XLSTAT
your data analysis solution

Préparé par :

- Elmannaoui Baraa
- Saital Abdelali

Encadré par :

- PR. M. El Merouani

Master spécialisé : Gestion Informatique de l'Entreprise

Année universitaire : 2016/2017

PLan

Introduction

1. **Installation**
2. **sélection des données**
3. **options**
4. **codage**
   - ❖ BOITE DE DIALOGUE
   - ❖ EXEMPLE
5. **régression linéaire**
   - ❖ DESCRIPTION
   - ❖ BOITE DE DIALOGUE
   - ❖ EXEMPLE
6. **Arima**
   - ❖ DESCRIPTION
   - ❖ BOITE DE DIALOGUE
   - ❖ RESULTATS
   - ❖ EXEMPLE

Introduction

XLSTAT est développé depuis plus de dix ans dans le but de rendre accessible au plus grand nombre un outil d'analyse de données et de statistique à la fois puissant, complet et convivial.

L'accessibilité vient de la compatibilité avec toutes les versions de Microsoft Excel aujourd'hui utilisées, de l'interface disponible en 7 langues (allemand, anglais, français, espagnol, italien, japonais, portugais) et de la mise à disposition sur le site www.xlstat.com d'une version d'évaluation utilisable 30 jours.

La puissance de XLSTAT vient à la fois du langage de programmation C++, et des algorithmes utilisés, qui sont le fruit des travaux de recherche de centaines de chercheurs statisticiens, mathématiciens ou informaticiens. Chaque développement d'une nouvelle fonctionnalité de XLSTAT est précédé d'une phase de recherche bibliographique approfondie, voire d'échanges avec les spécialistes des méthodes concernées.

La complétude de XLSTAT est le fruit d'une part de plus de dix ans de travail, et d'autre part d'échanges réguliers avec les utilisateurs, dont les idées et suggestions permettent de faire progresser le logiciel encore plus vite.

Enfin, la convivialité vient de l'interface, qui après quelques minutes de prise en main, rend facile et efficace l'utilisation de méthodes parfois très complexes qui requièrent dans d'autres logiciels des heures d'apprentissage.

L'architecture du logiciel a considérablement évolué au cours des 5 dernières années afin de prendre en compte les progrès d'Excel, et les problèmes de compatibilité entre les différentes plates-formes. Le logiciel s'appuie aujourd'hui sur le Visual Basic Application pour les interfaces et le C++ pour les calculs. Comme toujours, les équipes d'Addinsoft et des distributeurs de XLSTAT se tiennent à votre disposition pour répondre à toute question, ou pour prendre en compte vos remarques et suggestions afin de continuer à améliorer le logiciel.

1. installation :

Pour installer XLSTAT vous devez :

- Soit double-cliquer sur le fichier xlstat.exe téléchargé depuis le site www.xlstat.com ou depuis le site de l'un de nos partenaires, ou disponible sur le CD-Rom dont vous disposez,
- Soit insérer le CD-Rom à votre disposition et attendre que la procédure d'installation démarre automatiquement.

Si vos droits sont restreints sur l'ordinateur que vous utilisez, vous devez faire appel à un administrateur de la machine pour qu'il installe le logiciel. Une fois l'installation terminée, l'administrateur doit veiller à laisser un droit d'accès lecture/écriture aux éléments suivants :

- Dossiers du disque dur :
  - ✓ Dossier dans lequel se trouve Excel.exe
  - ✓ Dossier dans lequel se trouve les fichiers utilisateur, (ex : C:\...\Application Data\Addinsoft\XLSTAT\).

Le répertoire pour les fichiers utilisateur pourra être changé ultérieurement par une personne ayant des droits d'administrateur sur l'ordinateur. Pour cela, il suffit d'utiliser l'option correspondante dans l'onglet « Avancées » de la boîte de dialogue des options XLSTAT.

2. sélection des données

Comme pour l'ensemble des modules XLSTAT, la sélection des données se fait directement sur la feuille Excel, de préférence avec la souris. Les logiciels de statistique affichent classiquement des listes de variables à sélectionner ou non pour la méthode employée ou non. L'approche de XLSTAT est complètement différente puisque vous choisissez les données directement sur une ou plusieurs feuilles Excel.

Deux modes de sélection sont à votre disposition, sachant que pour chaque variable ou groupe de variables (par exemple d'une part la variable dépendante, d'autre part les variables quantitatives explicatives) vous pouvez opter pour l'un des modes. Les deux modes sont :

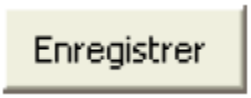
- Sélection par plage : vous sélectionnez avec la souris l'ensemble des cellules de la feuille Excel correspondant aux variables ou au tableau de données, après avoir cliqué dans la zone correspondante de la boîte de dialogue.

- Sélection par colonnes : ce mode de sélection ne peut être utilisé que si votre tableau de données commence sur la première ligne de la feuille Excel. Après avoir cliqué dans la zone de la boîte de dialogue correspondant à la sélection que vous voulez faire, vous devez cliquer sur le nom de la première colonne correspondant à votre tableau (A, B, C, …), puis sélectionner les autres colonnes en laissant le bouton droit de la souris enfoncé.
- Sélection par lignes : ce mode de sélection ne peut être utilisé que si votre tableau de données commence sur la première colonne de la feuille Excel (colonne A). Après avoir cliqué dans la zone de la boîte de dialogue correspondant à la sélection que vous voulez faire, vous devez cliquer sur le nom de la première ligne correspondant à votre tableau (1, 2, 3, …), puis sélectionner les autres lignes en laissant le bouton droit de la souris enfoncé.

3. options :

XLSTAT offre un nombre important d'options afin de vous permettre une utilisation personnalisée et optimale du logiciel.

Pour afficher la boîte de dialogue des options de XLSTAT, cliquez sur la commande « Options » du menu XLSTAT ou cliquez sur le bouton  de la barre d'outils XLSTAT.



: cliquez sur ce bouton pour enregistrer les modifications.

: cliquez sur ce bouton pour fermer la boîte de dialogue. Si vous n'avez pas préalablement enregistré vos modifications, elles ne seront pas prises en compte.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.

Onglet Générales :

Langue : utilisez cette option pour modifier la langue de l'interface de XLSTAT.

Entrées des boîtes de dialogue :

- Mémoriser pendant une session : activez cette option si vous souhaitez que XLSTAT mémorise le temps d'une session (ouverture / fermeture de XLSTAT) les différentes entrées des boîtes de dialogue.
  - ✓ Y compris pour les sélections de données : activez cette option si vous souhaitez que XLSTAT conserve pendant une session les sélections de données.
- Mémoriser d'une session à l'autre : activez cette option si vous souhaitez que XLSTAT mémorise les différentes entrées des boîtes de dialogue d'une session à l'autre.
  - ✓ Y compris pour les sélections de données : activez cette option si vous souhaitez que XLSTAT conserve aussi d'une session à l'autre les sélections de données. Cette option est particulièrement utile si vous travaillez souvent sur des feuilles Excel qui ont le même nom et une structure de données identiques.

Demander la confirmation des sélections : activez cette option si vous souhaitez que XLSTAT vous demande de confirmer les sélections de données après que vous avez cliqué sur le bouton OK des boîtes de dialogue. Si vous activez cette option, vous aurez la possibilité de vérifier le nombre de lignes et de colonnes sélectionnées pour l'ensemble des sélections actives.

Montrer seulement les fonctions actives dans les menus et les barres d'outils : Activez cette option si vous souhaitez que seules les fonctions actives correspondant à des modules auxquels la licence donne accès soient affichées dans le menu XLSTAT et les barres d'outils.

Onglet Sorties :

Position des nouvelles feuilles : si vous choisissez l'option de sortie « Feuille » dans les boîtes de dialogue des fonctions XLSTAT, utilisez cette option pour modifier la position des feuilles de résultats dans le classeur Excel.

Nombre de décimales : choisissez le nombre de décimales à afficher pour les résultats numériques. Notez que vous avez toujours la possibilité de voir par la suite un nombre de décimales inférieur ou supérieur en utilisant les options de formatage d'Excel.

p-value minimale : entrez la valeur p-value minimale en-dessous de laquelle la p-value est remplacée par « < p » où p est la p-value minimale

Afficher les titres en gras : activez cette option pour que XLSTAT affiche les titres des tableaux de résultats en gras.

Afficher l'en-tête des tableaux en gras : activez cette option pour que XLSTAT affiche entêtes des tableaux de résultats en gras.

Afficher la liste des résultats dans l'en-tête du rapport : activez cette option pour que XLSTAT affiche la liste des tableaux et graphiques de résultats dans l'en-tête du rapport.

Afficher le nom du projet dans l'en-tête du rapport : activez cette option pour que XLSTAT affiche le nom de votre projet dans l'en-tête du rapport, puis entrez le nom de votre projet dans le champ correspondant.

Elargir la première colonne du rapport par un facteur de X : activez cette option pour élargir automatiquement la première colonne du rapport de XLSTAT d'un facteur X. La valeur par défaut est 1, et correspond à laisser la largeur de la colonne inchangée.

Onglet Données manquantes :

Considérer les cellules vides comme des données manquantes : cette option est active par défaut et ne peut être désactivée. XLSTAT ne considère systématiquement qu'une cellule

vide dans une sélection correspond à une donnée manquante.

Considérer aussi les valeurs suivantes comme des données manquantes : si vous activez cette option, les valeurs indiquées dans la liste en dessous de l'option seront aussi considérées comme des données manquantes, que ce soit pour des données numériques ou des données nominales.

Considérer toute donnée textuelle comme une donnée manquante : cette option ne s'applique qu'aux sélections de données numériques. Quelle que soit la donnée textuelle rencontrée, elle sera considérée comme une donnée manquante. Si vous activez cette option soyez sûr que des données n'ont pas été converties par mégarde d'un format numérique en un format texte : vous risqueriez d'ignorer des observations alors qu'une rectification vous permettrait de les inclure dans les calculs.

Onglet Graphiques :

Afficher les graphiques sur des feuilles séparées : activez cette option pour que les graphiques soient affichés sur des feuilles graphiques séparées. Remarque : lorsque des graphiques sont affichés sur une feuille Excel standard, vous pouvez les convertir en feuille graphique séparée en les sélectionnant, puis en faisant un clic droit avec votre souris, puis en cliquant sur « Emplacement », puis en choisissant « sur une nouvelle feuille ».

Taille des graphiques :

- Automatique : choisissez cette option si vous souhaitez que XLSTAT détermine automatiquement la taille des graphiques en utilisant comme point de départ la hauteur et la largeur définies ci-dessous.
- Définie par l'utilisateur : activez cette option si vous souhaitez que XLSTAT affiche des graphiques dont la taille est exactement définie par les valeurs ci-dessous :
  - ✓ Largeur : entrez la valeur en points de la largeur des graphiques ;
  - ✓ Hauteur : entrez la valeur en points de la hauteur des graphiques.

