Notes d'économétrie

Ces notes sont en relation avec le cours d'économétrie donné à l'université de Namur et ont pour but de préciser quelques problèmes rencontrés en économétrie ainsi que leurs solutions, et également de donner quelques précisions sur d'autres points d'intérêts. Ce n'est donc pas un résumé de cours.

Quelles sont les qualités d'un bon estimateur ?

Biais

Une variable aléatoire fluctue autour de son espérance. On peut donc souhaiter que l'espérance de cette variable soit égale à θ, c'est-à-dire qu'en « moyenne » l'estimateur ne se trompe pas.

Lorsque l'espérance de l'estimateur $E(\theta)$ égale θ, i.e. le biais est égal à zéro, l'estimateur est dit **sans biais**.

Pour qu'un estimateur soit sans biais, il faut une **bonne spécification :** $E(\varepsilon_i) = 0$

Erreur quadratique moyenne

L'erreur quadratique moyenne (*mean squared error* en anglais) est l'espérance du carré de l'erreur entre la vraie valeur et sa valeur estimée doit être faible. $EQM(\theta) = \text{var}(\theta)$ si θ est un estimateur sans biais de θ.

Convergence ou consistance

On souhaite aussi pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur commise en prenant θ à la place de θ. Si c'est le cas, on dit que l'estimateur est **convergent** (on voit aussi ***consistant***), c'est-à-dire qu'il converge vers sa **vraie valeur**.

Si un estimateur est sans biais ou asymptotiquement sans biais et si sa variance tend vers 0, alors il est **convergent**.

Pour qu'un estimateur soit convergent (consistant), il faut

- Une bonne spécification. (Les erreurs sont des variables aléatoires). $E(\varepsilon_i) = 0$
- Exogénéité des variables. $cor(\varepsilon_i x_i) = 0$
- Pas de paramètres incidentaux (Le nbre de variables ne grandit pas avec N = effectif total)

Efficacité

La variable aléatoire fluctue autour de son espérance. Plus la variance $\text{var}(\hat{\theta})$ est faible, moins les variations sont importantes. On cherche donc à ce que la **variance soit la plus faible possible**. Un estimateur sans biais pour lequel la borne de Cramér-Rao devient égalité est dit **efficace**. Autrement dit, c'est **l'estimateur qui a la plus faible variance parmi tous les estimateurs sans biais**.

Un estimateur **efficient** est nécessairement convergent, mais l'inverse n'est pas vrai.

- Un estimateur peut être convergent sans être nécessairement efficient.

Un estimateur **efficient** n'est pas nécessairement consistant, ET l'inverse EST vrai.

- Un estimateur consistant ne doit pas nécessairement être efficient.

Pour qu'un estimateur soit efficace, il faut

- Qu'il soit sans biais. $E(\varepsilon_i) = 0$
- Homoscédasticité. $\text{var}(\varepsilon_i) = \sigma^2 = \text{constante}$
- Indépendance sérielle. $cor(\varepsilon_i \varepsilon_j) = 0, i \neq j$

Robustesse.

Il arrive que lors d'un sondage, une valeur extrême et rare apparaisse (par exemple un enfant de 10 ans mesurant 1,80 m). On cherche à ce que ce genre de valeur ne change que de manière très faible la valeur de l'estimateur. On dit alors que l'estimateur est **robuste**.

Exemple : En reprenant l'exemple d'un groupe d'enfants, la moyenne n'est pas un estimateur robuste car ajouter un enfant très grand modifiera beaucoup la valeur de l'estimateur. La médiane par contre n'est pas modifiée dans un tel cas.

Desired Properties	Requirement Assumptions
• $\underline{\hat{\beta}}_{OLS}$ is an unbiased estimate of $\underline{\beta}$	• $E[\epsilon_i] = 0$
• $\underline{\hat{\beta}}_{OLS}$ is an unbiased estimate of $\underline{\beta}$ • $\underline{\hat{\beta}}_{OLS}$ is the BLUE estimator	• $E[\epsilon_i] = 0$ • $Var(\epsilon_i) = \sigma^2 < \infty$, $\forall i$ • $Cov(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$
• $\underline{\hat{\beta}}_{OLS}$ is an unbiased estimate of $\underline{\beta}$ • $\underline{\hat{\beta}}_{OLS}$ is the BLUE estimator • $\underline{\hat{\beta}}_{OLS}$ is mathematically equivalent to the Maximum Likelihood Estimator $\underline{\hat{\beta}}_{MLE}$	• $E[\epsilon_i] = 0$ • $Var(\epsilon_i) = \sigma^2 < \infty$, $\forall i$ • $Cov(\epsilon_i, \epsilon_j) = 0$, $\forall i \neq j$ • $\epsilon_1, \dots, \epsilon_n \sim Normal$, are identically and independently distributed (i.i.d.)

Conséquences du non-respect des hypothèses de Gauss-Markov sur les estimateurs

Bonne spécification	$E(\varepsilon_i) = 0$	
	Omission d'une variable	Coefficients biaisés Ecart-types invalides
	Inclusion d'une variable non pertinente	Coefficients non efficaces. Ecart-types invalides
Exogénéité	$cor(\varepsilon_i x_i) = 0$	Coefficients biaisés. Coefficients non convergents
Homoscédasticité	$\text{var}(\varepsilon_i) = \sigma^2 = \text{constante}$	Ecart-types invalides
Absence d'autocorrélation (Indépendance sérielle)	$cor(\varepsilon_i \varepsilon_j) = 0, i \neq j$	Coefficients non efficaces. Ecart-types invalides

Note : quand les écart-types sont invalides, alors les tests post hoc deviennent inapplicables.

Types de données

Données en coupe transversale

Une base de données en coupe transversale est composée d'un échantillon d'individus, ménages, entreprises, villes, États, pays, ou autres unités, observés à un certain moment dans le temps. Il arrive que les données n'aient pas été recueillies exactement au même moment pour l'ensemble des unités d'observation. Par exemple, lors d'une enquête, plusieurs familles peuvent être interrogées au cours de différentes semaines d'une même année.

Tableau 1.1
Base de données en coupe transversale indiquant les salaires et d'autres caractéristiques individuelles

obsno	wage	educ	exper	female	married
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1

Séries chronologiques ou données temporelles.

Une base de séries chronologiques est composée d'une ou de plusieurs variables observées au cours du temps à plusieurs reprises. Comme exemples de séries chronologiques, on peut citer les prix des actions, l'offre de monnaie, l'indice des prix à la consommation, le produit intérieur brut, le taux d'homicides par an, et le chiffre d'affaires de l'industrie automobile.

Tableau 1.3
Salaire minimum, chômage et données associées pour le Puerto Rico

obsno	year	avgmin	avgcov	prunemp	prgnp
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1 015,9

Données empilées

Certaines bases de données ont à la fois des caractéristiques propres aux coupes transversales et aux séries chronologiques. Supposons par exemple que l'on mène aux États- Unis deux enquêtes sur les ménages, l'une en 1985 et l'autre en 1990. En 1985, nous tirons aléatoirement un échantillon de ménages à partir desquels nous obtenons des informations sur le revenu, l'épargne, la taille de la famille, etc. En 1990, un nouvel échantillon de ménages est tiré aléatoirement ; l'enquête est similaire et permet de récolter le même type de données. Afin d'accroître la taille de notre échantillon, peut combiner les deux années pour construire des données empilées

Tableau 1.4
Données empilées : les prix de l'immobilier pour deux années

obsno	year	hprice	proptax	sqrft	bdrms	bthrms
1	1993	85 500	42	1 600	3	2,0
2	1993	67 300	36	1 440	3	2,5
3	1993	134 000	38	2 000	4	2,5

Données de panel ou données longitudinales

Une base de données de panel (ou données longitudinales) contient des séries chronologiques pour chacune des unités reprises dans la coupe transversale. Par exemple, une telle base de données vous

permet d'observer le salaire, le niveau d'étude et l'expérience professionnelle d'un ensemble d'individus que l'on suit au cours du temps. Il est également possible de recueillir des informations sur la structure financière et les investissements pour un même groupe d'entreprises. Les données en panel peuvent aussi concerner des unités géographiques. Par exemple, considérant un ensemble fixe de comtés aux États- Unis, nous pouvons obtenir, pour les années 1980, 1985 et 1990, des données sur les flux d'immigration, les taux d'imposition, les taux de salaire, les dépenses publiques, etc.

La personne	An	Revenu	Âge	Sexe
1	2013	20 000	23	F
1	2014	25 000	24	F
1	2015	27 500	25	F
2	2013	35 000	27	M
2	2014	42 500	28	M
2	2015	50 000	29	M

Les Moindres Carrés Ordinaires MCO (OLS : Ordinary Least Square)

Le modèle de la population peut être écrit comme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

où Y est la variable expliquée, les X_i les variables explicatives et les $\beta_0, \beta_1, \beta_2, ... \beta_k$ sont les paramètres inconnus qui nous intéressent et u est le terme d'erreur aléatoire non observé, ou perturbation.
Cette équation représente le **modèle de la population**, aussi appelé **vrai modèle**.

Un modèle est dit **linéaire** s'il est linéaire par rapport aux β_i . Il n'est pas nécessaire que les relations avec les variables explicatives soient linéaires. Par exemple :

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \beta_3 X_3^2 + u$$

MCO sous forme matricielle

$$\boldsymbol{Y} = \boldsymbol{bX} + \boldsymbol{\varepsilon}$$

$\boldsymbol{Y}$	Vecteur des observations dépendantes
$\boldsymbol{b}$	Vecteur des coefficients de régression
$\boldsymbol{X}$	Matrice des données explicatives
$\boldsymbol{\varepsilon}$	Vecteur des résidus (iid)

Les **coefficients de régression** sont estimés par

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$\boldsymbol{b}$	Vecteur des coefficients estimés
$\boldsymbol{X}$	Matrice des données
$\boldsymbol{X}'$	Transposée de $\boldsymbol{X}$
$\boldsymbol{Y}$	Vecteur des données dépendantes

Variance des résidus

$$\boldsymbol{\sigma}_{\varepsilon}^2 = \frac{SCR}{n-k-1} = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k-1}$$

$\boldsymbol{\sigma}_{\varepsilon}^2$	Estimateur de la variance des résidus
SCR	Somme des carrés des résidus
$\boldsymbol{e}$	Vecteur des résidus
$\boldsymbol{e}'$	Vecteur transposé de $\boldsymbol{e}$
n	Nombre de données
k	Nombre de variables explicatives

Matrice de variance-covariance

$$\boldsymbol{\Omega}_{\boldsymbol{b}} = \boldsymbol{\sigma}_{\varepsilon}^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

Par construction, le modèle est linéaire en X (ou sur ces coefficients) et nous distinguons les hypothèses stochastiques (liées à l'erreur e) des hypothèses structurelles.

Hypothèses stochastiques

- H1 : les valeurs $x_{i,t}$ sont observées sans erreur.
- H2 : $E(\varepsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle.
- H3 : $E(\varepsilon_t^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante $(\forall t)$ (homoscédasticité).
- H4 : $E(\varepsilon_t\, \varepsilon_{t'}) = 0$ si $t \neq t'$, les erreurs sont non corrélées (ou encore indépendantes).
- H5 : $\mathrm{Cov}(x_{it}, \varepsilon_t) = 0$, l'erreur est indépendante des variables explicatives.

Hypothèses structurelles

- H6 : absence de colinéarité entre les variables explicatives, cela implique que la matrice $(X'\,X)$ est régulière et que la matrice inverse $(X'\,X)^{-1}$ existe.
- H7 : $(X'\,X)/n$ tend vers une matrice finie non singulière.
- H8 : $n > k+1$, le nombre d'observations est supérieur[1] au nombre des séries explicatives.

Signification des paramètres :

Note : toutes les interprétations se font *ceretis paribus* c'est-à-dire toutes autres choses étant égales par ailleurs (TACEPA).

Signification de l'intercept (la constante) β_0

1er cas : Son interprétation a un sens

C'est la valeur moyenne de la variable dépendante quand toutes les variables explicatives sont mises à zéro.

Exemple : $Score\ Examen = 60 + 2.5\ NbrHeureEtude$. Donc si *NbrHeureEtude* est égal à zéro, la cote moyenne sera de 60.

2eme cas : Son interprétation n'a pas de sens

On ne doit pas chercher une interprétation quelconque.

Exemple : $Surface = -30 + 3.5 Prix$. Une surface négative n'a pas de sens.

Modèle log-log et log-lin

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \beta_3 X_3^2 + \beta_4 D + u$$

Ces modèles permettent de diminuer l'effet des outliers.

- β_0 est la constante.
- β_1 est l'élasticité de Y par rapport X_1. Elle peut être interprétée comme le changement de β_1 % de Y salaire induite par un changement de 1% de X_1, toutes choses égales par ailleurs. Notons qu'il convient de parler d'élasticité partielle puisque la régression prend en compte d'autres variables.
- $\beta_2 = \dfrac{\partial Y / Y}{\partial X_1}$ alors $100\,\beta_2$ donne approximativement le changement en pourcentage de Y quand X_2 augmente d'une unité. Autrement dit, si X_2 augmente d'une unité alors Y augmente de

$100\,\beta_2\,\%$. Le résultat doit être en pourcentage puisqu'il s'agit d'une variation relative de *Y*. Notons que les variations de X_2 doivent être prises en termes absolus pour l'interprétation.

- β_3 dépend de X_3 qui est de forme quadratique. Son interprétation dépend du signe de β_3. $\beta_3 > 0$, on a une parabole à minima, donc un effet qui diminue avant ce minima et qui augmente après. $\beta_3 < 0$, on a une parabole à maxima, avec des effets inverses. L'extrema est déterminé par $-\frac{\beta_2}{2\beta_3}$. (C'est l'axe de la parabole). Voir plus loin la note sur la forme quadratique.
- β_4 : *D* étant une variable dichotomique (valant 0 ou 1). Quand *D* passe de 0 à 1, le taux de croissance de *Y* est : $g = \frac{Y_1 - Y_0}{Y_0} = e^{\beta_4} - 1$

Modèle lin-lin ou lin-log

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 X_2 + \beta_3 X_3^2 + u$$

- β_0 est la constante.
- $\beta_1 = \frac{\partial Y}{(\partial X_1 / X_1)}$: lorsque X_1 augmente de 100% (ou double), *Y* augmente de β_1 unités.

