Les régressions linéaires multiples: Tests d'hypothèses Partie 2

Problèmes courants:

- Les erreurs de spécification
- La multicolinéarité
- L’hétéroskedasticité
- L’autocorrélation

Les erreurs de spécification:

- Deux erreurs de spécification possibles :
  - —Le vrai modèle n'est pas linéaire
  - —Une ou plusieurs variables expliquant les variations de y ont été
    omises
- Importance du choix des variables et de la forme fonctionnelle du modèle
  - —Les analyses préliminaires
  - —La revue de littérature

Le vrai modèle n'est pas linéaire

- Cette erreur de spécification peut conduire à des problèmes d'autocorrélation des résidus et d'hétéroscédascité.
- Solutions : linéariser (ex: les log), employer une fonction différente (ex; fonction logistique...)

Variables omises

- Une ou plusieurs variables explicatives (importantes) ont été omises
- Deux origines possibles :
  - Manque de données (on est conscient)
  - Formulation incomplète du bloc explicatif, à droite de l'équation structurelle (on n'est pas conscient)
- Importance des analyses préliminaires et analyse de la littérature + analyse descriptive

Variables omises

- Plusieurs cas de figure possibles :
- Les variables omises (dans le terme d'erreur) ne sont pas corrélées avec les variables présentes
  - Conséquence : surestimation de la variance résiduelle = SCR élevée
- Les variables omises (dans le terme d'erreur) sont corrélées avec les variables présentes
  - Conséquence : Les estimateurs des variables non-omises et corrélées avec les variables omises seront biaisés
  - Biais => endogénéité, remise en cause de H6 (variables exogènes certaines)

Variables Omises:

Supposons le modèle avec la bonne spécification suivante :

— Ln(salaire) = β0 + β1 éducation + β2 .facultés innées + u

•Supposons que l'on n'observe pas les facultés innées, et que nous sommes contraints d'estimer le modèle suivant :

— Ln(salaire) = β0 + β1 éducation + u

•A quoi peut-on s'attendre ?

— Les deux variables explicatives éducation et facultés.innées sont positivement corrélés les personnes qui ont le plus de facultés sont celles qui font le plus d'études

— S2 est positif: plus les facultés innées sont grandes plus le taux de salaire augmente

— Biais positif possible surestimation de l'effet du niveau d'études sur la variation du taux de salaire lorsqu'on omet la variable x2 (facultés innées).

Variables Omises:

Résumé par l'exemple

Soit le vrai modèle spécfié avec deux variables explicatives :

En raison de notre ignorance ou par manque dünformation, nous sommes contraints d'estimer le modèle :

Si x1 et x2 sont corrélées voici les cas possibles :

	Corr(x1,x2)< 0	Corr(x1,x2)>0
B2< CO	Biais positif	Biais négatif
B2>0	Biais négatif	Biais positif

Variables Omises:

- Solutions :
  - Il est possible d'avoir les données
    - Revoir le bloc de variables explicatives
    - Importance des analyses descriptives, préliminaires, exploratoires
  - Les données ne sont pas disponibles
    - Controler la corrélation entre les variables omises et non omises en incluant dans le modèle une variable « proxy », i.e. une variable que l'on pense corrélée à la variable omise. exemple : niveau d'études du père/ de la mère
    - Méthode des variables instrumentales (VI)

Multicolinéarite:

- Rappel : Modèle incorporant des variables explicatives qui sont très corrélées entre elles
- Remise en cause des hypothèses de l'OLS
- Des variables sont parfaitement colinéaires s'il existe une relation linéaire entre elles.
- Exemple x1 = x2

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$
$$Y_i = \beta_0 + 2.\beta_1 x_{i2} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

- Conséquence : Inclure x1 et x2 équivaut à mettre deux fois la même variable dans le modèle.
- Perturbe l'estimation : on ne peut pas calculer 2 coefficients pour une même variable.

