VÉRIFICATION DES POSTULATS DU MODÈLE LINÉAIRE

Plan du chapitre 7

1. Importance des postulats
2. Vérification des postulats
3. Que faire si un des postulats n'est pas vérifié ?
4. Si rien ne marche

Importance des postulats

Premier postulat $E(\varepsilon_{\mathrm{n}}) = 0$

Le modèle consiste toujours à écrire la variable sous la forme :

$$\mathrm{Y_n} = \mu_{\mathrm{n}} + \varepsilon_{\mathrm{n}} \quad \text{avec} \quad \mu_{\mathrm{n}} = \sum \mathrm{x_{pn}} \theta_{\mathrm{p}}$$

Indispensable si l'on veut que le modèle soit plausible pour décrire la réalité.

Si les quatre postulats sont vérifiés, les estimateurs $\widehat{\theta}$ et $\widehat{\sigma}^2$ possèdent les propriétés suivantes :

— $E\left(\widehat{\theta}\right) = \theta$

— $\mathrm{Var}(\widehat{\theta}) = \sigma^2 (X'X)^{-1}$

— $\widehat{\theta}$ est gaussien

— $E(\widehat{\sigma}^2) = \sigma^2$

— $\widehat{\sigma}^2 \sim \frac{\sigma^2 \chi^2(\mathrm{N-P})}{\mathrm{N-P}}$

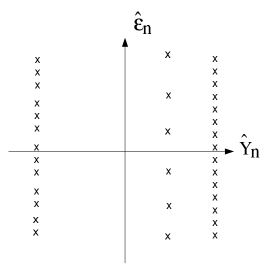
— $\widehat{\sigma}^2$ est indépendant de $\widehat{\theta}$

Grâce aux **quatre postulats**, il est possible de faire des tests sur l'estimateur $\widehat{\theta}$

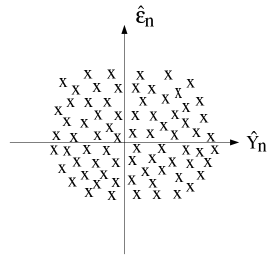
Vérification des postulats

Pour vérifier les postulats, on fait toujours une représentation graphique des résidus estimés.

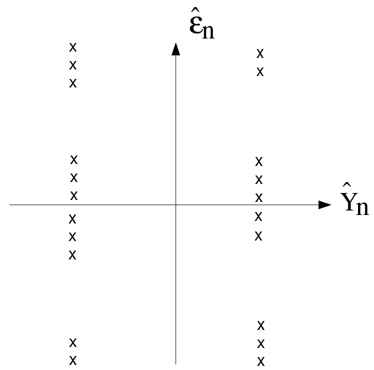
En analyse de variance



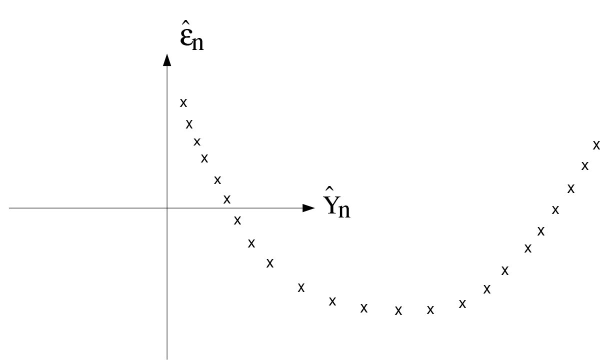
En regression



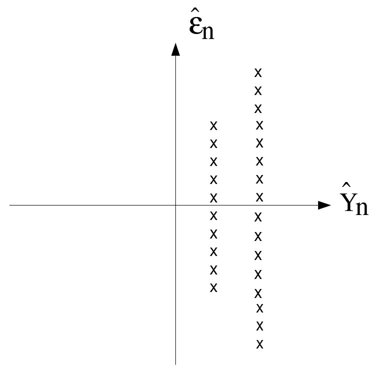
en analyse de la variance



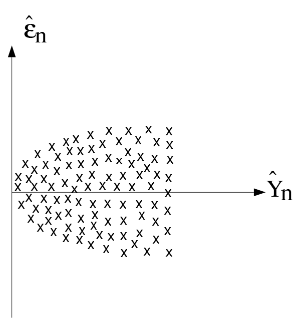
Le postulat $E(\varepsilon_n) = 0$ est-il vérifié ?
en regression

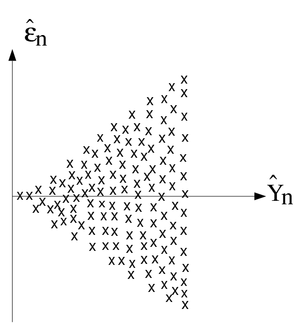


en analyse de la variance

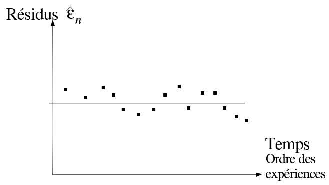


Le postulat $\mathrm{Var}(\varepsilon_n) = \sigma^2$ **est-il vérifié ?**
en regression



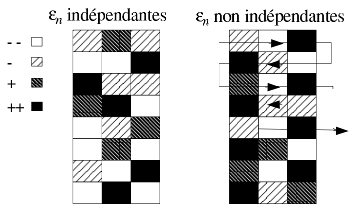


Représentation graphique de $\hat{\varepsilon}_n$ en fonction du temps ou de l'ordre des expériences



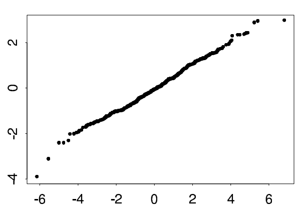
Les $\hat{\varepsilon}_n$ ne sont pas indépendants.

