7. Influence, résidus, validation

- Le critère des moindres carrés, comme la vraisemblance appliquée à une distribution gaussienne douteuse, est très sensible à des **observations atypiques**, hors norme ("outliers") c'est-à-dire qui présentent des valeurs trop singulières.
- L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant.
- Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les **observations influentes**, c'est-à-dire celles dont une faible variation des variables explicatives et réponse (couple (x_i, y_i) dans le cas de régression linéaire simple) induisent une modification importante des caractéristiques du modèle.
- Une fois ces observations repérées, il n'y a pas de remède universel : supprimer une valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (par exemple en norme L_1), ne rien faire, ...

La décision dépend du contexte et doit être argumentée.

7.1 Premier aperçu graphique

- L'inférence statistique relative à la régression (estimation par intervalle des coefficients, tests d'hypothèses, etc.) repose principalement sur les hypothèses liées au terme d'erreur ε qui résume les informations absentes du modèle. Il importe donc que l'on vérifie ces hypothèses afin de pouvoir interpréter les résultats.
- Rappelons brièvement les hypothèses liées au terme d'erreur :
  - sa distribution doit être symétrique, plus précisément elle suit une loi normale ;
  - sa variance est constante ;
  - les erreurs ε_i $(i = 1, ..., n)$ sont indépendantes.
- Pour inspecter ces hypothèses, nous disposons des erreurs observées (les résidus) produites par la différence entre les valeurs observées y_i et les prédictions ponctuelles de la régression $\hat{y}_i$:

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

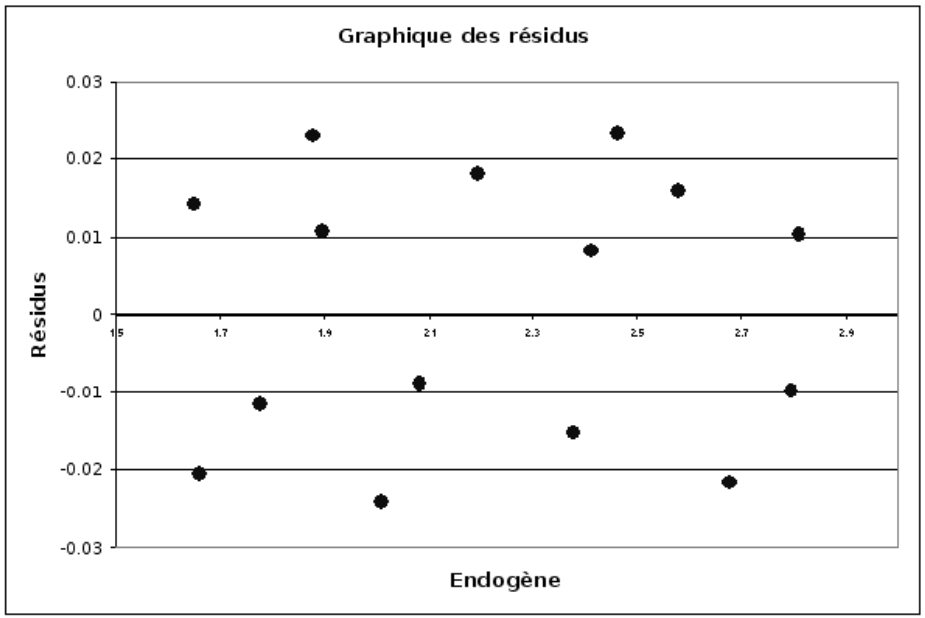
- Remarque : Dans un modèle avec constante, la moyenne des résidus est par construction égale à 0. Ce résultat ne préjuge donc en rien de la pertinence de la régression.

Premier aperçu graphique

- Aussi simpliste qu'il puisse paraître, le diagnostic graphique est pourtant un outil puissant pour **valider** une régression. Il fournit un nombre important d'informations que les indicateurs statistiques appréhendent mal.
- Toute analyse de régression devrait être immédiatement suivie des graphiques des résidus observés... car il y en a plusieurs.
- Avant d'énumérer les différents types de graphiques, donnons quelques principes généraux :
  - les résidus sont portés en ordonnée ;
  - les points doivent être uniformément répartis au hasard dans un intervalle sur l'ordonnée ;
  - aucun point ne doit se démarquer ostensiblement des autres ;
  - on ne doit pas voir apparaître une forme de régularité dans le nuage de points.
- Le type du graphique dépend de l'information que nous portons en abscisse.

Graphe des résidus en fonction de la réponse Y

- Ce type de graphique permet de se rendre compte de la qualité de la régression.
- Les résidus $\hat{\varepsilon}_i$ doivent être répartis aléatoirement autour de la valeur 0, ils ne doivent pas avoir tendance à prendre des valeurs différentes selon les valeurs de Y.
- Exemple de graphe des résidus "normal" :



