RAPPEL DES POSTULATS DU MODELE LINEAIRE

Modèle : $Y = X\theta + \varepsilon$

Postulats :

1. $\mathrm{E}(\varepsilon) = 0$
2. $\mathrm{var}(\varepsilon_i) = \sigma^2$ pour tout i
3. ε_i indépendantes
4. ε_i normalement distribuées (moins important)

L'utilisation du modèle linéaire suppose que ces quatre postulats soient vérifiés. Ces hypothèses sont contrôlables a posteriori. On utilise les résidus $\hat{\varepsilon}$ pour les contrôler.

Nous allons donc les passer en revue, en voyant pour chacun d'eux:

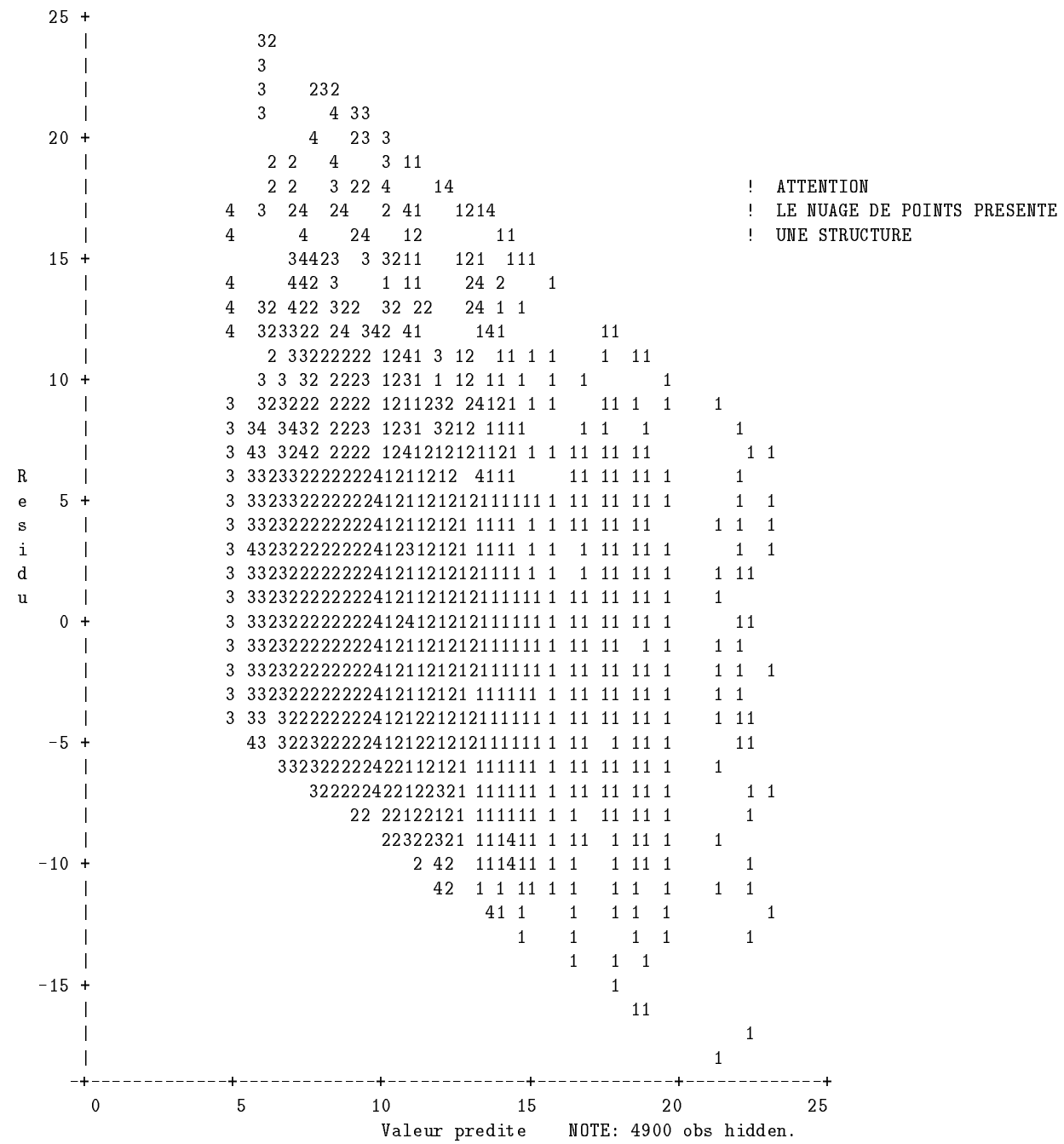
- ce qui se passe si le postulat n'est pas vérifié,
- les méthodes graphiques permettant de savoir s'il l'est,
- les remèdes éventuels.

