

ETUDE STATISTIQUE

Thème : Démographie

Sujet : Le nombre de naissances

Par Thibault Gibert, 2010.

Etude Statistique et création d'un modèle de régression multiple

Le sujet de cette étude est le nombre de naissance observé de par le monde.
Nous tenterons ici de mettre en avant d'éventuel lien entre le nombre de naissances et certaines caractéristiques propres aux pays.

Notre Etude est réalisée sur 136 pays, sans localisation géographique particulière ni tri sur la situation économique ou sociale.
On peut supposer que ce choix entraîne l'apparition de valeur aberrante en vu des possibles disparités en matière de développement des pays étudiés.

Composition de base du modèle :

L'objectif est d'observer de possibles concordances entre Le taux de natalité brut et les différentes variables explicatives sélectionnées et si possible établir un modèle probant.

Une variable endogène (Y) Le taux de natalité brut (par 1000 habitants)
Il s'agit du nombre de naissances enregistrées au cours d'une année par 1000 habitants.
Rq : Le fait de ramener le nombre de naissances totales à 1000 habitants permet d'inhiber l'effet de la taille de la population et facilite ainsi les comparaisons.

Sept variables exogènes dont:

X1: Population 15-64 ans en % de la pop totale.

La catégorie de 15 à 64 ans est généralement associée à la population active et constitue un élément intéressant dans le sens où il permet de comprendre la pyramide d'âge d'un pays.

Rq : Par hypothèse, on suppose qu'un pays ayant un % élevé sera plus enclin à avoir un taux de natalité brut plus élevé, les 1000 habitants étant plus en âge de procréer statistiquement.

X2: Indicateur de développement humain (IDH)

L'IDH est un indice composé qui mesure la qualité de vie moyenne de la population d'un pays.

X3: Espérance de vie à la naissance (année)

Il s'agit du nombre d'années que les personnes vivent en moyenne.

X4 : Taux de mortalité - enfants (par 1000 naissances)

Il s'agit du nombre d'enfant qui meurent avant d'atteindre 1 an. Le ratio est établi sur 1000 naissances au cours d'une année donnée.

X5 : % d'utilisation de Pratique Contraceptive parmi les Femmes Mariées Age : 15-49 / (Toutes les Méthodes)

Toutes les méthodes : Méthodes modernes et traditionnelles.

X6 : Taux d'activité des femmes, Age: [15+

Il s'agit d'un taux d'activité économique.

X7 : Le PNB par habitant en Dollar.

Rq : Le PNB par habitant sera par la suite utilisé comme critère de tri pour regrouper les pays en catégorie économiquement plus homogène.

Régression: (Les étapes du premier modèle seront détaillées précisément, les modèles suivants plus sommairement pour laisser place aux interprétations).

1°/ Etape : Composition de Base et 136 pays :

L'équation du Modèle va donc s'écrire:

$$Y = \beta_0 \cdot X_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4 + \beta_5 \cdot X_5 + \beta_6 \cdot X_6 + \beta_7 \cdot X_7 + \varepsilon$$

Avec X_0 : La constante = 1

β les différents paramètres.

ε un bruit aléatoire représentant le terme de l'erreur.

Dependent Variable: Y

Method: Least Squares

Date: 12/05/09 Time: 13:20

Sample: 1 136

Included observations: 136

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X7	-5.29E-05	3.01E-05	-1.760546	0.0807
X6	-0.013856	0.019789	-0.700181	0.4851
X5	-0.048259	0.020066	-2.405088	0.0176
X4	0.087610	0.020413	4.291914	0.0000
X3	-0.009986	0.063756	-0.156625	0.8758
X2	0.021400	0.041886	0.510920	0.6103
X1	-1.091758	0.074559	-14.64286	0.0000
C	91.56541	6.884757	13.29973	0.0000
R-squared	0.931899	Mean dependent var	23.66206	
Adjusted R-squared	0.928174	S.D. dependent var	12.01809	
S.E. of regression	3.220883	Akaike info criterion	5.234210	
Sum squared resid	1327.883	Schwarz criterion	5.405543	
Log likelihood	-347.9263	F-statistic	250.2221	
Durbin-Watson stat	2.015509	Prob(F-statistic)	0.000000	