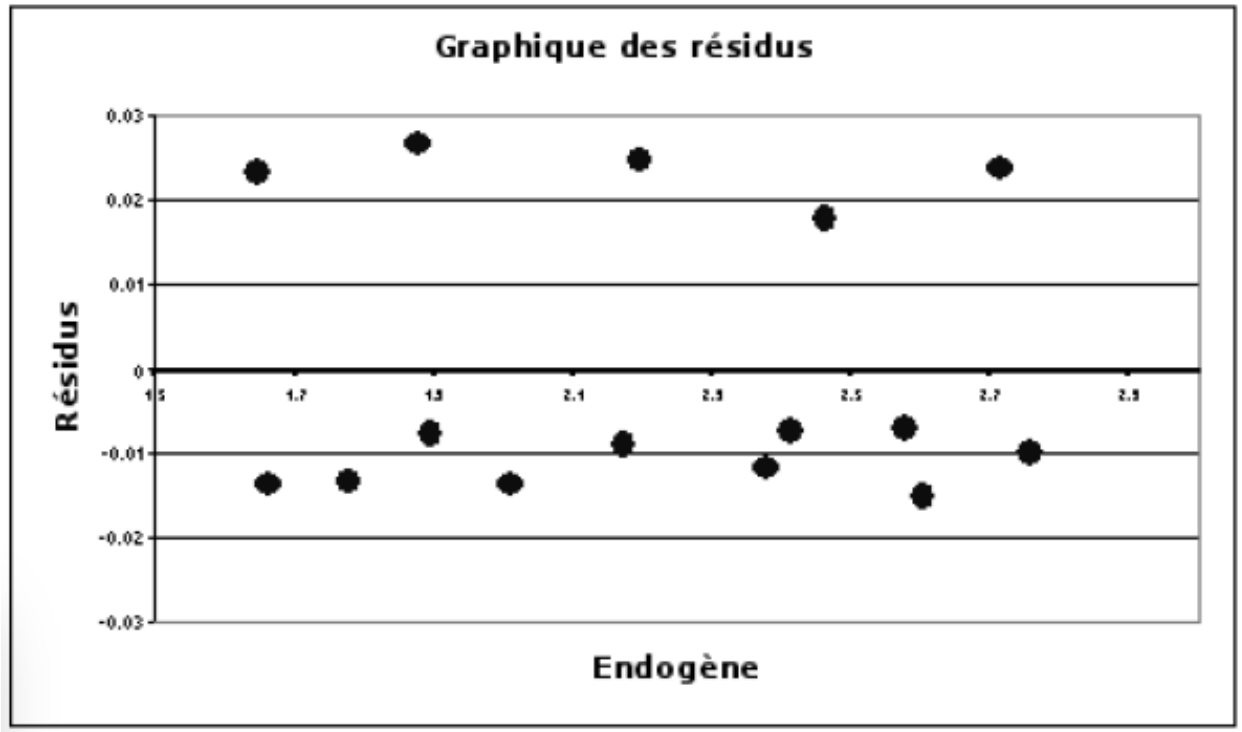
D'autres graphes des résidus...

- **Résidus en fonction de chaque variable explicative X_j**
  - Il doit être produit pour chaque variable explicative.
  - L'idée est de détecter s'il y a une relation quelconque entre le terme d'erreur et les variables explicatives.
  - Rappelons que les variables explicatives et les erreurs sont indépendantes par hypothèse (covariance nulle), cela doit être confirmé visuellement.

- **Résidus en fonction du temps (pour des données longitudinales)**
  - Dans le cas particulier des séries temporelles, nous pouvons produire un graphique supplémentaire en portant en abscisse la variable temps. Elle permet d'ordonner les valeurs d'une autre manière.
  - Il est alors possible de détecter une rupture de structure associée à une date particulière (ex. guerre, crise politique, choc économique, etc.), à une corrélation entre point de temps successifs (auto-corrélations) , ...

Asymétrie des résidus

Exemple de distribution des résidus asymétrique :



Asymétrie des résidus

L'asymétrie des résidus est notamment signe que la distribution des résidus ne suit pas la loi normale. Cette situation survient :

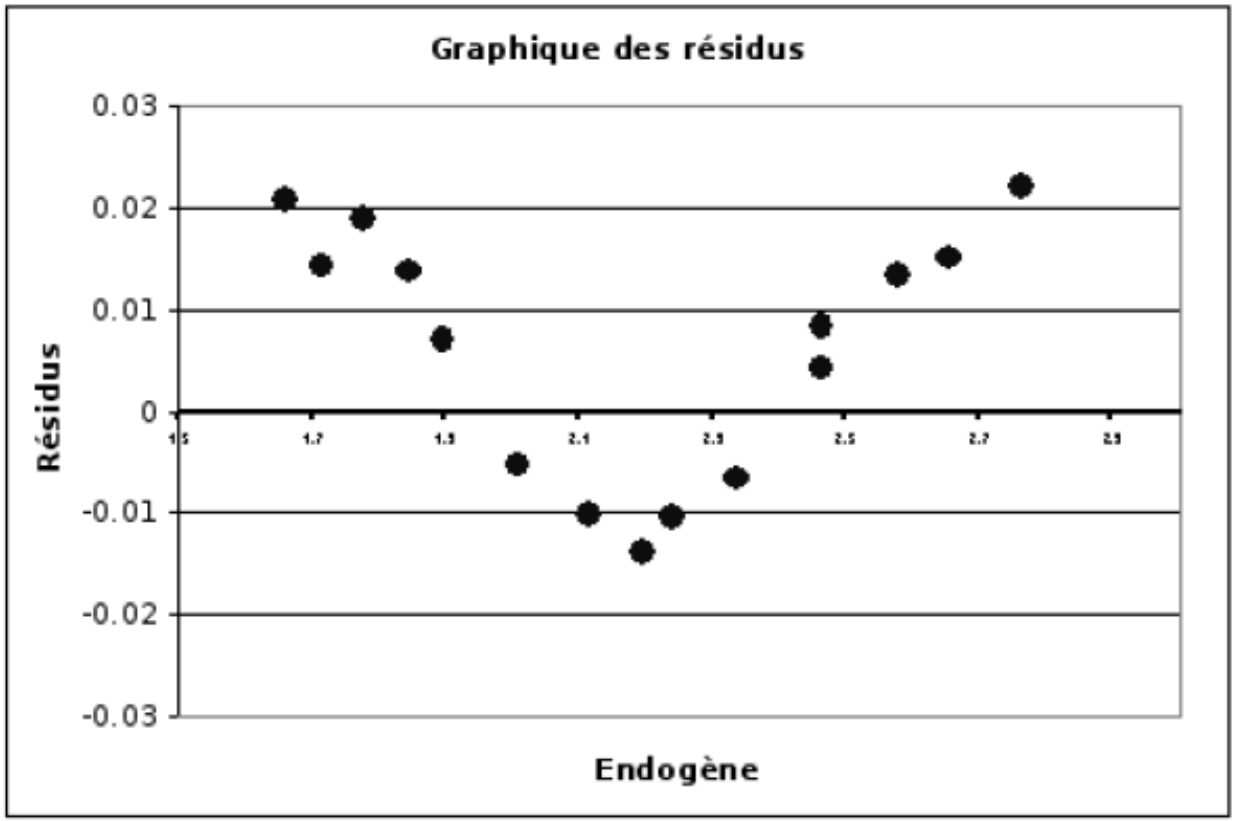
- lorsque certains points se démarquent des autres : ils sont alors souvent mal reconstitués par la régression.
  (La moyenne des résidus est mécaniquement égale à 0, mais la dispersion est très inégale de part et d'autre de cette valeur.)
- lorsque les données sont en réalité formées par plusieurs populations.
  (Exemple : en médecine, effectuer une régression en mélangeant les hommes et les femmes, sachant qu'ils réagissent de manière différente à la maladie étudiée).
- lorsqu'on est face à un problème de spécification, une variable explicative importante manque.
- ...