  Lorsque X_1 augmente d'1%, *Y* augmente de $\frac{\beta_1}{100}$ unités. Le résultat doit être en unités de *Y* puisqu'il s'agit d'une variation absolue de *Y*. Par contre les variations de X_1 doivent être prises en termes relatifs pour l'interprétation.
- β_2 : lorsque X_2 augmente de 1 unité *Y* augmente de β_2 unités.
- β_3 : dépend de X_3 qui est de forme quadratique. Son interprétation dépend du signe de β_3. $\beta_3 > 0$, on a une parabole à minima, donc un effet qui diminue avant ce minima et qui augmente après. $\beta_3 < 0$, on a une parabole à maxima, avec des effets inverses. Voir la note sur la forme quadratique

Résumé des principales interprétations (cerebis paribus)

Modèle	**Equation**	**Modification de X_1**	**Conséquence sur *Y***
Lin-Lin	$Y = \beta_0 + \beta_1 X_1$	X_1 augmente de 1 unité	*Y* augmente de β_2 unités.
Lin-Log	$Y = \beta_0 + \beta_1 \ln X_1$	X_1 augmente de 1%	*Y* augmente de $\frac{\beta_2}{100}$ unités
Log-Lin	$\ln Y = \beta_0 + \beta_1 X_1$	X_1 augmente de 1 unité	*Y* augmente de $100\beta_2\%$
Log-Log	$\ln Y = \beta_0 + \beta_1 \ln X_1$	X_1 augmente de 1%	*Y* augmente de $\beta_2\%$

La forme quadratique

Soit le modèle

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$$

Une forme quadratique implique une relation non-monotone.

- Il s'agit d'une parabole dont l'extrema est donné par $-\frac{\beta_1}{2\beta_2}$. Cet extrema est généralement appelé « **Point de retournement** ».
- Si $\beta_2 > 0$, on a une parabole à minima. Il y a donc un point où l'effet de X_1 est minimal. Si $\beta_2 > 0$, il y a un point où l'effet de X_1 est maximal.

		Signe du terme au carré	
		Positif	Négatif
Signe du terme simple	Positif	Type exponentielle	U-inversé
	Négatif	Forme en U	Type « falaise »

Effet « falaise »: diminution rapide d'un phénomène.

Quelques erreurs à ne pas commettre lorsqu'on spécifie un modèle.

Soit Y une variable à expliquer. X_1 une variable explicative. D_1 et D_2 deux variables dichotomiques avec $D_1 = 1 - D_2$

Exemple 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 D_1 + \beta_4 D_2 + u$$

Modèle mal spécifié car il y a une colinéarité entre D_1 et D_2. L'un est une combinaison linéaire de l'autre. (Corrélation = –1)

Exemple 2

$$Y = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_1^2) + \beta_3 D_1 + u$$

Modèle mal spécifié car il y a une colinéarité entre $\log(X_1)$ et $\log(X_1^2)$. L'un est une combinaison linéaire de l'autre : $2\log(X_1) = \log(X_1^2)$. (Corrélation = 1)

Exemple 3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 D_1 + \beta_4 D_1^2 + u$$

Modèle mal spécifié car le carré d'une variable binaire n'a pas de sens

Exemple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 \log(D_1) + u$$

Modèle mal spécifié car le log d'une variable binaire n'a pas de sens $(\log(0) = -\infty, \log(1) = 0)$

Modèles non linéaires

Pour que les résultats de régression en modèle linéaire aient les propriétés désirées, le **terme d'erreur** doit être **multiplicatif** dans le **modèle original** et **additif** dans le **modèle transformé**. Il doit également satisfaire aux conditions de Gauss-Markov et en particulier être normalement distribué dans le modèle transformé.

Modèle original	$Y = \beta_1 X^{\beta_2} e^u$
Modèle transformé	$\ln Y = \ln \beta_1 + \beta_2 . X + u$

Pour que u soit distribué normalement, il faut que $v = e^u$ soit distribué selon une log-normale.

Si le terme d'erreur est **additif** dans le modèle **original**, il n'est pas possible de linéariser. Il faut alors utiliser des méthodes spécifiques qui n'entrent pas dans l'objet de cette note.

Modèle original	$Y = \beta_1 X^{\beta_2} + u$
Modèle transformé	$\ln Y = \ln\left(\beta_1 X^{\beta_2} + u\right)$

Note

- Le modèle de régression avec variable dépendante binaire est appelé **modèle à probabilités linéaire** MPL (En anglais MLP)

Comment identifier les variables les plus importantes dans un modèle de régression ?

Ne comparez pas les coefficients de régression pour déterminer l'importance des variables

Les coefficients de régression décrivent la relation entre variables prédictives et réponse. La valeur du coefficient représente la variation moyenne de la réponse en fonction d'une augmentation d'une unité du prédicteur. Il est tentant d'en conclure que les variables ayant des coefficients de régression plus importants génèrent un changement plus important dans la réponse, mais cela pourrait vous induire en erreur. Les échelles des unités sont souvent différentes pour les variables prédictives, ce qui rend impossible une comparaison directe. Par exemple, si, pour un même jeu de données, vous utilisez des grammes dans un modèle et des kilogrammes dans un autre, les coefficients de régression pour le poids seront complétement différents alors que l'importance de la variable reste identique.
Les coefficients les plus importants peuvent être affectés par des effets d'échelle, et ne permettent pas d'identifier les variables prédictives les plus importantes.

Ne comparez pas les valeurs de *P* pour déterminer l'importance relative des variables

Des valeurs de *p* faibles indiquent qu'une variable est significative et qu'elle doit être conservée dans le modèle. Cependant, la valeur de *p* prend en compte des caractéristiques qui n'ont rien à voir avec l'importance d'une variable, par exemple la précision de l'estimation et la taille de l'échantillon. Même des effets peu importants dans le monde réel pourraient avoir des valeurs de *p* très faibles. Un effet statistiquement significatif risque d'avoir, en réalité, peu d'influence pratique. Les valeurs de *p* faibles ne permettent pas nécessairement d'identifier les variables prédictives qui sont importantes en pratique.

Solutions

- **Méthode** : Pour déterminer, les variables importantes, il faut soit utiliser des **Coefficients de régression normalisés/standardisés/codés** pour que les coefficients de régressions deviennent comparables. Soustraire la moyenne, puis divisez par l'écart-type permet de centrer les prédicteurs et de les placer sur une échelle comparable. Recherchez ensuite la variable prédictive avec la plus grande valeur absolue pour le coefficient codé.

- **Méthode 2 : Augmentation du R^2 lorsque la variable est ajoutée, en dernier, dans le modèle.** La valeur du R^2 représente la proportion de la variation de la réponse qui est expliquée par un modèle. Cette analyse du R^2 traite chaque variable comme la dernière entrée dans le modèle, l'augmentation représente le pourcentage de variation de la réponse, qui est expliqué intrinsèquement par cette variable et que les autres variables déjà présentes dans le modèle ne pouvaient pas expliquer. Intéressez-vous à la variable prédictive qui est associée à la plus grande augmentation du R^2.

- **Méthode 3 : test *F* marginaux de qualité d'ajustement**. Il s'agit de tester si l'ajout d'une variable améliore de façon significative la qualité de l'ajustement. Si l'augmentation de la qualité de l'ajustement n'est pas significative, c'est que la variable est de peu d'intérêt.

$$F\left(\text{coût, d.l. disponibles}\right)=\frac{\dfrac{\text{Amélioration}}{\text{Côut}}}{\dfrac{\text{Partie résiduelle non expliquée}}{\text{Degrés de liberté disponibles}}}$$

- **Amélioration** : réduction de la somme des carrés résiduels une fois le changement fait, dans le cas présent, lorsque le groupe de nouvelles variables est ajouté : $RSS_2 - RSS_1$.
- **Coût** : réduction du nombre de degrés de liberté restants après avoir fait le changement. Dans ce cas-ci, la perte est égale au nombre de nouvelles variables ajoutées, car il s'agit du nombre de nouveaux paramètres estimés.
- **Partie résiduelle non expliquée** : RSS_2
- **Degrés de liberté disponibles** : le nombre d'observations moins le nombre de paramètres estimés.

Comment juger de la qualité d'une régression ?

Régression linéaire.

Le R^2 ou coefficient de détermination.

$$R^2=\frac{ESS}{TSS}$$

ESS : Somme des carrés des écarts factoriels

TSS : Somme des carrés des écarts totaux

- Le coefficient de détermination exprime la variabilité qui est expliquée par le modèle.
- Le R^2 est le carré du coefficient de corrélation uniquement pour les régressions linéaires avec une seule variable explicative.
- *Notes :*
  - ***Le r ou coefficient de corrélation*** *(une seule variable explicative)* **n'exprime pas** la qualité de la régression mais mesure la force de liaison entre X et Y.
  $$r=\frac{\operatorname{cov}\left(X,Y\right)}{S_X S_Y}$$
  - Ne pas confondre le coefficient de corrélation avec les coefficients de régression qui sont le b_i.

Le R^2 ajusté.

$$R_a^2=1-\frac{n-1}{n-k}\left(1-R^2\right)$$

R^2 Coefficient de détermination

n Taille de l'échantillon

k Nombre de variables explicatives

- On note que R^2 augmente toujours quand le nombre de variables explicatives augmente.
- Le R^2 corrigé (ou "ajusté") peut être utilisé pour prendre en compte la perte de degrés de liberté lié à l'inclusion de variables explicatives dans le modèle estimé.
- Si l'addition d'une variable provoque une diminution du R^2 ajusté, c'est que cette variable n'est pas pertinente.

Test de Fisher

La validité du coefficient de détermination peut être testé par un test de Fisher :

$$F\left(k-1,n-k\right)=\frac{R^2}{1-R^2}.\frac{n-k}{k-1}$$

$F\left(k-1,n-k\right)$	Statistique F à $k-1,n-k$ dégrés de liberté
R^2	Coefficient de détermination
n	Nombre d'observations
k	Nombre de variables explicatives

Test du rapport de vraisemblance (LR)

Ce test est un test de la nullité des coefficients introduits dans le modèle. Souvent noté « Likelihood ratio » (LR), il correspond au log du ratio de vraisemblance entre le modèle non contraint (tous les paramètres sont libres) et le modèle contraint (certains ou tous les coefficients sont nuls, par exemple).

$$LR=2\left(\log L_U-\log L_R\right)\sim\chi^2_k$$

où L_U et L_R sont les vraisemblances des modèles non contraint et contraint. Sous l'hypothèse H_0, les contraintes sont satisfaites. La statistique de *LR* suit un χ^2 à k degrés de liberté (correspondant au nombre de contraintes posées – nombre de paramètres égalés à 0). Si on rejette H_0, c'est que le modèle non contraint est meilleur.

Ce test est aussi utilisé dans l'analyse TOBIT comme test de **biais de sélection**.

Régression logistique (Logit, Probit, Tobit)

La régression logistique s'applique au cas où :

- *Y* est qualitative à 2 modalités
- X_i qualitatives ou quantitatives

On utilise les 3 indicateurs suivants :

Le count R² ou pourcentage de prédictions correctes.

$$\text{Count } R^2=\frac{\text{Nombre de prédictions correctes}}{\text{Nombres dtotal d'observations}}$$

- Plus le count R^2 est élevé meilleur est le modèle.
- Il existe une version count R^2 ajusté qui tient compte du nombre du nombre attendus d'observations correctes à priori.
- Le count R^2 peut donner des valeurs anormalement élevées, principalement quand la variable prédite est relativement rare. Il faut alors ajuster le seuil à partir duquel une valeur sera considérée comme une prédiction correcte.

Le count R² ajusté

Le Count R² peut donner l'impression que le modèle donne de bonnes prédictions, alors que ce ne l'est pas. Dans un modèle binaire, sans connaissance au sujet des variables indépendantes, il est possible de prédire correctement au moins 50% des cas en choisissant la catégorie de résultat avec le plus grand pourcentage de cas observés. Par exemple, dans un échantillon on sait que 57% des femmes travaillent, si on prédit que toutes les femmes travaillent, on aura 57% de réussite.

TABLE 4.3 Classification Table of Observed and Predicted Outcomes for a Binary Response Model

Observed Outcome	*Predicted Outcome* $\hat{y}=1$	*Predicted Outcome* $\hat{y}=0$	*Row Total*
$y=1$	n_{11} :: correct	n_{12} :: incorrect	n_{1+}
$y=0$	n_{21} :: incorrect	n_{22} :: correct	n_{2+}
Column Total	n_{+1}	n_{+2}	N

Le count R² doit donc être ajusté pour tenir compte de la plus grande rangée marginale. Ce qui peut se faire via la formule

$$R^2_{\text{Adj Count}} = \frac{\sum_j n_{jj} - \max_r \left(n_{r+}\right)}{N - \max_r \left(n_{r+}\right)}$$

Où $\sum_j n_{jj}$ est le nombre correct d'observations. n_{r+} est la somme marginale de la rangée r. $\max_{r+}$ est la rangée présentant le maximum d'observations. N est le nombre total d'observations.
C'est la proportion de suppositions correctes au-delà du nombre qui serait correctement devinée en choisissant la rangée avec la plus grande somme marginale.

Le pseudo R² de McFadden pour les réponses binaires.

$$\text{McFadden} = 1 - \frac{\log L.\ full}{\log\ L.Constant}$$

$\log L.\ full$: Log vraisemblance du modèle estimé

$\log L.Constant$: log vraisemblance du modèle avec seulement une constante (l'intercept).

- Son interprétation est similaire à celle du R^2 en régression linéaire. Notez que les valeurs sont cependant souvent beaucoup plus basses. Par exemple des valeurs de 0,2 à 0,4 correspondent à une corrélation excellente.
- Il existe une version ajustée qui tient compte du nombre de variables explicatives.

$$\text{McFadden Ajusté} = 1 - \frac{\log L.\ full - NV}{\log\ L.Constant}$$

NV = nombre de variables explicatives du modèle.

Le critère d'information d'Akaike AIC

$$AIC = 2k - 2\ln L$$

L Vraisemblance

k Nombre de paramètres

- Quand de nombreux modèles doivent être comparés entre eux (MCO vs. Probit, Probit vs Probit ou Logit,…), le risque de rejeter l'hypothèse nulle alors qu'elle est vraie augmente substantiellement. Une solution possible (il y en a d'autres) consiste à comparer les modèles en utilisant l'*AIC* représente donc un compromis entre le biais (qui diminue avec le nombre de paramètres) et la parcimonie (nécessité de décrire les données avec le plus petit nombre de paramètres possible).
- Le meilleur modèle est celui possédant l'*AIC* le plus faible
- Quand le nombre de paramètres k est grand par rapport au nombre d'observations n i.e., si $n / k < 40$, il est recommandé d'utiliser l'*AIC* corrigé.

