1 Introduction

Econométrie – Pratique de la régression linéaire multiple avec le logiciel ''gretl''.

En faisant faire un travail d'analyse économétrique sur les ''Open Data'' à mes étudiants, où ils étaient libres d'utiliser l'outil de leur choix, j'ai constaté que plusieurs d'entre eux ont opté pour ''gretl''. Je connaissais, je l'avais testé vite fait par le passé, mais sans creuser davantage (cf. ma page de cours, section ''Logiciel gratuits''). Ces étudiants avaient parfaitement répondu au cahier des charges de l'étude, là où la majorité de leurs collègues avaient travaillé sous R. Visiblement, ''gretl'' semble proposer des fonctionnalités assez complètes, qui répondent aux attendus de mon enseignement d'économétrie de Licence en tous les cas. Nous allons examiner cela dans ce tutoriel.

Nous prenons comme repère la quatrième séance de TD (travaux dirigés sur machine) de mon cours. L'objectif est d'expliquer la nocivité des cigarettes (teneur en CO – monoxyde de carbone) à partir de leurs caractéristiques (teneur en nicotine, en goudron, poids). Les thèmes abordés sont : l'importation et la description des données, la pratique de la régression linéaire multiple avec l'estimation des paramètres du modèle et l'inspection des résultats. Mettre en parallèle les sorties de ''gretl'' et la correction du TD (traité sous R) accessible en ligne nous permettra de calibrer notre démarche et vérifier les résultats.

2 Le logiciel ''gretl''

''gretl'' est un logiciel de statistique ''open source'' essentiellement dédié à l'économétrie. La première version a été mise en ligne le 31 janvier 2000. La version la plus récente date du 24 janvier 2019 (ce tutoriel est rédigé en Mai 2019). Il peut être piloté par menu, c'est ce que nous mettrons en avant dans ce tutoriel. Mais il propose également d'un langage de script complet nommé ''hansl'', disposant des fonctionnalités essentielles d'un langage de programmation (types de données scalaires et matriciels, branchements conditionnels, boucles, définition de ses propres fonctions, récursivité, structures de données avancées, collections, …).

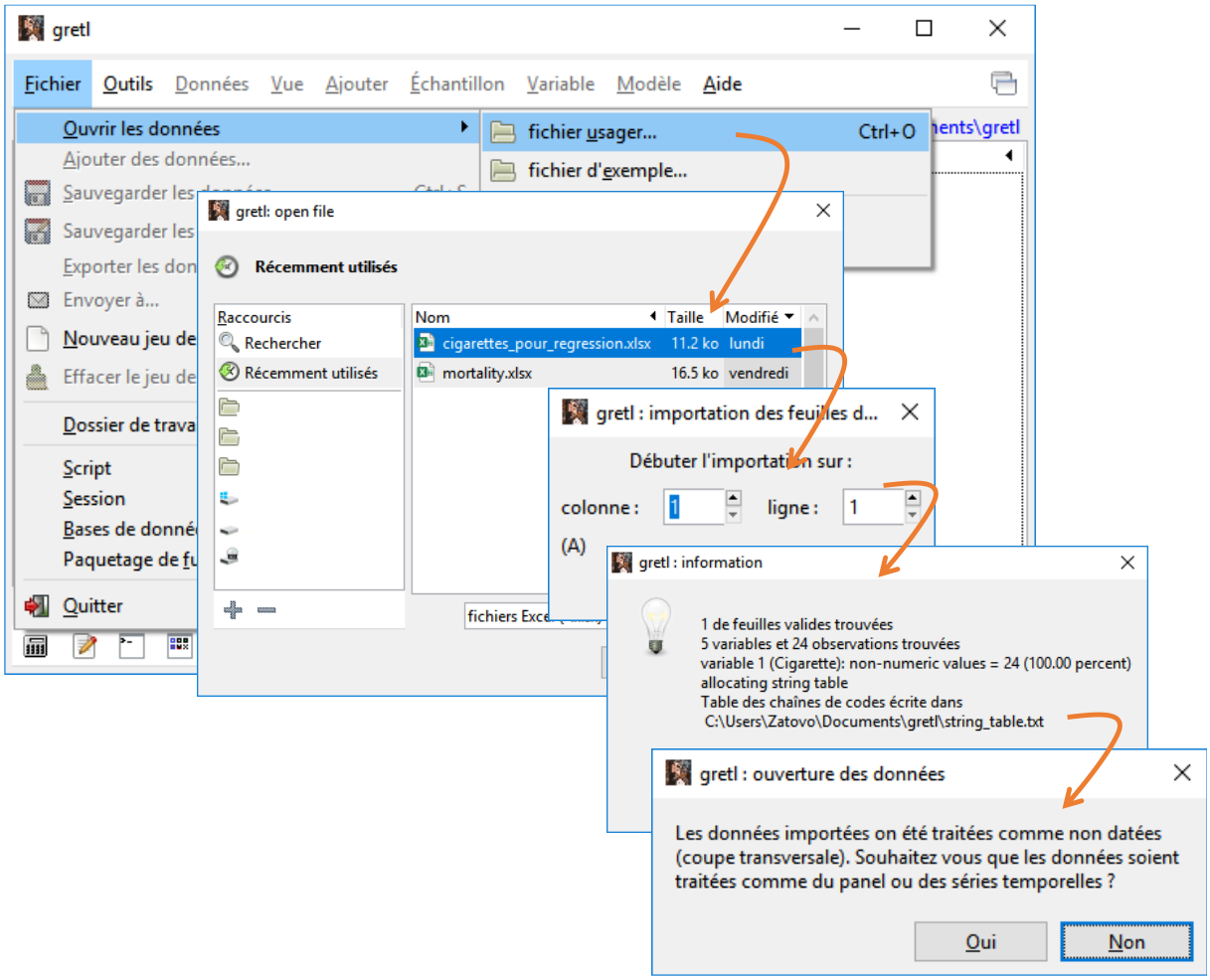
''gretl'' a fait l'objet de plusieurs publications dans des journaux scientifiques de référence, tels que ''Journal of Applied Econometrics'' (2003) ou encore ''Journal of Statistical Software'' (2008) (voir la page Wikipédia en anglais -- https://en.wikipedia.org/wiki/Gretl). Il est supporté par une communauté assez dynamique et donne lieu à des conférences tous les 2 ans depuis 2009 (http://www.gretlconference.org/ -- la prochaine aura lieu à Naples en Juin 2019).

Nous utilisons la version 2019a avec une interface en langue française dans ce tutoriel.

3 Importation et description des données

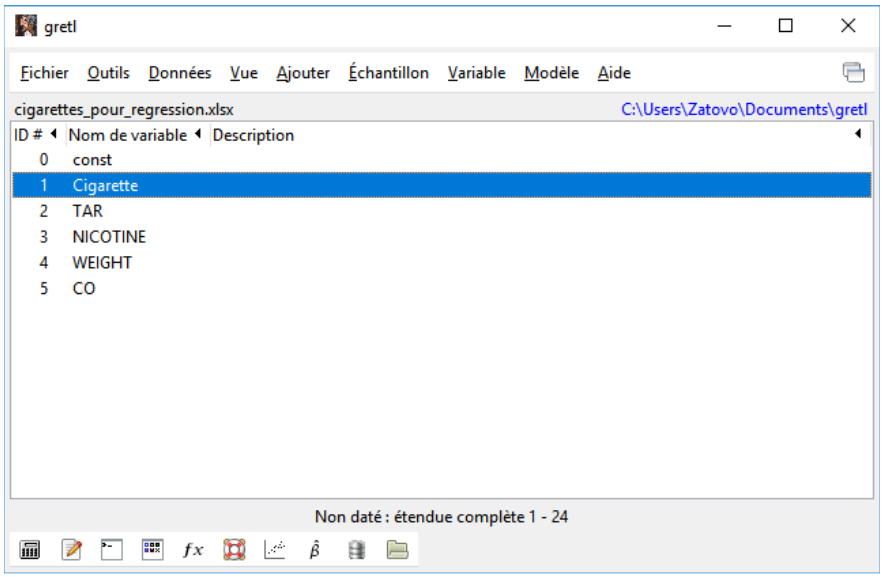
3.1 Importation du fichier

Nous utilisons le fichier ''**cigarettes_pour_regression.xlsx**'' de ma séance de TD n°4. On souhaite expliquer la nocivité des cigarettes (teneur en CO – monoxyde de carbone) à partir de leurs caractéristiques (teneur en nicotine, en goudron, poids). Pour accéder au fichier, nous actionnons le menu **FICHIER / OUVRIR LES DONNEES / FICHIER USAGER**. Une boîte de dialogue permet de sélectionner le fichier. Une succession de boîtes de dialogues vient préciser l'opération.



La dernière est importante. Elle nous permet d'opter pour une coupe transversale ou des données longitudinales. Nous concernant, nous avons un tableau ''individus x variables'', il n'y a pas d'ordre dans les lignes c.-à-d. nous pouvons interchanger les positions des individus dans le tableau sans que cela ne change la nature des données ou des traitements, nous avons bien une coupe transversale.

La liste des variables est énumérée dans la fenêtre principale. Gretl ne sait pas que ''cigarette'' est en réalité une étiquette. Il a ajouté la colonne ''**const**'' qui représentera la constante dans les régressions.



Pour une bonne compréhension des analyses que nous initierons par la suite, voici le tableau de données.

Cigarette	TAR	NICOTINE	WEIGHT	CO
Alpine	14.1	0.86	0.9853	13.6
Benson_Hedges	16	1.06	1.0938	16.6
CamelLights	8	0.67	0.928	10.2
Carlton	4.1	0.4	0.9462	5.4
Chesterfield	15	1.04	0.8885	15
GoldenLights	8.8	0.76	1.0267	9
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3
L_M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13
Marlboro	15.1	0.9	0.9316	14.4
Merit	7.8	0.57	0.9705	10
MultiFilter	11.4	0.78	1.124	10.2
NewportLights	9	0.74	0.8517	9.5
Now	1	0.13	0.7851	1.5
OldGold	17	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.007	15.9
TrueLight	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12	0.82	1.1184	14.9

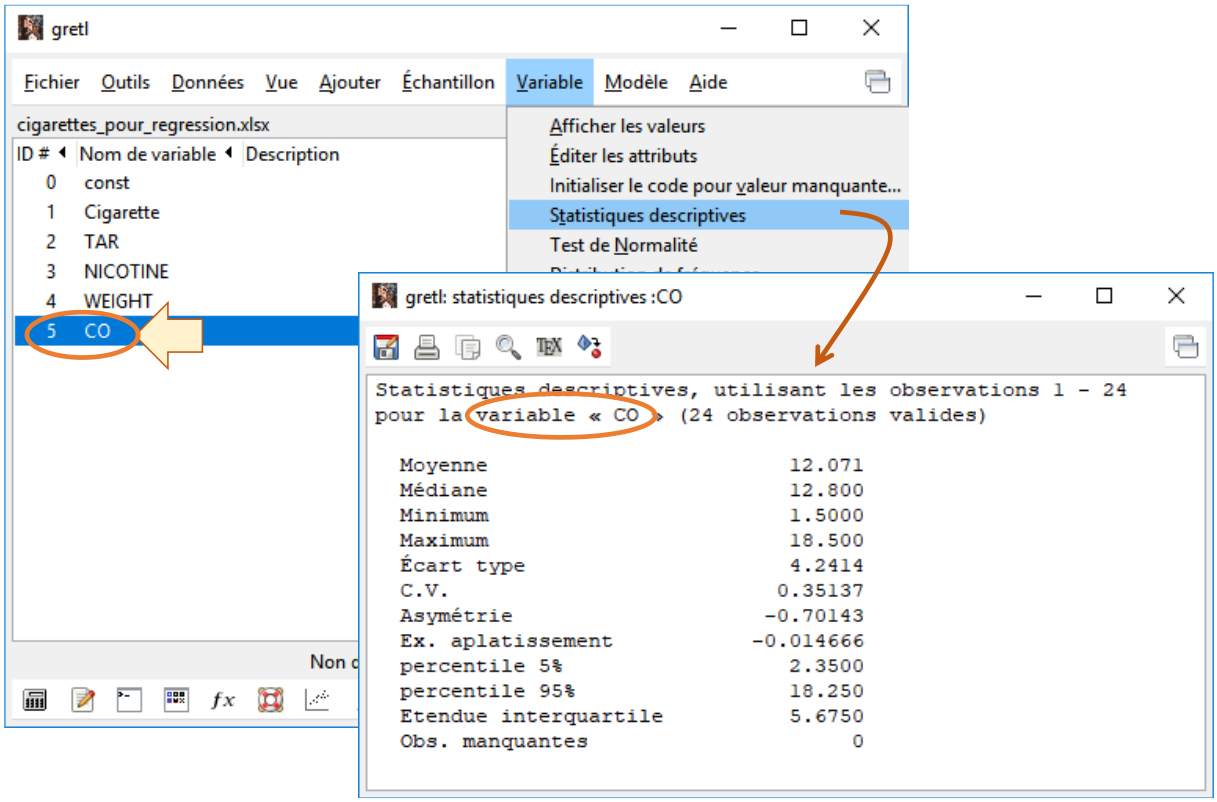
Figure 1 - Tableau de données ''cigarettes''

3.2 Description d'une variable

Le menu **VARIABLE** permet d'accéder à un ensemble de techniques statistiques applicables à une variable, celle qui est actuellement sélectionnée dans la fenêtre principale. Dans cette section, nous avons choisi de traiter l'endogène CO.