Estimation des paramètres : $(\beta := (X^T \cdot X)^{-1} X^T \cdot Y)$ « Coefficient »

L'équation du modèle devient donc :

$$Y = 91.56541 - 1.091758 X_1 + 0.021400 X_2 - 0.009986 X_3 + 0.087610 X_4 - 0.048259 X_5 - 0.013856 X_6 + -5.29E-05 X_7$$

Test de Student :

On va tester l'hypothèse de nullité des coefficients théoriques,

On teste l'hypothèse $H_0 B_j = 0$ contre l'hypothèse $H_1 B_j \neq 0$

si le Student calculé (en valeur absolue) < Student théorique (tabulé) alors on accepte H_0 , B_i n'est pas significativement différent de 0. La variable explicative X_j n'est pas contributive à l'explication de la variable Y et inversement.

Test de la signification des paramètres estimés au seuil de $\alpha = 5\%$

Le degré de liberté est de : $136-7-1=128$

Valeur du student théorique : 1.979 (Correspondant à un degré de liberté de 120)

Résultat pour un degré de liberté infini où la loi de student converge vers une loi normale avec comme valeur 1.96 .

Pour X1, X4,X5 la valeur du student, (**valeur en gras**) | « t-statistic » | > 1.979 on rejette H0 et accepte H1, leurs coefficients sont significativement différents de 0 et ces variables explicatives sont contributives à l'explication de la variable Y.

Pour X2,X3, et X6 la valeur du student < 1.979 on accepte H0 leurs coefficients ne sont pas significativement différents de 0 et ces variables ne sont pas contributives à l'explication de la variable Y.

X7 avec un student à -1.760546 peut être considéré comme contributif. (Voir par la suite après suppression des trois précédentes).

Coefficient de détermination :

Après lecture du coefficient de détermination et du coefficient de détermination ajusté l'on peut voir que

(R-carré X100) $\approx 93 \%$ de la variance totale de Y est expliqué par la régression.

Test de la qualité globale du modèle au seuil de 5% par le test de Fisher :

On va tester l'hypothèse de nullité de l'ensemble des coefficients Bj,

H0: $B_0=B_1=B_2..B_j=0$ Contre H1 au moins un des coefficients est non nul.

Cela revient à tester la nullité du coefficient de détermination.

Si fisher calculé < valeur du Fisher théorique alors H0 est accepté, il n'existe aucune relation linéaire significative entre la variable endogène et les variables exogènes. Le modèle n'est pas globalement explicatif.

Si fisher calculé > valeur du Fisher théorique alors on rejette H0. L'ensemble des variables exogènes a une influence sur la variable endogène. Le pouvoir explicatif du modèle est suffisant.

Test au seuil de $\alpha = 5\%$

La valeur du Fisher théorique est de 2.0868.

Le Fisher calculé est de 250.2221 <F-statistic> est plus grand que le fisher théorique 2.0868.

On rejette H0 le modèle est globalement explicatif.

Pour les deux observations précédentes (Coefficient de détermination et test de la qualité globale) les résultats sont à nuancer en vue du nombre important de variables exogènes.

Test sur le terme d'erreurs : ε

White Heteroskedasticity Test:

F-statistic	4.322616	Probability	0.000004
Obs*R-squared	45.34163	Probability	0.000036

Test Equation:

Dependent Variable: RESID^2

Method: Least Squares

Date: 12/05/09 Time: 14:11

Sample: 1 136

Included observations: 136

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	866.2016	221.4364	3.911740	0.0002
X7	0.000273	0.000833	0.328083	0.7434
X7^2	-1.86E-09	1.70E-08	-0.109829	0.9127
X6	-0.065379	0.556738	-0.117432	0.9067
X6^2	-0.000811	0.005609	-0.144631	0.8852
X5	0.704274	0.465433	1.513158	0.1328
X5^2	-0.010280	0.004592	-2.238763	0.0270
X4	0.474079	0.339362	1.396971	0.1650
X4^2	-0.003238	0.001804	-1.794867	0.0752
X3	-3.557371	3.081169	-1.154552	0.2505
X3^2	0.031917	0.026652	1.197523	0.2334
X2	-20.89679	46.66719	-0.447783	0.6551
X2^2	0.257256	0.583913	0.440573	0.6603
X1	-26.65958	6.773058	-3.936122	0.0001
X1^2	0.229817	0.053575	4.289624	0.0000
R-squared	0.333394	Mean dependent var	9.763845	
Adjusted R-squared	0.256266	S.D. dependent var	23.86640	
S.E. of regression	20.58238	Akaike info criterion	8.990472	
Sum squared resid	51259.75	Schwarz criterion	9.311720	
Log likelihood	-596.3521	F-statistic	4.322616	
Durbin-Watson stat	1.778344	Prob(F-statistic)	0.000004	

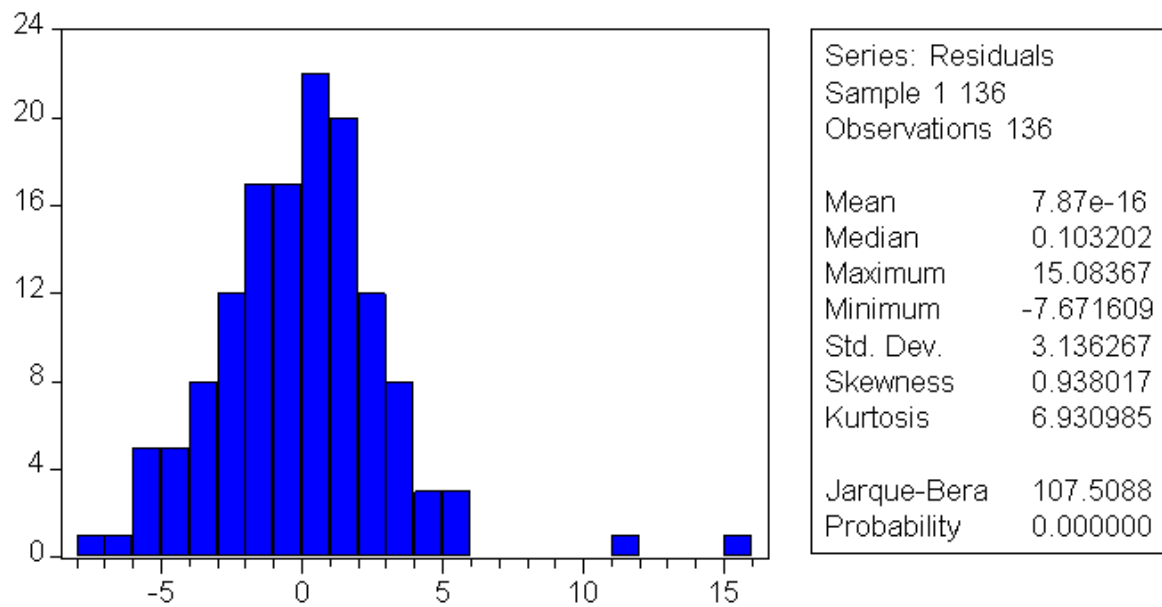
Hétéroscédasticité : Nous nous baserons sur le test de White,

Obs*R-squared est = 45.34163 < χ^2 avec une probabilité de 5% et un u de 1.6449 (loi normale réduite) = 155.1195 .

L'hypothèse d'homoscédasticité des erreurs est acceptée, les résidus sont homoscédastiques, c'est à dire que la variance de l'erreur est identique quelque soit l'observation.

Il s'agit d'une coupe transversale, les données étant simultanées les erreurs ne peuvent pas avoir d'effet de cumul, il n'y a donc pas autocorrélation des erreurs. Le test de Durbin-Watson n'est pas nécessaire.