Le Bayesian Information criterion (BIC)

$$BIC = k\ln(n) - 2\ln L$$

L Vraisemblance

k Nombre de paramètres

n Nombre d'observations

- Semblable au *AIC*. Plus le *BIC* est petit, meilleur est le modèle.
- Comme l'*AIC*, le *BIC* dépend de l'unité de mesure de la variable dépendante. Si la variable n'est pas à la même échelle, les critères ne sont pas applicables.

Le maximum de vraisemblance

Le maximum de vraisemblance est une méthode générale pour estimer les paramètres d'un modèle statistique. Par exemple, supposons que nous avons une série d'observations d'une variable aléatoire *Y* et un modèle statistique potentiel pour cette variable. Ce modèle peut inclure la dépendance de *Y* sur d'autres variables explicatives, ainsi qu'une distribution statistique pour la portion non-expliquée de la variation de *Y*.
En général, un tel modèle contient différents paramètres inconnus qui doivent être ajustés aux données observées. Selon le maximum de vraisemblance, les meilleures estimations des paramètres d'un modèle sont celles qui maximisent la probabilité des valeurs observées de la variable. Cette méthode peut être appliquée peu importe la forme mathématique du modèle.

Par exemple, selon la méthode du maximum de vraisemblance, l'estimateur de la moyenne d'une population normalement distribuée est la moyenne de l'échantillon.

On démontre que les coefficients des MCO correspondent au maximum de vraisemblance sauf les conditions 1) homoscédasticité, 2) normalité des résidus, 3) l'indépendance des erreurs. Notons que l'exogénéité et la normalité des variables explicatives ne sont pas des conditions requises.

La valeur maximale que peut prendre la fonction de vraisemblance est 1 (correspondant donc à un log = 0). La valeur minimale est – l'infini (log = – infini).

Notons que beaucoup de critères sont basés sur le maximum de vraisemblance (AIC, BIC, LR, Mc Fadden,…)

Multicolinéarité

Définition :

Il y a multicolinéarité s'il y a corrélation entre deux variables indépendantes. La multicolinéarité **ne va pas** à l'encontre des hypothèses de Gauss-Markov, mais pose néanmoins problème.

Dans une régression, la *multicolinéarité* est un problème qui survient lorsque certaines variables de prévision du modèle mesurent le même phénomène. Une multicolinéarité prononcée s'avère problématique, car elle peut **augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter**.

On dit que des variables sont multicolinéaires s'il existe une relation linéaire entre elles. C'est une extension du cas simple de la colinéarité entre deux variables. Par exemple, pour trois variables X_1, X_2, X_3, on dira qu'elles sont multicolinéaires si on peut écrire :

$$X_1 = aX_2 + bX_3$$

où a et b sont deux nombres réels.

Conséquences :

La multicolinéarité est un problème important puisqu'elle est à l'origine de la non convergence des estimateurs et donc de leur faible précision.

Les conséquences de **coefficients instables** peuvent être les suivantes :

- Les coefficients peuvent sembler non significatifs, même lorsqu'une relation significative existe entre le prédicteur et la réponse ;
- Les coefficients de prédicteurs fortement corrélés varieront considérablement d'un échantillon à un autre ; lorsque des termes d'un modèle sont fortement corrélés, la suppression de l'un de ces termes aura une incidence considérable sur les coefficients estimés des autres.
- Les coefficients des termes fortement corrélés peuvent même présenter le mauvais signe.

La multicolinéarité n'a aucune incidence sur l'adéquation de l'ajustement, (càd ne biaise pas l'estimation des coefficients), ni sur la qualité de la prévision. (Les écart-types des estimateurs et les t-tests restent valides). Cependant, les coefficients individuels associés à chaque variable explicative ne peuvent pas être interprétés de façon fiable.

Solutions :

Pour un modèle de régression tel que $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$, les solutions sont basées sur la formule qui donne la variance de population du coefficient b_2 :

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{n.Var(X_2)} \times \frac{1}{1 - r_{X_2,X_3}^2}$$

où r_{X_2,X_3}^2 est le coefficient de corrélation entre les variables explicatives.

- **Réduire** σ_u^2 en incluant au modèle des variables supplémentaires pertinentes.
- **Augmenter le nombre d'observations**.

Enquêtes : augmenter le budget, utiliser des clusters (Le partitionnement de données - data clustering - vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes)
Time séries : utiliser des données trimestrielles au lieu d'annuelles

- **Augmenter la variance des variables explicatives**. En pratique, cette option ne peut – et doit – être envisagée qu'au moment de la conception des enquêtes.
- **Combiner** les variables corrélées.
- **Enlever** certaines variables corrélées.
- Utiliser une **restriction** linéaire appropriée

Utilisation des restrictions linéaires

Si on soupçonne qu'une variable doit être reliée à Y mais que la t-value associée indique que la variable est non-significative, le problème peut être dû à une multicolinéarité avec une ou plusieurs autres variables. On peut alors imposer une restriction entre les variables, par exemple que $\beta_1 = \beta_2$ ce qui revient à définir un nouveau modèle de régression :

Modèle non restreint :	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_1$	
Modèle restreint :	$y = \beta_0 + \beta_1 (x_1 + x_2) + \beta_3 x_3 + \varepsilon_2$	si $\beta_1 = \beta_2$
	$y = \beta_0 + \beta_1 x_4 + \beta_3 x_3 + \varepsilon_2$	avec $x_4 = x_1 + x_2$

Il faut tester si la restriction (ou contrainte) est valide

$$F_{obs} = \frac{\frac{RSS_0 - RSS_1}{q}}{\frac{RSS_1}{n-k-1}} \geq 0$$

$RSS_1 = (\varepsilon' \varepsilon)_1 = RSS$ du modèle 1 (non restreint, valide sous H_1)

$RSS_0 = (\varepsilon' \varepsilon)_0 = RSS$ du modèle 0 (restreint, valide sous H_0)

$n =$ nombre d'observations dans le modèle 1

$k =$ nombre de variables explicatives dans le modèle 1

$q =$ nombres de restrictions/contraintes linéaires/

Test

Si $F_{obs} < F_{q,n+k-1} \Rightarrow AHO$

Si $F_{obs} > F_{q,n+k-1} \Rightarrow RHO$

Si la restriction est non valide, son utilisation mènera à des coefficients biaisés et des écart-types et tests invalides.

Notes : si $F_{obs} < 1$, la restriction est toujours valide.

On montre aussi que $t^2_{ddl} = F_{1,ddl}$

En R, il suffit de faire le test

anova(modèle non restreint, modèle restreint)

Pourquoi détecter la multicolinéarité ?

Il peut être utile de détecter des multicolinéarités au sein d'un groupe de variables notamment dans les cas suivants :

1. Pour identifier des structures dans les données et en tirer des décisions opérationnelles (par exemple, arrêter de mesurer une variable sur une chaîne de fabrication car elle est fortement liée à d'autres qui sont aussi mesurées) ;
2. Pour éviter des problèmes numériques lors de certains calculs. Certaines méthodes utilisent des inversions de matrices. L'inverse d'une matrice (p x p) ne peut être calculé que si elle est de rang p (ou régulière). Si elle est de rang inférieur, autrement dit s'il existe des relations linéaires entre ses colonnes, alors elle est singulière et non inversible.
3. Lorsqu'une régression linéaire multiple est exécutée sur des variables explicatives multicolinéaires, les estimations des coefficients peuvent être incorrectes. La fonctionnalité XLSTAT-régression linéaire permet de calculer automatiquement les statistiques de multicolinéarité sur les variables explicatives. Ainsi, l'utilisateur peut choisir d'éliminer les variables explicatives trop redondantes avec les autres. Notons que la régression PLS (régression par les moindres carrés partiels) n'est pas sensible à la multicolinéarité. Cependant, cette méthode est surtout utilisée à des fins de prédiction, et non pas pour des estimations de coefficients.

Comment détecter la multicolinéarité ?

Pour détecter les multicolinéarités et identifier les variables impliquées dans des multicolinéarités, on effectue des régressions linéaires de chacune des variables en fonction des autres. On calcule ensuite :

Le R^2 de chacun des modèles

- Si le R^2 vaut 1, alors il existe une relation linéaire entre la variable dépendante du modèle (le Y) et les variables explicatives (les X).

La tolérance pour chacun des modèles

- La tolérance vaut $(1–R^2)$. Elle est utilisée dans plusieurs méthodes (régression linéaire, régression logistique, analyse factorielle discriminante) comme un critère de filtrage des variables. Si une variable a une tolérance inférieure à un seuil fixé (la tolérance est calculée en prenant en compte les variables déjà utilisées dans le modèle), on ne la laisse pas entrer dans le modèle car sa contribution est négligeable et elle risquerait d'entraîner les problèmes numériques.

VIF (Variance Inflation Factor)

- Le VIF ou Variance Inflation Factor (Facteur d'inflation de la variance) est égal à l'inverse de la tolérance. Le VIF calcule le changement de la variance de chacun des facteurs lorsque on les introduit dans la régression.

$$VIF = \frac{1}{1 - R_j^2}$$

Le R_j^2 est le coefficient de détermination de la régression de la variable dépendante X_j sur toutes les autres variables sans inclure la variable dépendante. *Y*.

- Si un VIF est supérieur à 5 (certains auteurs proposent une valeur de 10) ou si la moyenne est beaucoup plus grande que 1, vous avez une multi-colinéarité élevée : la variation semblera plus grande et le facteur apparaîtra plus influent qu'il ne l'est. Le VIF donne un indice qui mesure dans quelle proportion la variance d'un coefficient est augmenté à cause de la colinéarité. Si VIF est plus proche de 1, alors le modèle est beaucoup plus robuste, car les facteurs ne sont pas influencés par la corrélation avec d'autres facteurs.
    - VIF = 1 Non corrélés
    - 1 < VIF < 5 Modérément corrélés
    - VIF > 5 Hautement corrélés
- La racine carrée du VIF de chaque coefficient nous donne le rapport entre la variance actuelle du coefficient et la variance si la variable explicative n'était pas corrélée aux autres variables du modèle.

Exemple

Commande **:** *estat vif*

```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
     hauteur |      4.70    0.212936
     largeur |      4.69    0.213289
         age |      1.06    0.940474
   lncatalog |      1.05    0.955959
       prive |      1.00    0.995945
-------------+----------------------
    Mean VIF |      2.50
```

Les variables dummy (ou dites indicatrices)

Une variable dummy (ou indicatrice) est une variable explicative particulière qui n'est composée que de 0 ou de 1. Cette variable est utilisée lorsque, dans un modèle, nous désirons intégrer un facteur explicatif binaire : « le phénomène a lieu ou n'a pas lieu » pour corriger, par exemple, d'une valeur anormale ; ou bien lorsque le facteur explicatif est qualitatif : « le genre d'un individu, homme ou femme ». Il s'agit donc d'incorporer une ou des variables explicatives supplémentaires au modèle spécifié initialement et d'appliquer les méthodes classiques d'estimation.
Le modèle de régression diffère selon l'apparition du phénomène par les valeurs d'un ou plusieurs coefficients alors que les autres paramètres sont identiques. En cas de modification structurelle d'un coefficient de régression, la variable muette affecte alors le coefficient de la ou des variables explicatives considérées. Par exemple, soit le modèle à deux variables explicatives x_{1t} et x_{2t} :

$$y_t = a_0 + a_1x_{1t} + a_2x_{2t} + b_0D_t + b_1D_t x_{1t} + b_2D_t x_{2t} + \varepsilon_t$$

- Si le phénomène existe, $D_t = 1$ et $D_t = 0$ sinon.
- Si $D_t = 0$, le modèle s'écrit : $y_t = a_0 + a_1x_{1t} + a_2x_{2t} + \varepsilon_t$
- Si $D_t = 1$, le modèle s'écrit : $y_t = (a_0 + b_0) + (a_1 + b_1)x_{1t} + (a_2 + b_2)x_{2t} + \varepsilon_t$
- Si $b_1 = b_2 = 0$, le modèle ne diffère que par la valeur du terme constant.

Domaine d'utilisation des variables indicatrices.

Le domaine d'utilisation des variables indicatrices est très vaste, nous pouvons citer : la correction des valeurs anormales, la modification structurelle (0 pour la période avant le changement structurel, 1 après le changement structurel), l'intégration de la saisonnalité, la caractérisation d'un individu (genre, situation matrimoniale…), l'intégration de facteurs qualitatifs (appartenance d'un pays à la zone euro, promotion non quantifiable…), etc.

Exemple : Variable qualitative
Il peut s'avérer important dans certaines spécifications de modèle de tenir compte de l'effet, sur la variable endogène, de variables qualitatives : être titulaire d'un diplôme, genre d'un individu, appartenance politique, etc. L'utilisation d'une variable indicatrice permet de segmenter les individus en deux groupes et de déterminer si le critère de segmentation est réellement discriminant. Attention, dans le cas de variables qualitatives à plusieurs modalités, par exemple la couleur des yeux (bleu, vert, marron, autres), ou bien la situation familiale (célibataire, marié, divorcé, veuf, autres), etc. Il convient alors de coder autant de variables indicatrices que de modalités moins une. En reprenant l'exemple de la couleur des yeux nous définissons trois variables indicatrices : bleu (= 1 si l'individu à les yeux bleus, 0 sinon), vert (= 1 si l'individu à les yeux verts, 0 sinon), marron (= 1 si l'individu à les yeux marrons, 0 sinon), le cas des autres individus n'appartenant pas à l'une des trois premières catégories est implicitement contenu dans le terme constant. Une erreur à ne pas commettre consiste à créer une seule variable explicative codée, par exemple, de la manière suivante : bleu = 1, vert = 2, marron = 3, …

La transformation BOX-COX

R TP6 et TP7

La transformation, BOX-COX permet de choisir **le meilleur modèle.**

De nombreuses procédures statistiques reposent sur la normalité des distributions. On sait généralement que des distributions très dissymétriques faussent les calculs. Transformer les variables de manière à se rapprocher de la distribution normale, ou tout du moins pour les symétriser, est parfois un préalable nécessaire avant toute analyse statistique. La transformation la plus communément répandue est la transformée de Box-Cox.
La transformée de Box-Cox est la transformation non linéaire de loin la plus rencontrée en statistique et en économétrie. Elle est définie comme :

$$y^{(\lambda)} = \frac{y^{\lambda} - 1}{\lambda} \quad \text{avec } y > 0$$

En particulier, si λ =–1,0 ou 1, on a

$$y^{(\lambda)} = \begin{cases} y - 1 & \text{si } \lambda = 1 \\ \ln(y) & \text{si } \lambda = 0 \\ 1 - 1/y & \text{si } \lambda = -1 \end{cases}$$

Dans le modèle dit « lambda », on applique sur la variable dépendante (y) et les variables indépendantes (x_i), la même transformation. Dans le modèle dit « thêta », on applique sur y et les x_i des transformations de paramètres θ et λ différents.