3.2.1 Statistiques descriptives

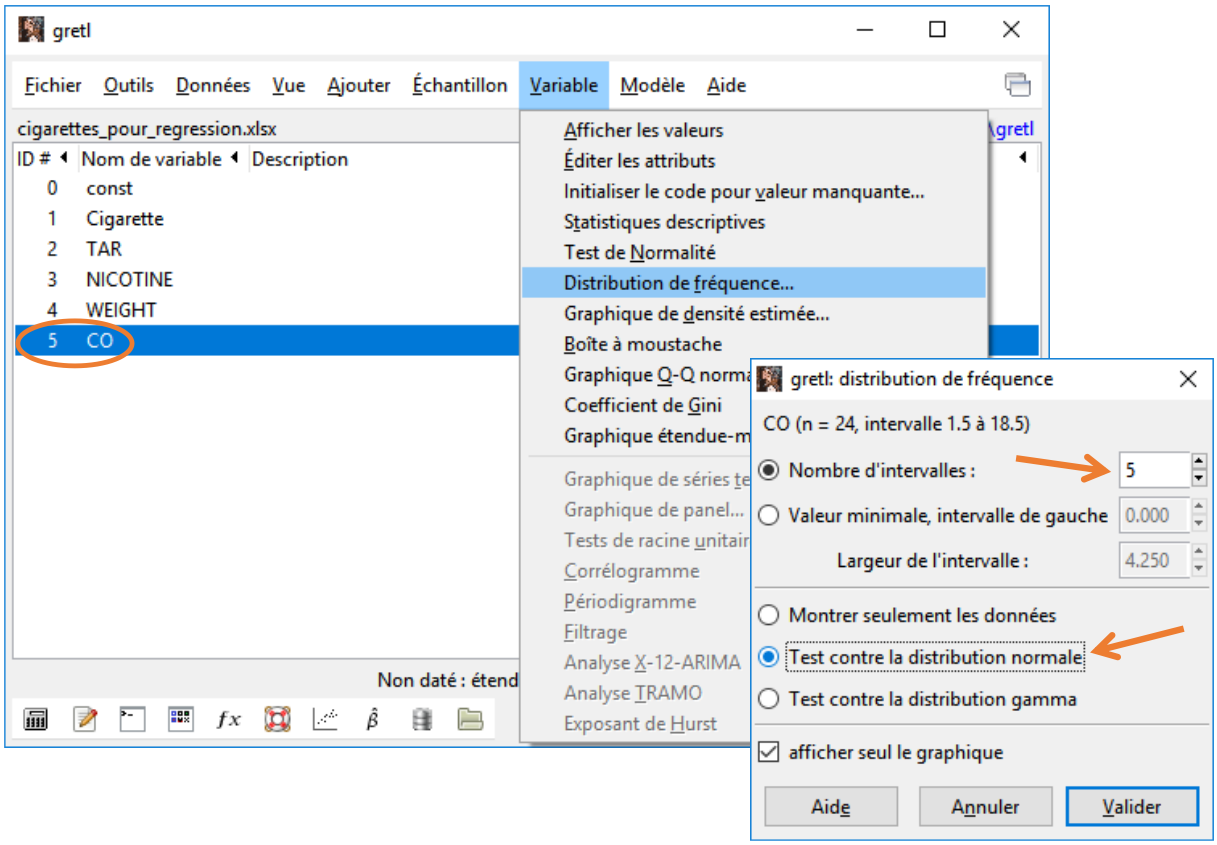
Nous accédons aux statistiques descriptives de CO en cliquant sur le menu **VARIABLE / STATISTIQUES DESCRIPTIVES**.



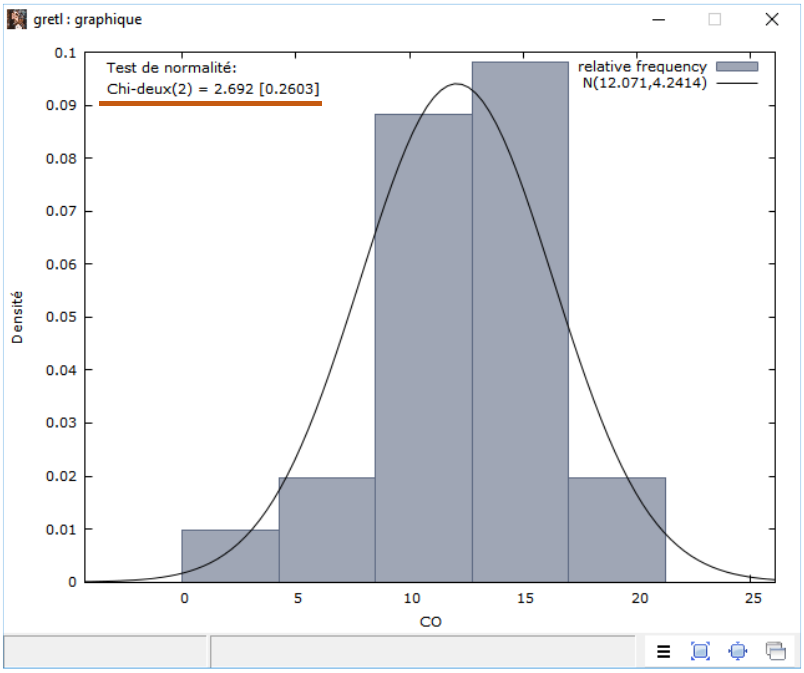
Nous disposons d'une série d'indicateurs : la moyenne, la médiane, etc.

3.2.2 Distribution de fréquences

VARIABLE / DISTRIBUTION DE FREQUENCE produit l'histogramme de fréquence. Le nombre de d'intervalles est paramétrable. Nous pouvons confronter la distribution empirique avec la distribution de la loi normale. Une courbe de densité est affichée en sortie dans ce cas. Un test de normalité est effectué. Il s'agit du test de Doornik-Hansen, qui est une version améliorée – plus précise sur les petits échantillons – du test de Jarque-Bera ("Tests de normalité", octobre 2011).

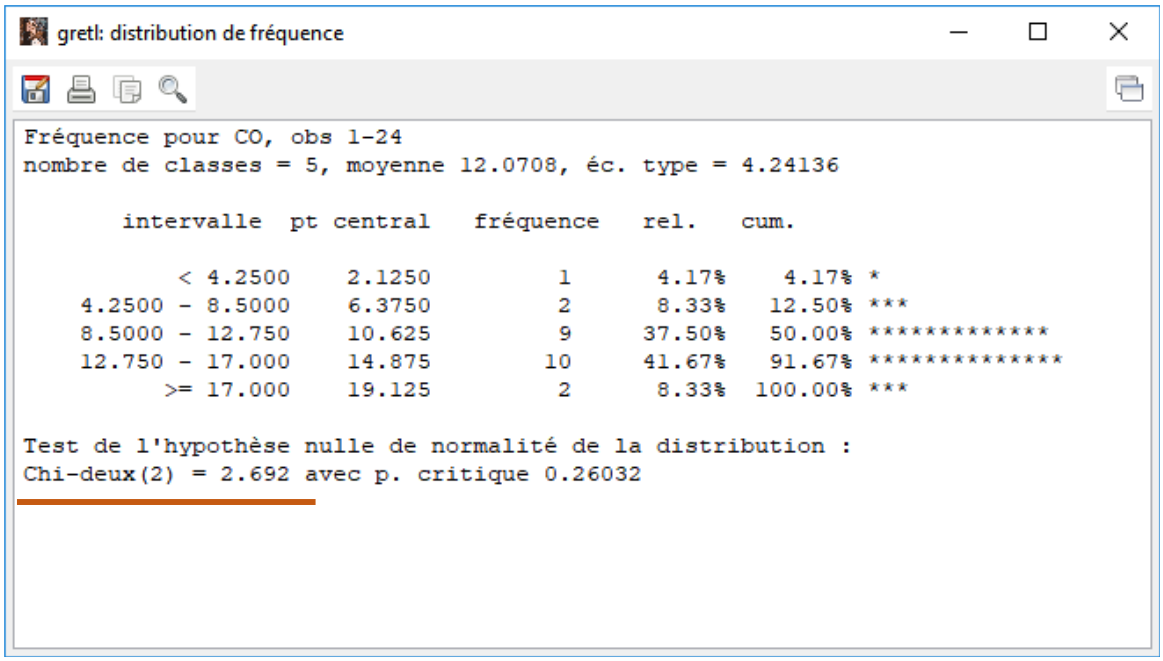


Nous disposons d'une sortie en mode graphique.



Ou en mode texte.

Remarque importante : le résultat du test de normalité ne dépend pas du nombre d'intervalles spécifié lors du paramétrage des calculs.

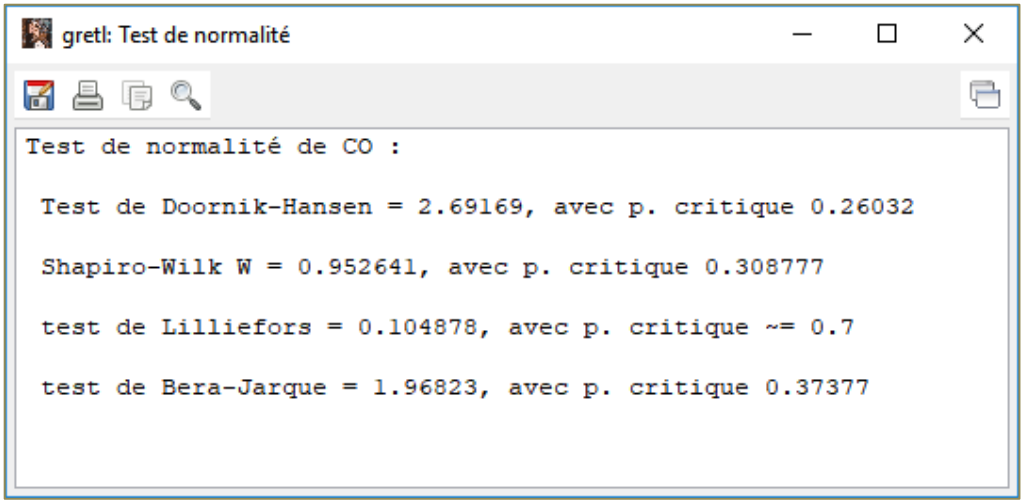


gretl: distribution de fréquence

```
Fréquence pour CO, obs 1-24
nombre de classes = 5, moyenne 12.0708, éc. type = 4.24136

       intervalle  pt central   fréquence   rel.   cum.

           < 4.2500    2.1250         1      4.17%    4.17% *
    4.2500 - 8.5000    6.3750         2      8.33%   12.50% ***
    8.5000 - 12.750    10.625         9     37.50%   50.00% **************
    12.750 - 17.000    14.875        10     41.67%   91.67% ***************
          >= 17.000    19.125         2      8.33%  100.00% ***

Test de l'hypothèse nulle de normalité de la distribution :
Chi-deux(2) = 2.692 avec p. critique 0.26032
```

3.2.3 Tests de normalité

Nous pouvons vérifier plus explicitement la compatibilité avec la distribution normale de la variable CO en activant le menu **VARIABLE / TEST DE NORMALITE**. Plusieurs approches sont proposées. Les procédures de Doornik-Hansen et de Shapiro-Wilk sont conseillées pour les petits effectifs.

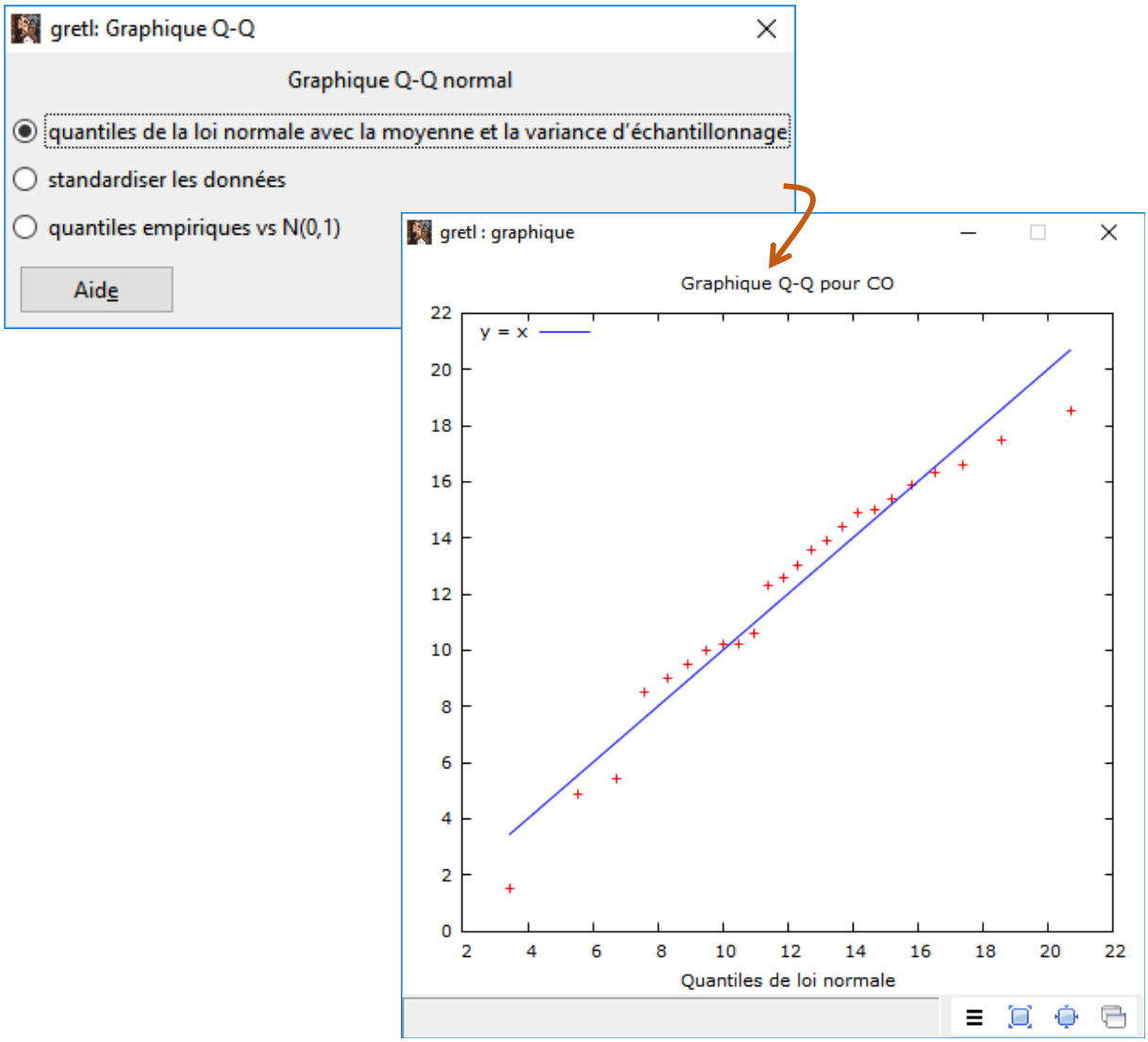


gretl: Test de normalité

```
Test de normalité de CO :

 Test de Doornik-Hansen = 2.69169, avec p. critique 0.26032

 Shapiro-Wilk W = 0.952641, avec p. critique 0.308777

 test de Lilliefors = 0.104878, avec p. critique ~= 0.7

 test de Bera-Jarque = 1.96823, avec p. critique 0.37377
```

3.2.4 Graphique Q-Q normal

Le graphique Q-Q normal est une représentation graphique qui permet de vérifier différemment la compatibilité d'une distribution empirique avec la loi normale. Contrairement aux procédures numériques, nous disposons d'une information visuelle sur la nature des éventuelles disparités constatées. Nous y accédons via le menu **VARIABLE / GRAPHIQUE Q-Q NORMAL**. Nous représentons les valeurs dans les unités originelles pour la définition des axes.



