# Économétrie II

## L3 Économétrie – L3 MASS

Prof. Philippe Polomé, U. Lyon 2

Année 2014-2015

# Rappel

1. $E(\epsilon_t) = 0 \; \forall t$ : **espérance nulle**
2. ✓ $var(\epsilon_t) = \sigma^2 \; \forall t$ : **Homoscédasticité**
3. ✓ $cov(\epsilon_t, \epsilon_s) = 0 \; \forall t \neq s$ : **Pas d'auto-corrélation**
4. ✓ $E(\epsilon_t x_t) = 0 \; \forall t$ : **Exogénéité**
5. ✓ La matrice X est de plein rang : **Pas de multicolinéarité**
6. Le modèle est **correctement spécifié**
7. La variable dépendante $Y$ est **continue**

# Table des matières

# Introduction

- 2 aspects : Régresseurs – Coefficients
- Forme fonctionnelle
  - Quels régresseurs faut-il inclure dans le modèle ?
  - Sous quelle forme (linéaire, polynomiale, logarithmique...) ?
  - Manque t-il des régresseurs ? Y en a t-il de trop ? Faut-il inclure une constante ?
  - Quelles conséquences ?
- Le modèle est-il effectivement linéaire en les coefficients ?
  - Les coefficients sont-il stochastiques ?
  - Non traité sauf erreurs de mesure en Ch 5

# Table des matières

# Quelle forme fonctionnelle ?

- Une régression linéaire (en les coefficients) peut donner de "bons" résultats même si la relation sous-jacente est non linéaire
  - Approximation locale – théorème de Taylor
    - Si le modèle est $Y = F(X, \beta) + \epsilon$ par exemple,
    - On peut l'approximer par $Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + R$ où $R$ est le **Reste** de Taylor, qui se trouvera dans le terme d'erreur $\epsilon$
  - C'est surtout pour des variables expliquées discontinues que le modèle doit vraiment être non-linéaire
- Une infinité de formes fonctionnelle est envisageable :
  - Logarithmes sur les variables expliquées ou explicatives
  - Formes quadratiques sur les régresseurs
  - Interactions entre régresseurs
  - ...

# Forme fonctionnelle : approche interprétative vs. test

- Interprétative
  - Que dit la théorie économique ?
    - Nature de la relation connue (exponentielle, linéaire. . . ) ?
    - Concavité / convexité attendue ?
  - Quelle interprétation peut-on dériver des résultats ?
    - $y = \alpha_1 + \beta_1 x$
    - $y = \alpha_2 + \beta_2 x + \delta_2 x^2$
    - $y = \alpha_3 + \beta_3 x_1 + \delta_3 x_2 + \gamma_3 x_1 x_2$
    - $y = \alpha_4 + \beta_4 \ln(x)$
    - $\ln(y) = \alpha_5 + \beta_5 \ln(x)$
    - ... interprétation de $\frac{\partial y}{\partial x}$, $\frac{\partial y}{\partial x_1}$, élasticités...

# Table des matières

# Test 1 : significativité

- Tester l'inclusion de termes quadratiques ou de termes croisés par des tests sur la significativité des coefficients associés
  - t-tests, F-tests
- Rapidement contraignant de tester toutes les combinaisons possibles

# Exemple : prix immobilier

- Données *hprice1.gdt* de Wooldridge dans Gretl
  - 88 prix de ventes déclarés de maisons, 1990, Boston
  - *price* = p = **p**rix de vente
  - *lotsize* = s = **s**urface de la propriété
  - *sqrft* = h = surface **h**abitable
  - *bedrooms* = c = nombre de **c**hambres
- $p = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 c + \epsilon$ ou bien
- $p = \beta_0 + \beta_1 s + \beta_{12} s^2 + \beta_2 h + \beta_{22} h^2 + \beta_3 c + \mu$
- Test $H_0 : \beta_{12} = \beta_{22} = 0$

# Exécution dans Gretl

- Menu “ajouter” “carrés des variables sélectionnées”
- Menu MCO $p = \beta_0 + \beta_1 s + \beta_{12} s^2 + \beta_2 h + \beta_{22} h^2 + \beta_3 c + \mu$
- Post-estimation : “Tests” “Omission de variables” : mettre $s^2$ et $h^2$
  - Test Wald : F(2, 82) = 14.5 ; p. valeur ≈ 4e-06
  - Donc R $H_0 : \beta_{12} = \beta_{22} = 0$
  - Les carrés sont significatifs : manquaient dans la forme linéaire ?