Les valeurs des paramètres θ et λ peuvent être quelconque et STATA recherche automatiquement les valeurs qui correspondent au maximum de vraisemblance. En plus STATA teste les valeurs –1, 0 et 1. Par exemple, dans l'exemple ci-dessous, on pourrait admettre les valeurs de 0, qui donnent une simple transformation logarithmique.

Pour R, voir TP6

```
/* HELP BOXCOX

        boxcox depvar [indepvars] [if] [in] [weight] [, options]

    options               description
    -------------------------------------------------------------------------
    Model
      noconstant          suppress constant term
      model(lhsonly)      left-hand-side Box-Cox model; the default
      model(rhsonly)      right-hand-side Box-Cox model
      model(lambda)       both sides Box-Cox model with same parameter
      model(theta)        both sides Box-Cox model with different parameters
      notrans(varlist)    nontransformed independent variables
*/
```

```
 */
.
boxcox prix catalog,model(theta)

                                                Number of obs   =       1529
                                                LR chi2(2)      =    2997.03
Log likelihood = -14288.974                     Prob > chi2     =      0.000

------------------------------------------------------------------------------
        prix |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     /lambda |   .1573892   .0188954     8.33   0.000     .1203549    .1944234
      /theta |   .1597368   .0138304    11.55   0.000     .1326296    .1868439
------------------------------------------------------------------------------


-------------------------------------------------------------
   Test              Restricted
    H0:            log likelihood       chi2       Prob > chi2
-------------------------------------------------------------
theta=lambda = -1     -17815.788      7053.63          0.000
theta=lambda =  0     -14355.316       132.68          0.000
theta=lambda =  1     -16092.839      3607.73          0.000
-------------------------------------------------------------
```

Parfois, il n'est pas possible de conclure en regardant les p-valeurs. On choisira alors la transformation dont le maximum de vraisemblance est le plus grand (log-likehood le plus proche de 0) ou dont la statistique de chi 2 est la plus faible. Concernant le chi 2, intuitivement, il y a relativement moins de chance au niveau de la transformation ayant le chi 2 le plus faible, d'atteindre et de dépasser la statistique théorique du test, et donc de rejeter cette transformation au profit des autres.

FITSTAT (commande Stata)

https://ideas.repec.org/c/boc/bocode/s407201.html

fitstat is a post-estimation command that computes a variety of measures of fit for many kinds of regression models. For all models, fitstat reports

- the log-likelihoods of the full and intercept-only models,
- the deviance (D). La déviance joue le même rôle que la somme des carrés résiduels dans la régression linéaire. Des valeurs élevées suggèrent que la ou les variables indépendantes ne sont pas utiles pour prédire la réponse.
- the likelihood ratio chi-square (G2),
- Akaike's Information Criterion (AIC),
- AIC*N,
- the Bayesian Information Criterion (BIC), and BIC'.

For all models except regress, fitstat reports

- McFadden's R^2,
- McFadden's adjusted R^2,
- the maximum likelihood R^2,
- Cragg & Uhler's R^2.

These measures all equal R^2 for OLS regression. fitstat reports R^2 and the adjusted R^2 after regress. fitstat reports the regular and adjusted count R^2 for categorical data models.
For ordered or binary logit or probit models, as well as models for censored data (tobit, cnreg, or intreg), it also reports McKelvey and Zavonia's R^2. In addition, it reports Efron's R^2 for logit or probit.

```
. fitstat

Measures of Fit for regress of prix

Log-Lik Intercept Only:   -17361.697     Log-Lik Full Model:       -16786.027
D(1527):                   33572.055     LR(1):                     1151.339
                                         Prob > LR:                    0.000
R2:                            0.529     Adjusted R2:                  0.529
AIC:                          21.959     AIC*n:                    33576.055
BIC:                       22375.527     BIC':                     -1144.006


Measures of Fit for probit of works

Log-Lik Intercept Only:   -54442.414     Log-Lik Full Model:       -47884.610
D(79943):                  95769.219     LR(14):                    13115.608
                                         Prob > LR:                    0.000
McFadden's R2:                 0.120     McFadden's Adj R2:            0.120
Maximum Likelihood R2:         0.151     Cragg & Uhler's R2:           0.203
McKelvey and Zavoina's R2:     0.234     Efron's R2:                   0.156
Variance of y*:                1.305     Variance of error:            1.000
Count R2:                      0.682     Adj Count R2:                 0.246
AIC:                           1.198     AIC*n:                    95809.219
BIC:                    -806732.834      BIC':                     -12957.558
```

Les hypothèses de Gauss-Markov

HYPOTHESES DE GAUSS-MARKOV

Hypothèses	
1) Bonne spécification	$\mathbf{E}(\varepsilon_i) = 0$
2) Exogénéité	$\mathbf{Cor}(\varepsilon_i, X_i) = 0$
3) Homoscedasticité	$\mathbf{Var}(\varepsilon_i) = \sigma^2$, **constant**
4) Indépéndence sérielle	$\mathbf{Cor}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$
5) Normalité	$\varepsilon_i \sim$ **Normal**
6) Pas de parameters incidentaux (K ne grandit pas avec N)	

(1), (2) et (6) sont nécessaires pour la CONVERGENCE, (3) and (4) sont nécessaires pour l'EFFICACITE et (5) est nécessaire pour les propriétés de petit échantillon

Si les conditions du théorème de Gauss-Markov sont respectées, les estimateurs des MCO sont des estimateurs dit **"BLUE" (*Best Linear Unbiased Estimator*),** c'est-à-dire qu'il est **sans biais** et à **variance minimale**. Ce sont donc des estimateurs **efficaces**.

Hypothèses

- H1 : le modèle est linéaire en x_t (ou en n'importe quelle transformation de x_t).
- H2 : les valeurs x_t sont observées sans erreur (x_t non aléatoire).
- H3 : $E(\varepsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle : en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle.
- H4 : $E(\varepsilon_t^2) = \sigma_\varepsilon^2$, la variance de l'erreur est constante[1] : le risque de l'amplitude de l'erreur est le même quelle que soit la période.
- H5 : $E(\varepsilon_t \varepsilon_{t'}) = 0$ si $t \neq t'$, les erreurs sont non corrélées (ou encore indépendantes) : une erreur à l'instant t n'a pas d'influence sur les erreurs suivantes.
- H6 : $\mathrm{Cov}(x_t, \varepsilon_t) = 0$, l'erreur est indépendante de la variable explicative.

Si on ajoute la condition de la normalité des erreurs, alors les estimateurs donnés par le **maximum de vraisemblance** et les **MCO** sont identiques.

Bonne spécification :

Définition

$E(\varepsilon_i) = 0$. La moyenne des résidus doit être égale à zéro.

Conséquences de la mauvaise spécification			
		True Model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted Model	$\hat{Y} = b_1 + b_2 X_2$	Spécification correcte, Pas de problème	Coefficients biaisés (en général). Ecart-types invalides.
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$	Coefficients non biasés (en général), mais inefficaces Ecart-types valides (en général)	Spécification correcte, Pas de problème

Raisons d'une mauvaise spécification :

- **Omission d'une variable pertinente.** ➔ Coefficients biaisés (en général). Écart-types invalides.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$Y = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\text{cov}(X_2, X_3)}{\text{var}(X_2)}$$

  - Lorsque la variable omise est corrélée avec une variable du modèle, alors elle va se retrouver dans le terme d'erreur, ce qui est un problème.
  - Lorsque la variable omise n'est pas corrélée avec une variable du modèle, les choses sont moins graves.
  - Si B est corrélé avec A *et* corrélé avec Y, cela entraînera un biais dans l'estimation du coefficient de A. Le diagramme suivant montre comment l'estimation du coefficient de A sera biaisée, en fonction de la nature de la relation avec B :

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with Y	Positive Bias	Negative Bias
B is negatively correlated with Y	Negative Bias	Positive Bias

- **Inclusion d'une variable non pertinente.** ➔ Coefficients non biaisés (en général), mais inefficaces. Écart-types valides (en général).

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$Y = b_1 + b_2 X_2 + b_3 X_3$$

$$\text{var}(b_2) = \sigma_{b_2}^2 = \frac{\sigma_u^2}{n\,\text{var}(X_2)} \times \frac{1}{1 - r_{X_2,X_3}^2}$$

- Plus la corrélation entre X_2 et X_3 est forte et plus var(b_2) est élevé
- L'introduction de variables non pertinentes conduit généralement à de la multicolinéarité.

Solution

- **Pour une variable omise** : Utilisation de variable **PROXY**
  - Pour des applications statistiques, un **proxy** ou une **variable proxy** (ou **variable de substitution** ou **variable substitutive**) est une variable qui n'est pas significative en soi, mais qui remplace une variable utile mais non observable ou non mesurable. Pour qu'une variable soit un bon proxy, elle doit avoir une **bonne corrélation**, pas nécessairement linéaire, avec la variable utile. Cette corrélation peut être positive ou négative.
    - Exemples :
      - Le PIB par habitant est souvent utilisé en tant que proxy pour des mesures du niveau de la vie ou de la qualité de la vie. On peut aussi utiliser la consommation par adulte.
      - Le nombre d'années d'études et le GPA (moyenne des notes d'un étudiant) constituent des proxies pour l'aptitude cognitive.
      - L'indice de masse corporelle (IMC) constitue un proxy pour le taux de graisse corporelle.
      - Les variations de taille sur une période donnée constituent un proxy pour le niveau des hormones dans le sang.
      - Les images satellites de la couleur des océans constituent un proxy pour la profondeur à laquelle pénètre la lumière.
      - La largeur des anneaux de croissance des arbres constitue un proxy pour l'historique des conditions environnementales.
      - La composition isotopique de l'oxygène dans les carottes de glace polaire en fonction de la profondeur constitue un proxy de la paléo-température moyenne en fonction du temps.
  - Soit la régression suivante :

    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u.$$

    Si nous ne pouvons avoir accès à X_2, une régression sans cette variable explicative mènerait à des estimations des coefficients biaisés et des écarts-types et tests invalides.

    Soit alors Z, une variable proxy de X_2 : $Z = \lambda + \mu X_2$, on obtient

    $$Y = \beta_0 + \beta_1 X_1 + \beta_2 (\lambda + \mu Z) + u \Rightarrow Y = \beta_0 + \beta_2 \lambda + \beta_1 X_1 + \beta_2 \mu Z$$

    Il ne sera donc pas possible d'estimer β_2. Cependant, les estimations, écarts-types et tests sur les autres variables et il sera aussi possible de dire si X_2 a un effet significatif.
- **Pour une variable excédentaire**, il suffit en général de la retirer (voir le paragraphe sur la détermination des variables importantes.)

Exogénéité :

Définition :

$Cor(\varepsilon_i, X_i) = 0$. La corrélation entre les variables explicatives X_i et le terme d'erreur doit être nulle. Autrement dit, les résidus sont indépendants des variables. Si cette condition n'est pas respectée alors les estimateurs seront **biaisés et non-convergeant**. On dit que l'on a un **biais d'endogénéité**.

Causes d'endogénéité :

- **Causalité inverse**. (Simultanéité). On cherche à estimer Y à partir de X_i alors que X_i est lui-même fonction de Y.
- **Variable Omise**.
  - Soit le vrai modèle

    $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

    Si on omet la variable explicative X_2, on estime ce modèle :

    $$Y = b_0 + b_1 X_1 + v$$

    Nous aurons alors un estimateur qui prend le vrai paramètre avec un biais.

    $$E(b_1) = \beta_1 + \boxed{\beta_2 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)}} \qquad \text{Situation avec biais}$$

    $$E(b_1) = \beta_1 \qquad \text{Situation sans biais.}$$

  - L'estimateur biaisé (b_1) surestimera le vrai coefficient si :
    - β_2 et la covariance sont deux positifs ou négatifs
  - L'estimateur biaisé (b_1) estimera correctement le vrai coefficient si :
    - β_2 est nul (donc X_2 n'explique par y)
    - la covariance est nulle
  - L'estimateur biaisé (b_1) sous-estimera le vrai coefficient si :
    - β_2 et la covariance sont de signes contraires.

- **Erreur de mesure**. (Les erreurs de mesure dans la variable dépendante Y sont moins graves que la variable explicatives X_i)

Solution :

- Méthode des **variables instrumentales** (VI).
  - Un **bon instrument** Z d'un variable X doit répondre à trois critères.
    - $\text{cov}(X, Z) \neq 0$: relation forte entre la variable instrumentale et la variable instrumentée. (C'est donc une variable proxy – voir remarque ci-après). Cette condition s'appelle parfois la **condition de rang**.
    - $\text{cov}(Z, u) = 0$: l'instrument ne peut être corrélé avec le terme d'erreur. C'est ce qu'on appelle la **condition d'orthogonalité**.
    - $Z \rightarrow X_i \rightarrow Y$: le seul lien qui peut exister entre Z et Y, c'est la variable X_i, variable endogène dans notre modèle et pas une variable exogène W. Autrement dit ***Z* ne peut pas** être une variable explicative en elle-même.
  - On peut parfois utiliser **plusieurs instruments** pour une même variable. Ceci permet d'avoir une estimation de X_i plus précise, et une deuxième étape plus facile à expliquer. (Les VI doivent évidemment être validés par une première étape).

- La méthode des VI permet d'obtenir des estimateurs **consistants** (= convergents) mais ils seront **biaisés.** Cependant, il est possible de **corriger le biais.** Par exemple pour le modèle : $Y = \beta_1 + \beta_2 X + u$, si Z est l'instrument de X alors $b_2^{IV} = \beta_2 + \frac{\text{cov}(Z,u)}{\text{cov}(Z,X)}$

- Remarques :
  - 1) Selon Wooldridge (le pape de l'économétrie), il n'est pas nécessaire que l'instrument soit corrélé avec la variable instrumentée, même si cela est préférable. L'important est que l'instrument ne soit pas corrélé avec le terme d'erreur. Par conséquent, il considère qu'une variable de substitution (= variable proxy), qui elle doit être fortement corrélée avec la variable omise pour pouvoir la simulée, n'est pas recommandée comme variable instrument. En effet, un instrument qui serait trop fortement corrélé avec la variable instrumentée, serait automatiquement corrélée au terme d'erreur puisque la variable instrumentée est corrélée avec le terme d'erreur.
  - 2) Si l'utilisation d'un instrument permet d'obtenir de meilleurs estimateurs, il faut néanmoins reste attentif, que les écart-types seront augmentés.

