Robustesse du modèle linéaire à la non normalité des erreurs

J.M. Azaïs *

22 mai 2012

Le modèle linéaire repose sur l'hypothèse a priori que les erreurs sont normales. Á part des cas très particuliers, comme certains modèles de génétique ou encore le cas où les observations sont en fait des moyennes, le statisticien n'a en général pas beaucoup d'arguments pour justifier cette normalité.

Par ailleurs tester la normalité a posteriori à partir des données n'est pas très commode dans la mesure où les erreurs ϵ_i ne sont pas observables et que l'on a à notre disposition seulement les résidus $\widehat{\epsilon}_i$ qui n'en sont qu'une estimation. Cette estimation détruit toutes les propriété d'indépendance et d'égalité de la variance sur laquelle sont basé les tests classiques comme le test de Kolmogorov-Smirnov ou le test de Anderson-Darling. L'hypothèse de normalité est donc le plus souvent invérifiable.

Fort heureusement, le modèle linéaire a une propriété de **robustesse** dès que le nombre d'observations est grand. C'est à dire que sous des hypothèses très générales, certaines des propriétés classiques du modèle linéaire restent vraies quand le nombre d'observations tend vers l'infini même si les données de départ ne sont pas normales. Cela est dû au fait que les estimateurs des moindres carrés peuvent être interprétés comme des moyennes pondérées qui vérifient loi des grands nombres et Théorème central limite.

1 Analyse de la variance à un facteur

Proposition 1 *on considère une famille de modèle d'analyse de la variance à un facteur*

$$Y_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \dots, I \quad j = 1, \dots, J$$

où les $\epsilon_{i,j}$ sont des variables i.i.d. centrées et admettant un moment d'ordre 3 : σ^2 fini.

On suppose que I est fixé et que le nombre de répétition J tend vers l'infini. Alors pour tout i

$$\sqrt{J}\big(\widehat{\mu}_i - \mu_i\big) \to N(0, \sigma^2).$$

*Université de Toulouse, Statistique et Probabilités, IMT, Mél : jean-marc.azais@math.univ-toulouse.fr

On a également la convergence en probabilité de

$$\widehat{\sigma}^2 = \frac{1}{n-I}\sum_{i,j}(Y_{ij} - Y_{i.})^2 \text{ vers } \sigma^2.$$

où $n = IJ$ *et*

$$Y_{i.} = 1/J \sum_j Y_{ij}.$$

On en déduit que la statistique $\widehat{F}$ *du test de Fisher définie par*

$$\widehat{F} = \frac{J\sum_i (Y_{i.} - Y..)^2/(I-1)}{\sum_{i,j}(Y_{ij} - Y_{i.})^2/(n-I)}$$

converge quand J *tens vers l'infini, sous l'hypothèse nulle d'égalité des moyennes vers une loi* $\chi^2(I-1)/(I-1)$

Comme la limite de la loi de Fisher $F_{(I-1),(n-I)}$ est une loi $\chi^2(I-1)/(I-1)$, en ce qui concerne les estimateurs et la statistique du test F, nous obtenons la même distribution limite que la distribution soit gaussienne ou non. Cela prouve la robustesse.

En pratique on peut également étudier le même problème par simulation. Pour ce faire, on peut comparer le niveau nominal du test F (calculé sous l'hypothèse gausienne) avec le niveau empirique du même test calculé de manière empirique par simulation d'un modèle d'analyse de la variance avec des erreurs non gaussiennes. Ici on peut généraliser l'étude au cas ou les nombres de répétitions n_i des différentes classes $i = 1, \ldots, I$ ne sont pas tous égaux. On appelle ce dernier cas, le cas " à effectifs déséquilibrés". On constate par des simulations que la réunion des facteurs suivants est nécessaire pour trouver un écart important entre le niveau réel et le niveau nominal du test F

- loi des erreurs disymétriques
- faible nombre de répétitions
- effectifs déséquilibrés.

Dans le cas déséquilibré on a

$$\widehat{F} = \frac{\sum_i n_i (Y_{i.} - Y..)^2/(I-1)}{\sum_{i,j}(Y_{ij} - Y_{i.})^2/(n-I)}$$

2 Loi des grands nombres dans le cas général

Exemple ...