Multicolinéarite:

- C’est une multicolinéarité prononcée qui peut être problématique
- NB : à l’opposé quand les variables explicatives ont une covariance nulle = Ortogonalité
  - Dans la réalité, les variables explicatives sont toujours plus ou moins corrélées entre elles mais cela ne pose problème que si la corrélation est forte

Multicolinéarite:

- Les effets principaux de la multicolinéarité
- Multicolinéarité parfaite
  - Problème d'identification , coefficient indéterminé
  - Ex: l'age et la date de naissance
- Multicolinéarité partielle
  - Peut augmenter la variance estimée des coefficients de régression perte de précision
  - Peut provoquer l'instabilité des coefficients : de faibles fluctuations de données (échantillons différents) peuvent entrainer de fortes variations de coefficients
  - Les coefficients peuvent sembler non significatifs et présenter le mauvais signe

- Rendent l'estimation Invalide.

Détection de la multicolinéarité

- Premières vérifications possibles:

—Certains coefficients de corrélation linéaire entre les variables explicatives sont très élevés (éditer la matrice des coefficients)

—Les tests de Student concluent à une non significativité des coefficients alors que le test de Fisher de significativité globable conclut à la significativité du modèle

Détection de la multicolinéarité

- Le test de Klein
  - —Comparaison du R2 calculé à partir de la régression du modèle général (y en fonction des k variables explicatives) avec ceux des régressions auxiliaires (chaque variable explicative sur les k-1 variables restantes)
  - —Si le R2 du modèle général est supérieur à chaque R2 des régressions auxiliaires, on rejette le risque de multicolinéarité

Multicolinéarite:

- Détection de la multicolinéarité
- Le facteur d'inflation de la variance ou VIF
  - Pour chaque régression auxiliaire, on calcule la tolérance :
    - TOL= 1-R2
  - La tolérance = part de la variance d'une variable explicative qui n'est pas expliquée par les autres variables explicatives du modèle
  - 0n calcule VIF = 1/T0L
  - Des valeurs élevées de VIF (>10) indiquent la présence de multicolinéarité

Multicolinéarite:

En cas de multi colinéarité : revoir le bloc explicatif et analyser à nouveaux ses variables. Ne pas oublier qu'un modèle est une représentation théorique de la réalité que l'on doit justifier avant de poser la moindre équation.

Mais commentfaire le choix entre 2 modèles ?

Choix entre deux modèles:

- Utilisation d'indicateurs
- Utilisation de procédures statistiques permettent de déterminer les variables à retirer/garder

Choix entre deux modèles:

- Choix entre 2 modèles
  - Si plusieurs modèles concurrents : les variables explicatives sont différentes, mais toutes significatives
  - Le meilleur modèle : celui composé des variables les plus corrélées avec la variable dépendante, et les moins corrélées entre elles
  - Critère pour déterminer le meilleur modèle
    - Maximisation du R2 ajusté
    - Minimisation du critère AIC,BIC et SIC

Choix entre deux modèles:

- Utilisation de procédures statistiques permettent de déterminer les variables à retirer/garder
  - Remarque : Démarches purement statistiques, sans raisonnement économique ou logique .
  - Attention à l'interprétation

Choix entre deux modèles:

Procédures de sélection automatique

Backward elimination == Élimination progressive

1. On part du modèle avec les k variables explicatives
2. On élimine de proche en proche les variables explicatives dont les t de Student sont en dessous du seuil critique

Stepwise regression == Sélection progressive

1. On sélectionne d'abord les variables dont le coefficient de corrélation simple est le plus élevé avec la variable y ;
2. Après chaque incorporation de variable, on élimine celles dont

