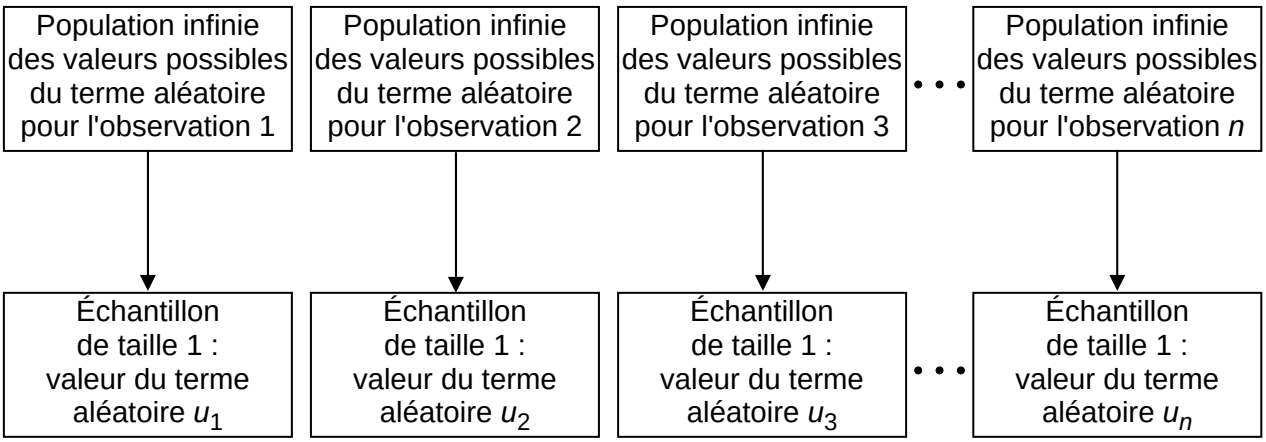
POPULATION(S) ET ÉCHANTILLON(S) DANS L'ANALYSE DE RÉGRESSION

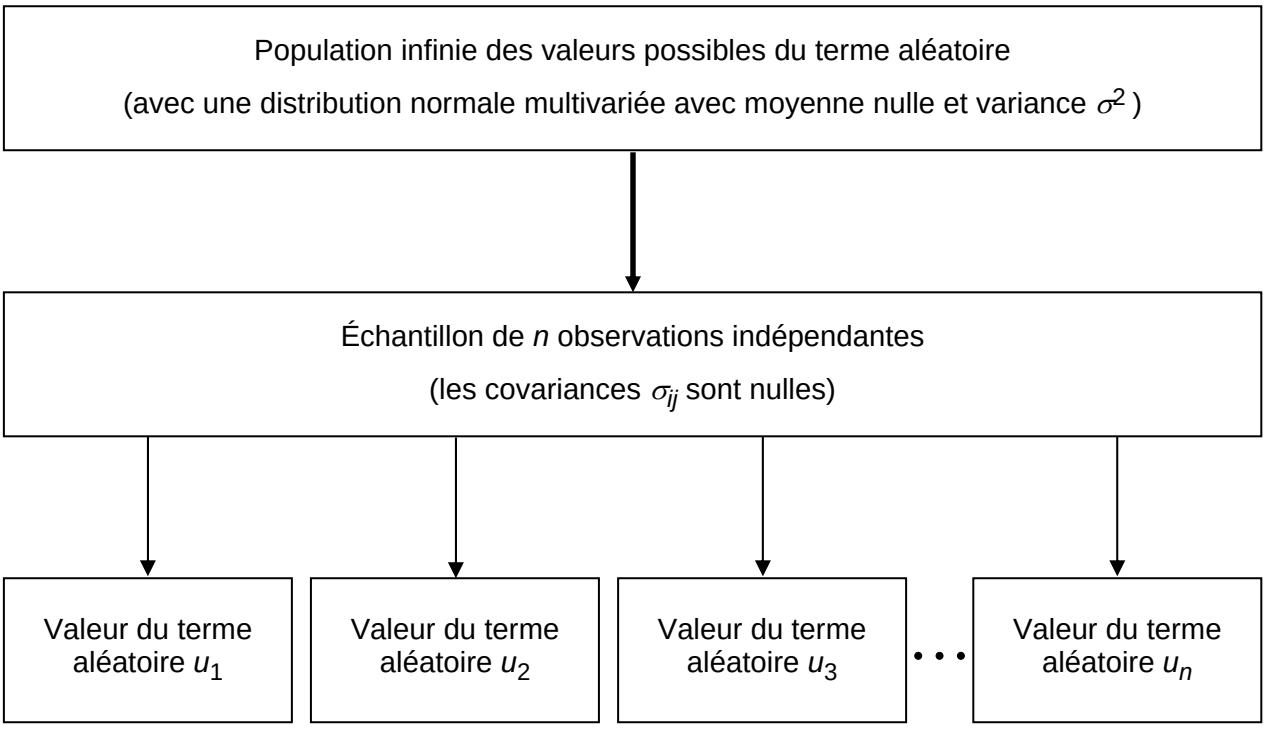
Cas général :

n échantillons de taille 1, tirés de *n* populations, une par observation



Modèle d'échantillonnage classique de la régression linéaire :

1 échantillon de *n* observations, tirés d'une même population



LE MODÈLE D'ÉCHANTILLONNAGE CLASSIQUE DE LA RÉGRESSION LINÉAIRE NORMALE IMPLIQUE NOTAMMENT QUE...

1. L'estimateur b_j du paramètre β_j est *non biaisé* :
   b_j a une distribution d'échantillonnage avec moyenne égale à β_j .
2. Il existe aussi un estimateur non biaisé de la variance d'échantillonnage $\sigma_{b_j}^2$:
   $s_{b_j}^2 =$ valeur estimée de la variance d'échantillonnage $\sigma_{b_j}^2$.
3. La variable $\dfrac{b_j - \beta_j}{s_{b_j}}$ a une distribution de Student avec $n - k$ degrés de liberté.

TEST D'UNE HYPOTHÈSE SIMPLE SUR LA VALEUR D'UN COEFFICIENT (TEST DE STUDENT BILATÉRAL)

Forme générale de l'hypothèse à tester

$H_0 : \beta_j = c$

Exemple de test bilatéral

$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$

$H_0 : \beta_2 = 1$

Si H_0 est vraie, alors le taux d'urbanisation (*PURB*/*PTOT*) est indépendant de la population totale *PTOT*, puisque

$$\ln\left(\frac{PURB}{PTOT}\right) = \ln PURB - \ln PTOT = \beta_1 + (\beta_2 - 1)\ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Variable-test : $t_{n-k} = \dfrac{b_j - c}{s_{b_j}}$

Noter l'analogie avec la variable-test de la moyenne :

$t_{n-1} = \dfrac{m_x - \gamma}{\left(s_x / \sqrt{n}\right)}$, où $\left(s_x / \sqrt{n}\right)$ est l'écart type d'échantillonnage de la moyenne.

Modèle aléatoire : Modèle classique de régression linéaire normale

La variable

$$\frac{b_j - \beta_j}{s_{b_j}}$$

a une distribution de Student avec $n - k$ degrés de liberté

Calcul de la valeur de la variable-test :

$t_{n-k} = \dfrac{0{,}971663 - 1}{0{,}0279321} = -1{,}0145$, avec $n - k = 64 - 4 = 60$

Probabilité critique (test bilatéral) = 0,314

Décision : normalement, non-rejet

TEST D'UNE HYPOTHÈSE SIMPLE SUR LA VALEUR D'UN COEFFICIENT (TEST DE STUDENT)

Test de l'hypothèse d'un coefficient nul

$H_0 : \beta_j = 0$

Exemple :

$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$

$H_0 : \beta_4 = 0$

Si on rejette H_0, cela signifie qu'il faut inclure la variable $(\ln GNPC)^2$ dans le modèle.