Tests :

Test de Fischer – Pour vérifier que l'instrument est bien corrélé avec la variable instrumentée

La méthode s'effectue en deux étapes (on parle parfois de Doubles Moindres Carrés – DMC - ou encore de 2SLS – 2 Stage Least Square).

- On régresse la variable X_i instrumentée en incluant Z et TOUTES les autres variables explicatives du modèle. Ce qui nous donne X_i (X_i prédit). Cette étape permet de vérifier que l'instrument Z est bon. (Voir aussi test de SARGAN)
  - L'instrument Z doit être significatif.
  - La F statistique doit être élevée.
- On régresse Y en mettant X_i à la place de X_i.

Test de SARGAN (Test de suridentification) – *pour vérifier que les instruments sont exogènes.*

Commande : *overid*
TP : 8

Le **test de Sargan** ou **test de Sargan-Hansen** est un test statistique permettant de tester une hypothèse de suridentification dans un modèle statistique. Il est aussi connu sous le nom de **test de Hansen** ou **test J**.

En pratique, on est souvent amené à effectuer des estimations d'une même équation en étendant ou restreignant la liste des variables instrumentales. En effet, on pouvait avoir intérêt à accroître le nombre de variables instrumentales dans la mesure où cela conduit à des estimateurs plus précis. Cependant, accroître indûment l'ensemble des variables instrumentales peut conduire à faire apparaître des biais dans l'estimation. Lorsqu'on dispose de plus de variables exogènes que nécessaires, on dit que le modèle est **suridentifié**. Le test de suridentification de Sargan est un test très important et très couramment utilisé permettant de contrôler qu'il n'y a pas d'incohérence dans le choix des variables instrumentales. Ce test, appelé test de suridentification, ou test de Sargan constitue un guide incontournable dans le choix des variables instrumentales.

- Permet de tester la validité d'un instrument et de vérifier son exogénéité. Ce test est valable sous **deux conditions**
  - au moins un des instruments est valable, c'est-à-dire qu'il est exogène.
  - il faut plus d'instruments que de variables instrumentées.
- Le test de Sargan est construit sur l'hypothèse que le terme d'erreur ne doit pas être corrélé avec l'ensemble des variables exogènes si les instruments sont valides. Si l'estimateur du terme d'erreur, n'est pas expliqué par les instruments, alors les instruments sont exogènes et donc valides.
- En pratique on utilise la méthode des double moindres carrés (DMC). On récupère le résidu de la régression et on régresse le carré des résidus sur les instruments. Si l'une des variables est corrélée au résidu, alors elle est significative ce qui veut dire qu'elle n'est pas exogène et donc contraire aux hypothèses.
- Un résumé du test de Sargan pour l'ensemble des variables exogènes est aussi obtenu par la statistique suivante à partir de la régression 2 :

$$S = N.R^2$$

que l'on compare à une loi de Chi2 au degré de liberté suivant (Nb instruments- Nb variables endogènes). Si c'est significatif, les instruments ne sont pas (tous) exogènes
- Si la p-value est >0.05, alors le modèle est correct. Si la p-value < 0.05, alors on a trop d'instruments.

Exemple

Commande

```
. qui ivregress 2sls mort5 (nb_ygsibs5 = twin5 time_to_fc) female urban i.yob RBR
birthsp i.yobm twin i.educlvl, first
. predict res2, res
. reg res2 twin5 time_to_fc female urban i.yob RBR birth_sp i.yobm twin i.educlvl
. di invchi2tail(1,0,05)
```

Le **test de Sargan** permet de tester l'exogénéité des résidus sous condition qu'au moins un des instruments soit valide. Ce test est uniquement valable pour 2 instruments. En effet, il faut plus d'instruments que de variables instrumentées.

$$\boxed{S = n \times R^2}$$

$$e^2 = \eta_0 + \eta_1 Z_1 + \eta_2 Z_2 + \eta_3 X_2 + \eta_4 X_3 + \epsilon$$

Stage 2

$$Y = \gamma_0 + \gamma_1 \widehat{X_1} + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon$$
$$\hat{Y} = g_0 + g_1 \widehat{X_1} + g_2 X_2 + g_3 X_3 + \epsilon$$
$$e = Y - \hat{Y}$$

Hypothèses du **test de Sargan** :

H_0: tous les instruments sont valides

H_A: au moins un des instruments n'est pas exogène (valide)

Si l'estimateur du terme d'erreur u n'est pas expliqué par Z_1 et Z_2, cela signifie que u et Z_i sont indépendants.

```
. reg res2 twin5 time_to_fc female urban i.yob RBR birth_sp i.yobm twin i.educlvl

      Source |       SS       df       MS              Number of obs =    4276
-------------+------------------------------           F( 20,  4255) =    0.02
       Model |   .03585195    20  .001792598            Prob > F      =  1.0000
    Residual |  370.671209  4255  .087114268            R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0046
       Total |   370.70706  4275  .086715102            Root MSE      =  .29515
```

Le R^2 est de 0.0001.

Dès lors, la formule de Sargan étant donnée par $\boxed{S = nR^2} \sim \chi^2$,

$$\leftrightarrow S = 4276 \times 0.0001 = 0.4276$$

On compare cette statistique de Sargan à la valeur critique d'une χ^2 de deux instruments et 1 variable instrumentée, et un taux de significativité de 5% :

```
. di invchi2tail(1,0.05) = 3.8414588
```

$$0.4276 < 3.8414$$

La statistique de Sargan étant inférieure à la valeur critique d'une Chi2 au seuil de 5%, on ne peut pas rejeter l'hypothèse nulle « tous les instruments sont valides ». Dès lors, la seconde hypothèse, à savoir, $Cov\,(Z, u) = 0$ est respectée.

```
. overid

Tests of overidentifying restrictions:
Sargan N*R-sq test        0.001  Chi-sq(1)     P-value = 0.9747
Basmann test              0.001  Chi-sq(1)     P-value = 0.9749
```

Le **test de Basmann** est équivalent au test de Sargan
Le **test de Hansen** est une version robuste du test de Sargan

Test de DURBIN-WU-HAUSMAN dit test de HAUSMAN pour vérifier que la variable suspecte est bien endogène.

Commande *endogenous*
TP 8

Le **test d'Hausman**, également connu comme le **test de Wu-Hausman** ou encore le **test de Durbin-Wu-Hausman** est un test statistique utilisé en économétrie pour comparer un estimateur convergent sous l'hypothèse nulle et l'hypothèse alternative, et un estimateur convergent et efficace (= de plus petit biais possible) sous l'hypothèse nulle mais non convergent sous l'hypothèse alternative.

Le test d'Hausman permet de tester l'hypothèse que l'estimateur des doubles moindres carrés est significativement différent de l'estimateur des moindres carrés ordinaires.

Le test est le suivant, on compare la différence entre les MCO[1] et la régression instrumentée. Si le paramètre estimé n'est pas différent alors il n'y a pas d'endogénéité.

- Le test permet de faire un choix entre plusieurs modèles.
  - H_0 pas de violation de la 4ème condition de Gauss-Markoff (pas de problème d'endogénéité). Donc les modèles MCO et IV[2] sont équivalents mais IV n'est jamais aussi précis que MCO (les coefficients sont légèrement biaisés mais ils sont efficients)
  - H_1 : problème d'endogénéité, Les coefficients estimés des MCO et IV sont significativement différents. MCO est inconsistant (c'est-à-dire pas convergent). On doit donc préférer le modèle IV qui est sans biais, mais les écarts-types sont gonflés.
- Sous l'hypothèse nulle, le test statistique a une distribution chi-carré avec un nombre de degrés de liberté égal au nombre de variables instrumentées.
- Note : Cependant, si le test statistique n'est pas significatif, cela ne veut pas nécessairement dire que l'hypothèse nulle est vraie. Cela pourrait être qu'elle est fausse mais que les instruments utilisés en IV sont si faibles que les différences entre les estimations IV et MCO ne sont pas significatives.

Exemple

Pour savoir quel modèle parmi ceux utilisés aux différentes questions ci-dessus : le modèle de probabilité linéaire, 2SLS avec un instrument ou 2SLS avec deux instruments, il convient d'utiliser le **test d'Hausman**. Ce dernier permet de décider lequel nous semble le plus adéquat.
⇨ Il faut donc **faire un choix** entre
- Les variables instrumentées : non-biais mais les écarts-types sont gonflés
- OLS : les coefficients sont légèrement biaisés mais ils sont efficients.

[1] MCO = Moindres Carrés Ordinaires (OLS = Ordinary Least Square)
[2] IV = Instrumented Variable = Variable instrumentée

Statistique de Hausman :

$$H = \frac{\left(\sum_{i=1}^{3}(g_i - a_i)\right)^2}{\sum_{i=1}^{3} Var\,(g_i - a_i)} \sim \chi^2(k)$$

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + u$$

$$OLS \quad \hat{Y} = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3$$

Hypothèses :

H_0 : OLS & IV (variable instrumentale) sont consistants
MAIS IV n'est jamais efficient → on doit préférer l'OLS à l'IV

H_a : OLS n'est pas consistant → IV

```
             ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       IV          OLS         Difference          S.E.
-------------+------------------------------------------------------------
  nb_ygsibs5 |     .431685     .0842671        .3474178         .095626
      female |   -.0280312    -.0224174       -.0056138        .0028189
       urban |    .0272445     .0293605        -.002116        .0043225
    1bn.yob |   -.1338397    -.1239154       -.0099243         .015053
      2.yob |   -.1272848    -.1284509        .0011661        .0146923
      3.yob |   -.1581412    -.1610545        .0029133        .0149981
      4.yob |   -.1549965    -.1666571        .0116606        .0156657
      5.yob |   -.1775763    -.1993391        .0217628        .0169077
        RBR |    .0281432    -.0139104        .0420537        .0126121
                    b = consistent under Ho and Ha; obtained from ivregress
         B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                 chi2(19) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =          13.20
               Prob>chi2 =         0.8282
```

Cette statistique de Hausman est de 13.20. On l'interprète en regardant la p-valeur associée. Celle-ci s'élève à 0.8282. Nous avons donc 82.82% de chance de nous tromper en rejetant l'hypothèse nulle. Par conséquent, nous ne rejetons pas l'hypothèse nulle. MLP et 2 SLS avec un instrument sont tous les deux consistants. IV est cependant moins efficace, on doit donc choisir MLP.

Résumé de la méthode par variables instrumentales.

Une procédure en deux étapes

Vrai modèle : $y = a_{vrai} + b_{vrai} x_{endo} + c_{vrai} x_2 + u$ avec $cov(x_{endo}, u) \neq 0$

- **Première étape :** on régresse la variable endogène à la fois sur l'instrument et sur les autres variables explicatives. NB : On met toutes les variables explicatives même non pertinentes en première étape.

$\boldsymbol{x_{endo}} = a_0 + a_1 \cdot z_{inst} + a_2 \cdot x_2 + u_{prem}$

- On récupère de cette première régression x'_{endo}, la prédiction de la variable endogène x_{endo} :

$x'_{endo} = a_0 + a_1 \cdot z_{inst} + a_2 \cdot x_2 = x_{endo} - u_{prem}$

- **Deuxième étape :** on introduit cette prédiction dans la régression à la place de x_{endo}

$y = a_{est} + \boldsymbol{b_{est} x'_{endo}} + c_{est} \cdot x_2 + u_{deux}$

- comme z_{inst} et x_2 ne sont pas corrélés avec le résidu u, alors x'_{endo} n'ést plus corrélé avec u, l'estimateur des variables instrumentales permet d'estimer sans biais b_{vrai} (et aussi a_{vrai} et c_{vrai}).

$E(b_{est}) = b_{vrai}$; $E(a_{est}) = a_{vrai}$; $E(c_{est}) = c_{vrai}$

La stratégie argumentative

- 1. Le raisonnement
- 2. Montrer que l'instrument est aussi bon qu'une ventilation aléatoire
  - Non corrélation avec les autres variables explicatives
- 3. La variable suspecte est elle endogène ?
  - Test d'endogénéité dit de Wu-Hausman
- 4. Les instruments sont ils bien exogènes ?
  - Test d'exogénéité des instruments (ou de validité jointe des deux instruments) dit de Sargan
- 5. Les instruments sont ils suffisamment puissants pour corriger le biais ?
  - Faiblesse des instruments

Homoscédasticité :

Définition

$Var(\varepsilon_i) = \sigma^2$, constant. La variance des résidus est une constante. Autrement dit, la variance des résidus ne varie pas, elle n'est pas spécifique à chaque individu. Si la distribution du terme d'erreur n'est pas la même pour toutes les observations, on a de l'**hétéroscédasticité**.

Conséquences de l'hétéroscédasticité :

- Les MCO restent **sans biais et convergents**, même en présence d'hétéroscédasticité. La qualité de l'estimation mesurée par R^2 ou Adjusted-R^2 reste valide. MAIS, la matrice de variance covariance des coefficients estimés est biaisée en présence d'hétéroscédasticité, donc les **écarts-types sont invalides**. On ne peut plus appliquer les tests d'hypothèses usuels post-estimation (*t* statistics, *F* statistics ou *LM* statistics)

Causes de l'hétéroscédasticité :

- Variables explicatives inobservées de variance différentes pour certains groupes (définis par des variables observées)
- Modèles à coefficients aléatoires
- Observations représentant des moyennes sur des sous-groupes.
- Répétition d'une même valeur de la variable à expliquer pour des valeurs différentes d'une variable explicative (ex : regroupement en tranches pour le revenu, etc.)