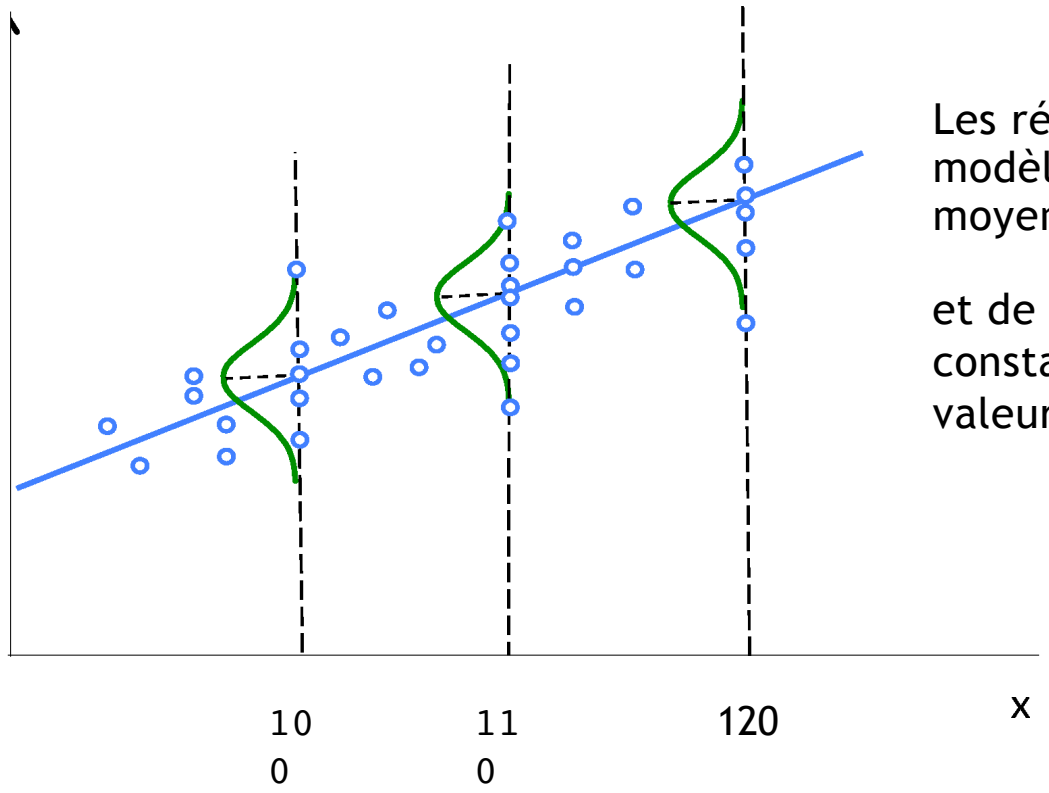
le t de Student est inférieur au seuil critique (non significatif)

Hétéroskedasticité :

- Remise en cause de l'hypothèse d' homoskedasticité
  - La variance des perturbations doit être constante
- Remise en cause de l'i.i.d des erreurs, de moyenne nulle et de variance constante
  - Remise en cause de l ' hypothese u~N(0, V)
  - Remise en cause de la loi suivie par l'estimateur ...

Hétéroskedasticité :