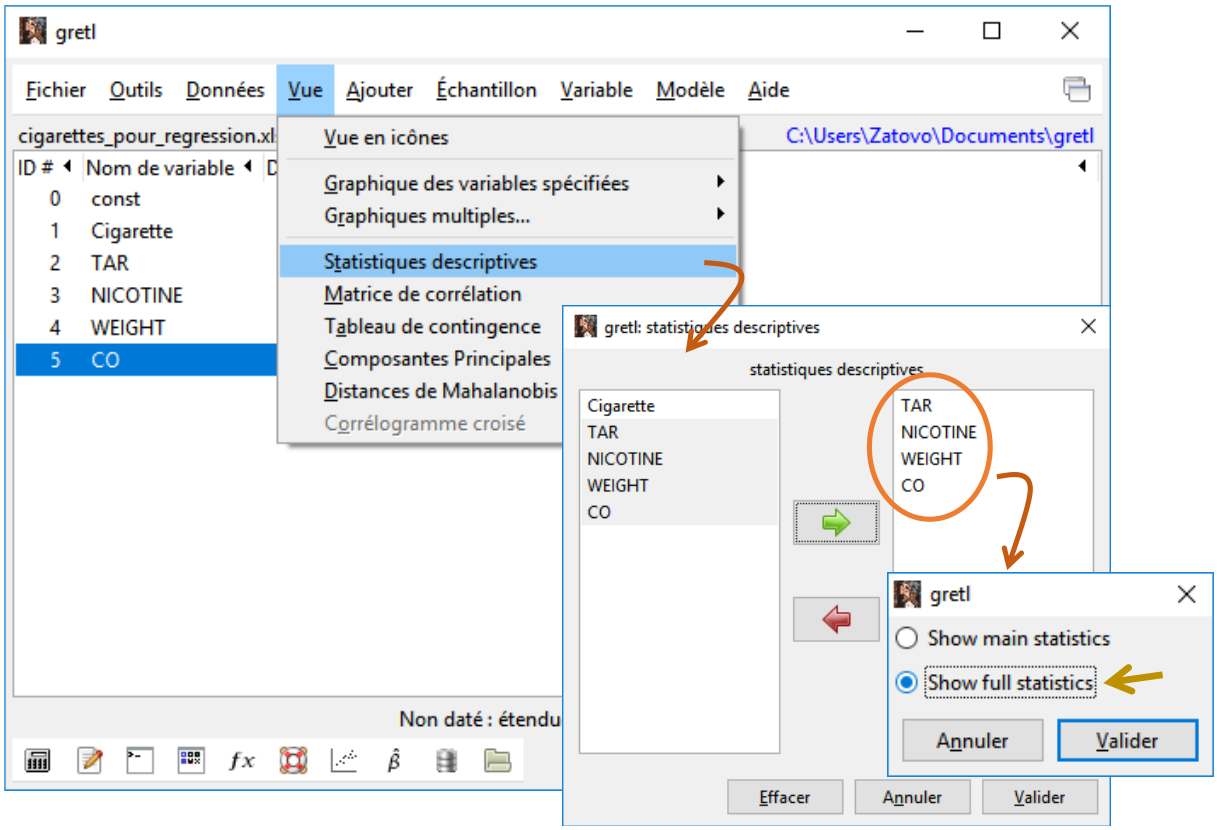
Les points forment plus ou moins une droite, la distribution est compatible avec la loi normale. Il y a quand même des petites divergences vers les queues de distribution.

3.3 Description d'un ensemble de variables

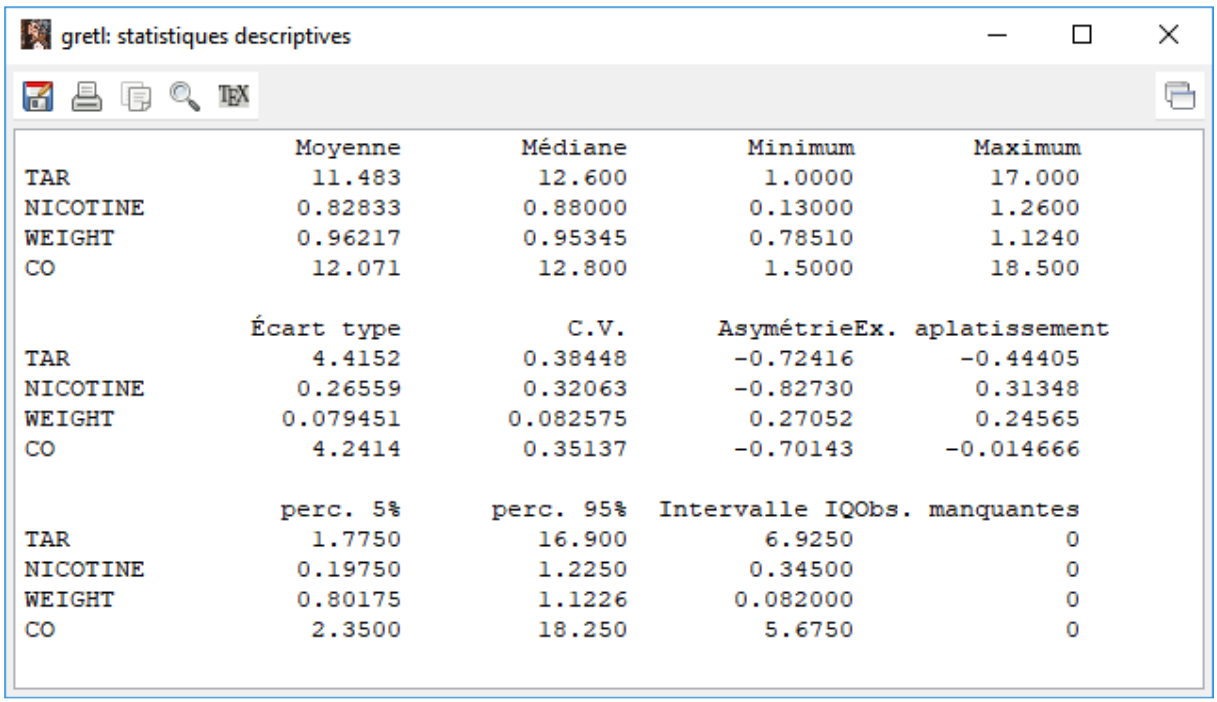
Nous passons par le menu **VUE** lorsque nous souhaitons calculer les indicateurs statistiques pour un ensemble de variables.

3.3.1 Statistiques descriptives

Pour les statistiques descriptives de (TAR, NICOTINE, WEIGHT et CO), nous cliquons sur **VUE / STATISTIQUES DESCRIPTIVES**. Une boîte de dialogue nous permet de préciser les variables à traiter. Nous optons ensuite pour un affichage détaillé.



Pour CO, nous avons exactement les mêmes valeurs que précédemment (section 3.2.1). Nous avons ici, en sus, les résultats pour les autres variables.



gretl: statistiques descriptives

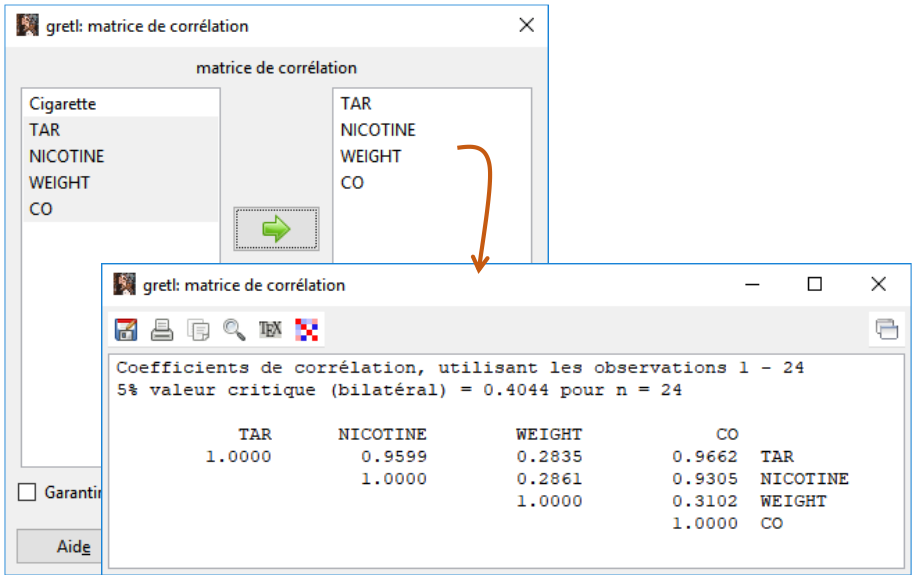
	Moyenne	Médiane	Minimum	Maximum
TAR	11.483	12.600	1.0000	17.000
NICOTINE	0.82833	0.88000	0.13000	1.2600
WEIGHT	0.96217	0.95345	0.78510	1.1240
CO	12.071	12.800	1.5000	18.500

	Écart type	C.V.	AsymétrieEx. aplatissement	
TAR	4.4152	0.38448	-0.72416	-0.44405
NICOTINE	0.26559	0.32063	-0.82730	0.31348
WEIGHT	0.079451	0.082575	0.27052	0.24565
CO	4.2414	0.35137	-0.70143	-0.014666

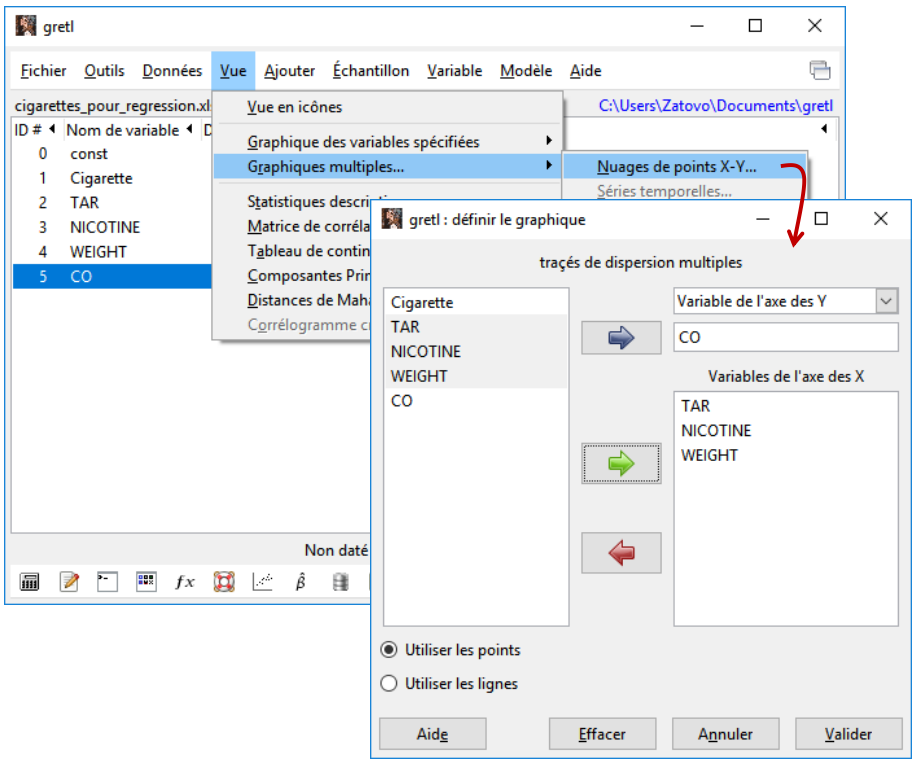
	perc. 5%	perc. 95%	Intervalle IQ	Obs. manquantes
TAR	1.7750	16.900	6.9250	0
NICOTINE	0.19750	1.2250	0.34500	0
WEIGHT	0.80175	1.1226	0.082000	0
CO	2.3500	18.250	5.6750	0

3.3.2 Matrice de corrélation

La matrice de corrélation (menu **VUE / MATRICE DE CORRELATION**) permet d'apprécier d'une part la force de la liaison des exogènes (TAR, NICOTINE, WEIGHT) avec l'endogène (CO), d'autre par la colinéarité entre les exogènes prises deux à deux. Dans notre cas, on s'attend à ce que TAR et NICONTINE soient pertinentes dans la régression linéaire multiple à venir (r_{co_tar} = 0.9662, $r_{co\_nicotine}$ = 0.9305), sachant quand-même qu'elles sont fortement redondantes ($r_{tar\_nicotine}$ = 0.9599).