Variable-test : $t_{n-k} = \dfrac{b_j - c}{s_{b_j}} = \dfrac{b_j}{s_{b_j}}$ quand $c = 0$

Modèle aléatoire : Modèle classique de régression linéaire normale

Calcul de la valeur de la variable-test : $t_{64-4} = \dfrac{-0{,}0453}{0{,}01345} = -3{,}368$, avec $n - k = 64 - 4 = 60$

Probabilité critique (test bilatéral) = 0,0013

Décision : normalement, rejet; il faut conserver la variable $(\ln GNPC)^2$ dans le modèle.

Test unilatéral d'une hypothèse simple sur la valeur d'un coefficient

Exemple :

$PLAR_i = K\, PURB_i^h$

$\ln PLAR_i = \ln K + h \ln PURB_i$

Hypothèse à tester : $H_0 : h = 1$

Hypothèse complémentaire : $H_A : h < 1$ (rejeter H_0 = accepter H_A)

H_A signifie que l'importance relative de la plus grande ville $\dfrac{PLAR}{PURB}$ est moindre dans les pays où la population urbaine *PURB* est plus grande.

Variable-test : $t_{n-k} = \dfrac{b_j - c}{s_{b_j}}$

Modèle aléatoire : Modèle classique de régression linéaire normale

Calcul de la valeur de la variable-test :

$$t_{64-2} = \frac{h-1}{s_h} = \frac{0{,}636 - 1}{0{,}0426} = -8{,}54$$

Probabilité critique (test unilatéral) < 0,0001

Décision : normalement, rejet

INTERVALLES DE CONFIANCE ET MARGES D'ERREUR QUANT À LA VALEUR D'UN COEFFICIENT (TEST DE STUDENT)

Intervalles de confiance et marges d'erreur

Intervalles de confiance avec un niveau de confiance de $(1-\alpha)$:

$$b_j - s_{bj}\ \theta_{n-k}(\alpha) < \beta_j < b_j + s_{bj}\ \theta_{n-k}(\alpha)$$

Marge d'erreur avec un niveau de confiance de $(1-\alpha)$:

$$\pm\, s_{bj}\ \theta_{n-k}(\alpha)$$

Exemple :

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Degrés de liberté = $n - k = 60$

Niveau de confiance = 0,99

$\theta_{64}(0{,}01) = 2{,}66 \Rightarrow$ **Valeurs critiques du *t* de Student** : –2,66 et +2,66

$b_j = -0{,}0453$ et $s_{bj} = 0{,}01345$

Intervalle de confiance de β_4 à 99 % :

$$-0{,}0453 - 0{,}01345 \times 2{,}66 < \beta_4 < -0{,}0453 + 0{,}01345 \times 2{,}66$$

$$-0{,}0811 < \beta_4 < -0{,}0095 \rightarrow$$ Noter que la valeur $\beta_4 = 0$ ne fait **pas** partie de l'intervalle

Marge d'erreur de β_4, à un niveau de confiance de 99 % :

$$\pm\, 0{,}01345 \times 2{,}66 = \pm\, 0{,}358$$

Les intervalles de confiance se déduisent ici exactement comme dans le cas d'un test d'hypothèse simple sur la moyenne : l'ensemble des hypothèses qui ne seraient pas rejetées à un niveau de signification de α est donné par

$$-\theta_{n-k}(\alpha) < \frac{b_j - c}{s_{b_j}} < +\theta_{n-k}(\alpha)$$

$$-\theta_{n-k}(\alpha)\ s_{b_j} < (b_j - c) < +\theta_{n-k}(\alpha)\ s_{b_j}$$
$$-b_j - \theta_{n-k}(\alpha)\ s_{b_j} < -c < -b_j + \theta_{n-k}(\alpha)\ s_{b_j}$$
$$+b_j + \theta_{n-k}(\alpha)\ s_{b_j} > +c > +b_j - \theta_{n-k}(\alpha)\ s_{b_j}$$
$$b_j - \theta_{n-k}(\alpha)\ s_{b_j} < c < b_j + \theta_{n-k}(\alpha)\ s_{b_j}$$

TEST D'UNE OU DE PLUSIEURS RELATIONS LINÉAIRES ENTRE DES COEFFICIENTS (TEST *F* DE FISHER)

Hypothèse de plusieurs coefficients nuls

Deux modèles rivaux :

$\ln PLAR = p' + q' \ln PTOT + r' \ln GNPC + t' (\ln GNPC)^2 + s \ln PURB$

$\ln PLAR = \ln K + h \ln PURB$

Trois hypothèses :

$H_1 : q' = 0$

$H_2 : r' = 0$

$H_3 : t' = 0$

Si ces trois hypothèses sont vraies, le second modèle est équivalent au premier

... avec $p' = \ln K$ et $s = h$.

Test *F* que les trois hypothèses sont vraies *simultanément* :

probabilité critique = 0,53 (avec 64 observations)

Décision : on ne peut pas rejeter le second modèle.

Une relation linéaire entre plusieurs coefficients

La fonction de production Cobb-Douglas

$Y = A K^B T^C$

$\log Y = \log A + B \log K + C \log T$

Hypothèse des rendements constants à l'échelle :

$H_0 : B + C = 1$

$Y = A (\lambda K_0)^B (\lambda T_0)^C = \lambda^{B+C} A K_0^B T_0^C = \lambda^{B+C} Y_0$

Si $B + C = 1$,

$Y = \lambda Y_0$

TEST D'UNE OU DE PLUSIEURS RELATIONS LINÉAIRES ENTRE DES COEFFICIENTS (TEST F DE FISHER)

En général

Le test de Fisher permet de tester toute hypothèse que l'on peut exprimer par une ou plusieurs contraintes linéaires sur les coefficients, de la forme :

$$\sum_{j=1}^{k} w_j \beta_j = w_1 \beta_1 + w_2 \beta_2 + \cdots + w_k \beta_k = c$$

Le *F* des logiciels

Test des (k – 1) hypothèses simultanées

$H_0 : \beta_2 = \beta_3 = ... \beta_k = 0$

C'est-à-dire test de l'hypothèse que le «vrai» $R^2 = 0$

SPÉCIFICATION D'UN MODÈLE ALÉATOIRE : LES CONDITIONS DU MODÈLE CLASSIQUE DE RÉGRESSION LINÉAIRE

H1. Pour chaque observation, la valeur du terme aléatoire est tirée d'une population théorique de moyenne nulle :
$E(u_i) = 0$ pour tous les i

H2a) Pour toutes les observations, les populations théoriques d'où sont tirées les valeurs des termes aléatoires ont la même variance :
$\sigma_i^2 = \sigma^2$ pour tous les i

H2b) Pour chaque observation, la valeur du terme aléatoire est statistiquement indépendante des valeurs des termes aléatoires des autres observations :
$\sigma_{ij} = 0$ pour toutes les combinaisons i,j où $i \neq j$

H3. Les variables indépendantes x_{ij} sont non aléatoires (en particulier mesurées sans erreur).

H4. Il y a moins de paramètres à estimer qu'il y a d'observations et il n'y a pas de redondance parmi les variables indépendantes.

PROPRIÉTÉS DE L'ESTIMATEUR DES MOINDRES CARRÉS DANS LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE : LE THÉORÈME DE GAUSS-MARKOV

1. L'estimateur des moindres carrés de β_j est non biaisé :
   la moyenne de la distribution d'échantillonnage de b_j est égale à β_j.
2. $\dfrac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \dfrac{SSR}{n-k}$ est un estimateur non biaisé de σ^2.
3. La méthode des moindres carrés produit aussi un estimateur *non biaisé* de la variance d'échantillonnage $\sigma_{b_j}^2$ de chacun des coefficients estimés b_j, et de la covariance d'échantillonnage de chaque paire de coefficients estimés, $\sigma_{b_j b_h}$.
4. L'estimateur des moindres carrés est l'estimateur linéaire qui a la plus grande efficacité relative (la plus petite variance d'échantillonnage) : il est *BLUE* («Best Linear Unbiased Estimate»).