V(u) est constante pour toute valeur de x : les résidus sont homoscédastiques

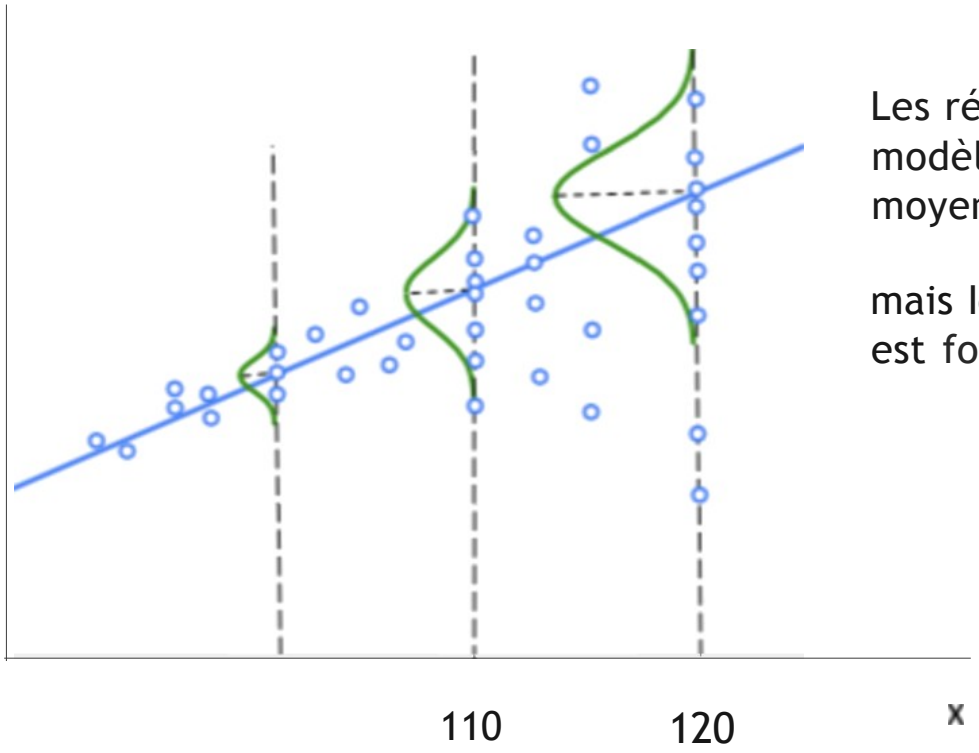


Les résidus du modèle sont de moyenne nulle

et de variance constante Pour toute valeur de X

Hétéroskedasticité :

V(u) n'est pas constante pour toute valeur de x : les résidus sont hétéroscédastiques



Les résidus du modèle sont de moyenne nulle

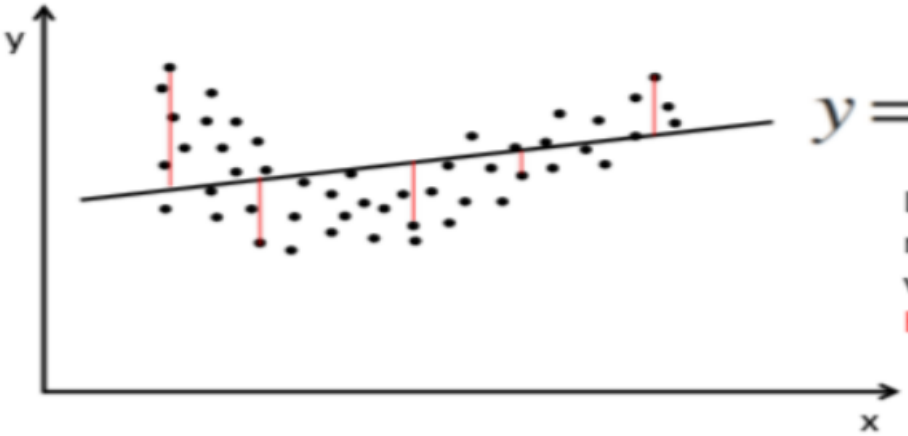
mais leur variance est fonction de X

Hétéroskedasticité :

Causes possibles :

- Mauvaise spécification du modèle
  - —Conséquence d'une variable omise
  - —La variance des résidus est dépendante des valeurs de cette variable omise

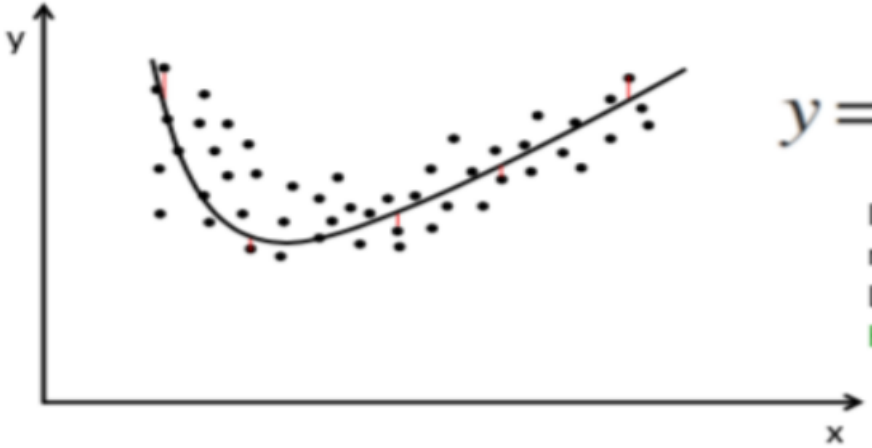
Hétéroskedasticité :



$$y = \beta_0 + \beta_1 x + u$$

L'omission d'une variable ($v = x^2$) rend la variance des résidus variable selon les valeurs de X

Hétéroscédasticité



$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

Ici, les résidus ont la même valeur moyenne pour toute valeur de X

La variance des résidus est constante

Homoscédasticité

Choix entre deux modèles:

Causes possibles :

- L'existence de clusters
  - Ex: les données d'enquêtes ménage, les données regroupant des pays, des entreprises de même secteur, etc.
  - Au sein du même cluster les observations peuvent se comporter
    d'une manière relativement similaire
  - Mais peuvent différer des observations des autres clusters.