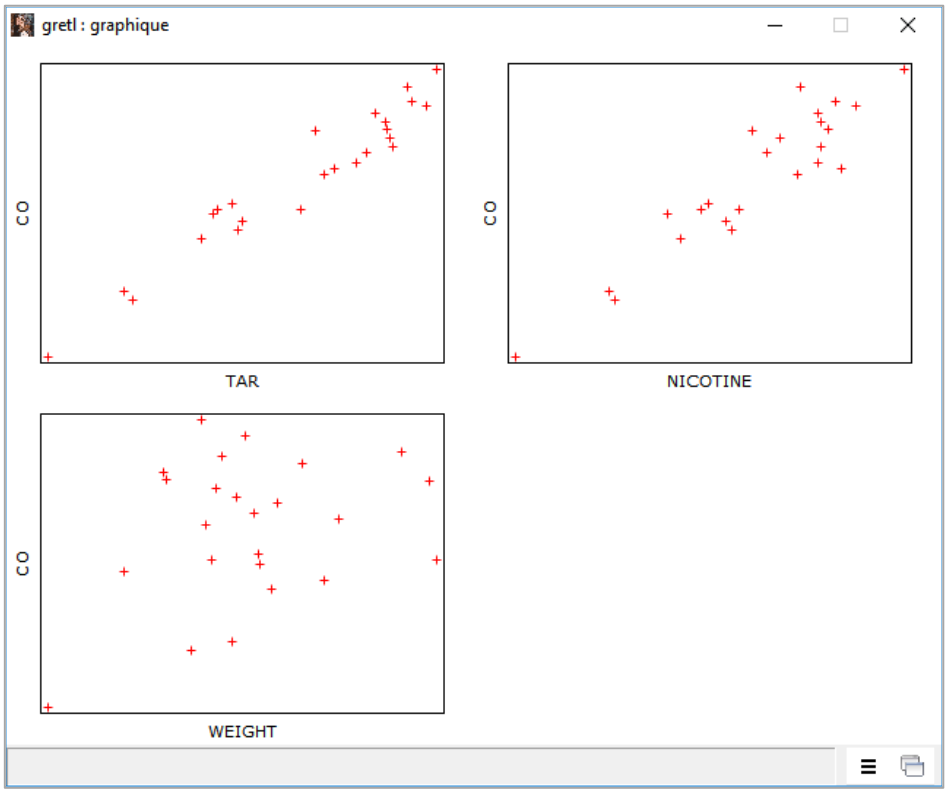


3.3.3 Nuages de points



Un graphique ''nuage de points'' permet d'apprécier visuellement la liaison entre les variables. Il est plus riche qu'un indicateur numérique en donnant des indications sur la forme de la liaison (linéaire ou non) et sur la présence éventuelle de points atypiques (qui peuvent fausser les calculs).

Dans notre exemple, les liaisons de CO avec TAR et NICOTINE sont confirmées. On observe également une marque de cigarette qui est très peu nocive (CO faible par rapport aux autres). En revenant sur le tableau de données (Figure 1), nous constatons qu'il s'agit de la marque NOW. Sans trop nous avancer (on ne va pas la supprimer de la base tout de suite), on peut légitimement penser qu'elle est susceptible de fausser les résultats de la régression à venir. Il faut garder cela à l'esprit.

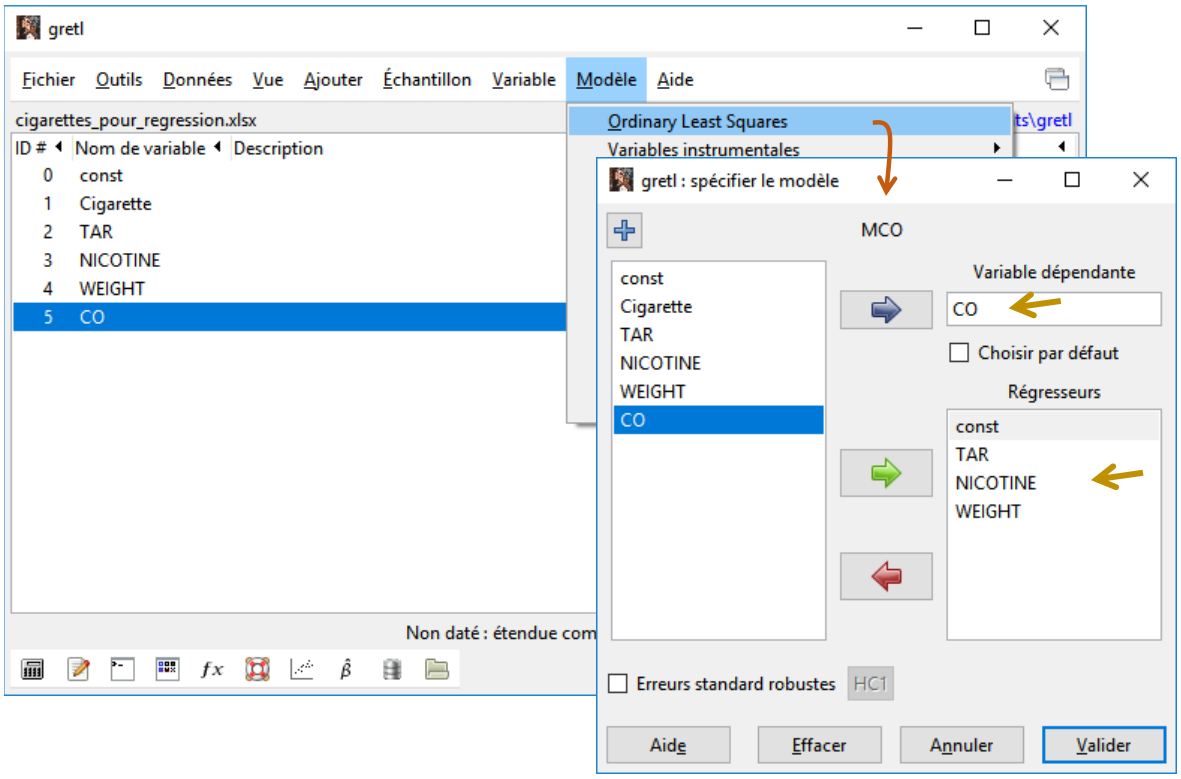


4 Régression et inspection des résultats

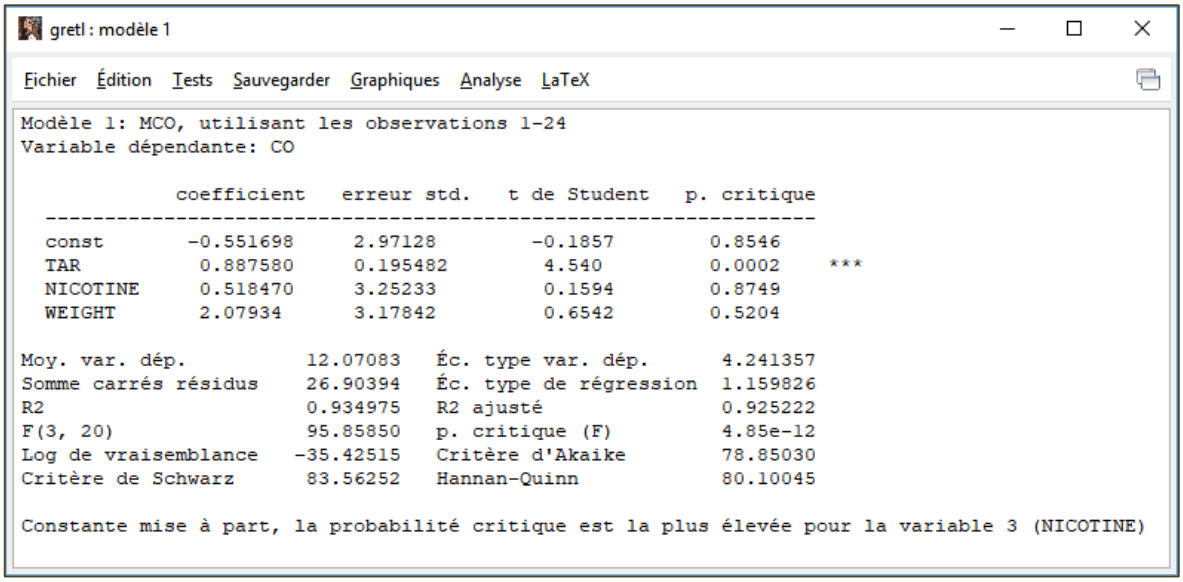
La plupart des concepts et indicateurs de la régression fournis par ''gretl'' sont décrits dans les supports de cours accessibles sur ma page d'économétrie. Sauf rares cas non recensés, je ne reviendrai pas dessus de manière détaillée.

4.1 Moindres carrés ordinaires

Le menu **MODELE / ORDINARY LEAST SQUARES** permet de lancer la régression avec la méthode des moindres carrés ordinaires (MCO). Nous désignons l'endogène (Variable dépendante) et les exogènes (Régresseurs). La constante ''const'' est automatiquement ajoutée, mais nous pouvons la retirer.



Une fenêtre spécifique apparaît. On notera de nouveau items dans la barre de menu (FICHIER, EDITION, **TESTS**, SAUVEGARDER, **GRAPHIQUES**, **ANALYSE**, LATEX).



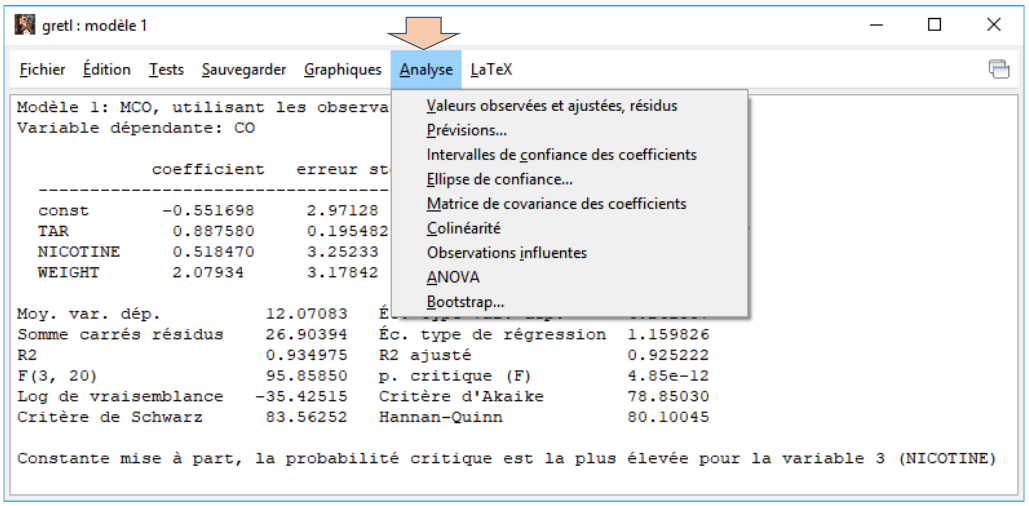
Pour ce qui est des résultats, nous avons les indicateurs usuels que nous connaissons pour la plupart (similaire au summary() de la procédure lm() de R), mis à part peut-être :

- La log-vraisemblance qui est décrite dans le support d'Adkins (novembre 2018, page 277, équation 8.8 ; $\sigma_i = \sigma_\varepsilon$, constant, ne dépend pas de l'observation dans le cas des MCO) ;

– Le critère d'Hannan-Quinn qui relève du même principe que l'AIC ou le BIC, mais est basé sur la log-vraisemblance.

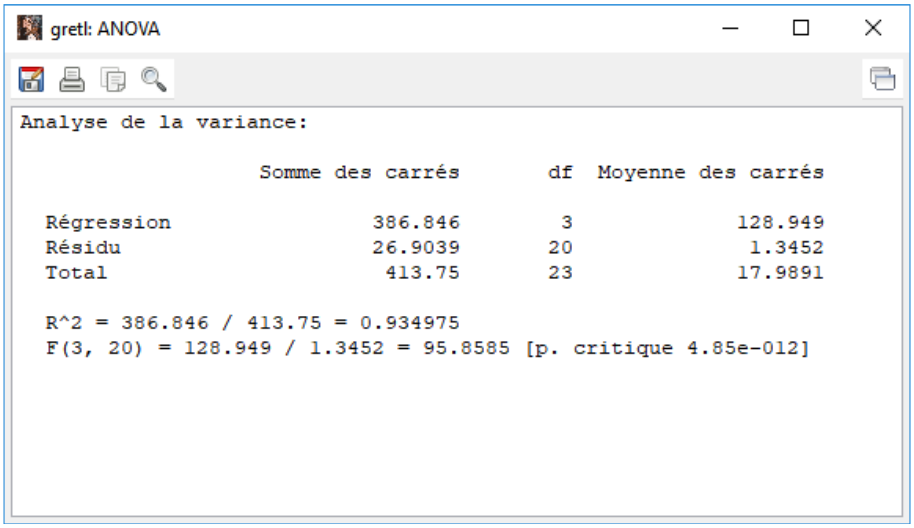
4.2 Analyse

Le menu ANALYSE donne accès à un ensemble de procédures permettant, entres autres, d'évaluer la qualité de la modélisation.



4.2.1 Tableau d'analyse de variance

Avec **ANALYSE / ANOVA**, nous obtenons le tableau d'analyse de variance. Le R2 (coefficient de détermination) et la statistique F du test de significativité globale sont repris ici.



4.2.2 Intervalle de confiance des coefficients

Nous affichons les intervalles de confiance des coefficients à 95% avec **ANALYSES / INTERVALLES DE CONFIANCE DES COEFFICIENTS**.

gretl : intervalles de confiance des coefficients

```
t(20, 0.025) = 2.086

      VARIABLE          COEFFICIENT      INTERVALLE DE CONF. A 95%

         const           -0.551698        -6.74968       5.64629
           TAR            0.887580        0.479813       1.29535
      NICOTINE            0.518470        -6.26577       7.30271
        WEIGHT            2.07934         -4.55072       8.70941
```

4.2.3 Détection de la colinéarité

Avec **ANALYSE / COLINEARITE**, nous disposons des informations sur le degré de redondance entre les exogènes. Un VIF (facteur d'inflation de la variance) supérieur à 10 est suspect. TAR et NICOTINE sont fortement liées, confirmant ce que nous avions constaté en calculant la matrice de corrélation.