LES CONDITIONS DU MODÈLE CLASSIQUE DE RÉGRESSION LINÉAIRE *NORMALE*

H5. Chacun des termes aléatoires a une distribution normale.

NOTE DE TERMINOLOGIE :
ÉCART TYPE, ERREUR TYPE, ETC.

1. Définitions générales

- **Écart type** de la distribution d'une variable aléatoire *v* dans la **population** :

  $\sigma_v = \sqrt{\sigma_v^2}$, avec $\sigma_v^2 = \frac{1}{n}\sum_i (v_i - \mu_v)^2$ (population finie)

  $\sigma_v^2 = E\left[(v_i - \mu_v)^2\right]$ (population finie ou infinie)

- **Écart type** d'un **échantillon**

  = écart type d'une variable *v* dans un échantillon,

  = écart type de la distribution d'une variable *v* dans un échantillon

  $s_v = \sqrt{s_v^2}$, avec $s_v^2 = \frac{1}{n-1}\sum_i (v_i - m_v)^2$

  s_v^2 est un estimateur non biaisé de σ_v^2

2. Application à une statistique et à sa distribution d'échantillonnage

Nous considérons désormais que

v est une **statistique** calculée sur les données d'un **échantillon**

Alors,

μ_V est la **moyenne** (l'espérance mathématique) de la **distribution d'échantillonnage** de *v*.

σ_v^2 est la **variance de la distribution d'échantillonnage** de *v* , autrement dit la variance d'échantillonnage de *v*.

σ_V est l'**écart type de la distribution d'échantillonnage** de *v*,
autrement dit l'**écart type d'échantillonnage** de *v*
ou l'**erreur type** de *v*
ou l'**erreur d'échantillonnage** (*sampling error*) de *v*.
Toutes ces expressions sont synonymes.

NOTE DE TERMINOLOGIE : ÉCART TYPE, ERREUR TYPE, ETC. (SUITE)

3. Estimateurs de l'erreur d'échantillonnage

- Erreur d'échantillonnage = paramètre inconnu de la distribution d'échantillonnage de la statistique v.
- Estimateur généralement dénoté s_v
- s_v = « l'estimateur de l'erreur d'échantillonnage » ou « la valeur estimée de l'erreur d'échantillonnage ».

4. Exemples

4.1 ERREUR D'ÉCHANTILLONNAGE DE LA MOYENNE

La moyenne de l'échantillon m_v a une distribution d'échantillonnage normale, avec une moyenne égale à la vraie moyenne μ_v et un écart type égal à $\sqrt{\frac{\sigma_v^2}{n}} = \frac{\sigma_v}{\sqrt{n}}$: c'est l'erreur d'échantillonnage de la moyenne.

L'estimateur de cette erreur d'échantillonnage est donné par $\sqrt{\frac{s_v^2}{n}} = \frac{s_v}{\sqrt{n}}$.

Structure du *t* de Student :

$$t \text{ de Student} = \frac{\text{Différence entre la moyenne de l'échantillon et la moyenne hypothétique de sa distribution d'échantillonnage}}{\text{Erreur d'échantillonnage estimée de la moyenne}}$$

NOTE DE TERMINOLOGIE : ÉCART TYPE, ERREUR TYPE, ETC. (FIN)

4.2 ERREUR D'ÉCHANTILLONNAGE DES COEFFICIENTS ESTIMÉS DANS LE MODÈLE CLASSIQUE DE LA RÉGRESSION LINÉAIRE NORMALE

La méthode des moindres carrés ordinaires produit

- des estimateurs b_j pour les coefficients β_j, [1];
- un estimateur s^2 pour la variance des termes d'erreur σ^2; cet estimateur est calculé au moyen de la formule

$$s^2 = \frac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \frac{SSR}{n-k} \; ;$$

- des estimateurs $s_{b_j}^2$ de la variance d'échantillonnage $\sigma_{b_j}^2$. des b_j [2].

Théorème de Gauss-Markov

Sous les hypothèses du modèle classique de la régression linéaire normale, les estimateurs b_j ont une distribution d'échantillonnage normale, avec une moyenne égale à β_j et une variance égale à $\sigma_{b_j}^2$:

σ_{b_j} est l'**erreur d'échantillonnage** de b_j et s_{b_j} est l'estimateur de σ_{b_j}.

Test* t *de Student

$\dfrac{b_j - \beta_j}{s_{b_j}}$ a une distribution de Student avec $n - k$ degrés de liberté.

Structure du *t* de Student :

$$t \text{ de Student} = \frac{\begin{array}{c}\textit{Différence entre la valeur estimée du coefficient et} \\ \textit{la moyenne hypothétique de sa distribution d'échantillonnage}\end{array}}{\textit{Erreur d'échantillonnage estimée du coefficient}}$$

---

1 En écriture matricielle, on a $\mathbf{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$.

2 La matrice estimée des variances-covariances est donnée par $(\boldsymbol{X'X})^{-1} s^2$.

ANALYSE DES RÉSIDUS
ERREUR DE SPÉCIFICATION (1)

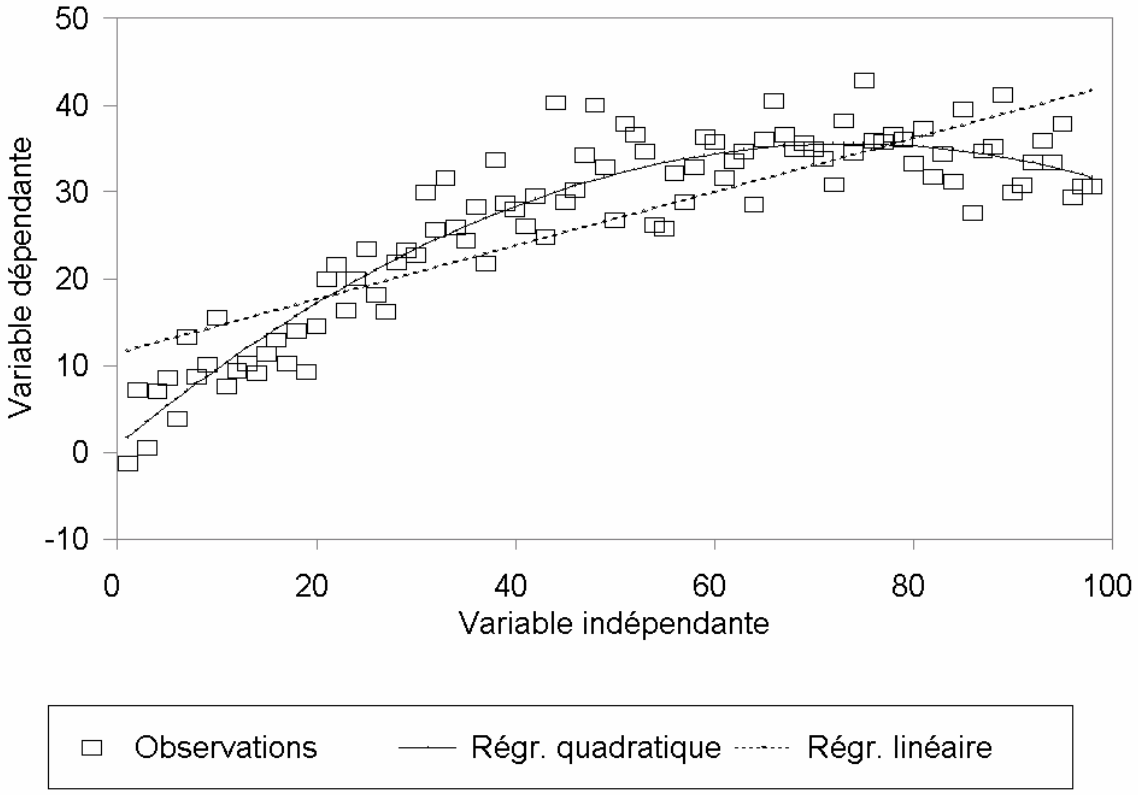
« Vrai » modèle :