Hétéroskedasticité :

- Conséquences
  - Estimateur MC0 calculable

Calculable : $\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$

Sans biais : $E(\hat{\beta}_1) = \beta_1$

Convergent : Quand $N \to \infty$, $V(\hat{\beta}_1) \to 0$.

- ... Mais estimateur MC0 inefficace
  - Variance non minimisée
  - Estimation de la variance biaisée, écart-types des paramètres différents
  - Perte d'efficacité, de précision l'OLS n'es plus BLUE => emploi de méthodes GLS
  - Souci d'inférence statistique et de validité des tests :
    - Les tests de Student (t ne suit plus une loi de Student) et Fisher (F ne suit plus une loi de Fisher)

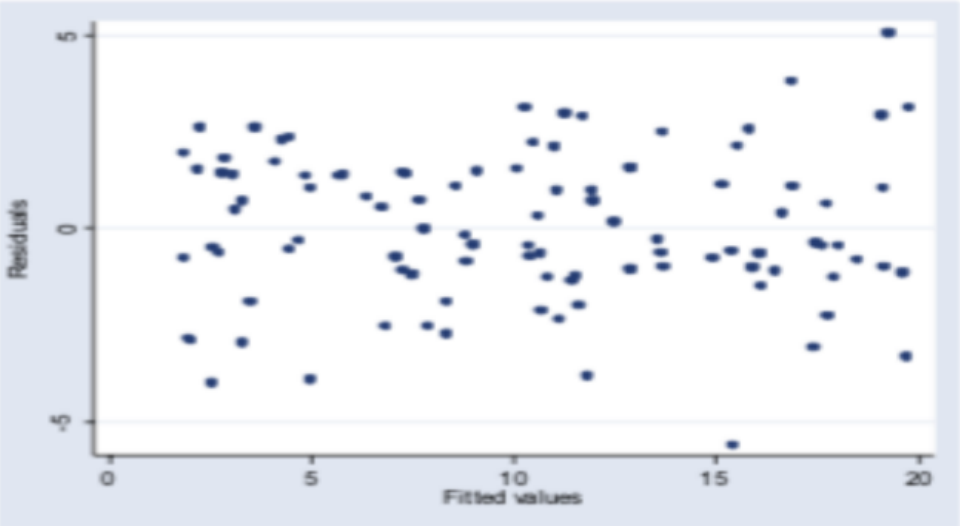
Hétéroskedasticité :

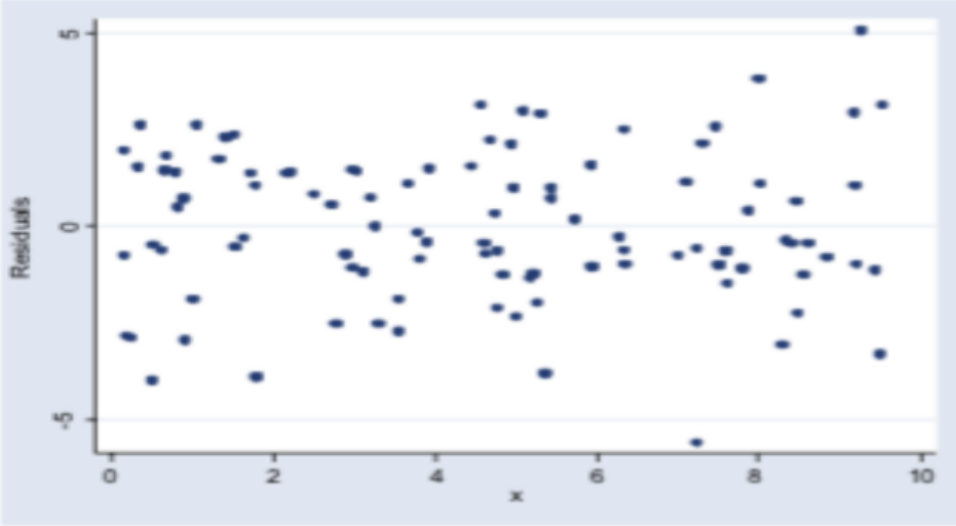
Tests de détection

- Approche visuelle
  - L'analyse des résidus vs y ou x¡
  - La distribution des résidus doit être aléatoire
- Approche formelle
  - Le test de Breusch-Pagan
  - Le test de White