Non-linéarité

- Si la relation étudiée est en réalité **non-linéaire**, elle ne peut pas être modélisée à l'aide de la régression linéaire multiple.
- Les résidus apparaissent alors en "blocs" au-dessus (prédiction sous-estimée) ou en-dessous (prédiction sur-estimé) de la valeur 0.
- On peut y remédier en ajoutant une variable transformée dans le modèle (par ex. en passant une des variables au carré, ou en utilisant une transformation logarithmique, etc.).
- On peut aussi passer à une régression non-linéaire...

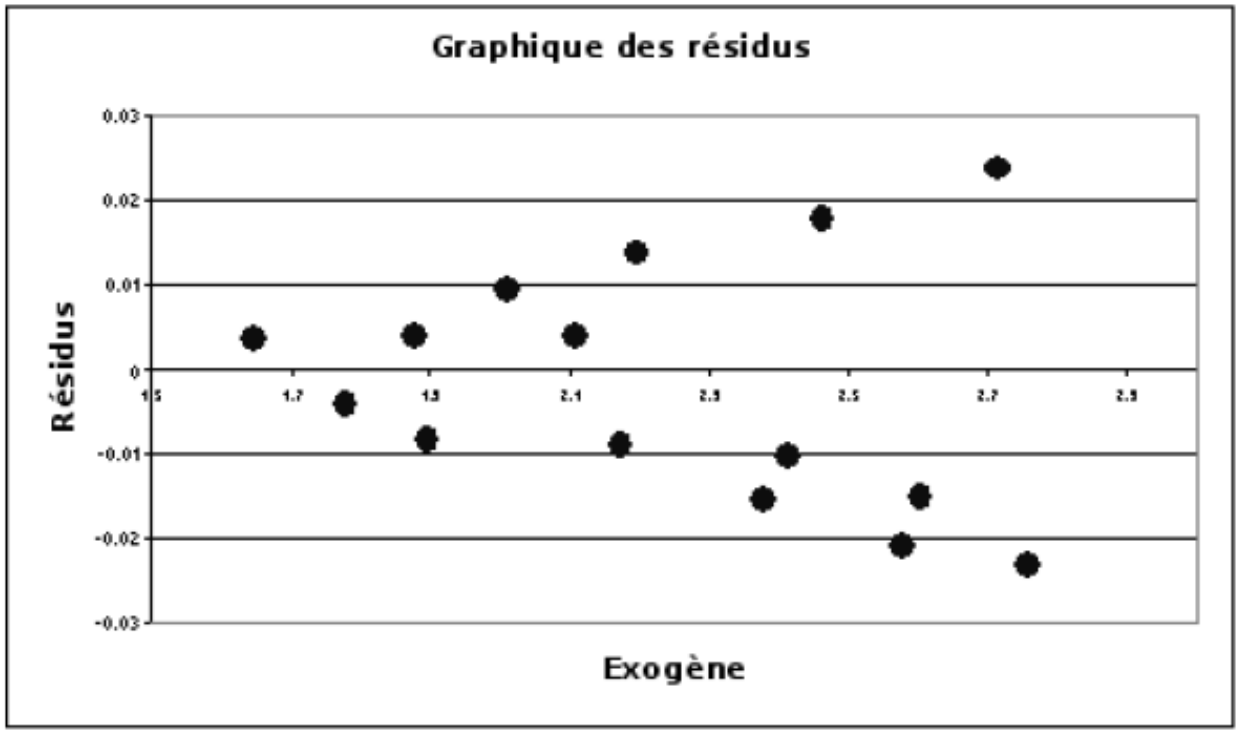
Non-linéarité

Exemple de graphe des résidus en fonction de la variable à expliquer (= variable endogène) pour lequel la relation à modéliser est non-linéaire :