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + u_i$$

Modèle incomplet :

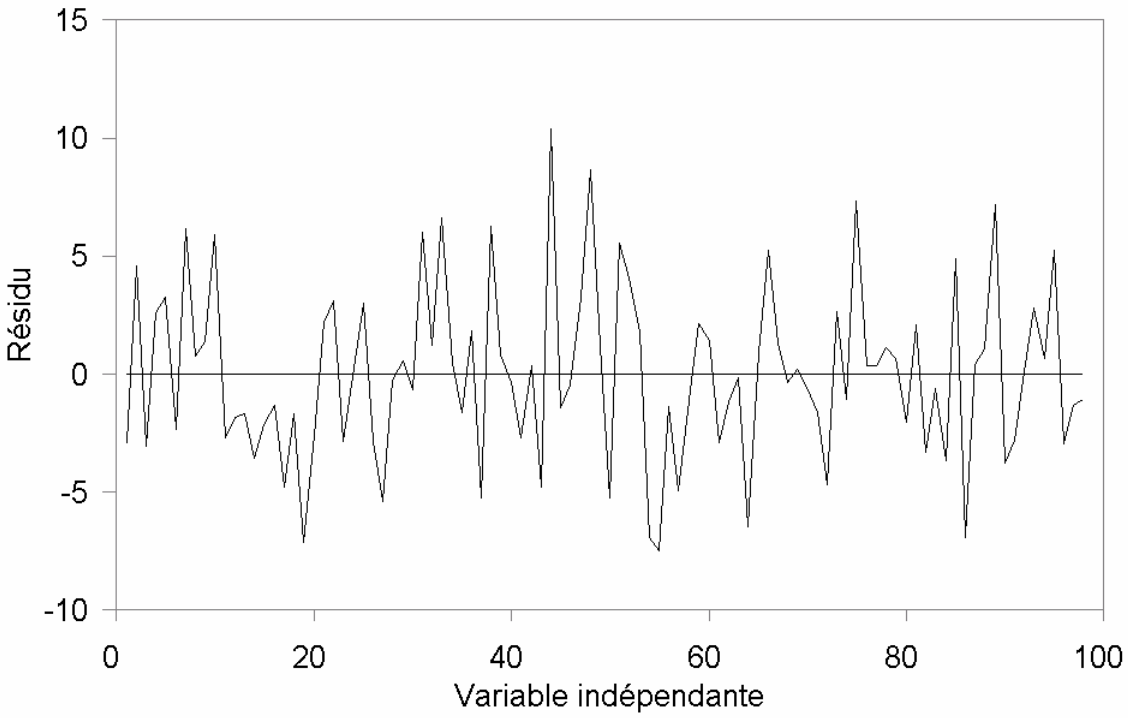
$$y_i = \alpha_0 + \alpha_1 x_i + u_i$$

Figure 6 - Observations et régressions
Une relation quadratique



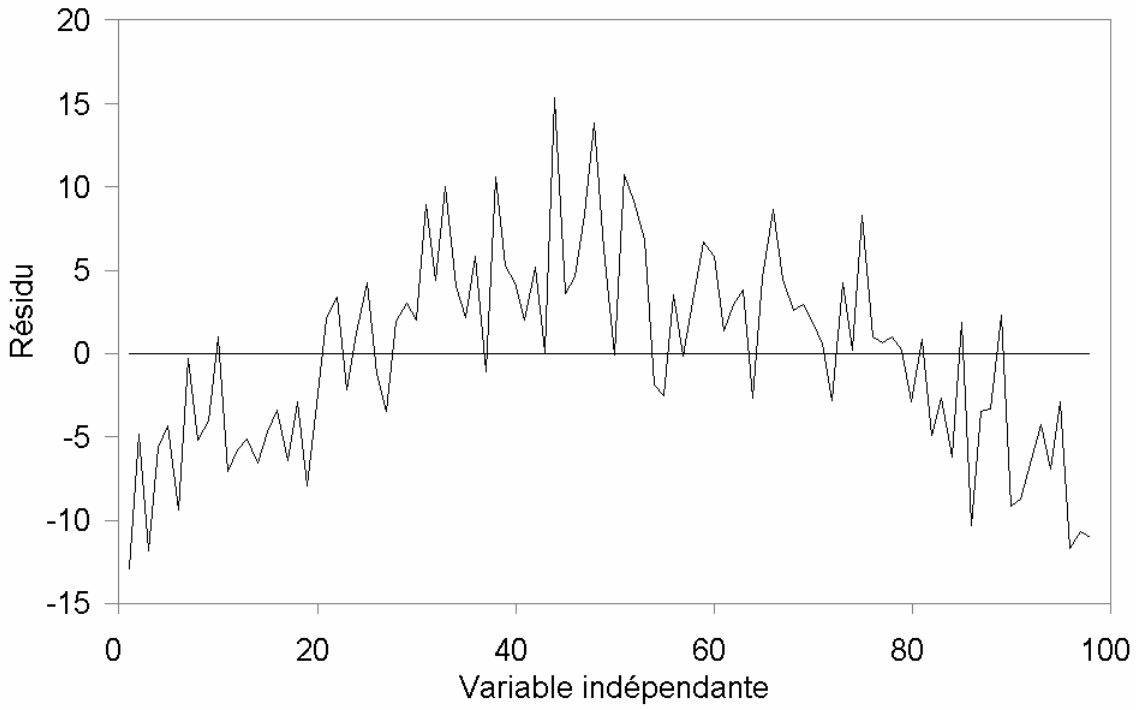
ANALYSE DES RÉSIDUS
ERREUR DE SPÉCIFICATION (2)

Figure 7 - Résidus régress.quadratique
Résidus sans erreur de spécification



ANALYSE DES RÉSIDUS
ERREUR DE SPÉCIFICATION (3)

Figure 8 - Résidus régression linéaire
Résidus avec erreur de spécification



ANALYSE DES RÉSIDUS
AUTOCORRÉLATION DES TERMES ALÉATOIRES (1)

Non respect de l'hypothèse H2b du modèle classique de la régression linéaire : $\sigma_{t,t-1} \neq 0$

Les données sous-jacentes aux résidus de régressions présentés aux figures 9 et 10 ont été générées au moyen de l'équation

$$y_t = x_t + 10\ \eta_t$$

où η_t a été généré au moyen d'un processus autorégressif de la forme

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$$

ε_t ayant une distribution proche de la normale [3]. Les figures sont répétées trois fois, avec trois valeurs différentes du paramètre α : 0,9, 0,6 et 0.

CONSÉQUENCES

- Forte variance d'échantillonnage : estimateurs moins précis
- Formules d'estimation de la variance des estimateurs sous-estiment la vraie variance : illusion de plus de précision
- Les tests statistiques ne sont plus valides.

TEST DE DÉTECTION

- Le plus connu est le **test de Durbin-Watson**

REMÈDES

- compléter le modèle en posant des hypothèses sur le mécanisme d'autocorrélation
- utiliser la méthode des ***moindres carrés généralisés***.

---

3 Plus exactement, il s'agit de la transformation logistique d'une variable ayant une distribution uniforme entre zéro et un.