Hétéroskedasticité :

- Approche visuelle
  - Distribution aléatoire des résidus $|\hat{y}_i$ ou $|x_i$

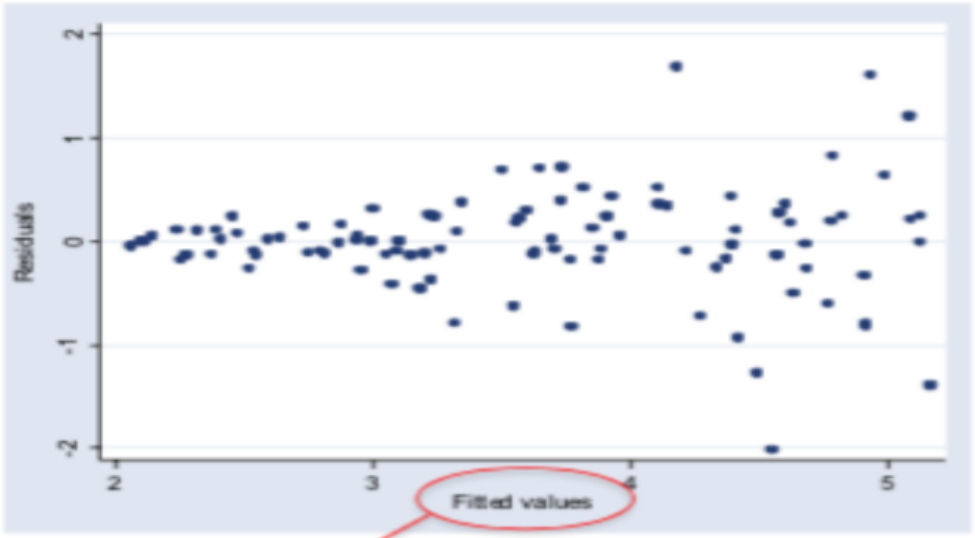




Homoscédasticité

Hétéroskedasticité :

- Approche visuelle
  - Distribution des résidus = $f(\hat{y}_i)$ ou = $f(x_i)$



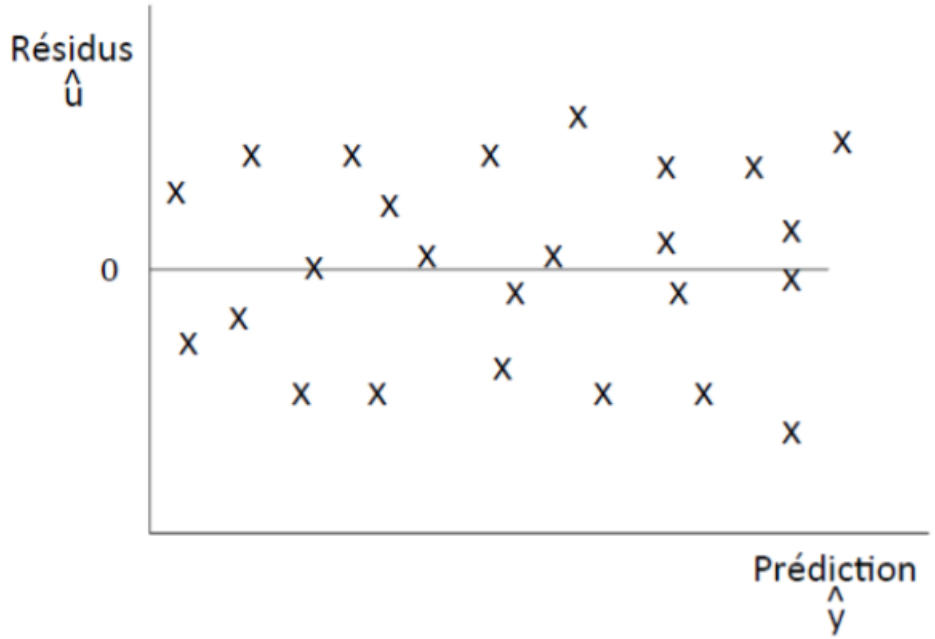
$\hat{y}_i$ révèle la présence générale d'hétérosc. dans le modèle

Hétéroscédasticité

La variable x_j explique la présence d'hétérosc. dans le modèle

Hétéroskedasticité :

- Approche visuelle



La distribution attendue !

