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

Les objectifs liés aux différents points sont les suivants

— Revoir les postulats et leur importance pour la loi des estimateurs

— Proposer des méthodes graphiques de vérification des postulats

— Proposer des méthodes pour *mieux* coller aux postulats

— Ouvrir sur les autres modules de FPSTAT

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 l'espérance des ε_n n'est pas nulle, μ_n n'est pas l'espérance de Y_n et la connaissance de θ ne permettra pas d'estimer ou de prédire Y_n.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-2

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

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

— 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 $\widehat{\theta}$

X est la matrice formée de colonnes indépendantes de la matrice du plan d'expérience. Cette matrice découle directement de l'écriture du modèle s'il est écrit sous une forme irréductible, elle est transformée par la recherche de colonnes indépendantes (ou par l'ajout de contraintes) si le modèle est sous forme réductible.

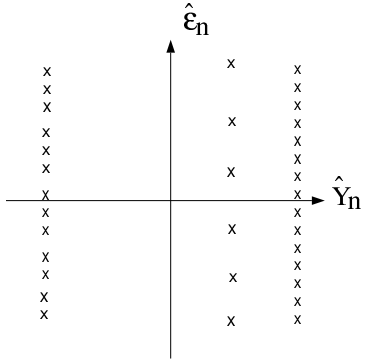
Si le postulat 4 n'est pas vérifié, $\widehat{\theta}$ est normal pour N grand et l'estimation de σ^2 reste valable si N-P est grand. Ce postulat est donc moins important que les trois autres.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-3

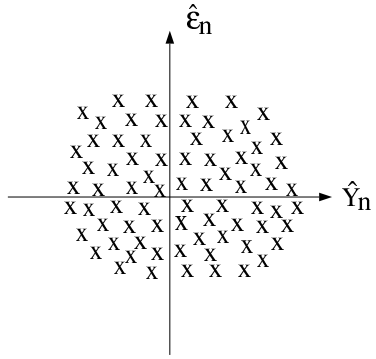
Vérification des postulats

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

En analyse de variance



En regression



Insister sur le fait que la vérification des postulats se fait *a posteriori* une fois les estimations calculées.

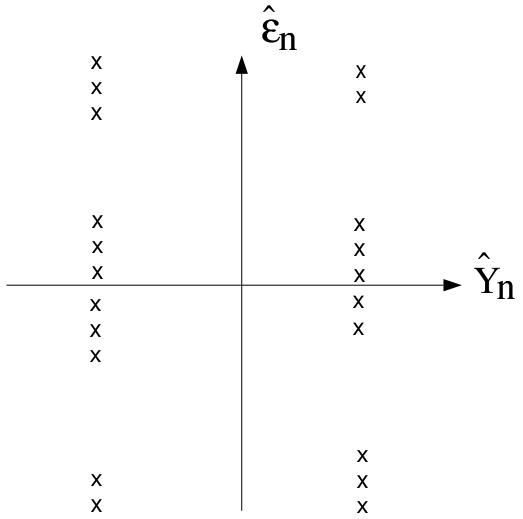
Les exemples présentés sur le transparent sont des exemples où les quatre postulats semblent à peu près vérifiés.

On ne dispose évidemment pas des ε_n et encore moins de leur distribution, mais uniquement pour chaque n d'un échantillon de taille 1 d'une estimation de ε_n ; nous supposerons que la distribution des ε_n est la même quel que soit n, et nous traiterons nos N échantillons de taille 1 comme un échantillon de taille N.