ANALYSE DES RÉSIDUS
AUTOCORRÉLATION DES TERMES ALÉATOIRES (2) : α=0,9

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

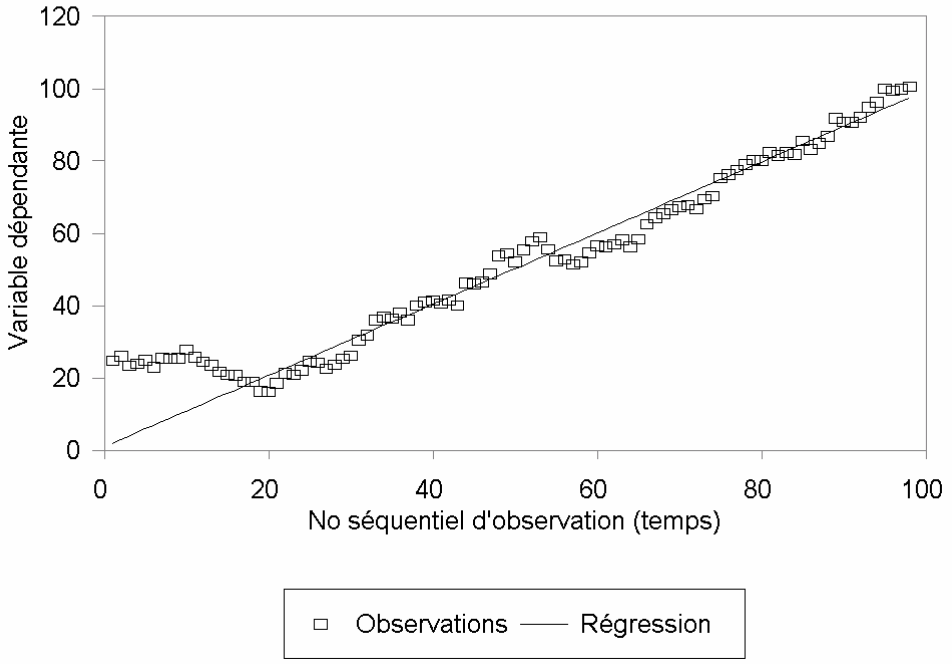
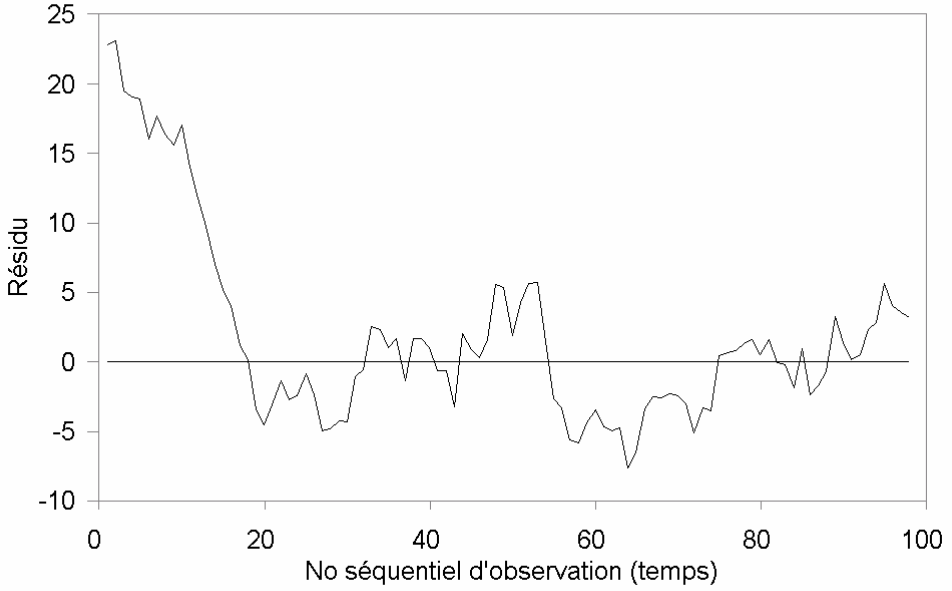


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



ANALYSE DES RÉSIDUS
AUTOCORRÉLATION DES TERMES ALÉATOIRES (3) : α=0,6

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

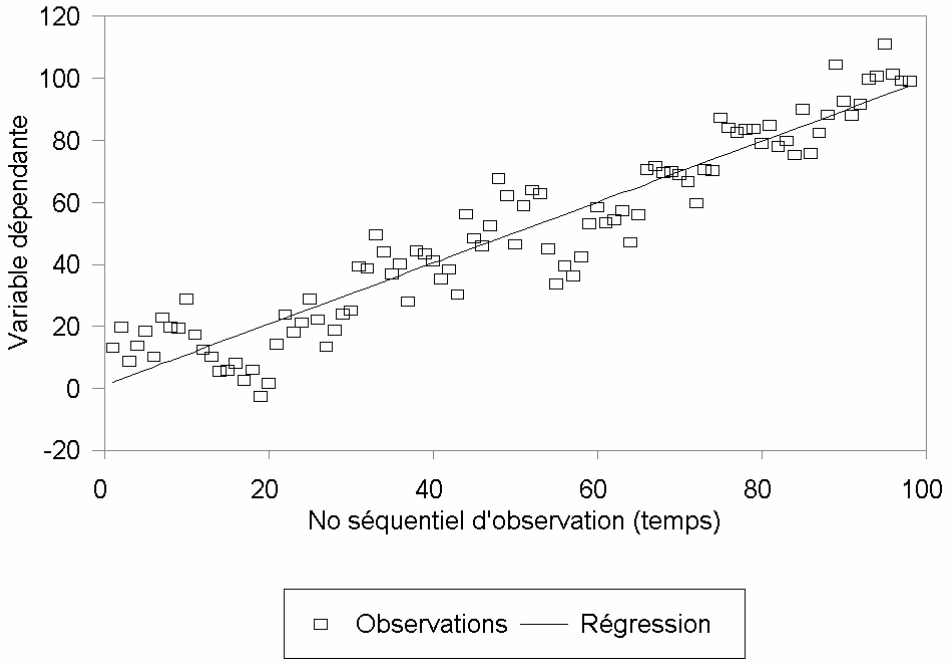
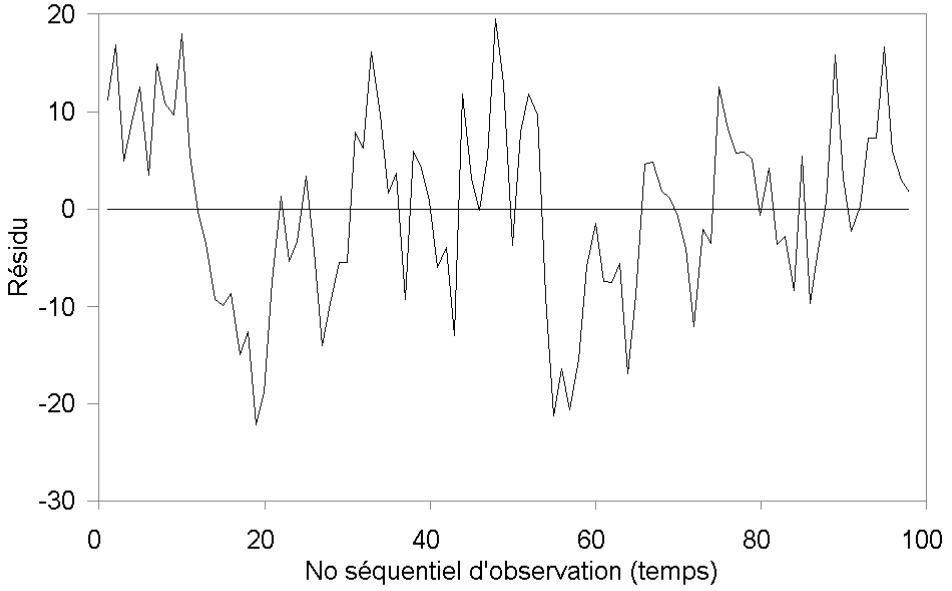