Les résidus sont distribués aléatoirement autour de $\hat{y}$

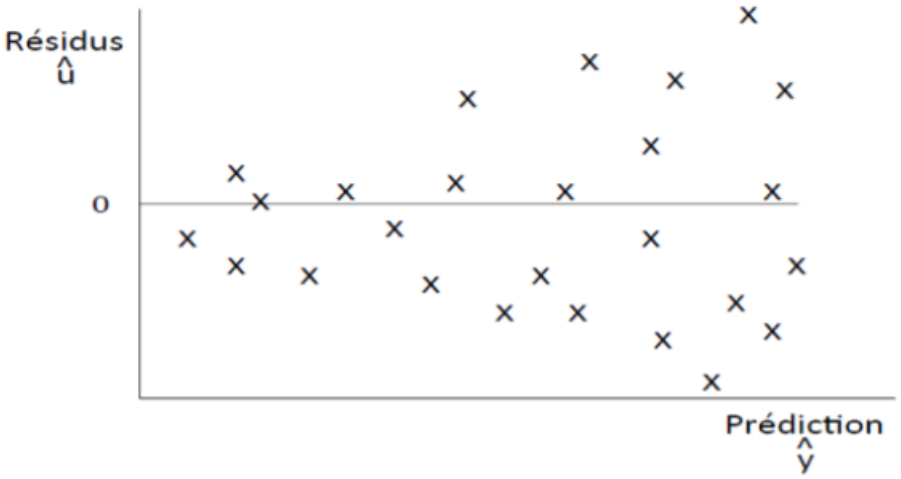
H_0 « spécification correcte »
= non rejetée

Rappel : $E(\hat{u}) = 0$

- Approche visuelle

  Prédiction et analyse des résidus

  - La spécification du modèle



Hétéroscédasticité

Les résidus ne sont pas distribués aléatoirement autour de $\hat{y}$

So

Tests de détection

- Le test de Breusch-Pagan
  - Intuition : la variance des résidus est fonction des variables explicatives

$H_0 : V(u_i) = \sigma^2, \forall n$
$\rightarrow$ Hypothèse d'homoscédasticité

vs

$H_1 : V(u_i) = \sigma_i^2 = \lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + ... + \lambda_k x_{ik}$
$\rightarrow$ Hypothèse particulière sur la forme de l'hétéroscédasticité