Remarque : Le deuxième postulat, homogénéité des variances, peut être vérifié a priori sur la variable mesurée Y.

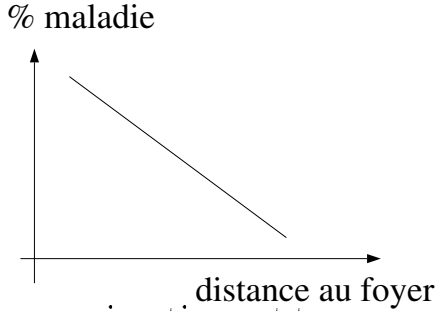
Premier postulat:
$E(\varepsilon) = 0$, c'est à dire que le modèle posé est correct. Si ce n'est pas vrai et que le modèle est faux, alors toute analyse basée sur un modèle faux sera bien évidemment fausse...

VERIFICATION $\mathrm{E}(\varepsilon) = 0$

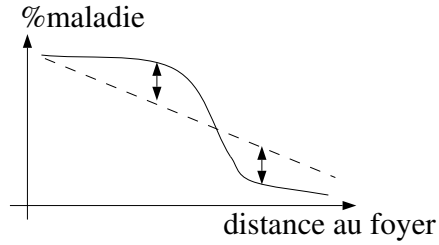
Représentation graphique $\hat{\varepsilon}$ en fonction de $\widehat{Y}$



Si on examine le graphique des résidus en fonction des valeurs prédites par le modèle, normalement on doit obtenir un nuage sans structure particulière.
Ici nous avons un exemple de nuage présentant une allure de droite de pente -1, avec des faibles valeurs plutôt surestimées par le modèle, et des fortes valeurs plutôt sous-estimées par le modèle. Dans ce cas le modèle posé n'était sûrement pas correct. Les mesures effectuées dans cet ensemble étaient des pourcentages de surface foliaire de blé attaquée par la rouille jaune, dans des parcelles où on avait introduit un foyer de maladie. Les mesures ont été réalisées en fin d'épidémie à différentes distances du foyer. Le modèle utilisé dans cette analyse supposerait que la relation entre maladie et distance au foyer était linéaire :



Cette approximation est trop grossière, on montre en fait en épidémiologie que les gradients de maladie ont la forme suivante :



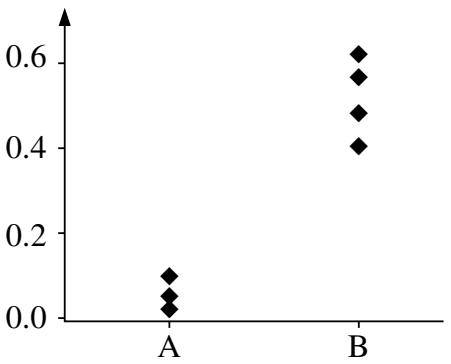
Si on ajuste la courbe à la droite en pointillés, on voit que l'on sous-estime la majeure partie des forts pourcentages de maladie et que l'on surestime les faibles, d'où la forme du nuage observé. Une transformation du régresseur "dist. au foyer" aurait pu améliorer le modèle posé. Dans le cas d'ANOVA, si le nuage présente une structure, réfléchir: n'aurait-on pas oublié un facteur ?