La variable de l'axe des x est le plus souvent la valeur prédite $\widehat{Y}_n$; mais elle pourra être autre chose en particulier dans le cas de la vérification de l'indépendance des ε_n.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-4

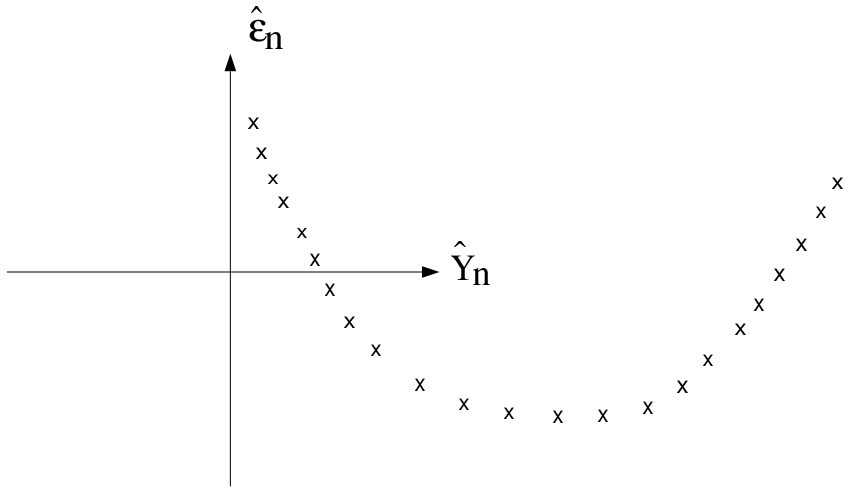
Le postulat $E(\varepsilon_n) = 0$ est-il vérifié ? en analyse de la variance



La structure particulière du graphe des résidus montre que le postulat $E(\varepsilon_n)$ n'est pas vérifié.

Il faut introduire un nouveau facteur à trois niveaux dont un contient tous les points où ε_n est négatif, un autre qui contient tous ceux où il est proche de zéro et enfin un troisième pour tous les points donnant un $\widehat{\varepsilon}_n$ positif.

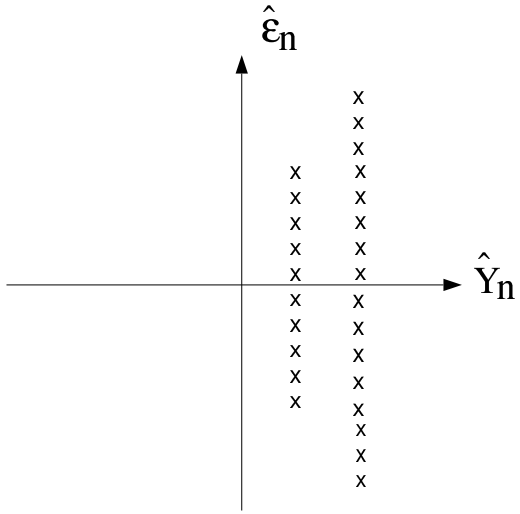
Le postulat $E(\varepsilon_n) = 0$ est-il vérifié ? en regression



On peut aussi refuser ce postulat en regression, si les résidus dont la moyenne est évidemment nulle sont distribués selon une structure particulière.

Nous avons ici une parabole ; il est clair que le résidu $\widehat{\varepsilon}_n$ dépend de la valeur du régresseur Z1 ; l'introduction du régresseur $Z1^2$ est nécessaire pour ne plus observer cette parabole.

Le postulat $\mathrm{Var}(\varepsilon_n) = \sigma^2$ est-il vérifié ? en analyse de la variance