Afficher des graphiques orthonormés : activez cette option pour que les graphiques issus d'analyses factorielles soient orthonormés. Cela permet d'avoir automatiquement des échelles identiques pour les abscisses et les ordonnées, et d'éviter des interprétations erronées du fait d'effets de dilatation artificiels.

Onglet Avancées :

Nombres aléatoires :

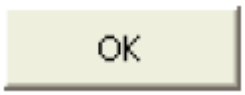
Fixer la graine à : activez cette option si vous voulez vous assurer que les résultats mettant en jeu des calculs sur des nombres aléatoires donnent toujours le même résultat. Entrez alors la valeur de la graine (le point de départ de génération des nombres aléatoires).

Chemin pour les fichiers utilisateurs : vous pouvez modifier le répertoire dans lequel doivent être enregistrés les fichiers utilisateurs en cliquant sur le bouton […] qui vous permettra de choisir le répertoire. Les fichiers utilisateurs comprennent les options définies dans cette boîte de dialogue et les options des boîtes de dialogues des différents outils. Le répertoire dans lequel sont enregistrés ces fichiers doit être accessible en lecture/écriture.

4. codage :

Utilisez cet outil pour recoder un tableau en utilisant un tableau de codage comprenant les valeurs initiales et les codes qui doivent les remplacer dans le nouveau tableau.

Boîte de dialogue :



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer calculs.

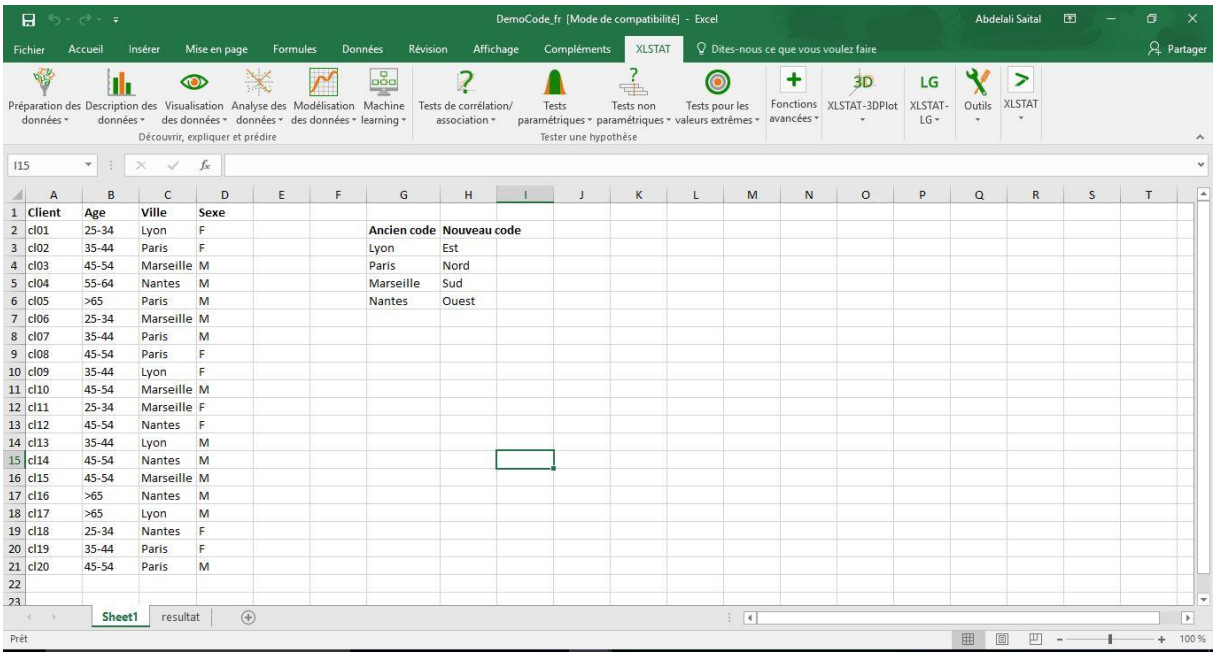
: cliquez sur ce bouton pour afficher l'aide.

: cliquez sur ce bouton pour rétablir les options par défaut.

: cliquez sur ce bouton pour effacer les sélections de données.

- Données : sélectionnez les données sur la feuille Excel. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.
- Tableau de codage : sélectionnez deux colonnes correspondant au tableau de codage. La première colonne doit contenir les valeurs telles qu'elles sont dans le tableau des données sélectionnées, et la seconde colonne les codes correspondants à utiliser dans le tableau recodé. Si des en-têtes ont été sélectionnés, veuillez vérifier que l'option « Libellés des colonnes » est activée.
- Libellés des colonnes : activez cette option si la première ligne des données sélectionnées (Données et tableau de codage) contient un libellé.
- Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.
- Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.
- Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.
- Afficher l'en-tête du rapport : désactivez cette option si vous souhaitez que le tableau disjonctif complet commence dès la première ligne de la feuille Excel (cas d'une sortie dans une feuille ou un classeur), et non après l'en-tête du rapport.

Exemple :



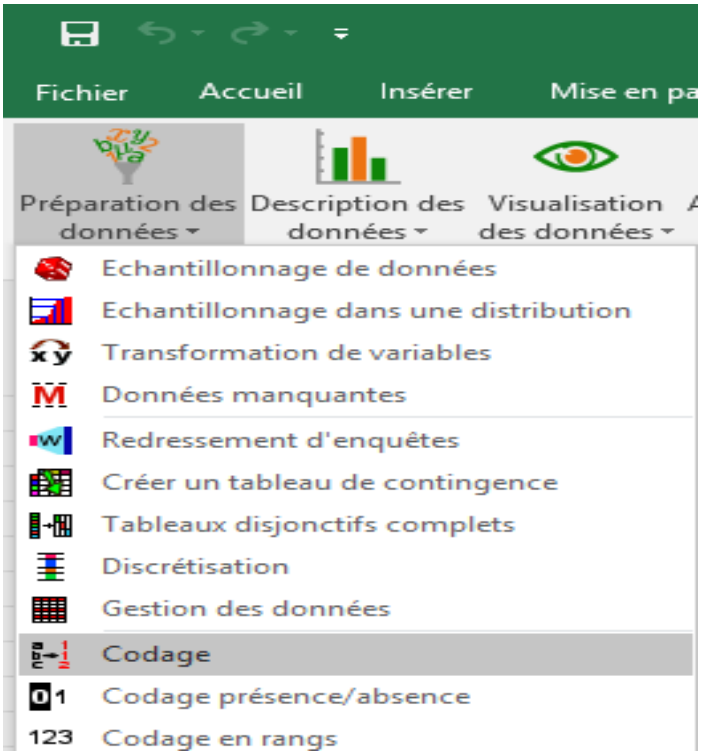
Vous allez trouver une feuille Excel contenant l'exemple de données et de résultats.

Les données correspondent à un échantillon de 20 clients avec des informations telles que leur catégorie d'âge, leur ville de résidence et leur genre.

Nous allons recoder la variable "Ville" à l'aide d'une table de codage contenant les nouvelles et anciennes valeurs.

Paramétrer le recodage des données avec une table de codage

Une fois XLSTAT lancé, cliquez sur l'icône **Préparation des données** et choisissez la fonction **Codage** ou dans la barre d'outils **Préparation des données** sélectionnez l'icône **Codage** (ci-dessous)



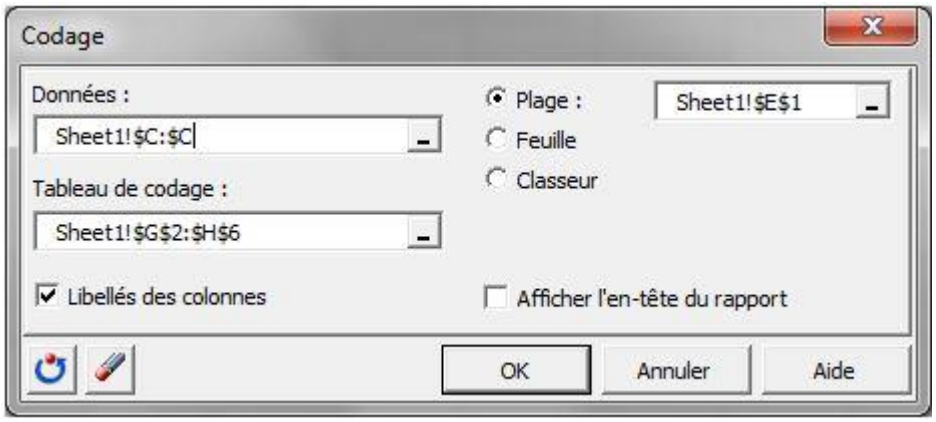
Une fois le bouton cliqué, la boîte de dialogue apparaît. Vous pouvez alors sélectionner la variable à recoder sur la feuille Excel. Sélectionnez la colonne C correspondant à la variable "Ville".

Sélectionnez ensuite la table de codage.

Notez qu'il faut aussi selectionner le libéllés des colonnes de cette table.

Pour obtenir les résultats accolés à la table de données, choisissez l'option **Plage**. Sélectionnez la **cellule E1** pour avoir la nouvelle variable accolée.

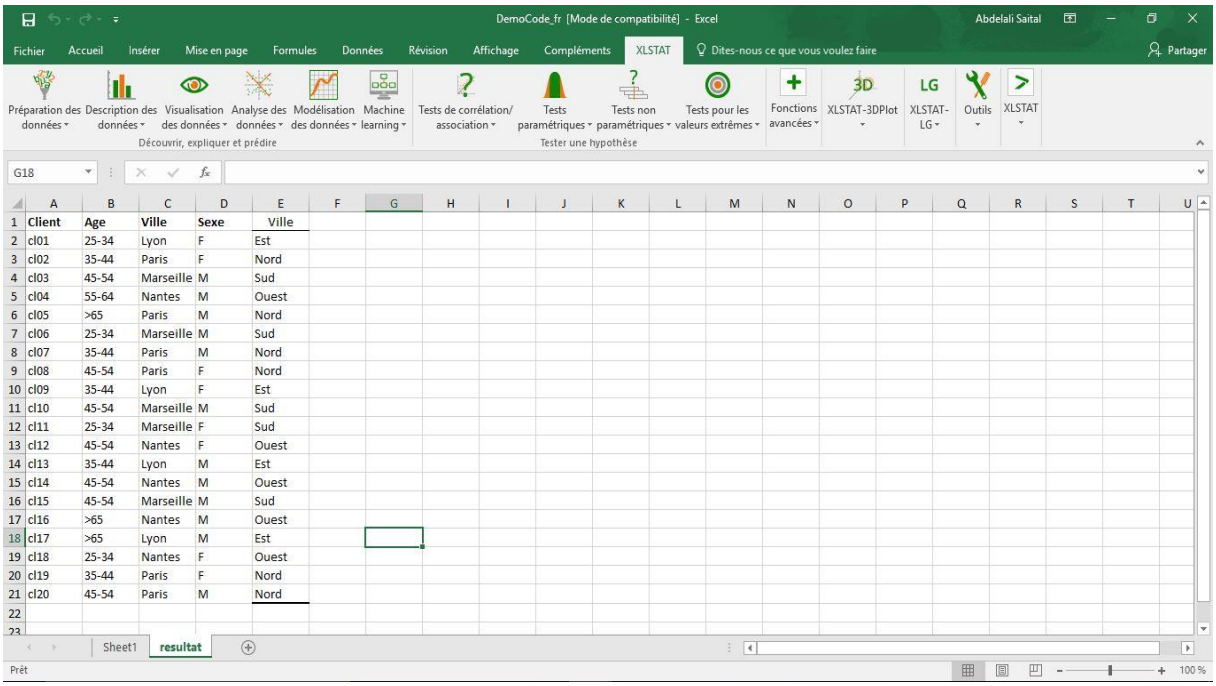
Pour ne pas avoir de texte au-dessus des résultats décochez l'option : **Afficher l'entête du rapport**.