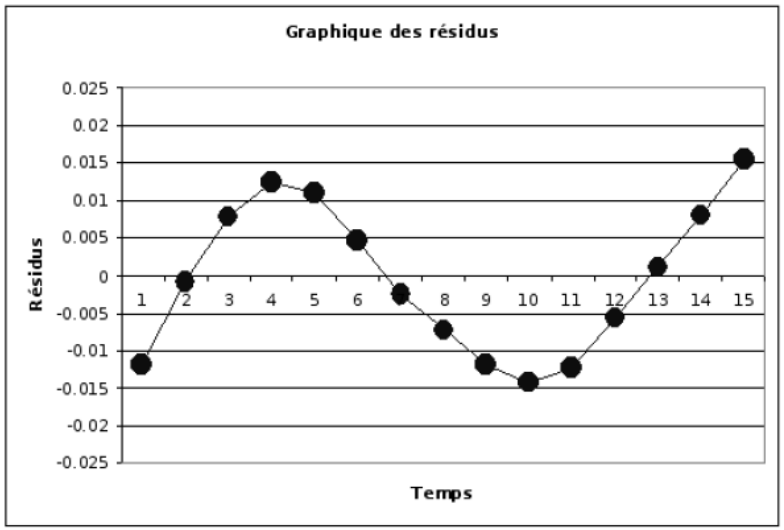
Hétéroscédasticité

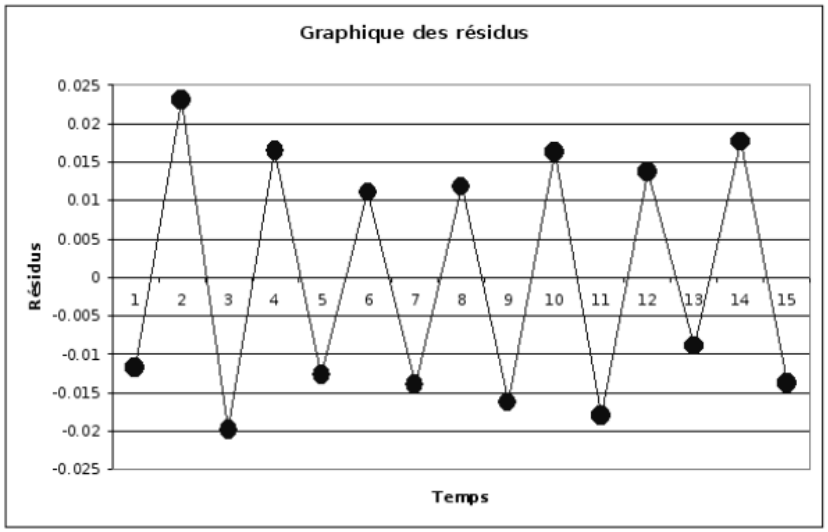
- Souvent associée à une des variables explicatives en abscisse, ce type de graphique indique que la variance des résidus n'est pas constante, et qu'elle dépend d'une des variables explicatives.
- Exemple sur ce graphe des résidus en fonction d'une variable explicative (= une variable exogène) :



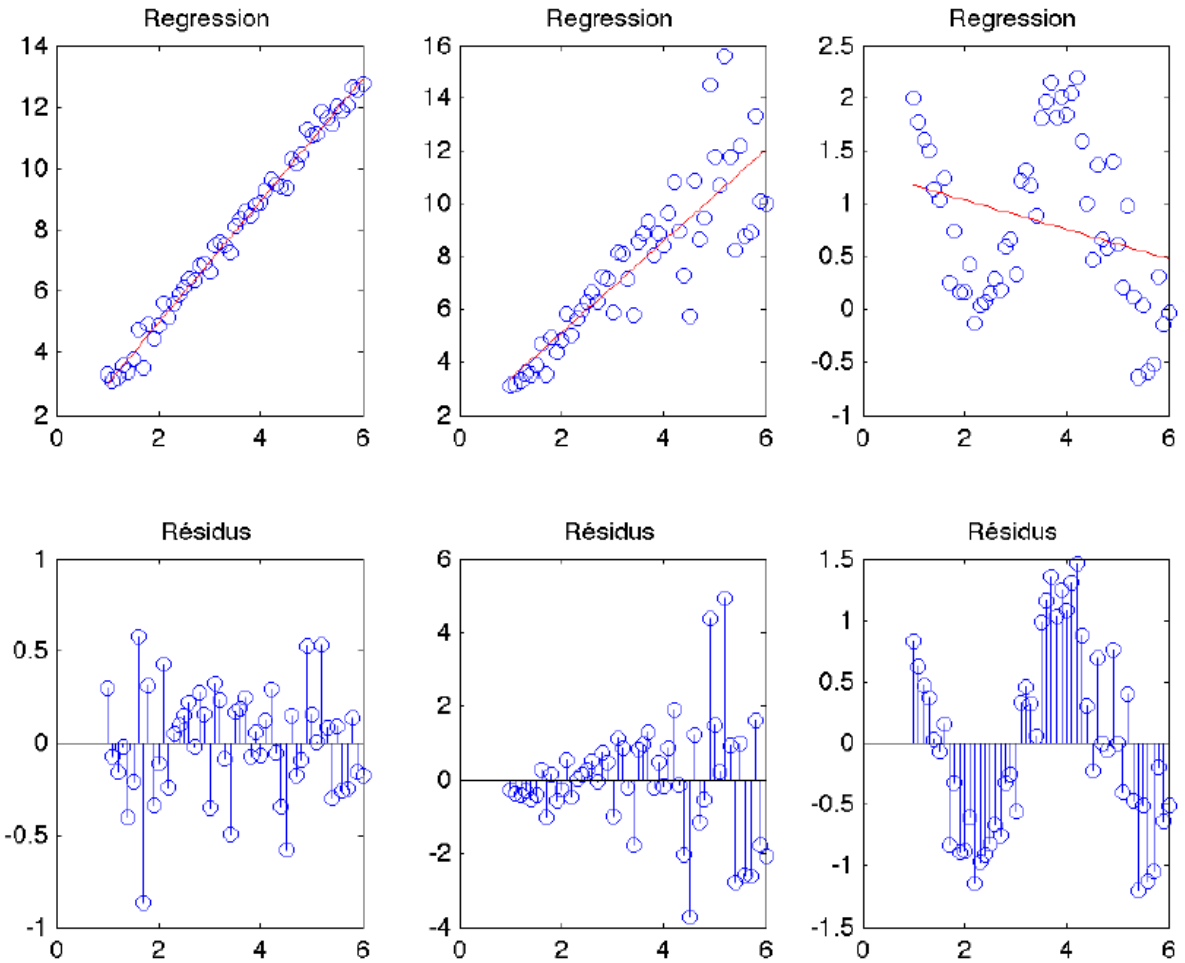
Auto-corrélation des résidus

- Ce problème est spécifique aux données longitudinales (mesurées sur des temps successifs).
- On peut alors essayer de détecter si les erreurs suivent un processus particulier au cours du temps à partir du graphe des résidus en fonction du temps.
- L'auto-corrélation peut être
  - positive : des "blocs" de résidus sont positifs ou négatifs (figure gauche)
  - négative : les résidus sont alternativement positifs et négatifs (figure droite).