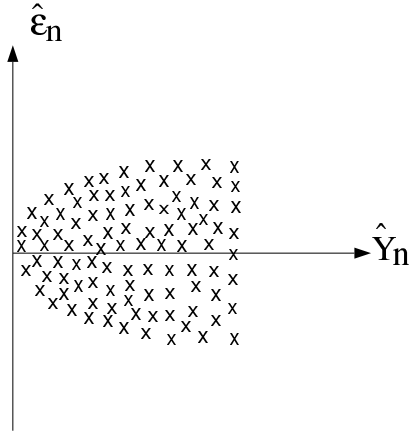


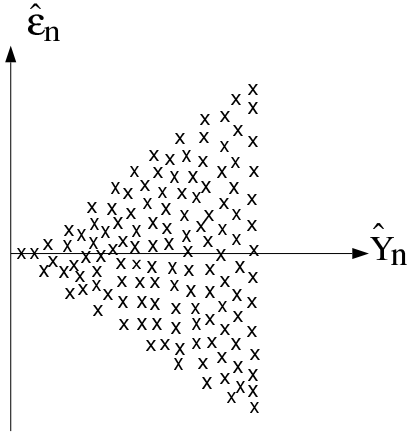
Analyse de variance à un facteur à deux niveaux

Il est clair que la variance n'est pas constante et dépend du niveau du facteur.

Cette situation est une situation très fréquente dans laquelle la variance du résidu croît avec sa moyenne.

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

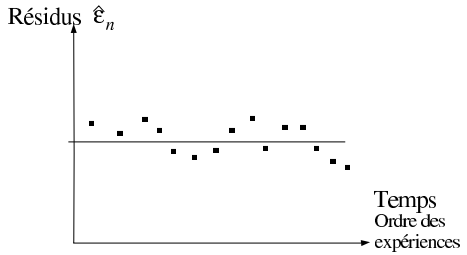




Il est clair que dans les deux cas, la variance dépend de la valeur prédite et que le postulat n'est pas vérifié.

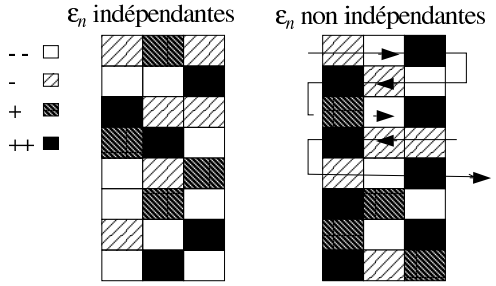
Les ε_n sont-ils indépendants ?

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



Il est très fréquent que la vérification de l'indépendance oblige à utiliser des variables non présentes dans l'analyse ; c'est le cas dans les cas présentés ici où les variables sont le numéro d'ordre des expériences ou la position dans un champ (paramètre bidimensionnel)

Le numéro d'ordre peut exprimer une évolution dans le temps ou une autre variable liée au temps (la position sur une droite dans le cas d'un transect).

En fonction de cette variable supplémentaire , on pourra faire un test des runs pour conclure si les suites observées sont aléatoires ou non. (voir module 1).

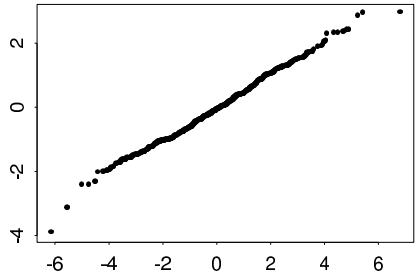
Dans le deuxième exemple de STATITCF ($\hat{\varepsilon}_n$ non indépendantes), les parcelles sont visitées dans l'ordre donné par les flèches : il apparaît clairement que les résultats vont en croissant dans chaque ligne.

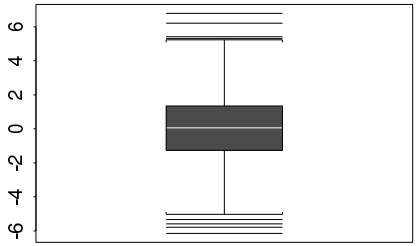
Il s'agit en réalité d'une moissonneuse-batteuse que l'on vide plus soigneusement à la fin de chaque rang qu'à la fin de chaque parcelle.

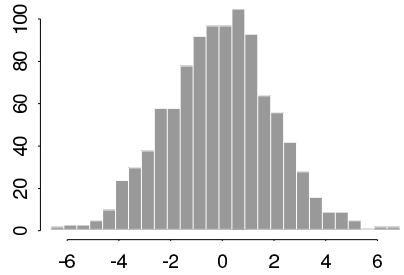
FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-9

Les ε_n sont-ils normalement distribués ?