Quand vous êtes prêt, cliquez sur **OK**. Les résultats s'affichent.

Résultats du codage des données

Vérifiez que la nouvelle colonne respecte bien le code de la table de codage.



	A	B	C	D	E
1	Client	Age	Ville	Sexe	Ville
2	cl01	25-34	Lyon	F	Est
3	cl02	35-44	Paris	F	Nord
4	cl03	45-54	Marseille	M	Sud
5	cl04	55-64	Nantes	M	Ouest
6	cl05	>65	Paris	M	Nord
7	cl06	25-34	Marseille	M	Sud
8	cl07	35-44	Paris	M	Nord
9	cl08	45-54	Paris	F	Nord
10	cl09	35-44	Lyon	F	Est
11	cl10	45-54	Marseille	M	Sud
12	cl11	25-34	Marseille	F	Sud
13	cl12	45-54	Nantes	F	Ouest
14	cl13	35-44	Lyon	M	Est
15	cl14	45-54	Nantes	M	Ouest
16	cl15	45-54	Marseille	M	Sud
17	cl16	>65	Nantes	M	Ouest
18	cl17	>65	Lyon	M	Est
19	cl18	25-34	Nantes	F	Ouest
20	cl19	35-44	Paris	F	Nord
21	cl20	45-54	Paris	M	Nord
22					

5. régression linéaire :

Utilisez ce module pour créer un modèle de régression linéaire simple ou multiple dans un but explicatif ou prédictif.

Description :

La régression linéaire est sans aucun doute la méthode statistique la plus utilisée. On distingue habituellement la régression simple (une seule variable explicative) de la régression multiple (plusieurs variables explicatives) bien que le cadre conceptuel et les méthodes de calculs soient identiques.

Le principe de la régression linéaire est de modéliser une variable dépendante quantitative Y, au travers d'une combinaison linéaire de p variables explicatives quantitatives, X1, X2, …, Xp.

Le modèle déterministe (ne prenant pas en compte d'aléa) s'écrit pour une observation i,

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

Où yi est la valeur observée pour la variable dépendante pour l'observation i, Xij est la valeur prise par la variable j pour l'observation i, et ε_i est l'erreur du modèle.

Le cadre statistique et les hypothèses qui l'accompagnent ne sont pas nécessaires pour ajuster ce modèle. Par ailleurs la minimisation par la méthode des moindres carrés (on minimise la somme des erreurs quadratiques ε^2_i) fournit une solution analytique exacte.

Néanmoins si l'on veut pouvoir tester des hypothèses et mesurer le pouvoir explicatif des différentes variables explicatives dans le modèle, un cadre statistique est nécessaire.

Les hypothèses de la régression linéaire sont les suivantes : les ε_i eurs suivent une même loi normale $N(0,\sigma)$ et sont indépendantes.

L'écriture du modèle complétée par cette hypothèse a pour conséquence que, dans le cadre du modèle de régression linéaire, les yi sont des réalisations de variables aléatoires de moyenne μi et de variance σ^2, avec :

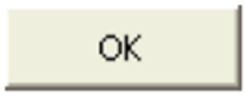
$$\mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

i

Si l'on souhaite utiliser les différents tests proposés dans les résultats de la régression linéaire il est recommandé de vérifier a posteriori que les hypothèses sous-jacentes sont bien vérifiées. La normalité des résidus peut être vérifiée en analysant certains graphiques ou en utilisant un test de normalité. L'indépendance des résidus peut être vérifiée en analysant certains graphiques ou en utilisant le test de Durbin Watson.

Boîte de dialogue :

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.


: cliquez sur ce bouton pour rétablir les options par défaut.


: cliquez sur ce bouton pour effacer les sélections de données.


: cliquez sur ce bouton pour changer la façon dont XLSTAT doit charger les données. Si la flèche est vers le bas, XLSTAT considère que les observations sont en lignes et les variables en colonnes. Si la flèche est vers la droite, XLSTAT considère que les variables sont en lignes et les observations en colonnes.

Onglet Général :

Y / Variables dépendantes :

Quantitatives : sélectionnez la ou les variables réponse que vous souhaitez modéliser. Si plusieurs variables sont sélectionnées, XLSTAT fera les calculs pour chacune des variables

indépendamment. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des variables » est activée.

X / Variables explicatives :

Quantitatives : sélectionnez la ou les variables qualitatives explicatives sur la feuille Excel.

Les données sélectionnées doivent être de type numérique. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Qualitatives : activez cette option si vous voulez inclure une ou plusieurs variables explicatives qualitatives dans le modèle. Dans ce cas, vous ne ferez plus de la régression linéaire, mais de l'ANCOVA. Sélectionnez alors la ou les variables correspondantes sur la feuille Excel. Les données sélectionnées peuvent être de tout type, mais les données numériques sont automatiquement considérées comme nominales. Si le libellé des variables a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des variables : activez cette option si la première ligne des données sélectionnées (variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations pour l'affichage des résultats. Si l'option « Libellés des variables » est activée, la première cellule de la sélection doit comprendre un en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (Obs1, Obs2, …).

Poids des observations : activez cette option si vous voulez pondérer les observations. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0. Un poids de 2 est équivalent à répéter deux fois la même observation. Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Poids dans la régression : activez cette option si vous voulez effectuer une régression par les moindres carrés pondérés. Si vous n'activez pas cette option, les poids seront tous considérés comme valant 1. Les poids doivent être impérativement supérieurs ou égaux à 0.

Si un en-tête de colonne a été sélectionné, veuillez vérifier que l'option « Libellés des variables » est activée.

Onglet Options :

Constante fixée : activez cette option pour fixer la constante du modèle de régression à une valeur que vous devez ensuite saisir (0 par défaut).

Tolérance : activez cette option pour permettre à l'algorithme de calcul de la régression OLS ne pas prendre en compte les variables qui seraient soit constantes soit trop corrélées avec d'autres variables déjà utilisées dans le modèle (0.0001 par défaut).

Interactions / Niveau : activez cette option pour inclure des interactions dans le modèle puis entrez le niveau maximum d'interaction (valeur comprise entre 1 et 4).

Intervalle de confiance (%) : entrez l'étendue en pourcentage de l'intervalle de confiance à utiliser pour les différents tests, et pour le calcul des intervalles de confiance autour des paramètres et des prédictions. Valeur par défaut : 95.

Sélection du modèle : activez cette option si vous souhaitez utiliser l'une des quatre méthodes de sélection proposées :

- Meilleur modèle : cette méthode permet de choisir le meilleur modèle parmi tous les modèles comprenant un nombre de variables variant de « Min variables » à « Max variables ». Par ailleurs le « critère » pour déterminer le meilleur modèle peut être choisi par l'utilisateur.
    - ✓ Critère : veuillez choisir le critère parmi la liste suivante : R^2 ajusté, Moyenne des Carrés des Erreurs (MCE), Cp de Mallows, AIC de Akaike, SBC de Schwarz, PC d'Amemiya.
    - ✓ Min variables : entrez le nombre minimum de variables à prendre en compte dans le modèle.
    - ✓ Max variables : entrez le nombre maximum de variables à prendre en compte dans le modèle.

Remarque : cette méthode peut entraîner des calculs longs car le nombre total de modèles explorés est la somme des Cn,k pour k variant entre « Min variables » et « Max variables », où Cn,k vaut n!/[(n-k)!k !]. Il est donc conseillé d'augmenter progressivement la valeur de « Max variables ».

- Stepwise : le processus de sélection commence par l'ajout de la variable ayant la plus forte contribution au modèle (le critère utilisé est la statistique t de Student). Si une seconde variable est telle que la probabilité associée à son t est inférieure à la « Probabilité pour l'entrée », elle est ajoutée au modèle. De même pour une troisième variable. A partir de l'ajout de la troisième variable, après chaque ajout, on évalue pour toutes les variables présentes dans le modèle quel serait l'impact de son retrait (toujours au travers de la statistique t). Si la probabilité est supérieure à la « Probabilité pour le retrait », la variable est retirée. La procédure se poursuit jusqu'à ce que plus aucune variable ne puisse être ajoutée/retirée.
- Ascendante : la procédure est identique à cette de la sélection progressive, hormis le fait que les variables sont uniquement ajoutées et jamais retirées.
- Descendante : la procédure commence par l'ajout simultané de toutes les variables. Les variables sont ensuite retirées du modèle suivant la procédure utilisée pour la sélection progressive.

Onglet Validation :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Jeu de validation : choisissez l'une des options pour définir le mode de sélection des observations utilisées pour la validation :

- Aléatoire : les observations sont sélectionnées de manière aléatoire. Le « Nombre d'observations » doit alors être saisi.
- N dernières lignes : les N dernières observations sont sélectionnées pour la validation.
- Le « Nombre d'observations » N doit alors être saisi.
- N premières lignes : les N premières observations sont sélectionnées pour la validation. Le « Nombre d'observations » N doit alors être saisi.

- Variable de groupe : si vous choisissez cette option, vous devez ensuite sélectionner une variable indicatrice composée de 0 pour les observations à utiliser pour le calcul du modèle, et de 1 pour les observations à utiliser pour la validation du modèle.