Solutions :

- Utilisation de la méthode MCP (Moindres Carrés Pondérés)➔Transformation de variable.
  - Passer au log permet parfois de supprimer l'hétéroscédasticité.
  - On peut aussi diviser la variable qui pose problème. Par exemple, faire une régression sur des prix au m² plutôt que sur les prix proprement dit.
- Utilisation du modèle LOGIT ou PROBIT pour les variables binaires.
- Utilisation de la méthode MCG (moindres carrés généralisés)

Tests :

Test de BREUSCH PAGAN

R : Voir TP7

Le **test de Breusch-Pagan** permet de tester l'hypothèse d'homoscédasticité du terme d'erreur d'un modèle de régression linéaire. Il cherche à déterminer la nature de la variance du terme d'erreurs : si la variance est constante, alors on a de l'homoscédasticité ; en revanche, si elle varie, on a de l'hétéroscédasticité.
Par exemple, on estime le modèle suivant

$$Y = \beta_0 + \beta_1 X + u$$

et on obtient alors les valeurs résiduelles : $\hat{u}$.
Les moindres carrés ordinaires (MCO) sont un estimateur faisant en sorte que la moyenne des résidus soit nulle. Ainsi, en supposant que la valeur des résidus ne dépend pas des variables explicatives, on peut exprimer la variance des résidus comme la valeur au carré des résidus. Si cette hypothèse n'est pas tenable, alors on pourrait par exemple exprimer la variance comme une relation linéaire entre les résidus et les variables explicatives. Un modèle de ce genre peut être testé en régressant les carrés des résidus sur **les variables explicatives** en utilisant une équation auxiliaire de la forme

$$u^2 = \gamma_0 + \gamma_1 X + v$$

C'est un test basé sur un test du χ^2 : si la statistique du test de Breusch-Pagan est supérieure à celle obtenue par le test du Chi-Deux, c'est-à-dire si la p-value est inférieure à un certain seuil (souvent 5%), alors on rejette l'hypothèse nulle d'homoscédasticité avec un risque d'erreur de première espèce de 5% (si on a choisi ce seuil).

Une des corrections possibles peut alors être l'utilisation des moindres carrés pondérés (si l'on connaît l'origine de l'hétéroscédasticité).

Le test se déroule en trois étapes :

- On estime le modèle par les MCO, ce qui nous permet d'obtenir les résidus que l'on élève ensuite au carré.
- On régresse le carré des résidus sur l'ensemble des *X*. S'il y a hétéroscédaticité, les coefficients des *X* ne seront pas tous significatifs. On récupère le R^2 de cette régression, et soit *K* le nombre de paramètres à estimer.
- On calcule la statistique de Breusch-Pagan :

  $BP = nR^2$

BP	Statistiques de Breusch-Pagan.
n	Nombre d'observations
R^2	Coefficient de détermination

  qui suit une loi du chi² : $\chi^2(K-1)$.
- Si la statistique de Breusch-Pagan est supérieure à celle lue dans la table du Chi-Deux pour un certain niveau de risque d'erreur de première espèce (5% étant la valeur généralement retenue), alors on rejette l'hypothèse nulle d'homoscédasticité.

Test de GOLDFELD-QUANDT

Commande Stata *: estat hettest*
R : Voir TP7

Le test de **Goldfeld-Quandt** teste la présence d'hétéroscédasticité. Ce test n'est valable que si l'une des variables est la cause de l'hétéroscédasticité et que le nombre d'observations est important. Il suppose aussi une forme spéciale d'hétéroscédasticité (typiquement que la variance augmente avec les X). Il se compose de trois étapes.

- L'échantillon est divisé en trois parties contenant, primo, les 3/8 des observations avec les valeurs les plus petites de la variable X, secundo, les 3/8 des observations avec les valeurs les plus grandes et, tertio, le ¼ des observations présentes au milieu de la distribution des X.
- On fait une régression sur la partie inférieure et la partie supérieure et on compare ensuite la somme des carrés résiduels pour les deux régressions : RSS_1 et RSS_2.
- La statistique du test est distribuée selon une loi de Fischer

$$F(n_2, n_1) = \frac{RSS_2 / n_2}{RSS_1 / n_1}$$

RSS_1, RSS_2 Σ des carrés des résidus des deux régressions ($RSS_2 > RSS_1$)

n_1, n_2 Dégrés de liberté des régressions (En général $n_1 = n_2$)

- Si $F(n_2, n_1) > F_{théorique, n_1, n_2, \alpha}$, on rejette H_0. Le modèle souffre d'hétéroscédasticité

Test de WHITE

R : Voir TP7

Le test de **WHITE** permet non seulement de déceler une éventuelle hétéroscédasticité, mais aussi d'identifier la forme que revêt cette hétéroscédasticité. Le test de White peut être fortement influencé par la présence d'outliers.

- Teste si la variance des erreurs d'un modèle de régression est constante (homoscédasticité).
- L'idée du test de White est de régresser le carré des résidus sur les variables explicatives, leur carré et tous les produits croisés.
- Il suffit ensuite de tester si e^2 dépend d'au moins un des termes de l'équation. Si on rejette l'hypothèse nulle qu'ils sont tous égaux à zéro, le test permettre de conclure qu'il existe de l'hétéroscédasticité et il faudra y remédier.
- Statistique

$nR^2 \sim \chi^2_{p-1}$

R^2 Coefficient de détermination de e^2

n Taille de l'échantillon

p Nombre de coefficients γ à déterminer

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$E(Y) = \hat{Y} = b_1 + b_2 X_2 + b_3 X_3 \qquad Y = \hat{Y} + e$$

$$e^2 = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_2^2 + \gamma_5 X_3^2 + \gamma_6 X_2 X_3 + \varepsilon$$

$$H_0 : \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = 0 \quad H_1 : \gamma_i \neq 0 \text{ pour au moins un } i$$

Exemple

```
.
. whitetst

White's general test statistic :  148.7045  Chi-sq(19)  P-value =  3.9e-22
```

Voir aussi : https://www.dummies.com/education/economics/econometrics/test-for-heteroskedasticity-with-the-white-test/

Notes :

1) Dans les cas où le test de White est statistiquement significatif, l'hétéroscédasticité n'en est pas nécessairement la cause, ce peut être également dû à une mauvaise spécification du modèle. En conséquence, le test de White apparaît non seulement comme un test d'hétéroscédasticité mais également comme un test de spécification. Lorsqu'aucun terme d'interaction n'est introduit (les produits entre tous les couples de variables) alors le test de White est un pur test d'hétéroscédasticité. Mais si ces produits croisés ne sont pas introduits alors c'est à la fois un test d'hétéroscédasticité et de biais de spécification.
2) La procédure de White permet également de définir une matrice variance-covariance « robuste » à l'hétéroscédasticité (matrice sandwich). On obtient alors des écarts-types « robustes ». Remarquons que les coefficients estimés restent strictement les mêmes que dans l'estimation sans l'option « robust ». L'estimation des coefficients reste donc inefficiente, mais non biaisée comme auparavant. Cette correction de White peut s'avérer peu performante en petit échantillon.

Indépendance sérielle et Autocorrélation.

Définition

$Cor\left(\varepsilon_i, \varepsilon_j\right) = 0, \quad i \neq j$. La corrélation entre les résidus est nulle. Les valeurs du terme d'erreur dans les observations de l'échantillon doivent être générées indépendamment les unes des autres.

Si $Cor\left(\varepsilon_i, \varepsilon_j\right) \neq 0, \quad i \neq j$ alors il y a **autocorrélation** ou **dépendance sérielle**.

Conséquence de la dépendance sérielle :

- Les estimateurs par la méthode MCO fournira des estimateurs **non efficaces et non biaisés**.
- L'erreur standard ne sera pas valide et donc l'inférence[3] ne sera pas valide également.

Modèles

- **Autocorrélation autorégressive d'ordre 1 (AR(1))** est le modèle le plus courant. Dans ce modèle, on retrouve comme variable explicative une variable dépendante retardée (Y_{-1}).

[3] En statistique, l'inférence associe un degré de probabilité à une action, une proposition, etc en liaison avec une autre action, proposition etc.

$Y_t = \beta_1 + \beta_2 X_t + u_t$ $u_t = \rho u_{t-1} + \varepsilon_t$ **Modèle Restreint** $Y_t = \beta_1(1-\rho) + \rho Y_{t-1} + \beta_2 X_t - \beta_2 \rho X_{t-1} + \varepsilon_t$ **Modèle Non Restreint** $Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 X_t + \lambda_3 X_{t-1} + \varepsilon_t$ **Restriction incorporée dans le processus AR(1)** $\lambda_3 = -\lambda_1 \lambda_2$	$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$ $u_t = \rho u_{t-1} + \varepsilon_t$ **Modèle Restreint** $Y_t = \beta_1(1-\rho) + \rho Y_{t-1} + \beta_2 X_{2t} - \beta_2 \rho X_{2t-1} + \beta_3 X_{3t} - \beta_3 \rho X_{3t-1} + \varepsilon_t$ **Modèle Non Restreint** $Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 X_{2t} + \lambda_3 X_{2t-1} + \lambda_4 X_{3t} + \lambda_5 X_{3t-1} + \varepsilon_t$ **Restrictions incorporées dans le processus AR(1)** $\lambda_3 = -\lambda_1 \lambda_2 \qquad \lambda_5 = -\lambda_1 \lambda_4$

En général le nombre de restrictions dans le modèle AR(1) est égale au nombre de variables explicatives.
Note : variable laguée = variable dépendante retardée. (Y_t dépend de $\underline{Y_{t-1}}$)

- L'AR(1) est autorégressive, car u_t dépend de ses propres valeurs passées, et de premier ordre, car elle dépend seulement de sa valeur précédente. u_t dépend aussi de t, une injection d'un caractère aléatoire nouvellement apparue au temps t, souvent décrite comme l'**innovation** au temps t.

- Les écarts-types estimés des MCO et les tests d'inférence statistiques ne sont pas fiables (même si les β_j restent sans biais). Les vrais écarts-types sont sous-estimés.
- L'autocorrélation positive est plus fréquente que l'autocorrélation négative.

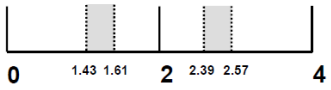
Test de DURBIN-WATSON

- Le test de **DURBIN-WATSON** permet de déterminer s'il y a de **l'autocorrélation d'ordre 1.**
- Le nombre d'observations doit être supérieur ou égal à 15. Le test de Durbin et Watson est un test présomptif d'indépendance des erreurs du fait qu'il utilise les résidus.
- Plus la statistique du test est proche de 2, et plus on peut avoir confiance qu'il n'y a pas d'autocorrélation. En général, on admet une valeur entre 1 et 3. Certains considèrent une valeur entre 1.5 et 2.5. En dessous on une autocorrélation positive et au-dessus une autocorrélation négative.

$d = \dfrac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$ **Pour de grands échantillons** $d \to 2 - 2\rho$ **No autocorrelation** $d \to 2$ **Severe positive autocorrelation** $d \to 0$ **Severe negative autocorrelation** $d \to 4$	e_t et e_{t-1} : résidus des MCO $H_0 : \rho = 0$ Pas d'autocorrélation. $H_1 : \rho > 0$

Exemple

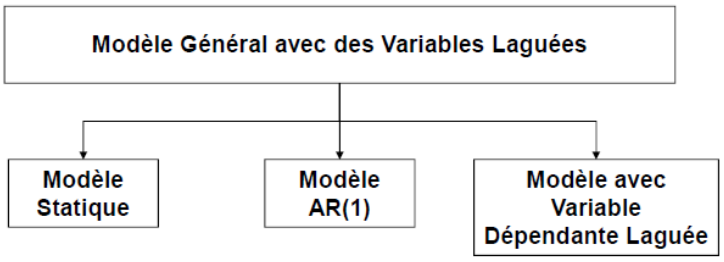
```
. reg  LGFOOD LGDPI LGPRFOOD

      Source |       SS       df       MS              Number of obs =      45
-------------+------------------------------           F(  2,    42) = 4347.62
       Model |  2.17339266     2  1.08669633            Prob > F      =  0.0000
    Residual |  .010497979    42  .000249952           R-squared     =  0.9952
-------------+------------------------------           Adj R-squared =  0.9950
       Total |  2.18389064    44  .049633878           Root MSE      =  .01581

------------------------------------------------------------------------------
      LGFOOD |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       LGDPI |   .6437701   .0259776    24.78   0.000     .5913451    .696195
   LGPRFOOD |  -.1107087   .0203594    -5.44   0.000    -.1517956   -.0696217
       _cons |   1.137597   .1372463     8.29   0.000     .8606231    1.414571
------------------------------------------------------------------------------

. dwstat

Durbin-Watson d-statistic(  3,     45) =  .8294176
```



La statistique *d* est très basse, sous d_L pour le test à 5% (1.35), donc nous rejetons l'hypothèse nulle qu'il n'y a pas d'autocorrélation.

Solution – Général vers spécifique

Voir TP 10



- Etape 1 : Nous commençons avec un **modèle général** comprenant toutes les variables laguées. Pour déterminer si notre modèle est statique ou dynamique, nous allons poser comme hypothèse nulle que les coefficients devant les variables laguées sont égaux à 0. (Une **variable laguée** est une variable retardée). Autrement dit, cela signifie qu'il n'y a pas d'autocorrélation. Si nous rejetons cette hypothèse nulle, notre modèle sera **dynamique**[4].

$$Y_t = \lambda_0 + \lambda_1 Y_{t-1} + \lambda_2 X_{1t} + \lambda_3 X_{1t-1} + \lambda_4 X_{2t} + \lambda_5 X_{2t-1} + \varepsilon_i \qquad (1)$$

$$H_0 : \lambda_1 = \lambda_3 = \lambda_5 = 0$$

Si on rejette H_o, on a un modèle **dynamique**.
Si on accepte H_0, on a un modèle **statique**
Attention : cette régression doit se faire par la commande dynlm() en R.

[4] Un modèle est dit **dynamique** s'il fait intervenir des variables décalées dans le temps contrairement aux modèles statiques. Un modèle dynamique fait donc intervenir des retards sur une ou plusieurs variables.
Si ces variables sont uniquement exogènes, on parlera de **modèles à retards échelonnés**.
Si ces variables retardées correspondent à l'endogène (*Y*), on parlera de **modèles autorégressifs**.
En général, on rencontre souvent des modèles autorégressifs et à retards échelonnés.