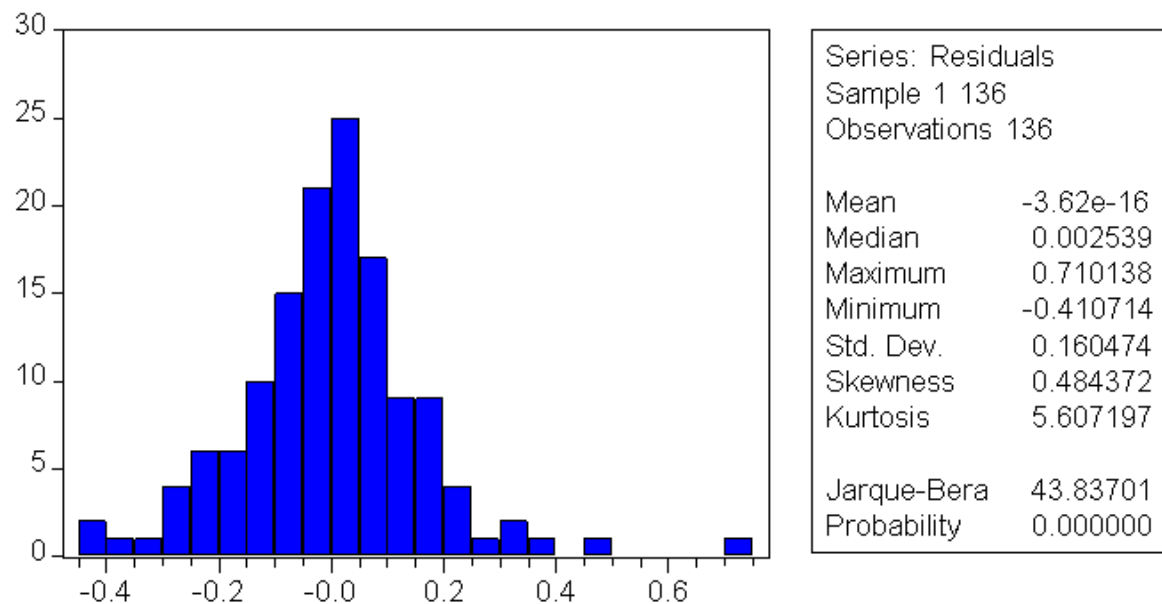
Normalité des résidus : Test de Jarque-Bera.



La valeur du Jarque-Bera est de 107.5088 qui est $>$ qu'un $\chi^2(2) = 5.99$
Les erreurs ne sont pas normalement distribuées.

Le Jarque-Bera étant largement au dessus de sa valeur critique, le modèle n'est donc pas valide.

Après un passage en Log de l'ensemble du modèle (Variable exogènes et endogène) :



La valeur du Jarque-Bera a diminué mais pas suffisamment : 43.83701.

Modèle après suppression des variables X2, X3, X6 :

Modèle 2 Nous allons supprimer X2,X3,X6 qui sont respectivement l'IDH, L'espérance de vie à la naissance et Le tx d'activité économique des femmes.

Le modèle devient donc $Y = \beta_0.X_0 + \beta_1.X_1 + \beta_4.X_4 + \beta_5.X_5 + \beta_7.X_7 + \varepsilon$

Dependent Variable: Y
Method: Least Squares
Date: 12/06/09 Time: 10:50
Sample: 1 136
Included observations: 136

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X7	-5.75E-05	2.75E-05	-2.091919	0.0384
X5	-0.047875	0.019802	-2.417643	0.0170
X4	0.088301	0.014772	5.977418	0.0000
X1	-1.084202	0.072729	-14.90737	0.0000
C	89.72851	4.912901	18.26385	0.0000

R-squared	0.931508	Mean dependent var	23.66206
Adjusted R-squared	0.929417	S.D. dependent var	12.01809
S.E. of regression	3.192903	Akaike info criterion	5.195810
Sum squared resid	1335.497	Schwarz criterion	5.302893
Log likelihood	-348.3151	F-statistic	445.4100
Durbin-Watson stat	2.028492	Prob(F-statistic)	0.000000

Le coefficient de détermination est toujours élevé.

Rq : Nous avons bien fait de garder la variable X7 dont le student -2.091919 est désormais supérieur en valeur absolu à 1.96. Le coefficient β_7 est significativement différent de 0 et la variable explicative est contributive à l'explication de la variable Y.

Il est en de même pour les variables X1,X4 et X5.