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



ANALYSE DES RÉSIDUS
AUTOCORRÉLATION DES TERMES ALÉATOIRES (4) : α=0

Figure 9 - Observations et régression
Autocorrélation des termes aléatoires

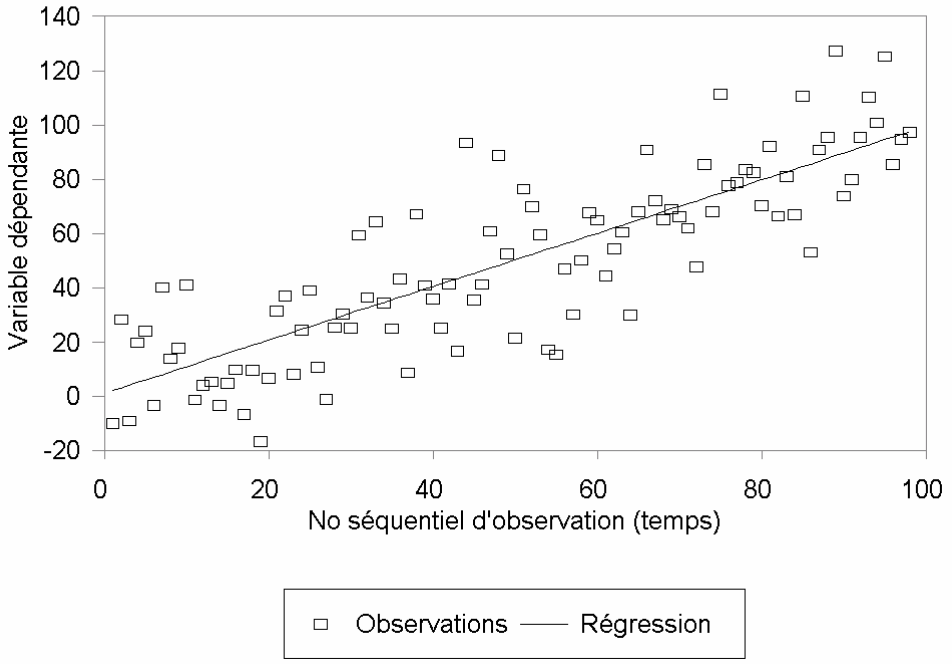
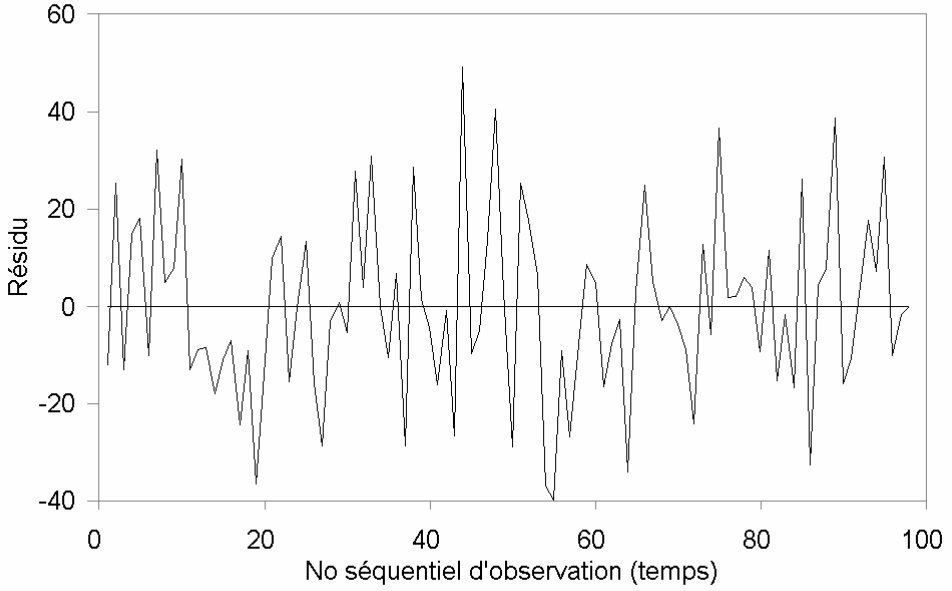


Figure 10 - Résidus de la régression
Autocorrélation des termes aléatoires



ANALYSE DES RÉSIDUS
HÉTÉROSCÉDASTICITÉ (1)

La variance de l'erreur est la même pour toutes les observations :

$\sigma_i^2 \neq \sigma_j^2$ pour $i \neq j$ (au lieu de $\sigma_i^2 = \sigma^2$ pour tous les i)

$\Rightarrow$ Non respect de l'hypothèse H2a du modèle classique de la régression linéaire.

Les données sous-jacentes aux résidus de régressions présentés aux figures 11 et 12 ont été générées au moyen de l'équation

$$y_i = x_i + 100\ \eta_i$$

où η_i a été généré au moyen de l'équation

$$\eta_i = 0{,}1\left(\varepsilon_i \sqrt{x_i}\right)$$

ε_i ayant une distribution proche de la normale [4].

Il est à noter que $x_i = i$, de sorte que les observations sont automatiquement rangées par ordre croissant de la variable indépendante x_i.

CONSÉQUENCES

- Forte variance d'échantillonnage : estimateurs moins précis
- Les tests statistiques ne sont plus valides.

TEST DE DÉTECTION

- Test de Goldfeld et Quandt (Theil, 1971, p. 196-199 ; voir aussi Kennedy, 1992, p. 126).

REMÈDES

- Correction du modèle d'échantillonnage par la transformation des données
- Exemple

  La variance est liée à variable indépendante x_{ik} et $\sigma_i^2 = x_{ik}\,\sigma^2$ approximativement.

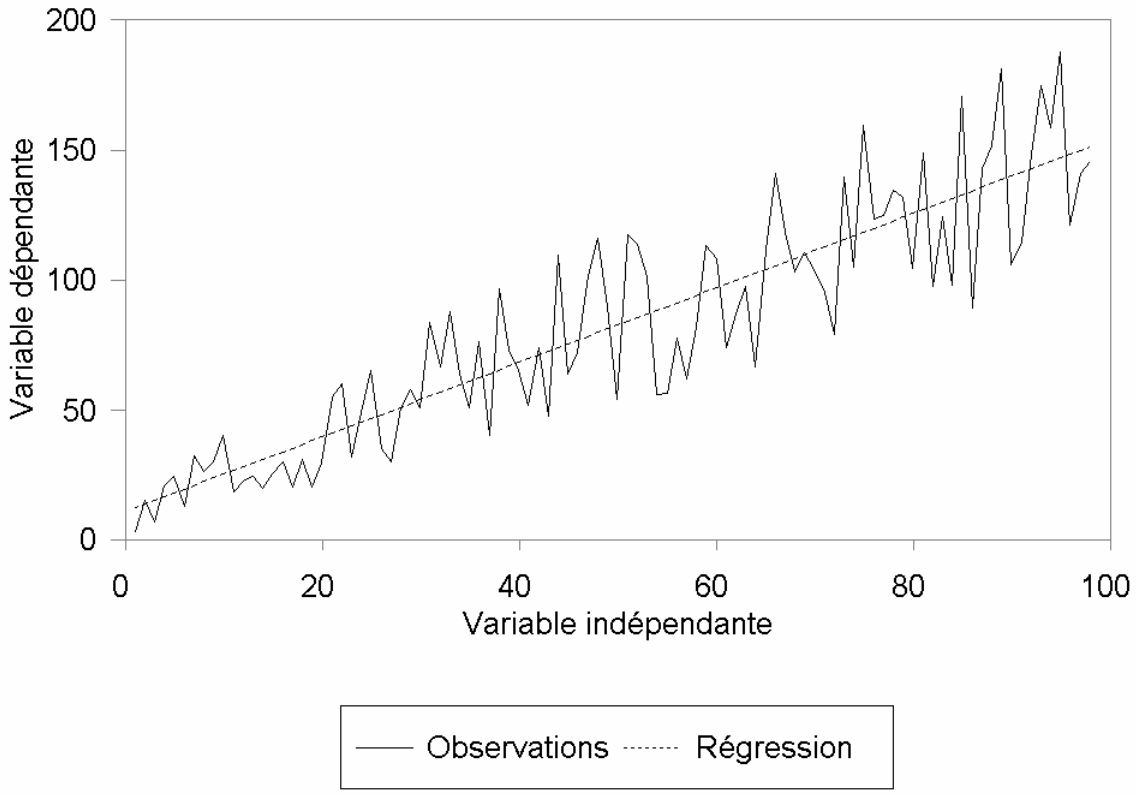
  La transformation $y'_i = \dfrac{y_i}{\sqrt{x_{ik}}}$ et $x'_{ij} = \dfrac{x_{ij}}{\sqrt{x_{ik}}}$ recrée l'homoscédasticité des conditions de Gauss-Markov