POSTULAT $\text{var}(\varepsilon_i) = \sigma^2$

Propriétés de l'estimateur des moindres carrés $\widehat{\theta}$ si ce postulat n'est pas vérifié:

- $\widehat{\theta} = (X'X)^{-1} X'Y$ gaussien
- $\text{E}(\widehat{\theta}) = \theta$ optimal parmi les estimateurs sans biais
- $\text{var}(\widehat{\theta}) = \sigma^2 \ (X'X)^{-1}$
- $SCR \sim \sigma^2 \chi^2_{(n-p)}$ tests de Fisher
- $CMR = \frac{SCR}{n-p}$ proche de σ^2 quand $n - p$ est grand

Exemple : taux de survie d'insectes à 2 insecticides A et B



Si ce n'est pas vrai, alors:

Rayez les points 3, 4 et 5

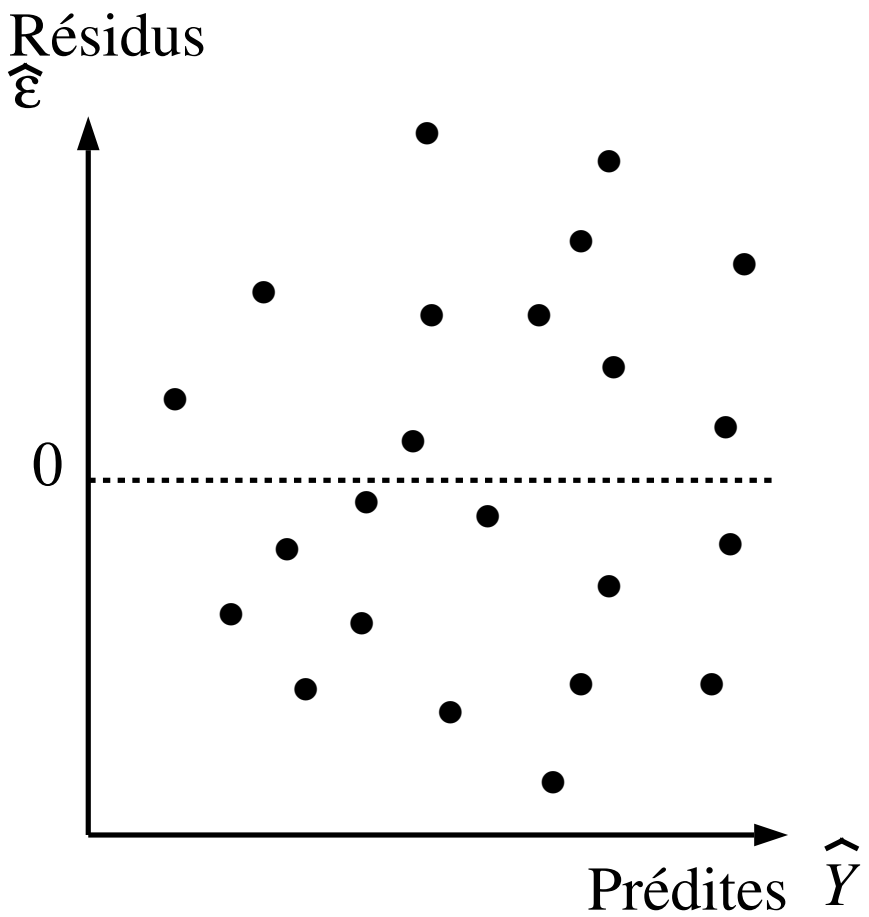
On a toujours un estimateur $\hat{\theta}$, sans biais, mais plus d'estimation de sa variance. Il n'est donc plus utilisable.
On ne peut plus faire de tests (ils reposent sur des lois dépendant de σ^2 qu'on ne sait plus calculer). Il est donc fondamental que ce postulat soit respecté.