Onglet Prédiction :

- Prédiction : activez cette option si vous souhaitez sélectionner des données à utiliser en mode prédiction. Si vous activez cette option, vous devez veiller à ce que les données de prédiction soient organisées comme les données d'estimation : mêmes variables, même ordre dans les sélections. En revanche vous ne devez pas sélectionner de libellés de variables : la première ligne des sélections décrites ci-dessous doit être une ligne de données.
- Quantitatives : activez cette option pour sélectionner la ou les variables quantitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.
- Qualitatives : activez cette option pour sélectionner la ou les variables qualitatives explicatives. La première ligne ne doit pas comprendre d'en-tête.
- Libellés des observations : activez cette option si vous voulez utiliser des libellés d'observations disponibles sur une feuille Excel pour l'affichage des résultats. La première ligne ne doit pas comprendre d'en-tête. Si vous n'activez pas cette option, des libellés seront automatiquement créés (PredObs1, PredObs2, …).

Onglet Données manquantes :

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Estimer les données manquantes : activez cette option pour estimer les données manquantes avant le début des calculs.

- Moyenne ou mode : activez cette option pour estimer les données manquantes en utilisant la moyenne (variables quantitatives) ou le mode (variables qualitatives) pour les variables correspondantes.
- Plus proche voisin : activez cette option pour estimer les données manquantes d'une observation en recherchant le plus proche voisin de l'observation.

Onglet Sorties :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives pour les variables sélectionnées.

Corrélations : activez cette option pour afficher la matrice de corrélation pour les variables quantitatives (dépendantes et explicatives).

Analyse de la variance : activez cette option pour afficher le tableau d'analyse de la variance.

Type I SS : activez cette option pour afficher le tableau de l'analyse de la variance de Type I

(Type I Sum of Squares).

Type III SS : activez cette option pour afficher le tableau de l'analyse de la variance de Type III (Type III Sum of Squares).

Press : activez cette option pour calculer et afficher le coefficient de Press.

Coefficients normalisés : activez cette option pour afficher les paramètres normalisés du modèle (coefficients bêta).

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

- Prédictions ajustées : activez cette option pour calculer et afficher les prédictions ajustées dans le tableau des prédictions et résidus.
- D de Cook : activez cette option pour calculer et afficher les distances de Cook dans le tableau des prédictions et résidus.

Onglet Graphiques :

Options communes :

Graphiques de régression : activez cette option pour afficher les graphiques de régression :

- Coefficients normalisés : activez cette option pour afficher sur un graphique les paramètres normalisés du modèle avec leur intervalle de confiance.
- Prédictions et résidus : activez cette option pour afficher les graphiques suivants :

  (1) Droite de régression : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

(2) Variable explicative versus résidus normalisés : ce graphique n'est affiché que s'il n'y a qu'une seule variable explicative, et que cette variable est quantitative.

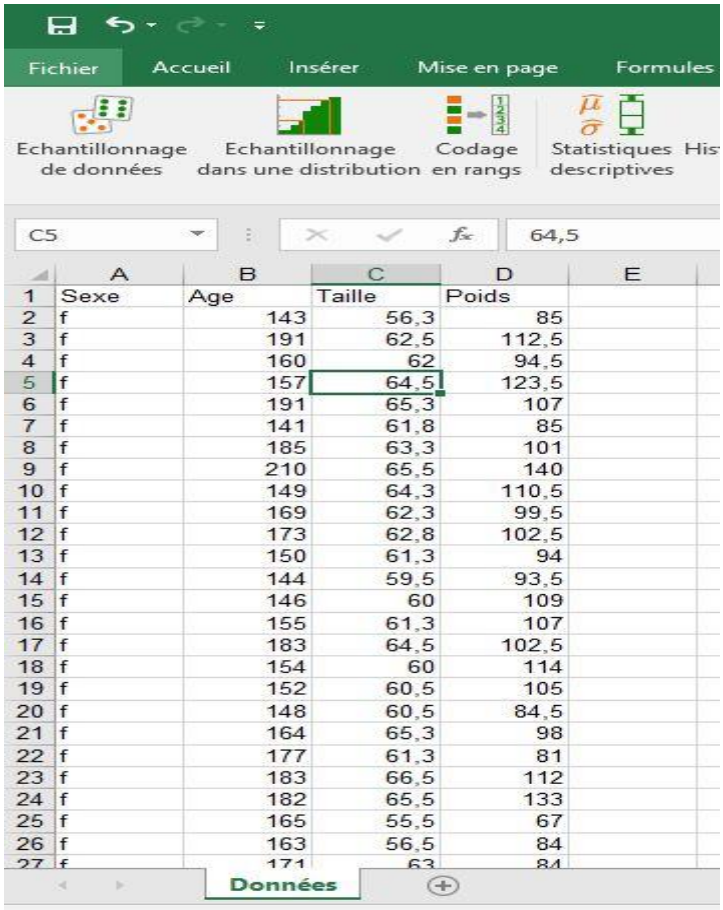
(3) Variable dépendante versus résidus normalisés.

(4) Prédictions pour la variable dépendante versus variable dépendante.

(5) Graphique en bâtons des résidus normalisés.

Intervalles de confiance : activez cette option pour afficher les intervalles de confiance sur les graphiques (1) et (4).

EXEPMLE :



	A	B	C	D	E
1	Sexe	Age	Taille	Poids	
2	f	143	56,3	85	
3	f	191	62,5	112,5	
4	f	160	62	94,5	
5	f	157	64,5	123,5	
6	f	191	65,3	107	
7	f	141	61,8	85	
8	f	185	63,3	101	
9	f	210	65,5	140	
10	f	149	64,3	110,5	
11	f	169	62,3	99,5	
12	f	173	62,8	102,5	
13	f	150	61,3	94	
14	f	144	59,5	93,5	
15	f	146	60	109	
16	f	155	61,3	107	
17	f	183	64,5	102,5	
18	f	154	60	114	
19	f	152	60,5	105	
20	f	148	60,5	84,5	
21	f	164	65,3	98	
22	f	177	61,3	81	
23	f	183	66,5	112	
24	f	182	65,5	133	
25	f	165	55,5	67	
26	f	163	56,5	84	
27	f	171	63	84	

Données

Jeu de données pour réaliser une régression linéaire simple

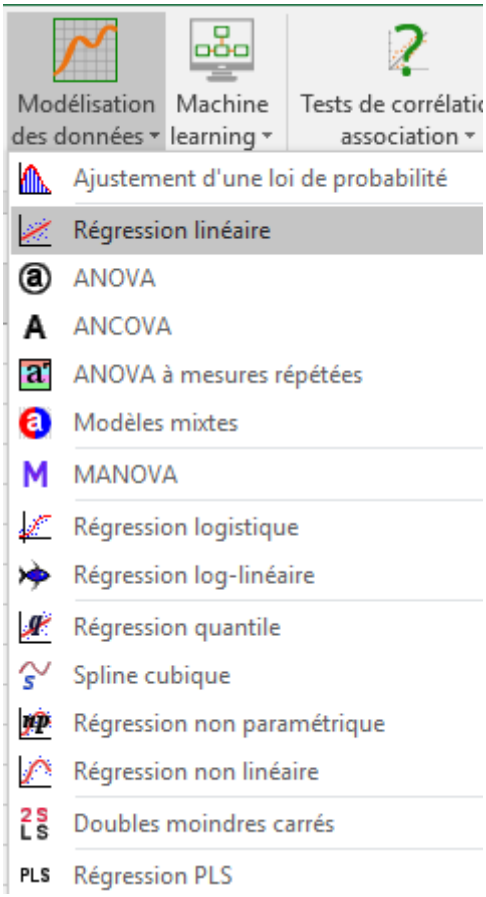
Vous allez trouver une feuille Excel contenant les données et les résultats de cet exemple.

En utilisant la régression linéaire simple, notre but est d'étudier comment le poids varie en fonction de la taille, et si une relation linéaire a un sens.

Nous nous limitons ici au cas des filles. Il s'agit ici d'une régression linéaire simple, car une seule variable explicative est utilisée (la taille).

Paramétrer une régression linéaire simple

Une fois XLSTAT lancé, choisissez la commande **XLSTAT / Modélisation / Régression linéaire** ou cliquez sur le bouton **Régression linéaire** de la barre d'outils **Modélisation**.

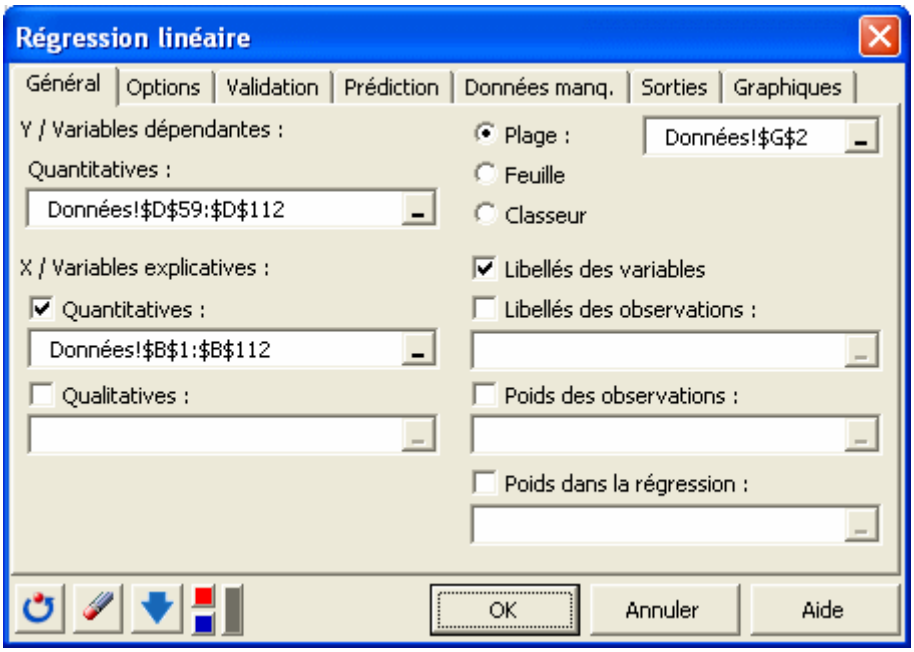


Une fois le bouton cliqué, la boîte de dialogue correspondant à la régression apparaît. Vous pouvez alors sélectionner les données sur la feuille Excel. La **Variable dépendante** correspond à la variable expliquée (ou variable à modéliser), qui est dans ce cas précis le "poids".

La **variable quantitative explicative** est ici la "taille". On veut ici expliquer la variabilité du poids par celle de la taille.

L'option **Libellés des colonnes** est activée car la première ligne des colonnes comprend le nom des variables.

Nous laissons l'option **Résidus** sélectionnée car nous analyserons les prédictions et les résidus pour valider l'hypothèse de normalité de la régression, et pour identifier des valeurs extrêmes.



Une fois que vous avez cliqué sur le bouton **OK**, les calculs commencent puis les résultats sont affichés.