- Etape 2 : On construit l'équation correspondant à l'AR1, dans laquelle l'autocorrélation devrait être éliminée, ce qui peut être vérifié par un test.
  - On part de l'équation sans lag

$$Y_t = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \varepsilon_i \qquad (2)$$

$$\rho Y_{t-1} = \alpha_0 + \alpha_1 X_{1t-1} + \alpha_2 X_{2t-1} + \varepsilon_i \qquad (3)$$

$$(2)-(3) \Rightarrow \quad Y_t = \alpha_0 + \rho Y_{t-1} + \alpha_1 X_{1t} - \rho\alpha_1 X_{1t-1} + \alpha_2 X_{2t} - \rho\alpha_2 X_{2t-1} + \varepsilon_i \qquad (4)$$

Cette équation (4) est l'équation dite **AR1** ou encore l'équation **restreinte (ou contrainte)**. *Important* : notons que cette équation est **non-linéaire.**

  - Par identification entre (1) et (4), on trouve

$$\rho = \lambda_1, \quad \alpha_1 = \lambda_2, \quad \alpha_2 = \lambda_4$$

Ainsi que

$$\lambda_3 = -\lambda_1\lambda_2 \quad \text{et} \quad \lambda_5 = -\lambda_1\lambda_4 \qquad (5)$$

Les deux relations (5) forment les **restrictions** appelées plus exactement **les restrictions de facteurs commun.** En d'autres termes, si on impose aux coefficients de l'équation (1) les restrictions (5), on obtient l'équation (4). En principe, il y a autant de restrictions que de variables indépendantes.
Attention : l'équation AR1 n'est pas une équation linéaire, et, en R, il faut utiliser la commande nls()

- Etape 3 : Si on a un modèle dynamique, nous devons faire un test qui nous permet de voir si le modèle AR1 serait le plus approprié. Il s'agit du **test de facteur commun**[5] car le test *F* n'est plus approprié puisqu'il s'agit d'un modèle non-linéaire. Nous allons donc utiliser une statistique alternative qui permet de comparer les RSS du modèle restreint et du modèle non-restreint. On se pose donc sur la significativité de la perte de validité d'approximation.

$$n \log \frac{RSS_t}{RSS_u} \sim \chi^2(\rho)$$

n	nombre d'observations
RSS_t	Somme des carrés des résidus dans le modèle restreint
RSS_u	Somme des carrés des résidus dans le modèle non restreint.
χ^2	Nombre de degré de libeté = nombre de restrictions.

H_0 : les restrictions sont valides ➔On garde le modèle AR(1)
H_1 : les restrictions sont non valides ➔On garde le modèle général.

  - H_0 : Si on a déterminé au point précédent qu'on gardait AR(1), il faut maintenant tester si le modèle souffre toujours d'autocorrélation. Voir test *h* de Durbin-Watson ou test de Breusch-Godfrey
  - H_1 : Si on a déterminé que le modèle AR(1) n'est pas significativement différent du modèle général, on peut conclure que l'autocorrélation d'ordre 1 se trouvait uniquement dans le terme d'erreur et on garde le modèle général.

[5] C'est le test de Sargan-Hendry-Mizon. On parle de facteurs communs au sens ou dans l'équation AR1 plusieurs des coefficients des variables possèdent un facteur commun ; Par exemple $\lambda_3 = -\lambda_1\lambda_2$ et $\lambda_5 = -\lambda_1\lambda_4$ possède λ_1 en commun.

- Etape 4 : Si on a déterminé que l'on ne gardait pas l'AR(1), on peut tester un modèle statique (2), dans lequel on peut introduire un lag uniquement que la variable dépendante.

$$Y_t = \alpha_0 + \rho Y_{t-1} + \alpha_1 X_{1t} + \alpha X_{2t} + \varepsilon_i$$

On vérifiera évidemment l'autocorrélation de ce modèle.

Test h de DURBIN-WATSON

- Si **une seule variable dépendante laguée qui est aussi une variable explicative**, on utilise de test de DURBIN-WATSON alternatif dit *h* de DURBIN-WATSON. (On ne peut pas utiliser le test de DURBIN-WATSON vu plus haut lorsqu'il y a des variables laguées dans le modèle, ce qui est le cas de AR(1))

$Y_t = \beta_1 + \beta_2 X_t + \beta_3 Y_{t-1} + u_t$ $u_t = \rho u_{t-1} + \varepsilon_t$ $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + \beta_3 Y_{t-2} + u_{t-1}$ $h = \hat{\rho}\sqrt{\frac{n}{1 - n s^2_{b_{Y(-1)}}}}$ $d \approx 2 - 2\rho$ $\hat{\rho} = 1 - 0.5d$ $h = (1 - 0.5d)\sqrt{\frac{n}{1 - n s^2_{b_{Y(-1)}}}}$	ρ : estimation de ρ = le paramètre dans le processus AR(1) n : nombre d'obervations $s^2_{b_{Y(-1)}}$: estimation de la variance du coefficient de la variable laguée

Le *h* de DW est normalement distribué.

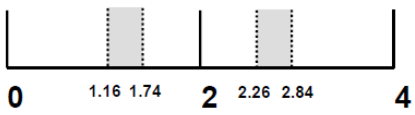
Exemple

```
. reg  LGHOUS LGDPI LGPRHOUS l.LGHOUS if DATE>=1964&DATE<=1994
```

Source	SS	df	MS	Number of obs = 31
Model	2.89802489	3	.966008296	F(3, 27) =18859.89
Residual	.001382947	27	.00005122	Prob > F = 0.0000
				R-squared = 0.9995
Total	2.89940784	30	.096646928	Adj R-squared = 0.9995
				Root MSE = .00716

LGHOUS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
LGDPI	.2625971	.0573576	4.58	0.000	.144909	.3802853
LGPRHOUS	-.0391084	.0136851	-2.86	0.008	-.0671879	-.0110289
LGHOUS L1.	.7875359	.0462763	17.02	0.000	.6925848	.882487
_cons	-.642853	.1954479	-3.29	0.003	-1.043879	-.241827

```
. dwstat

Durbin-Watson d-statistic(  4,    31) =  1.915793
```



. Durbin h=**0.2426**

Test de BREUSCH-GODFREY

R : Voir TP10
Commande R : bgtest()
Commande STATA : estat bgodfrey

- **Une ou plusieurs** variables indépendantes laguées. On utilise le test de **BREUSCH-GODFREY**. Ce test, fondé sur un test de Fisher de nullité de coefficients ou de Multiplicateur de Lagrange (« *LM test* »), permet de tester une **autocorrélation d'un ordre supérieur à 1** et reste valide en présence de la variable dépendante décalée en tant que variable explicative. L'idée générale de ce test réside dans la recherche d'une relation significative entre le résidu et ce même résidu décalé.
- **Attention :** Le test de BREUSCH-GODFREY n'est plus valable lorsqu'on introduit une variable dépendante (Y_{-1}) car il se base sur les résidus qui ne sont pas valable en cas d'autocorrélation.
- Le test se fait en trois étapes **:**
    - Estimation par les MCO du modèle et calcul du résidu u_t, puisque les erreurs sont inconnues, le test porte sur les résidus.
    - Estimation par les MCO de l'équation intermédiaire. n est le nombre d'observations disponibles (attention chaque décalage entraîne la perte d'une observation, ce qui explique dans la formule donnée plus loin le facteur $n - p$) pour estimer les paramètres du modèle et R^2 le coefficient de détermination.
    - Test d'hypothèses sur l'équation intermédiaire. H_0 est que tous les ρ sont nuls. Si on refuse l'hypothèse nulle, alors il existe un risque d'autocorrélation des erreurs à l'ordre p. Pour mener ce test, nous avons deux possibilités : soit effectuer un test de Fisher classique de nullité des coefficients ρ_i, soit recourir à la statistique LM qui est distribuée comme un χ^2 à p degrés de liberté ; si $(n-p)R^2 > \chi^2_{p,théorique}$ au seuil α, on rejette l'hypothèse 'indépendance des erreurs.

$Y_t = \beta_1 + \beta_2 X_t + u_t$ $\hat{u}_t = \rho_1 \hat{u}_{t-1} + \rho_2 \hat{u}_{t-2} + ... + \rho_p \hat{u}_{t-p} + \varepsilon_t$ $H_O : \rho_1 = \rho_2 = ... = \rho_p = 0$ $(n-p)R^2 \sim \chi_p^2$	n Nombre d'observations p Ordre de l'autocorrélation R^2 R^2 de la régression sur u_t χ_p^2 p degré de liberté

Exemple

```
. regress consump wagegovt
```

Source	SS	df	MS	Number of obs =	22
				F(1, 20) =	17.72
Model	532.567711	1	532.567711	Prob > F =	0.0004
Residual	601.207167	20	30.0603584	R-squared =	0.4697
				Adj R-squared =	0.4432
Total	1133.77488	21	53.9892799	Root MSE =	5.4827

consump	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
wagegovt	2.50744	.5957173	4.21	0.000	1.264796	3.750085
_cons	40.84699	3.192183	12.80	0.000	34.18821	47.50577

```
. estat bgodfrey, lags(1/2)

Breusch-Godfrey LM test for autocorrelation
```

lags(p)	chi2	df	Prob > chi2
1	14.264	1	0.0002
2	16.157	2	0.0003

Normalité :

$\varepsilon_i \sim \text{Normal}$. La répartition des résidus suit une loi normale.

La vérification de la normalité peut se faire via de nombreux tests : Kolmogorov-Smirnov, Shapiro, etc.

Pas de paramètres incidentaux.

Le nombre de variables explicatives *K* ne doit pas augmenter avec le nombre d'observations *N*.

Les modèles à variables dépendantes binaires.

Note : attention aux interprétations. *Y* est exprimé sur une échelle de 0 à 1. Quand *Y* passe de 40% à 44%, il faut dire que *Y* augmente de 4 **points de pourcentage** et pas de 4%. Par exemple, si $\beta_1 = -0.05$, alors si X_1 augmente de 1 unité alors *Y* diminue de 5 points de pourcentage.

Modèle de probabilité linéaire (MCO)

La variable dépendante ne peut prendre que deux valeurs, en général 0 ou 1. Le modèle est

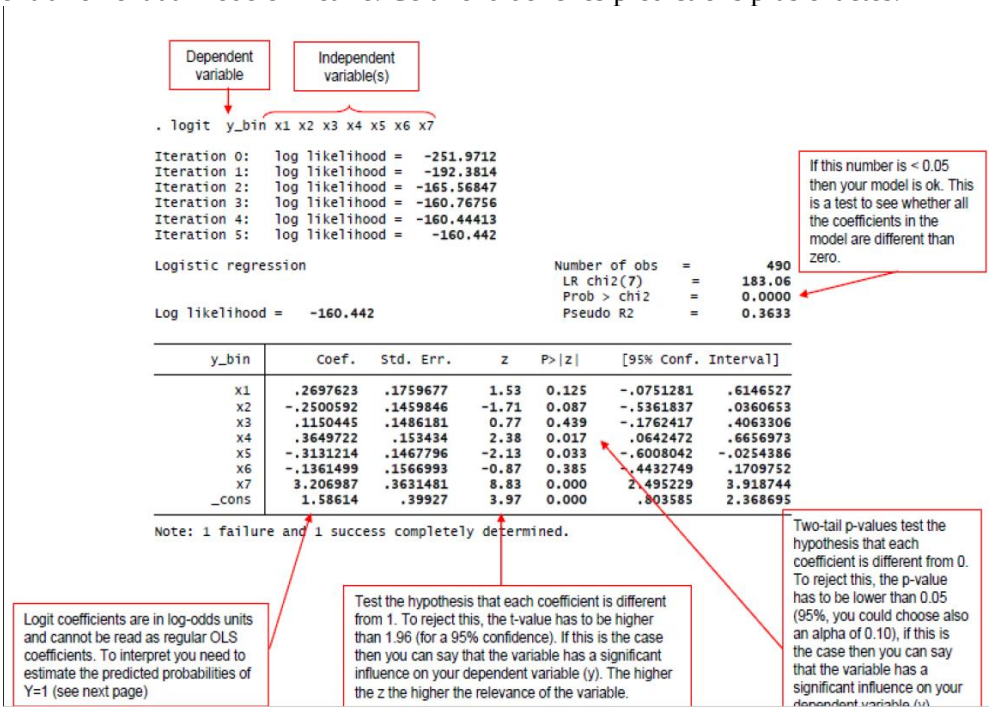
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i$$

Inconvénients

- Possibilité d'obtenir des estimations de probabilité anormale : $\hat{P} > 1$ ou $\hat{P} < 0$
- Les termes d'erreur ne sont pas distribués normalement : $u \backsim N(0, \sigma_u^2)$. En fait, la distribution du terme d'erreur n'est même continue.
- Les estimateurs sont biaisés.
- **Hypothèse d'homoscédasticité non remplie**. Les termes d'erreur sont spécifiques pour chaque individu. $\text{var}(u_i) = \sigma_i^2$. Un modèle dichotomique est toujours hétéroscédastique.

LOGIT, PROBIT.

- Les **coefficients** fournis par le logit ou probit sont **illisibles** en tant que tel. Seul le signe est lisible. En conséquence, ne jamais comparer deux coefficients. Par ailleurs, nous observons que l'**élasticité est variable**.
- L'avantage d'un modèle logit ou probit est que les valeurs obtenues ne dépassent jamais les bornes 0 et 1, tandis que le modèle linéaire peut donner des valeurs aberrantes, inférieures à 0% et supérieures à 100%. L'élasticité d'un modèle logit ou probit est plus vraisemblable qu'un modèle linéaire, puisque ce premier attribue un effet marginal pour chaque individu contrairement au modèle linéaire. Cela rend donc les prédictions plus exactes.



Effets marginaux

Il est important de calculer les effets marginaux car les fonctions logit et probit sont des fonctions non-linéaires. Si dans une modèle de probabilité linéaire, l'effet marginal est le même pour tous les individus, dans un modèle logit ou probit l'effet est différent pour chaque individu.

Effet marginal à la moyenne (ou **au point moyen**) : on évalue l'effet marginal en prenant la moyenne de chaque variable explicative. Ce qui revient à définir un individu moyen qui n'existe probablement.