Vérification:

Cela peut être évident, comme sur cet exemple de mesure des taux de survie d'insectes, suite à deux traitements insecticides A et B, dont l'un est beaucoup plus efficace que l'autre.

VERIFICATION $\text{var}(\varepsilon_i) = \sigma^2$

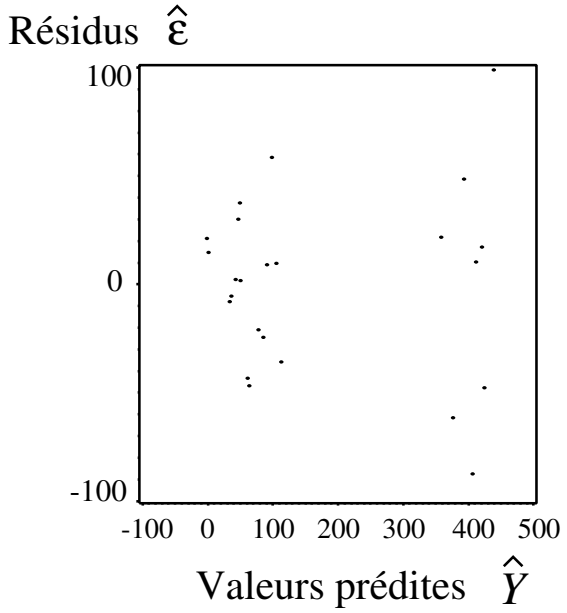
Représentation graphique $\hat{\varepsilon}$ en fonction de $\widehat{Y}$



Si la variance est homogène, on doit obtenir en faisant ce graphique un nuage qui ne présente aucune structure particulière.

VERIFICATION $\text{var}\,(\varepsilon_i) = \sigma^2$

Représentation graphique $\hat{\varepsilon}$ en fonction de $\widehat{Y}$



Ce nuage de points présente une structure:
$\text{var}\,(\varepsilon_i)$ n'est pas constante

Dans l'exemple suivant, la représentation graphique de $\hat{\varepsilon}$ en fonction de $\widehat{Y}$ montre que le postulat $var(\varepsilon_i) = \sigma^2$ n'est pas vérifié.

Cet exemple correspond à une expérience où l'on a dénombré des coquelicots dans des parcelles d'avoines traitées avec différents herbicides. On voit que la variance est plus faible pour les petites valeurs de $\widehat{Y}$, alors que la variance augmente quand $\widehat{Y}$ augmente.

TRANSFORMATIONS DE VARIABLES LORSQUE $\mathrm{var}(\varepsilon_i)$ N'EST PAS CONSTANTE

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 tranformation
- vérifier les résidus après transformation

$\propto$ signifie "proportionnel à".

Pour une transformation de Y, il est parfois possible de se rapprocher du postulat "les ε_i ont même variance".

Quelle transformation ?
Le choix doit être motivé, il ne se fait pas n'importe comment. La forme de la transformation dépend de la relation entre μ et σ. On peut avoir une idée a priori de la loi de Y.

- **Poisson:** souvent le cas de variables qui correspondent à des comptages.
- **log normale:** $\mu \propto \sigma$, c'est le cas par exemple lorsqu'on mesure des réponses résultant de processus de croissance ou de multiplication.
- **Binomiale:** les individus possèdent l'un ou l'autre de deux caractères opposés *(malade-sain)*.

  On mesure par exemple la proportion d'individus ayant le caractère.