Nous passerons ensuite directement au Test de la normalité des erreurs de Jarque-Bera :
Le résultat du Jarque Bera nous donne

Series: Residuals
Sample 1 136
Observations 136

Mean -2.77e-15
Median 0.150034
Maximum 15.11953
Minimum -7.526398
Std. Dev. 3.145246
Skewness 0.944900
Kurtosis 6.917239

Jarque-Bera 107.1913
Probability 0.000000

Transformation en Log =>

Series: Residuals
Sample 1 136
Observations 136

Mean -7.42e-16
Median -0.004940
Maximum 0.702096
Minimum -0.431764
Std. Dev. 0.166246
Skewness 0.490343
Kurtosis 5.178119

Jarque-Bera 32.33371
Probability 0.000000

> à 5.99 les erreurs ne sont pas normalement distribuées.

Malgré le passage en Log la valeur Du Jarque-Bera est > 5.99.

2°/ Etape : Etude sur les 30 pays au PNB/Hab. le plus élevé.

Comme signalé dans l'introduction, notre Etude est réalisée sur 136 pays, sans localisation géographique particulière ni tri sur la situation économique ou sociale. Comparer des pays avec de grands écarts en matière de développement, avec des situations démographiques très différentes entraîne certainement des problèmes de comparaison.

Nous allons donc utiliser X7 non plus comme une variable explicative mais comme critère de tri pour regrouper les pays en catégorie économiquement plus homogène.

Le modèle devient donc $Y = \beta_0.X_0 + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 + \beta_4.X_4 + \beta_5.X_5 + \beta_6.X_6 + \epsilon$

Dependent Variable: Y

Method: Least Squares

Date: 12/06/09 Time: 11:46

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X6	-0.100054	0.063543	-1.574585	0.1290
X5	-0.011979	0.052990	-0.226060	0.8232
X4	0.378590	0.098122	3.858365	0.0008
X3	0.329067	0.217079	1.515888	0.1432
X2	-0.024465	0.039394	-0.621028	0.5407
X1	-0.016821	0.186947	-0.089979	0.9291
C	-9.538037	24.37363	-0.391326	0.6992
R-squared	0.610650	Mean dependent var	11.80567	
Adjusted R-squared	0.509080	S.D. dependent var	3.996093	
S.E. of regression	2.799889	Akaike info criterion	5.098000	
Sum squared resid	180.3057	Schwarz criterion	5.424946	
Log likelihood	-69.47000	F-statistic	6.012133	
Durbin-Watson stat	2.245559	Prob(F-statistic)	0.000677	

On remarque une nette diminution du coefficient de détermination du supposé aux valeurs des variables explicatives qui se resserrent après tri. La variance de Y est désormais expliquée à 61% par le modèle.

Après test de Student apparaît comme significatif Seulement X4. Le taux de mortalité - enfants (par 1000 naissances). Variable qui était déjà nettement significative dans le premier modèle sans tri. Par contre X1 Le % de la population ayant entre 15-64 qui entrainait largement dans l'explication de notre premier modèle n'est plus d'une grande importance ici. Ce qui découle certainement du fait que les Pays regroupés par catégorie de richesse homogène ont sensiblement une pyramide des âges de même composition.

Même si un modèle de régression simple à une variable est peu intéressant, nous allons effectuer les Test avec une seule variable exogène X4.

Modèle : $\text{Log}(Y) = \beta_0 + \beta_1 X_0 + \beta_4 \text{Log}(X_4) + \epsilon$

Dependent Variable: LY

Method: Least Squares

Date: 12/06/09 Time: 12:17

Sample: 1 30

Included observations: 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LX4	0.310198	0.050847	6.100567	0.0000
C	1.895582	0.093319	20.31285	0.0000
R-squared	0.570664	Mean dependent var		2.426474
Adjusted R-squared	0.555330	S.D. dependent var		0.276775
S.E. of regression	0.184564	Akaike info criterion		-0.477305
Sum squared resid	0.953784	Schwarz criterion		-0.383892
Log likelihood	9.159574	F-statistic		37.21691
Durbin-Watson stat	1.984044	Prob(F-statistic)		0.000001

Le coefficient de détermination est de 0.55, La regression simple explique donc 55% de la variance de Y.

Le test du student nous indique que β_4 est significativement différent de 0 et que X4 rentre dans l'explication de la variable Y.