---

4 Plus exactement, il s'agit de la transformation logistique d'une variable ayant une distribution uniforme entre zéro et un.

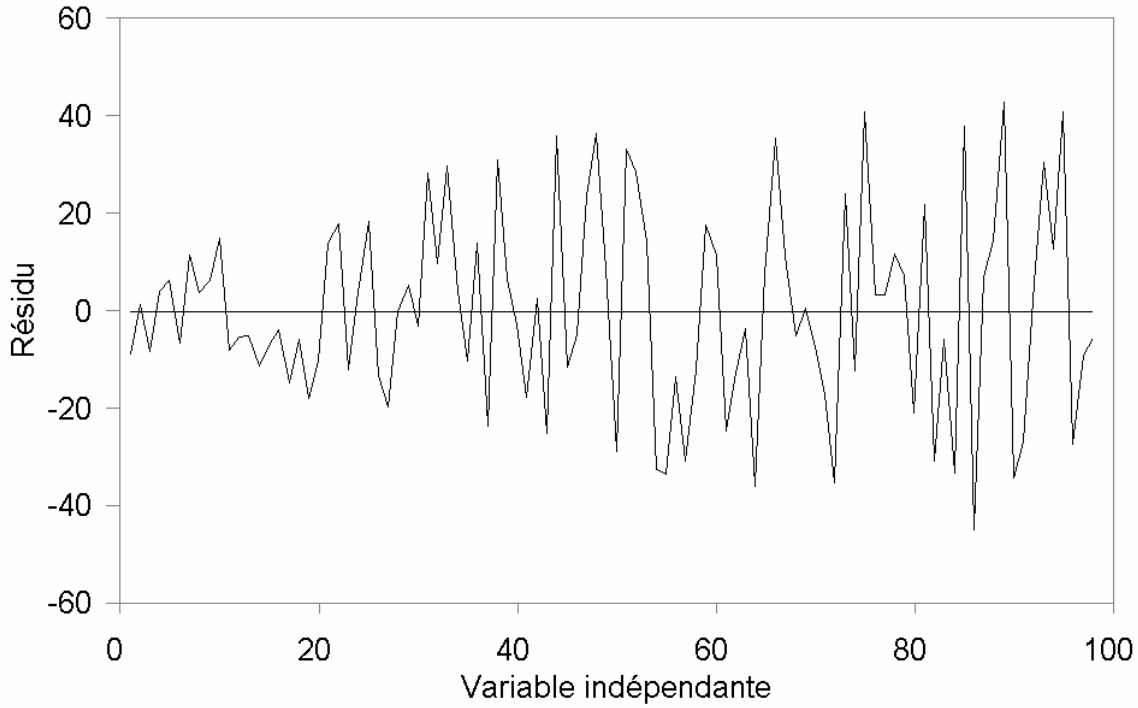
ANALYSE DES RÉSIDUS
HÉTÉROSCÉDASTICITÉ (2)

Figure 11 - Observations et régression
Hétéroscédasticité



ANALYSE DES RÉSIDUS
HÉTÉROSCÉDASTICITÉ (3)

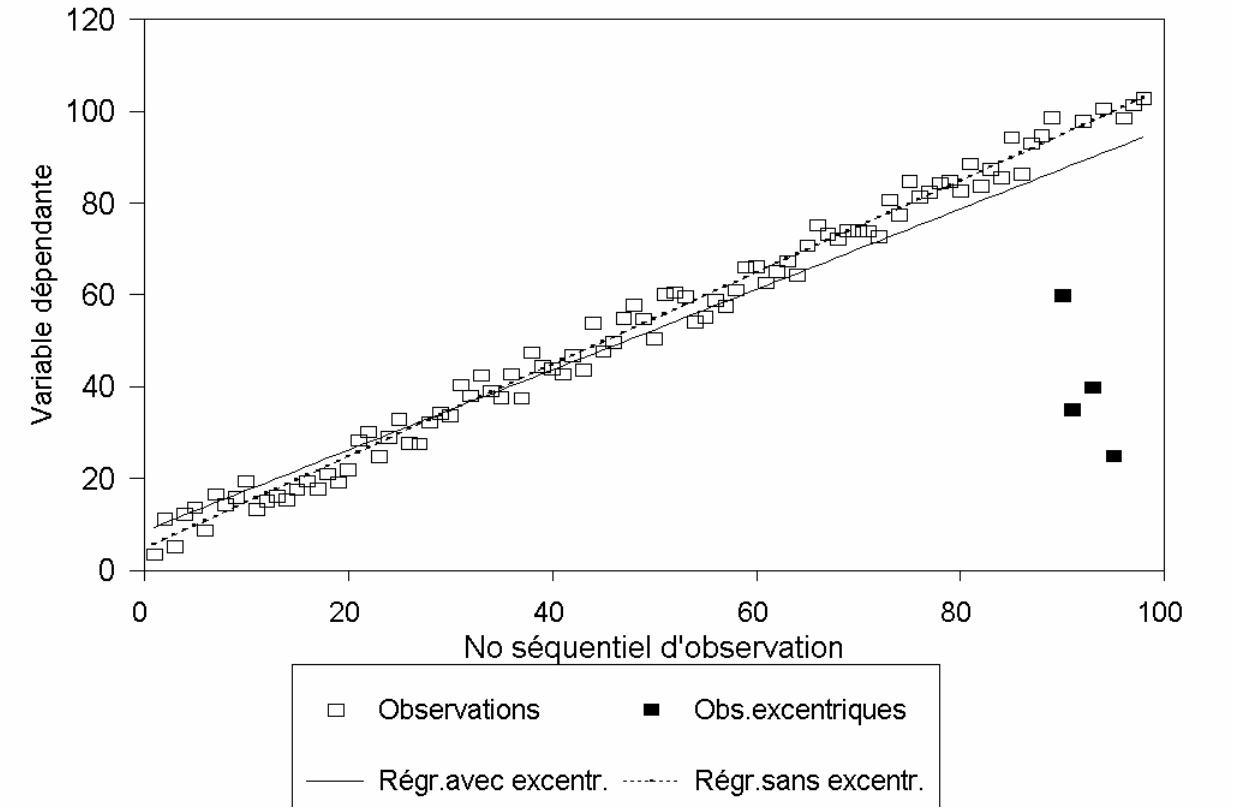
Figure 12 - Résidus de la régression
Hétéroscédasticité