TRANSFORMATIONS DE VARIABLES LORSQUE $\mathrm{var}(\varepsilon_i)$ N'EST PAS CONSTANTE

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 :

$k = 1$ $\quad \sigma \propto \mu$: $\quad Y \longrightarrow \mathrm{Log}\, Y$

$k = \frac{1}{2}$ $\quad \sigma \propto \mu^{1/2}$,

$\quad \sigma^2 \propto \mu$: $\quad Y \longrightarrow Y^{1-\frac{1}{2}} = \sqrt{Y}$

TROUVER k ?

si $\sigma \propto \mu^k$, alors $\sigma^2 \propto \mu^{2k}$

d'où $\sigma^2 = c\,\mu^{2k}$

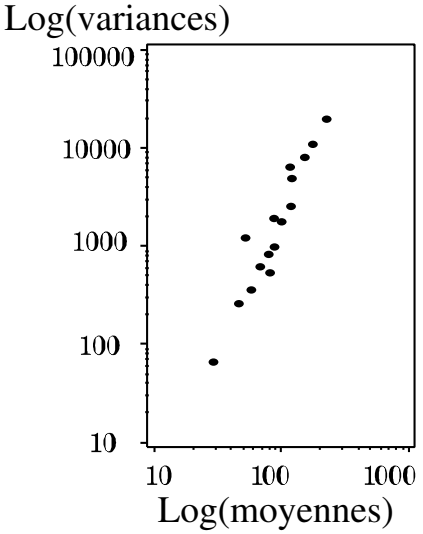
d'où $\mathrm{Log}\, \sigma^2 = c' + 2k\, \mathrm{Log}\, \mu$

Relation linéaire entre μ et σ^2, pente de la droite= $2k$.

On retrouve les transformations données précédemment.

EXEMPLE DE CHOIX D'UNE TRANSFORMATION DE VARIABLES LORSQUE var (ε_i) N'EST PAS CONSTANTE

Représentation graphique de Log σ^2 en fonction de Log μ



Comptages de mouches dans des pièges :

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

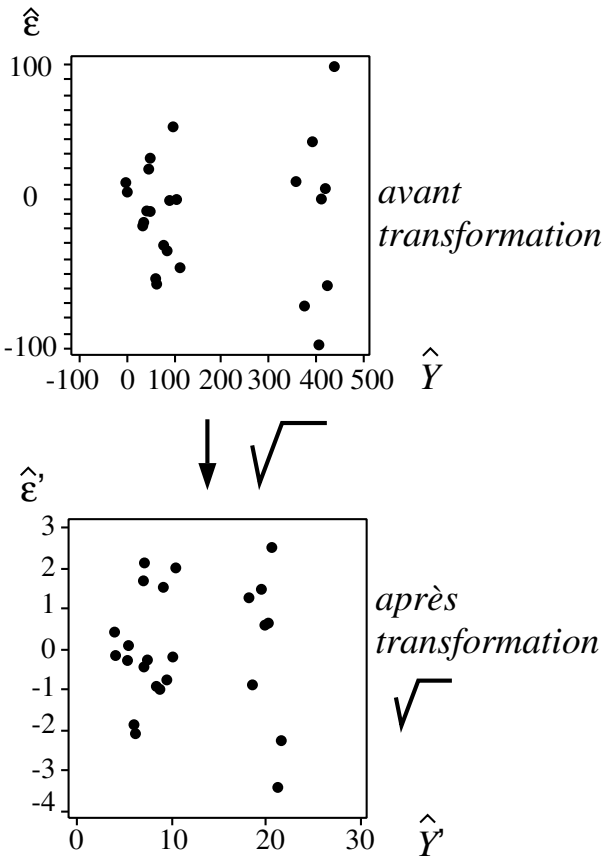
droite de pente très proche de 2

$\Longrightarrow$ transformation Log.

Remarque:
Il s'agissait de comptages, et la transformation pertinente est le Log, et non pas la racine carrée comme on aurait pu le penser si on applique automatiquement les transformations classiques présentées au transparent 11-6 (l'hétérogénéité des variances était ici très forte).

VERIFICATION $\mathrm{var}(\varepsilon_i) = \sigma^2$ APRES TRANSFORMATION DE VARIABLE

Représentation graphique de $\widehat{\varepsilon'}$ en fonction de $\widehat{Y'}$



Après avoir choisi et appliqué une transformation de variable, on vérifie sur la présentation graphique $\widehat{\varepsilon'}$ en fonction de $\widehat{Y'}$ que l'on a bien "stabilisé la variance", c'est à dire que le nuage obtenu ne présente plus de structure particulière.