<u>Interpréter les résultats de la régression linéaire simple</u>

Le premier tableau de résultats fournit les coefficients d'ajustement du modèle. Le R^2 (coefficient de détermination) donne une idée du % de variabilité de la variable à modéliser, expliqué par l a variable explicative. Plus ce coefficient est proche de 1, meilleur est le modèle.

Coefficients d'ajustement :	
R (coefficie	0.751
R² (coeffici	0.564
R²aj. (coefl	0.560
SCR	16614.585

Dans notre cas, 56% de la variabilité du poids est expliquée par la taille. Le reste de la variabilité est dû à des effets (autres variables explicatives) qui ne sont pas pris en compte dans cet exemple.

Le tableau d'analyse de la variance est un résultat qui doit être analysé attentivement (voir ci-dessous). C'est à ce niveau que l'on teste si l'on peut considérer que la variable explicative sélectionnée (la taille) apporte une quantité d'information significative au modèle (hypothèse nulle H0) ou non. En d'autres termes, c'est un moyen de tester si la moyenne de la variable à modéliser (le poids) suffirait à décrire les résultats obtenus ou non.

Evaluation de la valeur de l'information apportée par les variables (H0 = Y=Moy(Y)) :

Source	ddl	nme des ca	arrés moyer	F de Fisher	Pr > F
Modèle	1	21506.523	21506.523	141.094	< 0.0001
Résidus	109	16614.585	152.427		
Total	110	38121.108			

Le test du F de Fisher est utilisé. Etant donnée que la probabilité associée au F est dans ce cas inférieure à 0.0001, cela signifie que l'on prend un risque de se tromper de moins de 0.01% en concluant que la variable explicative apporte une quantité d'information significative au modèle.

Le tableau suivant fournit les détails sur le modèle et est essentiel dès lors que le modèle doit être utilisé pour faire des prévisions, des simulations ou s'il doit être comparé à d'autres résultats, par exemple les coefficients que l'on obtiendrait pour les garçons. Nous voyons que si le paramètre de la taille a un intervalle de confiance assez étroit, celui de la constante du modèle est assez large. L'équation du modèle est donnée sous le tableau. Le modèle indique que dans les limites de l'intervalle de variation de la variable taille données par les observations, à chaque fois que la taille augmente d'un inch, le poids augmente de 4 livres.

Estimation des paramètres du modèle :

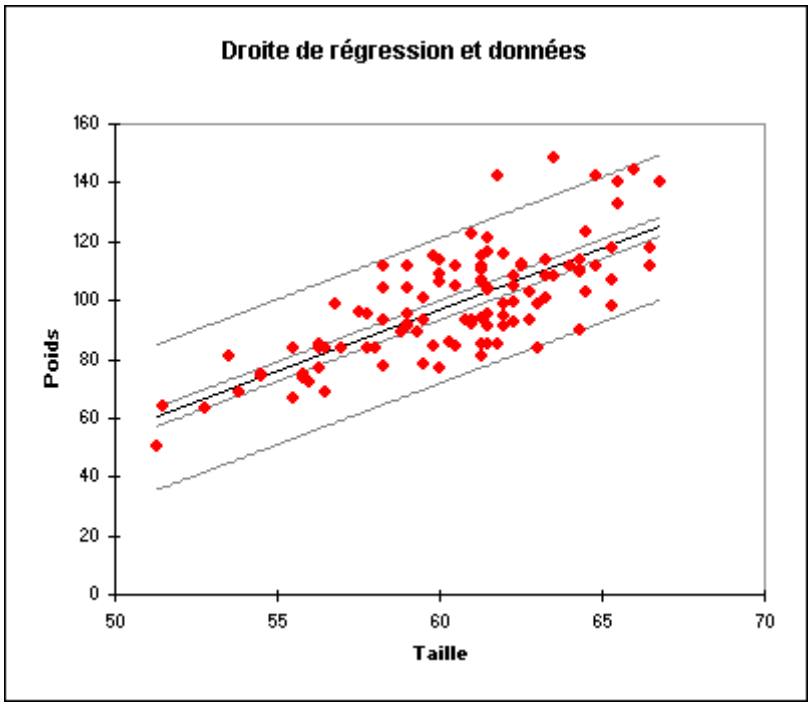
Paramètre	Valeur	Ecart-type	de Studen	Pr > t	ne inf. à 95	ne sup. à 95 %
Constante	-153.129	21.248	-7.207	< 0.0001	-195.242	-111.016
Taille	4.164	0.351	11.878	< 0.0001	3.469	4.858

*L'équation du modèle est : Y = -153.128910179442 + 4.16361172748231*Taille*

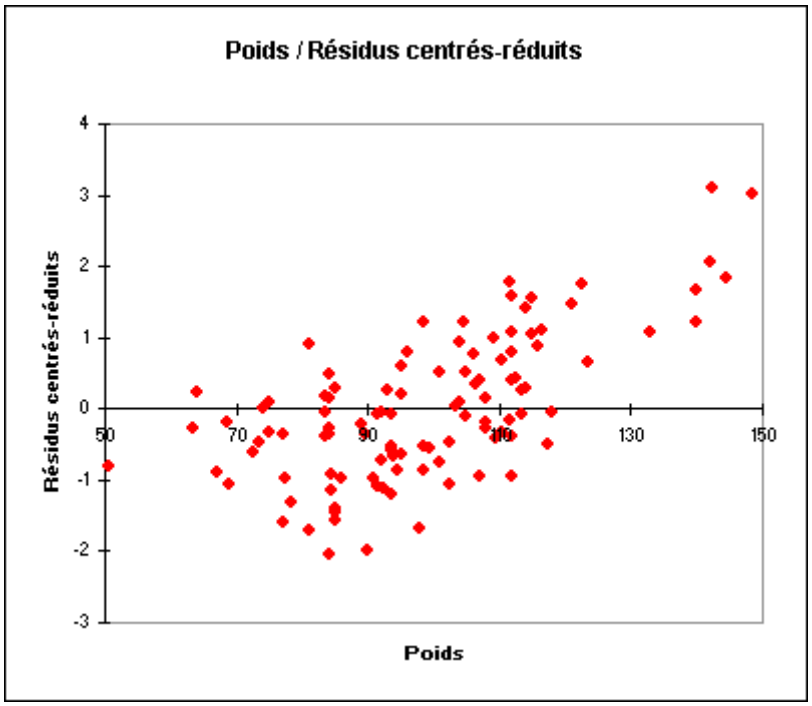
Le tabeau suivant présente l'analyse des résidus. Une attention particulière doit être portée aux résidus centrés réduits, qui, étant données les hypothèses liées à la régression linéaire, doivent être distribués suivant une loi normale N(0,1). Cela signifie, entre autres, que 95% des résidus doivent se trouver dans l'intervalle [-1.96, 1.96]. Etant donné le faible nombre de données dont on dispose ici, toute valeur en dehors de cet intervalle est révélatrice d'une donnée suspecte. Afin de mettre en évidence rapidement les valeurs se trouvant hors de l'intervalle [-1.96, 1.96], nous avons utilisé l'outil DataFlagger de XLSTAT.

Sur les 111 observations, cinq (26, 38, 64, 69, 77) sont hors de l'intervalle [-1.96, 1.96]. Cette analyse des résidus n'invalide donc pas l'hypothèse de normalité.

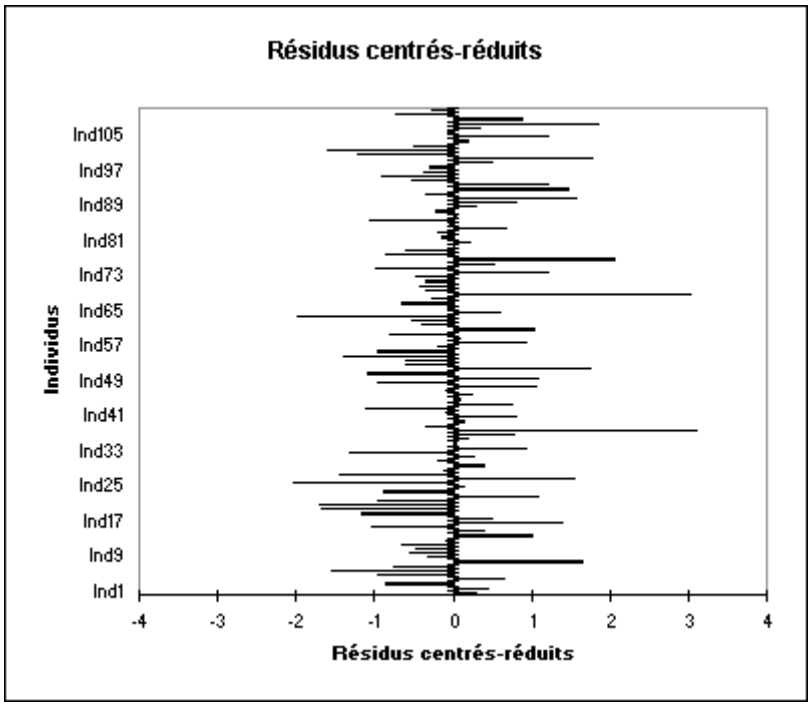
Le premier graphique permet de visualiser les données, la droite de régression, et les deux intervalles de confiance (le plus proche de la courbe est l'intervalle autour de la moyenne de l'estimateur, le second est l'intervalle autour de l'estimation ponctuelle aussi appelé intervalle de prédiction). On voit ainsi clairement une tendance linéaire, mais avec une forte variabilité autour de la droite. Les 5 valeurs suspectes sont en dehors du second intervalle de confiance.



Le troisième graphique semble indiquer que les résidus croissent en fonction du poids.



L'histogramme des résidus centrés réduits permet quant à lui de repérer rapidement et visuellement la présence de valeurs hors de l'intervalle [-2, 2].



En conclusion, la taille permet d'expliquer 56% de la variabilité du poids. Pour expliquer la variabilité restante, d'autres sources de variabitité doivent donc être prises en compte dans le modèle. Dans le tutoriel sur la régression linéaire multiple, l'âge est ajouté comme seconde variable explicative.

6. arima :

Utilisez cet outil pour ajuster un modèle ARMA (Autoregressive Moving Average), un modèle ARIMA (Autoregressive Integrated Moving Average) ou un modèle SARIMA (Seasonal Autoregressive Integrated Moving Average), et faire des prévisions sur la base de modèles dont les coefficients sont connus ou à estimer.

Description

Les modèles de la famille ARIMA permettent de représenter sous une forme succincte certains phénomènes variant avec le temps, et de faire des prévisions pour les valeurs futures du phénomène, avec un intervalle de confiance autour des prévisions.