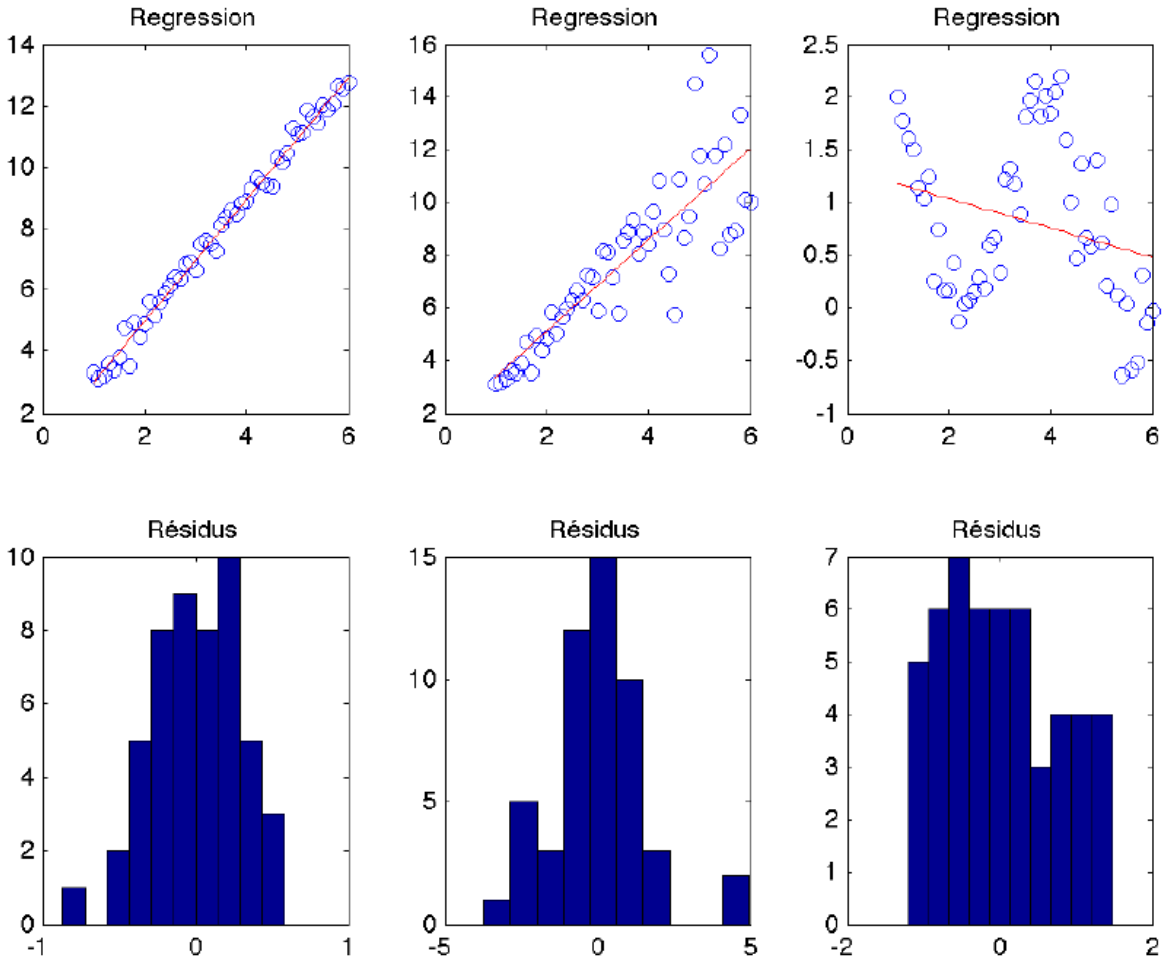




Exemples : graphes des résidus $\hat{e}$ vs variable explicative x

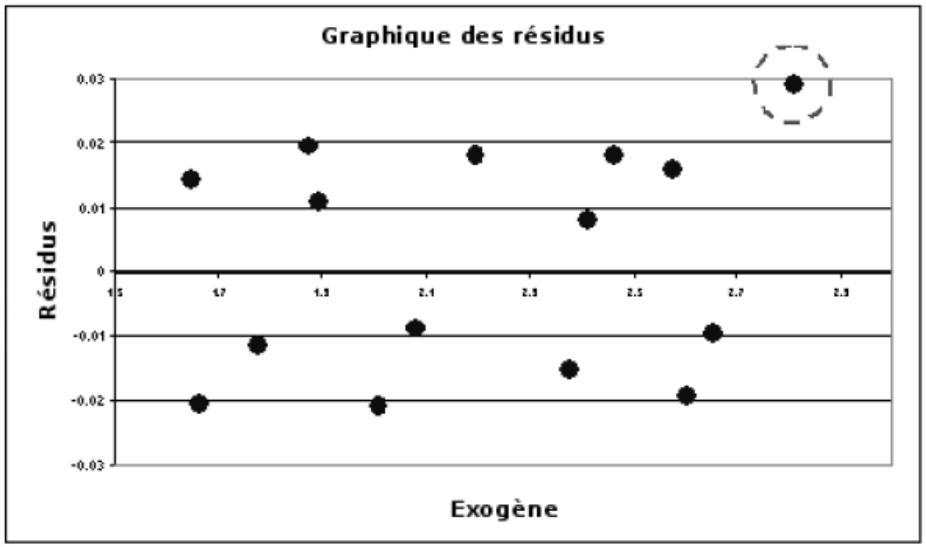


Exemples : histogramme des résidus



Points aberrants et points influents

- Par définition, un point **aberrant ou atypique** est une observation qui s'écarte résolument des autres.
- Cela peut être dû à une erreur de recueil des données, cela peut aussi correspondre à un individu qui n'appartient pas à la population étudiée.
- Dans le graphique de résidus, il s'agit de points éloignés des autres (que la variable en abscisse soit la réponse Y ou une variable explicative X_j)



Ici un point présente une valeur atypique pour une des variables explicatives (exogènes). De plus, cette valeur est mal reconstituée par la régression (le résidu est élevé).

Points aberrants et points influents

- Les points **influents** sont des observations qui pèsent exagérément sur les résultats de la régression. On peut les distinguer de plusieurs manières :
  - ils sont "isolés" des autres points, on constate alors que la distribution des résidus est asymétrique (voir prochaine figure) ;
  - ils correspondent à des valeurs extrêmes des variables, en cela ils se rapprochent des points atypiques.
- Bien souvent la distinction entre les points atypiques et les points influents est difficile. Elle est assez mal comprise : **un point peut être influent sans être atypique, il peut être atypique sans être influent**.
- Cela est difficilement discernable dans un graphique des résidus, il est plus approprié de passer par des calculs.
- Notez bien : La meilleure manière de le circonscrire est de **recalculer les coefficients de la régression en écartant le point** : si les résultats diffèrent significativement, en termes de prédiction ou terme de différence entre les coefficients estimés, le point est influent.