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







Le postulat de normalité est délicat à vérifier en pratique ; mais heureusement, on peut souvent travailler sans qu'il soit vérifié.
Il existe trois types de représentations graphiques qui permettent de savoir si on est proche ou non de la distribution normale. Ce sont des méthodes graphiques empiriques.

— Qq plot ou droite de Henry :
On représente en abscisses les quantiles de la loi normale et en ordonnées les quantiles de la loi des $\widehat{\varepsilon}_n$. Si les points sont alignés le long d'une droite ou à peu près, on supposera que les ε_n suivent une loi normale.

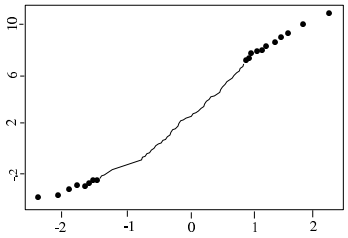
— Histogramme :
Les sommets doivent décrire approximativement une courbe de Gauss.

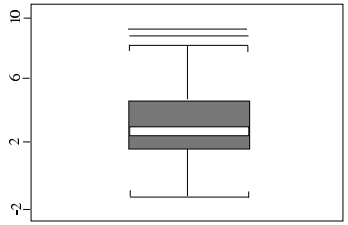
— Boite à pattes (en Anglais Box-plot) :
C'est un résumé de l'histogramme qui contient les valeurs extrêmes et les trois quantiles intérieurs.

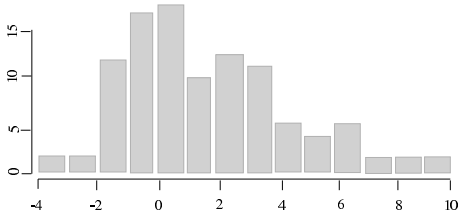
Les représentations graphiques du transparent correspondent à une distribution des $\boldsymbol{\varepsilon_n}$ à peu près normale.

Les ε_n sont-ils normalement distribués ?

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







On voit ici que les graphiques ne représentent pas une distribution normale

— La droite de Henry n'est pas une droite : elle est convexe au début et concave à la fin.

— L'histogramme montre une multimodalité, contrairement à celui d'une loi normale.

— La boîte à pattes montre que les données sont concentrées vers les valeurs faibles (ce que montre aussi l'histogramme).

Nos données proviennent en fait de la simulation d'un mélange de deux lois normales :

$$0.3N(5,2)+0.7N(0,2)$$

L'interprétation de ces représentations n'est pas toujours facile à faire :

— La courbe de Henry de ce transparent n'est pas très éloignée d'une droite et celle du transparent précédent quoique plus proche présente quelques irrégularités.

— De même, si l'histogramme de ce transparent est clairement non normal, celui du transparent précédent présente de réelles irrégularités ; une certaine habitude permet néanmoins d'éviter les erreurs d'interprétation.

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 : $E(\varepsilon_n) = 0$ **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 : $\mathrm{Var}(\varepsilon_n) = \sigma^2$ **constante**

La technique générale est de transformer la variable mesurée.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-12

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

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

Distribution de Poisson :

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

Distribution Log-normale :

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

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

A l'aide d'une transformation de Y, il est parfois possible de stabiliser la variance ; c'est à dire de rendre vrai le deuxième postulat.

La transformation doit être choisie en fonction des graphiques observés, et si l'on arrive à mettre en évidence une relation entre μ et σ, cette relation donnera des informations précieuses sur la transformation souhaitable.

Les distributions fréquemment rencontrées sont :

— Poisson : souvent le cas pour des variables correspondant à un comptage.

— Log normale : c'est le cas de variables correspondant à un phénomène de croissance multiplicative.