L'écriture mathématique des modèles ARIMA varie d'un auteur à l'autre, ceci impliquant notamment des différences pour les signes des coefficients. La notation utilisée dans XLSTATTime correspond à celle de la plupart des logiciels.

Soit {Xt} une série chronologique de moyenne μ. Si la série suit un modèle ARIMA (p,d,q)(P,D,Q)s, alors on peut écrire :

$$\begin{cases} Y_t = (1-B)^d \left(1-B^s\right)^D X_t - \mu \\ \phi(\mathrm{B})\Phi\left(\mathrm{B}^s\right)Y_t = \theta(\mathrm{B})\Theta\left(\mathrm{B}^s\right)Z_t, \quad Z_t \propto N(0,\sigma^2) \end{cases}$$

avec

$$\begin{cases} \phi(\mathrm{z}) = 1 - \sum_{i=1}^{p} \phi_i z^i, \quad \Phi(\mathrm{z}) = 1 - \sum_{i=1}^{P} \Phi_i z^i \\ \theta(\mathrm{z}) = 1 + \sum_{i=1}^{q} \theta_i z^i, \quad \Theta(\mathrm{z}) = 1 + \sum_{i=1}^{Q} \Theta_i z^i \end{cases}$$

p est l'ordre de la partie autorégressive du modèle.

q est l'ordre de la partie moyenne mobile du modèle.

d est l'ordre de différentiation du modèle.

D est l'ordre de différentiation du modèle pour la partie saisonnière.

s est la période du modèle (par exemple 12 si les données sont mensuelles et que l'on a repéré une cyclicité à l'échelle de l'année.

P est l'ordre de la partie autorégressive saisonnière du modèle.

Q est l'ordre de la partie moyenne mobile saisonnière du modèle.

Remarque 1 : le processus {Yt} est causal si et seulement si pour tout z tel que |z| <=1, (z) ≠0 et (z) ≠ 0.

Remarque 2 : si D=0, on se trouve dans le cas d'un modèle ARIMA (p,d,q). Dans ce cas, P, Q et s sont considérés comme étant nuls.

Remarque 3 : si d=0 et D=0, on se trouve dans le cas d'un modèle ARMA (p,q).

Remarque 4 : si d=0, D=0 et q=0, on se trouve dans le cas d'un modèle AR(p).

Remarque 5 : si d=0, D=0 et p=0, on se trouve dans le cas d'un modèle MA(q).

Si les coefficients des polynômes $\phi, \Phi, \theta, \Theta$ sont inconnus, une fois les paramètres (p,d,q), (P,D,Q) et s saisis, XLSTAT-Time permet d'estimer les coefficients des polynômes, puis de calculer différentes statistiques d'ajustement, et si l'utilisateur le souhaite, de calculer des prévisions de valeurs futures.

Si les coefficients des polynômes $\phi, \Phi, \theta, \Theta$ sont connus, l'utilisateur peut les saisir. XLSTAT calcule ensuite différentes statistiques d'ajustement, et si l'utilisateur le demande, des prévisions de valeurs futures.

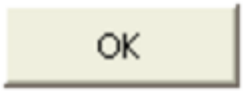
Dans le cas où D = 0, il est possible d'effectuer une estimation préliminaire des coefficients des polynômes □ et □ en utilisant la méthode proposée :

- Si q = 0, deux méthodes d'estimation préliminaire sont proposées. La première utilise l'algorithme de Yule-Walker, le seconde celui de Burg.
- Si p = 0, la méthode utilisée est l'algorithme des innovations.
- Si p ≠ 0 et q ≠ 0, la méthode utilisée est l'algorithme de Hannan-Rissanen.

Dans le cas où D ≠ 0, XLSTAT-Time effectue lui-même la recherche d'un point de départ raisonnable.

Boîte de dialogue

La boîte de dialogue est composée de plusieurs onglets correspondant aux différentes options disponibles tant pour la gestion des calculs que pour l'affichage des résultats. Vous trouverez ci-dessous le descriptif des différents éléments de la boîte de dialogue.



: cliquez sur ce bouton pour lancer les calculs.

: cliquez sur ce bouton pour fermer la boîte de dialogue sans effectuer les calculs.

: cliquez sur ce bouton pour afficher l'aide.

: cliquez sur ce bouton pour rétablir les options par défaut.

: cliquez sur ce bouton pour effacer les sélections de données.

Onglet Général :

Séries temporelles : sélectionnez la ou les séries temporelles dont vous voulez analyser le spectre. Si des en-têtes de colonnes ont été sélectionnés, veuillez vérifier que l'option « Libellés des séries » est activée.

Centrer : activez cette option pour centrer les séries avant de calculer le modèle.

Variance : activez cette option puis entrez la valeur de la variance si vous souhaitez imposer une variance des erreurs pour le modèle.

Données de date : activez cette option pour sélectionner des données de date. Ces données doivent être au format de data Excel, ou des valeurs numériques.

- Vérifier les intervalles : activez cette option si vous voulez que XLSTAT vérifie que les données de date sont bien régulièrement espacées.

Plage : si vous activez cette option, les résultats seront affichés à partir d'une cellule située dans une feuille existante. Vous devez alors sélectionner la cellule.

Feuille : activez cette option pour afficher les résultats dans une nouvelle feuille du classeur actif.

Classeur : activez cette option pour afficher les résultats dans un nouveau classeur.

Libellés des séries : activez cette option si la première ligne des données sélectionnées (Variables dépendantes et explicatives, libellés des observations, poids) contient un libellé.

Paramètres du modèle : entrez la valeur des différents ordres intervenant dans le modèle :

- p : entrez l'ordre de la partie autorégressive du modèle. Par exemple, entrez 1 pour un modèle AR(1) ou pour un modèle ARMA(1,2).
- d : entrez l'ordre de différentiation du modèle. Par exemple, entrez 1 pour un modèle ARIMA (0, 1,2).
- q : entrez l'ordre de la partie moyenne mobile du modèle. Par exemple, entrez 2 pour un modèle MA(2) ou pour un modèle ARIMA (1, 1,2).
- P : entrez l'ordre de la partie autorégressive saisonnière du modèle. Par exemple, entrez 1 pour un modèle ARIMA (1, 1,0) $(1, 1,0)^{12}$. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si D=0, on considère que P=0.

- D : entrez l'ordre de différentiation du modèle pour la partie saisonnière. Par exemple, entrez 1 pour un modèle ARIMA (0, 1,1) $(0, 1,1)^{12}$.
- Q : entrez l'ordre de la partie moyenne mobile saisonnière du modèle. Par exemple, entrez 1 pour un modèle ARIMA (0, 1,1) $(0, 1,1)^{12}$. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si D=0, on considère que Q=0.
- s : entrez la période du modèle. Vous ne pouvez modifier cette valeur que si $D \neq 0$. Si D=0, on considère que s=0.

Onglet Options :

Estimation préliminaire : activez cette option si vous souhaitez utiliser une méthode d'ajustement préliminaire. Cette option n'est disponible que si D=0.

- Yule-Walker : activez cette option pour estimer les coefficients du modèle autorégressif AR(p) avec l'algorithme de Yule-Walker.
- Burg : activez cette option pour estimer les coefficients du modèle autorégressif AR(p) avec l'algorithme de Burg.
- Innovations : activez cette option pour estimer les coefficients du modèle moyenne mobile MA(q) avec l'algorithme des Innovations.
- Hannan-Rissanen : activez cette option pour estimer les coefficients du modèle ARMA (p,q) avec l'algorithme de Hannan-Rissanen.

m/Auto : si vous choisissez la méthode des Innovations ou de Hannan-Rissanen, vous devez entrer la valeur m spécifique de chacun des algorithmes. Si vous choisissez Auto, XLSTAT détermine automatiquement quelle est la bonne valeur de m.

Coefficients initiaux : activez cette option pour sélectionner des valeurs initiales des coefficients du modèle.

- Phi : sélectionnez à ce niveau la valeur des coefficients correspondant à la partie autorégressive du modèle (y compris pour la partie saisonnière). Le nombre de valeurs sélectionné ici doit être égal à p+P.
- Theta : sélectionnez à ce niveau la valeur des coefficients correspondant à la partie moyenne mobile du modèle (y compris pour la partie saisonnière). Le nombre de valeurs sélectionné ici doit être égal à q+Q.

Optimiser : activez cette option pour estimer les coefficients selon l'une des deux méthodes proposées :

- Vraisemblance : activez cette option pour maximiser la vraisemblance.
- Moindres carrés : activez cette option pour minimiser la somme des carrés des erreurs.

Conditions d'arrêt :

- Itérations : entrez le nombre maximal d'itérations pour l'algorithme d'optimisation. Les calculs sont interrompus dès que le nombre maximal d'itérations est dépassé. Valeur par défaut : 500.
- Convergence : entrez la valeur seuil d'évolution maximale des communalités d'une itération à l'autre qui, une fois atteinte, permet de considérer que l'algorithme a convergé. Valeur par défaut : 0,00001.

Intervalles de confiance : entrez la valeur de l'intervalle de confiance pour les prédictions effectuées sur l'échantillon de validation et de prédiction.

Onglet Validation :

Validation : activez cette option si vous souhaitez utiliser une partie des données sélectionnées pour valider le modèle.

Pas de temps : entrez le nombre de pas de temps à la fin de la série sélectionnée qui doit être utilisé pour valider le modèle choisi.

Onglet Prédiction :

Prédiction : activez cette option pour effectuer des prédictions de nouvelles valeurs.

Pas de temps : entrez le nombre de pas de temps à prédire.

Onglet Données manquantes :

Ne pas accepter les valeurs manquantes : activez cette option pour que XLSTAT empêche la poursuite des calculs si des valeurs manquantes sont détectées.

Supprimer les observations : activez cette option pour supprimer les observations comportant des données manquantes.

Remplacer par la moyenne des valeurs précédente et suivante : activez cette option pour estimer les données manquantes par la moyenne de la première valeur précédente non manquante et de la première valeur suivante non manquante.

Onglet Sorties :

Statistiques descriptives : activez cette option pour afficher les statistiques descriptives des séries sélectionnées.

Coefficients d'ajustement : activez cette option pour afficher le tableau des statistiques d'ajustement du modèle.

Paramètres du modèle : activez cette option pour afficher le tableau des paramètres du modèle.

Prédictions et résidus : activez cette option pour afficher les prédictions et les résidus pour l'ensemble des observations.

Onglet Graphiques :

Afficher les graphiques : activez cette option pour afficher le graphique présentant les données originales et les prédictions du modèle, ainsi que le diagramme en bâtons des résidus.

Résultats

Statistiques simples : tableau dans lequel sont affichés le nombre d'observations, le nombre d'observations manquantes, le minimum, le maximum, la moyenne, la variance de la population (1/n) et l'écart type (1/n).