L'exemple ci-contre reprend le nuage présenté précédemment, correspondant à des comptages de coquelicots dans des parcelles d'avoine. On vérifie bien qu'après transformation $\sqrt{\ }$, la structure observée sur le précédent graphique a pratiquement disparu.

POSTULAT ε_i INDEPENDANTES

Propriétés de l'estimateur des moindres carrés $\widehat{\theta}$ si ce postulat n'est pas vérifié :

- $\widehat{\theta} = (X'X)^{-1} X'Y$ gaussien
- $\mathrm{E}(\widehat{\theta}) = \theta$ optimal parmi les estimateurs sans biais
- $\mathrm{var}(\widehat{\theta}) = \sigma^2 \ (X'X)^{-1}$
- $SCR \sim \sigma^2 \chi^2_{(n-p)}$ tests de Fisher
- $CMR = \frac{SCR}{n-p}$ proche de σ^2 quand $n - p$ est grand

Si ce n'est pas vrai, alors:

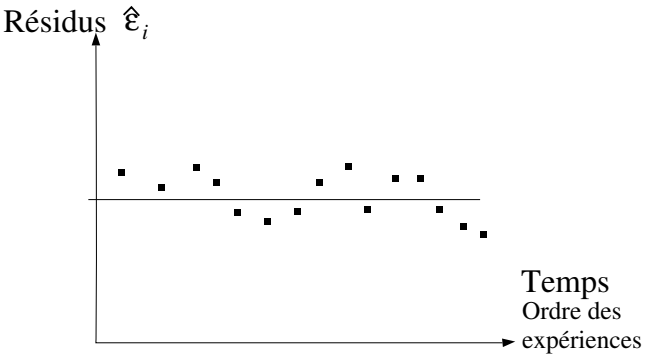
Rayez les points 3, 4 et 5

De la même façon que pour le postulat précédent, on perd la variance de $\hat{\theta}$ ainsi que l'estimation de σ^2 et les tests.

Il est donc indispensable que ce postulat soit vérifié.

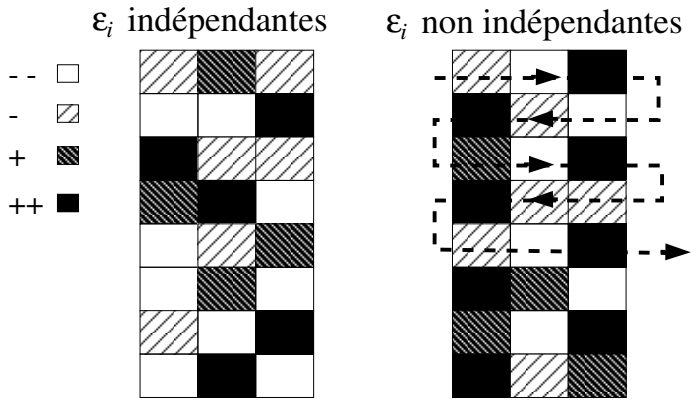
VERIFICATION ε_i INDEPENDANTES

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



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

Représentation graphique schématique des $\widehat{\varepsilon_i}$ dans l'espace (STATITCF)



Si le graphique soulève un doute sur l'indépendance, on peut faire un test de run (analyse du nombre de séquences positif-négatif, permettant de conclure que les séquences observées sont aléatoires ou non. Voir fiche module 1).

Sur le schéma du bas sont reportées les parcelles telles qu'elles sont disposées dans le champ. Les parcelles sont grisées d'autant plus foncé que les résidus sont élevés, ce qui permet de visualiser une répartition aléatoire ou non des résidus.

