées, on va considérer une autre extension de la régression logistique: la régression polytomique. Le but est désormais d'estimer la probabilité:

$$p_{ij} = P(Y_i = j \,|\, X) \qquad \text{pour } j = 1, \ldots, J \text{ et } i = 1, \ldots, n.$$

Pour ce faire, on va considérer à part une modalité, par exemple la modalité J (ce choix n'a aucune incidence) et on va effectuer $J - 1$ régressions logistiques. Ici nous ne prendrons que le cas de la fonction **logit** qui est le cas le plus simple:

$$\log\Big(\frac{p_{ij}}{p_{iJ}}\Big) = (X\theta^{(j)})_i \qquad \text{pour } j = 1, \ldots, J-1 \text{ et } i = 1, \ldots, n,$$

avec la condition $\sum_{j=1}^{J} p_{ij} = 1$, ce qui permet d'obtenir toutes les p_{ij}. Notons ici que pour chaque modalité j (excepté la modalité J) on associe un vecteur de paramètres $\theta^{(j)}$. Ainsi on aura:

$$p_{ij} = \frac{e^{(X\theta^{(j)})_i}}{1 + \sum_{k=1}^{J-1} e^{(X\theta^{(k)})_i}}$$

Cela signifie concrètement que l'on va estimer

$$\widehat{p}_{ij} = \frac{e^{(X\widehat{\theta}^{(j)})_i}}{1 + \sum_{k=1}^{J-1} e^{(X\widehat{\theta}^{(k)})_i}},$$

et $\widehat{p}_{iJ} = 1 - \sum_{j=1}^{J-1} \widehat{p}_{ij}$. Enfin, pour la prédiction d'une nouvelle valeur on utilisera la règle:

$$\widehat{Y}_{n+1} = \text{Argmax}_{j=1,\ldots,J} \widehat{p}_{(n+1),j}.$$

Comment peut-on estimer les différents vecteurs de paramètres? Une nouvelle fois on utilise un estimateur par maximum de vraisemblance. En effet, sous l'hypothèse que les variables Y_i sont indépendantes, la vraisemblance s'écrit:

$$L_{(\theta^{(1)},\ldots,\theta^{(J-1)})}(Y_1, \cdots, Y_n) = \prod_{i=1}^{n} \Big(\prod_{j=1}^{J} p_{ij}^{\mathbb{I}_{Y_i=j}} \Big).$$

Cela correspond à la vraisemblance d'une loi multinomiale. Comme dans le cas de la régression logistique, on utilisera un algorithme de type Newton-Raphson pour approcher l'estimateur de $(\theta^{(1)}, \ldots, \theta^{(J-1)})$ par maximum de vraisemblance.
Il est également possible de mettre en place des tests du rapport de vraisemblance, des tests de Wald, de la sélection de modèle.

5 Moindres carrés non linéaires

5.1 Présentation

Dans tout ce qui précède, nous avons supposé que le modèle liant les Y_i aux différents $X_i^{(k)}$ était linéaire. Mais si ceci peut être fait lors d'un premier travail de modélisation, il peut s'avérer plus intéressant d'écrire le modèle sous une forme plus générale:

$$Y_i = g_\theta(X_i^{(1)}, \cdots, X_i^{(p)}) + \varepsilon_i$$

où g_θ est une fonction connue, mais θ est un vecteur de paramètres inconnu (et comme précédemment les Y_i, $X_i^{(k)}$ sont connues, les ε_i inconnus, mais indépendants, centrés, de variance fini et constante). On cherchera à estimer θ. On est donc dans un cadre semi-paramétrique.

Remarque: Il est également possible de passer à un cadre non-paramétrique en posant comme modèle:

$$Y_i = g(X_i^{(1)}, \cdots, X_i^{(p)}) + \varepsilon_i$$

où g est inconnue, appartenant à un certain espace fonctionnel (typiquement $\mathbb{L}^2$). Il existe de nombreuses approches dans un tel cadre: estimation par noyaux (Nadaraya-Watson), estimation par splines, par régressions locales, par projection sur une base d'ondelettes,...