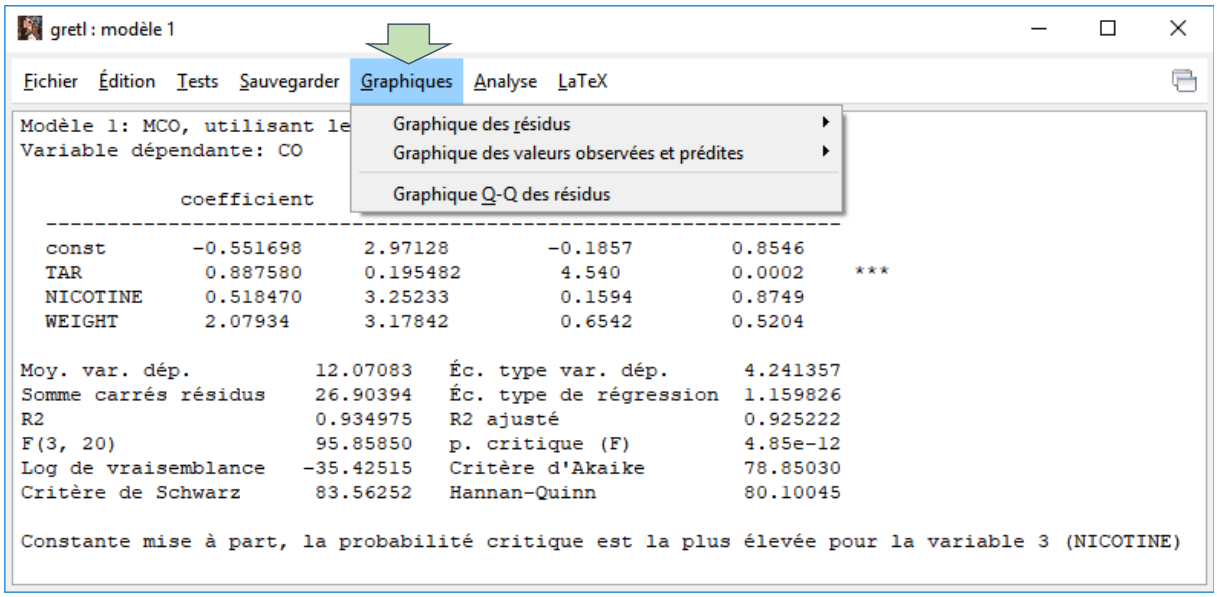
gretl : colinéarité

```
Facteurs d'inflation de variance
Valeur minimale possible = 1.0
Valeurs > 10.0 peut indiquer un problème de colinéarité

         TAR   12.736
    NICOTINE   12.757
      WEIGHT    1.090

VIF(j) = 1/(1 - R(j)^2), où R(j) est un coefficient de corrélation multiple
entre la variable j et les autres variables indépendantes

Belsley-Kuh-Welsch collinearity diagnostics:

                        --- variance proportions ---
    lambda      cond     const       TAR  NICOTINE    WEIGHT
     3.890     1.000     0.000     0.001     0.000     0.000
     0.103     6.152     0.017     0.033     0.011     0.013
     0.004    30.453     0.008     0.902     0.946     0.060
     0.003    35.389     0.975     0.065     0.042     0.927

  lambda = eigenvalues of X'X, largest to smallest
  cond   = condition index
  note: variance proportions columns sum to 1.0
```

Le diagnostic de Belsley-Kuh-Welsch s'appuie sur la décomposition de la matrice (X'X). L'interprétation n'est pas triviale. Il faut lire conjointement ''condition index'' et ''variance proportions'' des variables. Fort heureusement, un document accessible librement en ligne explicite la démarche et les formules (Adkins, Waters et Hill, juillet 2015).

4.3 Graphiques des résidus

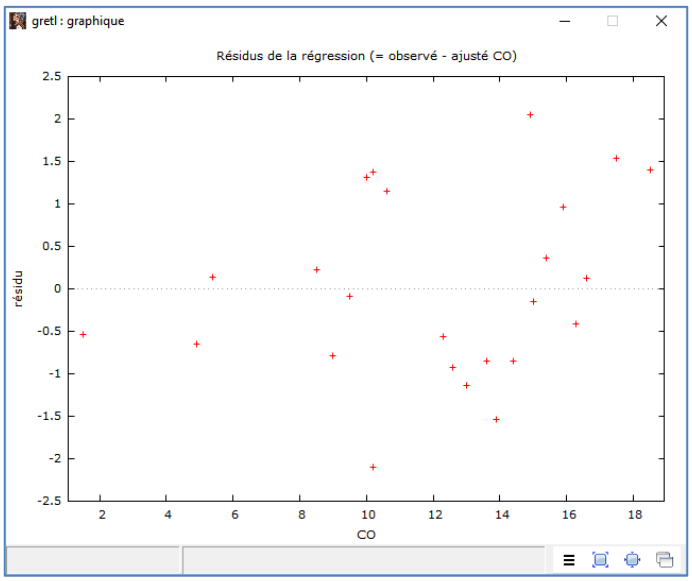
Le menu **GRAPHIQUES** nous permet d'inspecter les résidus.



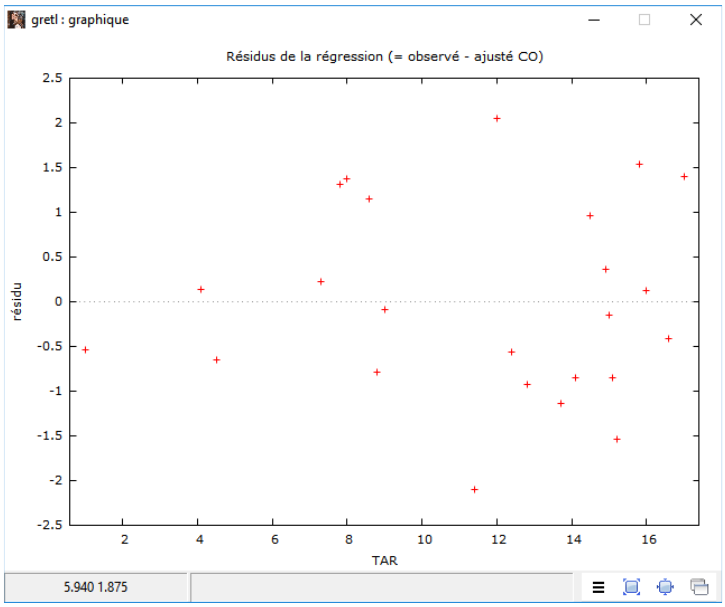
4.3.1 Résidus

Dans le graphique des résidus, les erreurs observées sont toujours en ordonnée. Tout dépend ensuite de ce que nous mettons en abscisse.

Nous faisons le choix de l'endogène CO dans un premier temps (menu **GRAPHIQUES / GRAPHIQUES DES RESIDUS / PAR RAPPORT A CO**). Nous observons ainsi s'il y a des plages de valeurs de l'endogène mal modélisées, ou encore s'il y a des points extrêmes.



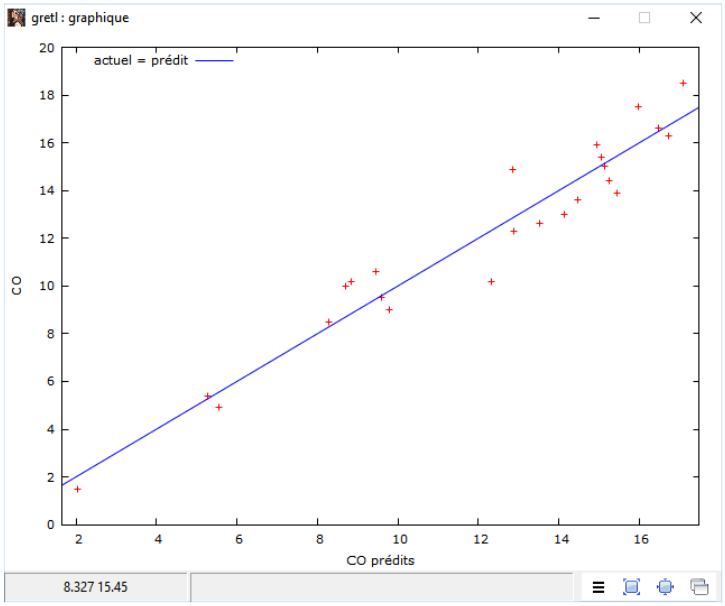
Puis nous construisons le graphique des résidus mais avec TAR en abscisse cette fois-ci (menu **GRAPHIQUES / GRAPHIQUES DES RESIDUS / PAR RAPPORT A TAR**).



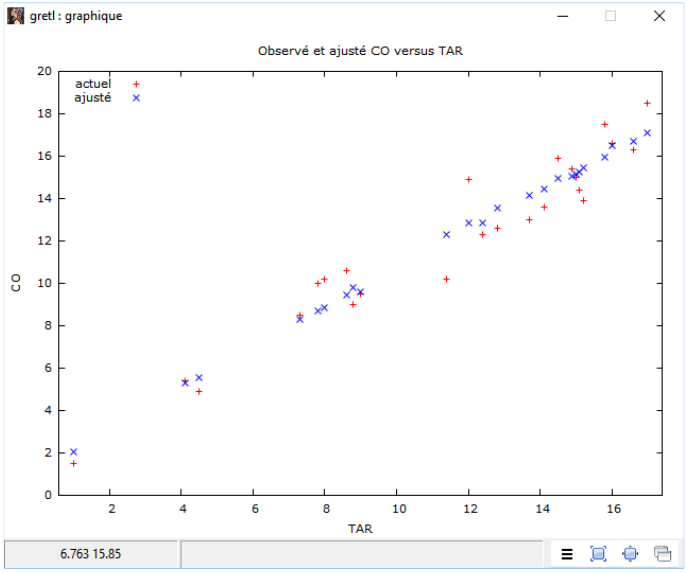
Dans ce cas, nous essayons de voir s'il y a des plages de valeurs de TAR qui entraînent une mauvaise modélisation. La nuance est importante. On peut se servir également de ce graphique pour identifier les éventuels phénomènes d'hétéroscédasticité (est-ce que la dispersion des résidus est variable en fonction des valeurs de TAR).

4.3.2 Valeurs observés vs. prédites de l'endogène

Les graphiques des valeurs observées et prédites mettent en relation y_i et $\hat{y}_i$. Si la modélisation est parfaite, nous avons une droite (menu **GRAPHIQUES / GRAPHIQUES DES VALEURS OBSERVEES ET PREDITES / OBSERVE ET AJUSTE**). Comme pour le graphique des résidus, l'objectif encore une fois est d'identifier des zones de Y qui seraient mal ajustées.



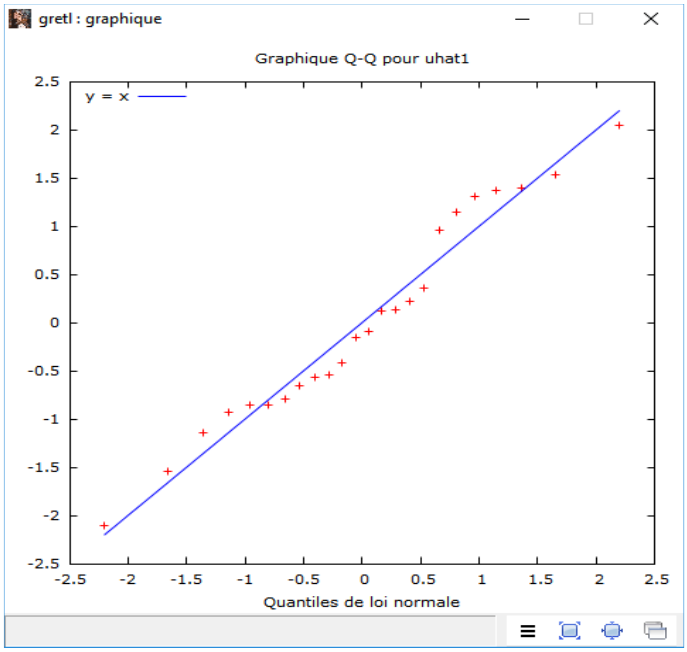
Nous pouvons produire la même opposition mais relativement aux valeurs d'une des exogènes (menu **GRAPHIQUES / GRAPHIQUES DES VALEURS OBSERVEES ET PREDITES / PAR RAPPORT A TAR**)



Nous constatons que la qualité de la modélisation est homogène relativement aux valeurs de TAR, qui est la seule variable pertinente du modèle rappelons-le.

4.3.3 Normalité des résidus – QQ-plot

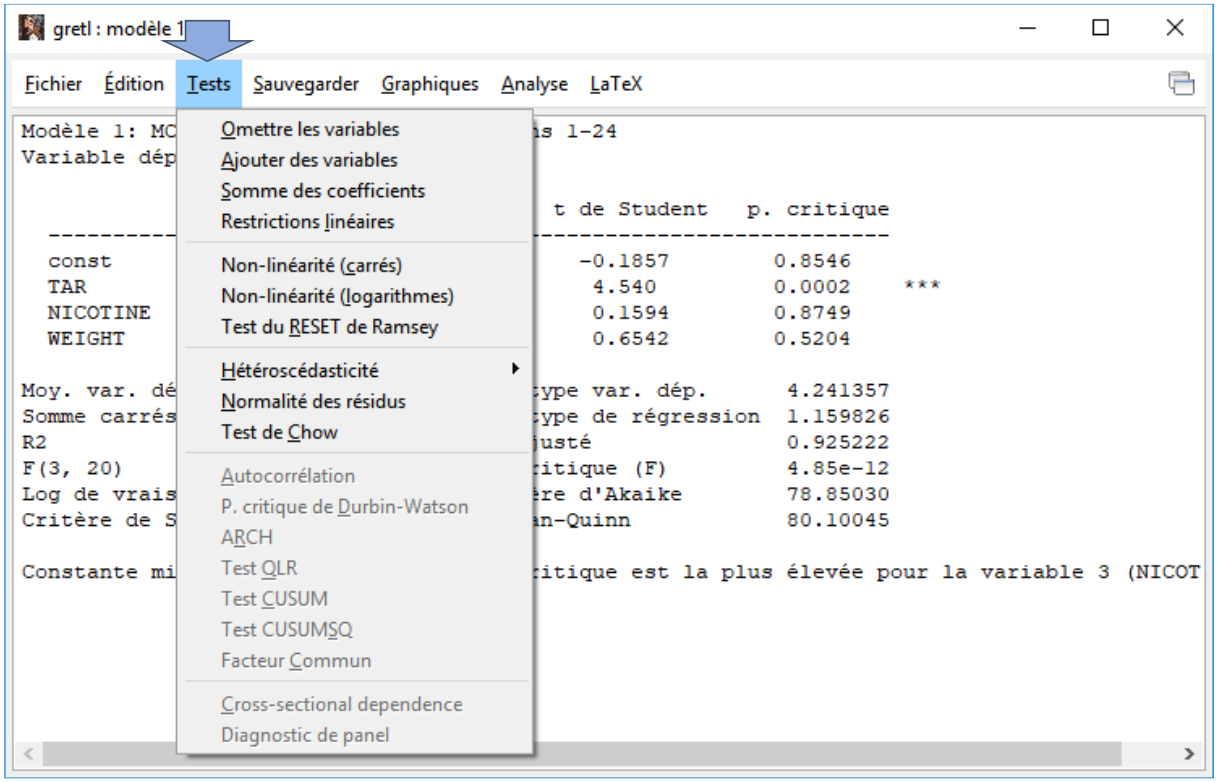
La normalité des erreurs est une des hypothèses essentielles des MCO. Vérifier si les résidus (les erreurs observées) sont compatibles avec ce prérequis est important. Nous le réalisons avec le graphique QQNORM (menu **GRAPHIQUES / GRAPHIQUE Q-Q DES RESIDUS**).