Notations nous allons considérer dans cette section et dans les suivantes, une suite de modèles linéaires de taille n :

$$Y^n = X^n\theta^n + \epsilon^n \quad n \to +\infty \tag{1}$$

où Y^n, ϵ^n sont des vecteurs de taille n, θ^n est un vecteur de taille $k_n < n$, X^n est une matrice n, k_n de plein rang k_n ; nous supposerons l'hypothèse (H) :

$\epsilon^n = (\epsilon_1, \ldots, \epsilon_n)$ où la suite ϵ_i est centrée et i.i.d. admettant un moment d'ordre 4

k_n peut être constant ou dépendre de n. Pour alléger les notations nous omettrons les exposants ^{n}

Considérons la convergence de la réponse

Proposition 2 – *Soit $\widehat{Y}_i$ la ième coordonnée de*

$X\widehat{\theta} = X(X'X)^{-1}X'Y = HY$. Nos conditions seront exprimées en fonction de H la "hat matrix", appelée ainsi car $HY = \widehat{Y}$. Alors
Pour i fixé dans $\{1, \cdots, n\}$,
$\left(\widehat{Y}_i^n - (X \cdot \theta)_i\right) \to 0$ en moyenne quadratique ssi $H_{ii}^n \to 0$.

– *Pour tout $i \in \{1, \cdots, n\}$,*
$\left(\widehat{Y}_i - (X \cdot \theta)_i\right) \to 0$ en moyenne quadratique ssi $\|H^n\| = \max_{1 \le i \le n} |H_{ii}^n| \to 0$.

Ces convergences en moyenne quadratique entraînent les convergences en probabilité.

Considérons maintenant l 'estimation de la variance

Proposition 3 *Supposons $k_n/n \to 0$ alors $\widehat{\sigma}^2$ converge en probabilité vers σ^2*

3 Théorème central limite

Les énoncés de cette section sont un peu compliqués car ils vont s'appliquer à des vecteurs de taille variable : le vecteur des paramètres θ et le vecteur des estimations $\widehat{Y} = X\widehat{\theta}$

Définition 1 *Nous dirons qu'une suite de vecteurs aléatoires Z^n dont la taille peut dépendre de n est asymptotiquement gaussienne si pour toute suite de combinaisons linéaires $(C^n)' \cdot Z^n$ non nulles, on a*

$$U^n := \frac{(C^n)' \cdot Z^n - \mathrm{E}((C^n)' \cdot Z^n)}{\sqrt{\mathrm{Var}\left((C^n)' \cdot Z^n\right)}} \Rightarrow \mathcal{N}(0,1) \text{ quand } n \text{ tend vers l'infini,}$$

où $\Rightarrow$ est la convergence en loi.

Le résultat est le suivant :

Théorème 1 *On considère la famille de modèles linéaires (1) sous l'hypothèse* **H**. *Alors, sous l'hypothèse $\|H^n\| = \max_{1 \le i \le n} |H_{ii}^n| \to 0$, les estimateurs $\widehat{Y}^n$ et $\widehat{\theta}^n$ sont asymptotiquement gaussiens.*

La démonstration de ce théorème est basée sur un théorème central limite pour des aires triangulaires sous condition de Lindeberg.

Théorème 2 (Théorème de la limite centrale de Lindeberg) *Soit* $(X_n)_{n\in\mathbb{N}}$ *une suite de variables aléatoires indépendantes et de même loi non gaussienne admettant un moment d'ordre* 2. *Sans perte de généralité, on peut supposer que* $\mathrm{E}(X_i) = 0$, $\mathrm{Var}(X_i) = 1$ *pour tout* $i \in \mathbb{N}$. *Considérons le "tableau triangulaire" suivant :*

$$(a_i^n)_{n\in\mathbb{N},i=1,\cdots,n} \quad \textit{tel que} \quad \sum_{i=1}^{n}(a_i^n)^2 = 1 \quad \textit{pour tout } n \in \mathbb{N}^*.$$

Alors on a :

$$Z_n = \sum_{i=1}^{n} a_i^n X_i \Rightarrow \mathcal{N}(0,1) \tag{2}$$

ssi

$$\max_{1\leq i\leq n} |a_i^n| \to 0. \tag{3}$$

4 Conclusion

Nous revenons sur le commentaire de l'introduction : si le nombre d'observation est grand les résidus $\widehat{\epsilon}$ sont proches des erreurs ϵ_i il serait donc possible dans ce cadre de faire une théorie asymptotique des test de normalité. les résultat ci dessus montrent qu cela serait sans intérêt : on pourrait tester la normalité seulement quand on n'en a pas besoin.

5 Suggestions

on pourra

- Démontrer les propositions 1 2 ou 3.
- Exposer les tests de normalité cités dans le texte et étudier leur propriétés (en terme de comparaison de niveau nominal et de niveau empirique) par simulation quand il sont appliqués aux résidus.
- Illustrer par une simulation les affirmations sur les propriétés de test F en analyse de la variance.
- Donner des exemples en analyse de la variance, en régression, pour des plans en blocs complets où la condition $\|H^n\| = \max_{1\leq i\leq n} |H_{ii}^n| \to 0$ du théorème est ou non vérifiée.