# Test 2 : Le test de Ramsey : RESET

- Idée simplificatrice : Au lieu d'inclure toutes les spécifications possibles des régresseurs, tester significativité de **fonctions** de la variable ajustée $\hat{y}$
  - Effet « combiné » des $X$
- Procédure en 4 étapes
  1. Estimation de la forme linéaire : $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$
  2. Valeurs ajustées : $\hat{y}$
  3. Estimation de : $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \nu$
  4. Test de $H_0 : \delta_1 = \delta_2 = 0$ ; test de Fisher $F \sim F_{2.n-k-3}$

# Exemple Gretl : prix immobilier

- Mêmes données que dans l'exemple précédent
- Ramsey "à la main"
  - Estimer $p = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 c + \epsilon$
  - $\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 h + \hat{\beta}_3 c$
    - Post-estimation : "sauvegarder" "valeurs ajustées"
  - Créer le carré de yhat puis le cube
    - "définir une nouvelle variable" : Cube_yhat=yhat^3
  - Estimer $p = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 c + \delta_1 \hat{p}^2 + \delta_2 \hat{p}^3 + \nu$
    - Les coef. de ce modèle ne sont pas interprétables
  - Tester $H_0 : \delta_1 = \delta_2 = 0$
    - R hypothèse : F(2, 82) = 4.67, avec p. valeur = 0.012
    - $\exists$ variabilité dans $p$ qui n'est pas expliquée par les variables incluses mais pourrait l'être par leurs carrés/cubes
- Ramsey en post estimation : "tests"
  - Résultats identiques

# Test 3 : Alternatives non-imbriquées/anidées/emboîtées

- Deux cas polaires

1. Mêmes régresseurs mais formes fonctionnelles $\neq$
   - Par ex. : $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$
   - Contre : $y = \delta_0 + \delta_1 \ln x_1 + ... + \delta_k \ln x_k + \nu$
2. Mêmes formes fonctionnelles mais régresseurs $\neq$
   - Par ex. A : $y = X\beta + \epsilon$
   - Contre B : $y = Z\delta + \nu$
   - avec $X \bigcap Z \neq \emptyset$ (pas nécessairement vide mais peut être vide)

- 2 approches
  - Minzon & Richard : **Principe "englobant"**
  - Davidson & MacKinnon : **Prédiction**

# 1. Mêmes régresseurs mais formes fonctionnelles $\neq$

## Approche Minzon & Richard : Modèle englobant

- Estimation d'un modèle complet incluant toutes les formes fonctionnelles des explicatives (possiblement sans interprétation)
  - $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \ln x_1 + ... + \delta_k \ln x_k + \mu$
  - On teste **2** hypothèses nulles (test F par ex.)
    - $H_{0X} : \beta_1 = ... = \beta_k = 0$
    - $H_{0lnX} : \delta_1 = ... = \delta_k = 0$
  - Si on R $H_{0X}$ mais pas $H_{0lnX}$ alors le modèle en ln est **validé par rapport au modèle en niveau**
- Nombre important de coef. à estimer
- Multicolinéarité (forte corrélation $x_m$ et $\ln x_m$)

# 1. Mêmes régresseurs mais formes fonctionnelles $\neq$

## Approche Davidson & MacKinnon : Prédiction

- Si $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$ est la bonne spécification
  - et donc $y = \delta_0 + \delta_1 \ln x_1 + ... + \delta_k \ln x_k + \nu$ la mauvaise
- Alors $\hat{y} = \hat{\delta}_0 + \hat{\delta}_1 \ln x_1 + ... + \hat{\delta}_k \ln x_k$ ne doit pas être significative dans $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \alpha \hat{y} + \epsilon$
  - Si $\alpha$ n'est pas significatif, le modèle en niveau est ***validé*** par rapport au modèle en ln
  - S'il est significatif, on ne peut conclure
- On procède pareillement avec l'autre modèle

# 2. Mêmes formes fonctionnelles mais régresseurs $\neq$

## Approche Minzon & Richard : Modèle englobant

- Soit $X_2$ les régresseurs présents dans A $y = X\beta + \epsilon$ mais pas dans B $y = Z\delta + \nu$
- Soit $Z_2$ les régresseurs présents dans B mais pas dans A
  - On estime le **modèle englobant selon B** : $Y = Z\delta + X_2\beta_2 + \nu_2$
    - On teste $H_0 : \beta_2 = 0$ (test en $F$)
  - Si $\beta_2$ n'est pas significativement $\neq 0$, alors A n'apporte rien de plus que B
    - On dit que B est englobant par rapport à A
    - Ce qui ***valide*** B
  - On teste pareillement le modèle englobant selon A : $Y = X\beta + Z_2\delta_2 + \epsilon_2$
- Mêmes remarques que pour le cas polaire 1
  - Beaucoup de coef., multicolinéarité