En résumé :

- Un bon résidu est un résidu sans structure ou sans structure apparente...
  - les résidus sont non structurés
    - leur variance est constante
    - ils sont indépendants des observations (x et y)
  - leur distribution est symétrique
  - il n'y a pas de point influent
- Les différentes figures à examiner sont :
  - Résidus $\hat{\boldsymbol{e}}$ vs variable à expliquer y
  - Résidus $\hat{\boldsymbol{e}}$ vs chaque variable explicative x_i,
  - histogramme des résidus $\hat{\boldsymbol{e}}$

7.2 Eléments de diagnostic

- a) Effet levier
- b) Résidus
  Résidus calculés, Résidus standardisés, Erreur de test, Résidus studentisés
- c) Mesure d'influence : Distance de Cook
- d) Régressions partielles
- e) Graphes

a) Effet levier

- Les $\hat{y}_i$ s'appellent les **valeurs ajustées** ou **valeurs prédites** par le modèle : $\hat{y}_i$ est une valeur approchée de y_i.
- $\hat{\boldsymbol{y}} = (\hat{y}_i)_{1 \leq i \leq n}$ est le vecteur des valeurs ajustées : c'est une observation de la variable aléatoire $\hat{\boldsymbol{Y}}$

$$\begin{aligned} \hat{\boldsymbol{Y}} &= \boldsymbol{X}\hat{\beta} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ &= \boldsymbol{H}\boldsymbol{Y} \end{aligned}$$

où $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ est appelée la "matrice chapeau" ou "Hat Matrix".
$\Rightarrow$ Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation $\hat{\boldsymbol{Y}}$: plus la valeur de h_{ii} est élevée, plus l'observation i a un **effet levier** important.

- La **stratégie** consiste le plus souvent à repérer les points tels que :

$$h_{ii} > \frac{2(p+1)}{n}.$$

b) Résidus

- Le vecteur des **résidus calculés** :

$$\hat{\boldsymbol{e}} = (\hat{e}_i)_{1 \leq i \leq n} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

est l'observation de la variable aléatoire $\hat{\varepsilon}$:

$$\begin{aligned} \hat{\varepsilon} &= \boldsymbol{Y} - \hat{\boldsymbol{Y}} \\ &= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} \end{aligned}$$

avec $\hat{\varepsilon} \sim \mathcal{N}\left(0, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\right)$

- Propriété : $\hat{\boldsymbol{Y}}$ et $\hat{\varepsilon}$ sont deux v.a. indépendantes ; $\hat{\varepsilon}$ et $\hat{\boldsymbol{\beta}}$ sont deux variables aléatoires indépendantes.

Résidus standardisés

Différents types de résidus sont définis afin d'affiner leurs propriétés.

- **Résidus calculés** : $\hat{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$

- Même si l'hypothèse d'homoscédasticité est vérifiée (les erreurs ε_i ont même variance), les résidus n'ont pas la même variance (voir diapo précédente : $Var(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii})$).

  Il est donc d'usage d'en calculer des versions **standardisées** afin de les rendre comparables (chacun divisé par l'estimation de l'écart-type).

  Résidus standardisés :

$$r_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}$$

Erreur de test

- Chaque observation i peut être utilisée simultanément pour évaluer la capacité prédiction du modèle. Dans ce cas, on mesure l'**erreur de test** :
  - Construction de l'échantillon $(x_{(-i)}, y_{(-i)})$ de taille $n-1$ (on retire la $i$ème observation (x_i, y_i))
  - Estimation de $\hat{\beta}_{(-i)}$ à partir de cet échantillon $(x_{(-i)}, y_{(-i)})$
  - Calcul de la prédiction : $\hat{y}_{(-i)i} = \hat{\beta}_{(-i)0} + x_{i1}\hat{\beta}_{(-i)1} + \ldots + x_{ip}\hat{\beta}_{(-i)p}$

  On obtient alors l'erreur de test pour l'observation i :

$$\hat{e}_{(-i)i} = y_i - \hat{y}_{(-i)i}$$

- **Faut-il calculer n modèles ?** (afin d'écarter successivement chaque observation 1, 2, ... i, ..., n) NON ! On peut montrer que :

$$\hat{e}_{(-i)i} = \frac{\hat{e}_i}{1 - h_{ii}}$$

- A partir du vecteur des erreurs de test, on peut calculer la somme des carrés de ces résidus, appelé le **PRESS de Allen** (Predicted Residual Sum of Squares) :

$$PRESS = \sum_{i=1}^{n} \hat{e}^2_{(-i)i}$$

Résidus studentisés

- En remplaçant $\hat{\sigma}^2$ dans la définition du résidu standardisé, par l'estimation de σ^2 obtenue sans la $i$ème observation :

$$\hat{\sigma}^2_{(-i)} = \frac{\hat{\boldsymbol{e}}^T_{(-i)}\hat{\boldsymbol{e}}_{(-i)}}{n-p-2}$$

on obtient les **résidus studentisés** :

$$t_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}^2_{(-i)}(1-h_{ii})}}$$

- Sous l'hypothèse de normalité (H_3), on montre que ces résidus suivent une **loi de Student à** $(n-p-2)$ **degrés de liberté**.

- Question : pourquoi $(n-p-2)$ et non $(n-p-1)$?

Résidus studentisés dans la pratique

- Dans la pratique, nous pouvons calculer les résidus studentisés t_i directement à partir des résidus standardisés r_i :

$$t_i = r_i \sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$$

- De trop grands résidus sont aussi des signaux d'alerte.

- En pratique, on considère qu'un résidu studentisé de **valeur absolue plus grande que 2** peut révéler un problème.

c) Mesures d'influence

- Les deux critères précédents (effet levier et taille des résidus) contribuent à déceler des observations potentiellement influentes par leur éloignement à $\overline{x}$ ou à la taille des résidus :
  - L'effet levier peut apparaître pour des observations dont les valeurs prises par les variables explicatives sont élevées (observation loin du barycentre $\overline{x}$).
  - De grands résidus signalent plutôt des valeurs atypiques de la variable à expliquer.