S'il n'y a pas indépendance et que les ε_i sont corrélées, il existe des méthodes de correction, mais qui sont complexes et dépassent le cadre de ce cours.
Toutefois, il faut remarquer que si l'on prend soin de randomiser correctement (voir module 4) cela permet de garantir une certaine structure de covariance pour les erreurs, qui n'est pas exactement l'indépendance, mais qui s'en rapproche suffisamment pour que l'utilisation du modèle linéaire soit possible.

POSTULAT ε_i NORMALEMENT DISTRIBUEES

Propriétés de l'estimateur des moindres carrés $\widehat{\theta}$ si ce postulat n'est pas vérifié :

- $\widehat{\theta} = (X'X)^{-1} X'Y$ gaussien — si n est grand
- $\mathrm{E}(\widehat{\theta}) = \theta$ optimal parmi les estimateurs ~~sans biais~~ — linéaires
- $\mathrm{var}(\widehat{\theta}) = \sigma^2 \ (X'X)^{-1}$
- $SCR \sim \sigma^2 \chi^2_{(n-p)}$ tests de Fisher
- $CMR = \frac{SCR}{n-p}$ proche de σ^2 quand $n - p$ est grand

Si ce n'est pas le cas,

corriger avec les $\boxed{\textit{encadrés}}$.

Si on a beaucoup de degrés de liberté résiduels ($n - p$ grand), σ^2 est correctement estimé, et les tests sont de niveau exact.

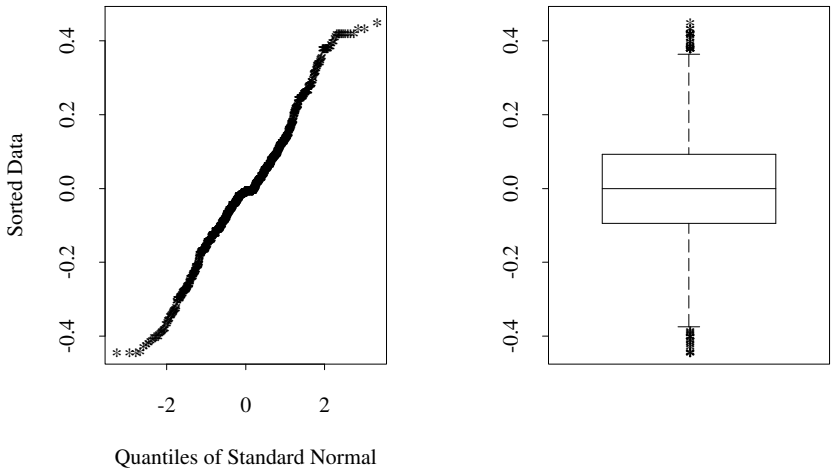
La normalité n'est donc pas indispensable, et en particulier si on a beaucoup de données, tout se passe comme si on avait la normalité.

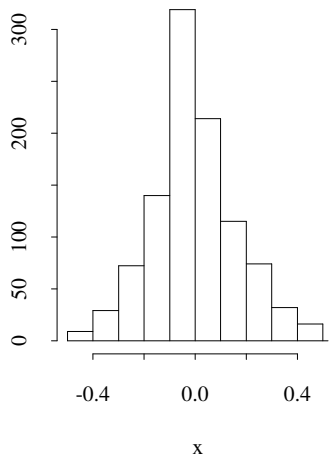
En pratique, l'analyse reste correcte même dans le cas de petits échantillons ($n = 10$).

$\hat{\theta}$ est optimal parmis les estimateurs linéaires (et non plus sans biais): cela signifie que si on connaissait précisemment la vraie loi, on aurait pu construire un estimateur plus performant.
Cependant, cela ne change rien aux autres propriétés de $\hat{\theta}$

VERIFICATION ε_i NORMALEMENT DISTRIBUEES

Représentation graphique de $\hat{\varepsilon}$





Le postulat de normalité est le plus délicat à vérifier en pratique. Fort heureusement beaucoup de propriétés restent vraies sans ce postulat.

