TP IV: Régression linéaire et erreurs non gaussiennes

Nous utiliserons le langage `R` pour ce TP. Le corrigé sera à faire sous forme de `.Rmd`.

1 Un peu de théorie

On dispose d'un échantillon de n couples $(x_i, y_i), i = 1 \cdots, n$ où

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdot + \beta_p x_{i,p} + \varepsilon_i, \text{ pour } i = 1, \cdots, n \tag{1}$$

où l'on suppose que le vecteur $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^t$ des résidus vérifie :
- indépendance: $\varepsilon_1, \cdots, \varepsilon_n$ sont indépendants et identiquement distribués,
- centrage: $\mathbb{E}\varepsilon_1 = 0$,
- homoscédasticité: $\mathbb{E}\varepsilon_1^2 = \sigma^2$.

Autrement dit, on ne suppose pas que la loi des erreurs est gaussienne. On veut étudier le comportement asymptotique des statistiques $\hat{\beta}$, l'estimateur de β par la méthode des moindres carrés, ainsi que des statistiques de tests classiques. On a le résultat théorique suivant

Théorème 1. On suppose que le modèle (1) est régulier et que les hypothèses précédentes sont vérifiées. On note π_X la matrice de projection sur $Im(X)$ et h_n le terme maximal de la matrice π_X

(*i*) Si $h_n \to 0$ lorsque $n \to +\infty$, $\hat{\beta}$ est asymptotiquement un vecteur gaussien.

(*ii*) En particulier si $\frac{1}{n^\alpha} X^t X \to Q$, où Q matrice définie positive et $\alpha > 0$, alors :

$$h_n \to 0 \quad \Longrightarrow \quad n^{\alpha/2}(\hat{\beta} - \beta) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 Q^{-1}\right)$$

(*iii*) On considère le problème de test $H_0 : K^t\beta = 0$ contre $H_1 : K^t\beta \neq 0$, où K^t matrice $k \times (p+1)$ de rang $\mathrm{Rg}(K) \leq p+1$. Alors on a :

$$h_n \to 0 \quad \Longrightarrow \quad \hat{F} = \frac{\left(K^t\hat{\beta}\right)^t \left[K^t \left(X^t X\right)^{-1} K\right]^{-1} K^t\hat{\beta}}{\mathrm{Rg}(K) s^2} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{\mathrm{Rg}(K)} \chi^2(\mathrm{Rg}(K))$$

Ce résultat, basé sur le Théorème de Lindeberg (qui généralise le TLC), assure la construction de tests asymptotiques de combinaisons linéaires des coefficients, notamment les tests de Fisher (global et partiel). Notons que d'après la théorie des probabilités, on sait que si $\hat{F} \sim \mathcal{F}[k, n-p]$, alors

$$\hat{F} \xrightarrow[n\to\infty]{\mathcal{L}} \frac{1}{k}\chi^2(k)$$

Ainsi les lois limites dans le théorème précédent sont les mêmes que celle que l'on peut obtenir dans le cas gaussien (cela est également valable pour les tests de Student). En conclusion, les résultats asymptotiques de la régression linéaire sont donc similaires sans l'hypothèse de normalité des résidus. C'est pour cela qu'en pratique, l'hypothèse de normalité n'est pas nécessaire lorsque n est assez grand.

2 Simulations

On va illustrer ces résultats en simulant des modèles de régression non-gaussiens. On considère un modèle de régression où $p = 3$ de la forme (1) où chaque $x_i = (x_{i,1}, x_{i,2}, x_{i,3})^t$ consiste en 3 réalisations de loi uniforme sur un intervalle $[0, 5]$. On prendra $\beta = (20, 2, 1, 0.5)^t$. Dans cette étude on se concentre sur l'erreur ε et on choisit $\varepsilon_1, \cdots, \varepsilon_n$ i.i.d. tel que $\varepsilon_i = we + (w - 1)e'$ où e, e' sont des variables aléatoires de loi exponentielle de paramètre $\lambda = 1$ et w une variable aléatoire uniforme sur $\{0, 1\}$.

1. Simuler un jeu de données avec $n = 100$ à l'aide du modèle décrit (écrire une fonction `generate_data(n)` qui prend en entrée un entier n et renvoie un n-échantillon). Faire la régression linéaire. Faire des graphiques.
2. En faisant varier n (prendre `seq(100, 2000, length = 50)`. Montrer que l'hypothèse de (i) est vérifiée expérimentalement.
3. En faisant varier n (prendre `seq(100, 10000, length = 50)`. Montrer que l'hypothèse de (ii) est vérifiée expérimentalement.
4. Calculer les réalisations de $\hat{\beta}, s^2$ et $F = \frac{MC_R}{MC_{res}}$ dans un modèle simulé.
5. Générer $m = 500$ modèles (1) et estimer la loi des statistiques précédentes. Vérifier les résultats théoriques.
6. Faites varier λ dans la régression avec bruit exponentiel. Que se passe-t'il?