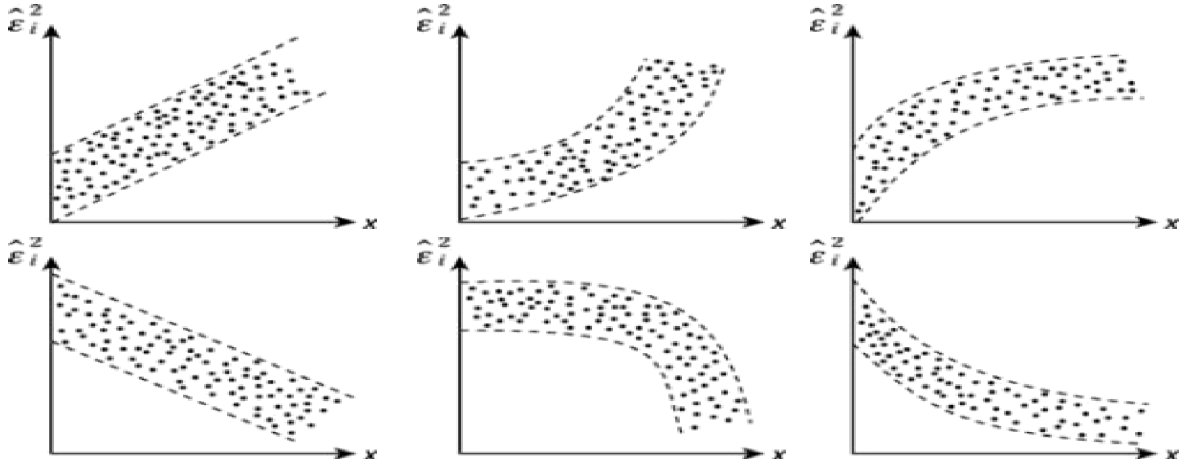
Le test de Breusch-Pagan

Les étapes

1. On estime le modèle initial par les MCO : $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + u_i$ et on récupère les résidus $\hat{u}_i^2$
2. On estime le modèle auxiliaire : $\hat{u}_i^2 = \lambda_0 + \lambda_1 x_{i1} + ... + \lambda_k x_{ik} + \vartheta_i$
3. On calcule le R^2 du modèle auxiliaire=R_a^2
4. On calcule la statistique de Breush-Pagan : $BPc = N \times R_a^2 \sim X_{k_a}^2$; avec k_a le nombre de variables explicatives du modèle auxiliaire
5. Si au seuil de 5%, $BPc < X_{k_a}^2$, on ne rejette pas l'hypothèse nulle d'homoscédasticité

Hétéroskedasticité :

- Le test de White
  - —Même principe que le test de Breusch-Pagan : la distribution des erreurs est une fonctions de la valeur des x, sauf qu'on ne sait pas quelle fonction.
  - —Intuition : plus général, formes fonctionnelles plus nombreuses



Hétéroskedasticité :

- Le test de White
  - On ajoute des termes croisés et termes au carré dans l'équation de régressions des résidus

$H_0 : V(u_i) = \sigma^2, \forall n$

$\rightarrow$ Hypothèse d'homoscédasticité

vs

$H_1 : V(u_i) = \sigma_i^2 = h(1, x_{i1}, x_{i2}, ..., x_{ik})$

$\rightarrow$ Hypothèse sur une certaine forme de l'hétéroscédasticité (pas forcément linéaire)

Hétéroskedasticité :

- Régression du carré du résidu sur les variables et leurs produits croisés

Les étapes

1. On estime le modèle initial par les MCO : $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + u_i$ et on récupère les résidus $\hat{u}_i^2$
2. On estime le modèle auxiliaire : $\hat{u}_i^2 = \lambda_0 + \lambda_1 x_{i1} + \lambda_2 x_{i2} + \ldots + \lambda_k x_{ik} + \lambda_{11} x_{i1}^2 + \lambda_{22} x_{i2}^2 + \ldots + \lambda_{kk} x_{ik}^2 + \lambda_{12} x_{i1} x_{i2} + \ldots + \lambda_{1k} x_{i1} x_{ik} + \ldots + \vartheta_i$
3. On calcule le R^2 du modèle auxiliaire$=R_a^2$
4. On calcule la statistique de White : $Wc = N \times R_a^2 \sim X_{k_a}^2$; avec k_a le nombre de variables explicatives du modèle auxiliaire
5. Si au seuil de 5%, $Wc < X_{k_a}^2$, on ne rejette pas l'hypothèse nulle d'homoscédasticité

Hétéroskedasticité :

- Le test de White :
  - Peu fiable avec un grand nombre de régresseurs
  - Nombre de régresseurs : nombre de variables explicatives + constante
  - $k_a + 1 = \frac{K(K+1)}{2}$
    - si 1 variable explicative : $k_a + 1 = cste, x, x^2$ donc $k_a = 3$ régresseurs du modèle auxiliaire
    - si 9 variables explicatives :
      $k_a + 1 = cste, x_{i1}, ..., x_{i10}, x_{i2}^2, ..., x_1 x_2, ...$ donc
      $k_a = \frac{10(10+1)}{2} = 55$ régresseurs du modèle auxiliaire

Mode de correction

Deux stratégies :

- Un estimateur corrigé
  - Trouver un estimateur plus efficient que MC0
  - Les moindres carrés généralisés MCG/GLS
  - Les moindres carrés généralisés réalisables MCQG/FGLS
- Une variance corrigée
  - La variance ou écart-types « robustes »
  - Conserver l'estimateur MC0
  - Corriger seulement la variance de β
  - La méthode de White