Si une estimation préliminaire et une optimisation ont été demandées, les résultats de l'estimation préliminaire sont affichés, suivis de ceux de l'optimisation. Si des coefficients initiaux ont été saisis, les résultats concernant ces coefficients sont d'abord affichés.

Coefficients d'ajustement :

- Observations : le nombre de données utilisées pour l'ajustement.
- SCE : la somme des carrés des résidus. Ce critère est minimisé lorsque l'option « Moindres carrés » est sélectionnée.
- Variance du bruit blanc : cette statistique est égale à SCE divisé par N. Dans certains logiciels cette statistique est désignée par sigma2.
- Variance du bruit blanc (estimée) : cette statistique est en principe égale à la précédente. Dans le cas des algorithmes de Yule-Walker et de Burg, une estimation légèrement différente est fournie.
- -2Log (Vrais.) : ce critère est minimisé dans le cas d'une optimisation basée sur le maximum de vraisemblance. Elle vaut l'opposé de deux fois le logarithme népérien de la vraisemblance.

- FPE : ce critère est dû à Akaike (Final Prediction Error). Ce critère est adapté pour les modèles autorégressifs.
- AIC : ce critère est dû à Akaike (Akaike Information Criterion).
- AICC : ce critère est dû à Brockwell (Akaike Information Criterion Corrected).
- SBC : ce critère est dû à Schwarz (Schwarz's Bayesian Criterion).

Paramètres du modèle :

Constante : le constant est systématiquement nul dans le cas de modèles ne comprenant pas de composante autorégressive. Dans le cas de modèles comprenant une composante autorégressive, la co$\mu.\phi(1).\Phi(1).$. La constante est aussi nul si l'option

« Centrer » n'est pas activée.

Le tableau suivant donne l'estimateur de chaque coefficient de chaque polynôme, ainsi que l'écart-type obtenu soit directement par la méthode d'estimation (estimation préliminaire) soit à partir de la matrice d'information de Fisher à l'issue de l'optimisation (désignée par Hess., pour Hessienne). Les écarts-types asymptotiques sont aussi calculés. Pour chaque coefficient et chaque écart-type est fourni un intervalle de confiance. Les coefficients sont identifiés de la manière suivante :

AR(i) : coefficient correspondant au coefficient d'ordre i du polynôme $\phi(z)$.

SAR(i) : coefficient correspondant au coefficient d'ordre i du polynôme $\Phi(z)$.

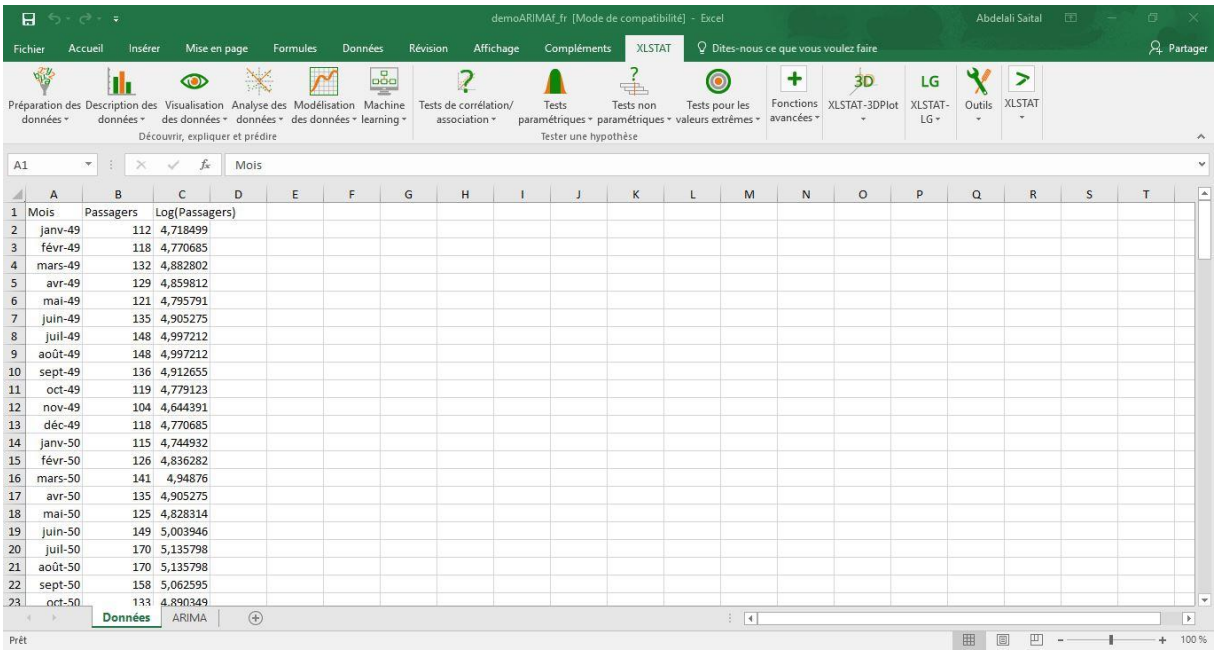
MA(i) : coefficient correspondant au coefficient d'ordre i du polynôme $\theta(z)$.

SMA(i) : coefficient correspondant au coefficient d'ordre i du polynôme $\Theta(z)$.

Prédictions et résidus : dans ce tableau sont affichés la série de départ, les prédictions calculées à partir du modèle, et les résidus correspondants. Si l'utilisateur l'a demandé, des prédictions pour les données de validation et pour les valeurs futures sont calculées, ainsi que les écart-types et les intervalles de confiance correspondants.

Graphiques : deux graphiques sont affichés. Le premier graphique permet de visualiser les données, les valeurs calculées à partir du modèle, les prévisions de validation et des valeurs futures, de même que les intervalles de confiance. Le second graphique permet de visualiser les résidus du modèle.

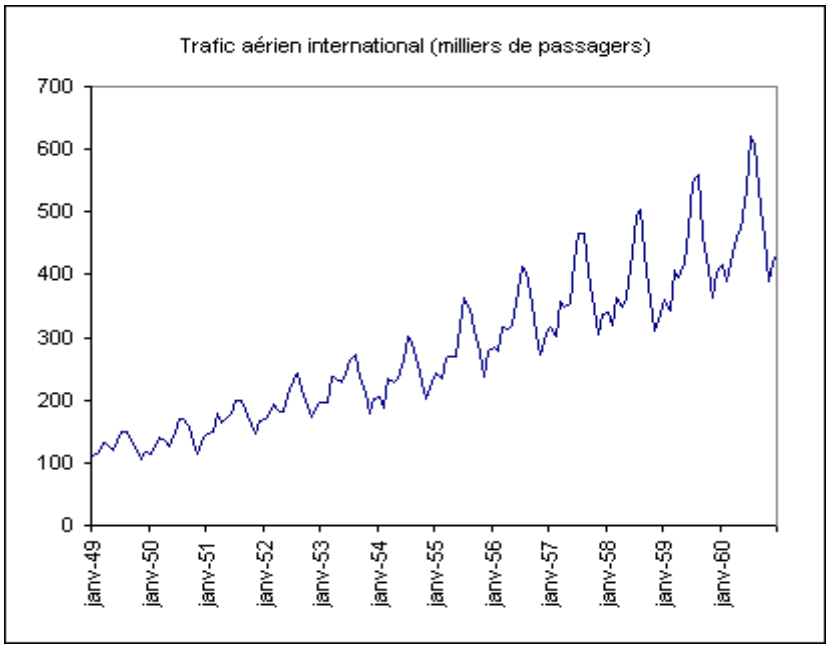
Exemple :



Jeu de données pour ajuster **un modèle ARIMA :**

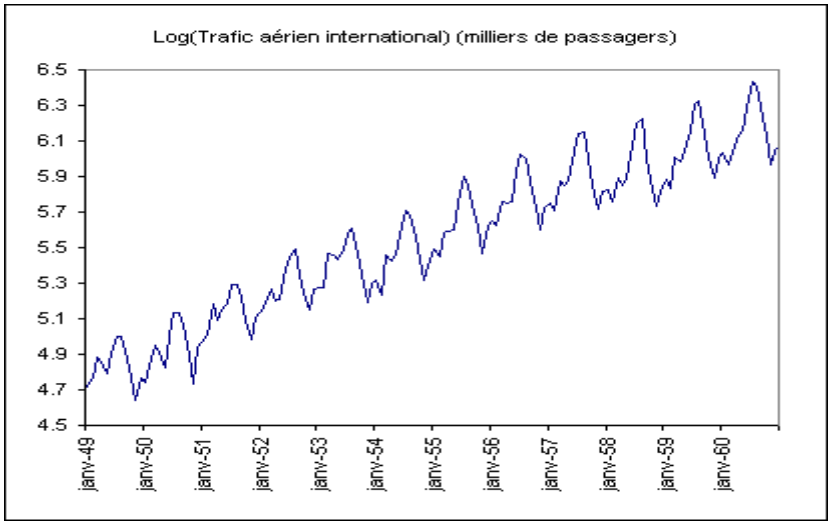
Vous allez trouver une feuille Excel contenant les données et les résultats de cet exemple.

Le but de l'analyse est d'ajuster le modèle sur les données des 11 premières années puis de prédire le trafic de l'année 1960 avec le modèle.



On note sur ce graphique que le nombre de passagers à tendance à augmenter régulièrement, que l'on retrouve chaque année un cycle similaire, mais que les variations à l'intérieur d'une même année sont de plus en plus fortes.

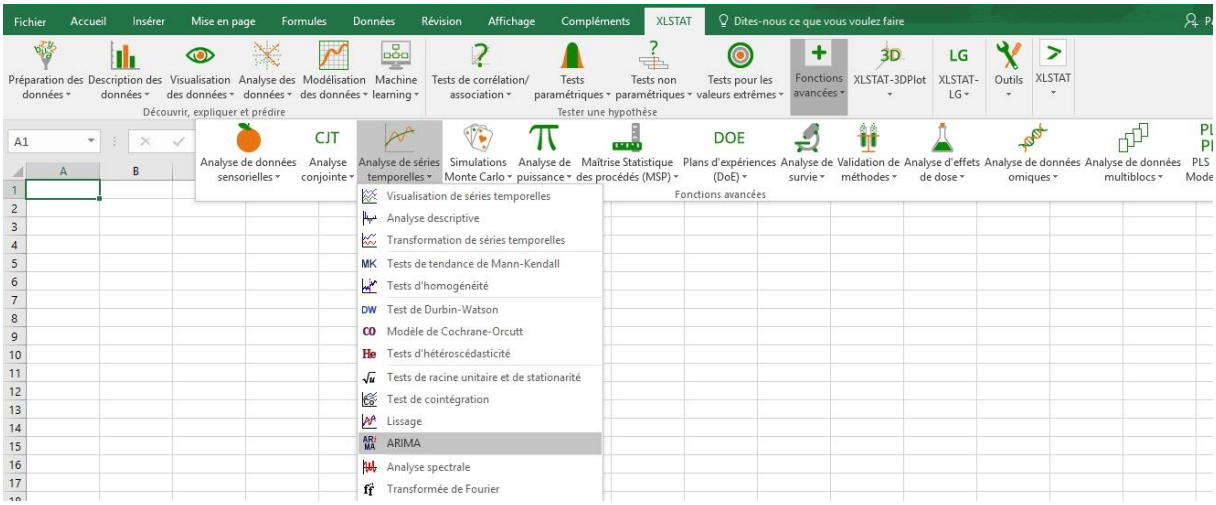
Afin de supprimer l'augmentation des variations intra-annuelles nous prenons le logarithme népérien des données. Nous pouvons vérifier sur le graphique ci-dessous que l'augmentation des variations intra-annuelles est nettement réduite.