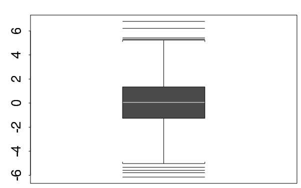
Représentation graphique schématique des $\hat{\varepsilon}_n$ dans l'espace

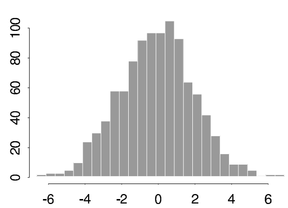


Les ε_n sont-ils normalement distribués ?

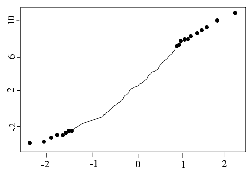
Représentations graphiques de la distribution de $\hat{\varepsilon}_n$

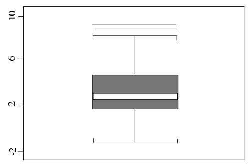


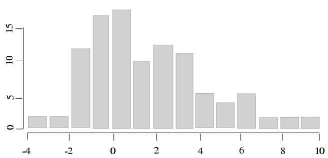




Représentations graphiques de la distribution de $\widehat{\varepsilon}_{\mathrm{n}}$







Que faire si un des postulats n'est pas vérifié ?

Si un postulat n'est pas vérifié, les données récoltées gardent leur valeur que faire alors ?

Postulat 1 : $E(\varepsilon_{\mathrm{n}}) = 0$ **pour tout n**

— en analyse de la variance : introduire de nouveaux facteurs

— en régression : penser à des régresseurs non linéaires et à des produits de régresseurs.

Postulat 2 : $\mathrm{Var}(\varepsilon_{\mathrm{n}}) = \sigma^2$ **constante**

La technique générale est de transformer la variable mesurée.

Lorsque l'on dispose d'une information sur la distribution de Y

Distribution de Poisson :

$$\sigma^2 \propto \mu \;:\; Y \longrightarrow \begin{cases} \sqrt{Y} \\ \sqrt{Y + \text{cste}} \end{cases}$$

Distribution Log-normale :

$$\sigma \propto \mu \;:\; Y \longrightarrow \begin{cases} \operatorname{Log} Y \\ \operatorname{Log}(Y+1) \end{cases}$$

Distribution Binomiale :

$$\sigma \propto \sqrt{\mu(1-\mu)} \;:\; Y \longrightarrow \arcsin \sqrt{Y}$$

Ne pas appliquer sans discernement

- regarder les résidus avant transformation
- choisir éventuellement une transformation
- vérifier les résidus après transformation

Transformations de variables lorsque $\mathrm{Var}(\varepsilon_n)$ n'est pas constante

Lorsque l'on ne dispose pas d'information sur la distribution de Y

Si l'écart-type est une fonction puissance de la moyenne $\sigma \propto \mu^k$,
on définit une famille de transformations :
pour $k \neq 1$, Y^{1-k}
pour $k = 1$, $\operatorname{Log} Y$

Exemples

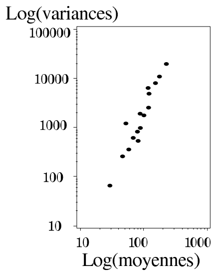
k=1	$\sigma \propto \mu$:	$Y \longrightarrow \operatorname{Log} Y$
$k = \frac{1}{2}$	$\sigma \propto \mu^{1/2}$,	
	$\sigma^2 \propto \mu$:	$Y \longrightarrow Y^{1-\frac{1}{2}} = \sqrt{Y}$

Trouver k ?

si $\sigma \propto \mu^k$, alors $\sigma^2 \propto \mu^{2k}$
d'où $\sigma^2 = c\,\mu^{2k}$
d'où $\operatorname{Log} \sigma^2 = c' + 2k \operatorname{Log} \mu$

Relation linéaire entre $\operatorname{Log} \mu$ et $\operatorname{Log} \sigma^2$, pente de la droite $= 2k$.

variables lorsque $\mathrm{Var}(\varepsilon_n)$ **n'est pas constante**



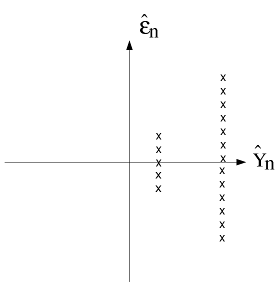
Comptages de mouches dans des pièges :

4 appâts 4 blocs 5 répétitions	Calcul de 16 moyennes et de 16 variances

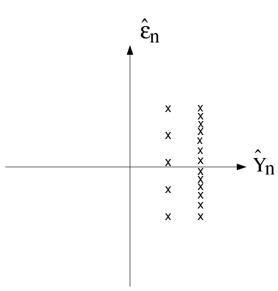
droite de pente très proche de 2
$\Longrightarrow$ transformation Log.

Vérification de $\mathrm{Var}(\varepsilon_n) = \sigma^2$ après transformation des variables

résidus avant transformation



résidus après transformation $Y \rightarrow \sqrt{Y}$



Postulat 3 : les ε_n sont indépendants

Les solutions sortent du cadre de ce cours. On peut utiliser :

— La randomisation avant l'expérience (voir module)

— Le modèle mixte ou les séries chronologiques.

Postulat 4 : les ε_n ont une distribution normale

— Chercher s'il existe des données aberrantes (ou outliers) expliquant le défaut de normalité

— On peut souvent faire quand même des tests si N et N-P sont grands.

Si rien ne marche

Si aucune transformation ne permet d'avoir des résidus satisfaisants, on peut encore utiliser :

— un modèle linéaire généralisé (voir module)

— un modèle non linéaire.