On se contentera ici d'évoquer le cas unidimensionnel suivant:

$$Y_i = g_\theta(X_i) + \varepsilon_i. \tag{3}$$

On va supposer ici qu'il n'est pas possible de directement linéariser le modèle, comme par exemple dans le cas où $g_\theta(x) = \theta_1\, x^{\theta_2}$ par un passage au logarithme.

Exemples: Citons par exemple:

- La fonction de Lorentz: $g_\theta(x) = \dfrac{\theta_1}{1 + \left(\frac{x-\theta_2}{\theta_3}\right)^2}$.
- Un modèle de type exponentiel: $g_\theta(x) = \theta_1 + \theta_2\, x^{\theta_3}$.

5.2 Estimation par moindres carrés

Le but est donc d'estimer θ dans le modèle (3). Nous allons proposer une méthode qui s'inspire de l'estimation par maximum de vraisemblance pour un échantillon gaussien et qui prolonge le cadre de la régression linéaire: c'est la méthode des moindres carrés. Il s'agira donc de minimiser (en θ):

$$S(\theta) = \sum_{i=1}^{n} \big(Y_i - g_\theta(X_i)\big)^2,$$

et ainsi:

$$\widehat{\theta} = \mathrm{Argmin}_{\theta \in \mathbb{R}^d} S(\theta).$$

Contrairement au cas linéaire, il peut y avoir plusieurs solutions à cette minimisation. Par exemple, dans le cas de la fonction de Lorentz, on voit tout de suite que l'on peut aussi bien prendre θ_3 ou $-\theta_3$. Dans l'autre exemple, on vient bien que si $\theta_2 = 0$ alors on peut choisir θ_3 comme l'on veut. La bijectivité de l'application $\theta \to g_\theta$, et plus particulièrement au voisinage du vrai paramètre θ^*, est donc la clé de cette unicité de $\widehat{\theta}$.

Pour minimiser $S(\theta)$ et contrairement au cas linéaire, il n'y a pas, en général, de formule explicite. On en revient à la minimisation d'une fonction à plusieurs variables. On en vient donc à déterminer la solution des équations normales:

$$0 = \frac{\partial g_\theta}{\partial \theta}(X_i)\big(Y_i - g_\theta(X_i)\big) \qquad \text{pour tout } i = 1, \cdots, n,$$

ce qui peut encore s'écrire matriciellement:

$$0 = {}^t\dot{G}_\theta(X)\big(Y - G_\theta(X)\big)$$

avec $G_\theta(X) = (g_\theta(X_i))_i$, $Y = (Y_i)_i$ et $\dot{G}_\theta(X)$ la matrice $\big(\frac{\partial g_\theta}{\partial \theta_j}(X_i)\big)_{ij}$.

Comment faire pour résoudre cette équation? Posons $\widehat{\theta}$ une solution de l'équation normale, solution que l'on supposera pour l'instant unique, ce qui signifie que:

$$0 = {}^t\dot{G}_{\widehat{\theta}}(X)\big(Y - G_{\widehat{\theta}}(X)\big). \tag{4}$$

Deux possibilités:

1. On utilise, comme dans le cas de la régression logistique, une méthode de résolution de type Newton-Raphson;
2. On "linéarise" le système (c'est-à-dire que l'on en le résolvant de manière itérative (donc on considère une suite $(\widetilde{\theta}^k)_k$). Pour cela on utilise la formule de Taylor

$$\begin{aligned} G_{\widehat{\theta}}(X) &\simeq G_{\widetilde{\theta}^k}(X) + \dot{G}_{\widetilde{\theta}^k}(X)\big(\widehat{\theta} - \widetilde{\theta}^k\big) \\ &\simeq G_{\widetilde{\theta}^k}(X) + \dot{G}_{\widetilde{\theta}^k}(X)\big(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\big) \end{aligned}$$