Une deuxième façon de faire est de calculer l'effet marginal de chaque individu (qui existe) et de prendre la moyenne des effets marginaux. On obtient l'**effet marginal moyen**. Cette méthode est en général préférable.

```
Marginal effects after probit
      y  = Pr(insurance) (predict)
         =  .37617223
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
  retire*|   .0371605      .01965     1.89   0.059  -.001362  .075683   .624766
     age |    .079193      .04496     1.76   0.078  -.008923  .167309   66.9139
    age2 |  -.0006188      .00033    -1.86   0.063  -.001272  .000034   4490.98
hstatusg*|   .0700977      .02064     3.40   0.001   .029636  .110559   .704616
hhincome |   .0004543      .00015     3.09   0.002   .000166  .000743   45.2639
educyear |   .0300312       .0031     9.69   0.000   .023954  .036108   11.8986
  female*|  -.0325755      .01936    -1.68   0.092  -.070522  .005371   .477854
 married*|    .125826      .02056     6.12   0.000   .085533  .166119   .733001
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Rappel : ne jamais interpréter le coefficient de l'effet marginal, mais uniquement son signe. Par exemple, de 0.08 ne signifie pas une probabilité ou un effet plus grand qu'une valeur de 0.05, mais simplement une probabilité plus grande par rapport à la référence de base.

Comment évaluer les différents modèles ?

- Count R^2
- McFadden R^2
- AIC

Pour la définition de ces critères, au paragraphe « Comment juger de la qualité d'une régression ? »

LIKEHOOD RATIO test: LR test.

Ce test permet de savoir si on améliore de façon significative en incluant des variables explicatives

H_0 : l'inclusion des variables n'améliore pas la fonction de vraisemblance.

H_1 : l'inclusion de variables améliore la vraisemblance.

Démarche :

- Estimer le modèle avec constante uniquement : $Probit_0$ et récupérer la valeur de la fonction log de vraisemblance.
- Estimer le modèle complet avec toutes les variables : *Probit*
- Statistique du test

$$LR = 2\left(\log L\left(Probit\right) - \log L\left(Probit_0\right)\right)$$

Suit une distribution du χ^2 avec comme nombre de degré de liberté le nombre de variables ajoutées.

WALD test.

Le test de Wald est souvent utilisé pour tester la significativité individuelle des coefficients (comme le test de Student avec les MCO)
Soit le modèle :

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{matrix}$$

Statistique

$$z = \frac{b_1}{se\left(b_1\right)}$$

La statistique *z* suit une loi normale (comme la distribution au sein de la population est inconnue)
Le test peut être utilisé également de manière plus générale pour les tests sur plusieurs coefficients. Dans ce cas la statistique de Wald est (les β étant des vecteurs) :

$$W = \left(b - \beta_{H_O}\right)\left(\text{var}\left(b\right)\right)^{-1}\left(b - \beta_{H_0}\right)'$$

La statistique de Wald suivra alors une distribution de χ^2 avec comme ddl le nombre de paramètres estimés. On peut tester par exemple si l'ensemble des coefficients sont différents de 0 (comme le test *F*)

TOBIT : modèle de régression tronqué ou censuré

La variable dépendante est sujette à **une limite supérieure ou inférieure** (parfois les deux). Si une variable dépendante est censurée (c'est-à-dire qu'un résultat positif n'est pas observé pour toutes les observations), elle provoque une concentration d'observations à des valeurs nulles. Si ce n'est pas pris en compte dans la procédure d'estimation, une estimation des moindres carrés ordinaires produira des estimations biaisées des paramètres. Avec les variables dépendantes censurées, il y a violation de l'hypothèse de Gauss-Markov de corrélation nulle entre les variables indépendantes et le terme d'erreur.

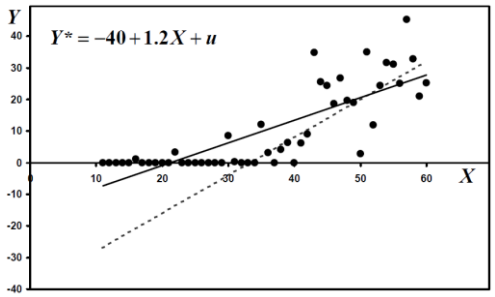
Utilisation des MCO

Par exemple soit le modèle :

$$Y = \beta_0 + \beta_1 X + u \qquad \text{avec} \quad Y \geq 0$$

Dans ce cas :

- Les estimateurs seront biaisés et non convergents.
- L'estimateur de β_0 est biaisé vers le bas
- L'estimateur de β_1 est biaisé vers le haut.



- Notons que la sélection des valeurs de Y non négatives entraîne un biais de sélection.

Tobit

Le modèle Tobit est un mélange mixte entre une régression linéaire pour les valeurs positives et une régression Logit ou Probit pour les autres observations.

Inverse du ratio de Mill

En statistiques, l'**inverse du ratio de Mill** est le ratio de la densité de probabilité sur la fonction de répartition d'une variable aléatoire.

$$m = \frac{\text{Densité de probabilité}}{\text{Fonction de répartition}}$$

Une application courante du ratio Mills inverse (parfois aussi appelé « risque de non-sélection ») se produit dans l'analyse de régression pour tenir compte **d'un biais de sélection possible**.
Lorsque l'inverse du ratio de Mill n'est pas significatif, cela veut dire qu'il n'existe pas de biais de sélection et la MCO peut être utilisée. Cela veut aussi dire que l'échantillon a été tiré aléatoirement.

La procédure de Heckman

James Heckman a proposé une procédure d'estimation en deux étapes utilisant le ratio inverse de Mills pour corriger le biais de sélection.
Dans un premier temps, une régression pour observer un résultat positif de la variable dépendante est modélisée avec un modèle probit. Le rapport Mills inverse doit être généré à partir de l'estimation d'un modèle probit, un logit ne peut pas être utilisé. Le modèle probit suppose que le terme d'erreur suit une distribution normale standard. Les paramètres estimés sont utilisés pour calculer le ratio inverse de Mills, qui est ensuite inclus comme variable explicative supplémentaire dans l'estimation de la MCO.

Dans un deuxième temps, on régresse les valeurs positives de la variable dépendante (si *Y* doit être positif) avec les différentes variables explicatives et l'inverse du ratio de Mill.

Finalement, on compare les résultats d'une régression simple, d'une régression corrigée du biais de censure, d'une estimation faite avec les moindres carrées généralisés corrigés du biais de censure, et, d'une estimation faite avec le maximum de vraisemblance.

Séries temporelles - Stationnarité

Avant le traitement d'une série chronologique, il convient d'en étudier les caractéristiques stochastiques. Si ces caractéristiques – c'est-à-dire son espérance et sa variance – se trouvent modifiées dans le temps, la série chronologique est considérée comme non stationnaire ; dans le cas d'un processus stochastique invariant, la série temporelle est alors stationnaire. De manière formalisée, le **processus stochastique y_t est stationnaire** si :

$E(y_t) = E(y_{t+m}) = \mu \; \forall t$ et $\forall m$, la moyenne est constante et indépendante du temps ;

$\text{var}(y_t) < \infty \; \forall t$, la variance est finie et indépendante du temps ;

$\text{cov}(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)] = \gamma_k$, la covariance est indépendante du temps.

Note : Le processus ne doit pas nécessairement suivre une distribution normale.

Il apparaît, à partir de ces propriétés, qu'un processus de bruit blanc[1] ε_t dans lequel les ε_t sont indépendants et de même loi $N(0,\sigma_{2\varepsilon})$ est **stationnaire**.
Une série chronologique est donc stationnaire si elle est la réalisation d'un processus stationnaire. Ceci implique que la série ne comporte ni tendance, ni saisonnalité et plus généralement aucun facteur n'évoluant avec le temps.
On parle d'**indépendance sérielle** si les termes d'erreur des différentes observations ne sont pas corrélés.

Test de DICKEY et FULLER Augmentés (ADF)

Voir TP10
R : package tseries: adf.test()
Package aTSA : stationary.test()

Les tests de Dickey-Fuller (DF) permettent de mettre en évidence le caractère stationnaire ou non d'une chronique par la détermination d'une tendance déterministe ou stochastique.

Liste des principaux tests utilisés

Test ou paramètre	**Description**
Test t de Student	Test de significativité des coefficients
Test F de Fisher	Test de variances
Test du χ^2	Test d'adéquation à une loi, ou de conformité ou d'indépendance.
Test de Kolmogorov-Smirnov	Test de normalité ou de conformité à une loi donnée.
Test de Shapiro	Test de normalité
Test F marginaux de qualité d'ajustement	On test si l'ajout d'une variable améliore de façon significative la qualité de l'ajustement
R^2	Le coefficient de détermination exprime la variabilité qui est expliquée par le modèle
R^2 ajusté	Le coefficient de détermination ajusté permet de ternir compet de la perte de degré de liberté lié à l'inclusion de variables explicatives supplémentaires.
r	Le coefficient de régression mesure la force de liaison entre Y et X
Test de Fisher du R^2	Test la validité du coefficient de détermination
Test du rapport de vraisemblance ou LR test ou likehood Ratio test	Test la nullité des coefficients introduit dans le modèle, donc permet de savoir si on améliore de façon significative en incluant des variables explicatives supplémentaires.
R Count	Calcul le pourcentage de prédictions correctes. Utilisé pour les variables binaires.
R Count ajusté	Calcul le pourcentage de prédictions correctes après ajustement. Utilisé pour les variables binaires.
Pseudo R^2 de Mc Fadden	Pour les variables binaires est l'équivalent du R^2. Basé sur le maximum de vraisemblance
Pseudo R^2 de Mc Fadden ajusté	Mc Fadden ajusté pour tenir compte du nombre de variables explicatives.
AIC	Critère d'information d'Akaike AIC. Basé sur le maximum de vraisemblance.
BIC	Critère d'information bayésien. Basé sur le maximum de vraisemblance.
VIF	Permet de mesurer la multicolinéarité.
BOX-COX	Transformation Box-Cox permet tester le maximum de vraisemblance de plusieurs modèles.
Test de Fisher pour instrument	Permet de vérifier que l'instrument est bien corrélé avec la variable instrumentée.
Test de Sargan ou test de suridentification ou test de Sargan-Hansen ou test de Hansen ou test J	Permet de tester la validité d'un instrument et son exogénéité. Permet aussi de vérifier la cohérence des instruments.
Test de Durbin-Wu-Hausman ou test de Hausman ou test d'endogénéité.	Permet de tester si une variable suspecte est bien endogène.
Test de Breusch Pagan	Test d'hétéroscédasticité qui consiste à régresser le carré des résidus des MCO sur les variables explicatives.
Test de Goldfeld Quandt	Test la présence d'hétéroscédasticité en divisant l'échantillon en plusieurs parties.
Test de White	Test visant à évaluer la présence d'hétéroscédasticité dans les erreurs d'un modèle et qui consiste à régresser les résidus des MCO au carré sur les variables indépendantes ainsi que leur carré.
Test de Durbin Watson	Permet de déterminer une autocorrélation d'ordre 1 qui est test

	présomptif d'indépendance des erreurs.
Test de nullité des coefficients des variables laguées.	Permet de déterminer si on a un modèle dynamique ou statique.
Test de facteur commun ou test de Sargan-Hendry-Mizon	Permet de déterminer le modèle AR1 est plus approprié que le modèle dynamique générale.
Test h de Durbin-Watson	Si une seule variable laguée, pour vérifier que le modèle AR1 ne souffre pas d'autocorrélation.
Test de Breusch Godfrey	Si une ou plusieurs variables indépendantes laguées, ce test permet de tester une autocorrélation d'un ordre supérieur à 1.
Test de Wald	Ce test est l'équivalent du test de Student pour les modèles à variables binaires (probit, logit). Il permet de tester la significativité individuelle des coefficients du modèle.
Inverse du ratio de Mill	Ce ratio est utilisé dans la procédure de Heckman pour tenir compte du biais de sélection dans le modèle Tobit.
Procédure de Heckman	Procédure pour utilisée pour corriger le biais de sélection dû à la troncature fortuite ou pour corriger toute forme de données manquantes.
Test de Dickey Fuller augmenté ou test ADF	Le test de Dickey-Fuller (DF) permet de mettre en évidence le caractère stationnaire ou non d'une chronique par la détermination d'une tendance déterministe ou stochastique.

Résumé de quelques problèmes typiques rencontrés en économétrie

Problème	Définition	Conséquences	Détection	Solution
Mauvaise spécification	$E(\varepsilon_i) = 0$. La moyenne des résidus n'est pas égale à zéro.	• Omission d'une variable pertinente. ➔ Coefficients biaisés (en général). Ecart-types invalides. • Inclusion d'une variable non pertinente. ➔ Coefficients non biaisés (en général), mais inefficaces Ecart-types valides (en général).		Pour une variable omise, utiliser une variable proxy. Pour une variable excédentaire, il suffit en général de la retirer.
Multicolinéarité élevée.	Deux (ou plus) variables indépendantes d'une régression présente une relation linéaire entre elles.	a) *augmentation de la variance estimée de certains coefficients* lorsque la colinéarité entre les variables explicatives augmente (le *t* de Student diminue) b) *instabilité des estimations des coefficients* des moindres carrés, des faibles fluctuations concernant les données entraînent des fortes variations des valeurs estimées des coefficients ; c) *en cas de multicolinéarité parfaite*, l'*estimation des coefficients est alors impossible et leur variance est infinie*.	Pairwise correlation coefficients Variance inflation factor (VIF)	1. Augmenter le nombre d'observations. 2. Changer le modèle. Utiliser le critère AIC pour sélectionner le meilleur modèle. 3. Éliminer des variables redondantes.
Hétéroscédasticité	$Var(\varepsilon_i) = \sigma^2 \neq$ constant La variance du terme erreur change lorsqu'un changement intervient dans les variables indépendantes.	Les MCO restent sans biais et convergents R^2 ou Adjusted-R^2 reste valide. MAIS, la matrice de variance covariance des coefficients estimés est biaisée Les tests d'hypothèses usuels post-estimation (*t* statistics, *F* statistics ou *LM* statistics) ne sont plus applicables	Park test Goldfeld-Quandt test Breusch-Pagan test White test	1. Weighted least squares (WLS). 2. Robust standard errors. 3. Transformer le modèle (Box-Cox).
Endogénéité	$cor(\varepsilon_i, X_i) \neq 0$. La corrélation entre les variables explicatives X_i et le terme d'erreur n'est pas nulle. Autrement dit, les résidus sont dépendants des variables	Les estimateurs seront **biaisés et non-convergeant**. (**biais d'endogénéité)**. Les causes sont : • Causalité inverse. • Variable Omise. • Erreur de mesure.	Test de Fischer (validation des instruments) Test de Sargan (test d'exogénéité des erreurs et de suridentification) Test de Hausman (compare les modèles avec et sans instruments)	Utilisation des variables instruments (VI)

Autocorrélation	Une relation identifiable (positive ou négative) existe entre les valeurs de l'erreur d'une période et les valeurs de l'erreur d'une autre période.	Idem hétéroscédasticité. Les coefficients des MCO ne sont pas efficients Les erreurs standards sont biaisées Les tests d'hypothèses ne sont pas applicables	Geary or runs test Durbin-Watson test Breusch-Godfrey test	1. Cochrane-Orcutt transformation. 2. Prais-Winsten transformation. 3. Newey-West robust standard errors.