— Binomiale : c'est le cas lorsque l'on mesure la proportion d'individus pour un caractère qui prend les valeurs 0 ou 1.

Remarque :
Il faut être prudent en régression ; en effet une transformation de la variable peut obliger à faire une transformation des régresseurs .

Ainsi lorsque l'on remplace la régression $Y = \mu + \alpha Z$ par la régression $\mathrm{Log(Y)} = \nu + \beta Z$, on peut se trouver dans une situation où un seul régresseur n'est pas suffisant ; dans ce cas , on pourra ajouter un régresseur en Z^2 qui représentera le début du développement de Taylor de Log(Y)

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 $k = 1$, $\mathrm{Log}\, Y$

Exemples

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

Trouver k ?

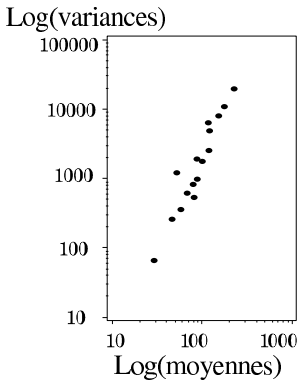
si $\sigma \propto \mu^k$, alors $\sigma^2 \propto \mu^{2k}$
d'où $\sigma^2 = c\,\mu^{2k}$
d'où $\mathrm{Log}\,\sigma^2 = c' + 2k\,\mathrm{Log}\,\mu$

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

Remarque :
La transformation arsin$\sqrt{Y}$ stabilise correctement la variance lorsque le dispositif expérimental est proche d'un dispositif équirépété.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-14

Exemple de choix d'une transformation de 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.

C'est en analyse de variance, que l'on utilise cette technique pour le choix d'une transformation.

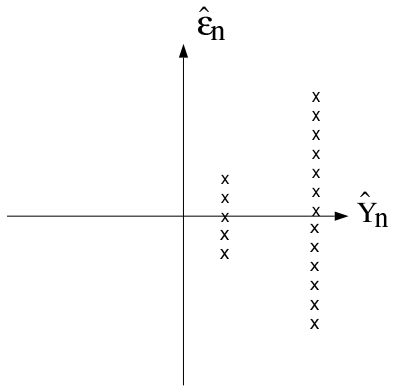
En fonction de la structuration des données (1 facteur, 2 facteurs, ...) on estime par niveau i de la structuration : $\widehat{\mu}_i$ et $\widehat{\sigma}_i = \frac{\sum_{r=1}^{n_i} \widehat{\varepsilon}_{ir}^2}{n_i}$.

La représentation graphique de $\mathrm{Log}(\widehat{\mu}_i)$ en fonction de $\mathrm{Log}(\widehat{\sigma}_i)$ permet d'estimer la pente de la droite.

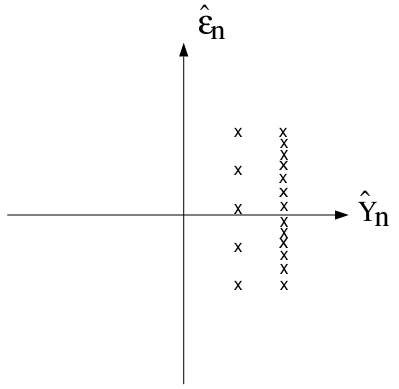
FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-15

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

résidus avant transformation



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



Après avoir appliqué une transformation de variable, on vérifie, avec les nouveaux résidus $\widehat{\varepsilon}_n$ représentés sur un graphique en fonction de $\widehat{Y}$ que la variance est bien stabilisée c'est à dire qu'il n'y a pas de structure particulière du nuage.

On constate ainsi qu'après avoir appliqué la transformation $Y' = \sqrt{Y}$ la structure observée disparaît à peu près complètement.

Une telle vérification est toujours nécessaire après une transformation.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-16

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.

FPSTAT INRA 1997 VII-Vérification des Postulats du Modèle Linéaire-17

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.