# Interprétation des tests de spécification

- Il est possible qu'aucune des deux spécifications n'apparaisse dominer l'autre :
  - Les deux spécifications sont rejetées
    - Significativité des coefficients pour les deux alternatives dans le test de Davidson et MacKinnon
    - $H_0 : \beta_1 = ... = \beta_k = 0$ et $H_0 : \delta_1 = ... = \delta_k = 0$ sont acceptées dans le test de Minzon et Richard
    - Nécessité de spécifier un modèle plus complet
  - Les deux sont acceptées
    - Les deux spécifications sont « également » acceptables.
    - On peut comparer les valeurs des $R^2$ pour choisir la spécification.
    - Les données sont "trop molles" pour pouvoir trancher (si tant est qu'il faut trancher)
- Rejeter une spécification contre une autre ne signifie pas que la deuxième est la « bonne »
  - Elle pourrait être rejetée contre une troisième

# Exemple Davidson & MacKinnon : prix immobiliers

- Mêmes données que dans l'exemple précédent
- Estimer (par exemple) modèle A :
  $\ln p = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 c + \epsilon$
  - Prédire "in-sample" $\hat{\ln p} = \hat{\beta}_0 + \hat{\beta}_1 s + \hat{\beta}_2 h + \hat{\beta}_3 c$ via Post-estimation : "sauvegarder" "valeurs ajustées"
- Estimer modèle B : $\ln p = \beta_0 + \beta_1 \ln s + \beta_2 \ln h + \beta_3 \ln c + \mu$ et prédire de façon semblable, on note la prédiction $\tilde{\ln p}$ (avec un tilde)
- Tests de validation
  - Estimer modèle A avec $\tilde{\ln p}$ :
    $\ln p = \beta_0 + \beta_1 s + \beta_2 h + \beta_3 c + \alpha_1 \tilde{\ln p} + \epsilon$
    - Si $\alpha_1$ n'est pas significatif, le modèle A est validé
  - Estimer modèle B avec $\hat{\ln p}$ :
    $\ln p = \beta_0 + \beta_1 \ln s + \beta_2 \ln h + \beta_3 \ln c + \alpha_2 \hat{\ln p} + \epsilon$
    - Si $\alpha_2$ n'est pas significatif, le modèle B est validé

# Table des matières

# Régresseur omis (hétérogéneité)

- Modèle correctement spécifié $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
  - Modèle estimé $Y = \beta_0 + \beta_1 x_1 + \nu$
- Alors l'effet du régresseur manquant se retrouve dans l'**erreur** du modèle estimé : $\nu = \beta_2 x_2 + \epsilon$

1. Régresseur manquant n'est corrélé avec **aucun** régresseur présent
   - Hétéroscédasticité vraisemblablement si $var(x_{2t}) \neq var(x_{2s}), t \neq s$
   - Peut-être autocorrélation si $corr(x_{2t}, x_{2s}) \neq 0, t \neq s$
     - Terme d'erreur n'a plus une espérance nulle : ci-dessous
2. Régresseur manquant **est** corrélé à un régresseur présent
   - Alors, en plus des problèmes ci-dessus : **inconsistance**

- **Exemple** de régresseur manquant avec & sans endogénéité : fichier tableur Regr manquant.ods

# Régresseur superflu

- Si le modèle correctement spécifié est $Y = \beta_0 + \beta_1 x_1 + \nu$
- Mais que le modèle estimé est $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- Si le modèle $Y = \beta_0 + \beta_1 x_1 + \nu$ était correctement spécifié au départ, le terme d'erreur est un bruit blanc
  - $\nu = \beta_2 x_2 + \epsilon$, $\nu$ bruit blanc, comme $\beta_2 = 0$, $\epsilon$ est aussi un bruit blanc
    - En particulier, il n'est pas corrélé (avec quoi que ce soit)
  - Donc $E\left(\hat{\beta}_2\right) = 0$
  - Mais si corrélation$(x_1, x_2) \neq 0$, alors il peut apparaître de la **multicolinéarité**, qui induit une perte de significativité de tous les régresseurs (1er semestre)
- **Exemple** fichier tableur Regr manquant.ods