On peut maintenant ajuster un modèle ARIMA(0,1, 1)$(0,1,1)^{12}$ qui semble approprié pour tenir compte à la fois de la composante tendancielle et de la cyclicité annuelle observées.

Paramétrer un modèle ARIMA

Pour activer la boîte de dialogue des méthodes de lissage, lancez XLSTAT, puis sélectionnez la commande **XLSTAT / Fonction avancée / Analyse des séries temporelles/ARIMA**.



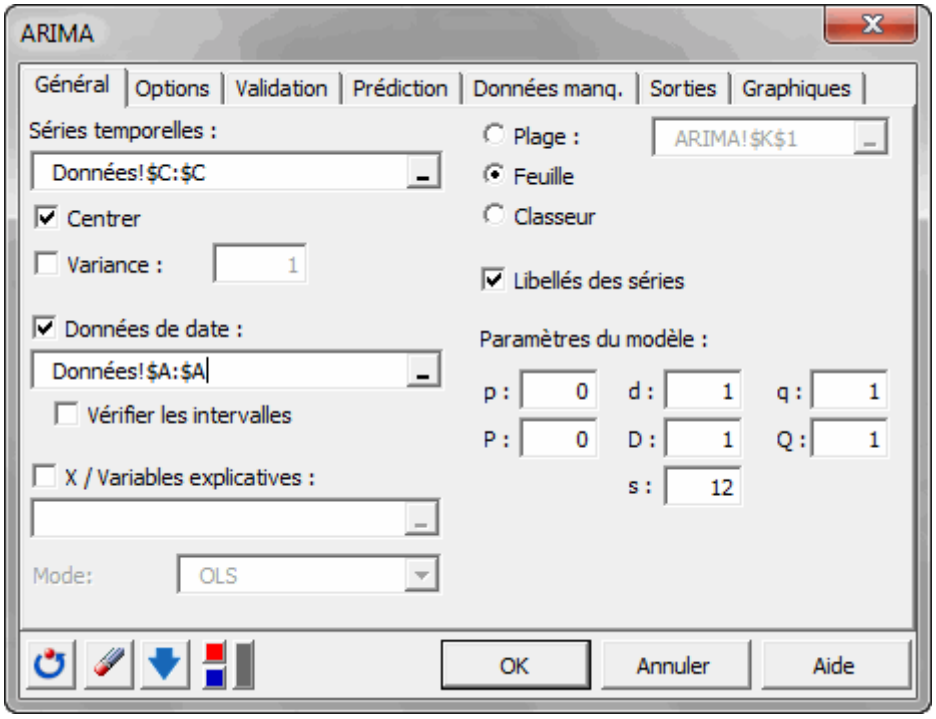
Une fois le bouton cliqué, la boîte de dialogue des méthodes de lissage apparaît.

Vous pouvez alors sélectionner les données sur la feuille Excel. La **Série à analyser** correspond à la série étudiée, les données "Log(Passagers)".

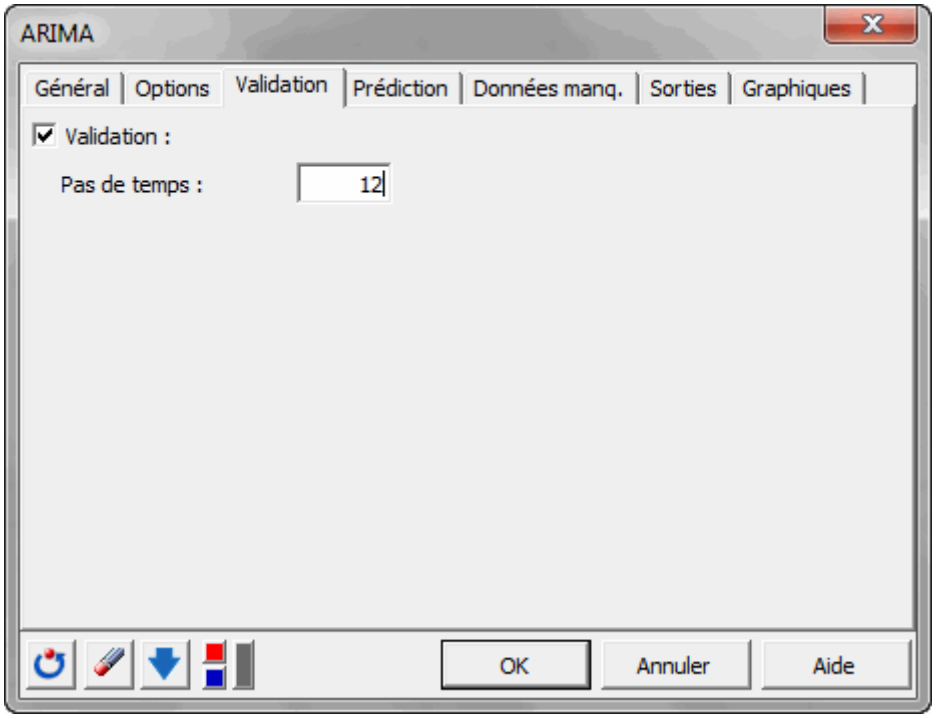
On laisse l'option **Centrer** activée afin de permettre à XLSTAT de centrer automatiquement la série.

Après avoir sélectionner la colonne des données, définissez le type de modèle ARIMA à ajuster en entrant les ordres du modèle **(p,d,q)(P,D,Q)^{s}** . La période de la série est fixée à 12 car le trafic semble connaître des cycles annuels (12 mois).

L'option **Libellés des colonnes** est activée car la première ligne de la série comprend le nom de la série.



Dans l'onglet **validation**, nous mettons la valeur **12** car nous voulons que les 12 derniers mois correspondant à l'année 1960 ne soient pas pris en compte pour l'ajustement du modèle, mais que les prévisions soient calculées pour cette période (validation du modèle).



Une fois que vous avez cliqué sur le bouton **OK**, les calculs commencent puis les résultats sont affichés.

Interpréter les résultats d'un modèle ARIMA

Le premier tableau fournit des statistiques simples pour la série sélectionné. Un tableau permettant d'évaluer la qualité du modèle après optimisation est ensuite fourni. Ces différents indices permettent éventuellement de comparer différents modèles entre eux.

Résultats après optimisation (Log(Passagers)) :

Coefficients d'ajustement :

Observations	132,000
SCE	0,156
MAPE(Diff)	185,004
MAPE	0,470
Variance BB	0,001
Variance BB (estimée)	0,001
-2Log(Vrais.)	-447,253
FPE	0,001
AIC	-441,253
AICC	-441,045
SBC	-432,916

Dans le tableau suivant sont affichés les paramètres du modèle. On note que les paramètres MA(1) et SMA(1) sont significativement différent de 0, leur intervalle de confiance à 95% ne comprenant pas la valeur 0. Les intervalles de confiance sont calculés sur la base de la matrice hessienne après optimization, comme il est proposé par la plupart des logiciels.

Le résultat asymptotique est aussi affiché afin de donner une idée de l'éloignement de la série par rapport à un cas idéal. La constante du modèle est fixée, et est une fonction de la moyenne de la série.

Paramètres du modèle :

Paramètre	Valeur	Ecart-type Hess.	inférieure	supérieure	Ecart-type asympt.	inférieure	supérieure (95
Constante	0,000	0,001	-0,002	0,002	0,031	-0,060	0,060

Paramètre	Valeur	Ecart-type Hess.	inférieure	supérieure	Ecart-type asympt.	inférieure	supérieure (95
MA(1)	-0,348	0,094	-0,533	-0,164	0,090	-0,525	-0,172
SMA(1)	-0,562	0,077	-0,714	-0,410	0,080	-0,720	-0,405

Le modèle ARIMA s'écrit alors :

$Y(t) = 0.000+Z(t-1)-0.348.Z(t-1)-0.562.Z(t-12)+0.196*Z(t-13)$ avec Z(t) est un bruit blanc $N(0, 0.001)$ $Y(t)=(1-B)(1-B^{12})X(t)$, et X(t) est la série de départ.

L'équation permettant de calculer des prévisions pour la série X(t) est : X(t+1) = Y(t+1)+X(t)+X(t-11)-X(t-12)

Après le tableau donnant les valeurs des paramètres du modèle, un tableau fournit les résultats de l'ajustement, avec la série originale et la série correspondant au modèle. En raison de contraintes liées au modèle, nous ne disposons pas de prévisions pour les treize premières valeurs. Elles sont arbitrairement fixées à la valeur de la série observée. Pour les douze dernières observations, les prévisions (Validation) du modèle sont affichées avec un intervalle de confiance.

oct-59	6,009	6,009	0,000	-0,009			
nov-59	5,892	5,872	0,020	0,542			
déc-59	6,004	5,987	0,017	0,473			
janv-60	6,033	6,039	-0,006	-0,155	0,036	5,968	6,110
févr-60	5,969	5,989	-0,020	-0,555	0,043	5,904	6,074
mars-60	6,038	6,146	-0,108	-2,971	0,049	6,049	6,242
avr-60	6,133	6,119	0,014	0,395	0,055	6,012	6,226
mai-60	6,157	6,160	-0,003	-0,077	0,060	6,043	6,276
juin-60	6,282	6,305	-0,023	-0,622	0,064	6,179	6,430
juil-60	6,433	6,433	-0,001	-0,014	0,068	6,300	6,567
août-60	6,407	6,446	-0,039	-1,084	0,072	6,305	6,588
sept-60	6,230	6,267	-0,036	-1,006	0,076	6,118	6,416
oct-60	6,133	6,136	-0,003	-0,083	0,080	5,980	6,292
nov-60	5,966	6,008	-0,042	-1,159	0,083	5,846	6,171
déc-60	6,068	6,115	-0,046	-1,274	0,086	5,946	6,284

Sur le graphique ci-dessous, on peut visuellement confirmer que les prévisions sont bien ajustées aux données.