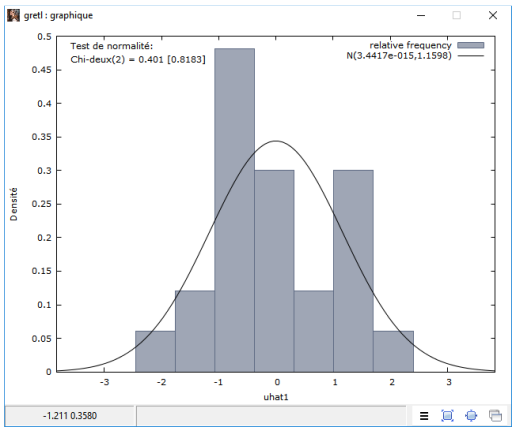
Les points forment une droite, le postulat de normalité ne peut être rejeté.

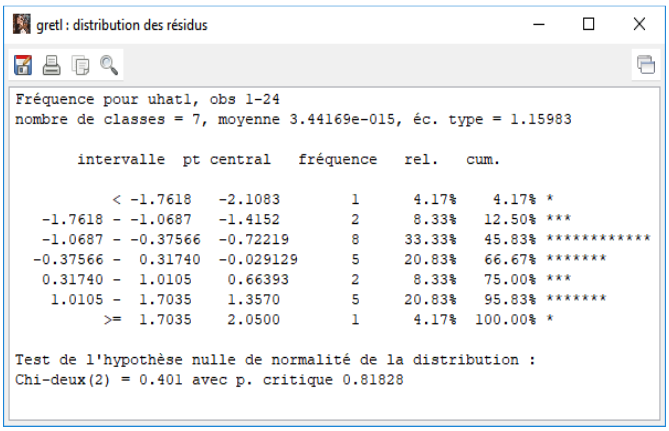
4.3.4 Normalité des résidus – Test

Le même test de normalité des résidus peut être réalisé à l'aide de procédure numériques (menu **TESTS / NORMALITE DES RESIDUS**).



Gretl affiche l'histogramme de fréquence et le résultat du test de Doornik-Hansen (section 3.2.2). La normalité des erreurs ne peut pas être remise en cause pour notre régression.

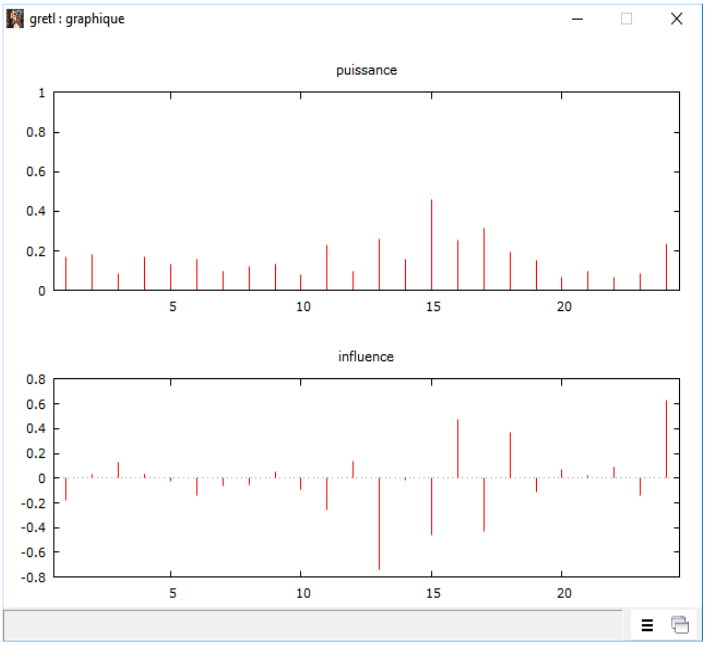


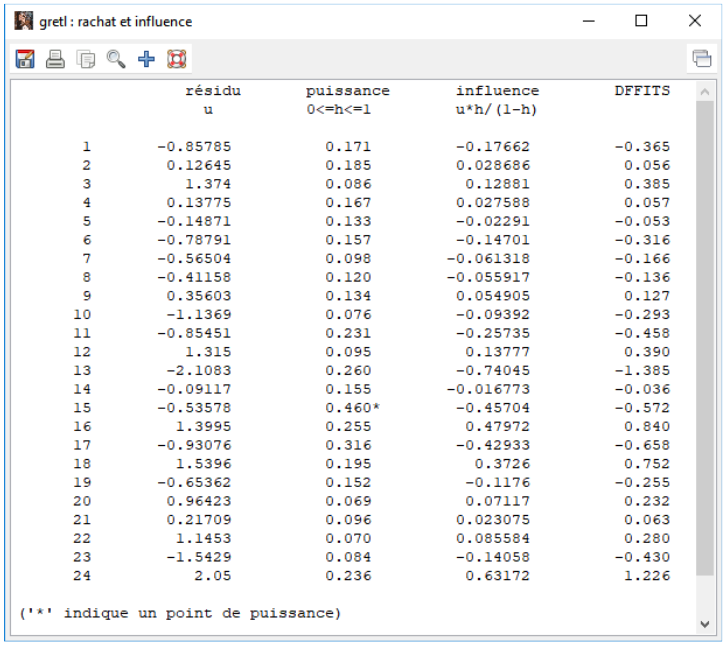


4.4 Détection des points atypiques

Avec le menu **ANALYSE / OBSERVATIONS INFLUENTES**, nous pouvons identifier les observations atypiques (qui s'écartent fortement des autres) et influentes (si nous les retirons, le résultat de la régression serait passablement différent).

Nous disposons sous forme graphique et numérique : du levier (puissance dans la terminologie gretl), de quelque chose qui s'apparente au résidu studentisé (influence, mais la formule et les valeurs sont différentes), et du DFFITS.





	résidu u	puissance 0<=h<=1	influence u*h/(1-h)	DFFITS
1	-0.85785	0.171	-0.17662	-0.365
2	0.12645	0.185	0.028686	0.056
3	1.374	0.086	0.12881	0.385
4	0.13775	0.167	0.027588	0.057
5	-0.14871	0.133	-0.02291	-0.053
6	-0.78791	0.157	-0.14701	-0.316
7	-0.56504	0.098	-0.061318	-0.166
8	-0.41158	0.120	-0.055917	-0.136
9	0.35603	0.134	0.054905	0.127
10	-1.1369	0.076	-0.09392	-0.293
11	-0.85451	0.231	-0.25735	-0.458
12	1.315	0.095	0.13777	0.390
13	-2.1083	0.260	-0.74045	-1.385
14	-0.09117	0.155	-0.016773	-0.036
15	-0.53578	0.460*	-0.45704	-0.572
16	1.3995	0.255	0.47972	0.840
17	-0.93076	0.316	-0.42933	-0.658
18	1.5396	0.195	0.3726	0.752
19	-0.65362	0.152	-0.1176	-0.255
20	0.96423	0.069	0.07117	0.232
21	0.21709	0.096	0.023075	0.063
22	1.1453	0.070	0.085584	0.280
23	-1.5429	0.084	-0.14058	-0.430
24	2.05	0.236	0.63172	1.226

('*' indique un point de puissance)

La marque de cigarette n°15 (Now) est différente des autres (puissance élevée, on s'en doutait dès les premiers graphiques) ; la n°13 (MultiFilter) est mal modélisée (influence élevée en valeur absolue) c.-à-d. la CO ajustée par le modèle (12.3) est largement supérieure à la CO observée (10.2), d'où un résidu négatif de **-2.1083**.

4.5 Tests généralisés

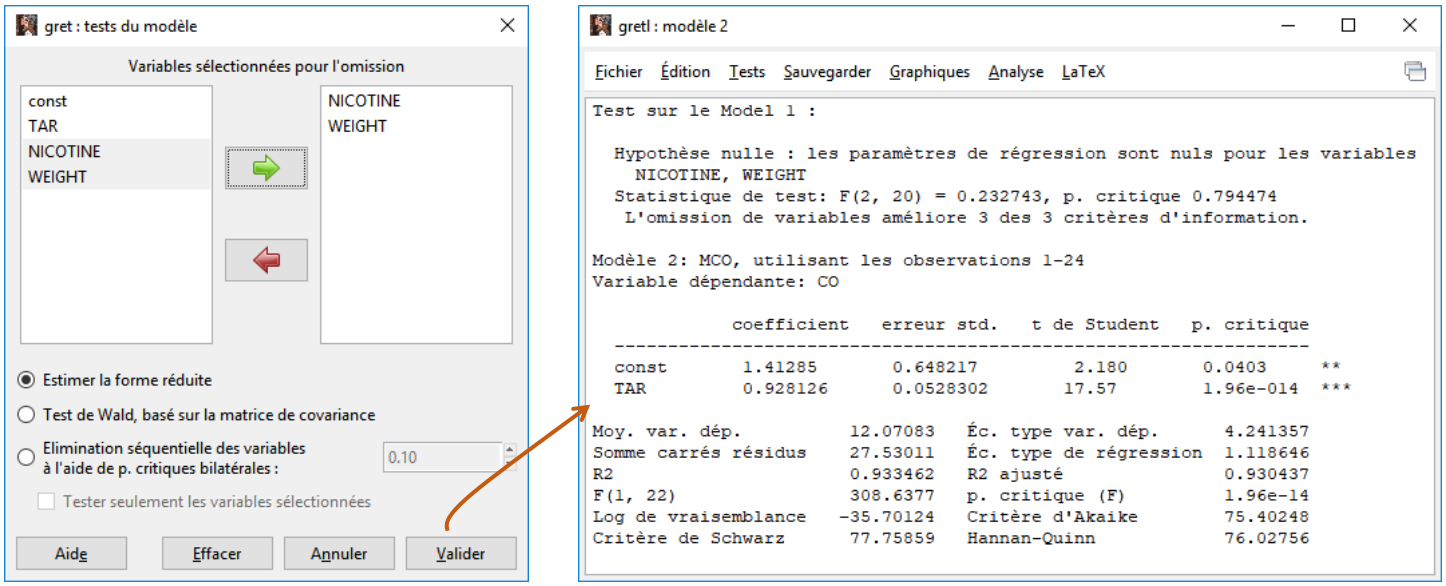
Des outils pour réaliser des tests avancés sur les coefficients sont proposés. J'adopte la notation de Gretl pour désigner notre régression :

CO = b1 + b2 x TAR + b3 x NICOTINE + b4 x WEIGHT

4.5.1 Test de nullité d'un groupe de coefficients

Le premier test consiste à vérifier la nullité simultanée des coefficients (b3 et b4) en opposant notre première régression avec l'ensemble des exogènes et une seconde où seule TAR serait présente. Formellement, nous vérifions si l'hypothèse (H0 : b3 = b4 = 0) est contredite ou non par les données.

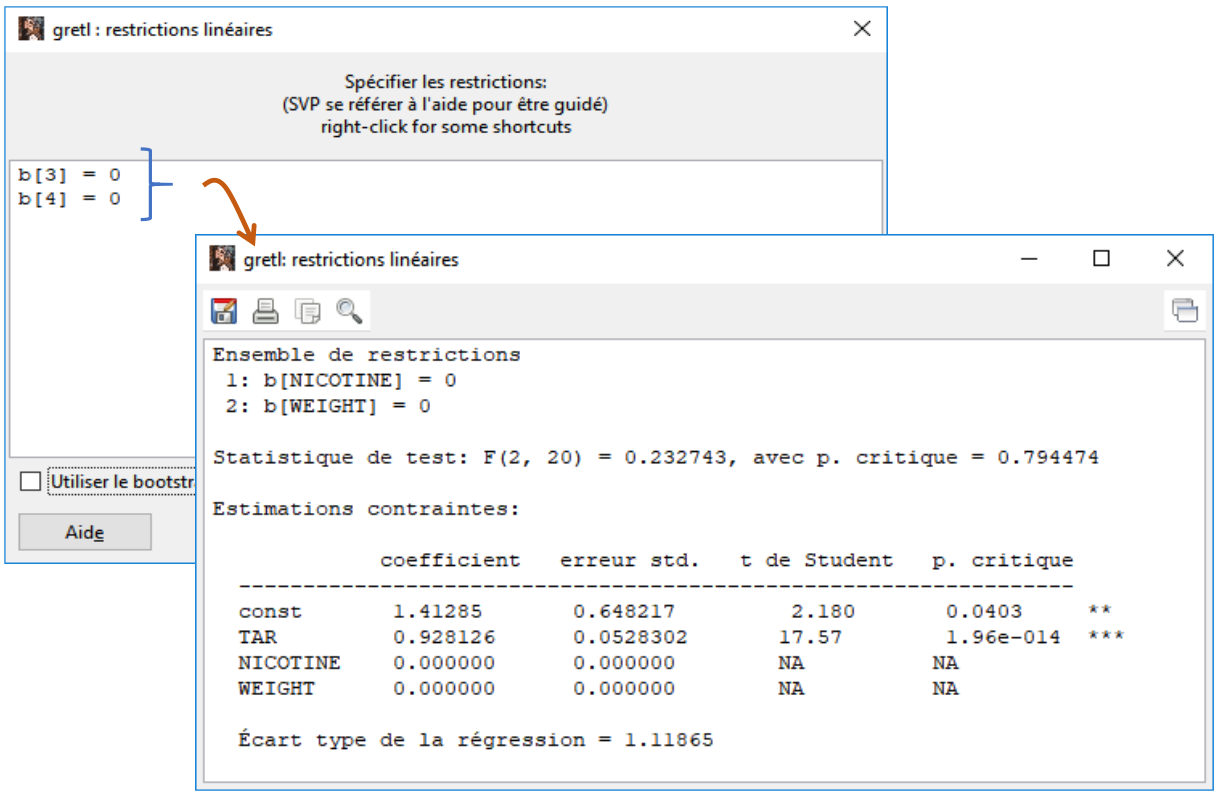
Nous actionnons le menu **TESTS / OMETTRE LES VARIABLES**. Dans la boîte de paramétrage, nous désignons les variables à retirer, à savoir NICOTINE et WEIGHT. Gretl réalise la régression sans ces variables et oppose les coefficients de détermination entre les deux régressions (cf. support de cours, section 10.4).