# Étendue du problème

- Inhérent à toute analyse économétrique
  - Certains régresseurs sont inobservables ou difficilement mesurables
    - dynamisme, charisme, capital social d'un individu, esprit d'équipe dans une entreprise ...
  - Certains régresseurs sont indisponibles
    - questions non posées, réponses biaisées sur des sujets sensibles ...
- Il faut bien rester conscient que certaines variables peuvent capter des effets plus larges "confondants" que ce pourquoi elles sont incluses dans le modèle

# Table des matières

# Sélection de régresseur

- En pratique, quel est l'ensemble des régresseurs pertinents ?
  - La théorie n'aide pas toujours car se concentre sur quelques régresseurs
  - p.e. équation de demande d'un bien
    - devrait dépendre du prix et du revenu de l'acheteur,
    - peut aussi dépendre d'autres caractéristiques du bien (packaging, marketing...) ou de l'acheteur (profil sociologique)...
- **Sélection successive** forward step : intégrer régresseurs 1 à 1
  - Dans beaucoup de logiciels
  - À exclure car statistiquement incorrect
    - Les régressions antérieures peuvent présenter de l'inconsistance suite à l'absence de regresseurs pertinents
    - Les tests effectués dans la régression $s$ sont conditionnels aux décisions prises sur les régressions antérieures

# Méthodologie la plus couramment admise

- Partir d'un ensemble général de régresseurs sur base d'arguments économiques / théoriques / intuitifs mais pas statistiques
  - Éviter multicolinéarité
  - Éviter data mining (inclusion de régresseurs sur base d'erreurs de type I ou faux positifs)
- Enlever des régresseurs non significatifs si on juge le modèle trop complexe ou trop colinéaire
  - Mais pas une obligation : intéressant de montrer la non-significativité
  - Utiliser le test de Wald / F (nullité de plusieurs coefficients) et le test des modèles non-emboîtés

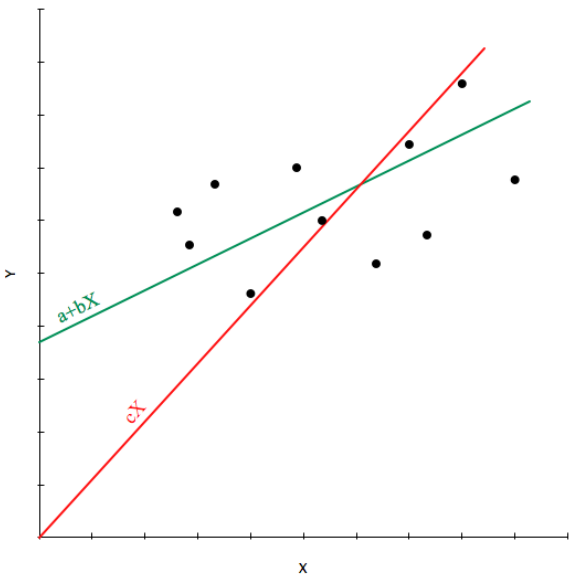
# Table des matières

# $E(\epsilon) \neq 0$

- Ne peut être détecté car $\sum_{i=1}^{n} \hat{\epsilon} = 0$ dès que le modèle comporte une constante (sans preuve)
- Nature du problème
  - $\forall \epsilon$ on peut écrire $\epsilon = a + \mu$, avec $E(\mu) = 0$
  - donc $Y = \beta_0 + \beta_1 X + \epsilon = \beta_0 + \beta_1 X + a + \mu = \gamma + \beta_1 X + \mu$
- $\hat{\gamma}$ biaisé & inconsistant pour $\beta_0$ et/ou $a$ : problème d'**identification**
- Pas d'effet sur autres coefficients
  - **Sauf** si $E(\epsilon) \neq 0$ à cause de l'omission d'un régresseur
  - Dans ce cas, si le régresseur omis est corrélé aux régresseurs présents...

# Flexibilité & rôle de la constante

- Sans point autour du zéro, l'intercept est estimé avec une variance importante
  - Que $E(\epsilon)$ soit $= 0$ ou non n'a alors guère d'importance
  - $\beta_0$ alors paramètre de flexibilité : permet que la pente s'ajuste




- Pour cette raison, en général on inclut une constante, même non significative, dans tout modèle