- Ces informations sont synthétisées dans des critères synthétiques évaluant directement l'influence d'une observation sur certains paramètres : les prédictions $\hat{y}$, les paramètres $\hat{\beta}_0$, $\hat{\beta}_1$, le déterminant de la matrice de covariance des estimateurs.

  $\Rightarrow$ Tous ces indicateurs proposent de **comparer** un paramètre estimé **sans la $i$ème observation** et ce même paramètre estimé **avec toutes les observations**.

c) Mesures d'influence

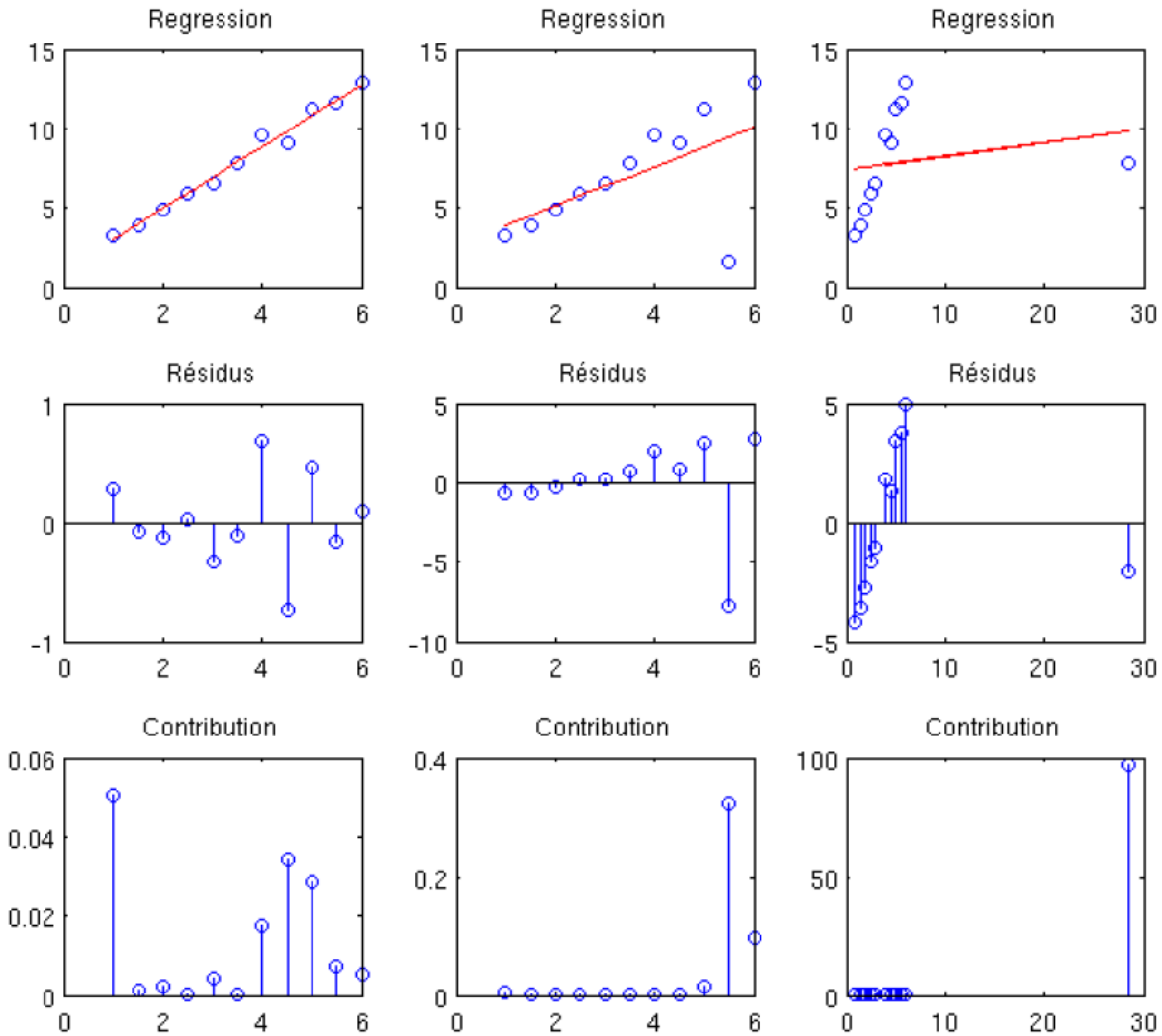
- Un des critères les plus couramment utilisés est la **distance de Cook** :

$$\begin{aligned} D_i &= \frac{(\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(-i)})^T (\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_{(-i)})}{\hat{\sigma}^2 (p+1)} \\ &= \frac{\sum_{j=1}^{n} (\hat{y_{(-i)j}} - \hat{y}_j)^2}{\hat{\sigma}^2 (p+1)} \\ &= \frac{h_{ii}}{(1-h_{ii})} \frac{r_i^2}{(p+1)} \end{aligned}$$

  pour $i = 1, ..., n$, qui mesure l'influence d'une observation sur l'ensemble des prévisions en **prenant en compte effet levier et importance des résidus**.

- **La stratégie de détection** consiste le plus souvent à :
  - repérer les points influents **en comparant les distances de Cook avec la valeur** 1,
  - puis à **expliquer cette influence** en considérant, pour ces observations, leur résidu ainsi que leur effet levier.