(en supposant que $(\widetilde{\theta}^k)_k$ converge vers $\widehat{\theta}$). On peut ainsi écrire que $Y - G_{\widehat{\theta}}(X) \simeq Y - G_{\widetilde{\theta}^k}(X) - \dot{G}_{\widetilde{\theta}^k}(X)\big(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\big)$. Les équations normales s'approchent alors par les équations:

$$0 = {}^t\dot{G}_{\widetilde{\theta}^k}(X)\big(Y - G_{\widetilde{\theta}^k}(X) - \dot{G}_{\widetilde{\theta}^k}(X)\big(\widetilde{\theta}^{k+1} - \widetilde{\theta}^k\big)\big).$$

Le but ici est de trouver $\widetilde{\theta}^{k+1}$, ou encore $\Delta^{k+1} = \widetilde{\theta}^{k+1} - \widetilde{\theta}^{k+1}$. On retombe sur des équations similaires au cas linéaire puisqu'en posant $Z^k = Y - G_{\widetilde{\theta}^k}(X)$ (connu), l'équation précédente devient:

$$0 = {}^t\dot{G}_{\widetilde{\theta}^k}(X)\big(Z^k - {}^t\dot{G}_{\widetilde{\theta}^k}(X)\,\Delta^{k+1}\big) \quad\Longrightarrow\quad \Delta^{k+1} = \big({}^t\dot{G}_{\widetilde{\theta}^k}(X)\,\dot{G}_{\widetilde{\theta}^k}(X)\big)^{-1}{}^t\dot{G}_{\widetilde{\theta}^k}(X)\,Z^k.$$

(en supposant bien-sûr que les matrices $\big({}^t\dot{G}_{\widetilde{\theta}^k}(X)\,\dot{G}_{\widetilde{\theta}^k}(X)\big)^{-1}$ existent). Par itération, si la valeur initiale $\widetilde{\theta}^0$ est bien choisie suffisamment près de $\widehat{\theta}$ (ce qui peut se faire à partir d'une grille), si $\theta \mapsto \dot{G}_\theta(X)$ ne s'annule pas dans un voisinage de $\widehat{\theta}$, la suite $(\widetilde{\theta}^k)$ converge vers $\widehat{\theta}$.

5.3 Etude asymptotique de l'estimateur des moindres carrés non-linéaires

Supposons que l'on ait su obtenir $\widehat{\theta}$, solution de (4). Les propriétés permettant d'obtenir aisément l'espérance et la variance de $\widehat{\theta}$ dans le cas d'un modèle linéaire ne sont plus applicables dans le cas non linéaire. Cependant, sous certaines hypothèses, on peut retrouver des propriétés asymptotiques assez proches du cas linéaire:

Proposition. *On suppose que:*

1. Θ *est un ouvert borné de* $\mathbb{R}^p$*;*
2. *Pour* $\theta_1, \theta_2 \in \Theta$, $\frac{1}{n}\sum_{i=1}^n \big(g_{\theta_1}(X_i) - g_{\theta_2}(X_i)\big)^2 \underset{n\to+\infty}{\longrightarrow} K(\theta_1,\theta_2) < \infty$ *et* $K(\theta_1,\theta_2) \neq 0$ *si* $\theta_1 \neq \theta_2$*;*
3. *Pour tout* $x = X_i$*, la matrice Hessienne de* g_θ, $\Big(\frac{\partial^2 g_\theta}{\partial\theta_i\partial\theta_j}(x)\Big)$*, existe et est continue au voisinage de* θ^**;*
4. *Pour tout* θ *dans un voisinage de* θ^**, la matrice* $I(\theta) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \dot{G}_\theta(X_i)^t\dot{G}_\theta(X_i)$ *existe et est inversible;*
5. *L'erreur* $(\varepsilon_i)_i$ *est une suite de vaiid centrées, telles que* $var\varepsilon_i = \sigma^2$ *et* $\mathbb{E}(\varepsilon_i^4) < \infty$*.*

Alors:

$$\sqrt{n}\,\big(\widehat{\theta} - \theta^*\big) \underset{r\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}\big(0\,,\,\sigma^2\,I(\theta^*)^{-1}\big).$$