Les trois types de représentations graphiques suivants permettent de savoir si l'on est éloigné ou non de la distribution normale.

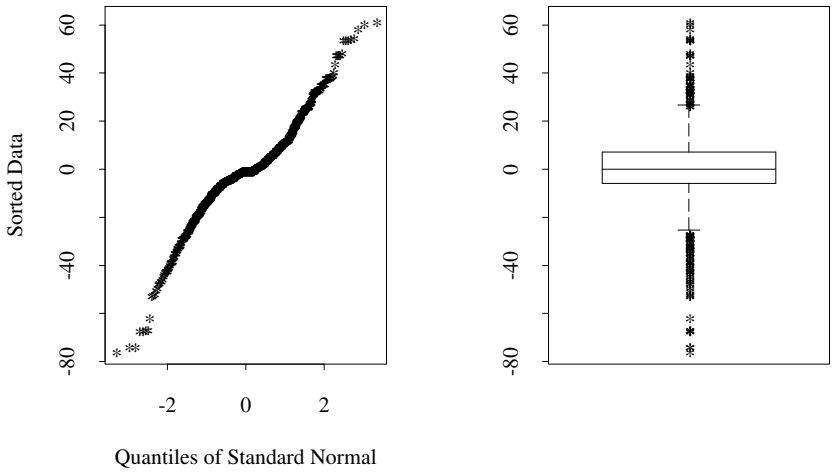
Qq plot ou *droite de Henry*:
En abscisse, les quantiles de la loi normale, en ordonnées, les quantiles de la loi des ε_i.
Si les $\varepsilon_i \sim \mathcal{N}$, on doit obtenir une droite, qui est à peu près correcte sur cet exemple.

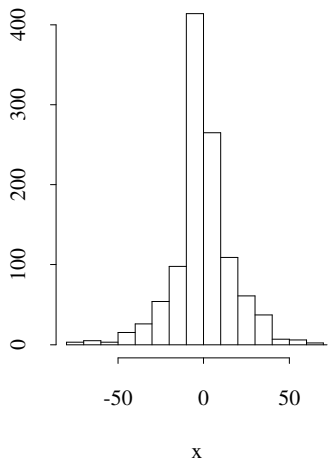
Il faut compléter cette représentation par une ***boite à pattes*** et un *histogramme*, qui sont également ici tout à fait corrects.

On peut conclure pour cet exemple que les ε_i sont normalement distribuées.

VERIFICATION ε_i NORMALEMENT DISTRIBUEES

Représentation graphique de $\hat{\varepsilon}$





Ici grâce aux trois graphiques sur les $\widehat{\varepsilon_i}$, on voit qu'on ne peut accepter l'hypothèse de la normalité pour les ε_i:

- Qq plot: on n'obtient pas une droite de pente 1
- Box plot et histo: distribution symétrique mais queues de distribution très étalées.

Remarque: Souvent lorsqu'on utilise une transformation de variable Y', pour que le postulat $var(\varepsilon_i') = \sigma^2$ soit vérifié, alors les ε_i' sont normalement distribués (alors que les ε_i ne l'étaient pas).

Remarque: **Résidus suspects**

Si les erreurs sont normalement distribuées, de variance σ^2 et centrées sur 0, alors les résidus réduits (divisés par σ) suivent une $\mathcal{N}(0,1)$, donc leurs valeurs, à 5 % près, doivent être approximativement comprises entre -2 et $+2$.

si on détecte des individus "suspects" avec des valeurs absolues de résidus réduits très supérieurs à 2:

- retouner aux données brutes:
  - erreur de mesure,
  - de transcription,
  - etc...

  $\hookrightarrow$ correction éventuelle.
- s'il n'y a pas d'erreur $\Rightarrow$ décision:
  - justification objective $\longrightarrow$ supprimer. (animal malade, parcelle inondée, tube contaminé, etc ...)
  - pas de justification $\longrightarrow$ conserver.