Exemples


Regression
Regression
Regression
Résidus
Résidus
Résidus
Contribution
Contribution
Contribution

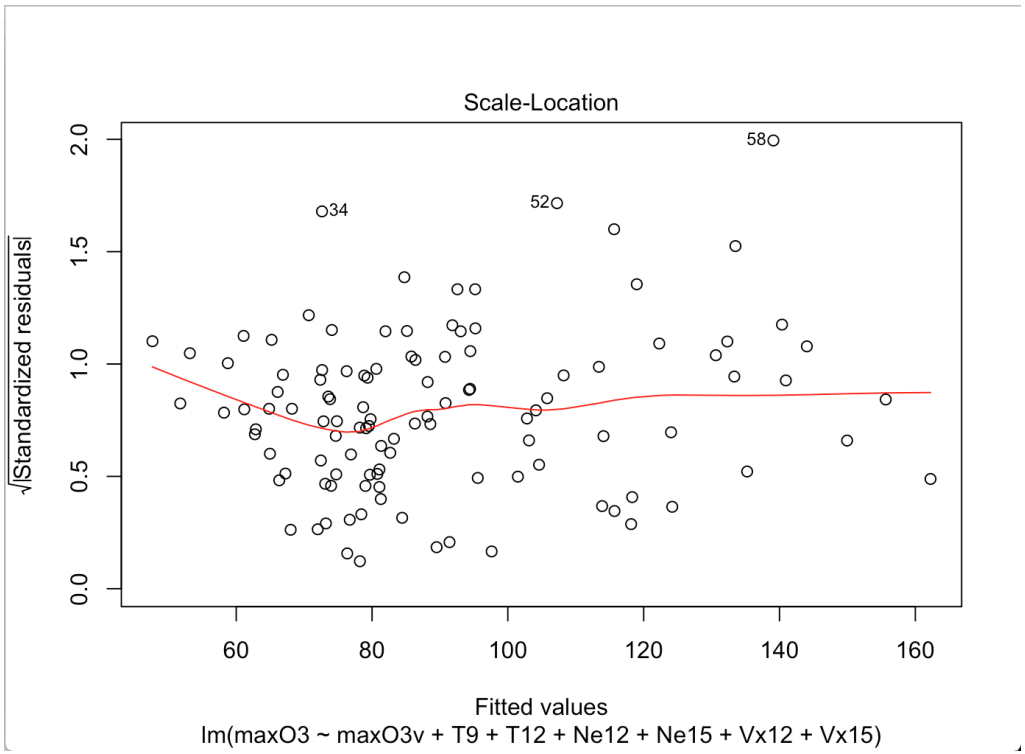
d) Régressions partielles

- Un modèle de régression multiple est une technique linéaire. Il est raisonnable de s'interroger sur la **pertinence du caractère linéaire de la contribution d'une variable explicative** à l'ajustement du modèle. Ceci peut être réalisé en considérant une régression partielle. On calcule alors deux régressions :
  - la régression de Y sur les variables $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ dans laquelle la jème variable est omise. Soit $r_{y(j)}$ le vecteur des résidus obtenus.
  - la régression de X_j sur les variables $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$. Soit $r_{x(j)}$ le vecteur des résidus obtenus.
- La comparaison des résidus par un graphe (nuage de points $r_{y(j)} \times r_{x(j)}$) permet alors de représenter la nature de la liaison entre X_j et Y **conditionnellement aux autres variables explicatives** du modèle.

e) Analyse graphique

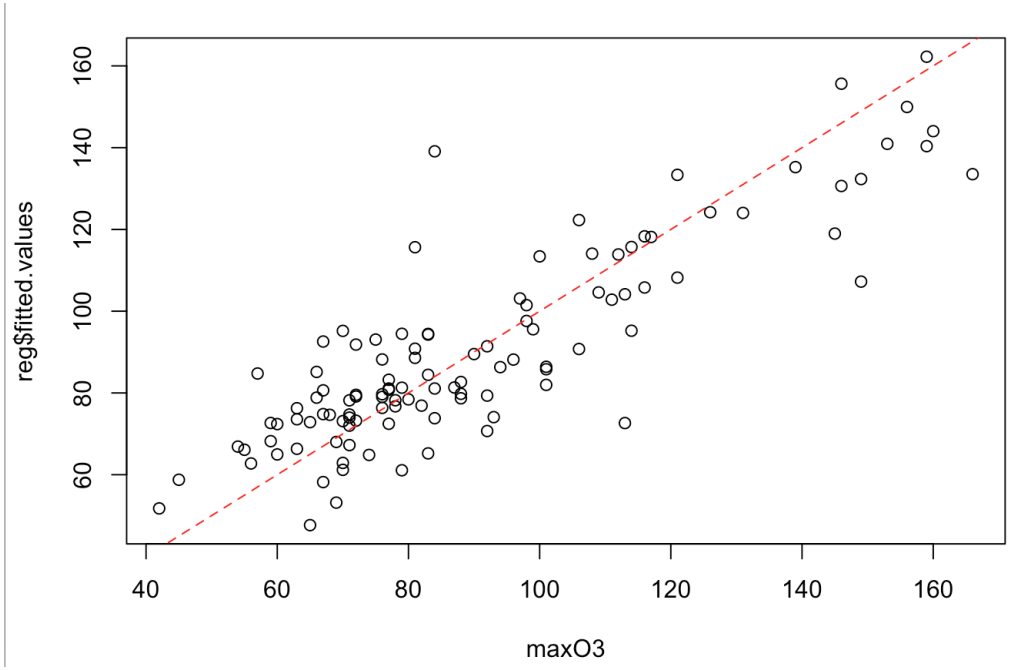
Différents graphiques permettent de contrôler le bien fondé des hypothèses de linéarité, d'homoscédasticité, éventuellement de normalité des résidus.

- Le premier considère **le nuage de points des résidus standardisés ou studentisés croisés avec les valeurs prédites**. Les points doivent être uniformément répartis entre les bornes -2 et $+2$ et ne pas présenter de formes suspectes.



e) Analyse graphique

- **Le deuxième croise les valeurs observées de Y avec les valeurs prédites** $\hat{Y}$. Il illustre le coefficient de détermination R^2. Les points doivent s'aligner autour de la première bissectrice.



e) Analyse graphique

- La qualité, en terme de linéarité, de l'apport de chaque variable est étudiée par des **régressions partielles**.
- Chaque graphe de résidus peut être complété par une estimation fonctionnelle ou régression non-paramétrique (loess, noyau, spline) afin d'en facilité la lecture.

e) Analyse graphique

- Le dernier trace **la droite de Henri (Normal QQ-plot) des résidus** dont le caractère linéaire de la représentation donne une idée de la normalité de la distribution.