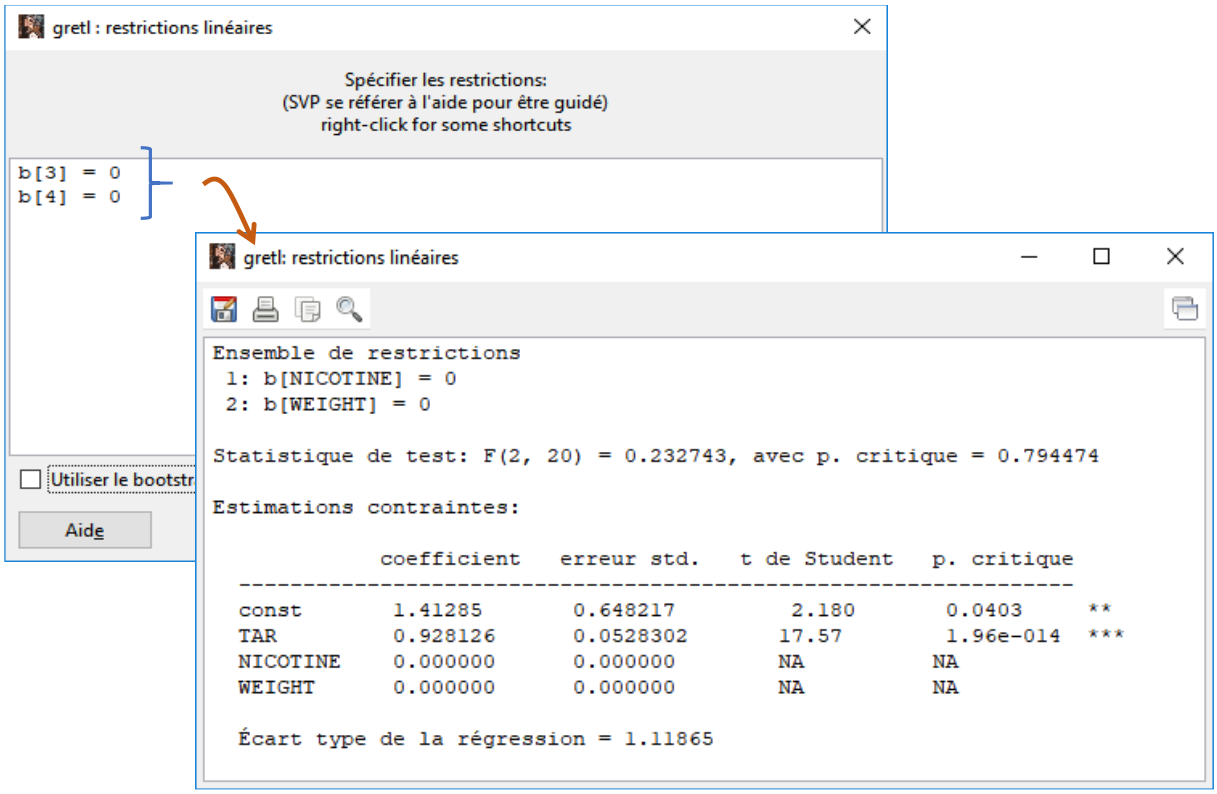
La statistique de test F = 0.232743 suit une loi de Fisher à (2, 20) degrés de liberté. Il n'est pas possible de rejeter l'hypothèse nulle.

Pour corroborer cette conclusion, gretl nous indique que les 3 critères d'information (Schwarz, Akaike et Hannan-Quinn) sont également améliorés (prennent des valeurs plus faibles) sur la seconde régression avec TAR comme seule exogène.

4.5.2 Restrictions linéaires sur les coefficients



Nous pouvons réaliser le même test en l'exprimant sous forme de contraintes linéaires sur les coefficients (cf. support de cours, section 11.3). Avec le menu **TESTS / RESTRICTIONS LINEAIRES**, nous indiquons explicitement les contraintes sur les coefficients (attention sur les indices du vecteur des coefficients b).



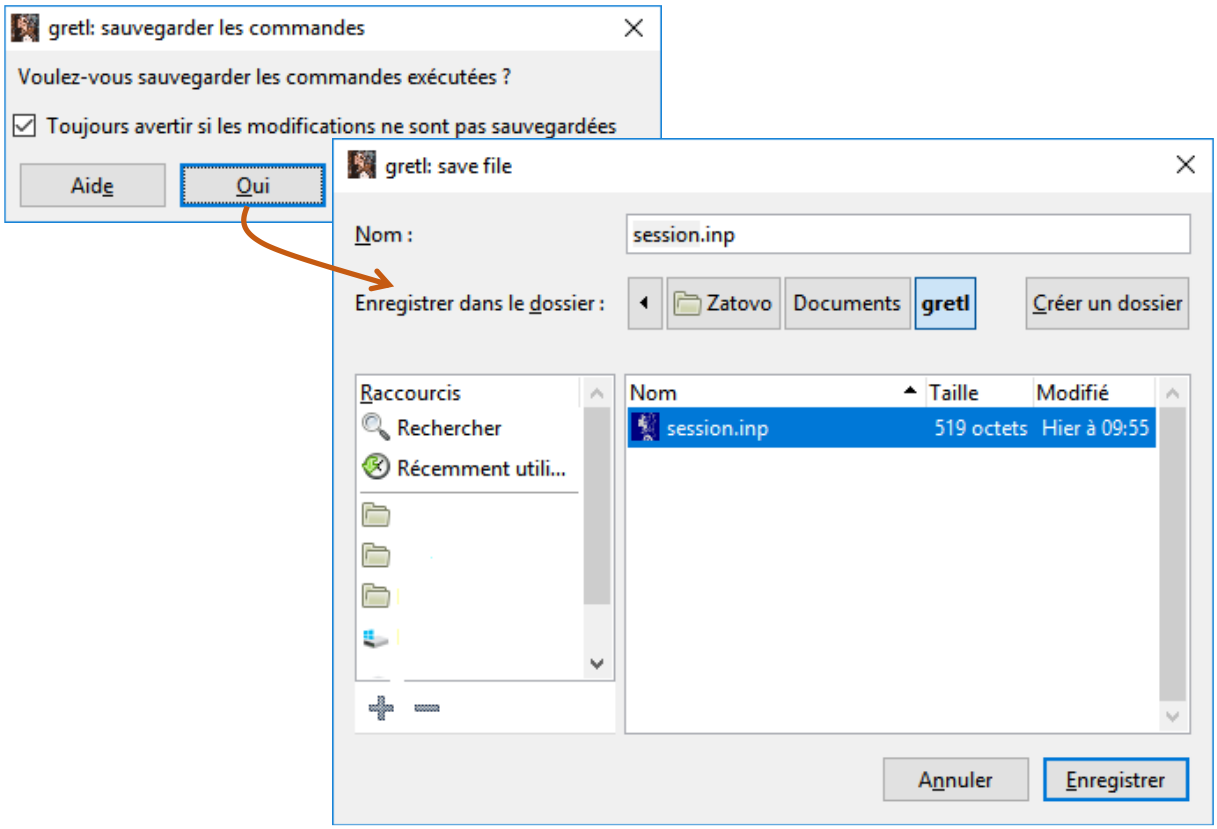
Les résultats (F et p-value) sont bien évidemment identiques à ceux de la stratégie précédente.

5 Fichier de script

Nous pouvons programmer avec le langage ''hansl'' sous Gretl. En lisant attentivement la documentation (Cottrell et Lucchetti, janvier 2019), je constate qu'il est complet et intègre toutes les structures utiles pour la programmation statistique. Après, s'investir dans l'apprentissage d'un nouveau langage pour un outil aussi spécifique mérite réflexion.

J'y vois quand même un avantage certain sur au moins un des aspects de l'utilisation du logiciel, la retranscription des actions sous la forme d'un script de commandes. Ainsi, si nous souhaitons reproduire les mêmes actions sur une version mise à jour des données, il suffit de réexécuter les instructions dans l'ordre sans avoir à se préoccuper des menus à actionner et se rappeler des paramètres que l'on a pu spécifier dans telle ou telle boîte de dialogue.

A la fermeture du logiciel, gretl nous demande si nous souhaitons sauvegarder nos commandes.

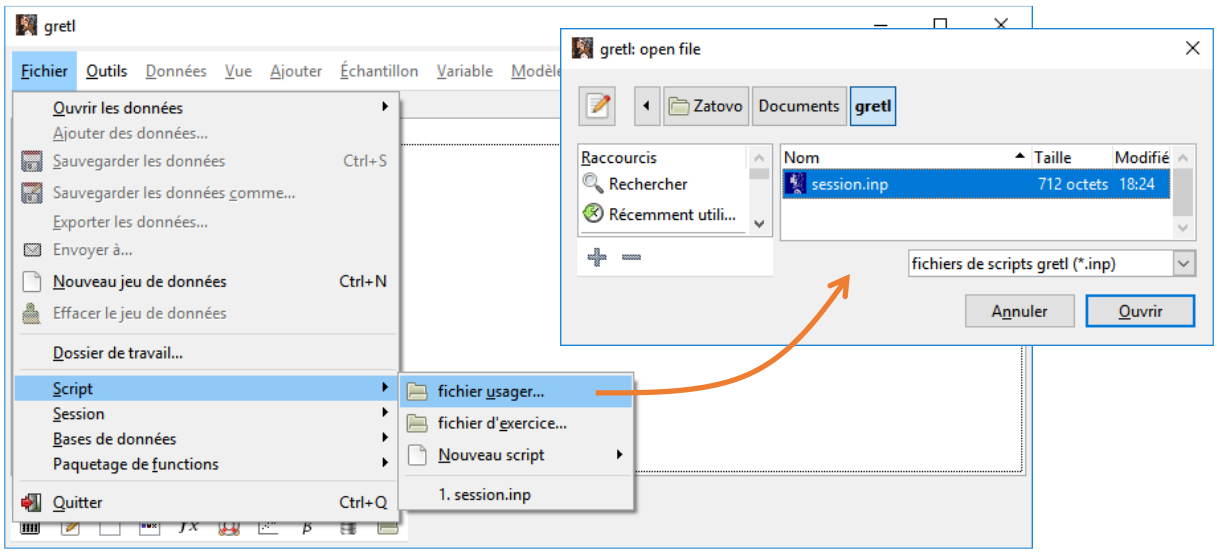


J'ai fait OUI, et j'ai indiqué l'emplacement du fichier "**session.inp**" qui est un fichier texte que l'on peut ouvrir dans un éditeur quelconque. Nous avons ceci pour notre analyse.

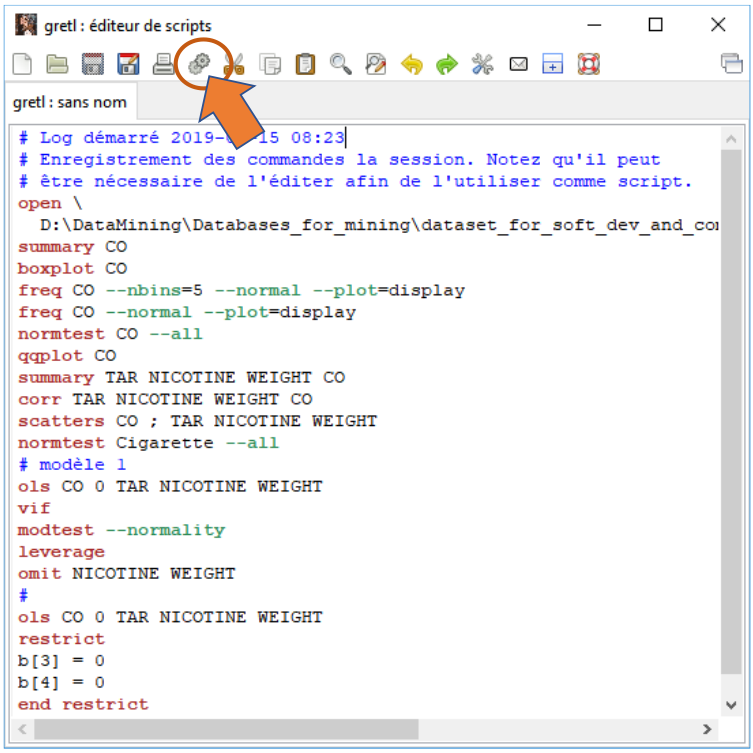
```
# Log démarré 2019-05-15 08:23
# Enregistrement des commandes la session. Notez qu'il peut
# être nécessaire de l'éditer afin de l'utiliser comme script.
open \
  D:\gretl\cigarettes_pour_regression.xlsx
summary CO
boxplot CO
freq CO --nbins=5 --normal --plot=display
freq CO --normal --plot=display
normtest CO --all
qqplot CO
summary TAR NICOTINE WEIGHT CO
corr TAR NICOTINE WEIGHT CO
scatters CO ; TAR NICOTINE WEIGHT
normtest Cigarette --all
# modèle 1
ols CO 0 TAR NICOTINE WEIGHT
vif
modtest --normality
leverage
omit NICOTINE WEIGHT
```

```
#
ols CO 0 TAR NICOTINE WEIGHT
restrict
b[3] = 0
b[4] = 0
end restrict
```

Au redémarrage du logiciel, j'a actionné le menu **FICHIER / SCRIPT / FICHIER USAGER** et j'ai désigné le fichier ''**session.inp**'' à charger.



Une nouvelle fenêtre avec le code apparaît.