Le test de White pour tester l'hétéroscédasticité :

White Heteroskedasticity Test:

F-statistic	5.081519	Probability	0.013394
Obs*R-squared	8.204150	Probability	0.016538

Obs*R-squared = 8.204150 est inférieur à un $\chi^2(28)=41.34$
La variance de l'erreur est identique quelque soit l'observation.

Test de la normalité des erreurs : Test de Jarque-Bera,

Series: Residuals

Sample 1 30

Observations 30

Mean 4.02e-16
Median -0.012821
Maximum 0.431024
Minimum -0.306882
Std. Dev. 0.181354
Skewness 0.449719
Kurtosis 2.968092

Jarque-Bera 1.012507
Probability 0.602749

$< \chi^2(1)$ à 5% = 3.84 Les erreurs sont donc normalement distribuées.

3°/ Etape : Etude sur les 30 pays aux PNB/Hab le plus faible.

Dependent Variable : Y
Method : Least Squares
Date : 12/06/09 Time : 12 :51
Sample : 1 30
Included observations : 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X6	-0.001784	0.042442	-0.042026	0.9668
X5	-0.100234	0.041172	-2.434490	0.0231
X4	0.058319	0.021771	2.678791	0.0134
X3	-0.129401	0.091916	-1.407817	0.1726
X2	-5.374135	7.548065	-0.711988	0.4836
X1	-1.130190	0.188844	-5.984774	0.0000
C	106.2953	11.89038	8.939603	0.0000
R-squared	0.958640	Mean dependent var	37.48933	
Adjusted R-squared	0.947850	S.D. dependent var	9.358470	
S.E. of regression	2.137132	Akaike info criterion	4.557770	
Sum squared resid	105.0487	Schwarz criterion	4.884716	
Log likelihood	-61.36655	F-statistic	88.84830	
Durbin-Watson stat	1.773525	Prob(F-statistic)	0.000000	

Après le Test de Student on sait que X1,X4,X5 rentre dans l'explication de Y.

Modèle avec X1,X4,X5 :

Dependent Variable : Y
Method : Least Squares
Date : 12/06/09 Time : 13 :01
Sample : 1 30
Included observations : 30

Variable	Coefficient	Std. Error	t-Statistic	Prob.
X5	-0.082331	0.035084	-2.346667	0.0268
X4	0.079414	0.018296	4.340503	0.0002
X1	-1.338897	0.145304	-9.214484	0.0000
C	105.6188	8.409924	12.55883	0.0000
R-squared	0.949595	Mean dependent var	37.48933	
Adjusted R-squared	0.943779	S.D. dependent var	9.358470	
S.E. of regression	2.218987	Akaike info criterion	4.555544	
Sum squared resid	128.0214	Schwarz criterion	4.742370	
Log likelihood	-64.33316	F-statistic	163.2734	
Durbin-Watson stat	1.806056	Prob(F-statistic)	0.000000	

Après suppression des variables qui ne rentraient pas dans l'explication de Y, le test de Student nous permet de dire que X1, X4 et X5 sont bel et bien contributives à l'explication de la variable Y pour les pays aux PNB/hab le plus faible, Ce qui n'était pas le cas pour X1 et X5 pour les pays au PNB/hab le plus fort. Nous commenterons cette différence après avoir effectué les différents tests afin de valider le modèle.

Le coefficient de détermination ajusté est de : Adjusted R-squared = 0.943779
94% de la variance totale de Y est expliqué par la régression.

Test sur le terme d'erreur :