ANALYSE DES RÉSIDUS

OBSERVATIONS EXCENTRIQUES (*OUTLIERS*) (1)

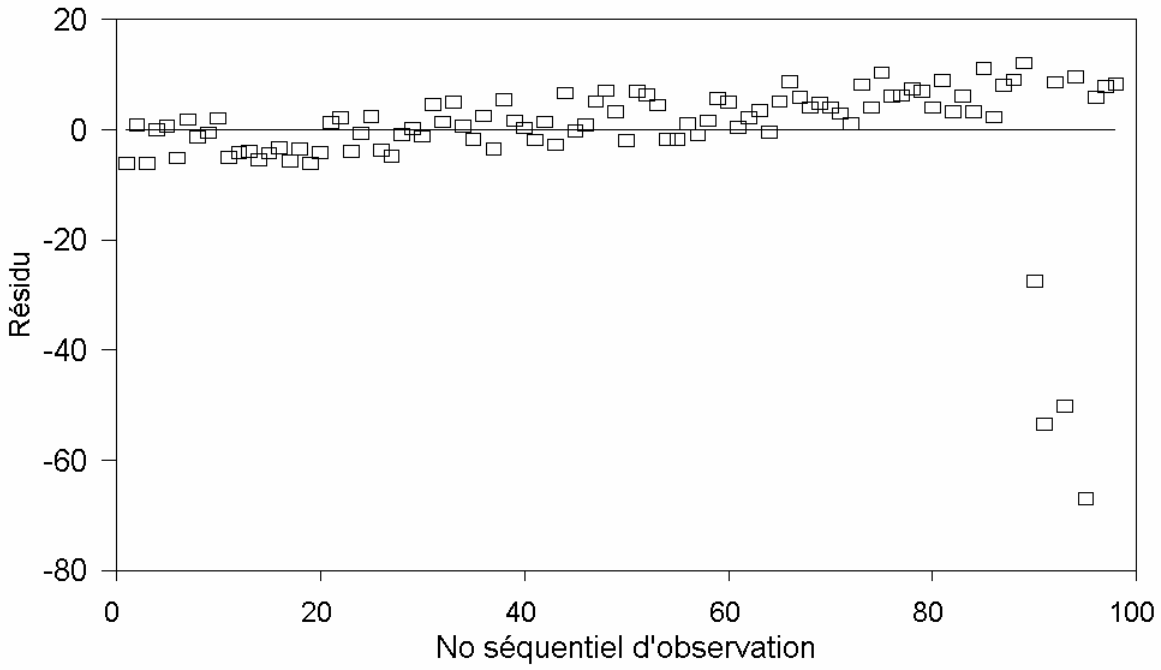
Figure 13 - Observations et régression
Observations excentriques (Outliers)



ANALYSE DES RÉSIDUS

OBSERVATIONS EXCENTRIQUES (*OUTLIERS*) (2)

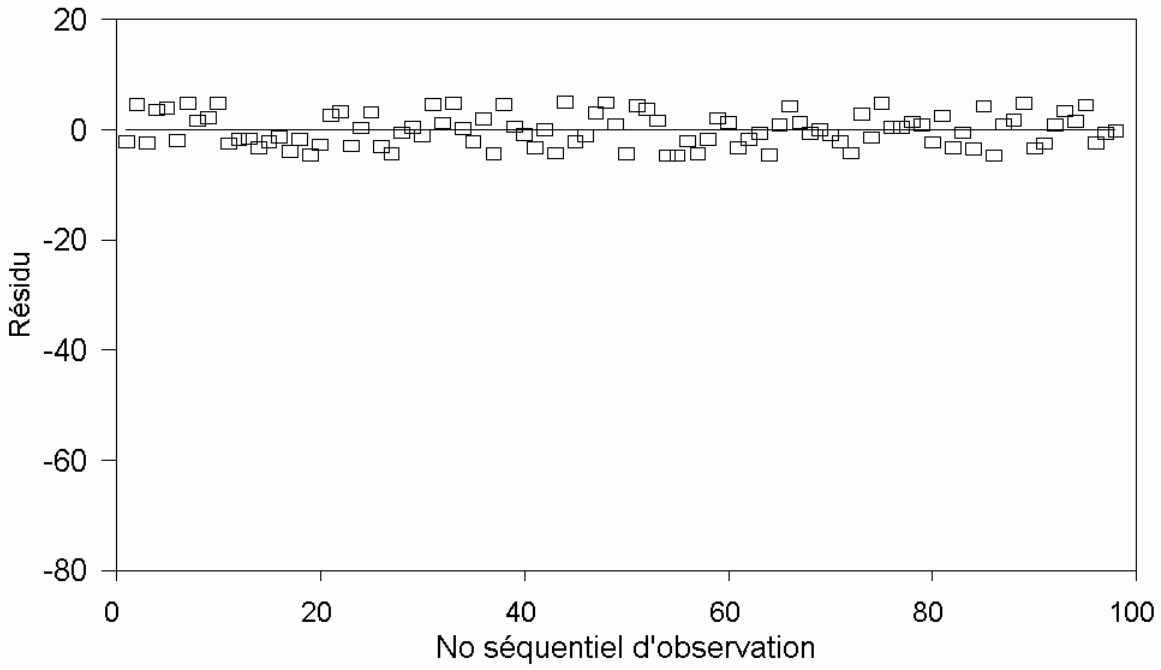
Figure 14 - Résidus des régressions
avec observ. excentriques (Outliers)



ANALYSE DES RÉSIDUS

OBSERVATIONS EXCENTRIQUES (*OUTLIERS*) (3)

Figure 15 - Résidus des régressions
sans observ. excentriques (Outliers)



MULTICOLLINÉARITÉ

1. Multicollinéarité stricte

- Contradiction avec l'hypothèse H4 du modèle classique de la régression linéaire :
  redondance parmi les variables indépendantes
- Peut résulter d'une erreur de spécification lorsque le modèle contient des variables muettes (*dummy variables*).
- Aucun problème de détection :
  diagnostiquée à cause de l'impossibilité des calculs d'estimation.

2. Multicollinéarité approximative

- Lorsque l'une des variables indépendantes est fortement correliée à une autre
  ou à une combinaison linéaire des autres
- Variable « presque » redondante :
  apporte peu d'information supplémentaire

CONSÉQUENCES

- Précision des estimateurs faible : variances d'échantillonnage s_{bj}^2 grandes.
- On ne peut pas bien séparer l'influence des variables qui sont correliées entre elles.
- Dans le cas de deux variables, cela peut se manifester de la façon suivante :
  aucune des deux variables n'a un coefficient significativement différent de zéro ;
  mais si l'on retire les deux variables, le modèle contraint ainsi défini est rejeté par le test *F*.

TESTS DE DÉTECTION

- Deux à deux : coefficients de corrélation simple entre variables indépendantes
- Souvent plus complexe et il faut recourir à des analyses plus raffinées
  (voir Kennedy, 1992, p. 180, à propos du *condition index*).

REMÈDES

- Dans certains cas, une ou plusieurs variables indépendantes de trop dans le modèle.
- Souvent, il faut accepter de « vivre avec » et renoncer à séparer l'influence des variables correliées.

MULTICOLLINÉARITÉ

FAUT-IL ÉLIMINER UNE VARIABLE ?

- C'est parfois une erreur grave.
- Exemple :
  - Variable dépendante *Y*
  - Facteurs inobservables *A*, *B* et *C* (ne peuvent pas figurer dans le modèle)
  - Variables indépendantes observables X_1 et X_2 correliées
    à cause de l'influence commune de *B*
  - Mais écarter X_1 ou X_2 = éliminer facteur sous-jacent *A* ou *C*