Nous pouvons réexécuter la totalité des traitements en cliquant sur l'icône ''engrenage'' (cf. copie d'écran ci-dessus) situé dans la barre d'outils. Plusieurs fenêtres vont surgir, l'une d'entre elles regroupe toutes les sorties textuelles. Voici son contenu :

```
gretl version 2019a
Session courante: 2019-05-15 18:42

# Log démarré 2019-05-15 08:23
# Enregistrement des commandes la session. Notez qu'il peut
# être nécessaire de l'éditer afin de l'utiliser comme script.
? open \
  D:\gretl\cigarettes_pour_regression.xlsx

1 de feuilles valides trouvées
5 variables et 24 observations trouvées
variable 1 (Cigarette): non-numeric values = 24 (100.00 percent)
allocating string table
Table des chaînes de codes écrite dans
 C:\Users\Zatovo\Documents\gretl\string_table.txt
Liste des 6 variables :
  0) const        1) Cigarette    2) TAR          3) NICOTINE
  4) WEIGHT       5) CO

? summary CO

Statistiques descriptives, utilisant les observations 1 - 24
pour la variable « CO » (24 observations valides)

  Moyenne                        12.071
  Médiane                        12.800
  Minimum                        1.5000
  Maximum                        18.500
  Écart type                     4.2414
  C.V.                          0.35137
  Asymétrie                    -0.70143
  Ex. aplatissement           -0.014666
  percentile 5%                  2.3500
  percentile 95%                 18.250
  Etendue interquartile          5.6750
  Obs. manquantes                     0

? boxplot CO
écrit C:\Users\Zatovo\Documents\gretl\gpttmp01.plt

? freq CO --nbins=5 --normal --plot=display

Fréquence pour CO, obs 1-24
nombre de classes = 5, moyenne 12.0708, éc. type = 4.24136

       intervalle  pt central   fréquence   rel.   cum.

           < 4.2500    2.1250         1      4.17%    4.17% *
    4.2500 - 8.5000    6.3750         2      8.33%   12.50% ***
    8.5000 - 12.750    10.625         9     37.50%   50.00% *************
```

```
    12.750 - 17.000    14.875        10     41.67%   91.67% **************
          >= 17.000    19.125         2      8.33%  100.00% ***

Test de l'hypothèse nulle de normalité de la distribution :
Chi-deux(2) = 2.692 avec p. critique 0.26032

? freq CO --normal --plot=display

Fréquence pour CO, obs 1-24
nombre de classes = 7, moyenne 12.0708, éc. type = 4.24136

       intervalle  pt central   fréquence   rel.   cum.

           < 2.9167    1.5000         1      4.17%    4.17% *
    2.9167 - 5.7500    4.3333         2      8.33%   12.50% ***
    5.7500 - 8.5833    7.1667         1      4.17%   16.67% *
    8.5833 - 11.417    10.000         6     25.00%   41.67% *********
    11.417 - 14.250    12.833         5     20.83%   62.50% *******
    14.250 - 17.083    15.667         7     29.17%   91.67% **********
          >= 17.083    18.500         2      8.33%  100.00% ***

Test de l'hypothèse nulle de normalité de la distribution :
Chi-deux(2) = 2.692 avec p. critique 0.26032

? normtest CO --all
Test de normalité de CO :

 Test de Doornik-Hansen = 2.69169, avec p. critique 0.26032

 Shapiro-Wilk W = 0.952641, avec p. critique 0.308777

 test de Lilliefors = 0.104878, avec p. critique ~= 0.7

 test de Bera-Jarque = 1.96823, avec p. critique 0.37377

? qqplot CO
écrit C:\Users\Zatovo\Documents\gretl\gpttmp02.plt

? summary TAR NICOTINE WEIGHT CO

                  Moyenne        Médiane        Minimum        Maximum
TAR                11.483         12.600         1.0000         17.000
NICOTINE          0.82833        0.88000        0.13000         1.2600
WEIGHT            0.96217        0.95345        0.78510         1.1240
CO                 12.071         12.800         1.5000         18.500

               Écart type           C.V.      AsymétrieEx. aplatissement
TAR                4.4152        0.38448       -0.72416       -0.44405
NICOTINE          0.26559        0.32063       -0.82730        0.31348
WEIGHT           0.079451       0.082575        0.27052        0.24565
CO                 4.2414        0.35137       -0.70143      -0.014666

                 perc. 5%      perc. 95%  Intervalle IQObs. manquantes
TAR                1.7750         16.900         6.9250              0
NICOTINE          0.19750         1.2250        0.34500              0
```

```
WEIGHT            0.80175         1.1226       0.082000              0
CO                 2.3500         18.250         5.6750              0

? corr TAR NICOTINE WEIGHT CO

Coefficients de corrélation, utilisant les observations 1 - 24
5% valeur critique (bilatéral) = 0.4044 pour n = 24

           TAR      NICOTINE        WEIGHT            CO
        1.0000        0.9599        0.2835        0.9662  TAR
                      1.0000        0.2861        0.9305  NICOTINE
                                    1.0000        0.3102  WEIGHT
                                                  1.0000  CO

? scatters CO ; TAR NICOTINE WEIGHT
écrit C:\Users\Zatovo\Documents\gretl\gpttmp03.plt

? normtest Cigarette --all
Test de normalité de Cigarette :

 Test de Doornik-Hansen = 1.93788, avec p. critique 0.379485

 Shapiro-Wilk W = 0.958858, avec p. critique 0.415911

 test de Lilliefors = 0.0733283, avec p. critique ~= 1

 test de Bera-Jarque = 1.45003, avec p. critique 0.484316

# modèle 1
? ols CO 0 TAR NICOTINE WEIGHT

Modèle 1: MCO, utilisant les observations 1-24
Variable dépendante: CO

             coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------
  const       -0.551698     2.97128        -0.1857        0.8546
  TAR          0.887580     0.195482        4.540         0.0002    ***
  NICOTINE     0.518470     3.25233         0.1594        0.8749
  WEIGHT       2.07934      3.17842         0.6542        0.5204

Moy. var. dép.          12.07083   Éc. type var. dép.      4.241357
Somme carrés résidus    26.90394   Éc. type de régression  1.159826
R2                      0.934975   R2 ajusté               0.925222
F(3, 20)                95.85850   p. critique (F)         4.85e-12
Log de vraisemblance   -35.42515   Critère d'Akaike        78.85030
Critère de Schwarz      83.56252   Hannan-Quinn            80.10045

Constante mise à part, la probabilité critique est la plus élevée pour la variable 3
(NICOTINE)

? vif

Facteurs d'inflation de variance
Valeur minimale possible = 1.0
```

```
Valeurs > 10.0 peut indiquer un problème de colinéarité

         TAR   12.736
    NICOTINE   12.757
      WEIGHT    1.090

VIF(j) = 1/(1 - R(j)^2), où R(j) est un coefficient de corrélation multiple
entre la variable j et les autres variables indépendantes

Belsley-Kuh-Welsch collinearity diagnostics:

                         --- variance proportions ---
    lambda      cond     const       TAR  NICOTINE    WEIGHT
     3.890     1.000     0.000     0.001     0.000     0.000
     0.103     6.152     0.017     0.033     0.011     0.013
     0.004    30.453     0.008     0.902     0.946     0.060
     0.003    35.389     0.975     0.065     0.042     0.927

  lambda = eigenvalues of X'X, largest to smallest
  cond   = condition index
  note: variance proportions columns sum to 1.0

? modtest --normality

Fréquence pour uhat1, obs 1-24
nombre de classes = 7, moyenne 3.44169e-015, éc. type = 1.15983

       intervalle  pt central   fréquence   rel.   cum.

           < -1.7618   -2.1083        1      4.17%    4.17% *
   -1.7618 - -1.0687   -1.4152        2      8.33%   12.50% ***
   -1.0687 - -0.37566  -0.72219       8     33.33%   45.83% ************
  -0.37566 -  0.31740  -0.029129      5     20.83%   66.67% *******
   0.31740 -  1.0105    0.66393       2      8.33%   75.00% ***
    1.0105 -  1.7035    1.3570        5     20.83%   95.83% *******
          >=  1.7035    2.0500        1      4.17%  100.00% *

Test de l'hypothèse nulle de normalité de la distribution :
Chi-deux(2) = 0.401 avec p. critique 0.81828

? leverage
                  résidu       puissance       influence        DFFITS
                    u          0<=h<=1         u*h/(1-h)

       1       -0.85785          0.171         -0.17662         -0.365
       2        0.12645          0.185         0.028686          0.056
       3          1.374          0.086          0.12881          0.385
       4        0.13775          0.167         0.027588          0.057
       5       -0.14871          0.133         -0.02291         -0.053
       6       -0.78791          0.157         -0.14701         -0.316
       7       -0.56504          0.098        -0.061318         -0.166
       8       -0.41158          0.120        -0.055917         -0.136
       9        0.35603          0.134         0.054905          0.127
      10        -1.1369          0.076         -0.09392         -0.293
      11       -0.85451          0.231         -0.25735         -0.458
```

```
      12          1.315          0.095          0.13777          0.390
      13        -2.1083          0.260         -0.74045         -1.385
      14       -0.09117          0.155        -0.016773         -0.036
      15       -0.53578          0.460*        -0.45704         -0.572
      16         1.3995          0.255          0.47972          0.840
      17       -0.93076          0.316         -0.42933         -0.658
      18         1.5396          0.195           0.3726          0.752
      19       -0.65362          0.152          -0.1176         -0.255
      20        0.96423          0.069          0.07117          0.232
      21        0.21709          0.096         0.023075          0.063
      22         1.1453          0.070         0.085584          0.280
      23        -1.5429          0.084         -0.14058         -0.430
      24           2.05          0.236          0.63172          1.226

('*' indique un point de puissance)

Critère de validation croisée = 41.3413

? omit NICOTINE WEIGHT
Test sur le Model 1 :

  Hypothèse nulle : les paramètres de régression sont nuls pour les variables
    NICOTINE, WEIGHT
  Statistique de test: F(2, 20) = 0.232743, p. critique 0.794474
   L'omission de variables améliore 3 des 3 critères d'information.

Modèle 2: MCO, utilisant les observations 1-24
Variable dépendante: CO

             coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------
  const       1.41285       0.648217         2.180       0.0403     **
  TAR         0.928126      0.0528302       17.57        1.96e-014  ***

Moy. var. dép.          12.07083   Éc. type var. dép.      4.241357
Somme carrés résidus    27.53011   Éc. type de régression  1.118646
R2                      0.933462   R2 ajusté               0.930437
F(1, 22)                308.6377   p. critique (F)         1.96e-14
Log de vraisemblance   -35.70124   Critère d'Akaike        75.40248
Critère de Schwarz      77.75859   Hannan-Quinn            76.02756

#
? ols CO 0 TAR NICOTINE WEIGHT

Modèle 3: MCO, utilisant les observations 1-24
Variable dépendante: CO

             coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------
  const       -0.551698     2.97128        -0.1857        0.8546
  TAR          0.887580     0.195482        4.540         0.0002    ***
  NICOTINE     0.518470     3.25233         0.1594        0.8749
  WEIGHT       2.07934      3.17842         0.6542        0.5204

Moy. var. dép.          12.07083   Éc. type var. dép.      4.241357
```

```
Somme carrés résidus    26.90394   Éc. type de régression  1.159826
R2                      0.934975   R2 ajusté               0.925222
F(3, 20)                95.85850   p. critique (F)         4.85e-12
Log de vraisemblance   -35.42515   Critère d'Akaike        78.85030
Critère de Schwarz      83.56252   Hannan-Quinn            80.10045

Constante mise à part, la probabilité critique est la plus élevée pour la variable 3
(NICOTINE)

? restrict
? b[3] = 0
? b[4] = 0
? end restrict
Ensemble de restrictions
 1: b[NICOTINE] = 0
 2: b[WEIGHT] = 0

Statistique de test: F(2, 20) = 0.232743, avec p. critique = 0.794474

Estimations contraintes:

             coefficient   erreur std.   t de Student   p. critique
  ---------------------------------------------------------------
  const       1.41285       0.648217         2.180       0.0403     **
  TAR         0.928126      0.0528302       17.57        1.96e-014  ***
  NICOTINE    0.000000      0.000000        NA          NA
  WEIGHT      0.000000      0.000000        NA          NA

  Écart type de la régression = 1.11865
```

Pas mal, vraiment. Nous retrouvons tous les résultats de notre session d'analyse. On notera que les sorties graphiques sont exportées dans des fichiers temporaires créés dans notre espace de documents.

6 Conclusion

Non, je n'ai pas viré ma cuti. Je continuerai à m'investir dans R et Python parce que ce sont des compétences (à la fois logiciels de statistique / machine learning et langage de programmation) que peuvent valoriser mes étudiants auprès de leurs futurs employeurs. Mais l'hégémonie n'est jamais une bonne chose. Il faut continuer à s'intéresser aux solutions alternatives. Une bonne manière de ne pas être dépendant d'un outil est justement d'en connaître et tester un grand nombre.

Gretl est un logiciel de qualité. Il propose des fonctionnalités et des techniques de modélisation prédictive puissantes. Malheureusement, son positionnement fortement marqué ''économétrie'' restreint son audience. De plus, il n'y a pas ou prou de tutoriels détaillés avec des exemples de traitements didactiques en français (j'en ai trouvé peu, assez succincts souvent, en effectuant une recherche Google à la mi-mai 2019). C'est là un frein considérable pour la diffusion du logiciel auprès

de la communauté francophone. Dommage parce que sa prise en main est relativement facile. Pour ma part, je me suis initié à Gretl en très peu de temps. Une fois que j'ai (re)trouvé mes marques, il m'a été facile de reproduire les traitements (et les résultats) que j'ai pu réaliser par ailleurs sous R. Finalement, comme souvent, la vraie barrière à l'entrée reste le langage de programmation qui repose sur une syntaxe et des structures qui nécessitent un certain apprentissage….

7 Références

''gretl'' – Gnu Regression, Econometrics and Time-series Library
http://gretl.sourceforge.net/

A. Cottrell, R. Lucchetti, ''A Hansl Primer'', Janvier 2019 ;
http://ricardo.ecn.wfu.edu/pub/gretl/manual/PDF/hansl-primer-a4.pdf

A. Cottrell, R. Lucchetti, ''Gretl User's Guide'', Mars 2019 ;
http://gretl.sourceforge.net/gretl-help/gretl-guide.pdf

L.C. Adkins, ''Using gretl for Principles of Econometrics'', 5th Edition, Novembre 2018 ;
http://www.learneconometrics.com/gretl/poe5/using_gretl_for_POE5.pdf

R. Rakotomalala, ''Cours économétrie'' ;
https://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html