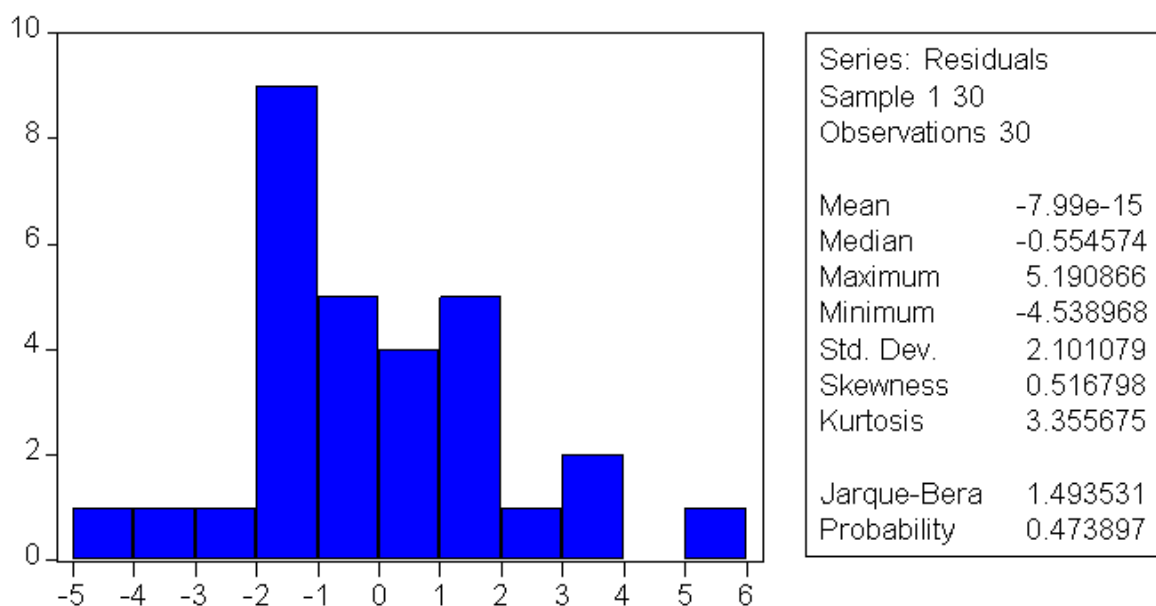
Test de l'hétéroscédasticité, Test de White :

White Heteroskedasticity Test:

F-statistic	0.603166	Probability	0.725004
Obs*R-squared	4.078662	Probability	0.666032

Notre Obs*R-squared = 4.078662 est inférieur à un $\chi^2(26)$ au seuil de 5% égal à 38.88
L'hypothèse d'homoscédasticité des erreurs est vérifiée ce qui signifie que la variance de l'erreur est identique pour toute l'observation.

Test de l'hypothèse de normalité des erreurs : Test de Jarque-Bera :



Le résultat du test est = à $1.493531 < \chi^2(2) = 5.99$
Les erreurs sont donc normalement distribuées.

La méthode d'estimation des moindres carrés ordinaires est donc valide.

En conclusion nous pouvons établir une relation quasi certaine entre le taux d'enfants qui n'atteignent pas 1 an sur 1000 naissances et le taux de natalité brut, variable qui entre dans l'explication au sein des 3 sous cas : Tout les Pays, Pays aux PNB par habitant les plus élevés, Pays aux PNB par habitant les plus faibles. Il y a corrélation positive entre les deux, plus un pays a un taux de mortalité des enfant de moins de 1 an élevé plus son nombre de naissance sera élevé.

Le % de la population totale ayant entre 15 et 64 ans (X1) apparait comme significative dans le cas où l'on prend en compte l'ensemble des pays, car sont présent de forts écarts dans la composition des pyramides des âges des différents pays du monde, écart qui entraîne des différences sur le nombre de naissances. (Age de procréation, il aurait été encore plus intéressant d'avoir une tranche d'âge de 15-49 ans)

Pour les pays au PNB/habitant les plus élevés le % de cette tranche d'âge au sein de la population varie trop peu pour avoir un sens dans la régression.

Par contre pour les Pays au PNB/habitant les plus faibles, celui-ci entre dans l'explication du taux de natalité brut car des disparités existent entre ces pays, dû a des états de transition démographique différents et justifiant ainsi l'importance de cette variable pour cette catégorie économique de pays.

Il en est de même pour la variable % d'utilisation des moyens de contraception chez les femmes (X5) qui n'a de sens que dans notre dernier cas. En effet au sein des pays les plus pauvres existent de fortes différences en ce qui concerne ces pratiques.

Une relation décroissante lie cette variable au taux de natalité, ce qui est logique puisque par hypothèse plus le % de femme utilisant des moyens de contraceptions est élevé moins il y a de chance qu'arrive des grossesses « non voulues » et donc le taux de natalité brut diminue.